「發展鄉鎮逐時天氣預報」計畫之高解析度統計預報技術研究(1)
-- 引入最小絕對壓縮挑選機制(LASSO)改善建模成效

陳雲蘭[1] 王政忠[1] 劉欣怡[1] 馮智勇[2] 薛宏宇[2] 劉家豪[2]
[1]中央氣象局氣象預報中心
[2]多采科技

摘 要

在降尺度統計迴歸建模過程，需要對大量的動力模式多層場資料進行有效的篩選，而過去對於統計模型的建立，一般常採用最小平方法以多元線性迴歸模式來進行預測，然而受限於其線性模型的誤差及其處理高維度變數的能力，直接運用傳統逐步迴歸的變數選擇方法進行天氣預測仍有改善空間。為提升統計預報成效及改善統計預報建模程序，我們在先前的研究中引入使用壓縮係數的概念，提出可將同時具有係數壓縮與變數挑選的 LASSO 方法應用到天氣預報上，並經測試證實可以有效提升預報模式的穩定性。本研究藉由 99-100 年 FIFOW 計劃甫剛完成的統計建模發展暨作業環境，以實際作業預報實例資料對新引入 LASSO 方法與氣象上傳統常用的統計迴歸方法進行比較，結果證實 LASSO 確實能改善建模程序並提升預報準確度。

一、前言

「發展高解析度統計預報技術」是本局 99 年至 100 年所執行「發展鄉鎮逐時天氣預報」(FIFOW)計畫中的重要工作項目之一，除了先採用理想模式預測法(Perfect-Prog)的統計降尺度預報策略，結合動力數值天氣預報，完成預報作業系統雛型並提供初步指引的產製之外，亦同時進行相關研發改善統計技術，以期提供更可靠的高解析網格統計預報指引。此 2 年在此計畫下所進行科研大致可分為 2 個方向：(1) 發展高解析度統計預報技術 (2) 改善統計方法與迴歸技術。

1.1 發展高解析度統計預報技術

面對有限的觀測點，要提供高解析度格點預報，一般有 2 種做法：一為先預報測站點的天氣再根據其值對高解析網格點進行估計，目前美國作業單位即採用此方法(Glenn,2009)；另一種則為先對高解析網格點求取觀測分析值，再來進行統計降尺度預報。目前 2 種方法各有擁護者，後者的可行性依賴高解析網格點的分析技術，而採先預報再內插的前者則依賴空間內插技術以及測站點預報模型的表現。相較於其他國家，台灣測站的密度並不低，如果能納入更多資料源增加預報測站並搭配提升測站預測能力的工作，在引入合適的空間統計技術下，上述先預報再內插的降尺度策略較為經濟，有效率上的優勢。

FIFOW 計劃分別在高解析網格點分析技術及高解析度統計預報技術的研發投入心力，得以同時進行上述兩種策略以提供評估比較。因此，一方面我們利用該計畫研發完成的近 5 年高解析度網格地面分析資料提供統計建模，直接預報所需高解析網格點。另一方面，我們也開發空間內插技術，引用克利金空間統計方法(李，2009)進行先預報測站再內插的高解析度統計預報，此法比起一般使用的距離內插更為合理，另外，也可直接處理溫度在地形上的垂直變化。

同時，為了能更理解空間內插技術的應用成效，我們也研究了美國作業單位所採用的 BCDG 方法(Glenn,2009)，希望用以與克利金空間統計方法所製結果進行比較。過程中，我們發現 BCDG 法有一些可以改善的地方。此方面的詳細結果，另文於本研討會另一報告(馮等，2011)。

1.2 改善統計方法與迴歸技術

對於高解析度格點預報，不管是採用先內插再進行格點統計預報，或是先預報測站再內插至格點，都會經過統計模型建置的程序，其中包含一些資料處理、模型假設或是變數選取等設計常是發展統計預報技術的核心議題。在 FIFOW 計畫初期，為了對統計建模資料有合適的分類處理，我們做了分季方式的討論，並確認了以逐月分別建模的可行性(陳等，2009)。後續的研究則著重在統計迴歸模型的改善。

