

<https://doi.org/10.3799/dqkx.2020.032>



基于随机森林的滑坡空间易发性评价： 以三峡库区湖北段为例

吴润泽¹, 胡旭东², 梅红波^{3*}, 贺金勇³, 杨建英¹

1. 中国地质调查局武汉地质调查中心, 湖北武汉 430205

2. 武汉大学遥感信息工程学院, 湖北武汉 430079

3. 中国地质大学资源学院, 湖北武汉 430074

摘要: 滑坡空间易发性分析有助于开展滑坡防灾减灾工作, 训练有效的滑坡预测模型在其中扮演重要角色. 以三峡库区湖北段为研究区, 选取高程、坡度、斜坡结构、土地利用类型、岩土体类型、断裂距离、路网距离、河网距离、以及归一化植被指数这9个影响因子建立滑坡空间数据库, 采用集成学习中的随机森林算法进行滑坡易发性评价. 结果显示, 随机森林抽样训练的方式有利于确定较优的训练参数, 保证随机森林在不过拟合的情况下取得满意的拟合能力和泛化能力. 随机森林绘制的滑坡易发性分级图显示出合理的空间分布, 其中73.35%的滑坡分布在较高和极高级别区域. 而巴东县北部、秭归县中部以及夷陵区南部等区域显示出较高的易发性级别. 性能评估及易发性统计结果均表明随机森林是一种出色的算法, 在滑坡空间预测领域具有较好的适用性.

关键词: 滑坡; 易发性; 随机森林; 三峡库区.

中图分类号: P642.2

文章编号: 1000-2383(2021)01-321-10

收稿日期: 2020-02-21

Spatial Susceptibility Assessment of Landslides Based on Random Forest: A Case Study from Hubei Section in the Three Gorges Reservoir Area

Wu Runze¹, Hu Xudong², Mei Hongbo^{3*}, He Jinyong³, Yang Jianying¹

1. Wuhan Center of Geological Survey, CGS, Wuhan 430205, China

2. School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

3. School of Earth Resources, China University of Geosciences, Wuhan 430074, China

Abstract: Landslide spatial susceptibility assessment can assist to conduct the prevention and mitigation of landslides, in which the application of effective landslide models plays a significant role. Taking Hubei section of the Three Gorges Reservoir Area as study area, nine influencing factors including elevation, slope angle, slope structure, land use, engineering rock group, distance to faults, distance to roads, distance to rivers, and normalized difference vegetation index were selected to establish the landslide spatial database. Then the random forest ensemble algorithm was used to assess landslide susceptibility. The results show that the sampling training scheme of random forest benefits to search suitable training parameters, and enables the random forest achieve desirable fitting ability and generalization skill when avoiding the over-fitting problem. The landslide susceptibility mapping results developed by random forest show a reasonable spatial distribution, where 73.35% of the landslides are located in highly and very highly susceptible areas. Furthermore, the areas of northern Badong county, central Zigui county and the southern Yiling district

基金项目: 三峡库区后续地质灾害防治信息系统建设(No. 0001212012AC50001).

作者简介: 吴润泽(1979—), 男, 工程师, 主要从事地质灾害防治信息化研究. ORCID: 0000-0003-1063-1431. E-mail: wurunze@163.com

***通讯作者:** 梅红波, ORCID: 0000-0001-6377-3877. E-mail: hbmei@cug.edu.cn

引用格式: 吴润泽, 胡旭东, 梅红波, 等, 2021. 基于随机森林的滑坡空间易发性评价: 以三峡库区湖北段为例. 地球科学, 46(1): 321-330.

show a higher susceptibility level. The performance evaluation and statistical results of susceptibility show that the random forest is an excellent algorithm, and has a good applicability in the field of landslide spatial prediction.

Key words: landslide; susceptibility; random forest; Three Gorges reservoir area.

0 引言

滑坡是常见的地质灾害类型之一,具有突发性强、分布广、破坏性大等特点,在世界范围内对不同的国家造成巨大且无法挽回的生命财产损失。而近些年来随着人类社会的快速发展及人类活动范围的不断扩大,由滑坡造成的危害也日趋严重。研究显示,发展中国家比发达国家更容易受到地质灾害的影响(Alcántara-Ayala, 2002)。作为最大的发展中国家,中国一直以来饱受滑坡地质灾害的威胁。因此,利用高效可靠的评价技术进行滑坡区域易发性评价,精确的识别滑坡地质灾害高危高发区,提高地质灾害预报水平,为区域地质灾害防灾减灾工作提供依据,是一项非常有意义的工作。

纵观国内外学者使用的滑坡易发性评价技术,其大致可以归纳为知识驱动法、数据驱动法以及两者相结合的方式(Tehrany *et al.*, 2013)。知识驱动法直观简单,评估过程容易理解,在技术不发达的年代得到了广泛的运用(Fenti *et al.*, 1979)。其缺点是主观依赖性太强,评价结果无法比较。随着计算机技术的发展,数据驱动法的运用越来越广泛(Song *et al.*, 2012)。其中机器学习法以其擅长处理因子间的非线性关系而著名,适用于滑坡易发性分析,如李远远等(2018)将确定性系数与支持向量机相结合运用到云南省地质灾害多发区段的泸水县地区。武雪玲等(2016)在蚁群算法的基础上使用支持向量机分析三峡库区长江干流岸坡滑坡易发性预测精度。郭子正等(2018, 2019)将传统的统计方法证据权法与BP神经网络结合起来进行滑坡易发性分析。Zare *et al.* (2013)对多层感知器和径向基函数算法这两种神经网络模型进行比较,结果证明径向基函数算法具有更出色的空间预测能力。Felicísimo *et al.* (2013)指出分类与回归树(CART, classification and regression tree)具有很高的探测能力,是一种非常有潜力的机器学习方法,尤其适用于地理空间上的滑坡地质灾害预测任务。此外,C4.5决策树、朴素贝叶斯、贝叶斯网络、逻辑回归等机器学习算法在滑坡地质灾害易发性评价领域也有出色的表现。

