

機器學習期末報告

台北市PM2.5短期模型

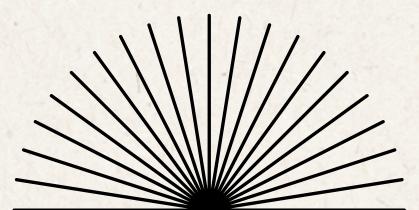
12/31 第十一組

資三B 12156223 吳承瑀

資三B 12156206 章祖綸

資三B 12156217 陳翡翠

資三B 12156229 高碩辰



目錄

- | | |
|----|----------------|
| 一、 | 引言 |
| 二、 | 問題陳述 |
| 三、 | 資料來源與預處理 |
| 四、 | 研究步驟與方法 |
| 五、 | 氣象變數和PM2.5的相關性 |
| 六、 | 所有變數與PM2.5的相關性 |

目錄

七、	研究方法與評估
八、	12/31日之模型效能驗證
九、	LSTM氣象測站預測
十、	研究結果與分析
十一、	結論
十二、	參考文獻
十三、	分工表

一、引言

提案概述：

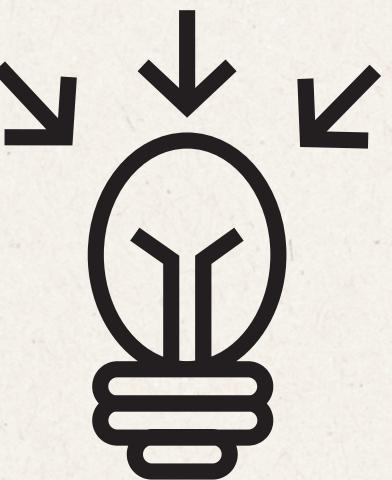
本專案以政府與合法公共資料庫的資料為基礎，建置PM2.5指數短期預測模型，協助使用者提前因應空氣品質變化。

重要性：

提前預測可降低暴露風險、優化戶外活動與營運調度、提升公共衛生與政策溝通效率，並作為智慧城市與永續治理的決策依據。

受益對象：

一般民眾與高風險族群、學校與運動場館、醫療與長照機構、地方政府與防災單位、通風/空調與場館營運業者、研究與教育單位。



二、問題陳述

- **核心問題：**

對特定時間提供準確且可驗證的 PM2.5 預測，作為行動與健康管理依據。

- **為何重要：**

可以提前通知少外出或是穿戴口罩等防護用具避免受到汙染。

- **對目標族群影響：**

1. 高風險族群：對汙染產生疾病極為容易。
2. 學校/運動場館：無法及時調整戶外課程與比賽。
3. 醫療機構：預警不足，加劇急診與慢性病加重風險。

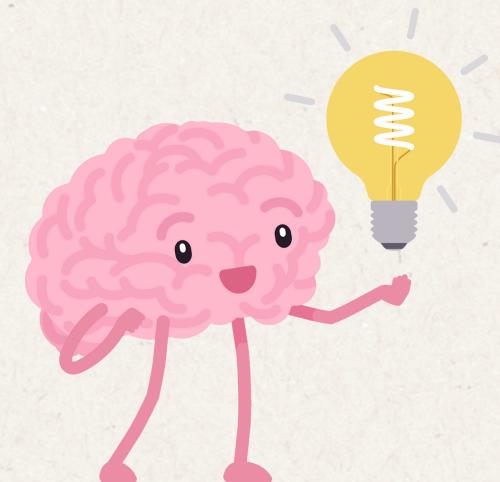


三、資料來源與預處理

資料來源：

- 環境部空氣品質監測網 (台北市部分行政區PM2.5數值)
- 台北市環境品質資訊網 (台北市部分行政區PM2.5數值)
- CODiS氣候觀測資料(台北市行政區氣象資料，包括氣溫，相對溼度，風速，風向，降雨量)

2018-2024，11和12月的小時級資料



三、資料來源與預處理

氣象資料標準化：

- 統一氣象參數格式，處理異常值（如：負值或超出極限之數值）。
- 對齊各測站之時間戳記，確保 PM2.5 與氣象資料在時空上的一致性。
- 缺失值處理：針對感測器故障造成的數據中斷，進行線性補值。



三、資料來源與預處理

滯後變數建立：

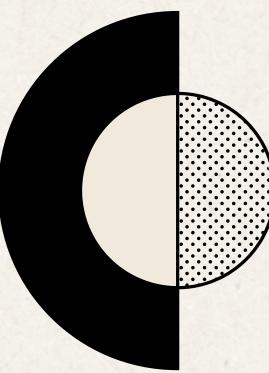
為了反映空氣污染的持續性與延遲效應。

建立了以下特徵：

- PM25_Lag_1h / 2h：短期變動趨勢。
- PM25_Lag_24h：捕捉日週期性的濃度規律。
- 意義：使模型能夠參考「過去的狀態」來預測「未來的濃度」。



三、資料來源與預處理



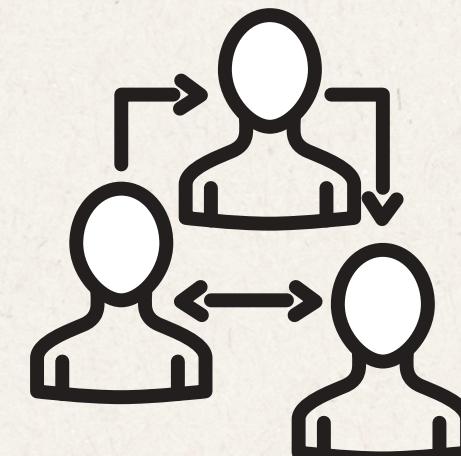
非線性與類別特徵轉換

- 風向向量化 (Wind Direction Encoding)：
 - 將 0-360 度的風向角度轉換為 Wind_Sin 與 Wind_Cos。
 - 原因：解決 360 度與 0 度在數值上遙遠但在地理上一致的斷層問題。
- 類別變數 One-Hot Encoding：
 - 測站編碼 (Station)：將 11 個測站轉為獨立特徵，使模型學習不同地點的環境特性。
 - 星期編碼 (Day of Week, DOW)：區分平日與週末的活動模式差異（如交通流量）。

三、資料來源與預處理

【特徵舉例格式】：

- 風向向量化 (Wind Vector)：
 - Wind_Sin : 0.42 / Wind_Cos : -1.18 (解決0度到360度的連續性問題)
- 測站 One-Hot 編碼 (Station)：
 - Station_中山站 : True , 其餘測站欄位皆為 False
- 星期編碼 (Day of Week)：
 - DOW_4 : True(代表該數據發生在週五)



三、資料來源與預處理

特徵縮放 (Feature Scaling)：

- 對所有數值型特徵（氣象、滯後 PM2.5）進行 Z-score 標準化，消除單位量綱影響（如溫度與風速的數值級距不同）。

資料結構：

- 最終特徵矩陣包含：時序滯後項、氣象因子、週期性特徵、空間編碼特徵。

產出目標：

建立一個高維度、多因子且已清洗完成的FINAL_MODEL_TRAINING_DATA。

PM2.5_V	PM25_La	PM25_La	PM25_La	RAINFAI	WIND_SFRH	AMB_TE	Month	Hour	Wind_Sin	Wind_Cos	
7	-0.64914	-0.27358	0.111042	1.049386	0.874168	1.323242	-0.02094	11	0	0.425516	-1.18277
9	-0.52415	-0.64887	0.237475	0.081597	1.436251	1.323242	-0.04452	11	1	0.408301	-1.2088
10	-0.27417	-0.52378	0.363907	0.081597	0.405766	1.323242	0.002634	11	2	0.425516	-1.18277
11	-0.14918	-0.27358	0.111042	0.404193	0.124724	1.323242	0.049786	11	3	0.62636	-0.76134
4	-0.02419	-0.14849	0.616772	0.404193	1.24889	1.323242	0.096937	11	4	0.390634	-1.23453
3	-0.89912	-0.02339	1.248935	0.081597	1.24889	1.323242	0.096937	11	5	0.372522	-1.25994

四、研究步驟與方法

Step 1	我們先使用隨機森林、XGBoost、RNN、LSTM模型分析2024/12/20到2024/12/29的PM2.5。
Step 2	再聚焦於2024/12/31，分析模型對小區間的預測準確度。
Step 3	接著針對每個氣象測站，用準確度最高的模型分析2024/12/20到2024/12/29的PM2.5。
Step 4	先根據以上2024年的模型分析結果，找出準確度最高的兩個模型，接著再預測 2026-2028年台北市的PM2.5狀況。

五、氣象變數和PM2.5的相關性

圖表中所有的柱狀圖都是向下的（數值為負），這代表這些天氣變數與 PM2.5 呈現負相關。

RH (相對濕度)係數：-0.32

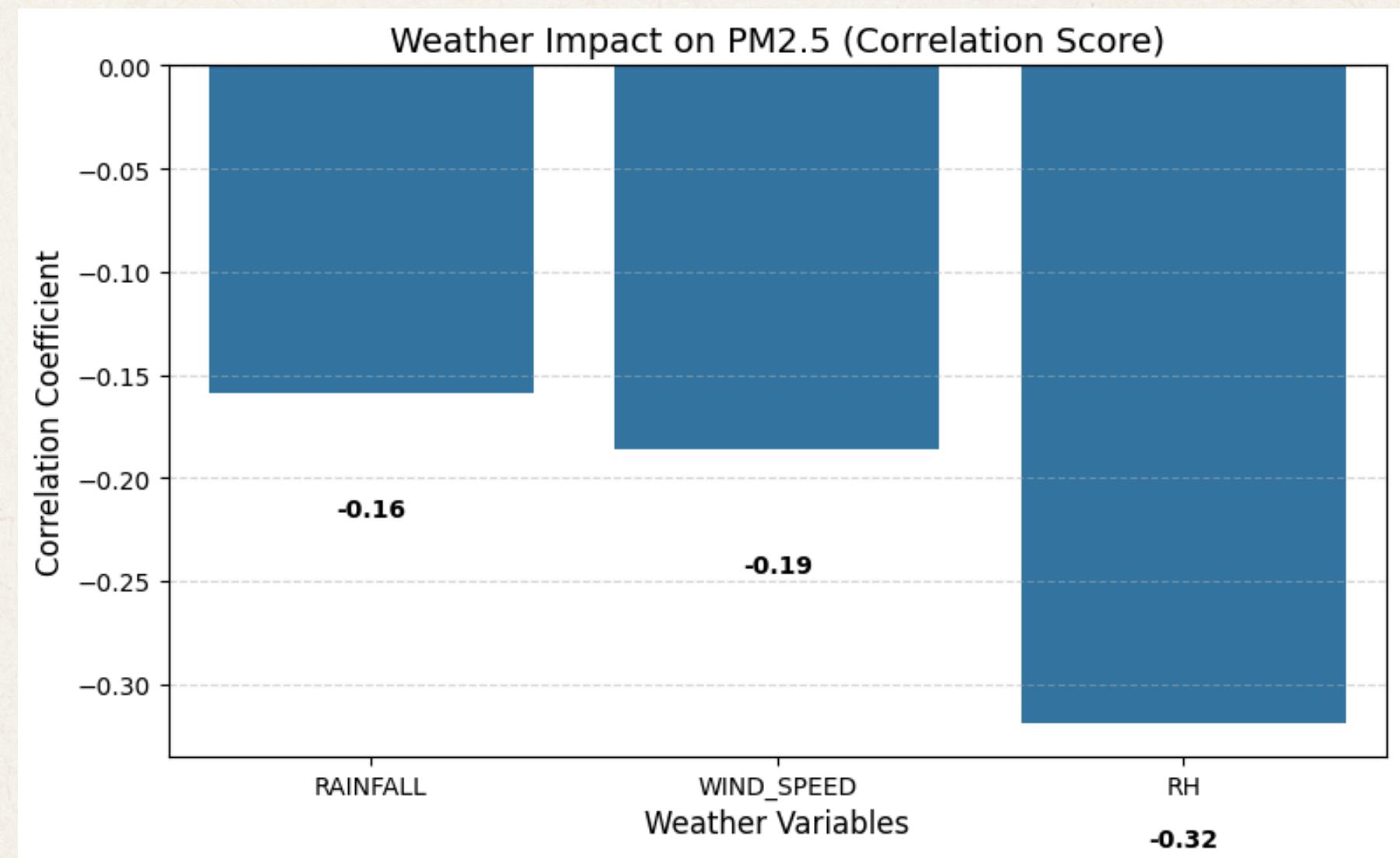
- 原因：當空氣濕度高時，水氣容易吸附懸浮微粒，使其變重而沉降，減少空氣中飄浮的 PM2.5。

