•滑坡堰塞湖灾害机理与防控•

DOI:10.15961/j.jsuese.202201271

本刊网刊

滑坡易发性预测建模的不确定性:不同“非滑坡样本”选择方式的影响

黄发明[1], 曾诗怡[1], 姚　池[1*], 熊浩文[1], 范宣梅[2], 黄劲松[3]

(1.南昌大学 工程建设学院, 江西 南昌 330031; 2.成都理工大学 地质灾害防治与地质环境保护国家重点实验室, 四川 成都 610059; 3.纽卡斯尔大学 岩土科学与工程卓越研究中心, 纽卡斯尔 2287)

摘　要: 滑坡易发性预测建模中如何选择非滑坡是影响建模结果的重要不确定因素。为研究不同非滑坡选择方式的影响规律, 拟用5种方式, 即全区随机、坡度低于5°区域、滑坡缓冲300 m外区域、信息量(IV)法、半监督法来选择出与滑坡等比例的非滑坡样本; 进一步将各选择方式与随机森林(RF)耦合构建随机RF、低坡度RF、缓冲区RF、IV–RF及半监督RF等模型。以江西南康区为例, 获取高程、岩性、公路密度等19种环境因子和233个滑坡编录, 将滑坡编录划分为2 598个滑坡栅格单元构建上述耦合模型的输入–输出数据集。再采用预测精度和易发性指数分布等指标分析其建模不确定性。进一步针对耦合模型预测的滑坡易发性指数分布不合理等问题, 在半监督RF建模时采用滑坡与非滑坡比例为1∶2的样本集开展建模并与1∶1等比例样本集工况作对比。结果表明: 1)低坡度RF、缓冲区RF、IV–RF和半监督RF等模型的预测精度均大幅优于随机RF模型, 可见准确选择非滑坡样本对易发性建模至关重要; 2)半监督RF模型选择非滑坡样本的建模性能最优, 且半监督RF在滑坡∶非滑坡=1∶2比其在1∶1时预测的易发性指数分布规律更准确可信。后续研究中有必要更深入探索滑坡与非滑坡样本的比例问题。

关键词: 滑坡易发性预测; 非滑坡样本选择; 半监督机器学习; 信息量; 随机森林

中图分类号: P642.22　　**文献标志码:** A　　**文章编号:** 2096-3246(2024)01-0169-14

滑坡对生命和财产造成的破坏十分严重且影响范围较大[1–2]。滑坡易发性预测作为滑坡风险评估的基础非常重要, 在GIS和机器学习等技术快速发展的背景下, 利用多学科交叉融合方式开展滑坡易发性建模已成为滑坡风险评估的有效工具之一[1,3–4]。

当前, 滑坡易发性建模(LSP)过程主要包括获取滑坡编录与环境因子、划分模型训练/测试集、确定合适的机器学习模型、分析预测结果的不确定性等步骤[5]。其中, 存在各种影响建模结果的不确定性, 例如数据测量、不同联接方法、不同数据驱动模型的不确定性等等, 已有研究分析了部分因素的不确定性[6–8]。根据机器学习建模原理可知, 由滑坡和非滑坡样本共同组成的训练/测试集作为机器学习的核心在整个建模过程中至关重要。其中滑坡样本通常根据历史滑坡编录或遥感影像及航片进行选取, 其存在的不确定性较小[9–10]。而非滑坡样本通常无法直接获取, 文献显示大多通过采集“伪”负样本来代替非滑坡样本且目前没有统一的选择方式[11–13]。因此, 非滑坡的选择相对于滑坡样本而言不确定性更大, 是影响模型训练/测试集质量的关键因素之一, 合理地选择可信度高的非滑坡样本有利于降低建模不确定性[11]。

现有研究大多在整个研究区内未发生滑坡的区域中随机选择非滑坡样本[14–15]。一般而言, 滑坡在河道、沟谷等低坡度区域内发生的概率较小, 因此, 可利用高分辨影像解译低坡度属性区并从中随机选择非滑坡[16–17]。此外, 缓冲区控制采样法也常应用于滑

收稿日期:2022 – 11 – 20

基金项目:国家自然科学基金项目(41807285; 42377164; 42272326)

作者简介:黄发明(1988—), 男, 副教授, 博士. 研究方向: 滑坡易发性预测. E-mail: faminghuang@ncu.edu.cn

* **通信作者:**姚　池, E-mail: chi.yao@ncu.edu.cn

网络出版时间:2023 – 05 – 31 15 : 08 : 00　　**网络出版地址:**https://link.cnki.net/urlid/51.1773.tb.20230529.1853.003

坡易发性预测，即从滑坡面缓冲区以外的区域选择非滑坡样本[11,13,18]。对于上述非滑坡选择方式，全区随机选择虽然避开了已知滑坡点，但是难以保证非滑坡样本的可靠性，导致样本误差较大，进而将误差传递给了易发性建模。从低坡度属性区中选择虽然改善了非滑坡的稳定性，但使采样工作被坡度因子主导。另外缓冲区外选择非滑坡对缓冲半径的确定并无统一标准，缓冲区过大或过小均会造成建模的不确定性，且采集的滑坡点不同将影响缓冲区的位置从而改变选择范围[11]。

信息量（IV）法和半监督法选择非滑坡样本的原理类似，即极低和低易发区内发生滑坡概率较小，在此范围内选择非滑坡样本的可靠性更高[9,19]。信息量法不需要非滑坡样本也能得到初始滑坡易发性分区[20]。半监督法结合了全监督和无监督的优点，在仅有少量已标记样本的情况下能够利用隐藏在大量无标签样本中的数据分布信息来提升学习性能[21]。

综上所述，由于上述各类选择方式的主观性和随机性较强，导致获取的非滑坡样本不具有足够的代表性，降低测试集的质量，从而影响后续建模性能。本文以江西省南康区为例，采用5种非滑坡样本选择方式，即全区随机选择[14]、坡度小于5°的特定属性区内随机选择[16]、滑坡面缓冲300 m外随机选择[11]、信息量法[20]以及半监督法[21]；将得到的非滑坡样本与机器学习耦合构建模型预测易发性；对比5种方式的易发性结果，进而探究不同非滑坡选择方式对建模不确定性的影响规律；比较不同机器学习的应用效果，发现随机森林(RF)所需输入参数和调整较少，且预测精度较高[8,22–23]；故最终构建随机RF、低坡度RF、缓冲区RF、IV–RF和半监督RF模型进行易发性预测。

1 滑坡易发性预测建模方法

1.1 研究思路

本文构建随机RF、低坡度RF、缓冲区RF、IV–RF和半监督RF模型开展易发性建模并对比分析。具体流程如图1所示：1）获取研究区滑坡编录信息并采集19种基础环境因子数据，利用频率比相关性分析得到各因子的值；2）基于环境因子相关数据，利用IV模型、随机RF模型进行初始易发性分区，将极低和低易发区作为选择范围获取高可靠性的非滑坡样本；3）合并获得的非滑坡样本与历史滑坡样本，耦合构建随机RF、低坡度RF、缓冲区RF、IV–RF和半监督RF模型进行易发性预测；4）对比5种选择方式耦合的RF模型的预测结果，采用ROC（receiver operation characteristic curves）曲线和易发性指数分布规律分析建模不确定性，并探讨半监督RF模型中样本非对称的情况。

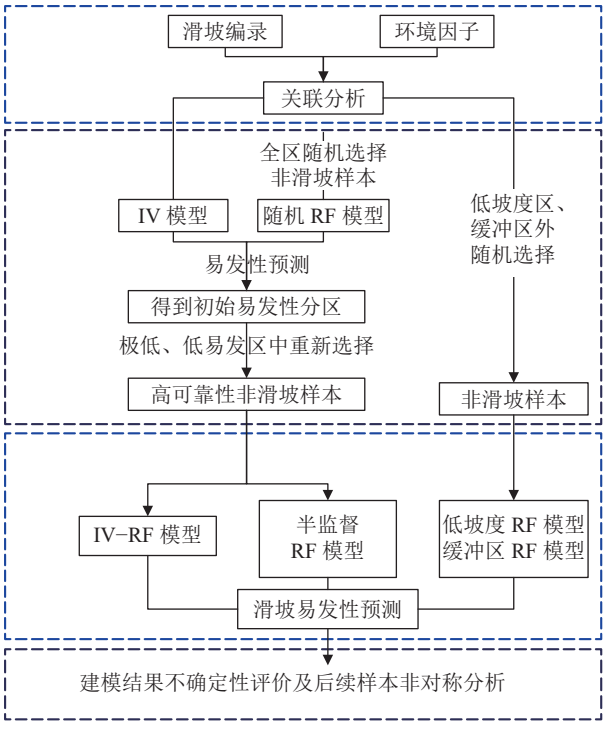


图 1 不同非滑坡样本选择方式的滑坡易发性预测建模流程

Fig. 1 Flow chart of LSP modeling under different non–landslide sample selection methods

