國立中央大學 統計研究所

統計實務_書面報告

中國五大城市 PM2.5 之統計分析

第二組

小組成員：陳睦璿 蘭禹筑 康楹婕 徐欣迪 劉懿萱

指導教授：鄒宗山 所長

一、 依城市地理位置由北至南分析 2013-2015 年 PM2.5 濃度

1.1 瀋陽市 PM2.5 濃度分析

瀋陽市以自然資源為基礎的重工業城市，素有「東北魯爾」的美譽，其 PM2.5 為五座城市中較高者，並且波動程度較大。根據相關資料可以將 PM2.5 的影響因素分為人為因素與自然因素。人為因素中以供暖期、重工業排放、秸稈燃燒、汽機車排放為主；自然因素則與瀋陽氣候條件、地理位置等息息相關。瀋陽屬溫帶半濕潤大陸性氣候，春季乾燥多風、夏季溫熱多雨等，造就春夏兩季低 PM2.5 濃度；此外，其地形多為平地、低山、丘陵，地勢起伏小使得空氣流通性差，不易汙染物擴散。

從 PM2.5 與時間變數的關係，可以發現季節相比其他時間變數與 PM2.5 濃度有較明顯的相關性，冬季 PM2.5 汙染程度最為嚴重，春夏季則較為輕微；當逆溫現象生成且風速不大時，利於空氣汙染物的濃度提高。

從 PM2.5 濃度均值與連續型氣候變數的散佈圖(詳見圖一)中可以發現 PM2.5 濃度與風速、溫度、露點溫度呈負相關，與氣壓、相對濕度呈正相關；但除了風速、降雨，大致皆呈現水平帶狀分布，顯示相關性強度皆不高。在 PM2.5 濃度較高的時間，通常與低風速、紊亂風向對應(詳見圖二)。反之，低濃度則與實際風速超過 4 公尺/秒且穩定持續的風向對應。綜上，風速、風向、降雨為主要影響瀋陽市空氣質量的氣候因素。

1.2 北京市 PM2.5 濃度分析

北京市為中國首都，位處華北平原的西北邊緣，其 PM2.5 平均濃度為五座城市中較高者，自 2013 至 2015 年雖有逐年下降趨勢，但在五個城市中變化最不明顯。經外媒分析，北京空汙緊隨經濟規劃而變動，金融業與旅遊業為當時主要產業，交通繁忙造成持續汽機車汙染物排放，致使 PM2.5 濃度下降趨勢較不顯著。

從 PM2.5 濃度均值與時間變數的關係，可以發現在逆溫現象較容易生成的時間，皆有 PM2.5 較高的現象。以季節為時間尺度可以發現冬季的 PM2.5 較夏季高；以天為尺度則有凌晨及傍晚的 PM2.5 濃度會較其他時段略高的現象，但差異並不顯著。另外，因北京屬北方城市，冬季供暖期燃煤亦是造成 PM2.5 濃度較高的可能成因。

從 PM2.5 濃度均值與連續型氣候變數(風向外的氣候變數)的散佈圖(詳見圖三)中我們發現，露點、溫度、壓力與 PM2.5 平均濃度並沒有顯著的線性趨勢，濕度則與 PM2.5 有較明顯的正向關係，累積風速、累積降雨則與 PM2.5 為反向關係。在濃度均值較高之時間，對應的溫度與露點大多落於攝氏 0 度附近，濕度則是在 60-100%的高濕度區間，累積風速與累積降雨則是在較低的區間。

從 PM2.5 濃度均值與風向、風速等關係圖(詳見圖四)中可以發現，吹拂西北風時多數時候可使濃度下降或在低濃度區間震盪，吹拂東南風時則無論吹拂時間長短或實時風速高低皆較無法使 PM2.5 濃度下降。另外，在吹拂西北風

時，瞬時的大風並不會使 PM2.5 濃度有上升的現象。綜上，可以發現風向在北京市對 PM2.5 濃度有較關鍵的影響力，究其原因可能與北京市周邊的城市規劃有關；北京市的東南方為重工業發展區，使得東南風吹拂時會夾帶汙染物，進而造成 PM2.5 濃度上升的現象。

1.3 上海市 PM2.5 濃度分析

上海氣候屬於亞熱帶季風氣候，四季分明，日照充分，雨量充沛，5 月至 9 月間為上海的汛期，降水量達全年的 60%左右。每年的 7 月至 8 月進入伏旱天氣，較之日常月份顯得潮濕酷熱。

上海燃煤主要集中在電力行業且分散燃煤使用較少（冬季沒有採暖），工業的比重仍然偏重，工業門類齊全、生產經濟體量和能源消耗量較大，因此，上海工業生產對 PM2.5 的貢獻率不低，上海揚塵的貢獻率略低於京津冀地區，與近年來加強揚塵治理，且所在長三角區域氣候濕潤，受乾燥天氣下大風揚沙及沙塵傳輸影響較小有關。

從 PM2.5 濃度均值與時間變數的關係，以季節為時間尺度可以發現 PM2.5 具週期性變化，呈冬季濃度高、夏季濃度低等現象。以月份為時間尺度，可以發現 PM2.5 平均濃度最低是發生在 9 月，最高是發生在 12 月。

從 PM2.5 濃度均值與氣候變數的關係(詳見圖五)，可以發現與多數變數的線性關係較不明顯，只與累積風速與累積降雨有較明顯的反向關係。另外，可以發現在露點、溫度、壓力、濕度等變數與 PM2.5 的散佈圖中，資料大多呈水平帶狀的分布，較難看出變數與 PM2.5 的關聯。因上海的地理位置關係，當吹拂東南風與東北風時多可降低 PM2.5 濃度或使其在低濃度區間內震盪，且紊亂的風向較容易致使 PM2.5 濃度上升。與《上海地區風趨雨氣候特徵分析》提及風趨雨現象相同，在 2013-2015 年降雨或高風速通常會伴隨著 PM2.5 濃度的降低，若兩者加乘則 PM2.5 濃度的降低會更為明顯。綜上，主要影響上海空氣質量的氣候變數為風速與降雨(詳見圖六)。

1.4 成都市 PM2.5 濃度分析

成都地處四川盆地西部，且屬亞熱帶濕潤季風氣候，有秦嶺與大巴山形成天然屏障，阻擋了冬季來自北方的冷空氣，使得冬天的成都相較於其他縣市仍是比較溫暖的，但夏季亦有熱氣不易擴散的悶熱問題。人口密度高、民眾車輛持有量高、高度工業化，造成生活與工業汙染物排放量大；進而使得擁有豐富自然、礦產資源與高森林覆蓋率的成都，其空氣品質長期仍處於較差的狀態。

