ARDSWC(113.06 版)

計畫編號：保發-5.1-保-01-06-001(48)

運用深度學習與衛星遙測資料強化土地違規利用判釋

Enhancing the Detection of Unauthorized Land Use by Artificial Intelligence Modeling and Remote Sensing Data

執 行 單 位：國立中央大學

執 行 期 間：113 年 01 月 01 日至 113 年 12 月 31 日

計畫主持人 ：曾國欣 教授

農業部農村發展及水土保持署　編印

中華民國 113 年 12 月

運用深度學習與衛星遙測資料強化土地違規利用判釋

摘要

隨著經濟發展及社會變遷，土地利用型態日趨複雜，為有效防止不當或違法開發，自民國 107 年起政府單位便開始實行了國土利用監測工作，以確保土地資源的合理利用。隨著國土計畫法的完善與國土利用監測整合作業的施行，衛星影像也被運用來加速掌握土地變遷的資訊，國土利用監測至今已然成為協助國家土地管理的重要工具。

然而衛星影像判釋變異點的過程，難以直接區分變異點的合法與非法性，因此往往需耗費人力與各項資源成本實地訪查變異點是否合法，故本計畫提出了一項基於人工智慧技術的解決方案，旨在利用人工智慧模型預測山坡地的變異點是否高機率為違規變異點，以期加速變異點查復效率。為了達成此目標，計畫收集並整合了多種數據來源，包含歷年的山坡地變異點數據，及其他具空間特性之資料，如：高程、坡度、坡向、道路資訊、人口數與人口密度等等。透過整合這些資料並進行資料處理，可順利將各種分散的資料集進行清理與特徵提取，形成有利於人工智慧模型訓練的特徵因子。

透過該項計畫，期望能找出非違規變異點在空間上分佈的相關性，建構出有利於模型進行違規與非違規分類問題的關鍵因素，這將有助於提升對變異點的識別能力，協助相關單位事先排除部分非違規變異點去進行勘查，有效地降低對合法變異點的檢核頻率，以節省寶貴的人力資源並提升查復效率。

關鍵詞：變異點、人工智慧模型、土地管理

Enhancing the Detection of Unauthorized Land Use by Deep Learning and Remote Sensing Data

Abstract

As economic development and social changes progress, land utilization has become increasingly complex. To effectively prevent inappropriate and unauthorized land use, government agencies have implemented National Land Use Monitoring since 2018 to ensure the rational use of land resources. With the improvement of the Spatial Planning Act and the implementation of Integrated National Land Use Change Monitoring, satellite imagery has been utilized to accelerate the detection of land use. National land use monitoring has now become an essential tool in assisting national land management.

Since interpreting variation points from satellite imagery cannot directly distinguish between legal and illegal changes, on-site investigations often require significant human resources and costs to determine legality. Therefore, this project proposes an artificial intelligence-based solution using AI models to predict whether land use in hillside areas is highly likely to be unauthorized. This approach is expected to accelerate the efficiency of land use verification. To achieve this goal, this project has collected and integrated various data sources, including historical variation point datasets and other spatially characteristic information such as elevation, slope, aspect, road information, population, and population density.

By integrating and processing these data, we can successfully clean and extract features from various dispersed datasets, forming advantageous features for AI-model training. Through this project, we aim to identify the spatial distribution correlations of unauthorized land use and then construct important features beneficial for models to classify illegal and legal cases. This will help improve the identification ability of land use, assisting relevant agencies in filtering out authorized development in advance, effectively reducing the frequency of checks on legal variation points, thereby saving valuable human resources and greatly enhancing verification efficiency.

This research project would be helpful in optimizing land resource management and protection, contributing to the sustainable development of national land. It would hold practical application value.

Keywords: Land Use, AI modeling, Land Management

目次

摘要.....I

Abstract.....II

目次.....IV

表次.....VI

圖次.....VII

第一章　緒論 一-1

第一節　計畫介紹 一-1

1.　全程目標..... 一-1

2.　研究範圍..... 一-1

第二節　資料作業與項目 一-2

1.　資料目的..... 一-2

2.　資料收集..... 一-2

(1)　變異點資訊..... 一-2

(2)　數值地形模型..... 一-5

(3)　人口資料..... 一-5

(4)　全臺道路資料..... 一-6

第三節　機器學習模型 一-7

1.　LightGBM..... 一-7

2.　CatBoost..... 一-9

3.　Random Forest (RF) 一-11

第二章　工作執行方法與步驟 二-1

第一節　工作執行方法 二-1

1.　資料前處理..... 二-1

2.　模型建構..... 二-6

3. 數據分析 二-7
第二節 工作執行步驟 二-11
1. 流程圖 二-11
2. 工作內容 二-12
第三章 工作進度與交付項目 三-1
第一節 計畫甘特圖 三-1
第二節 計畫交付項目 三-1
第四章 研究成果 四-1
1. 模型效能評估 四-1
2. 最低違規機率 四-3
3. 顯著因子 四-4
4. 不同時長訓練結果對比 四-5
5. 歷年各模型違規變異點誤判情形 四-7
6. 主成份分析（PCA）成果 四-10
7. 違規變異點機率分布地圖 四-12
第五章 結論與建議 五-31
參考文獻 參-1
附錄 附錄-1
附錄一、期初審查會議紀錄暨回覆辦理情形 1
附錄二、期末聯合審查會議紀錄暨回覆辦理情形 5

表次

表 1-1、每年各縣市變異點數量統計 一-4
表 2-1、資料缺失值比例..... 二-3
表 2-2、每年違規與非違規變異點數量統計 二-5
表 4-1、整體準確度..... 四-1
表 4-2、109 年 RANDOM FOREST 模型預測結果 四-1
表 4-3、生產者精度..... 四-2
表 4-4、最低違規機率..... 四-4
表 4-5、顯著因子..... 四-4
表 4-6、民國 108 年之數據作為訓練集 四-5
表 4-7、以民國 109 年之數據作為訓練集 四-6
表 4-8、以民國 110 年之數據作為訓練集 四-6
表 4-9、以民國 111 年之數據作為訓練集 四-6
表 4-10、以民國 112 年之數據作為訓練集 四-7
表 4-11、各縣市違規變異點誤判之數量統計 四-9

圖次

圖 1-1、變異點分佈圖，(A)民國 108 年、(B)民國 109 年、(C)民國 110 年、(D)民國 111 年、(E)民國 112 年 一-3
圖 1-2、數值地形模型示意圖 一-5
圖 1-3、道路圖資分佈圖 一-7
圖 2-1、將變數標準化（中心化） 二-9
圖 2-2、求出兩個主方向 二-10
圖 2-3、投影到主成份空間 二-10
圖 2-4、工作流程圖 二-11
圖 3-1、計畫甘特圖 三-1
圖 4-1、歷年各縣市變異點每一模型之誤判機率統計 四-8
圖 4-2、主成份投影圖 四-11
圖 4-3、碎石圖 (SCREE PLOT) 與累積解釋變異量圖 四-11
圖 4-4、長時長 LIGHTGBM 變異點違規機率圖 四-13
圖 4-5、長時長 CATBOOST 變異點違規機率圖 四-14
圖 4-6、長時長 RF 變異點違規機率圖 四-15
圖 4-7、以民國 108 年單一年份訓練 LIGHTGBM 之違規機率圖 . 四-16
圖 4-8、以民國 109 年單一年份訓練 LIGHTGBM 之違規機率圖 . 四-17
圖 4-9、以民國 110 年單一年份訓練 LIGHTGBM 之違規機率圖 . 四-18
圖 4-10、以民國 111 年單一年份訓練 LIGHTGBM 之違規機率圖 四-19
圖 4-11、以民國 112 年單一年份訓練 LIGHTGBM 之違規機率圖 四-20
圖 4-12、以民國 108 年單一年份訓練 CATBOOST 之違規機率圖 . 四-21
圖 4-13、以民國 109 單一年份訓練 CATBOOST 之違規機率圖 ... 四-22
圖 4-14、以民國 110 單一年份訓練 CATBOOST 之違規機率圖 ... 四-23
圖 4-15、以民國 111 單一年份訓練 CATBOOST 之違規機率圖 ... 四-24

圖 4-16、以民國 112 單一年份訓練 CatBoost 之違規機率圖 ... 四-25
圖 4-17、以民國 108 年單一年份訓練 RF 之違規機率圖 四-26
圖 4-18、以民國 109 單一年份訓練 RF 之違規機率圖， 四-27
圖 4-19、以民國 110 單一年份訓練 RF 之違規機率圖， 四-28
圖 4-20、以民國 111 單一年份訓練 RF 之違規機率圖， 四-29
圖 4-21、以民國 112 單一年份訓練 RF 之違規機率圖， 四-30

第一章　緒論

第一節　計畫介紹

1. 全程目標

隨著人口密度的增長與經濟型態的多元發展，在我國土地面積且自然資源有限的情況下，確保國土能夠永續經營與開發是國家刻不容緩的目標。為有效遏止土地違法利用，內政部國土管理署、農村發展及水土保持署（簡稱農村水保署），以及經濟部水利署等單位，近年來持續運用衛星影像及遙測技術進行國土利用監測作業，以協助進行地表變異點的辨識。

在判釋技術上，目前主要借重前後期高解析度衛星影像比對，透過光譜（spectral）與紋理（texture）等特徵上的變化，可查找地表變異點並進一步通報地方政府及相關主管機關，以派遣勘查人員至現地進行違規變異點的查報與後續追蹤。雖然以人工辨識的方式可有效圈繪出變異點範圍與樣態，有效濾除因自然災害或常態性季節變化，如農地輪作、植物物候等，但過程中受限於背景資料不足，在人工變異通報上難以分辨合法與非法情事，受資料面限制，絕大多數衛星影像辨識的變異點多為合法。有鑑於此，為了減少對合法變異點進行現地查核所造成人力資源與時間成本的浪費，本計畫擬利用機器學習工具測試創新的研究方法，旨在運用人工智慧技術提升變異點的查復效率。

2. 研究範圍

在空間範圍上，本計畫根據水保署提供的全臺山坡地變異點位資料，以臺澎金馬地區作為資料集收集範圍。在時間範圍上，

透過近五年(108-112)的資料，依測試規劃分別拆分為獨立的訓練及驗證資料，作為評斷模式精度與可行性的依據。

第二節　資料作業與項目

1. 資料目的

為了訓練用於預測變異點違規機率的機器學習模型，除了變異點本身的資訊外，也須考量到與變異點成因相關的各種人文與環境因素數據，以利模型進行學習。本計畫收集到的各項數據與圖層包含了變異點資訊、數值地形模型、人口資料與道路資料。

2. 資料收集

(1) 變異點資訊

透過國土利用監測作業的衛星影像變異自動化判釋與通報工作，農村水保署提供了民國 108 年至 112 年間每月全臺山坡地的變異點位資訊，包含了變異點的座標資訊與違規檢核結果，將這五年間每一年各月份檢測到的所有變異點資料各自集結起來後，各縣市統計之變異點數量如表 1-1 所示，在臺灣的分佈情形如圖 1-1 所示，其中藍色點代表非違規變異點，紅色點為違規變異點。各縣市近年在違規變異點的數量上互有消長，主要與通報頻率、影像數量以及使用需求有關，並非直接或間接顯示土地違規利用的趨勢，亦與通報及查報力度無關。

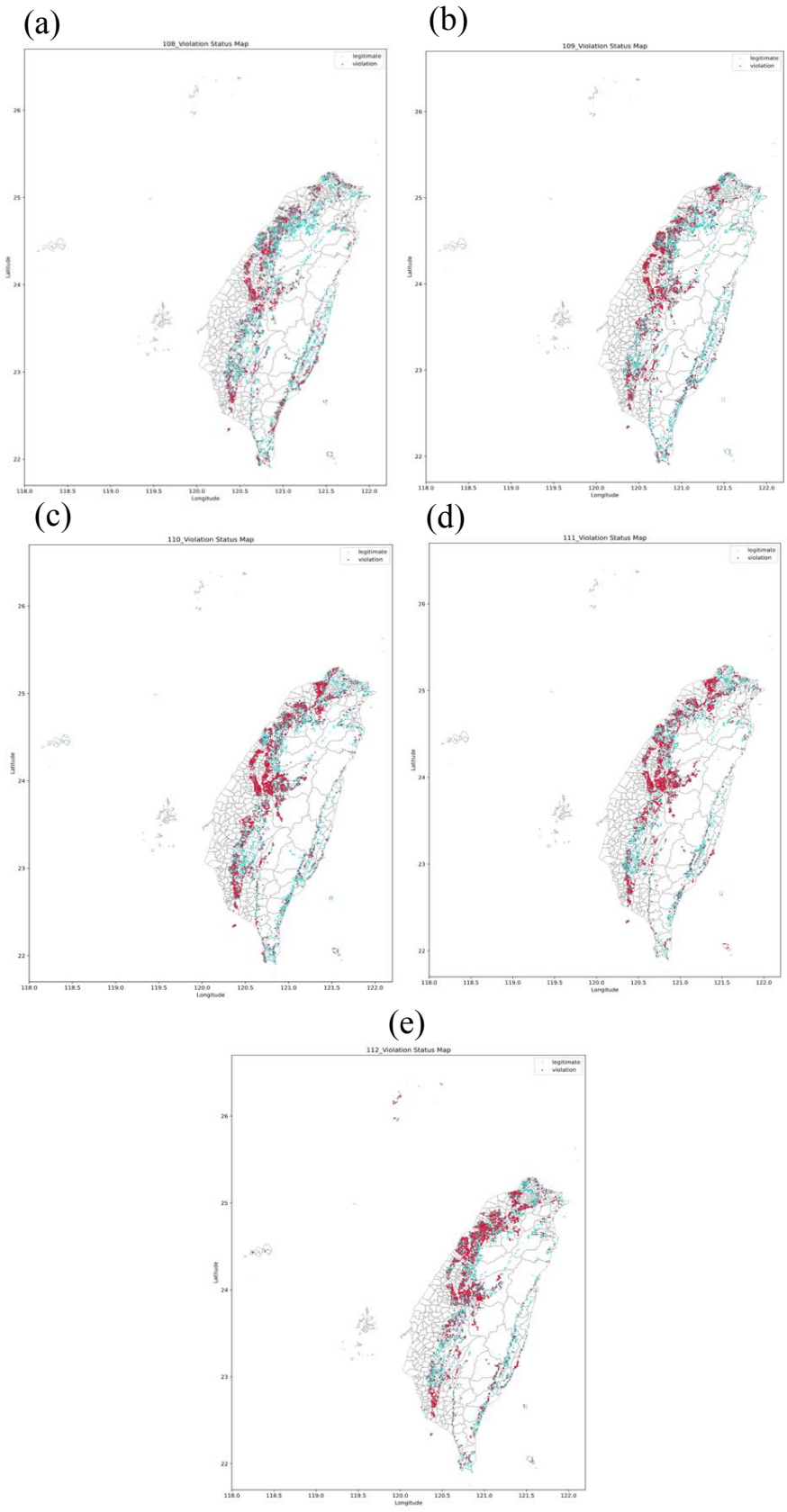

圖 一-1、變異點分佈圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 111 年、(e)民國 112 年

表 一-1、每年各縣市變異點數量統計

地區	108 年 變異點數量	109 年 變異點數量	110 年 變異點數量	111 年 變異點數量	112 年 變異點數量
臺北市	53	38	79	188	159
新北市	630	849	1030	826	911
桃園市	318	361	451	497	593
高雄市	751	801	822	640	542
臺南市	508	511	862	601	480
臺中市	589	886	865	773	780
基隆市	68	32	46	59	43
新竹市	92	67	57	52	109
嘉義市	8	15	18	10	3
宜蘭縣	280	157	259	256	227
花蓮縣	409	432	298	301	418
南投縣	1039	1211	2305	2065	1810
屏東縣	503	444	453	236	256
苗栗縣	1230	1852	1118	963	1743
雲林縣	87	56	65	114	127
新竹縣	987	769	908	596	1434
嘉義縣	439	424	477	351	389
彰化縣	216	159	229	199	127
臺東縣	998	820	798	590	754
金門縣	0	0	8	8	26
連江縣	0	0	0	0	23
澎湖縣	0	0	0	0	0
總計	**9205**	**9884**	**11148**	**9325**	**10954**

(2) 數值地形模型

考慮到變異點資料集中，部分變異點的成因是由於山崩與土石流的發生，根據前人對山崩預測的研究 (Goetz et.al ,2015; Brenning, 2005)，可知高程、坡度與坡向等等因素均為顯著的崩塌影響因子，因此，本計畫通過政府資料開放平臺，收集了全臺 20 公尺解析度的數值地形模型，如圖 1-2 所示，以用於進一步的分析與運用。

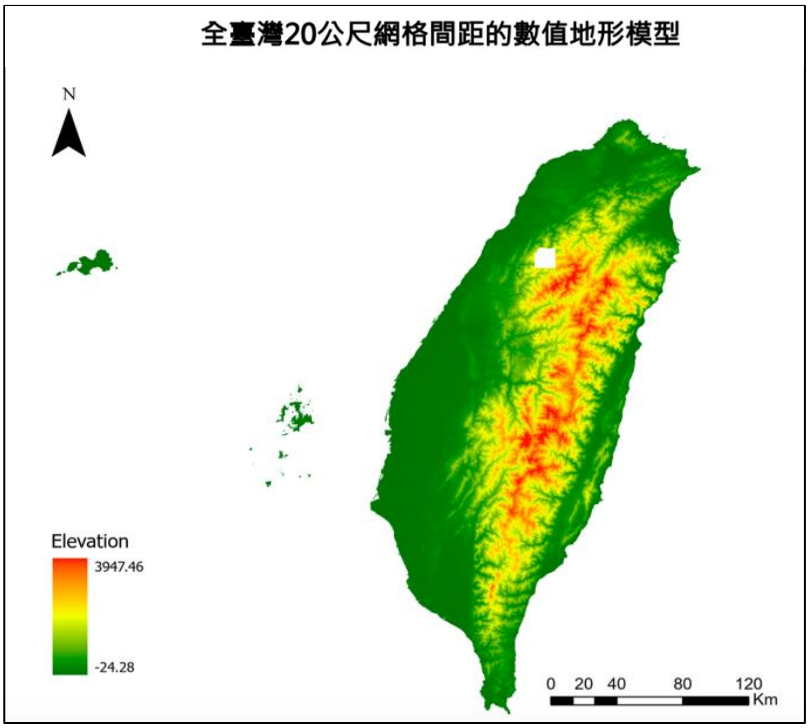

圖 一-2、數值地形模型示意圖

(3) 人口資料

隨著社會經濟的逐步發展，根據內政部全國人口資料庫統計(https://gis.ris.gov.tw)可知，臺灣近年來的總人口數於 109 年達到峰值，居民的生活環境逐漸受到了壓縮。人口增長會使得較低海拔地區的國土開發率逐漸升高 (Reis,

