國立臺灣大學工學院環境工程學研究所

碩士論文

Graduate Institute of Environmental Engineering

College of Engineering

National Taiwan University

Master Thesis

應用多變量分析及機器學習技術於老街溪水質評估

Assessment of Lao-Jie River Water Quality Using Multivariate Statistics and Machine Learning Techniques

鄭奕晴

Yi-Ching Cheng

指導教授：于昌平 博士

Advisor: Chang-Ping Yu, Ph.D.

中華民國 107 年 8 月

August 2018

doi:10.6342/NTU201803135

國立臺灣大學碩士學位論文
口試委員會審定書

應用資料探勘技術於老街溪水質評估

Assessment of Lao-Jie River Water Quality

Using Data Mining Techniques

本論文係鄭奕晴君(學號 r04541210)在國立臺灣大學環境工程學研究所完成之碩士學位論文，於民國 107 年 6 月 26 日承下列考試委員審查通過及口試及格，特此證明

論文審查委員：

于昌平 — 于昌平博士
國立台灣大學環境工程學研究所副教授

郭獻文 — 郭獻文博士
東海大學環境科學與工程學系副教授

黃郁慈 — 黃郁慈博士
中原大學環境工程學系副教授

指導教授：于昌平

所　　長：林立宏

doi:10.6342/NTU201803135

誌謝

交出一本碩士論文，結束這幾年的研究所生涯，雖然成果離預想的曠世巨作等級有段距離，至少是本對得起自己的作品 (有沒有誤人子弟就不知道了)。寫作期間經歷了認清自己的極限後，從牛角中鑽出來的過程，測試了自學的能力，啟動腦內學術研究區塊，最終收獲了名為「研究生看不到盡頭的寫作日常」等寶貴經歷。感謝一路走來遇到的所有人們，讓我自由發展的人、指點迷津的人、給建議的人、分享經驗的人、拿著鞭子蘿蔔的人和默默守護支持的人，你們貢獻的靈感和力量，促成這本論文的誕生。

每個選擇都有它的意義，先進入職場學到了寫文章的邏輯和上台報告的膽量，工作一段時間再重回校園後會更珍惜學習的機會；選擇這個時間轉換為學生身分，是為了在家人需要時有更多彈性時間陪伴；來到環工所是為了遇到一群同樣愛地球(和十萬元)的朋友、自帶聖光的指導老師及可愛的實驗室夥伴們。感謝路上遇到的所有挑戰和過程，這些經歷打磨成現在的我，對於自己在人生岔路口的選擇和堅持感到驕傲。

回顧這些時間學習到的知識，似懂非懂的工程技術、不知從何下手的規劃管理理論、實驗消耗的資源及產出的廢棄物、同學手中不減的塑膠袋和拋棄式餐具，PM2.5 從哪裡來？玻璃吸管有沒有比較環保？地球的問題一樣難解。期待接下來的旅程能讓我更了解環境的真相，不管以後在什麼地方做了什麼樣的工作，一定不會忘記想為地球做點好事的初衷。

鄭奕晴 2018.8.13

doi:10.6342/NTU201803135

摘要

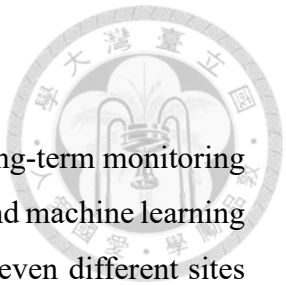
本研究目的在於探索長期的水質監測數據，使用多變量分析及機器學習等資料探勘方式，探索河水中污染物隨時間的變化及彼此的關聯。蒐集桃園市老街溪流域 2002 至 2016 年共計 15 年期間水質監測數據，河川主流長約 37 公里，分析範圍包括主流老街溪及支流大坑缺溪所設置之 7 點測站，每月各蒐集 10 至 32 項水質參數。

所產生的龐大水質資料集(總共約 21,194 個觀測值)將以多變量分析方法中的主成分分析、因素分析及群集分析方法進行水質評估，除了水質特徵識別外，還加入了時間軸，探討水質隨時間之變化。經主成分及因素分析，萃取出的 6 個因素可解釋資料集 70％變異量，因素依序為複合污染物、降雨沖刷、工業排水污染(半導體業、印刷電路板業等)及工業常見金屬材料等污染來源；群集分析將 7 個測站分類為 3 個群組，分別為支流，上游群組及下游群組，高度污染的支流匯入主流後，影響中下游水質，導致上下游群組組成逐年變化。

機器學習亦可用於水質監測集的資料探勘上，本研究為判斷水中銅濃度超標與否及評估河川污染程度指標(RPI)，同時利用決策森林模型及類神經網路模型等兩種技術，針對上述議題分別建立模型。在判斷水中銅濃度超標與否的議題上，決策森林模型之正確率較高(0.83)，同時可得知懸浮固體、導電度及點位因素是判斷超標與否的重要決策指標；而在評估 RPI 數值上，同樣是決策森林模型的評估誤差較小，平均絕對誤差及平均絕對誤差百分比分別為 0.352 及 0.087，並可得知生化需氧量及氨氮為重要的決策資訊。

關鍵字：水質監測、水質評估、多變量分析、主成分分析、因素分析、群集分析、機器學習、決策森林、類神經網路

doi:10.6342/NTU201803135



Abstract

This study investigated the water quality of river basin from a long-term monitoring dataset using data mining techniques, such as multivariate statistical and machine learning techniques. Water quality of Lao-Jie River basin was monitored at seven different sites from mainstream and tributary Da-Keng-Que creek, with 10-32 water quality parameters collected every month for 15 years (2002–2016).

Multivariate statistical techniques, such as Principal Components Analysis (PCA), Factor Analysis (FA) and Cluster Analysis (CA), were applied to evaluate the water quality of the large size monitoring dataset (21,194 observations). PCA/FA identified six factors that explains 70 % of the variance in the dataset. These six factors indicated the source of the pollutions might originate from complex pollutions, rain erosion, industrial wastewater effluent (like semiconductor industry and printed circuit board industry), and industrial metal pollution. Furthermore, CA classified seven sampling sites into three groups: tributary, upstream groups, and downstream groups, while members in upstream and downstream groups change by year due to highly polluted tributary.

Machine learning can also be used for data exploration in water quality monitoring datasets. This study addressed two water quality assessment issues, the concentration of copper and the river pollution index (RPI), by using both decision forest and neural network techniques respectively. In terms of the concentration of copper, decision forest has a higher accuracy (0.83), and elucidates that suspended solids, electrical conductivity, and sampling sites are important in determining whether the copper concentration in the water is standard-exceeded or not. On the other hand, for the assessment of the RPI, decision forest model also has a lower mean absolute error and mean absolute percentage error (0.352 and 0.087), and BOD as well as ammonia play important roles in decision-making information.

Keywords: Water Monitoring; Water Quality Assessment; Multivariate Statistical Techniques; Principal Components Analysis; Factor Analysis; Cluster Analysis; Machine Learning; Decision Forest; Neural Network

doi:10.6342/NTU201803135

目錄

doi:10.6342/NTU201803135

doi:10.6342/NTU201803135

圖目錄

doi:10.6342/NTU201803135

表目錄

doi:10.6342/NTU201803135

第一章　緒論

1.1 研究動機及目的

監測資料對於維持良好河川水質的重要性

河川水源可作為灌溉、飲用或工業水源，有些甚至具有運輸價值，因此河川流域周遭總是聚集了高密度的人口。河川流域周遭密集的人為活動也讓水質處於易被污染的狀態，如沿岸住家或工廠的廢水排放，除此之外，河川水質也易受各種因素影響，如地表逕流、氣候狀況、流域地質等，都是造成河川污染的原因。河川水質的好壞會直接地影響周遭居民的健康，以及水中生物的生存，環境管理單位需要可靠的監測資料及水質分析訊息，以維持河水狀況良好。爰此，蒐集長期完整的監測數據，以及獲得有用的分析資訊是管理河川很重要的一環。
(Noori, Sabahi, Karbassi, Baghvand, & Taati Zadeh, 2010; Vega, Pardo, Barrado, & Debán, 1998)

河川水質監測現況

台灣的河川水質監測最早可回溯至 1976 年，由 1975 年 9 月成立之台灣省水污染防治所，開始辦理台灣全國水質調查，期望建立河川完整水質變化長期資料庫，以做為水污染防治計畫採取措施之基本依據，至 2002 年改為由行政院環境保護署環境監測與資訊處統籌辦理全國水質監測工作，在河川流域檢測對象包括 54 條流域，資料內容涵蓋水質物化特性及重金屬濃度等監測數值，部分數值每個月量測(如酸鹼值、溶氧、生化需氧量、氨氮等)，部分數值每季量測(如總磷、硝酸氮及金屬濃度等)。本研究分析對象為桃園市老街溪 2002 年至 2016 年間，7 個河川水質監測站紀錄之水質監測資訊。

水質監測數據繁多分析不易問題及解決方法

doi:10.6342/NTU201803135

長期的水質調查監測提供豐富的水體變化資料，眾多水質參數隨時間累積成龐大的資訊量，讓結果解釋或分析更具難度。為了能從大量複雜的資料中萃取出有用資訊，可透過多變量分析方法協助簡化並詮釋資料。(Papaioannou et al., 2010)

本研究使用多變量分析方法中的主成分分析(principal component analysis)及群集分析(Cluster Analysis)方法，主成分分析被廣泛地運用在流域相關的分析研究上，這種分析方法是藉由辨別變數間的相關性，達到減少變數的數量，流域的性質便可以較少的變數表現，以便進行後續水質於時間或空間變化之描述，並可更進一步達到污染源頭辨別。群集分析則多被應用在探索樣本間的相似程度，在沒有任何既定假設的前提下，將樣本特性交由統計方法依其物理上的距離重新定義群組分類，藉此發掘的各樣本間不易被發現的相似特性。本研究分析水質監測數據之目標在於發掘變項間關聯性，進而得出污染排放模式，推測污染來源，配合採樣點間相似度資訊，以更深入了解河川整體變化，期望能為提升水質監測規劃或河川整治政策提供有用資訊。(Olsen, Chappell, & Loftis, 2012; Vega et al., 1998)

目前水質監測規劃不足處及解決方法

現今河川水質監測執行已是每月例行工作，經了解目前河川水質採樣規劃，發現每個水質參數量測頻率不盡相同，有些項目為每月量測(如酸鹼值、溶氧、生化需氧量及氨氮等)，有些項目則為每季量測(如總磷、硝酸氮及金屬濃度等)。本研究鎖定在會對人體產生負面影響，但在監測計畫中屬於每季量測一次的水中金屬濃度值，經彙整各金屬濃度超標頻率，最後選定以金屬銅濃度作為判斷標的，嘗試使用每月量測水質項目，作為判斷當月銅濃度是否超標之分類資訊及依據。

另外，河川污染程度指標(River Pollution Index, RPI)是台灣目前用來判定河川污染的重要指數，RPI 指標是以水中溶氧量、生化需氧量、懸浮固體、與氨氮等 4 項水質參數依特定點數對照表平均而得。但其中生化需氧量測量不易，採樣後需再花 5 天實驗時間以得到河川的污染程度指標，在資訊傳播快速的今日，應考量有

doi:10.6342/NTU201803135

別以往獲得 RPI 指標的方式。

針對以上構想，本研究嘗試以機器學習方法建置分類及回歸模型，以用做判斷金屬濃度是否超標，及評估 RPI 指標之推測數值。模型架構使用類神經網路模型及決策森林模型為主體，類神經網路模型因為效能佳，操作使用具極大彈性，而成為機器學習領域中最常被應用的模型之一，決策樹/決策森林模型則以構造簡單易於解釋等特色，亦廣為被應用於各種領域。目的在於提供新的方法方便需要者能在沒有金屬濃度量測預定之月份得到超標與否的判斷數值，或能更即時的得到河川污染程度指標資訊。(Couto, Vicente, Machado, Abelha, & Neves, 2012; Winkler, Haltmeier, Kleidorfer, Rauch, & Tscheikner-Gratl, 2018)

doi:10.6342/NTU201803135

1.2 研究架構

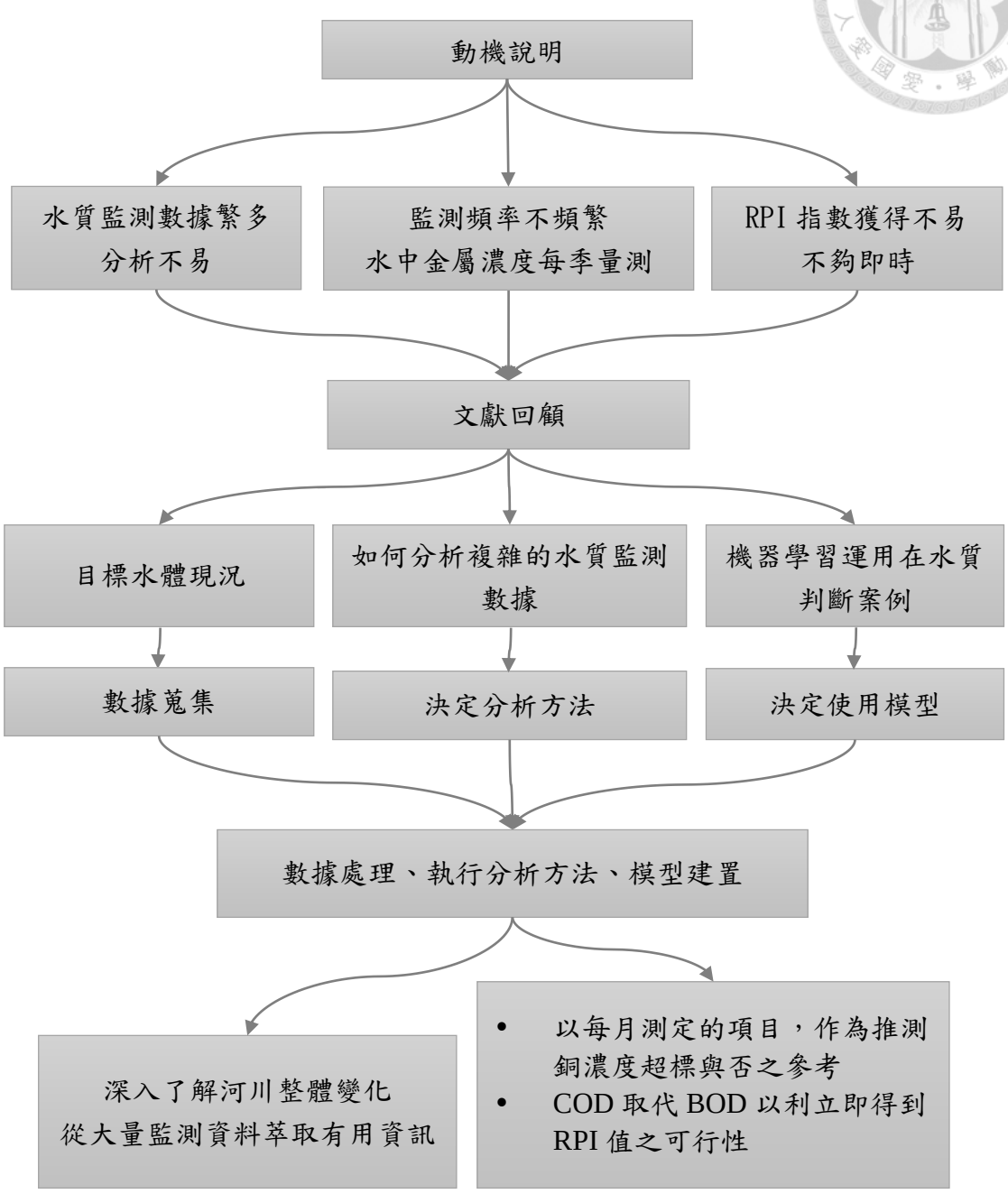


圖 1- 1　研究架構流程圖

doi:10.6342/NTU201803135

第二章　文獻回顧

在進行資料探勘前需先調查目前能蒐集到之資訊，本章 2.1 節包括老街溪歷年水質回顧，資料蒐集重點有河川周遭地理環境，可能污染源、污染現況及目前應變對策等；接著調查如何從複雜又大量的水質監測資料中萃取有用的資訊，2.2 節彙整文獻資料，顯示多變量分析方法常被應用於水質分析領域之案例；最後一部分為調查機器學習方法應用於水質判斷評估的案例，2.3 節介紹目前蒐集到成功使用在類別分類或是數值判別，且與環境工程領域相關的模型案例。

2.1 桃園老街溪背景資訊

老街溪地理位置

老街溪發源於桃園市龍潭區深窩子地區，注入龍潭大池後沿途流經平鎮區、中壢區、大園區等行政區，主要支流河川為大坑缺溪，接近出海口時有田心仔溪匯入，最終於大園區許厝港出海，主流全長約 37 公里，流域面積約 81.59 平方公里。老街溪流域流經地形主要為台地，受潮汐影響不大，許厝港一號橋為老街溪之感潮終點，相關地理位置如下圖**錯誤! 找不到參照來源。**。

doi:10.6342/NTU201803135

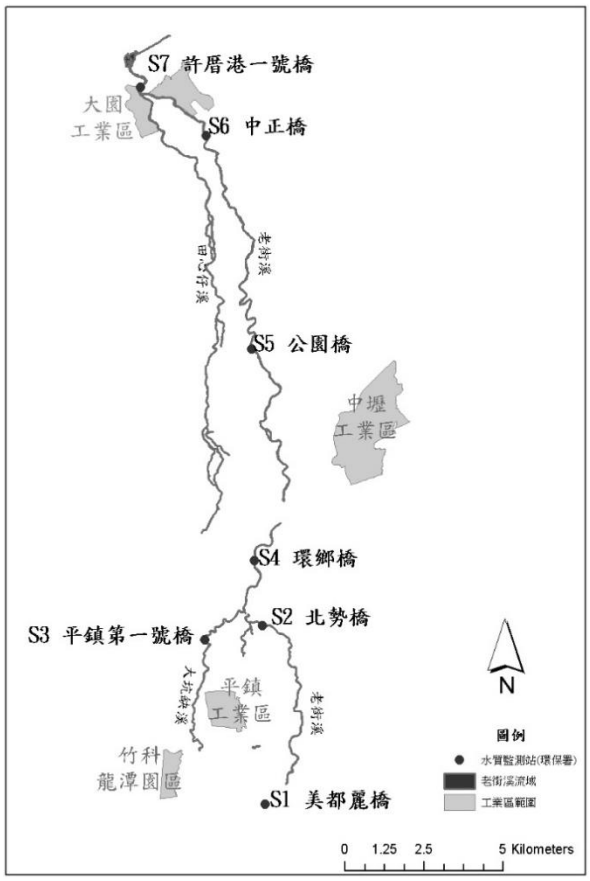


圖 2- 1　老街溪與周遭地理位置

水質指標選擇—河川污染程度指標(River Pollution Index, RPI)及河川水質指數(Water Quality Index, WQI)

河川污染程度指標(RPI)為評估河川污染程度的重要指標，該指標是由生化需氧量(BOD)、溶氧量(DO)、氨氮(NH_3-N)及懸浮固體(SS)之濃度對照點數換算表後，將 4 個參數對應點數相加後平均換算而成，RPI 指標小於 3 者為輕度污染或無污染、RPI 指標在 3 到 6 之間表示中度污染，而 RPI 指標倘大於 6 時則表示該測點為嚴重污染，詳細點數如表 2- 1。

表 2- 1　RPI 指標計算點數對照表

水質\分類	未(稍)受污染	輕度污染	中度污染	嚴重污染
溶氧量 mg/L	DO≧6.5	6.5＞DO≧4.6	4.5≧DO≧2.0	DO＜2.0

doi:10.6342/NTU201803135

水質\分類	未(稍)受污染	輕度污染	中度污染	嚴重污染
生化需氧量 mg/L	$BOD_5 \leqq 3.0$	$3.0 < BOD_5 \leqq 4.9$	$5.0 \leqq BOD_5 \leqq 15.0$	$BOD_5 > 15.0$
懸浮固體 mg/L	$SS \leqq 20.0$	$20.0 < SS \leqq 49.9$	$50.0 \leqq SS \leqq 100$	$SS > 100$
氨氮 mg/L	$NH_3\text{-}N \leqq 0.50$	$0.50 < NH_3\text{-}N \leqq 0.99$	$1.00 \leqq NH_3\text{-}N \leqq 3.00$	$NH_3\text{-}N > 3.00$
點數	1	3	6	10
污染指數積分值(S)	$S \leqq 2.0$	$2.0 < S \leqq 3.0$	$3.1 \leqq S \leqq 6.0$	$S > 6.0$

資料來源：行政院環境保護署全國環境水質監測資訊網

河川水質指數(Water Quality Index, WQI)可涵蓋較多影響河川水質項目，包括溶氧、大腸桿菌群、pH 值、生化需氧量、氨氮、懸浮固體及總磷等，各水質項目有獨立之公式計算污染點數，數值範圍為由 0 至 100。相較之下，PRI 指數判別河川水質參數較少，僅有 DO、BOD、NH_3-N 及 SS 等 4 項參數，且數值範圍小(1-10)，易受水質參數變化影響，相對地 WQI 值範圍較大，能更細緻地評估河川水質改善成效。惟因 WQI 指數計算指標之一的總磷指數，在目前監測計畫中被列為一季測試一次之項目，目前台灣仍以 RPI 指數作為辨別水質污染程度的指標。(桃園縣政府環境保護局, 2015; 張祚楨, 2013)

老街溪歷年水質及污染來源

彙整環境保護署全國環境水質監測資訊網公布老街溪 2002 年至 2016 年監測資料，老街溪流域 7 個測站之 RPI 指標多呈現中度污染，即 RPI 指標落於 3 至 6 間，RPI 指標偏高者為 S3 平鎮第一號橋測站及 S7 許厝港一號橋測站，有較多採樣點的 RPI 指標高於 6，屬於嚴重污染，如下圖 2- 2。

doi:10.6342/NTU201803135

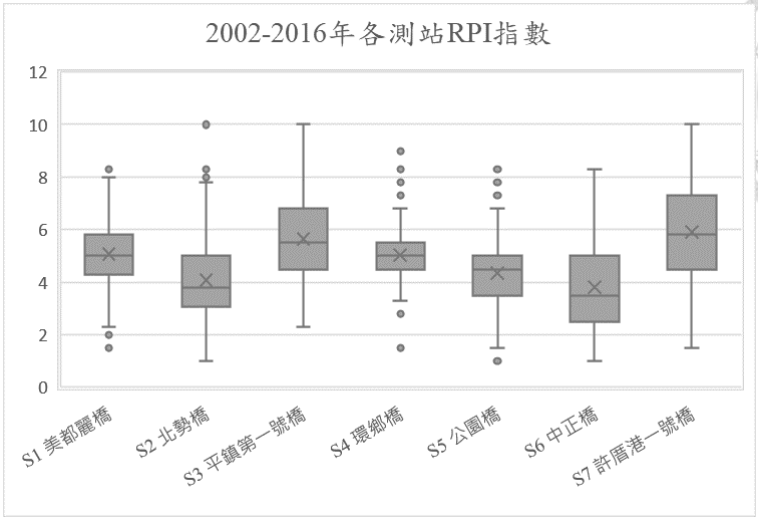


圖 2- 2　老街溪各測站 RPI 指標分布

依據桃園縣政府環保局調查報告，老街溪全河段分為 9 個集污區，分別計算得到區域推估污染物排放量如表 2- 1。影響下游許厝港一號橋河段之主要污染來源為大園一期工業區內專管排放之事業廢水，推估一日排放生化需氧量 3,514 公斤，氨氮 635 公斤，其次為田心仔溪沿岸零星散布之畜牧及住宅所產生之污水。中游受支流平鎮一號橋測站接收大坑缺溪接收之污染物影響，推估一日排放生化需氧量 490 公斤，氨氮 1,459 公斤，污染來源來自上游新竹科學工業園區龍潭園區、平鎮工業區事業廢水、平鎮山子頂都市計畫區生活污水與沿岸之零星畜牧廢水。(桃園縣政府環保局, 2011; 楊于嫻, 2014)

doi:10.6342/NTU201803135

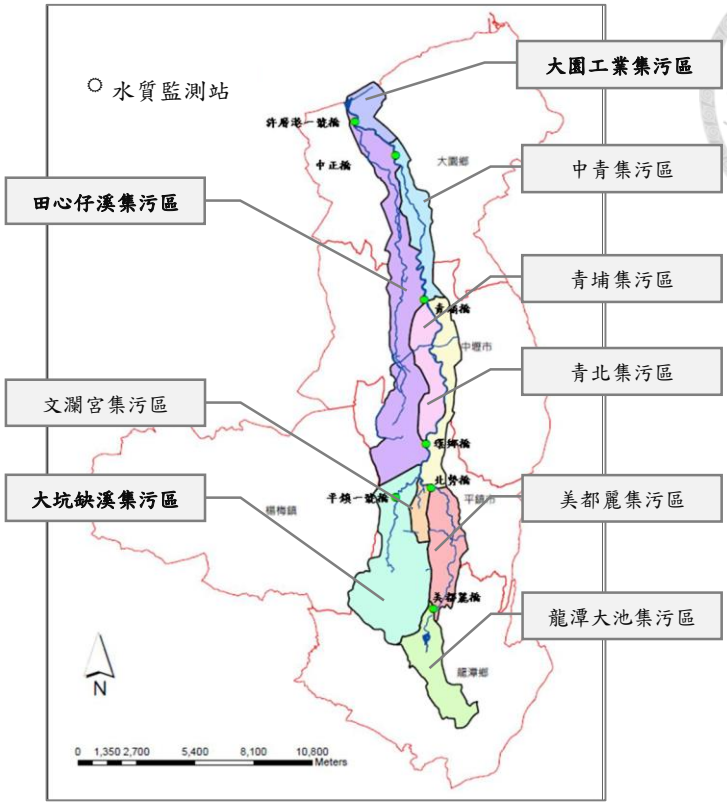


圖 2- 3　老街溪流域集污區分布圖

資料來源：老街溪流域水質改善暨整治策略(桃園縣政府環保局, 2011)

