應用縱向資料K-means群集方法之臺灣雨量分區研究

陳翠玲[1,4]　紀雍華[2]　陳孟詩[3]　林沛練[4]

中央氣象局科技中心[1]　中央氣象局預報中心[2]　中央氣象局第三組第二科[3]　中央大學大氣科學系[4]

摘　　要

本文是中央氣象局104年度「災害性天氣監測與預報作業建置計畫」的一部分。台灣的氣候分區是採用中央氣象局的分區法，使用的方法多數以氣候理論為基礎，統計為輔的方式進行分析，較少使用純數值統計方法的研究。因此收集客觀的數值資料，並以適當的統計方法進行分析是可行的。特別在近十年來，氣象局增加了測站的設置，可以幫助進行類挖礦的氣象研究，其中群聚分析就是此一方向的研究方法。因此受益於豐富資料與更多高階的統計方法，考慮氣候資料的時間序列特性，在此使用適用於縱列資料特性的K-means群聚分析來做台灣氣候分區的分析，並評估方法的可行性。

在此，採用了2000年至2013年臺灣超過百個測站的雨量與溫度的時序觀測資料來進行臺灣氣候分區實驗。另用縱列資料特性來使用K-means方法，進行臺灣的雨量與溫度分區，大略結果可以區分：以臺灣的溫度區分多受到地形影響，而雨量則明顯的受到季節影響。此計畫中，使用的K-Means法有機會在未來更精緻的氣候分區設置上成為基礎。

關鍵字：群聚分析(Cluster Analysis)、縱列研究(Longitudinal Study)、K-Means法、EM演算法

一、前言

目前中央氣象局分區法，將臺灣以中央山脈微分界，大略為四區：基隆市、大臺北地區至新竹為北區；大台中區域至嘉義為中部地區;將台南往南延伸至恆春為南部地區;以及從宜蘭、花蓮與台東為東部地區－的分類法。這樣的氣候以地理、經度（中央山脈）與緯度的分區是有其氣候理論基礎的[2, 3]。

臺灣氣候分區的研究已經有相當的歷史，蔣丙然利用柯本氣候分類法[3]將臺灣分為6區、陳正祥則是按照桑士偉氣候分類法將臺灣分為八大區域[4]。另外結合統計方法來進行，如郭文鑠使用測站逐月降水量與逐年的平均氣溫相關係數比較法進行分區[5]、吳明進與陳幼麟使用主成分分析，再利用群聚實驗分析做的氣候分區[6]；較近期邱祈榮等結合了GIS與崔瓦氣候分類系統[7]，使用氣象局測站的溫度與雨量資料，並應用統計迴歸與克利金方法分別對溫度與雨量的資料來進行網格化的全台分區實驗方法。

過去的臺灣氣候分區研究，多屬於氣象水文專門為主，少直接使用資料數據的統計分析。本研究主要重點在空間與時間上：氣象局在近十年設置多達五百至六百的人工、自動測站，有充足的測站分布全臺，另外則是分析的變數：雨量、溫度皆為時間序列資料，因此可直接使用時間序列資料的群集分析。

有大量的資料條件下，雨量與溫度如果使用數值平均則太粗略，在統計數學理論發展下，可使用高維度資料軌跡（Trajectory）方式進行分析，這就是縱向資料研究或列隊研究。縱向資料研究在醫學、社會科學與精算領域中使用頻率相當高，主要因為科技進步下能擷取的資料型態維度增加而產生的。此研究方法有幾項優點，首先它不會有校準的問題，可以做特殊時間上因素測量；第二縱向資料可看為重覆測量性質，可以減輕特殊狀況發生的錯誤性偏估。

本研究使用純統計方法的資料分析，並使用適合的統計方法自行「說話」的方式來進行分區實驗。在此使用了2000至2013年在臺灣多處的雨量與溫度測站資料，使用純統計的k-means群聚分析來進行臺灣的分區實驗。

二、K-Means法的集群分析

集群分析的方法相當的多樣性：包括了層階集群分析(hierarchical clustering)、光譜集群分析(spectral clustering)、k-means群集分析到貝式(Bayesian)群集分析法等[10]。在此選擇k-means原因是(1)不需要群集的初始常態參

