國立雲林科技大學研究生論文上傳同意書

工業工程與管理系 研究所 工工 組研究生 周映竹

論文已完成口試委員修改建議，請同意論文上傳學校圖書館。

謹陳

論文指導教授：陳奕中 (請簽章)

是否提前畢業：☐ 是，於___月畢業。(非學期公告離校時間，請以實際畢業離校月份書寫)

☑ 否。

學生簽章：周映竹

學生學號：M10621243

申請日期：108, 8, 23

備註：
1. 指導教授簽章後請連同論文一併上傳圖書館審查，據以確認論文已修改完成。
2. 同意書請放置於論文電子檔第一頁(封面前一頁)，紙本論文內請勿裝訂此同意書。
3. 延後公開請另填延後公開申請書。

國立雲林科技大學工業工程與管理系工業工程組

碩士論文

MS in Industrial Engineering and Management

Department of Industrial Engineering and Management
National Yunlin University of Science & Technology

Master Thesis

基於社群網路資料之旅遊區辨識

Tourist Area Identification Based on Social Network Data

周映均

Ying-Chun, Chou

指導教授：陳奕中 博士

Advisor: Yi-Chung Chen, Ph.D.

中華民國 108 年 6 月

June 2019

國立雲林科技大學
研究所碩士班學位論文考試委員會審定書

本論文係**周映均**君在本校 **工業工程與管理系**提論文「**基於社群網路資料之旅遊區辨識**」合於碩士資格水準，業經本委員會評審認可，特此證明。

口試委員：

陳奕中	陳奕中
陳思翰	陳思翰
傅家啓	傅家啓
劉傳銘	劉傳銘
陳以錚	陳以錚

指導教授：陳奕中　陳奕中

所　長：

中　華　民　國　108　年　7　月　18　日

摘要

近年來促進台灣旅遊發展的議題持續被重視，台灣各城市的旅遊區人潮擁有不平衡的現象，因此本研究目的為在不影響熱門景點營運下將遊客吸引至較少遊客的潛力景點。然而隨著科技得進步，大部分遊客的選擇深受著網際網路上顯示的旅遊資訊，並且相關研究更證明社群網路的資訊富含使用者行為特徵。本研究開發新的 AP-DBSCAN 分群演算法，將具有關聯的顯性/隱性景點與地理標籤照片聯繫成旅遊區。而當遊客選擇一個目的地時，顯然表示該使用者具有前往該目的地附近的可能性。最後本研究建立一個框架不僅促進該旅遊區內的潛力遊客與聯盟商家進行互動，對於政府而言更能根據不同的旅遊區給予更合適的資源。

關鍵字：隱性景點、旅遊區概念模型、社群網路、DBSCAN 演算法

Abstract

There have been concerted efforts in recent years to promote the development of tourism in Taiwan. An initial examination of tourist areas in Taiwan revealed a considerable imbalance in the number of tourists in different regions of the country. Our objective in this study was to direct tourists toward less popular areas with tourism potential without affecting the tourist trade in current hotspots. Many individuals obtain most of their travel-related information on the internet, and travel preferences are strongly affected by their feeds on social networks. Researchers have demonstrated that user behavior can be derived from the messages they leave on social networks. In this paper, we develop the AP-DBSCAN algorithm in conjunction with textual trajectory data to identify popular tourist areas, based on noticed attractions, unnoticed attractions, and the geo-coordinates of photos. Clearly, when a tourist selects a given tourist destination, he/she is a potential customer for the surrounding areas. The proposed scheme is meant to facilitate cooperation among operators in specific areas to promote themselves based on the preferences of tourists in surrounding areas. This data could also help to guide the government in the allocation of resources for the promotion of tourism.

Keywords: unnoticed attraction, tourist area conceptual model, social network, DBSCAN algorithm

誌謝

在碩士班就讀的日子哩，一份碩士論文能如期完成，背後要感謝的人很多，因此藉由此機會為引領我、協助我完成碩士論文的每一個人致上最真摯的感謝。

首先感謝開啟我在資料科學領域興趣的啟蒙老師鍾震耀老師及陳思翰老師，告訴我資料科學的奧妙以及其偉大之處，儘管身在不同處也一直關心者以及給予協助。然而，最為感謝的莫過於陳奕中老師，在我僅靠者初淺的資料科學知識，願意接受我讓我成為實驗室的一員，在陳奕中老師不辭辛勞的指導下，不僅是專業知識上的培養、解惑、報告表達、解決問題的思維協助，老師更提供許多專案參與機會、比賽及出國發表讓我與不同國家及領域的人互動，在這兩年的研究過程中受益良多。

在實驗室的日子裡，不管在課業或是實驗室的研究，時常與大家有高度的合作，日子雖然雖甜苦辣，但是也非常充實，感謝教導我寫程式與研究資料分析的陳駿宏學長、羅志豪學長，感謝東錡、欣霖、泊樣、昱瑋、士哲以及碩一及政傑的學弟妹們，特別感謝東錡，每當我遇到瓶頸時，在技術方面與想法上都給予極大的耐心與協助，讓我持續前進。

感謝我的父母、妹妹與弟弟，除了支持我繼續學習也會適時關心我的狀態，讓我在求學的過程中能無後顧之憂的學習。最後感謝一路上讓我成長的每個人 讓我順利拿到碩士學位。願以此篇論文獻給你們，與你們分享這份喜悅與榮耀。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究背景及動機

根據世界旅遊及旅行理事會 World Travel and Tourism Council (WTTC)於 2018 年全球經濟影響報告中，提到目前旅遊業佔全球 GDP 共 10.4%，其 2017 至 2018 年間 GDP 成長幅度，更超越了製造業與金融服務業，龐大的經濟流動之下，旅遊觀光便成了不可或缺的議題。

推動觀光為近年來台灣的重要目標，政府祭出春遊補助等補貼措施，但對台灣整體觀光收入的貢獻卻非常有限。究其原因，是因為旅遊人口分布不均，大部分民眾只認識如日月潭、九份或太魯閣國家公園等大型景點，或是一些新聞上爆紅的景點，如彎腰郵筒等，相比對其他宣傳力不足的小景點通常不認識，如日月潭旁邊的年梯步道或是雲林高鐵站的旁的美人樹大道通常沒聽過，使得遊客人流數目較嫌少。而每當放假時，大部分民眾只會一而再再而三的去一些已經去過的大型景點，而那些宣傳力不足的小景點通常乏人問津。於此情況下，部分大型景點因為遊客常去而失去新鮮感而造成消費低落,另一面小景點則因為無人前往而，該景點沒有收入，最終造成整體觀光產值的低落。

1.2 研究目的

本研究提出了旅遊區辨識與遊客人流數目建模兩大概念來克服上述的問題。首先，旅遊區辨識意指將民眾常去旅遊的區域標記出來。而該區域內會包含三個部分，包含了顯性景點、隱性景點，以及民眾在去這些顯性或隱性景點時會活動的範圍，其中顯性景點是指大部分人都知道的景點，隱性景點是指相對於顯性景點而言，固然是一個景點，但並非大部分的人都知道的。以下舉台灣知名景點日月潭進行說明，如圖 1 可以看到伊達邵碼頭、年記做不復賣、伊達邵渡假旅店景點被包在同一個區域中，代表民眾在出遊時會將這些景點視為同一區塊而在同次旅行中前往，其中伊達邵碼頭為顯性景點，而伊達邵渡假旅店與年記做不復賣為

圖 1 日月潭伊達邵碼頭旅遊區

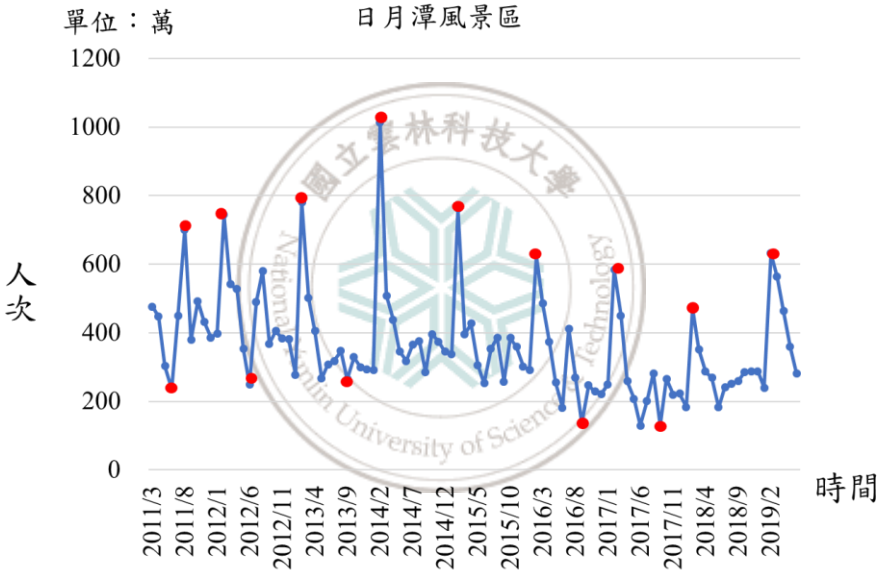

圖 2 遊客人流數目變化

隱性景點，即為在伊達邵碼頭顯性景點附近的一家特色小吃及旅館，部分遊客在遊玩伊達邵碼頭顯性景點後可能會一併前往，也有部分遊客遊玩伊達邵碼頭顯性景點後會直接忽略該景點而離開。圖 1 中旅遊區周圍不規則變化的曲線則代表遊客在這區活動的範圍，在該區中可以看到某些旅遊區邊界距離景點較遠，有些則較近，這是因為小吃店或紀念品商店可能分布不均，而大部分分布在某一區而鮮少人在另一區；緊接在遊客人流數目建模部分，本研究盡研究之力推算該旅遊區不同時間的可能遊客人流數，亦即有多少遊客前往該地區，以圖 2 說明日月潭地區的遊客人流數變化圖，在該圖中我們可以看到 2 月與 8 月時遊客人流數較多，

因為該區在這兩個月時是旅遊旺季。而特別的是在 6 月時，遊客人流數目有突然的多變的情況，這是因為該區在 6 月時有颱風侵襲，所以遊客人流數目突然減少，而在 10 至 11 月時雖然是旅遊淡季，但因為縣市政府有辦活動，所以遊客人流數目上升。本研究在辨識出旅遊區後，進行遊客人流數的建模，並依此推估未來遊客人流數。

藉由上述兩大作法，本研究預想協助台灣地區改善因旅遊人數分布不均而造成觀光產值低落的問題。憑藉旅遊區辨識的結果，整理出台灣地區大部分的顯性景點與隱性景點，並讓民眾在查詢後了解原來台灣並非只有那些主要的顯性景點而已。更重要的是，也可以藉由旅遊區辨識出來的區域告知民眾隱性景點與顯性景點的關係，讓民眾在前往這些顯性景點的時候也能考慮前往附近的隱性景點，並刺激民眾額外的消費。此外若對於政府而言，旅遊區辨識的結果也可以讓政府更加了解一個地區觀光資源的狀況，並作為是否要投入更多資源改善觀光情況的依據。最後本研究進行的旅遊區遊客人流數目建模則可以協助遊客了解未來每個旅遊區域的遊客人流數，並做出合理的出遊區域選擇，以避免因人潮擁擠、交通壅塞等旅遊體驗品質不佳，影響遊客出遊心情[1]。進一步說明，此作法也能協助政府將大型景點的人潮移動至其他有潛力、旅遊品質佳的景點，不僅能讓著名景點回歸原有旅遊體驗品質，也可以提昇其他景點的人潮量並帶動該地方經濟。

在我們前述所提到的旅遊區辨識概念，過往曾有些學者進行過研究。首先，Czernek-Marszałek [2]提出了如果不同產業與區域串成一個旅遊區，則可以旅遊業者就可以推出屬於該旅遊區的特色活動、旅遊套票等來吸引遊客到訪。至於政府則可因為旅遊區的產生，提供不同旅遊區合適的資源配給，以達到旅遊資源分配最佳化的目的。接著，Buhalis and Law [3]與 Navío-Marc *et al.* [4]曾提及若把地圖間景點的遊玩關聯性串連起來並輸入系統中，則當遊客 A 到景點 X 旅遊後，系統便可以快速建議 A 一般人到 X 景點旅遊後，還會到附近的 Y 景點與稍遠的 Z 景點旅遊，並吸引 A 前去 Y 與 Z 旅遊。Lee *et al.* [5]運用 Flickr 地理標籤，搭配 DBSCAN

演算法來找出人群分佈熱點，利用關聯規則來幫這些熱點建立關聯性，並因此建立人群的分布與移動區域。Kennedy *et al.* [6]則是先用視覺化的方式觀察 Flickr 照片中的地理標籤的可能分群數目，接著利用 *K*-means 演算法來將這些地理標籤分群並識別為旅遊區。整體而言，上述方法大多能解決他們各自所提出的問題，但卻無法解決本論文所提出的問題。因為對[3]、[4]、[5]這三篇的方法而言，他們是先討論景點間的關聯性後，再將高關聯性的景點整合成旅遊區。而這樣所繪製出來的旅遊區可能空間上分布很散亂且區域很大，與本篇論文所提到之概念不合。而 Kennedy *et al.* [6]提到的概念雖然與本論文欲解決問題類似，但他的旅遊區切割方式僅考慮 Flickr 照片的點位，並沒有考慮景點的位置。在這個狀況下，他們所描繪出的旅遊區可能與景點真實位置有所偏差，與現實狀況不吻合。鑒於上述新方法應該被提出來改善上述缺點。

本研究提出了以下順序來達成旅遊區辨識與遊客人流數目建模的目標，包含了(1)建立台灣地區的顯性景點資料庫、(2)找出台灣地區的隱性景點、(3)旅遊區的辨識，以及(4)利用深度學習演算法替每個旅遊區的遊客人流數目變化進行建模，並藉此推估每個旅遊區的未來遊客人流數目變化。現介紹本論文每部分工作的執行概念如下。首先為顯性景點面，本篇論文的定義是所有人都熟知的景點，但要分別定義每個景點是否為人熟知並不容易，固本研究使用台灣觀光局與水土保持局等提供的景點資料作為顯性景點，因為這些官方資料庫大多只收集台灣地區長久存在的大型景點，而這些景點也通常是為人熟知。接著是找尋台灣地區隱性景點面，本篇論文使用社群網路文章曾經提過的景點，但卻沒有出現在顯性景點資料庫者來做為隱性景點。會有這樣的作法，是因為這些景點符合前述隱性景點的定義，這些景點相對於顯性景點來說名氣較差，只有部分人知道，而非所有人都知道者。而社群網路文章上所提到的景點，便是因為他們被文章撰寫者所知，所以才會在文章上被提及，而來閱讀這些文章的讀者通常是不知道這些景點或是對這些景點不熟才會來閱讀的，明顯符合隱性景點的定義。第三，本論文會使用 Flickr