表 2- 2　老街溪集污區污染排放列表

集污區	BOD (kg/day)	SS (kg/day)	NH_3-N (kg/day)
龍潭大池集污區	342.2	362.4	128.9
美都麗集污區	79.7	104.7	41.2
文瀾宮集污區	206.2	349.7	95.9
大坑缺溪集污區	489.9	398.9	1,459.2
青北集污區	738.3	1,435.5	408.1
中青集污區	32.6	210.5	10.9
大園工業區集污區	3,514.3	2,315.1	635.1
田心仔溪集污區	348.3	703.9	133.4
總和	5,751.3	5,880.7	2,912.7

資料來源：老街溪污染總量管制模式評估計畫專案工作計畫

(行政院環境保護署, 2011)

doi:10.6342/NTU201803135

針對老街溪污染狀況之應變措施

一、列管排放源

檢視老街溪水體接收的污染物排放情形，由行政院環境保護署提供之各類水體污染物排放總量資料，統計期間為每半年一次，目前可得最新資料為 2016 年上下半年之排放資料，申報項目包括排水量、生化需氧量(BOD)、化學需氧量(COD)、懸浮固體(SS)及金屬項目等，彙整如下表 2- 3 及表 2- 4。

在 2016 年間，總申報排放化學需氧量(COD)達 109 萬公斤，而金屬排放量部份，排放申報量前五名者為銅(2,661 公斤)、鎳(570 公斤)、鋅(311 公斤)、總鉻(164 公斤)及鉛(85 公斤)，而硒及銀全年總排放申報量則不到 1 公斤。

表 2- 3　老街溪 2016 年水體接收的污染物排放情形

統計期	申報家數	排放總量(公斤)						
		排放水量	BOD	COD	SS	銅	鎳	鋅
上半年	129	7,744,386	88,711	543,995	1,671	1,270	258	214
下半年	88	7,548,227	90,637	549,483	184	1,391	312	97
總計	217	15,292,613	179,348	1,093,479	1,855	2,661	570	311

資料來源：各類水體污染物排放總量(行政院環境保護署, 2018d)

表 2- 4　老街溪 2016 年水體接收的污染物排放情形-續

統計期	排放總量(公斤)										
	總鉻	鉛	硼	鐵	六價鉻	鎘	錳	總汞	砷	銀	硒
上半年	112	36	7.22	22	13	1.63	6.39	0.83	0.34	-	-
下半年	52	49	74	32	1.45	5.61	-	0.36	0.67	0.20	-
總計	164	85	81	54	15	7.23	6.39	1.20	1.01	0.20	-

資料來源：各類水體污染物排放總量(行政院環境保護署, 2018d)

於行政院環境保護署列管污染源資料查詢系統，查詢位於老街溪流域之龍潭

doi:10.6342/NTU201803135

區、平鎮區、中壢區及大園區中，列管類別為水污染類的事業名單。經整理出前 10 大行業別及在各行政區之列管家數如下表 2-5，另外蒐集水中常見金屬污染及可能排放源之產業別如表 2-6。

表 2-5　列管行業別於各行政區統計

行業別	大園區	中壢區	平鎮區	龍潭區	總計
總計	**272**	**622**	**259**	**177**	**1330**
建設營造業	54	246	68	59	427
廢水處理業	8	65	32	8	113
畜牧業	33	14	10	6	63
印刷電路板製造業	7	27	13	3	50
金屬加工處理業	25	11	8	6	50
印染整理業	25	12	6	0	43
化學製品製造業	13	13	4	3	33
食品製造業	6	10	8	3	27
暫無分類	5	15	2	5	27
電子零組件製造業	2	11	11	2	26

資料來源：列管污染源資料查詢系統(行政院環境保護署, 2018c)

表 2-6　水中常見金屬污染物及其污染來源

金屬	污染來源工業別	主要用途及備註
鉻(Cr)	鋼鐵業、電鍍業、皮革業、金屬表面處理業	不鏽鋼添加物、金屬防鏽用
銅(Cu)	電子業、印刷電路板工業、銅電鍍工廠、廢五金回收業	印刷電路板材料、電線材料
鎳(Ni)	鋼鐵工業、電鍍業、化學品製造業	鋼鐵製造添加物、電鍍材料
鋅(Zn)	電鍍業、金屬表面處理業	金屬防鏽用、電鍍材料
砷(As)	化工業、半導體電子業、礦業冶金業、煉焦業	砷為製造半導體之材料、冶煉業之砷污染來源多因礦石材料含砷量過高導致
鎘(Cd)	電鍍業、礦業冶金業、塑膠製造業	塑膠硬化劑、電鍍材料
汞(Hg)	製造業	用於溫度計、日光燈、水銀電池製造

doi:10.6342/NTU201803135

金屬	污染來源工業別	主要用途及備註
鉛(Pb)	製造業、氯鹼工業	用於鉛蓄電池製造、映像管等玻璃製成材料
銀(Ag)	化工業、電子業、醫療業	常被用作化學催化劑，另因其良好導電性,在電子業多用來製造導體。

資料來源：環境品質調查資料空間變異分析之探討(林倩如, 2006)

二、訂定大坑缺溪污染排放總量管制標準及推動污染整治計畫

鑒於老街溪支流大坑缺溪為科學園區龍潭廠、平鎮工業區及龍潭工業區的廢水排放承受水體，因為事業廢(污)水過於集中，超過該段河川可以承受的污染量，該河段之河川污染等級常顯示為嚴重污染。依水污染防治法第 9 條規定，針對需特予保護水體，與地方政府合作訂定總量排放標準，並於 103 年 12 月發布實施。下表 2- 7 為大坑缺溪放流水規定水質項目及限值，既設事業之生化需氧量限值 15 至 25mg/L 及氨氮限值 30 至 75mg/L，相較放流水標準中規定事業廢水的生化需氧量限值 50 至 150mg/L 及氨氮限值 10 至 150mg/L 更為嚴格。

表 2- 7　桃園市大坑缺溪放流水規定水質項目及限值

項目		限值 (mg/L)	適用對象
生化需氧量	最大值	10	新設者
	最大值	25	既設者
	七日平均值	15	既設者
氨氮		10	新設者
		30、75 (依行業別而不同)	既設者

資料來源：桃園市大坑缺溪放流水標準

除了支流大坑缺溪已經訂定總量管制標準外，對於老街溪整體治理，桃園市政府於 2008 至 2015 年間完成了老街溪整體規劃、礫間曝氣工程等整治行動，詳如下表 2- 8。

doi:10.6342/NTU201803135

表 2- 8　老街溪污染整治歷程

年度	事件	說明
2008	推動「老街溪及南崁溪流域污染整治調查及水岸活化整體規劃計畫」	掌握南崁溪及老街溪各污染源及重點河段水體水質狀況
2011	老街溪開蓋工程完工	拆除老街溪中壢市區河段上方加蓋之商場，並整治河岸空間
2013	「新勢公園礫間接觸曝氣氧化工程」完工	截流老街溪上游環鄉橋至延平路老街溪橋人口密集段之晴天排水，處理後放流回老街溪
2014	公告「桃園縣大坑缺溪放流水標準」 ※於 2015 年 12 月 3 日重新公告「桃園市大坑坎溪放流水標準」	針對以大坑缺溪作為中承受水體的列管事業及污水下水道系統，訂定生化需氧量及氨氮排放標準
2015	老街溪整治計畫後續工程完工 華映友達公司達成廢水零排放目標	華映友達公司達成廢水零排放目標，位於老街溪龍潭段之排放專管於年底進行封管儀式

資料來源：水質保護網(行政院環境保護署, 2018a)

doi:10.6342/NTU201803135

2.2 多變量分析案例

多變量分析方法可應用於分析水質資料在時間及空間的變化趨勢，常見方法有主成分分析(Principal component Analysis, PCA)、群集分析(Cluster Analysis, CA)、因素分析(Factor Analysis, FA)和判別分析(Discrimination Analysis, DA) 等，在不同的提問及視研究目的需要，產出的結果不同，所選用的方法也不同，各方法的目的及應用彙整，如下表 2- 9 (Bierman, Lewis, Ostendorf, & Tanner, 2011)：

表 2- 9　多變量分析方法及其適用類別

方法	目的	應用及範例
主成分分析	數據探索及減少變項	從數量多的變項中尋找具相似變化趨勢者，組合產生彼此無關且數量較少的主成份，以利後續分析
群集分析	數據探索及驗證	將具有相似特性的樣本分類至同一群
因素分析	解釋現象	找出影響水質變項的隱藏因素
辨別分析	解釋現象及預測	找出目標群組間決定性的差異變項

在水質分析的應用上，多變量分析方法可評估水質變化情形，並確認污染源，且可更進一步提出改善水質資料管理對策，包括最佳化水質監測計畫及規劃未來污染防治(Papaioannou et al., 2010)。以下介紹以多變量分析評估水質之案例，方法以主成分分析、因素分析及群集分析為主，評估對象包括地下水、地面水及海域水質等。

2.2.1 主成分分析/因素分析

發展與意義

主成分分析(principal component analysis, PCA) 是由 Pearson 於 1901 年提出，後來由 Hotelling 再加以發展的統計方法(林清山, 1991)。於主成分分析中，可將 m 個變數加以轉化，使所得線性組合而得的 P 個主成分變異數變為最大(m>P)，成份間彼此無關，利用 PCA 使變項變為無關的數個成份分數，以利後續統計分析。PCA 被廣泛地運用在流域相關的分析研究上，這種分析方法是藉由辨別變數間的相關

doi:10.6342/NTU201803135

性，達到減少變數的數量，便可以較少的變數表現整體河川水質特性，以便進行後續對流域水質在空間上的樣態描述，或源頭污染源的辨別。(Olsen et al., 2012)

因素分析方法是 Spearman 所創用，為主成分分析的推廣與發展，同樣可以用來降低變數的維度，雖然在功用有類似的地方，但兩種方法概念不同，於抽取因素時，主成分分析及因素分析(Factor Analysis，FA)的過程是相同的，但不同處在於相關係數矩陣對角線上的數值，於 PCA 時對角線數值為 1，而在 FA 的相關係數矩陣上的對角線微小於 1 之數值，此矩陣稱為「縮減成相關係數矩陣」或「調整相關係數矩陣」(林清山, 1991)，採用此方法的研究者興趣在探討觀察變項間是否能以數個潛在變項(latent variable)來代表觀察變項之間的關係。(傅粹馨, 2002)

應用案例

Vega 等人收集西班牙境內 Pisuerga 河 3 個測站為期 2 年半的水質資料，試圖判辨河水水質變化的原因。利用 PCA 減少維度的功能，將代表水質的 22 個物化性質，轉換為 3 個有意義的主成分，主成分 1 的重要成員為礦物質(硬度、鈣、氯離子及硫酸根等)，主成分 2 的重要成員為生化需氧量、化學需氧量及酸鹼值等人為污染有關成分，主成分 3 為溫度及溶氧。可解釋整體數據的 67.8 %變化。並以 ANOVA 分析發現，主成分 1 與時間變化相關，主成分 2 隨採樣點變化。(Vega et al., 1998)

Roger 等人針對美國伊利諾州的伊利諾河水監測數值進行主成分分析，想要評估環境樣態，並找出造成水質污染的源頭。該研究不只收集地表河川及底泥樣本，還有周圍田野逕流、廢水處理廠放流水甚至畜牧廢水等，在 2 年半間在 279 個點位上採集 621 個樣本，測量 26 個物化特性。經過 PCA 程序後得到第 1 個主成分對於總資料的解釋力有 38 %，前 3 個主要因素對於總資料集解釋力累積達 56 %。由第 1 及第 2 主成分的組成可看出其所代表的特性，第 1 主成分中最重要的因素

doi:10.6342/NTU201803135

有鐵、大腸桿菌、糞生大腸菌、腸球菌、總有機碳、總磷、銅及鋅等，這些物質與禽類養殖廠的排放污水有相似的組成；第 2 主成分的重要因素有氯、鈉、硫酸根離子等，這些項目常出現於污水處理廠的放流水中。將各資料點位轉換成新的主成分數值後，將之散布在由第 1 及第 2 主成分為資料軸架構成的 2 維平面圖上，依點位所在的位置，可分辨出其物化性質組成較接近畜牧廢水或是污水廠排放的廢水，同時檢視污染源對於水質的影響。(Olsen et al., 2012)

Singh 等人利用不同的多變量分析方法於印度 Gomti 河川水質分析上，嘗試將不同的多變量分析方法運用在北部印度 Gomiti 河的監測上，目的是希望能優化監測網路及監測參數的數量，在不減損有用資訊的傳達下，減少監測參數的數目，以節省監測經費。該研究蒐集 5 年 8 個點位 24 個參數的資料，共觀測到 1 萬 7,790 筆數據。針對如此複雜的資料集，主成分分析協助識別影響水質的重要因素，萃取出的主成分有 6 個，可以反映出 71%資料結構變化；其後該團隊針對 PCA 定義出的 6 個主成分再以 Varimax rotation 方法做轉軸，得到了新的 6 個主要因素，前 3 項分別為礦物質因素(EC, Cl, K, Na)，人為污染因素(BOD, COD, DO, pH)，土壤組成因素(Ca, 鈣離子硬度)，可解釋的變異量分別為 17.6%, 16.2%, 12.0%。然而主成分分析/因素分析並無達到顯著減少參數的效果，6 個主成分包含了 14 個參數，代表仍需 14 個參數才能解釋資料集 71%變化。(Singh, Malik, Mohan, & Sinha, 2004)

2.2.2 群集分析

群集分析屬於非監督式樣態識別的技術，分為階層式(Hierarchical cluster)及非階層式(Non-hierarchical cluster)兩類，目的為協助觀察值的分類，在水質分析上較常被應用者為階層式群集分析，分析結果通常以樹狀圖表示，分類於同一群組表現整條河川中不同空間採樣點具有相似性，群組間樹枝的距離則表示差異的程度。

在印度 Gomiti 河的研究中，8 個採樣點數據經過 z-scale 標準化後以階層式群

doi:10.6342/NTU201803135

集分析，分成了 3 個顯著的群組，群組內成員具有相似的特性且來源的環境背景相似，群集 1 的點位均來自河川上游，為低度污染地區，群集 2 的點位採樣於接收污水廠出流及兩條高度污染支流會合之地區，為高度污染區域，群集 3 點位分布於河川下游，屬中度污染地區。該研究將群集分析之結果應用在規劃採樣點策略上，倘需要快速地知道 Gomiti 河川水質狀況，從每個群組各挑一個點位(總共挑 3 個)分析得出結果，應能代表分析完整監測網路 8 個點位的結果。這也證明群集分析技術可以用來優化採樣點位的設置策略，以較少的量點位得到同樣具代表性的測量結果。

Vega 等人的研究同樣運用階層式凝聚群集分析方法，配合 Ward's 距離計算方法(Hierarchical agglomerative clustering by the Ward's method)，將 3 處地點在 3 年間所得監測數據分為 3 個集合，第 1 個集合為人為污染與礦物質都低的組合，水質最好的點位也在這一組，第 2 個集合為人為污染物與礦物質都高的群組，水質最差的幾個觀測點位也出現在此組，另一個集合為人為污染物低但礦物質高的群組，作者亦觀察到乾季量測的點位多也分布於第 2 及 3 組。

多變量分析是非常有用的工具，可以從大量而複雜的監測資料中，萃取有價值的水體變化資訊。主成分及因素分析主要目標為針對水質變項做分類，簡化資料使分析者能更方便直觀地發掘其中隱含的關聯性，並進一步確認污染物來源；群集分析主要功能為針對樣本做分群，透過群聚的結果發現觀察值間內在結構的相似程度，並推測發掘觀察值間隱含的共同性。多變量分析的結果可用做研提改善水質監測管理規劃之對策，具體來說包括最佳化水質監測規劃及作為未來污染防治政策參考資訊等。(Nosrati & Van Den Eeckhaut, 2012; Papaioannou et al., 2010; 王嶽斌, 2015)

doi:10.6342/NTU201803135

2.3 機器學習應用於水質評估

隨著計算機科技的發展，數值模式經常被用來模擬水流或水質的變化，傳統的水質模式將重點放在解決特定問題的演算法上，發展至今水質模式的可行性及和技術已相當成熟，數值分析方法可分為有限元素方法、差分方法、邊界元素法等，方程式的架構可以是一階、二階或是更高的階層。為了得到最準確的預測數值，選擇適合的模型和設定合適邊界條件是首要任務，然而這必須對模型建構細節具有相當程度的了解，造成了模式建立者與使用者間的隔閡，反映出傳統水質模式對使用者不夠友善，後續數據解讀困難等問題。

水質模式的發展經過多個世代的演變，在第三代模型以前，只有熟悉數學模型的特定使用者可以執行，例如使用複雜的二維或三維有限差分模型(two-dimensional or three-dimensional finite difference numerical models) 模擬潮汐流或特定的水質現象。第四代模型藉由較智慧且人性化的前端功能(智慧選單功能等)，讓使用者可以依據現況調整模式，使水質模式可用於更廣泛的環境狀況，現在則到了第五代模式，這代模式融合了人工智慧及水質動力資訊於單一操作系統，提供輔助予沒有經驗的使用者，使模式的使用對入門者更為友善。(K. Chau & Chen, 2001; K. W. Chau, 2006)

本研究嘗試利用機器學習方法以基本水質指標判斷水中金屬銅濃度之超標與否，及以 COD 參數取代 BOD 參數判斷河川污染指標之數值，針對兩項課題搜尋類似的研究如下所示。

2.3.1 利用機器學習方法進行分類

Couto 等人利用類神經網路及決策樹方法，以葡萄牙 Odivelas 水庫 2001 至 2010 年計 10 年間水質監測資料為輸入，建置水質污染等級分類模型，將水質分為 A 至 E 五個等級，並比較 2 種方法的正確率，結果發現分類正確率均可達 97.4%。

doi:10.6342/NTU201803135

此時類神經網路使用 4 個水質參數(生化需氧量、溶氧量、總懸浮固體及氧化能力)，而決策樹需使用上述 4 個水質參數加上導電度及溫度，共 6 個參數才能達到同樣的正確率，不過總結來說決策樹和類神經網路對於水質分類均可以有不錯的表現。(Couto et al., 2012)

Winkler 等人利用決策樹模型來預測管線損壞機率，有鑑於使管線損壞的物理機制非常複雜，現今也沒有完全被掌握，所以該研究藉由歷年管線故障的記錄，以決策樹方法建立管線損壞預測模型，研究地點為奧地利的一處中型城市，約有 9 萬 5 千位居民，管線總長達 851 公里，管線損壞紀錄由 1983 年開始。因為數據量龐大，彙整資訊經過前處理後再進行抽樣，最後取出 3743 筆數據進行後續建模分析，使用的參數有年份、型式、直徑、壓力、長度、材質、閥數、損壞記錄…等 12 個項目，最終獲得模型之正確率達 0.96，曲線下面積達 0.93，這個模型正實際被運用於中型城市，預測 5 至 10 年內管線故障損壞的機率。(Winkler et al., 2018)

2.3.2 利用機器學習方法進行數值判斷

Chou 等人有感於傳統用於計算卡爾森指數的水質參數如總磷、葉綠素 a 及沙奇盤深度等不易取得，爰嘗試用其他較易取得的參數如溫度、溶氧量、懸浮固體、化學需氧量及氨氮等指標，以機器學習方法製作輔助預測卡爾森指數之模式，該團隊蒐集了台灣 20 個水庫在 1995 至 2016 年間的監測數據，使用了類神經網路、支持向量機、分類回歸樹及線性回歸等方法，經排列組合及比較不同的模型優化方法後，發現類神經網路有最好的表現，總體來說其根均方誤差(RMSE)，平均絕對誤差(MAE)及絕對誤差百分比(MAPE)分別為 3.941、3.131 及 6.786%，該研究也得出，以新參數配合類神經網路模型來計算卡爾森指數是可行的，對有特別需求的使用者，此方法降低了實驗及測量的複雜度，並也可以達到高的預測準確率。 (Chou, Ho, & Hoang, 2018)

doi:10.6342/NTU201803135

Heddam 和 kisi 亦嘗試利用機器學習方法預測溶氧的濃度，該團隊使用方法為極限學習機(Extreme Machine Learning, EML)方法，此方法是類神經網路其中一種架構。輸入參數有 2 組，一組為水溫、導電度、酸鹼值及濁度等水質參數，另一組為時間變項，如年月日時間等參數，經比較後發現，使用水質參數做為預測溶氧時，配合 Optimally Pruned Extreme Learning Machine (OP-ELM)模型會有最佳的預測效果，其根均方誤差(RMSE)為 0.172，平均絕對誤差(MAE)為 0.116；而使用時間參數做為輸入值時，預測效果最佳的模型為 Extreme Learning Machine with Radial Basis Activation Function (R-ELM)，其根平方誤差(RMSE)為 0.186，平均絕對誤差(MAE)為 0.243，雖然以時間參數作為預測溶氧的效果沒有以水質參數做為預測戰術的效果好，但是在水質參數缺失的狀況下，時間參數也可以替代做為預測之輸入值。(Heddam & Kisi, 2017)

由上述文獻回顧結果可得知，預測水質數值及指標分類的研究都是有成功案例，可以嘗試運用於本研究的目標，可以發現無論是用於分類或是數值預測，類神經網路是非常受歡迎的機器學習方法，後續將選用類神經網路及文獻中提到過的決策樹/決策森林模型，做為建立水中金屬濃度超標與否判斷，以及河川污染程度指標評估模型之方法。

doi:10.6342/NTU201803135

第三章　分析材料與方法

3.1 研究方法流程

圖 3- 1 為第三章分析材料與方法之撰寫架構，內容分為 3.2 資料蒐集及前處理，3.3 資料分析方法，內容包括描述統計及多變量分析方法之運算說明，3.4 水質判斷模型建置小節則介紹模型建置流程及評估模型效果方法。

圖 3- 1　研究方法流程圖

doi:10.6342/NTU201803135

3.2 資料蒐集及前處理

3.2.1 資料蒐集

資料來源及研究分析材料

台灣的河川水質監測最早可回溯至 1976 年，由 1975 年 9 月成立之台灣省水污染防治所，開始辦理台灣全國水質調查，期望能建立河川完整水質變化長期資料庫，以做為水污染防治計畫採取措施之基本依據，至 2002 年改為由行政院環境保護署環境監測與資訊處統籌辦理全國水質監測工作，水質監測成果均公布於行政院環境保護署全國水質監測資訊網，該網站提供台灣 54 條河川流域監測資訊，另有水庫、海岸及地下水水質相關資料。(行政院環境保護署, 2016)

本研究擷取桃園市老街溪流域 2002 至 2016 年(總計 15 年)間河川水質監測資訊作為分析材料。

測站分布

行政院環境保護署目前在老街溪流域設有 7 處水質測站，6 站位於老街溪，1 站位於支流大坑缺溪（S3 平鎮第一號橋測站），各測站資訊如表 3- 1，地理相關位置如圖 3- 2。

表 3- 1　老街溪河川水質測站資訊

測站名	設站年度	行政區	水體分類	經度	緯度
許厝港一號橋	1979	大園區	丙	121.1778230	25.0775620
中正橋	1976	大園區	丙	121.1963970	25.0637580
公園橋上游 (原為青埔橋)	1976	中壢區	丙	121.2093650	25.0033540
環鄉橋 (原為宋屋)	1976	平鎮區	丙	121.2100950	24.9433770
北勢橋	1976	平鎮區	丙	121.2124420	24.9248290
平鎮第一號橋	1988	平鎮區	丙	121.1961610	24.9209930
美都麗橋	1988	龍潭區	丙	121.2132910	24.8742700

資料來源：全國環境水質監測資訊網(行政院環境保護署, 2018b)

doi:10.6342/NTU201803135

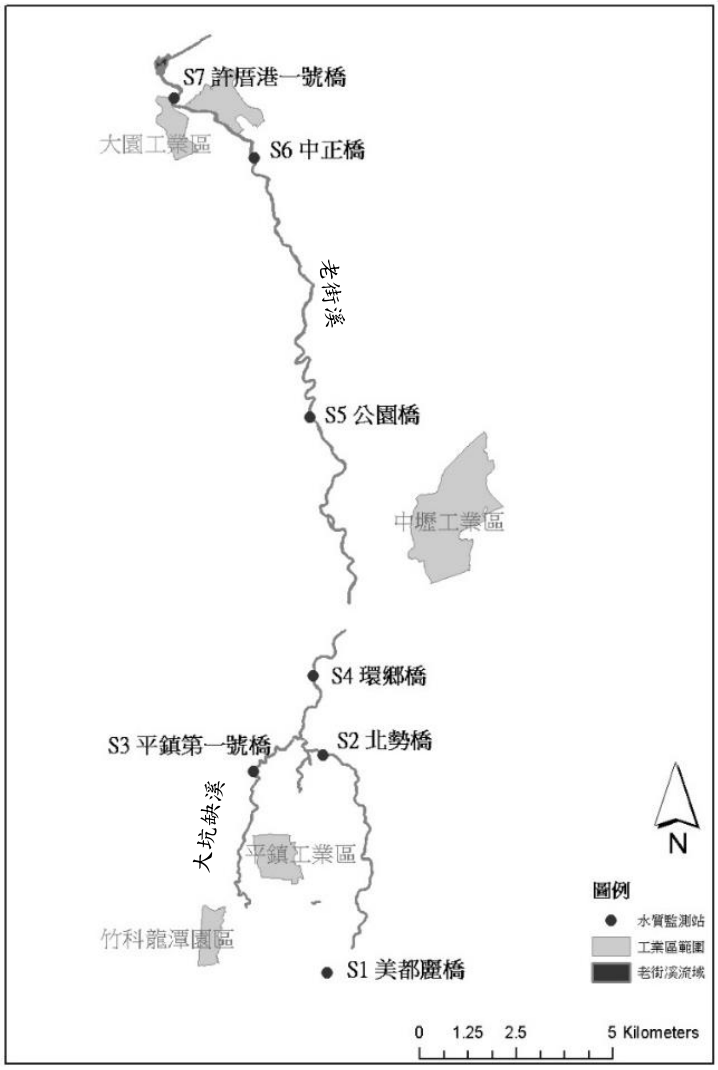


圖 3- 2　老街溪流域河川水質測站地理位置圖

水質監測項目之資料特性

清點目標數據資料集，2002 至 2016 年間老街溪 7 個測站、採集樣本總數 1,260 個，表列測量變項有 32 個，但其中部分變項非每個月都有量測數值，經整理數據的測量頻率樣態如下，本研究分析對象將以季測及月測的項目為主。

