

# 臺灣各縣市登革熱好發與否研究分析

統計111 H24089014 許祐誠

統計111 H24089014 許祐誠

統計111 H24089014 許祐誠

統計111 H24089014 許祐誠

統計111 H24089014 許祐誠

## 一、摘要

### （一）背景動機與統計方法：

本研究主要探討氣候因素是否對於登革熱之好發有所影響，資料取自 2014 年至 2021 年臺灣各縣市罹患登革熱病例之資料，擷取本土確診案例部分並合併當月份之氣溫、降雨及氣壓等相關變數。變數篩選方面，透過多變量的主成份分析方法將資料降維，選取解釋變異高的因子以利後續分析；由於本資料案例中超過八成的資料為零值，因此主要採用零膨脹模型中的卜瓦松以及負二項式模型作為本研究之主要分析方法；此外我們也透過將意義相近的變數做因子命名，將各項因子分析組合放入樹狀模型以求得分類結果。

### （二）結果：

在使用廣義線性模型方面，在我們配適的所有 6 個模型中，透過模型選擇及驗證後選擇**零膨脹負二項式模型**中配適最佳的 Model 5 作為結果說明，而在此模型當中，通過模型預測以及實際案例比較後，得出臺灣南部秋季與冬季為登革熱好發之季節，其中又以秋季最為好發，而春季則是四個季節中發生登革熱最低的季節，且臺灣北中南東各區域之間，登革熱的不發生率沒有顯著差異，也就是臺灣各區皆有潛在發生登革熱的可能性。

在樹狀模型方面，通過因子分析的兩種方法-最大概似估計法、主成分法，並加以考慮病例數是否出現 0，共 4 種因子分析組合之情況；再配合將 4 種分析組合套入決策樹、梯度提升決策樹模型最終得到 8 種分析組合結果。最終樹狀模型所預測之結果大致上與使用廣義線性模型的結論相符，當地區為臺灣南部，且月份為接近秋冬之際時，會有登革熱疫情較為好發的情況產生。

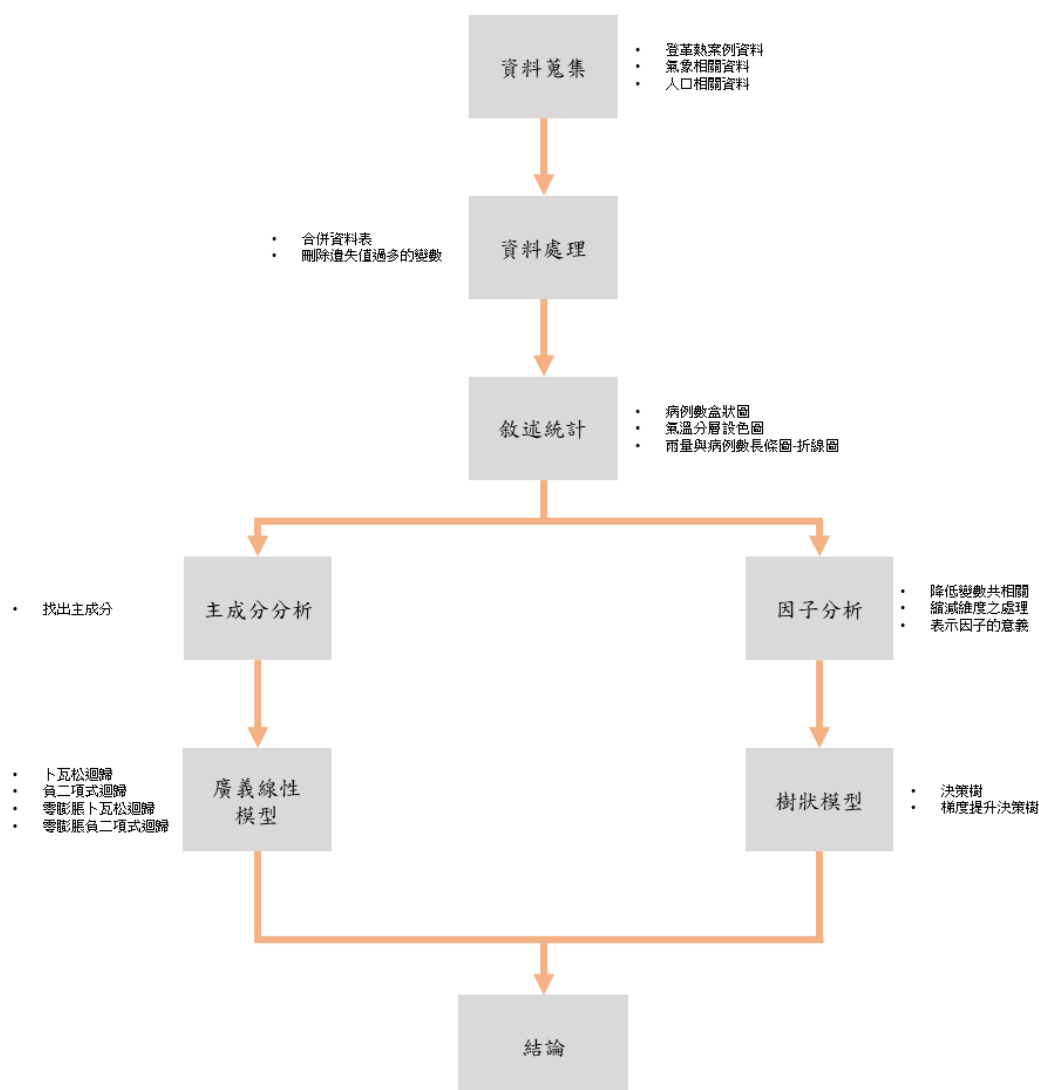
（三）結論：透過廣義線性模型以及因子選擇輔以樹狀模型等統計方法得出臺灣四大地區中以南部地區（嘉義縣、嘉義市、臺南市、高雄市、屏東縣）為登革熱高好發區，其中又以秋季（9 月至 11 月）為最好發季節。

## 二、背景動機

關於我們研究此题目的動機是有鑒於就讀位於臺南的成大，生活在登革熱高好發的南部地區，並曾耳聞系上曾經因為使用積水容器有被校方記點的情形。近幾年學校一直有在對於登革熱進行大規模的防疫以及全面性的消毒，可知校方對於登革熱防治的重視。對於登革熱疫情近幾年來的病例數多寡令我們所好奇且關心，而導致登革熱好發的成因也是我們希望深入探討的部分，比方像是溫度、濕度的變化是否對於登革熱的引發產生顯著的影響，甚至我們更希望能夠進一步探討如氣壓、雲量等等更詳細的外部因素是否對於疫情的擴散有所影響。

而登革熱的好發研究之所以重要，是由於臺灣位於亞熱帶地區，這種濕熱的環境是孕育蚊子良好的溫床，萬一患上了登革病毒，將引起諸多不適，最嚴重甚至引起器官壞死，極為嚴重。而在 2014 及 2015 年則發生了歷年來最嚴峻的登革熱疫情，病例數突破了萬例，疫情集中於高雄市以及臺南市，也就是我們所生活的都會區，與我們息息相關，因此也顯現了登革熱研究的重要性。

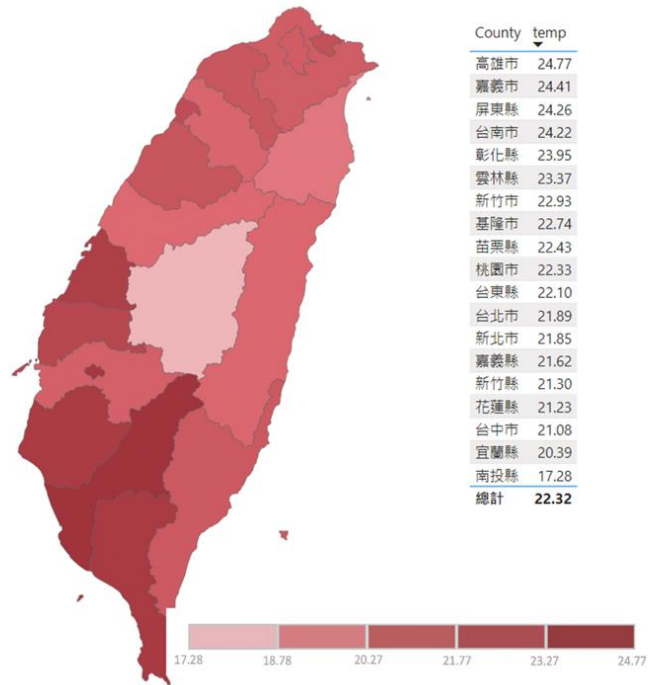
## 三、分析步驟流程圖



#### 四、方法學應用

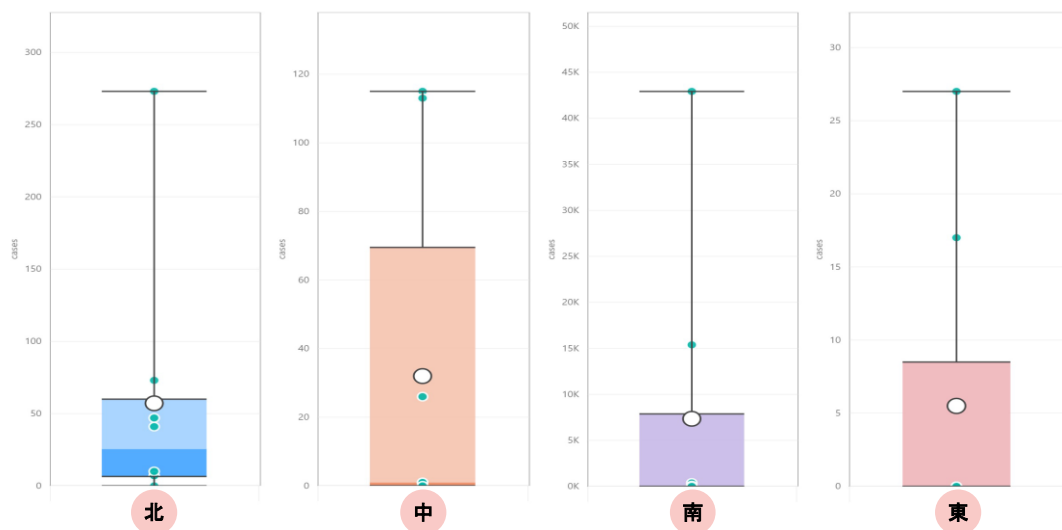
##### (一) 敘述統計 (descriptive statistics)

圖一、臺灣各縣市平均氣溫



由圖一可以得知，在臺灣各地區均溫最高的城市為南部地區的高雄市、嘉義市以及屏東縣，年均溫約落在 24.5 度，而均溫最低的城市位於海拔較高的南投縣，均溫約為 17.3 度。

圖二、臺灣四大地區分區病例總數之盒狀圖



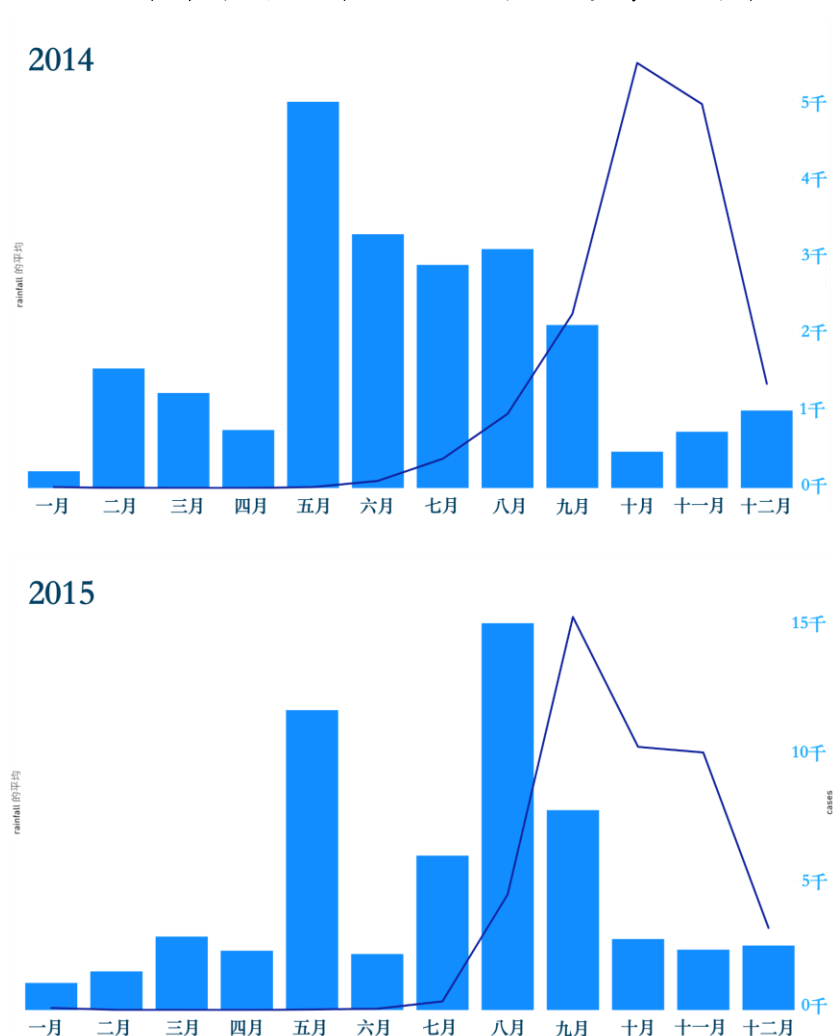
由圖二可以了解到，病例總數最高的地區為南部地區，最高值落在登革熱疫情爆發的 2015 年，病例總數將近 45000 例，而各年度病例平均總數則約落在

8000 例；病例總數最低的地區則是花東地區，最高值僅 28 例，平均病例總數則約為 6 例。其中，我們的分區規則以及季節切割的月份如下表所示。

發病區域(region)	發病縣市(County)
北區	臺北市、新北市、基隆市、宜蘭縣、桃園市、新竹縣、新竹市
中區	苗栗縣、臺中市、彰化縣、南投縣、雲林縣
南區	嘉義縣、嘉義市、臺南市、高雄市、屏東縣
東區	花蓮縣、臺東縣

發病季節(season)	發病月份(Month)
春	2~4 月
夏	5~7 月
秋	8~10 月
冬	11~1 月

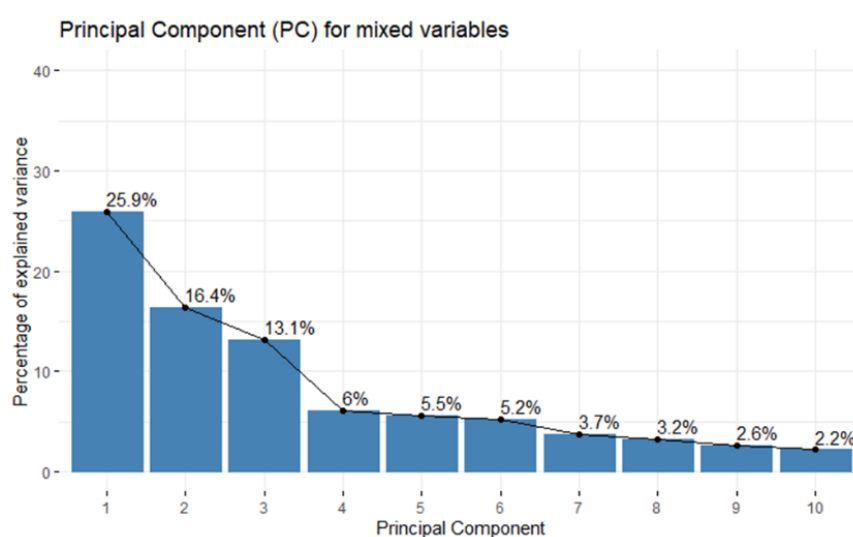
圖三、平均月雨量與各月份病例總數長條及折線圖  
(以 2014 與 2015 年為例，其餘年度放至附錄以供參考及比較)



圖三為各年度平均月雨量與各月份病例總數統計圖，長條圖為平均月雨量，折線圖為病例數總數，依據各年度的資料可以推論出臺灣在夏季 7、8 月將水豐沛的雨季後一個月，也就是約在 9、10 月將會爆發登革熱的疫情，與參考文獻之研究結果基本一致。

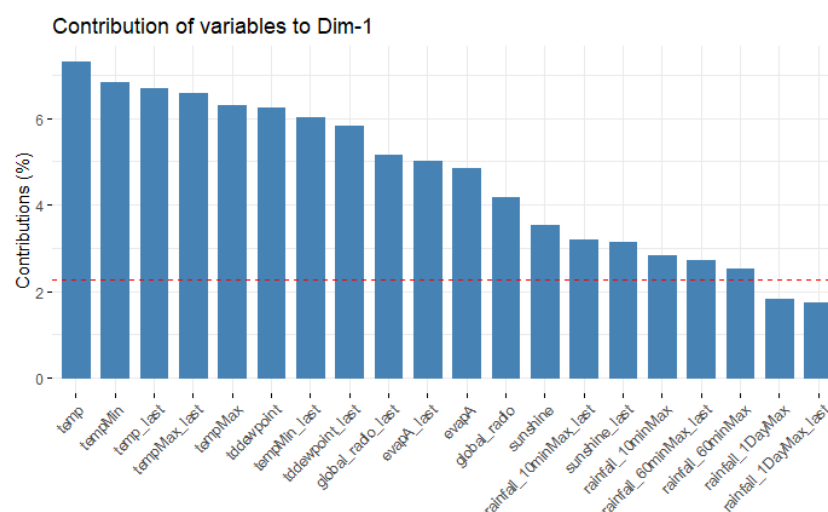
## (二) 主成分分析 (Principal Components Analysis, PCA)