2008)，考量到違規變異點主要由人為活動引起，本研究亦收集了全臺最小統計區的空間單元資料，包含人口數與人口密度，並後續利用模型的訓練探討人口資料對違規變異點的影響。

(4) 全臺道路資料

由於道路是人類活動的重要指標，考慮到距離道路較近的山坡地更有可能受到人為活動的影響，進而增加違法使用的可能性，故本研究也將變異點與道路的最短距離作為訓練特徵進行機器學習。目前一共收集到了由交通部公路局提供的全臺之 (1)國道、 (2)省道、 (3)快速道路、 (4)直轄市區道、 (5)市道、 (6)縣道與 (7)鄉道等等道路圖資，並額外收錄了由(8) 臺北與(9)臺中市政府所提供地較為詳細的市區內道路圖資，以及由(10)雪霸國家公園管理處提供的登山步道圖資，如圖 1-3 所示。

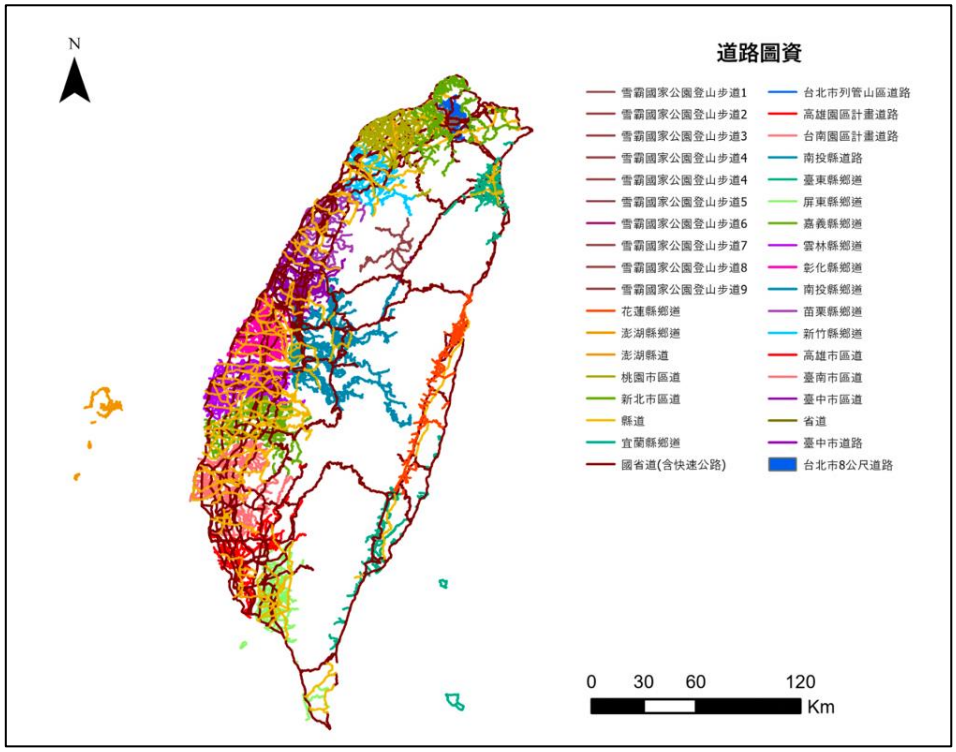

圖 一-3、道路圖資分佈圖

第三節　機器學習模型

經過相關文獻的調查後，本計畫選擇三種機器學習算法，分別為 LightGBM、CatBoost 和 Random Forest(簡稱 RF)。這三個模型各自具有獨特的特點和應用優勢，將使用相同的資料集對這些模型進行訓練，並比較其預測結果，以選出表現最佳之機器學習方法。透過該過程，期望能夠找出最適合該研究計畫目標的模型，從而提高預測的準確度和可靠性。

1. LightGBM

1989 年時 Schapire 提出了 boosting 演算法的概念(Schapire, 1990)，這種方法通過將多個弱學習演算法組合在一

起，提升其準確性，達到與強學習算法相同的計算程度。1995 年時，Freund 和 Schapire 再次提出了 AdaBoost（適應性提升）演算法（Freund et al., 1997），解決了早期 boosting 演算法的一些困難。隨後，Friedman 提出了梯度提升決策樹（GBDT）的概念（Friedman, 2001），通過整合多個弱分類器（決策樹）來發展最優模型。GBDT 衍生出的工具包括由陳天奇開發的 XGBoost 模型（Chen et al., 2016），該模型能夠大幅提升迭代運算的效率與預測準確度。然而，XGBoost 在處理龐大數據集或高維度特徵資料時，其運算效率與空間擴展性仍存在不足。這主要是因為該方法首先需要計算所有數據的特徵值，以估計可能的決策樹分割點以進行增益計算，將使得記憶體消耗過大。

於是在 2017 年時，微軟提出了更高效的 LightGBM 模型（Ke et al., 2017），該方法在 GBDT 演算法上增加了基於梯度的單側採樣技術（GOSS）和互斥特徵捆綁（EFB）兩種創新方法，從而具有更高的訓練效率、較低的記憶體使用以及更好的準確性。

不同於 XGBoost 運用之樹向生長策略，LightGBM 採用了葉向生長策略，能協助處理大規模的資料。此外，GOSS 能夠幫助減小樣本維度，降低在樹建構過程中尋找最佳分割點的計算成本。具體來說，假設訓練時分裂節點上的樣本為 I，設置兩個比例，a 為大梯度樣本抽樣比，b 為小梯度樣本抽樣比，訓練時在每一次迭代時對樣本進行預測，主要透過兩個步驟來執行：

(1) 按照梯度大小將樣本降序排列；

(2) 取出前（a × |I|）個大梯度樣本，再從剩下的小梯度樣本中隨機抽取樣本（b × |I|）個，並乘上常數因子（(1-a)/b）來放大採樣，以補償數據分佈的影響。

此外，GOSS之公式可表示為：

$$GOSS = a \times |I| + b \times |I| \times \left(\frac{1-a}{b}\right) \quad (1)$$

透過GOSS方法，LightGBM能專注於學習對模型訓練貢獻最大的資料特徵因素，同時仍利用剩餘因素來指導決策樹的建構過程。而EFB旨在減少訓練過程中的特徵數量，藉由設置閾值判斷互斥的特徵，而後將互斥的特徵綑綁在一起，從而減少決策樹中分裂節點考慮的特徵數量，藉此縮短訓練的時長及減少記憶體的消耗。

EFB之運算方式可表示為：

(1)設 F 為特徵集， E 為互斥特徵集合，則：

$$E = F_1, F_2, \ldots, F_n$$

其中$F_i \cap F_j = \emptyset$，對所有$i \neq j$。

(2)互斥特徵捆綁透過將互斥特徵集合 E 中的特徵進行綁定來減少特徵數量：

$$EFB = \sum_{i=1}^{n} \mathbb{I}(F_i) \quad (2)$$

其中𝕀為特徵選擇函數，用於判斷哪些特徵應該捆綁在一起。

2. CatBoost

CatBoost的名稱是Category與Boost之組合，是針對處理類別特徵問題所設計的監督式學習模型，由俄羅斯最大的搜尋引擎公司Yandex推出，與LightGBM一樣，皆屬於基於決策樹梯度

提升（Gradient Boosting）的演算法，通過構建新樹來擬合模型的梯度。

傳統的梯度提升算法在建立每一棵新樹時，都會使用目前模型的梯度來進行估計。然而，由於每一步的梯度都是使用基於相同數據點來估計，這將導致估計梯度在特徵空間中產生偏移，從而導致過擬合。為了解決這個問題，CatBoost 使用了有序提升(ordered boosting) 和有序目標統計 (ordered target statistics) 來改進梯度提升決策樹 (GBDT) 的算法(Dorogush et al., 2018)。這種方法能有效地解決梯度提升中存在的預測偏移問題，同時也能達到比 LightGBM 更快的訓練速度和更好的準確度。以下是 CatBoost 的簡單演算法 (Prokhorenkova et al, 2018)：

(1) 在梯度提升迭代的過程中，會建構一系列的函數近似F^t，其中h^t是弱學習器（基本預測器），目標要達到最小預期損失：

$$h^t = argminE\left[L\left(y, F^{(t-1)}(x) + h(x)\right)\right] \quad (3)$$

(2) 梯度的估計公式為：

$$g^t(x,y) = \frac{\partial L(y,s)}{\partial s}\Big|_{s=F^{(t-1)}(x)} \quad (4)$$

(3) 弱學習器的訓練目標是最小化預期損失，$h^t(x)$近似於$-g^t(x,y)$，使用最小二乘法進行逼近：

$$h^t = arg\ minE[(-g^t(x,y) - h(x))^2] \quad (5)$$

(4) 弱學習器的形式表示為：

$$h(x) = \sum_{j=1}^{J} R_j 1_{x \in R_j} \quad (6)$$

其中R_j代表決策樹的葉子節點所對應的不相交區域。在進行預測時，模型會根據輸入樣本的特徵值將其劃分到對應的區域（即葉子節點），並返回該區域對應的預測值。

3. Random Forest (RF)

RF 模型運用的決策樹方法，被廣泛應用於資料探勘和預測分析中 (Song et al, 2015)。決策樹方法可依據特徵將資料分割，並包含多種不同的演算法 (Hastie et al., 2009)，其中包括分類樹（適用於分類問題）和回歸樹（適用於回歸問題），而 CART（Classification and Regression Trees）是一種通用的決策樹算法，可以同時用於分類和回歸問題。決策森林是一個包含多個決策樹的集成學習模型 (Tong et al.,2003)，而 RF 是在決策森林的基礎上進一步引入了特徵的隨機選取方法，這可以看作是對 CART 算法的進一步改進 (Rigatti, 2017)。

RF 每一次構建一棵樹時，都會隨機抽樣一部分樣本和特徵來建立多棵決策樹，其精確度取決於個別樹狀分類器的強度和相關性。強度是指樹狀分類器的平均準確率，相關性是指樹狀分類器之間的相關係數 (Breiman, 2001)。

RF 的演算法如下 (Ho, 1995)：

(1) 計算在節點v_j上類別c的條件機率：

$$P\left(c\middle|v_j(x)\right)=\frac{P\left(c\middle|v_j(x)\right)}{\sum_{l=1}^{n=1}P\left(c_l\middle|v_j(x)\right)} \quad (7)$$

(2)計算類別c的平均預測機率：

$$g_c(x)=\frac{1}{t}\sum_{j=1}^{t}P\left(c\middle|v_j(x)\right) \quad (8)$$

其中，P是機率， c是類別，v_j是節點，t 是決策樹的數量，g_c是平均預測。隨機森林因其高精度、低計算成本和易於使用等優點，已被廣泛應用於遙測影像分類中，能夠適應不同的分類目標，例如土地覆蓋 (Colditz, 2015)、植被 (Wang et al., 2015)等。

第二章　工作執行方法與步驟

第一節　工作執行方法

本研究目的是預測山坡地變異點是否違規，以及判別臺灣地區可能存在違規開發的高機率區域。為了達到此一目標，結合了多種數據來源如：臺灣山坡地的地理變異點數據，其中涵蓋了不同年度通報點位的地理坐標（X, Y）、違規狀態（違規或非違規）、以及其他環境特徵（例如高程、坡度、坡向、曲率、人口數、人口密度等）來進行特徵萃取，並使用機器學習模型進行分類預測。

1. 資料前處理

(1) 缺失數據填補

在本研究中，由於從政府開放資料平台(https://data.gov.tw) 取得的資料集中，或有資料收集錯誤或記錄不完整等原因，某些觀測值的變量存在缺失數據，如表 2-1 所示。缺失數據會在數據分析和解釋中引發多種挑戰和複雜性，並可能影響模型的預測能力與演算法的性能。如果模型基於不完整的數據學習，將可能無法很好地泛化到新數據，從而影響結果的可靠性。為了解決這些問題，本研究使用了兩種主要的插補方法：反距離加權插值法（IDW）和迭代插補法。

1. 反距離權重插值法(IDW)

數字高程模型（DEM）的缺失數據使用此方法進行插值，假設彼此距離較近的點比距離較遠的點具有更多的相關性和相似性。在 IDW 方法中，相鄰點之間的相關性和相

似性比率與它們之間的距離成反比，具體運算公式如下(Setianto et al., 2015)：

$$Z_0 = \frac{\sum_{i=1}^{N} Z_i \cdot d_i^{-n}}{\sum_{i=1}^{N} d_i^{-n}} \quad (9)$$

其中，Z_0 是目標點的預測值，Z_i 是已知點的值，d_i 是目標點與已知點之間的距離，N是加權指數。本研究中，數值高程模型（DEM）的資料空值區域即是使用此方法進行內插計算。

2. 迭代內插法

對於人口數、人口密度、坡度等其他數據，本研究使用了迭代內插法，這是基於多元填補鏈式方程（Multiple Imputation by Chained Equations, MICE）的內插方法(Stef van Buuren et al., 2011)。迭代內插法（Rubin, 1987）是一種處理複雜缺失數據問題的首選方法，通過產生多個合理的內插數據集並綜合每個數據集的結果，能夠確保對缺失數據的不確定性進行適當處理 (Resche-Rigon et al., 2018)。

其主要差補方式是將具有缺失值的每個變量建模為其他變量的函數，從初始內插開始，通過迭代這些條件密度來抽取內插值。簡單的運算公式如下：

$$Z_j^{(t)} = M_j\left(X, Z_j^{(t-1)}\right) + \varepsilon_j^{(t)} \quad (10)$$

$$d_j^{(t)} = \left|Z_j^{(t)} - Z_j^{(t-1)}\right| \quad (11)$$

$$d_{max}^{(t)} \leq Threshold \quad (12)$$

其中，t 是迭代次數，X是除了Z_j之外的所有觀測值，$Z_j^{(t)}$是Z_j的內插值，M_j是用於預測缺失變量對其他變量影響的模型，$\varepsilon_j^{(t)}$屬於誤差項。

表 二-1、資料缺失值比例

	Column Name	Missing Count	Percentage (%)
1	Elevation	23	0.05
2	Slope	23	0.05
3	Aspect	23	0.05
4	Curvature	26	0.05
5	Population	26	0.05
6	Population Density	26	0.05

(2)　道路距離計算

在本研究中，藉由變異點與道路之地理座標資訊，使用歐幾里德距離公式迭代計算點到道路之間的最小距離，使其作為訓練的參數之一，公式如下：

$$d(A,B) = \sqrt{(X_B - X_A)^2 - (Y_B - Y_A)^2} \qquad (13)$$

在取得與道路之間的最短距離後，又根據距離大小將其分成多個區間，並賦予每個區間一個等級，形成交通便利度特徵。若變異點與道路最短距離介於 0-1000 公尺之間，則每 50 公尺為一級距，1000-4000 公尺之間則每 500 公尺為一級距，4000 公尺以上則皆劃分為同一級，藉由此方法提高了數據的實用性。

(3)坐標特徵處理

為了進一步分析變異點的空間分佈特徵，本研究使用 DBSCAN（Density-Based Spatial Clustering of Applications with Noise）聚類方法，將地理坐標（X, Y）進行分組，DBSCAN 是一種基於密度的聚類算法，其主要優勢在於能夠識別任意形狀的聚類並區分噪聲點。由表 1-1 可知，五年來所有縣市統計出的變異點數量中最少為 3 個，故本研究定義一個聚類中至少包含的最小點數量為 3，並定義點之間的最大距離為 10 公尺，如若一個變異點的鄰域內包含的點數量大於或等於 3，則該點便可假設為核心點，並且該鄰域內的所有點（包括核心點本身）都屬於同一聚類。由於考慮到空間上違規開發土地的人類行為不會太分散，故違規變異點之間應具有空間上分佈的相對密集性，如果一個變異點既不是核心點也不在任何核心點的鄰域內，則該點將被標記為噪聲點。通過這一方法，可以協助模型更好地理解山坡地變異點的空間分佈特徵。

(4)重採樣

在現實情境的問題中，資料集通常具有不平衡的類別，由農村水保署提供的變異點資料經統計後，違規變異點與非違規變異點數量如表 2-2 所示。可以看到五年來非違規變異點的數量比違規變異點的數量平均多了約 3.03 倍。不平衡的數據會影響模型的效能，因為演算法可能會偏向多數類別而忽略少數類別，從而導致對少數類別錯誤分類的情形更為嚴重。由於使用過採樣技術合成違規變異點的新樣本時，模型

出現了的過擬合的現象，而考慮到使用欠採樣會有丟棄掉非違規變異點中有用的資料之風險，但違規樣本數量又足以提供有效訓練，故本計畫選擇了 EasyEnsemble 的方法，隨機欠採樣非違法變異點數據。EasyEnsemble 能從多數類別中取樣子集，並使用 AdaBoost 在每個子集上訓練分類器，最後將其組合以產出最終的數據集，此方法能夠良好地探索到欠採樣過程中可能被忽略的多數類別數據，是一種能夠專注於在欠採樣過程中經常被忽略的多數類別數據來解決類別不平衡問題的有效方法 (Schubert et al., 2017)。EasyEnsemble 的 AdaBoost ensemble 集成模型公式如下：

$$H(x) = sgn\left(\sum_{j=1}^{S_i} \alpha_{i,j} h_{i,j}(x) - \sum_{i=1}^{T} \theta_i\right) \qquad (14)$$

其中，S_i為迭代次數，$h_{i,j}$為弱分類器，$\alpha_{i,j}$是權重，θ_i是分類決策的閾值。

表 二-2、每年違規與非違規變異點數量統計

變異點個數	108 年	109 年	110 年	111 年	112 年
非違規	7417	7645	8092	6577	7873
違規	1788	2239	3056	2748	3081
總計	9205	9884	11148	9325	10954
非違規/違規	**4.1482**	**3.4145**	**2.6479**	**2.3934**	**2.5553**

2. 模型建構

本研究選擇了三種機器學習模型(LightGBM、CatGBM、RF)來預測變異點是否為違規。首先將五年的數據集進行迭代抽取，每一年都輪流作為測試集，而後剩餘四年的數據合併作為訓練數據集。其中，又將訓練數據集分為 70%為訓練集和 30%為驗證集，驗證集能夠協助評估模型性能，計算模型精確度，並分析各個縣市的預測準確率。

(1) 訓練結果比較

在模型參數調整過程中，採用了混淆矩陣來評估模型的表現。具體來說，關注了以下幾個指標：

- 真陰性（TN）：表示非違規變異點被正確預測為非違規的數量。這一指標能夠協助了解模型在識別非違規變異點方面的能力。
- 真陽性（TP）：表示違規變異點被正確預測為違規的數量。這一指標對於評估模型在檢測違規變異點方面的精度至關重要。
- 生產者精度（Producer's Accuracy）：表示模型在預測違規變異點時的準確性，即真陽性數量 （TP） 佔實際違規變異點數量的比例。這一指標能夠反映出模型在實際應用中的可用性和可靠性。

此外，還計算了每個模型預測違規變異點的最小機率。若所預測出的機率低於該機率，則可完全不需派遣勘查人員進行實地訪查。透過比較各模型預測出之最低違規變異點機率的高低和穩定性，也可以判斷模型的穩健度。

