「臺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

陳雲蘭[1]、薛宏宇[2]、呂致穎[2]、陳品妤[2]、詹智雄[2]、沈里音[2]
中央氣象局氣象預報中心[1]、中央氣象局氣象科技中心[2]

摘　要

臺灣長期氣候資料整集分析計畫是中央氣象局為提昇氣候與氣候變遷應用服務能力所規劃四年(103~106 年) 工作計畫的一個子項部分，此子計畫工作目標是透過對臺灣長期氣候資料的整集、處理及分析，以及建立長時間高解析度的氣候資料，逐步建立本局氣候資訊應用服務的基礎，進而提昇本局對臺灣氣候資料及氣候與氣候變遷資訊服務的能力與品質。計畫工作由氣溫記錄著手，整理累積已近 20 年的自動氣象站觀測值，對於長期資料難免缺遺問題，計畫首年(103 年)已發展了合理補整的方法，使城鄉尺度應用所需的高解析網格化資料可在參考測站時空採樣均一的條件下產製。計畫後續工作持續在資料品質檢測技術尋求改進，本研究提出透過與觀測值與估計參考值的迴歸關係產出的迴歸離群值(regression outlier)來協助檢測，同時引入刪除式殘差(Studentized Residuals)的應用來量化觀測值的離群程度。結果顯示成效良好，可含蓋一般使用總體大型誤差檢驗(Gross error check) 、氣候背景範圍檢驗(Climatological test)、改變率檢驗(Rate-of-change limits) 及連續型檢驗(Continuous no-observed-change with time limits)等所得檢測結果。

一、前言

高品質的多年長期觀測資料是提供各類氣候資訊應用服務的基礎，有多年連續觀測的氣象數據可以提供氣候背景的基本認識、提供氣候變化的診斷與分析，乃至預報應用。多年氣象資料也是各學科領域探討環境影響力之所需，從許多應用研究可知氣象因素是影響生活環境、社會經濟活動、自然資源變化的常見關鍵因子。

中央氣象局對於提供台灣多年長期觀測資料的服務，過去集中在近 30 個有人駐守、資料經嚴謹品管的氣象觀測站記錄，這些測站分布於全台，其中至少有 20 個測站的觀測已超過半個世紀，甚至有 10 個站已達百年，大致可滿足對於台灣各大範圍分區長期氣候變化的分析需求。除了這些有人員駐守的主要測站，中央氣象局自 1980 年代開始規畫陸對全台佈建自動氣象站，首批從大台南及大台北開啟，從西部到東部，由疏至密，過去卅年經過數期佈建及汰換計劃，目前已有超過 300 多個自動氣象站。雖然現有測站有近半比例是近 10 年內所建置, 但也有百餘站已連續記錄氣象超過十數年，如此多年長期觀測資料已顯現出可為氣候資訊應用的價值，而且大比例提高空間分布密度的測站數量也使鄉鎮尺度氣候資訊服務變得可能，近年各界在這種時空高解析產品的需求已見增長。不過，由於自動站因遠地儀器維修不易及電訊傳遞有時不良等資料接收問題，其資料品質相對於人工測站較不穩定，因此在運用自動站資料於各類應用分析之前，對於過往觀測收集資料的正確檢核就成了一個非常重要的前置資料處理工作。

本研究探討自動氣象站長期氣溫觀測值合理性檢測方法，是中央氣象局「臺灣長期氣候資料整集分析計畫」下的一個相關工作，此計畫自 103 年開始執行，總計 4 年的工作目標希望透過對臺灣長期氣候資料的整集、處理及分析，以及建立長時間高解析度的氣候資料，逐步建立本局氣候資訊應用服務的基礎，進而提昇本局對臺灣氣候資料及氣候與氣候變遷資訊服務的能力與品質。計畫工作由氣溫記錄著手，整理累積已近 20 年的自動氣象站觀測值，對於長期資料難免缺遺問題，計畫首年(103 年)已發展了合理補整的方法，使城鄉尺度應用所需的高解析網格化資料可在參考測站時空採樣均一的條件下產製(陳等，2014)。計畫第 2 年(104 年)除了進一步納入對雨量及溼度資料的處理之外，同時也持續在資料品質檢測技術上尋求改進，本文將主要說明近期在這方面工作的一個改進與突破，提出透過觀測值與估計參考值的迴歸關係產出的迴歸離群值(regression outlier)來協助資料正確性

