應用主成份迴歸分析方法發展長期預報

陳雲蘭
氣象預報中心
中央氣象局

摘　　要

爲改進月長期統計預報技術，本研究改變以往月長期統計預報中預報因子的資料型態，引用主成份分析法（Principal Component Analysis）將預報因子場的格點資料轉換成代表大範圍空間特性並以變異數大小排序的主成份，再利用各主成份之間相互正交的特性，建立線性迴歸預報模式。

預報模式建立的過程使用交錯驗證法（cross-validation），即依次抽取任一資料年作爲預報目標，並使用其餘的資料來發展模式。再利用所有的預報結果，檢視其預報成效。評估方式主要依據現行長期預報作業三分法，計算預報命中率及技術得分，另外比較預報值及實際值的相關程度及模式在定量方面的誤差情形。透過高命中率及正值的得分，挑選最佳的預報因子組合。

主成份迴歸預報法相對於傳統格點資料統計迴歸法是一個顯著的改變。針對長期預報而言，希望掌握的是大尺度的訊息，因此將格點資料轉換成含有大範圍資料訊息的主成份，再來建立迴歸方程以應用在長期統計預報作業，理論上是可以有所助益的。本研究發展並測試主成份迴歸法的成效後，有以下的結論：

（1）主成份迴歸法具有大量濃縮資料的功能，對經常需處理大量資料的長期統計預報作業相當有幫助。

（2）由於各主成份之間相互獨立，避免了預報因子之間關連性的問題，最後的預報值只需將各主成份對預報元的貢獻作線性相加即可。這種簡單線性組合的特性可便利任意挑選主成份因子，也提昇了大量測試的效率。

（3）透過主成份迴歸法的過程可幫助分析診斷的工作。模式計算出的係數分布圖提供影響預報元的預報因子在空間分布的情形，使我們對模式中掌握的預報訊息有清楚的圖象概念，這也可改善以往無法對傳統模式產品解釋的問題。

對長期預報作業而言，高頻的訊息較難對預報有正的貢獻，主成份迴歸法利用主成份分析，去掉大量的雜訊，主要是希望掌握大尺度的訊息，從本研究的測試證實主成份迴歸方法確實可應用至長期預報作業上。

一、前言

目前月以上的長期預報在全世界的各個作業單位都仍是極具挑戰性的工作，要有很好的掌握並不容易。近年來以動力理論發展模式的系集預報（Ensemble Forecast）產品已逐漸被寄與厚望來改善長期預報，但是其發展仍未臻成熟，成效尚未能有明顯的突破，且短期內可能的幫助僅限於6~14天的預測，對於月以上的趨勢掌握仍屬有限，因此月預報方法的研發仍然以統計方法爲主軸。

統計預報結果的好壞取決於資料訊息與統計預報方法。爲改進統計模式以有效截取可用預報訊息，本研究改變以往使用預報因子場格點資料建立迴歸方程的傳統方式，引用主成份迴歸法改採具備大範圍空間代表性的主成份因子來發展統計模式。這是一種改變資料處理的方式，希望藉由此種資料的前置處理，有效掌握隱含於資料中的訊息，以期能對月長期預報成效有所突破。除了在對預報因子的資料型態做重大突破外，由於採用主成份因子，各預報因子之間相互獨立，因此在建立迴歸模式時可直接將各因子對預報元的貢獻直接相加，免除逐步迴歸的問題。另外，統計模式發展期間往往需要大量的測試，此正交特性便利於預報因子間的任意組合，可有效幫助測試工作的進行。

主成份迴歸預報在資料應用型態及模式方法上較以往而言，均爲一重大改變，其理論基礎相當完整並已有國外作業單位及學者進行研究應用（陳，1993, Vautard,1996，Yu,1997），本局在過去二年

