國立清華大學教學發展中心學生讀書會

成果報告

學期：103 學年度第一學期

讀書會編號：10310G122

計畫主題：同好讀書會

讀書會名稱：資料分析與統計建模

小組成員：

陳雅汝，吳宗祐，吳周駿，王建勛

目錄

壹、 計畫簡介

一、 選讀書籍……3

二、 計劃目的與執行方式……4

三、 預期成效 ……6

貳、 讀書會內容

一、 第一次……7

二、 第二次 ……10

三、 第三次 ……13

四、 第四次 ……15

參、 個人心得感想

一、 陳雅汝心得 ……17

二、 吳宗祐心得 ……18

三、 吳周駿心得 ……19

四、 王建勛心得 ……20

肆、 評量

一、 陳雅汝評量 ……21

二、 吳宗祐評量 ……22

三、 吳周駿評量 ……23

四、 王建勛評量 ……24

伍、 對本校讀書會計畫的建議 …… 25

壹、 計畫簡介

小組名稱	資料分析與統計建模

小組成員資料

召集人姓名	系所	學號	身分
吳周駿	工學院 / 工業工程與工程管理學系	102034606	碩博生
組員姓名	**系所**	**學號**	**身分**
吳宗祐	理學院 / 統計學研究所	102024501	碩博生
王建勛	理學院 / 統計學研究所	103024514	碩博生
陳雅汝	理學院 / 統計學研究所	103024502	碩博生

計畫內容

一、計畫主題

同好讀書會

二、選讀書籍

書名	作者	出版社
An Introduction to Statistical Learning: with Applications in R	G. James, D. Witten, T. Hastie and R. Tibshirani.	Springer-Verlag.
The Elements of Statistical Learning	T. Hastie, R. Tibshirani and J. Friedman.	Springer-Verlag.
Data Mining for Business Intelligence	Galit Shmueli, Nitin R. Patel and Peter C. Bruce	Wiley & Sons

期刊名	出版頻率(次/年)
Artificial Intelligence	12
Expert Systems with Applications	18
Pattern Recognition Letters	16

三、計劃目的與執行方式

1.培養組員之資料分析能力，以順利參加第一屆半導體大數據分析競賽

重要性：

本組團隊將參加今年 10 月舉辦的大數據分析競賽，競賽以半導體製造程序中記錄的資料為主要分析對象，由於半導體元件必須經過近千道的製程，怎樣在這些快速蒐集而來且潛在共線性的變數中，快速找出產品良率與參數設定及機台配置之間的最佳模式，是將先進製程良率提高，並進一步迅速達成量產的關鍵。

執行方法：

從九月份開始，本組將優先練習資料準備部分，透過資料結構化、遺失值的補值、資料視覺化的呈現等來對資料的分佈有初步的認識，接著在初賽以前，本組會透過文獻的閱讀以及既有方法的比較找出適當的模型方法，並對於預測結果進行比較進而尋找出最適當的模型，在初賽到複賽這段時間，本組還會額外加強巨量資料庫管理與分析的能力，向業界導師學習 Hadoop 等分析工具，期盼能夠順利通過複賽為校爭光。

2.學習資料科學家應具備之專業知識與所需具備之分析能力

重要性：

資料科學家所應具備的能力屬於全面性統整能力，從一開始的資料庫處理、資料的敘述統計，到中期的模型選擇，以及最後的結果呈現與解釋都必須同時掌握許多能力，期望本組能夠融合統計所與工業工程所的專長，兼備理論與實務操作能力，對於學習的科目有高熟練度的掌控。

執行方法：

首先，我們會在指導老師-統計所鄭少為教授的指導下，學習正統的統計分析步驟，從資料準備開始，透過既有的方法下去進行資料整理，在分析的部分，本組會以軟體內建的方式或是選讀教科書建議的方式下去做初步的嘗試，把結果整理成表格和教授討論，其間遇到新穎的方法我們將會按照組員與各自的專長下去分配主題進行講解，最後本組將練習如何在假設聽眾沒有統計分析背景的基礎下，進行報告與分析步驟的解釋。

3.提升統計建模創新能力

重要性：

不管是多變量分析、線性模式，這些課程都提供很多實用的方法進行分析，然而這些既有模型背後都有自己的假設和使用的背景，值得注意的是這些方法往往不適用於過大的樣本數，特別是

這次的大數據分析，除了融合連續和間斷型資料外也要同時分析多個維度的資料，因此，本組希望在既有理論的基礎下，進一步修正出可以適用於這次競賽的新型態模型。

執行方法：

除了前面提到學習常見的分析工具外，我們會透過近期文獻的閱讀來探究現在有哪些新的方法可以適用這次比賽的資料，此外，組員們將會配合這學期所學習的課程加強方法論的部分，在創新建模部分，我們會特別加強在統計學習理論方面的學習，不管是監督式學習或者是降維度的方法我們都會涉獵，同時在類神經網路的部分，我們會透過與教授的討論想辦法想出可以解釋的地方。