數設定的假設。(2) k-means在數值收斂方法上相對穩定[9]。它歸屬於EM演算法(Expectation-Maximization)[8] --將每一個觀察質分派至最接近其已決定的群集中心的群集中。在E階段與M階段交錯中，最佳化的群集會在觀測值被分派的群集越來越固定中決定。

現在假設有興趣的變數有 n 個測站，假設變數為 Y ，時間軸常度為 t 。 i 測站在時間點 l 的 Y 值為 y_{il} ；對 i 測站的雨量軌跡為 $y_i = (y_{i1}, y_{i2}, \ldots, y_{it})$ 。在此目標是將資料集切割成 k 個內部同性質的群。其中k-means在將每一個觀察值或單測站被測量的雨量軌跡， y_i ，分派至『距離最近』的群集中。在這裡距離的定義取歐氏距離(Euclidean distance): $d^E(y_i - y_j) = \sqrt{\sum_{l=1}^{t}(y_{il}-y_{jl})^2}$ 。

k-means方法的問題為群集數需要主觀的認定，處置方法為在初始條件下，計算不同的分區 k 結果，再以客觀準則來決定最佳群集數。以下定義最佳群集標準，假設 n_m 為在 m 群集的(雨量或溫度)軌跡數， $\bar{y}_m$ 為群集 m 的均軌跡(mean trajectory)， $\bar{y}$ 則為資料集的均軌跡。令 v' 向量為 v 向量的轉置，則群集間共異變數矩陣為 $B = \sum_{m=1}^{k} n_m(\bar{y}_m - \bar{y})(\bar{y}_m - \bar{y})'$ 。而 $tr(B)$ 代表 B 矩陣對角系數總和。另外群集內的共變異數矩陣為 $W = \sum_{m=1}^{k}\sum_{l=1}^{n_m}(y_{ml} - \bar{y}_m)(y_{ml} - \bar{y}_m)'$ 。令 $tr(W)$ 為 W 矩陣對角系數總和。從 B 與 W 的定義來看，當 $tr(B)$ 值高代表群集間的差異大，當 $tr(W)$ 值小代表群集內同質性高。因此最佳化的 k 群集標準 $c(k)$ 值(Calinski and Harabasz criterion)定義[13、14]為

$$c(k) = \frac{trace(B)}{trace(W)} \cdot \frac{n-k}{k-1}$$

由 $tr(B)$ 與 $tr(W)$ 來看最佳化的 k 群集為最大化的 $c(k)$ 。另外在選取 k 值時，另外計算 $k-1$ 與 k 的群集間變異量增減的百分比為考量群集數 k 參考。

k-means演算流程使用R進行演算[15]，其中參數初始化進行資料準備，隨機的選擇 k 個中心則進入第一個迴圈i重覆執行EM演算法直到交替的期態值群集中心跟最大化概似函數的結果收斂；第二的k迴圈則是每一個初始條件下，進行不同k分區數的群集分析。

三、資料來源與處理

這裡分析的資料源自氣象局的人工、自動測站，時間自2000年1月至2013年12月。溫度測站共114個，雨量測站共238個的觀測資料分別進行臺灣氣候分區實驗。測站分布可參考圖 2、圖 11等。資料的時間單位為月累積雨量與月平均溫度，長度為168筆。另外也進行了雙變數的分區實驗，而雙測站數為99。資料遺缺沒有進行外插，而是使用Gower矯正[11]的歐氏距離 $d^{E,G}$ ，假設雨量軌跡長度t的 y_i 與 y_j ，假設 y_{il} 與 y_{jl} 無資料則命令 $w_{ijl}=0$ ，否則 $w_{ijl}=1$ 。則

$$d^{E,G}(y_i - y_j) = \sqrt{\frac{t}{\sum_{l=1}^{t} w_{ijl}} \sum_{l=1}^{t} w_{ijl} \cdot (y_{il} - y_{jl})^2}$$

