應用k-mean群集分析的台灣氣候分區實驗

陳翠玲[1]　陳孟詩[2]　沈里音[3]　陳雲蘭[1]
氣象科技研究中心[1]　第三組第二科[2]　第三組第一科[2]
中央氣象局

摘　要

本研究接續104年應用縱向資料K-means群集方法之台灣雨量分區研究的延續。資料已更新延長至1998年至2016年的雨量、溫度資料，並且雨量站的測站數增加，擁有更完整的資料來進行全台氣候分區實作。利用 k-means cluster analysis 演算法，使用月累積雨量、月平均溫度為變數進行單變數與雙變數的氣候分區。本實驗因為新資料的更新，雨量分區的地理分布較舊資料更為顯著，而溫度的差異較小。實驗的結果顯示溫度與雨量的分區都有季節上顯著的不同；其中，雨量有相當顯著的地理區域特徵，溫度則是有顯著的緯度、海拔的差異。另外，本實驗增加相對雨量的降水比值與相對溫度的距平值為分區變數進行實驗，其部分結果也會摘節在本文中。

關鍵字：群聚分析、縱列研究、K-Mean法、EM演算法

一、前言

延續104年台灣雨量分區研究一文[1]，將更新的資料與地理上更完整分布的測站資訊再進行台灣氣候分區實驗。前研究與本文的差異以下列三點敘述：第一、舊資料為未經檢覈的原始資料，在分區前需要事先篩選資料，且其完整度不能低於98%，因為缺漏的資料需要以線性內插取得，而新資料是經由氣象局局內經過檢覈、無缺漏的資料[2、3]。第二、新資料時間長度自1998年至2016年，較舊資料多兩年以上。第三、參考**錯誤! 找不到參照來源。**，測站數中雨量測站數318較舊資料測站數多80站；溫度、雙變數測站分布在新舊資料上則差異不超過5站，因此新資料在空間上與時間上都較完整，以群集分析這樣無監督性的資料挖礦而言，有分析上的優勢。

群集分析視為探索性的資料分析演算，非統計推論工具。如**錯誤! 找不到參照來源。**所示，將在資料上每一個物件進行分群在兩個條件之下：第一，在群集內每一個物件彼此之間的變異量或距離越小越好；同時，第二，群集之間離間的距離越大越好，意同離間分開差異大的群集。群集分析使用的演算法不只一種，本文沿用104年的K-Means方法，隨機給定k個群集中心的條件下，EM演算法將收斂最後k個中心點，並以收斂後的中心點收集與其距離最近的群集成員，來完成群集分析。

實驗資料是同104年文中類似時間序列的縱列資料，如**錯誤! 找不到參照來源。**所示，以台北測站2016年月均溫為例，可以看到在同一個地點上重複測量而產生的軌跡變化；其溫度受季節影響，而產生溫度高低變化，特別是測量時間點如果是從1998年到2016年可能還會有年際等週期現象的特性可供觀察。這樣的資料將拉長的時間外加上地理空間的延展資訊，如**錯誤! 找不到參照來源。**所示，可以使用群集分析，讓相同特性的資料整合，結合已知的氣候特性來檢覈群集的合理性，或檢視有無新的氣候資訊可供分析。

本文規劃如下，下節簡述k-mean法之分析，數學內容已在[1]詳述，在此就不再贅述；第三節簡述資料來源與處理，後整理群集分析的實驗結果，並簡述其分析之結論與未來展望。

二、k-mean法的集群分析

本部分因為已在104相關的研究中[1]以詳述，因此在此只將演算法簡化敘述之。參考圖 4，給定k-Mean的意思是：第一、可以為隨機在資料中選取k個點為資料中心(圖4左1)；第二、進行全資料與k個點的距離運算，最後資料中與三個不同分配的中心點最近的歸為一群(圖四左2)；第三、此一步驟重新計算新分配的資料的中心點(圖四右2)；第四、最後則為在重新計算每筆資料與新中心點的距離再進行重新分配。上面步驟會在迴圈內直到新舊的中心點能夠收斂。此iris[1]資料大約在這些步驟重複13次左右收斂(圖四右1)。K-Mean的EM演算法原理可以參考[4]。

資料與中心的距離運算仍使用歐式距離的定義。參考**錯誤! 找不到參照來源。**，判斷分群的好壞可以以下函式表示

$$C_{opt} = \frac{\sum_{i=1}^{k} D_{b_i}/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} D_{w_j}/(n-k)}$$

。其中，可視為分子D_b為k個分群的變異量和，因此其權重為$k-1$；相對每一群體內的變異量和，在分

[1] Google Iris flower data set.