一、 月測變項：RPI、氣溫、水溫、酸鹼值(pH)、導電度(EC)、溶氧量(包括滴定法、電極法及溶氧百分比)、生化需氧量(BOD)、化學需氧量(COD)、懸浮固體(SS)、大腸桿菌群(coliform)、氨氮(NH_3-N)，計有 13 項。

二、 季測變項：總磷(TP)、硝酸氮(NO_3-N)、鎘(Cd)、鉛(Pb)、六價鉻(Cr^{6+})、砷(As)、

doi:10.6342/NTU201803135

汞(Hg)、銅(Cu)、鋅(Zn)、錳(Mn)、銀(Ag)，計有 11 項。

三、 (半)年測或不定期測量變項：濁度(turbidity)、鉻(Cr)、總氮(TN)、總凱氏氮(TKN)、硒(Se)、總有機碳(TOC)、亞硝酸氮(NO_2-N)，計 7 項。

四、 感潮河段加測變項：氯鹽(Cl)，計 1 項。

常規每月量測的變項正常狀況應有 1,260 筆資料點，季量測變項應有 413 筆資料點，下表 3-2 彙整各變項詳細資料數據量、低於偵測極限量(比率)及遺失值比率等，可作為後續資料結果判讀之參考。表 3- 3 則提供水質變項檢測方式及變項英文簡寫及單位等資訊，後續提到水質變項將多以英文簡寫表示。

doi:10.6342/NTU201803135

表 3- 2　水質測項及數據缺失或低於偵測極限狀況

水質測項	2002 - 2016 年、7 測站 樣本數 1260 個、變項 32 個、21,194 觀測值			
	(1) 數據量 (個/單一測項)	(2) 低於偵測極限量 (個)	低於偵測極限比率 (2)/(1)	缺失數據量 (個/單一測項)
RPI	1260	0	0%	0
氣溫	1008	0	0%	0
水溫、pH、EC、BOD、COD	1260	0	0%	0
DO_titra	756	0	0%	504
DO	1092	0	0%	168
SS	1260	3	0%	0
coliform	1260	4	0%	0
NH_3-N	1260	0	0%	0
Cl	148	0	0%	1112
TP	404	0	0%	856
TOC	364	0	0%	896
NO_3-N	413	0	0%	847
NO_2-N	336	0	0%	924
Cd	431	385	89%	829
Pb	413	232	56%	847
Cr^{6+}	413	407	99%	847
As	413	9	2%	847
Hg	413	399	97%	847
Cu	413	3	1%	847
Zn	413	0	0%	847
Mn	413	1	0%	847
Ag	413	382	92%	847
Se	336	334	99%	924
TN	61	0	0%	1199
TKN	140	0	0%	1120
Turbidity、Cr	0	0	-	1260

doi:10.6342/NTU201803135

表 3- 3　水質測項資訊及其量測方法

水質測項 (英文簡寫)	單位	測量方法編號	測量方法名稱
RPI	無	(計算)	無
氣溫	℃	無	無
水溫	℃	NIEA W217.51A	溫度計法
pH 值 (pH)	無	NIEA W424.52A	電極法
導電度 (EC)	μmho/cm	NIEA W203.51B	導電度計法
溶氧電極法 (DO)	mg/L	NIEA W455.52C	電極法
溶氧滴定法 (DO_titra)	mg/L	NIEA W421.54C	疊氮化物修正法
生化需氧量 (BOD)	mg/L	NIEA W510.55B	五日恆溫培養法
化學需氧量 (COD)	mg/L	NIEA W515.54A NIEA W516.55A	重鉻酸鉀迴流法 高鹵離子重鉻酸鉀迴流法
懸浮固體 (SS)	mg/L	NIEA W210.58A	水中總溶解固體及懸浮固體檢測方法─103℃～105℃乾燥
大腸桿菌群 (Coliform)	CFU/100 mL	NIEA E202.55B	濾膜法
氨氮 (NH_3-N)	mg/L	NIEA W448.51B NIEA W437.52C	靛酚比色法 靛酚法
氯鹽(Cl)	mg/L	NIEA W407.51C	硝酸銀滴定法
總磷(TP)	mg/L	NIEA W427.53B	分光光度計/維生素丙比色法
總有機碳 (TOC)	mg/L	NIEA W532.52C	過氧焦硫酸鹽加熱氧化/紅外線測定法
硝酸鹽氮 (NO_3-N)	mg/L	NIEA W436.52C NIEA W415.53B	鎘還原流動注入分析法 離子層析法
亞硝酸鹽氮 (NO_2-N)	mg/L	NIEA W436.52C NIEA W418.53C	鎘還原流動注入分析法 分光光度計法
鎘(Cd) 鉛(Pb) 銅(Cu)	mg/L	NIEA W308.22B NIEA W311.53C NIEA W313.53B	鉗合離子交換樹脂濃縮法 感應耦合電漿原子發射光譜法 感應耦合電漿質譜法
砷(As)	mg/L	NIEA W313.53B NIEA W434.54B NIEA W435.53B	感應耦合電漿質譜法 自動化連續流動式氫化物原子吸收光譜法 批次式氫化物原子吸收光譜法
汞(Hg)	mg/L	NIEA W330.52A	冷蒸氣原子吸收光譜法
六價鉻(Cr^{6+})	mg/L	NIEA W320.52A	比色法
錳(Mn)	mg/L	NIEA W308.22B NIEA W311.53C	鉗合離子交換樹脂濃縮法

doi:10.6342/NTU201803135

水質測項 (英文簡寫)	單位	測量方法編號	測量方法名稱
鋅(Zn)		NIEA W313.53B	感應耦合電漿原子發射光譜法 感應耦合電漿質譜法
銀(Ag)	mg/L	NIEA W311.53C NIEA W313.53B	感應耦合電漿原子發射光譜法 感應耦合電漿質譜法
硒(Se)	mg/L	NIEA W341.51B NIEA W303.51A NIEA W340.51A	自動化連續流動式氫化物原子吸收光譜法 石墨爐式原子吸收光譜法 氫化硒原子吸收光譜法
濁度	NTU	NIEA W219.50T	濁度計法
鉻(Cr)	mg/L	NIEA W306.50A	火焰式原子吸收光譜法
總氮(TN)	mg/L	(計算)	無
總凱氏氮 (TKN)	mg/L	NIEA W420.51B	分光光度計法

資料來源：台灣河川水質年報、環境水質年報

(台灣省水污染防治所, 1975; 行政院環境保護署, 2016)

doi:10.6342/NTU201803135

3.2.2 資料前處理

為利後續分析，每列樣本中不能有缺失值或小於偵測極限備註等非數值資料點，加上因為水質監測數值之單位及數值大小不一，因此需經過資料前處理，方能進行後續之多變量分析或建置模型，資料前處理分為缺失值及小於偵測極限數值處理及數值標準化等步驟。

一、缺失值處理

最常見的處理方法是刪除存在缺失值的樣本，其缺點是會損失有用的資訊；也可以使用替代的方法，以該測項的平均值替代缺失值；另外還有更複雜的方式，例如以回歸或是迭代的方法估算缺失值。本研究使用刪除法，移除測項有缺失的樣本。

二、小於偵測極限值處理

小於偵測極限數值可以零值取代、以偵測極限值本身取代，或以 1/2 偵測極限值取代等 3 種方法。本研究部分金屬測項之低於偵測極限值之資料點數超過總樣本數之 9 成，為區別低於偵測極限數值，經考量後採用以 1/2 偵測極限值取代之處理方式。(Olsen et al., 2012)

三、數值標準化

標準化動作可以將數值過小的變項影響放大，並減少大數值的變項影響，平衡範圍過小或過大數值對於分析的影響，更可以減少變數不同單位狀況對於分析的影響，使資料庫的數據成為無因次的數值。(Singh et al., 2004)

群集分析和主成分分析方法均須使用標準化後的數值，本研究透過 Z-score 方法標準化數據，以避免因不同的參數其數據範圍大不相同，而造成分析上的誤差。Z-score 標準化公式如下：

$$z = \frac{x - \mu}{\sigma}$$

x 為需要被標準化的原始分數

μ 為母體平均值

σ 為母體的標準差，且 σ 不為 0

doi:10.6342/NTU201803135

3.3 資料分析方法

資料分析段落分為單變項的描述及多變量統計分析兩部分。3.3.1 介紹描述性統計常用之盒方圖，解釋構造及其代表意義；再以多變量分析對河水的現況詳加探索，3.3.2 小節介紹多變量分析方法中的主成分分析(Principal components Analysis, PCA)及因素分析(Factor Analysis, FA)之運作方式、數學原理及分析流程，3.3.3 小節介紹群集分析方法。

3.3.1 描述統計(盒方圖)

盒方圖(Boxplot)，又稱為盒鬚圖或箱型圖，是一種可用來表現數據分布狀況的統計圖，組成有盒子本體、延伸出的最大最小值線條及離群點構成，如圖 3- 3。

盒子本體的上下邊界分別為第 1 四分位數(Q_1)及第 3 四分位數(Q_3)，盒子的長度為四分位間距($Q_3 - Q_1 = \Delta Q$) ，盒子中間的橫線代表的第 2 四分位數，也就是中位數，十字或交叉記號為平均數。最大值及最小值應位於特定區間，最大值區間為 $Q_3+1.5\Delta Q$，最小值區間為 $Q_1-1.5\Delta Q$，倘超過上開區間的數值，則將被視為離群值(outliner)，以資料點方式呈現。(Pagano & Gauvreau, 2018)

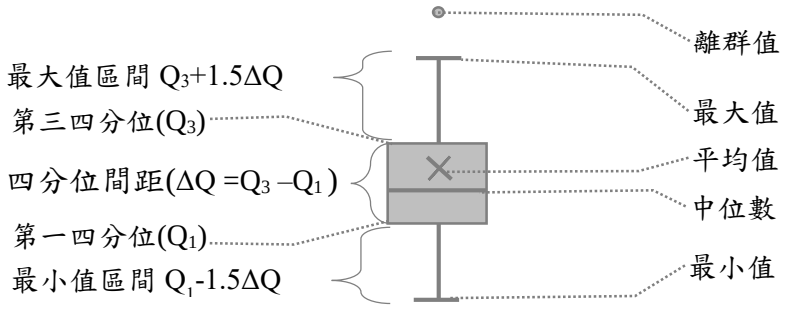


圖 3- 3　盒方圖結構代表統計意義示意圖

3.3.2 主成分分析/因素分析

主成分分析及因素分析方法(Principal component Analysis/Factor Anlysis)主要目的在於以簡潔、精確的方法探討資料集中多個變項間是否能以少數的潛在因素解釋，用潛在的因素來描述眾多變項之間的交互關係，並解釋變項間的關聯性以發

doi:10.6342/NTU201803135

掘影響水質變化的隱藏因素。本研究參照 Singh 等人的分析步驟，先使用主成分分析方法萃取主成分，將變項轉變為數量較少的主成分(Principal components, PCs)，再藉由轉軸方法，進一步減少主成分分析結果中不顯著的變項，以最終得到的變項因素(Varifactors, VFs)解釋變項間關係。(Shrestha & Kazama, 2007; Singh et al., 2004)

主成分分析運作概念

主成分分析技術在於從原變數的共變數矩陣或相關矩陣中，萃取特徵值或特徵向量，目標是將 p 個 X 變項給予不同的加權(w)，以組合成 q 個(q<p)新的主成分變項 PC，即主成分為變項的線性組合表現(正昌, 2005)，常用來表示主成分的公式如下：

$$PC_1 = w_{11}X_1 + w_{12}X_2 + w_{13}X_3 + \cdots + w_{1p}X_p$$

$$PC_2 = w_{21}X_1 + w_{22}X_2 + w_{23}X_3 + \cdots + w_{2p}X_p$$

...

$$PC_q = w_{q1}X_1 + w_{q2}X_2 + w_{q3}X_3 + \cdots + w_{qp}X_p$$

主成分分析步驟

1. 將原始變數轉變為相關係數矩陣 R：

$$\mathrm{R}=\begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{bmatrix}$$

$$\mathrm{r} = \frac{[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})]/(n-1)}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(n-1)}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{(n-1)}}}$$

2. 以相關矩陣 R 進行分析，在 ω'ω=1 的條件下，求 ω'Rω 的極大值。

doi:10.6342/NTU201803135

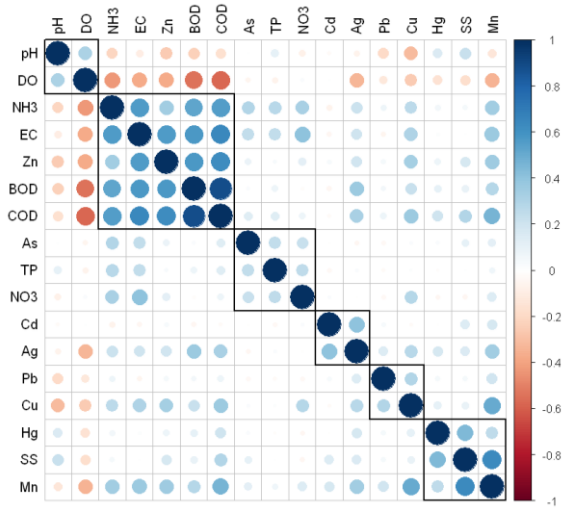


圖 3- 4　相關係數矩陣示意圖

(藍點為正相關、紅點為負相關、圓點大小代表相關程度高低)

3. 承上公式及相關矩陣 R 可解出特徵向量 ω，特徵向量為 X 變項之加權係數，帶入主成分計算公式，計算主成分之變異數為特徵值 δ。

$$\mathrm{Var}\,(PC_q) = \delta_q$$

4. 各主成分解釋的比例為：

$$\frac{\delta_q}{trR} = \frac{\delta_q}{p}$$

5. 確定主成分數目，通常可依據以下原則：

(1) 依 Kaiser's 弱下限法(weakest lower bound)，留特徵值 δ 大於 1 之主成分。

(2) 採用陡坡法(Scree test)，依據因素變異量遞減的斜率來決定。

因素分析轉軸步驟及結果解釋

1. 以主成分分析方法確定的主成分數目為基礎，進行因素轉軸，使因素的意義更加明顯清晰。

2. 選擇轉軸方式：因素轉軸有兩類，分別為正交轉軸(orthogonal rotation)和斜交轉軸(oblique rotation)，依據 Olsen 等人彙整 24 篇關於水質主成分分析的文獻，大部分研究使用轉軸法為正交轉軸的最大變異方法(Varimax)，本研究亦以正交轉軸最大變異方法作為轉軸法。

doi:10.6342/NTU201803135

3. 結果判讀：

Liu et al.等人將變項對於因素的負荷分類為強、中及弱三種等級，當負荷值之絕對值大於 0.75 時表示該水質變項與因素間具有強烈相關性，當負荷值之絕對值在 0.5 至 0.75 間代表具有相關性中等，若低於 0.5 則表示因素與變項間相關性弱。(Liu, Lin, & Kuo, 2003)

本研究取相關性中等至強烈的變項(及負荷值大於 0.5 者)，作為萃取出因素中相關的變項成員，並觀察各個因素的成員組成邏輯及原因，進而以共通性命名。

取樣適切性量表數(Kaiser-Meyer-Olkin Measure of Sampling Adequacy, KMO)

所蒐集的水質樣本是否適合進行主成分/因素分析，可從 KMO 指標(Kaiser-Meyer-Olkin measure)來判斷，KMO 值又稱「抽樣適切性量測值」(measure of sampling adequacy)。當 KMO 大於等於 0.80 時表示變項間的共同因素越多，適合進行主成分/因素分析，大於等於 0.60 時表示尚可接受，倘低於 0.50 時，則表示因為指標變數間的共同因素過於凌亂，導致該資料及不適合進行分群做主成分/因素分析 (Kaiser & Rice, 1974)。

本研究使用 R 語言 psych 套件中 KMO 指令，計算細節如下(Revelle, 2017)，算是中 R 為資料集中變數之相關矩陣：

$$S^2=\text{diag}(R^{-1})^{-1} \text{ and } Q=SR^{-1}S$$

$$\text{sumr}^2=\sum R^2 \text{ and } \text{sumq}^2=\sum Q^2$$

$$\text{KMO} = \frac{\text{sumr}^2}{\text{sumr}^2 + \text{sumq}^2}$$

分析程式及套件資訊

- 分析平台：R version 3.4.4。
- 分析套件：主成分分析使用內建 prcomp 函數、KMO 判別使用 psych 套件(Revelle, 2017)中 kmo( ) 函數、因素分析轉軸使用 psych 套件中 principal( ) 函數。

doi:10.6342/NTU201803135

3.3.3 群集分析

本研究使用階層式群集分析，表現採樣點間的關係，並以 5 年為個單位，分析近 15 年間，河川測點間相似性的變化。

群集分析概念

群集分析是一種非監督式的樣態識別技術，他不需要先設定假設，可揭露數據集內部的結構及發現數據集內隱含的行為，分析方法可分為階層式(Hierarchical Method)及非階層式(Nonhierarchical Method)兩大類。階層式群集分析是最被廣為運用的方法，這個方法是由最相近的資料結合在一起，持續連接，連接到層級越高代表相似度越低。群集的分類無任何預設立場，純粹以數學角度將水質參數轉換成距離特性，再以阿基里德距離是用來判定兩樣本相似性的，不同的距離代表兩樣本的相似度是不一樣的。(Singh et al., 2004)

階層式群集分析方法

階層式群集運算方法有分離分層法及凝聚分層法等 2 種，凝聚分層法為較常被應用，開始時，每一個體自成一群，之後依序將距離最近的 2 個個體合成一群，一步步地使群組越變越少，直到所有的個體結合成一群，最後通常以樹枝狀圖表示觀測值的親緣遠近關係。(正昌, 2005)

計算群組間距離之方法

華德法(Wards Method，又稱「華德最小變異法」)距離計算方法如下列公式：

$$d_{A,B} = n_A\|\bar{x}_A - \bar{\bar{x}}\|^2 + n_B\|\bar{x}_B - \bar{\bar{x}}\|^2$$

分析程式及套件資訊

分析平台：R version 3.4.4
分析方法：Hierarchical Clustering，使用內建階層群集分析功能 hclust( ) 函數。
視覺化：使用 factoextra 套件 (Kassambara & Mundt, 2017)

doi:10.6342/NTU201803135

3.4 機器學習模型建置

本節內容包括 3.4.1 類神經網路模型及決策樹/決策森林等 2 種模型之運作流程介紹，3.4.2 說明建置完成模型之效果評估方法，評估方法分別為預測目標為類別及預設目標為數值等兩類模型之評估方式，最後以 3.4.3 介紹本研究所使用之機器學習分析平台及展示實際操作介面。

3.4.1 模型運作流程介紹

類神經網路模型因為效能佳，操作使用具極大彈性，而成為機器學習領域中最常被應用的模型之一，決策樹/決策森林模型則以構造簡單易於解釋等特色，亦廣為被應用於各種領域，本研究使用上述 2 種類模型作為判斷水質指標之方法。以下介紹上述模型之開發歷史、模型結構、運作原理及機制及流程執行參數設定等資訊。

一、類神經網路模型 (Artificial Neural Network)

類神經網路的概念首次出現於 1943 年由 McCulloch 和 Pitts 所提出，然而被實際應用在研究上則是到了 1986 年，由 Rumelhart 等人發展成前饋式(feedforward)倒傳遞演算法(Back-propagation Algorithm)。

模型結構

不同類型的類神經網路主要差異在於結構及決定網路裡神經元的啟動權重的運算方式不同。例如多層類神經網路(Multilayer perceptron, MLP)，其結構為在輸入層和輸出層間有一個以上的隱藏層(Palani, Liong, & Tkalich, 2008)，類神經網路基本架構如下圖 3- 5。

doi:10.6342/NTU201803135

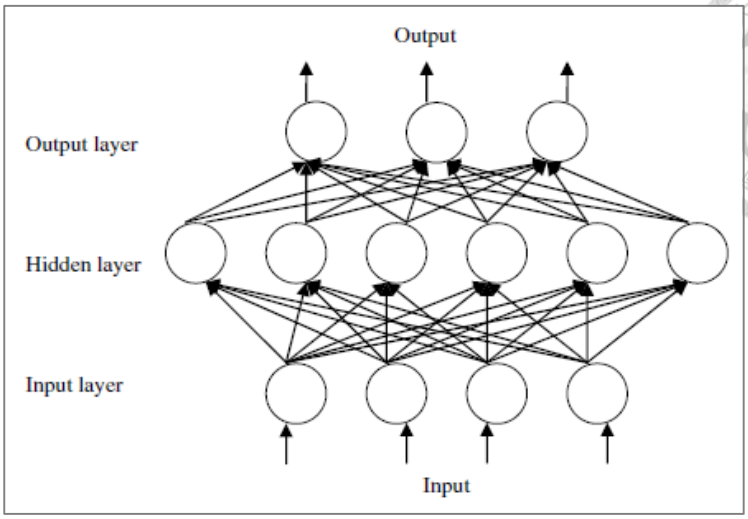


圖 3- 5　類神經網路基本架構

(資料來源：Chau,2016)

運作原理及機制

類神經網路的運作方式為模仿人類神經元連接及傳導的模式，適用於複雜、非線性的的方程式解。生物的神經系統由神經元構成，彼此間透過電流傳遞訊號。是否傳遞訊號、取決於神經細胞接收到的訊號量，當訊號量超過了某個閾值(Threshold)時，細胞體就會產生電流、通過突觸傳到其他神經元。

人工神經元(an artificial neuron)仿照生物神經元的設計，每個人工神經元有一個激發函數的公式，當對神經元接收個輸入值(input)，輸入值包括訊號值和權重，所有的輸入值經過激發函數的運算加總後，會決定要不要把訊號傳給下一個神經元(如圖 3- 6)，經過一系列的傳遞後，傳到最後一層輸出層即可輸出預測結果。

訓練模型之行為是一個調整輸入權重的過程，如果第一次輸出結果有誤，倒傳遞神經系統會回傳誤差訊號，對剛走過的每一個神經元調整權重，如此重複數次後，機器就學會如何辨識正確的結果。

doi:10.6342/NTU201803135

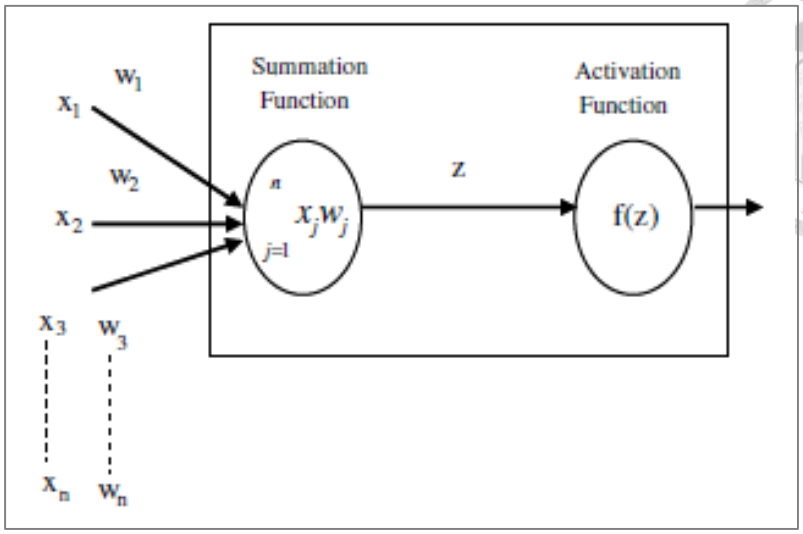


圖 3- 6　人工神經元結構圖

(資料來源：Palani, Liong, & Tkalich, 2008))

流程及執行參數設定

類神經網路模型建立步驟如下：

1. 決定模型參數

(1) 隱藏層及節點數量

(2) 學習速率(learning rate, η)及動量(momentum, α)

(3) 初始權重(Initial weights)

(4) 停止準則(Stopping criteria)

2. 選擇輸入的參數及欲分析目標(Selection of input variable and target object)

3. 分割資料集(Data partition)

4. 評估模型表現(Model performance evaluation)

二、決策樹/決策森林 (Decision Tree/Decision Forest)

簡介及模式結構

決策樹模型是最為廣用的分類法，他最主要的優勢是方法簡單且運算效率高，訓練出來的結果能視覺化，因此也容易解釋，例如由根節點開始遇到

doi:10.6342/NTU201803135

的第 1 個分支是最重要的問題，如圖 3- 7，該例子顯示判斷水質污染等及最重要的因素是「生化需氧量是否大於 20 mg/L」，越不重要的問題會位於分支末端，這讓決策者可快速將重點集中在最重要決定性的因素上。因此決策樹方法被廣泛應用於醫學、生物、天文或商業等各種領域(Winkler et al., 2018)。

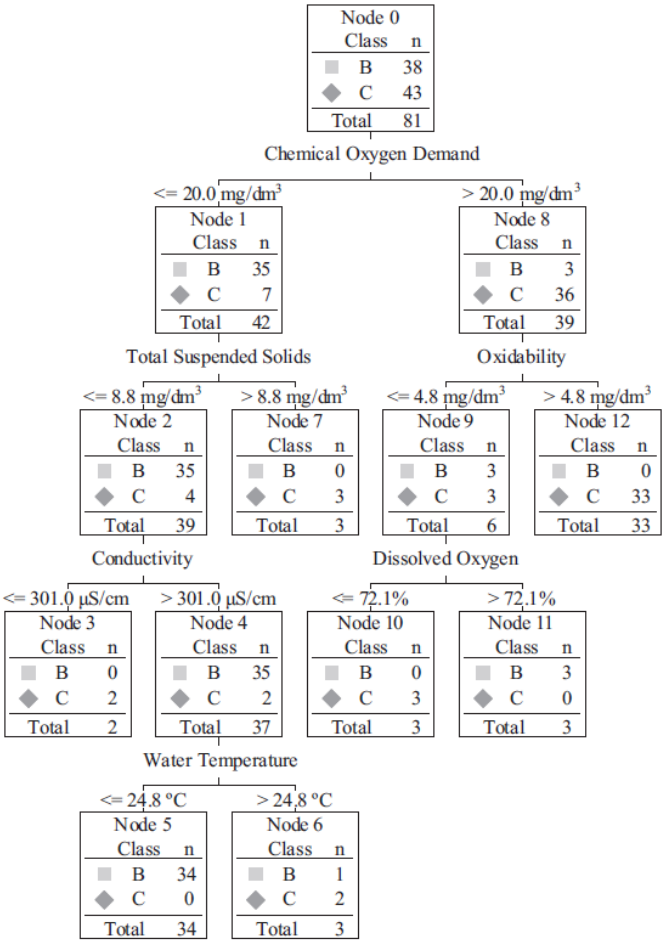


圖 3- 7　預測水質污染的決策樹模型

資料來源：(Couto et al., 2012)