我們想透過主成分分析將資料降維度，把多個變數的資料降維成少數幾個主成分，希望能用少數個主成分反映原始資料絕大部份的資訊，再將我們所認為解釋變異足夠大的主成分代入模型，做後續分析。

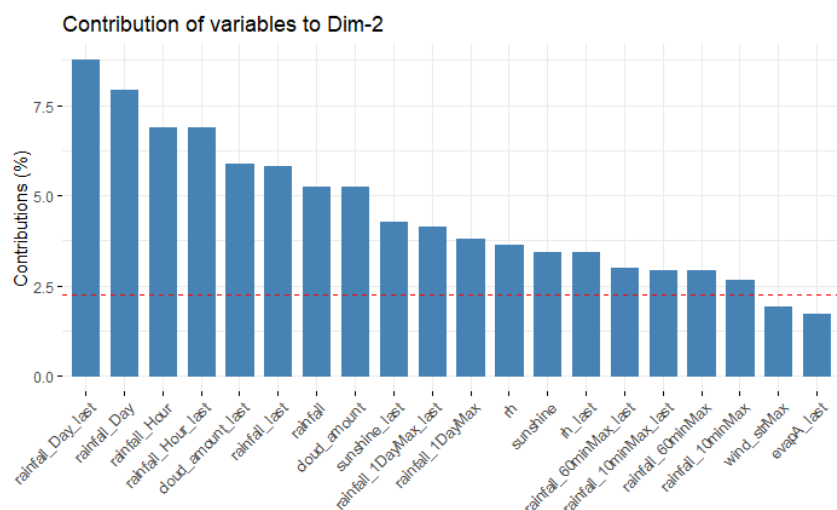


由上圖可以看到，除了前三個主成分解釋變異大於 10%，其餘後面的主成分能提供的解釋變異量都不大，因此我們決定使用前三個主成分帶入模型作分析。

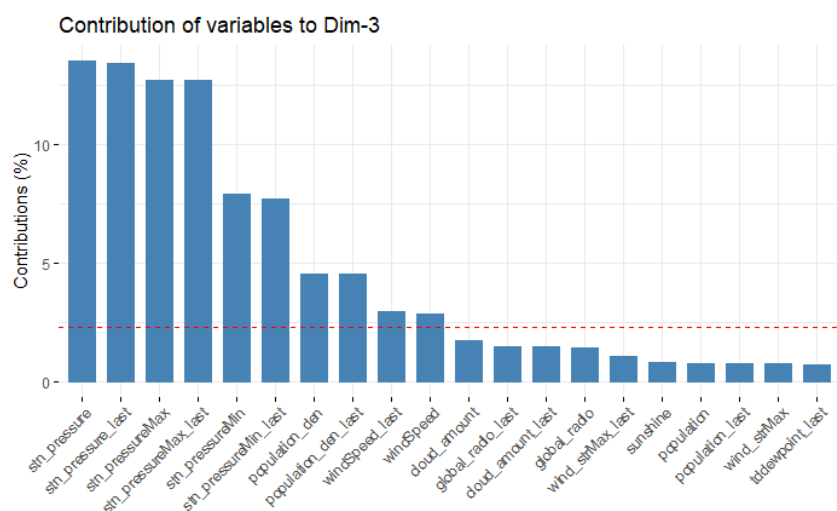
以下是前三個主成分的各個變數在主成分的貢獻比例圖



(第一主成分)



(第二主成分)



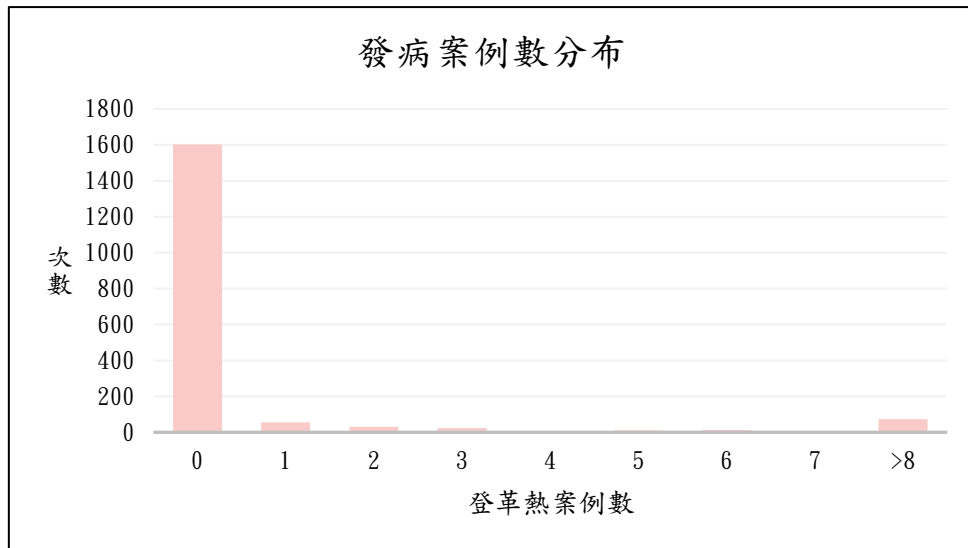
(第三主成分)

我們可以看到第一個主成分主要貢獻變數大多是溫度相關；第二主成分則是雨量相關；第三主成分則是氣壓相關，最後我們將這三個主成分代入模型進行分析。

### (三) 廣義線性模型 (Generalized Linear Model, GLM)

#### 1. 模型配適

由於案例個數為計數型的資料，所以應採用卜瓦松迴歸 (Poisson Regression) 或是負二項式迴歸 (Negative Binomial Regression)，且由於案例個數中，有超過八成的資料為零，因此尚須使用零膨脹模型 (Zero-Inflated Model)。零膨脹模型與傳統計數型模型的比較如附錄二之整理。各種模型配適的過程如附錄三所整理，以下僅列零膨脹模型配適的部分。



(1) 零膨脹卜瓦松模型模型：(Model 3)

● Count

$\log(E(\text{cases} | X))$

$$= \beta_0 + \beta_{\text{region}}I(\text{region}) + \beta_{\text{season}}I(\text{season}) + \beta_{\text{temp}_{\text{last}}}x_{\text{temp}_{\text{last}}} + \beta_{\text{rainfall}_{\text{last}}}x_{\text{rainfall}_{\text{last}}}$$

● Zero

$$\text{Logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{\text{region}}I(\text{region}) + \beta_{\text{season}}I(\text{season})$$

Model 3(count)	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	-9.4742	0.1062	-89.216	< 0.0001
$\beta$ region 北	-0.3896	0.0803	-4.8543	< 0.0001
$\beta$ region 東	-0.2179	0.1749	-1.2456	0.2129
$\beta$ region 南	4.7851	0.0643	74.4711	< 0.0001
$\beta$ season 春季	-4.3747	0.1348	-32.4554	< 0.0001
$\beta$ season 秋季	-0.1818	0.0211	-8.6069	< 0.0001
$\beta$ season 夏季	-2.8181	0.0267	-105.654	< 0.0001
$\beta$ temp_last	0.4642	0.0037	127.0435	< 0.0001
$\beta$ rainfall_last	-0.001	< 0.0001	-69.4346	< 0.0001

Model 3(zero)	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	1.9263	0.2620	7.3512	< 0.001
$\beta$ region 北	-0.6524	0.2334	-2.7951	0.0052
$\beta$ region 東	-0.1018	0.3651	-0.2788	0.7804
$\beta$ region 南	-0.4219	0.2292	-1.8412	0.0656
$\beta$ season 春季	0.7356	0.4069	1.8077	0.0707
$\beta$ season 秋季	-0.5109	0.2252	-2.2689	0.0233
$\beta$ season 夏季	-0.2683	0.2415	-1.1111	0.2665

計數的部分中，大多數的變數皆呈現顯著，從係數估計的方向來看，與中區相比，北區與東區發生登革熱案例數較少，南區較多；而以季節來說，冬季發生登革熱案例數較多；且當上個月的氣溫較高且雨量較少時，可能也會有較多的登革熱案例。模型第二個部分為估計案例數為零的機率，僅有秋季顯著，且係數為負，也就是秋季登革熱不發生的勝算比相較於冬季低，也就是秋季可能較容易有登革熱的疫情。

## (2) 零膨脹負二項式模型：(Model 4)

### ● Count

$\log(E(\text{cases} | X))$

$$= \beta_0 + \beta_{\text{region}}I(\text{region}) + \beta_{\text{season}}I(\text{season}) + \beta_{\text{temp}_{\text{last}}}x_{\text{temp}_{\text{last}}} + \beta_{\text{rainfall}_{\text{last}}}x_{\text{rainfall}_{\text{last}}}$$

### ● Zero

$$\text{Logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{\text{region}}I(\text{region}) + \beta_{\text{season}}I(\text{season})$$

Model 4(count)	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	-7.5128	1.1654	-6.4465	< 0.0001
$\beta$ region 北	-0.9894	0.4149	-2.3845	0.0171
$\beta$ region 東	-0.6409	0.7392	-0.867	0.386
$\beta$ region 南	4.2013	0.5168	8.1293	< 0.0001
$\beta$ season 春季	-4.4576	0.6637	-6.7159	< 0.0001
$\beta$ season 秋季	0.3298	0.4795	0.6877	0.4916
$\beta$ season 夏季	-1.4634	0.5689	-2.5724	0.0101
$\beta$ temp_last	0.3421	0.0601	5.6887	< 0.0001
$\beta$ rainfall_last	0.0001	0.0008	0.1373	0.8908
Log(theta)	-2.4569	0.1398	-17.5761	< 0.0001

Model 4(zero)	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	0.4018	0.5400	0.7439	0.4569
$\beta$ region 北	-14.8979	891.84135	-0.0170	0.9870
$\beta$ region 東	-0.4747	1.0072	-0.4713	0.6374
$\beta$ region 南	-0.0259	0.4456	-0.0581	0.9536
$\beta$ season 春季	-0.1337	0.7698	-0.1737	0.8621
$\beta$ season 秋季	-0.6734	0.4340	-1.5515	0.1208
$\beta$ season 夏季	-0.2665	0.4557	-0.5848	0.5587

模型計數的部分，大多數的變數皆呈現顯著，從係數估計的方向來看，與中區相比，北區與東區發生登革熱案例數較少，南區較多；而以季節來說，與冬季相比，秋季發生登革熱案例數較多，而春季及夏季則較少；且當上個月的



氣溫較高且雨量多時，可能也會有較多的登革熱案例。模型第二個部分為估計案例數為零的機率，各個地區及季節的顯著性都極弱，表示無論臺灣各區域的各個季節都還是會有登革熱發生的機會存在。另外，由上表也可發現，theta 呈現顯著不為 0，也就是說，我們應該採用負二項迴歸才較正確。

### (3) 零膨脹負二項式模型：(Model 5)

除了上述模型外，我們也根據相關文獻以及資料的有限程度，手動嘗試各種模型。

#### ● Count

$$\log(E(\text{cases} | X))$$

$$= \beta_0 + \beta_{\text{region}}I(\text{region}) + \beta_{\text{season}}I(\text{season}) + \beta_{\text{temp}_{\text{last}}}x_{\text{temp}_{\text{last}}} + \beta_{\text{rainfall}_{\text{last}}}x_{\text{rainfall}_{\text{last}}}$$

#### ● Zero

$$\text{Logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{\text{region}}I(\text{region})$$

Model 5(count)	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	-7.7779	1.1373	-6.8389	< 0.0001
$\beta$ region 北	-0.8943	0.4070	-2.1972	0.028
$\beta$ region 東	-0.5997	0.7008	-0.8557	0.3922
$\beta$ region 南	4.2647	0.5066	8.4179	< 0.0001
$\beta$ season 春季	-4.4432	0.5593	-7.945	< 0.0001
$\beta$ season 秋季	0.4553	0.4652	0.9786	0.3278
$\beta$ season 夏季	-1.4132	0.5480	-2.5789	0.0099
$\beta$ temp_last	0.0001	0.0008	0.1692	0.8656
$\beta$ rainfall_last	0.3459	0.0594	5.8259	< 0.0001
Log(theta)	-2.4764	0.1399	-17.7024	< 0.0001

Model 5(zero)	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	-0.1618	0.4733	-0.3419	0.7324
$\beta$ region 北	-12.2067	178.9677	-0.0682	0.9456
$\beta$ region 東	-0.5114	1.0811	-0.4730	0.6362
$\beta$ region 南	0.1767	0.4552	0.3882	0.6979

觀察模型配適的結果，在地區與季節上，和 Model 4 結論類似，南部較中部案例顯著地多，而北部與東部則是案例較少，秋季與夏季較冬季的案例多，而春季則是較少。並且，當上個月的氣溫較高且雨量多時，可能也會有較多的登革熱案例。而模型第二個部分為估計案例數零的機率的部分，各個地區及季節的顯著性都極弱，表示無論臺灣各區域的各個季節都還是會有登革熱發生的機會存在。

#### (4) 零膨脹負二項式模型模型：（Model 6）

我們也利用上述主成分分析所找出的貢獻度前三大的主成分，第一主成分可以解釋成溫度因子，第二主成分可以解釋成雨量因子，第三主成分可以解釋成氣壓因子，使用這三個主成分來配適模型。

##### ● Count

$\log(E(\text{cases} | X))$

$$= \beta_0 + \beta_{\text{region}}I(\text{region}) + \beta_{\text{season}}I(\text{season}) + \beta_{PC1}x_{PC1} + \beta_{PC2}x_{PC2} + \beta_{PC3}x_{PC3}$$

##### ● Zero

$$\text{Logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{\text{region}}I(\text{region}) + \beta_{\text{season}}I(\text{season})$$

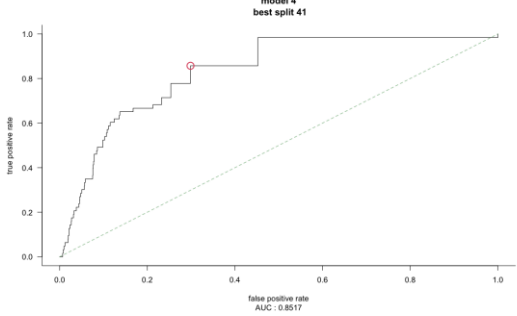
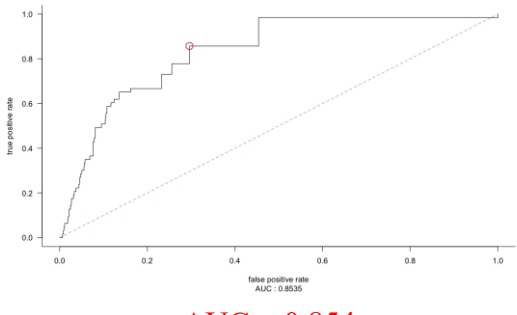
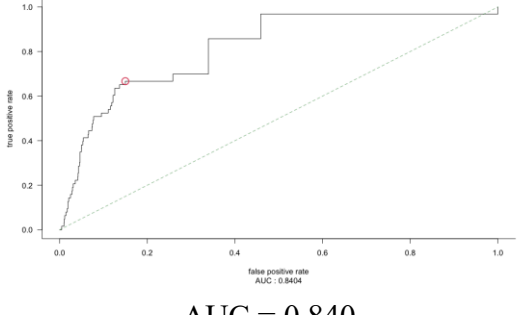
Model 6(count)	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	-1.2531	0.4658	-2.6903	0.0071
$\beta$ region 北	-0.2144	0.4118	-0.5206	0.6026
$\beta$ region 東	-0.9098	0.6121	-1.4865	0.1371
$\beta$ region 南	4.8210	0.4610	10.4582	<0.0001
$\beta$ season 春季	-4.4845	0.5154	-8.7012	<0.0001
$\beta$ season 秋季	2.1955	0.3495	6.2813	<0.0001
$\beta$ season 夏季	0.7060	0.3878	1.8206	0.0687
$\beta$ pc1	1.2154	2.8057	0.4332	0.6649
$\beta$ pc2	0.9219	2.0628	0.4469	0.6549
$\beta$ pc3	-5.9025	1.8395	-3.2087	0.0013
Log(theta)	-3.1380	0.0791	-39.6775	<0.0001

Model 6(zero)	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	-9.4683	355.1227	-0.0267	0.9787
$\beta$ region 北	-4.8934	428.6847	-0.0114	0.9909
$\beta$ region 東	-0.8067	374.7736	-0.0022	0.9983
$\beta$ region 南	-5.5975	117.6204	-0.0476	0.9620

觀察模型配適的結果，在地區與季節上，和前幾個模型結論類似，南部較中部案例多，而北部與東部則是案例較少，秋季與夏季較冬季的案例多，而春季則是較少。在主成分的部分，氣溫因子的係數為正，表示氣溫愈多，登革熱案例愈多；雨量因子的係數為負，表示降雨愈多，登革熱案例愈少；氣壓因子係數為負，表示氣壓愈高，登革熱案例愈少，氣壓的結論比較特別，可以再做進一步地討論。而模型第二個部分為估計案例數零的機率的部分，各個地區及季節的顯著性都極弱，表示無論臺灣各區域的各個季節都還是會有登革熱發生的機會存在。