母的$\sum_{j=1}^{n_i} D_{w_j}$，其權重為$n-k$。C_{opt}最大化時視為最佳分群結果。考慮k-Mean的初始k中心點是隨機選取時，它進行不只一次的k-Mean實驗，一方確定演算法能收斂，並進一步確定最佳的分群結果。

每一個實驗項目都至少進行20次分群，每一次分群皆會有一個C_{opt}值參考，進一步再選擇最佳分群的實驗結果。

三、資料來源與處理

本文延續104年研究[1]，分析資料皆來自氣象局內經由檢覈之資料[2、3]，時間自1998年1月至2016年12月。溫度測站110站，雨量測站則為318，雙變數則是101個測站。資料時間單位為月累積雨量/月平均溫度。

雨量與溫度資料再製成為降雨比/溫度距平值來進行分群。降雨比是以1998年到2016年每一個月的月累積雨量排序取medium值，再以月為單位，月累積雨量時間序列同步除以當月的medium值。溫度距平值也是在每一個月平均溫度取平均值，月平均溫度時間序列同步減去其當月平均值為其取得方法。

四、分區實驗的輸出與結果

本文將時間軸上分為5個季節，另外每一個季節、不同變數進行一個分區實驗。季節分為：春天(SP)2-4月、梅雨(PR)5-6月、颱風(TY)7-9月、秋天(FA)10-11月、冬天(WI)12-1月；變數為：(月累積)雨量(RR)、月平均溫度(TT)、雨量/溫度(RT)；另外有在第三節中討論的降雨比、溫度距平值。因為群集分析變數與時間軸皆可以自由變動，並且因此得出需多不同的分析結果，因此本文只節述部分的內容，其中以季節為時間軸簡述其分區的結果整理如下，其餘的會整理為R程式與空間，並總整為一CWB Shiny App在GitHub上[2]。

在詳述實驗結果前，回顧C_{opt}之定義，它是在每一個實驗皆會產生的判斷值。在進行分區數k=2,…,9的實驗下，每一個k皆會重複實驗20次，每次都會產生一個C_{opt}值，如圖 5所示：以季節為春天、變數為雨量為例，它表示當分區數k=2時研判為最佳化的分區數。如此進行實驗步驟是因為群集分析與k-mean演算法性質的緣故；群集分析是為無監督之統計方法，沒有前置知識介入，意即k為未知；而k-mean則是因為起始的中心點為隨機選取，要確保其數值演算結果收斂並最佳化的手段，因此重複進行實驗。

以C_{opt}的定義為前提，實驗分區的結果整理參考表 1、表 2與表 3。雨量上，不管為全數值或Ratio值，它的最佳化分區數目多為2-3區左右；溫度在全數值時分區數目多可至7-8區甚至更多、Anomaly值則2-3區。雙變數則落在2-4區，這關乎於哪一個變數影響C_{opt}值更劇，這個部分後面會在敘述。接下來則是個變數在各季節中最佳的分區，與其測站之分部。

雨量(RR)為變數，在(R1)春天的最佳分區結果參考圖 5可觀察到以最大化C_{opt}的原則下，最佳分區數是k=2。測站的分布顯示如**錯誤! 找不到參照來源。**，月累積雨量的測站分布是以南北為趨勢，降雨比的分區是以西邊與台灣其他區域；參考**錯誤! 找不到參照來源。**，每一個分區的平均/中心曲線圖，雨量顯示在北邊的1998年至2016年的春季累計降雨量較南邊測站大，降雨比則顯示西部降雨比較其他區域高。

(R2)梅雨在月累積雨量與降雨比的最佳分區數皆為2。參考圖 8與圖 9，兩個不同的雨量值測站分布差異相對的不同，雨量分區在南部與山區測站的月累積降雨較其他地區大；降雨比則是北部測站相較於台灣其他區域的降雨比大。

(R3)颱風季節也都是k=2有最佳的分區。參考圖 10與圖 11，月累積雨量與降雨比在颱風季測站分布差異不大，皆為南部測站的月累積雨量與降雨比較大。南部測站在颱風季會經歷較大的雨量與雨量變化較劇烈。

