

## Python 資料科學實作

### 資料分析洞察報告：PM2.5 空氣品質分析 – 以 104 年新竹監測站資料為例

#### ● 摘要說明

本篇報告以環保署空氣品質監測網提供之 104 年新竹監測站資料為例，分析 PM2.5 空氣品質相關因子與 PM2.5 預測模型建立(以新竹地區資料)，藉以瞭解目前新竹地區 PM2.5 影響狀況及重要組成因素。

#### ● 資料集內容

資料來源網址(<http://taqm.epa.gov.tw/taqm/tw/AqiForecast.aspx>) 環保署空氣品質監測網，此資料為 104 年度每小時新竹監測站收集之空氣品質相關量測數據，其中包括測項如下表所示：

測項簡稱	單位	測項名稱
SO2	ppb	二氧化硫
CO	ppm	一氧化碳
O3	ppb	臭氧
PM10	$\mu\text{g}/\text{m}^3$	懸浮微粒
PM2.5	$\mu\text{g}/\text{m}^3$	細懸浮微粒
NOX	ppb	氮氧化物
NO	ppb	一氧化氮
NO2	ppb	二氧化氮
THC	ppm	總碳氫化合物
NMHC	ppm	非甲烷碳氫化合物
CH4	ppm	甲烷
UVB	UVI	紫外線指數
AMB_TEMP	°C	大氣溫度
RAINFALL	mm	雨量
WIND_SPEED	m/sec	風速(以每小時最後 10 分鐘算術平均)
WIND_DIREC	degress	風向(以每小時最後 10 分鐘向量平均)
WS_HR	m/sec	風速小時值(以整個小時算術平均)
WD_HR	degress	風向小時值(以整個小時向量平均)