1.2 非滑坡样本选择

全区随机选择时先剔除整个研究区内的已知滑坡区域，从得到的无滑坡区中随机选择非滑坡样本[14]。从低坡度区域选择时根据研究区的坡度特征设置合理的坡度值，筛选出坡度小于该值的区域后从中随机选择单元作为非滑坡样本[16]。结合研究区地理环境以及历史滑坡信息确定缓冲距离的大小后基于历史滑坡面数据利用ArcGIS 10.2创建滑坡缓冲区，选择整个研究区缓冲区以外的区域作为非滑坡样本的选择范围[11]。

信息量法选择非滑坡首先利用信息量模型计算获得各环境因子的信息量值，对其进行叠加后得到总信息量值[20]。在ArcGIS 10.2中运用自然断点法对全区总信息量值进行初始分区。由于信息量值越高表明发生滑坡的概率越大，在极低、低易发区内进行非滑坡样本的选择。

在全区随机选择非滑坡进行易发性建模的基础上建立半监督法。由于初始预测得到的极低、低易发区内的栅格单元发生滑坡易发性小，在此区域内进行非滑坡样本的采集更加合理，提高了非滑坡样本的可信度。

1.3 随机森林（RF）模型

RF模型基于决策树算法通过独立采样和随机选择特征变量构建多个决策树模型进行预测和分类得到综合分析结果[22]：1）从原始训练集样本中进行有放回的重复采样以获得与原始样本特征数目相同的

样本，作为决策树根节点处训练集；2）从N个特征中随机选取n个($n \ll N$)为决策树节点的分裂建立特征集并择取其中一个作为某节点的分裂属性；3）决策树上每一节点按2）中进行分裂并建立此类大量决策树形成随机森林[8]。由于建立一组决策树进行预测会产生泛化误差的限制值，使用RF模型能有效避免模型过拟合问题的出现，显著提高模型的有效性和优越性[20]。

1.4　滑坡易发性建模不确定性评价

1.4.1　基于ROC曲线的精度分析

采用ROC曲线分析易发性建模精度，能有效降低因测试集差异而产生的干扰，使模型性能评估工作更客观[24]。由式（1）～（2）计算的真阳性率（R_{TPR}）和假阳率（R_{FPR}），分别代表分类器识别滑坡的准确程度[22]：

$$R_{TPR}=\frac{n_{TP}}{n_{TP}+n_{FN}} \tag{1}$$

$$R_{FPR}=\frac{n_{FP}}{n_{FP}+n_{TN}} \tag{2}$$

式中，n_{TP}、n_{FN}分别为被正确识别为滑坡的滑坡点个数和被错误识别为非滑坡的滑坡点个数，n_{FP}、n_{TN}分别为被错误识别为滑坡的非滑坡点个数和被正确识别为非滑坡的滑坡点个数。结合ROC曲线下面积S_{AUC}对模型进行量化分析，反映出随机挑选的结果中滑坡样本排名高于非滑坡样本的概率[8]。S_{AUC}值一般在0.5～1.0范围内，越接近1.0,说明该模型的预测性能更优越。利用式（3）计算S_{AUC}：

$$S_{AUC}=\frac{\sum_{i=1}^{n_0} r_i-n_0(n_1+1)/2}{n_0 n_1} \tag{3}$$

式中，n_0、n_1分别为非滑坡与滑坡样本个数，r_i为第i个非滑坡样本在整个测试样本中的排序。

1.4.2　基于混淆矩阵的精度评价

基于混淆矩阵衍生得到的Kappa系数（K_C）和总体分类精度（O_A）是滑坡易发性模型精度评价的重要指标。Kappa系数通常用于一致性检验，而总体分类精度能够直接反映模型分类正确的比例，由式（4）～（5）计算得到：

$$O_A=\frac{n_{TP}+n_{TN}}{n_{TP}+n_{TN}+n_{FP}+n_{FN}} \tag{4}$$

$$K_C=\frac{O_A-P_e}{1-P_e} \tag{5}$$

式（5）中，K_C系数值通常在0～1.0范围内。当K_C在0.6～1.0时，则说明模型具有高可靠性；当K_C大于0.8时，说明预测结果与实际较一致[20]。P_e为期望一致率，即两次检验结果由于偶然机会所造成的一致率：

$$P_e=\frac{(n_{TP}+n_{FP})(n_{TP}+n_{FN})+(n_{TN}+n_{FN})(n_{TN}+n_{FP})}{(n_{TP}+n_{FP}+n_{TP}+n_{FN})^2} \tag{6}$$

1.4.3　滑坡易发性指数分布

滑坡易发性指数分布特征主要通过均值和标准值两个指标进行分析，二者分别反映了易发性指数分布的平均水平和离散趋势[20]。均值较小，表明极低和低易发区包含了大部分易发性指数，结合高S_{AUC}精度，进一步表明此时建模的不确定性更小；标准差大，说明整体易发性指数的分散程度高，结合高S_{AUC}精度，进一步表明滑坡易发性指数的可识别性强，且与野外滑坡实际分布情况更契合[25]。

2　研究区概况及环境因子选取

2.1　南康区概况及其滑坡编录

如图2所示，南康区地处江西省赣州市西部，属中亚热带季风湿润气候。年均降雨约1 443.2 mm，雨量充沛但分布不均，境内水资源丰富。地处山脉区间高度范围为96～995 m，呈纵长横狭之势，总面积约1 844.96 km^2。地形地貌以丘陵、山地为主，且章江、上犹江两岸分布有较平整的河谷平原。根据南康区自然资源部门所知，1970—2010年累计发生约233处滑坡。当地滑坡以中小型规模为主，大部分为牵引式滑坡，滑体主要是第四纪堆积层。滑坡空间分布较均匀，北部和南部的低山、高丘陵山区、中部低丘陵区（红层盆地）为滑坡多发区，其中地层界线交界处、道路两侧及植被分布较少的区域分布有较多滑坡。而在受人类活动影响较小的植被丰富地区滑坡数量更少，如镜坝—三江—龙华河谷阶地。为避免原始滑坡点的空间位置误差引起建模不确定性，在绘制滑坡样本的边界时将边界向外合理且尽可能准确地扩展。

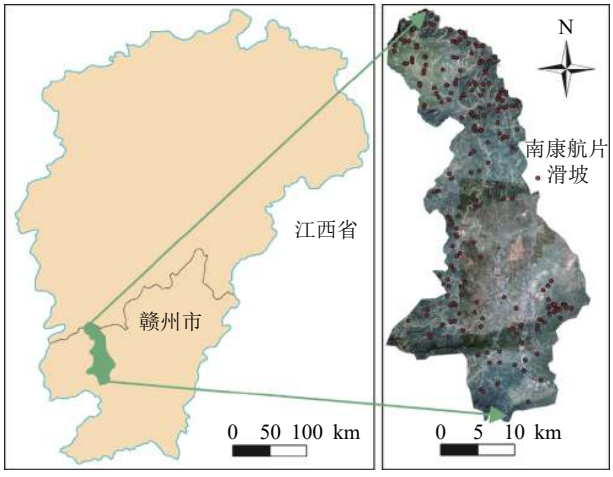


图 2　南康区概况及滑坡编录

Fig. 2　Overview of Nankang District and landslide inventory

相关文献综述表明南康区内诱导滑坡发生的主要原因是强降雨，其次人类活动也对滑坡的发生具有一定的影响[21]。

2.2 数据源

研究采用的数据源主要包括：1）南康区自然资源局历史滑坡编录资料及地质勘察报告；2）30 m分辨率的数字高程模型（DEM）用于获取地形地貌和水文环境等环境因子；3）采用 1∶10^5比例尺的地质图提取岩性因子；4）采用30 m分辨率的Landsat TM8遥感影像1景（2013.07.03，轨道号119/041）提取地表覆被因子；5）从中国科学院地理数据贡献平台获取研究区的年均降雨量、人口密度、GDP等数据用于提取相关环境因子。本文采用30 m的分辨率表达DEM和遥感影像，符合国家基础空间数据库的建库标准且能有效反映地形地貌特征，同时能够避免出现因栅格数过多造成模型计算困难的问题[26]。最终将233个滑坡编录划分成2 598个滑坡栅格单元，即易发性建模时共有2 598个滑坡样本。

