東海大學資訊管理研究所

碩士學位論文

基於遞迴神經網路之深度學習

建立 $PM_{2.5}$ 預測模型

A $PM_{2.5}$ Prediction Model

Based on Deep Learning with Recurrent Neural Network

指導教授：姜自強 博士

研 究 生：莊宜庭 撰

中華民國 108 年 7 月

東海大學資訊管理學系碩士學位

考試委員審定書

資訊管理學系研究所 莊宜庭 君所提之論文

基於遞迴神經網路之深度學習建立PM2.5預測模型

經本考試委員會審查，符合碩士資格標準。

學位考試委員會 召集人：葉銘民 (簽章)

委　員：楊朝棟

丁川康

姜自強

中 華 民 國 108 年 7 月 24 日

致謝

光陰飛縱，歲月流逝，在研究所的日子轉眼間就過去哩，這也代表我的人生即將邁向下一個里程，在攻讀研究所的過程中除了學習到很多知識外，也學習到做人處事的道理與態度。

首先要特別感謝指導教授姜自強老師，他從一開始論文題目的選定到最後口試等各階段，皆給予我許多的協助、指導與叮嚀，在經過無數次和老師每週報告進度的批改與指教，才使得本論文得以順利完成，在此致上最誠摯的謝意。另外感謝臺北大學資訊工程學系沈榮麟教授與嘉義大學應用數學系鄭富國教授在研討會上給予許多寶貴意見及看法，在此敬上最深的謝意。最後，在論文口試時，本校資工系楊朝棟教授和許瑞愷教授、成功大學工程科學系黃悅民教授、銘傳大學資訊管理學系丁明勇教授、關島大學李鳳霖教授給予許多寶貴的意見與指導，使本論文能更臻完善，在此深表感謝。

在攻讀研究所的這段期間，感謝學長佳鴻、復榮及學姊昕穎在課程上的指導及幫助。感謝和我一起從碩一就努力奮鬥到最後的同仁們：順利、玟蒨、思諭、廷睿、梓豪、瑄儒、晉瑋、耕希。感謝學弟奉原、煥捷在我有難時拉了我一把，感謝姜家學弟育棋、家君、有平在平常時給予我些許幫助。

最後，則要感謝我的家人在精神上給予我支持與關心，並提供良好的學習環境，使我能專心學習與研究，毫無後顧之憂，您們無怨無悔的付出是我成長的動力，在此將這份學位榮耀分享給您們。

論文名稱：基於遞迴神經網路之深度學習建立 $PM_{2.5}$ 預測模型

校所名稱：東海大學資訊管理學系研究所

畢業時間：民國 108 年 07 月

研究生：莊宜庭　　指導教授：姜自強

論文摘要：

近年來由於許多的研究發表都驗證空污會嚴重影響到人體健康，再加上媒體報導許多有關空污的議題，因此更讓民眾開始重視它的存在。本研究以 2018 年環保署的空氣品質即時污染指標的資料來做分析，採用五種補值的方法進行補值，藉由主成分分析和相關係數各別找出影響 $PM_{2.5}$ 濃度的主要相關變數（單因子：PM_{10}、SO_2、NO_X、NO_2、CO，雙因子：NO_X＋NO_2＋CO、SO_2＋PM_{10}），並利用遞迴神經網路（RNN）的長短期記憶模型（LSTM）來建立預測未來 8 小時的 $PM_{2.5}$ 濃度模型。根據研究結果顯示，豐原測站的預測值與真實值之誤差大部分都有落在合理的 MAPE（0.2~0.5）範圍內。另外在補值法方面是以線性插值法最好。

關鍵詞：空氣污染、主成分分析、相關係數、深度學習、遞迴神經網路、長短期記憶模型

Title of Thesis：A PM2.5 Prediction Model

Based on Deep Learning with Recurrent Neural Network

Name of Institute：Tunghai University, Graduate Institute of Information Management

Graduation Time：07/2019

Student Name：Yi-Ting Chuang Advisor Name：Tzu-Chiang Chiang

Abstract：

In recent years, many studies have verified that air pollution will seriously affect human health. In addition, the media reported many issues concerning air pollution, so people have begun to pay attention to its existence. This study analyzes the data of the Environmental Protection Administration air quality immediate pollution indicators in 2018. Five methods are used to deal with the missing values. The main correlation variables affecting the PM25 concentration are identified by principal component analysis and correlation coefficients (single factor: PM10, SO2, NOX, NO2, CO, two-factor: NOX+NO2+CO, SO2+PM10), and the Long-Short Term Memory Model (LSTM) of the Recurrent Neural Network (RNN) was used to model the PM25 concentration model for the next 8 hours. According to the research results, most of the errors between the predicted and true values of Fengyuan Station fall within the reasonable range of MAPE (0.2~0.5). In addition, the best way to deal with the missing value is linear interpolation.

Keywords：Air pollution , Principal Components Analysis, Correlation Analysis, Deep learning , RNN, LSTM.

目錄

圖目錄

表目錄

第一章　緒論

1.1　研究背景、動機

近幾年來，臺灣人民對健康意識逐漸提升，同時也開始注重生活環境品質。在早期的時候，臺灣空污的情況雖然嚴重，但很少有這方面的報導，民眾都把因污染所造成的灰濛濛天氣，當成普通的雲霧[1]。直到 2015 年左右，民眾才逐漸意識到這個問題與嚴重性[2]。

臺灣的空氣污染，主要可分為境內產生（如工廠、發電廠）和境外傳輸（如中國大陸）。其中，在 $PM_{2.5}$ 污染部分有 66%是由國內所產生[3]。臺灣的地形也是惡化臺灣空氣污染的重要因素之一。例如北臺灣的臺北市和新北市被一座接一座的山團團圍繞；而臺灣西部沿海的城市（如臺中市、臺南市、高雄市）和工業區的東邊則是有高聳入雲、重疊連綿的中央山脈等，每當在冬天時就會阻擋東北季風來吹散中南部空氣中的懸浮微粒[4]。

行政院環境保護署於 2016 年推出的空氣品質新指標（AQI），它是依據各空氣污染物的濃度而分為六個等級，簡單的說就是依據二氧化硫（SO_2）、一氧化碳（CO）、臭氧（O_3）、懸浮微粒（PM_{10}）、細懸浮微粒（$PM_{2.5}$）、二氧化氮（NO_2）對人體健康影響之濃度大小來分等級的，並且搭配 6 種顏色的方式呈現（表 1-1），提供民眾一看就懂的指標和顏色，作為當天是否外出活動的參考依據，例如當 AQI 指標數值落在 101~150（橘色）區時，表示敏感性族群若要外出時建議盡量減少參與體力消耗型的活動或戶外活動，同時也要注意身體情況；當 AQI 指標數值落在 151~200（紅色）區時，表示所有族群（包含一般健康民眾、敏感性族群）在外出時更要注意戶外活動及身體情況[5]。

表 1-1 空氣品質新指標（AQI）

對健康影響	良好	普通	對敏感族群不健康	對所有族群不健康	非常不健康	危害
空氣品質指標(AQI)	0~50	51~100	101~150	151~200	201~300	301~500

（來源：環保署空氣品質監測網[6]）

經過行政院環境保護署的統計,從2016年12月開始實施空氣品質新指標(AQI)後到2017年2月20日這段期間（80天)，光是達到不健康的「不良橘」部分，北臺灣地區就有12天，中臺灣地區則是有37天，南臺灣地區竟然有76天[7]。

另外，國內外針對空氣污染是否影響民眾健康的相關研究（表1-2）得知，空氣污染的確是會影響民眾的健康，其中以$PM_{2.5}$的影響最為明顯。

表 1-2 相關研究及結果之文獻整理

發佈時間	研究領域及結果
2018/10	世界衛生組織（WHO）於在首屆全球空氣污染和健康大會中發表「亞太地區空氣污染狀況」的報告中指出，全球每年約有700萬人是死於空氣污染所引發的疾病，也就是說10個人中有9個人呼吸到的都是髒空氣[8]。因此需要擴大應對措施，以預防疾病和死亡發生。
2018/8	美國德州大學奧斯汀分校（University of Texas at Austin）在《環境科學與技術快報期刊》發布了最新的研究中指出，由空氣污染所引發的細懸浮微粒$PM_{2.5}$問題，使得所有人類的預期壽命平均減短1年，亞洲地區的民眾深受其害[9]。
2017/6	衛生福利部公布2017年國人十大死因中，排名前三名的依舊是癌症、心臟疾病、肺炎。其中在死亡之首的「癌症」裡，排名前三名依序為肺癌、肝癌、直腸癌。十大死因死亡人數中，相較於2016年，癌症增加277人、肺炎增加268人較為明顯[10]。可以看出肺癌已位居國人死亡率最高的癌症，而造成肺癌的主因之一的空氣污染已成為國人不可輕忽的死亡威脅因素。
2017/6	根據臺灣大學公衛學院與衛生福利部國民健康署合作研究的「臺灣地區歸因於可介入危險因子之主要疾病死亡負擔」中，排名為「前三大」的危險因子依序為高血糖、抽菸、高血壓。但最令人驚訝的是第四名竟是「$PM_{2.5}$暴露」[11]。

（來源：網路新聞）

由於學者專家們的研究和媒體們的報導，使得「空污」一直是備受關注的議題之一，因為空氣污染對健康有很多潛伏的影響，包括身體的細微生理變化，到明顯的疾病症狀，例如：咳嗽、氣喘、胸痛、胸悶、鼻子過敏等。當哮喘或慢性呼吸系統疾病患者接觸到空氣污染時，病情便會加重。雖然不同的人受到空氣污染影響的程度是取決於不同的因素，但是不同年齡的人都會受到惡劣的空氣品質影響，而空氣污染對兒童和老人的影響則更大[12]。

另外，就目前國內對 $PM_{2.5}$ 濃度預測方法的相關研究（表 1-3）來看，預測的方法有很多種，卻很少會有人用空污資料來做經過各種補值後所做的預測比較。

表 1-3 相關研究及結果之文獻整理

學者	研究領域及結果	預測方法
李政霖（2019）	利用 LoRa 無線通訊技術來開發一套低功耗的即時空氣品質動態監控系統並結合 LSTM 模型來預測未來 $PM_{2.5}$ 濃度。[13]	LSTM RFNN BP [比較]
顧芷瑄（2018）	使用自迴歸隱馬可夫模型（AR-HMM）與其他類型（監督式學習的支援向量機（SVM）、非監督式學習的隱馬可夫模型（HMM））相比，發現它除了在觀察值中彼此有相依的關係存在，且在預測 $PM_{2.5}$ 濃度上有較好的預測表現。[14]	AR-HMM SVM HMM [比較]
林冠名（2018）	在大數據平臺下開發三種機器學習方法為決策樹回歸（DTR）、梯度增強樹回歸（GBTR）和支持向量回歸（SVR），透過結合 SVR 的 ARIMA 補值方法，建立一個 8 小時的空氣品質預測模型。[15]	DTR GBTR SVR [比較]
林俞均（2018）	提出一個模糊類神經網路的空氣品質預測系統，透過歷史資料進行訓練，輸入的資料型態必須為時間序列，這樣才能依照時間的變化來預測空氣品質與環境因子的未來走向。[16]	神經模糊系統模型

表 1-4 相關研究及結果之文獻整理（續）

學者	研究領域及結果	預測方法
黃彥齊（2018）	在預測方面，以西屯測站為預測目標，先篩選出預測變數，再以 LSTM 及 BPN 分別進行預測。[17]	LSTM BPN[比較]
盧俊源（2017）	使用自我迴歸整合移動平均、類神經網路、支援向量回歸及兩種混合模式（ARIMA-ANN 及 ARIMA-SVR）來預測臺灣六都 $PM_{2.5}$ 的濃度。[18]	如左邊敘述 [比較]

（來源：臺灣博碩士論文知識加值系統）

1.2 研究目的

本研究以環保署在臺中市地區設置的空氣品質監測站為評估對象，利用各監測站所監測的空氣污染濃度（SO_2、CO、O_3、PM_{10}、$PM_{2.5}$、NO_2、NO_X、NO、WindSpeed、WindDirec）之即時資料，藉由統計分析來找出特徵值，再利用遞迴神經網路建立預測 $PM_{2.5}$ 濃度模型。因此本研究的目的可分為以下幾項：

1. 比較五種補遺法
2. 利用統計分析找出特徵值

   將各監測站污染物濃度進行主成分分析法與相關係數找出影響 $PM_{2.5}$ 的主要相關變數，也就是特徵值。
3. 利用遞迴神經網路建立預測 $PM_{2.5}$ 濃度模型

   使用 R 語言的 keras 和 TensorFlow 套件來建立預測 $PM_{2.5}$ 濃度模型，先投入特徵值做訓練，待訓練完後再投入特徵值之未來數值來預測未來 8 小時 $PM_{2.5}$ 濃度值，最後藉由平均絕對誤差百分比（MAPE）來評估預測的準確度。

1.3 論文大綱

本論文之研究，主要是找出影響 $PM_{2.5}$ 的相關變數及建立預測 $PM_{2.5}$ 模型來測試哪種組合較好。因此，針對各章節主要內容作簡要的描述如下：

I. 第二章的文獻探討，說明空氣污染物種類、簡單介紹機器學習、深度學習和遞迴神經網路（RNN），並整理相關的研究論文，作為本研究欲探討問題之依據。

II. 第三章的研究方法，是針對本研究的資料來源、資料預處理和採用何種方法來做分析...等的詳細之說明、解釋。

III. 第四章的研究結果與分析，詳細描述、解釋各分析方法之輸出結果，並給予簡單扼要的表格或視覺化圖形呈現。

IV. 第五章的結論與未來研究方向，是根據研究分析結果，彙整後並提出結論及未來可以改善的方向，供後續相關研究之學者作為參考依據。

第二章　文獻探討

本章內容分為二大部分，首先將近年來會影響 $PM_{2.5}$ 濃度的空氣污染物做簡單介紹，接著說明機器學習和深度學習，最後說明選用遞迴神經網路之原因和優缺點。

2.1　影響 $PM_{2.5}$ 之相關變數探討

2.1.1　空氣污染物介紹

1.　PM_{10} 及 $PM_{2.5}$（懸浮微粒與細懸浮微粒）

懸浮微粒（particulate matter, PM），是指漂浮在空氣中類似灰塵的微小粒子。PM 以粒徑來區分大小，其中 PM_{10} 及 $PM_{2.5}$ 分別指的是 10 微米與 2.5 微米以下的粒子[19]。PM 的計量單位以微克/立方公尺（$\mu g/m^3$）表示之。它的主要來源包括道路揚塵、汽機車排放廢氣、露天燃燒雜物、營建施工等。當我們在呼吸的同時，PM 會隨著人體的呼吸系統進入體內，由於 $PM_{2.5}$ 比 PM_{10} 更容易穿透肺部氣泡，並直接進入血管中隨著血液循環到全身，而導致諸多的疾病發生，如咳嗽、誘發哮喘、慢性支氣管炎、慢性肺炎，甚至是心血管疾病和肺癌[19]。若 PM 又附著其他污染物，如各種化學物質或重金屬，甚至細菌、黴菌或病毒，將更加深呼吸系統之危害[20]。