## 2.模型選擇與比較

從候選模型(Model 4, Model 5, Model 6)進行比較，選擇模型的標準是以 AIC 及 BIC 的大小衡量，以及接收者操作特徵曲線（Receiver Operating Characteristic Curve, ROC 曲線）計算曲線下面積(Area under the Curve, ROC)的大小來比較各模型的優劣。由上表可知，Model 5 有最小的 AIC 及 BIC，且有最大的 AUC，因此我們選擇了 Model 5 來解釋我們的資料。

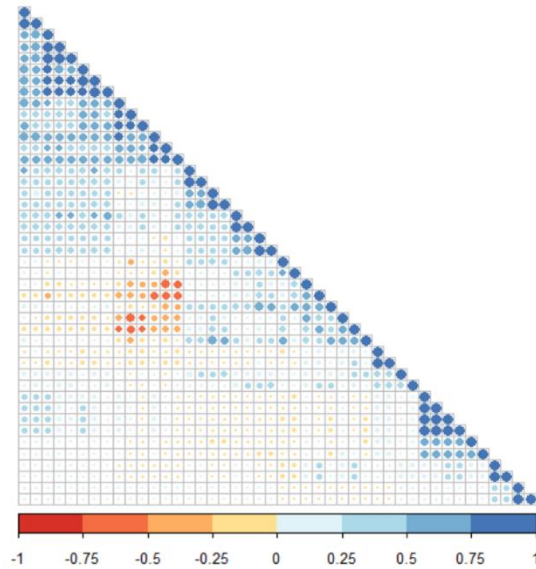
Model	AIC	BIC	ROC & AUC
Model 4	2779.27	2861.75	 <p>AUC = 0.852</p>
Model 5	2730.43	2807.41	 <p>AUC = 0.854</p>
Model 6	2733.85	2733.85	 <p>AUC = 0.840</p>

### （四）樹狀模型

#### 1.資料前處理（僅敘述因子分析、決策樹有用到的預處理方法）

##### ● 變數過多

我們所採用的資料變數有 25 個，包含諸多氣候（雨量、氣壓等）、社會科學（人口數、縣市區域等）類型的變數，透過參考文獻的建議，我們亦將一個月前的資料做為變數，因此變數便有 50 個。

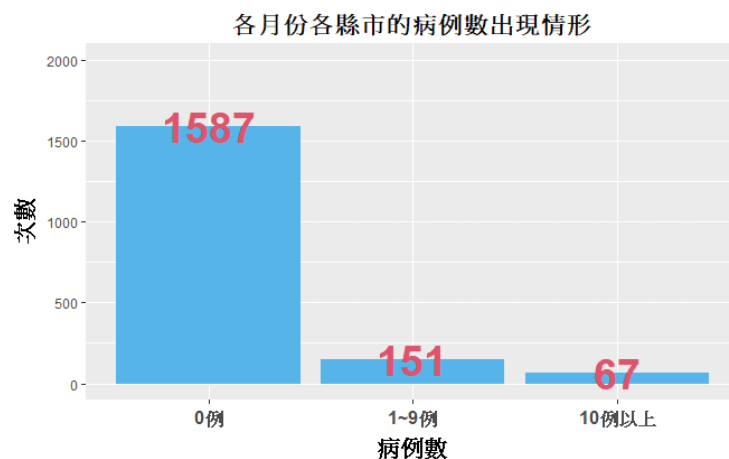


再從上面本資料各變數的 Correlation Plot 可以看到，很多變數之間都有高度相關性，變數之間存在共線性的問題，因此這會造成我們的分類結果不夠好，而我們研究的目的是想透過樹狀模型進行分群，並且能找到影響分群結果的變數、變數如何影響。為了達到我們的目的，我們試圖找尋既能降低變數間的相關性，又能保留變數變異、進行分群，甚至具有預測力的方法。

### ● 反應變數改為類別型、考慮資料不平等的處理

本資料的反應變數—某月份某縣市登革熱病例數為可數型 (countable)，在事前嘗試模型配適時，我們直接將模型，由於各病例數出現次數差異大，病例數愈多出現次數愈少，且病例數只差一點沒有太大意義，模型很難用有限的變數判斷出其中的差異，因此我們將病例數分為 0 例、1~9 例、10 例以上。

此分類考量了資料不平等的問題，選擇 0 例、9 例作為切割點，並為了使模型能訓練得好，僅分成 3 類。然而，由於病例數為 0 的觀察值個數過多，為避免資料不平等造成模型訓練上的困難，我們在後續的分析中考慮了兩種資料選取的情況—「納入病例數為 0」的資料、「排除病例數為 0」的資料，分別進行分析。下圖為各個病例數類別的出現次數長條圖。



## ● 資料正規化處理

本資料的解釋變數包含雨量、氣壓、日照時數等變數，由於其表示單位不一，若以原始資料直接進行因子分析，會使結果有偏差，因此我們對連續型變數進行標準化，而類別型變數則進行 Label Encoding，由於 GBDT 模型的輸入限制，再放入該模型前我們更將類別型變數進行了 One Hot Encoding。

## 2.降低變數共相關與縮減維度之處理--因子分析

我們使用因子分析的兩種方法，「最大概似法 (MLM)」與「主成分法 (PCM)」，前面提到分析分成考慮/不考慮病例數為 0 的情況，兩種情況會各進行 MLM、PCM 的因子分析，因此會有 4 種組合，我們直接以 R 程式進行分析演算。

進行因子分析時我們僅考慮連續型變數，首先取出資料正規化後屬於連續型變數的部分資料，代入 psych 套件的 fa.parallel 函式挑選最適因子個數，此函式提供三種挑選的準則，我們選擇 parallel 法挑選。選好最適因子個數後，即進行因子分析，MLM 法使用 psych 套件的 fa 函式，並將 fm 設為 ml，PCM 法使用 psych 套件的 principal 函式，而為了讓因子能更具代表性，我們進行了直交旋轉，採用「因子變異最大化 (Varimax)」的方法。輸入以上函式設定後即能直接得到因子分數（於輸出結果的 score 部分），我們以少量因子個數的因子分數，代替為數眾多的原始資料變數，而我們還會依據因子負載量（原始資料變數與因子的相關性）命名因子，以使分群結果能有足夠解釋力。命名因子的方法及結果呈現於下方結論的章節。

## 3.如何表示因子的意義—因子寫成變數的線性組合

因子負載量可說是變數寫成因子的線性組合的係數，然而我們將連續型變數以因子替代後，我們在意的是如何表示因子的意義，若能將因子寫成變數的線性組合則能達到我們的目的，所幸 fa 函式與 principal 函式的均可輸出因子寫成變數的線性組合的係數（於輸出結果的 weights 部分），我們在下方結論的決策樹圖表說明中，會附上樹狀圖節點中有出現的因子如何以變數的線性組合表示，然而因原始資料的變數眾多，我們僅列出係數大於 0.05，或最多 6 個變數的線性組合。

## 4.分群演算--決策樹、梯度提升決策樹：

進行因子分析後，我們以因子取代連續型變數，達到縮減維度的效果，隨後將因子與編碼後的類別型變數放入決策樹 (Decision Tree, DT) 模型中訓練。由於模型配適結果不如預期，我們又嘗試以決策樹的改良模型—梯度提升決策樹 (Gradient Boost Decision Tree, GBDT) 進行分析，並比較兩模型的預測準確度有無顯著差異。前述因子分析會產出 4 種組合的結果，每種結果都會放入決策樹及梯度提升決策樹模型中訓練，因此最終會有 8 種結果。

## 5.模型驗證、調整--交叉驗證、網格搜尋法調整參數

一開始我們將全部資料均當作訓練資料放入模型，然而無法得知模型的好壞，更發生決策樹無法進行分支的情況，而且本研究的目標是希望能將訓練出的模型用在預測下一年度的登革熱發生情況，因此我們將資料重新切分，先隨機取出 20%的資料作為測試資料，剩下的 80%資料作為訓練及驗證用。

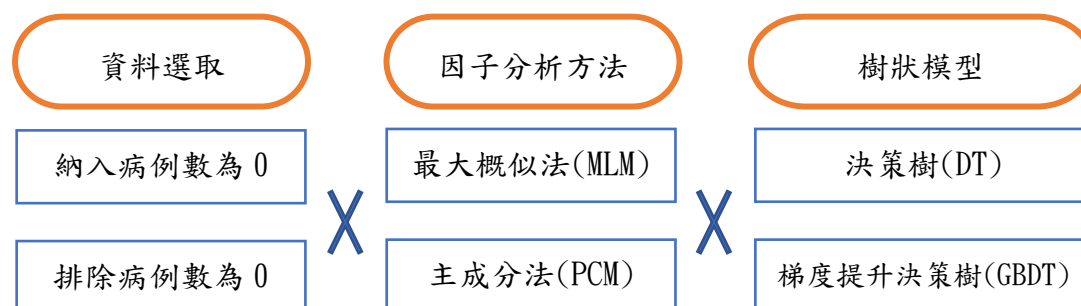
考量前述資料不平等的問題，我們進行了 2 種驗證方法，對於「排除病例數為 0」的情況，採用 10-fold Cross Validation，即訓練資料隨機切成 10 等份，每次取出 1 份作為驗證資料，其餘 9 份作為訓練資料，每份將輪流作為驗證資料。而對於「包含病例數為 0」的情況，則採用 Stratified 10-fold Cross Validation，即在每個類別（0 例、1~9 例、10 例以上）均各自進行 10-fold Cross validation，可避免資料不平等為模型帶來的偏差，並將切分完的資料再進行 SMOTE 演算使資料更加平衡。

我們進行交叉驗證的目的是為了能夠找到合適的模型，為了使模型更合理，我們同時利用網格搜尋法（Grid Search）進行參數調整，而我們進行了三階段的調整，即利用前一階段的最佳參數作為下一階段的參數搜尋範圍設定，逐步縮小參數搜尋範圍，重複三次的搜尋，以找到更合理的參數設定。

以上的模型調整方法都透過了 caret 套件的 trainControl 和 train 函式進行分析。

## 6.模型評估指標—Accuracy、AUC

下圖為 8 種分析組合的組合方法，我們想以此方式進行組合有其意義。由於此資料有 8 成以上的病例數都是 0 例，因此將資料選取分成考慮/不考慮病例數為 0 的情況，想觀察兩種情況對於結果（預測效果評估指標、變數重要性、決策樹分支）有無差異。而儘管我們運用因子分析，將為數眾多的變數改以因子表示，但因子分析取出因子的方法會影響各因子的負載量、組成、如何解釋，因此我們採用 MLM、PCM 兩種因子分析方法，欲比較兩種方法的預測效果是否有差異，並以較好的方法之結果作視覺化呈現。雖然決策樹的樹狀圖結果可以很清楚地展示模型如何分支，然而決策樹僅用一棵樹分群，結果可能有偏差，預測效果可能也不太好，因此我們還考慮了梯度提升決策樹模型，想比較兩模型的預測效果，預期能看到梯度提升決策樹的逐步學習處理能改善預測效果。



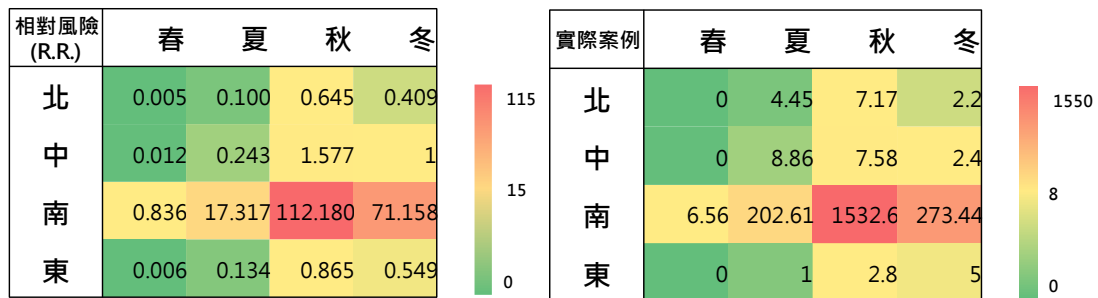


## 五、結論

### (一) 廣義線性模型

由於上述配適模型以及模型選擇比較的結果，Model 5 為我們所認為最佳的模型，因此對此模型做進一步的說明。

左表為透過 Model 6 所估計的每個地區每個季節登革熱案例數的相對風險熱度圖，這邊的相對風險定義為，模型所估計的案例數與冬季中部（基準）估計的案例數的比值；右表為實際資料的平均每月病例數（僅計算案例數大於零的平均）的熱度圖。



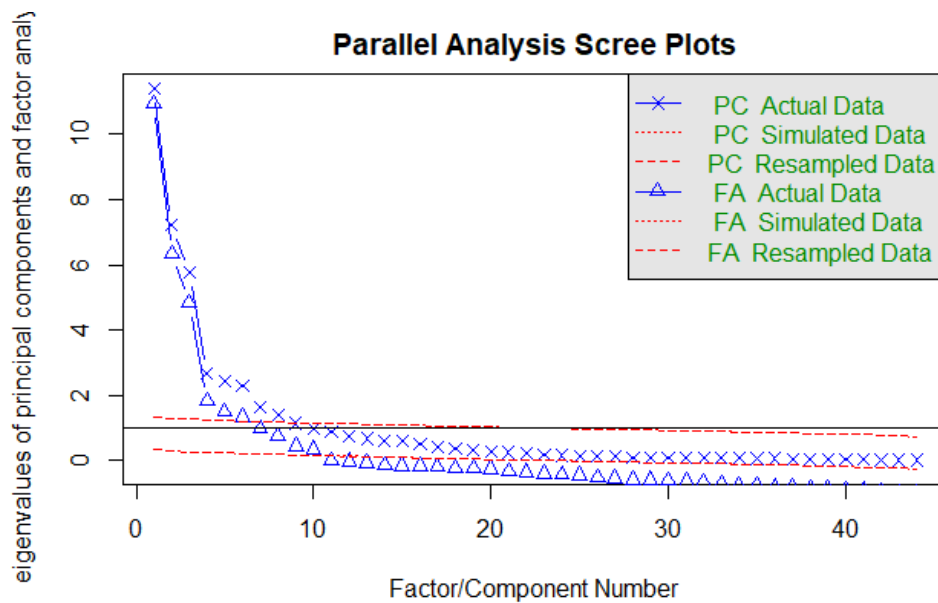
從兩個表比較，可以發現模型估計的效果還算不錯，相對風險的大小大致有呈現實際案例的情況。並且，可以發現南部秋季與冬季的相對風險最高，實際發生的平均案例也最高，代表南部秋季登革熱相當的盛行。至於各區域的春季相對風險都是四季最低，實際發生的平均案例也是同樣的情形，代表春季是幾乎不會有登革熱的案例發生。

### (二) 樹狀模型

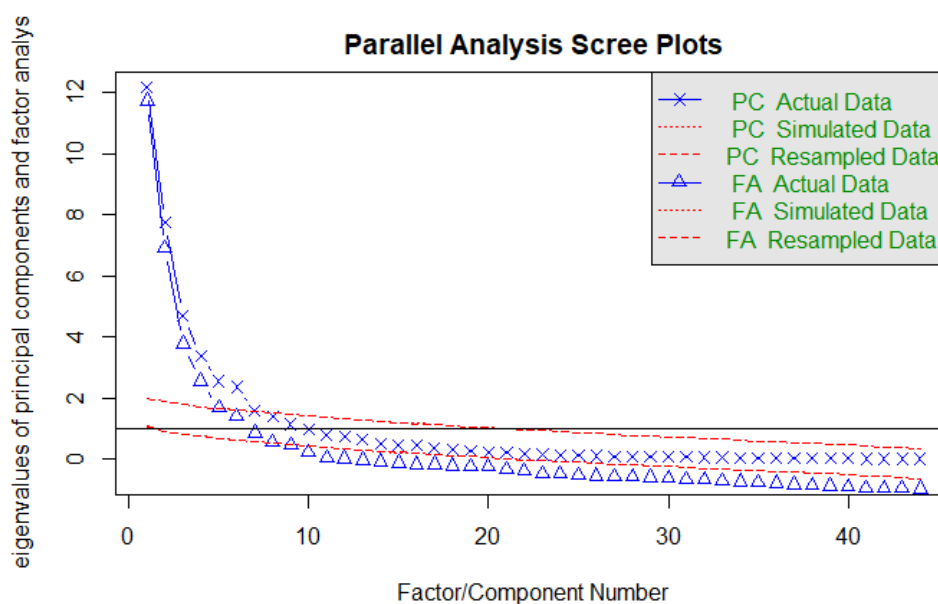
#### 1. 因子分析之最適因子個數

由於我們使用兩種因子分析的子方法，最大概似法(MLM)與主成分法(PCM)，且又分別考慮病例數為 0 與不考慮的情況，共 4 種組合，以下僅呈現我們認為最終分類之最好結果，其餘 3 種情況則放在附錄。

