

國立成功大學統計學系
科技部 大專學生研究計畫

象徵性資料的時間序列應用
以南臺灣PM_{2.5}濃度為例

Symbolic Interval Time Series Models in Application to the
Level of PM_{2.5} on Southern Taiwan

指導教授：林良靖 博士

執行計畫學生：宋豪

中華民國一百一十年三月

摘要

隨著台灣經濟發展，社會環保意識抬頭，PM_{2.5}儼然成為當今政府與專家們傾力研究的領域。因觀測濃度存在許多遺失值，傳統PM_{2.5}的統計分析，通常以天或週為單位，取平均值後，進行時間序列的建模，但此作法將損失資料的變異性。本研究採用象徵性資料分析的概念，保留區間內部變異，將南台灣每週的PM_{2.5}觀測值整理成區間數值的形式，配適自區間式迴歸模型與異質變異數自區間式迴歸模型，並利用區間變數特有的蒲公英圖，展現不同地區之間，污染物質的區間相關程度。分析結果顯示，中點相關性與全距相關性，深受地理因素影響；污染物質大約以高屏溪為界，分為兩個相異的流通體系，且各自擁有不同的變異情況，這是文獻使用向量自迴歸模型所看不到的現象。

關鍵字：自區間式迴歸模型；象徵性資料分析；蒲公英圖。

1 緒論

近幾年來，空氣汙染的防治，早已成為國際社群的關注要點，其中又以隱形的殺手—細懸浮微粒(PM_{2.5})濃度，最被日常百姓所關心。細懸浮微粒，是指瀰漫在空氣當中的粒狀污染物，粒徑在2.5微米以下，僅有頭髮直徑的1/28。由於極為細小的緣故，PM_{2.5}具有強大的穿透力，人體的鼻毛、黏膜、支氣管纖毛皆無法捕捉；因此，附帶著戴奧辛、以及重金屬等有毒物質，PM_{2.5}可輕鬆地穿透肺部氣泡，進入血管隨著血液循環遍佈全身。面對如此嚴重的威脅，國內外學者紛紛投入PM_{2.5}的分析當中，如Yao et al. (2012)利用陸面測站所蒐集到的地理與大氣資料，對中國北方的PM_{2.5}建立預測模型；Cesari et al. (2018)分解懸浮微粒的化學構成，探討義大利南部地區，空氣品質的影響因素；Emami et al. (2018)針對美國紐約地區的PM_{2.5}以及其組成成分，進行多變量的長時間分析。

自1982年起，迄今39年間，我國環保署保存了相當豐富的空氣品質監測數據，涵蓋每日每小時，各種空氣污染物的時間序列資料，蒐集頻率非常的密集，但由於觀測機台維修或故障的緣故，時常發生缺失與無效的狀況，傳統的解決方法，例如Gao and Tsay (2019)，以天或週為單位，取平均值後，再進行時間序列的建模。然而，此作法會損失區間的某些資訊，例如「區間內變異性」等。另一方面，考慮使用象徵性資料分析(Symbolic Data Analysis, SDA)的

概念，取出時段內最高與最低濃度，作為區間資料的上下界，並配適區間式時間序列模型，我們便可從資料內部，提取更多的訊息，優化模型對PM_{2.5}的解釋能力，進而揭露測站之間的真實相關性。

象徵性資料的重要價值，在於合併大量資料的過程中，盡可能地保存原始資料所提供的訊息。Billard and Diday (2003)首先提出SDA的基礎概念，將變數型態用列表(List)、區間(Interval)與分布(distribution)等樣態，重新進行詮釋，並將變數區分為多值變數(multivalued variable)、區間值變數(interval-valued variables)、模態多值變數(modal multivalued variables)與模態區間值變數(modal interval-valued variables)。Brito (2014)從資料探勘的角度，彰顯SDA在巨量資料分析的重要性，並統整近年來SDA所發展出的統計推論方法。在SDA眾多的應用中，實用性最大的，莫過於區間值變數，許多研究也被相繼提出，例如Arroyo et al. (2010)將區間值的時間序列重新整理成區間中心與區間半徑的變數，再配適向量自迴歸模型(vector autoregressive model, VAR model)；Lin and González-Rivera (2016)在常態分佈的假設下，將區間值時間序列視為不同時間的最大順序統計量與最小順序統計量，並給予常態分佈的參數與區間個數有自迴歸效應，建立區間式時間序列的模型。近期的研究中，Lin et al. (2021)也以資料點服從常態分布為前提，將順序統計量與傳統的自迴歸模型(Autoregressive model)結合，提出自區間式迴歸模型(auto-interval-regressive model, AIR model)，並針對時間序列的異質變異性，進一步提出異質變異數的自區間式迴歸模型(heteroscedastic volatility auto-interval-regressive model, HVAIR model)；Zhang and Lin (2020)則在區間中點與區間全距，服從多維常態分佈的假設下，設計線段圖(segment plot)與蒲公英圖(dandelion plot)兩視覺化方法，同時呈現區間資料的相關性。

以象徵性資料分析的應用為核心，本研究使用Lin et al. (2021)提出的區間式時間序列模型，以及Zhang and Lin (2020)提出的視覺化方法，探討高屏空品區，2006年至2015年期間，15個測站PM_{2.5}的區間相關性。實際分析中，我們採用Cleveland et al. (1990)提出的分解方法，將區間濃度分解為趨勢項，季節項與剩餘項，並針對剩餘項區間值的極大值與極小值，配適AIR模型與HVAIR模型，進而計算模型產生的殘差區間，繪製蒲公英圖，推演不同測站的區間相關性。最終，我們將中點相關性與全距相關性以數值表示，並與傳統VAR模型產生的殘差變異係數矩陣相互比對，呈現不同分析手法的差異性。

在文章的編排上，本研究第二章將介紹濃度區間值的符號定義、時間序列分解法、VAR模型與優化流程、AIR與HVAIR模型，以及線段圖與蒲公英圖的繪製流程與相關性計算法。在第三章中，透過蒙地卡羅模擬，我們驗證區間式時間序列模型對參數的估計能力，以及不同步數下的預測誤差。在第四章中，我們應用區間式時間序列模型，分析南台灣各測站PM_{2.5}的區間相關性，同時與傳統分析結果相互比較。最後，在第五章我們針對分析結果進行討論，並給予未來研究一些建議與方向。

2 文獻回顧與探討

本章節將針對PM_{2.5}濃度的資料型態，進行符號定義，並詳述後續分析所運用的統計模型與方法，包含時間序列的分解方法、向量自迴歸模型、區間式自迴歸模型、以及區間式相關性分析。

2.1 符號定義

在環保署公開的PM_{2.5}觀測值中，主要以「小時」作為紀錄的時間單位。因此，統整各測站每小時的觀測值後，令 $X_{i,j}$ 表示第*i*測站第*j*筆觀測值，我們可將資料的架構以矩陣表示，定義如下：

$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p,1} & X_{p,2} & \cdots & X_{p,n} \end{pmatrix},$$

其中*p*表示測站總數、*n*表示各測站的資料總數。理想狀況下，我們可針對各測站每小時的觀測濃度，進行精確的時間序列分析。然而，由於測站故障與定期維修的緣故，每小時的資料時常出現異常與闕漏的情形，因此在分析之前，多數文獻選擇將一段時間內的資料合併，以Gao and Tsay (2019)為例，該研究以週為固定的時間區段，計算出平均數代表當週的整體趨勢。已知一週共有 $7 \times 24 = 168$ 筆觀測值，令*m*表示總週數，則第*t*週週平均濃度的符號定義如下：

$$y_{i,t} = \frac{\sum_{j=168(t-1)+1}^{168t} X_{i,j} \times I_{i,j}}{\sum_{j=168(t-1)+1}^{168t} I_{i,j}}, \quad i = 1, \dots, p, \quad t = 1, \dots, m, \quad (1)$$

且

$$I_{i,j} = \begin{cases} 1, & \text{if } X_{i,j} \text{ is a valid data point;} \\ 0, & \text{otherwise.} \end{cases}$$

另一方面，在象徵性資料的分析中，我們不再以單一的點作為分析主體，取而代之，提取每週日均濃度的最大值與最小值，構成週區間濃度的上下界；以區間值來涵蓋週內觀測值所蘊含的資訊。第*t*週日均濃度的集合，定義如下：

$$\mathbf{W}_{i,t} = \left\{ \mu_d \mid \mu_d = \frac{\sum_{j=168(t-1)+24(d-1)+1}^{168(t-1)+24d} X_{i,j} \times I_{i,j}}{\sum_{j=168(t-1)+24(d-1)+1}^{168(t-1)+24d} I_{i,j}}, \quad d = 1, \dots, 7. \right\},$$

其中

$$I_{i,j} = \begin{cases} 1, & \text{if } X_{i,j} \text{ is a valid data point;} \\ 0, & \text{otherwise.} \end{cases}$$

則第*t*週週區間濃度可被表示為：

$$\mathbf{y}_{i,t} = [\text{Max}(\mathbf{W}_{i,t}), \text{Min}(\mathbf{W}_{i,t})], \quad i = 1, \dots, p, \quad t = 1, \dots, m. \quad (2)$$

2.2 時間序列分解方法

Seasonal and Trend decomposition using Loess (STL)透過局部加權迴歸的應用，結合傳統迴歸的簡潔性和非線性迴歸的靈活性，達到穩健且快速的時序分解，由 Cleveland et al. (1990)提出。令 Y_t 表示單維度的時間序列，則STL將序列分成季節項(S_t)、趨勢項(T_t)與剩餘項(R_t)，如下所示：

$$Y_t = S_t + T_t + R_t. \quad (3)$$

做為一個疊代的分解方法，STL由兩個環節所構成：「內循環」與「外循環」，每當內循環疊代一次，季節項與趨勢項即更新一次。令 $S_t^{(k)}$ 與 $T_t^{(k)}$ 表示第*k*次疊代所產生的季節項與趨勢項，則第*k*+1次的季節項 $S_t^{(k+1)}$ 與趨勢項 $T_t^{(k+1)}$ ，可藉由下列流程取得：

步驟1：去除趨勢。將原序列扣除估計的趨勢項 $T_t^{(k)}$: $X_t^{\text{detrend}} = X_t - T_t^{(k)}$ 。

步驟2：週期性子序列的平滑化(smoothing)。在序列 X_t^{detrend} 中，每個週期性子序列皆經由局部加權迴歸進行平滑，獲得初步的季節項 $C_t^{(k+1)}$ 。

步驟3：低通濾波器(Low-pass Filter)。針對 $C_t^{(k+1)}$ 再次使用低通濾波器，與局部加權迴歸進行平滑，獲得剩餘的序列 $L_t^{(k+1)}$ 。

步驟4: 週期性子序列的趨勢去除。將初步季節項 $C_t^{(k+1)}$ 減去**步驟3**所獲得的剩餘序列，防止低頻信息進入季節項，藉此獲得第 $k + 1$ 次的季節項：
 $S_t^{(k+1)} = C_t^{(k+1)} - L_t^{(k+1)}$ 。

步驟5: 去除季節項。將原序列扣除估計的季節項 $S_t^{(k+1)}$ ：
 $X_t^{deseason} = X_t - S_t^{(k+1)}$ 。

步驟6: 趨勢平滑化。針對**步驟5**所獲得的序列 $X_t^{deseason}$ ，進行局部加權迴歸的平滑，獲得第 $k + 1$ 次的趨勢項 $T_t^{(k+1)}$ 。

當內循環完成後，外循環便接續啓動。外循環利用內循環估計的季節項與趨勢項，計算出剩餘項的大小：

$$R_t^{(k+1)} = X_t - T_t^{(k+1)} - S_t^{(k+1)}, \quad (4)$$

同時運用該項數值，計算權重係數。當進行下一階段的疊代時，這些權重係數將用於離群值的判斷，以降低離群值對內循環的影響程度。最終，透過不斷疊代，以及內外循環的相互配合，我們完成時間序列的分解，將序列拆解為趨勢項，季節項與剩餘項。

2.3 向量自迴歸模型

2.3.1 模型架構

Sims (1980)提出向量自我迴歸模型 (vector auto-regressive model, VAR model)，捨棄先驗理論的必要性，以一組迴歸方程式探討變數間的交互效果。令 $\mathbf{Y}_t = (y_{1t}, \dots, y_{kt})^\top$ 表示 t 時間點下， k 維的時間序列，則VAR(p)模型定義如下：

$$\mathbf{Y}_t = \phi_0 + \phi_1 \mathbf{Y}_{t-1} + \dots + \phi_p \mathbf{Y}_{t-p} + \mathbf{a}_t, \quad p > 0, \quad (5)$$

其中 ϕ_0 為 $(k \times 1)$ 的常數向量、 $\phi_i, i = 1, \dots, p$ 為 $(k \times k)$ 的係數矩陣、 \mathbf{a}_t 為 $(k \times 1)$ 的誤差向量，服從多維常態分配 $\mathcal{MN}(0, \Sigma)$ ，且 Σ 為正定矩陣。倘若使用後移算子 (back-shift operator) B ，經簡單的移項，定義式(5)可重新表示如下：

$$(\mathbf{I} - \phi_1 B - \dots - \phi_p B^p) \mathbf{Y}_t = \phi_0 + \mathbf{a}_t, \quad (6)$$

