2025 年 第 25 卷 第 15 期
2025, 25(15):06220-10
科 学 技 术 与 工 程
Science Technology and Engineering
ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2405249
引用格式:刘少泽, 崔美娟, 付晓祎, 等. 顾及缓冲区范围与负样本优化的随机森林地质灾害易发性评价[J]. 科学技术与工程, 2025, 25(15): 6220-6229.
Liu Shaoze, Cui Meijuan, Fu Xiaoyi, et al. Random forest evaluation of geological hazard susceptibility considering buffer range and negative sample optimization[J]. Science Technology and Engineering, 2025, 25(15): 6220-6229.

顾及缓冲区范围与负样本优化的随机森林地质灾害易发性评价

刘少泽[1], 崔美娟[1], 付晓祎[2], 唐宗源[3*]
(1. 河北地质大学华信学院, 石家庄 050000; 2. 河北省国控矿业开发投资有限公司, 石家庄 050000;
3. 河北地质大学地质调查研究院, 石家庄 050000)

摘 要 为了提升地质灾害易发性评价的精度,以浙江省杭州市富阳区为研究区,提出了考虑缓冲区优化策略的随机森林地质灾害易发性评价方法。首先,选取归一化植被指数、距道路距离、距断层距离、汛期降雨量、坡度、坡向、起伏度、距水系距离和岩性共 9 个评价因子,通过多重共线性分析确保因子的独立性。其次,构建 0.5、1、1.5、2 km 4 种缓冲区范围,以随机采样方法生成负样本点,避免正负样本交叉污染,提高样本代表性和模型的区分能力,并增设一组无缓冲区的随机采样点作为参照。接着,采用随机森林算法对地质灾害易发性模型进行训练和测试,结果表明缓冲区优化策略能够显著提升模型预测精度,且缓冲区的设定存在最优边界。其中,1 km 缓冲区对应的模型 AUC(area under curve)值最高为 0.815,表明该缓冲区设定下采集的负样本能够更加准确地区分地质灾害特征。最后,基于最优缓冲区与随机森林算法(random forest,RF)模型的易发性评价结果,高易发区主要集中在西北部和东南部的山区地带,且频率比随着易发性等级升高而不断增大,验证了这一方法的科学性,可为富阳区的地灾防治工作提供依据。

关键词 地质灾害; 缓冲区; 随机森林; 易发性评价; 地理信息
中图法分类号 P694; **文献标志码** A

Random Forest Evaluation of Geological Hazard Susceptibility Considering Buffer Range and Negative Sample Optimization

LIU Shao-ze[1], CUI Mei-juan[1], FU Xiao-yi[2], TANG Zong-yuan[3*]
(1. Huaxin College of Hebei GEO University, Shijiazhuang 050000, China;
2. Hebei State-controlled Mining Development Investment Co., Ltd., Shijiazhuang 050000, China;
3. Geological Survey and Research Institute, Hebei GEO University, Shijiazhuang 050000, China)

[**Abstract**] In order to improve the accuracy of geological hazard susceptibility assessment, Fuyang District in Hangzhou, Zhejiang Province was taken as the research area and a random forest method was proposed for evaluating geological hazard susceptibility, considering buffer zone optimization strategies. Firstly, nine evaluation factors were selected: normalized difference vegetation index, distance to roads, distance to faults, rainfall during the flood season, slope, aspect, ruggedness, distance to water systems, and lithology. Multicollinearity analysis was conducted to ensure the independence of the factors. Secondly, buffer zones of 0.5 km, 1 km, 1.5 km, and 2 km were constructed. Negative sample points were generated using random sampling to avoid cross-contamination between positive and negative samples, enhance sample representativeness, and improve the model's discrimination capability. An additional set of random sampling points without buffer zones was also established for comparison. The random forest algorithm was then used to train and test the geological hazard susceptibility model. Results indicated that the buffer zone optimization strategy significantly improved the model's predictive accuracy and that there was an optimal boundary for the buffer zone. The model's AUC(area under curve) value was highest at 0.815 for the 1 km buffer zone, indicating that negative samples collected within this buffer zone could more accurately distinguish geological hazard characteristics. Finally, based on the susceptibility evaluation results of the optimal buffer zone and the random forest model, high susceptibility areas were mainly concentrated in the mountainous regions in the northwest and southeast. The frequency ratio increased with the susceptibility level, validating the scientific validity of this method. This approach can provide a ba-

收稿日期: 2024-07-12 **修订日期**: 2024-10-29
基金项目: 河北省青年科学基金(D2022403027)
第一作者: 刘少泽(1989—),男,汉族,河北石家庄人,硕士,讲师。研究方向:矿产地质调查,地质灾害。E-mail:qiu_la_si@163.com。
*通信作者: 唐宗源(1990—),男,汉族,河北邯郸人,博士,讲师。研究方向:岩石地球化学、地质灾害。E-mail:t_izzy@sina.com。

sis for geological hazard prevention and control in Fuyang District.
[**Keywords**]　geological hazard; buffer zone; random forest; susceptibility assessment; geographic information

浙江省地处东部沿海地区,省内山区、丘陵地形广泛分布,气候条件复杂,各类地质灾害(如崩塌、滑坡、泥石流等)频繁发生,对人民生命和财产安全构成了严重威胁,极大地破坏了区域生态系统的稳定性[1]。地质灾害易发性评价作为灾害防治领域的一种重要手段,能够基于遥感卫星影像数据[2-4]、数理统计[5-6]、人工智能算法[7-10]等方式,结合历史地质灾害点的数据输出区域地质灾害风险等级的分布情况,为现场的灾害风险调查及隐患点排查提供有效依据。

目前,基于地理信息系统(geographic information system, GIS)和遥感(remote sensing,RS)技术的地质灾害易发性评价,按照类型划分可以分为启发式方法(经验模型)[11]和数据驱动方法(机器学习算法等)[12],二者之间最大的区别在于权重的分配过程是否能够降低人为主观性;而按照评价模型的数量进行划分,则可分为单一模型和组合模型这两大类[13-15],进一步地,单一模型可分为统计分析方法和机器学习算法。其中,信息量法[16]、确定性系数法[17-19]、证据权法[20]等属于统计分析方法,K 近邻算法[21]、朴素贝叶斯[22]、人工神经网络[23]、决策树[23]、支持向量机[24]、随机森林[25]等属于机器学习算法。

