

國立陽明交通大學
運輸與物流管理學系

碩士論文

透過資料融合以梯度提升決策樹方法預測高速公路旅
行時間之研究

Predicting Freeway Travel Time by Using Gradient
Boosting Decision Tree Method Through Data Merging

研究生：李敏聞

指導教授：王晉元

中華民國一 一 〇 年 七 月

透過資料融合以梯度提升決策樹方法預測高速公路旅行時間之研究

Predicting Freeway Travel Time by Using Gradient
Boosting Decision Tree Method Through Data Merging

研 究 生：李敏聞

Student：Min-Wen, Li

指導教授：王晉元

Advisor：Jin-Yuan, Wang

國 立 陽 明 交 通 大 學

運輸與物流管理學系

碩 士 論 文

A Thesis

Submitted to Department of Transportation and Logistics Management

College of Management

National Yang Ming Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Traffic and Transportation

July 2021

Hsinchu, Taiwan, Republic of China

中 華 民 國 一 一 〇 年 七 月

誌謝

能順利完成論文，首先要感謝我的指導教授 王晉元老師，雖然我在大學時就讀的並不是交通運輸相關科系，但老師這兩年來一步步帶領我，讓我快速累積運輸知識，並且在參與計劃案中更加認識運輸產業的現況。此外，我從老師身上學到如何獨立思考並解決問題，也加強了自己的時間管理能力。期許自己能帶著老師這兩年所教導我的做事態度面對未來的人生的旅途。

感謝 蘇昭銘老師和 黃家耀老師擔任我的口試委員，在口試時給予我指導與建議，讓我的論文能更佳完善。感謝碩士班就讀期間所有授課老師的諄諄教誨，讓我能順利完成學業。

謝謝每一位王 lab 好夥伴，77、凱肉、士軒、阿丁、筱寧、明展。我們一起忙東忙西當 OR 助教，也常常一起為計劃案苦惱，甚至熬夜看日出趕進度，雖然當時累個半死但現在想想其實也蠻有趣好笑的！在冬暖夏涼的 1103 裡有你們一起玩耍、閒聊、耍廢、唱歌、拼拼圖，甚至拖我一起去打球運動，大概是碩班中最幸運的一件事了！我也要特別感謝品蓁學姊，在我苦惱時陪我聊天，給我建議，帮大家處理行政上的大小事。謝謝學長姐，ㄣㄣ、佑芳、子勤、王允，傳承給我們很多資料減少我們徬徨的時間。也謝謝學弟妹語芳、日隆、子賢、欣妤，在我們為論文忙得焦頭爛額的時候分擔很多助教的工作。

謝謝同屆的運管朋友們，從大學就開始互相加油的旅遊+咖啡廳夥伴仲仲，相見恨晚的成大寶貝們采馨、玄鈺、家瑄、雅方、怡婷，不吵架就無法對話的宗翰，還有一起玩耍、吃美食、打球、開趴的威瑜、庭緯、晟華、先維、啓洋、祐德、貞瑩、瑀妮、庭嘉、宣理等。兩年來與大家共同創造了許多美好回憶，希望未來大家一切順利，同學會時再一起聊聊這些日子的各種荒誕小事吧！

謝謝大學時期的好麻吉，654 的老王與老楊，大學一起住了四年，雖然唸了不同研究所，但不定期聊天、訴苦、煮湯圓、看綜藝，總是能讓我充滿力量繼續面對論文。謝謝樹這一年來包容我的各種情緒，陪我面對各種生活中的喜怒哀樂。

最後也是最重要的，我要感謝我的家人。在我的求學階段總是給予我最大的支持與鼓勵，讓我無憂的完成學業。謝謝我的父母，讓我在新竹生活的這六年，每次回家都可以軟爛的耍廢充電。謝謝你們把我培養成今天的模樣，辛苦了，我愛你們！

李敏聞 謹誌

2021.07.05 於新竹交大

透過資料融合以梯度提升決策樹方法預測高速公路旅行時間之研究

學生：李敏聞

指導教授：王晉元

國立陽明交通大學運輸與物流管理學系碩士班

摘要

旅行時間是道路上最重要的旅行資訊之一，若能準確預測未來的旅行時間，用路人可提早規劃行程，管理單位亦能事先規劃交通疏導措施。本研究之目的為融合多元資料，建立機器學習模型，用以預測高速公路未來的旅行時間。

本研究利用高速公路總局之公開資料進行分析，提出兩種融合 ETC 資料以及 VD 資料之方法，並將資料輸入梯度提升決策樹中，經過參數校估，進行旅行時間預測。實證研究結果顯示，本研究所提出之資料融合方法二相較於單純 ETC 資料可以提升 82%~97% 的準確率，顯示本研究所使用之資料融合方法，可以有效提升整體預測準確度。

關鍵字：旅行時間預測、梯度提升決策樹、資料融合

Predicting Freeway Travel Time by Using Gradient Boosting Decision Tree Method Through Data Merging

Student : Min-Wen, Li

Advisor : Jin-Yuan, Wang

Department of Transportation and Logistics Management National Yang
Ming Chiao Tung University

Abstract

Travel time is one of the most important travel information. If we can accurately predict the travel time, travelers can plan their schedule appropriately to avoid traffic jam, and the traffic management agencies can implement effective traffic management actions. The purpose of this research is to integrate multiple data and propose a machine learning model to predict highway travel time.

This study adopts the gradient boosting decision tree (GBDT) model as well as two proposed methods of fusing ETC and VD data to predict travel time. The empirical testing results show that the proposed GBDT model and data fusion method can improve the accuracy of travel time prediction by 82% to 97% compared with that of using only ETC data. It shows that the integration of multiple data sources can result in a better prediction accuracy.

Keywords: travel time prediction, gradient boosting decision tree, data fusion

目錄

誌謝	I
摘要	II
ABSTRACT.....	III
圖目錄.....	VI
表目錄.....	VII
第一章 緒論	1
1.1 研究背景與動機.....	1
1.2 研究目的.....	2
1.3 研究範圍.....	2
1.4 研究流程.....	2
第二章 文獻回顧.....	4
2.1 高速公路旅行時間預測回顧.....	4
2.2 機器學習用於交通預測.....	5
2.3 小結.....	8
第三章 研究方法	10
3.1 問題描述.....	10
3.2 資料處理.....	10
3.2.1 VD 資料處理	11
3.2.2 ETC 資料處理	13
3.3 資料融合.....	16
3.4 梯度提升決策樹（GRADIENT BOOSTING DECISION TREE, GBDT）	21
3.4.1 樹狀模型	21
3.4.2 梯度提升演算法（Gradient Boosting algorithm）	21
3.4.3 梯度提升決策樹演算法	25
3.5 本研究模式建構.....	27
第四章 實證研究.....	33
4.1 實證研究範圍.....	33
4.2 模式建立.....	34
4.3 預測結果與分析.....	37
4.4 延遲加總分析.....	40
4.4.1 製作時空表	40
4.4.2 延遲加總.....	42

4.4.3 分析結果.....	43
4.5 路段延伸分析.....	45
4.6 小結.....	56
第五章 結論與建議	57
第六章 參考文獻.....	59

圖目錄

圖 1.4.1 研究流程與架構圖	3
圖 3.1 研究方法流程圖	10
圖 3.2.1.1 VD 資料處理流程圖	12
圖 3.2.2.1 ETC 資料處理流程圖	14
圖 3.3.1 設備編號示意圖 (a)南下 (b)北上	16
圖 3.3.2 路段示意圖	17
圖 3.4.1 梯度提升演算法概念圖	21
圖 3.4.2.1 集成學習示意圖	22
圖 3.4.2.2 引導聚集演算法流程圖	22
圖 3.4.2.3 增強演算法流程圖	23
圖 3.5.1 模式設計流程圖	27
圖 3.5.2 五折交叉驗證	28
圖 4.2.1 路段切割示意圖	35
圖 4.3.1 北上平面路段預測結果	37
圖 4.3.2 南下平面路段預測結果	37
圖 4.3.3 北上高架路段預測結果	38
圖 4.3.4 南下高架路段預測結果	38
圖 4.4.1 交流道間旅行時間計算流程	40
圖 4.5.1 路段示意圖(含交流道)	45
圖 4.5.2 延伸路段示意圖	46
圖 4.5.3 各時段旅行時間結果	47

表目錄

表 2.3.1 機器學習優缺點	8
表 3.2.1.1 高公局 VD 五分鐘欄位說明	11
表 3.2.1.2 VD 資料示意表	12
表 3.2.1.3 VD 合併車道資料示意表	13
表 3.2.2.1 M04A 欄位說明	13
表 3.2.2.2 M05A 欄位說明	14
表 3.2.2.3 M04A 選取欄位	14
表 3.2.2.4 M05A 選取欄位	15
表 3.2.2.5 M04A 與 M05A 合併資料示意表	15
表 3.2.2.6 ETC 合併車種資料示意表	15
表 3.3.1 VD 資料示意表	17
表 3.3.2 ETC 資料示意表	17
表 3.3.3 資料融合方法一結果	18
表 3.3.4 資料融合方法二結果	19
表 3.3.5 單純 ETC 資料	20
表 3.4.2.1 損失函數種類	24
表 3.4.3.1 GBDT 參數	26
表 3.5.1 GBDT 模組重要參數	29
表 4.1.1 研究範圍設備統計	33
表 4.2.1 資料融合方法一輸入變數與預測目標	34
表 4.2.2 資料融合方法二輸入變數與預測目標	34
表 4.2.3 單純 ETC 資料輸入變數與預測目標	35
表 4.2.4 各研究路段切割路段數統計	36
表 4.2.5 參數校估結果（節錄）	36
表 4.4.1.1 原始時空表	40
表 4.4.1.2 時空表-設備合併後	41
表 4.4.1.3 時空表-時間問題	41
表 4.4.1.4 時空表-時間差補後	42
表 4.4.2.1 延遲加總示意表	42
表 4.4.3.1 ETC 設備至 ETC 設備誤差比較表	43
表 4.4.3.2 尖離峰誤差比較表	44
表 4.5.1 平面路段結果	47
表 4.5.2 北上平面路段統計資料	50
表 4.5.3 南下平面路段統計資料	51
表 4.5.4 高架路段結果	51
表 4.5.5 北上高架路段統計資料	55

表 4.5.6 南下高架路段統計資料55

第一章 緒論

1.1 研究背景與動機

在分秒必爭的忙碌社會，旅行時間是非常重要的交通資訊。許多人在出行前會習慣查詢旅行時間，評估後選擇能最快到達目的地的路線。尤其當有移動距離較長的城際旅運輸需求時，用路人通常會更早進行規劃。旅行時間資訊可以讓用路人更了解路況，預先規劃出發時間。

旅行時間的預測為十分重要的研究課題。若用路人無法事先預估旅行時間，為了避免遲到可能會過早出門造成時間浪費，亦可能過晚出門，使得無法準時抵達目的地。因此，提供準確的旅行時間資訊不但能降低道路交通負荷，亦能使用路人有更佳的出行體驗。

旅行時間的預測方法通常為用歷史資料進行預測，了解未來一段時間後某起點至迄點的旅行時間。預測結果可供用路人規劃何時出發，用最少的行車時間抵達目的地。例如，台灣高速公路總局推出之「行車時間預測」網站，民眾選擇日期、出發時間、起點與迄點交流道，即可得知預估的旅行時間。

進行旅行時間預測之研究多為單一資料來源（Goudarzi, 2018；Wu et al., 2004；黃冠維, 2018），然使用多資料來源可以降低風險，彌補資料天生的缺陷（翟雅嶠等人, 2009）。因此本研究希望能利用交通資料開放平台，結合多資料來源，藉以提升旅行時間預測的準確度。

以高速公路旅行時間預測為例，目前相關研究大多依賴車輛偵測器（Vehicle Detector，簡稱 VD）資料，雖然可以得到即時的行車速度，但設備一旦故障，容易產生嚴重偏誤。台灣高速公路電子道路收費系統（Electronic Toll Collection，簡稱 ETC）已於 2013 年全面啟用，不只能有效節省每位用路人的旅行時間，也使高速公路收費更為簡易公平。此外該系統全天候蒐集大量旅行時間的歷史資料，且準確率高。因此本研究希望能善用 ETC 資料，與 VD 資料，相互結合後，達到高準確度的旅行時間預測。

預測的方法有十分多種，其中利用機器學習進行預測，是目前的研究趨勢。人工智慧（Artificial Intelligence，簡稱 AI）一詞誕生於 1956 年的達特茅斯夏季人工智慧研究計畫（Dartmouth Summer Research Project on Artificial Intelligence），在經歷了多年的發展後，近年再次迎來研究熱潮。此熱潮因為硬體計算能力大幅提升，大數據資料研究突飛猛進，加上機器學習（Machine

Learning，簡稱 ML）與深度學習（Deep Learning，簡稱 DL）的概念先後被提出而出現。因此本研究希望能利用機器學習方法提升預測的準確度。

本研究期望發展一個有效的方法，融合多資料來源（VD 與 ETC），並運用機器學習方法建構高速公路旅行時間預測模型。本研究希望能彌補單一資料來源所會面臨的問題，並且使用大量數據訓練機器學習模型，經過參數校估，讓旅行時間預測能達到高準確度，以提升用路人的出行體驗。

1.2 研究目的

本研究目的旨在利用多資料來源建立機器學習模型，用以預測高速公路旅行時間。由於旅行時間是旅運者最在乎的旅行資訊之一，且過往研究中可以發現，使用多資料來源較單一資料來源更為可靠，可加強模型預測的準確度，最後加上近年來機器學習的蓬勃發展，使得預測準確度提升。因此本研究期望找出有效的資料結合方法，建立一套完整的資料庫。並以此資料庫為基礎，利用機器學習模型預測未來高速公路之旅行時間，建立旅行時間預測模式。

1.3 研究範圍

本研究選用同時具有 VD 與 ETC 設備之台灣高速公路作為研究範圍。由於本研究之資料來源為台灣高速公路總局的開放資料，因此研究路段需同時具備 VD 與 ETC 資料方能進行。目前台灣共有 12 條高速公路，其中國道一號平面路段、國道一號高架路段、國道三號與國道五號裝有 ETC 系統。

1.4 研究流程

根據研究目的與範圍，繪製本研究之研究流程如圖 1.4.1 所示，並詳細說明各步驟如下所列：

1. 文獻回顧

進行高速公路旅行時間預測之文獻回顧。首先了解國內外針對高速公路的旅行時間預測之研究方法。接著回顧多種用於預測的機器學習方法。整理總結後，供本研究後續參考。

2. VD 資料處理

從高速公路總局的開放資料中提取研究範圍的 VD 資料，對資料進行分析與過濾。

3. ETC 資料處理

從高速公路總局的開放資料中提取研究範圍的 ETC 資料，對資料進行分析與過濾。

4. 資料結合，建立資料庫

嘗試找出有效的資料結合方法，補足 VD 與 ETC 資料各自的缺點，並將此方法自動化，建立完整資料庫。

5. 模式設計

分別對單純 ETC 資料、單純 VD 資料、資料融合方法一與資料融合方法二，共四種資料庫，根據文獻回顧後之結果，設計機器學習模型，調整參數。

6. 實證研究

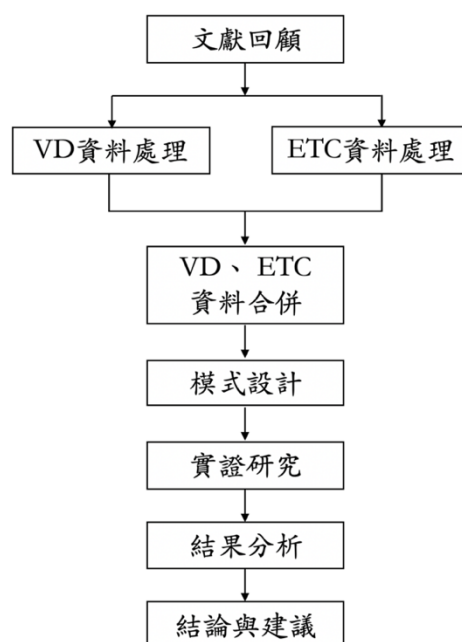
以台灣國道一號新竹交流道（含）以北之路段為研究對象，進行實證研究。從資料庫中提取歷史資料放入機器學習模型，並得出結果。

7. 分析結果

分析本研究模式之結果，與高速公路總局所開發的「行車時間預測」系統以及真實旅行時間進行三方比較。

8. 結論與建議

整理本研究之結論與貢獻，並提出建議供後續研究參考。



第二章 文獻回顧

本章節分為兩個部分進行文獻回顧：第一部分探討國內外高速公路旅行時間預測相關研究，並了解資料處理與結合的方法；第二部分回顧多種用於預測的機器學習方法，並比較優劣勢。最後，總結文獻回顧，做為本研究後續參考。

2.1 高速公路旅行時間預測回顧

Nanthawichit (2003) 提出一種處理探測車輛 (probe-vehicle) 和固定偵測器 (fixed-detector) 數據的方法，並用卡爾曼濾波器 (Kalman filter) 進行估計車流量、空間平均速度和密度等交通數據。利用模擬軟體 INTEGRATION 生成，研究路段為日本東京都會高速公路的一部分 (約 18,197 英尺)。結果顯示，在不同的情況下該融合方法均能有效的結合兩種資料，且加入探測車輛的預測更為準確。