其中 \mathbf{I} 為 $(k \times k)$ 的單位矩陣。在時間序列 \mathbf{Y}_t 為弱平穩的條件下，運用普通最小平方法，我們可針對係數矩陣 $\phi_i, i = 1, \dots, p$ ，進行估計。

2.3.2 優化流程

Gao and Tsay (2019)融合結構模型 (structural model)與潛在因素 (latent factor)的概念，對於多變量時間序列的分析，提出一套系統化的建模流程。令 $\mathbf{y}_t = (y_{1t}, \dots, y_{kt})^\top$ 表示 t 時間點下， k 維的時間序列，該研究將時間序列拆解為三個要素，定義如下：

$$\mathbf{y}_t = \boldsymbol{\mu}_t + \mathbf{s}_t + \boldsymbol{\eta}_t, \quad (7)$$

其中 $\boldsymbol{\mu}_t = (\mu_{1t}, \dots, \mu_{kt})^\top$, $\mathbf{s}_t = (s_{1t}, \dots, s_{kt})^\top$, $\boldsymbol{\eta}_t = (\eta_{1t}, \dots, \eta_{kt})^\top$ 分別表示時間序列的趨勢項、季節項與剩餘項。對於所有 $1 \leq i \leq k$ ，該研究使用 d_0 次多項式與 k_0 個三角級數的線性組合，各別估計趨勢項與季節項，如下所示：

$$\mu_{it} = \alpha_{i0} + \alpha_{i1}t + \dots + \alpha_{id_0}t^{d_0} \text{ and } s_{it} = \sum_{j=1}^{k_0} [\beta_{ij} \cos(\rho_j t) + \gamma_{ij} \sin(\rho_j t)], \quad (8)$$

其中 d_0 與 k_0 為非負整數， $\rho_j = 2\pi j/s$ ， s 為季節效應的週期。最後，針對弱平穩的剩餘項，該研究套用特殊的因素模型，將其重新表示成兩矩陣的內積：

$$\boldsymbol{\eta}_t = \tilde{\mathbf{L}} \begin{bmatrix} \mathbf{f}_t \\ \boldsymbol{\epsilon}_t \end{bmatrix} = \tilde{\mathbf{L}}_1 \mathbf{f}_t + \tilde{\mathbf{L}}_2 \boldsymbol{\epsilon}_t, \quad (9)$$

其中 $\tilde{\mathbf{L}} = (\tilde{\mathbf{L}}_1, \tilde{\mathbf{L}}_2)$ 為 $(k \times k)$ 的可逆矩陣， $\mathbf{f}_t = (f_{1t}, \dots, f_{rt})^\top$ 、 $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \dots, \epsilon_{vt})^\top$ ，且滿足 $r + v = k$ 。透過該模型的統計性質，我們可推斷 \mathbf{f}_t 是由 $\boldsymbol{\eta}_t$ 的共同因子 (common factor) 所構成，且 $\boldsymbol{\epsilon}_t$ 為時間序列的隨機誤差。因此，倘若針對 \mathbf{f}_t 配適 VAR 模型，便可降低模型所需的參數數量，同時提升模型的可辨識性 (identifiability)。

2.4 區間式時間序列模型

2.4.1 自區間式迴歸模型

Lin et al. (2021) 在常態分佈的假設下，從順序統計量的觀點，提出自區間式迴歸模型。令 $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t})^\top$ 表示雙變量的時間序列，則 AIR(p) 模型的關係式，定義如下：

$$\mathbf{Y}_t = \phi_1 \mathbf{Y}_{t-1} + \dots + \phi_p \mathbf{Y}_{t-p} + \mathbf{A}_t, \quad (10)$$

其中 $\phi_i, i = 1, \dots, p$ 為模型參數， $\mathbf{A}_t = (A_{u,t}, A_{l,t})^\top$ 為獨立同分布的白噪音誤差；上界 $A_{u,t} = \max\{A_{1,t}, \dots, A_{n,t}\}$ 與下界 $A_{l,t} = \min\{A_{1,t}, \dots, A_{n,t}\}$ ，皆由隨機變數 $A_{i,t}$ 所組成，且 $A_{i,t} \stackrel{iid}{\sim} N(0, \sigma^2)$, $i = 1, \dots, n$ 。

透過極大化對數概似函數(log-likelihood function)，我們同時估計關係式(10)中的自迴歸係數，以及 $A_{i,t}$ 的變異數大小。令觀測區間為 $\mathbf{X}_1, \dots, \mathbf{X}_m$ ，參數 $\boldsymbol{\theta} = (\phi_1, \dots, \phi_p, \sigma^2)$ ，AIR(p)的對數概似函數，定義如下：

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \text{constant} + (n-2) \sum_{t=p+1}^m \log \int_{(X_{l,t}-\phi_1 X_{l,t-1}-\cdots-\phi_p X_{l,t-p})}^{(X_{u,t}-\phi_1 X_{u,t-1}-\cdots-\phi_p X_{u,t-p})} f_\sigma(z) dz - \frac{(m-p)}{2} \log \sigma^2 \\ &\quad - \sum_{t=p+1}^m \frac{(X_{u,t} - \phi_1 X_{u,t-1} - \cdots - \phi_p X_{u,t-p})^2 + (X_{l,t} - \phi_1 X_{l,t-1} - \cdots - \phi_p X_{l,t-p})^2}{\sigma^2},\end{aligned}$$

其中 f_σ 表示 $N(0, \sigma^2)$ 的機率密度函數。

針對未來趨勢的預測，AIR(p)模型 h 步的預測式定義如下：

$$\hat{\mathbf{X}}_t(h) = E \left(\begin{array}{c|c} X_{u,t+h} & \\ \hline X_{l,t+h} & F_t \end{array} \right) = \phi_1 \hat{\mathbf{X}}_t(h-1) + \cdots + \phi_p \hat{\mathbf{X}}_t(h-p) + E \left(\begin{array}{c} A_{u,t+h} \\ A_{l,t+h} \end{array} \right), \quad (11)$$

其中 F_t 為 σ -域 (sigma field)，涵蓋 t 時間點之前的所有資訊。由於 $A_{u,t}$ 與 $A_{l,t}$ 為常態分佈下，極大值與極小值的順序統計量，兩隨機變數的期望值不為零，故該研究採用Blom (1958)提出的近似方法，估計兩變數的期望值，定義如下：

$$\begin{aligned}E(A_{u,t}) &\approx \Phi^{-1} \left(\frac{n-\alpha}{n-2\alpha+1} \right) \times \sigma, \\ E(A_{l,t}) &\approx \Phi^{-1} \left(\frac{1-\alpha}{n-2\alpha+1} \right) \times \sigma.\end{aligned} \quad (12)$$

其中 $\Phi(x)$ 表示 $N(0, 1)$ 的累積分布函數， $\alpha = 3/8$ (參考Blom, 1958)。將近似式(12)帶入預測式(11)，我們便可得到完整的預測公式。

2.4.2 異質變異數自區間式迴歸模型

在商業分析中，資料的變異數往往隨著時間而有所改變。為了詮釋時間序列的異質變異性，該研究重新設計變異數的架構，進一步提出HVAIR模型。HVAIR(p,q)模型的關係式，定義如下：

$$\mathbf{Y}_t = \phi_1 \mathbf{Y}_{t-1} + \cdots + \phi_p \mathbf{Y}_{t-p} + \mathbf{A}_t, \quad 1 \leq t \leq m, \quad (13)$$

其中 $\mathbf{A}_t = (A_{u,t}, A_{l,t})^\top$ ， $A_{u,t}$ 與 $A_{l,t}$ 亦由常態隨機變數 $A_{i,t}$ 的極大值與極小值所構成，但與AIR模型相異的地方，在於 $A_{i,t}$ 的變異數 σ_t^2 隨時間變化，滿足下列等式：

$$\sigma_t^2 = \beta_0 + \beta_1 \gamma_{t-1}^2 + \cdots + \beta_q \gamma_{t-q}^2 \quad (14)$$

且

$$\gamma_{t-j}^2 = \left(\frac{Y_{1,t-j} - Y_{2,t-j}}{\Phi^{-1} \left(\frac{n-\alpha}{n-2\alpha+1} \right) - \Phi^{-1} \left(\frac{1-\alpha}{n-2\alpha+1} \right)} \right)^2, \quad j = 1, \dots, q. \quad (15)$$

令觀測區間為 $\mathbf{X}_1, \dots, \mathbf{X}_m$ ，參數 $\boldsymbol{\theta} = (\phi_1, \dots, \phi_p, \beta_0, \beta_1, \dots, \beta_q)$ ，則 HVAIR(p, q) 模型的對數概似函數，定義如下：

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \text{constant} + (n-2) \sum_{t=p+1}^m \log \int_{(X_{l,t}-\phi_1 X_{l,t-1}-\cdots-\phi_p X_{l,t-p})}^{(X_{u,t}-\phi_1 X_{u,t-1}-\cdots-\phi_p X_{u,t-p})} f_{\sigma_t}(z) dz - \frac{1}{2} \sum_{t=p+1}^m \log \sigma_t^2 \\ &\quad - \sum_{t=p+1}^m \frac{(X_{u,t} - \phi_1 X_{u,t-1} - \cdots - \phi_p X_{u,t-p})^2 + (X_{l,t} - \phi_1 X_{l,t-1} - \cdots - \phi_p X_{l,t-p})^2}{\sigma_t^2},\end{aligned}$$

其中 f_{σ_t} 為 $N(0, \sigma_t^2)$ 的機率密度函數。

在預測方面，與 AIR 模型的結果相似，HVAIR 模型的預測關係式亦可用方程式(11)表達，但因為異質變異性的緣故，近似式(12)中的變異數，皆須用 σ_t 取代，如下所示：

$$\begin{aligned}E(A_{u,t}) &\approx \Phi^{-1}\left(\frac{n-\alpha}{n-2\alpha+1}\right) \times \sigma_t, \\ E(A_{l,t}) &\approx \Phi^{-1}\left(\frac{1-\alpha}{n-2\alpha+1}\right) \times \sigma_t.\end{aligned}\tag{16}$$

2.5 區間相關性分析

Zhang and Lin (2020) 提出線段圖(segment plot)與蒲公英圖(dandelion plot)，以同時呈現區間資料獨有的相關性。令 X 與 Y 為兩區間變數，則區間資料可以用下表呈現：

X	Y
$(x_{u,1}, x_{l,1})$	$(y_{u,1}, y_{l,1})$
⋮	⋮
$(x_{u,n}, x_{l,n})$	$(y_{u,n}, y_{l,n})$

其中 $x_i = (x_{u,i}, x_{l,i})$ 與 $y_i = (y_{u,i}, y_{l,i})$ 代表兩區間變數第 i 筆的觀測值。此外，令 $x_{ic} = \frac{x_{u,i} + x_{l,i}}{2}$ 、 $y_{ic} = \frac{y_{u,i} + y_{l,i}}{2}$ 為區間中點， $x_{ir} = x_{u,i} - x_{l,i}$ 、 $y_{ir} = y_{u,i} - y_{l,i}$ 為區間全距，則兩區間變數可再定義為 $X = (X_c, X_r)$ 以及 $Y = (Y_c, Y_r)$ 。透過探討兩區間變數間，中心與全距的關聯，我們能展現六種相異的區間相關性，如表1所示。

在兩區間變數的中心與全距，服從多維常態分佈的假設下，為了避免單位造成影響，該研究將 X_c, X_r, Y_c, Y_r 各別標準化，並分別以 V_c, V_r, W_c, W_r 代表標準化後的變數，進行後續的呈現。

表 1: 區間資料間六種區間相關性。

符號	種類	統計意義
ρ_{cc}	中點相關性	X_c 與 Y_c 之間的相關性
ρ_{rr}	全距相關性	X_r 與 Y_r 之間的相關性
ρ_{cr}	互相關性	X_c 與 Y_r 之間的相關性
ρ_{rc}	互相關性	X_r 與 Y_c 之間的相關性
ρ_{xx}	自相關性	X_c 與 X_r 之間的相關性
ρ_{yy}	自相關性	Y_c 與 Y_r 之間的相關性

在線段圖的繪製中，該研究以中心值為橫坐標，全距值為縱座標，將成對的 (V_c, V_r) 、 (W_c, W_r) 資料點繪於圖上，並以線段連接。圖上的一個點，即代表一筆區間資料。透過對線段圖的觀察，我們得以從點的散布狀況，推斷樣本自相關性的強弱，並發掘中心與全距間可能存在的關係。如圖 1(a) 中，區間變數 X 的中心與全距間，存在明顯的二次曲線關係。

而在蒲公英圖的繪製中，該研究將原先線段圖中的線段平移，使區間變數 W 的所有點座標 (W_c, W_r) ，皆移動至原點 $(0, 0)$ ；令 $d_c = V_c - W_c$ ， $d_r = V_r - W_r$ ，則該行為相當於將 (d_c, d_r) 資料點繪於圖上，並連結點座標與原點，完成的圖形即為圖 1(b)。