(R4)秋天在月累積雨量值上k=2與3的C_{opt}值其實差異不大，因此在這裡選取k=3的降雨分布為參考，如圖 12顯示，分為北區、東區與中部南部三個分區為主。圖 13可知秋天降雨北部大於東部，南部降雨較稀少。一樣參考同圖 12圖 13，降雨比分區則分為東西兩個區塊，其中東區的降雨比較西區大。

(R5)冬天在月累積雨量分區數是k=2，降雨比k=3較佳。參考圖 14，月累積雨量只有在北部與其他區域的分布，北部降雨明顯較多於其他區域，如圖 15所表示的；而降雨比則是分為北部、東部與中南部，北部則有較大的降雨比率的趨勢。

以溫度(TT)為變數討論如下。以月平均溫度進行分區的k值都相當的多，且C_{opt}值也偏大；主要是溫度會被區分出來的可以為高度、緯度、溫差的差異而被分列為不同的群體。未陳列所有溫度實驗結果，請參考GitHub上的相關資訊或洽詢作者。

參考圖 16.，以冬天為例，台中到台南一區為溫度第2高的區域，與紅色的東北角台灣西部沿岸區域第3高的區域差異不大，然而第2高的溫度溫差較第3高的區域溫差小。另外參考圖 17.，相對於以春天為冬天實驗的對照組，春天的溫度溫差相對較大，溫度最高的高雄以南測站與第2高的南部測站的差異是因為溫差差異上被區分。

相對於溫度分區，依春天、梅雨、颱風、秋天、冬天季節順序其結果可以參考圖 18至圖 22，溫度距平值的分區數k少很多，除了颱風季外其餘k=2。去

[2] 請接洽第一作者:tlchen@cwb.gov.tw

除高度、緯度，甚至溫差條件下距平值所產生的差異，是讓距平值分區結果分區數顯著的減少的原因。春天、梅雨、秋天的溫度距平值，參考圖 18、圖 19、圖 21，兩區的平均/中心曲線差異不大；颱風與冬天，分別參考圖 20與圖 22，的差異較明顯，另外這兩個季節最佳化的C_{opt}值也較高。

雙變數，在雨量/溫度或是降雨比/溫度距平的例子裡，分區數k至多4；並且C_{opt}值都偏低。當訊息變多，沒有前置知識的無監督方法在區分群體上會比較困難；以圖 23為例，在雙變數(雨量/溫度)最佳的分區數k=3有最大的C_{opt}值，k=3的結果中可以看見有一部分北部測站與南部測站在溫度上比較接近(參考圖 23內的TT圖)，然而在降雨上(參考圖 23內的RR圖)有相當的區隔，同樣的，高山測站在雨量上與北部測站可能差異不大(RR圖)，然而在溫度上(TT圖)的差異卻讓這兩區分隔，這些區隔皆在區與區之間的歐式距離決定，亦即變異量上的差異，這樣是否符合氣候物理的解釋，或者需要更多的前置知識來導引都需要再討論。

另外比較雙變數，雨量/溫度(圖 24)與降雨比/溫度距平值(圖 25)在春天分區上的差異。雨量/溫度在春天最佳分區數k=4，參考圖 24的RR圖可以看出分區上歐式距離所分出來的群集雨量的因素大於溫度；同季節在降雨比/溫度距平值的最佳分區數落在k=3，參考圖 25TT圖上各群集的溫度距平值平均/中心曲線圖較無法看出區別，各區明顯的區別也落在RR圖中，顯示分區因為降雨比的影響較多。

五、結論

單純以數值方法，應用長時間的雨量溫度等氣候變數來區分，不同的季節會有不同的表現。其中，可以觀察出雨量在不同季節上有強烈的地域性，溫度則是被影響因素較多而產生較大的分區數，雙變數會減少分群數主要是因為變數眾多，比較需要外力，如前置知識導引來進行較精確的分區。另外在降雨比/溫度距平值的分析上，還需要更多氣候背景知識的加入，讓分群研究的詮釋能夠更清楚，以擴展更多挖礦統計數值方法研究。

六、參考文獻

1. 陳翠玲、紀雍華、陳孟詩、林沛練, 2015: 應用縱向資料K-Means群集方法之台灣雨量分區研究. 104年天氣分析與預報研討會
2. 陳雲蘭、薛宏宇、呂致穎、陳品妤、詹智雄、沈里音, 2015: 「台灣長期氣候資料整集分析」計畫研究(1)– 自動氣象站長期氣溫觀測值合理性檢測方法探討及分析. 104年天氣分析與預報研討會
3. 陳品妤、陳雲蘭、沈里音, 2015: 「台灣長期氣候資料整集分析」計畫研究(2)– 自動站雨量累計於後記錄值的問題分析及處理
4. A.P. Dempster, N.M. Laird, and D.B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc., 39B, 1-39.