何聰儒 (2009) 使用群集分析與線性迴歸方法，將車輛偵測器與探針車回傳之資料進行分析。以國道一號為實證研究對象，將車輛偵測器所蒐集之交通量與速度資料進行分群，再使用線性迴歸推估資料與探針車之間的關聯，最後以 ETC 蒐集之旅行時間資料當作實際旅行時間計算誤差，得出結果於平日誤差約 5~24%，於假日誤差約 1~14%。

Chang et al. (2016) 利用歷史資料預測未來一週高速公路的旅行時間。研究使用台灣高速公路總局 VD 與 ETC 資料，在 JAVA 環境中建立一套預測模型，並使用卡爾曼濾波器 (Kalman filter) 預測短期、傅里葉變換 (Fourier transform) 進行長期預測。最後與真實資料對比，結果顯示 MAPE 誤差不大於 10%。Chang 再進行分析時，首先對 VD 與 ETC 之原始資料進行調整與處理，然本研究希望能借由機器學習方法，在不更動設備原始資料的前提下直接進行分析。

Oh et al. (2018) 針對基於模式 (model-based) 的分析方法進行研究，並透過預測範圍、準確性、效率、適用性和穩健性進行評估，最後和數據驅動 (data-driven) 分析進行比較。結果表示基於模式的研究方法在研究範圍較長 (30~120 分鐘) 以及面對擁擠時有較好的表現，且能較數據驅動的分析方法有更好的效率。但基於模式的研究方法預先設定許多限制，導致當假設不成立時的預測能力大幅下降。

吳耿輝 (2018) 提出混合深度學習方法，利用高速公路 VD 資料與 CMS 自然語言格式資料，使高速公路旅行時間預測更為準確。首先利用卷積神經網路 (convolutional neural network, CNN) 處理時空資訊，再用自然語言處理中嵌入 (embedding) 技術處理非結構化的文字資料，如：交通資訊顯示板 (CMS)。結果顯示，比起單純的 Long Short-term memory 或 CNN 方法，該研究提出之提出混合深度學習方法皆有更準確的預測值。

Miao et al. (2020) 比較三個統計模型 (ARIMA、時空模型、向量自迴歸模型) 和三個機器學習模型 (支持向量機、反向傳播神經網路、多線性迴歸) 在預測旅行時間的表現。研究建立三個週期函數，分別對不同時間 (5 分鐘, 10 分鐘和 15 分鐘) 進行預測。使用美國休斯頓裡六個 AVI 旅行時間數據，路段總長約 12 英里。結果發現機器學習有較好的預測能力，且支持向量機、反向傳播神經網路的結果明顯較多線性迴歸佳。

2.2 機器學習用於交通預測

李開復於《AI 新世界》(2020) 指出目前人工智慧技術適合用來解決具與人互動低、高度重複性、勞力密集等特性的問題，且若問題本身具高度結構化與資料密集，人工智慧的應用就可以充分達到它的價值。

人工智慧適合應用在交通產業上。交通運輸相關規劃與管理針對的是運具、號誌、道路設施等，較不需與人面對、交流，屬於社交性弱的產業。且交通決策行為每天不斷重複發生，例如：號誌時制的調整，每個路口每個週期皆在重複；大眾運輸業的排班調度亦是每日循環重複，因此具有重複性高的特性。再加上臺灣地區交通繁忙，每日皆產生大量資料，包含 ETC、路面偵測器、閉路電視、信令資料等，屬於資料密集產業。因此交通產業是適合仰賴機器學習進行分析的。

以下將回顧四種目前常用於交通預測的機器學習方法：

1. 神經網路 (Neural network, 簡稱 NN)

神經網路為人工智慧的一環，學習人類神經元的傳遞方式，期望電腦能完成達到且超過人腦的正確決策。將大量的資料輸入，經過多層神經元的計算以及反向傳播 (Backward-Propagation) 的優化，最後輸出結果。目前已發展出許多方法，並且被運用在預測，例如：RNN、LSTM、CNN、DBN。

Duan et al. (2016) 使用長短期記憶 (Long Short-Term Memory, LSTM) 深度學習建立旅行時間預測模型。研究中旅行時間資料來自英國公

路局。模式結果有很好的準確率，作者認為考慮時間序列的深度學習方法適合用於在交通分析。

Goudarzi (2018) 利用人工神經網路 (Artificial Neural Network, ANN) 進行旅行時間預測。研究使用 Google Maps Application Programming Interface (API) 收集路段旅行時間數據。將結果與最近鄰近 (Nearest Neighbor, NN)，窗口最近鄰近 (Windowed Nearest Neighbor, WNN)、線性迴歸進行比較，顯示人工神經網路 ANN 預測準確度最高。

黃冠維 (2018) 使用台灣桃園地區市區道路的車輛偵測器資料，利用卷積神經網路建立路況預測模型。預測研究路段未來 5 至 30 分鐘的行車速度，進而了解未來路況。研究針對不同預測範圍、空間範圍、歷史資料長度、預測時階長度，設計五個群組以了解每個變因的影響，結果顯示該模型在大多情況下有良好預測結果。

2. 支撐向量機 (Support Vector Machine, 簡稱 SVM)

SVM 是一種監督式機器學習。此方法以統計學為基礎，期望在一群資料的多種特徵中找出最佳決策邊界 (decision boundary) 使不同類別資料間的邊界最大化。分類完成後，即可透過特徵，預測新資料屬於哪一類。

Wu et al. (2004) 使用支持向量迴歸 (Support Vector Regression, 簡稱 SVR) 方法進行旅行時間預測，並將結果與實際交通數據進行比較。作者利用高速公路總局鋪設的線圈資料，間接得到旅行時間。作者找出一組能有效預測旅行時間的 SVR 參數，結果顯示在時間序列問題中 SVR 有良好的表現。

Vanajakshi et al. (2007) 使用美國德克薩斯州聖安東尼奧市交通管理中心的數據，比較 SVM 與 ANN 在短期旅行時間預測的準確率。用過去四天資料當作訓練集，將過去十分鐘的資料輸入，預測未來兩分鐘到一小時的旅行時間。結果顯示，SVM 與 ANN 都有不錯的表現，但相對於 ANN，SVM 對訓練資料量較不敏感，因此當訓練數據不足時，SVM 能有更好的表現。

3. 樹狀模型 (Tree-based)

樹狀模型的機器學習方法是監督學習的一種。由兩個部分構成：分支 (branches) 和節點 (nodes)。從樹的頂端開始，比較不同特徵，利用分支為資料進行標籤。此集成方法相較於其他機器學習方法，可以在不需過多資料預處理下，提供可解釋的結果，並指出變量的重要性與相關性，因此被廣泛運用在預測領域 (Zhang et al., 2015)。

Zhang et al. (2015) 使用梯度提升迴歸樹法 (gradient boosting regression tree method, GBM) 進行旅行時間預測。作者提出了使用 GBM 的特點：首先，在交通狀態改變 (從變順暢為壅擠) 時，仍可以正常處理資料；第二，此法可以自動選擇相關變量，擬合模型和調整參數，但當複雜度上升時須考量計算成本；此外，也具備了可解釋結果的優勢。本文最後將建置好的 GBM 模型與 ARIMA 和隨機森林 (Random forest, RF) 進行比較，結果顯示 RF 與 GBM 模型均優於傳統統計模型 ARIMA。

Li et al. (2016) 蒐集貨運公司每天收集的大量軌跡數據，使用梯度提升決策樹 (GBDT) 模型與貝葉斯優化 (Bayesian optimization) 進行旅行時間預測。結果表明，預測準確性達到 80% 以上。另外，發現若加入前 5 分鐘的即時平均速度，可以提高準確度。若加入更多即時的平均速度，即使發生異常事件時，預測準確度也能提升約 2%。

Fan et al. (2018) 使用台灣高速公路電子收費站的數據來預測高速公路行駛時間。為處理不同交通狀況，作者在 Apache Hadoop 平台上開發出兩種利用隨機森林構建大數據分析機器學習。在正常交通狀況時使用一站式旅行時間預測 (One-destination travel time prediction, 簡稱 OTTP)，計算方法為每小時在每個門架之間產生隨機森林。在交通壅擠時使用自適應旅行時間預測 (Adaptive travel time prognosis, 簡稱 ATTP) 配對即時交通流量，對 OTTP 進行修正。

Cheng et al. (2018) 利用模擬軟體 VISSIM 取得道路交通量、車輛類型、速度分配、車輛比例和旅行時間等交通資訊，利用梯度提升決策樹模式預測未來的旅行時間。作者將建構好的模式與反向傳播神經網絡和支持向量機模型比較，結果顯示 GBDT 模型能更準確的預測旅行時間。但由於資料來源為模擬數據，因此有其先天限制性。

4. k-Nearest Neighbor (k-NN)

k-Nearest Neighbor 是一種監督式機器學習。透過計算不同特徵值之間的距離進行分類。其中 k 值可決定與鄰近多少個資料進行分析，因此 k 值為一個重要參數。

朱志杰(2013)利用台灣國道 VD 以及自動車輛辨識 (Automatic Vehicle Identification, 簡稱 AVI) 資料，進行旅行時間預測。研究使用國道一號台中-新竹、新竹-台北南北向為對象，使用 K-NN 方法比對交通狀況，接著校估參數以提升整體預測準確度。研究結果證實加入 AVI 資料可以提升預測準確度。

劉軒寧(2016)利用台灣國道 VD 資料透過變動 k 值之 k -NN 法進行旅行時間預測。研究範圍為國道一號新竹至台北北上路段，並僅取用每個星期五當作分析對象。結果顯示相較於固定 k 值的 k -NN 法，尤其在車流量較大的時後，此研究有更高的準確率。但由於資料量較少，因此建議後續可考慮增加資料量。

2.3 小結

綜合以上，本研究希望能使用多資料來源進行高速公路旅行時間預測，並且發展一套合適的資料融合方法，能對台灣的高速公路旅行時間預測有所貢獻。

由於交通資料每日蒐集，累積龐大的歷史資料，因此已有不少研究運用機器學習進行旅行時間的預測。相較於傳統統計研究，機器學習方法能自我學習，因此大多有較佳的預測表現，但同時機器學習方法需要校估許多參數，不同的輸入資料特徵也容易造成變異。四種機器學習方法各有優缺點，如下表 2.3.1 所示。

表 2.3.1 機器學習優缺點

方法	優點	缺點
NN	● 影像處理佳。 ● 能處理複雜非線性關係。 ● 延伸出許多種類，可處理不同問題。	● 需校估大量參數。 ● 輸出結果難以解釋。 ● 若學習時間長，容易陷入區域最佳解，而非整體最佳解。
k-NN	● 對異常值較不敏感。 ● 可用於非線性分類。	● 計算時間成本大。 ● 資料不平衡時誤差大。 ● K 值選擇影響大。
SVM	● 適合解決高維度問題。	● 效率較低。 ● 對缺失資料敏感。 ● 無法解決非線性問題。
Tree-based	● 可處理迴歸與分類問題。 ● 可了解變數重要程度。 ● 多棵樹加權穩健度較高。	● 需校估參數。 ● 需避免過度適配。

樹狀模型（Tree-based）方法能增加模式穩健度，且在變數較多的情況下亦能針對各變數進行分析，了解各變數對整體模式的重要比例。由於本研究除了期望能有效預測高速公路旅行時間，也希望能了解特徵變數對於預測值的重要程度，並加以解釋，對於未來的交通研究有所貢獻。因此本研究選用樹狀模型方法進行建模，將於第三章詳細介紹研究方法。

第三章 研究方法

本章節依照研究流程詳述本研究的研究方法，共分為五個部分，如圖 3.1 所示。第一部分為本研究之問題描述；第二部分為資料處理，詳細敘述本研究對 VD 與 ETC 兩資料集的各別處理方法；第三部分為資料融合方法，提出本研究發展之資料融合方法；第四部分為本研究採用之樹狀模型方法，梯度提升決策樹演算法之介紹；最後，第五部分則詳述本研究的梯度提升決策樹模式建構之方法與流程。

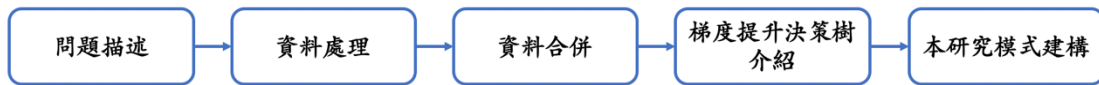


圖 3.1 研究方法流程圖

3.1 問題描述

本研究欲探討高速公路旅行時間之預測。期望融合高速公路總局之 VD 與 ETC 歷史資料，建立資料庫，接著建立一個機器學習模型，預測未來高速公路旅行時間。

3.2 資料處理

本研究使用交通部數據匯流平台之開放資料進行分析。選用以下三種資料做為本研究分析使用：高公局 VD 五分鐘統計資料、高公局 eTag_M04A（以下簡稱 M04A）與高公局 eTag_M05A（以下簡稱 M05A）。

由於本研究之目的為利用機器學習方法進行多資料來源之旅行時間預測，因此本研究在清洗資料時，將不會對原始資料進行插補。原因為考量到高速公路上的 VD 偵測器設備與 ETC 門架維修不易，本研究期望能在最小幅度的資料清洗下，使用原始設備所蒐集之資料，利用機器學習模型進行有效的預測。

本節將分成兩個部分，第一部分說明 VD 資料的處理方式，第二部分說明 ETC 資料的處理方式。

3.2.1 VD 資料處理

根據交通部數據匯流平台網站中，「高公局 VD 五分鐘統計資料」之詮釋資料，可得知該資料欄位如下表 3.2.1.1 所示：

表 3.2.1.1 高公局 VD 五分鐘欄位說明

欄位名稱	欄位說明
VdID	VD 設備代碼
LinkID	基礎路段代碼
LaneID	車道代碼
LaneType	車道種類
Speed	平均速率偵測值(單位:kph)
DataCount	合格資料筆數
SVolume	小車流量
FixedSVolume	統計修正後 小車流量
TVolume	卡車聯結車車流量
FixedTVolume	統計修正後 連結車流量
LVolume	大車流量
FixedLVVolume	統計修正後 大車流量
MVolume	機車流量
FixedMVolume	統計修正後 機車流量
TotalVolume	總流量
FixedTotalVolume	統計修正後 總流量
Occ	佔有率(單位:%)
K	平均車流密度
Q	平均流量
V	平均速度
Datacollecttime	資料蒐集時間
UpdateTime	本平台資料更新時間(yyyy-MM-ddHH:mm:ss)
InfoDate	資料代表日期(yyyy-MM-dd)

VD 資料處理流程圖如下圖 3.2.1.1 所示，詳細說明各步驟如下所列：

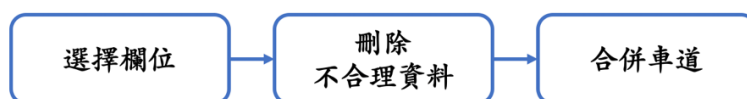


圖 3.2.1.1 VD 資料處理流程圖

1. 選擇欄位

首先選擇資料收集時間（Datacollecttime）、VD 設備代碼（VdID）、車道代碼（LaneID）欄位。

接著，流量部分有車種流量欄位（TotalVolume）與統計修正後的欄位（FixedTotalVolume），由於各車種原始流量欄位有小於零的不合理的資料，導致 TotalVolume 欄位可靠度下降，因此本研究選用高公局計算之統計修正後的欄位進行分析。速度部分則選用平均速率偵側值（Speed）進行分析。

另外，由於佔有率（Occ）、平均車流密度（K）、平均流量（Q）與平均速度（V）為經過計算後的資料，非原始蒐集資料，因此本研究不採用。

2. 刪去不合理資料

將流量小於 0 之不合理資料予以刪除。由於車流量不可能為負值，因此本研究將發生此現象資料刪除。

3. 合併車道

在相同時間（Datacollecttime）下，相同 VD 設備代碼（VdID）之不同車道（LaneID）資料進行合併計算，如下表 3.2.1.2 所示範例。

表 3.2.1.2 VD 資料示意表

Datacollecttime	VdID	LaneID	FixedTotalVolume	Speed
2020-06-0100:05:00	nfbVD-N1-N-1.950-M-LOOP	0	8	103
2020-06-0100:05:00	nfbVD-N1-N-1.950-M-LOOP	1	5	98
2020-06-0100:05:00	nfbVD-N1-N-1.950-M-LOOP	2	3	100

速度資料依照流量加權平均進行合併，並重新命名為「VD 速度」欄位，以表 3.2.1.2 為例，VD 速度為 $(103*8+98*5+100*3)/(8+5+3) = 100.875$ 。總交通量則為各車道之加總，並重新命名為「VD 交通量」欄位，以表 3.2.1.2 為例，VD 交通量為 $8+5+3 = 16$ 。完成後刪去車道（LaneID）欄位，結果如下表 3.3.1.3 所示。