WIND_SPEED (風速)係數：-0.19

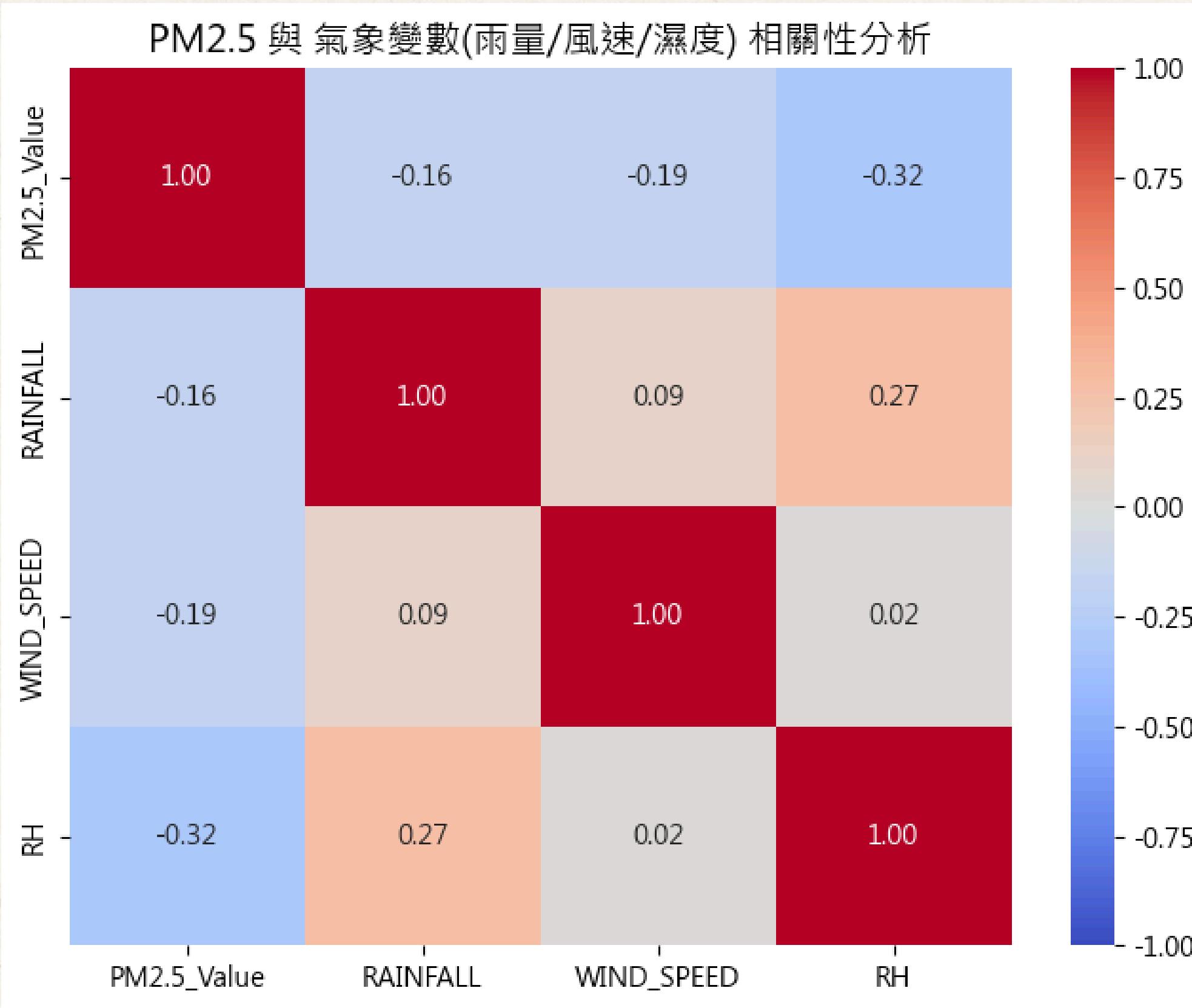
- 原因：強風有助於空氣流通，能將聚集的污染物吹散，降低濃度。

RAINFALL (降雨量)係數：-0.16

- 原因：下雨有洗刷空氣的效果，能將污染物帶離地面。



Pearson Correlation



RH (濕度) vs PM2.5 :

係數約 -0.32

- 三者中影響最大。濕度越高，PM2.5 越低。

WIND_SPEED (風速) vs PM2.5 :

係數約 -0.19

- 風越大，擴散條件越好，PM2.5 越低，但關聯性沒有想像中強。

RAINFALL (降雨量) vs PM2.5 :

係數約 -0.16

- 弱負相關。下雨雖然會洗空氣，但數據顯示其線性關聯性最弱。

六、所有變數與 PM2.5的相關性

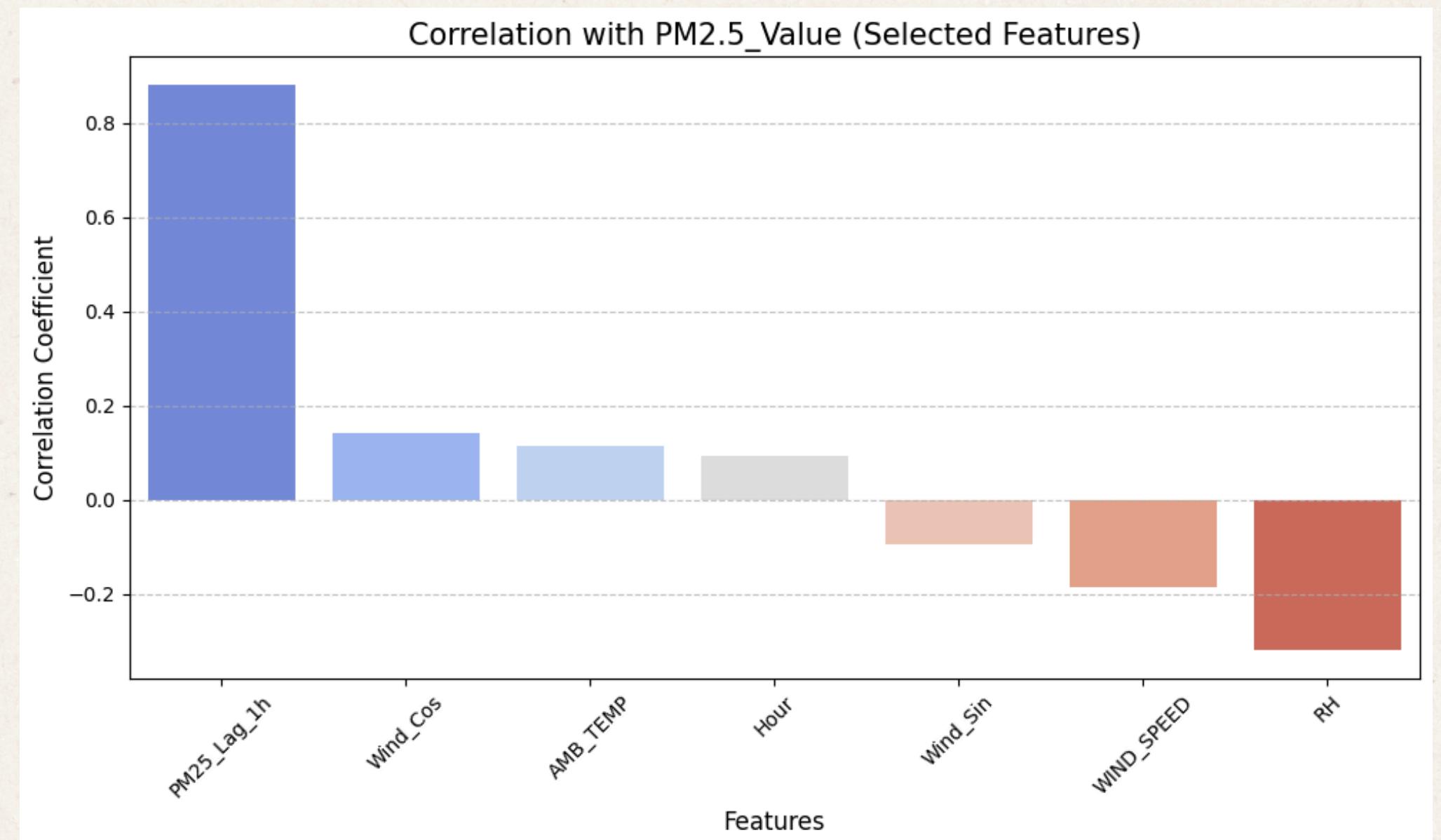
極強的正相關：

PM25_Lag_1h (一小時前的 PM2.5) 係數
超過 0.8。

數據意義：

這代表過去的空氣品質是預測未來空氣品質的最強指標。空氣污染具有持續性，不會突然消失。如果一小時前空氣很糟，現在通常也很糟。

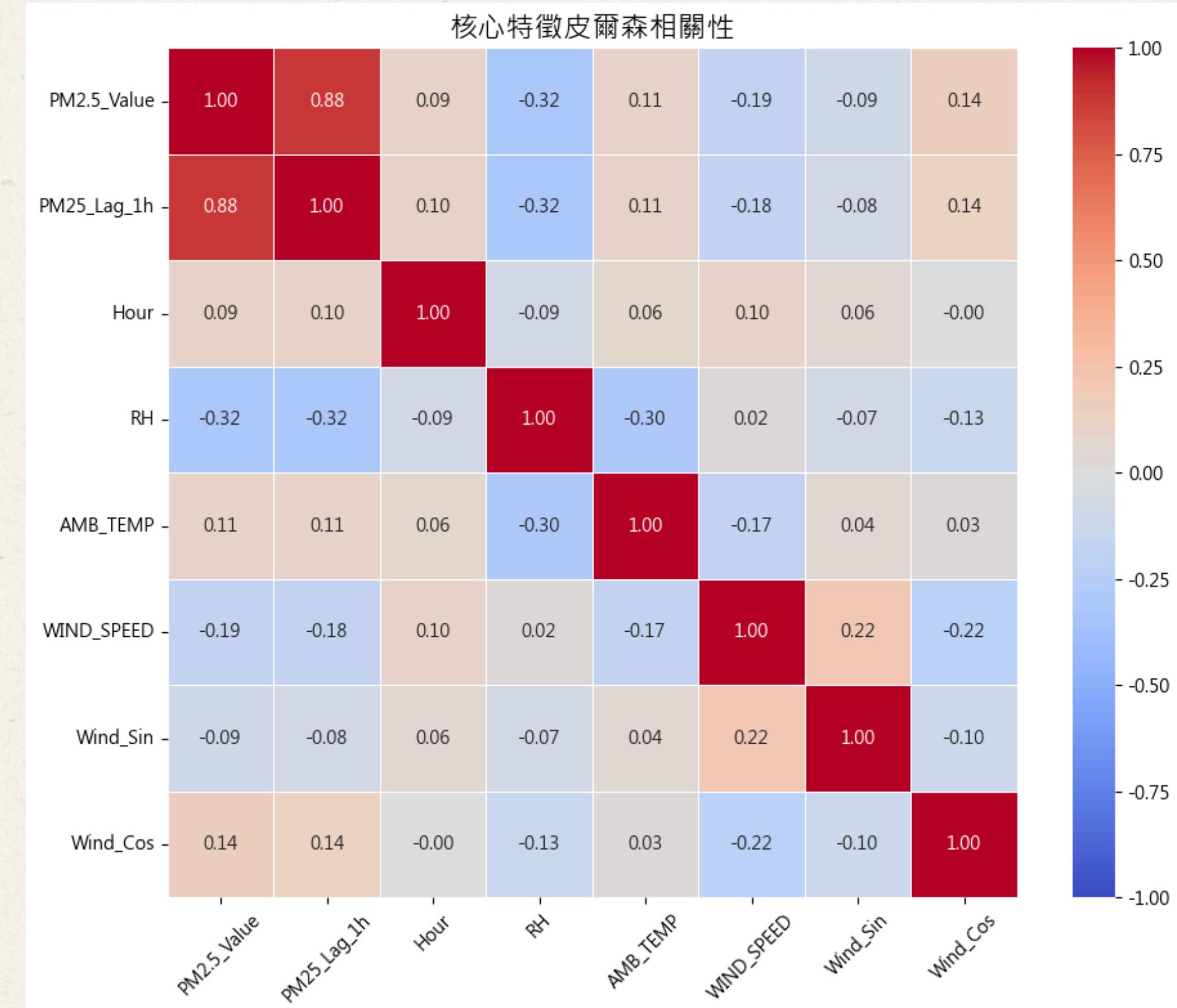
- 相對濕度 (RH)、風速 (WIND_SPEED) 和降雨量 (RAINFALL) 都與 PM2.5 呈負相關。
- AMB_TEMP (溫度) & Hour (小時)：呈現微弱的正相關。這可能與人類活動規律有關。