也加以應用並提供短期降水機率預報作業的參考（林，1994），此時應用到重視大尺度訊息的長期預報，應更能符合此統計方法的特性。

二、使用資料與研究方法

（一）使用資料

本研究使用美國國家環境預報中心1958-1996年500百帕月平均場重新分析資料作爲預報因子來預測台北站月溫度與累積雨量的情形。進行分析之前先將資料作標準化處理並除去長期趨勢變化。測試後並可證實此種處理有助於突顯有效的訊息。在測試資料的時間長度方面，爲配合延遲相關的計算，以預報元而言，取1960-1996共37年個案進行預報測試。

（二）研究方法

爲瞭解所選取預報因子場對預報元可提供的訊息程度，首先製作500百帕高度場與台北站各月溫度與雨量的各延遲相關圖（未附圖）。相關圖的分析可幫助選取合適的預報因子場落後月份（或季節平均）及資料的範圍。在選定較合適的預報資料後，以主成份迴歸法來建立模式（說明見第三節）並引用交錯驗證法（Cross Validation）來檢視其預報成效。該作法是依次抽取任一資料年作爲預報目標，並使用其餘的資料來發展模式，再利用所有的預報結果評估預報成效。如此比傳統作法中建立單一模式尋找最佳預報值的評估方式更爲嚴格，因爲在一組資料中建立一套符合原始資料分布的方程，不能保證其在發展資料外的預測能力。也就是說以交錯驗證法來檢視預報的成效可比較接近真實的技術（True Skill）（Michaelsen，1987）。

預報的技術評估主要配合目前本局月長期展望的作業方式，有關月溫度與累積雨量的預報是以分類等級來發布，即由累年的觀測值依大小作排序，並取個數30%、 40%、 30%的範圍，分別定義爲偏高（或偏多，＋）、正常（O）、偏低（或偏少，－）三個等級。因此本研究的主要檢驗方式是將模式預報值換算爲三分法的預測再製作列聯表，分別計算命中率（NSK）與命中技術得分（TSK）。（戚，1978）。另外由交錯驗證過程所得的預報值與實際值的相關程度（Cor）可觀察模式對實際變化趨勢掌握的程度。爲瞭解模式在定量預報方面的能力，也以白氏得分（Brier,1950）方法計算定量技術得分（陳，1996）。

三、主成份迴歸法在長期預報上的應用

（一）主成份分析方法簡介

主成份分析（Principal Component Analysis），簡單的說，是將可能有相互關係存在的多個變數，改成由所有變數組成的線性組合，產生新的座標，且在這個座標系中所有的因子相互獨立。此種爲所有變數綜合特性的新組合，即稱爲主成份。在氣象上此分析法常被稱爲經驗正交分析（Empirical Orthogonal Function Analysis），其中「經驗」代表訊息來自所使用資料，「正交」則代表因子之間無相關的特性（Rudolph, 1988）。

對主成份與原始格點資料的關係，我們可以有以下的定義，令$[S]=[A][X]^T$；或者

$$\left.\begin{aligned} S_1 &= A_{11}X_1 + A_{12}X_2 + \cdots\cdots + A_{1g}X_g = \sum_{i=1}^{g} A_{1i}X_i \\ S_2 &= A_{21}X_1 + A_{22}X_2 + \cdots\cdots + A_{2g}X_g = \sum_{i=1}^{g} A_{2i}X_i \\ &\cdots\cdots \\ S_p &= A_{p1}X_1 + A_{p2}X_2 + \cdots\cdots + A_{pg}X_g = \sum_{i=1}^{g} A_{pi}X_i \end{aligned}\right\} \quad (1)$$

其中 $[S]_{p*1}$:爲主成份矩陣

$[A]_{p*g}$:爲權重係數矩陣

$[X]_{g*1}$:爲資料矩陣

亦即對資料陣列乘上一個權重，使每個格點資料有其特定的權重值。接下來我們希望按照主成份所含訊息的多寡，依序分爲第一主成份、第二主成份...，由此可將訊息集中到前面少數幾個主成份，如此可用較少數的主成份來代表原有的多個變數，將變數數目予以減少。