表 3.2.1.3 VD 合併車道資料示意表

Datacollecttime	VdID	VD 交通量	VD 速度
2020-06-0100:05:00	nfbVD-N1-N-1.950-M-LOOP	16	100.875

3.2.2 ETC 資料處理

由於本研究之目的為預測平均旅行時間，因此在 ETC 資料中選用具平均旅行時間的 M04A 資料以及具平均車速資料之 M05A 資料。根據交通部數據匯流平台所公告 M04A 與 M05A 詮釋資料，可得欄位如下表 3.2.2.1 與 3.2.2.2 所示：

表 3.2.2.1 M04A 欄位說明

欄位名稱	欄位說明
TimeInterval	報表產製時間; 依每 5 分鐘為時階統計產出報表
GantryFrom	上游偵測站編號
GantryTo	下游偵測站編號
VehicleType	車種：31(小客車)、32(小貨車)、41(大客車)、42(大貨車)、5(聯結車)
TravelTime	平均旅行時間，依車種統計相鄰上下游偵測站之平均旅行時間，計算單位：秒
Volume	交通量
Authority	資料來源單位
SrcUpdateTime	來源端平台資料更新時間(yyyy-MM-dd HH:mm:ss)
UpdateTime	本平台資料更新時間(yyyy-MM-dd HH:mm:ss)
InfoTime	資料代表時間(yyyy-MM-dd HH:mm:ss)
InfoDate	資料代表日期(yyyy-MM-dd)

表 3.2.2.2 M05A 欄位說明

欄位名稱	欄位說明
TimeInterval	報表產製時間; 依每 5 分鐘為時階統計產出報表
GantryFrom	上游偵測站編號
GantryTo	下游偵測站編號
VehicleType	車種：31(小客車) 、32(小貨車) 、41(大客車) 、42(大貨車) 、5(聯結車)
SpaceMeanSpeed	平均車速，計算單位：Km/Hr
Volume	交通量
Authority	資料來源單位
SrcUpdateTime	來源端平台資料更新時間(yyyy-MM-dd HH:mm:ss)
UpdateTime	本平台資料更新時間(yyyy-MM-dd HH:mm:ss)
InfoTime	資料代表時間(yyyy-MM-dd HH:mm:ss)
InfoDate	資料代表日期(yyyy-MM-dd)

ETC 資料處理流程圖如下圖 3.2.2.1 所示，詳細說明各步驟如下所列：



圖 3.2.2.1 ETC 資料處理流程圖

1. 選取欄位

選取 M04A 與 M05A 共同欄位： TimeInterval、GantryFrom、GantryTo、VehicleType 與 Volume，為下一步之合併用。另外選取 M04A 的 TravelTime 欄位；M05A 的 SpaceMeanSpeed 欄位，為後續分析資料用。如下表 3.2.2.3 與 3.2.2.4 所示範例。

表 3.2.2.3 M04A 選取欄位

TimeInterval	GantryFrom	GantryTo	VehicleType	Volume	TravelTime
2020-06-01 00:05:00	01F-002.9N	01F-001.7N	31	7	45
2020-06-01 00:05:00	01F-002.9N	01F-001.7N	32	10	44

表 3.2.2.4 M05A 選取欄位

TimeInterval	GantryFrom	GantryTo	VehicleType	Volume	SpaceMeanSpeed
2020-06-01 00:05:00	01F-002.9N	01F-001.7N	31	7	95
2020-06-01 00:05:00	01F-002.9N	01F-001.7N	32	10	98

2. 合併資料

利用 M04A 與 M05A 之共同欄位，進行兩資料的合併。當共同欄位皆相同，表示為相同時間相同門架所蒐集的資料，以表 3.2.2.3 以及 3.2.2.4 的資料為例，其合併結果如表 3.2.2.5 所示。

表 3.2.2.5 M04A 與 M05A 合併資料示意表

TimeInterval	GantryFrom	GantryTo	VehicleType	Volume	TravelTime	SpaceMeanSpeed
2020-06-01 00:05:00	01F-002.9N	01F-001.7N	31	7	45	95
2020-06-01 00:05:00	01F-002.9N	01F-001.7N	32	10	44	98

3. 合併車種

將相同時間且相同門架區間的資料進行車種合併。以表 3.2.2.5 為例，由於在 2020-06-01 00:05:00 時間下，從代號 01F-002.9N 門架出發至代號 01F-001.7N 門架的資料中，有兩不同車種（VehicleType）類別，分別為代號 31 與 32。為配合 VD 資料，本研究將 ETC 資料進行車種合併。

合併方法如下：首先，將各車種交通量相加成為最終總流量，並重新命名為「ETC 交通量」，以此例為 $7+10=17$ 。接著，速度資料則將不同車種之值依照交通量之加權平均進行合併，並重新命名為「ETC 速度」，以此例為 $(95*7+98*10)/(7+10) = 96.76$ 。最後，旅行時間資料則依照不同車種之交通量加權平均進行合併，並重新命名為「ETC 旅行時間」，以此例為 $(45*7+44*10)/(7+10) = 44.4$ 。刪除車種（VehicleType）欄位後，最終結果如下表 3.2.2.6 所示。

表 3.2.2.6 ETC 合併車種資料示意表

TimeInterval	GantryFrom	GantryTo	ETC 交通量	ETC 旅行時間	ETC 速度
2020-06-01 00:05:00	01F-002.9N	01F-001.7N	17	44.4	96.76

3.3 資料融合

為了進行資料融合，本研究首先對 VD 與 ETC 設備進行編號。依照經緯度依序編號，如下圖 3.3.1 所示。藍線表示 ETC 門架位置；橘點表示 VD 設備位置，橘點右側之南 01 等編號為 VD 設備之名稱。依照行駛方向，南下者，由北而南編號；北上者，則由南而北編號（圖 3.3.1 為示意圖，非實際情況）。

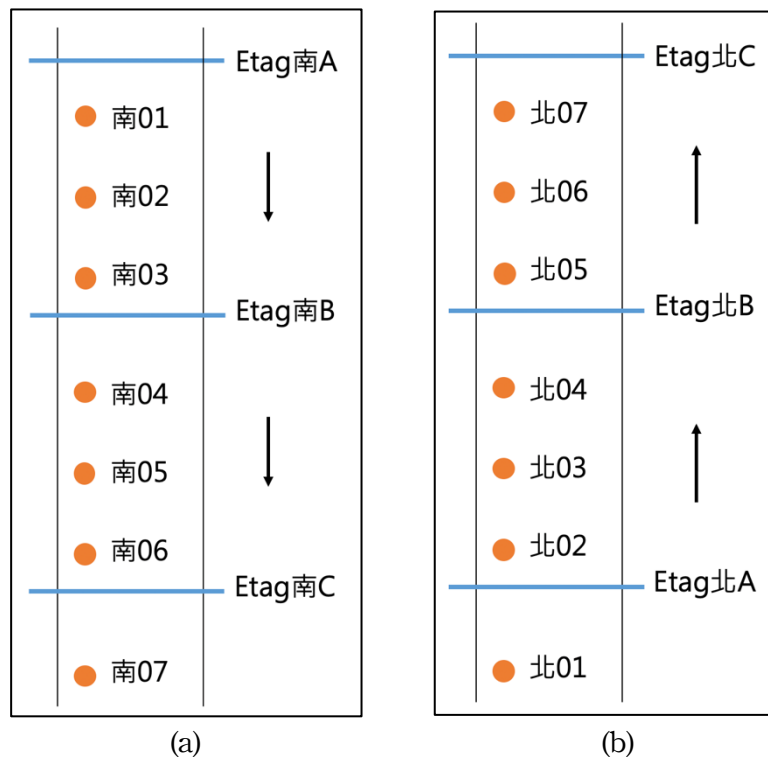


圖 3.3.1 設備編號示意圖 (a)南下 (b)北上

資料融合旨在增加分析資料之可用性，進而使預測更準確。Chang et al. (2016) 中提及 ETC 資料雖有較高的準確度，但收費器中間缺乏空間訊息；而 VD 設備佈設較為密集，使空間資料較齊全，但因為設備老舊而有許多資料遺失與缺漏的問題。兩種資料各有優缺點，本研究期望透過資料融合，使兩資料能互相彌補缺點，讓預測更準確。

為解決兩資料來源的問題，本研究開發了兩種融合方法。圖 3.3.2 為路段示意圖，此圖以南下路段為例，藍色線為 ETC 門架位置，橘色點為 VD 設備位置，而綠色字所示 m, n, o, p 表示為設備間的行駛距離。

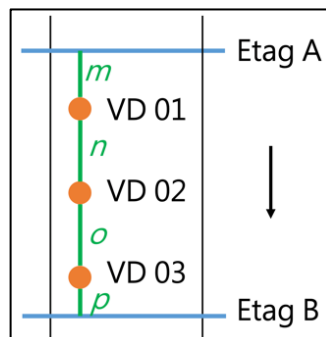


圖 3.3.2 路段示意圖

下表 3.3.1 為 VD 資料示意表，Datacollecttime 欄位為資料蒐集之時間，以五分鐘為單位，VdID 欄位為 VD 設備代碼，VD 交通量與 VD 速度欄位分別為交通量與速度整理後的數據。

表 3.3.1 VD 資料示意表

Datacollecttime	VdID	VD 交通量	VD 速度
2020-06-01 00:05:00	VD01	d	g
2020-06-01 00:05:00	VD02	e	h
2020-06-01 00:05:00	VD03	f	i

下表 3.3.2 為 ETC 資料示意表，TimeInterval 欄位為資料蒐集之時間，GantryFrom 為起點門架代碼，GantryTo 為迄點門架代碼，ETC 交通量、ETC 速度和 ETC 旅行時間欄位分別為交通量、速度和旅行時間整理後的數據。

表 3.3.2 ETC 資料示意表

TimeInterval	GantryFrom	GantryTo	ETC 交通量	ETC 速度	ETC 旅行時間
2020-06-01 00:05:00	EtagA	EtagB	a	b	c

ETC 資料為收費門架起迄點（圖 3.3.2 為例，起點為 EtagA，迄點為 EtagB）的路段資料，而 VD 資料則為偵測器所在位置（為圖 3.3.2 中 VD01、VD02 與 VD 03）之點資料，此因素為兩資料融合時所會面對之最大挑戰。

下文將利用圖 3.3.2 路段示意圖、表 3.3.1 VD 資料示意表與表 3.3.2 ETC 資料示意表為例，逐一舉例說明並介紹本研究所使用之資料融合方法：

1. 資料融合方法一

本融合方法之想法為以 ETC 資料為基礎，加入兩 ETC 門架間所有 VD 設備（以本例為 VD01、VD02、VD03）之平均訊息。

在相同時間下（2020-06-01 00:05:00），方法一「VD_交通量」欄位為 VD 資料示意表中「VD 交通量」的平均值，以此例為 $\frac{d+e+f}{3}$ ；而方法一「VD_速度」則為 VD 資料示意表中「VD 速度」資料平均值，其餘原始 ETC 資料則維持不變。如下表 3.3.3 所示。

表 3.3.3 資料融合方法一結果

日期	時間	上游	下游	ETC 交通量	ETC 速度	VD_交通量	VD_速度	ETC 旅行時間
2020-06-01	00:05:00	EtagA	EtagB	a	b	$\frac{d+e+f}{3}$	$\frac{g+h+i}{3}$	c

2. 資料融合方法二

由於方法一可能造成兩 ETC 門架間的 VD 資料因為進行平均而失去多元性，因此方法二的想法為將 VD 資料轉換成路段資料，並與 ETC 資料進行融合。

融合後的結果，第一筆為 EtagA 作為起點，VD01 作為迄點，第二筆則是 VD01 作為起點，VD02 作為迄點，以此類推。融合後之平均交通量與平均速度為該起迄點原始資料之平均值，以第一筆資料為例，平均交通量為 ETC 交通量與 VD01 交通量 d 之平均值 $\frac{a+d}{2}$ ，平均速度則為 ETC 速度 b 與 VD01 速度 g 之平均值 $\frac{b+g}{2}$ ，以此類推。由於路段被切割，因此捨棄原始 ETC 資料中的旅行時間欄位，改利用起迄點距離與平均速度相除之結果作為每筆資料之平均旅行時間，以第一筆資料為例，平均旅行時間為 EtagA 與 VD01 之間的距離 m 除以平均速度 $\frac{b+g}{2}$ ，以此類推。最終資料融合方法二結果如表 3.3.4 所示。

表 3.3.4 資料融合方法二結果

日期	時間	上游	下游	平均交通量	平均速度	平均旅行時間
2020-06-01	00:05:00	EtagA	VD01	$\frac{a+d}{2}$	$\frac{b+g}{2}$	$\frac{m}{\frac{b+g}{2}}$
2020-06-01	00:05:00	VD01	VD02	$\frac{d+e}{2}$	$\frac{g+h}{2}$	$\frac{n}{\frac{g+h}{2}}$
2020-06-01	00:05:00	VD02	VD03	$\frac{e+f}{2}$	$\frac{h+i}{2}$	$\frac{o}{\frac{h+i}{2}}$
2020-06-01	00:05:00	VD03	EtagB	$\frac{f+a}{2}$	$\frac{i+b}{2}$	$\frac{p}{\frac{i+b}{2}}$

3. 單純 ETC 資料

為了證明本研究所開發資料融合方法是否有效，因此本研究亦將利用原始的資料進行建模。單純 ETC 資料如表 3.3.5 所示。

表 3.3.5 單純 ETC 資料

TimeInterval	上游	下游	ETC 交通量	ETC 速度	ETC 旅行時間
2020-06-01 00:05:00	EtagA	EtagB	a	b	c

4. 單純 VD 資料

由於原始 VD 資料為點資料形式，無旅行時間之欄位，若利用距離為 VD 資料加上旅行時間欄位，則有違單純使用 VD 資料之原則。且由於 VD 設備位置無法與其他三種資料形式對應，因此本研究不採用單純 VD 資料。

本研究將針對資料融合方法一、資料融合方法二，與單純 ETC 資料，共三種資料，分別建立模型，並且相互比較、分析結果。

3.4 梯度提升決策樹 (Gradient Boosting Decision Tree, GBDT)

本研究使用梯度提升決策樹進行建模，本節將循序漸進地介紹此方法。梯度提升決策樹顧名思義為「梯度提升」與「決策樹」之結合，如下圖 3.4.1 概念圖所示。

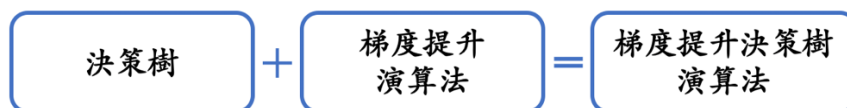


圖 3.4.1 梯度提升演算法概念圖

本節分為三個部分，第一部分介紹樹狀模型的基本概念與決策樹，第二部分介紹梯度提升的原理與計算方法，第三部分則介紹「梯度提升」與「決策樹」結合後的梯度提升決策樹模型。

3.4.1 樹狀模型

樹狀模型是一種探究資料的監督式機器學習方法。監督式學習的訓練資料可表示為： $\{(X^1, y^1), (X^2, y^2), (X^3, y^3) \dots (X^N, y^N)\}$ ，其中 N 表示訓練資料總筆數； X^i 表示每筆資料的特徵值； y^i 表示目標值，也就是預測的標的。

樹狀模型從頂端的根 (root) 開始，透過節點 (node) 的變數產生分支 (branches)。節點繼續往下分支直到無法再分支或達停止條件後終止，該最後的節點稱為葉節點 (leaf node)。因此，在建構決策樹時通常需要進行修剪 (pruning)，以避免過多分支產生過度擬合的問題。每筆資料進入完成的決策樹後，只會對應一條路徑與一個葉節點。

目前已發展出許多以樹狀模型為基礎之模式，如決策樹 (Decision Tree)、隨機森林 (Random Forest)、條件推論樹 (Conditional Inference Tree)、梯度提升決策樹等。

樹狀模型可以解決迴歸以及分類問題。當目標變數為一定性變數時，決策樹即為分類樹 (classification tree)；而目標變數為定量變數時，決策樹則為迴歸樹 (regression tree)。本研究目標為預測出旅行時間，而旅行時間屬於連續變數，因此選用迴歸樹方法進行。

3.4.2 梯度提升演算法 (Gradient Boosting algorithm)

梯度提升演算法為利用集成學習 (ensemble learning) 找到最佳解的一種方法。集成學習是由多個較弱的學習器結合而成的學習法，如圖 3.4.2.1 所示。相

較於一般的模型屬於單一的學習器，集成學習則是由多個較簡單的學習器所組成。根據 Yang et al. (2017)，集成學習的優點是可以獲得比單一學習器更好的預測結果。