。雖然有方法可以處理資料遺缺，然而在有充沛的測站條件下，所有資料皆有98%資料完整性再佐以Gower矯正的歐氏距離。

四、分區實驗的輸出與結果

每一個實驗輸出皆有最佳化 k 群集標準 $c(k)$ 值(如圖 1、圖 5與圖 4)、分區測站分布圖(如圖 2、圖 11)以及分群平均軌跡曲線圖(如圖 3)。參考圖 1標準 $c(k)$ 值圖，每一個曲線上數字 k 代表每一次進行k-means分區都重覆進行20次，而每一次的 k 分群就有相對應的測站分群分布圖與分群均軌跡曲線圖。然而因為版面空間的限制，無法討論每一個條件下所有可能的分群結果。所以每一個 k 輸出結果將限制在 $c(k)$ 值最高的那一個分群。對實驗更多的結果有興趣的可以參考[1]。以下為詮釋分區實驗結果與討論。

(一)一月累積雨量分區實驗結果

在圖 1中， $k=5$ 為(2000-2013年的)一月累積雨量(資料長度14筆)分區實驗的最佳值， $k=6$ 為第二佳。考慮一月累積雨量分區可以分為5、6區，因為參考圖 4說明 $k=6$ 較 $k=5$ 的群間變異量減少大約4%， $k=6$ 並沒有減少 $k=5$ 群間的變異量太多。另外，一月累積雨量最佳的 $c(k=5)$ 值測站的分布如圖 2所表示，五個分區大部分集中於北緯24度以南的測站，除了部分顯示為綠色在高山的測站外。另外四個分區則集中在中北部臺灣測站，特別是東北角與宜蘭測站。參考對應圖 2的均軌跡圖 3，解釋降雨的分布，紅色就是北緯24度以南測站是一月累積降雨最少的區域，其次是黃綠色分布在桃竹苗與部分臺北的測站，東北角測站與宜蘭與北部山區測站則是降雨最多。這裡清楚顯示出一月的東北季風所帶來的北部降雨與南部的少雨的特徵。

(二)月累積雨量臺灣分區實驗

在選擇詮釋雨量分區的便利性考量，並未將雨量資料進行標準化處理。分區實驗結果如圖 5到圖 9所表示如下。

雨量的資料並不如溫度資料一致化，因此在分區上其$c(k)$值較低，這在接下來的溫度分區中會討論。雖然最好的分區數是k=3，在此仍然可以參考圖 9尋找$k>3$的可能性。其中所顯示出來的趨勢因為颱風的緣故使得代表綠色的B區單獨區分出來。圖 7的k=6則是花東測站被單獨區分(參考圖 9此時變異量較k=5少8%)；而圖 8的k=7則是多出北緯24度以南至恆春半島測站(變異量較k=6少3%)都是可以參考的可能分區提案。

(三)月平均溫度臺灣分區實驗

實驗結果如圖 10到圖 12所示，這裡使用BoxPlot而不是軌跡圖顯示是為了分區的詮釋與視覺上的便利。如圖 11溫度的分區無疑有相當大受到高度的影響，因此可以在同一個行政縣市中有不只一個分區的可能。另外，溫度也受緯度影響。從圖 12可看出-高雄以南(青色E區)、北部測站(紅色A區)、北緯23-24度測站(黃綠色C區)都是溫度較高的區，但溫差順序A>C>E-的特徵讓這樣的分區顯現出來。其次是北部山區測站(黃色B區)、東部部分測站(綠色D區)溫度較低，其餘則因高度分布不同的三個F、G、H、I區。

另外溫度分月的分區實驗，如圖 13所表示的一、五、七、十一月的結果，顯示分月臺灣溫度區域也會有變化，另外每一個月最佳化的$c(k)$值也都至少在k=8以上[1]。

(四)雙變數(溫度/雨量)臺灣分區實驗

雙變數的分析加入了重覆測量(Repeated Measurement)模式[12]，因為每一個測站不只是重覆測量多變數，因此較前單變數的分區有不一樣特性，可參考[1]。這裡資料仍然是資料遺缺標準仍不超過2%，考量標準化處理讓詮釋困難化，因此並未進行標準化處理。其中雙變數的進行使得$c(k)$值比雨量更低，可參考圖 14與圖 5的比較。在此k=2有最高的$c(k)$值，但是如參考k=3、4、5彼此之間的變異量從7%、11%到4%。