2.　SO_2（二氧化硫）

硫氧化物是硫的氧化合物總稱，其主要成分有二氧化硫（SO_2）和三氧化硫（SO_3）等[21]，其中 SO_2 是空氣主要污染物之一，它是呈酸性、無色的氣體，聞起來非常臭且刺鼻。由於大部分的煤和石油都含有硫化合物，因此在燃燒時會產生 SO_2。當 SO_2 溶於水中，會形成亞硫酸，也就是酸雨的主要成分，它同時還是製造硫酸的主要原料[22]。硫氧化物對人體的傷害主要是刺激人的呼吸系統。當我們在呼吸的同時，它會隨著人體的呼吸系統進入體內，首先會刺激到上呼吸道粘膜表層的第十對腦神經末稍，引起支氣管反射性地收縮和痙攣，導致咳嗽和呼吸道阻力增加。

當呼吸道的抵抗力減弱時，便會誘發慢性呼吸道疾病，甚至引起肺水腫和肺心性疾病 [21]。

3. NO_X（氮氧化物）

氮氧化物主要包括一氧化氮（NO）及二氧化氮（NO_2），其產生原因是來自於燃燒過程中，空氣中的氮或燃料中的氮化物氧化而成[23]，像是汽機車排放、發電廠、煉鋼廠…等。在 NO 部分，它為無色無味的氣體，稍溶於水，在燃燒過程中所產生的氮氧化物以 NO 為主要成份；與光化學反應後會變成 NO_2 [23]。在 NO_2 部分，它為赤褐色的氣體且具有刺激性的味道，易溶於水，與水反應後會變成亞硝酸和硝酸、與光化學反應並吸收陽光後會分解成 NO 和氧、在空氣中可氧化成硝酸鹽，也是造成雨水酸化的原因之一[23]。此外，NO_X 具有神經性毒性，其中就屬 NO_2 的傷害特別明顯。NO_2 會破壞人體的中樞神經，長期吸入的話會引發腦性麻痺，手腳萎縮等傷害，而大量吸入的話會導致記憶力喪失、四肢癱瘓、甚至是死亡。NO 也會與血紅蛋白結合而引起高鐵血紅蛋白血症[24]。

4. O_3（臭氧）

是一種由氮氧化物、碳氫化合物及日光照射後產生的二次污染物，具有強氧化特性。若在高濃度的臭氧環境下，會對人體的呼吸系統產生刺激性，初期的症狀為眼部刺激、嘴巴乾燥、咳嗽、肺部阻塞、呼吸急促，免疫力較差的人甚至會出現胸痛。特別是針對小孩、老人、病人或戶外運動者有較大影響，同時對於植物，包括農作物有不良影響，對於人造材料，如橡膠（輪胎等）及油漆等，都能造成傷害[25]。

5. CO（一氧化碳）

它是一種無色無味的氣體，尤其是每到冬天時就會發生的一氧化碳中毒案件就是明顯的例子。在大氣當中，CO 是少量存在的氣體，但因人為方面導致 CO 的增加，舉例來說，市區內的汽機車輛在剛發動時的燃燒所產生，因此市區內的 CO 濃度總是跟交通流量、地點及時間有密切關係。此外，像是一般家庭的熱水器或瓦斯爐的燃燒以及電廠、石油廠等都會產生 CO。最常見的一氧化碳中毒症狀，如

頭痛、噁心、嘔吐、頭暈和虛弱等感覺。常期暴露在一氧化碳的環境中可能會嚴重損害到心臟和中樞神經系統。

2.1.2 空氣污染物之相關研究

本節將針對國內目前對於 $PM_{2.5}$ 與其他的空氣污染物（如 PM_{10}、NO、NO_2、O_3…等）之間是否互相影響的研究文獻加以整理（表 2-1）。

表 2-1 相關研究及結果之文獻整理

學者	研究領域及結果	分析方法
李建業（2018）	整理並分析 2010~2014 年間環保署的中部空品區的空氣品質監測資料，想了解中部空品區 $PM_{2.5}$ 與其他指標污染物之特性。研究結果顯示，從相關性分析的角度來看，在中部地區的空氣品質裡頭 $PM_{2.5}$ 與 PM_{10}、NO、NO_2、O_3、CO、SO_2 呈正相關，而與溫度、降雨、相對濕度、風速呈負相關。從因子分析的角度來看，得知影響中部地區的空氣品質有四個因子，分別是 $CO+NO_X$、PM_{10}、O_3、SO_2 [26]。	相關係數 因子分析
陳柏丞（2017）	整理並分析 2013~2015 年間臺中市地區的五個測站（大里、沙鹿、忠明、西屯與豐原）之各種污染物濃度變化，並探討 $PM_{2.5}$ 與其他空氣污染物間的關係。研究結果顯示，發現 $PM_{2.5}$ 除了跟 PM_{10} 關係最密切外，跟 CO、SO_2 及 NO_X 也有相當大的關係，跟 O_3 及 NMHC 則是稍微有關係 [27]。	相關係數
陳正暐（2014）	整理並分析 2009~2012 年間中部空品區四縣市的 $PM_{2.5}$ 主要排放來源，並探討 $PM_{2.5}$ 與各項空氣污染物（NMHC、CO、SO_2、NO_X、O_3）之間關係。研究結果顯示，在相關性分析的部分，CO 及 O_3+NO_X 相關性高於 SO_2、NO_X 及 NMHC [28]。	相關係數 線性回歸 單因子變異數

表 2-2 相關研究及結果之文獻整理（續）

學者	研究領域及結果	分析方法
邱瑞仙（2008）	整理並分析 2004~2005 年六個自動空氣品質監測站的 PM_{10}、O_3、SO_2、CO、NO_X 各污染物資料，藉由統計分析的方式，解析各監測站測值的時空分布特性。以 Pearson 相關係數矩陣分析桃園縣空氣品質測站與污染物濃度的相關性，在研究結果顯示，所有測站的 PM_{10} 濃度和 O_3 濃度受到相同污染型態或擴散因子的影響。五權和中壢測站在 SO_2 相關性高，主要是測站污染物排放源或排放時間型態相似。CO 各測站間相關性高，但中壢交通測站與其他各站相關性較低，可能是中壢測站受到車輛排放直接影響大，其他測站則是受到經大氣環境混合的影響。NO_2 及 NO_X 測站間相關性高的因素為空間傳輸分布上相似 [29]。	相關係數 集群分析 主成分分析
吳奎縉（2007）	整理並分析 2003 年中部空品區四縣市其 $PM_{2.5}$ 主要的污染來源及排放量。研究結果顯示，在相關係數分析部分，NMHC、SO_2、NO_2、CO、O_3 及 O_3+NO_2 這 6 種空氣污染物與 $PM_{2.5}$ 具有相關性，在經回歸分析後又以 SO_2、NO_2、CO 及 O_3+NO_2 較具顯著性，由於 SO_2、CO、NO_2 為機動車輛排放廢氣中較為明顯的污染物，且 O_3+NO_2 主要發生在工業區、都會區和交通要道，另屬於交通量較大之測站如忠明、大里、南投測站，其 NMHC 也具有相關，由此可見機動車輛為重要污染源之一 [30]。	相關係數 線性回歸

（來源：臺灣博碩士論文知識加值系統）

從表 2-1 到表 2-2 發現，$PM_{2.5}$ 的確與其他空氣污染物之間是互相影響，就分析方法來講，大部分的研究都採用相關係數來看空氣污染物之間的關係。因此本研究將採用相關係數和主成分分析來作為找出特徵值的方法。

2.2 機器學習與深度學習

目前我們所知道的人工智慧（Artificial Intelligence，AI）已從類神經網路進展到「機器學習」，可以運用在過濾垃圾郵件、分析人類的行為並投放相關廣告、無人車自動駕駛…等領域[31]。近來年，由於「深度學習」技術的突破，再加上由 Google DeepMind 開發的人工智慧圍棋程式 AlphaGo 打敗韓國棋士並得到圍棋冠軍，使得深度學習成為 AI 學門中應用最廣與最快速的領域。

2.2.1 機器學習

機器學習（Machine Learning），簡單來說就是要讓機器（電腦）具有學習能力以及從資料中自動學習規則，並利用規則對新的資料進行預測；其主要目的是設計和分析出可以自動學習的演算法，讓機器（電腦）可以從過去的資料或經驗當中建立一個模型（Model），而學習（Learning）就是執行此模型，並利用訓練資料集（Training Dataset）來建立模型[32]。

常見的機器學習可以分成監督式學習、非監督式學習、半監督式學習，詳細說明如下：

1. 監督式學習（Supervised Learning）：

所有的資料都要有標準答案（也就是「標註」（Label）），可以提供機器學習在輸出時判斷誤差時使用，這樣預測出來的值就會比較精準，這就像在學校考模擬考一樣，學生在考後可以拿到模擬考的解答與自己的答案互作比對看誤差，這樣在考正式考試時的成績會比較好[31]。

2. 非監督式學習（Unsupervised Learning）：

所有的資料都沒有標準答案，無法提供機器學習在輸出時判斷誤差時使用，機器得必須自己尋找答案，這樣預測出來的值就會比較不準，就好比當學生考完模擬考後卻沒有拿到解答，這樣無法與自己的答案互作比對看誤差，導致後來在考正式考試時的成績就會比較差[31]。

3. 半監督式學習（Semi-supervised learning）：

少部分的資料有標準答案，可以提供機器學習在輸出時判斷誤差時使用；大部分的資料沒有標準答案，機器得必須自己尋找答案，等於是將監督式與非監督式學習的優點做結合[31]。

2.2.2 深度學習

就一般來說，深度學習（Deep learning）是指具有層面性的機器學習演算法，它能透過一層又一層的處理將大量混亂又無規則的訊號逐漸轉為有用的資訊並解決問題。但通常人們所提到的深度學習，則是指一種特定的機器學習演算法—「深度神經網路」（Deep Neural Network）[33]。

在談深度神經網路之前，得先知道神經網路的基本原理。神經網路包含著許多神經元（neuron），而神經元的功能是負責接受資料或傳遞資料。而基本的神經網路包含三層神經元，除了輸入層和輸出層以外，中間還有一層的隱藏層，負責傳遞並處理資料。其中在隱藏層的部分，它可以有一層以上，當有數個以上的隱藏層的神經網路就會被稱為深度神經網路。在圖 2-2 的右邊是一個深度神經網路的示意圖，也許它只有兩個隱藏層看起並沒有很深，但在實際上神經網路可以高達數十層至數百層，非常具有「深度」[33]。

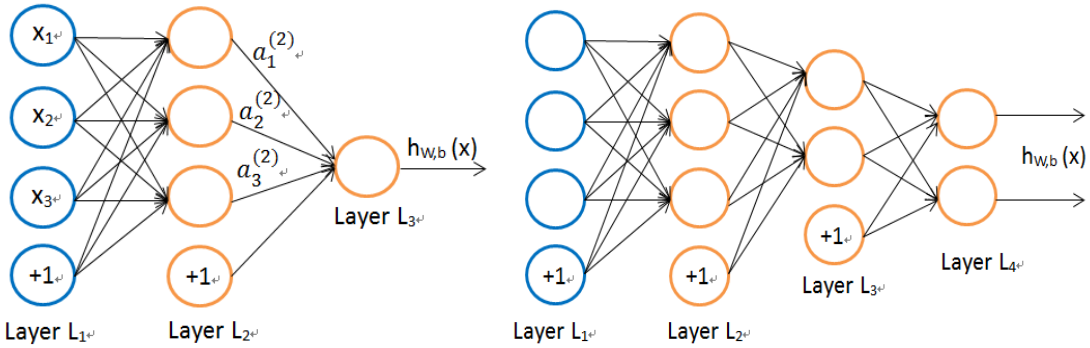


圖 2-1 神經網路（左）和深度神經網路（右）的示意圖

（來源：CASE PRESS/臺大科學教育發展中心[33]）

深度神經網路在不同層之間的神經元是具有不同權重值的連結，由輸入層加乘連結的權重逐層向後傳遞資料。每一個神經元會將輸入資料乘以權重值後進行

線性加總，若加總超過給定的閥值（threshold）或偏移值（bias），則會依非線性激勵函數（activation function）產生輸出。神經元之間連結的權重值及神經元的閥值是可以調整的，因為權重值（含閥值或偏移值）的調整就是深度學習最核心的過程[34]。

本研究以監督式（supervised）深度學習的方法來調整權重值，其說明如下：監督式深度學習的方法可分為訓練（training）和預測（prediction）兩個階段。在訓練階段中，會不斷輸入具有標記（label）的訓練資料（training data），其中標記是對應輸入資料的真實輸出結果，它通常是依據真實情況的實際觀察而加上的。在最初狀態下，權重值是由隨機方式所給定的，因此在輸入訓練資料後會輸出與標記具有誤差的預測結果。一般我們以代價函數（cost function）或是損失函數（loss function）來表示誤差大小。例如，公式（2-1）就是以均方差（mean squared error, MSE）函數為代價函數來表示輸入一批訓練資料後誤差的大小[34]。

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

在公式（2-1）中，n 為輸入資料筆數，$\hat{y}_i$為第 i 筆輸入資料之標記（label），y_i為第 i 筆輸入資料之預測結果[34]。

若能夠調整權重值到最小化之代價函數，就表示能夠讓預測值與標記的誤差值降至最小。因為代價函數相對於權重值的梯度（gradient）為代價函數成長最快速的方向，為了讓代價函數最小化，可以使用梯度下降（gradient descent）法來計算權重值的修正量，使代價函數沿著梯度反方向來修正，如公式（2-2）所示：

$$\Delta W = -\eta \cdot \frac{\partial E}{\partial W} \qquad (2\text{-}2)$$

其中，E 為誤差代價函數，W 為權重值，$\frac{\partial E}{\partial W}$ 為代價函數對權重值偏微分（Partial Derivative）（也就是梯度），ΔW為權重值修正量，η 為學習率（learning rate），可控制權重值的修正量之幅度大小，但這會影響代價函數的最小化過程之收斂速度。另外利用微積分的連鎖律（chain rule），將輸出層的誤差代價函數值以反向方式傳遞到各隱藏層以修正各層的權重值，就被稱為誤差倒傳遞（error backpropagation）[35] [34]。

梯度下降法是可以達到最小化代價函數的優化器（optimizer）之一，但有一個容易陷入局部最佳解（local optimum）的問題，而無法達成全域最佳解使代價函數最小化。也因為這樣，後來有不少改良版的優化器被學者所提出，可以提高達成全域最佳解的機會並加快其達成速度，如 Momentum、AdaGrad、RMSprop、AdaDelta 等[36]。這些改良的優化器就包含加入動量或自動調節學習率的概念，前者在參數更新時，會加上前一次更新動量的考量；後者則是當梯度大，就會減少學習率，反之，就會增加學習率[34]。

2.2.3 長短期記憶遞迴神經網路

在深度學習中，遞迴神經網路（recurrent neural network, RNN）[37]是較適合處理與時間序列相關性的資料。與一般神經網路不同的地方在於，RNN 強調資料間存在著時間相依性，因此在網路中加入迴圈（loop），可讓當前的輸出做為下一次的輸入，所以在處理資料時可以考慮到之前的輸出結果。這就像是網路有了「記憶」一樣，會依照過去的記憶決定接下來的輸出結果。如圖 2-3 所示，將一個 RNN 展開後，可看作是時間序列的資料由左往右輸入，其中 x_0、x_1、x_2、…、x_t 為每個時間序列的輸入資料，h_0、h_1、h_2、…、h_t 為每個時間序列的輸出，t 為輸入的時間[34]。

