國立政治大學統計研究所

碩士論文

多重群集的偵測研究

A Study of Methods for Detecting Multiple Clusters

研究生：黃柏誠

指導教授：余清祥 教授、蔡紋琦 教授

中華民國一百零一年六月

謝辭

這篇論文的完成除了感謝老闆指導外，還必須要感謝小馬學長及蔡丞庭學長。小馬學長仿若我的第二個老闆，除了給予我空間統計上的指導外，亦分享他對統計的一些觀點，使我對統計這門學問有更深的看法；蔡丞庭學長便有如苦海明燈般，對於我不太懂的程式都能即時給予建議，讓我能即時完成老闆對我的要求，雖然他有著神龍見首不見尾之稱，但幸運的是每次都能聯絡上他。

除了小馬學長及蔡丞庭學長為本篇論文首要功臣外，還必須感謝我大學同學林育德、鳥勳、VOLVO 哥及囧柏。這些好人好事代表能夠在百忙之中聽我發牢騷及陪我玩樂，讓我擁有短暫的快樂後，能夠繼續面對我的論文。在政治大學這短暫的求學過程中，感謝這些人的陪伴，讓我的求學生涯顯得更加有趣，在未來的道路上也希望大家不論是就職或求學都能順利。

摘要

檢測某些地區是否有較高的疾病發生率，亦即群集(Cluster)現象，是近年來空間統計(Spatial Statistics)在流行病學的主要應用之一，常見的偵測方法包括 SaTScan (Kulldorff, 1995)及 Spatial Scan Statistic (Li et al., 2011)。這些方法多半大都採用一次性偵測，也就是比較疑似群集之內外相對風險(Relative Risk)，如此確實可提高計算效率，同時檢視所有疑似群集。然而，一次性偵測會受到群集外其他發生率較高群集的影響，對於相對風險較小群集的偵測能力過於保守(Zhang et al., 2010)。

本文以多重群集偵測為研究目標，以逐次分析的方式修正 SaTScan 等群集偵測方法，逐一篩選出發生率較高的顯著群集，並探討逐次分析在使用上的時機及限制。除了透過電腦模擬，測試逐次群集分析的改進效果，我們也分析臺灣地區的癌症死亡率，比較偵測結果的差異。研究發現，逐次群集偵測確實能提高相對風險較小群集的偵測能力，像是在相對風險不大於 1.6 的群集時尤其有效，但若相對風險大於 1.6 時，SaTScan 的偵測能力不受多重群集的影響。

關鍵詞：群集偵測、空間統計、逐次分析、電腦模擬

Abstract

Cluster detection, one of the major research topics in spatial statistics, has been applied to identify areas with higher incidence rates and is very popular in many fields such as epidemiology. Many famous cluster detection methods are proposed, such as SaTScan (Kulldorff, 1995) and Spatial Scan Statistic (Li et al., 2011). Most of these methods adapt the idea for comparing the relative risk inside and outside the suspected clusters. Although these methods are efficient computationally, clusters with smaller relative risk are not easy to be detected (Zhang et al, 2010).

The goal of this study is to apply the idea of sequential search into SaTScan, in order to improve the power of detecting clusters with smaller relative risk, and to explore the limitation of sequential method. The computer simulation and empirical study (Taiwan cancer mortality data) are used to evaluate the sequential SaTScan. We found that the Sequential method can improve the power of cluster detection, especially effective for the cases where the clusters with relative risk not greater than 1.6. However, the sequential method also suffers from identifying false clusters.

Keywords：Cluster detection, Spatial statistics, Sequential method, Computer simulation

目錄

表目錄

圖目錄

第一章　緒論

第一節 研究動機

在流行病學的研究資料裡，往往包含大量的地理訊息，但在傳統的流行病學的討論中，此訊息往往較不受研究者重視。隨著地理資訊系統(Geographic Information System, GIS)及空間統計(Spatial Statistics)的蓬勃發展，流行病學結合空間上的分析越來越普遍，而檢測某些地區是否有較高的疾病發生率，亦即群集(Cluster)現象，是空間統計在流行病學上主要的應用之一。

Kulldorff and Nagarwalla (1995)的 SaTScan 及 Tango and Takahashi (2005)的 FlexScan 為目前空間統計檢測群集的主要方式。SaTScan 採用圓形窗格堆疊的方式搭配概似函數的想法找出顯著的區塊，在檢定力上比過去方法有效，但在群集形狀上以圓形及橢圓形偵測效果較佳。FlexScan 以相鄰區為連結決定區塊並搭配概似函數的想法找出顯著的區塊，所以在檢定非圓形群集上較具彈性，但計算上較為耗時。這些方法被除了被應用在流行病上，亦在人文、地質、生態、天文、地震等等，例如：以流行病學而言，空間中的群集可以代表某些疾病發生率較高的地區，醫療資源及預防措施應該往這些相對風險(Relative Risk)較高的區塊集中；以政治版圖來看，若某一政黨得票率在某些地區較高，敵對政黨就必須在這些區塊特別注意。

多數的群集偵測方法採取一次性偵測，比較疑似群集之內外相對風險，同時檢視所有疑似群集，若研究目標為找出最顯著群集，如此確實可提高計算效率。然而，如果目標在於找出所有顯著群集，相對風險較小群集會受到其他發生率較高群集的影響，一次性偵測在相對風險較小群集會過於保守(Zhang et al, 2010)。

假定空間存在兩個群集，群集中心分別為(4.5,4.5)及(12.5,12.5)，相對風險分別 1.40 及 2.0，非群集區塊相對風險為 1.0，群集占研究區域面積比例為 2.25% 及 24.25%，並將研究區域設定為 20*20 個單位正方形所組成的方型研究區域，

相鄰區塊的距離為一單位長，共有 400 個區塊，每一個區塊具有相同的人口數一萬人，所以總人口為 400 萬人。並以普瓦松隨機生成觀察值(λ=10* Relative Risk)，並使用 R 套件中 SpatialEpi 模擬 1000 次。從圖 1-1 可以看出一次性偵測在相對風險較小群集過於保守。

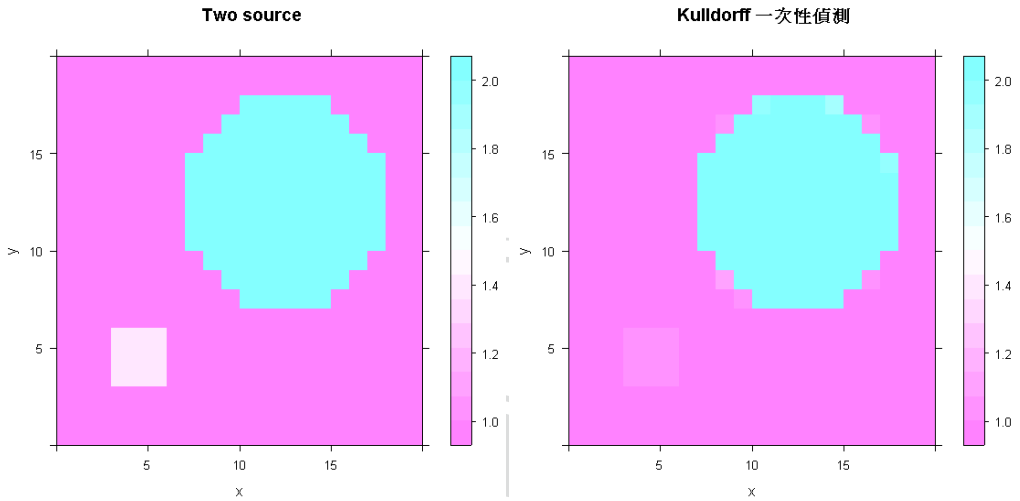


圖 1-1、Kulldorff 兩個群集偵測結果示意圖

註 1：左圖為原始設定群集示意圖；右圖為模擬 1000 次 kulldorff 偵測結果示意圖

目前對於多重群集的檢測方式主要有兩種方式，第一種為上面提到的一次性偵測，比較疑似群集之內外相對風險，同時檢視所有疑似群集；另外還有一種一次性偵測，乃透過外部資訊得知群集數目，在檢定時便同時考慮群集個數，所以在偵測時相對風險較小群集不會受到其他發生率較高群集的影響；第二種為逐次分析，亦即考慮顯著群集存在下檢測其它群集，偵測上亦能排除顯著群集的影響。所以在多重群集的存在下，如何正確檢測出所有顯著群集，為目前多重群集偵測主要研究的目標。

第二節 研究目的

在上一節研究動機提到，目前在多重群集的偵測上，以一次性偵測及逐次分析為主。一次性偵測若在得知群集個數下，透過修改 Scan Statistic，確實能避免

風險較小群集受到其他發生率較高群集的影響，然而在大多數的情況下，無法得知準確的群集數目，此時利用一次性偵測同時比較所有疑似群集的內外風險，相對風險較小群集會受到其他發生率較高群集的影響，在檢測上會過於保守。

以降低顯著群集影響的方向思考，此問題類似於矩陣在尋找特徵值的過程中，透過 Power Method 找到最大特徵值後，如何尋找其他特徵值。此問題可以藉由特徵值平移的概念，降低較大特徵值的影響後，重新尋找新的特徵值，新的特徵值加上原始平移量及為原始特徵值(Lloyd et al., 1997)。若仿照尋找特徵值的概念，在檢定的過程中，將最顯著的群集移除資料，減少對其他群集在偵測時的影響，對新的研究區域檢測以尋找其他的群集，此概念即為逐次分析。

本文主要以多重群集偵測為研究目標，以 Kulldorff and Nagarwalla (1995)的 SaTScan 作為檢測群集的方式，並配合逐次分析的方式修正群集偵測方法。雖然 Zhang et al. (2010)有提出類似的想法，但主要是針對方法的比較及對於群集檢定力的改善，這一部分說明在第二章文獻探討的部分會有更詳細的說明。除此之外，本文亦提供逐次分析在方法上的限制及特點並配合電腦模擬測試逐次群集分析的改進效果，也進一步分析臺灣地區的癌症死亡率，並比較偵測結果的差異。

第二章　文獻探討

在前面一章緒論裡，初步介紹了群集的概念，指出在多重群集偵測上現有方法可能的限制，本章第一節將先整理空間統計中群集偵測的常見作法，包括對於群集類型的分類，並且詳細介紹常用的偵測方法 SaTScan；第二節則介紹多重群集的處理方法，並討論方法適用時機。

第一節 群集檢測方法

在檢測群集的問題上，Marshall (1991)指出兩個主要的議題，第一項為研究區域中是否存在不尋常高的事件數，第二項則是研究區域中是否發生群集及發生的位置。一般而言，群集檢測方法依目的的不同可分為總體檢定(Global Test)、局部檢定(Local Test)及焦點檢定(Focused Test)。

(一) 總體檢定

總體檢定是指在不考慮群集位置的狀況下，檢定整體研究區域是否有群集的傾向。而使用空間自相關(Spatial Autocorrelation)的量度以檢驗空間單元與其相鄰的空間單元的屬性間是否具相似性是較為典型的做法。

Moran I 與傳統在計算皮爾森相關係數類似，差別在於多考慮一個加權矩陣，通常可以以 0 或 1 代表空間單元間是否相鄰或採兩空間單元間距離的倒數作為加權矩陣(Moran, 1950)。Moran I 的值介於-1 到 1，值越接近 1 表示正向空間自相關程度越強，Moran I 值越接近 0，則相鄰之空間單元間相關性低。當空間型態為常態隨機分布時，對於「空間是否有自相關性存在」，可進一步檢定 Moran I (Cliff and Ord, 1981)。此方法除了可應用在連續型資料亦可應用在離散型資料，在使用上較為廣泛。