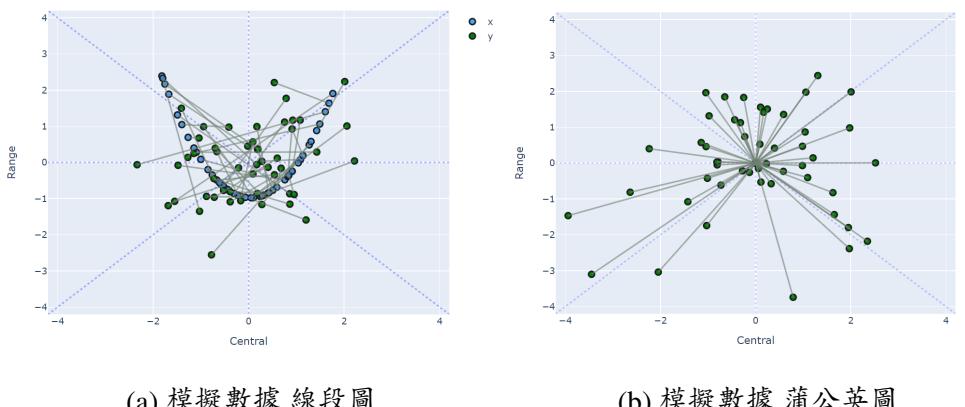


圖 1: 模擬區間資料的視覺化，其中 $X_c \sim N(0, 3)$, $X_r = X_c^2$, $Y_c \sim N(0, 3)$, $Y_r \sim N(3, 1)$ ，兩變數各生成 100 筆。

特別的是，在蒲公英圖中，計算每個線段在「水平方位」、「鉛直方位」、「對角線」與「非對角線」的平均投影長度，我們便可藉由其期望值與實際值，判斷中點相關性、全距相關性與兩互相關性的強弱，並據此繪製導讀多邊形(Guided polygon)，協助區間相關性的視覺化。

令 p 表示平均投影長度，則當 $|p - \frac{2}{\sqrt{\pi}}|$ 越大時，投影長所對應的區間相關性越為強烈，且 $p < \frac{2}{\sqrt{\pi}}$ 表正相關、 $p > \frac{2}{\sqrt{\pi}}$ 表負相關、 $p = \frac{2}{\sqrt{\pi}}$ 表相關性獨立。因此，令 p_h, p_v, p_d, p_o 分別代表水平、鉛直、對角線與非對角線的平均投影量，該研究在圖上以紅色線段連結 $(-p_h, 0), (-p_o, p_o), (0, p_v), (p_d, p_d), (p_h, 0), (p_o, -p_o), (0, -p_v), (-p_d, -p_d)$ 8個點，並以黑色線段連結 $(-\frac{2}{\sqrt{\pi}}, 0), (-\frac{2}{\sqrt{\pi}}, \frac{2}{\sqrt{\pi}}), (0, \frac{2}{\sqrt{\pi}}), (\frac{2}{\sqrt{\pi}}, \frac{2}{\sqrt{\pi}}), (\frac{2}{\sqrt{\pi}}, 0), (\frac{2}{\sqrt{\pi}}, -\frac{2}{\sqrt{\pi}}), (0, -\frac{2}{\sqrt{\pi}}), (-\frac{2}{\sqrt{\pi}}, -\frac{2}{\sqrt{\pi}})$ 做為比較依據。如果在某方位，紅色多邊形相比之下越是內縮，則代表對應的相關性越是正向；相反的，越是超出，則代表對應的相關性越是負向。圖2為加上導讀多邊形的蒲公英圖，不論在水平或鉛直方向，紅線與黑線皆近乎重疊，代表兩區間資料呈現獨立的中點相關性與全距相關性。

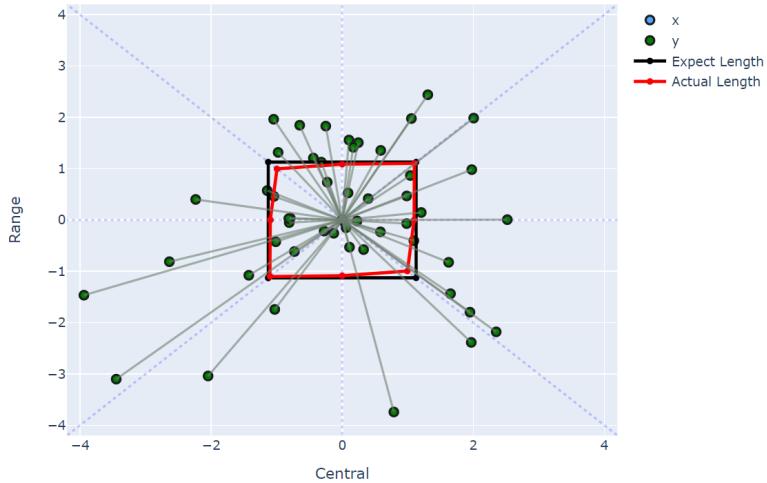


圖 2：蒲公英圖與導讀多邊形，其中 $X_c \sim N(0, 3)$, $X_r = X_c^2$, $Y_c \sim N(0, 3)$, $Y_r \sim N(3, 1)$ ，兩變數各生成100筆。

3 模擬結果

在本章節中，將透過自迴歸關係式(10)與(13)，模擬區間值時間序列的數據，進而在不同的配置中，驗證模型對參數的估計能力，並且探討參數對模型預測

的影響。每種設定下的模擬，皆重複進行1000次，並應用粒子群演算法(particle swarm optimization, PSO)，極大化模型的對數概似函數。同時，我們使用相對誤差(Relative error，簡稱RE)，作為衡量估計準確度的標準。相對誤差RE的定義如下：

$$RE = \frac{RMSE}{| True Value |},$$

其中RMSE為均分根誤差 (root mean square error)的縮寫。

3.1 參數估計

在樣本模擬自AIR模型時，固定參數 ϕ 與 σ ，我們便可討論時間點個數(m)與單一時間點下樣本個數(n)對參數估計的影響。表2為不同時間個數(m)與樣本數(n)的組合下，AIR(1)模型的估計誤差。 ϕ 的估計誤差皆低於6%，而 σ 的估計誤差皆低於3%。當時間個數(m)增加時，參數的估計誤差皆明顯降低，但單純地增加樣本數(n)，卻無法有效提高估計的準確度，顯示增加時間點個數(m)，對模型的估計有正向的效用。

表 2: 當 $\sigma = 0.02$ 、 $\phi = -0.7$ ，不同時間個數(m)與樣本數(n)的組合下，AIR(1)模型對參數估計的相對誤差。

m	100	250	500	750	250	250	500
n	1000	1000	1000	1000	500	250	500
$\hat{\phi}$	0.0585	0.0362	0.0261	0.0215	0.0359	0.0377	0.0260
$\hat{\sigma}$	0.0263	0.0163	0.0119	0.0098	0.0162	0.0175	0.0119

固定時間個數(m)與樣本數(n)，我們亦可探討參數 ϕ 與 σ 的真實數值，對彼此估計的影響。表3呈現不同 σ 的組合下，AIR(1)模型的估計誤差，顯示 σ 真實數值的大小，並不影響 ϕ 值與自身的估計；但 ϕ 真實數值的大小，似乎同時影響自身與 σ 值的估計。

表 3: 當 $m = 250$ 、 $n = 1000$ 、 $\phi = -0.7$ 與 0.7 ，不同 σ 的組合下，AIR(1) 模型對參數估計的相對誤差。

$\phi = -0.7$										
σ	0.0001	0.001	0.005	0.01	0.02	0.04	0.08	0.1	0.5	1
$\hat{\phi}$	0.0364	0.0358	0.0373	0.0368	0.0378	0.0372	0.0366	0.0365	0.0365	0.0370
$\hat{\sigma}$	0.0165	0.0160	0.0167	0.0166	0.0171	0.0168	0.0165	0.0163	0.0164	0.0167
$\phi = 0.7$										
σ	0.0001	0.001	0.005	0.01	0.02	0.04	0.08	0.1	0.5	1
$\hat{\phi}$	0.0135	0.0133	0.0135	0.0137	0.0133	0.0138	0.0132	0.0136	0.0139	0.0139
$\hat{\sigma}$	0.0339	0.0333	0.0340	0.0340	0.0330	0.0345	0.0330	0.0337	0.0346	0.0347

表 4 為不同 ϕ 的組合下，AIR(1) 模型的估計誤差。在 $\phi < 0$ 時， σ 的估計誤差普遍較低，且誤差隨著 ϕ 值的增加而提高；而在 $\phi > 0$ 時， σ 的估計誤差相對較高，且誤差亦隨著 ϕ 值的增加而提高。 ϕ 的估計誤差，隨著 $|\phi|$ 增加而降低。

表 4: 當 $m = 250$ 、 $n = 1000$ 、 $\sigma = 0.02$ 與 0.5 ，不同 ϕ 的組合下，AIR(1) 模型對參數估計的相對誤差。

$\sigma = 0.02$									
ϕ	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
$\hat{\phi}$	0.0201	0.0283	0.0366	0.0445	0.0566	0.0694	0.0946	0.1353	0.2622
$\hat{\sigma}$	0.0111	0.0140	0.0166	0.0181	0.0202	0.0215	0.0235	0.0239	0.0254
$\sigma = 0.02$									
ϕ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$\hat{\phi}$	0.0046	0.0088	0.0130	0.0200	0.0298	0.0434	0.0615	0.1061	0.2379
$\hat{\sigma}$	0.0448	0.0378	0.0326	0.0321	0.0321	0.0307	0.0282	0.0283	0.0281
$\sigma = 0.5$									
ϕ	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
$\hat{\phi}$	0.0206	0.0284	0.0370	0.0472	0.0548	0.0717	0.0936	0.1376	0.2558
$\hat{\sigma}$	0.0114	0.0141	0.0166	0.0191	0.0199	0.0221	0.0234	0.0245	0.0250
$\sigma = 0.5$									
ϕ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$\hat{\phi}$	0.0045	0.0076	0.0130	0.0191	0.0288	0.0406	0.0597	0.1066	0.2301
$\hat{\sigma}$	0.0439	0.0328	0.0326	0.0303	0.0305	0.0288	0.0274	0.0281	0.0273

最後，我們針對結構較複雜的HVAIR模型進行模擬。表 5 為不同參數設定下，HVAIR(1,1) 模型的估計誤差。當 $\phi < 0$ 時， β_0 的估計較為準確，而 β_1 的估計

則相對不準確。 β_1 的估計誤差，隨著 β_1 真實數值的增加而明顯減少。當 $|\phi|$ 增加時， ϕ 的估計誤差也隨之降低，此現象與AIR(1)模型的模擬結果相似。

表 5: 當 $m = 250$ 、 $n = 1000$ ，不同參數設定下，HVAIR(1,1)模型對參數估計的相對誤差。

真實數值			相對誤差			真實數值			相對誤差		
ϕ	β_0	β_1	$\hat{\phi}$	$\hat{\beta}_0$	$\hat{\beta}_1$	ϕ	β_0	β_1	$\hat{\phi}$	$\hat{\beta}_0$	$\hat{\beta}_1$
-0.8	0.04	0.2	0.0354	0.0322	0.4535	0.2	0.04	0.2	0.1095	0.1032	0.2500
-0.8	0.04	0.8	0.0375	0.0452	0.1291	0.2	0.04	0.8	0.1252	0.2131	0.0593
-0.5	0.04	0.2	0.0660	0.0463	0.4562	0.5	0.04	0.2	0.0298	0.1210	0.0791
-0.5	0.04	0.8	0.0657	0.0639	0.1216	0.5	0.04	0.8	0.0528	0.2740	0.0624
-0.2	0.04	0.2	0.1487	0.0637	0.4076	0.8	0.04	0.2	0.0212	0.1564	0.0820
-0.2	0.04	0.8	0.1533	0.1006	0.1064	0.8	0.04	0.8	0.0325	0.3240	0.0613

3.2 預測能力

為了評估AIR模型的預測能力，我們使用均距離誤差(Mean distance error, MDE)衡量兩區間向量的差異性。令 $\mathbf{X}_t = [X_{u,t}, X_{l,t}]$ 為實際區間的上下界， $\hat{\mathbf{X}}_t = [\hat{X}_{u,t}, \hat{X}_{l,t}]$ 為預測區間的上下界，則MDE定義如下：

$$MDE = \sqrt{\frac{\sum_{t=1}^T (|X_{u,t} - \hat{X}_{u,t}|^2 + |X_{l,t} - \hat{X}_{l,t}|^2)}{2T}},$$

其中 T 為預測的時間步數。

表6為不同 ϕ 值與預測步數(h)的配置下，AIR(1)模型的平均預測誤差；括弧內的數值，為預測誤差的變異數。當僅僅預測一步時， ϕ 的真實數值並不影響預測的準確度，預測誤差均落在30%左右，但預測步數增加至一步以上時，預測誤差隨著 $|\phi|$ 增加而提高，且預測誤差的變異程度，亦隨著 $|\phi|$ 增加而增加。

表 6: 當 $m = 250$ 、 $n = 1000$ 、 $\sigma = 1$ ，不同 ϕ 值與預測步數(h)的配置下，AIR(1)模型的平均預測誤差；括弧內的數值，為預測誤差的變異數。

$\sigma = 1$									
ϕ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$h = 1$	0.3077 (0.1753)	0.3028 (0.1667)	0.3072 (0.1724)	0.3089 (0.1716)	0.3147 (0.1826)	0.3015 (0.1688)	0.3067 (0.1830)	0.3112 (0.1748)	0.3019 (0.1728)
$h = 2$	0.3831 (0.1907)	0.3614 (0.1647)	0.3660 (0.1675)	0.3483 (0.1509)	0.3429 (0.1502)	0.3252 (0.1439)	0.3365 (0.1438)	0.3300 (0.1410)	0.3284 (0.1382)
$h = 3$	0.4328 (0.1937)	0.4145 (0.1828)	0.3921 (0.1626)	0.3665 (0.1441)	0.3621 (0.1351)	0.3528 (0.1309)	0.3436 (0.1228)	0.3346 (0.1200)	0.3284 (0.1150)