上的照片資料集來勾勒每個旅遊區的範圍。此處使用 Flickr 照片資料集是因為該資料集內擁有大量使用者上傳的照片，且這些照片都附有拍攝的經緯度座標。更重要的是，這些照片大多是遊客們出遊的照片，因此能充分表示遊客們出遊時曾經到過的位置，若我們能有效分析每個景點周遭遊客曾經拍照的地點，則必然對我們勾勒旅遊區範圍有莫大的幫助。最終，在我們勾勒出旅遊區範圍後，我們會收集這區域內的遊客人流數歷史數據，用長短期記憶模型對這些歷史數據建模，最終訓練好的深度學習模型則會用來預測該區域未來的遊客人流數目。

然而，上述所提到的工作中，最具有挑戰性的是第二項工作「從社群網路文章中找出台灣地區的隱性景點」與第三項工作「從 Flickr 照片資料集進行旅遊區的辨識」，現分別介紹這兩項的挑戰與本論文的處理方式如下。首先是「從社群網路文章中找出台灣地區的隱性景點」這部份，雖然過往已有許多學者嘗試將社群網路上的文字資訊轉換成旅遊資訊，但與本論文相關者幾乎沒有。例如 Majid *et al.* [7]運用 Flickr 資料中附有地理標籤與短文及天氣資料製作個人化景點推薦之推薦系統。He *et al.* [8]將個人檔案中附有多屬性資訊資料，如：興趣旅遊類型、個人旅遊景點順序，旅程開始即抵達地點等文字資料，運用 Latent Dirichlet Allocation 模型，找出人們旅遊的順序及關鍵景點名稱，建立旅遊路線推薦系統。上述研究在他們各自所提出的問題上表現都不錯，然而我們發現他們大多只使用文章中已經定義好的地理標籤來進行研究，而非從文章中找到未知的地理標籤並進行後續研究，故上述研究都無法直接應用在本論文中。那為了克服上述的問題，本論文開發了一整套的演算法來從社群網路文章中找到未知的地理標籤，並為後續步驟所用。至於第三項工作「從 Flickr 照片資料集進行旅遊區的辨識」，過往也曾有 Kennedy *et al.* [6]與 Lee *et al.* [5]分別利用 *K*-means 與 DBSCAN 演算法從 Flickr 照片資料集中找出旅遊區。然而這兩個知名的演算法可能無法直接用於本演算法中，因為這兩個演算法都是直接對一群空間資料做分群，但本論文卻必須以顯性景點與隱性景點為中心，之後再以這些景點周遭的 Flickr 資料來描繪旅遊區的範圍。明顯的，我

們研究的目標與過往研究的目標有非常大的不同。為了改善這樣的困難，本論文則是開發新的 AP-DBSCAN 分群演算法來進行旅遊區辨識，該演算法的細節則會在論文中介紹。最終在我們克服上述兩個困難後，我們將能有效辨識出台灣地區的旅遊區。如此當該區之業者及政府，得知相互之關聯性後，可建立所屬該區之在地特色，如推出旅遊體驗套票、地方特色伴手禮等。此外，若我們能預測這些旅遊區未來的遊客人流量，則我們也能協助地方業者與政府做出適當的決策，在淡季時想辦法提昇旅遊人數，而在旺季時有效維持旅遊品質。而我們的實驗模擬則可驗證我們所提出方法的有效性。

1.3 研究範圍及相關定義

本研究為時間空間議題，由於資料取得的限制以及興趣點的時效性，故將空間範圍定義為位於南投日月潭；時間範圍為 2011/3/1-2018/02/28。於現況中社群網路中含有網路寫手，而該情況在研究範圍內。

顯性景點定義：可於各觀光局、農村風情網以及易遊網查詢到的興趣點；隱性景點定義：遊客分享至 PTT 之 Tai-Travel 版的興趣點；旅遊區定義：距離相近的顯性/隱性景點運用半監督式機器學習分群演算法共構出的區域範圍。旅遊區概念模型定義：旅遊區概念模型亦針對使用者愈想探討的主題做設計，而本研究以預估該旅遊區人數預測為例。

第二章 文獻探討

本章介紹數個與本論文相關的研究，包含 2.1 節的社群網路資料應用、2.2 節的空間資料分群演算法，以及 2.3 節的時間序列預測原理介紹與應用

2.1 社群網路資料應用

社群網路中資料型態眾多，最常被人應用的是(1)文字識別(2)地理標籤照片應用兩種，現介紹這兩種資料被應用的狀況如下。

文字識別：顧名思義，讓電腦辨析一段句子、文章的意思，目前常見的方法是利用 LDA 進行關鍵字查詢，He *et al.* [8]將個人檔案中附有多屬性資訊資料，如：興趣旅遊類型、個人旅遊景點順序，旅程開始即抵達地點等文字資料，運用 LDA 模型，找出人們旅遊的順序及關鍵景點名稱，建立旅遊路線推薦系統，關於這類研究皆需要將原始文本分解找出具有意義的詞彙進行組合，而本研究著重於文章中找尋景點文字，常見的識別工具如 Python 開源中文斷詞程式庫 Jieba 斷詞及中研院 CKIP 斷詞系統，兩者最大差異在於 Jieba 自由且開放可設立專屬於自己的詞庫進行斷詞，而中研院 CKIP 相對封閉不過斷詞相對精準，綜合評估後本研究所需景點資料名稱組合，若使用 Jieba 斷詞需要自行定義許多景點名稱，且找尋斷詞後的景點名稱位置也是相對繁瑣以及 Jieba 與 CKIP 的慣用用語不同，故採用中研院 CKIP 斷詞系統，找出常見的詞性組合作為景點名稱的篩選。

地理標籤照片應用：近年來社群網路發達地理標籤照片也隨之普及，在這樣的科技背景下許多相關研究利用這些資料找尋使用者的旅行軌跡、辨識身分，如 Mukhina *et al.* [9]運用 Instagram 照片中地理標籤資料，計算持續上傳時間 15 天以上且為距離相近的地理位址辨別為當地居民，反之為遊客；Chareyron *et al.* [10]運用 Flickr 照片中記錄攝影師地理標籤與時間軸，在拍攝數量與造訪數量進行比率[10]，當比率小表示主要拍攝路徑機率愈低，反之亦然。如此，我們便能上述方法提取攝影師的主要拍攝路徑。

2.2 空間資料分群演算法

空間資料分群的演算法大致上可以分為切割式、階層式、密度式，以及網格式四種。

切割式：將切割資料點為 *K* 群之方法。MacQueen [11]提出 *K*-means 演算法，其以資料重心為基礎判斷，使用者必須定義資料集欲切割的群體數 *K*。Liu *et al.* [12]為了提升工業生產效能，找到中國工業的能源消耗使用情形，運用 *K*-means 找到每個公司能源與水的使用平衡，搭配關聯規則找出生產過程、公司與能源使用的關聯性，協助確立個體公司環境績效。

階層式：以樹狀結構呈現層級特性之方法，可分成凝聚法(Agglomerative)及分裂法(Divisive)。凝聚法是由下而上(Bottom-up)凝聚而成，演算法一開始先將資料庫內每一筆資料當作一個群聚，接著依資料屬性的相似度開始做合併，每次合併兩個相似度最高的群聚，直到所設定的終止群聚數目為止。分裂法是為由上而下(Top-down)分裂，將資料庫內所有的資料視為同一個群聚，然後將資料相似度低的分裂呈不同群聚，直到群聚數目為所設定之終止數目為止。而上述兩種分類中，又以凝聚法最常為人所用，如 AGNES 演算法就是凝聚法的經典案例。AGNES 演算法最初將每個資料點作為一個簇，由下開始讀取資料點，而該資料點會根據同一準則將相似度一樣者依照讀取順序逐步合併，反複進行直到滿足設定的簇數。

密度式：於特定單位面積下，所擁有資料點數之方法。經典密度式演算法為 DBSCAN，Ester *et al.* [13]提出目標半徑(*Eps*)內及最小數量(*MinPts*)下擴增凝聚點範圍，其演算法為將資料點標記為點位 *P* 後，當點位 *P* 符合 Eps 與 *MinPts* 規則下，該點位 *P* 便能成為一個簇 *C*，反之該點位 *P* 內點數不足，並辦定為雜訊(Noise)，接者讀取 *N* 內的點位 *P'*再計算 *EPS* 內點數 *N'*是否達到 *MinPts*，最後將 *N* 與 *N'*進行聯繫將條件內的點位 *P'*重複判斷是否為同一簇 *C*，直到不能再向外擴張群聚範圍後，再隨機讀取任意尚未被判斷過的點位 *P*，直到每一點位皆有所屬屬性。

網格式：此為將資料轉成固定大小網格之網格式方法。STING 演算法為通過多解析度的聚類技術而形成的演算法。STING 演算法通常會將網格劃分為矩形單元並針對不同級別的解析度形成了一個層次結構。通常高層底下存在多個低層單元，這些單位通常用於計算並儲存相關的屬性資訊。在 STING 演算法當中由於每個單位中的資訊都是獨立的，固查詢資料時並不需要依賴彙總的資訊同時有利於進行處理和增量與資訊更新。

小結：過往研究中大多使用基於中心的 *K*-Means 與基於密度的 DBSCAN 進行分群，然而 *K*-Means 於分群時需決定應分的群數，第一回隨機找點後透過找尋每群的中心點而重複執行計算與所有點位的距離，直到前一回找出的中心點與下一回相同才停止。DBSCAN 於分群時決定距離(*EPS*)及最小點數(*MinPts*)兩項參數，透過隨機決定一個點位後，計算點位間的密集度與挑出未滿密度閥值的異常點的特性，儘而得出分群結果。兩者方法比較之下儘管有 DBSCAN 資料的運算較低與排除異常點的特性，但是兩者於分群上皆為隨機讀點，較無直接的景點與遊客活動範圍關係上的關聯。因此應建立較為貼近景點與遊客活動的分群演算法。

2.3 時間序列預測原理介紹與應用

以時間序列研究的工具有許多，如基礎遞迴式類神經網路、長短期記憶型的類神經網路、迴歸模型等方法，本章節將列舉大多數研究者使用的模型背景研究如下。

ARMA：自迴歸移動平均模型(Autoregressive moving average model)是由自迴歸(AR)模型與移動平均(MA)模型組成，AR 模型主要為目前的數值與過去數值間的關係，在計算資料點 *k* 前的 *n* 個數值皆會乘上一個權重，並且加上常數項與白噪音後便可計算出 *k* 值。而 MA 模型為計算資料點 *k* 前 *n* 個數值的平均，再與前 *n* 個皆與計算出來的平均相減來獲得誤差，同時乘上 *n* 個資料點對應的權重值，最後再與平均相加獲得當前 *k* 的值。過去研究中 Zhang *et al.* [14]運用 ARMA 模型對

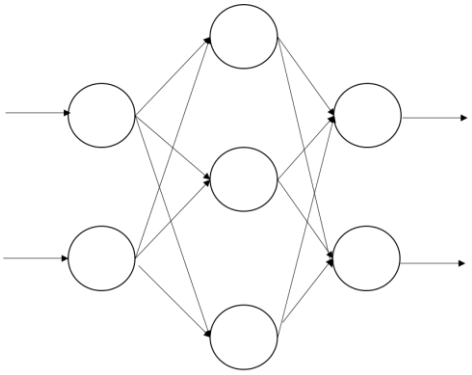

圖 3 基礎類神經網路架構圖

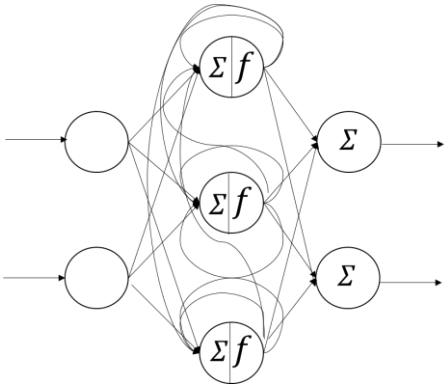

圖 4 Recurent Neural Network 架構圖

網路攻擊頻率進行一項長期預測，Liu *et al.* [15]以 ARMA 為基底進行了單變量時間序列預測。

隱藏式馬可夫模型：馬可夫模型為藉由當前事件的狀態來決定下一個狀態的機率分布，而隱藏式馬可夫模型增加了事件其他隱藏因子的考量，雖無法觀察目前的狀態，但可以藉由各個關聯的變數來推測當前的狀態。Tan *et al.* [16] 運用隱藏式馬可夫模型的特點建立 Pure Birth Process 模型，來預測網路影片的熱門程度。

類神經網路：類神經為一項常被用來在時間序列預測的工具，如金融投資、天氣預測等，相較線性的模型，類神經擁有較佳的非線性連續資料結果，類神經為由輸入層、隱藏層及輸出層所組成，輸入層根據神經元數與權重，相乘後建立矩陣，並根據使用者設定的激活函式在進入下一層時進行非線性轉換。由於每顆神經元皆與進行權重相乘，因此模型具有大量參數需要運算，使用者在建構類神經時跟依照需求調整神經元個數、激活函式、類神經層數等，因此具有良好的記性性來適應各種結構的資料。

用來進行時間序列的類神經網路有許多種類型，從圖 3 的基礎類神經網路、圖 4 的傳統的 Recurrent Neural Network (RNN)、Gated Recurrent Unit (GRU)至 LSTM 等。根據需求的變化，開發更多變體的神經網路。Chen *et al.* [17]應用模糊與類神經的概念建構 Kernel Regression Model 以及 Recurrent Neuro−Fuzzy Model 已進行預測；Wang and Chen [18]提出高運算速度的 Hammerstein-Wiener Recurrent Neural Network 結構進行預測。