但此方法在研究區域中人口密度有異質性(Non-homogenuous)時，Moran I 判

定空間是否有自相關性存在會受到人口大小影響，所以為了解決此問題，Oden (1995)的I^*_{pop}將人口納入考量並重新改寫 Moran I 及調整 Moments，配合計算 Z-scores 或蒙地卡羅亦可檢定空間是否有自相關性。Waldhör (1996)則在檢定 Moran I 時，在變異數上允許每一區不同，而每一區的變異數與該區人口數成反比，重新對 Moran I 的 Moments 作調整。

另外，Tango (1995)的 Excess Events Test 採用距離矩陣(Distance Matrix)衡量區域間是否相近，計算第 i 區事件數超過第 i 區期望發生數與第 j 區事件數超過第 j 區期望發生數，兩者相乘的加權總和，權重為指數型函數。由於必須決定參數 λ(與空間群集大小有關)，λ 一般不知道，所以透過檢定不同的 λ 值以便判斷是否有群集現象，但又衍生出重複檢定上的問題，故 Tango (2000)提出 Maximized Excess Events Test(MEET)修正此問題，在檢測形成群集的現象上檢定力較佳 (Song and Kulldorff, 2005)。Tango (2000)的 MEET 及 Oden (1995)的I^*_{pop}兩種方法在檢測 Global Clustering 檢定力較佳，而原始的 Moran I 較差(Jackson et al., 2009)。

(二) 局部檢定

局部檢定用來偵測在研究區域中發生群集現象的位置。在總體檢定中，可以評估 Global Clustering，在局部檢定中亦有類似作法。檢測局部區域自相關程度 Local Indices of Spatial Association(LISA)，以 Moran I 而言，此時不再只計算一個值，必須針對每一個地理單位(Geographic Unit)如：鄉、鎮，計算 Moran I。或是以概似函數配合蒙地卡羅檢定研究區域是否存在群集，如 Kulldorff and Nagarwalla (1995)的 SaTScan 及 Tango and Takahashi (2005)的 FlexScan。

Kulldorff and Nagarwalla (1995)的 SaTScan 是目前被廣為使用的方法之一，同時解決過去在多重檢定及人口密度上的異質性的問題。此方法主要採用圓形窗格堆疊的方式搭配概似函數的想法找出顯著的區塊，在檢定力上也比過去方法更

為有效(Kulldorff et al., 2003; Song and Kulldorff, 2003)，但此方法對於非圓形或橢圓形的群集形狀，檢定力較差(Kulldorff et al., 2006)。有鑑於此，陸陸續續有其他學者針對奇形怪狀的群集提出解決之道，如 Demattei et al. (2007)及 Cucala (2009)從維度縮減的角度，將 Spatial Data 轉換單一維度，並透過 Scan Statistic 進行單一維度的分析，但此方法在轉換的過程中可能會造成點順序上的不一致，因而偵測上產生過多的 False Positive。

Tango and Takahashi (2005)的 FlexScan 與 SaTScan 最大的不同點在於 SaTScan 採用圓形窗格堆疊的方式而 FlexScan 以鄰區相連結的方式執行群集偵測，所以為了避免找到局部解，在 K 個鄰近區塊上都必須討論所有的可能性，因此在對奇形怪狀的群集偵測時有不錯的效果(王泰期, 2006)，但有計算繁複的問題，除此之外 FlexScan 只在群集大小不大的情況下偵測效果較佳(Tango and Takahashi, 2005)。

(三) 焦點檢定

焦點檢定用於檢測某一特定位置周圍是否有顯著較高的事件發生率，在使用時，通常需要了解疾病的影響範圍、傳染的強度等等，由於在近幾年並沒有新方法的提出，近期焦點檢定的方法上多採用 Stone (1988)的 Stone's Test 及 Diggle (1990)的 Diggle's Test，以 Stone's Test 為主(Auchincloss et al. 2012)。Diggle's Test 為一種適合度檢定(Goodness-of-fit Test)，比較原始資料的空間分布及在控制可能的汙染源位置下所生成的分布是否一致(Diggle, 1990; Diggle, 1994)，但其主要用於 Individual-level Data。Stone (1988)提出 Maximum Likelihood Ratio (MLR)及 Poisson Maximum (Pmax) tests，兩種方法皆假設隨著離可能的汙染源位置越遠風險就越小的情況下使用 Isotonic Regression Estimator。Stone 提到 MLR Test 的檢定力通常大於 Pmax 的檢定力，但當群集範圍較小時，Pmax 的檢定力會大於 MLR test 的檢定力。由於在 MLR 檢定方法中，估計各區塊相對風險值可能會出現小

於 1 的值，所以 Bithell(1995)加入相對風險至少為 1 的條件，發現 MLR 在此狀況下 P-value 較小，而 Pmax 無明顯變化。

由於本文探討群集分析採用 Kulldorff and Nagarwalla (1995)的 SaTScan，所以底下介紹 SaTScan 在群集分析時的步驟如下：

Step1：建立 I 個格子點，使這些格子點能包含全部的研究區域。

Step2：訂定擴散半徑 R，以格子點為中心向外圓形擴張半徑為 R，找出多個圓形區塊，作者建議最大半徑 R 所涵蓋總人口數不要超過 50%。

Step3：計算每個圓形區塊的概似比，以獲得最大概似比。

Step4：針對每個格子點重複 Step2 到 Step3

Step5：利用 Monte Carlo 檢定哪些區塊顯著。

透過上述步驟，即可檢定研究區域是否有群集及群集的位置。SaTScan 在面對群集的相對風險較小時，主要是以群集面積來判斷，亦即必須要達到一定的區塊檢定才會顯著，如相對風險為 1.2 下，群集區塊必須達研究區域 30%以上，Power 才會達 90%以上；隨著相對風險的提高下，群集區塊所佔研究區域就不需達到 30%，以相對風險為 1.6 而言，只要達 15% Power 便達 90%以上。所以當群集相對風險很低，隨著所占研究區域面積比例下降，檢定力也會隨之下降，且檢定顯著的群集都容易涵蓋過多的 False Positive。在提高相對風險後，偵測涵蓋過多的 False Positive 才會得到改善。在下一章研究方法會介紹 SaTScan 的檢定統計量並改寫逐次分析下，SaTScan 的檢定統計量。

第二節 多重群集檢測方法

在一般情況下，研究區域往往存在不只一個群集，如何使用局部檢定找出所有顯著群集，為局部檢定中一個重要的議題。目前在局部檢定中，多重群集的偵測方式以一次性偵測及逐次分析為主。一次性偵測若在得知群集個數下，透過改

寫 Scan Statistic，確實能避免風險較小群集受到其他發生率較高群集的影響，然而在大多數的情況下，無法得知準確的群集數目，此時利用一次性偵測同時比較所有疑似群集的內外風險，相對風險較小群集會受到其他發生率較高群集的影響，在檢測上會過於保守(Zhang et al., 2010)。故 Zhang et al. (2010)提出以 Sequential Method 亦即逐次分析，如此確實能修正相對風險較低的群集的 Type I Error。

然而此方法無法消除其它潛在的群集對最顯著群集的影響 (Li et al., 2011)，故 Li et al. (2011)提出新的 Spatial Scan Statistic，此方法屬於一次性偵測的一種，透過逐次分析初步決定潛在群集個數後，進一步透過作者的 Spatial Scan Statistic 重新檢定。另一方面 Wan et al. (2012)提出新的演算法，在不規則的多重群集上有不錯的檢定力，但此方法在運算複雜度上 $O(N^4)$，相當耗時且此方法只能針對累計型資料作分析，所以對於 Individual-level Data 必須先轉換。

在研究動機裡提到，逐次分析以降低顯著群集的影響為思考方向，此問題類似於從矩陣中尋找特徵值的過程，透過 Power Method 找到最大特徵值後，如何尋找其他特徵值。此問題可以藉由特徵值平移的概念，降低較大特徵值的影響後，重新尋找新的特徵值，新的特徵值加上原始平移量及為原始特徵值。若仿照尋找特徵值的概念，在檢定的過程中，將最顯著的群集移除資料，減少對其他群集在偵測時的影響，對新的研究區域檢測以尋找其他的群集，此概念即為逐次分析。

雖然此概念已經被人提出來，但 Zhang et al. (2010)主要針對在採用逐次分析的前提下，比較移除最顯著的群集、移除最顯著群集並用平均發生事件數取代、或是移除最顯著群集並將周圍區塊一併移除，這三種方法對於相對風險較低的群集 Type I Error 的修正，並建議直接移除最顯著的群集在群集偵測結果上便可達到修正的目的。但對於此方法有何使用限制及在何種情況下使用較佳並未詳細討論，所以本文即著重在這兩方面上進行討論。本文以 Kulldorff and Nagarwalla (1995)的 SaTScan 作為檢測群集的方法，以逐次分析的方式修正 SaTScan，並在第四章的電腦模擬分析比較改進的成果。

第三章　研究方法及假設

在上一章最後一段有提到逐次分析的主要想法與尋找特徵值過程類似，此概念基本上可套用在 SaTScan 或是 FlexScan 等群集偵測的方法上。雖然 FlexScan 在群集的形狀上，相對於 SaTScan 較具彈性，但計算上較為耗時，所以本文以 Kulldorff and Nagarwalla (1995)的 SaTScan 作為群集檢測方式。由於在逐次分析的概念上與 Zhang et al. (2010)的概念類似，所以以下將會參考 Zhang et al. (2010)的模型。

第一節 空間統計模型

在大多數疾病群集研究中，經常為累計型資料(Aggregate Data)，像是臺灣地區常以縣市、鄉鎮市區、村里等為單位。首先將研究區域分割成 K 個小區塊，也就是地理分割，通常會是普查研究區塊(Census Tract)。考慮 Kulldorff and Nagarwalla (1995)的 SaTScan，以下令 Z 表示選取的區塊，G 代表研究區域，o(Z)表示區塊內發生的事件數，o(G)表示研究區域內發生的事件數，n(Z)表示區塊內的人口數，n(G)表示研究區域內的人口數。虛無假設可寫為區塊 Z 內的事件發生率與區塊與 Z 外的事件發生率相同，對立假設可寫為區塊 Z 內的事件發生率大於區塊 Z 外的事件發生率。所以 SaTScan 檢定方法的統計量可表示如下：

$$S = \max_Z \frac{\left(\frac{o(Z)}{n(Z)}\right)^{o(Z)}\left(\frac{o(G)-o(Z)}{n(G)-n(Z)}\right)^{o(G)-o(Z)}}{\left(\frac{o(G)}{n(G)}\right)^{o(G)}} I\left(\frac{o(Z)}{n(Z)} > \frac{o(G)-o(Z)}{n(G)-n(Z)}\right), \text{otherwise } S=1$$

其中$\frac{o(Z)}{n(Z)}$代表區塊 Z 內的事件發生率，$\frac{o(G)-o(Z)}{n(G)-n(Z)}$代表區塊 Z 外的事件發生率，o(.)可被視為一隨機變數，過去研究中大多假設 o(.)服從二項分配或普瓦松分配。

若空間存在多重群集，並採取逐次方法(Sequential Method)進行群集檢測，則應如何改寫檢定統計量，以下以兩個群集為例。在群集沒有相互重疊的情況下，對於最顯著的群集 Z，其檢定統計量如下：

$$S = \max_Z \frac{\left(\frac{o(Z)}{n(Z)}\right)^{o(Z)} \left(\frac{o(G)-o(Z)}{n(G)-n(Z)}\right)^{o(G)-o(Z)}}{\left(\frac{o(G)}{n(G)}\right)^{o(G)}} I\left(\frac{o(Z)}{n(Z)} > \frac{o(G)-o(Z)}{n(G)-n(Z)}\right)$$