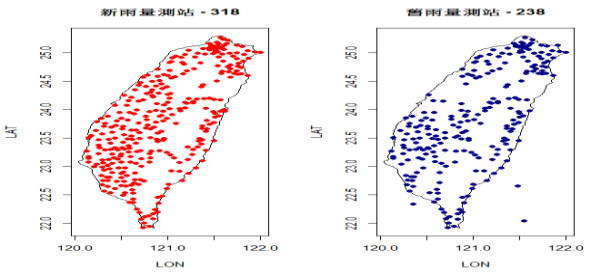

圖 1. 新舊雨量測站分布圖，左邊新資料雨量測站數較右邊舊測站多 80 個測站。

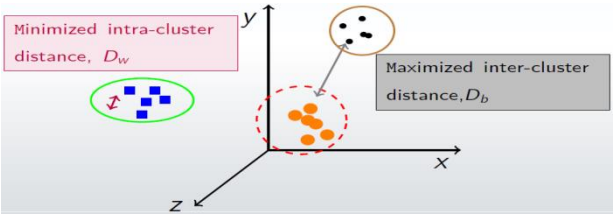

圖 2. 圖示群集分析概念，最小化群內物件彼此的距離(D_w)，同時最大化群間的距離(D_b)。在分群上將彼此相似物件、成分等聚集，同時分開不同性質彼此相異的群集。

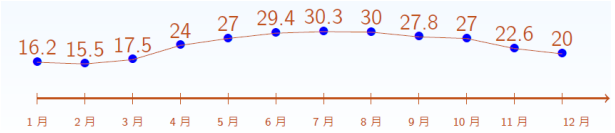

圖 3. 台北2016年月均溫縱列資料圖示，圖中資料顯示台北，例如在5 月，月均溫為27 度。

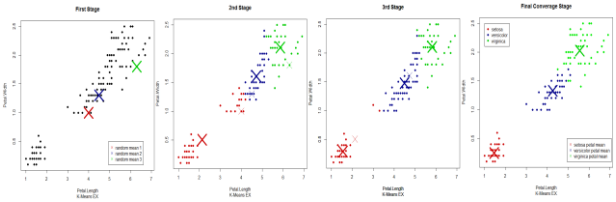

圖 4. K-Means 方法演算過程示意圖。

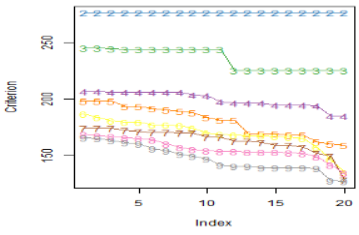

圖 5. 分區實驗C_{opt}圖，表示每一個季節變數進行2-9 個分

區實驗，並每一個實驗皆重複實驗 20 次的結果。這裡以春天雨量的分區為例，分為 2 區有最佳的C_{opt}值，其次為 3、4 等分區數。(每一個曲線上皆有顯示不同的分區數 k 的C_{opt}值。)