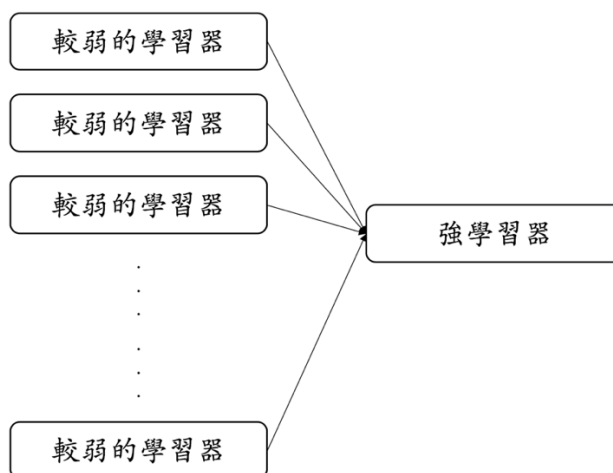


圖 3.4.2.1 集成學習示意圖

集成學習可分為兩大類：引導聚集演算法（Bootstrap Aggregating，簡稱 Bagging）與增強演算法（Boosting）。以下將分別敘述：

1. 引導聚集演算法

首先，對原始資料進行隨機抽樣，抽出 n 筆（ n 為自訂）資料成為訓練資料集。重複對原始資料進行隨機抽樣，直到獲得 K 份（ K 為自訂）訓練資料集。由於每個訓練集皆為原始資料的隨機抽樣，在 K 份訓練資料集中，可能有原始資料從未被選取，也可能重複被選取。接著，對 K 份訓練資料集各自進行建模，獲得 K 個模型與 K 個預測值。最後，若是迴歸問題則將 K 個預測值做簡單平均獲得最終結果；若是分類問題，最終結果則是 K 個預測值中的多數解。如圖 3.4.2.2 所示。

引導聚集演算法加上決策樹模型即為常見的隨機森林（Random forest）方法。

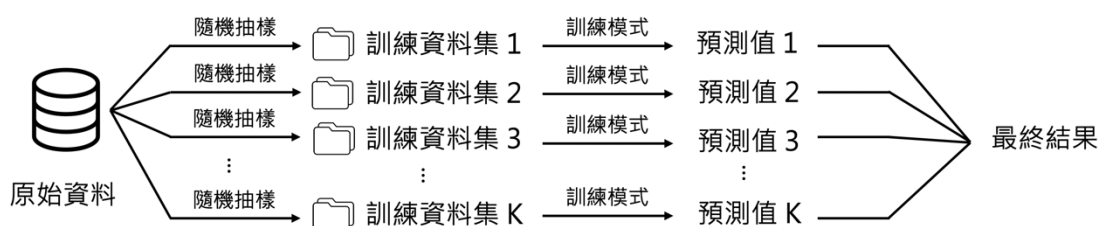


圖 3.4.2.2 引導聚集演算法流程圖

2. 增強演算法

首先，從原始資料中抽取 n 筆資料，形成訓練資料集，並建立第一個模型，獲得該模型的預測值。接著根據該模型的預測結果調整訓練資料集，形成更新後的訓練資料集。

接著利用此更新後的訓練資料集建立第二個模型，並獲得該模型的預測值。接著根據該模型的預測結果調整訓練資料集，再一次形成更新後的訓練資料集。重複此步驟 M 次（ M 為自訂）後，將每此的預測值進行處理獲得最終結果。如圖 3.4.2.3 所示。

從上述流程敘述可以發現，增強演算法模型之間互相相關，每個模型皆是建立在上一個模型的結果之下，且此法須決定如何調整訓練資料集、訓練模式的方法以及如何將每一次的預測值計算成最終結果。常見的有自適應增強演算法（Adaptive Boosting，簡稱 Adaboost）和本研究使所用之梯度提升決策樹。

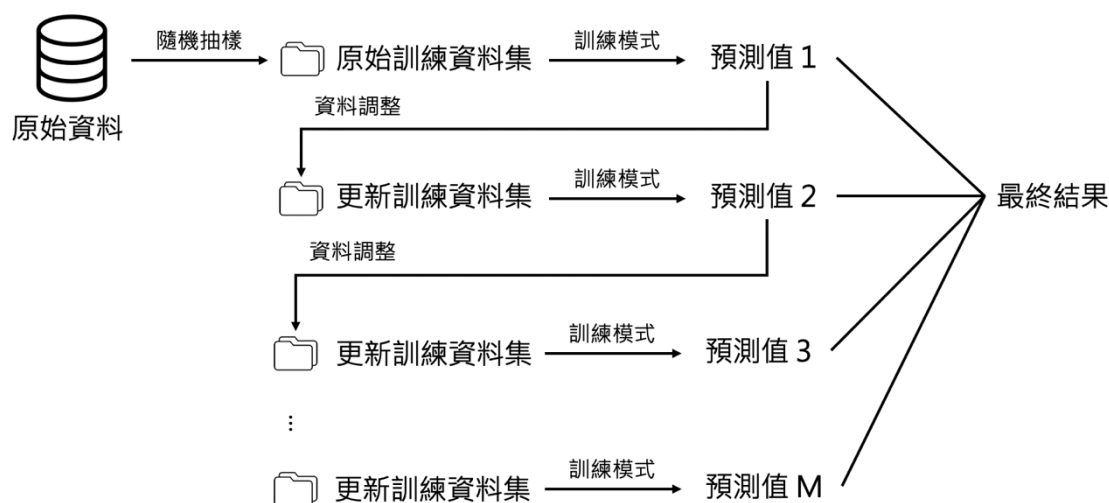


圖 3.4.2.3 增強演算法流程圖

本研究所使用之梯度提升演算法目標是利用訓練資料，找出特徵值與目標值之間的函數關係 $F(X)$ 。為求出最佳解，梯度提升演算法使用損失函數的概念來評估每個模型之預測結果。損失函數 $L(y, F(x))$ 可以代表模式的可信程度，當損失函數越小，表示模式越精確。負梯度向量可以理解成結果與真實答案的殘差值，並以此殘差為依據，調整下一個模型的訓練資料樣本。目前已發展出許多計算損失函數的方法，下表 3.4.2.1 節錄自 Cheng et al. (2018) 中常應用於計算損失函數與其負梯度值：

表 3.4.2.1 損失函數種類

名稱	損失函數	負梯度向量
Squared loss function	$\frac{1}{2}[y_i - f(x_i)]^2$	$y_i - f(x_i)$
Absolute loss function	$ y_i - f(x_i) $	$\text{sign}[y_i - f(x_i)]$
Huber-M loss function	$\begin{cases} \frac{1}{2}[y_i - f(x_i)]^2, & y_i - f(x_i) \leq \delta_m \\ \delta_m \left(y_i - f(x_i) - \frac{1}{2}\delta_m \right), & y_i - f(x_i) > \delta_m \end{cases}$	$\begin{cases} y_i - f(x_i), & y_i - f(x_i) \leq \delta_m \\ \delta_m \text{sign}[y_i - f(x_i)], & y_i - f(x_i) > \delta_m \end{cases}$

由於負梯度向量值可視為模式與目標值之間的殘差值，因此梯度提升演算法在調整完訓練資料後，即為在學習上個模式預測結果與目標值之間的殘差，所以梯度提升演算法在計算最終結果時，使用的方法為將每一次的預測值加總。

以較簡單的例子說明：如果希望利用梯度提升演算法學習受測者的體重，而受測者真實體重為 60 公斤。建立第一個模型後得到預測值為 53 公斤，從此預測值可得殘差值為 $60-53=7$ 公斤，因此第二個模型的目標則是學習 7 公斤。接者建立第二個模型後得到預測值為 7 公斤，從此預測值可得殘差值為 $7-7=0$ 公斤，表示已學習完畢。兩次預測值的總和 $53+7$ 則為最終的預測值 60 公斤。

3.4.3 梯度提升決策樹演算法

梯度提升決策樹演算法即為「梯度提升」與「決策樹」之結合。結合方法為：將決策樹做為 3.2.2 節梯度提升演算法中所提到之模型，並選用適當之損失函數後，進行梯度提升演算法之計算。

建立 GBDT 模型時需進行參數校估，表 3.2.3.1 為建立 GBDT 所需之參數，並詳細說明如下所列：

1. 單一決策樹中節點個數 (J)

由於 GBDT 是以多個決策樹為基礎的改良模型，因此 GBDT 限制每棵決策樹有相同的節點數，以避免每棵樹過度生長或修剪。對於不同問題， J 的值可以透過實驗找出，在 Cheng et al. (2018) 中建議 J 為 5~8。

2. 學習率 (η)

學習率可視為搜尋可行解時，模型沿著損失梯度下降的幅度，其值為介於 0、1 之間的數值。當學習率越小，表示損失梯度每下降一小幅度即進行一次搜尋，雖可避免錯過最佳解，但所需時間即被拉長，且可能出現過度擬合情形。反之當學習率越大，雖可節省學習時間，但有較大的機會錯過最佳解，產生學習不完整的情況。設定學習率的目的是為了防止模式過度擬合。學習率與決策樹個數 (M) 需達到平衡。通常會建議使用較大的決策樹個數，搭配較小的學習率 ($\eta < 0.1$)。

3. 子樣本比例 (frd)

子樣本比例為另一種防止過度擬合的參數，子樣本比例為介於 0、1 之間的數值。使用引導聚集演算法的隨機森林模型，在隨機抽樣子樣本時，使用可重複選取的方式，如 3.4.2 節中所述。而 GBDT 在採樣子樣本時，採取不重複選取的方式，即只採用一定比例的原始資料進行模式計算，以防止模式過度擬合。當 frd 等於 1，表示所有樣本皆納入；當 frd 小於 1，則表示只有該比例的樣本納入子樣本。 frd 建議設定在 0.5~0.8 之間。

4. 屬性採樣個數 (S_a)

在構建模型的過程中，每個節點都可以選擇用多少個屬性進行模式建構。

5. 決策樹個數 (M)

在 GBDT 中，每次迭代皆會產生一棵決策樹，因此決策樹個數又可解釋為迭代次數。決策樹個數越大，雖可以提升準確度，但訓練時間將會拉長。

表 3.4.3.1 GBDT 參數

參數	解釋
J	單一決策樹中節點個數
η	學習率
frd	子樣本比例
S_a	屬性採樣個數
M	決策樹個數

3.5 本研究模式建構

本研究所使用之程式環境為 python spyder 4.1.5，梯度提升決策樹模式建構則選用 sklearn 套件中的 GradientBoostingRegressor 模組。下圖 3.5.1 為本研究模式設計之流程，並將逐一說明。

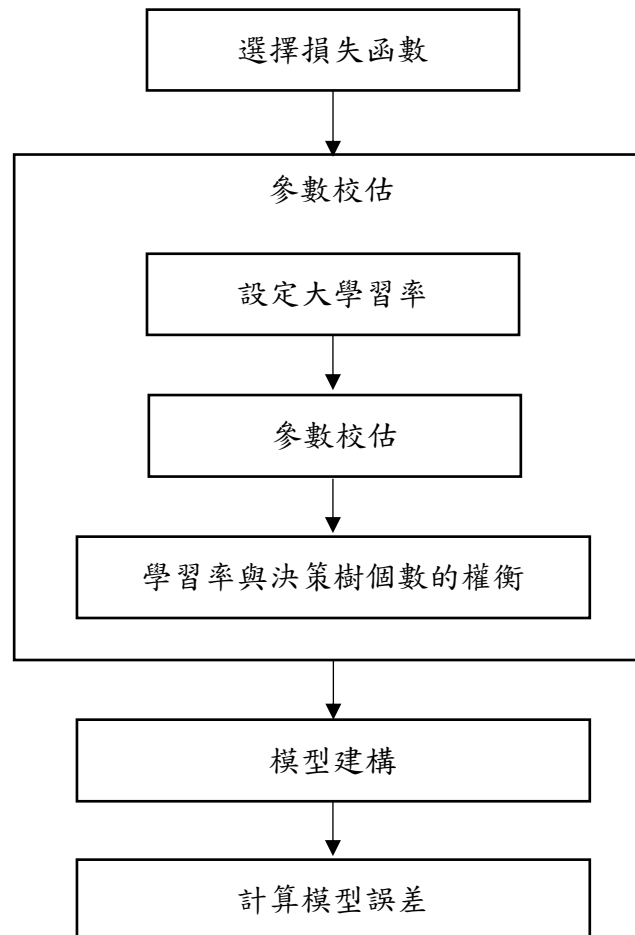


圖 3.5.1 模式設計流程圖

1. 選擇損失函數

GBDT 模式需選擇一種損失函數進行學習器的評估。根據 Cheng et al. (2018)，此方法中最常使用的最小平方迴歸（least squares regression）法計算損失函數。此法將預測值與正確值的誤差平方進行平均，可用以估計誤差之平均值。

2. 參數校估

校估參數時，本研究使用交叉驗證（cross validation）增加模型穩定性。依據交叉驗證之折數，可稱為 n 折交叉驗證，較常使用的為三折、五折與十折。

以五折交叉驗證為例，如圖 3.5.2 所示，首先將訓練資料分成五等分，並選擇其中一份（編號五）作為驗證集，其餘四份做為訓練集，並求得準確率 E_1 。接著選擇另一份（編號四）作為驗證集，其餘四份做為訓練集，並求得準確率 E_2 。依此類推共執行五次，最後將五次準確率進行平均，其值即為此次交叉驗證之準確率。

訓練資料	1	1	1	1	1
	2	2	2	2	2
	3	3	3	3	3
	4	4	4	4	4
	5	5	5	5	5
每折準確率	E_1	E_2	E_3	E_4	E_5
整體準確率	$E = \frac{1}{5} \sum_{i=1}^5 E_i$				

圖 3.5.2 五折交叉驗證

下表 3.5.1 為 GBDT 所需校估之重要參數表，並附上在研究所使用之 GradientBoostingRegressor 模組中對各個參數的默認值，以下進行詳細的解釋：

- (1) 學習率：為 0~1 之間的數值，學習率過大易造成訓練不完整；但當學習率過小時，訓練時間會被拉長，且可能出現過度擬合。模組中默認值為 0.1。
- (2) 決策樹個數：整個模型中所使用之決策樹總個數。須與學習率達到權衡，以防止過度擬合。模組中默認值為 100。
- (3) 決策樹的最大深度：限制每棵決策樹的可生長的最大節點數量，防止決策樹過度生長產生過度擬合問題。模組中默認值為 3。

- (4) 節點所需最少樣本數：形成一個節點時所需的最少樣本數，低於此樣本數即無法產生新節點。模組中默認值為 2。
- (5) 葉節點所需最少樣本數：形成一個葉節點所需的最少樣本數，當低於此樣本數即無法形成一個葉節點。模組中默認值為 1。
- (6) 最大屬性採樣數：每個節點在進行預測時，選擇之屬性個數。此數值小於等於所有屬性數。模組中默認值為屬性個數值。
- (7) 子樣本數：採樣子集合時所採取的樣本比例，當值等於 1 時，代表採納所有樣本。模組中默認值為 1。

表 3.5.1 GBDT 模組重要參數

變量	解釋	默認值
學習率 (learning_rate)	為 0~1 之間的數值。	0.1
決策樹個數 (n_estimators)	整個模型中所使用之決策樹總個數。	100
決策樹的最大深度 (max_depth)	限制節點的數量，以防止過度擬合。	3
節點所需最少樣本數 (min_samples_split)	一個節點所需最少樣本數。	2
葉節點所需最少樣本數 (min_samples_leaf)	一個葉節點所需最少樣本數。	1
最大屬性採樣數 (max_features)	每個節點選擇之屬性個數。	等於所有屬性數
子樣本數 (subsample)	採樣子集合時採取的樣本比例。	1

本研究搭配使用 sklearn 中 GridSearchCV 套件，進行最佳參數搜尋，由於本研究使用之模型共需校估七個參數，若是同時進行所需耗費的時間成本過大，因此本研究首先設定大學習率，並依序進行個別參數校估，最後進行學習率與決策樹個數的調整：

- (1) 設定大學習率：由於決策樹個數 (n_estimators) 與學習率 (learning_rate) 之間存在權衡關係。決策樹個數較大時，學習率

較低，反之決策樹個數較少時，須提高學習率。為防止過度擬合，校估時會先以較大的學習率配上較小的決策樹個數進行。

- (2) 參數校估：利用 GridSearchCV 套件並設定十折交叉驗證進行最佳參數校估。此部分共需校估六個參數，本研究第一步先確定總共使用多少棵決策樹（決策樹個數），第二步則限制每棵決策樹的生長節點數（決策樹的最大深度），第三步為同時限制每個節點與葉節點至少需要多少樣本數（節點所需最少樣本數、葉節點所需最少樣本數），第四步決定每棵決策樹建模時採用的屬性個數（最大屬性採樣數），透過以上四個步驟完成決策樹的定義。最後則決定建模時使用的資料比例（子樣本數）。以下將進行個步驟的說明：

第一步、搜索決策樹個數

首先將決策樹個數外其於所有參數設定為默認值。進行最佳決策樹個數值的搜尋。

第二步、搜索決策樹的最大深度

將決策樹個數設定為第一步所得之結果，接者將除了決策樹的最大深度外其於所有參數設定為默認值。進行最佳決策樹的最大深度值的搜尋。

第三步、搜索節點所需最少樣本數與葉節點所需最少樣本數

由於節點所需最少樣本數與葉節點所需最少樣本數兩參數皆會影響決策樹的分支情況，因此需要同時進行校估。

將決策樹個數設定為第一步所得之結果，決策樹的最大深度為第二步所得之結果。接著將除了節點所需最少樣本數以及葉節點所需最少樣本數外其餘參數設定為默認值。同時進行節點所需最少樣本數以及葉節點所需最少樣本數值的搜尋。