運作原理及機制

決策樹是種層級式的分支架構，包含形成分支的節點及作為分類終點的樹葉節點，每個內部節點(非樹葉節點)表示以某個屬性作為測試條件，每個分支代表該測試的輸出，而每個樹葉節點(或終端節點)存放一個類別編號(所

doi:10.6342/NTU201803135

欲預測的目標項目)。

其運作方法為建立一個根節點後，根據待解決問題條件選擇最佳的分支屬性及分支值，沿著分支路線將資料分至互斥子集，重複前面步驟直到停止條件為止。範例說明見圖 3- 8，如果要分開藍圓點與綠方塊，首要條件是以長度大於 42 的條件做第一層分支的節點，第二層則是以年齡大於 24 的條件做為分支節點，以 2 節點便可完成分組。

而決策森林則是決策樹的增強方法，經過抽樣先製作許多決策樹模型，這些訓練好的決策樹集合成決策森林，每棵決策樹的分支及節點安排不盡相同，對於輸入資料的預測結果也可能不一，決策森林將蒐集每棵樹的結果後進行投票，以多數決產生的預測值做為最終結果。

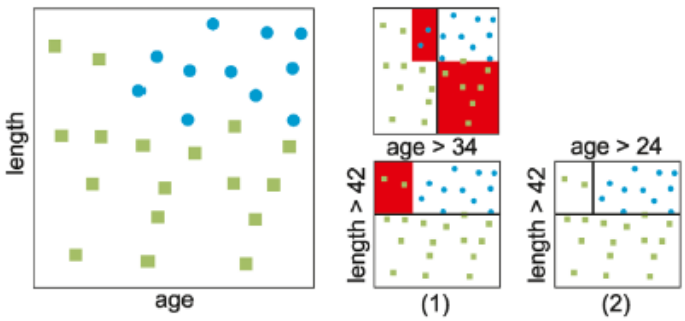


圖 3- 8　決策樹運作範例

資料來源：(Winkler et al., 2018)

流程及執行參數設定

1. 決定模型參數
   (1) 重複抽樣方法(Resampling method)
   (2) 決策森林中樹的數量(Number of decision tree)
   (3) 決策樹的高度(Maximum depth of the decision trees)
2. 選擇輸入的參數及欲分析目標(Selection of input variable and target object)
3. 分割資料集(Data partition)：分割資料集成訓練資料集、測試資料集
4. 評估模型表現(Model performance evaluation)

doi:10.6342/NTU201803135

3.4.2 模型效果評估

本研究使用兩種不同類型的模型，判斷水中銅濃度是否超標，使用二元分類模型，推估 RPI 指標數值則使用回歸模型。因這 2 類模型產出結果分別為類別變項及連續數值，所以評估模型效果方法亦不相同，以下就 2 種輸出結果之評估方法進行說明：

一、二元分類模型評估

銅濃度超標或是未超標為此二元模型分類判斷之標的，需依據預測結果及實際值歸納出判斷分析表格如下表 3-4，再依表格內容計算模型正確率、精確度、查全率及 F1 分數，以評估模型分類效果。(Azure, 2018)

表 3-4　二元分類模型判斷分析表格

預測 / 實際	超標	未超標
超標	正確正例 True Positive, TP	錯誤負例 False Negative, FN
未超標	錯誤正例 False Positive, FP	正確負例 True Negative, TN

一、正確率(Accuracy)：測試資料集分類正確的比例，值越接近 1 表示模型越準確。

$$\text{正確率(Accuracy)} = \frac{TP + TN}{TP + TN + FP + FN}$$

二、精確度(Precision)：測試資料集中被預測為超標的樣本，其分類正確的比例，值越接近 1 表示當得到預測值為超標時，樣本也確實為超標之正確率越高。

$$\text{精確度(Precision)} = \frac{TP + TN}{TP + TN + FP + FN}$$

三、查全率(Recall)：完整資料集(包括建模及測試資料集)中分類正確比例。

$$\text{查全率(Recall)} = \frac{\text{完整資料集中分類正確樣本}}{\text{完整資料集中所有樣本}}$$

doi:10.6342/NTU201803135

四、F1 分數(F1 Score)：精確度及查全率之調和平均數，通常用以評估兩不同模型之分析效果。

$$\text{F1 分數(F1 Score)} = \frac{2 \times \text{精確度(Precision)} \times \text{查全率(Recall)}}{\text{精確度(Precision)} + \text{查全率(Recall)}}$$

五、ROC 曲線下方的面積(Area under the Curve of ROC, ROC AUC)：表示若隨機抽取一個陽性樣本和一個陰性樣本，模型能正確判斷超標樣本的值高於未超標樣本之機率，AUC 值越接近 1 的模型，正確率越高。

二、回歸模型評估

使用回歸模型評估 RPI 數值，需比較各樣本預測結果中實際值及預測值的差值，以下為各種比較計算公式，均用以評估模型分類效果(劉應興, 1997)。

1. 絕對誤差平均(Mean Absolute Error, MAE)：預測值和實際值的差異絕對值平均，值越低表示預測值及實際值間誤差越小。

$$\text{MAE} = \frac{1}{N}\sum\left|X_{observed} - X_{predicted}\right|$$

2. 平均絕對值誤差百分比(Mean Absolute Percentage Error , MAPE) ：本評估指標為相對數值，不受測量值與預估值單位與大小之影響，能客觀得獲得預估值與實際值間之差異程度。

$$\text{MAPE} = \frac{1}{N}\sum\frac{\left|X_{observed} - X_{predicted}\right|}{X_{observed}}$$

3. 均方根誤差(Root Mean Square Error, RMSE)：預測值和實際值差異的平方平均後再開根號，值越低表示預測值及實際值間誤差越小。

$$\text{RMSE} = \sqrt{\frac{\sum(X_{observed} - X_{predicted})^2}{N}}$$

doi:10.6342/NTU201803135

4. 決定係數(coefficient of determination, R^2)：為相關係數的平方，用以解釋依變項與自變項間之線性關係的強弱，其值越接近 1 表示模型的預測效能越高。

$$R^2 = 1 - \left(\frac{\sum(X_{observed} - X_{predicted})^2}{\sum(X_{observed} - X_{mean\ observed})^2}\right)$$

doi:10.6342/NTU201803135

3.4.3 模型建置平台

本研究使用微軟公司開發之 Azure Machine Learning Studio 作為模型建置之平台，該平台以雲端服務方式提供使用者操作，採用容易使用的瀏覽器為操作介面，只要在工作環境中拖放執行步驟模塊，並依需求設定訓練參數，即可完成模型建置，對於初學者來說是個容易上手的工具。

水中金屬銅超標判斷模型建置

使用 Azure Machine Learning Studio 建立銅金屬超標與否判斷的流程如下圖 3-9，首先將數據及存入雲端系統後，即可以拖曳的方式將數據集模塊置入工作空間，接下來進行遺失值清除，標準化及選定匯入的水質參數及預測目標後，將目標銅濃度超標與否判斷欄位變為類別變項等前處理作業。數據前處理完成後，即進行資料分割，將 70 %的數據做為訓練模型用，剩餘 30%數據做為測試模型用。因為判斷目標為二元類別，所以訓練模型是由各種二元分類的方法中選擇出二元決策森林(Two-class Decision Forest)及二元神經網路(Two-Class Neural Network)2 種模型，模型訓練參數以系統推薦的預設數值為主(如下表 3- 5)，最終進行模型可行與否的評估。

doi:10.6342/NTU201803135

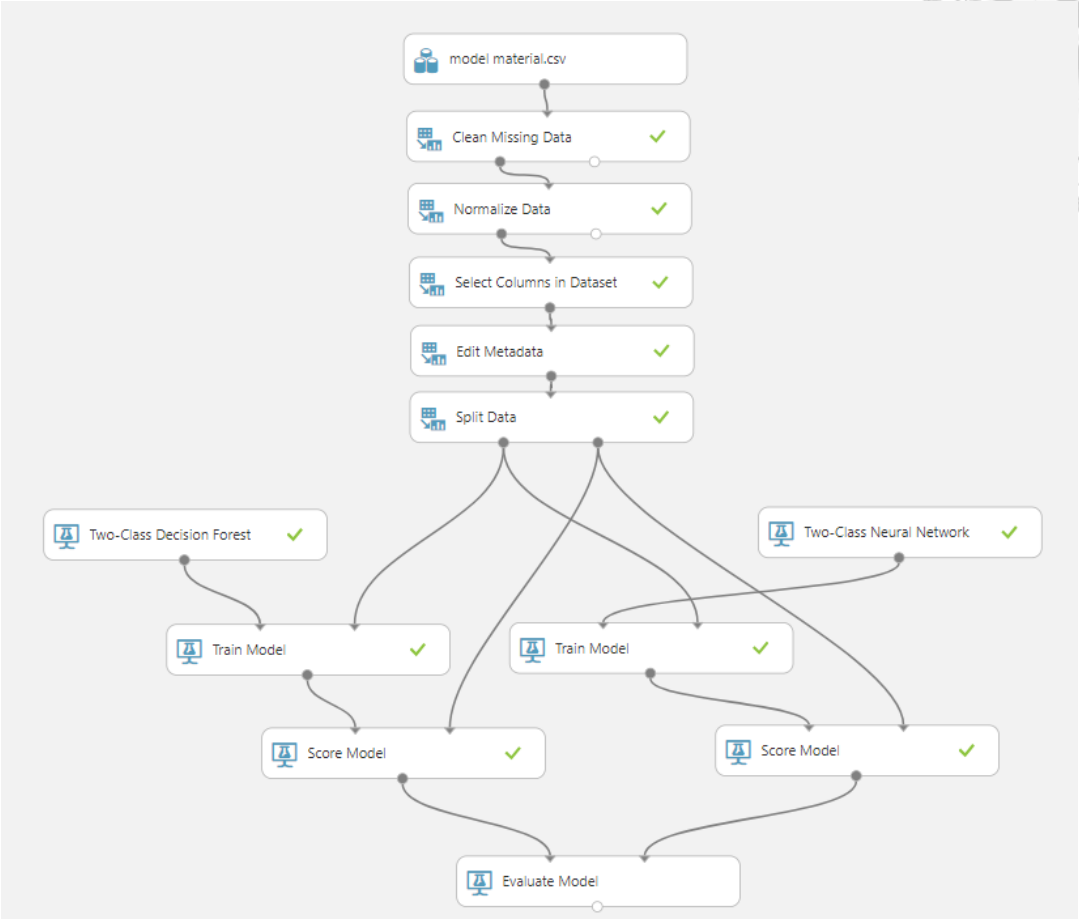


圖 3- 9　水中金屬銅超標判斷模型建置流程

doi:10.6342/NTU201803135

表 3- 5　水中金屬銅超標判斷模型訓練參數設定

Two-Class Decision forest		Two-Class Neural Network	
Create trainer mode	Single Parameter	Create trainer mode	Single Parameter
Resampling method	Bagging	Hidden layer specification	Fully-connected case
Number of decision tree	200	Number of hidden nodes	100
Maximum depth of the decision trees	32	Learning rate	0.5
Number of random splits per node	128	Number of learning iterations	100
Minimum number of samples per leaf node	1	The initial learning weights diameter	0.1
-	-	The momentum	0
-	-	The type of normalizer	Min-Max normalizer
-	-	Shuffle examples	Checked
-	-	Random number seed	168
Allow unknown values for categorical features	Checked	Allow unknown values for categorical features	Checked

doi:10.6342/NTU201803135

RPI 污染指標值判定模式建立

使用 Azure Machine Learning Studio 建立河川污染程度指標判斷的流程如下圖 3- 10，同樣將數據集存入雲端系統後，即可以拖曳的方式將數據集模塊置入工作空間，因為所使用的參數為基礎測項，每個月均有測值，所以略過清除遺失值的步驟，選擇目標欄位後進行標準化，因為欲判斷的項目為數值，亦不須使用資料及編輯工具將項目變為類別變項。數據前處理完成後，即進行分割，將 70 %的數據做為訓練模型用，因為欲預測的目標為數值，所以使用回歸類別的模組，使用模組為決策森林回歸(Decision Forest Regression)及神經網路回歸(Neural Network Regression)，其模型訓練參數設定以系統推薦的預設數值為主(如下表 3- 6)，惟有隱藏層神經節數目部分，依據 Hecht-Nielsen 等學者建議設定為小於 2 倍輸入參數量加 1(Hecht-Nielsen, 1987)，即建置模型所需參數有 4 個(COD、NH_3-N、SS、DO)，則隱藏層神經節數目設定為 9，將測試集數據載入測試模型後，最終進行模型可行與否的評估。

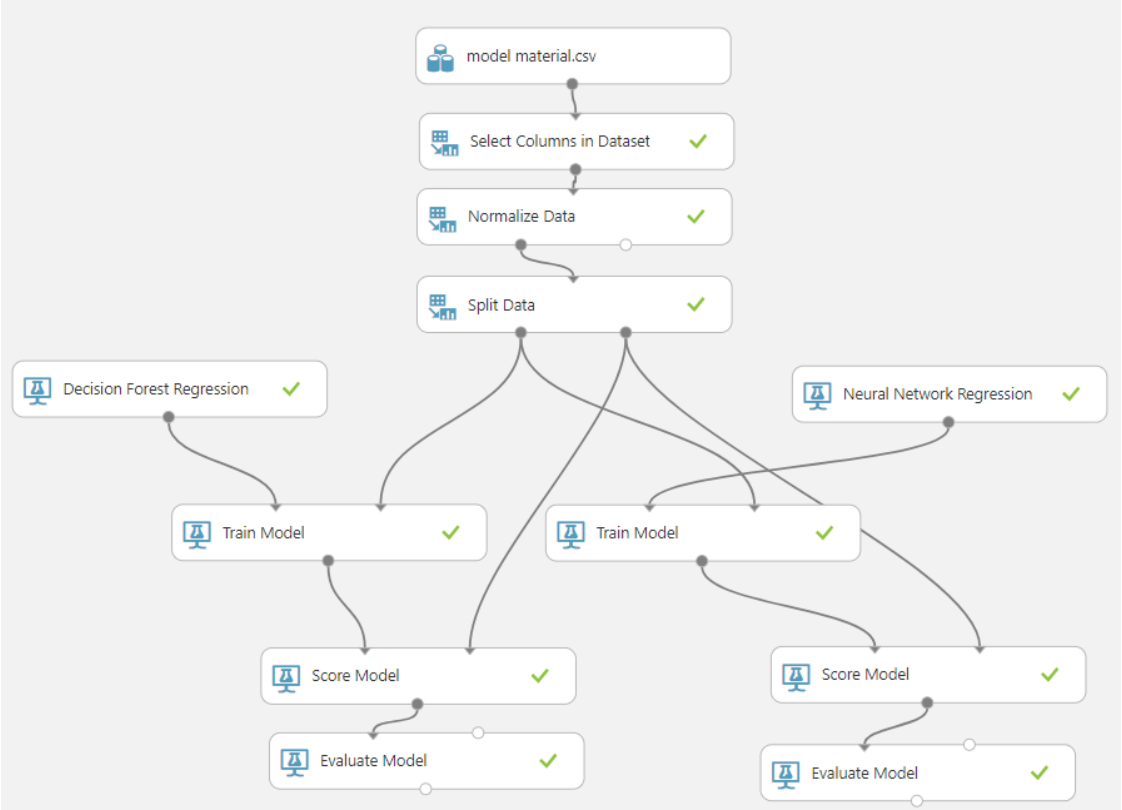


圖 3- 10　RPI 污染指標值判定模式建立流程圖

doi:10.6342/NTU201803135

表 3- 6　RPI 污染指標值判定模式訓練模型參數設定

Decision Forest Regression		Neural Network Regression	
Create trainer mode	Single Parameter	Create trainer mode	Single Parameter
Resampling method	Bagging	Hidden layer specification	Fully-connected case
Number of decision tree	64	Number of hidden nodes	9
Maximum depth of the decision trees	32	Learning rate	0.005
Number of random splits per node	128	Number of learning iterations	100
Minimum number of samples per leaf node	1	The initial learning weights diameter	0.1
-	-	The momentum	0
-	-	The type of normalizer	Min-Max normalizer
-	-	Shuffle examples	Checked
-	-	Random number seed	168
Allow unknown values for categorical features	Checked	Allow unknown values for categorical features	Checked

doi:10.6342/NTU201803135

第四章　分析結果

本章分為三節說明，分別為 4.1 老街溪水質資訊，以單測項敘述水質變化，4.2 節為多變量分析，以主成分分析/因素分析及群集分析方法說明整體水質變化趨勢，4.3 節為機器學習判斷模型建置，以機器學習方法建構水中銅濃度是否超標分類模型及水質 RPI 指標判斷模型，並呈現其運作結果及可行性，完整架構如下圖 4- 1。

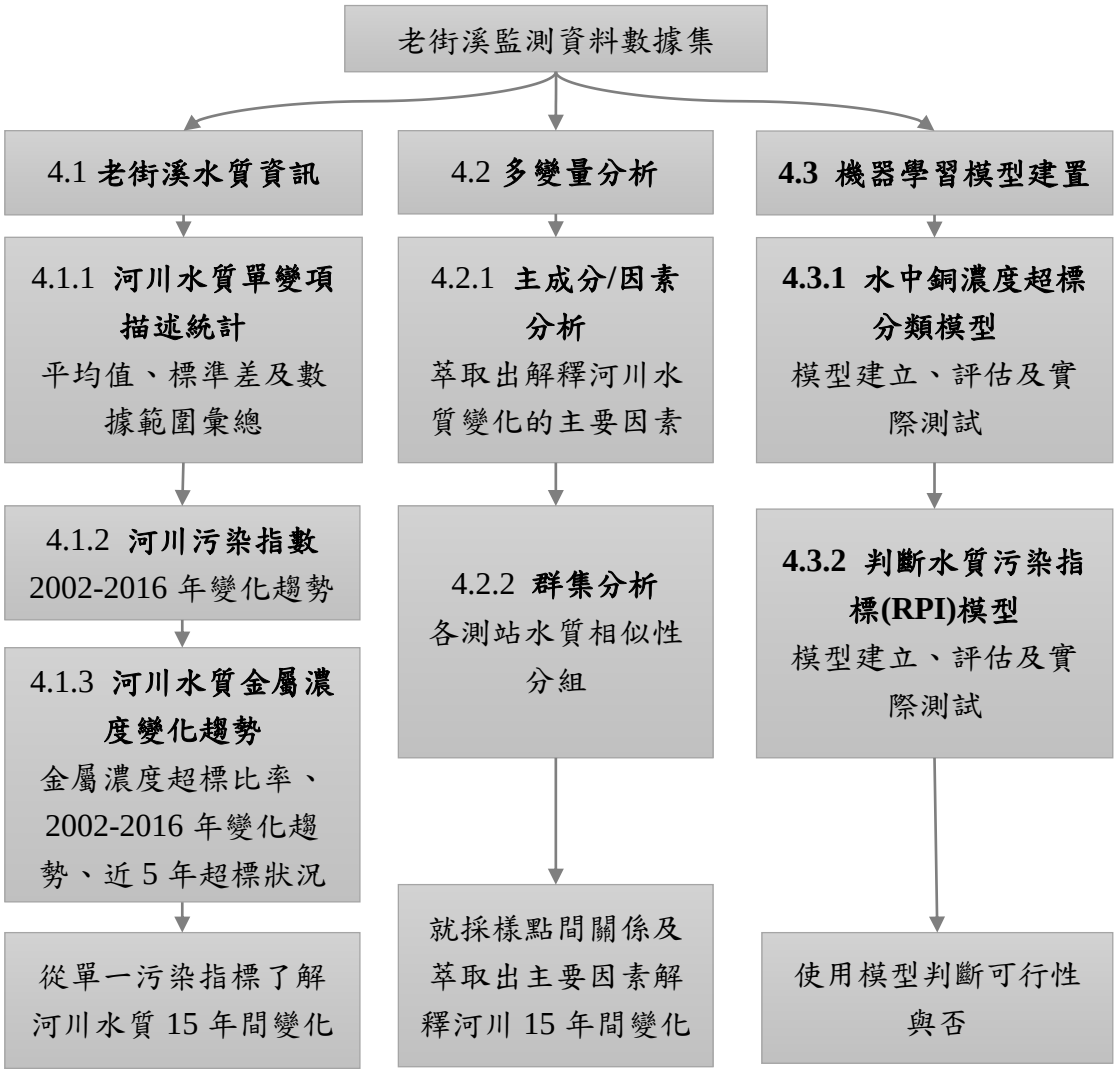


圖 4- 1　分析結果章節架構圖

doi:10.6342/NTU201803135

4.1 老街溪水質資訊

經蒐集 2002 - 2016 年計 15 年間，桃園市老街溪流域中 7 個測站之水質，總樣本數 1,260 個，包含水質測項 32 個，合計有 21,194 個資料點，本節將由全年度描述統計資料、歷年 RPI 指標及金屬超標現象作探討。

於 4.1.1 小節呈現 2002-2016 年 15 年間的水質測項平均值、標準差及最大最小值彙整表格，4.1.2 小節探討歷年度各測站 RPI 之變化，並將構成 RPI 的成員如溶氧量(DO)、懸浮固體(SS)、生化需氧量(BOD)及氨氮(NH_3-N)等水質測項一併作圖分析，以找出影響水質 RPI 指標最大的測項為何者。4.1.3 小節則就水中金屬濃度作探討，分析各河段水質超過環保署訂定標準的現況。

4.1.1 河川水質單變項描述統計

彙整 2002 - 2016 年間，於各採樣點測得之所有河川水質監測項目平均值、標準差及最大最小值如表 4- 2 及表 4- 3。觀察各水質測項在各採樣點之平均測得數值，以粗黑體字型表現測項最大平均值之落點，各水質測項平均值最大點位多出現於 S3 平鎮第一號橋或 S7 許厝港一號橋點位，如表 4- 1。

表 4- 1　水質測項平均最高值於各測站之分布

測站	測項平均最高項目
S1 美都麗橋	鉛(Pb)
S2 北勢橋	酸鹼值(pH)、懸浮固體(SS)、硒(Se)
S3 平鎮第一號橋	氨氮(NH_3-N)、總磷(TP)、硝酸氮(NO_3-N)、銅(Cu)、錳(Mn)
S4 環鄉橋	大腸桿菌群(Coliform)、砷(As)
S5 公園橋	溶氧(DO)、酸鹼值(pH)
S6 中正橋	水溫、酸鹼值(pH)
S7 許厝港一號橋	河川污染程度指標(RPI)、導電度(EC)、生化需氧量(BOD)、化學需氧量(COD)、總有機碳(TOC)、亞硝酸氮(NO_2-N)、鉻(Cr^{6+})、銀(Ag)及鋅(Zn)

doi:10.6342/NTU201803135

表 4- 2　2002-2016 年間水質參數描述統計值(非金屬部分)

