台灣各縣市登革熱

好發與否研究分析

組員：許祐誠、[redacted]

目錄
Content

研究動機

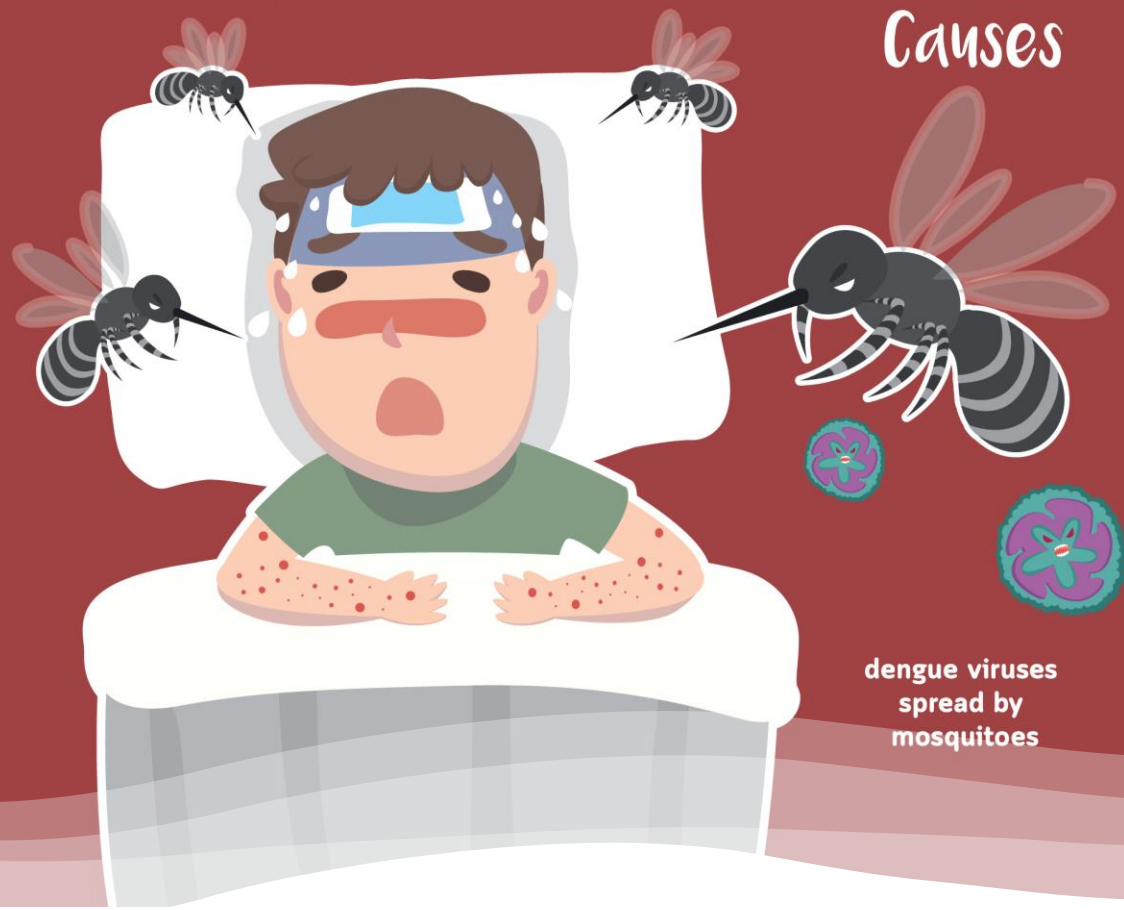
研究動機

有鑑於就讀成大，生活在登革熱高好發區的台南，對於登革熱疫情的病例數的多寡是我們所好奇且關心的，因此我們想要探討有那些天氣因素可能會影響病例數目的多寡，例如溫度高低以及降雨多寡是否對於登革熱的引發有所關聯、或是北中南各縣市罹患登革熱的人數比例等等。

登革熱簡介

- 登革熱（Dengue fever），是一種由**登革病毒**所引起的急性傳染病，這種病毒會**經由蚊子傳播**給人類。登革病毒，可能引起宿主不同程度的症狀，比方說如**發燒**、**出疹**的典型登革熱，或出現嗜睡、躁動不安、肝臟腫大等警示徵兆，甚至可能導致嚴重出血或嚴重器官損傷的**登革熱重症**。
- 登革熱的好發地區，主要集中在**熱帶、亞熱帶**等有**埃及斑蚊和白線斑蚊**分布的國家，隨著全球化發展，各國之間相互流通越發頻繁，登革熱也開始向各國蔓延，成為嚴重的公共衛生問題。
- 臺灣位於**亞熱帶**地區，像這樣**有點熱、又有點溼的環境**，正是蚊子最喜歡的生長環境，為**登革熱流行高風險地區**。

Dengue Fever



Causes

dengue viruses spread by mosquitoes

Symptoms

Headache

High fever

Muscle, bone and joint pain

Pain behind the eyes

Rash

Fatigue

Bleeding from nose

Nausea and vomiting

敘述統計

變數介紹

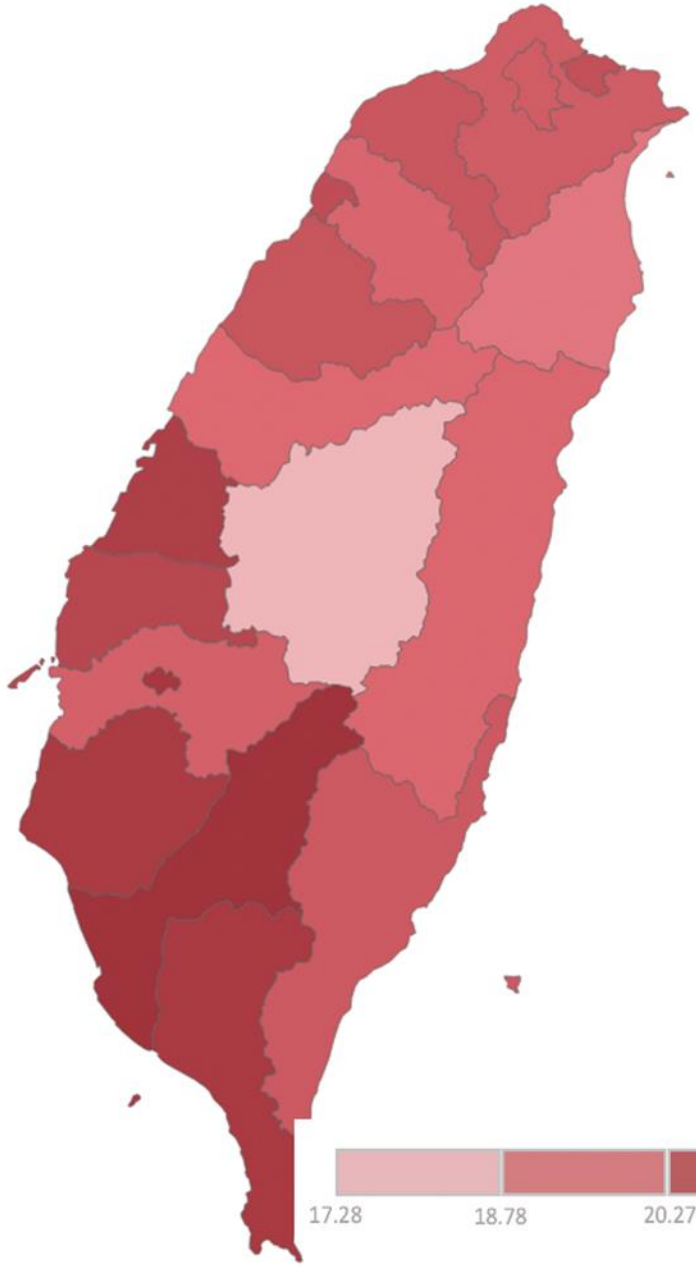
變數類型		變數名稱	意義
類別型		Year	發病年分
		Month	發病月份
		County	發病縣市
		Cases	案例數
		cloud_amount	雲量
連續型	天氣相關	temp, temp_Max, temp_Min, tddewpoint	氣溫相關變數
		rainfall, rainfall_Hour, rainfall_Day, rainfall_10minMax, rainfall_60minMax, rainfall_1DayMax	降雨相關變數
		stn_pressure, stn_pressureMax, stn_pressure_Min	氣壓相關變數
		windspeed, wind_strMax	風速變數
		sunshine, global_radio	日照相關變數
		rh	相對溼度
		evapA	蒸發量
	人口相關	population, population_den	人口數、人口密度

