Group1
108024522 劉軒成
108024503 莊仕祺
108024466 劉倍銘
108024703 陳昱瑋
108024520 林敬皓
108024507 張文騰

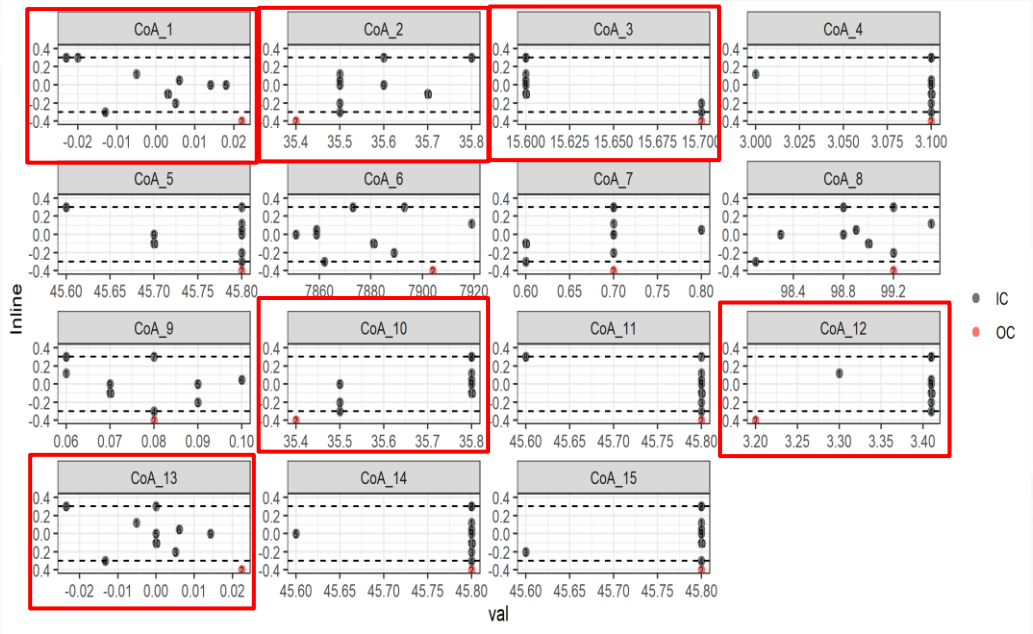
Original problem

1.能否收到物料的COA時就知道inline會異常?

2.每天收到這麼多批原物料的COA資料，能不能從中盲測出哪一批料會有問題?

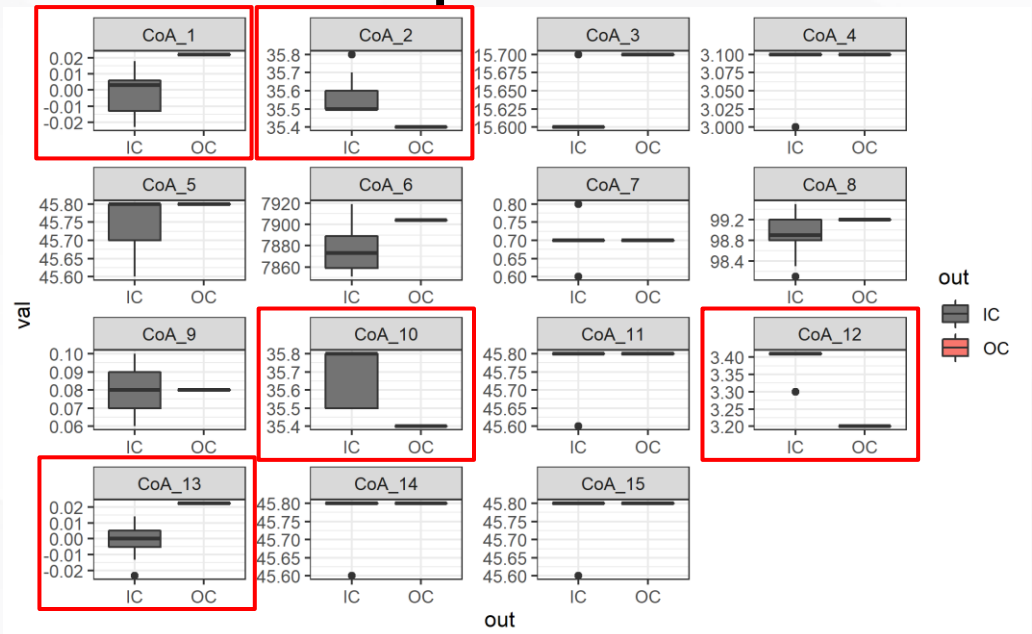
Preliminary data analysis-1

Inline 1 CoA vs Inline



- ***IC***為inline值正常的9個批次、***OC***代表inline出現異常的第2批次的數值
- 可以發現在CoA 1、2、10、12、13上，CoA的值與Inline有線性的pattern

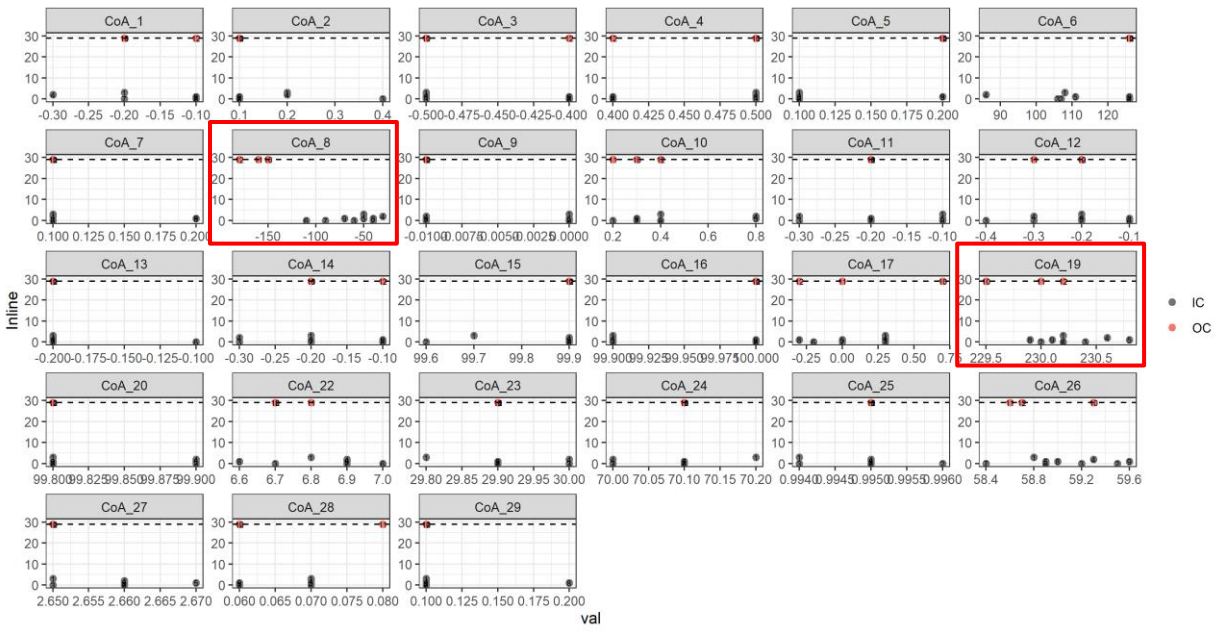
Inline 1 CoA boxplot



- 第2批次在CoA 1、2、10、12、13的數值和其他9個批次較有差異

Preliminary data analysis-2

Inline 2 CoA vs Inline



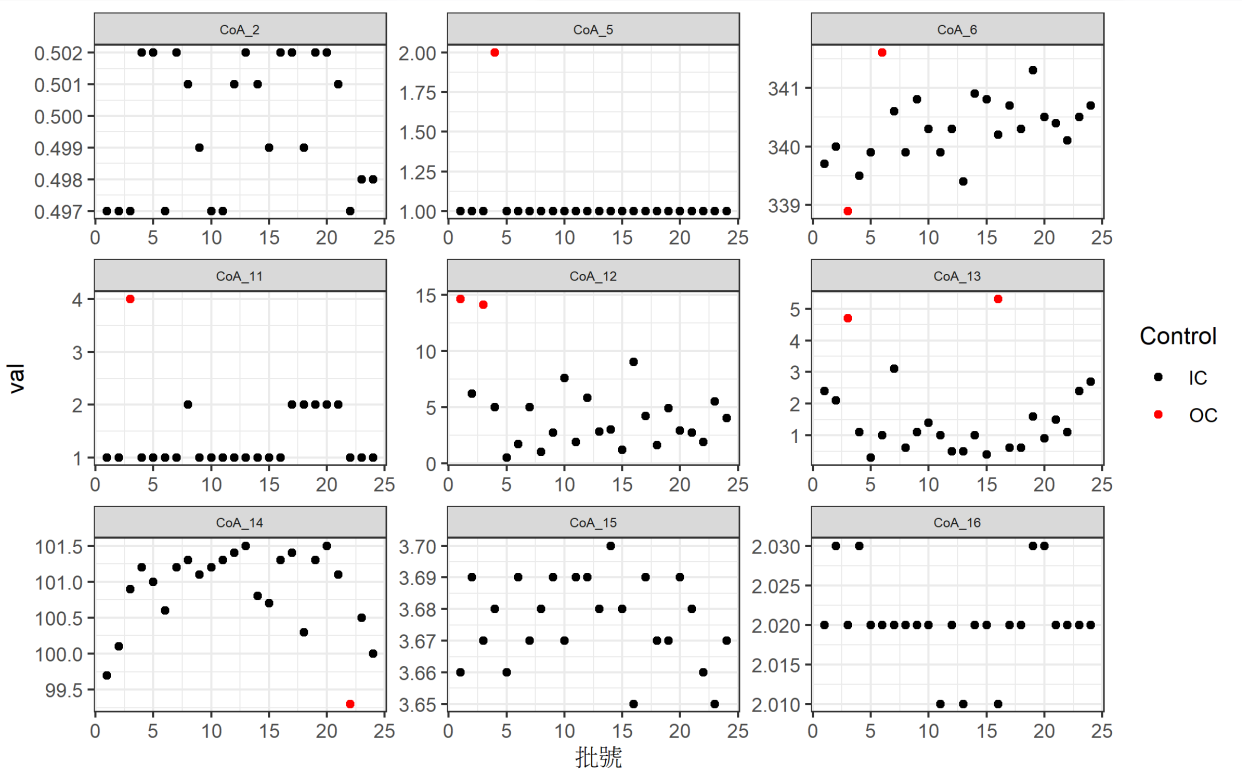
Inline 2 CoA boxplot

- ***IC***為inline正常的9個批次、***OC***代表inline異常的第10、11、12批次的數值
- 可以發現在CoA 8、19上，CoA的值與Inline有線性的pattern