$\sigma = 1$									
ϕ	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
$h = 1$	0.2889 (0.1581)	0.3017 (0.1683)	0.3052 (0.1676)	0.3085 (0.1698)	0.3079 (0.1790)	0.3082 (0.1755)	0.3052 (0.1816)	0.3001 (0.1615)	0.3128 (0.1807)
$h = 2$	0.3817 (0.1802)	0.3670 (0.1760)	0.3558 (0.1617)	0.3471 (0.1564)	0.3378 (0.1498)	0.3362 (0.1408)	0.3264 (0.1375)	0.3251 (0.1389)	0.3296 (0.1405)
$h = 3$	0.4293 (0.1852)	0.4131 (0.1767)	0.3825 (0.1525)	0.3694 (0.1450)	0.3542 (0.1328)	0.3480 (0.1252)	0.3459 (0.1231)	0.3369 (0.1174)	0.3353 (0.1172)

4 實證資料分析

本章節將探討高屏地區的空氣品質，藉由不同測站所收集到的PM_{2.5}資料，分別配適傳統的VAR模型(參考Gao and Tsay, 2019)，以及第2.4節提到的區間數值時間序列模型(AIR模型與HVAIR模型)，將配適完後的殘差計算出相關係數矩陣，藉此看出不同測站空間中的關聯性。

圖3為完整的研究流程圖，其中資料預處理的部分敘述在第4.1節；時間序列趨勢的除去，在第4.2節中呈現；而模型配適的流程，以及各測站相關性的探討，將分別於第4.3節與第4.4節，給予完整的描述。

為了呈現的簡潔性，之後的分析僅使用大寮測站的PM_{2.5}觀測值，作為分析標的；剩餘測站的分析流程將省略，僅將結果整理成表格，且依此結果進行討論。

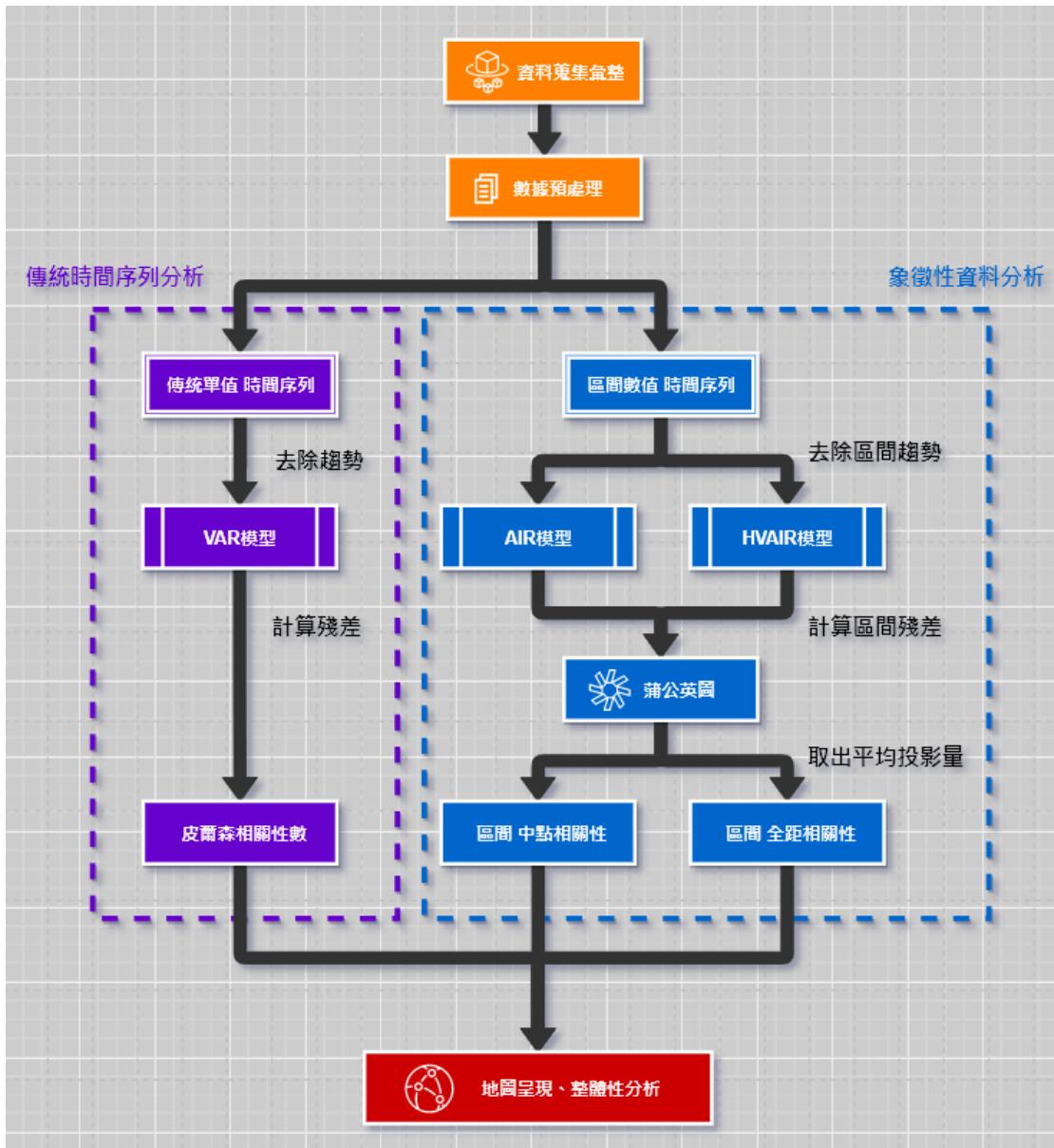


圖 3: 研究流程圖。

4.1 數據選擇與資料處理

本研究使用行政院環保署，於開放平台公布的空氣品質監測小時值資料，作為PM_{2.5}數據的主要來源。分析範圍為高屏空氣品質區的15個測站，包含潮州、大寮、鳳山、復興、恆春、林園、美濃、楠梓、屏東、前金、前鎮、橋頭、仁武、小港以及左營測站，如圖4所示，時間橫跨2006年1月1日至2015年12月31日(在本研究執行的期間，適逢環保署官網的大更新，2013年前的資料，如今似乎沒有公開，所以我們將資料上傳至github上，以提供有需要的人士: <https://github.com/halosung/Symbolic-Data-Analysis>)。

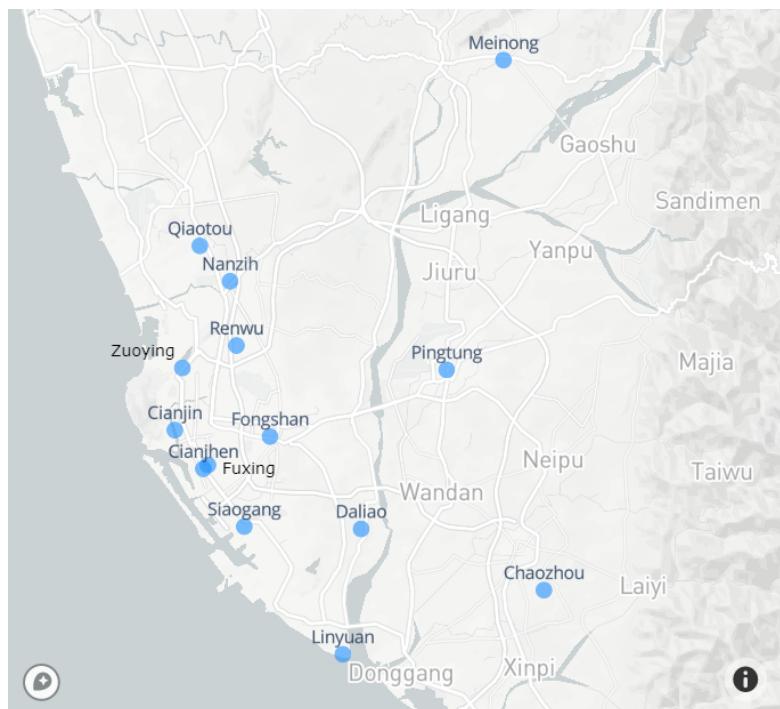
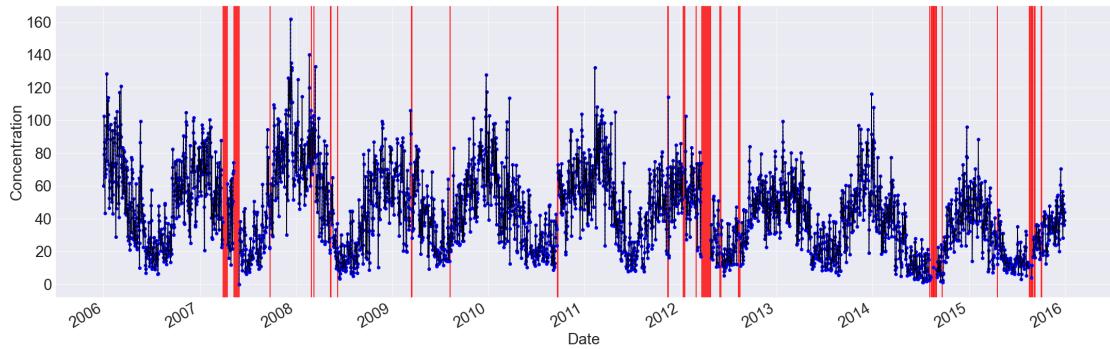
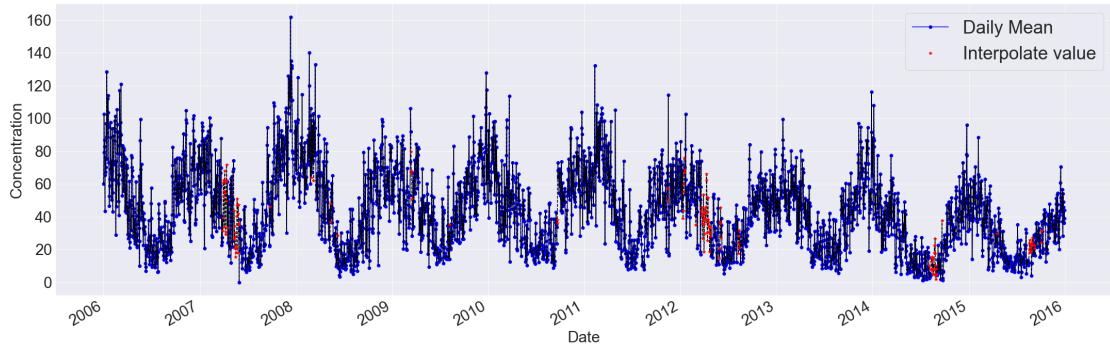


圖 4: 各測站地理位置。恆春測站與其他測站相距甚遠，故在地圖上予以省略。

在初步處理中，每小時觀測值的時間序列，共有5588筆遺失值。由於缺失筆數不少，加上觀測濃度的趨勢複雜，不易填補，我們先以「日」為單位，用計算日平均的方式，對資料進行合併。圖5(a)為日均濃度的時間序列圖，其中紅色區域表示缺失值發生的時間點，共有148筆。在缺失值大幅降低的情況下，我們採用Seasonally splitted missing value imputation的插補方式(其中R的指令為na_seasplit(data, find_frequency=TRUE))，針對日遺失值進行填補，如圖5(b)所示，填補值以紅點代表。



(a) 填補前 日平均濃度



(b) 填補後 日平均濃度

圖 5: 大寮觀測站 $\text{PM}_{2.5}$ 日均濃度時間序列圖。

資料完備的情況下，我們再以「週」為單位合併資料，並進行方根轉換，完成必要的前處理。針對單維度的時間序列，我們將每週的日平均濃度，再取一次平均，形成週平均濃度的時間序列，等同定義式(1)所示；另一方面，針對區間值時間序列，我們提取每週日平均濃度的極大值與極小值，構成區間值的上下界，如定義式(2)所示。圖6呈現轉換後，週平均濃度的時間序列，與週區間時間序列的上下界。平均濃度與區間濃度的上下界間，存在高度相似的趨勢。

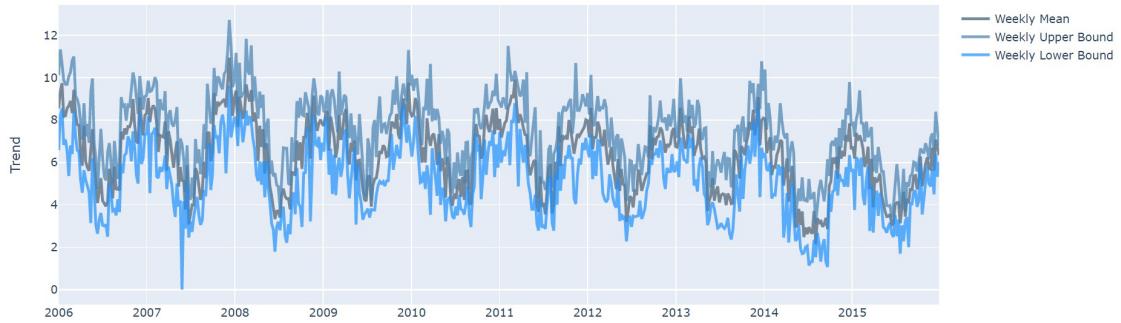


圖 6: 方根轉換後，大寮測站週平均濃度與週區間濃度的時間序列圖。

4.2 去除趨勢

時間序列分析中，經常運用時間序列分解法，將單維度的時間序列，拆解為趨勢項，季節項與剩餘項。遵循相同的理念，我們除去週資料的趨勢項與季節項，僅對平穩的剩餘項進行後續的建模。由於15個測站的週平均濃度與週區間濃度的上下界間，具有高度相似的趨勢，我們先估計週平均濃度的趨勢項及季節項，再將其從週平均與週區間資料中扣除，藉此一併去除區間值時間序列，上下界的趨勢。