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

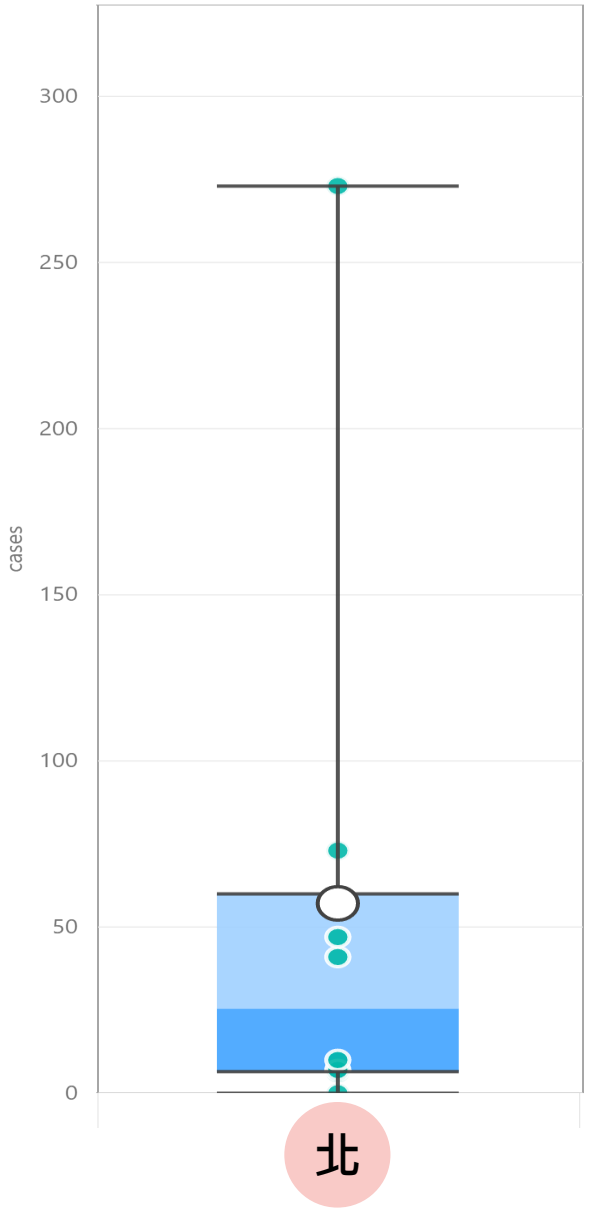
變數介紹

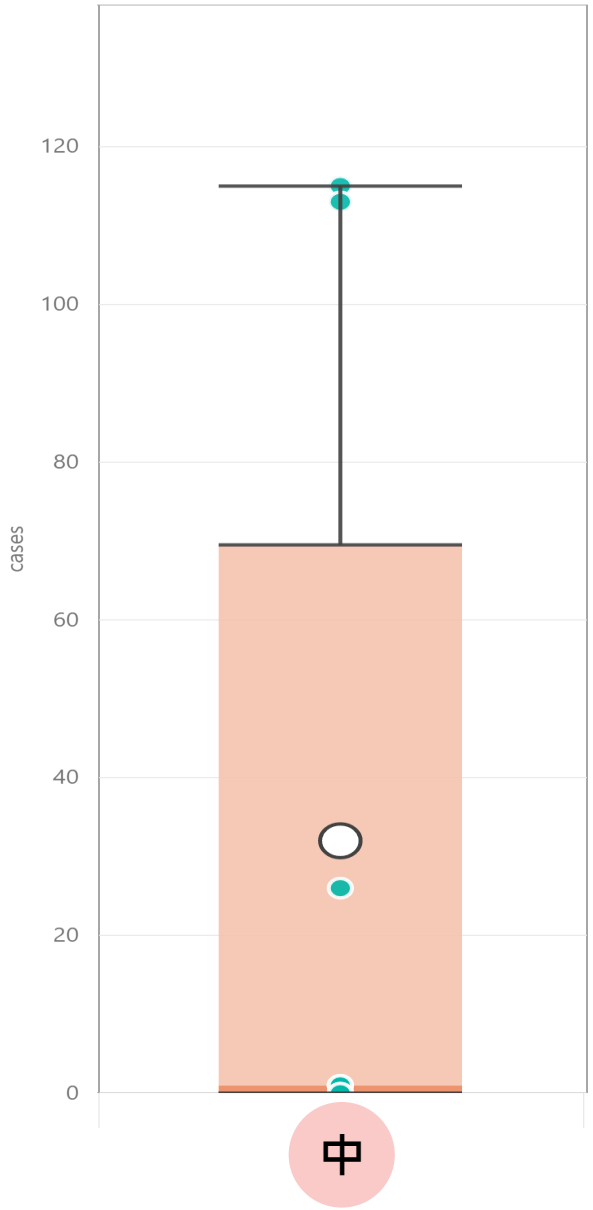
各縣市平均氣溫

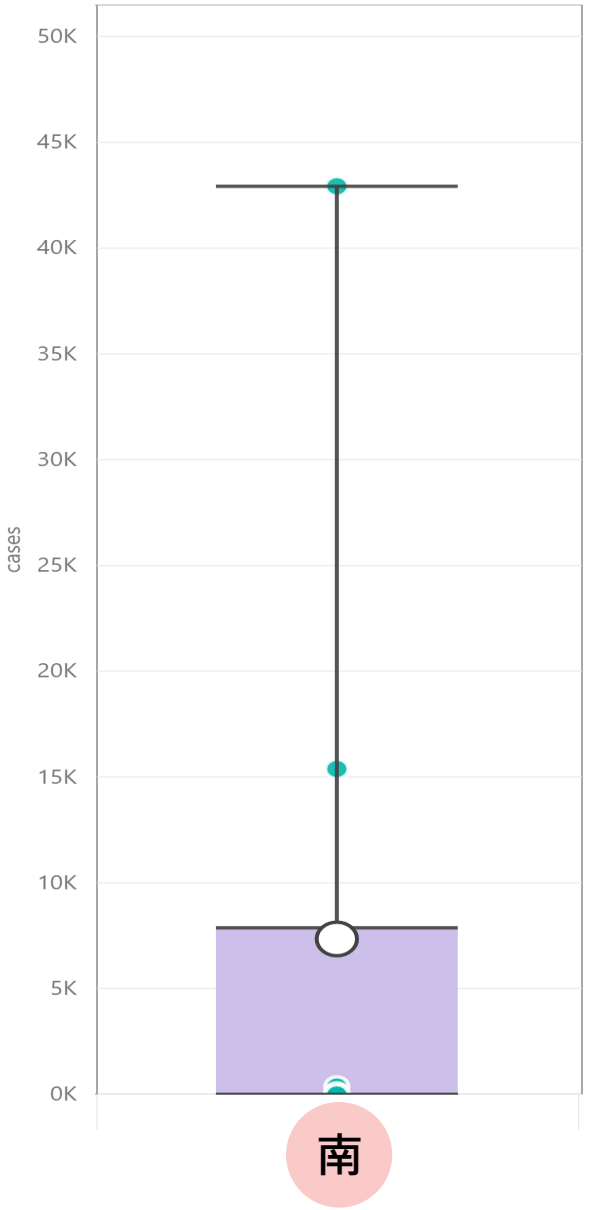


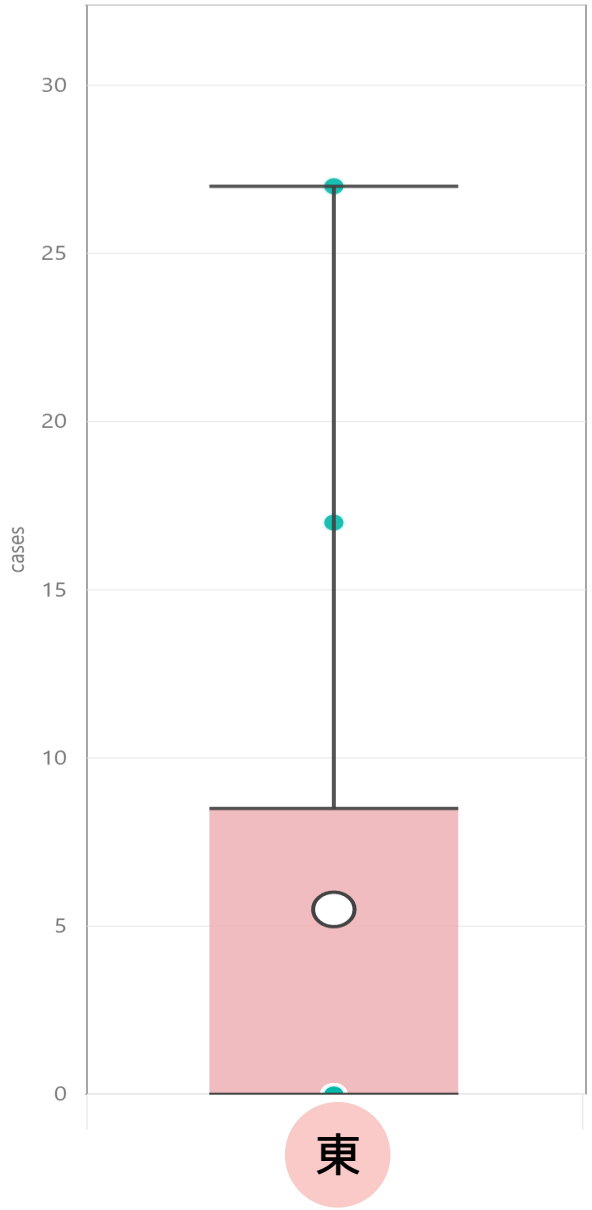
County	temp
高雄市	24.77
嘉義市	24.41
屏東縣	24.26
台南市	24.22
彰化縣	23.95
雲林縣	23.37
新竹市	22.93
基隆市	22.74
苗栗縣	22.43
桃園市	22.33
台東縣	22.10
台北市	21.89
新北市	21.85
嘉義縣	21.62
新竹縣	21.30
花蓮縣	21.23
台中市	21.08
宜蘭縣	20.39
南投縣	17.28
總計	**22.32**

