

國立臺灣大學生物資源暨農學院生物環境系統工程學系

碩士論文

Department of Bioenvironmental Systems Engineering

College of Bioresources and Agriculture

National Taiwan University

Master Thesis

應用機器學習建立台灣本土土壤保水曲線轉換方程

式：以中部農業土壤為例

Establish machine learning-based localized pedotransfer
functions of soil water retention curves: a case study of
agricultural soils in central Taiwan.

馮博煜

Bo-Yu Fung

指導教授：許少瑜 博士

Advisor: Shao-Yiu Hsu, Ph.D.

中華民國 112 年 6 月

June, 2023



國立臺灣大學碩士學位論文 口試委員會審定書

應用機器學習建立台灣本土土壤保水曲線轉換方程式：

以中部農業土壤為例

Establish machine learning-based localized pedotransfer
functions of soil water retention curves: a case study of
agricultural soils in central Taiwan.

本論文係 馮博煜 君 (R10622006) 在國立臺灣大學
生物環境系統學研究所完成之碩士學位論文，於民國 112
年 6 月 21 日承下列考試委員審查通過及口試及格，特此
證明

口試委員：

許少瑜

(指導教授)

許健輝

詹國偉

胡明哲

系主任：

余心龍

謝誌

感謝我的指導教授許少瑜老師提供一個優質的學習環境，且每週都抽空與我進行深入討論，使我無論是在研究問題抑或是論文撰寫上都獲得大量幫助，也因此才能順利完成此本論文。也感謝口試委員臺灣大學胡明哲教授、廖國偉教授以及農試所許健輝博士，不吝地提供本論文許多寶貴的回饋。特別感謝鄭克聲教授能在百忙之中抽出時間，指點我關於統計方面的知識。

感謝許少瑜老師研究室的每一位成員對我的幫助。尤其感謝蔡義誌博士的協助，除了在研究進行、撰寫論文方面都不遺餘力的幫忙外，有時甚至花上一整天的時間討論目前所碰到的困難、疑問該如何解決及處理，都給予我研究上極大的支持與鼓勵。感謝研究室的大哥兼夥伴哲佑，無論是一同共進午餐後漫步在大街上的時事哈拉(抨擊)，還是在許老師辦公室進行激烈的學術討論，都使我留下許多珍貴而難忘的回憶；感謝凱翔時常順路送我回家，在路途中交流彼此研究碰到的各種問題，以及分享其他趣聞(梗圖、minecraft)，為煩悶的生活增添幾分樂趣；感謝祥玲協助進行壓力鍋、篩分析、比重計等土壤實驗，給予我研究上莫大的協助；感謝胡傑、昱豪、維歆學長讓我們研究室每次出差都充滿著歡樂的氣氛；感謝群展學長總是可以點出我不曾想過的問題，提出許多簡報以及口試上的建議。

最後感謝我的家人給予我全力的支持與陪伴，使我能夠專心致志地完成碩士學業。再次致上最深的謝意，感謝各位。

中文摘要



本研究以台灣中部農業土壤為範例，以土壤質地(砂、粉、黏含量)、總體密度與有機質含量作為解釋變數(或稱輸入特徵)，應用多變量線性迴歸、隨機森林、類神經網路三種不同架構建立台灣本土土壤保水曲線的轉換方程式(pedotransfer function, PTF)，並進一步分析 PTF 的輸入特徵重要性與不確定性。為了釐清不同輸入特徵對於 PTF 最終預測結果的影響力，本研究透過排列重要性的分析，得知 PTF 在基質勢能(matric potential)大於 -0.1bar 時，以總體密度作為最重要的輸入特徵；隨著基質勢能的減少總體密度重要性下降，粉粒與有機質含量重要性則逐漸上升。而為了評估非線性 PTF 的預測結果(不為常態分佈且具有較高偏度)之不確定性，本研究使用無母數自助法評估 PTF 的誤差 95%信賴區間與 95%預測區間。藉由衡量預測區間之實際涵蓋率，確認土壤含水量的實際涵蓋率皆介於 $95\pm 1\%$ ，再次檢核無母數自助法可以有效評估非線性模型的不確定性。預測區間的建立除了提供 PTF 預測值的可靠度資訊外，此預測區間未來可用於檢驗土壤量測資料有效性的依據。最後，本研究成果也顯示三種不同架構的本土 PTF，針對台灣土壤進行轉換的結果皆優於美國農業部開發之 PTF—Rosetta3，再次確認建立台灣本土 PTF 的必要性。

關鍵字：土壤轉換方程式、土壤保水曲線、排列重要性、不確定性分析、無母數自助法。

ABSTRACT



This study takes agricultural soils in central Taiwan as an example and uses soil texture (sand, silt, clay content), bulk density, and organic matter content as explanatory variables (or input features) to develop pedotransfer function (PTF) for the soil water retention curve using three different frameworks: multiple linear regression, random forest, and artificial neural networks. The study further analyzes the uncertainty and feature importance of PTF. Through permutation importance analysis, the study reveals that bulk density is the most important feature when the matric potential is larger than -0.1 bar in PTF predictions. As the matric potential decreases, the importance of bulk density decreases while the importance of silt and organic matter content gradually increases. To evaluate the uncertainty of the nonlinear PTF predictions, which do not follow a normal distribution and have higher skewness, the study uses the nonparametric bootstrap method to assess the 95% confidence intervals of PTF's error and 95% prediction intervals of PTF. By measuring the coverage probability of the prediction intervals, the study confirms that the coverage probability of soil water content is approximately $95\pm1\%$, validating the effectiveness of the nonparametric bootstrap method in assessing the uncertainty of nonlinear models. In addition to providing reliability information for PTF predictions, the establishment of prediction intervals can also be used as a basis for testing the validity of soil measurement data. Finally, the results of this study demonstrate that the three different frameworks of local PTF outperform the PTF developed by United States Department of Agriculture(USDA), Rosetta3, in converting Taiwanese soils, reaffirming the necessity of developing local PTF for Taiwan.

Keywords: pedotransfer function, soil water retention curve, permutation importance,

uncertainty analysis, nonparametric bootstrap.



目錄



口試委員會審定書	#
謝誌	i
中文摘要	ii
ABSTRACT	iii
目錄	v
圖目錄	vii
表目錄	x
第 1 章 緒論	1
1.1 研究背景及動機	1
1.2 研究目的	2
第 2 章 文獻回顧	3
2.1 土壤保水曲線	3
2.1.1 土壤水總勢能	3
2.1.2 與土壤含水量之關係	3
2.1.3 VG model	5
2.2 土壤轉換方程式(Pedotransfer Function, PTF)	9
2.2.1 點型與連續型 PTF	9
2.2.2 建立 PTF 的工具或方法	11
2.2.3 作為預測因子的土壤性質	14
2.2.4 轉換表現評估指標	17
2.2.5 不確定性分析	19
2.2.6 Rosetta	21
第 3 章 研究方法	23
3.1 研究架構及流程	23
3.1.1 土壤資料	24
3.1.2 土壤數據整理	28
3.1.3 Vw15bar 補值	31
3.2 模型選擇與建置	33

3.2.1	多變量線性迴歸(Multiple Linear Regression, MLR)	33
3.2.2	隨機森林(Random Forest, RF)	34
3.2.3	類神經網路(Artificial Neural Network, ANN)	36
3.3	特徵重要性	38
3.3.1	基尼重要性(Gini importance)	38
3.3.2	排列重要性	40
3.4	不確定性分析	42
3.4.1	誤差期望值	42
3.4.2	預測區間	44
第 4 章	結果與討論	47
4.1	訓練、測試資料集探索與描述	47
4.2	模型訓練	56
4.2.1	多變量線性迴歸	56
4.2.2	隨機森林	59
4.2.3	類神經網路	60
4.2.4	土壤含水量訓練集預測結果	61
4.3	模型預測	62
4.3.1	土壤含水量測試集預測結果	62
4.3.2	參數重要性	62
4.3.3	誤差期望值之信賴區間	66
4.3.4	模型預測區間	67
4.4	與 Rosetta3 之比較	70
4.5	增加額外解釋變數的影響	73
4.6	點型與連續型 PTF 之不確定性比較	76
第 5 章	結論與建議	79
5.1	結論	79
5.2	建議	80
參考文獻		81
附錄		92

圖目錄

圖 2.1 土壤保水曲線與土壤含水量關係示意圖(Matsuoka <i>et al.</i> , 1999).....	5
圖 2.2 土壤含水量與水力傳導係數關係示意圖(Sojka <i>et al.</i> , 2009).....	6
圖 2.3 土壤含水量與土壤保水曲線、飽和水力傳導係數關係圖	8
圖 2.4 運用前饋神經網路建立田間含水量的 PTF 範例(Wösten <i>et al.</i> , 2001).....	13
圖 2.5 運用決策樹建立田間含水量的 PTF 範例(Amorim <i>et al.</i> , 2022)	14
圖 2.6 USDA 土壤三角質地圖(Shirazi & Boersma, 1984).....	15
圖 2.7 總體密度隨土壤含水量改變示意圖(Kim <i>et al.</i> , 2019).....	16
圖 2.8 自助法(Bootstrap)示意圖(Efron & Tibshirani, 1994).....	21
圖 3.1 研究架構圖	23
圖 3.2 研究流程圖	24
圖 3.3 農試所得出土壤保水曲線之實驗儀器	26
圖 3.4 農試所民國 101、102 年所有土壤調查地點及數量	28
圖 3.5 資料前處理流程圖	29
圖 3.6 模型建置與後續分析流程圖	33
圖 3.7 隨機森林示意圖	34
圖 3.8 單一神經元(Neuron)示意圖	36
圖 3.9 類神經網路架構.....	37
圖 3.10 基尼雜質計算示意圖	38
圖 3.11 擾亂原始特徵順序之示意圖	40
圖 3.12 無母數自助法推估預測區間之流程圖	46
圖 4.1 訓練資料集(共 1822 筆)土壤三角質地圖.....	47
圖 4.2 測試資料集(共 792 筆)土壤三角質地圖.....	48
圖 4.3 各解釋變數的盒狀圖	50
圖 4.4 各解釋變數之標準分數盒狀圖	51
圖 4.5 各反應變數之土壤含水量盒狀圖	52
圖 4.6 訓練集與測試集土壤含水量標準差與平均值散佈圖	53
圖 4.7 隨機森林模型在不同基質勢能之解釋變數基尼重要性(訓練集資料).....	59
圖 4.8 類神經網路模型在不同基質勢能之訓練曲線	60

圖 4.9 三種模型在不同基質勢能之反應變數排列重要性(測試集)	63
圖 4.10 MLR 與 RF 在不同基質勢能之反應變數排列重要性(測試集).....	65
圖 4.11 三種模型土壤含水量預測誤差盒狀圖(測試集).....	66
圖 4.12 RF 模型之土壤飽和含水量預測區間(測試集).....	68
圖 4.13 台灣本土 PTF 與 Rosetta3 土壤含水量轉換誤差盒狀圖(測試集).....	70
圖 4.14 台灣本土 PTF 與 Rosetta3 轉換與實際土壤保水曲線弗雷歇距離盒狀圖(測試集).....	72
圖 4.15 RF 新模型之土壤飽和含水量預測區間(測試集).....	74
圖 4.16 土壤含水量與基質勢能關係圖	77
圖 A.1 應用 MLR 轉換台灣土壤含水量結果(訓練集).....	92
圖 A.2 應用 MLR 轉換台灣土壤含水量結果(測試集).....	92
圖 A.3 應用 MLR 轉換台灣土壤含水量的誤差(綠色為訓練集、棕色為測試集)...	93
圖 A.4 應用 MLR 轉換台灣土壤含水量的預測區間與實際涵蓋率(訓練集).....	93
圖 A.5 應用 MLR 轉換台灣土壤含水量的預測區間與實際涵蓋率(測試集).....	94
圖 A.6 應用 RF 轉換台灣土壤含水量結果(訓練集).....	94
圖 A.7 應用 RF 轉換台灣土壤含水量結果(測試集).....	95
圖 A.8 應用 RF 轉換台灣土壤含水量的誤差(綠色為訓練集、棕色為測試集).....	95
圖 A.9 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(訓練集).....	96
圖 A.10 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(測試集).....	96
圖 A.11 應用 ANN 轉換台灣土壤含水量結果(訓練集).....	97
圖 A.12 應用 ANN 轉換台灣土壤含水量結果(測試集).....	97
圖 A.13 應用 ANN 轉換台灣土壤含水量的誤差(綠色為訓練集、棕色為測試集).....	98
圖 A.14 應用 ANN 轉換台灣土壤含水量的預測區間與實際涵蓋率(訓練集).....	98
圖 A.15 應用 ANN 轉換台灣土壤含水量的預測區間與實際涵蓋率(測試集).....	99
圖 A.16 應用 Rosetta(Zhang & Schaap, 2017)轉換台灣土壤含水量結果(訓練集)...	99
圖 A.17 應用 Rosetta(Zhang & Schaap, 2017)轉換台灣土壤含水量結果(測試集).....	100
圖 A.18 應用 RF 轉換台灣土壤含水量結果(VwPF253 為額外解釋變數，訓練集).....	100
圖 A.19 應用 RF 轉換台灣土壤含水量結果(VwPF253 為額外解釋變數，測試集).....	100

集).....	101
圖 A.20 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(VwPF253 為額外解釋變數，訓練集).....	101
圖 A.21 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(VwPF253 為額外解釋變數，測試集).....	102
圖 A.22 應用 RF 轉換台灣土壤含水量結果(VwPF253、Vw15bar_VG 為額外解釋變數，訓練集).....	102
圖 A.23 應用 RF 轉換台灣土壤含水量結果(VwPF253、Vw15bar_VG 為額外解釋變數，測試集).....	103
圖 A.24 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(VwPF253、Vw15bar_VG 為額外解釋變數，訓練集).....	103
圖 A.25 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(VwPF253、Vw15bar_VG 為額外解釋變數，測試集).....	104
圖 A.26 應用預測區間檢核土壤含水量與基質勢能關係之範例	104

表目錄



表 2.1 以多變量線性迴歸所建立之點型 PTF 範例(Rawls et al., 1982).....	10
表 2.2 以多變量線性迴歸所建立之連續型 PTF 範例(Li et al., 2007)	11
表 2.3 不同學者建立共 19 個 PTF(Guber & Pachepsky, 2010).....	17
表 3.1 台灣中部土樣 5187 筆之前 10 筆資料表	25
表 3.2 中部土壤不完整資料於各屬性中無缺值數量表	28
表 3.3 土壤質地與顆粒密度、總體密度及孔隙率之關係表(McCarty et al., 2016; Rowell, 1994)	30
表 3.4 8 個不同基質勢能的土壤含水量所對應的隨機森林架構	36
表 4.1 解釋變數與反應變數之皮爾森積動差相關係數與 p 值表	54
表 4.2 解釋變數與反應變數之斯皮爾曼等級相關係數與 p 值表	55
表 4.3 訓練集之變異數膨脹因子表	56
表 4.4 訓練集解釋變數之間的皮爾森相關係數表	57
表 4.5 各基質勢能預測土壤含水量之多變量線性迴歸截距與係數表	58
表 4.6 三種模型預測各基質勢能土壤含水量均方根誤差和決定係數(訓練集).....	61
表 4.7 三種模型預測各基質勢能土壤含水量均方根誤差和決定係數(測試集).....	62
表 4.8 三種模型預測誤差期望值之 95%信賴區間表(測試集)	67
表 4.9 三種模型在不同基質勢能下的土壤含水量預測區間實際涵蓋率	68
表 4.10 RF 預測各基質勢能土壤含水量均方根誤差和決定係數(6 個解釋變數).....	73
表 4.11 RF 預測各基質勢能土壤含水量均方根誤差和決定係數(7 個解釋變數).....	73
表 4.12 本土連續 PTF 之 95%預測區間的實際涵蓋率表	76

第1章 緒論



1.1 研究背景及動機

土壤水分特性曲線(soil water characteristic curve, SWCC)或稱土壤保水曲線(soil water retention curve, SWRC)為一種描述土壤含水量與基質勢能的函數,對於降雨入滲、地表逕流、農業灌溉、土壤物質傳輸等各方面,均扮演極其重要的角色。土壤保水曲線可藉由擬合實際土壤含水量與基質勢能的量測點獲得,也因此學者發展各式實驗方法以量測不同基質勢能下的土壤含水量。高基質勢能下的量測方法包括濾紙法、蒸發法、懸水柱法(hanging water column)以及張力計實驗,而低基質勢能下的量測方法則有露點法以及壓力鍋實驗(Schelle *et al.*, 2013)。然而上述實驗皆需耗費大量時間進行,短則 2~3 天而長則 2~3 週。因此在前人不斷精進量測土壤含水量實驗的同時,也有一群學者研究以另一種替代方式,希望以較易測得之土壤物理、化學特性,透過函數轉換為不同基質勢能下的土壤含水量或直接轉換為土壤保水曲線。此函數轉換即稱為土壤轉換方程式或稱土壤轉換函數(pedotransfer function, PTF)。

PTF 依照轉換目標可分為點型(Point)與連續型(Continuous),若最終轉換結果為土壤含水量則為點型 PTF;若最終轉換結果為土壤保水曲線參數則為連續型 PTF。PTF 的輸入參數可根據使用者的需求進行調整,Wösten *et al.* (2001)即整理出 PTF 最為廣泛應用的數個參數,分別為粒徑分佈、總體密度或孔隙率、有限的土壤含水量資料(單點或雙點)與有機質或有機碳含量。PTF 也因可自由選定輸入參數,具備高度工程與實務應用性。一旦 PTF 建立完成,即可避免施作成本較高的實驗量測。也因此,在 PTF 此一概念被提出後,數十年間國外相關的研究如雨後春筍般湧現。較為早期的研究如:Hall *et al.* (1977)、Gupta & Larson (1979)、Rawls *et al.* (1982)、Salchow *et al.* (1996)即根據不同土壤特性建立點型 PTF。隨著點型 PTF 發展漸趨成熟,學者逐漸將目光投注在建立連續型 PTF(Li *et al.*, 2007; Rajkai *et al.*, 2004; Saxton & Rawls, 2006; Schaap *et al.*, 2001)。同一時間,也有部分學者致力於研究預測結果的不確定性(Christiaens & Feyen, 2001; Iman, 1992; Minasny & Mc Bratney, 2002; Minasny *et al.*, 1999; Schaap & Leij, 1998; Vereecken *et al.*, 1992)。

過去大量的土壤學者根據不同地區的土壤資料,建立各式各樣的 PTF。然而隨

著土壤領域的應用需求逐漸增高，僅呈現 PTF 的最佳預測結果，已無法滿足現行應用需求 Iman (1992)。透過不確定性分析來衡量預測結果的可靠性，對於最終評估投入的成本與風險至關重要。因此，PTF 必須考慮預測結果的不確定性，並且釐清不確定性的主要因素。國外 PTF 的不確定性評估常應用蒙地卡羅模擬分析 (Minasny & Mc Bratney, 2002)。蒙地卡羅模擬需要假設樣本分佈並評估樣本分佈參數，若 PTF 預測結果與實際值之誤差不為常態或其他常見之統計分佈，即無法直接應用蒙地卡羅模擬分析 PTF 的不確定性。為了改善此問題，部分學者改採自助法評估 PTF 的不確定性。例：美國農業部根據類神經網路所開發之 PTF—Rosetta(Schaap *et al.*, 2001; Schaap *et al.*, 2004; Zhang & Schaap, 2017)。透過自助法即可在不需任何假設前提下，評估 PTF 轉換誤差不為常態或其他常見之統計分佈的不確定性。

然而 PTF 的轉換成效極大部分取決於訓練時所輸入的土壤，台灣目前關於 PTF 的研究較少，轉換成效可能有限。劉滄琴等(2007)即指出，利用國外 PTF 評估台灣土壤的飽和水力傳導係數，可能無法完全反映台灣土壤的區域性因素。而以土壤保水曲線作為轉換目標的本土 PTF 研究則更加稀少。雖有部分學者已就台灣不同區域之土壤，分別建立土壤保水曲線的 PTF(吳晟哲, 2010; 林可薇, 2012; 林俐玲等, 2013; 洪靖惠, 2008)。然而前述研究採取的土壤資料筆數皆較少(小於 100 筆)，樣本的代表性是否足夠為其隱憂。且其對於 PTF 的分析僅止於預測結果驗證，尚無對於預測結果的不確定性分析。對於現行土壤保水曲線的應用，僅根據模型輸出呈現最佳預測結果已逐漸不敷需求。因此台灣有必要根據大量之土壤資料，建立本土土壤保水曲線的 PTF，並進行預測結果的不確定性分析。

1.2 研究目的

根據前述的研究背景與動機，本研究目的具體陳述如下：

- (1). 根據台灣土壤樣本建立本土土壤保水曲線 PTF。
- (2). 建立本土土壤保水曲線 PTF 的不確定性分析。

第2章 文獻回顧



2.1 土壤保水曲線

2.1.1 土壤水總勢能

土壤內的水勢能可分為三個部分，包括壓力勢能(pressure potential)、重力勢能(gravitational potential)、滲透勢能(osmotic potential)，如式(2.1)所示(Mullins *et al.*, 2000)。

$$\psi_t = \psi_p + \psi_g + \psi_o \quad (2.1)$$

$$\psi_p = \psi_c + \psi_m \quad (2.2)$$

$$\psi_h = \psi_p + \psi_g = \psi_c + \psi_m + \psi_g \quad (2.3)$$

$$\psi_w = \psi_o + \psi_m \quad (2.4)$$

ψ_t 為土壤內的水總勢能。 ψ_p 為壓力勢能。 ψ_g 為重力勢能。 ψ_o 為滲透勢能。

ψ_c 為毛細勢能(capillary potential)。 ψ_m 為基質勢能。

ψ_h 為水力勢能(hydraulic potential)。 ψ_w 為水勢能(water potential)。

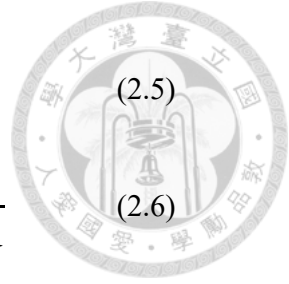
壓力勢能中又包含毛細勢能以及基質勢能，如式(2.2)所示。基質勢能總小於0且範圍介於 $-\infty \sim 0$ ，土壤內含水量越高則基質勢能也愈大，常使用的單位為巴(bar)或公分水柱(cm-water)。重力勢能與壓力勢能之和又稱為水力勢能，如式(2.3)所示。土壤間不同位置的水力勢能差值即會決定土壤間水的流動方向。基質勢能與滲透勢能之和又稱為水勢能，如式(2.4)所示。水勢能的高低會直接影響，土壤液—汽相交界面在平衡時的相對濕度大小。

2.1.2 與土壤含水量之關係

土壤含水量為土壤中所含的水量，其包括兩種不同的物理定義。第一種以重量為基礎表示，如式(2.5)。第二種則以體積為基礎表示，如式(2.6)。一般而言以第二種方式即式(2.6)，表示土壤含水量(Or *et al.*, 2002)。

$$\theta_m = \frac{\text{水重(mass of water)}}{\text{乾土重(mass of dry soil)}} = \frac{\text{溼土重} - \text{乾土重}}{\text{乾土重}} \quad (2.5)$$

$$\theta_v = \frac{\text{水體積(volume of water)}}{\text{土總體積(bulk volume of water)}} = \frac{\text{水重/水密度}}{\text{乾土重/總體密度}} \quad (2.6)$$



式(2.5)、(2.6)可以透過式(2.7)進行轉換。

$$\theta_v = \theta_m \frac{\rho_b}{\rho_w} \quad (2.7)$$

其中 ρ_b 為總體密度(即烘乾土重與原始體積的比率，單位為克/立方公分， g/cm^3)、 ρ_w 為水密度(在攝氏 20°C 為 1000 公斤/立方公尺 ， 1000 kg/m^3 at 20°C)。

土壤含水量多寡與土壤種類、孔隙結構、質地組成、基質勢能等密切相關。其中基質勢能為影響土壤含水量多寡的關鍵因素。目前已有多種不同的實驗可以量測不同基質勢能範圍的土壤含水量。基質勢能小於 1bar 的量測方法以濾紙法、蒸發法、懸水柱法以及張力計實驗為主；基質勢能介於 $1\sim 15\text{bar}$ 的量測方法則有露點法以及壓力鍋實驗(Schelle *et al.*, 2013)。採取現地土樣後，即可透過實驗取得特定基質勢能下的土壤含水量散佈點。若將所有散佈點以一個數學函數進行曲線擬合，此擬合曲線即為土壤保水曲線。因此土壤保水曲線具有許多不同的候選函數，本研究將以目前應用最為廣泛的 VG model(van Genuchten, 1980)作為研究與討論主軸(函數形式及推導將於 2.1.3 小節詳細闡述)。若以 VG model 擬合土壤含水量散佈點，可繪製曲線圖如圖 2.1 所示。

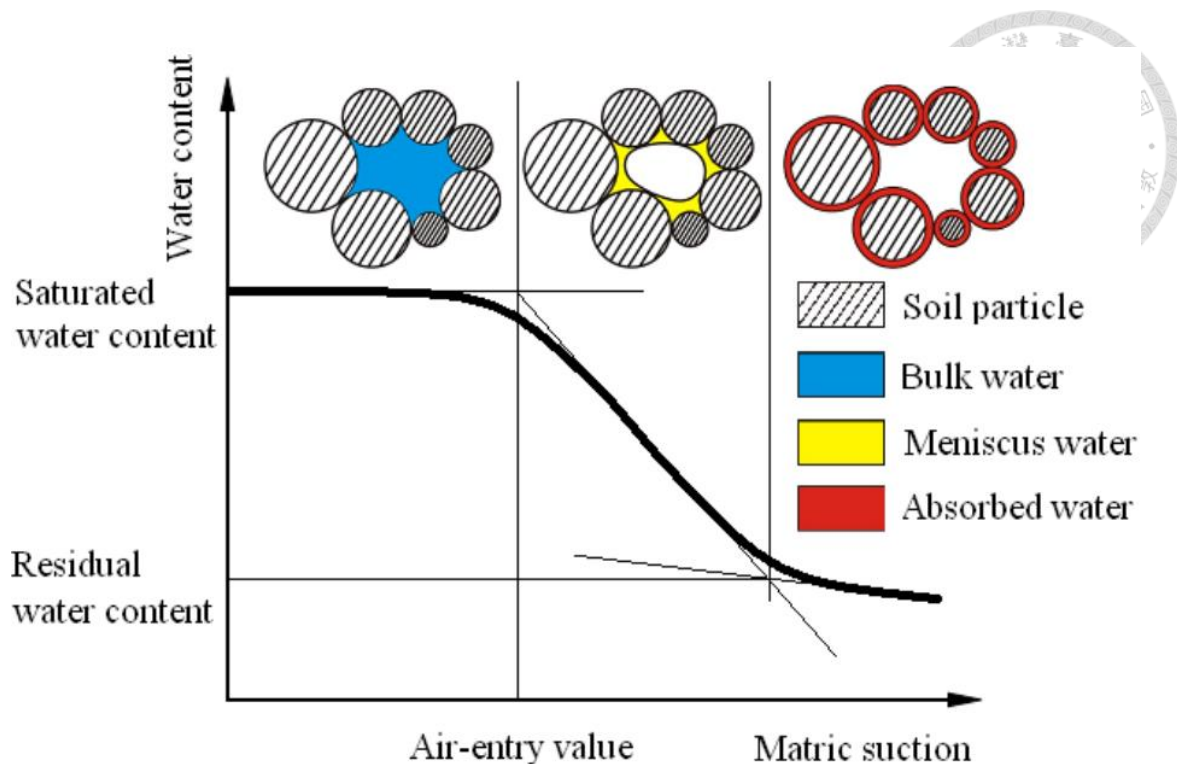


圖 2.1 土壤保水曲線與土壤含水量關係示意圖(Matsuoka *et al.*, 1999)。橫軸為基質勢能，單位為公分水柱或巴。縱軸為土壤含水量，單位為體積含水量(cm^3/cm^3)。

透過圖 2.1 得知，VG model 為描述土壤含水量與基質勢能之間變化關係的 S 型曲線，其上下界可分別對應於土壤的飽和含水量與殘餘含水量。S 型曲線的轉折點則可由空氣進入壓力(air-entry pressure)決定。Brooks & Corey (1964)定義空氣進入壓力為第一個或最大的孔隙開始排退水的基質勢能。Matsuoka *et al.* (1999)依照基質勢能的範圍，將土壤含水量分成三個部分：基質勢能大於空氣進入壓力時，土壤含水量主要由土壤結構所控制(孔隙水)；基質勢能降低至接近殘餘含水量情況時，土壤含水量則改由土壤質地所控制(吸附水)；若基質勢能介於兩者之間，土壤含水量則由孔徑分佈不均導致之介面張力差異所控制(半月水)。Dexter (2004)也提出類似的看法，以土壤保水曲線反曲點為分界，區分孔隙水與吸附水。

2.1.3 VG model

水力傳導係數受到土壤含水量的影響。在土壤含水量較低時，水僅能藉由土壤固體顆粒周圍的薄水膜，流經曲折且狹小的通道；而在土壤含水量較高時，水流經的通道為飽和或接近飽和的孔隙，大幅增加水力傳導係數(Sojka *et al.*, 2009)。圖 2.2

即描述土壤含水量與水力傳導係數之關係。

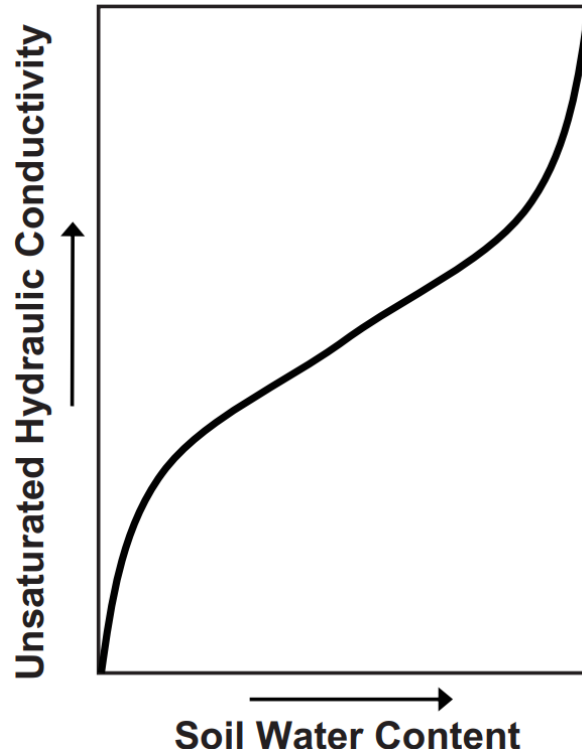


圖 2.2 土壤含水量與水力傳導係數關係示意圖(Sojka et al., 2009)。橫軸為土壤含水量。縱軸為非飽和水力傳導係數。土壤含水量與水力傳導係數呈現正相關的趨勢。

土壤保水曲線與土壤含水量—水力傳導係數曲線的第一個轉折點皆與空氣進入壓力相關。土壤保水曲線和土壤含水量—水力傳導係數曲線，皆與土壤基質勢能相關且皆為 S 型曲線。

Mualem (1976)根據土壤保水曲線與水力傳導係數的關係，推導出預測水力傳導係數的方程式，如式(2.8)。

$$K_r = \theta^{1/2} \left[\int_0^\theta \frac{1}{h(x)} dx / \int_0^1 \frac{1}{h(x)} dx \right]^2 \quad (2.8)$$

$$K_r = \frac{K}{K_{sat}} \quad (2.9)$$

$$\theta = \frac{\theta - \theta_r}{\theta_s - \theta_r} \quad (2.10)$$

式(2.8)~(2.10)中， h 為基質勢能的絕對值。 θ 為無因次之土壤含水量。 K 、 K_{sat} 分別為非飽和、飽和水力傳導係數。 θ 為土壤當前含水量。 θ_r 為土壤殘餘含水量。 θ_s 為土壤飽和

含水量。

為了求解式(2.8)，需要一個土壤保水曲線函數，連接基質勢能與土壤含水量之關係。van Genuchten (1980)於是參考前人所提出的式(2.11)(Ahuja & Swartzendruber, 1972; Endelman *et al.*, 1974; Haverkamp *et al.*, 1977)，將其無因次化並增加額外未知數重新改寫為式(2.12)，即為 VG model 的雛形。

$$\theta = \frac{\alpha(\theta_s - \theta_r)}{\alpha + h^n} + \theta_r \quad (2.11)$$

$$\theta = \left[\frac{1}{1 + (\alpha h)^n} \right]^m \quad (2.12)$$

式(2.11)、(2.12)中， α 、 m 、 n 為待定之參數且 $h > 0$ 。

然而式(2.12)將 m 、 n 視為兩個完全獨立的參數，與目前廣泛使用的 VG model($m = 1 - 1/n$)仍有些許不同。van Genuchten (1980)欲使用式(2.12)求解式(2.8)，於是將式(2.12)整理為式(2.13)，並將式(2.13)代入式(2.8)中即可得式(2.14)。

$$h = \frac{1}{\alpha} \left(\frac{1 - \theta^{\frac{1}{m}}}{\theta^{\frac{1}{m}}} \right)^{\frac{1}{n}} \quad (2.13)$$

$$K_r = \theta^{1/2} \left[\frac{f(\theta)}{f(1)} \right]^2 \quad (2.14)$$

$$f(\theta) = \int_0^\theta \left[\frac{x^{1/m}}{1 - x^{1/m}} \right]^{1/n} dx \quad (2.15)$$

式(2.15)中的 x 即為 θ ，此為避免 θ 、 θ 同時使用造成混淆。

若假設 $x = y^m$ 則式(2.15)經過變數變換可化簡為式(2.16)。

$$f(\theta) = m \int_0^{\theta^{1/m}} y^{m-1+1/n} (1 - y)^{-1/n} dy \quad (2.16)$$

式(2.16)為一種不完全貝塔函數(Incomplete Beta-function)的特殊形式，在一般情況下無法導出其解析解(analytic expression)或封閉解(closed-form expression)。van Genuchten (1980)即令 $k = m - 1 + 1/n$ 將其代入式(2.16)，即可利用不完全貝塔函數

的性質，快速地計算不定積分。若在 $k = 0$ 的特殊情況下，則式(2.16)可化簡為式(2.17)，式(2.14)則可表示為式(2.18)。

$$f(\theta) = 1 - \left(1 - \theta^{\frac{1}{m}}\right)^m \quad (m = 1 - 1/n) \quad (2.17)$$

$$K_r(h) = \frac{\{1 - (\alpha h)^{n-1}[1 + (\alpha h)^n]^{-m}\}^2}{[1 + (\alpha h)^n]^{m/2}} \quad (m = 1 - 1/n) \quad (2.18)$$

式(2.18)即為 Mualem's model 基於 VG model 的解析解。此時連接基質勢能與土壤含水量函數的式(2.12)，即可根據 $(m = 1 - 1/n)$ 與式(2.10)的資訊，整理而得式(2.19)。此即為目前最廣為使用的土壤保水曲線函數—VG model。

$$\theta(h) = \theta_r + \frac{\theta_s - \theta_r}{[1 + (\alpha h)^n]^{1-1/n}} \quad (2.19)$$

透過 Mualem's model 式(2.18)以及 VG model 式(2.19)即可建立土壤含水量與飽和水力傳導係數、土壤保水曲線之關係如圖 2.3 所示。

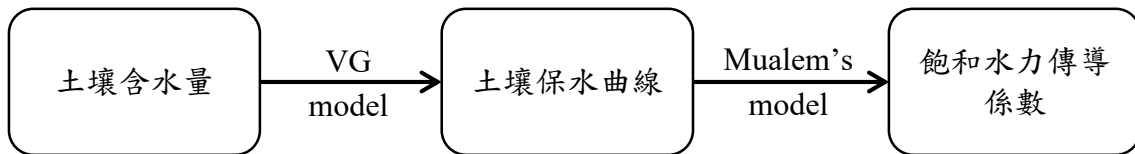


圖 2.3 土壤含水量與土壤保水曲線、飽和水力傳導係數關係圖

2.2 土壤轉換方程式(Pedotransfer Function, PTF)

透過實驗雖然可以直接量測真實土壤現況，但是土樣採取與實驗過程需耗費大量時間與人力，不利於現地大範圍應用或後續的數值模擬(Gupta & Larson, 1979; Rawls *et al.*, 1982)。且土壤受到地形、溫度、雨量、生物等多重因素的影響，導致空間、時間分佈高度不均質，使實驗量測值的代表性有所侷限。因此數十年來，大量土壤物理學家創建數種不同的 PTF。PTF 此一名詞由 Bouma (1989) 首先提出，並定義為將現有土壤數據轉換為所需土壤數據的轉換方程式。利用較易測得之土壤性質，進一步推估土壤保水曲線、水力傳導係數(Ahuja *et al.*, 1985; Arya & Dierolf, 1992; Arya & Paris, 1981; Bloemen, 1980; Costa *et al.*, 2013; Dashtaki *et al.*, 2010; Fooladmand, 2011; Li *et al.*, 2007)。本研究將著重於討論土壤保水曲線的 PTF。

2.2.1 點型與連續型 PTF

PTF 根據其轉換結果可以分為兩類，點型 PTF(point PTF)和連續型 PTF(continuous PTF)。點型 PTF 屬於離散預測，主要預測數個關鍵的含水量，包含田間含水量(Field Capacity, FC)與永久凋萎點(Permanent Wilting Point, PWP)。田間含水量對於粗糙質地土壤，其基質勢能約介於 $-5 \sim -10$ kPa；中等質地土壤，其基質勢能約為 -33 kPa；細緻質地土壤，其基質勢能約為 -50 kPa(Colman, 1947; Jamison & Kroth, 1958; Rivers & Shipp, 1978)。而永久凋萎點則與土壤質地較無直接關係，不同土壤其基質勢能皆約為 -1500 kPa(Richards & Weaver, 1943)。點型 PTF 對於單點預測較佳但缺點即為無法預測所有基質勢能的土壤水分含量。Gupta & Larson (1979)、Rawls *et al.* (1982)、Salchow *et al.* (1996)等人建立的 PTF 皆屬於點型 PTF。表 2.1 即為 Rawls *et al.* (1982)根據不同解釋變數所建立的點型 PTF 結果。

表 2.1 以多變量線性迴歸所建立之點型 PTF 範例(Rawls et al., 1982)。表格中解釋變數依序為砂、粉、黏、有機質含量，總體密度以及田間含水量與永久凋萎點。反應變數則是不同基質勢能下的土壤含水量。

Metric potential bars	Intercept	Sand, %	Silt, %	Clay, %	Organic matter, %	Bulk density, g/cm ³	0.33 bar water retention, cm ³ /cm ³	15 bar water retention, cm ³ /cm ³	Correlation coefficient, R
----- Regression coefficients -----									
	a	b	c	d	e	f	g	h	
-0.04	0.7899	-0.0037			0.0100	-0.1315			0.58
	0.6275	-0.0041			0.0239			-0.08	0.57
	0.1829				-0.0246	-0.0376	1.89	-1.38	0.77
-0.07	0.7135	-0.0030		0.0017		-0.1693			0.74
	0.4829	-0.0035			0.0263			0.25	0.74
	0.8888	-0.0003			-0.0107		1.53	-0.81	0.91
-0.10	0.4118	-0.0030		0.0023	0.0317				0.81
	0.4103	0.0031			0.0260			0.41	0.81
	0.0619	-0.0002			-0.0067		1.34	-0.51	0.95
-0.20	0.3121	-0.0024		0.0032	0.0314				0.86
	0.3000	-0.0024			0.0235			0.61	0.89
	0.0319	-0.0002					1.01	-0.06	0.99
-0.33	0.2576	-0.0020		0.0036	0.0299				0.87
	0.2391	-0.0019			0.0210			0.72	0.92
-0.60	0.2065	-0.0016		0.0040	0.0275				0.87
	0.1814	-0.0015			0.0178			0.80	0.94
	0.0136					-0.0091	0.66	0.39	0.99
-1.0	0.0349		0.0014	0.0055	0.0251				0.87
	0.1417	-0.0012			0.0151			0.85	0.96
	-0.0034				0.0022		0.52	0.54	0.99
-2.0	0.0281		0.0011	0.0054	0.0200				0.86
	0.0986	0.0009			0.0116			0.90	0.97
	-0.0043				0.0026		0.36	0.69	0.99
-4.0	0.0238		0.0008	0.0052	0.0190				0.84
	0.0649	-0.0006			0.0085			0.93	0.98
	-0.0038				0.0026		0.24	0.79	0.99
-7.0	0.0216		0.0006	0.0050	0.0167				0.81
	0.0429	-0.0004			0.0062			0.94	0.98
	-0.0027				0.0024		0.16	0.86	0.99
-10.0	0.0205		0.0005	0.0049	0.0154				0.81
	0.0309	-0.0003			0.0049			0.95	0.99
	-0.0019				0.0022		0.11	0.89	0.99
-15.0	0.0260			0.0050	0.0158				0.80

Sand (%) + silt (%) + clay (%) = 100 Sand = 2.0–0.5 mm Silt = 0.05–0.002 mm Clay < 0.002

$O_x = a + b \times \text{sand} (\%) + c \times \text{silt} (\%) + d \times \text{clay} (\%) + e \times \text{organic matter} (\%) + f \times \text{bulk density} (\text{g}/\text{cm}^3) + g \times 0.33 \text{ bar moisture} (\text{cm}^3/\text{cm}^3) + h \times 15 \text{ bar moisture} (\text{cm}^3/\text{cm}^3)$

O_x = predicted water retention (cm³/cm³) for a given metric (x) potential

a–n = regression coefficients

連續型 PTF 則預測土壤保水曲線函數的參數。連續型 PTF 對於後續進行數值模擬或其他應用較便利於點型 PTF(Zacharias & Wessolek, 2007)。多數連續型 PTF 以 van Genuchten (1980)的 VG-model 與 Brooks & Corey (1964)的 BC-model，兩種土壤保水曲線函數的參數作為最後轉換結果。Li *et al.* (2007)、Saxton & Rawls (2006)、Schaap *et al.* (2001)、(Rajkai *et al.*, 2004)等人建立的 PTF 即屬於連續型 PTF。表 2.2 即為 Li *et al.* (2007)根據不同解釋變數所建立的連續型 PTF 結果。

表 2.2 以多變量線性迴歸所建立之連續型 PTF 範例(Li et al., 2007)。方程式中解釋變數 SAND 為砂、SILT 為粉、CLAY 為黏、SOM 為有機質含量、BD 為總體密度。反應變數則是飽和水力傳導係數 K_s 與 VG model 的 3 個參數， θ_s 為飽和含水量、 α 則與空氣進入壓力倒數相關、 n 與土壤孔隙分佈相關， θ_r 在此假設為 0。SSE 為殘差平方和，SSR 為總平方和與殘差平方和之差值。

