九十六年道路交通
安全與執法研討會
中華民國96年9月

運用資料探勘於公車旅行時間推估模式之建構－以桃園縣蘆竹鄉中正路至南崁路路段為例

朱松偉[1]　林佑霖[2]

摘 要

先進大眾運輸系統中的即時車輛到站資訊可以改善大眾運輸系統中的派遣任務，並且提高乘客搭乘大眾運輸運具意願。因此旅行時間的預估就變的非常重要。在旅行時間推估上多以路面上佈設之車輛偵測器（VD）所蒐集之資訊為主，但VD之埋設需經路面挖掘、VD佈設、重鋪柏油等程序，佈設不容易且損壞率偏高，亦普遍存在維護困難的問題；故要在市區佈設大量偵測器需耗費較大成本，使得透過探針車蒐集即時且連續之交通資訊收集變成可能，因此若能透過探針車提供交通管理上所需之相關數據，實為成本較低且應用性更廣之選擇。所以，本研究自2006/08/01至2006/08/31期間所累積之大量時空資料中，運用SQL Server 2005軟體，並結合資料探勘分析方法，選取桃園縣蘆竹鄉中正路至南崁路路段，利用探針車實際預估市區公車到站時間，透過其所傳回大量真實GPS時空資料，發展一套資料的處理流程，再透過資料探勘方法建構出模式並加以探討。而本研究顯示透過長時間探針車資料，可有效區格研究區內不同星期類型的尖峰/離峰時間。

關鍵字：探針車、資料探勘、旅行時間推估

壹、前言

智慧型運輸系統(Intelligent Transportation System，ITS) 係將電子、通訊、資訊、電腦、以及控制等技術加以整合而成，為一種可以提昇運輸機動性、能源效率以及環保，進而改善交通運輸問題的先進科技系統。我國交通部自推動電信自由化以來，交通運輸相關單位開始注重引進資訊、通訊等科技並結合交通運輸之專業

[1]清雲科技大學行銷與流通管理系助理教授（聯絡地址：桃園縣中壢市健行路 229 號，電話：03-4581196 轉 7502，E-mail:swchu@cyu.edu.tw）
[2]清雲科技大學經營管理所碩士生（聯絡地址：桃園縣中壢市健行路 229 號，E-mail:9434006@cyu.edu.tw）

知識，隨著世界智慧型運輸系統發展的潮流將先進的科技整合到運輸領域中。而先進交通管理系統（Advanced Traffic Management System，ATMS）之發展因應ITS而生，該系統係在現有的道路上，進行交通狀況之預測、交通管理策略之分析、評估與執行整體性的交通管理，並將相關資訊傳送給用路人，以達到運輸效率最佳化與運輸安全之目的，為ITS發展的基礎與核心。

公車是大眾交通運輸載具之一，其到站的時間經常是搭乘者無法掌握且必要花費的時間，長久以來傳統的市區公車站牌上所能提供的班車相關資訊並不能夠滿足搭乘者的需要，這些站牌上往往僅列出了該站站名、營運路線圖、頭末班車發車時間以及班車間距或起站發車時刻，造成搭乘者如要搭乘不常搭乘的路線、服務班次較少或車班相距較長的客運路線時，必須花費更多不明確的時間等候，因而耗費許多寶貴時間。所以如果能夠提供搭乘者正確的乘車資訊，例如預估到站時間以及預估旅行時間，這將能使搭乘者節省較多的等候時間，也會增加民眾搭乘大眾運輸載具的意願。另一方面，近年來利用裝置具備全球衛星定位系統(Global Position System，GPS)、無線通訊等功能車上單元（On Board Unit，OBU）之探針車（Probe Vehicle，PV）來蒐集交通資訊，隨著實際蒐集交通資訊探針車數量之增加以及相關理論與應用等研究成果之發展，而逐步變成即時交通資訊蒐集之主流。因此，若能透過探針車提供ATMS上所需之相關數據，實為成本較低且應用性更廣之選擇。且透過探針車所傳回之GPS資料，發展一套資料的處理流程，藉由大量之GPS資料解析可以真實呈現道路狀況來產生研究或驗證資料，以期能夠提供使用者準確且符合即時現況之車輛旅行時間推估。