已知週資料的循環週期約為一年(52週)，我們參考Gao and Tsay (2019)的分析手法，在此資料中，選取 $d \in \{0, 1, 2\}$, $k \in \{1, \dots, 25\}$, $s = 52$ 為(8)式的候選模型，並透過貝氏信息量準則，挑選 $\hat{d} = 2$, $\hat{k} = 3$ ，意即以二次方程式估計序列的趨勢項，並以三個三角級數的線性組合估計序列的季節項。圖7同時呈現去除趨勢後的剩餘項，以及剩餘項的自我相關函數(Autocorrelation function, ACF)；在ACF中，剩餘項仍殘有些微的季節性效應，顯示此分解方法，無法完全將趨勢從剩餘項中去除。為了避免模型建立的困難，我們採用第2.2節提到的時間序列分解方法—STL。

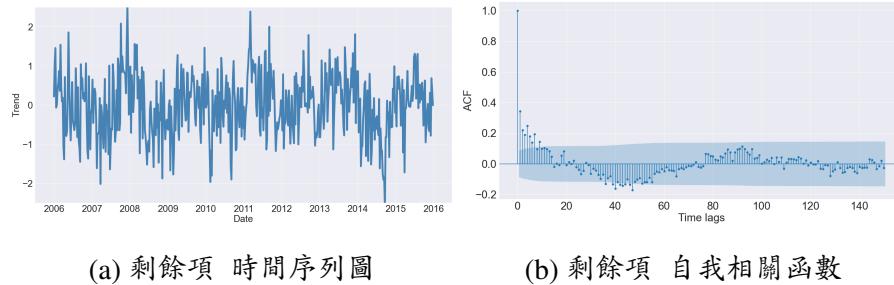
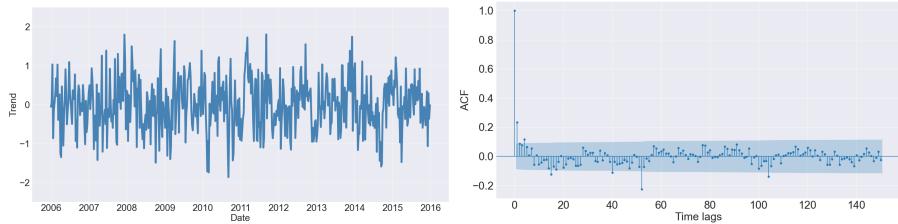


圖 7: Gao and Tsay (2019)時間序列分解法。

圖8為週平均濃度的STL分解結果。同時比對兩分解方法剩餘項的自我相關函數，我們可以發覺，STL複雜的估計結構，較能有效的將趨勢效應自剩餘項中排除，故本研究選擇STL作為去除趨勢的主要方法。圖9為去除趨勢後，週平均濃度與週區間濃度的時間序列圖；區間上界多為正數，下界多為負數。



(a) 剩餘項 時間序列圖

(b) 剩餘項 自我相關函數

圖 8: STL時間序列分解法。

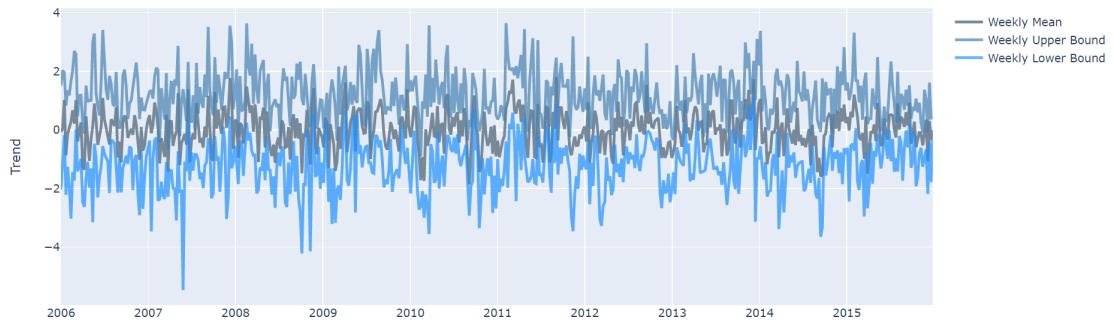


圖 9: 去除趨勢後，大寮測站週平均濃度與週區間濃度的時間序列圖。

圖10為15個測站，趨勢項的時間序列圖。我們可以發現，在2010年中期後，各測站PM_{2.5}的濃度走勢，皆有明顯下降的趨勢，並在2015年達到最低的數值；此現象與國家的管制規範，有著密切的關聯：以2010年的「反國光石化運動」為濫觴，空污議題逐漸成為民間的關注焦點，更進一步影響政府部門，成為重要的施政目標，高屏空品區的空氣品質，因此獲得相當程度的改善。

圖11呈現15個測站，季節項的時間序列圖。PM_{2.5}濃度在每年冬季時達到高峰，而在夏季時大幅降低；此明顯的季節效應，與地形、風向習習相關：受西伯利亞冷高壓的影響，台灣冬季盛行東北季風，南部地區因位於雪山與中央山脈的背風面，大氣流相當穩定，污染物質擴散不易，又加上季風夾帶北部與中部的污染物在此堆積，造成冬季污染最為嚴重；反之，夏季盛行的西南季風，使得高屏空品區對流良好，擴散能力提升，污染程度便隨之下降，與冬季形成強烈的對比。

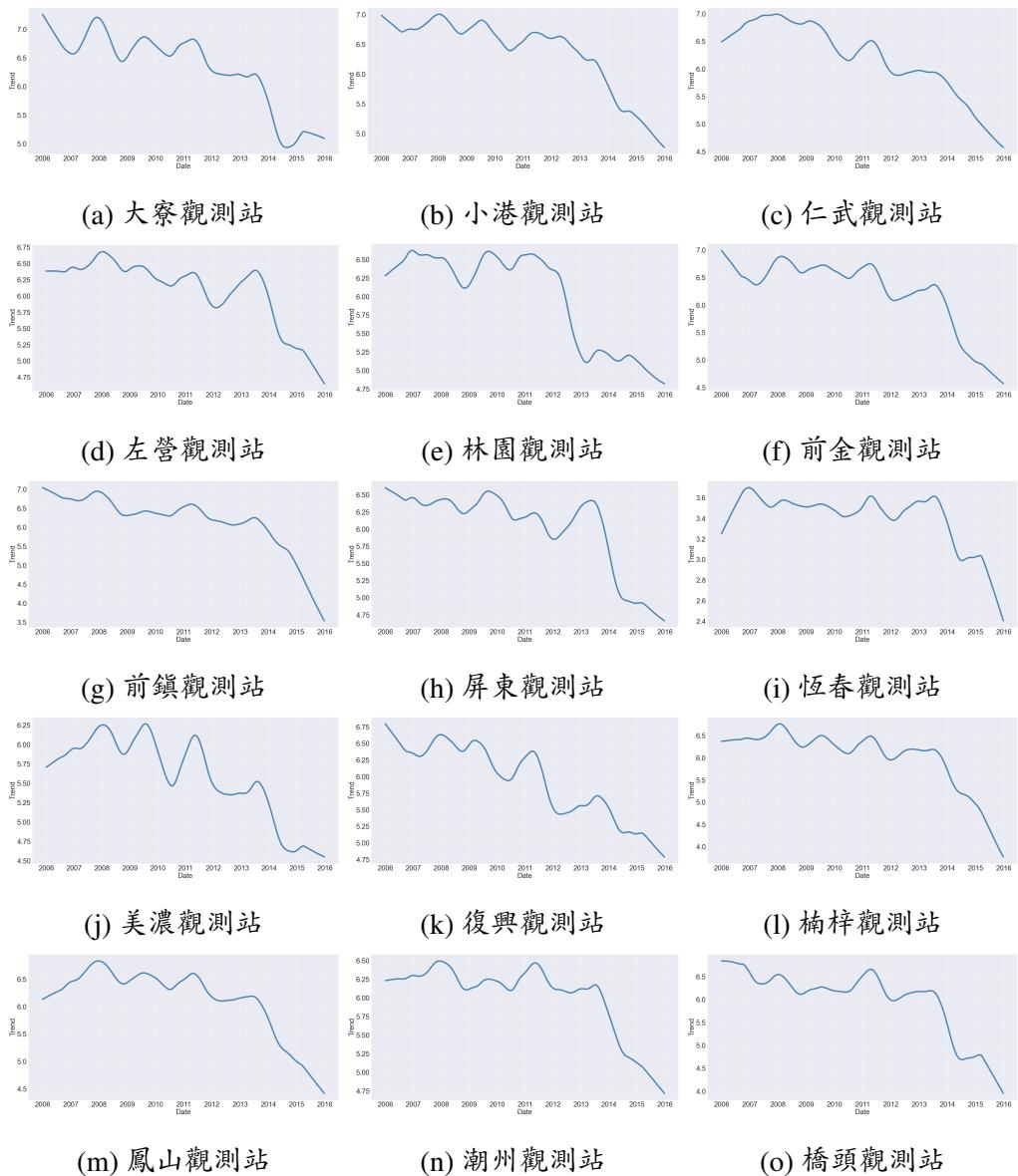


圖 10: 高屏空品區 各測站PM_{2.5}濃度的趨勢項。

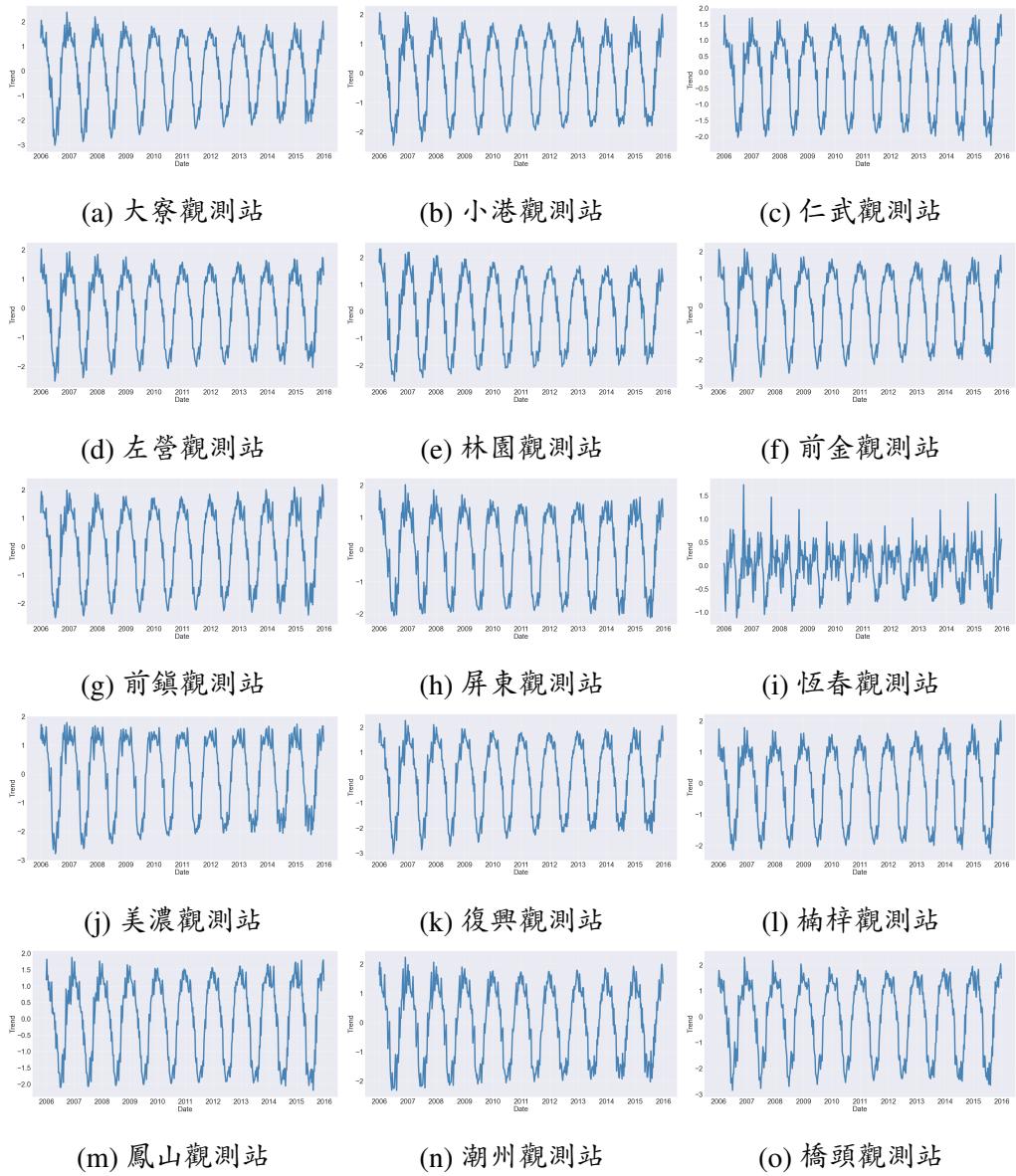


圖 11: 高屏空品區 各測站PM_{2.5}濃度的季節項。除恆春觀測站外，其餘觀測站皆具有相似的趨勢。

4.3 模型配適

針對平穩的時間序列，我們依據不同的變數型態，選擇合適的時序模型。首先，我們參考Gao and Tsay (2019)的結果，使用VAR(1)模型，同時對15個測站的平均濃度建模；而在區間值時間序列中，我們考慮AIR模型與HVAIR模型的可能性，分別對15個測站的濃度區間，配適AIR(p)模型與HVAIR($p,1$)模型，在自迴歸項數 $p \in \{1, 2, \dots, 25\}$ 的限制下，以MDE為評斷標準，選擇配適誤差最小的模型，作為測站的代表模型。