在降尺度統計迴歸建模過程，需要對大量的動力模式多層場資料進行有效的篩選，而過去對於統計模型的建立，一般常採用最小平方法以多變量線性迴歸模式來進行預測，然而受限於其線性模型的誤差及其處理高維度變數的能力，直接運用傳統逐步迴歸的變數選擇統計方法進行天氣預測仍有

改善空間。我們希望善加利用資訊截取及變數選擇等技術來改善建模品質，並期提升模型的預測能力。針對問題，我們引入壓縮係數法(王與陳，2009)，將同時具有係數壓縮與變數挑選的 LASSO 方法應用到天氣預報上，經測試已證實可以有效提升預報模式的穩定性，尤其，LASSO 搭配邏輯斯迴歸模型在降水機率預報上，更見成效(王與陳，2010)。

本報告是說明 FIFOW 計畫下發展高解析度統計預報技術研究的系列報告之一，主要將說明引入 LASSO 改善建模成效的情形。以下第二節將先介紹傳統的統計預報建模方法及潛在問題，第三節則簡略介紹引用 LASSO 方法的優勢及應用，有關壓縮係數法更深入的說明，請詳見王與陳(2009,2010)的報告。在第四節中我們使用作業實例資料對 LASSO 與傳統建模方式進行比較，最後說明未來發展方向。

二、統計預報建模方法

不談對資料的前置分類處理，統計建模過程包含模型假設及變數挑選 2 個議題，換言之，建模之前需考慮 2 個設計，1 個是對套用模型的假設，另一個則是針對變數挑選的方法。預報模式的優劣可說是所選擇套用模型與變數挑選方法的集成結果。

2.1 常用多元線性迴歸方法

2.1.1 模型假設與係數估計

迴歸建模首要工作是要決定合適的模型，決定的關鍵除了根據預報目標性質，亦要盡可能去維持原始建模資料特性，讓預報模型具有合理性。在統計預報方法中，線性迴歸模型是最簡便的模型，迴歸係數的估計是用最小平方法(OLS)，其特點是不偏估計，亦即所估數值的平均值與期望值相同。氣象上很多像是溫度等連續型的天氣要素，以線性迴歸模型都能有不錯的估計。對於像是降雨事件之有無等二元化變數則需使用邏輯斯迴歸模型，迴歸係數改用最大概似估計法(MLE)來估計。不過，氣象上很多應用者爲求方便，常見仍是直接使用線性迴歸模型來進行二元化變數的預報。王與陳(2010)在引用 LASSO 搭配邏輯斯迴歸模型對降水機率預報的研究中，對使用線性迴歸建立機率預報模型的不合理性有充分的討論。

2.1.2 預報變數挑選方法

面對氣象資料具高維度的預報變數，常見的挑選方法爲前序選取(Forward Selection，後稱 FS)法。在每次選取過程利用判定係數(R-Square)等準則決定哪些變數可以進入模型(陳等，2009)。FS 進行的是逐步挑選的程序，應用者可以依滿足點決定停止挑選的時機。雖然說加入的變數愈多可使模型的擬合度愈高，但一般而言，我們希望建模擬合度高，又不希望對建模資料過度擬合(Overfitting)反而導致帶入預報資料時發生不穩定，因此理論上選取預報變數個數的多寡需要經過評估。不過在氣象預報作業中爲求方便，通常不作細考，而依經驗值直接給定。例如美國作業單位大約挑選 10~15 個，本局先前發展的 MOS 有些甚至只用不到 5 個。

2.1.3 共線性問題

另外，在多元線性迴歸中，如果迴歸模式中預測變數之間有太高的相關時，就會造成共線性的問題，將可能影響模型係數的估計，甚至產生迴歸係數與相關係數正負符號不一致的不合理現象。這是因爲變數間的高共線性使估計參數的求解行列式值很接近零而產生奇異（singularity），造成估計值的不穩定。對於共線性問題的處理，一般有下列幾種建議，(1)刪去高共線性的重覆變數，亦即將彼此相關係數較高的預測變項只取一個重要變項投入分析，(2)主成份迴歸分析 (3)使用帶有偏差的迴歸分析法(Biased Regression Analysis)，例如脊迴歸等壓縮係數法。在天氣尺度統計預報應用中，上述前 2 項包含降低變數間的共線性或採用主成份分析的建議較常見，至於像是脊迴歸等壓縮係數法則較少被應用，本研究所引入的 LASSO，即是此類方法，將於下節說明。