臺灣四大地區分區病例總數

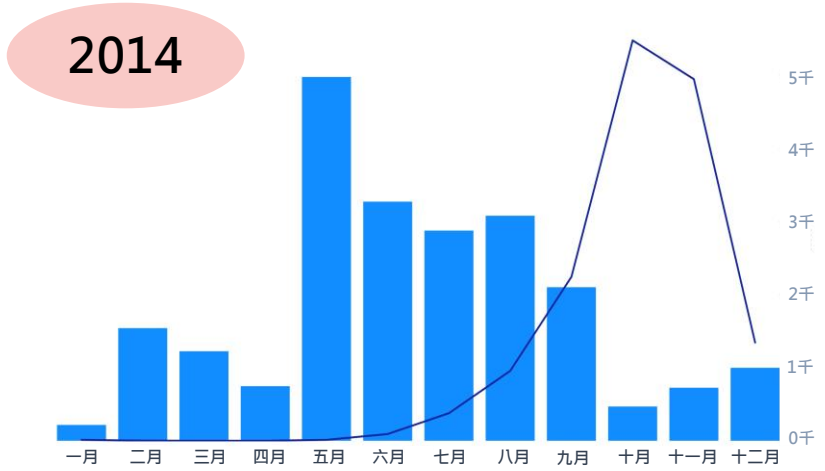


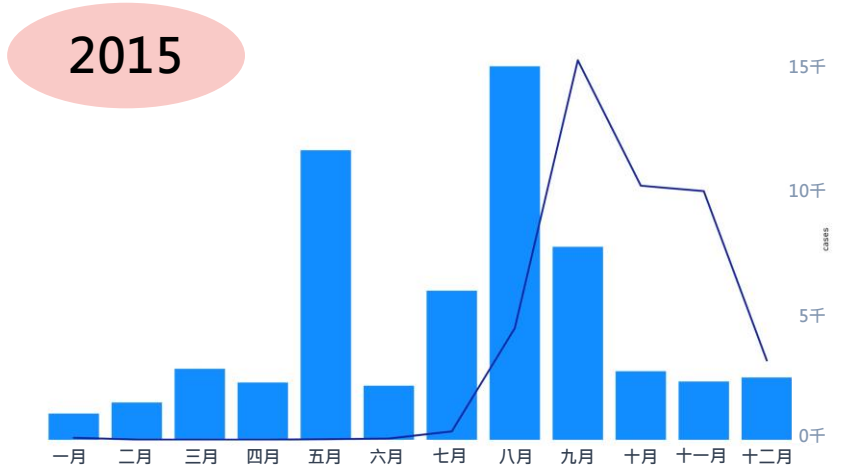


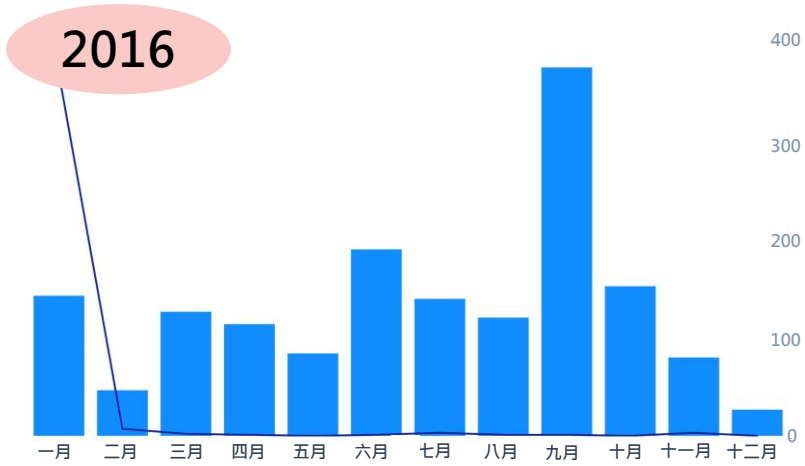


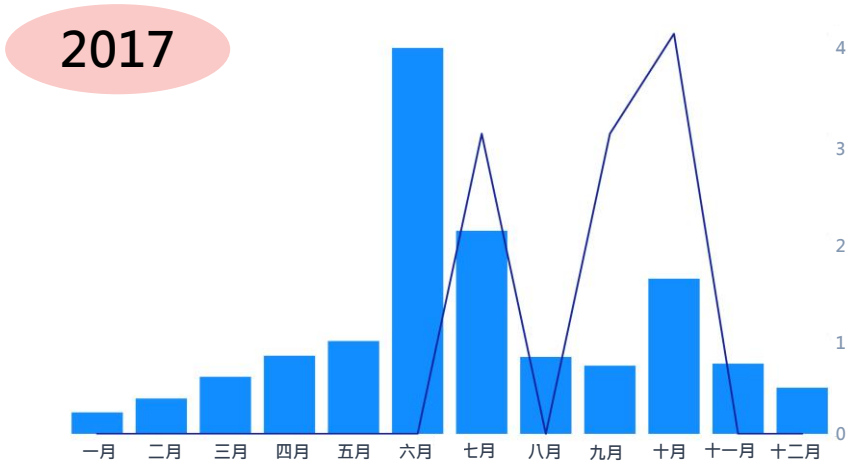


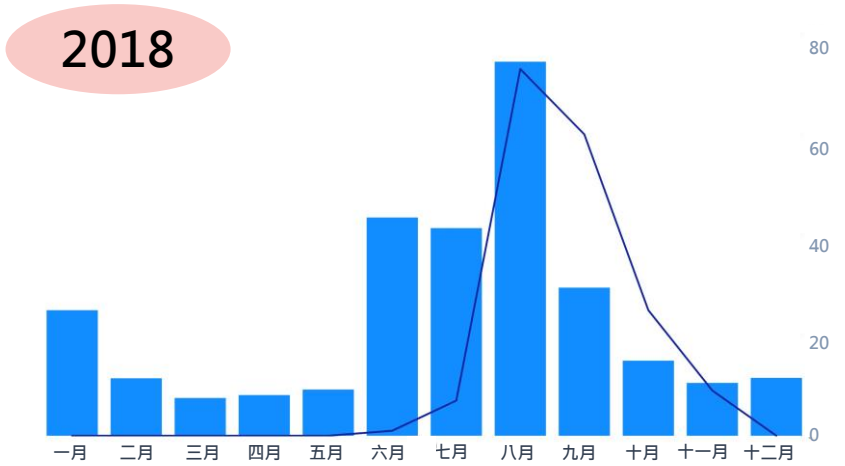
各年度雨量＆病例數

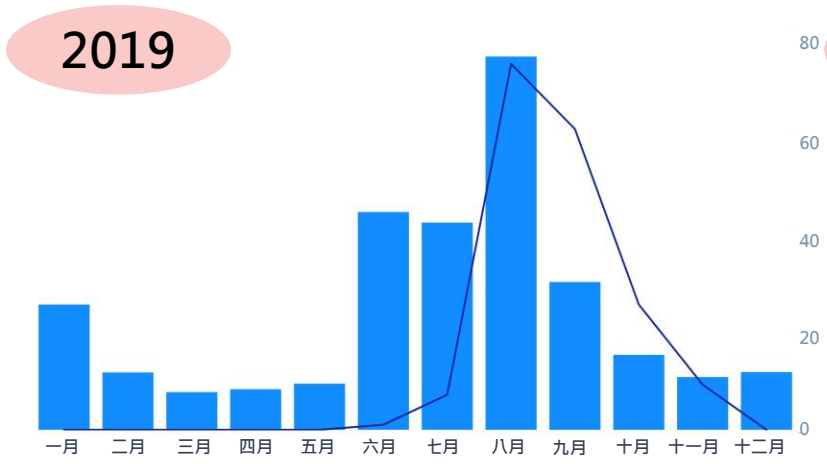

2014
5千
4千
3千
2千
1千
0千
一月 二月 三月 四月 五月 六月 七月 八月 九月 十月 十一月 十二月

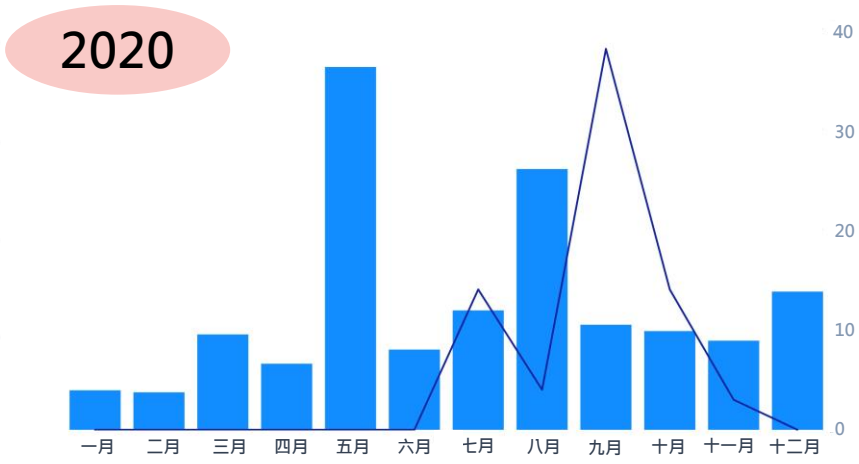

2015
15千
10千
5千
0千
cases
一月 二月 三月 四月 五月 六月 七月 八月 九月 十月 十一月 十二月

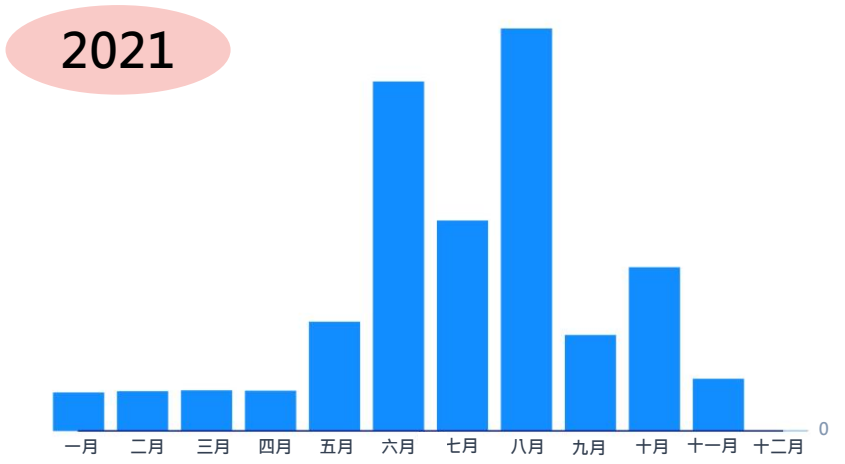

2016
400
300
200
100
0
一月 二月 三月 四月 五月 六月 七月 八月 九月 十月 十一月 十二月


2017
4
3
2
1
0
一月 二月 三月 四月 五月 六月 七月 八月 九月 十月 十一月 十二月


2018
80
60
40
20
0
一月 二月 三月 四月 五月 六月 七月 八月 九月 十月 十一月 十二月


2019
80
60
40
20
0
一月 二月 三月 四月 五月 六月 七月 八月 九月 十月 十一月 十二月

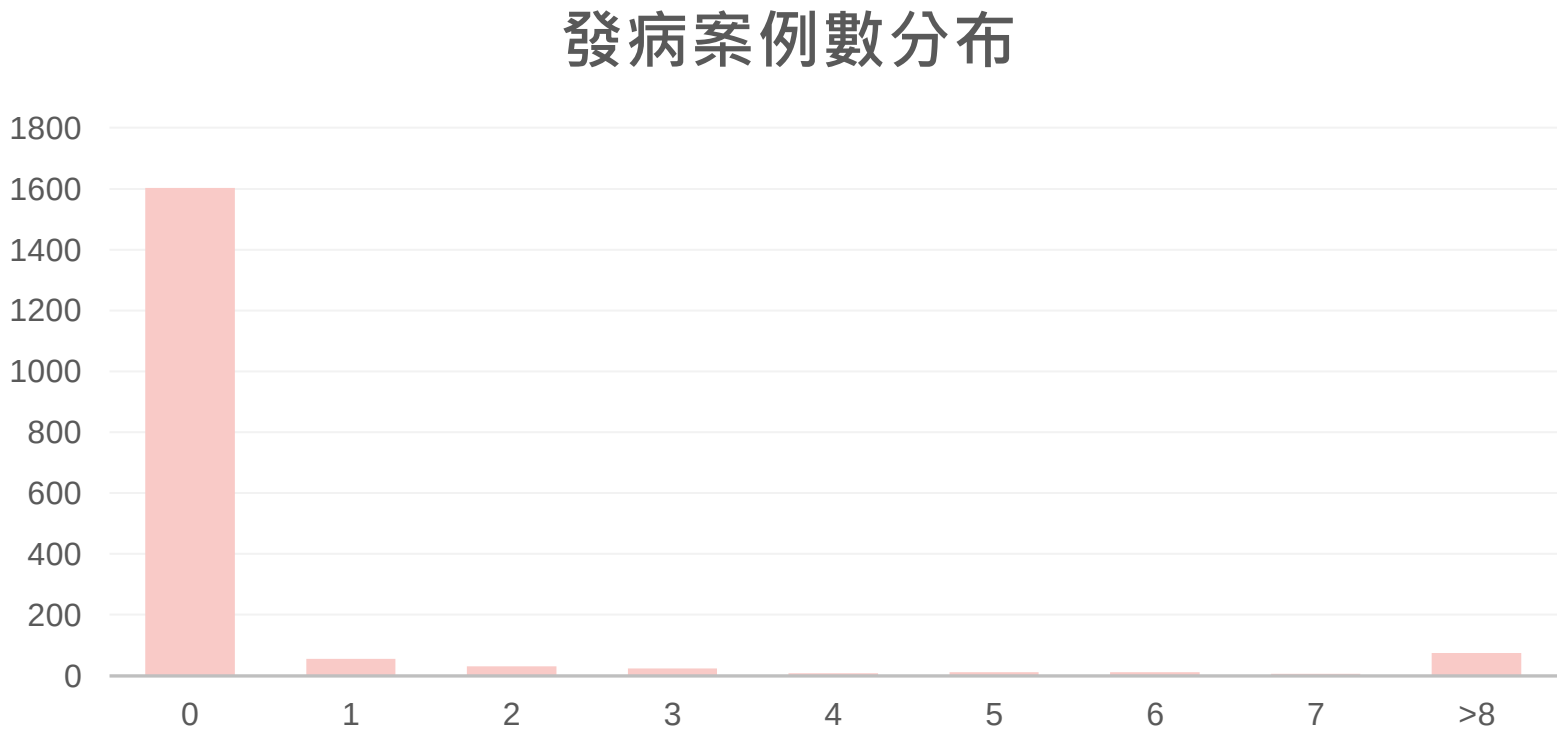

2020
40
30
20
10
0
一月 二月 三月 四月 五月 六月 七月 八月 九月 十月 十一月 十二月


2021
一月 二月 三月 四月 五月 六月 七月 八月 九月 十月 十一月 十二月

GLM

Zero-Inflated Model (零膨脹模型)

- Zero-Inflated Model零膨脹模型，運用在計數資料的實際研究中，**觀察事件發生數中含有大量的零值**。例如由於疾病發生率很低，觀察值有許多零。這種數據數資料中的零值過多，超出了Poisson分布或Negative Binomial等一般離散分布的預測能力。
- 相較於一般計數模型， Zero-Inflated Model多了一個參數π來估計不發生率。



Zero-Inflated Poisson vs. Poisson Distribution

$\boldsymbol{Y \sim ZIP(\pi, \mu)}$	$\boldsymbol{Y \sim Poi(\mu)}$
$P(Y=y)=\begin{cases}\pi+(1-\pi)\frac{e^{\mu}\mu^{y}}{y!}, & y=0\\ (1-\pi)\frac{e^{\mu}\mu^{y}}{y!}, & y=1,2,3,\dots\end{cases}$	$P(Y=y)=\frac{e^{\mu}\mu^{y}}{y!}, y=0,1,2,3,\dots$
$P(Y=0)=\pi+(1-\pi)e^{\mu}$	$P(Y=0)=e^{\mu}$
When $\boldsymbol{\pi=0}$, $\boldsymbol{Y \sim Poi(\mu)}$.	

Zero-Inflated Negative Binomial vs. Negative Binomial Distribution

$\boldsymbol{Y \sim ZINB(\pi, \mu, \alpha^{-1})}$	$\boldsymbol{Y \sim NB(\mu, \alpha^{-1})}$
$P(Y=y) = \begin{cases} \pi + (1-\pi)\frac{\Gamma(y+\alpha-1)}{y!\Gamma(\alpha-1)}\left(\frac{\alpha^{-1}}{\mu+\alpha^{-1}}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}, & y=0 \\ \frac{\Gamma(y+\alpha^{-1})}{y!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\mu+\alpha^{-1}}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}, & y=1,2,3,\dots \end{cases}$	$P(Y=y) = \frac{\Gamma(y+\alpha-1)}{y!\Gamma(\alpha-1)}\left(\frac{\alpha^{-1}}{\mu+\alpha^{-1}}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}, y = 0,1,2,3,\dots$
$P(Y=0) = \pi + (1-\pi)(1+\alpha\mu)^{-1/\alpha}$	$P(Y=0) = (1+\alpha\mu)^{-1/\alpha}$
When $\boldsymbol{\pi = 0}$, $\boldsymbol{Y \sim NB(\mu, \alpha^{-1})}$	

Zero-Inflated Poisson vs. Zero-Inflated Negative Binomial Distribution

$Y \sim ZIP(\pi, \mu)$	$Y \sim ZINB(\pi, \mu, \alpha^{-1})$
$P(Y=y) = \begin{cases} \pi + (1-\pi)\frac{e^{\mu}\mu^{y}}{y!}, & y = 0 \\ (1-\pi)\frac{e^{\mu}\mu^{y}}{y!}, & y = 1,2,3,\ldots \end{cases}$	$P(Y=y) = \begin{cases} \pi + (1-\pi)\frac{\Gamma(y+\alpha-1)}{y!\Gamma(\alpha-1)}(\frac{\alpha^{-1}}{\mu+\alpha^{-1}})^{\alpha^{-1}}(\frac{\mu}{\mu+\alpha^{-1}})^{y}, y = 0 \\ \frac{\Gamma(y+\alpha^{-1})}{y!\Gamma(\alpha^{-1})}(\frac{\alpha^{-1}}{\mu+\alpha^{-1}})^{\alpha^{-1}}(\frac{\mu}{\mu+\alpha^{-1}})^{y}, & y = 1,2,3,\ldots \end{cases}$
$E(Y) = (1-\pi)\mu$	$E(Y) = (1-\pi)\mu$
$Var(Y) = (1-\pi)\mu$	$Var(Y) = (1-\pi)\mu(1+\mu(\pi+\alpha))$
When $\alpha = 0$, $Y \sim ZIP(\pi, \mu)$	

Zero-Inflated Regression

library(pscl)

1. Poisson/Negative Binomial Regression：

$$\log(\mu) = \alpha + \beta X$$

α：X=0時，$\mu = e^{\alpha}$

β：X多1，μ多e^{β}倍

2. Logistic Regression：（logit link）

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X \ ; \ \pi = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

α：X=0時，本質是0的odds為e^{α}

β：X多1，本質是0的odds為e^{β}倍

Fitted Model 1&2 (drop cases = 0)

Model 1 : (Poisson Regression)

$$log(E(cases \mid X))$$
$$= \beta_0 + \beta_{region} I(region) + \beta_{season} I(season) + \beta_{temp\_last} x_{temp\_last} + \beta_{rainfall\_last} x_{rainfall\_last}$$

Model 2 : (Negative Binomial Regression)

$$log(E(cases \mid X))$$
$$= \beta_0 + \beta_{region} I(region) + \beta_{season} I(season) + \beta_{temp\_last} x_{temp\_last} + \beta_{rainfall\_last} x_{rainfall\_last}$$

Fitted Model 1&2 (drop cases = 0)

Theta : 0.4294

Std Err. : 0.0348

Model 1	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-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
β_{temp_last}	0.4628	0.0037	126.5948	< 0.0001
$\beta_{rainfall\_last}$	-0.0010	< 0.0001	-63.6723	< 0.0001

Model 2	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-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
β_{temp_last}	0.2330	0.0498	4.6755	< 0.0001
$\beta_{rainfall\_last}$	0.0004	0.0005	0.8078	0.4192

Fitted Model 3 (zero-inflated Poisson regression)

Count :

$$log(E(cases \mid X))$$

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

Zero :

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

```
zeironfl(cases ~ region + season + temp_last + rainfall_last | region + season, dt, dist = "poisson")
```

- base line : 中區冬季
- 與中區相比，
  北區發生登革熱機率小；
  南區發生登革熱機率較大。
- 而季節來說，秋冬之際發生登革熱的機率較大。
- 與前述模型有類似的結論

COUNT	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-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
β_{temp_last}	0.4642	0.0037	127.0435	< 0.0001
$\beta_{rainfall\_last}$	-0.001	< 0.0001	-69.4346	< 0.0001

ZERO	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	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

- base line : 中區冬季
- 就地區來看，北區不發生率的odds 顯著較低。
- 就季節來看，秋季不發生率的odds顯著較低。

Fitted Model 4 (zero-inflated negbin regression)

Count :

$$log(E(cases \mid X))$$

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

Zero :

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

```
zeroinfl(cases ~ region + season + temp_last + rainfall_last | region + season , dt, dist =  "negbin" )
```

- base line : 中區冬季
- 與中區相比，
  北區發生登革熱機率小；
  南區發生登革熱機率較大。
- 而季節來說，秋冬之際發生登革熱的機率較大。
- 與前述模型有類似的結論
- 而theta 估計為顯著因此使用 negbin 比 poisson 恰當。