- 在CoA 8上有明顯數值分佈差異
- 在進行inline1、2兩組資料判別時，上述較有差異的CoA可能會有較大的影響

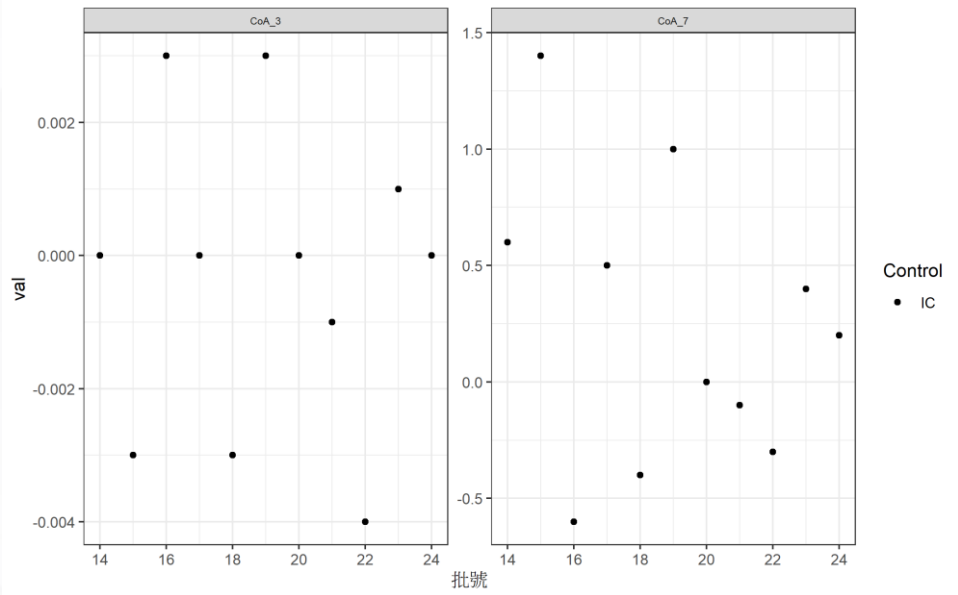
Preliminary data analysis-3

Material 1 批次vs CoA



- 紅點代表該批次在這個CoA中**數值距離平均超過兩個標準差**，因此我們認定其為異常值並特別挑出

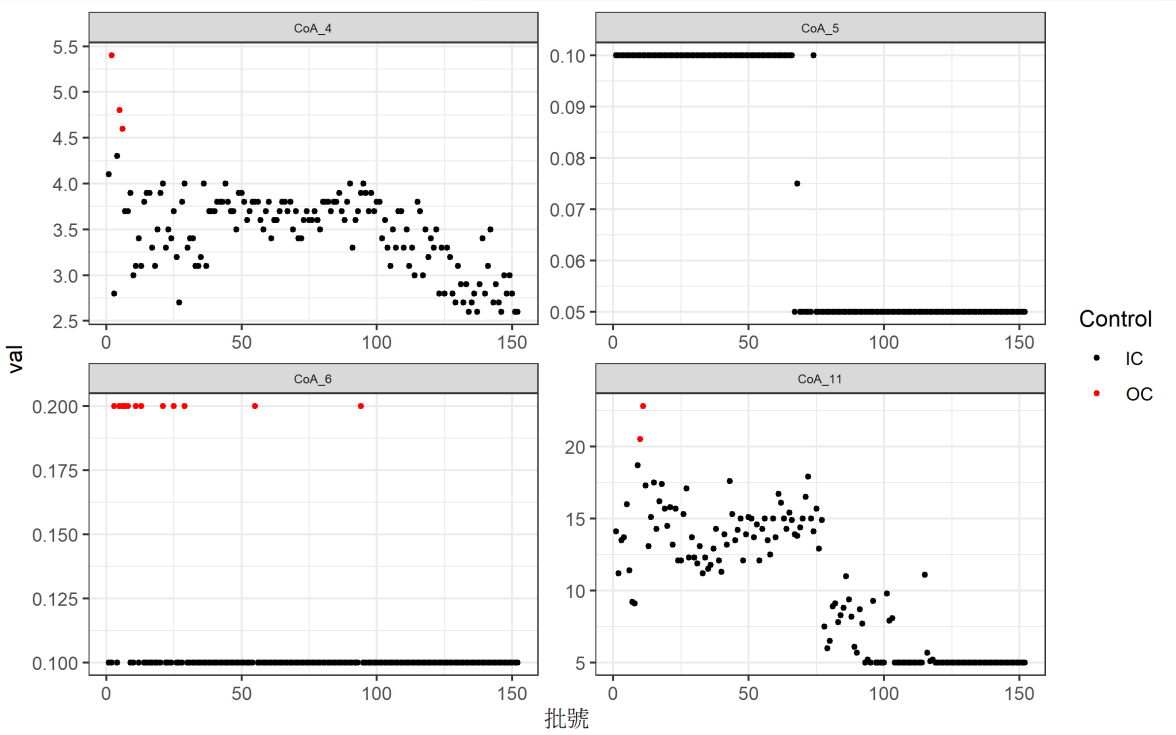
Missing CoA



- 此方法會挑出數值較為極端的批次，或是在較多固定值的CoA中挑出少數偏離的點
- CoA3、7為Material1缺值的變數，發現數值都判定為***IC***

Preliminary data analysis-4

Material 2 批次vs CoA

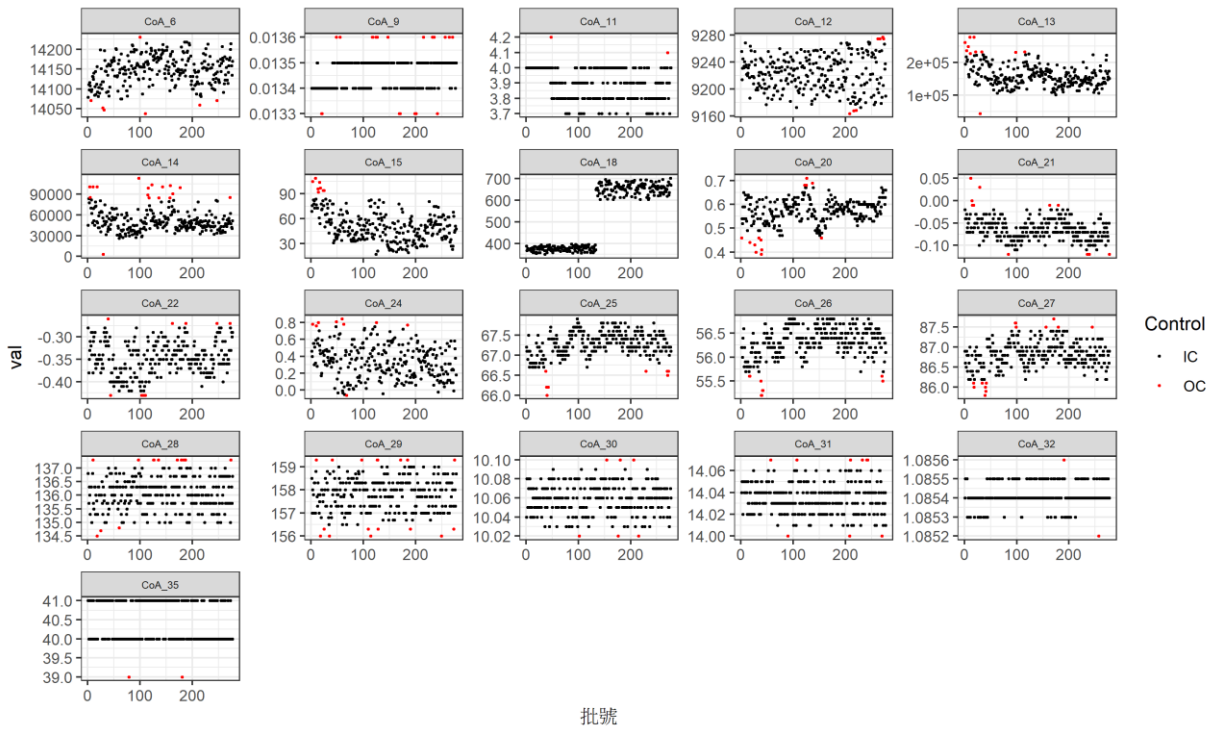


- 隨批次增加，前後的CoA數值有很大的趨勢變化
- 從CoA_6發現：
  當CoA值為兩組固定數值時，兩個標準差的方法會將比例較少的一方視為異常值
- 從CoA_5發現：
  當數值分散比例接近時，兩個標準差的方法會使所有批次皆被視為正常