Model Parameters	Regression equations	SSR	SSE	R^2_{adj}
$\ln(\theta_s)$	$-1.531 + 0.212 \cdot \ln(\text{SAND}) + 0.006 \cdot \text{SILT} - 0.051 \cdot \text{SOM} - 0.566 \cdot \ln(\text{BD})$	0.355	0.206	0.61
$\ln(\alpha)$	$-67.408 - 0.040 \cdot \text{SILT} - 0.670 \cdot \ln(\text{SILT}) - 2.189 \cdot \text{SOM} + 1.410 \cdot \ln(\text{SOM}) + 78.400 \cdot \text{BD} - 121.331 \cdot \ln(\text{BD})$	120.0	76.12	0.57
n	$1.488 + 0.002 \cdot \ln(\text{SILT}) + 0.013 \cdot \text{CLAY} - 0.248 \cdot \ln(\text{CLAY}) + 0.048 \cdot \ln(\text{SOM}) + 0.451 \cdot \ln(\text{BD})$	0.642	0.524	0.51
$\ln(K_s)$	$13.262 - 1.914 \cdot \ln(\text{SAND}) - 0.974 \cdot \ln(\text{SILT}) - 0.058 \cdot \text{CLAY} - 1.709 \cdot \ln(\text{SOM}) + 2.885 \cdot \text{SOM} - 8.026 \cdot \ln(\text{BD})$	25.94	10.13	0.66

2.2.2 建立 PTF 的工具或方法

決定 PTF 的輸入參數與最終輸出結果後，尚需要選擇一種方法建立輸入和輸出的關係。目前廣用於建立 PTF 的方法包括多變量迴歸分析(Multiple regression, MR)、類神經網路(Artificial neural network, ANN)、K 近鄰(K Nearest Neighbors)、迴歸樹等。以下將著重介紹最常應用於建立 PTF 的多變量迴歸分析、類神經網路與迴歸樹演算法。

多變量迴歸可分為線性與非線性迴歸。早期學者建立 PTF 的方式幾乎全使用多變量線性迴歸(Multiple linear regression, MLR)，Gupta & Larson (1979)、Rawls *et al.* (1982)等人所建立的 PTF 即屬於此類型。而後 Rawls & Brakensiek (1985)改採多變量非線性迴歸建立 PTF。使用迴歸分析的優點為，可以使用逐步迴歸法找到最重要的輸入參數。然而多變量迴歸也具有不少缺點，包括方程式只能模擬特定形狀的

關係。且一旦輸入大量數據進行訓練時，對於整體數據迴歸分析往往無法得到令人滿意的解釋成果。為了改善此情形，Lin *et al.* (1999)嘗試使用主成分分析尋找新輸入參數。主成分分析所尋找的新輸入參數是原始輸入參數的線性組合，並可以顯著提升解釋變數對於反應變數變異性的解釋力。然而即使如此，迴歸分析的結果仍然有所侷限。因此更多學者改採其他機器學習的方法建立 PTF，例如：類神經網路、和迴歸樹。

類神經網路為建構變數間複雜關係時所常用的模型，尤其是當解釋與反應變數具有高度非線性關係時，類神經網路具有廣泛的應用性(Maren *et al.*, 1990; McCord-Nelson & Illingworth, 1991)。類神經網路由許多相互關聯且可簡單計算的元素組成，稱為節點或神經元。神經元具有多種類型與連接方式，不同的組合選擇可構成截然不同的神經網路架構(Hecht-Nielsen, 1989)。其中最常被應用於建立 PTF 的架構為前饋神經網路(feedforward neural network)，圖 2.4 即為根據前饋神經網路所建構的 PTF 範例。類神經網路的優勢在於，可藉由改變神經元的組成以及層與層之間相互連結的範圍，從而影響彼此且進一步模擬複雜系統的行為。一旦建立並訓練完成此網路架構，即可有效聯繫輸入與輸出值。當輸入變數數量大於三個時，類神經網路通常可以獲得優於迴歸分析的結果，即使迴歸分析採用更為複雜的多項式迴歸(Hecht-Nielsen, 1989)。也因此許多學者改採用前饋神經網路建立 PTF，例如：Pachepsky *et al.* (1996)、Schaap & Bouten (1996)、Koekkoek & Boeltink (1999)即是應用此架構建立 PTF。

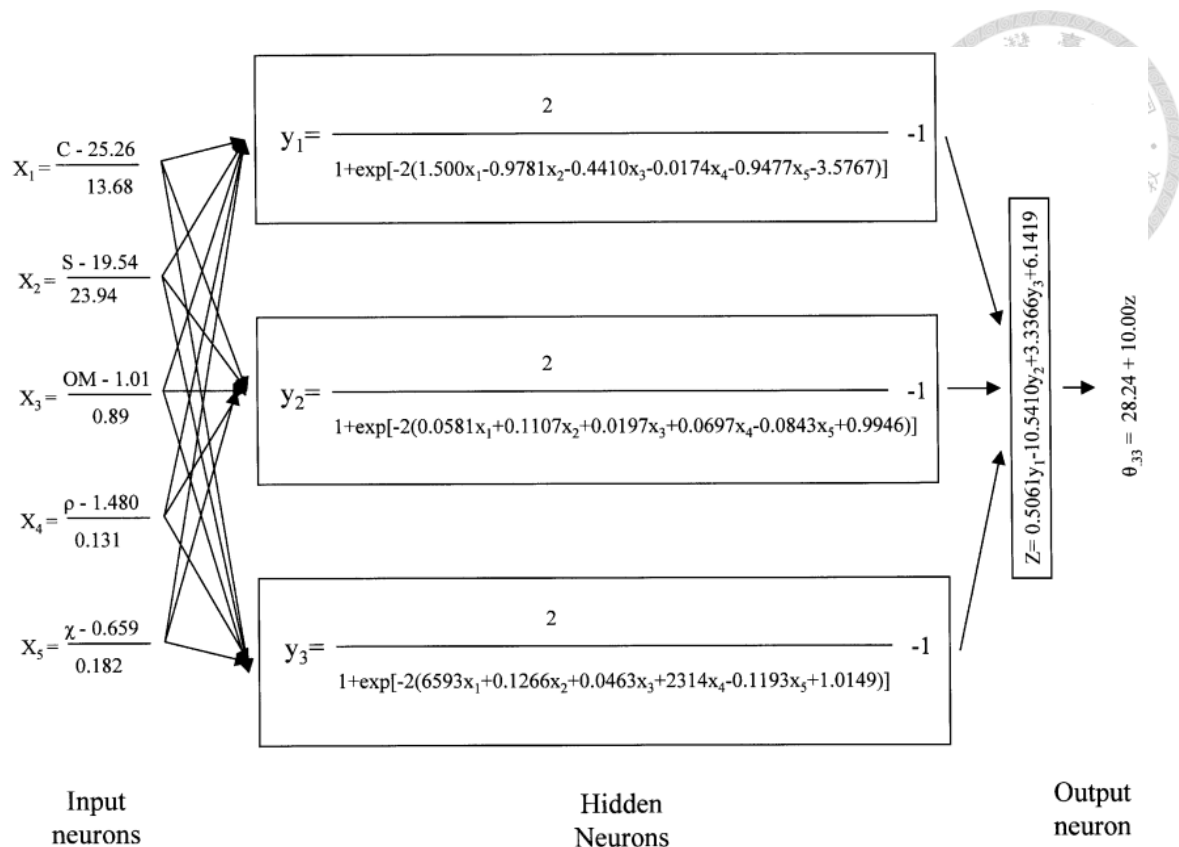


圖 2.4 運用前饋神經網路建立田間含水量的 PTF 範例(Wösten et al., 2001)。其中 C 為黏、S 為砂、OM 為有機質含量、 ρ 為總體密度、 χ 為活性指數。活性指數為塑性指數與黏含量的比值。

分類和迴歸樹(Classification and Regression Trees, CART)演算法最早由 Breiman (1984)所提出。CART 演算法最大的特色為，將輸入資料依照階層的篩選條件作循環切割(recursive partition)，且所有解釋變數皆可在不同的篩選條件中被重複使用。CART 模型訓練主要藉由最小化雜質(impurity)或最大化資訊增益(information gain)完成。模型最終訓練完成後為一顆帶有許多分支和多個終端節點的樹，若模型最終結果為多顆具有分支和終端節點的樹則為隨機森林(Random forest, RF)。相較於類神經網路，CART 具有數個優點包括：低計算成本、模型具有高度解釋力。McKenzie & Jacquier (1997)最早使用 CART 模型開發 PTF。圖 2.5 即為根據決策樹所建構的 PTF 範例。

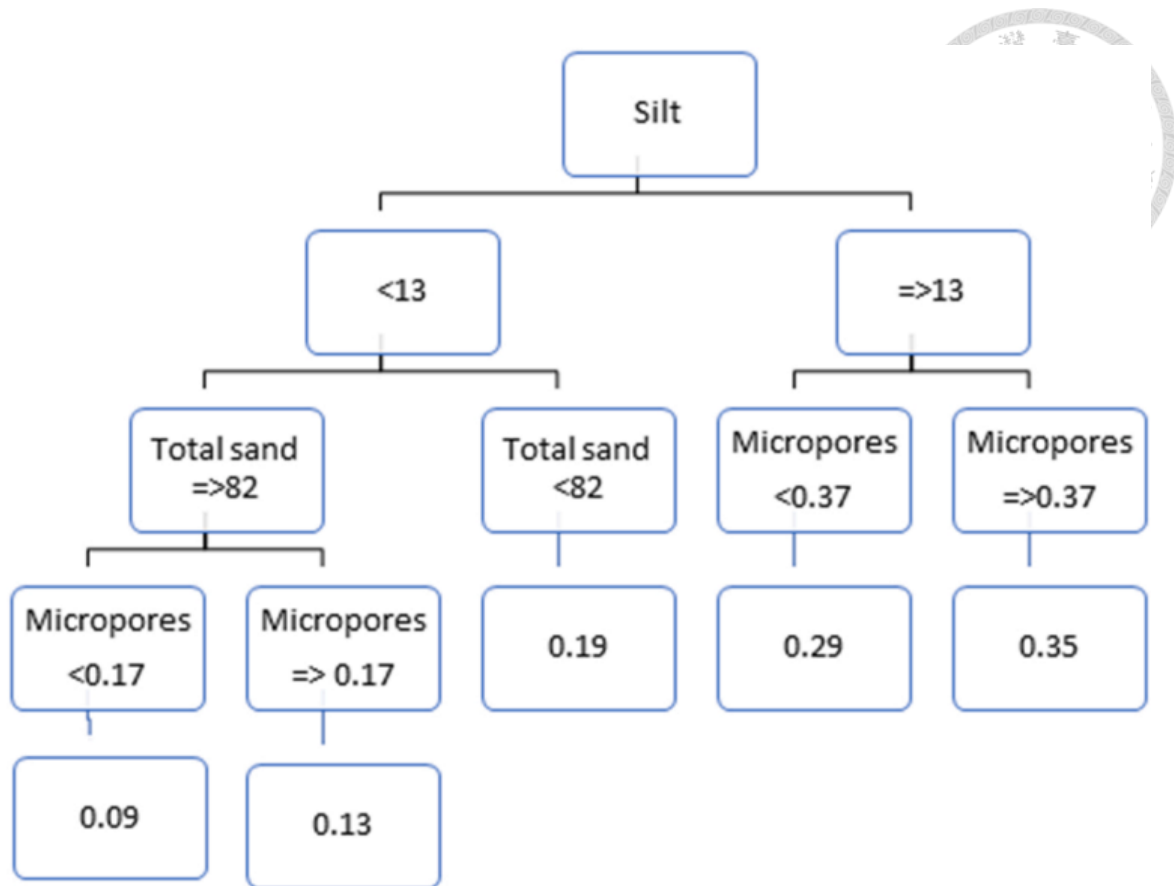


圖 2.5 運用決策樹建立田間含水量的 PTF 範例(Amorim et al., 2022)。Silt 為土壤粉含量。Total sand 為土壤砂含量。Micropores 定義為基質勢能 -6kPa 時的土壤含水量。

2.2.3 作為預測因子的土壤性質

土壤質地為較容易測得且時常被應用於建立 PTF 的土壤性質。Bloemen (1980) 首先引入土壤粒徑的中位數作為土壤質地分類的指標。Shirazi & Boersma (1984)則進一步提出三角質地圖作為美國農業部(United States Department of Agriculture, USDA)的土壤質地分類依據，如圖 2.6 所示。

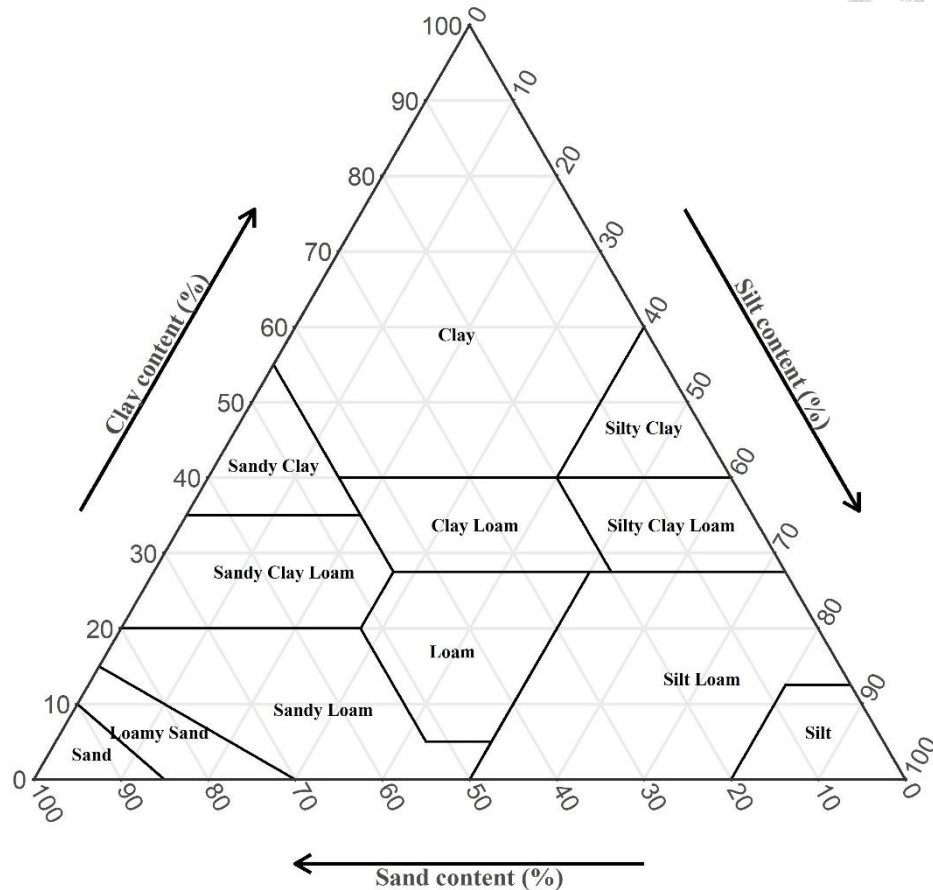


圖 2.6 USDA 土壤三角質地圖(Shirazi & Boersma, 1984)

McCuen *et al.* (1981)透過統計分析表明，BC model 的參數隨著美國農業部土壤質地分類(USDA soil texture)成系統性的變化。Campbell & Shiozawa (1992)隨後將此概念用於連結土壤質地與土壤水力特性。Jaynes & Tyler (1984)、El-Kadi (1985)發現僅使用土壤質地預測砂質土壤的飽和水力傳導係數，已有相當優異的結果。Puckett *et al.* (1985)進一步點出，砂質土壤以外的其他土壤，以黏含量為最主要與飽和水力傳導係數相關的土壤性質。Saxton *et al.* (1986)則根據前人的分析，依照土壤質地建立土壤水力特性的 PTF。

總體密度或孔隙率為另一個時常被應用於建立 PTF 的土壤性質。與土壤質地相同，總體密度或孔隙率皆為較容易量測之土壤性質。總體密度與顆粒密度、孔隙率直接相關，三者之關係可以式(2.20)表示。

$$\phi = \frac{V_v}{V_T} = 1 - \frac{BD}{\rho_p} \quad (2.20)$$

ϕ 為土壤孔隙率，單位為無因次。 V_v 為土壤孔隙體積。 V_T 為土壤總體積。 ρ 為總體密度，單位為 g/cm^3 。 ρ_p 為顆粒密度，單位為 g/cm^3 。

孔隙率為衡量土壤結構的一個指標，其數值大小控制土壤含水量的上限，即土壤飽和含水量。Dexter (2004) 即指出低基質勢能的土壤含水量由土壤結構所主宰。若顆粒密度為一定值(約略介於 $2.6 \sim 2.8 \text{g/cm}^3$ 之間)，則透過(2.20)式可以得知總體密度與土壤結構直接相關，也因此總體密度與土壤含水量密切相關。Berndt & Coughlan (1977) 也指出開裂黏土的總體密度將隨土壤含水量變化而改變，示意圖如圖 2.7 所示。

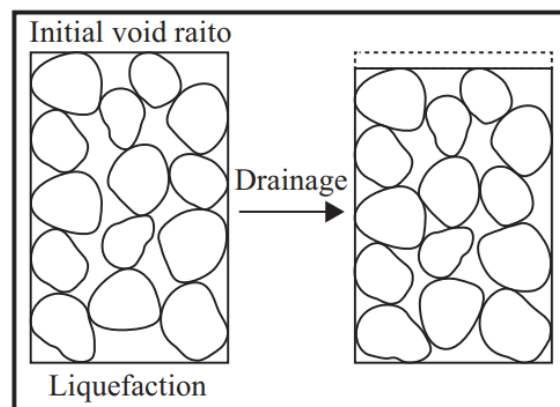


圖 2.7 總體密度隨土壤含水量改變示意圖(Kim et al., 2019)

Gupta & Larson (1979) 即已嘗試運用總體密度、粒徑分佈和有機質含量推估土壤保水曲線。Aina & Periaswamy (1985)、Bruand *et al.* (1996)、Rajkai & Várallyay (1992) 等學者，皆使用孔隙率或總體密度作為建構 PTF 的解釋變量。

有機質或有機碳含量也時常作為 PTF 的解釋變量。Bloemen (1980) 在研究中證明了總體密度和有機質含量之間的相關性，有機質含量也因此與土壤含水量具有部分相關性。Rawls *et al.* (1982)、Rawls *et al.* (1983)、Wösten *et al.* (1999) 等學者，皆引入有機質或有機碳含量作為建構 PTF 的額外變量。Guber & Pachepsky (2010) 收集並整理多位學者建立的 PTF，如表 2.3。

表 2.3 不同學者建立共 19 個 PTF(Guber & Pachepsky, 2010)。Model 表示最終輸出結果為 VG model、BC model、土壤含水量(WH)。OC、BD 依序表示有機碳和總體密度。Depth 表示土壤採樣深度。

No.	PTF	Region	Number of soil samples	Model ^{b)}	Input soil property						Reference
					Sand	Silt	Clay	OC	BD	Depth	
1	Saxton <i>et al.</i> PTF	USA	5 320	BC	+ ^{c)}		+		+		Saxton <i>et al.</i> (1986)
2	Campbell and Shizawa PTF	No particular	6	BC	+		+		+		Campbell and Shizawa (1992)
3	Rawls and Brakensiek PTF	USA	5 320	BC	+		+		+		Rawls and Brakensiek (1985)
4	Williams <i>et al.</i> PTF	Australia	196	BC	+		+		+		Williams <i>et al.</i> (1992)
5	Williams <i>et al.</i> PTF	Australia	196	BC	+		+	+	+		Williams <i>et al.</i> (1992)
6	Oosterveld and Chang PTF	Alberta, Canada	298	BC	+		+		+	+	Oosterveld and Chang (1980)
7	Mayr and Jarvis PTF	UK	306	BC	+	+	+	+	+		Mayr and Jarvis (1999)
8	Wösten <i>et al.</i> PTF	Europe	4 030	VG	+	+	+			+	Wösten <i>et al.</i> (1999)
9	Varallyay <i>et al.</i> PTF	Hungary	230	VG			+		+		Varallyay <i>et al.</i> (1982)
10	Vereecken <i>et al.</i> PTF	Belgium	182	VG	+		+	+	+		Vereecken <i>et al.</i> (1989)
11	Wösten <i>et al.</i> PTF	Europe	4 030	VG		+	+	+	+	+	Wösten <i>et al.</i> (1999)
12	Tomasella and Hodnett PTF	Brazil	196	WH -> VG ^{d)}		+	+	+			Tomasella and Hodnett (1998)
13	Rawls <i>et al.</i> PTF	USA	5 320	WH -> VG	+	+	+	+	+		Rawls <i>et al.</i> (1983)
14	Gupta and Larson PTF	Central USA	43		+	+	+	+	+		Gupta and Larson (1979)
15	Rajkai and Varallyay PTF	Hungary	270	WH -> VG	+		+	+	+		Rajkai and Varallyay (1992)
16	Petersen <i>et al.</i> PTF	Pennsylvania, USA	1 267	WH -> VG			+				Petersen <i>et al.</i> (1968)
17	Bruand <i>et al.</i> PTF	Central France	20	WH			+				Bruand <i>et al.</i> (1994)
18	Canarache PTF	Romania	Unknown	WH			+		+		Canarache (1993)
19	Hall <i>et al.</i> PTF	England, Wales, UK	261	WH	+	+	+		+		Hall <i>et al.</i> (1977)

^{a)}OC = organic carbon; BD = bulk density.

^{b)}BC = Brooks and Corey; VG = van Genuchten; WH = water content at selected capillary pressure.

^{c)}Required input variable.

^{d)}Point predictions based on VG parameters.

從表 2.3 中可以發現，不同地區的土壤學者，皆應用土壤質地的砂、粉和黏土含量，其中 1~3 項作為 PTF 輸入變數。且大多數學者會額外採用總體密度作為 PTF 的額外變量。若土壤性質資料完整，則會再額外追加有機質含量作為 PTF 的第二個額外變量。

2.2.4 轉換表現評估指標

PTF 作為推估土壤保水曲線的轉換方程式，可根據實測值與預測值之間的關係進行評估。若實測值應用於訓練 PTF，則可藉由訓練完成之 PTF 預測值與實測值之關係，評估 PTF 的轉換準確性(此即使用訓練集資料訓練模型)；反之，若實測值未應用於訓練 PTF，則可藉由先前已完成訓練的 PTF 預測值與實測值之關係，評估 PTF 的轉換可靠性(此即使用測試集資料檢驗模型)。以下將列舉出文獻中，時常應用於評估轉換表現的指標。



1. 相關係數

相關係數有多種不同的型式，其中皮爾森積動差相關係數(Pearson product-moment correlation coefficient，以下簡稱為 ρ)與斯皮爾曼等級相關係數(Spearman's rank correlation coefficient，以下簡稱為 ρ_s)為兩種最常見的相關係數，其計算方法如(2.21)、(2.22)式所示。

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2} \sqrt{\sum_i^n (Y_i - \bar{Y})^2}} \quad (2.21)$$

$$\rho_{s_{X,Y}} = \frac{cov[R(X), R(Y)]}{\sigma_{R_X} \sigma_{R_Y}} \quad (2.22)$$

$cov(X,Y)$ 為兩個變數的共變異數。 σ_X 、 σ_Y 為兩變數的標準差。 $R(X)$ 、 $R(Y)$ 則分別是 X 、 Y 兩變數轉換後的等級變數。 σ_{R_X} 、 σ_{R_Y} 則為兩等級變數的標準差。

皮爾森相關係數主要用於度量變數之間的線性相關程度。解釋變數與反應變數的皮爾森相關係數絕對值越高($|\rho| \rightarrow 1$)，即表示兩者具有高度線性相關，其中正值為線性正相關，負值則為線性負相關。斯皮爾曼相關係數則以單調函數度量變數之間的相關程度。解釋變數與反應變數的斯皮爾曼相關係數越高($|\rho_s| \rightarrow 1$)，即表示兩者具有高度單調(monotone)相關，其中正值為單調正相關，負值則為單調負相關。若已知變數之間不為線性變化，則可利用斯皮爾曼相關係數衡量兩者的單調性，斯皮爾曼相關係數絕對值越高即表示兩者變化越趨一致性，兩者也愈可能存在某種函數關係。Cornelis *et al.* (2001)、Sharma *et al.* (2006)即分別使用皮爾森積動差相關係數與斯皮爾曼等級相關係數，作為 PTF 的轉換表現評估指標。然而相關係數無法有效考慮資料離群值的影響為其一大隱憂。

2. 決定係數

決定係數(determination coefficient，常以 R^2 表示)為統計學中，用於衡量模型的反應變數變異可由解釋變數詮釋的比例。若預測模型為線性迴歸，決定係數即為皮爾森積動差相關係數的平方。決定係數的計算方法如式(2.23)所示。

$$R^2 = 1 - \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - \bar{y}_i)^2} = 1 - \frac{SS_{res}}{SS_{tot}} \quad (2.23)$$

y 代表樣本實際值。 $\hat{y}$ 代表樣本預測值。 N 代表樣本總數量。 SS_{res} 代表殘差平方和。 SS_{tot} 代表總平方和。

Stumpp *et al.* (2009)即使用決定係數作為 PTF 的轉換表現評估指標。然而應用決定係數作為模型表現評估指標，可能具有以下缺點：一旦大量增加預測模型的解釋變數數量，則訓練模型過程極有可能發生過度擬合的問題。模型因為過度擬合造成實際預測能力下降，仍有可能獲得過高的決定係數，導致最終錯誤判讀模型表現。

3. 誤差指標

相關係數與決定係數對於評估模型表現有其侷限性，因此多數學者改採誤差指標評估 PTF 的轉換表現。過去學者建立 PTF 所使用的誤差指標包含，平均誤差(mean error, ME)(Nemes *et al.*, 2003)、平均絕對誤差(mean absolute error, MAE)(Weynants *et al.*, 2009)、均方根誤差(root mean square error, RMSE)(Pachepsky *et al.*, 1998; Patil *et al.*, 2012; Van Alphen *et al.*, 2001)，其計算方法如式(2.24)~(2.26)所示。其中又以 RMSE 作為評估指標最為合適。因為其計算實際值和預測值之間差異的平方，賦予離群值對於 RMSE 計算結果更高的權重。有助於模型改善預測值整體之離散程度，也因此 RMSE 為最多數研究者所使用的評估指標(Patil & Singh, 2016)。

$$ME = \frac{\sum_i^N (y_i - \hat{y}_i)}{N} \quad (2.24)$$

$$MAE = \frac{\sum_i^N |y_i - \hat{y}_i|}{N} \quad (2.25)$$

$$RMSE = \sqrt{\frac{\sum_i^N (y_i - \hat{y}_i)^2}{N}} = \sqrt{\frac{SS_{res}}{N}} \quad (2.26)$$

2.2.5 不確定性分析

土壤轉換方程式預測結果的不確定性可能來自模型偏差、參數不確定性和輸入變數的測量誤差(Minasny & Mc Bratney, 2002)。Vereecken *et al.* (1992)即評估土壤轉換方程式對於土壤水力參數的不確定性，並衡量不確定性對於後續其他模型

模擬結果的影響。Schaap & Leij (1998)根據原始土壤資料的準確性，考量依據土壤數據庫所建立的土壤轉換方程式不確定性。Christiaens & Feyen (2001)則分析不同土壤轉換方程式的不確定性及其後續在水文模型中的應用結果。上述學者皆考慮由模型誤差或參數不確定性所引起的不確定性。Minasny *et al.* (1999)即指出模型參數不確定性遠較輸入變數的誤差小，且發現輸入變數的誤差，將直接影響土壤轉換方程式最終的預測結果。

四種不同的統計理論可應用於不確定性分析，包括：Delta method、貝氏定理、mean-variance estimation 以及自助法(Khosravi *et al.*, 2011)。其中 Delta method、貝氏定理需假設誤差為常態分佈，且分析過程的計算量龐大也為其一大缺點(需計算複雜的多階導數或黑塞矩陣[Hessian matrix])。mean-variance estimation 則透過計算樣本平均值、標準差，並假設誤差為常態分佈進行計算。分析過程簡單且計算快速為其優點，然而若模型預測值與真值具有較大的偏差，將會嚴重錯估模型不確定性為其最大缺點(Dybowski & Roberts, 2001)。自助法則不需事先假設誤差為任一分佈，可以更加準確地考慮真實誤差為偏態分佈的情況。相較為其他三個方法，雖然自助法分析過程的計算量為四個方法中最為龐大的，但也因自助法的分析過程相對容易，因此為四個方法中最廣為使用的不確定性分析方法。本研究即使用自助法進行不確定性分析，以下將針對自助法進行詳細說明。

自助法(Bootstrap method)是一種從樣本中進行均勻地重複抽樣，以形成新的自助分佈(Bootstrap distribution)，並以此分佈評估樣本的種種資料特性，如圖 2.8 所示。詳細流敘述程如下：

1. 從原先樣本中以取後放回的方式，隨機抽樣產生一次新的自助樣本(樣本大小與原先樣本相同)。
2. 計算自助樣本的各式統計量。
3. 重複 1、2 過程多次，並以自助樣本的統計分佈估計樣本的統計分佈。

自助法主要由兩大原理構成，1.插入原則(plug-in principle)、2.蒙地卡羅方法(Monte Carlo method)。重複抽取樣本的过程需符合原樣本分佈情形，此即蒙地卡羅方法的應用。以自助分佈評估樣本的資料特性，此即插入原則。以自助法推估樣本統計分佈或區間估計具有數個優點，包括：

1. 不需假設樣本來自何種母體分佈。

2. 不需考慮樣本分佈之偏度(skewness)是否為 0。
3. 對於偏度為 0 之雙峰或三峰值樣本仍可使用。
4. 樣本數量較大時，以自助法推估之信賴區間將優於傳統以常態或 t 分佈推估之結果(Hesterberg, 2011)。
5. 不必重複收集樣本即可有效地評估樣本背後可能之母體分佈形狀(shape)、峰度(kurtosis)、偏度等。

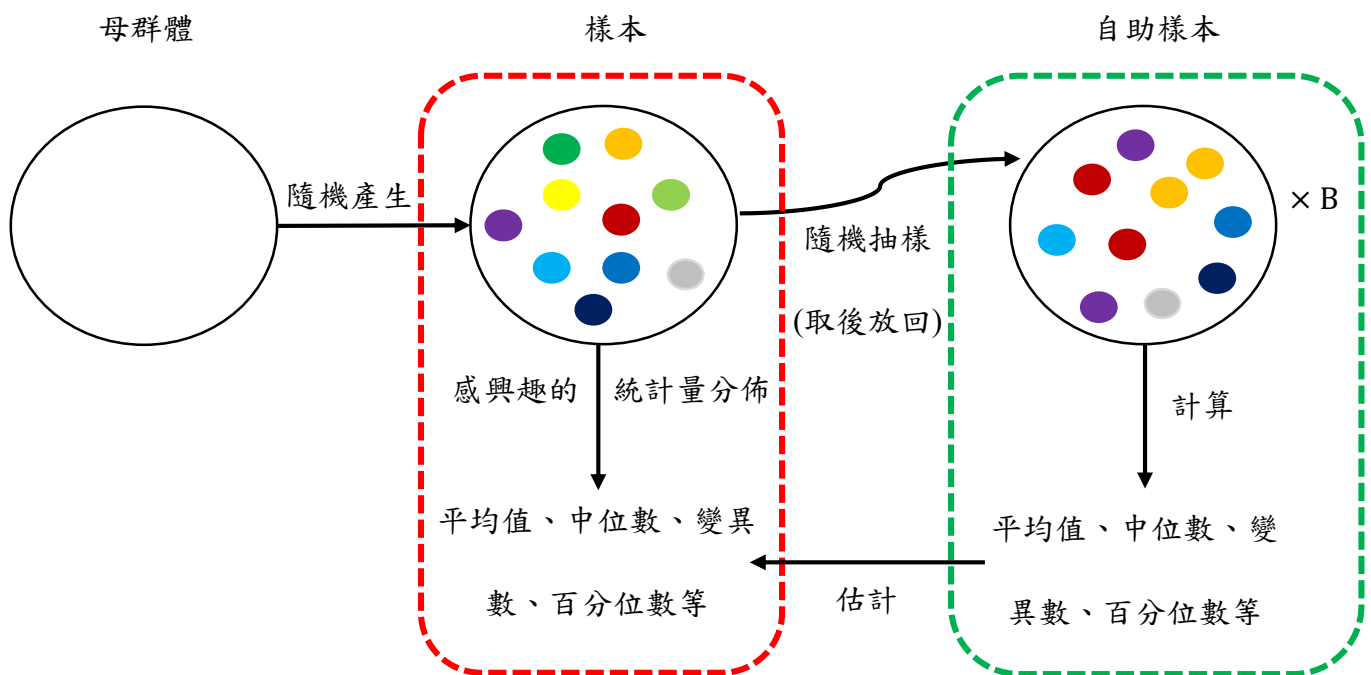


圖 2.8 自助法(Bootstrap)示意圖(Efron & Tibshirani, 1994)

2.2.6 Rosetta

Rosetta1(Schaap *et al.*, 2001)為美國農業部根據北美、歐洲土壤以類神經網路方式所建置之 PTF。也因其內置於 HYDRUS 數值模式內，故為目前廣受應用的 PTF。Rosetta1 為一種連續型 PTF，其轉換結果為 VG model 的 4 個參數與飽和水力傳導係數。Rosetta1 為一分層 PTF(共 5 層)可供使用者根據自身需求，選擇合適的解釋變數輸入並預測土壤水力參數。Rosetta1 尚根據自助法(60 或 100 個自助樣本)求取土壤水力參數預測值的標準差，並以此為不確定性的分析指標。之後 Schaap *et al.* (2004)、Zhang & Schaap (2017)等學者，將 Rosetta1 模式部分參數與架構修改後，提出新的 Rosetta 版本：Rosetta2、3。其中以 Zhang & Schaap (2017)提出的 Rosetta3

為最新且預測效果最佳的 Rosetta 版本。Zhang & Schaap (2017)提出一組改進的分層 PTF，將土壤保水曲線和飽和水力傳導係數兩個子模型統整為一個模型。並以 Rosetta1 預測結果的不確定性，作為新的訓練資料並對整個 Rosetta3 模型重新進行校正。Rosetta3 一樣採用自助法對 PTF 預測結果進行不確定性分析，且輸出結果同為土壤水力參數的算術平均值與標準差。與 Rosetta1 不同的是，Rosetta3 的自助樣本提升為 1000 個，此舉可有助於顯著降低自助法本身的變異性。Zhang & Schaap (2017)尚根據 Rosetta3 的預測結果，指出土壤水力參數並非常態分佈，而是屬於重尾 α 穩定分佈 (heavy-tailed α -stable distributions)。然而算術平均值對大多數預測結果均只有微小的估計偏差。因此，Zhang & Schaap (2017)仍然建議使用算術平均值和(協)方差作為估計分佈的統計量。

第3章 研究方法



3.1 研究架構及流程

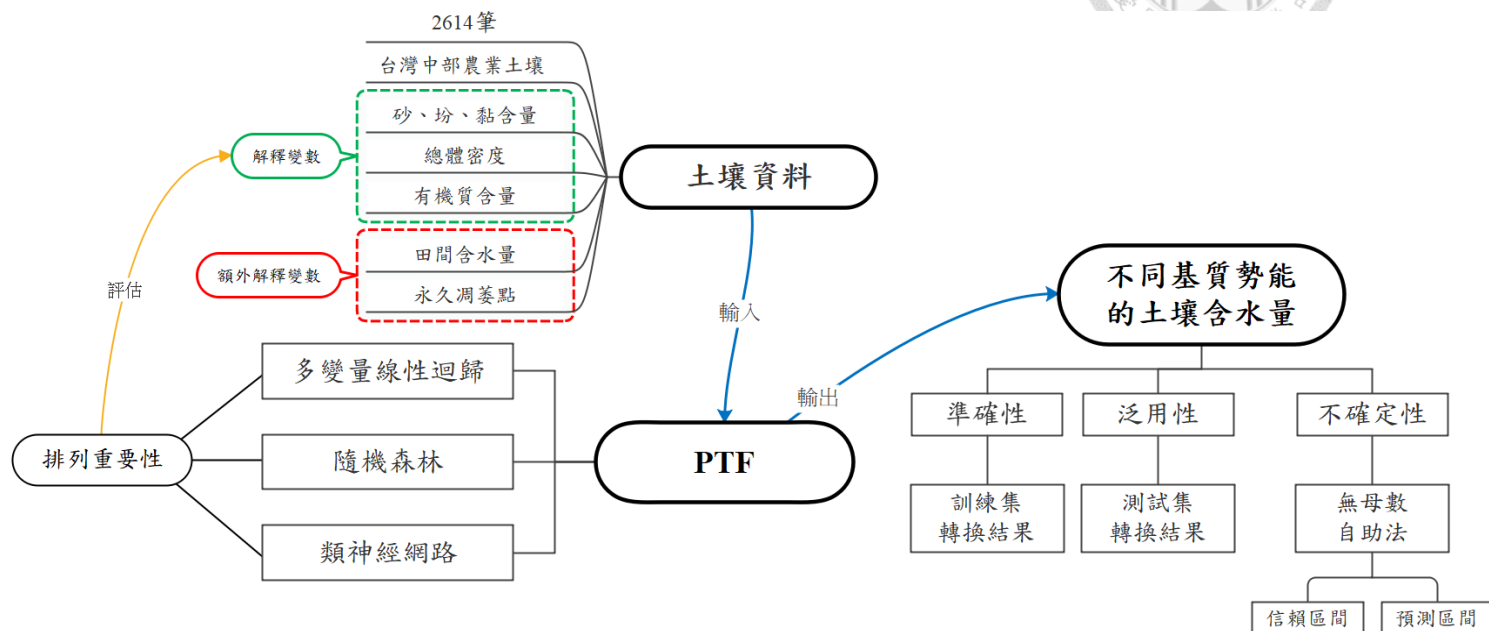


圖 3.1 研究架構圖

本研究的研究架構如圖 3.1 所示。本研究欲開發台灣本土的 PTF，選定適當且容易取得的土壤物理、化學特性作為解釋變數，以預測不同基質勢能下的土壤含水量。回顧過去數十年關於 PTF 解釋變數選定的國外文獻(Clapp & Hornberger, 1978; Gupta & Larson, 1979; Patil & Singh, 2016; Rajkai & Várallyay, 1992; Rawls *et al.*, 1982; Wösten *et al.*, 2001)，並綜合考量台灣土壤資料庫各項屬性齊全程度，最終選定模型的解釋變數為土壤質地之砂粉黏比例(Sand、Silt、Clay，以下簡稱 SA、SI、CL)、總體密度(Bulk Density，以下簡稱 BD)、有機質含量(Organic matter，以下簡稱 OM)。土壤含水量預測模型則使用目前國外文獻最常使用的兩種模型：MLR、ANN(Abdelbaki, 2020; Baker & Ellison, 2008; Costa *et al.*, 2013; Dashtaki *et al.*, 2010; Pachepsky *et al.*, 1996; Rawls *et al.*, 1982; Saxton & Rawls, 2006; Schaap *et al.*, 2001)。另外考量到 RF 的優點與應用性(Breiman, 2001; Ho, 1995)，本研究也將額外採取此方式作為預測模型，總共分成三種模型。MLR、RF 以及 ANN 之優缺點與研究建模的詳細流程，將分述於小節 3.2.1、3.2.2、3.2.3。

本研究所使用的程式語言為 R 以及 python，進行包括資料前處理、模型建置、特徵重要性與後續誤差分析。分析細節將依序詳述於 3.1.2 小節、3.2、3.3、3.4 節，本研究之完整流程圖如圖 3.2 所示。

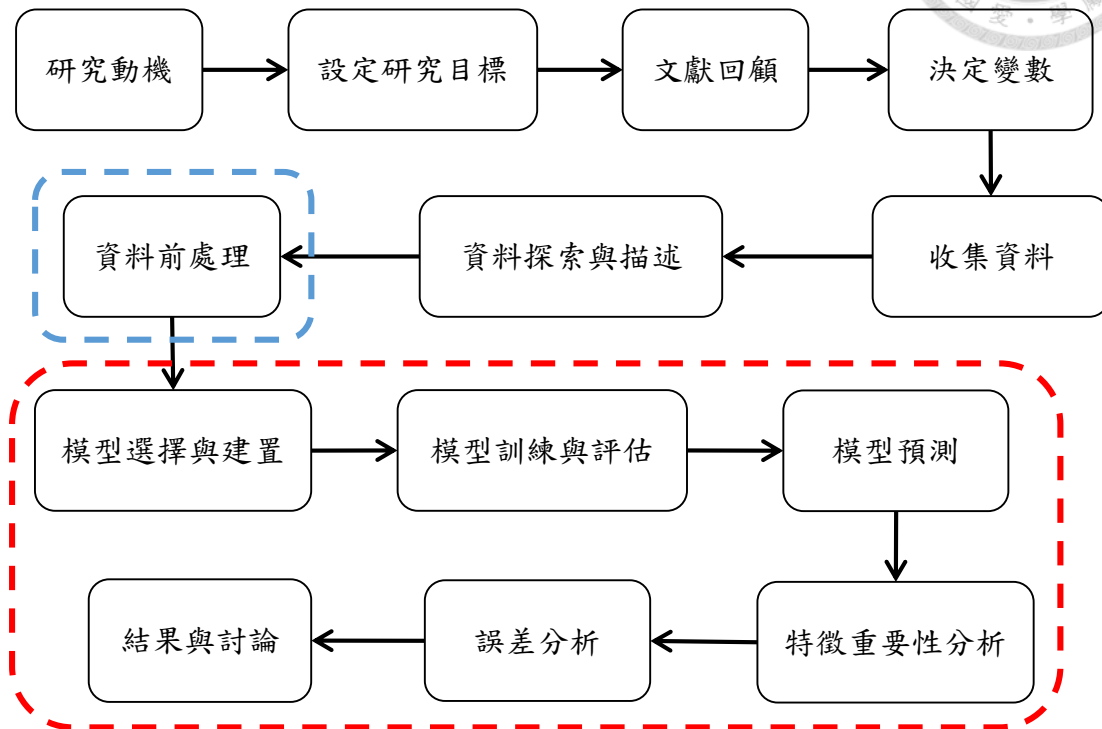


圖 3.2 研究流程圖。藍色與紅色虛線框詳細內容分別補充在圖 3.5、圖 3.6。

3.1.1 土壤資料

本研究所使用的資料為民國 101、102 年台灣中部土壤調查數據共 5187 筆，台灣中部土壤調查數據由行政院農業委員會農業試驗所(以下簡稱為農試所)提供，部分資料呈現如表 3.1，更多土樣資料處理的細節將在 3.1.2 小節討論。