FIFOW 計畫在發展初期先參照過去經驗進行統計預報指引的產製，所採用的建模設計乃依線性迴歸模型搭配前序選取法，預報變數方面則依經驗先固定挑 10 個。對於共線性的處理方面，則另設計 2 個參數，1 個把關變數對預測天氣的相關顯著程度，另一個則是控制已挑選變數間的共線性。可知上述現行做法需要某種程度的主觀來決定預報建模參數，雖然這些參數的選擇影響未必關鍵，但事前給予控制參數爲預報成效分析增添不確定性。另外，由於迴歸模型是以建模資料爲依據，而逐步迴歸方法對資料很敏感，很可能因爲資料略有變化即影響了最後建立的模型。對於這些問題，我們引入由 Tibshrani(1996)提出的 LASSO 估計方法，此法能有效地處理高共線性的資料，不需主觀給予控制參數，並且透過對係數的壓縮特性可增加預報模型的穩健性。

2.2 最小絕對壓縮挑選機制的應用

氣象資料有具高共線性的特徵，如上節所提，處理共線性的方式之一可以考慮使用具偏差的迴歸分析法，山脊型迴歸是這類方法常被提及的一種，其作法是在最小平方法估計係數時，另外加入一個 2 維的限制式，使其係數的平方值總和限制在一個最佳特定值之內，如此透過對於係數的壓縮，可以降低模型對資料的敏感性，同時透過額外加入的限制式，可以解決當共線性高時行列式值很接近 0 的問題。

LASSO 是另一種具偏差的迴歸分析法，與上脊型迴歸方法不同的是，其所加入的是 1 維的限制式，限制的對象是所有係數值的絕對值總和，這樣的設計在數學上求解其實是比較困難的，不過卻能獲致另外一個極大好處，就是有機會將部份係數壓縮至零值，同時解決建模過程中需要決定挑選變數的問題。

$$\min_{\beta}\|Y-X\beta\|^2 \quad subject \quad to \quad \|\beta\|=\sum_{j=1}^{d}|\beta_j|\le t$$

與上節所介紹傳統常用的統計迴歸預報方法相比，LASSO 模型的穩健性及更簡化更客觀的預報變數決定程序對於改善天氣統計預報存在可能性。為了實證，我們在先前的研究(王與陳，2009)，以台北日均溫的預報為例，對 LASSO、複迴歸法、逐步迴歸法等傳統常用的統計預報方法進行評比，觀察到 LASSO 不僅解決了建立迴歸模式時面臨大量解釋變數的問題，並且其更穩定的預報方程不易受資料變動而影響結果，也提高了模式的預報準確度。

LASSO 的提出最初是針對線性模型估計式的改善，不過，LASSO 限制式的想法也可以應用在處理二元化資料的邏輯斯迴歸模型，做法上只要在概似函數式加入迴歸係數限制式，藉由數值方法求得邏輯斯迴歸係數估計。為了解 LASSO 搭配邏輯斯迴歸模型對降水機率預報的應用效益，我們在先前的研究(王與陳，2010)設計了 4 組不同模型假設與變數挑選的建模組合，結果發現雖然以 FS 搭配邏輯斯迴歸模型的建模擬合度最佳，但預報成效卻最差，呈現 overfitting 的問題。以 LASSO 搭配邏輯斯迴歸模型的預報成效則為 4 組中最佳，其建模擬合度居次，亦即仍高於其餘 2 組使用線性假設的模型。該研究說明以線性假設模型進行降水機率預報不僅在數學上有不合理之處，過程需要後製處理，也顯得比較麻煩。可是即使正確更換邏輯斯迴歸模型解決了合理性的問題，若是搭配了 FS，反而又會因 overfitting 遭遇使預報成效更差的狀況。我們推測或許這也是許多氣象應用者棄邏輯斯迴歸模型而仍採用線型假設的原因。我們的研究說明其實這樣的問題在應用 LASSO 取代 FS 的變數挑選法之後即可以被解決。