對於次要顯著群集，其檢定統計量如下：

$$S = \max_{Z'} \frac{\left(\frac{o(Z')}{n(Z')}\right)^{o(Z')} \left(\frac{o(G')-o(Z')}{n(G')-n(Z')}\right)^{o(G')-o(Z')}}{\left(\frac{o(G')-o(Z')}{n(G')-n(Z')}\right)^{o(G')-o(Z')}} I\left(\frac{o(Z')}{n(Z')} > \frac{o(G')-o(Z')}{n(G')-n(Z')}\right)$$

，$G' = G - Z$

若存在更多群集，則以此類推。

第二節 逐次分析

在第二章文獻探討的最後一段有稍微闡述逐次分析的想法，並於上一節裡提到，在群集彼此不重疊的情況下，可以寫出逐次分析的檢定統計量。關於逐次分析時的操作步驟將於本節詳細介紹。

逐次分析法概念亦如同迴歸分析中的逐步回歸。逐步回歸中的 Forward-entry Procedure 是選定一個標準，按自變數對 y 的貢獻大小由大到小依次挑選進入模式，每選入一個變數進入模式，則重新計算模式外各自變數對 y 的貢獻，直到方程外變數均達不到入選標準，沒有自變數可被引入模式為止。若仿照逐步回歸檢測變數步驟，逐次分析在檢測群集時，可採取以下步驟：

Step1：以 SaTScan 檢測研究區域是否存在顯著群集

Step2：將檢定最為顯著的群集自研究區域中移除

Step3：對新的研究區域重新使用 SaTScan 進行檢測

Step4：重複 Step2 至 Step3 直到沒有任意群集被檢測出來

第三節 本文研究特色

本文主要強調逐次分析及一次性偵測主要的使用時機及使用上可能的限制，不同於 Zhang et al. (2010)。Zhang et al. (2010)雖然提出此概念並應用在群集的偵測上，但著重於在採用逐次分析的前提下比較移除最顯著的群集並用平均發生事件數取代、或是不取代直接移除這兩種方法上，對於相對風險較低的群集在 Power 上的修正，並得到直接移除最顯著的群集在群集偵測結果上便可達到修正的目的。在研究動機時便有談到當有一群集占研究區域比例面積過大時，便可能在檢測群集產生過於保守的問題。

本文於第四章電腦模擬將有系統的探討群集占研究區域比例面積、群集個數、群集相對風險這三個因素在檢測群集時可能會產生的影響，並透過逐次分析與一次性偵測，比較兩者在檢測群集結果上的差異。

第四章 電腦模擬比較分析

本文的目標在於如何檢測研究區域中的多重群集，藉由 Kulldorff and Nagarwalla (1995)的 SaTScan 可採取一次性偵測或逐次分析兩種策略進行群集檢測。第二章的文獻探討裡提到，採取一次性偵測的方式有兩種，一種透過外部資訊獲得準確群集數目，在檢定時同時考慮多個群集，如此可避免相對風險小的群集受其它發生率較高群集的影響，然而在大多數的情況下，無法得知準確的群集數目，此時利用一次性偵測同時比較所有疑似群集的內外風險，相對風險較小群集會受到其他發生率較高群集的影響，檢測結果會較為保守。而逐次分析透過移除最顯著的群集，在檢測的其它的群集不受其影響，盡可能使 P-value 回到正常值。

由檢定統計量來看，無論是單一群集或是多重群集，群集的相對風險及群集所占研究區域的比例皆會影響群集偵測時的正確性，有關不同相對風險下 SaTScan 的群集偵測能力已有完整的討論，所以本章第二節會著重在群集所占研究區域的比例對群集偵測的影響上作討論，並先假設單一群集的情況作為比較的基準，進一步探討在兩個群集時，群集間相對風險及個別群集所占研究區域的比例對群集偵測時的影響。此外本文主要以多重群集為研究目標，所以在第三節會將群集個數增加，進一步探討兩個群集時的推論是否適用於群集個數很多的情況，並比較一次性偵測及逐次分析。

第一節 偵測結果的衡量方式

本文主要以 Power、Positive Predictive Value(PPV)、Sensitivity 及 Error Rate 作為衡量偵測結果的方式。因為 Power、PPV、Sensitivity 皆為過去在疾病群集偵測上常用的方式，本節主要針對 Error Rate 進行說明及為何使用此一衡量方式。

表 4-1、Compare Criteria

		Predicted class	
		Yes	No
Actual class	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

TP 為群集格點且同時被偵測出來，TN 為非群集格點且同時未被偵測出來，FN 為群集格點但卻未被偵測出來，FP 為非群集格點但卻被偵測出來。在單一群集的情況下，通常以檢測出的群集與原始設定群集作交集，若交集數大於 0，則以此群集作為與原始設定群集比較的依據，此時的 FP 為非群集格點但卻被偵測出的個數；但在多重群集時，檢測出的群集有兩種以上的身分，亦即涵蓋的格點可能同時包含多個原始設定的群集，此時要討論檢測的群集該屬於那個原始設定群集較為困難，所以在這種情況下，FP 認定的方式如下：將偵測出的群集與原始設定的某個群集作交集，並將所有偵測出的群集與原始設定的某個群集交集元素個數大於 0 的群集作聯集，成為新的集合，這個新的集合扣除與原始設定群集交集的元素，剩餘元素即為 FP。假定研究區域存在兩個群集分別為 OC1 及 OC2，其中 OC1：1、2、3、4、5；OC2：6、8、9。偵測出的群集為分別為 DC1、DC2、DC3，其中 DC1：1、5、6、7；DC2：2、3、8、9；DC3：10、11、12。顯然只有 DC1 及 DC2 與 OC1 有交集，所以新的集合為{1、5、6、7、2、3、8、9}，扣除與原始設定群集交集的元素所以 OC1 的 FP 為{7}。

PPV、Sensitivity、Error rate 定義如下：

$$\text{PPV} = \frac{\text{TP}}{\text{TP+FP}}; \text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}; \text{Error rate} = \frac{\text{FN+FP}}{\text{TP+FN+FP}} ;$$

$$\text{Power} = \frac{\text{檢測群集與原始設定群集交集大於 0 的次數}}{\text{模擬次數}}$$

PPV 可以視為偵測結果中包含正確的比例；Sensitivity 可以視為被正確偵測出的

比例。傳統在計算 Error Rate 時，分母會包含 TN，但由於研究區域範圍很大，而群集占研究區域面積比例又不大時，此時 TN 的影響會很大，不利於偵測結果上的比較所以本文在定義 Error Rate 上不考慮 TN。

在第二章文獻探討裡提到，一次性偵測同時比較所有疑似群集的內外風險，相對風險較小群集會受到其他發生率較高群集的影響，檢測結果會較為保守。本文定義內外風險差異縮減比例，作為衡量的標準。

內外風險差異縮減比例 =

$$1-\frac{1}{\sum(\text{檢測出的群集占研究區域面積比例}\times\text{群集相對風險})+1*\text{剩餘研究區域比例}}$$

理論上在計算對疑似群集的影響程度必須要扣除疑似群集占研究區域面積比例，但疑似群集占研究區域面積比例無法得知，所以透過此種方式計算只能獲得最小可能內外風險差異縮減比例。另外，若使用一次性偵測檢測所得出的群集及群集相對風險，計算內外風險差異縮減比例會比實際縮減值小。內外風險差異縮減比例值越小，代表偵測群集時受其它群集影響越小。

另外，流行病學與空間統計所定義相對風險不同，所以本文在相對風險的定義如下：

$$RR=\frac{\frac{c}{E[c]}}{\frac{C-c}{C-E(c)}}$$

c:群集範圍內的觀察值總和；C 為研究區域中所有觀察值總和

第二節 群集所占研究區域面積比例

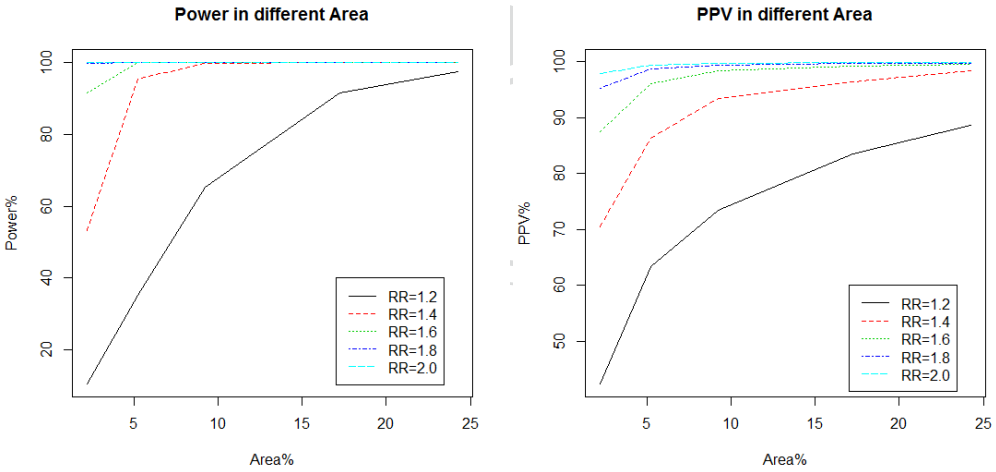
本文在採取電腦模擬分析上採用格點空間資料模型，為了模擬上的方便將研究區域設定為 20*20 個單位正方形所組成的方型研究區域，相鄰區塊的距離為一單位長，共有 400 個區塊，底下舉例說明。將此研究區域以座標平面表示，研究區域的 X 座標由 0 到 20，Y 座標亦為由 0 到 20，每一區塊的座標以區塊內中心點座標表示，離原點最近的區塊座標為(0.5,0.5)，最遠的區塊座標為(19.5,19.5)。

設定每一個區塊具有相同的人口數一萬人，所以總人口為 400 萬人。並以普瓦松隨機生成觀察值(λ=10)。為了避免因隨機亂數不同所造成的差異，本文在電腦模擬分析上皆採取相同一組隨機亂數。由於 SatScan 在電腦模擬上不方便，因此採用 R 套件中的 Spatialepi 進行模擬，並設定顯著水準 α=0.1，模擬 1000 次。

(一)單一群集

前面有提到，群集的相對風險及群集所占研究區域面積比例皆會影響群集偵測結果，過去文獻指出 SaTScan 在群集相對風險 1.6 以上，皆有不錯的偵測能力，而相對風險介於 1.1 至 1.5 表現較為不佳，相對風險介於 1.5 至 1.8 偵測能力逐漸提高。而探討固定相對風險下，群集占研究區域面積比例對群集偵測的影響之前必須先有單一群集的偵測結果作為衡量基準，所以接下來假設單一群集下固定相對風險，討論群集所占研究區域面積比例對偵測結果的影響。

假定群集中心所在座標為(12.5,12.5)，考慮群集的半徑為 1.5、2.5、3.5、4.5，此時群集所占研究區域面積比例為 2.25%、5.25%、9.25%、17.25%、24.25%，而群集範圍內以普瓦松隨機生成觀察值(λ=10*RR)。



