



行政院環境保護署

111 年度土壤及地下水污染整治基金補助研究與模場試驗專案

利用人工智慧結合政府跨部會多元資料 建立地下水硝酸鹽氮污染潛勢預測模型

期末報告(定稿)

主辦單位： 行政院環境保護署

專案執行單位：國立中央大學／應用地質研究所

專案主持人：陳瑞昇 教授

專案執行期間：111 年 09 月 01 日起至
112 年 08 月 31 日止

中華民國 112 年 08 月 印製





專案基本資料表

專案性質	☐ 實驗性質 ☒ 非實驗性質	專案類別(單選)	☐ 先導型 ☒ 研究型 ☐ 模場型		
研究主題	☒ 調查 ☐ 整治 ☐ 其他				
申請機構系所	國立中央大學/應用地質研究所				
機構地址	320317 桃園市中壢區中大路 300 號				
專案主持人	陳瑞昇	職等／職稱	教授		
協同主持人	洪炯宗	職等／職稱	教授		
協同主持人	梁菁萍	職等／職稱	教授		
專案名稱	中文	利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型			
	英文	Combining artificial intelligence and inter-ministerial and inter-agency multivariate data to establish a predictive model for groundwater nitrate pollution vulnerability			
	關鍵字	硝酸鹽氮，人工智慧，人工神經網路，多元資料，預測變量，反應變數			
執行期程	自民國 111 年 9 月 1 日起至民國 112 年 8 月 31 日止				
專案主持人	姓名：陳瑞昇 E-mail：jschen@geo.ncu.edu.tw		專線：03-2807427 手機：0932360361		
專兼任人員	姓名：陳琤雯 E-mail：akeyuke.0109@gmail.com		專線：03-4227151#65865 手機：		
經費分析總表 (僅模場試驗專案需填寫兩年度金額)	專案預估經費		第一年金額	第二年金額	編列說明
	1.	人事費用	504,900		(1~6 項相加之 50%為限)
	2.	設備使用含維護費			(與計畫實驗相關)
	3.	耗材與主要費用	17,940		(與計畫主體相關)
	4.	其它研究相關費用	4,160		(含差旅與租賃費用)
	5.	雜支費用	28,000		(1~6 項相加之 5%為限)
	6.	行政管理費	55,000		(1~6 項相加之 10%為限)
	7.	自籌款			(自行籌備款項)
	申請補助金額(1~6 項)		610,000		總金額：610,000
	計畫總金額(1~7 項)		610,000		總金額：610,000

專案主持人(簽名及蓋章)：_____

日期：_____



中文摘要

地下水為我國淡水資源供給的重要來源，高硝酸鹽氮濃度地下水會威脅人體健康且是地下水質保護關注的議題，預測高濃度硝酸鹽氮地下水發生位置的工作，對於水資源達到保護與發展減緩衝擊的策略。以數據驅動化的機器學習為基礎的人工智慧方法在近年已成功應用於準確地預測地下水硝酸鹽氮污染情勢。本計畫目標為發展結合機器學習與跨部會政府多元資料建立地下水硝酸鹽氮污染潛勢預測模式，機器學習所需的預測變量資料庫包括來自政府跨部會土地利用、水文、土壤型式、水文地質與硝酸鹽氮輸入等資料。考慮的 4 個機器學習方法包括線性迴歸、隨機森林、極限梯度提昇與人工神經網路。4 個機器學習模型的功能評估藉由 15 折交叉驗證進行，利用 Shapley 值法來篩選主要影響的預測變量，結果顯示果樹土地利用為最重要的影響變數，接著為未飽和層粉土厚度與未飽和層礫石厚度，此結果建議應該教育果農應減少化肥或雞糞的使用，推廣有機果樹種植以有效減少地下水硝酸鹽氮污染。



英文摘要

Groundwater is a vital source of the fresh water supplies in Taiwan. Presence of high nitrate nitrogen concentration can pose threat to human health and is already a major concern for groundwater quality protection. Efforts for predicting the locations of high nitrate nitrogen groundwater are required for the protection of water resources and development of mitigation strategies. Most recently, artificial intelligence (AI) techniques like data-driven machine learning (ML) approaches have proven their worth for successively and accurately characterizing nitrate nitrogen groundwater pollution occurrence. In this project, we aim to develop groundwater nitrate nitrogen pollution vulnerability prediction model by combining machine learning and inter-ministerial and inter-agency multivariate data. The data set for predictor variables used for machine learning includes land use, hydrological conditions, soil type, hydrogeology and nitrate concentration collected from inter-ministerial and inter-agency. Four machine learning methods including linear regression (LR), random forest (RF), extreme gradient boosting method (XGBboost) and artificial neural network (ANN) are used to develop the AI prediction model. Fifteen-fold cross-validation is performed to evaluate the prediction performance. Ultimately, dominant predictors that significantly influence the nitrate nitrogen concentration are examined using Shapley value evaluation method. Land use type of fruit farming is the most important factors, follows by silt thickness of unsaturated soil and gravel thickness of unsaturated soil. This finding implies that fruit farmers should educate the farmer to reduce the use of chemical fertilizers and chicken manure and promote the organic fruit farming for effectively minimizing nitrate nitrogen groundwater contamination.



目次

一、前言	1
二、研究方法與過程	12
2.1 政府跨部會預測變量與反應變數資料	13
2.1.1 地下水水質	13
2.1.2 地質材料	18
2.1.3 土地利用	20
2.1.4 其他部會資料	23
2.1.5 由計算方法取得的預測變量	23
2.2 ANN 模式	24
三、結果與討論	26
3.1 模型建置的資料前處理	26
3.2 人工智慧方法與評估指標	33
3.3 模型建置結果分析	38
四、結論與建議	56
五、參考文獻	57
附錄一 111 年度專案成果績效自評表	59
附錄二 審查意見回復對照表	64



圖次

圖 1 影響地下水硝酸鹽氮濃度的各種因子.....	2
圖 2 序列結合因子分析與群集分析綜合評估土地利用與水文地質特性對地下水水質的影響流程圖.....	4
圖 3 因子 1(海水鹽化)因子得分的地理視覺化.....	5
圖 4 因子 3(硝酸鹽氮污染)因子得分的地理視覺化.....	6
圖 5 經濟部水利署的 123 口觀測井經群集分析可區分為 3 個群集的地理視覺化.....	7
圖 6 本專案示意圖.....	11
圖 7 研究流程圖.....	12
圖 8 經濟部水利署 82 口地下水水質監測的觀測井位置圖.....	14
圖 9 環保署 360 口地下水水質監測井位置圖.....	15
圖 10 經濟部水利署與環保署共計 442 口地下水水質監測井或觀測井位置圖.....	16
圖 11 第二次國土利用調查分類系統監測井週圍土地利用圖(第一級分類).....	21
圖 12 第二次國土利用調查分類系統(第二級分類, 農業進一步區分為農作、養殖、畜牧與農業四類).....	21
圖 13 第二次國土利用調查分類系統井週圍土地利用圖(第三級分類, 農作再區分為稻作、旱作、果樹與廢耕地四類).....	22
圖 14 以井周圍圓形區域距離內各類土地利用組成面積的百分比做為數值資料, 便進行數值相關計算(此處考慮半徑 1,000m 範圍內之土地利用).....	22
圖 15 ANN 架構示意圖.....	25
圖 16 硝酸鹽氮濃度取對數前後的資料分佈, 左圖為取對數前的硝酸鹽氮濃度分佈; 右圖為取對數後的硝酸鹽氮濃度分佈.....	26
圖 17 硝酸鹽氮濃度取對數前後的資料分佈, 左圖為底數是自然對數的硝酸鹽氮濃度分佈; 右圖為底數是 10 的硝酸鹽氮濃度分佈.....	27



- 圖 18 各個土地利用的資料分佈圖，由左而右、由上到下，依序是稻作利用土地、旱作利用土地、果樹利用土地、廢耕地利用土地、畜牧利用土地、農業其他利用土地、農業利用土地、森林利用土地、交通利用土地、水利利用土地、建築利用土地、公共利用土地、遊憩利用土地、礦鹽利用土地、其他利用土地..... 29
- 圖 19 各個水文地質組成的資料分佈圖，由左而右、由上到下，分別是水文地質礫石組成、水文地質砂土組成、水文地質粉土組成、水文地質黏土組成..... 30
- 圖 20 其他部會資料，由左而右、由上到下，分別是降雨量、總硬度 (mg/L)、總溶解固體物 (mg/L)、氯鹽 (mg/L) 31
- 圖 21 不同 ANN 建模結果，上半部三張圖由左到右分別是使用線性整流函數、雙曲正切激活函數、恆等函數，且神經元皆是(200, 100)之架構；下半部三張圖由左到右分別是神經元(200, 100)、(150, 50)、以及(75, 50)，且激活函數為線性整流函數之結果..... 38
- 圖 22 MLR 模型預測結果..... 38
- 圖 23 RF 採用不同參數的迴歸模型預測結果，由左到右，由上到下分別是樹木數量 100、80、75、及 60 之結果..... 39
- 圖 24 由左到右分別是 XGBoost 採用樹木數量 100、80、及 60 的實驗結果..... 39
- 圖 25 未取對數之硝酸鹽氮濃度不同模型於訓練（左）與測試（右）資料集的準確率..... 41
- 圖 26 ANN 建模結果，第一列的三張圖由左到右分別是使用線性整流函數、雙曲正切激活函數、恆等函數，且神經元皆是(350, 350)之架構；第二列兩張圖由左到右分別是神經元(150, 150)、以及(150, 50)；第三列兩張圖由左到右分別是神經元(500, 500)、以及(1500, 1500)，且激活函數為恆等函數之結果..... 43
- 圖 27 LR 模型預測結果..... 44
- 圖 28 RF 採用不同參數的迴歸模型預測結果，由左到右，分別是樹木數量



100、75、及 60 之結果	44
圖 29 由左到右分別是 XGBoost 採用樹木數量 100、50、及 10 的實驗結果	44
圖 30 取對數之硝酸鹽氮濃度不同模型於訓練（左）與測試（右）資料集的準確率	46
圖 31 上半部兩張圖由左到右分別是 LR 模型預測結果，RF 採樹木數量 20 之架構；下半部兩張圖由左到右分別是 XGBoost 採樹木數量 20 深度為 2 和神經元 (256, 128,64,32)，且激活函數為線性整流函數的 ANN 之結果	47
圖 32 上半部兩張圖由左到右分別是 LR 模型預測結果，RF 採樹木數量 20 之架構；下半部兩張圖由左到右分別是 XGBoost 採樹木數量 20 深度為 2 和神經元 (256, 128,64,32)，且激活函數為線性整流函數的 ANN 之結果	48
圖 33 訓練資料中未採用過採樣（左）與採用過採樣（右）資料集的 RMSE	48
圖 34 測試資料中未採用過採樣（左）與採用過採樣（右）資料集的 RMSE	49
圖 35 不同模型未採用過採樣時於訓練（左）與測試（右）資料集的準確率	51
圖 36 不同模型採用過採樣時於訓練（左）與測試（右）資料集的準確率	53
圖 37 採用底數為 1.05 之對數（左）與採用自然對數（右）在各模型上測試集的差異	53
圖 38 以自然對數（左）與採用底數為 10 之對數轉換（右）在各模型上測試集的差異	53
圖 39 ANN 模型預測類別結果（左）與實際硝酸鹽氮濃度（右）	54
圖 40 單次模型之 Shapely Value 結果	55



表次

表 1 行政院環保署飲用水水質標準與農委會之灌溉水質標準	17
表 2 沉積物粒徑分類	19
表 3 各預測變量之平均、最大、及最小值.....	28
表 4 未取對數之硝酸鹽氮濃度預測迴歸模型評估結果	40
表 5 不同模型於訓練資料集的混淆矩陣.....	40
表 6 不同模型於測試資料集的混淆矩陣.....	41
表 7 取對數之硝酸鹽氮濃度預測迴歸模型評估結果.....	45
表 8 不同模型於訓練資料集的混淆矩陣.....	45
表 9 不同模型於測試資料集的混淆矩陣.....	46
表 10 取對數之硝酸鹽氮濃度預測迴歸模型評估結果.....	49
表 11 不同模型於訓練資料集的混淆矩陣.....	50
表 12 不同模型於測試資料集的混淆矩陣.....	50
表 13 取對數之硝酸鹽氮濃度預測迴歸模型評估結果.....	51
表 14 不同模型於訓練資料集的混淆矩陣.....	52
表 15 不同模型於測試資料集的混淆矩陣.....	52
表 16 使用 Shapley Value 分析出來重要預測變量前三名。	55



一、前言

台灣許多地區地下水資源蘊藏量非常豐富，在地面水不足夠供給水資源使用需求時，就會進行地下水開發做為民生、工業與農業用水(包含灌溉、畜牧與養殖)等目的使用，以穩定水資源供應，促進國家經濟發展。根據經濟部水利署水利統計年報顯示，台灣全年總用水量約為 170 億立方公尺，其中水庫運用水量與引水河川水量約供給全年總用水量的 68%，因此地下水資源使用量約已達台灣全年總用水量的 1/3(經濟部水利署，2021)。另外，受到氣候變遷影響而改變水文循環，因此也影響乾旱事件與極端降水發生頻率，因此也造成地表水做為水資源穩定供給的不確定性，也造成地下水在扮演水資源供給有著更重要的角色。

地下水資源開發要滿足家戶、農業灌溉、漁業養殖與工商業使用等各類目的之用水需求，除要考慮水量是否足夠外，水質更是一個地下水能否安全使用最關鍵要素。自然過程中的水岩交互作用(water-rock interaction)過程，常常會釋放危害人體的物質進入地下水體，而人類活動的居住、工業、農漁業也會不經意產生危害人體物質，這些危害物質會隨著降雨入滲而進入地表下的土壤與地下水含水層系統，而使得地下水水質劣化，全球許多國家面臨地下水水質劣化問題，地下水水質劣化會對生態環境與人體健康造成重大的危害。

為了掌握全台灣地下水水質概況，經濟部水利署與行政院環保署目前經常性進行台灣之地下水水質監測，環保署與經濟部水利署長期的地下水水質監測結果顯示部份監測井中的特定水質參數超過環保署飲用水水質標準。

如圖 1 所示，人為活動包括都市化、工業製造與農業活動可能產生各種不同型式的污染物，這些污染物會隨著降雨或灌溉的入滲水進入地表下，首先會經過未飽和層土壤(unsaturated soil)，當這些入滲水經過未飽和土壤，未飽和層的材料組成、厚度與物理與化學特性都會影響入滲水的地下水水質。繼續往下滲漏(percolation)的入滲水最終到達淺層非受壓含水層(unconfined aquifer)，伴隨入滲水的污染物將會影響含水層的地下水水質。含水層地下水流特性、礦物顆粒性質與水化學特性都可能改變地下水中的污染物濃度，也就是說地下水質特性是地下水系統中的各種物理、化學與生物過程所造成的地下水水質的變化，同時人為活動也會造成地下水水質的變化，因此在探討地下水水質特性需考慮地表的人為活動與地表下水文地質特性。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型



圖 1 影響地下水硝酸鹽氮濃度的各種因子

考慮反應人類活動的最重要特徵就是土地利用方式(land use practice)，土地使用方式會產生不同污染來源，因而直接或間接影響地下水水質，專案執行團隊(陳瑞昇專案主持人與梁菁萍協同主持人)以經濟部水利署的 123 口淺層觀測井為基礎，發展序列結合因子分析(factor analysis)與群集分析(cluster analysis)進行全地下水水質特性綜合評估(Liang et al., 2022)，探討土地利用與水文地質特性對地下水水質的影響(如圖 2)，相關研究成果發表於環境科學領域頂尖國際期刊 Science of the Total Environment (IF=10.753, Rank=26/279 in ENVIRONMENTAL SCIENCES)。該文章先利用因子分析(Ca^{2+} 、 Mg^{2+} 、 Na^{+} 、 K^{+} 、 As 、 Cl^{-} 、 SO_4^{2-} 、 Fe^{2+} 、 Mn^{2+} 、 $\text{NO}_3^{-}\text{-N}$ 、 TOC 、 HCO_3^{-} 等 12 項地下水水質參數；水稻田、旱作、果樹、水產養殖、畜牧、森林、建築物公共用地、河道 8 項內政部國土測繪中心土地利用型態資料；礫石、粗(中)砂、細砂、黏土(粉砂)4 項中央地質調查所非飽和層地質材料分類)揭示影響地下水水質的土地利用方式(主要人為活動所造成的來源)和非飽和層(水文地質過程)，辨認出 4 個主要的因子(地下水水質特性)。

因子 1 命名為海水鹽化(seawater salinization)，其特徵為與水產養殖土地利用呈現中度因子負荷，因子 1 在沿海地區的高因子得分(圖 3)是由水產養殖的過度抽水造成的，需進行重點土地利用管理，以防止海水侵入沿海含水層。因子 2 命名為砷富集(arsenic enrichment)，因子 2 與土地利用和水文地質特性相關性不高，顯示出其可能為自然過程中水岩作用的產物。因子 3 命名為硝酸鹽氮污染



一、前言

(nitrate pollution)，與果樹土地利用和非飽和層的礫石厚度呈現高度相關，在濁水溪沖積扇頂地區和屏東平原東北部山區都發現因子 3 的高得分(圖 4)，其管理方法可教育果農減少化肥施用及推廣有機果樹種植。因子 4 命名為鐵錳的還原溶解(reduction dissolution of Fe^{2+} 和 Mn^{2+})，土地利用和水文地質特性對因子 4 的影響可以忽略不計。

對 4 個主要因子的因子得分進行群集分析，可將水利署 123 口監測井分為 3 個群集(圖 5)，群集 1 分群為低污染區(low polluted zone)，因此可以不須特別關注。群集 2 分群為硝酸鹽污染區(nitrate polluted zone)，群集 2 須減少區域的農用化學品使用。群集 3 分群為海水鹽化、砷富集與鐵錳的還原溶解混合污染區(hybrid zone of seawater salinization, arsenic enrichment and reduction dissolution of Fe^{2+} 和 Mn^{2+})，在集群 3 區域須優先實施密集監測。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

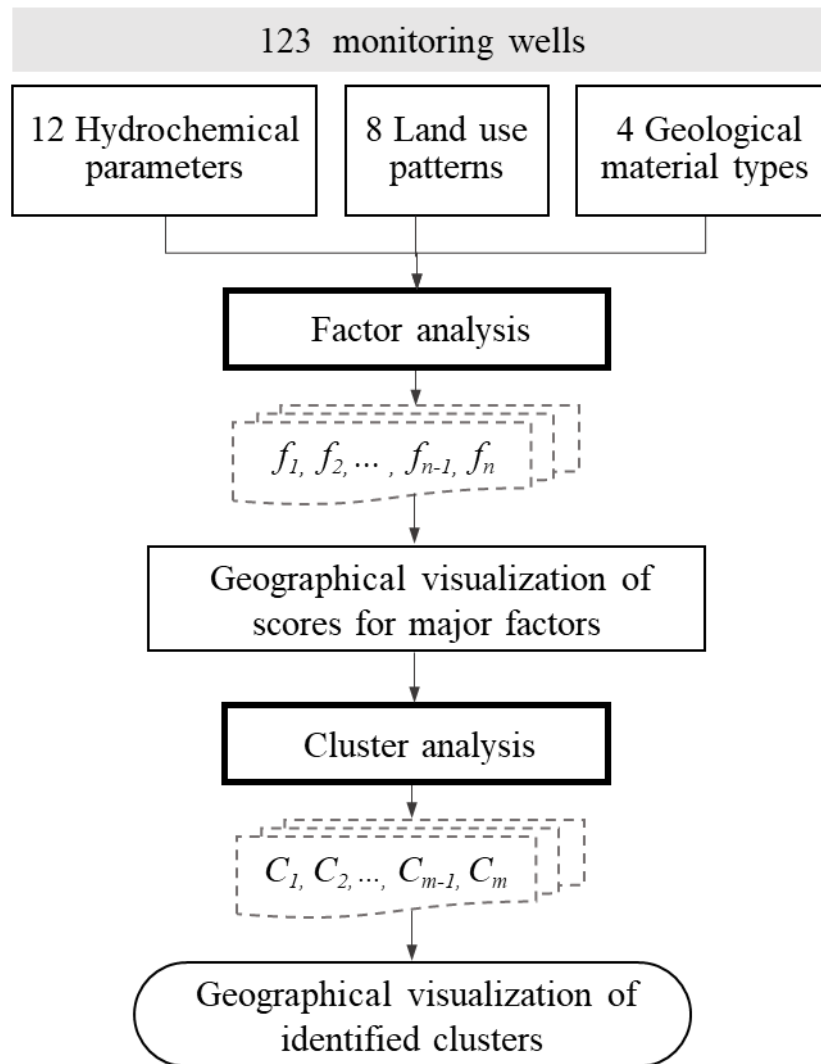


圖 2 序列結合因子分析與群集分析綜合評估土地利用與水文地質特性對地下水水質的影響流程圖

(引用自 Liang et al. (2022))

<https://www.sciencedirect.com/science/article/abs/pii/S0048969722052342>)