雨量	分區數	C_{opt}
SP	2	275
PR	2	176
TY	2	141
FA	3	449
WI	2	435

降雨比	分區數	C_{opt}
SP	2	171
PR	2	431
TY	2	364
FA	2	426
WI	3	170

表 1. 分季並以雨量為變數的分區實驗結果。

溫度	分區數	C_{opt}
SP	7	525
PR	8	550
TY	7	590
FA	8	600
WI	7	540

距平值	分區數	C_{opt}
SP	2	154
PR	2	103
TY	3	258
FA	2	35
WI	2	200

表 2. 分季並以溫度為變數的分區實驗結果。

雨量/溫度	分區數	C_{opt}
SP	4	83
PR	2	86
TY	2	100
FA	3	97
WI	3	110

降雨比/距平值	分區數	C_{opt}
SP	3	49
PR	2	91
TY	2	99
FA	2	90
WI	3	65

表 3. 分季並以雨量/溫度、降雨比/溫度距平值為多變數的分區實驗結果。

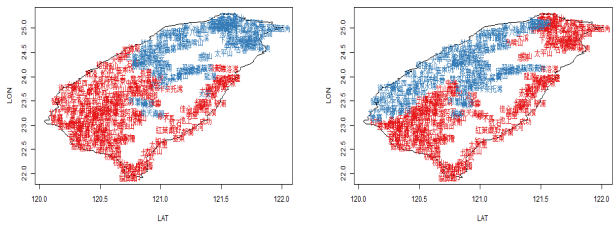

圖 6. 左為雨量，右為降雨比，季節為春天的分區數在 k=2 擁有最佳分區結果的測站分布。

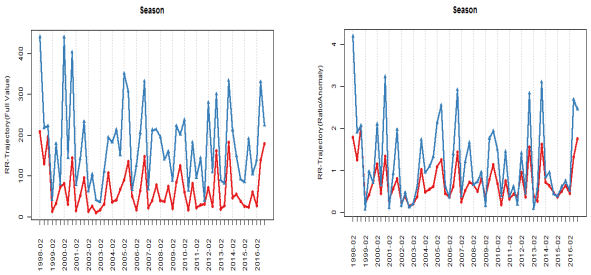