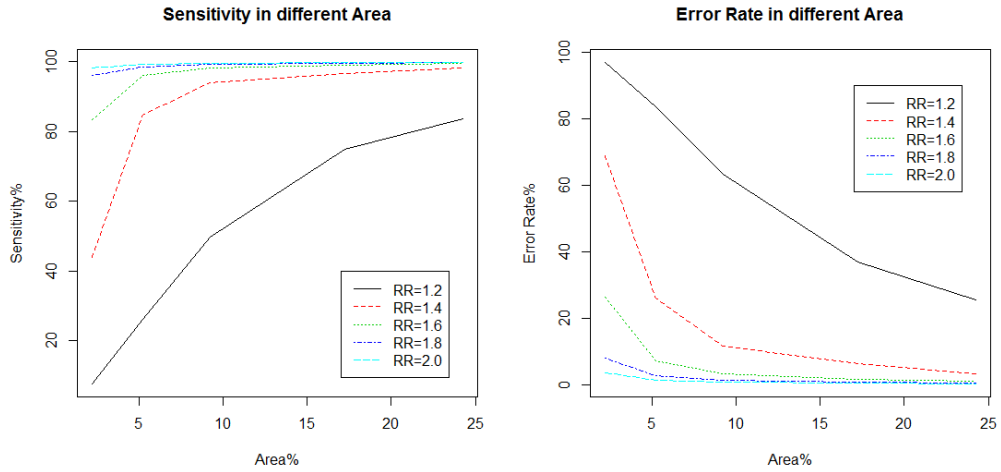


圖 4-1、SaTScan 單一群集偵測結果

註 1：Area%代表群集占研究區域面積比例以百分比計

透過圖 4-1 可以確認，提高群集所占研究區域面積比例有利於群集的偵測，對於相對風險越低的群集，其所占研究區域面積比例必須越高，Error Rate 才會降低到一定值；相對風險 2.0 以上，Area%的增加對 Error Rate 的降低已無太大幫助，亦即群集的相對風險達 2.0 以上，即便 Area%不大，亦能正確的偵測該群集，與過去文獻在相對風險的討論上一致。

(二)兩個群集

在上面單一群集的情況裡得到相對風險越高及群集所占研究區域面積比例越高皆有助於群集的偵測，但當研究區域存在不只一個群集時，採取一次性偵測的方式，對於次要顯著的群集會較為保守，亦即研究區域裡其它群集的存在，對於檢測的群集而言可視為一種門檻，但如果門檻太高將會無法檢測出該群集，至於「門檻」多高才會對檢測結果造成影響？所以接下來將假設在兩個群集下，變動其中一個群集所占研究區域面積比例，且固定另一個群集所占研究區域面積比例，討論一次性偵測的群集檢測結果，並進一步使用逐次分析，比較兩者差異。

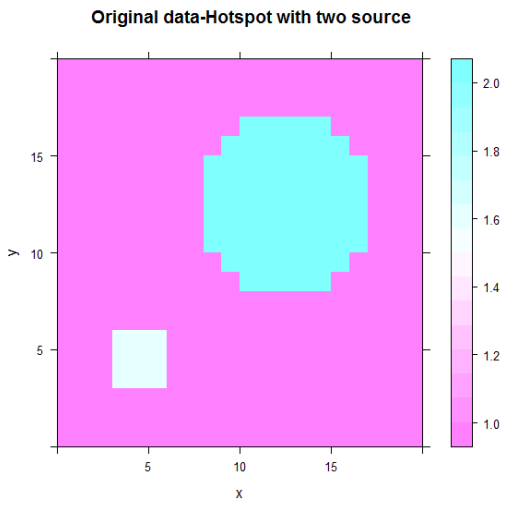


圖 4-2、兩個群集示意圖

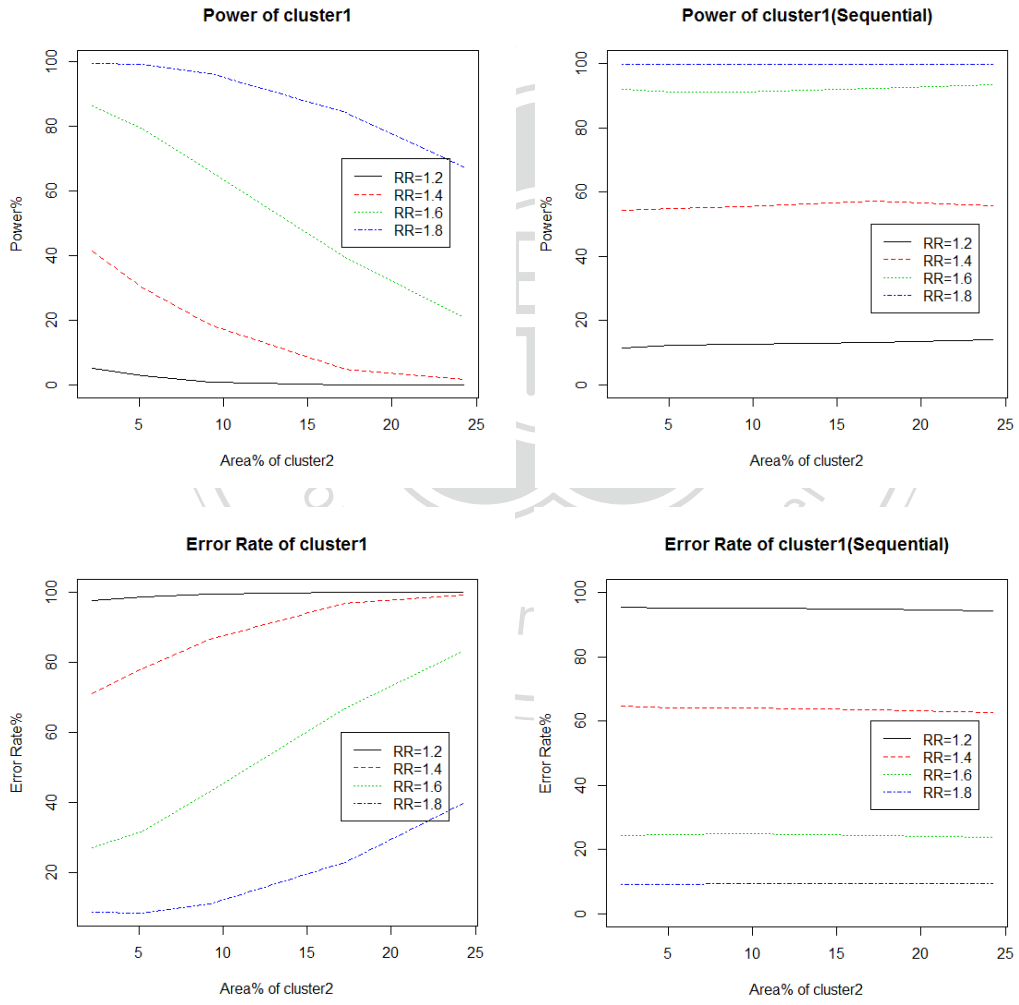


圖 4-3、相對風險較高群集占研究區域面積比例對群集偵測上的影響

註 1：固定 cluster2 的相對風險為 2.0，並變動其所占研究區域面積比例，討論 cluster1 的偵測結果

表 4-2、RR=2.0 所造成內外風險縮減比例

RR=2.0 所占研究區域面積比例	最小內外風險縮減比例	實際內外風險縮減比例
2.25%	2.20%	2.25%
5.25%	4.99%	5.10%
9.25%	8.47%	8.64%
17.25%	14.71%	15.00%
24.25%	19.52%	19.88%

假定 cluster1 中心座標為(4.5,4.5)的半徑為 1.5(群集所占研究區域面積比例為 2.25%)，cluster2 中心座標為(12.5,12.5)，考慮 cluster2 的半徑為 1.5、2.5、3.5、4.5(群集所占研究區域面積比例為 2.25%、5.25%、9.25%、17.25%、24.25%)，而群集範圍內以普瓦松隨機生成觀察值(λ=10*RR)。

由圖 4-3 可以看出當 cluster1 的相對風險為 1.8，cluster2 所占研究區域面積比例達 12.5%以上時，一次性偵測會有過於保守的問題，且 Error rate 的提高主要原因為 FN 過多。另外，透過計算內外風險縮減比例可知對於 cluster1 而言，在估計相對風險時，相對風險將降為 1.60 以下，此時採用逐次分析便能有效讓其回到單一群集的偵測結果。若 cluster1 的相對風險介於 1.4 至 1.6，cluster2 即便所占研究區域面積比例不高(5.25%以下)，配合表 4-2 可知在估計 cluster1 的相對風險時，相對風險將降為 1.36 至 1.51，採用逐次分析在 Power 上能夠改善 10%~15%，在 Error rate 的部分能夠改善 3%~10%。

至於 cluster1 的相對風險為 1.4 以下，逐次分析所能改善的幅度最多回到單一群集的偵測結果，但在單一群集的偵測結果上 Power 及 Error rate 的結果亦不甚理想，以至於並無太大的改善。以下考慮相對風險較小群集所占研究區域面積比例對相對險較大群集會有何影響。

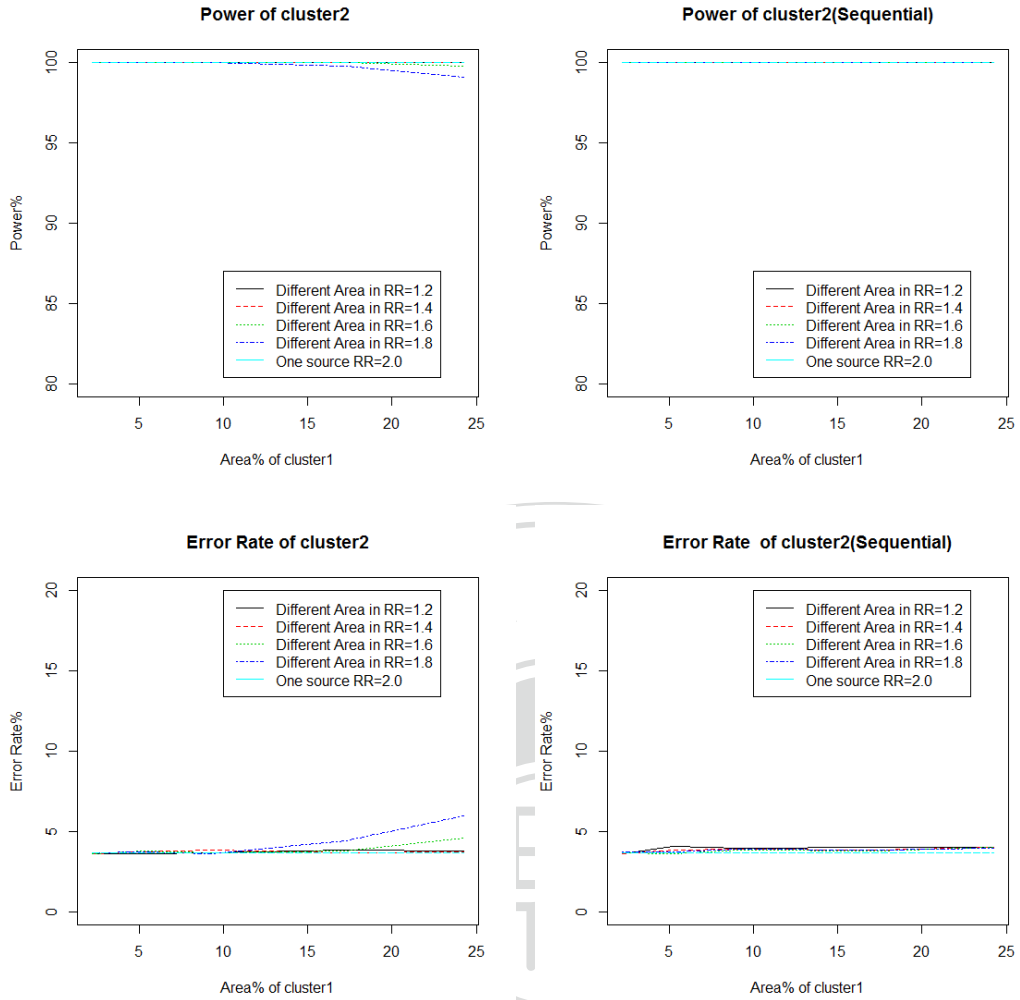


圖 4-4、相對風險較小群集占研究區域面積比例對群集偵測上的影響

註 1：固定 cluster2 的相對風險為 2.0 及固定其所占研究區域面積比例 2.25%，討論 cluster2 的偵測結果

註 2：one source RR=2.0 為假定在單一群集下，SaTScan 偵測結果

假定 cluster1 中心座標為(12.5,12.5)的半徑為的半徑為 1.5、2.5、3.5、4.5(群集所占研究區域面積比例為 2.25%、5.25%、9.25%、17.25%、24.25%)，cluster2 中心座標為(4.5,4.5)，考慮 cluster2 的半徑為 1.5 (群集所占研究區域面積比例為 2.25%)，而群集範圍內以普瓦松隨機生成觀察值(λ=10*RR)。