從 PM2.5 濃度均值與時間變數的關係，可以發現 2013-2015 年 PM2.5 的平均濃度有逐年下降的趨勢，其變異程度也隨之變小，且與工業及生活汙染物排放量的下降趨勢一致。以季節為尺度可以發現具有週期性的變化，秋冬季的明顯高於夏季；但以日為時間尺度則 PM2.5 並沒有明顯的變化。

從 PM2.5 濃度均值與氣候變數的關係(詳見圖七)，發現不論是濕度、溫

度、露點、雨量、風向及風速對 PM2.5 的影響皆不顯著。潮濕悶熱的盆地地形、高人口密度與工業化程度、長期吹靜風、秋冬季降雨量低與太陽輻射弱等因素，使得污染物無法順利擴散。但自 2013 年起，成都實施《大氣污染防治行動計畫》後，透過減少人為因素之排放，使得空氣品質得以有逐年顯著的改善。

1.5 廣州市 PM2.5 濃度分析

廣州市位於中國東南的沿海城市，地勢東北高、西北低，背山面海，導致天氣受大陸和海洋的影響明顯，平均溫度高且多降雨。廣州市天氣污染狀況整體較輕微，2013 年至 2015 年變化并沒有太大變化，PM2.5 濃度大部分時間都落於 100 以下。

從 PM2.5 濃度均值與時間變數的關係，可以發現污染具有明顯的季節性，5 月-10 月污染程度較輕，11 月-4 月污染程度較重。造成這種現象的主要原因是夏季地表溫度較高，空氣對流較快，不易形成逆溫現象，使得 PM2.5 容易擴散；次要原因則是夏季降水較多，暴雨對於 PM2.5 的清除具有一定的沖刷作用。

從 PM2.5 濃度均值與氣候變數的關係(詳見圖八)，發現露點溫度、壓力等因素影響並不明顯。PM2.5 與濕度的正相關趨勢較爲顯著；受到地理位置的影響，來自海洋的溫暖潮濕的風會攜帶大量水氣，濕度過高時，大量的水氣反而會吸附空氣中的污染物，不斷加重 PM2.5 濃度。廣州的風向較爲紊亂，累積風速、實際風速較低，吹拂效果不明顯，使得具有高降水量的廣州在 PM2.5 濃度不具有風趨雨現象。綜上，主要影響廣州空氣質量的氣候變數為溫度、濕度、降雨。

二、 模型建立

2.1 前情提要

由資料可以發現 PM2.5 濃度以小時為單位紀錄，因此在建模階段我們選擇使用時間序列模型 ARIMA 作為模型預測(流程詳見附件六)，同時我們提供 KNN 模型作為比較模型，利用 2013-2014 年的資料預測 2015 年 PM2.5 日平均濃度，並計算 MAE 以評估兩模型在各城市的預測能力。

2.2 資料整理

根據前述資料分析發現 PM2.5 濃度一日之內的變化在五個城市皆無顯著差異，因此我們將一日內的資料取平均以合併小時資料至每日資料。並且針對遺失值我們採取 KNN 插值法進行補值，避免直接刪除遺失值造成時間不連續等問題。因遺失值占比並不高，且 KNN 補值法所補之值皆非邊界之值，所以不予探討此補值法適切程度。

2.3 預測結果

	MAE		差值 (KNN-ARIMA)
城市	KNN 模型	ARIMA 模型	d
廣州	13.48	7.51	5.97
上海	24.36	19.45	4.91
成都	34.45	17.35	17.1
瀋陽	35.69	29.18	6.51
北京	54.28	44.21	10.08

2.4 結果討論

就各城市預測能力而言，若以 KNN 為預測模型，則廣州>上海>成都>瀋陽>北京；若以 ARIMA 為預測模型，則廣州>成都>上海>瀋陽>北京。共同趨勢為南方城市 PM2.5 預測能力好於北方城市，可能原因在於南方空氣品質優於北方城市，造成 PM2.5 的波動性有落差，通常北方城市有長達 6 個月之供暖期，亦多為重工業基地，造成廢氣排放量大，因此 PM2.5 的穩定性較差。

根據兩模型的預測能力差距可以發現，ARIMA 的預測能力優於 KNN，推測可能的原因在於使用時間序列模型而非利用變數間的相關性可能較適合該筆資料。

三、 結論

各個城市依據政策規劃而有不同的產業分布，不同的地理位置及周邊環境都使得各城市間有許多異同。在與時間變數的關係中，五個城市有較類似的結果；除了明顯的季節週期性外，自 2013 至 2015 年均有逐年下降的趨勢，並且一日內的變化幅度不大。在與氣候變數的關係中，則有較多的相異。降雨在上海及廣州市有顯著的反向關係，風速與風向在瀋陽、北京、上海與 PM2.5 濃度均值有較明顯的變動關係，濕度在廣州有較明顯的正向關係。壓力則是在五個城市皆沒有明確的與 PM2.5 的關係。瀋陽與北京為五座城市中唯二有供暖期的城市，其 PM2.5 平均濃度也是城市中較高者；成都則因地形、氣候、地理位置等多重因素導致 PM2.5 濃度長年較高的現象。上海與廣州則因臨海 PM2.5 濃度可受到海洋的調節，加之地處南方具豐沛的降雨而有總體 PM2.5 濃度較低的現象。