COUNT	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-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
β_{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

ZERO	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	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

- base line : 中區冬季
- 各區域各季節登革熱的不發生率沒有顯著不同。

Fitted Model 5

Count :

$$log(E(cases \mid X))$$

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

Zero :

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

```
zeroinfl(cases ~ region + season + temp_last + rainfall_last | region , dt, dist =  "negbin" )
```

- base line : 中區冬季
- 與中區相比，
  北區發生登革熱幾率小；
  南區發生登革熱機率較大。
- 而季節來說，秋冬之際發生登革熱的機率較大。
- 與前述模型有類似的結論
- 而theta 估計為顯著因此使用 negbin 比 poisson 恰當。

COUNT	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-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
$\beta_{rainfall\_last}$	0.0001	0.0008	0.1692	0.8656
β_{temp_last}	0.3459	0.0594	5.8259	< 0.0001
Log(theta)	-2.4764	0.1399	-17.7024	< 0.0001

ZERO	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-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

- base line : 中區冬季
- 各區域各季節登革熱的不發生率（π）沒有顯著不同。

Fitted Model 6 (with PC)

Count：

$$log(E(cases \mid X)) = \beta_0 + \beta_{region} I(region) + \beta_{season} I(season) + \beta_{PC1} x_{PC1} + \beta_{PC2} x_{PC2} + \beta_{PC3} x_{PC3}$$

pc1 : 溫度相關；pc2 : 雨量相關；pc3 : 氣壓相關

Zero：

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

```
zeroinfl(cases ~ region + season + pc1+ pc2 +pc3 | region, dt, dist = "negbin")
```

- base line : 中區冬季
- 與中區相比，
  北區發生登革熱幾率小；
  南區發生登革熱機率較大。
- 而季節來說，秋冬之際發生登革熱的機率較大。
- 與前述模型有類似的結論
- 而theta 估計為顯著因此使用 negbin 比 poisson 恰當。

COUNT	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-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
β_{pc1}	1.2154	2.8057	0.4332	0.6649
β_{pc2}	0.9219	2.0628	0.4469	0.6549
β_{pc3}	-5.9025	1.8395	-3.2087	0.0013
Log(theta)	-3.1380	0.0791	-39.6775	<0.0001

ZERO	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-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

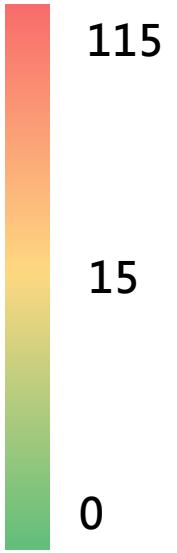
- base line : 中區冬季
- 各區域各季節登革熱的不發生率（π）沒有顯著不同。

模型比較

	Model	AIC	BIC	AUC
Model 4	• $log(cases) = \beta_0 + \beta_{region}I(region) + \beta_{season}I(season) + \beta_{temp\_last}x_{temp\_last} + \beta_{rainfall\_last}x_{rainfall\_last}$ • $log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_{region}I(region) + \beta_{season}I(season)$	2779.27	2861.75	0.852
Model 5	• $log(cases) = \beta_0 + \beta_{region}I(region) + \beta_{season}I(season) + \beta_{temp\_last}x_{temp\_last} + \beta_{rainfall\_last}x_{rainfall\_last}$ • $log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_{region}I(region)$	勝 2730.43	勝 2807.41	勝 0.854
Model 6	• $log(cases) = \beta_0 + \beta_{region}I(region) + \beta_{season}I(season) + \beta_{PC1}x_{PC1} + \beta_{PC2}x_{PC2} + \beta_{PC3}x_{PC3}$ • $log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_{region}I(region)$	2733.85	2733.85	0.840

Heatmap of Fitted Model 5

相對風險 (R.R.)	春	夏	秋	冬
北	0.005	0.100	0.645	0.409
中	0.012	0.243	1.577	1
南	0.836	17.317	112.180	71.158
東	0.006	0.134	0.865	0.549



相對風險：估計的案例數冬季中部（基準）估計的案例數的比值

Fitted Model 台南市&高雄市

Count：

$$\boldsymbol{cases} = \boldsymbol{\beta_0} + \boldsymbol{\beta_{season}} \boldsymbol{I(season)} + + \boldsymbol{\beta_{temp\_last}} \boldsymbol{x_{temp\_last}} + \boldsymbol{\beta_{rainfall\_last}} \boldsymbol{x_{rainfall\_last}}$$

Zero：

$$\boldsymbol{Logit(\pi)} = \boldsymbol{log}\left(\frac{\boldsymbol{\pi}}{\boldsymbol{1-\pi}}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_{高雄市}} \boldsymbol{I}(高雄市)$$

```
zeroinfl(cases ~ region + season + temp_last + rainfall_last | region , dt, dist = “negbin”)
```

- base line : 冬季
- 季節來說，秋冬之際發生登革熱的機率較大。
- 與前述模型有類似的結論
- 而theta 估計為顯著因此使用 negbin 比 poisson 恰當。

COUNT	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-6.2422	3.526	-1.7704	0.0767
$\beta_{season夏季}$	-1.6617	1.3239	-1.2551	0.2094
$\beta_{season春季}$	-3.8508	0.892	-4.3172	< 0.0001
$\beta_{season秋季}$	1.0593	1.3115	0.8076	0.4193
$\beta_{rainfall\_last}$	-0.0019	0.0018	-1.0387	0.299
β_{temp_last}	0.4726	0.1562	3.0252	0.0025
Log(theta)	-2.5963	0.1738	-14.9375	< 0.0001

ZERO	Est.	S.E.	Z	P-value
$\beta_{(Intercept)}$	-1.1125	0.8151	-1.3648	0.1723
$\beta_{County高雄市}$	-8.3627	105.4532	-0.0793	0.9368

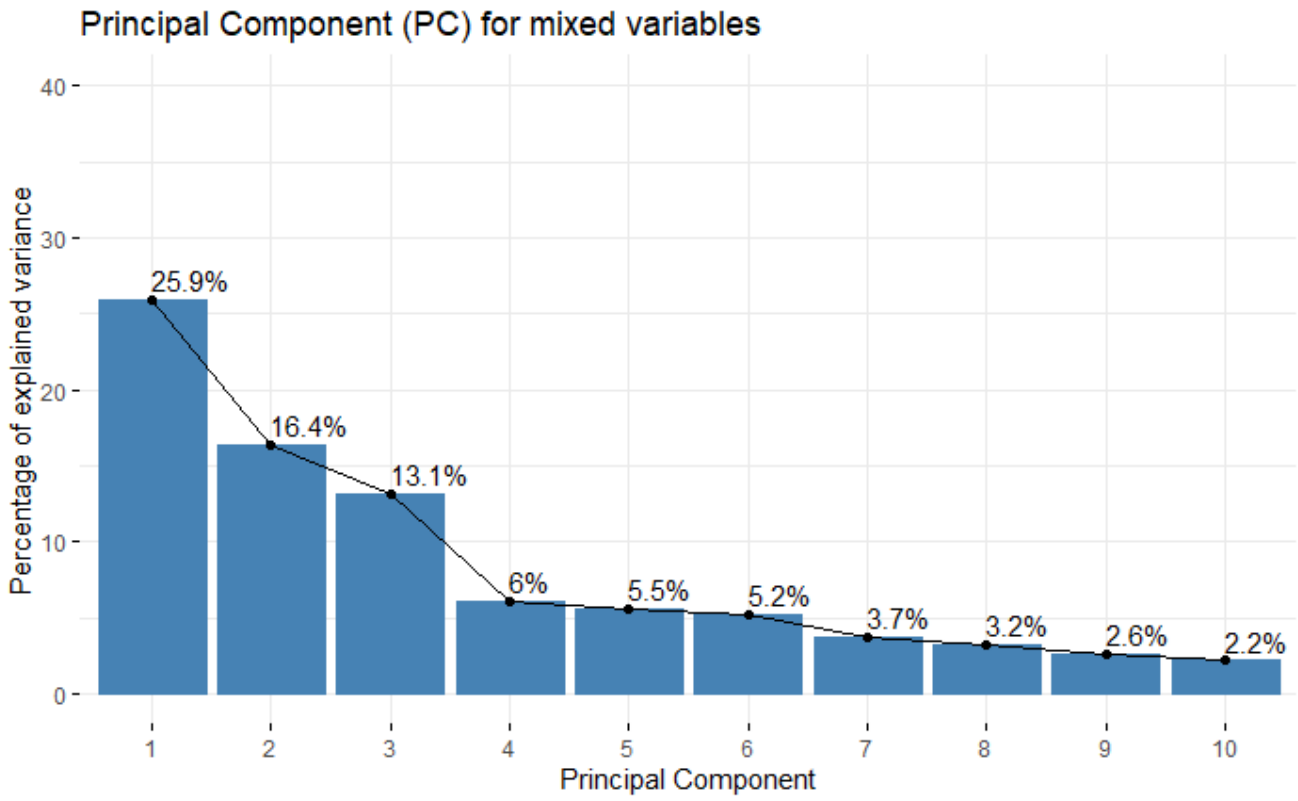
- base line : 臺南市
- 臺南市與高雄市登革熱的不發生率（π）沒有顯著不同。

