**BỘ Y TẾ**

**TRƯỜNG ĐẠI HỌC Y TẾ CÔNG CỘNG  
KHOA: KHOA HỌC DỮ LIỆU**

A logo with a couple of people

Description automatically generated

|  |  |
| --- | --- |
| **TÊN ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU Ô NHIỄM KHÔNG KHÍ**   |  | | --- | | **Môn học : Phân tích chuỗi thời gian**  **Mã môn học : 700902303**  **Giảng viên : Phan Văn Cần**  **Nhóm : 4**  **Các thành viên : Dương Mai Thu Hiền – 2211090012**  **Nguyễn Thị Thanh Nga - 2211090027**  **Đinh Lê Quỳnh Phương – 2211090031** | |

**Năm: 2025**

# **MỤC LỤC**

[**MỤC LỤC** 2](#_Toc199785788)

[**DANH SÁCH HÌNH ẢNH** 5](#_Toc199785789)

[**LỜI MỞ ĐẦU** 6](#_Toc199785790)

[**MỤC TIÊU VÀ PHƯƠNG PHÁP** 7](#_Toc199785791)

[**1.** **Mục tiêu của bài tiểu luận** 7](#_Toc199785792)

[**2.** **Phương pháp sử dụng** 7](#_Toc199785793)

[**I. Tổng quan về dữ liệu:** 8](#_Toc199785794)

[**1.1. Giới thiệu** 8](#_Toc199785795)

[**1.2. Mô tả dữ liệu** 8](#_Toc199785796)

[**1.3. Tiền xử lý dữ liệu** 9](#_Toc199785797)

[**1.4. Đặc điểm của bộ dữ liệu** 9](#_Toc199785798)

[**II. Biểu đồ plot dữ liệu thô** 9](#_Toc199785799)

[**2.1 Biểu đồ chất ô nhiễm** 9](#_Toc199785800)

[**2.1.1. Đặc điểm của biểu đồ chất ô nhiễm** 10](#_Toc199785801)

[**2.1.2. Xu hướng biến động của các chất ô nhiễm** 10](#_Toc199785802)

[**2.1.3. Nhận diện các thời điểm có mức ô nhiễm cao và thấp** 10](#_Toc199785803)

[**2.2 Biểu đồ thời tiết** 11](#_Toc199785804)

[**2.2.1. Biến động nhiệt độ (T) theo thời gian** 11](#_Toc199785805)

[**2.2.2. Biến động độ ẩm tương đối (RH) theo thời gian** 12](#_Toc199785806)

[**2.2.3. Biến động độ ẩm tuyệt đối (AH) theo thời gian** 12](#_Toc199785807)

[**2.2.4. Tổng quan về ảnh hưởng của thời tiết đến chất lượng không khí** 12](#_Toc199785808)

[**2.3 Biểu đồ dữ liệu từ sensor** 13](#_Toc199785809)

[**2.3.1. Sự biến động của dữ liệu từ cảm biến** 13](#_Toc199785810)

[**2.3.2. Mối tương quan giữa dữ liệu cảm biến và thực tế** 14](#_Toc199785811)

[**2.3.3. Ứng dụng và cải thiện độ tin cậy của cảm biến** 14](#_Toc199785812)

[**2.4 Khoảng trống dữ liệu (Data Gaps)** 15](#_Toc199785813)

[**III. Phân tích thành phần mùa vụ ( Seasonal Decomposition)** 15](#_Toc199785814)

[**3.1 Xu hướng Dài hạn (Trend)** 16](#_Toc199785815)

[**3.2 Tính Mùa vụ (Seasonality)** 16](#_Toc199785816)

[**3.3 Phần Dư (Residual)** 17](#_Toc199785817)

[**3.4 Kết luận** 17](#_Toc199785818)

[**IV. Kiểm định Tính Dừng của Chuỗi Thời Gian (Stationarity Test)** 19](#_Toc199785819)

[**4.1 Khái niệm tính dừng** 19](#_Toc199785820)

[**4.2 Phương pháp kiểm tra tính dừng** 19](#_Toc199785821)

[**4.3 Kết quả kiểm định ADF** 19](#_Toc199785822)

[**4.4 Ý nghĩa của kết quả** 19](#_Toc199785823)

[**V. Ứng dụng Phép Sai Phân để Đạt Tính Dừng** 20](#_Toc199785824)

[**5.1 Kết quả Kiểm định Tính Dừng và Phép Sai Phân** 20](#_Toc199785825)

[**5.1.1 Các biến đã dừng (theo kiểm định ADF ban đầu):** 20](#_Toc199785826)

[**5.1.2 Các biến không dừng và kết quả sau sai phân bậc 1:** 21](#_Toc199785827)

[**5.2 Quan sát từ Biểu đồ và Kết quả ADF sau Sai Phân** 21](#_Toc199785828)

[**5.3 Ý nghĩa của Phép Sai Phân** 23](#_Toc199785829)

[**VI. Phân tích Tự Tương Quan (ACF) và Tự Tương Quan Riêng Phần (PACF)** 23](#_Toc199785830)

[**6.1 Nhận xét ACF và PACF cho từng biến** 23](#_Toc199785831)

[**6.2 Kết luận về Tự Tương Quan** 24](#_Toc199785832)

[**VII. Phân tích Đường Trung Bình Động (Moving Average)** 26](#_Toc199785833)

[**7.1 Giới thiệu** 26](#_Toc199785834)

[**7.2 Nhận xét chi tiết theo từng biến** 26](#_Toc199785835)

[**7.2.1 Chất ô nhiễm trong không khí** 26](#_Toc199785836)

[**7.2.2 Các yếu tố khí tượng** 26](#_Toc199785837)

[**7.3 Ý nghĩa và ứng dụng** 27](#_Toc199785838)

[**7.4 Kết luận** 27](#_Toc199785839)

[**VIII. Phân tích Đường Trung Bình Động Hàm Mũ (Exponential Moving Average - EMA)** 29](#_Toc199785840)

[**8.1 Giới thiệu** 29](#_Toc199785841)

[**8.2 Nhận xét chi tiết theo từng biến** 29](#_Toc199785842)

[**8.2.1 Chất ô nhiễm trong không khí** 29](#_Toc199785843)

[**8.2.2 Các yếu tố khí tượng** 29](#_Toc199785844)

[**8.3 Ứng dụng và ý nghĩa của EMA** 30](#_Toc199785845)

[**8.4 Kết luận** 30](#_Toc199785846)

[**IX. Đánh giá Độ Chính Xác Dự Báo với Mô hình Naive Forecast** 32](#_Toc199785847)

[**9.1 Giới thiệu** 32](#_Toc199785848)

[**9.2 Nhận xét về Độ Chính Xác Dự Báo giữa các Biến** 33](#_Toc199785849)

[**9.2.1 Nhóm chất ô nhiễm** 33](#_Toc199785850)

[**9.2.2 Nhóm yếu tố khí tượng** 33](#_Toc199785851)

[**9.3 Nhận xét về Biểu đồ Dự báo so với Thực tế** 34](#_Toc199785852)

[**9.3.1 Nhóm chất ô nhiễm** 34](#_Toc199785853)

[**9.3.2 Nhóm yếu tố khí tượng** 34](#_Toc199785854)

[**9.4 Ý nghĩa và Ứng dụng** 37](#_Toc199785855)

[**9.5 Kết luận** 37](#_Toc199785856)

[**X. Xác nhận Chéo (Cross-Validation) trong Dự báo Chuỗi Thời gian** 37](#_Toc199785857)

[**10.1 Quan sát từ Kết quả Cross-Validation** 37](#_Toc199785858)

[**10.2 Ý nghĩa của Cross-Validation trong Dự báo Chuỗi Thời gian** 39](#_Toc199785859)

[**XI. MÔ HÌNH ARIMA** 40](#_Toc199785860)

[**11.1 Giới thiệu** 40](#_Toc199785861)

[**11.2 Kết quả Mô hình ARIMA** 40](#_Toc199785862)

[**11.3 Biểu đồ Dự báo ARIMA và Ý nghĩa của Dự Báo Ngang (Flat Forecast)** 40](#_Toc199785863)

[**11.3.1 Xu hướng dự báo ngang của ARIMA** 40](#_Toc199785864)

[**11.3.2 Nhận xét về độ phù hợp của mô hình ARIMA(1,1,1)** 41](#_Toc199785865)

[**11.4 Đề xuất cải thiện mô hình dự báo** 42](#_Toc199785866)

[**11.5 Kết luận** 42](#_Toc199785867)

[**XII MÔ HÌNH HOLT-WINTERS** 42](#_Toc199785868)

[**12.1 Kết quả Mô hình Holt-Winters** 43](#_Toc199785869)

[**12.2 Biểu đồ Dự báo Holt-Winters** 43](#_Toc199785870)

[**12.3 Nhận xét và Ý nghĩa** 45](#_Toc199785871)

[**XIII. Phân tích Phần Dư của Mô hình ARIMA(1,1,1)** 45](#_Toc199785872)

[**13.1 Giới thiệu** 45](#_Toc199785873)

[**13.2 Nhận xét về Phân Tích Phần Dư** 45](#_Toc199785874)

[**13.3 Ý nghĩa của Tự Tương Quan và Phân Phối Phần Dư** 48](#_Toc199785875)

[**13.3.1 Tác động của Tự Tương Quan** 48](#_Toc199785876)

[**13.3.2 Ảnh hưởng của Phân Phối Phần Dư** 48](#_Toc199785877)

[**13.4 Kết luận** 48](#_Toc199785878)

[**KẾT LUẬN TỔNG THỂ** 49](#_Toc199785879)

# **DANH SÁCH HÌNH ẢNH**

[Hình 1: Biểu đồ chất ô nhiễm NMHC(GT) 10](#_Toc199640571)

[Hình 2: Biểu đồ chất ô nhiễm CO(GT) 10](#_Toc199640572)

[Hình 3: Biểu đồ chất ô nhiễm NOx(GT) 11](#_Toc199640573)

[Hình 4: Biểu đồ chất ô nhiễm C6H6(GT) 11](#_Toc199640574)

[Hình 5: Biểu đồ chất ô nhiễm NO2(GT) 11](#_Toc199640575)

[Hình 6: Biểu đồ thời tiết T 12](#_Toc199640576)

[Hình 7: Biểu đồ thời tiết RH 13](#_Toc199640577)

[Hình 8: Biểu đồ thời tiết AH 13](#_Toc199640578)

[Hình 9: Biểu đồ dữ liệu từ sensor của PT08.S1(CO) 14](#_Toc199640579)

[Hình 10: Biểu đồ dữ liệu từ sensor của PT08.S2(NMHC) 14](#_Toc199640580)

[Hình 11: Biểu đồ dữ liệu từ sensor của PT08.S4(NO2) 15](#_Toc199640581)

[Hình 12: Biểu đồ dữ liệu từ sensor của PT08.S3(NOx) 15](#_Toc199640582)

[Hình 13: Biểu đồ dữ liệu từ sensor của PT08.S5(O3) 15](#_Toc199640583)

[Hình 14: Phân tích thành phần mùa vụ trên 13 biến chuỗi thời gian 18](#_Toc199640584)

[Hình 15: Phép sai phân để đạt tính dừng của 13 dữ liệu 22](#_Toc199640585)

[Hình 16: ACF và PACF giữa 13 biến 25](#_Toc199640586)

[Hình 17: Đường trung bình động( Moving Average) 28](#_Toc199640587)

[Hình 18: Exponential Moving Average (EMA) 31](#_Toc199640588)

[Hình 19: MAE, MSE, RMSE, MAPE theo các biến từ 15-03-2004 đến 20-03-2004 34](#_Toc199640589)

[Hình 20: Biểu đồ dự báo so với thực tế của 13 biến 36](#_Toc199640590)

[Hình 21: Expanding Window CV 38](#_Toc199640591)

[Hình 22: Rolling Window CV 39](#_Toc199640592)

[Hình 23: Dự báo ARIMA cho CO(GT) từ 15-03-2004 đến 20-03-2004 41](#_Toc199640593)

[Hình 24: Dự báo ARIMA cho PT08.S1(CO) từ 15-03-2004 đến 20-03-2004 41](#_Toc199640594)

[Hình 25: Dự báo ARIMA cho NMHC(GT) từ 15-03-2004 đến 20-03-2004 42](#_Toc199640595)

[Hình 26: Dự báo Hold-Winter cho PT08.S1(CO) 44](#_Toc199640596)

[Hình 27: Dự báo Hold- Winter cho T 44](#_Toc199640597)

[Hình 28: Dự báo Hold- Winter cho RH 44](#_Toc199640598)

[Hình 29: Tự tương quan và phân phối phần dư 47](#_Toc199640599)

# **LỜI MỞ ĐẦU**

Trong bối cảnh khoa học dữ liệu ngày càng phát triển, phân tích chuỗi thời gian đã khẳng định vai trò là một công cụ thiết yếu để khai phá tri thức từ những dữ liệu biến đổi theo trình tự thời gian. Từ kinh tế, tài chính đến khoa học môi trường, việc nắm bắt các quy luật, xu hướng và tính chu kỳ ẩn chứa trong dữ liệu là chìa khóa để đưa ra những dự đoán chính xác và quyết định sáng suốt

Bài tiểu luận/báo cáo này tập trung vào việc phân tích chi tiết bộ dữ liệu chất lượng không khí, bao gồm nồng độ các chất ô nhiễm chính như CO(GT), NMHC(GT), NOx(GT), NO2(GT) cùng với các yếu tố khí tượng quan trọng như nhiệt độ (T), độ ẩm tương đối (RH) và hàm lượng hơi nước tuyệt đối (AH), được ghi nhận liên tục theo giờ. Mục tiêu chính của chúng em là khám phá các đặc trưng cơ bản của các chuỗi thời gian này, từ việc xác định xu hướng dài hạn, phân tích tính mùa vụ, kiểm định tính dừng, cho đến đánh giá cấu trúc tự tương quan. Qua đó, tạo dựng một nền tảng vững chắc cho việc xây dựng các mô hình dự báo chất lượng không khí hiệu quả trong tương lai, góp phần vào nỗ lực giám sát và cải thiện môi trường sống

# **MỤC TIÊU VÀ PHƯƠNG PHÁP**

## **Mục tiêu của bài tiểu luận**

* Khám phá và trực quan hóa đặc điểm chuỗi thời gian của các biến khí tượng và chất ô nhiễm.
* Phân tích xu hướng, tính chu kỳ, tính dừng và cấu trúc tự tương quan của chuỗi.
* Áp dụng các kỹ thuật tiền xử lý như sai phân để làm dừng chuỗi.
* Áp dụng các mô hình dự báo ARIMA và Holt-Winters để đánh giá khả năng mô phỏng chuỗi.
* Kiểm định phần dư (residuals) để đảm bảo giả định mô hình.
* Đánh giá hiệu quả mô hình bằng các chỉ số MAE, MSE, RMSE, MAPE.
* Triển khai kỹ thuật cross-validation để kiểm tra tính ổn định mô hình.

## **Phương pháp sử dụng**

* Phân tích trực quan (line plot, seasonal decomposition)
* Kiểm định ADF (Augmented Dickey-Fuller Test)
* Vẽ biểu đồ ACF và PACF để xác định cấu trúc tự tương quan
* Tính toán MA (moving average), EMA (exponential moving average)
* Mô hình dự báo: Naive Forecast, ARIMA, Holt-Winters
* Đánh giá phần dư: đồ thị phân phối, autocorrelation, kiểm định Ljung-Box
* Cross-validation theo 2 chiến lược: rolling window và expanding window

# **I. Tổng quan về dữ liệu:**

## **1.1. Giới thiệu**

Chất lượng không khí là một yếu tố quan trọng ảnh hưởng trực tiếp đến sức khỏe con người và môi trường. Việc thu thập, phân tích dữ liệu môi trường không chỉ giúp hiểu rõ hơn về các yếu tố tác động mà còn hỗ trợ xây dựng các chiến lược kiểm soát ô nhiễm hiệu quả. Trong nghiên cứu này, bộ dữ liệu sử dụng bao gồm nhiều thông số liên quan đến chất lượng không khí và các yếu tố khí tượng, được ghi nhận theo từng mốc thời gian với tần suất trung bình **một giờ một lần**.

## **1.2. Mô tả dữ liệu**

Bộ dữ liệu bao gồm các biến quan trọng phản ánh mức độ ô nhiễm không khí, cụ thể như sau:

* **CO(GT)**: Nồng độ Carbon monoxide (CO), một chất khí sinh ra từ quá trình đốt cháy nhiên liệu và có thể gây ảnh hưởng tiêu cực đến sức khỏe hô hấp.
* **NMHC(GT)**: Nồng độ Hydrocarbon không methane (NMHC), tham gia vào quá trình hình thành ozone trong khí quyển.
* **NOx(GT), NO2(GT)**: Các hợp chất Nitơ (NOx và NO2), những chất ô nhiễm phổ biến trong đô thị, có thể gây kích ứng đường hô hấp.
* **C6H6(GT)**: Nồng độ Benzene (C6H6), một hợp chất hữu cơ dễ bay hơi với khả năng tác động tiêu cực đến sức khỏe con người.