表7呈現不同模型的配適誤差、赤池訊息量準則(AIC)與貝氏訊息量準則(BIC)。我們可以觀察到，隨著自迴歸項數的增加，配適誤差先升後降，在HVAIR(25,1)模型的配適下，達到最小值1.046。根據前述的選模標準，我們選擇誤差最小的HVAIR(25,1)模型，作為解釋模型。為了檢視模型的配適程度，我們同時繪製該模型，配適區間值的時間序列圖，如圖12所示。不難察覺，該模型最大的特點，在於成功捕捉真實區間值的大致趨勢，但針對急劇震動的上下界，卻顯得力有未逮。以提升解釋能力為發想，我們嘗試搭配「移動視窗」的技巧，重新選擇模型，期望優化配適結果。

表 7: 區間式時間序列模型的配適結果，最終選取模型以顏色標記。

Model	MDE	AIC	BIC	Model	MDE	AIC	BIC
AIR(1)	1.300	30128.367	30136.883	HVAIR(1,1)	1.233	28770.239	28783.012
AIR(5)	1.699	16997.975	17023.521	HVAIR(5,1)	1.320	17346.811	17376.614
AIR(10)	1.480	14291.971	14338.806	HVAIR(10,1)	1.167	14998.725	15049.817
AIR(15)	1.314	14291.971	14338.806	HVAIR(15,1)	1.087	13933.197	14005.577
AIR(20)	1.217	12217.824	12307.235	HVAIR(20,1)	1.065	13267.293	13360.962
AIR(25)	1.149	11793.541	11904.241	HVAIR(25,1)	1.046	13039.644	13154.601

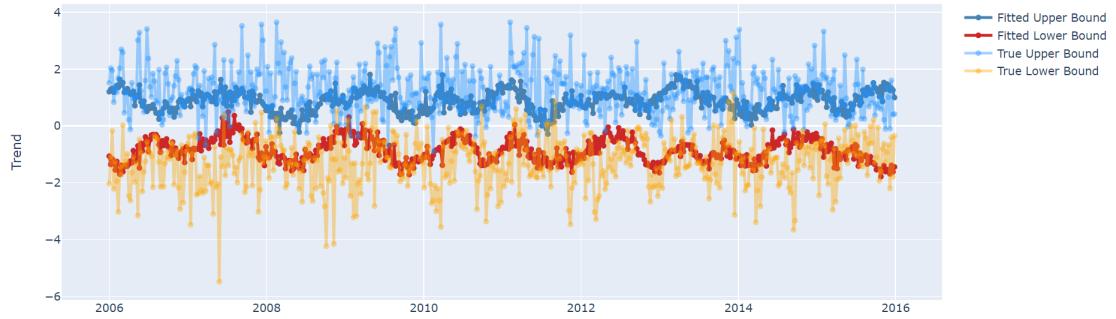


圖 12: 大寮測站 HVAIR(25,1)配適結果。

「移動視窗」的設計，如圖13所示。固定時間長度為一年，移動間隔為半年的情況下，我們將時間序列劃分成20個時間窗，並針對各時間窗的區間資料，同時配適AIR(p)模型與HVAIR($p,1$)模型，在自迴歸項數 $p \in \{1, 2, \dots, 25\}$ 的限制下，改以平均MDE做為判斷基準，尋找整體表現最佳的模型。

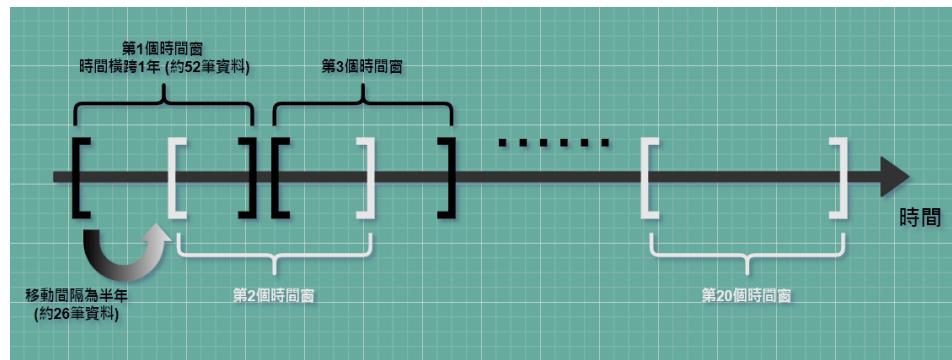


圖 13: 移動視窗示意圖。

表8為加入移動視窗後，不同模型的配適結果。隨著自迴歸項數的增加，配適的平均誤差亦先升後降，在AIR(25)模型的配適中，達到最小值0.806。由於此數值明顯優於HVAIR(25,1)模型的結果，我們改選AIR(25)模型，搭配移動視窗，構成最終的建模流程。圖14為AIR(25)模型，配適區間值的時間序列圖。我們可以發現，加入移動視窗的處理後，估計區間的上下界，更加符合真實資料的震盪趨勢，顯示模型的解釋能力，確實獲得顯著的提升。

表 8: 區間式時間序列模型移動視窗的配適結果，最終選取模型以顏色標記。括弧內為各判斷指標的標準差，但由於AIC與BIC的標準差相同，故僅附上前者的標準差。

Model	MDE	AIC	BIC	Model	MDE	AIC	BIC
AIR(1)	1.246(0.118)	2929.395(674.606)	2933.298	HVAIR(1,1)	1.184(0.115)	2804.099(718.223)	2809.952
AIR(5)	1.535(0.298)	1479.129(208.264)	1490.837	HVAIR(5,1)	1.248(0.160)	1517.071(234.726)	1530.730
AIR(10)	1.215(0.174)	1117.771(89.979)	1139.235	HVAIR(10,1)	1.064(0.117)	1162.128(98.545)	1185.543
AIR(15)	0.990(0.144)	900.024(48.374)	931.244	HVAIR(15,1)	0.928(0.092)	940.727(60.628)	973.898
AIR(20)	0.882(0.130)	747.313(34.991)	788.289	HVAIR(20,1)	0.886(0.115)	785.69(50.463)	828.619
AIR(25)	0.806(0.177)	616.374(22.297)	667.107	HVAIR(25,1)	0.858(0.127)	644.812(33.878)	697.496

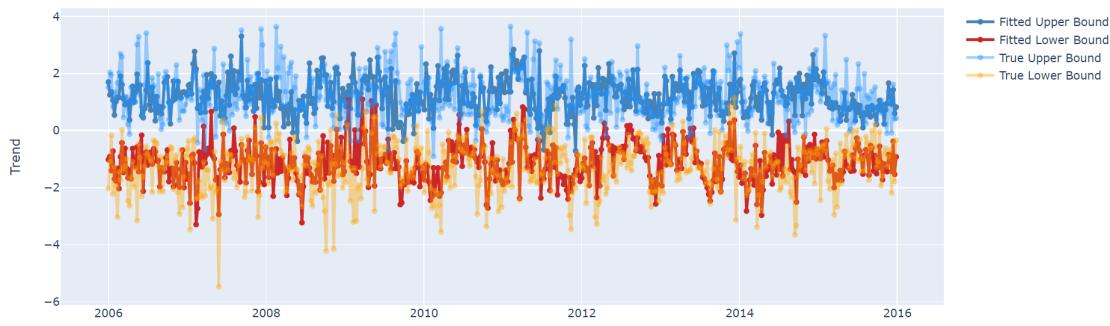


圖 14: 大寮測站 AIR(25)移動視窗配適結果。

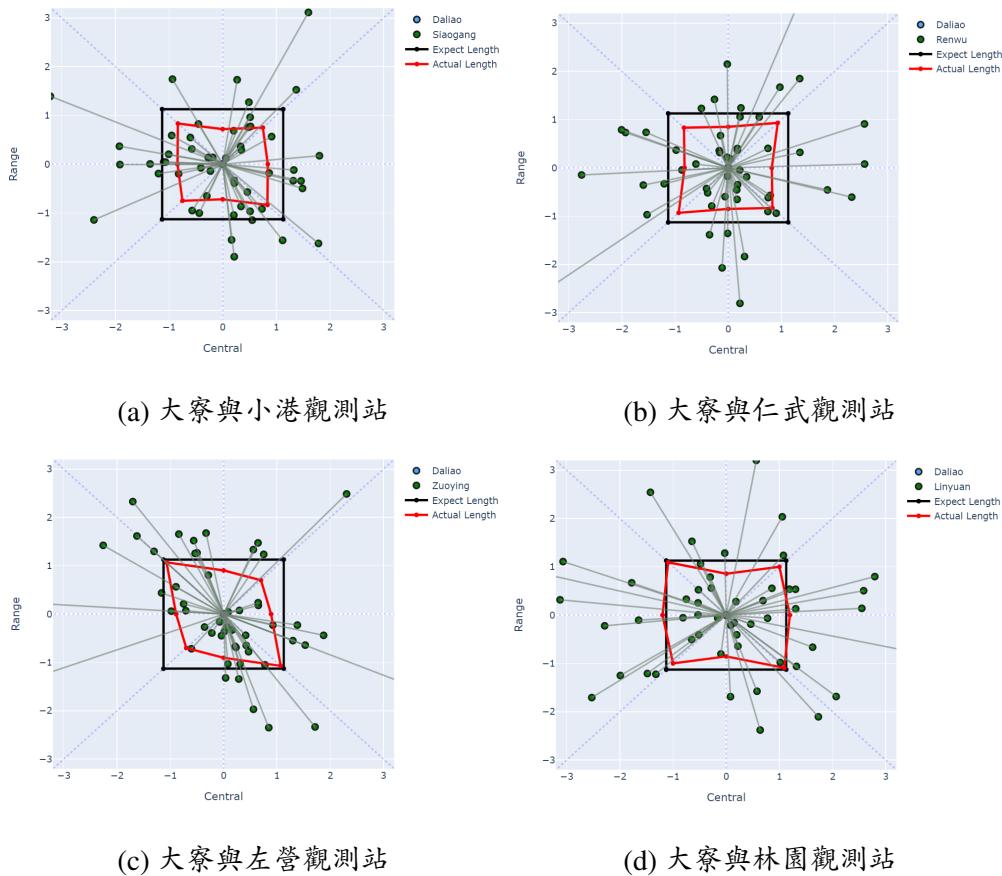
同時比較表7與表8，在相同模型的配適下，藉由移動視窗的技巧，我們能大幅降低平均的配適誤差，此現象彰顯象徵性資料分析的特色—相比於傳統的資料形式，象徵性資料能在相同的時間範圍內，保留更多的資訊，故針對PM_{2.5}濃度建立模型，我們僅需要少量的觀測值，便能優化模型對真實區間的擬合，增進模型對細微變化的解釋能力。同時，為了便於後續的比較，在週平均濃度的建模中，我們亦加入移動視窗的技巧，將時間序列劃分成20個時間窗，並針對各時間窗的資料，配適VAR(1)模型。

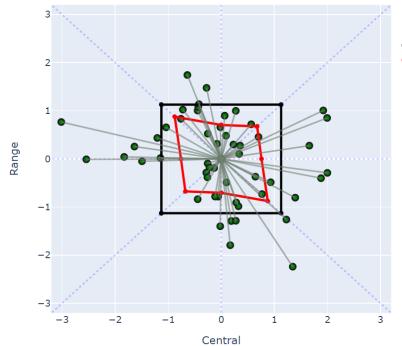
4.4 區間相關性視覺化

模型配適完成後，我們計算模型剩餘的殘差，並使用不同的呈現手法，描繪其時間序列的關聯性。在區間值時間序列中，本研究採用Zhang and Lin (2020)提出的視覺化方法，藉由蒲公英圖，尋找兩區間資料中心與全距的相關性。由於我們使用移動視窗的配適方法，相鄰時間窗所產生的殘差間，有部分時間範圍

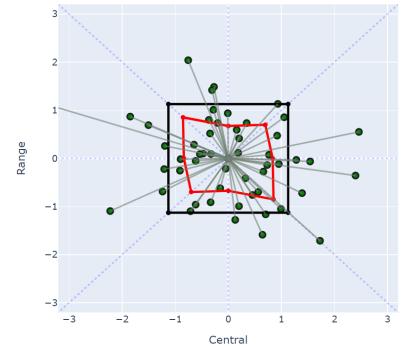
重疊，為了分析的合理性，我們每次固定一個時間窗，在相同時間範圍下，比較15個測站的相關性。