然而,在探索滑坡预测的精度方面,许多学者付诸巨大努力用以提升预测模型的性能。其中,集成学习是机器学习的分支之一,其通过组合不同的机器学习模型从而获得更强大的预测能力,因此受到越来越广泛的重视。本文基于集成学习中著名的随机森林(RF, random forest)算法构建滑坡易发性评价模型,旨在全面了解随机森林这种集成算法在滑坡易发性评价领域中的适用性。为此,选择地质灾害防治工作典型示范地区三峡库区湖北段作为研究区,进行滑坡易发性定量区划分析。

1 随机森林

随机森林由Breiman (2001)首次提出,可以看作是Bagging集成学习的一个拓展。随机森林以决策树为基本模型通过构造不同的训练数据集以及不同的特征空间来产生一系列具有差异性的决策树模型,组合策略方面通常使用投票或者取平均值得到最终的决策。

随机森林首先使用自助法(bootstrap)重采样技术,从原始训练样本集 T 中有放回地重复随机抽取 n 个样本(一般为 $2/3$)生成新的训练样本集合,每一个独立抽取的训练样本用于训练一棵树,基于自助样本集生成的 n 个决策树组成森林。其中每棵树具有相同的分布,分类误差取决于每一棵树的分类能力和它们之间的相关性。而剩下未被抽取的数据集被称为袋外数据(OOB, out of bag),其误差是一种无偏估计,可以用来验证模型的性能好坏以防止过拟合。

随机森林的泛化误差 P^* 定义为:

$$P^* \leq \frac{\rho(1-s^2)}{s^2}, \quad (1)$$

其中: ρ 为决策树相关度平均值; s 为决策树的平均强度。

由式(1)中可以看出,若要增强随机森林的泛化性能,需要降低各个基础决策树模型之间的相关度或者增强决策树的强度。为此在随机森林算法中引入特征变量的随机扰动,即参与每棵树的分裂节点可能是不同的。因此,在随机森林算法中训练样本和每个节点变量都是随机产生的。

决策树算法是一个庞大的家族,本文所使用的随机森林是基于分类与回归树构建的(Breiman *et al.*, 1984). 在建立每一棵分类与回归树的过程中,每个节点的分裂过程依靠计算分裂后的样本“纯度”来完成,分类与回归树采用基尼系数来衡量这种所谓的“纯度”,即随机森林通过基尼指数进行树的分裂完成决策. 基尼系数越小则代表样本的纯度越高,树划分的效果越好. 假设样本集 T 中包含 k 个类别,则样本集的基尼系数可以表示为:

$$\text{gini}(T) = 1 - \sum_{i=1}^k p_i^2, \quad (2)$$

其中: p_i 是 T 中包含类 i 的概率. 若将 T 划分为两个子集 T_1 和 T_2 ,则划分后的基尼系数可以表示为:

$$\text{gini}(T_1, T_2) = \frac{|T_1|}{|T|} \text{gini}(T_1) + \frac{|T_2|}{|T|} \text{gini}(T_2), \quad (3)$$

其中: $|*|$ 代表当前样本集中元素的数量.

2 研究区概况及数据源

2.1 研究区概况

作为地质灾害研究典型示范区,三峡库区一直

以来备受关注(郭子正等, 2019; 史绪国等, 2019). 本次研究区为三峡库区湖北段(图1), 该区属长江中上游, 包含恩施巴东县、宜昌兴山县、秭归县及夷陵区三县一区, 占地面积约 11 360 km², 地理坐标为东经 110°06'~东经 111°40'以及北纬 30°14'~北纬 31°34'. 研究区地处中国地形第二阶梯与第三阶梯过渡带, 区内高峰林立, 重峦叠嶂, 最大高程分布达 3 000 m. 该区气候温暖, 气候类型主要为亚热带季风气候, 多年年均气温约 17 °C. 雨量充沛, 多年年均降雨约 1 320 mm. 断裂构造发育, 交错分布. 地质岩性主要为碳酸盐岩、变质岩及松散堆积物. 研究区内河流众多, 水资源丰富, 主要为长江及其支流, 构成了复杂的地表径流网络. 复杂的地理、地质、水文条件使得研究区滑坡地质灾害现象频发, 危害十分严重(李松林等, 2018).

2.2 数据源

基于地质灾害调查资料结合遥感影像以及 GIS 工具, 建立研究区滑坡空间数据库. 主要数据来源为: (1) 研究区 30 m 分辨率的数字高程模型(DEM)数据(<http://www.gscloud.cn>), 用来提取高程、坡度等地形信息; (2) 武汉地调中心提供的三峡库区构造纲要图, 用来提取构造信息; (3) 湖北省自然资源厅提供的研究区第二次全国土地调查数据, 用来获