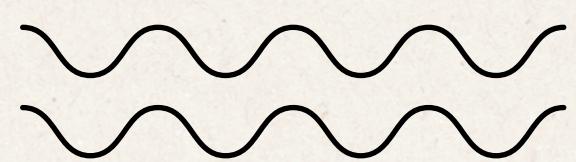
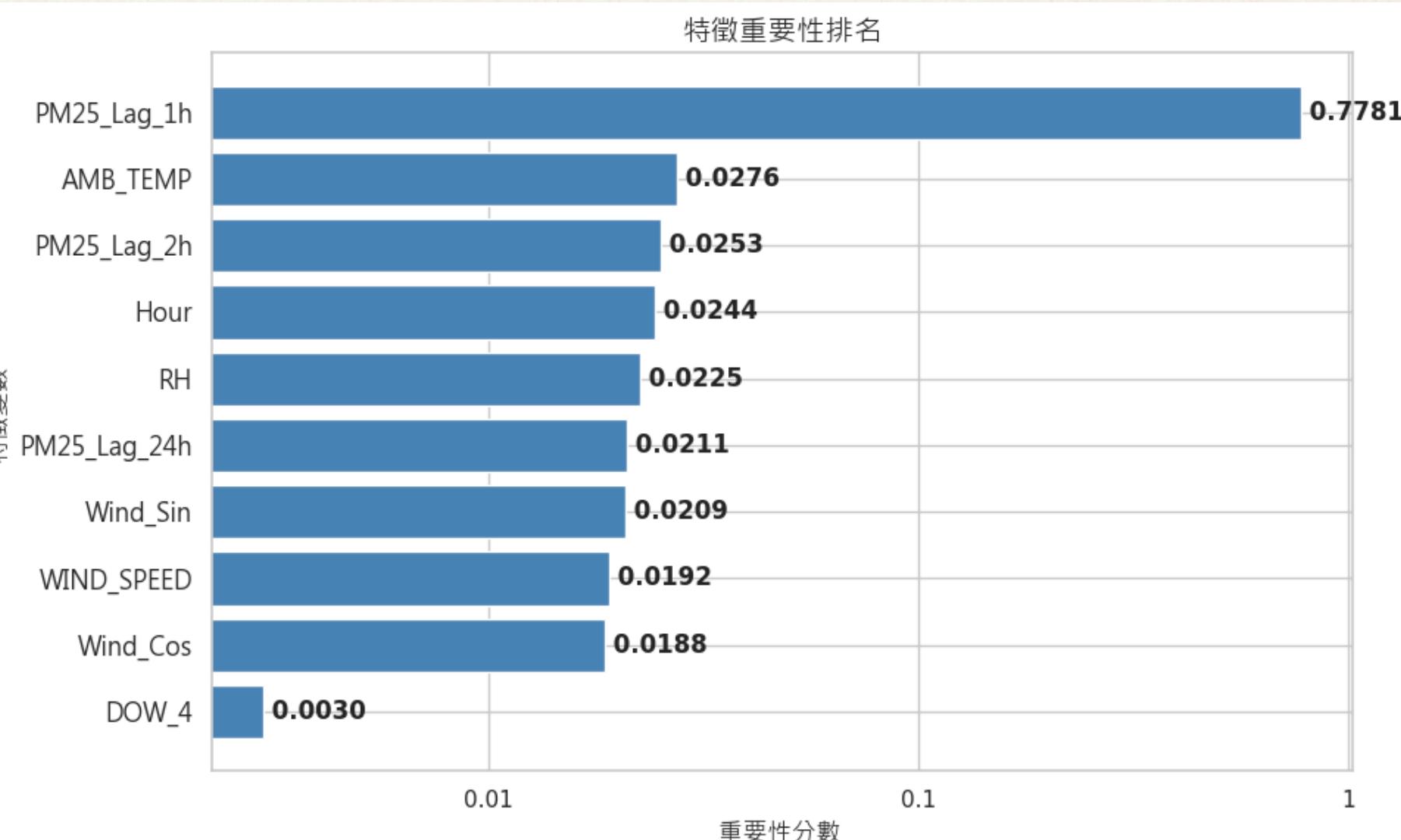
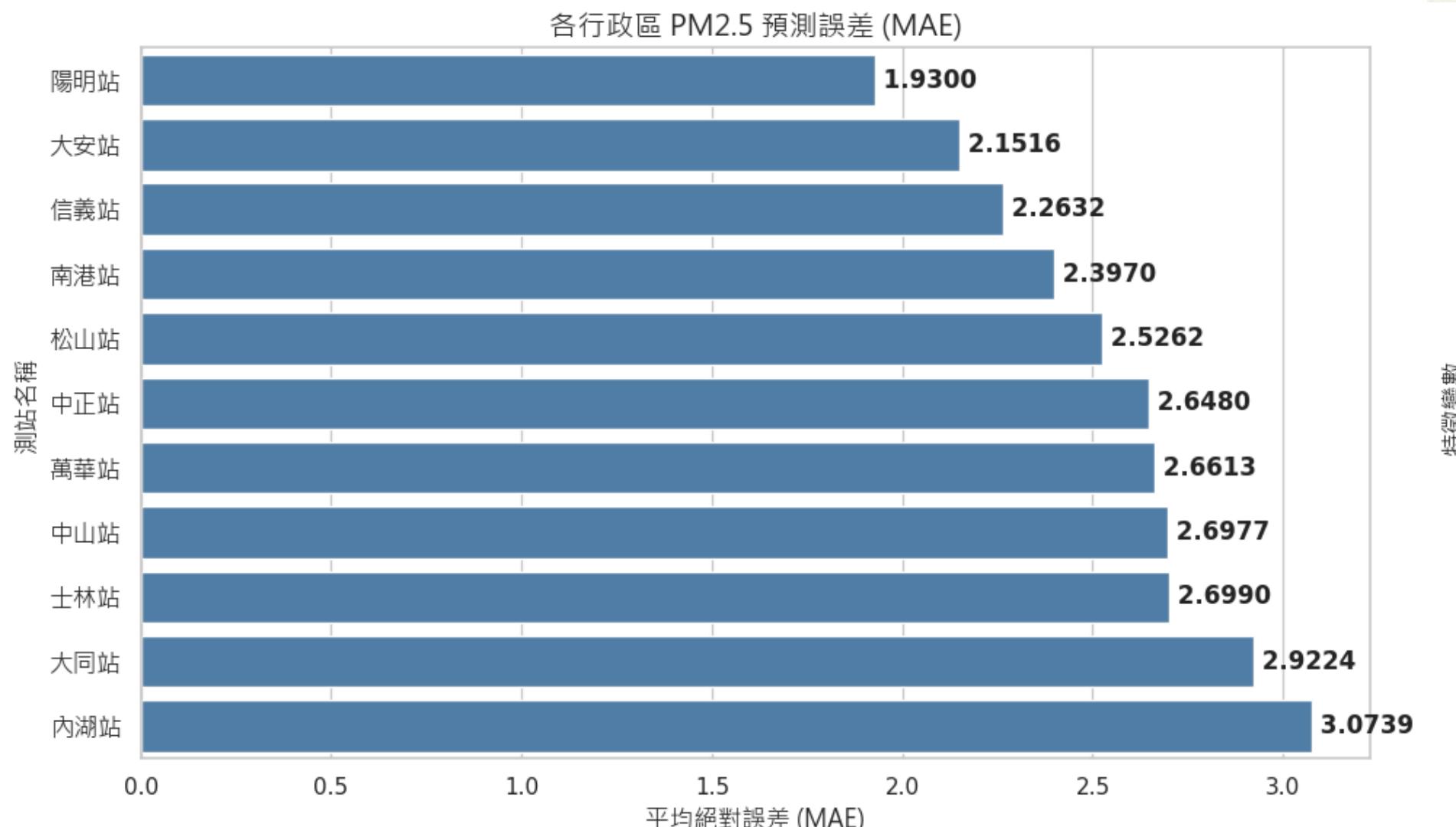
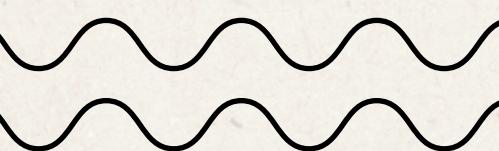
六、所有變數與 PM2.5 的相關性

我們首先分析了氣象因素，發現濕度是氣象條件中與 PM2.5 負相關最強的變數 (-0.32)。然而，若加入歷史數據比較，會發現前一小時的 PM2.5 濃度才是最關鍵的預測指標 (0.88)，顯示空氣汙染具有高度的持續性。

各變數與 PM2.5 的相關係數排序	
PM2.5_Value	1.000000
PM25_Lag_1h	0.880065
Wind_Cos	0.142573
AMB_TEMP	0.113771
Hour	0.092221
Wind_Sin	-0.094213
WIND_SPEED	-0.186197
RH	-0.319169