4.提升自我充實之能力

重要性：

此次參加競賽，同時配合讀書會進度，還要同時進行修課與研究，時間管理和良好的自我控制能力是非常重要的，而要通過這趟學習之旅，除了有導師的帶領和組員間的配合外，還要培養自我充實的能力，組員本身要能夠有自我學習的能力，才能夠自我提升也才能教導其他隊員。

執行方法：

在選讀書目的章節部分，本組會透過輪流的方式來帶領其他組員選讀該主題，此外，也要提供自己額外找的資料以及文獻來進行討論，除了每週規定的進度外，我們也會透過鼓勵的方式希望組員可以自主提供額外的想法進行討論，由於這次閱讀量相當龐大，本讀書會也會透過類似翻轉教室的形式，希望組員們可以在讀書會開始前都對該週的教學有一定的認識，聚會時間會以討論為主。

5.提升與業界導師互動之能力

重要性：

這次競賽的兩大困難點，首先是方法上要配合資料，而且隨者競賽的演進相信難度會越來越加深，再來是對於資料本身的解釋，由於本組成員都是統計相關背景，對於半導體製程不像電機系同學那樣熟悉，這時侯業師就扮演著溝通的橋樑，透過與業師的互動，相信我們可以對於分析更添加實用性和真實性。

執行方法：

首先，我們會透過邀請在半導體先關產業工作的學長姐回來，對於工作內容或是資料本身進行簡短的講解，同時，我們也會透過寫信或是參觀的方式對於這次競賽的主辦方進行互動，最重要的

一點是，我們要透過這些媒介了解主辦方真正想要的東西，並且把我們的發現完整地講解給他們知道。

四、單元的內容與活動方式

日期	時間	地點	預定導讀人	進行討論內容
2014-10-24 五	13:00:00 至 16:00:00	綜合三館八樓 840 教室	張智翔	資料處理
2014-11-07 五	13:00:00 至 16:00:00	綜合三館八樓 840 教室	張智翔	變數選擇
2014-11-28 五	13:00:00 至 16:00:00	綜合三館八樓 840 教室	張智翔	統計建模
2014-12-05 五	13:00:00 至 16:00:00	綜合三館八樓 840 教室	張智翔	解釋與評估

五、導讀人基本資料

姓名	服務單位/職稱
張智翔	國立清華大學統計學研究所/研究生

六、預期成效

希望經由一學期讀書會的訓練，每位組員都能夠達成預期目標。在讀書會結束後，每位參與讀書會的成員將評比上述 5 項目標實際達成的分數，同時，也會請導師進行評分，整理之後我們會將評比結果呈現於最後成果報告，透過量化結果檢視讀書會對於每位成員所帶來的實際幫助。

在培養組員之資料分析能力，以順利參加第一屆半導體大數據分析競賽部分，最主要是希望本組可以順利進入決賽，為學校爭取榮譽。在學習資料科學家應具備之專業知識與所需具備之分析能力方面，希望透過這次學習之旅，開啟組員進入資料科學家這個領域的大門，繼續學習朝這個目標發展。在提升統計建模創新能力部分，除了需熟練既有的方法外，還要想辦法熟悉前端研究，並且把相關方法應用到資料當中。在提升自我充實之能力方面，希望提升時間管理以及自我學習的能力，同時要兼顧研究生的本分可以兼顧研究和競賽的進行。最後，在提升與業界導師互動之能力部分，最主要是希望組員們除了研究外也要提升跨領域、產業的互動能力，可以把自己的專業和別人交流。