(2) 違規潛勢圖繪製

為了直觀地展示模型預測結果，最後將所有變異點的預測違規機率繪製在臺灣地圖上呈現。藉由該方法可以直接地呈現違規變異點的主要分佈區域，並且可以與真實的違規與非違規變異點分佈情況進行對比（如圖 1-1 所示）。這種可視化方法有助於觀察哪些地區的預測尚有待改進，以及哪些地區的變異點被誤判為違規。通過這樣的分析，可以幫助我們思考數據的收集方向和處理方法，以進一步提高模型的準確性和穩健性。

(3) 統計違規變異點誤判數量

為了進一步了解模型的預測表現，也統計了五年間每一年作為測試集資料時，全臺各縣市中實際為違規變異點但卻被模型判為非違規變異點的數量。這些統計數據能夠協助了解哪些縣市需要重點關注，並思考可以再收集哪些資料以期能夠提升這些容易誤判地區的違規變異點預測準確度。透過這種分析，能夠識別數據不足或特徵表現差的區域，並針對這些地區進行更具針對性的數據收集與模型改進。

3. 數據分析

主成分分析（Principal Component Analysis, 簡稱 PCA），是一種將高維數據降至低維的統計方法，由英國數學家卡爾·皮爾森發明，有助於降低大型資料集的維度，同時保留盡可能多的變異數。其核心思想是通過線性變換，計算數據協方差矩陣的特徵值和特徵向量，找到一組新的正交基底，使得數據在這組基底上的投影

具有最大變異性。藉由將原始變數轉換為較小的一組不相關變數（主成分），PCA 能在不遺失重要資訊的情況下簡化資料 (Abdi et al., 2010)。

在進行 PCA 之前，數據通常需要進行預處理。首先是標準化，即使每個變量的平均值為零：

$$X = P\,\Sigma\,Q^T \tag{15}$$

其中，

P是左奇異向量矩陣，Q是右奇異向量矩陣，Σ是奇異值矩陣。而數據點偏離重心(中心點)的程度被稱為慣性（Inertia），慣性在 PCA 中代表了數據的總變異量，提供了數據分散程度的量化指標，在標準化數據之後通常需要進行慣性計算：

$$\gamma_j^2 = \sum_i^I x_{i,j}^2 \tag{16}$$

其中，總慣性等於所有特徵值的和$I = \sum \lambda_l$，每個主成分解釋的變異比例可以通過其特徵值與總慣性的比值來計算，較大的慣性值表示該主成分捕獲了數據中更多的變異，透過該比例可得知每個主成分保留了多少原始數據的信息：

$$\tau_l = \lambda_l / I \tag{17}$$

在計算完每個主成分的相對重要性後，需要將原始數據從原始特徵空間轉換到主成分空間，得到每個觀測值在新座標系統（主成分）上的投影位置，該步驟又稱因子得分(Factor Scores)：

$$F = P\Sigma = P\Sigma Q^T Q = XQ \tag{18}$$

其中，F是因子得分矩陣，X是原始數據矩陣：$X = P\Sigma Q^T$，Q又可視為對原始資料矩陣變換為因子分數的投影矩陣。綜和上述，整體的幾何解法步驟包含（圖 2-1 ~ 2-3）：

(1) 將變數標準化，

(2) 求點雲的主方向（稱為第一個分量），使得點到該分量的距離平方和最小，再加入與第一個分量正交的第二個分量，使得距離平方和最小，

(3) 旋轉圖形，投影到主成份空間。

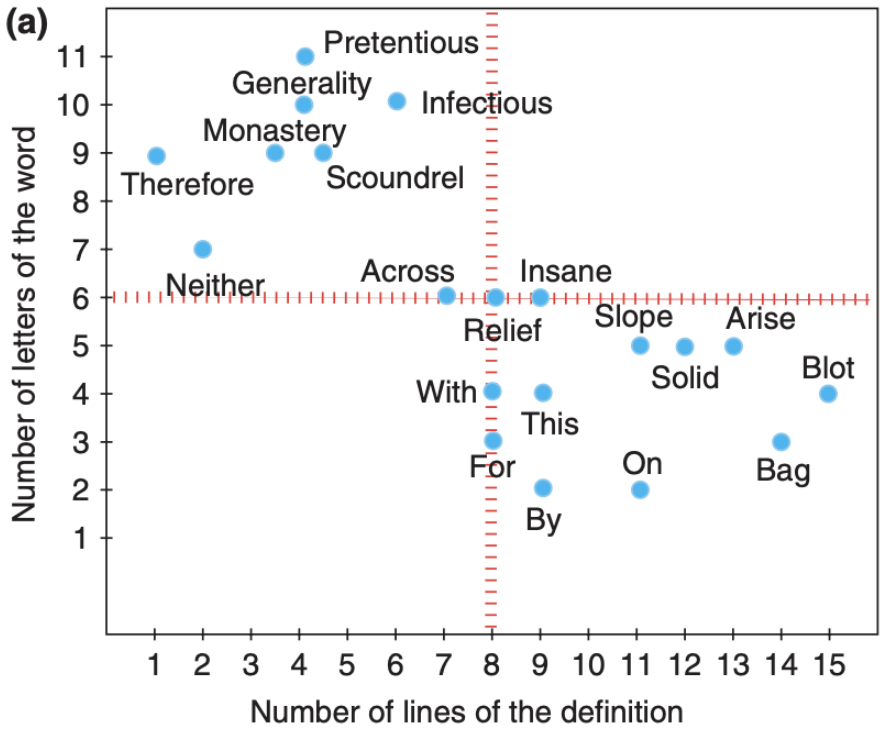

圖 二-1、將變數標準化（中心化）

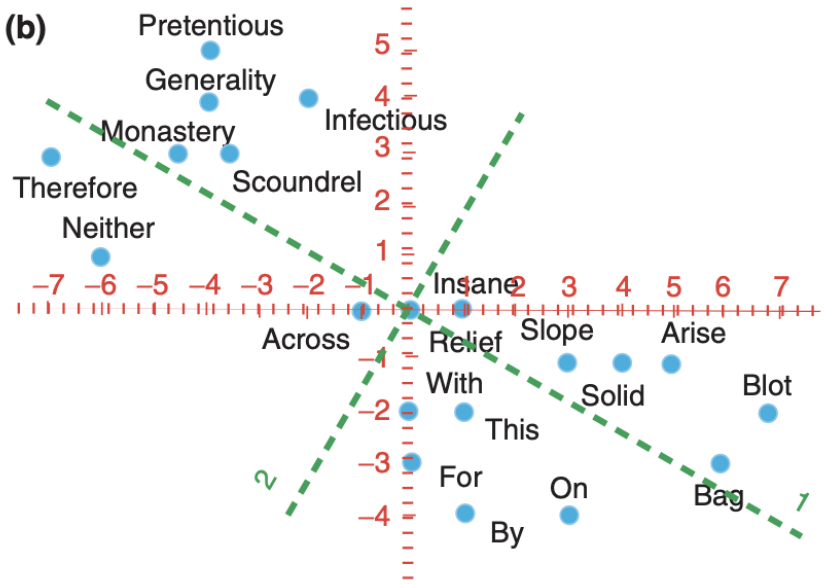

圖 二-2、求出兩個主方向

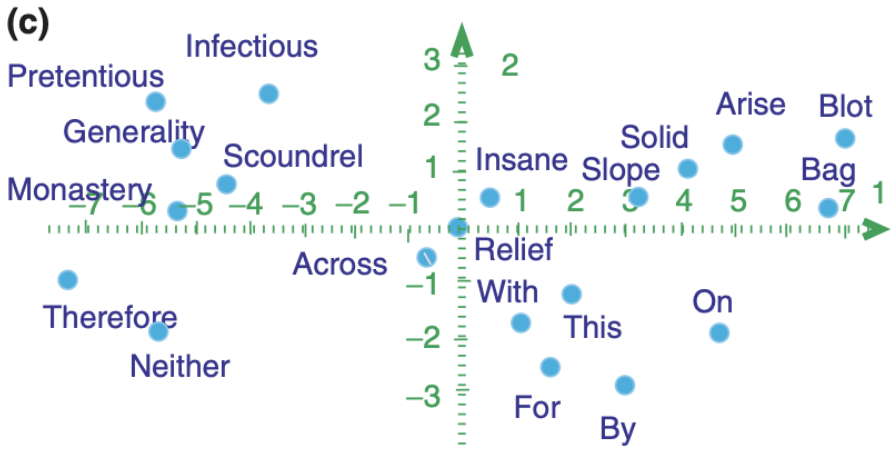

圖 二-3、投影到主成份空間

第二節　工作執行步驟

1. 流程圖

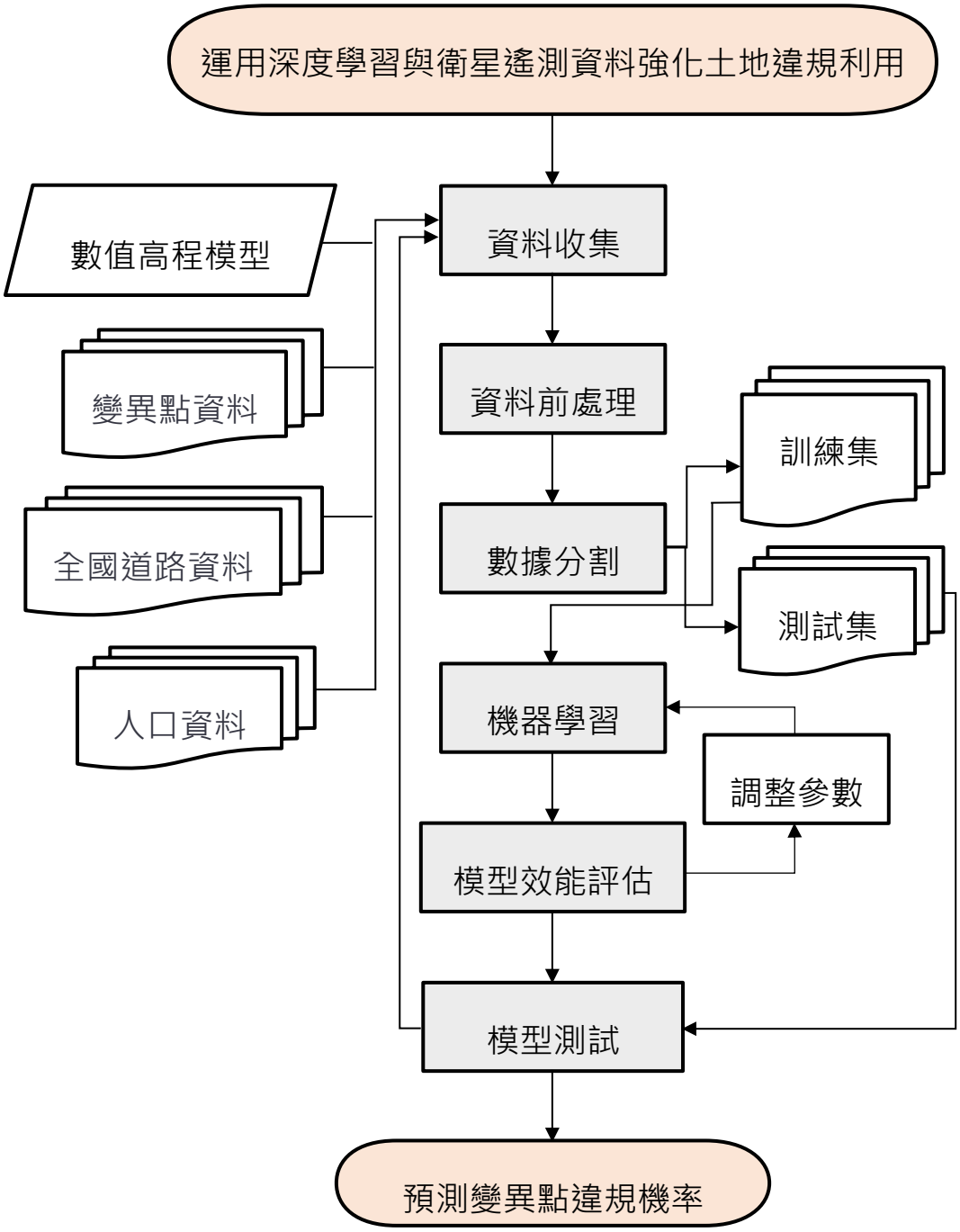

圖 二-4、工作流程圖

2. 工作內容

(1) 本計畫執行至目前所收集到的資料包含了：

1. 民國 108-112 年間的山坡地變異點資料。
2. 20 公尺數值高程模型。
3. 人口資料（包含最小統計區之人口數與人口密度）。
4. 全國道路資料（包含國道、省道、快速道路、區道、市/縣道與鄉道）。

(2) 對於資料的處理流程步驟大致為：

1. 資料彙整：整合來自不同來源的資料。
2. 資料前處理：資料清理與特徵提取。
3. 數據分割：將處理後的資料分為訓練集、驗證集與測試集。

(3) 模型過程：

1. 機器學習：使用訓練集訓練模型。
2. 模型效能評估：使用驗證集評估模型的表現。
3. 模型測試：使用測試集進行最終測試。
4. 調整參數：根據評估結果動態調整模型參數。

(4) 輸出結果：

1. 預測變異點違規機率：模型最終輸出對測試集資料中每個變異點預測的違規機率，用於判斷土地被違規利用的可能性。

2. 繪製變異點違規機率圖：可協助直觀的觀察哪一處的變異點預測精度仍須提升，藉此思考未來還需要再收集和何種資料以增進預測準確度。

3. 統計違規變異點誤判數量：整體工作的重點都可專注在找到與違規變異點高度相關的因素。能夠了解哪些縣市具有更多違法土地開發，以及了解哪些縣市的違規變異點無法讓模型正確學習到，將有助於啟發資料的收集方向，同時找出模型表現不佳之問題節點，協助研究的進程。

第三章　工作進度與交付項目

第一節　計畫甘特圖

該圖顯示各項工作之開始執行日期以及完成日期。

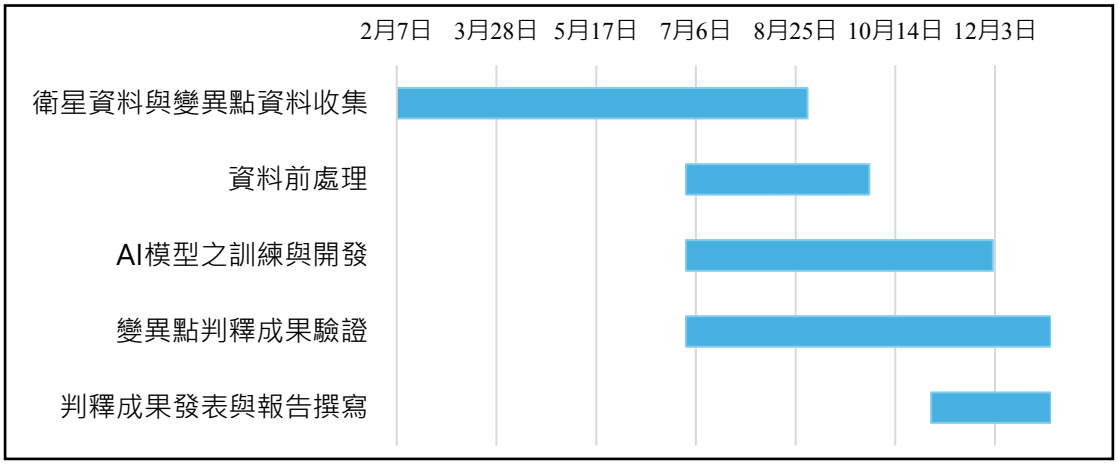

圖 三-1、計畫甘特圖

第二節　計畫交付項目

本計畫於期中階段交付期中報告一份，於期末階段交付期末成果報告，內容包含最佳機器學習或深度學習進行違規變異點判識成果，視模型穩定性可提供水保署相關軟體以供後續參用。

第四章　研究成果

1. 模型效能評估

分析三種模型的整體準確度(Overall Accuracy)(如表 4-1)，結果顯示：當以民國 108 年與 109 年作為測試資料，其餘年度作為訓練資料時，LightGBM 模型展現出最高的準確率。而當採用相同方法，分別以民國 110 年至 112 年這三年為測試資料時，則是 Random Forest 模型的準確度最高。

然而，以民國 109 年 Random Forest 模型的預測結果為例(如表 4-2)，我們發現整體預測準確率實際上受到了數據不平衡的影響：由於非違規變異點的數量遠多於違規變異點，即使模型將絕大多數違規變異點誤判為非違規，整體準確度仍然會偏高。因此，若僅觀察整體預測準確度，將無法真實反映模型的實際預測能力。

表 四-1、整體準確度

Overall Accuracy	108	109	110	111	112
LightGBM	**0.766**	**0.713**	0.497	0.522	0.540
CatBoost	0.528	0.503	0.510	0.533	0.524
RF	0.661	0.638	**0.599**	**0.608**	**0.617**

表 四-2、109 年 Random Forest 模型預測結果

109 Testing (單位：變異點個數)	Truth Data		
	非違規	違規	Total

Predicted Result	非違規	5407	1335	6742
	違規	2242	909	3151
	Total	7649	2244	9893

為了更準確評估模型在預測違規變異點時的表現，本研究採用生產者精度（Producer's Accuracy）作為性能評估指標（如表 4-3）。生產者精度是混淆矩陣中的重要指標之一，用於衡量實際屬於某個類別被正確分類的比例。舉例來說，若有 100 個實際違規案例，模型能正確識別出其中 80 個，則生產者精度為 80%（80/100），表示模型能成功找出 80%的實際違規案例。

在土地違規偵測的應用中，高生產者精度代表模型能有效識別出大部分的實際違規案例。根據目前的初步研究成果顯示，在預測違規變異點方面，CatBoost 模型表現最為優秀，而 Random Forest（RF）模型的表現則相對較差。

表 四-3、生產者精度

Producer's Accuracy	LightGBM	CatBoost	RF
108 Testing	0.436	**0.575**	0.434
109 Testing	0.417	**0.656**	0.405
110 Testing	0.635	**0.659**	0.491
111 Testing	0.552	**0.563**	0.403
112 Testing	0.492	**0.521**	0.419

2. 最低違規機率

該計畫目的是為了能夠減少派駐到非違規變異點進行實地勘查所造成之人力與時間成本的浪費，為此，本研究針對每一個模型預測的最低違規機率進行了統計與整理（如表 4-4 所示）。當變異點位的預測機率低於此最低違規機率時，即可免除派遣人員進行實地勘查的需求。

觀察表 4-4 可發現，雖然目前三個模型產出的最低違規機率都不高，因此能夠排除勘查的變異點數量較為有限。然而，在這五年期間的所有模型中，CatBoost 的最低違規機率預測值始終維持最高，這意味著使用 CatBoost 模型能夠協助排除最多無勘查必要的非違規變異點。這個結果也間接顯示，相較於其他兩個模型，CatBoost 模型具有更完善的學習成效。