檢測，同時引入刪除式殘差(Studentized Residuals)的應用來量化觀測值的離群異常程度。

資料的品質控管過程可以分有多個層次，一般在即時接收階段會有一些簡易的檢核處理，例如依物理特性給定合理性範圍檢查大型錯誤 (Gross error check ; limit check) 等。進入資料處理中心整集儲存階段則會再有進一步的檢核處理，若以單一測站資訊出發，可能進行背景範圍檢驗(Climatological test)、改變率檢驗(Rate-of-change limits) 、連續型檢驗(Continuous no-observed-change with time limits) 、或是使用同站多項觀測變數檢查資料一致性等。若從運用測站之間的空間關係出發，則可採用鄰近站相關檢核(buddy check)技術。本研究鑑於單由單一測站資訊並無法對錯誤異常值與實受天氣影響的顯著離群值做出有效的區隔，於是探討空間關係檢核技術的可利用性及運用方式，希望發展更有效的資料檢查方法，以幫助長期氣候資料整集過程所需建立的資料品管流程。

二、資料及研究方法

本研究希望利用測站之間的關連性來發展資料品質檢查方法。就氣溫而言，個別測站與鄰近站常因受同一空氣團影響而有一致的冷暖起伏變化，可作為相互驗證的參考。此類由空間關係所發展的驗證方式，一般會對尋找高相關鄰近站做出定義，再根據所選出鄰近站的量測數值以內插方法估算待檢測測站的推測值，提供判斷觀測數值合理性的比對參考。設計這樣的一個資料檢查過程，有 3 個重要問題需要處理：(1) 鄰近站的合適選擇 (2) 內插估計方法的合適選擇 (3) 觀測值與檢測參考推估值的比對程序。這些問題的處理方式皆可能影響檢測的成效。本研究所屬計畫在先前的工作已對於空間內插問題做過詳盡探討(李，2009；馮等，2012)，以實例驗證克利金方法優於其他內插工具，其不只以數學最佳化的方式求出內插權重係數，在對於具影響力測站的選擇方面，也能根據資料客觀求算，而不似多數方法主觀事先給定。對於長期氣候資料整集過程所需的資料檢查方法，本研究繼續應用克利金方法來求算資料檢測所需的參考推估值。至於對於觀測值與檢測參考推估值的比對程序方面，則提出透過觀測值與估計參考值的迴歸關係產出的迴歸離群值(regression outlier)的做法，並引入刪除式殘差(Studentized Residuals)的應用來量化觀測值的離群異常程度，以下分節對使用資料及研究方法做進一步說明。

(一) 使用資料與目標測站

在「臺灣長期氣候資料整集分析計畫」第一年(103 年)針對氣溫資料的處理工作中，已挑選出 110 個測站(圖 1)做為首波分析目標，這些測站除了需為具備 15 年以上的長期觀測記錄之外，其有效記錄值的比例亦需達到 85%以上。本研究配合計畫探討氣溫觀測值品質檢測方法，將針對此 110 站分析 1998 年至 2014 年期間所有小時觀測資料的合理性。需要檢測的資料總筆數以概算來說為：17 年 x 365 天 x 24 小時 x 110 站。

這些待檢測的資料乃取自中央氣象局資料處理科，換言之，本文研究所檢測氣溫觀測記錄即是本局目前對外所直接提供的資料。在所選出的 110 個目標站中，有 22 站是有人員駐守觀測的主要測站，其氣溫記錄已經過繁覆檢查，準確度應已極佳，其他的 88 個自動站的量測記錄僅經過簡易檢測，則是本研究的主要分析目標。