一、前言

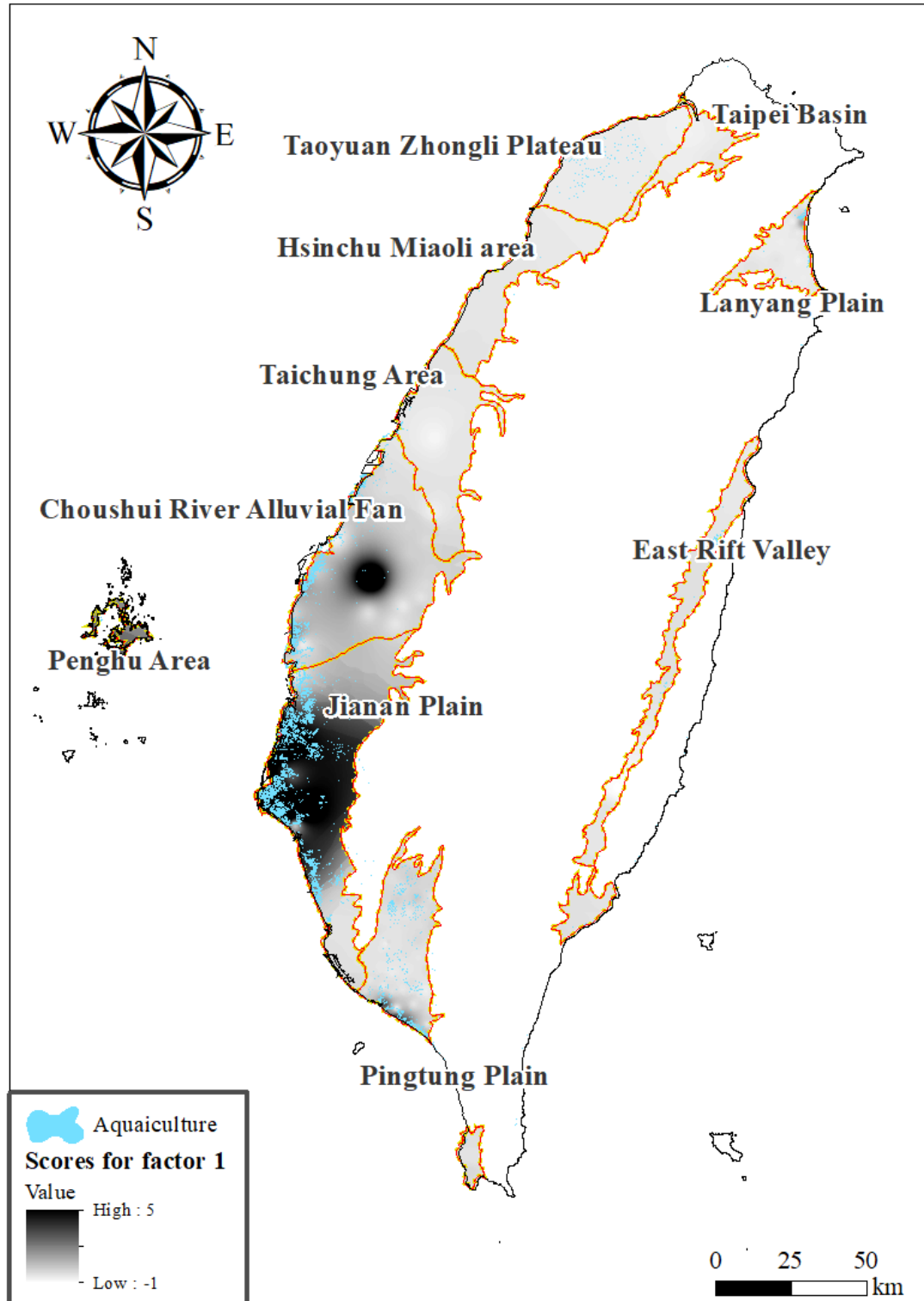


圖 3 因子 1(海水鹽化)因子得分的地理視覺化



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

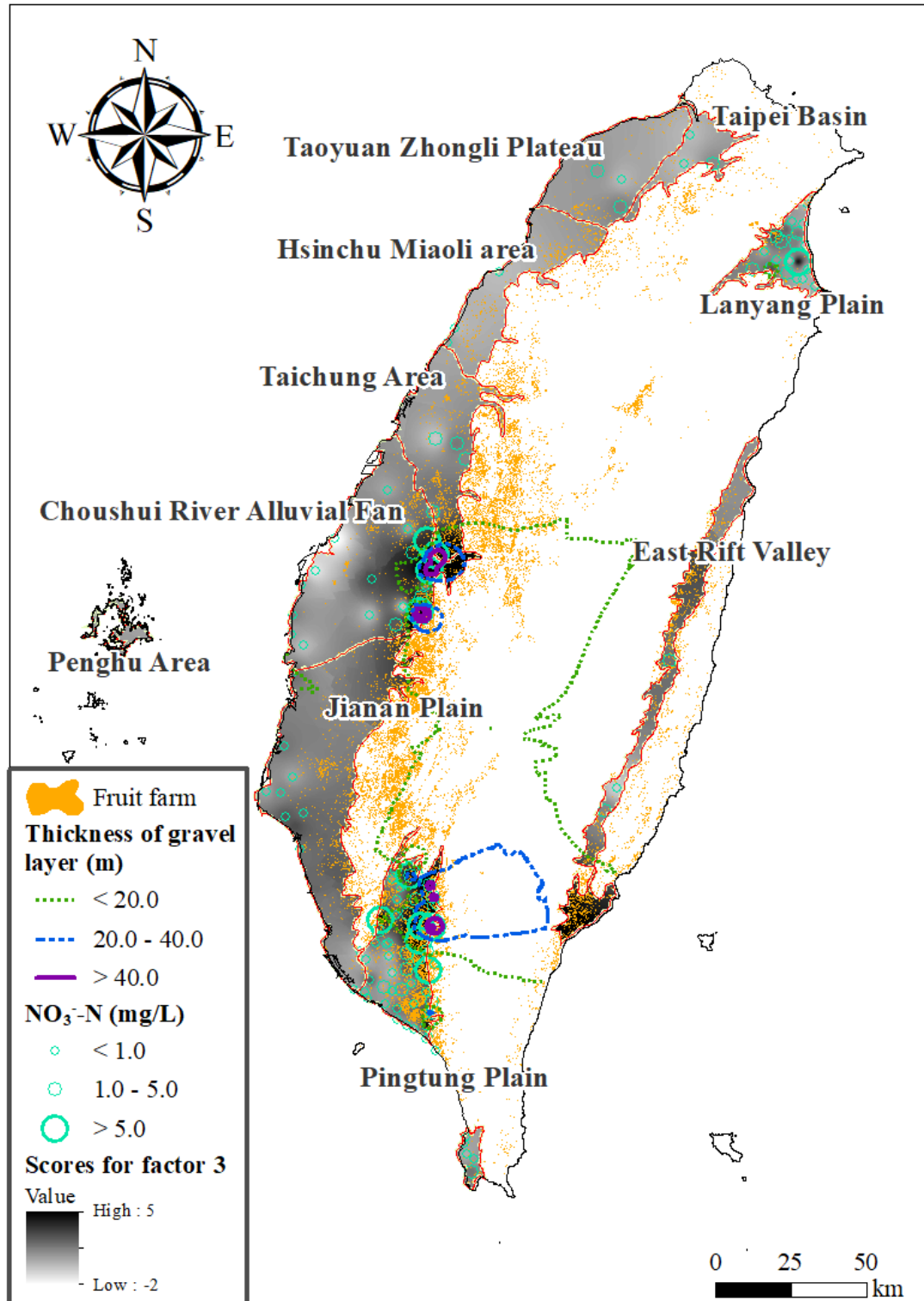


圖 4 因子 3(硝酸鹽氮污染)因子得分的地理視覺化



一、前言

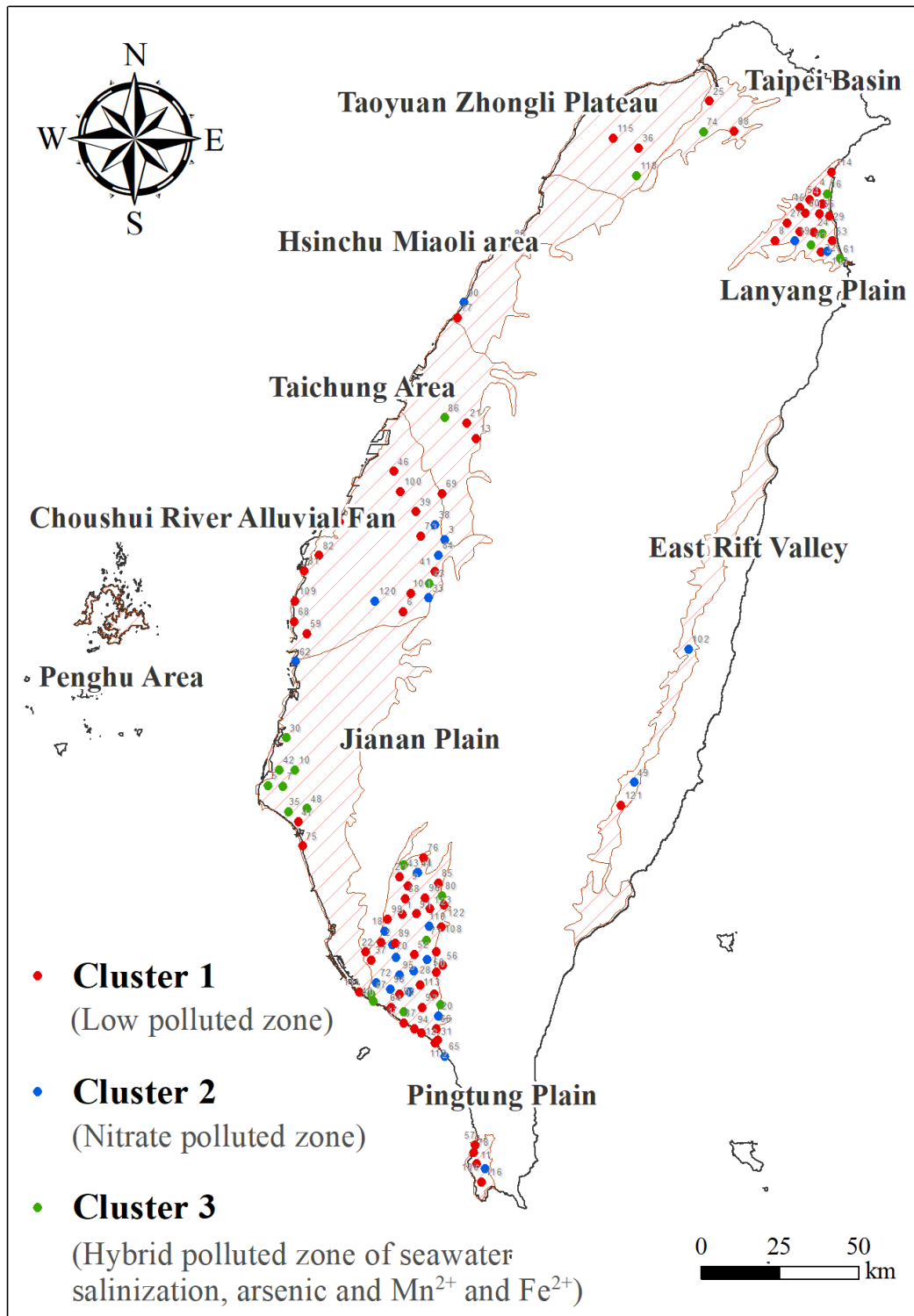


圖 5 經濟部水利署的 123 口觀測井經群集分析可區分為 3 個群集的地理視覺化



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

硝酸鹽氮為含氮無機化合物，是植物生長必需營養來源，易溶於水，人類多從飲用水或食物中攝入硝酸鹽氮，依攝入硝酸鹽氮量多寡可分為急性中毒及慢性危害。硝酸鹽氮本身雖對人體無害，但當經由攝食吃入硝酸鹽氮後，25%的硝酸鹽氮經口腔中的唾液或胃中的胃酸可將其轉換成亞硝酸鹽氮，而人體吸收過量時，亞硝酸會和血紅素結合，使血紅蛋白中的二價鐵離子(Fe^{2+})轉變成三價鐵(Fe^{3+})，造成紅血球中的血紅蛋白異常喪失輸送氧氣的功能，導致細胞組織缺氧產生高鐵血紅蛋白症，就是俗稱的藍嬰症(blue baby syndrome)，主要會發生在腸胃功能還是不完整的嬰幼兒、孕婦及患腸胃炎的人士。國際癌症研究機構(International Agency for Research on Cancer, IARC)確定攝入硝酸鹽氮和亞硝酸鹽氮對人類可能致癌，在2010年將其列為2A的致癌等級(2A類致癌物為對人體致癌的可能性較高的物質或混合物，在動物實驗中發現充分的致癌性證據)。

根據環保署431口地下水水質監測井資料顯示地下水硝酸鹽氮污染主要位在屏東平原及嘉南平原一帶，2019年全臺灣地下水硝酸鹽氮濃度超過飲用水標準(10mg/L)的井總共有11口，其中屏東平原就佔了6口，濁水溪沖積扇有2口，嘉南平原、新苗地區各有1口，而超過中濃度範圍(1mg/L)的井有204口，同樣也是屏東平原超標最多，有63口。地下水中硝酸鹽氮的生成與人類活動高度相關，地表人類活動的土地利用，如稻作、旱作及果樹使用的肥料含有氮、磷等物質，畜牧業動物的排泄物與人類生活化糞池產生的污水也都含有大量的含氮物質，工業產生的廢水也為硝酸鹽氮之可能來源。降雨或灌溉會造成這些含氮物質跟著雨水入滲進入土壤，進而進入土壤下方的地下水系統，造成硝酸鹽氮之地下水污染，台灣許多農業地區出現嚴重的硝酸鹽氮污染，此類污染主要源自於附近地表農業活動的旱作種植與過量施肥。

為保護與永續利用珍貴的地下水資源，世界各國家均採取各種積極作為保護或改善地下水水質。地下水水質監測可以提供地下水水質資訊，然而地下水水質監測工作通常需要設置地下水監測井，並定期進行地下水採樣與水質參數分析，這些工作通常須花費大量金錢、人力與時間，因此規劃地下水水質監測站網時，如何在有限的經費、人力與時間下決定地下水水質監測井設置地點的優先順序，是地下水水質保護與永續利用的重要工作。在有限的現地監測(採樣)資料，推估未監測(採樣)點的地下水水質空間變異分布，以了解易發生地下水污染的高潛勢地區，進一步優先設置監測井加強地下水水質監測，根據地下水水質監測結果採取有效的地下水質保護與管理措施，才能達到有效撙節費用，妥善運用監測資源，提高地下水資源永續經營工作的效能。



一、前言

地下水污染潛勢預測為環境科學領域長期重要研究工作，相關的研究常需要在現有地下水水質監測資料下對未進行監測地區的地下水水質濃度進行預測，傳統的空間內插法如反距離加權法(inverse distance weighting, IDW)與 kriging 法為常使用的方法，而多變量統計也是相當成熟的技術，可以探討多個水質與人類活動或土地利用的相互關係，得出污染潛勢的地理視覺化(geographical visualization)。美國環境保護署(United States Environmental Protection Agency)與國家水井會(National Water Well Association, 1987)聯合發展 DRASTIC 模式，用來評估地下水污染潛勢。DRASTIC 模式考慮包含 7 個地質與水文地質參數，其中 D 為地下水位深度(Depth to water)，R 為淨補注量(Net Recharge)、A 為含水層介質(Aquifer media)、S 為土壤介質(Soil media)、T 為地形(Topography)、I 為通氣層介質(Impact of vadose zone)及 C 為水力傳導係數(hydraulic Conductivity)，雖然 DRASTIC 模式對地質與水文地質有完整的考慮，但是在權重與得分的給定較為主觀且沒有考慮到硝酸鹽氮的來源，因此在與現地監測的硝酸鹽氮濃度進行比較，其預測結果常常會與現地監測結果不一致。

人工智慧(Artificial Intelligence, AI)泛指利用電腦模擬人的思想與行為以解決問題，近年來更因電腦運算能力的大幅提升，人類得以藉由電腦的大量資料(數據)中找出規律性並建立模型(model)，進而學習避免過去發生的錯誤，使得 AI 已廣泛應用於各個領域。數據驅動(data-driven)的機器學習(Machine Learning, ML)為 AI 重要的一個分支，亦為現今熱門的資訊科學研究，機器學習的蓬勃發展，不僅可有效率地分析大量資料，也可用來建立更精確的模型，協助做出適當決策。機器學習包含人工神經網路(Artificial Neural Network, ANN)、支持向量機(Support Vector Machine, SVM)、隨機森林(Random Forest, RF)等。過去二十多年來，隨著 AI 的精進，以機器學習為基礎的 AI 技術也開始廣泛應用在各個領域，同時展現在地下水水質的預測能力比傳統 kriging 有更好的預測能力。Chowdhury et al. (2010)比較 ANN 與 kriging 預測孟加拉砷濃度的空間分布的準確度，其結果顯示 ANN 預測的正確率為 67.23%，而 kriging 則只為 50.65%，很明顯 ANN 較於 kriging 有多達 17%的預測準確度。計畫主持人團隊(Liang et al., 2021)利用 ANN 方法於蘭陽平原的砷濃度空間變異分析，並與 kriging 預測結果比較，ANN 與 OK 的決定係數(coefficients of determination, R^2)分別為 0.55 與 0.49，顯示 ANN 比 kriging 有較佳的預測。Man et al. (2021)利用深度學習(deep learning)預測污染場址的土壤重金屬空間分布，經 kriging 的結果比較，ANN 的準確度會比 kriging 多 32%。上述研究說明 ANN 的 AI 方法在進行水質資料的空間變異分析有較好的預測準確度。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

AI 方法也廣泛應用於地下水硝酸鹽氮污染潛勢預測，以環境科學領域頂尖期刊 Science of the Total Environment 近年來發表的幾個歐美國家應用 AI 建立硝酸鹽氮污染潛勢預測模型的研究工作說明。Rodriguez-Galiano et al. (2014)利用 AI 技術的 RF 建立西班牙南部農業環境的硝酸鹽污染潛勢預測模型，研究中考慮了 24 個解釋變量(explanatory variable)，24 個預測變量分類為「驅動力(driving forces)」、「內部特性(intrinsic properties)」、「含水層特性(aquifer properties)」、「地下水流特性(groundwater flow properties)」、「土壤(soil)」、「遙感變數(remotely sensed variables)」與「物理-化學變數(physical-chemical variables)」。「驅動力」主要包括與河川的距離、與灌溉渠道距離、與都市距離、與污水排放區距離、與污水處理場距離以及與畜牧場距離，「內部特性」主要為坡度，「含水層特性」則包括含水層底部高程、地下水位面高程、飽和含水層厚度，「地下水流特性」則包括流通係數、水利傳導係數、水力梯度 90、水力梯度 0、絕對水力梯度，「土壤」則包括土壤組成與通氣層厚度，「遙感變數」則為地表覆蓋與標準化植被指數，「物理-化學變數」包括 pH、電導度與溫度。而硝酸鹽濃度做為反應變數(response variable)，研究結果顯示電導度為最重要的預測變量，然後依序為畜牧養殖、絕對水力梯度、流通係數、地下水位高程等。另外 RF 的結果與邏輯式迴歸(logistic regression, LR)結果比較發現 RF 有較好的預測準確率。Ransom et al. (2017)利用增長回歸樹(boosted regression tree, BRT)結合 145 個預測變量(=解釋變數)資料庫建立美國加州 Central Valley 含水層的硝酸鹽濃度預測模型，研究結果顯示最重要的 5 個預測變量分別為溶氧濃度低於 0.5ppm 與錳超過 50ppb 的機率、1975 年間現地未飽和層的氮輸入(nitrogen input)、1971-2000 年間降雨與蒸發散的差值、1992 年全部景觀硝酸鹽的輸入，最後，25 個預測變量被選用做為硝酸鹽(經對數轉換)濃度的預測模型。Ransom et al. (2022)利用極限梯度提升法(eXtreme Gradient Boosting, XGBoost) 結合 187 個預測變量(=解釋變數)資料庫建立美國本土的硝酸鹽濃度預測模型，研究結果顯示井深度、氣候、土壤、水文與硝酸鹽輸入為最重要的預測變量。

由於飲用含高濃度硝酸鹽氮地下水對人體會有不良健康效應，因此須了解地下水高硝酸鹽氮污染潛勢區，以便採取相關地下水積極保護措施，維護居民健康。上述文獻回顧清楚展現 AI 結合大量的解釋變數(預測變量)資料可有效建立地下水的硝酸鹽氮污染潛勢預測模型，了解硝酸鹽氮高污染潛勢區域，並在高污染潛勢區域加強預警教育，或進行土地管理或加強設置監測井密度定期進行地下水水質採樣與分析，可提高監測經費效益，並達成保護地下水水質與維護民眾用水安全。因此，本研究目的為利用 AI 結合政府跨部會多元資料建立全國硝酸鹽氮污染潛勢預測模型，根據所建立的 AI 地下水硝酸鹽氮污染潛勢預測模型劃分地下



一、前言

水硝酸鹽氮高污染潛勢地圖，提供政府各單位作為採取預警與保護措施的參考依據，本專案示意圖如圖 6。

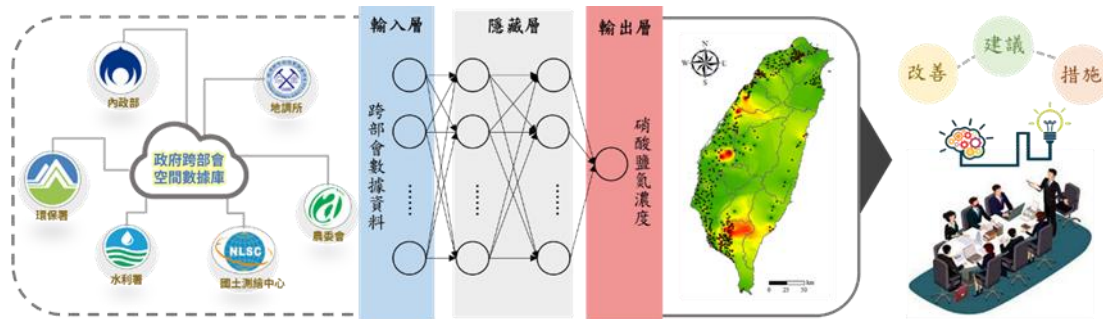


圖 6 本專案示意圖

因此本專案將達成下列工作目標：

1. 藉由 AI 結合政府跨部會多元資料(預測變量)建立地下水硝酸鹽氮污染潛勢預測模型。
2. 篩選顯著影響地下水硝酸鹽氮濃度的主要預測變量並優化地下水硝酸鹽氮污染潛勢預測模型。
3. 建立地下水硝酸鹽氮污染潛勢地理視覺化地圖。
4. 根據硝酸鹽氮污染潛勢地理視覺化地圖在高潛勢區以篩選出的顯著影響地下水硝酸鹽氮濃度的主要預測變量為參考，擬定地下水硝酸鹽氮污染改善措施。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

二、研究方法與過程

本研究目的為藉由 AI 結合政府跨部會多元資料(預測變量)建立全國地下水硝酸鹽氮污染潛勢預測模型，根據所建立的 AI 地下水硝酸鹽氮污染潛勢預測模型劃分高污染潛勢地圖，提供政府單位作為採取預警與保護措施的參考依據。如前所述，地面人為活動會產生各種危害物質進入含水層而改變地下水水質，這些危害物質為地下水污染的來源，因此建立地下水硝酸鹽氮 AI 預測模型首先要考慮可能產生這些污染源的相關變數，硝酸鹽氮污染的來源可能來自農業施肥與畜牧、生活廢污水、工業污水、大氣氮氧化物沉降與地表河川污染。除各種人為活動的可能影響硝酸鹽氮濃度外，硝酸鹽氮會隨入滲水進入地表下經過未飽和層土壤而最終到達淺層含水層，未飽和層與含水層的相關水文地質特性也會影響地下水硝酸鹽氮濃度。因此要藉由收集來自政府單位(包括內政部、經濟部水利署、經濟部地調所、環保署、農委會與內政部國土測繪中心等)可能影響地下水硝酸鹽氮濃度的人為活動、水文地質與地下水水質等資料作為預測變量，並以硝酸鹽氮濃度(包括經濟部水利署與環保署)做為反應變數，以此預測變量與反應變數進行訓練與測試發展地下硝酸鹽氮污染潛勢的 AI 預測模式，經訓練與測試所發展預測模式可用來描繪出可能的硝酸鹽氮高污染潛勢區域與範圍地圖，並以此地圖研提地下水水質預警與保護措施，圖 7 展示本研究流程圖，以下就相關研究方法與資料說明。

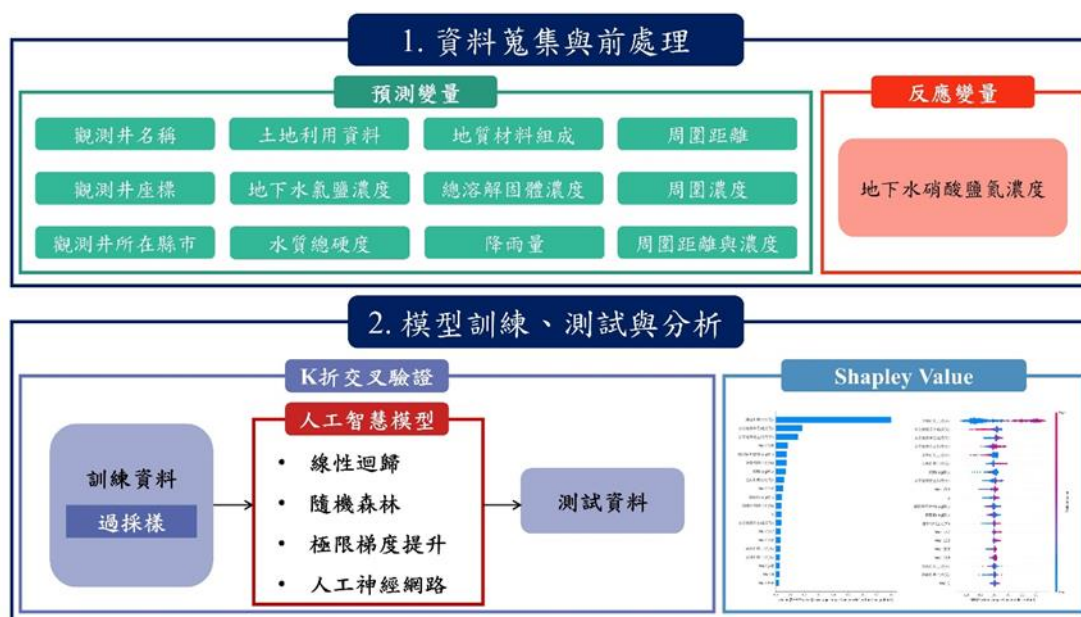


圖 7 研究流程圖



2.1 政府跨部會預測變量與反應變數資料

2.1.1 地下水水質

目前台灣地區經常性地下水水質監測基本上有兩大系統：環保署與經濟部水利署，本專案同時考慮環保署與經濟部水利署的地下水水質監測資料。

採用地下水水質資料來自 2018 年經濟部水利署淺層 82 口地下水觀測井資料(如圖 8)，監測頻率約一年一次，經濟部水利署的地下水水質參數監測項目有大腸桿菌群(CFU/100 mL)、氯鹽(mg/L)、硫酸鹽(mg/L)、硝酸鹽氮(mg/L)、氨氮(mg/L)、硫化物(mg/L)、總有機碳(mg/L)、銅(mg/L)、鎘(mg/L)、鉛(mg/L)、鋅(mg/L)、鉻(mg/L)、鎳(mg/L)、鐵(mg/L)、錳(mg/L)、鈣(mg/L)、鎂(mg/L)、鈉(mg/L)、鉀(mg/L)、碳酸氫根(mg/L)、碳酸根(mg/L)、砷(mg/L)、汞(mg/L)、總溶解固體(mg/L)、硬度(mg/L)。