表 3.1 台灣中部土樣 5187 筆之前 10 筆資料表

樣本編號	TWD97 X 座標	TWD97 Y 座標	砂粒(%)	粉粒(%)	黏粒(%)	質地	BD(g/cm ³)	OM(%)
01S2GK04011	208985	2680493	41.25	36.09	22.66	Loam	1.75	2.86
01S2GK04051	209681	2681759	39.58	37.11	23.30	Loam	1.69	4.13
01S2GK05031	210328	2680671	42.00	39.37	18.63	Loam	1.50	1.82
01S2GK05051	211231	2681447	36.88	39.50	23.62	Loam	1.56	1.88
01S2GK14061	208687	2677527	46.30	39.05	14.65	Loam	1.48	1.88
01S2GK14081	209189	2679723	40.07	38.02	21.91	Loam	1.73	2.13
01S2GK20081	223167	2677346	33.39	43.18	23.43	Loam	1.58	3.47
01S2GK23041	205914	2674917	28.64	43.85	27.51	Clay Loam	1.65	2.36
01S2GK23081	205793	2675495	32.91	41.20	25.90	Loam	1.43	2.13
01S2GK24021	208091	2677033	44.48	34.69	20.83	Loam	1.67	2.51

台灣中部土壤調查數據共 5187 筆，可根據採集土樣的方式不同，再細分為土塊法採樣的資料共 499 筆、土扣法採樣的資料共 885 筆以及未分類資料共 3803 筆。土塊法為在現地以採樣鏟採取大於所需土壤的完整土塊，其後將土塊以 1:4 的 SARAN 樹脂包覆使其形狀固定；土扣法則使用劈管採樣器以敲擊的方式扣入土壤中，待扣入深度達指定值時再取出整個採樣器。土塊法採集之土壤樣本較接近現地情形，然而採集與後續包覆過程繁複且不易操作為其主要缺點。土扣法則因敲擊過程產生高度擾動，土壤樣本代表性降低，然而具有採集快速、操作容易等優勢。

本研究中所採用的土壤資料包括土壤保水曲線(反應變數)，以及土壤質地、有機質含量、統體密度(解釋變數)。其中土壤保水曲線實驗數據來自砂箱及壓力鍋實驗。砂箱(Sand/kaolin box for pF determination 08.02.SA)由 Royal Eijkelpkamp 公司生產(圖 3.3(a))。量測原理為隨著基質勢能的減少，砂箱內的土壤含水量逐漸減少，最終藉由對土壤秤重換算得出各基質勢能下的土壤含水量。此方法適用於基質勢能較大之情況(通常為 $\geq -1\text{bar}$)，較小基質勢能之土壤含水量量測則須依靠壓力鍋實驗；壓力鍋(15 BAR CERAMIC PLATE EXTRACTOR)由 SOILMOISTURE 公司生產(圖 3.3(b))。量測原理為外加壓力對土壤樣本進行加壓並持續一或數天，期間等待土壤樣本排水完畢至鍋內系統平衡，再將對土壤樣本取出並進行稱重進而換算得出土壤含水量。

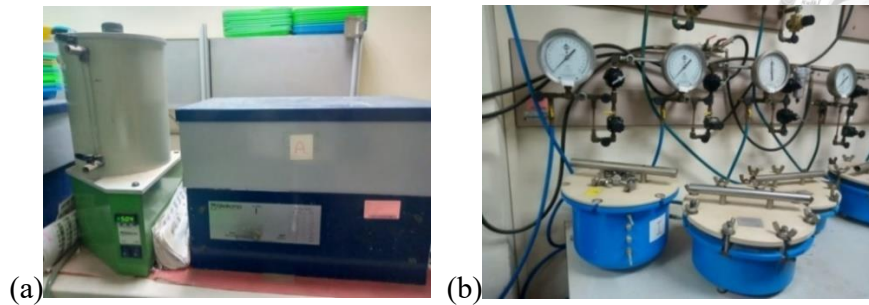


圖 3.3 農試所得出土壤保水曲線之實驗儀器。(a)砂箱、(b)壓力鍋。

土壤質地(粒徑分析)的實驗數據則來自以下分析(Kilmer & Alexander, 1949)：

1. 配製所需試劑：

- (1). Hydrogen peroxide (雙氧水) H_2O_2 (30~35%)
- (2). Sodium citrate solution (檸檬酸鈉) 0.3M $\text{Na}_3\text{C}_6\text{H}_5\text{O}_7 \cdot 2\text{H}_2\text{O}$ (88.4 g/L)
- (3). Sodium bicarbonate buffer solution (碳酸氫鈉) 1M NaHCO_3 (84g/L)
- (4). Sodium dithionite (連二亞硫酸鈉) $\text{Na}_2\text{S}_2\text{O}_4$ -hydrosulphite
- (5). Saturated NaCl solution (飽和食鹽水) 溶解度 360g/L(20°C)
- (6). Sodium hexametaphosphate HMP $((\text{NaPO}_3)_6)$ (六偏磷酸鈉)50g/L

2. 實驗步驟

- (1). 秤取土壤樣本 10g。
- (2). 加入 10ml 去離子水與 5ml H_2O_2 移除有機質。
- (3). 加入 40ml Sodium citrate solution 與 5ml Sodium bicarbonate buffer solution 以移除鐵氧化物。
- (4). 土壤樣本加入 200ml 去離子水並重複過濾直至含鐵氧化物之溶液被清洗乾淨。
- (5). 烘乾土壤樣本。
- (6). 以 10ml HMP 充分濕潤土壤後再加入 175ml 去離子水使土壤分散。
- (7). 將分散完畢之土壤溶液過 0.048mm 篩網，並將通過篩網的土壤溶液以雷射粒徑分析儀進行分析。

另外，將裝有砂粒的 0.048mm 篩網及底座放置烘箱烘乾。

(8). 將烘乾後的砂粒以 1mm、0.495mm、0.25mm、0.1mm、0.048mm 篩網放在搖篩機上均勻過篩後，秤各篩網內砂重。

(9). 將雷射粒徑分析儀的圖譜依照下列粒徑範圍個別積分運算：
Clay：0~2um、Fine silt：2~20um、Coarse silt：20~50um、Fine sand：50~200um、Coarse sand：200~2000um

統計出每個樣本各項比例佔比，取樣本兩重複平均值，再配合烘乾土重、篩砂重等秤重數據，計算出樣本的重量及百分比。

最後則分析具有高度保水性的土壤有機質含量，分析方法採用乾灰化法 (Davies, 1974)。

1. 實驗步驟：

(1). 秤取坩鍋淨重(W_1)，再秤取 2.0 g 已風乾過篩土壤(< 2 mm)樣本於坩鍋中秤重(W_2)，最後則放入烘箱以 105°C 烘乾。

(2). 待烘箱降溫後再取出土壤樣本放置於乾燥皿中，待其完全冷卻後再依序秤重(W_3)。

(3). 放入烘箱以 400°C 烘乾。待烘箱降溫後再取出土壤樣本放置於乾燥皿中，待其完全冷卻後再依序秤重 (W_4)，並依式(4.)計算有機質含量。

$$(4). OM(\%) = \left[\frac{(W_3 - W_4)}{W_3 - W_1} \right] \times 100\%$$

農試所提供之土壤數據內容包含 1997 台灣大地基準經緯座標(TWD97 XY)、粒徑分析、砂坩黏粒百分比、總體密度、孔隙率、土壤保水曲線對應含水量實驗數據、飽和水力傳導係數、陽離子交換容量、有機質含量等。若再根據土壤資料屬性的完整程度，可另外分為完整資料共 1384 筆與不完整資料 3803 筆，所有土壤調查地點與數量如圖 3.4 所示。

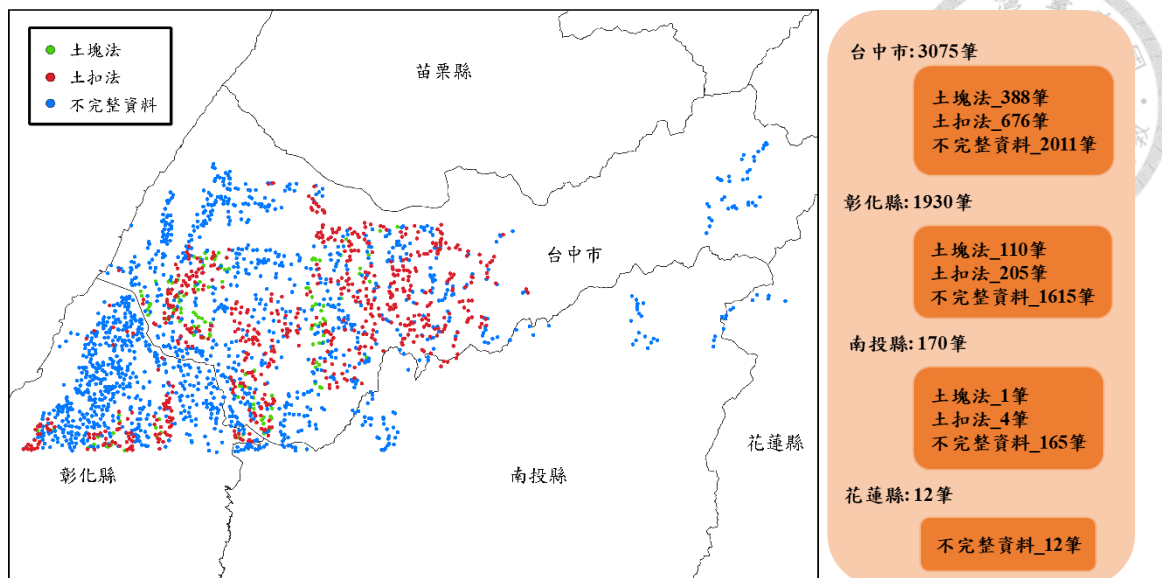


圖 3.4 農試所民國 101、102 年所有土壤調查地點及數量。綠點為土塊法，紅點表示為土扣法，藍點為不完整資料採樣之資料。

完整資料的所有屬性皆有分析結果，不完整資料則是部分屬性有所缺值，以下列出不完整資料各屬性無缺值之數量，如表 3.2 所示。第一欄為不完整資料的總筆數，其他欄位分別表示該屬性有分析結果的筆數。

表 3.2 中部土壤不完整資料於各屬性中無缺值數量表

總數量	團粒穩定度	粒徑分析	總體密度 (土塊法)	總體密度 (土扣法)	飽和水力傳導係數	土壤保水曲線	陽離子交換容量	有機質含量	pH 值
3803	1700	2927	1872	548	1143	2766	3116	3355	3353

3.1.2 土壤數據整理

台灣中部土樣資料項目完整，僅少數採樣地點如：高山地區，無法完整取得土壤樣本。其中以土塊法與土扣法採集之資料最為完整。本研究欲使用之每一筆土壤資料屬性，包含解釋變數(SA、SI、CL、BD、OM)、反應變數(土壤含水量)，皆不能有所缺值。其中反應變數共包含 8 個不同基質勢能(0、-0.05、-0.1、-0.33、-0.5、-1、-2 以及 -15bar)的土壤含水量，以下依序表示為 Vw、VwPF17、VwPF20、VwPF253、VwPF27、Vw1bar、Vw2bar、Vw15bar。資料前處理的流程圖如圖 3.5 所示。

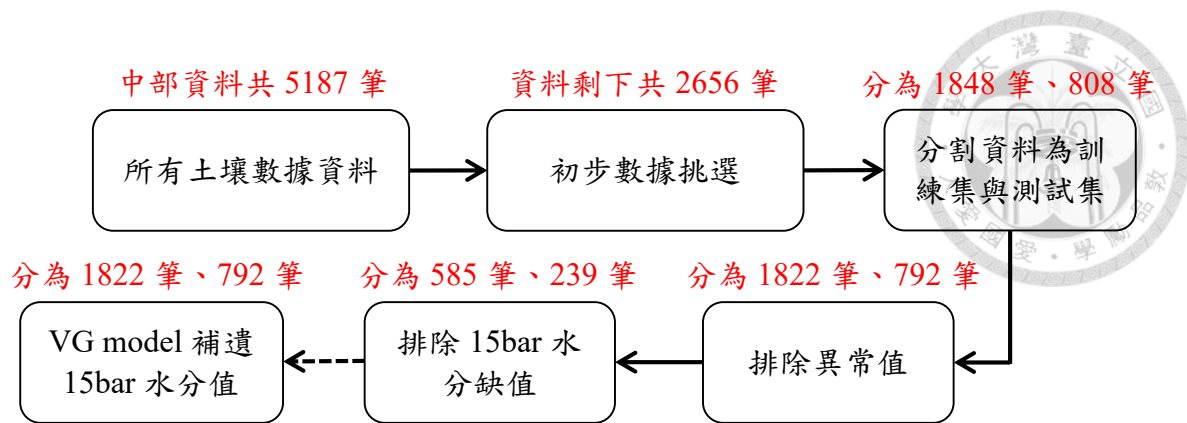


圖 3.5 資料前處理流程圖。虛線箭頭表示將前一步驟已去除之土壤數據，重新以 VG model 進行補遺，細節將於 3.1.3 小節闡述。

資料前處理的詳細步驟如下：

1. 使用資料為農試所提供之台灣中部土壤共 5187 筆。
2. 依據解釋變數的選擇，選擇欲使用的土壤特性(SA、SI、CL、BD、OM)，並確保每個欄位皆無缺失。
3. 將完整的資料集以隨機函數進行切割，分為訓練集 70%(1848 筆)、測試集 30%(808 筆)。為確保切割前後的樣本統計特徵不變(或相近)，控制隨機函數的亂數種子產生過程如下：
 - (1). 以任一亂數種子決定隨機函數並將資料集分割。
 - (2). 以各基質勢能的土壤含水量，計算兩資料集的樣本分位數(包含 10%、20%……90%)。
 - (3). 計算兩資料集樣本分位數之 RMSE。
 - (4). 記錄本次 RMSE，並改變亂數種子重新分割資料集(重複前述 3 個步驟)。
 - (5). 反覆迭代 100 次並持續記錄每次的 RMSE。
 - (6). 以 RMSE 最小的一次作為數據切割使用的亂數種子。
4. 異常值排除主要處理總體密度(bulk density, BD)與有機質含量(organic matter, OM)，過程如下：
 - (1). 刪除訓練、測試資料集中 OM 為負值的異常土壤。
 - (2). 刪除訓練、測試資料集中 $OM \geq 35\%$ 的土壤。
 - (3). 刪除訓練、測試資料集中 $BD < 1$ 的土壤。



OM 與 BD 的刪除閾值選定方式則分別敘述如下：

- A. 對於大部分高山土壤，OM 約為表土總重量的 1%~6%，含有 12%~18%OM 的土壤則常被歸類為有機土壤(Troeh & Thompson, 2005)。OM 超過 35%的土壤則會被歸類為淤泥土(muck soil)或泥炭土(peat soil)(F.E. Allison, 1973)。淤泥土與泥炭土和一般礦物風化所形成之土壤有極大的不同，本研究主要研究對象為礦物所形成之土壤，因此刪除 OM 大於 35%以排除淤泥土或泥炭土的土壤資料。
- B. McCarty *et al.* (2016)、Rowell (1994)根據不同區域的土壤，整理土壤在不同質地地下所對應的顆粒密度、總體密度及孔隙率範圍如表 3.3。從表 3.3 中可以看到，一般礦物形成之土壤以及耕作土壤其總體密度最小值約為 1g/cm^3 。本研究主要研究對象即為礦物所形成之土壤，因此刪除訓練、測試資料中 $\text{BD} < 1$ 的土壤。

表 3.3 土壤質地與顆粒密度、總體密度及孔隙率之關係表(McCarty *et al.*, 2016; Rowell, 1994)

Soil Type	Particle density (g/cm^3)	Bulk density (g/cm^3)	Porosity (cm^3/cm^3)
Mineral soils, plowed horizons			
fine-medium texture	2.60	1.0-1.6	0.69-0.46
coarse texture	2.60	1.4-1.8	0.46-0.35
Subsoil and parent materials	2.65	1.5-1.8	0.43-0.32
Grassland and woodland, surface horizon	2.4	0.8-1.2	0.67-0.50
Peats	1.4	0.1-0.3	0.93-0.79

從圖 3.5 中可以看到，若排除反應變數(V_w 、 V_{wPF17} 、...、 V_{w2bar} 、 V_{w15bar})無值之情況，將會使可用資料大幅縮減。其中缺值最為嚴重的為 V_{w15bar} ，為了最大化利用土壤資料，將會以 VG model 外插 V_{w15bar} 的方式進行補值，詳細內容將會在 3.1.3 小節中進行討論。

3.1.3 Vw15bar 補值

Vw15bar 為基質勢能-15bar 時所測得之土壤水分含量，又可稱為永久凋萎點，為植物能夠從土壤中獲取水分而不枯萎之最小土壤含水量(Cassel & Nielsen, 1986)。且 Vw15bar 對於土壤保水曲線、非飽和水力傳導係數也有密切的關係(Ghanbarian-Alavijeh & Millán, 2009)，因此評估 Vw15bar 對於後續水文與農業方面應用極其重要。然而透過圖 3.5 可以發現，若排除 Vw15bar 缺值則資料量將從 1822+792 筆銳減為 585+239 筆，顯示出 Vw15bar 雖具有高度應用性，但可用資料卻高度稀缺的窘境。此外，資料數量的減少也對於後續建模有顯著的影響，將使模型泛用性降低。土壤資料庫中不乏許多難以取得的數據，需要經過曠日廢時的實驗，若僅因為缺少一欄而捨棄其他大量土壤資料，對於台灣土壤資料庫的應用實屬可惜與浪費。

因此本研究即利用 van Genuchten (1980)所提出的 VG model 作為土壤保水曲線的連續函數，如式(3.1)所示。

$$\theta(h) = \theta_r + \frac{(\theta_s - \theta_r)}{[1 + (\alpha h)^n]^m}, \quad m = 1 - \frac{1}{n} \quad (3.1)$$

$\theta(h)$ 為基質勢能在 $-h$ cm-water 時的土壤含水量。 θ_s 、 θ_r 分別為飽和含水量與殘餘含水量，單位為 cm^3/cm^3 ， α 為模型參數並與空氣進入壓力值之倒數有關，單位為 $1/\text{cm}^3$ ， n 為孔徑分佈(pore-size distribution)的量度，單位為無因次。

藉由數個已知土壤含水量以曲線擬合的方式，求解 VG model 的 4 個參數(θ_s 、 θ_r 、 α 、 n)。再以擬合參數與欲補值之基質勢能壓力一同代入式(3.1)，即可求得 VG model 所估計之土壤含水量。以下為使用 VG model 進行 Vw15bar 補值之具體操作方法：

1. 設定 VG-model 參數擬合上下限值

$$\{\theta_s | 0 < \theta_s < 1\}、\{\theta_r | 0 < \theta_r < 1\}、\{\alpha | \alpha > 0\}、\{n | n > 0\}$$

2. 以原始資料擬合得到 VG-model 參數。
3. 以得到的 VG-model 參數外插 Vw15bar。

上述方法雖可有效進行 Vw15bar 補值，但因擬合時使用的原始資料絕大部分僅為基質勢能 $\geq -2\text{bar}$ 的土壤含水量，殘餘含水量(θ_r)擬合結果幾乎皆為 0 值，較不符合真實土壤情況。本研究為避免此一情況產生，欲以 Vw nbar 取代殘餘含水量(θ_r)，



其中 $n \geq 15$ 且為一個常數。 n 值可透過以下分析而定。

- i. 每一筆資料皆搜尋一個 n 值範圍滿足(3.2)不等式。

$$\frac{2}{10} V_{w,15bar} \leq V_{w,nbar} < \frac{8}{10} V_{w,15bar} \quad (3.2)$$

$V_{w,15bar}$ 為每一筆資料的原始觀測含水量或經過 VG-model 補值之含水量。

- ii. 尋找一個 n 值，使全部資料的 8 成可以滿足(3.2)式。
iii. 若滿足 2.條件的 n 值為一個區間，則選定區間最大值作為最終 n 值。

(3.2)不等式主要界定殘餘含水量(θ_r)與 V_{w15bar} 關係，上下界係數則藉由分析無缺值資料殘餘含水量(θ_r)與 V_{w15bar} 關係，並以試誤法經過多次嘗試而定。

最後承繼前一部分 $V_{w\ nbar}$ 分析結果，進行 VG-model 參數擬合改善如下：

4. 以相同初始條件(除了 θ_r 下限改為 $V_{w\ nbar}$)進行第二次擬合(重複 1~3 步驟)。
5. 若第一次擬合之 $\theta_r \neq 0$ 則採用第一次參數擬合結果外插 V_{w15bar} 。
6. 若第一次擬合之 $\theta_r = 0$ 則採用第二次參數擬合結果外插 V_{w15bar} 。

經過此一系列的分析，一旦 $V_{w\ nbar}$ 決定後即可完全避免最終擬合殘餘含水量(θ_r)結果為 0。最後以 VG-model 進行 V_{w15bar} 的補值，往後的符號表示將以 V_{w15bar_VG} 代表經過 VG-model 外插補值而得之 V_{w15bar} 。

3.2 模型選擇與建置

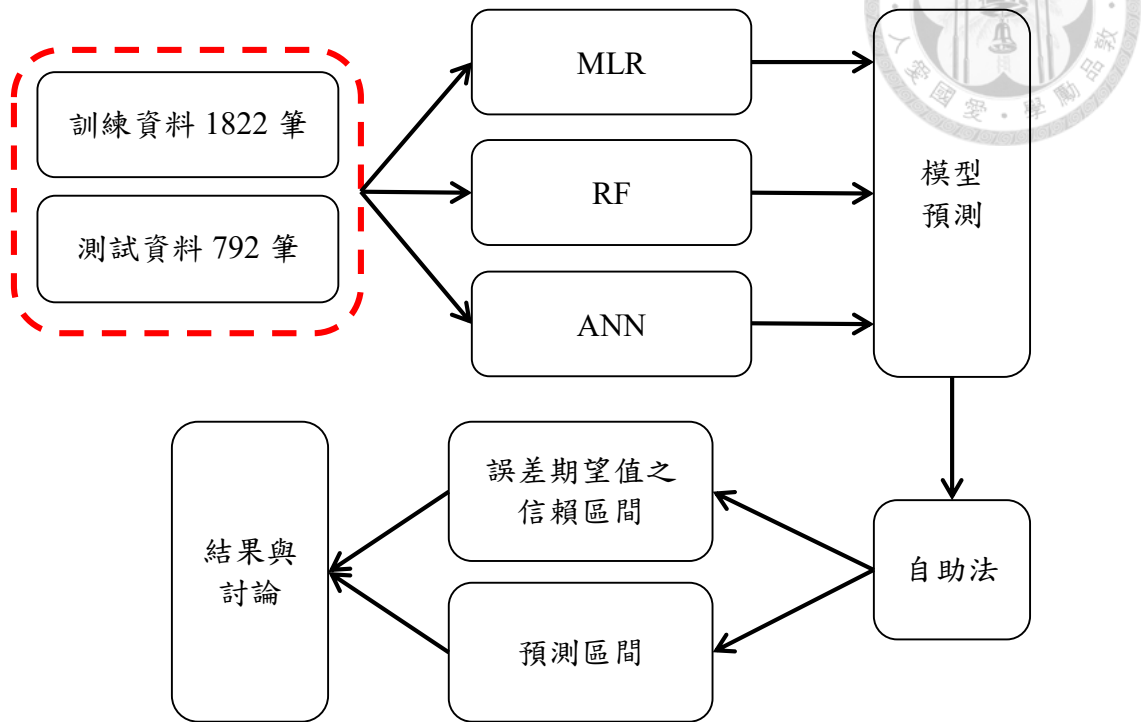


圖 3.6 模型建置與後續分析流程圖。MLR、RF、ANN 的內容將於 3.2.1、3.2.2、3.2.3 小節詳細陳述。

本研究中共有三個不同模型，皆使用相同的解釋、反應變數。選擇 RMSE 作為模型表現評估指標，衡量模型預測值與真實值之間的誤差，並藉由 RMSE 數值大小比較三種模型間的預測表現。

3.2.1 多變量線性迴歸(Multiple Linear Regression，MLR)

線性迴歸為監督式學習的一種，為統計學上常見的一種模型，若解釋變數數量大於一個則為多變量線性迴歸(Allison, 1999; Berry *et al.*, 1985; Draper & Smith, 1998)，如式(3.3)。

$$y = C_0 + C_1x_1 + C_2x_2 + C_3x_3 + \cdots \quad (3.3)$$

其中 C_0 為迴歸方程式的截距， C_1 、 C_2 、 C_3 為迴歸方程式的係數，截距與係數可以透過最小平方法求得。

本研究中共有 8 個不同的基質勢能之土壤含水量，每個基質勢能之土壤含水量皆可建立一個多變量線性迴歸方程式(因此共有 8 個多變量線性迴歸方程式)，每個方程式中包含 5 個輸入變因(解釋變數)、1 個輸出結果(反應變數)，如式(3.4)。

$$\theta(\psi) = C_0 + C_1 \times SA + C_2 \times SI + C_3 \times CL + C_4 \times BD + C_5 \times OM \quad (3.4)$$

迴歸方程式中的每一個解釋變數皆需線性獨立，因此在求解係數前應先檢驗共線性問題。本研究透過計算變異數膨脹因子(Variance Inflation Factor, VIF)，檢核解釋變數間是否存在高度共線性(Akinwande *et al.*, 2015; Craney & Surles, 2002)，如式(3.5)。

$$VIF_i = \frac{1}{1 - R_i^2} \quad (3.5)$$

其中 VIF_i 為第 i 個解釋變數之變異數膨脹因子， R_i^2 為將第 i 個解釋變數視為應變數，其他變數當成解釋變數建立迴歸模型，所得之決定係數(Coefficient of determination)，其計算方式如式(2.23)。

本研究使用 R 語言中的 car 模組分析變異數膨脹因子以及共線性問題，而多變量線性迴歸方程式則應用 python 中 sklearn 的 LinearRegression 模組進行求解。

3.2.2 隨機森林(Random Forest, RF)

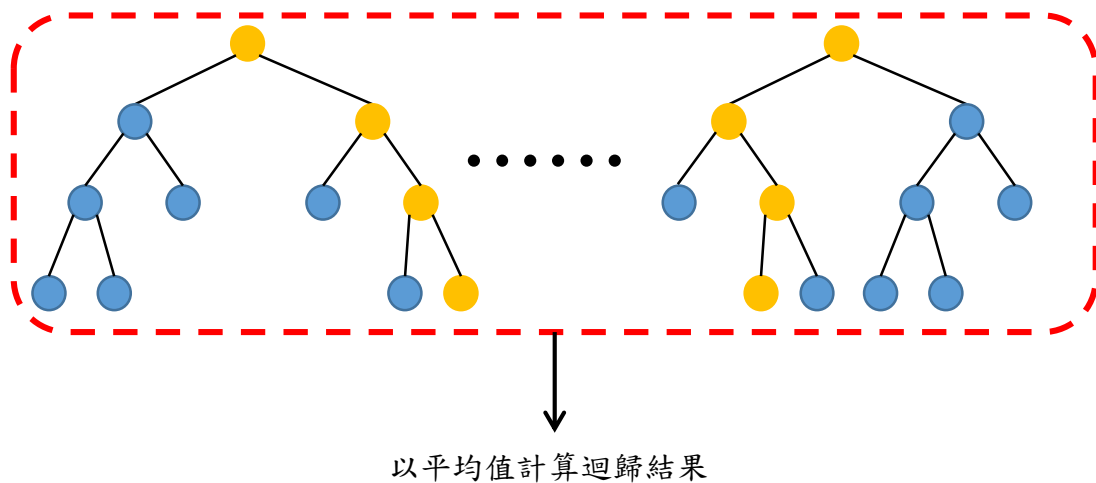


圖 3.7 隨機森林示意圖

隨機森林為監督式學習的一種，由多個決策樹(Decision tree)組合而成(Breiman, 2001; Ho, 1995)，示意圖如圖 3.7 所示。隨機森林相較於決策樹，具有兩個完全不同的特性：第一個為隨機取樣，訓練過程中每棵樹皆會從訓練集隨機重複抽取 n 筆資料(n 為訓練集資料筆數)作為新的訓練樣本。第二個為隨機挑選特徵，每棵樹皆會從資料中隨機挑選 $1 \sim k$ 個特徵作為解釋變數(k 為特徵數量，本研究設定每棵樹的挑選的特徵數量上限為 $\sqrt{k}$)。最終輸出結果時則運用機器學習中的整體學習或稱集成學習(ensemble learning)，透過多個學習器進而提升整體學習性能。

隨機森林主要藉由特徵重要性的選取進行模型擬合，因此訓練集數據不需進行標準化(即使標準化後其預測效力仍然相同，因標準化僅影響變數數值而不影響變數分佈，故不會改變特徵重要性)。輸入隨機森林模型的解釋變數不需考慮變數之間的共線性，因其擬合過程僅考慮最小化基尼雜質，若多個變數為完全線性相依，隨機森林將隨機選擇一項特徵對節點進行分支。但是若解釋變數有共線性問題時，線性相依的解釋變數其中一個基尼重要性(見 3.3.1 小節)將遠大於其他變數。

隨機森林的其他優點包括：

1. 可避免單一決策樹造成的過度擬合問題(over fitting)。
2. 可以評估變數對於預測結果的重要性。
3. 輸入資料較為多元時，仍然可以產生高準確度的分類器。
4. 可輸入並處理大量的變數。
5. 學習過程相較於其他機器學習方法快速。

隨機森林可分為迴歸與分類兩大類，而本次研究以迴歸作為主要用途。建構隨機森林模型使用的程式語言為 python，模組為 sklearn 中的 RandomForestRegressor。隨機森林的架構則以 sklearn 中的 RandomizedSearchCV，利用隨機網格的方式再給定的空間中劃分選取以進行超參數(Hyperparameter)最佳化。網格由超參數空間共四個條件組成並分述如下：

1. 森林中樹的數量(10~200)。
2. 每棵樹的最大深度(10~ n ， n 為所有葉子皆少於最小分割樣本數)。
3. 拆分內部節點所需的最小樣本數(2~20)。
4. 形成葉節點所需的最小樣本數(1~10)。

最後經過 100 次的迭代，且每次迭代都包含 3 次的交叉驗證，即可求出當前基質勢能下隨機森林模型的最佳架構，表 3.4 即列出以隨機森林建立之 8 個不同基質勢能的土壤含水量預測模型架構。



表 3.4 8 個不同基質勢能的土壤含水量所對應的隨機森林架構

	Vw	VwPF17	VwPF20	VwPF253	VwPF27	Vw1bar	Vw2bar	Vw15bar_VG
樹總量	160	170	170	80	170	120	200	120
拆分節點的最小樣本數	3	8	8	6	8	17	11	17
形成葉節點的最小樣本數	8	10	10	8	10	10	10	10
每棵樹的最大深度	10	10	10	10	10	70	100	70

3.2.3 類神經網路(Artificial Neural Network，ANN)

類神經網路為監督式學習的一種，透過神經元的排列組成 1 層(layer)，並層層疊加形成一個完整的網路(LeCun & Bengio, 1995; LeCun *et al.*, 1989)，單一神經元的示意圖如圖 3.8。類神經網路的架構由一層輸入層、一個或多個隱藏層以及一層輸出層組成。本次研究中所使用的為前饋式全連接神經網路(Feedforward fully connected neural network)，即每一層中的每個神經元皆與上下層的每個神經元連接，且參數從輸入層向輸出層單向傳播，示意圖如圖 3.9 所示。

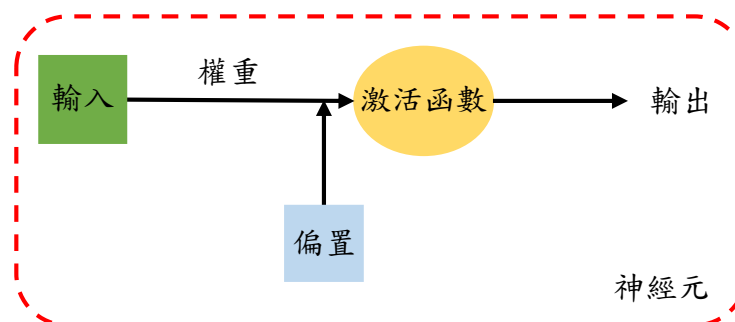


圖 3.8 單一神經元(Neuron)示意圖。類神經網路即透過調整每個神經元中的權重(weight)與偏差(bias)達成最佳化。

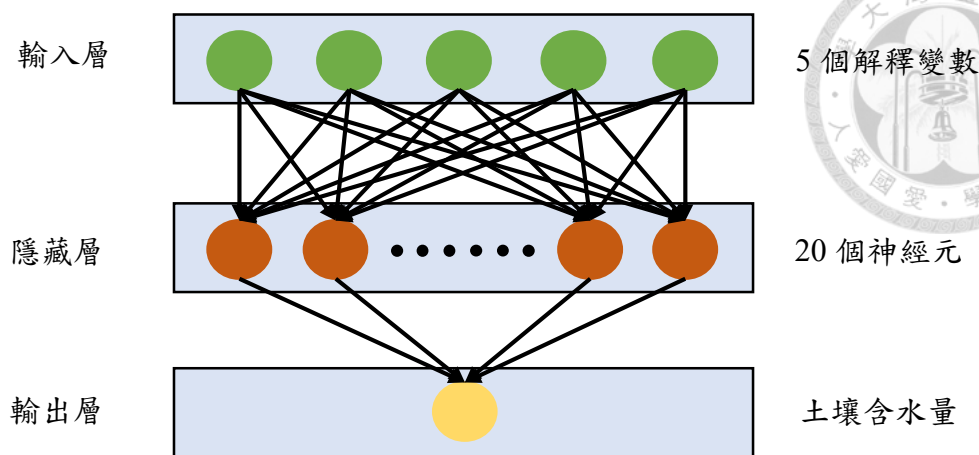


圖 3.9 類神經網路架構

本研究共輸入 5 個解釋變數，因此輸入層的維度為 5。隱藏層只有一層且包含 20 個神經元。輸出為土壤水分含量，因此輸出層的維度為 1。建構類神經模型所使用的程式語言為 python，使用 keras 中的 Sequential、Dense 模組。模型中超參數分述如下：

1. 隱藏層數為 1。
2. 隱藏層中的神經元數目為 20 個。
3. 使用的激活函數(activation function)為線性整流函數(Rectified Linear Unit，ReLU)。
4. 優化器(Optimizer)採用 Adam 進行。
5. 不採用丟棄法(Dropout)。

超參數透過試誤法而確立，試誤法過程為最佳化訓練集預測結果且必須不過度擬合(即受驗證集結果好壞控制)。是否過度擬合則透過繪製模型訓練曲線進行判定。



3.3 特徵重要性

3.3.1 基尼重要性(Gini importance)

若有一群資料其屬性類別共有 J 種，且欲透過決策樹輸入 n 個特徵 ($X_1, X_2, X_3, \dots, X_n$) 進行分類預測。則可藉由基尼雜質(Gini impurity)最小化的過程，擬合決策樹模型並有效完成分類預測，並依據不同特徵所減少的基尼雜質，進行特徵重要性評估。

決策樹中節點 t 的基尼雜質定義如(3.6)所示(Ishwaran, 2015)。

$$I(t) = \sum_{j=1}^J p_j(t) [1 - p_j(t)] = 1 - \sum_{j=1}^J p_j(t)^2 \quad (3.6)$$

$I(t)$ 為節點 t 的基尼雜質。 $p_j(t)$ 為節點 t 中屬於 j 的類別頻率(class frequency)。

若有兩個箱子內含 4 顆球，每顆球顏色可能為紅、綠、黃，則藉由(3.6)式可計算每個箱子的基尼雜質如圖 3.10 所示。

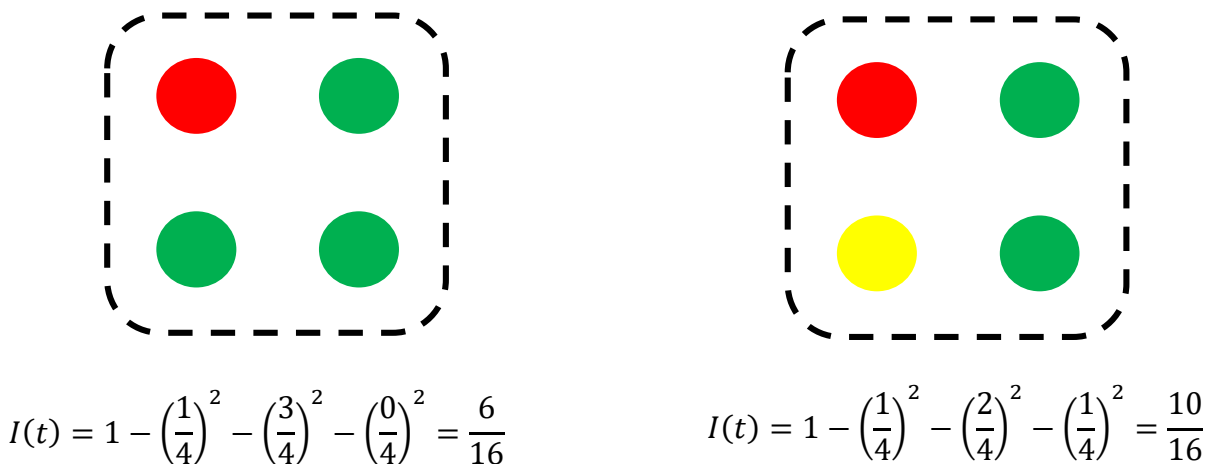


圖 3.10 基尼雜質計算示意圖

圖 3.10 中基尼雜質較大之箱子內含球的顏色分佈較為均勻(或稱較為混亂)，基尼雜質較小之箱子內含球的顏色分佈較為不均(或稱較為單一)。決策樹即是透過不斷降低基尼雜質進而將各屬性進行有效分類。決策樹中的節點每次進行分支時，所減少的基尼雜質(Gini impurity decrease)可由(3.7)式計算而得(Ishwaran, 2015)。

$$\Delta I(t) = \frac{N_t}{N} \times \left(I(t) - \frac{N_{tr}}{N_t} I(t_r) - \frac{N_{tl}}{N_t} I(t_l) \right) \quad (3.7)$$

$I(t_r)$ 為節點 t 向右分裂之子節點的基尼雜質。 $I(t_l)$ 為節點 t 向左分裂之子節點的基尼雜質。 N_t 為 t 節點樣本數量。 N 為總樣本數量。 N_{tr} 為節點 t 向右分裂的子節點所包含的樣本數量。 N_{tl} 為節點 t 向左分裂的子節點所包含的樣本數量。

決策樹中之每個特徵的重要性(variable importance)則可藉由式(3.8)計算而得 (Ishwaran, 2015)。

$$V(X_i) = \frac{\sum_{X_i} \Delta I(t)}{\sum \Delta I(t)} \quad (3.8)$$

$V(X_i)$ 為特徵 X_i 的重要性。 $\sum_{X_i} \Delta I(t)$ 為以變數 X_i 進行分裂之所有節點的基尼雜質減少量總和。 $\sum \Delta I(t)$ 為所有節點的基尼雜質減少量總和。

最後再將決策樹中之每個特徵的重要性，依照(3.9)式調整為 0~1 的數值。

$$G_{decision\ tree}(X_i) = \frac{V(X_i)}{\sum_{i=1}^n V(X_i)} \quad (3.9)$$

若一個隨機森林共包含 K 棵決策樹，則隨機森林之基尼重要性可透過累加每一棵決策樹的特徵重要性而得，計算方法如式(3.10)所示。將所有特徵的基尼重要性依照降序排列後，最高者即為該模型最重要的特徵。

$$G_{random\ forest}(X_i) = \frac{\sum_{k=1}^K G_{decision\ tree_k}(X_i)}{\sum_{k=1}^K \sum_{i=1}^n G_{decision\ tree_k}(X_i)} = \quad (3.10)$$

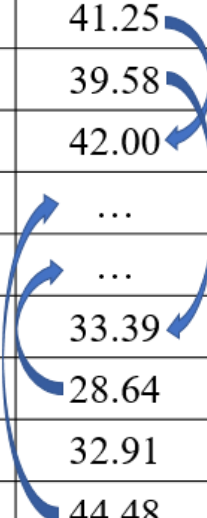
基尼重要性的優點為計算過程快速且可描述隨機森林透過訓練集擬合之特徵重要性。由於隨機森林擬合標準即是將模型基尼雜質最小化，隨機森林模型擬合完成時，輸入特徵的基尼重要性即可依照(3.6)~(3.10)式同步計算完成，不需再進行額外的運算。然而時常高估特徵重要性為基尼重要性的最大缺點，特別是隨機森林仰賴某一特徵進行預測時，該特徵的基尼重要性即會被放大許多。此外，因為隨機森林的基尼重要性依據訓練集資料進行計算，若隨機森林過擬合(over-fitting)而無法準確預測測試集資料的目標變量時，先前計算的基尼重要性極有可能產生偏誤

(Breiman, 2001)。

基於上述種種原因且基尼重要性僅可應用於決策樹模型，本研究僅將基尼重要性作為輔助指標，而將 3.3.2 小節的排列重要性作為主要指標。本研究計算基尼重要性的程式模組則與 3.2.2 小節隨機森林所使用的模組相同皆為 sklearn。

3.3.2 排列重要性

排列重要性(Permutation Importance)最早是由 Breiman (2001)提出並應用在隨機森林模型，用以衡量模型特徵的重要性。隨後 Fisher *et al.* (2018)將此方法拓展並應用到衡量任意模型特徵的重要性。排列重要性主要想法為，透過排列改變原始之特徵數值，擾動原始已擬合完成模型，並計算擾動後模型之預測誤差是否增加，以此衡量特徵的重要性。圖 3.11 即表示擾亂砂粒此特徵的示意圖。



編號	砂粒	校正坩粒	校正黏粒	BD	OM
1	41.25	36.09	22.66	1.75	2.86
2	39.58	37.11	23.30	1.69	4.13
3	42.00	39.37	18.63	1.50	1.82
...	...	...	...	...	...
...	...	...	...	...	...
1809	33.39	43.18	23.43	1.58	3.47
1810	28.64	43.85	27.51	1.65	2.36
1811	32.91	41.20	25.90	1.43	2.13
1812	44.48	34.69	20.83	1.67	2.51

圖 3.11 擾亂原始特徵順序之示意圖。BD、OM 分別代表總體密度與有機質含量。藍色箭頭表示一次完整的擾動，特徵所移動的方向。

每擾動一次可獲得新的擾動樣本，並可透過擾動樣本計算排列重要性(可重複擾動多次計算多次排列重要性)。若打亂其值而增加模型誤差，則該特徵便是較為重要的。因為在此種情況時，模型仰賴該特徵進行預測。若打亂其值但模型誤差增加幅度不大或保持不變，則該特徵便是較不重要的。因為在此種情況時，該特徵對