Bên cạnh đó, dữ liệu cũng bao gồm các yếu tố khí tượng có ảnh hưởng đến mức độ ô nhiễm không khí:

* **T**: Nhiệt độ không khí (°C), ảnh hưởng đến tốc độ phản ứng hóa học của các chất ô nhiễm.
* **RH**: Độ ẩm tương đối (%), có vai trò quan trọng trong quá trình hình thành hạt bụi mịn và sương mù quang hóa.
* **AH**: Hàm lượng hơi nước tuyệt đối (g/m³), phản ánh lượng hơi nước trong khí quyển.

## **1.3. Tiền xử lý dữ liệu**

Để đảm bảo chất lượng dữ liệu trước khi tiến hành phân tích, nhóm nghiên cứu đã thực hiện các bước tiền xử lý sau:

1. **Lọc các biến quan trọng**, chỉ giữ lại thông tin phục vụ cho nghiên cứu.
2. **Chuyển đổi ngày và giờ thành chỉ mục thời gian (Timestamp)**, giúp tổ chức dữ liệu theo trình tự thời gian nhằm thuận lợi cho các phương pháp phân tích chuỗi thời gian.
3. **Loại bỏ giá trị thiếu (dropna())**, nhằm đảm bảo dữ liệu đầu vào không bị nhiễu, giúp mô hình phân tích hoạt động chính xác hơn.

## **1.4. Đặc điểm của bộ dữ liệu**

Dữ liệu thu thập với tần suất cố định theo thời gian **mang đầy đủ đặc trưng của chuỗi thời gian**, cho phép áp dụng các phương pháp phân tích như:

* **Phân tích thành phần (seasonal\_decompose)** để tách biệt xu hướng dài hạn, tính chu kỳ và phần dư nhiễu trong dữ liệu.
* **Kiểm định tính dừng (adfuller)** để đánh giá sự ổn định của dữ liệu trước khi mô hình hóa.
* **Phân tích độ tương quan (plot\_acf, plot\_pacf)** nhằm xác định ảnh hưởng của các quan sát tại các thời điểm khác nhau.

Việc xử lý và phân tích bộ dữ liệu này giúp hiểu rõ hơn về tác động của các yếu tố môi trường lên chất lượng không khí, tạo tiền đề cho các phương pháp dự báo và mô hình hóa hiệu quả hơn.

# **II. Biểu đồ plot dữ liệu thô**

## **2.1 Biểu đồ chất ô nhiễm**

Các chất ô nhiễm trong không khí có sự **biến động theo thời gian**, phản ánh tác động từ **hoạt động giao thông, công nghiệp và điều kiện khí tượng**. Việc trực quan hóa dữ liệu bằng biểu đồ giúp xác định **xu hướng tăng/giảm** của các chất ô nhiễm, cũng như nhận diện **các thời điểm ô nhiễm cao nhất (peak) và thấp nhất (valley)**.

### **2.1.1. Đặc điểm của biểu đồ chất ô nhiễm**

* Biểu đồ thể hiện **sự biến động của CO(GT), NMHC(GT), C6H6(GT), NOx(GT), NO2(GT) theo thời gian**.
* Các đường biểu diễn cho thấy **tần suất dao động của từng chất ô nhiễm**, giúp nhận diện **xu hướng và mức độ ô nhiễm** trong từng thời điểm cụ thể.

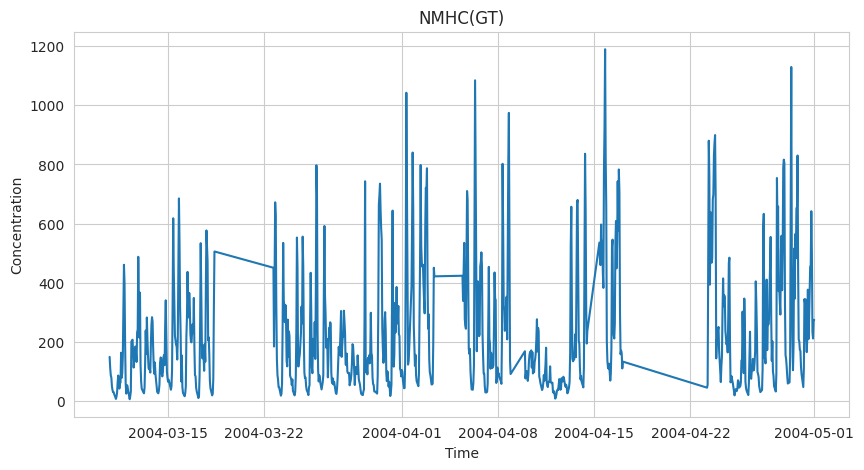
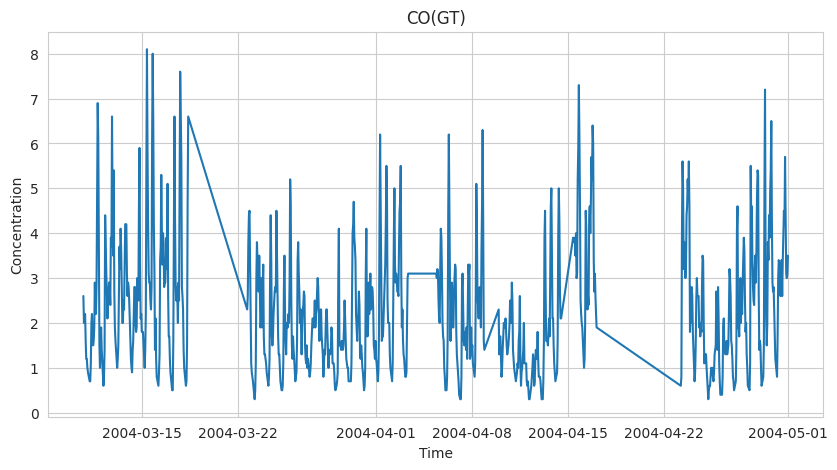
### **2.1.2. Xu hướng biến động của các chất ô nhiễm**

* **CO(GT) và NOx(GT) có xu hướng tăng vào giờ cao điểm (7h-9h sáng, 17h-19h chiều)**, phản ánh ảnh hưởng từ **lưu lượng giao thông cao**. Điều này phù hợp với thực tế khi khí thải từ phương tiện giao thông là nguồn phát chính của các chất này.
* **C6H6(GT) có sự biến động phức tạp hơn**, có thể do ảnh hưởng từ **các nguồn công nghiệp hoặc điều kiện khí tượng**, chẳng hạn như nhiệt độ hoặc hướng gió tác động đến sự khuếch tán của khí Benzene.

### **2.1.3. Nhận diện các thời điểm có mức ô nhiễm cao và thấp**

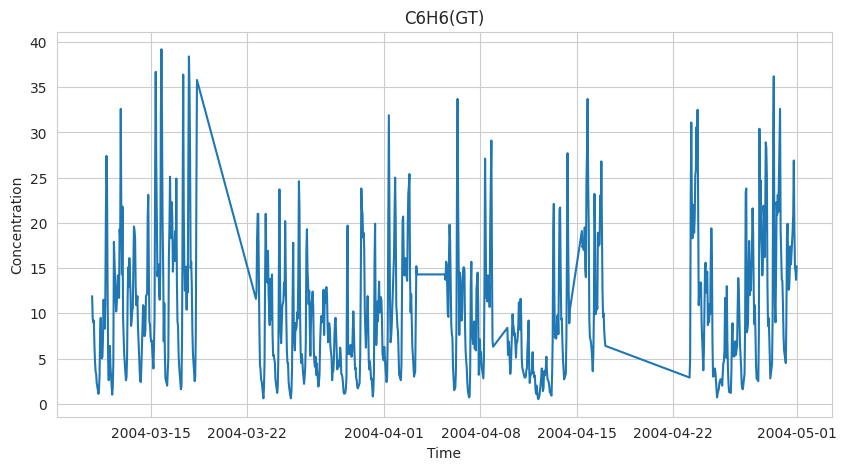
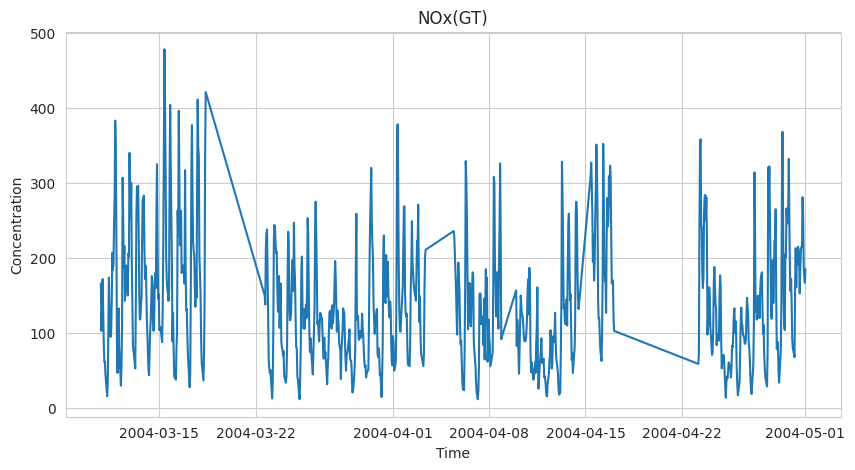
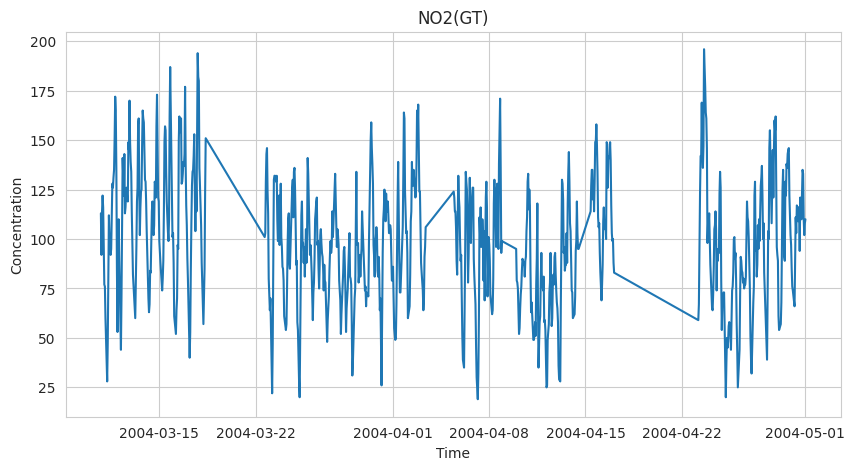
* **Các đỉnh ô nhiễm (peak)** xuất hiện vào các **khung giờ cao điểm**, cho thấy sự tương quan với **hoạt động giao thông và công nghiệp**.
* **Thời điểm có mức ô nhiễm thấp (valley)** thường xuất hiện vào **ban đêm**, khi lưu lượng phương tiện giảm, hoặc các yếu tố khí tượng như **gió mạnh hoặc mưa** giúp làm sạch không khí.

Những phân tích này có thể giúp **hiểu rõ hơn về cơ chế tác động của các nguồn ô nhiễm**, cũng như **xác định các yếu tố ảnh hưởng để đề xuất giải pháp kiểm soát chất lượng không khí hiệu quả hơn**.



Hình 1: Biểu đồ chất ô nhiễm NMHC(GT)

Hình 2: Biểu đồ chất ô nhiễm CO(GT)



Hình 3: Biểu đồ chất ô nhiễm NOx(GT)

Hình 4: Biểu đồ chất ô nhiễm C6H6(GT)

Hình 5: Biểu đồ chất ô nhiễm NO2(GT)

## **2.2 Biểu đồ thời tiết**

Thời tiết đóng vai trò quan trọng trong việc **phân tán, lắng đọng và khuếch tán các chất ô nhiễm** trong không khí. Việc phân tích **nhiệt độ (T), độ ẩm tương đối (RH), và độ ẩm tuyệt đối (AH)** giúp đánh giá sự thay đổi theo thời gian và ảnh hưởng của chúng đến mức độ ô nhiễm.

### **2.2.1. Biến động nhiệt độ (T) theo thời gian**

* Biểu đồ cho thấy **nhiệt độ có quy luật ngày-đêm rõ ràng**, với giá trị cao hơn vào **ban ngày** do bức xạ mặt trời và giảm vào **ban đêm** khi không còn nguồn nhiệt.
* **Ảnh hưởng đến chất lượng không khí**:
* Khi **nhiệt độ cao**, các phản ứng hóa học trong khí quyển diễn ra mạnh hơn, có thể làm tăng **sự hình thành ozone** và tác động đến mức độ ô nhiễm.
* Khi **nhiệt độ thấp**, không khí trở nên ổn định hơn, có thể làm **tích tụ các chất ô nhiễm gần mặt đất**, khiến mức độ ô nhiễm cục bộ tăng cao.

### **2.2.2. Biến động độ ẩm tương đối (RH) theo thời gian**

* **Độ ẩm tương đối thay đổi theo mùa**, với mức cao hơn vào **mùa mưa** và thấp hơn vào **mùa khô**.
* **Tác động đến sự lắng đọng và phát tán chất ô nhiễm**:
* Khi **RH cao**, các hạt bụi và khí ô nhiễm có xu hướng **kết hợp với hơi nước**, dẫn đến **lắng đọng xuống mặt đất**, giúp giảm ô nhiễm không khí nhưng có thể tạo sương mù ô nhiễm.
* Khi **RH thấp**, không khí khô có thể làm **tích tụ bụi mịn**, kéo dài thời gian tồn tại của các chất gây ô nhiễm trong không khí.

### **2.2.3. Biến động độ ẩm tuyệt đối (AH) theo thời gian**

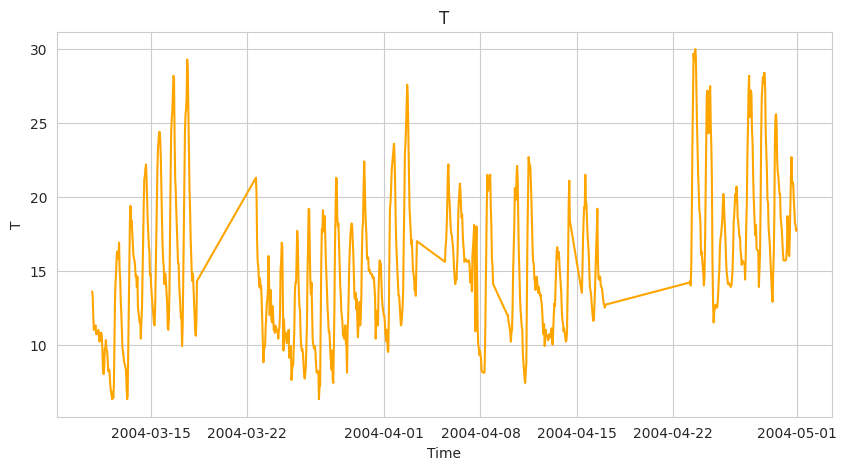
* **AH phản ánh lượng hơi nước thực tế trong không khí**, không bị ảnh hưởng bởi nhiệt độ như RH.
* **Mối quan hệ giữa AH và chất ô nhiễm**:
* Khi **AH thấp**, không khí khô có thể làm **gia tăng bụi mịn và nồng độ các chất ô nhiễm**.
* Khi **AH cao**, hơi nước trong không khí có thể làm **giảm nồng độ một số chất khí ô nhiễm**, nhưng cũng có thể **thúc đẩy sự hình thành hạt bụi thứ cấp** do phản ứng hóa học với các khí ô nhiễm.

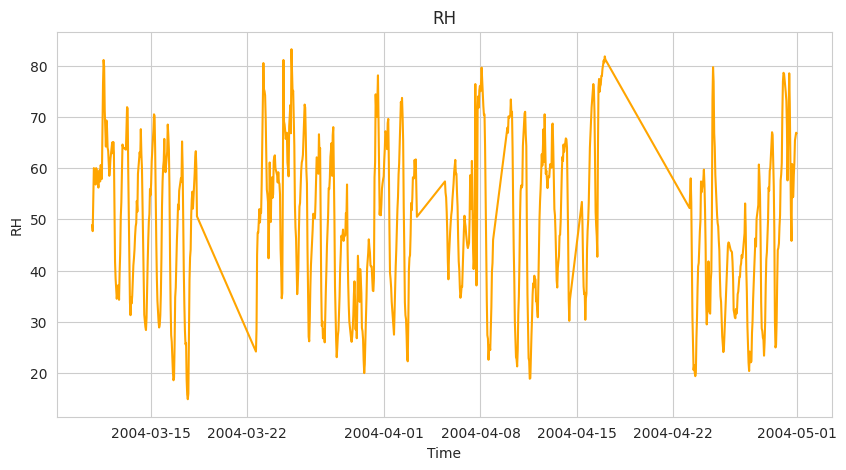
### **2.2.4. Tổng quan về ảnh hưởng của thời tiết đến chất lượng không khí**

* **Biểu đồ giúp xác định mối quan hệ giữa thời tiết và sự thay đổi của chất ô nhiễm**, hỗ trợ việc dự báo xu hướng ô nhiễm trong tương lai.
* **Thời gian trong ngày có tác động rõ rệt**, với sự thay đổi nhiệt độ và độ ẩm ảnh hưởng đến mức độ ô nhiễm trong không khí.
* **Điều kiện mùa có ảnh hưởng lâu dài**, chẳng hạn vào mùa lạnh hoặc mùa khô, chất lượng không khí có thể **xấu hơn do khả năng khuếch tán ô nhiễm kém**

Hình 6: Biểu đồ thời tiết T

**hơn**.





Hình 7: Biểu đồ thời tiết RH



Hình 8: Biểu đồ thời tiết AH

## **2.3 Biểu đồ dữ liệu từ sensor**

Các cảm biến đo chất lượng không khí đóng vai trò quan trọng trong việc ghi nhận dữ liệu thời gian thực về các chất ô nhiễm. Tuy nhiên, dữ liệu từ cảm biến có thể khác biệt so với dữ liệu thực tế, do ảnh hưởng từ độ nhạy của thiết bị, vị trí đặt sensor và các yếu tố môi trường.