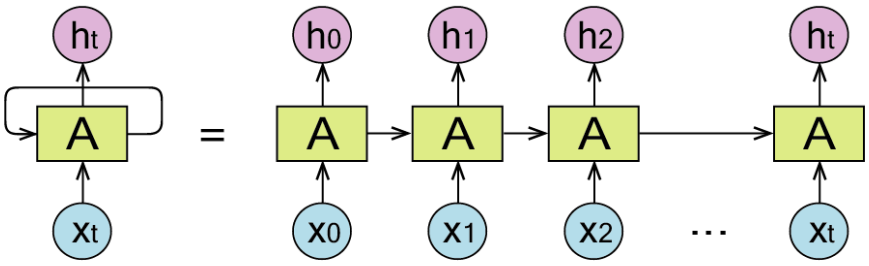


圖 2-2 遞迴神經網路 RNN 的展開（來源：colah's blog [37]）

在理論上 RNN 架構可以處理任何長度與時間相關性的資料，然而實際上在訓練時所用的倒傳遞演算法，卻會讓此架構產生梯度爆炸（gradient exploding）或是梯度消失（gradient vanishing）的問題，隨著時間的增加，權重的更新大到變成無窮大、非數值的情況，或權重的更新小到變成趨近於 0 的情況，都會讓神經網路

很難反應時間中相隔較遠資料的相依性[34]。

而長短期記憶（long short-term memory, LSTM）[38]是目前 RNN（Recurrent Neural Network）最常使用的模型，於 1997 年由 Hochreiter 與 Schmidhuber 提出。當傳統的 RNN 在進行梯度下降法時，修正量會隨著時間的間隔變長而衰減，而導致權重值無法適度更新，所以無法將神經網路訓練到最好。然而 LSTM 架構卻可以解決這個問題，因為 LSTM 能夠掌握長時間中相隔久遠資料的相依性。LSTM 架構內有三閘（gate）一單元（cell），分別是：輸入閘（input gate）、輸出閘（output gate）、遺忘閘（forget gate）、和記憶單元（memory cell）。LSTM 是以「閘」的概念來控制訊息是否加入到記憶單元或從記憶單元中移除，每個閘有各自的輸入權重值，可以經過訓練調整後而儲存最適當的歷史訊息於儲存單元中[34]。

在標準的 RNN 架構中，重複使用的神經元只是一個非常簡單的結構，例如輸入當前的輸入和前一次的輸出，並依照個別權重值加總後透過 tanh 函數輸出，如圖 2-4 所示[37]。而 LSTM 的結構就比較複雜，每個 LSTM 單元使用四個神經元，以特殊的方式交互作用，其基本架構如圖 2-5 到圖 2-7 所示[37]。LSTM 中最關鍵的是記憶單元（memory cell），它的功能類似一個暫存區，可以暫存先前輸入資料所產生的狀態，之後神經元可以根據先前的狀態而去計算出不同的輸出值。「遺忘閘」是決定哪些訊息應該從記憶單元中丟棄與丟棄比重；「輸入閘」是決定記憶單元內要保存哪些新訊息與保存比重，可分成兩種：更新舊狀態和新增臨時狀態；最後由「輸出閘」更新記憶單元的狀態並同時進行輸出。因為 LSTM 可以有效儲存歷史資訊，因此非常適合用於處理和預測時間序列中間隔和延遲非常久的重要事件[34]。

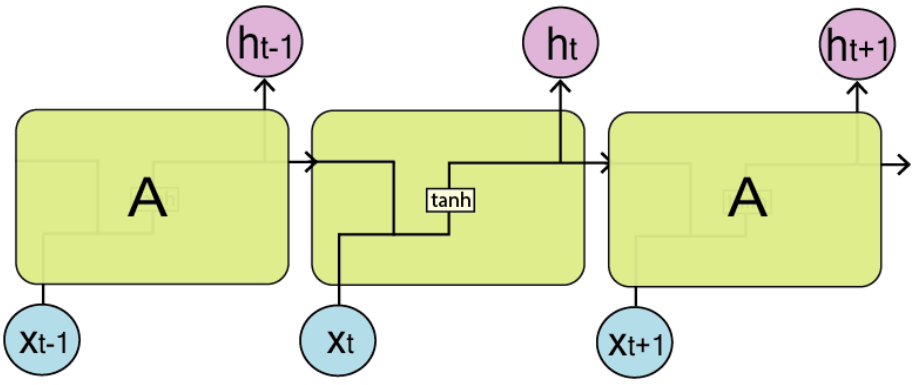


圖 2-3 標準遞迴神經網路（RNN）之基本架構（來源：colah's blog [37]）

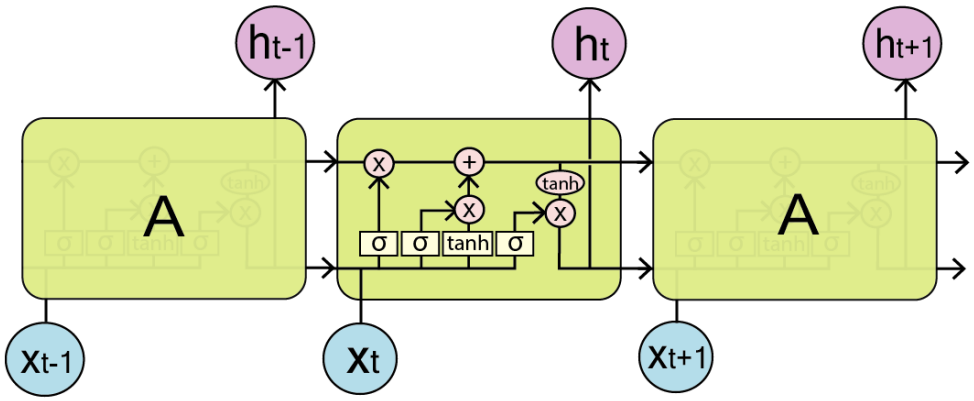
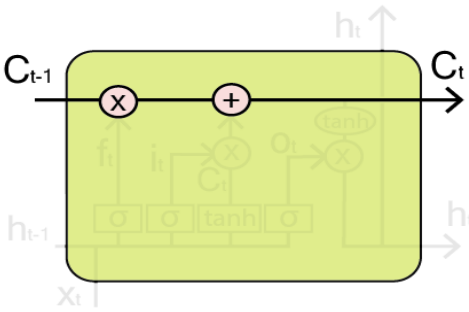
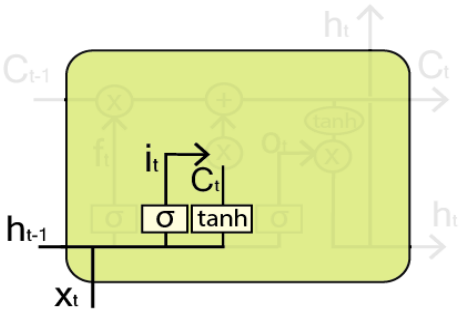


圖 2-4 長短期記憶（LSTM）之基本架構（來源：colah's blog [37]）

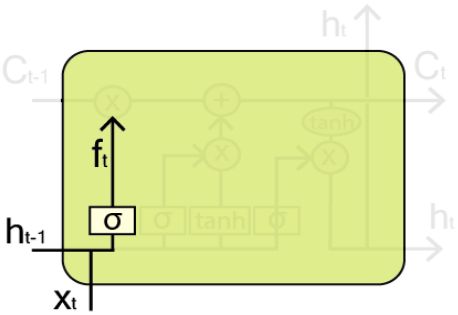
Memory Cell



Input Gate



Forget Gate



Output Gate

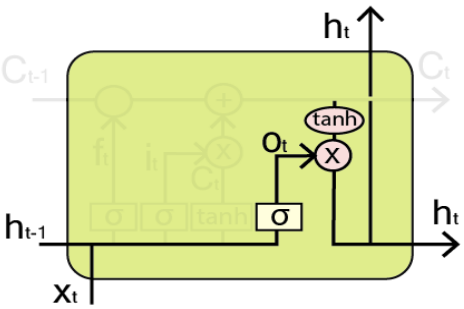


圖 2-5 長短期記憶（LSTM）之基本架構（續）（來源：colah's blog [37]）

2.2.4 深度學習之相關研究

本節將針對國內目前使用遞迴神經網路（RNN）的長短期記憶模型（LSTM）所作的預測 $PM_{2.5}$ 之相關研究文獻加以整理（表 2-3）。

表 2-3 相關研究及結果之文獻整理

學者	研究領域及結果
李政霖（2019）	本研究以 LoRa 無線通訊技術開發一套低功耗的即時空氣品質動態監控系統並結合 LSTM 模型來預測未來 $PM_{2.5}$ 濃度。
	預測方法：開發+結合 LSTM
Zhao, Jiachen Deng, Fang Cai, Yeyun Chen（2018）	本研究提出一種長期短期記憶-完全連接（LSTM-FC）神經網路，使用歷史空氣品質預測 48 小時內特定空氣品質監測站的 $PM_{2.5}$ 污染數據、氣象數據、天氣預報數據和星期幾。本研究的預測模型由兩部分組成：(1) 使用基於長期短期記憶（LSTM）的時間模擬器來模擬 $PM_{2.5}$ 污染的局部變化；(2) 使用基於神經網絡的空間組合來捕獲空間中心站 $PM_{2.5}$ 污染與鄰近站點之間的相關性。我們在 2014/05/01 至 2015/04/30 期間對包含北京 36 個空氣質量監測站記錄的數據集評估我們的模型，並與人工神經網絡（ANN）和長期短期記憶（LSTM）模型進行比較。結果顯示，LSTM-FC 神經網路模型提供了更好的預測性能[39]。
	預測方法：開發 LSTM-FC 並與 ANN 比較
蔡宜廷（2018）	本研究以 ETL（Extract-Transform-Load）的框架整理了 2013~2017 年環保署與氣象局提供的歷史數據，透過 AKIMA 進行補值且將資料分類成三種不同污染源的資料集，利用 Aggregation Model 的方式，分別使用這三種資料集建立 LSTM 子神經網路，來得到三種不同預測特徵資料，透過融合層結合這三種不同的預測特徵，並輸出到全連接層中透過反向傳播分別給予其隱藏層不同的權重，最後得到未來 8 小時預測的 $PM_{2.5}$ 數值。在與現有的方法 ANN 與 LSTM 比較後，第一個小時的準確率在 RMSE 上比 LSTM 減少 0.15 誤差、在 MAE 誤差減少 0.11，與 ANN 比較後在 RMSE 誤差減少 0.75、在 MAE 誤差減少 0.54[40]。
	預測方法：ANN 與 LSTM 比較、預測未來 8 小時的 $PM_{2.5}$ 數值

表 2-4 相關研究及結果之文獻整理（續）

學者	研究領域及結果
Tsai, Yi-Ting Zeng, Yu-Ren Chang, Yue-Shan（2018）	本研究提出了一種使用 LSTM（長短期記憶）的 RNN（遞迴神經網路）預測 $PM_{2.5}$ 濃度的方法。利用 Keras，這是一個用 Python 編寫的高級神經網路 API，能夠運行在 Tensorflow 之上，構建一個神經網絡，並通過 Tensorflow 運行帶有 LSTM 的 RNN。在訓練數據部分是從 2012 年至 2016 年從台灣環保局（EPA）獲取，並合併為 20 維數據；預測測試數據則是 2017 年。我們進行了實驗，以評估台灣 66 個站點未來 4 小時 $PM_{2.5}$ 濃度的預測值。結果顯示，該方法可以有效預測 $PM_{2.5}$ 的值[41]。
	預測方法：RNN、LSTM
李靜芳（2018）	本研究以 RNN 進行空氣污染 $PM_{2.5}$ 濃度的分析和自動化預測。在實驗中建立一個 RHadoop 的分布式計算環境來分析空氣污染。除了將即時資料透過 MySQL 資料庫進行存取，也利用了 Sqoop 對 HBase 進行歷史數據快速存取。此外，我們也針對資料缺值補值進行討論及實驗，求出不影響預測精度或能提升預測精度的補值方法。在實驗中使用 MAPE 將 $PM_{2.5}$ 的短期預測精度進行量化，將 MAPE 控制在 0.2 至 0.5 區間。在不影響精度下，針對 RNN 各個參數進行實驗和校調，進一步開發 RNN 自動化訓練程式[42]。
	預測方法：RNN、資料缺值補值進行討論及實驗。
Bui, Tien-Cuong Le, Van-Duc Cha, Sang-Kyun（2018）	本研究採用具有長短期記憶的遞迴神經網路（RNN）作為框架，利用大邱、首爾、北京和沈陽的空氣污染和氣象資料的時間序列數據。此外，我們使用編碼器-解碼器模型，它類似於機器理解問題，作為我們預測機器的關鍵部分。最後，我們研究了各種配置的預測精度。我們的實驗在預測遠期時間步長時會阻礙在預測模型上集成多層 RNN 的效率[43]。
	預測方法：RNN、LSTM

表 2-5 相關研究及結果之文獻整理（續）

學者	研究領域及結果
盧慧鴻（2018）	本研究以物聯網技術建構以空氣品質感測器連接手機應用程式的應用，主要分為兩個步驟，第一個步驟為建立預測模型，本研究使用 LSTM 方法建立預測模型，為了建立適合的預測模型，設計了六組實驗，且每組實驗有五種不同的時間序列，探討環保署監測站資料與微型監測站資料與時間序列的改變在預測模型上的表現。第二個步驟為開發系統，以 Android 為作業系統開發手機應用程式[44]。
	預測方法：LSTM
Athira, V Geetha, P Vinayakumar, R Soman, KP （2018）	本研究的空氣汙染物以 PM_{10} 為主，以空氣污染和氣象時間序列 AirNet 數據來分析，使用遞迴神經網路（RNN）、長期短期記憶（LSTM）和門控遞迴單位（GRU）進行預測。為了找出最佳架構，研究了不同 RNN 模型及其變化的拓撲和模型參數的廣泛分析。通過改變[0.01,0.5]範圍內的學習率，每個實驗執行多達 1000 次。從該研究中觀察到，這三個模型在預測中表現相對較好[45]。
	預測方法：RNN、LSTM、GRU　比較
鍾玉峰 張文鎰 蔡惠峰 蘇威智 （2018）	本研究以深度學習技術為基礎，探討 $PM_{2.5}$ 濃度預測方法。蒐集環保署於全臺各地設置的空氣品質監測站紀錄資料，利用深度學習技術，以深度遞迴神經網路（RNN）訓練學習各項監測資料與 $PM_{2.5}$ 濃度在空間與時間的關係性，預測未來數十小時之間的 $PM_{2.5}$ 濃度，可作為空氣污染擴散預防及民眾戶外活動參考的簡要指標。本論文亦探討於輸入資料完整度與 RNN 模型結構不同的情況下，訓練出的模型預測表現的差異，也探討在輸入資料加入未來時間風場模擬預測數據，加強關聯性，以提升 $PM_{2.5}$ 濃度預測結果精確度[46]。
	預測方法：利用 RNN 來預測未來數十小時之間的 $PM_{2.5}$ 濃度。