LASSO 被提出距今只有十多年，應用還不算廣，在氣象預報領域也幾乎未見。我們在 FIFOW 計畫下研究 LASSO 方法，認為此法對於改善預報成效及簡化作業流程很有潛力，目前仍繼續開發其在氣象可應用之處，其中一個仍在進行中的工作為順序型邏輯斯迴歸模型的應用(王與陳，2011)。

三、作業實例比較分析

經過先前的研究及測試，我們認為 LASSO 的在氣象上具高度實用價值。本研究接續藉由 99-100 年 FIFOW 計劃甫剛完成的統計建模發展暨作業環境，以實際作業預報實例資料對新引入的 LASSO 方法與氣象上傳統常用的統計迴歸方法進行比較。目前預報方程乃以 2007~2010 年資料建模，提供對 2011 年的預報計算。以下將以 2011 年上半年 6 個月份為例，比較傳統建模方式與新引入 LASSO 方法的差別。根據預報對象資料屬性的不同，我們又依連續型變數及機率型變數分別討論。其中連續型變數將以正點氣溫、最高氣溫，夜間低溫作為代表，機率型變數以降水機率預報為例。

模型間的比較乃透過對建模擬合情形以及模式預測情形的觀察。在連續型變數方面，簡單利用相關係數及預報誤差的均方根(RMS)分別從趨勢及定量方面來比較擬合度。對於機率型預報變數則是使用白氏得分及 AUC(Area Under Curve)值來分別檢視模式的定量誤差及分類能力。有關校驗方法的詳細介紹，可參見王與陳(2010)。

3.1. 連續型預報變數

若預報目標的數值在實數域中變化，是所謂的連續型變數，例如溫度類的天氣要素。常見的建模設計是以線性模型作為假設模型。此節我們設計 4 種建模方式，希望充分了解傳統作法下各種控制參數的影響及使用 LASSO 的可能效益。

在傳統作法中，為了解決共線性問題及挑選合適變數，我們在建模過程設計 3 個控制參數，第 1 個是各影響變數與預報對象相關值的 p-value，可決定剔除不重要變數的程度，其值愈小，表示剔除不重要變數的程度愈大。第 2 個是檢查變數間共線性的容忍值，其值愈大，表示對於共線性的把關程度愈大。第 3 個則是挑選變數的個數，根據經驗目前作業我們皆選 10 個。如上節所提，使用傳統方法，需要去決定最佳控制參數，不然就是要經過多次測試或透過經驗，主觀的選擇一個大約可接受的值。

以下比對分析，我們所計算的 4 種模型，前 3 種是傳統常見的線性模型搭配前序選取法挑選變數，所不同者在於控制參數的選擇，其中第 1 種以 FS10_P0 稱之，其變數個數固定取 10 個，p-value 設為 0.001，亦即只有顯著相關者才有機會進入模型。第 2 種以 FS10_P1 稱之，變數亦固定取 10 個，p-value 設為 1.001，等於是不先過濾變數與預報對象的相關程度。第 3 種以 FS30_P1 稱之，同樣不先考慮相關程度，但把變數提高至 30 個。最後第 4 種則是 LASSO 方法，其事先亦不特別處理相關性。本研究對多站進行了比對，其結果類似，以下說明將僅以台北站的測試結果作為代表。

圖 1 至圖 3 為建模擬合校驗比較圖，預報目標依序為上午 8 時正點氣溫、白天高溫、夜間低溫，圖 4 至圖 6 則為對

應的預測情形，其中長條圖爲4年建模資料的標準差，代表資料變異氣候背景值，上方的曲線爲模型估計值與實測值的相關係數，下方曲線則爲估計均方根誤差。

觀察正點氣溫的模擬，4 種模型的建模擬合程度皆相當高，各月份RMS在1度左右，大約只有資料平均變動的1/3；相關值在0.93~0.98之間，1, 2月略高於其他月份。更仔細比較，FS30_P1因變數選得較多，建模擬合程度略高於其他模式，不過反應到預報上，FS30_P1卻是相對最差的，顯示出過度擬合的現象。相反的，LASSO雖然建模擬合略低於其他3個FS模式，但是其預報卻是最佳的。至於FS10_P0與FS10_P1的比較，在6個月內互有些微領先，顯示差異不大。