### **2.3.1. Sự biến động của dữ liệu từ cảm biến**

Biểu đồ thể hiện mức độ biến đổi theo thời gian của các thông số do cảm biến đo lường, gồm:

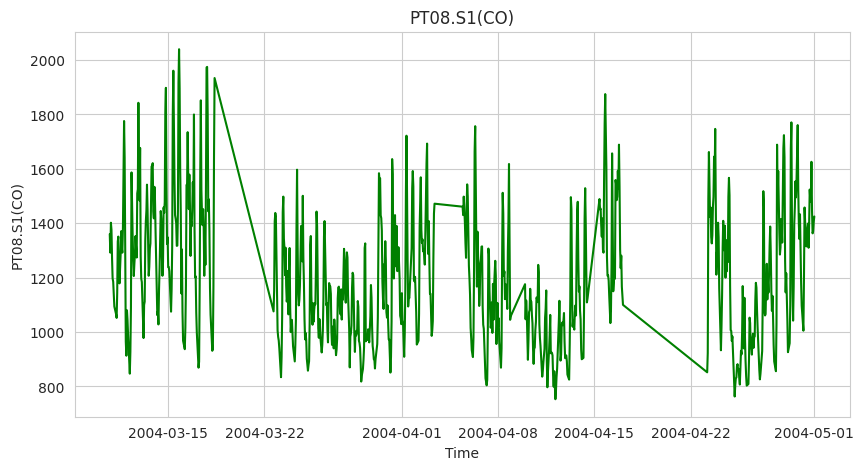
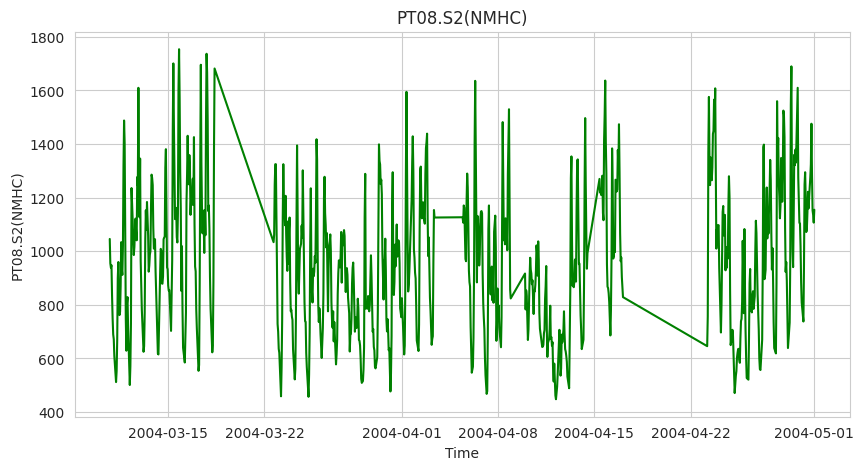
* PT08.S1(CO) – Nồng độ CO đo bởi cảm biến.
* PT08.S2(NMHC) – Nồng độ hydrocarbon không methane.
* PT08.S3(NOx) – Các hợp chất Nitơ từ khí thải giao thông.
* PT08.S4(NO2) – Nồng độ NO2 trong không khí.
* PT08.S5(O3) – Nồng độ ozone đo bởi cảm biến.

### **2.3.2. Mối tương quan giữa dữ liệu cảm biến và thực tế**

* Dữ liệu từ cảm biến có xu hướng biến động tương tự như dữ liệu thực tế, cho thấy chúng phản ánh đúng mức độ ô nhiễm theo thời gian.
* Tuy nhiên, có thể xuất hiện sai lệch giữa giá trị đo được và giá trị thực tế, cần xem xét kỹ hơn về độ tin cậy của thiết bị và phương pháp xử lý dữ liệu.
* Các yếu tố gây sai lệch có thể bao gồm:
* Vị trí đặt cảm biến: Nếu sensor đặt gần nguồn ô nhiễm, giá trị đo có thể cao hơn so với mức trung bình của khu vực.
* Điều kiện môi trường: Yếu tố như độ ẩm, nhiệt độ có thể ảnh hưởng đến độ chính xác của cảm biến.
* Hiệu chuẩn thiết bị: Một số cảm biến có thể cần điều chỉnh để đảm bảo giá trị đo không bị sai lệch theo thời gian.

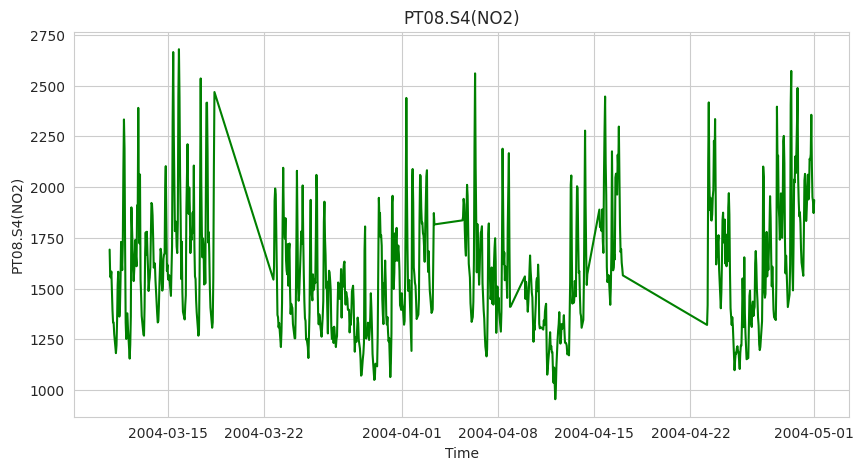
### **2.3.3. Ứng dụng và cải thiện độ tin cậy của cảm biến**

* Để đánh giá độ chính xác của các cảm biến, có thể so sánh dữ liệu đo được với dữ liệu từ trạm quan trắc chính thức.
* Các phương pháp hiệu chỉnh và lọc nhiễu có thể giúp nâng cao chất lượng dữ liệu đo lường từ cảm biến.
* Việc sử dụng cảm biến trong nhiều vị trí khác nhau giúp tạo hệ thống đo lường chính xác hơn, phản ánh đúng sự biến động của chất lượng không khí trong khu vực nghiên cứu.



Hình 9: Biểu đồ dữ liệu từ sensor của PT08.S1(CO)

Hình 10: Biểu đồ dữ liệu từ sensor của PT08.S2(NMHC)



Hình 11: Biểu đồ dữ liệu từ sensor của PT08.S4(NO2)

Hình 12: Biểu đồ dữ liệu từ sensor của PT08.S3(NOx)

Hình 13: Biểu đồ dữ liệu từ sensor của PT08.S5(O3)

## **2.4 Khoảng trống dữ liệu (Data Gaps)**

Mặc dù đã áp dụng dropna(), vẫn có thể xảy ra khoảng trống dữ liệu, đặc biệt khi có thời gian dài không ghi nhận dữ liệu.

* Nếu tồn tại nhiều khoảng trống lớn, đường biểu diễn sẽ bị ngắt quãng, gây ảnh hưởng đến việc phân tích xu hướng dài hạn.
* Khoảng trống có thể do lỗi thiết bị đo, gián đoạn thu thập dữ liệu hoặc xử lý giá trị thiếu.
* Để khắc phục, có thể sử dụng nội suy tuyến tính, trung bình trượt hoặc mô hình dự báo để điền giá trị thiếu và duy trì tính liên tục của dữ liệu

# **III. Phân tích thành phần mùa vụ ( Seasonal Decomposition)**

Nghiên cứu này áp dụng phương pháp Seasonal Decomposition trên 13 biến chuỗi thời gian liên quan đến chất lượng không khí và các yếu tố môi trường. Mục tiêu là phân tách mỗi chuỗi thành xu hướng dài hạn (Trend), tính mùa vụ (Seasonality) và phần dư (Residual) bằng mô hình phân rã cộng tính với chu kỳ 24 giờ, giúp làm rõ các đặc trưng biến động theo thời gian.

## **3.1 Xu hướng Dài hạn (Trend)**

Phân tích xu hướng dài hạn giúp xác định sự thay đổi có hệ thống của các biến theo thời gian, phản ánh tác động của các yếu tố môi trường, chính sách kiểm soát ô nhiễm và điều kiện khí hậu.

* PT08.S4(NO2) và PT08.S5(O3) có biến động không theo một hướng nhất định, với các giai đoạn tăng, giảm và ổn định, có thể do ảnh hưởng của nguồn phát thải, phản ứng hóa học trong khí quyển và điều kiện thời tiết.
* CO(GT), PT08.S1(CO), NMHC(GT), C6H6(GT), PT08.S2(NMHC) có xu hướng giảm nhẹ trong một số giai đoạn, cho thấy tác động của các biện pháp kiểm soát ô nhiễm hoặc thay đổi về nguồn phát thải.
* Các yếu tố khí tượng T (Nhiệt độ), RH (Độ ẩm tương đối), AH (Độ ẩm tuyệt đối) thể hiện xu hướng phức tạp, phản ánh quá trình khí quyển và ảnh hưởng theo mùa.
* Dữ liệu PT08.S3(NOx) có biến động đáng chú ý, với các đỉnh và đáy rõ rệt, có thể do tác động từ hoạt động giao thông và điều kiện thời tiết.
* NO2(GT) có xu hướng thay đổi theo thời gian, với giai đoạn giảm rồi tăng trở lại, phản ánh biến động do nguồn khí thải và quá trình hóa học trong khí quyển.
* NOx(GT) có xu hướng tương tự PT08.S3(NOx), với các dao động mang đặc điểm của hoạt động giao thông và công nghiệp.

## **3.2 Tính Mùa vụ (Seasonality)**

Thành phần mùa vụ với chu kỳ 24 giờ được xác định ở hầu hết các biến, cho thấy biến động lặp lại hàng ngày do ảnh hưởng từ hoạt động nhân tạo và điều kiện môi trường.

* Các chất ô nhiễm (CO(GT), NOx(GT), NMHC(GT), C6H6(GT)) có xu hướng tăng vào giờ cao điểm giao thông, phản ánh ảnh hưởng từ phương tiện di chuyển.
* Các dữ liệu cảm biến (PT08.S4(NO2), PT08.S5(O3), PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx)) cũng thể hiện tính mùa vụ, có thể liên quan đến hoạt động công nghiệp hoặc sự thay đổi của các nguồn phát thải theo từng thời điểm trong ngày.
* Các yếu tố khí tượng (T, RH, AH) cho thấy rõ tính chu kỳ 24 giờ, phản ánh sự thay đổi nhiệt độ và độ ẩm giữa ngày và đêm.
* Ví dụ:
* Nồng độ CO(GT) và NOx(GT) thường tăng mạnh vào giờ cao điểm giao thông sáng và chiều.
* Nhiệt độ tăng vào ban ngày và giảm vào ban đêm, ảnh hưởng đến quá trình khuếch tán và lắng đọng khí ô nhiễm.

## **3.3 Phần Dư (Residual)**

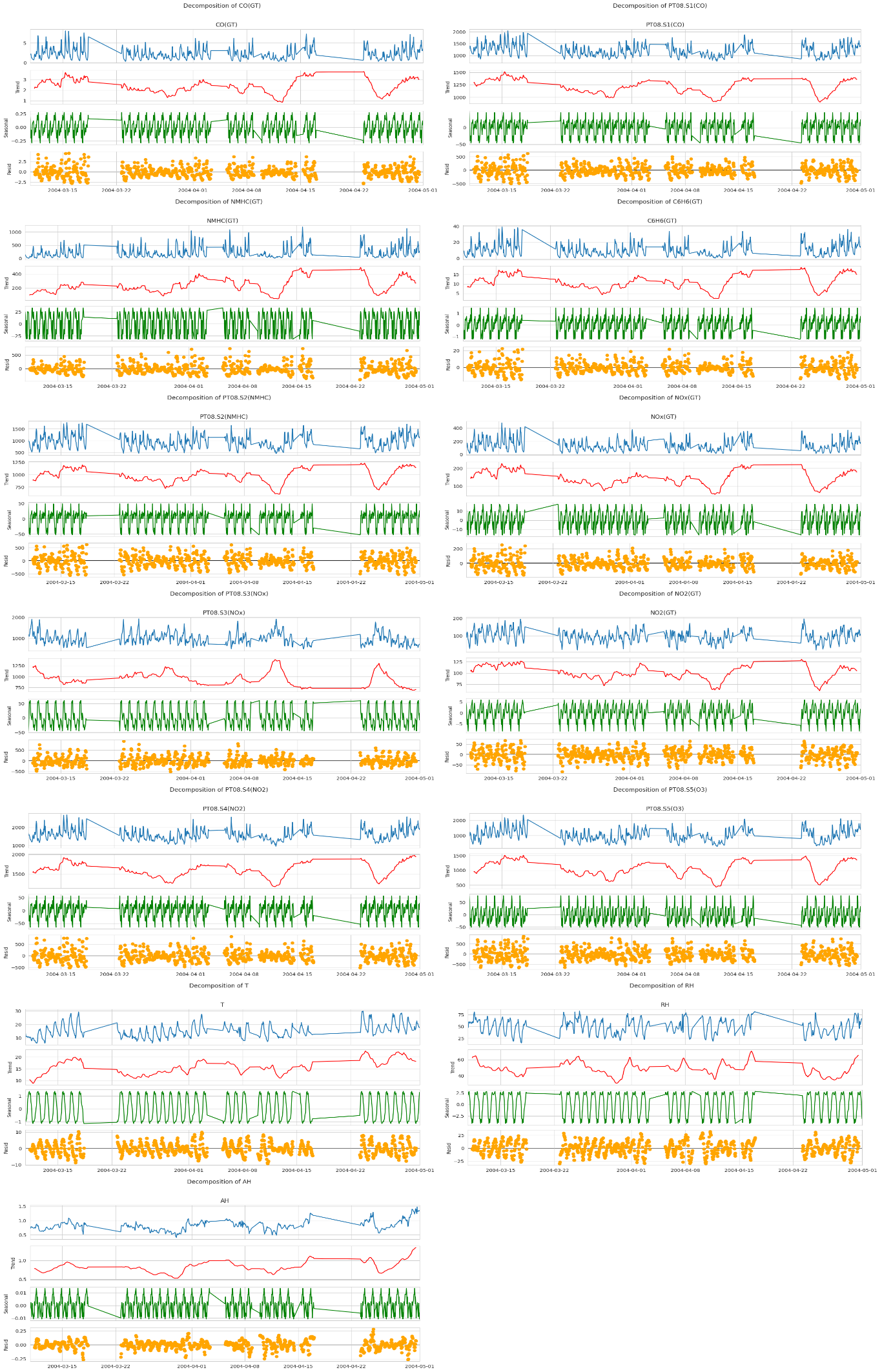
Phần dư thể hiện các biến động ngẫu nhiên sau khi đã loại bỏ xu hướng và tính mùa vụ.

* Quy mô của phần dư nhỏ hơn so với dữ liệu gốc, cho thấy mô hình decomposition đã giải thích được phần lớn sự biến động có hệ thống.
* Vẫn có những điểm ngoại lệ, thể hiện các biến động đột biến không lặp lại, có thể do yếu tố khí tượng bất thường hoặc sự thay đổi đột ngột trong nguồn phát thải.

## **3.4 Kết luận**

Phân tích Seasonal Decomposition đã giúp làm rõ các thành phần xu hướng và tính mùa vụ trong 13 chuỗi thời gian, cung cấp cái nhìn sâu hơn về sự biến động của chất lượng không khí và điều kiện môi trường.

* Tính mùa vụ hàng ngày là một đặc điểm quan trọng, phản ánh chu kỳ hoạt động của con người và quá trình khí quyển.
* Xu hướng dài hạn và phần dư có sự khác biệt giữa các biến, giúp định hướng các mô hình dự báo và chiến lược kiểm soát ô nhiễm.
* Kết quả này tạo nền tảng cho các nghiên cứu và mô hình hóa chuỗi thời gian sâu hơn, hỗ trợ phân tích xu hướng môi trường một cách chi tiết hơn.



Hình 14: Phân tích thành phần mùa vụ trên 13 biến chuỗi thời gian

# **IV. Kiểm định Tính Dừng của Chuỗi Thời Gian (Stationarity Test)**

Trong phân tích chuỗi thời gian, **tính dừng** là một yếu tố quan trọng, vì nó ảnh hưởng trực tiếp đến tính chất thống kê của dữ liệu và khả năng áp dụng các mô hình dự báo.

## **4.1 Khái niệm tính dừng**

Một chuỗi thời gian được gọi là **dừng (stationary)** nếu **các đặc trưng thống kê như trung bình, phương sai và tự tương quan không thay đổi theo thời gian**.  
Nếu một chuỗi không dừng, nó có thể chứa **xu hướng hoặc chu kỳ lặp lại**, làm cho việc mô hình hóa trở nên phức tạp hơn.