#### 1. 檔案格式說明

a.每個檔案第 1 列為表頭，格式說明如下：

監測日期（西元年/月/日），測站名稱，測項簡稱，監測時間（0-23 時）

b.第 2 列以後為測值，內容如表頭所述，單位如下。

c.0 時：指 0:00-0:59，2 時：指 2:00 - 2:59...，23 時：指 23:00-23:59。

## 2. 普通測站資料註記說明：

- # 表示儀器檢核為無效值。 \* 表示程式檢核為無效值。  
 x 表示人工檢核為無效值。 NR 表示無降雨。  
 空白 表示缺值。

## ● 分析方法

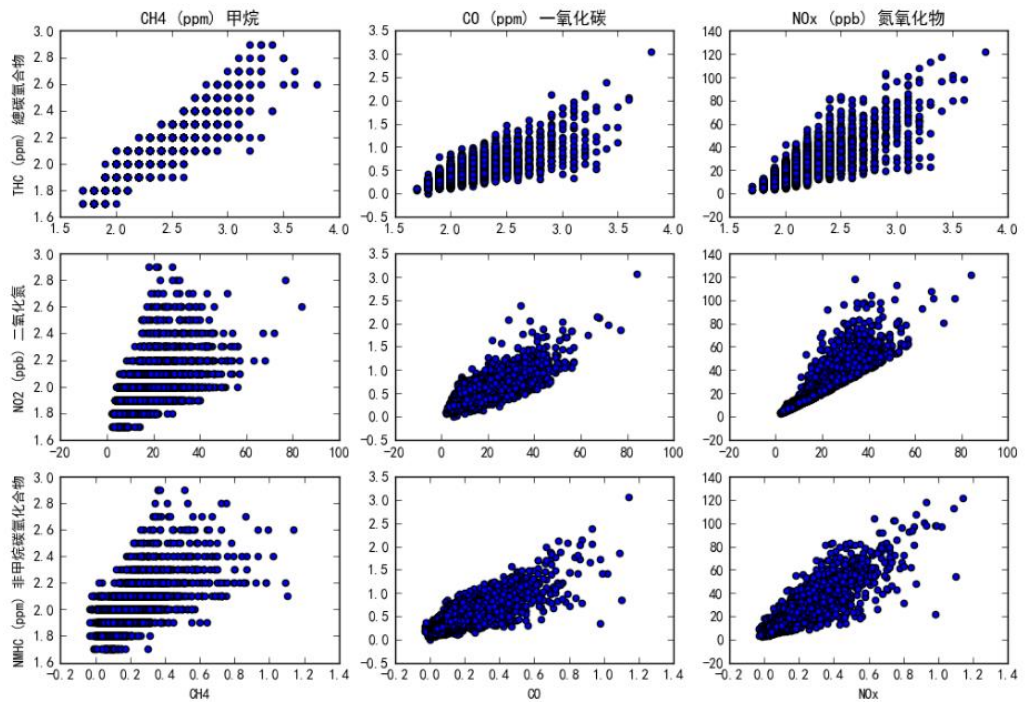
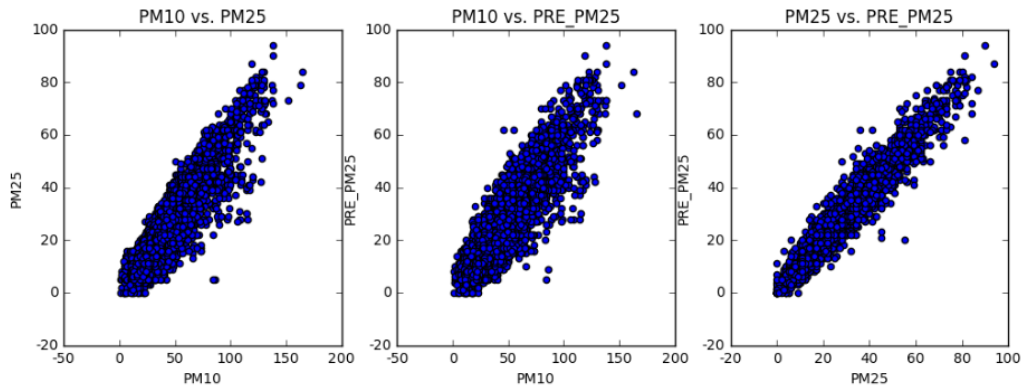
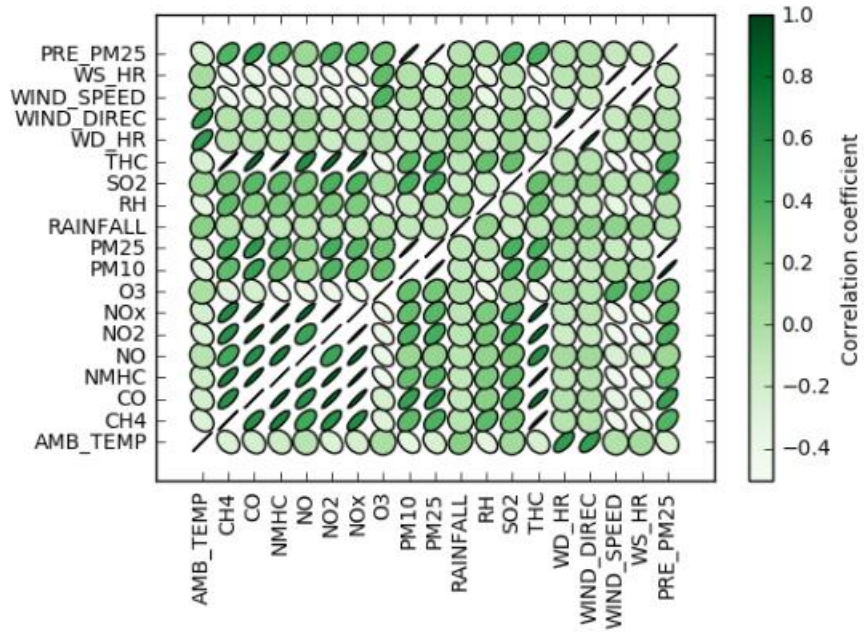
### 1. 資料預處理：