當k=2，如圖 15所示，這兩區的趨勢接近紅色A區溫度、雨量皆大於青藍色的B區。如果考慮圖 16的k=3，則多出在東北角宜蘭測站的區域，除了在k=2的藍色的區域一樣的特徵外，紅色與綠色的區域差別在:綠色的區域溫差與降雨量皆較紅色區域大。

五、結論與建議

溫度k-means分區實驗結果普遍能得到的分區數都較雨量多，溫度分區數多大於8區以上；雨量則較少大約3到6區左右；雙變數則更少大約3到4區，有可能是雨量資料影響緣故。群集分析是無導引式的機器學習，分區的數值演算過程中完全沒有經、緯度為變數的情況下，自然的雨量在分區上產生強烈的地域性，是值得觀察的特徵。雙變數的分區可能需要有更多的資訊加入。

吾人認為因為詮釋結果的需要，標準化並非必要。然而在月份也會影響變數分區結果的情形下，從純粹統計方法上考量更需要的是加入能有輔助資料的機器學習，例如貝氏的群集分析法是建議的研究方向。

六、參考文獻

1. 交通部中央氣象局，2015:「104年度災害性天氣監測與預報作業建置計畫-氣候模式及應用作業委託辦理發展專案工作項目4[1A]」

2.陳國彥，1987：”溫濕圖與柯本氣候分類。”師大學報，29：517-536。

3.蔣丙然，1954：”臺灣氣候誌。”臺灣研究叢刊，26。

4.陳正祥，1957：”氣候之分類與分區。”中央氣象局，氣象學報，第3卷第2期1-9。

5.郭文鑠，1980：”臺灣農業氣候區域規劃。”中央氣象局，氣象學報，第27卷第1期16-28。

6.吳明進、陳幼麟，1993：”臺灣的氣候分區。”大氣科學，第21期第1號，55-66。

7. 邱祈榮、梁玉琦、賴彥任、黃名媛，2004:”臺灣農業氣候區域規劃。” 臺灣地理資訊學刊第一期:41-62。

8. Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: “Maximum likelihood from incomplete data via the EM algorithm”. J. Roy. Stat. Soc., 39B, 1–39

9.J.A. Hartigan (1975). Clustering algorithms. John Wiley & Sons, Inc.

10.E.W. Forgy (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics 21: 768–769.

11.Hunt.D., Jorgensen. M., 2003: Mixture model clustering for mixed data with missing information, Computational Statistics and Data Analysis 41, 429-440.

12.Littell,R.C.,2007:"Repeated measures analysis with clustered subjects." Stat and Data Analy. SAS Global Forum.
13.Shim,Y., J Chung, I-C Choi, 2005: "Comparision study of cluster validity indices using a nonhierarchical clustering algorithm."IEEE comp. soc. pp199-204
14.Calinski, T., J.Harabasz, 1974:"A dendrite method for cluster analysis."Communication in Statistics 3, 1-27.
15.Cenolini, C., B. Falissard, 2011:"kml: a package to cluster longitudinal data."Comp. Met. & Prog. Biom., v104, 3., 112-121.