首先透過 psych 套件 fa.parallel 函式挑選最適合因子個數，內建 Parallel 法決定最適數量，由於我們是針對因子分析，因此僅需觀察圖中三角形藍線與下方紅虛線的交會點。



上圖為「納入病例數為 0」的情況下，主成分與因子分析的各主成分、因子之 eigenvalue。圖中可以看到，在選到第 11 個因子時，我們資料的實際特徵值（Actual Data，三角形藍線）小於重抽樣多次迭代後計算的特徵值平均（Resampled Data，下方紅虛線），所以我們最終選定 eigenvalue 前 10 大的因子，做後續的分析。



上圖為「排除病例數為 0」的情況下，主成分與因子分析的各主成分、因子之 eigenvalue。圖中可以看到，在選到第 9 個因子時，我們資料的實際特徵值（Actual Data，三角形藍線）小於重抽樣多次迭代後計算的特徵值平均（Resampled Data，下方紅虛線），所以我們最終選定 eigenvalue 前 8 大的因子，做後續的分析。



下表為資料選取與因子分析方法構成的 4 種組合中，各組合的最適因子個數與累積解釋變異比例，可發現各組合的累積解釋變異比例都在 80% 左右。

資料選取	納入病例數為 0		排除病例數為 0	
因子分析方法	MLM	PCM	MLM	PCM
k=最適因子個數	10	10	7	7
前 k 個因子累積解釋變異比例	79.1%	83.8%	72.4%	78.3%

## 2. 解釋因子意義（為因子命名）：

雖然由我們使用的 psych 套件中的 fa.diagram 函式，即能以視覺化方式看到因子對應各變數的負載量，然而由於我們取用的因子個數眾多，因此以各負載量數值大小命名之。

Loadings:	ML3	ML4	ML9	ML6	ML5	ML10	ML7	ML1
temp	0.778	0.139	0.229	0.507		0.161		
tempMax	0.641	0.203	0.230	0.559		0.121	-0.134	
tempMin	0.790		0.239	0.439		0.188		
rainfall	0.171		0.902			0.151		
rainfall_Hour	-0.141		0.495	-0.457	0.332		0.201	0.224
rainfall_Day		-0.120	0.752	-0.164	0.294	0.134	0.107	
rainfall_10minMax	0.261		0.665	0.263		0.162		
rainfall_60minMax	0.248		0.697	0.195		0.155		
rainfall_1DayMax	0.216		0.839			0.167		
stn_pressure	0.114	0.980						
stn_pressureMax		0.938						
stn_pressureMin		0.580	-0.149					0.117
tdewpoint	0.446	0.267	0.313	0.633		0.143		
windSpeed		0.142		-0.124			0.924	
wind_strMax	0.137		0.312			0.107	0.479	0.126
sunshine	0.268		-0.205	0.709	-0.388	0.109		
rh		-0.107	0.506		0.250			-0.160
evapA	0.310	0.128		0.791	-0.193			-0.112
global_radio	0.285	-0.129		0.811	-0.238			
cloud_amount		0.146	0.392	-0.326	0.529		0.123	
population_den		0.207						0.954
population								0.133

temp_last	0.918	0.131	0.109	0.143	-0.195	0.211		
tempMax_last	0.854	0.201	0.130	0.196	-0.195	0.170		
tempMin_last	0.900				-0.154	0.244		
rainfall_last	0.171		0.187		0.236	0.896		
rainfall_Hour_last	-0.192			-0.234	0.605	0.367	0.171	0.225
rainfall_Day_last		-0.142	0.268		0.613	0.543		
rainfall_10minMax_last	0.366		0.260	0.185		0.571		
rainfall_60minMax_last	0.321		0.204	0.163		0.634		
rainfall_1DayMax_last	0.229		0.172			0.878		
stn_pressure_last		0.983						
stn_pressureMax_last		0.943						
stn_pressureMin_last		0.581						0.123
tdewpoint_last	0.552	0.261	0.170	0.371		0.241		
windSpeed_last	-0.140	0.160					0.930	
wind_strMax_last						0.353	0.496	0.121
sunshine_last	0.319			0.348	-0.787			
rh_last		-0.126			0.473	0.353		-0.176
evapA_last	0.416	0.122	0.112	0.500	-0.448			-0.125
global_radio_last	0.464	-0.162	0.170	0.488	-0.579			
cloud_amount_last		0.123	0.139		0.826	0.129		
population_den_last		0.207						0.954
population_last								0.133

上圖為資料「納入病例數為 0」的情況下，使用 MLM 進行因子分析後，各變數對應到各因子的負載量（loading），由於我們因子分析時都有加上「因子變異最大化」的處理，因此給定同個因子時，各變數的因子負載量會呈現大者更大，小者更小的趨勢，使我們能容易察覺負載量較大的變數，能更方便為因子命名。

首先可看到上圖紅圈圈起的部分，此即為該行因子中，負載量相對較大的變數（通常只看負載量絕對值>0.5 的部分），可看到第一行的 ML3 因子中，負載量相對大的變數包含 temp, tempMax, tempMin, temp\_last, tempMax\_last, tempMin\_last，這些變數就是前一個月的平均溫度、每日最高溫度、每日最低溫度，及當月的平均溫度、每日最高溫度、每日最低溫度，由於當月份與前月份有關溫度的變數的負載量都較大，因此命名為「溫度因子」，以此類推，第二行 ML4 因子會命名為「氣壓因子」。然而，第三行的 ML9 因子對應到的較大負載量僅有當月份有關降雨量的變數，因此會命名為「當月雨量因子」，第四行 ML6 因子命名為「當月日照量蒸發量因子」，第五行 ML5 因子命名為「前月降雨日數因子」。由於因子個數眾多，僅列至此。

### 3.8 種分析組合的模型評估指標

下表為前述 4 種因子分析組合放入決策樹（DT）、梯度提升決策樹（GBDT）兩模型，並進行模型驗證、調整後，8 種組合的預測資料評估指標，提供 Accuracy, AUC 兩種指標的結果。

資料選取	納入病例數為 0				排除病例數為 0			
模型	DT		GBDT		DT		GBDT	
因子分析方法	MLM	PCM	MLM	PCM	MLM	PCM	MLM	PCM
Accuracy	0.8694	0.8611	0.8806	0.8222	0.6512	0.7209	0.7674	0.7209
AUC	0.7794	0.7034	0.7479	0.774	0.8227	0.673	0.8773	0.8132

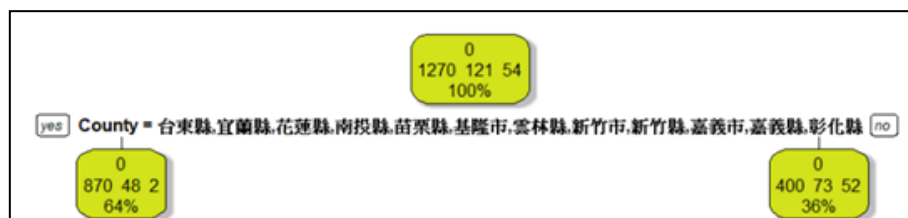
首先考慮「納入病例數為 0」的情況下的 4 種分析組合，我們可以發現 DT 與 GBDT 兩模型的預測效果差不多，GBDT 的迭代學習沒有讓模型的預測準確度改善很多，推測可能是因子能解釋的變異不夠多，使得模型就算經過多次驗證、參數調整後，仍未有顯著的改善。就不同因子分析方法上的比較，經由 Accuracy，可發現 MLM 法的預測準確度均較高，而 AUC 則是 4 種組合均為類似的結果。

再來考慮「排除病例數為 0」的情況下的 4 種分析組合，我們可以發現 GBDT 模型的結果都比 DT 模型好，可見 GBDT 模型有改善一些 DT 模型無法訓練到的地方。而就不同因子分析方法上的比較，可發現使用 DT 模型時，PCM 法的 Accuracy 較高，然而 MLM 法的 AUC 比 PCM 法的高出許多，使用 GBDT 模型時，MLM 法的 Accuracy, AUC 都比較好。

綜合以上兩種模型評估指標的結果，我們認為 MLM 法大多數的結果比

PCM 法好，然而 DT 模型與 GBDT 模型的結果沒有顯著的差異，我們在下方的視覺化結果呈現中，DT 模型均會展示 MLM 法分析的結果，而 GBDT 模型在納入病例數為 0 的情況下將展示 PCM 法的結果，排除病例數為 0 的情況下將展示 MLM 法的結果。

#### 4. 決策樹



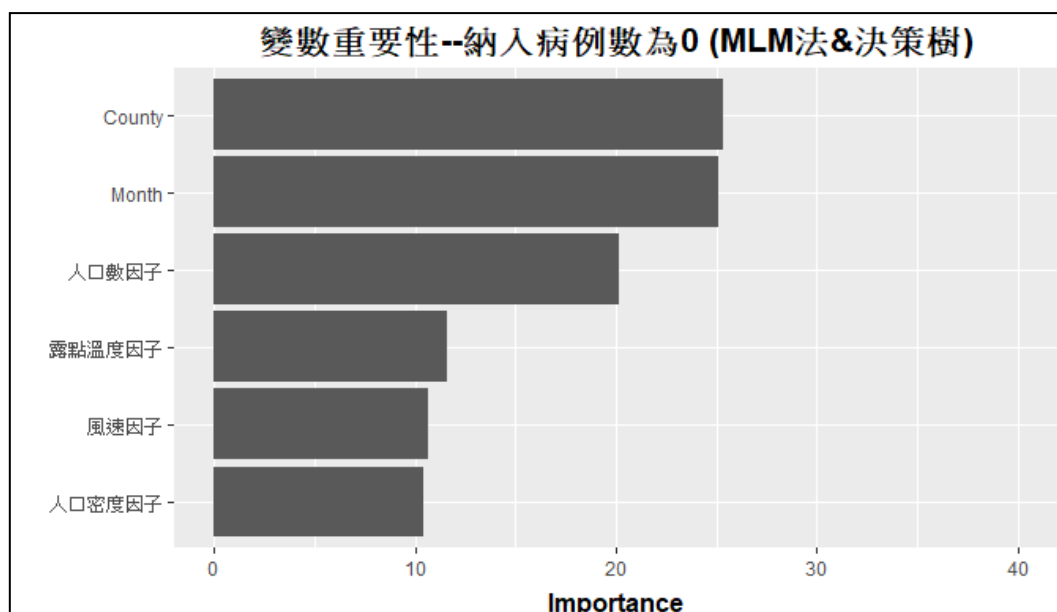
首先介紹節點內數值的意義，節點上層呈現的是該節點觀測值中出現頻率最高的類別，中層呈現的是該節點各個類別的出現次數，在資料納入病例數為 0 的情況下會出現 3 個數值，分別是 0 例、1~9 例、10 例以上出現的次數，而在資料排除病例數為 0 的情況下會出現 2 個數值，分別是 1~9 例、10 例以上出現的次數，而節點下層呈現的是該節點的觀測值總數占全部觀測值的比例，可知若此比例愈大且各類別出現次數差異大，表示此節點較為清晰。

至於顏色的呈現，在資料納入病例數為 0 的情況下，當節點顏色愈黃，表示該節點出現 0 例的比例較高，顏色愈接近綠色，表示該節點出現 1~9 例的比例較高，愈接近青色，表示該節點出現 10 例以上的比例較高。而在資料排除病例數為 0 的情況下，節點顏色愈黃，表示該節點出現 1~9 例的比例較高，愈接近青色，表示該節點出現 10 例以上的比例較高。除了末端節點，每個節點下方均有一行條件式，當滿足該條件式，會分到左下方節點，當不滿足時，則會分到右下方節點。

由於原始樹狀圖分支過多，以下的樹狀圖均為事後剪枝（post-prune）的結果，我們設定層數最高至 6 層、每個葉節點需至少包含 15 個觀測值，最底層每個節點需至少包含 5 個觀測值。

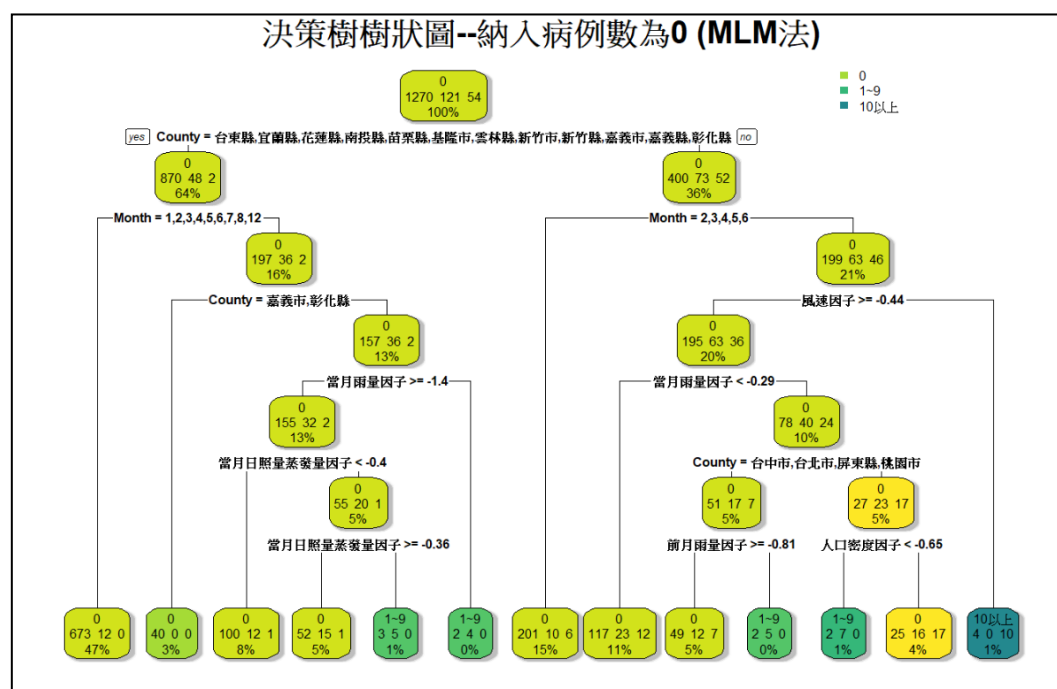
# (1)納入病例數為 0 的情況 (MLM 法結果)

## ● 變數重要性圖



上圖為資料納入病例數為 0 的情況下，MLM 法的決策樹模型的重要變數/因子圖，僅列出重要性前 6 大的因子，重要程度愈高表示在分支時產生的變異/資訊量愈大，上圖未顯示的人口數因子亦為重要程度高的因子，此因子未出現在樹狀圖中可能是因為縣市與人口數有些共相關，使得僅有縣市因子出現在樹狀圖中。

## ● 樹狀圖



上圖為資料納入病例數為 0 的情況下，MLM 法的決策樹模型樹狀圖。可發現最上層的幾個節點都是根據縣市 (County)、月份 (Month) 分支，代表這

兩個因子相當重要，可發現在秋季（9~11月）時，傾向分到病例數非0的組別。我們可以發現，當縣市為台南或高雄、月份在1月或7~12月、風速因子不小於0.44時，傾向分到病例數為10例以上。

整體上我們可以發現，這個樹狀圖分群分得不是很好，僅縣市、月份兩個因子能將大部分病例數為0例的情況分開，當月雨量因子亦能將一些病例數為1~9例的情況分開，但多數仍會與病例數為0例的觀測值混雜在一起，雖然第四列分節的因子均為當月雨量因子，然而無法看出此因子數值愈大愈傾向分到哪個組別，可推測本研究資料的解釋變異不足，或是考慮病例數為0例的情況會使解釋增加困難。

下表為上圖於分節時所出現的5個因子寫成變數的線性組合之表示，以下表示式之解釋變數均以還原成標準化前的原始變數。

風速因子 =

$$0.563 * \text{windSpeed\_last} + 0.422 * \text{windSpeed} - 0.07 * \text{stn\_pressure\_last} - 0.057 * \text{stn\_pressure}$$

當月雨量因子 =

$$0.431 * \text{rainfall} - 0.258 * \text{temp\_last} + 0.222 * \text{rainfall\_1DayMax} + 0.147 * \text{tddewpoint} + 0.141 * \text{stn\_pressure\_last} + 0.14 * \text{rainfall\_Day}$$

當月日照量蒸發量因子 =

$$- 0.811 * \text{temp\_last} + 0.578 * \text{temp} + 0.39 * \text{tddewpoint} + 0.161 * \text{global\_radio} - 0.156 * \text{tempMin\_last} + 0.155 * \text{evapA}$$