另一部分地下水水質資料來自 2018 年行政院環保署全台 360 口地下水水質監測井資料(如圖 9)，監測頻率約每一個月一次，環保署水質參數監測項目有導電度($\mu\text{mho}/\text{cm}25^\circ\text{C}$)、酸鹼值、溶氧(電極法)(mg/L)、氧化還原電位(mV)、總硬度(mg/L)、總溶解固體物(mg/L)、氯鹽(mg/L)、氨氮(mg/L)、硝酸鹽氮(mg/L)、硫酸鹽(mg/L)、總有機碳(mg/L)、砷(mg/L)、鎘(mg/L)、鉻(mg/L)、銅(mg/L)、鉛(mg/L)、鋅(mg/L)、鐵(mg/L)、錳(mg/L)、汞(mg/L)、鎳(mg/L)、總酚(mg/L)、氯鹽(mg/L)、鈉(mg/L)、鉀(mg/L)、鈣(mg/L)、鎂(mg/L)、總鹼度(mg/L)。

將經濟部水利署與環保署資料合併，統整 2018 全年度監測站水質數據，作為地下水水質監測之基本資料來源，共計 442 口(如圖 10)，行政院環保署飲用水水質標準與農委會之灌溉水質標準(表 1)。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

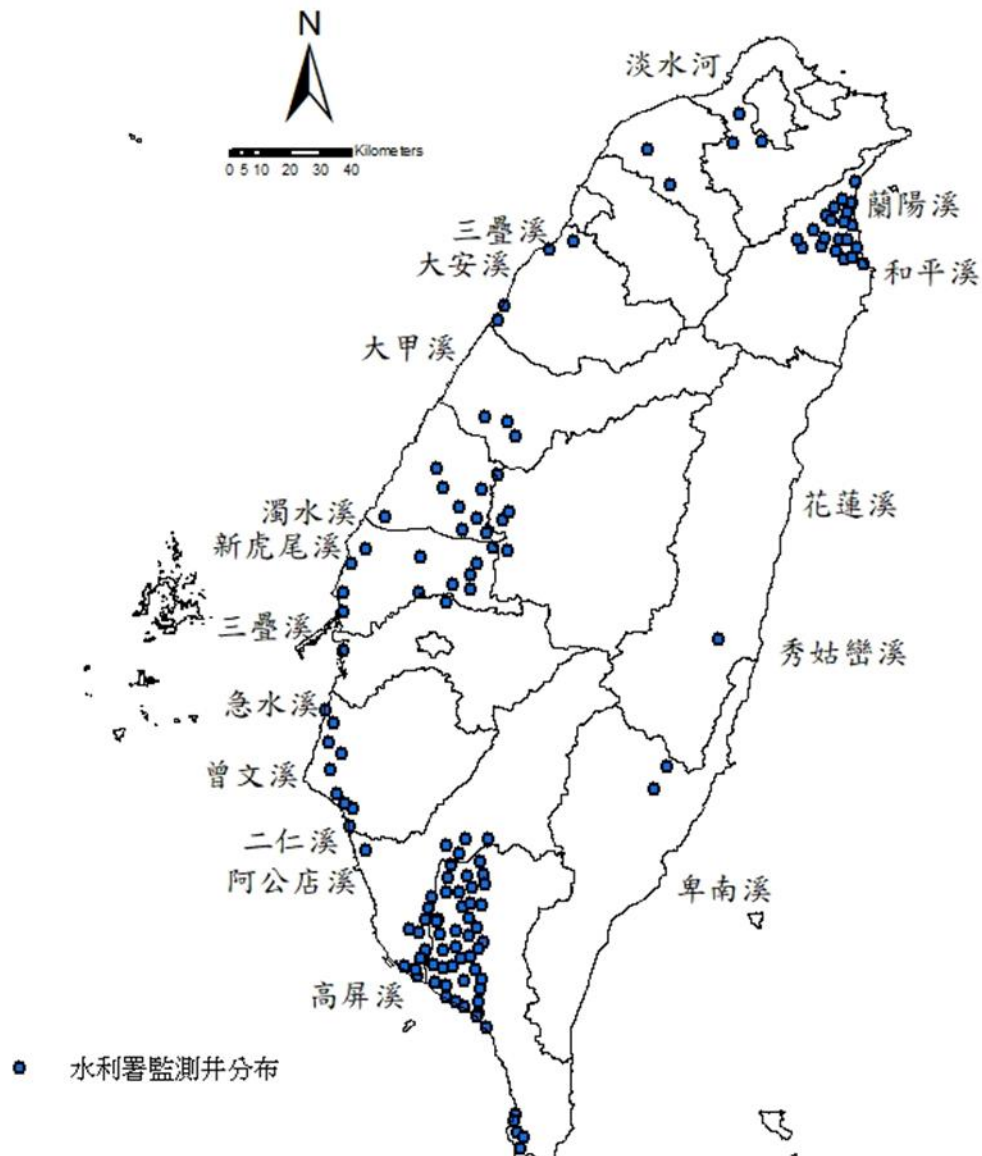


圖 8 經濟部水利署 82 口地下水水質監測的觀測井位置圖



二、研究方法

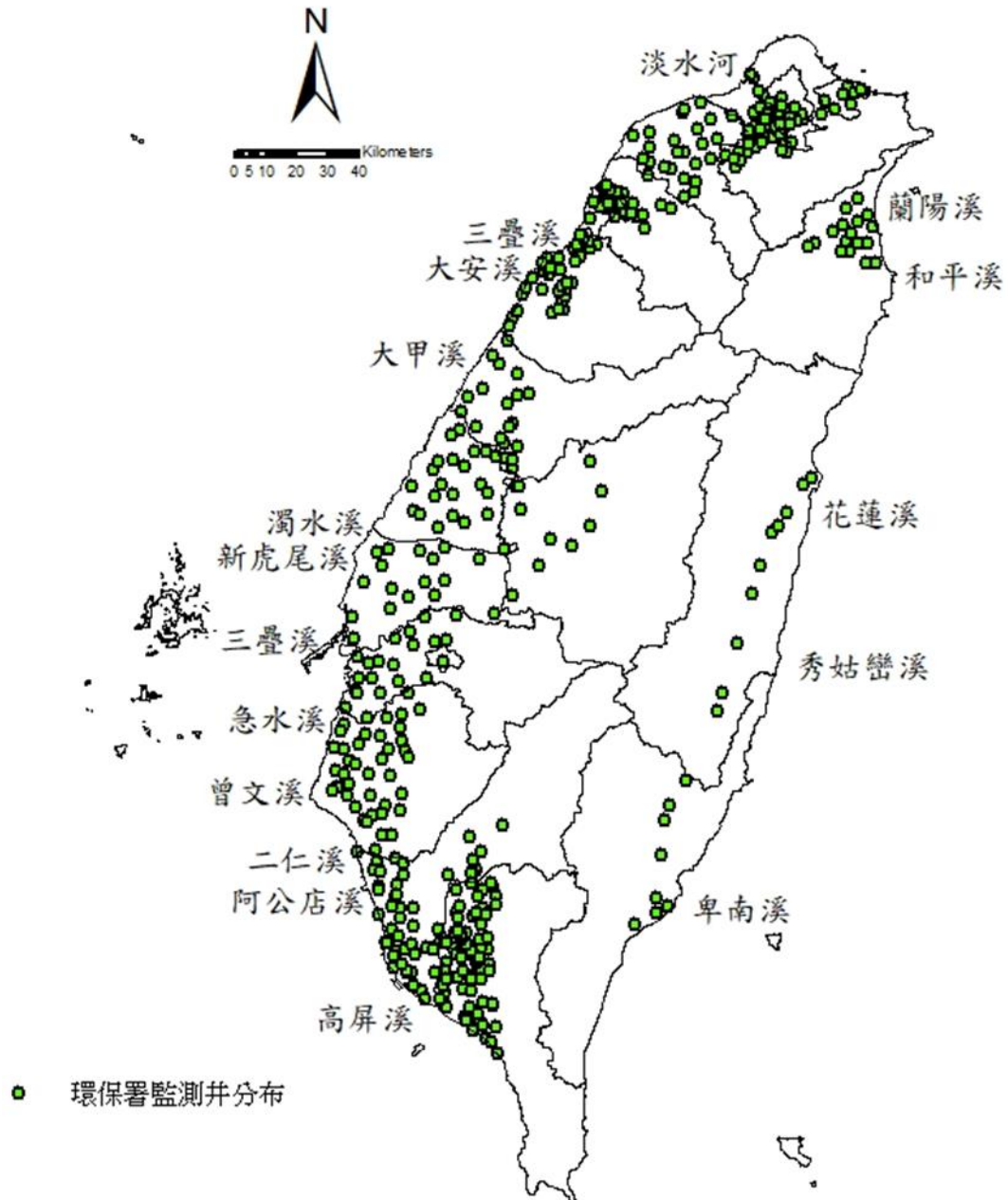


圖 9 環保署 360 口地下水水質監測井位置圖



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

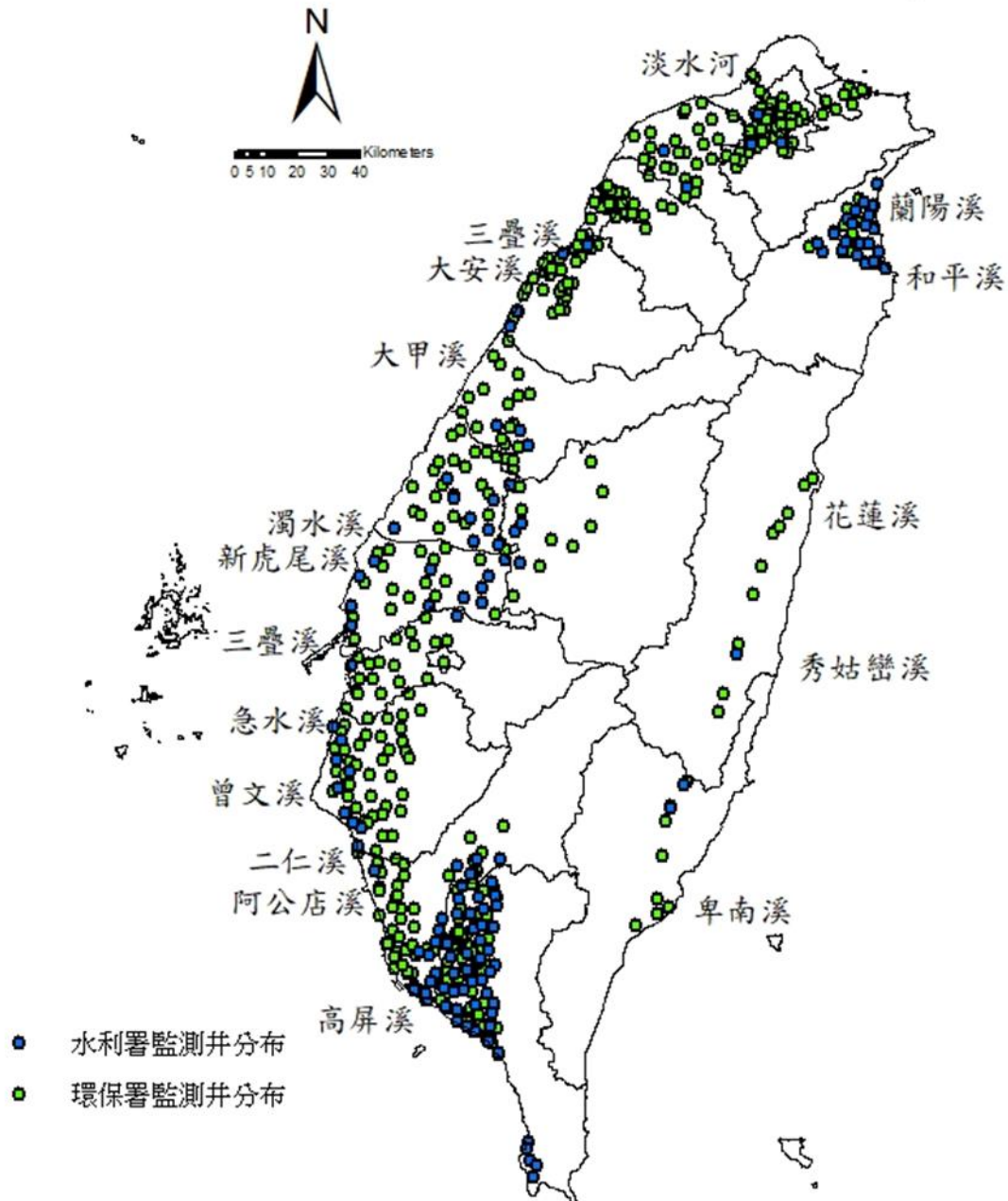


圖 10 經濟部水利署與環保署共計 442 口地下水水質監測井或觀測井位置圖



二、研究方法

表 1 行政院環保署飲用水水質標準與農委會之灌溉水質標準

水質項目	飲用水水質標準(mg/L)	灌溉水質標準(mg/L)
As[1]	0.01	0.05
Cd[1]	0.005	0.01
Cr[1]	0.05	0.1
NO ₃ ⁻ -N[2]	10.0	-
Fe	0.3	5.0
Mn	0.05	0.2
Cu	1.0	0.2
Zn	5.0	2.0
Cl ⁻	250	175
SO ₄ ²⁻	250	200
NH ₄ ⁺ -N	0.1	-



2.1.2 地質材料

當地面污染物隨入滲水進入地表下經由未飽和層，未飽和層的材料特性對污染物的遷移會有重要影響，當地層材料組成為礫石時，顆粒間的孔隙較大，地下水較易向下入滲且流動較為快速。反之，若是地層中的黏土較多，因黏土顆粒的孔隙小，不利於地下水的流動。本專案考慮非飽和層地質材料的影響，將沉積物依粒徑大小分類成礫石、砂土、粉土及黏土(如表 2)，水利署監測井本身就有記錄井之岩心資料，因此經濟部水利署的觀測井的地質材料直接取用經濟部水利署-水文資訊網紀錄之鑽探的岩心紀錄。環保署監測井因著重在地下水水質，並未記錄監測井之岩心資料，因此使用井周遭的中央地質調查所-工程地質資料庫資料作為未飽和層地質材料資料。由相關資料庫取得岩心資料可計算在 442 口井的未飽和層中礫石、砂土、粉土與黏土的各自厚度，將各材質厚度後除以未飽和層厚度(監測井地下水水位深度)，就可以計算各井的地質材料組成百分比。



二、研究方法

表 2 沉積物粒徑分類

分類	名稱	粒徑(mm)
水文地質礫石組成(%)	巨礫	>250.0
	粗礫	50.0~250.0
	中礫	10.0~50.0
	細礫	2.0~10.0
水文地質砂土組成(%)	極粗砂	1.00~2.00
	粗砂	0.50~1.00
	中砂	0.25~0.50
水文地質粉土組成(%)	細砂	0.125~0.25
	極細砂	0.063~0.125
水文地質黏土組成(%)	粉砂	0.004~0.063
	黏土	<0.004

資料來源:中央地質調查所



2.1.3 土地利用

人為活動會產生地下水污染來源，反應人類活動的最重要特徵是土地利用，因此土地利用型式會與地下水水質高度相關，為納入考慮土地利用，計畫中將採用內政部國土測繪中心的第二次國土利用調查分類系統，此分類系統總共包含三級分類(第一級分類、第二級分類與第三級分類)，第一級分類區分為包含農業(01)、森林(02)、交通(03)、水利(04)、建築(05)、公共(06)、遊憩(07)、礦鹽(08)與其他使用土地共 9 類(如圖 11)。第二級分類則將農業(01)又分為農作(0101)、水產養殖(0102)、畜牧(0103)、農業附帶設施(0104)等 4 類，森林(02)則再區分成天然林(0201)、人工林(0202)、其他森林利用土地(0203)等 3 類(如圖 12)。在第三級分類將農作(0101)又分為稻作(010101)、旱作(010102)、果樹(010103)與廢耕地(010104)等 4 類，而畜牧(0103)則再分為畜禽舍(010301)與牧場(010302)2 類(如圖 13)。

本計畫使用第一級分類，和第三級分類。第一級分類包含農業利用土地(%)、森林利用土地(%)、交通利用土地(%)、水利利用土地(%)、建築利用土地(%)、公共利用土地(%)、遊憩利用土地(%)、礦鹽利用土地(%)、其他利用土地(%)。第三級分類部分，我們使用第一級農業利用土地下的第三級分類。包含稻作利用土地(%)、旱作利用土地(%)、果樹利用土地、廢耕地利用土地(%)、畜牧利用土地(%)，除了上述五種第三級分類，其餘在農業利用土地底下的第三級分類皆被劃分到農業其他利用土地(%)。

由於土地利用資料為分類資料，較難取得數值資料進行分析，因此本研究採取 Heißerer et al. (2016) 所提出以計算井周圍圓形區域距離內各類土地利用組成面積的百分比做為數值資料分便進行數值相關計算，本研究採用井周圍半徑 1000 英尺分別計算 9 項土地利用型式的面積百分比。如圖 14 所示，以環保署位在東華大學的監測井為例，本研究以井為中心向外延伸 1,000 公尺為半徑畫圓，計算圓形區域內各土地利用分類面積占圓面積的百分比。



二、研究方法

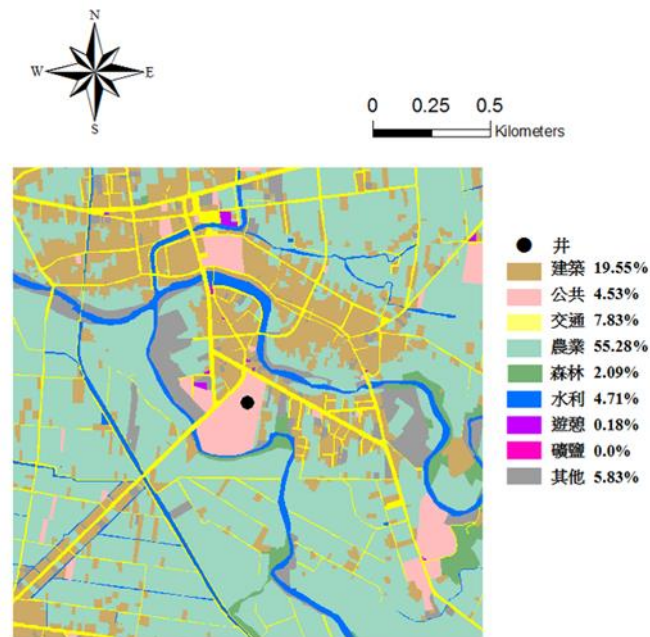


圖 11 第二次國土利用調查分類系統監測井週圍土地利用圖(第一級分類)

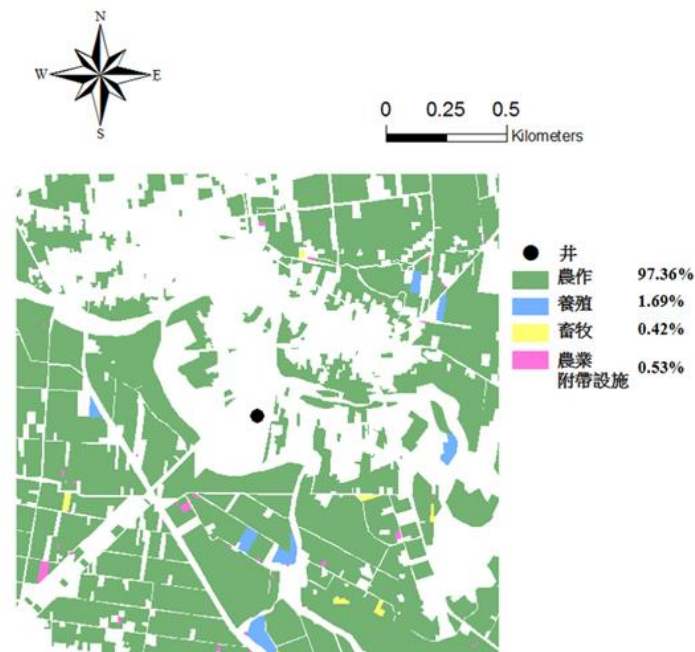


圖 12 第二次國土利用調查分類系統 (第二級分類，農業進一步區分為農作、養殖、畜牧與農業四類)



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

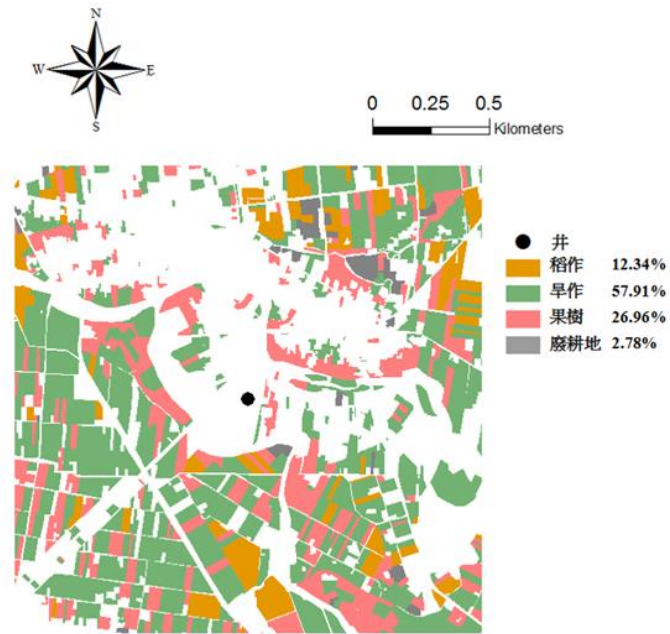


圖 13 第二次國土利用調查分類系統井週圍土地利用圖(第三級分類，農作再區分為稻作、旱作、果樹與廢耕地四類)

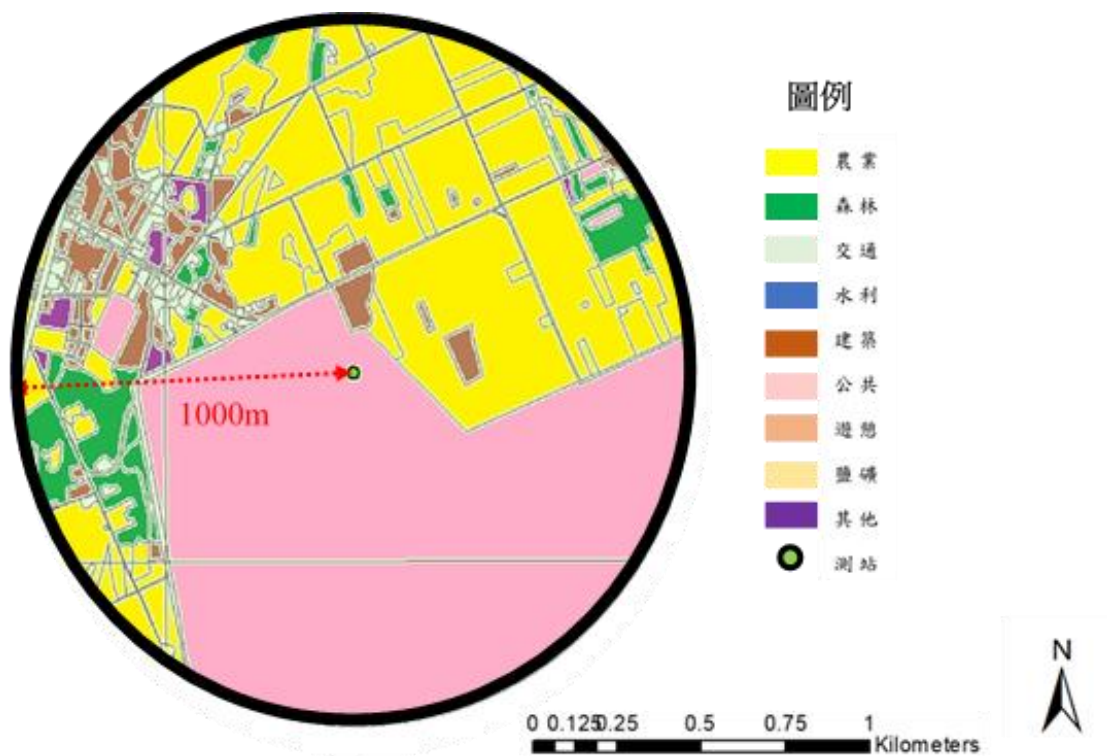


圖 14 以井周圍圓形區域距離內各類土地利用組成面積的百分比做為數值資料，便進行數值相關計算(此處考慮半徑 1,000m 範圍內之土地利用)