LSTM：Long Short Term Memory Network 是由 RNN 改良而來，而 RNN 由 Schuster *et al.* [19] 於 1997 年提出，兩者主要差異在於隱藏層的部分，遞迴類神經網路除了將該次輸出傳遞給下一層外，還會將該次輸出返回給下一次輸入造成「記憶」的功能，這樣的作用可以促使下一次輸入可以取得上一次輸入的「記憶」做為參考，如圖 4 所示，因為有記憶的功能，所以可以同時考量過去時間的資訊，讓訓練結果更為準確，但也因為 RNN 在模型訓練的過程由於模型設計考量到之前的記憶，當輸入的資料越來越多，在不斷將上一次輸出返回給一次輸入進而造成「梯度爆炸」的發生，因此，Hochreiter and Schmidhuber [26]於 1997 年提出 LSTM 的模型架構，然而現今主流的 LSTM 架構是由 Gers *et al.* [20]於 1999 年提出的，新增了三種不同的 Gate 是去調整模型的權重。首先是修改 Input gate，在訓練過程中，讓模型自我學習判斷是否將這次的資訊輸入；第二是修改 Memory cell，判斷在經過計算後的資訊是否要儲存起來，給予下一個輸入做使用；最後是 Forget gate，Forget gate 可以自我學習判斷是否將儲存的資料清掉，可以影響下一個輸入的結果。

旅遊議題中的時間序列預測：此議題大部分是探討在連續時間下建立時間序列模型來探討旅遊區的人數變化。如 Ö nder [21]調查了潛在遊客在網路關鍵字和圖像查詢，以預測維也納的旅遊需求。Chu [22]使用 ARMA 模型預測九個亞太國家旅遊人數的建模和預測旅遊人數，Cho [23]則比較統計指數平滑法、ARIMA、Artificial Neural Networks (ANN)，預測六個國家(美國、英國、新加坡、日本、台灣和韓國)到香港的遊客到訪情形，結果為 ANN 準確率最高，所以該作者建議使用 ANN 來作為旅遊上時間序列的預測模型。然而就上述的研究看來，目前旅遊議題中時間序列預測的方法尚未使用到目前常用的方法，因此有持續精進的必要性。

第三章 研究方法

本章內容可分為 Offline 與 Online 兩部分，如圖 5 所示。首先 Offline 從資料處理至旅遊區人數預測，共可分為三大節進行說明，3.1 節為蒐集全台景點資訊，包含顯性(一般旅遊網站)及隱性(社群網路)景點，3.2 節為整合景點名稱及地理標籤資料，辨析旅遊區的劃分，3.3 節為旅遊區遊客人流數目時間序列建模。在前三節完成 Offline 步驟後，3.4 節則是描述 Online 部分如何進行旅遊區遊客人流的預測。

3.1 蒐集全台景點資訊

全台景點資訊分布主要可從一般旅遊網站及社群網路蒐集，兩者資訊處理方式有所不同，故本節次可分為 3.1.1 景點資料介紹及 3.1.2 顯性/隱性景點資料清洗及統整進行說明。

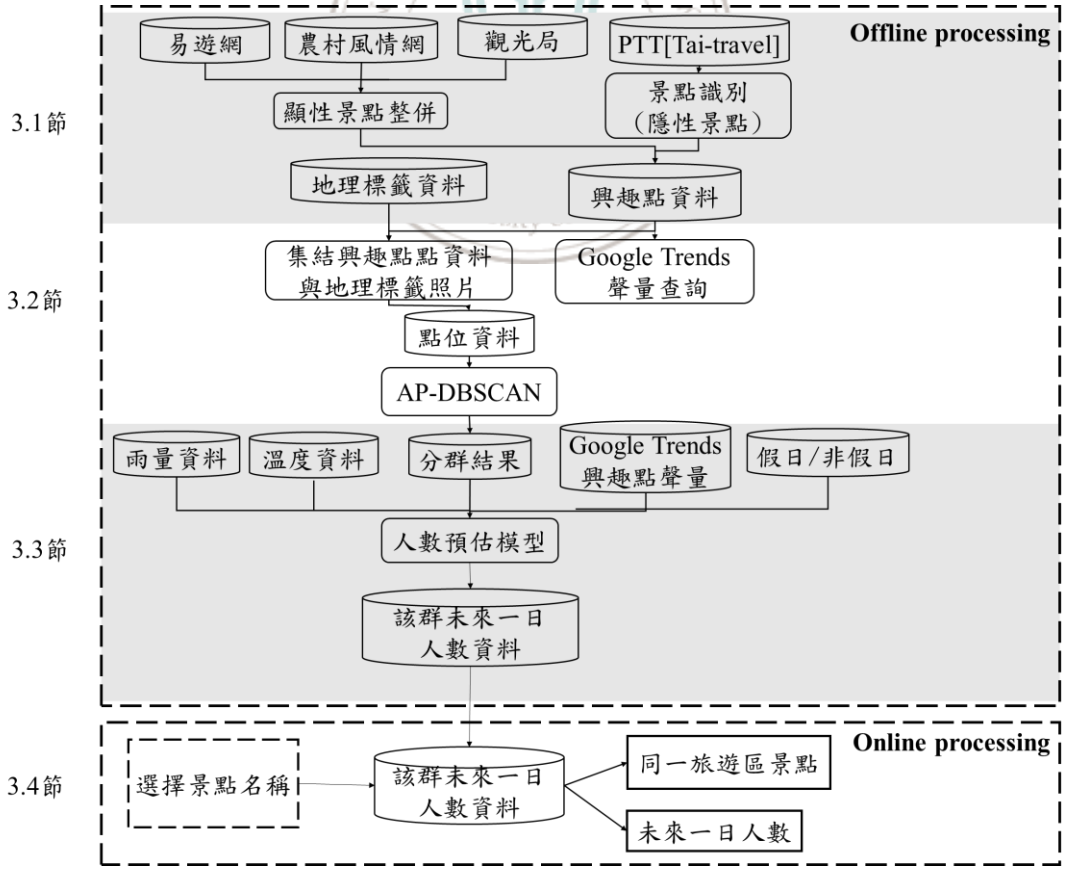

圖 5 研究流程圖

圖 6 交通部觀光局介面 (資料來源：交通部觀光局旅遊網站 https://www.taiwan.net.tw/)

圖文網誌版：http://blog.yam.com/inoueminami/article/65098769

因為老婆跟高中同學的聚會，

使得我們又再次來到清境農場散心。

除了跑到很低調的貝卡巧克力莊園參觀，

也到了之前沒去的小瑞士花園，

圖 7 PTT 網路爬蟲圖

3.1.1 景點資料介紹

在旅遊的過程中，遊客大多會將旅遊過程分享於社群網路中，為了盡量蒐集遊客於旅遊過程中所留下的足跡及較全面的景點資訊，本研究除了使用合作單位所提供農村風情網、易遊網旅遊景點資料及 Flickr 資料集外，也運用網路爬蟲技術採集交通部觀光局所提供的景點名稱與座標位置增進景點資料的完整度，如圖 6 所示；同時採集台灣熱門 BBS 網站 PTT 中 Tai-Travel 討論版(圖 7)內文，擷取網友們分享的旅遊經驗，從中找出他們曾經到訪的景點資訊，以彌補一般常見旅遊網站所缺漏的旅遊景點資料。

3.1.2 顯性/隱性景點資料清洗及統整

隨著社群媒體的興起，讓許多「隱藏版」景點被大家挖掘，特別在旅遊版網路日誌上，許多使用者於該社群網路平台上分享相關資訊，紀錄人們於旅遊時的旅遊心得，本研究預想從使用者發佈於網路上的內文中判別該使用者曾經造訪或提及過的景點名稱，找尋尚未被記錄在一般旅遊網站上的隱性景點，圖 8 為隱性景點的找尋方式。

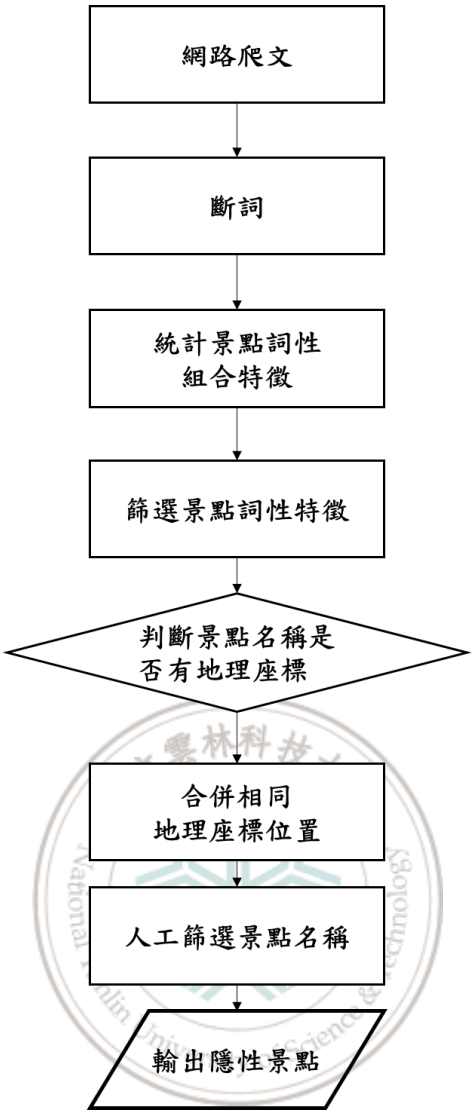

圖 8 找尋隱性景點流程圖

首先我們將交通部觀光局所提供的景點名稱運用中研院 CKIP 斷詞系統進行斷詞，統計景點常見的組合詞性，並以景點名稱常見的詞性組合(ex: Na + Nb)作為篩選出候選景點名稱。其次我們將擷取下來的文章運用中研院 CKIP 斷詞系統，依照中文句構詞性切割出有意義的單字詞，結果如圖 9 所示，但由於此步驟並無法確立所有文字組合都是景點名稱，所以將這些候選景點再手動放入 Google API 查詢該文字組合是否回傳地理座標，並清除無地理座標之文字組合及合併相同地理座標，最後人工篩選景點名稱放入隱藏景點資料集。

使得(VL) 我們(Nh) 又(D) 再次(D) 來到(VCL) 清境(Na) 農場(Nc) 散心(VA) 。(PERIODCATEGORY) 除了(P) 跑到(VCL) 很(Dfa) 低調(VH) 的(DE) 貝卡(Nb) 巧克力(Na) 莊園(Na) 參觀(VC) ，(COMMACATEGORY) 也(D) 到(VCL) 了(Di) 之前(Ng) 沒(D) 去(VCL) 的(DE) 小(VH) 瑞士(Nc) 花園(Nc) ，(COMMACATEGORY)

圖 9 斷詞結果示意圖

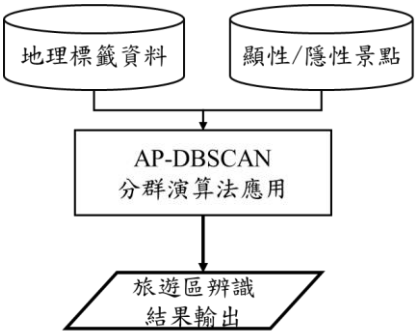

圖 10 旅遊區辨識流程圖

表 1 Flickr 資料格式

PhotoId	OwnerId	TakenTime(拍照時間)	緯度	經度
25418700653	101082573@N05	2016/3/25 00:33	23.7519	120.25
25929989882	91280188@N08	2016/3/25 02:02	24.1645	120.639
25959259461	91280188@N08	2016/3/25 06:02	24.1246	120.675
27659102400	12192309@N03	2016/3/25 10:06	23.869	120.947
27860749761	12192309@N03	2016/3/25 10:07	23.869	120.947
27836308642	12192309@N03	2016/3/25 10:08	23.869	120.947
27659101850	12192309@N03	2016/3/25 10:09	23.869	120.947
27860749291	12192309@N03	2016/3/25 10:09	23.869	120.947
27903589786	12192309@N03	2016/3/25 10:24	23.8684	120.948

3.2 整合景點名稱及地理標籤資料，辨析旅遊區的劃分

在確立所有隱性景點名稱皆具意義後，我們接著會將顯性及隱性景點統整為景點資料集，搭配地理標籤照片資料以我們自行設計的 AP-DBSCAN 分群演算法，把較為緊密的景點以及地理標籤劃分成同一旅遊區，藉此看出遊客於該景點的活動空間分佈。本節分為 3.2.1 Flickr 地理標籤照片資料集說明；3.2.2 建構 AP-DBSCAN 演算法，以及 3.2.3 的演算法參數設定。

3.2.1 地理標籤照片資料及說明

Flickr [6]為知名網路照片分享平台之一，他最大的特色是可以讓使用者上傳 5K 高畫質照片，而這點在所有照片分享平台中非常少見，因此該平台吸引了許多旅遊者上傳照片，並因此收集到大量的照片資料集。而對本研究來說，如此龐大的照片資料集最適合用來解決本論文所提出的問題，因此本研究使用 Flickr 資料集來做為旅遊區識別的基礎。目前 Flickr 所釋放出來的照片資料集格式如表 1 所示，共計有五個維度，其中 PhotoId 為照片 ID，OwnerId 為使用者 ID，TakenTime 為照

片拍攝時間、拍照所在緯度及經度。至於本論文所使用之資料集則以 2011/03/01 開始到 2018/02/28 為止，在台灣南投日月潭地區所拍攝之照片為主。

3.2.2Appointed-DBSCAN 分群演算法模型說明

Appointed-DBSCAN (AP-DBSCAN)演算法為本論文所提出，適用於在已擁有某些興趣點與大量空間資訊點的地圖上，以興趣點為中心來對空間資訊點進行分群的演算法。舉本論文之問題為例，若我們已知某些興趣點(i.e.顯性景點與隱性景點)及大量空間資訊點(i.e. Flickr上的地理標籤照片)，則我們會以這些景點為中心，來對 Flickr 上的地理標籤照片進行分群，其分群後的每一群必然會包含一個以上的景點，且群內的 Flickr 地理標籤照片必然是與這些景點有關聯者。而這些分群的結果即為本論文所提及的旅遊區。