前月雨量因子 =

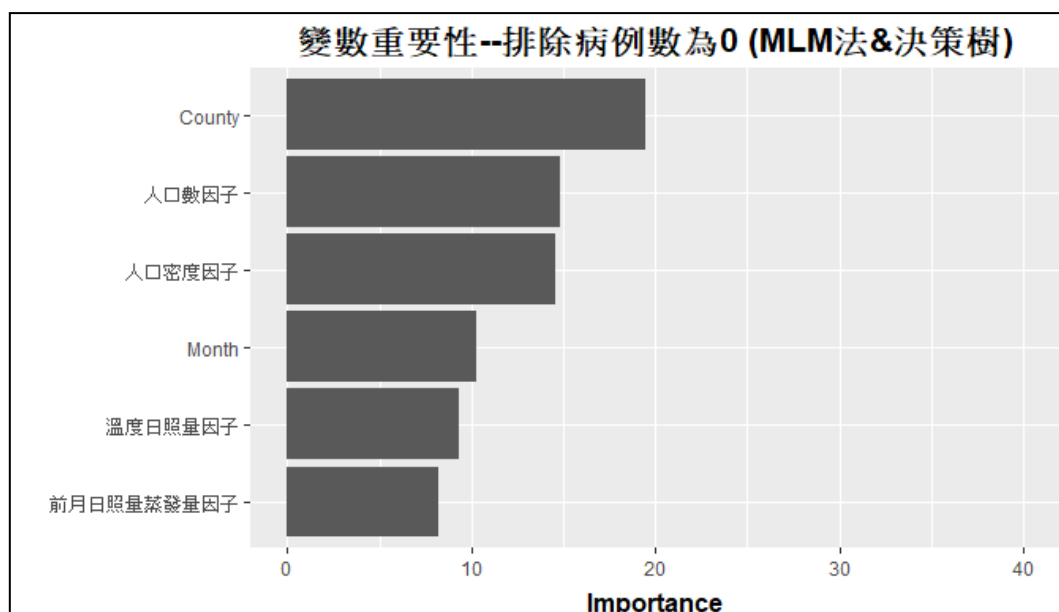
$$0.694 * \text{rainfall\_last} + 0.323 * \text{rainfall\_1DayMax\_last} - 0.253 * \text{temp} + 0.176 * \text{stn\_pressure} - 0.13 * \text{tddewpoint} + 0.096 * \text{sunshine\_last}$$

人口密度因子 =

$$0.544 * \text{population\_den} + 0.544 * \text{population\_den\_last} - 0.089 * \text{stn\_pressure} - 0.088 * \text{stn\_pressure\_last} - 0.08 * \text{population} - 0.08 * \text{population\_last}$$

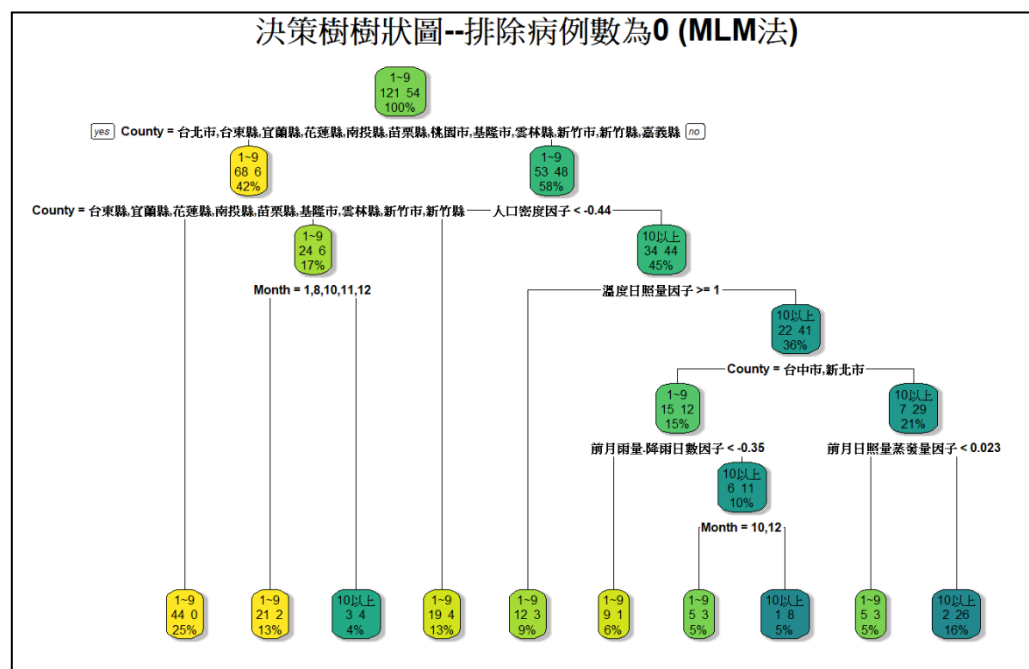
## (2)排除病例數為 0 的情況 (MLM 法結果)

### ● 變數重要性圖



上圖為資料排除病例數為 0 的情況下，MLM 法的決策樹模型的重要變數/因子圖，僅列出重要性前 5 大的因子，重要程度愈高表示在分支時產生的變異/資訊量愈大，可發現縣市、人口數、人口密度等因子均為重要程度高的因子，此結果與納入病例數為 0 的情況類似，比較不同的是在排除病例數為 0 的情況下，與日照量、雨量有關的因子（溫度日照量因子、前月日照量蒸發量因子）的重要程度都較高。

### ● 樹狀圖



上圖為資料排除病例數為 0 的情況下，MLM 法的決策樹模型樹狀圖。與納入病例數為 0 的結果相比，可發現首先分枝的準則同樣為縣市，（分到左側條件

的)縣市同樣沒有台中、台南、高雄、屏東等人口數較多且位於中南部的縣市，而對於病例數傾向較多的縣市(第一次分支分到右側)，第二次分支準則為人口密度因子，可知人口密度較低的縣市傾向分到病例數較少的組別。

我們觀察到本圖左側，當縣市為東部縣市(宜蘭、花蓮、台東)與中北部非核心縣市(新竹縣市、苗栗、基隆、南投、雲林)等，均完全沒有登革熱病例，而若非這些縣市(即台北、桃園、嘉義)，且在2~7,9月時，傾向分到病例數較多的組別。

而若觀察到本圖右側，可發現當縣市非台中、台北(可看出應為台南、高雄、屏東等南部縣市)、前月日照量蒸發量因子數值較大時，傾向分到病例數較多的組別，而當縣市為台中、台北時，前月雨量、降雨日數因子較大時，也傾向分到病例數較多的組別。由此結果與下表的因子之變數線性組合可觀察到，前月份的雨量愈多、當月份溫度愈低、前月份最大日雨量愈多、前月份蒸發量愈大，均會傾向分到病例數較多的組別，此結果也應證了我們所參考的文獻中提到了前月雨量會影響當月登革熱疫情的情形。

下表為上圖於分節時所出現的4個因子寫成變數的線性組合之表示，以下表示式之解釋變數均以還原成標準化前的原始變數。

人口密度因子 =

$$0.539 * \text{population\_den\_last} + 0.539 * \text{population\_den} - 0.133 * \text{temp} \\ - 0.103 * \text{population\_last} - 0.103 * \text{population} + 0.089 * \text{sunshine\_last}$$

溫度日照量因子 =

$$0.747 * \text{temp} + 0.074 * \text{global\_radio\_last} + 0.063 * \text{sunshine\_last} \\ + 0.056 * \text{tempMax} - 0.053 * \text{rainfall\_last}$$

前月雨量、降雨日數因子 =

$$0.499 * \text{rainfall\_last} - 0.322 * \text{temp} + 0.240 * \text{rainfall\_1DayMax\_last} \\ - 0.237 * \text{stn\_pressure\_last} + 0.146 * \text{stn\_pressure} + 0.139 * \text{rainfall\_Day\_last}$$

前月日照量蒸發量因子 =

$$- 1.158 * \text{temp} + 0.5 * \text{sunshine\_last} + 0.41 * \text{global\_radio\_last} \\ + 0.248 * \text{rainfall\_last} + 0.165 * \text{rainfall\_1DayMax\_last} + 0.152 * \text{evapA\_last}$$

## 5. 梯度提升決策樹

決策樹的結果很清楚直覺，然而光靠一棵樹解釋樹狀分支並不簡單，因此我們想知道梯度提升決策樹的結果與決策樹的差別。隨機森林使用 Bagging 法，綜合多個樹模型結果，可以降低單一樹模型的高變異性並提升預測正確率，且是由「去相關性」的樹模型所組成的集成演算法，對於觀測值、變數都會進行隨機子集選取。而梯度提升決策樹又是隨機森林的改良，有別於隨機森

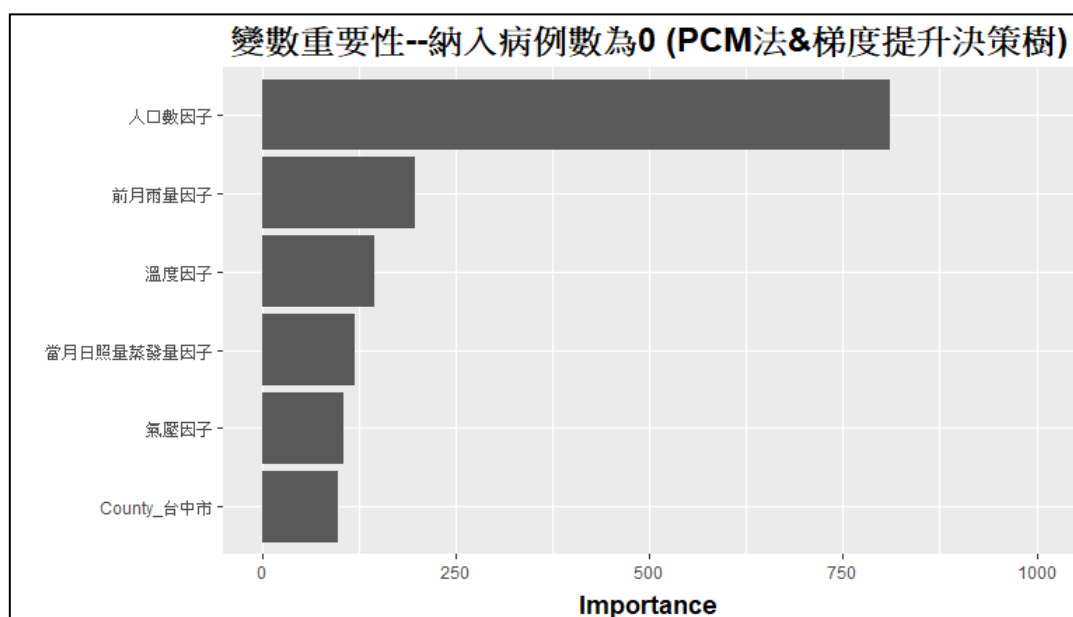


林集成眾多深且獨立的樹模型，梯度提升決策樹使用 Boosting 法，集成諸多淺且弱連續的樹模型，每次生成的新的決策樹都是要修正前面舊的樹在預測上的錯誤，除了可減少單一決策樹產生的變異，透過逐步學習修正還能減少誤差。

由於梯度提升決策樹是用非常多棵樹所構成的模型，我們無法以樹狀圖清楚展示此模型的結果，然而，我們可以觀察模型中每個解釋變數對於反應變數（病例數）的邊際效應/貢獻（Partial Dependence），R 的 pdp 套件中的 Partial 函式即能進行變數邊際變數對 y 效應圖（Partial Dependence Plot）的繪製，然而本資料的反應變數為多元類別（multi-label），因此我們選擇自行設計 Partial Dependence Plot，而我們僅呈現重要性前 6 高的變數的 Partial Dependence Plot，每張子圖 x 軸的範圍是該解釋變數的第 25 百分位數與第 75 百分位數區間，y 軸表示分到各類別的機率。圖中線條顏色的設定與前面決策樹樹狀圖類似，在資料納入病例數為 0 的情況下，病例數為 0 例、1~9 例、10 例以上三個類別分別以黃、淺綠、深綠/青色表示，而在資料排除病例數為 0 的情況下，病例數為 1~9 例、10 例以上二個類別分別以黃、淺綠表示。

### (1) 納入病例數為 0 的情況（PCM 法結果）

- 變數重要性圖（注意：此圖 x 軸與其他變數重要性圖不同）

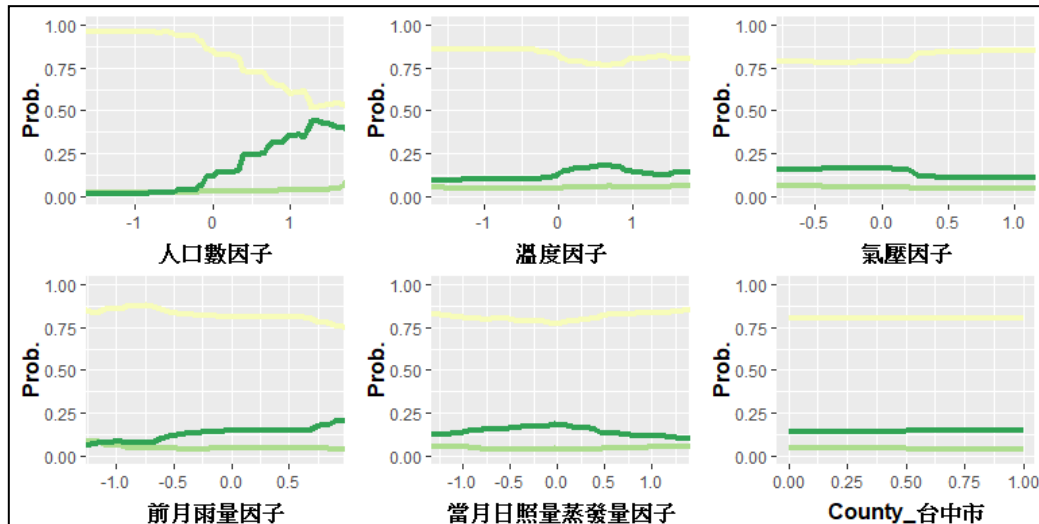


上圖為資料納入病例數為 0 的情況下，PCM 法的梯度提升決策樹模型的變數重要性圖。我們可以發現，此模型重要程度前 6 高的變數與決策樹模型不太一樣，在決策樹模型中，縣市（County）、月份（Month）都是重要程度非常高的變數，在此不同，可能是因為梯度提升決策樹在建模過程中會直接將類別型變數轉化為 dummy variable，使得類別型變數中的每個 level 都變成了新的變數，雖然這能讓我們更加了解哪些縣市、哪些月份較為重要，但我們無法對這些變數在決策樹與梯度提升決策樹的結果中作比較，且在 Partial Dependence Plot 的結果呈現中，資訊會顯得較少。



而我們可看到，人口數因子是重要性最高的變數，且重要程度比其他變數高出許多，而在前兩個決策樹模型與此模型中，人口數因子與前月雨量因子都是重要性很高的變數，與決策樹模型結果較不同的地方是，氣壓因子在此模型亦為較重要的變數。

● Partial Dependence Plot



上圖為資料納入病例數為 0 的情況下，PCM 法的梯度提升決策樹模型的 Partial Dependence Plot，黃、淺綠、深綠/青三種顏色的線分別表示病例數為 0 例、1~9 例、10 例以上三個類別。首先我們可以發現每個子圖的黃線都沒有與其他線交會，這代表不論每個解釋變數的邊際變動量為何，分類至病例數為 0 的機率均為最大，這非常有可能是病例數為 0 的情況過多的因素。然而當人口數因子數值愈大，分到 10 例以上的機率會顯著增加，接近分到 0 例的機率。前月雨量因子的數值愈大，分到 10 例以上的機率亦有增加。

下表為上圖所提到的因子寫成變數的線性組合之表示，以下表示式之解釋變數均以還原成標準化前的原始變數，可酌以參考。

人口數因子 =

$$0.478 * \text{population} + 0.478 * \text{population\_last} + 0.086 * \text{rh\_last} + 0.07 * \text{rh} \\ - 0.069 * \text{rainfall\_10minMax\_last} - 0.065 * \text{rainfall\_60minMax\_last} \\ - 0.063 * \text{rainfall\_10minMax}$$

溫度因子 =

$$0.283 * \text{tempMin\_last} + 0.267 * \text{temp\_last} + 0.241 * \text{tempMax\_last} \\ + 0.216 * \text{tempMin} + 0.187 * \text{temp} + 0.133 * \text{tempMax}$$

氣壓因子 =

$$0.222 * \text{stn\_pressureMax\_last} + 0.221 * \text{stn\_pressure\_last} \\ + 0.219 * \text{stn\_pressureMax} + 0.218 * \text{stn\_pressure} + 0.128 * \text{stn\_pressureMin} \\ + 0.126 * \text{stn\_pressureMin\_last}$$

前月雨量因子 =

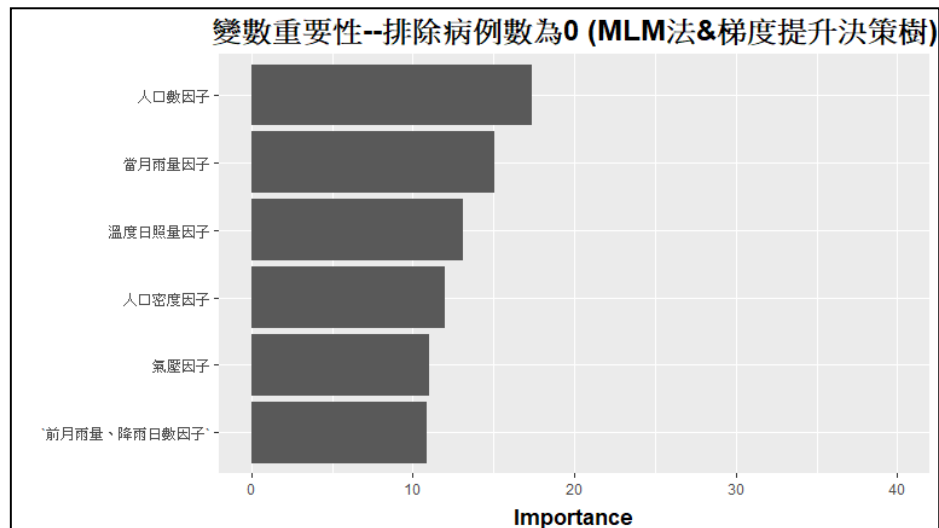
$$0.346 * \text{rainfall\_60minMax\_last\_last} + 0.315 * \text{rainfall\_1DayMax\_last} \\ + 0.309 * \text{rainfall\_10minMax\_last} + 0.259 * \text{rainfall\_last} \\ + 0.145 * \text{wind\_strMax\_last} - 0.118 * \text{tempMax}$$