貳、研究方法

本文主要係以決策樹技術建構旅行時間之推估模式，而決策樹是一個樹結構（Tree structure），和一般資料結構中的樹一樣，有節點、樹葉等結構，它可以產生易於了解的規則，因其易於讓人了解，故為現今相當受歡迎而普通應用的資料挖掘技術。決策樹利用樹狀結構圖的方式，來推演一連串的決策問題。每個決策樹均由根部開始發展，分支用來決定每一筆資料該進入下一層哪一個子節點，如此反覆進行直到所有資料均到達葉節點為止。由根節點（Root Node）到葉節點（Leaf Node）所形成的規則，做為分類或預測的依據。

在決策樹技術方面，Han & Kamber [1]指出現今決策樹的演算法大致有C5.0、CART、CHAID與QUEST 這四種演算法。CHAID 演算法的目的主要是在每次分割時利用卡方檢定（chi-square test）來計算節點中類別的p-value，以p-value 大小來決定

決策樹是否繼續生長，所以不需要再做修剪樹的動作，CHAID 的一個問題是它無法處理連續型資料，在本研究的資料庫中存在著許多連續型數值的資料，所以並不適用在我們的研究中。CART（Classification and Regression Trees ）演算法是一個二元（binary）分割的方法，應用於資料屬性為連續型的資料型態，每次分割將資料分為兩個子集合，以gini index 評估資料的分散程度，作為選擇分割條件的依據。本研究採用以亂度(Entropy)為基礎之決策樹，為Quinlan[2]所開發的決策樹演算法ID3(Iterative Dichotomiser 3, C5.0/C4.5 的前身)。C4.5是Quinlan改善他自己所發展出來的ID3 演算法，C4.5與ID3最大的不同就是C4.5改進了ID3不能處理連續型數值的問題，所以C4.5可以算是ID3的後續版本。他以資訊增益(Information Gains)作為分岔準則，但是發現應用在實際案例上時，資訊增益會偏好選擇選項數較多的變數作為多，且容易造成過度學習的效應。為了改善這項系統偏誤，Quinlan[2]重新定義出「增益比值(Gain Ratio)」的計算公式來取代原有的分岔準則，但無論是哪個版本，最根本的內容還是透過所謂亂度(Entropy)的概念作為決策樹的分岔準則。以下便為ID3資訊獲利計算方式加以說明。假設我們有一個資料組，有兩個類別標籤 A 跟 B ，a 為類別 A 的個數，b 為類別 B 的個數，對一個給定的資料組所需要的期望資訊 $I(A,B)$ 的計算公式如下：

$$I(A,B)= -\frac{a}{a+b}\log_2\frac{a}{a+b}-\frac{b}{a+b}\log_2\frac{b}{a+b} \qquad (12)$$

接著計算每個屬性的熵值（entropy），以屬性 C 來說，a_i 代表屬性 C 中類別 A 被劃分的子集個數， b_i 代表屬性 C 中類別 B 被劃分的子集個數，所以根據由 C 劃分成子集的熵值或期望值由下式算出：

樣本分類所需的期望資訊：

$$E(C)=\sum_{i=1}^{v}\frac{a_i+b_i}{a+b}I(a_i,b_i) \qquad (13)$$

藉由下列的公式，我們可以分別計算出屬性 C 及其餘屬性的Gain 值之後，決策樹會依序挑選Gain值較高的屬性，開始建構決策樹。

$$Gain(C)= I(a,b)-E(C) \qquad (14)$$

到了C4.5中解決了ID3無法處理連續型數值並以資訊獲利正規化或稱為獲利比的方式緩和產生過多分枝的缺點，正規化的方法是將原來的資訊獲利值除以分割資訊值（split information），如下式所述：

$$Gain\quad ratio = \frac{Gain(C)}{split(C)} \qquad (15)$$

$$split(C) = -\sum_{i=1}^{n} \frac{Si}{S} \times \log_2\left(\frac{S_i}{S}\right) \qquad (16)$$