圖 7. 對應**錯誤! 找不到參照來源。**中 k=2 兩區對應的平均/中心曲線。

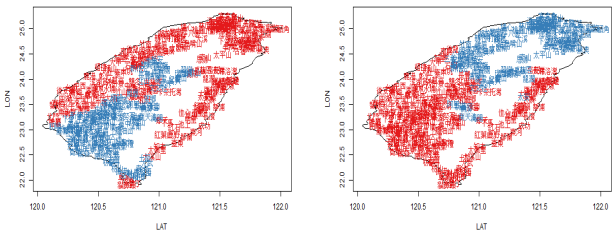

圖 8. 左為雨量，右為降雨比，季節為梅雨的分區數在 k=2 擁有最佳分區結果的測站分布。

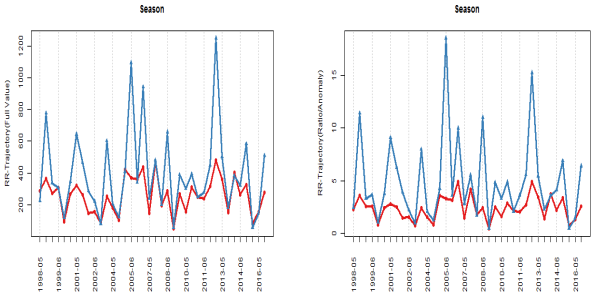

圖 9. 對應圖 8 中 k=2 兩區的平均/中心曲線。

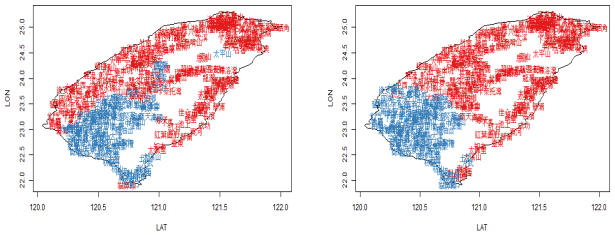

圖 10. 左為雨量，右為降雨比，季節為颱風的分區數在 k=2 擁有最佳分區結果的測站分布。

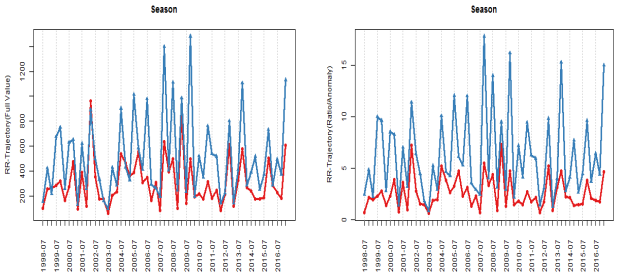

圖 11. 對應圖 10 中 k=2 兩區的平均/中心曲線。

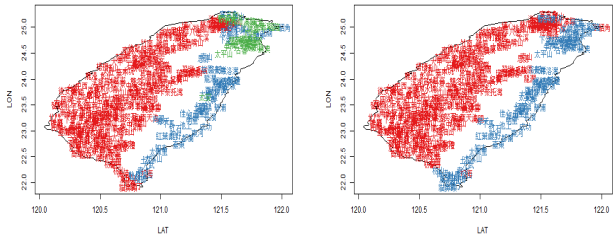

圖 12. 左為雨量，右為降雨比，季節為秋天的分區數分別在 k=3、k=2 擁有最佳分區結果的測站分布。

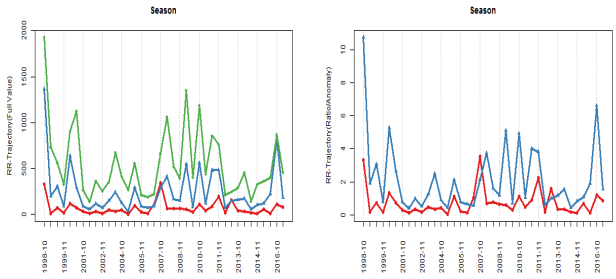

圖 13. 對應圖 12 中分別 k=3 與 2 的平均/中心曲線。

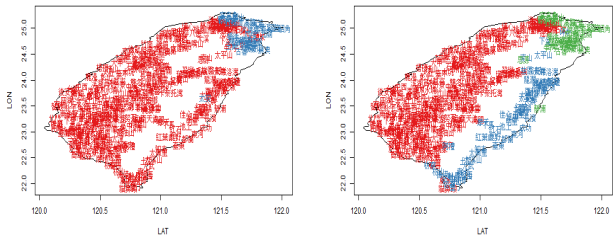