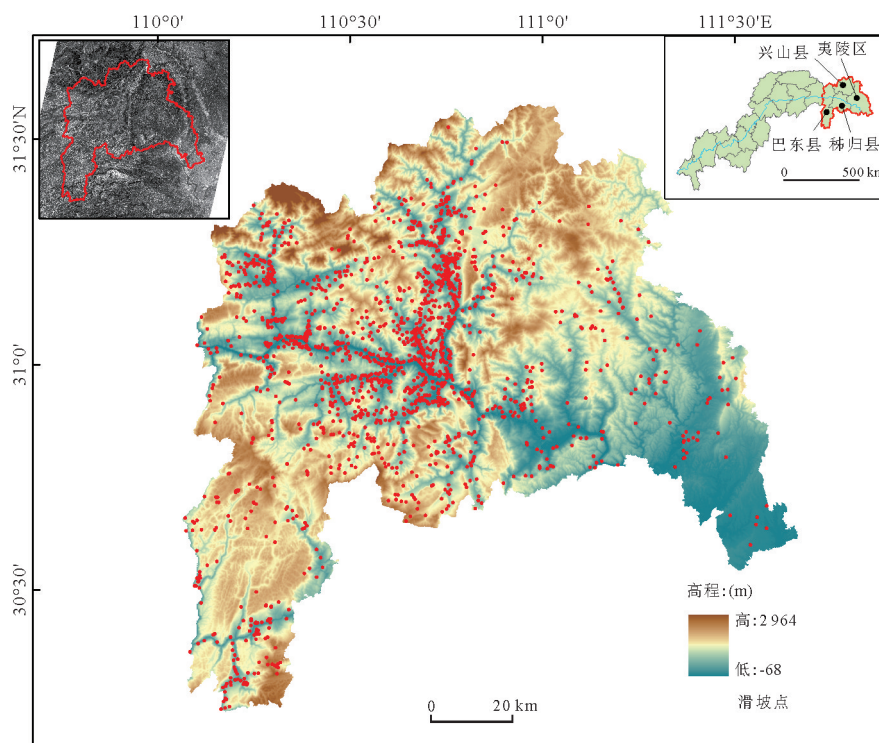


图1 研究区地理位置以及滑坡分布

Fig.1 Locations of study area and landslides

取区内土地利用类型、路网及水系等信息;(4)研究区 30 m 分辨率 Landsat8 OLI 影像(<http://www.gscloud.cn>)用来计算归一化植被指数(NDVI)。其中,影像数据由条代号/行编号分别为 125/38 与 125/39 的两景影像镶嵌获得,成像时间均为 2015 年 4 月 14 日,平均含云量均为 0.01(遥感影像缩略图见图 1);(5)武汉地调中心提供的三峡库区 1:65 万地质灾害分布图,用来提取工程岩组信息,并确定滑坡 1 778 处,滑坡类型以中小型为主,其中特大型滑坡 184 处、大型滑坡 245 处、中型滑坡 695 处、小型滑坡 654 处。介于本次研究的重点是大范围区域易发性评价,因此,本文并未区分考虑滑坡规模和类型。将滑坡位置的坐标以点的形式导入到 GIS 环境中以绘制实际滑坡分布图(图 1)。

3 研究方法

3.1 评价因子

滑坡的发生涉及到地理地质、气象水文、人类活动等众多因素,是各种外界条件和内在条件互相影响、协同作用下产生的结果,是一个复杂的非线性过程。因此分析影响滑坡的评价因子在滑坡空间预测中是非常关键的一步(冯杭建等,2016)。高程因为其影响坡体的势能变化、坡体的稳定性、温度气候分布、植被类型、降雨、甚至是人类活动等众多因素,成为间接的、不可或缺的地形因素(肖桐,2007)。坡度控制斜坡重力、压力等应力条件(刘丽娜等,2014)。斜坡结构反映了坡体的坡度、坡向以及岩层倾向的空间组合关系,不同类型的斜坡结构导致坡体失稳的机制和变形模式不同,该因子被认为是研究三峡库滑坡空间发育规律的关键性因子(李松林等,2018)。归一化植被指数是反映植被覆盖的一个重要指标,而植被根系作用可以加固边坡稳定性(蒋锐,2010)。边坡失稳的概率随岩性的不同而发生变化(向灵芝和崔鹏,2014),边坡岩石的坚硬强度、变形程度取决于岩石的岩性和结构特征。断裂构造产生的松散碎屑物质不仅为滑坡发生提供物质基础,其形成的裂隙和节理也有利于水文条件累积,侵蚀坡体(武尚等,2012)。河流的侵蚀是影响地质灾害的重要因素(聂娟等,2014)。近些年来人类活动对自然环境的破坏成为诱发地质灾害的主要因素之一(王丽丽,2016)。强烈的人类干预能够快速破坏坡体平衡状态造成坡体失稳发生滑坡等活动,而道路建设是人类活动的代名词。此外

不合理的土地利用方式也是加剧滑坡灾害发生的主要外部因素之一(郭慧娟等,2010)。

结合研究区特点、滑坡发生机制、可利用的数据源以及前人研究,筛选出高程、坡度、斜坡结构、土地利用类型、岩土体类型、归一化植被指数、路网距离、河网距离、断裂距离等 9 个因子分析研究区内的滑坡易发性。这些数据包含连续型因子和离散型因子,结合各因子分布规律利用等间距离散化将连续型因子分割为若干个区间,以统一数据类型(表 1)。特别提出的是,由于研究区内路网分布密集,将 $[0, 100]$ 区段划分为 $[0, 50]$ 以及 $[50, 100]$ 进行分析更为合理。本文使用频率比值(frequency ratio, FR)来衡量因子各区间的信息量(Adition *et al.*, 2018)。该比值可以表示为各因子类别所包含的滑坡数比例(percentage of landslide, PL)与对应类别所占面积比例(percentage of area, PA)之比。从表 1 中可以看出,200~400 m 的高程、 $20^{\circ}\sim 25^{\circ}$ 的坡度范围、水平坡、小于 -0.1 的 NDVI、路网距离小于 50 m、河网距离小于 500 m、断裂距离小于 1 000 m、居民地土地利用类型、松散岩土类岩土体类型分别在各所属因子中占据最高的频率比值。

本文的基本研究单元为栅格单元,是滑坡空间易发性分析中的一种常用模型单元,具有划分方案简单快捷、便于模型计算等优点。此外,考虑到数据获取来源不同导致了各因子尺度上的差异,为了统一数据尺度,将所有与滑坡地质灾害相关的特征图层划分为 12 626 025 个像素大小为 30×30 m (与遥感影像像素大小保持一致)的栅格单元为进一步分析做准备。