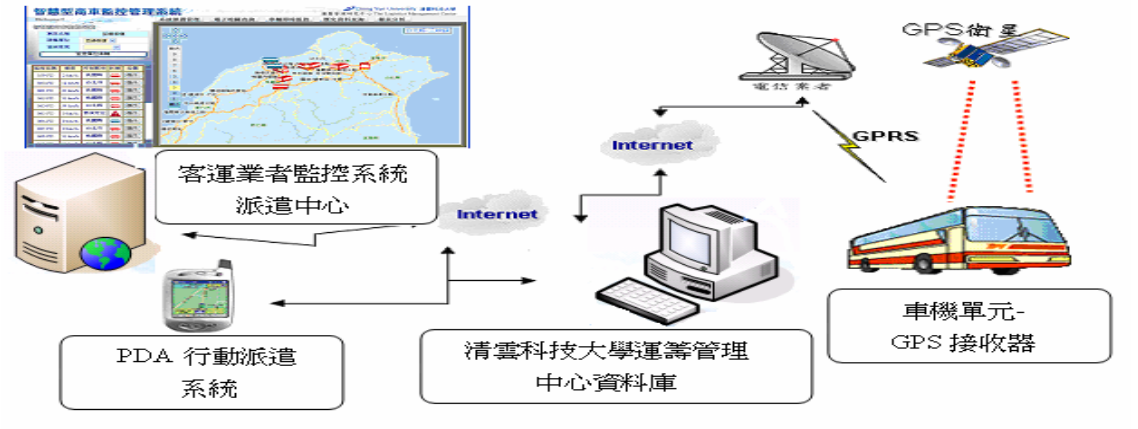
其中S_i是根據屬性A中各類別劃分後的資料子集合，S為所有資料集的總數；而C4.5在處理連續型數值時，會先將數值排序，在依次地計算各別的獲利比（Gain ratio），挑選最大的獲利比為數值的分割點。

早在1960年代開始便有許多文獻以樹狀結構方式進行資料分析，只不過早期的分析工具僅能處理連續型的變數，對於間斷型變數的分析，因為無法同步處理，所以難達到理想的成效。故因此，近幾年來許多學者都在想該如何同時解決連續型和間斷型變數的分析，於是便孕育了迴歸樹（Regression Tree）的誕生，相對也大幅度拓展決策樹的應用範圍。迴歸樹分析方法係整合決策樹與迴歸的一種分析方法。其分析過程可區分為兩各步驟，首先，透過決策樹先產生樹狀規則分岔，而分岔準則的目標就是要使得分岔後樣本的連續數值變異數降低，再接下來，將在純化過的樹狀分岔樣本中，各自建立一條線性方程式，以此過程達到預測間斷變數及連續變數之功能，此即迴歸樹分析方法之原理[3]。

本研究所採用的迴歸樹分析方法係資料探勘技術的一種。而本研究所採用的資料探勘軟體為SQL Server 2005，它有別於SQL Server 2000，主要因SQL Server 2000的決策樹只能處理類別變數的分類問題，而無法解決連續變數的問題。SQL Server 2005的決策樹演算法中，除了原先就有的分類樹演算法之外，還另外增加了迴歸樹演算法，讓決策樹具備預測連續變數的功能，此功能即符合本研究之預測旅行時間之要求，故採用SQL Server 2005之迴歸樹分析方法。

參、資料蒐集與分析

為建立旅行時間之預測模式，本研究之探針車資料利用清雲科技大學運籌管理研究中心之商車營運系統資料庫，進行空間資料探勘作業，該資料庫主要配合某客運業者行駛班次、路線作為基本資料，該客運為大型車且為固定路線，選取桃園縣蘆竹鄉中正路至南崁路路段，研究範圍長度總長共1.167公里，其餘之定義則在後列所述。資料蒐集流程如下圖一。