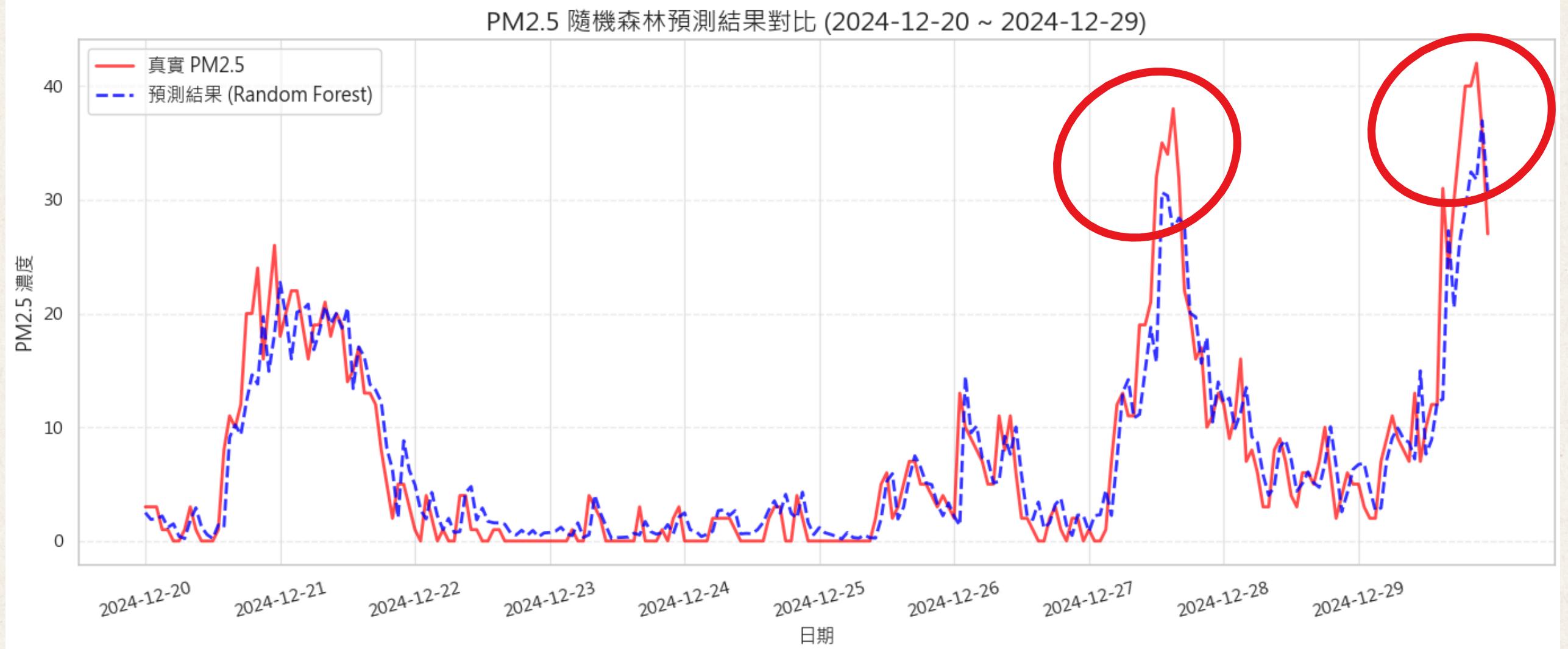
七、研究方法與評估-隨機森林



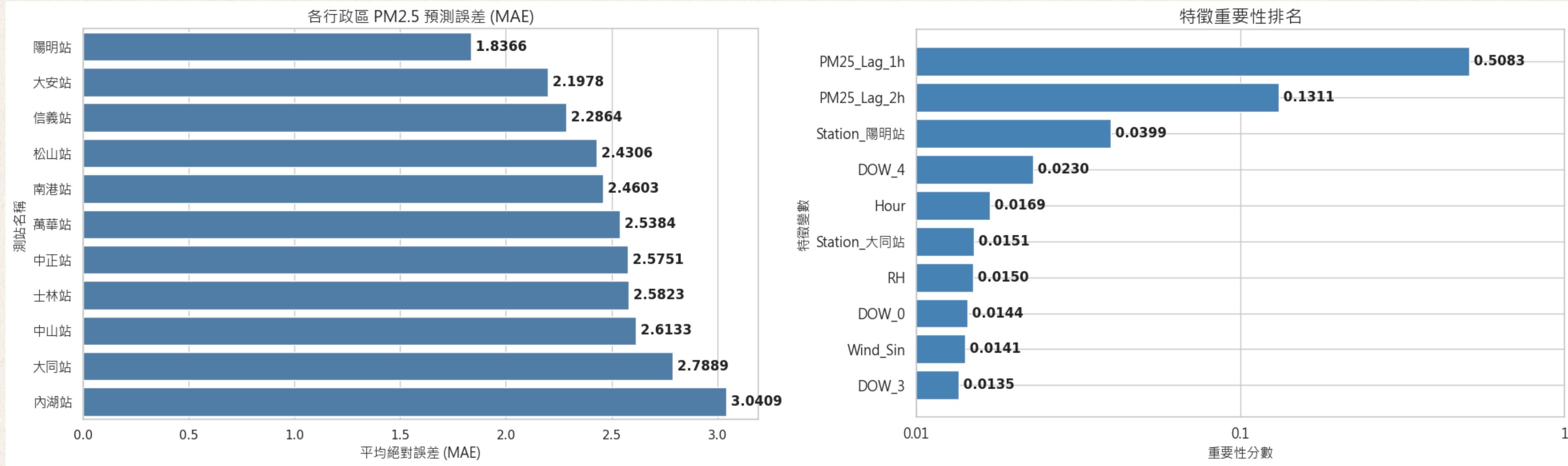
七、模型評估-隨機森林

隨機森林是並行訓練多棵樹，然後取平均。如果你的訓練資料中確實有高污染的歷史數據，那麼森林中會有許多樹預測出高值。平均下來，整體預測值就能衝得很高。

R^2 Score: 85.48% MAE:2.2986



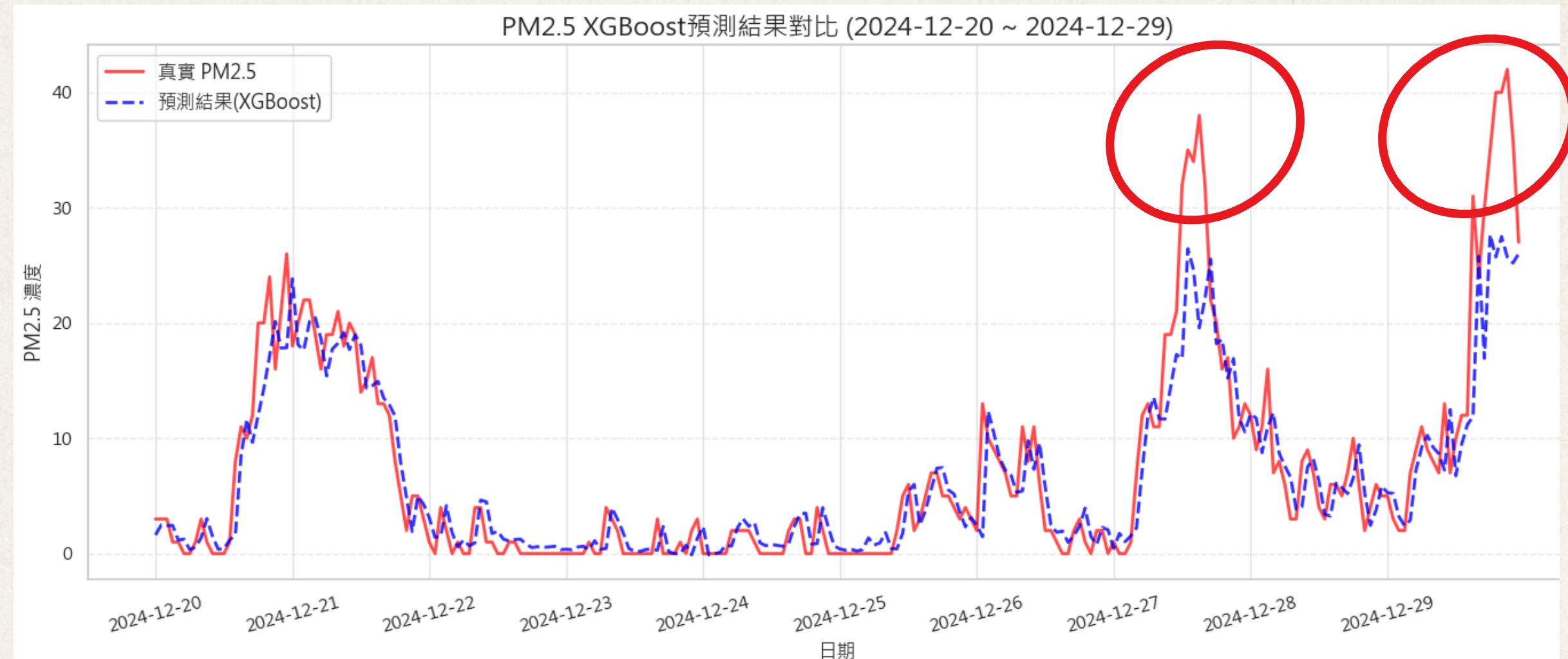
七、研究方法與評估-XGBoost



七、模型評估-XGBoost

由於PM25_Lag_1h的相關係數最高。當PM2.5平穩變化時，預測很準。但突然飆升時，Lag_1h的數值還是低的。模型看到前一小時是20，它傾向於預測接下來是22或23，追不上真實數據。

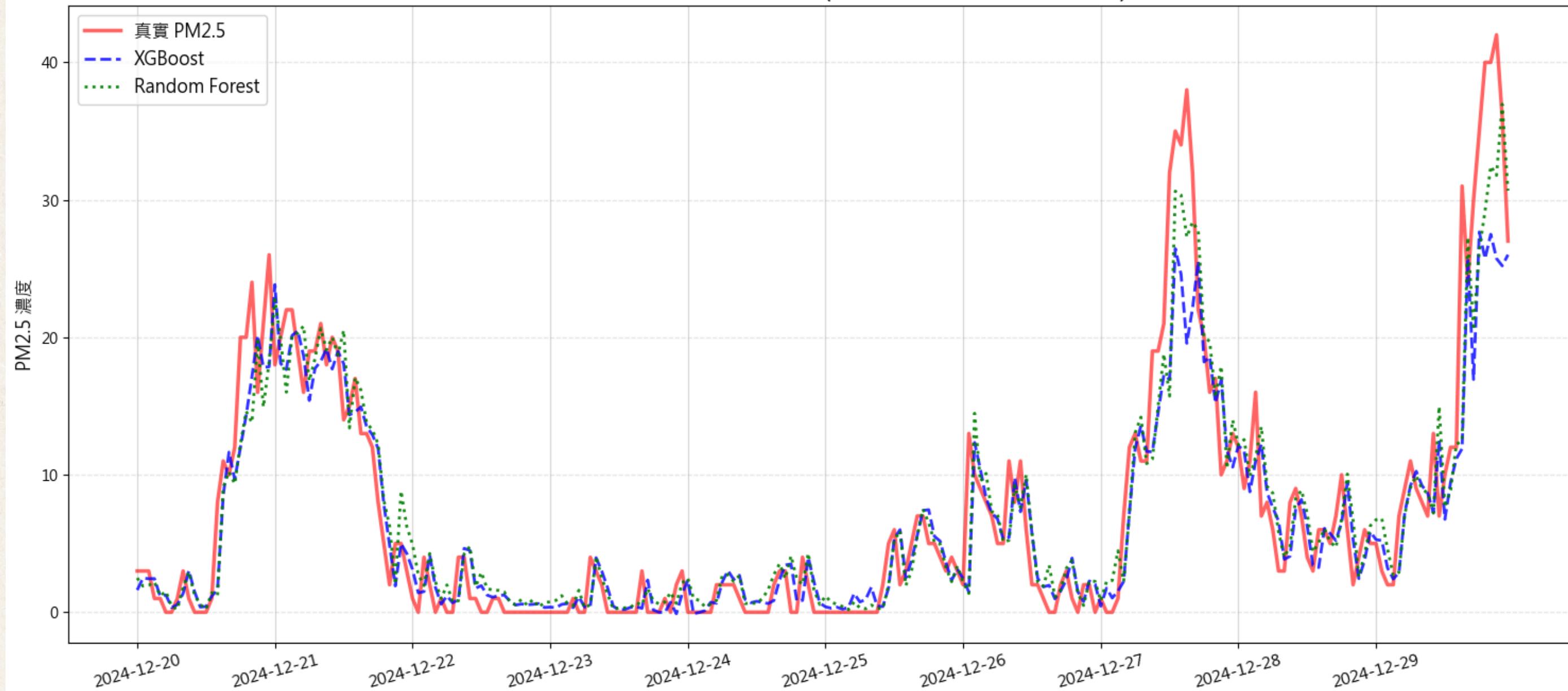
R^2 Score: 82.43% MAE:2.3596



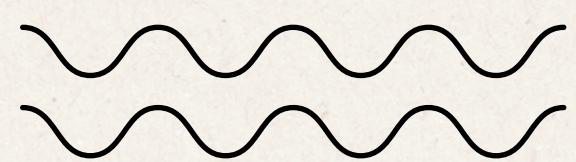
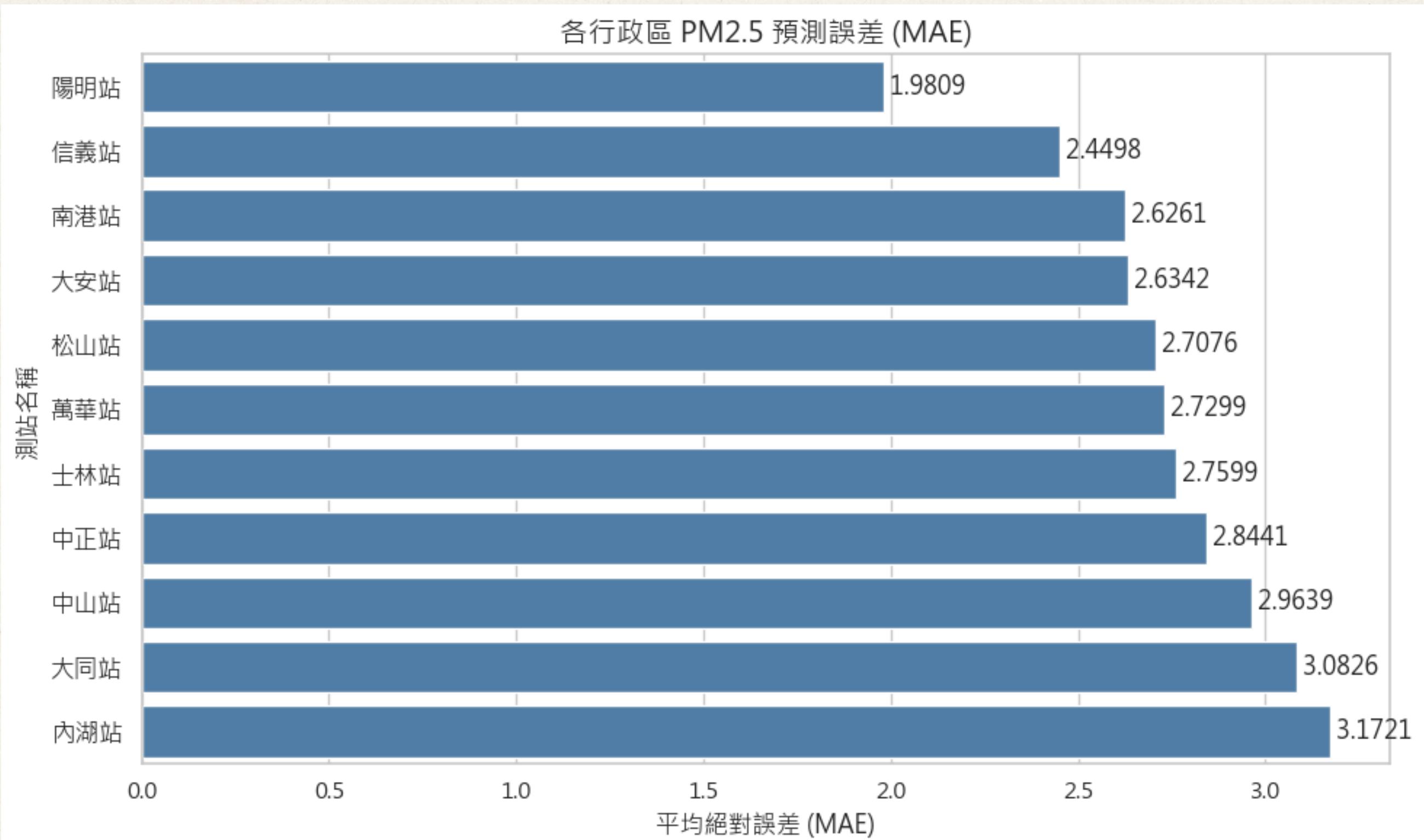
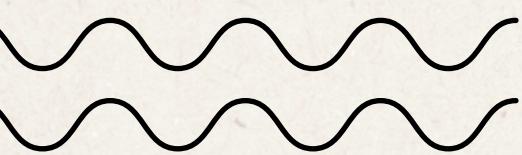
Showcase