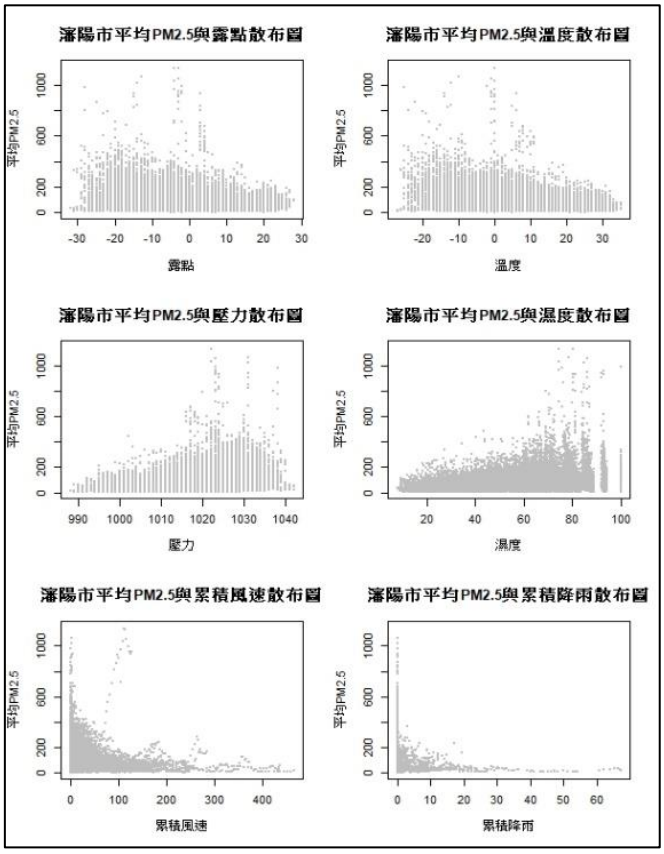
PM2.5 成因複雜性高，直觀上 PM2.5 平均濃度與氣候變數等應有所關聯，但在各城市中皆可發現並非所有氣候變數皆與 PM2.5 有顯著的關聯；可能進而導致考量氣候變數的 KNN 模型其預測能力不如考慮時間相關性的 ARIMA 模型。從探索性資料分析與建模預測結果皆顯示僅考量氣候變數分析 PM2.5 資料之不足，因此我們認為分析 PM2.5 資料除了從氣候變數等自然因素著手，更應搭配人口、產業、政策等人為因素，方能使得分析更臻完善。

四、 參考資料

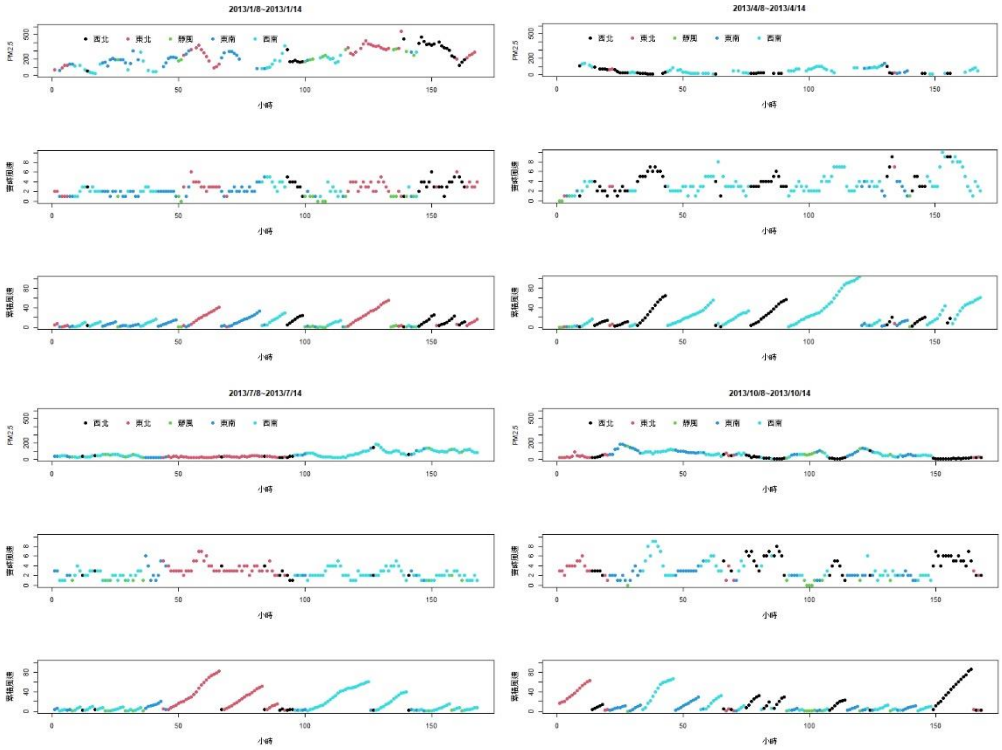
1. Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang and Song Xi Chen (2021).《Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating》. *Proc. R. Soc. A 471: 20150257.*
2. 袁楊森，劉大猛，車瑞俊，董雪玲 (2007).《北京市秋季大氣顆粒物的污染特徵研究》. *Ecology and Environment, 16(1)：18-25.*
3. 朱倩茹，劉永紅，徐偉嘉，黃敏 (2013)，《廣州 PM2.5 汙染特徵及影響因素分析》，*中國環境監測.，2013 年第 02 期。*
4. 每日頭條(2018-10-15)，同處東北，為啥長春空氣比瀋陽好，https://reurl.cc/V5G06R
5. 解放日報(2015-01-18)，PM2.5"基因譜"解讀：北京上海污染源差異大，http://env.people.com.cn/BIG5/n/2015/0118/c1010-26404420.html
6. 中外對話(2019-12-30)，先抑後揚：中國環保十年歷程回首，https://reurl.cc/4346RY
7. 穆海振，《上海地區風趨雨氣候特徵分析》，*西安建築科技大學學報(自然科學版)，2019 年第 01 期。*

五、 參考圖片

5.1 瀋陽市參考圖片

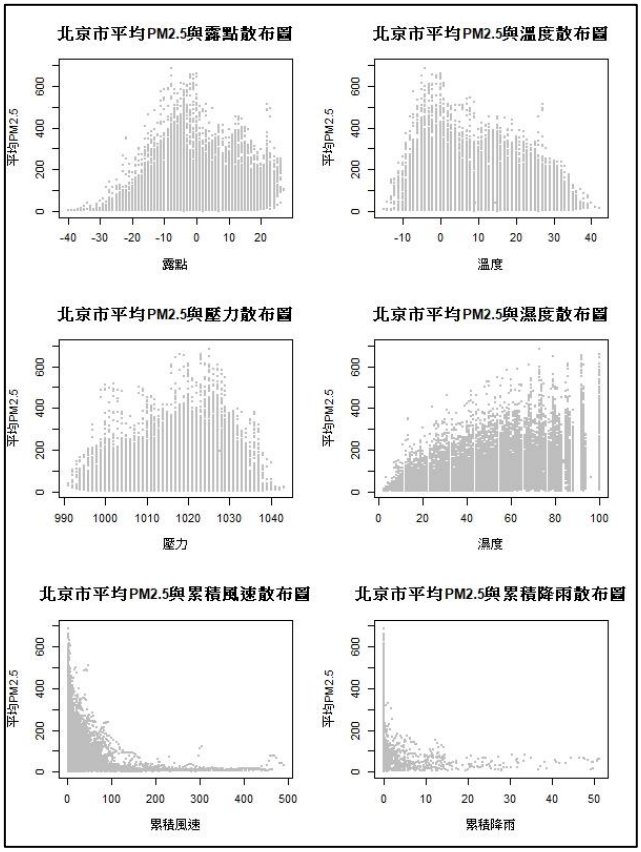


(圖一)