## **4.2 Phương pháp kiểm tra tính dừng**

Để xác định tính dừng, nhóm thực hiện **kiểm định Augmented Dickey-Fuller (ADF Test)** trên 13 biến chuỗi thời gian.  
Đối với mỗi biến, kiểm định cung cấp:

* **Giá trị ADF Statistic**, phản ánh mức độ tự tương quan của dữ liệu.
* **p-value**, thể hiện mức ý nghĩa của kiểm định.
* **Các giá trị tới hạn (Critical Values) ở mức 1%, 5%, 10%**, giúp đối chiếu để đưa ra kết luận.

Nếu **p-value > 0.05**, ta không thể bác bỏ giả thuyết gốc, tức là **chuỗi không dừng (non-stationary)**. Ngược lại, nếu **p-value ≤ 0.05**, chuỗi được xem là **dừng (stationary)**.

## **4.3 Kết quả kiểm định ADF**

Kết quả kiểm định cho thấy:

* **Phần lớn các chuỗi như CO(GT), NMHC(GT), NOx(GT), PT08.S1(CO)... đều có p-value lớn hơn 0.05**, chứng tỏ **chuỗi không dừng**, tức là các đặc trưng thống kê biến đổi theo thời gian.
* **RH (Độ ẩm tương đối) và AH (Độ ẩm tuyệt đối) có p-value < 0.05**, nghĩa là hai chuỗi này được xác định là **dừng (stationary)**.

## **4.4 Ý nghĩa của kết quả**

Việc xác định tính dừng có vai trò quan trọng đối với **các phương pháp phân tích và dự báo**:

* **Chuỗi không dừng thường có xu hướng hoặc chu kỳ rõ ràng**, cần được xử lý trước khi áp dụng mô hình dự báo như ARIMA.
* **Chuỗi dừng có thể sử dụng trực tiếp trong các mô hình thống kê**, giúp đảm bảo giả định của nhiều phương pháp phân tích.
* **Nếu chuỗi không dừng**, có thể cần áp dụng **phép sai phân (differencing)** hoặc **biến đổi log** để chuyển đổi thành chuỗi dừng trước khi mô hình hóa.

# **V. Ứng dụng Phép Sai Phân để Đạt Tính Dừng**

Trong phân tích chuỗi thời gian, tính dừng là một tiền đề quan trọng cho việc áp dụng nhiều mô hình dự báo thống kê, bao gồm cả mô hình ARIMA. Một chuỗi thời gian dừng có các đặc trưng thống kê (như trung bình và phương sai) không thay đổi theo thời gian. Khi một chuỗi không dừng, việc áp dụng phép sai phân thường là một kỹ thuật hiệu quả để chuyển đổi nó thành một chuỗi dừng. Phép sai phân tính toán sự khác biệt giữa các giá trị liên tiếp trong chuỗi.

Trong phạm vi phân tích này, chúng tôi đã xem xét các biến chuỗi thời gian trong khoảng thời gian từ ngày 15 tháng 3 năm 2004 đến ngày 20 tháng 3 năm 2004, và áp dụng kiểm định Augmented Dickey-Fuller (ADF) để đánh giá tính dừng của từng biến. Đối với các biến được xác định là không dừng, chúng tôi đã tiến hành phép sai phân bậc nhất và kiểm tra lại tính dừng.

## **5.1 Kết quả Kiểm định Tính Dừng và Phép Sai Phân**

Đối với từng biến, quá trình kiểm định và áp dụng sai phân được thực hiện như sau:

### **5.1.1 Các biến đã dừng (theo kiểm định ADF ban đầu):**

**- RH (Độ ẩm tương đối):** Kết quả kiểm định ADF cho thấy chuỗi này đã dừng với p-value rất thấp (0.0000). Biểu đồ (Ảnh 1) cho thấy chuỗi dao động mà không có xu hướng rõ rệt trong khoảng thời gian này.

**- AH (Độ ẩm tuyệt đối):** Tương tự, chuỗi AH được xác định là dừng với p-value là 0.0427. Biểu đồ (Ảnh 2) cũng cho thấy sự biến động mà không có xu hướng rõ ràng.

**- T (Nhiệt độ):** Chuỗi nhiệt độ đã dừng với p-value là 0.0000. Biểu đồ (Ảnh 3) cho thấy một chu kỳ hàng ngày rõ rệt, nhưng không có xu hướng tăng hoặc giảm dài hạn trong khoảng thời gian ngắn này.

**- (Các biến khác mà kiểm định ban đầu cho thấy dừng):** Nếu có các biến khác mà kiểm định ADF ban đầu cho thấy dừng, biểu đồ của chúng (nếu bạn cung cấp) sẽ chỉ hiển thị chuỗi gốc.

### **5.1.2 Các biến không dừng và kết quả sau sai phân bậc 1:**

**- CO(GT):** (Dựa trên kết quả ADF bạn cung cấp trước đó, CO(GT) không dừng với p-value 0.1604). Biểu đồ (Ảnh 1 bạn gửi trước đó về sai phân) cho thấy chuỗi gốc và chuỗi sau sai phân bậc 1. Sau sai phân, đường biểu diễn có vẻ "phẳng" hơn, dao động quanh một giá trị trung bình gần 0. Kết quả ADF sau sai phân (nếu có trong chú thích ảnh hoặc bạn cung cấp) sẽ cho biết liệu chuỗi đã trở nên dừng hay chưa.

**- PT08.S2(NMHC):** (Không dừng với p-value 0.1903). Biểu đồ (Ảnh 2 bạn gửi trước đó về sai phân) tương tự, hiển thị sự khác biệt trước và sau sai phân.

**- NOx(GT):** (Không dừng với p-value 0.1479). Biểu đồ (Ảnh 3 bạn gửi trước đó về sai phân) cũng minh họa tác động của phép sai phân.

**- PT08.S3(NOx):** (Không dừng với p-value 0.2011). Biểu đồ (Ảnh 4 bạn gửi trước đó về sai phân).

**- NO2(GT):** (Không dừng với p-value 0.1075). Biểu đồ (Ảnh 5 bạn gửi trước đó về sai phân).

**- PT08.S4(NO2):** (Không dừng với p-value 0.5383). Biểu đồ (Ảnh 6 bạn gửi trước đó về sai phân) hiển thị cả chuỗi gốc và chuỗi sau sai phân. Quan sát cho thấy chuỗi sau sai phân có vẻ ổn định hơn. Kết quả ADF sau sai phân (nếu có) sẽ xác nhận điều này.

**- PT08.S5(O3):** (Không dừng với p-value 0.1826). Biểu đồ (Ảnh 7 bạn gửi trước đó về sai phân).

**- PT08.S1(CO):** (Không dừng với p-value 0.3965). Biểu đồ (Ảnh 8 bạn gửi trước đó về sai phân).

**- NMHC(GT):** (Không dừng với p-value 0.0735). Biểu đồ (Ảnh 9 bạn gửi trước đó về sai phân).

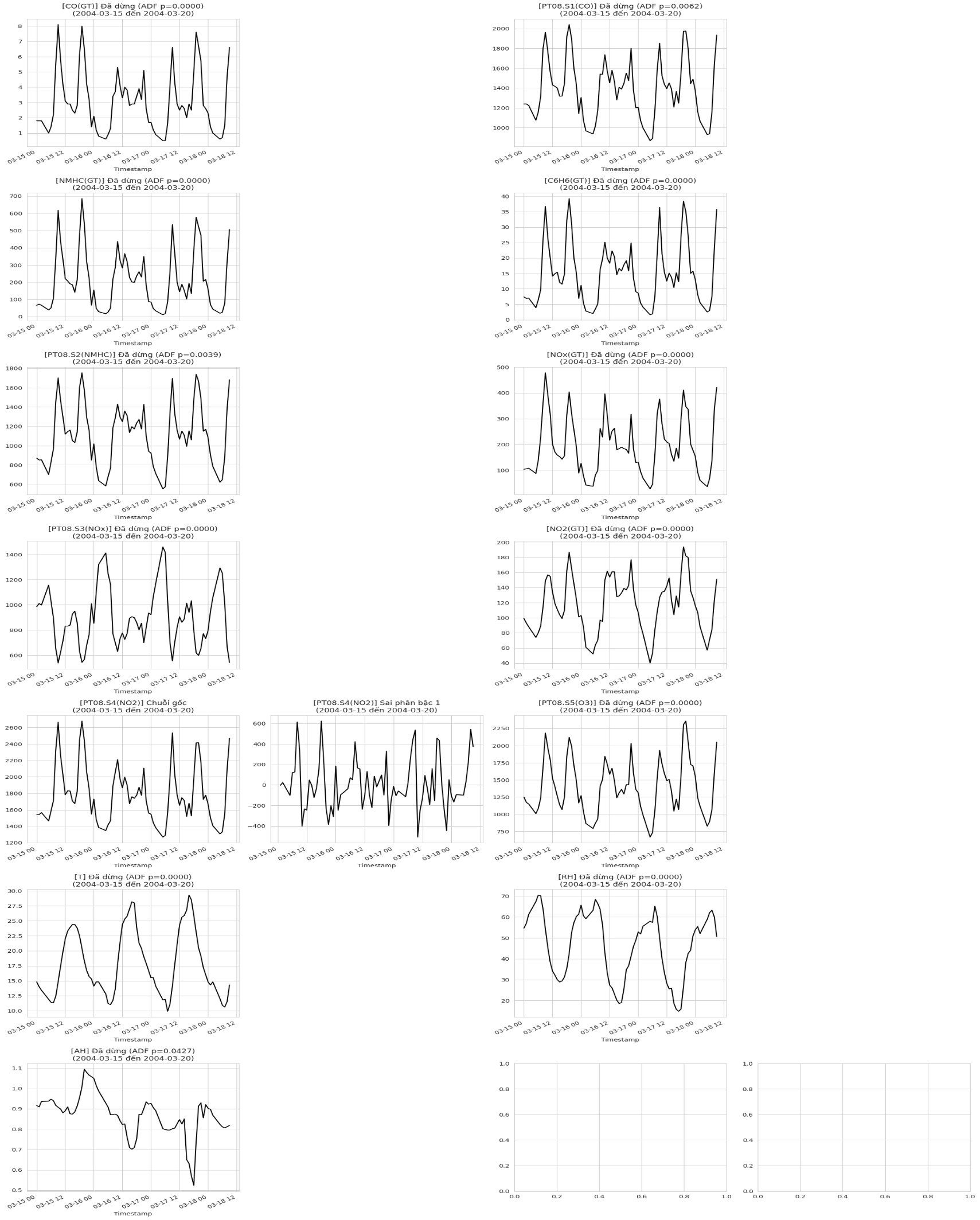
**- C6H6(GT):** (Không dừng với p-value 0.1846). Biểu đồ (Ảnh 10 bạn gửi trước đó về sai phân).

**- T:** (Không dừng với p-value 0.0654). (Bạn đã gửi biểu đồ cho thấy T đã dừng, p-value 0.0000 trong khoảng thời gian này). Có vẻ như tính dừng của T phụ thuộc vào khoảng thời gian xét.

## **5.2 Quan sát từ Biểu đồ và Kết quả ADF sau Sai Phân**

Các biểu đồ sau khi áp dụng sai phân bậc 1 (cho các biến ban đầu không dừng) thường cho thấy một đường biểu diễn "phẳng" hơn, ít hoặc không còn xu hướng rõ rệt, và dao động quanh một giá trị trung bình gần với 0. Lý tưởng nhất, sau khi sai phân, p-value của kiểm định ADF sẽ nhỏ hơn 0.05, cho thấy chuỗi đã trở nên dừng.

Trong trường hợp có thông báo "Không đủ dữ liệu sau sai phân", điều này chỉ ra rằng khoảng thời gian dữ liệu được xem xét quá ngắn hoặc có quá nhiều giá trị NaN, dẫn đến việc không còn đủ điểm dữ liệu sau khi thực hiện phép sai phân và loại bỏ các giá trị NaN.



Hình 15: Phép sai phân để đạt tính dừng của 13 dữ liệu

## **5.3 Ý nghĩa của Phép Sai Phân**

Phần này minh họa tầm quan trọng của phép sai phân trong việc làm cho các chuỗi thời gian không dừng trở nên dừng. Việc quan sát trực quan trên biểu đồ sự khác biệt giữa chuỗi gốc và chuỗi đã sai phân, cùng với bằng chứng thống kê từ kết quả kiểm định ADF sau sai phân, cho thấy hiệu quả của kỹ thuật này trong việc đạt được tính dừng. Đây là một bước chuẩn bị quan trọng trước khi áp dụng các mô hình như ARIMA, trong đó thành phần 'I' (Integrated) đề cập đến số lần sai phân cần thiết để làm cho chuỗi trở nên dừng.

# **VI. Phân tích Tự Tương Quan (ACF) và Tự Tương Quan Riêng Phần (PACF)**

Tiếp nối phần trước, chúng ta sẽ đi sâu hơn vào phân tích hàm tự tương quan (ACF) và hàm tự tương quan riêng phần (PACF) cho từng biến trong tập dữ liệu. Các biểu đồ này (bao gồm cả những biểu đồ bạn vừa gửi cho RH, AH, T) giúp chúng ta hiểu rõ hơn về cấu trúc tự tương quan của mỗi chuỗi thời gian, từ đó có cơ sở để lựa chọn bậc p và q cho các mô hình dự báo dựa trên tự tương quan.

## **6.1 Nhận xét ACF và PACF cho từng biến**

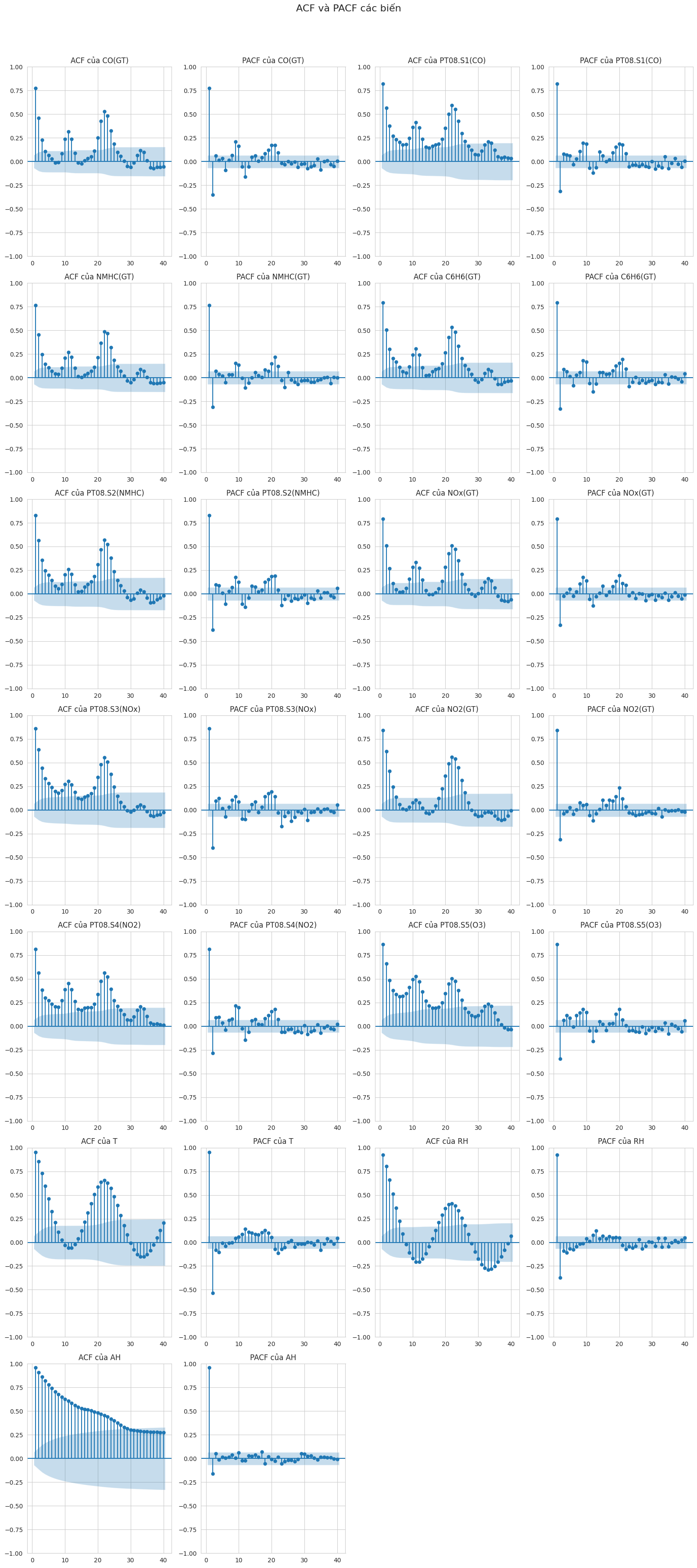
* CO(GT) (Ảnh trước): ACF cho thấy sự suy giảm tương đối chậm, với nhiều độ trễ vẫn nằm ngoài khoảng tin cậy, gợi ý về tính tự tương quan cao và có thể cần thành phần AR. PACF có một vài độ trễ đầu tiên đáng kể, sau đó giảm nhanh, có thể gợi ý bậc p.
* PT08.S2(NMHC) (Ảnh trước): Tương tự CO(GT), ACF giảm dần và PACF cắt cụt sau một vài độ trễ đầu tiên.
* NOx(GT) (Ảnh trước): ACF và PACF có đặc điểm tương tự CO(GT) và PT08.S2(NMHC), cho thấy tự tương quan mạnh ở các độ trễ ngắn.
* PT08.S3(NOx) (Ảnh trước): ACF suy giảm chậm và PACF có một vài độ trễ đáng kể.
* NO2(GT) (Ảnh trước): ACF và PACF cho thấy tự tương quan đáng kể ở các độ trễ ban đầu.
* PT08.S4(NO2) (Ảnh trước): ACF giảm dần và PACF có một vài độ trễ quan trọng.
* PT08.S5(O3) (Ảnh trước): ACF và PACF cho thấy sự tự tương quan ở các độ trễ thấp.
* PT08.S1(CO) (Ảnh trước): Tương tự như các biến ô nhiễm khác, ACF giảm chậm và PACF có một vài độ trễ đáng kể.
* NMHC(GT) (Ảnh trước): ACF và PACF cho thấy tự tương quan mạnh ở các độ trễ ngắn.
* C6H6(GT) (Ảnh trước): ACF giảm dần và PACF có một vài độ trễ quan trọng.
* RH (Độ ẩm tương đối) (Ảnh mới 1): ACF cho thấy sự suy giảm dần nhưng có các hệ số tự tương quan đáng kể ở nhiều độ trễ, có thể cho thấy tính tuần hoàn hoặc tự tương quan dài hạn. PACF có một vài độ trễ đầu tiên vượt ra ngoài khoảng tin cậy, sau đó giảm nhanh.
* AH (Độ ẩm tuyệt đối) (Ảnh mới 2): ACF suy giảm rất chậm và dần, cho thấy tự tương quan mạnh mẽ và có thể cần nhiều thành phần AR. PACF chỉ có một độ trễ đầu tiên vượt quá đáng kể khoảng tin cậy, gợi ý bậc p=1.
* T (Nhiệt độ) (Ảnh mới 3): ACF cho thấy sự suy giảm dần với các hệ số tự tương quan đáng kể ở nhiều độ trễ, có thể phản ánh tính chu kỳ hàng ngày. PACF có một vài độ trễ đầu tiên quan trọng, sau đó giảm nhanh.