3.2 滑坡易发性评价

一般来说,滑坡的空间预测可以看作是一个二元分类过程(Bennett *et al.*, 2016),在建模过程中需要准备等量的负例数据集(非滑坡)和正例数据集(滑坡)(Kornejady *et al.*, 2017)。负例数据集通过 ArcGIS 软件随机产生。之后正例数据集与负例数据集均以 70/30 的比例划分为训练样本集以及测试样本集两部分。训练集用来训练模型的参数,然后在测试集上测试训练好的模型。在 R 软件中输入 9 个滑坡影响因子,使用随机森林构建滑坡预测模型。模型性能评估将从训练集以及测试集两方面进行,因为训练集衡量的是模型的拟合优度而测试集可以体现模型的泛化能力。至于性能度量指标,选取准确率、均方根误差、kappa 系

表 1 评价因子汇总表
Table 1 The summary of conditioning factors

高程(m)				坡度(°)				斜坡结构			
类别	PA(%)	PL(%)	FR	类别	PA(%)	PL(%)	FR	类别	PA(%)	PL(%)	FR
≤200	7.44	10.22	1.37	≤5	6.22	2.47	0.40	水平坡	0.40	0.52	1.30
200~400	9.21	30.96	3.36	5~10	11.05	6.49	0.59	顺向坡	18.33	18.50	1.01
400~600	11.84	20.05	1.69	10~15	15.21	14.59	0.96	顺斜向坡	17.83	18.27	1.02
600~800	13.75	14.13	1.03	15~20	17.34	22.11	1.28	横向坡	32.79	33.20	1.01
800~1000	15.28	11.72	0.77	20~25	16.15	23.09	1.43	逆斜向坡	15.56	14.47	0.93
1 000~1 200	15.60	7.01	0.45	25~30	12.69	16.94	1.34	逆向坡	15.09	15.05	1.00
1 200~1 400	13.37	3.73	0.28	30~35	9.04	8.50	0.94				
≥1 400	13.51	2.18	0.16	35~40	5.88	3.10	0.53				
				≥40	6.43	2.70	0.42				
路网距离(m)				NDVI				断裂距离(m)			
类别	PA(%)	PL(%)	FR	类别	PA(%)	PL(%)	FR	类别	PA(%)	PL(%)	FR
≤50	15.99	30.96	1.94	≤-0.1	2.12	3.27	1.54	≤1000	19.75	17.92	0.91
50~100	12.52	21.65	1.73	-0.1~0	1.85	2.81	1.52	1 000~2 000	16.07	13.10	0.81
100~200	17.96	22.63	1.26	0~0.1	4.53	6.66	1.47	2 000~3 000	12.68	10.45	0.82
200~300	11.97	9.88	0.83	0.1~0.2	10.27	11.60	1.13	3 000~4 000	9.63	8.62	0.89
300~400	8.36	5.05	0.60	0.2~0.3	19.00	17.69	0.93	4 000~5 000	7.74	6.72	0.87
400~500	6.11	3.22	0.53	0.3~0.4	27.02	23.95	0.89	5 000~6 000	6.59	8.27	1.26
500~600	4.66	1.84	0.39	0.4~0.5	25.24	23.43	0.93	6 000~7 000	5.71	9.25	1.62
600~700	3.60	1.03	0.29	≥0.5	9.96	10.57	1.06	7 000~8 000	4.85	5.51	1.14
700~800	2.91	1.21	0.41					8 000~9 000	4.10	4.77	1.16
≥800	15.91	2.53	0.16					≥9 000	12.88	15.39	1.19
河网距离(m)				土地利用类型				岩土体类型			
类别	PA(%)	PL(%)	FR	类别	PA(%)	PL(%)	FR	类别	PA(%)	PL(%)	FR
≤500	17.34	21.83	1.26	农地	14.38	17.98	1.25	松散岩土类	12.09	24.99	2.07
500~1 000	17.04	18.90	1.11	园地	6.68	22.80	3.41	碎屑岩类	21.63	38.94	1.80
1 000~1 500	15.72	14.88	0.95	居民地	2.97	8.90	3.00	碳酸盐岩类	52.40	29.47	0.56
1 500~2 000	13.33	9.13	0.69	林地	72.47	43.54	0.60	岩浆岩及变质岩类	13.88	6.61	0.48
2 000~2 500	10.30	8.39	0.81	草地	0.56	0.52	0.93				
2 500~3 000	7.79	6.32	0.81	其它	2.95	6.26	2.12				
3 000~3 500	5.65	6.84	1.21								
≥3 500	12.84	13.73	1.07								

数以及 AUC 这 4 个指标.前三者是基于统计的指标,分别用来衡量样本被正确分类的比例、预测值与真值之间的偏差以及模型的一致性检验. AUC 表示的是受试者工作特征(ROC)曲线下面积,由一系列不同假设下的特异度和灵敏度计算得到,它们是常用的且具有代表性的指标.

一旦滑坡评价模型建立后,它们将为整个研究区的每一个栅格单元计算出唯一的概率值,该值可以表示为滑坡易发性指数,介于 0~1 之间.易发性指数越大指示该地区越脆弱,越容易发生滑坡.

4 结果与分析

模型调参是一项关键且费时的步骤,一方面要保证建立的模型保持优良的性能,另一方面需要注意过拟合问题.随机森林包含两个重要的参数,分别为 ntree(森林中决策树的数量)以及 mtry(每次随机抽取的变量个数).根据随机森林的算法原理可知,该算法每次训练仅使用了 2/3 的数据,而剩下的数据(OOB)则可以用来计算当前迭代过程中的模型误差.换句话说,随机森林可以进行内部评估,且评估过程伴随着训练过程一起进行.通过设置不同的 ntree 以及 mtry,可以得到一系列 OOB 误差估计