定義主成份後，由於我們的要求是在所有的個案中，主成份的變異數（V_{ss}）有最大值。另外，爲避免變異數的無限放大，有邊界條件 $\sum_{i=1}^{g} A_{Ki}^2 = 1$ 的限制，由以上的條件，可得以下方程組：

$$\left.\begin{aligned} & S_k = \sum_{i=1}^{g} A_{ki}X_i \\ & V_{ss}\text{有最大值} \\ & [A][A]^T = I \end{aligned}\right\} \quad (2)$$

由 $V_{SS} = \frac{1}{M}\sum [S][S]^T$

$= \frac{1}{M}\sum [AX][AX]^T$

$= \frac{1}{M}\sum A(XX^T)A^T$

$= A\frac{\sum XX^T}{M}A^T$

$= AV_{XX}A^T$

其中M爲樣本數，V_{XX}爲原網格座標資料的變異-互變異矩陣，我們可利用V_{XX}求解V_{SS}；即解$(V_{XX}-\lambda I)A=0$固有結構問題（Eigen Structure Problem）。由V_{XX}矩陣解出的固有值（λ）（Eigen Value）就等於主成份的變異數，而固有向量（Eigen Vector）就等於權重係數（$[A]$）。再由定義主成份的公式代入原來的變數值，即可求得由各變數線性組合後的主成份$[S]$，其中向量S的元素組就稱爲主成份。

透過主成份分析，我們依然保留了所有的資料訊息，只不過將說明訊息的座標轉到變數相互獨立的主成份上。因爲各主成份按變異數大小排列的緣故，使得前面的主成份有較多的訊息，而後面的主成份訊息微弱，也就是說透過主成份分析，可將訊息集中，由前面少數幾個主成份來代表原本多數但可能含有大量重覆訊息的變數，達到資料精簡的目的。

（二）主成份迴歸預報的應用

主成份迴歸預報法依循複迴歸的基本公式，使用多個獨立說明變數（X）來解釋目的變數（Y）。由於多因子間可能存在的相關性，常是各種複迴歸模式需克服的問題。本研究利用主成份各因子之間相互正交的特性可完全避免此一問題，解決傳統複迴歸使用逐步迴歸的繁雜性，並可提高計算效率以便利建立及測試迴歸模式的過程。

在建立模式的過程中，主要利用場量資料作主成份分析，其中的計算部分引用處理固有值問題的作法，求解出固有向量，作爲主成份與格點資料之間轉換的工具。在模式發展過程，迴歸模式是建立在主成份因子與預報元之間的關係，最後爲符合實際作業上線的需求，仍需利用固有向量將主成份還原成格點座標上的對應值，而得到預報元與格點資料的迴歸係數。並且透過比較由主成份因子所得迴歸係數推算的估計值與由格點資料的迴歸係數推算的估計值，可檢驗模式在資料型態轉換過程的正確性，增加對模式的信心。

（三）迴歸係數的反演過程

主成份迴歸方法利用主成份的座標代替傳統的格點因子，因此模式建立之後可獲得的迴歸係數是描述主成份與預報元之間的關係。但在實際預報作業上，對預報元的估計仍需使用格點資料爲輸入資料，因此需將模式所得迴歸係數（$\mathrm{Re}\,gPY'$）反演回到格點資料與預報元的關係（$\mathrm{Re}\,gX'Y'$）。