相较而言,机器学习算法对于数据的依赖程度较高,虽然能够有效克服人为主观性在评价过程中的影响,但对数据本身的质量提出了更高的要求。因此,如何构建一个科学合理的数据集直接影响算法模型的预测精度。由于地质灾害易发性评价本质上属于二分类问题,通常将地质灾害的历史清单作为正样本数据集,而负样本数据集一般是通过现场调查、遥感影像解译或者数据点随机构建[26]。受各种客观条件的限制,负样本数据集的构建通常采用随机方式,而这种方法本身存在极大的随机性,忽视了正负样本点之间的空间距离,难以有效区分二者之间的特征差异,降低了负样本的可靠性,进而影响模型的预测精度[27-28]。目前,许多研究并未充分考虑负样本在随机采样过程中存在的问题,而只是关注模型本身的精度高低[29-30]。因此,为了能够有效地克服这一问题,现引入灾害点的缓冲区,通过限定负样本的采集范围,确保正负样本点保持一定的最小空间距离,避免因距离过近导致的特征混淆。由于随机森林算法具有高效的特征选择及数据处理能力,将其用于地质灾害易发性评价能够有效解决二分类问题,尤其是在高维数据的分析上,随机森林可通过集成学习的方法提高模型的泛化能力,并在多样化的地质特征处理中表现出较高的准确性和稳定性。

为了能够更好地提升地质灾害易发性评价模型的预测精度,基于随机森林算法通过设定负样本缓冲区的范围,优化随机选取负样本的方式,并以浙江省杭州市富阳区作为研究区,结合受试者工作曲线(receiver operating characteristic, ROC)及频率比等数理统计方法对预测模型的精度进行评估,进而为富阳区的灾害防治工作提供依据。

1　研究区概况和数据来源

1.1　研究区概况

如图 1 所示,杭州市富阳区位于浙江省西北部,全区总面积为 1 821.08 km^2,地处钱塘江中游平原与天目山脉过渡带,境内有富春江和东洲岛等景区。天目山脉绵亘西部,呈东西走向,东部属于钱塘江河谷低山丘陵区,富春江为主要河流,西入东出,斜贯区境中部。地势由东南、西北向中部倾斜,山地丘陵、平原盆地、水域范围分别占区境总面积的 78.61%、16.36%、5.02%。年降雨量为 1 000 ~ 1 200 mm,雨量主要集中在梅雨季节和台风季节,即 6—9 月。富阳地区的地质灾害类型主要是滑坡,其次是崩塌和泥石流等。灾害主要发生在降雨强度较高以及人类工程活动频繁的区域,在暴雨和台风季节尤为频繁,规模以中小型为主。因此,开展研究区的地质灾害易发性评价对于当地的防灾减灾工作而言是十分必要的[31]。

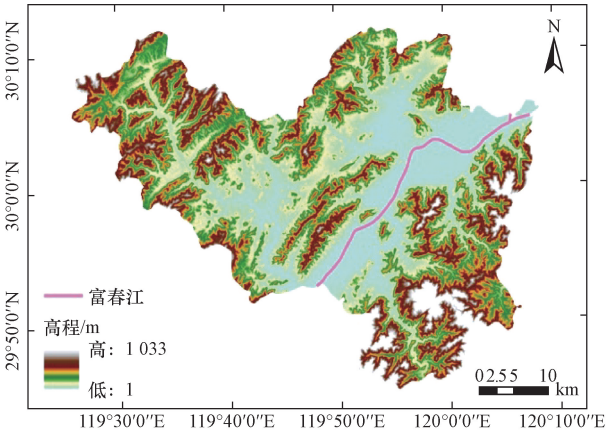

图 1　研究区概况

Fig. 1　Overview of the study area

1.2　数据来源

本文中采用的主要数据源如表 1 所示。由于不同栅格数据影像的分辨率各不相同,为了便于后续

表 1 数据来源和数据类型

Table 1 Data sources and data types

基础数据	数据来源	数据格式
灾害点数据	资源环境科学数据平台(http://www.resdc.cn/Default.aspx)	矢量
DEM	地理空间数据云平台(https://www.gscloud.cn)	栅格
地层岩性数据	资源环境科学数据平台 1:25 万地质图(http://www.resdc.cn/Default.aspx)	矢量
水系和路网数据	全国地理信息资源目录服务系统(https://www.webmap.cn/main.do?method=index)	矢量
NDVI	地理空间数据云平台(https://www.webmap.cn/main.do?method=index)	栅格
降雨数据	国家青藏高原科学数据中心平台(https://data.tpdc.ac.cn)	栅格

的评价单元计算,通过 Arcgis 中的重采样工具将所有的栅格影像分辨率统一为 30 m。

1.3 评价因子选取

在特定的地质环境条件下容易孕育各类特殊的地质灾害,如崩塌、滑坡、泥石流等,而按照孕灾条件进行划分,可将其分为地形条件、物源条件以及水源条件这三类。根据研究区的地形地貌及水文地质等条件,结合相关文献及前人经验[20,32-33],本文选取了 9 个评价因子,包括归一化植被指数(normalized difference vegetation index,NDVI)、距道路距离、距断层距离、汛期降雨量、坡度、坡向、起伏度、距水系距离、岩性。

为了避免评价因子之间存在共线性的关系,研究使用逐步回归法对所选因子进行多重共线性检验。通过容忍度(toleration,TOL)和方差膨胀因子(variance inflation factor, VIF)检验各特征因子的相关性。由表 2 可知,所选特征因子的容忍度均大于 0.1,方差膨胀因子均小于 10,表明所选因子之间的共线性程度低,有着较强的独立性。各个评价因子的分布情况如图 2 所示。

(1)NDVI。植被覆盖程度对地质灾害的发生有着重要影响。植被覆盖度较低的区域,受到的河流侵蚀作用强烈,土体稳定性差,大量裸露的岩石在长期风化作用下,容易产生松散固体物质,诱发各种地质灾害。研究区整体植被覆盖程度高,东北部和中部为城市生活区,富春江横亘其中,故该区域内的 NDVI 值低[图 2(a)]。

(2)距道路距离。道路附近区域由于人类活动频繁,土体受到扰动,稳定性降低,容易诱发各类地质灾害。研究区内道路分布广泛,距离城市越近的区域道路密度越高,在东西部的山区地带因为高速公路的建设,削山填谷、坡脚开挖的现象在公路沿线普遍存在,为地质灾害的发生提供了有利的物源条件[图 2(b)]。

(3)距断层距离。断层对地质灾害的发育演化具有重要作用。距离断层近的区域,地质活动频繁,土体稳定性差,容易发生滑坡、地震等地质灾害。从图 2(c)可以看出,距离断层较近的区域主要分布在研究区的西北部和东南部,这一分布特点构成了富阳市核心区“两山夹一江”的地理格局。这些区域的地质环境并不稳定,易受断层活动带的影像。而距离断层较远的区域其地质活动较少,相对稳定。

(4)汛期降雨。降雨对地质灾害的发育演化具有重要作用。降雨量大的区域,河流冲刷作用强,土体稳定性差,容易产生松散固体物源,诱发各类地质灾害。本文中采用研究区 5—9 月的汛期降雨量数据来研究降雨对地质灾害的影响[图 2(d)],汛期降雨量较大的区域主要分布在研究区的北部和南部的山区地带,这些区域的降雨量显著高于其他地区,且地质环境更为复杂。汛期降雨量较小的区域则主要分布在中部和北部的城区地带,该区域内的降雨量相对较少,地质灾害发生的风险较低。