表 2-6 相關研究及結果之文獻整理（續）

學者	研究領域及結果
Li, Xiang Peng, Ling Yao, Xiaojing Cui, Shaolong Hu, Yuan You, Chengzeng Chi, Tianhe （2017）	本研究提出了一種新的長期短時記憶神經網絡擴展（LSTME）模型，該模型本質上考慮了空間污染物濃度預測的時空相關性。長期短期記憶（LSTM）層用於從歷史空氣污染物數據中自動提取特徵值，並將輔助數據（包括氣象數據和時間戳數據）合併到所提出的模型中以提高性能。取 2014/1/1 至 2016/5/28 在北京市的 12 個空氣品質監測站的每小時 $PM_{2.5}$ 濃度數據用於驗證其有效性。建立的 LSTME 模型與時空深度學習（STDL）模型、時間延遲神經網絡（TDNN）模型、自回歸移動平均（ARMA）模型、支持向量回歸（SVR）模型和傳統 LSTM NN 模型進行比較，結果顯示，LSTME 模型優於其他模型[47]。
	預測方法：LSTME、STDL、TDNN、ARMA、SVR 和傳統 LSTM NN 做比較
Kim, Min Han Kim, Yong Su Lim, JungJin Kim, Jeong Tai Sung, Su Whan Yoo, ChangKyoo （2016）	開發並比較多元線性迴歸（MLR）、神經網路（NN）和遞迴神經網路（RNN）的數據驅動預測方法，用於地鐵站的室內空氣品質。RNN 模型可以通過將昨天污染物的先前時間資料添加到模型來預測地鐵站平台處的空氣污染物濃度。為了優化預測模型，使用偏最小二乘法（PLS）的投影（VIP）中的變量重要性來選擇關鍵輸入變量作為預處理步驟。將預測模型應用於來自遠程監測系統數據（TMS）的真實室內空氣品質數據集，展示了一些非線性動態行為，結果顯示所選擇的關鍵變量對模型的預測性能具有強烈影響。它顯示 RNN 模型具有對非線性和動態系統進行建模的能力，RNN 模型的預測結果比其他的預測模型具有更好的建模性能和更高的可解釋性[48]。
	預測方法：開發並比較 MLR、NN、RNN

（來源：臺灣博碩士論文知識加值系統＋GOOGLE 學術）

從表 2-3 到表 2-6 發現，就預測方法來講，大部分都採用 RNN 的 LSTM，且程式語言幾乎是 Python 為主，很少人用 R 語言來寫，因此本研究在撰寫程式部分採用 R 語言，在預測方法部分採用遞迴神經網路（RNN）的長短期記憶模型（LSTM）。

第三章　研究方法

3.1　研究工具及軟體介紹

「R 語言」，是一種自由軟體程式語言的操作環境，主要功能是統計分析、繪圖和資料探勘[49]。其中，統計分析的主要目的是查看資料分佈的狀態，而資料探勘則較偏向分析，例如關聯（Association）、分類（Classification）、預測（Prediction）等模型，皆為 R 的領域。另外在深度學習方面，可以利用 R 語言透過 Anaconda 來使用 keras 或 TensorFlow 套件，學習預測訓練或開發全新的深度學習模型。接著說明本研究的實驗環境及軟體工具：

1. 在電腦（1 台）方面其配備是：

   處理器 = AMD Ryzen 5 2600 Six-Core Processor　3.40 GHz

   記憶體(RAM) = 16.0 GB、核心數目 = 6、邏輯處理器 = 12

2. 在軟體方面是使用：

   Anaconda3 2019.03(Python 3.7.3 64-bit)

   Python 3.7.3 (32-bit)

   R for Windows 3.6.1

   Rstudio 1.2.1335

3.2　研究設計及研究流程

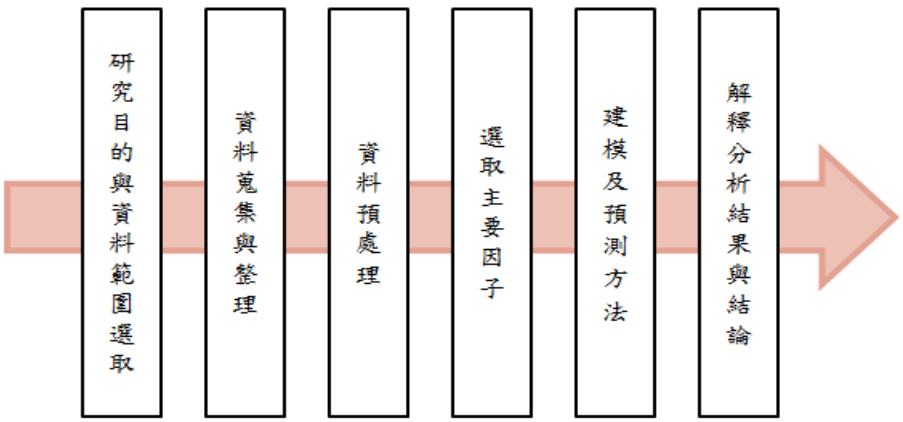


圖 3-1 研究設計與流程圖（來源：本研究）

3.3 資料來源

行政院環境保護署於 1980 年設立空氣品質監測站網，1987 年設置空氣品質監測站來監測 $PM_{2.5}$ 及其他相關因子濃度，之後便陸續進行測站的校調及增設[2]，其監測結果於環保署全球資訊網上提供即時的空氣品質資料，以及提供歷時、歷年的監測資料[19]。截至今天為止，在臺灣地區共設立約 78 個測站，依測站類型可分為一般（共 60 站）、工業（共 5 站：頭份、線西、麥寮、臺西、前鎮）、交通（共 6 站：鳳山、三重、中壢、永和、復興、大同）、國家公園（共 2 站：恆春、陽明）和背景（共 5 站：萬里、觀音、三義、橋頭、富貴角）。以資料量來計算，1 個測站在一年的資料量是 365 天 ×24 小時 ＝ 8,760 筆，相對的 78 個測站 × 8760 筆 ＝ 683,280 筆。因為資料量大，於是本研究參考中興大學環工系莊秉潔教授所公布的 2018 上半年全臺縣市的空污排名[50]，來取得排名位居中間的縣市，也就是臺中市（一般測站的資料量：5 個測站 × 365 天 × 24 小時 ＝ 43,800 筆）。

本研究的資料來源為環保署於 2016 年成立的環境資源資料庫（圖 3-2），以環保署在臺中市地區設立的五個測站（大里、西屯、忠明、沙鹿、豐原）為目標，研究範圍以 2018 年 1 月 1 日至 12 月 31 日的空氣品質即時污染指標的資料為主（圖 3-3）。

圖 3-2 環保署之環境資源資料庫（來源：環保署的環境資源資料庫）

SiteName	PublishTime	SO2	CO	O3	PM10	PM25	NO2	Nox	NO	WS	WD
忠明	2018/1/1 00:00	2.5	0.44	28	48	29	14	15	1	2.4	38
豐原	2018/1/1 00:00	0.9	0.48	31	30	10	4.9	6	1	2.1	25
大里	2018/1/1 00:00	1.5	0.62	23	40	19	13	15	2	1.5	50
西屯	2018/1/1 00:00	2	0.35	40	56	28	7.8	8	1	2.8	9.1
沙鹿	2018/1/1 00:00	1.8	0.31	37	51	23	9.1	9	0	4.4	7

圖 3-3 空氣品質即時污染指標的資料內容（來源：本研究）

3.4 資料預處理

在做資料分析前，可以先將資料做預處理（preprocessing）的動作，也就是資料清理、資料集成、資料轉換和資料歸約的動作，光是這些動作就佔了 70% 的工作量[51]。剛提到的資料處理技術都得在資料探勘前使用，因為這樣在做資料探勘分析時便可以大幅提升模式的優劣程度，降低實際探勘時所需的時間[52]。

在進行資料分析時，最常遇到的問題就是離群值（Outlier）和遺失值（Missing Value）的處理。尤其是遺失值的部分，當重要的特徵變數有遺失值時，是無法輕易忽略的。例如，在建置回歸模型時，資料內若有任一個特徵值是 NA（Not Available）時，整列資料就將被忽略不使用，這樣肯定會使我們失去很多資訊[53]。

3.4.1 離散值

在挖掘資料的過程中，「資料內是否存在離群值」這點是很重要的，因為它可能會嚴重影響到資料分析的結果，甚至會影響到建立模型的準確度，因此判斷離群值的方法便相當重要。以下將介紹判斷離群值的方法－箱型圖。

箱型圖（Box plot），又稱為盒鬚圖，於 1977 年美國著名統計學家約翰・圖基（John Tukey）所發明[54]，它的功能是用來顯示數據分布情況的統計圖，組成要素如下：

1. 最大值（max）（箱型上虛線的端點）
2. 最小值（min）（箱型下虛線的端點）
3. 中位數（median）（涵蓋前 50%資料之位置）
4. 第一四分位數（涵蓋 25%之資料）（Q1）
5. 第三四分位數（涵蓋 75%之資料）（Q3）

其中，Q1 與 Q3 的差值稱為四分位距（Interquartile range, IQR）

繪製箱型圖時，需決定籬笆（fence），籬笆為第一四分位數的 −1.5×IQR 與第三四分位數的 +1.5×IQR [55]。

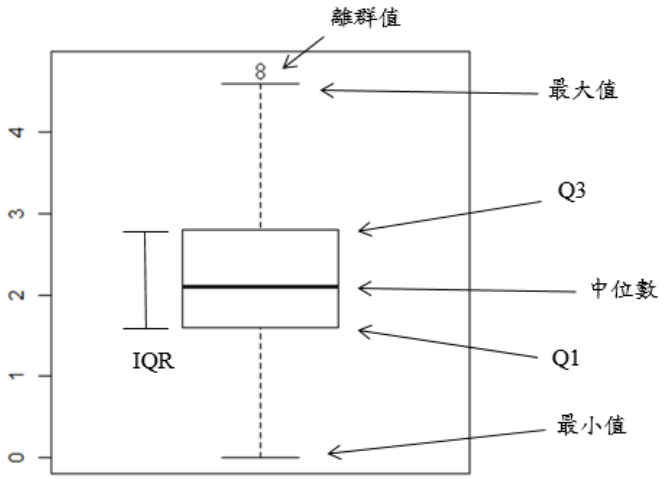


圖 3-4 箱型圖（來源：本研究）

繪製箱型圖（圖 3-4）前，會先找出最大值、最小值、中位數、第一四分位數以及第三四分位數。其中，在計算第一四分位數時，會將 n 個觀測值由小到大進行排序，計算 i=（25/100）×n，若 i 為整數，第一四分位數為第 i 大及第 i+1 大的觀測值之平均；若 i 不為整數，則取下一個大於 i 之整數為第一四分位數之觀測值位置。箱型圖以第一及第三四分位數劃出盒子，再沿著盒子左右劃出肖線（Whiskers），肖線兩端為籬笆內的資料最大值及最小值。若有觀測值落在箱型圖的籬笆外，則會將之視為離群值。在本研究裡，藉由箱型圖來找出離群值並用 NA 取代。

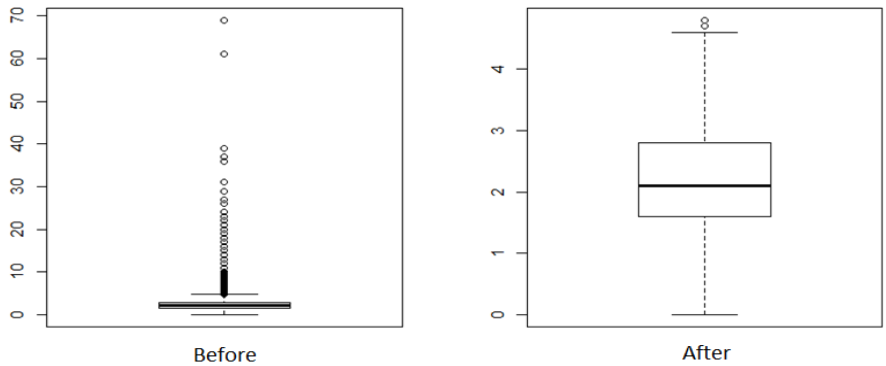


圖 3-5 SO2：使用前（左）、使用後（右）（來源：本研究）

3.4.2 遺失值

一般來說，會先判斷遺失的資料在整體資料中佔的比例是多少，若只占極小比例的話，是可以直接刪除資料列。但如果特徵變數中有超過5%的遺失比例時，就得進行遺失值處理。本研究將採用五種填補遺失值的方法，分別如下：

1. Hmisc 套件內的 approxExtrap()函數。
2. Hmisc 套件內的 aregImpute()函數。
3. MICE 套件內的 mice()函數。
4. MissForest 套件內的 missForest()函數。
5. rpart 套件內的 rpart()函數。

3.4.3 預測精準度

在衡量預測精準度（Forcast Accuracy）的方法有很多種，如 MAPE、RSME、MSE、MAE，而本研究是以 MAPE 來作為比對預測精準度。

MAPE 代表平均絕對誤差百分比（Mean Absolute Percentage Error, MAPE），當 MAPE 的值越小時，表示預測值與實際值之間的誤差越小，結果也就越好。其計算方式如公式（3-1）所示，其中$\hat{y}_t$為預測值，y_t為實際值，n 為預測值數目[56]。

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \qquad (3\text{-}1)$$

其評估標準如表 3-1 所示：

表 3-1 MAPE 預測準確度之評估標準

MAPE	< 0.1	0.1-0.2	0.2-0.5	> 0.5
預測能力	高精準度	良好	合理	不正確

（來源：水土保持學報 第 37 卷 第 02 期[56]）

本研究將根據五種補遺失值的方法，測試補值後得到的精準度（表 3-2）。

表 3-2 測試補值的精準度

補值方法 / 空污變數	補值一	補值二	補值三	補值四	補值五
SO_2	0.174732	0.486200	0.408471	0.545962	0.452754
CO	0.131512	0.280142	0.201281	0.296069	0.204462
O_3	0.278137	0.888141	0.630194	0.893389	0.575736
PM_{10}	0.189726	0.484627	0.381544	0.548388	0.429024
PM_{25}	0.313639	0.690342	0.514275	0.711591	0.494610
NO_2	0.173754	0.115238	0.090504	0.147465	0.100753
NO_X	0.166486	0.095142	0.077485	0.124018	0.089866
NO	0.281833	0.574349	0.315454	0.362334	0.220668
WS	0.285439	0.499426	0.409254	0.567364	0.400127
Mean	0.221695	0.457067	0.336496	0.466287	0.329778
Ranking	1	4	3	5	2

註：補值一：approxExtrap、補值二：aregImpute、補值三：mice
補值四：missForest、補值五：rpart

（來源：本研究）

根據表 3-2 發現，補值補的最好的是線性差值（也就是補值一），因為它的補值較接近真實值外，還有它的計算方式：只利用兩點的對應值來推算兩點之間的對應值，雖然兩點對應值本身往往受到各種偶然因素的影響，所以線性插值的結果可能誤差較大。

3.5 影響 $PM_{2.5}$ 之相關變數的選取

本研究將透過主成分分析和相關係數來找出影響 $PM_{2.5}$ 之主要變數。

3.5.1 主成分分析（Principal components analysis, PCA）

1. 主成分分析之定義

主成分分析法是一種將變數較多的資料，加以縮減來產生一些主要成份的較少變數方法，而在產生的過程中，須達成幾項重要目標：（1）代表性：盡可能保有原變數的資訊。（2）獨立性：新的變數之間必須要相互獨立。（3）精簡性：變數量已適當縮減[57]。