利用模式中求解的固有向量$[A]$，可作爲主成份與預報因子場格點資料之間的轉換工具，我們可得以下的推演過程：

$$Y(t)=\sum_{\#=1,npc}\mathrm{Re}\,gPY'(\#)*PC(\#,t)$$

$$=\sum_{\#=i,npc}\mathrm{Re}\,gPY'(\#)*\sum_{ii=1,ngg}A(\#,ii)*X'(ii,t)$$

$$=\sum_{ii=1,ngg}\left(\sum_{\#=1,npc}\mathrm{Re}\,gPY'(\#)*A(\#,ii)\right)*X'(ii,t)$$

$$=\sum_{ii=1,ngg}M(ii)*X'(ii,t)$$

$$=\sum_{ii=1,ngg}\mathrm{Re}\,gX'Y'*X'(ii,t)$$

其中$M(ii)=\sum_{\#=1,npc}\mathrm{Re}\,gPY'(\#)*A(\#,ii)=\mathrm{Re}\,gX'Y'$

即代表還原到格點座標上的係數。

四、實例測試結果

本節以台北站五月的降水預報來說明主成份迴歸分析方法的模式建立過程及其可應用價值。在雨量與500百帕高度場單點相關分析中，發現以前一年10月至12月三個月平均的高度場與五月份的降水相關最爲顯著（圖一），因此選取該平均資料作爲預報因子，並以主成份迴歸方法進行模式測試。

接下來本文以8個測試結果（表一）來說明尋找最佳預報方程的過程。主要的測試重點在資料範圍的變動及挑選主成份的影響情形。由於太多的預報因子可能會提高模式的不穩定度（Michaelsen，1987），測試中最多只選取6個主成份，挑選原則是儘可能使用具較大變異量的主成份，因爲變異量過小的主成份其預報訊息可能太弱而無法與雜訊分離。所以只挑選排名在前面15名以內者。

在第1個測試中以整個北半球範圍的資料進行分析，並且主觀選取前面6個主成份，結果在37次的預報中只有13次預報正確，命中技術得分爲負值（-0.06），顯示預報能力稍差（圖二）。爲使與預報元的相關訊息集中在主要的主成份，需對資料進行調整。由圖一分析預報訊息可能偏重在太平洋區，因此在後面的測試中只截取100°E-80°W的資料範圍進行主成份分析，結果顯示第二個主成份型式幾乎完全說明了相關圖中傳達的訊息（圖三）。因此同樣是選取前6個主成份當說明變數，在第二

個測試中三分命中技術可達到+0.15，已顯示出相當高的得分（圖四）。

由於含有較大變異數的主成份未必表示能對預報元提供較多的預報訊息，因此在第3個測試到第8個測試將根據各主成份與預報元相關的程度，逐一將相關值較高的主成份（pc2、pc6、pc13、pc14、pc9、pc1）加入迴歸模式進行測試。以第3個測試而言，只選取具最高相關值的第二個主成份製作迴歸，由於其與預報元的相關值達-0.51，因此雖只是單迴歸方程，其命中技術得分已可出現正值，預測值與實際值的相關程度也可達+0.4，可說已具備作業應用價值。在第四個測試之後，隨著有效說明變數的增加，預報成效的提升可在四個檢驗方式中看出。以第8個測試而言，在加入前面6個較高相關的主成份後，命中得分可達+0.18，而且不論在預報趨勢相關程度（+0.55）或定量技術得分（+29.7%）上，都有相當高的成績（圖五）。

經由上面的測試結果可知，若大尺度環流存在對台灣地區天氣因子預報的訊息，藉由主成份迴歸方法，可將有效的資料訊息集中，提供具應用價值的預報結果。除此之外，將主成份迴歸係數反演回格點座標上，則得格點迴歸係數圖（圖六），可提供我們對模式中掌握的預報訊息有清楚的圖象概念，這也可改善以往無法對傳統模式產品解釋的問題。

五、結語

主成份迴歸預報法可大量濃縮資料，並各成份因子正交的特性可改進統計複迴歸的應用技巧。本研究主要的目的即在引用此法應用至月長期預報作業，測試其在月均溫與月累積雨量的預報成效。經由簡單的單一預報因子場測試，初步已證明其具應用價值。惟爲同時截取所有可能的有效預報訊息，以下是未來可以繼續嚐試的工作：