(5)坡度。坡度较大的区域,水流汇聚和发散作用强,土体稳定性差,容易发生滑坡等地质灾害。研究区内坡度较陡的区域主要分布在研究区的西南部和东北部,该区域以山地和丘陵地形为主,地形起伏大、坡度陡峭,且地质环境相对复杂。坡度较缓的区域主要分布在研究区的中部和东部,这些区域的地形平缓,土体相对稳定,地质灾害的风险较低[图 2(e)]。

(6)坡向。坡向是影响地质灾害发生和分布的重要因素之一。不同坡向的区域,其阳光照射、植被覆盖、土壤湿度等环境条件存在显著差异,进而影

表 2 地质灾害易发性评价因子多重共线性分析

Table 2 Multicollinearity analysis of geological hazard susceptibility assessment factors

评价因子	NDVI	距道路距离	距断层距离	汛期降雨	坡度	坡向	起伏度	距水系距离	岩性
VIF	2.31	1.57	1.08	1.59	1.67	1.32	1.97	1.72	1.46
TOL	0.46	0.28	0.75	0.94	0.73	0.64	0.77	0.81	0.45

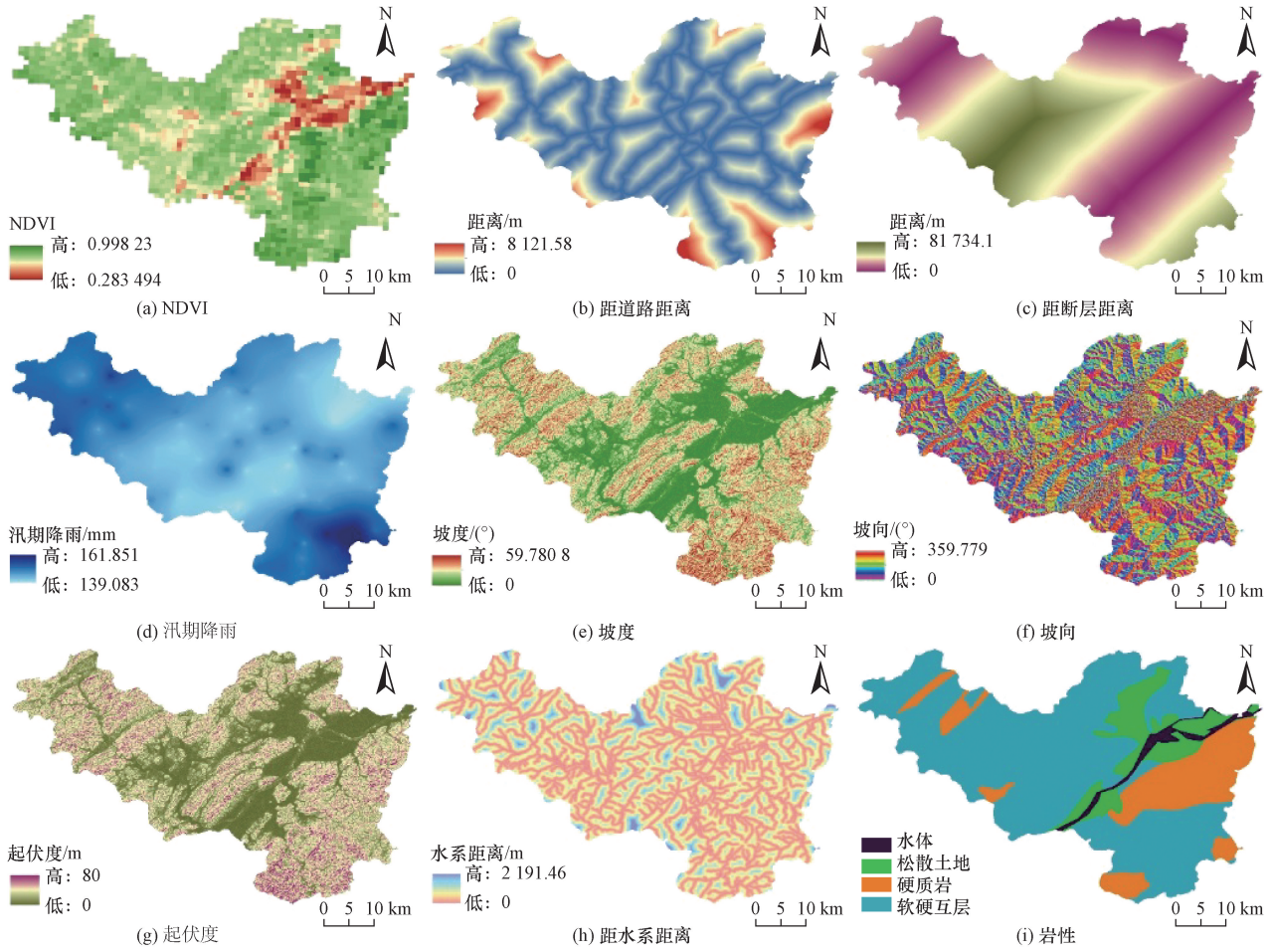

图 2　研究区评价因子

Fig. 2　Evaluation factors of the study area

响滑坡等地质灾害的风险。一般而言,南向坡可能受到的阳光照射较多,植被覆盖较好,土壤湿度较低,地质灾害风险较小。北向坡受到的阳光照射较少,植被覆盖较差,土壤湿度较高,地质灾害风险较大[图 2(f)]。

(7)起伏度。起伏度大的区域,地形变化剧烈,坡度陡峭,容易形成滑坡等地质灾害。从图 2(g)可以看出,起伏度较大的区域主要分布在研究区的西北部和东南部,这些区域的地形起伏剧烈,地质环境复杂,潜在的地质灾害风险较高。起伏度较小的区域主要分布在研究区的中部和南部,这些区域的地形相对平缓,地质环境较为稳定,地质灾害的风险较低。

(8)距水系距离。距水系较近的区域,地表径流对斜坡体的冲刷作用越强,长期冲刷作用会导致周围山体的土壤流失,临空面增大,土体稳定性变差的情况,进而诱发滑坡和泥石流等地质灾害。从图 2(h)可以看出,研究区内水系发达、河网分布广泛,水文活动频繁,地质灾害的潜在风险较高。

(9)岩性。不同岩体的力学性质和层间结构面的状态决定了斜坡的稳定性。通常情况下,软岩或软硬互层结构的岩土体抗剪强度低、风化程度高、稳定性差,容易演化为滑动面的主要部位。这些岩体在外力作用下容易发生变形和破裂,导致斜坡失稳,进而引发滑坡等地质灾害。研究区主要以软硬互层为主土体性质相对软弱;中部平原地带存在较大面积的松散土体,属于河流冲积平原;东部山区地带硬质岩层广泛分布,相较而言土体性质更为坚固稳定,不利于地质灾害的发生[图 2(i)]。