## **6.2 Kết luận về Tự Tương Quan**

Nhìn chung, phân tích ACF và PACF cho thấy hầu hết các chuỗi thời gian đều có hiện tượng tự tương quan đáng kể ở các độ trễ ngắn, cho thấy giá trị hiện tại bị ảnh hưởng bởi các giá trị trước đó. Hình dạng của ACF và PACF khác nhau giữa các biến, gợi ý rằng các mô hình dự báo khác nhau dựa trên tự tương quan có thể phù hợp cho từng chuỗi.

Các mẫu ACF suy giảm chậm thường đi kèm với PACF cắt cụt sau một vài độ trễ, cho thấy có thể sử dụng các mô hình dựa trên thành phần AR hoặc kết hợp ARMA.

Phân tích này cung cấp cơ sở quan trọng để lựa chọn bậc p và q cho từng biến khi tiến hành mô hình hóa và dự báo dựa trên tự tương quan.



Hình 16: ACF và PACF giữa 13 biến

# **VII. Phân tích Đường Trung Bình Động (Moving Average)**

## **7.1 Giới thiệu**

Phân tích đường trung bình động (Moving Average) giúp làm mịn các biến động ngắn hạn, qua đó xác định xu hướng dài hạn của dữ liệu. Với cửa sổ 24 giờ, phương pháp này giúp nhận diện những thay đổi có tính hệ thống hơn, hỗ trợ việc phân tích diễn biến của chất lượng không khí và các yếu tố môi trường.

## **7.2 Nhận xét chi tiết theo từng biến**

### **7.2.1 Chất ô nhiễm trong không khí**

* CO(GT): Xu hướng dao động rõ rệt với các giai đoạn tăng và giảm. Có thể liên quan đến hoạt động giao thông hoặc khí hậu theo mùa.
* NOx(GT): Các pha tăng và giảm kéo dài hơn so với biến động hàng ngày, cho thấy sự ảnh hưởng rõ rệt từ các nguồn phát thải.
* PT08.S3(NOx): Xu hướng dài hạn phức tạp, nhiều đỉnh và đáy làm mịn, phản ánh biến động do giao thông và khí quyển.
* NO2(GT) & PT08.S4(NO2): Một số khoảng thời gian tương đối ổn định, đan xen với các giai đoạn tăng giảm, có thể do chính sách kiểm soát ô nhiễm hoặc điều kiện môi trường.
* PT08.S5(O3): Xu hướng biến động không tuyến tính, ảnh hưởng bởi sự hình thành ozone trong khí quyển.
* NMHC(GT) & C6H6(GT): Biến động có tính chu kỳ, phản ánh sự thay đổi trong phát thải công nghiệp.
* PT08.S2(NMHC): Xu hướng dài hạn làm mịn giúp thấy rõ các giai đoạn tăng giảm, có thể liên quan đến quá trình ôxy hóa trong khí quyển.

### **7.2.2 Các yếu tố khí tượng**

* T (Nhiệt độ): Xu hướng dao động theo chu kỳ ngày-đêm, phản ánh sự thay đổi giữa ban ngày và ban đêm.
* RH (Độ ẩm tương đối): Xu hướng biến động rõ ràng theo thời gian, duy trì độ ẩm cao hoặc thấp trong từng giai đoạn, ảnh hưởng đến sự khuếch tán chất ô nhiễm.
* AH (Độ ẩm tuyệt đối): Có xu hướng tăng dần trong giai đoạn cuối của dữ liệu, thể hiện sự thay đổi về hơi nước trong không khí theo thời tiết.

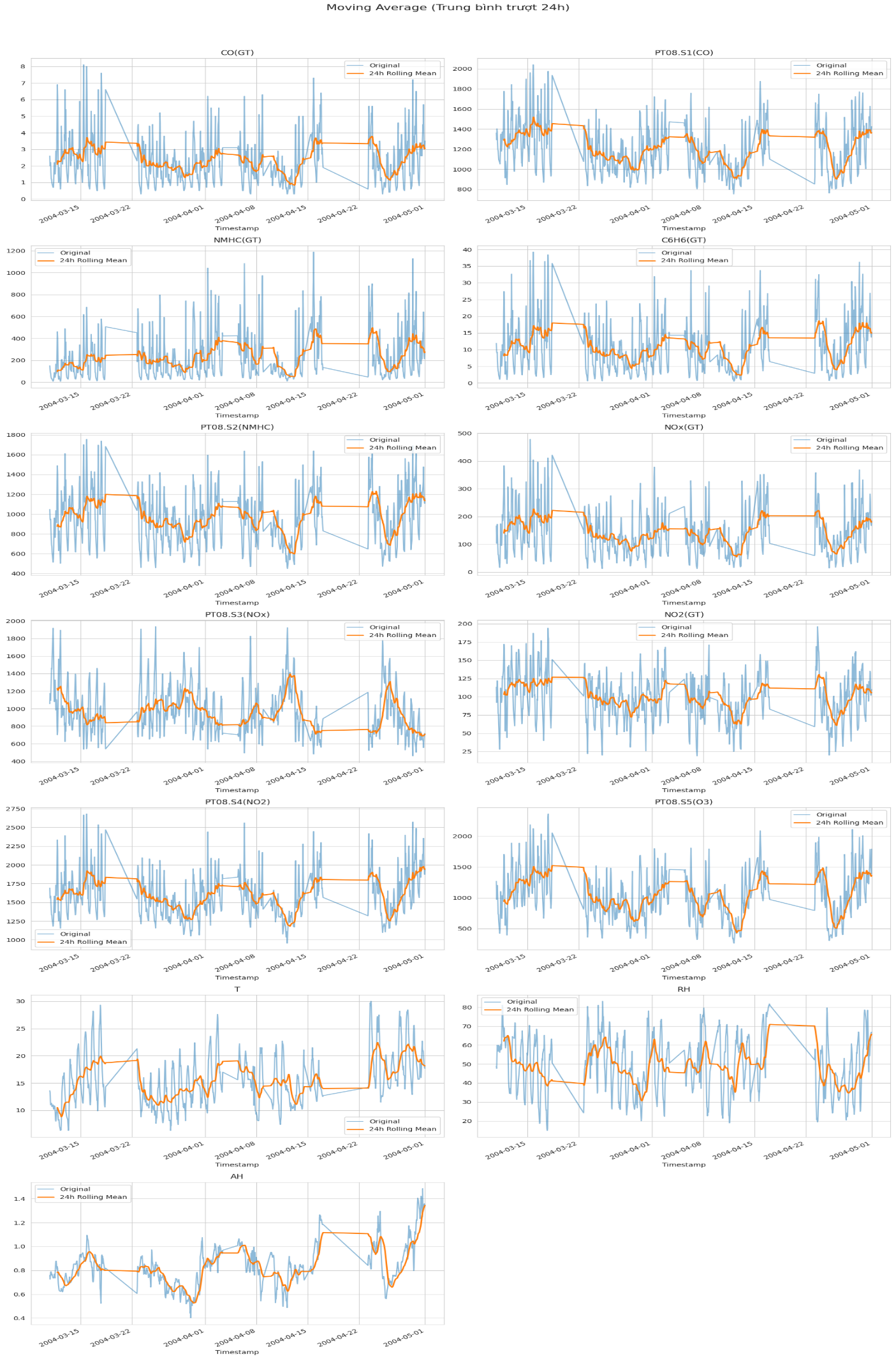
## **7.3 Ý nghĩa và ứng dụng**

Việc sử dụng đường trung bình động 24 giờ giúp làm rõ xu hướng dài hạn, loại bỏ nhiễu từ các biến động ngắn hạn. Điều này mang lại giá trị thực tiễn trong:

* Dự báo chất lượng không khí, giúp xác định xu hướng ô nhiễm theo thời gian.
* Kiểm soát và đánh giá chính sách môi trường, hỗ trợ điều chỉnh các biện pháp quản lý ô nhiễm hiệu quả hơn.
* Cải thiện mô hình dự báo, bằng cách sử dụng dữ liệu đã được làm mịn để tăng độ chính xác của các mô hình phân tích chuỗi thời gian.

## **7.4 Kết luận**

Phân tích đường trung bình động đã giúp nhận diện xu hướng dài hạn trong dữ liệu, làm rõ hơn ảnh hưởng của chất ô nhiễm và yếu tố môi trường theo thời gian. Kết quả này tạo nền tảng cho các nghiên cứu tiếp theo về dự báo, kiểm soát chất lượng không khí và tối ưu hóa các mô hình phân tích môi trường.



Hình 17: Đường trung bình động( Moving Average)

# **VIII. Phân tích Đường Trung Bình Động Hàm Mũ (Exponential Moving Average - EMA)**

## **8.1 Giới thiệu**

Phương pháp Đường Trung Bình Động Hàm Mũ (Exponential Moving Average - EMA) giúp làm mịn dữ liệu bằng cách gán trọng số cao hơn cho các điểm dữ liệu gần hiện tại, giúp phản ứng nhanh hơn với các biến động cục bộ. Với hệ số span = 24, EMA được áp dụng trên 13 biến để làm nổi bật xu hướng dài hạn trong dữ liệu chất lượng không khí và khí tượng.

## **8.2 Nhận xét chi tiết theo từng biến**

### **8.2.1 Chất ô nhiễm trong không khí**

* CO(GT): EMA làm mịn dữ liệu, phản ánh xu hướng dao động có hệ thống hơn. Các đỉnh (peak) và đáy (valley) trở nên rõ ràng hơn, giúp nhận diện sự thay đổi theo thời gian.
* NOx(GT): Xu hướng tăng và giảm kéo dài, phản ánh biến động phát thải khí NOx theo giờ cao điểm. EMA bám sát dữ liệu thực tế nhanh hơn so với Rolling Mean.
* PT08.S3(NOx): Đường EMA cho thấy sự tương tác giữa nguồn phát thải và các yếu tố khí tượng, làm nổi bật các điểm biến động mạnh.
* NO2(GT) & PT08.S4(NO2): Xu hướng tổng thể ổn định hơn khi được làm mịn, nhưng vẫn giữ độ nhạy với biến động ngắn hạn, phù hợp để theo dõi tác động của kiểm soát ô nhiễm.
* PT08.S5(O3): EMA làm nổi bật biến động O3 theo thời gian, giúp nhận diện các giai đoạn có mức ozone cao hoặc thấp bất thường.
* NMHC(GT) & C6H6(GT): Xu hướng dao động rõ rệt, có thể phản ánh ảnh hưởng từ các nguồn công nghiệp hoặc điều kiện khí tượng.
* PT08.S2(NMHC): EMA cho thấy các chu kỳ tăng giảm ngắn hạn, đồng thời giữ được xu hướng tổng thể.

### **8.2.2 Các yếu tố khí tượng**

* T (Nhiệt độ): EMA giúp làm rõ sự thay đổi nhiệt độ theo các giai đoạn ấm hơn và mát hơn, phản ứng nhanh với các biến động đột ngột.
* RH (Độ ẩm tương đối): Xu hướng thay đổi có hệ thống hơn, giúp nhận diện các chu kỳ độ ẩm theo thời gian.
* AH (Độ ẩm tuyệt đối): EMA theo sát xu hướng tăng dần ở giai đoạn cuối, đồng thời giữ được độ nhạy với biến động ngắn hạn.

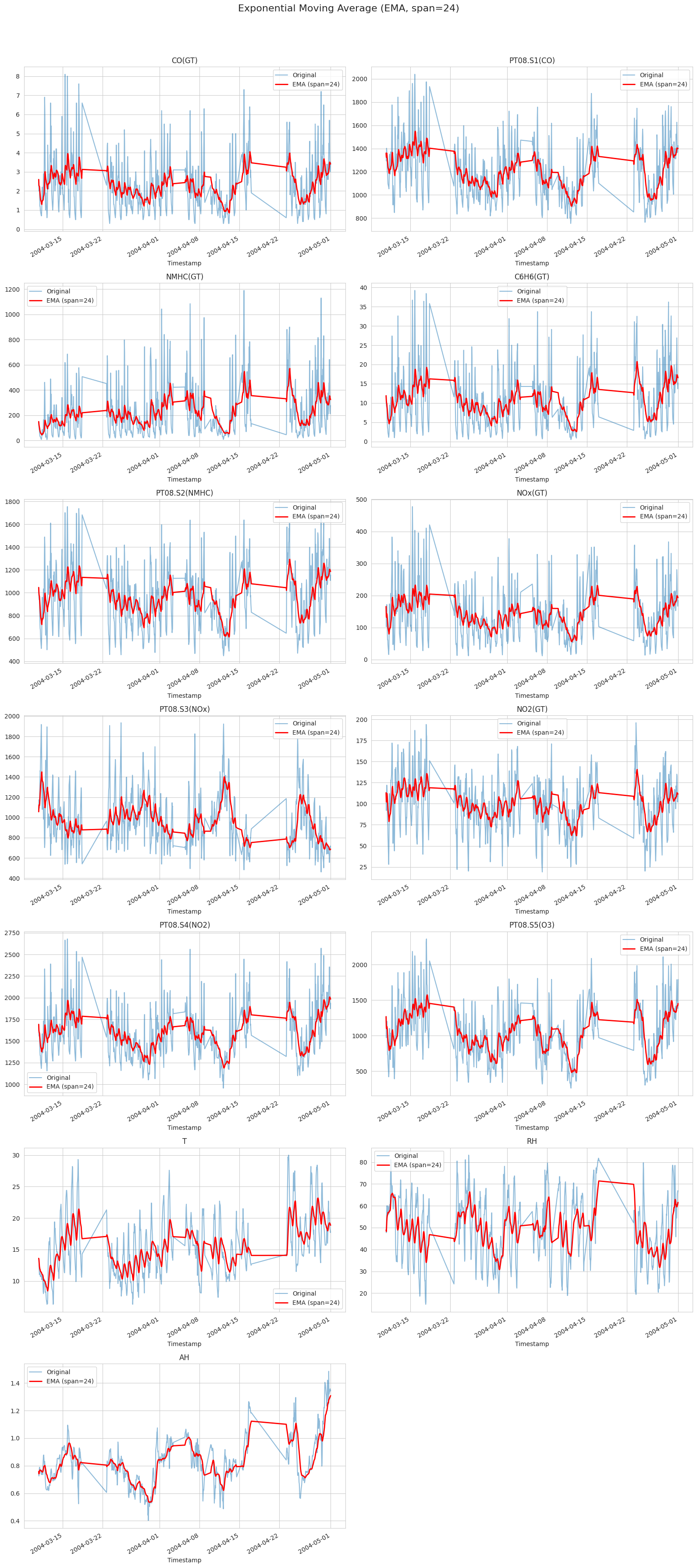
## **8.3 Ứng dụng và ý nghĩa của EMA**

Việc sử dụng EMA mang lại một số lợi ích quan trọng trong phân tích và kiểm soát chất lượng không khí:

* EMA giúp theo dõi các biến động ô nhiễm một cách nhanh chóng, phù hợp cho các hệ thống cảnh báo ô nhiễm theo thời gian thực.
* Làm rõ xu hướng dài hạn, hỗ trợ mô hình hóa và dự báo chất lượng không khí trong các nghiên cứu khí tượng.
* So với Rolling Mean, EMA phản ứng nhanh hơn với các thay đổi đột ngột, giúp xác định các sự kiện ô nhiễm bất thường hoặc điều kiện khí hậu cực đoan.
* Ứng dụng trong quản lý môi trường, giúp nhận diện các thay đổi nhỏ nhưng có tác động lớn đến không khí đô thị.