2.1.4 其他部會資料

根據 2018 年中央氣象局觀測資料查詢系統資料，由於觀測井/監測井的座標位置與中央氣象局測站的空間座標位置不完全相同，因此，採用距離各觀測井/監測井最近的氣象觀測站，計算觀測站的年降水量來使用。此外，本研究亦採用環保署在 2018 年的總溶解固體濃度、水質總硬度、地下水氯鹽濃度，此資料並未有空間座標位置不一致的問題。

2.1.5 由計算方法取得的預測變量

利用觀測水井與訓練資料彼此間的距離與其硝酸鹽氮濃度計算後，取得三種預測變量，分述如下：

- (a) 考量兩水井距離較遠，影響較小，因此，計算兩水井距離之後，將此數值取倒數，並移除最小的數值，作為新的預測變量。
- (b) 將周圍觀測水井的硝酸鹽氮濃度取對數後，移除最小的數值，其餘作為新的預測變量。
- (c) 基於前兩項的精神，綜合考量周圍觀測水井的距離與硝酸鹽氮濃度，將兩數值相除，做為新的預測變量。

例如，若訓練資料有 a、b、c 三口水井，則 a 水井會有三種(每次訓練只選一種或都不選)新增特徵的方式。

- (a) 新增 $1/a$ 與 b 水井距離、 $1/a$ 與 c 水井距離。
- (b) 新增 b 水井硝酸鹽濃度/a 與 b 水井距離、c 水井硝酸鹽濃度/a 與 c 水井距離。
- (c) 新增 b 水井的硝酸鹽濃度、c 水井的硝酸鹽濃度。



2.2 ANN 模式

本專案主要採用 ANN 進行 AI 模型的建構，ANN 是由一個非線性的神經元(運算單元)和連結這些神經元的權重所組合而成，是一種模仿生物神經網路(動物的中樞神經系統，特別是大腦)的結構和功能的數學模式或計算模式。簡言之，ANN 是從大數據中學習並建立模式。ANN 具有三個特性：學習、回想與歸納推演，學習是指 ANN 依學習演算法，自數據中學習，經反覆的運算，調整網路架構、活化函數、修正權重、門檻值，得到最好的模型，目前 ANN 已有多樣化且發展成熟的學習演算法；回想是當 ANN 接收到一個輸入值，會根據網路架構與之前的數據產生一個輸出值，且由於計算機科學的進步使得 ANN 的回想速度越來越快，若 ANN 的效能越好，其回想的過程更快速有效；最後，歸納推演是從一個系統中的局部觀察，而可以描述出其整體的特性，也就是找出盡可能適用於全部類別物件的重要因子，萃取特徵的能力，並提供一個有效率的記憶與儲存模式。

本計畫將以多層感知機(**M**ulti**L**ayer **P**erceptron, MLP)為架構建立硝酸鹽氮污染潛勢預測模型，圖 15 為 ANN 示意圖，包含輸入層、兩層隱藏層和輸出層等三種不同層來構成，其中，每層都由若干個神經元組成，各層的神經元以鍵結值來連接，以監督式學習方式來處理輸入與輸出之間的非線性映射關係。本計畫採用基於倒傳遞演算法(Backpropagation Algorithm)之倒傳遞神經網路(**B**ack **P**ropagation **N**eural **N**etwork, BPNN)，BPNN 網路訓練分成兩階段，分別為前饋階段與倒傳遞階段。前饋階段時，輸入向量由輸入層引入，以前饋方式經由隱藏層傳到輸出層，並計算出網路輸出值，此時，網路的鍵結值都是不變的；倒傳遞階段時，網路的鍵結值則根據錯誤更正法則來進行修正，藉由鍵結值的修正，以使網路的輸出值趨向於期望輸出值。

本專案以收集來自政府跨部會的資料作為 ANN 之輸入層，並以硝酸鹽氮濃度、是否超標等作為輸出層之預測目標，以此訓練與測試硝酸鹽氮污染潛勢 AI 預測模式。本專案採用 K 折交叉驗證(K-fold cross validation, K-fold CV)方式評估 ANN 模型好壞，K 折交叉驗證將資料集切割為 K 個子集合，訓練模型時利用 K-1 個子集合訓練模型，被排除的子集合作為驗證模型的數據，針對此驗證數據及計算模型之評估指標，重複 K 次，亦即每個子集合將被驗證一次，最後，根據 K 次的評估指標之平均值作為模型表現之依據。

除 ANN 模式外，本專案也另外考慮 LR、RF 與 XGBoost。



二、研究方法

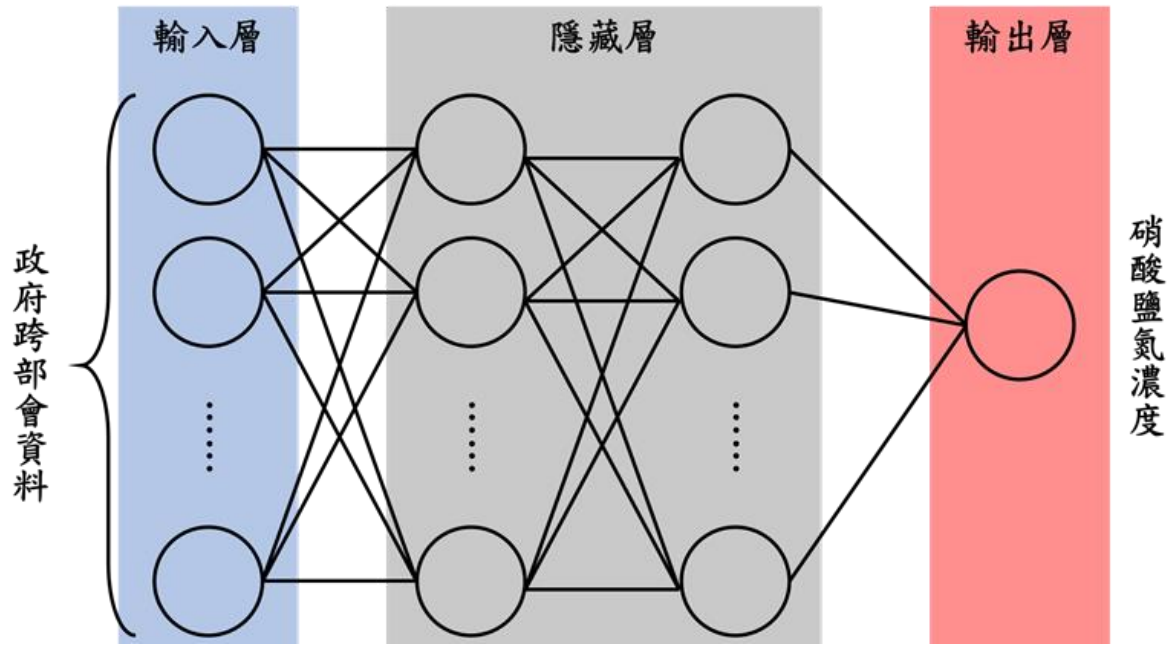


圖 15 ANN 架構示意圖



三、結果與討論

3.1 模型建置的資料前處理

在進行模型的建置須先進行資料的前處理，以下就資料前處理說明：

為了表示不同監測井的差異，使模型能學習不同監測井的資訊，將測站名稱和測站所在縣市名稱利用標籤編碼（label encoding）方式，轉換成可用來訓練模型的數值資料，本計畫共計 442 口井，因此各水井將採用 0 至 441 分別標示編碼。本計畫總共包含臺灣本島的 17 個縣市，因此，各縣市將採用 0 至 17 分別標示。

對硝酸鹽氮首先進行繪製濃度分布圖以了解資料概況，由於濃度常常會有多個數量級的變化，因此分別考慮橫座標的硝酸鹽濃度為未取對數(算術尺度)與取對數兩種方式繪製。圖 16 左側之硝酸鹽氮濃度分布圖顯示有離群值，為避免因離群值影響模型之訓練，移除該離群值。另一方面，為觀察轉換硝酸鹽氮濃度對模型訓練之影響，利用對數轉換硝酸鹽氮濃度，由於每次量測極限略有不同，將硝酸鹽氮濃度低於 0.02 之數值，皆當作 0.02 之後，再取對數轉換，轉換結果如圖 16 右圖所示。

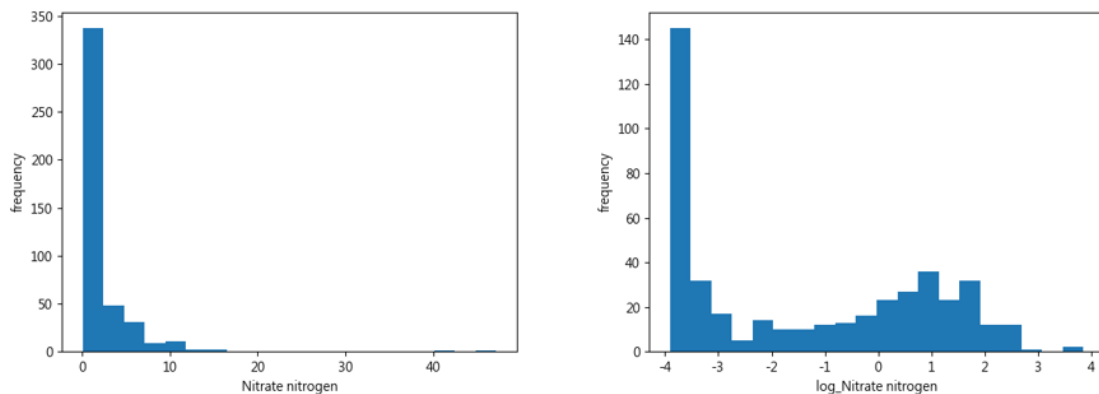


圖 16 硝酸鹽氮濃度取對數前後的資料分佈，左圖為取對數前的硝酸鹽氮濃度分佈；右圖為取對數後的硝酸鹽氮濃度分佈



三、結果與討論

針對主要的土地利用以及水文地質組成等預測變量，表 3 展示了用於建構硝酸鹽氮濃度預測模型各預測變量之平均、最小與最大值，圖 14 與圖 15 分別展示土地利用以及水文地質組成之分布圖。

此外，為了觀測用不同底數取對數轉換會不會對訓練結果造成影響，因此做了底數為自然對數和底數為 10 的對數轉換。圖 17 為兩種底數結果取對數後的硝酸鹽氮濃度分佈圖(左圖為底數為自然對數取對數後的硝酸鹽氮濃度分布圖，右圖為底數為 10 取對數後的硝酸鹽氮濃度分布圖)。

針對主要的土地利用以及水文地質組成等預測變量，表 3 展示了用於建構硝酸鹽氮濃度預測模型各預測變量之平均、最小與最大值，圖 18、圖 19 與圖 20 分別展示土地利用以及水文地質組成以及其他部會資料之分布圖。

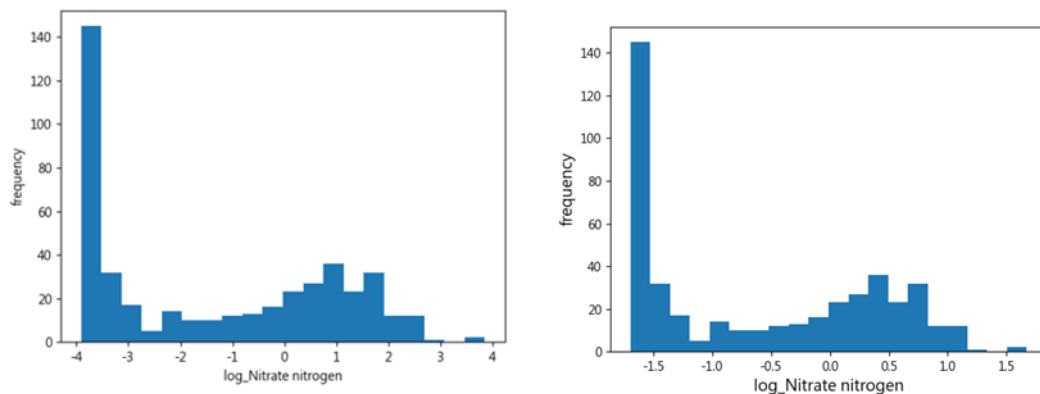


圖 17 硝酸鹽氮濃度取對數前後的資料分佈，左圖為底數是自然對數的硝酸鹽氮濃度分佈；右圖為底數是 10 的硝酸鹽氮濃度分佈



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

表 3 各預測變量之平均、最大、及最小值

預測變量	平均值	最大值	最小值
農業利用土地(%)	40.118	94.097	0.000
稻作利用土地(%)	11.685	79.108	0.000
旱作利用土地(%)	13.300	74.151	0.000
果樹利用土地(%)	9.400	87.781	0.000
廢耕地利用土地(%)	1.029	19.291	0.000
畜牧利用土地(%)	0.690	10.485	0.000
農業利用土地其他(%)	4.121	77.198	0.000
森林利用土地(%)	6.075	89.249	0.000
交通利用土地(%)	12.179	74.863	1.689
水利利用土地(%)	5.882	67.989	0.000
建築利用土地(%)	21.823	89.532	0.176
公共利用土地(%)	4.017	44.508	0.000
遊憩利用土地(%)	2.207	29.598	0.000
礦鹽利用土地(%)	0.430	47.226	0.000
其他利用土地(%)	6.913	47.502	0.000
水文地質礫石組成(%)	20.056	100.000	0.000
水文地質砂土組成(%)	37.093	100.000	0.000
水文地質粉土組成(%)	26.876	100.000	0.000
水文地質黏土組成(%)	16.018	100.000	0.000
降雨量	2152.639	4219.5	745.0
總硬度 (mg/L)	572.742	17300.0	12.8
總溶解固體物 (mg/L)	2476.710	95400.0	15.5
氯鹽 (mg/L)	1175.848	49500.0	1.3



三、結果與討論

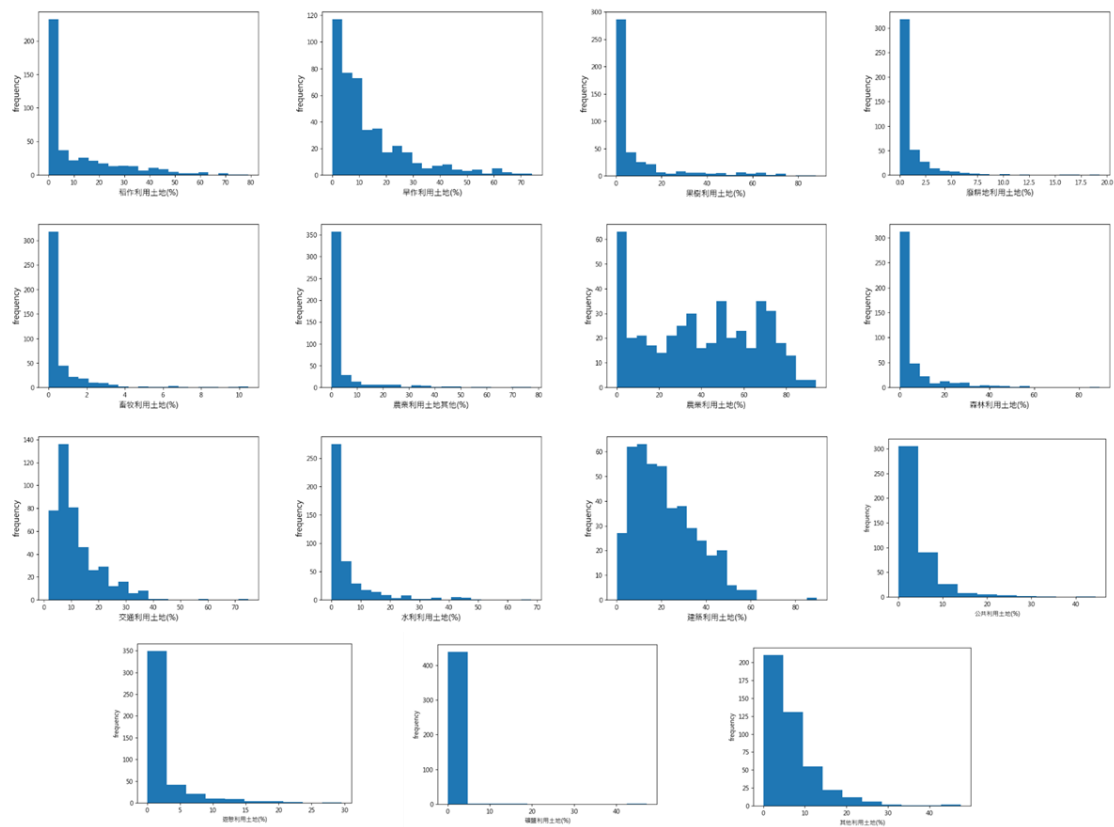


圖 18 各個土地利用的資料分佈圖，由左而右、由上到下，依序是稻作利用土地、旱作利用土地、果樹利用土地、廢耕地利用土地、畜牧利用土地、農業其他利用土地、農業利用土地、森林利用土地、交通利用土地、水利利用土地、建築利用土地、公共利用土地、遊憩利用土地、礦鹽利用土地、其他利用土地



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

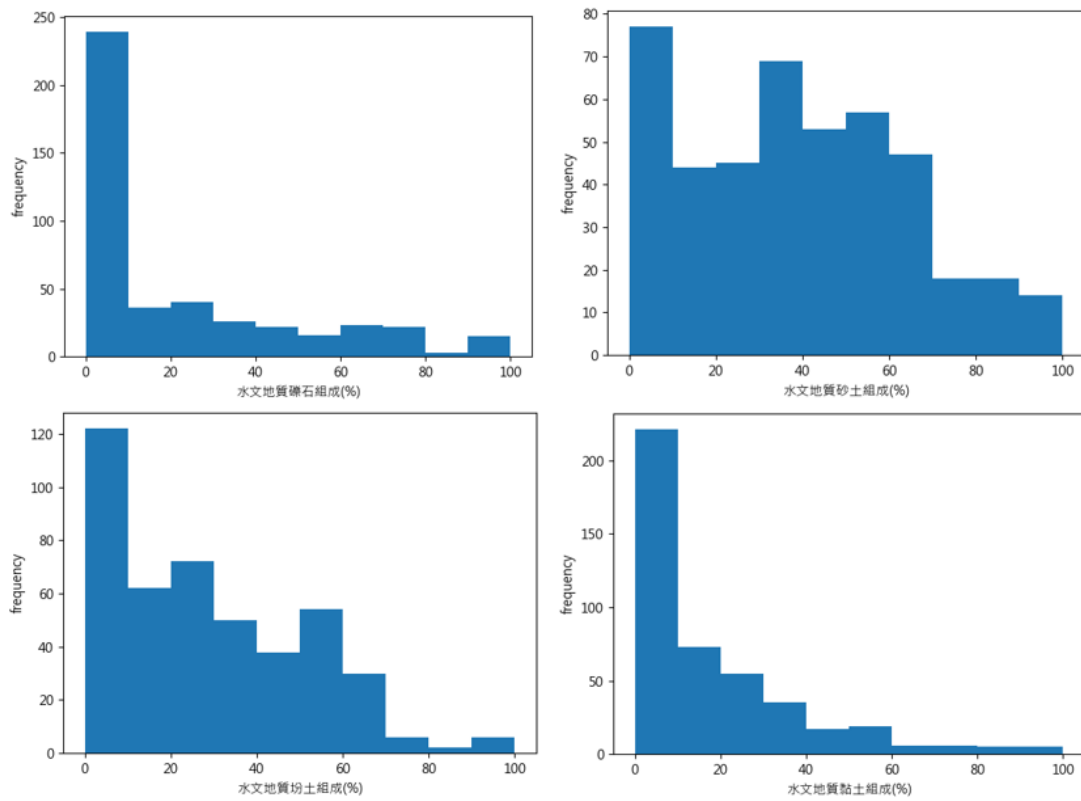


圖 19 各個水文地質組成的資料分佈圖，由左而右、由上到下，分別是水文地質礫石組成、水文地質砂土組成、水文地質粉土組成、水文地質黏土組成



三、結果與討論

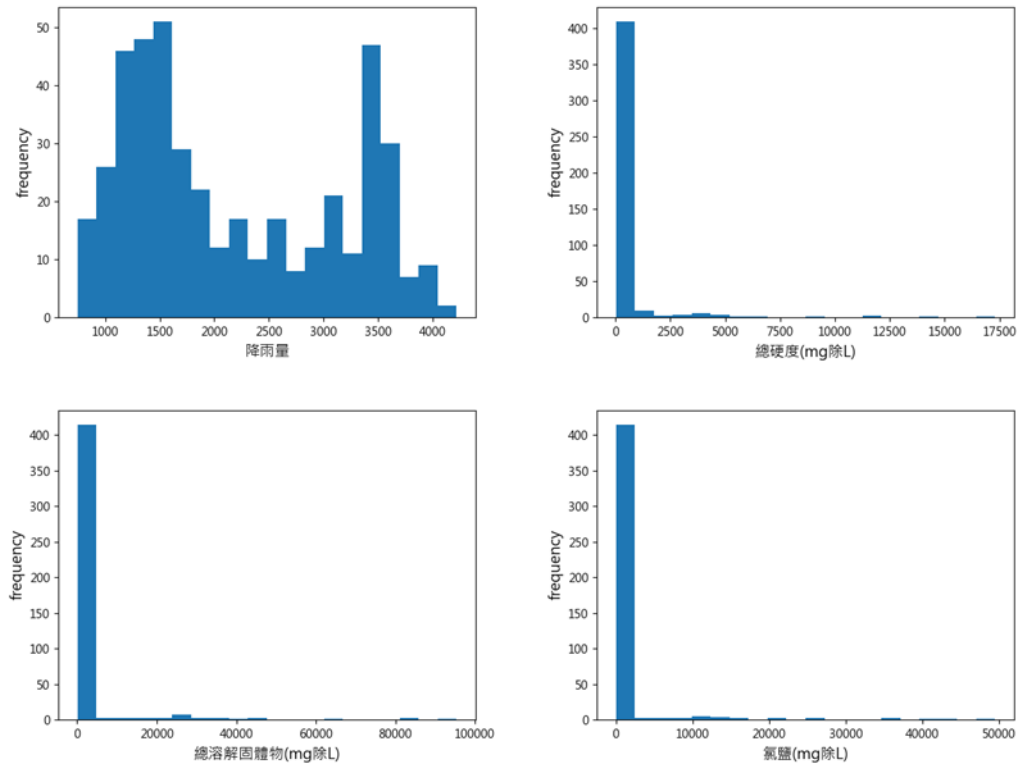


圖 20 其他部會資料，由左而右、由上到下，分別是降雨量、總硬度 (mg/L)、總溶解固體物 (mg/L)、氯鹽 (mg/L)



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

為避免用於建置模型之土地利用與水文地質組成的資料差異過大，本計畫使用 4 種方式對各資料進行標準化，分述如下：

(a) 最小—最大值標準化 (Min-Max normalization)

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

其中， x 表示某個土地利用或是水文地質組成資料， x_{scaled} 表示該轉換資料， $x_{\min}$ 與 $x_{\max}$ 分別為某個土地利用或是水文地質組成資料之最小與最大值。

(b) Z-Score 標準化 (Z-score normalization)

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \quad (2)$$

其中， x 表示某個土地利用或是水文地質組成資料， x_{scaled} 表示該轉換資料， μ 與 σ 分別為某個土地利用或是水文地質組成資料之平均值和標準差。

(c) Maximum Absolute 標準化 (Maximum Absolute Scaling)

$$x_{\text{scaled}} = \frac{x}{x_{\max}} \quad (3)$$

其中， x 表示某個土地利用或是水文地質組成資料， x_{scaled} 表示該轉換資料， $x_{\max}$ 為某個土地利用或是水文地質組成資料之最大值。

(d) Robust 標準化 (Robust Scaling)

$$x_{\text{scaled}} = \frac{x - \text{median}(x)}{3\text{rd quartile}(x) - 1\text{st quartile}(x)} \quad (4)$$

其中， x 表示某個土地利用或是水文地質組成資料， x_{scaled} 表示該轉換資料， $3\text{rd quartile}(x)$ 與 $1\text{st quartile}(x)$ 分為某個土地利用或是水文地質組成資料之第一四分位數和第三四分位數。