2　研究方法

提出考虑缓冲区优化策略的随机森林地质灾害易发性评价方法,通过在不同的缓冲区范围外随机创建非地质灾害点(负样本点),构建与地质灾害点(正样本点)相同数量的负样本数据集,以保证每一次的模型训练其样本比例都是均衡的。基于不同缓冲区下的样本数据集,采用随机森林算法(random forest,RF)对地质灾害易发性模型进行训练测试,评估不同缓冲区对评价模型预测精度的影响,进而获取最优采样范围以提升研究区地质灾害易发性评价的可靠度,技术流程如图 3 所示。

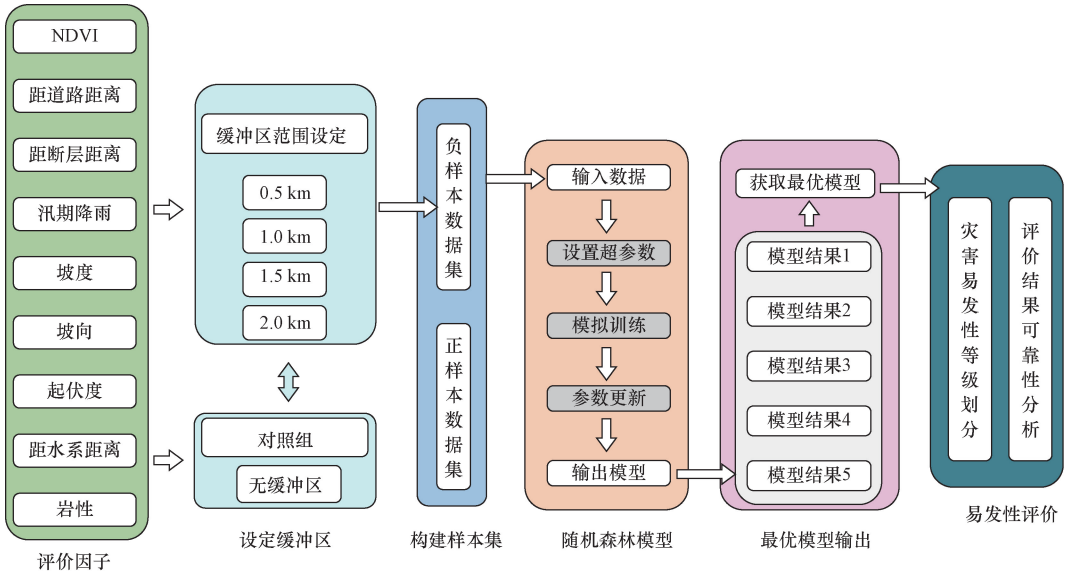

图 3 技术流程图

Fig. 3 Technical flowchart

2.1 评价单元

在地质灾害易发性的评价中,主要有流域单元、斜坡单元和栅格单元这 3 种。相较于前面两者,采用栅格单元具有以下几个优势。

(1)栅格单元方法能够充分利用地理信息系统(GIS)的强大数据处理和空间分析能力。通过将研究区域划分为多个大小相同的栅格单元,可以更精确地表达空间异质性,提高评价结果的空间分辨率。

(2)栅格单元方法便于数据的集成和管理。地质灾害易发性评价通常需要整合多种数据源,如地形、地质、降雨、土地利用等信息。将这些数据转换为栅格格式后,可以通过栅格计算实现多因子的综合分析和评价。

(3)栅格单元方法具备较好的模型扩展性和通用性。不同区域的地质灾害类型和诱发因素可能不同,但通过设定适当的评价指标和权重,可以在不同区域内应用同一评价模型,确保评价结果的可比性和一致性。

结合这些特点,选用分辨率为 30 m×30 m 的栅格单元作为评价单元,将研究区总共划分为 2 292×1 652 个栅格单元。

2.2 缓冲区构建

缓冲区分析作为 ArcGIS 软件的重要空间分析功能,是指在给定地理对象(如点、线或面)周围创建的一定距离的区域,常用于噪声影响范围、污染物扩散、热点吸引力等空间分析,基本内涵是空间要素能够扩散、影响的范围。对于点对象而言,将会生成以点为中心的圆形区域,能够有效地处理相交、合并等拓扑问题,有助于进一步划分受影响的区域。因此,将缓冲区这一概念引入地质灾害易发性评价的非灾害点择取过程中,则可以防止样本交叉污染,减少样本偏差,提高模型的区分能力。

具体而言,研究通过在正样本点周围创建不同大小的缓冲区,控制负样本点的空间分布区域,确保负样本点样本不会落在地质灾害影响区的范围内,进而避免样本交叉污染。如果负样本点都集中在灾害点附近,则可能导致模型过于保守,无法准确识别真正的高风险区域;使用缓冲区控制采样法则可以更均匀地分布样本,减少偏差,提高模型的稳健性,避免因样本选择偏差引起的易发性高估或低估。因此,设定合理的缓冲区范围,测试哪种缓冲区距离的负样本点在模型中表现最佳,可使正样本点和负样本点在地形、植被和土地利用等方面的差异更加明显,有助于算法模型在训练过程中学到更具区分性的特征,提高预测性能。

在实际计算中,一般是基于矢量数据(shp)进行分析,其计算方法有 Euclidean(欧氏缓冲区)和 Geodesic(测地线缓冲区)两种方式。由于研究区的范围相对较小(1 821 km^2),本文中选用欧氏缓冲区的方法进行计算,该方法以二维笛卡尔坐标系为基准计算缓冲距用于表示两点之间直线距离,投影后将以区域图斑的形式显示。

研究区境内的历史地质灾害点(含崩塌、滑坡、泥石流)数量为 137 个,即图 4 的正样本点,这些灾害点主要分布在富阳区境内的西北部和东南部。结合研究区的空间大小,分别设定了 0.5、1、1.5、2 km 这 4 种缓冲区的范围(图 4),负样本点之间的距离不小于 500 m,在 ArcGIS 中运用随机采样的方法