明顯由圖 4-4 可以看出 cluster1 的相對風險為 1.6 以下即便群集所占研究區域達 20%以上，對 cluster2 在偵測上的影響不大，配合內外風險縮減比例，即可以看出對 clsuetr2 在估算相對風險時，皆在 1.6 以上，所以 Power、Error rate 變化幅度很小都在 1%左右；當 cluster1 的相對風險提高至 1.8 以上，群集所占研究

區域亦要達 15%以上,才能對 cluster2 產生影響,但影響程度亦不大,約 1%~2%,此時使用逐次分析改善效果並不明顯。

表 4-3、相對風險較小群集造成內外風險縮減比例

RR	所占研究區域面積比例	最小內外風險縮減比例	實際內外風險縮減比例
1.2	24.25%	4.63%	4.73%
1.4	17.25%	6.45%	6.59%
1.4	24.25%	8.84%	9.03%
1.6	9.25%	5.26%	5.37%
1.6	17.25%	9.38%	9.57%
1.6	24.25%	12.70%	12.96%
1.8	5.25%	2.06%	2.10%
1.8	9.25%	6.89%	7.04%
1.8	17.25%	12.13%	12.37%
1.8	24.25%	16.25%	16.56%

(三)模擬小結

綜合以上,從比較內外風險來看,相對風險較高的群集所占研究區域面積比例越高,一次性偵測確實不利於相對風險較小的群集;而相對風險較小的群集所占研究區域面積比例越高,也會對相對風險較高的群集偵測結果上產生影響,通常影響一次性偵測結果的程度必須視群集相對風險及其所占研究區域面積比例而定,配合計算內外風險縮減比例下,更可看出上述結果。另外,使用逐次分析確實能夠降低顯著群集的影響,但逐次分析改善幅度最多只能回到單一群集的偵測結果,所以對於相對風險較小的群集(RR=1.4 以下),改善程度相當有限,相對風險介於 1.4~1.6 改善幅度較大。

第三節 群集個數

過去文獻在討論多重群集時，群集個數大都介於兩個到三個，三個以上便較少提及，而以群集數來看，以台灣人有氣喘或呼吸道疾病的比率，群集數便有可能超過三個，所以群集個數亦是本文所探討重點。而在上面電腦模擬中除了討論相對風險大小及群集所占研究區域面積比例對一次性偵測在兩個群集時，偵測的結果及逐次分析所能改善程度。但以上推論在空間存在「多」個群集下是否依然適用，抑或產生不同於兩個群集時所偵測的結果?所以在考慮群集間彼此的距離，及所有群集所占研究區域面積比例，本文在這小節將群集個數擴大至九個，每個群集所占研究區域面積比例為 2.25%，所有群集占研究區域面積比例達 20.25%。前面在兩個群集時有提到逐次分析在相對風險為 1.6 時，改善幅度最大，相對風險為 2.0 以上時，改善幅度較不明顯，所以接下來進一步討論所有群集的相對風險為 1.6 及所有群集的相對風險為 2.0，檢視一次性偵測及逐次分析的偵測結果。

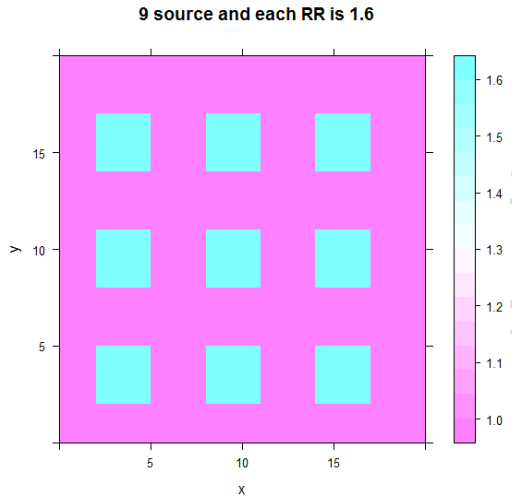


圖 4-5、RR 均為 1.6 所有群集的示意圖

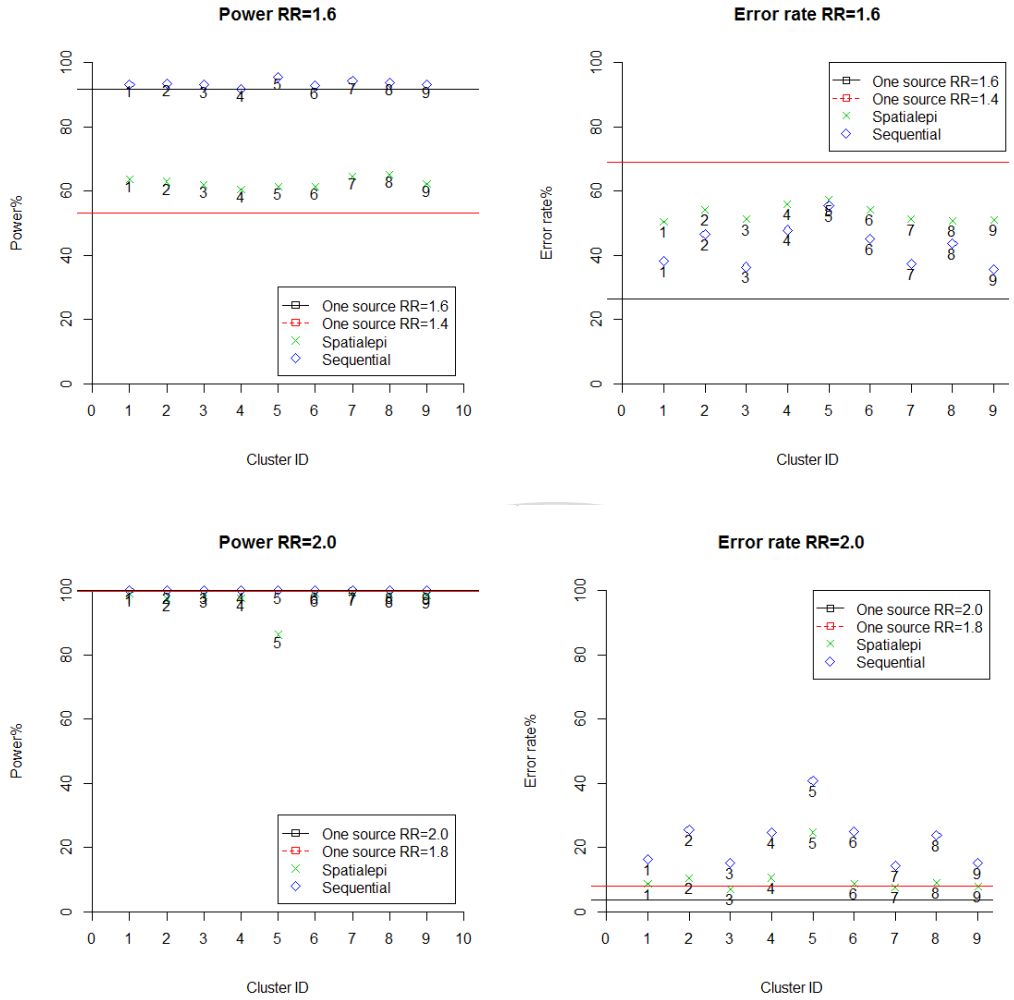


圖 4-6、RR 均為 1.6 及 RR 均為 2.0 下所有群集的偵測結果

註 1：圖上標記的數字 1~9 分別代表群集的編號

註 2：one source RR=1.4(1.6、1.8、2.0)為假定在單一群集下，SaTScan 偵測結果

假設空間存在 9 個群集，且相對風險皆相同(RR=1.6 或 RR=2.0)，群集大小相同(3*3)，分別將群聚中心按風險大小分別放置於座標 cluster1：(3.5,3.5)、cluster2：(9.5,3.5)、cluster3：(15.5,3.5)、cluster4：(3.5,9.5)、cluster5：(9.5,9.5)、cluster6：(15.5,9.5)、cluster7：(3.5,15.5)、cluster8：(9.5,15.5)、cluster9：(15.5,15.5)，且群集範圍內以普瓦松隨機生成觀察值(λ=10*RR)。

由圖 4-6 可以看出，理論上在已知所有群集的相對風險均為 1.6 及所占研究區域面積比例均為 2.25%，可以估算每一個群集內外風險差異縮小降為 1.44，而對照在單一群集下的偵測結果，對於一次性偵測而言偵測結果一致。另外使用逐

次分析在 Power 上確實可一回到單一群集的偵測結果，但在 Error rate 部分略高。同理，當所有群集的相對風險均為 2.0 亦可以估算內外風險差異降為 1.69，此時一次性偵測在 Power 上並無因保守而下降，在 Error rate 上優於逐次分析約 3%~8%。所以初步可以知道逐次分析確實能提高相對風險較小群集的偵測能力，像是在相對風險不大於 1.6 的群集時尤其有效，但若相對風險大於 2.0 時，一次性偵測較不受多重群集的影響。底下進一步假設群集的相對風險皆不同下，不同方法的檢測結果。

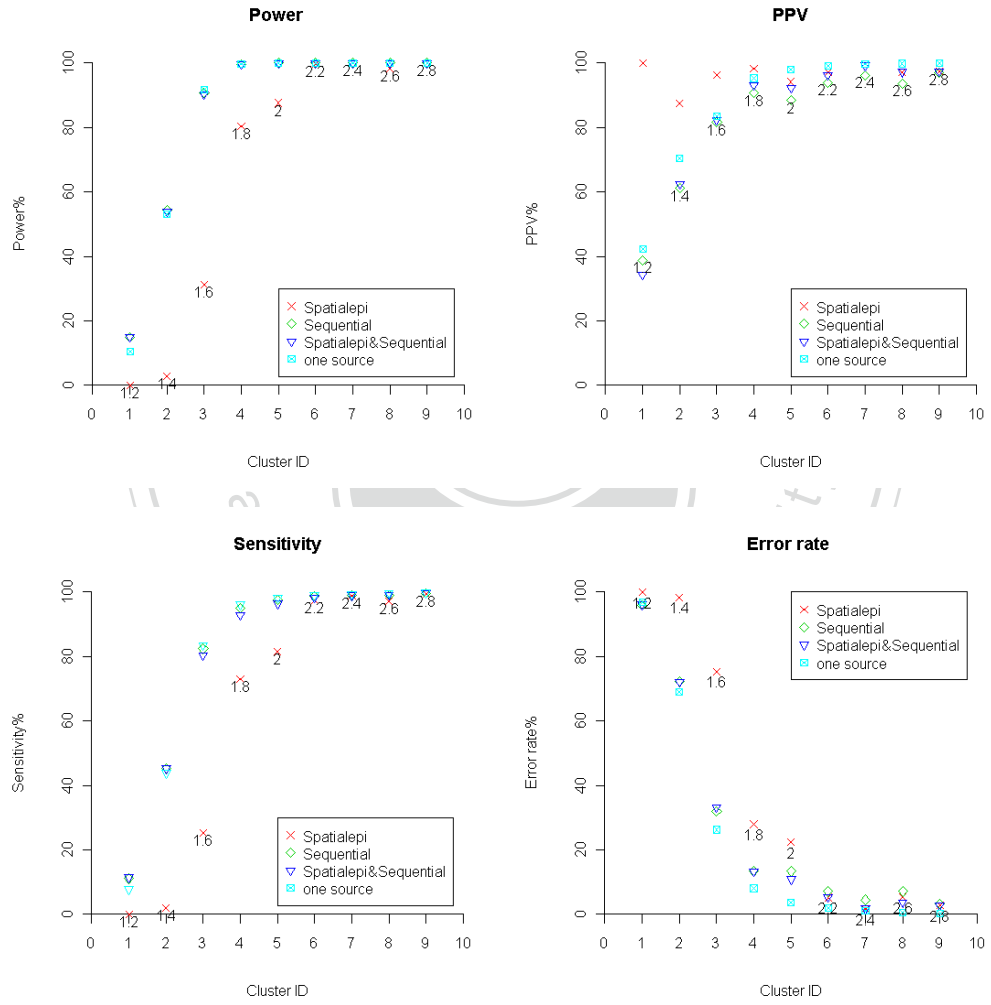


圖 4-7、不同相對風險下所有群集的偵測結果

註 1：Spatialepi & Sequential 加入以內外對風險值是否大於 1.6 作為使用一次性偵測與逐次分析分界點

註 2：圖上標記的數字 1.2~2.8 分別代表 9 個群集的相對風險

註 3：圖 PPV 中相對風險為 1.2 只有一個觀察值

註 4：one source 為假定在單一群集下，SaTScan 偵測結果

假設空間存在 9 個群集，且相對風險由 1.2 到 2.8，每個群集相對風險差距為 0.2 且群集大小相同(3*3)，分別將群聚中心按風險大小分別放置於座標 cluster1：(3.5,3.5)、cluster2：(9.5,3.5)、cluster3：(15.5,3.5)、cluster4：(3.5,9.5)、cluster5：(9.5,9.5)、cluster6：(15.5,9.5)、cluster7：(3.5,15.5)、cluster8：(9.5,15.5)、cluster9：(15.5,15.5)。