LASSO比FS好的情形同樣可見於高低溫的實例分析。對於白天高溫這種可能受多重因素影響的天氣因子，FS模型取30個變數明顯比只取10個的建模擬合程度來得好，但預報卻較不佳。不過，LASSO的變數也大約取了近30個，但由於壓縮係數發揮穩健效果，在預報時比較不會有不穩定的情形。這樣的結果與之前的研究結論一致，再次驗證LASSO的確優於傳統OLS與FS的組合，除了可提供預報準確度，而且也免去了事先決定變數挑選原則等控制參數的程序。

3.2. 二元類別機率型預報變數

在天氣預報的項目中，降雨的有無是以機率的型式來發布預報，不過此項資料的觀察值是事件發生與否的二元化資料，由二元化資料進行機率預報的方式可使用邏輯斯迴歸的假設模式，王與陳(2010)在這方面已有清楚討論。

雖然直接將降雨有無的二元化資料直接轉爲機率值的0%及100%，再將其視爲連續型變數以線型模型處理，也是可以發展出預測模型，但王與陳(2010)已論述其中數理上不盡合理之處。惟包含美國的很多作業應用或研究，仍常見直接以線型模式來處理降水機率預報。我們的研究認爲可應用邏輯斯迴歸模型配合LASSO得到改善。爲觀察成效，我們將其與2種使用FS的模型相比，其一稱爲FS_LOGI，乃使用邏輯斯迴歸模型，另一稱爲FS_OLS，使用線性模型。

在擬合圖方面(未附)，可見以FS搭配LOGI的吻合度最佳，但反應到預報則仍以LASSO模型的表現爲較佳。透過不同降雨氣候背景的6個測站的預報分析(圖7至圖12)皆顯示同樣結果。

四、未來方向

因應小區域預報及數位化預報服務趨勢的需求，我們需面對處理大量預報目標點的挑戰。在追求改善預報的過程，我們朝儘量避免逐站逐點去尋求最佳化程序的方向思考。經過研究及實例測試，我們已證明LASSO不只能提升預報準確度，又可免除主觀設定預定參數，同時其建模效率也遠比逐次法好，而且不只是在線型模型、非線型模型，甚至是多類別的預報皆可應用，非常適合引入氣象預報作業。99-100年的FIFOW計劃已先使用傳統方法建置模型發展及預報作業環境，接下來我們將把LASSO方法也加入正式作業環境，在更多的實例比較確認之後，擬改以LASSO爲主來提供預報指引。與此同時，發展高解析度統計預報技術的各項研究也將持續進行，除了點預報，我們也將加入場預報的思維，LASSO在這方面也能提供應用，屆時仍會是我們求發展所依賴的方法之一。

五、參考文獻

Glahn, B., K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009a: The gridding of MOS. Wea. Forecasting, 24, 520–529.

Tibshirani, R. 1996; Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267–288.

王政忠、陳雲蘭，2009：最小絕對壓縮挑選機制(LASSO)於天氣分析迴歸預報的應用。天氣分析與預報研討會論文彙編，中央氣象局，314-319。

王政忠、陳雲蘭，2010：邏輯斯迴歸(Logistic Regression)模型輔以最小絕對壓縮挑選機制(LASSO)於降水機率預報之應用。天氣分析與預報研討會論文彙編，中央氣象局，236-241。

王政忠、陳雲蘭，2011：「發展鄉鎮逐時天氣預報」計畫之高解析度統計預報技術研究(2) --引用順序型邏輯斯迴歸模型進行雲量多類別預報試驗。天氣分析與預報研討會論文彙編，中央氣象局。