貳、 讀書會內容

資料分析與統計建模 第一次讀書會：資料處理

<table>
<tr><td colspan="3">時間：103 年 10 月 24 日（星期五）下午 13:00 至 16:00</td></tr>
<tr><td>地點：綜合三館 8 樓 840</td><td>導讀人：張智翔</td><td>召集人：吳周駿</td></tr>
<tr><td colspan="3">成員：陳雅汝、吳宗祐、吳周駿、王建勛</td></tr>
<tr><td colspan="2">報告人：陳雅汝</td><td>記錄人：吳宗祐</td></tr>
<tr><td colspan="3">報告主題：資料分析與統計建模-資料處理</td></tr>
<tr><td colspan="3">主題內容：
當統計分析人員拿到一筆資料時，所做的第一件事並不是立即將資料丟進模型，觀察其分析結果，而是要先了解資料型態，知道各個變數的意義，並藉由數值方法與圖形方法，觀察資料的平均值、變異數、分布型態、是否存在有共線性、變數內是否保有什麼特殊 pattern 等等，從這個步驟開始下手，與資料對話，挖掘隱藏在資料背後的寶藏，以提供接下來做模型時使用。
但真實的資料並不是我們所想樣的這麼完美，通常都存在有遺失值，只是多或少的差別。若是遺失值的個數在總資料內占的比例較少，一般情況下會選擇直接刪去這些遺失值；但如果遺失值的個數在總資料內占的比例較多，就要思考應該用什麼樣的方法才能在不改變資料想說的話的條件下，成功補值。例如我們這次的半導體製程練習資料，在補值時用了兩種方法，其一是了解半導體製程的背景，因為半導體製程在相同批量下，所使用的工具以及參數設定的值可能是相同的，利用這項背景知識對這些變數進行補值；而其他變數因為比較沒有上述特性，因此我們決定取該變數相同批次時使用其他方法進行補值。
上述對資料的觀察都是侷限於 X 與 X 之間的關係，但在接下來做模型預測時，我們關心的並非 X，而是如何用 X 去預測 Y，因此在這個步驟我們仍須觀察 X 與 Y 之間的關係。在做預測時，testing data 與 training data 兩者間的內插、外插關係尤其重要，若是 testing data 中，外插太多，則模型中 X 的共線性就變成了一個不可忽略的特性，且模型不宜建構得太複雜；反之，則可以忽略 X 的共線性(因為我們練習的這組資料只注重 Y 值的預測，因此不論模型是否有重複考慮到性質相似的 X，只要預測的準都無所謂)，且在內插較多的情況下，即使使用了較複雜的模型，也不易導致 overfitting 的情況產生。
在理解了內插與外插對於接下來的分析的重要性後，我們總共想了 3 種方法去分析內插與外插，其一是在多變量分析課程裡學到的距離矩陣；其二是老師提供的類別行距離計算；其三是觀察 Y 的區間範圍。最後我們選擇使用類別型的距離來做內插、外插的分析。
所使用的方法主要是在計算類別型資料的距離，相同類別時距離設為 0，不同類別時距離設為 1，任兩個樣本下的距離總和即為此兩個樣本之間距離。將練習資料丟進 R 裡做出距離矩陣，並用這此距離矩陣畫出圖型。</td></tr>
</table>

從圖形中可以發現，黑色(training data)的點都集中在左邊，而紅色(testing data)的點則分為左右兩邊，很明顯的可以看出，在還沒補 NA 值之前，這筆練習資料的外插(右邊紅色的那群)相當嚴重；而圖二則顯示，在補完 NA 值之後，外插消失了。除了看內插、外插外，還可以從圖二看出，資料很明顯的被分為兩群，因此我們也可以將這項發現利用在之後的建模上，例如直接把資料分兩群，分別對這兩群資料做模型預測。

上述的所有步驟在統計分析上統稱為 initial data analysis[1]，主要目的是找出有代表性的數據，以提供接下來的變數選擇使用。

心得感想：

在做分析這筆練習資料前，我其實也曾經做過資料分析，但是當時並沒有很著重於 initial data analysis 這一塊，反而只在乎模型該如何建立，總是在不斷的嘗試各種模型，最後從各種嘗試的模型中挑出最好的模型呈現，但是這一次的讀書會，藉由指導老師鄭少為的帶領，讓我了解 initial data analysis 的重要性，必須先了解數據才能更深入的探索他所蘊藏的寶藏，這就如同交朋友的道理，必須先了解每個人的個性，才能更深入地進行心靈層面的交流，進而變成知己。

指導老師在這一次的讀書會中，帶領我們做 initial data analysis，不斷的提供我們意見、對我們的分析方法提出質疑、幫助我們釐清問題、了解統計分析的精髓、教我們如何用數學方式定義問題等等，這些都讓我們受益良多，而我也在這一次的讀書會中，學習到了新的統計知識－漢明距離。

整個讀書會的過程就是四個組員與指導老師之間互相提問，利用各自的背景知識替彼此解決問題，這種討論方式是我以前沒接觸過的，而我也因此成長了許多，感謝這次的讀書會提供我這麼好的平台學習統計分析。

下次進度：資料分析與統計建模-變數選擇

參考資料：

[1] 清華大學 STAT6910 統計實習 Basic Procedure of Statistic 上課影片
http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat6910_2/index.html
[2] Applied Multivariate Statistical Analysis Chapter12.6
Multidimensional scaling (RICHARD A.JOHNSON, DEAN W.WICHERN)
[3] Hammong distance
http://www.maths.manchester.ac.uk/~pas/code/notes/part2.pdf