固定時間範圍為2012年，圖15為各測站與大寮測站的蒲公英圖。根據導讀多邊形的提示，多數測站與大寮測站間，存在正向的區間相關性，但程度不盡相同。在水平方位(x 軸)上，紅色多邊形分別在前金、復興、鳳山與潮州測站的蒲公英圖中，較為內縮，顯示這些測站與大寮測站間，存有正向且明顯的中點相關性；此外，在林園測站的蒲公英圖中，多邊形相對突出，代表林園與大寮測站間具有略微負向的中點相關性。而在鉛直方位(y 軸)上，紅色多邊形分別在小港、前金、前鎮、美濃與復興測站的蒲公英圖中，較為內縮，顯示這些測站與大寮測站間，擁有正向且明顯的全距相關性。

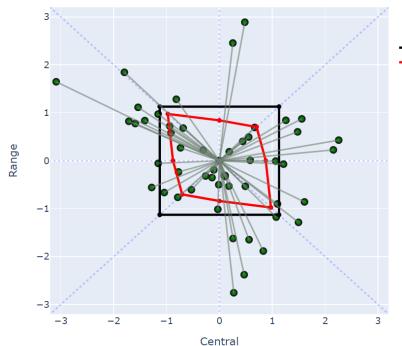




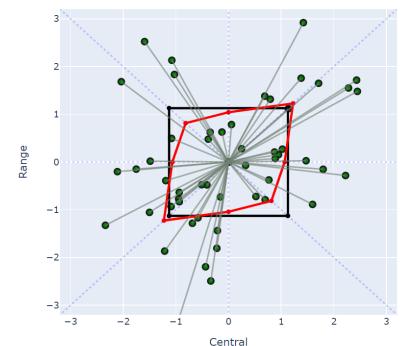
(e) 大寮與前金觀測站



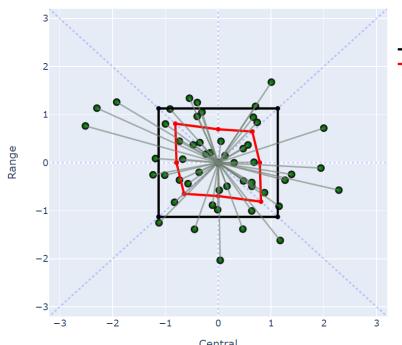
(f) 大寮與前鎮觀測站



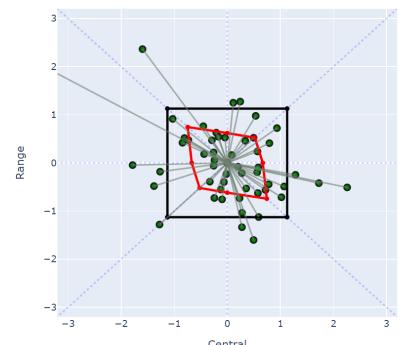
(g) 大寮與屏東觀測站



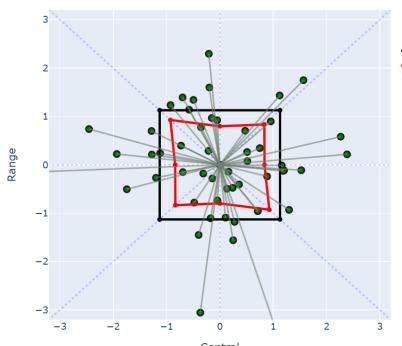
(h) 大寮與恆春觀測站



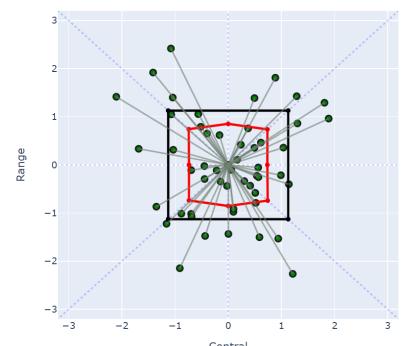
(i) 大寮與美濃觀測站



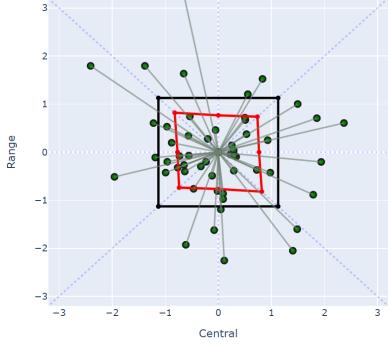
(j) 大寮與復興觀測站



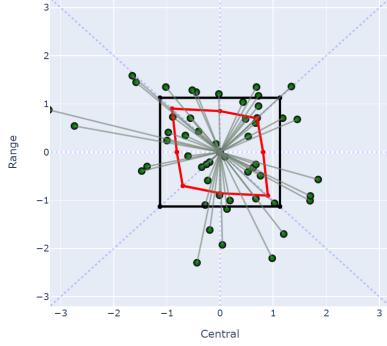
(k) 大寮與楠梓觀測站



(l) 大寮與鳳山觀測站



(m) 大寮與潮州觀測站



(n) 大寮與橋頭觀測站

圖 15: 各測站與大寮測站的蒲公英圖。

在蒲公英圖的應用下，我們得以迅速了解，兩區間資料的區間相關性。然而，隨著分析測站的增加，倘若一次僅進行兩兩資料集的比較，我們會需要更多空間，呈現測站彼此的關聯性。參考Zhang and Lin (2020)的描述，線段水平與鉛直的平均投影量，可反映兩區間資料的中點相關性與全距相關性。因此，我們改以平均投影量的數值大小，直接表示相關性的強烈程度；如此一來，透過紀錄測站間平均投影量的數值，我們即可藉由矩陣的形式，同時呈現15個測站的區間相關性。值得注意的是，由於存在20個時間窗，每種不同的相關性都可生成20個矩陣。為了探討各測站整體的相關性，我們將20個矩陣取平均，呈現各測站長時間的平均相關性。

令變數 p 表示矩陣中任兩測站間的平均投影量，為了增進閱讀的直觀性，我們經由簡單的線性轉換 $p' = \frac{2}{\sqrt{\pi}} - p$ ，使用新的指標 p' 衡量兩測站的區間相關性，且 p' 的值域範圍應滿足 $p' \in (-\infty, \frac{2}{\sqrt{\pi}}]$ 。在對應的相關性中， $p' = \frac{2}{\sqrt{\pi}}$ 表示兩者完全正相關， $p' = 0$ 表示兩者無關， $p' < 0$ 表示兩者負相關。同時，為了聚焦在較強烈的區間相關性，我們亦針對 p' 的值域範圍進行切割與定義，並對中度以上的區間相關性，進行後續的探討。

倘若在正數範圍內， p' 與相關係數 ρ 之間，存在一一對應的比例關係，則 p' 中度相關的範圍，應約略介於 $0.45(\frac{2}{\sqrt{\pi}} \times 0.4)$ 與 $0.67(\frac{2}{\sqrt{\pi}} \times 0.6)$ 之間，故我們以0.45為分界線，特別討論大於0.45的區間相關性。表9及表10分別為中點相關性矩陣與全距相關性矩陣，並使用顏色標示中度以上的區間相關性。其中，無論在中點相關性或是全距相關性，屏東與潮州測站間的關聯，皆為所有測站的組合中，最為強烈者。

表 9: 中點相關性矩陣。

測站名稱	美濃	橋頭	楠梓	仁武	左營	屏東	前金	鳳山	復興	前鎮	小港	大寮	潮州	林園	恆春
美濃	—	0.3812	0.3493	0.3560	0.2955	0.4480	0.3899	0.3606	0.3270	0.3144	0.4218	0.3692	0.3809	0.2759	0.1413
橋頭	0.3812	—	0.4272	0.4169	0.4290	0.3310	0.5540	0.4100	0.4325	0.4226	0.4533	0.3666	0.2894	0.2788	0.1540
楠梓	0.3493	0.4272	—	0.5255	0.4940	0.3005	0.4061	0.4370	0.4532	0.4606	0.3757	0.3864	0.3418	0.3117	0.1728
仁武	0.356	0.4169	0.5255	—	0.5173	0.3257	0.4369	0.5029	0.4785	0.5034	0.3798	0.4145	0.3884	0.3496	0.1742
左營	0.2955	0.429	0.4940	0.5173	—	0.3026	0.4408	0.4880	0.5074	0.5168	0.3963	0.4139	0.3176	0.3168	0.1825
屏東	0.448	0.331	0.3005	0.3257	0.3026	—	0.3924	0.3327	0.3235	0.3459	0.4268	0.3943	0.6424	0.3211	0.1336
前金	0.3899	0.554	0.4061	0.4369	0.4408	0.3924	—	0.4251	0.4676	0.4688	0.5628	0.3710	0.3186	0.3116	0.1689
鳳山	0.3606	0.4100	0.4370	0.5029	0.4880	0.3327	0.4251	—	0.4823	0.4985	0.4201	0.4715	0.3781	0.3617	0.1411
復興	0.327	0.4325	0.4532	0.4785	0.5074	0.3235	0.4676	0.4823	—	0.6185	0.4419	0.4719	0.3573	0.3427	0.1880
前鎮	0.3144	0.4226	0.4606	0.5034	0.5168	0.3459	0.4688	0.4985	0.6185	—	0.4387	0.4221	0.3796	0.4025	0.1806
小港	0.4218	0.4533	0.3757	0.3798	0.3963	0.4268	0.5628	0.4201	0.4419	0.4387	—	0.3797	0.3376	0.3209	0.1743
大寮	0.3692	0.3666	0.3864	0.4145	0.4139	0.3943	0.3710	0.4715	0.4719	0.4221	0.3797	—	0.4446	0.3357	0.1408
潮州	0.3809	0.2894	0.3418	0.3884	0.3176	0.6424	0.3186	0.3781	0.3573	0.3796	0.3376	0.4446	—	0.3426	0.1499
林園	0.2759	0.2788	0.3117	0.3496	0.3168	0.3211	0.3116	0.3617	0.3427	0.4025	0.3209	0.3357	0.3426	—	0.1478
恆春	0.1413	0.1540	0.1728	0.1742	0.1825	0.1336	0.1689	0.1411	0.1880	0.1806	0.1743	0.1408	0.1499	0.1478	—

表 10: 全距相關性矩陣。

測站名稱	美濃	橋頭	楠梓	仁武	左營	屏東	前金	鳳山	復興	前鎮	小港	大寮	潮州	林園	恆春
美濃	—	0.3554	0.2999	0.3155	0.2874	0.4221	0.3219	0.344	0.2799	0.3086	0.3605	0.3147	0.3784	0.304	0.0397
橋頭	0.3554	—	0.4403	0.4286	0.3953	0.2703	0.5146	0.4226	0.4176	0.4252	0.4551	0.3234	0.2339	0.2970	0.0309
楠梓	0.2999	0.4403	—	0.4998	0.4280	0.2321	0.3864	0.4186	0.3642	0.3993	0.3488	0.3007	0.2374	0.3100	0.0603
仁武	0.3155	0.4286	0.4998	—	0.4914	0.2797	0.4256	0.5284	0.4299	0.4589	0.3868	0.3300	0.3074	0.3517	0.0735
左營	0.2874	0.3953	0.4280	0.4914	—	0.2329	0.4379	0.5015	0.4722	0.5212	0.3956	0.3154	0.2374	0.3319	0.0541
屏東	0.4221	0.2703	0.2321	0.2797	0.2329	—	0.3247	0.3241	0.2820	0.2956	0.3573	0.3237	0.6553	0.3058	0.0602
前金	0.3219	0.5146	0.3864	0.4256	0.4379	0.3247	—	0.4446	0.4539	0.5013	0.5464	0.3194	0.2754	0.3182	0.0442
鳳山	0.344	0.4226	0.4186	0.5284	0.5015	0.3241	0.4446	—	0.4446	0.4539	0.5013	0.5464	0.3194	0.2754	0.3182
復興	0.2799	0.4176	0.3642	0.4299	0.4722	0.2820	0.4539	0.5005	—	0.5005	0.5482	0.4302	0.4259	0.3395	0.3904
前鎮	0.3086	0.4252	0.3993	0.4589	0.5212	0.2956	0.5013	0.5482	0.6274	—	0.5068	0.4104	0.3152	0.3963	0.0447
小港	0.3605	0.4551	0.3488	0.3868	0.3956	0.3573	0.5464	0.4302	0.4448	0.5068	—	0.3525	0.2907	0.3643	0.0463
大寮	0.3147	0.3234	0.3007	0.3300	0.3154	0.3237	0.3194	0.4259	0.4134	0.4104	0.3525	—	0.3373	0.3294	-0.0052
潮州	0.3784	0.2339	0.2374	0.3074	0.2374	0.6553	0.2754	0.3395	0.2909	0.3152	0.2907	0.3373	—	0.3446	0.0457
林園	0.3040	0.2970	0.3100	0.3517	0.3319	0.3058	0.3182	0.3904	0.3571	0.3963	0.3643	0.3294	0.3446	—	0.0632
恆春	0.0397	0.0309	0.0603	0.0735	0.0541	0.0602	0.0442	0.0597	0.0498	0.0447	0.0463	-0.0052	0.0457	0.0632	—