例如：在比較企業間的競爭力時，其中一項就是財務績效的程度，然而在衡量財務績效程度的變數可能多達數十種（如資產報酬率、每股營收、…等），此時單就財務績效程度而言，我們可以將所有變數透過主成分分析，縮減成少數具代表性的主成分，也可以進一步將每間公司在這些主成分的分數，透過一些加權的方式，計算出每間公司財務績效程度的分數，然後再加以排序比較[57]。

2. 主成分分析之計算步驟

例如：某份原始資料中有 x_1、x_2、…、x_p 等 P 個變數，針對這些變數進行主成分分析，則計算步驟如下[58]：

步驟 1：將原始變數進行標準化動作

先計算變數 x_1、x_2、…、x_p 所對應之平均值 $\bar{x}_1$、$\bar{x}_2$、…、$\bar{x}_p$ 和標準差 s_1、s_2、…、s_p，並藉由平均值和標準差將變數 x_1、x_2、…、x_p 進行標準化，且將標準化後所得之數值依序設為 u_1、u_2、…、u_p。其公式（3-2）所示[58]：

$$u_1 = \frac{x_1 - \bar{x}_1}{s_1} \text{ ，} u_2 = \frac{x_2 - \bar{x}_2}{s_2} \text{ ，…，} u_p = \frac{x_p - \bar{x}_p}{s_p} \qquad (3\text{-}2)$$

注意：在分析資料時，當變數的單位不同時，應考慮將此變數的資料先進行標準化（standardize）然後再做主成分分析。這麼做的主要原因是，如果變數具有不同的單位，則當某些變數的單位改變時，同樣的，分析所得到的主成分也會跟著改變[57]。

步驟 2：對標準化求相關係數矩陣

求取任兩變數間的相關係數：　$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

藉由相關係數計算相關係數矩陣：　$R = \begin{bmatrix} 1 & r_{x_1x_2} & \cdots & r_{x_1x_p} \\ r_{x_2x_1} & 1 & \cdots & r_{x_2x_p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{x_px_1} & r_{x_px_2} & \cdots & 1 \end{bmatrix}$

其中$r_{x_1x_p}$為變數x_1及變數x_p之相關係數，且$r_{x_1x_p} = r_{x_px_1}$ [58]。

步驟 3：計算特徵值和特徵向量

得到相關係數矩陣 R 後，將相關係數矩陣 R 代入公式（3-3）中之矩陣 A，來取得特徵值（eigenvalue）λ_i ，之後再將特徵值 λ_i 代入公式（3-4）中，來取得與特徵值 λ_i 相對應之特徵向量（eigenvector）X_i [58]。

$$|A - \lambda I| = 0 \qquad (3\text{-}3)$$

$$(A - \lambda \cdot I)\, X = 0 \qquad (3\text{-}4)$$

步驟 4：求取新成分

得到特徵值和特徵向量後，以相關係數矩陣 R 的第 1 特徵值 λ_1（也是最大特徵值）所對應的特徵向量 X_1 來計算第 1 個新成分 $y_1 = a_{11}u_1 + a_{21}u_2 + \cdots + a_{p1}u_p$（其中$X_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}$），接著，以相關係數矩陣 R 的第 2 特徵值 λ_2 所對應的特徵向量 X_2 來計算第 2 個新成分 $y_2 = a_{12}u_1 + a_{22}u_2 + \cdots + a_{p2}u_p$ ；之後便以此類推，也就是可依照前面所敘述的步驟逐步來計算第 i 個新成分 $y_i = a_{1i}u_1 + a_{2i}u_2 + \cdots + a_{pi}u_p$（i=3、4、…、p）[58]。

步驟 5：計算各新成分的貢獻率與累積貢獻率

在得到各新成分後，因為各新成分的解釋力都有所差異，因此為了知道各新成分的解釋力，接著計算各新成分 y_i 的貢獻率 $b_i = \frac{\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_p} = \frac{\lambda_i}{p}$（也就是代表解釋比例），並求算累積貢獻率 $c_i = \sum_{j=1}^{i} b_j$（也就是將各新變數的貢獻率給予加總）[58]。

步驟 6：主成分個數的選取

決定主成分的個數可以直接由特徵值本身數值大小的角度來決定，而選取的方法可以是：

1. 取特徵值大於全部平均值者。
2. 取特徵值大於 1 者（適用於標準化資料）。
3. 透過陡坡圖（scree plot），選取開始變平緩的點所對應的個數。
4. 經由正式的統計檢定（例如：Bartlett test）來決定。[57]

3.5.2 相關分析（Correlation Analysis）

1. 相關分析的定義

相關分析探討的是兩個變數之間的關聯程度（degree of association），普遍運用在各個學科，社會科學中的人文教育和管理學科（企管、資管、人管、行銷…等），例如：身高和體重、血壓和年齡…等。常用的相關分析有 Pearson 積差相關係數、Ø相關係數、點二系列相關、Spearman 等級相關和淨相關。因為 Pearson 積差相關係數（Pearson product-moment correlation coefficient）是適用於 2 個變數都是連續變數，也可以是區間變數（interval scale）或比率變數（ratio scale），因此本研究將採用「Pearson 積差相關係數」來做分析[59]。

2. 相關係數的特性

（1）相關係數的大小（magnitude）

表示在兩個變數之間，相關程度的強弱。當相關係數的絕對值愈大時，代表相關程度愈強；相對的，當相關係數的絕對值愈小時，代表相關程度愈弱；另外，當相關係數的值為 0 時，代表零相關，也就是沒有相關[59]。根據 Pearson 所設計出來的相關指標，其值介於 -1 到 +1 之間，則相關程度如表 3-3 所示：

表 3-3 相關係數與相關程度對照表

相關係數絕對值	約=1	0.7~0.99	0.4~0.69	0.1~0.39	0.01~0.09	約=0
相關程度	完全相關	高度相關	中度相關	低度相關	接近無相關	沒有相關

（來源：SPSS 操作與應用問卷統計分析實務[60] [61]）

（2）相關係數的方向（direction）

表示在兩個變數之間，是正相關還是負相關。當相關係數為正值時，代表兩個變數中的一個變數增加時，另一個變數也會增加；相對的，當相關係數為負值時，代表兩個變數中的一個變數增加時，另一個變數就會減少；反之亦然。[59]

3. 相關係數的計算

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{1}{N}\sum(X-u_x)(Y-u_y)}{\sqrt{\frac{1}{N}\sum(X-u_x)^2}\sqrt{\frac{1}{N}\sum(Y-u_y)^2}} \quad (3\text{-}5)$$

ρ_{XY}：X,Y 變數的相關係數　　σ_Y：Y 的標準差

σ_{XY}：X 與 Y 的共變數　　u_x：X 的平均值

σ_X：X 的標準差　　u_y：Y 的平均值

注意：ρ_XY 計算的是 X 和 Y 的總相關，值的大小位於±1 之間。[59]

第四章　研究結果與分析

4.1　主成分分析

主成分分析的主要目的：降低資料維度。而本研究卻是拿它來找出主要會影響 $PM_{2.5}$ 之相關變數。先將五種補值方法做補值後匯出，再將所有會影響 $PM_{2.5}$ 之相關變數透過主成分分析來找出主要因子，並利用 R 語言在處理主成分分析的套件來做運算(stats 套件的 prcomp()函數，它是專門在處理主成分分析的相關運算)，步驟說明如下：

步驟 1：投入變數做主成分分析

將 9 個空氣污染物（SO_2、CO、O_3、PM_{10}、NO_2、NO_X、NO、WindSpeed、WindDirec）變數進行主成分分析後，得到 9 個主成分資料（圖 4-1）。

```
Standard deviations (1, .., p=9):
[1] 1.9598876 1.2834844 0.9910755 0.9017403 0.7881490 0.7140348 0.5889548 0.4539629
[9] 0.1793540

Rotation (n x k) = (9 x 9):
             PC1         PC2          PC3         PC4         PC5         PC6
SO2  -0.25883712 -0.40903355 -0.063723908 -0.12301284  0.8086054  0.28896159
CO   -0.44485569 -0.04565146  0.191962060  0.13035407 -0.2462625  0.05088035
O3    0.19832147 -0.60854924  0.110274174  0.17692085 -0.1153137 -0.27690518
PM10 -0.27296415 -0.47569135  0.315580421 -0.06849428 -0.1269926 -0.42809007
NO2  -0.47355274  0.02413228  0.123375238  0.06081545 -0.1821554  0.29627471
Nox  -0.48182459  0.06402874  0.006913776  0.16140562 -0.1312025  0.19366934
NO   -0.27612972  0.17086826 -0.494281683  0.54493456  0.2329869 -0.53008023
WS    0.28803407 -0.25515227  0.002882195  0.71889323 -0.0890765  0.46933155
WD   -0.07515015 -0.36744028 -0.766646265 -0.29727387 -0.3787816  0.16562346
             PC7          PC8          PC9
SO2  -0.031265687 -0.090438481 -0.001946202
CO   -0.218049145 -0.797029408  0.011933924
O3   -0.652575891  0.176179766  0.002589438
PM10  0.622500043  0.085614243 -0.026486143
NO2  -0.107321785  0.417185300  0.670615119
Nox  -0.123455004  0.369082935 -0.729013410
NO    0.007543882  0.008523288  0.133756707
WS    0.321707009 -0.058362449  0.003099574
WD    0.088340346 -0.066717969  0.007740764
```

圖 4-1 主成分資料_9 個氣污染物（來源：本研究）

從圖 4-1 來看，Standard deviations 是每個主成分的標準差，當它進行平方後就是我們要的特徵值。Rotation 是每個主成分的負荷量(loading vector)。

步驟 2：陡坡圖和特徵值

由於主成分分析會將特徵值最大的因素先萃取出來，藉由陡坡圖（Scree plot）或特徵值（eigenvalue）來判斷，也就是說，當特徵值大於 1 時，就是我們需選擇的主成分。從圖 4-2 來看，用線標示出特徵值=1 的地方後，明顯出現兩個主成分（PC1 和 PC2），這就是我們要的東西。

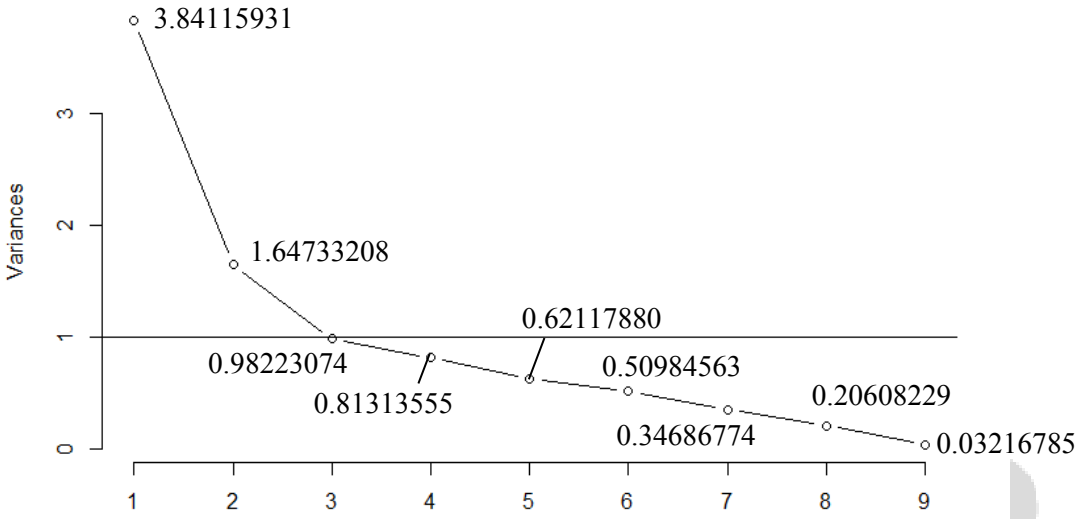


圖 4-2 陡坡圖和特徵值_補值一（approxExtrap）（來源：本研究）

步驟 3：因素負荷量圖和分數散佈圖

先將萃取出來的主成分進行轉軸後得到的因素負荷量，繪製成因素負荷量圖，透過此圖可以更清楚的看出他們之間的關係。在圖 4-3 中，從 PC1 裡找到兩個組合（NO_X＋NO_2＋CO、NO＋PM_{10}＋SO_2），從 PC2 裡找到兩個組合（PM_{10}＋SO_2＋WD、CO＋NO_2＋NO_X）。

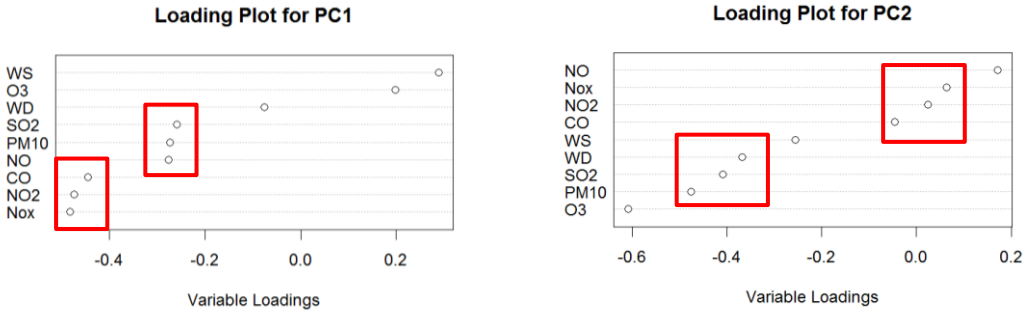
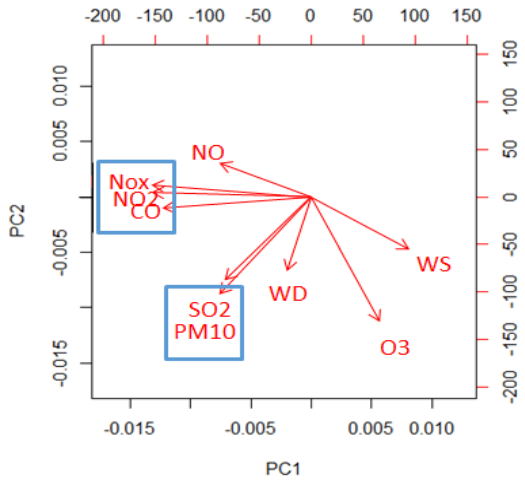


圖 4-3 補值一（approxExtrap）_因素負荷量圖

（PC1：第一主成分、PC2：第二主成分）（來源：本研究）



另外，也可以透過分數散佈圖來判別重要的因子，如圖 4-4 所示，紅色文字及箭頭表示變數位置及方向，從裡頭可以找到兩個組合：

（1）NO_X、NO_2、CO。

（2）SO_2、PM_{10}。

圖 4-4 補值一（approxExtrap）_分數散佈圖

（PC1：第一主成分、PC2：第二主成分）（來源：本研究）

本研究將五種補值方法進行主成分分析後，將得到的特徵值做整理（表 4-1）。

表 4-1 五種補值方法之特徵值

補值方法 / 主成分	補值一	補值二	補值三	補值四	補值五
PC1	**3.84115931**	**3.63769347**	**3.57963620**	**3.7249375**	**3.6562380**
PC2	**1.64733208**	**1.62303535**	**1.59779483**	**1.6252141**	**1.5685929**
PC3	0.98223074	1.01373788	1.10198616	0.9852986	0.9923317
PC4	0.81313555	0.84379911	0.80233610	0.8413293	0.8688188
PC5	0.62117880	0.68459188	0.61171561	0.6557507	0.7156597
PC6	0.50984563	0.57032986	0.52308938	0.5413577	0.5880143
PC7	0.34686774	0.37682250	0.45414184	0.3662837	0.3694739
PC8	0.20608229	0.23097682	0.31884733	0.2325436	0.2225190
PC9	0.03216785	0.01901314	0.01045254	0.0272847	0.0183518