## **8.4 Kết luận**

Phân tích EMA giúp làm nổi bật xu hướng dài hạn và biến động cục bộ của chất lượng không khí, cho phép theo dõi tác động của nguồn phát thải theo thời gian. So với phương pháp Rolling Mean, EMA cung cấp cái nhìn chính xác hơn về các thay đổi nhanh chóng, tạo nền tảng cho dự báo ô nhiễm và điều chỉnh chính sách kiểm soát môi trường.



Hình 18: Exponential Moving Average (EMA)

# **IX. Đánh giá Độ Chính Xác Dự Báo với Mô hình Naive Forecast**

## **9.1 Giới thiệu**

Mô hình **Naive Forecast** là một phương pháp đơn giản trong dự báo chuỗi thời gian, giả định rằng **giá trị trong tương lai bằng giá trị quan sát gần nhất**. Để đánh giá hiệu suất của mô hình này, nghiên cứu sử dụng **khoảng thời gian từ 15/03/2004 đến 20/03/2004** và phân tích dự báo trên **13 biến** thông qua các chỉ số sai số và biểu đồ so sánh với thực tế.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | |  | | --- | |  |  |  | | --- | | **Biến** | | |  | | --- | |  |  |  | | --- | | **MAE** | | |  | | --- | |  |  |  | | --- | | **MSE** | | |  | | --- | |  |  |  | | --- | | **RMSE** | | |  | | --- | |  |  |  | | --- | | **MAPE** | |
| 1 | NMHC(GT) | 89.780822 | 14380.328767 | 119.918009 | 50.762425 |
| 2 | C6H6(GT) | 5.361644 | 49.750411 | 7.053397 | 39.093318 |
| 3 | |  | | --- | |  |  |  | | --- | | CO(GT) | | 1.038356 | 1.932603 | 1.390181 | 34.698536 |
| 4 | |  | | --- | |  |  |  | | --- | | NOx(GT) | | 59.821918 | 6073.246575 | 77.931037 | 32.914889 |
| 5 | |  | | --- | |  |  |  | | --- | | NO2(GT) | | 17.232877 | 464.821918 | 21.559729 | 16.040810 |
| 6 | PT08.S5(O3) | 216.589041 | 72466.506849 | 269.196038 | 15.213976 |
| 7 | |  | | --- | |  |  |  | | --- | | PT08.S2(NMHC) | | 165.369863 | 42938.136986 | 207.215195 | 14.636411 |
| 8 | |  | | --- | |  |  |  | | --- | | PT08.S3(NOx) | | 118.684932 | 22155.123288 | 148.845972 | 14.183467 |
| 9 | |  | | --- | |  |  |  | | --- | | PT08.S1(CO) | | 162.863014 | 42241.931507 | 205.528420 | 11.333021 |
| 10 | |  | | --- | |  |  |  | | --- | | RH | | 4.497260 | 29.993288 | 5.476613 | 11.128858 |
| 11 | PT08.S4(NO2) | 196.287671 | 63694.917808 | 252.378521 | 10.301313 |
| 12 | T (Temperature) | 1.505479 | 3.361781 | 1.833516 | 8.715343 |
| 13 | AH (Absolute Humi) | 0.031321 | 0.002716 | 0.052118 | 3.835637 |

## **9.2 Nhận xét về Độ Chính Xác Dự Báo giữa các Biến**

Phân tích các chỉ số **MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), và MAPE (Mean Absolute Percentage Error)** cho thấy sự khác biệt đáng kể về khả năng dự đoán giữa nhóm **chất ô nhiễm** và **các yếu tố khí tượng**.

### **9.2.1 Nhóm chất ô nhiễm**

Các biến **CO(GT), NOx(GT), PT08.S3(NOx), NO2(GT), PT08.S4(NO2), PT08.S5(O3), PT08.S1(CO), NMHC(GT), C6H6(GT), PT08.S2(NMHC)** có xu hướng **MAPE cao hơn**, cho thấy tỷ lệ sai lệch trung bình của mô hình Naive khi dự báo các chất ô nhiễm là đáng kể. Nguyên nhân có thể do:

* **Tính biến động mạnh** của các chất ô nhiễm, đặc biệt là **NMHC(GT) và C6H6(GT)**, dẫn đến sai số lớn hơn tại các thời điểm có sự thay đổi đột ngột.
* **Tác động từ nhiều yếu tố bên ngoài**, chẳng hạn như giao thông, công nghiệp, hoặc điều kiện khí tượng, khiến mô hình Naive không phản ánh đúng xu hướng thực tế.
* **PT08.S4(NO2) có MAPE thấp hơn so với các biến còn lại**, do có xu hướng ổn định hơn trong khoảng thời gian đánh giá.

### **9.2.2 Nhóm yếu tố khí tượng**

Các biến **T (Nhiệt độ), RH (Độ ẩm tương đối), AH (Độ ẩm tuyệt đối)** có **MAPE thấp hơn**, cho thấy mô hình Naive dự đoán chúng với tỷ lệ sai lệch nhỏ hơn. Nguyên nhân có thể do:

* **Tính tuần hoàn của dữ liệu khí tượng**, giúp mô hình Naive dự báo tốt hơn các biến có sự biến động ổn định.
* **RH (Độ ẩm tương đối) và AH (Độ ẩm tuyệt đối) có sai số thấp hơn** do tính chất khí tượng ít bị ảnh hưởng bởi các yếu tố ngẫu nhiên như khí thải từ phương tiện giao thông.

****

Hình 19: MAE, MSE, RMSE, MAPE theo các biến từ 15-03-2004 đến 20-03-2004

## **9.3 Nhận xét về Biểu đồ Dự báo so với Thực tế**

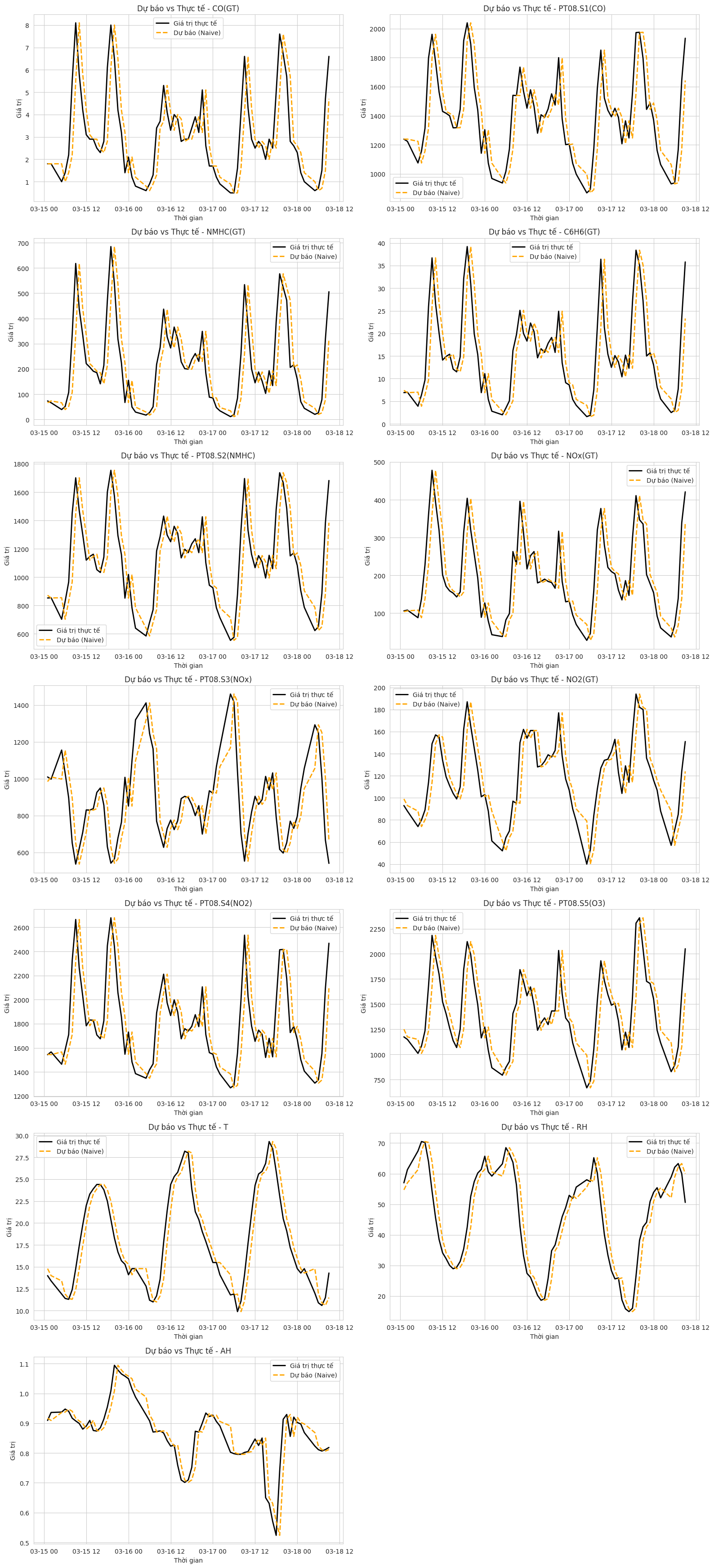
Các biểu đồ so sánh **đường giá trị thực tế và đường dự báo Naive** giúp làm rõ hơn hiệu suất của mô hình trên từng biến.

### **9.3.1 Nhóm chất ô nhiễm**

* Biểu đồ cho các biến **CO(GT), NOx(GT), PT08.S3(NOx), NO2(GT), PT08.S4(NO2), PT08.S5(O3), PT08.S1(CO), NMHC(GT), C6H6(GT), PT08.S2(NMHC)** cho thấy rõ **độ trễ của đường dự báo Naive**, đặc biệt tại các điểm có sự thay đổi nhanh chóng.
* Ở **NMHC(GT) và C6H6(GT)**, khoảng cách giữa đường dự báo và đường thực tế thường lớn hơn tại **các đỉnh và đáy nhọn**, dẫn đến sai số cao.

### **9.3.2 Nhóm yếu tố khí tượng**

* Biểu đồ cho thấy đường dự báo Naive **bám sát đường thực tế hơn** đối với **T (Nhiệt độ), RH (Độ ẩm tương đối), AH (Độ ẩm tuyệt đối)**.
* Tuy nhiên, **vẫn có độ trễ**, đặc biệt ở những điểm ngoặt. Ví dụ:
* Ở **Nhiệt độ (T)**, đường dự báo phản ứng chậm với **các đỉnh và đáy hàng ngày**.
* Ở **Độ ẩm tương đối (RH) và Độ ẩm tuyệt đối (AH)**, sự thay đổi đột ngột vẫn khiến mô hình Naive có sai số, mặc dù nhỏ hơn so với các chất ô nhiễm.



Hình 20: Biểu đồ dự báo so với thực tế của 13 biến

## **9.4 Ý nghĩa và Ứng dụng**

Việc đánh giá mô hình **Naive Forecast** giúp nhận diện những hạn chế và tiềm năng của phương pháp dự báo đơn giản này:

* **Mô hình Naive hoạt động tốt hơn đối với các yếu tố khí tượng**, nơi mà dữ liệu có tính tuần hoàn và ít biến động bất thường.
* **Đối với các chất ô nhiễm, mô hình này có độ sai lệch cao**, đặc biệt với các biến có tính biến động mạnh như **NMHC(GT) và C6H6(GT)**.
* **Các mô hình phức tạp hơn như ARIMA hoặc LSTM có thể cải thiện độ chính xác**, bằng cách xử lý xu hướng và tính mùa vụ tốt hơn.
* **Ứng dụng thực tiễn**: Kết quả này có thể giúp các nhà nghiên cứu quyết định **khi nào nên dùng mô hình Naive** và khi nào cần một phương pháp dự báo tiên tiến hơn.

## **9.5 Kết luận**

Phân tích độ chính xác dự báo của mô hình **Naive Forecast** cung cấp **cái nhìn nền tảng về khả năng dự đoán dữ liệu chất lượng không khí và khí tượng**. Mặc dù phương pháp này hoạt động tốt hơn trên nhóm **yếu tố khí tượng**, nó gặp hạn chế đáng kể khi dự đoán các **biến động mạnh của chất ô nhiễm**.

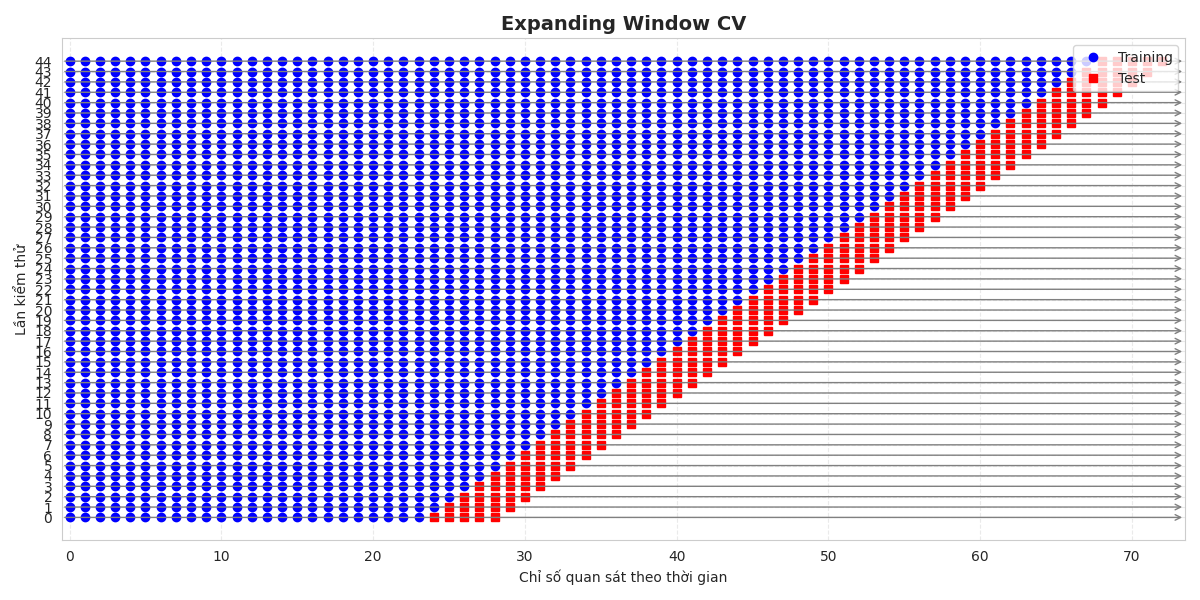
Kết quả này nhấn mạnh **sự cần thiết của các mô hình dự báo phức tạp hơn**, đặc biệt với dữ liệu có độ biến động cao. Việc kết hợp mô hình Naive với **các phương pháp xử lý xu hướng và tính mùa vụ**, như **ARIMA, LSTM**, có thể mang lại khả năng dự báo chính xác hơn, giúp **quản lý ô nhiễm và điều chỉnh chính sách môi trường hiệu quả hơn**.

# **X. Xác nhận Chéo (Cross-Validation) trong Dự báo Chuỗi Thời gian**

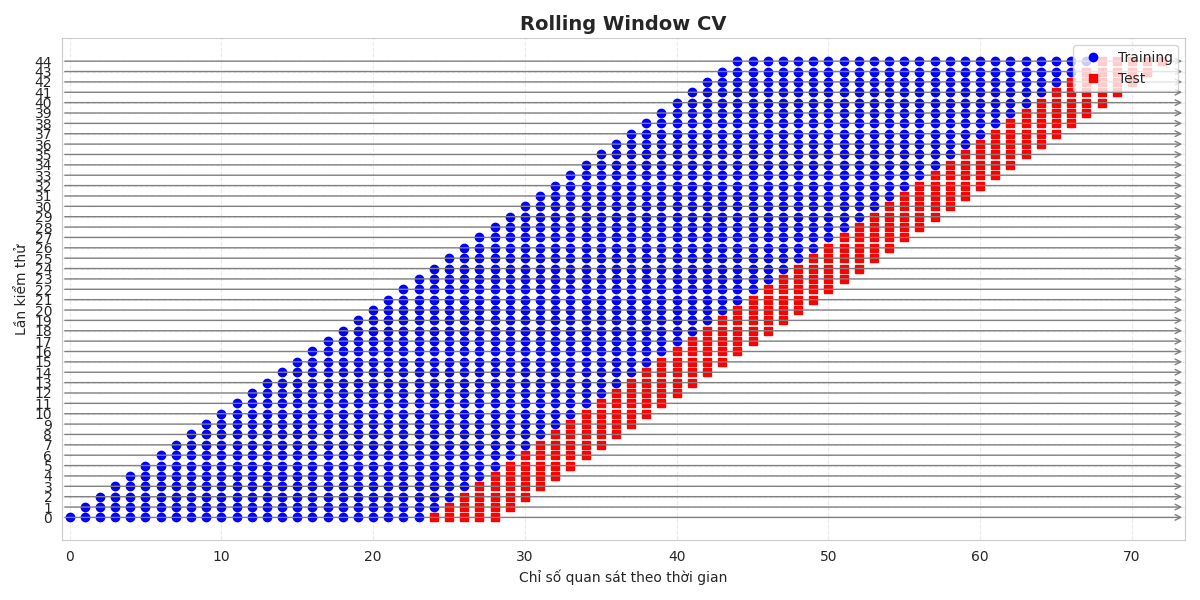
Để đánh giá một cách khách quan khả năng tổng quát hóa của mô hình dự báo Naive, chúng tôi đã áp dụng kỹ thuật xác nhận chéo (Cross-Validation) với hai phương pháp phổ biến cho chuỗi thời gian: Expanding Window và Rolling Window. Cross-validation giúp ước lượng hiệu suất của mô hình trên các phân đoạn dữ liệu khác nhau, từ đó đánh giá độ ổn định và tin cậy của nó.