模型	R2 Score	MAE (平均誤差)
XGBoost	0.8243	2.3596
Random Forest	0.8548	2.2986

XGBoost vs Random Forest (2024-12-20 ~ 2024-12-29)



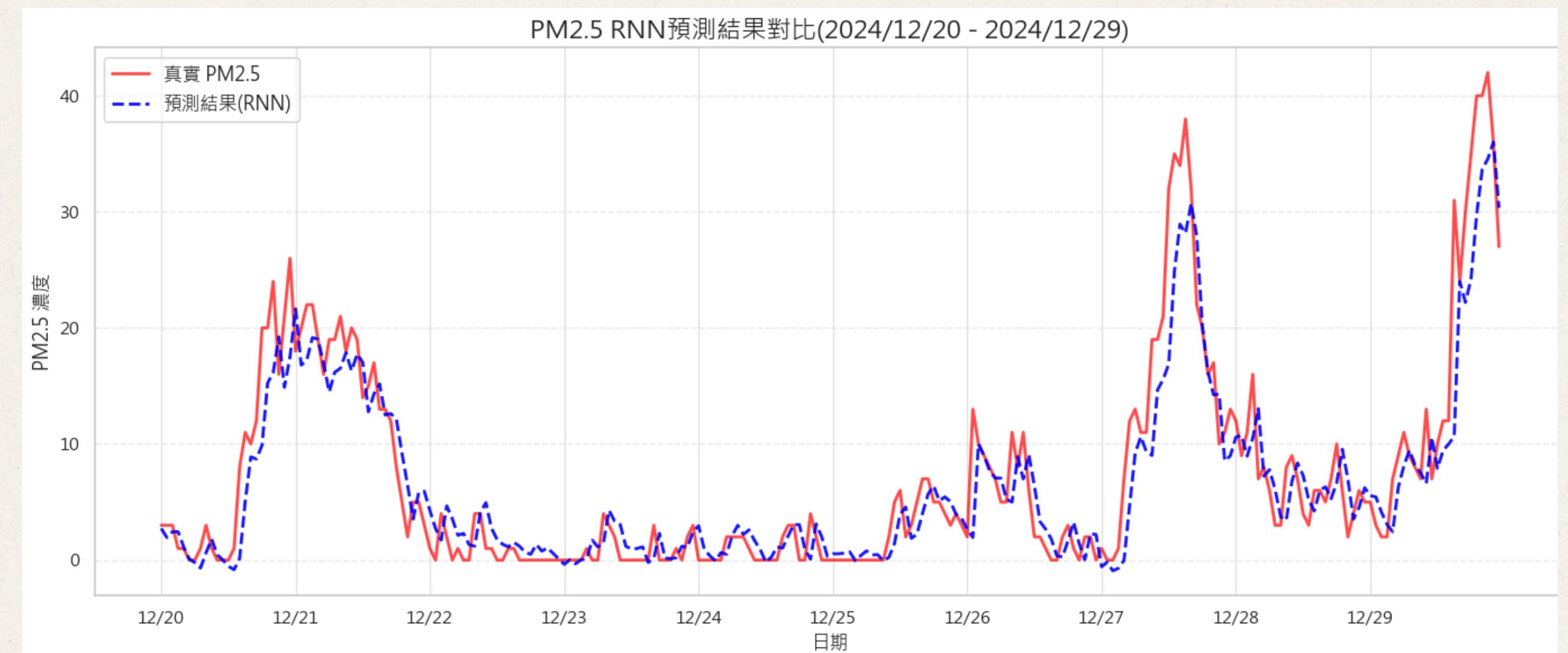
七、研究方法與評估-RNN



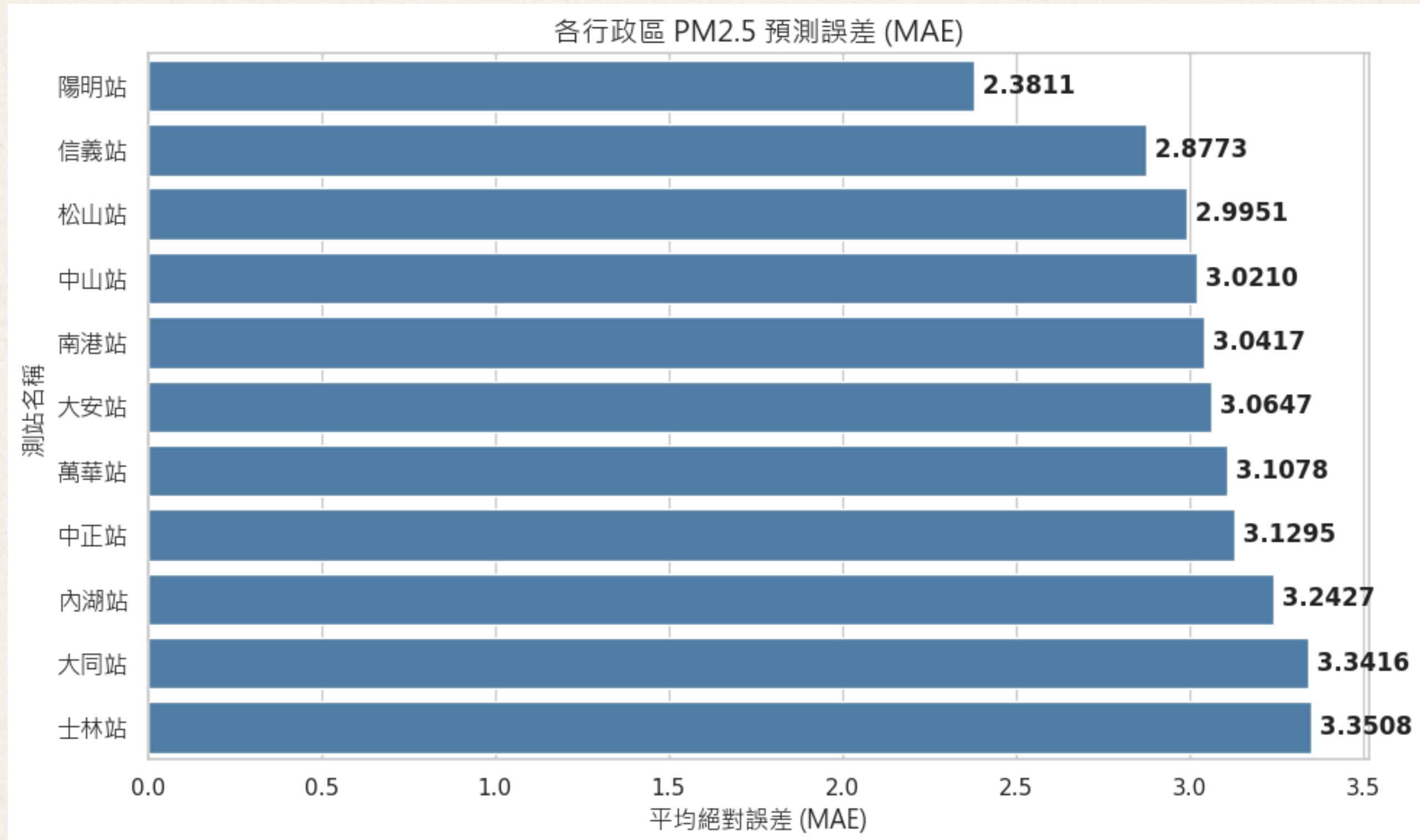
七、模型評估-RNN

R^2 Score: 86.08% MAE: 2.2400

因為 PM2.5 是典型的時間序列資料，RNN、LSTM 的記憶機制對應了這個特性，可以更好的記住前一小時的 PM2.5 數值。

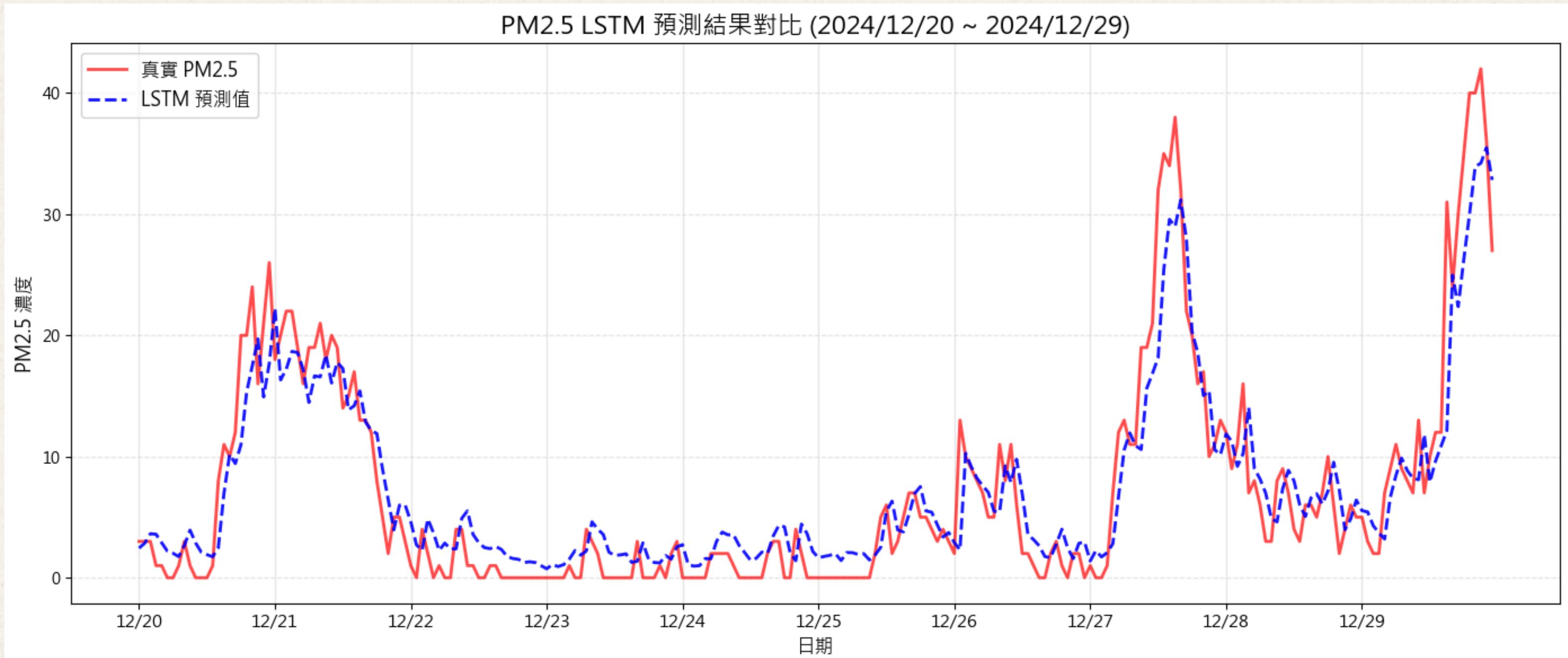


七、研究方法與評估-LSTM