為了納入各測站的地理資訊，進行綜合性的評估，我們在地圖上標示各測站的地理位置，並針對擁有中度相關以上的兩測站，以線段連接。圖16為中點相關性的地圖呈現，透過地圖上交錯的紫色線段，我們不難察覺，中點相關性似乎與測站所在地區密切關聯。

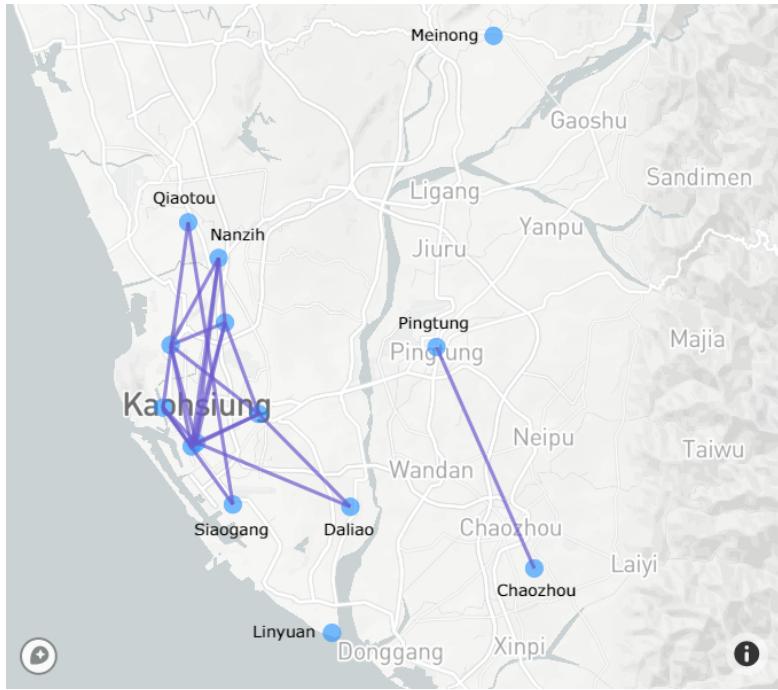


圖 16: 區間中點相關性的地圖呈現，約以高屏溪為界線，劃分為兩個體系。

倘若從統計概念解釋，區間中點的相關性，不僅代表兩區間值時間序列，中點趨勢的協同變化程度，亦同時表示長時間範圍下，污染物質的傳遞情況。地圖中，越接近高雄市中心的測站點，其彼此間的連線便越趨複雜，顯示市中心可能存在較為嚴重的污染源，且污染物質的流通範圍廣闊，包含北方的橋頭區，與南方的小港、大寮區。美濃區及林園區，則與市中心的距離較遠，污染狀況相對獨立。約以高屏溪為分界線，在屏東縣境內，屏東市與潮州鎮間的連線，橫跨狹長的內陸地區，表示污染物質的傳遞範圍，涵蓋重要的人口稠密區；參考表9羅列的相關性數值，兩地區間存在強烈的相關性，我們合理推測，倘若增強其中一區域的空汙防治，則內陸鄉鎮市的空氣品質皆會同步改善；而恆春地區因位處台灣的最南端，與其他14個測站相距甚遠，故其污染狀況較為獨立。

圖17為全距相關性的地圖呈現。與中點相關性相同，全距相關性亦出現地區性的效應，分別在高雄市與屏東縣境內，形成相異的網絡。

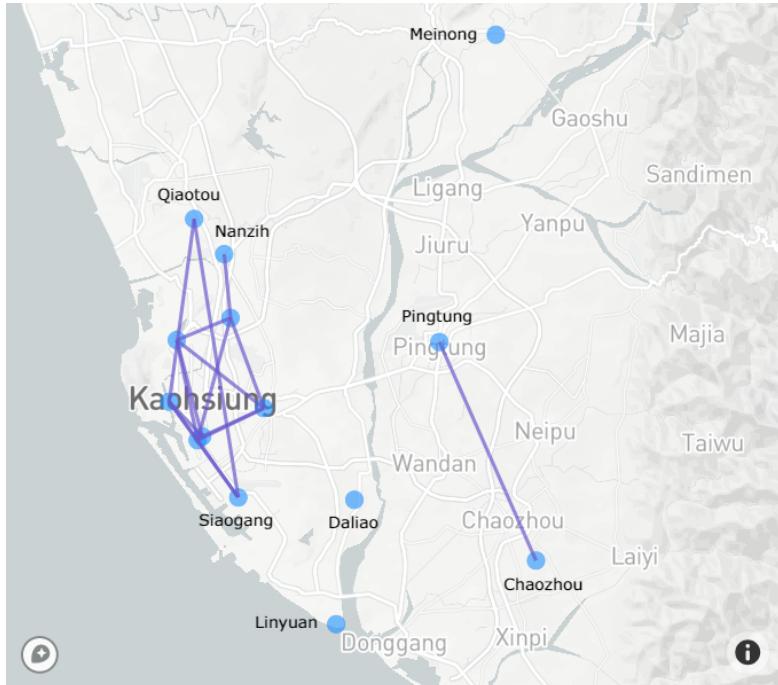


圖 17: 區間全距相關性的地圖呈現，約以高屏溪為界線，劃分為兩個體系。

$PM_{2.5}$ 的觀測數值，深受風向、風速、人類活動等因素影響，倘若發生突發性的變化，則區間資料的全距大小，勢必因此增加。為了觀察短時間範圍內，突發性因子對空氣品質的影響範圍，我們可以透過全距相關性，探討兩區間值時間序列，變異數的相關程度。地圖中，高雄西半部地區的線段相對密集，顯示突發性的變化，將造成市區空氣品質的不穩定，影響範圍包含北方的橋頭區，與南方的小港區。而在屏東縣境內，線段範圍亦涵蓋重要的人口稠密區，顯示內陸地區皆受到相同的突發因子影響，使得觀測濃度的區間大小，隨著時間而產生同步性的增減。

值得一提的是，圖16與圖17間最大的不同之處，在於大寮區與高雄市區的區間關聯。在中點相關性中，不難發現，大寮區與市區的測站間，存在中度相關性，然而在全距相關性中，卻顯得較為獨立。我們可以推測，長時間範圍下，污染物質會流通於大寮區與市區之間，但針對短時間的突發性變化，大寮地區敏感度較低，觀測濃度的變異性，不易受突發因素影響。

針對單維度的時間序列，我們比照相同的分析流程，計算每個時間窗下，各測站殘差的相關係數矩陣，再將20個矩陣取平均，挑選中度相關以上的兩測站，以線段連結的方式，繪製於地圖上。圖18為VAR模型殘差相關性的地圖呈現，顯示幾乎任兩測站間皆有關聯，我們無法透過相關性分析，正確解釋行政區內部，污染物質傳遞的特性；即便將相關性的篩選標準改為高度相關以上，我們仍會得到相同的結論。

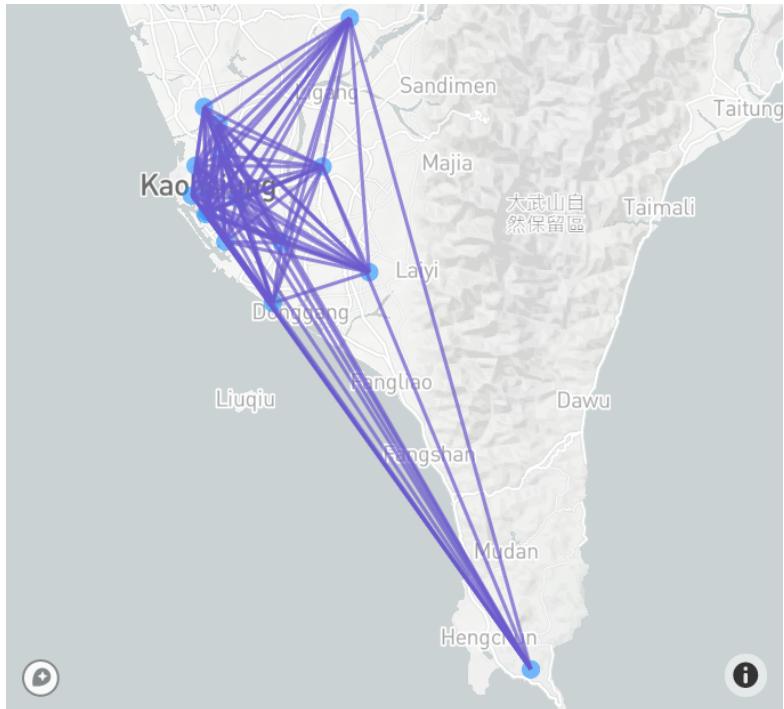


圖 18: VAR模型殘差相關性的地圖呈現，幾乎任兩側站都有相關。

分析結果迥異的主因，我們猜測為「資訊的保留程度」不同所造成。使用週平均濃度來進行分析，等同於把168筆資料濃縮於單一個數值，在資訊被過分簡化的前提下，資料原有的自迴歸效應無法被模型完全解釋，使部分趨勢殘留到殘差項中，形成所有測站皆相互相關的假象；相反的，使用象徵性資料的型態，以週區間資料保存週內觀測濃度所提供的資訊，我們便能在損失訊息量為最小的條件下，利用區間式時間序列模型，除去資料的自迴歸效應，進而使殘差項揭露不同區域間，最真實的相關性。

5 結論與建議

本研究驗證區間式時間序列模型的估計能力，並採用象徵性資料分析的思維，探討南台灣15個測站， $PM_{2.5}$ 濃度的區間相關性，藉此推論不同地區污染物質的傳遞模式與變異特性。透過地圖的呈現，我們發覺中點相關性與全距相關性，皆具有明顯的區域效應；污染物質分別在高雄與屏東境內，形成不同的流通網路。這些現象無法在傳統VAR模型的分析結果中得出，也無形間彰顯象徵性資料在空氣污染防治的應用潛能。

針對本研究的分析流程，我們提供以下改善建議與未來的發展方向：1)區間相關性的定義。由於目前尚未有研究在常態分佈的假設下，推導兩區間變數的相關係數，本研究先以蒲公英圖上線段的平均投影量，代表區間相關性的數值大小，再與傳統的皮爾森相關係數相互比較。如何嚴謹定義區間變數的相關性，是本研究未來的努力目標之一。2)自迴歸項數 p 的縮減。在建模的流程中，我們發現各測站所選擇的最終模型，自迴歸項數 p 的大小均介於15與25之間，顯示AIR模型與HVAIR模型需要龐大的參數，才能達到較好的解釋能力。因此，我們猜測 $PM_{2.5}$ 的週區間濃度可能存在移動平均效應，並期望在關係式(10)與(13)中加入區間式的移動平均項，使配適模型更為精簡。

參考文獻

- [1] Arroyo, J., González-Rivera G, Maté C (2010) Forecasting with interval and histogram data. Some financial applications. *Handbook of empirical economics and finance*, 247–280.
- [2] Billard, L., and Diday, E. (2003). From the Statistics of Data to the Statistics of Knowledge. *Journal of the American Statistical Association*, **98**, 470-487.
- [3] Brito, P. (2014). Symbolic Data Analysis: Another Look at the Interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, **4**, 281-295.
- [4] Blom, G. (1958). *Statistical Estimates and Transformed Beta-Variables*. Wiley.

- [5] Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, **6(1)**, 3–33.
- [6] Cesari, D., DeBenedetto, G. E., Bonasoni, P., Busetto, M., Dinoi, A., Merico, E., Chirizzi, D., Cristofanelli, P., Donateo, A., Grasso, F. M., Marinoni, A., Pennetta, A., and Contini, D. (2018). Seasonal variability of PM_{2.5} and PM₁₀ composition and sources in an urban background site in Southern Italy. *Science of the Total Environment*, **612**, 202–213.
- [7] Emami, F., Masiol, M., and Hopke, P. K. (2018). Air pollution at Rochester, NY: Long-term trends and multivariate analysis of upwind SO₂ source impacts. *Science of the Total Environment*, **612**, 1506–1515.
- [8] Gao, Z., and Tsay, R. S. (2019). A Structural-Factor Approach to Modeling High-Dimensional Time Series and Space-Time Data. *Journal of Time Series Analysis*, **40(3)**, 343-362.
- [9] Lin, W., and González-Rivera, G. (2016). Interval-Valued Time Series Models: Estimation Based on Order Statistics Exploring the Agriculture Marketing Service Data. *Computational Statistics and Data Analysis*, **100**, 694-711.
- [10] Lin, L.-C., Chien, H.-L., and Lee, S. (2021). Symbolic Interval-Valued Data Analysis for Time Series Based on Auto-Interval-Regressive Models. *Statistical Methods & Applications*, **30**, 295-315. <https://doi.org/10.1007/s10260-020-00525-7>.
- [11] Sims, C.A. (1980). Macroeconomics and reality. *Journal of Econometrica*, **48**, 1-48
- [12] Yao, L., Lu, N., and Jiang, S. (2012). Artificial Neural Network (ANN) for Multi-source PM_{2.5} Estimation Using Surface, MODIS, and Meteorological Data. *The 2012 International Conference on Biomedical Engineering and Biotechnology*, IEEE, 1228–1231.
- [13] Zhang, M., and Lin, D. K. J. (2020). Visualization for Interval Data. manuscript.