（1）測試單場多Lag組合。經由本研究雖可找出某一落後時間對預報元較具預報訊息的一個延遲相關預報因子場，但並不代表其他的延遲相關時間就完全沒有預報的訊息。主成份迴歸法對資料的消化能力有助於對大量資料的處理，因此可考慮引用所有延遲時間的預報因子場，模式將會把各Lag中有共同解釋訊息的因子合併，在提供更充足的資料訊息後，可能提高預報的準確度。此即主成份迴歸的進階使用－Mutiple PC-Reg。

（2）測試多場多Lag組合。同（1）之作法，亦爲多重主成份迴歸，但爲引用多種可提供預報元訊息的預報因子場資料。如海面溫度、700百帕高度場、高低層風場等。由於預報結果的好壞除受模式影響外，主要仍要依賴預報因子資料訊息的提供，以目前的研究而言，只採用北緯10度至80度的500百帕高度場資料，大部份爲中高緯度地區的訊息，如此可能是不足以說明影響臺灣地區的天氣變化的，因此加入足夠的資料訊息或許可幫助對預報元的掌握。

六、參考文獻

林秀雯、陳雲蘭，1994：利用主成份迴歸分析探討單一場量與台北降雨機率的關係。天氣分析與預報研討會論文彙編（83），453-458。

戚啓勳、嚴夢輝，1978：氣象統計學。復興書局,288-308。

陳建民、張智北，1993：主成份分析應用在完全預報及數值模式預報之迴歸統計之研究。天氣分析與預報研討會論文彙編（82），401-401。

陳雲蘭，1996："應用主成份迴歸分析方法發展長期預報"，中央氣象局研究發展專題報告。

Brier G. W., 1950 : Verification of forecasts expressed in terms of probability, Mon. Wea. Rev., 78, 1-3.

Michaelsen, J., 1987: "Cross-validation in statistical climate forecast models". J. Climate Appl. Meteor., 26, 1589-1600.

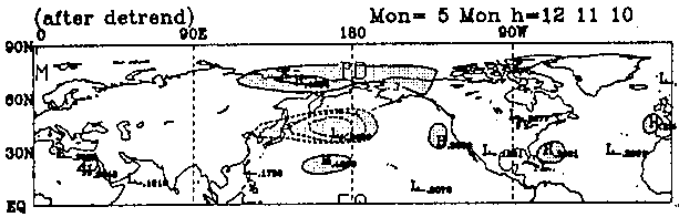
Rudolph W. Preisendorfer, Curtis D. Mobley, 1988：Principal Component Analysis in Meteorology and Oceanography. Elsevier Science Publishers B.V.,1-425.

Vautard, R., Pires, C., Plaut, G., 1996: "Long-range atmospheric predictability using space-time principal components". Mon. Wea. Rev., 124, 288-307.

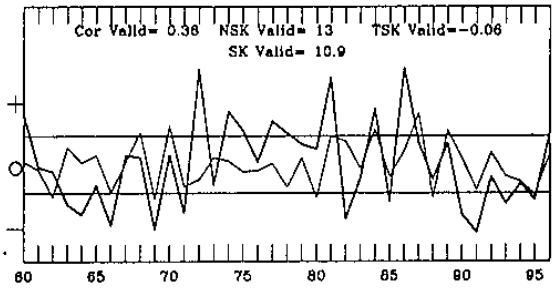
Yu, Z.-P., Chu, P.-S., Schroeder, T., 1997: "Predictive skills of seasonal to annual rainfall variations in the U.S. affiliated Pacific islands: canonical correlation analysis and multivariate principal component regression approaches". J. Climate., 10, 2586-2599.