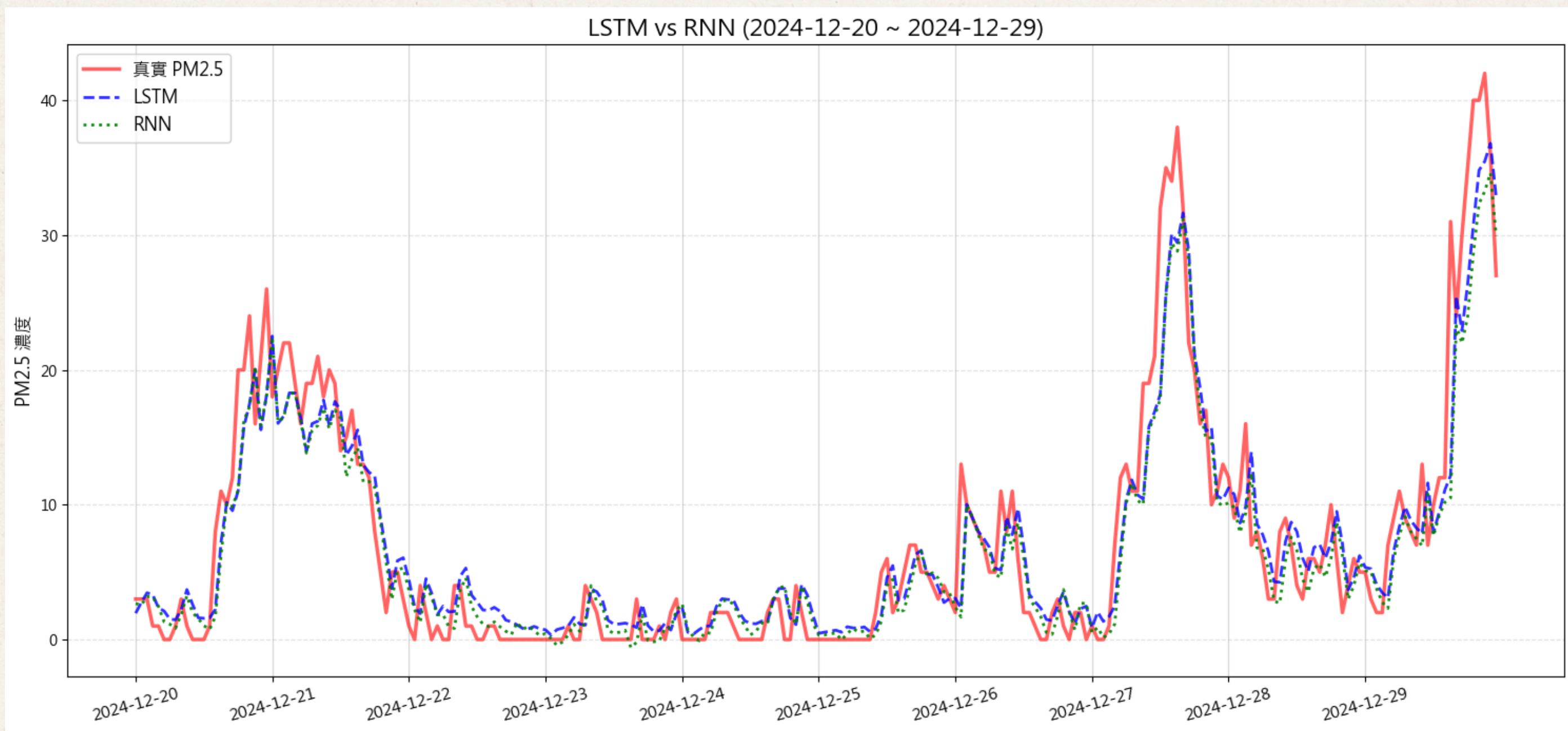
七、模型評估-LSTM

R^2 Score: 87.48% MAE: 2.2609



Showcase

模型	R2 Score	MAE (平均誤差)
LSTM	0.8748	2.2694
RNN	0.8608	2.2400





我們使用同樣的四種模型聚焦分析2024/12/31當天的PM2.5，一小時為一單位，我們想要看在短時間內模型的準確度。

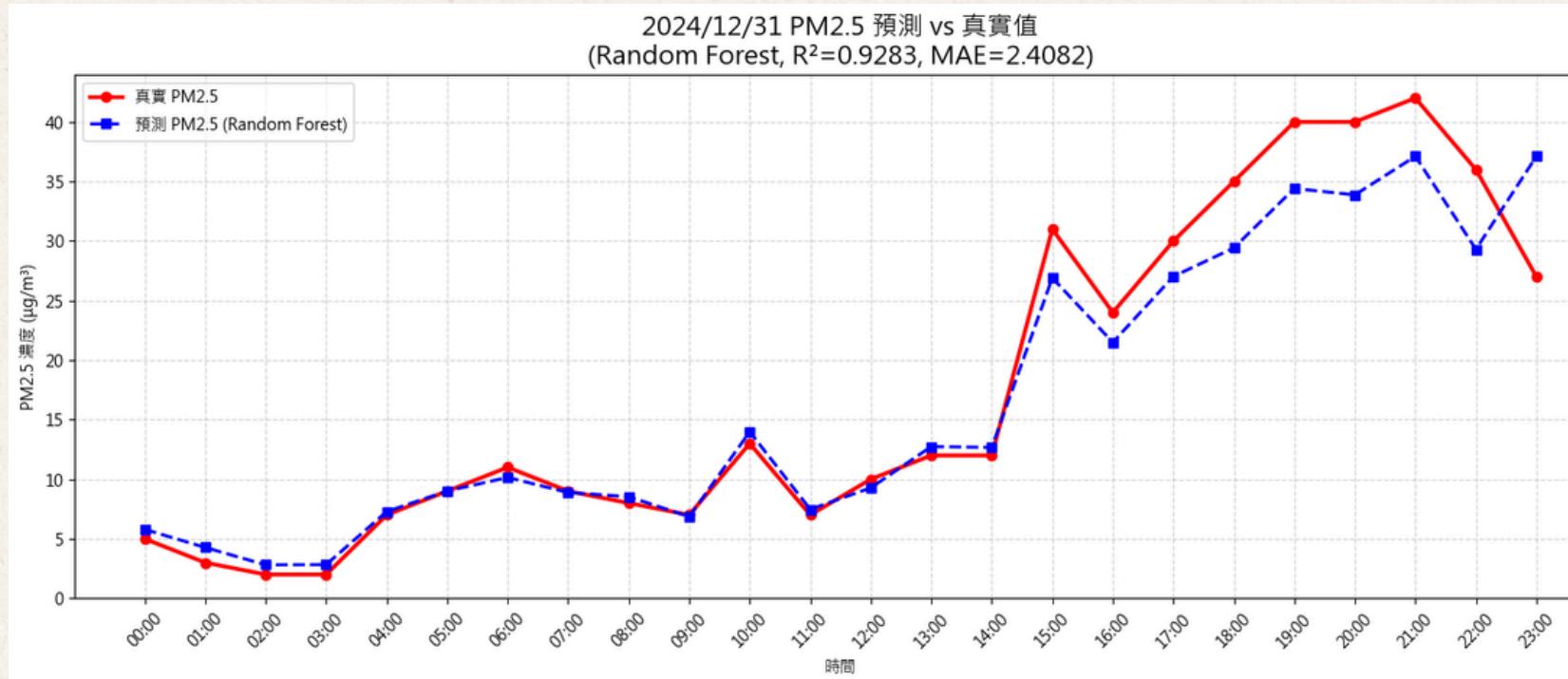


lee

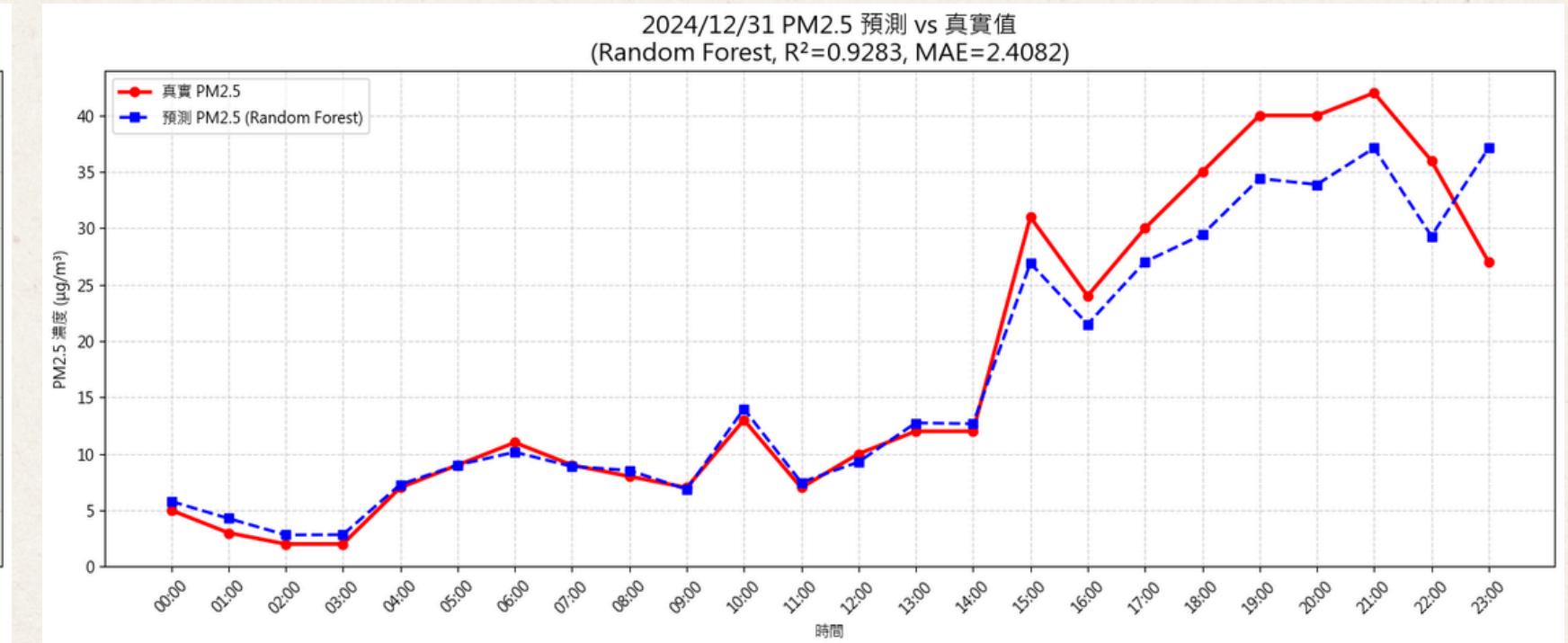
八、12/31之Random Forest效能驗證

聚焦2024/12/31當天的PM2.5

使用Random Forest方法



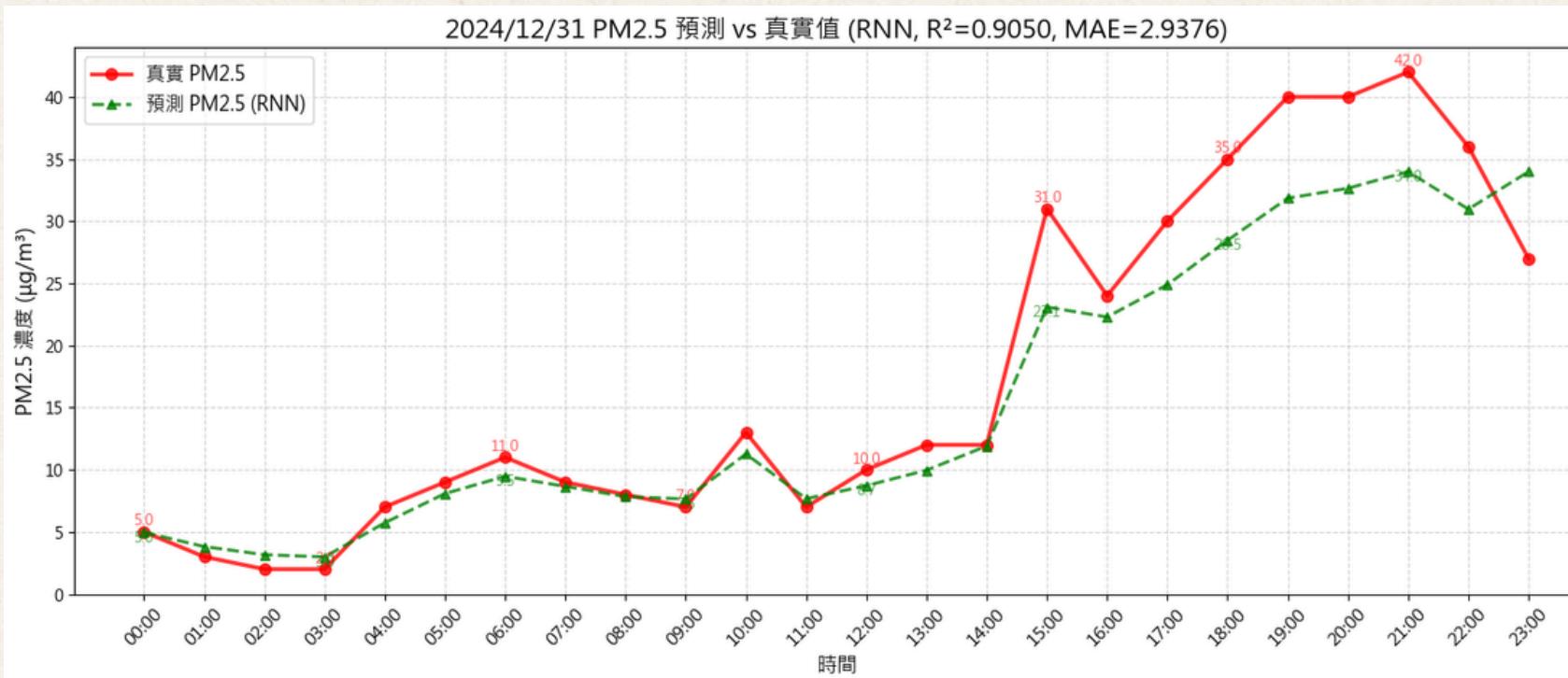