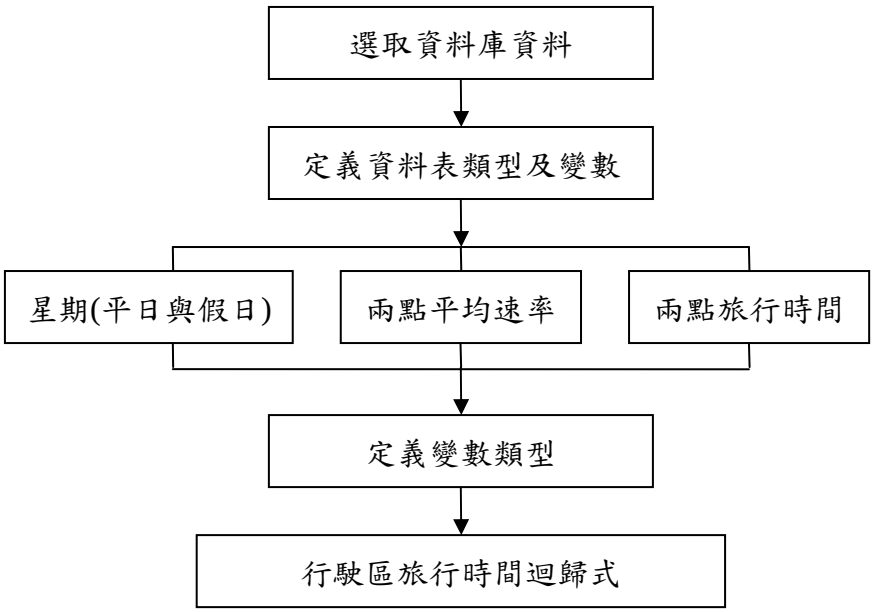


圖一 探針車資料蒐集流程圖

本章節將對研究中將會使用到的相關定義作一註解說明，本研究所使用或定義之各公式所使用之單位，時間為秒(Second)、距離為公尺(Meter)。其中本研究一共有2個停靠站牌，故共有2個停等區，由於道路特性，其站牌點與交通號誌點的距離相距甚近，故本研究不另將交通號誌納入停等點之列。另外，由於GPS會有些許偏移誤差所以在選取資料時，本研究使用之GIS軟體為Arc view，為了快速篩選出本研究之研究範圍資料，故利用Arc view所提供之分析功能，也就是環域分析工具進行分析，並以市區道路的道路中軸線向外各延伸35公尺，作為本研究篩選資料的範圍。其主要之原因為35公尺之環域分析可以包含住95%以上之資料，在此本研究認為超越35公尺環域範圍的資料為漂移嚴重的定位資料，故本研究將之剔除。在本研究中各分割路段及停等區域的資料選取皆是以相同方法來獲得。

肆、模式建立

本段說明決策樹中所使用的輸入與輸出變數，本研究所欲預測的模型大致上可分為行駛路段預測模型、停等區發生停等狀況預測模型以及停等區發生非停等狀況模型等三種。在輸入變數的的選取上，藉由不同的輸入變數將會產生出不同的模型，然而變數的有效性與否也將會影響模式的表現，在藉由多次反覆的測試與排列組合中，本研究最後選定以星期兩點平均速率與兩點平均旅行時間作為預測模型之輸入變數，再者由於本研究以推估旅行時間為主，故在行駛路段輸出變數選定為行駛區旅行時間，在停等區方面則是停等區旅行時間(如圖二)。