第四步、搜索最大屬性採樣數

將決策樹個數設定為第一步所得之結果，決策樹的最大深度為第二步所得之結果，節點所需最少樣本數和葉節點所需最少樣本數則為第三步所得之結果。接著將除了最大屬性採樣數外其餘參數設定為默認值。進行最佳最大屬性採樣數值的搜尋。

第五步、搜索子樣本數

將決策樹個數設定為第一步所得之結果，決策樹的最大深度為第二步所得之結果，節點所需最少樣本數和葉節點所需最少樣本數為第三步所得之結果，最大屬性採樣數則為第四步所得之結果。接著進行最佳子樣本數值的搜尋。

- (3) 學習率與決策樹個數的權衡：完成參數校估後，將所得之參數結果帶入模型，並嘗試將學習率下降，決策樹個數提升，其餘參數不變，找到學習率與決策樹個數的權衡。

3. 模型建構

利用參數校估之結果建立 GBDT 模型。

4. 計算模型誤差

將模型所預測之結果與測試資料進行比較計算誤差。以下為常見之誤差計算方法。

y ：測試資料之值

$\hat{y}$ ：模型所預測之結果

n ：資料筆數

- (1) 均方誤差(MSE, Mean Square Error)

$$MSE(y, \hat{y}) = \frac{1}{n} \left(\sum_{i=1}^n (y_i - \hat{y}_i)^2 \right)$$

- (2) 均方根誤差(RMSE, Root Mean Square Error)

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \left(\sum_{i=1}^n (y_i - \hat{y}_i)^2 \right)}$$

- (3) 平均絕對誤差(MAE, Mean Absolute Error)

$$MAE(y, \hat{y}) = \frac{1}{n} \left(\sum_{i=1}^n |y_i - \hat{y}_i| \right)$$

(4) 平均絕對百分比誤差(MAPE, Mean absolute percentage error)

$$MAPE(y, \hat{y}) = \frac{1}{n} \left(\sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right) \times 100$$

第四章 實證研究

本章節將分為五個部分：第一部分為實證研究範圍，第二部分為針對四個實證研究路段分別進行建模，第三部分則為分析結果，第四部分為延遲加總分析，而第五部分則是為了增加本研究之應用性，因此進行路段延伸之分析。

4.1 實證研究範圍

本研究設定實證研究之範圍為國道一號新竹交流道（含）以北之雙向路段（0~98k），且包含五股楊梅高架與汐止五股高架。因此可將實證研究範圍細分為四個路段，分別為：北上平面高速公路、南下平面高速公路、北上高架路段與南下高架路段。

本研究利用過去三個月的歷史資料（2020 年 6 月～2020 年 8 月）進行分析，預測未來一個月（2020 年 9 月）內每日的旅行時間。

VD 與 ETC 設備統計如，表 4.1.1 所示。南下路段中，ETC 設備共 35 個，包含 7 個位於高架路段；VD 設備共 211 個，包含 57 個位於高架路段。北上路段中，ETC 設備共 36 個，包含 8 個位於高架路段；VD 設備共 207 個，包含 53 個位於高架路段。

表 4.1.1 研究範圍設備統計

設備種類	路段	南下（個）	北上（個）
ETC	平面路段	28	28
	高架路段	7	8
	總計	35	36
VD	平面路段	154	154
	高架路段	57	53
	總計	211	207

4.2 模式建立

本研究依照 3.3 節中所述資料融合方法將資料進行處理後分別建模。由於原始 VD 資料無法在相同機種下進行比較，因此本研究的實證研究共使用三個資料庫，分別為資料融合方法一、資料融合方法二與原始 ETC 資料，進行模式建構。由於 GBDT 為監督式學習，因此需決定輸入變數與預測目標，以下逐一說明：

以 ETC 代號 01H-020.0N 至 01H-017.4N 路段為例，該路段中有三個 VD 設備，分別為代號 nfbVD-N1H-N-19.511-M-RS、nfbVD-N1H-N-19.000-N-LOOP 與 nfbVD-N1H-N-18.060-M-LOOP。

1. 資料融合方法一

資料處理完成之結果為下表 4.2.1，黃色底之欄位為輸入變數：ETC 交通量、ETC 速度、VD_交通量與 VD_速度。橘色底則為預測目標欄位：ETC 旅行時間。

表 4.2.1 資料融合方法一輸入變數與預測目標

日期	時間	上游	下游	ETC 交通量	ETC 速度	VD_交通量	VD_速度	ETC 旅行時間
2020-06-01	00:05:00	01H-020.0N	01H-017.4N	21	97	42	107	95

2. 資料融合方法二

資料處理完成之結果為下表 4.2.2，黃色底之欄位為輸入變數：平均交通量與平均速度。橘色底則為預測目標欄位：平均旅行時間。

表 4.2.2 資料融合方法二輸入變數與預測目標

日期	時間	上游	下游	平均交通量	平均速度	平均旅行時間
2020-06-01	00:05:00	01H-020.0N	nfbVD-N1H-N-19.511-M-RS	32	102.77	17.12
2020-06-01	00:05:00	nfbVD-N1H-N-19.511-M-RS	nfbVD-N1H-N-19.000-N-LOOP	40	107.74	17.07
2020-06-01	00:05:00	nfbVD-N1H-N-19.000-N-LOOP	nfbVD-N1H-N-18.060-M-LOOP	41	107.56	31.45
2020-06-01	00:05:00	nfbVD-N1H-N-18.060-M-LOOP	01H-017.4N	34	102.59	23.15

3. 單純 ETC 資料

資料處理完成之結果為下表 4.2.3，黃色底之欄位為輸入變數：平均交通量與平均速度。橘色底則為預測目標欄位：平均旅行時間。

表 4.2.3 單純 ETC 資料輸入變數與預測目標

日期	時間	上游	下游	ETC 交通量	ETC 速度	ETC 旅行時間
2020-06-01	00:05:00	01H-020.0N	01H-017.4N	21	97	95

為了使三種資料間可以互相比較，本研究將四個研究路段依照 ETC 所在位置切割成數個路段，並分別依序編號。如圖 4.2.1 所示，橘色點為 VD 設備位置，其中南 01~南 06 為其編號；藍色直線則為 ETC 設備所在位置。依照 ETC 設備所在位置將路段切割，左側 A、B 為切割之路段編號。下表 4.2.4 為各研究路段切割路段數統計。

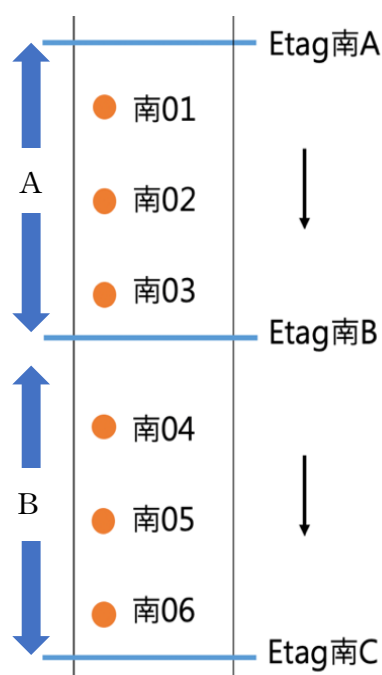


圖 4.2.1 路段切割示意圖

表 4.2.4 各研究路段切割路段數統計

研究路段	北上平面	南下平面	北上高架	南下高架
切割路段數	27	27	7	6

將切割完成之路段依照 3.4 節所定義之流程進行參數校估，下表 4.2.5 為結果之節錄，以原始 ETC 資料中路段編號 A 為例。

表 4.2.5 參數校估結果（節錄）

資料	原始 ETC 資料			
路段	北上	北上高架	南下	南下高架
A	n_estimators =120 max_depth =4 min_samples_split =60 min_samples_leaf =1 max_features =2 subsample =0.7 learning_rate =0.05	n_estimators =30 max_depth =6 min_samples_split =20 min_samples_leaf =1 max_features =2 subsample =0.7 learning_rate =0.1	n_estimators =70 max_depth =3 min_samples_split =130 min_samples_leaf =5 max_features =2 subsample =0.55 learning_rate =0.1	n_estimators =140 max_depth =4 min_samples_split =30 min_samples_leaf =1 max_features =2 subsample =0.75 learning_rate =0.05

從上述節錄結果中可發現，不同研究路段之參數值皆所差異。將三種資料完成參數校估後，進行 GBDT 模式建立，接著將預測值與實際值進行誤差計算，結果於 4.3 節中呈現。

4.3 預測結果與分析

本研究將三種資料分別對四個實證研究路段進行 GBDT 模式建構，並且將旅行時間預測值與旅行時間正確值進行誤差計算。

下圖 4.3.1 至 4.3.4 為各研究路段之預測結果，其中橫軸為切割路段編號，縱軸為誤差（MAPE）值，藍線表示方法二，橘線表示方法一，綠線則表示原始 ETC 資料。

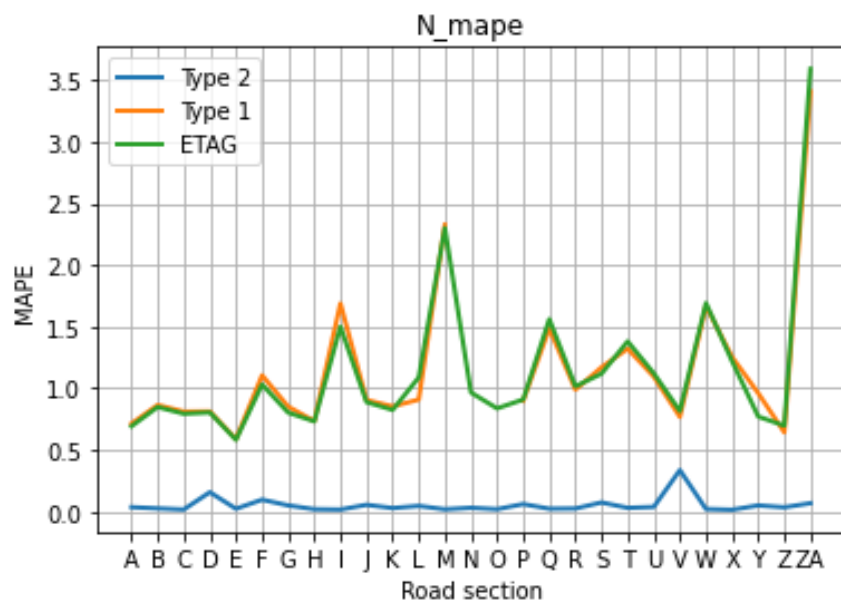


圖 4.3.1 北上平面路段預測結果

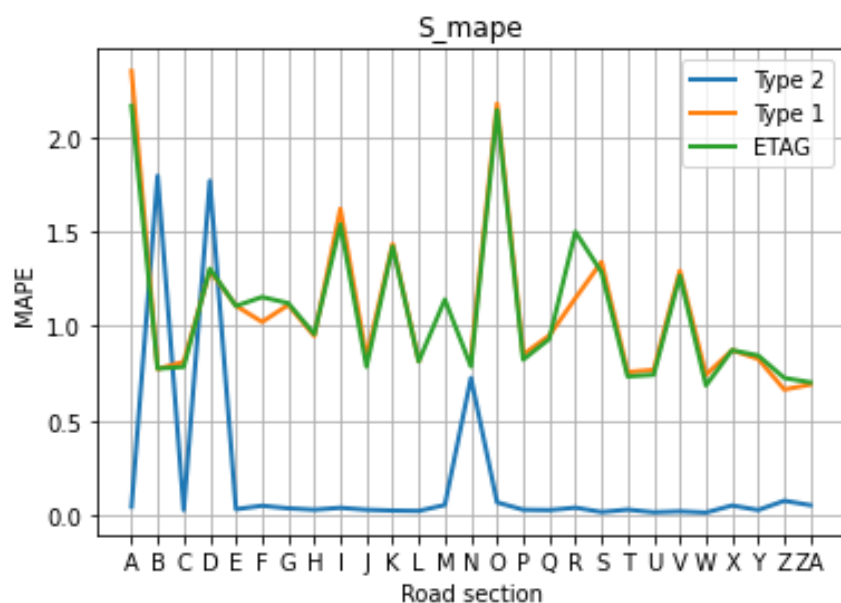


圖 4.3.2 南下平面路段預測結果

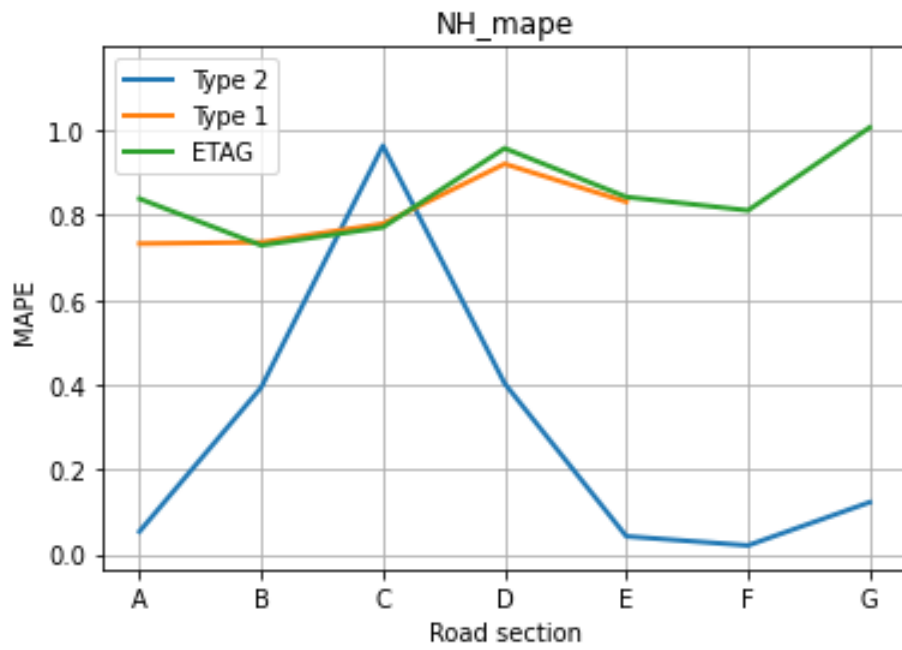


圖 4.3.3 北上高架路段預測結果

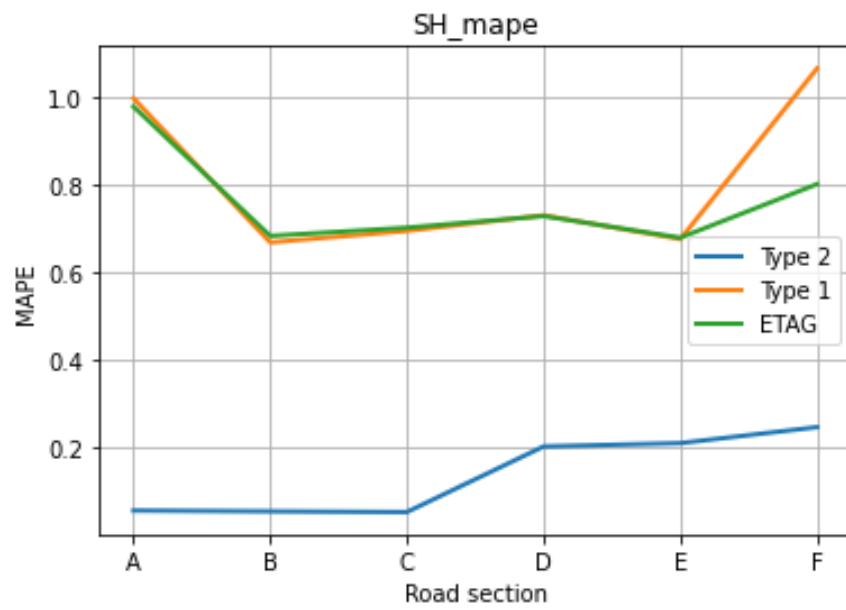


圖 4.3.4 南下高架路段預測結果

由上述四張圖可知以下幾點：

1. 資料融合方法一與原始 ETC 資料預測誤差結果相近。表示資料融合方法一並沒有將 VD 設備資料的優勢展現。由於此融合方法直接將 ETC 間所有 VD 設備資料合併，使 VD 設備的多樣性下降，導致結果與原始 ETC 資料相近。
2. 資料融合方法二明顯有較好的預測準確度。由圖可知，資料融合方法二的預測誤差較低，相較之下能更準確的預測旅行時間。
3. 資料融合方法二有些許資料誤差較大。少部分路段，如南下平面高速公路中 B、D 路段、北上高架高速公路中 C 路段之誤差值較高，原因為 VD 設備資料不齊全，有些設備出現無回傳資料的狀況，導致在預估時部分路段會出現較高的誤差。未來可針對此現象將該部分路段的 VD 資料重新審視，找出有問題之 VD 設備，提供交通主管單位進行維修。
4. 資料融合方法一有資料缺少的問題。如北上平面高速公路 O 與 P 路段、南下平面高速公路 N 路段和北上高架高速公路中 G 路段，原因為該路段內 VD 設備因故障或其他因素無回傳資料，又因為 GBDT 為監督式模型，當路段建模時輸入變數有缺少即無法進行模式建構。