其他：

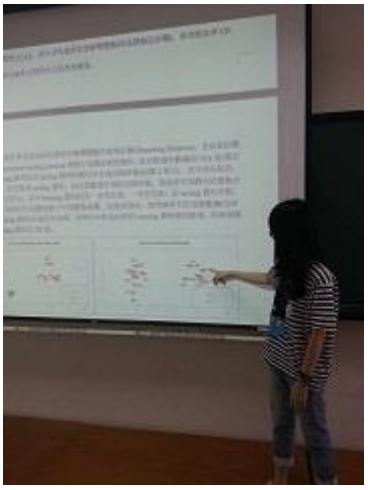
閱讀章節：

An Introduction to Statistical Learning: with Applications in R G. James, D. Witten, T. Hastie and R. Tibshirani. Springer-Verlag. **Chapter 1 to Chapter 2**

The Elements of Statistical Learning T. Hastie, R. Tibshirani and J. Friedman. Springer-Verlag. **Chapter 1 to Chapter 2**
Data Mining for Business Intelligence Galit Shmueli, Nitin R. Patel and Peter C. Bruce Wiley & Sons **Chapter 1 to Chapter 3**

讀書會照片：





資料分析與統計建模 第二次讀書會：變數選擇

<table>
<tr><td colspan="3">時間：103 年 11 月 7 日（星期五）下午 13:00 至 16:00</td></tr>
<tr><td>地點：綜合三館 8 樓 840</td><td>導讀人：
張智翔</td><td>召集人：吳周駿</td></tr>
<tr><td colspan="3">成員：陳雅汝、吳宗祐、吳周駿、王建勛</td></tr>
<tr><td colspan="2">報告人：吳宗祐</td><td>記錄人：吳周駿</td></tr>
<tr><td colspan="3">報告主題：資料分析與統計建模-變數選擇</td></tr>
<tr><td colspan="3">主題內容：

這次分析的資料是有關於一百個站點的紀錄，每個站點都有提供參數設定等資訊，其中在不同的站點相同的變數所代表意義也不同，意思是同樣是 Tool1，站點一和站點二所代表的 Tool 是不一樣的。除此之外，同一個站點裡面的 Chamber 和 Carrier 也會隨著不同 Tool 有所差異，例如:站點 1 中某個 Lot 使用 Tool1 和 Tool2，同時都使用 Chamber1 的話，這情況的 Chamber1 其實不是指同一個東西，所以變數之間是有所謂 nested(巢狀)結構。

由於變數眾多的情況下，很容易就出現共線性的情形，再加上從之前的報告中，可以發現資料有兩群的結構，需要避免不適合外插的模型，共線性的挑選變得更加重要，故先採用類別型變數相關性計算的方法，發現到具有完全共線性的情形發生，所以最後我們採用刪取這些變數完全共線姓的變數。

變數之間都是類別型的結構，加上資料本身數量不多，無法把所有變數同時放入很多種模型，然後使用 AIC 或者是 R 平方當作篩選的依據，選出具有代表性的變數。篩選變數在這種情況下，就會變得困難許多。因此讓篩選出來的變數，有較強的公信力，我們採取了用多種模型方法來選擇代表性變數，也就是挑選出來的變數會在這些方法中都是具有重要的。

在此我們選擇的模型來挑選變數的方法有，例如以下的方法。

<table>
<tr><td>線性迴歸</td><td>由於這個方法的缺點是變數維度不能超過資料個數，因此我們將不同站點的變數，分開來做模型，故一百個站點就有一百個模型，其中變數之間的交互作用也會進入模型當中，採取逐步迴歸(stepwise)的方法，以及觀察 R 平方比較大的模型，來挑選具有代表性的站點以及其變數。</td></tr>
<tr><td>線性迴歸樹</td><td>這個方法變數維度可以超過資料個數，把所有變數的主效應都丟進模型，觀察跑出來得模型中，哪些變數會優先當作考慮，代表這些變數影響力較大。</td></tr>
<tr><td>隨機森林</td><td>這個方法的挑選方式同線性迴歸樹。</td></tr>
</table>
</td></tr>
</table>

除了使用模型來決定變數外，還要把觀察到的變數額外做一些數值(單獨把那些變數作 ANOVA，觀察其 AIC 和 R 平方)和圖形(依據該變數的不同 level 畫 boxplot)方法的呈現，這種對話形式的觀察資料，有助於我們了解資料，挑出代表性變數，以及之後模型的建構。