➢ 兩個標準差的判定方法可能需要視CoA不同而調整

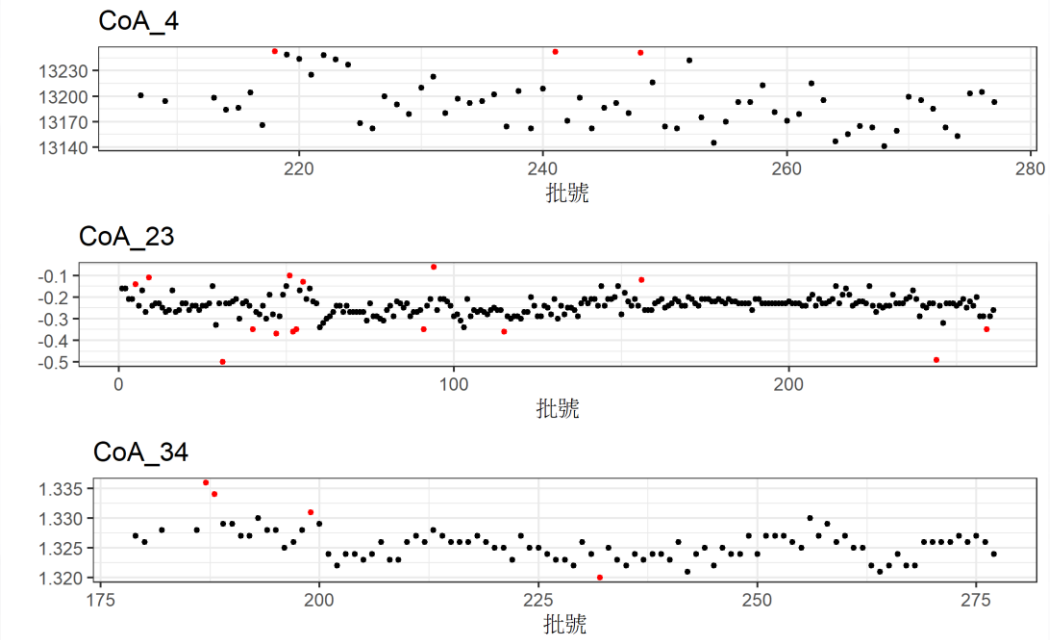
Preliminary data analysis-5

Material 3 批次vs CoA



- 在CoA 18中同樣出現隨批次而改變趨勢的現象
- 在最後幾組CoA中，數值大多集中在某些特定值上

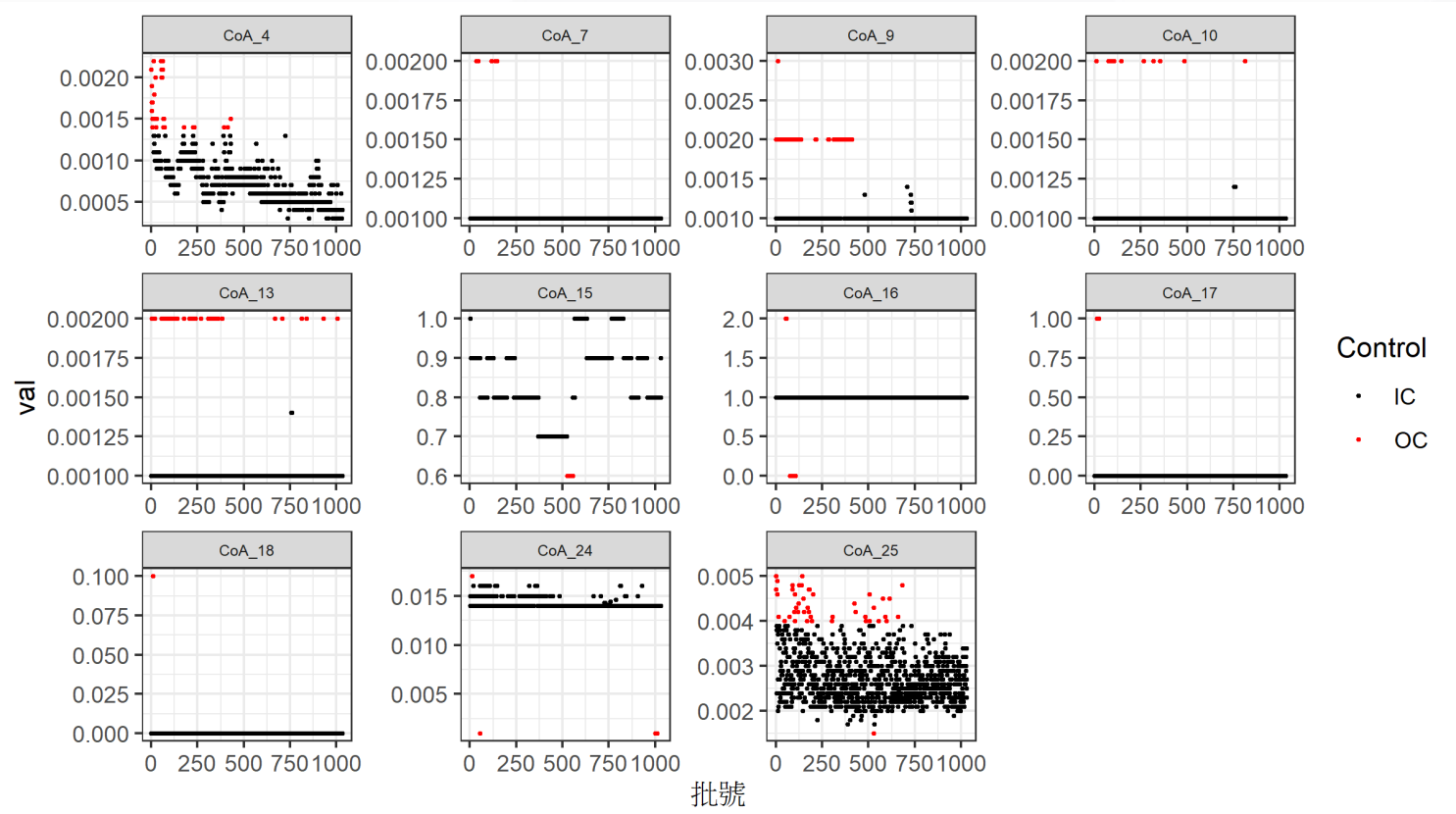
Missing CoA



- 上圖為Material3缺值的CoA，發現缺的批號都不同，可以發現出現***OC***的批號較多出現在前段

Preliminary data analysis-6

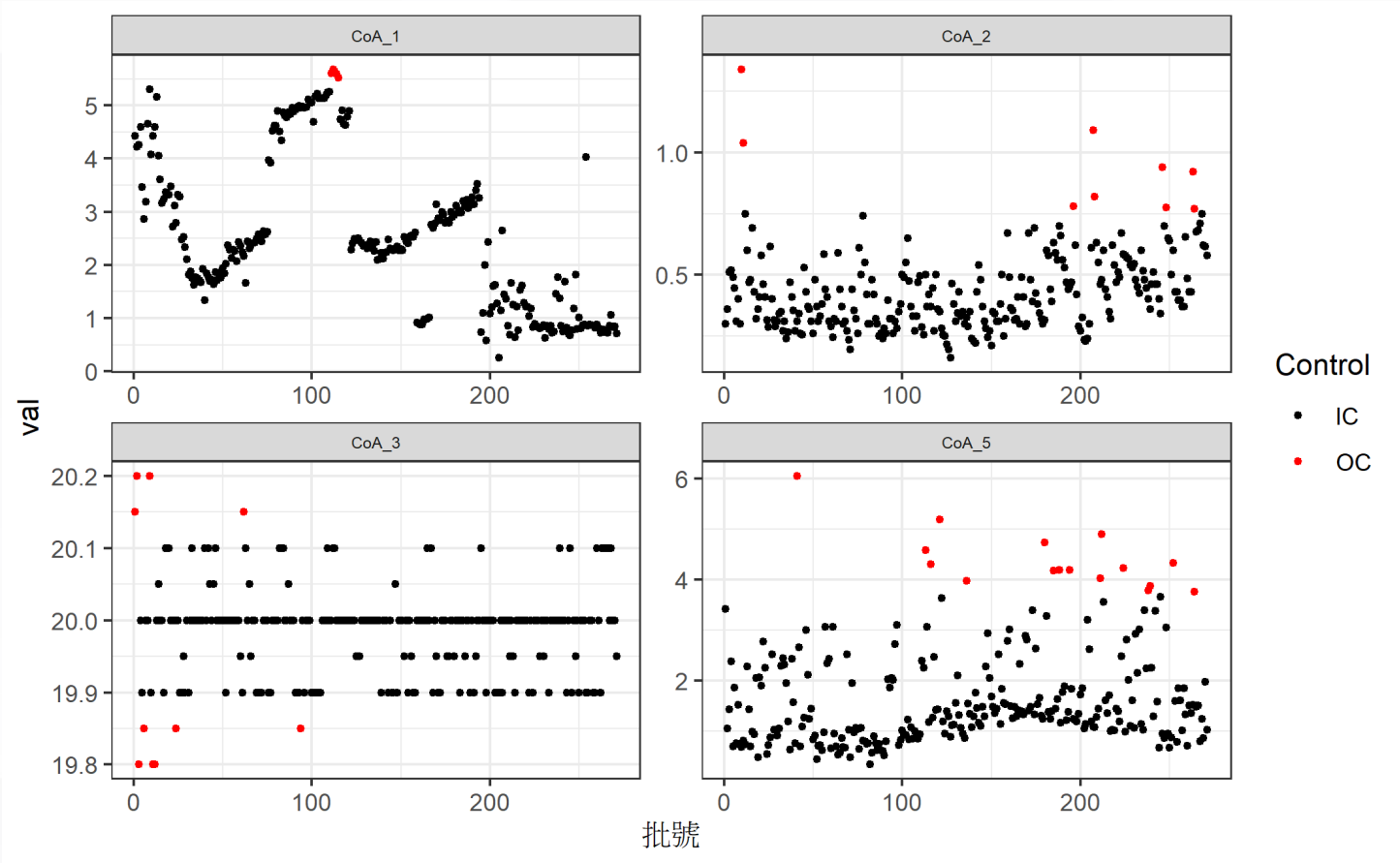
Material 4 批次vs CoA



- CoA 4的數值隨批次增加而下降
- CoA 25的數值在前半段較常出現較大值，但整體來說都落在0.003左右
- 其他CoA皆為幾組固定數值，並出現少數偏離固定值的批次