圖二 模式架構圖

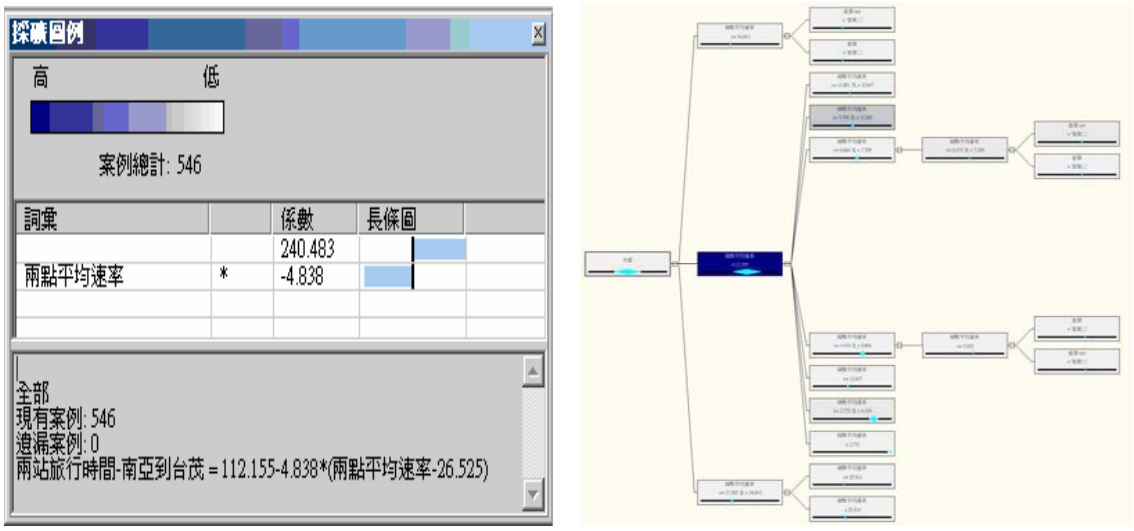
在確定研究範圍之後，本研究選取某客運業者行駛台北到桃園蘆竹鄉中正路段路線圖中之兩重點站為範圍，並將路段依照站牌點劃分為停等區及路段行駛區，其中路段由南到北依序為台茂站到南亞站行駛路段、南亞站到溪洲站間行駛路段，共2個路段為分析路段。

4.1 台茂站到南亞站行駛路段決策樹分析

本研究的目的為推估旅行時間，故設定行駛區旅行時間為預測之變數，輸入變數則選擇星期、兩點平均速率以及兩點旅行時間，決策樹建構完成後，共分割成五層(如圖三)，共產生二十個分支節點，雖然在決策樹中每個節點都有其迴歸式產生，然而可用之有效迴歸式為各分支節點的葉子節點，故本路段所得出之預估模式共有十三個，以下將以下就包含根節點及各葉子節點預估模式詳列於下。

由上式可以發現自變數-星期在大多被分支成星期二及非星期二的節點，扣除根節點後在二十個節點中共佔了八個，其次為星期四及非星期四的節點共佔了四個；其中星期五及非星期五的節點各為二個，星期一及非星期一的節點各為二個，另外可以發現星期三並沒有在這個路段中被分支節點出來。