水質參數	統計概要	S1 美都麗橋	S2 北勢橋	S3 平鎮第一號橋	S4 環鄉橋	S5 公園橋	S6 中正橋	S7 許厝港一號橋
RPI	Mean±SD	5.1 ± 1.5	4.1 ± 1.6	5.6 ± 1.5	5 ± 1.1	4.3±1.3	3.8±1.7	**5.9**±1.8
	Range	1.5~8.3	1~10	2.3~10	1.5~9	1~8.3	1~8.3	1.5~10
水溫	Mean±SD	22.3 ± 4.1	22.9 ± 5.1	23.7 ± 4.5	23.5 ± 4.8	24.7 ± 5.4	**25** ± 5.9	24.6 ± 5.7
	Range	12.2~29.8	12.2~32.2	14.3~32.5	12.5~32.5	13.1~35.1	12.3~36.1	12.6~35.1
pH	Mean±SD	6.9 ± 0.4	**7.6** ± 0.6	7.4 ± 0.3	7.5 ± 0.3	**7.6** ± 0.4	**7.6** ± 0.4	7.3 ± 0.7
	Range	3.5~8.2	5.7~10.6	6.6~9.2	4.8~8.4	6.4~8.9	5.9~9	2.9~9.2
EC	Mean±SD	385 ± 442	508 ± 265	1567 ± 446	930 ± 633	872 ± 599	743 ± 621	**1580** ± 677
	Range	130~5960	175~3730	311~2490	250~8750	243~8230	233~8740	365~5550
DO	Mean±SD	5.2 ± 2	7.3 ± 1.9	5.6 ± 1.7	6.6 ± 1.3	**7.8** ± 1.8	6.7 ± 1.6	5.5 ± 2.5
	Range	0.8~11.9	0.9~10.5	0.2~9.8	2.1~9.5	3.8~15.7	3~13.7	0.1~11.9
BOD	Mean±SD	9.5 ± 7.7	8.2 ± 7.9	13.2 ± 10.9	10.4 ± 6.1	7.7 ± 8.7	7.6 ± 6.5	**23** ± 23.6
	Range	1.7~48.7	2.3~67.1	1.3~65.7	2.6~45	1.9~96.2	1.5~71.3	3.2~114
COD	Mean±SD	34.8 ± 25.7	32.5 ± 22.2	63.6 ± 42.9	47.5 ± 29.2	36.4 ± 30.3	33.2 ± 18.1	**87.3** ± 78.8
	Range	9.9~194	10.8~179	16.9~281	5.8~259	11~335	14~180	17.1~666
SS	Mean±SD	17.2 ± 51.1	**68.2** ± 248	65.3 ± 224	31 ± 61.3	43 ± 135.3	41.2 ± 119.7	57.8 ± 202.2
	Range	<3~675	3.9~2010	4.8~2570	<3~604	2.6~1470	2.4~1370	4.2~2720
coliform	Mean±SD	344857 ± 724549	231963 ± 753812	255060 ± 1082118	**243603** ± 461818	88182 ± 316945	53712 ± 92050	216997 ± 509431
	Range	<10~5400000	220~9500000	15~9500000	25~4400000	320~3200000	260~840000	<10~4300000
NH_3-N	Mean±SD	6.24 ± 3.77	3.13 ± 2.65	**9.74** ± 7.38	6.98 ± 3.73	4.61 ± 2.8	2.19 ± 2.03	6.66 ± 7.87
	Range	0.8~21	0.03~14	0.48~38.4	0.17~24.7	0.35~14.6	0.02~9.58	0.12~63

doi:10.6342/NTU201803135

水質參數	統計概要	S1 美都麗橋	S2 北勢橋	S3 平鎮第一號橋	S4 環鄉橋	S5 公園橋	S6 中正橋	S7 許厝港一號橋
TP	Mean ±SD	0.92 ± 0.6	0.65 ± 0.38	**4.29** ± 7	1.81 ± 1.63	1.15 ± 0.8	0.86 ± 0.37	1.17 ± 0.52
	Range	0.252~2.88	0.151~1.85	0.331~45	0.412~9.18	0.37~4.15	0.355~2.1	0.113~2.27
TOC	Mean±SD	9.08 ± 5.94	7.3 ± 3.02	15.54 ± 11.35	11.82 ± 6.59	8.35 ± 2.66	8.05 ± 2.71	**16.75** ± 8.28
	Range	3.02~30.6	3.7~20	4.25~78.4	4.47~35.8	3.52~17.8	3.28~16.8	4.8~38.9
NO_3-N	Mean±SD	2.25 ± 2.34	6.19 ± 3.69	**10.25** ± 7.76	4.81 ± 2.11	3.94 ± 1.66	4.26 ± 1.61	5.68 ± 4.07
	Range	0.37~14.1	0.37~15.4	0.89~56	0.74~12	1.12~10.5	0.76~8.76	1.76~23.3
NO_2-N	Mean±SD	0.124 ± 0.071	0.467 ± 0.291	0.473 ± 0.312	0.404 ± 0.167	0.56 ± 0.343	0.488 ± 0.253	**0.969** ± 1.185
	Range	0.008~0.327	0.032~1.29	0.067~1.51	0.114~0.966	0.116~2.05	0.105~1.16	0.064~8.01

doi:10.6342/NTU201803135

表 4- 3　2002-2016 年間水質參數描述統計值(金屬部分)

水質參數	統計概要	S1 美都麗橋	S2 北勢橋	S3 平鎮第一號橋	S4 環鄉橋	S5 公園橋	S6 中正橋	S7 許厝港一號橋
Cd	Mean ± SD	0.001 ± 0	0.001 ± 0.002	0.001 ± 0	0.001 ± 0	0.001 ± 0.001	0.001 ± 0.002	0.001 ± 0
	Range	<0.001	<0.001~0.011	<0.001	<0.001	<0.001~0.007	<0.001~0.016	<0.001~0.001
Pb	Mean ± SD	**0.084** ± 0.226	0.006 ± 0.005	0.019 ± 0.058	0.009 ± 0.01	0.006 ± 0.006	0.01 ± 0.022	0.011 ± 0.014
	Range	<0.003~1.28	<0.003~0.015	<0.003~0.427	<0.003~0.058	<0.003~0.023	<0.003~0.164	<0.003~0.063
Cr^{6+}	Mean ± SD	0.002 ± 0.002	0.002 ± 0.002	0.006 ± 0.033	0.003 ± 0.009	0.002 ± 0.002	0.002 ± 0.002	**0.024** ± 0.136
	Range	<0.002~0.005	<0.002~0.005	<0.002~0.258	<0.002~0.073	<0.002~0.005	<0.002~0.005	<0.002~1.05
As	Mean ± SD	0.002 ± 0.004	0.002 ± 0.002	0.012 ± 0.012	**0.005** ± 0.004	0.003 ± 0.002	0.003 ± 0.001	0.003 ± 0.001
	Range	<0.0005~0.0299	<0.0005~0.0129	<0.0005~0.0475	0.0007~0.0205	0.00025~0.011	0.001~0.0061	0.0005~0.0058
Hg	Mean ± SD	0.0002 ± 0.0001	0.0002 ± 0.0002	0.0002 ± 0.0001	0.0002 ± 0.0001	0.0002 ± 0	0.0002 ± 0.0002	0.0002 ± 0.0001
	Range	<0.0003~0.0008	<0.0003~0.0014	<0.0003~0.0009	<0.0003~0.0007	<0.0003	<0.0003~0.0015	<0.0003~0.0007
Cu	Mean ± SD	0.683 ± 1.658	0.069 ± 0.065	**1.539** ± 2.677	0.449 ± 0.811	0.109 ± 0.159	0.072 ± 0.106	1.273 ± 3.055
	Range	0.002~9.17	<0.005~0.34	0.04~12.3	0.012~4.39	0.008~0.868	0.007~0.518	0.013~18.4
Zn	Mean ± SD	0.03 ± 0.024	0.117 ± 0.114	0.301 ± 0.294	0.14 ± 0.139	0.062 ± 0.043	0.049 ± 0.044	**0.564** ± 0.654
	Range	0.008~0.108	0.015~0.534	0.065~2	0.038~0.972	0.007~0.266	0.012~0.272	0.024~2.82
Mn	Mean ± SD	0.13 ± 0.152	0.159 ± 0.197	**0.252** ± 0.154	0.188 ± 0.108	0.159 ± 0.248	0.102 ± 0.095	0.206 ± 0.206
	Range	<0.005~0.863	0.023~1.05	0.072~0.713	0.061~0.613	0.026~1.92	0.016~0.518	0.038~1.06
Ag	Mean ± SD	0.001 ± 0.001	0.001 ± 0.001	0.001 ± 0.001	0.001 ± 0.001	0.001 ± 0.001	0.001 ± 0.003	**0.002** ± 0.002
	Range	<0.001~0.005	<0.001~0.004	<0.001~0.009	<0.001~0.004	<0.001~0.0025	<0.001~0.02	<0.001~0.016
Se	Mean ± SD	0.0006±0.0002	**0.0007**±0.0003	0.0006±0.0002	0.0006±0.0002	0.0006±0.0002	0.0006±0.0002	0.0006±0.0002
	Range	<0.001~0.001	0.0005~0.002	<0.001~0.001	<0.001~0.001	<0.001~0.001	<0.001~0.001	<0.001~0.001

doi:10.6342/NTU201803135

4.1.2 河川污染程度指標變化趨勢

河川污染程度指標(River Pollution Index, RPI)為辨別河川污染程度的重要指標，計算方法為先測量氨氮(NH_3-N)、溶氧量(DO)、懸浮固體(SS)及生化需氧量(COD)濃度後，分別對照點數量表得到每個測項之污染點數，將所得總點數平均後即可得RPI 指標，指數落於 2.0 以下者為未(稍)受污染，2.1 至 3.0 間為輕度污染、3.1 至 6.0 間為中度污染，積分大於 6.0 以上者為嚴重污染。

歷年河川污染程度指標

經觀察 2002-2016 年間老街溪流域各測站 RPI 指標之變化，發現河川整體 RPI 平均趨勢為下降，自 2011 年以後，各測站均平均點數未出現嚴重污染之結果，惟整體 RPI 指標大部分仍落於中度污染等級，S2 北勢橋、S5 公園橋及 S6 中正橋是 RPI 指標較低、水質較好的點位，如圖 4-2，以下將就組成河川污染程度指標的水質測項個別探討變化趨勢，找出關鍵影響 RPI 指標之變項。

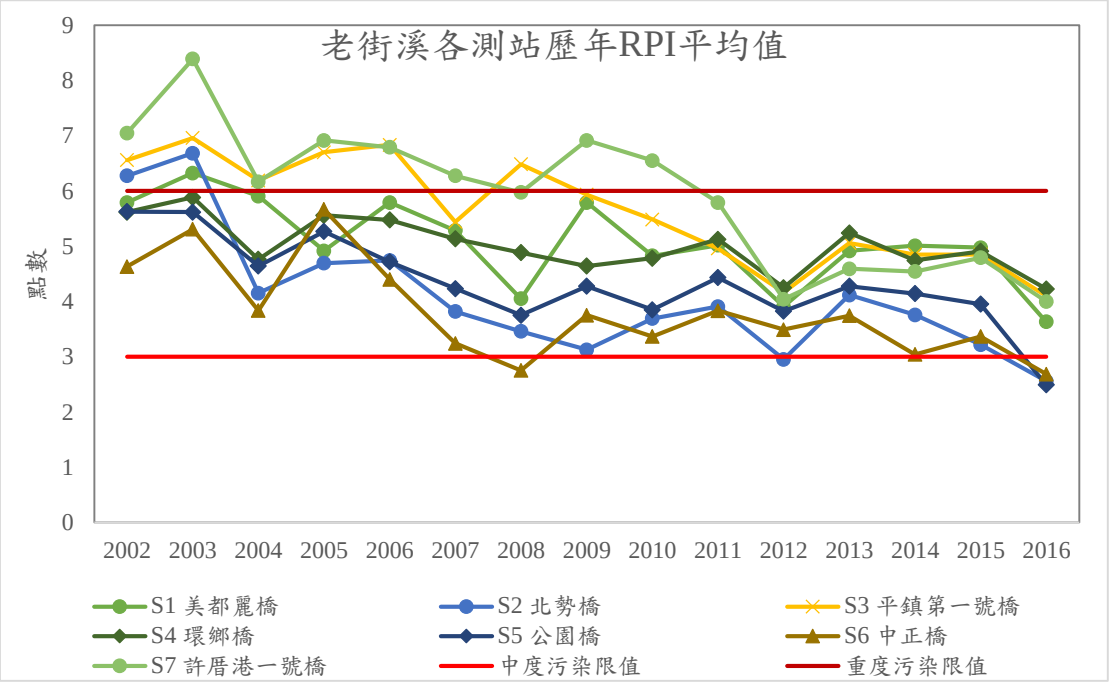


圖 4-2 老街溪各測站歷年 RPI 平均值

doi:10.6342/NTU201803135

探討水質指標對 RPI 指標之影響

就組成 PRI 指數之水質指標(DO、SS、NH_3-N 及 BOD)分別繪製歷年於各測站之測值平均折線圖，以辨別影響 RPI 指標變化的主要原因。

一、各測站 DO 平均值折線圖如圖 4- 3，測值大於 4.5 mg/L 為輕度污染，介於 2 至 4.5 mg/L 屬中度污染，數值小於 2 mg/L 為嚴重污染，由歷年 DO 平均值可看出，少有觀測點落於中度或嚴重污染限值區間。

二、各測站 SS 平均值折線圖如圖 4- 4，測值小於 50 mg/L 為輕度污染，介於 50 至 100 mg/L 為中度污染，大於 100 mg/L 即為嚴重污染，由歷年 SS 平均值可看出，2006 年以前落於中度或嚴重污染之平均值較多，但 2007 年以後平均值大部分屬輕度污染或無污染，少部分測站受極端測值影響使平均值落於嚴重或中度污染。

三、各測站 NH_3-N 平均值折線圖如圖 4- 5，測值小於 1 mg/L 為輕度污染，介於 1 至 3 mg/L 為中度污染，大於 3 mg/L 即為嚴重污染，由歷年 NH_3-N 平均值可看出，大部分平均值落於嚴重污染，近年雖稍有下降，惟仍多落於中度污染。

四、各測站 BOD 平均值折線圖如圖 4- 6，測值小於 5 mg/L 為輕度污染，介於 5 至 15 mg/L 為中度污染，大於 15 mg/L 即為嚴重污染，由歷年 BOD 平均值可看出，各測站平均值多落於中度污染區間。

由此可知，老街溪流域各測站 RPI 指標多落於中度或重度污染，可歸因於 NH_3-N 及 BOD 濃度過高，至於 DO 及 SS 對於 RPI 指標較無不良之影響。

doi:10.6342/NTU201803135

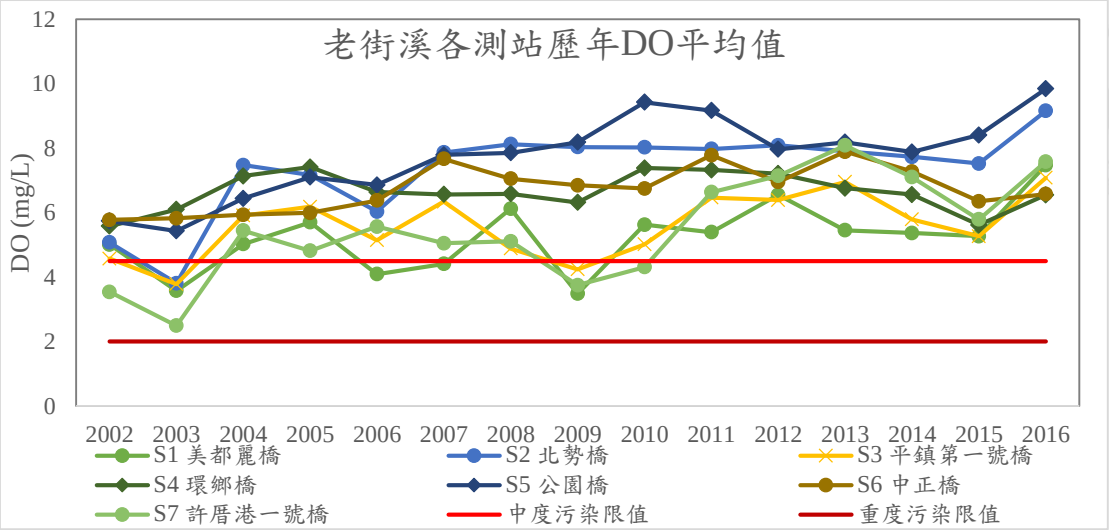


圖 4- 3　老街溪各測站歷年 DO 平均值

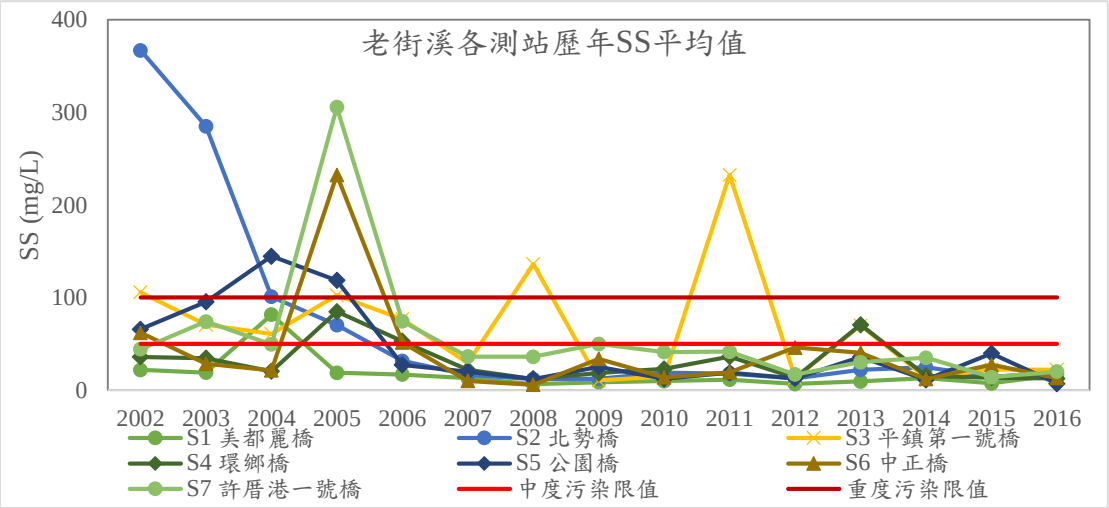


圖 4- 4　老街溪各測站歷年 SS 平均值

doi:10.6342/NTU201803135

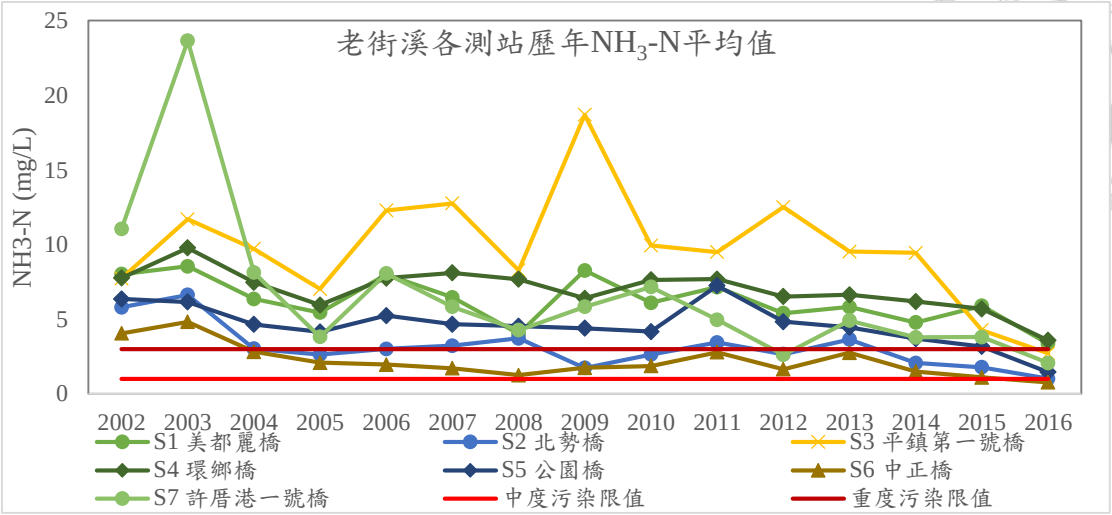


圖 4- 5　老街溪各測站歷年 NH_3-N 平均值

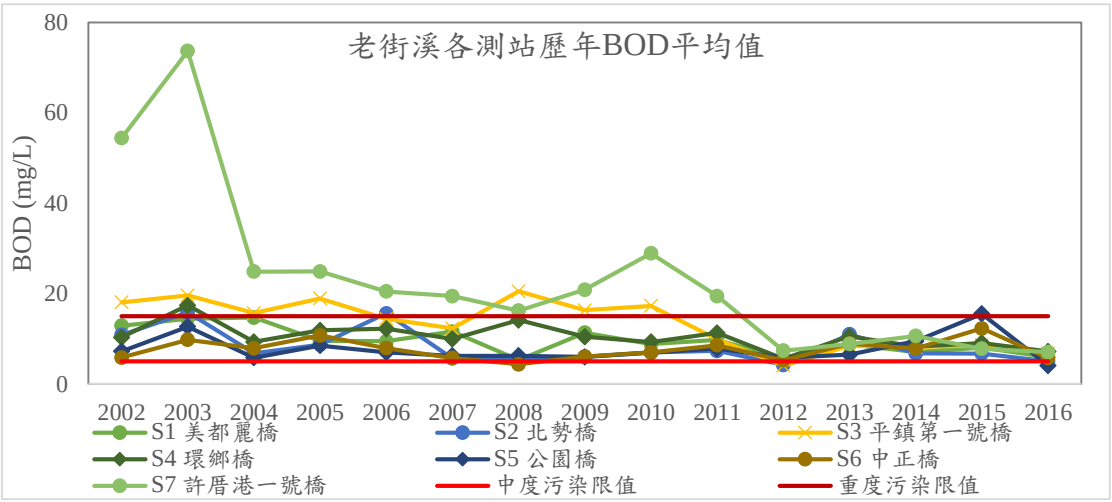


圖 4- 6　老街溪各測站歷年 BOD 平均值

doi:10.6342/NTU201803135

4.1.3 河川水質金屬濃度變化趨勢

依照水污染防治法附件訂定保護人體健康相關環境基準，規定水中金屬濃度超標與否之基準，各採樣點測得超過基準的水質變項累計超標次數如表 4-4，彙整 2002 年至 2016 年監測資訊可得知，總超標率最高排名前三者為錳(87 %)、銅(69 %)及鉛(29 %)，鋅及六價鉻總超標率分別為 8 %及 1 %，而砷、汞、銀、鎘、硒等測項於資料蒐集期間無超標紀錄。

就採樣點位而言，於 S3 平鎮一號橋、S4 環鄉橋及 S7 許厝港一號橋有較高機會測得超標的數值，其中在 S3 點位測得之金屬錳及銅超標率為 100 %。

表 4-4　水中金屬濃度累計超標次數

測站編號		基準值	S1	S2	S3	S4	S5	S6	S7	各金屬測值總超標率 各金屬總超標點位數/總測點數*100%
測點總數			59	59	59	59	59	59	59	
超標點位數	Mn	0.05	47	50	**59**	**59**	49	40	**56**	87%
	Cu	0.03	17	39	**59**	**53**	39	29	**51**	69%
	Pb	0.01	17	15	**19**	**20**	14	17	**19**	29%
	Zn	0.5	0	1	7	3	0	1	23	8%
	Cr^{6+}	0.05	0	0	1	1	0	0	3	1%
	As	0.05	0	0	0	0	0	0	0	0%
	Hg	0.001	0	0	0	0	0	0	0	0%
	Ag	0.05	0	0	0	0	0	0	0	0%
	Cd	0.005	0	0	0	0	0	0	0	0%
	Se	0.01	0	0	0	0	0	0	0	0%

進一步觀察歷年金屬濃度平均值，將水中金屬濃度超標率前 3 高之項目錳、銅及鉛濃度繪製歷年平均濃度折線圖，如圖 4-7、圖 4-9 及圖 4-11。可發現各金屬濃度隨時間變化趨勢相似，平均值高峰均集中於 2006 年以前，錳及銅雖然於 2006 年後平均測值大幅降低，但多數測值仍大於保護人體基準值(Mn 為 0.05 mg/L、Cu 為 0.03 mg/L)，如圖 4-8 及圖 4-10；金屬鉛於 2007 年以後之監測值均未超過保護人體基準值 0.01 mg/L，如圖 4-12。