(图 2). 本文 n_{tree} 的初始值设为 1 000, $mtry$ 的默认值为变量个数的均方根(即 3). 为了探讨 $mtry$ 对训练过程的影响, 本文小幅度改变 $mtry$ 取值, 将其取值范围设为 2~5. 图 2 中横坐标表示 n_{tree} , 黑色实线表示 OOB 总误差, 红色虚线表示正例(滑坡)的 OOB 误差, 绿色虚线表示负例(非滑坡)的 OOB 误差. 从图 2 中可以看出, $mtry$ 的取值影响随机森林的收敛速度, 同时也直接影响了随机森林的训练质量. 如当 $mtry$ 取值为 2 时, OOB 总误差很快在决策树数量为 100 时接近收敛, 然而正例与负例的 OOB 误差在训练开始阶段极难达到平衡; 当 $mtry$ 取值为

3 时, 随机森林在接近 420 颗决策树的情况下得到收敛; 当 $mtry$ 取值为 4 时, 随机森林在决策树数量为 180 时就得以收敛; 而当 $mtry$ 取值为 5 时, 训练的效果似乎并不理想, 正例与负例的 OOB 误差之间总是存在差距, 随机森林将更多的滑坡划分为非滑坡, 这样训练出的模型并不适用于滑坡地质灾害的空间预测任务. 尽管 $mtry$ 取值不同影响模型的收敛速度, 图 2 的结果均一致性指示, 增加单体决策树的数量在一定程度上有利于降低训练误差, 然而一味增加树的数量并不总是利于随机森林的预测精度. 综合考虑模型的复杂度以及 OOB 误差, 本文最终