由於利用 Maximum Absolute 進行標準化達到最佳結果，因此本計畫後續成果將以此進行建模與分析。

除了前述之土地利用與水文地質組成，本計畫加以考量監測井之經緯度座標，並且以臺灣中心點(經度：120.98202，緯度：23.97387)為原點轉換座標¹。

¹ 臺灣經緯度原點：<https://zh.wikipedia.org/zh-tw/%E5%8F%B0%E7%81%A3%E7%B6%93%E7%B7%AF%E5%BA%A6%E5%8E%9F%E9%BB%9E>



3.2 人工智慧方法與評估指標

本專案除使用 ANN 外，也考慮其它 3 種人工智慧模型，包括 LR、RF 與 XGBoost 分別觀察硝酸鹽濃度取對數前與取對數後之差異。

1. 4 種模型於未加入各部會資料和計算方法取得的預測變量時硝酸鹽氮濃度未取對數之設定：

- (a) 線性迴歸：

利用 python 的 sklearn 提供的 LinearRegression 方法建構硝酸鹽氮濃度預測模型。

- (b) 隨機森林：

利用 python 的 sklearn 提供的 RandomForestRegressor 方法建構硝酸鹽氮濃度預測模型。採用樹木數量 100、80、75、60 進行測試，結果顯示，數量 75 之預測結果最佳。因此，將樹木的數量設定為 75。

- (c) 極限梯度提升：

利用 python 的 XGBoost 建構硝酸鹽氮濃度預測模型。採用樹木數量 100、80、60 進行測試，結果顯示，數量 60 之預測結果最佳，若再增加，反而導致過度擬合(overfitting)，因此，為了避免此情形，將樹木的深度設定為 2，樹木數量設定為 100。

- (d) 人工神經網路：

利用 python 的 sklearn 提供的 MLPRegressor 方法建構硝酸鹽氮濃度預測模型。採用三種激活函數 (Activation function) 進行測試，包含線性整流函數 (rectified linear unit, relu)、雙曲正切激活函數 (hyperbolic tangent activation function, tanh)、恆等函數(identity_)，結果顯示線性整流函數為激活函數時模型預測結果較佳。此外，測試兩層神經元個數為(200, 100)、(150, 50)、(75, 50) 等情形，結果顯示 (75, 50) 的預測表現最佳，同時，神經元個數不足會導致模型無法收斂。因此，採用兩層神經元個數分別為 75 與 50 的架構，並使用線性整流函數作為激活函數、自適應矩估計 (adaptive moment estimation, Adam) 作為優化器、正則化參數設為 0.001、初始學習率設為 0.001、最大迭代次數為 2000、以及損失值容忍閥值 (表示小於該值時停止訓練) 為 0.0001。

2. 4 種模型於未加入各部會資料和計算方法取得的預測變量時硝酸鹽氮濃度取



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

對數之設定：

(a) LR：

利用 python 的 sklearn 提供的 LinearRegression 方法建構硝酸鹽氮濃度預測模型。

(b) RF：

利用 python 的 sklearn 提供的 RandomForestRegressor 方法建構硝酸鹽氮濃度預測模型。採用樹木數量 100、75、60 進行測試，結果顯示，數量 100 之預測結果最佳。因此，將樹木的數量設定為 100。

(c) XGboost：

利用 python 的 XGBoost 建構硝酸鹽氮濃度預測模型。採用樹木數量 100、50、10 進行測試，為避免過度擬合，將樹木數量設定為 50。

(d) ANN：

利用 python 的 sklearn 提供的 MLPRegressor 方法建構硝酸鹽氮濃度預測模型。採用三種激活函數(activation function)進行測試，包含線性整流函數(rectified linear unit, relu)、雙曲正切激活函數(hyperbolic tangent activation function, tanh)、恆等函數(identity)，結果顯示，若採用評估迴歸模型之指標，三個激活函數沒有太大差異，然而，正切激活函數容易預測單一相同的數值。若採用評估分類模型之指標，恆等函數較線性整流函數佳。此外，在活化函數為線性整流函數時，測試兩層神經元個數分別為 (1500, 1500)、(500, 500)、(350, 350)，結果顯示 (150, 150) 的預測表現最佳，同時，神經元個數不足會導致模型無法收斂。因此，採用兩層神經元個數分別為 150 與 150 的架構，並使用恆等函數作為激活函數、自適應矩估計(adaptive moment estimation, Adam)作為優化器、正則化參數設為 0.001、初始學習率設為 0.055、最大迭代次數為 2000、以及損失值容忍閾值（表示小於該值時停止訓練）為 0.0001。

3. 4 種模型於加入各部會資料和計算方法取得的預測變量時硝酸鹽氮濃度取對數之設定：

(a) LR：

利用 python 的 sklearn 提供的 LinearRegression 方法建構硝酸鹽氮濃度預測模型。



(b) RF：

利用 python 的 sklearn 提供的 RandomForestRegressor 方法建構硝酸鹽氮濃度預測模型。因將樹木的數量設定為 20。

(c) XGBoost：

利用 python 的 XGBoost 建構硝酸鹽氮濃度預測模型。將樹木數量設定為 20。深度設置為 2。

(d) ANN：

利用 python 的 sklearn 提供的 MLPRegressor 方法建構硝酸鹽氮濃度預測模型。採用線性整流函數 (Rectified linear unit, relu) 神經元個數為 (256, 128, 64, 32) 使用自適應矩估計 (Adaptive moment estimation, Adam) 作為優化器，初始學習率設為 0.001、最大迭代次數為 1500、以及損失值容忍閾值 (表示小於該值時停止訓練) 為 0.0001。

為了確認 27 個預測變量對模型的訓練皆有幫助，因此我們做了遞迴特徵消除(recursive feature elimination, RFE)。在 RFE 當中，首先會先訓練一個機器學習模型，在此我們用 XGBoost Model，並將所有的預測變量放進去模型算各個預測變量對於預測硝酸鹽濃度的重要程度。接著從對於預測硝酸鹽濃度重要程度較小的預測變量開始刪除，並在預測變量集合上不斷重複的回上述步驟，直到最終達到所需的預測變量數量為止。。

本專案藉由上述方法，測試模型用 N 個($27 \geq N \geq 1$)重要度最高的預測變量建模對於預測硝酸鹽濃度的影響。結果發現若是不消除任意一個預測變量對於模型的準確度會是最好的。顯示此次實驗收集的每一個預測變量皆是重要的。

本計畫採用交叉驗證技術建立和驗證預測結果，驗證使用 K 折交叉驗證法。在 K-fold 的方法中，我們一開始會將資料切割成 K 份。接著訓練 K 次，每次訓練都從這 K 份中挑選 K-1 份會用於做訓練集資料，剩下的 1 份作為測試集資料。並且這 K 次的訓練中，測試集資料不會重複。此次計畫採用之 K 值為 5，也就是每次用其中 80% 的資料當訓練集，20% 的資料當測試集。在加入其他部會資料後，我們將 K 擴充成 15，以擴充訓練資料集的大小。

為了讓除了高濃度 (硝酸鹽濃度大於 10) 的資料可以在 15 個 fold 內平均分佈，我們對除了高濃度 (硝酸鹽濃度大於 10) 的資料資料做了 Kmeans 分群。在 Kmeans 的方法中，我們會先決定要分成 K 群，然後在資料中先選出 K 個點當作群集中心。接下來我們會將每一個資料點分群到離自己最近的群集中心，全部分



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

完之後重新計算各群的群集中心。反覆迭代下去直到群集中心不變或該次每群的群集中心與上次每群的群集中心小於誤差值為止。此次計畫 K 值為 3，在分完 3 群後，我們將每群的資料平均分配到 15 個 fold 中。並且為了確保每個 fold 皆具有高濃度資料，我們將資料中的每個高濃度水井平均分配到 15 個 fold 中。

由於濃度大於 10 以上的資料集只有 15 口水井，因此我們對資料做了過採樣 (oversampling) 的處理，將分成三類的硝酸鹽濃度資料做虛擬少數類別過抽樣 (Synthesized **M**inority **O**versampling **T**echnique)，此方法會對少樣的樣本進行人工合成，以增加少樣樣本，用以增加少樣樣本分類的準確率。

評估的指標包含均方根誤差 (Root Mean Square Error, RMSE)，以及 R^2 其定義如下：

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (5)$$

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

其中 n 為資料筆數， $\hat{y}$ 為預測硝酸鹽氮濃度， $\bar{y}$ 為資料中硝酸鹽氮的平均濃度，y 為真實硝酸鹽氮濃度。

硝酸鹽氮濃度是否超標為此計畫關注之重點之一，因此，本計畫除了採用前述用於評估迴歸模型的指標，也採用評估分類模型的指標，用於評估預測模型是否可成功預測該區域之硝酸鹽氮濃度超標。過去文獻發現當硝酸鹽濃度超過 1.37mg/L 時，會與兒童的腦癌之間有關係(Weng et al., 2011)，同時，根據環保署標準，地下水的硝酸鹽濃度若超過 10mg/L，則視為地下水遭硝酸鹽污染，因此，本計畫將硝酸鹽氮濃度分為三群，分別是低於 1.37mg/L 的第一類、濃度介於 1.37mg/L 與 10mg/L 的第二類、以及濃度超過 10mg/L 的第三類。在加入各部會的資料後，我們將分類改為低於 1mg/L 的第一類、濃度介於 1mg/L 與 10mg/L 的第二類、以及濃度超過 10mg/L 的第三類，根據此分類依據，分類模型之評估指標採用準確率 (Accuracy)，定義如下：

$$\text{準確率} = \frac{\text{正確預測個數}}{\text{資料總個數}} \quad (7)$$



三、結果與討論

此外，為了充分了解不同預測變量對於預測硝酸鹽濃度有最大的影響，因此也利用了 Shapley 值(Shapley value)評估。Shapley 值方法是根據博弈論中解決合作對策問題的概念，此方法根據玩家們(預測變量)在遊戲(模型)中得到的總支出(預測結果)算出每個玩家(預測變量)的貢獻度。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

3.3 模型建置結果分析

對未取對數之硝酸鹽氮濃度迴歸模型預測，圖 21 至圖 24 為刪除離群值後，採用 ANN、LR、RF 與 XGBoost 建構 AI 硝酸鹽氮濃度預測模型的實驗結果，其中每個演算法與參數皆是採用 5 折交叉驗證，圖中為 5 次訓練中最佳的結果。表 4 為 4 個模型採用迴歸模型評估指標的結果，指標於訓練及測試資料之差異，顯示 RF、XGB 以及 ANN 過擬合之情況較嚴重。

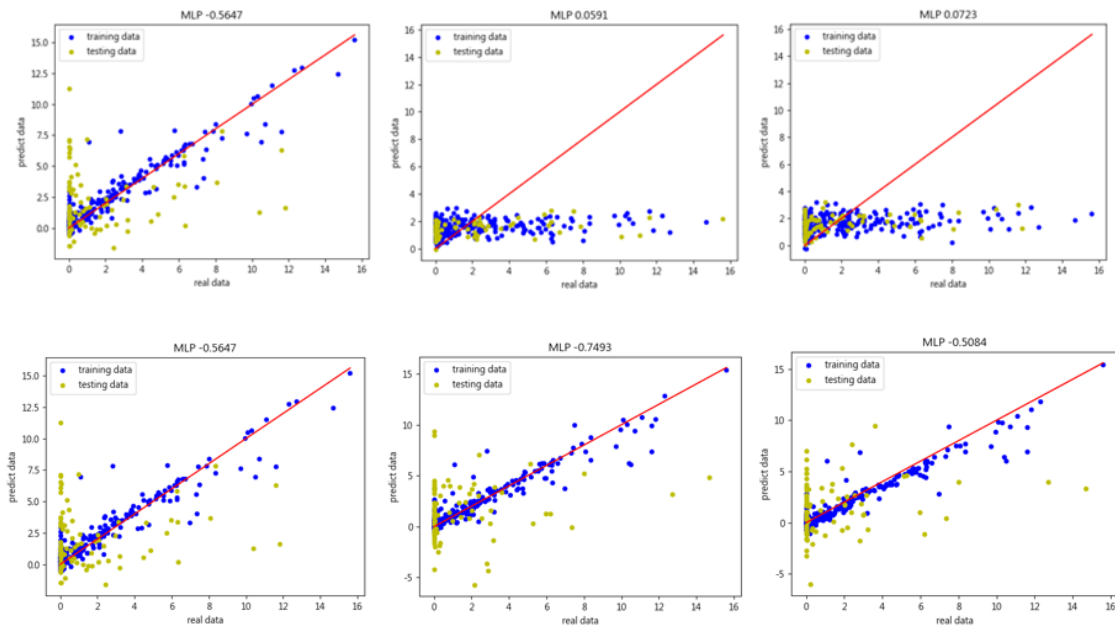


圖 21 不同 ANN 建模結果，上半部三張圖由左到右分別是使用線性整流函數、雙曲正切激活函數、恆等函數，且神經元皆是(200, 100)之架構；下半部三張圖由左到右分別是神經元(200, 100)、(150, 50)、以及(75, 50)，且激活函數為線性整流函數之結果

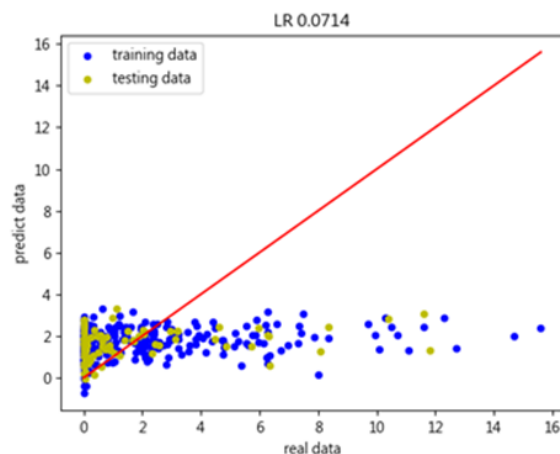


圖 22 MLR 模型預測結果



三、結果與討論

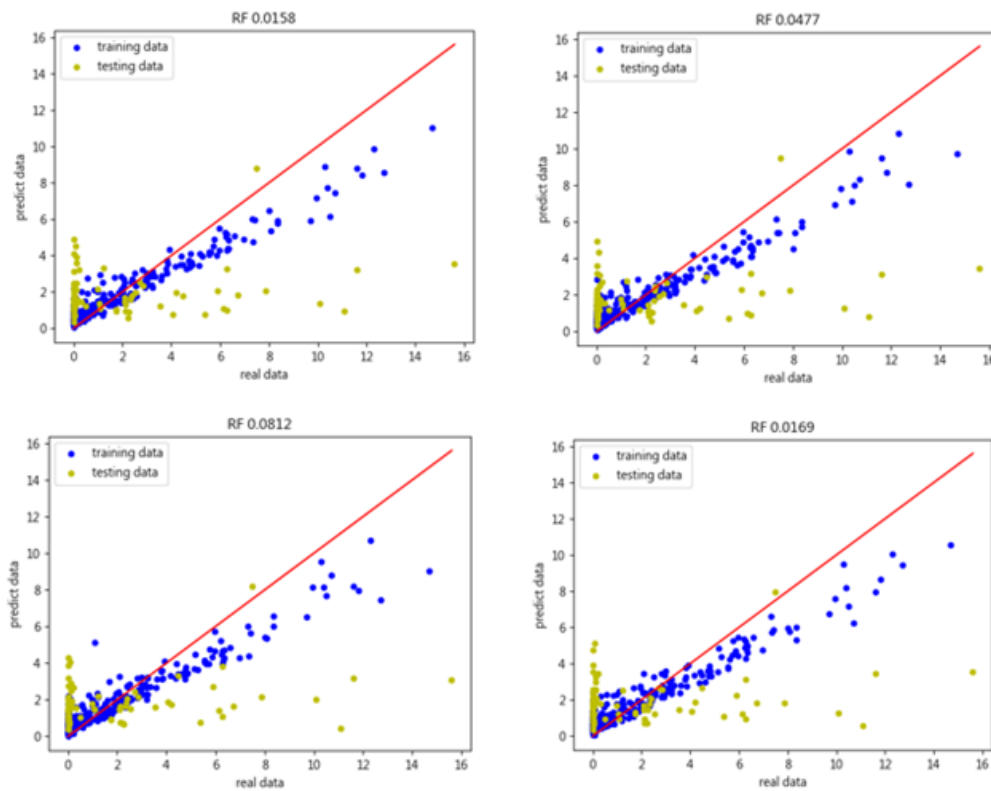


圖 23 RF 採用不同參數的迴歸模型預測結果，由左到右，由上到下分別是樹木數量 100、80、75、及 60 之結果

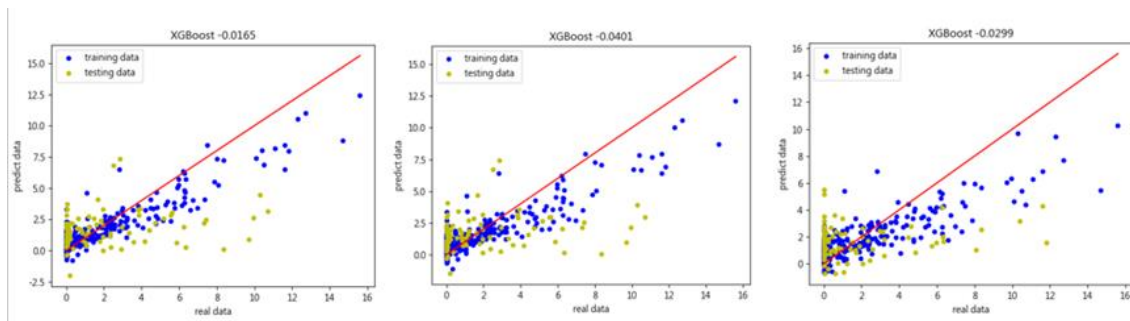


圖 24 由左到右分別是 XGBoost 採用樹木數量 100、80、及 60 的實驗結果



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

表 4 未取對數之硝酸鹽氮濃度預測迴歸模型評估結果

	訓練資料		測試資料	
	RMSE	R ²	RMSE	R ²
LR	2.606(0.0022)	0.0685(0.0001)	2.7294(0.0508)	-0.0382(0.0039)
RF	1.0262(0.0006)	0.8555(0.0)	2.7921(0.0508)	-0.0872(0.0064)
XGBoost	1.2024(0.0027)	0.8016(0.0002)	2.8529(0.0486)	-0.1395(0.0154)
ANN	0.8989(0.0241)	0.8856(0.0015)	4.1298(0.1595)	-1.3852(0.0832)

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。

不同方法建構的迴歸模型未取對數之硝酸鹽氮濃度預測結果如表 5 與表 6，於訓練及測試資料集中以分類指標評估的混淆矩陣(confusion matrix)，由於未取對數之硝酸鹽濃度高數值資料筆數較少，且與低數值的資料差異大，模型較易高估。圖 25 展示訓練與測試資料集 5 次交叉驗證的準確率盒圖，此結果顯示未取對數之硝酸鹽氮濃度預測模型於測試集的準確率明顯較訓練資料集差，此結果也與前述之迴歸模型結果一致，ANN、XGBoost、以及 RF 的分類準確度於訓練與測試資料有明顯差別，凸顯這些方法已過度擬合。

表 5 不同模型於訓練資料集的混淆矩陣

		方法	真實硝酸鹽氮濃度值 (y)		
			≤1.37	1.37<y<10	≥10
預測硝酸鹽氮 濃度 ($\hat{y}$)	≤1.37	LR	41	117	0
	1.37< $\hat{y}$ <10		9	96	0
	≥10		0	11	0
	≤1.37	RF	135	26	0
	1.37< $\hat{y}$ <10		1	104	0
	≥10		0	8	1
	≤1.37	XGBoost	112	49	0
	1.37< $\hat{y}$ <10		7	98	0
	≥10		0	7	2
	≤1.37	ANN	92	66	0
	1.37< $\hat{y}$ <10		25	80	0
	≥10		4	7	0

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。



三、結果與討論

表 6 不同模型於測試資料集的混淆矩陣

		方法	真實硝酸鹽氮濃度值 (y)		
			≤ 1.37	$1.37 < y < 10$	≥ 10
預測硝酸鹽氮 濃度 ($\hat{y}$)	≤ 1.37	LR	10	30	0
	$1.37 < \hat{y} < 10$		0	27	0
	≥ 10		0	2	0
	≤ 1.37	RF	12	25	0
	$1.37 < \hat{y} < 10$		6	21	0
	≥ 10		1	3	0
	≤ 1.37	XGBoost	18	19	0
	$1.37 < \hat{y} < 10$		7	20	0
	≥ 10		0	4	0
	≤ 1.37	ANN	16	24	0
	$1.37 < \hat{y} < 10$		5	22	0
	≥ 10		0	2	0

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。

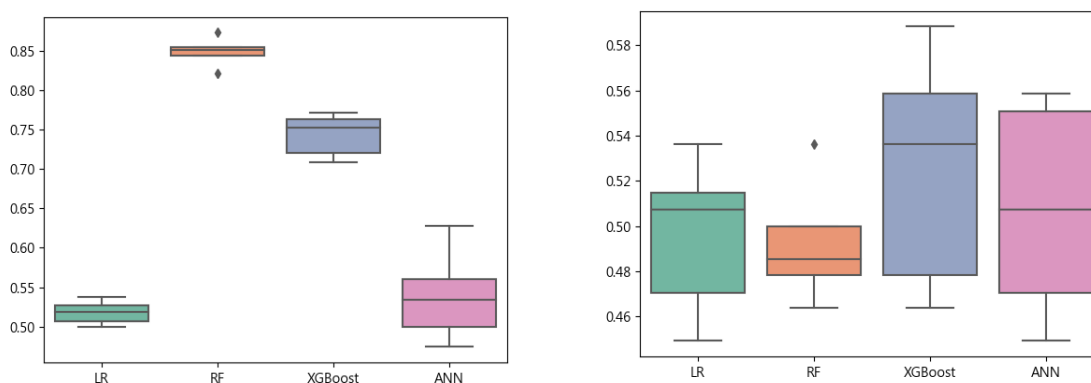


圖 25 未取對數之硝酸鹽氮濃度不同模型於訓練（左）與測試（右）資料集的準確率



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

對取對數之硝酸鹽氮濃度迴歸模型建置，並未過濾離群值，圖 26 至圖 29 為採用 ANN、LR、RF、以及 XGBoost 建構硝酸鹽氮濃度預測模型的實驗結果，同樣地，每個演算法與參數皆是採用 5 折交叉驗證，圖中為 5 次訓練中最佳的結果。結果顯示 RF 與 XGBoost 建構經對數轉換的硝酸鹽氮濃度迴歸模型，過擬合之問題仍嚴重，後續計畫執行將持續尋求適當方法解決。表 6 為 4 個模型採用迴歸模型評估指標的結果，同樣地，指標於訓練及測試資料之差異，顯示 RF、及 XGBoost 以及 ANN 過擬合之情況較嚴重。