降維、樹狀模型

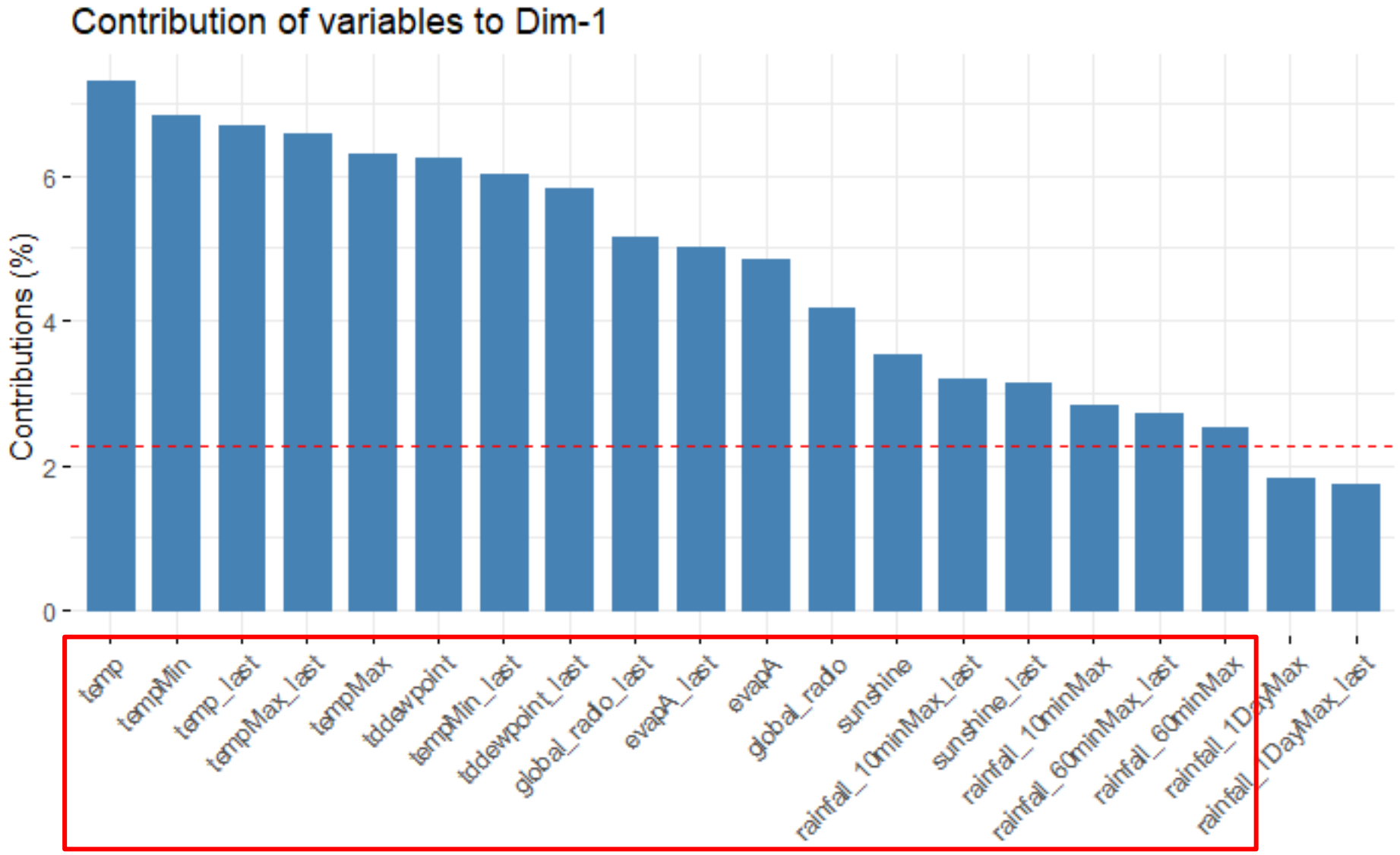
PCA Method

每個主成分解釋多少數據的變異

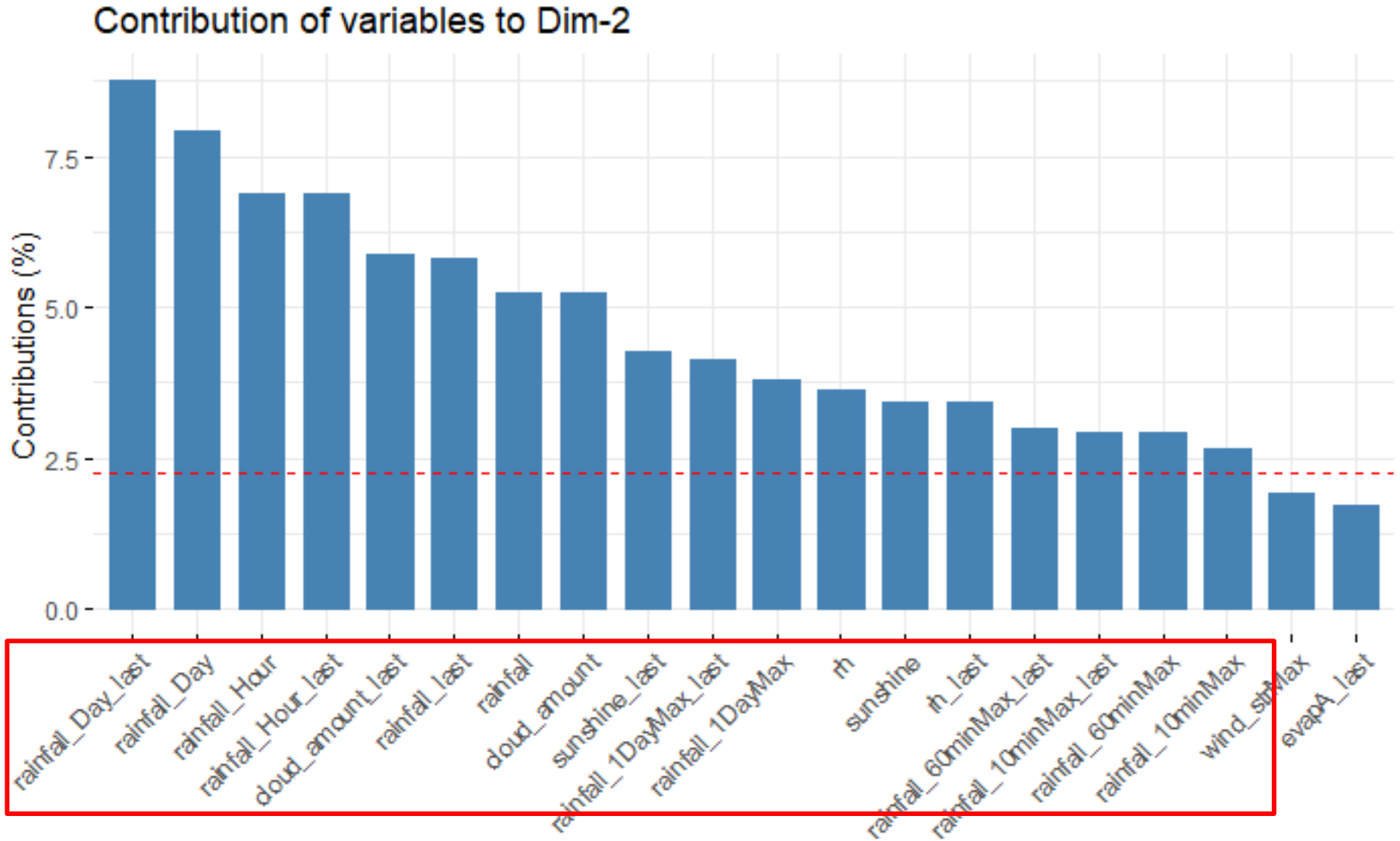
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	11.3983979108619	25.905449797413	25.90545
Dim.2	7.2051222600818	16.375277863822	42.28073
Dim.3	5.7663062205686	13.105241410383	55.38597
Dim.4	2.6576542988521	6.040123406482	61.42609
Dim.5	2.4381876989787	5.541335679497	66.96743
Dim.6	2.2692584510458	5.157405570559	72.12483
Dim.7	1.6399783395874	3.727223499062	75.85206
Dim.8	1.3939648352338	3.168101898259	79.02016
Dim.9	1.1307963885651	2.569991792193	81.59015
Dim.10	0.9549735487983	2.170394429087	83.76055
Dim.11	0.8764704005309	1.991978183025	85.75252
Dim.12	0.7425993581277	1.687725813927	87.44025
Dim.13	0.6661947644501	1.514079010114	88.95433
Dim.14	0.6020774565419	1.368357855777	90.32269
Dim.15	0.5735341887291	1.303486792566	91.62617
Dim.16	0.5080897202236	1.154749364144	92.78092
Dim.17	0.3886992761527	0.883407445802	93.66433
Dim.18	0.3666272335771	0.833243712675	94.49757
Dim.19	0.3123497227452	0.709885733512	95.20746
Dim.20	0.2538562396558	0.576945999218	95.78441
Dim.21	0.2375739007237	0.539940683463	96.32435
Dim.22	0.1989965992500	0.452264998295	96.77661
Dim.23	0.1614511667904	0.366934469978	97.14355
Dim.24	0.1468309123237	0.333706618918	97.47725
Dim.25	0.1375056946204	0.312512942319	97.78976
Dim.26	0.1234960405494	0.280672819430	98.07044
Dim.27	0.1111611381245	0.252638950283	98.32308
Dim.28	0.0902652528609	0.205148301957	98.52823
Dim.29	0.0791177208419	0.179813001913	98.70804
Dim.30	0.0753023631039	0.171141734327	98.87918
Dim.31	0.0685324407498	0.155755547159	99.03494
Dim.32	0.0645372725535	0.146675619440	99.18161
Dim.33	0.0633189217307	0.143906640297	99.32552



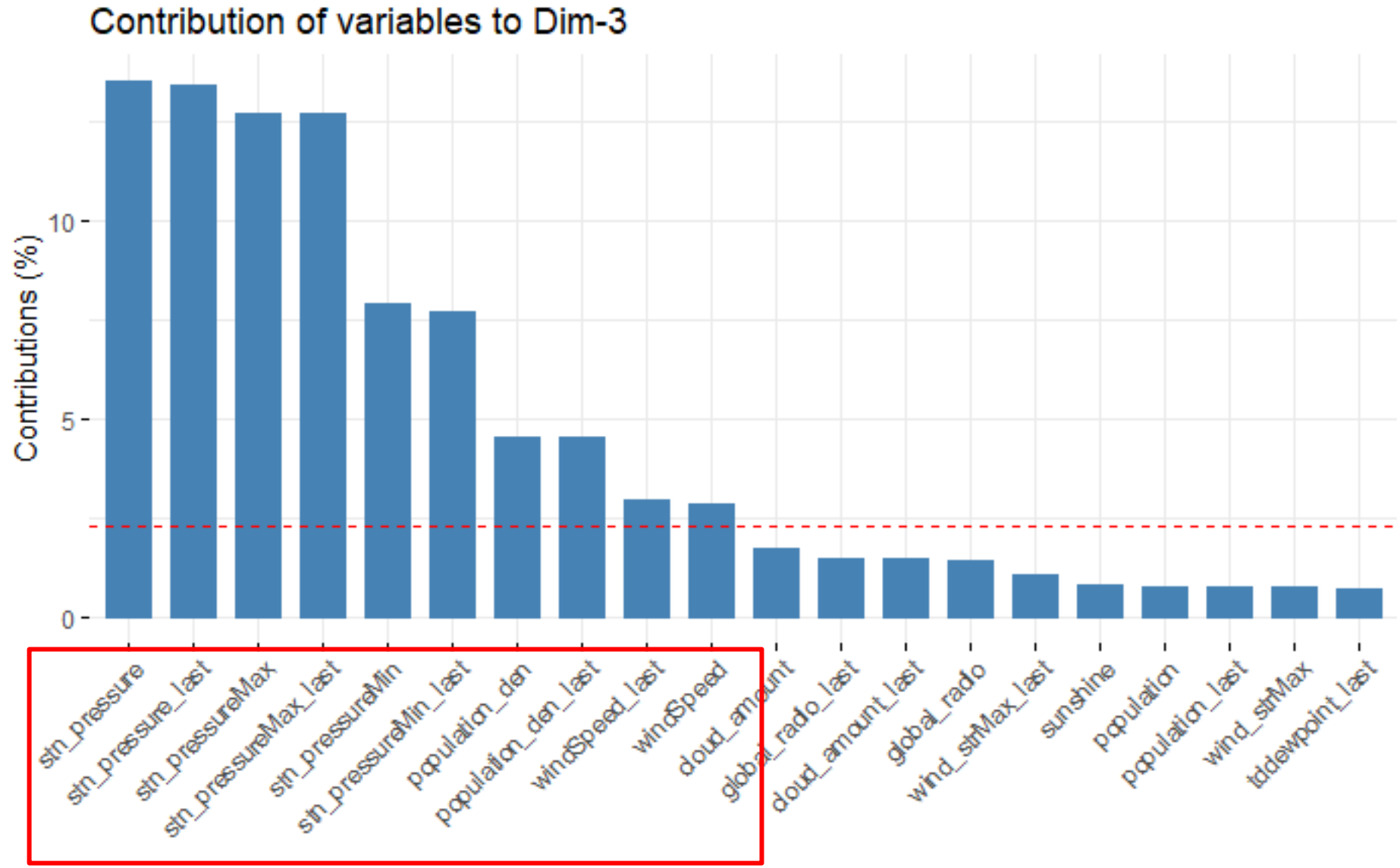
每個變數在主成分的貢獻比例 (超過平均貢獻視為重要變數)



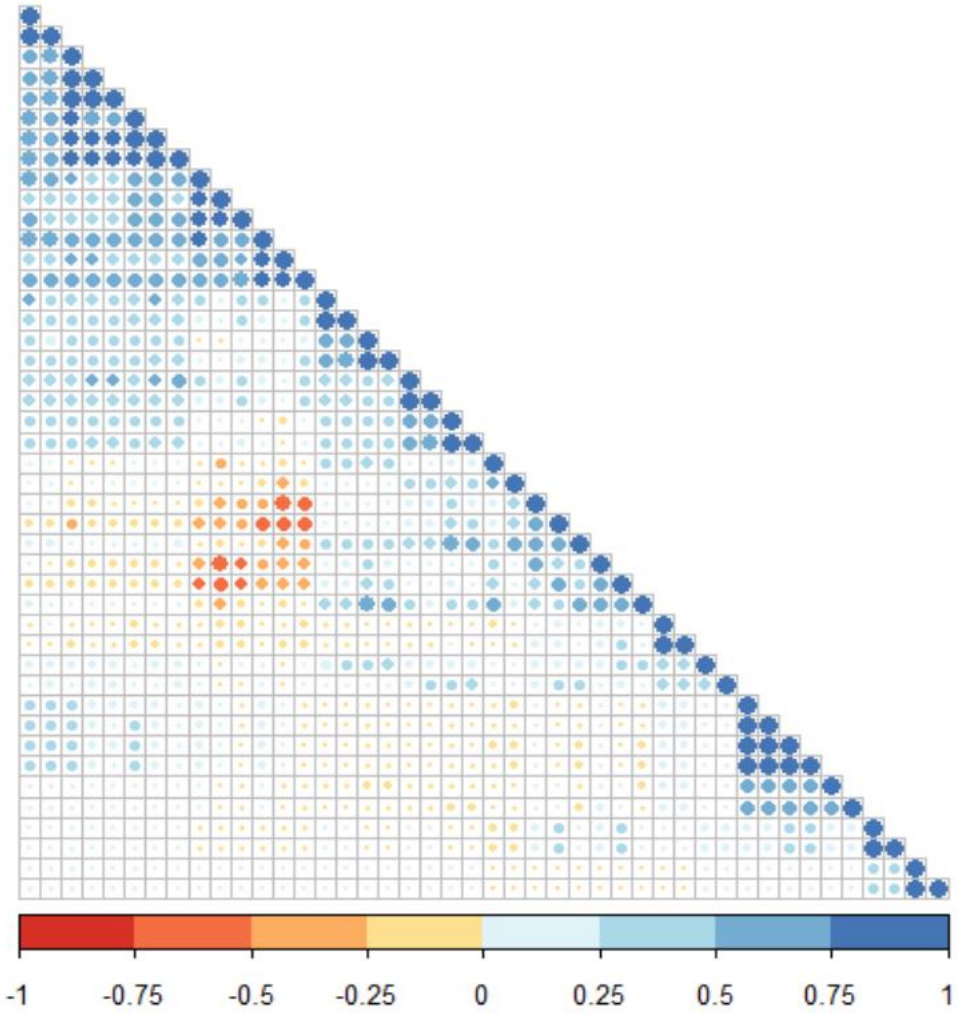
每個變數在主成分的貢獻比例 (超過平均貢獻視為重要變數)



每個變數在主成分的貢獻比例 (超過平均貢獻視為重要變數)



Correlation Plot of Variables



降維：Factor Analysis（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
- 對資料作摘要、變數篩選、轉軸、因子命名

1. 特徵值 (eigenvalues)：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

2. 因素負荷量 (factor loadings)：個別變數與因素之間的相關性，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

$$\text{變項 } 1 = \text{權重}_{11} * \text{因素 } 1 + \text{權重}_{12} * \text{因素 } 2 + \text{權重}_{12} * \text{因素 } 3 + \cdots + \text{誤差}_1$$

$$\text{變項 } 2 = \text{權重}_{21} * \text{因素 } 1 + \text{權重}_{22} * \text{因素 } 2 + \text{權重}_{23} * \text{因素 } 3 + \cdots + \text{誤差}_2$$