2.3 环境因子频率比分析

滑坡的发生是由多种内部因素和外部因素共同作用的结果[27–28]。参考江西省内其他与南康类似的研究区相关文献资料，考虑相关环境因子的客观实在性、统计继承性等原则，从数据源中获取地形地貌、基础地质、气象水文、地表覆被4个方面的19个环境因子数据用于易发性建模[29]。选择频率比（F_R）模型处理滑坡与环境因子间的非线性响应关系来反映环境因子对滑坡易发性的影响程度[27]。当F_R>1时，说明该因子所属区间有利于滑坡孕育；当F_R<1，则说明不利于滑坡孕育。利用ArcGIS 10.2的自然断点法，将连续型的环境因子划分为8个子区间见表1[8]，具体分布如图3所示。

2.3.1 地形地貌因子

基于DEM利用ArcGIS 10.2获取其他地形地貌因子。由表1和图3（a）、（b）可知，高程在163.4～360.9 m、坡度在6.2°～21.2°内，F_R>1，表明中等程度海拔和坡度的区域是南康区滑坡的主要发生地。剖面曲率和平面曲率分别体现垂直方向、水平方向的地形复杂程度[23]。当剖面曲率介于1.3～10.4、平面曲率小于28.7时，F_R>1，易造成滑坡的发生。地形起伏度从宏观角度反映研究区的地貌特征，其在20.0～100.1 m内滑坡发生的概率较大。

2.3.2 基础地质因子

岩性通过直接影响基岩和堆积体的力学性质来干扰滑坡的孕育[16,26]。地质调查显示，研究区内主要出露有变质岩、碳酸盐岩、碎屑岩，其中变质岩和碎屑岩的F_R均>1。滑坡密度表示一定范围内所包含的滑坡点数量[27]，滑坡密度大的区域是滑坡的高易发区。斜坡形态包括凹形坡、直线形坡、复合形坡和凸形坡4类。由表1可知，研究区内凹形坡和复合形坡有利于滑坡的发育。土壤黏/砂粒含量与水的渗透、侵蚀联系密切[30]。当土壤表层结构中黏粒比砂粒含量更低，底层黏粒比砂粒含量更高时土壤中水分的渗透作用增强，加重了斜坡体重量，从而促进滑坡面的形成[31]。

2.3.3 气象水文因子

边坡受降雨冲刷易发生软化，且雨水下渗会改变坡体内部的力学性质[32]。由表1可知，降雨量高的区域滑坡发生的概率也更高。采用沟壑密度和改进的归一化差异水体指数（MNDWI）因子反映水文环境对滑坡的影响。沟壑密度定义为单位面积内沟壑河道的长度之和[33]。沟壑越密集的区域，受降雨、水系的侵蚀作用更加严重，滑坡发生的概率更高。MNDWI则能有效突显影像中的水体信息，揭示水体微细特征。

2.3.4 地表覆被因子

归一化建筑指数（NDBI）能有效表示出研究区内建筑用地的信息，当NDBI在0.56～0.75范围内时有利于滑坡发生[25]。归一化植被指数（NDVI）反映区域内植被生长情况和覆盖程度，覆盖度高的区域通常滑坡发生的可能性较小[21]。总辐射包括水平地表所接受太阳的直接辐射与漫射辐射，通过影响植被生长和土壤湿度间接作用于滑坡的发生[8]。人口密度和GDP密度分别表示单位面积内人口数量和经济的分布特征。公路密度体现了研究区内公路修建的密集程度，道路修建过程中的开挖切坡行为会改变坡体的自然结构，破坏边坡坡脚的稳定从而促进滑坡的发生[16]。

3 不同非滑坡选择下的易发性结果

全区随机选择的方式从无滑坡区内随机选择与滑坡样本等量的2 598个非滑坡样本。根据南康区历史滑坡的地理特征和相关文献[16–17]，认为坡度小于5°的属性区发生滑坡的概率较小，故筛选研究区坡度低于5°的栅格从中选择2 598个单元作为非滑坡。从缓冲区外选择时基于2 233个历史滑坡面创建距离大小为300 m的滑坡缓冲区，在该范围以外随机选择2 598个非滑坡样本。

信息量法加权处理环境因子信息量值后，得到总信息量值的范围为–21.15～8.83。分区后在极低、低易发区中随机选择2 598个栅格单元作为非滑坡样本。

采用半监督法选择时从随机RF模型易发性分区的极低、低易发区中随机选择2 598个栅格作为可靠性更高的非滑坡样本。另外在进行样本非对称分析时，为构建滑坡与非滑坡比例为1∶2的样本集，将非滑坡样本个数增加至5 196个栅格。

表 1　部分基础环境因子 F_R 值

Tab. 1　Frequency ratios of some environmental factors

基础环境因子	变量值	类型	全区栅格数/个	栅格比例/%	滑坡内栅格数/个	坡内栅格比例 /%	F_R值
坡度/(°)	0～2.9	连续型	570 276	27.64	35	1.35	0.05
	2.9～6.2		465 255	22.55	369	14.20	0.63
	6.2～9.7		342 869	16.62	639	24.60	1.48
	9.7～13.2		276 438	13.40	731	28.14	2.10
	13.2～16.9		200 695	9.73	525	20.21	2.08
	16.9～21.2		125 840	6.10	248	9.55	1.56
	21.2～26.9		62 731	3.04	50	1.92	0.63
	26.9～47.0		18 898	0.92	1	0.04	0.04
沟壑密度	0～0.2	连续型	105 908	5.13	55	2.12	0.41
	0.2～0.4		158 398	7.68	230	8.85	1.15
	0.4～0.5		214 485	10.40	360	13.86	1.33
	0.5～0.6		257 696	12.49	362	13.93	1.12
	0.6～0.7		361 880	17.54	563	21.67	1.24
	0.7～0.8		361 004	17.50	508	19.55	1.12
	0.8～0.9		379 524	18.40	325	12.51	0.68
	0.9～1.1		224 107	10.86	195	7.51	0.69
地层岩性	变质岩	离散型	815 922	39.55	1 453	55.93	1.41
	碳酸盐岩		688 947	33.40	336	12.93	0.39
	碎屑岩		546 081	26.47	809	31.14	1.18
	水域		12 052	0.58	0	0	0
NDVI	0～0.26	连续型	11 978	0.58	0	0	0
	0.26～0.40		68 643	3.33	35	1.35	0.40
	0.40～0.49		144 291	6.99	147	5.66	0.81
	0.49～0.56		261 494	12.68	325	12.51	0.99
	0.56～0.63		454 414	22.03	664	25.56	1.16
	0.63～0.69		528 457	25.62	755	29.06	1.13
	0.69～0.75		430 066	20.85	595	22.90	1.10
	0.75～1.00		163 658	7.93	77	2.96	0.37
年均降雨量/mm	1 244～1 260.7	连续型	988 032	47.89	767	29.52	0.62
	1 260.7～1 288.0		268 563	13.02	252	9.70	0.75
	1 288.0～1 322.9		222 824	10.80	159	6.12	0.57
	1 322.9～1 356.4		142 997	6.93	200	7.70	1.11
	1 356.4～1 385.2		88 902	4.31	310	11.93	2.77
	1 385.2～1 414.1		130 065	6.30	323	12.43	1.97
	1 414.1～1 446.0		120 523	5.84	274	10.55	1.81
	1 446.0～1 484.0		101 096	4.90	313	12.05	2.46
土壤黏/砂含量	15.2～20.8	连续型	40 258	1.95	31	1.19	0.61
	20.8～22.4		191 256	9.27	105	4.04	0.44
	22.4～23.5		200 373	9.71	130	5.00	0.52
	23.5～25.2		718 403	34.82	884	34.03	0.98
	25.2～27.2		339 452	16.45	381	14.67	0.89
	27.2～29.4		230 867	11.19	400	15.40	1.38
	29.4～31.4		281 905	13.66	560	21.56	1.58
	31.4～35.0		60 488	2.93	107	4.12	1.40
公路密度/(km·km^{-2})	0～0.9	连续型	370 601	17.96	215	8.28	0.46
	0.9～1.8		535 714	25.97	1 150	44.26	1.70
	1.8～2.8		414 650	20.10	752	28.95	1.44
	2.8～3.7		299 942	14.54	238	9.16	0.63
	3.7～4.6		254 921	12.36	187	7.20	0.58
	4.6～5.8		131 311	6.37	42	1.62	0.25
	5.8～7.9		40 545	1.97	14	0.54	0.27
	7.9～11.9		15 318	0.74	0	0	0