由於此分析之每個路段距離較短，為每個 ETC 設備至下個 ETC 設備，為了能使三種資料的比較更有意義，本研究透過延遲加總的方法，將分析路段拉長至研究路段上的第一個 ETC 設備至最後一個 ETC 設備，並於 4.4 節將詳細說明。

4.4 延遲加總分析

由於本研究所使用之機器學習方法為監督式學習的一種，須在有正確資料的情況下方能進行預測，如資料有所缺漏則無法產生預測結果。本研究發現四種路段於9月9、10、11、12、18、25、26日，由於VD資料缺漏，導致全日皆無法產生旅行時間預測資料。由於相同路段的車輛行為具相似的型態，並以一週為單位，因此本研究選用2020年9月1日至2020年9月7日連續七天做為此分析的資料來源。

本節將依照圖4.4.1之流程進行延遲加總，圖中各步驟將於下文依序說明。

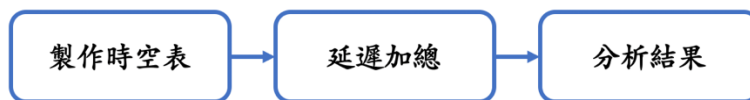


圖 4.4.1 交流道間旅行時間計算流程

4.4.1 製作時空表

下表4.4.1.1為原始時空表，此表以資料融合方法二之北上高架路段為例，每一列代表設備至設備的預測旅行時間，欄位則為出發時間(五分鐘為間隔)。

由表可知，在2020/9/1 3:20時設備編號19因故無回傳資料，因此該時段僅有設備18至設備20的預測值，而沒有設備18設備至19設備以及設備19至設備20的預測值；而2020/9/1 3:30時設備編號15因故無回傳資料，因此該時段僅有設備14至設備16的預測值，而沒有設備14設備至15設備以及設備15至設備16的預測值。

表 4.4.1.1 原始時空表

	2020/9/1 3:15	2020/9/1 3:20	2020/9/1 3:25	2020/9/1 3:30	2020/9/1 3:35	2020/9/1 3:40
14-15	1.181452021	1.185817107	1.317999698		1.214623703	1.261641138
14-16				2.325968208		
15-16	0.882233372	0.885271236	0.909648006		0.877615524	0.873404786
16-17	0.829970783	0.820378935	0.799321758	0.818657555	0.838374137	0.783000179
17-18	0.711561034	0.742151493	0.753498454	0.711344636	0.769966699	0.682209864
18-19	1.164092244		1.288786326	1.206540998	1.237467734	1.170563777
18-20		2.006587777				
19-20	0.756448405		0.77621568	0.734426171	0.770215272	0.7737927
20-21	0.419752945	0.432979677	0.386400995	0.392136019	0.382784222	0.40839152

為解決設備問題，本研究將對原始時空表進行合併。以上表為例，將設備 14 至設備 15 的旅行時間與設備 15 至設備 16 旅行時間加總，形成設備 14 至設備 16 的旅行時間，以解決設備 15 資料缺少的問題。合併後的結果如下表 4.4.1.2 所示。

表 4.4.1.2 時空表-設備合併後

	2020/9/1 3:15	2020/9/1 3:20	2020/9/1 3:25	2020/9/1 3:30	2020/9/1 3:35	2020/9/1 3:40
14-16	2.06368539	2.07108834	2.2276477	2.32596821	2.09223923	2.13504592
16-17	0.82997078	0.82037893	0.79932176	0.81865756	0.83837414	0.78300018
17-18	0.71156103	0.74215149	0.75349845	0.71134464	0.7699667	0.68220986
18-20	1.92054065	2.00658778	2.06500201	1.94096717	2.00768301	1.94435648
20-21	0.41975294	0.43297968	0.38640099	0.39213602	0.38278422	0.40839152

將設備合併後，發現有少部分資料缺漏，如下表 4.4.1.3 所示。設備 34 至設備 35 於 2020/9/1 4:05 預測值為空值，由於本研究使用的機器學習模式為的監督式學習的一種，此學習法僅能在有正確解答的環境下進行預測，因此若原始資料中有缺漏，即會導致無法預測的情況發生。

表 4.4.1.3 時空表-時間問題

	2020/9/1 3:55	2020/9/1 4:00	2020/9/1 4:05	2020/9/1 4:10
34-35	0.65857403	0.66338177		0.7397996
35-37	1.014690843	1.03324561	1.11865563	1.16242482
37-39	0.473661009	0.47057112	0.47329224	0.49484515
39-40	0.301250993		0.30370356	0.31997745
40-41	0.293193621		0.33235156	0.322015
41-42	0.521463422	0.56270144	0.59838685	0.54776339
42-43	0.392077247	0.40145744	0.4015961	0.39823275

為解決時間問題，本研究將對完成設備合併的時空表進行差補。以上表為例，2020/9/1 4:05 時設備 34 至設備 35 的旅行時間計算方法為相同設備間前後五分鐘(2020/9/1 4:00 與 2020/9/1 4:10)的旅行時間平均。差補後結果如下表 4.4.1.4 所示。

表 4.4.1.4 時空表-時間差補後

	2020/9/1 3:55	2020/9/1 4:00	2020/9/1 4:05	2020/9/1 4:10
34-35	0.65857403	0.66338177	0.701590685	0.7397996
35-37	1.014690843	1.03324561	1.11865563	1.16242482
37-39	0.473661009	0.47057112	0.47329224	0.49484515
39-40	0.301250993	0.302477276	0.30370356	0.31997745
40-41	0.293193621	0.312772590	0.33235156	0.322015
41-42	0.521463422	0.56270144	0.59838685	0.54776339
42-43	0.392077247	0.40145744	0.4015961	0.39823275

本研究利用上述合併與差補原則，個別完成資料融合方法一、資料融合方法二與單純 ETC 資料對四個研究路段之時空表。

4.4.2 延遲加總

完成時空表建置後，利用時空表進行延遲加總，以計算出各時間出發的旅行時間。利用下表 4.4.2.1 進行延遲加總說明，每一列代表起迄點的預測旅行時間(以分鐘為單位)，欄位則為出發時間(五分鐘為間隔)。

假設目標起迄點分別為 A 點與 E 點，出發時間為 00:00，由表可知 A 點到 B 點需花費 3.5 分鐘，因小於五分鐘的時間間隔，因此下一個路段 B 點至 C 點依然使用 00:00 之旅行時間。抵達 C 點時的總旅行時間為 5.2 (3.5+1.7)分鐘，超過了五分鐘的時間間隔，因此下一個路段 C 點至 D 點則使用 00:05 的旅行時間。抵達 D 點時的總旅行時間為 7.7 (3.5+1.7+2.5)分鐘，因小於五分鐘的時間間隔，因此下一個路段 D 點至 E 點依然使用 00:05 之旅行時間。由此可得出抵達 E 點時的總旅行時間為 9.8 (3.5+1.7+2.5+2.1)分鐘。以此類推即可求出各時間出發之總旅行時間。

表 4.4.2.1 延遲加總示意表

	00:00	00:05	00:10	00:15
A-B	3.5	4.1	3.9	3.7
B-C	1.7	1.5	1.9	1.6
C-D	2.7	2.5	2.8	2.6
D-E	1.9	2.1	2.0	1.9

4.4.3 分析結果

下表 4.4.3.1 為四種實證研究路段中第一個 ETC 設備至最後一個 ETC 設備旅行時間誤差值比較表。

表 4.4.3.1 ETC 設備至 ETC 設備誤差比較表

		MSE	RMSE	MAE	MAPE
北上平面路段	ETC	2556.79258	50.5647366	23.5201491	0.531269
	方法一	—	—	—	—
	方法二	21.0437518	4.58734693	0.58575021	0.013299
北上高架路段	ETC	113075.625	336.267194	62.7984181	1.667876
	方法一	—	—	—	—
	方法二	326.596068	18.0719691	1.0456158	0.053378
南下平面路段	ETC	9067.61777	95.2240399	28.9998028	0.621986
	方法一	—	—	—	—
	方法二	3123.13607	55.8850255	5.721027	0.111418
南下高架路段	ETC	443.504074	21.0595364	7.49878406	0.456308
	方法一	192.841252	13.8867293	6.98705862	0.429097
	方法二	54.6520053	7.39269946	0.37878799	0.021655

由於在 4.3 節之分析結果中發現資料融合方法一有資料卻漏的問題發生，因此資料融合方法一在北上平面路段、北上高架路段與南下平面路段三個實證研究路段無法完成時空表時，因此上表中以「—」表示。

從上表可發現，在四種誤差計算方法下，資料融合方法二皆相較原始 ETC 資料有較小的誤差。以 MAPE 值來說，資料融合方法二較單純 ETC 資料減少了 82%~97% 的誤差值。

為了了解資料融合方法二在本研究所發展的模式是否具有尖離峰的預測準確度差異，因此本研究針對尖峰時段（本研究定義早尖峰為 7 點~10 點出發，午尖峰為 16 點~20 點出發）與離峰時段（除了尖峰時段，其餘時間皆為離峰時段）分別進行旅行時間預測準確度分析，MAPE 誤差結果如下表 4.4.3.2 所示：

表 4.4.3.2 尖離峰誤差比較表

	尖峰時段	離峰時段
北上平面路段	0.012837	0.013578
北上高架路段	0.041121	0.060756
南下平面路段	0.058577	0.142079
南下高架路段	0.006614	0.030708

由上表可知，資料融合方法二在本研究所發展之模式下，尖峰時段的誤差值較離峰時段低，表示此模式在尖峰時段亦有良好的預測表現。

使用本研究所發展之資料融合方法二，利用 GBDT 模式建模，能有效降低旅行時間預測的誤差，此外在尖峰時段亦有良好的預測結果，因此該資料結合方法為有效的。為了增加模式的應用性，本研究將利用此資料融合方法二進行路段延伸之分析。

4.5 路段延伸分析

對於一般用路人，交流道至交流道的旅行時間才是常用且有意義的資訊。然而為了進行各融合方法之比較，上述實證研究所計算之旅行時間為 ETC 設備至 ETC 設備，對於用路人而言此資訊較不直觀。

在 4.4 節實證研究結果中可發現，利用 GBDT 模式進行高速公路旅行時間預測，在本研究所發展之資料融合方法二下，可以有效降低四個路段的旅行時間預測誤差。

綜合上述，為了使本研究之可應用性提升，本節將利用資料融合方法二，進行路段延伸，將路段從 ETC 設備至 ETC 設備，拉長至交流道與交流道之間。如下圖 4.4.1 所示，此路段示意圖中，黃色方塊為交流道位置，橘色圓點為 VD 設備位置，藍色直線為 ETC 設備位置。

實證研究中所計算的旅行時間路段為 ETC 設備(藍線)之間，以圖 4.5.1 為例，為 EtagA 至 EtagB 以及 EtagB 至 EtagC。如此計算之結果對於用路人較無幫助，因此本節欲將路段改為交流道至交流道。

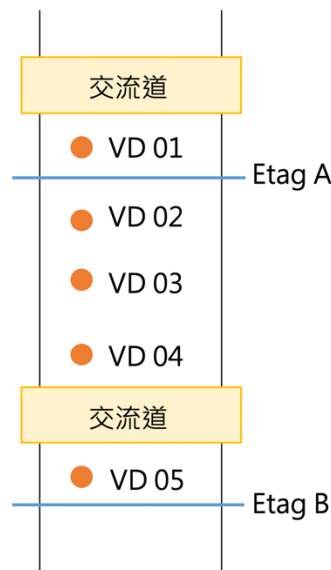


圖 4.5.1 路段示意圖(含交流道)

本研究期望進行未來任意時間自研究路段內任意交流道出發，開往任意交流道之旅行時間預測。由於起迄組合數量眾多，加上出發時間為每五分鐘均可進行預測，十分密集，因此本研究以四種研究路段之第一個交流道至最後一個交流道為例進行說明。

上述 4.4.1 節中已完成時空表的建置，在完成之時空表中選取兩個交流道內所有設備。接著加上交流道至第一個設備之間的旅行時間，即為下圖 4.5.2 中綠色箭頭之路段。

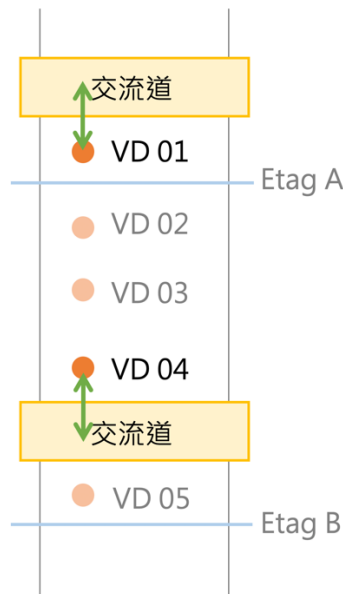


圖 4.5.2 延伸路段示意圖

此段旅行時間則使用最靠近的設備段之預測旅行速度回推。以圖 4.5.2 為例，交流道至 VD01 的旅行時間為該段距離除以 VD01 至 EtagA 的旅行速度，而 VD04 至交流道的旅行時間為該段距離除以 VD03 至 VD04 的旅行速度。計算完成後將所得之結果加入時空表中，完成路段延伸時空表建置。接著依照 4.4.2 節所述之延遲加總計算方法進行旅行時間計算。

各時間出發之旅行時間計算完成後，利用每一天的結果繪製出下圖 4.5.3。此圖以 9 月 1 日北上平面路段為例，橫軸為出發時間，每天共有 288 個五分鐘，因此依序將時間編號，呈現 0、1、……、287。縱軸為旅行時間預測值單位為分鐘。

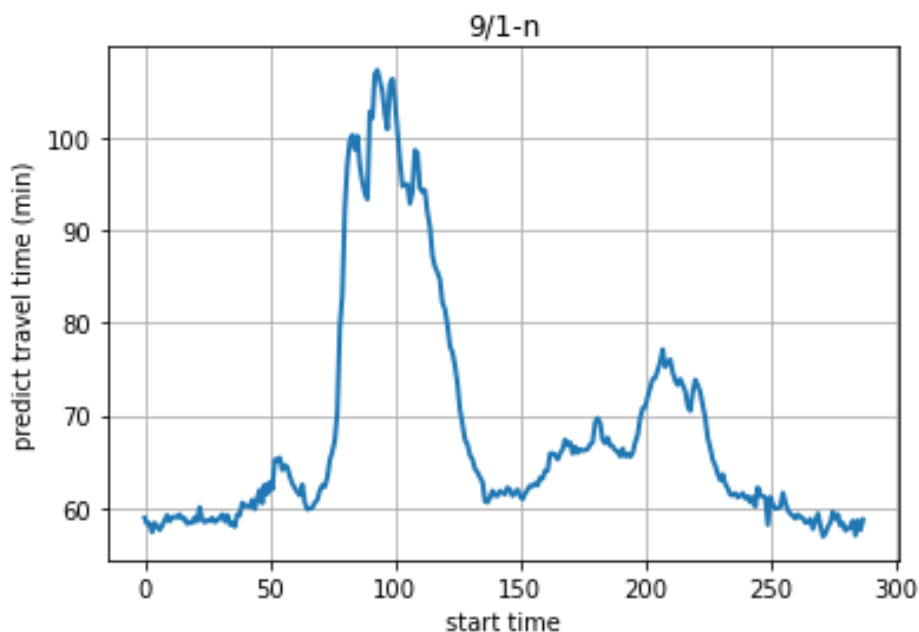


圖 4.5.3 各時段旅行時間結果

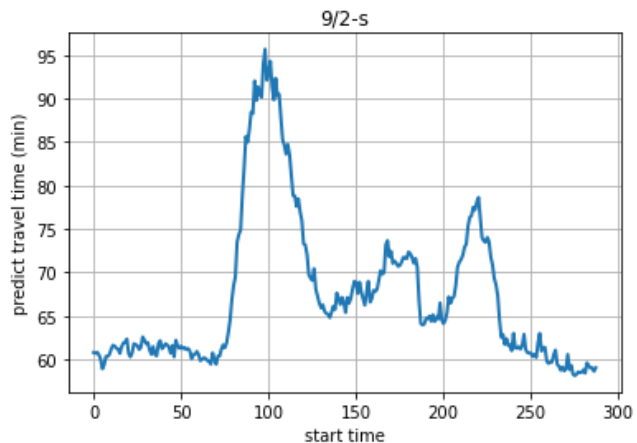
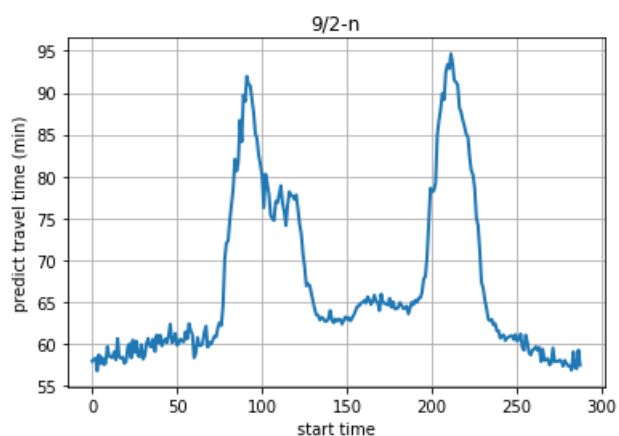
由於 9 月 1 日為週二上班日，因此圖 4.4.4.1 可看出明顯的早尖峰與午尖峰時段，該日平均旅行時間約為 68 分鐘，標準差約為 13 分鐘。旅行時間大於兩倍標準差的出發時間為早晨 6:40 至 9:30，此段時間從新竹交流道愈開往基隆交流道為最壅塞的時段。

