**多變量分析期末報告**

**Multivariate Analysis Final Report**

**公車路線資料**

**組員：鍾興潔/ 統博一**

**黃三騰/610611105 統碩一**

**林子祥/610611102 統碩一**

**蔡伊婷/410311306 應數四**

* **摘要**

因偏鄉地廣人稀、道路狹小、接駁需求分散，相較於傳統公車定時定班服務，最近新推出的方式更適合偏鄉地區，依據乘客所申請的站點規劃路線，更能提高公車效率。

像是台北的「跳蛙公車」就是讓乘客、公車業者及政府三方透過網路平台交流意見，乘客利用手機APP就可以提出公車搭乘需求，公車業者與路線主管機關依照民眾提出的特地時間，地點，快速規劃並開通公車路線，多條快速、直捷的快速公車路線。跳蛙公車不僅節能環保，而且在路線規劃與價格以達到最佳化，但路線開通需要達到一定的人數與天數，這對於偏鄉地區來說非常困難。

因此出現了「小黃公車」其結合了跳蛙公車與計程車，取代大眾客運無法到達的地方。此規劃大大的改善民眾生活的便利性。像是提升就醫的方便性、也降低了空車率。

我們用了PCA將資料的維度降低但仍然保留站點之間的重要訊息、接下來從直觀上我們看出大概能分成四群,所以用Hierarchical進行分群，先將站點投影在PCA上的座標位置,彼此間的歐式距離定義出距離矩陣再使用single linkage進行分割，由圖可以觀察最佳的分群數目是4個。也使用另外一種分群法K-means其主要目標是要在大量高維的資料點中找出具有代表性的資料點，然後在根據這些群中心，進行資料分類，使用這些方法主要是將各個族群進行常用站點分析，找出那些站點對於學生是重要的或是對於老人。

最後希望能為學生規劃專車、針對敬老票特別高的站點做為小黃公車停靠的據點，希望能讓偏鄉地區的民眾有更方便的交通。

目錄

[ **摘要** 2](#_Toc517020746)

[**第一章** **前言** 4](#_Toc517020747)

[**第二章** **分析方法** 4](#_Toc517020748)

[I. 敘述統計 4](#_Toc517020749)

[II. PCA 4](#_Toc517020750)

[III. Hierarchical 5](#_Toc517020751)

[IV. K-means 5](#_Toc517020752)

[**第三章** **問題與討論** 6](#_Toc517020753)

[**第四章** **分析結果** 7](#_Toc517020754)

1. **前言**

在偏鄉地區,而我們知道對於老人以及小孩來說主要的交通方式就是搭乘大眾運輸工具。然而在我們的資料中顯示，且有少數停靠站更是鮮少有人使用，搭乘率主要都是敬老票與學生票，因此我們希望能夠針對族群找出對於特定族群重要的站點。

1. **分析方法**
2. 敘述統計

希望能透過敘述統計的圖表來瞭解資料，從時間與站點的點圖，路線11一天有三班車。從ID個數的長條圖發現有六個人整年平日搭車達到一百多次,最高的達到一百四十幾次，而總搭乘數達到八千三百多次,由此可見主要的客戶還是以散客為主，與我們猜測主要由固定客搭乘的假設違背，將資料用圖形的方式呈現，讓我們能快數的瞭解資料型態。

1. PCA

n=72(站點)p=591(搭車模式)，將72個站點投影到591個維度上無法直接看出差異性，因此透過PCA將資料降維，前三個PC就能有66.2%的解釋力，從圖明顯可以看出有四群，因此我們希望透過不同的分群方法能分出我們看到的四群

1. Hierarchical

我們使用聚合式階層分群法(agglomerative hierarchical clustering)。利用歐氏距離作為測量相似性(measure of similarity)。

第一步: 將每一筆資料視為一個聚類 .

第二步: 找出所有聚類之間，距離最接近的兩個

第三步: 將 合併成一個新的聚類.

第四步: 重複第二步及第三步直到所有聚類合併成一個或其他停止條件滿足.

其中第二步的聚類之間的距離，我們定義為不同群聚中最接近兩點間的距離(single-linkage):

1. K-means

K-means為一種分群方法，使用資料的距離作為分群依據。在此篇使用的距離為歐氏距離。

首先設定k 個初始均值點 ,

第一步: 考慮

使得每個被分配到某一個聚類.

第二步: 以第一步分群後各個聚類內的均值點，作為新的k個均值點。

第三步: 重複著第一步和第二步直到一些停止條件達成。

1. **問題與討論**

* 在對**”路線11站點與使用次數”**之矩陣做PCA降維時發現需要使用到第11 PC才能夠有超過60%的解釋力，由於維度較高，不便畫在圖形上做觀察，因此相較於需要事先判斷分成 k 群的k-means分群方法，我們更傾向於使用Hierarchical Clustering 的分群方式，進行下一步的分析。
* 在觀察**”路線11站點與使用次數”**之矩陣時發現，有部分乘客所使用的站點是相同的，因此我們將其視為一種**”搭車模式”**並將具有相同搭車模式之乘客累加起來，此舉所附帶的好處便是能夠進一步降低資料的維度。
* 在對整理過後的**”路線11站點與搭車模式”**之矩陣做PCA時發現，此時只需要使用到第3 PC就能夠有超過60%的解釋力，因此我們先將DATA投影在由PC1-PC3所建構出的3D圖上，再從不同角度去觀察，判斷分成幾群，並使用k-means分群方法，去觀察會將哪先車站分到不同的群裡面，同時我們也有使用Hierarchical Clustering 的分群方式兩相對照，觀察分群的結果有何不同。
* 對**” 路線11站點與搭車模式”**之矩陣做完分群之後，我們將分群結果對照**”各站點使用量”**之長條圖，發現有部分使用量非常多的站點並沒有被分出來，因此我們猜測 :

1. 可能是因為對**”搭車模式”**做分群，而該站點之使用量是由多種不同搭車模式提供，造成此站點沒有被區分出來。
2. 是否因為不同的族群(老人、學生)搭車習慣不一樣導致此結果。
   * 依照乘車票種進一步將**”路線11站點與搭車模式”**分為

**”路線11(敬老票)站點與搭車模式”**與

**”路線11(學生票)站點與搭車模式”**進行分析。

* 乘客刷卡記錄只占總搭乘率之20%
* 每天公車發車時間相同，但透過畫出一整年非假日乘客搭乘公車刷卡記錄的點圖，並按照公車行經站點排序，發現乘客搭車時間誤差明顯偏高，初步猜測是因為每日的路況都不盡相同，而造成搭乘時間有較大的變動。

1. **分析結果**

資料顯示GS、HP、CL2、JP這四個站對於學生來說是重要的，我們猜測這四個站點的上、下學時間會有大量的學生搭乘，如果能在這些站開發新的路線能舒緩人潮堵塞。對於老人而言，SCS、GS、CM、MF、LYTO這些站是重要的,如果能在這幾個站點設立據點，如此一來能使得民眾搭車更加方便。