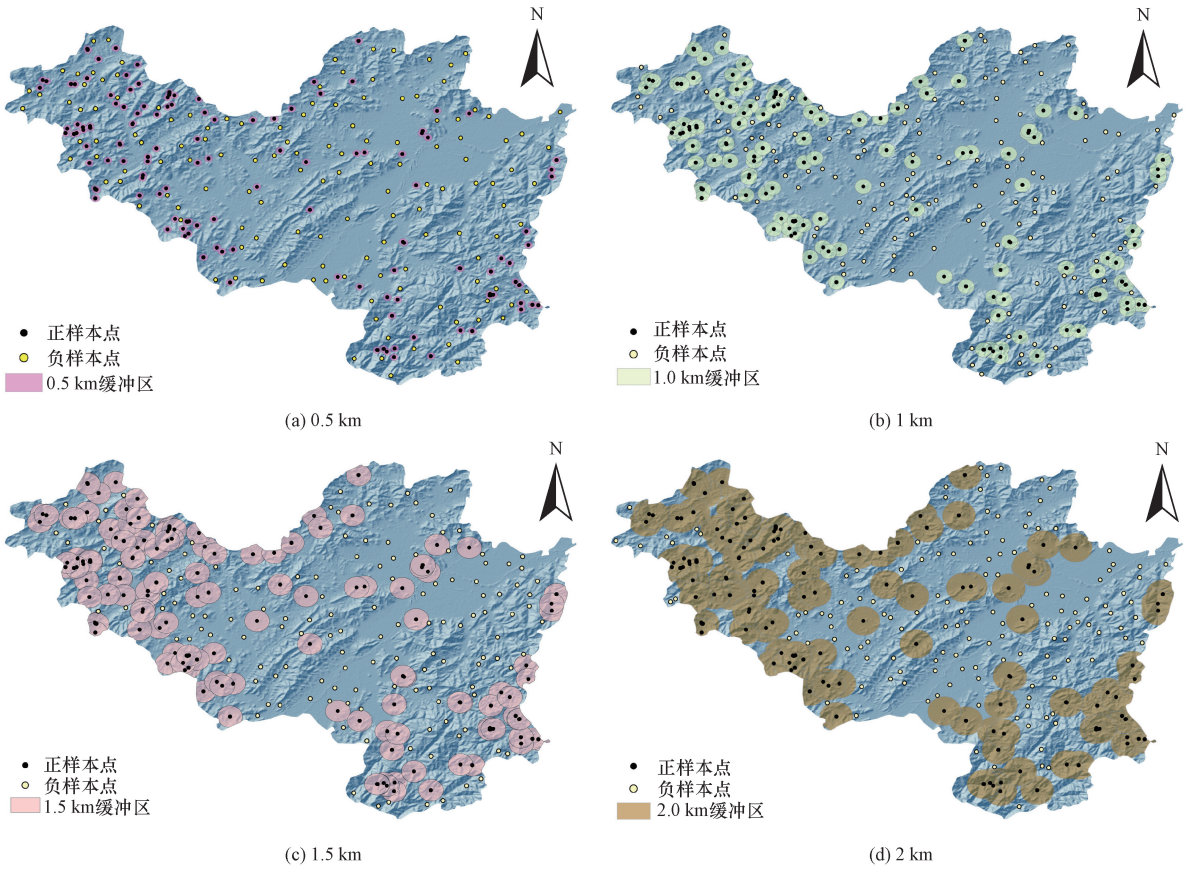

图 4　缓冲点及样本分布情况

Fig. 4　Distribution of buffer points and samples

依次生成与正样本点相同数量的负样本点,共 4 个数据集。

2.3　随机森林模型构建

2.3.1　随机森林算法原理

随机森林(random forest,RF)是一种高精度的分类器算法,主要由多个决策树(decision tree,DT)组成,能够有效处理原始信号中的噪声和异常值。RF 中的每个决策树 $T(i)$ 会生成一个分类结果 $R(i)$,然后通过对所有决策树的结果进行投票,得票最多的 $R(i)$ 就是最终的分类结果。常见的决策树构建算法包括 CLS、ID3、C4.5 和 CART 等[16, 26]。

RF 算法通常采用 Bagging 重采样集成算法。与单一分类器相比,Bagging 集成法通过对多个训练样本进行多次分类训练,从而得到多个分类结果,并最终汇总为一个分类结果。

因此,RF 算法是基于决策树和 Bagging 重采样算法的多分类器。Bagging 重采样算法使用有放回的采样方式,使得每个子数据集与原始数据集具有相同的数据量,并且数据元素在不同的子数据集中可以重复出现,这增加了数据选取的随机性[29]。

2.3.2　建模流程

地质灾害易发性评价是通过分析每个评价因子的空间分布情况,并考虑研究区内不同评价因子对于地质灾害的影响程度及关联关系,运用不同的计算方法确定地质灾害发生的可能位置及其易发程度的评价方法。本文中将 9 个评价因子作为易发性评价的条件属性(样本特征),并将其与对应的正负样本点(样本标签)共同组成初始数据集。

RF 算法的基本步骤如下:

(1)构建训练样本数据集。

$$T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\} \quad (1)$$

式(1)中:T 为整个训练数据集;$(\boldsymbol{x}_i, y_i)$ 对应一个训练样本;$\boldsymbol{x}_i$ 表示第 i 个样本的特征向量,即输入数据;y_i 表示第 i 个样本的标签或目标值,即输出数据。

通过 Bagging 对样本数据集 T 进行重采样,形成自助样本集 θ_k。

(2)构建二叉树。决策树模型为 $\{h(a, \theta_k, k = 1, 2, \cdots, K)\}$,在 scikit-learn 算法库中调用 CART 算法,获取样本集 θ_k 对应的二叉树模型结果。

(3)循环迭代。重复执行步骤(1)、步骤(2),直至所有特征属性都已使用,并在训练集基础上生成能够准确分类的决策树。然后,将所有生成的决策树进行组合,最终构建基于随机森林(RF)的监测模型。

(4)判断样本类别。对于测试集上的样本,计算各棵树的投票结果确定各样本对应的类别,公式为

$$c = \arg\max_c \left\{ \frac{1}{K} \sum_{k=1}^{K} I[h(a,\theta_k) = c] \right\} \tag{2}$$

式(2)中:c 为类别;K 为决策树的数量;h 为决策树模型;a 为输入样本;θ_k 为第 k 棵决策树的参数;I 为某个指示器函数。

(5)准确性判别。对于所有测试样本,经过式(2)计算后,将生成混淆矩阵 $\boldsymbol{C}$。其中,元素 $\boldsymbol{C}(i,j)$ 表示被分类为类别 j 的测试样本实际属于类别 i 的总次数。如果 $i=j$,则表示该类测试样本被正确分类。

结合上述计算流程,研究基于网格搜索法对随机森林的参数进行优化,优化后的 RF 模型参数值如表 3 所示。

表 3 RF 参数设置

Table 3 RF parameter settings

估计器数量	最大深度	最小样本分割数	最小叶子节点数
60	3	8	5

3 评价结果与分析

3.1 不同缓冲区下的模型精度

为了评估不同缓冲区的正负样本集在评价模型上的预测精度,增设一组不设缓冲区的随机采样数据集。然后基于随机森林算法模型的预测结果,采用受试者工作特征曲线(receiver operating characteristic,ROC)对各组数据集的模型精度进行验证[34]。其中,ROC 曲线通常用于评估二分类模型的性能,它通过绘制模型在不同阈值下的真阳性率(true positive rate, TPR)与假阳性率(false positive rate, FPR)之间的关系,展示分类器的表现。其基本原理是对于每一个可能的分类阈值,计算对应的 TPR 和 FPR,并在二维坐标系中绘制曲线图。当 ROC 曲线越接近左上角时,模型的分类性能越好。与此同时,曲线下方的面积(area under the curve, AUC)越接近 1 表示模型预测效果越佳,而面积小于等于 0.5 时,表明该模型预测性能差且无实际应用价值。以上 5 组数据集的 ROC 曲线及其 AUC 值如图 5 和表 4 所示。