表 4.5.1 為每日平面路段之結果，可發現週間(週一至週五)的旅行時間皆出現明顯早午尖峰，其中週一至週四早尖峰較午尖峰明顯，而週五的午尖峰較早尖峰明顯。假日(週六週日)的尖峰時間持續較長，且集中於午尖峰。

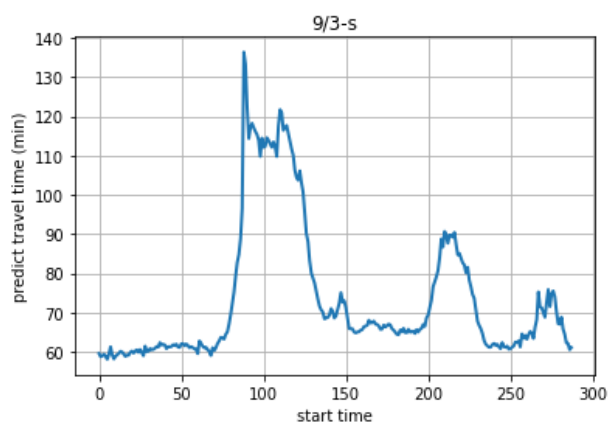
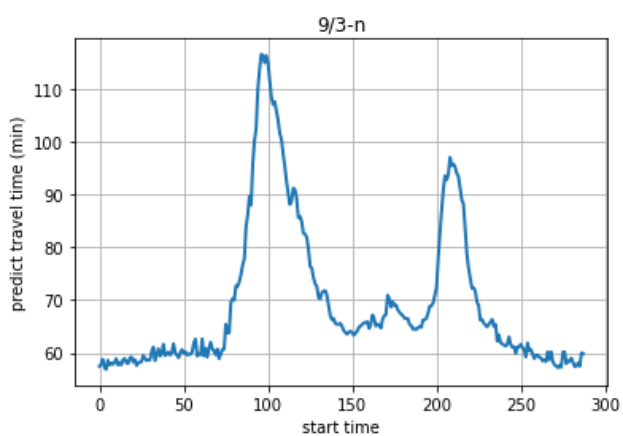
表 4.5.1 平面路段結果

	北上平面	南下平面
9 月 1 日 星期二		

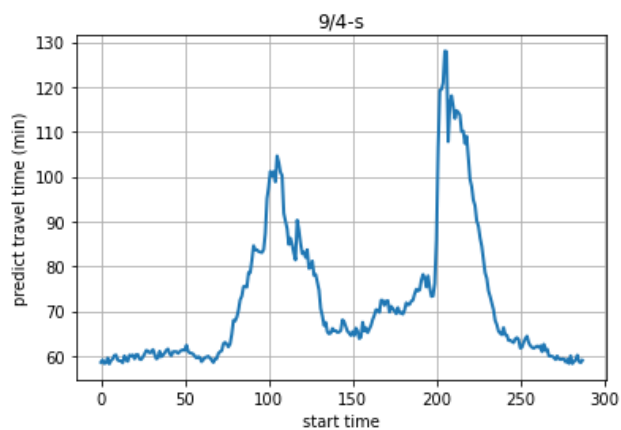
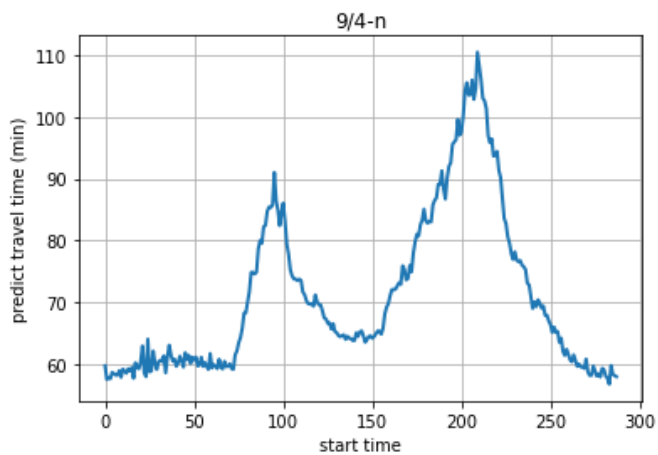
9月2日星期三

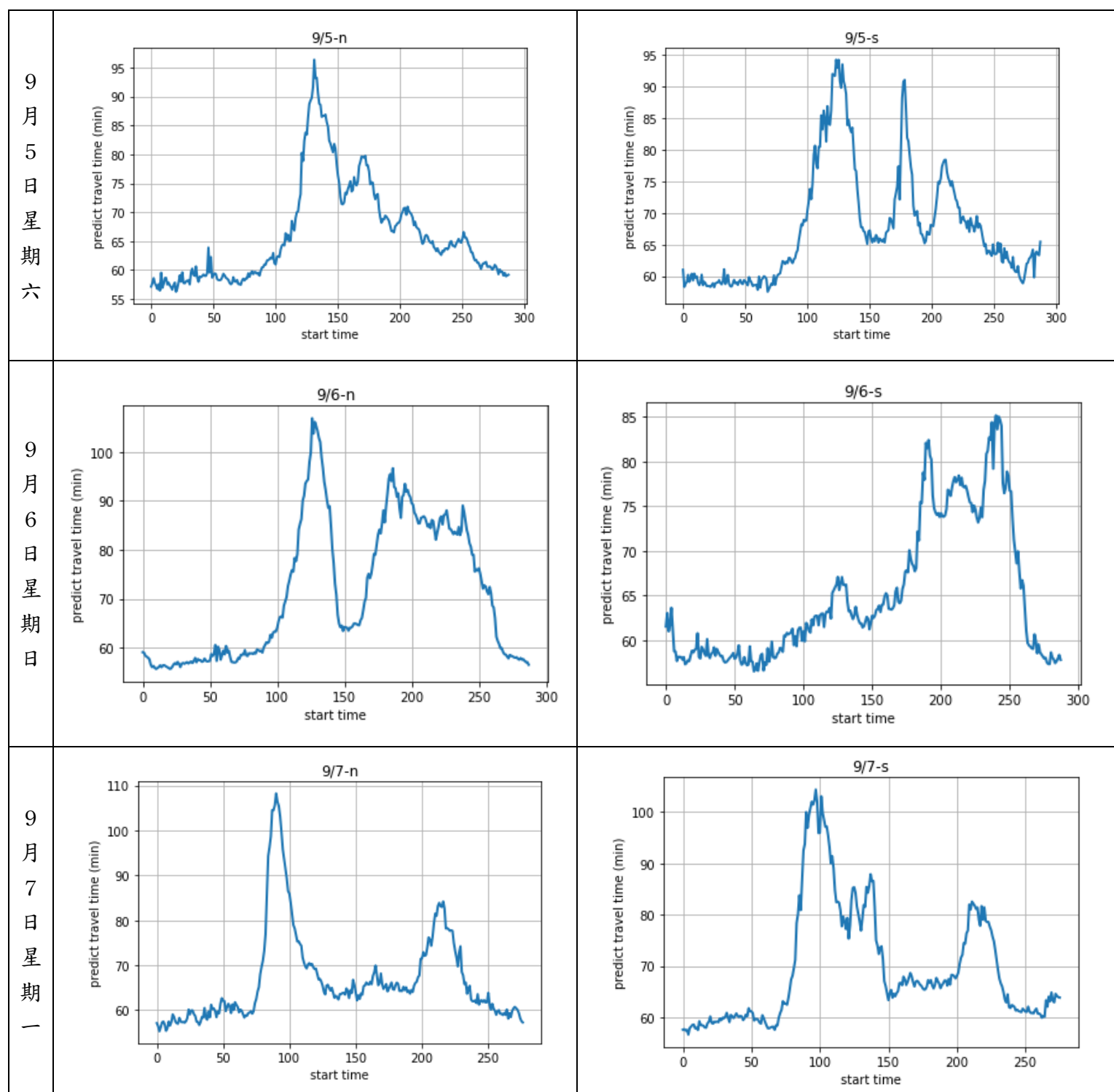


9月3日星期四



9月4日星期五





下表 4.5.2 為北上平面高速公路路段七天之統計資料，而表 4.5.3 為南下平面高速公路路段七天之統計資料。表格中分別計算了每日平均旅行時間、標準差、最長旅行時間、最長旅行時間發生時間點與早午尖峰時段的開始與結束之時間點。

本研究以每日平均旅行時間加上一倍標準差為基準，旅行時間超過該基準者視為尖峰發生。其中尖峰時間若為空白者，即表示該時段並無超過基準的旅行時間發生。

從表格中可以發現以下幾點相同處：

1. 週間(週一至週五)早尖峰時間較週末(週六與週日)早。由於週間早尖峰發生之原因為通勤，因此時間較早。
2. 週五早尖峰不明顯，但午尖峰持續時間較長。
3. 週日午尖峰持續時間長，皆持續至晚間。推測為出遊返家人潮眾多造成。
4. 最長旅行時間均發生於週四中午，其次為週五下午。

從表格中亦可以發現以下幾點不同處：

1. 北上路段平均旅行時間最長的為週日，然而南下路段週日之平均旅行時間為最短。
2. 北上路段週日標準差較大，然而南下路段較小。

表 4.5.2 北上平面路段統計資料

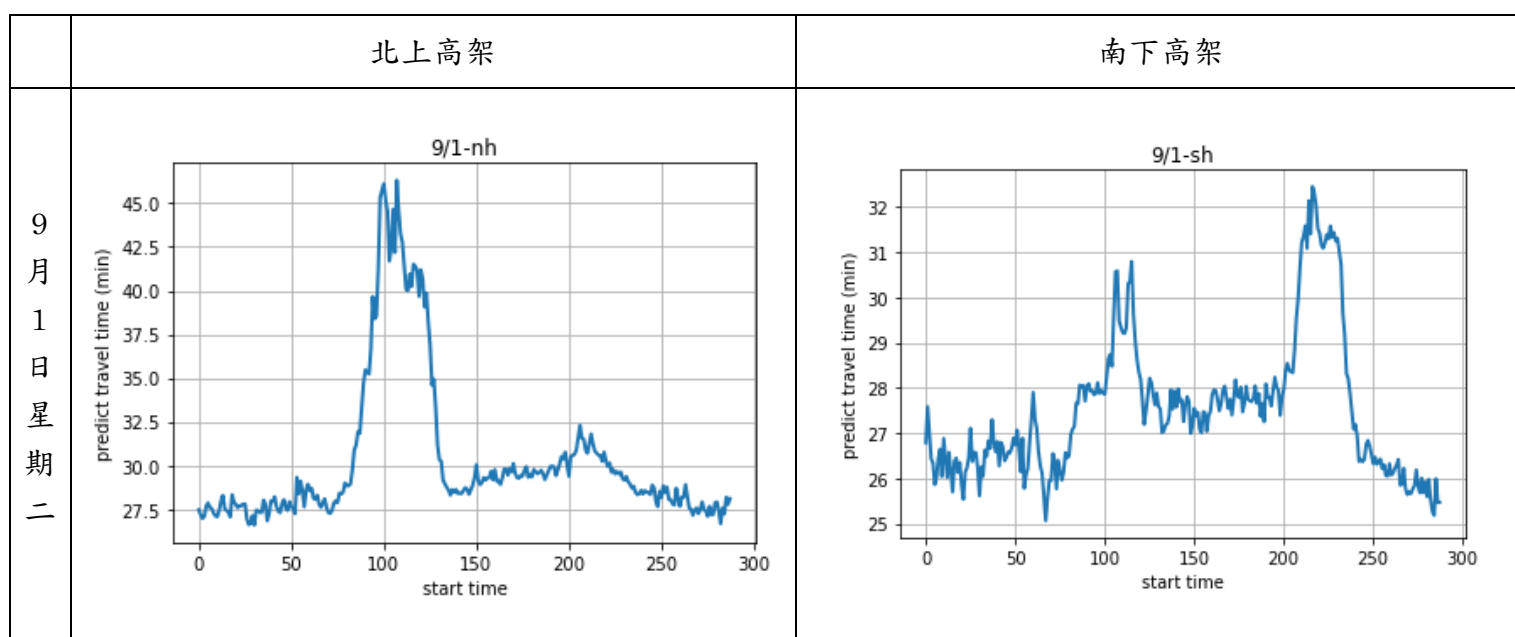
	2021/9/1 (二)	2021/9/2 (三)	2021/9/3 (四)	2021/9/4 (五)	2021/9/5 (六)	2021/9/6 (日)	2021/9/7 (一)
平均旅行時間(分)	68.26	66.95	69.48	71.11	66.10	71.38	66.89
標準差	12.76	9.88	14.11	12.99	8.73	14.20	10.44
最長旅行時間(分)	107.24	94.64	116.48	110.43	96.37	106.90	108.22
最長旅行時間之 出發時間	7:45	17:35	8:00	17:25	10:55	10:30	7:30
旅行時間之一倍 標準差(分)	81.02	76.83	83.59	84.11	74.84	85.59	77.34
早尖峰起始	6:35	6:55	7:15	7:35	10:05	9:50	6:55
早尖峰結束	10:00	10:00	10:00	8:20	12:35	11:35	8:40
午尖峰起始		16:35	16:55	14:55	13:20	14:55	17:25
午尖峰結束		18:45	18:00	18:35	14:50	20:00	18:35