三、結果與討論

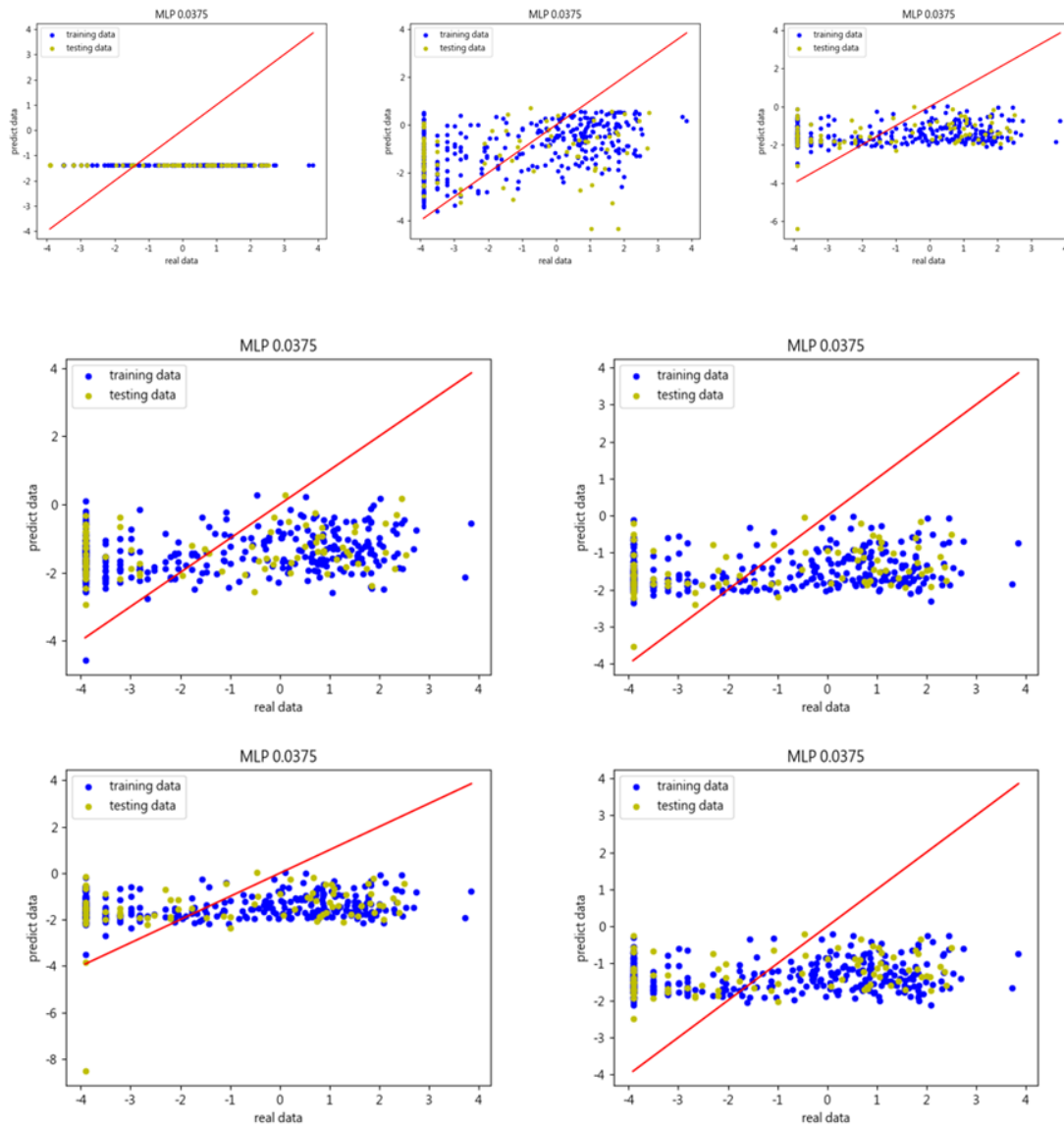


圖 26 ANN 建模結果，第一列的三張圖由左到右分別是使用線性整流函數、雙曲正切激活函數、恆等函數，且神經元皆是(350, 350)之架構；第二列兩張圖由左到右分別是神經元(150, 150)、以及(150, 50)；第三列兩張圖由左到右分別是神經元(500, 500)、以及(1500, 1500)，且激活函數為恆等函數之結果



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

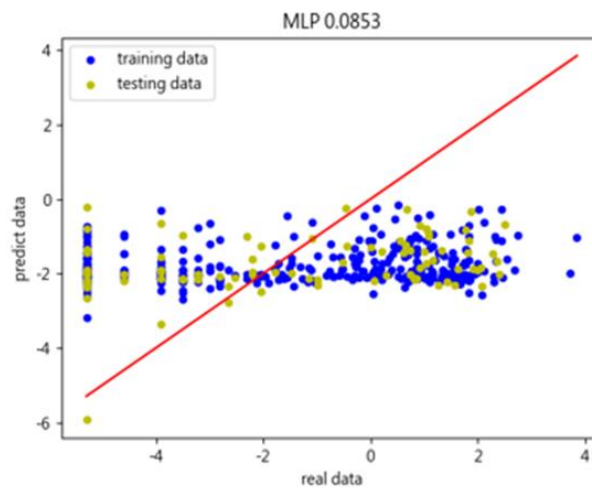


圖 27 LR 模型預測結果

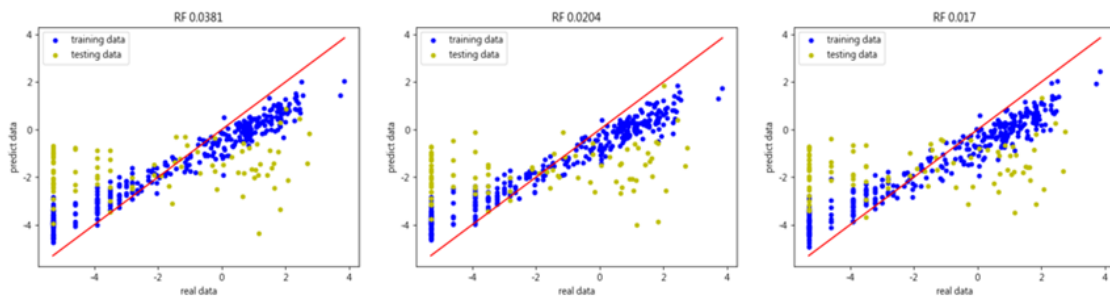


圖 28 RF 採用不同參數的迴歸模型預測結果，由左到右，分別是樹木數量 100、75、及 60 之結果

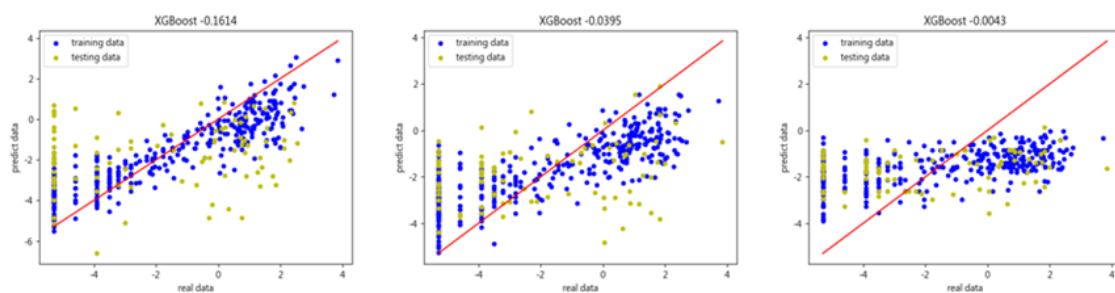


圖 29 由左到右分別是 XGBoost 採用樹木數量 100、50、及 10 的實驗結果



三、結果與討論

表 7 取對數之硝酸鹽氮濃度預測迴歸模型評估結果

	訓練資料		測試資料	
	RMSE	R ²	RMSE	R ²
LR	2.2149(0.0003)	0.0831(0.0002)	2.3072(0.0044)	-0.0124(0.0039)
RF	0.8878(0.0001)	0.8527(<0.001)	2.4178(0.0086)	-0.1132(0.0094)
XGBoost	1.4922(0.0005)	0.5837(0.0002)	2.5048(0.0091)	-0.1939(0.0088)
ANN	2.2325(0.0003)	0.0685(0.0002)	2.3217(0.0048)	-0.0256(0.0050)

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。

表 8 與表 9 分別展示了取對數之硝酸鹽氮濃度不同方法建構的迴歸模型，於訓練及測試資料集中以分類指標評估的混淆矩陣，儘管 ANN 於前述的迴歸模型表現不佳，但分類結果顯示較佳。由於硝酸鹽氮濃度經對數轉換，數值範圍較小，有利於 ANN。圖 30 展示訓練與測試資料集五次交叉驗證的準確率盒圖，結果顯示，取對數之硝酸鹽氮濃度模型於測試資料集的表現明顯較未取對數之硝酸鹽氮濃度模型佳，準確率最高可達到 0.75，人工神經網路也達到 0.73 之準確度。

表 8 不同模型於訓練資料集的混淆矩陣

		方法	真實硝酸鹽氮濃度值 (y)		
			≤1.37	1.37<y<10	≥10
預測硝酸鹽氮 濃度 ($\hat{y}$)	≤1.37	LR	239	1	0
	1.37< $\hat{y}$ <10		100	1	0
	≥10		12	0	0
	≤1.37	RF	232	1	0
	1.37< $\hat{y}$ <10		58	51	0
	≥10		1	11	0
	≤1.37	XGBoost	229	1	0
	1.37< $\hat{y}$ <10		85	26	0
	≥10		5	8	0
	≤1.37	ANN	240	0	0
	1.37< $\hat{y}$ <10		101	0	0
	≥10		12	0	0

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

表 9 不同模型於測試資料集的混淆矩陣

		方法	真實硝酸鹽氮濃度值 (y)		
			≤ 1.37	$1.37 < y < 10$	≥ 10
預測硝酸鹽氮 濃度 ($\hat{y}$)	≤ 1.37	LR	54	1	0
	$1.37 < \hat{y} < 10$		29	2	0
	≥ 10		3	0	0
	≤ 1.37	RF	62	0	0
	$1.37 < \hat{y} < 10$		22	1	0
	≥ 10		3	0	0
	≤ 1.37	XGBoost	59	6	0
	$1.37 < \hat{y} < 10$		17	4	0
	≥ 10		1	1	0
	≤ 1.37	ANN	55	0	0
	$1.37 < \hat{y} < 10$		31	0	0
	≥ 10		3	0	0

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。

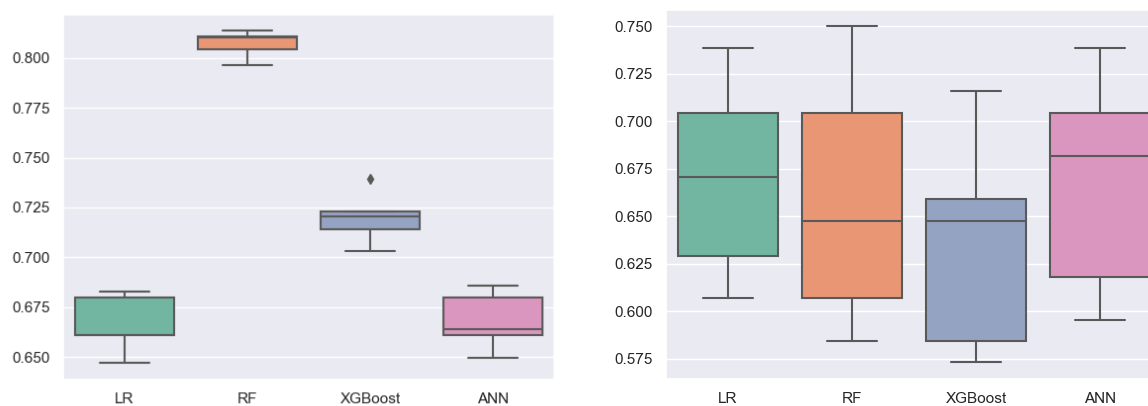


圖 30 取對數之硝酸鹽氮濃度不同模型於訓練（左）與測試（右）資料集的準確率



三、結果與討論

本專案加以考量納入其他部會資料作為預測變量，圖 31 為未採用過採樣後 ANN、LR、RF、以及 XGBoost 建構硝酸鹽氮濃度預測模型的實驗結果，圖 32 為採用過採樣後 ANN、LR、RF、以及 XGBoost 建構硝酸鹽氮濃度預測模型的實驗結果。其中每個演算法與參數皆是採用 15 折交叉驗證。圖 33、圖 34 為四個模型在未採用過採樣和有採用過採樣時，採用迴歸模型評估指標的測試資料和訓練資料的盒狀圖結果，指標於訓練及測試資料之差異。可以看到兩者差異不大。

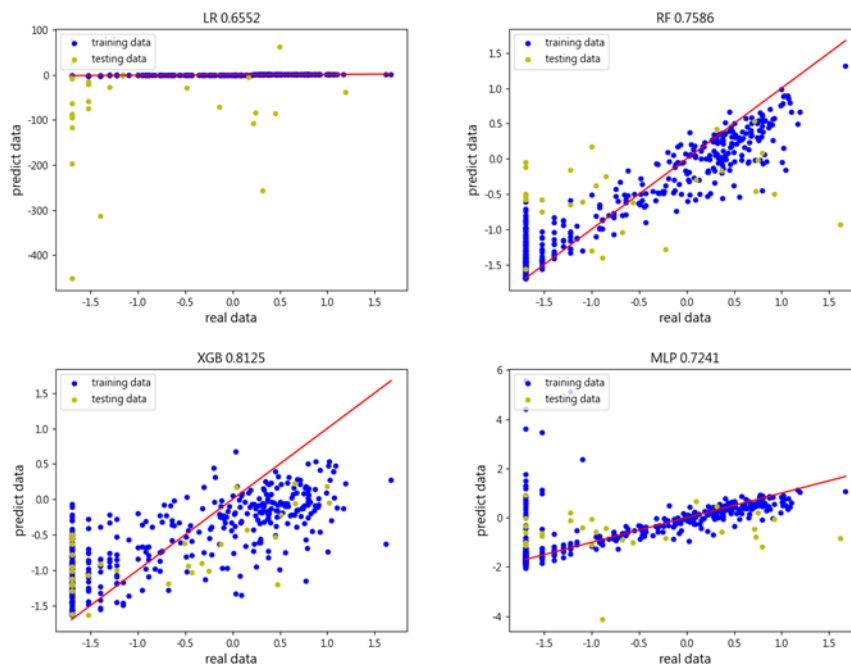


圖 31 上半部兩張圖由左到右分別是 LR 模型預測結果，RF 採樹木數量 20 之架構；下半部兩張圖由左到右分別是 XGBoost 採樹木數量 20 深度為 2 和神經元 (256, 128,64,32)，且激活函數為線性整流函數的 ANN 之結果



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

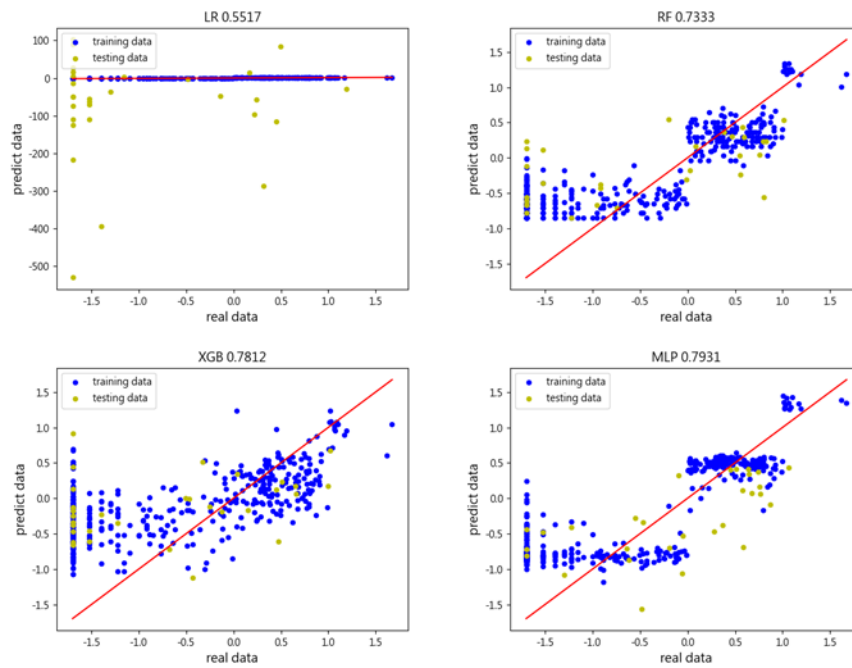


圖 32 上半部兩張圖由左到右分別是 LR 模型預測結果，RF 採樹木數量 20 之架構；下半部兩張圖由左到右分別是 XGBoost 採樹木數量 20 深度為 2 和神經元 (256,128,64,32)，且激活函數為線性整流函數的 ANN 之結果

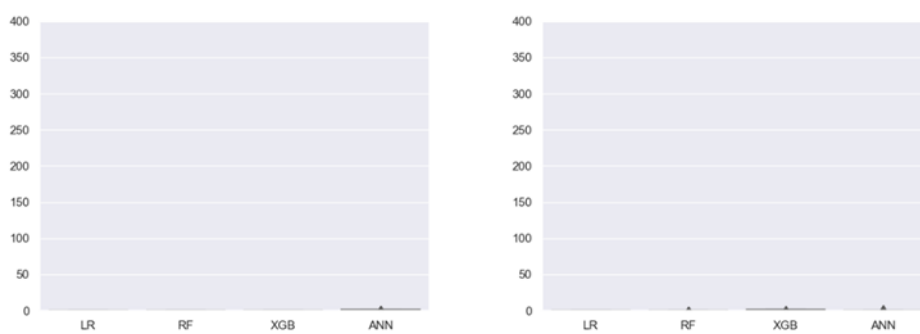


圖 33 訓練資料中未採用過採樣（左）與採用過採樣（右）資料集的 RMSE



三、結果與討論

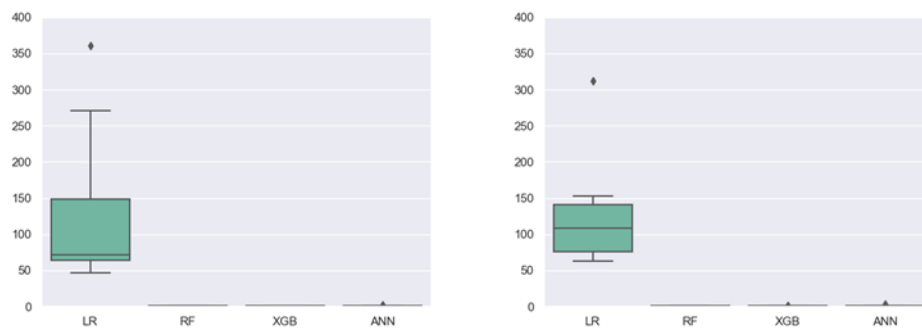


圖 34 測試資料中未採用過採樣（左）與採用過採樣（右）資料集的 RMSE

由於硝酸鹽氮濃度分布之特性，本專案比較有無採用過採樣方法，首先，表 10 為 4 個模型採用分類模型評估指標的結果，指標於訓練及測試資料之差異。表 11、表 12 分別展示了不同方法建構的迴歸模型，於訓練及測試資料集中以分類指標評估的混淆矩陣，由於高數值資料較少，因此模型容易低估水井硝酸鹽數值。圖 35 展示訓練與測試資料集十五次交叉驗證的準確率盒圖，此結果顯示未採用過採樣之硝酸鹽氮濃度預測模型於測試集的準確率明顯較訓練資料集差，但相較未加入其他部會資料的結果，準確率有明顯提升。若僅採用土地利用和水文地質組成作為特徵進行硝酸鹽預測，因特徵不足，導致 ANN 效果不彰，因此，本研究採用更多其他部會資料，以提升模型之表現。

表 10 取對數之硝酸鹽氮濃度預測迴歸模型評估結果

	訓練資料	測試資料
	Accuracy	Accuracy
LR	0.9987(0.0)	0.3592(0.0266)
RF	0.8756(0.0001)	0.629(0.0059)
XGBoost	0.7324(0.0002)	0.642(0.0008)
ANN	0.8303(0.0085)	0.6042(0.0041)

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

表 11 不同模型於訓練資料集的混淆矩陣

		方法	真實硝酸鹽氮濃度值 (y)		
			≤ 1.37	$1.37 < y < 10$	≥ 10
預測硝酸鹽氮 濃度 ($\hat{y}$)	≤ 1	LR	252	0	0
	$1 < \hat{y} < 10$		0	144	0
	≥ 10		0	0	14
	≤ 1	RF	250	42	1
	$1 < \hat{y} < 10$		2	102	12
	≥ 10		0	0	1
	≤ 1	XGBoost	254	90	3
	$1 < \hat{y} < 10$		4	50	11
	≥ 10		0	0	0
	≤ 1	ANN	242	42	2
	$1 < \hat{y} < 10$		10	95	9
	≥ 10		0	7	3

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。

表 12 不同模型於測試資料集的混淆矩陣

		方法	真實硝酸鹽氮濃度值 (y)		
			≤ 1.37	$1.37 < y < 10$	≥ 10
預測硝酸鹽氮 濃度 ($\hat{y}$)	≤ 1	LR	17	6	1
	$1 < \hat{y} < 10$		0	0	0
	≥ 10		6	2	0
	≤ 1	RF	22	8	0
	$1 < \hat{y} < 10$		1	0	1
	≥ 10		0	0	0
	≤ 1	XGBoost	13	11	1
	$1 < \hat{y} < 10$		4	1	0
	≥ 10		0	0	0
	≤ 1	ANN	19	7	0
	$1 < \hat{y} < 10$		4	1	1
	≥ 10		0	0	0

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。



三、結果與討論

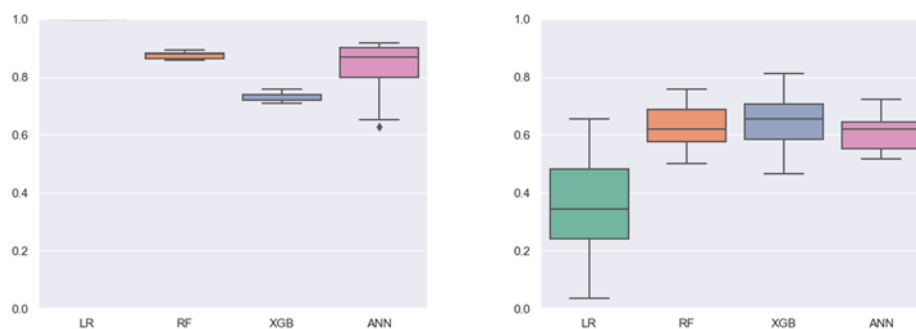


圖 35 不同模型未採用過採樣時於訓練（左）與測試（右）資料集的準確率

表 13 為 4 個模型採用分類模型評估指標的結果，指標於訓練及測試資料之差異。圖 36 分別展示了不同方法建構的迴歸模型，於訓練及測試資料集中以分類指標評估的準確率盒圖。此結果顯示未使用過採樣跟使用過採樣的準確率並未有太大差別。表 14、表 15 展示訓練與測試資料集十五次交叉驗證的混淆矩陣，可以看出雖然有無使用過採樣的準確率差不多，但相較於未採樣過採樣的資料，過採樣過後的結果使中濃度（ $1 < < 10$ ）資料正確辨認類別的機率提高。

表 13 取對數之硝酸鹽氮濃度預測迴歸模型評估結果

	訓練資料	測試資料
	Accuracy	Accuracy
LR	1.0(0.0)	0.3432(0.024)
RF	0.9871(0.0)	0.6281(0.0063)
XGBoost	0.7736(0.0003)	0.6257(0.0048)
ANN	0.9812(0.0002)	0.6135(0.0055)

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

表 14 不同模型於訓練資料集的混淆矩陣

		方法	真實硝酸鹽氮濃度值 (y)		
			≤ 1.37	$1.37 < y < 10$	≥ 10
預測硝酸鹽氮 濃度 ($\hat{y}$)	≤ 1	LR	252	0	0
	$1 < \hat{y} < 10$		0	144	0
	≥ 10		0	0	14
	≤ 1	RF	251	2	0
	$1 < \hat{y} < 10$		1	142	0
	≥ 10		0	0	14
	≤ 1	XGBoost	213	39	0
	$1 < \hat{y} < 10$		39	104	7
	≥ 10		0	1	7
	≤ 1	ANN	251	6	0
	$1 < \hat{y} < 10$		1	138	0
	≥ 10		0	0	14

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。

表 15 不同模型於測試資料集的混淆矩陣

		方法	真實硝酸鹽氮濃度值 (y)		
			≤ 1.37	$1.37 < y < 10$	≥ 10
預測硝酸鹽氮 濃度 ($\hat{y}$)	≤ 1	LR	17	6	1
	$1 < \hat{y} < 10$		0	0	0
	≥ 10		6	2	0
	≤ 1	RF	20	5	0
	$1 < \hat{y} < 10$		3	3	1
	≥ 10		0	0	0
	≤ 1	XGBoost	19	2	0
	$1 < \hat{y} < 10$		4	6	1
	≥ 10		0	0	0
	≤ 1	ANN	16	4	1
	$1 < \hat{y} < 10$		7	3	0
	≥ 10		0	1	0