(二) 資料檢測參考推估值

資料的品管是將觀測值與一合理預期的推估值進行比對，本研究運用克利金內插技術，以去一法(Take one out)逐站逐時使用該站以外的全部 109 個測站資料求算出推估值。這個計算過程是本項工作最耗時的程序，需執行內插計算的總次數與上節所述需要檢測的資料總筆數相同，亦即所有的觀測值皆能對應到一個檢測所需的推估值。目前實作執行此項總計 17 年逐時資料的克利金內插計算大約耗時 5 天，不過，可特別一提的是，此項去一法推估值數據同時也應用於長期氣候資料整集過程所需的補遺參考值，以及網格化不確定性分析，因此原本就已是計畫過程必須產出的基本分析資料，換言之，並不需額外產出資料檢測專用資料。

在長期氣候資料整集計畫下，對於去一法推估值另外進行了修正偏差計算(T1O_BC)，以增進補遺參考值的合適性。本研究亦使用經過偏差修正值的去一法推估值來提供資料檢測，目的是希望在分析過程中同時多一些機會觀察克利金內插值作為補遺參考值的合適性。

(三) 迴歸離群值檢定法與刪除式殘差

在一組記錄資料中，離群值是指顯著不同於其他觀測樣本的資料數據，常被作為辨認資料異常的標的，例如氣候背景範圍檢驗(Climatological test)即是以氣候平均背景值來提取氣候離群值。不過，因為天氣的高變異特性，氣象記錄資料在特殊或劇烈天氣影響下，也常會表現出背離氣候背景範圍的離群現象。本研究鑑於單從氣候背景資訊所得離群值並不能明確區分確實錯誤與可能合理的資料，乃思考藉由具可信度的參考推估值來檢測觀測數據的合理性，並運用迴歸方法由迴歸離群值來尋找可疑的錯誤數據。在兩組互相成對的數

據中，迴歸線可表示此兩組數據的平均統計關係，迴歸離群值是遠離迴歸線的樣本，反應出該樣本成對數據之間的關係與大多數樣本不同，可提供作為檢測錯誤數據的資訊。

迴歸離群程度可用迴歸殘差表示，本研究參考迴歸診斷分析常用指標，使用經標準化的刪除式殘差(或稱學生化殘差, Studentized Residuals)，此殘差指標服從 t 分佈，可幫助標準化定義合理值範圍(Yang and Xu, 2009)。

三、結果分析

本研究所提出的資料品質檢測做法是一種整體性的檢查過程，亦即可一次性地整體檢查所有待檢測的 17 年的資料。其分析執行步驟可整理如下：(1) 求算檢測參考推估值：由克利金內插工具以去一法求算。(2) 求算經標準化的刪除式殘差：逐站逐時分月進行迴歸分析，求算每一筆資料的迴歸殘差值。(3) 分析迴歸殘差值，進行異常值診斷及判定。上述前 2 項步驟屬於分析資料的準備，第 3 項則是本研究探討迴歸離群值檢定法適用性的核心工作。本節將說明此部分實作分析情形。

在求算出殘差值之後，理想上如果能夠找到一個可區辨錯誤異常值的殘差閥值，那麼在第 3 個步驟對於異常值的判定也就只是一個挑出殘差值超出閥值的簡單動作了。換言之，找出可區辨錯誤異常值的殘差閥值才是整個檢測程序的關鍵。本研究運用經標準化、服從 t 分佈的刪除式殘差，目的也是希望能比較容易地訂出可辨認錯誤異常，同時適用於各站、各季節的一致性標準閥值。理論上，刪除式殘差值將大部分落在正負 3 之間，刪除式殘差值太大者，則為可疑值。但是，如何認定殘差值太大的量化閥值仍不能簡單論斷，需要經過實作分析來幫助正確認識。

對於本計畫所做 110 站共 17 年所有小時氣溫資料分析中，絕大多數的刪除式殘差值是在正負 5 以內，迴歸離群程度小。以台北站的一個分析為例(圖 2)，由資料散布圖可見所分析 17 年的 2 月份早上 8 點氣溫實測資料與克利金估計值皆相近，迴歸離群值大部分落在正負 2 之間，只有幾筆絕對值超過 3，但也不及 5。由於台北站是有人駐守的測站，資料經過較為慎密的檢查流程，出錯相對機會較小，透過這些正確性較高的測站的刪除式殘差值分布圖，可幫助我們對合理殘差值域的認識。