註：補值一：approxExtrap、補值二：aregImpute、補值三：mice

補值四：missForest、補值五：rpart

PC1：第一主成分、PC2：第二主成分、…..、PC9：第九主成分

（來源：本研究）

表 4-2 五種補值方法之因素負荷量圖

主成分 / 補值方法	PC1：主成分 1	PC2：主成分 2
補值一：approxExtrap	NO_X＋NO_2＋CO NO＋PM_{10}＋SO_2	PM_{10}＋SO_2＋WD CO＋NO_2＋NO_X
補值二：aregImpute	NO_X＋NO_2＋CO PM_{10}＋NO＋SO_2	SO_2＋WD CO＋NO_2＋NO_X
補值三：mice	NO_X＋NO_2＋CO NO＋PM_{10}＋SO_2	SO_2＋WD CO＋NO_2＋NO_X
補值四：missForest	NO_X＋NO_2 PM_{10}＋SO_2＋NO	PM_{10}＋SO_2＋WD NO_2＋NO_X
補值五：rpart	NO_X＋NO_2＋CO PM_{10}＋NO＋SO_2	WD＋SO_2 NO_2＋NO_X

（來源：本研究）

表 4-3 五種補值方法之分數散佈圖

組合 / 補值方法	組合 1	組合 2
補值一：approxExtrap	NO_X＋NO_2＋CO	SO_2＋PM_{10}
補值二：aregImpute	NO_X＋NO_2＋CO	SO_2＋PM_{10}
補值三：mice	NO_X＋NO_2＋CO	SO_2＋PM_{10}
補值四：missForest	NO_X＋NO_2	SO_2＋PM_{10}
補值五：rpart	NO_X＋NO_2＋CO	SO_2＋PM_{10}

（來源：本研究）

本研究根據表 4-2 和表 4-3 的分析，找出重要影響 $PM_{2.5}$ 的相關變數為：（1）NO_X＋NO_2＋CO。（2）SO_2＋PM_{10}。

4.2 相關分析

相關分析的目的：探討兩個變數之間的關聯程度，如：身高和體重...等。而本研究是要藉由它來找出空污變數與 $PM_{2.5}$ 之間的關聯程度（強到弱）。

先將五種補值方法做補值後匯出，再將所有會影響 $PM_{2.5}$ 之相關變數透過 Pearson 相關分析來找出主要因子，整理後如表 4-4 所示：

表 4-4 相關係數

補值方法 / 空污變數	補值一	補值二	補值三	補值四	補值五
$PM_{2.5}$					
SO_2	0.42	0.36	0.38	0.41	0.28
CO	0.43	0.39	0.4	0.34	0.38
O_3	0.22	0.22	0.21	0.2	0.2
PM_{10}	0.79	0.76	0.77	0.79	0.78
NO_2	0.39	0.34	0.36	0.43	0.34
NO_X	0.34	0.29	0.31	0.37	0.31
NO	0.02	-0.01	0.02	0	0.03
WindSpeed	-0.18	-0.14	-0.15	-0.2	-0.16
WindDirec	0.14	0.14	0.13	0.1	0.13

註：補值一：approxExtrap、補值二：aregImpute、補值三：mice

補值四：missForest、補值五：rpart

（來源：本研究）

將表 4-4 轉換成相關程度後，可發現幾種組合：

表 4-5 相關程度

相關程度 / 補值方法	高度相關	中度相關	低度相關
補值一：approxExtrap	PM_{10}	CO、SO_2	NO_2、NO_X
補值二：aregImpute	PM_{10}		CO、SO_2、NO_2
補值三：mice	PM_{10}	CO	SO_2、NO_2、NO_X
補值四：missForest	PM_{10}	NO_2、SO_2	NO_X、CO
補值五：rpart	PM_{10}		CO、NO_2、NO_X

（來源：本研究）

本研究根據表 4-5 的分析，找出重要影響 $PM_{2.5}$ 的相關變數為：（1）PM_{10}。（2）SO_2。（3）CO、NO_2、NO_X。

4.3 時間序列

時間序列是一組隨著時間的先後順序所取得的資料，如：每小時的空污資料、每天股票價格資料、每週的銷售資料等。而本研究卻是拿它來做空污變數的未來 8 小時的預測值。另外，由於本研究使用五種補值方法，在此以補值方法一：approxExtrap 的大里測站做解釋，從圖 4-5 可觀察到 SO_2、CO、PM_{10}、$PM_{2.5}$、NO_2、NO_X 在 2018 年的狀態和趨勢。

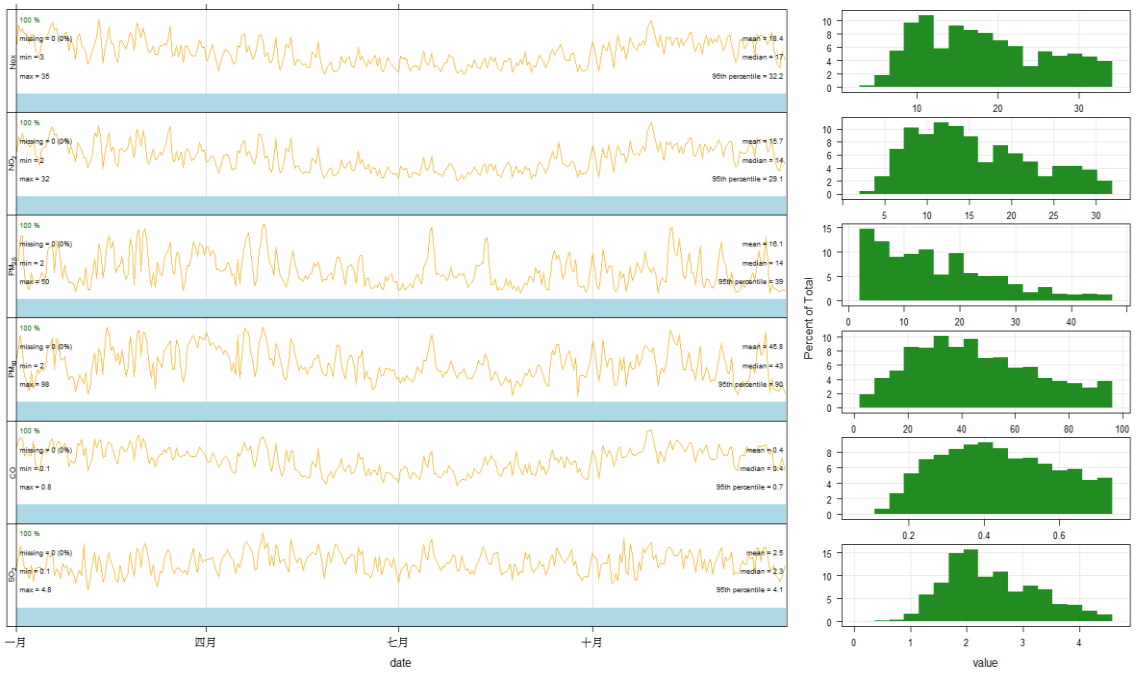


圖 4-5 各項汙染物之時間序列、頻數分布和統計分析

（平均值、中位/95%分位值、最小/大值、缺失值資料量及佔比）

（來源：本研究）

從圖 4-6 可觀察到由於測量的單位不同所以無法做比較，而圖 4-7 是經過標準化後的各項汙染物（SO_2、CO、PM_{10}、$PM_{2.5}$、NO_2、NO_X）之時間變化圖。從裡頭可觀察到：在「週與日變化」或「週變化」的部分，發現在星期日和星期一的空污較低，星期二到五較高。而在「日變化」部分，發現在 0 到 6 小時的空污較低，18 到 23 小時的空污較高。而在「月變化」部分，發現在 9 月到 12 月和 1 月到 5 月的空污較高，懷疑跟季節的風向有關係，因為冬天（12 月到隔年 2 月）和春天（3 月到 5 月）是吹東北季風，同時它可能會將來自中國大陸的境外污染給吹進臺灣，造成全臺空氣品質急速惡化，也就是氣象主播說的今天空氣「紫爆」了！

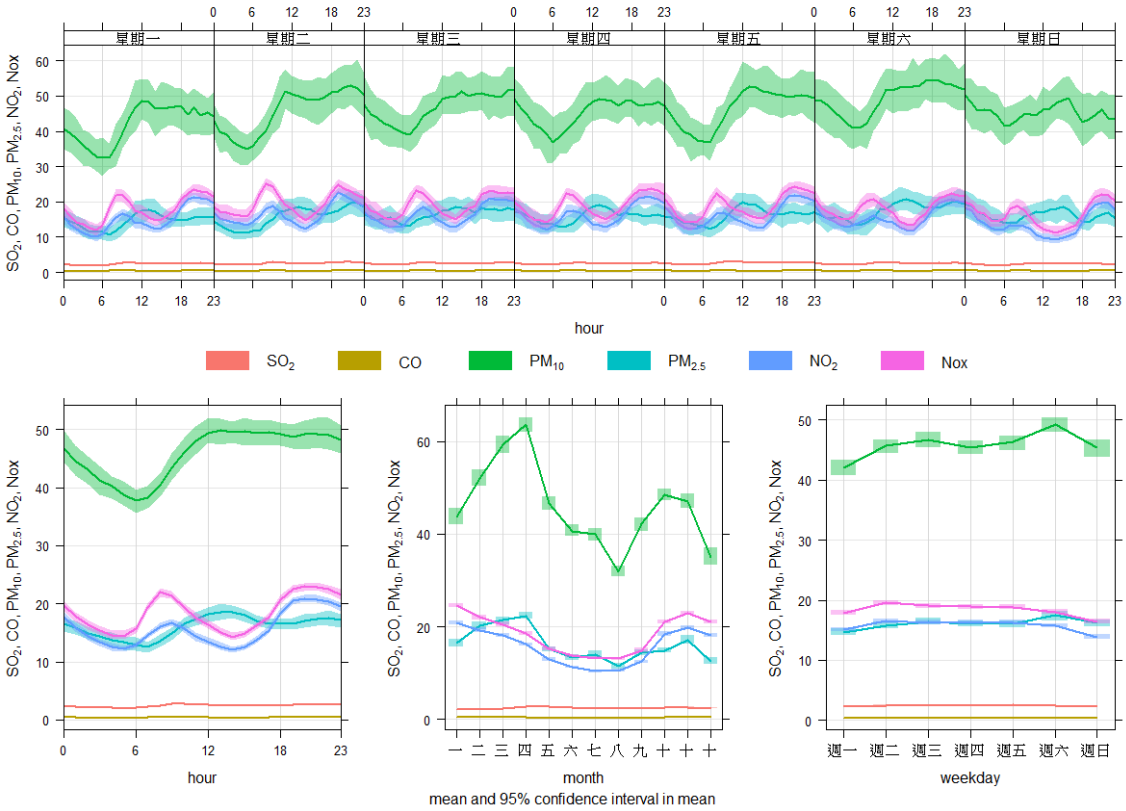


圖 4-6 各項汙染物之時間變化圖-均值及其 95%置信區間

（top-週與日變化，bottom-日變化、月變化、週變化）

（來源：本研究）

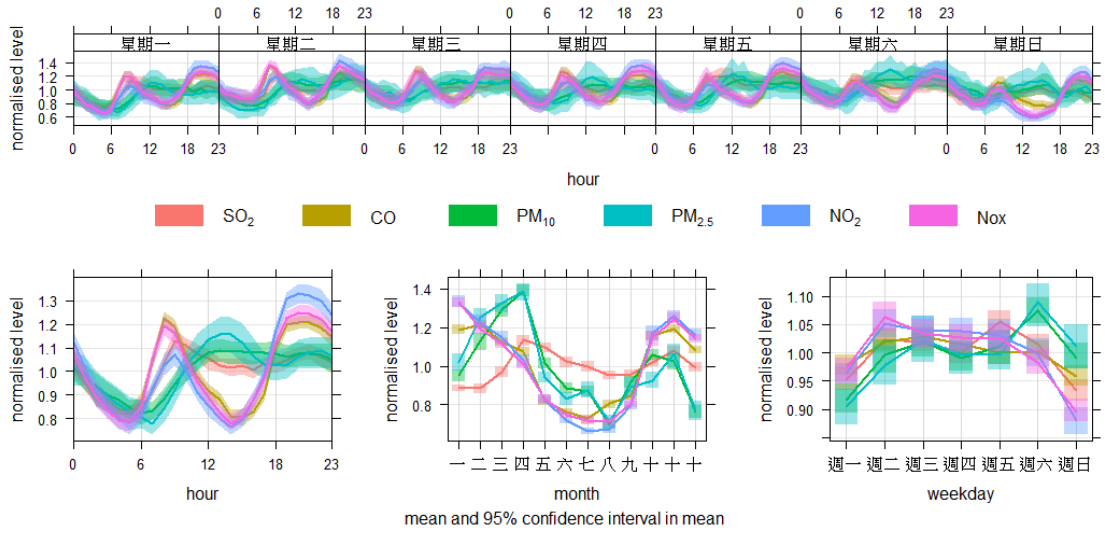


圖 4-7 標準化的各項汙染物之時間變化圖（均值及其 95%置信區間）

（top-週與日變化，bottom-日變化、月變化、週變化）

（來源：本研究）

反之，夏天（6 月到 9 月）是吹西南季風，同時它可能也將臺灣內部的交通與工業等因素產生的境內污染給吹到中國大陸去了。

本研究在未來預測值部分，利用 R 語言在處理時間序列的套件（forecast 套件的 auto.arima()函數，它是使用 ARIMA 模型做預測）的方法來取得，投入資料的時間以預測 2019/1/1 00:00~07:00 的前三個月（10、11、12 月）的 00:00~23:00 為主，如圖 4-8 所示：黑色線條代表原始數值的曲線，藍色線條代表預測出來的未來趨勢，並提供較高及較低的信賴區間，而圈起來的是我們要的未來數值。