由图 5 可以发现,除了无缓冲区(0 km)的数据集,其余设定缓冲区的数据集在 RF 模型上预测精度都比较高,平均 AUC 值为 0.8。从曲线的形态上对比,离左上角最近的曲线主要为 1.0 km 和 1.5 km 这两个缓冲区的数据集,其次为 2.0 km 和 0.5 km 的缓冲区数据集。在一定程度上,说明对于研究区的负样本点采集区域而言,以 1.0 ~ 1.5 km 作为最小采集边界能够有效地划分灾害点的影响范围,且不因位置间隔过小或过大,出现正负样本点的属性特征差异较为微弱或过分显著,导致模型出现欠拟合或过拟合的情况。

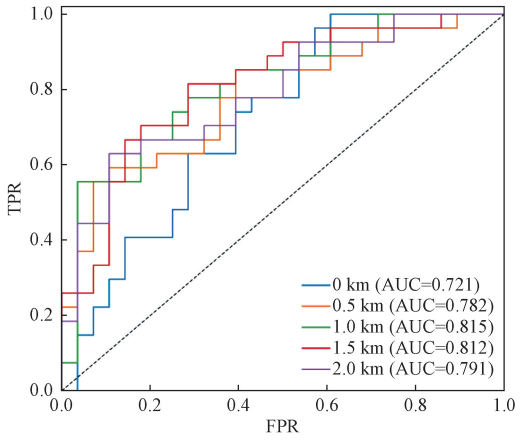

图 5 ROC 曲线

Fig. 5 ROC curve

表 4 模型精度对比

Table 4 Model accuracy comparison

数据集	准确率	精确率	召回率	F_1 分数
无缓冲区(0 km)	0.712	0.682	0.712	0.697
缓冲区(0.5 km)	0.755	0.738	0.760	0.749
缓冲区(1.0 km)	0.785	0.769	0.792	0.780
缓冲区(1.5 km)	0.781	0.765	0.788	0.776
缓冲区(2.0 km)	0.760	0.744	0.768	0.756

进一步地,结合各数据集的模型预测精度(AUC 值)可知,缓冲区 1.0 km 的 AUC 值最高为 0.815,表明模型在该缓冲区下拥有最佳的预测性能,表明以这一距离作为最小采集边界,随机生成的样本点能够更加准确地捕捉与地质灾害发生相关的特征,从而提高了模型的预测能力。而随着缓冲区的增加,AUC 值并未持续增加,反而在缓冲区 1.5 km(AUC = 0.812)和 2.0 km(AUC = 0.791)时略有下降,说明过大的缓冲区可能包含了过多无关的信息或噪声,反而降低了模型的性能。此外,较小的缓冲区(0.5 km)的 AUC 值(0.782)也比最佳缓冲区的低,这可能是因为缓冲区过小使其无法涵盖足够多的相关信息,RF 模型在训练过程中难以识别更加复杂的特征属性值。

表 4 中,当缓冲区范围为 1 km 时,模型具有最高的准确率,表明此时模型正确分类的比例最高,

且此时的模型具有最高的召回率(0.792),表明模型识别正样本的能力最强;当缓冲区的范围为 1.5 km 时,虽然精确率与 1.0 km 的相近,但是在召回率和 F_1 分数上均偏低,表明该模型识别正样本的能力相对较弱,存在一定的局限性。

由此可知,缓冲区大小对于地质灾害易发性评价模型的准确性存在着显著影响,选择合适的缓冲区可以显著提高模型的预测能力。基于当前的 5 个评价指标可知,1 km 缓冲区是研究区构建负样本数据集的最佳选择。

3.2　最优模型的结果检验与分析

由上可知,基于 1.0 km 缓冲区数据集训练得到的随机森林为最优模型,将研究区各个评价指标的栅格图层数据输入至这一训练好的评价模型,可输出全区范围内各个栅格单元的易发性指数。然后,运用自然断裂法将全区的易发性分为极低易发区[0.081,0.317]、低易发区[0.317,0.451]、中易发区[0.451,0.552]、高易发区[0.552,0.601]、极高易发区[0.601,0.760]5 个等级,结果如图 6 所示。

由图 6 可知,高易发区主要集中在研究区的西北部和东南部,该区域地势险峻,断层发育显著,部分地区岩土体质松散破碎,存在着丰富的松散物质,为各类地质灾害的发育提供了有利的孕灾条件;同时,高速公路在这些山区丘陵地带的广泛建设,大规模的切割和平整以及削坡挖方,极大地破坏了原有的地质结构和斜坡稳定性,增加地质灾害发生的风险。相较而言,极低、低易发区则主要集中在中部的平原地带,该区域地势平缓、坡度小,且中部有富春江横贯其中,具有良好的天然排水条件,使得汛期降雨能够及时排走,有效减小土体内部的动水压力,从而降低地陷、滑坡和泥石流等不良地质现象的发生概率。

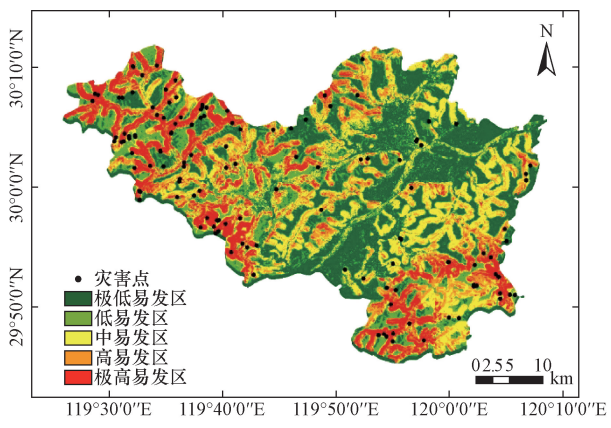

图 6　研究区易发性评价图

Fig. 6　Susceptibility assessment map of the study area

为进一步分析模型输出的易发性等级与历史地质灾害点分布情况的吻合程度,本文中采用频率比确定研究区易发性评价结果的合理性与准确性,若极高、高易发区频率比较大,低易发区频率比较小,则说明评价结果是合理的。其中,地质灾害频率比的计算公式为

$$P_i = \frac{N_i / \sum_i^n N_i}{S_i / \sum_i^n S_i} \tag{3}$$

式中:P_i 为等级 i 的频率比;N_i 为等级 i 的灾害点数量;S_i 为等级 i 的沟域面积。

从表 5 可知,研究区地质灾害极高易发区频率比最大为 2.433,远大于极低易发区的 0.345,且随着易发性等级增高,频率比逐渐增大。此外,极高易发区和高易发区的频率比较大,低易发区和极低易发区的频率比较小,表明富阳区地质灾害易发性评价结果与历史地质灾害的分布情况相符,故研究提出的方法是合理可靠的。

表 5　地质灾害易发性评价结果统计表

Table 5　Statistical table of geological hazard susceptibility assessment results

易发性等级	面积/km²	面积比/%	灾害点数量	灾害点数量比/%	频率比
极低易发区	539.763	29.639	14	10.219	0.345
低易发区	406.533	22.323	22	16.058	0.719
中易发区	345.630	18.979	24	17.518	0.923
高易发区	272.358	14.955	30	21.898	1.464
极高易发区	256.795	14.101	47	34.307	2.433