## **10.1 Quan sát từ Kết quả Cross-Validation**

Kết quả cross-validation được thể hiện qua bảng các chỉ số trung bình và biểu đồ minh họa quy trình CV.

* **Bảng Kết quả CV:** Bảng (được in ra từ code) so sánh hiệu suất trung bình (MAE, MSE, RMSE, MAPE) của mô hình Naive Forecast khi được đánh giá bằng phương pháp Expanding Window và Rolling Window trên các biến. Các giá trị trung bình này cung cấp một cái nhìn tổng quan và ổn định hơn về hiệu suất của mô hình so với việc chỉ đánh giá trên một tập train/test duy nhất. Chúng ta có thể so sánh trực tiếp hiệu suất của Naive Forecast giữa hai phương pháp CV cho từng biến.
* **Biểu đồ Minh họa CV:** Các biểu đồ trực quan hóa cách dữ liệu được chia thành các tập huấn luyện (màu xanh lam) và tập kiểm thử (màu đỏ) trong mỗi lần lặp của quá trình CV.
  + **Rolling Window CV:** Biểu đồ cho thấy một cửa sổ dữ liệu có kích thước cố định trượt qua chuỗi thời gian. Trong mỗi lần lặp, một phần dữ liệu liên tiếp được sử dụng để huấn luyện, và phần tiếp theo ngay sau đó được dùng để kiểm thử. Phương pháp này giả định rằng các quan sát gần đây có ảnh hưởng lớn nhất đến dự báo.
  + **Expanding Window CV:** Biểu đồ minh họa rằng tập huấn luyện ngày càng mở rộng theo thời gian, bao gồm tất cả các dữ liệu từ điểm bắt đầu cho đến trước tập kiểm thử hiện tại. Tập kiểm thử thường có kích thước cố định và dịch chuyển theo thời gian. Phương pháp này phù hợp khi chúng ta tin rằng việc sử dụng toàn bộ lịch sử dữ liệu sẽ cải thiện khả năng dự báo.

Hình 21: Expanding Window CV



Hình 22: Rolling Window CV

## **10.2 Ý nghĩa của Cross-Validation trong Dự báo Chuỗi Thời gian**

Việc thực hiện cross-validation là một bước quan trọng để đánh giá tính tin cậy của mô hình dự báo. Bằng cách đánh giá mô hình trên nhiều phân đoạn dữ liệu khác nhau, chúng ta có thể có được ước tính chính xác hơn về hiệu suất thực tế của nó khi áp dụng trên dữ liệu mới.

* Expanding Window CV phù hợp với các chuỗi thời gian mà việc học từ lịch sử dữ liệu dài hơn có thể cải thiện khả năng dự báo.
* Rolling Window CV thích hợp hơn cho các chuỗi thời gian mà các mẫu gần đây có xu hướng quan trọng hơn, hoặc để đánh giá hiệu suất của mô hình trong một "horizon" dự báo cụ thể khi mô hình được cập nhật thường xuyên.

Kết quả từ phần này cung cấp một đánh giá toàn diện hơn về hiệu suất của mô hình Naive Forecast trên các phần khác nhau của dữ liệu, giúp chúng ta có một đường cơ sở (baseline) vững chắc hơn trước khi khám phá các mô hình dự báo phức tạp hơn. Việc so sánh kết quả giữa Expanding và Rolling Window CV cũng có thể cung cấp thông tin về đặc điểm của chuỗi thời gian đang được phân tích.

# **XI. MÔ HÌNH ARIMA**

## **11.1 Giới thiệu**

Mô hình **ARIMA (Autoregressive Integrated Moving Average)** là một phương pháp dự báo chuỗi thời gian kết hợp **thành phần tự hồi quy (AR), tích hợp (I) và trung bình trượt (MA)**. Trong phần này, nghiên cứu thử nghiệm mô hình ARIMA(1,1,1) cho ba biến **CO(GT), PT08.S1(CO), và NMHC(GT)** trong khoảng thời gian từ **2004-03-15 đến 2004-03-20** để đánh giá khả năng dự báo.

## **11.2 Kết quả Mô hình ARIMA**

Mô hình ARIMA(1,1,1) được áp dụng trên ba biến, với các kết quả chính bao gồm:

* **Hệ số AR (ar.L1) và MA (ma.L1)** không có ý nghĩa thống kê ở mức **alpha = 0.05**, tức là các thành phần này không đóng vai trò quan trọng trong việc mô hình hóa dữ liệu.
* **Phương sai sai số (sigma2)** cho thấy **PT08.S1(CO) có phương sai sai số lớn nhất**, tức là biến này có mức độ dao động khó dự báo hơn.
* **Các chỉ số đánh giá mô hình như AIC và BIC** giúp so sánh mức độ phù hợp của mô hình với dữ liệu.

## **11.3 Biểu đồ Dự báo ARIMA và Ý nghĩa của Dự Báo Ngang (Flat Forecast)**

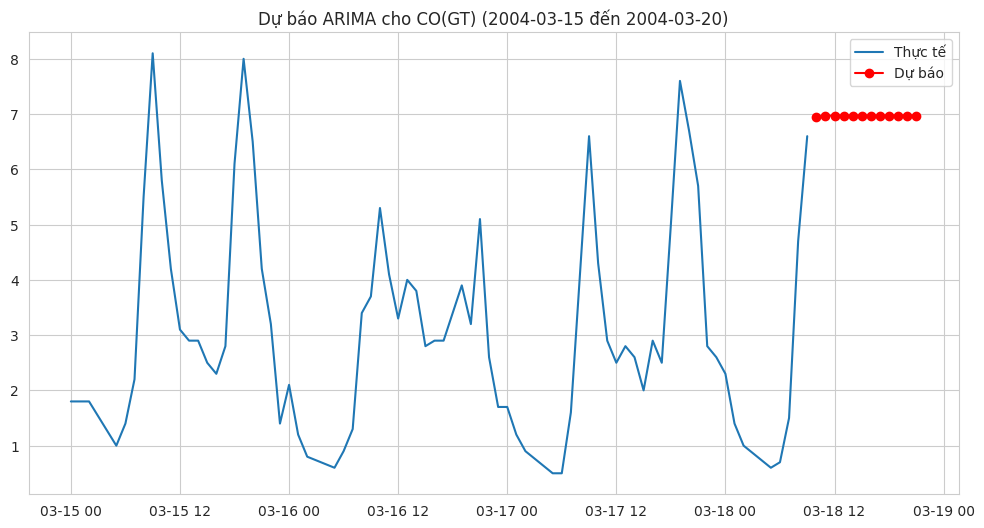
Biểu đồ so sánh **đường giá trị thực tế (màu xanh lam)** với **đường dự báo ARIMA (các chấm tròn màu đỏ)** làm rõ khả năng dự báo của mô hình:

### **11.3.1 Xu hướng dự báo ngang của ARIMA**

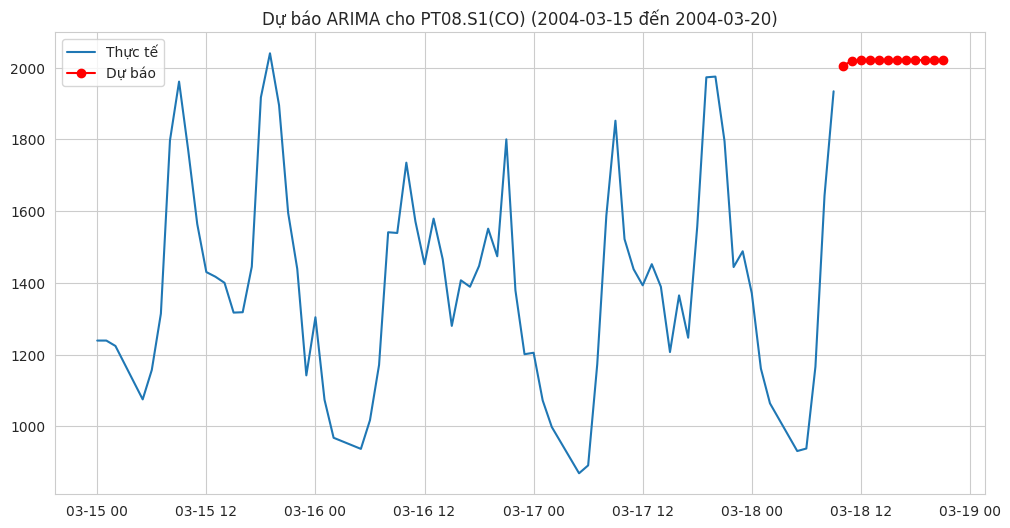
* **Dự báo ARIMA cho NMHC(GT), CO(GT) và PT08.S1(CO) đều thể hiện xu hướng ngang sau điểm cuối của dữ liệu thực tế**.
* **Dự báo ngang (flat forecast) có thể phản ánh rằng dữ liệu không chứa xu hướng rõ ràng hoặc tính mùa vụ không đáng kể**, dẫn đến việc mô hình không thể nắm bắt biến động dài hạn.
* Nếu dữ liệu có sự thay đổi mạnh mẽ theo thời gian, nhưng mô hình **chỉ đưa ra dự báo phẳng**, điều này cho thấy **cấu trúc ARIMA(1,1,1) chưa đủ linh hoạt để phản ánh động thái của dữ liệu**.

### **11.3.2 Nhận xét về độ phù hợp của mô hình ARIMA(1,1,1)**

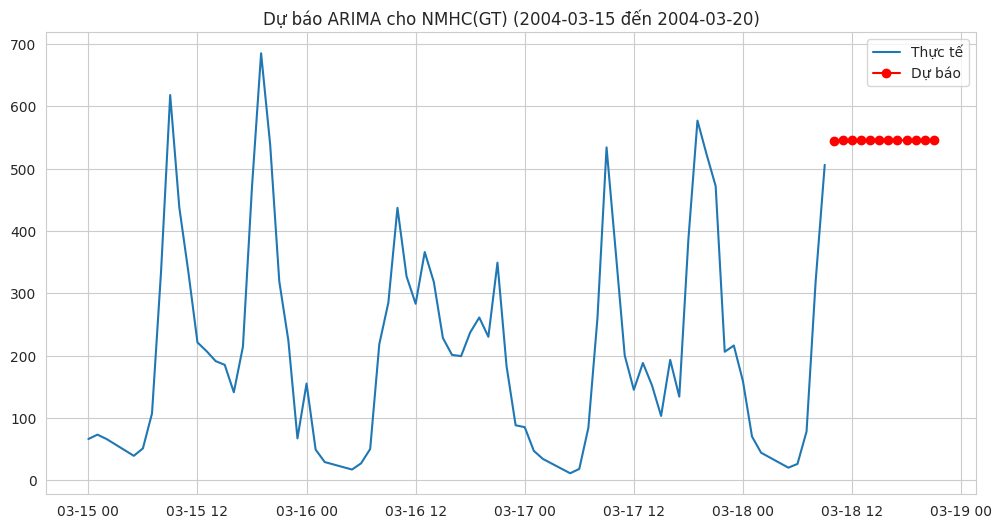
* **NMHC(GT) và CO(GT) có biến động mạnh theo thời gian**, nhưng dự báo của ARIMA vẫn đi ngang, chứng tỏ mô hình không nhận diện được xu hướng dài hạn hoặc các biến động theo chu kỳ.
* **PT08.S1(CO) có phương sai sai số lớn nhất**, cho thấy mô hình gặp khó khăn khi dự báo biến này.



Hình 23: Dự báo ARIMA cho CO(GT) từ 15-03-2004 đến 20-03-2004



Hình 24: Dự báo ARIMA cho PT08.S1(CO) từ 15-03-2004 đến 20-03-2004



Hình 25: Dự báo ARIMA cho NMHC(GT) từ 15-03-2004 đến 20-03-2004

## **11.4 Đề xuất cải thiện mô hình dự báo**

* **Thử nghiệm các bậc ARIMA khác** (ví dụ ARIMA(2,1,2) hoặc ARIMA(0,1,2)) để kiểm tra mô hình phù hợp hơn với dữ liệu.
* **Xem xét mô hình SARIMA (Seasonal ARIMA)** nếu dữ liệu có chu kỳ lặp lại theo thời gian.
* **Ứng dụng mô hình phi tuyến như LSTM hoặc XGBoost** để dự báo các biến động phức tạp hơn.

## **11.5 Kết luận**

Mô hình **ARIMA(1,1,1) chưa thể hiện hiệu suất cao trong việc dự báo ba biến này**, với các hệ số không có ý nghĩa thống kê và **đường dự báo ngang sau điểm cuối của dữ liệu thực tế**. Kết quả này gợi ý rằng:

* **Dữ liệu có thể không có xu hướng rõ ràng hoặc mô hình đang bỏ sót các đặc trưng quan trọng**.
* **Cấu hình ARIMA(1,1,1) có thể chưa phải lựa chọn tốt nhất**, cần thử nghiệm các bậc khác hoặc mô hình phức tạp hơn.
* **Việc kết hợp các phương pháp như SARIMA hoặc LSTM có thể giúp nâng cao độ chính xác dự báo**, đặc biệt với dữ liệu biến động mạnh.

# **XII MÔ HÌNH HOLT-WINTERS**

Trong phần này, chúng ta đã xây dựng và đánh giá mô hình Holt-Winters cho ba biến được cho là có tính thời vụ: PT08.S1(CO), T, và RH, trong khoảng thời gian từ 2004-03-15 đến 2004-03-20. Mô hình Holt-Winters phù hợp với các chuỗi thời gian có cả xu hướng và tính thời vụ. Chúng ta đã cấu hình mô hình với thành phần xu hướng cộng tính (trend='add') và thành phần thời vụ cộng tính (seasonal='add'), với chu kỳ thời vụ là 24 giờ (do dữ liệu được thu thập theo giờ).

## **12.1 Kết quả Mô hình Holt-Winters**

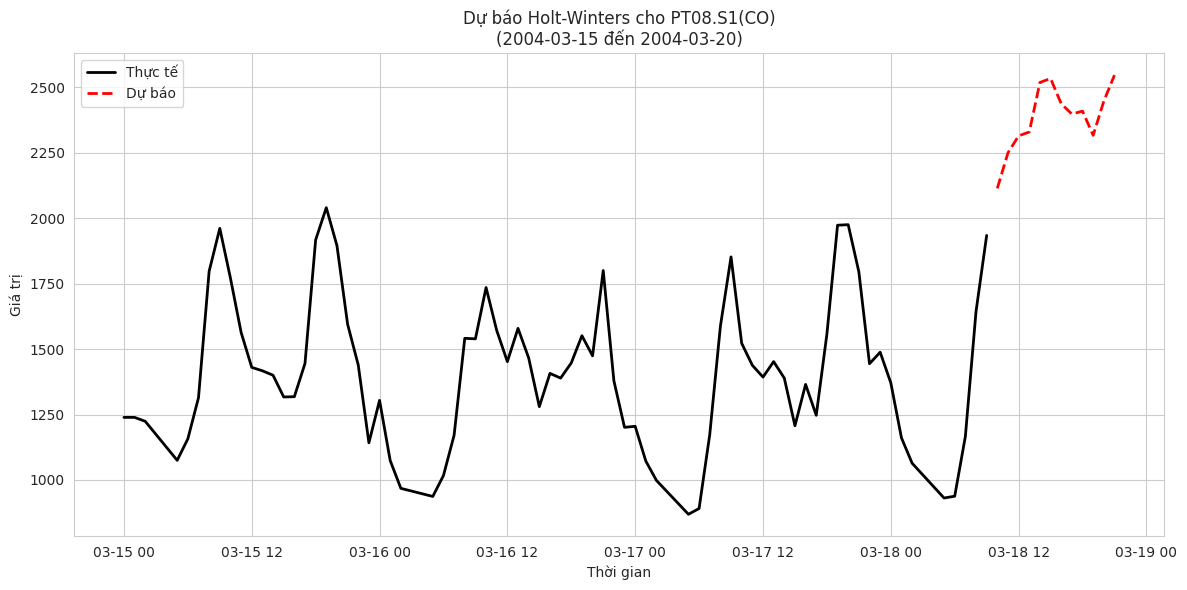
Kết quả tóm tắt của mô hình Holt-Winters cho từng biến đã được in ra, cung cấp thông tin về các hệ số làm trơn (smoothing level, smoothing trend, smoothing seasonal), các giá trị khởi tạo (initial level, initial trend, initial seasons), và các chỉ số đánh giá như SSE (Sum of Squared Errors) và AIC.

* **PT08.S1(CO):** Mô hình Holt-Winters cho biến này đã được tối ưu hóa, với các hệ số làm trơn và các giá trị khởi tạo được ước tính. Giá trị SSE là 2198089.101 và AIC là 818.129.
* **T (Nhiệt độ):** Tương tự, mô hình cho nhiệt độ cũng đã được tối ưu hóa với SSE là 91.705 và AIC là 71.874. Các hệ số làm trơn cho thấy mức độ ảnh hưởng của các thành phần khác nhau đến dự báo.
* **RH (Độ ẩm tương đối):** Mô hình Holt-Winters cho độ ẩm tương đối có SSE là 934.228 và AIC là 243.638.