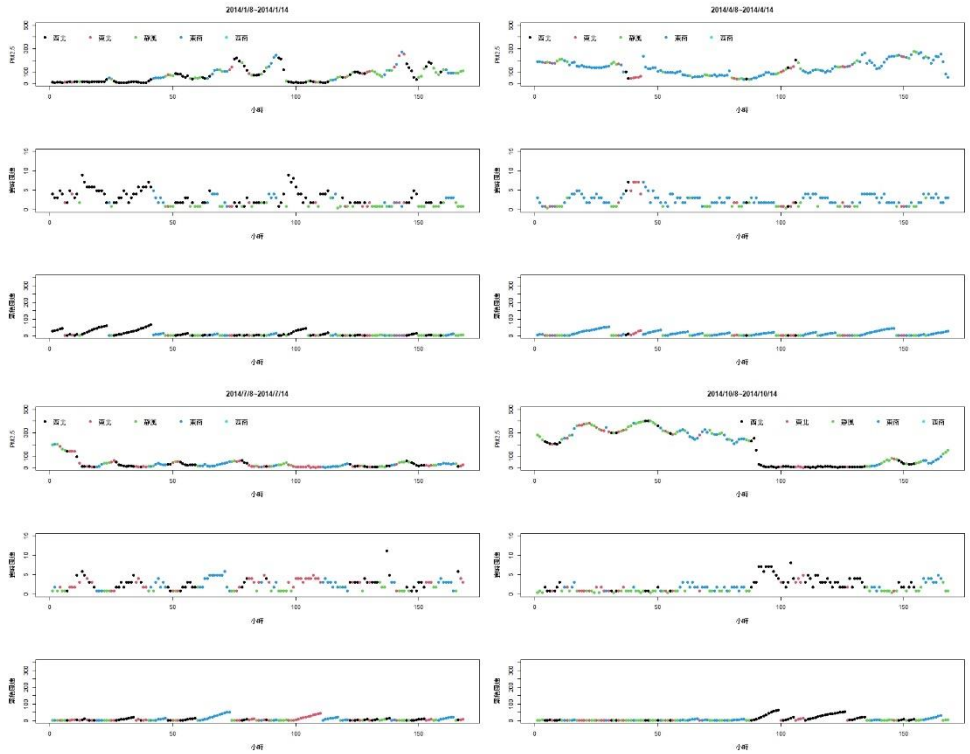


(圖二)

5.2 北京市參考圖片

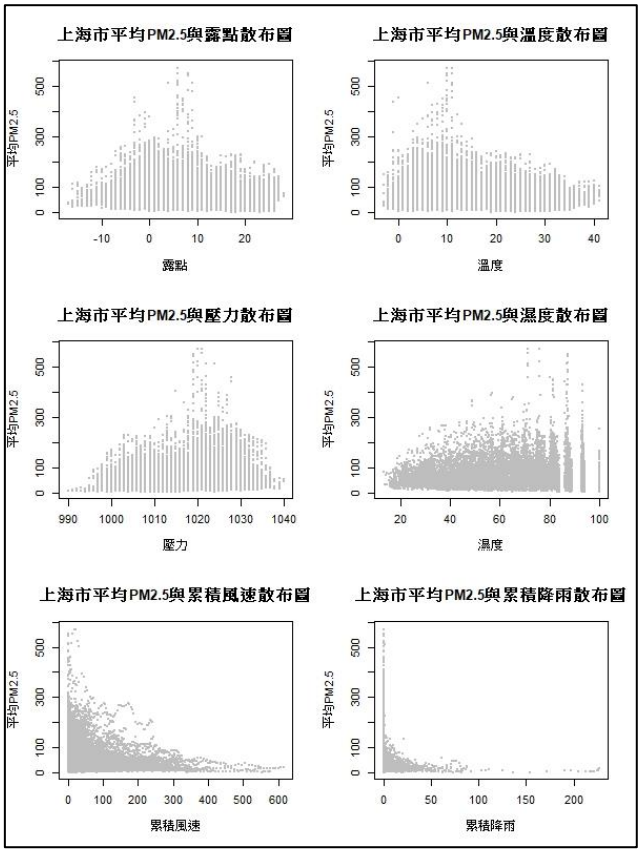


(圖三)

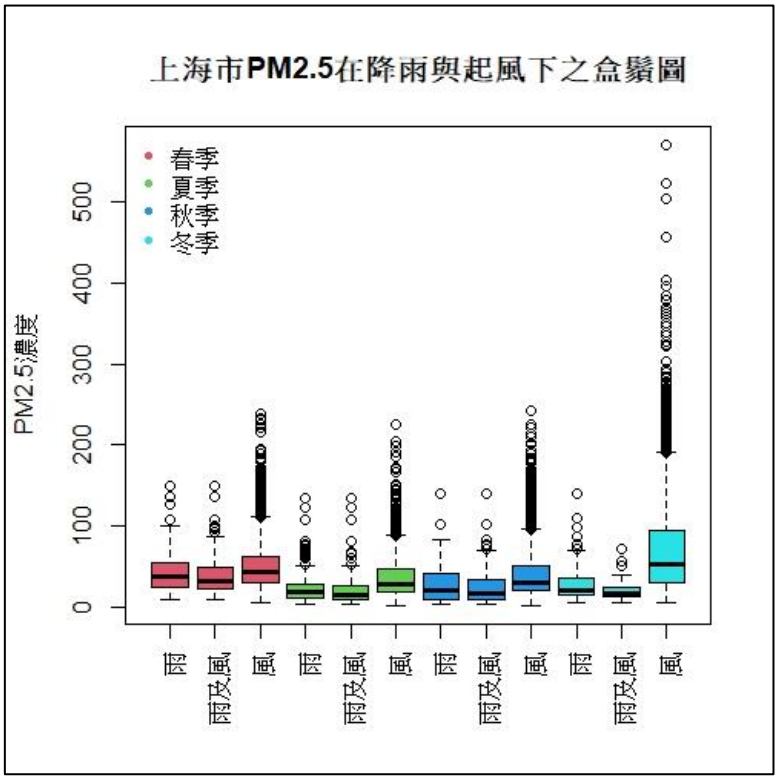


(圖四)