Preliminary data analysis-7

Material 5 批次vs CoA

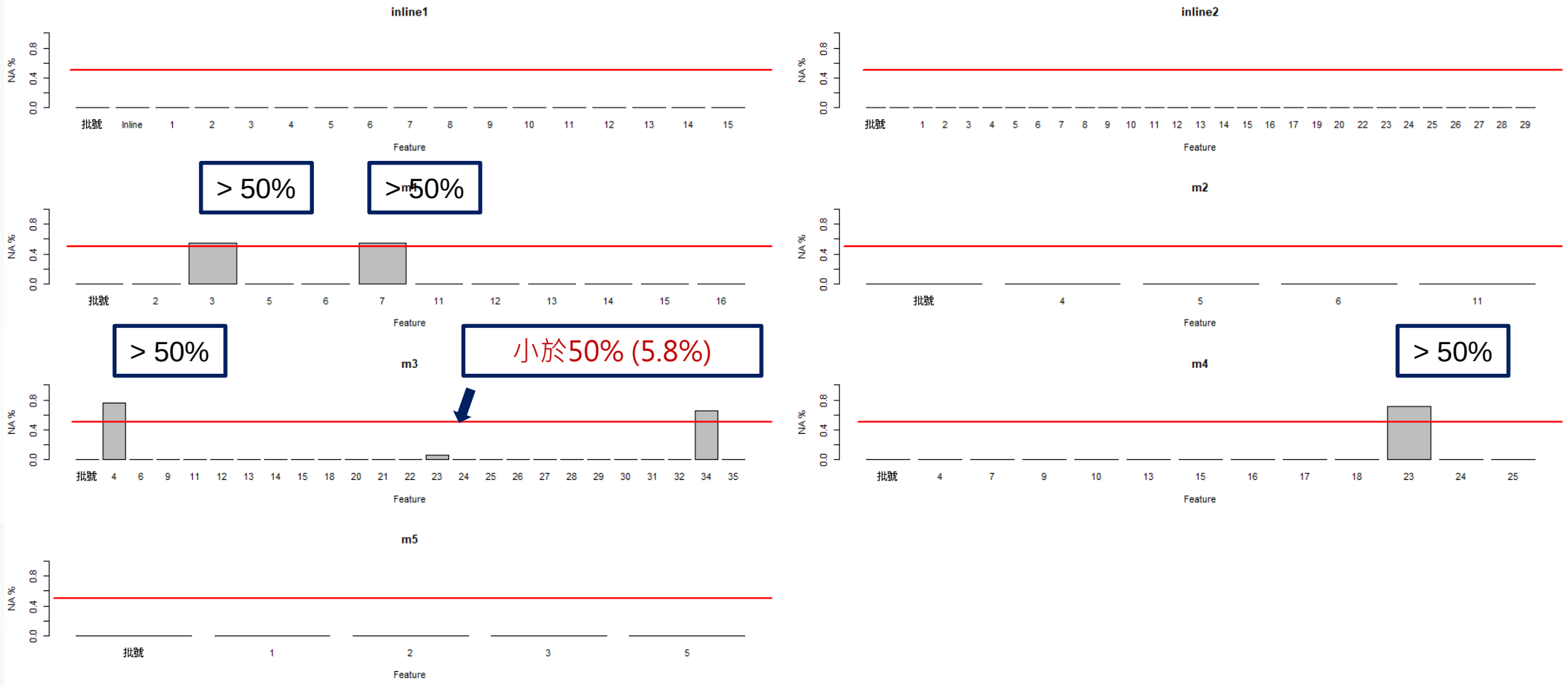


- CoA 1的數值隨批次出現上下的變動

Missing Value Treatment

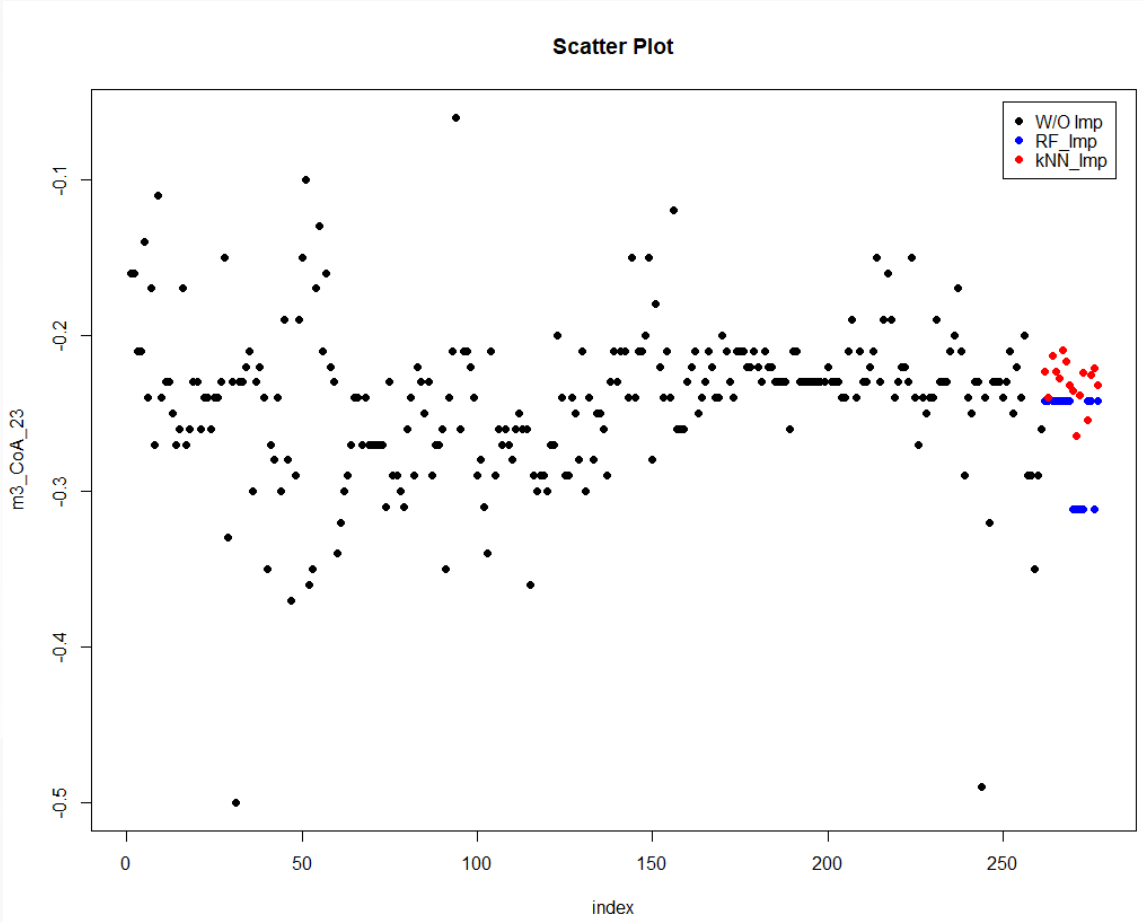
1. 只對 **Material3** 的 **CoA23** 建模插值, 其餘皆使用平均數插值

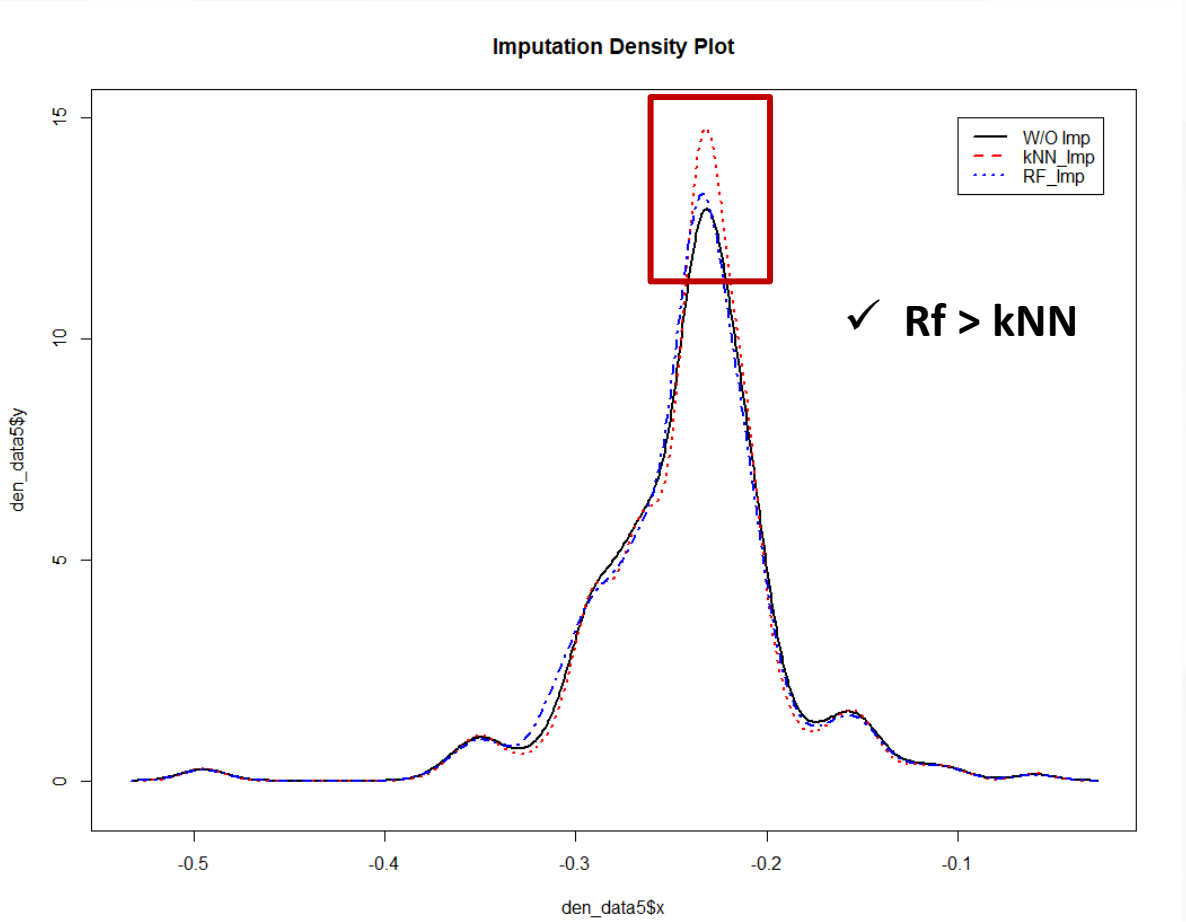
Red Line：y = 0.5



Missing Value Treatment

2. Material3: n = 277
   - 使用 kNN, Random Forest 做插值





Analysis Method

- **Problem1.**
  - Linear model
- **Problem2.**
  - Isolation Forest
  - Hierarchical Clustering
  - Multivariate SPC

Problem 1

- Forward selection via **Orthogonal Greedy Algorithm(OGA)**
- 因資料筆數不足，OGA只對變數的主效應挑選，並不考慮交乘效應
- 適配出的模型如下:
  - $\hat{y}_{inline1} = 47.78 - 3.11\boldsymbol{x_{CoA_3}} - 7.98\,\boldsymbol{x_{CoA_1}} + 1.18\,\boldsymbol{x_{CoA_7}}$
  - $\hat{y}_{inline2} = -39.56 - 0.22\,\boldsymbol{x_{CoA_8}} - 146.15\,\boldsymbol{x_{CoA_{13}}}$
- Cross Validation:
  - 模型固定下，給定某個cut point，leave one out做配適
  - 對所有的批號記錄預測結果正確與否，算出平均的accuracy和1-FDR
  - 找出令accuracy與1-FDR加總最高的cut point

Problem 1

Inline1

$R^2 = 0.941$	$\hat{\beta}$	$\hat{\beta}$ 標準差	T value	p-value
截距	47.7864	8.1809	5.841	0.001110
CoA_3	-3.1106	0.5166	-6.021	0.000947
CoA_1	-7.9880	1.5680	-5.094	0.002234
CoA_7	1.1816	0.4426	2.670	0.037048

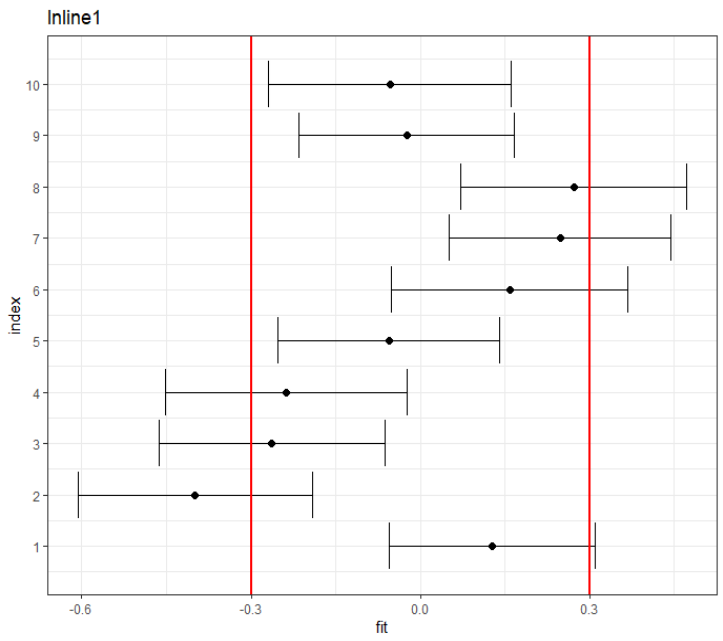
Inline2

$R^2 = 0.8803$	$\hat{\beta}$	$\hat{\beta}$ 標準差	T value	p-value
截距	-39.5618	10.5992	-3.733	0.00468
CoA_8	-0.2268	0.0285	-7.956	0.000231
CoA_{13}	-146.1529	51.4725	-2.839	0.01942