圖 14. 左為雨量，右為降雨比，季節為冬天的分區數分別在 k=2、k=3 擁有最佳分區結果的測站分布。

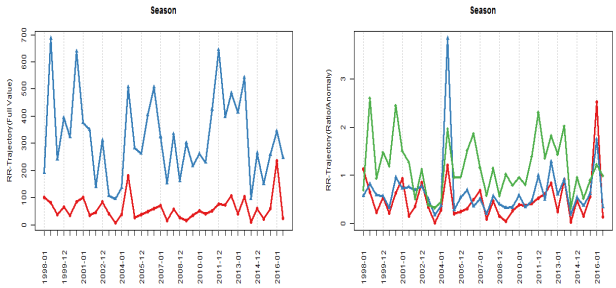

圖 15. 對應圖 14 中 k=2 與 3 的平均/中心曲線。

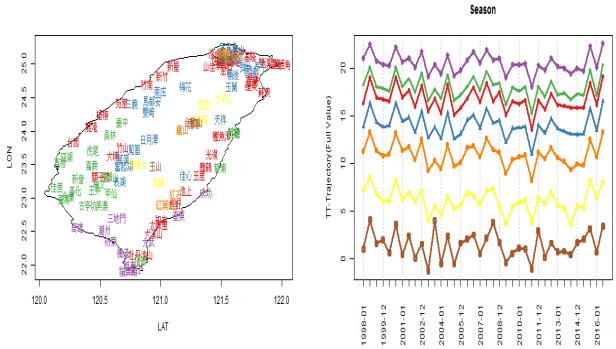

圖 16. 左圖為測站在冬天溫度的分區數 k=7 的最佳分區結果，右圖為對應左圖測站不同分區的平均/中心曲線。

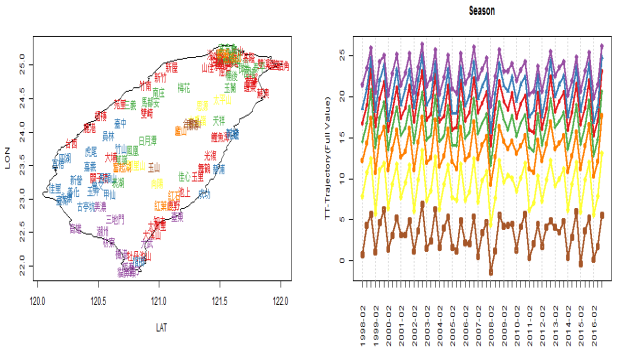

圖 17. 左圖為測站在春天溫度的分區數 k=7 的最佳分區結果，右圖為對應左圖測站不同分區的平均/中心曲線。

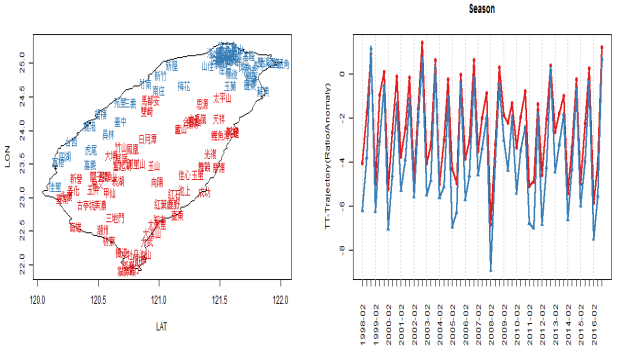

圖 18. 左圖為測站在春天溫度距平值的分區數 k=2 的最佳分區結果，右圖為對應左圖測站不同分區的平均/中心曲線。

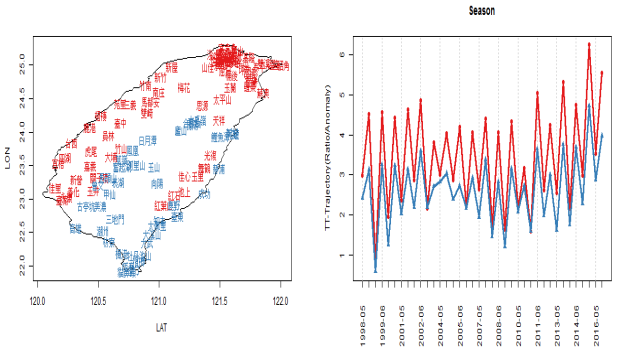

圖 19. 左圖為測站在梅雨溫度距平值的分區數 k=2 的最佳分區結果，右圖為對應左圖測站不同分區的平均/中心曲線。

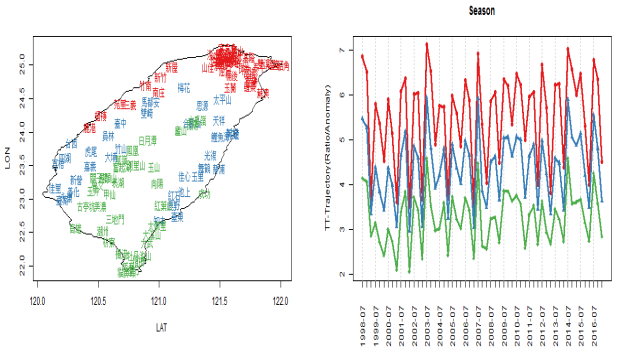

圖 20. 左圖為測站在颱風溫度距平值的分區數 k=3 的最佳