表 四-4、最低違規機率

最低違規機率	LightGBM	CatBoost	RF
108	0.17	**0.18**	0.08
109	0.17	**0.64**	0.12
110	0.19	**0.27**	0.08
111	0.18	**0.19**	0.12
112	0.18	**0.20**	0.05

3. 顯著因子

顯著因子(或稱特徵重要性)是指在機器學習模型中對預測結果影響最大的變數或特徵(如表 4-5)。分析本研究所收集的數據發現，坐標特徵與高程對於所有模型而言，均為最重要的兩大顯著因子。這個結果證實了本研究使用聚類算法所衍生出的座標特徵成功協助了模型進行學習。

又結合高程資訊可以推測，在地勢與山坡地高程條件允許的情況下，違規開墾國土的情形往往會聚集在距離某個核心點一定範圍內的鄰域中。這個推論也可以從表中得到進一步的驗證：與地勢相關的坡度因子同樣穩定地出現在前四大顯著因子之中。

表 四-5、顯著因子

顯著因子	LightGBM	CatBoost	RF
1	Coordinate	Coordinate	Coordinate

2	Elevation	Elevation	Elevation
3	Slope	Population Density	Slope
4	Population Density	Slope	Aspect
5	Road Distance	Aspect	Curvature

4. 不同時長訓練結果對比

本研究比較了短時長與長時長數據對模型學習效果與預測能力的影響。長時長部分採用四年數據作為訓練集，剩餘一年作為驗證集，其訓練結果如前述討論（表 4-3）。短時長部分則選擇單一年度數據進行訓練，並用來預測其餘四年的違規與非違規變異點（如表 4-6 至 4-10）。

從表中數據可以明顯觀察到，無論選擇民國 108 年至 112 年間的哪一年作為訓練數據，僅使用單年數據訓練的模型，其生產者精度都低於使用長時長數據集訓練的結果。這項發現顯示，累積多年資料進行訓練確實有助於提升模型的準確度。

表 四-6、民國 108 年之數據作為訓練集

Producer's Accuracy	108 Training		
Testing Year	LightGBM	CatBoost	RF
109	0.26	0.24	0.17
110	0.31	0.29	0.19
111	0.31	0.30	0.21
112	0.29	0.29	0.24

表 四-7、以民國 109 年之數據作為訓練集

Producer's Accuracy	109 Training		
Testing Year	LightGBM	CatBoost	RF
108	0.23	0.21	0.30
110	0.30	0.29	0.38
111	0.31	0.30	0.34
112	0.29	0.28	0.21

表 四-8、以民國 110 年之數據作為訓練集

Producer's Accuracy	110 Training		
Testing Year	LightGBM	CatBoost	RF
108	0.22	0.21	0.29
109	0.26	0.24	0.37
111	0.32	0.30	0.39
112	0.30	0.29	0.33

表 四-9、以民國 111 年之數據作為訓練集

Producer's Accuracy	111 Training		
Testing Year	LightGBM	CatBoost	RF
108	0.26	0.24	0.17
109	0.31	0.29	0.19
110	0.31	0.30	0.21
112	0.29	0.29	0.24

表 四-10、以民國 112 年之數據作為訓練集

Producer's Accuracy	112 Training		
Testing Year	LightGBM	CatBoost	RF
108	0.22	0.20	0.28
109	0.25	0.23	0.31
110	0.30	0.28	0.33
111	0.31	0.30	0.36

5. 歷年各模型違規變異點誤判情形

透過個別分析模型對各縣市的預測結果，可以了解模型對哪些縣市的變異點識別力較弱（如表 4-11），對於這些識別力較弱的地區，可考慮額外進行調查，以深入分析其人文及產業結構，以利調整研究方案，並為各縣市制定具體的配套措施與目標。

歷年資料顯示，南投縣與苗栗縣的變異點數量一直位居全臺前二名，而在本島地區中，嘉義縣的變異點數量則始終最少（如表 1-1）。就三種模型的訓練結果而言（如圖 4-1），南投縣與苗栗縣的違規變異點被誤判為非違規變異點的數量也同樣位居前二。此外，台中市、新竹縣與高雄市也都是容易被誤判的地區。以上偏山區的縣市之誤判情形，或與山坡地道路受限本計畫無引用農路、林道與部落間聯絡道有關，有待後續研究持續加強。

以上結果顯示，南投縣與苗栗縣的土地違規情形需要特別檢核。而對於其他縣市，在進行數據收集與模型訓練前，應先做好前期調查工作，分析該縣市是否在某些面向上與其他縣市有明顯差異需要特別關注。這些前期工作也能幫助我們針對不同縣市進行客製化的模型訓練。

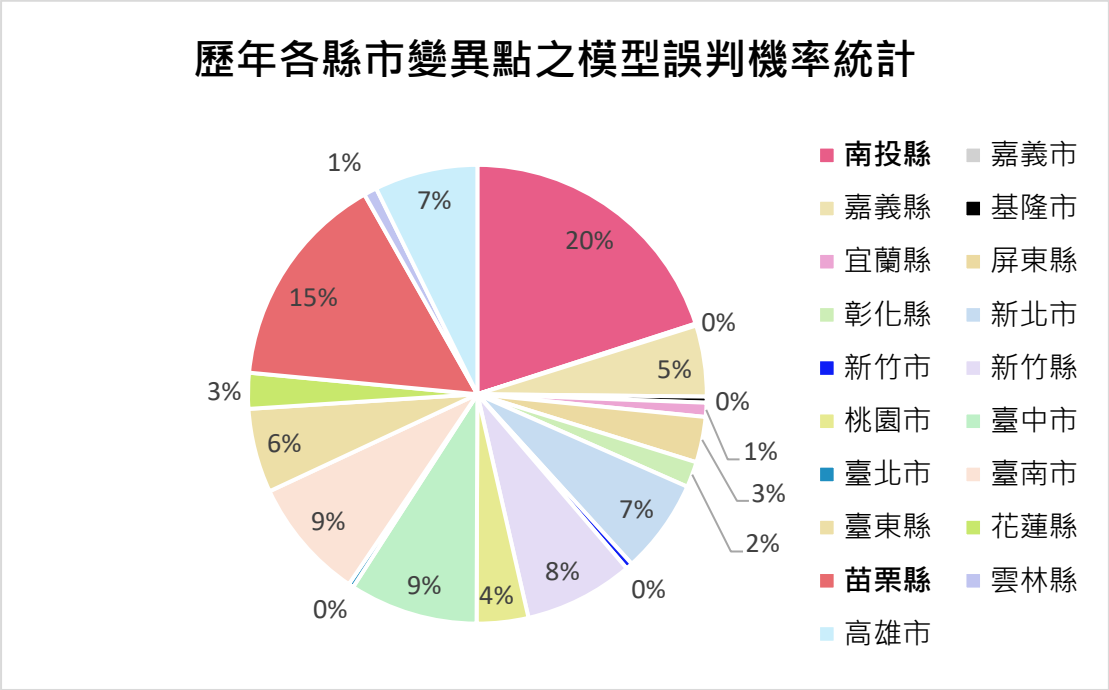

圖 四-1、歷年各縣市變異點每一模型之誤判機率統計

表 四-11、各縣市違規變異點誤判之數量統計

年度	縣市	LightGBM	CatGBM	RF
		違規變異點誤判之數量		
108	南投縣	**586**	**543**	**456**
	嘉義市	5	4	2
	嘉義縣	232	237	206
	基隆市	31	25	22
	宜蘭縣	84	58	46
	屏東縣	178	130	209
	彰化縣	130	132	113
	新北市	188	162	146
	新竹市	9	10	11
	新竹縣	464	345	151
	桃園市	85	79	74
	臺中市	330	312	278
	臺北市	4	2	11
	臺南市	352	332	264
	臺東縣	328	298	379
	花蓮縣	165	108	107
	苗栗縣	**752**	**701**	**327**
	雲林縣	44	41	36
	高雄市	288	277	289
109	南投縣	**683**	**710**	**599**
	嘉義市	11	9	6
	嘉義縣	280	276	215
	基隆市	18	13	9
	宜蘭縣	31	22	16
	屏東縣	140	179	172
	彰化縣	79	77	76
	新北市	240	263	242
	新竹市	53	12	17
	新竹縣	448	338	150
	桃園市	108	99	99
	臺中市	430	445	396
	臺北市	6	3	6
	臺南市	350	358	234
	臺東縣	257	318	303
	花蓮縣	128	138	112
	苗栗縣	**1258**	**1131**	**520**
	雲林縣	26	26	24
	高雄市	372	421	381
110	南投縣	**1383**	**1321**	**1098**
	嘉義市	10	11	14
	嘉義縣	250	271	243
	基隆市	21	23	15
	宜蘭縣	68	48	39
	屏東縣	232	174	184
	彰化縣	74	77	81
	新北市	403	390	384
	新竹市	10	47	11
	新竹縣	376	471	219
	桃園市	207	204	208
	臺中市	401	391	347
	臺北市	8	9	11
	臺南市	624	594	511
	臺東縣	339	195	259
	花蓮縣	138	85	99
	苗栗縣	**648**	**748**	**372**
	金門縣	1	3	2
	雲林縣	23	26	27
	高雄市	409	383	352

年度	縣市	LightGBM	CatGBM	RF
		違規變異點誤判之數量		
111	南投縣	**1068**	**1088**	**981**
	嘉義市	6	7	4
	嘉義縣	196	189	160
	基隆市	27	20	20
	宜蘭縣	59	48	42
	屏東縣	102	70	93
	彰化縣	80	78	81
	新北市	355	352	323
	新竹市	14	14	17
	新竹縣	203	215	176
	桃園市	214	215	204
	臺中市	492	487	352
	臺北市	27	22	38
	臺南市	460	451	325
	臺東縣	183	150	211
	花蓮縣	85	66	91
	苗栗縣	**512**	**511**	**392**
	金門縣	2	4	1
	雲林縣	52	57	51
	高雄市	342	328	309
112	南投縣	**1112**	**1029**	**895**
	嘉義市	2	2	2
	嘉義縣	239	216	179
	基隆市	12	16	15
	宜蘭縣	42	33	37
	屏東縣	108	84	118
	彰化縣	45	48	55
	新北市	350	335	366
	新竹市	29	30	36
	新竹縣	568	586	502
	桃園市	207	229	230
	臺中市	531	512	427
	臺北市	25	26	43
	臺南市	362	328	263
	臺東縣	263	239	295
	花蓮縣	107	136	140
	苗栗縣	**867**	**870**	**754**
	連江縣	17	17	16
	金門縣	6	6	18
	雲林縣	70	56	63
	高雄市	264	256	253

6. 主成份分析（PCA）成果

主成份分析在理解資料的底層結構方面扮演著關鍵角色。透過觀察主成分的投影圖（圖 4-2），點的分布模式能夠展示數據的結構特徵。當點在第一和第二主成分平面上呈現清晰的分群時，表示這兩個主成分能有效區分數據中的不同類別。然而，若點的分布較為分散或混雜，則可能需要考慮增加主成分數量或採用其他分析方法。

從本研究的分析結果可知（圖 4-2），數據結構具有低解釋變異性:第一主成分(Principal Component 1)僅解釋了 22.87%的變異量，第二主成分(Principal Component 2)也只解釋了 14.66%的變異量，顯示模型無法有效捕捉數據的主要特徵。另外,進一步觀察碎石圖(圖 4-3 左）可以看出，各個主成份的解釋變異比例下降較為緩慢；而累積解釋變異圖（圖 4-3 右）則顯示需要 7-8 個主成分才能達到 90%以上的解釋率，這些都反映了本研究數據具有高維度的複雜性。

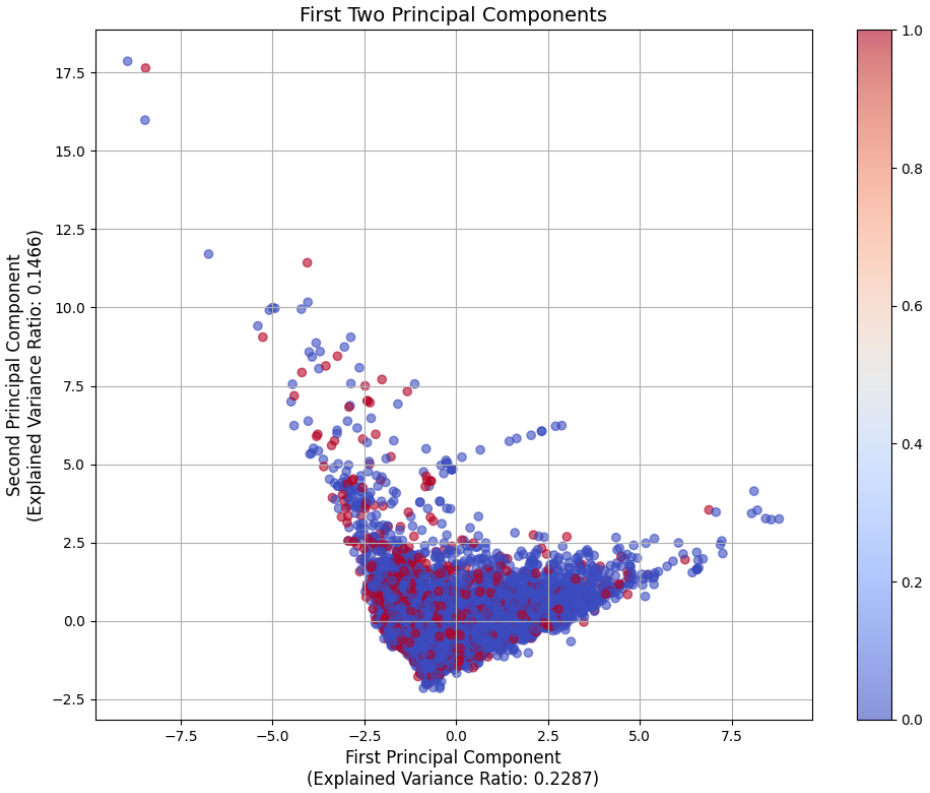

圖 四-2、主成份投影圖

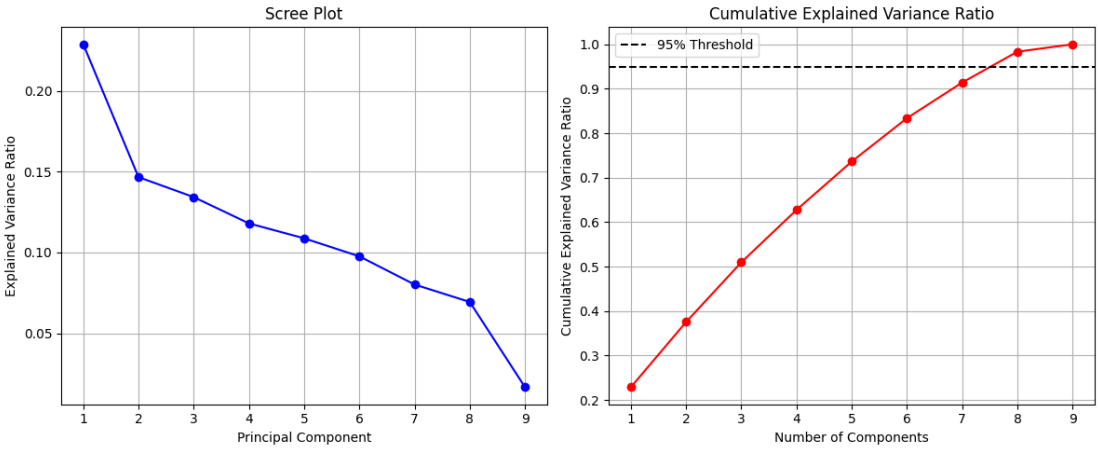

圖 四-3、碎石圖（Scree Plot）與累積解釋變異量圖

7. 違規變異點機率分布地圖

本研究首先以長時長與短時長兩種方法進行數據分析預測。在長時長方法中（圖 4-4 ~ 4-6）為所有模型採用迭代的方式，分別將民國 108-112 年中的單一年度作為測試集，其餘四年數據合併作為訓練集，用以訓練模型辨識違規變異點，並繪製違規變異點機率分佈地圖。地圖上每個點的機率值代表該位置可能為違規土地利用的程度。在三種模型中，Random Forest 模型的表現最不理想（圖 4-6）。其產製的機率分布地圖顯示，幾乎所有變異點都被判定為非違規變異點（圖中多為冷色區域，代表違規機率偏低），導致無法有效篩選出真正需要實地勘查的違規利用地點，該結果無法達成減少人力資源消耗並提高違規變異點檢核效率的研究目標。

在短時長方法部分（圖 4-7 至 4-21），我們選擇民國 108-112 年中的單一年份作為訓練集，用於預測其他四年的數據，以此來評估數據收集時間長短對預測精度的影響。根據統計結果（表 4-6 至 4-10），單一年份的訓練數據量不足以讓模型進行有效學習，而從機率分布地圖可以觀察到，短時長數據預測結果在相鄰區域皆呈現均勻的機率分布，缺乏區域識別性；除此之外，預測為違規土地利用點位的機率普遍偏低，因此驗證短時長的方法無法有效協助政府相關單位篩選出非違規點位，故也無法達到降低成本消耗的研究目標。