當月日照量蒸發量因子 =

$$0.323 * \text{global\_radio} + 0.287 * \text{sunshine} + 0.238 * \text{evapA} \\ - 0.186 * \text{cloud\_amount} - 0.181 * \text{tempMin\_last} - 0.168 * \text{temp\_last}$$

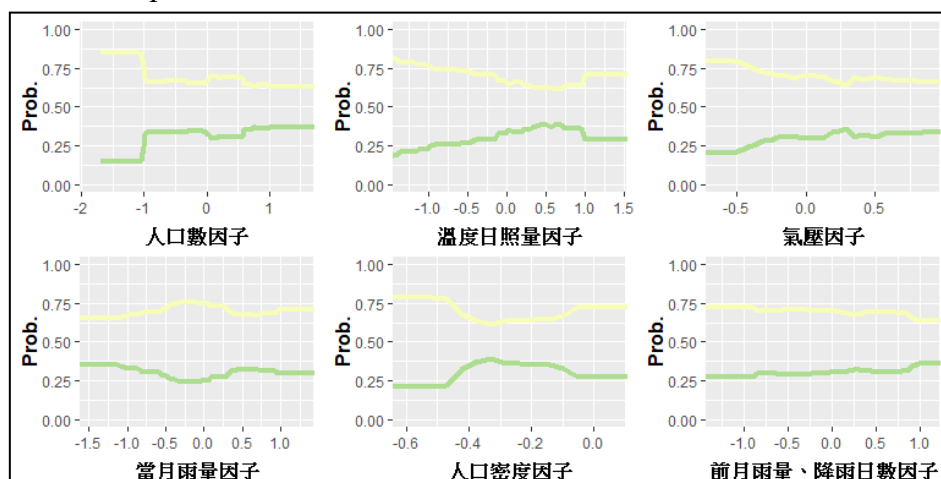
## (2)排除病例數為0的情況（MLM法結果）

### ● 變數重要性圖



上圖為資料排除病例數為0的情況下，MLM法的梯度提升決策樹模型的變數重要性圖。我們可以發現，人口數同樣是重要程度最高的變數，且溫度相關因子、前月雨量因子仍是重要性頗高的變數，此結果與納入病例數為0的PCM法梯度提升決策樹模型結果很相似。

### ● Partial Dependence Plot



上圖為資料排除病例數為 0 的情況下，MLM 法的梯度提升決策樹模型的 Partial Dependence Plot，黃、淺綠二種顏色的線分別表示病例數 1~9 例、10 例以上二個類別。我們可以發現，不論每個解釋變數的邊際變動量為何，分類至病例數為 1~9 例的機率均比分到 10 例以上的機率高，這可能也是由於資料不平衡的因素，我們也猜想或許病例數以 9 例為分界不夠適合。而我們可以大致看出，當人口數因子、氣壓因子、前月雨量與降雨日數因子等三個變數的數值愈大時，分類到 10 例以上的機率/勝算比愈高。而與前面納入病例數為 0 的結果比較，我們可以發現當線條有交會處的情形，該變數重要程度會相當高（可參見納入病例數為 0 時 MLM 法的梯度提升決策樹模型的變數重要性圖）。

下表為上圖所提到的因子寫成變數的線性組合之表示，以下表示式之解釋變數均以還原成標準化前的原始變數，可酌以參考。

人口數因子 =

$$0.534 * \text{population\_last} + 0.534 * \text{population} - 0.008 * \text{stn\_pressure\_last} - 0.069 * \text{population\_den\_last} - 0.069 * \text{population\_den} - 0.045 * \text{temp}$$

溫度日照量因子 =

$$0.747 * \text{temp} + 0.074 * \text{global\_radio\_last} + 0.063 * \text{sunshine\_last} + 0.056 * \text{tempMax} - 0.053 * \text{rainfall\_last} + 0.051 * \text{population\_den\_last}$$

氣壓因子 =

$$0.567 * \text{stn\_pressure\_last} + 0.264 * \text{stn\_pressure} + 0.205 * \text{stn\_pressureMax\_last} - 0.196 * \text{temp} + 0.065 * \text{rainfall\_last} - 0.064 * \text{population\_last}$$

當月雨量因子 =

$$0.332 * \text{rainfall\_Day} + 0.199 * \text{rainfall} - 0.162 * \text{stn\_pressure} + 0.155 * \text{rainfall\_Hour} - 0.112 * \text{sunshine} + 0.105 * \text{cloud\_amount}$$

人口密度因子 =

$$0.539 * \text{population\_den\_last} + 0.539 * \text{population\_den} - 0.133 * \text{temp} - 0.103 * \text{population\_last} - 0.103 * \text{population} + 0.089 * \text{sunshine\_last}$$

前月雨量、降雨日數因子 =

$$0.499 * \text{rainfall\_last} - 0.322 * \text{temp} + 0.24 * \text{rainfall\_1DayMax\_last} - 0.237 * \text{stn\_pressure\_last} + 0.146 * \text{stn\_pressure} + 0.139 * \text{rainfall\_Day\_last}$$

## 6. 模型結果討論

由於我們的目的是探討各分析組合的好壞，以下分別討論三種操縱變因的結果。

### (1) 資料選取

從以上的結果可以明顯發現，當我們將資料納入病例數為 0 的情況時，決策樹與梯度提升決策樹的結果都很難討論，決策樹樹狀圖的分支很細卻沒有將各類別的分群分好，梯度提升決策樹的 Partial Dependence Plot 也呈現不論解釋變數的變化為何，分類至病例數為 0 的機率均為最高的情形。至於模型預測效果（Accuracy, AUC）的結果在不同資料選取情形的比較則不作討論，因為兩種情形的病例數類別數不同，在模型配適上不能混為一談。我們認為，資料排除病例數為 0 的情況作分析較為適合，儘管排除這些資料會降低資料可解釋的意義，然而這些資料產生的偏差可能已遠遠超過可解釋的變異，也許我們進行一些資料不平衡或是異常值的處理後，仍可以考慮納入病例數為 0 的情況作分析。

### (2) 因子分析方法

由於我們最後的視覺化僅選出 MLM / PCM 兩種因子分析方法中較好的結果作呈現及解釋，因此因子分析方法的比較主要在於預測資料的評估指標。從預測效果可以發現，當控制資料選取方法、樹狀模型兩個變因為固定時，MLM 法的效果普遍較好。然而，當我們在命名因子時，比較兩種方法分析後的因子負載量、因子寫成變數線性組合之係數後發現，PCM 法通常能使同類型變數（例如：降雨、氣壓、日照、人口）的負載量更大，不同類型變數則更小，而因子寫成變數線性組合的係數也是同類型變數的係數較大，不同類型較小，可以推測 PCM 法能使因子內的變數共同特性更加顯著，但因子內也犧牲了一些能使預測效果更好的變數訊息。因此我們認為，兩種因子分析方法都有各自的優點，MLM 法能增加預測準確度，PCM 法能增加因子的可解釋程度。

### (3) 樹狀模型

由預測資料的評估指標我們可以發現，當資料納入病例數為 0 的情況時，決策樹與梯度提升決策樹的結果沒有顯著差別，然而當資料排除病例數為 0 的情況時，梯度提升決策樹的兩項評估指標皆有所提升。

在決策樹的視覺化呈現中，我們發現即便我們作了模型剪枝，但資料納入病例數為 0 的情況並沒有把各個類別清楚分開，且枝節仍繁茂，產生了解釋上的困難。排除病例數為 0 的情況則較能將兩種病例數類別的情況分開，我們也能看出縣市、人口密度因子是較能使模型分支的變數。

在梯度提升決策樹的視覺化呈現中，我們發現到雖然有些解釋變數與傾向分類到病例數較多的組別的機率有共同的趨勢，然而由於資料不平衡的因素，使傾向分類到病例數較少的組別的機率總是最高，資料不平衡是我們使用這些樹狀模型的方法上必須再特別注意的問題。

#### (4)總結

觀察到各模型的重要變數，我們可以發現不論使用何種模型進行分析，縣市、月份、人口相關（人口數、人口密度）等變數均為最重要的變數之一，而且由決策樹樹狀圖可知道這幾個變數對於分支很重要，不是屢次出現在圖中節點的條件就是分群後產生的子節點有很大的變化。

而我們也能發現溫度、雨量都是影響病例數分群的重要變數，尤其在雨量的部分，因子分析時通常可分成當月份的降雨因子與前月份降雨因子，而且前月份的降雨因子的重要程度比想像中的高，且大致與病例數成正比，這點驗證了文獻所提出前月份的降雨會影響到下個月的登革熱病例數之論述。

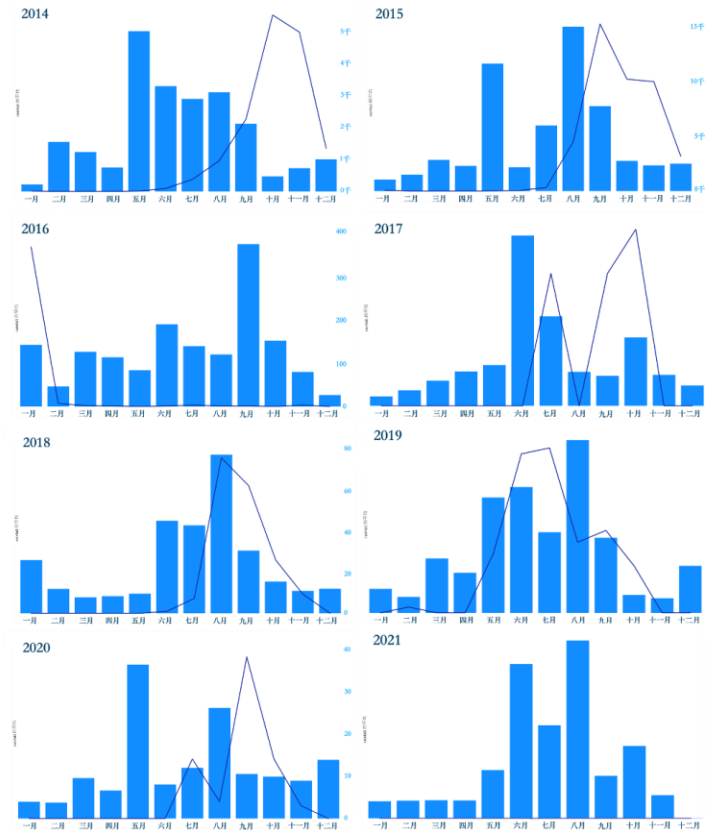
統整決策樹與梯度提升決策樹模型的結果，我們可大致推論，就季節方面，秋天發現登革熱的情況最多，冬春之際最少；就縣市方面，南部縣市與中北部都會區發生登革熱的情況較多。而前月雨量愈多、當月或前月日照量愈低、人口數愈多，愈傾向發生登革熱，至於溫度、氣壓等因子雖然也是重要變數，然而沒有觀察到它們對於登革熱發生病例數有無明顯趨勢。

#### 六、參考文獻

1. NCSS Statistical Software--NCSS Statistical Software  
[https://ncss-wpengine.netdna-ssl.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/Zero-Inflated\\_Negative\\_Binomial\\_Regression.pdf](https://ncss-wpengine.netdna-ssl.com/wpcontent/themes/ncss/pdf/Procedures/NCSS/Zero-Inflated_Negative_Binomial_Regression.pdf)
2. Github-- GLM with zero-inflated data  
[https://fukamilab.github.io/BIO202/04-C-zero-data.html#zero-inflated\\_poisson\\_glm](https://fukamilab.github.io/BIO202/04-C-zero-data.html#zero-inflated_poisson_glm)
3. Github--Parameter tuning caret  
<https://csantill.github.io/RTuningModelParameters/>
4. Rpubs--caret demonstration with kaggle bikeshare data (I)  
<https://rpubs.com/chengjiun/52658>
5. Github-- Handling Class Imbalance with R and Caret - An Introduction  
<http://dpmartin42.github.io/posts/r/imbalanced-classes-part-1>
6. 台部落--R 語言主成分和因子分析篇  
<https://www.twblogs.net/a/5b81fdc42b71772165af1435>
7. 登革熱流行病學－登革熱在台灣之流行  
<https://www.airitilibrary.com/Publication/alDetailedMesh?docid=02575655-198901-5-1-1-11-a>

## 七、附錄

### ● 附錄一：2014 至 2021 年度平均月雨量與病例總數趨勢圖表



### ● 附錄二：四種迴歸模型的比較

分配	參數	$P(Y=0)$	$E(Y)$	$Var(Y)$
卜瓦松分配 (Poisson Distribution)	$\mu$	$e^{-\mu}$	$\mu$	$\mu$
負二項式分配 (Negative Binomial Distribution)	$\mu, \alpha$	$(1 + \alpha\mu)^{-1/\alpha}$	$\mu$	$\mu(1 + \alpha\mu)$
零膨脹卜瓦松分配 (Zero-Inflated Poisson Distribution)	$\mu, \pi$	$\pi + (1 - \pi)e^{-\mu}$	$(1 - \pi)\mu$	$(1 - \pi)\mu$
零膨脹負二項式分配 (Zero-Inflated Negative Binomial Distribution)	$\mu, \alpha, \pi$	$\pi + (1 - \pi)(1 + \alpha\mu)^{-1/\alpha}$	$(1 - \pi)\mu$	$(1 - \pi)\mu(1 + \mu(\pi + \alpha))$

上表為我們所使用的四種模型的分配簡易比較，大致可以歸納為下列幾點：

(一) 零膨脹的分配相較於傳統計數型分配多了一個參數 $\pi$ ，此參數的意義可以理解成，有 $\pi$ 的機率觀察值必為零，而剩下的 $1 - \pi$ 則分給傳統計數的卜瓦松或負二項式分配。

(二) 負二項式分配相較於卜瓦松分配多了一個參數 $\alpha$ ，來表示期望值和變異數不同的地方。

由於我們資料中 0 的個數佔了總樣本超過 80%，因此，我們的模型多半採用零膨脹的分配，多了一個參數 $\pi$ ，來表示實際不發生率。零膨脹的模型式同時估計兩個估計式，一個是計數的部分，使用卜瓦松迴歸或負二項式迴歸，另一個是估計不發生率 $\pi$ 的部分，採用羅吉斯迴歸，模型如下：

$$\begin{aligned} \text{count: } \log(\mu) &= \alpha + \beta X \\ \text{zero: } \pi &= \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}} \end{aligned}$$

### ● 附錄三：傳統計數型模型配適過程

#### (一) 卜瓦松迴歸 (Poisson Regression)

由於資料中，「零」的個數過多，卜瓦松迴歸無法配適，因此我們將案例數為零的資料去除後配適，結果如下。

模型：(Model 1)

$\log(E(\text{cases} | X))$

$$= \beta_0 + \beta_{\text{region}} I(\text{region}) + \beta_{\text{season}} I(\text{season}) + \beta_{\text{temp}_{\text{last}}} x_{\text{temp}_{\text{last}}} + \beta_{\text{rainfall}_{\text{last}}} x_{\text{rainfall}_{\text{last}}}$$

Model 1	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	-9.4044	0.1053	-89.2881	< 0.0001
$\beta$ region 北	-0.3737	0.0781	-4.7852	< 0.0001
$\beta$ region 東	-0.1340	0.1634	-0.8203	0.4121
$\beta$ region 南	4.7499	0.0627	75.7486	< 0.0001
$\beta$ season 春季	-4.3369	0.1310	-33.0945	< 0.0001
$\beta$ season 秋季	-0.1789	0.0211	-8.4779	< 0.0001
$\beta$ season 夏季	-2.8105	0.0266	-105.467	< 0.0001
$\beta$ temp_last	0.4628	0.0037	126.5948	< 0.0001
$\beta$ rainfall_last	-0.0010	< 0.0001	-63.6723	< 0.0001

#### (二) 負二項式迴歸 (Negative Binomial Regression)

同樣地，我們將案例數為 0 的資料去除後配適負二項式迴歸，結果如下。  
模型：(Model 2)

$\log(E(\text{cases} | X))$

$$= \beta_0 + \beta_{\text{region}}I(\text{region}) + \beta_{\text{season}}I(\text{season}) + \beta_{\text{temp}_{\text{last}}}x_{\text{temp}_{\text{last}}} + \beta_{\text{rainfall}_{\text{last}}}x_{\text{rainfall}_{\text{last}}}$$