於預測值所佔的比重遠小於其他特徵。排列重要性分析流程與演算法如下(Fisher *et al.*, 2018)：

1. 輸入訓練完成的任意模型 $\hat{f}$ 、特徵矩陣 X 、預測目標真實值向量 y 以及誤差估計函數 $E(y, \hat{f})$ 。
2. 計算原始模型誤差 $e_{orig} = E(y, \hat{f}(X))$ ，例如模型均方根誤差 RMSE。
3. 對於每一個特徵 $j \in \{1, 2, 3, \dots, p\}$ 進行以下操作。
 - (1). 藉由排列數據 X 中的特徵 j 產生特徵矩陣 X_{perm} ，此時特徵 j 和預測目標真實值向量 y 之間的關係受到擾動而破壞。
 - (2). 計算擾動模型誤差 $e_{perm} = E(y, \hat{f}(X_{perm}))$ 。
 - (3). 計算擾動模型誤差和原始模型誤差之比值 $FI_j = \frac{e_{perm}}{e_{orig}}$ ，或兩者之差值 $FI_j = e_{perm} - e_{orig}$ ，以衡量該特徵之排列重要性。
4. 將所有特徵的排列重要性 FI 依照降序排列，最高者即為該模型最重要的特徵。

另外，需要特別注意的是，排列重要性並不反應該特徵本身與目標預測真實值的關係，而是反應該特徵對於特定模型預測目標值的重要性。本研究使用擾動模型誤差和原始模型誤差之差值計算排列重要性，並依序分析三種不同模型(MLR、RF、ANN)之 5 個解釋變數(SA、SI、CL、OM、BD)的排列重要性。

解釋變數的排列重要性計算，透過 python 程式語言中 sklearn 函式庫的 inspection.permutation_importance 模組進行實作。誤差函數採用均方根誤差 RMSE，且迭代計算 1000 次排列重要性，並繪出不同解釋變數的排列重要性分佈。



3.4 不確定性分析

3.4.1 誤差期望值

若假設不同基質勢能下的每個土壤含水量皆可用一個預測模型描述，其中包含土壤含水量的噪音項(noise)(Kumar & Srivistava, 2012)，如式(3.11)。

$$y(x) = \psi(x) + \epsilon(x) \quad (3.11)$$

x 為輸入的解釋變數。 $y(x)$ 為實際土壤含水量。 $\epsilon(x)$ 為土壤含水量的噪音項，滿足獨立同分佈且其期望值 0，標準差為一有限值。 $\psi(x)$ 為預測模型，為決定性函數(即相同 x 代入函數會得到完全相同的數值而無變異性)且滿足連續、 (p, C) 光滑函數條件。

(p, C) 光滑函數條件定義如下(Kohler *et al.*, 2009)：

$$\text{令 } p = k + \gamma, \quad k \in \mathbb{N}_0, \quad 0 < \gamma \leq 1, \quad \text{令 } C > 0$$

函數 $m: \mathbb{R}^d \rightarrow \mathbb{R}$ ，若對於所有 $k_1, \dots, k_d \in \mathbb{N}_0$ 且 $k = k_1 + \dots + k_d$ ， m 的偏導數式(3.12)存在且滿足式(3.13)則稱函數 m 為 (p, C) 光滑函數。

$$\frac{\partial^k m}{\partial x_1^{k_1} \dots \partial x_d^{k_d}} \quad (3.12)$$

$$\left| \frac{\partial^k m}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}(x) - \frac{\partial^k m}{\partial x_1^{k_1} \dots \partial x_d^{k_d}}(z) \right| \leq C \cdot \|x - z\|^\gamma \quad (x, z \in \mathbb{R}^d) \quad (3.13)$$

若未來有一新解釋變數值代入式(3.11)，並將式(3.11)中無變異性與變異性部分依序拆解如式(3.14)、(3.15)。再將式(3.14)、(3.15)相加即可得未來實際土壤含水量，如式(3.16)所示。

$$\psi(x_0) = \hat{y}_n(x_0) + \eta(x_0) \quad (3.14)$$

$$\epsilon(x_0) = \eta_n(x_0) + \varepsilon(x_0) \quad (3.15)$$

$$y(x_0) = \psi(x_0) + \epsilon(x_0) = \hat{y}_n(x_0) + \eta(x_0) + \eta_n(x_0) + \varepsilon(x_0) \quad (3.16)$$

x_0 為未來解釋變數值。 $\hat{y}_n(x_0)$ 為模型新預測值。 $\eta(x_0)$ 為模型新預測誤差。 $\eta_n(x_0)$ 為模型新預測值的變異性。 $\varepsilon(x_0)$ 為新觀測值的樣本變異性，假設 $\eta_n(x_0)$ 、 $\varepsilon(x_0)$ 兩者獨立。

在本研究中自助法即欲模擬 $\eta(x_0)$ 、 $\eta_n(x_0)$ 、 $\varepsilon(x_0)$ ，以求取模型信賴區間與預測區間。 $\eta(x_0)$ 由模型無法詮釋的部分所組成，例：若模型改以土壤砂、粉、黏比例預測不同基質勢能下的土壤含水量時，解釋變數詮釋反應變數的能力下降， $\eta(x_0)$ 數值即會隨著詮釋能力下降而增大。 $\eta_n(x_0)$ 則來自擬合模型所造成的變異性，例：給定相同樣本並以不同模型進行擬合與預測時，以多變量線性迴歸建立之模型變異性，將小於以隨機森林或類神經網路建立之模型變異性。 $\varepsilon(x_0)$ 為與生俱來的隨機性(inherent randomness)，無法透過其他額外方法或手段減少其變異性。

利用無母數自助法模擬 $\eta(x_0)$ 得預測誤差估計值 $\hat{\eta}(x_0)$ 並計算其期望值 $E[\hat{\eta}(x_0)]$ 。最後根據選定的信心水準 α ，選取估計之預測誤差期望值 $E[\hat{\eta}(x_0)]$ 的 $[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}]$ 百分位數作為 $(1 - \alpha)$ 雙尾預測誤差期望值信賴區間。本研究選擇信心水準為 0.05，並計算 95%雙尾預測誤差期望值，誤差期望值之信賴區間分析方法敘述如下(Davison & Hinkley, 1997; Efron & Tibshirani, 1994)：

1. 根據不同基質勢能建立不同推估土壤含水量的模型。
2. 以三種不同模型(MLR、RF、ANN)分別建立並以原始樣本訓練模型。
3. 以三種模型分別推估土壤含水量並減去實際土壤含水量，得到三種模型的不同基質勢能下 1822 筆的土壤含水量推估誤差值。
4. 將此 1822 筆誤差值視為一次樣本。
5. 採用自助法(Bootstrap method)，將樣本進行取後放回的隨機重複採樣 B 次，每一次隨機重複採樣所產生的自助樣本皆須與原本樣本大小相同。
6. 計算 B 次樣本的平均值並排序，2.5~97.5 百分位數即為顯著水準為 0.05 下，樣本誤差期望值的 95%雙尾信賴區間。

其中 B 為一任意數量，Efron & Tibshirani (1994)建議 B 至少需要大於 200，而 Hesterberg (2011)則建議至少大於 1000。本研究設定 B 為 1000 次。每個不同基質勢能的三種模型(MLR、RF、ANN)，皆可利用無母數自助法，分別計算當次樣本預測誤差期望值之信賴區間。



3.4.2 預測區間

式(3.16)中已包含所有未來土壤含水量預測值的不確定性，利用無母數自助法模擬 $y(x_0)$ ，即可有效評估模型預測區間。式(3.16)中每一項數值評估方式如下：

- 先以訓練集資料進行模型訓練並得到其估計值 $\hat{y}_n(x_0)$ 。
- 再利用推估之訓練集誤差，評估模型預測誤差和樣本誤差之和 $\eta(x_0) + \varepsilon(x_0)$ 。
- 最後利用無母數自助法產生自助樣本，評估模型預測值變異性 $\eta_n(x_0)$ 。

然而若僅以訓練集誤差評估 $\eta(x_0) + \varepsilon(x_0)$ 則容易陷入低估的情形，原因為訓練過程模型傾向於過擬合。因此必須將訓練集額外分割出一部分作為驗證集(不加入訓練)，以驗證集誤差評估 $\eta(x_0) + \varepsilon(x_0)$ 作為修正。但驗證集誤差往往會過大，因為自助樣本平均只包含約 2/3 的訓練集數據，表示預測結果將被人為地惡化。

於是本研究參考 Hastie *et al.* (2009) Section 7.11 節部分內容，使用無資訊誤差率(no-information error rate)、相對過擬合率(relative overfitting rate)，計算模型預測誤差和樣本誤差之和 $\eta(x_0) + \varepsilon(x_0)$ 的權重。三者的計算方式依序如式(3.17)、(3.18)、(3.19)所示。

$$\hat{\gamma} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (y(x_i) - \hat{y}_n(x_j))^2 \quad (3.17)$$

$$\hat{R} = \frac{\text{error}_{val} - \text{error}_{train}}{\hat{\gamma} - \text{error}_{train}} \quad (3.18)$$

$$\hat{w} = \frac{0.632}{1 - 0.368\hat{R}} \quad (3.19)$$

$$\eta(x_0) + \varepsilon(x_0) = (1 - \hat{w}) \times \text{error}_{train} + \hat{w} \times \text{error}_{val} \quad (3.20)$$

n 為樣本數、 $\hat{\gamma}$ 為無資訊誤差率、 $\hat{R}$ 為相對過擬合率。 error_{train} 為訓練集之誤差、 error_{val} 為驗證集之誤差。 $\hat{w}$ 為模型預測誤差和樣本誤差之和的權重。

考量實際情況，無資訊誤差率的計算量將隨著樣本數的增加而快速成長。因此本研究僅考慮 $y(x_i)$ 、 $\hat{y}_n(x_j)$ 兩者的隨機排列作為計算無資訊誤差率的替代方案。最後依序累加 $\hat{y}_n(x_0)$ 、 $\eta(x_0)$ 、 $\varepsilon(x_0)$ 、 $\eta_n(x_0)$ 得預測估計值 $\hat{y}(x_0)$ ，再依升序排列並

根據選定的信心水準 α ，選取預測估計值 $\hat{y}(x_0)$ 的 $\left[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\right]$ 百分位數作為 $(1 - \alpha)$ 雙尾預測區間。本研究選擇信心水準為 0.05，並計算 95%雙尾預測區間，演算法流程可參酌圖 3.12，每一步驟的詳細過程如下(Hastie *et al.*, 2009; Kumar & Srivistava, 2012)：

1. 將訓練集進行取後放回的隨機重複採樣，產生樣本大小與原先訓練集相同的自助樣本。
2. 選擇訓練集中未能被抽取至自助樣本的資料作為驗證集。
3. 以 1.自助樣本之解數變數作為輸入、反應變數作為輸出進行模型訓練。模型訓練完成後同步進行以下操作。
 - i. 將驗證集之解釋變數代入訓練完畢模型，推估驗證集之反應變數，並計算驗證樣本之反應變數誤差值。
 - ii. 將測試集之解釋變數代入訓練完畢模型中推估測試集之反應變數。
4. 重複操作 1、2.步驟共 B 次。
5. 將 B 次測試集之反應變數推估值減去其推估平均值。此即評估(3.15)式中模型預測值的變異性 $\eta_n(x_0)$ 。
6. 以訓練集之解數變數作為模型輸入、反應變數作為輸出進行模型訓練，並計算其反應變數推估值與誤差。
7. 以訓練集之反應變數實際值與推估值代入(3.17)式計算無資訊誤差率。
8. 建立 1~100 間隔為 1 的百分位數區間，以選取訓練集、驗證集之反應變數誤差(避免兩者選取樣本數不一致)。
9. 利用 8.建立的百分位數選取訓練集、驗證集反應變數誤差，代入(3.18)、(3.19)計算相對過擬合率及誤差權重。
10. 透過式(3.20)計算模型預測誤差和樣本誤差之和 $\eta(x_0) + \varepsilon(x_0)$ 。
11. 將樣本大小為 B 的模型預測值變異性 $\eta_n(x_0)$ ，與樣本大小為 100 的模型預測誤差和樣本誤差之和 $\eta(x_0) + \varepsilon(x_0)$ 相加。並列出所有可能的 $\hat{y}(x_0)$ 相加組合，其樣本大小為 $B \times 100$ 。其中 B 為一任意數量，本研究為求與 3.4.1 小節設定一致將 B 設定為 1000 次。
12. 計算 $\hat{y}(x_0)$ 的 2.5、97.5 百分位數，此範圍即為顯著水準 0.05 的 95%預測區間。

無母數自助法，分

自助樣本



無母數自助法，分

自助樣本

第4章 結果與討論



4.1 訓練、測試資料集探索與描述

土塊法、土扣法以及不完整的土壤資料合併後，經過 3.1 節的分析後，即可分成本研究中所使用的訓練及測試資料集。若依據美國農業部(USDA)土壤質地進行初步分類，可將訓練與測試資料集繪成三角質地圖如圖 4.1、圖 4.2 所示。

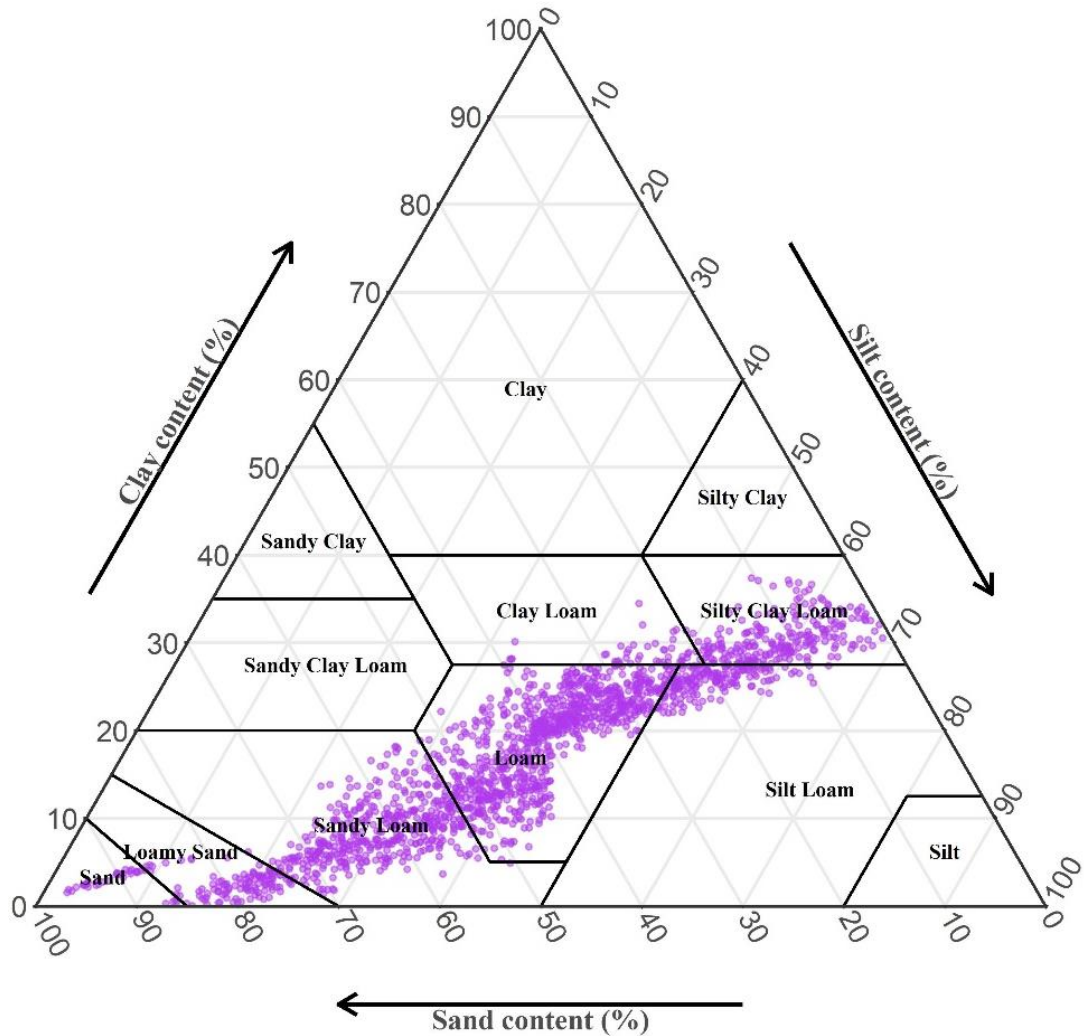


圖 4.1 訓練資料集(共 1822 筆)土壤三角質地圖

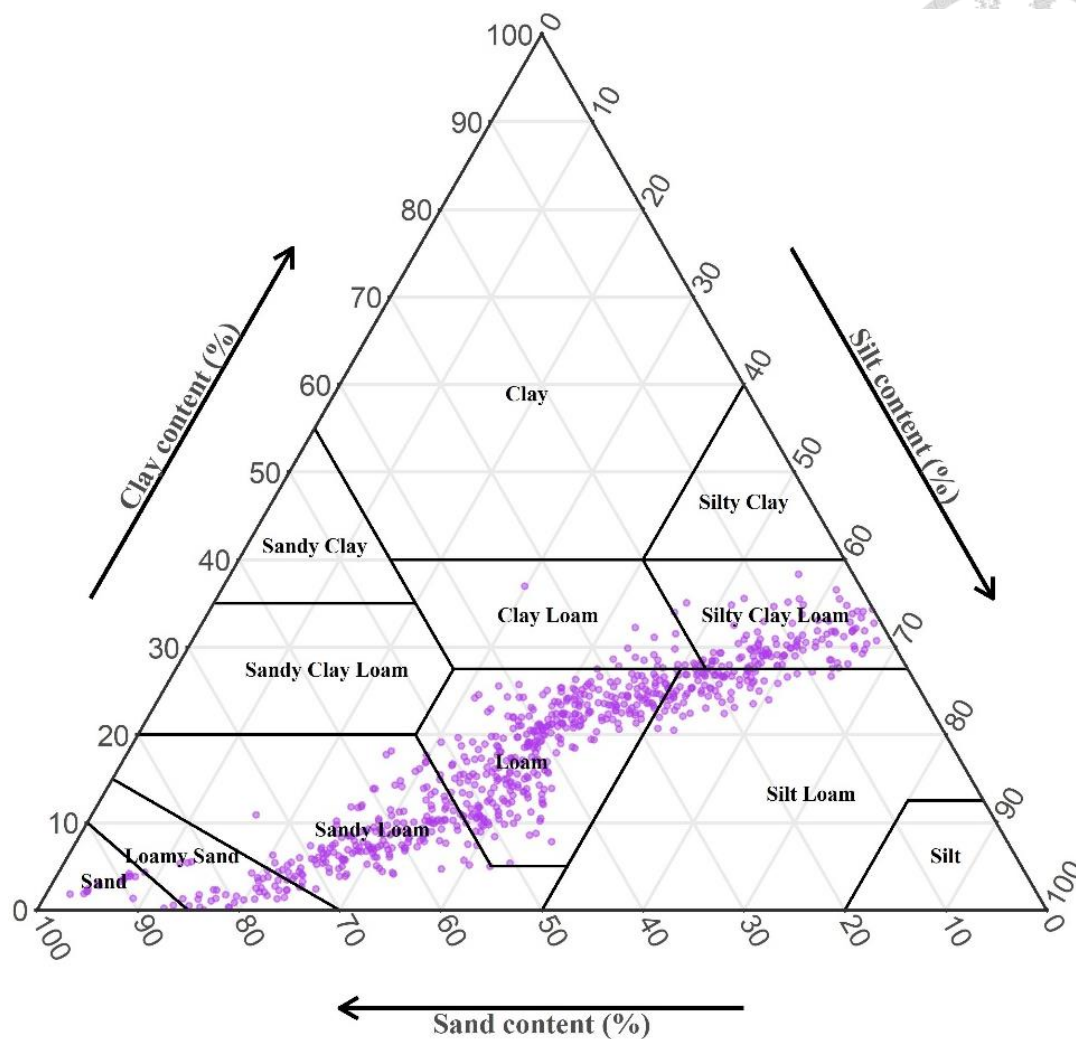


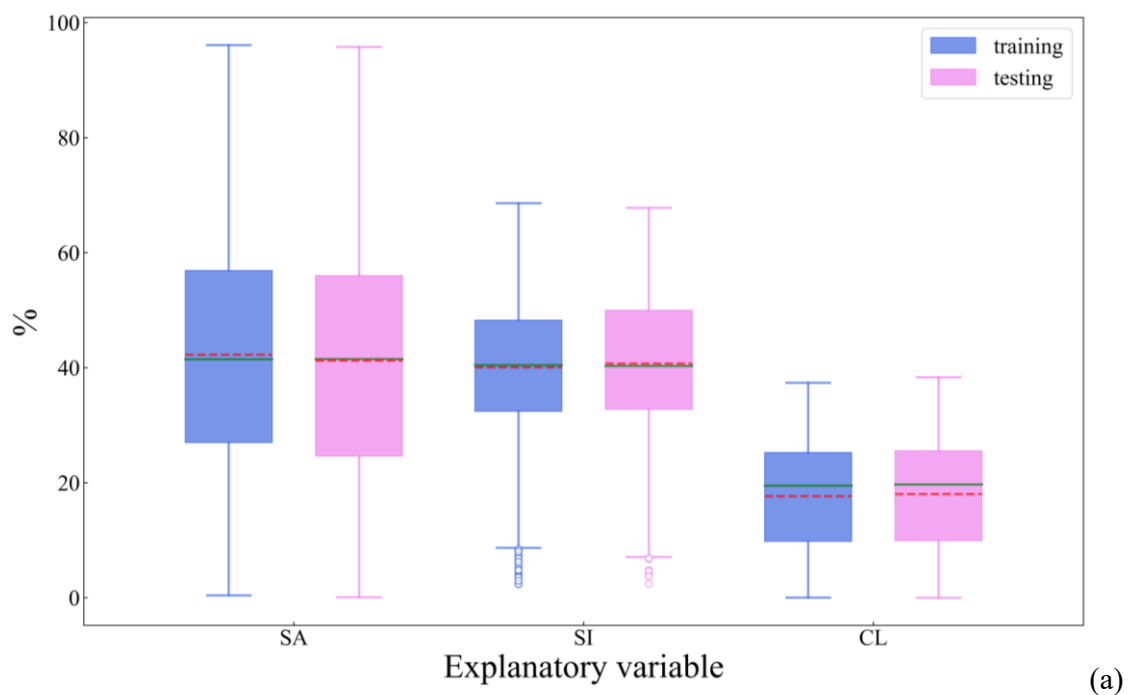
圖 4.2 測試資料集(共 792 筆)土壤三角質地圖

從圖 4.1、圖 4.2 中可以看出，無論是哪一類資料皆以壤土(Loam)、砂質壤土(Sandy Loam)、粉質黏壤土(Silty Clay Loam)為主，其次則為黏質壤土(Clay Loam)、粉質壤土(Silt Loam)、壤質砂土(Loamy Sand)。不同種類之土壤特徵描述如下 (Division of Soil Survey, 1993)：

- 砂土：砂粒肉眼可見，以手觸摸則有明顯之砂粒感。
- 粉土：粉粒肉眼不可見，以手摩擦則有粉末感，但加水呈塑性、無黏性但易斷裂。
- 黏土：黏粒具有黏性，潮濕時黏稠而有拉力且不易斷裂。
- 壤土：具備砂、粉、黏土此三者部分性質，但特徵不明顯。

其他像是壤質砂土則是以砂土的性質為主並具有部分壤土性質，黏質壤土則是以壤土的性質為主並具有部分黏土性質，其他更多的土壤性質可以相同方法推得。經土壤三角質地圖初步分類後，可得知農試所採集的土樣大部分皆與壤土密切相關。壤土相較於其他土壤較適合作為農業用途，壤土相較於砂土可以保留更多的水分及養分，且相較於黏土具備更好的排水及通風性。另外，可發現土樣分佈呈現帶狀且兩者分佈情況相似，可作為 3.1.2 小節之分析是否正確的檢核。

本研究欲繪製訓練集與測試集各解釋變數的盒狀圖，以檢視各解釋變數的離群值與分佈情況，如圖 4.3 所示。



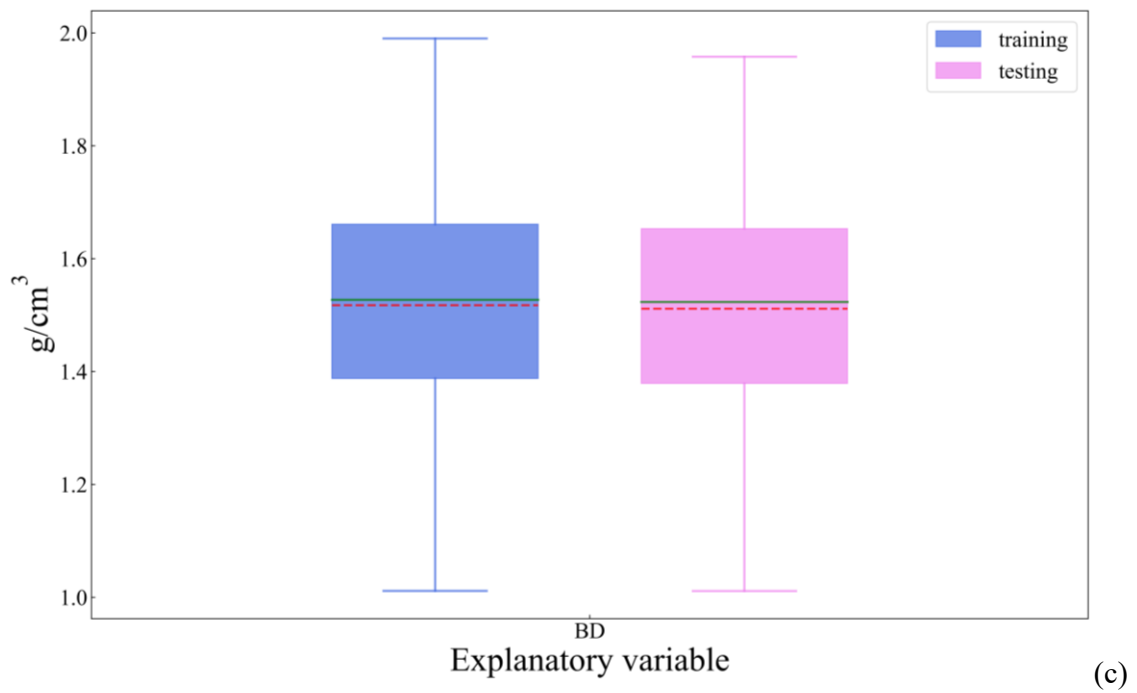
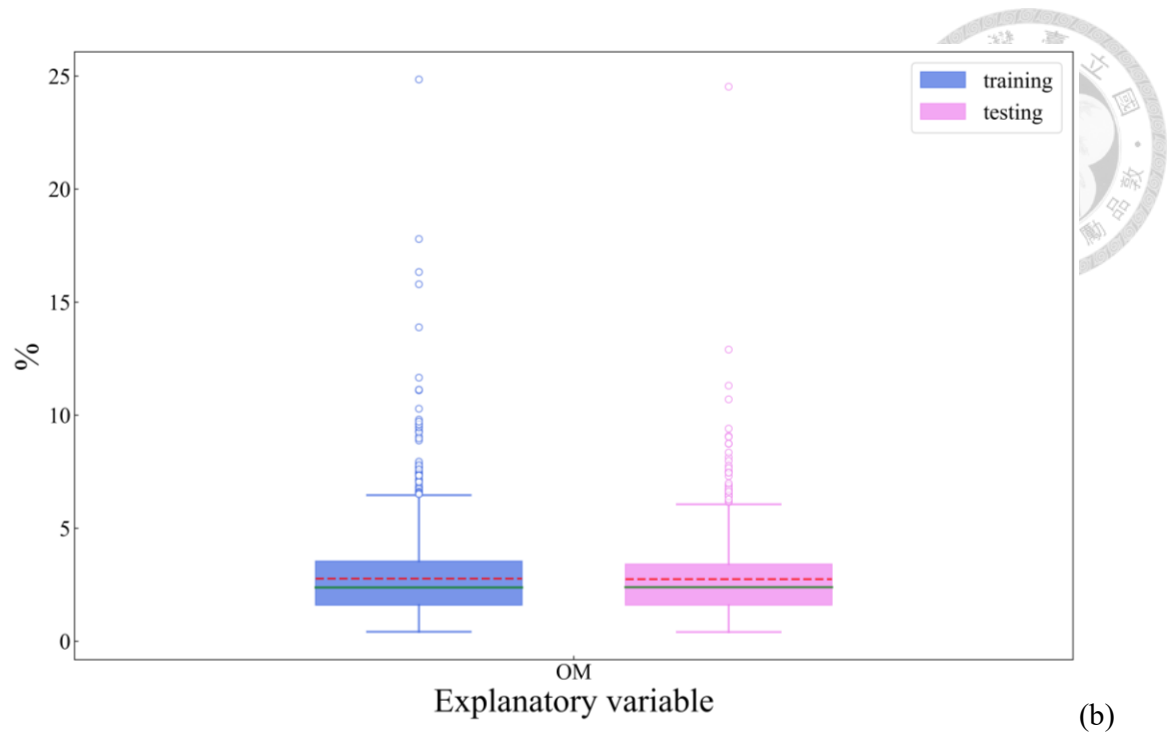


圖 4.3 各解釋變數的盒狀圖(a)砂、坩、黏含量、(b)有機質含量、(c)總體密度。藍色與紫色分別代表訓練、測試集資料。盒中的紅色虛線代表平均值，綠色實線代表中位數。盒中的上下邊界為最大、最小值。

然而不同的解釋變數皆有各自的單位(SA、SI、CL、OM 為重量百分率、BD 為 g/cm^3)，若分別繪製盒狀圖則平均數、中位數、最大值、最小值等統計數值，將會

受到單位的影響而無法直接依照數值進行後續比較。因此本研究先將各解釋變數透過式(4.1)轉換為標準分數(standard score)再繪製成盒狀圖進行分佈比較。

$$z = \frac{x - \bar{x}}{S} \quad (4.1)$$

z 為標準分數， x 為樣本平均值， S 為樣本標準差。

標準化(Standardization)為一種線性轉換，因此解釋變數的原始分佈形狀將與轉換後的分佈形狀一致。且此過程將使解釋變數無因次化，可將不同單位的解釋變數(例如：BD 單位為 g/cm^3 、其他 4 個解釋變數單位則為重量百分比)擺置相同的座標軸上，利於一同比較各解釋變數的分佈情形。使用轉換後的標準分數可繪製盒狀圖如圖 4.4 所示。可以發現不論是訓練集或測試集都有以下幾種情形：

- SA、CL 彼此分佈較為接近。
- SI、BD 彼此分佈較為接近但 SI 具有少數離群值。
- OM 分佈迥異於其他 4 種解釋變數，且分佈具有明顯的右偏(即有許多高於最大值的離群值)。

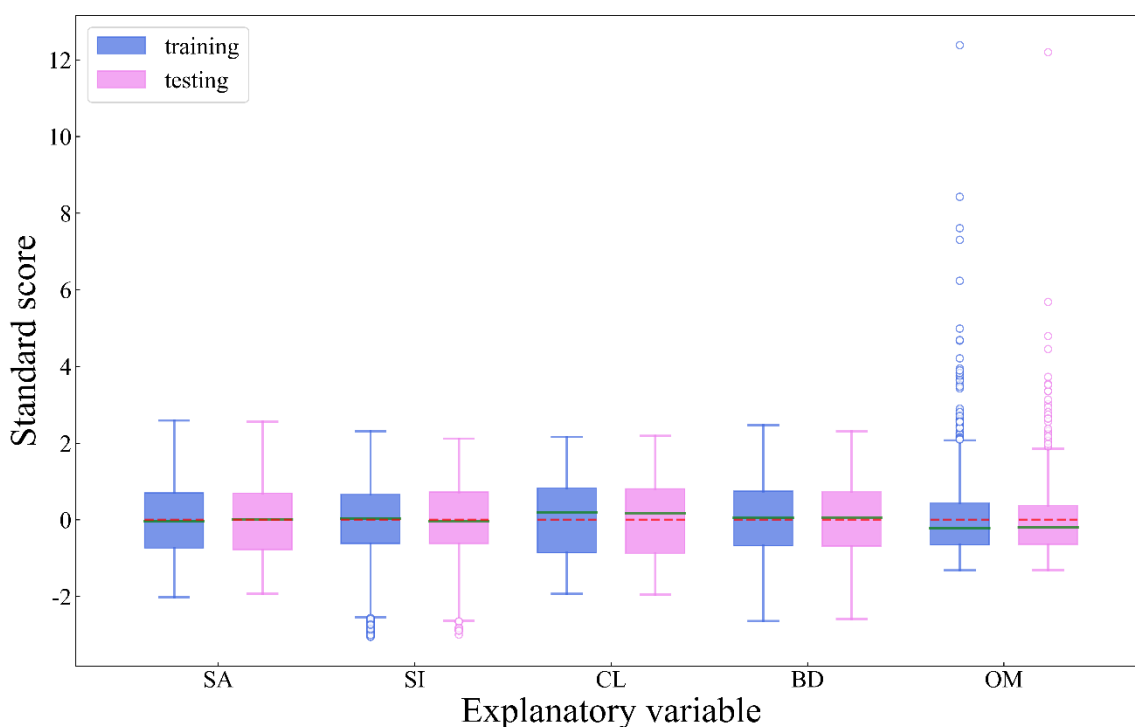


圖 4.4 各解釋變數之標準分數盒狀圖。橫軸為 5 個解釋變數、縱軸為解釋變數經過

標準化(Standardization)後之標準分數(standard score)。



最大、最小值的計算方式如式(4.2)、(4.3)所示：

$$MAX = [Q_{3/4} + 1.5 \times IQR, \max(x)]_{min} \quad (4.2)$$

$$MIN = [Q_{1/4} - 1.5 \times IQR, \min(x)]_{max} \quad (4.3)$$

x 為樣本， $Q_{3/4}$ 為第 3 四分位(upper quartile)， $Q_{1/4}$ 為第 1 四分位(lower quartile)， IQR 為四分位間距(interquartile range)，即 $Q_{3/4} - Q_{1/4}$ 。最大、最小值以外的紅點則為離群值。

若將反應變數(土壤含水量共 8 個不同基質勢能)分佈以盒狀圖方式呈現如圖 4.5 所示。

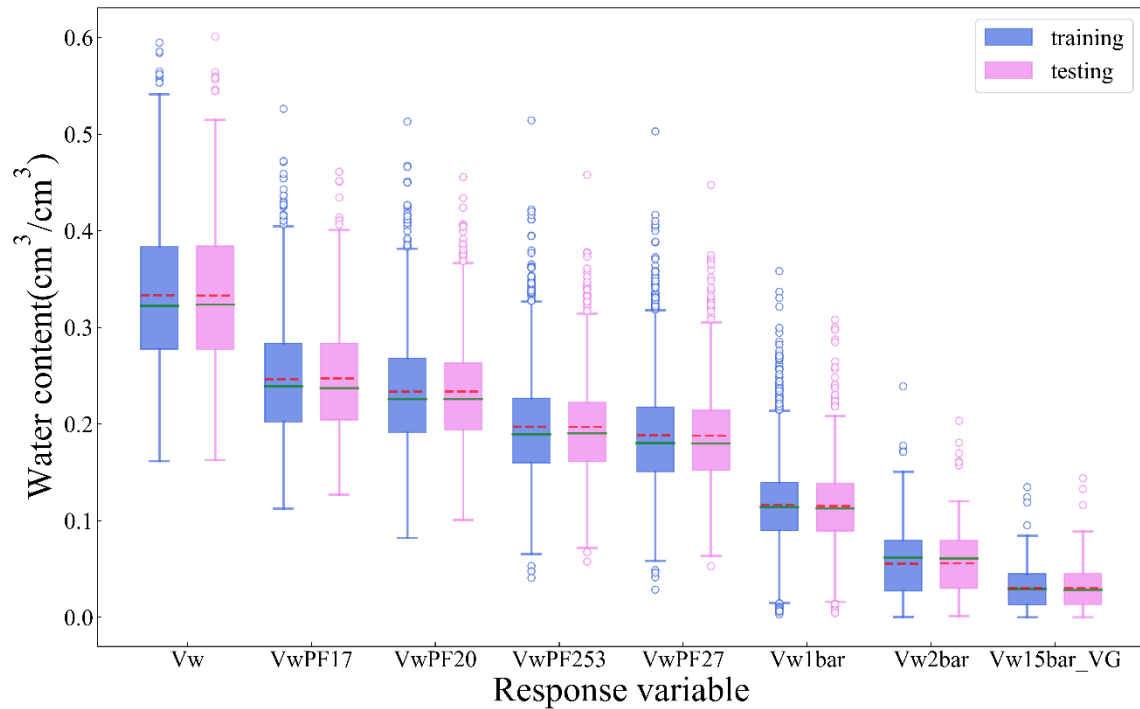


圖 4.5 各反應變數之土壤含水量盒狀圖。橫軸為 8 個反應變數、縱軸為土壤含水量，單位為體積含水量(cm^3/cm^3)。藍色與紫色分別為訓練、測試集資料。盒中紅色虛線為平均值，綠色實線為中位數。

從圖 4.5 中可發現，無論是訓練集或測試集，土壤含水量隨著基質勢能的減少，其中位數、平均值都隨之遞減。標準差則因為分佈越來越集中，因此也呈現遞減的趨勢。若將各基質勢能的土壤含水量平均值與標準差點繪在圖上可得圖 4.6。

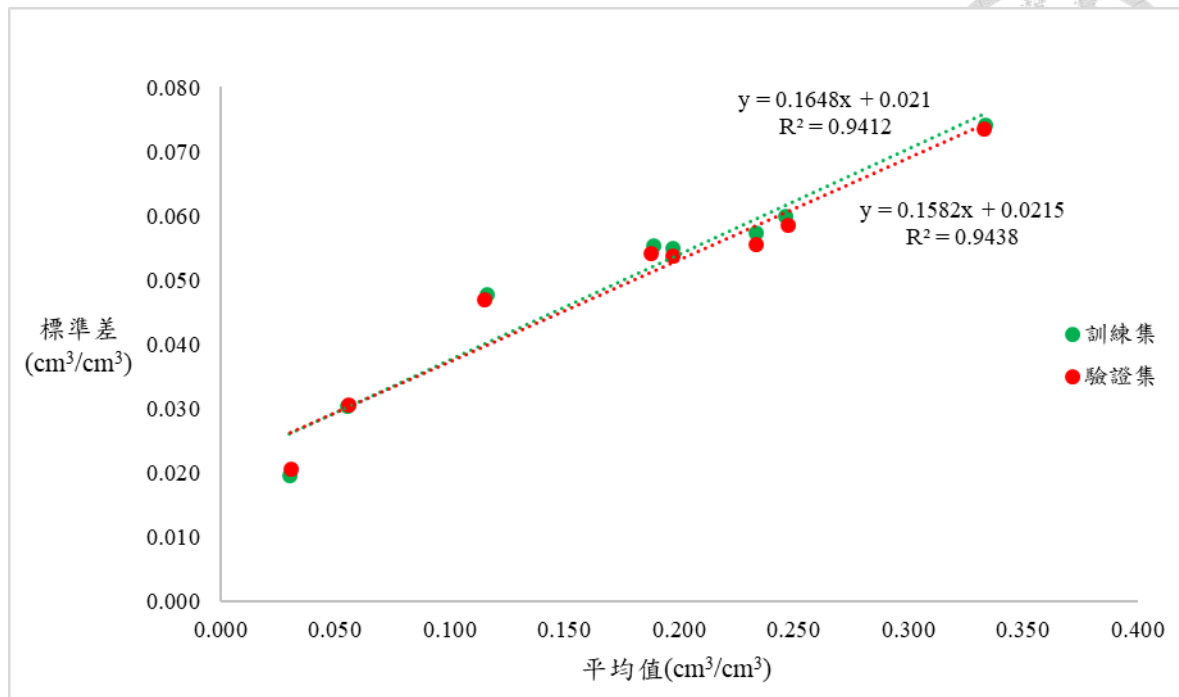


圖 4.6 訓練集與測試集土壤含水量標準差與平均值散佈圖。橫軸為土壤含水量平均值，縱軸為土壤含水量標準差，單位皆為 cm^3/cm^3 。綠點與紅點分別代表訓練集與測試集在不同基質勢能測得之土壤含水量。

從圖 4.6 可發現，標準差與平均值呈現顯著的線性正相關，具有明顯的比例效應(Proportional Effect)(Manchuk *et al.*, 2009)。從不同基質勢能的土壤含水量具有顯著比例效應可以推測，各模型對於不同基質勢能的土壤含水量預測值，其變異性應當隨著基質勢能減少而遞減，此可與圖 4.5 以及 4.2.4、4.3.1 小節之分析相互參酌。

另外從圖 4.5 也可發現不同土壤含水量之間的重疊範圍廣大，若依照物理-半經驗模型如 VG model，將此 8 個土壤含水量視為同一種土壤在不同基質勢能的條件下，並以 VG model 參數作為反應變數，則模型訓練與建構時極有可能造成顧此失彼的情況，即 V_w 擬合良好但 V_{w1bar} 、 V_{w2bar} 等擬合較差。因此本研究為了避免上述情況發生，參考 Rawls *et al.* (1982) 建立點型 PTF 的方式，假設 8 個土壤含水量彼此之間互為獨立關係。再依序建置 8 個模型依序對應於 8 個不同基質勢能的土壤含水量，每一個模型僅預測單一基質勢能下的土壤含水量。

本研究採取三種不同的預測模型(MLR、RF、ANN)，並以 SA、SI、CL、OM、BD 作為解釋變數，土壤含水量作為反應變數。若欲初步檢核解釋變數的選擇是否

合理，可透過計算解釋變數與反應變數之相關係數。透過(2.21)、(2.22)式即可分別計算解釋變數與反應變數之間的兩種相關係數。若再進一步假設顯著水準(significance level, 即發生第一型錯誤的機率)為 0.05, 則可對兩種相關係數進行顯著性檢定, 檢核兩種相關係數為 0 的機率, 此即顯著性檢定的 p 值(p-value)。計算結果如表 4.1、表 4.2 所示。

表 4.1 解釋變數與反應變數之皮爾森積動差相關係數與 p 值表。/前數值代表訓練集, /後數值代表測試集。括號內數值為 p 值。黃色區域為訓練、測試集 p 值皆大於顯著水準。淺藍色區域為僅測試集 p 值大於顯著水準。淺橘色區域為僅訓練集 p 值大於顯著水準。