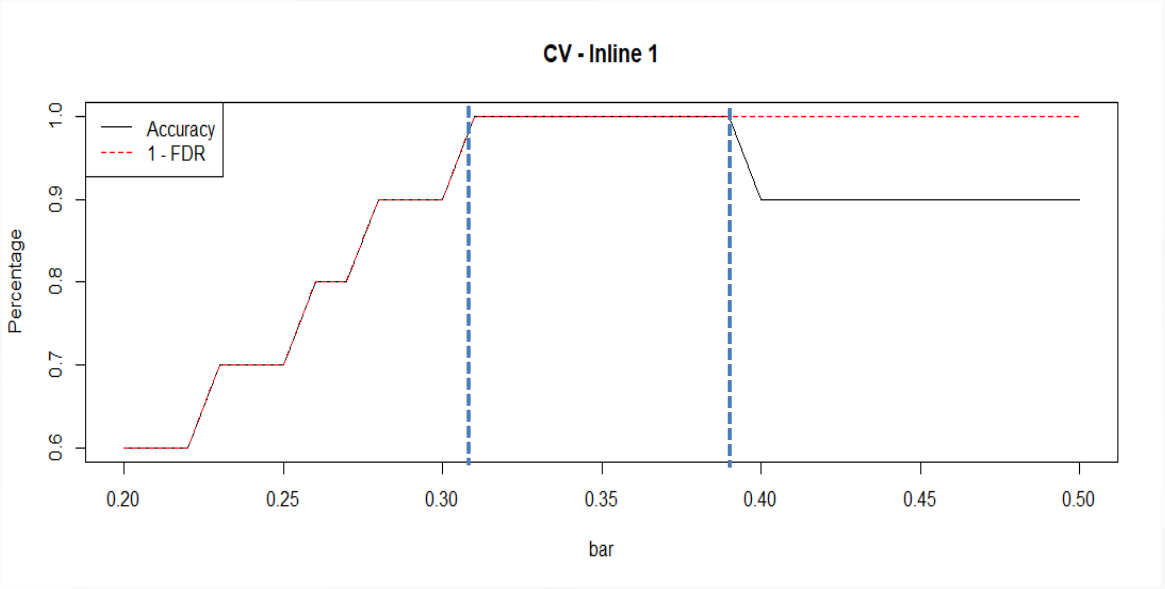
Problem 1-Linear model

Inline1

- 配適值及預測信賴區間



- 交叉驗證以最小化FDR及最大化正確率

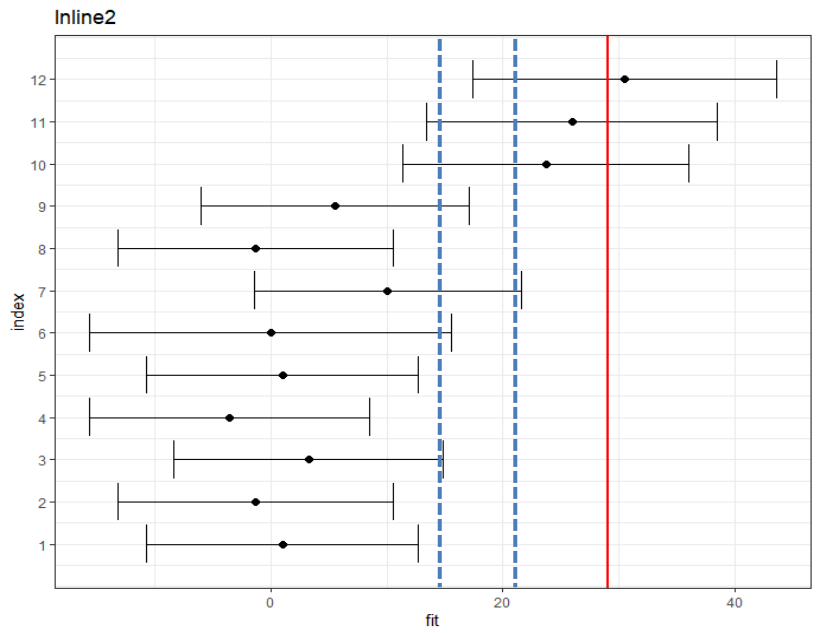


- 可以觀察到批號2的fitted value超過0.3 ，其他批號皆落在±0.3內
- 透過CV，把門檻定在±0.31~0.39 (藍色虛線) ，則可以完全區分異常批號

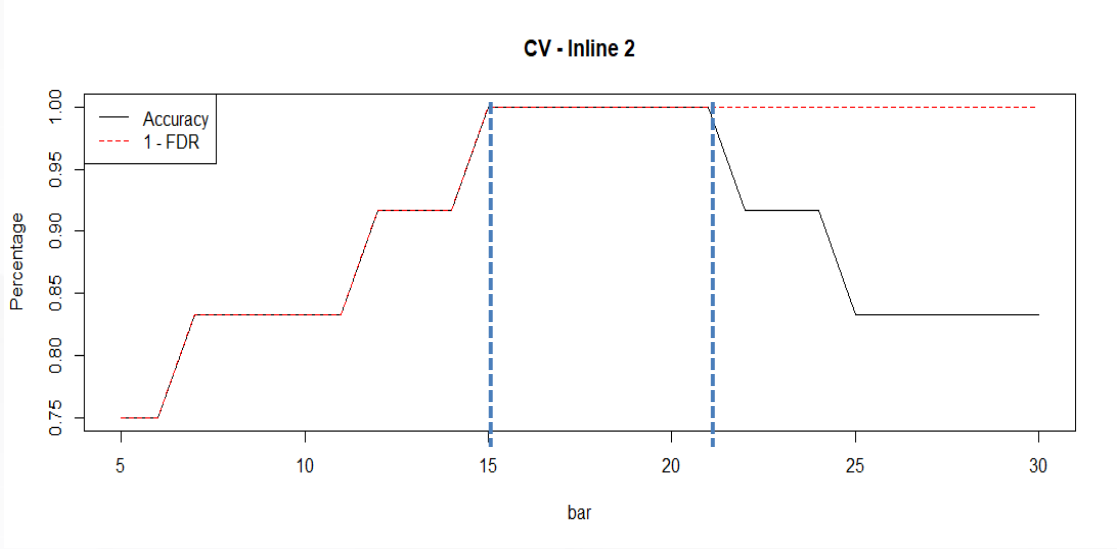
Problem 1-Linear model

Inline2

- 配適值及預測信賴區間



- 交叉驗證以最小化FDR及最大化正確率



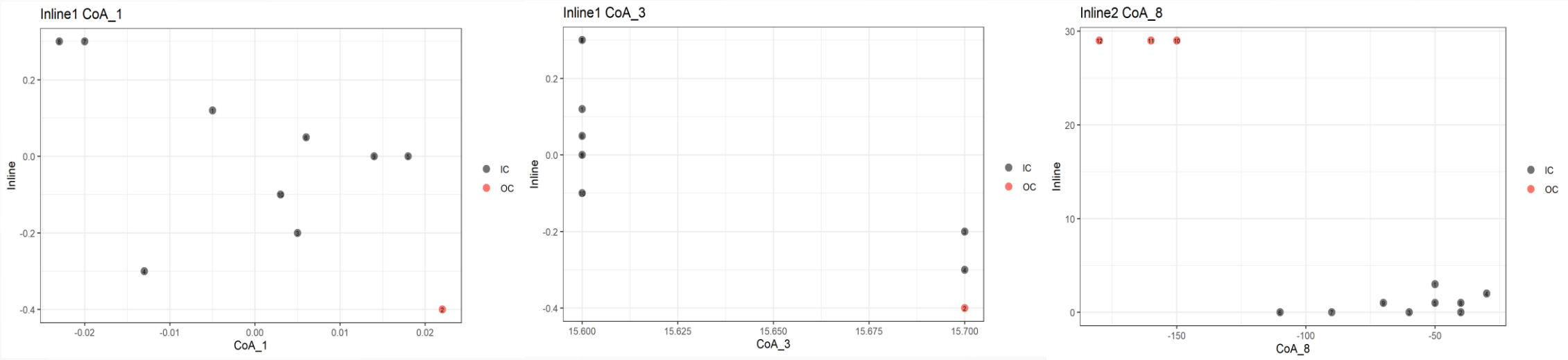
- 可以觀察到批號10、11、12的fitted value皆較其他批號大，紅線值為29
- 透過CV，把門檻定在15~21 (藍色虛線) ，則可以完全區分異常批號

Information Summary

Problem1

- 在資料探索時，Inline1的CoA_1和CoA_3，Inline2的CoA_8有線性的趨勢，且OGA也將其選出作為重要變數

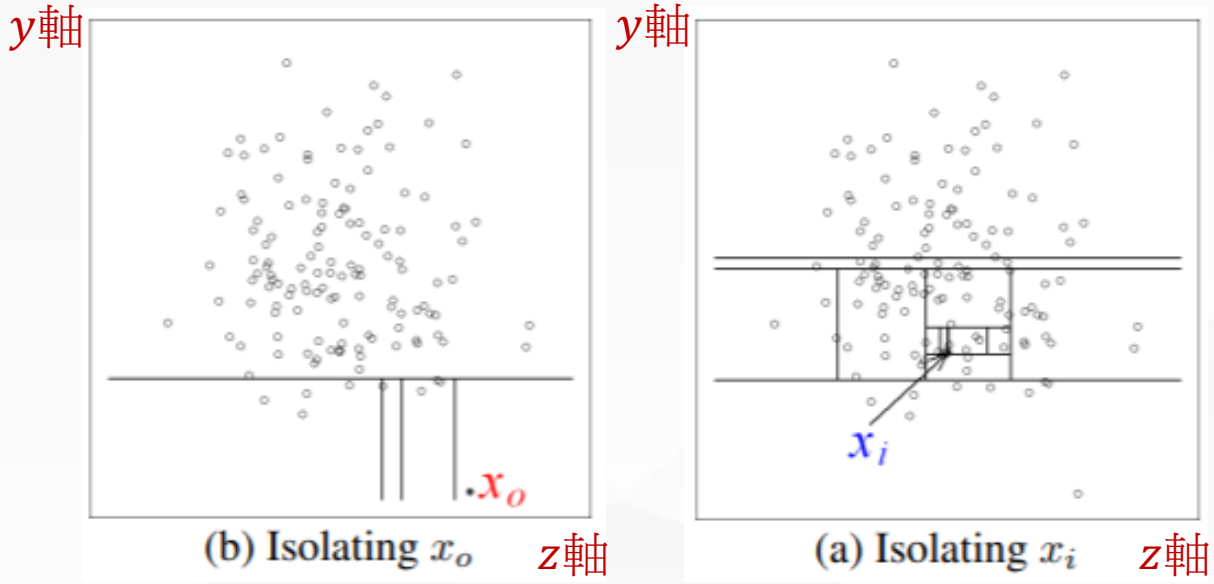
- 選擇這些解釋變數大致能解釋Inline的結果($\boldsymbol{R_1}^2 = 0.941, \boldsymbol{R_2}^2 = 0.8803$)