李天浩，2009：應用克利金法建立高解析度網格點氣象數據之研究。交通部中央氣象局委託研究計畫成果報告。

陳重功、羅存文、王惠民與賀介圭，2000：中央氣象局統計預報系統的發展。氣象學報，41，p18-33。

陳雲蘭、王政忠與張琬玉，2009：統計迴歸模式季內時間取樣差異測試。中央氣象局自行研發計畫成果報告第CWB98-1A-03號，16頁。

馮智勇、李天浩、陳雲蘭、高慧萱，2011：「發展鄉鎮逐時天氣預報」計畫之高解析度統計預報技術研究(3) --BCDG空間內插方法分析與應用。天氣分析與預報研討會論文彙編，中央氣象局。

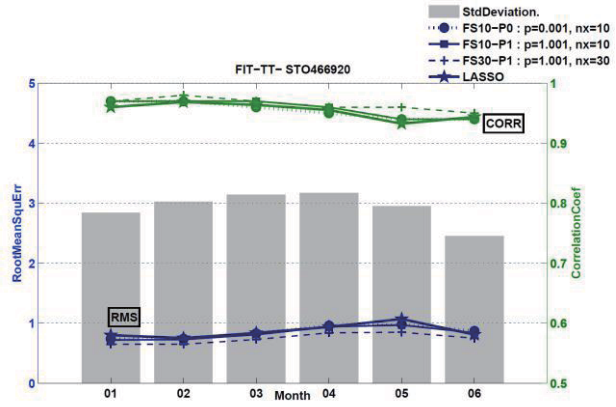

圖 1：4 種統計模型的建模擬合度比較圖，以預測台北上午 8 時氣溫爲例。圖中長條圖爲 1 至 6 月各月份氣溫資料在建模期間的標準差，上方的曲線爲模型估計值與實測值的相關係數，下方曲線則爲估計均方根誤差。其中帶星記符號的粗實線是本研究所提倡的 LASSO 壓縮係數模型，另 3 組爲傳統常見以逐步迴歸方法建立的模型。