原始資料格式如下表所示，資料是以日期/測站/測項/小時 0-23 為欄位，資料中包括了一些空值或註記的無效資料，因此，在匯入 Python 之前，先利用 SQL 指令將欄位格式轉為：日期/小時/測項(AMB\_TEMP,CH4,CO,NMHC,NO,..etc.)之資料格式，以利 Python 匯入後進行後續分析。

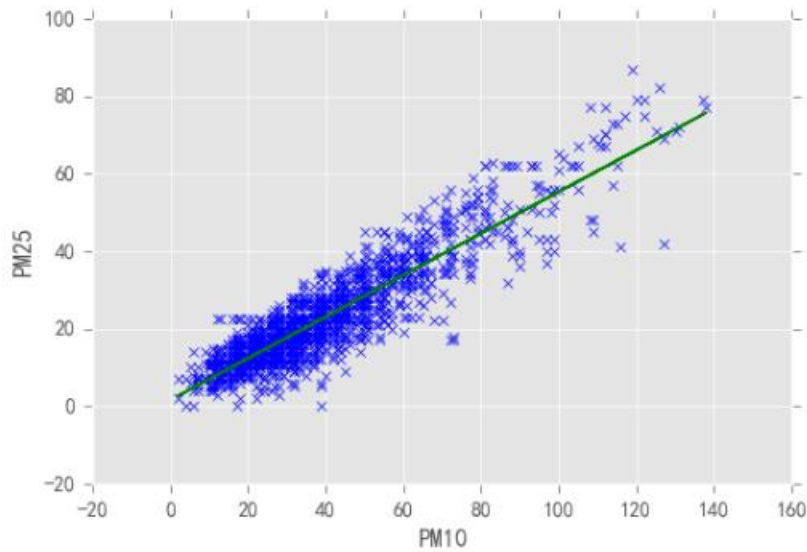
日期	測站	測項	0	1	2	3	4	5	6
2015/1/1	新竹	AMB_TEMP	17#	18	17#	15	14	14	13
2015/1/1	新竹	CH4	1.9#	1.9#	1.9#	2		1.9	1.9
2015/1/1	新竹	CO	0.47#	0.51#	0.47#	0.69	0.62	0.53	0.53
2015/1/1	新竹	NMHC	0.13#	0.11#	0.13#	0.14	0.13	0.11	0.13
2015/1/1	新竹	NO	2.8#	2.8#	2.8#	0.8	0.7	0.6	1
2015/1/1	新竹	NO2	18	16	15	15	13	11	13
2015/1/1	新竹	NOx	19	17	16		14	12	14
2015/1/1	新竹	O3	36	36	37	35	35	35	32
2015/1/1	新竹	PM10	132	165	163	152	134	119	104
2015/1/1	新竹	PM2.5	68	84	79	73	65	55	45

### 2. 因子相關性分析：

利用相關性分析初步瞭解各因子(測項)之相關程度，以利後續挑選適合因子來建立預測模型。



### 3. 迴歸模型建立：



第一組迴歸模型：PM2.5 與 PM10 之間的相關模型

各變項參數：

$$Y: (PM2.5) = a (PM10) + b$$

[[ 0.5372407]]

均方誤差 (Mean squared error, MSE): 30.02

第二組多元迴歸模型：THC 與 CH4、CO、NMHC、NO2、NOx，氮/氧化合物之間的相關模型

各變項參數(normalized):

$$Y: (THC) = a (CH4) + b (CO) + c (NMHC) + d (NO2) + e (NOx) + f$$

[[ 0.11987631 0.00190413 0.11319619 0.00163634 0.00029829]]

均方誤差 (Mean squared error, MSE): 0.00

第三組多元迴歸模型：PM2.5 與 WD\_HR、WIND\_DIREC、WIND\_SPEED、WS\_HR 風速/向的模型

各變項參數(normalized):