使用XGBoost方法



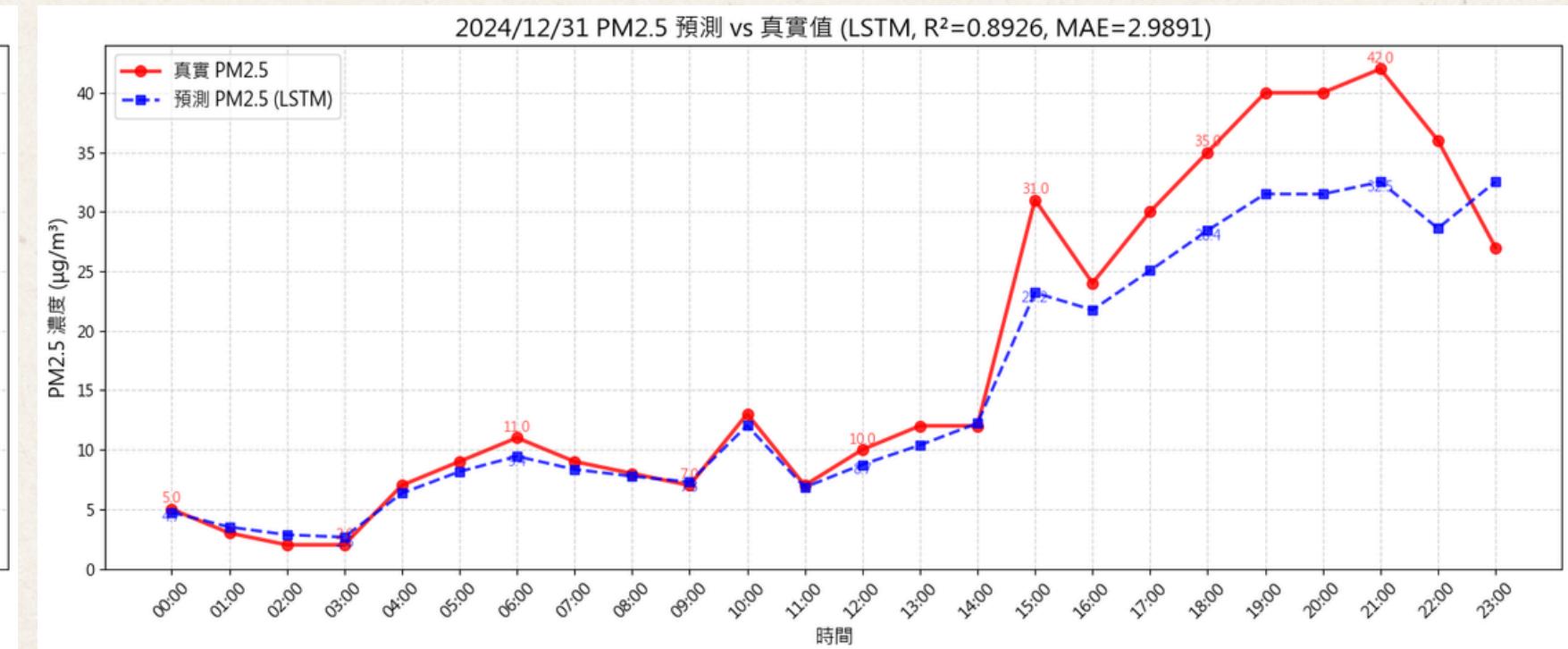
八、12/31之Random Forest效能驗證

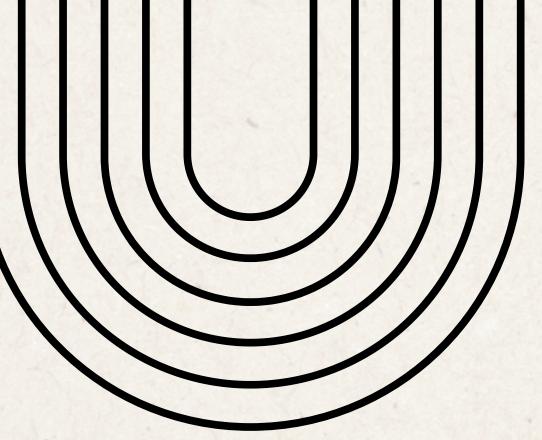
聚焦2024/12/31當天的PM2.5

使用RNN方法



使用LSTM方法





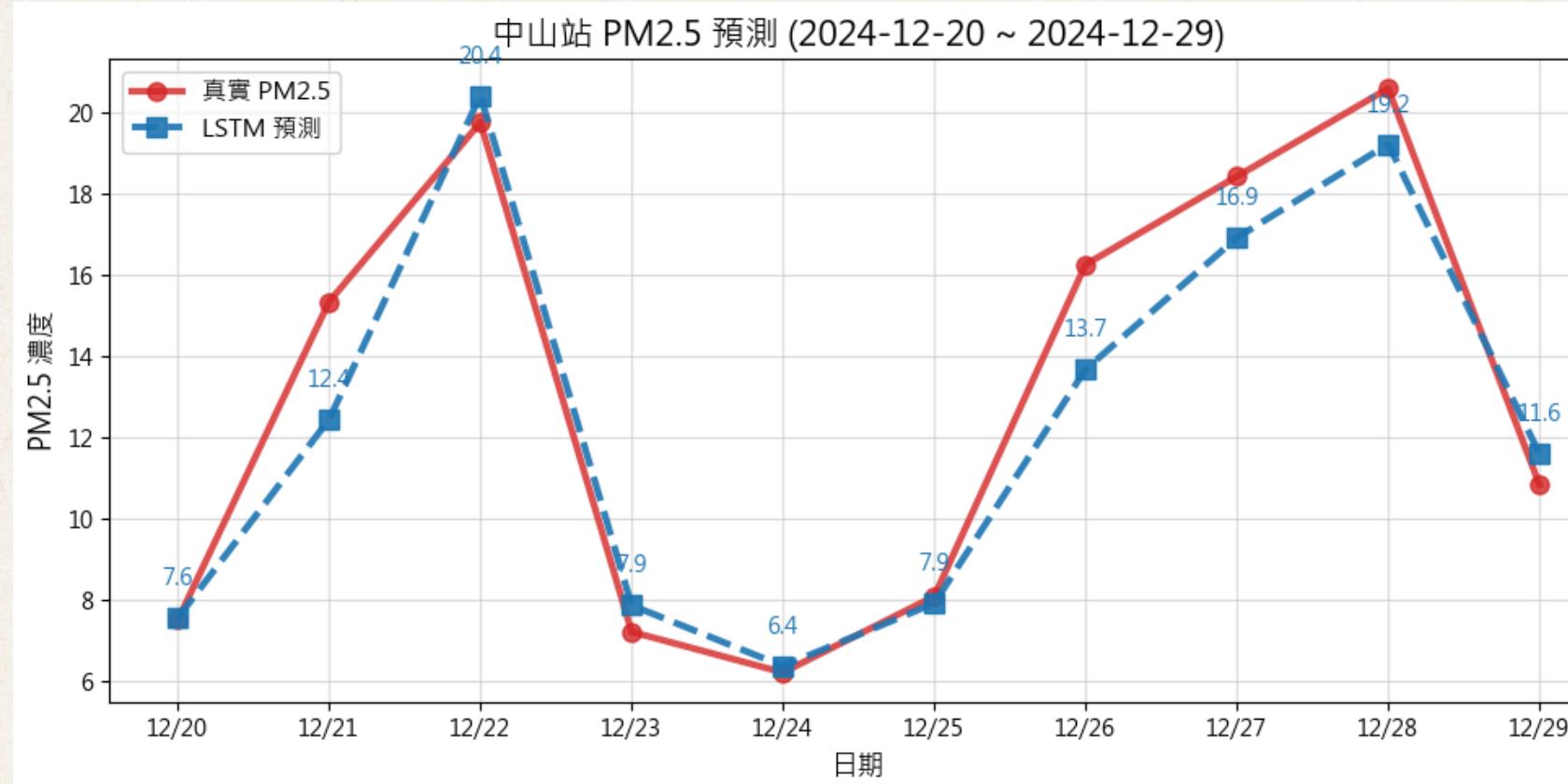
四種方法之效能驗證—結果

- 我們發現所有模型在傍晚至夜間（16:00-23:00）誤差明顯增加，我們推測是日落後大氣擴散條件變差，加上下班尖峰的排放增加，導致汙染物快速累積，形成模型難以捕捉的非線性暴增。
- 根據我們前面分析的結果，由於LSTM的模型準確度都非常高，因此我們決定使用LSTM分析氣象站點的PM2.5，以其中三個測站舉例。