doi:10.6342/NTU201803135

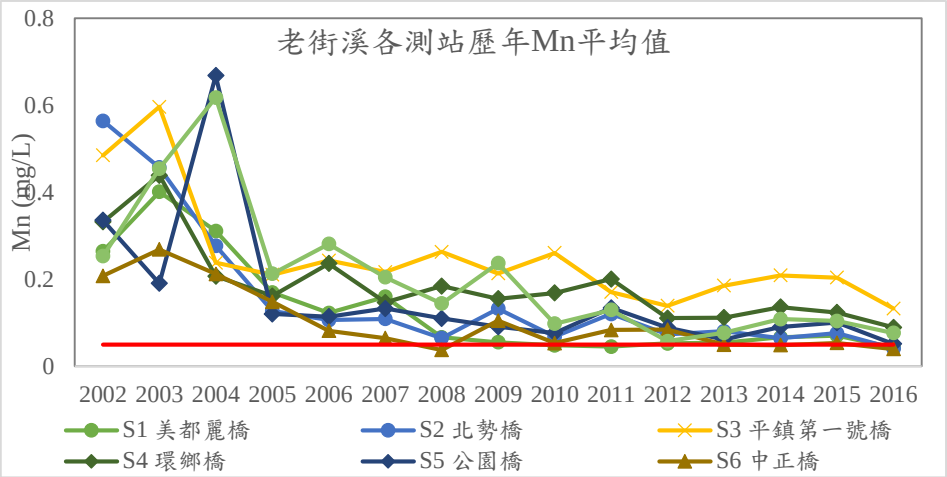


圖 4- 7　老街溪各測站歷年 Mn 平均值

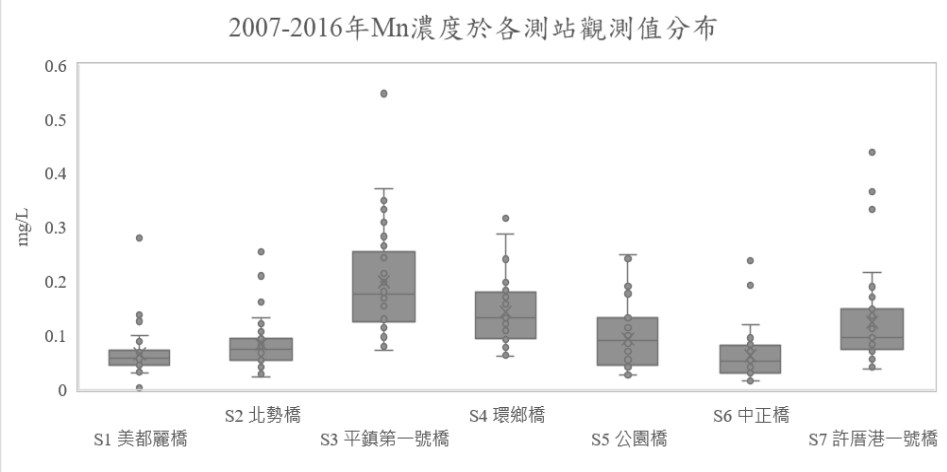


圖 4- 8　2007-2016 年 Mn 濃度於各測站觀測值分布

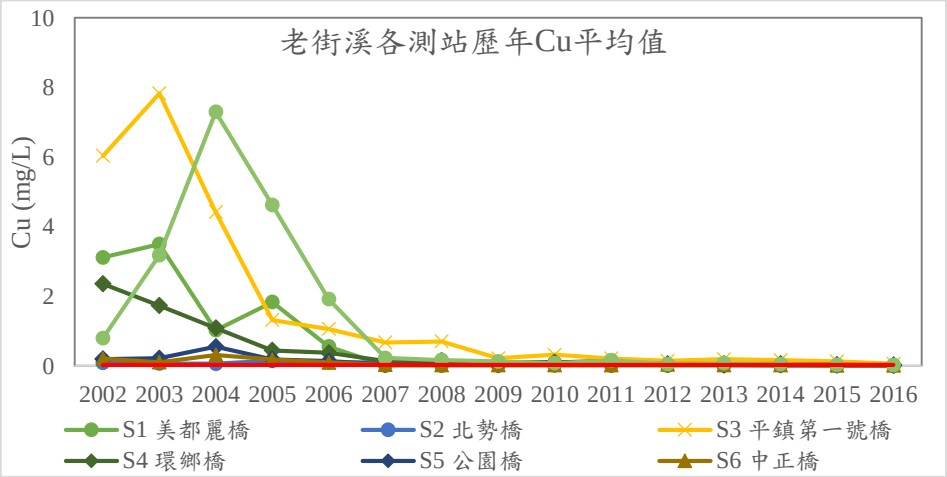


圖 4- 9　老街溪各測站歷年 Cu 平均值

doi:10.6342/NTU201803135

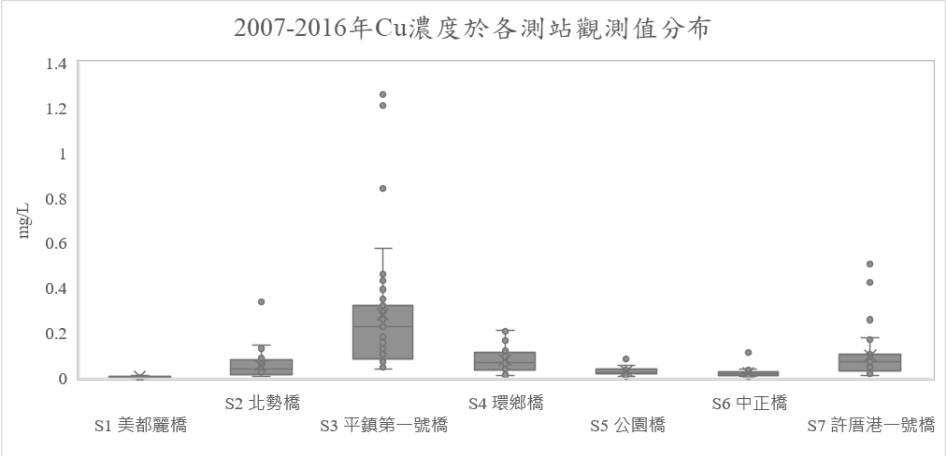


圖 4- 10　2007-2016 年 Cu 濃度於各測站觀測值分布

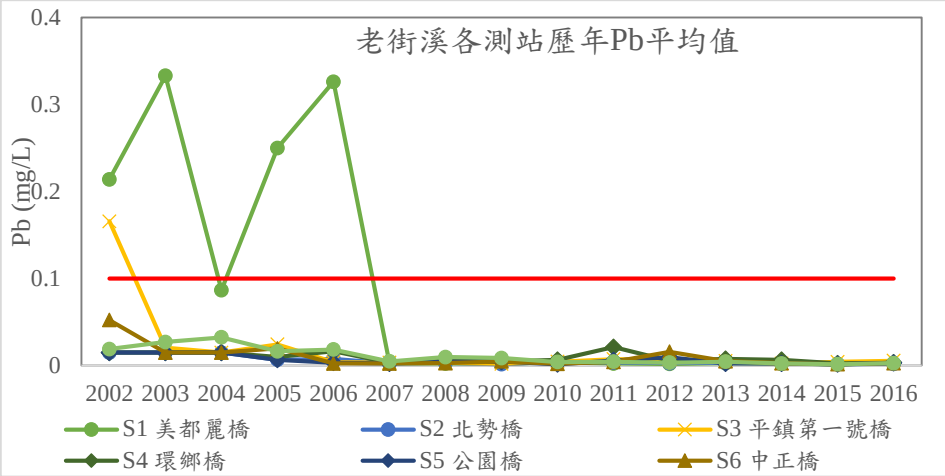


圖 4- 11　老街溪各測站歷年 Pb 平均值

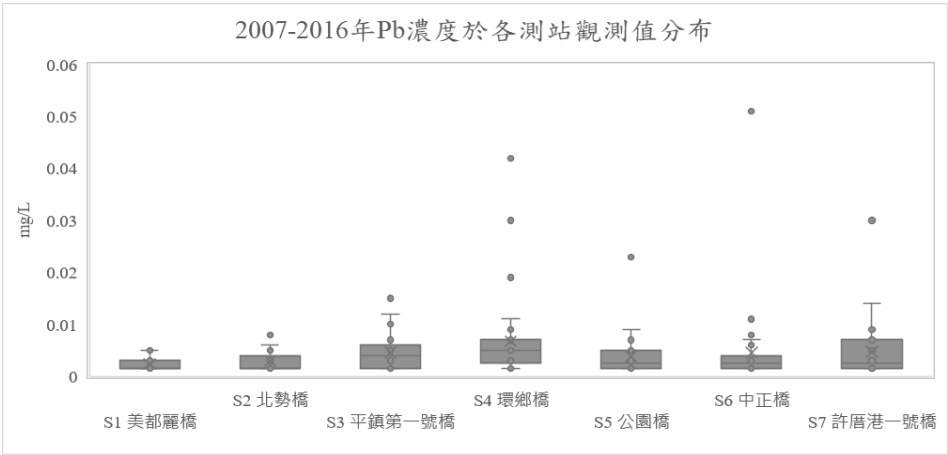


圖 4- 12　2007-2016 年 Pb 濃度於各測站觀測值分布

doi:10.6342/NTU201803135

4.2 多變量分析

4.2.1 主成分分析/因素分析

本節希望能藉由主成分分析及因素分析觀察老街溪水質之變化，首先以主成分分析方法萃取主成分，後經選定主成分數目並進行轉軸後將得到因素數目。分析結果將探討各因素的組成成員，及推測該組合是因為何種污染源影響而成，其後以因素代替原本的水質參數，分析因素分數在各點位之分布及隨著時間之變化。

分析數據及步驟

起初加入分析的目標變數為水溫、pH、EC、DO、BOD、COD、TP、NH_3-N、NO_3-N、SS、Hg、Mn、Cd、Ag、As、Pb、Cu、Zn、Cr^{6+}、Coliform 等 20 個水質測項，經 PCA 觀察各個水質測項對於主成分之貢獻之後，先後依序刪除 Cr^{6+}、水溫、coliform 等幾項貢獻不明顯的項目，最後以 17 個水質項目，401 筆樣本，總共 6,817 個數據值，進行後續 PCA/FA 分析。上開資料集經 KMO 取樣適切性量表數判讀結果為 0.71，代表適合做後續 PCA/FA 分析。

以主成分分析方法萃取因素並決定因素數目

表 4- 5 為以主成分分析方法萃取出來之前 6 個主成分，這六個主成分可以解釋整體資料及變異量之 70 %，在本階段各水質變項於主成分負荷偏低，將於決定因素數目後，以轉軸方法拉大因素負荷間的分布，將貢獻不明顯變項去除，使旋轉主成分(因素)的意義更加明顯且更容易解釋。(林倩如, 2006)

表 4- 5　主成分分析各水質變項負荷表(取前六個主成分)

水質變項	水質變項對於主成分負荷					
	PC1	PC2	PC3	PC4	PC5	PC6
pH	-0.14	0.14	-0.47	0.27	-0.10	0.04
EC	0.34	-0.27	-0.12	0.05	-0.08	0.21

doi:10.6342/NTU201803135

水質變項	水質變項對於主成分負荷					
	PC1	PC2	PC3	PC4	PC5	PC6
DO	-0.31	-0.04	-0.18	-0.15	0.00	0.28
BOD	0.36	-0.07	0.18	0.34	-0.09	-0.04
COD	0.40	-0.01	0.03	0.26	0.01	0.00
TP	0.09	-0.25	-0.38	-0.08	-0.13	-0.45
NH_3-N	0.32	-0.24	-0.10	-0.02	-0.02	-0.21
NO_3-N	0.12	-0.30	-0.28	-0.44	-0.01	0.33
SS	0.15	0.48	-0.36	0.07	0.17	0.09
Hg	0.11	0.38	-0.25	0.15	0.24	-0.08
Mn	0.30	0.31	-0.16	-0.22	0.13	0.13
Cd	0.03	0.28	0.00	-0.28	-0.68	0.07
Ag	0.22	0.28	0.10	-0.12	-0.49	-0.15
As	0.10	-0.20	-0.39	-0.12	0.01	-0.27
Pb	0.08	0.17	0.24	-0.38	0.27	-0.49
Cu	0.25	0.06	0.13	-0.43	0.28	0.20
Zn	0.32	-0.05	0.10	0.13	0.01	0.33
Stand.dev	2.19	1.40	1.29	1.19	1.09	0.98
特徵值	4.77	1.96	1.67	1.43	1.19	0.95
可解釋變異量	0.28	0.12	0.10	0.08	0.07	0.06
累積解釋變異量	0.28	0.40	0.49	0.58	0.65	0.70

選擇主成分/因素數目最常用的方法兩種，第一種為選擇特徵值大於 1 的主成分數目，另一個常用的方法為觀察陡坡圖，尋找斜率突然變緩當下的因素數目，由下圖 4- 13 可看出從因素 6-7 所構成的斜率略較因素 5-6 構成的斜率平緩，以陡坡圖選擇因素則以 6 個為宜。考量第 6 的主成分其特徵值仍有 0.95，相當接近 1，爰選擇 6 個主成分做後續分析。

doi:10.6342/NTU201803135

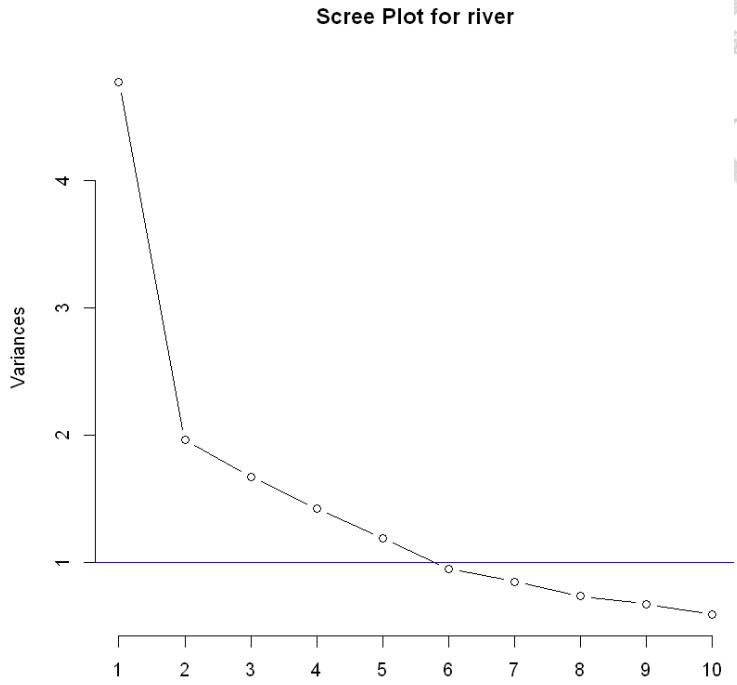


圖 4- 13　主成分分析陡坡圖(Scree Plot)

進行轉軸後之新產生因素說明

因素分析可將原本不明顯的負荷值分出明顯區別，負荷小的變項會被排除因素組成。設定轉軸因素為 6 個，經最大變異正交轉軸法(Varimax)轉軸後，萃取出的 6 個因素可解釋資料及總變異量的 70%，詳如表 4- 6。以下就萃取出的因素推測其由來：

一、第 1 因素可解釋整體資料集 23%的變異量，與水質參數中的 BOD、COD、Zn、EC、DO 及 NH_3-N 等相關，BOD、COD 及 NH_3-N 為常見污染物，來源很多，可能由工廠排水、畜牧業或是家庭污水排放所致，鋅及導電度推測可能來源為金屬表面處理業使用鋅做為防鏽所致，上述有機污染源排入河川造成水中溶氧下降，導致溶氧的負荷值為負相關，綜上述原因將第 1 因素命名為複合污染物。(桃園縣政府環保局, 2011; 楊于嫻, 2014)

doi:10.6342/NTU201803135

二、第 2 因素可解釋整體資料集 12%變異量，與之相關的水質參數有 SS、Hg 及 Mn，Mn 為地殼中常見成分，推測河水中 Mn 濃度及 SS 升高，可能因降雨沖刷導致，Hg 之測值有 90%以上均低於偵測極限，其變動趨勢與 SS 及錳相符。

三、第 3 因素可解釋整體資料集 9%變異量，與之相關的水質參數有 TP 和 As，這 2 種污染物推測可能來自工業廢水，例如電子業半導體製造過程中，砷化鎵是製作半導體材料，而清洗廢水中包含磷酸、硫酸及硝酸等廢水。(經濟部工業局, 1993)

四、第 4 因素可解釋整體資料集 9%變異量，與之相關為 NO_3-N 及 Cu，這 2 種污染物推測可能來自工業廢水，例如印刷電路板製作過程排出，Cu 為印刷電路之原料，而硝酸為清洗電路板之常用溶劑之一。(吳孟育, 2005)

五、第 5 因素可解釋整體資料集 9%變異量，與之相關者為 Pb、pH 及 Cu。與 Pb 為高度正相關，與 Cu 及 pH 為中度相關，又與 pH 相關性為負相關，亦即 Pb 及 Cu 濃度上升時 pH 將下降，Pb、Cu 及酸為工業常用原料，因此將因素命名為工業用金屬材料。

六、第 6 因素可解釋整體資料集 8%變異量，有關者為 Cd 和 Ag，此 2 項金屬濃度測值低於偵測極限者，均占總測量點數的 92%，水中的 Cd 可能來自電鍍或塑膠業，Ag 可能來自導電材料製作或化工用催化劑等，因此亦將因素命名為工業用金屬材料。

doi:10.6342/NTU201803135

表 4- 6　因素分析各變項負荷表

水質變項	水質變項對於旋轉主成分(因素)之負荷					
	RC1	RC2	RC3	RC4	RC5	RC6
BOD	**0.92**	0.02	0.01	-0.09	0.03	0.06
COD	**0.89**	0.23	0.09	0.06	0.06	0.04
Zn	**0.72**	0.09	-0.13	0.3	-0.02	0.02
EC	**0.71**	-0.03	0.28	0.42	-0.14	0
DO	**-0.68**	-0.11	-0.08	0.16	-0.31	-0.1
NH_3-N	**0.61**	-0.01	0.48	0.19	0.16	0.01
SS	0.08	**0.89**	0	0.06	-0.04	0.12
Hg	0.09	**0.72**	0.02	-0.12	0.03	-0.04
Mn	0.33	**0.61**	0.03	0.39	0.27	0.23
TP	0.07	-0.03	**0.78**	0.01	-0.03	0.03
As	0.05	0.09	**0.66**	0.16	-0.02	-0.06
NO_3-N	0.01	-0.1	0.36	**0.78**	-0.05	0.01
Cu	0.27	0.17	-0.1	**0.58**	**0.52**	0.03
Pb	-0.05	0.1	0.06	-0.06	**0.82**	0.05
pH	-0.3	0.37	0.2	-0.19	**-0.56**	-0.02
Cd	-0.11	0.02	-0.04	0.08	-0.06	**0.90**
Ag	0.33	0.14	0.01	-0.06	0.18	**0.74**
特徵值	3.92	1.98	1.57	1.54	1.53	1.44
可解釋變異量	0.23	0.12	0.09	0.09	0.09	0.08
累積解釋變異量	0.23	0.35	0.44	0.53	0.62	0.70
因素來源推測	複合污染物(工廠廢水、畜牧業或家庭污水)	降雨沖刷	工業廢水排放(列如半導體業)	工業廢水排放(列如印刷電路板業)	工業金屬材料(鉛銅酸)	工業金屬材料(鎘及銀)

doi:10.6342/NTU201803135

因素與採樣點間關係及污染源推測

得到各因素之變項負荷表後，可將各樣本數值轉變為因素分數，新產生之 6 個因素將取代原本 17 個水質變項，以因素分數描述各採樣點位之水質變化，意即每個採樣點的特徵由 17 個水質變數減少成以 6 個新產生的因素代表。所有樣本以採樣點位作為橫軸，將新產生因素於各點位分數變化以盒方圖呈現如下圖 4- 14，以下說明各因素在不同採樣點位的分數變化趨勢。

一、因素 1 複合污染物於 S7 許厝港一號橋及 S3 平鎮第一號橋分數偏高，這 2 個點位亦是大園工業區(影響 S7)、科技園區龍潭廠(影響 S3)及平鎮工業區(影響 S3)的工業排水的接收點位，S7 點位同時接受了田心仔溪沿岸畜牧廢水的匯入(桃園縣政府環保局, 2011; 楊于嫻, 2014)。另發現因素 1 在各點位分數的分布與 RPI 指標分布趨勢相同(對照 2-1 節圖 2- 2)，均為 S7 許厝港一號橋(下游)>S3 平鎮工業區(支流)>S1 美都麗橋=S4 環鄉橋>S2 北勢橋=S5 公園橋=S6 中正橋，可由此得知除了 RPI 本身成員(BOD、DO、NH_3-N)外，EC、COD 及 Zn 與 RPI 指標之變化趨勢相近。

二、因素 2 降雨沖刷的分數，在各採樣點間無明顯差異，各點位均出現少數高濃度離群值，經查多是因為極高的 SS 濃度所致。

三、因素 3 推測與工業廢水排放相關(如半導體產業廢水)，對照各行政區列管產業，電子零組件製造業多分布於平鎮區及中壢區。本因素在 S3 平鎮第一號橋的分數明顯較其他點位高，經查為高濃度 As 所致，污染可能來自平鎮區的半導體產業污染排放。

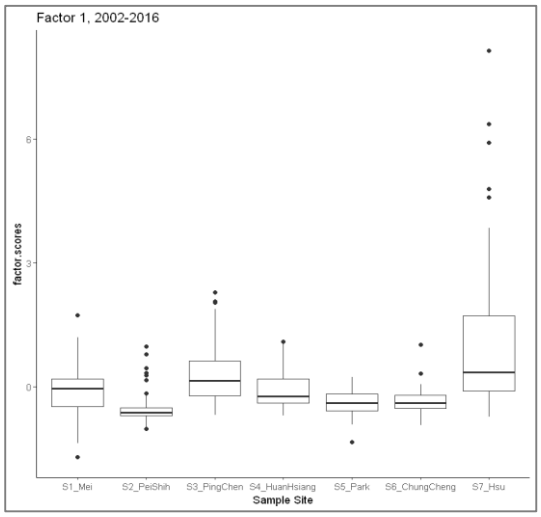
四、因素 4 推測與工業廢水排放相關(如印刷電路板產業廢水)，對照各行政區列管產業，印刷電路板產業亦多分布於中壢區及平鎮區。本因素在 S3 平鎮一號橋(平鎮區)、S7 許厝港一號橋(大園區)及 S2 北勢橋(平鎮區)有較高的分數，推測可能為上游中壢區及平鎮區之印刷電路板業污染排放所致。

doi:10.6342/NTU201803135

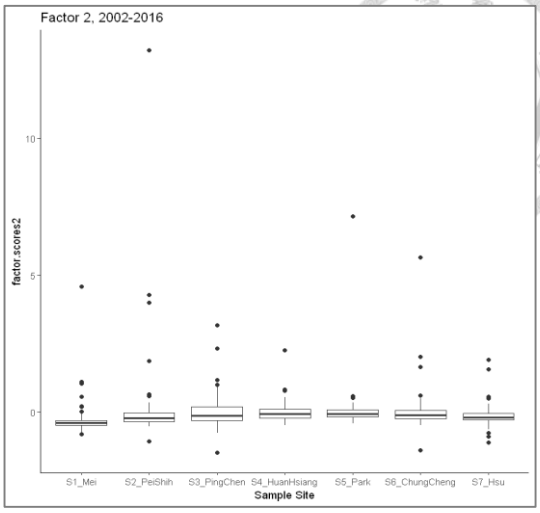
五、因素 5 為 Pb、pH 及 Cu 等工業製程常見金屬材料，本因素在 S1 美都麗橋有較高的平均分數，經查主要受高濃度 Pb 所致，另外 S7 許厝港一號橋出現較多高濃度離群值。

六、因素 6 為 Cd 及 Ag 等金屬，這 2 金屬測值多低於偵測極限，且在各採樣點之分數無明顯差別。

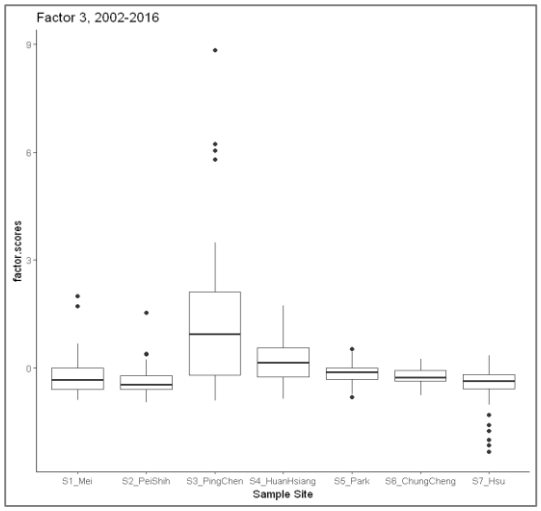
doi:10.6342/NTU201803135



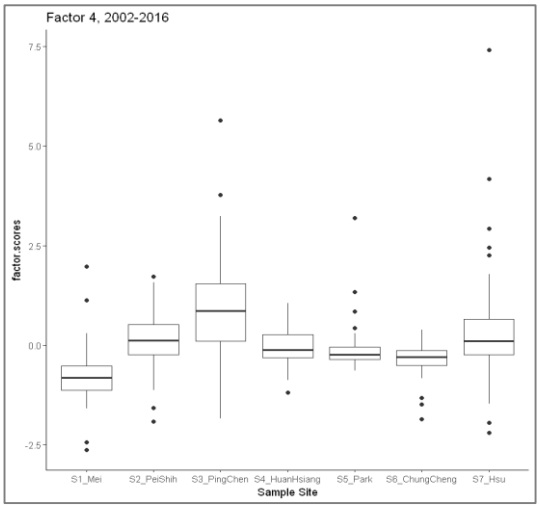
因素 1　複合污染物
(BOD、COD、Zn、EC、DO、NH_3-N)



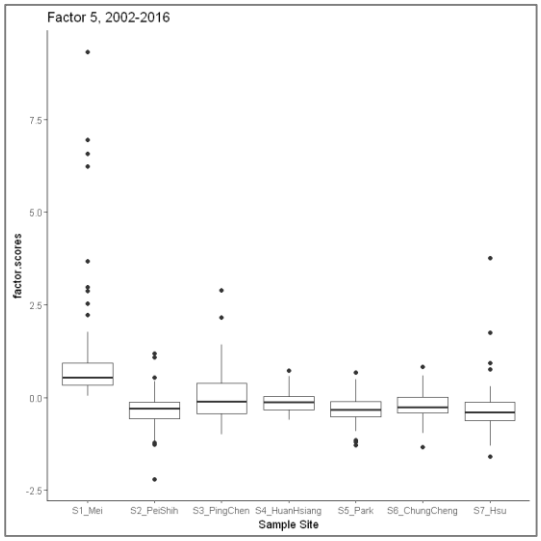
因素 2　降雨沖刷
(SS、Hg、Mn)



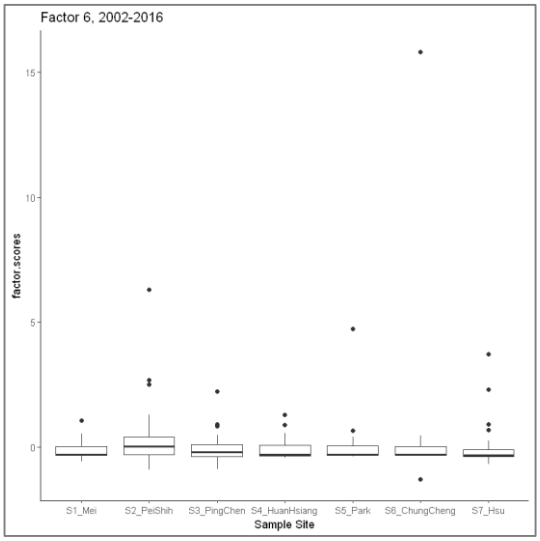
因素 3　工業廢水排放如半導體業
(TP、As)



因素 4　工業廢水排放如印刷電路板業
(Cu、NO_3-N)



因素 5　工業金屬材料(Pb、pH、Cu)



因素 6　工業金屬材料(Cd、Ag)

圖 4- 14　因素分數於各採樣點位之分數圖

doi:10.6342/NTU201803135

因素分數隨時間的變化

所蒐集樣本採樣期間自 2002 至 2016 年為止，以 5 年為單位區分樣本成 2002-2006 年、2007-2011 年及 2012-2016 年等 3 個群體，將各因素分數兩兩組合成散布圖如圖 4- 15，觀察各因素分數隨時間之變化趨勢。

一、因素 1(複合污染物)與因素 2(降雨沖刷)：

在 2002 至 2006 年期間，複合污染物和降雨沖刷分數有較多的高濃度離群值，隨著時間的推進，這兩項因素分數都呈現減少的趨勢，高濃度離群值數目也大幅降低，推測可能因 2008 年起推動老街溪河川整治計畫，及總量管制標準訂定公告，使污染物排放被控制。

二、因素 3(工業廢水排放如半導體業)與因素 4(工業廢水排放如印刷電路板業)：

工業廢水排放(如半導體業)污染分數隨時間推進並無明顯變化，工業廢水排放(如印刷電路板業)污染分數在縱軸分布廣，表示各樣本分數差異大，隨著時間推進，高分群點位(多屬於 S7 及 S3 點位)的分數下降，低分群點位(多屬於 S1 點位)分數上升，點位分布有集中之趨勢，推測可能因為印刷電路板業的污染量排放改善，使高分群分數降低；但原為低分群可能因印刷電路板業進駐，而導致污染分數增加。