...

$$\text{變項 } 50 = \text{權重}_{501} * \text{因素 } 1 + \text{權重}_{502} * \text{因素 } 2 + \text{權重}_{503} * \text{因素 } 3 + \cdots + \text{誤差}_3$$

Factor Analysis vs. Principle Component Analysis

- 主成份分析 (PCA) 與因子分析(FA) 是利用不同的方法來減少變數數量

	主成分(PCA)	因子分析(FA)
方法	將多個變數，縮減到 少數個主成份，在這同時儘量保留變數的variation	將多個變數用少數個factor所組成的線性關係表示
解釋	variance	covariance

8種分析組合

- 為何要使用8種分析組合？

1. 此資料有8成病例數都是0例 → 資料選取分成考慮 / 不考慮病例數為0的情況
2. 因子分析的方法會影響各因子的負載量、因子如何解釋 → 比較MLM、PCM兩種方法的預測效果
3. 決策樹僅用一棵樹分群，結果可能有偏差 → 考慮GBDT模型，比較兩模型的預測效果

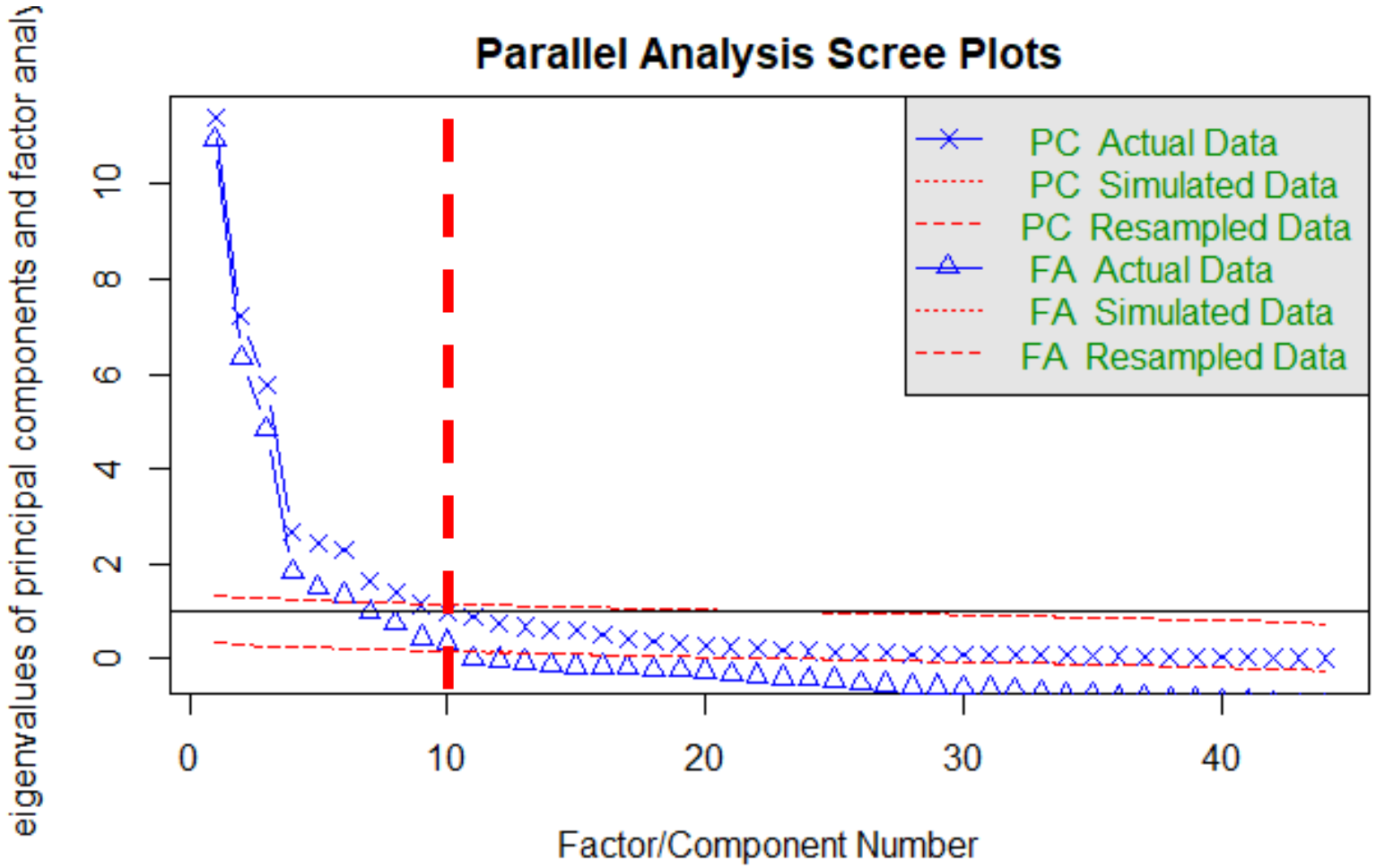
8種分析組合

- 目標：

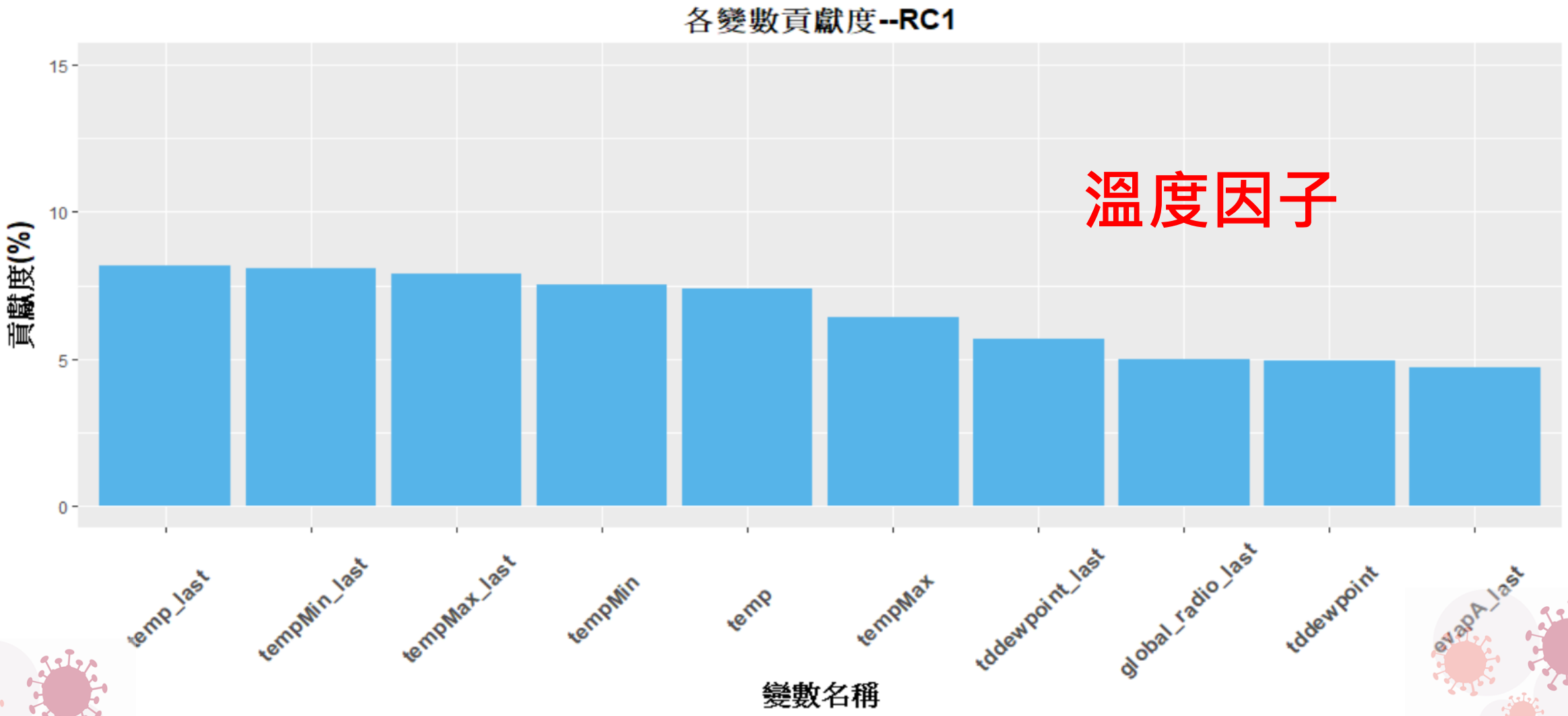
1. 觀察資料考慮 / 不考慮病例數為0的情況對結果有無差異
2. 觀察MLM / PCM 哪種因子分析方法對預測結果較好
3. 觀察DT / GBDT 哪種模型的預測結果較好
4. 透過DT / GBDT 模型結果視覺化觀察變數重要性、變數對於分群的影響

如何選擇因子個數?