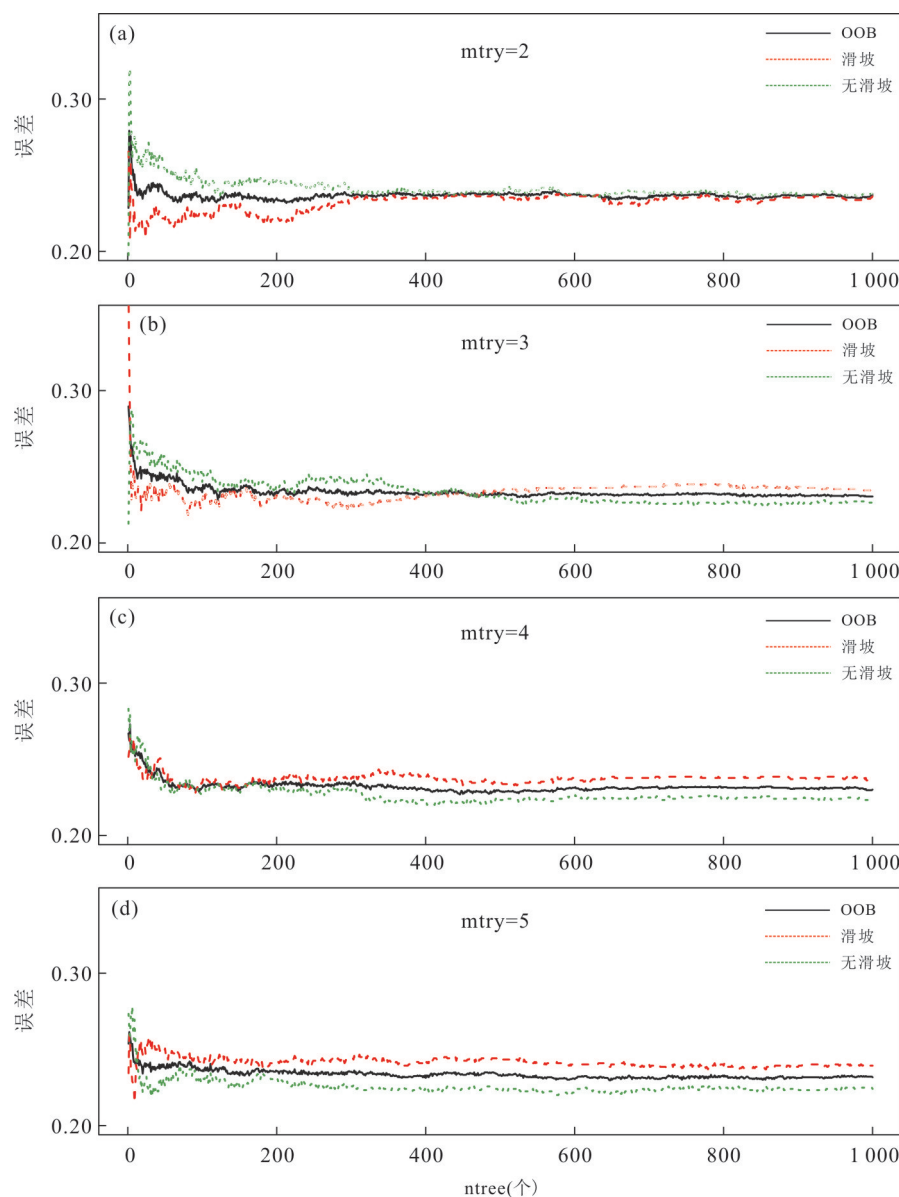


图 2 随机森林 OOB 误差曲线图

Fig.2 OOB error curves of random forest

a. $mtry=2$; b. $mtry=3$; c. $mtry=4$; d. $mtry=5$

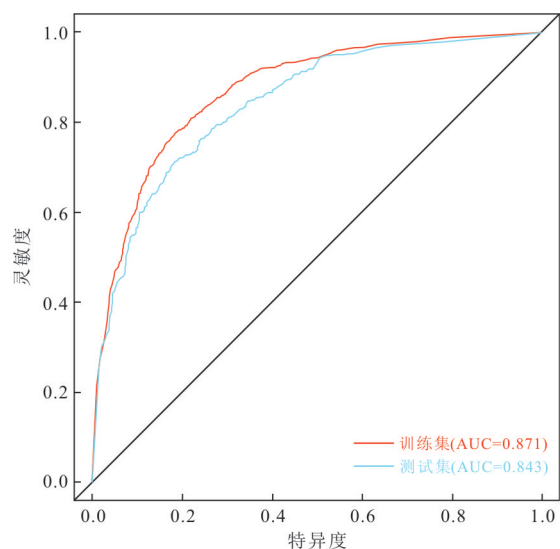


图3 随机森林ROC曲线

Fig.3 The ROC curve of random fores

表2 随机森林性能统计表

Table 2 Performance of random forest

性能度量指标	准确度	均方根误差	kappa系数
训练集	0.791	0.390	0.582
测试集	0.753	0.417	0.507

将ntree和mtry的值分别设为180和4。

模型性能方面(表2和图3),首先在训练集上,随机森林准确率、均方根误差、kappa系数以及AUC值分别为0.791、0.390、0.582、0.871。测试集上,准确率、均方根误差、kappa系数以及AUC值分别达到0.753、0.417、0.507、0.843。结果显示,随机森林表现出良好的预测性能,且没有出现拟合问题。从算法机理角度考虑,相比于单体分类与回归树,随机森林对样本和特征进行随机抽取,可以理解为对一个数据集的行和列均进行了采样。单棵树的分类能力可能较小,然而在随机产生大量的决策树后,由于这些决策树各不相同,其中的一些决策树可能对某些样本具有较高的识别能力因而适用于那部分样本,最终的表现是由所有的树构成的森林将适合处理各种样本,一个测试样本可以通过每一棵树的分类结果经统计后选择最可能的分类。随机森林以此种方式实现基分类器的多样化,获得更广阔的假设空间,因此体现出强大的分类能力。

将随机森林模型计算出的易发性指数导入到ArcGIS中,利用ArcGIS中几何间隔法将易发性指数划分为5个等级(极低、较低、中等、较高、极高)以绘制研究区易发性评价图(图4)。可以看出,易发性

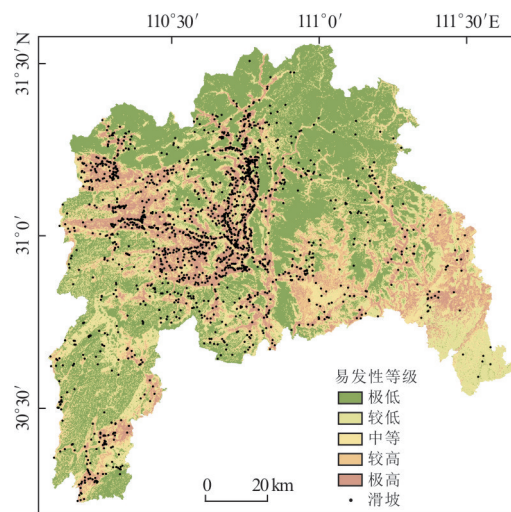


图4 研究区滑坡易发性评价图

Fig.4 Landslide susceptibility maps in the study area

级别极低和较低区域整体上以片状分布,而极高和较高易发性区域主要以线状集群分布,符合滑坡发育的特点。研究区内易发性极高的区域多分布在长江水系及其支流、路网密集、高程较低处,尤其集中在秭归县大部分地区,说明研究区内水系、人类活动和地形均对滑坡的孕育发生起到较大作用,这与评价因子各区间的频率比结果相一致。此外因子频率比的结果还表明,适合条件的坡度($20^{\circ}\sim 30^{\circ}$)、植被覆盖率低($NDVI < -0.1$)、松散的岩土结构对边坡失稳均具有促进作用。这些因子的协同作用同时也体现在夷陵区南部地区,从易发性评价结果来看,该区域实际发生的滑坡分布并不多。然而由于该区高程分布较低且近水系,适宜居民居住,一些居民区、园地和路网密集区多分布在此处,人类活动较为频繁。加之植被覆盖相对稀少,坡体失去植被根系保护作用,而该区分布的岩石结构类型较为松散,为坡体失稳提供了物理条件。各方面不利因素导致该区易发性级别较高。

模型拟合能力以及泛化能力评估结果指示随机森林具有较好的预测精度,然而这种性能度量不能反映出易发性指数的空间分布模式。因此,为了验证随机森林易发性评价结果的合理性,将实际分布的滑坡位置作为验证数据,统计各个易发性等级下滑坡数、滑坡百分比、栅格数、栅格百分比以及滑坡密度(表3)。结果显示,研究区大部分区域(60.31%)的易发性等级为较低或极低。而73.35%的滑坡发生在极高和较高易发区,该值可以考虑为满足空间分布验证精度要求。此外,滑坡密度作为

表 3 易发性评价灾害点分布统计表

Table 3 The summary of landslide locations in corresponding susceptibility classes

易发性等级	滑坡数	滑坡百分比 (%)	栅格数	栅格百分比 (%)	滑坡密度
极高	1 040	58.47	1 839 800	14.57	4.01
较高	265	14.88	1 475 600	11.69	1.27
中等	218	12.23	1 695 650	13.43	0.91
较低	125	7.00	2 232 150	17.68	0.40
极低	132	7.41	5 382 825	42.63	0.17

滑坡百分比与易发性等级百分比的比值,反映了单位面积内滑坡地质灾害的分布信息.随着易发性等级的升高,滑坡密度值相应增加,表明随机森林的易发性评价结果是合理的 (Pham *et al.*, 2016).

5 结论

本文基于多源数据、GIS 环境以及先进的数据挖掘技术,以三峡库区湖北段作为研究区,选取高程等 9 个因子建立滑坡易发性评价模型,模拟结果显示随机森林取得较好的预测结果,绘制的易发性评价图符合实际情况.