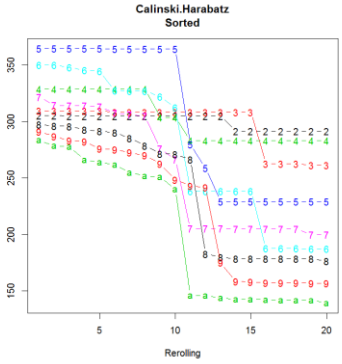

圖 1 一月 k= 2,…,10 累積雨量分區實驗結果，最佳化群集 $c(k)$ 值圖。 $c(k)$ 曲線已排序，曲線上數字 k 代表分區數。

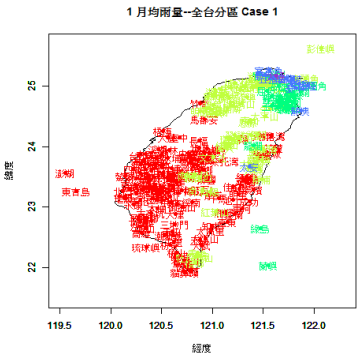

圖 2 一月雨量分群 k= 5 實驗，雨量測站根據經緯度顯示，其中顯示的紅色、綠色等代表不同的分群分區。圖中顯示最大的分區在北緯 24 度以南。

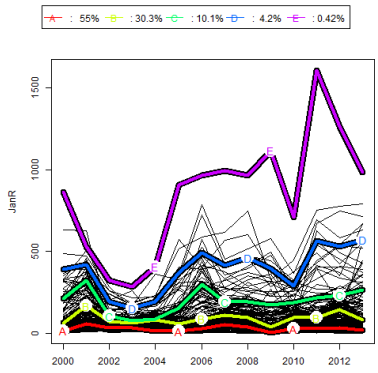

圖 3 一月雨量分群 k= 5 實驗的分群平均軌跡圖，曲線的顏色對應圖 2 的測站分布顏色；例如紅色曲線對應圖 2 的紅色測站區域的一月平均累積雨量。

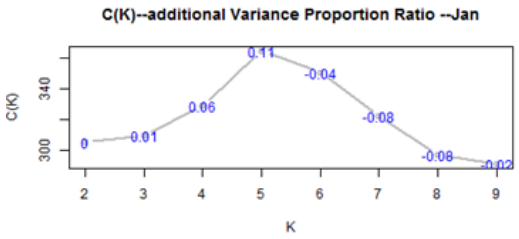

圖 4 對應圖 1 每一個 k 最佳的 $c(k)$ 值，曲線上的藍色字體為每當 k 增加 1，增加/減少的解釋變異量百分比。其中 k= 5 較 k= 4 增加了解釋變異量的 11%。