三、因素 5(工業金屬材料鉛銅酸)、因素 6(工業金屬材料銀及鎘)：

工業用金屬中鉛銅酸因素分數分布範圍廣，鎘及銀因素分布範圍雖然較一致，但時常出現高濃度離群值(多出現於 S2 點位)，隨時間推進，2 個因素的分數在採集到樣本中均漸降低且集中，在因素 5 工業用金屬(鉛銅酸)因素中仍可看出不同採樣點位之分數有差異(高分點位多位於 S1)。

doi:10.6342/NTU201803135

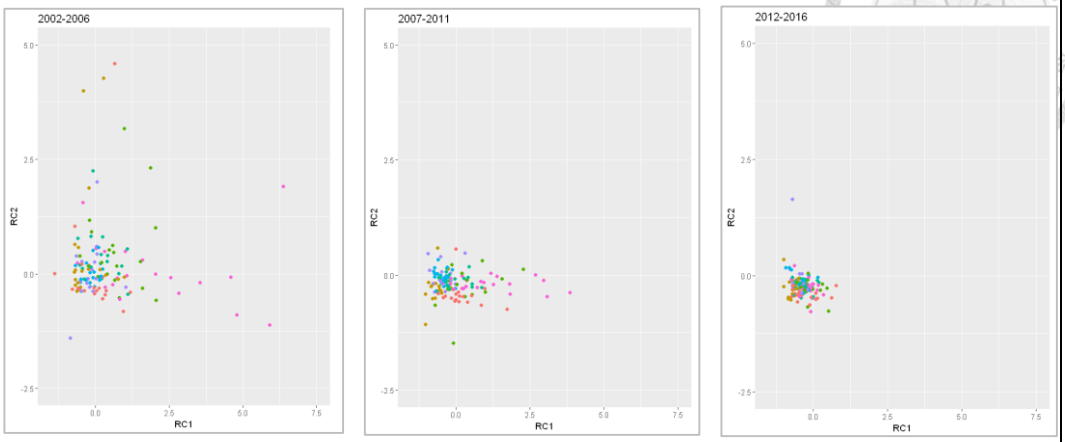
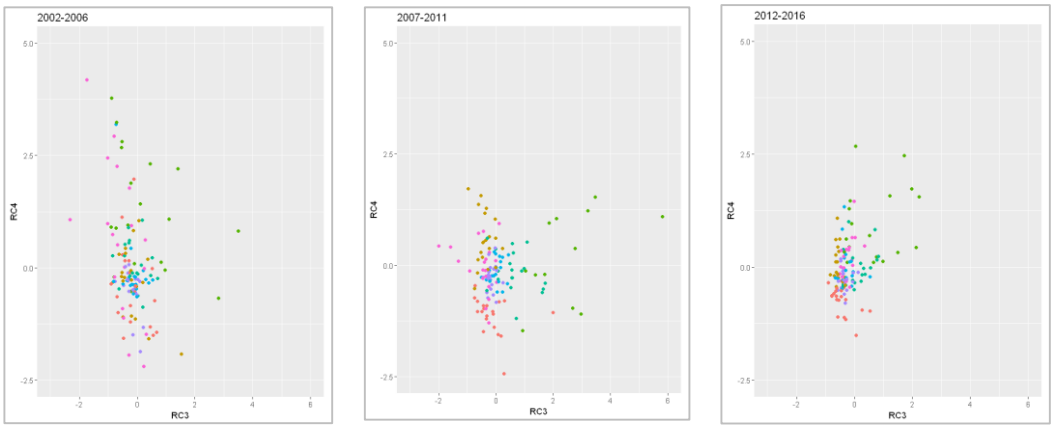
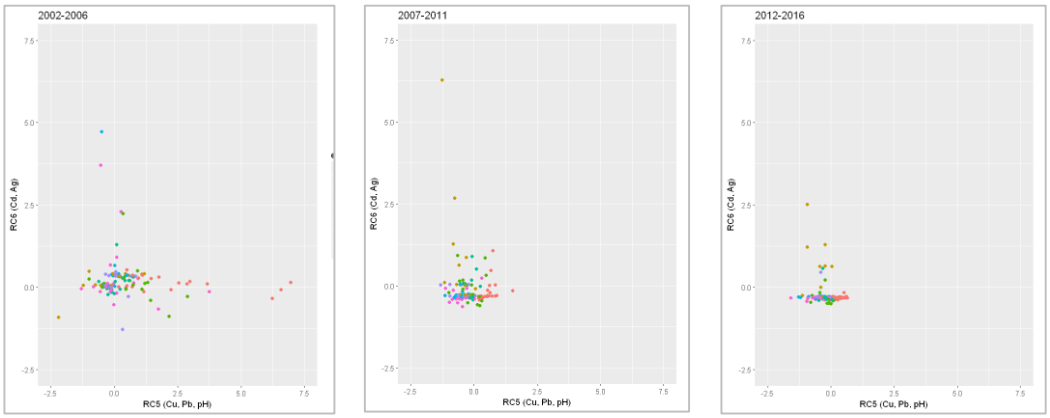
	2002-2006 年	2007-2011 年	2012-2016 年
因素1及2			
	橫軸(因素 1)：複合污染物 縱軸(因素 2)：降雨沖刷		
因素3及4			
	橫軸(因素 3)：工業廢水(如半導體業) 縱軸(因素 4)：工業廢水(如印刷電路板業)		
因素5及6			
	橫軸(因素 5)：工業金屬材料(鉛銅酸) 縱軸(因素 6)：工業金屬材料(銀鎘)		

圖 4- 15　因素分數隨時間變化圖

doi:10.6342/NTU201803135

4.2.2 群集分析

蒐集 2002 至 2016 年之水溫、pH、EC、DO、BOD、COD、TP、NH_3-N、NO_3-N、SS、Hg、Mn、Cd、Ag、As、Pb、Cu、Zn、Cr^{6+}、Coliform 等 20 個參數，計算總平均製作群集分析樹枝圖如圖 4- 16，另外製作群集分布對應採樣點位示意圖如圖 4- 17，示意圖中呈現採樣點位與地理位置現況之關係圖，老街溪流向為由上游 S1 美都麗橋點位流至 S7 許厝港一號橋點位，支流大坑缺溪採樣點位於 S3 平鎮第一號橋，之後匯流進入 S4 環鄉橋上游。

分析 15 年間各點位各水質參數平均之群集關係，可分出 3 個群集，S3 及 S7 為第一個群集，S1 獨自成立一個群集，剩下來 S2、S4、S5 及 S6 群組成同一群集，觀察群集間差異，可看出 S3、S7 均為嚴重污染的工業區排放口下游，S3 上游為竹科龍潭園區及平鎮工業區，S7 採樣點上游為大園工業區，S1 採樣點老街溪流域為最上游的採樣點，剩餘 S2、S4、S5 及 S6 點位位於河流中段。

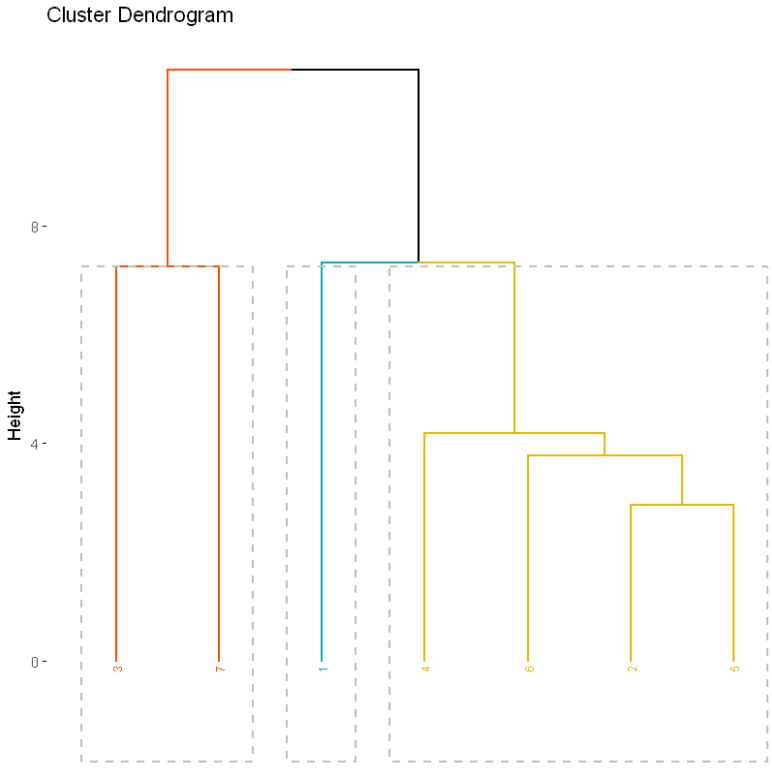


圖 4- 16　群集分析樹狀圖(2002-2016 年)

doi:10.6342/NTU201803135

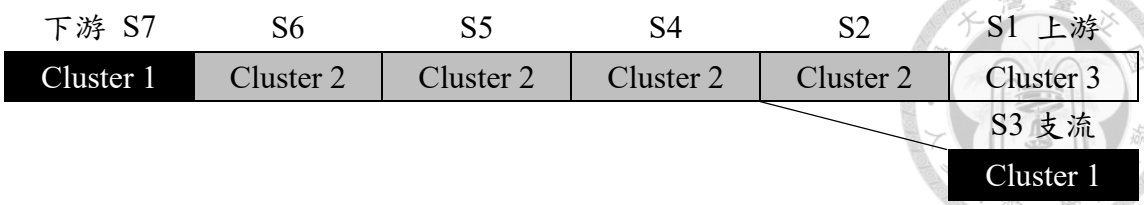


圖 4- 17　群集分布對應採樣點位示意圖(2002-2016 年)

另外考量 15 年間水質是一直有在變動,以下再細分以分 5 年平均做群集分析,其樹狀圖及示意圖如圖 4- 18。可看出支流 S3 平鎮第一號橋測點，不論何時都自成 1 個群集，變化最大的地方在於群集 2 及群集 3 間的消長，群集 3 始終包含上游點位，本研究把他歸納為較無污染點群集，群集 2 一開始即包含大園工業區下游的點位 S7，本研究把他歸納為有污染的點位，群集成員組成隨時間變化說明如下：

一、在第 1 個 5 年(2002 至 2006 年)期間，群集 3(上游群組)成員包括除了支流 S3 和下游 S7 點為外的剩餘 5 個點位，表示除了工業區排放口下游的 S3 及 S7 點位外，其餘點位水質相似。

二、但到了第 2 個 5 年(2007 至 2012 年)，支流的匯入點 S4 環鄉橋離開群集 3(上游群組)，被歸入了群集 2(下游群組)。

三、而到了最近 5 年(2012 至 2016 年間)，群集 3(上游群組)中的 S5 及 S6 點位亦轉移至群組 2(下游群組)。

就近 15 年的變化來看，較無污染的群集 3(上游群組)成員由 5 個降為 2 個，持續在群組 3(上游群組)的點位者有 S1 及 S2，中游三個點位 S4、S5 及 S6 分批轉入群組 2(下游群組)，顯示出隨時間推進，這些中游點位的水質與下游水質漸趨相似之趨勢。推測造成此現象的原因，可能與高度污染的支流水質(S3)有關，支流大坑缺溪於採樣點位 S4 上游地點匯入主要河川，水中高濃度污染物逐漸影響下游水質，S4 點位於 2007 至 2011 年區間首當其衝被影響，分類轉變為下游群組，接連著最近 5 年則是 S5 和 S6 亦被納入了下游群組。

doi:10.6342/NTU201803135

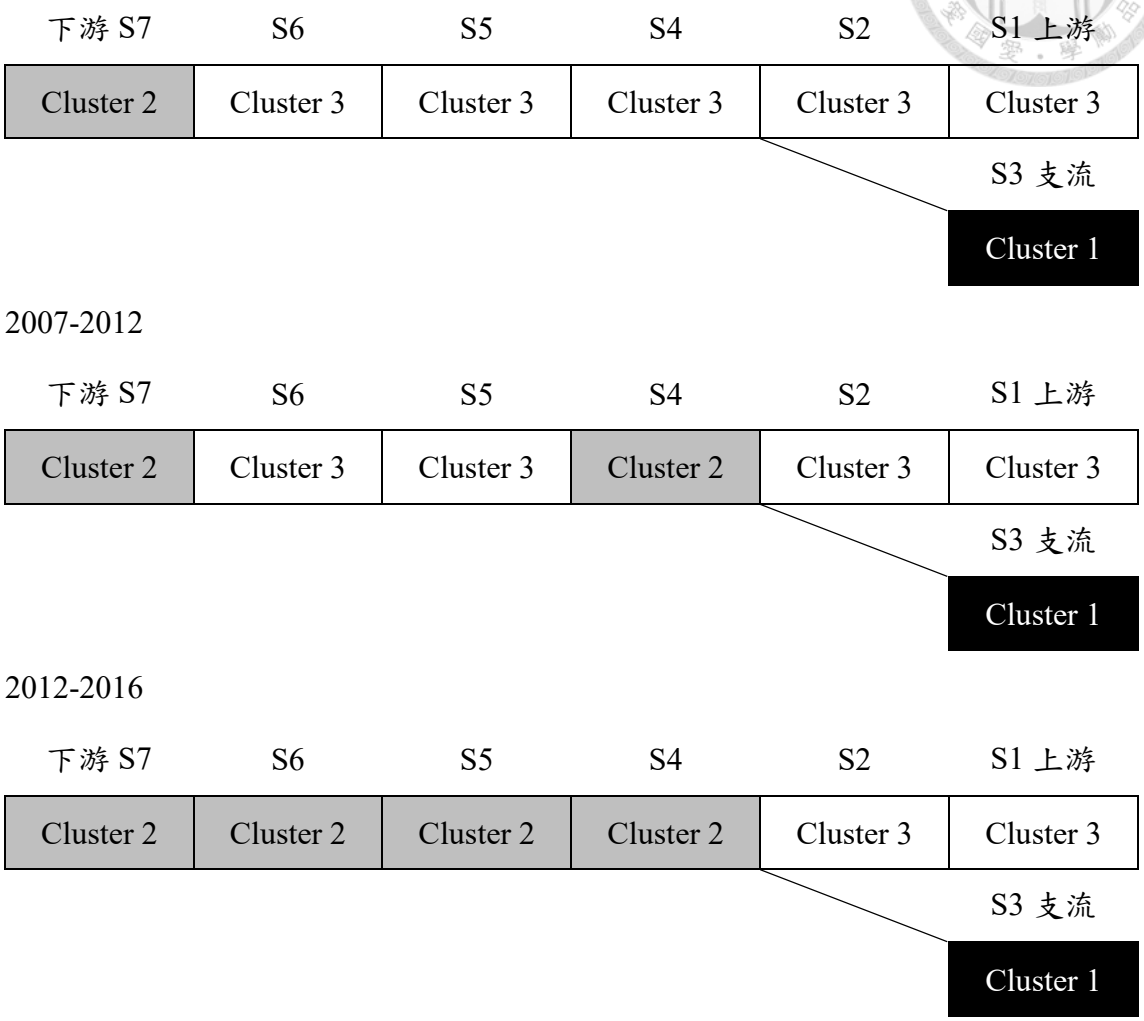


圖 4- 18　2002-2016 年間每 5 年群集分布狀況

doi:10.6342/NTU201803135

4.3 機器學習

機器學習研究分為兩主題：第一個主題為「以例行量測數值預測金屬銅濃度超標可行性」，第二個主題為「以 COD 代替 BOD 預測水質污染分類指標(RPI)之可行性」。本研究將使用 2002 年至 2016 年之水質監測資料為基礎，建立模型，並評估後，再以 2017 年 1 月至 2018 年 4 月期間共計 16 個月之數據，實際測試模型效果。

4.3.1 以每月例行量測水質參數判斷水中銅濃度超標可行性

水中高濃度的金屬對人體有害，據此水污染防治法訂定保護人體健康相關環境基準，水中金屬測值是否超過規定標準是大眾關切的重點，但並非每個月都有觀測數值，目前仍是以每季 1 次的頻率監測河川水質重金屬濃度。老街溪流域水中金屬濃度，除了錳以外，銅的超標頻率為最高的，詳最見前表 4-4。又依據桃園市政府環境保護局表示，錳為地殼中主要元素，大多以非溶解性之氧化態存在，自然水體中含量本高，一般河川水質超標比率不低，故以超標率第 2 高的金屬銅作為判斷目標。

銅金屬在法規中限值為 0.03 mg/L，表 4- 7 為 2002-2016 年間水中銅濃度在各點位被量測到超標的次數，可以發現在 S3 點位量測到的銅數值持續超標。分年度觀察其超標狀況，可以發現雖然水中銅濃度超標率雖逐年下降，2016 年年度總超標率仍有 21％，尤其 S3 點位仍為四季均超標情形，S4 及 S7 點位也有一季被測得超標，本研究將測試以模型判斷金屬銅濃度超標與否之構想之可行性。

表 4- 7　各測站水中銅濃度歷年超標點次及比率統計

年度	測量點次(次數)	各測站水中銅濃度超標點次(次數)								年度總超標率
		S1	S2	S3	S4	S5	S6	S7	總計	
2002	28	4	3	4	4	4	4	4	27	96%
2003	28	3	2	4	4	4	4	4	25	89%
2004	28	4	3	4	4	4	4	4	27	96%
2005	28	4	4	4	4	4	4	4	28	100%

doi:10.6342/NTU201803135

年度	測量點次(次數)	各測站水中銅濃度超標點次(次數)								年度總超標率
		S1	S2	S3	S4	S5	S6	S7	總計	
2006	28	2	4	4	4	4	4	4	26	93%
2007	28	0	4	4	4	2	2	4	20	71%
2008	28	0	4	4	4	3	1	4	20	71%
2009	28	0	4	4	4	1	1	4	18	64%
2010	28	0	4	4	4	3	1	4	20	71%
2011	28	0	4	4	4	3	2	4	21	75%
2012	28	0	2	4	3	3	1	2	15	54%
2013	28	0	1	4	4	1	1	4	15	54%
2014	28	0	0	4	4	2	0	3	13	46%
2015	21	0	0	3	1	1	0	1	6	29%
2016	28	0	0	4	1	0	0	1	6	21%
總計	413	17	39	59	53	39	29	51	總超標 287 點次	
各點位總超標率		29%	66%	100%	90%	66%	49%	86%	總超標率 69%	

模型建置步驟

藉由機器學習方法，以每月都會測量的水質測項，判斷水中銅濃度是否超標。輸入資訊為點位、溫度、酸鹼值、溶氧量、生化需氧量、氨氮、生化需氧量、懸浮固體及大腸桿菌群等數值，再將水中銅濃度以 0.03 mg/L 為界，分為超標與沒有超標 2 個族群，超標與否的二元類別變項即作為預測目標。

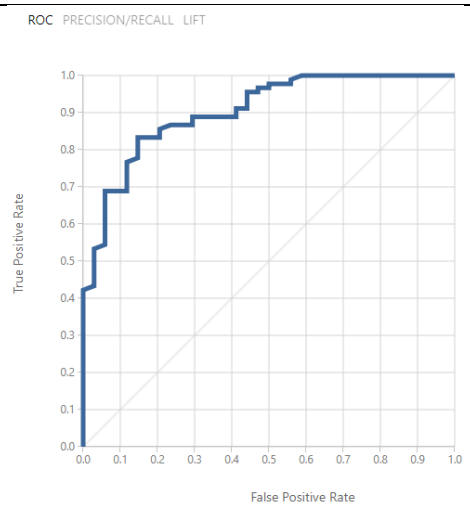
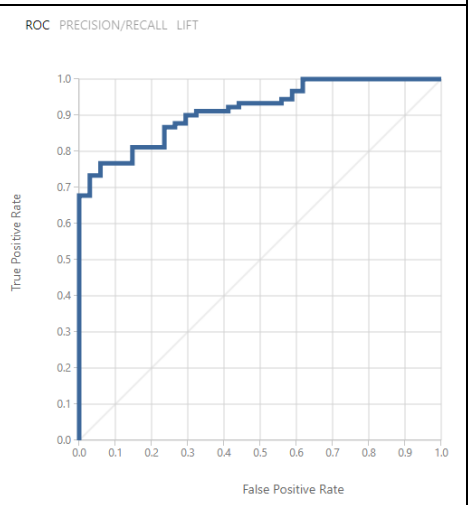
413 個樣本經過標準化後，即將數據及分成 289 筆(70%)資料用作訓練，124 筆資料(30%)用作測試集後，將選定訓練模型以訓練集訓練後得到 ROC 曲線下面積及正確率等數值，以評估模型可用性。

模型效果評估

預測標的為二元分類變項，本研究使用的方法以二元神經網路(Two-Class Neural Network)及二元決策森林(Two-class Decision Forest)為主。模型預測結果評估如下表 4- 8：

doi:10.6342/NTU201803135

表 4- 8　判斷水中銅濃度超標模型之效果評估

模型 / 指標	二元決策森林模型 Two-class Decision Forest			二元神經網路模型 Two-Class Neural Network		
判斷評估表格	預測 / 實際	超標	未超標	預測 / 實際	超標	未超標
	超標	78	12	超標	80	10
	未超標	9	25	未超標	9	25
Accuracy	0.831			0.847		
Precision	0.897			0.899		
Recall	0.867			0.889		
F1 Score	0.881			0.881		
ROC						
AUC	0.903			0.923		

決策樹(森林)構造解讀

可由列表看出類神經網路模型之評估指標值均略高於決策森林，顯示效果較佳，惟神經網路之判斷邏輯可視為黑盒子，無法解構。二元決策森林模型是以分層分枝方法，建立分類流程規則以達到判斷金屬銅超標與否之目標，越上層節點所用的水質參數對於判斷水中銅是否超標越重要。

決策樹之示意圖如圖 4- 19，以本次產生的模型為例，抽樣決策森林中 10 棵決策樹，第 1 層節點的水質參數為 SS 者機率最高，10 個節點中占 6 個，第 2 層節點

doi:10.6342/NTU201803135

之水質參數以 EC 和是否為 S3 點位之判斷節點各占 4 個機率最高、其次為 SS 及是否為 S1 點位之判斷節點各占 3 個為次高，詳如表 4- 9。

總結來說，對於銅濃度超標與否之判斷，SS、EC、是否為點位 3 或點位 1 等參數為重要的判斷依據。

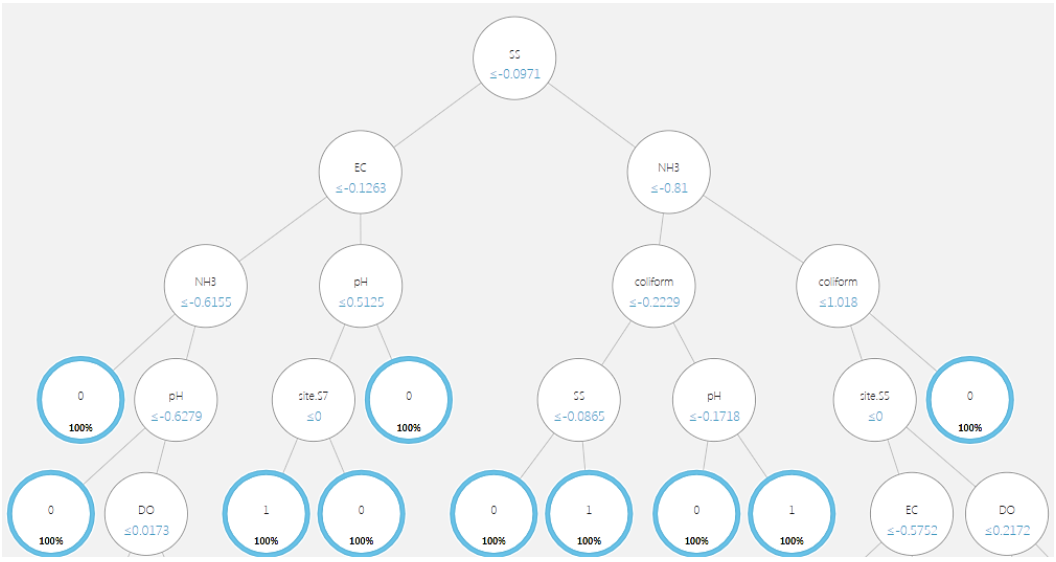


圖 4- 19　判斷銅濃度超標之決策樹示意圖

表 4- 9　判斷銅濃度超標之決策樹節點內容

層級	節點數目							
	SS	BOD	EC	Site3	Site1	NH_3-N	Coliform	水溫
第一層(共 10 節點)	6	1	2	0	0	1	0	0
第二層(共 20 節點)	3	2	4	4	3	2	1	1

doi:10.6342/NTU201803135

實作—判斷水中金屬濃度超標與否

上一節建置完成之類神經網路及決策森林模型，其正確率及 AUC 等指標效果不差，遂使用 2017 年至 2018 年 4 月年監測數據輸入此 2 模型，再次進行評估及判斷。先以監測數值有包含銅濃度之 42 筆數據輸入模型，比較本節與上節之預測正確率；再以剩餘 10 個月沒有包含銅濃度的監測數值輸入模型，討論其結果，及兩模型判斷的一致性。

一、使用已有銅濃度數據樣本驗證模型效果

先以有銅濃度測值之 6 個月的數據做驗證，總數據量為 42 筆，超標數為 14 筆，未超標數 28 筆。以決策森林預測銅濃度超標與否，分類正確筆數有 37 筆，錯誤筆數又 5 筆，正確率為 88%，而類神經網路之分類正確率為 74%，整體來說決策森林方法表現較好，判斷結果詳如表 4- 10。

表 4- 10　兩模型判斷銅濃度超標結果比較-實際測試(2017-2018)

方法	二元決策森林模型		二元神經網路模型	
預測狀況 / 實際狀況	超標	未超標	超標	未超標
超標	13	1	13	1
未超標	4	24	10	18
正確率	37/42=88%		31/42=74%	

決策森林判斷失誤之點位

模式分類為未超標，實際上是超標者有 1 筆

測站名稱	時間	RPI	水溫	pH	EC	DO	BOD	COD	SS	Coliform	NH_3-N	Cu
S5 公園橋	201804	3.5	27.8	8.11	1280	10	6.8	26.9	14.8	4400	1.85	0.044