我們所設計的 AP-DBSCAN 演算法是改良 DBSCAN 演算法而來，其執行步驟有相似的地方也有不相似之處，現介紹如下。首先，AP-DBSCAN 演算法與 DBSCAN 執行步驟相似，皆需要設定兩項參數：目標鄰近區域的半徑 *EPS* 以及最小存在於該鄰居區域的資料點數 *minPts*，而這兩項參數都能讓研究者依據目標分析的城鎮不同而調整。接著，AP-DBSCAN 演算法與 DBSCAN 演算法最大的差異在於 DBSCAN 為非監督式分群演算法，且在初始挑選資料點位時採隨機挑選分群點位 *P* 的方式進行，而本研究所提出的演算法為半監督式分群演算法，且用來分群的點位 *P* 是以使用者所指定的興趣點為主。

AP-DBSCAN 演算法的虛擬程式碼如圖 11 所示。一開始演算法會在使用者指定的空間範圍內找出興趣點 *P* (line 1)，並以這些興趣點為中心向外找尋鄰近的空間資訊點 *p*'數目(line 3)，如圖 12 所示。在這步驟中，為了節省演算法的運算時間，我們不使用傳統 DBSCAN 每個資料點都必須兩兩計算距離的作法，而是使用了正方形篩選法(RegionQuery)先限制資料範圍並因此減少距離的計算次數，如圖 13 所示。又因為本研究所使用的資料為地理空間上的資料，所以我們是採用歐幾里得距離(Euclidean Distance)來計算兩點之間的距離，其公式(1)所示。最終，此步驟找

到的空間資訊點數目會儲存在 *Neighborpts* 這個變數中。

Algorithm 1 : Appointed-DBSCAN

```
def Main_ExpandCluster(Spot,Coordinate,eps,MinPts):
    for P in range(len(Spot)):
        Coordinate = Coordinate is unvisited
        NeighborPts = RegionQuery(P, Coordinate,eps,data)
        if len(NeighborPts)>=MinPts
          mark Spot.C'[P] = P
          p'=0
          while p'<len(NeighborPts):
            sub_NeighborPts=RegionQuery(p',Coordinate,eps,NeighborPts)
            if len(sub_NeighborPts) >= MinPts:
               NeighborPts = NeighborPts + sub_NeighborPts
               NeighborPts. C.index in Coordinate = P if NeighborPts. C==unvisited
            else:
               NeighborPts. C.index in Coordinate = P if NeighborPts. C ==unvisited
            p'!=1
        else:
          mark Spot.C'[P] = Noise
    return NeighborPts.C, Spot.C'
def RegionQuery(P,Coordinate,eps,data):
    Coordinate_region=Coordinate [(Max Latitude)&(Min Longitude)&(Min Latitude)&(Max
    Longitude)]
    for x in range(len(Coordinate_region)):
        dist=Euclidean Distance(P, x)
    if dist<eps:
        NeighborPts.append(Coordinate_region[x])
    return NeighborPts
```

圖 11 AP-DBSCAN 演算法的虛擬程式碼

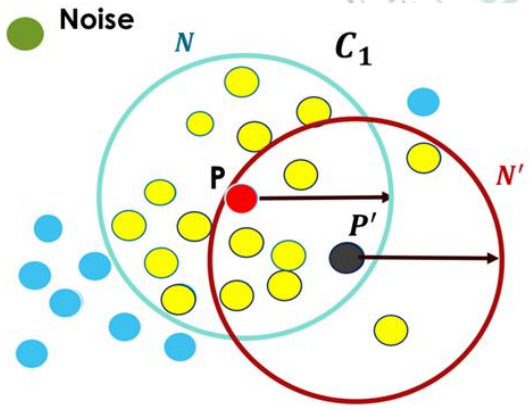

圖 12 AP-DBSCAN 概念示意圖

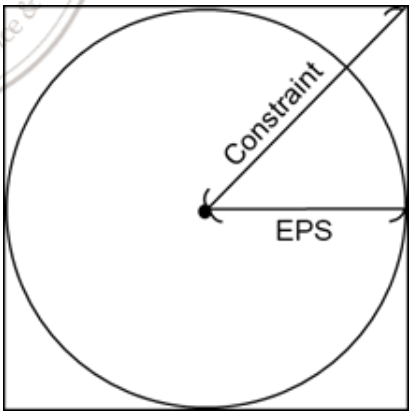

圖 13 限制式範圍

$$dist(P,X) = \sum_{i=1}^{n} \sqrt{(P - X_i)^2} \qquad (1)$$

接著，我們會判斷 *Neighborpts* 的數值是否大於使用者所設定的鄰近點數目門檻值 *Minpts*。若有，則該興趣點 P 便認定可成立群集 C，反之則不足以成立一個

群集 *C*，並判定為雜訊(Noise)(i.e. line 15-16)。接著，若一個興趣點 *P* 能成立一個群集 *C*，則我們會將 *P* 周圍的空間資訊點或是其他使用者指定的興趣點視為新的興趣點(i.e. line 4-5)，而演算法則會持續判斷這些新興趣點的周遭空間資訊點數目是否周遭有超越 *Minpts*，若有，則會把新的鄰近點加入原有的群集 C 中，否則則結束這個新興趣點的檢查並轉移檢查其他的興趣點(i.e. line 7-14)。整個步驟則會持續到演算法中沒有新的興趣點為止。

本研究提出的 AP-DBSCAN 演算法不只保有 DBSCAN 具有密集度越高愈容易形成一群、密度愈低或零散視為雜訊的特性。更重要的是，我們可以依照使用者所設定的興趣點，將其鄰近有關聯的空間資料點分配到同一個群體中，解決了傳統 DBSCAN 無法考慮興趣點的問題。而這樣的演算法恰能用於本論文在已知某些顯性景點與隱性景點的狀況下，找尋與這些景點有關聯的 Flickr 照片資料標籤，並依據這些資料標籤範圍繪出這些景點旅遊區的範圍。

3.2.3AP-DBSCAN 演算法參數設定

一般來說與 DBSCAN 類似的演算法，其效能通常與使用者設定的 MinPts 與 Eps 絕對相關。當單一 MinPts 下，Eps 愈小其 Noise 愈多；反之 Eps 愈大，Noise 愈少。而本篇論文所提出的 AP-DBSCAN 演算法也會有類似的問題，因此本論文將使用 Rousseeuw [24] 提出之 Silhouette Coefficient 來做為 AP-DBSCAN 演算法效能的判斷。這個演算法能夠協助我們計算群間及群內資料點的相似度，其計算出來的相似度數值會介於[-1,1]。當係數愈靠近 1 時，表示群體間距離遠且群內點位數多，是較好的分群結果。而對我們的演算法來說，我們會在多組實驗下，挑選分群結果之 Silhouette Coefficient 最靠近 1 者為最終參數。

3.3 旅遊區遊客人流時間序列建模與預測

在完成各景點旅遊區的劃分後，本研究以旅遊區的概念進行時間單位的人數預估，其預估概念如圖 14 所示。本部分演算法的目的不在於精準預測每天旅遊區

人數量，而是提供人數疏密程度供民眾參考，畢竟影響人數因素有許多，而本研

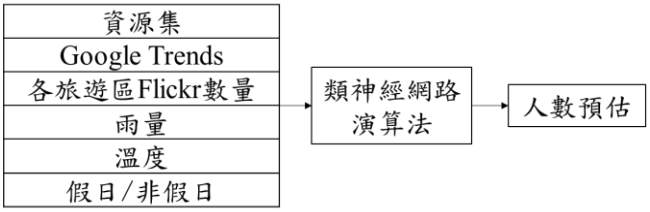

圖 14 旅遊區概念模型圖

究所能蒐集到的資訊有限，不可能做到精準預測，但所獲得的結果仍然可以提供旅遊相關單位為參考資訊。以下首先說明本研究所收集到用來預估遊客人流數目的資料來源，接著說明我們所使用的模型。

3.3.1 遊客人流數目量預測的資料集

本研究針對影響旅遊區遊客人流主要因素資料進行蒐集，共可分為五種，分別是 Google Trends 景點聲量、該旅遊區的歷史氣溫、雨量、假日非假日及 Flickr 地理標籤照片，其中前四項為預測模型的輸入資料，最後一項則為模型的輸出資料。

我們所收集的第一種資料為 Google trends 資料集，所謂的 Google trends 資料集是統計每個關鍵詞每天在 Google 上被查詢次數的開放資料集。一般來說民眾在出遊前都會透過搜尋引擎搜尋景點資訊，像是交通方式、營業時間、注意事項等。雖然民眾查了不一定會去，但仍有許多研究者認為 Google trends 能充分表達民眾出遊的意向而有參考價值。因此本研究會使用這個資料集來預測旅遊區未來的遊客人流數目。接著，我們第二與第三種收集的資料為氣溫與雨量。會使用這些資料，主要是因為天氣因素通常被視為遊客是否會成行的關鍵[5]。至於第四種資料為人事行政局公告之行事曆，其紀錄假日及非假日的日期為遊客於安排旅遊日期的關鍵考量之一，最後第五種資料為 Flickr 上該旅遊區當天的拍照人數量，會使用此資料當做模型的輸出，則是因為拍照人數量能夠準確反應出有多少遊客真的前往當地而有參考價值。當然我們也必須了解到不一定每位遊客都會把照片上傳到 Flickr 上，所以該數值有時會失真並影響到預測結果。現介紹我們所收集到的五種

資料如下：

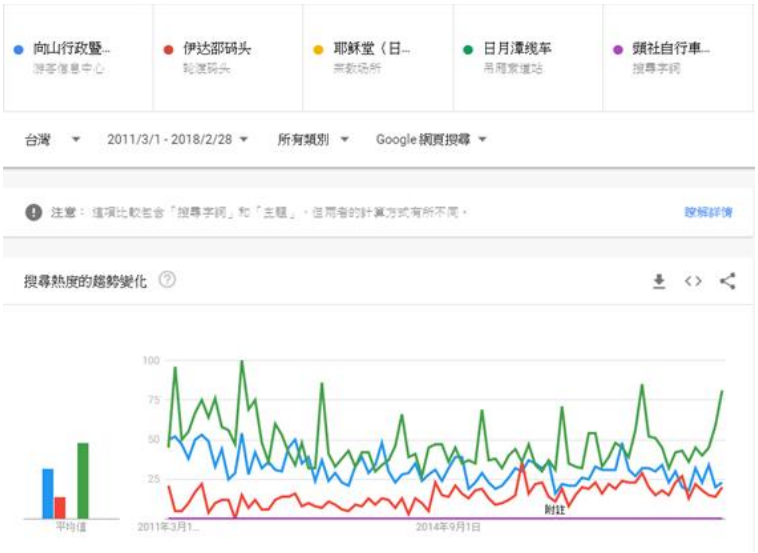

圖 15 GoogleTrends 介面

表 2 聲量資料欄位

日期	苗栗勻淨湖(值域範圍 0-100)
2016/09/01	75
2016/09/02	20
2016/09/03	10
2016/09/04	15
2016/09/05	18

- **Google Trends 興趣點聲量**

資料集說明：同一旅遊區通常包含許多顯性景點與隱性景點，但我們通常無法得知哪個景點的 Google trends 聲量數據對遊客人流預測最有幫助。故本研究在預測未來遊客人流時會將所有景點的 Google trends 聲量數據都納入考量。圖 15 為 Google Trends 平台中輸入該旅遊區內所屬景點之關鍵字時間序列聲量服務示意圖，而我們從 Google Trends 平台上下載到的資料格式則如表 2 所示：

資料前處理-連比例：由於各興趣點P_i的相對聲量於 Google Trend 採集時最多只能搜尋 5 項，故於單次聲量下載須設置共同興趣點，再經由資料連比例得出所有興趣點。連比例公式如下

$$P_i : P_{i+1} = a : b \text{則} P_{i+1} : P_{i+2} = b : c \qquad (2)$$

- **每日雨量歷史紀錄**

本論文會使用中央氣象局之紀錄，從中取出目標旅遊區每天的天氣狀況，並

作為後續預測使用，其中我們所收集到的資料格式則如表 3 所示。

表 3 該日雨量資料欄位

日期	雨量(mm)
2016/09/01	0
2016/09/02	0
2016/09/03	0
2016/09/04	0
2016/09/05	0

表 4 該日溫度資料欄位

日期	溫度(^{O}C)
2016/09/01	20
2016/09/02	20
2016/09/03	19
2016/09/04	19
2016/09/05	19

表 5 假日/非假日資料欄位

日期	假日：0/非假日：1
2016/09/01	0
2016/09/02	0
2016/09/03	1
2016/09/04	1
2016/09/05	0

表 6 Flickr 數量欄位

日期	群 1 內 Flickr 拍照數量
2016/09/01	20
2016/09/02	15
2016/09/03	150
2016/09/04	114
2016/09/05	3

- **每日溫度歷史紀錄**

本論文會使用中央氣象局之紀錄，從中取出目標旅遊區每天的天氣狀況，並作為後續預測使用，我們所收集到的資料格式則如表 4 所示。

- **假日/非假日歷史紀錄**

本論文採用行政院人事行政總處每年公布之辦公日曆表來判斷一天是否為假日，若為假日為 0，非假日則為 1，其資料格式則表 5 所示。

- **目標旅遊區內的歷史 Flickr 拍照數量**

本論文採用目標旅遊區內的歷史 Flickr 拍照數量來做為模型預測的目標。而該數值則是統計每日拍照位置在該旅遊區的數目，並整理成如表 6 的格式。

3.3.2 旅遊區遊客人流數目時間序列模型之建立

本論文使用深度學習模型中，知名的 LSTM 模型來進行旅遊區遊客人流數目時間序列模型之建立。本論文會使用這個模型，是因為其有三種不同的 Gate 能調整模型的權重，且其已被多方驗證在時間序列上的建模效果極佳。他所擁有的三個 Gate 分別是，Input gate、Memory cell，Forget gate，現分別介紹這三個 Gate 的概念如下。首先是 Input gate，他能在訓練過程中讓模型自我學習判斷是否將這次的資訊輸入。第二個是 Memory cell，他會負責判斷計算後的資訊是否要儲存起來，並給予下一個輸入做使用。最後是 Forget gate，他能夠協助模型自我判斷是否將儲