	Selected pc	Cor	Nsk	Tsk	Sk
1	pc1+..6 (N.H.)	0.36	13/37	-0.06	10.9%
2	pc1+..6 (regional)	0.30	18/37	+0.15	3.7%
3	pc2	0.40	16/37	+0.04	15.7%
4	pc2+6	0.41	16/37	+0.03	17.0%
5	pc2+6+13	0.48	17/37	+0.10	22.6%
6	pc2+6+13+14	0.54	16/37	+0.08	28.7%
7	pc2+6+13+14+9	0.58	17/37	+0.14	32.7%
8	pc2+6+13+14+9+1	0.55	18/37	+0.18	29.7%

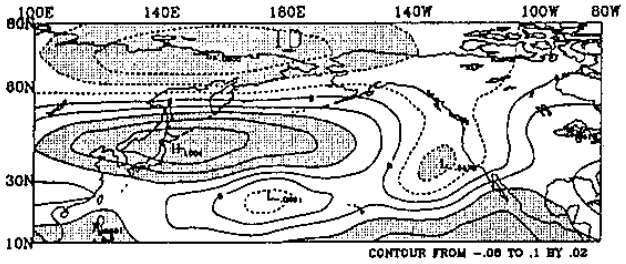
表一、利用前一年 10 月至 12 月 500 百帕平均高度場預報台北站五月雨量中,8 個測試模式的預報成效。其中 Cor 表示預測值與實際值的相關程度， Nsk 表示預報命中率， Tsk 表示三分法的命中技術得分， Sk 表示定量誤差的技術得分。8 個測試中除第一個測試的資料範圍爲北半球外，其餘皆使用經選取的區域資料。



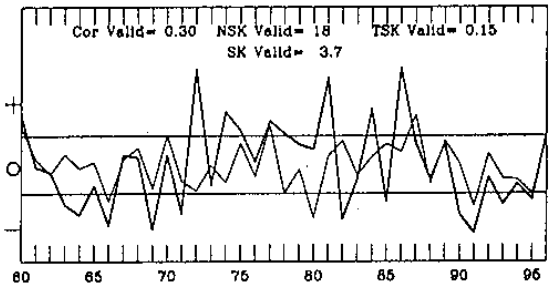
圖一、台北五月降雨資料與前一年10月至12月500百帕平均高度場單點相關圖



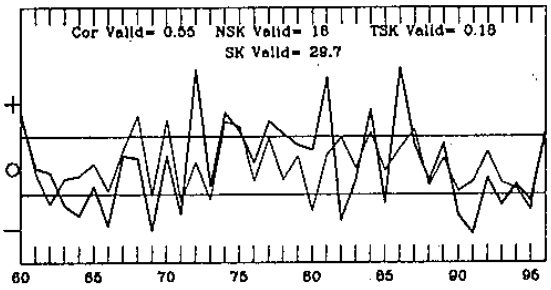
圖二、第1個測試中交錯驗證法的時間序列。其中粗曲線表實際值，細曲線表預測值。二條細直線區分出三分法中的多雨（＋）、正常（○）及少雨（－）範圍。



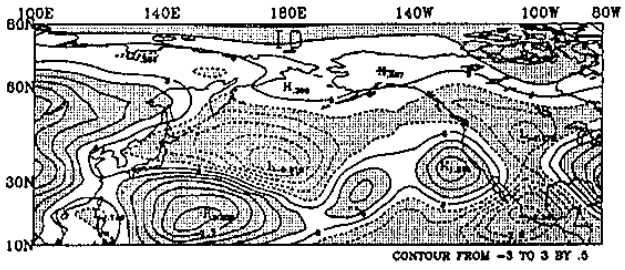
圖三、使用經選取區域資料作主成份分析後的第二個主成份型式。該主成份與台北五月雨量有最高的相關值(-0.5)。



圖四、同圖二，但爲第2個測試。



圖五、同圖四，但爲第8個測試。



圖六、使用第8個測試模式所得的迴歸係數圖。