	Vw	VwPF17	VwPF20	VwPF253	VwPF27	Vw1bar	Vw2bar	Vw15bar_VG
SA	0.149/0.231 (0.000/0.000)	-0.039/0.039 (0.099/0.276)	-0.128/-0.066 (0.000/0.063)	-0.313/-0.274 (0.000/0.000)	-0.323/-0.278 (0.000/0.000)	-0.415/-0.377 (0.000/0.000)	-0.435/-0.438 (0.000/0.000)	-0.378/-0.382 (0.000/0.000)
SI	-0.181/-0.265 (0.000/0.000)	0.030/-0.048 (0.199/0.175)	0.120/0.059 (0.000/0.095)	0.297/0.262 (0.000/0.000)	0.301/0.263 (0.000/0.000)	0.398/0.368 (0.000/0.000)	0.451/0.446 (0.000/0.000)	0.393/0.394 (0.000/0.000)
CL	-0.094/-0.167 (0.000/0.000)	0.047/-0.023 (0.044/0.525)	0.128/0.070 (0.000/0.048)	0.311/0.270 (0.000/0.000)	0.326/0.278 (0.000/0.000)	0.405/0.361 (0.000/0.000)	0.379/0.393 (0.000/0.000)	0.328/0.336 (0.000/0.000)
BD	-0.460/-0.470 (0.000/0.000)	-0.352/-0.375 (0.000/0.000)	-0.323/-0.339 (0.000/0.000)	-0.271/-0.250 (0.000/0.000)	-0.283/-0.263 (0.000/0.000)	-0.196/-0.174 (0.000/0.000)	0.168/0.148 (0.000/0.000)	0.067/0.045 (0.004/0.208)
OM	0.231/0.202 (0.000/0.000)	0.237/0.212 (0.000/0.000)	0.278/0.262 (0.000/0.000)	0.350/0.321 (0.000/0.000)	0.360/0.331 (0.000/0.000)	0.363/0.372 (0.000/0.000)	0.083/0.159 (0.000/0.000)	0.114/0.186 (0.000/0.000)

表 4.2 解釋變數與反應變數之斯皮爾曼等級相關係數與 p 值表

	Vw	VwPF17	VwPF20	VwPF253	VwPF27	Vw1bar	Vw2bar	Vw15bar_VG
SA	0.181/0.264 (0.000/0.000)	-0.003/0.081 (0.894/0.023)	-0.087/-0.010 (0.000/0.780)	-0.278/-0.232 (0.000/0.000)	-0.289/-0.241 (0.000/0.000)	-0.402/-0.364 (0.000/0.000)	-0.426/-0.421 (0.000/0.000)	-0.386/-0.398 (0.000/0.000)
SI	-0.208/-0.300 (0.000/0.000)	-0.007/-0.097 (0.778/0.006)	0.075/-0.004 (0.001/0.906)	0.258/0.218 (0.000/0.000)	0.263/0.224 (0.000/0.000)	0.388/0.364 (0.000/0.000)	0.451/0.442 (0.000/0.000)	0.409/0.419 (0.000/0.000)
CL	-0.133/-0.216 (0.000/0.000)	0.018/-0.060 (0.453/0.091)	0.099/0.026 (0.000/0.473)	0.286/0.236 (0.000/0.000)	0.303/0.247 (0.000/0.000)	0.397/0.348 (0.000/0.000)	0.370/0.382 (0.000/0.000)	0.339/0.359 (0.000/0.000)
BD	-0.416/-0.449 (0.000/0.000)	-0.313/-0.350 (0.000/0.000)	-0.281/-0.308 (0.000/0.000)	-0.216/-0.199 (0.000/0.000)	-0.227/-0.213 (0.000/0.000)	-0.133/-0.128 (0.000/0.000)	0.189/0.186 (0.000/0.000)	0.081/0.074 (0.001/0.038)
OM	0.129/0.145 (0.000/0.000)	0.147/0.164 (0.000/0.000)	0.207/0.227 (0.000/0.000)	0.303/0.290 (0.000/0.000)	0.318/0.304 (0.000/0.000)	0.337/0.350 (0.000/0.000)	0.128/0.183 (0.000/0.000)	0.171/0.213 (0.000/0.000)

從表 4.1、表 4.2 中塗色區域可以發現，訓練與測試集之解釋變數 SA、SI、CL 與反應變數 VwPF17、VwPF20，線性相關或是單調相關程度皆遠低於其他反應變數。此即說明土壤質地對於基質勢能為 -0.05 、 -0.1bar 的土壤含水量相關性低。可預期土壤質地在 VwPF17、PF20 的標準化迴歸方程式係數也會較 BD、OM 低。



4.2 模型訓練

4.2.1 多變量線性迴歸

本研究欲透過 SA、SI、CL、OM、BD 作為解釋變數，建立多變量線性迴歸模型用以預測不同基質勢能的土壤含水量。因此模型訓練前必須先確認各解釋變數是否皆為線性獨立，本研究即透過計算 VIF 進行共線性分析。透過式(3.5)即可計算不同解釋變數的 VIF。若已知 SA、SI、CL 三個解釋變數為完全線性相依(三者相加總是為 1)，可預期式(3.5)所計算出之 SA、SI、CL 的 VIF 皆為無限大(因為 $R^2 = 1$)。因此每次僅選擇 4 個解釋變數計算 VIF，且不能同時包含 SA、SI、CL 以避免 VIF 為無限大，並依序計算所有可能的組合，如表 4.3。

表 4.3 訓練集之變異數膨脹因子表

	SA	SI	CL	OM	BD
VIF (without SA)	NA	4.05	4.27	1.51	1.37
VIF (without SI)	11.41	NA	11.76	1.51	1.37
VIF (without CL)	22.11	21.57	NA	1.51	1.37

註：每一列表示一次 VIF 計算組合，NA 表示這一系列之 VIF 計算不考慮此變數。

從表 4.3 中可以看見若剔除 SA 則可以最有效地消除共線性影響，因此往後之多變量迴歸分析皆令式(3.4)中的 $C_1 = 0$ ，使式(3.4)改寫為式(4.4)。

$$\theta(\psi) = C_0' + C_1' \times SI + C_2' \times CL + C_3' \times BD + C_4' \times OM \quad (4.4)$$

其中 C_0' 為修改後之迴歸方程式截距， C_1' 、 C_2' 、 C_3' 、 C_4' 為修改後之迴歸方程式係數。

另外從表 4.3 中可發現，SA、SI、CL 似乎不影響 OM、BD 的變異數膨脹因子計算。若欲檢核解釋變數之間的線性相關性，則可透過計算 5 個解釋變數之間的皮爾森積動差相關係數，如表 4.4。

表 4.4 訓練集解釋變數之間的皮爾森相關係數表

	SA	SI	CL	OM	BD
SA	1	-0.98	-0.95	-0.23	0.00
SI	-0.98	1	0.86	0.18	0.03
CL	-0.95	0.86	1	0.29	-0.03
OM	-0.23	0.18	0.29	1	-0.50
BD	0.00	0.03	-0.03	-0.50	1

從表 4.4 中可以看到 SA、SI、CL 對於 OM 影響程度相近，而對於 BD 幾乎都無影響力。也因此無論是剔除 SA、CL、OM 中的哪一項，透過式(3.5)對於 OM、BD 的變異數膨脹因子計算皆無顯著影響。

多變量線性迴歸模型經過訓練集資料最佳化後，可得式(4.4)在不同基質勢能下預測土壤含水量的迴歸係數與截距，如表 4.5 所示(無括號之數值)。從表 4.5 中可以發現，基質勢能較高的土壤含水量(Vw、VwPF17、VwPF20、VwPF253、VwPF27)預測值主要貢獻來自於截距。此即表示 SA、SI、CL、OM、BD 這 5 個解釋變數，對於較低基質勢能的土壤含水量無明顯線性關係，解釋變數無法準確捕捉數值變化趨勢。此外，因為各迴歸係數帶有解釋變數的單位，因此無法直接透過迴歸係數的數值，比較變數之間的重要性。

表 4.5 各基質勢能預測土壤含水量之多變量線性迴歸截距與係數表。其中無括號之數值表示迴歸係數與截距，括號之數值表示標準化迴歸係數。

	intercept	SI	CL	BD	OM
Vw	0.63913 (0.000)	-0.00182 (-0.304)	0.00123 (0.151)	-0.16921 (-0.438)	0.00087 (0.021)
VwPF17	0.38318 (0.000)	0.00019 (0.040)	-0.00011 (-0.017)	-0.09844 (-0.315)	0.00255 (0.076)
VwPF20	0.31976 (0.000)	0.00056 (0.122)	-0.00013 (-0.021)	-0.07790 (-0.261)	0.00420 (0.131)
VwPF253	0.20977 (0.000)	0.00089 (0.200)	0.00045 (0.075)	-0.04858 (-0.170)	0.00640 (0.208)
VwPF27	0.20742 (0.000)	0.00075 (0.167)	0.00072 (0.118)	-0.05218 (-0.181)	0.00637 (0.205)
Vw1bar	0.07550 (0.000)	0.00105 (0.272)	0.00051 (0.097)	-0.01890 (-0.076)	0.00665 (0.248)
Vw2bar	-0.04549 (0.000)	0.00118 (0.480)	-0.00022 (-0.067)	0.03417 (0.216)	0.00212 (0.125)
Vw15bar_VG	-0.01522 (0.000)	0.00070 (0.442)	-0.00018 (-0.084)	0.01124 (0.110)	0.00125 (0.114)

若將訓練集之解釋變數與反應變數先進行標準化，再進行多變量線性迴歸模型擬合，則可得標準化迴歸截距與係數，如表 4.5 所示(括號內數值)。所有變數均先經過標準化後再進行迴歸，因此表 4.5 各解釋變數之截距均為 0。表 4.5 之係數表示解釋變數每增加一個標準差的數值，反應變數所增減的標準差數值。從表 4.5 可以發現，BD 對於預測基質勢能較高的土壤含水量(Vw、VwPF17)貢獻較大。若再根據(2.20)式可得知，顆粒密度固定時 BD 與孔隙率為一對一之函數關係，此即表示土壤結構對於基質勢能較高的土壤含水量影響較大。土壤質地(SI、CL)則對於預測基質勢能較低的土壤含水量(Vw2bar、Vw15bar_VG)貢獻較大。土壤質地對於土壤孔隙水影響較小而對於吸附水影響則較大，表示低基質勢能時土壤結構對於土壤含水量的影響力下降，而土壤質地則是影響力上升。

標準化迴歸係數具有計算快速且準確反應各解數變數重要性的優點，然而無法反應各解數變數重要性的變異性為其缺點。排列重要性則可完整分析各解數變數重要性的分佈，但是計算較為費時。本研究將以排列重要性作為特徵重要性的主

要衡量指標(將於 4.3.2 小節與其他模型進行詳細比較)。



4.2.2 隨機森林

使用 SA、SI、CL、OM、BD 作為解釋變數，建立隨機森林模型用以預測不同基質勢能的土壤含水量。隨機森林屬於決策樹模型，因此在擬合隨機森林模型時，可同步計算出輸入變數對於此模型的基尼重要性，如圖 4.7。

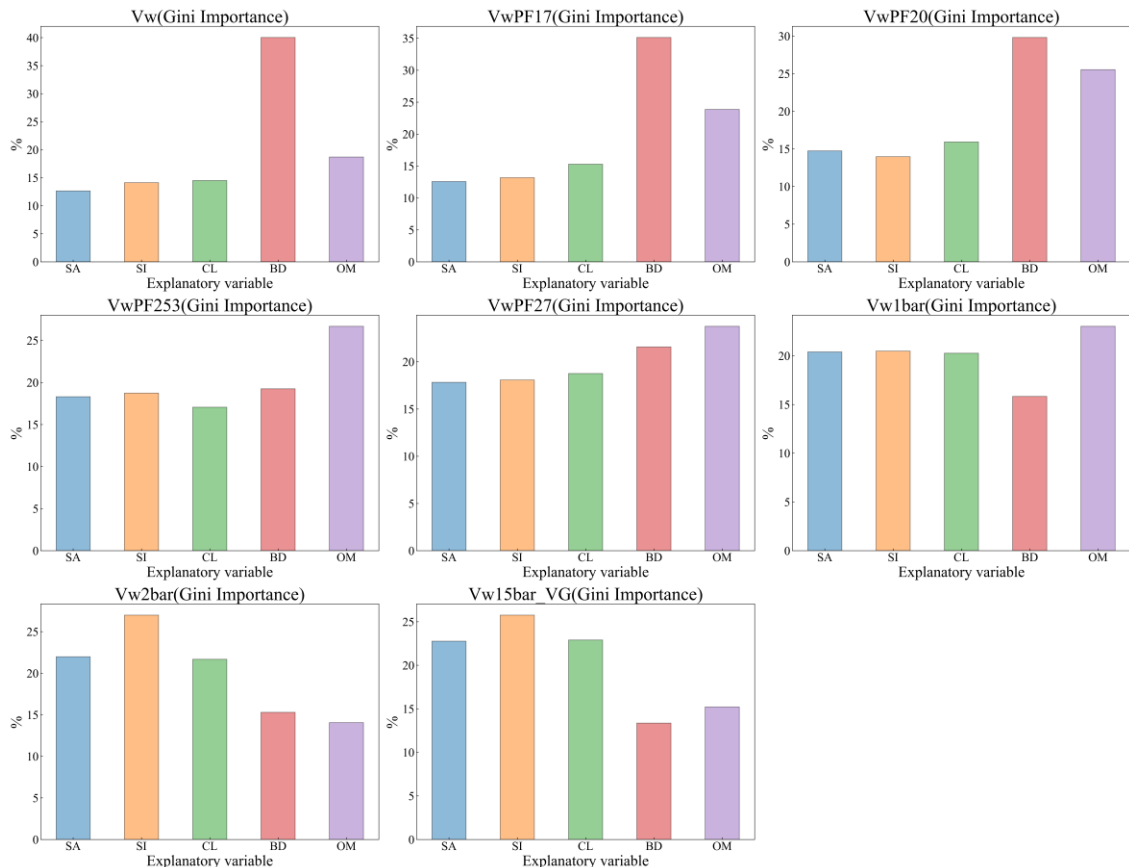


圖 4.7 隨機森林模型在不同基質勢能之解釋變數基尼重要性(訓練集資料)。橫軸為 5 個不同的解釋變數。縱軸為基尼重要性數值，單位為無因次。

圖 4.7 為隨機森林模型擬合過程所計算的基尼重要性，透過圖 4.7 可以瞭解哪一些解釋變數對於降低隨機森林的基尼雜質扮演重要角色，例：基質勢能 $\geq -0.5\text{bar}$ (VwPF27)以 BD、OM 為主要考量的解釋變數；基質勢能為 -1bar 時所有解釋變數皆具一定影響力，但是 BD 比重下降；基質勢能為 -2 、 -15bar 時則以土壤質地(砂、粉、黏比例)為主要考量的解釋變數。然而基尼重要性存在數個疑慮，此點於 3.3.1 小節已詳細闡述。因此基尼重要性僅作為特徵重要性的輔助指標，將以

排列重要性作為特徵重要性的主要指標進行後續分析(將於 4.3.2 小節與其他模型進行詳細比較)。



4.2.3 類神經網路

使用 SA、SI、CL、OM、BD 作為解釋變數，建立類神經網路模型用以預測不同基質勢能的土壤含水量。透過檢核類神經網路模型在不同基質勢能之訓練曲線如圖 4.8，可以了解模型目前的訓練程度為過擬合抑或是欠擬合。

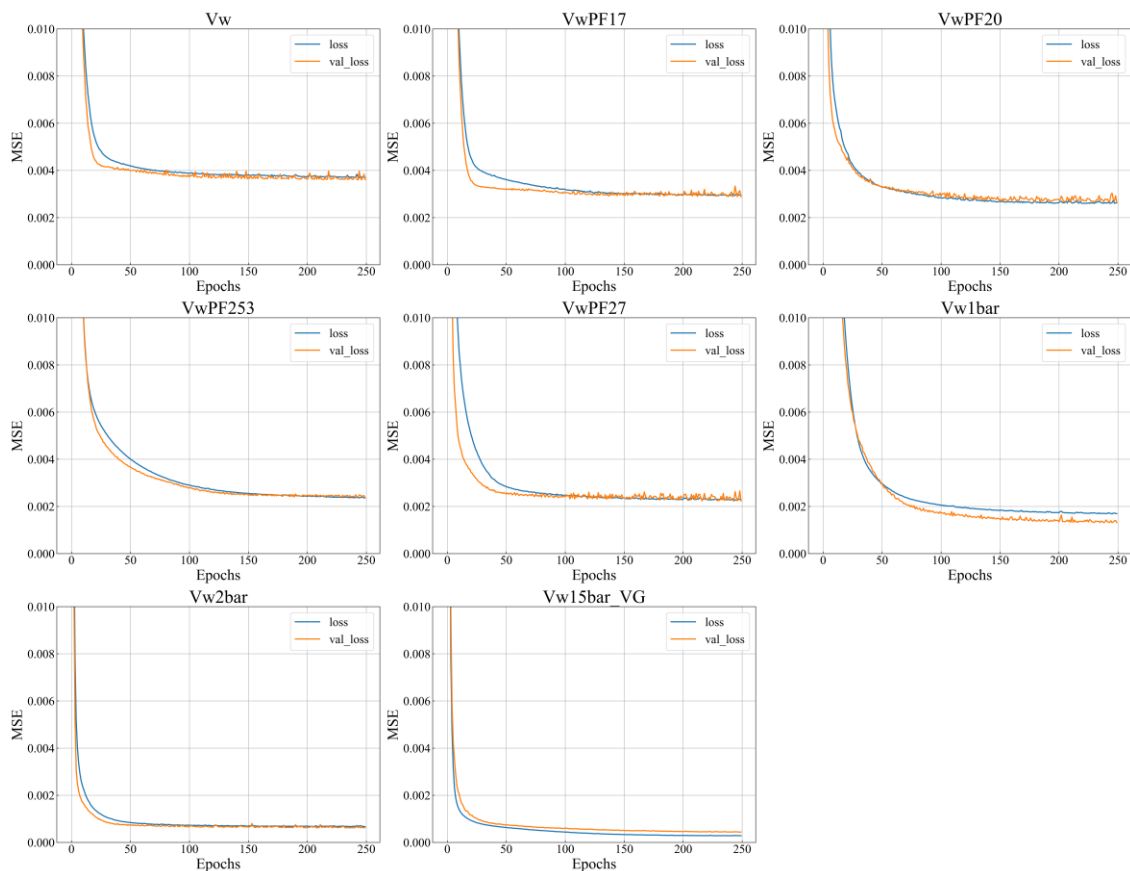


圖 4.8 類神經網路模型在不同基質勢能之訓練曲線。橫軸為類神經模型完整訓練的次數(或稱期)。縱軸為土壤含水量的均方誤差，單位為 $(\text{cm}^3/\text{cm}^3)^2$ 。藍線表示訓練集資料，橘色線表示驗證集資料。將原始訓練集(1822 筆)依照 80%、20%的比例分成模型訓練集與模型驗證集資料

從圖 4.8 可以發現各土壤含水量之模型訓練集與驗證集，皆已達到穩定的趨勢且兩條曲線數值接近。驗證集雖有部分震盪，但其幅度小且相對穩定，表示此時模型介於過擬合(overfitting)和欠擬合(underfitting)之良好擬合狀態。



4.2.4 土壤含水量訓練集預測結果

以訓練集資料分別代入三種訓練完畢之模型，評估不同模型的訓練結果，如表 4.6 所示(訓練集 1822 筆土壤資料，對應 8 個不同基質勢能的含水量預測散佈圖，可參酌圖 A.1、圖 A.6、圖 A.11)。觀察表 4.6 中可以發現，RMSE 隨著基質勢能減少而降低，但 R^2 則無明顯趨勢。其中 R^2 以式(2.23)、RMSE 以式(2.26)計算可得。表示殘差平方和 SS_{res} 與總平方和 SS_{tot} 同步降低，才能滿足 RMSE 與 R^2 的預測趨勢。此結果與圖 4.6 相似，可能原因為資料變異性隨著基質勢能減少而降低。表 4.6 陳列三種不同模型的訓練結果。RF 對於各基質勢能的預測土壤含水量皆有較低的 RMSE 與較高的 R^2 ，訓練結果其次的模型則為 ANN，最差則為 MLR。然而若模型訓練過程產生過擬合情形，則訓練結果將大幅高估實際預測能力，進而導致最終結果的誤判。因此除了比較訓練集預測結果之外，尚須結合測試集預測結果，方可確認模型的實際預測能力。(表 4.7 呈現測試集預測結果，分析內容將在 4.3.1 小節進行討論)

表 4.6 三種模型預測各基質勢能土壤含水量均方根誤差和決定係數(訓練集)

	model	Vw	VwPF17	VwPF20	VwPF253	VwPF27	Vw1bar	Vw2bar	Vw15bar_VG
RMSE (cm ³ /cm ³)	MLR	0.064	0.056	0.053	0.049	0.049	0.041	0.026	0.018
	RF	0.055	0.050	0.047	0.042	0.043	0.035	0.023	0.015
	ANN	0.061	0.054	0.051	0.049	0.048	0.040	0.026	0.018
R²	MLR	0.25	0.13	0.13	0.20	0.21	0.25	0.24	0.17
	RF	0.45	0.31	0.32	0.41	0.40	0.45	0.45	0.37
	ANN	0.33	0.18	0.20	0.21	0.26	0.29	0.29	0.18



4.3 模型預測

4.3.1 土壤含水量測試集預測結果

以測試集資料分別代入三種訓練完畢之模型，評估不同模型的泛化能力，結果如表 4.7 所示(測試集 792 筆土壤資料，對應 8 個不同基質勢能的含水量預測散佈圖，可參酌圖 A.2、圖 A.7、圖 A.12)。觀察表 4.7 中可以發現，RMSE 隨著基質勢能減少而降低，但 R^2 則無明顯趨勢。表示殘差平方和 SS_{res} 與總平方和 SS_{tot} 同步降低，才能滿足 RMSE 與 R^2 的預測趨勢。此結果與 4.2.4 小節相似，可能原因為資料變異性隨著基質勢能減少而降低。從表 4.7 中可以發現 RF 對於預測各基質勢能的土壤含水量皆有較低的 RMSE 與較高的 R^2 (僅 Vw15bar_VG 預測表現略遜於 MLR)。若再同時綜合考量三種不同模型訓練集結果(表 4.6)，則 RF 仍然是預測結果最為良好的 PTF。

另外，值得注意的是，類神經網路在訓練集表現較多變量線性迴歸優異，但測試集表現卻較差。可能原因與類神經網路模型參數較多、自由度較高有關。即使圖 4.8 透過驗證集檢核模型為良好擬合，但從表 4.7 的測試集預測結果可以發現，類神經網路可能仍有些微的過擬合導致模型泛化能力較差。

表 4.7 三種模型預測各基質勢能土壤含水量均方根誤差和決定係數(測試集)

	model	Vw	VwPF17	VwPF20	VwPF253	VwPF27	Vw1bar	Vw2bar	Vw15bar_VG
RMSE (cm ³ /cm ³)	MLR	0.062	0.054	0.052	0.049	0.049	0.041	0.027	0.019
	RF	0.062	0.053	0.051	0.047	0.047	0.039	0.027	0.019
	ANN	0.064	0.055	0.052	0.049	0.049	0.041	0.027	0.019
R^2	MLR	0.28	0.14	0.13	0.17	0.18	0.23	0.24	0.18
	RF	0.29	0.16	0.17	0.24	0.24	0.29	0.24	0.17
	ANN	0.25	0.10	0.12	0.16	0.19	0.23	0.21	0.10

4.3.2 參數重要性

MLR、RF、ANN 可依照 3.3.2 小節的流程，將訓練完成之模型代入測試集數據，計算 1000 次參數排列重要性，並將結果以盒狀圖呈現如圖 4.9。

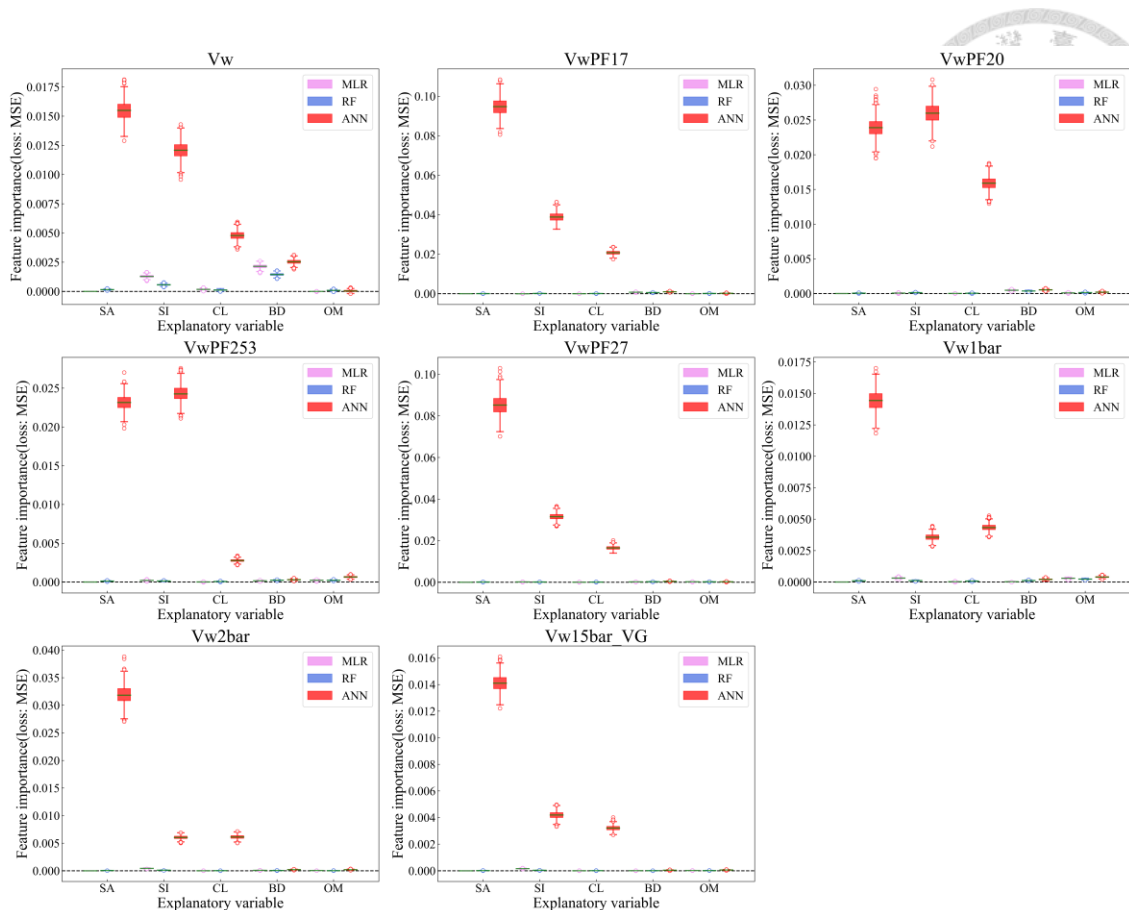


圖 4.9 三種模型在不同基質勢能之反應變數排列重要性(測試集)。橫軸為 5 種不同的解釋變數(MLR 為解決共線性問題而排除 SA，因此其值為 0)。縱軸為擾動解釋變數所增加的模型均方誤差 MSE，單位為土壤含水量比值平方(cm^3/cm^3)²。虛線表示模型均方誤差 MSE 不變。模型均方誤差 MSE 增加較高者，其解釋變數重要性也較大。

透過比較不同模型間相同反應變數的排列重要性，可以衡量不同模型預測能力的穩定性。穩定性較高之模型較不易因解釋變數的微小擾動，導致最終預測結果產生大幅度的改變。從圖 4.9 可以發現 ANN 各反應變數的排列重要性明顯高於其他兩種模型。以相同擾動程度的反應變數代入模式進行土壤含水量預測，ANN 的模型均方誤差增加量遠高於另外兩者。此即表示 ANN 預測結果相對於 MLR、RF 較不穩定。此原因可能與 ANN 具有較多模型參數有關，導致其模型自由度較高但是泛化能力相對較弱。泛化能力較弱之模型，對於未來應用於全新的土壤資料進行轉換與預測時，其表現將低於原先模型訓練的預期結果。

若比較同一模型內不同反應變數的排列重要性，則可以衡量該模型不同反應變數對於模型預測值的影響。從圖 4.9 可以發現 ANN 在所有基質勢能下，SA、SI、

CL 的排列重要性皆遠高於 BD、OM，代表其重要性遠高於 BD、OM。在高基質勢能時，此結果與 Dexter (2004)所提出的結論不同(基質勢能較高時，土壤含水量多寡與 BD 相關程度較高)。可能原因與前述相同，類神經模型參數較多、自由度較高，使得模型穩定度較其他兩種模型低。也因此類神經網路的參數重要性，仍需進一步的證據檢核此趨勢是否合理。

若進一步比較 MLR 與 RF 反應變數的排列重要性，可將圖 4.9 重新放大繪製成圖 4.10。從圖 4.10 中可以發現以下事項。兩個模型在預測高基質勢能(V_w 、 V_{wPF17} 、 V_{wPF20} 、 V_{wPF253} 、 V_{wPF27})土壤含水量時，反應變數的排列重要性趨勢接近，且解釋變數以 BD 最為重要。隨著基質勢能減少 BD 重要性下降而 SI、OM 重要性上升；預測低基質勢能(V_{w1bar} 、 V_{w2bar} 、 V_{w15bar_VG})土壤含水量時，MLR 的反應變數以 SI 最為重要(從表 4.1 可見 SI 與 V_{w1bar} 、 V_{w2bar} 、 V_{w15bar_VG} 之間的相關係數較高、表 4.5 為其迴歸結果)；RF 的反應變數同樣以 SI 最為重要，但其比例不如 MLR 相對其他反應變數高。也因此 RF 相對於 MLR 模型預測能力的穩定性更高。

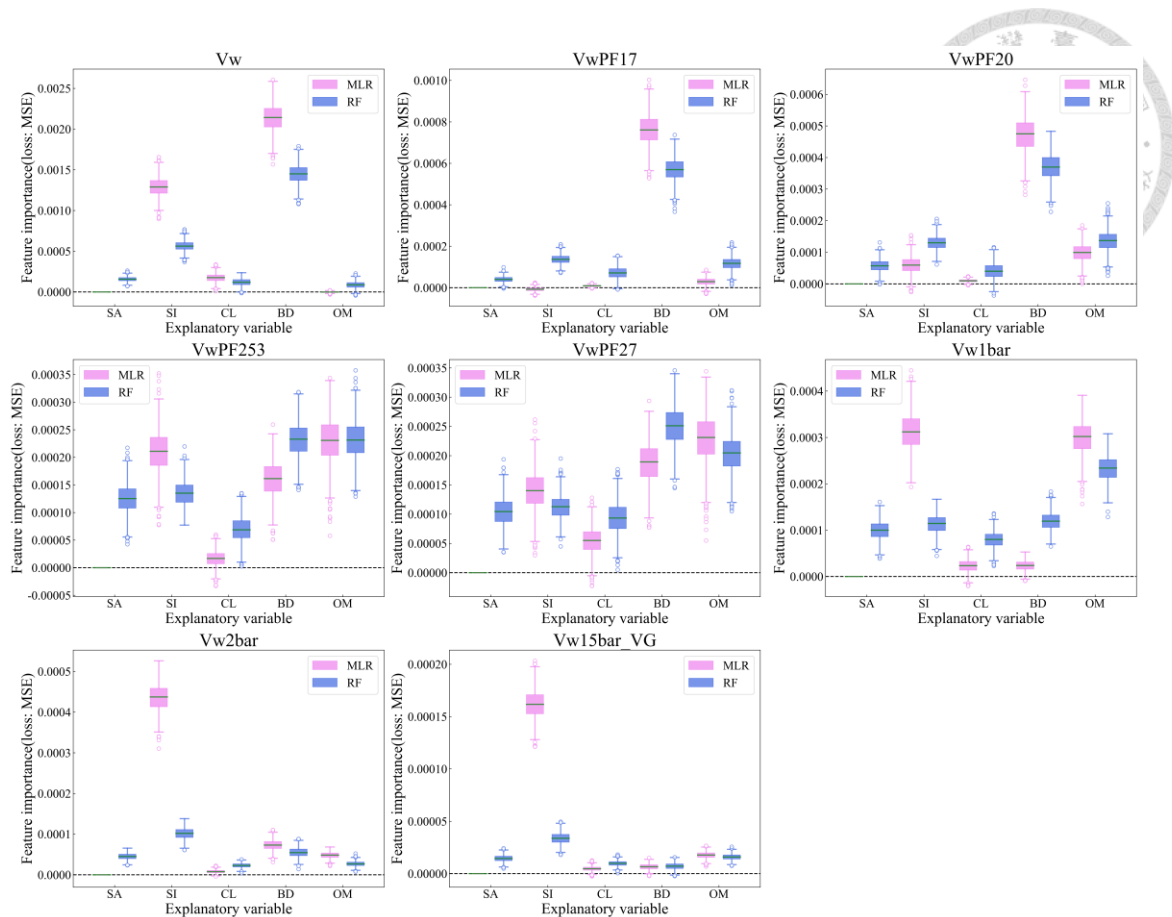


圖 4.10 MLR 與 RF 在不同基質勢能之反應變數排列重要性(測試集)。橫軸為 5 種不同的解釋變數(MLR 為解決共線性問題而排除 SA，因此其值為 0)。縱軸為擾動解釋變數所增加的模型均方誤差 MSE，單位為土壤含水量比值平方(cm^3/cm^3)²。虛線表示模型均方誤差 MSE 不變。模型均方誤差 MSE 增加較高者，其解釋變數重要性也較大。

從上述參數重要性結果可總結如下：預測高基質勢能之土壤含水量時，以 BD 為最重要之參數；而預測低基質勢能之土壤含水量時，則以 SI、OM 為最重要之參數。Yang *et al.* (2014)、Liu *et al.* (2020) 點出 SI 和 OM 為影響土壤比表面積最重要的因素。前人的眾多研究皆提及在基質勢能較低時，土壤吸附水與比表面積密切相關(Banin & Amiel, 1970; Grismer, 1987; Petersen *et al.*, 1996; Ross, 1978)。而隨著基質勢能逐漸降低，土壤含水量組成由孔隙水逐漸轉變為由吸附水所構成，因此 SI、OM 的參數重要性也隨著基質勢能降低而逐漸增加，與本研究的參數重要性結果吻合。



4.3.3 誤差期望值之信賴區間

MLR、RF、ANN 可依照 3.4.1 小節的流程，將訓練完成之模型代入測試集數據衡量推估誤差，並將其分佈以盒狀圖繪出，如圖 4.11。

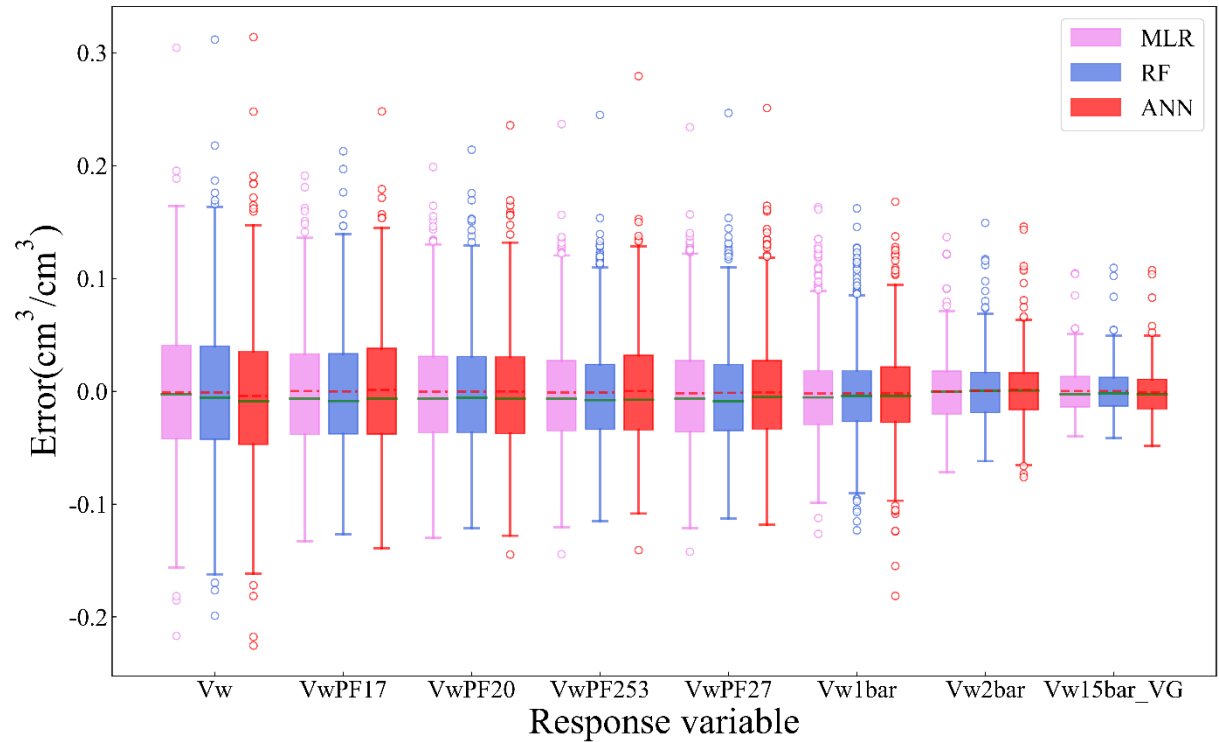


圖 4.11 三種模型土壤含水量預測誤差盒狀圖(測試集)。橫軸表示不同基質勢能。縱軸為土壤含水量預測誤差，單位為體積含水量(cm^3/cm^3)。

從圖 4.11 中可以發現，三種方法預測各基質勢能下的土壤含水量，誤差期望值皆近似於 0，且隨著基質勢能減少其誤差變異性遞減。此結果與 4.1 節反應變數分佈相似(圖 4.5、圖 4.6)，可能表示(3.16)式模型總變異量中的模型參數不確定性較小，模型總變異量由資料不確定性所控制。

本研究欲檢核三種模型的預測結果是否為不偏估計，可根據誤差期望值的信賴區間進行假設檢定(以下假設顯著水準為 0.05)。其中虛無假設敘述如下：模式預測結果為偏誤估計。透過無母數自助法可將圖 4.11 中的誤差樣本，以取後放回的方式重複產生 1000 次自助樣本，並計算此 1000 次自助樣本的平均值。再將此 1000 次自助樣本平均值以升序方式排列，取其 2.5 百分位數和 97.5 百分位數，即可得以自助法評估之 95%信賴區間，如表 4.8 所示。從表 4.8 中可以發現，三種模型預

測誤差之期望值其 95%信賴區間皆涵蓋 0。因此可以拒絕原先的虛無假設，推斷三種模型的預測結果為不偏估計。

表 4.8 三種模型預測誤差期望值之 95%信賴區間表(測試集)。每一格中的數值分別代表土壤含水量預測誤差的 95%信賴區間的下界與上界，單位為體積含水量 (cm^3/cm^3)。

	Vw	VwPF17	VwPF20	VwPF253	VwPF27	Vw1bar	Vw2bar	Vw15bar_V G
MLR	-0.00507~ 0.00317	-0.00360~ 0.00381	-0.00408~ 0.00340	-0.00441~ 0.00249	-0.00515~ 0.00180	-0.00478~ 0.00088	-0.00195~ 0.00180	-0.00113~ 0.00152
RF	-0.00530~ 0.00295	-0.00355~ 0.00364	-0.00372~ 0.00360	-0.00407~ 0.00240	-0.00494~ 0.00184	-0.00438~ 0.00114	-0.00167~ 0.00205	-0.00108~ 0.00151
ANN	-0.00841~ 0.00003	-0.00252~ 0.00513	-0.00361~ 0.00385	-0.00319~ 0.00384	-0.00449~ 0.00245	-0.00485~ 0.00111	-0.00091~ 0.00298	-0.00251~ 0.00023

4.3.4 模型預測區間

應用 3.4.2 小節的方式，若假設顯著水準為 0.05，以自助法衡量三種模型不同基質勢能下，每個土壤含水量預測值的 95%預測區間。三種模型在不同基質勢能的土壤含水量皆可繪製一張散佈圖如圖 4.12 所示(此處僅列出一張圖作為範例，若全部列出將有 $3 \times 8 \times 2 = 48$ 張圖)。圖 4.12 為 RF 的土壤飽和含水量預測區間。從圖 4.12 可以發現 RF 土壤飽和含水量預測區間上下界會有輕微震盪。原因為應用自助法的隨機性，隨著抽樣次數的增加其振幅下降。從圖 4.12 中也可發現，此隨機性不會對整體預測範圍產生偏移或其他明顯影響，因此可以有效應用估計模型預測區間。三種模型在不同基質勢能下，兩種資料集所有土壤含水量預測區間的實際涵蓋率如表 4.9 所示。

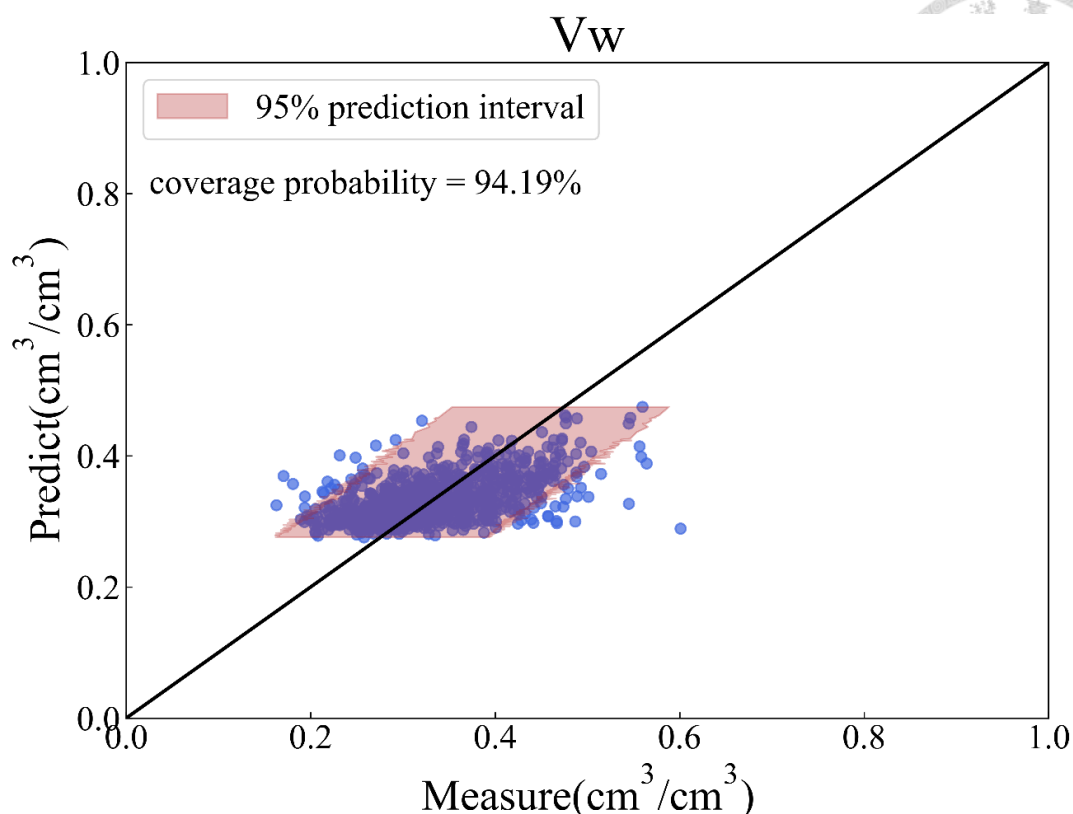


圖 4.12 RF 模型之土壤飽和含水量預測區間(測試集)。藍點為 792 筆土壤資料的飽和含水量預測值與量測值散佈點。橫軸為土壤含水量量測值，縱軸為土壤含水量預測值，單位皆為 cm^3/cm^3 。圖中黑色實線為 45°線(量測值=預測值)。紅色區塊為連接每個土壤樣本的 95%預測區間之上下界所圍出的區域。實際涵蓋率(coverage probability)則標注在散佈圖中。