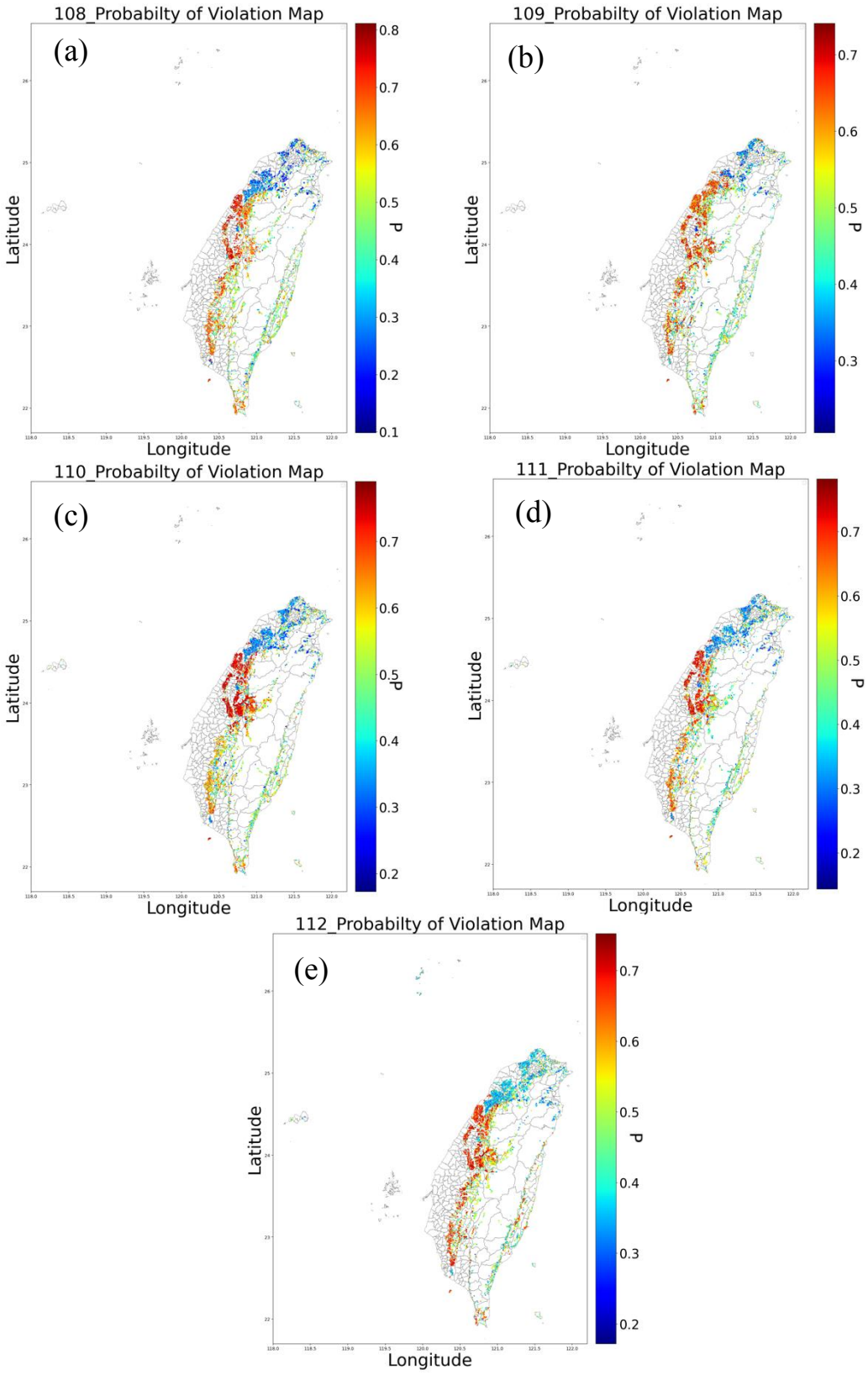

圖 四-4、長時長 LightGBM 變異點違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 111 年、(e)民國 112 年

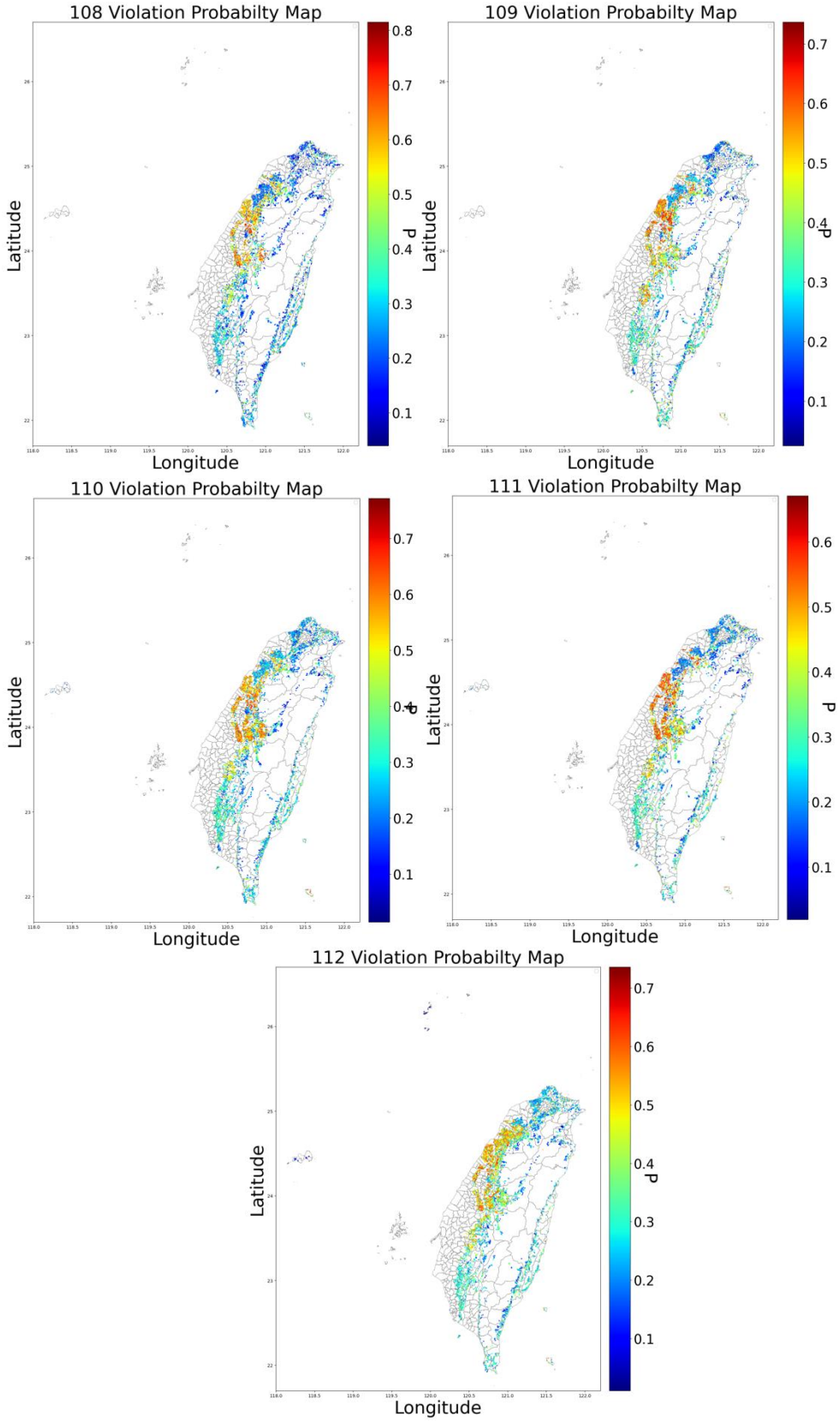

圖 四-5、長時長 CatBoost 變異點違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 111 年、(e)民國 112 年

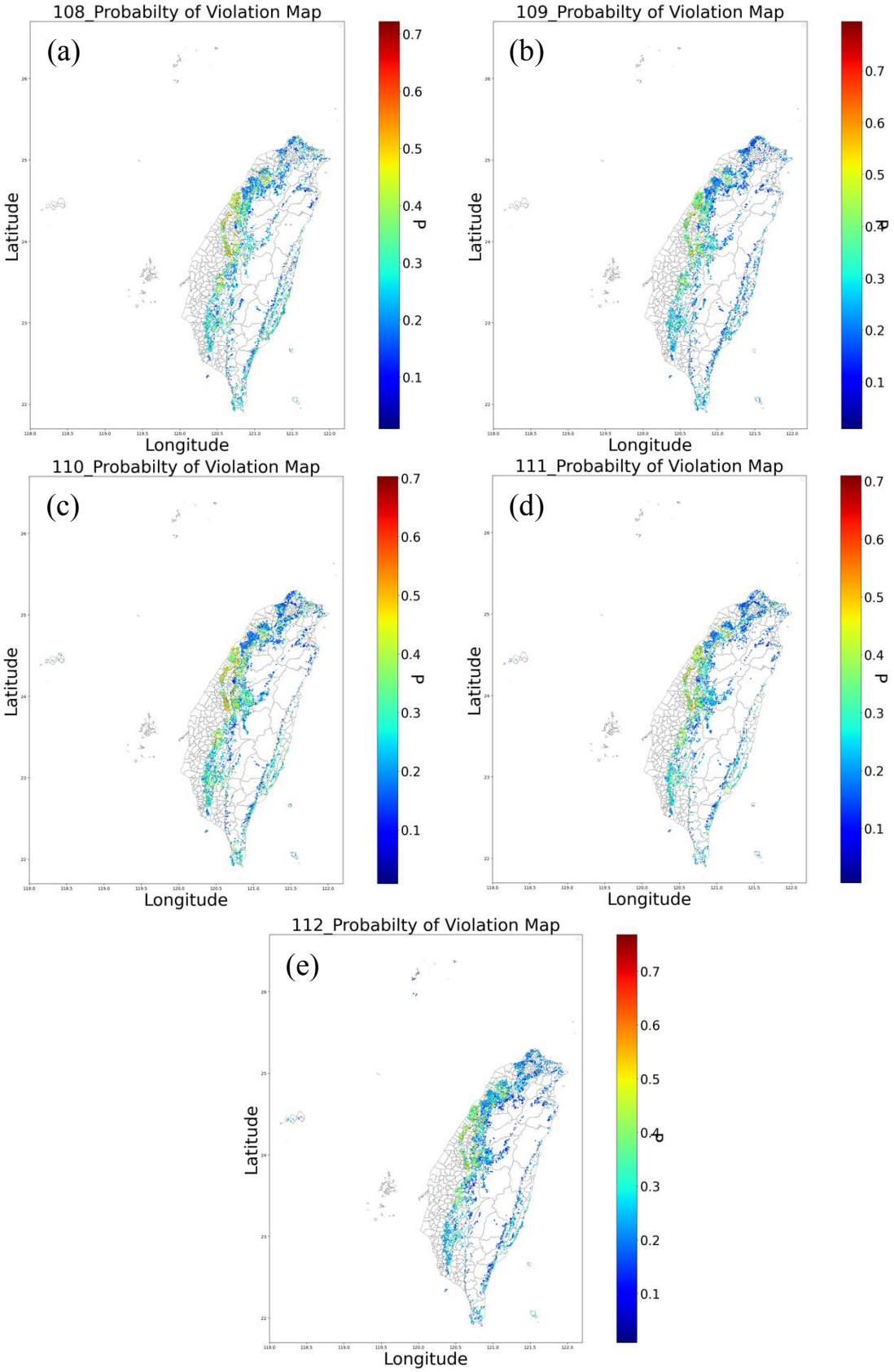

圖 四-6、長時長 RF 變異點違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 111 年、(e)民國 112 年

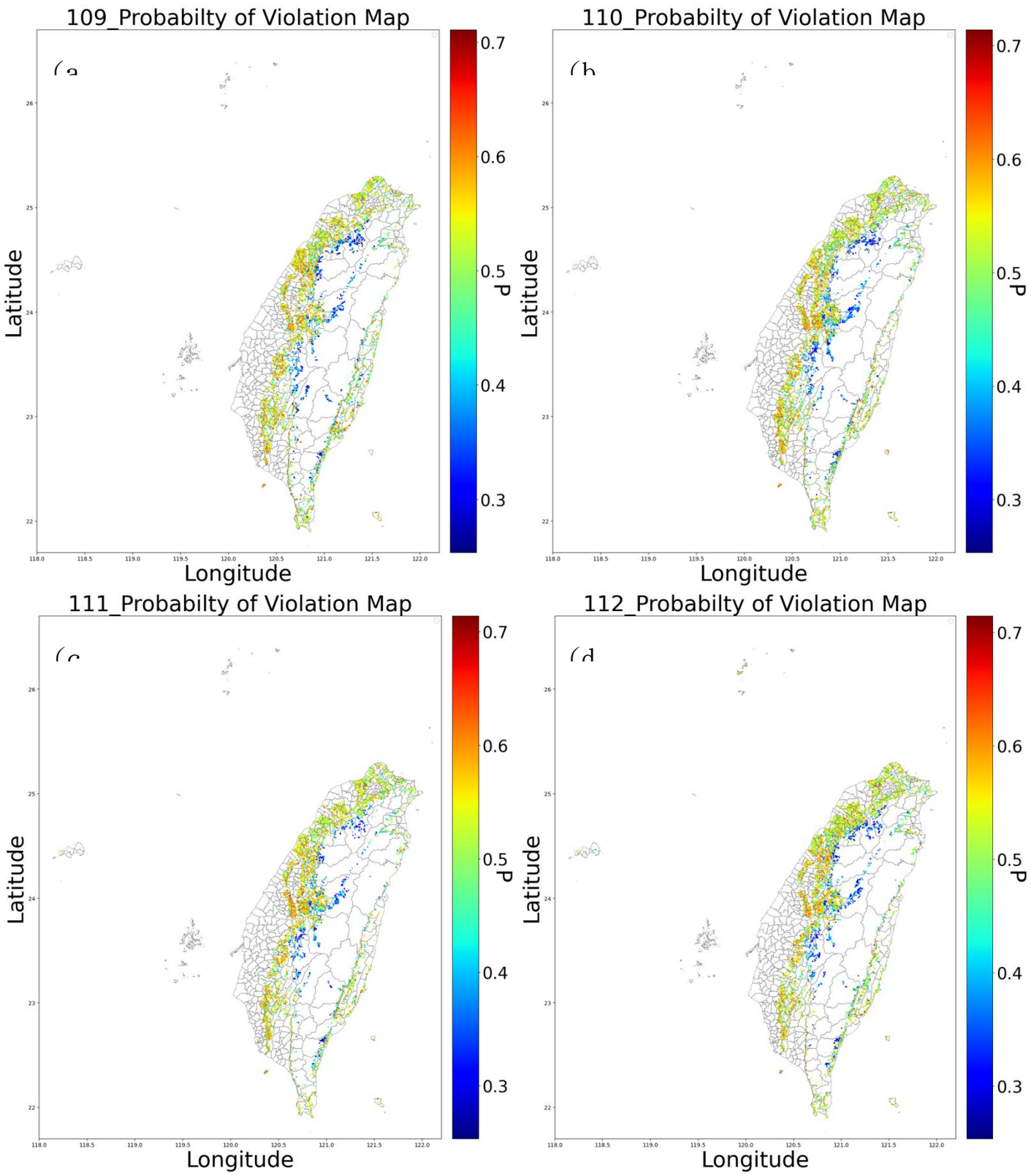

圖 四-7、以民國 108 年單一年份訓練 LightGBM 之違規機率圖，(a)民國 109 年、(b)民國 110 年、(c)民國 111 年、(d)民國 112 年

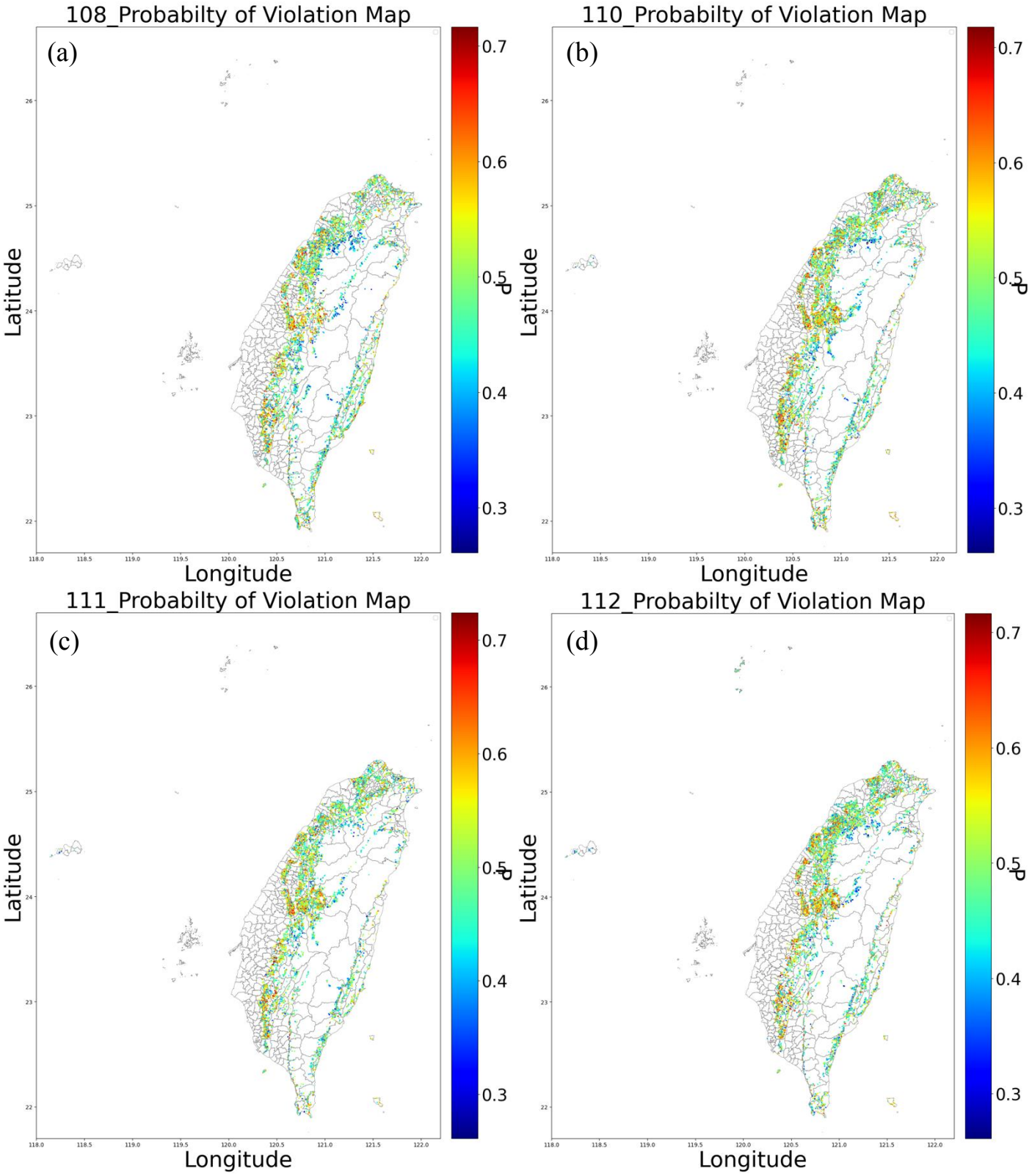

圖 四-8、以民國 109 年單一年份訓練 LightGBM 之違規機率圖，(a)民國 108 年、(b)民國 110 年、(c)民國 111 年、(d)民國 112 年