心得感想：

以往在解決共線性問題的時候，都是遇到連續型的變數，可以直接使用相關係數的方法來觀察，但這次的資料中，所有的變數皆為離散型的變數，故不能使用相關係數的方法，在此有考慮過是否要使用「熵」指標，但最後選擇老師提供的其他方法來觀察共線性，因為在討論共線性的問題，類別行的方法有較好的解釋意義以及方法是更簡單明瞭。

模型選擇變數的方法中，使用的模型也都是以前沒使用過的，因此在使用上以及分析模型結果，都是需要花時間去思考以及查相關資料，在這邊有蠻多的想法會產生，例如預估模型的建構思維差異，以及不同模型挑出不同變數時，該如何解釋其中差異。

下次進度：資料分析與統計建模-統計建模

參考資料：

[1] Canonical correlation analysis 相關文獻

http://download.springer.com/static/pdf/825/chp%253A10.1007%252F978-3-540-72244-1_14.pdf?auth66=1415797322_93163d4a7bdb3a4bab8e429552059fbf&ext=.pdf

[2] Regression tree 相關文獻

http://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf

[3]Random forest 相關文獻

http://www.biomedcentral.com/content/pdf/1471-2105-7-3.pdf

其他：

閱讀章節：

An Introduction to Statistical Learning: with Applications in R G. James, D. Witten, T. Hastie and R. Tibshirani. Springer-Verlag. **Chapter 3 to Chapter 7**

The Elements of Statistical Learning T. Hastie, R. Tibshirani and J. Friedman. Springer-Verlag. **Chapter 3 to Chapter 7**

Data Mining for Business Intelligence Galit Shmueli, Nitin R. Patel and Peter C. Bruce Wiley & Sons **Chapter 4 to Chapter 5**

讀書會照片：

資料分析與統計建模 第三次讀書會：統計建模

時間：103 年 11 月 28 日（星期五）下午 13:00 至 16:00		
地點：綜合三館 8 樓 840	導讀人：張智翔	召集人：吳周駿
成員：陳雅汝、吳宗祐、吳周駿、王建勛		
報告人：吳周駿		記錄人：王建勛
報告主題：資料分析與統計建模-統計建模		

主題內容：

在統計建模部分，首先要決定模型的目的是要進行解釋還是進行預測，如果目的是要進行解釋的話，可能平衡的指標就會定義在 R 平方越大越好，還有所建模的資料 Training Data 也要盡可能的多，如果目的在於預測，那可能就必須透過交互驗證的方式挑選適當的模型，可能的原因有所建的模型比較偏好所使用的資料進而導致模型的準確度有偏差，因此模型的目的會導致所使用的方法不同。

再者，自變數和應變數的關係是否可以解釋也是一大重點，典型的例子如線性迴歸和類神經網路。線性迴歸是典型預測因子和反應有清楚關係的例子，應變數可以寫成由自變數所組成的數學函數，而類神經網路就如同一個黑盒子，使用者不能解釋變數與反應之間的關係，對於得到的預測值也沒有辦法有比較客觀的評估，不過優點是在大部分的情況下預測結果的誤差較小，因此使用者必須在誤差和解釋能力上進行權衡以及取捨。

不可否認的是變數挑選有時候和建模是同步進行的，例如在變數太多的情況下可以透過逐步迴歸的方式同時進行變數挑選以及建模，而使用者可做的是設定一個可靠的進入門檻以及停止機制，此外，在一些機器學習的方法中，有時候需要使用者自行挑選一些變數，這時候就必須參考專家知識，或者是背景知識來幫助初步挑選可能的變數。

在模型的挑選與比較上，常見的方法是建立數個模型再透過誤差評比的方式來挑選一個較佳的模型，而在模型的解釋方面在下次的讀書會會有所討論。

心得感想：

在還沒學習資料挖礦之前，我一直不知道解釋和預測的差別，我以為模型解釋能力越高預測能力就愈好，完全不知道有偏誤的可能性存在，在交互驗證方面，以前只知道要分成訓練資料以及測試資料，現在可以用雙向的方式進行模型的比較與選擇。在建模方面，一直以來都是利用傳統的統計方法例如線性迴歸的方式來進行預測，然而遇到的資料變化越來越大，以前熟悉的方法碰到不同的資料並不能直接套用以前的模式進行分析，必須對資料進行轉變又或者是要透過更進階的方式來建立適當的模型，例如這次碰到的半導體製程資料就和之前碰過的資料完全不一樣，在一開始的時候很難想到要如何處理，但是透過文獻的閱讀以及教授的提點，並搭配新方法的學習，分析的情況也越來越進入軌道，而在統計學習這一塊還有很多方法還沒辦法掌握，特別像是參加預測模型競賽，常常評斷的標準是預測的錯誤率，傳