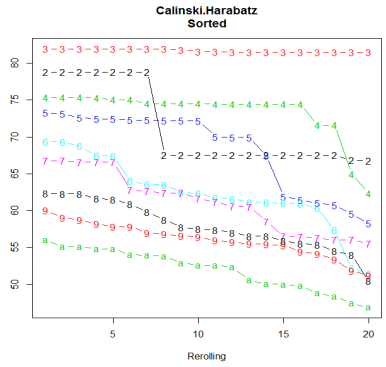

圖 5 k= 2,…,10 月累積雨量分區實驗結果，最佳化群集 $c(k)$ 值圖；其餘參考圖 1。

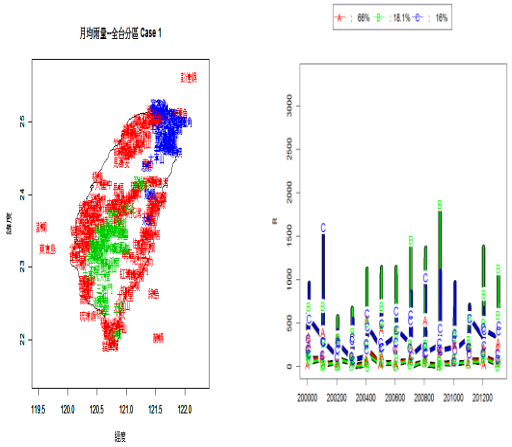

圖 6 左圖為最佳 $c(k)$ 值的雨量分群 k= 3 實驗測站分布結果；顏色與測站意義可參考圖 2。右圖為對應左圖的均累積雨量軌跡圖。

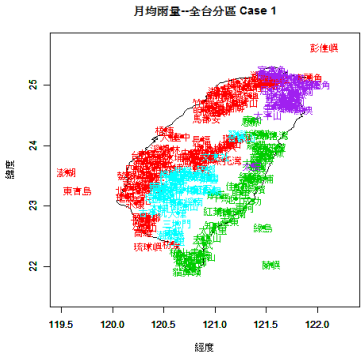

圖 7 最佳 $c(k)$ 值的雨量分群 k= 4 實驗測站分布結果；顏色與測站意義可參考圖 2。

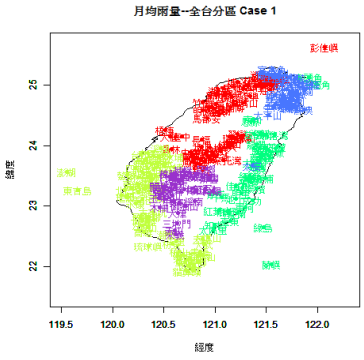

圖 8 最佳 $c(k)$ 值的雨量分群 k= 5 實驗測站分布結果；顏色與測站意義可參考圖 2。

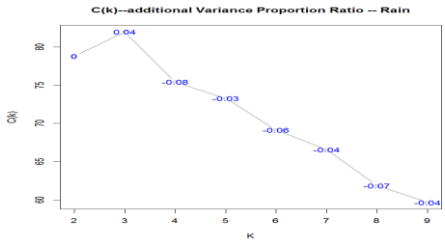

圖 9 對應圖 5 的每一個 k 最佳化的 $c(k)$ 值，曲線上的藍色字體為每當 k 增加 1，增加/減少的解釋變異量百分比。

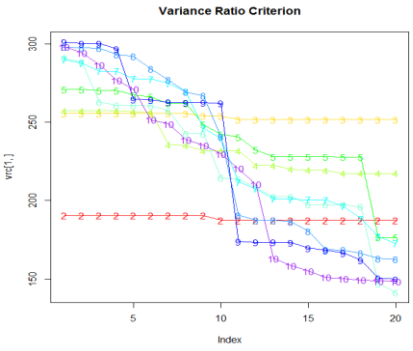

圖 10 k= 2,…,10 月平均溫度分區實驗結果，最佳化群集 $c(k)$ 值圖。

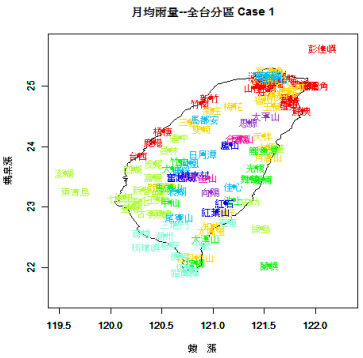

圖 11 最佳 $c(k)$ 值的雨量分群 k= 9 實驗測站分布結果。

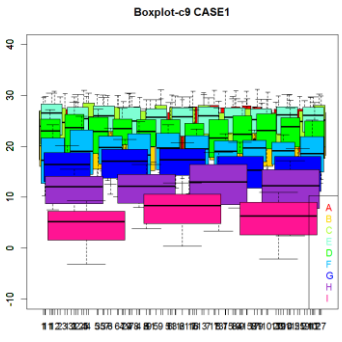

圖 12 對應圖 11 的測站分群 Box 圖，是平均溫度曲線外

另一個參考分群的指標。其中代表紅色的 A 區與代表青色的 C 區高溫差不多，然而被分開來的主要因素是 C 區溫差較 A 區小。其餘 F、G、H、I 區都是高度較高的測站。