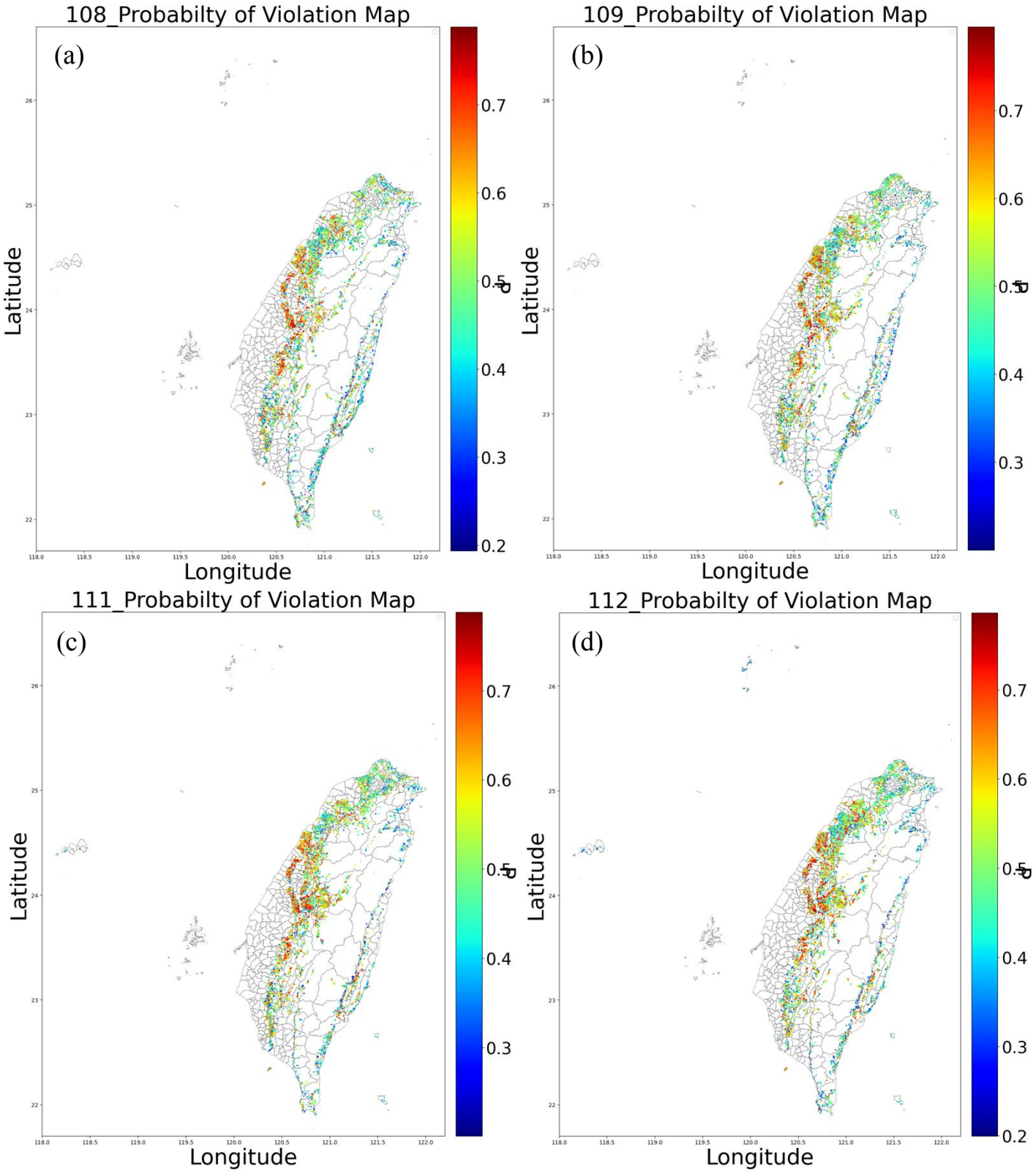

圖 四-9、以民國 110 年單一年份訓練 LightGBM 之違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 111 年、(d)民國 112 年

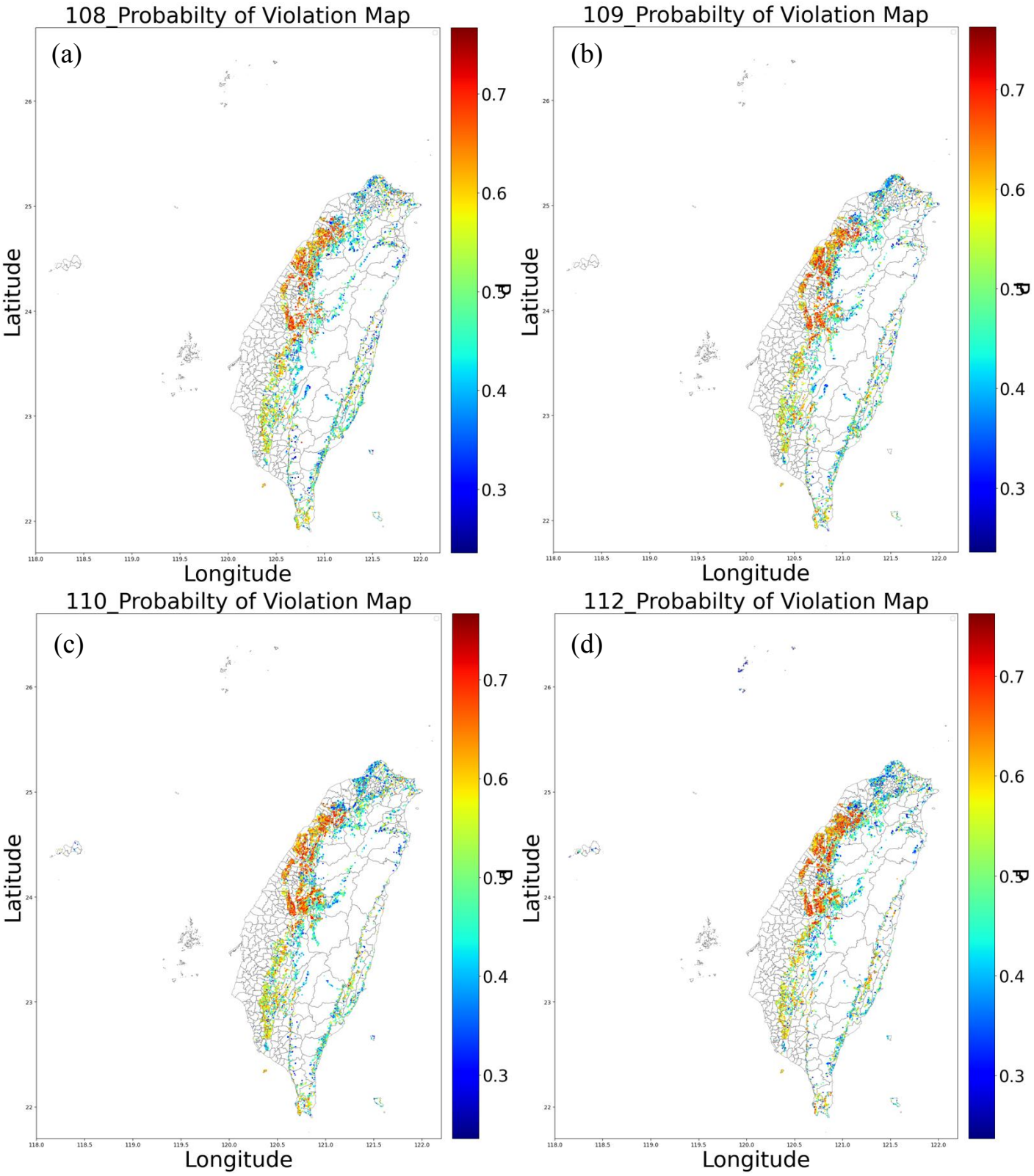

圖 四-10、以民國 111 年單一年份訓練 LightGBM 之違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 112 年

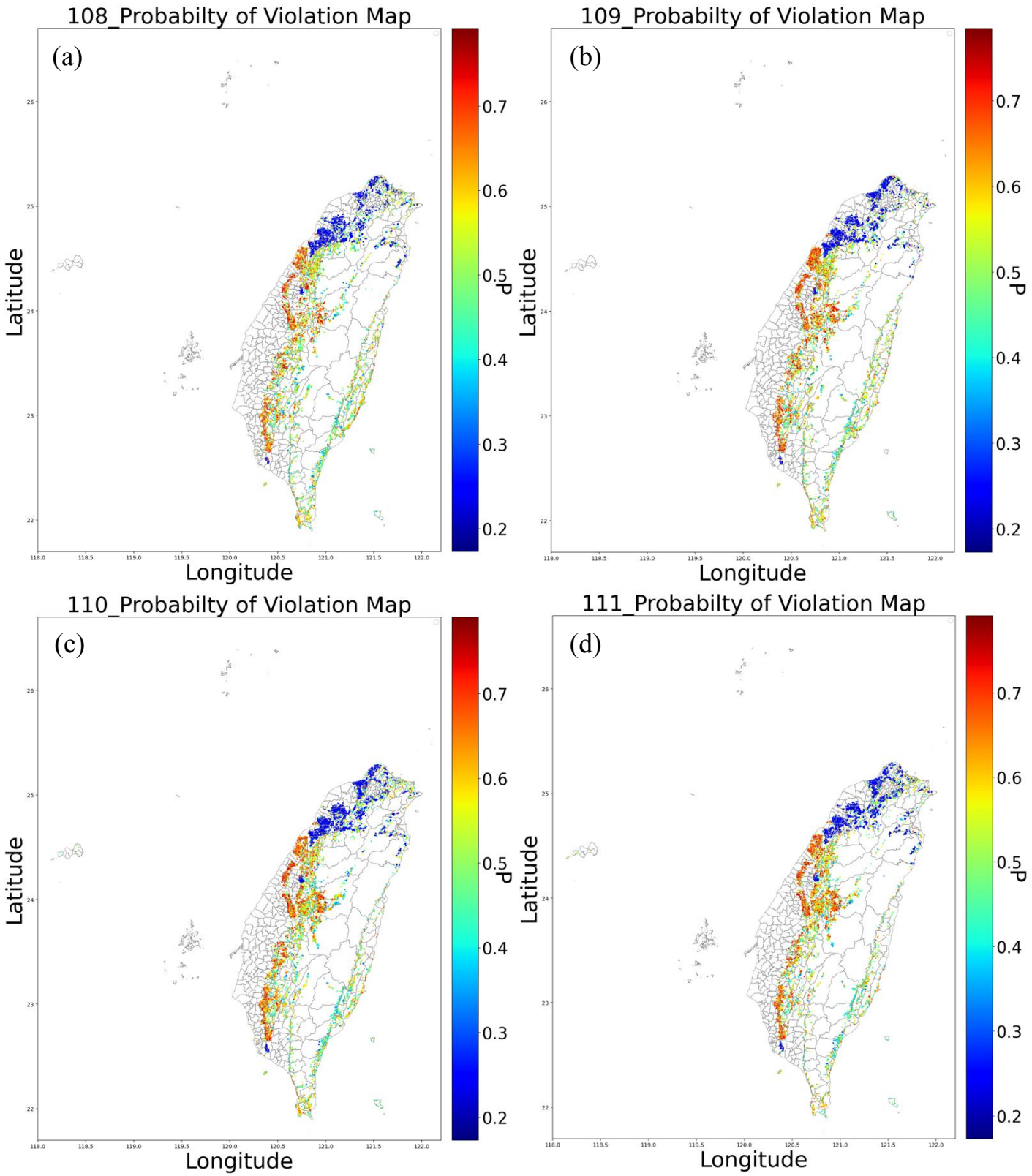

圖 四-11、以民國 112 年單一年份訓練 LightGBM 之違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 111 年

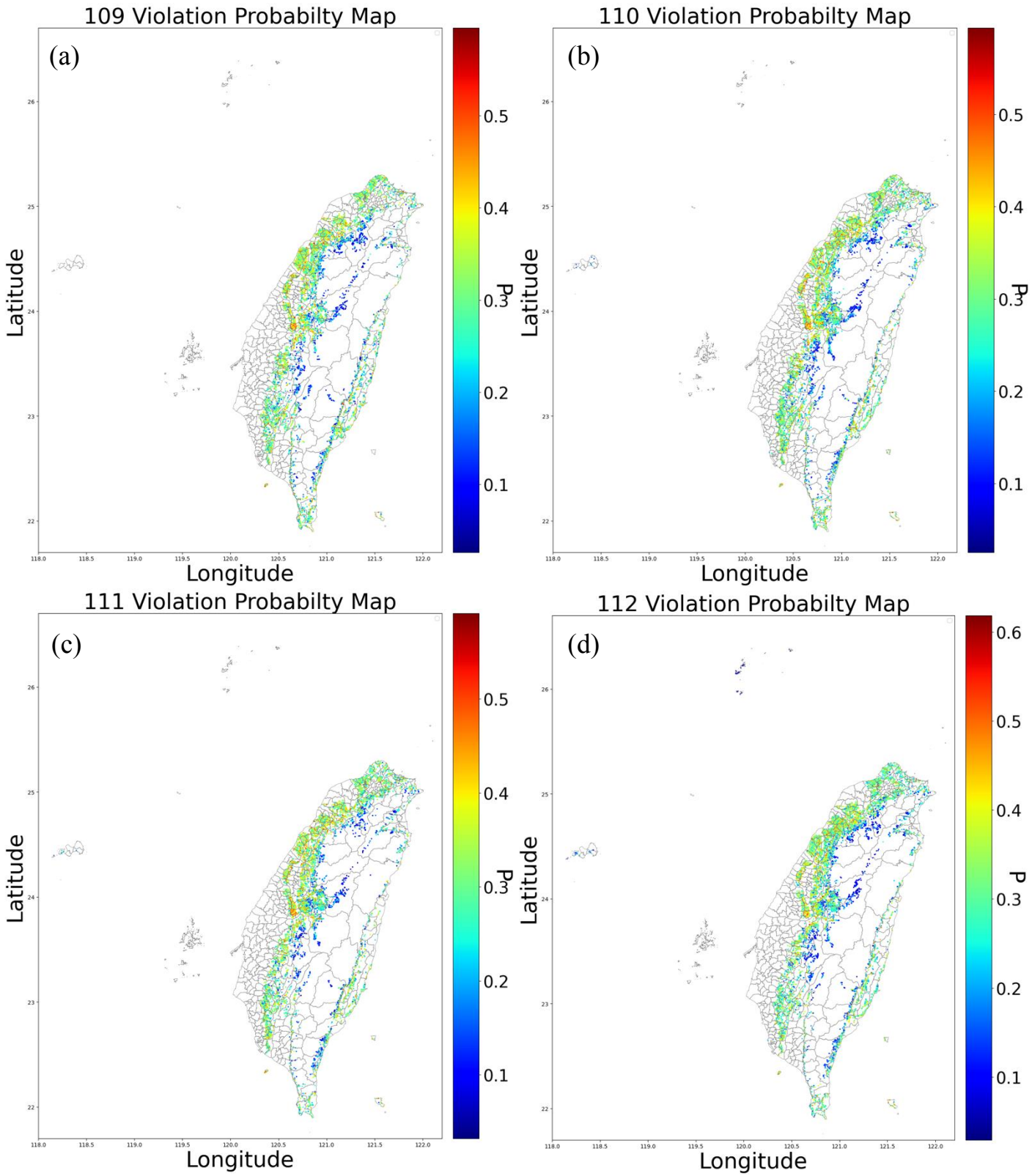

圖 四-12、以民國 108 年單一年份訓練 CatBoost 之違規機率圖，(a)民國 109 年、(b)民國 110 年、(c)民國 111 年、(d)民國 112 年

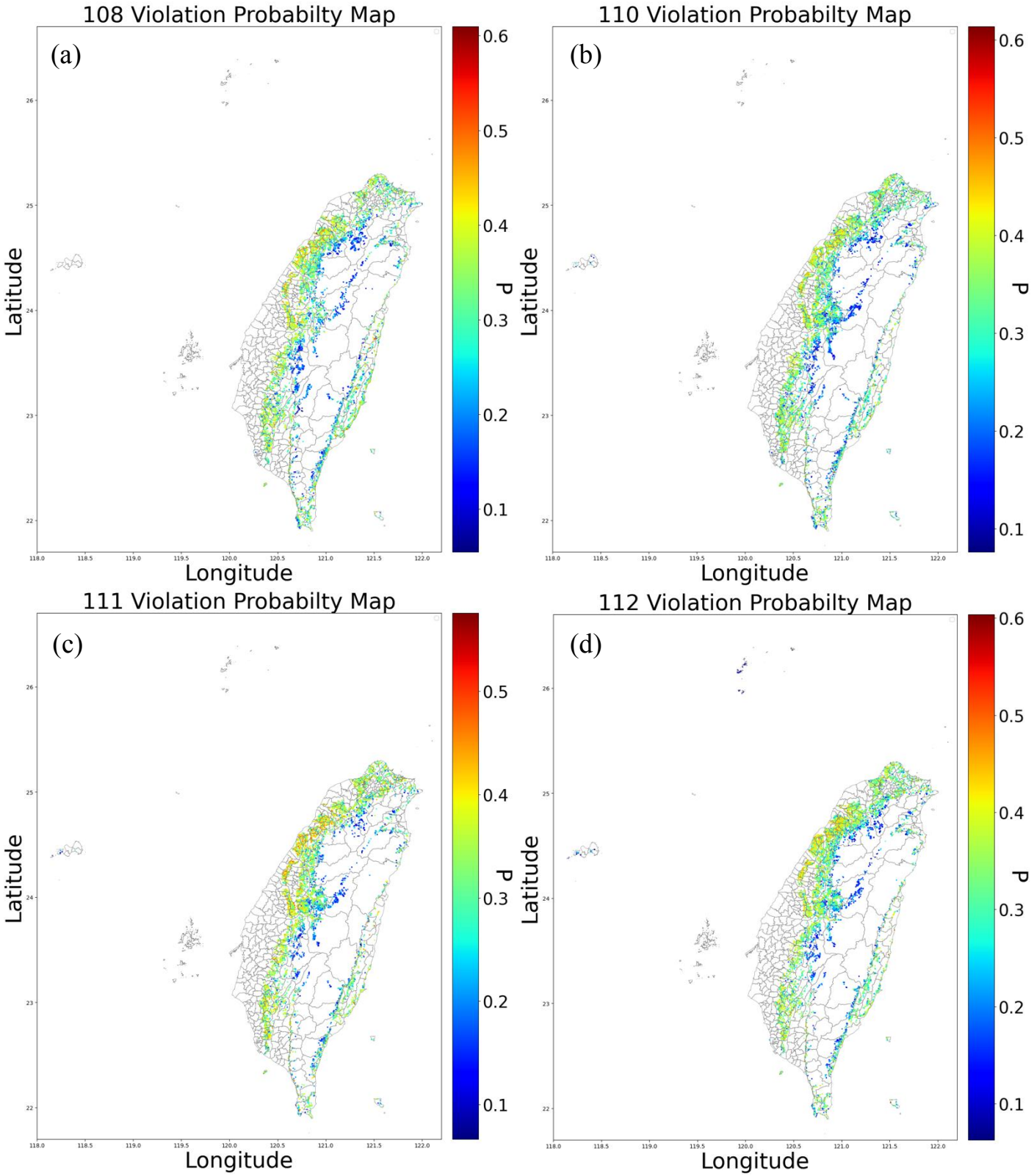

圖 四-13、以民國 109 單一年份訓練 CatBoost 之違規機率圖，(a)民國 108 年、(b)民國 110 年、(c)民國 111 年、(d)民國 112 年