註：LR：線性迴歸；RF：隨機森林；XGBoost：極限梯度提升；ANN：人工神經網路。



三、結果與討論

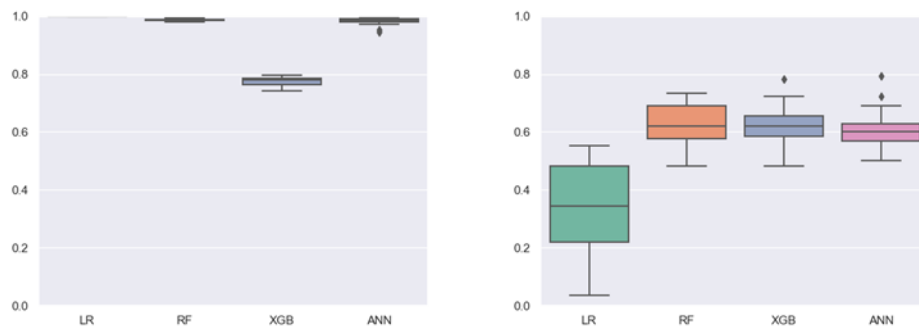


圖 36 不同模型採用過採樣時於訓練（左）與測試（右）資料集的準確率

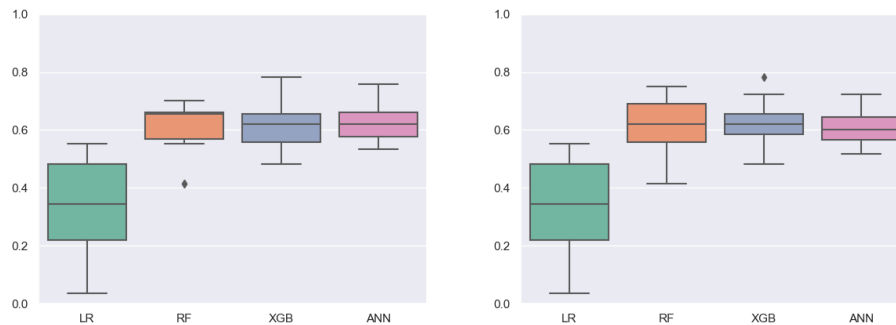


圖 37 採用底數為 1.05 之對數（左）與採用自然對數（右）在各模型上測試集的差異

圖 37、圖 38 為硝酸鹽濃度在採用不同底數之對數後，在各模型上的測試集訓練準確率，結果顯示，若採用底數為 1.05 的結果和採用底數為自然對數或者採用底數為 10 之對數結果並無太大差異。

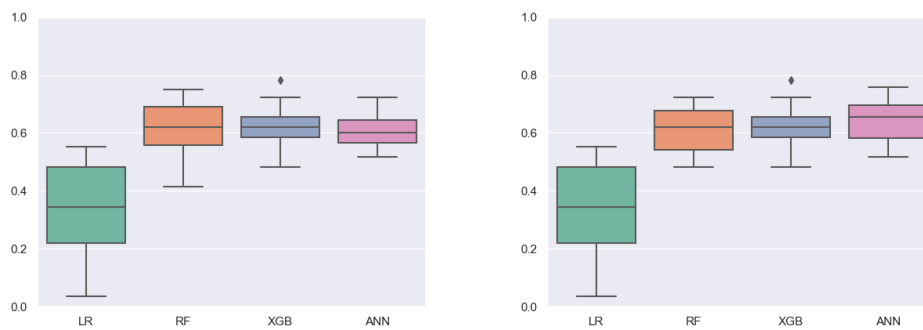


圖 38 以自然對數（左）與採用底數為 10 之對數轉換（右）在各模型上測試集的差異



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

藉由 15 交叉驗證調校模型參數後，根據最佳實驗結果，利用所有資料訓練模型，並預測全台之硝酸鹽氮濃度之類別，預測結果如圖 38，左圖與右圖分別為 ANN 預測結果與真實的硝酸鹽氮濃度，結果顯示有相當高之一致性。

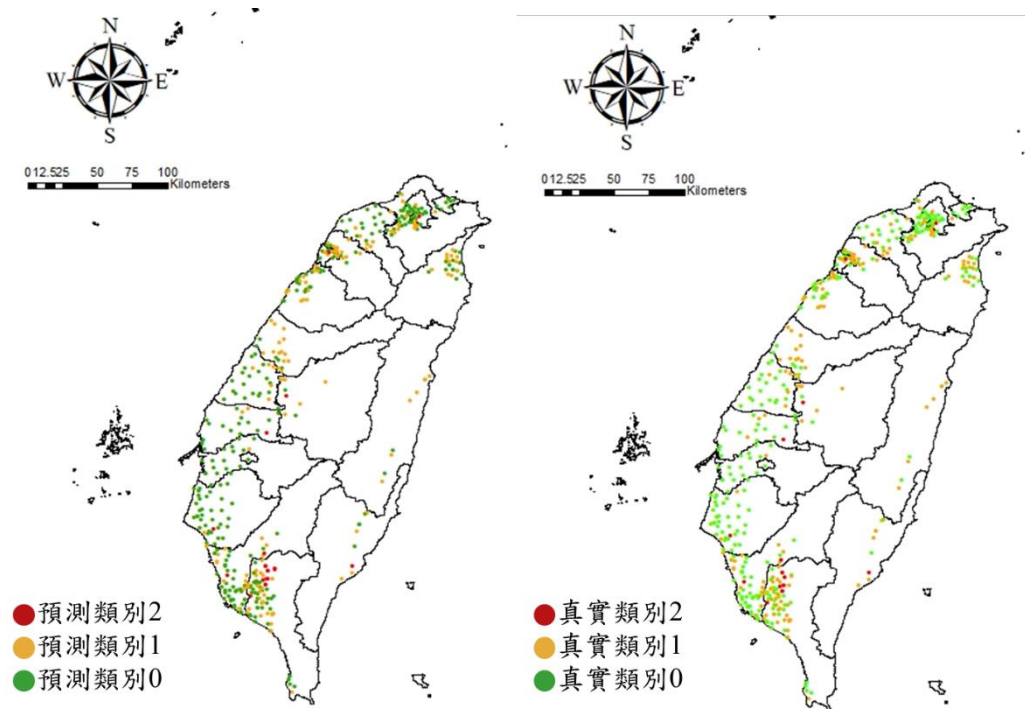


圖 39 ANN 模型預測類別結果（左）與實際硝酸鹽氮濃度（右）

除了前述模型之建置，本研究進一步採用 Shapley Value 分析不同預測變量對硝酸鹽氮濃度之影響。表 16 為計算模型在 15 次交叉驗證後，前三名的特徵，其中果樹利用土地(%)在 15 次模型內皆是第一名，顯示在臺灣土地的硝酸鹽濃度與果樹利用有重要的相關聯性。圖 38 為單次模型 Shapley Value 之結果，可以從圖上看出若是果樹利用土地(%)越高，則對於硝酸鹽濃度的預測會是正相關的。此外，進一步比較 LR 中果樹利用土地(%)、水文地質礫石組成(%)、水文地質粉土組成(%)的迴歸係數，15 次交叉驗證之結果顯示，果樹利用土地的平均迴歸係數為 3，與硝酸鹽濃度為正相關關係，水文地質粉土組成的平均迴歸係數為 3，與硝酸鹽濃度為負相關，此兩項都與 Shapley Value 結果一致。惟水文地質之礫石組成的平均迴歸係數為 3，顯示與硝酸鹽濃度為負相關，此結果與 Shapley Value 的結果稍微不同。



三、結果與討論

表 16 使用 Shapley Value 分析出來重要預測變量前三名

排名	重要預測變量
1	果樹利用土地 (%)
2	水文地質粉土組成 (%)
3	水文地質礫石組成 (%)

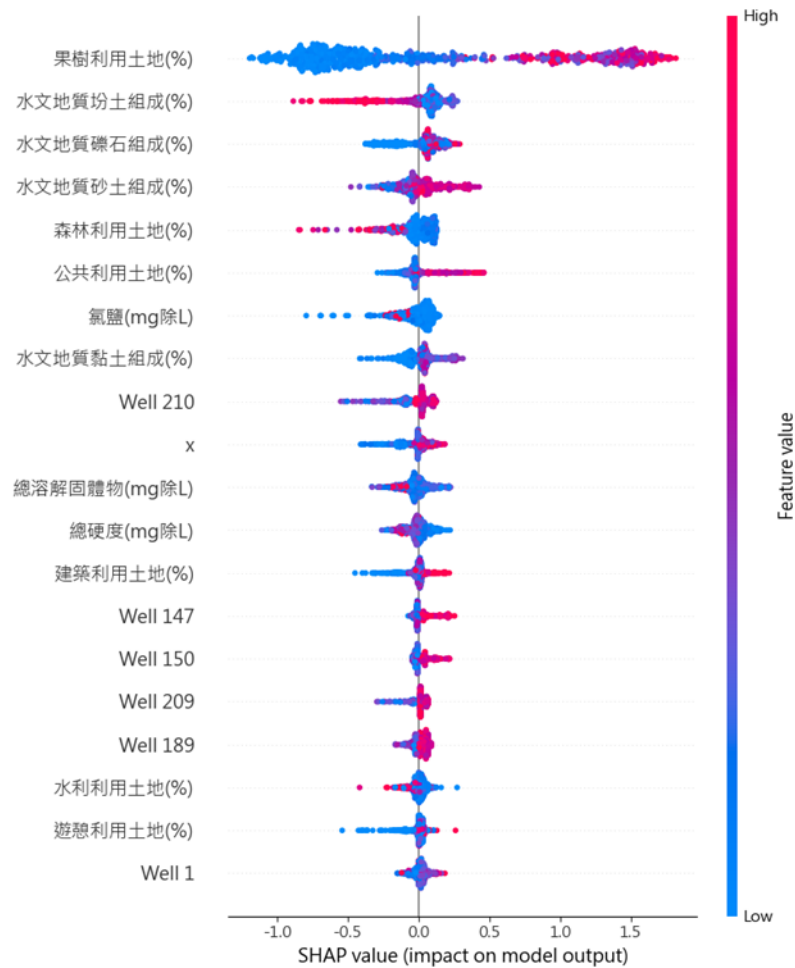


圖 40 單次模型之 Shapely Value 結果



四、結論與建議

本專案利用利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型。考慮的政府跨部會資料包括地下水水質觀測井資訊、土地利用、水文地質組成，以及藉由計算方法得到的預測變量，共計有 27 個預測變量。採用 4 種人工智慧模型，包括線性迴歸(LR)、隨機森林(RF)、極限梯度提升(XGBoost)與人工神經網路(ANN)。計畫中分分別建構有無經對數轉換之硝酸鹽氮濃度預測模型。此外，由於硝酸鹽氮濃度資料分布的特性，本計畫另外加以採用了過採用技術，以解決資料不平衡之問題，人工智慧模式的訓練與測試採用 15 折交叉驗證。

研究結果顯示在模型訓練部份，LR 有最高的準確率，再來為 RF 與 ANN，XGBoost 則是最差。然而在測試資料，RF 有最高的準確率，然後依序為為 XGBoost 與 ANN，LR 則有最低的準確率。在採用過採用技術時，LR 與 RF 在過採用技術後對模型準確率的提昇有所幫助，而 XGBoost 與 ANN 則沒有，顯示在高硝酸鹽濃度資料不足下，資料之擴增對於不同人工智慧方法的模型效能提升仍有所助益。進一步針對模式準確率針對不同硝酸鹽濃度分級進行分析，結果顯示，模型能準確預測較低硝酸鹽氮濃度分級，硝酸鹽氮濃度分級的預測準確。

其主要原因為低硝酸鹽氮濃度分級資料的數量較多，同時，此次計畫自各部會整合之數據位置並不一致，其中，導致空間差異為氣象監測站數據，換言之，缺乏在該水質監測井位置之數據，僅能以最接近之數據取代之，間接影響模型精確度。本專案另外也建立了硝酸鹽氮污染地圖，政府可根據此硝酸鹽氮污染地圖排除較低濃度分級之區域，而於高濃度分進一步優先設置監測井進行地下水水質監測，達到有效擷節監測資源與費用，或教育教育果農減少化肥施用及推廣有機種植，提高地下水資源永續管理工作的效能。



五、參考文獻

- Heißerer, T., Haslauer, C.P., Bárdossy, A., 2016. Including land use information for the spatial estimation of groundwater quality parameters – 1. Local estimation based on neighbourhood composition. *J. Hydrol.*, 535, 688-698.
- Liang, C. P., Lin, T.C., Suk, H., Wang, C.H., Liu, C.W., Chang, T.W., Chen, J.S., 2022. Comprehensively assessing impact of land use and hydrogeological properties on groundwater quality in Taiwan using factor analysis, *Sci. Total Environ.*, 851, Part 1, 158-135.
- Rodriguez-Galiano, V., Paula Mendes, M., Garcia-Soldado, M. J., Chica-Olmo, M., Ribeiro, L., 2014, Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain), *Sci. Total Environ.*, 476-477, 189-206. <https://doi.org/10.1016/j.scitotenv.2014.01.001>
- Ransom, K.M., Nolan, B.T., Traum, J.A., Faunt, C.C., Bell, Gronberg, J.A.M., Wheeler, D.C., Zamora, C., Jurgens, R.B., Schwarz, G.E., Belitz, Eberts, S.M., Harter, K.T., 2017, A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA, *Sci. Total Environ.*, 601–602, 1160-1172. <https://doi.org/10.1016/j.scitotenv.2017.05.192>
- Weng, Hsu-Huei, Tsai, Shang-Shyue, Wu, Trong-Neng, Sung, Fung-Chang, Yang, Chun-Yuh, 2011. Nitrates in drinking water and the risk of death from childhood brain tumors in Taiwan. *J. Toxicol. Environ Health A.* 74 (12), 769-778.



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

研究進度及預期完成之工作項目（甘特圖）

年月 工作項目	111/ 9	111/ 10	111/ 11	111/ 12	112/ 1	112/ 2	112/ 3	112/ 4	112/ 5	112/ 6	112/ 7	112/ 8	備 註
跨部會多元資料收集與 數據資料庫建立													
以 ANN 建立 AI 硝酸鹽 氮污染潛勢預測模型						※							
篩選主要影響預測變量 優化 AI 硝酸鹽氮污染潛 勢預測模型													
建立污染潛勢地圖與研 提管制建議													
撰寫期末報告												※	
工作進度估計百分比 (累 積 數)	10 %	20 %	30 %	40 %	45 %	50 %	60 %	70 %	80 %	90 %	95 %	100 %	
預定查核點	期中	1.彙整跨部會多元資料及建立資料庫。 2.以 ANN 建立 AI 硝酸鹽氮污染潛勢預測模型。											
	期末	1.篩選主要影響變數與建立污染潛勢地圖。 2.撰寫期末報告。											
說明： 1、工作項目請視專案性質及需要自行訂定。預定進度以粗線表示其起迄日期。 2、「工作進度百分比」欄係為配合管考作業所需，累積百分比請視工作性質就以下因素擇一估計訂定：(1)工作天數，(2)經費之分配，(3)工作量之比重，(4)擬達成目標之具體數字。 3、「預定查核點」，請在條形圖上標明※符號，並在「預定查核點」欄具體註明關鍵性工作要項。 4、以 12 個月作規劃，其中期中報告書提送要件需達計畫執行進度 50%以上，期末報告書需於計畫結束前 1.5 個月提送。													



土壤及地下水污染整治基金補助研究與模場試驗專案

附錄一 111 年度專案成果績效自評表

填表日期：112 年 08 月 28 日

一、專案基本資料

專案性質	☐ 實驗性質 ☒ 非實驗性質	專案類別	☐ 先導型 ☒ 研究型 ☐ 模場型
研究主題	☐ 整治 ☒ 調查 ☐ 其他		
執行機構	國立中央大學/應用地質研究所	專案主持人	陳瑞昇
專案名稱	利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型		
專案執行期程	☐ 申請階段 ☐ 期中 ☒ 期末		

二、成果績效自評

「計畫總預估數」應與計畫審查核定值相符，請執行單位依實際達成之量化成果填寫於欄位中。

(一) 學術面

項目		目標達成程度	申請 預估數	期中 達成數	期末 達成數	結案後半年 達成率	備註 (說明未達成 原因或學術產 出發表名稱)
A 學術 產出 及 活動	1.國內投稿 (篇數)	(1)論文	0	0	0		
		(2)研討會論文	1	0	0		準備投稿
	2.國外投稿 (篇數)	(1)期刊論文	1	0	0		準備投稿
		(2)國際研討會 論文發表	1	0	1		投稿 2023 AGU Fall Meeting
	3.報告 (篇數)	(1)技術報告	0	0	0		
		(2)研究報告	1	1	1		
	4.專著 (本數)						
	5.辦理學術 會議(場數)	(1)研討/說明會	0	0	0		
		(2)成果發表會	0	0	0		
		(3)論壇	0	0	0		
B 人才 培育	6.研發改良 技術(項數)	(1)已開發技術	0	0	0		
		(2)技術平台	0	0	0		
	7.研發人員 (人數)	(1)碩士	2	2	2		
		(2)博士	0	0	0		
		(1)跨領域團隊	1	1	1		
		(2)跨機構團隊	1	1	1		
	8.研究團隊 (個數)	(3)形成研究中 心	0				
		(4)形成實驗室	0				



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

目標達成程度			申請 預估數	期中 達成數	期末 達成數	結案後半年 達成率	備註 (說明未達成 原因或學術產 出發表名稱)
項目							
9.其他指標 (請自行命名)		(請自填)					



(二) 產業面

項目\目標達成程度				申請 預估數	期中 達成數	期末 達成數	結案後半年 達成率	備註 (說明未達成原因或專利、技術轉移相關詳細資料)
A 智慧 財產權	1.專利 (件數)	已 核 准	發明	0	0	0		
			新型/設計	0	0	0		
			合計	0	0	0		
		申 請 中	發明	0	0	0		
			新型/設計	0	0	0		
			合計	0	0	0		
B 研發 技術轉 移	2.先期技術 成果移轉	件數		0	0	0		
		授權金(仟元)		0	0	0		
		衍 生 利 益 金 (仟元)		0	0	0		
	3.技術移轉 (專利)	件數		0	0	0		
		授權金(仟元)		0	0	0		
		衍生利益金 (仟元)		0	0	0		
	4.技術移轉 (應用技術)	件數		0	0	0		
		授權金(仟元)		0	0	0		
		衍 生 利 益 金 (仟元)		0	0	0		
	5.可移轉 產業技術	(1)技術(件數)		0	0	0		
		(2)品種/系 (件數)		0	0	0		
C 產學研 合作	6.促成合作 研究	件數		0	0	0		
		金額(仟元)		0	0	0		
	7.促成投資	件數		0	0	0		
		投資金額 (仟元)		0	0	0		
	8.促成取得 業界科專	件數		0	0	0		
		業界投資金額 (仟元)		0	0	0		
9.其他指標 (請自行命名)			(請自填)	0	0	0		



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

(三) 政策面

目標達成程度			申請預估數	期中達成數	期末達成數	結案後半年達成率	備註 (說明未達成原因或其他詳細資料)
項目							
A 服務 便民	1.技術服務	次數	0	0	0		
		收入(仟元)	0	0	0		
	2.諮詢服務	次數	0	0	0		
		收入(仟元)	0	0	0		
B 政策 推動	3.協助政府推動 (件數)	(1)政策	1	0	0		
		(2)法規	0	0	0		
C 技術 效益	4.整治技術提升(%)		0	0	0		
	5.整治成本降低(%)		0	0	0		
	6.提升能源效率(%)		20%	20%	20%		
7.其他指標 (請自行命名)		(請自填)					

三、請依前述學術面、產業面、政策面等預期量化成果，具體敘明研究成果對本署政策推動之助益。(200字為限)

學術面(如國內外研討會、期刊投稿件數、或人才培育碩博士生說明)
1. 投稿 SCI 頂尖期刊 1 篇、參與國際研討會發表論文 1 篇。 2. 培育 2 名碩士生。
產業面(如合作研發產業、申請專利、洽談技術移轉廠商件數說明)
無。
政策面(如整治費用降低、特定污染物整治效益提升、或提供政府作為監測/管制標準、污染址管理等政策及法規研訂之參考)
1. 可繪製地下水硝酸鹽氮污染潛勢地圖，在地下水硝酸鹽氮污染高潛勢地區，優先設置監測井進行地下水水質採樣與監測，達到撙節費用，妥善運用監測資源。 2. 篩選顯著影響地下水硝酸鹽氮濃度的主要預測變量，用以擬定地下水硝酸鹽氮污染改善的相關策略。



四、主要研究人力

請依照「專案主持人」、「協同主持人」、「專任人員」、「兼任人員」及「臨時工」等類別之順序分別填寫

姓名	服務機構/系所	職稱	在本研究專案內擔任之具體工作性質、項目及範圍	執行期間
陳瑞昇	國立中央大學 應用地質研究所	教授	策畫計畫流程、結果研判與撰寫期中/期末報告。	111/09/01～ 112/08/31
洪炯宗	國立中央大學 資訊工程學系	教授	指導 ANN 模式發展。	111/09/01～ 112/08/31
梁菁萍	輔英科技大學 護理系	教授	協助資料收集、結果判釋與協助研提管制建議。	111/09/01～ 112/08/31
李柏峯	國立中央大學 應用地質研究所	碩士生	協助 ANN 模式建立與 ANN 模式優化。	111/12/01～ 112/01/31
黃昶輔	國立中央大學 應用地質研究所	碩士生	繪製污染潛勢地理視覺化地圖。	111/12/01～ 112/01/31
林禹昇	國立中央大學 應用地質研究所	碩士生	資料收集、處理資料補遺。	111/12/01～ 112/03/31
鍾佳儒	國立中央大學 資訊工程學系	博士後研究員	ANN 模式建立與 ANN 模式優化。	111/09/01～ 112/08/31

五、產業界資源投入表

專案執行若有與產業界合作，請執行單位依實際量化成果填寫於欄位中。

投入資源類別	數量	說明
1.自籌款	元	
2.人力	人月	
3.設備		
4.其他資源 (請自行增列)	(請自填)	



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

附錄二 審查意見回復對照表

行政院環境保護署「土壤及地下水污染整治基金補助研究與模場試驗專案」

☐構想書 ☐申請計畫書 ☐期中報告
☐修正計畫書 ☒期末報告

審查意見回覆對照表

計畫年度	111 年度	計畫類型	☐ 先導型 ☒ 研究型 ☐ 模場型
計畫類別	☒ 調查 ☐ 整治 ☐ 其他	主持人：陳瑞昇	NO：SC7
計畫名稱	利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型		
審查意見		執行單位回覆	
委員一			
1. 本計畫研究目的為利用 AI 結合政府跨部會多元資料建立全國硝酸鹽氮污染潛勢預測模型，根據所建立的 AI 地下水硝酸鹽氮污染潛勢預測模型劃分地下水硝酸鹽氮高污染潛勢地圖，提供政府各單位作為採取預警與保護措施的參考依據，尚具體說明執行之研究流程及方法。 2. 本計畫研究內容與計畫目的尚稱符合。本研究專案除使用 ANN 外，也考慮其它 3 種人工智慧模型，包括 MLR、RF 與 XGBoost 分別觀察硝酸鹽濃度取對數前與取對數後之差異，包括 4 種模型於未加入各部會資料和計算方法取得的預測變量時硝酸鹽氮濃度未取對數、取對數設定及 4 種模型於加入各部會資料和計算方法取得的預測變量時硝酸鹽氮濃度取對數之設定，然其中 LR 模型皆僅說明利用 python 的 sklearn 提供的 LinearRegression 方法建構硝酸鹽氮濃度預測模型，建議強化 LR 模型結果說明。 3. 本計畫無執行進度落後，執行進度與預定進度查核表與標準達成情形相符合，達結案標準。 4. 本研究內容已提出結論，宜強化未來地下水硝酸鹽氮污染管制方向建議。 5. 成果績效自評學術面產出僅達成 1 篇研究報告，建議強化國內外投稿論文及研討會發表文章。		1. 謝謝委員肯定。 2. LR 模型的結果相對於其他 3 方法不理想，因此未特別說明。 3. 謝謝委員肯定。 4. 已強化未來地下水硝酸鹽氮污染管制方向建議。 5. 謝謝委員意見，除 1 篇研究報告，目前已投稿全球最大的地球科學領域國際會議(2023 美國地球地理秋季會議)海報論文 1 篇，另外也著手準備投稿 1 篇 SCI 期刊論文。 6. 謝謝委員肯定。	