注:频率比 = 灾害点数量比/面积比。

4　结论

基于多源数据和 ArcGIS 软件,采用缓冲区优化策略和随机森林模型相结合的研究手段,对浙江省杭州市富阳区进行了地质灾害易发性评价,得到如下结论。

(1)在区域地质灾害易发性评价中,引入缓冲区优化策略能够有效降低传统随机采样中负样本代表性不足的问题。其中,0.5、1.0、1.5、2.0 km 缓冲区的 AUC 值均比不考虑缓冲区的随机采样方法高,结果表明这一优化策略能够显著提升评价模型的预测精度。

(2)基于不同缓冲区的最小边界随机生成的样本点,其所能反映的地质灾害特征信息有着显著差异,对于模型评价而言,存在最优边界的特定缓冲区。实验结果表明,0.5、1。0、1.5、2.0 km 缓冲区的 AUC 值存在先上升后下降的特点,在 1 km 的缓

冲区设定下 AUC 值达到最大为 0.815,表明 1.0 km 是这一研究区采集负样本点的最小边界。

(3)采用缓冲区优化策略和随机森林模型相结合的研究手段能够科学合理地反映研究区易发性等级空间分布情况。结果表明,研究区的风险区域主要集中在西北部和东南部,且随着易发性等级升高频率比不断增大,证明本文所构建评价模型合理可靠。

参 考 文 献

[1] 周诗凯, 刘正华, 余丰华, 等. 浙江省地质灾害气象风险预警一体化建设的探索与实践[J]. 中国地质灾害与防治学报, 2024, 35(2): 21-29.
Zhou Shikai, Liu Zhenghua, Yu Fenghua, et al. Exploration and practice of integrated construction of meteorological risk early warning for geological disasters in Zhejiang Province[J]. The Chinese Journal of Geological Hazard and Control, 2024, 35(2): 21-29.

[2] 寸得欣, 令狐昌卫, 马一奇, 等. 基于 GIS 和加权信息量模型的富源县地质灾害易发性评价[J]. 科学技术与工程, 2024, 24(18): 7563-7573.
Cun Dexin, Linghu Changwei, Ma Yiqi, et al. Evaluation of susceptibility to geological hazard in Fuyuan County based on GIS and weighted information model[J]. Science Technology and Engineering, 2024, 24(18): 7563-7573.

[3] 樊昱初, 吕义清, 杨娜. 基于 SABS-InSAR 的保德矿区地表形变监测与分析[J]. 科学技术与工程, 2024, 24(3): 941-951.
Fan Yuchu, Lü Yiqing, Yang Na. Monitoring and analysis of surface deformation in Baode Mining area based on SABS-InSAR[J]. Science Technology and Engineering, 2024, 24(3): 941-951.

[4] 赵超英, 刘晓杰, 张勤, 等. 甘肃黑方台黄土滑坡 InSAR 识别、监测与失稳模式研究[J]. 武汉大学学报(信息科学版), 2019, 44(7): 996-1007.
Zhao Chaoying, Liu Xiaojie, Zhang Qin, et al. Research on identification, monitoring, and instability model of loess landslides in Heifangtai, Gansu, based on InSAR[J]. Journal of Wuhan University (Information Science Edition), 2019, 44(7): 996-1007.

[5] 陈攀, 葛永刚, 孙庆敏, 等. 基于小流域单元的泥石流易发性评价[J]. 科学技术与工程, 2022, 22(29): 12764-12771.
Chen Pan, Ge Yonggang, Sun Qingmin, et al. Debris flow susceptibility assessment based on catchment [J]. Science Technology and Engineering, 2022, 22(29): 12764-12771.

[6] 王峰, 杨帆, 江忠荣, 等. 基于沟域单元的康定市泥石流易发性评价[J]. 中国地质灾害与防治学报, 2023, 34(3): 145-156.
Wang Feng, Yang Fan, Jiang Zhongrong, et al. Debris flow susceptibility assessment based on valley units in Kangding City[J]. The Chinese Journal of Geological Hazard and Control, 2023, 34(3): 145-156.

[7] 李辉, 翟星, 李琛曦, 等. 河北省泥石流灾害易发性云模型评价方法: 以邢台赵沟村泥石流为例[J]. 科学技术与工程, 2024, 24(25): 10884-10891.
Li Hui, Zhai Xing, Li Chenxi, et al. Development law of debris flow disaster and the evaluation method of lts susceptibility: take mudslide in Zhaogou Village, Xingtai as an example[J]. Science Technology and Engineering, 2024, 24(25): 10884-10891.

[8] Wang J Q, Sun P F, Chen L L, et al. Recent advances of deep learning in geological hazard forecasting[J]. Computer Modeling in Engineering & Sciences, 2023, 137(2): 1381-1418.

[9] 夏辉, 殷坤龙, 梁鑫, 等. 基于 SVM-ANN 模型的滑坡易发性评价——以三峡库区巫山县为例[J]. 中国地质灾害与防治学报, 2018, 29(5): 13-19.
Xia Hui, Yin Kunlong, Liang Xin, et al. Landslide susceptibility assessment based on SVM-ANN model: a case study of Wushan County in the Three Gorges Reservoir area[J]. The Chinese Journal of Geological Hazard and Control, 2018, 29(5): 13-19.

[10] 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能, 2014, 27(4): 327-336.
He Qing, Li Ning, Luo Wenjuan, et al. A review of machine learning algorithms in the context of big data[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(4): 327-336.

[11] Chauhan S, Sharma M, Arora M K, et al. Landslide susceptibility zonation through ratings derived from artificial neural network[J]. International Journal of Applied Earth Observation and Geoinformation, 2010, 12(5): 340-350.

[12] Fang Z C, Wang Y, Peng L, et al. Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping[J]. Computers & Geosciences, 2020, 139: 104470.

[13] Youssef K, Shao K, Moon S, et al. Landslide susceptibility modeling by interpretable neural network[J]. Communications Earth and Environment, 2023, 4(1): 162.

[14] Nachappa T, Ghorbanzadeh O, Gholamnia K, et al. Multi-hazard exposure mapping using machine learning for the state of Salzburg, Austria[J]. Remote Sensing, 2020, 12(17): 2757.

[15] 刘璐瑶, 高惠瑛. 基于证据权与 Logistic 回归模型耦合的滑坡易发性评价[J]. 工程地质学报, 2023, 31(1): 165-175.
Liu Luyao, Gao Huiying. Landslide susceptibility assessment based on coupling of evidence weight and logistic regression models [J]. Journal of Engineering Geology, 2023, 31(1): 165-175.

[16] 刘亚静, 刘红健. 基于信息量-随机森林模型的地震带地质灾害易发性评价: 以松潘-较场地震带为例[J]. 科学技术与工程, 2024, 24(1): 143-154.
Liu Yajing, Liu Hongjian. Evaluation of geological hazard susceptibility in seismic zone based on information data-RF model: a case study of Songpan-Jiaochang Seismic Zone[J]. Science Technology and Engineering, 2024, 24(1): 143-154.