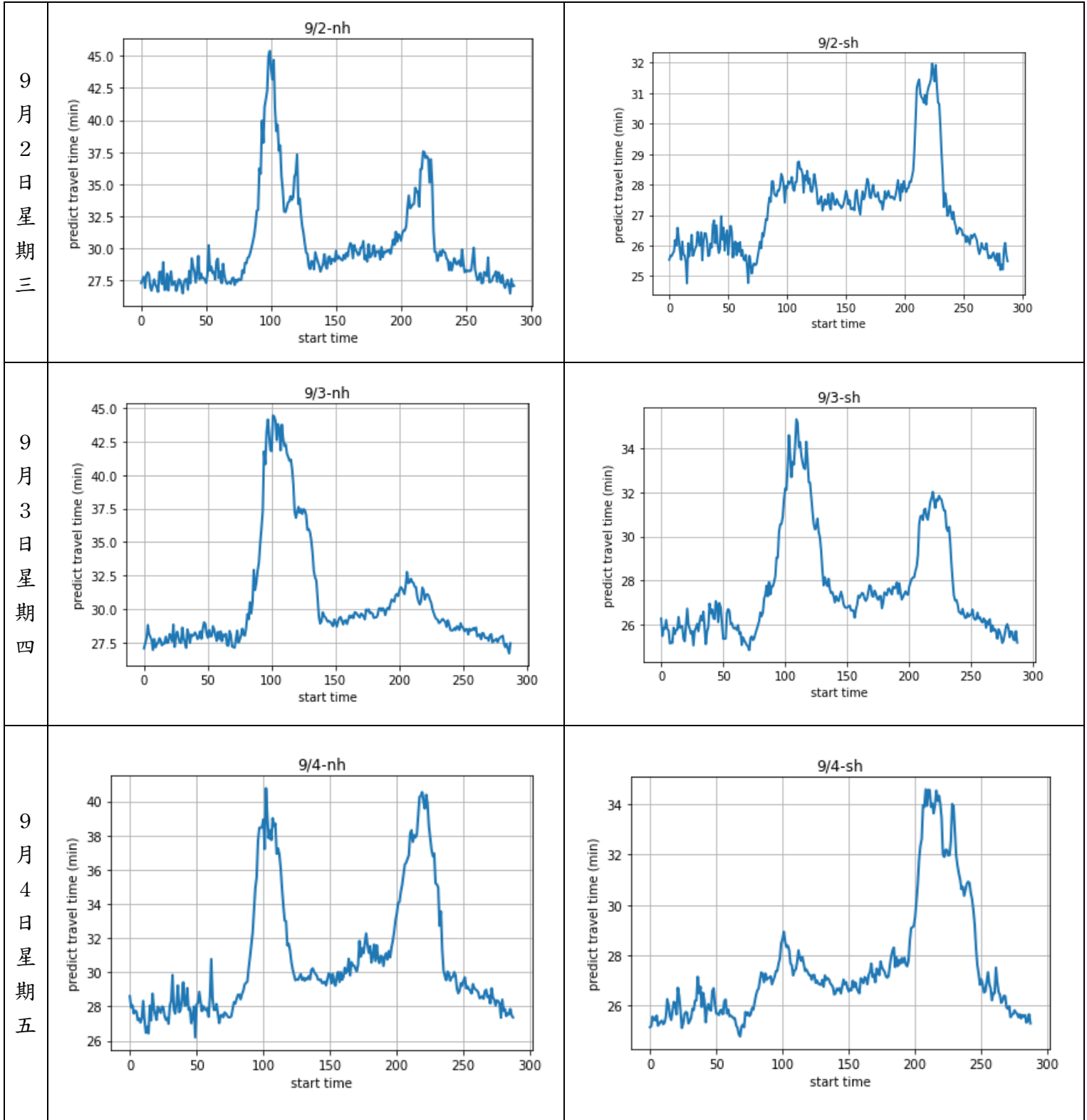
表 4.5.3 南下平面路段統計資料

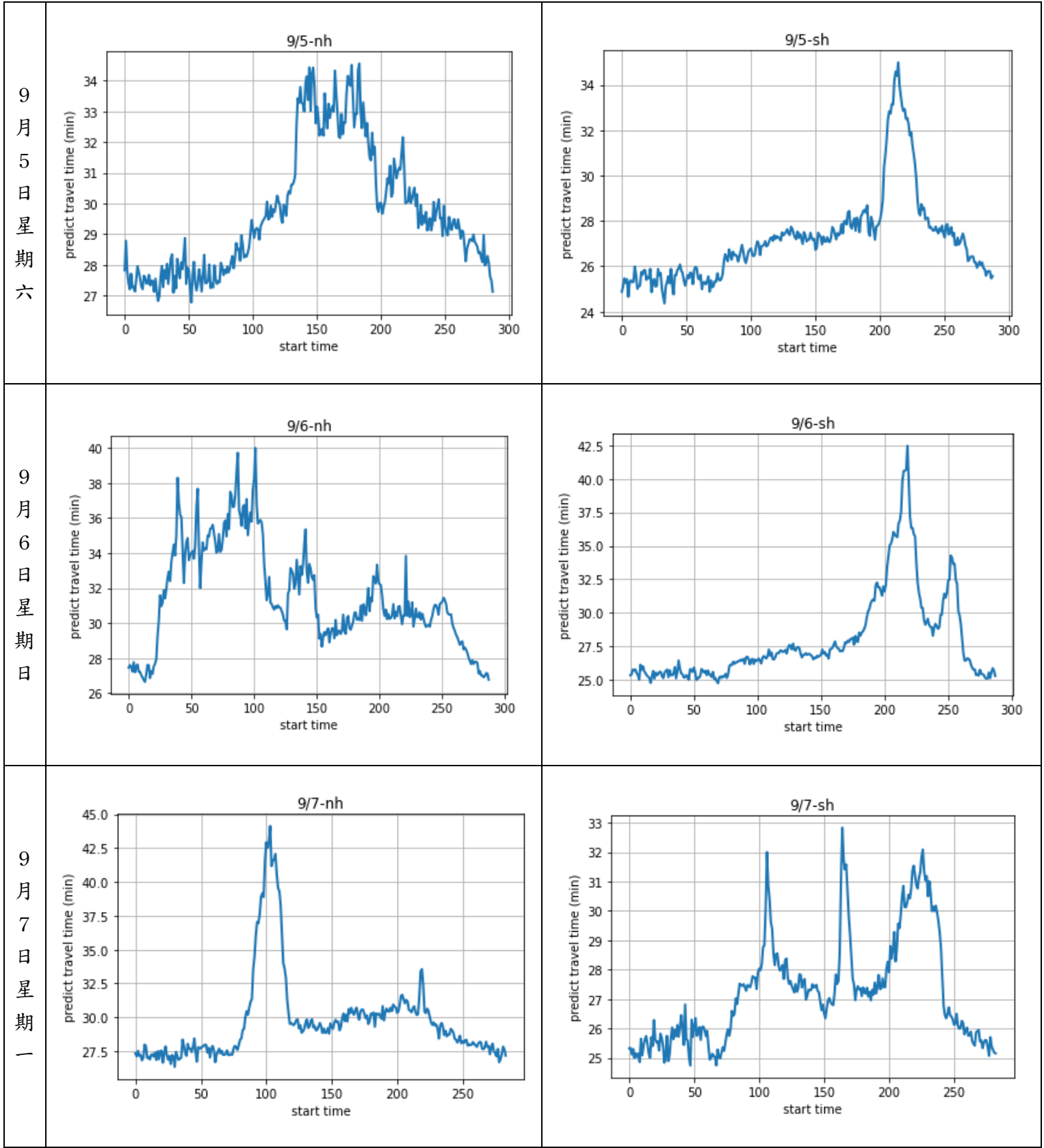
	2021/9/1 (二)	2021/9/2 (三)	2021/9/3 (四)	2021/9/4 (五)	2021/9/5 (六)	2021/9/6 (日)	2021/9/7 (一)
平均旅行時間(分)	72.02	67.3	73.43	72.06	67.63	65.05	69.70
標準差	11.87	8.75	17.73	15.62	9.11	7.81	11.65
最長旅行時間(分)	102.05	95.67	136.39	128.02	94.26	85.14	104.35
最長旅行時間之 出發時間	17:40	8:10	7:20	17:05	10:15	20:00	8:05
旅行時間之一倍 標準差(分)	83.90	76.15	91.16	87.69	76.75	72.87	81.35
早尖峰起始	6:55	7:05	7:15	8:15	8:45		7:05
早尖峰結束	9:30	9:50	10:25	9:50	11:30		9:30
午尖峰起始	17:20	17:55		16:45	14:25	15:25	17:25
午尖峰結束	19:00	18:25		18:45	15:15	21:00	18:20

表 4.5.4 為每日高架路段之結果，可發現北上高架路段的午尖峰於週一至週四較不明顯，此點與而南下高架路段相反，推測早晨上班時以北上通勤者居多，而下班時段則出現相反車流，南下返家。此外北上高架路段假日(週六週日)的尖峰時間持續較長，而南下高架路段的週間與週末的差異在於早尖峰，週末無早尖峰。

表 4.5.4 高架路段結果







下表 4.5.5 為北上高架路段七天之統計資料，而表 4.5.6 為南下高架路段七天之統計資料。表格中分別計算了每日平均旅行時間、標準差、最長旅行時間、最長旅行時間發生時間點與早午尖峰時段的開始與結束之時間點。

與平面路段相同，本研究以每日平均旅行時間加上一倍標準差為基準，旅行時間超過該基準者視為尖峰發生。其中尖峰時間若為空白者，即表示該時段並無超過基準的旅行時間發生。

從表格中可以發現以下幾點相同處：

1. 兩路段最長的平均旅行時間皆發生於週日。
2. 週間早尖峰時間較週末早。
3. 週五與週六午尖峰持續時間較長。

從表格中亦可以發現以下幾點不同處：

1. 北上高架路段最長旅行時間多發生於上午，南下高架路段則多發生於下午。與前述推測之早晨北上通勤者居多，而下班時段出現相反車流相同。
2. 南下高架路段週日自晚間 20:35~21:25 亦發生尖峰現象。推測為週日返家人潮所造成，因此於晚餐時間後再次出現尖峰現象。
3. 南下高架路段週一下午 13:35~14:10 亦發生尖峰現象。從表 4.4.4.4 中可發現此段時間出現短暫旅行時間高峰，由於發生於週間，因此推測可能為模型造成之誤差。

表 4.5.5 北上高架路段統計資料

	2021/9/1 (二)	2021/9/2 (三)	2021/9/3 (四)	2021/9/4 (五)	2021/9/5 (六)	2021/9/6 (日)	2021/9/7 (一)
平均旅行時間(分)	30.34	30.16	30.58	30.72	29.70	31.54	29.75
標準差	4.39	3.62	4.23	3.57	2.07	2.89	3.31
最長旅行時間(分)	46.31	45.38	44.46	40.76	34.56	40.00	44.13
最長旅行時間之 出發時間	8:55	8:15	8:25	8:30	15:15	8:25	8:35
旅行時間之一倍 標準差(分)	34.74	33.78	34.81	34.29	31.78	34.43	33.06
早尖峰起始	7:30	7:35	7:35	7:50	11:10	6:15	7:30
早尖峰結束	10:35	10:10	10:55	9:30	12:00	8:55	9:30
午尖峰起始		17:10		16:55	12:05		18:10
午尖峰結束		18:40		19:15	16:15		18:15

表 4.5.6 南下高架路段統計資料

	2021/9/1 (二)	2021/9/2 (三)	2021/9/3 (四)	2021/9/4 (五)	2021/9/5 (六)	2021/9/6 (日)	2021/9/7 (一)
平均旅行時間(分)	27.53	27.19	27.69	27.53	27.20	27.99	27.26
標準差	1.57	1.47	2.40	2.33	1.99	3.43	1.84
最長旅行時間(分)	32.45	31.97	35.32	34.59	34.98	42.46	32.83
最長旅行時間之 出發時間	18:00	18:35	9:05	17:20	17:50	18:10	13:40
旅行時間之一倍 標準差(分)	29.10	28.66	30.09	29.86	29.19	31.42	29.10
早尖峰起始	8:45	9:05	7:55				8:45
早尖峰結束	9:40	9:10	10:35				9:10
午尖峰起始	17:15	17:20	17:20	16:45	16:55	16:05	17:00**
午尖峰結束	19:30	19:15	19:20	20:15	19:00	18:55*	19:55

*20:35~21:25 亦發生尖峰現象，詳見內文解釋

**13:35~14:10 亦發生尖峰現象，詳見內文解釋

4.6 小結

本研究利用兩種資料融合方法以及單純 ETC 資料分別進行 GBDT 模式建模用以預測旅行時間。並發現資料融合方法一與單純 ETC 資料兩種資料之誤差相近，而資料融合方法二有效提升旅行時間預測的準確度。

為使本研究的應用性提升，進而將資料融合方法二進行路段延伸分析，完成交流道至交流道的旅行時間預測模型。此模型可預測研究範圍內任意出發時間(五分鐘為間隔)各交流道之間的旅行時間，方便用路人使用，並預先決定出行時間，用以減少旅行時間。

第五章 結論與建議

本研究的目的是為利用多資料來源建立機器學習模型，並用以預測高速公路旅行時間。本研究使用高速公路局之 ETC 和 VD 開放資料進行分析，首先分別對原始資料進行前處理，接著開發兩種資料融合方法，並且建立 GBDT 模型，用三種資料進行旅行時間預測，最後比較三種資料對旅行時間預測的準確度。

為增加本研究之應用性，本研究利用準確度最高之資料融合方法二進行後續路段延伸的分析。透過本研究之成果，交通主管機關可更準確的預測旅行時間，讓用路人有更好的出行體驗。本章節分為兩個部分，首先根據本研究之研究成果提出結論，接著提出未來進行相關研究之建議。

5.1 結論

1. 本研究利用 ETC 和 VD 資料預測高速公路未來一個月內之旅行時間。本研究利用過去三個月的歷史資料（2020 年 6 月 1 日～2020 年 8 月 31 日）進行分析，預測未來一個月（2020 年 9 月）內每日的旅行時間。
2. 本研究實證研究範圍選擇國道一號新竹交流道（含）以北之雙向路段，包含五股楊梅高架與汐止五股高架。針對北上與南下兩種行車方向以及是否為高架路段，將實證研究路段區分為：北上平面高速公路、南下平面高速公路、北上高架路段與南下高架路段，共四種。
3. 本研究在不調整原始資料的前提下發展兩種融合 VD 資料與 ETC 資料的方法。透過資料融合提升資料多元性，並且改善 VD 設備常出現的資料異常問題以及 ETC 資料空間稀疏問題。
4. 本研究透過參數校估建立 GBDT 模式，並應用於旅行時間預測。利用模式設計之流程，進行七個參數的校估。
5. 經過實證研究，可發現在每段 ETC 設備至 ETC 設備的預測結果中，資料融合方法一與原始 ETC 資料有相近的結果，而資料融合方法二有明顯較低的誤差值。
6. 經過實證研究，可發現四種路段中第一個 ETC 設備至最後一個 ETC 設備的預測結果顯示：資料融合方法二相較於單純 ETC 資料可以提升 82%～97% 的準確率。

7. 本研究所開發的資料融合方法二，利用 GBDT 方法建構旅行時間預測模型，可以有效提升預測準確率。並且在四種實證研究路段中皆有良好的表現。
8. 本研究為加強應用性，進行路段延伸分析，將旅行時間預測從 ETC 設備至 ETC 設備延伸成交流道至交流道。可預測任意時間出發，從任意交流道至交流道的旅行時間，此預測可供用路人參考，並提前規劃。

5.2 建議

1. 由於 VD 原始資料缺漏多，未來進行研究時可針對 VD 原始資料進行全面性檢討，提出改善資料品質方法。
2. 未來可加入非重現性資料（如車禍）一同進行分析，增加模式準確度。
3. 未來進行研究時可針對 VD 原始資料進行預分析與建模，使融合後的資料有更佳的品質。
4. 未來可嘗試選擇不同的變數進行建模，提升模型準確度。
5. 未來可針對不同假期進行分別建模，提升假期之旅行時間預測準確度。
6. 未來可以針對不同路段進行實證研究分析。

第六章 參考文獻

- [1] Chun-Hsin Wu, Jan-Ming Ho, and D. T. Lee(2004).Travel-Time Prediction With Support Vector Regression. IEEE Transactions on intelligent transportation systems, vol.5, no.4, December 2004.
- [2] Fan, S.S., Su, C., Nien, H. et al. Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection. Soft Comput 22, 5707 – 5718 (2018).
- [3] Forough Goudarzi (2018). Travel Time Prediction: Comparison of Machine Learning Algorithms in a Case Study. 2018 IEEE 20th International Conference on High Performance Computing and Communications.
- [4] J. H. Friedman, ‘ ‘Greedy function approximation: A gradient boosting machine,’ ’ Ann. Statist., vol. 29, no. 5, pp. 1189 – 1232, Oct. 2001, doi: 10.1214/aos/1013203451.
- [5] Juan Cheng, Gen Li, and Xianhua Chen (2018).Research on Travel Time Prediction Model of Freeway Based on Gradient Boosting Decision Tree. 2018 IEEE.
- [6] Lelitha Vanajakshi, and Laurence R. Rilett (2007). Support Vector Machine Technique for the Short Term Prediction of Travel Time. IEEE Intelligent Vehicles Symposium Istanbul, Turkey, June 13-15, 2007.
- [7] Nanthawichit C, Nakatsuji T, Suzuki H. Application of Probe-Vehicle Data for Real-Time Traffic-State Estimation and Short-Term Travel-Time Prediction on a Freeway. Transportation Research Record. 2003;1855(1):49-59. doi:10.3141/1855-06
- [8] Senyan Yang, Jianping Wu, Yiman Du, Yingqi He, Xu Chen, "Ensemble Learning for Short-Term Traffic Prediction Based on Gradient Boosting Machine", Journal of Sensors, vol. 2017, Article ID 7074143, 15 pages, 2017. <https://doi.org/10.1155/2017/7074143>
- [9] Simon Oh, Young-Ji Byon, Kitae Jang, and Hwasoo Yeo (2018). Short-term Travel-time Prediction on Highway: A Review on Model-based approach. KSCE Journal of Civil Engineering (2018) 22(1):298-310
- [10] Tang-Hsien Chang, Albert Y. Chen, Yu-Ting Hsu, Chien-Li Yang (2016). Freeway travel time prediction based on seamless spatio-temporal data fusion: case study of the freeway in Taiwan. Transportation Research Procedia 17 (2016) 452 – 459.

- [11] Xia Li, Ruibin Bai (2016).Freight Vehicle Travel Time Prediction Using Gradient Boosting Regression Tree. 2016 15th IEEE International Conference on Machine Learning and Applications.
- [12] Xu Miao, Bing Wu, Yajie Zou, and Lingtao Wu (2020). Examining the Impact of Different Periodic Functions on Short-Term Freeway Travel Time Prediction Approaches. Journal of Advanced Transportation.
- [13] Yanjie Duan, Yisheng Lv, and Fei-Yue Wang. Travel Time Prediction with LSTM Neural Network. 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC).
- [14] Yanru Zhang, Ali Haghani(2015). A gradient boosting method to improve travel time prediction. Transportation Research Part C 58 (2015) 308 - 324.
- [15] 翟雅嶠, 翁劍成, 榮建, 劉小明 (2009) 多元數據融合的區間車量速度預測算法研究, 交通信息與安全 2009 年第 3 期第 27 卷總 148 期。
- [16] 何聰儒 (2009) 利用群集分析與線性迴歸推估高速公路旅行時間之研究。
- [17] 劉軒寧 (2016) 變動 k 值之 k-NN 法 應用於高速公路旅行時間預測之研究。
- [18] 黃冠維 (2018) 利用卷積神經網路預測市區道路路況之研究。
- [19] 朱志杰 (2013) 使用車輛偵測器和自動車輛辨識之資料預測高速公路旅行時間。
- [20] 吳耿暉 (2018) 以深度學習預測國道短期旅行時間。