5.3 上海市參考圖片

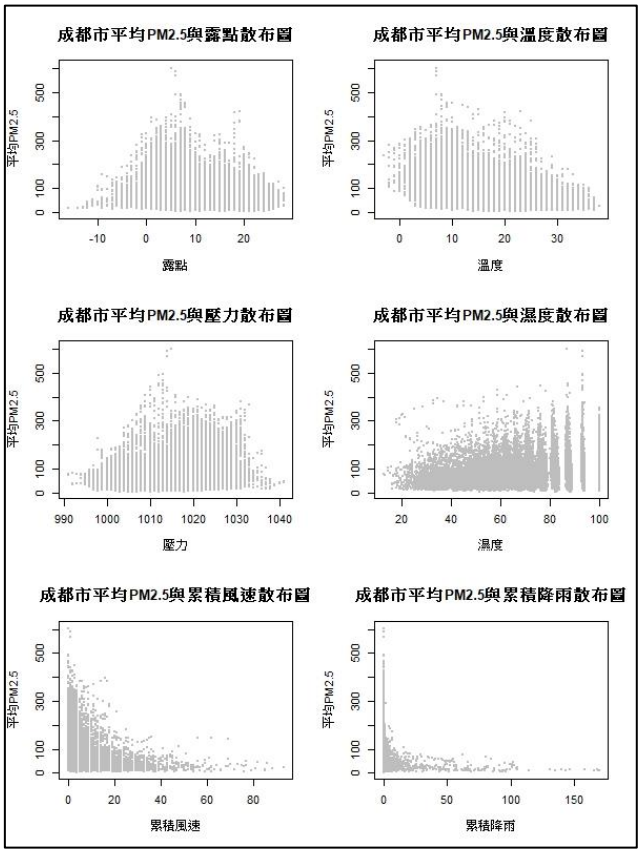


(圖五)



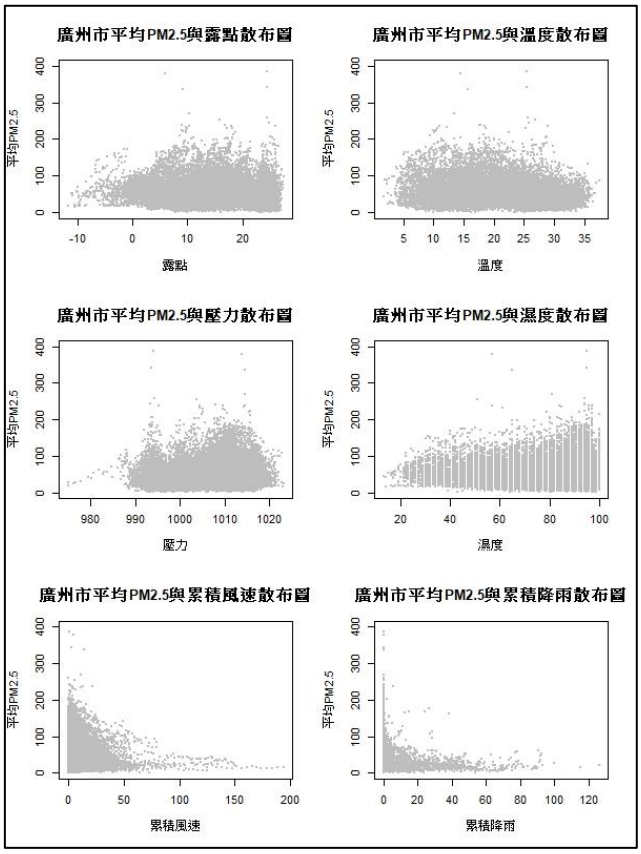
(圖六)

5.4 成都市參考圖片



(圖七)

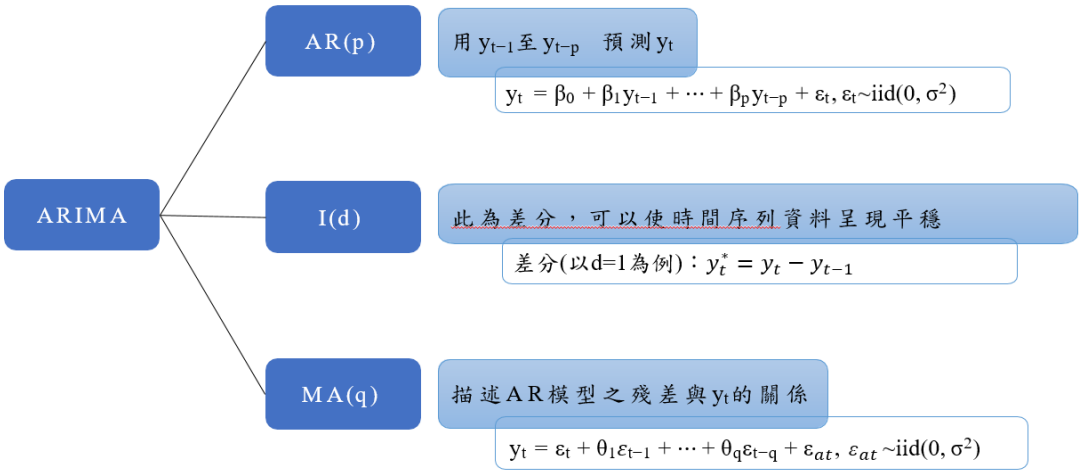
5.5 廣州市參考圖片



(圖八)

六、 ARIMA 模型假設與建模流程

(一)、模型假設



1. 模型參數定義

   p ： p 階 AR 模型，表示使用前 p 筆的資料建立自迴歸模型。

   d ：使時間序列平穩所需差分次數。

   q ： q 階 MA 模型，表示使用前 q 筆的 AR 模型的殘差建立 MA 模型。

2. 名詞解釋

   - Dies down(拖尾)：隨著時差之增加而漸漸消失之趨勢
   - Cut off(截尾)：在 lag 為 q 之後切斷之趨勢（即當所有 lag 大於 q 時）
   - 平穩： 資料不會隨觀測時間變化而變化，序列趨近水平帶狀並保持固定的方差
   - d(差分)：計算相鄰觀測值的差值，並以此差值形成新的時間序列資料，目的是使得資料不會產生特定趨勢，資料平均數呈現水平帶狀(平穩)
   - ACF(自相關函數)：反映同一份資料在不同時刻下，$Y_{t-1}, \ldots, Y_{t+k-1}$之間的相關程度
   - PACF(偏自相關函數)：移除$Y_{t-1}, \ldots, Y_{t+k-1}$的線性相關下，兩觀測值的線性相關程度