随机森林作为著名的集成学习算法,相比传统技术,具有对参与评价的变量限制小、无需考虑数据尺度和数据分布、计算效率高、取得高精度的同时超参数调试代价相比神经网络、深度学习等明显较低等优点.通过绘制 OOB 误差曲线进一步说明了随机森林易于训练的特性,通过提供可视化的方式,方便了解训练参数对于训练过程的影响,可以快速建立有效的随机森林模型.此外,抽样训练也使得随机森林能够在模型建立的同时进行内部评估,有利于缓解过拟合趋势.OOB 误差曲线结果表明,虽然集成决策树的机制可以产生强大的决策能力,然而一味的增加决策树数量并不总是有效,超过一定数量时效果甚至适得其反,这取决于 *mtry* 的取值.模型性能方面,通过性能度量指标,验证了随机森林在训练集上获得了满意的拟合能力,准确率、均方根误差、kappa 系数以及 AUC 值分别为 0.791、0.390、0.582、0.871.而在测试集上,随机森林的准确率、均方根误差、kappa 系数以及 AUC 值分别达到 0.753、0.417、0.507、0.843.研究区内滑坡易发性制图显示随机森林取得较为合理的分级结果,极低、较低、中等、较高、极高级别分别占到研究区总面积的 42.63%、17.68%、13.43%、11.69%、

14.57%.其中,73.35%的滑坡分布在极高和较高易发区内.性能评估以及易发性统计结果均表明随机森林是一种出色的算法,适用于滑坡易发性分析.而最终的易发性区划也指示,三峡库区湖北段内的巴东县北部、秭归县中部以及夷陵区南部需要重点防范,提前做好防灾减灾规划.

滑坡地质灾害的空间预测是一个复杂的非线性过程,提高模型的精度对于滑坡预测任务具有重要意义.探索评价因子的分布特点有助于理解滑坡发生的机制.下一步工作需要建立更完善的滑坡评价因子体系,定量化表达评价因子的重要程度,以便探索更合理可靠的滑坡预测模型.

References

Adition, A., Kubota, T., Shinohara, Y., 2018. Comparison of GIS - Based Landslide Susceptibility Models Using Frequency Ratio, Logistic Regression, and Artificial Neural Network in a Tertiary Region of Ambon, Indonesia. *Geomorphology*, 318: 101–111. <https://doi.org/10.1016/j.geomorph.2018.06.006>

Alcántara-Ayala, I., 2002. Geomorphology, Natural Hazards, Vulnerability and Prevention of Natural Disasters in Developing Countries. *Geomorphology*, 47(2/3/4): 107–124. [https://doi.org/10.1016/s0169-555x\(02\)00083-1](https://doi.org/10.1016/s0169-555x(02)00083-1)

Bennett, G. L., Miller, S. R., Roering, J. J., et al., 2016. Landslides, Threshold Slopes, and the Survival of Relict Terrain in the Wake of the Mendocino Triple Junction. *Geology*, 44(5): 363–366. <https://doi.org/10.1130/g37530.1>

Breiman, L., Friedman, J. H., Olshen, R. A., et al., 1984. Classification and Regression Trees. Chapman & Hall, New York.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45 (1): 5–32. <https://doi.org/10.1023/A:1010933404324>.

Feng, H.J., Zhou, A.G., Yu, J.J., et al., 2016. A Comparative Study on Plum-Rain-Triggered Landslide Susceptibility Assessment Models in West Zhejiang Province. *Earth Science*, 41(3):403–415 (in Chinese with English abstract).

Felicísimo, Á. M., Cuartero, A., Remondo, J., et al., 2013. Mapping Landslide Susceptibility with Logistic Regression, Multiple Adaptive Regression Splines, Classification and Regression Trees, and Maximum Entropy Methods: A Comparative Study. *Landslides*, 10(2): 175–189. <https://doi.org/10.1007/s10346-012-0320-1>