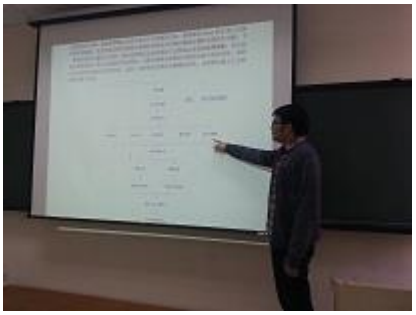
統統計手法可能分析流程以及假設的建立很扎實，但是預測結果就是沒有比類神經網路之類的黑盒子方法好，因此在模型的結果評估方面還要多下一點功夫。

下次進度：資料分析與統計建模-解釋與評估

參考資料：
多變量分析（陳順宇 著）
迴歸分析（陳順宇 著）
類神經網路：MATLAB 的應用（羅華強 著）

其他：
閱讀章節：
An Introduction to Statistical Learning: with Applications in R G. James, D. Witten, T. Hastie and R. Tibshirani. Springer-Verlag. **Chapter 8 to Chapter 11**
The Elements of Statistical Learning T. Hastie, R. Tibshirani and J. Friedman. Springer-Verlag. **Chapter 8 to Chapter 11**
Data Mining for Business Intelligence Galit Shmueli, Nitin R. Patel and Peter C. Bruce Wiley & Sons **Chapter 6 to Chapter 11**

讀書會照片：



資料分析與統計建模 第四次讀書會：解釋與評估

<table>
<tr><td colspan="3">時間：103 年 12 月 5 日（星期五）下午 13:00 至 16:00</td></tr>
<tr><td>地點：綜合三館 8 樓 840</td><td>導讀人：張智翔</td><td>召集人：吳周駿</td></tr>
<tr><td colspan="3">成員：陳雅汝、吳宗祐、吳周駿、王建勛</td></tr>
<tr><td colspan="2">報告人：王建勛</td><td>記錄人：陳雅汝</td></tr>
<tr><td colspan="3">報告主題：資料分析與統計建模-解釋與評估</td></tr>
<tr><td colspan="3">主題內容：
前幾次的討論主要針對半導體製造的品質管理去進行產品缺失的原因分析，並將資料分析的流程導入產品異常的分析手法當中，目的是要找出可能影響產品良率的潛在變數，以協助未來產品良率的預測。從分析結果發現，半導體製程的機台和腔室之間的交互作用可能會影響到最後的良率，因此在交互作用的處理上就顯得格外重要，此外良率的表現也有可能在時間上有特殊的規律，因此也要加入時間變數到模型當中進行討論。
在解釋的部分可以根據實際的分析結果，挖掘出產品發生問題的背後原因，並輔助決策者提出具體的改善措施。因此資料分析還要和其他擁有專業知識的人員進行搭配，才可以清楚抓出關鍵因子。在自動化偵測的建置方面，未來可以考慮更長時間的資料投入去建立自動化的機制。
在建模方面，可以嘗試對不良品再進行分類並建立決策樹模型或是 SVM（Support Vector Machine）二元分類法再進行預測與模型比較。在預測方面，則可考慮建立多個配適模型，並以其預測值之加權平均來降低 overfitting 的風險。在納入資料方面，可以加入機台維護紀錄來當作未來預測之修正。</td></tr>
<tr><td colspan="3">心得感想：
接續上一次讀書會討論統計建模的方法之後，本次讀書會著重在對於模型的預測的解釋能力做評估，我們發現在資料量很大的情況下，光以單一的統計模型來預測可能太過樂觀，若採取建立多個統計模型之後在將預測值取平均可能會較是一個保守穩健的方式，在解釋評估上的也不能單就程式的得到的結果套用在測試資料上，可能還要具備該領域的知識，考慮所得的預測值是否合理。
此外本次讀書會我們也邀請了業界導師 Sam 加入，主要和我們分享在面對大數據的情況，如何運用資料庫篩選我們所需要的資訊，再做統計分析，因為傳統的資料量並不大，一般電腦都還能夠負荷，但是當資料量大到一個程度，程式跑一跑可能就會當機，如果能妥善使用資料庫軟體，將能夠縮短建模和分析的時間，對於本次大數據比賽非常有幫助。
在這樣討論的過程中我們也發現，一個資料科學團隊不能夠只有統計背景的人，是需要一個跨領域的整合，很難一個人懂資工又懂統計又能夠了解該領域的專業知識，所以溝通能力在這個過程也扮演了很重要的角色，我相信經過這四次的讀書會對於每一位成員在學習大數據的分析方法之外，也能夠體認多元工作環境的未來性。</td></tr>
</table>