(二)、建模流程

3. 以 2013-2014 年的資料作為訓練集，畫出原始資料每日 PM2.5 平均濃度的折線圖、ACF 與 PACF，觀察資料的變異數與平均數是否平穩。
4. 若變異數與平均數不平穩：

   (1) 變異數不平穩的處理方式：將每日 PM2.5 平均濃度經過函數轉換後，再以折線圖呈現，觀察轉換後資料的變異數是否平穩，轉換方式如下所列：

$$y^* = \log(y)$$

$$y^* = \sqrt[2]{y}$$

$$y^* = \sqrt[4]{y}$$

接續，根據上述折線圖選出變異數最平穩的轉換方式後再處理平均數不平穩的問題。

(2) 平均數不平穩的處理方式：對轉換後的資料進行差分，再畫出折線圖、ACF(Auto Correlation Function,自相關函數)與 PACF(Partial Autocorrelation Function,偏自我相關函數)觀察資料變異數與平均數是否平穩。

5. 資料變異數與平均數呈現平穩的狀態後，檢定資料是否具有 ARCH Effect，若有 ARCH Effect 則表示變異數非同質，需使用 GARCH 模型，若無 ARCH Effect 則使用 ARIMA 模型。此處使用 Engle's ARCH Test，關於此檢定假設如下：

H_0：$\rho_1 = \rho_1 = \cdots = \rho_m$ (absence of ARCH Effect)
H_1：$\exists \rho_k \neq 0$
其中，m = maximum number of lags included in the test；ρ_i= the population auto correlation function of the squared time series (y_t), $1 \leq k \leq m$

若拒絕虛無假設，意謂資料存在 ARCH Effect，應選用 ARCH Model；若不拒絕虛無假設，則意謂資料不存在 ARCH Effect，可使用其他模型，此次我們選擇以 ARIMA 做為其他模型。

6. 使用 ARIMA(p,d,q) 模型的步驟：
(1) 根據資料的 PACF 與 ACF 決定使用 ARIMA、AR 或是 MA。
(2) 若 PACF 呈現 dies down 的狀態，則使用 MA 模型，並以 ACF 決定 MA 模型的參數 q；若 ACF 呈現 dies down 的狀態，則使用 AR 模型，並以 PACF 決定 AR 模型的參數 p。
(3) 若 PACF 與 ACF 相仿，則 p=1,q=1，使用 ARIMA(1,1)模型。
(4) 執行步驟(2)、(3)的資料若有進行差分，則需使用 ARIMA 模型，其中 p 根據 PACF 決定，q 根據 ACF 決定，d 為資料差分次數。
(5) 檢定配適完 ARIMA 模型後的殘差變異數是否同質且獨立，其中，藉由 Engle's ARCH Test 進行變異數同質檢定：

檢定統計量：$\frac{1}{n}\sum(\varepsilon_i - \bar{\varepsilon_i})^2$ *F* statistic for the regression on the squared residuals.

另外，藉由 Ljung-Box test 進行變異數獨立性檢定， Ljung-Box test 用於檢驗某個時間段內的一序列觀測值是否為隨機獨立的觀測值，於時間序列資料可檢驗其殘差是否存在滯後相關，其檢定假設如下：

H_0：$\beta_1 = \beta_1 = \cdots = \beta_m$ (殘差之間獨立)

H_1：$\exists\beta_k \neq 0$

其中，m = maximum number of lags included in the test；β_i= the correlations in the population from which the time series (y_t), $1 \leq k \leq m$

應用於 ARIMA 模型其統計量計算方式如下：

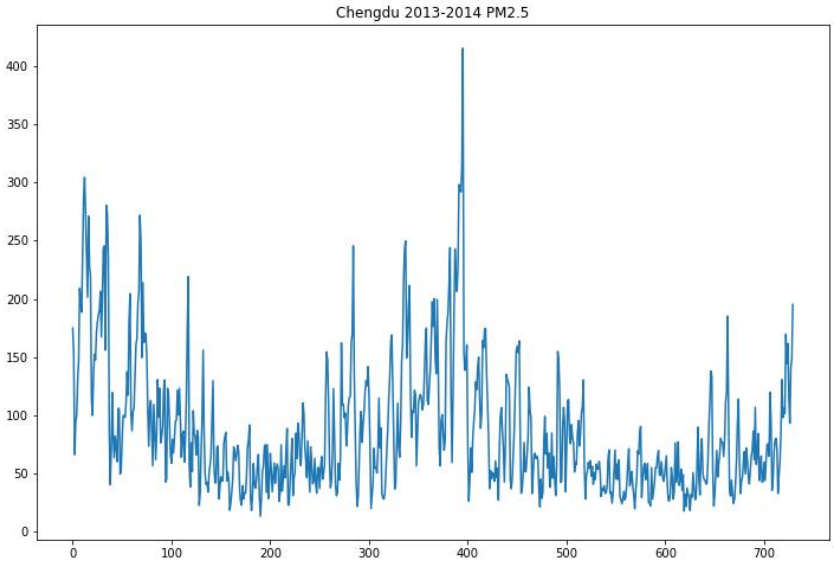
$$Q(m) = T(T+2) \sum_{i=1}^{m} \frac{\widehat{\rho_i}(a_t{}^2)}{T-i}$$

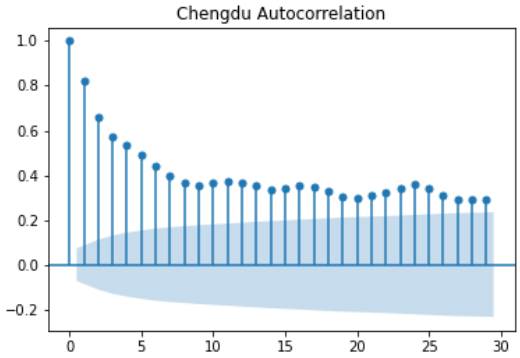
其中 T 是樣本大小，m 是人為選定的，a_t是殘差，而$\widehat{\rho_i}(a_t{}^2)$是$a_t{}^2$的 i 階自相關函數 (ACF)。