- Fenti, V., Silvano, S., Spagna, V., 1979. Methodological Proposal for an Engineering Geomorphological Map. Forecasting Rockfalls in the Alps. *Bulletin of the International Association of Engineering Geology*, 19(1): 134—138. <https://doi.org/10.1007/bf02600465>
- Guo, H.J., Tang, N.Q., Lin, J.B., 2010. Sensibility Analysis of Land-Use and Landslide Hazard Based on GIS in Xianyou County. *Journal of Fujian Agriculture and Forestry University (Natural Science Edition)*, 39(4):417—420 (in Chinese with English abstract).
- Guo, Z.Z., Yin, K.L., Fu, S., et al., 2018. Evaluation of Landslide Susceptibility Based on GIS and WOE-BP Model. *Earth Science*, 44(12): 4299—4312 (in Chinese with English abstract).
- Guo, Z.Z., Yin, K.L., Liu, Q.L., et al., 2019. Rainfall Warning of Creeping Landslide in Yunyang County of Three Gorges Reservoir Region Based on Displacement Ratio Model. *Earth Science*, 45(2): 672—684 (in Chinese with English abstract).
- Jiang, R., 2010. Research on Generation Mechanism and Service of Slope Disaster Warning Information in Kunming City(Dissertation). Kunming University of Science and Technology, Kunming (in Chinese with English abstract).
- Kornejady, A., Ownegh, M., Bahremand, A., 2017. Landslide Susceptibility Assessment Using Maximum Entropy Model with Two Different Data Sampling Methods. *CATENA*, 152: 144—162. <https://doi.org/10.1016/j.catena.2017.01.010>
- Li, S.L., Xu, Q., Tang, M.G., et al., 2018. Study on Spatial Distribution and Key Influencing Factors of Landslides in Three Gorges Reservoir Area. *Earth Science*, 45(1): 341—354 (in Chinese with English abstract).
- Li, Y.Y., Mei, H.B., Ren, X.J., et al., 2018. Geological Disaster Susceptibility Evaluation Based on Certainty Factor and Support Vector Machine. *Journal of Geo-Information Science*, 20(12): 1699—1709 (in Chinese with English abstract).
- Liu, L.N., Xu, C., Xu, X.W., et al., 2014. GIS-Based Landslide Hazard Evaluation Using AHP Method in the 2013 Lushan Earthquake Region. *Journal of Catastrophology*, 29(4): 183—191 (in Chinese with English abstract).
- Nie, J., Lian, J., Hu, Z.W., 2014. Spatial Variation of Landslides in Wenchuan Earthquake - Stricken Areas. *Geographical Research*, 33(2): 214—224 (in Chinese with English abstract).
- Shi, X.G., Xu, J.H., Jiang, H.J., et al., 2019. Slope Stability State Monitoring and Updating of the Outang Landslide, Three Gorges Area with Time Series InSAR Analysis. *Earth Science*, 44(12): 4284—4292 (in Chinese with English abstract).
- Pham, B. T., Pradhan, B., Tien Bui, D., et al., 2016. A Comparative Study of Different Machine Learning Methods for Landslide Susceptibility Assessment: A Case Study of Uttarakhand Area (India). *Environmental Modelling & Software*, 84: 240—250. <https://doi.org/10.1016/j.envsoft.2016.07.005>
- Song, Y. Q., Gong, J. H., Gao, S., et al., 2012. Susceptibility Assessment of Earthquake-Induced Landslides Using Bayesian Network: A Case Study in Beichuan, China. *Computers & Geosciences*, 42: 189—199. <https://doi.org/10.1016/j.cageo.2011.09.011>
- Tehrany, M. S., Pradhan, B., Jebur, M. N., 2013. Spatial Prediction of Flood Susceptible Areas Using Rule Based Decision Tree (DT) and a Novel Ensemble Bivariate and Multivariate Statistical Models in GIS. *Journal of Hydrology*, 504: 69—79. <https://doi.org/10.1016/j.jhydrol.2013.09.034>
- Wang, L.L., 2016. Feature Processing Method in Rainfall-induced Landslide Susceptibility Assessment (Dissertation). Zhejiang University, Hangzhou (in Chinese with English abstract).
- Wu, S., Li, L.W., Li, S.J., 2012. Analysis on the Influence of Coupling Relationship between Rock Mass Structural Plane and Water on Slope Stability. *Journal of Jingchu University of Technology*, 27(2): 33—36 (in Chinese with English abstract).
- Wu, X.L., Shen, S.Q., Niu, R.Q., 2016. Landslide Susceptibility Prediction Using GIS and PSO-SVM. *Geomatics and Information Science of Wuhan University*, 41(5): 665—671 (in Chinese with English abstract).
- Xiang, L.Z., Cui, P., 2014. Triggering Factors Susceptibility of Earthquake-Induced Collapses and Landslides in Wenchuan County. *Journal of Sichuan University (Engineering Science Edition)*, 42(5): 105—112 (in Chinese with English abstract).
- Xiao T., 2007. GIS-Based Research on Spatial Simulation of Landslide in Lanzhou City (Dissertation). Lanzhou University, Lanzhou (in Chinese with English abstract).
- Zare, M., Pourghasemi, H. R., Vafakhah, M., et al., 2013. Landslide Susceptibility Mapping at Vaz Watershed (Iran) Using an Artificial Neural Network Model: A Comparison between Multilayer Perceptron (MLP) and Radial Basic Function (RBF) Algorithms. *Arabian Journal of Geosciences*, 6(8): 2873—2888. <https://doi.org/>

10.1007/s12517-012-0610-x

附中文参考文献

- 冯杭建, 周爱国, 俞剑君, 等, 2016. 浙西梅雨滑坡易发性评价模型对比. 地球科学, 41(3): 403—415.
- 郭慧娟, 唐南奇, 林金宝, 2010. 基于 GIS 的仙游县土地利用与滑坡灾害敏感性分析. 福建农林大学学报(自然科学版), 39(4): 417—420.
- 郭子正, 殷坤龙, 付圣, 等, 2018. 基于 GIS 与 WOE-BP 模型的滑坡易发性评价. 地球科学, 44(12): 4299—4312.
- 郭子正, 殷坤龙, 刘庆丽, 等, 2019. 基于位移比模型的三峡库区云阳县域内蠕变型滑坡降雨预警. 地球科学 45(2): 672—684.
- 蒋锐, 2010. 昆明市斜坡灾害预警信息生成机理及服务研究(博士学位论文). 昆明: 昆明理工大学.
- 李松林, 许强, 汤明高, 等, 2018. 三峡库区滑坡空间发育规律及其关键影响因子. 地球科学, 45(1): 341—354.
- 李远远, 梅红波, 任晓杰, 等, 2018. 基于确定性系数和支持向量机的地质灾害易发性评价. 地球信息科学学报, 20(12): 1699—1709.
- 刘丽娜, 许冲, 徐锡伟, 等, 2014. GIS 支持下基于 AHP 方法的 2013 年芦山地震区滑坡危险性评价. 灾害学, 29(4): 183—191.
- 聂娟, 连健, 胡卓玮, 2014. 汶川地震灾区滑坡空间特征变化分析. 地理研究, 33(2): 214—224.
- 史绪国, 徐金虎, 蒋厚军, 等, 2019. 时序 InSAR 技术三峡库区藕塘滑坡稳定性监测与状态更新. 地球科学, 44(12): 4284—4292.
- 王丽丽, 2016. 降雨型滑坡地质灾害易发性评价中的特征处理方法. 杭州: 浙江大学.
- 武尚, 李利文, 李世佳, 2012. 岩体结构面与水耦合关系对边坡稳定性的影响分析. 荆楚理工学院学报, 27(2): 33—36.
- 武雪玲, 沈少青, 牛瑞卿, 2016. GIS 支持下应用 PSO-SVM 模型预测滑坡易发性. 武汉大学学报(信息科学版), 41(5): 665—671.
- 向灵芝, 崔鹏, 2014. 汶川县地震诱发崩滑灾害影响因素的敏感性分析. 四川大学学报(工程科学版), 42(5): 105—112.
- 肖桐, 2007. 基于 GIS 的兰州市滑坡空间模拟研究(硕士学位论文). 兰州: 兰州大学.