存的資料清掉，來減少對模型的影響。

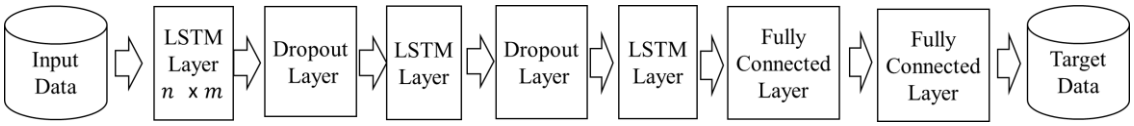

圖 16 LSTM 模型架構圖

本研究 LSTM 模型架構如圖 16 所示，共有 1 層輸入層、3 層 LSTM 層、2 層 Dropout 層，2 層全連接層，並在最後輸出預測結果。不過由於本論文並未設計新的 LSTM 模型，故我們不再贅述裡面的內容，讀者們可以參考[20][27]這兩篇論文來獲得更多的細節。此處我們僅就本論文客製化的模型輸入與輸出進行介紹。首先輸入層會有 $m\times(4+n)$個輸入，其中 m 表示我們考慮的歷史資料時間，4 代表雨量、溫度、假日/非假日，以及該旅遊區過往拍照的照片數量，n 則代表該旅遊區中有多少顯性及隱性景點，而這些景點的 Google trends 數值將會被個別輸入模型中。而在輸出結果部分，我們主要預測該旅遊區隔日會有多少使用者上傳 Flickr 照片地理標籤資料。會使用這筆資料，是因為本論文在完成之前尚無法獲得任何真實遊客人流資料，故僅能使用這筆最接近真實遊客人流數目的資料來替代真實的遊客人流數目。

3.4Online 說明

在完成上述的歷史遊客人流數量建模後，我們即可使用訓練好的模型進行旅遊區遊客人流數量的預估。要使用這個功能，一開始使用者必須從我們的系統中挑選他希望預測的旅遊區。接著，我們的系統會自動抓取使該旅遊區遊客人流預測模型所需的輸入，包含 Google Trends 資料、每日雨量歷史資料、每日溫度歷史資料、假日/非假日歷史資料，以及前幾日的目標旅遊區內的歷史 Flickr 拍照數量。而在輸入資料準備完成後，系統即會利用訓練好的模型進行未來遊客人流預估，並將預估結果輸出給民眾。

第四章 實驗與討論

此章節介紹我們的驗證實驗，包含了 4.1 節的資料集介紹、4.2 節的顯性及隱性景點結果、4.3 節不同分群參數結果驗證，以及 4.4 節的旅遊區每群人數預估結果與其他模型的比較。此外，本研究所有實驗都是在 64 位元作業系統上運行 Windows 10 的 Python 和 3.6GHz 的 Intel Core i7-8700 處理器，8 GB 內存以及顯示卡為 Intel® UHD Graphics 630 實現的。

4.1 資料集說明

本研究挑選南投縣日月潭風景區為研究區域(如圖 17 所示)，其挑選原因為日月潭為環狀地形，遊客可搭乘遊艇觀賞美景，也可騎乘腳踏車環湖四周，遊客活動皆在日月潭四周，也因此較適合本研究驗證隱性景點之搜尋結果、我們所設計之分群演算法效能，以及遊客人流數目量預估模型之準確性。

在資料集部分，本研究用來找尋隱性景點的社群網路文章為 PTT[Tai-travel]版文章資料集，會使用該資料集是因為 PTT 為目前台灣最大的 BBS 站，每日使用人口都在 15 萬上下，且該平台資料集易於取得且無限制，適合用於研究上。至於旅遊區辨識的資料則使用 Flickr 資料集，其中我們會把拍照地點在經度範圍

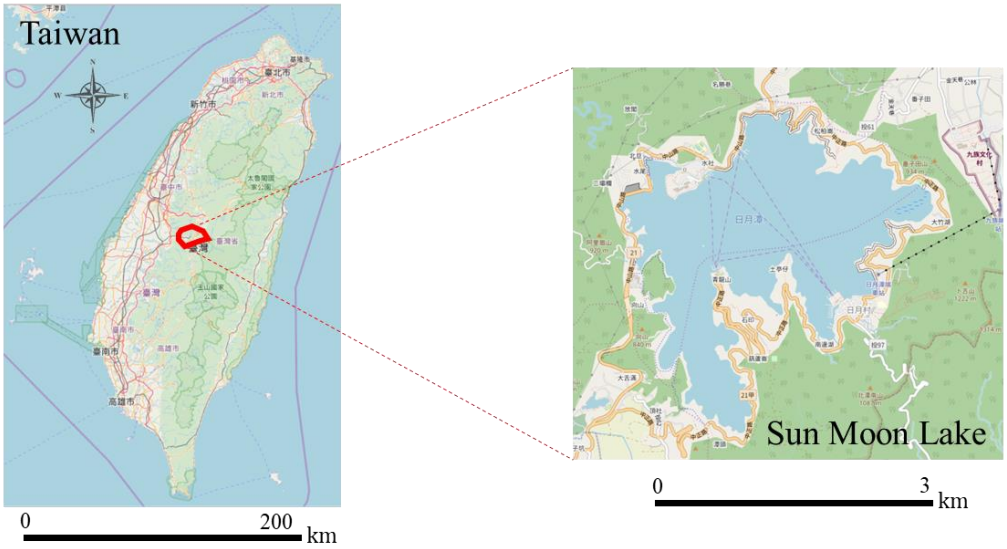

圖 17 日月潭風景區示意圖(底圖為 Open Street Map)

[120.884807, 120.949609]與緯度範圍[23.826394, 23.877340]視為在日月潭周邊拍照的照片。而在遊客人流建模與預測部分，如同我們於 3.3 節所討論的，我們會收集 Google Trends 資料集、雨量與溫度資料集，以及是否放假的資料集來做為建模的基礎。最終，所有的資料集其收集時間皆為 2011 年 3 月 1 日至 2018 年 2 月 28 日逐日資料，共 2557 日。會使用這個時間區段資料，是因為這個時間區段是目前所有資料集能抓取到的時間範圍交集。而在這個時間區段下，本研究共收集到 PTT[Tai-travel]版上共 10230 篇文章(不限於我們所設定的日月潭區域)，以及 Flickr 上照片 16,686 筆(僅限於我們所設定的日月潭區域)。

4.2 顯性及隱性景點結果

藉由本論文所設計的隱性景點萃取方式，我們共能從 PTT[Tai-travel]上的文章找到顯性景點 26 個，以及隱性景點共 9 個，如表 7 與表 8 所示。而這些景點在地圖上的視覺化結果則如圖 18 所示，其上的文字則為隱性景點的名稱。在該圖中，我們可以看到右側方框區有阿拉丁廣場和娜魯灣劇場兩個隱性景點，而這兩個景點都是九族文化村的知名內部景點，但官方一般不會把他們當成旅遊景點介紹，也不會被所有民眾熟知，所以屬於我們定義的隱性景點範圍內。而這樣的結果也說明了本研究方法的確能從社群網路文章中找到一些隱性的景點。然而，本研究發現到本研究所找出來的結果數量較預期的少。推測有兩個原因，(1)某些旅遊景點命名較為特殊且不符合詞學的構造，因此以本論文所使用的詞性規則難以找出這些景點。(2)本研究為受限於 CKIP 斷詞系統，故在斷詞及篩選結果上數量較少，未來建議可增設景點詞庫及斷詞演算法協助景點辨識。

4.3 AP-DBSCAN 結果

4.3.1 最佳參數選擇

本實驗經由 Silhouette Coefficient 找出較好的分群結果，各參數結果如表 9。在該表中，除了 Eps =0.1km 且 MinPts=1000 找不到結果時，其他組 Eps 與 MinPts

表 7 本論文方法從上述三個網站所擷取到的顯性景點

編號	顯性景點	經度	緯度
1	頭社自行車步道	120.896614	23.833705
2	向山自行車步道	120.901731	23.852933
3	向山遊客中心	120.902116	23.851145
4	日月潭國家風景區	120.903039	23.852166
5	日月潭頭社休閒農業區	120.903589	23.831724
6	耶穌堂	120.910185	23.863551
7	日月潭涵碧樓步道	120.910617	23.863615
8	拉魯島	120.911111	23.855556
9	水社碼頭	120.911825	23.864503
10	玄光寺	120.913604	23.852097
11	青龍山步道	120.913966	23.850292
12	日月潭	120.915913	23.857334
13	水社親水步道	120.916465	23.868553
14	玄奘寺	120.917130	23.847372
15	日月潭慈恩塔	120.920393	23.842364
16	日月潭文武廟	120.927429	23.869839
17	伊達邵碼頭	120.929674	23.849741
18	德化社	120.930253	23.848579
19	正心書院 明德宮	120.930300	23.847633
20	甘為霖紀念禮拜堂	120.930471	23.847908
21	孔雀園	120.931687	23.868312
22	日月潭伊達邵	120.932666	23.849410
23	日月潭纜車	120.934739	23.852248
24	大林休閒農業區	120.943621	23.875952
25	水社大山步道	120.946889	23.851107
26	九族文化村	120.947861	23.869107

表 8 本論文方法從 PTT[Tai-travel]版上所擷取到的隱性景點

編號	顯性景點	經度	緯度
1	伊達邵渡假旅店	120.9326	23.84771
2	年梯步道	120.927	23.86944
3	阿拉丁廣場	120.9464	23.86804
4	阿婆茶葉蛋	120.9131	23.85249
5	日月潭水社遊客中心	120.9111	23.86646
6	娜魯灣劇場	120.9501	23.86224
7	松鶴園大飯店	120.9119	23.86616
8	雲品溫泉酒店	120.9235	23.87092
9	日月潭民宿漁村別館	120.9299	23.84813

都能找到分群的結果。但在比較這些結果後，我們可以看到(0.5km,1000)、(0.5km,500)、(0.3km,500)相較於其他參數而言，擁有組間差異大、組內差異小優勢。本研究運用這三種參數與分群演算法 DBSCAN 套件及 K-Means 套件進行比較。

圖 18 日月潭興趣點地理化結果

表 9 輪廓係數表

Eps MinPts	0.0179656 (2km)	0.0089828 (1km)	0.0044914 (0.5km)	0.00269484 (0.3km)	0.00089828 (0.1km)
1000	0.2943	0.2951	0.4819	0.1460	X
500	0.2943	0.271	0.4218	0.4370	-0.0500
200	0.2943	0.2020	0.2743	0.3840	0.1507
100	0.2943	0.2020	-0.0063	0.0691	0.2454
50	0.2943	0.2020	0.0790	0.0837	0.3478
10	0.2943	0.2020	0.2300	0.0837	0.1795

4.3.2 AP-DBSCAN 結果視覺化

- 參數與興趣點形成之旅遊區

本研究將興趣點資料集放入 AP-DBSCAN 演算法，並從興趣點資料集中顯性景點依序讀取至隱性景點，進行上述三項 AP-DBSCAN 演算法運算。參數一(0.5km,500)共分成 3 群、參數二(0.3km,500)共分成 6 群，最後為參數三(0.5km,1000)分成 4 個旅遊區；各參數視覺化後如圖 19 至 21，將旅遊區分門別類成對應的顏色以及所屬的 POI、核心 POI 以及雜訊 POI。

以下分為 3 個面向說明，首先為參數對範圍的影響，由三項參數可以發現日月潭風景區大致上分為三大區域，由圖 21 的玄光寺至文武廟、伊達邵碼頭以及九族文化村；半徑 EPS 的影響，如表 10 顯示三項參數於不同旅遊區及雜訊的 Flickr 點位數以及興趣點數量，隨著縮減半徑 EPS，如參數一(0.5km,500)至參數二(0.3km,500)，縮小 EPS 且 MinPts 不變，除了總群數增加也剔除更多雜訊，更能看

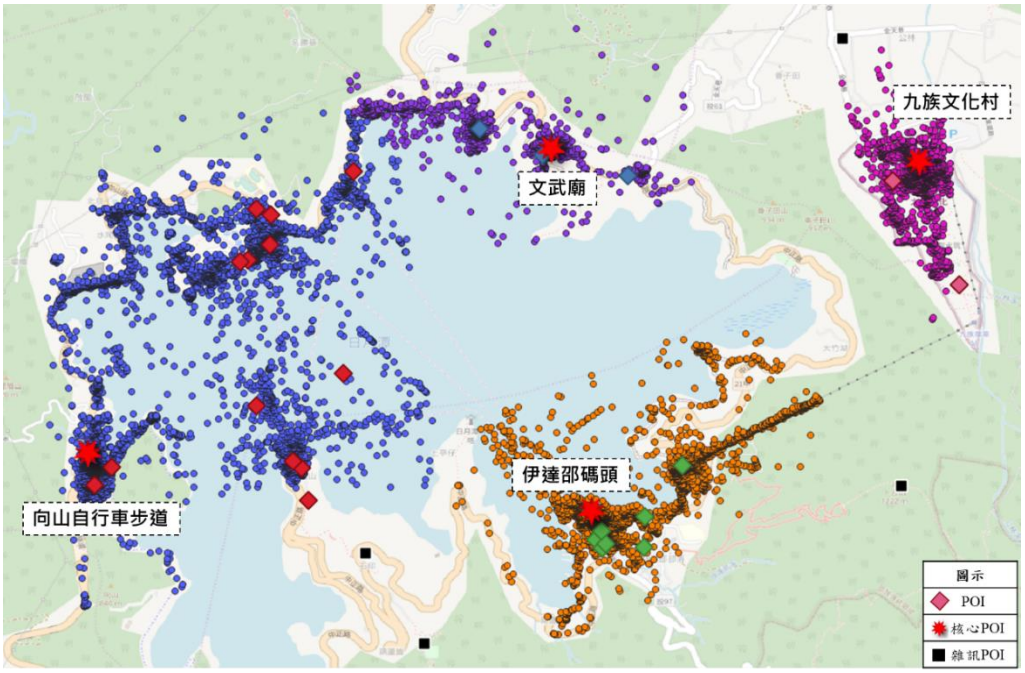

圖 19 AP-DBSCAN (0.5km,1000)