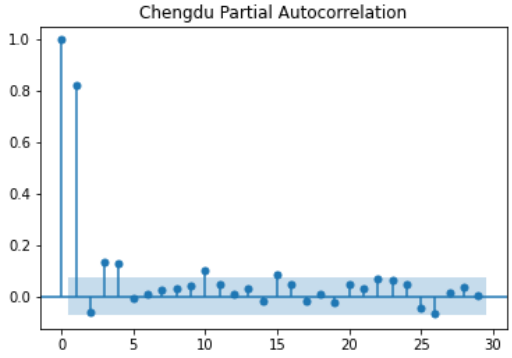
(6) 運用 one step 的方式計算預測值，並計算 MAE (Mean Absolut Error)。

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}$$

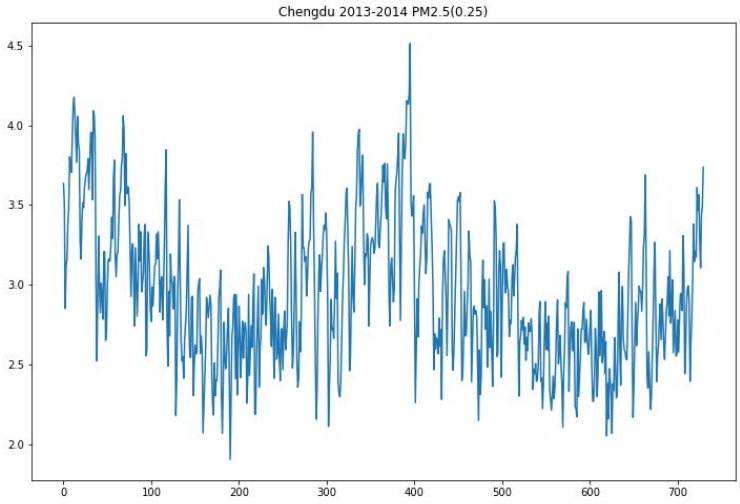
7. 以下以成都為例詳述建模流程，其餘城市依照建模流程進行建模與評估。

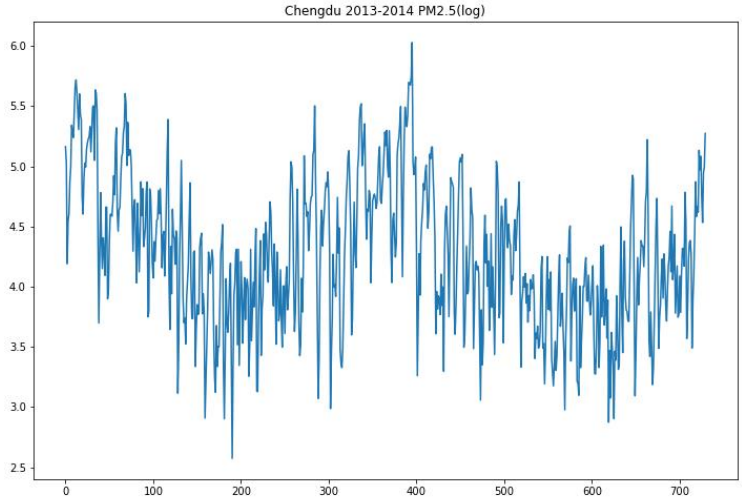


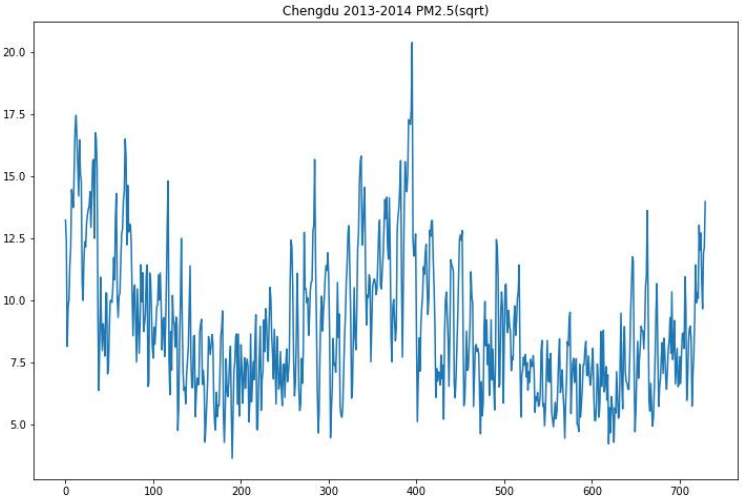




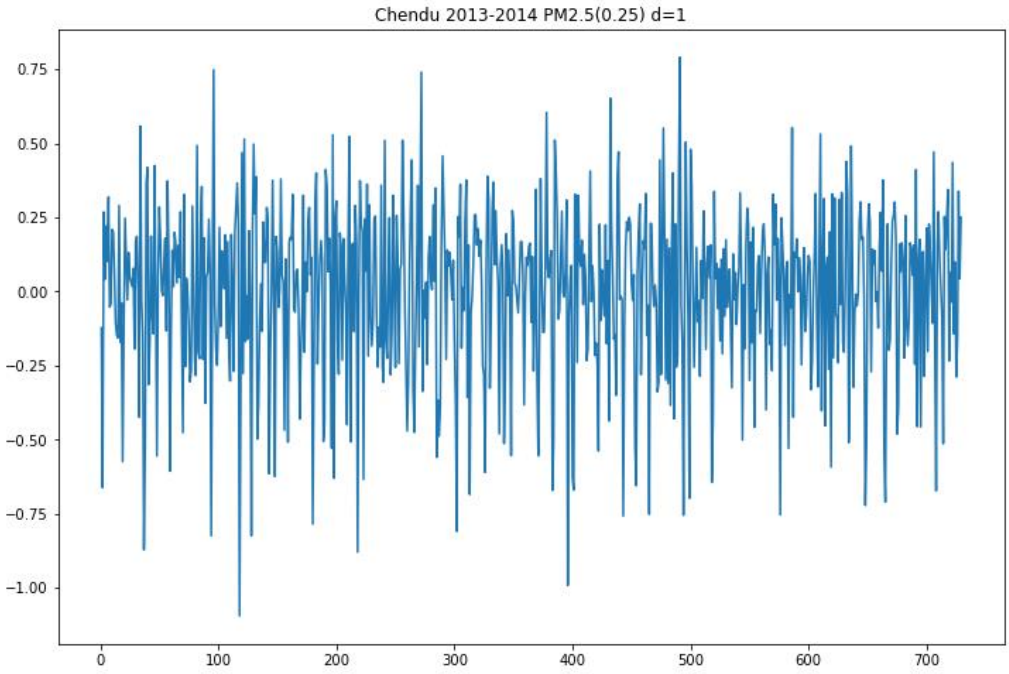
將成都 2013-2014 年的資料作為訓練集，畫出日均 PM2.5 濃度的折線圖、ACF 與 PACF，從折線圖來看，資料有明顯的高峰且具有大幅震盪，呈現變異數不平穩的狀況，且從 ACF 可以看出時間序列的自相關係數下降十分緩慢，呈現不平穩的狀態，再者，從折線圖的波動幅度不一致的狀況來看，平均數也非平穩的狀態。針對變異數不平穩的問題，我們先將 PM2.5 分別取 log、開根號、開四次根號，以觀察轉換後的 PM2.5 變異數是否能趨於平穩。