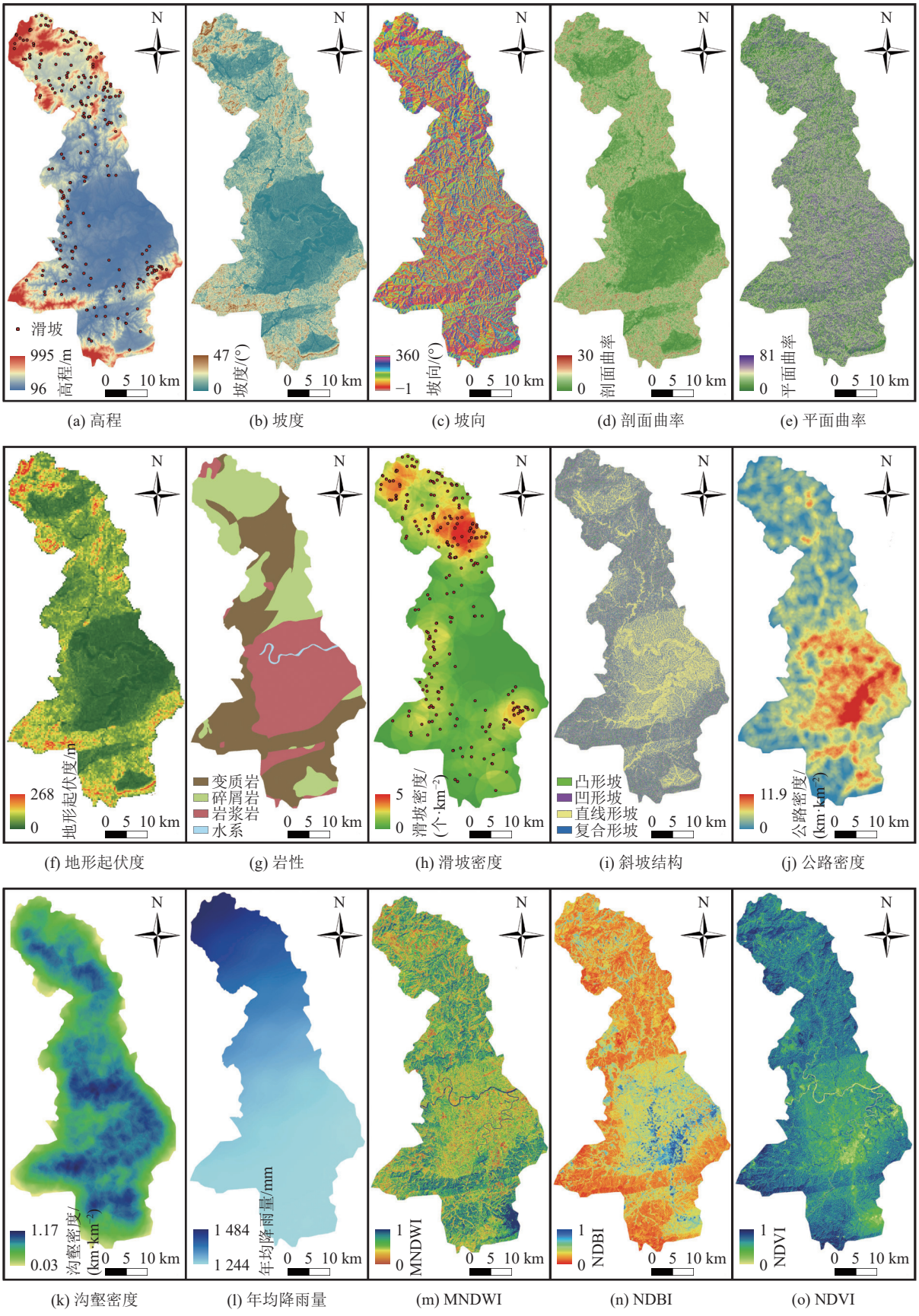


图 3　滑坡基础环境因子

Fig. 3　Basic environmental factors of landslide

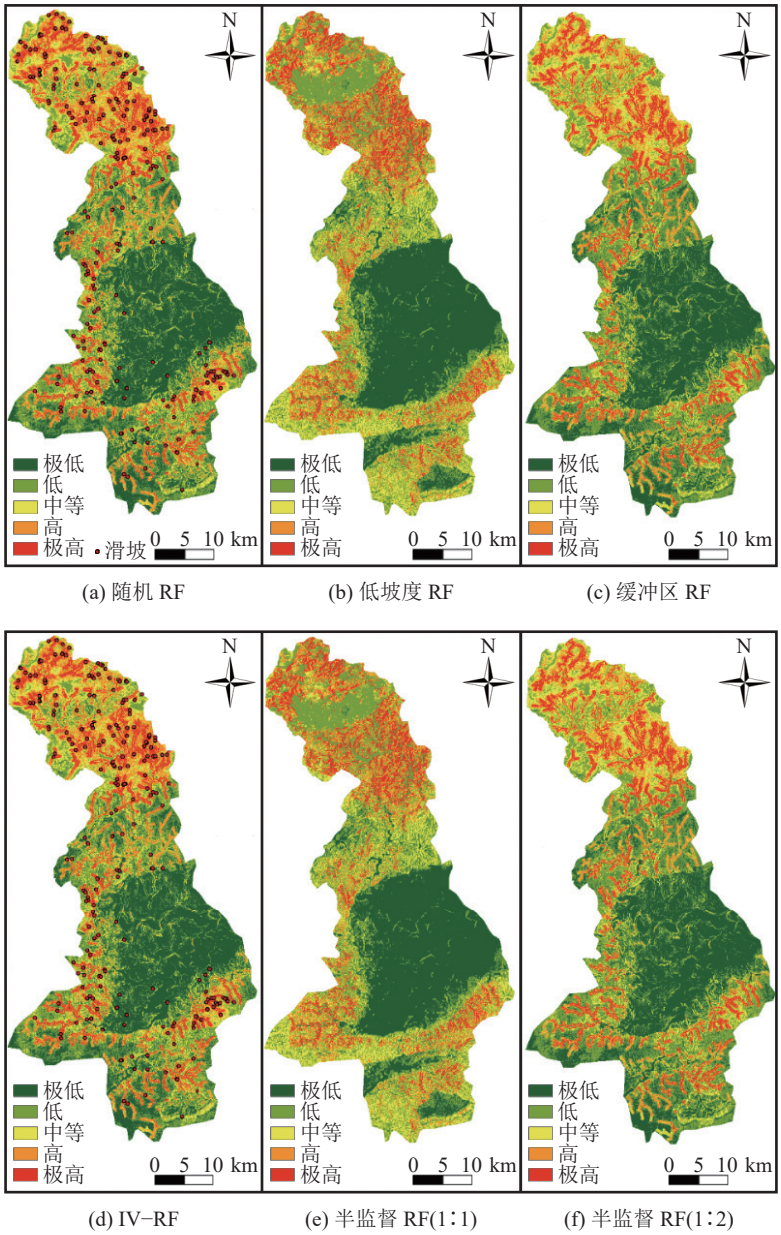
3.1　RF的易发性建模

整个研究区采用30 m的分辨率划分为2 063 002个栅格单元，将所有环境因子频率比分析后重新赋值的结果作为RF模型的输入变量。利用ArcGIS 10.2转换得到的2 598个滑坡栅格单元易发性并将其赋值为1，各方式选择的非滑坡样本易发性赋值为0，二者共同组成预测模型的输出变量[8]。联接滑坡与非滑坡样本及环境因子的FR值后合并构成等比例样本集作为训练/测试集，按7∶3随机划分为两部分，其中70%用于模型训练，30%用于模型测试[21]。

利用Python 3.8.8的Pandas、NumPy、Scipy等库对数据进行读取、计算和预处理，以及Scikit–learn库实现RF模型的机器学习过程[34]。由于构建RF模型时决策树的数量将对模型的整体精度造成影响，根据重复实验验证得最优的RF决策树数目并在模型中应用该参数进行预测[20]。

3.2　滑坡易发性预测结果

基于5种选择方式得到的非滑坡样本，利用Python 3.8.8对整个南康区栅格单元进行易发性预测，将易发性指数导入ArcGIS 10.2中制图。为方便对比不同选择方式的预测结果，结合易发性指数分布规律和自然间断点法将预测的易发性指数均按10%、10%、20%、30%和30%的比例划分为极低、低、中等、高和极高5个级别[12]。不同非滑坡样本选择方式的建模结果如图4和表2。



(a) 随机 RF　(b) 低坡度 RF　(c) 缓冲区 RF

(d) IV–RF　(e) 半监督 RF(1∶1)　(f) 半监督 RF(1∶2)

图 4　基于不同非滑坡样本选取方法的滑坡易发性

Fig. 4　Landslide susceptibility maps under different non–landslide selection methods

表 2 基于5种非滑坡样本选择方式的易发性评价等级的统计结果

Tab. 2 Statistical results of susceptibility classification based on five non–landslide sample selection methods