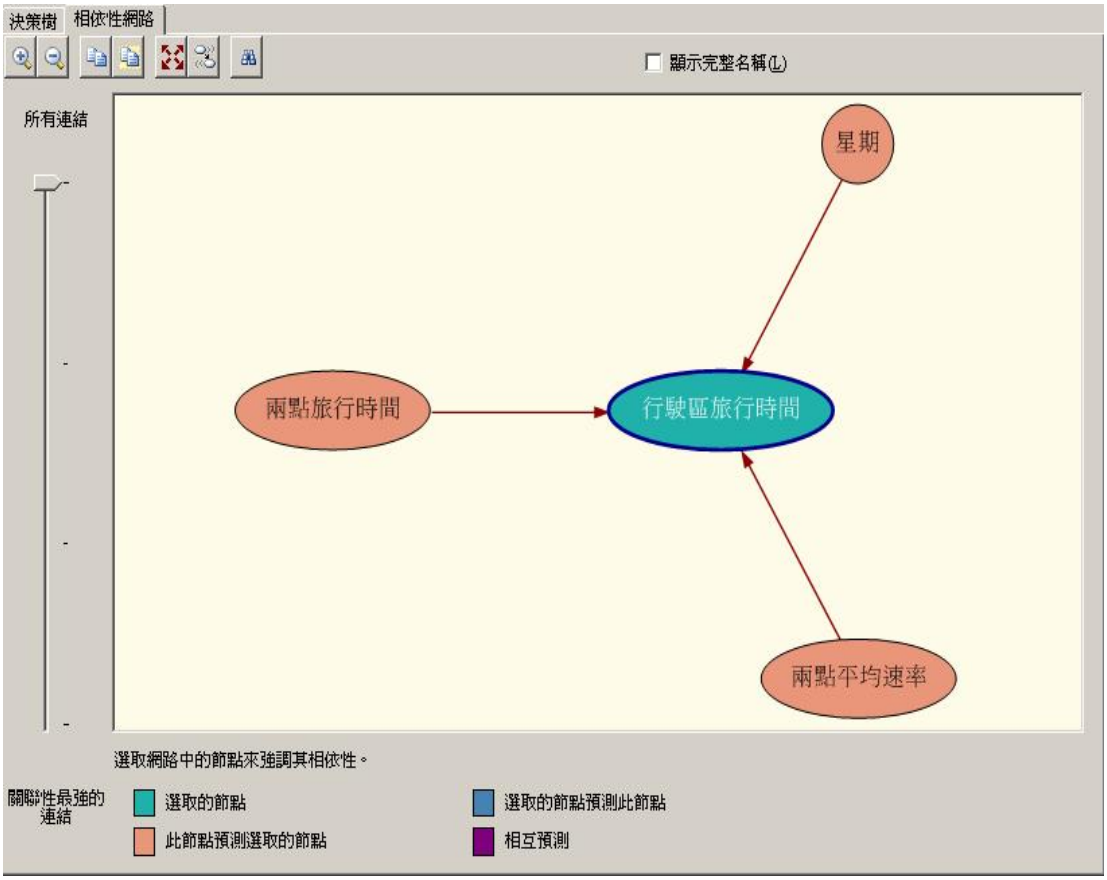
由於決策樹分割成五層，並由決策樹中得知兩點平均速率共被分成十三種狀況，故在迴歸式的使用上可依兩點平均速率的不同而運用不同之迴歸式推算旅行時間。舉例來說，當我們得知今天為星期二、兩點平均速率為 19.52 時，則在台茂站到南亞站可使用兩點平均速率 >= 18.004 及 < 22.564 及 星期不為星期四這個旅行時間預估模式，將上述資料代入行駛區旅行時間 = 121.383-5.987*(兩點平均速率-20.243)後得知當車輛進入此路段大約所需花費的行駛區旅行時間預估為 117.0544 秒。

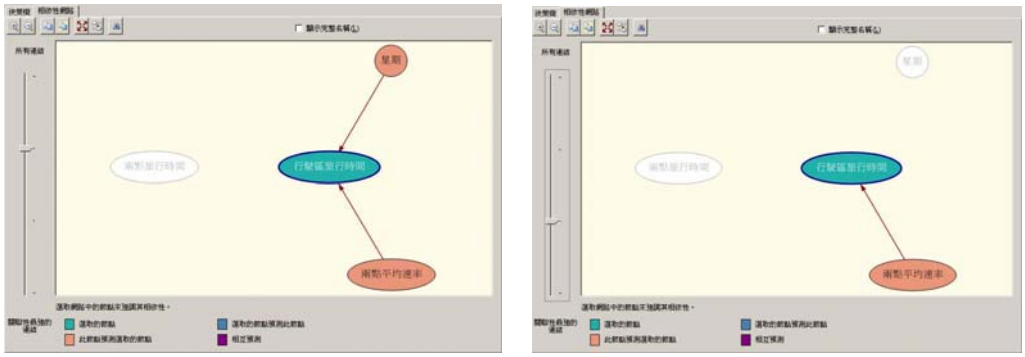


圖三 台茂站到南亞站行駛路段決策樹

4.2 台茂站到南亞站行駛路段相依性分析

藉由相依性網路可以得知自變數和應變數之關係，當關聯性越強則其相依性越強，由圖 4.4 可以得知當將關聯性遞增時自變數-兩點旅行時間對於行駛區旅行時間之間的箭頭消失；當我們再將關聯性增強時，由圖中也可得知自變數-星期對於行駛區旅行時間之間的箭頭消失，也就是在台茂站到南亞站行駛路段中兩點平均速率對於行駛區旅行時間的關連性最強，星期對於行駛區旅行時間的關連性次之，然而兩點旅行時間對於行駛區旅行時間的關連性在這三個自變數中是最弱的。