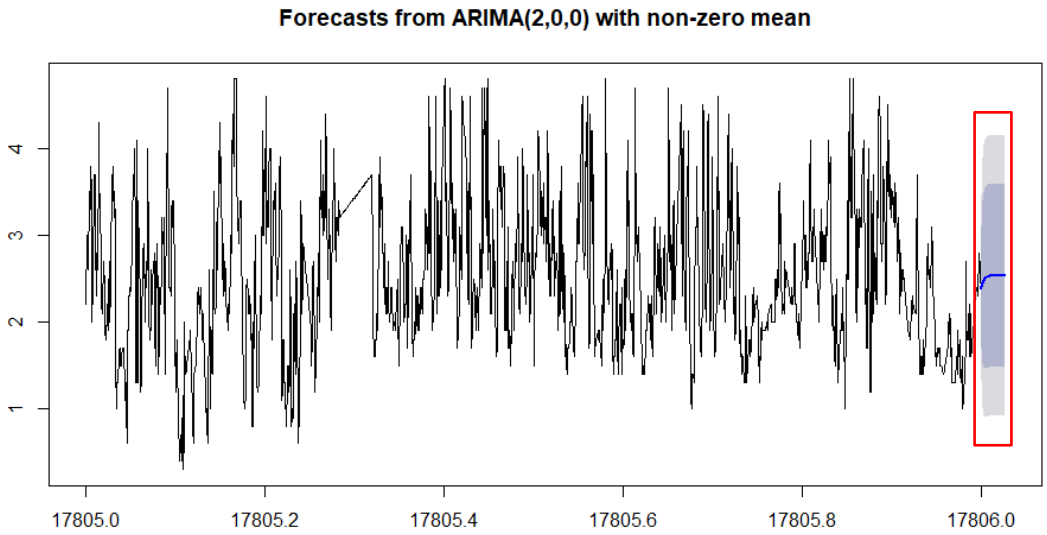


圖 4-8 2019/1/1 未來趨勢預測_SO_2（來源：本研究）

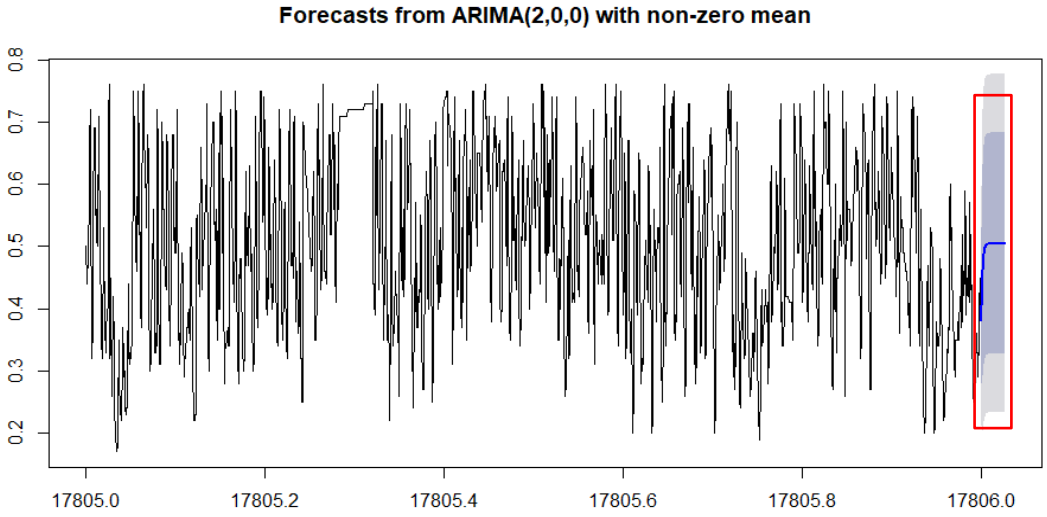


圖 4-9 2019/1/1 未來趨勢預測_CO（來源：本研究）

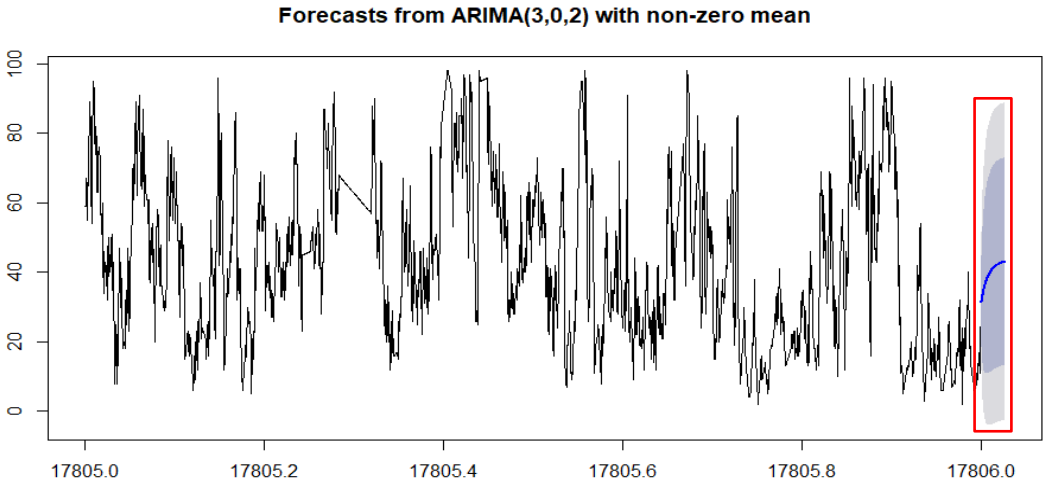


圖 4-10　2019/1/1 未來趨勢預測_PM_{10}（來源：本研究）

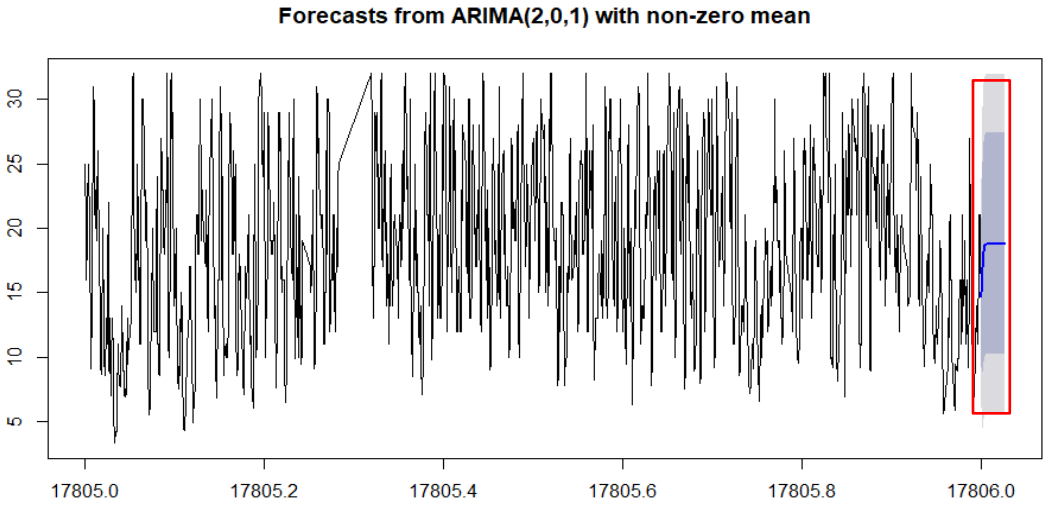


圖 4-11　2019/1/1 未來趨勢預測_NO_2（來源：本研究）

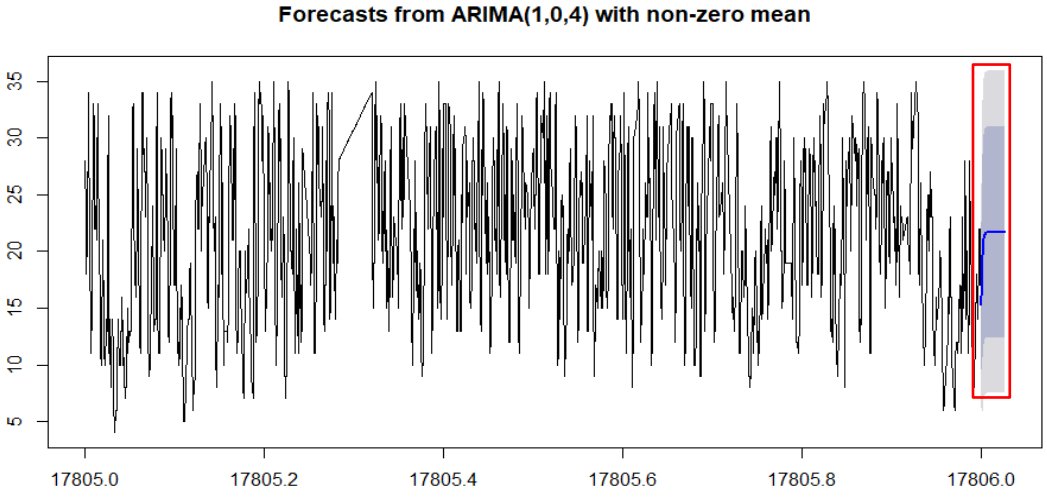


圖 4-12　2019/1/1 未來趨勢預測_NO_X（來源：本研究）

4.4 遞迴神經網路

遞迴神經網路（recurrent neural network, RNN）是適合處理與時間序列相關性的資料，但缺點是會產生梯度爆炸（gradient exploding）或是梯度消失（gradient vanishing）的問題。而長短期記憶（long short-term memory, LSTM）是目前 RNN 最常使用的模型，因為它可以解決 RNN 的問題。所以本研究採用 RNN 的 LSTM 模型來建立預測 $PM_{2.5}$ 的模型。

4.1 建立模型

本研究利用 R 語言的 keras 套件來建立預測模型，其說明如下：

1. 訓練和驗證：

[整筆資料丟入] 1 個測站的資料量=365×24=8760 筆，在訓練和驗證的分配比例是 8：2，也就是訓練 7008 筆、驗證 1752 筆。

[分批丟入] 取要預測某天 8 小時的前一個月的資料（31×24=744 筆），分成訓練及驗證（736 筆）和測試（8 筆）。

採批次方式：第一次（第 1~200 筆）、第二次（第 2~201 筆）、……、第 545 次（第 537~736 筆），這樣每個小時都能訓練到，每批為 200 筆資料。在訓練和驗證的分配比例是 8：2，也就是訓練 140 筆、驗證 60 筆。其中「每批次取 200 筆」，是根據 12 月份的風速與 $PM_{2.5}$ 濃度之關係來決定的。從圖 4-13 發現，大部分時當風速增強時，$PM_{2.5}$ 濃度就會下降；反之，當風速減弱時，相對的 $PM_{2.5}$ 濃度就會增加。

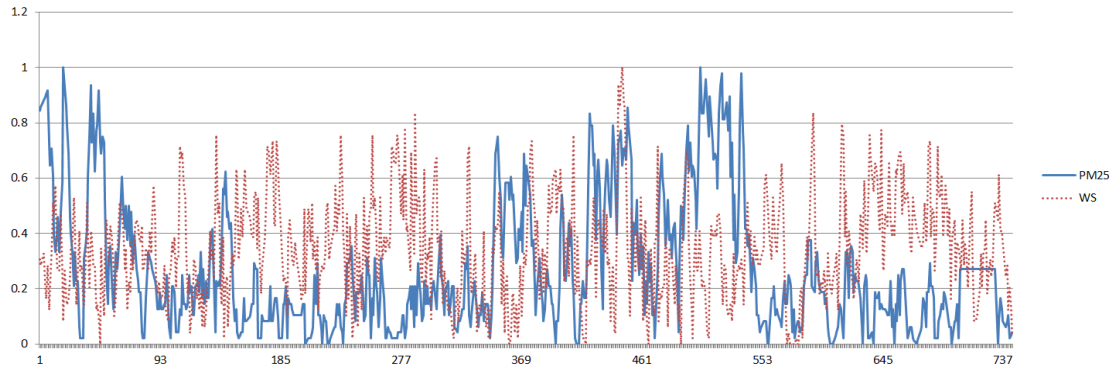


圖 4-13 豐原測站_風速與 $PM_{2.5}$ 變化（來源：本研究）

2. 自變數（X）與依變數（Y）：

[整筆資料丟入]

將 $PM_{2.5}$ 設為依變數，會影響 $PM_{2.5}$ 之相關變數（SO_2、CO、PM_{10}、NO_2、NO_X、$PM_{10} + SO_2$、$NO_X + NO_2 + CO$）設為自變數。

[分批丟入]

將資料採用時間步（t-1、t）方式，將 t：$PM_{2.5}$ 設為依變數。t-1：$PM_{2.5}$、SO_2、CO、PM_{10}、NO_2、NO_X、$PM_{10} + SO_2$、$NO_X + NO_2 + CO$ 設為自變數。

3. 其餘參數設定：

經過數次的測試後，將學習率（Learning Rate, lr）設為 0.01、學習率隨之衰減的比率（decay）設為 1e-3、batch_size 設為 24、epochs 設為 100。

本研究將建立好的預測 $PM_{2.5}$ 模型，投入 X 和 Y 後開始進行訓練，待訓練完成後再做驗證與 MAPE 評估。本研究以豐原測站為目標所做的 5 種補值方法與 7 個模型驗證之誤差值做整理（表 4-31 到表 4-32）：

表 4-6 豐原測站_5 種補值方法與 7 個模型驗證誤差_[整筆]

補值方法 / 模型	補值一 approxExtrap	補值二 aregImpute	補值三 mice	補值四 missForest	補值五 rpart
SO_2	0.8989	0.9344	0.9132	0.9065	0.8921
CO	1.0299	1.0188	0.9968	0.9744	0.9668
PM_{10}	0.7809	0.8312	0.8133	0.8051	0.7460
NO_2	1.0080	1.0012	0.9797	0.9710	0.9594
NO_X	1.0049	0.9937	0.9752	0.9639	0.9370
$PM_{10} + SO_2$	0.7820	0.8456	0.8102	0.8154	0.7546
$NO_X + NO_2 + CO$	1.0026	1.0011	0.9827	0.9657	0.9526

（來源：本研究）

表 4-7 豐原測站_5 種補值方法與 7 個模型驗證誤差_[分批]

補值方法 / 模型	補值一 approxExtrap	補值二 aregImpute	補值三 mice	補值四 missForest	補值五 rpart
SO_2	0.2223	0.2455	0.2304	0.2345	0.2259
CO	0.2405	0.2381	0.2226	0.2125	0.2209
PM_{10}	0.2465	0.2637	0.2435	0.2326	0.2382
NO_2	0.2123	0.2403	0.2149	0.2259	0.2117
NO_X	0.2355	0.2394	0.2397	0.2426	0.2139
$PM_{10} + SO_2$	0.2795	0.3012	0.2956	0.3149	0.2814
$NO_X + NO_2 + CO$	0.2916	0.3083	0.2900	0.3218	0.2865

（來源：本研究）

4.2 預測模型

利用 R 語言內的 keras 套件，將前面建立好的模型，藉由 predict()來做 $PM_{2.5}$ 未來 8 小時的預測。在預測值部分可分成：[整筆資料丟入]：利用第三節時間序列來找出這些影響 $PM_{2.5}$ 之相關變數的未來 8 小時數值。[分批丟入]：取前 8 小時的資料來做未來相關變數的預測值。本研究以豐原測站為目標所做的 5 種補值方法與 7 個模型預測之誤差值做整理（表 4-33 到表 4-34）：

表 4-8 豐原測站_5 種補值方法與 7 個模型預測誤差_[整筆]

補值方法 / 模型	補值一 approxExtrap	補值二 aregImpute	補值三 mice	補值四 missForest	補值五 rpart
SO_2	1.9291	2.0859	1.9801	2.1422	1.9334
CO	2.3497	1.9084	2.2041	1.6657	2.1422
PM_{10}	2.1752	2.1672	1.8764	2.3618	1.9558
NO_2	2.3514	2.2649	2.2089	2.1217	2.1517
NO_X	2.1538	2.1049	2.2021	2.3053	2.1913
$PM_{10} + SO_2$	2.5716	2.5246	2.1957	2.7573	2.1485
$NO_X + NO_2 + CO$	2.7865	2.0585	2.3435	2.3038	2.5557