仿照先前估算內外相對風險下降幅度，此時對於檢測的群集，內外相對風險的差異減少約 20%，所以若採用本文建議方法，則約略以原始相對風險為 2.0 作為使用逐次分析與一次性偵測的分界點。由圖 4-7 可以得知在相對風險界於 1.6 以下，採用逐次分析下改善幅度最大，在相對風險 2.0 以上的群集是否採用逐次分析對偵測的結果並無太大改善。

第四節 模擬結論與建議

透過上面的模擬，可以確認無論是單一群集或是多重群集，群集的相對風險及群集所占研究區域的比例皆會影響群集偵測時的正確性。以群集所占研究區域的比例而言，占研究區域面積比例越高的群集在偵測上會較易被偵測出，相對風險越低的群集若要提高 Power，只能透過增加所占研究區域面積比例。

以群集的相對風險而言，即便群集所占研究區域面積比例很低，相對風險越高在偵測上無論透過何種方式皆能被偵測出來，但當群集所占研究區域面積比例提高時，內外風險差異減少的程度將會決定偵測的結果，在前面模擬亦應證了當群集的相對風險在 1.6 以下，內外風險差異減少 8%以上，一次性偵測就會過於保守，此時採用逐次分析，便能避免 Power 過低，Error Rate 亦能降低。

另外，若要使用本文建議方式來估計內外風險差異減少程度，由於一次性偵測所獲得的群集的相對風險會比實際值低，此時可以把研究區域移除檢測出的群集，重新計算群集的相對風險，如此便可以排除其他群集的影響，相對風險會較為接近實際值，此時重新計算內外風險縮減比例會較佳。

以「多」個群集的角度來看，逐次分析透過移除顯著群集減少對檢測的群集在偵測上的影響，一旦執行次數很多時，移除過多區塊將導致研究區域破碎，對下一個群集在偵測時，探索區域方向會受到限制，偵測結果可能會受到影響。

本文認為在進行逐次分析前，可以先透過一次性偵測將疑似群集挑出並估算其相對風險，若同時存在相對風險小於 1.6 及大於 1.6 的群集，則在移除相對風險大於 1.6 的群集後，重新對新的研究區域執行逐次分析。

模擬所使用的電腦配備為 i5 CPU 及 3.49GB RAM，並加入此準則後的逐次分析與逐次分析在計算時間上較為省時。以本文在電腦模擬情境假設為例，假設研究區域存在 9 個群集，群集占研究區域面積比例相同，且相對風險為 1.2 至 2.8，不加入本文的準則，計算上須達 9 至 10 小時；若加入相對風險 1.6 以下再採用逐次分析，可節省至少 3 小時。從模擬分析的角度來看，逐次分析在加入此準則後，降低運算複雜度，減少模擬分析所需的時間。

第五章　實證分析

本文在第四章的電腦模擬，討論群集所占研究區域面積比例及群集的相對風險，對一次性偵測的偵測結果上確實有影響，也探討在「多」個群集下，一次性偵測與逐次分析的優缺點及限制。實證上，建議使用逐次分析的時機為群集相對風險 1.6 以下最為有效，1.6 以上一次性偵測較不受多重群集影響，所以本章將根據此準則，分析臺灣地區癌症死亡率，比較一次性偵測、逐次分析的分析結果，群集偵測比較將包括研究區域中發生群集現象的位置。

本章使用的資料來自行政院衛生署民國 99 年的鄉鎮市區(Township)層級全年齡組癌症標準化死亡率資料(以 2000 年 WHO 世界人口年齡結構為基準)，在不考慮個年齡組及特定癌症的情況下，分析總體癌症死亡率在研究區域中的群集現象。由於鄉鎮市區形狀的劃分並不規則，計算每個鄉鎮市區的中心並不容易，本文以鄉鎮市區公所的位置作為每個區塊的中心。

第一節　實證資料介紹(台灣鄉鎮市區分布)

自 2010 年 12 月 25 日起，新北市（臺北縣改制）、臺中市（臺中縣市合併改制）、臺南市（臺南縣市合併改制）及高雄市（高雄縣市合併改制）等四直轄市成立。原縣下轄之鄉鎮市改制為區，臺灣共有 368 個鄉鎮市區（153 鄉、41 鎮、17 市、157 區）。本文在以民國 99 年鄉鎮市區層級全年齡組癌症標準化死亡率資料作空間群集分析時，僅考慮台灣本島的鄉鎮市區，扣除外島金門縣、連江縣、澎湖縣、綠島鄉、蘭嶼鄉、琉球鄉，臺灣本島共有 349 個鄉鎮市區（138 鄉、38 鎮、16 市、157 區，如圖 5-1）。

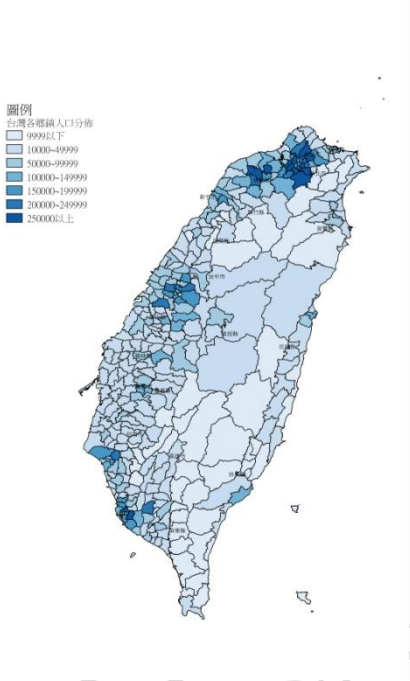


圖 5-1、台灣各鄉鎮市區人口分佈

從圖 5-1 可看出，台灣人口主要集中在少數鄉鎮市區，以台北市而言所占台灣地區面積僅 0.76%，但人口卻高達 11.38%，而花蓮縣占台灣面積為 12.90%，人口卻僅占 1.49%，在人口結構上各地區差異甚大。

由於台灣各鄉鎮市區年齡結構、性別結構特性都不一樣，若直接以粗死亡率作為當作死亡率高低的判斷標準，則老年人口比例較高的地區，死亡率通常比較高。所以為了避免判讀的誤差，必須透過標準化的方法調整人口組成，使其具有比較性。而所謂標準化死亡率(Standard Mortality Ratio, SMR)，即在進行不同族群的比較時，調整族群間的差異後所計算出來的數值(Lilienfeld and Stolley, 1994)。所以本文為了防止判讀的誤差，資料採用衛生署民國 99 年的鄉鎮市區層級全年齡組癌症標準化死亡率(以 2000 年 WHO 世界人口年齡結構為基準)。標準死亡率及鄉鎮市區癌症死亡人數計算方式如下：

$$\text{年齡別死亡率} = \frac{\text{某年齡層癌症新診斷或死亡人數}}{\text{某年齡層人口數}} \times 100000$$

$$標準化死亡率 = \frac{\sum（某特定年齡別死亡率 \times 該年齡層標準人口數）}{\sum 某特定年齡層標準人口數} \times 100000$$

$$鄉鎮市區癌症死亡人數 = 標準化死亡率 \times 鄉鎮市區人口數 \div 100000$$

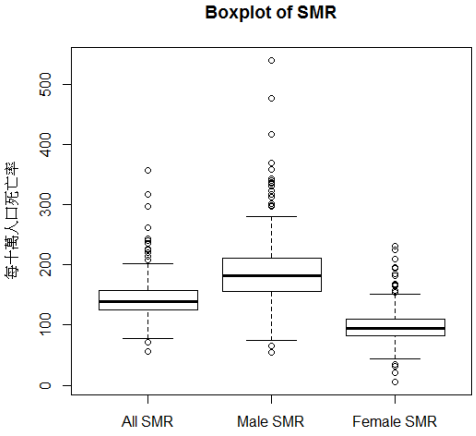


圖 5-2、標準化死亡率盒鬚圖

以圖 5-2 來看，無論是以台灣整體、男性及女性標準化死亡率，分配皆為右偏，其中又以台灣男性癌症標準化死亡率最為右偏，女性癌症標準化死亡率最接近對稱，所以在空間群集分析結果可能會包含相對風險大於 1 及相對風險小於 1 的群集。

在使用台灣的鄉鎮市區人口資料時，與第四章電腦模擬設定不同。以研究區域而言，先前設定為四方形，且每一格點距離皆相同；而對照實證資料，台灣地形狹長，鄉鎮市區集中分佈在西部。在人口數上，先前設定為每一格點皆為 1 萬人；而對照實證資料，台灣在人口結構上各鄉鎮市區差異甚大。以群集而言，先前設定皆避免靠近邊緣，以免在群集偵測上造成檢定力下降，在空間統計為邊緣效應(Edge Effect)；而對照實證資料，群集有可能發生在本島邊緣，以上這些因素都有可能導致實證分析結果與模擬結果間差異。

第二節 99 年癌症死亡率分析結果

本節分析民國 99 年的鄉鎮市區層級全年齡組癌症標準化死亡率資料(以 2000 年 WHO 世界人口年齡結構為基準),由於各鄉鎮市區形狀的劃分並不規則，計算每個鄉鎮市區的中心並不容易，本文以各個鄉鎮市區公所作為每個區塊的中心。底下本文使用 R 中的 Spatialepi 對台灣整體、台灣女性、男性各鄉鎮市區每十萬人口癌症死亡率檢測群集，並進一步比較一次性偵測、逐次分析，並列出相對風險大於 1 的群集。另外，在第一節時提到台灣各鄉鎮市區人口差異甚大，所以在探索區域上,取縣市人口占台灣人口最大比例 0.16,作為探索區域人口上限。首先針對 99 年台灣整體癌症死亡率進行群集檢測。