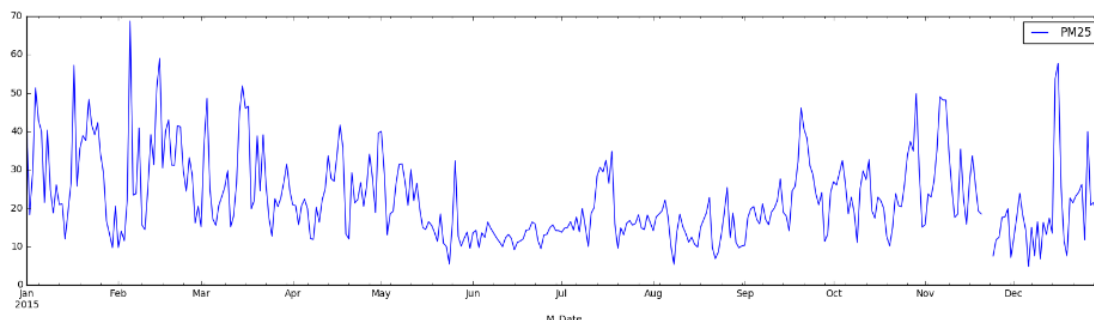
$$Y: (PM2.5) = a (WD\_HR) + b (WIND\_DIREC) + c (WIND\_SPEED) + d (WS\_HR) + e (PRE\_PM25) + f$$

[[ -0.29075421 0.15607838 0.05713568 -0.34364879 12.25110573]]

均方誤差 (Mean squared error, MSE): 12.28

## ● 資料分析結果與洞察

我們由 2015 年度每日 PM2.5 均值趨勢線圖(如下)可知一月至三月(春)及九月中旬至十二月(秋冬)，新竹地區的 PM2.5 均值較四月至八月(夏)高，由氣象資料推測九月中旬後，受大陸東北季風帶來的境外移入 PM2.5 霾害影響，大約比夏季多 30 ug/m<sup>3</sup>。



由相關分析圖可以知道目前空氣品質測項可分為三種類型，第一類由 PM2.5/PM10/PRE\_PM2.5 組成，第二類由 THC、CH<sub>4</sub>、CO、NMHC、NO<sub>2</sub>、NO<sub>x</sub> 等，氮/氧化合物組成，第三類為風速、風向的測量數據。我們進一步建立迴歸模型來瞭解此三種類型量測因子具有一定的相關性，PM10 為 PM2.5 較大顆粒的懸浮微粒，污染源較 PM2.5 來得多 (註 1)，而氮氫化合物大多由交通工具排放，它需要一些反應時間後，一部份會結成懸浮微粒，與 PM2.5 有間接相關。另外，風向與風速在直覺上會影響 PM2.5 濃度高低，但在相關分析中，卻沒有看到有高度相關，我們推估應是風向/風速量測數據為大氣巨觀的數據，PM2.5 較受台灣地區平原、海洋、山脈之間對流影響，故在測度上沒有明顯呈現出來。另外也因為此為單一量測站(新竹)所得到的數據，尚不足以細部代表整個新竹地區的空氣品質狀況，量測點太少造成量測數據不夠的問題。

另外為使 PM2.5 能夠給相關單位有一管理的參考數據，建議應建立一套 PM2.5 與用電量對價關係模型，來做為管理 PM2.5 危害的衡量機制，透過電價的附加成本來使污染源節制空汙排放，這部份應可以做進階研究。

註 1. 台灣常見的污染源：工廠 (30%)、交通 (30%)、境外移入 (30%)、溢散源 (10%) ex. 餐飲。

參考資料: g0v 台灣空汙地圖 <http://airmap.g0v.asper.tw/>

LASS (Location Aware Sensing System) 環境感測器網路系統 <http://lass-net.org/>