模型	易发性级别	随机RF	低坡度RF	缓冲区RF	IV–RF	半监督RF(1∶1)	半监督RF(1∶2)
全区栅格数	极低	616 910	620 509	621 957	619 707	597 233	1 314 580
	低	518 089	512 803	510 232	623 012	522 179	185 297
	中	411 937	410 366	417 291	409 461	452 337	104 720
	高	309 247	314 817	307 034	207 269	304 671	308 872
	极高	206 819	204 507	206 488	203 553	186 582	149 533
全区栅格百分比/%	极低	29.90	30.08	30.15	30.04	28.95	63.72
	低	25.11	24.86	24.73	30.20	25.31	8.98
	中	19.97	19.89	20.23	19.85	21.93	5.08
	高	14.99	15.26	14.88	10.05	14.77	14.97
	极高	10.03	9.91	10.01	9.87	9.04	7.25
滑坡内栅格数	极低	20	75	21	108	4	141
	低	57	379	52	349	30	207
	中	239	385	225	734	224	239
	高	614	480	660	351	651	1 118
	极高	1 668	1 279	1 640	1 056	1 689	893
坡内栅格百分比/%	极低	0.77	2.89	0.81	4.16	0.15	5.43
	低	2.19	14.59	2.00	13.43	1.15	7.97
	中	9.20	14.82	8.66	28.25	8.62	9.20
	高	23.63	18.48	25.40	13.51	25.06	43.03
	极高	64.20	49.23	63.13	40.65	65.01	34.37
F_R值	极低	0.03	0.10	0.03	0.14	0.01	0.09
	低	0.09	0.59	0.08	0.44	0.05	0.89
	中	0.46	0.74	0.43	1.42	0.39	1.81
	高	1.58	1.21	1.71	1.34	1.70	2.87
	极高	6.40	4.97	6.31	4.12	7.19	4.74

由图4和表2可知，随着滑坡易发性级别的提高，其对应的FR值也逐渐增大，各方式预测的极高和高易发区内均包含了大部分的滑坡栅格单元。由此可见，5种选择方式预测的易发性图整体上相似，但对比细节发现预测结果间仍存在差异。使用等比例样本集（滑坡∶非滑坡=1∶1）的情况下，随机RF、低坡度RF、缓冲区RF、IV–RF和半监督RF模型中极高和高易发区的历史滑坡占比分别为87.83%、67.71%、88.53%、54.16%和90.07%。其中，半监督RF预测的极高和高易发区中包含的历史滑坡数量最多，表明半监督RF的易发性结果与已知滑坡的分布特征更加吻合，具有更优的建模性能。

4 滑坡易发性预测结果不确定性分析

4.1 模型精度评价

4.1.1 ROC精度评价

上述6种模型的ROC曲线及S_{AUC}如图5所示。由图5可见：随机RF和缓冲区RF模型的ROC曲线相近且凸出程度较低，S_{AUC}分别为0.895、0.896；低坡度RF模型S_{AUC}高达0.973，可见，在坡度<5°的区域内，选择非滑坡样本显著提升了模型精度；半监督RF和IV–RF模型的ROC曲线均高于上述3种方式的ROC曲线，其中IV–RF的S_{AUC}为0.990，滑坡与非滑坡的比例分别为1∶1和1∶2时，半监督RF的S_{AUC}高达0.997和0.999，可见半监督RF模型的性能更佳，同时其在滑坡∶非滑坡=1∶2的比例条件下预测精度最高。

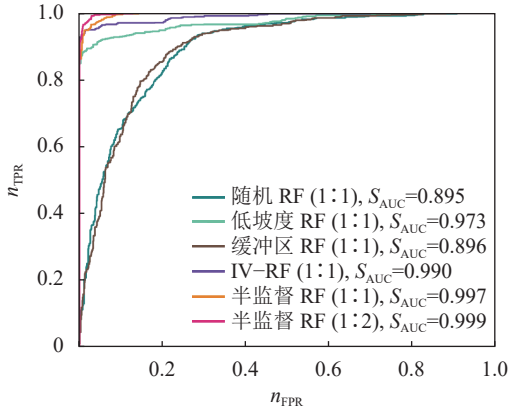


图 5 不同非滑坡样本选择方式模型的ROC曲线

Fig. 5 ROC curves of different non–landslide sample selection models

4.1.2　精度统计指标

各耦合模型的精度统计指标如表3所示。由表3可知：K_C和O_A的大小排序均为随机RF<缓冲区RF<低坡度RF<IV–RF<半监督RF(1∶1)<半监督RF(1∶2)。结果表明，相比于随机RF和缓冲区RF，其他4种耦合模型均具有更高的预测精度。其中，半监督RF(1∶1)模型的K_C为94.9%，高于低坡度RF和IV–RF模型，进一步证明了半监督RF模型具有更高的可靠性。同时，半监督RF(1∶2)模型的K_C达到95.6%，且O_A为98.0%，相比于半监督RF(1∶1)模型精度进一步提升。

4.2　不同非滑坡样本选择方式的易发性指数分布

将易发性指数在[0,1]范围内均分为100个区间进行统计分析，如图6所示。由图6可见，均值越小且标准差越大，说明建模过程中存在的不确定性越小。

表 3　不同耦合模型验证指标

Tab. 3　Validation indicators of coupled different models

模型	K_C/%	O_A/%
随机RF	68.2	84.1
低坡度RF	90.6	95.3
缓冲区RF	71.5	85.7
IV–RF	94.5	97.3
半监督RF(1∶1)	94.9	97.5
半监督RF(1∶2)	95.6	98.0

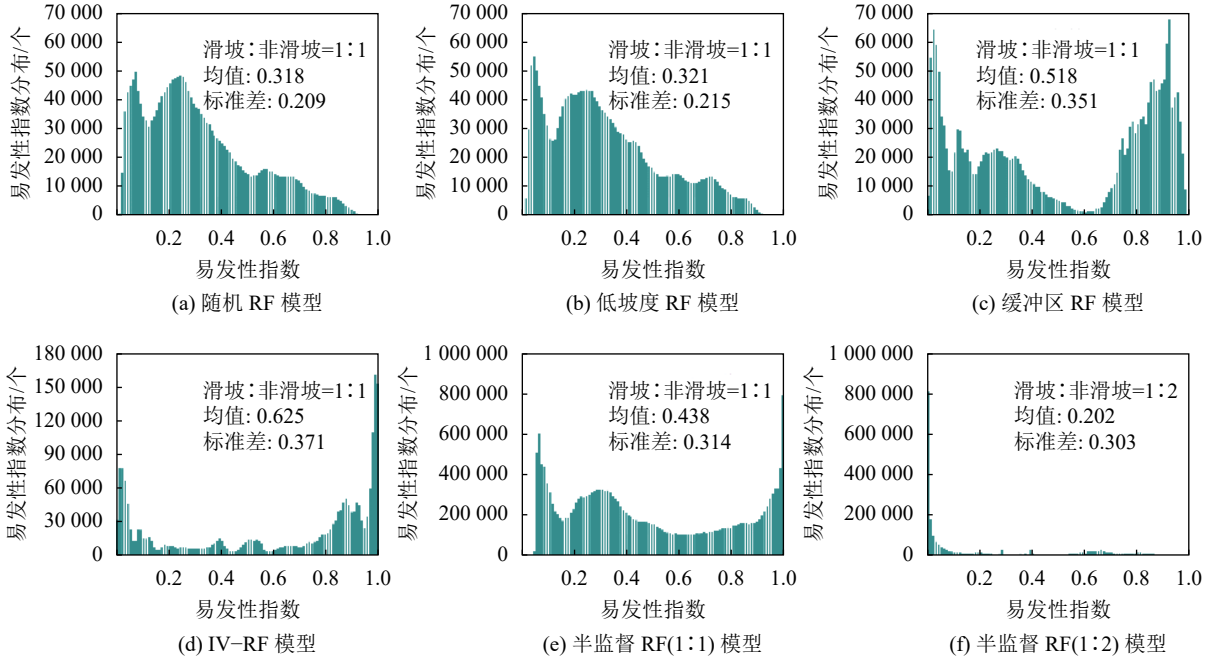


图 6　不同非滑坡样本选择方式对应的易发性指数分布

Fig. 6　Susceptibility indexes distribution of different non–landslide sample selection methods

1)随机RF与缓冲区RF模型的易发性指数分布规律较一致，整体上呈现随易发性指数增大而分布逐渐减小的趋势，且在极低易发区内均出现一段小幅增加。低坡度RF、IV–RF和半监督RF模型(滑坡∶非滑坡=1∶1)的易发性指数分布表现为低易发区和高易发区的分布较集中，而中间易发区分布较少。