表 4.9 三種模型在不同基質勢能下的土壤含水量預測區間實際涵蓋率。每一格中的數值由上而下分別代表訓練集、測試集的實際涵蓋率(coverage probability)。

	Vw	VwPF17	VwPF20	VwPF253	VwPF27	Vw1bar	Vw2bar	Vw15bar_V G
MLR	95.01% 94.95%	95.17% 95.20%	95.01% 96.46%	95.17% 94.82%	95.17% 94.57%	95.17% 95.20%	95.06% 95.20%	95.06% 94.19%
RF	97.69% 94.19%	96.82% 95.33%	97.04% 95.96%	97.09% 94.44%	97.20% 94.70%	96.82% 94.95%	97.64% 94.95%	97.91% 94.07%
ANN	96.32% 93.69%	96.93% 94.70%	96.76% 95.45%	96.49% 94.19%	96.32% 94.70%	96.38% 94.57%	96.98% 94.32%	95.28% 95.08%

檢核實際量測之含水量是否落在預測區間內，即可評估預測區間的實際涵蓋率，如表 4.9。從表 4.9 中可以發現三種模型的測試集實際涵蓋率皆可逼近 95%。而訓練集則因式(3.17)~(3.20)的修正，使得非線性模型 RF、ANN 之訓練集預測區間被高估(部分貢獻來自誤差較大之驗證集)，與 4.3.4 小節預期結果一致。透過檢核實際涵蓋率也可發現，自助法可以有效評估非線性模型的預測區間。

4.4 與 Rosetta3 之比較

本研究為檢核建立台灣本土 PTF 的必要性，使用目前受到國內外廣泛應用的 PTF—Rosetta3(Zhang & Schaap, 2017)進行台灣中部農業土壤的轉換，並與本研究先前建立之台灣本土 PTF 轉換結果比較。本研究使用 Rosetta3 第二階層模型，分別輸入台灣土壤資料的 SA、SI、CL、BD 作為解釋變數，進行台灣土壤保水曲線轉換。而為了對本土 PTF 與 Rosetta3 進行公平比較，將先前所分析之台灣土壤保水曲線 PTF 刪減 OM 解釋變數後，剩下 SA、SI、CL、BD 共 4 個解釋變數，使解釋變數數量與 Rosetta3 相同。再依照圖 3.6 之流程重新擬合三種不同的本土 PTF，計算各基質勢能土壤含水量誤差。

本研究則分別從點以及連續型兩個不同層面進行本土 PTF 與 Rosetta3 轉換誤差的比較。點的部分為了契合農試所提供之原始土壤資料，本研究根據 Rosetta3 轉換後的參數，求取各基質勢能下的土壤含水量，並分析其預測誤差。以測試集內每一筆土壤資料，8 個不同基質勢能的土壤含水量誤差的 RMSE 作為評估標準。將本土 PTF 與 Rosetta3 之 RMSE 分佈以盒狀圖方式呈現，如圖 4.13 所示。

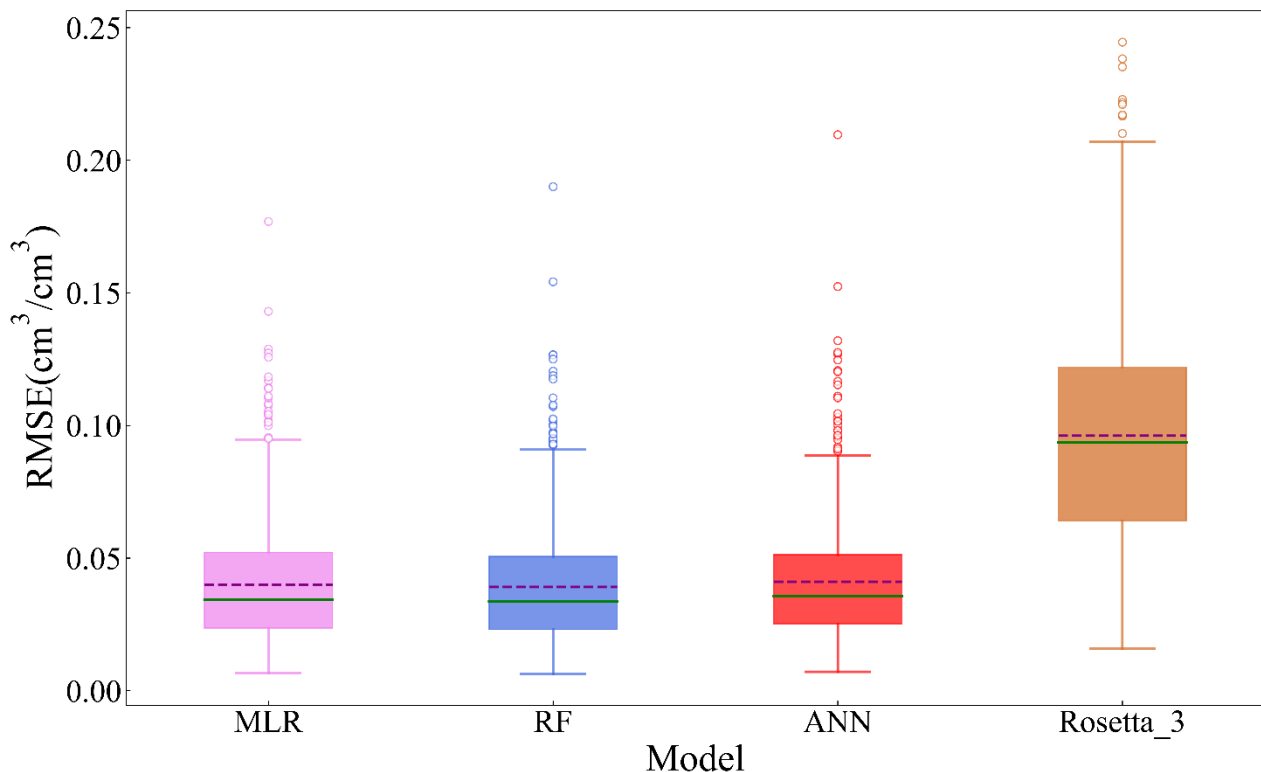


圖 4.13 台灣本土 PTF 與 Rosetta3 土壤含水量轉換誤差盒狀圖(測試集)。橫軸為四種

不同的 PTF。縱軸為土壤含水量的均方根誤差，單位為體積含水量(cm^3/cm^3)。

從圖 4.13 可以發現，若根據 RMSE 的平均值、中位數與整體分佈形狀，選擇轉換結果最為優良的台灣本土 PTF，則以 RF 為最佳的 PTF(平均值與中位數最小，整體分佈與 MLR、ANN 相近)。另外，無論是哪一種台灣本土 PTF(MLR、RF、ANN)，其表現明顯優於 Rosetta3 模型。

若欲比較 Rosetta 和本土 PTF 對於土壤保水曲線的轉換結果，則可透過計算轉換後與實際土壤保水曲線之弗雷歇距離(Fréchet distance)進行比較。弗雷歇距離是一種用於比較兩個曲線或軌跡之間相似度的距離度量(Fréchet, 1906)。其計算方法為在兩個曲線上各選取一個起始點，然後移動這兩個點，使它們可同時沿著各自曲線前進且不折返，在連續移動的過程中，兩個點的移動速度可任意決定。最後則依序計算每一個移動過程中兩個點之間的連線長度，最長的連線長度即為弗雷歇距離。

弗雷歇距離的計算是一個複雜問題，需要考慮曲線上的所有點以及它們之間的連接方式。在實際應用中，通常使用近似或離散演算法來計算弗雷歇距離，以提高計算效率。本研究即透過 python 模組 frechetdist 計算離散弗雷歇距離(Eiter & Mannila, 1994)。

以測試集資料的 8 個實測土壤含水量進行 VG model 參數最佳化即可得實際土壤保水曲線。同理，以測試集資料代入重新訓練完成之三種本土 PTF，得到 8 個土壤含水量預測值後，再進行 VG model 參數最佳化即可得三種本土 PTF 的土壤保水曲線轉換結果。Rosetta3 則因轉換結果直接為 VG model 參數因此不需再做其他處理。最後計算以三種本土 PTF 和 Rosetta3 兩種方式轉換的土壤保水曲線與實際土壤保水曲線的弗雷歇距離，並將其結果繪製成盒狀圖如圖 4.14。

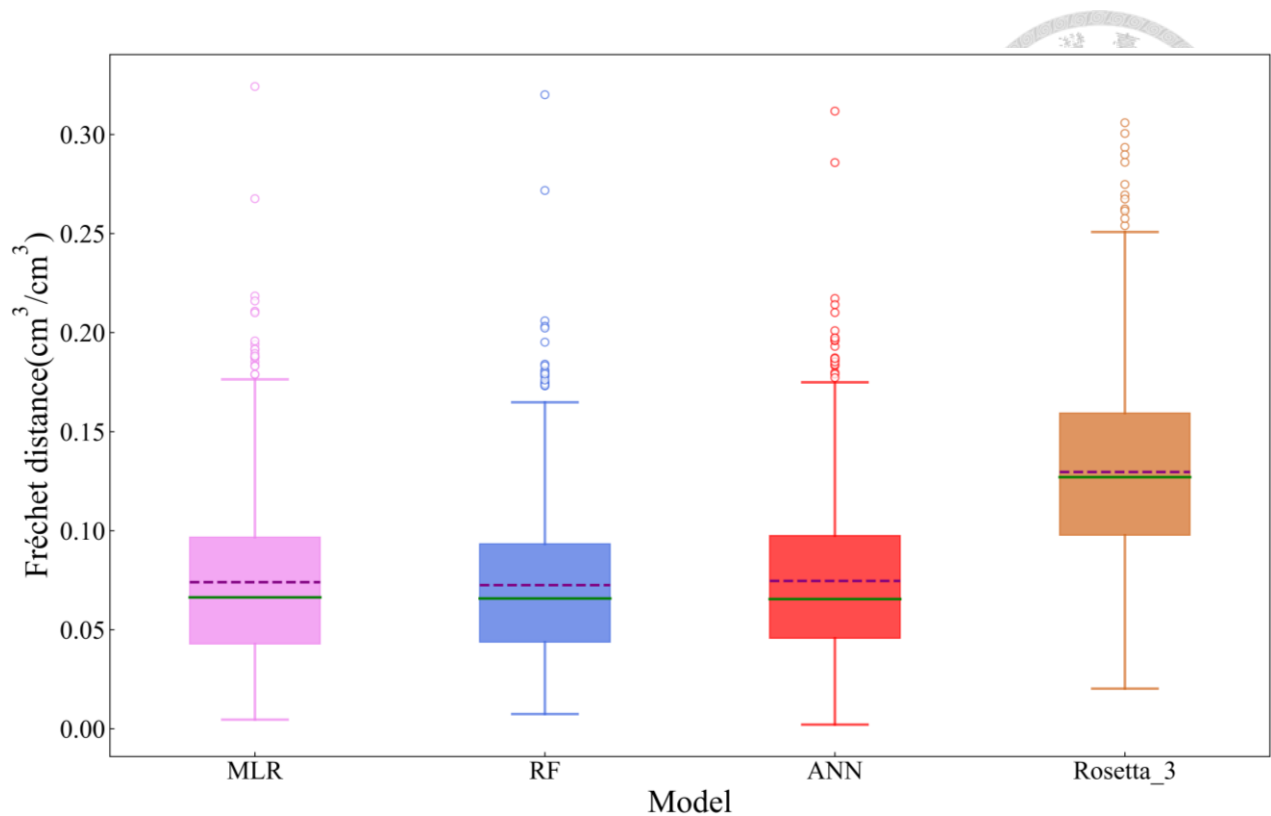


圖 4.14 台灣本土 PTF 與 Rosetta3 轉換與實際土壤保水曲線弗雷歇距離盒狀圖(測試集)。橫軸為四種不同的 PTF。縱軸為土壤含水量的均方根誤差，單位為體積含水量 (cm^3/cm^3)。

從圖 4.14 可以發現，若根據弗雷歇距離的平均值、中位數與整體分佈形狀，選擇轉換結果最為優良的台灣本土 PTF，則以 RF 為最佳的 PTF(平均值與中位數最小，整體分佈也較集中)。另外，無論是哪一種台灣本土 PTF(MLR、RF、ANN)，其表現也明顯優於 Rosetta3 模型。

綜合圖 4.13、圖 4.14 皆可發現本土 PTF 明顯優於 Rosetta3 的轉換結果。探究其主要原因可能為 Rosetta3 是基於類神經網路所建立的 PTF。訓練類神經網路時，有較高的機會學習一些訓練資料的變異性，然而此訓練資料的變異性與台灣本土土壤的變異性迥異。因此造成 Rosetta3 應用在台灣本土農業土壤的泛化能力降低，也再次凸顯建立台灣本土 PTF 的重要性。



4.5 增加額外解釋變數的影響

根據 4.3.1 小節的分析可以得知三種模型的轉換結果其決定係數皆小於 0.5，表示原先 5 個解釋變數僅可詮釋 $\leq 50\%$ 的反應變數變異性。若欲提升本土 PTF 的轉換結果，則可根據現有的土壤資料屬性，增加額外的解釋變數並重新訓練本土 PTF。本研究選擇轉換結果最佳的 RF 進行額外分析，參考 Rawls *et al.* (1982)、Rosetta3 作法選擇田間含水量(VwPF253)或永久凋萎點(Vw15bar_VG)作為額外輸入解釋變數。重新進行 4.3.1 小節分析新訓練之 PTF 轉換結果如表 4.10 所示。

表 4.10 RF 預測各基質勢能土壤含水量均方根誤差和決定係數(6 個解釋變數)。每一個儲存格內的數值從上而下依序代表訓練集、測試集。

	Vw	VwPF17	VwPF20	VwPF253	VwPF27	Vw1bar	Vw2bar	Vw15bar_VG
RMSE (cm ³ /cm ³)	0.036	0.029	0.023	0.009	0.014	0.027	0.022	0.015
	0.044	0.032	0.027	0.016	0.019	0.032	0.026	0.018
R ²	0.76	0.77	0.84	0.98	0.94	0.69	0.45	0.43
	0.64	0.70	0.77	0.91	0.88	0.54	0.26	0.23

比較表 4.7 與表 4.10 的 RF 轉換表現可發現。加入田間含水量(VwPF253)作為額外的解釋變數，對於預測基質勢能 $\geq -1\text{bar}$ 的土壤含水量，其結果有顯著提升；預測基質勢能 -2 、 -15bar 的土壤含水量其結果則無明顯改變。若再額外加入永久凋萎點(Vw15bar_VG)作為輸入解釋變數，重新進行 4.3.1 小節分析新訓練之 PTF 轉換結果如表 4.11 所示。

表 4.11 RF 預測各基質勢能土壤含水量均方根誤差和決定係數(7 個解釋變數)。每一個儲存格內的數值從上而下依序代表訓練集、測試集。

	Vw	VwPF17	VwPF20	VwPF253	VwPF27	Vw1bar	Vw2bar	Vw15bar_VG
RMSE (cm ³ /cm ³)	0.036	0.024	0.023	0.013	0.015	0.014	0.008	0.003
	0.043	0.030	0.027	0.020	0.021	0.023	0.015	0.008
R ²	0.77	0.84	0.84	0.95	0.93	0.92	0.94	0.97
	0.65	0.73	0.77	0.87	0.84	0.76	0.76	0.85

若比較表 4.10、表 4.11 可發現。加入永久凋萎點(Vw15bar_VG)作為第二個額外的解釋變數，對於預測基質勢能 $\leq -2\text{bar}$ 的土壤含水量，其結果有顯著提升。綜合上述分析結果可以發現，加入田間含水量或永久凋萎點，對於預測各基質勢能的土壤含水量有顯著提升。可能導致此現象的原因分為下列兩點：1.不同基質勢能的土壤含水量之間，具備某種程度的相關性。基質勢能相近的土壤含水量相關性強，隨著基質勢能差距增加兩者相關性逐漸下降。2.土壤含水量量測時可能產生一些系統誤差，使得解釋變數的詮釋能力降低。然而若加入單點或雙點的土壤含水量進行模型擬合，擬合過程即可考慮部分系統誤差，因此最終土壤含水量預測結果得以進行修正。

另外，本研究也額外探討土壤含水量預測值的總變異性，是否會受到額外增加的解釋變數影響。承繼與表 4.10、表 4.11 相同之模型，重新進行 4.3.4 小節分析模型之預測區間，並選擇飽和土壤含水量進行繪製，如圖 4.15 所示。

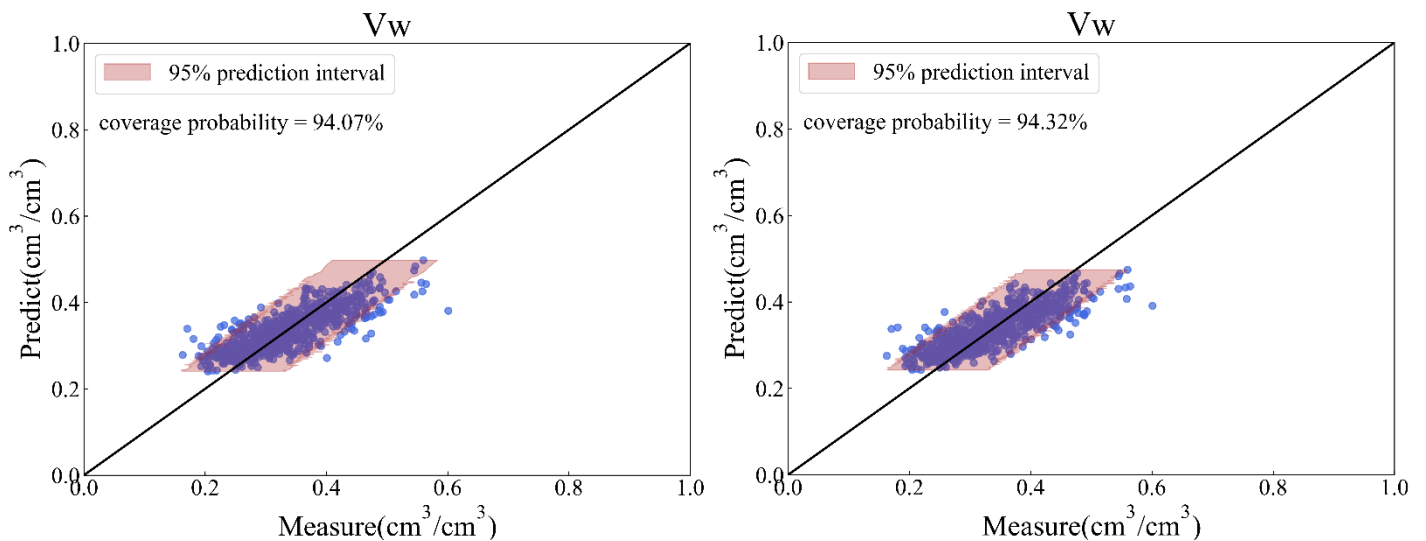


圖 4.15 RF 新模型之土壤飽和含水量預測區間(測試集)。左圖為加入田間含水量、右圖為加入田間含水量與永久凋萎點，作為額外解釋變數訓練 RF 的預測區間。

比較圖 4.12、圖 4.15 兩者的 95%預測區間寬度(粉紅色區域寬)可發現。圖 4.12 預測區間寬度約為 $0.22\text{cm}^3/\text{cm}^3$ ，圖 4.15 則為 $0.17\text{cm}^3/\text{cm}^3$ ，顯示新模型的預測值總變異性降低。從 3.4 節不確定性分析的式(3.16)中可以得知，預測值的變異性主要由 $\eta(x_0)$ 、 $\eta_n(x_0)$ 、 $\varepsilon(x_0)$ 此三部分所構成。Minasny *et al.* (1999)已指出模型參數不確定性遠較輸入變數的誤差小，表示 $\eta_n(x_0)$ 數值在總變異性中占比小。而 $\varepsilon(x_0)$

則無法透過其他方法減少其變異性，可以得知增加額外解釋變數不會影響其數值。因此可判斷導致預測值總變異性降低的主要原因為 $\eta(x_0)$ 下降(即模型預測表現顯著提升)。

綜上所述，新增田間含水量或永久凋萎點作為額外解釋變數，對於 PTF 整體預測能力與單點預測能力都有顯著的提升。未來即可根據土壤資料屬性齊全程度，選擇建立階層式的 PTF 以提升轉換結果。



4.6 點型與連續型 PTF 之不確定性比較

Rosetta3 透過自助法提供轉換參數的算術平均值與標準差。然而 Rosetta3(連續型 PTF)所提供的土壤保水曲線參數不確定性分析，難以應用於農試所提供之不同基質勢能下的單點土壤含水量。因此本研究尚針對連續型 PTF 與點型 PTF 的不確定性進行比較。額外建立一個本土連續型 PTF，其解釋變數與本土點型 PTF(本研究分析之三種 PTF)相同皆為 5 個，反應變數則改為 VG model 的 4 個參數。接著採取無母數自助法(與 Rosetta3 相同為 1000 自助樣本)，評估本土連續型 PTF 轉換參數的 95%預測區間(為求計算方便，此處考慮 4 個參數為互相獨立)。再以此預測區間計算參數的實際涵蓋率，如表 4.12 所示。

表 4.12 本土連續 PTF 之 95%預測區間的實際涵蓋率表。前四欄為考慮單一參數下，4 個不同參數之實際擬合值，分別介於該參數之預測區間內的實際涵蓋率；最後一欄則為同時考慮 4 個參數下，4 個參數之實際擬合值，皆分別介於 4 個參數之預測區間內的實際涵蓋率。

	θ_s	θ_r	$\text{Log}(\alpha)$	$\text{Log}(n)$	ALL
本土連續 PTF	94.95%	94.70%	95.71%	94.07%	83.33%

從表 4.12 中可發現，無母數自助法準確評估參數預測區間(95±1%)，與先前表 4.9 的結果相似。唯一不同之處為，VG model 需要 4 個參數才可完整描述土壤保水曲線。將 VG model 應用於預測特定基質勢能下的土壤含水量時，也需要 4 個參數才可完整描述。因此本研究尚追加分析 4 個 VG model 參數皆介於預測區間的實際涵蓋率。若此 4 個參數彼此互相獨立，則理想涵蓋率約為 $(95\%)^4 = 81.45\%$ 。從表 4.12 的最後一欄可發現實際涵蓋率略大於理想涵蓋率，此即暗示 4 種參數彼此具有相關性而不完全為互相獨立。

若選擇測試集其中一筆土壤(編號 01S2FK00021)作為範例，結合本土點型與連續型 PTF 的預測與不確定性分析結果，如圖 4.16 所呈現。

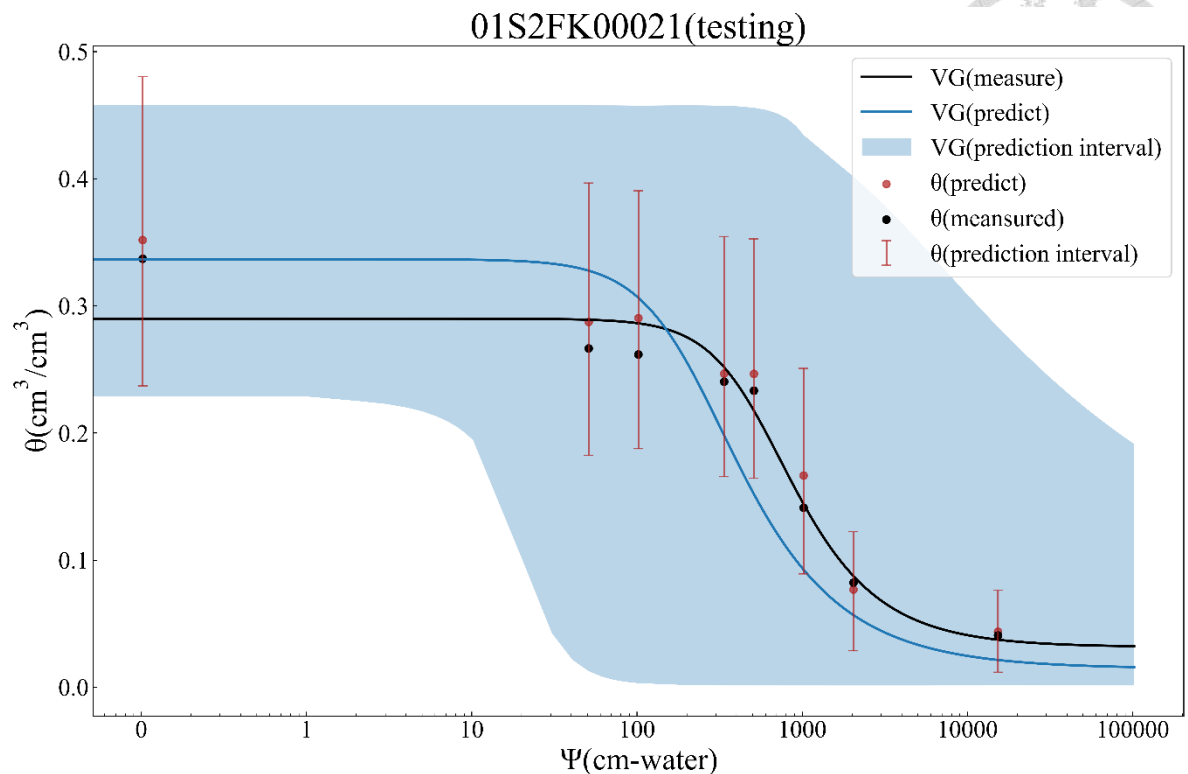


圖 4.16 土壤含水量與基質勢能關係圖。圖標題 01S2FK00021 為測試集中一筆土壤資料的編號。橫軸為基質勢能的絕對值，單位為公分水柱(cm-water)。縱軸為土壤含水量，單位為體積含水量(cm³/cm³)。黑點為實際土壤含水量，黑色線為以實際土壤含水量所擬合之 VG model 曲線。藍色線為本土連續型 PTF 預測的 VG model 曲線。紅點為本土點型 PTF 預測的土壤含水量，紅色誤差線則代表應用無母數自助法評估之 95%預測區間。藍色區塊則為考慮 4 個 VG model 參數，皆介於連續型 PTF 之預測區間時，所有參數組合預測之土壤含水量上下限。(依照對數尺度將基質勢能絕對值的範圍 0~100000 以等間距的方式，離散為 500 個點再求取每個點在給定參數預測區間時的土壤含水量最大、小值)

觀察圖 4.16 中可以發現，點型與連續型 PTF 的預測結果接近(紅點與藍色線)，然而其不確定性則不盡相同。隨著基質勢能減少，土壤含水量各點 95%預測區間逐漸縮小；參數 95%預測區間所涵蓋的土壤含水量範圍，則在曲線轉折處明顯增加。整體而言除了土壤飽和含水量外，參數 95%預測區間所涵蓋的土壤含水量範圍，皆較土壤含水量各點 95%預測區間寬廣。先前的 4.3.4 小節中，已評估土壤含水量各點 95%預測區間的實際涵蓋率皆為 95±1%。因此可以合理推斷，參數 95%預測區間所涵蓋的土壤含水量範圍，其實際涵蓋率將大於 95±1%。此即說明，若

應用連續型 PTF 進行土壤保水曲線轉換，其參數不確定性可能高估個別土壤含水量的不確定性。

若進一步探究導致此現象的原因，則可能與 VG model 曲線函數有關。VG model 屬於多變量非線性函數，因此若欲準確計算在特定含水量的不確定性，則應考量 4 個參數的聯合分佈，而非依序考量個別參數的分佈。若僅考量個別參數的分佈，則可能因參數之間的相關性程度高而高估最終的預測區間範圍。然而 4 個參數的聯合分佈不易估計，也成為連續型 PTF 評估特定基質勢能下，土壤含水量預測區間的最大障礙。

第5章 結論與建議



5.1 結論

本研究以台灣中部農業土壤作為示範，挑選土壤質地(砂、粉、黏含量)、總體密度與有機質含量作為五個解釋變數，且應用多變量線性迴歸、隨機森林和類神經網路共三種不同方式，分別建立台灣本土土壤保水曲線 PTF，並依序探討本土 PTF 的準確性與泛用性、參數重要性以及不確定性。

透過切割資料集的方式，本研究以訓練集、測試集資料分別衡量本土 PTF 的轉換準確性與泛用性。研究結果指出，應用隨機森林建立之 PTF，其準確性與泛用性最高；而根據類神經網路建立之 PTF，雖然其準確性優於以多變量線性迴歸建立之 PTF，然而其泛用性卻較其他兩種本土 PTF 低，本研究推測可能與其具有較多待擬合參數和較高的自由度有關。

本研究透過排列重要性的計算，發現多變量線性迴歸與隨機森林所建立的台灣本土 PTF，其解釋變數具有類似的變化趨勢：在基質勢能較大時，總體密度為影響土壤含水量最重要的參數；隨著基質勢能減少總體密度的重要性逐漸下降，取而代之的是粉粒與有機質含量的重要性逐漸上升。

當 PTF 不為線性模型時，自助法可提供一種有效且準確評估信賴區間與預測區間的方法。本研究應用無母數自助法，評估預測誤差期望值之信賴區間，以及單一預測值的預測區間。透過土壤含水量的實際涵蓋率作為檢核標準，確認無母數自助法對於非線性模型的預測區間估計相當精準(測試集的預測區間幾乎皆為 $95\pm1\%$)。

最後，本研究使用國內外廣為應用的 PTF—Rosetta3 轉換台灣中部農業土壤，並與先前建立之三種本土 PTF 進行比較。研究結果指出，三種本土 PTF 的土壤保水曲線轉換 RMSE 平均值與中位數皆僅有 Rosetta3 的 1/3，且其分佈皆較 Rosetta3 小且集中，此結果也再一次說明建立台灣本土 PTF 的必要性。

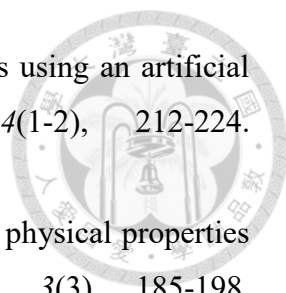
5.2 建議

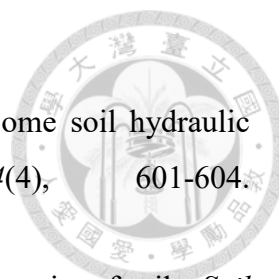
1. 本研究依據台灣中部農業土壤建置本土 PTF，尚缺少台灣其他區域的土壤資料。若應用於其他區域進行土壤轉換，仍須更進一步的研究與討論。未來若有新量測之數據(例：北部、南部資料等)也可依照本研究分析流程，重新訓練、更新本土 PTF 以提升轉換結果的泛用性。
2. 本研究建置之本土 PTF 為點型 PTF，轉換結果為各基質勢能的土壤含水量。若欲將此結果應用於灌溉或水文模式，必須先選擇合適的土壤保水曲線函數(例：VG-model、BC-model)，再進一步以轉換結果搭配非線性規劃求得函數之參數，並以此參數進行模式後續應用。
3. 本研究建置之本土 PTF 輸入變數為土壤基本性質，對於預測土壤含水量能力有限，僅能詮釋土壤含水量部分變異性。本研究尚以田間含水量、永久凋萎點作為兩個額外輸入變數，發現可顯著提升本土 PTF 的轉換結果。未來即可根據轉換需求，選擇是否額外蒐集田間含水量、永久凋萎點以增進轉換結果。
4. 傳統量測土壤保水曲線過程繁複而耗時，且完成後也無法檢核土壤保水曲線是否因為實驗操作不當而產生偏差。本研究所推估之土壤含水量預測區間，以統計理論衡量土壤含水量的實際涵蓋範圍，可作為實驗量測曲線的檢核依據。未來即可透過土壤基本性質，評估傳統量測的土壤保水曲線可靠性，改善台灣土壤資料庫數據品質。

參考文獻



- Abdelbaki, A. M. (2020), Assessing the best performing pedotransfer functions for predicting the soil-water characteristic curve according to soil texture classes and matric potentials, *European Journal of Soil Science*, 72(1), 154-173. <https://doi.org/10.1111/ejss.12959>
- Ahuja, L., J. Naney, R. Williams (1985), Estimating soil water characteristics from simpler properties or limited data, *Soil Science Society of America Journal*, 49(5), 1100-1105. <https://doi.org/10.2136/sssaj1985.03615995004900050005x>
- Ahuja, L. R., D. Swartzendruber (1972), An improved form of soil-water diffusivity function, *Soil Science Society of America Journal*, 36(1), 9-14. <https://doi.org/10.2136/sssaj1972.03615995003600010002x>
- Aina, P., S. Periaswamy (1985), Estimating available water-holding capacity of western Nigerian soils from soil texture and bulk density, using core and sieved samples, *Soil Science*, 140(1), 55-58. <https://doi.org/10.1097/00010694-198507000-00007>
- Akinwande, M. O., H. G. Dikko, A. Samson (2015), Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis, *Open journal of statistics*, 5(07), 754. <https://doi.org/10.4236/ojs.2015.57075>
- Allison, P. D. (1999), *Multiple regression: A primer*, Pine Forge Press.
- Amorim, R. S. S., J. A. Albuquerque, E. G. Couto, M. Kunz, M. F. Rodrigues, L. d. C. M. da Silva, J. M. Reichert (2022), Water retention and availability in Brazilian Cerrado (neotropical savanna) soils under agricultural use: Pedotransfer functions and decision trees, *Soil and Tillage Research*, 224, 105485. <https://doi.org/10.1016/j.still.2022.105485>
- Arya, L., T. S. Dierolf (1992), Predicting soil moisture characteristics from particle-size distributions: An improved method to calculate pore radii from particle radii, paper presented at Proceedings of the International Workshop on Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils, University of California Press Riverside, CA.
- Arya, L. M., J. F. Paris (1981), A physicoempirical model to predict the soil moisture characteristic from particle-size distribution and bulk density data, *Soil Science Society of America Journal*, 45(6), 1023-1030. <https://doi.org/10.2136/sssaj1981.03615995004500060004x>

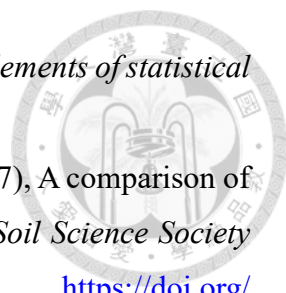
- 
- Baker, L., D. Ellison (2008), Optimisation of pedotransfer functions using an artificial neural network ensemble method, *Geoderma*, 144(1-2), 212-224.
<https://doi.org/10.1016/j.geoderma.2007.11.016>
- Banin, A., A. Amiel (1970), A correlative study of the chemical and physical properties of a group of natural soils of Israel, *Geoderma*, 3(3), 185-198.
[https://doi.org/10.1016/0016-7061\(70\)90018-2](https://doi.org/10.1016/0016-7061(70)90018-2)
- Berndt, R., K. Coughlan (1977), The nature of changes in bulk density with water content in a cracking clay, *Soil Research*, 15(1), 27-37. <https://doi.org/10.1071/SR9770027>
- Berry, W. D., W. D. Berry, S. Feldman, D. Stanley Feldman (1985), *Multiple regression in practice*, Sage. <https://doi.org/10.4135/9781412985208>
- Bloemen, G. W. (1980), Calculation of Hydraulic Conductivities of Soils from Texture and Organic matter Content, *Zeitschrift für Pflanzenernährung und Bodenkunde*, 143(5), 581-605. <https://doi.org/10.1002/jpln.19801430513>
- Bouma, J. (1989), Using soil survey data for quantitative land evaluation, *Advances in soil science*, 9(1989), 177-213. https://doi.org/10.1007/978-1-4612-3532-3_4
- Breiman, L. (1984), *Classification and regression trees*, Routledge.
<https://doi.org/10.2307/2530946>
- Breiman, L. (2001), Random forests, *Machine learning*, 45, 5-32.
- Brooks, R., A. Corey (1964), Hydraulic properties of porous media. Hydrology Paper No. 3, *Civil Engineering Department, Colorado State University, Fort Collins, CO*.
- Bruand, A., O. Duval, H. Gaillard, R. Darthout, M. Jamagne (1996), Variabilité des propriétés de rétention en eau des sols: importance de la densité apparente, *Etude et Gestion des sols*, 31(1), (1) 27-40.
- Campbell, G., S. Shiozawa (1992), Prediction of hydraulic properties of soils using particle-size distribution and bulk density data, *Indirect methods for estimating the hydraulic properties of unsaturated soils*, 317-328.
- Cassel, D., D. Nielsen (1986), Field capacity and available water capacity, *Methods of soil analysis: part 1 physical and mineralogical methods*, 5, 901-926.
<https://doi.org/10.2136/sssabookser5.1.2ed.c36>
- Christiaens, K., J. Feyen (2001), Analysis of uncertainties associated with different methods to determine soil hydraulic properties and their propagation in the distributed hydrological MIKE SHE model, *Journal of hydrology*, 246(1-4), 63-81.

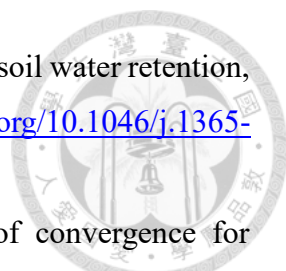


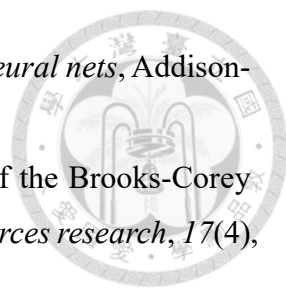
- [https://doi.org/10.1016/S0022-1694\(01\)00345-6](https://doi.org/10.1016/S0022-1694(01)00345-6)
- Clapp, R. B., G. M. Hornberger (1978), Empirical equations for some soil hydraulic properties, *Water resources research*, 14(4), 601-604.
<https://doi.org/10.1029/WR014i004p00601>
- Colman, E. (1947), A laboratory procedure for determining the field capacity of soils, *Soil Science*, 63(4), 277-284. <https://doi.org/10.1097/00010694-194704000-00003>
- Cornelis, W. M., J. Ronsyn, M. Van Meirvenne, R. Hartmann (2001), Evaluation of pedotransfer functions for predicting the soil moisture retention curve, *Soil Science Society of America Journal*, 65(3), 638-648.
<https://doi.org/10.2136/sssaj2001.653638x>
- Costa, A. d., J. A. Albuquerque, J. A. d. Almeida, A. d. Costa, R. V. Luciano (2013), Pedotransfer functions to estimate retention and availability of water in soils of the state of Santa Catarina, Brazil, *Revista Brasileira de Ciência do Solo*, 37, 889-910.
<https://doi.org/10.1590/S0100-06832013000400007>
- Craney, T. A., J. G. Surles (2002), Model-dependent variance inflation factor cutoff values, *Quality engineering*, 14(3), 391-403. <https://doi.org/10.1081/QEN-120001878>
- Dashtaki, S. G., M. Homaei, H. Khodaverdiloo (2010), Derivation and validation of pedotransfer functions for estimating soil water retention curve using a variety of soil data, *Soil use and management*, 26(1), 68-74. <https://doi.org/10.1111/j.1475-2743.2009.00254.x>
- Davies, B. E. (1974), Loss-on-ignition as an estimate of soil organic matter, *Soil Science Society of America Journal*, 38(1), 150-151.
<https://doi.org/10.2136/sssaj1974.03615995003800010046x>
- Davison, A. C., D. V. Hinkley (1997), *Bootstrap methods and their application*, Cambridge university press. <https://doi.org/10.1017/CBO9780511802843>
- Dexter, A. (2004), Soil physical quality: Part I. Theory, effects of soil texture, density, and organic matter, and effects on root growth, *Geoderma*, 120(3-4), 201-214.
<https://doi.org/10.1016/j.geoderma.2003.09.004>
- Division of Soil Survey (1993), *Soil survey manual*, US Department of Agriculture.
- Draper, N. R., H. Smith (1998), *Applied regression analysis*, John Wiley & Sons.
- Dybowski, R., S. J. Roberts (2001), Confidence intervals and prediction intervals for feed-forward neural networks, edited, Cambridge University Press.