模式分類為超標，事實上未超標者有 4 筆。

測站名稱	時間	RPI	水溫	pH	EC	DO	BOD	COD	SS	Coliform	NH_3-N	Cu
S1 美都麗橋	201804	4.5	19	7.01	278	6.7	6.4	22.3	13.3	68000	4.68	0.003
S2 北勢橋	201709	5	31	6.7	596	3.2	45.5	107	23.5	6000000	0.06	0.007
S4 環鄉橋	201709	4.3	31	7.5	1050	5.9	4.7	32	18.2	180000	3.73	0.027

doi:10.6342/NTU201803135

測站名稱	時間	RPI	水溫	pH	EC	DO	BOD	COD	SS	Coliform	NH_3-N	Cu
S7 許厝港	201709	3.5	33.3	8.6	2270	6.7	6.1	36.8	13.6	67000	2.55	0.026

神經網路判斷失誤之點位

模式分類為未超標，實際上是超標者有 1 筆，該筆數據與決策森林模型判斷錯誤為同一筆。

測站名稱	時間	RPI	水溫	pH	EC	DO	BOD	COD	SS	Coliform	NH_3-N	Cu
S5 公園橋	201804	3.5	27.8	8.11	1280	10	6.8	26.9	14.8	4400	1.85	0.044

模式分類為超標，事實上未超標者有 10 筆，其中環鄉橋 2017 年 9 月數據與決策森林模型同樣判斷錯誤，其餘判斷失誤樣本與決策森林模型均不同。

測站名稱	時間	RPI	水溫	pH	EC	DO	BOD	COD	SS	Coliform	NH_3-N	Cu
S2 北勢橋	201804	1.5	21.4	7.73	483	10.3	4.5	18.9	14.9	13000	0.3	0.009
S2 北勢橋	201706	1.5	32.8	8.8	434	10.2	3.1	18.7	10.6	6100	0.15	0.007
S2 北勢橋	201703	1	16.6	7	498	9.6	2.9	13	9.8	68000	0.33	0.013
S4 環鄉橋	201709	4.3	31	7.5	1050	5.9	4.7	32	18.2	180000	3.73	0.027
S4 環鄉橋	201706	2.8	31.9	7.8	972	6.9	3.1	24.1	7.6	45000	2.02	0.025
S5 公園橋	201801	2.3	16	7.62	622	9.6	2.5	17.7	11	24000	1.23	0.019
S5 公園橋	201703	2	18.5	7.4	1200	10.6	3.7	17	7.1	37000	0.93	0.028
S6 中正橋	201709	2.5	34.3	7.8	710	5.2	4.6	27.9	11.4	7500	0.53	0.010
S6 中正橋	201703	2.8	17.6	7.3	709	8.6	5.3	24.7	9.1	64000	0.8	0.012
S7 許厝港	201801	2.8	15.5	7.6	1080	8.9	4.7	23.1	8.3	18000	1.92	0.028

決策森林判斷失誤點位為 5 筆，神經網路判斷失誤樣本為 11 筆，其中有 2 筆是 2 模型均判斷錯誤的，其餘判斷失誤樣本均不相同。此結果顯示各種模型對於不同的水質狀況判斷的效果不一，但當 2 模型判斷結果一致時，判斷失誤的機率大幅下降，以本例來說，2 模型針對同一樣本，判斷結果相同的點位數有 30 筆，其中失誤者僅占 2 筆(6.7 %)。

決策森林在建置模型時預估準確率與實測準確率差異原因探討

doi:10.6342/NTU201803135

決策森林模型於測試銅濃度有無超標模型，原本建模評估正確率 0.85，實測時正確率僅 0.74。推測為早期檢測數據不穩定而影響模型建置，爰捨去 2008 年(含)以前之數據，以近 8 年資料作為建模依據，建置模型數據及測試數據分別為 152 筆及 65 筆，總共 217 筆數據，較 15 年份數據少了 196 筆。以 2009 年(含)以後之數據建置模型結果如下，銅濃度超標判斷之正確率為 0.83，雖然相較先前少了 2 個百分比，但實測時正確率達到 0.81，建模預估之準確率與實測之準確率差距變小，顯示 2008 年以前不穩定之檢測數據可能影響模型建置。

二、判斷未包含銅濃度數據的樣本比較 2 模型判斷結果一致性

2017 年至 2018 年 4 月間歷經 16 個月，其中 6 個月份有銅濃度的監測數值，用於上一節測試模型使用，其餘 10 個月份並無水中銅濃度測項，本研究嘗試將 10 個月中 7 測站各蒐集的 10 個樣本分別以 2 種模型中執行，下表 4- 11 顯示 2 模型判斷結果及判斷結果為一致性彙整。

表 4- 11　兩模型判斷銅濃度超標結果比較-推測(2017-2018)

點位	決策森林判斷超標點位個數	類神經網路判斷超標點位個數	判斷結果為一致之點位數	判斷結果一致且結果為超標點位
S1	1	0	9	0
S2	4	5	7	3
S3	10	10	10	10
S4	9	9	8	8
S5	5	5	10	5
S6	4	5	7	3
S7	9	6	7	6
總計	42	40	58	35

因為本群數據無銅濃度測值可驗證預測結果是否正確，在此僅討論兩模型判斷結果的相似性，依據上一段落的結論，當 2 模型有一致的結果時，錯誤機率大幅下降，倘需將本模型結果實際應用在採樣輔助上，建議優先考慮表 4- 11 末欄「判斷結果一致且結果為超標點位」中 35 個樣本之採樣或驗證。

doi:10.6342/NTU201803135

4.3.2 以 COD 代替 BOD 判斷水質污染指標(RPI)之可行性

河川污染程度指標(RPI)是重要評估河川水質的參考指標，其計算方式為溶氧量、懸浮固體、氨氮及生化需氧量的點數。但其中生化需氧量測定需耗費 5 日實驗時間，造成無法即時得到污染指數之狀況。本研究嘗試以 COD 取代 BOD 指標，優化 RPI 指標取得即時性。

數據蒐集及建置流程

1,260 個樣本經過標準化及將數據集分成 882 筆(70%)資料用作訓練，378 筆資料(30%)用作測試集，後以決策森林回歸方法(Decision Forest Regression)及神經網路回歸方法(Neural Network Regression)建立判斷模型，並以平均絕對誤差(Mean Absolute Error, MAE)、根均方誤差(Root Mean Squared Error, RMSE)、平均絕對誤差百分比(Mean Absolute Percentage Error, MAPE)及決定係數(Coefficient of Determination, R^2)等指標作為模型效果判斷依據。

模型效果評估

由測試集 378 筆數據測試決策森林回歸模型判斷 RPI 數值之效果，由分布圖(圖 4- 20)可看出實際值與判斷值具明顯線性關係，且決定係數接近 1，代表判斷值不會離實際值太遠。類神經網路回歸模型部分，由分布圖可看出實際值與判斷值不具明顯線性關係。在實際 RPI 數值小於 3 的點位(無污染或輕度污染)，出現高估狀況，但對於實際 PRI 值大於 4 的點位(中度污染或重度污染)，判斷值則多出現低估現象。

下表 4- 12 顯示 2 種模型的評估指標值，經比較後發現決策森林回歸模型的表現為最好，誤差值為最小，顯示模型判斷值與實際值差異不大，而決定係數接近 1，可以看出實際值與判斷值間具有線性關係，模型的預測能力高。

doi:10.6342/NTU201803135

表 4- 12　兩模型推估 RPI 水質指數效果評估

評估模型指標	決策森林回歸模型 (Decision Forest Regression)	神經網路回歸模型 (Neural Network Regression)
平均絕對誤差(MAE)	0.352	0.818
根均方誤差(RSME)	0.464	1.064
平均絕對誤差百分比(MAPE)	0.087	0.235
決定係數(R^2)	0.929	0.643

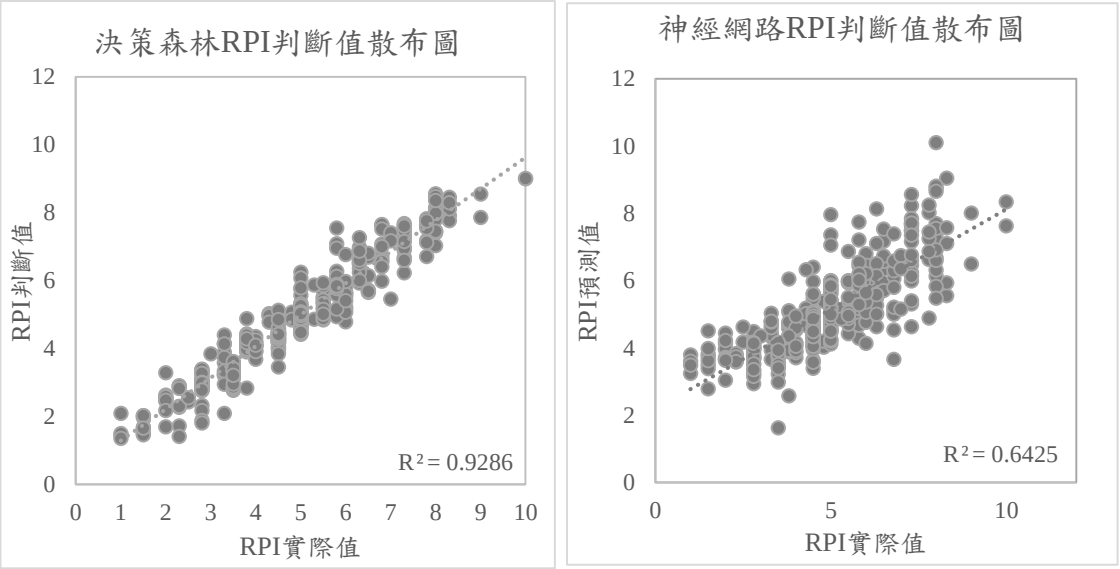


圖 4- 20　RPI 指標實際值及判斷值散布圖-模型建置測試

決策樹(森林)構造解讀

可由列表看出決策森林回歸模型之評估指標值高於神經網路回歸模型，顯示決策森林模型效果較佳，解讀決策樹之分類流程規則，越上層節點所用的水質參數對於判斷 RPI 指標越為重要。

以本次產生的模型為例，抽樣決策森林中 10 棵決策樹，第 1 層節點的水質參數為 COD 者機率最高，10 個節點中占 8 個，第 2 層節點之水質參數以 NH_3-N 之判斷節點占 9 個機率最高、其次為 COD 及 SS 之判斷節點分別占 6 個及 5 個為次

doi:10.6342/NTU201803135

之，詳如表表 4-13。總結來說，對於 RPI 指標之判斷，最重要的判斷依據為 COD，其次為 NH_3-N，而 DO 對於判斷 RPI 指數最不重要。

表 4-13　判斷銅濃度超標之重要決策樹節點內容

層級	節點數目			
	COD	NH_3-N	SS	DO
第一層(共 10 節點)	8	2	0	0
第二層(共 20 節點)	6	9	5	0

實作一評估以 COD 取代 BOD 判斷 RPI 指標可行性

比較 2 個數值推估的回歸模型，決策森林回歸模型的表現優於神經網路回歸模型，遂拿決策森林回歸方法用作 2017 年至 2018 年 4 月年間 RPI 數據判斷，並檢視效果是否合乎想像，以及確認以 COD 代替 BOD 之構想是否可行。

測試資料 2017 年 1 月至 2018 年 4 月之觀測數據，總數據量為 112 筆，總結來說，以決策森林回歸模型判斷水質污染指標 RPI 值之效果較神經網路回歸模型佳，決策森林回歸模型之平均絕對誤差(MAE)為 0.379、根均方誤差(RMSE)為 0.116、平均絕對誤差百分比(MAPE)為 0.141 及決定係數(R^2)為 0.877，如表 4-14，實際值及預測值散布圖如下圖 4-21，其絕對誤差最大值為 1.274，表示預測值最遠和實際值僅相差 1.274，因此證明利用本模型判斷 RPI 指標數值應為可行。

表 4-14　兩模型推估 RPI 水質指數結果比較-實際測試(2017-2018)

評估模型指標	決策森林回歸模型	類神經回歸模型
平均絕對誤差(MAE)	0.379	0.865
根均方誤差(RSME)	0.116	1.101
平均絕對誤差百分比(MAPE)	0.141	0.376
決定係數(R^2)	0.877	0.442

doi:10.6342/NTU201803135

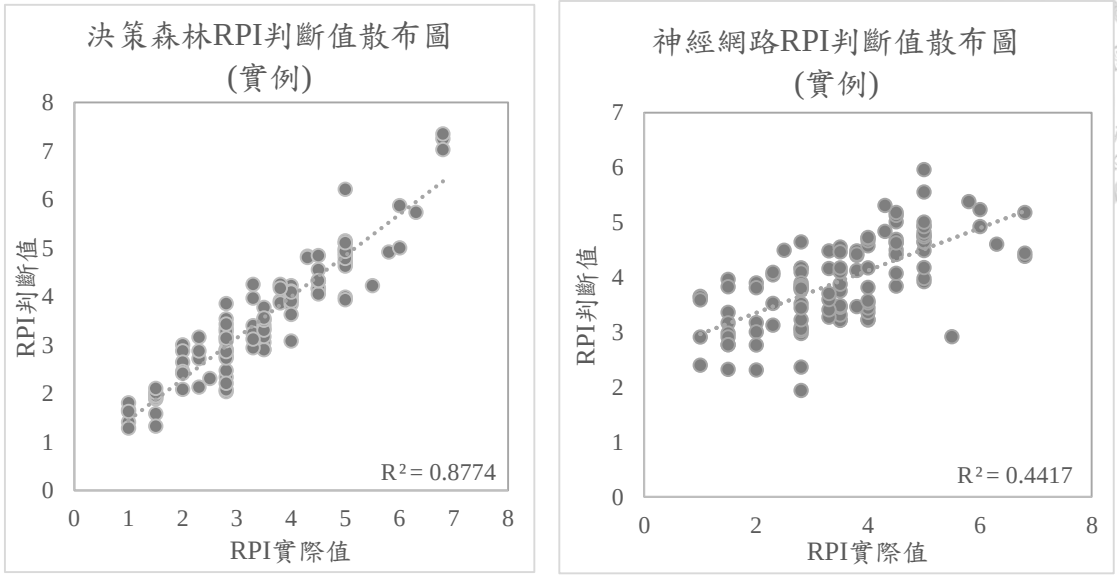


圖 4- 21　RPI 指標實際值及判斷值散布圖-實際測試(2017-2018)

doi:10.6342/NTU201803135

第五章　結論與建議

本研究目的在於以資料探勘的方式，分析樣本數量多變項又複雜的水質監測資料，期望能從水質資料中發掘水質變項或採樣點間較難直接察覺的關聯性。了解變項間關聯性，可進而得出污染排放模式，推測污染來源，配合採樣點間相似度資訊，可以更深入解釋河川整體變化。另一部分為使用機器學習方法建置判斷模型，針對目前水質監測的不足之處，如不頻繁的水中金屬濃度量測頻率，或為得到河川污染程度指標需經耗時過久的實驗過程等狀況，提供新的方法予需要者能在沒有金屬濃度測量的月份得到超標與否的判斷數值，或能更即時地得到河川污染程度指標資訊。

水質監測數據分析結論

一、污染物平均值最高值集中於支流 S3 平鎮一號橋測站和下游 S7 許厝港一號橋測站

以單一污染指標描述老街溪 2002 至 2016 年水質變化，比較水質測項在各測站的平均值，藉以看出污染於河川流域中分布狀況。經統計分析後得知污染物平均值最高值集中於 S3 測站和 S7 測站，水質測項平均值最高點為落於 S3 測站者有氨氮、總磷、硝酸氮、銅、錳，落於 S7 測站者有河川污染程度指標、導電度、生化需氧量、化學需氧量、總有機碳、亞硝酸氮、鉻、銀及鋅。

二、老街溪河川污染程度指標多呈現中度污染，主要原因為高濃度的氨氮及生化需氧量所致

以老街溪歷年河川污染程度指標(RPI)來看，各採樣點的 RPI 指標隨時間推近呈現下降的趨勢，惟大部分測站仍屬於中度污染，可歸因於氨氮及生化需氧量濃度過高。測站 S2 北勢橋、S5 公園橋及 S6 中正橋是 RPI 指標較低、水質較好的點位。

doi:10.6342/NTU201803135

三、**水中的金屬錳、銅及鉛超標次數最多，且超標樣本多來自支流 S3 平鎮一號橋、S4 環鄉橋及 S7 許厝港一號橋點位**

水中金屬測值總超標率依序排名為錳(87 %)、銅(69 %)、鉛(29 %)、鋅(8 %)及六價鉻(1 %)，各金屬濃度隨時間變化趨勢相似，平均值高峰均集中於 2006 年以前，錳及銅雖然於 2006 年後平均測值大幅降低，但多數測值仍大於保護人體基準值，金屬鉛於 2007 年後之監測值均未超過保護人體基準值。

四、**主成分分析及因素分析結果**

經主成分及因素分析，萃取出的 6 個因素可解釋資料及總變異量的 70%，萃取出的因素依序有複合污染物，降雨沖刷，工業廢水排放(如半導體業、印刷電路板業)，工業常見金屬材料等。

複合污染物污染主要位於 S7 許厝港一號橋，半導體產業污染主要分布於 S3 平鎮第一號橋及 S4 環鄉橋，印刷電路板業污染則是在 S2 北勢橋、S3 平鎮一號橋及 S7 許厝港一號橋有較高的分數。

隨時間變化，複合污染物排放、降雨沖刷污染、工業金屬材料污染均有逐年降低趨勢，顯示污染物有效被控制，印刷電路板污染分數漸趨向平均值集中，顯示高濃度污染排放有效被控制，但也可能因為產業進駐導致原本較無污染的點位污染分數提升，而半導體產業污染排放則無明顯趨勢。

五、**依據 2012 至 2016 年監測資訊，可將老街溪水質測站分為 3 個群集，分別為上游群集(S1、S2)、支流(S3)及中下游群集(S4、S5、S5 及 S7)**

各採樣點位水質特性持續隨時間變化，觀察每 5 年 1 筆之群集樹狀圖的分類結果，自 2002 年起，支流 S3 平鎮第一號橋未曾與主要河流被分類為同一群集，顯示其水質具獨特性；另外，與上游測站特性相似的點位逐年減少，相反地，與下游特性相似的點位逐年增加，推測可能因污染程度高的支流匯入主流後影響。最近 5 年(2012-2016)監測資料進行群集分析後，得到上游群組 2 站、下游群組 4 站及支流 1 站的結果。

doi:10.6342/NTU201803135

機器學習建置模型

一、以每月例行量測數值預測金屬銅濃度超標之分類之可行性

評估建置完成的模型，二元決策森林模型及二元神經網路模型之正確率分別達到 0.831 及 0.847。實際測試結果則以二元決策森林模型效果較佳，正確率達 88 %，用以判斷水中金屬銅濃度超標與否應為可行。

由決策森林中運算完成之決策樹結構可得知，對於銅濃度超標與否之判斷，SS、EC 及採樣點位等參數為重要的判斷依據。

二、以 COD 代替 BOD 判斷水質污染指標(RPI)之可行性

評估建置完成的模型，決策森林回歸模型有較低的平均絕對誤差(MAE)、根均方誤差(RSME)及平均絕對誤差百分比(MAPE)，分別為 0.352、0.464 及 0.087，且有較高的決定係數(R^2)為 0.929。實際測試 PRI 數值評估結果，其絕對誤差最大值為 1.274，表示預測值最遠和實際值僅相差 1.274，因此證明利用本模型判斷 RPI 指標數值應為可行。

由決策森林中運算完成之決策樹結構可得知，對於 RPI 指標之判斷，最重要的判斷依據為 COD，其次為 NH_3-N。

建議

一、本研究採用探索式方法分析水質資訊，以因素分析方法追溯影響河川水質之污染源頭，除了可能為畜牧或家庭廢水外，工業廢水排放(如金屬表面處理業、半導體產業污染或電路印刷業等)亦為重要污染來源，可進一步於污染分數較高之採樣點位進行工業放流水的採樣驗證。

二、機器學習模型評估結果顯示，各種模型在不同的水質狀況下，其判斷的效果不同，倘欲精進模型學習效果，可以多種機器學習技術針對同一議題製作多種模型，再利用整體學習(Ensemble Learning)的方法，比對彼此的結果後進行投票或重複運算以得到最佳結果，可精進分類或數值判斷的效果。

doi:10.6342/NTU201803135

doi:10.6342/NTU201803135

第六章　參考文獻

Azure, M. (2018). Evaluate Model. Retrieved from https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model

Bierman, P., Lewis, M., Ostendorf, B., & Tanner, J. (2011). A review of methods for analysing spatial and temporal patterns in coastal water quality. *Ecological Indicators, 11*(1), 103-114.

Chau, K., & Chen, W. (2001). A fifth generation numerical modelling system in coastal zone. *Applied Mathematical Modelling, 25*(10), 887-900.

Chau, K. W. (2006). A review on integration of artificial intelligence into water quality modelling. *Mar Pollut Bull, 52*(7), 726-733.

Chou, J.-S., Ho, C.-C., & Hoang, H.-S. (2018). Determining quality of water in reservoir using machine learning. *Ecological Informatics, 44*, 57-75.

Couto, C., Vicente, H., Machado, J., Abelha, A., & Neves, J. (2012). Water quality modeling using artificial intelligence-based tools. *International Journal of Design & Nature and Ecodynamics, 7*(3), 300-309.

Hecht-Nielsen, R. (1987). *Kolmogorov's mapping neural network existence theorem.* Paper presented at the Proceedings of the IEEE International Conference on Neural Networks III.

Heddam, S., & Kisi, O. (2017). Extreme learning machines: a new approach for modeling dissolved oxygen (DO) concentration with and without water quality variables as predictors. *Environ Sci Pollut Res Int, 24*(20), 16702-16724.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and psychological measurement, 34*(1), 111-117.

Kassambara, A., & Mundt, F. (2017). Package 'factoextra,'. *R topics documented*, 75.

Liu, C.-W., Lin, K.-H., & Kuo, Y.-M. (2003). Application of factor analysis in the

doi:10.6342/NTU201803135

assessment of groundwater quality in a blackfoot disease area in Taiwan. *Science of The Total Environment, 313*(1-3), 77-89.

Noori, R., Sabahi, M. S., Karbassi, A. R., Baghvand, A., & Taati Zadeh, H. (2010). Multivariate statistical analysis of surface water quality based on correlations and variations in the data set. *Desalination, 260*(1-3), 129-136.

Nosrati, K., & Van Den Eeckhaut, M. (2012). Assessment of groundwater quality using multivariate statistical techniques in Hashtgerd Plain, Iran. *Environmental Earth Sciences, 65*(1), 331-344.

Olsen, R. L., Chappell, R. W., & Loftis, J. C. (2012). Water quality sample collection, data treatment and results presentation for principal components analysis--literature review and Illinois River Watershed case study. *Water Res, 46*(9), 3110-3122.

Pagano, M., & Gauvreau, K. (2018). *Principles of biostatistics*: Chapman and Hall/CRC.

Palani, S., Liong, S. Y., & Tkalich, P. (2008). An ANN application for water quality forecasting. *Mar Pollut Bull, 56*(9), 1586-1597.

Papaioannou, A., Mavridou, A., Hadjichristodoulou, C., Papastergiou, P., Pappa, O., Dovriki, E., & Rigas, I. (2010). Application of multivariate statistical methods for groundwater physicochemical and biological quality assessment in the context of public health. *Environmental monitoring and assessment, 170*(1-4), 87-97.

Revelle, W. R. (2017). psych: Procedures for personality and psychological research.

Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling & Software, 22*(4), 464-475.

doi:10.6342/NTU201803135

Singh, K. P., Malik, A., Mohan, D., & Sinha, S. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)--a case study. *Water Res, 38*(18), 3980-3992.

Vega, M., Pardo, R., Barrado, E., & Debán, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water research, 32*(12), 3581-3592.

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W., & Tscheikner-Gratl, F. (2018). Pipe failure modelling for water distribution networks using boosted decision trees. *Structure and Infrastructure Engineering*, 1-10.

王嶽斌. (2015). *整合多變量方法評估水體及底泥品質時空特性之研究*.

台灣省水污染防治所. (1975). *台灣河川水質年報(中華民國六十五年)*.

正昌. (2005). *多變量分析方法: 統計軟體應用*: 五南圖書出版股份有限公司.

行政院環境保護署. (2011). *老街溪污染總量管制模式評估計畫專案工作計畫*.

行政院環境保護署. (2016). *2016 年環境水質年報定稿*.

行政院環境保護署. (2018a). 水質保護網. Retrieved from https://water.epa.gov.tw/River_laochieh.aspx

行政院環境保護署. (2018b). 全國環境水質監測資訊網. Retrieved from https://wq.epa.gov.tw/Code/Theme/Overall.aspx

行政院環境保護署. (2018c). 列管污染源資料查詢系統. Retrieved from https://prtr.epa.gov.tw/

行政院環境保護署. (2018d). 環境資源資料庫. Retrieved from https://erdb.epa.gov.tw/DataRepository/PollutionProtection/AllWaterPollutantEmissions.aspx

吳孟育. (2005). 印刷電路板業含銅廢液銅化合物之回收與轉化. *崑山科技大學環境工程研究所學位論文*, 1-107.

doi:10.6342/NTU201803135

林倩如. (2006). 環境品質調查資料空間變異分析之探討. *臺灣大學生物環境系統工程學研究所學位論文*, 1-149.

林清山. (1991). *多變量分析統計法. 台北市: 東華.*

桃園縣政府環保局. (2011). *老街溪流域水質改善暨整治策略.*

桃園縣政府環境保護局. (2015). 河川水質監測項目說明. Retrieved from https://www.tydep.gov.tw/tydep/static/river/main4.html

張祚楨. (2013). 河川水質管理之水質指標評估. *淡江大學水資源及環境工程學系碩士班學位論文*, 1-65.

傅粹馨. (2002). 主成份分析和共同因素分析相關議題之探究. *教育與社會研究.*

楊于嫻. (2014). 都市河川復育之研究-以老街溪為例. *成功大學水利及海洋工程學系學位論文*, 1-83.

經濟部工業局. (1993). 行業製程減廢及污染防治技術－半導體業介紹.

劉應興. (1997). *應用線性迴歸模型. 台北市: 華泰文化事業股份有限公司.*

doi:10.6342/NTU201803135