附錄二

6. 本研究計畫藉由篩選顯著影響地下水硝酸鹽氮濃度的主要預測變量，用以擬定地下水硝酸鹽氮污染改善的相關策略。所繪製地下水硝酸鹽氮污染潛勢地圖，提供政府單位作為採取預警與保護措施之參考依據。	
委員二 1. 本計畫目標擬結合機器學習與跨部會政府多元資料建立地下水硝酸鹽氮污染潛勢預測模式。計畫執行成果有助於未來預測模式的精度。 2. 本計畫擬收集來自政府單位(包括內政部、經濟部水利署、經濟部地調所、環保署、農委會與內政部國土測繪中心等)，建議補充說明所收集預測變量資料的具體內容及其對反應變量預測模式的影響。 3. 請補充說明採用底數為自然對數和底數為10的對數轉換的依據。 4. 建議參數納入季節性雨季的影響。	1. 謝謝委員肯定。 2. 謝謝委員意見，已補充說明所收集預測變量資料的具體內容及其對反應變量預測模式的影響。 3. 採用指數主要是藉由對數轉換減低濃度的變化率，另外根據機率分布圖應該也比較接近對數常態分布，而非一般常態分布，自然對數和底數為10主要是測試兩種常用的底數。 4. 本計畫主要目標為預測濃度的空間分布，未考慮預測時變，有考濃度的時間變化，因此只考慮年降雨做為預測變量，無考慮季節性雨季的影響。
委員三 1. 本計畫以人工智慧方法及跨部會政府資料庫，探討我國高硝酸鹽氮濃度主要預測變量，建構地下水水質硝酸鹽氮高污染潛勢圖，期能協助規劃地下水監測重點區，文獻收集具體詳細。實驗流程方法及結果討論說明清晰。執行進度與預定進度查核相符。執行項目、內容與成果與計劃書相符。 2. 圖38顯示果樹利用土地(%)在15次模型內皆是第一名，顯示在臺灣土地的硝酸鹽濃度與果樹利用有重要的相關聯性。果樹利用包含蔬菜及糧食作物區域嗎？還是僅限果樹利用？不同作物施用之肥料種類頻率與量不同。 3. 農用化學氮肥主要分成銨態氮肥，銨態氮肥包括碳酸氫銨(NH_4HCO_3)、硫酸銨$\{(\text{NH}_4)_2\text{SO}_4\}$、氯化銨($\text{NH}_4\text{Cl}$)等。硝態氮肥包括硝酸鈉($\text{NaNO}_3$)、硝酸鈣$\{\text{Ca}(\text{NO}_3)_2\}$、硝酸銨($\text{NH}_4\text{NO}_3$)等。亦有有機肥(主要為雞豬糞，目前環保署推廣豬沼液)。也就是說有可能農用肥料中之銨(或氮氣)進入地下，再被轉換成硝酸鹽。環保監測井也	1. 謝謝委員肯定。 2. 由於主要果樹用地的由內政部國土測繪中心定義，其定義為係指從事水果及乾果種植、栽培而以收穫其果實為目的之土地。包括李、杏、柿、栗、枇杷、橄欖、木瓜、楊桃、鳳梨、檳榔、葡萄、椰子、柑桔類、番石榴、梅、桃、棗、梨、芒果、胡桃、蘋果、龍眼、香蕉、蓮霧、荔枝、番荔枝、百香果，其並未包含稻作，稻作為另一類土地分類，係指從事稻米栽培之土地，包括水稻、陸稻，目前果樹土地已經是第三級分類，並未進一步考慮不同的果樹作物進行分類。 3. 謝謝委員建議，後續將考慮測試氮氣的影響。 4. 謝謝委員肯定。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

常測出高濃度氮氮，所以將地下水中氮氮納入模式中，是否可能增加污染潛勢預測的準確度？ 4. 此模式若能用於預測目前硝酸鹽尚符合但未來會極有可能超過標準之區域，將有助於地下水中硝酸鹽的管理，例如增設監測井、調整監測頻率或是肥料管理。	
委員四 1. 果樹利用土地學最重要的預測變量，但這並不表示其造成原因就是化肥施用，許多果樹利用土地可能質地較粗或是使用土雞糞所造成，摘要及結論建議修改。 2. 現有之資料來源有空間之差異性，這也是現有模型之瓶頸，請於報告中說明。	1. 謝謝委員指正，已修正摘要與結論，加入可能質地較粗或是使用土雞糞造成。 2. 由於相關資料來自政府單位因此容易有空間不一致性，其中，導致空間差異為氣象監測站數據，換言之，缺乏在該水質監測井位置之數據，僅能以最接近之數據取代之，間接影響模型精確度。已於報告中說明。
委員五 1. 有針對預測變量分析的期中意見，利用 Shapley Value 分析篩選關鍵特徵。而此方式所篩選出的預測變量，是否可嘗試做為初始模型建立之特徵？建議可進一步探討與比較。 2. 工作項目「建立污染潛勢地圖」成果僅呈現於 P.53，建議給予更完整的描述。 3. 工作項目「篩選主要影響預測變量優化 AI 硝酸鹽氮污染潛勢預測模型」之篩選主要影響預測變量的部分，執行尚可再深入。 4. P.35 文章提到使用 27 個預測變量為何？源自哪個資料庫？建議說明之。 5. 模型建立過程應著重跨不同部門之關鍵預測變量組合試誤並探討，較符合計畫之目的。	1. 謝謝委員意見，後續將利用 Shapley Value 分析篩選的關鍵特徵做為初始模型建立的特徵探討與比較預測結果的準確率。 2. 謝謝委員意見，將對建立污染潛勢地圖成果給予更具體說明。 3. 謝謝委員意見，儘管此次計畫並未採用特徵選取進行模型優化，然而，此次計畫發現若僅採用土地利用和水文地質組成作為特徵進行硝酸鹽預測，因特徵不足，導致 ANN 效果不彰，因此，本研究採用更多其他部會資料，以提升模型之表現。未來將可移除部分特徵了解其影響以進一步深化篩選主要影響預測變量優化 AI 硝酸鹽氮污染潛勢預測模型 4. 謝謝委員意見，已於報告中進一步說明 5. 謝謝委員意見，本計畫依序考量土地利用(內政部)、地質材料(中央地質調查所)、雨量(氣象局)與水質濃度進行模型之建置，其中，若僅採用土地利用和水文地質組成作為特徵進行硝酸鹽預測，因特徵不足，導致 ANN 效果不彰，因此，本計畫進一步納入雨量與其他濃度進行建模，提升模型精確度。同時，本計畫利用 Shapley Value 評估特徵對於模型預測之影響，顯示從不同部門取得的特徵有不同程度的貢獻，後續可加入跨不同部門之關鍵預



附錄二

	測變量組合進行測試。
委員六 1. 報告書第 11 頁第二段說明自然過程中的水岩交互作用(water-rock interaction)過程，常常會釋放危害人體的物質進入地下水體，請問會產生哪些危害物質，其原因為何？ 2. 報告中完整敘明，所採行各項實驗結果。	1. 地下水化學污染物可能來自人為活動與天然過程兩種不同方式，其中大部分為人為活動(例如：農業、工業或人類居住區)，另外有些污染物來自地下水與岩石交互作用(water-rock interaction)，導致含水層岩石或沉積物中化學物質釋放進入地下水。雖然有多種天然存在的元素，但這其中只有極少數有可能成為污染物，主要取決於這些元素在岩石和沉積物中的濃度、在某些環境條件下的溶解度以及它們以可溶形式存在且濃度對人類有毒等三種影響因子。地下水中最常見的可溶性離子有 Na^+、Mg^{2+}、Ca^{2+}、Cl^-、HCO_3^- 與 SO_4^{2-}，這些物質會影響水的適飲性，但不會對人體健康構成直接威脅。Fe 和 Mn 可能存在缺氧的地下水環境，許多證據顯示飲用水中的 Mn 會導致人體神經系統問題。在所有水岩交互作用相關污染物中，As 和 F 顯然是對人類健康的最大威脅，全世界有數百萬人受到影響。 2. 謝謝委員肯定。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

行政院環境保護署「土壤及地下水污染整治基金補助研究與模場試驗專案」

☐構想書 ☐申請計畫書 ☒期中報告
☐修正計畫書 ☐期末報告

審查意見回覆對照表

計畫年度	111 年度	計畫類型	☐ 先導型 ☒ 研究型 ☐ 模場型
計畫類別	☒ 調查 ☐ 整治 ☐ 其他		主持人：陳瑞昇 NO：SC7
計畫名稱	利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型		
審查意見		執行單位回覆	
委員一			
7.	本計畫目標為發展結合機器學習與跨部會政府多元資料建立地下水硝酸鹽氮污染潛勢預測模式，其執行之研究流程及方法說明宜再強化說明。	1.	謝謝委員指正，期末報告已強化執行之研究流程及方法之說明。
8.	有關計畫執行之研究進度及預期完成項目甘特圖，計劃書內並未載明目前實際執行進度，建議應補充之。	2.	謝謝委員指正，已修正。
9.	本計畫研究內容與計畫目的尚稱符合。根據研究內容所提出之討論與建議，目前結果顯示，受限於資料多寡，模型訓練易導致過擬合問題，從而導致測試資料集表現不佳，因此，後續專案執行將著重於過擬合問題，以提升預測精準度，宜具體呈現如何優化地下水硝酸鹽氮污染潛勢預測模型內容。	3.	謝謝委員指正，已針對相關過擬合問題，嘗試不同方法改進，包括將 5 折驗證法改為 15 折驗證法、預測變量的正規化與採用了過採用技術減少高濃度分級資料較少的問題，但有部份屬於資料數量的問題所導致的預測表現不佳也詳實於報告中說明。
10.	本研究目的為藉由 AI 結合政府跨部會多元資料(預測變量)建立全國地下水硝酸鹽氮污染潛勢預測模型，根據所建立的 AI 地下水硝酸鹽氮污染潛勢預測模型劃分高污染潛勢地圖，提供政府單位作為採取預警與保護措施的參考依據。	4.	謝謝委員肯定，本專案有繪製高污染潛勢地圖可提供政府根據此硝酸鹽氮污染地圖排除較低濃度分級之區域，而於高濃度分進一步優先設置監測井進行地下水水質監測，達到有效擲節監測資源與費用。



附錄二

委員二 5. 本計畫目標為發展結合機器學習與跨部會政府多元資料建立地下水硝酸鹽氮污染潛勢預測模式。 6. 建議可以增加來自政府跨部會多元資料範圍，涵蓋土地利用、水文、土壤型式、水文地質與硝酸鹽氮輸入等資料。 7. 建議可以比較政府不同部會(例如環保署、水利署)於相近地點的監測資料。以了解監測資料的一致性。 8. 建議補充說明，內政部、經濟部水利署、經濟部地調所、環保署、農委會與內政部國土測繪中心等資料內容的量化數據是否具有相關性，彼此之間如何比對。 9. 圖 15，利用對數轉換硝酸鹽氮濃度的方法依據應加以論述說明。 10. 目前結果顯示，受限於資料多寡，模型訓練易導致過擬合問題，從而導致測試資料集表現不佳。請補充說明未來解決問題的規劃。	1. 謝謝委員肯定。 2. 謝謝委員建議，本計畫已以增加來自政府跨部會多元資料範圍，涵蓋土地利用、水文、土壤型式、水文地質與硝酸鹽氮輸入等資料外，並依這些資料經計算產生更多預測變量作為輸入使用。 3. 謝謝委員指正，已進行相關資料比較，了解其一致性。 4. 謝謝委員指正，許多資料為單一部會所產生資料，有些資料有跨部會間相關性，已於報告中說明。 5. 謝謝委員指正，對數轉換硝酸鹽氮濃度的方法已於期末報告加以論述說明。 6. 謝謝委員指正，已針對相關過擬合問題，嘗試不同方法改進，包括將 5 折驗證法改為 15 折驗證法、預測變量的正規化與採用了過採用技術減少高濃度分級資料較少的問題，但有部份屬於資料數量的問題所導致的預測表現不佳也詳實於報告中說明。
委員三 5. 文獻收集具體詳細。模式架構設計及模擬結果討論 說明清晰。執行進度與預定進度查核相符。	謝謝委員肯定。
委員四 本研究將土地利用之特徵以組成面積之百分比表示，並做為預測變量。此舉將複雜的環境資訊有效引入人工智慧模型進行訓練，有助於未來模型之應用性。以下幾點建議供參： 3. 建議預測變量應更多元，目前模型數入之預測變量僅使用土地利用與水文地質（14 項），	謝謝委員肯定。 1. 謝謝委員建議，將環保署與水利署之地下水水質資料一併考慮進行模型訓練，並利用 Shapley Value 分析篩選出重要預測變量前三名土地利用果樹、水文地質粉土與礫石。 2. 謝謝委員建議，基於資料較少的考量，已將



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

建議將環保署與水利署之地下水水質資料一併考慮進行模型訓練。 4. 建議 K 折交叉驗證與特徵篩選(Recursive Feature Elimination)之結果選定可進一步深入探討與說明。例如：K 值選定是否已經是最佳值、探討多組較佳特徵組合進行模型訓練。 5. 從混淆矩陣圖表發現，不同模型的訓練集及與測試集資料不同，建議於程式碼設定固定之隨機種子 (Random Seed)，讓實驗手法有一致性。 6. 資料總筆數為 443 且輸入之預測變數為 14 項進行模型訓練，容易發生維度災難，建議嘗試主成分分析進行降維改善問題。	期中報告的 5 折驗證改為 15 折驗證，改善模型訓練與測試，另外有利用 Shapley Value 分析篩選出來重要預測變量前三名土地利用果樹、水文地質粉土與礫石。 3. 謝謝委員意見，因為期中報告前採用 5 折驗證法，但資料總數目不是剛好是 5 的倍數，所以 5 個分組的資料數目不會完全相同，又每一機器模型最佳的驗證的訓練組與測試組可能不同，因此自然每個模型的訓練的資料數會不同。 4. 謝謝委員建議，研究團隊依據水利署地下水水質資料，因子分析的結果進行資料數量的減少，並以因子得分的結果進行機器學習訓練，也獲得可接受的結果，本專案主要將環保署與水利署之地下水水質資料一併考慮，並利用相關政府部會資料進行模型訓練，並利用 Shapley Value 分析篩選出重要預測變量前三名土地利用果樹、水文地質粉土與礫石。
委員五 6. 表 8 四種模型之總數量不同，請說明。 7. 過擬合之情形，後續將如何改善？	1. 謝謝委員意見，因為期中報告前採用 5 折驗證法，但資料總數目不是剛好是 5 的倍數，所以 5 個分組的資料數目不會完全相同，又每一機器模型最佳的驗證的訓練組與測試組可能不同，因此自然每個模型的訓練的資料數會不同，表 8 主要是四種模型最佳時的資料數量，因此表 8 四種模型之總數量會有所不同。 2. 謝謝委員指正，已針對相關過擬合問題，嘗試不同方法改進，包括將 5 折驗證法改為 15 折驗證法、預測變量的正規化與採用了過採用技術減少高濃度分級資料較少的問題，但有部份屬於資料數量的問題所導致的預測表現不佳也詳實於報告中說明。



附錄二

委員六

1. 前言 P.8 第一段第三行:何謂距離反比加權法?
2. 報告內容完整及各圖表清楚。
3. 建議將所獲得之資料，提供給各主管機關，以利掌握所需資訊。
4. 台灣常見的地下水污染除了硝酸鹽氮外，是否還有其他常見之污染物？

1. 距離反比加權法又稱為反距離加權(inverse distance weighting, IDW) 為一種在有已知的離散點變數觀測資料的情形下進行未知點的變數插值的確定性算法。賦給未知點的值是用已知點的值的加權平均數計算得出的。權重是以距離反比計算。反距離加權法為常用的方法。
2. 謝謝委員肯定。
3. 謝謝委員建議。
4. 台灣常見的地除了硝酸鹽氮外，另外砷也為地下水常見的之污染物，砷一般認為是地質生成的污染物，台灣的砷污染與其衍生的公共衛生問題在全世界的科學研究都有舉足輕重的角色，計畫主持人與共同主持人梁菁萍過去對於砷所造成的飲用水與養殖魚貝的攝食健康風險累積一定的研究成果發表於頂尖期刊。另外台灣有許多含氯地下水污染場址，計畫執行團隊在開發含氯地下水污染傳輸模式與健康風險評估軟體有世界最先進的成果發表於最頂尖期刊，為解析模式發展出互動式圖形使用者介面的軟體(MUST, Multi-Species Transport Analytical Model, <https://camrdaprod.ncu.edu.tw/products-landing/must/>)，時也經中央大學智財委員會同意進行全球商品化銷售，已技轉美商傑明與業興環境科技公司，MUST也獲環保署111年環境科技論壇暨成果發表會惟一應用及推廣度優勝獎。



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

行政院環境保護署「土壤及地下水污染整治基金補助研究與模場試驗專案」

☒階段成果報告審查意見 ☐期中報告 ☐修正計畫書 ☐期末報告 **審查意見回覆對照表**

計畫年度	111 年度	計畫類型	☐ 先導型 ☒ 研究型 ☐ 模場型
計畫類別	☒ 調查 ☐ 整治 ☐ 其他	主持人：陳瑞昇 NO：SC7	
計畫名稱	利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型		
審查意見		執行單位回覆	
1. 經審閱階段成果報告已依申請計畫書定稿本執行,符合預期執行進度。		謝謝委員肯定。	



行政院環境保護署「土壤及地下水污染整治基金補助研究與模場試驗專案」

☐申請計畫書 ☐期中報告 ☒修正計畫書 ☐期末報告 **審查意見回覆對照表**

計畫年度	111 年度	計畫類型	☐ 先導型 ☒ 研究型 ☐ 模場型
計畫類別	☒ 調查 ☐ 整治 ☐ 其他		主持人：陳瑞昇 NO：SC7
計畫名稱	利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型		
審查意見		執行單位回覆	
委員一 2. 計畫主持人具備豐富水文質研究經驗，協同主持人具人工智慧經驗，具備本計畫執行能力。 3. 本計畫透過豐富地下水質資料，有助於地下水整治成效分析探討。		1. 謝謝委員肯定。 2. 謝謝委員肯定。	
委員二 1. 本計畫結合 AI 與部會資料庫的大數據，開發預測地下水中氮系污染物之污染潛勢模型，研究對於台灣地下水保護，具有參考性。 2. 主持人在地下水模式領域研究多年，累積相當經驗，預期研究成果可以發表學術期刊論文。 3. 為更強化模式的應用性，建議強化報告中 7.8 節研提改善措施部分應用，以產出各種污染源管制情境，以協助相關主管機關控制污染。		1. 謝謝委員肯定。 2. 謝謝委員肯定。 3. 謝謝委員寶貴意見，已於計畫書「7.8 研提地下水硝酸鹽氮污染改善措施」強化模式的應用性。	
委員三 1. 本計畫結合人工智慧方法及跨部會政府資料庫，探討我國高硝酸鹽氮濃度主要預測變量，建構地下水水質硝酸鹽氮高污染潛勢比圖，並規劃地下水污染改善策略，有效提升我國水資源永續管理效能。 2. 有關研提地下水硝酸鹽氮污染改善措施，		1. 謝謝委員肯定。 2. 謝謝委員寶貴意見，已於計畫書「7.8 研提地下水硝酸鹽氮污染改善措施」修正。	



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

建議後續可補充目前國內外相關現地研究技術，如厭氧生物處理及電透析技術，並根據各技術去除效率、產水品質及消耗成本等面向進行綜合評析，有助於強化我國地下水污染改善策略。	
委員四 研究主題符合土壤及地下水污染整治基金補助精神、探討課題新穎、整體研究架構明確方法可行、技術未來應用及推廣性應加強。	謝謝委員肯定與寶貴意見，已加強說明技術未來應用及推廣性於「7.8 研提地下水硝酸鹽氮污染改善措施」。
委員五 執行方法規劃得宜。	謝謝委員肯定。
委員六 1. 本案擬利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型。 2. 計畫構想明確。 3. 未來成果可作為主管機關參考。 4. 建議補充說明如何進行跨部會的資料庫整合，所跨部會的範圍應進行評估資料庫的內容。	1. 謝謝委員肯定。 2. 謝謝委員肯定。 3. 謝謝委員肯定。 4. 謝謝委員寶貴意見，已於計畫書加強如何進行跨部會資料整合與資料的補遺於「7.3 跨部會資料整合與補遺」。
委員七 1. 本計畫對於研提改善地下水硝酸鹽污染之措施或策略，並不構具體，如何達到政策或決策之參考，請再加強補充說明。 2. 請補充說明本計畫 AI 監測或預測分析之執行方法、實施步驟，以及建立模型之驗證方式。 3. 請補充說明當資料缺漏或不足時之因應作為，或對模型建立之影響。	1. 謝謝委員肯定與寶貴意見，已加強說明研提改善地下水硝酸鹽污染之措施或策略於「7.8 研提地下水硝酸鹽氮污染改善措施」。 2. 謝謝委員寶貴意見，已於計畫書說明預測分析之執行方法、實施步驟，以及建立模型之驗證方式。 3. 謝謝委員寶貴意見，已於計畫書說明當資料缺漏或不足時之因應作為於「7.3 跨部會資料整合與補遺」。



附錄二

委員七

p.11-43 耗材電腦周邊、儲存設備建議以年度攤提或租賃編列。

謝謝委員寶貴意見，遵照辦理



利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型

行政院環境保護署「土壤及地下水污染整治基金補助研究與模場試驗專案」

☒構想書 ☐申請計畫書 ☐期中報告 ☐修正計畫書 ☐期末報告 **審查意見回覆對照表**

計畫年度	111 年度	計畫類型	☐ 先導型 ☒ 研究型 ☐ 模場型
計畫類別	☒ 調查 ☐ 整治 ☐ 其他	主持人：陳瑞昇教授 NO：	
計畫名稱	利用人工智慧結合政府跨部會多元資料建立地下水硝酸鹽氮污染潛勢預測模型		
審查意見		執行單位回覆	
委員一 1.研究目標/範疇應再檢核。 2.研究方法、探討變數不確定性甚高，應請精確規劃。		1. 謝謝委員意見，計畫書已檢核研究目標與範疇，並重新調整研究目標與範疇使整體研究成果更有對目前地下水管理工作有更具體貢獻。 2. 謝謝委員意見，計畫書已補充詳細說明研究方法與流程並詳細列出與所使用的探討變數(預測變量)。	
委員二 1.以機器學習方法，採用何種電腦軟體(如:matlab)或 open sources 應說明，建議未來應該公開程式模式。		1.謝謝委員意見，計畫會同時採用 Python 與 Matlab 做為機器學習方法的電腦軟體，但考慮未來公開性，最後應該會以 Python 程式公開提供做為參考依據。	
委員三 1.本計畫擬以跨部會資料為基礎，應用人工智慧分析硝酸鹽汙染潛勢。研究結合大數據、人工智慧，具有創新性。 2.台灣地下水中除硝酸鹽，氮系汙染物氨氮也是重要汙染物，建議可以移併納入分析。		1.謝謝委員肯定。 2.謝謝委員意見，計畫可以同時考慮氨氮。	
委員四 1.有關專案未來性，本研究計畫採用人工智慧方法 ANN 進行模型建構，建議宜針對我國政府跨部會資料面數據不完善疑慮進行補充說明，有助於後續透過大數據模型完善其數據量化值，以提升該研究計畫之未來可行性。		1.謝謝委員意見，計畫書已對不同部會資料的整合以利 ANN 模型建構進行補充說明。	



附錄二

委員五 1.研究主題符合土壤及地下水污染整治基金補助精神、探討課題新穎、整體研究架構明確方法可行、技術未來應用及推廣性應加強。	1.謝謝委員意見，計畫會加強技術未來應用及推廣性說明。計畫將探討顯著影響硝酸鹽氮濃度的主要預測變量並依據這些篩選出的主要影響變量，發展減緩硝酸鹽氮污染的有效策略，這些策略可能包括藉由法律規範控制污染、公眾提高社區意識與改變可能導致硝酸鹽氮污染的人為活動。
委員六 1.輸入資料是否也含時間所產生的變異，若是，其解析度。預測結果與驗證來修正預測模型。 2.水利署水質監測(地下水)是否在同一水層，可先確認。	1.謝謝委員意見，我們參考相關文獻在考慮時間的變異，以地下水質變化相對較為緩慢，且採樣時間常常不一致，計畫在今年會從空間的變異性來先進行驗證，暫時排除時間的變異性。 2.由於硝酸鹽氮較屬於人為污染物，會以第一含水層為優先。
委員七 1.請補充說明本計畫污染(硝酸鹽)潛勢預測模型之驗證方法。 2.對於本計畫蒐集之跨部會多元資料為基礎，所建立之預測模型，實施計畫過程之可能瓶頸為何？未來建立模型之靈敏度及準確定之判斷依據為何？	1. 謝謝委員意見，驗證將採用 K 折交叉驗證，評估指標包含平均絕對誤差(Mean Absolute Error, MAE)、平均平方誤(Mean Square Error, MSE)、均方誤差(Root Mean Square Error, RMSE) 及決定係數 (R²)。 2. 可能瓶頸為資料的空間不一致性，未來建立模型之靈敏度會針對探討變數進行篩選，並參考國際相關經驗建立模型之靈敏度及準確定之判斷依據。