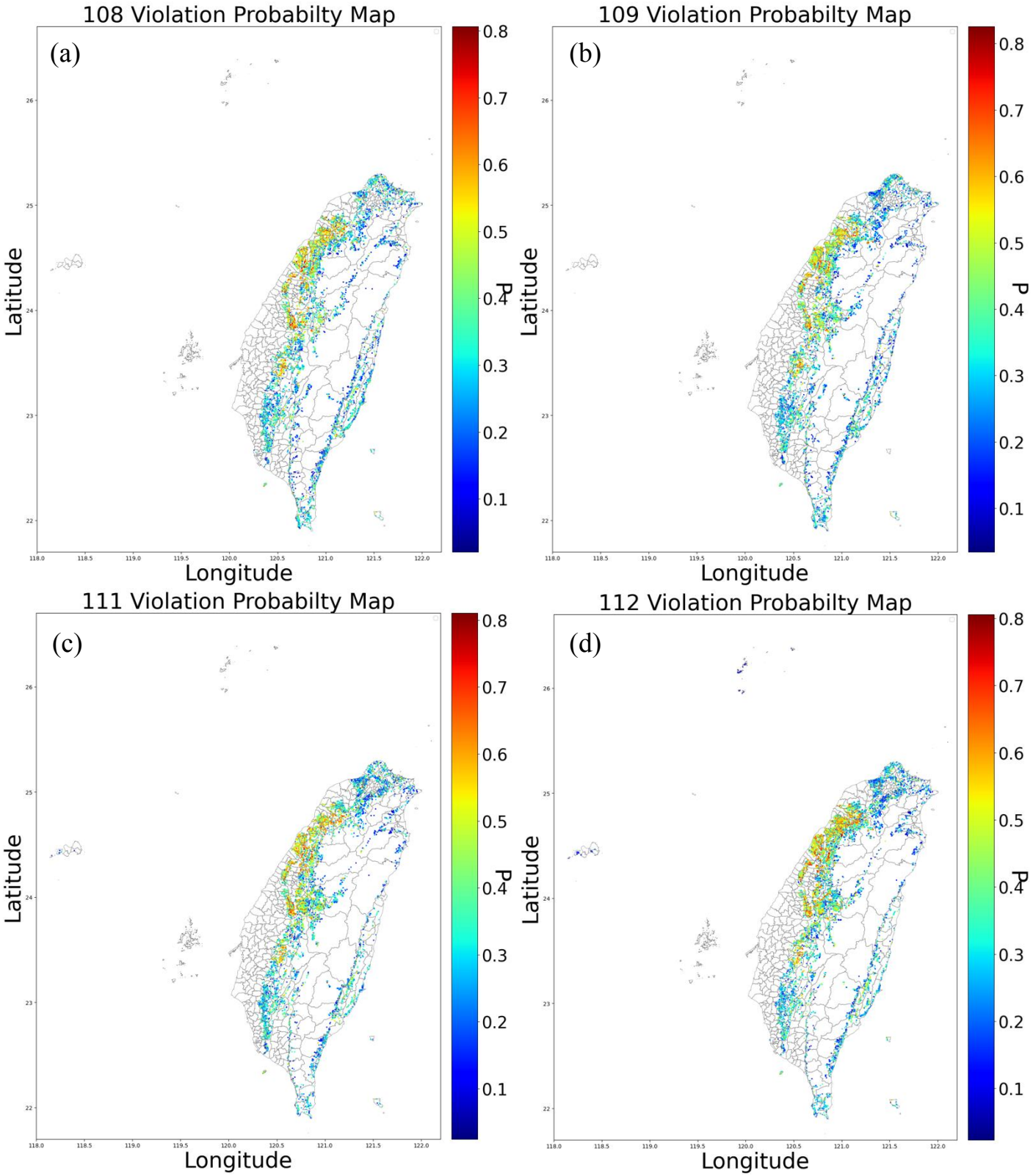

圖 四-14、以民國 110 單一年份訓練 CatBoost 之違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 111 年、(d)民國 112 年

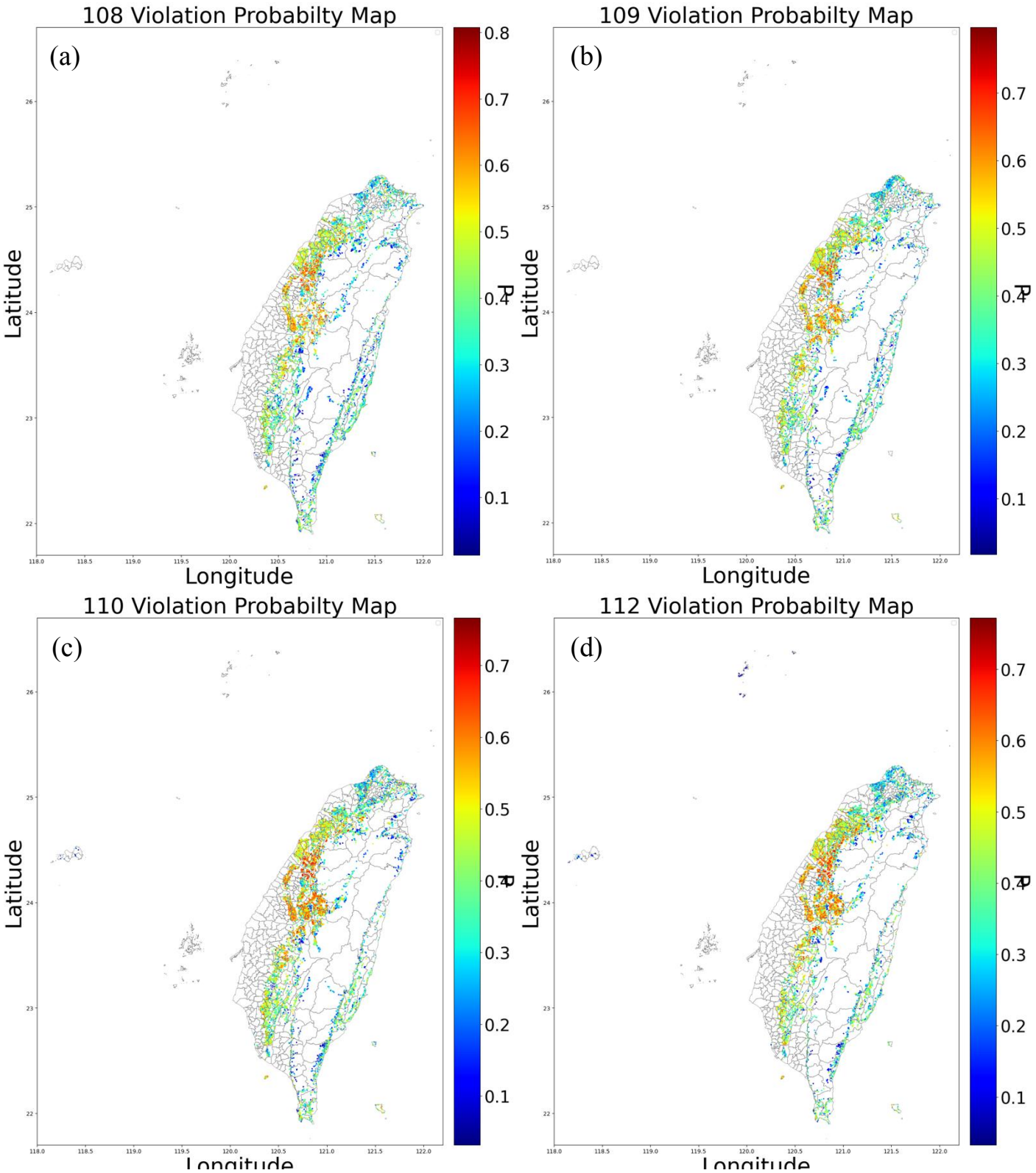

圖 四-15、以民國 111 單一年份訓練 CatBoost 之違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 112 年

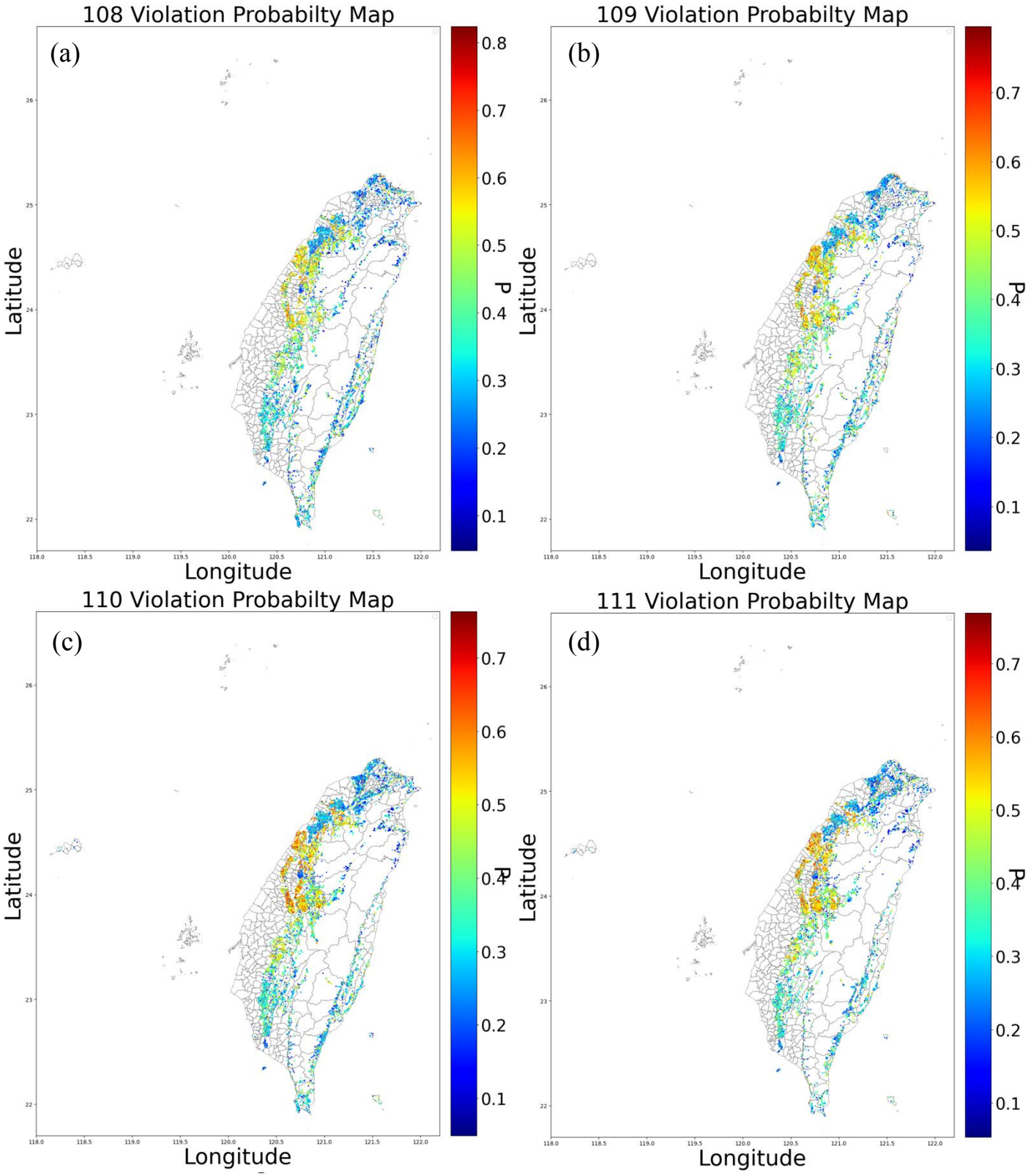

圖 四-16、以民國 112 單一年份訓練 CatBoost 之違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 111 年