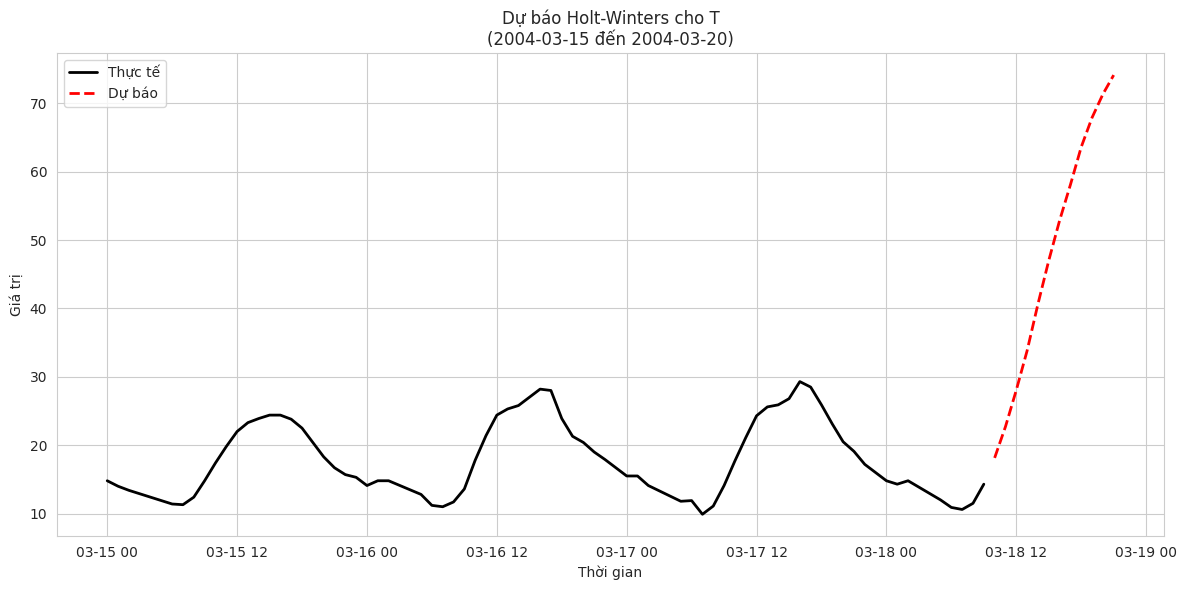
## **12.2 Biểu đồ Dự báo Holt-Winters**

Các biểu đồ bạn cung cấp hiển thị đường giá trị thực tế (màu đen) và 12 bước dự báo tiếp theo từ mô hình Holt-Winters (đường màu đỏ nét đứt).

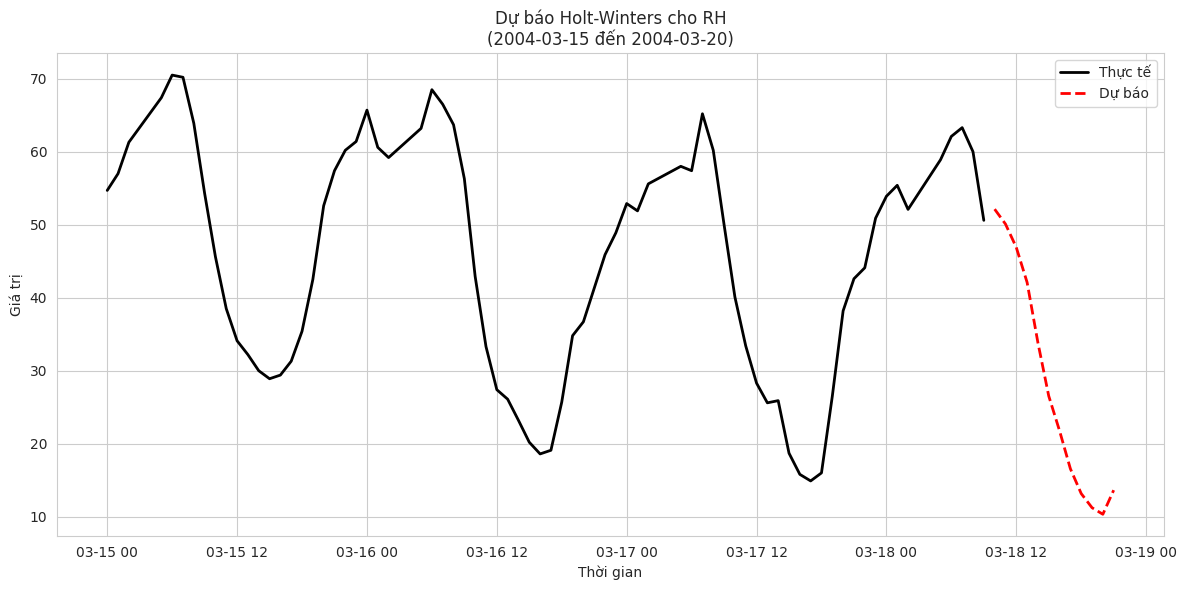
* **Dự báo Holt-Winters cho RH:** Đường dự báo có vẻ tiếp nối xu hướng giảm ở cuối dữ liệu thực tế, sau đó lại tăng lên. Mô hình dường như đang cố gắng nắm bắt tính thời vụ.
* **Dự báo Holt-Winters cho PT08.S1(CO):** Đường dự báo cho thấy sự dao động, có thể là do mô hình đang cố gắng học theo tính thời vụ quan sát được trong dữ liệu huấn luyện.
* **Dự báo Holt-Winters cho T:** Đường dự báo cho nhiệt độ cho thấy một xu hướng tăng mạnh, có thể phản ánh một chu kỳ hoặc xu hướng mà mô hình đã học được từ dữ liệu.



Hình 26: Dự báo Hold-Winter cho PT08.S1(CO)



Hình 27: Dự báo Hold- Winter cho T



Hình 28: Dự báo Hold- Winter cho RH

## **12.3 Nhận xét và Ý nghĩa**

Việc áp dụng mô hình Holt-Winters cho các biến có tính thời vụ đã cho thấy khả năng của mô hình trong việc tiếp nối các mẫu và xu hướng. So với mô hình ARIMA đơn giản (bậc (1, 1, 1)) đã thử nghiệm trước đó, các đường dự báo từ Holt-Winters có vẻ phản ánh các đặc trưng của dữ liệu (như tính thời vụ) rõ ràng hơn.

Tuy nhiên, cần lưu ý rằng việc lựa chọn các tham số (trend, seasonal, seasonal\_periods) có ảnh hưởng lớn đến hiệu suất của mô hình Holt-Winters. Để đánh giá đầy đủ hiệu quả của mô hình, cần thực hiện đánh giá định lượng bằng các chỉ số sai số trên một tập kiểm thử riêng hoặc thông qua cross-validation.

Các bước tiếp theo có thể bao gồm việc tinh chỉnh các tham số của mô hình Holt-Winters, so sánh hiệu suất của nó với các mô hình khác (như ARIMA với bậc phù hợp hơn, hoặc các mô hình phức tạp hơn), và đánh giá độ chính xác của các dự báo đã thực hiện

# **XIII. Phân tích Phần Dư của Mô hình ARIMA(1,1,1)**

## **13.1 Giới thiệu**

Phân tích phần dư của mô hình **ARIMA(1,1,1)** giúp đánh giá mức độ phù hợp của mô hình với dữ liệu chuỗi thời gian. Nếu phần dư **vẫn còn tự tương quan**, điều này có thể cho thấy mô hình chưa mô hình hóa tốt toàn bộ biến động của dữ liệu, dẫn đến **dự báo thiếu chính xác**. Ngoài ra, việc kiểm tra **phân phối phần dư** giúp xác định liệu sai số có tuân theo phân phối chuẩn hay không, ảnh hưởng đến **hiệu suất của mô hình dự báo**.

## **13.2 Nhận xét về Phân Tích Phần Dư**

**Nhóm 1: CO(GT), PT08.S1(CO), NMHC(GT), C6H6(GT), PT08.S2(NMHC)**

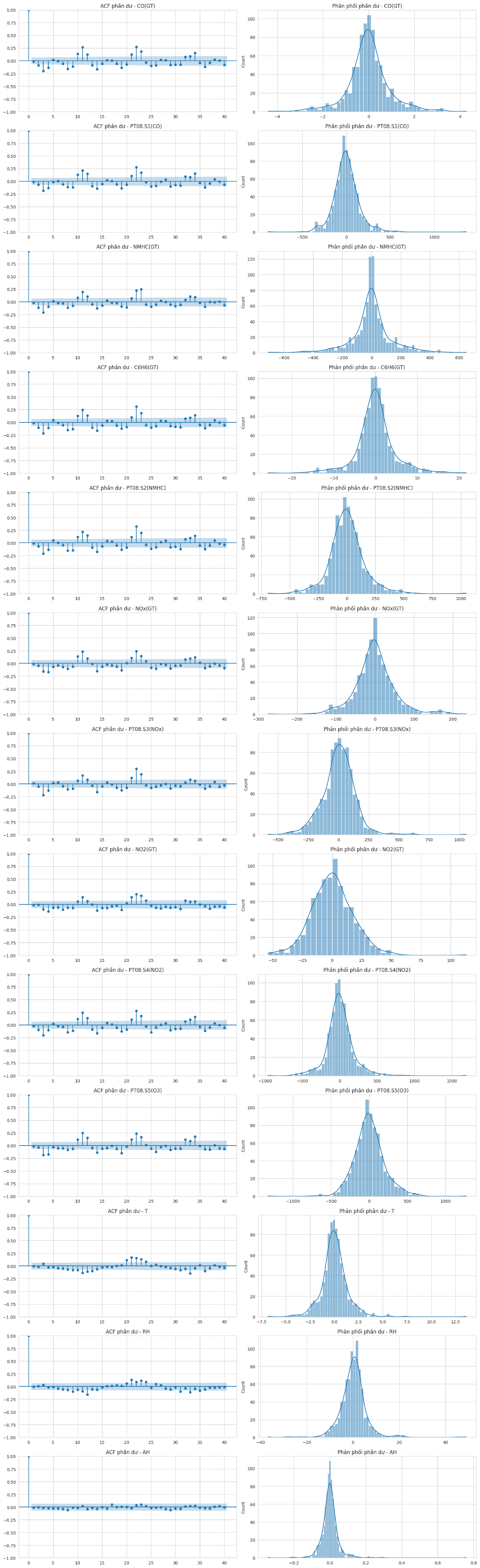
* **ACF Phần Dư:** Biểu đồ ACF cho thấy **nhiều spike vượt ra ngoài vùng tin cậy**, đặc biệt ở các lags nhỏ. Điều này phản ánh **vẫn còn sự tự tương quan trong phần dư**, chứng tỏ mô hình ARIMA(1,1,1) chưa nắm bắt hoàn toàn các mẫu trong dữ liệu.
* **Phân Phối Phần Dư:** Phần dư tập trung quanh giá trị trung bình, nhưng **có sự lệch nhẹ (skewness) hoặc đuôi dày hơn so với phân phối chuẩn**. Điều này có thể làm tăng **biến động của sai số dự báo**, ảnh hưởng đến độ chính xác mô hình.
* **Nhận xét chung:** Mô hình ARIMA(1,1,1) chưa phải là lựa chọn tối ưu cho các biến này, vì vẫn còn **sự tự tương quan trong phần dư và phân phối không lý tưởng**, dẫn đến dự báo không chính xác.

**Nhóm 2: NOx(GT), PT08.S3(NOx), NO2(GT), PT08.S4(NO2), PT08.S5(O3)**

* **ACF Phần Dư:** Biểu đồ ACF tiếp tục cho thấy **sự tự tương quan đáng kể**, đặc biệt với các lags nhỏ, cho thấy mô hình chưa giải thích hết sự thay đổi của dữ liệu.
* **Phân Phối Phần Dư:** Phân phối phần dư có vẻ **tập trung quanh giá trị trung bình nhưng đuôi nặng**, có thể ảnh hưởng đến độ chính xác của dự báo.
* **Nhận xét chung:** Kết quả phần dư của nhóm này tương tự nhóm 1, cho thấy **cần điều chỉnh mô hình để giảm tự tương quan và tối ưu hóa phân phối phần dư**.

**Nhóm 3: T, RH, AH**

* **T (Nhiệt độ):**
  + **ACF Phần Dư:** Có một vài **spike vượt ra ngoài vùng tin cậy ở lags xa**, cho thấy vẫn còn tự tương quan.
  + **Phân Phối Phần Dư:** Có xu hướng tập trung quanh giá trị trung bình nhưng có thể có **đuôi dày hơn**, ảnh hưởng đến dự báo.
* **RH (Độ ẩm tương đối):**
  + **ACF Phần Dư:** Tương tự T, biểu đồ ACF có **một vài spike**, gợi ý mô hình chưa xử lý tốt toàn bộ dao động.
  + **Phân Phối Phần Dư:** Có thể hơi lệch, ảnh hưởng đến chất lượng dự báo.
* **AH (Độ ẩm tuyệt đối):**
  + **ACF Phần Dư:** Biểu đồ ACF cho AH **tốt hơn so với các biến khác**, phần lớn các lags nằm trong vùng tin cậy.
  + **Phân Phối Phần Dư:** Có vẻ **đối xứng hơn**, phản ánh rằng ARIMA(1,1,1) có thể mô hình hóa AH tốt hơn.
* **Nhận xét chung:** **AH có phần dư phù hợp hơn** so với T và RH, nhưng cả ba biến vẫn có thể cải thiện bằng cách điều chỉnh thông số mô hình.



Hình 29: Tự tương quan và phân phối phần dư

## **13.3 Ý nghĩa của Tự Tương Quan và Phân Phối Phần Dư**

### **13.3.1 Tác động của Tự Tương Quan**

* Nếu phần dư **vẫn còn tự tương quan**, mô hình có thể **bỏ sót các thông tin quan trọng về xu hướng và mẫu trong dữ liệu**, dẫn đến **dự báo thiếu chính xác**.
* **Tự tương quan cao** có thể làm cho sai số dự báo **có xu hướng theo thời gian**, thay vì phân bố ngẫu nhiên như giả định.

### **13.3.2 Ảnh hưởng của Phân Phối Phần Dư**

* Nếu phần dư **không tuân theo phân phối chuẩn**, mô hình ARIMA có thể **gặp khó khăn trong việc dự báo chính xác**, đặc biệt khi có **đuôi phân phối dài**.
* **Sai số không ngẫu nhiên** có thể dẫn đến **sai số hệ thống trong mô hình**, làm giảm hiệu suất dự báo.

## **13.4 Kết luận**

Phân tích phần dư của mô hình ARIMA(1,1,1) cho thấy **vẫn còn sự tự tương quan đáng kể trong dữ liệu**, dẫn đến **sai số dự báo cao**. Đồng thời, **phân phối phần dư không hoàn toàn chuẩn**, ảnh hưởng đến độ tin cậy của mô hình.

Để cải thiện mô hình, cần **thử nghiệm các bậc ARIMA khác và điều chỉnh thông số** để giảm tự tương quan và tối ưu hóa phân phối phần dư. Ngoài ra, **việc áp dụng các kỹ thuật xử lý dữ liệu trước** (ví dụ: biến đổi log hoặc sai phân bậc hai) có thể giúp mô hình hoạt động tốt hơn.

# **KẾT LUẬN TỔNG THỂ**

Thông qua quá trình phân tích chuỗi thời gian chi tiết, nhóm nghiên cứu đã xác định được những đặc điểm quan trọng của dữ liệu chất lượng không khí và khí tượng.

Dữ liệu cho thấy đặc trưng rõ ràng của chuỗi thời gian, với tính chu kỳ ngày/đêm, xu hướng dài hạn, và cấu trúc tự tương quan mạnh. Các biến khí thải như CO, NOx, NO2 bị tác động mạnh bởi hoạt động giao thông và công nghiệp, trong khi các yếu tố khí hậu như nhiệt độ và độ ẩm phản ánh tính chất tự nhiên nhiều hơn.

Nhiều chuỗi chưa dừng, thể hiện qua sự thay đổi trung bình và phương sai theo thời gian, yêu cầu các phép xử lý như sai phân để đảm bảo điều kiện cho các mô hình dự báo như ARIMA. Chu kỳ 24 giờ rõ rệt ở cả nhóm chất ô nhiễm và khí tượng, nhấn mạnh tác động từ hoạt động con người và quy luật khí hậu.

Sự tự tương quan mạnh ở nhiều độ trễ phản ánh mối liên hệ chặt chẽ giữa dữ liệu quá khứ và hiện tại, đóng vai trò quan trọng trong việc phát triển mô hình dự báo chính xác hơn. Tuy nhiên, các mô hình được áp dụng như ARIMAvà Holt-Winters mới chỉ phản ánh được một phần xu hướng và chu kỳ, vẫn còn hạn chế đối với các biến có biến động mạnh và phần dư còn tự tương quan. Kết quả này gợi mở tiềm năng của việc ứng dụng các mô hình phi tuyến hoặc phương pháp học máy để nâng cao độ chính xác dự báo.

Những phân tích trên không chỉ giúp cải thiện khả năng dự báo chất lượng không khí, mà còn là cơ sở để định hướng chiến lược kiểm soát ô nhiễm và thiết lập chính sách môi trường hiệu quả hơn trong tương lai.