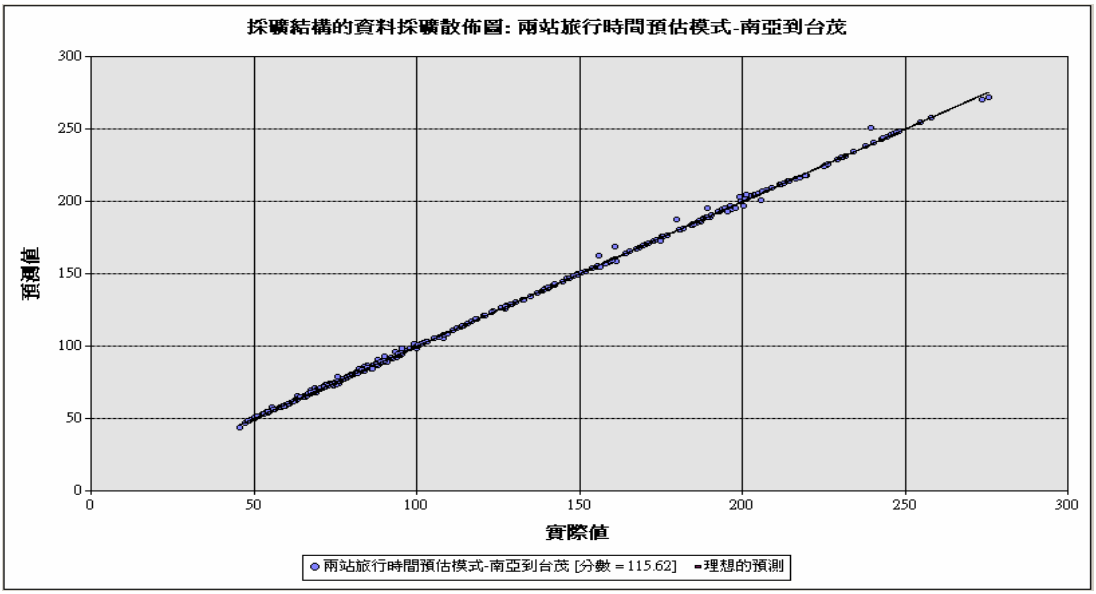
圖四 台茂站到南亞站行駛路段相依性

將本研究所需的決策樹建構完成後，在路段的行駛區部分一共建構了 2 個決策樹模型。在所建構之模型分別使用另一組測試資料來做模式的實證分析，在資料方面則是採用兩個星期共六天之資料來當做驗證資料，除了維持每週天數一樣之外，也讓驗證資料有一定的數量。

4.3 台茂站到南亞站行駛路段散佈圖與折線圖分析

本節將比較驗證資料使用所構建出的迴歸式其產生的預測值與實際值做比較。

如圖 4.5，藉由資料探勘的散佈圖可以初步看出模式是否準確，藉由訓練資料產生行駛區旅行時間迴歸模式之後，將驗證資料帶入所符合的迴歸式後之實際值與模式預測值做相互比較，其中圖中四十五度線為預測線，故由圖中可知模式所建構出的決策樹在實際值旅行時間為 150 秒以下時其模式所預測之預測值與實際值差異不大，在實際值超過 150 秒時開始有微幅波動，然而在實際值到達 200 秒時其實際值與預測值間的差異波動增加，所以可以發現當行駛區旅行時間越長時，模式所做出之預估可能會發生較大的差異，但整體上來說，本模式預估能力之準確度應尚可接受。



圖五 台茂站到南亞站行駛路段散佈圖

在本段則是將模式相關數據作一比較整理，首先先來看行駛路段的部份，由表一中可以得知行駛路段其平均實際旅行時間與平均預估旅行時間的差異均不到一秒，這顯示了行駛路段部份其資料分佈平均且預估模式有著相當不錯的效果。