<https://doi.org/10.1017/CBO9780511543494.013>

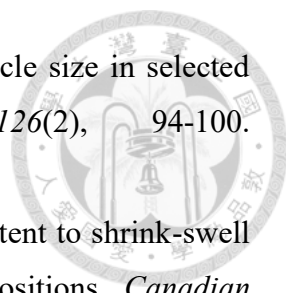
- Efron, B., R. J. Tibshirani (1994), *An introduction to the bootstrap*, CRC press.
- Eiter, T., H. Mannila (1994), Computing discrete Fréchet distance.
- El-Kadi, A. I. (1985), On estimating the hydraulic properties of soil, Part 2. A new empirical equation for estimating hydraulic conductivity for sands, *Advances in water resources*, 8(3), 148-153. [https://doi.org/10.1016/0309-1708\(85\)90055-7](https://doi.org/10.1016/0309-1708(85)90055-7)
- Endelman, F. J., G. E. Box, J. R. Boyle, R. R. Hughes, D. R. Keeney, M. L. Northup, P. G. Saffigna (1974), Mathematical modeling of soil--water--nitrogen phenomena, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).
- F.E. Allison (1973), *Developments in Soil Science*, Elsevier.
- [https://doi.org/https://doi.org/10.1016/S0166-2481\(08\)70587-7](https://doi.org/https://doi.org/10.1016/S0166-2481(08)70587-7)
- Fisher, A., C. Rudin, F. Dominici (2018), All models are wrong but many are useful: variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance, *arXiv preprint arXiv:1801.01489*, 1.
- Fooladmand, H. R. (2011), Pedotransfer functions for point estimation of soil moisture characteristic curve in some Iranian soils, *African Journal of Agricultural Research*, 6(6), 1586-1591.
- Fréchet, M. M. (1906), Sur quelques points du calcul fonctionnel, *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 22(1), 1-72.
- Ghanbarian-Alavijeh, B., H. Millán (2009), The relationship between surface fractal dimension and soil water content at permanent wilting point, *Geoderma*, 151(3-4), 224-232. <https://doi.org/10.1016/j.geoderma.2009.04.014>
- Grismer, M. (1987), Water vapor adsorption and specific surface, *Soil Science*, 144(3), 233-236. <https://doi.org/10.1097/00010694-198709000-00010>
- Guber, A. K., Y. A. Pachepsky (2010), Multimodeling with pedotransfer functions: Documentation and user manual for PTF Calculator (CalcPTF), version 3.0, *USDA Rep., Beltsville Agricultural Research Center, Beltsville, Md*, 26.
- Gupta, S., W. Larson (1979), Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density, *Water resources research*, 15(6), 1633-1635. <https://doi.org/10.1029/WR015i006p01633>
- Hall, D., M. Reeve, A. Thomasson, V. Wright (1977), Water retention, porosity and density of field soils.

- 
- Hastie, T., R. Tibshirani, J. H. Friedman, J. H. Friedman (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer.
- Haverkamp, R., M. Vauclin, J. Touma, P. Wierenga, G. Vachaud (1977), A comparison of numerical simulation models for one-dimensional infiltration, *Soil Science Society of America Journal*, 41(2), 285-294. <https://doi.org/10.2136/sssaj1977.03615995004100020024x>
- Hecht-Nielsen, R. (1989), *Neurocomputing*, Addison-Wesley Longman Publishing Co., Inc.
- Hesterberg, T. (2011), Bootstrap, *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6), 497-526. <https://doi.org/10.1002/wics.182>
- Ho, T. K. (1995), Random decision forests, paper presented at Proceedings of 3rd international conference on document analysis and recognition, IEEE.
- Iman, R. L. (1992), Uncertainty and sensitivity analysis for computer modeling applications, *ASME AEROSP DIV PUBL AD.*, ASME, NEW YORK, NY(USA), 1992, 28, 153-168.
- Ishwaran, H. (2015), The effect of splitting on random forests, *Machine learning*, 99, 75-118. <https://doi.org/10.1007/s10994-014-5451-2>
- Jamison, V., E. Kroth (1958), Available moisture storage capacity in relation to textural composition and organic matter content of several Missouri soils, *Soil Science Society of America Journal*, 22(3), 189-192. <https://doi.org/10.2136/sssaj1958.03615995002200030001x>
- Jaynes, D. B., E. J. Tyler (1984), Using soil physical properties to estimate hydraulic conductivity, *Soil Science*, 138(4), 298-305. <https://doi.org/10.1097/00010694-198410000-00007>
- Khosravi, A., S. Nahavandi, D. Creighton, A. F. Atiya (2011), Comprehensive review of neural network-based prediction intervals and new advances, *IEEE Transactions on neural networks*, 22(9), 1341-1356. <https://doi.org/10.1109/TNN.2011.2162110>
- Kilmer, V. J., L. T. Alexander (1949), Methods of making mechanical analyses of soils, *Soil Science*, 68(1), 15-24. <https://doi.org/10.1097/00010694-194907000-00003>
- Kim, J., T. Kawai, M. Kazama (2019), Minimum void ratio characteristic of soils containing non-plastic fines, *Soils and Foundations*, 59(6), 1772-1786. <https://doi.org/10.1016/j.sandf.2019.08.001>

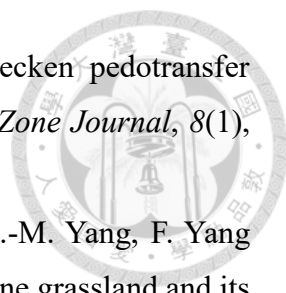
- 
- Koekkoeck, E., H. Booltink (1999), Neural network models to predict soil water retention, *European Journal of Soil Science*, 50(3), 489-495. <https://doi.org/10.1046/j.1365-2389.1999.00247.x>
- Kohler, M., A. Krzyżak, H. Walk (2009), Optimal global rates of convergence for nonparametric regression with unbounded data, *Journal of Statistical Planning and Inference*, 139(4), 1286-1296. <https://doi.org/10.1016/j.jspi.2008.07.012>
- Kumar, S., A. N. Srivistava (2012), Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection, paper presented at The 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.
- LeCun, Y., Y. Bengio (1995), Convolutional networks for images, speech, and time series, *The handbook of brain theory and neural networks*, 3361(10), 1995.
- LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel (1989), Backpropagation applied to handwritten zip code recognition, *Neural computation*, 1(4), 541-551. <https://doi.org/10.1162/neco.1989.1.4.541>
- Li, Y., D. Chen, R. White, A. Zhu, J. Zhang (2007), Estimating soil hydraulic properties of Fengqiu County soils in the North China Plain using pedo-transfer functions, *Geoderma*, 138(3-4), 261-271. <https://doi.org/10.1016/j.geoderma.2006.11.018>
- Lin, H., K. McInnes, L. Wilding, C. Hallmark (1999), Effects of soil morphology on hydraulic properties I. Quantification of soil morphology, *Soil Science Society of America Journal*, 63(4), 948-954. <https://doi.org/10.2136/sssaj1999.634948x>
- Liu, J., Z. Wang, F. Hu, C. Xu, R. Ma, S. Zhao (2020), Soil organic matter and silt contents determine soil particle surface electrochemical properties across a long-term natural restoration grassland, *Catena*, 190, 104526. <https://doi.org/10.1016/j.catena.2020.104526>
- Manchuk, J., O. Leuangthong, C. Deutsch (2009), The proportional effect, *Mathematical Geosciences*, 41, 799-816. <https://doi.org/10.1007/s11004-008-9195-z>
- Maren, A. J., C. T. Harston, R. M. Pap (1990), *Handbook of neural computing applications*, Academic Press. <https://doi.org/10.1016/C2013-0-11292-5>
- Matsuoka, H., T. Luo, Y. Yao (1999), Soil mechanics, *Morikita Shuppan Co., Ltd*, 61.
- McCarty, L. B., L. R. Hubbard, V. L. Quisenberry (2016), *Applied soil physical properties, drainage, and irrigation strategies*, Springer. <https://doi.org/10.1007/978-3-319-24226-2>

- 
- McCord-Nelson, M., W. T. Illingworth (1991), *A practical guide to neural nets*, Addison-Wesley Longman Publishing Co., Inc.
- McCuen, R., W. Rawls, D. Brakensiek (1981), Statistical analysis of the Brooks-Corey and the Green-Ampt parameters across soil textures, *Water resources research*, 17(4), 1005-1013. <https://doi.org/10.1029/WR017i004p01005>
- McKenzie, N., D. Jacquier (1997), Improving the field estimation of saturated hydraulic conductivity in soil survey, *Soil Research*, 35(4), 803-827. <https://doi.org/10.1071/S96093>
- Minasny, B., A. Mc Bratney (2002), Uncertainty analysis for pedotransfer functions, *European Journal of Soil Science*, 53(3), 417-429. <https://doi.org/10.1046/j.1365-2389.2002.00452.x>
- Minasny, B., A. B. McBratney, K. L. Bristow (1999), Comparison of different approaches to the development of pedotransfer functions for water-retention curves, *Geoderma*, 93(3-4), 225-253. [https://doi.org/10.1016/S0016-7061\(99\)00061-0](https://doi.org/10.1016/S0016-7061(99)00061-0)
- Mualem, Y. (1976), A new model for predicting the hydraulic conductivity of unsaturated porous media, *Water resources research*, 12(3), 513-522. <https://doi.org/10.1029/WR012i003p00513>
- Mullins, C. E., K. Smith, C. Mullins (2000), Matric potential, *Soil and environmental analysis: Physical methods. eds ka smith and ce mullins*, 65-93.
- Nemes, A., M. Schaap, J. Wösten (2003), Functional evaluation of pedotransfer functions derived from different scales of data collection, *Soil Science Society of America Journal*, 67(4), 1093-1102. <https://doi.org/10.2136/sssaj2003.1093>
- Or, D., J. M. Wraith, A. Warrick (2002), Soil water content and water potential relationships, *Soil physics companion*, 1, 49-84. <https://doi.org/10.1201/9781420041651-6>
- Pachepsky, Y., W. Rawls, D. Gimenez, J. Watt (1998), Use of soil penetration resistance and group method of data handling to improve soil water retention estimates, *Soil and Tillage Research*, 49(1-2), 117-126. [https://doi.org/10.1016/S0167-1987\(98\)00168-8](https://doi.org/10.1016/S0167-1987(98)00168-8)
- Pachepsky, Y. A., D. Timlin, G. Varallyay (1996), Artificial neural networks to estimate soil water retention from easily measurable data, *Soil Science Society of America Journal*, 60(3), 727-733.

- <https://doi.org/10.2136/sssaj1996.03615995006000030007x>
- Patil, N., D. Pal, C. Mandal, D. Mandal (2012), Soil water retention characteristics of vertisols and pedotransfer functions based on nearest neighbor and neural networks approaches to estimate AWC, *Journal of irrigation and drainage engineering*, 138(2), 177-184. [https://doi.org/10.1061/\(ASCE\)IR.1943-4774.0000375](https://doi.org/10.1061/(ASCE)IR.1943-4774.0000375)
- Patil, N. G., S. K. Singh (2016), Pedotransfer functions for estimating soil hydraulic properties: A review, *Pedosphere*, 26(4), 417-430. [https://doi.org/10.1016/S1002-0160\(15\)60054-6](https://doi.org/10.1016/S1002-0160(15)60054-6)
- Petersen, L., P. Moldrup, O. Jacobsen, D. Rolston (1996), Relations between specific surface area and soil physical and chemical properties, *Soil Science*, 161(1), 9-21. <https://doi.org/10.1097/00010694-199601000-00003>
- Puckett, W., J. Dane, B. Hajek (1985), Physical and mineralogical data to determine soil hydraulic properties, *Soil Science Society of America Journal*, 49(4), 831-836. <https://doi.org/10.2136/sssaj1985.03615995004900040008x>
- Rajkai, K., G. Várallyay (1992), Estimating soil water retention from simpler properties by regression techniques, paper presented at Proc. Int. Workshop on Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils. University of California, Riverside.
- Rajkai, K., S. Kabos, M. T. Van Genuchten (2004), Estimating the water retention curve from soil properties: comparison of linear, nonlinear and concomitant variable methods, *Soil and Tillage Research*, 79(2), 145-152. <https://doi.org/10.1016/j.still.2004.07.003>
- Rawls, W., D. Brakensiek, B. Soni (1983), Agricultural management effects on soil water processes part I: Soil water retention and Green and Ampt infiltration parameters, *Transactions of the ASAE*, 26(6), 1747-1752. <https://doi.org/10.13031/2013.33837>
- Rawls, W. J., D. L. Brakensiek (1985), Prediction of soil water properties for hydrologic modeling, paper presented at Watershed management in the eighties, ASCE.
- Rawls, W. J., D. L. Brakensiek, K. Saxton (1982), Estimation of soil water properties, *Transactions of the ASAE*, 25(5), 1316-1320. <https://doi.org/10.13031/2013.33720>
- Richards, L., L. Weaver (1943), Fifteen-atmosphere percentage as related to the permanent wilting percentage, *Soil Science*, 56(5), 331-340. <https://doi.org/10.1097/00010694-194311000-00002>

- 
- Rivers, E., R. Shipp (1978), Soil water retention as related to particle size in selected sands and loamy sands, *Soil Science*, 126(2), 94-100. <https://doi.org/10.1097/00010694-197808000-00005>
- Ross, G. (1978), Relationships of specific surface area and clay content to shrink-swell potential of soils having different clay mineralogical compositions, *Canadian Journal of Soil Science*, 58(2), 159-166. <https://doi.org/10.4141/cjss78-020>
- Rowell, D. (1994), *Soil science: methods and applications*, Department of Soil Science, University of Reading, Reading, UK.
- Salchow, E., R. Lal, N. Fausey, A. Ward (1996), Pedotransfer functions for variable alluvial soils in southern Ohio, *Geoderma*, 73(3-4), 165-181. [https://doi.org/10.1016/0016-7061\(96\)00044-4](https://doi.org/10.1016/0016-7061(96)00044-4)
- Saxton, K., W. J. Rawls, J. S. Romberger, R. Papendick (1986), Estimating generalized soil-water characteristics from texture, *Soil Science Society of America Journal*, 50(4), 1031-1036. <https://doi.org/10.2136/sssaj1986.03615995005000040054x>
- Saxton, K. E., W. J. Rawls (2006), Soil water characteristic estimates by texture and organic matter for hydrologic solutions, *Soil Science Society of America Journal*, 70(5), 1569-1578. <https://doi.org/10.2136/sssaj2005.0117>
- Schaap, M. G., W. Bouten (1996), Modeling water retention curves of sandy soils using neural networks, *Water resources research*, 32(10), 3033-3040. <https://doi.org/10.1029/96WR02278>
- Schaap, M. G., F. J. Leij (1998), Database-related accuracy and uncertainty of pedotransfer functions, *Soil Science*, 163(10), 765-779. <https://doi.org/10.1097/00010694-199810000-00001>
- Schaap, M. G., F. J. Leij, M. T. Van Genuchten (2001), Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions, *Journal of hydrology*, 251(3-4), 163-176. [https://doi.org/10.1016/S0022-1694\(01\)00466-8](https://doi.org/10.1016/S0022-1694(01)00466-8)
- Schaap, M. G., A. Nemes, M. T. Van Genuchten (2004), Comparison of models for indirect estimation of water retention and available water in surface soils, *Vadose Zone Journal*, 3(4), 1455-1463. <https://doi.org/10.2113/3.4.1455>
- Schelle, H., L. Heise, K. Jänicke, W. Durner (2013), Water retention characteristics of soils over the whole moisture range: A comparison of laboratory methods, *European*

- Journal of Soil Science*, 64(6), 814-821. <https://doi.org/10.1111/ejss.12108>
- Sharma, S. K., B. P. Mohanty, J. Zhu (2006), Including topography and vegetation attributes for developing pedotransfer functions, *Soil Science Society of America Journal*, 70(5), 1430-1440. <https://doi.org/10.2136/sssaj2005.0087>
- Shirazi, M. A., L. Boersma (1984), A unifying quantitative analysis of soil texture, *Soil Science Society of America Journal*, 48(1), 142-147. <https://doi.org/10.2136/sssaj1984.03615995004800010026x>
- Sojka, R., G. Lehrsch, S. Kostka, J. Reed, A. Koehn, J. Foerster (2009), Soil water measurements relevant to agronomic and environmental functions of chemically treated soil, *Journal of the ASTM International*, 6(1), 1-20. <https://doi.org/10.1520/JAI101497>
- Stumpp, C., S. Engelhardt, M. Hofmann, B. Huwe (2009), Evaluation of pedotransfer functions for estimating soil hydraulic properties of prevalent soils in a catchment of the Bavarian Alps, *European journal of forest research*, 128, 609-620. <https://doi.org/10.1007/s10342-008-0241-7>
- Troeh, F. R., L. M. Thompson (2005), *Soils and soil fertility*, Blackwell Iowa.
- Van Alphen, B., H. Booltink, J. Bouma (2001), Combining pedotransfer functions with physical measurements to improve the estimation of soil hydraulic properties, *Geoderma*, 103(1-2), 133-147. [https://doi.org/10.1016/S0016-7061\(01\)00073-8](https://doi.org/10.1016/S0016-7061(01)00073-8)
- van Genuchten, M. T. (1980), A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Science Society of America Journal*, 44(5), 892-898. <https://doi.org/10.2136/sssaj1980.03615995004400050002x>
- Vereecken, H., J. Diels, J. Van Orshoven, J. Feyen, J. Bouma (1992), Functional evaluation of pedotransfer functions for the estimation of soil hydraulic properties, *Soil Science Society of America Journal*, 56(5), 1371-1378. <https://doi.org/10.2136/sssaj1992.03615995005600050007x>
- Wösten, J., Y. A. Pachepsky, W. Rawls (2001), Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics, *Journal of hydrology*, 251(3-4), 123-150. [https://doi.org/10.1016/S0022-1694\(01\)00464-4](https://doi.org/10.1016/S0022-1694(01)00464-4)
- Wösten, J., A. Lilly, A. Nemes, C. Le Bas (1999), Development and use of a database of hydraulic properties of European soils, *Geoderma*, 90(3-4), 169-185. [https://doi.org/10.1016/S0016-7061\(98\)00132-3](https://doi.org/10.1016/S0016-7061(98)00132-3)

- 
- Weynants, M., H. Vereecken, M. Javaux (2009), Revisiting Vereecken pedotransfer functions: Introducing a closed-form hydraulic model, *Vadose Zone Journal*, 8(1), 86-95. <https://doi.org/10.2136/vzj2008.0062>
- Yang, F., G.-L. Zhang, J.-L. Yang, D.-C. Li, Y.-G. Zhao, F. Liu, R.-M. Yang, F. Yang (2014), Organic matter controls of soil water retention in an alpine grassland and its significance for hydrological processes, *Journal of hydrology*, 519, 3086-3093. <https://doi.org/10.1016/j.jhydrol.2014.10.054>
- Zacharias, S., G. Wessolek (2007), Excluding organic matter content from pedotransfer predictors of soil water retention, *Soil Science Society of America Journal*, 71(1), 43-50. <https://doi.org/10.2136/sssaj2006.0098>
- Zhang, Y., M. G. Schaap (2017), Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3), *Journal of hydrology*, 547, 39-53. <https://doi.org/10.1016/j.jhydrol.2017.01.004>
- 吳晟哲 (2010), 土壤水力性質轉換函數之研究—以石門水庫集水區為例, 國立中興大學水土保持學研究所. <https://doi.org/10.6845/NCHU.2010.01263>
- 林可薇 (2012), 以連續土壤轉換函數預測 van Genuchten 模式參數之研究, 國立中興大學水土保持學研究所. <https://doi.org/10.6845/NCHU.2012.00282>
- 林俐玲, 陳威竹, 林可薇, 曹舜評 (2013), 以群集分析加強 van Genuchten 模式參數推估之研究, *中華水土保持學報*, 44(4), 324-334. [https://doi.org/10.29417/JCSWC.201312_44\(4\).0006](https://doi.org/10.29417/JCSWC.201312_44(4).0006)
- 洪靖惠 (2008), 土壤水分特性曲線參數與物理性質關係之研究, 國立中興大學水土保持學研究所. <https://doi.org/10.6845/NCHU.2008.00207>
- 劉滄琴, 彭宗仁, 范家華, 郭鴻裕 (2007), 應用土壤轉換方程式 (PTF) 評估台灣平地土壤之飽和水力傳導度, *西太平洋地質科學*.

附錄

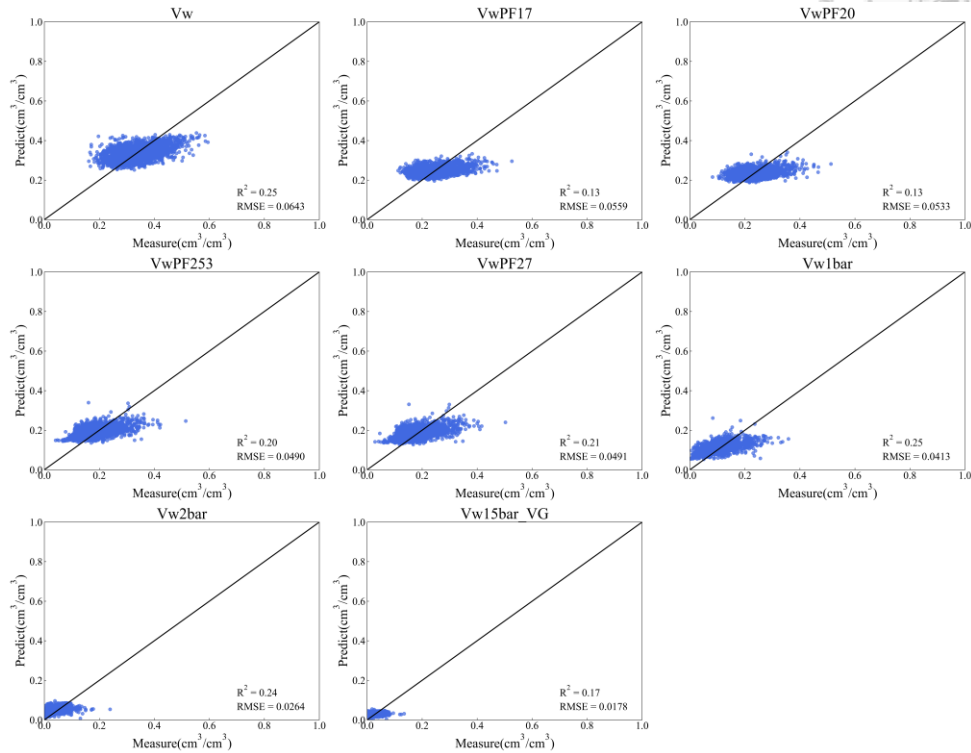
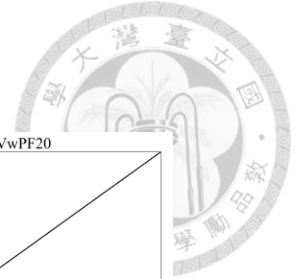


圖 A.1 應用 MLR 轉換台灣土壤含水量結果(訓練集)

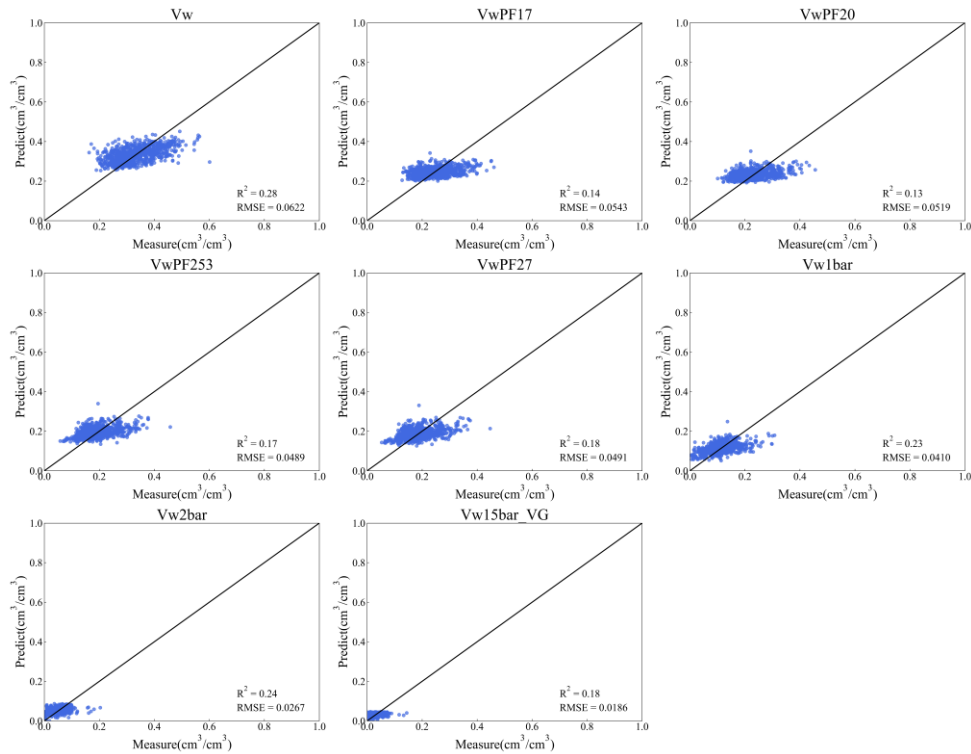


圖 A.2 應用 MLR 轉換台灣土壤含水量結果(測試集)

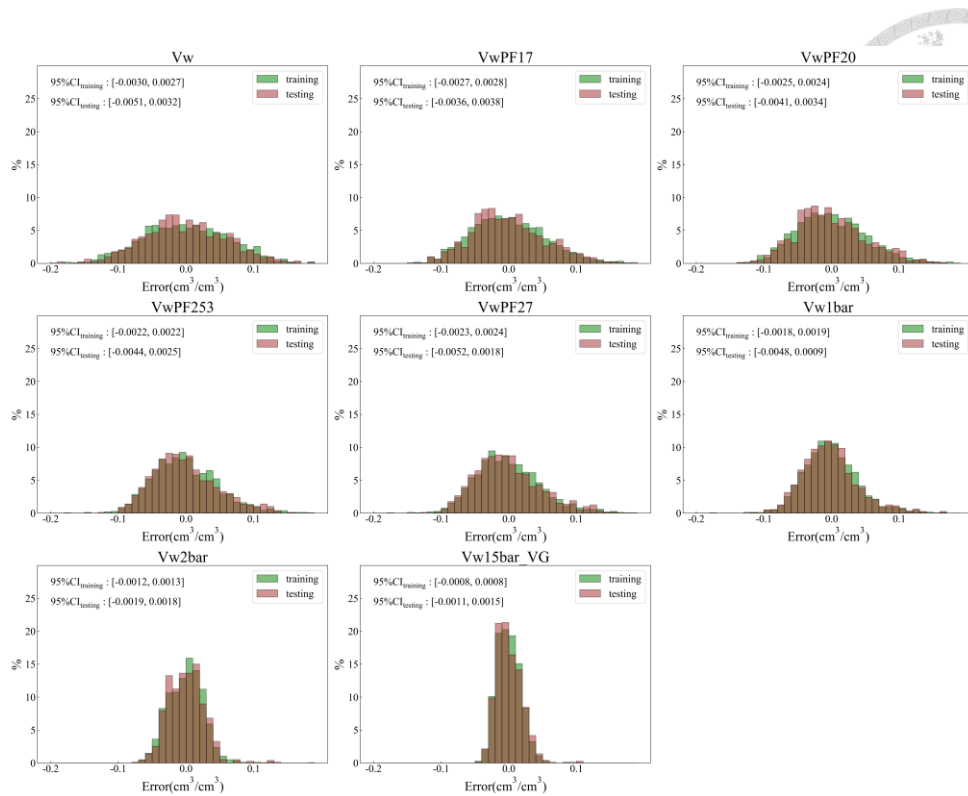


圖 A.3 應用 MLR 轉換台灣土壤含水量的誤差(綠色為訓練集、棕色為測試集)

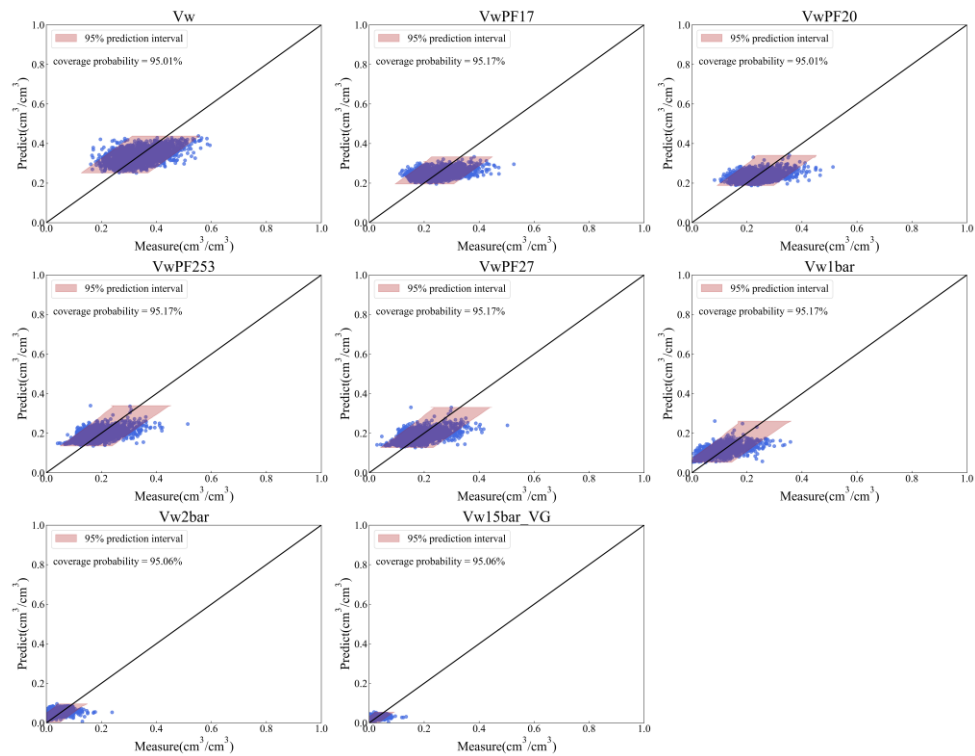


圖 A.4 應用 MLR 轉換台灣土壤含水量的預測區間與實際涵蓋率(訓練集)

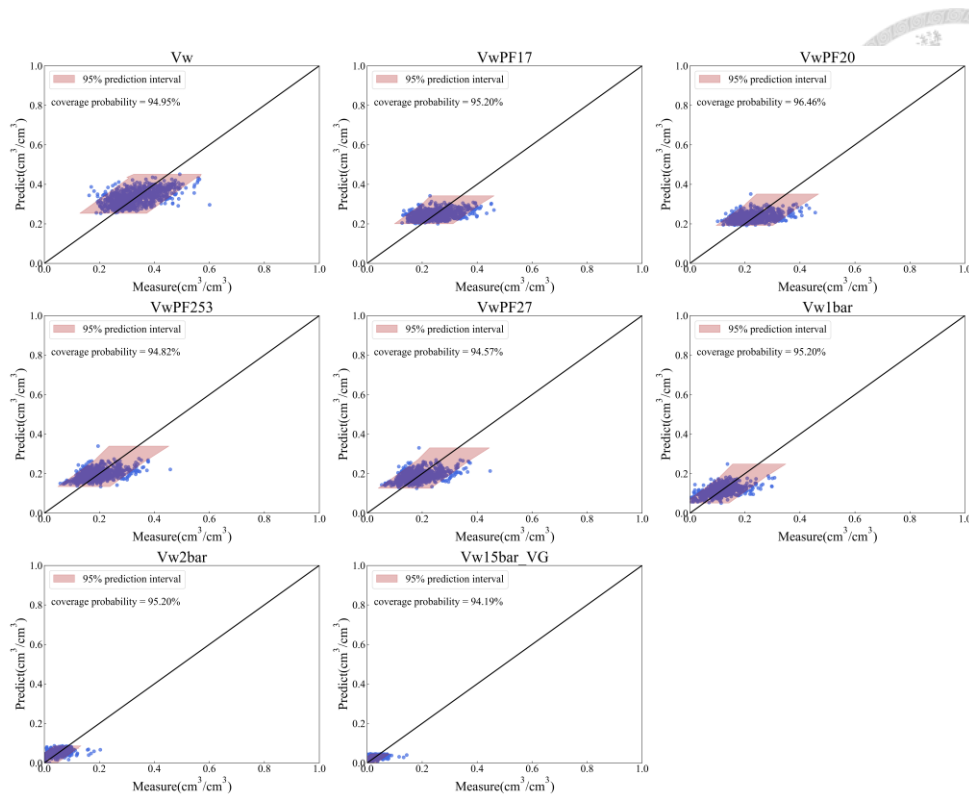


圖 A.5 應用 MLR 轉換台灣土壤含水量的預測區間與實際涵蓋率(測試集)

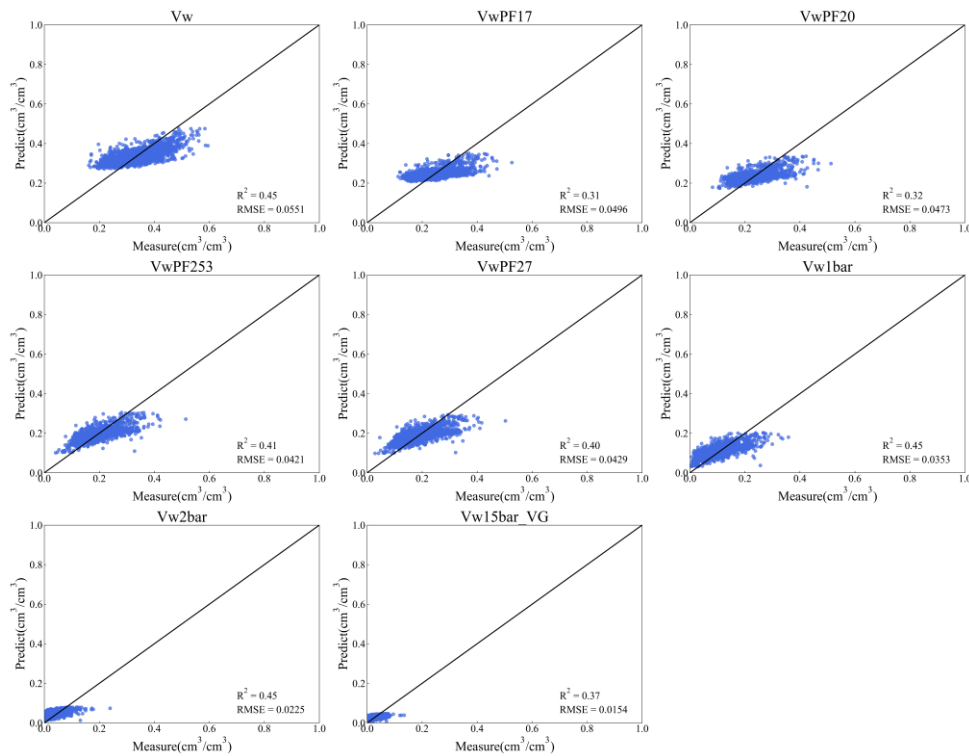


圖 A.6 應用 RF 轉換台灣土壤含水量結果(訓練集)

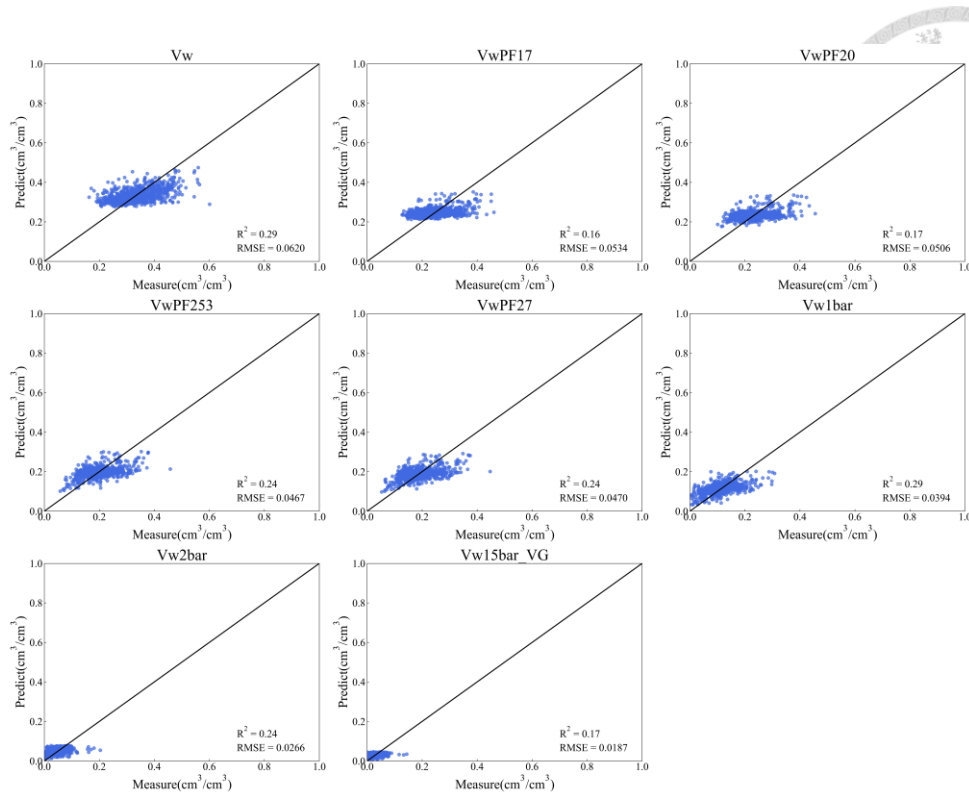


圖 A.7 應用 RF 轉換台灣土壤含水量結果(測試集)

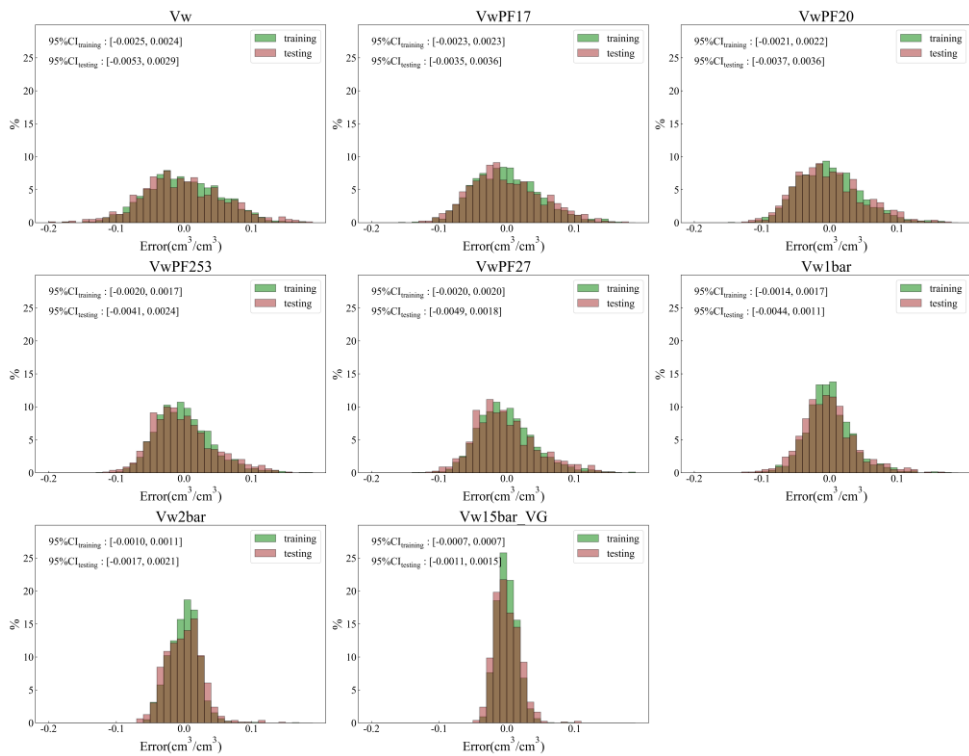


圖 A.8 應用 RF 轉換台灣土壤含水量的誤差(綠色為訓練集、棕色為測試集)

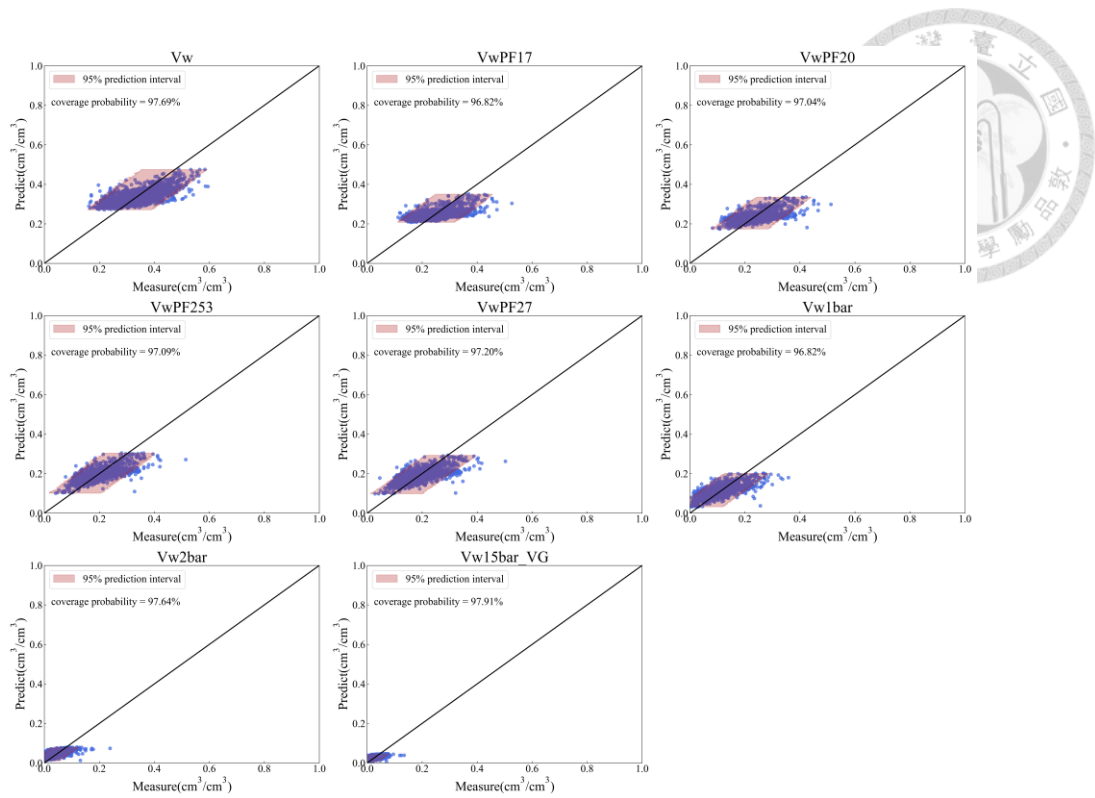


圖 A.9 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(訓練集)