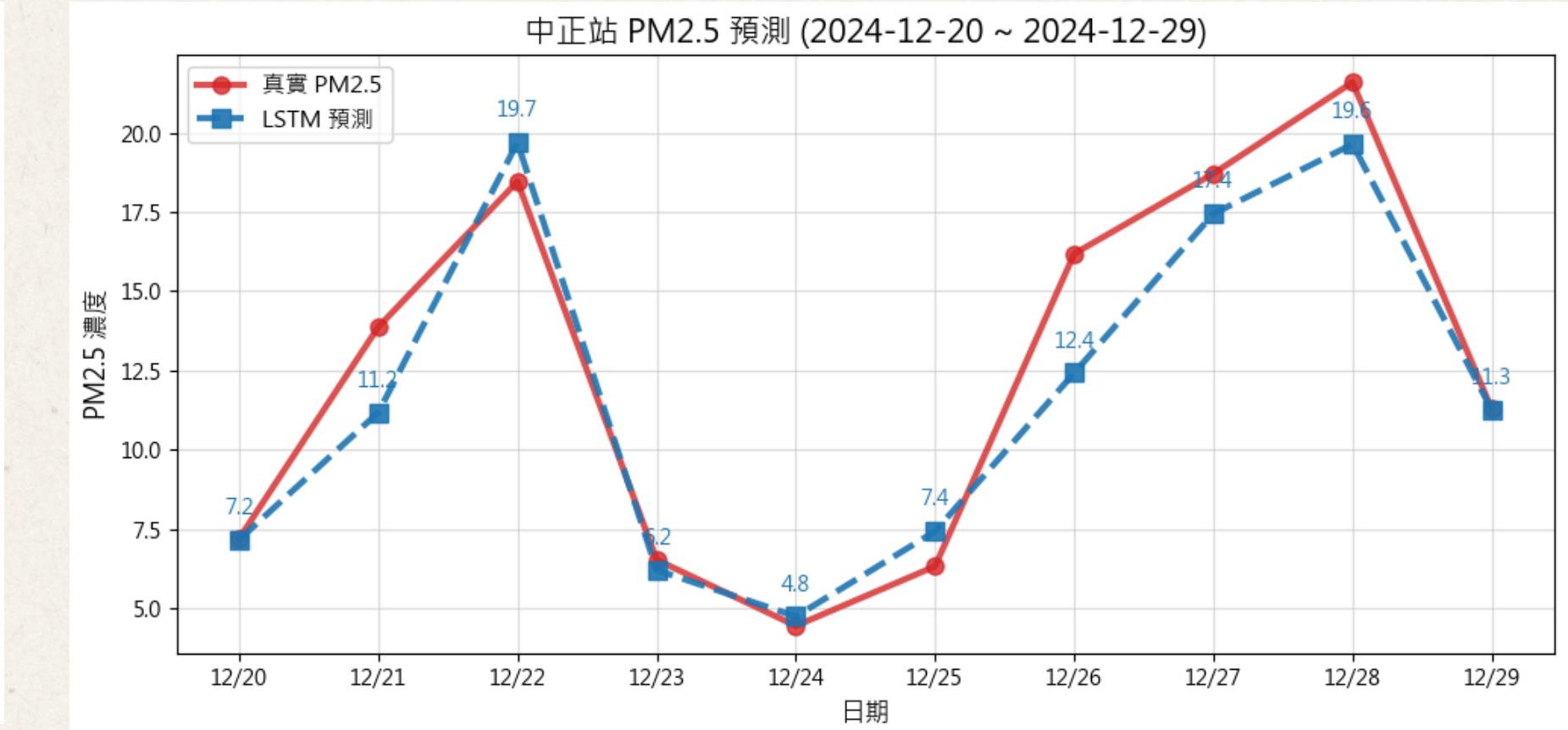
九、LSTM 氣象測站預測

針對各氣象測站2024/12/20-2024/12/29使用LSTM方法

中山站



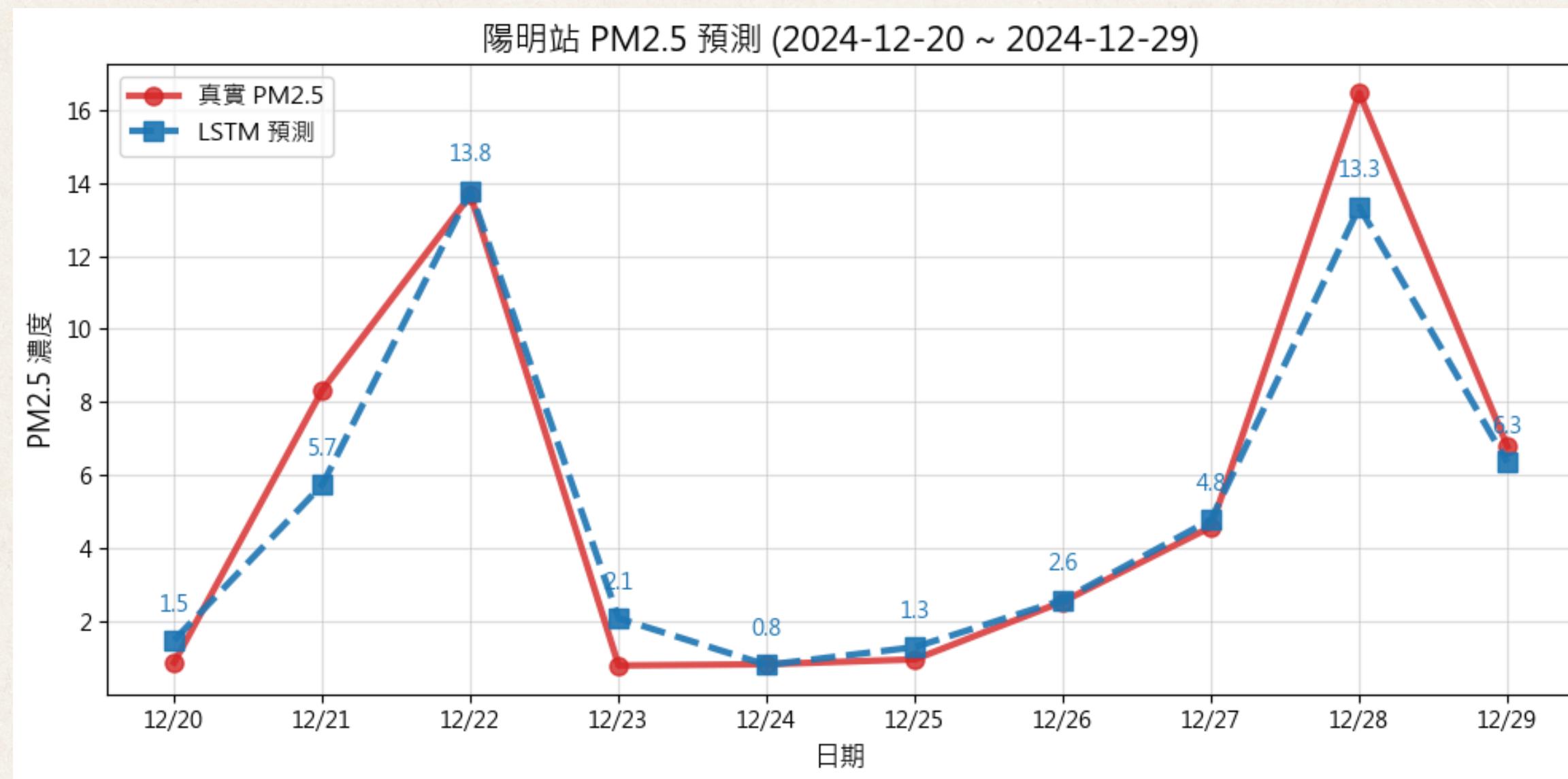
中正站



九、LSTM 氣象測站預測

針對各氣象測站2024/12/20-2024/12/29使用LSTM方法

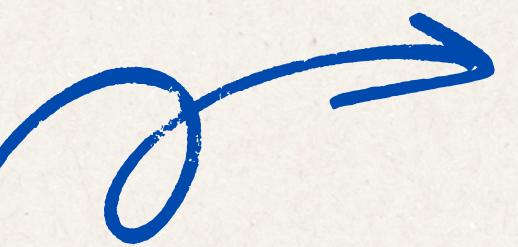
陽明站





Finally

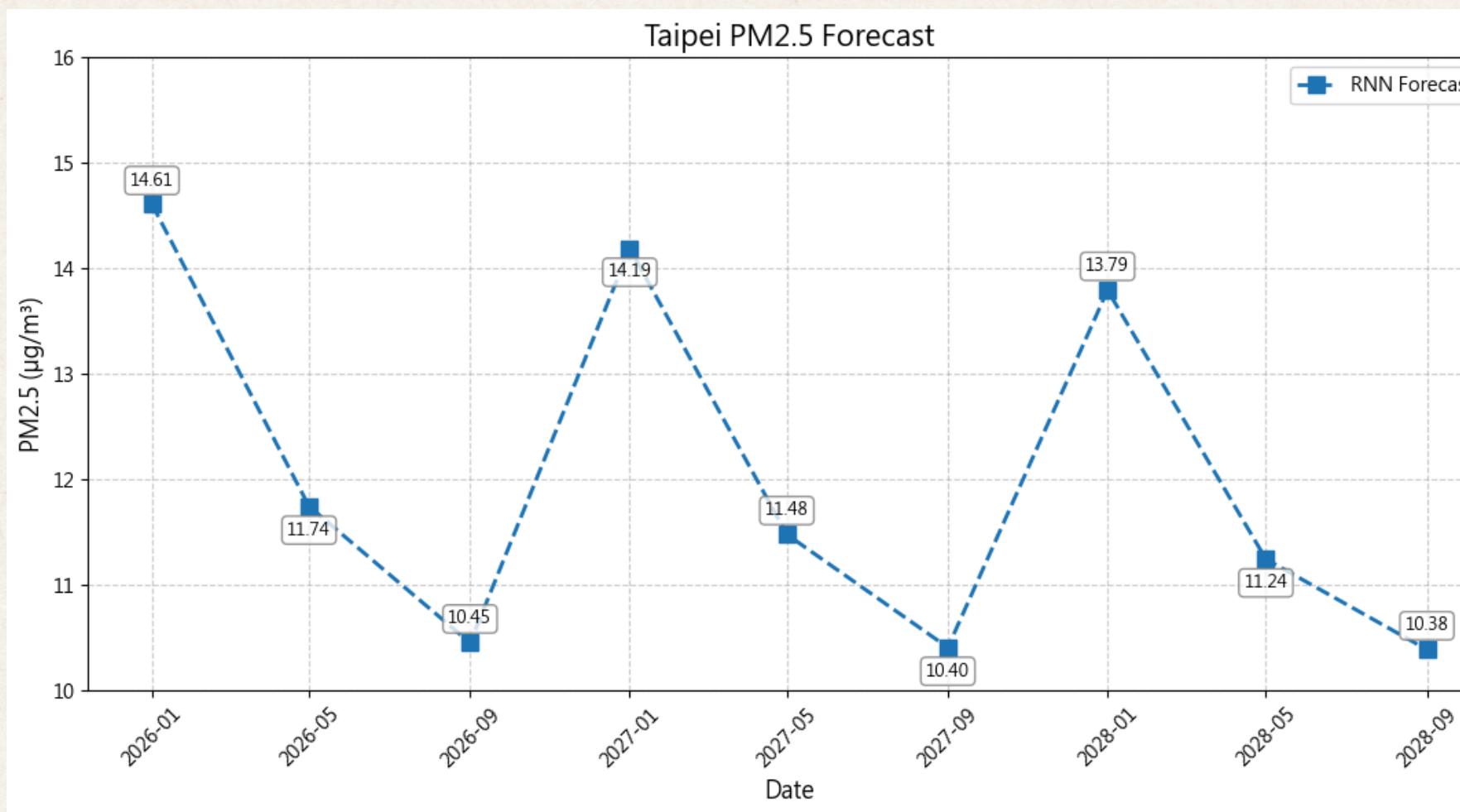
我們挑選模型準確度較高的RNN和LSTM來做2026-2028
台北市PM2.5的預測。



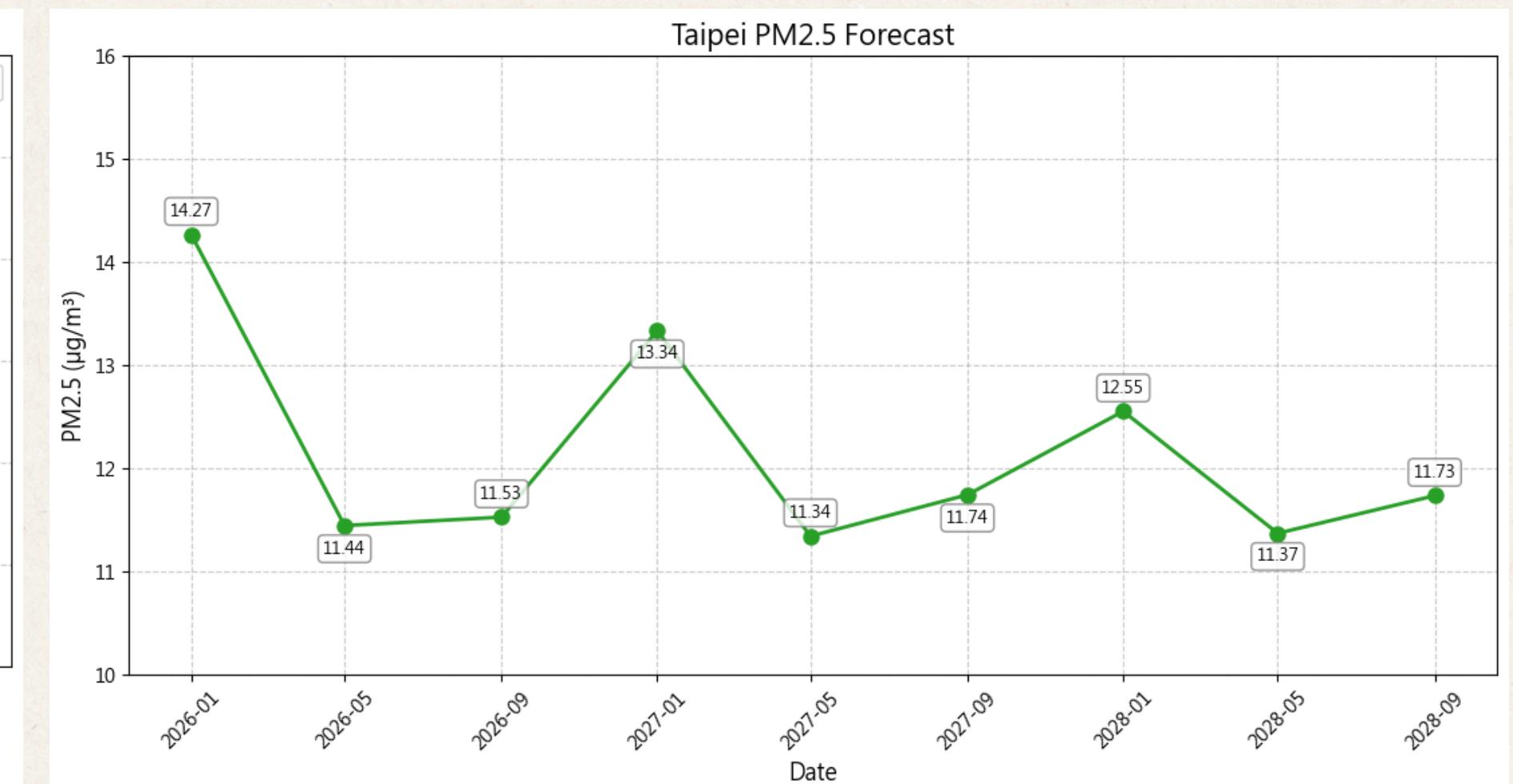
lee

2026-2028台北市PM2.5使用RNN預測

使用RNN方法



使用LSTM方法



十、研究結果與分析

- 根據相關性矩陣與特徵重要性分析，我們發現「前一小時的PM2.5 數值」與當下數值的相關係數高達 0.88，遠高於相對濕度或風速。這證實了「空氣污染具有極強的時間延續性」。
- 利用表現最穩定的RNN 與LSTM模型對台北市未來三年進行預測，結果顯示PM2.5濃度將呈現季節性波動。每年1月至3月仍是污染高峰期，數值約落在13-14；而夏季受對流旺盛影響，數值將回到10-11左右。整體而言，模型預測未來三年台北市的PM2.5平均濃度將維持在11-14的區間。

十一、結論

我們運用機器學習與深度學習技術，建立了針對台北市 PM2.5 的短期預測模型。研究證實，透過整合歷史數據與氣象變數，能夠有效預測空氣品質變化。在模型比較中，**隨機森林**在捕捉「突發性高污染」事件上表現最為優異，適合用於防災預警；而 **RNN**和**LSTM**則在整體趨勢的穩定性上略勝一籌，適合用於「長期」的趨勢分析。

十二、參考文獻

- 環境部氣候
- 政府公開平台(PM2.5)
- Codis氣候觀測資料查詢
- A Machine Learning-Based Ensemble Framework for Forecasting PM2.5 Concentrations in Puli, Taiwan
- 基於深度學習技術應用於空氣品質PM2.5預測

十三、分工表



12156206 章祖綸	word製作、評估模型預測PM2.5、特徵篩選
12156217 陳翡翠	收集資料、資料預處理、處理程式碼
12156223 吳承瑀	ppt製作、處理程式碼
12156229 高碩辰	word製作、評估模型預測PM2.5、特徵篩選

Thank you !