[17] 李远远, 梅红波, 任晓杰, 等. 基于确定性系数和支持向量机的地质灾害易发性评价[J]. 地球信息科学学报, 2018, 20(12): 1699-1709.
Li Yuanyuan, Mei Hongbo, Ren Xiaojie, et al. Geological hazard susceptibility assessment based on certainty coefficient and support vector machine[J]. Journal of Geo-Information Science, 2018, 20(12): 1699-1709.

[18] Long N, De Smedt F. Analysis and mapping of rainfall-induced landslide susceptibility in A Luoi district, Thua Thien Hue Province, Vietnam[J]. Water, 2019, 11(1): 51.

[19] Wu Y L, Li W P, Wang Q Q, et al. Landslide susceptibility assessment using frequency ratio, statistical index and certainty factor models for the Gangu county, China[J]. Arabian Journal of Geosciences, 2016, 9(2): 1-16.

[20] 胡燕, 李德营, 孟颂颂, 等. 基于证据权法的巴东县城滑坡灾害易发性评价[J]. 地质科技通报, 2020, 39(3): 187-194.
Hu Yan, Li Deying, Meng Songsong, et al. Landslide susceptibility assessment of Badong County town based on evidence weight method[J]. Geological Science and Technology Bulletin, 2020, 39(3): 187-194.

[21] 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.
Huang Faming, Yin Kunlong, Jiang Shuihua, et al. Landslide susceptibility assessment based on cluster analysis and support vector machines[J]. Chinese Journal of Rock Mechanics and Engineering, 2018, 37(1): 156-167.

[22] 杨灿, 刘磊磊, 张遗立, 等. 基于贝叶斯优化机器学习超参数的滑坡易发性评价[J]. 地质科技通报, 2022, 41(2): 228-238.
Yang Can, Liu Leilei, Zhang Yili, et al. Landslide susceptibility assessment based on Bayesian optimized machine learning hyperparameters[J]. Geological Science and Technology Bulletin, 2022, 41(2): 228-238.

[23] 田乃满, 兰恒星, 伍宇明, 等. 人工神经网络和决策树模型在滑坡易发性分析中的性能对比[J]. 地球信息科学学报, 2020, 22(12): 2304-2316.
Tian Naiman, Lan Hengxing, Wu Yuming, et al. Performance comparison of artificial neural network and decision tree models in landslide susceptibility analysis[J]. Journal of Geo-Information Science, 2020, 22(12): 2304-2316.

[24] 张越, 宋炜炜. 基于BP神经网络和决策树的昆明市东川区滑坡空间易发性评价[J]. 国土与自然资源研究, 2023(2): 67-70.
Zhang Yue, Song Weiwei. Landslide spatial susceptibility assessment of Dongchuan District, Kunming City, based on BP neural network and decision tree[J]. Land and Natural Resources Research, 2023 (2): 67-70.

[25] 毛伊敏, 周昭飞, 彭喆, 等. 基于不确定多分类支持向量机在滑坡危险性预测的应用[J]. 江西理工大学学报, 2016, 37(3): 102-108.
Mao Yimin, Zhou Zhaofei, Peng Zhe, et al. Application of uncertain multi-class support vector machine in landslide hazard prediction[J]. Journal of Jiangxi University of Science and Technology, 2016, 37(3): 102-108.

[26] 杜鹏, 陈宁生, 伍康林, 等. 基于随机森林模型的藏东南地区滑坡易发性评价及主控因素分析[J]. 成都理工大学学报(自然科学版), 2024, 51(2): 328-344.
Du Peng, Chen Ningsheng, Wu Kanglin, et al. Landslide susceptibility assessment and main controlling factors analysis in southeast Tibet based on random forest model[J]. Journal of Chengdu University of Technology (Science & Technology Edition), 2024, 51(2): 328-344.

[27] Hu Q, Zhou Y, Wang S, et al. Machine learning and fractal theory models for landslide susceptibility mapping: case study from the Jinsha River basin[J]. Geomorphology, 2020, 351: 106975.

[28] 徐胜华, 刘纪平, 王想红, 等. 熵指数融入支持向量机的滑坡灾害易发性评价方法——以陕西省为例[J]. 武汉大学学报(信息科学版), 2020, 45(8): 1214-1222.
Xu Shenghua, Liu Jiping, Wang Xianghong, et al. Landslide susceptibility assessment method integrating entropy index and support vector machine: a case study of Shaanxi Province[J]. Journal of Wuhan University (Information Science Edition), 2020, 45(8): 1214-1222.

[29] 林荣福, 刘纪平, 徐胜华, 等. 随机森林赋权信息量的滑坡易发性评价方法[J]. 测绘科学, 2020, 45(12): 131-138.
Lin Rongfu, Liu Jiping, Xu Shenghua, et al. Landslide susceptibility assessment method based on random forest weighted information[J]. Science of Surveying and Mapping, 2020, 45(12): 131-138.

[30] 刘纪平, 梁恩婕, 徐胜华, 等. 顾及样本优化选择的多核支持向量机滑坡灾害易发性分析评价[J]. 测绘学报, 2022, 51(10): 2034-2045.
Liu Jiping, Liang Enjie, Xu Shenghua, et al. Landslide susceptibility analysis and evaluation using multi-core support vector machine with optimized sample selection[J]. Acta Geodaetica et Cartographica Sinica, 2022, 51(10): 2034-2045.

[31] 朱建军, 李志伟, 胡俊. InSAR变形监测方法与研究进展[J]. 测绘学报, 2017, 46(10): 1717-1733.
Zhu Jianjun, Li Zhiwei, Hu Jun. InSAR deformation monitoring methods and research progress [J]. Acta Geodaetica et Cartographica Sinica, 2017, 46(10): 1717-1733.

[32] 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.
Huang Faming, Yin Kunlong, Jiang Shuihua, et al. Landslide susceptibility assessment based on cluster analysis and support vector machines[J]. Chinese Journal of Rock Mechanics and Engineering, 2018, 37(1): 156-167.

[33] 王佳佳, 殷坤龙, 肖莉丽. 基于GIS和信息量的滑坡灾害易发性评价——以三峡库区万州区为例[J]. 岩石力学与工程学报, 2014, 33(4): 797-808.
Wang Jiajia, Yin Kunlong, Xiao Lili. Landslide susceptibility assessment based on GIS and information value method: a case study of Wanzhou District in the Three Gorges Reservoir area[J]. Chinese Journal of Rock Mechanics and Engineering, 2014, 33(4): 797-808.

[34] 李益敏, 杨蕾, 魏苏杭. 基于小流域单元的怒江州泥石流易发性评价[J]. 长江流域资源与环境, 2019, 28(10): 2419-2428.
Li Yimin, Yang Lei, Wei Suhang. Debris flow susceptibility assessment based on small watershed units in Nujiang Prefecture [J]. Resources and Environment in the Yangtze Basin, 2019, 28(10): 2419-2428.