分區結果，右圖為對應左圖測站不同分區的平均/中心曲線。

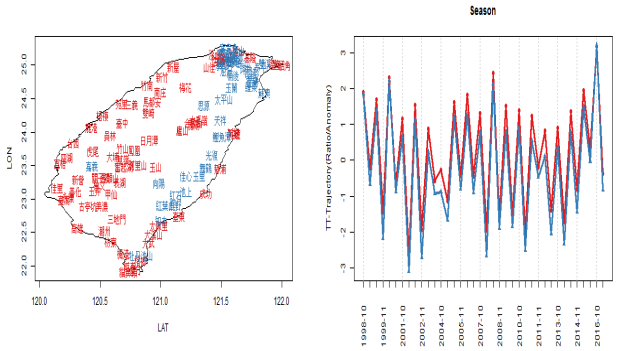

圖 21. 左圖為測站在秋天溫度距平值的分區數 k=2 的最佳分區結果，右圖為對應左圖測站不同分區的平均/中心曲線。

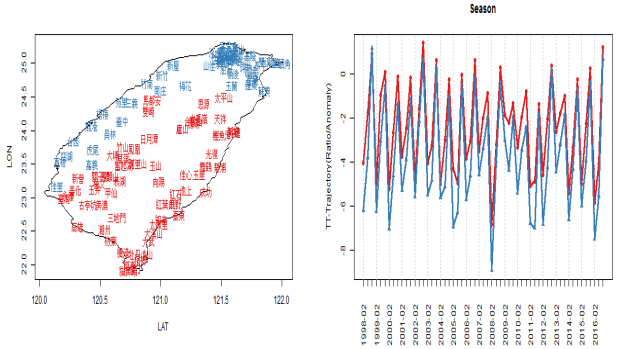

圖 22. 左圖為測站在冬天溫度距平值的分區數 k=2 的最佳分區結果，右圖為對應左圖測站不同分區的平均/中心曲線。

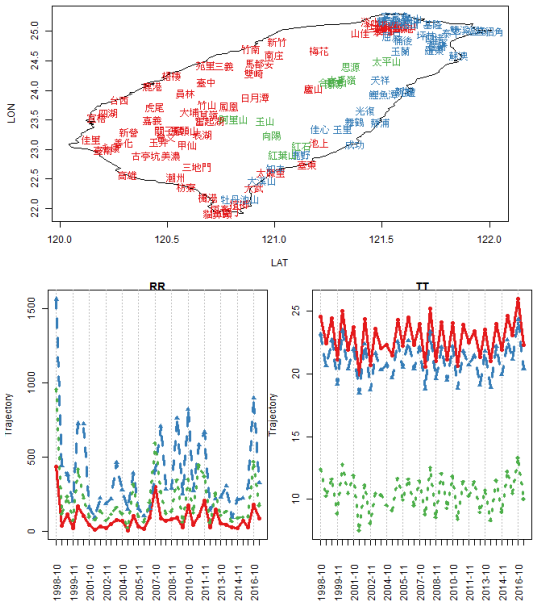

圖 23. 雨量/溫度雙變數在秋天的最佳分區結果落在 k=3，其測站的分布，以及對應之變數雨量(RR 圖)、溫度(TT 圖) 平均/中心曲線圖。

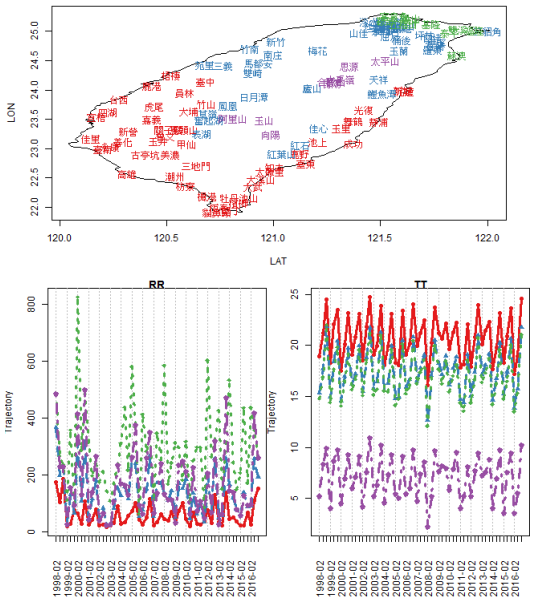

圖 24. 雨量/溫度雙變數在春天的最佳分區結果落在 k=4，其測站的分布，以及對應之變數雨量(RR 圖)、溫度(TT 圖) 平均/中心曲線圖。

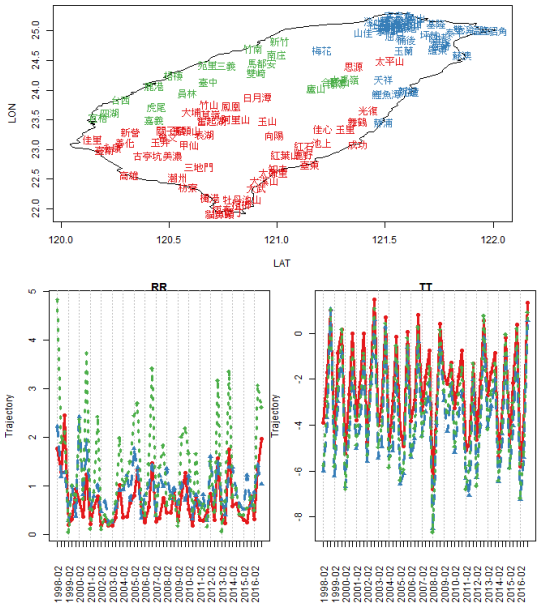

圖 25. 降雨比/溫度距平值雙變數在春天的最佳分區結果落在 k=3，其測站的分布，以及對應之變數雨量(RR 圖)、溫度(TT 圖) 平均/中心曲線圖。