2)各耦合模型的易发性指数按均值大小排序为：均值(IV–RF) > 均值(低坡度) > 均值(半监督1∶1) > 均值(缓冲区) > 均值(随机) > 均值(半监督1∶2)。按标准差大小排序为：标准差(IV–RF) >标准差(低坡度) >标准差(半监督1∶1) > 标准差(半监督1∶2) > 标准差(缓冲区) > 标准差(随机)。滑坡∶非滑坡=1∶1时各耦合RF模型的均值普遍较大。其中，随机RF和缓冲区RF的均值较小，分别为0.318和0.321；但二者的标准差也较小，分别为0.209和0.215，说明利用这两种方法进行预测时对易发性的区分度不高。低坡度RF和IV–RF的均值为0.518和0.625，标准差分别为0.351和0.371，相对于其他非滑坡选择方式，其均值和标准差均较大。半监督RF的均值为0.438，小于低坡度RF和IV–RF；且标准差为0.314，大于随机RF和缓冲区RF，综合而言半监督RF模型的预测性能更优。在滑坡∶非滑坡=1∶2时，半监督RF模型的均值为0.202，在6种模型中其均值最小，且易发性指数大部分位于极低和低易发区内，说明利用少量的高易发性指数能反映出尽量多的历史滑坡编录信息[8]。

4.3　滑坡环境因子重要性分析

环境因子的重要性是评估各因子对滑坡发生的影响程度的指标之一。将基础环境因子中具有较高

重要性的因子称为滑坡易发性主控因子。在滑坡易发性预测过程中分析基础环境因子的重要性程度对易发性研究起到参考作用。本文利用Python 3.8.8对6种耦合模型中的19种环境因子进行分析，获得相应的重要性后，通过origin 2018软件处理得到各基础环境因子的重要性排序，如图7所示。

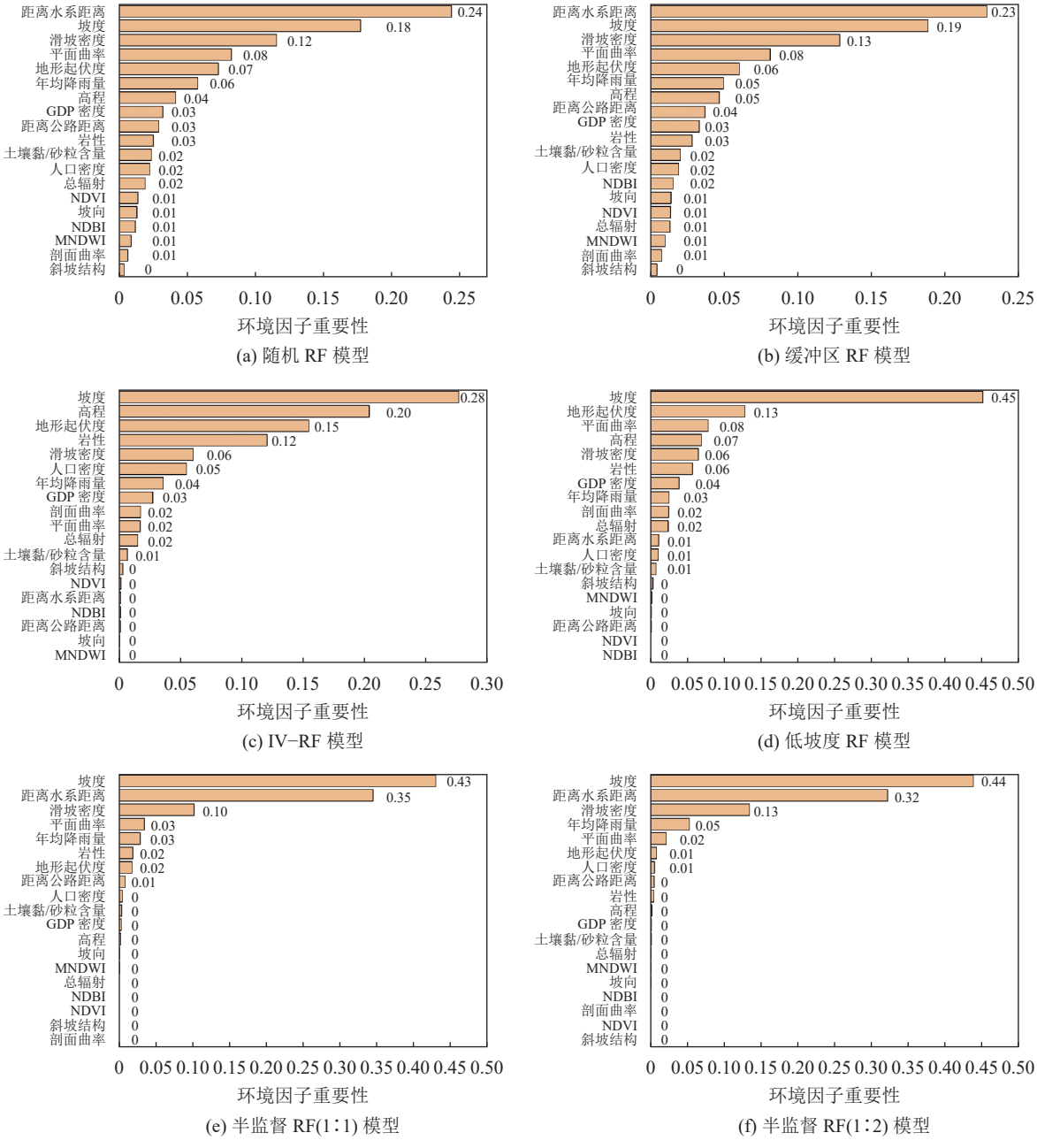


图 7　不同非滑坡样本选择方式对应的环境因子重要性排名

Fig. 7　Environmental factor importance ranking of different non–landslide sample selection methods

综合对比图7(a)～(f)可知，各耦合模型的环境因子重要性程度大同小异。随机RF和缓冲区RF中的环境因子重要性排序大致相同，两种模型中距离水系距离、坡度和滑坡密度等因子的重要性程度均较大。低坡度RF和IV–RF在因子重要性方面表现相似，坡度、地形起伏度、高程、岩性和滑坡密度等因子的重要性程度较大。而低坡度RF模型中坡度因子重要性占比更大，间接反映了非滑坡样本根据坡度特征进行选择。半监督RF模型中坡度因子对模型的影响程度最为显著，其次是距离水系距离、坡度、滑坡密度和年均降雨量等因子。

总体而言，坡度是南康区滑坡发生最重要的主控因子之一，在所有耦合模型中其因子重要性均较大。而MNDWI、NDBI和NDVI等环境因子对南康区滑坡事件发生的影响较小。结合南康区的滑坡成因以及地理环境条件可知半监督RF模型计算的环境因

子重要性更加契合实际，可信度更高。

5 讨　论

5.1 滑坡易发性分区结果

上述5种不同的非滑坡样本选择方式预测的滑坡易发性结果整体上类似。由于研究区内中等程度高程与坡度的地形地貌有利于边坡堆积层的形成从而促进滑坡发育，由表1可知，中等程度高程和坡度地区的F_R均大于1。结合图3(a)～(b)和图4观察发现，滑坡的极高和高易发区常分布于此类地区。降雨量丰富且沟壑密度大的区域内，地表/地下水的渗流现象明显，滑带湿润情况严重会导致抗滑力显著下降[22]，年均降雨量超过1 322.9 mm、沟壑密度在0.2～0.8的区域包含较多的极高和高易发区。当地的碎屑岩及变质岩原本的结构应力平衡易被破坏使坡体的力学强度下降[8]，碎屑岩和变质岩对应的频率比分别为1.18和1.41，对比图3(g)和图4可知，这两种地层岩性区域滑坡易发性较高。观察NDVI环境因子可知，上述区域通常植被覆盖程度不高，间接加速了岩体的风化作用。南康区的滑坡极低和低易发区普遍海拔较高，且降雨作用不明显，水流下渗对斜坡体形成的影响较小，同时植被覆盖程度较高，有效降低了滑坡发生概率[21]。

5.2 不同非滑坡样本选择方式下的易发性建模

随机RF模型选择非滑坡样本减少人为的干扰，其预测精度整体效果尚可且操作简便，故目前大多数研究采用随机方式选择非滑坡[15]。但由于该方法未考虑到无历史滑坡区域存在高易发性样本的可能，将影响非滑坡样本的质量和可靠性[35]。

从低坡度属性区选择非滑坡样本一定程度上提高了建模精度，目前已有部分研究应用此方式开展，例如Kavzoglu等[7]选择在坡度小于5°的地区进行采样。利用此方法时需要结合研究区地形确定合适的坡度范围以得到更可靠的非滑坡样本。对比图3(b)和图4发现，该模型预测结果中低坡度区内易发性指数均较低，而其他区域的易发性指数普遍较大，表明利用此方法选取得到的非滑坡样本进行预测一定程度的降低了模型地泛化能力，难以合理识别全区滑坡易发性。实际情况中低坡度地区与低易发区不能完全划等号，可见在低坡度区内选择非滑坡样本存在显而易见的缺陷[36]。