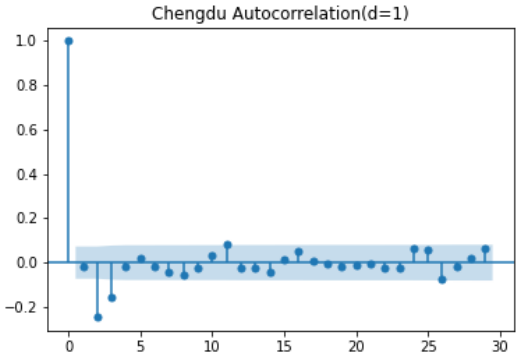


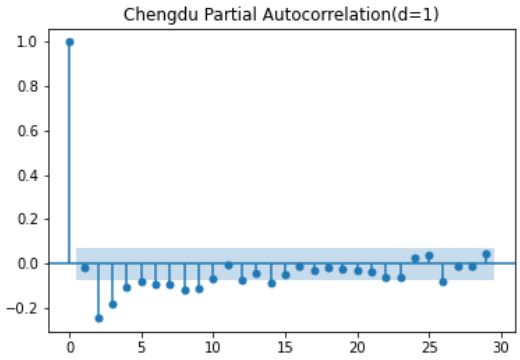




因為 PM2.5 開四次根號的的震盪幅度比取 log 再小一點，所以選用開四次根號後的 PM2.5 作為建模的變數。由於轉換後的 PM2.5 仍有平均數不平穩的現象，因此先做一次差分後畫出折線圖、ACF 與 PACF 觀察資料經轉換後又差分的變異數與平均數是否呈現平穩的狀態。







從折線圖來看，平均數與變異數皆呈現平穩的狀態，且從 ACF 可以看出時間序列的自相關係數快速的下降，呈現平穩的狀態，因此，在變異數與平均數平穩的狀態下，再使用開四次根號後的 PM2.5 進行 ARCH Effect Test，以決定應使用 ARCH 模型或 ARIMA 模型，由於檢定開四次根號後的 PM2.5 其 ARCH

Effect Test 的 p-value 為 0.558，因此傾向不拒絕虛無假設，故選用 ARIMA 模型以配適資料，並根據轉換後的 ACF 與 PACF 決定參數，因此，p、d、q 分別為 1、1、2。

ARIMA Model Results

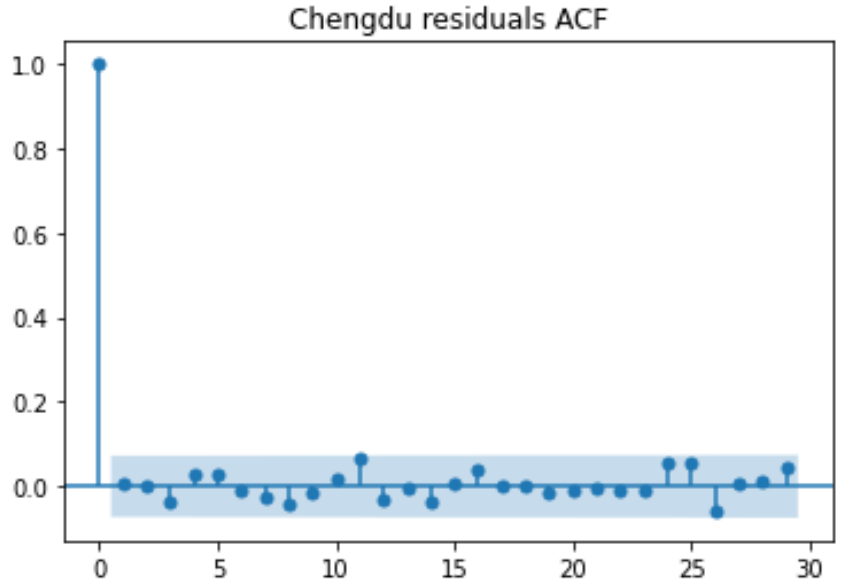
Dep. Variable:	D.PM_mean	No. Observations:	729
Model:	ARIMA(1, 1, 2)	Log Likelihood	-53.369
Method:	css-mle	S.D. of innovations	0.260
Date:	Wed, 29 Dec 2021	AIC	116.739
Time:	17:29:45	BIC	139.697
Sample:	1	HQIC	125.597

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0006	0.001	-0.375	0.708	-0.003	0.002
ar.L1.D.PM_mean	0.4818	0.055	8.798	0.000	0.375	0.589
ma.L1.D.PM_mean	-0.6519	0.057	-11.531	0.000	-0.763	-0.541
ma.L2.D.PM_mean	-0.2704	0.047	-5.712	0.000	-0.363	-0.178

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	2.0753	+0.0000j	2.0753	0.0000
MA.1	1.0642	+0.0000j	1.0642	0.0000
MA.2	-3.4754	+0.0000j	3.4754	0.5000

因 ARCH Effect Test 的 p-value 為 0.447，因此不拒絕虛無假設，即推估為同質變異數，又根據殘差的 ACF 與 Ljung-Box Test Lag30 的殘差之 p-value 為 0.92 的結果，因此不拒絕虛無假設，即推估殘差之間獨立，故符合 ARIMA 模型假設。



將結果以表格如下呈現：

	北京	成都	廣州	上海	瀋陽
轉換方式	$y^* = \log(y)$	$y^* = \sqrt[4]{y}$	$y^* = \log(y)$	$y^* = \sqrt[4]{y}$	$y^* = \sqrt[4]{y}$
ARIMA(p,d,q)	(1,0,1)	(1,1,2)	(1,1,2)	(0,1,2)	(0,1,2)
MAE	44.209	17.350	7.511	19.446	29.178