參考資料：

多變量分析（陳順宇 著）

迴歸分析（陳順宇 著）

類神經網路：MATLAB 的應用（羅華強 著）

其他：

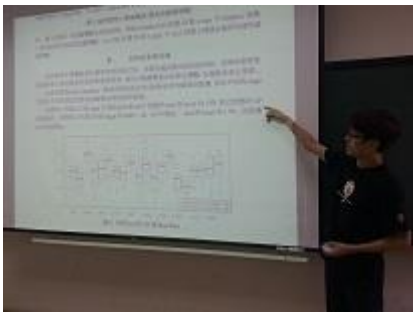
閱讀章節：

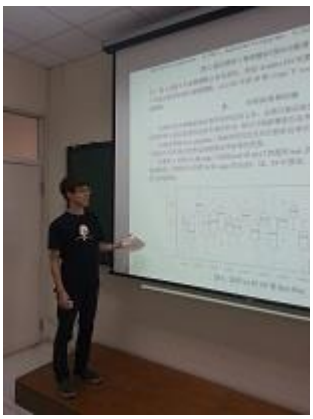
An Introduction to Statistical Learning: with Applications in R G. James, D. Witten, T. Hastie and R. Tibshirani. Springer-Verlag. Chapter 1 to Chapter 11

The Elements of Statistical Learning T. Hastie, R. Tibshirani and J. Friedman. Springer-Verlag. Chapter 1 to Chapter 11

Data Mining for Business Intelligence Galit Shmueli, Nitin R. Patel and Peter C. Bruce Wiley & Sons Chapter 1 to Chapter 11

讀書會照片：





參、 個人心得感想

陳雅汝心得：

大學四年也是在清華大學就讀，但一直沒有機會參與讀書會相關的活動，直到研究所碩一，找到了一群志同道合的夥伴們一起努力，成功申請了這次的讀書會，感謝清華大學提供我這麼好的環境，讓我能夠在在清華的日子裡，參與這麼有意義的活動，使我擁有一個比以往都還要充實的碩一上生活。

這次的讀書會讓我學到很多，組員間的討論切磋、老師的授業解惑，再加上我們有另外找一些業界的軟體分析師，教導我們將自己覺得可行的分析方法成程式，更方便使用，除了吸取很多學業上的知識外，我覺得自己在分析數據的思考邏輯上也有顯著的進步，寫程式的能力也增進不少，這些種種，我想在我的未來必定有所用處，要再次感謝清華大學給我這個機會，讓我有機會像這麼多厲害的人學習他們的專業能力。

吳宗祐心得：

之前在大四生碩一的暑假曾經有參與過類似這樣的讀書會活動，但那次是由老師自己親自主辦，所以講者和討論的內容都是老師自己掏腰包或者請人來講，因此之前就有想要自己來辦一個類似像這樣的讀書會，讓我們可以有課堂以外的機會學習一些我們有興趣的事情，借由這次不同系所的成員，以及不同年級的合作，我們成功申請了這次的讀書會，在此很感謝清華大學的教學發展中心，可以在這邊給予我們這樣的機會和環境。

我們主動參與的讀書會，比起一般上課分組的團隊合作，可以讓我們每個人心態更加積極。我們每個人專長都有所差異，所以討論的時候更能互相補足之間的缺點，在困惑之餘，得以請教學校師長以及借由申請的資源來聘請外面業師來講解，讓我們可以除了學生之間的討論外，也有機會聘請業界的講師來教導我們一些，不同於學校的一些觀點，以及在業界實用上的方法。

吳周駿心得：

參加了第二年的讀書會，特別的是這次的身分不只是組員還是讀書會招集人，這和過往經驗十分不同，以前就是把自己該報告的部分準備好就可以，參加讀書會僅須扮演一個跟隨者的角色，而招集人必須掌控每次讀書會的進度以及確保每次讀書會紀錄以及心得有完整的呈現當週所討論的內容。此外還有和指導老師以及導讀人配合的部分，時常經過老師指導後會有新增閱讀的進度要做，而招集人除了要記錄下來外，還要分配內容給所有組員參與，並適當的修改下次討論的內容並對於進度作回報，此外在讀書會後期我們有邀請校外專業人士來參與討論，是一個十分難得的經驗，統計分析是一個理論與實務並重的科目，透過老師的介紹可以知道一些文獻上既有的問題以及解決方法，和業界進行溝通可以知道實際處理資料的方法，和分析師對話可以了解巨量資料在實際分析上會遇到的問題。最後要感謝學校給我們機會參與那麼棒的活動，這次活動能夠成功且順利地進行特別要謝謝教學發展中心的專員們非常細心與耐心地指導我們申請流程以及後續報告繳交等內容，希望那麼有意義的活動能夠持續辦下去，學弟妹能夠受惠於這個活動，最後想說-申請加入讀書會就對了！