Problem 2-Isolation Forest

1. 方法假設：異常資料是**少數** $(10-20\%)$ 且**特別**的

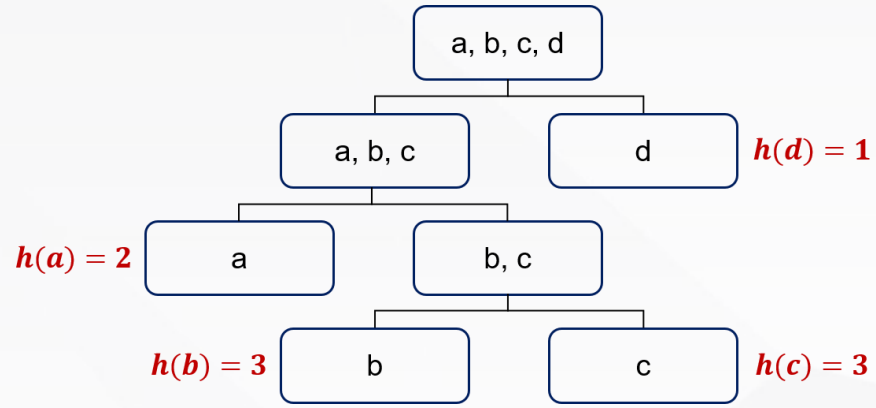
2. 想法：若一樣本 x **越早**被**切割**出去，代表 x 越可能為**異常點** (**不用計算距離**)



3. 作法：**隨機**從 p 維中選取一維，再從那個維度的全距中**隨機**選取一值做切分

Problem 2-Isolation Forest

4. 路徑長度：$\boldsymbol{h(x)}$。如下示意：



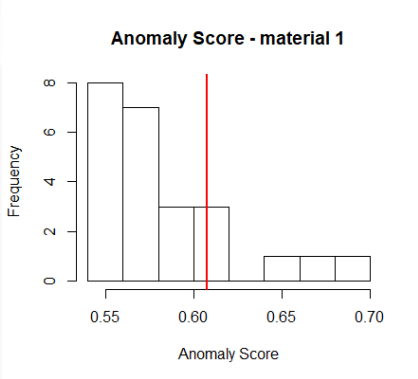
5. Bagging：取後放回種 m 棵樹，計算每個樣本點的平均 $\boldsymbol{h(x)}$

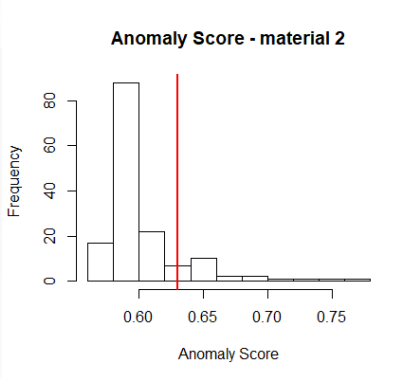
6. 異常分數：$\boldsymbol{s(x,n)} = 2^{-E(h(x))/c(n)}$ 其中 $c(n)$ 是給定n下的平均 $\boldsymbol{h(x)}$，即為對$h(x)$做正規化

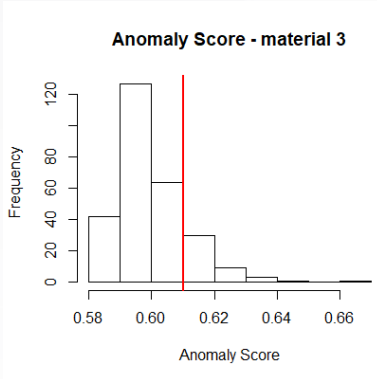
7. 決策法則：$s(x,n) \in [0,1]$ 越接近1越可能為異常，$\boldsymbol{s(x,n) > c}$, 其中 c 為切分點 (人為決定)

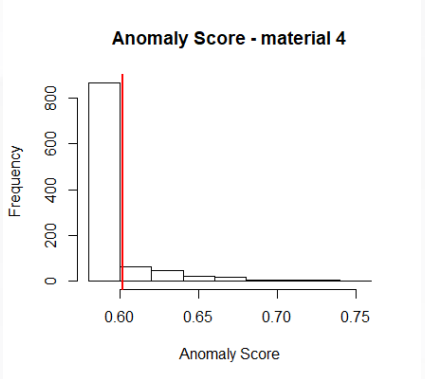
Problem 2-Isolation Forest

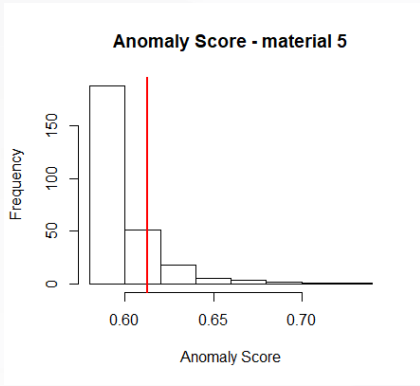
1. 以85分位點作為切分點











2. 最可能異常的批號 (95分位點)：

1 - (3, 4)

2 - (2, 5, 6, 3, 11, 94, 27, 13)

3 - (18, 39, 126, 10, 8, 125, 233, 117, 274, 1, 41, 124, 40, 215)

4 - (75, 91, 813, 812, 57, 55, 106, 89, 1, 3, 12, 143, 59, 95, 62,...)

5 - (9, 2, 11, 113, 10, 121, 207, 264, 12, 1, 205, 41, 112, 3)

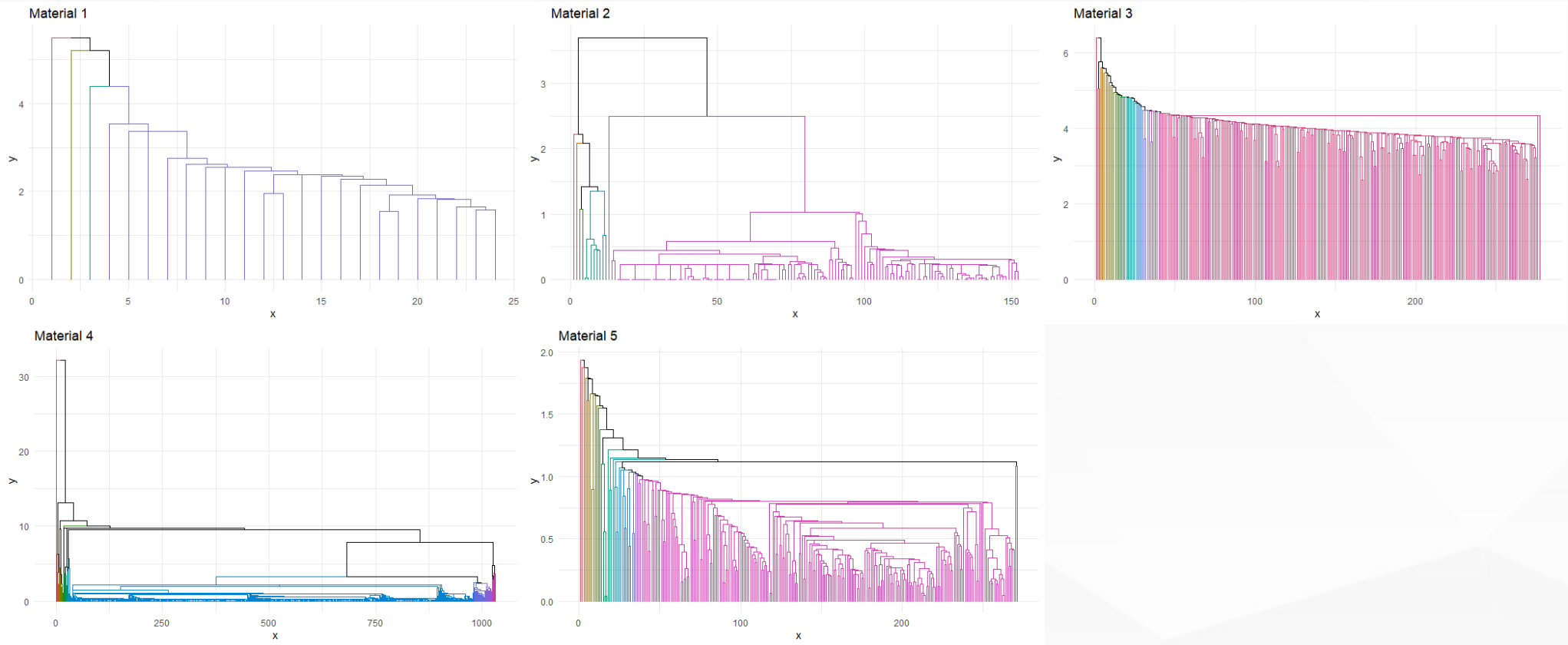
3. 參數使用：

	總樣本數	樣本數 / 樹	種樹量	樹的深度
Material1	24	24	100	4.58
Material2	152	128	100	7
Material3	277	256	100	8
Material4	1031	256	100	8
Material5	271	256	100	8

Problem 2-Hierarchical clustering

- 資料處理 : 標準化
- 距離 : 歐式距離
- 群聚算法 : single linkage
- Goal :小群的批號總數相加至多到15%

Problem 2-Hierarchical clustering



	m1	m2	m3	m4	m5
clusters	4	8	40	30	30
Anomalous %	12.50%	8.55%	14.80%	8.73%	15.13%

Problem 2-Multivariate SPC

- Multivariate SPC-$Hotelling's\ T^2\ Shewhart\ chart$:

$$T_i^2 = (X_i - \bar{X})'\left(\boldsymbol{S^2}\right)^{-1}(X_i - \bar{X})$$

$\bar{X}$ 為所有批號的CoA平均, $\boldsymbol{S^2}$ 為對應之共變異矩陣

- 定義p為CoA總數，M為批號總數量，則Control limit為:

$$Control\ limit = \frac{(M-1)^2}{M} Beta_{1-\alpha}(\frac{p}{2}, \frac{M-p-1}{2})$$

- 若$T_i^2 > Control\ limit$，則稱批號i為異常批號(OC)，其餘為正常批號(IC)

Problem 2-Multivariate SPC

新批號預測:

- 移除前一階段的OC資料後，針對新進批號X_{i+1}，同樣計算:

$$T_{i+1}^2 = (X_{i+1} - \overline{X'})' \left(\boldsymbol{S'^2}\right)^{-1} (X_{i+1} - \overline{X'})$$