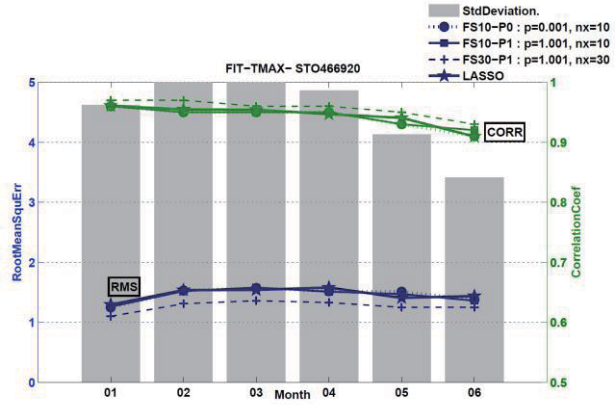

圖 2：同圖 1，但爲以預測台北白天高溫爲例。

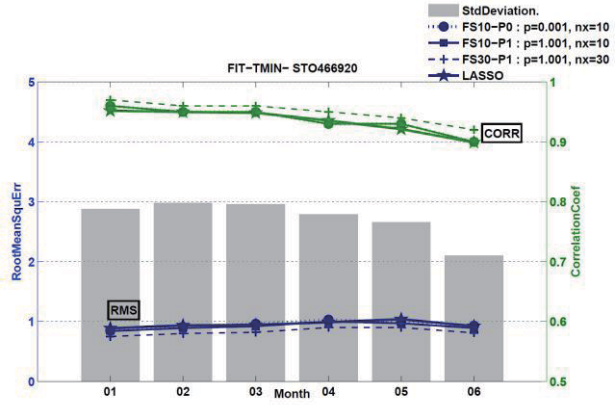

圖 3：同圖 1，但爲以預測台北夜間低溫爲例。

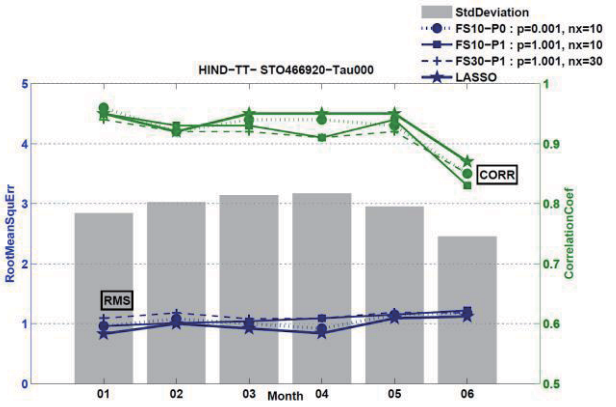

圖 4：4 種統計模型對於 2011 年 1 月至 6 月的預測成效比較圖。圖中長條圖仍爲 1 至 6 月各月份氣溫資料在建模期間的標準差，上方曲線爲模型預測值與實測值的相圖係數，下方曲線爲預測均方根誤差。

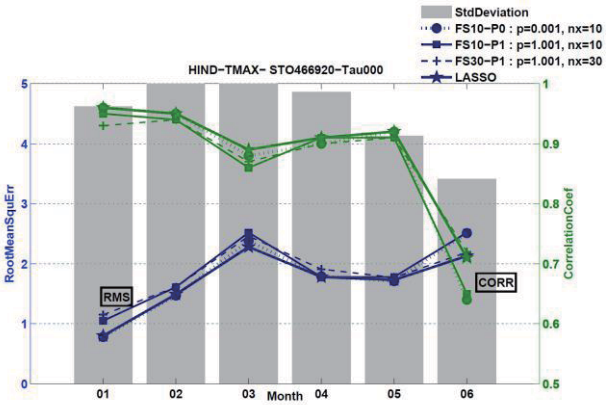

圖 5：同圖 4，但爲以預測台北白天高溫爲例。

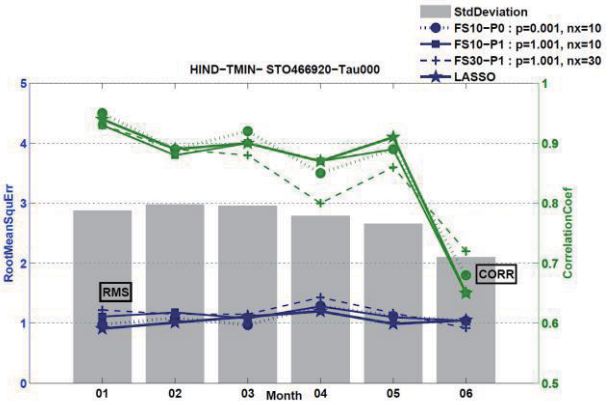

圖 6：同圖 4，但爲以預測台北夜間低溫爲例。

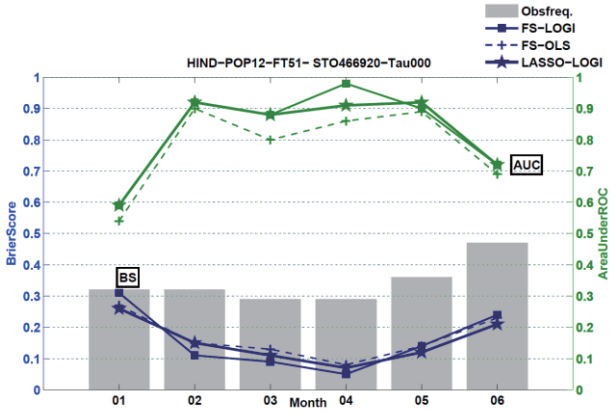

圖 7：3 種統計模型的預報成效比較圖，以預測台北白天降水機率爲例。圖中長條圖爲 1 至 6 月各月份降雨資料在建模期間的發生頻率，上方曲線爲 AUC 值，下方曲線爲白氏得分。