Model 2	係數估計	標準誤	Z	p 值
$\beta$ (Intercept)	-3.5242	1.0572	-3.3337	0.0009
$\beta$ region 北	-0.5838	0.3223	-1.8111	0.0701
$\beta$ region 東	-0.5320	0.5090	-1.0452	0.2959
$\beta$ region 南	3.8327	0.3345	11.4572	< 0.0001
$\beta$ season 春季	-3.8084	0.6109	-6.2339	< 0.0001
$\beta$ season 秋季	0.0904	0.4225	0.2140	0.8306
$\beta$ season 夏季	-1.1783	0.4942	-2.3845	0.0171
$\beta$ temp_last	0.2330	0.0498	4.6755	< 0.0001
$\beta$ rainfall_last	0.0004	0.0005	0.8078	0.4192

而負二項式迴歸所估計變異數和平均數之差的部分的 95%信賴區間為 [0.036, 0.50]，不包含 0，變異數和平均數不相同，因此，這筆資料應以負二項式迴歸配適較佳。

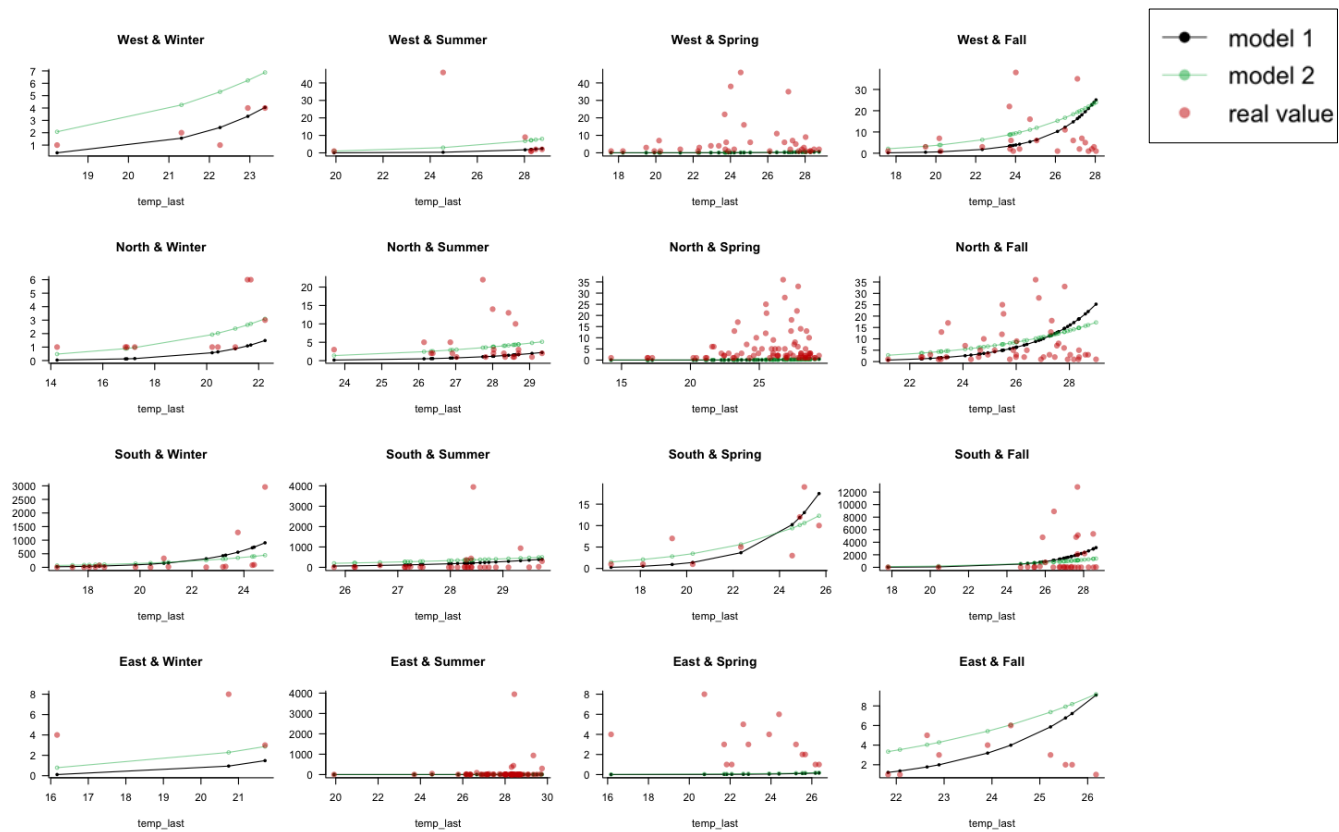
Theta	標準誤
0.4294	0.0348

### (三) 卜瓦松迴歸及負二項式迴歸之比較

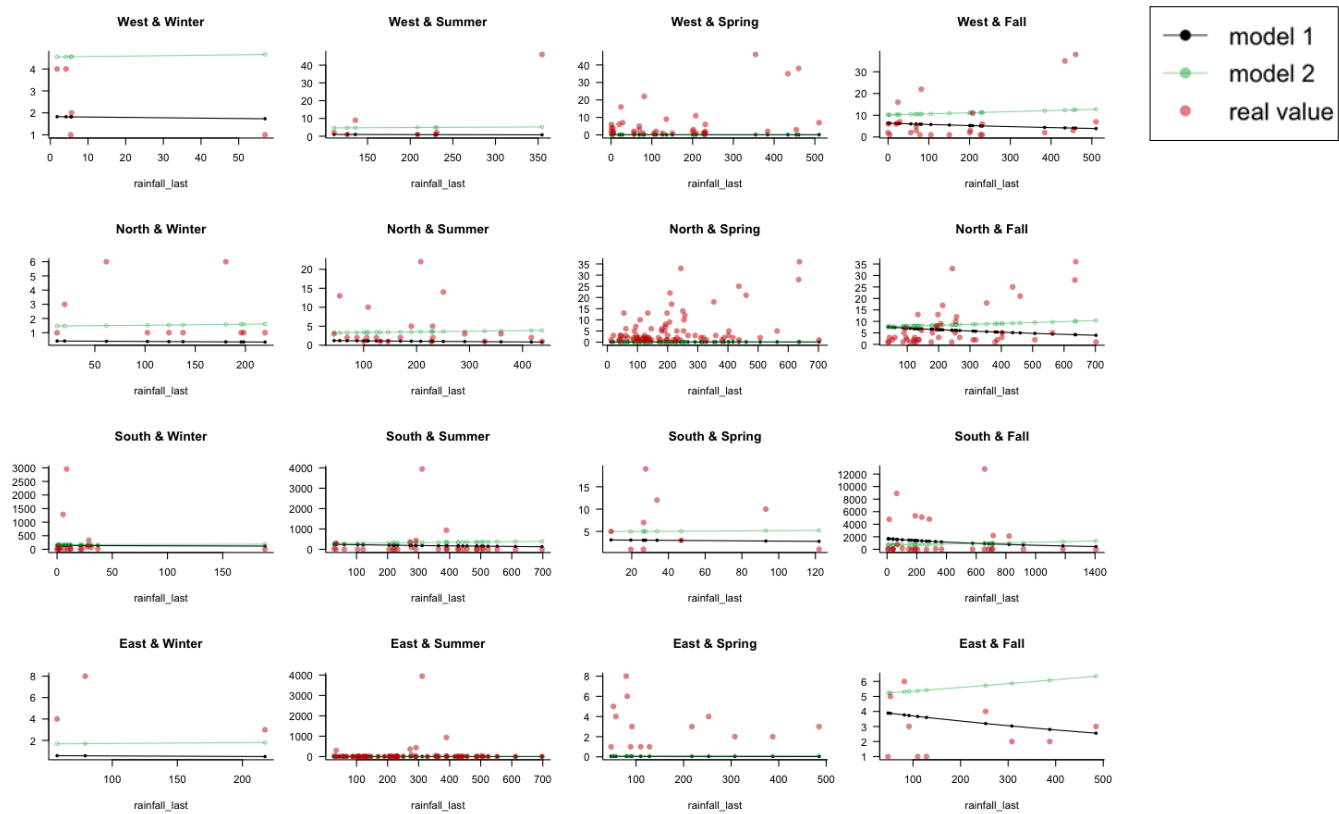
比較卜瓦松迴歸以及負二項式迴歸的係數正負號方向，可以發現

1. 卜瓦松迴歸秋季的係數為負，而負二項式秋季的係數為正，表示卜瓦松所估計的模型認為秋季比冬季登革熱案例數少，而負二項式的模型認為秋季比冬季登革熱案例數多。
2. 卜瓦松迴歸之中，前月雨量的係數為負，而負二項式前月雨量的的係數為正，表示卜瓦松所估計的模型認為雨量多則登革熱案例數少，而負二項式的模型認為雨量多則登革熱案例數多。
3. 將模型的預測值與實際值做比較，如下兩圖所示。第一張圖為臺灣各地區在各季節下，前一個月的氣溫與案例數之間的關係，第二張圖為臺灣各地區在各季節下，前一個月的雨量與案例樹之間的關係。其中，紅點為實際值，黑點與綠點分別為 Model 1(卜瓦松迴歸)以及 Model 2(負二項式迴歸)的預測值。





(橫軸：上個月氣溫，縱軸：案例數)



(橫軸：上個月雨量，縱軸：案例數)

#### ● 附錄四：方法學介紹--因子分析

因子分析是以少數幾個因子來解釋一群相互之間有關係存在的變數，每個變數除了受共同因素(Common Factor)的影響外，尚有獨特因素 (Specific Factor)，簡言之就是以數個因子來解釋一大群變數。

**因子分析應用：**

(一) 找出潛藏因子：從一大堆變數找出少數幾個共同因子，用這幾個因子來將變數簡化。

(二) 篩選變數：且因子分析還有篩選變數的作用，透過因子分析能找出幾群內部相關性高的變數族群，在每一個族群中挑選一兩個變數當做該族群的代表 (或是就以因子本身當代表)，可避免變數間的共線性問題，便於我們使用回歸分析。

(三) 對資料作摘要：以少數幾個因素解釋大部分變數的變異

我們在做因子分析時，最常會碰到，因素負荷量 (factor loadings) 與特徵值 (eigenvalues) 這兩個詞，這邊就來解釋一下這兩個詞的意義。

- 特徵值：在作因素分析時，每一個因素都會得到一個 eigenvalues，而這個值表示在所有的變數裡面，這個因素可以解釋多少的 variance，所以如果你有十個變數，所有因素 eigenvalues 加總應該要等於十。換言之，如果特徵值太小，表示此 factor 只能解釋非常少部分的變數，所以這對減少變數數量並沒有什麼幫助，如果特徵值小於 1，表示此因素解釋少於 1 個變數，當然也就不適合放在納入考量了。
- 因素負荷量 (factor loadings)：此詞簡單地說就是個別變數與因素之間的相關性 (沒轉軸前)，數值介於-1 至 1 之間，這些變數在這個因素裡面的 weight 有多少，或是這個變數多接近這個因素。最後最重要的是我們要決定，哪些變數可歸入某個因素？

#### ● 附錄五：方法學介紹--梯度提升決策樹

梯度提升決策樹 (gradient boost decision tree, GBDT) 是一種 boosting 的方法，每一次的訓練集不變，訓練集之間的選擇不是獨立的，每一是選擇的訓練集都是依賴上一次學習的結果，根據錯誤率 (給予訓練樣本不同的權重) 取樣，由於集成諸多淺且弱連續的樹模型，每個樹模型會以之前的樹模型為基礎去學習和精進，結果通常是難以擊敗的。GBDT 意指每棵樹皆為 Decision Tree，樹也可以是其他的樹模型，此時可稱模型為梯度提升機 (Gradient Boost Machine)，然而通常樹模型為 Decision Tree。

Decision tree 通常為一棵複雜的樹，而在 Boosting 則是產生非常多棵的樹，但是每一棵的樹都很簡單的決策樹，Boosting 希望新的樹可以針對舊的樹預測不太好的部分做一些補強。最終我們要把所有簡單的樹合在一起才能當最後的預測輸出。

- Gradient Boosting 除了 Boosting 一般擁有的性質外，還具備一些優點，
- Gradient Boosting 可以應用在許多不同的（可微分）Loss Function 上
  - 利用不同的 Loss Function，我們可以處理 Regression / Classification / Ranking 等不同的問題
  - 通常提供無與倫比的預測準確性。
  - 很大的靈活性，可以優化不同的損失函數，並提供幾個超參數調整選項，使函數非常靈活。
  - 無需數據預處理 - 通常適用於分類和數值。
  - 處理缺失的數據 - 不需要插補。

## ● 附錄七：方法學介紹-- k-fold cv, stratified k-fold cv 介紹

### （一）k-fold cv (k 折交叉驗證)：

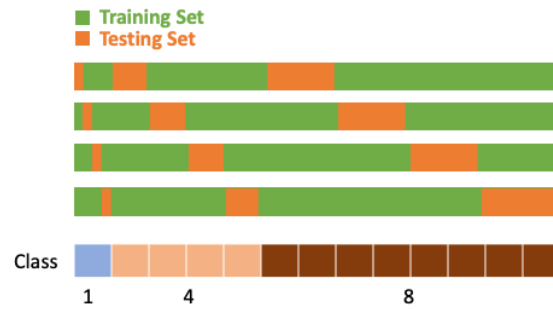
在 K-Fold 的方法中我們會將資料切分為 K 等份，K 是由我們自由調控的，以下圖為例：假設我們設定 K=10，也就是將 Training data 切割為十等份，這意味著相同的模型要訓練十次，每一次的訓練都會從這十等份挑選其中九等份作為訓練資料，剩下一等份未參與訓練並作為驗證集。

因此訓練十次會有十個不同驗證集的 Error，這個 Error 通常稱作 loss 也就是模型評估方式，最後再把這十次的 loss 加總起來取平均就可以當成最終結果。透過這種方式，不同分組訓練的結果進行平均來減少方差，因此我們模型的性能對數據的劃分就不會那麼敏感。



### （二）stratified k-fold cv (分層交叉驗證):

每個 Fold 都是按照類別的比例抽出來的，假設這個分類任務一共有三個類別 A、B、C，它們的比例是 1:4:8。那麼每個 fold 中的 A、B、C 的比例也必須是 1:4:8。其實實現方式也非常簡單，首先依序把 A、B、C 類別的數據隨機分成 k 組，最後再把它們合併依照比例起來，就得到了 k 組滿足 1:2:10 的數據了。

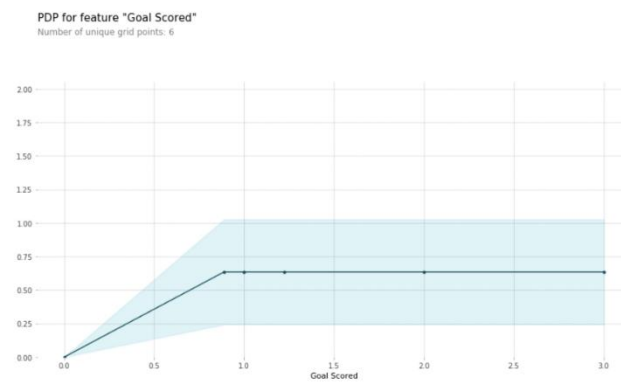


其優點包含：

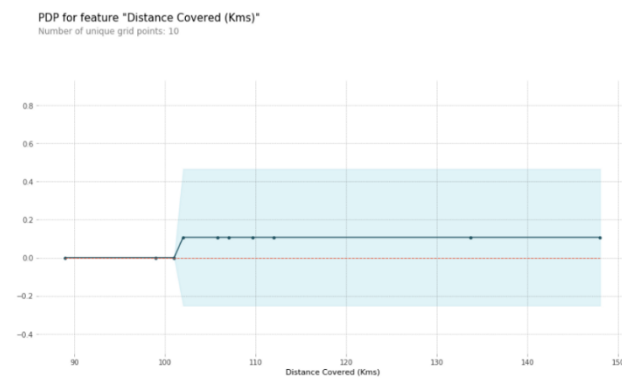
1. 優於一般的 k-fold，因為測試集能充分代表整體數據。
2. 預測結果的方差也會變小，使得交叉驗證的 error 更為可靠。
3. 對於資料不平衡的數據很有用。

### ● 附錄八：方法學介紹--邊際變數對 y 效應圖

邊際變數對 y 效應圖（Partial Dependence Plot），可以顯示每一個自變數的變化是如何影響預測表現，以下舉例說明。



y 軸代表變數 Goal Scored 的變化所導致的預測值變化，藍色是信賴區間，因此上圖說明隨著 Goal Scored 的增加也明顯提高預測的機率，但是超過 1.0 之後的就沒有明顯的影響了。