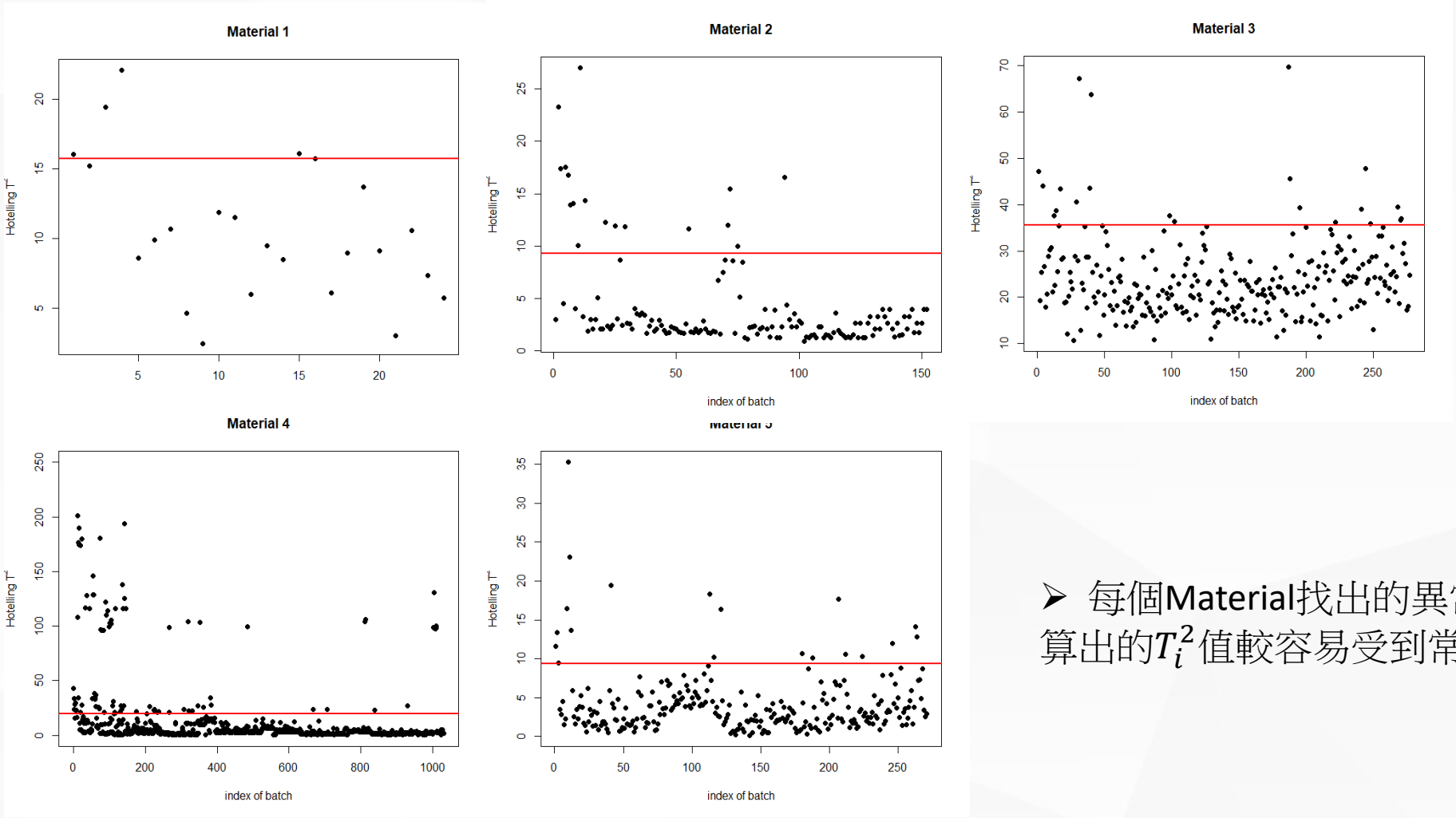
$\overline{X'}$ 為所有IC批號重新計算的CoA平均, $\boldsymbol{S'^2}$ 為對應之共變異矩陣

- 預測用的Control limit為:

$$Control\ limit = \frac{p(M-1)(M+1)}{(M-\mathrm{p})M} F_{1-\alpha}(\mathrm{p}, \mathrm{M}-\mathrm{p})$$

- 若$T_{i+1}^2 > Control\ limit$，則預測批號i+1為異常批號(OC)

Problem 2-Multivariate SPC



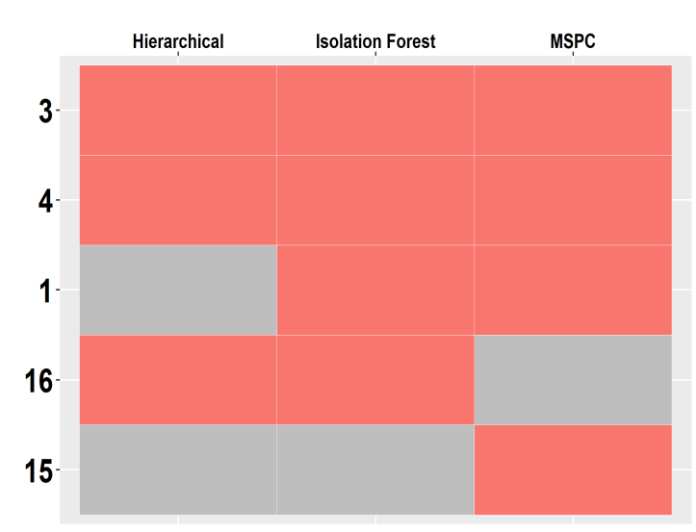
- 每個Material找出的異常批號比例都不相同
算出的T_i^2值較容易受到常數型CoA's的影響

Method Comparison

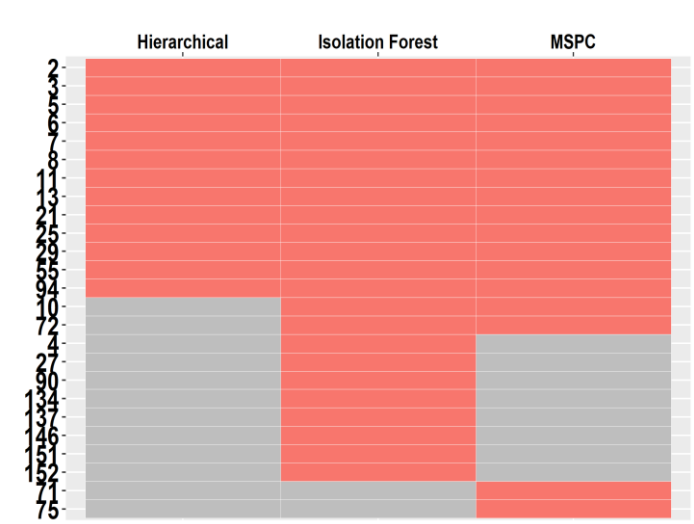
	Multivariate SPC	Isolation Forest	Hierarchical Clustering
耗時	短 勝	依據森林大小	長（任兩資料點都要算）
挑選兩側極端值	可 勝	可 勝	弱
單變量影響結果	嚴重	尚可 勝	尚可 勝
其他	變量分配不符下，Control Limit的power和type I error不準確	以單變量分割，而無法考慮多變量分割	將鄰近點圈入，未必具有偵測outliers的效果

Information Summary-各方法異常批次比較

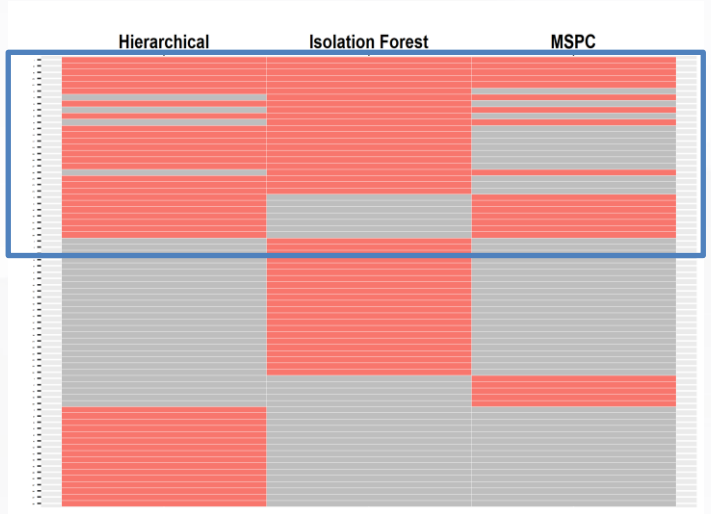
Material1



Material2



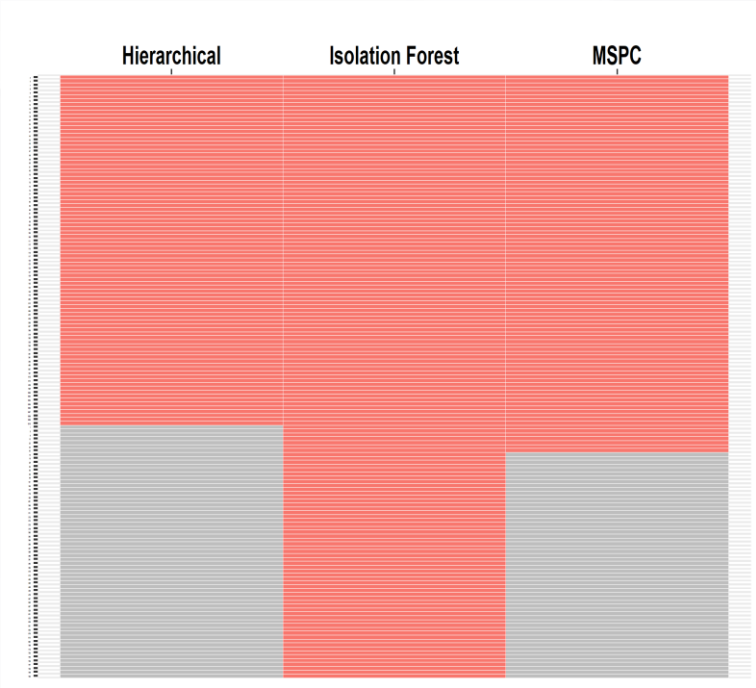
Material3



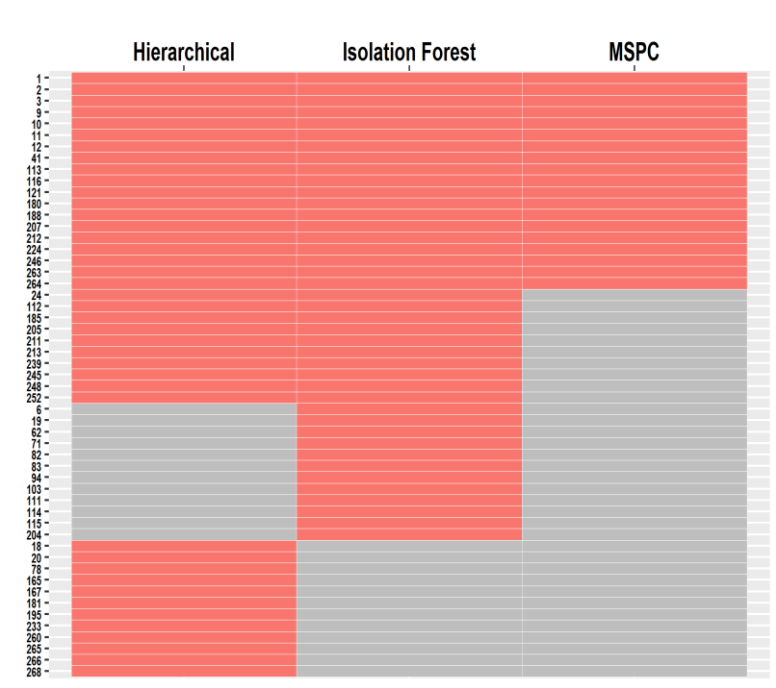
1. Material1批次較少，只有24個
2. Material2中偵測到的異常批號大致相同，Isolation Forest偵測到較多不同於其他方法的批次
3. Mateial3中，三方法偵測到的異常批號差異較大，僅有5個批號皆被偵測異常

