多 變 量 分 析 期 末 報 告

Multivariate Analysis Final Report

公 車 路 線 資 料

Bus Route Data

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，因為沒有大都會區多元的交通運輸設施(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較沒有效率。

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定時間、地點，快速規劃並開通公車路線。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的優點，取代大眾客運。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，讓偏鄉地區的民眾有更方便的交通。

第一章 前言

在偏鄉地區，老人以及學生主要的交通方式就是搭乘大眾運輸工具。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族群找出對於特定族群重要的站點。

第二章 分析方法

I. 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，讓我們能快速的瞭解資料型態。

II. PCA

Principal Component Analysis，主成分分析。這裡用來將原本高維度的資料進行降維，但盡量減少在降維過程中遺失的重要訊息。

III. 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 $\mathrm{C_i}, i = 1, \dots, n$。

第二步: 找出所有聚類之間，距離最接近的兩個 C_i, C_j。

第三步: 將C_i, C_j 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):

$$\mathrm{D}(\mathrm{C_i}, \mathrm{C_j}) = \min_{a \in C_i, b \in C_j} d(a, b) \text{ , where d}(\cdot)\text{: Euclidean distance}$$

IV. 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 $m_1^{(1)}, m_2^{(1)}, \ldots, m_k^{(1)}$，接下來

第一步: 考慮

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \le \left\| x_p - m_j^{(t)} \right\|^2 \ \forall\, 1 \le j \le k \right\},\ i = 1, \ldots, k$$

使得每個x_p被分配到某一個聚類$S_i^{(t)}$。

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 $m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$ ，　$i = 1, \ldots, k$。

第三步: 重複著第一步和第二步直到停止條件達成。

第三章 分析

從刷卡時間與上車下車站點的散布圖，看出路線 11 一天有三個班次。從 ID 個數的長條圖，且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，由此可見大部分的客戶還是以散客為主，並非主要由固定乘客所搭乘。

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，將在上下車站點一樣的人歸為同種模式，可將資料矩陣轉成搭車模式矩陣。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並未被分出來。

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;對於老人而言，SCS、GS、CM、MF、LYTO 這些站是重要的。

第四章 問題與討論

- 透過畫出”**路線 11 整年非假日乘客搭乘公車刷卡記錄**”的散布圖，並按照公車行經站點排序，我們發現該路線每日發三班車，但乘客搭車時間存在誤差，初步認為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。

- 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，因此我們先將資料投影在由 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，去觀察哪些車站分到同一群，同時也使用 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，明顯看出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。

第五章 結論

經由分析結果我們發現 GS、HP、CL2、JP 這四個站對學生是重要的，究其原因，我們利用網路地圖搜尋這四個站點附近的建築，發現這四個站點附近確實都有國中、小，此外，對於老人較重要的五個站點，SCS、GS、CM、MF、LYTO，我們也發現，這些站點附近較多功能性建築，例如:醫院、市場、公務機關…等。

針對這些重要站點，我們希望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。

然而我們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，並能夠套用我們分析出的結果，是值得未來進一步探討的。