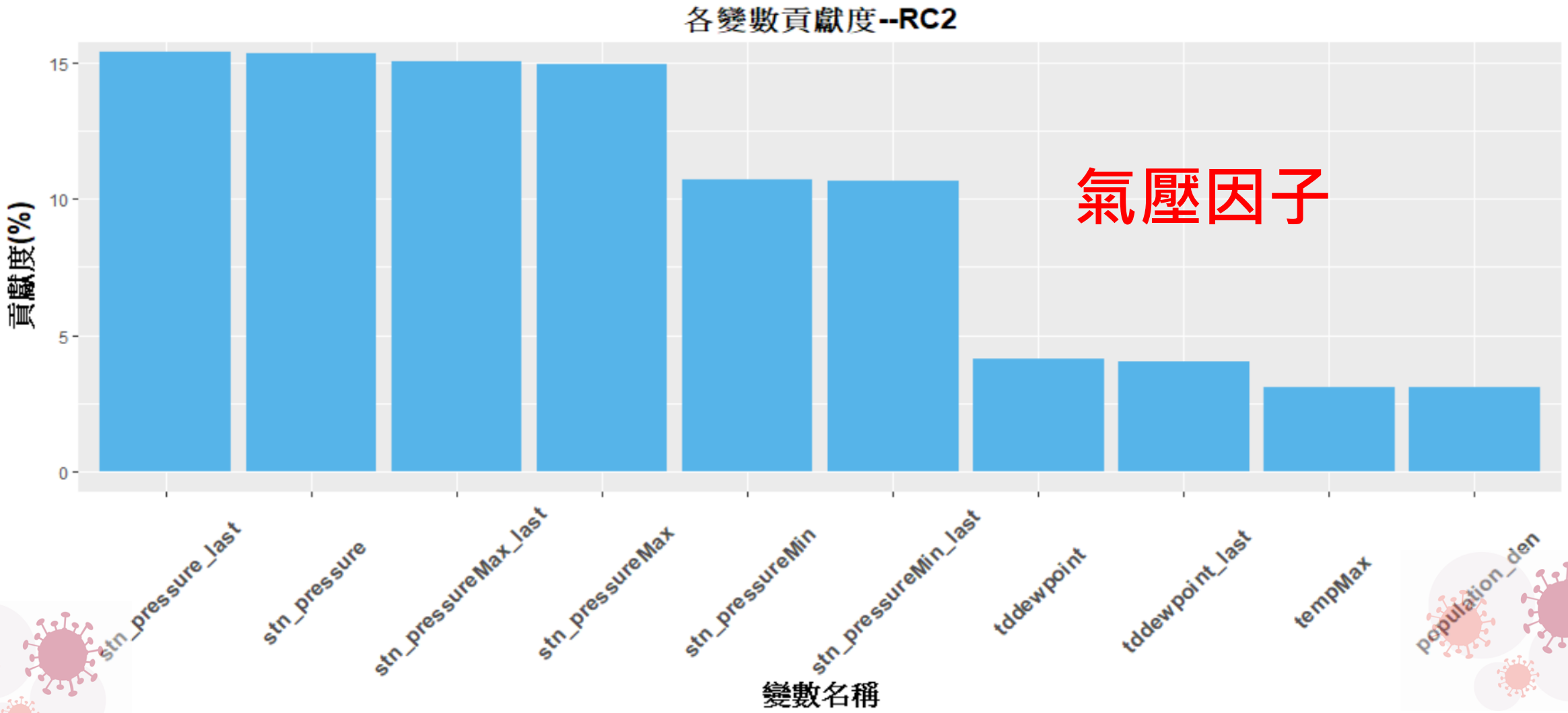
- 選擇因子套件 psych::fa.parallel



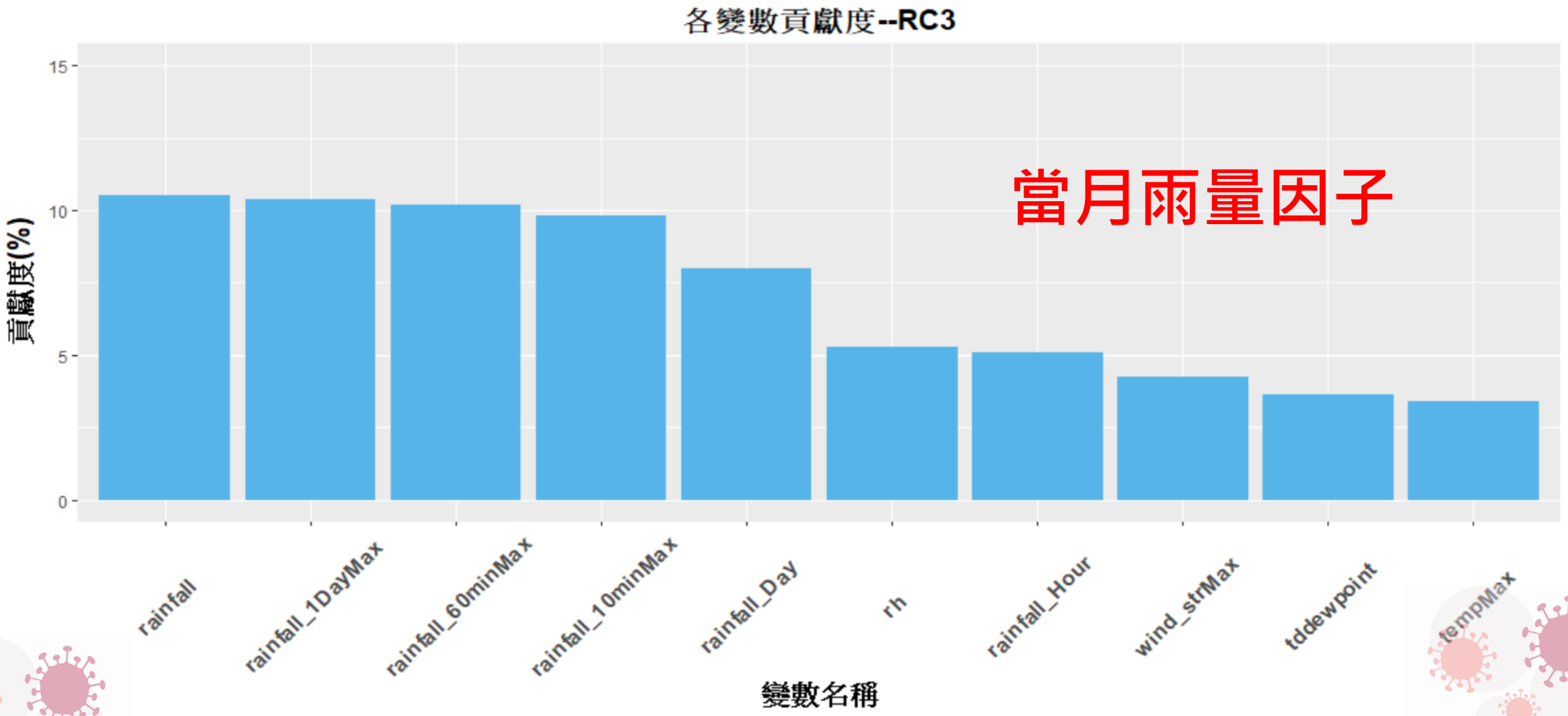
解釋因子


各變數貢獻度--RC1
貢獻度(%)
15
10
5
0
temp_last
tempMin_last
tempMax_last
tempMin
temp
tempMax
tddewpoint_last
global_radio_last
tddewpoint
evapA_last
變數名稱
溫度因子

如何命名因子


各變數貢獻度--RC2
貢獻度(%)
15
10
5
0
stn_pressure_last
stn_pressure
stn_pressureMax_last
stn_pressureMax
stn_pressureMin
stn_pressureMin_last
tddewpoint
tddewpoint_last
tempMax
population_den
變數名稱
氣壓因子

如何命名因子


各變數貢獻度--RC3
貢獻度(%)
15
10
5
0
rainfall
rainfall_1DayMax
rainfall_60minMax
rainfall_10minMax
rainfall_Day
rh
rainfall_Hour
wind_strMax
tddewpoint
tempMax
變數名稱
當月雨量因子

模型驗證&評估指標

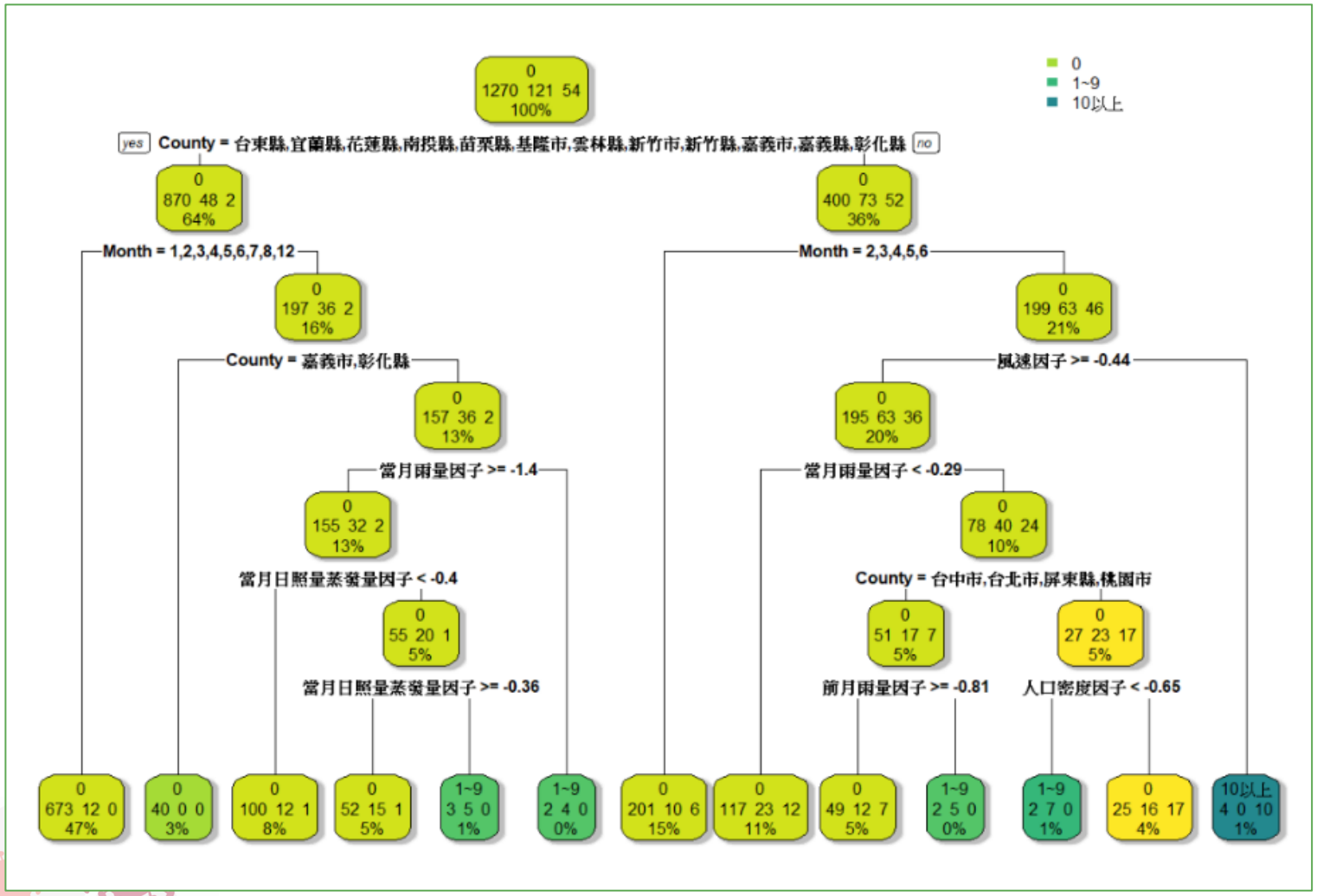
- 模型驗證
  - 20% test, 80% 作Cross Validation (train+validation)
  - 納入病例數為0：10-fold CV，排除病例數為0：Stratified 10-fold CV+SMOTE imbalance
- 評估指標
  - Accuracy, AUC

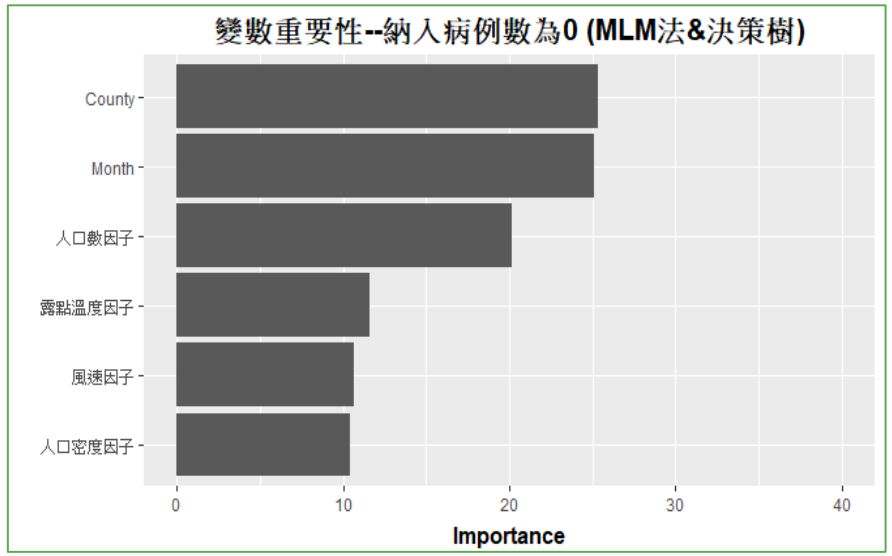
8種分析組合的預測效果

- 納入病例數為0的情況
  - DT / GBDT 模型效果差不多，MLM效果較好
- 排除病例數為0的情況
  - GBDT模型效果明顯比DT好，DT模型的PCM法效果較好，GBDT模型則是MLM法較好

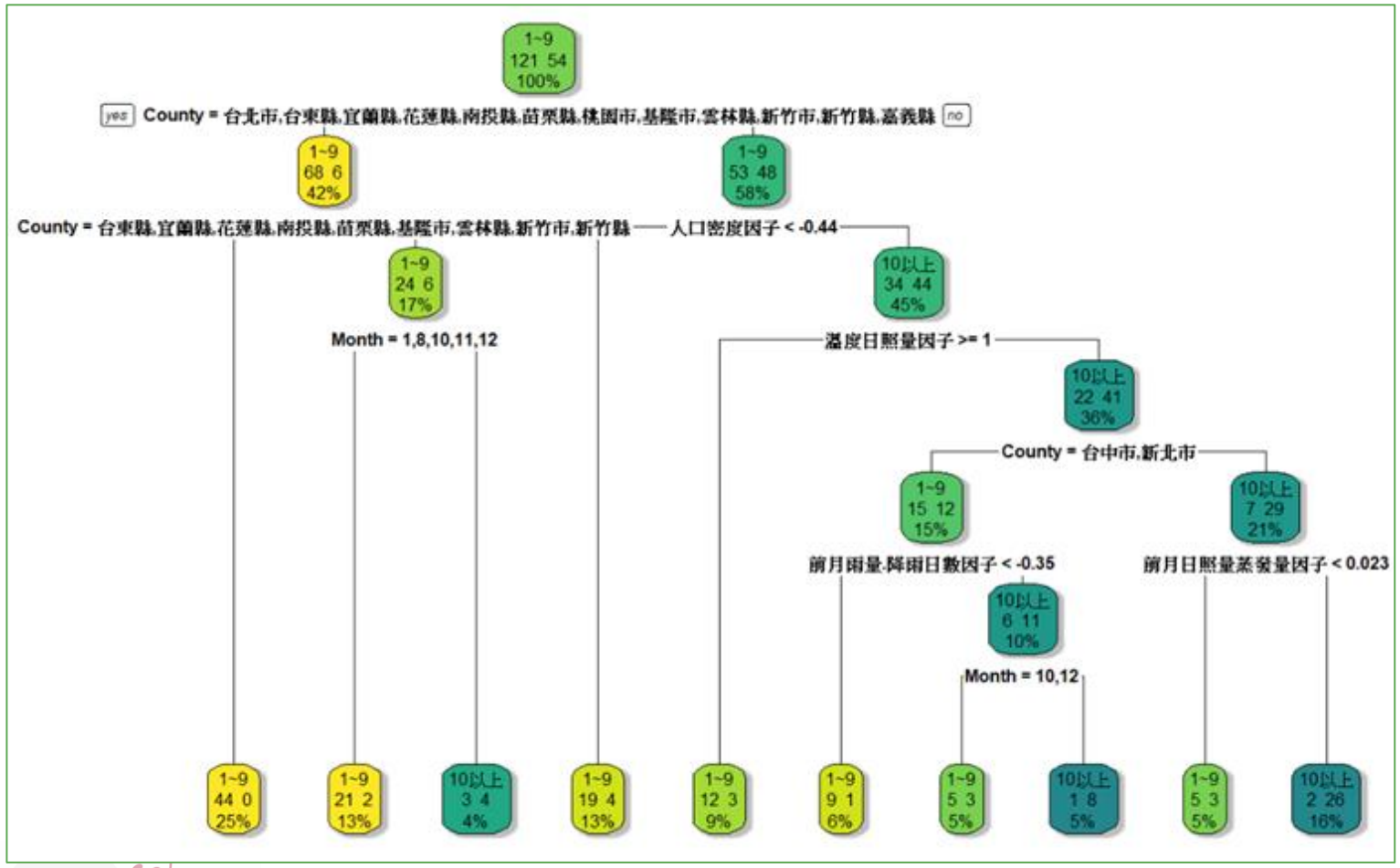
資料選取	納入病例數為0				排除病例數為0			
模型	DT		GBDT		DT		GBDT	
因子分析方法	MLM	PCM	MLM	PCM	MLM	PCM	MLM	PCM
Accuracy	0.8694	0.8611	0.8806	0.8222	0.6512	0.7209	0.7674	0.7209
AUC	0.7794	0.7034	0.7479	0.774	0.8227	0.673	0.8773	0.8132