刪除式殘差值大於正負 5 以上可說是迴歸離群程度較大的，由散布圖分析通常亦可見這些資料點脫離於群體分布密集區，尤其是殘差值大於 10 以上的更是明顯遠離迴歸線。以桶後及檳榔 2 個自動站的分析為例(圖 3、圖 4)，可見大部份資料的刪除式殘差值皆落在正負 2 內較小的值域，而幾筆超過 10 以上的離群資料，經過人工覆檢也確認為錯誤記錄。不過，雖然許多刪除式殘差值大於正負 10 以上的觀測值可被證實為錯誤記錄，我們也發現有所例外，例如焚風天氣的影響就可能使迴歸離群程度衝高。

對於可區辨錯誤異常值的殘差閥值，如果要能達到同時適用於各站、各季節的一致性標準，我們經過多個案例的分析結果顯示刪除式殘差值大於正負 20 以上的觀測值可被證實全為錯誤記錄。雖然此迴歸離群閥值感覺過大，但在這個閥值的設定下，我們有效地正確過濾出 54 筆錯誤值，包含許多誤植為 0 度的觀測值，及明顯不合理的過高或過低值。與只給定一個高低範圍檢查大型錯誤的方法相比，這個看似過大的閥值不僅已可包含大型錯誤的檢查，並能更多的提列出明顯錯誤值。

錯誤資料的檢查常常不是能一網打盡的事，在技術上只能儘量的發展或設計可增加有效檢查錯誤量的方法，上述 20 這個明顯過大的殘差閥值顯然有再下調的空間，但必須面對如何在高迴歸離群值中區分可能合理值與真正錯誤值的問題。從個案分析中，我們知道許多刪除式殘差值大於正負 10 以上的離群值是錯誤資料，因此嘗試把區辨錯誤異常值的殘差閥值下調至 10，但加入該觀測值需同時滿足氣候離群值的條件，設定其離群程度需大於 4 個標準差。這樣的設計仍是想先提取較明顯離群值的概念，把可能被區分的錯誤資料先提取出來。下調閥值後共提列出 289 筆可疑離群值，經過人工覆檢比對，幾乎都被確認為錯誤值，少數幾筆仍未能區分的合理正確資料為焚風個案。

在此研究之前，本計畫對於資料的檢查乃以氣候離群值為主，且檢查過程需配合人工參考輔助圖型來進行診斷。本研究探討迴歸離群值檢查法的可行性，目前初步雖然只有先設較大的閥值提取較明顯離群值，分析結果已顯示其良好成效，不只所檢查出的錯誤值記錄可完全含蓋並超出先前做法所得結果，整體性的檢查設計也大幅簡化查核程序。

四、討論與結語

本研究利用迴歸離群值來辨識觀測資料中的錯誤記錄值，初步分析結果已顯示其可行性，為求進一步降低誤判風險及儘可能減少人工核對的必要性，後續仍需要繼續尋求強化及改進，例如，我們已知在某些地理背景因素下，使用鄰近站資訊的克利金內插方法或有成效表現不佳的可能(詹等，2015；薛等，2015)，未來將考慮同時引入數值模式結果來提供另一組資訊獨立的檢測參考推估值，以強化檢測系

統。另外，如何充分利用測站所觀測其他氣象要素協助資料檢測，亦是接續努力的方向。

五、參考文獻

Yan, X. and Su, X.: Linear Regression Analysis: Theory and Computing,World Scientific, Singapore, 2009.