缓冲区RF模型能降低非滑坡样本的错误率，本研究对比不同缓冲距离的预测结果后，选择以300 m作为缓冲距离展开研究。研究中最佳缓冲距离的确定与所选研究区的环境特征、数据源等有关，需要反复实验才能更好地确定[18]。不同研究区之间对缓冲距离的选择可能存在较大差异，如鲍帅[13]、Lucchese[18]等选择以1 km作为缓冲距离，而缪亚敏等[11]将200～500 m范围作为最佳缓冲距离。然而由于人为局限了非滑坡样本的空间范围，易使非滑坡样本分布不够均衡。

针对上述非滑坡样本选择方式中存在的问题，现有研究中还存在基于自组织映射神经网络[12]、DBSCAN[13]的聚类分析法、目标空间外向化采样法[35]等均能获得更加可靠的非滑坡样本。本文进一步分析信息量法和半监督法以探索更高效准确地选择非滑坡样本。构建IV–RF模型进行预测时对易发性结果的可识别性效果不佳。结合精度和易发性指数分布规律可知，半监督法相比于其他4种方式具有更高的预测精度且更强的易发性指数分布规律性。

事实上，研究各类模型均是基于RF算法，差别在于不同选择方式得到的非滑坡样本可靠程度不同，从而影响滑坡易发性预测建模过程中训练测试集的质量。半监督模型的优势在于减小了模型训练和测试过程中由于非滑坡样本的质量产生的误差。半监督机器学习根据初次易发性预测结果，从极低和低易发区进行更加准确的采样工作，一定程度上提高了训练测试集的质量，从而提高了模型预测的精度且降低了建模的不确定性。

5.3 半监督机器学习中滑坡与非滑坡不同比例的预测建模

虽然各种方式构建等比例的滑坡–非滑坡样本开展易发性建模的预测精度均较高，但图6(a)～(e)显示其普遍存在易发性指数均值较大等问题。为进一步避免滑坡易发性指数分布不合理等问题，考虑采用不同的滑坡/非滑坡比例以尝试解决该问题。由于滑坡区域占少数，而非滑坡区域占多数，通过扩大非滑坡比例使模型更贴近研究区内真实的滑坡与非滑坡的数量关系[37]。对比5种非滑坡选择方式可知，半监督法的效果最好，故本文利用更具代表性的半监督法构建滑坡∶非滑坡=1∶2的样本集进行易发性建模，并与等比例样本集下的半监督RF模型对比。

由图5可知，相较于等比例样本集，利用滑坡∶非滑坡=1∶2比例的样本集进行易发性建模的不确定性最低。其表现为1∶2比例下的结果中易发性指数主要分布在低和极低易发区且均值显著降低，使滑坡易发性指数分布更合理。当然，本文仅讨论了利用滑坡∶非滑坡=1∶2时构建半监督RF模型的情况，滑坡与非滑坡比例的问题有待研究，比如滑坡∶非滑坡=1∶3、1∶4、1∶5、1∶6等比例下的建模情况仍需进一步探索。

6 结　论

1）利用低坡度、缓冲区、信息量法、半监督法等

方式选择非滑坡样本进行滑坡易发性预测建模时，构建的耦合RF模型具有比随机RF模型更高的预测精度。可见利用其他方式选择更可靠的非滑坡样本对提升易发性建模性能具有显著作用，准确的非滑坡样本有利于降低建模不确定性。

2）5种非滑坡选择方式耦合模型中半监督RF建模结果的精度高于IV–RF模型，其次是低坡度RF、缓冲区RF、随机RF模型。半监督RF模型结果中的均值和标准差分别为0.438和0.314，均值相对较小且标准差较大，其不确定性较小。半监督RF模型计算得到的滑坡环境因子重要性结果更贴合实际，半监督RF模型的滑坡易发性预测性能更优。

3）对比滑坡与非滑坡不同比例的工况显示，滑坡∶非滑坡=1∶2的半监督RF模型预测得到的滑坡易发性的均值显著减小到0.202，且获得的预测精度和Kappa系数最高，分别达到0.999和95.6%。由此可见，采用滑坡∶非滑坡=1∶2的比例建模能获得更准确可靠的滑坡易发性指数分布规律。

参考文献：

[1] Nath S K,Sengupta A,Srivastava A.Remote sensing GIS-based landslide susceptibility & risk modeling in Darjeeling–Sikkim Himalaya together with FEM-based slope stability analysis of the terrain[J].Natural Hazards,2021,108(3):3271–3304.

[2] Wang Yi,Fang Zhice,Niu Ruiqing,et al.Landslide susceptibility analysis based on deep learning[J].Journal of Geo-Information Science,2021,23(12):2244–2260.[王毅,方志策,牛瑞卿,等.基于深度学习的滑坡灾害易发性分析[J].地球信息科学学报,2021,23(12):2244–2260.]

[3] Chen Wei,Li Yang.GIS-based evaluation of landslide susceptibility using hybrid computational intelligence models[J].Catena,2020,195:104777.

[4] Napoli M,Carotenuto F,Cevasco A,et al.Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability[J].Landslides,2020,17(8):1897–1914.

[5] Zhao Yu,Wang Rui,Jiang Yuanjun,et al.GIS-based logistic regression for rainfall-induced landslide susceptibility mapping under different grid sizes in Yueqing,Southeastern China[J].Engineering Geology,2019,259:105147.

[6] Kim H G,Lee Dong kun,Park C,et al.Estimating landslide susceptibility areas considering the uncertainty inherent in modeling methods[J].Stochastic Environmental Research and Risk Assessment,2018,32(11):2987–3019.

[7] Qin Chengzhi,Bao Lili,Zhu Axing,et al.Uncertainty due to DEM error in landslide susceptibility mapping[J].International Journal of Geographical Information Science,2013,27(7):1364–1380.

[8] Li Wenbin,Fan Xuanmei,Huang Faming,et al.Uncertainties of landslide susceptibility modeling under different environmental factor connections and prediction models[J].Earth Science,2021,46(10):3777–3795.[李文彬,范宣梅,黄发明,等.不同环境因子联接和预测模型的滑坡易发性建模不确定性[J].地球科学,2021,46(10):3777–3795.]

[9] Liang Zhu.Comprehensive Application and Study of Machine Learning in Susceptibility Evaluation of Shallow Landslides[D].Jilin:Jilin University,2021.[梁柱.机器学习在浅层滑坡敏感性评价中的综合应用与研究[D].吉林:吉林大学,2021.]

[10] Zhao Pengxiang,Masoumi Z,Kalantari M,et al.A GIS-based landslide susceptibility mapping and variable importance analysis using artificial intelligent training-based methods[J].Remote Sensing,2022,14(1):211.

[11] Miao Yamin,Zhu Axing,Yang Lin,et al.Sensitivity of BCS for sampling landslide absence data in landslide susceptibility assessment[J].Journal of Mountain Science,2016,34(4):432–441.[缪亚敏,朱阿兴,杨琳,等.滑坡危险度评价对BCS负样本采样的敏感性[J].山地学报,2016,34(4):432–441.]

[12] Huang Faming,Yin Kunlong,Jiang Shuihua,et al.Landslide susceptibility assessment based on clustering analysis and support vector machine[J].Chinese Journal of Rock Mechanics and Engineering,2018,37(1):156–167.[黄发明,殷坤龙,蒋水华,等.基于聚类分析和支持向量机的滑坡易发性评价[J].岩石力学与工程学报,2018,37(1):156–167.]

[13] Bao Shuai,Liu Jiping,Wang Liang.Landslide susceptibility evaluation based on combined DBSCAN cluster sampling and SVM classification[J].Technology for Earthquake Disaster Prevention,2021,16(4):625–636.[鲍帅,刘纪平,王亮.联合DBSCAN聚类采样和SVM分类的滑坡易发性评价[J].震灾防御技术,2021,16(4):625–636.]

[14] Imtiaz I,Umar M,Latif M,et al.Landslide susceptibility mapping:Improvements in variable weights estimation through machine learning algorithms—a case study of upper Indus River Basin,Pakistan[J].Environmental Earth Sciences,2022,81(4):112.

[15] Chen Wenwu,Zhang Shuai.GIS-based comparative study of Bayes network,Hoeffding tree and logistic model tree for landslide susceptibility modeling[J].Catena,2021,203:105344.

[16] Kavzoglu T,Sahin E K,Colkesen I.Landslide susceptibility mapping using GIS-based multi-criteria decision analysis,support vector machines,and logistic regression[J].Landslides,2014,11(3):425–439.

[17] Choi J,Oh H J,Won J S,et al.Validation of an artificial neural network model for landslide susceptibility mapping[J].Environmental Earth Sciences,2010,60(3):473–483.