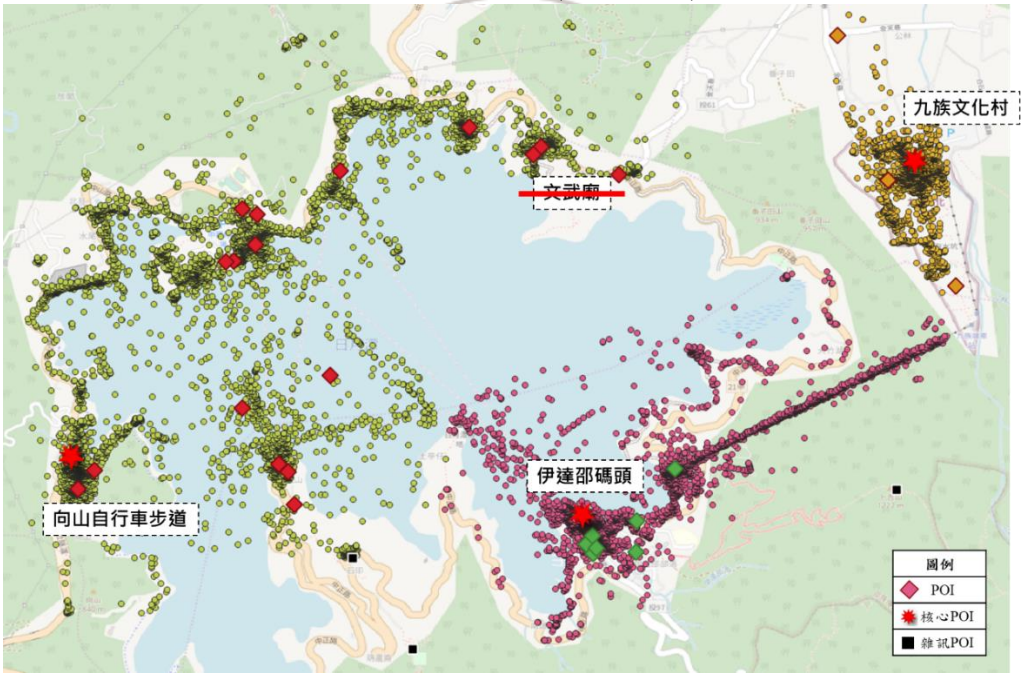

圖 20 AP-DBSCAN (0.5km,500)

出密度高的興趣點，反之亦然；最後為 MinPts 最小點位數的影響，如表 10 中參數三(0.5km,1000)至參數一(0.5km,500)，縮小 MinPts 且 EPS 不變之下密度閥值更為寬裕，單一旅遊區中有更多的關聯興趣點，如圖 21 中文武廟被併為與向山自行車步道同一旅遊區。

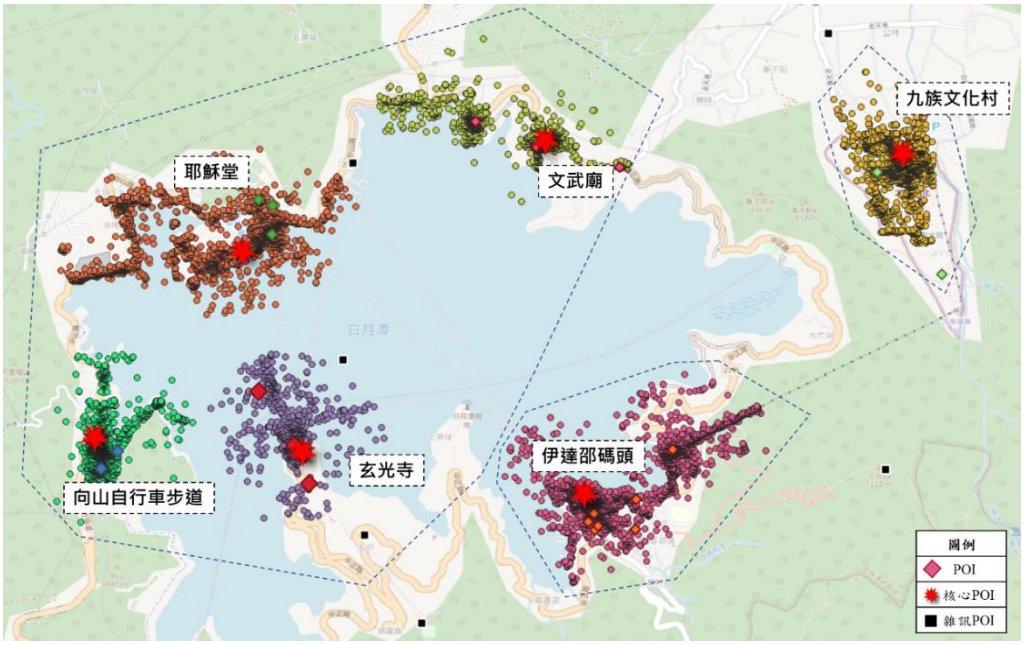

圖 21 AP-DBSCAN (0.3km,500)

表 10 各參數分群結果統計

(a) 參數為(0.5km,1000)

0.5km/1000	群 1	群 2	群 3	群 4	雜訊
點位數量	6677	2153	3206	2018	2635
景點數量	14	4	3	8	4

(b) 參數為(0.5km,500)

0.5km/500	群 1	群 2	群 3	雜訊
點位數量	2017	3778	9197	1701
景點數量	18	8	4	5

(c) 參數為(0.3km,500)

0.3km/500	群 1	群 2	群 3	群 4	群 5	群 6	雜訊
點位數量	2041	2909	1158	1905	2748	1922	4040
景點數量	3	5	4	4	8	3	6

- 雜訊與不同參數下的結果

根據 AP-DBSCAN 的判斷中，當興趣點被判斷為雜訊時，表示該興趣點半徑 EPS 範圍內的 Flickr 點位數不滿足設定的 MinPts 最小點位數，而觀察圖 21 到圖 23 中的雜訊時分為兩種結果。一種情況為不同參數下皆有雜訊及所屬旅遊區的結果，如大林休閒農業區。另一種是皆為雜訊的興趣點，如玄奘寺、慈恩塔以及水社大山步道。本研究於今年 2 月實地訪查後，針對參數二(0.3km,500)中水社親水步道進行合理性說明，在圖 23 中可觀察到 Flickr 點位位於水社親水步道頭尾兩側，經實地訪查結果，Flickr 點位密集處左側有商圈，如旅館、腳踏車店等，而該興趣點為一條長型的步道延伸至耶穌堂，形成遊客於左側商圈借腳踏車，騎乘腳踏車於該步道致右側旅遊區沿路欣賞日月潭風景後，折返回左側商圈歸還。

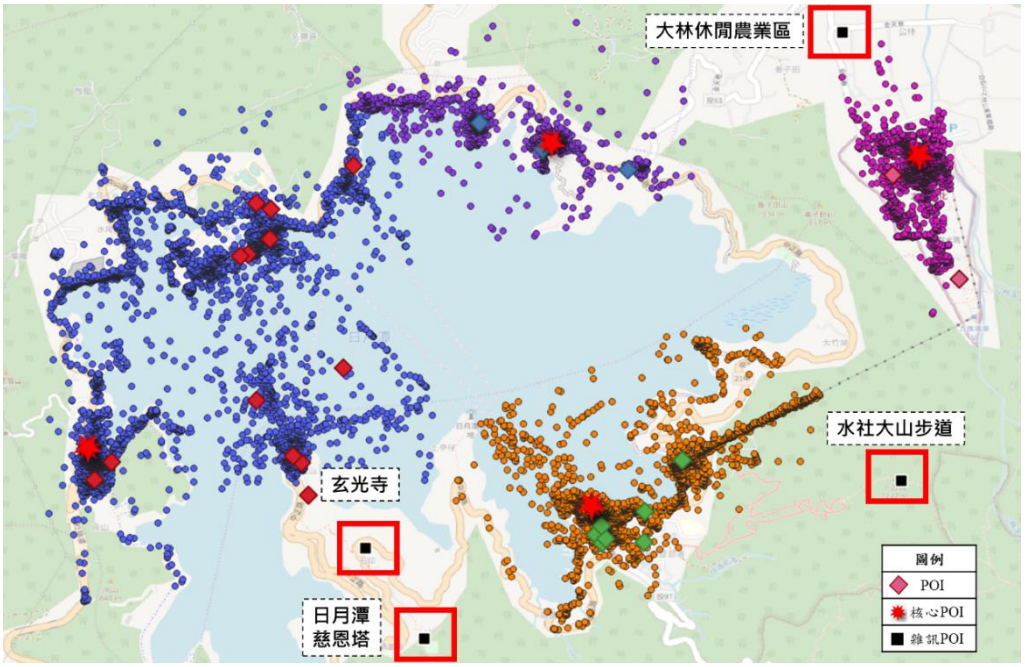

圖 22 AP-DBSCAN (0.5km,1000)雜訊圖

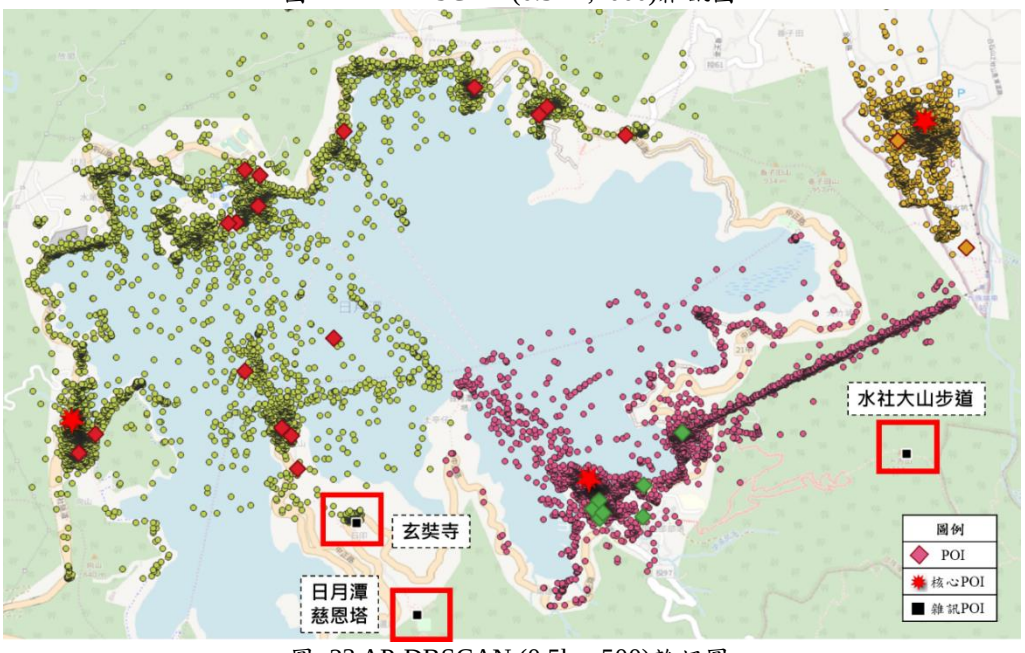

圖 23 AP-DBSCAN (0.5km,500)雜訊圖

另一種情況對九族文化村進行舉例，考量不同參數設定影響分群結果，觀察大林休閒農業區，圖 22 到圖 24 的結果有兩項參數被判別為雜訊，一項參數被辨別為擁有所屬旅遊區，經網站查詢大林休閒農業區內有許多農村體驗活動，如：採菇體驗，除體驗之外也有民宿，說明該農業區為長時間體驗型的興趣點，而九族文化村為遊樂園，也為長時間體驗型的興趣點，遊客於安排活動時部分遊客不會將兩個興趣點編列為同一天行程。

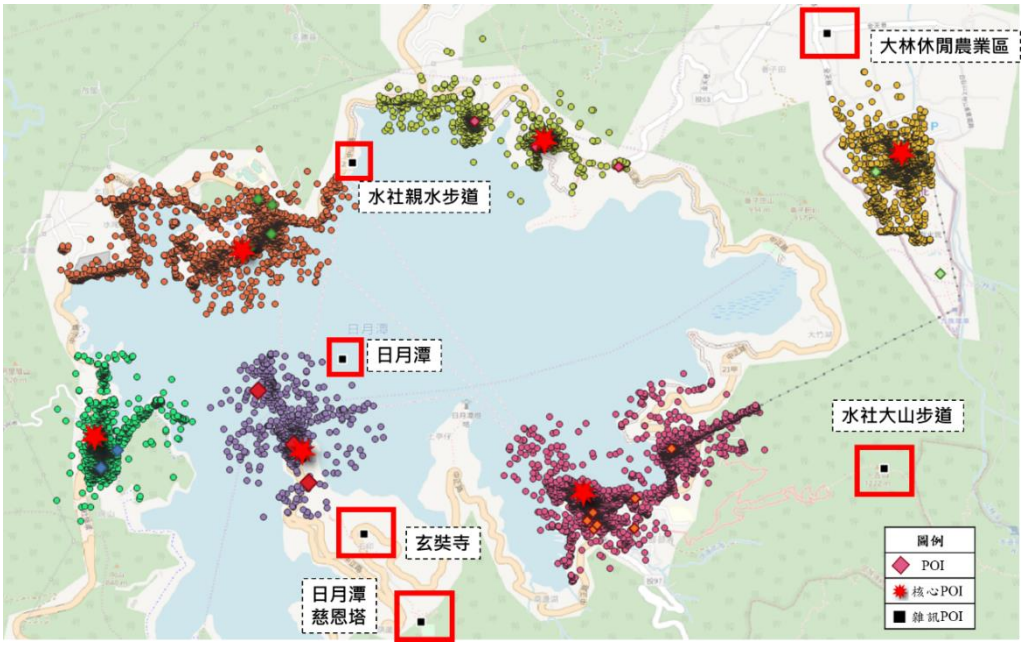

圖 24 AP-DBSCAN (0.3km,500)雜訊圖

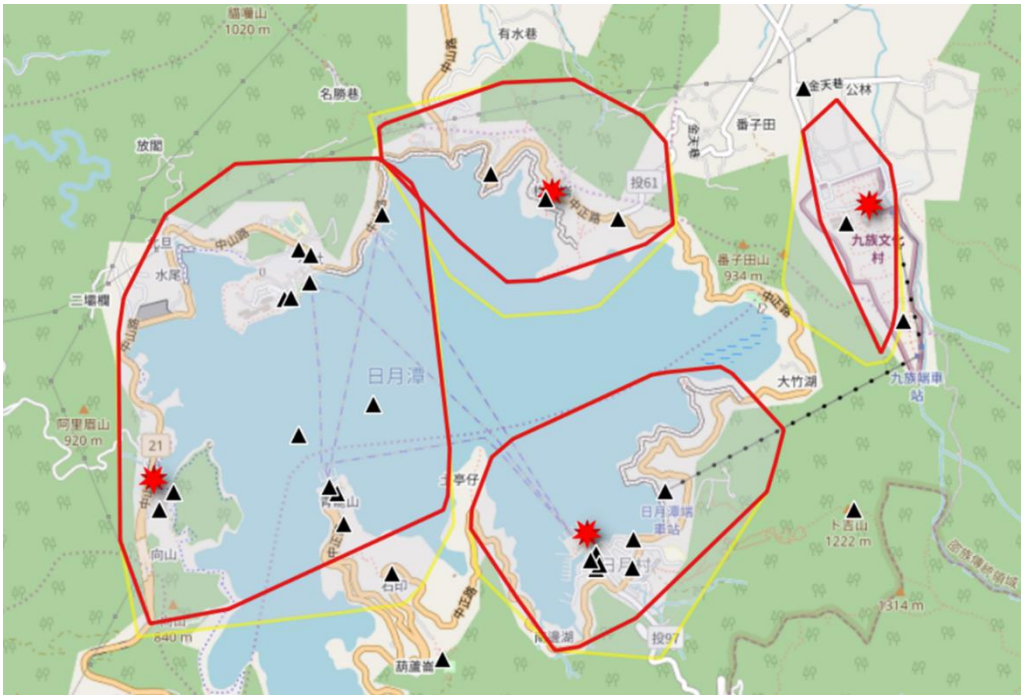

圖 25 DBSCAN VS AP-DBSCAN (0.5km, 1000)