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。

陳雲蘭、陳品妤、詹智雄、沈里音、馮智勇、劉家豪、林佑蓉，2014：台灣自動氣象站氣溫資料補遺方法探討及網格化分析。103 年天氣分析與預報研討會論文彙編，中央氣象局。

馮智勇、劉家豪、陳雲蘭，2012：客觀分析法地面溫度案例分析。天氣分析與預報研討會論文彙編，中央氣象局。

詹智雄、陳雲蘭、馮智勇、劉家豪，2015：「臺灣長期氣候資料整集分析」計畫研究(3) —大台北區測站氣溫空間關係探討。天氣分析與預報研討會論文彙編，中央氣象局。

薛宏宇、詹智雄、陳雲蘭、馮智勇、劉家豪，2015：「臺灣長期氣候資料整集分析」計畫研究(4) —探討 NWP 資料對檢測氣溫觀測合理性的可應用性。天氣分析與預報研討會論文彙編，中央氣象局。

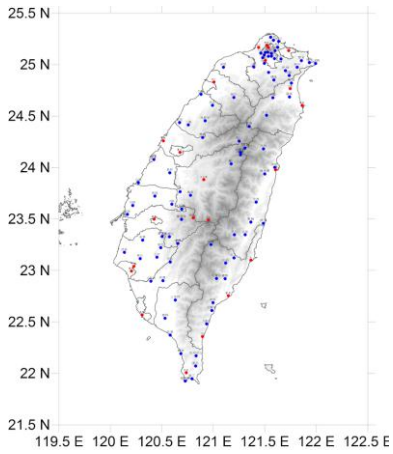

圖 1：本研究按資料高齊備率原則所挑選出的 110 個網格化參考站。

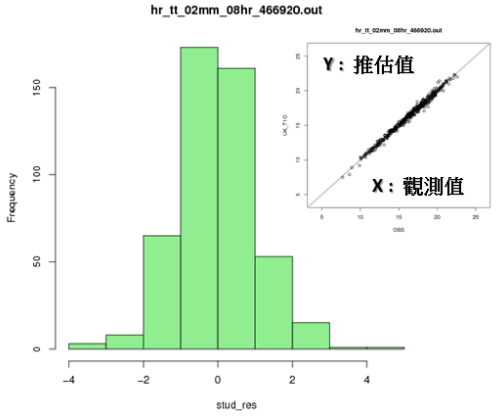

圖 2：以台北站 2 月份早上 8 時的資料為例，長條圖表示本研究所進行殘差分析在各值域的次數分布。嵌入的散布圖可表現觀測值與資料檢測所使用參考推估值的對應情形。

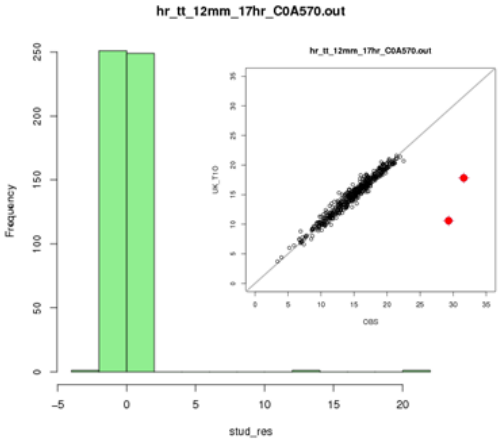

圖 3：同圖 2，但為桶後自動站，分析案例為 12 月份 17 時。

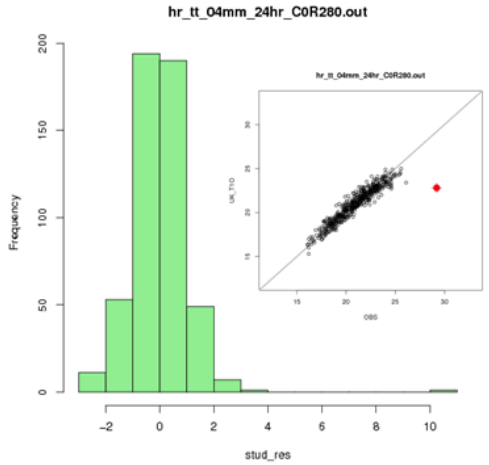

圖 4：同圖 2，但為檳榔自動站，分析案例為 4 月份 24 時。