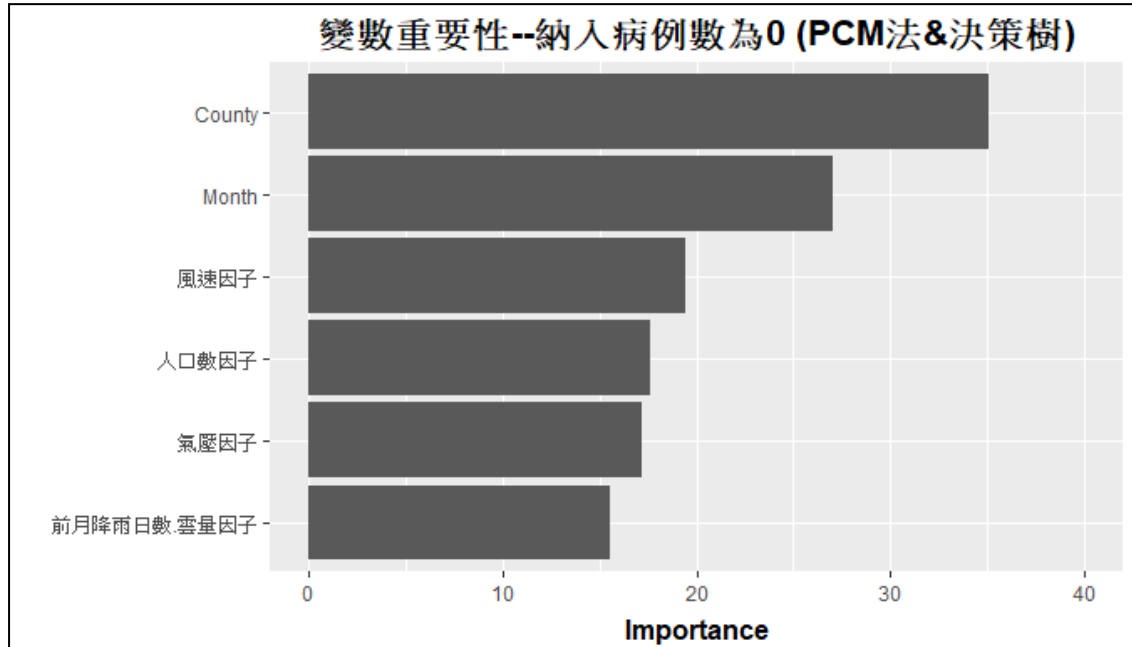


再看到上圖，這張圖可以看到，100 多之後會有最大預測值。但值不高，約 0.1，跟 Goal Scored 比較是相對低影響的。

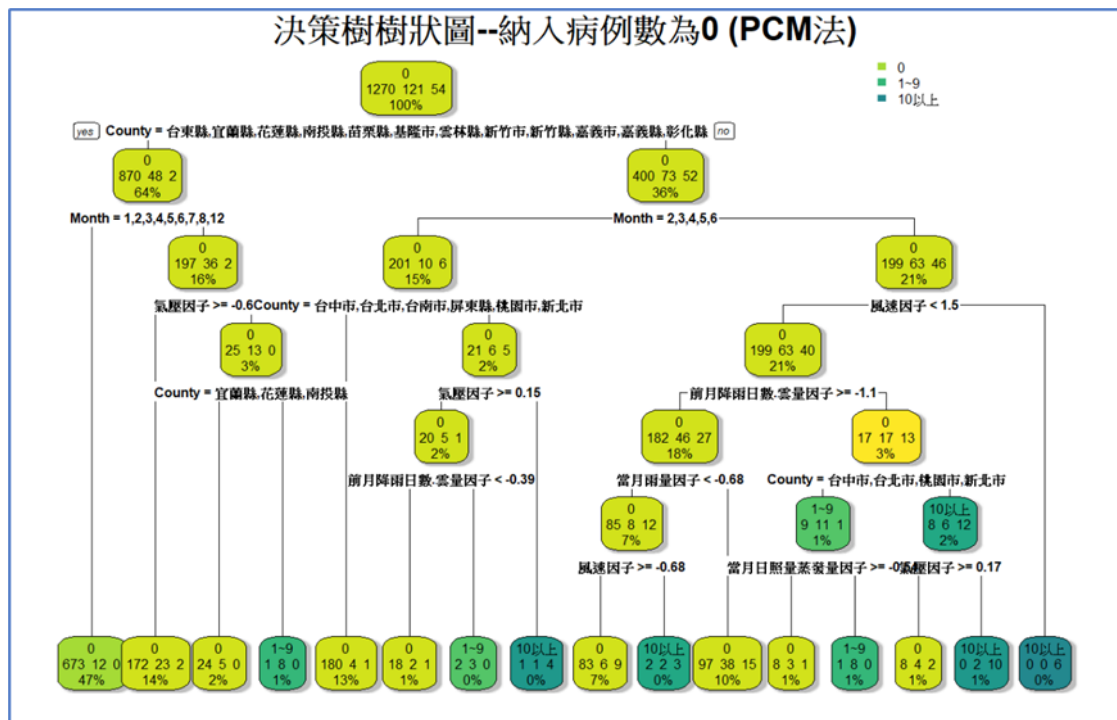
● 附錄九：決策樹模型的其餘 2 種結果（僅以圖示）

（一）資料納入病例數為 0--PCM 法

1. 變數重要性圖

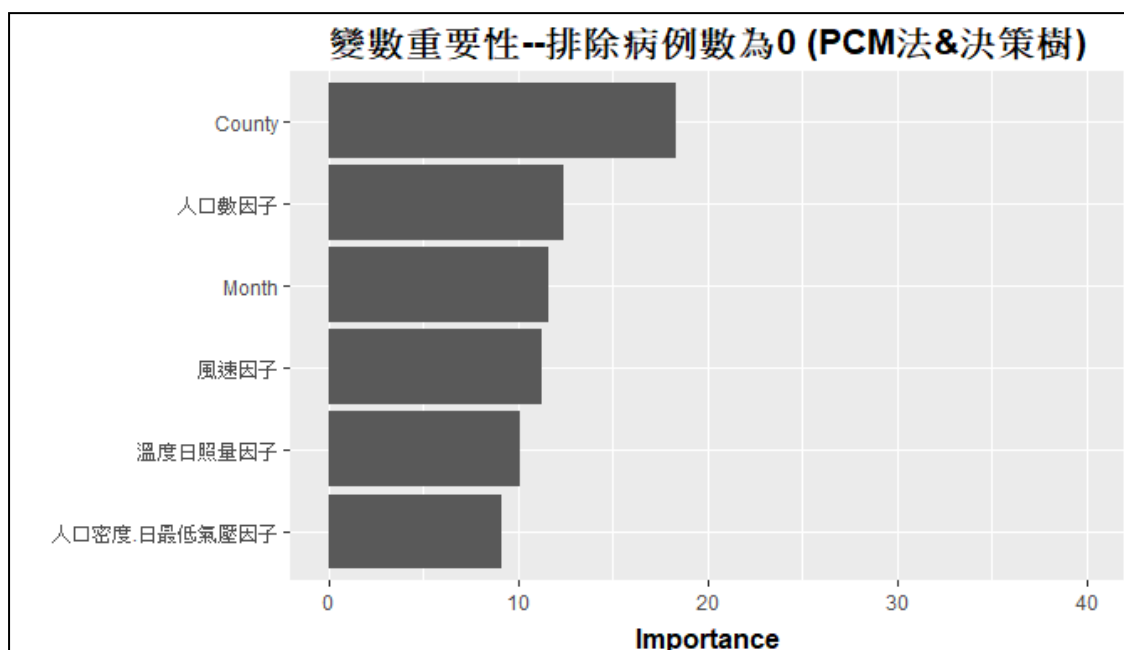


2. 樹狀圖

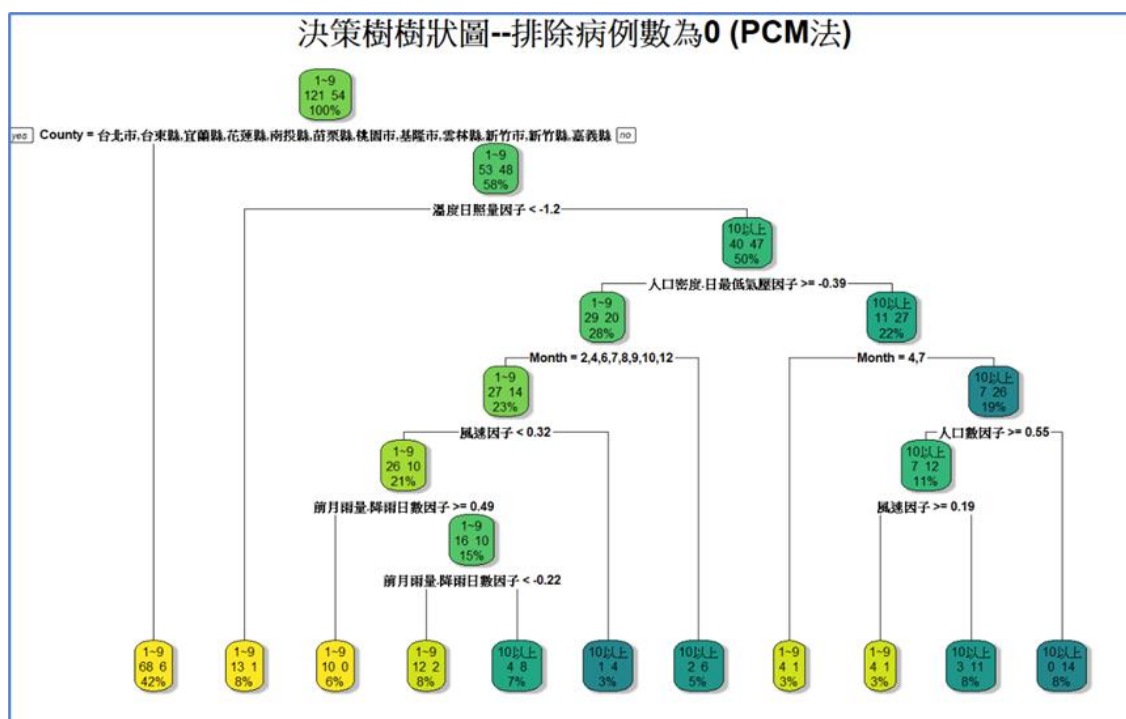


## (二) 資料排除病例數為 0--PCM 法

### 1. 變數重要性圖



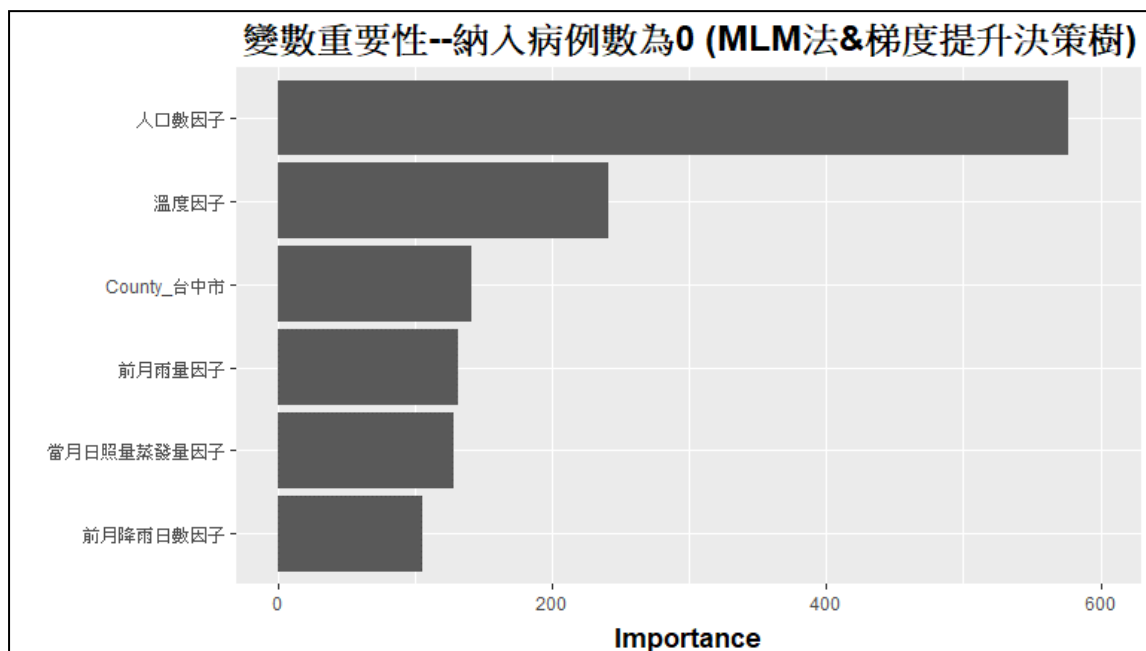
### 2. 樹狀圖



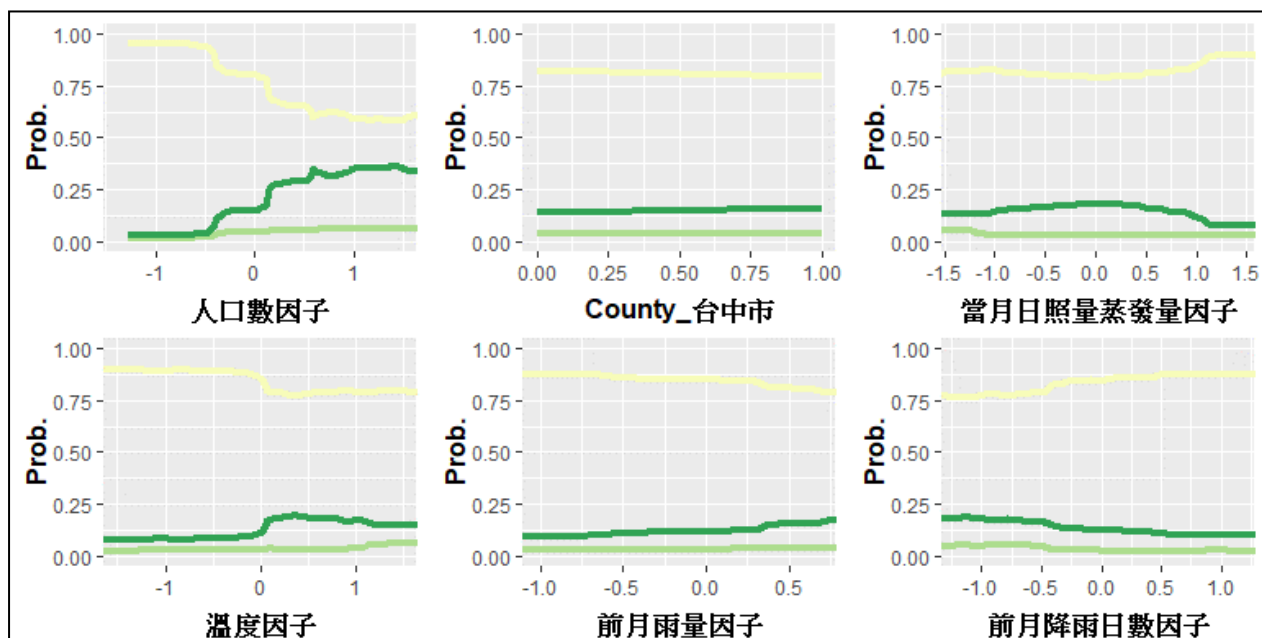
● 附錄十：梯度提升決策樹的其餘 2 種結果（僅以圖示）

（一）資料納入病例數為 0--MLM 法

1. 變數重要性圖

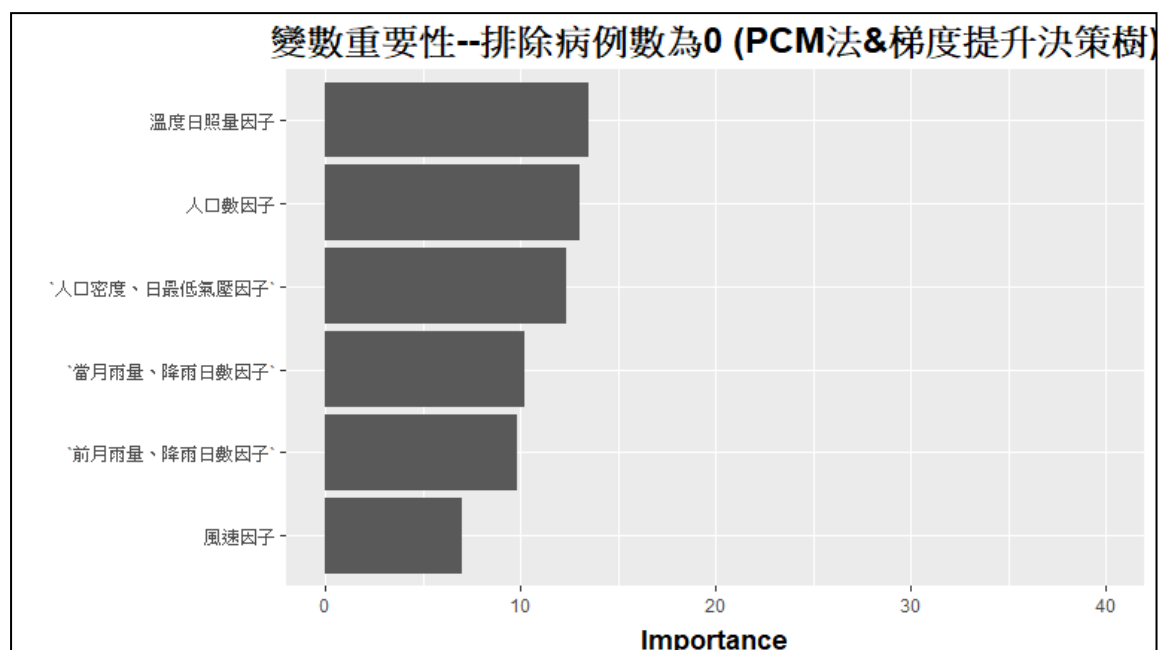


2. Partial Dependence Plot



## (二) 資料排除病例數為 0--PCM 法

### 1. 變數重要性圖



### 2. Partial Dependence Plot