- AP-DBSCAN 分群演算法與其他方法的比較

本節會比較本論文所提出方法與 DBSCAN 套件及 *K*-means 套件之比較，並且應用 QGIS convexholl 功能協助觀察與分群的差異與合理性。以下首先討論我們方法與 DBSCAN 套件比較之成果，其中兩個方法分別在(*Eps*, *Minpts*)參數分別為(0.5km, 1000)、(0.5km, 500)，以及 (0.3km, 500)的比較結果分別列於圖 25 到圖 27，

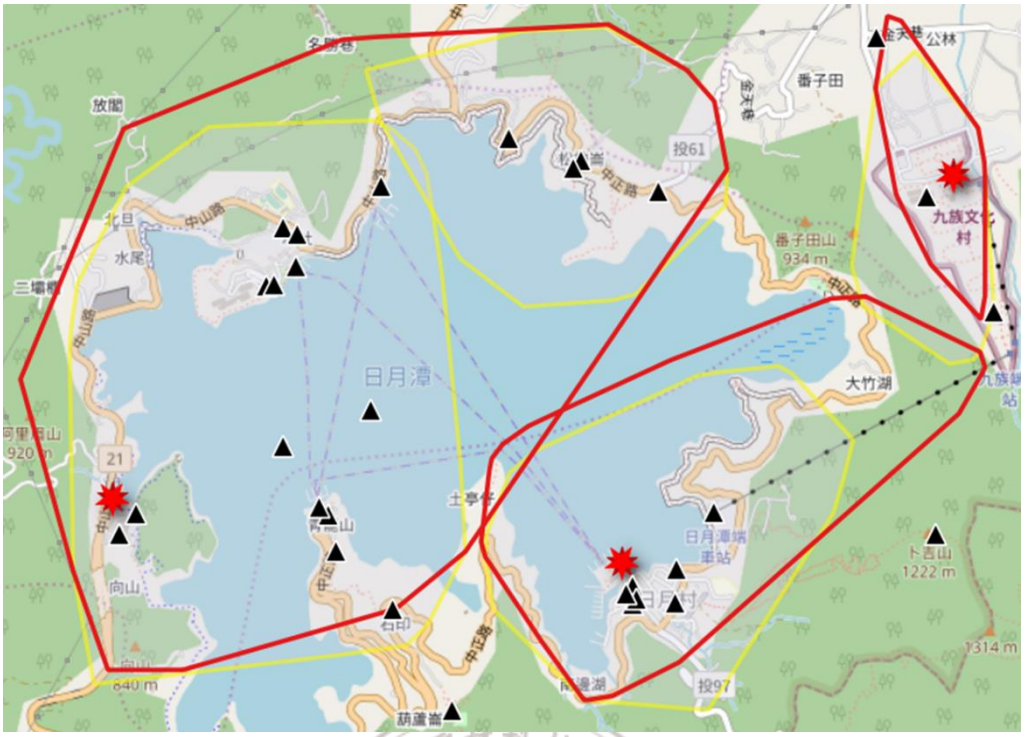

圖 26 DBSCAN VS AP-DBSCAN (0.5km, 500)

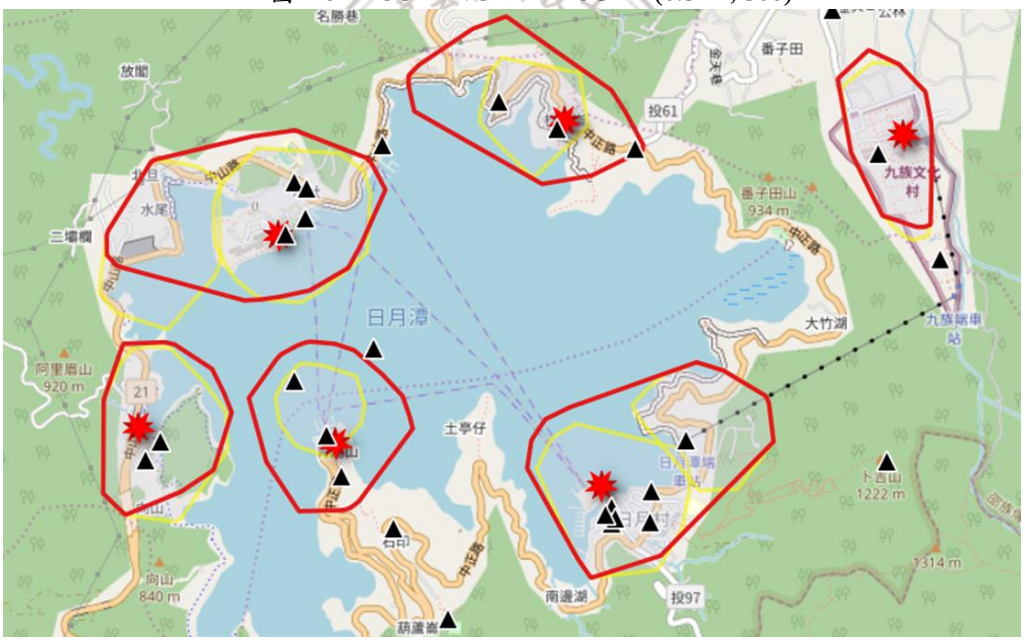

圖 27 DBSCAN VS AP-DBSCAN (0.3km, 500)

其中紅色圈圈代表本論文演算法的結果，紅色圈圈則代表 DBSCAN 得結果。於原理上 DBSCAN 套件為隨機讀取點位，除每次分群結果會不同外，也無法由套件得知選擇的核心點，而 AP-DBSCAN 為指定讀取點且非為套件因此使用者可記錄下每次分群的過程，於理論基礎上 AP-DBSCAN 擁有較好的分群穩定性。圖 25 左側的向山自行車步道旅遊區，AP-DBSCAN 將玄奘寺視為雜訊，而 DBSCAN 將此分

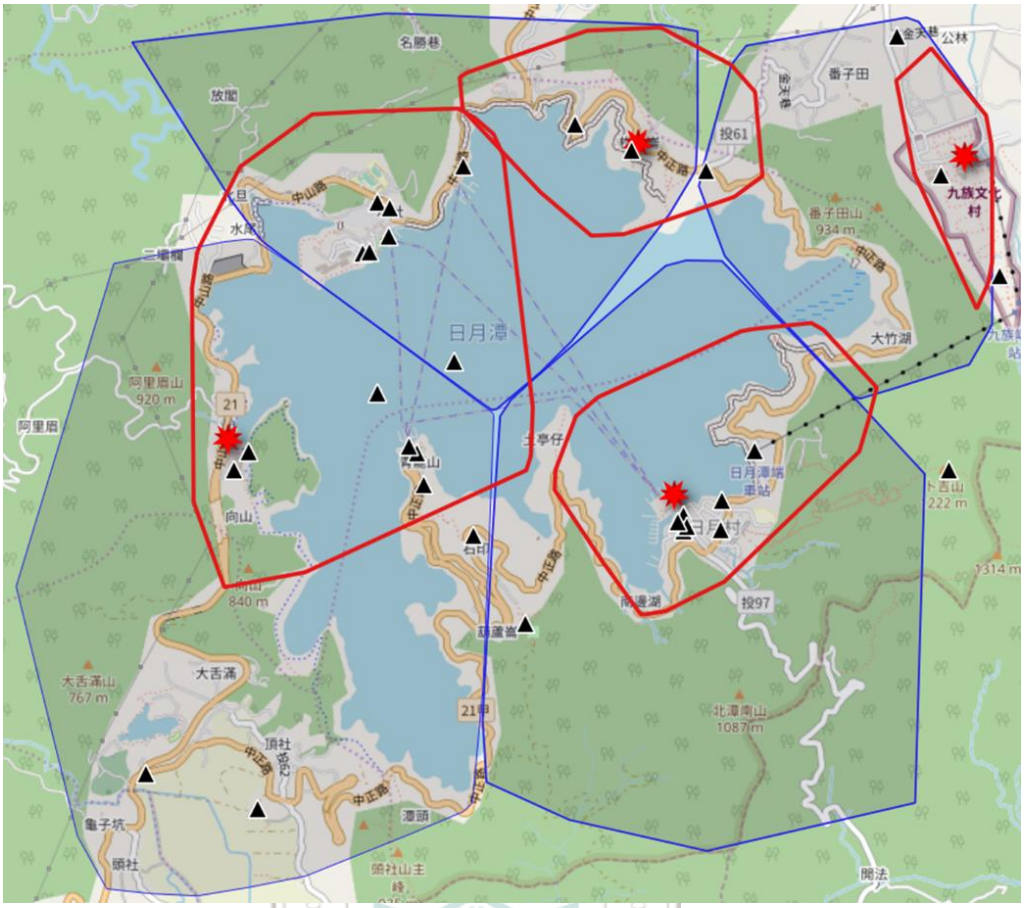

圖 28 AP-DBSCAN(0.5km, 1000)與 K-means(4 群)比較

為同一群，經實地訪查，於玄光寺旁有一個碼頭，大部分遊客皆為搭乘遊艇至該處，因此僅能靠雙腳於附近遊憩，不會移動到玄奘寺，由此可發現 AP-DBSCAN 相較於 DBSCAN 分群上，較符合向山自行車步道人群聚集的實際意義。另於圖 25 到圖 27 九族文化村的分群結果，DBSCAN 將九族文化村旁的中正路也納為同一群，於實際道路可達性及遊客活動方式而言，並不合理。

接著我們比較我們方法與 *K*-means 套件分群之成果，其比較圖如圖 28 到圖 30 所示，其使用參數分別是 (1)AP-DBSCAN(0.5km,1000) 與 *K*-means(4 群)、(2)AP-DBSCAN(0.5km,500) 與 *K*-means(3 群) 及 (3)AP-DBSCAN(0.3km,500) 與 *K*-means(6 群)。