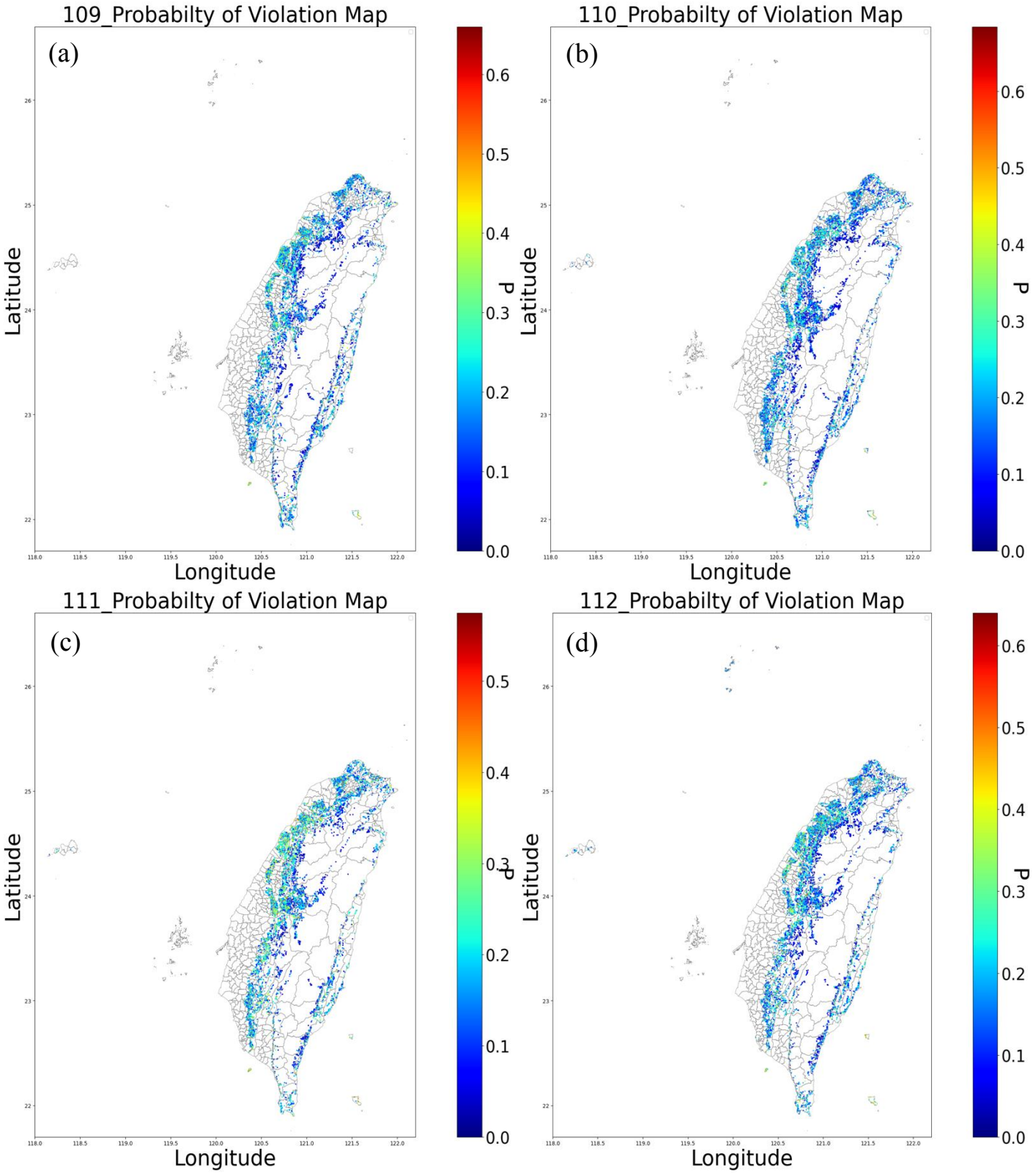

圖 四-17、以民國 108 年單一年份訓練 RF 之違規機率圖，(a)民國 109 年、(b)民國 110 年、(c)民國 111 年、(d)民國 112 年

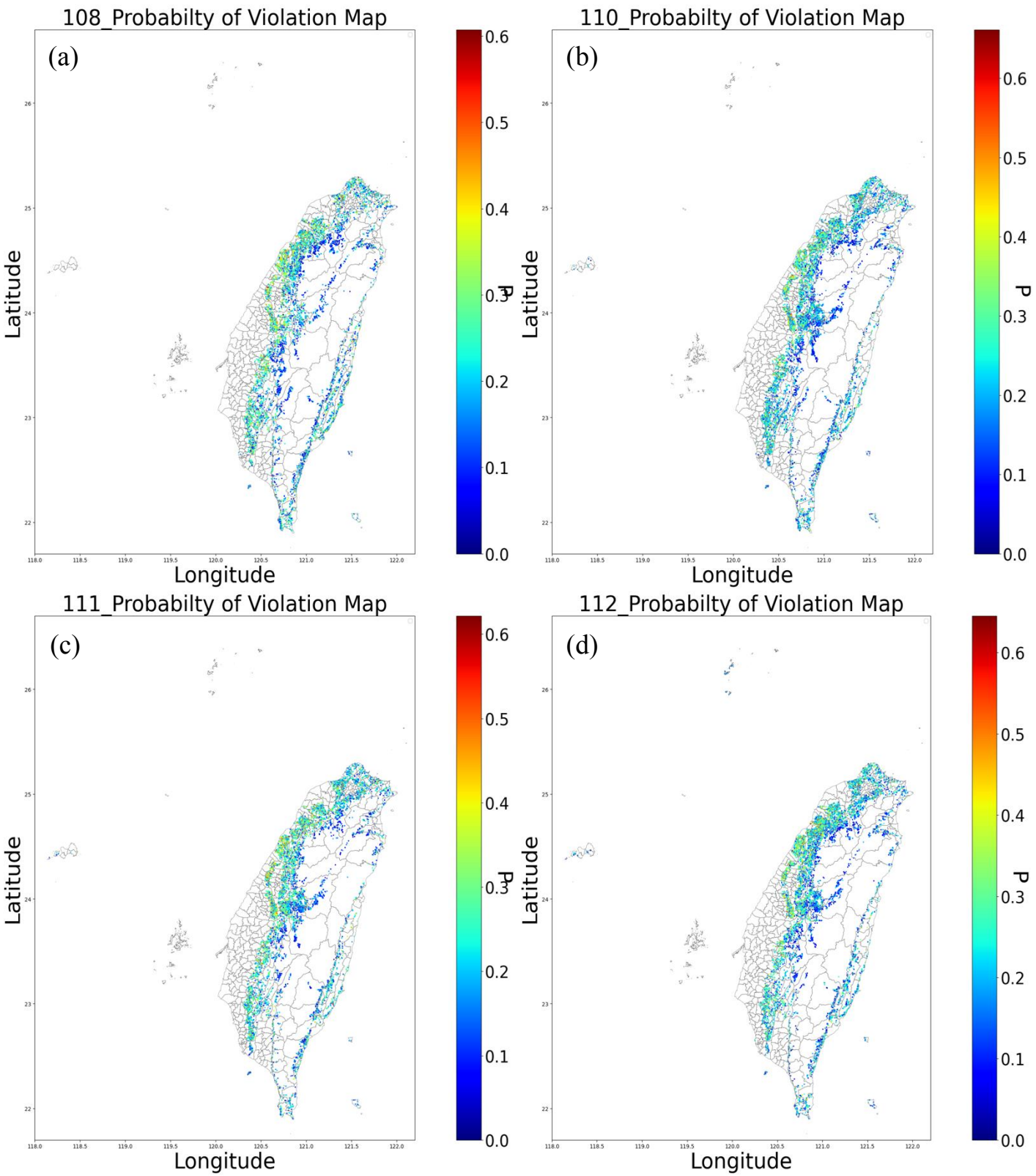

圖 四-18、以民國 109 單一年份訓練 RF 之違規機率圖，(a)民國 108 年、(b)民國 110 年、(c)民國 111 年、(d)民國 112 年

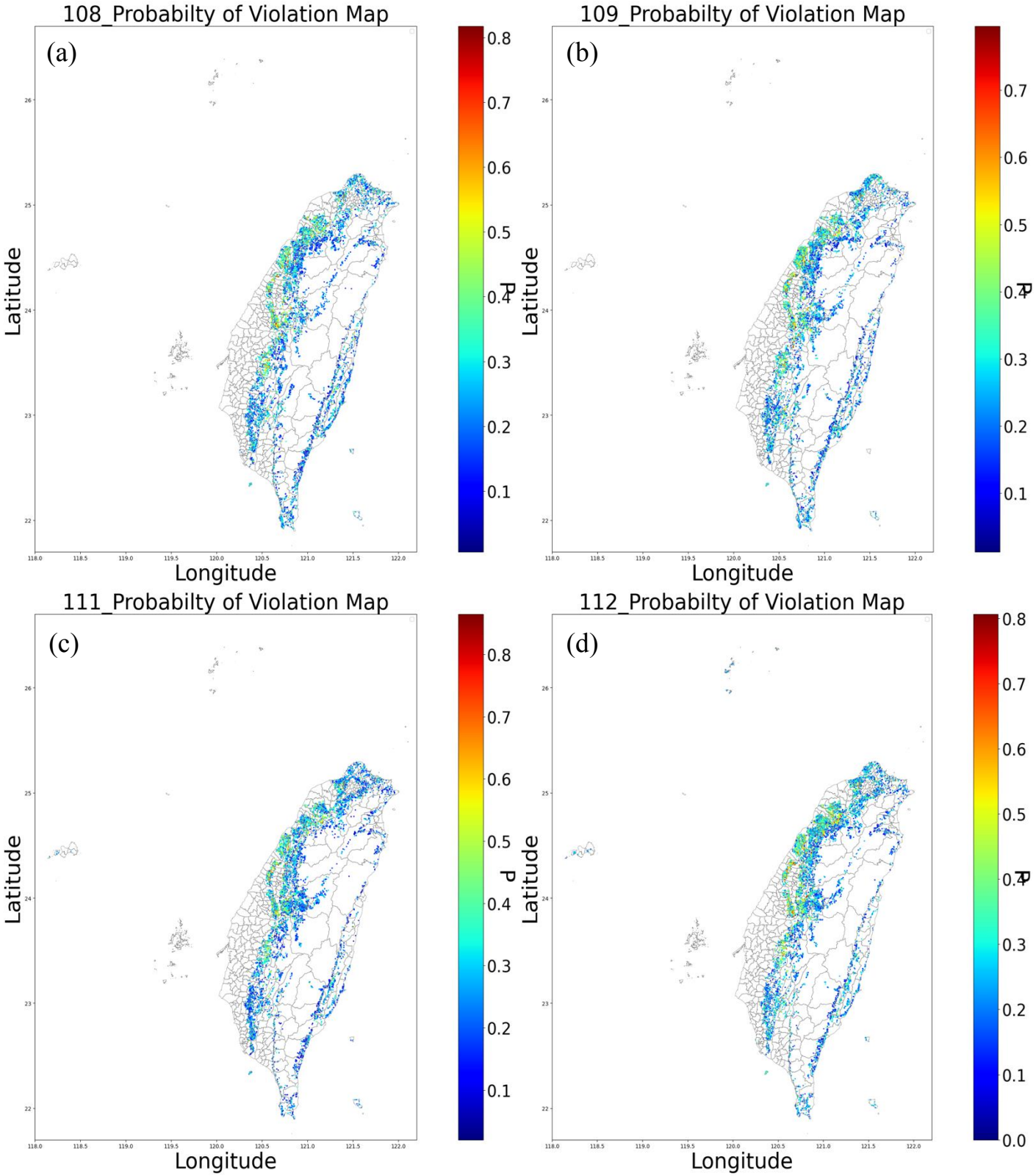

圖 四-19、以民國 110 單一年份訓練 RF 之違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 111 年、(d)民國 112 年

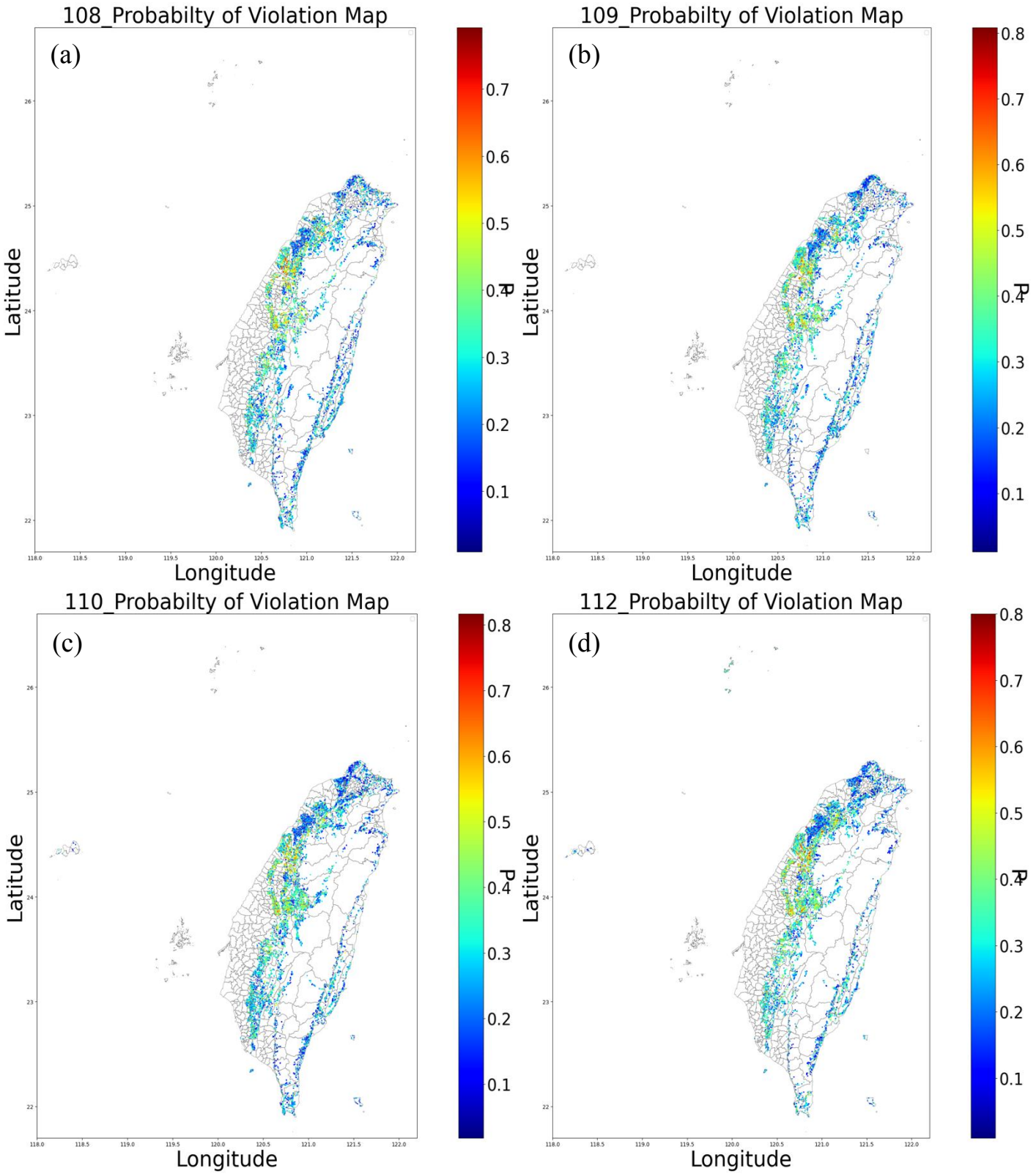

圖 四-20、以民國 111 單一年份訓練 RF 之違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 112 年

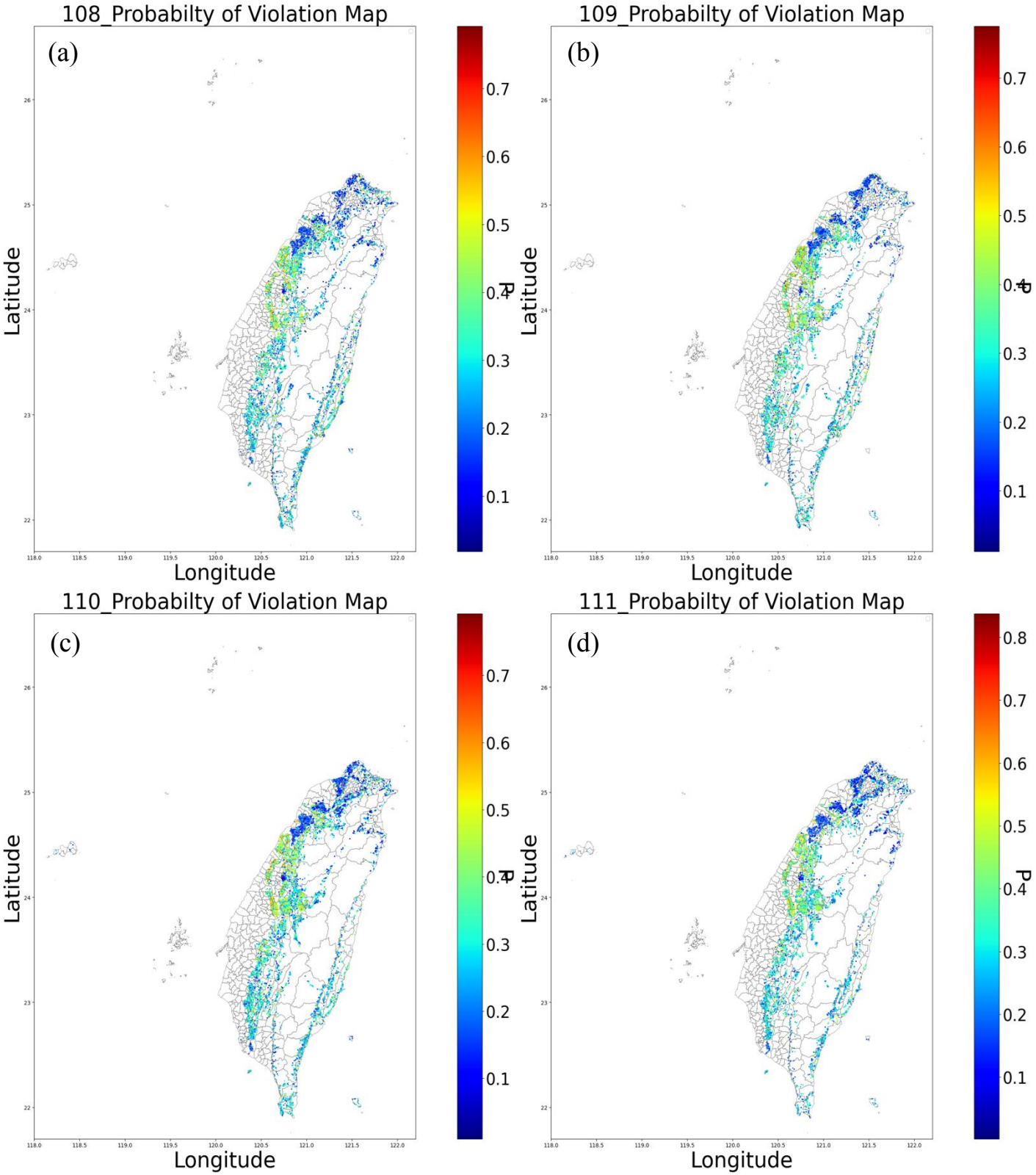

圖 四-21、以民國 112 單一年份訓練 RF 之違規機率圖，(a)民國 108 年、(b)民國 109 年、(c)民國 110 年、(d)民國 111 年

第五章　結論與建議

在本研究進行的各式測試中可以發現對於違規變異點的判識精度不高，無論各種年度組合僅達50%左右，不排除數據中可能包含大量異常值與雜訊，同時必須考慮違規變異點具有高度複雜度的特性，不同違規樣態背後或有極其複雜之地方政治、文化及其他干擾因素，導致模型難以有效區分違規與非違規的特徵。此外基於PCA之分析，未來研究方向應著重於重新評估資料的選擇，並加強數據收集的質與量，以確保模型能更準確地捕捉違規與非違規行為的特徵，進而提升預測精度。

本計畫執行迄今耗費較多時間與人力於資料收集、格式整理、除錯過濾等步驟，此繁瑣程序亦為機器學習與深度學習過程中常見之隱形成本。有鑒於人工智慧使用愈趨頻繁，後續對於地理資訊資料標準化、資料介接普及化以及單一入口資料平台等需求將日益升高。因此建議業務單位可研擬非機敏資料的開放資料平台，如土地利用、人口、道路直接或介接提供與業務範疇相關之常用資料，可擴大各型計畫成果的影響力，並觸及更多相關領域的資料分析人員。同時，獲取更多變異點周邊資料有助於深度學習模型抓取非線性特徵，有助提升判識精度。

參考文獻

Goetz, J. N., Brenning, A., Petschko, H., & Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & geosciences*, *81*, 1-11.

Brenning, A. (2005). Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, *5*(6), 853-862.

Reis, S. (2008). Analyzing land use/land cover changes using remote sensing and GIS in Rize, North-East Turkey. *Sensors*, *8*(10), 6188-6202.

Schapire, R. E. (1990). The strength of weak learnability. Machine learning, 5, 197- 227.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. arXiv preprint

arXiv:1810.11363.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Tong, W., Hong, H., Fang, H., Xie, Q., & Perkins, R. (2003). Decision forest: combining the predictions of multiple independent decision tree models. Journal of Chemical Information and Computer Sciences, 43(2), 525-531.

Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Colditz, R. R. (2015). An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. Remote Sensing, 7(8), 9655-9681.

Wang, H., Zhao, Y., Pu, R., & Zhang, Z. (2015). Mapping Robinia pseudoacacia forest health conditions by using combined spectral, spatial, and textural information extracted from IKONOS imagery

and random forest classifier. remote Sensing, 7(7), 9020-9044.

Setianto, A., Triandini, T., 2015. COMPARISON OF KRIGING AND INVERSE DISTANCE WEIGHTED (IDW) INTERPOLATION METHODS IN LINEAMENT EXTRACTION AND ANALYSIS. Journal of Applied Geology 5.. https://doi.org/10.22146/jag.7204

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). “mice: Multivariate Imputation by Chained Equations in R”. Journal of Statistical Software 45: 1-67.

Rubin DB (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York.

Resche-Rigon, M., & White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. Statistical methods in medical research, 27(6), 1634-1649.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3), 1-21.

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.

附錄

附錄一、期初審查會議紀錄暨回覆辦理情形

項次	審查意見	回覆辦理情形
報告內容審查意見：		
一	依本計畫進度於7月期間應有進行少數成果分析，則計畫內文之格式亦應有結果討論，而非逕自說明結論與建議，請審視；又部分文獻引用僅以中括號標示，而無作者標示或僅有2個字之引用，不易閱讀對照，建請檢視。	已於期末報告修正。
二	報告書內文有未說明即直接列出圖及表，或列出圖及表未進行說明情形，請檢視。又相關附圖不清晰，亦請調整。	已於期末報告修正。
三	1-2 違規變異點建請就新增、改正(善)後減除酌加說明。	已補充說明。

項次	審查意見	回覆辦理情形
四	1-6 臺灣近年來人口樹整體上升趨勢，是否正確，請檢視。	已修正文字說明為「…臺灣近年來的總人口數於109年達到峰值…」。
五	2-3 表2-1內之缺失值比例以萬分之5，對於選用數據來源應相關關聯說明。	已新增資料來源為政府開放資料平台(https://data.gov.tw)。
六	2-9 圖2-1之模型測試建議更改圖塊。	已於期末報告修正。
七	2-11違規樣態應可作為模型表現不佳之問題節點，建請納入討論。	已於期末報告主成分分析章節加入討論內容。
八	4-1依據表4-1與表4-2出現不同生產者精度，而此值與整體準確度又有何關聯，且非違規變異點應是委辦計畫單位所需，建請再就偏差審酌進行相關說明。	已於期末報告補充說明。

項次	審查意見	回覆辦理情形
九	4-2依表4-3最低違規機率在偏高卻有偏差低時，即可假定為較完善模型學習，請補充依據。	已於期末報告修正。
十	4-3表4-4應有顯著因子之徵值來進行比較說明，而非僅以字體顏色說明顯著性。	本表中顯著因子為整合三種模型評估而得。
十一	4-4違規變異點被模型誤判之原因，遺在日後期末報告呈現。山坡地道路受限本計畫無引用農路、林道與部落間聯絡道，建請酌量納入。	已於期末報告補充說明。
十二	4-6圖4-1~圖4-3如何視出違規及非違規等變異點請檢視，又相關圖4-3有RF將變異點皆判斷為非違規變異點之情形，卻有其整體準確度偏高情形，是否再確認立論。	已於期末報告補充說明「…(圖中多為冷色區域，代表違規機率偏低)…」。
十三	經計算模型誤判為違規行為，是否得以再區分誤判之態樣，以保持模型的準確性和有效性。	在無新增訓練資料情況下，難以重新訓練獲得不同成果。

項次	審查意見	回覆辦理情形
十四	本案建議業務單位可研擬非機敏資料的開放資料平臺，基於運用相關技術的同時，仍需遵守保障個人隱私與數據安全等相關法規，爰本案是否有建議那些必要常用資料數據，以供業務單位研析是否屬非機敏資料。	已於期末報告修正。

附錄二、期末聯合審查會議紀錄暨回覆辦理情形

項次	審查意見	回覆辦理情形
報告內容審查意見：		
一	若機器學習無法反映特徵之間的關係，建議可以嘗試利用深度學習模型(擅長抓取非線型特徵)。	已於建議中補充說明後續精進作法。
二	目前使用的資料是否是以人類可能抵達的位置來進行初步判釋為高機率違規行為?又老師提及今年度以全台資料進行分析判釋,可能會影響結果(資料混雜模糊重點資訊),倘只選取小區域進行分析,有沒有可能因為不同縣市的發展樣態不同(如東部主要為農業發展、都市區大多是建築開發等)，而導致無法適用於全台?	本案使用資料皆為往年通報及現場勘查後之回報資料，故納入道路及人口資料作為訓練樣本測試其相關性。 因各縣市之違規特徵或有些許不同，故本案建議後續可縮小測試區域，針對特定違規樣態可能發生的區域及熱點進行方法學探索，若有初步成果再拓展至其他縣市並進行模型調整。