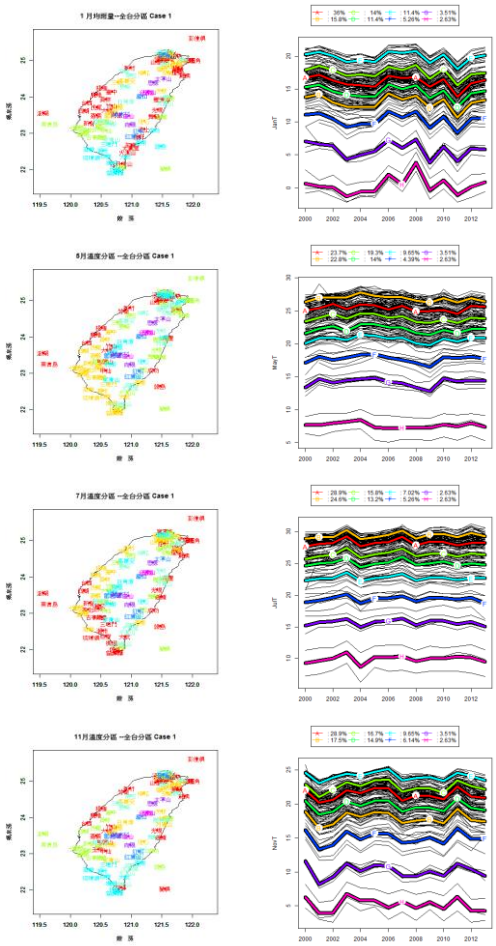

圖 13 月平均溫度分區實驗結果，由上到下分別為一月、五月、七月、十一月。左欄為最佳 $c(k)$ 值的雨量分群 $k=8$ 實驗測站分布結果。右欄為對應左欄的月平均溫度均軌跡。

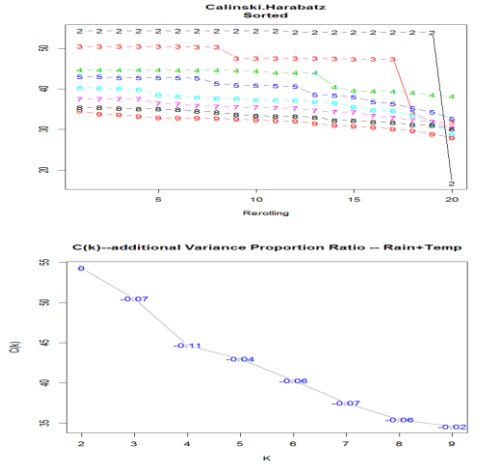

圖 14(上圖) $k=2,\ldots,10$ 雙變數分區實驗結果，最佳化群集的 $c(k)$ 值圖。(下圖)對應上圖的每一個 k 最佳化的 $c(k)$ 值，同圖 4、圖 9。

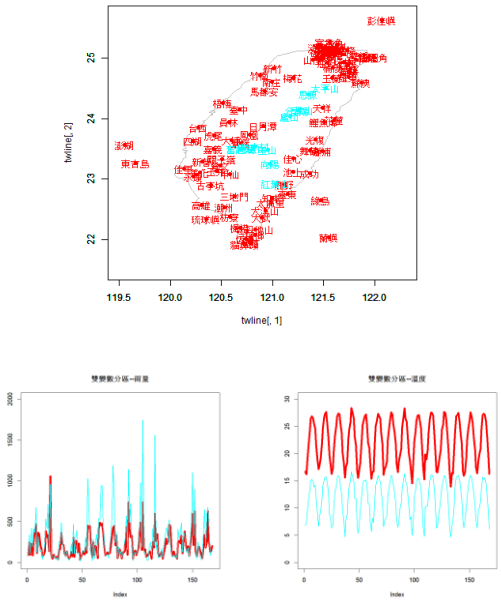

圖 15(上圖) 最佳 $c(k)$ 值的雨量分群 $k=2$ 實驗測站分布結果。(下圖)為對應上圖的雙變數均軌跡，左為平均累積雨量，右為平均平均溫度軌跡。

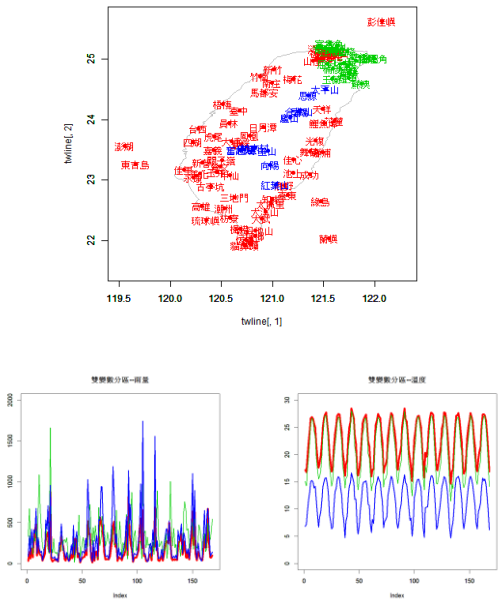

圖 16 雙變數分區 $k=3$ 的結果，其它參考圖 15 說明。