王建勛心得：

本學期因為參加台積電所舉辦為期一個學期的半導體大數據比賽而舉辦了這樣的資料分析與統計建模讀書會，經過這四次讀書會，我們確實地走過整個統計分析與建模的流程，比賽順利晉級到複賽。

過程中我們面臨到許多問題，一開始不太了解半導體數據的背景知識，需要請教該領域的專家，這也是我們第一次處理這麼龐大的數據，使用的數據分析方法不能夠僅限於傳統的統計分析，學習建立巨量資料庫再進行分析，將會減少許多資料讀取的時間，因此在第四次讀書會時我們邀請的業界導師教導我們如何使用 Hadoop 和 R 結合，以彌補我們讀書會成員缺少資工領域背景的不足，我發現一個資料科學團隊的成員必須具備能夠互補的能力，每一位成員能夠針對同一個問題從不同的角度給予建議，是一個很好的跨領域學習機會。

總而言之，本學期的讀書會讓我對於統計分析有更深入的了解，也是一個很難得的團隊合作經驗，很感謝學校鼓勵學生自組讀書會，提供完善的環境和經費上的支援。

肆、 評量

陳雅汝評量

完全不同意=1 非常同意=5

時間	2014-10-24 五	2014-11-07 五	2014-11-28 五	2014-12-05 五
讀書會能夠準時開始與準時結束	4	4	4	4
討論內容和研究主題相互配合	5	5	4	5
讀書會進度能夠符合預期目標	4	5	5	5
報告人與組員討論參與程度熱烈	3	4	5	4
讀書會內容引起未來研究興趣	5	3	5	4
能夠學習到新知識或是新觀念	5	5	5	5

整體評量：4 分

小組互評：
吳宗祐：4 分
吳周駿：4 分
王建勛：4 分

吳宗祐評量

完全不同意=1 非常同意=5

時間	2014-10-24 五	2014-11-07 五	2014-11-28 五	2014-12-05 五
讀書會能夠準時開始與準時結束	5	5	5	5
討論內容和研究主題相互配合	4	4	3	4
讀書會進度能夠符合預期目標	5	5	5	5
報告人與組員討論參與程度熱烈	4	5	5	4
讀書會內容引起未來研究興趣	4	4	4	5
能夠學習到新知識或是新觀念	4	3	3	4

整體評量：5 分

小組互評：

陳雅汝：5 分

吳周駿：5 分

王建勛：4 分

吳周駿評量

完全不同意=1 非常同意=5

時間	2014-10-24 五	2014-11-07 五	2014-11-28 五	2014-12-05 五
讀書會能夠準時開始與準時結束	5	4	3	4
討論內容和研究主題相互配合	5	5	5	5
讀書會進度能夠符合預期目標	5	5	5	5
報告人與組員討論參與程度熱烈	3	4	4	4
讀書會內容引起未來研究興趣	5	5	5	5
能夠學習到新知識或是新觀念	4	4	4	4

整體評量：5 分

小組互評：

陳雅汝：5 分

吳宗祐：5 分

王建勛：5 分

王建勛評量

完全不同意=1 非常同意=5

時間	2014-10-24 五	2014-11-07 五	2014-11-28 五	2014-12-05 五
讀書會能夠準時開始與準時結束	4	4	4	4
討論內容和研究主題相互配合	4	4	4	4
讀書會進度能夠符合預期目標	3	4	4	5
報告人與組員討論參與程度熱烈	4	4	4	4
讀書會內容引起未來研究興趣	3	3	4	4
能夠學習到新知識或是新觀念	4	5	5	4

整體評量：4 分

小組互評：

陳雅汝：4 分

吳宗祐：4 分

吳周駿：4 分

伍、 對本校讀書會計畫的建議

1. 在更改次數方面，更改場地的部分希望可以和總次數分開，又或者是總更改次數可以等於讀書會總次數，這次使用的更改次數都和場地有關，由於系所有較高的使用權限因此撞期會被迫更改讀書會日期
2. 這次校外導讀人找得太慢，沒辦法事先就在申請的時候加入，也導致交通費等費用沒有辦法編列在預算當中，因此建議學校可以在校外導讀人的部分可以臨時申請，同時在交通費的部分可以選擇列講師費用，因為找竹科的專業人員來講解不太需要額外的交通費
3. 在成果報告繳交部分應該可以利用類似進度心得的格式，按照每個章節繳交，同時平日心得因為和部落格連結在編輯的時候會有格式跑掉的問題，不知道可不可以改成繳交電子檔案就好，不要貼到網頁上編輯