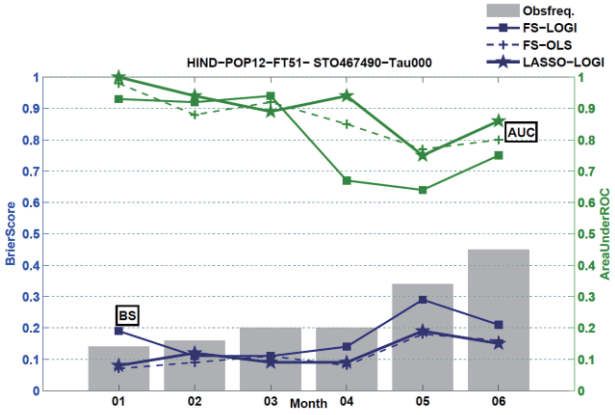

圖 8：同圖 7，但爲對台中站的預測成效比較圖。

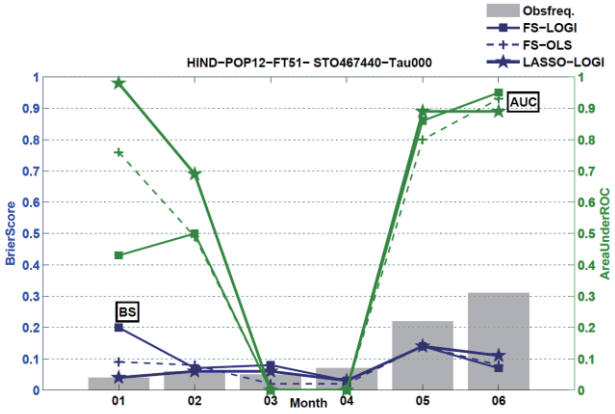

圖 9：同圖 7，但爲對高雄站的預測成效比較圖。

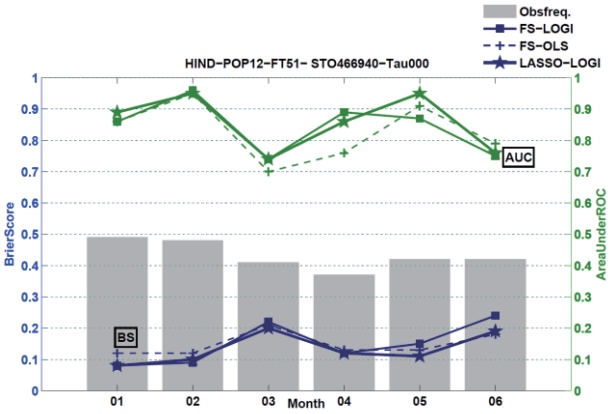

圖 10：同圖 7，但爲對基隆站的預測成效比較圖。

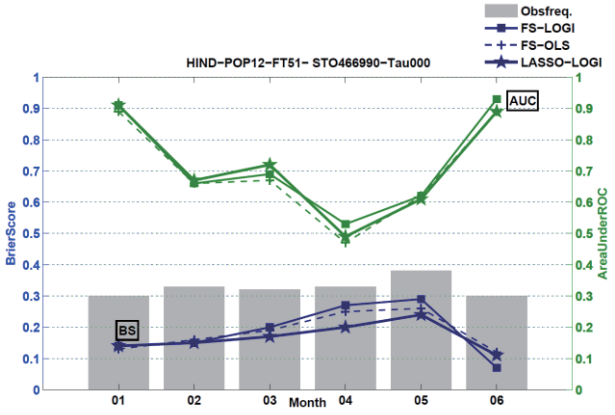

圖 11：同圖 7，但爲對花蓮站的預測成效比較圖。

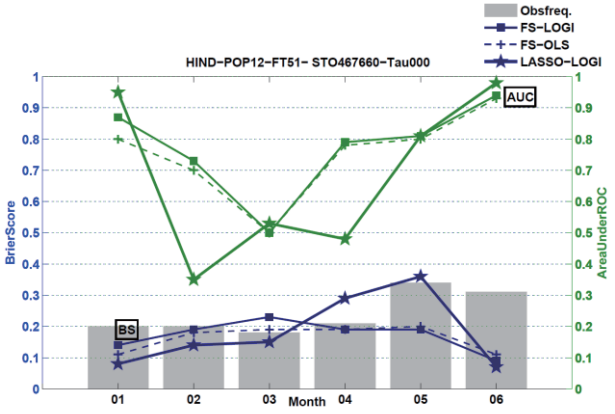

圖 12：同圖 7，但爲對台東站的預測成效比較圖。

。