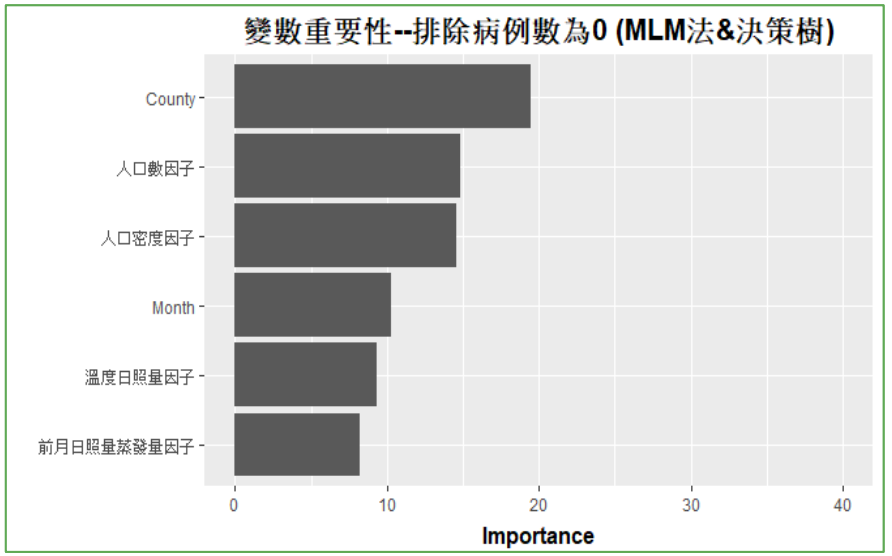
Decision Tree ： MLM法 (資料：納入病例數為0)



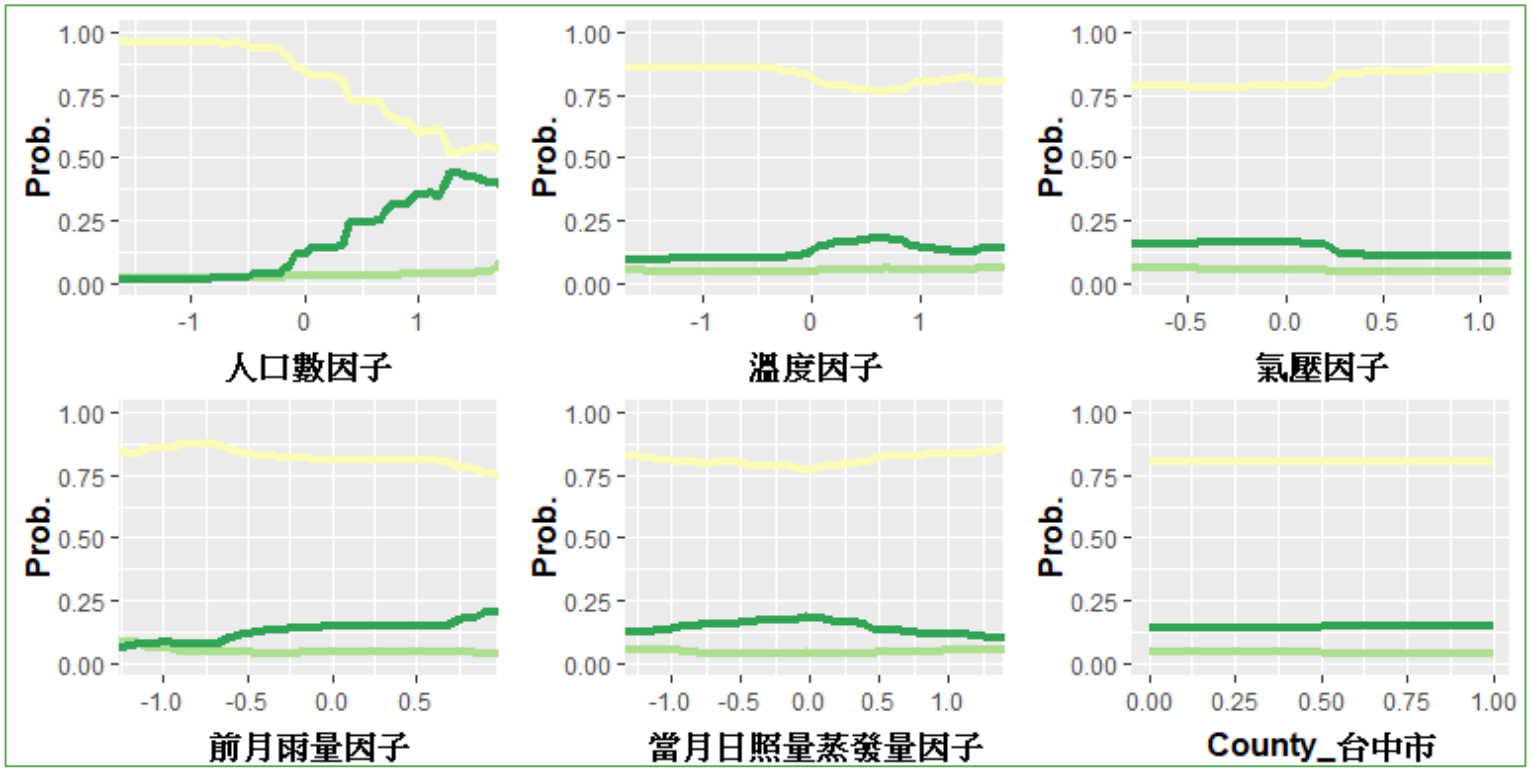


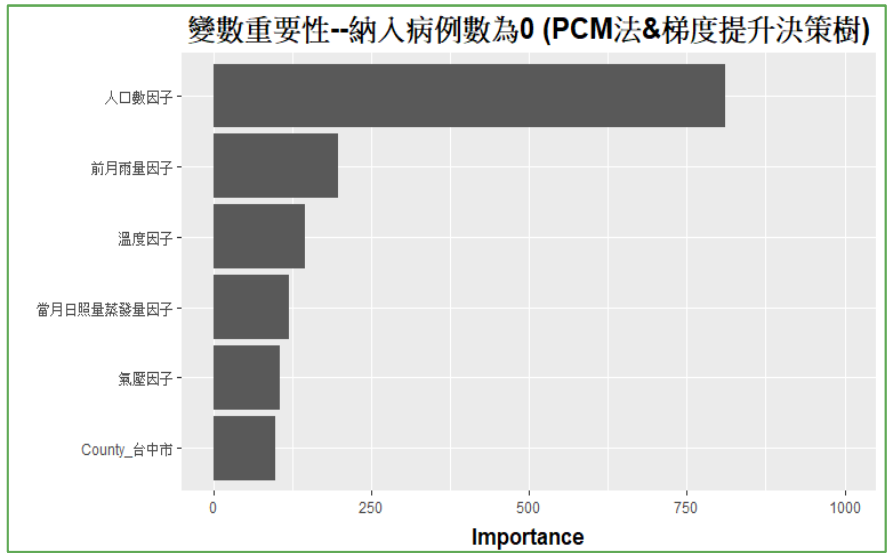
Decision Tree ： MLM法 (資料：排除病例數為0)



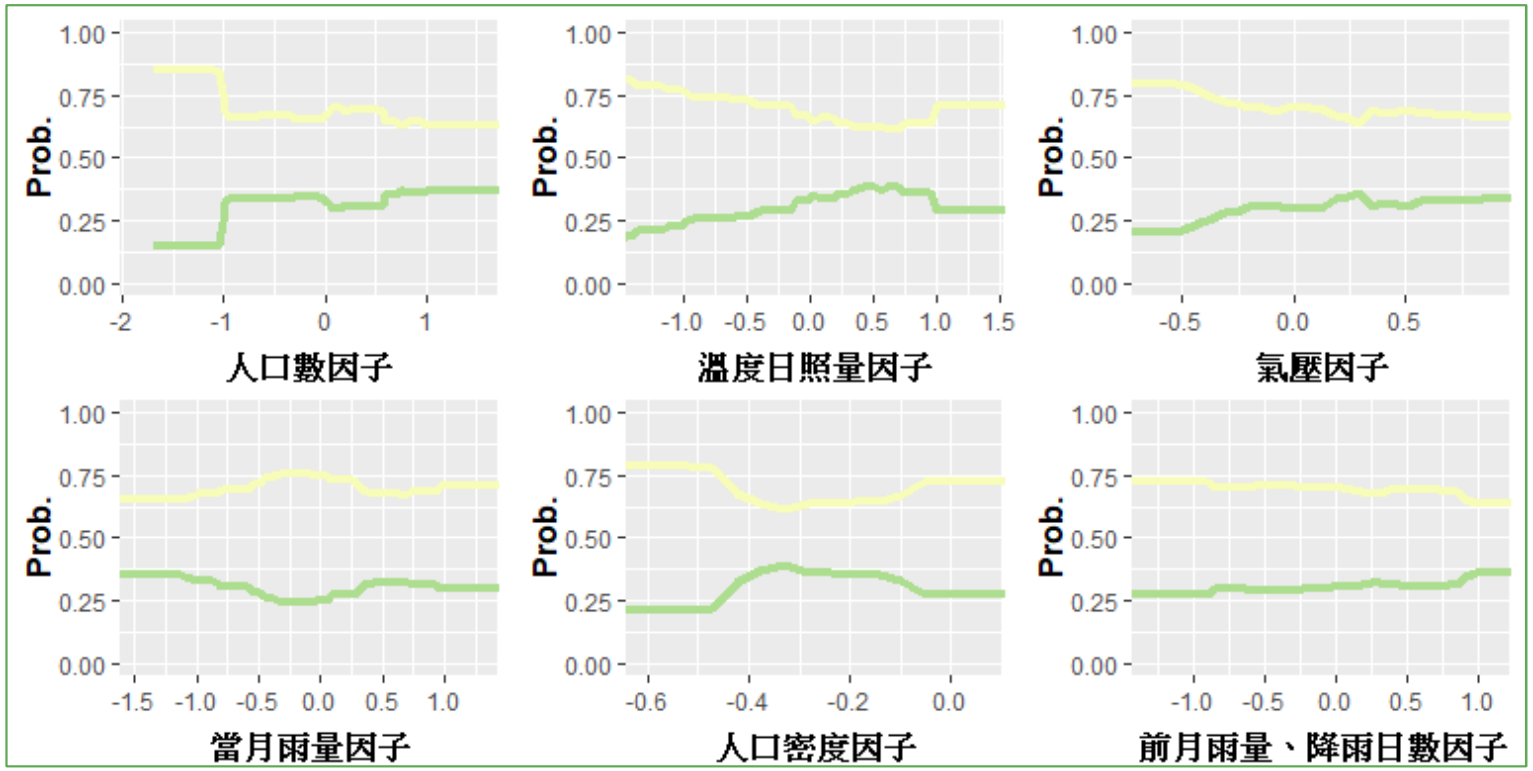


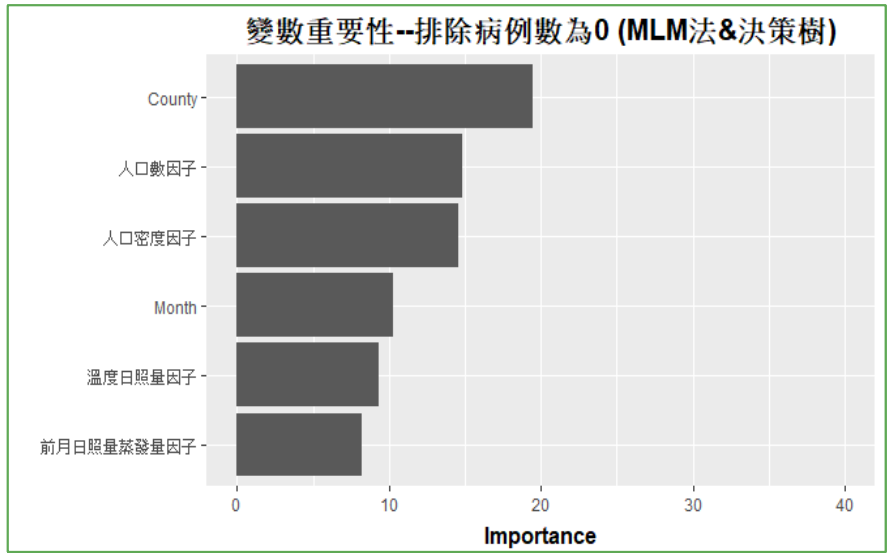
GBDT： PCM法 (資料：納入病例數為0)





GBDT： MLM法 (資料：排除病例數為0)





結論

- 由所配適的GLM模型來看，在**秋季及冬季**時，**南部**地區若前一個月**降水多**且**氣溫高**，則須嚴防登革熱的爆發。
- 由Zero配飾的模型來看，各地區各個月分登革熱的不發生率沒有顯著差異，所以**各地仍須嚴防登革熱的發生**。
- 由決策樹來看，雖然分類的狀況不佳，但大致可抓到**人口**與**月份**是重要的因子，**8到10月人口數較多**的地方，登革熱案例數會較多；人口數較少且降雨量較少的地區，登革熱案例數會較少。

The End