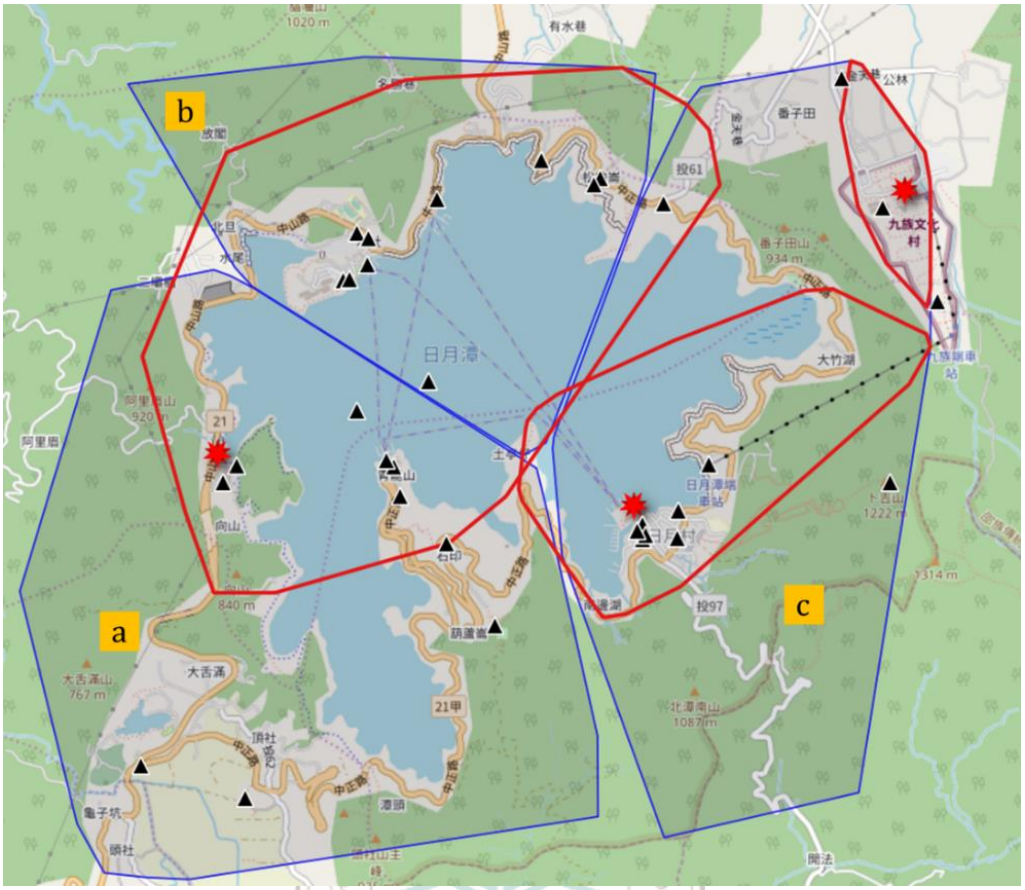

圖 29 AP-DBSCAN(0.5km, 500)與 K-means(3 群)比較

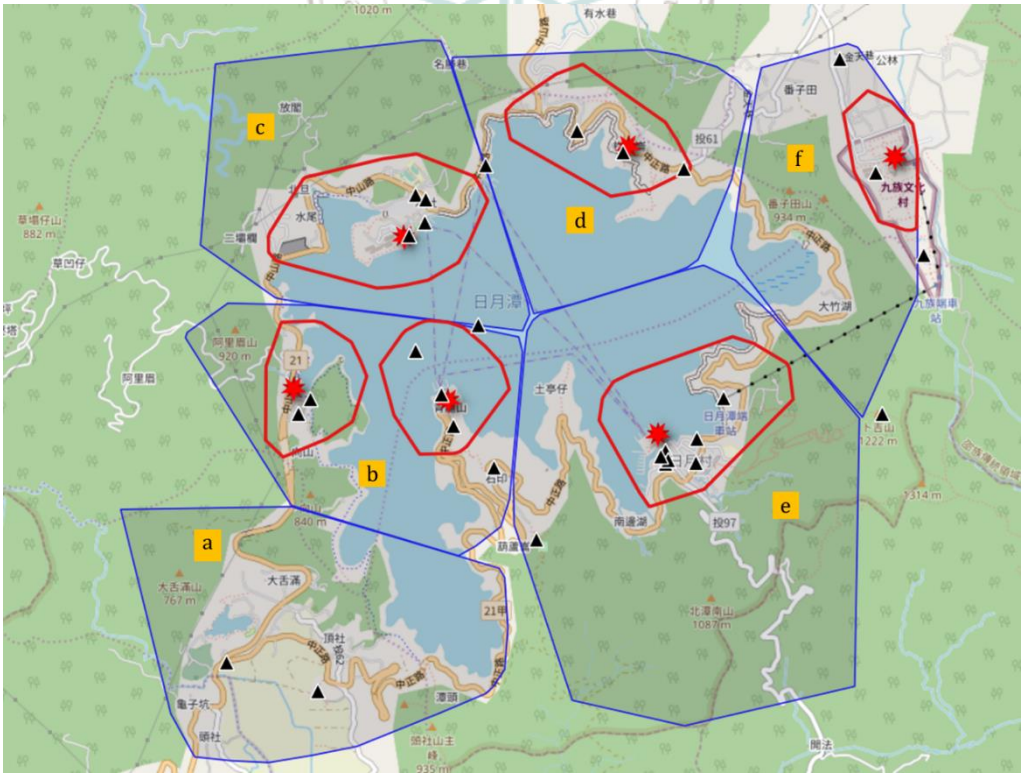

圖 30 AP-DBSCAN(0.3km, 500)與 K-means(6 群)

表 11 機器學習參數表

各模型結果	參數
隨機森林	max_depth：40 min_samples_split：0.1 min_weight_fraction_leaf：0.1
決策樹	max_depth：10 min_samples_split：0.5 min_weight_fraction_leaf：0.2

K-means 於分群上為隨機找尋起始位置，將每一個點位皆分至某一群體中，對於現實遊客活動中，容易將偏頗的點位也分成同一群體以及易將遊客活動上的密集區切割成不同群，較不具實際含意。反觀於 AP-DBSCAN 而言可於指定讀取某點位後根據密度可達性，去除雜訊同時考量密度可達性，形成較好的分群結果。

4.4 遊客人流數目預測模型之驗證

本小節內容以向山自行車步道旅遊區為目標進行驗證，其中該區是由 AP-DBSCAN 參數設定為(0.5km, 1000)時所辨識出的結果。而我們總共使用了 5 種模型來預估未來 1 日會出現在 Flickr 上的照片數量。

我們所使用的模型分為兩類，第一類為類神經網路結構模型，分別為本論文欲使用的 LSTM 模型[20][27]、知名的 GRU 模型[25][26]，以及傳統的 RNN 模型[17][18]。而為了對三個模型進行公平的比較，每個模型我們都會使用 42 個輸入，包含了雨量、溫度、假日/非假日，以及向山自行車步道旅遊區 35 個顯性與隱性景點的 Google trends 數據，所以每個模型都會有 38 個輸入。這些模型的隱藏層會有四層，而每一層所使用的神經元數量分別為(30,15,5,5)，其中在第二與第三層、第三層與第四層間我們會分別設定 dropout=0.2 與 0.1。整個模型則訓練 100 個 Epoch。第二類模型則為機器學習迴歸模型，分別為隨機森林[28][29]及決策樹[30][31]，其參數設定則如表 11 所示。為挑選出較好的模型，本論文評分績效採用預測模型常見的均方根誤差(Root Mean Squared Error, RMSE)及平均絕對誤差(Mean Absolute Error, MAE)進行效能比較，符號表示 M 為資料筆數、y_i為預測值、$y_{\hat{i}}$為真實值以及 i 為第幾筆資料。當指標值愈低時表示有好的績效。

表 12 初步 LSTM 預估模型結果

模型結果	排名	Train RMSE	排名	Test RMSE	排名	Train MAE	排名	Test MAE	總分 (愈低愈好)
隨機森林	4	13.61	3	12.91	4	6.04	4	6.13	**15**
決策樹	5	14.13	1	12.13	5	6.54	2	5.94	**13**
GRU	2	11.60	5	14.78	2	4.81	5	6.34	**14**
RNN	1	11.39	4	14.33	1	4.34	3	5.98	**9**
LSTM	3	11.82	2	12.33	3	4.89	1	5.68	**9**

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - y_i)^2} \tag{3}$$

$$MAE = \frac{1}{M}\left(\sum_{i=1}^{M}|y_i - y_i|\right) \tag{4}$$

本研究將實驗績效表 12 結果進行排序及加總，總排名次序為 LSTM=RNN>決策樹>GRU>隨機森林。說明類神經網路模型較適合用於此框架，而 RNN 與 LSTM 中又以 LSTM 在測試指標中最佳，因此 LSTM 在未來擁有較好的適應性，而這點也證實了本論文所選擇之 LSTM 模型確實能有效對旅遊區未來一日的遊客人流數目進行預測。

第五章 結論與未來研究

近年來台灣地區旅遊發展的議題逐漸受到政府與學者的重視，然而即便政府與學者投入了大量心力與資源仍然無法克服不同景點因為名氣差異太大而造成出遊人數不均的問題。本論文提出了兩大概念來克服上面的問題，包含了(1)藉由社群網路的文章找出台灣地區的隱性景點，結合政府已有的顯性景點，並藉由 Flickr 的照片標籤資料集來描繪出每個旅遊區的範圍。(2)收集各項資料來替我們所辨識出的旅遊區建立遊客人流模型，並協助政府與民眾預測該旅遊區未來每天的遊客人流數目目。藉由(1)的概念，我們將能整理出台灣地區目前所有的景點，將這些景點整合成多個旅遊區後推銷給民眾，所以民眾會以旅遊區為單位前去遊玩，並平衡同一區域內景點因為名氣不同而造成的來客量差距以及促進整體的消費。藉由(2)的概念，我們也可以協助民眾了解未來每個旅遊區域的遊客人流數，做出合理的出遊區域選擇。如此某些熱門旅遊區將不再擁擠並可維護景點品質，而原本要去這些熱門旅遊區卻又沒去的遊客會轉向其他人數較少的旅遊區，並平衡台灣不同旅遊區間的差異。

為了達成上述兩大概念，本論文提出了一連串的作法，包含了(1)利用政府觀光資料庫建立台灣地區的顯性景點資料庫、(2)從 PTT[Tai-Travel]版上的旅遊文章中找出台灣地區的隱性景點、(3)利用前述的顯性與隱性景點與 Flickr 的照片地理標籤描繪出台灣地區的旅遊區，以及(4)利用深度學習理論中的 LSTM 模型替每個旅遊區的遊客人流數目變化進行建模。而在這些作法中，因為第二項與第三項作法較為困難，所以本論文額外設計了兩個新演算法來達成。最終，本論文使用了台灣南投日月潭地區 7 年的資料來進行驗證，我們總共從 PTT[Tai-Travel]版上的旅遊文章找到 26 個顯性景點以及 9 個隱性景點。同時，我們也驗證了本論文所提出的 AP-DBSCAN 能比傳統空間分群演算法中的 K-means 演算法與 DBSCAN 演算法更能有效進行旅遊區的辨識。最終，我們也比較了現階段最常被人使用的五種模

型來對旅遊區的遊客人流數目進行建模與預測，並得到本論文所使用的 LSTM 模型是其中表現最優秀的結果。

在未來本研究可以分成四大方面精進，包含了(1)在資料集方面蒐集更具有代表性的社群網誌(Dcard)及富含使用者行為的相關軌跡資料(GPS 軌跡)來進行分析，使後續研究的成果更加完整與貼近現實情況。(2)在演算法方面使用空間資料結構演算法與台灣地區道路資料集來加速空間分群演算法的計算。(3)在應用端提供政府更多面的建議，包含旅遊區的設施建構的評估、是否要開拓新的道路以串連不同旅遊區等。以及(4)在遊客人流預測部分考慮更多的輸入，像是旅遊區的特殊活動或是新聞報導是否有提及旅遊區等來提昇預測的精準度。相信在完成這些工作後，本論文的內容將能有效協助政府與民眾了解台灣地區的觀光資源與關聯，進一步改善台灣地區景點旅遊人數分布不均的問題，並提昇台灣地區整體的觀光收益。

參考文獻

[1] J. K. S. Jacobsen, N. M. Iversen, and L. E. Hem, "Hotspot crowding and over-tourism: Antecedents of destination attractiveness," *Annals of Tourism Research*, vol. 76, pp. 53-66, 2019.

[2] K. Czernek-Marszałek, "Cooperation evaluation with the use of network analysis," *Annals of Tourism Research*, vol. 72, pp. 126-139, 2018.

[3] D. Buhalis and R. Law, "Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research," *Tourism Management*, vol. 29, no. 4, pp. 609-623, 2008.

[4] J. Navío-Marco, L. M. Ruiz-Gómez, and C. Sevilla-Sevilla, "Progress in information technology and tourism management: 30 years on and 20 years after the internet - Revisiting Buhalis & Law's landmark study about eTourism," *Tourism Management*, vol. 69, pp. 460-470, 2018.

[5] I. Lee, G. Cai, and K. Lee, "Exploration of geo-tagged photos through data mining approaches," *Expert Systems with Applications*, vol. 41, no. 2, pp. 397-405, 2014.

[6] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How flickr helps us make sense of the world: context and content in community-contributed media collections," *presented at the Proceedings of the 15th ACM international conference on Multimedia*, 2007.

[7] A. Majid, L. Chen, H. T. Mirza, I. Hussain, and G. Chen, "A system for mining interesting tourist locations and travel sequences from public geo-tagged photos," *Data & Knowledge Engineering*, vol. 95, pp. 66-86, 2015.

[8] Z. He, Z. Wu, B. Zhou, L. Xu, and W. Zhang, "Tourist Routs Recommendation Based on Latent Dirichlet Allocation Model," in *2015 12th Web Information System and Application Conference (WISA)*, 2015, pp. 201-206.

[9] K. D. Mukhina, S. V. Rakitin, and A. A. Visheratin, "Detection of tourists attraction points using Instagram profiles," *Procedia Computer Science*, vol. 108, pp. 2378-2382, 2017.

[10] G. Chareyron, J. Da-Rugna, and B. Branchet, "Mining tourist routes using flickr traces," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 2013, pp. 1488-1489.

[11] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, no. 14, pp. 281-297: Oakland, CA, USA.

[12] G. Liu, J. Yang, Y. Hao, and Y. Zhang, "Big data-informed energy efficiency

assessment of China industry sectors based on K-means clustering," *Journal of Cleaner Production*, vol. 183, pp. 304-314, 2018.

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd*, 1996, vol. 96, no. 34, pp. 226-231.

[14] Q. Zhang, H. Wang, J. Dong, G. Zhong, X. J. I. G. Sun, and R. S. Letters, "Prediction of sea surface temperature using long short-term memory," in *2017 IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1745-1749, 2017.

[15] S. Liu, S. Gu, and J. J. C. J. o. E. Peng, "Self-adaptive Processing and Forecasting Algorithm for Univariate Linear Time Series," *Chinese Journal of Electronics*, vol. 26, no. 6, pp. 1147-1153, 2017.

[16] Z. Tan, Y. Wang, Y. Zhang, and J. J. I. T. o. B. Zhou, "A novel time series approach for predicting the long-term popularity of online videos," in *2016 IEEE Transactions on Broadcasting*,vol. 62, no. 2, pp. 436-445, 2016.

[17] Y.-P. Chen, S.-N. Wu, and J.-S. Wang, "A hybrid predictor for time series prediction," in *2004 IEEE International Joint Conference on Neural Networks*, 2004, pp. 1597-1602: IEEE.

[18] J.-S. Wang and Y.-C. Chen, "A Hammerstein-Wiener recurrent neural network with universal approximation capability," in *2008 IEEE International Conference on Systems, Man and Cybernetics*, 2008, pp. 1832-1837: IEEE.

[19] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in *1997 IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1997.

[20] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *9th International Conference on Artificial Neural Networks: ICANN* ,p. 850 - 855,1999.

[21] Ö nder, I. "Forecasting tourism demand with Google trends: Accuracy comparison of countries versus cities," *International Journal of Tourism Research*, vol. 19, no. 6, pp. 648-660, 2017.

[22] F.-L. Chu, "Forecasting tourism demand with ARMA-based methods," *Tourism Management*, vol. 30, no. 5, pp. 740-751, 2009.

[23] V. Cho, "A comparison of three different approaches to tourist arrival forecasting," *Tourism Management*, vol. 24, no. 3, pp. 323-330, 2003.

[24] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.

[25] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values," *Scientific Reports*, vol. 8, no. 1, pp. 6085, 2018.

[26] X. Zhang, F. Shen, J. Zhao, and G. Yang, "Time Series Forecasting Using GRU Neural Network with Multi-lag After Decomposition," in *Neural Information Processing*, Cham, 2017, pp. 523-532: Springer International Publishing.

[27] S. Hochreiter and J. J. N. c. Schmidhuber, "Long short-term memory," *Neural computation* ,vol. 9, no. 8, pp. 1735-1780, 1997.

[28] O. Mutanga, E. Adam, and M. A. Cho, "High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm," *International Journal of Applied Earth Observation and Geoinformation*, vol. 18, pp. 399-406, 2012.

[29] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1862-1874, 2015.

[30] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761-1768, 2007.

[31] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15273-15285, 2011.