（來源：本研究）

表 4-9 豐原測站_5 種補值方法與 7 個模型預測誤差_[分批]

補值方法 / 模型	補值一 approxExtrap	補值二 aregImpute	補值三 mice	補值四 missForest	補值五 rpart
SO_2	0.4521	0.3984	0.4747	0.3621	0.3837
CO	0.4428	0.4153	0.3780	0.6315	0.3967
PM_{10}	0.4367	0.4408	0.5111	0.4890	0.3728
NO_2	0.4364	0.3775	0.4626	0.4140	0.3676
NO_X	0.4365	0.3936	0.4440	0.4107	0.3835
$PM_{10} + SO_2$	0.4615	0.4495	0.4185	0.3188	04111
$NO_X + NO_2 + CO$	0.4536	0.4491	0.3896	0.4600	0.4295

（來源：本研究）

第五章　結論與未來研究方向

5.1　結論

由於近幾年空汙議題的盛行，到目前為止，國內研究出很多預測 $PM_{2.5}$ 濃度的方法（例如用 LSTM、迴歸…等），但做的都是預測方法的比較，又或者做的純粹是分析而已（例如用相關分析、因子分析…等），卻很少會有人用空污資料來做經過各種補值後所做的預測比較。本研究將空污資料做資料預處理時採用 5 種處理遺失值的方法（approxExtrap、aregImpute、mice、missForest、rpart）來補值，補完後匯出會產生 5 個檔案（.csv），再藉由主成分分析和相關分析找出影響 $PM_{2.5}$ 之相關變數（單因子：PM_{10}、SO_2、NO_X、NO_2、CO，雙因子：NO_X＋NO_2＋CO、SO_2＋PM_{10}），接著建立 LSTM 模型，之後再把未來數值投入建立好的模型內，觀察參數的設定找出最佳的誤差率。其中，未來數值指的是利用時間序列來找出相關變數的未來數值又或者是取預測前 8 小時的資料為未來數值。根據研究結果顯示，豐原測站的預測值與真實值之誤差大部分都有落在合理的 MAPE（0.2~0.5）範圍內，這就可以確定此模型是可以被研究的。另外在補值法方面是以線性插值法最好。

5.2　未來研究方向

為了使本研究的內容更加充分，以下將詳述未來可研究發展之方向：

1.　加入其他變數

目前所找到的特徵值部分，是由 SO_2、CO、O_3、PM_{10}、NO_2、NO_X、NO、WindSpeed、WindDirec 下去分析而找出來的，但除了這 9 個因子以外還可再加入其他變數，如：氣溫、雨量、THC、NMHC、RH…等天氣因子。

2.　參數的調整

在 LSTM 內的參數（units、batch_size、epochs…等）設定、層數多寡、優化器的選擇…等都會影響訓練及預測結果。而這次本研究所設定的參數值，在

批次資料部分雖然驗證誤差和預測誤差的值都落在 MAPE 認定合理之評估標準內，但還不是最好的，所以還得再花時間做調整。

3. 預測時間比較

既然已建模並預測出未來 8 小時 $PM_{2.5}$ 濃度數值，那麼也可以用來嘗試做預測 4 小時、6 小時、12 小時、24 小時的預測和實際值差多少，利用 MAPE 認定的評估標準來看幾小時是最好的。

參考文獻

[1] Chen, C.F.（2017），〈【環保無藍綠】台中空污惡化還是改善了？誰的功與過？〉，六都春秋 LADO POST，線上資料，https://www.ladopost.com/newsDetail1.php?ntId=3&nId=1999，2019/02/27。

[2] 〈空污「早在 2010 就紫爆」竟被當成「霧」〉，《蘋果日報》，線上資料，https://web.archive.org/web/20180308073047/https://tw.appledaily.com/new/realtime/20180104/1272291/，2019/02/27。

[3] Huang, Abby（2017），〈1217 反空汙大遊行，什麼才是全台灣最大的汙染源？〉，The News Lens 關鍵評論網，線上資料，https://www.thenewslens.com/article/85695，2019/02/27。

[4] 〈臺灣空氣污染〉，維基百科，線上資料，https://zh.wikipedia.org/wiki/%E8%87%BA%E7%81%A3%E7%A9%BA%E6%B0%A3%E6%B1%A1%E6%9F%93，2019/03/02

[5] 〈[資訊]環保署自 105 年 12 月 1 日採用空氣品質指標（AQI）〉，國家環境毒物研究中心，線上資料，http://nehrc.nhri.org.tw/toxic/news.php?cat=news&id=349，2019/03/01。

[6] 〈空氣品質指標〉，行政院環境保護署，線上資料，https://taqm.epa.gov.tw/taqm/tw/b0201.aspx，2019/03/01。

[7] 董俞佳、林佩均、張明慧（2017），〈PM2.5 災區 南部人幾乎每天吸髒空氣〉，聯合新聞網，線上資料，https://news.housefun.com.tw/news/article/208274155010.html，2019/03/02。

[8] 〈WHO 最新數據！每年 700 萬人死於空污疾病〉，《民視新聞》，線上資料，https://www.ftvnews.com.tw/news/detail/2018A30I20M1，2019/03/03。

[9] 〈研究：PM2.5 空污縮短人類預期壽命 亞洲影響最劇〉，《自由時報》，線上資料，https://news.ltn.com.tw/news/world/breakingnews/2530451，2019/03/03。

[10] 洪毓琪（2018），〈106 十大死因：癌症連 36 年最奪命！子宮頸癌再入榜〉，華人健康網，線上資料，https://www.top1health.com/Article/57875，2019/03/03。

[11] 羅真（2017），〈台灣慢性病危險因子排名 第四竟是 PM2.5 暴露〉，聯合新聞網，線上資料，https://health.gvm.com.tw/article.html?id=63779，2019/03/02。

[12] 〈空氣污染對健康的影響〉，香港特別行政區政府衛生署衛生防護中心，線上資料，https://www.chp.gov.hk/tc/healthtopics/content/460/3557.html，2019/03/02。

[13] 李政霖（2019），《即時空氣品質動態監測系統結合 LSTM 模型預測 PM2.5 濃度之應用》，國立臺北科技大學電機工程系碩士論文，出版。

[14] 顧芷瑄（2018），《探討能源決策管理：應用機器學習於空氣汙染預測之研究》，國立清華大學工業工程與工程管理學系碩士論文，出版。

[15] 林冠名（2018），《在大數據平台使用機器學習方法預測空氣汙染》，國立臺北大學資訊工程學系研究所碩士論文，出版。

[16] 林俞均（2018），《應用自適應神經模糊推理技術預測空氣品質指數》，國立中山大學電機工程學系研究所碩士論文，出版。

[17] 黃彥齊（2018），《類神經網路應用於空氣品質預測與異常偵測之研究》，國立交通大學環境工程系碩士論文，出版。

[18] 盧俊源（2017），《應用 ARIMA、計算智能及混合方法以預測 PM2.5 濃度-以台灣六都為例》，輔仁大學統計資訊學系應用統計碩士在職專班碩士論文，出版。

[19] 陳亦云、許嘉真（2018），〈PM2.5、SO2、臭氧，怎麼避免這些空氣汙染物的傷害？〉，HEHO：Health & Hope，線上資料，https://heho.com.tw/archives/23255，2019/04/01。

[20] 吳和桔（2018），〈淨化空氣品質之探究與實作〉，《科學研習月刊》，第 57 卷第 10 期，台灣網路科教館，線上資料，https://www.ntsec.edu.tw/LiveSupply-Content.aspx?cat=6841&a=6829&fld=&key=&isd=1&icop=10&p=1&lsid=14086，2019/04/01。

[21] 〈硫氧化物〉，華人百科，線上資料，https://www.itsfun.com.tw/%E7%A1%AB%E6%B0%A7%E5%8C%96%E7%89%A9/wiki-9927306-1844285，2019/04/01。

[22] 〈二氧化硫〉，維基百科，線上資料，https://zh.wikipedia.org/wiki/%E4%BA%8C%E6%B0%A7%E5%8C%96%E7%A1%AB，2019/04/01。

[23] 〈空氣污染物簡介〉，《汙染小百科》，高雄市政府環境保護局空氣品質管理中心，線上資料，https://www.ksaqmc.com.tw/PollutionInfo.aspx，2019/04/01。

[24] 〈氮氧化物〉，維基百科，線上資料，https://zh.wikipedia.org/wiki/%E6%B0%AE%E6%B0%A7%E5%8C%96%E7%89%A9，2019/04/01。

[25] 〈各項汙染物〉，行政院環境保護署，線上資料，https://taqm.epa.gov.tw/taqm/tw/b0202.aspx，2019/04/01。

[26] 李建業（2018），《2010-2014 中部空品測站 PM2.5 與其它空品指標污染物之分析》，國立中興大學環境工程學系碩士論文，出版。

[27] 陳柏丞（2017），《台中市 PM2.5 與其他空氣污染物之關係》，東南科技大學營建與空間設計系營建科技與防災碩士班碩士論文，出版。

[28] 陳正暐（2014），《中部空品區 PM2.5 排放污染源分析》，國立中興大學環境工程學系碩士論文，出版。

[29] 邱瑞仙（2008），《桃園地區空氣污染物濃度相關性及地理分布》，國立中央大學環境工程研究所碩士在職專班碩士論文，出版。

[30] 吳奎縉（2007），《中部空品區 PM2.5 污染來源分析》，國立中興大學環境工程學系所碩士論文，未出版。

[31] 曲建仲（2017），〈翻轉人類未來的 AI 科技：機器學習與深度學習〉，TechNews 科技新報，線上資料，http://technews.tw/2017/10/05/ai-machine-learning-and-deep-learning/，2019/04/02。

[32] 李仁鐘（2015），《應用 R 語言於資料分析：從機器學習、資料探勘到巨量資料》，台北市：松崗。

[33] 陳奕廷（2016），〈機器學習與人工神經網路（二）：深度學習（Deep Learning）〉，CASE 報科學，線上資料，https://case.ntu.edu.tw/blog/?p=26340，2019/04/20。

[34] 許哲昇、江振瑞（2018），《基於長短期記憶遞迴神經網路深度學習之剩餘可用壽命預測》，NCS 2017 全國計算機會議：國立東華大學，P761-766。

[35] G. E. H. David E. Rumelhart, Ronald J. Williams., "Learning representations by back-propagating errors," Nature, 323,533-536., 1986.

[36] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv preprint arXiv:1609.04747, 2016.

[37] 〈Understanding LSTM networks〉，colah's blo，online，http://colah.github.io/posts/2015-08-Understanding-LSTMs/，2019/04/21。

[38] a. J. S. S. Hochreiter, "Long short-term memory," Neural Computation 9(8):1735-80., 1997.

[39] J. Zhao, F. Deng, Y. Cai, and J. J. C. Chen, "Long short-term memory-Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction," vol. 220, pp. 486-492, 2019.

[40] 蔡宜廷（2018），《基於聚合神經網路之空氣污染預測與分析》，國立臺北大學資訊工程學系碩士論文，出版。

[41] Y.-T. Tsai, Y.-R. Zeng, and Y.-S. Chang, "Air pollution forecasting using RNN with LSTM," in *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, 2018, pp. 1074-1079: IEEE.

[42] C.-F. LEE, "Air Quality Monitoring and Analysis with Automatic Forecasting Using Machine Learning," COMPUTER SCIENCE DEPARTMENT, TUNGHAI UNIVERSITY, 2018.

[43] T.-C. Bui, V.-D. Le, and S.-K. J. a. p. a. Cha, "A Deep Learning Approach for Forecasting Air Pollution in South Korea Using LSTM," 2018.

[44] 盧慧鴻（2018），《以 LSTM 預測細懸浮微粒值規畫最佳行徑路線》，國立暨南國際大學資訊管理學系碩士論文，出版。

[45] V. Athira, P. Geetha, R. Vinayakumar, and K. J. P. c. s. Soman, "DeepAirNet: Applying recurrent networks for air quality prediction," vol. 132, pp. 1394-1403, 2018.

[46] 鍾玉峰、張文鎰、蔡惠峰、蘇威智（2018），《基於深度學習之環境空污智能預測》，TANET2018 臺灣網際網路研討會：國立中央大學，P1185-1190。

[47] X. Li *et al.*, "Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation," vol. 231, pp. 997-1004, 2017.

[48] M. H. Kim, Y. S. Kim, J. Lim, J. T. Kim, S. W. Sung, and C. J. K. J. o. C. E. Yoo, "Data-driven prediction model of indoor air quality in an underground space," vol. 27, no. 6, pp. 1675-1680, 2010.

[49] 〈R 語言〉，維基百科，線上資料，https://zh.wikipedia.org/wiki/R%E8%AF%AD%E8%A8%80，2019/05/05。

[50] 連珮宇（2018），〈空汙排名出爐！北市居六都最佳 台中首度擠進前十〉，聯合新聞網，線上資料，https://udn.com/news/story/7314/3353040，2019/04/02。

[51] 〈填補遺漏值（Missing Value）方法〉，R-自學筆記，線上資料，https://derek97072002rlanguage.blogspot.com/2017/03/20170303.html，2019/05/05。

[52] 〈資料預處理〉，華人百科，線上資料，https://www.itsfun.com.tw/%E8%B3%87%E6%96%99%E9%A0%90%E8%99%95%E7%90%86/wiki-6252113-4020592，2019/05/05。

[53] 〈Missing Value Treatment|遺失值處理|統計 R 語言〉，Jam Jam 果醬珍珍健忘女孩的學習筆記和生活雜記，線上資料，https://www.jamleecute.com/missing-value-treatment-%E9%81%BA%E5%A4%B1%E5%80%BC%E8%99%95%E7%90%86/，2019/05/05。

[54] 〈箱形圖〉，維基百科，線上資料，https://zh.wikipedia.org/wiki/%E7%AE%B1%E5%BD%A2%E5%9C%96，2019/05/07。

[55] 陳丘原（2016），〈離群值的檢測（Detection of Outliers）〉，科學 Online 高瞻自然科學教學資源平台，線上資料，http://highscope.ch.ntu.edu.tw/wordpress/?p=73655，2019/05/07。

[56] 蔡大偉（2005），〈不同模式之預測能力研究〉，《水土保持學報》，第 37 卷，第 2 期，第 127-138 頁，線上資料，http://ir.lib.nchu.edu.tw/bitstream/11455/78721/1/37-2-2.pdf，2019/05/07。

[57] 林師模、陳苑欽（2013），《多變量分析：管理上的應用（二版）》，台北市：雙葉書廊。

[58] 許邦輝（2006），《以主成分分析法為基礎之文件自動分類模式》，國立清華大學工業工程與工程管理學系碩士論文，出版。

[59] 蕭文龍（2016），《統計分析入門與應用：SPSS 中文版＋SmartPLS 3（PLS_SEM）》，台北市：碁峯資訊。

[60] 吳明隆（2009），《SPSS 操作與應用-問卷統計分析實務》，臺北市：五南。

[61] 靳孟霖（2015），《12 年國教的認知及學習態度對國中數學學習成效之影響：利用線性模型分析》，東吳大學數學系碩士論文，出版。