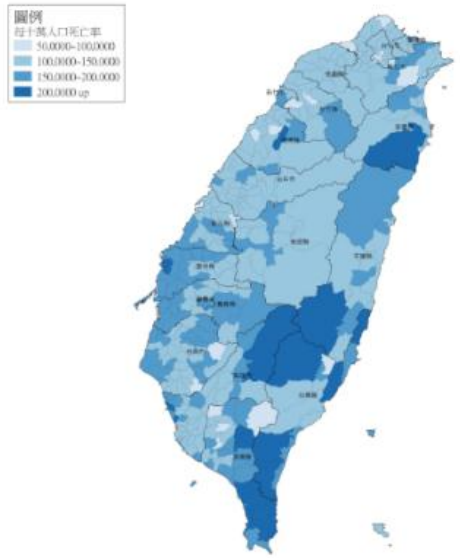


圖 5-3、台灣整體各鄉鎮市區每十萬人口癌症死亡率

註 1：圖中劃分的顏色區塊為每十萬人口死亡率以 50 為級距

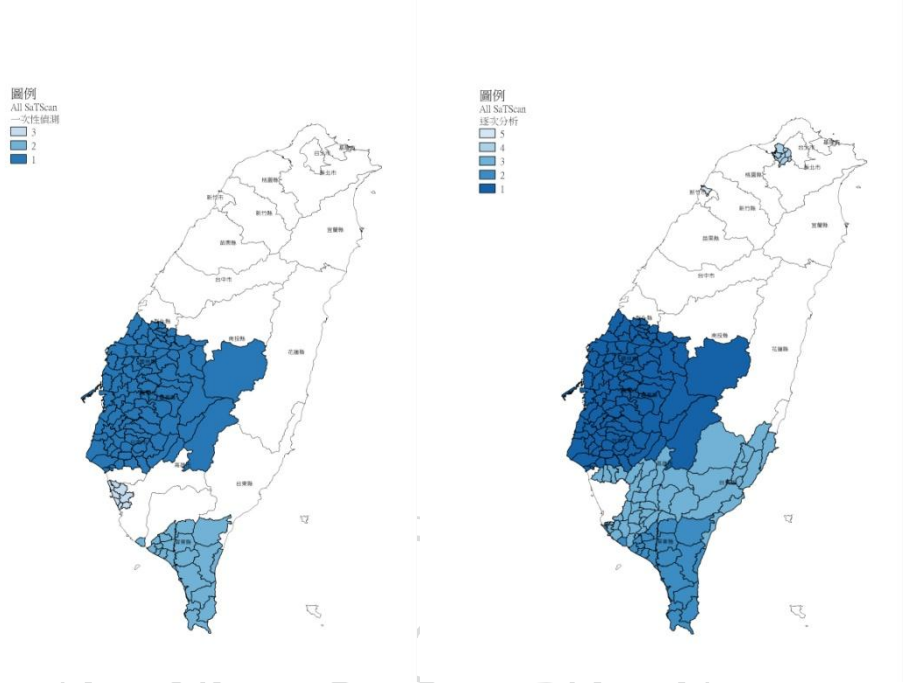


圖 5-4、台灣整體癌症死亡率群集分佈(由左至右：一次性、逐次)

註 1：數字代表偵測優先順序

在第四章電腦模擬中，提到群集所占研究區域及相對風險大小會影響群集的偵測結果，而透過一次性偵測並無發現存在異常低的群集。對照圖 5-3 可以發現屏東、雲林、嘉義、台南、高雄、新竹市北區、宜蘭等地區標準死亡率較高，而對照 5-4，一次性偵測到的群集所占台灣面積為 32.54%，群集 1 所占研究區域達 25.24%，且相對風險為 1.18，群集 2 所占研究區域達 6.72%，且相對風險為 1.34，群集 3 所占研究區域達 0.58%，且相對風險為 1.26，內外風險縮減比例為 6.52%。一次性偵測與逐次分析所偵測到的結果，皆和標準死亡率較高的區塊有重複，但逐次分析並未偵測到高雄市部分區域，而偵測到新竹市北區及新北市某些區塊。逐次分析在將前 3 個顯著群集移除研究區域後，高雄市剩餘區塊便很小，且受到探索區域人口上限限制，所以在偵測過程中可能導致不顯著。底下進一步分析台灣女性各鄉鎮市區每十萬人口癌症死亡率的群集分佈。

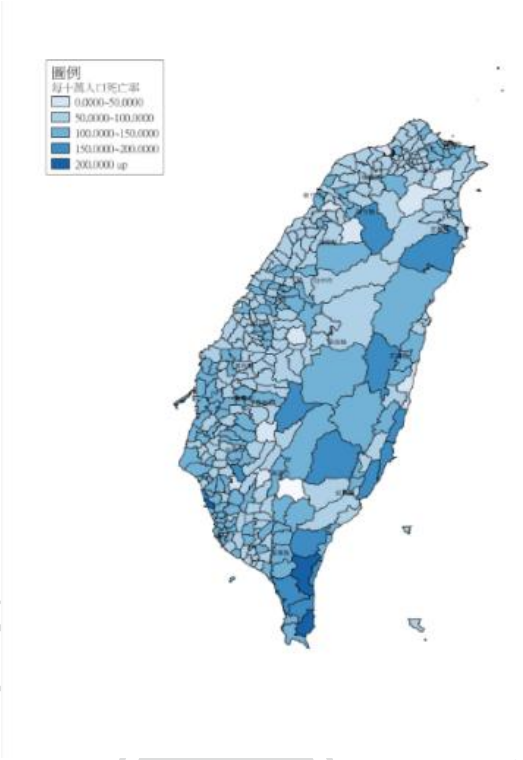


圖 5-5、台灣女性各鄉鎮市區每十萬人口癌症死亡率

註 1：圖中劃分的顏色區塊為每十萬人口死亡率以 50 為級距

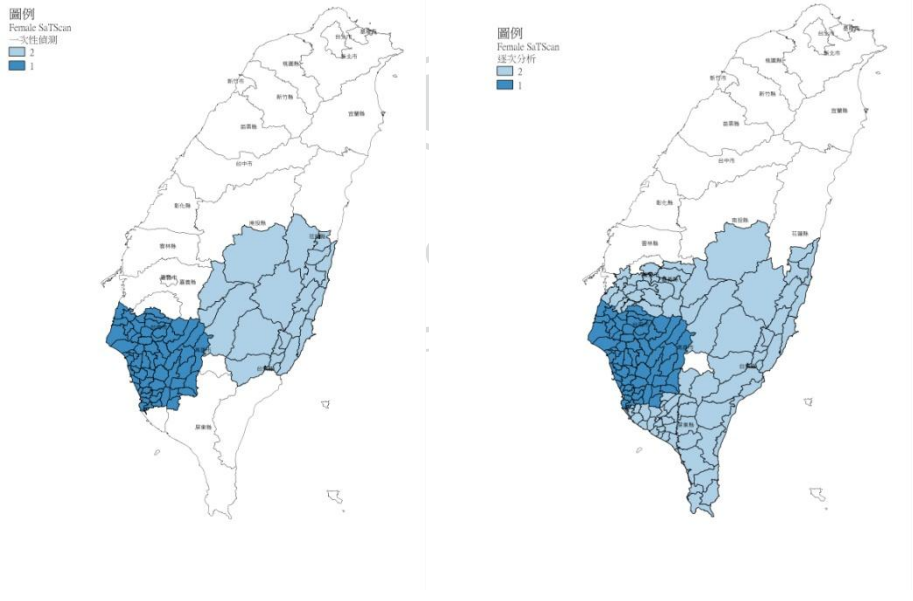


圖 5-6、台灣女性癌症死亡率群集分佈(由左至右：一次性、逐次)

註 1：數字代表偵測優先順序

註 2：由於高雄市茂林區於 2010 年並無癌症死亡人口，所以無法得知標準死亡率

對照圖 5-5 可以發現屏東、雲林、嘉義、台東、宜蘭等地區標準死亡率較高，

對照 5-6，一次性偵測到的群集所占台灣面積為 31.30%，群集 1 所占研究區域達 9.70%，且相對風險為 1.15，群集 2 所占研究區域達 21.60%，且相對風險為 1.48，內外風險縮減比例為 10.57%。一次性偵測與逐次分析所偵測到的結果，皆和標準死亡率較高的區塊有重複，而逐次分析所偵測到群集 2 的區塊範圍較大，有可能受邊緣效果影響。底下進一步分析台灣男性各鄉鎮市區每十萬人口癌症死亡率的群集分佈。

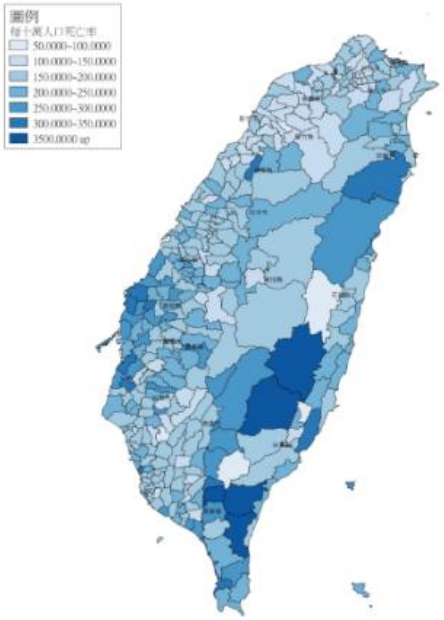


圖 5-7、台灣男性各鄉鎮市區每十萬人口癌症死亡率

註 1：圖中劃分的顏色區塊為每十萬人口死亡率以 50 為級距

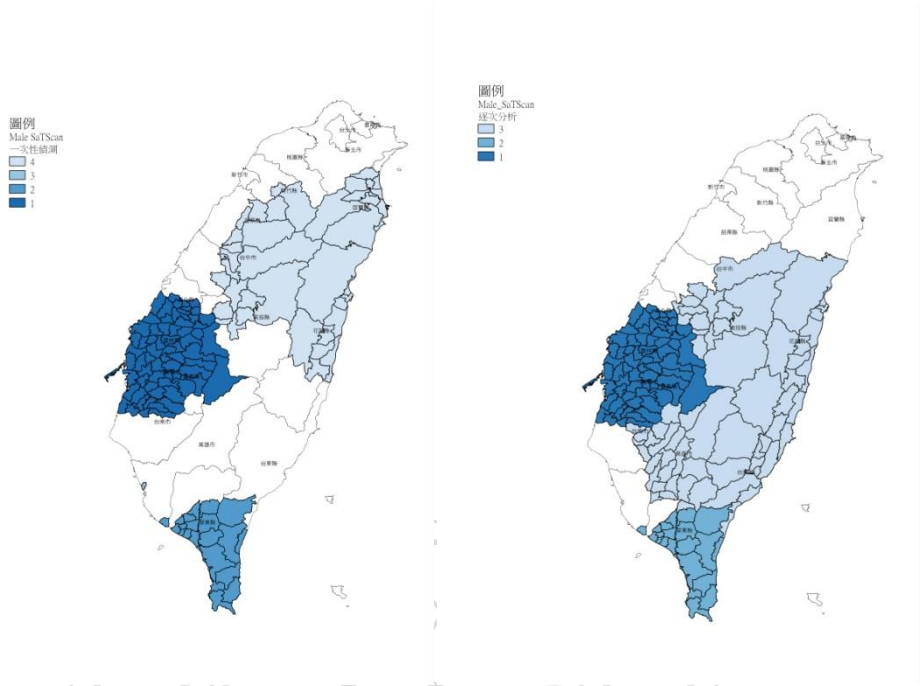


圖 5-8、台灣男性癌症死亡率群集分佈(由左至右：一次性、逐次)

註 1：數字代表偵測優先順序

對照圖 5-7 可以發現屏東、雲林、嘉義、台東、宜蘭等地區標準死亡率較高，而對照 5-8，一次性偵測到的群集所占台灣面積為 49.0%，群集 1 所占研究區域達 13.00%，且相對風險為 1.29，群集 2 所占研究區域達 6.72%，且相對風險為 1.38，群集 3 所占研究區域達 0.03%，且相對風險為 1.98，群集 4 所占研究區域達 29.17%，且相對風險為 1.14，內外風險縮減比例為 9.45%。一次性偵測與逐次分析所偵測到的結果，皆和標準死亡率較高的區塊有重複，逐次分析偵測群集 3 則可能涵蓋過多 False Positive。