[18] Lucchese L V,de Oliveira G G,Pedrollo O C.Investigation

of the influence of nonoccurrence sampling on landslide susceptibility assessment using Artificial Neural Networks[J]. Catena,2021,198:105067.

[19] Zhou Ping,Deng Hui,Zhang Wenjiang,et al.Landslide susceptibility evaluation based on information value model and machine learning method:A case study of Lixian County, Sichuan Province[J].Scientia Geographica Sinica,2022,42(9): 1665–1675.[周萍,邓辉,张文江,等.基于信息量模型和机器学习方法的滑坡易发性评价研究——以四川理县为例[J]. 地理科学,2022,42(9):1665–1675.]

[20] Zhou Xiaoting,Huang Faming,Wu Weicheng,et al.Regional landslide susceptibility prediction based on negative sample selected by coupling information value method[J].Advanced Engineering Sciences,2022,54(3):25–35.[周晓亭,黄发明,吴伟成,等.基于耦合信息量法选择负样本的区域滑坡易发性预测[J].工程科学与技术,2022,54(3):25–35.]

[21] Huang Faming,Pan Lihan,Yao Chi,et al.Landslide susceptibility prediction modelling based on semi-supervised machine learning[J].Journal of Zhejiang University (Engineering Science),2021,55(9):1705–1713.[黄发明,潘李含,姚池,等.基于半监督机器学习的滑坡易发性预测建模[J].浙江大学学报(工学版),2021,55(9):1705–1713.]

[22] Kainthura P,Sharma N.Machine learning driven landslide susceptibility prediction for the Uttarkashi region of Uttarakhand in India[J].Georisk:Assessment and Management of Risk for Engineered Systems and Geohazards,2022,16(3): 570–583.

[23] Pourghasemi H R,Sadhasivam N,Amiri M,et al.Landslide susceptibility assessment and mapping using state-of-the art machine learning techniques[J].Natural Hazards,2021,108(1): 1291–1316.

[24] Cantarino I,Carrion M A,Goerlich F,et al.A ROC analysis-based classification method for landslide susceptibility maps [J].Landslides,2019,16(2):265–282.

[25] Huang Faming,Tao Siyu,Li Deying,et al.Landslide susceptibility prediction considering neighborhood characteristics of landslide spatial datasets and hydrological slope units using remote sensing and GIS technologies[J].Remote Sensing,2022,14(18):4436.

[26] He Sanwei,Pan Peng,Dai Lan,et al.Application of kernel-based Fisher discriminant analysis to map landslide susceptibility in the Qinggan River delta,Three Gorges,China[J]. Geomorphology,2012,171/172:30–41.

[27] Li Wenyan,Wang Xile.Application and comparison of frequency ratio and information value model for evaluating landslide susceptibility of loess gully region[J].Journal of Natural Disasters,2020,29(4):213–220.[李文彦,王喜乐.频率比与信息量模型在黄土沟壑区滑坡易发性评价中的应用与比较[J].自然灾害学报,2020,29(4):213–220.]

[28] Aditian A,Kubota T,Shinohara Y.Comparison of GIS-based landslide susceptibility models using frequency ratio,logistic regression,and artificial neural network in a tertiary region of Ambon,Indonesia[J].Geomorphology,2018,318: 101–111.

[29] Huang Faming,Hu Songyan,Yan Xueya,et al.Landslide susceptibility prediction and identification of its main environmental factors based on machine learning models[J].Bulletin of Geological Science and Technology,2022(2):79–90. [黄发明,胡松雁,闫学涯,等.基于机器学习的滑坡易发性预测建模及其主控因子识别[J].地质科技通报,2022(2): 79–90.]

[30] Nhu V H,Hoang N D,Nguyen H,et al.Effectiveness assessment of Keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area[J].Catena,2020,188:104458.

[31] Bhandari B P,Dhakal S.Compositional analysis and phase relations of soil mass from the active landslides of Babai River watershed,Siwalik zone of Nepal[J].Engineering Geology,2020,278:105851.

[32] Medina V,Hürlimann M,Guo Zizheng,et al.Fast physically-based model for rainfall-induced landslide susceptibility assessment at regional scale[J].Catena,2021,201:105213.

[33] Chen Zhuo,Song Danqing,Juliev M,et al.Landslide susceptibility mapping using statistical bivariate models and their hybrid with normalized spatial-correlated scale index and weighted calibrated landslide potential model[J].Environmental Earth Sciences,2021,80(8):1–19.

[34] Sun Deliang,Wen Haijia,Wang Danzhou,et al.A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm[J].Geomorphology,2020,362:107201.

[35] Miao Yamin,Zhu Axing,Yang Lin,et al.A new method of pseudo absence data generation in landslide susceptibility mapping[J].Geography and Geo-Information Science,2016, 32(4):61–67.[缪亚敏,朱阿兴,杨琳,等.滑坡危险度制图中一种新型的负样本采样方法[J].地理与地理信息科学, 2016,32(4):61–67.]

[36] Guo Guo,Chen Yun,Li Minghui,et al.Study on the relationship between development probability and slope of soil landslide[J].Journal of Engineering Geology,2013,21(4): 607–612.[郭果,陈筠,李明惠,等.土质滑坡发育概率与坡度间关系研究[J].工程地质学报,2013,21(4):607–612.]

[37] Zhang Junyi,Ding Yuekai,Sun Deliang.Landslide susceptibility evaluation based on different sample proportion and super parameter optimization:Take Wulong district of Chongqing municipality as an example[J].Journal of Chongqing Normal University (Natural Science),2022,39(5):47–57.[张军以,丁悦凯,孙德亮.基于不同样本比例与超参数优化的滑坡易发性评价——以重庆市武隆区为例[J].重庆师范大学学报(自然科学版),2022,39(5):47–57.]

Uncertainties of Landslide Susceptibility Prediction Modeling: Influence of Different Selection Methods of "Non-landslide Samples"

HUANG Faming[1], *ZENG Shiyi*[1], *YAO Chi*[1*], *XIONG Haowen*[1], *FAN Xuanmei*[2], *HUANG Jinsong*[3]

(1.School of Infrastructure Eng., Nanchang Univ., Nanchang 330031, China;
2.State Key Lab. of Geohazard Prevention and Geoenvironment Protection, Chengdu Univ. of Technol., Chengdu 610059, China;
3.ARC Centre of Excellence for Geotechnical Sci. and Eng., Univ. of Newcastle, Newcastle 2287, Australia)

Abstract: How to select non-landslide samples for landslide susceptibility prediction (LSP) modeling is an important uncertainty affecting the LSP results. To study the influence of different non-landslide sample selection methods on LSP modeling, five sampling methods were proposed (Randomly selected from the whole area, from the specific attribute area with a slope lower than 5°, from the area outside buffer zone which is 300 m from each landslide, selected by information value method, selected by Semi-supervised machine learning) with the same number of landslide grid units, and coupled with Random Forest (RF) to construct random selection-RF, low-slope RF, buffer-based RF, IV-RF, and semi-supervised RF models for LSP. Taking Nankang County of Jiangxi province as the study area, a total of 19 environmental factors such as elevation, slope, population density, and road density were acquired, and 233 landslide inventories were obtained. The landslide inventory was divided into 2 598 grids as landslide samples to construct the input-output of the above-coupled model. Then, the prediction accuracy and the distribution characteristics of predicted landslide susceptibility indexes were used to analyze the LSP modeling uncertainty. To further solve the problem of unreasonable distribution of landslide susceptibility indexes predicted by the coupled model, a sample set with a 1∶2 ratio of landslide to non-landslide was used for LSP, and the condition of the sample set with equal proportion was compared in semi-supervised RF. Results showed that: 1) The prediction accuracy of models such as low-slope RF, buffer-based RF, IV-RF, and semi-supervised RF was substantially better than that of the random selection-RF model, suggesting that accurate selection of non-landslide samples was critical for LSP. 2) The modeling performance of the semi-supervised RF was optimal, which predicted the distribution characteristics of landslide susceptibility indexes more accurately and reliably at landslide∶non-landslide = 1∶2 than at 1∶1. It is necessary to explore the ratio of landslide to non-landslide samples in depth in future studies.

Key words: landslide susceptibility prediction; non-landslide samples selection; semi-supervised machine learning; information value; random forest

（编辑　张凌之）

引用格式:Huang Faming,Zeng Shiyi,Yao Chi,et al.Uncertainties of landslide susceptibility prediction modeling: Influence of different selection methods of "non-landslide samples"[J].Advanced Engineering Sciences,2024,56(1):169–182.[黄发明,曾诗怡,姚池,等.滑坡易发性预测建模的不确定性：不同"非滑坡样本"选择方式的影响[J].工程科学与技术,2024,56(1):169–182.]