然而，在二個行駛路段中以台茂站到南亞站間行駛路段的平均旅行時間約 112.152 秒，其次南亞站到溪洲站間行駛路段只需約 66.030 秒。

表一 行駛路段實際與預估平均旅行時間比較表

項目	平均實際旅行時間(秒)	平均預估旅行時間(秒)
台茂站到南亞站	112.152	112.468
南亞站到溪洲站	66.030	66.688

資料來源：本研究整理

有預測必有誤差，要達到百分之百的準確度是相當困難的，所以當檢視模式其預估能力時，除了整體準確率之外，還能夠藉由模式高估或低估的現象來看出模式其預測之特性。當設定能夠接受之預測模式誤差率為正負 5%以下時，則當準確率達 95%以上時視為預估準確。

從表二發現當可接受誤差率為 5%時，行駛路段如有較高的高估百分比與低估百分比的現象發生，推測其發生的原因事由於行駛路段的距離較短使得資料量下降，由於資料量的不足將會影響到模式之建構亦會影響後續的驗證結果，本研究在此兩路段中所得到之資料量夠充沛並不影響到模式的表現。

表二 行駛路段預估模式高低估百分比較表

項目	高估百分比(%)	低估百分比(%)	準確百分比(%)
台茂站到南亞站	3.50%	3.18%	93.31%
南亞站到溪洲站	2.70%	4.82%	92.49%

資料來源：本研究整理

最後，藉由驗證資料的輸入，直接驗證訓練資料所建構出之行駛路段旅行時間模式，在表三中可以發現 2 個行駛路段模式都有 90%以上的模式平均準確率表現，所以本研究所建構之 2 個行駛路段旅行時間預估模式其效果尚可接受。

表三 行駛路段預估準確率比較表

項目	台茂站到南亞站	南亞站到溪洲站
準確率(%)	98.50%	98.30%
誤差率(%)	1.50%	1.70%

資料來源：本研究整理

伍、結論

本研究以大量、連續性之探針車資料，結合資料探勘之方法與技術，選取桃園縣蘆竹鄉中正路至南崁路路段，實際推算於市區道路行駛之旅行時間與範圍，經過嚴格資料篩選與處理流程後，利用多筆時空採樣資料，研究結果顯示透過較高時、空解像力之資料，不僅可有效區隔研究區內不同日期類型的尖峰/離峰時間以及行駛速率，而本研究在模式操作過程之中，可以透過更新每月之探針車蒐集資料，套用在模式上，可換算出另一參數值，再由參數值推估出公車於到站之旅行時間，可作為用路人之前資訊。且本研究也提出一套由GPS原始資料藉由資料篩選、整併、轉換等等的程序並且一直到建構預估模式的處理流程與方法，這對於有意利用此類擁有大量真實GPS資料庫來做研究的研究者有者相當大的正向幫助。

本研究依照研究範圍以公車兩重點站間為主，在建構模式的過程中，選定星期、兩點旅行時間與兩點平均速率做為輸入變數，而行駛區旅行時間為推估模式之輸出變數，其中共完成2個決策樹總模型，在各模型中依照條件代用不同的迴歸式，即可得到旅行時間推估。在2個行駛路段的模式表現中，模式準確率高達95%以上，本研究所發展之路段推估模式尚可滿意。

對於後續研究者提出以下幾點研究建議，首先，雖然本研究在行駛路段的模式推估上有著尚可的表現，但唯獨僅以公車其營運路段中的兩重點站作為研究範圍，對於完整營運路線的推估上仍有待模式發展，希望後續研究者能夠以本研究的模式為基礎，繼而將營運路線完整串聯起來，相信對於業者、搭乘者都有著莫大的助益。再者，由於本研究最後僅以決策樹建構預估模式，建議後續研究者能夠增加利用其他資料探勘模式來加以比較。

參考文獻

1. Han, J. and Kamber M., (2000), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.
2. Quinlan, J. R., (1993), *C4.5: Program for Machine Learning*, Morgan Kaufmann.
3. 尹相志(民國 94)，SQL Server 2005 資料探勘聖經，初版，台北：學貫行銷出版。