第三節 實證應用小結

本文在先前建議使用逐次分析的時機為群集相對風險 1.6 以下最為有效，1.6 以上一次性偵測較不受多重群集影響，而從 99 年癌症死亡資料可以發現實際資料不會如同先前模擬會有高相對風險的情形發生，都只有低相對風險的群集發生。從模擬結果可知，相對風險在介於 1.1 至 1.2 之間的群集在檢測時可能會涵蓋過

多 False Positive，對照在實證分析結果上，探索區域人口上限設定為 0.5 幾乎可以涵蓋半個台灣，可以預期在檢測結果並不理想。所以在探索區域人口上限調降為 0.16，透過 SaTScan 執行一次性偵測與逐次分析亦發現兩者在群集偵測結果上略有差異。實證發現，無論就整體癌症死亡率、女性癌症死亡率或男性癌症死亡率，三者在群集分佈上都有重疊的區域如雲林、嘉義、台南、高雄、屏東花東地區。

另外，在群集的偵測的結果上，群集範圍皆涵蓋台灣本島邊緣，檢測結果可能受邊緣效果影響，且群集的相對風險皆不高。雖然本文在探索區域人口上限縮小至 0.16，但透過 SaTScan 偵測到的群集面積幾乎達台灣本島 30%~50%，而逐次分析所檢測出的群集面積占台灣本島比例更高，對照第四章電腦模擬可知，在群集相對風險很低(如 1.2)，檢測結果容易涵蓋過多 False Positive。

過去麥寮六輕、恆春核三廠等區域，被人質疑該區癌症死亡率較高，而蔡丞庭 (2011)透過焦點檢定，檢測 2009 年麥寮六輕、恆春核三廠癌症死亡率，亦發現這些地區周圍具有較高癌症死亡率，雖然本文在檢測群集的結果上亦包含這些地區，但可能涵蓋過多 False Positive，而透過縮減探索區域的方式雖然可以達到降低 False Positive，但顯然在這個參數上還可以做調整。另外，關於 SaTScan 檢測出的群集相對風險偏低時會涵蓋過多 False Positive 的問題，過去亦有學者提出可以使用 Permutation test 的方式做調整，若使用逐次分析搭配縮減探索區域範圍及 Permutation test 或許可以達到不錯效果，但可以預期加入 Permutation test 後，逐次分析耗費的時間會很大。

第六章　結論與建議

第一節 結論

空間群集偵測在空間統計上，廣泛的被應用在各領域，而 Kulldorff and Nagarwalla (1995)的 SaTScan 及 Tango and Takahashi (2005)的 FlexScan 是最常被使用的群集偵測方法。多數的群集偵測方法採取一次性偵測，比較疑似群集之內外相對風險，同時檢視所有疑似群集，如果研究目標為找出最顯著群集，如此確實可提高計算效率。然而，如果目標在於找出所有顯著群集，相對風險較小群集會受到其他發生率較高群集的影響，在檢測結果上會較保守。所以在多重群集的檢測上亦有學者提出不同的偵測的方式如 Li et al. (2011)提出新的 Spatial Scan Statistic 及 Zhang et al. (2010)的 Sequential method。

本文以檢測所有群集為目標，並以 Kulldorff and Nagarwalla (1995)SaTScan 作為檢測群集的主要方式，並比較群集占研究區域面積比例、群集相對風險及群集個數對於一次性偵測及逐次分析在檢測結果上的影響。透過電腦模擬可知，在進行一次性偵測時，當相對風險大於 1.0 的群集(亦即 Cluster)占研究區域面積比例越高，便會減少其他群集內外風險差異，內外風險差異減少幅度越大，在檢測群集時會越保守，此時透過逐次分析，在檢測結果上能回到單一群集時的檢測結果。

以群集相對風險而言，當檢測的群集相對風險介於 1.3~1.6，使用逐次分析結果會優於一次性偵測，在 Power 上可以改善 10%~15%，Error Rate 可以改善 3%~10%；若檢測的群集相對風險在 1.2 以下，即便透過逐次分析亦無法得到有效改善；若檢測的群集相對風險在 1.6 以上，一次性偵測在群集檢測結果上，受多重群集影響不大。以「多」個群集的角度來看，逐次分析透過移除顯著群集減少對檢測的群集在偵測上的影響，一旦執行次數很多時，移除過多區塊將導致研究區域破碎，對下一個群集在偵測時，探索區域方向會受到限制，偵測結果可能

會受到影響。

本文也以民國 99 年的鄉鎮市區全年齡組癌症標準化死亡率資料，驗證逐次分析的效果，與 Kulldorff and Nagarwalla (1995)的 SaTScan 比較台灣整體、女性、男性的癌症標準化死亡率。透過逐次分析與一次性偵測，兩者皆發現台灣整體、女性及男性群集分佈達研究區域 30%~50%，某些區塊無法透過一次性偵測檢測出來，亦被逐次分析偵測到。另外，群集範圍皆涵蓋台灣本島邊緣，可能產生邊緣效果，且群集相對風險皆不高，容易涵蓋過多 False Positive，所以在檢測結果上可能會與實際結果產生偏差。

第二節 討論及建議

在 Zhang et al. (2010)提出 Sequential method 修正多重群集的偵測，實證分析結果上，對於對風險較低的群集，透過逐次分析可以讓 P-value 回到正常值，與本文在電腦模擬所得到結果相同，在實證分析上透過逐次分析亦發現到一次性偵測所無法檢測到的群集。

雖然本文發現逐次分析可改善 SaTScan 對相對風險較小（例如：1.6 以下）群集的偵測，但發現逐次分析也有待改進的地方。首先，因為逐次分析的作法類似逐步迴歸，藉由移除顯著群集以凸顯風險較小群集，但若移除過多區塊將導致研究區域破碎，對下一個群集在偵測時，探索區域方向會受到限制，檢測結果可能會受到影響，或許可以透過降低縮減探索區塊範圍。第二個問題則是群集偵測經常遇到的多重檢定，逐一篩選顯著群集會增加型一誤差，除了可使用常見的 Bonferroni 修正，也可嘗試逐次分析中 Repeated Significance Test (Armitage et al., 1969)或α-spending Function (DeMets and Lan, 1994)，依照檢定次數（或是顯著群聚數）調整顯著水準。

以 Kulldorff and Nagarwalla (1995)的 SaTScan 而言，透過逐次分析檢測出相對風險較高(RR=1.6 以上)的群集，產生誤判的可能性較小，通常會發生誤判的情

形可能來自於相對風險較小的群集。在統計學上對於多重檢定，最常採取的作法是降低每次檢測時的顯著水準，但由於在 SaTScan 在檢測群集過程中，相對風險較小的群集乃透過增加群集範圍達到檢定顯著，通常檢定顯著後 P-value 都很小，只降低顯著水準效果可能不明顯，若有充分資訊(如疾病傳染的範圍)，較可以避免對相對風險較小的群集產生誤判。但若沒有充分的事前資訊，在縮減探索區域範圍必須要特別小心。

關於 SaTScan 檢測出的群集相對風險偏低時會涵蓋過多 False Positive 的問題，除了可以透過縮減探索區域範圍外，過去亦有學者提出可以使用 Permutation test 的方式做調整。若使用逐次分析搭配縮減探索區域及 Permutation test 或許可以達到不錯效果，但可以預期加入 Permutation test 後，逐次分析耗費的時間會很大。

以氣喘罹病率、台灣藍綠政治版圖為例，多重群集存在的可能性便很高，研究者未來在探討這些議題上，若目標在於找出所有顯著群集，方法的使用上便可以考慮使用逐次分析，探討空間群集時，便可避免相對風險較小群集受到其他發生率較高群集的影響。

過去在空間群集檢測上，大多在單一群集下討論，從群集相對風險到群集形狀不同下的檢測結果都有學者探討。在多重群集的空間檢測尚存在很多問題，如逐次分析在群集形狀上的差異對檢測結果影響程度、移除區塊面積的限制及如何降低顯著水準，這些在空間群集檢測上有待各位學者討論研究。

參考文獻

Auchincloss, A.H., Gebreab, S.Y., Mair, C. and Diez Roux, A.V. (2012). A Review of Spatial Methods in Epidemiology, 2000–2010, Annual Review of Public Health, 33:107–22

Bithell J.F. (1995). The choice of test for detecting raised disease risk near a point source, Statistics in Medicine 14:2309–2322.

Cliff, A. and Ord, J.K. (1981). Spatial Processes: Model and Applications, London: Pion.

Cucala, L. (2009). A flexible spatial scan test for case event data, Computational Statistics and Data Analysis 53: 2843–2850.

Demattei, C., Molinari, N. and Daures, J.P. (2007). Arbitrarily shaped multiple spatial cluster detection for case event data, Computational Statistics and Data Analysis 51:3931–3945.

DeMets, DL and Lan, KKG (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, 13:1341-1352

Diggle, P.J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point, Journal of the Royal Statistical Society 153:349-362.

Diggle, P.J. and Rowlinson, B.S. (1994). A conditional approach to point process modeling of elevated risk, Journal of the Royal Statistical Society 157:433-440.

Fairbanks, K. and Madsen, R. (1982). P values for tests using a repeated significance test design, Biometrilca, 69, 1, pp. 69-74

Jackson, M.C., Huang, L., Luo, J., Hachey, M. and Feuer, E. (2009). Comparison of

tests for spatial heterogeneity on data with global clustering patterns and outliers, International Journal of Health Geographics 8:55.

Kulldorff M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference, Statistics in Medicine 14: 799–810.

Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2006). An elliptic spatial scan statistic, Statistics in Medicine 25: 3929–3943.

Kulldorff, M., Tango, T. and Park, P.J. (2003). Power comparisons for disease clustering tests, Computational Statistics and Data Analysis 42: 665–684.

Li, X-Z, Wang, J-F, Yang, W-Z, Li, Z-J and Lai, S-J. (2010). A spatial scan statistic for multiple clusters, Mathematical Biosciences, 233: 135–142.

Lilienfeld, D.E. and Stolley, P.D. (1994). Foundations of Epidemiology (3rd Ed.). Oxford University Press

Lloyd, .N, Trefethen and David, .Bau, III. (1997). Numerical Linear Algebra, SIAM

Song, C. and Kulldorff, M. (2003). Power evaluation of disease clustering tests, International Journal of Health Geographics 2.

Song, C. and Kulldorff, M. (2005). Tango's maximized excess events test with different weights, International Journal of Health Geographics Dec 15: 4:32.

Stone R.A. (1988). Investigations of excess environmental risks around putative sources: statistical problems and a proposed test, Statistics in Medicine 7:649–660.

Tango, T. (1995). A class of tests for detecting general and focused clustering of rare diseases, Statistics in Medicine 14: 2323-2334.

Tango, T. (2000). A test for spatial disease clustering adjusted for multiple testing, Statistics in Medicine 19:191-204.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters, International Journal of Health Geographics 4.

Waldhor T. (1996). The spatial autocorrelation coefficient Moran's I under heteroscedasticity, Statistics in Medicine 15(7-9)：887-892.

Wan, Y., Pei, T., Zhou, C., Jiang Y., Qu, C. and Qiao, Y. (2012). ACOMCD: A multiple cluster detection algorithm based on the spatial scan statistic and ant colony optimization, Computational Statistics and Data Analysis 56：283–296.

Zhang, Z., Assunçã o, R. and Kulldorff, M. (2010). Spatial scan statistics adjusted for multiple clusters, Journal of Probability and Statistics Article ID: 642379.

王泰期, 2006。 疾病群集檢測方法及檢定力比較, 政治大學碩士論文

蔡承庭, 2011。 焦點檢定方法比較, 政治大學碩士論文