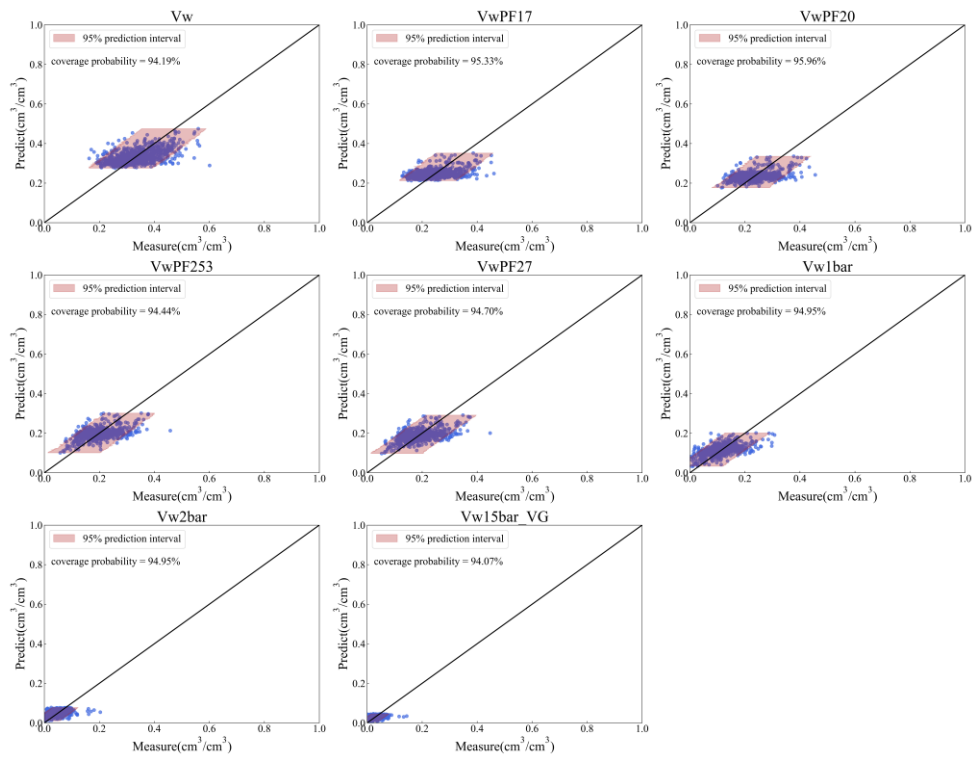


圖 A.10 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(測試集)

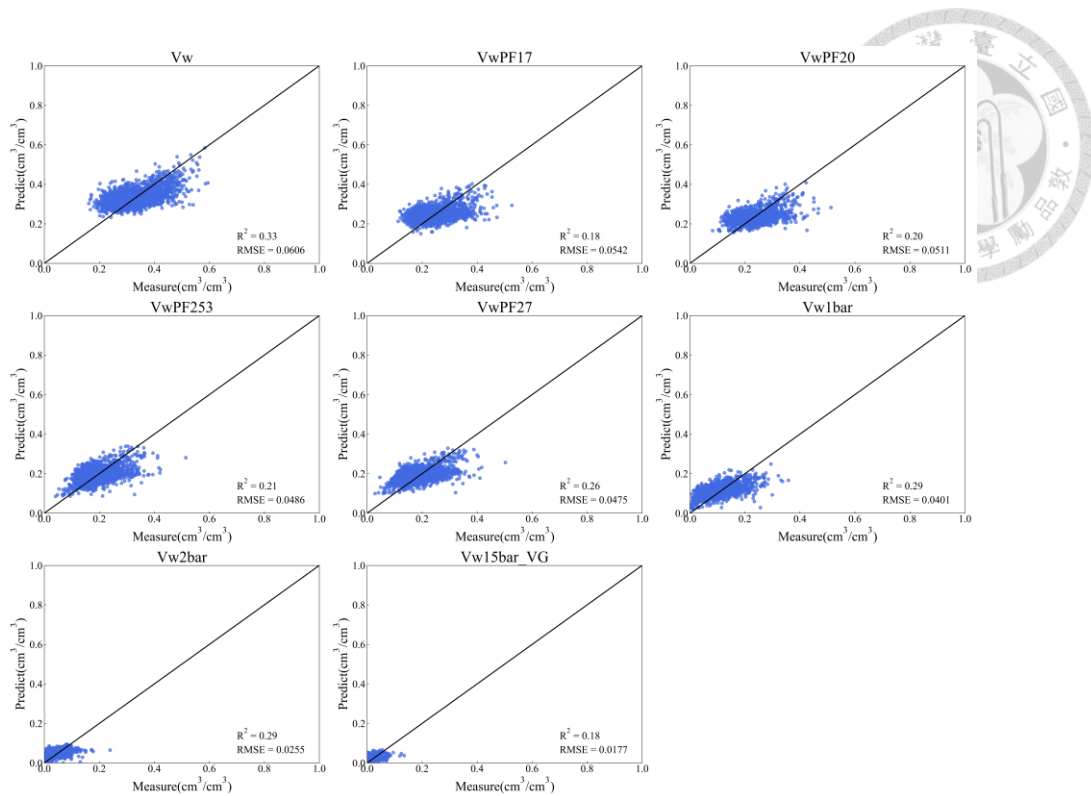


圖 A.11 應用 ANN 轉換台灣土壤含水量結果(訓練集)

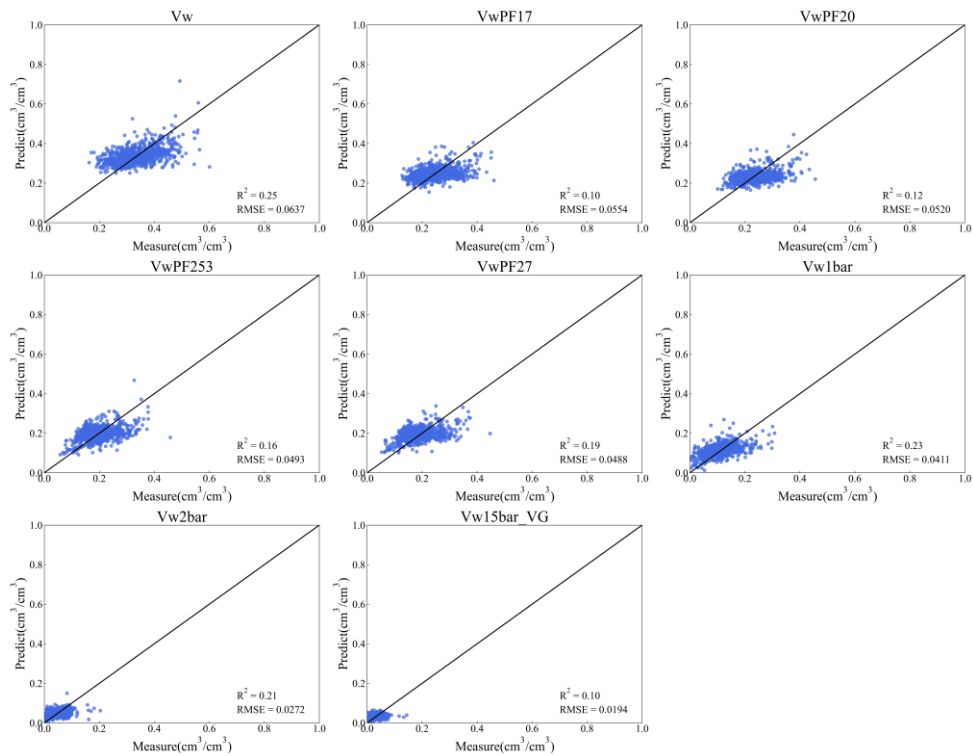


圖 A.12 應用 ANN 轉換台灣土壤含水量結果(測試集)

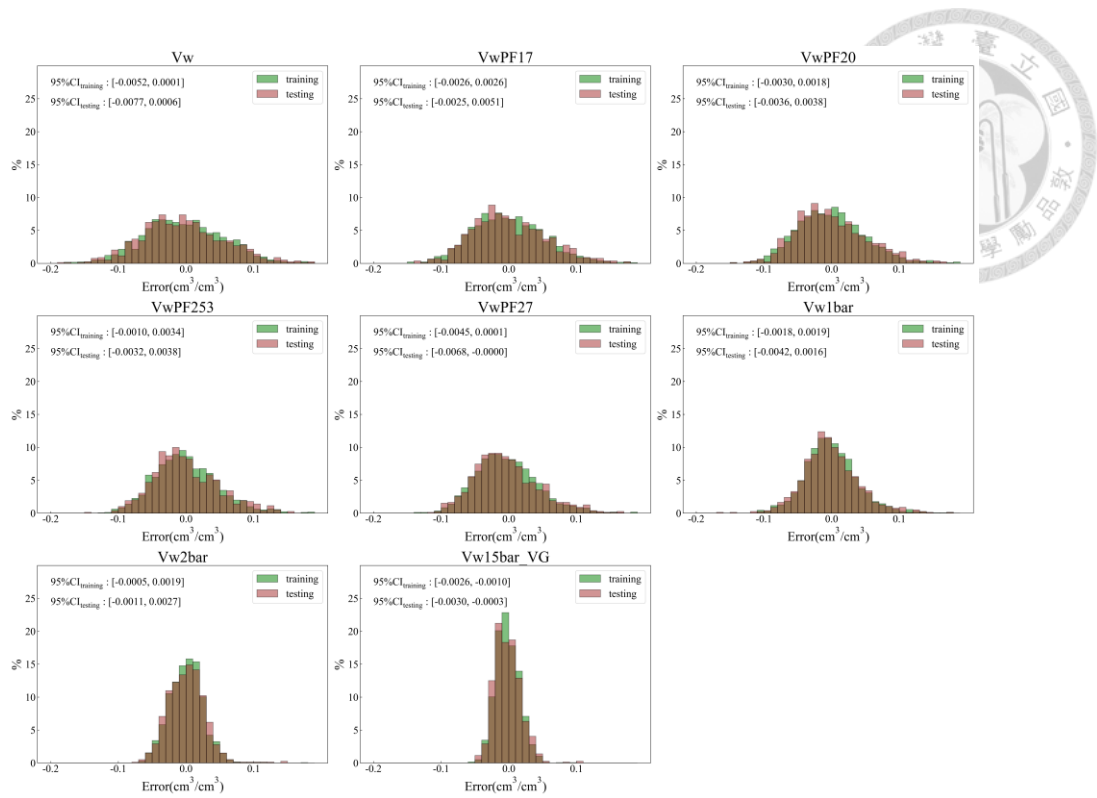


圖 A.13 應用 ANN 轉換台灣土壤含水量的誤差(綠色為訓練集、棕色為測試集)

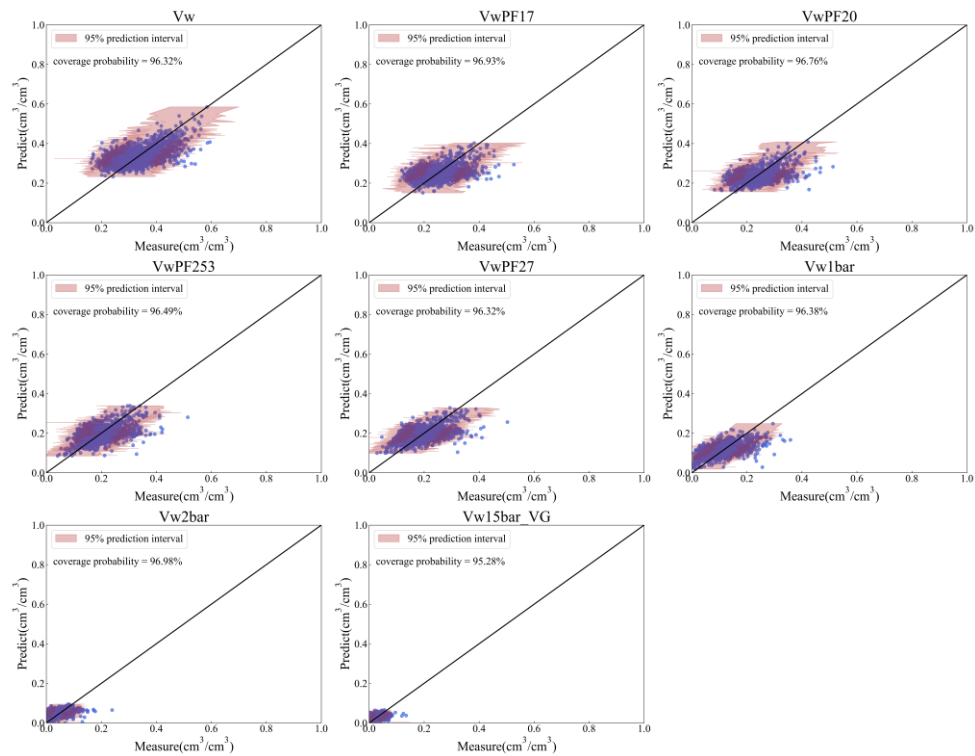


圖 A.14 應用 ANN 轉換台灣土壤含水量的預測區間與實際涵蓋率(訓練集)

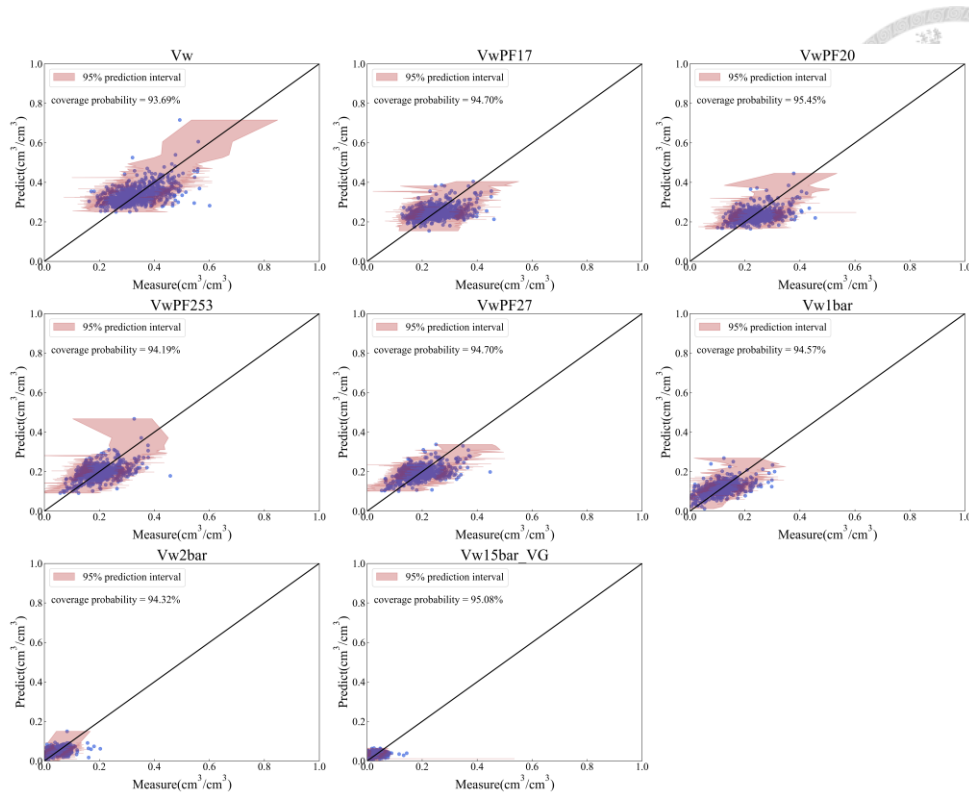


圖 A.15 應用 ANN 轉換台灣土壤含水量的預測區間與實際涵蓋率(測試集)

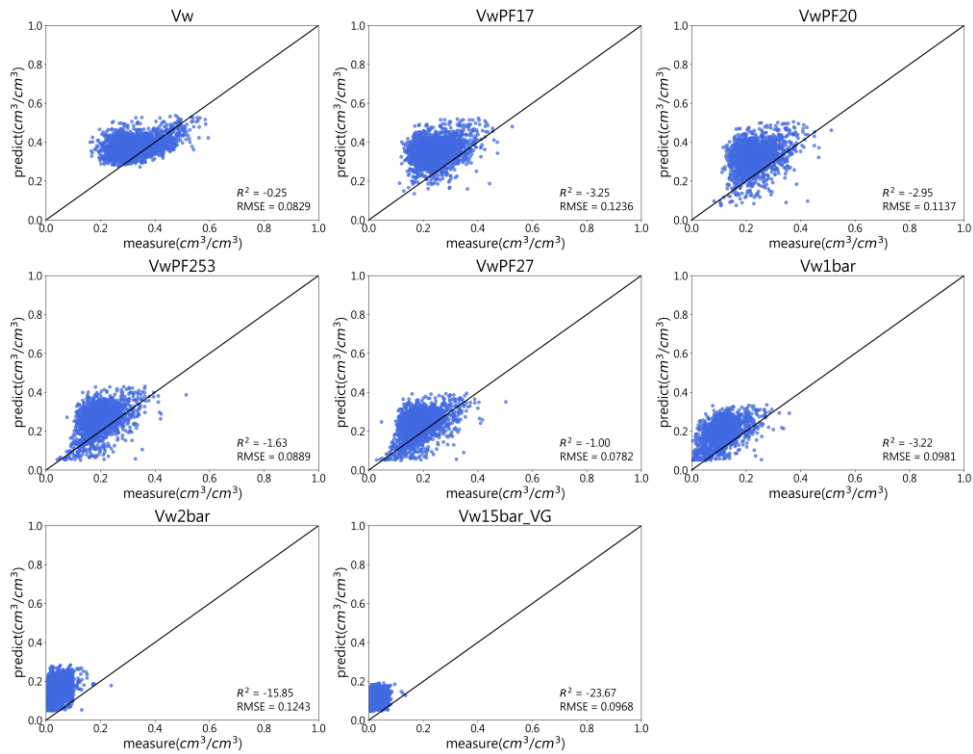


圖 A.16 應用 Rosetta(Zhang & Schaap, 2017)轉換台灣土壤含水量結果(訓練集)

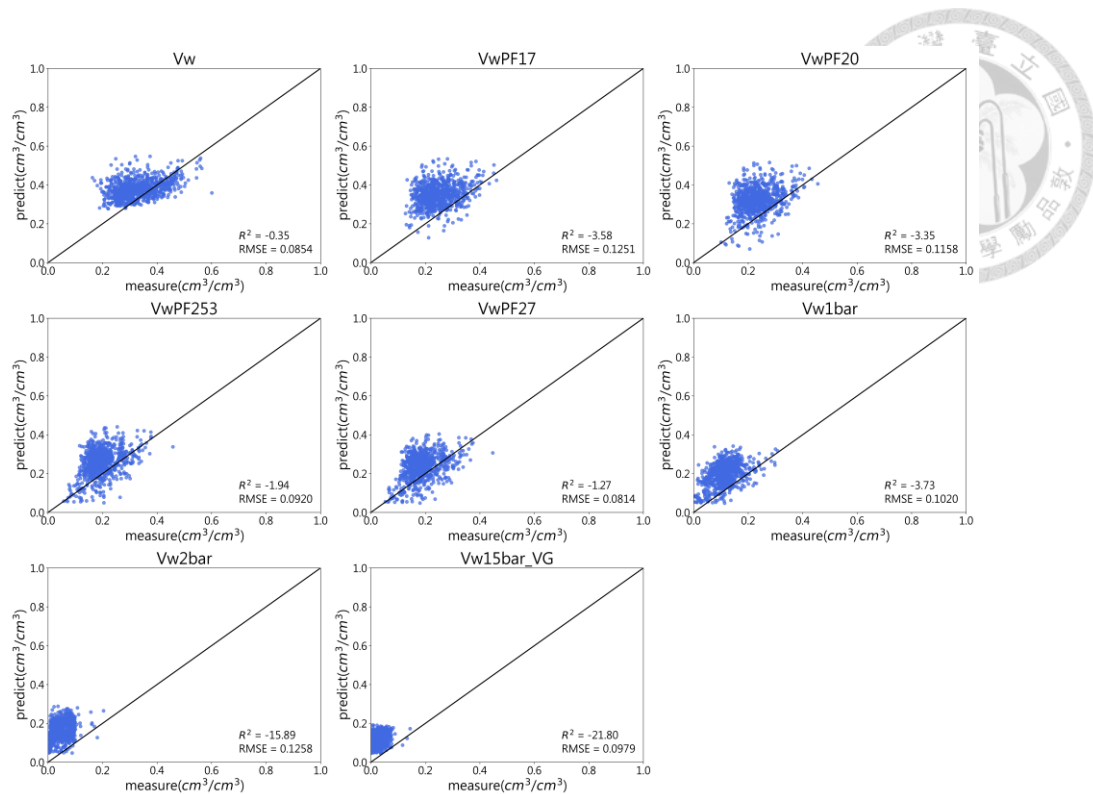


圖 A.17 應用 Rosetta(Zhang & Schaap, 2017)轉換台灣土壤含水量結果(測試集)

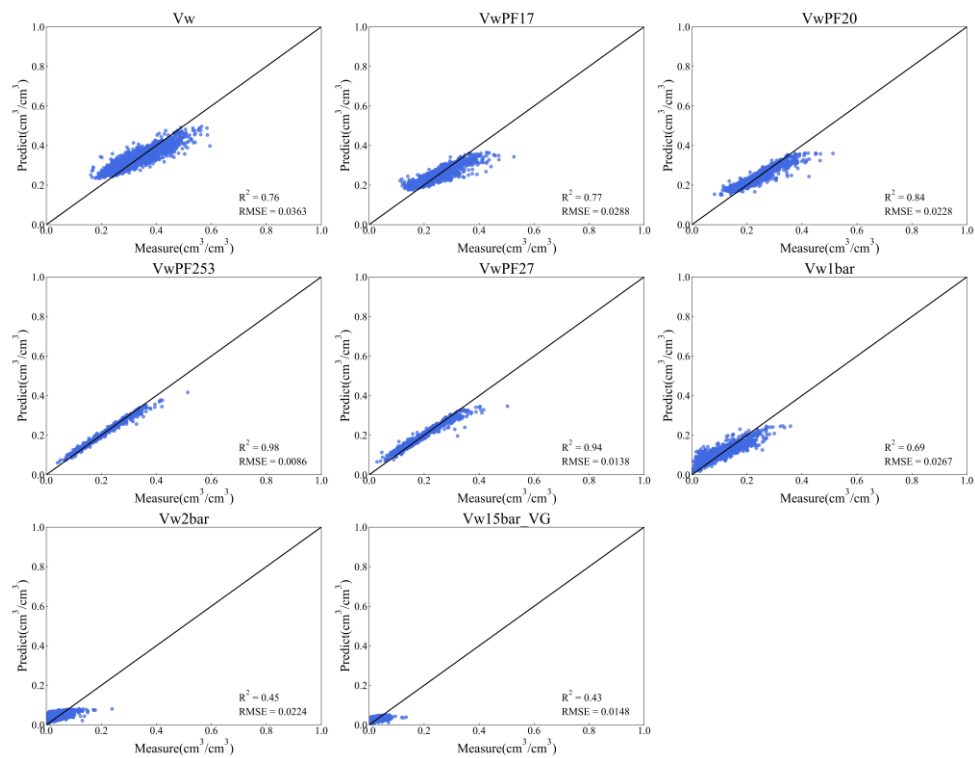


圖 A.18 應用 RF 轉換台灣土壤含水量結果(VwPF253 為額外解釋變數，訓練集)

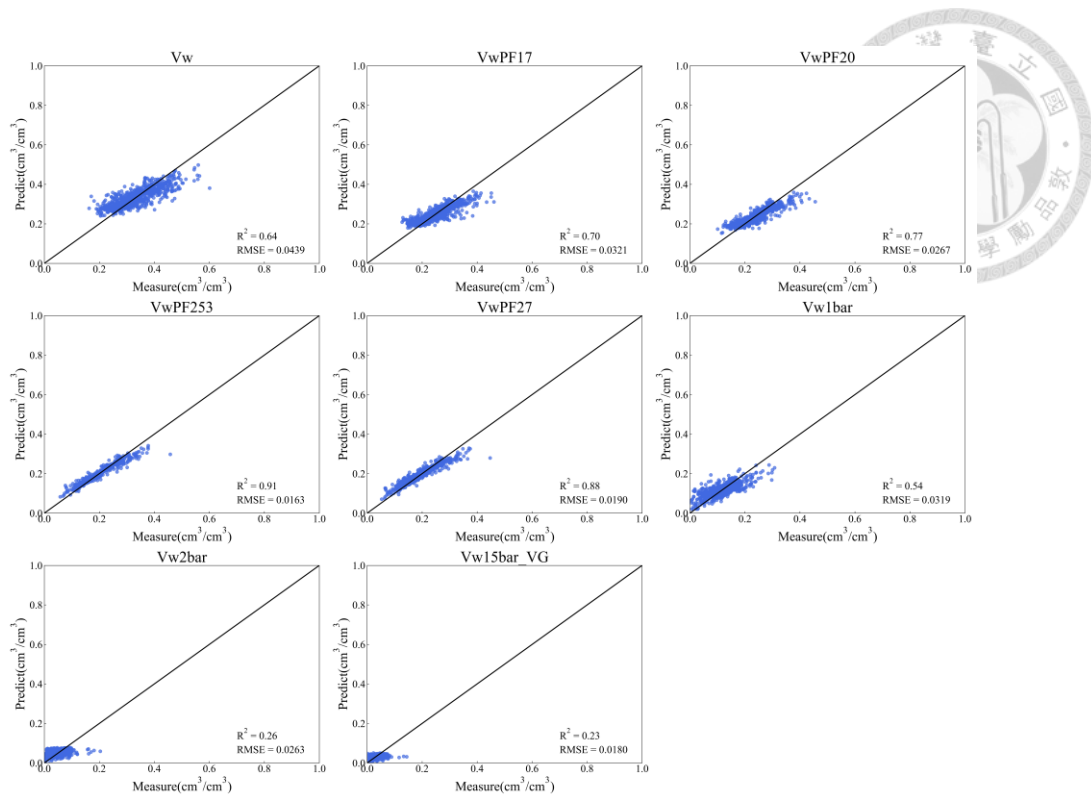


圖 A.19 應用 RF 轉換台灣土壤含水量結果(VwPF253 為額外解釋變數，測試集)

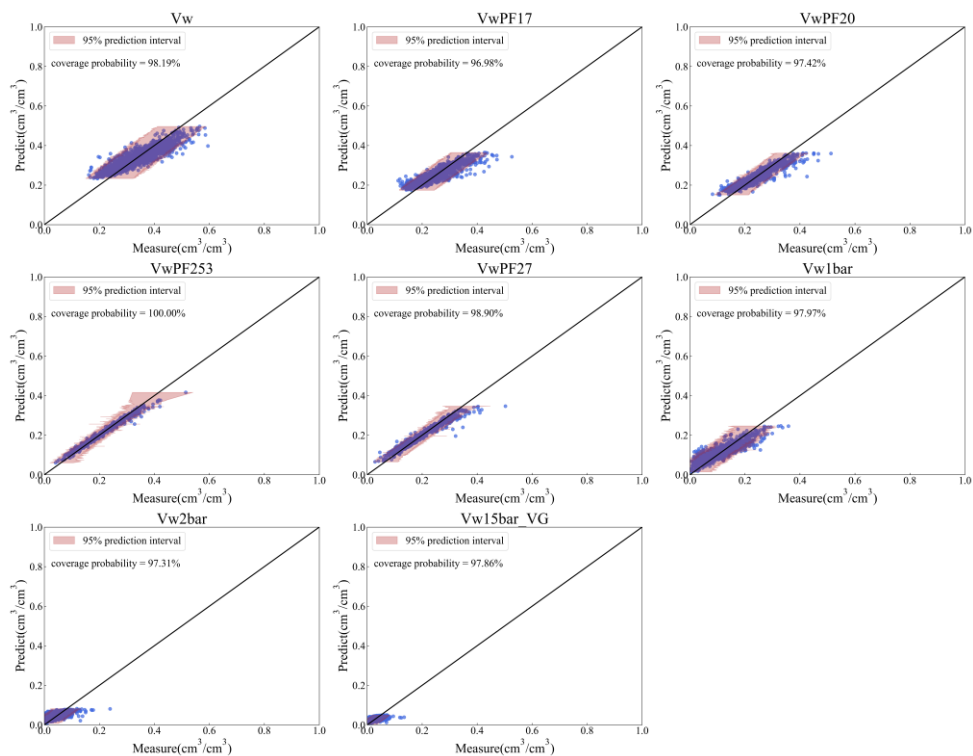


圖 A.20 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(VwPF253 為額外解釋變數，訓練集)

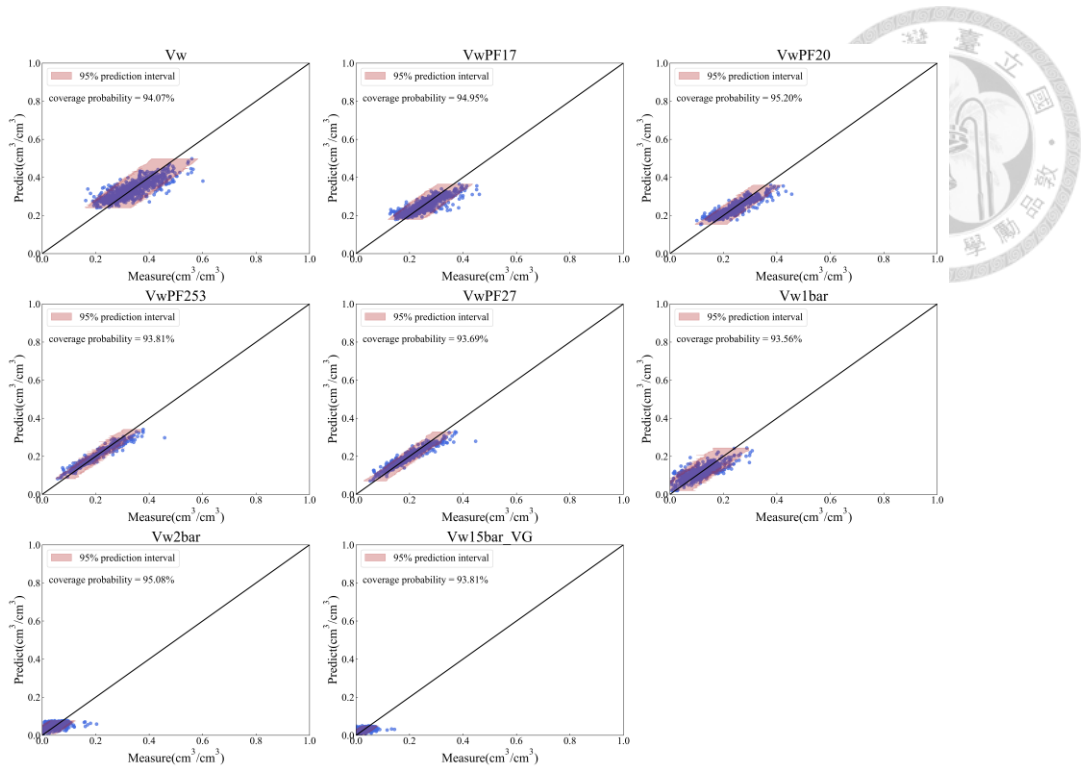


圖 A.21 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(VwPF253 為額外解釋變數，測試集)

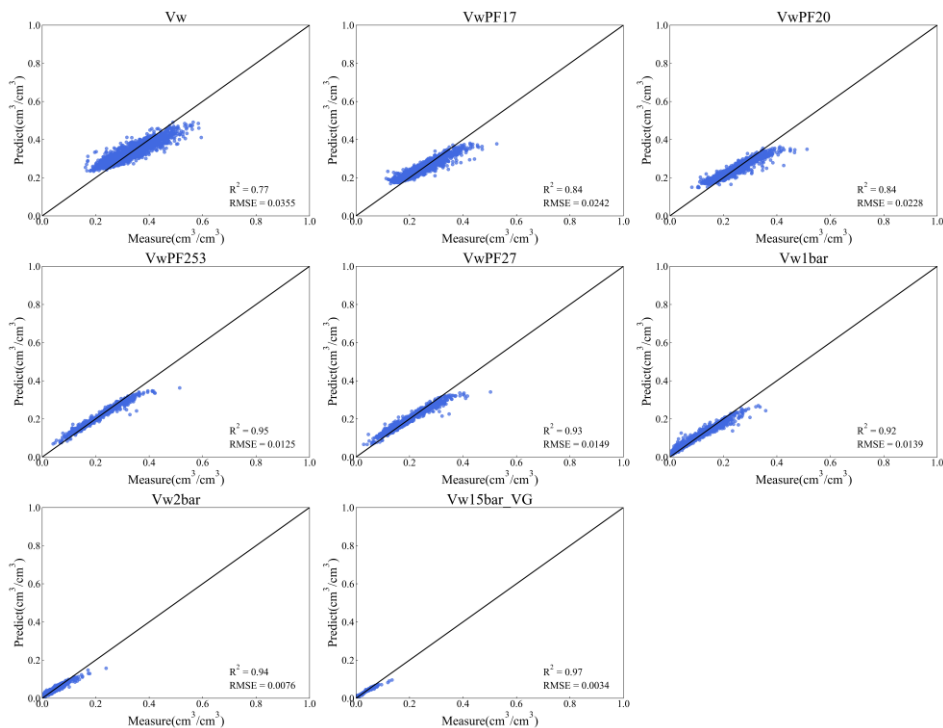


圖 A.22 應用 RF 轉換台灣土壤含水量結果(VwPF253、Vw15bar_VG 為額外解釋變數，訓練集)

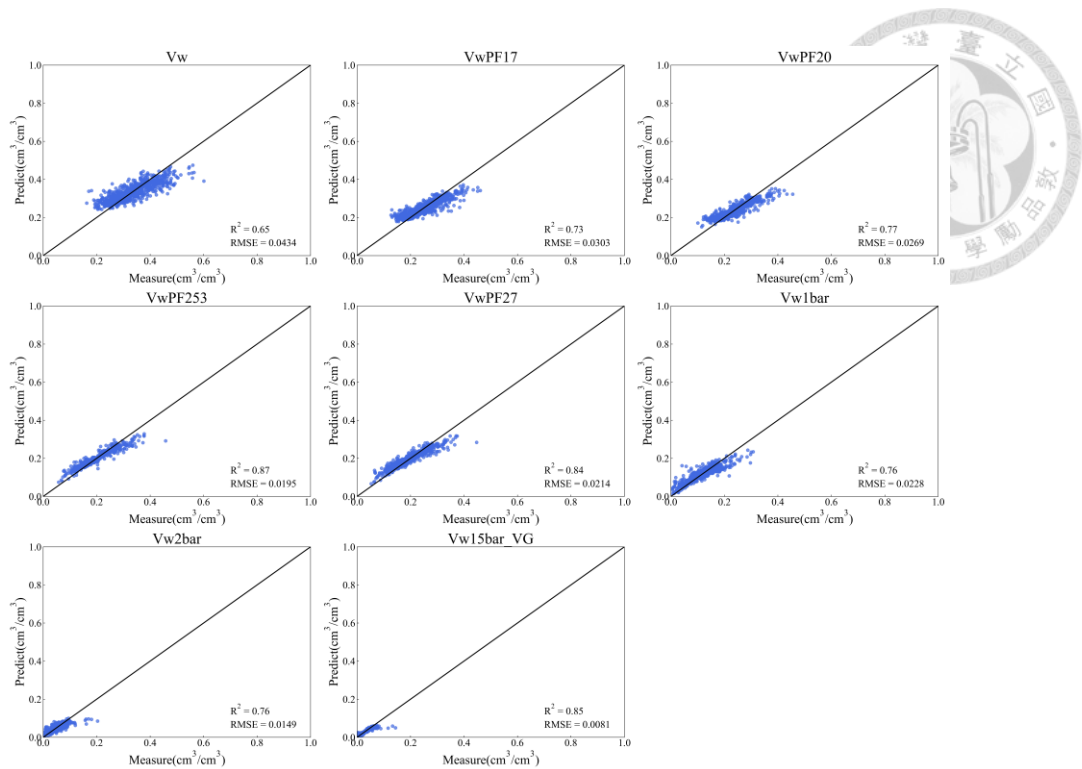


圖 A.23 應用 RF 轉換台灣土壤含水量結果(VwPF253、Vw15bar_VG 為額外解釋變數，測試集)

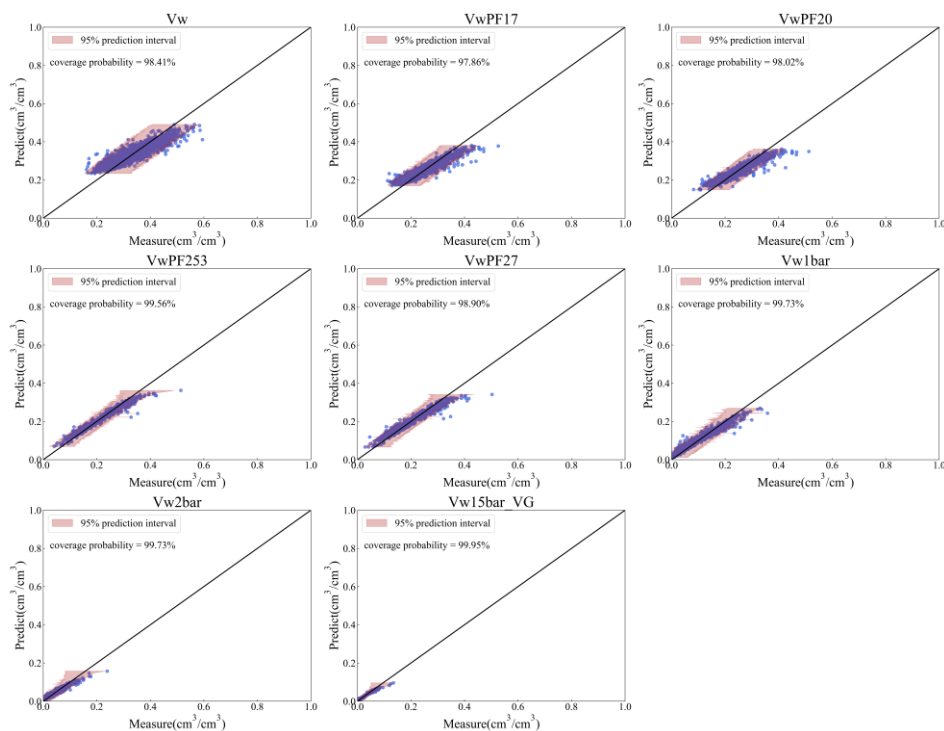


圖 A.24 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(VwPF253、Vw15bar_VG 為額外解釋變數，訓練集)

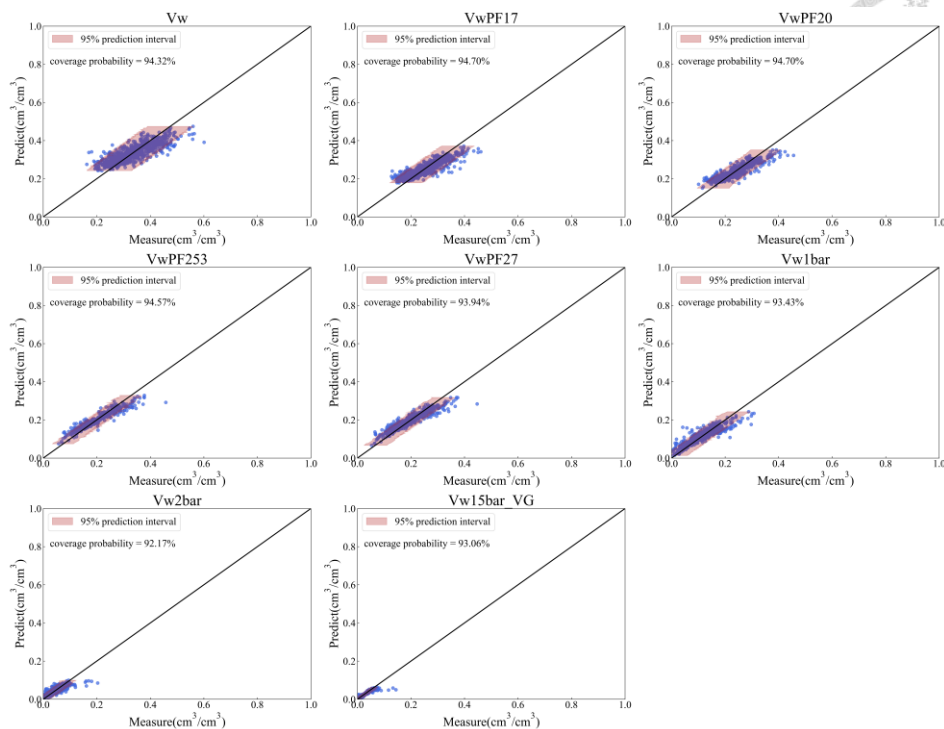


圖 A.25 應用 RF 轉換台灣土壤含水量的預測區間與實際涵蓋率(VwPF253、Vw15bar_VG 為額外解釋變數，測試集)

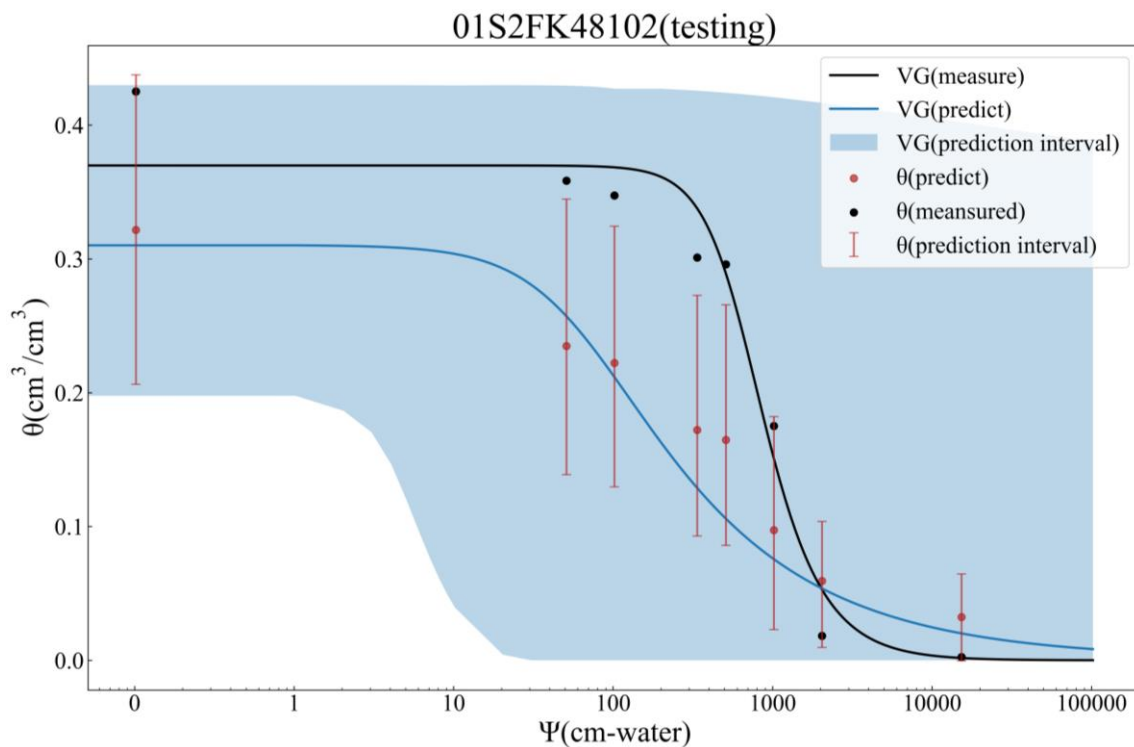


圖 A.26 應用預測區間檢核土壤含水量與基質勢能關係之範例