**多變量分析期末報告**

**Multivariate Analysis Final Report**

**公車路線資料**

**組員：鍾興潔/810611001 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**

傳統公車定時定班服務，是目前廣為人知且歷史悠久的大眾運輸工具，尤其在偏鄉地區，因為沒有大都會區特別規畫的其它交通運輸設施(如: 捷運、輕軌列車)，使得民眾不得不使用傳統公車的服務方式，但偏鄉地廣人稀、道路狹小、接駁需求分散，使用傳統公車的服務可能較沒有效率。

最近政府最近新推出了一些有別於傳統公車的設計，使得民眾搭車方式可以更靈活，例如:台北的「跳蛙公車」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。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。

因此另一種「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，並且更適合偏鄉地區，可以提升就醫的方便性、也降低了空車率。

本篇文章在探討對於偏鄉地區，傳統的公車搭配「小黃公車」的可能性。我們收集了某條公車路線的乘客搭車刷卡紀錄，並且我們使用主成分(PCA)降維，從直觀將客群分為四群，同時我們也使用了Hierarchical進行分群，和 K-means交互驗證我們的分群結果，使用這些方法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。

最後的分析結果，我們分析出對於學生而言有四個重要站點，而對老人而言有五個重要站點，希望未來能為學生規劃專車、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,老人以及學生為主要的交通方式就是搭乘大眾運輸工具。並且在我們的資料中顯示，有少數停靠站是鮮少有人使用，搭乘的次數主要來自敬老票與學生票，這與我們主觀的認知是一致的，因此我們希望能夠針對這兩個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

透過敘述統計的圖初步瞭解資料。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

Principal Component Analysis，主成分分析。這裡用來將原本複雜的資料降維，同時保留資料原始變異程度，達到降低資料的維度以及遺失的訊息壓到最低。

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 , 接下來

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停止條件達成。

1. **分析**

從刷卡時間與上車下車站點的散布圖，發現路線11一天有三個班次。從ID個數的長條圖，發現有六個人整年平日搭車次數達到一百多，最高的有一百四十幾次。而總搭乘數為八千三百多，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違背。

將每一個站點的使用次數視為樣本，變數部分代表該站點的使用次數被哪些刷卡紀錄所貢獻，得到n=72(站點)，p=8380(被使用次數)的資料矩陣。

由於許多刷卡紀錄所使用的站點是一樣的，意味著將相同的加起來合併，得到路線11的搭車模式。在n=72(站點)，p=591(搭車模式)，無法將72個站點的散布圖畫出來，因此透過PCA將資料降維。發現前三個PC就有66.2%的解釋力，畫圖可以看出資料約為四個群體。

接著分別利用K-means及Hierarchical將這四群分出來，同時對照著站點的使用次數長條圖。確實使用次數高的站點有被分到同一群，但也有使用次數高的站點被分到那些使用次數低的站點中。

初步猜測是由於不同族群的搭車習慣不同造成，因此我們將分析的對象改為分別對敬老票以及學生票進行上述步驟。

結果發現對於學生而言，GS、HP、CL2、JP這四個站是重要的，我們推論這四個站點的上、下學時間會有大量的學生搭乘，如果能在這些站開發專車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使用「小黃公車」，如此一來能使得民眾搭車更加方便。

1. **問題與討論**

* 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時間與站點的散布圖，並按照公車行經站點排序，發現乘客搭車時間有部分誤差，但對分析並沒有太大影響。
* 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，無法畫在圖形上做觀察，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。

1. **結論**

我們透過以上方法分析出路線11的重要站點，並初步規劃出學生族群專車，以及敬老族群的「小黃公車」，期望能使公車的使用率以及效率達到最大。

取得的資料為乘客刷卡記錄，而這些紀錄只占總搭乘率之20%，是否刷卡記錄和整體的搭車紀錄表現相符，是值得未來進一步探討的。