Information Summary-各方法異常批次比較

Material4



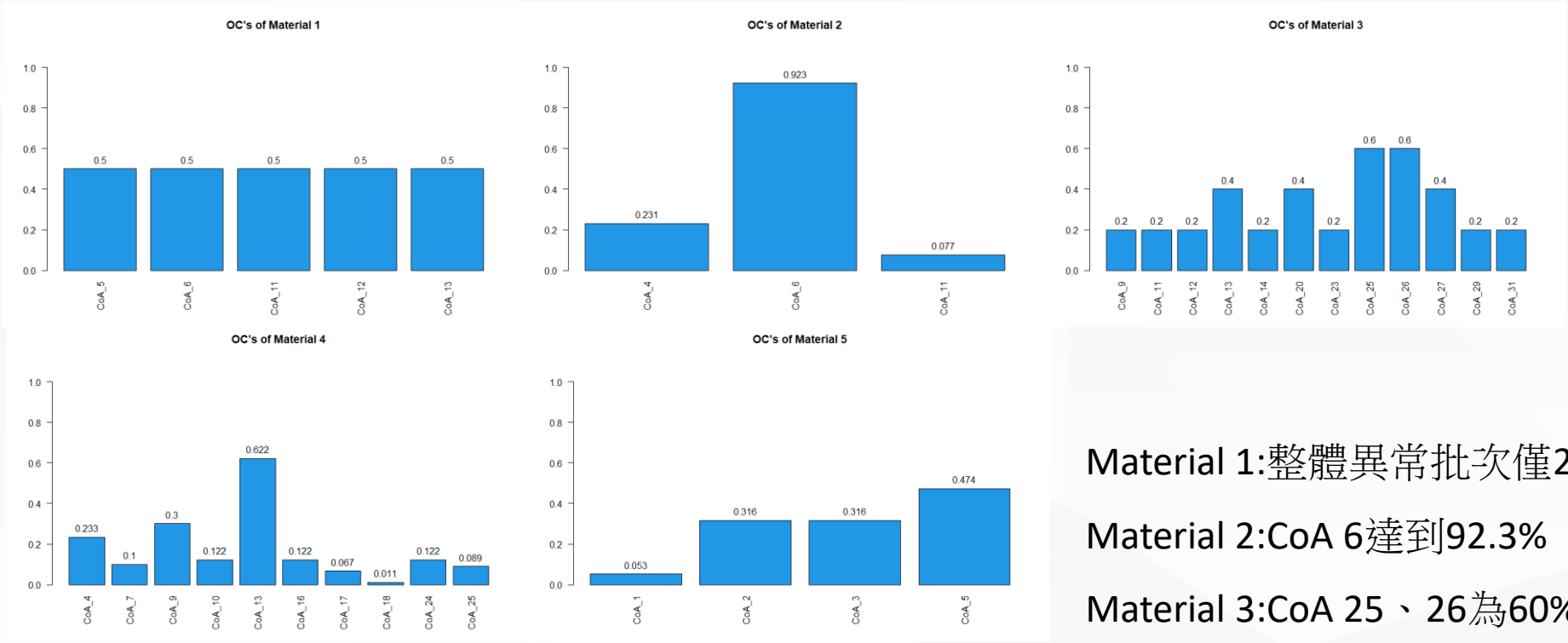
Material5



	Material1	Material2	Material3	Material4	Material5
Hierarchical	12.5%	8.55%	14.8%	8.73%	15.13%
Isolation Forest	16.67%	15.13%	15.88%	15.03%	15.13%
MSPC	16.67%	11.18%	7.58%	9.41%	7.01%
Intersection	8.33%	8.55%	1.81%	8.73%	7.01%

1. Material4中偵測到的異常批號大致相同，Isolation Forest偵測到較多不同於其他方法的批次
2. Mateial5中，MSPC選到的異常批次相對另外兩個方法少，另外兩個方法則有偵測到各別不同的批次
3. 把三個方法的異常批次取交集，定義為我們認為最有可能異常的批次

Information Summary-CoA Importance



圖片:
交集的異常批號，在各CoA上原本就是異常值的比例
比例高代表該CoA對Material整體分析的影響程度較大

Material 1:整體異常批次僅2筆，較無法判斷

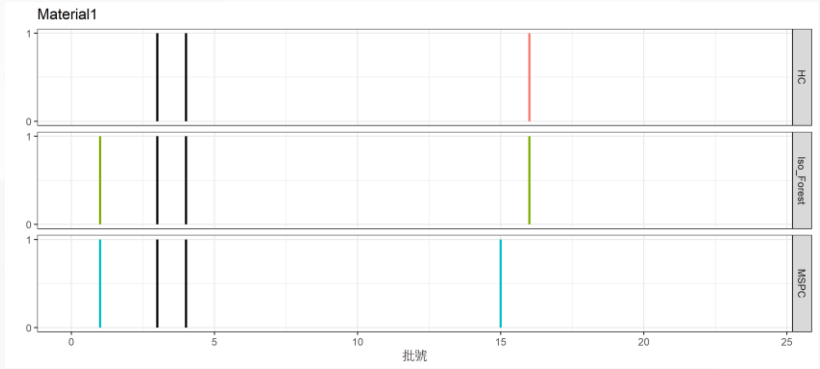
Material 2:CoA 6達到92.3%

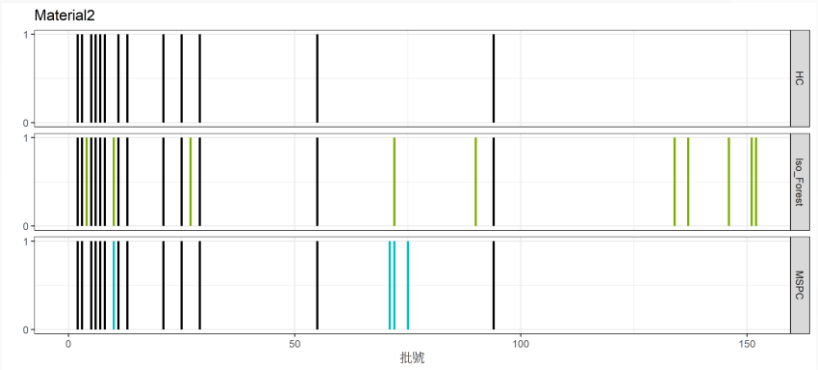
Material 3:CoA 25、26為60%，但整體異常批次僅5筆，較難判斷

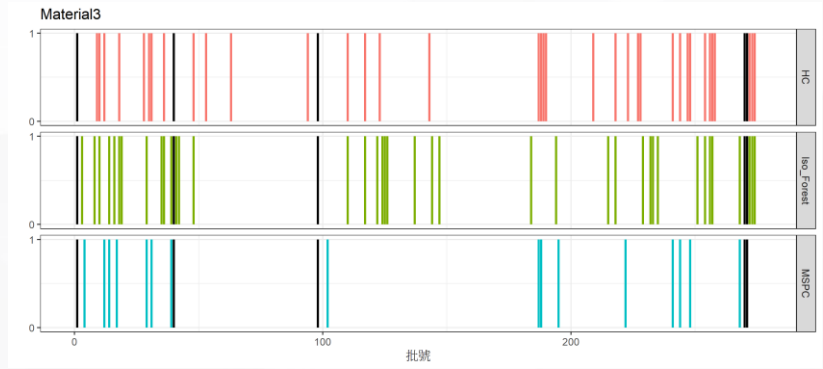
Material 4: CoA 13達到62.2%，數值較高

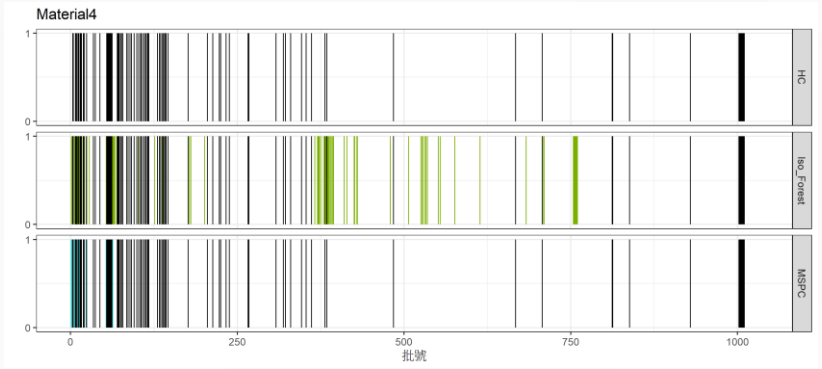
Material 5:CoA 2、3超過30%，CoA 5達到47.4%，明顯高於CoA 1

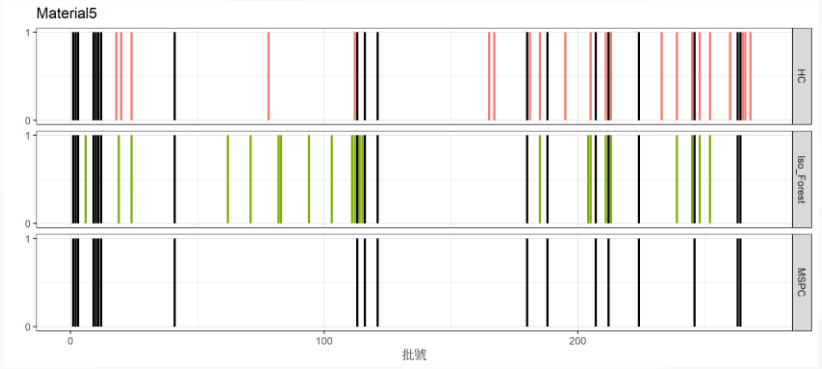
Information Summary-異常批次趨勢











- Hierarchical
- Isolation Forest
- MSPC
- Intersection

1. Material2、4、5， 三個方法找到的異常批次較接近
2. Material3三個方法找到的異常批次差異較大
3. Material2、4有較高比例的異常批次發生在前期

Decision making

- 未來資料預測:
  - Problem 1:
    - Regression : 預測response值後判斷是否異常
  - Problem 2:
    - Isolation forest : 將新資料代入模型計算Anomaly score
    - Hierarchical clustering : 將資料label後，再採用KNN分類
    - Shewhart chart : 移除異常批號後重新建立control chart，並針對新批號計算T^2值
    - 共同監控，例如任一組出現異常即判定為異常批次

Appendix

三種方法的異常批號取交集后列表:

Object	Abnormal data
Material 1	3,4
Material 2	2,3,5,6,7,8,11,13,21,25,29,55,94
Material 3	1,40,98,270,271
Material 4	3,4,7,8,9,11,12,14,15,16,19,20,24,34,37,44,54,55,56,57,58,59,60,61,62,70,71,72,73,75,77,78,84,86,89,91,95,99,102,105,106,109,111,113,115,116,117,130,133,135,137,139,140,142,143,146,176,205,213,223,225,233,238,266,267,308,319,322,330,346,353,361,381,384,484,667,707,812,813,838,929,1002,1003,1004,1005,1006,1007,1008,1009,1010
Material 5	1,2,3,9,10,11,12,41,113,116,121,180,188,207,212,224,246,263,264

THANKS !