中国地质灾害与防治学报
THE CHINESE JOURNAL OF GEOLOGICAL HAZARD AND CONTROL
中国科技核心期刊
《中国科学引用文数据库》来源期刊
Caj-cd规范获奖期刊

样本选取对地质灾害易发性评价的影响

陈建平，辛亚波，王泽鹏，陈　伟，万长园，刘云艳，黄俊杰

Effect of sample selection on the susceptibility assessment of geological hazards: A case study in Liulin County, Shanxi Province

CHEN Jianping, XIN Yabo, WANG Zepeng, CHEN Wei, WAN Changyuan, LIU Yunyan, and HUANG Junjie

在线阅读 View online: https://doi.org/10.16031/j.cnki.issn.1003-8035.202210037

DOI: 10.16031/j.cnki.issn.1003-8035.202210037
陈建平, 辛亚波, 王泽鹏, 等. 样本选取对地质灾害易发性评价的影响——以山西柳林县为例[J]. 中国地质灾害与防治学报, 2024, 35(3): 152-162.
CHEN Jianping, XIN Yabo, WANG Zepeng, et al. Effect of sample selection on the susceptibility assessment of geological hazards: A case study in Liulin County, Shanxi Province[J]. The Chinese Journal of Geological Hazard and Control, 2024, 35(3): 152-162.

样本选取对地质灾害易发性评价的影响
——以山西柳林县为例

陈建平[1], 辛亚波[1], 王泽鹏[1], 陈 伟[2], 万长园[1], 刘云艳[1], 黄俊杰[1]
(1. 辽宁工程技术大学矿业学院, 辽宁 阜新 123000;
2. 辽宁工程技术大学环境学院, 辽宁 阜新 123000)

摘要: 非地质灾害样本的合理选取对地质灾害易发性预测准确度的提高具有重要意义。文章以柳林县为例, 选取适宜的影响因子, 基于GIS技术采用随机森林模型进行易发性评价。以地质灾害与非地质灾害比例为1∶1、1∶1.5、1∶3、1∶5、1∶10和非地质灾害点距已知灾害点100, 500, 800, 1 000 m为选取条件交叉结合共创建20组模型进行分析。结果表明: (1)通过误差指标、混淆矩阵和ROC曲线检验, 样本比例和距已知灾害点距离变化对地质灾害易发性评价结果有较大影响。随着样本比例变小, 距已知灾害点距离增加, 各模型平均绝对误差和均方根误差整体下降, 准确率整体上升。各模型ROC曲线下面积值均大于0.8, 均有较好的预测效果。当样本比例小于1∶3时, 距已知灾害点距离增加对模型误差和准确率影响较小, 变化趋于稳定。综合判断样本比例为1∶10、距已知灾害点1 000 m为最适合研究区模型。(2)高和极高易发区主要分布在中部及北部道路和河流两侧的地区, 是柳林县防灾减灾的重点区。(3)样本选取差异导致易发性结果不同主要是因为建模过程中随机森林模型对数据特征的采集及判断发生变化, 样本是否具有代表性发生变化。这些研究成果对当防灾减灾工作的实施具有重要意义。

关键词: 非地质灾害; GIS; 随机森林; 易发性; 误差; 混淆矩阵; ROC
中图分类号: P642.22 **文献标志码:** A **文章编号:** 1003-8035(2024)03-0152-11

Effect of sample selection on the susceptibility assessment of geological hazards: A case study in Liulin County, Shanxi Province

CHEN Jianping[1], XIN Yabo[1], WANG Zepeng[1], CHEN Wei[2], WAN Changyuan[1], LIU Yunyan[1], HUANG Junjie[1]
(1. *College of Mining*, *Liaoning Technical University*, *FuXin*, *Liaoning* 123000, *China*;
2. *College of Environment*, *Liaoning Technical University*, *FuXin*, *Liaoning* 123000, *China*)

Abstract: The rational selection of non-geological hazard samples is of great significance to improve the accuracy of geological hazard susceptibility prediction. This study uses Liulin County as a case study, where appropriate impact factors were selected, and the random forest (RF) model was employed for susceptibility assessment based on GIS technology. A total of twenty sets of models were created by varying the ratio of geological hazard to non-geological hazard points (1∶1, 1∶1.5, 1∶3, 1∶5 and

收稿日期: 2022-10-25; **修订日期:** 2023-02-05 **投稿网址:** https://www.zgdzzhyfzxb.com/
基金项目: 国家自然科学基金项目(51604140)
第一作者: 陈建平(1971—), 男, 山西保德人, 地质资源与地质工程专业, 博士, 副教授, 主要从事工程地质、水文地质等方向研究。E-mail: 13804181164@139.com
通讯作者: 王泽鹏(1997—), 男, 山西长治人, 资源与环境专业, 硕士研究生, 主要从事地质工程和地理信息系统方向研究。E-mail: 449139098@qq.com

1∶10) and the distance from non-geological hazard points to known hazard points (100,500,800,1 000 m). The results demonstrate that: (1) Through error index, confusion matrix, and ROC curve tests, the sample proportion and distance from the known hazard point significantly influenced the geological hazard susceptibility evaluation. As the sample proportion decreased and the distance from known hazard points increased, the overall *MAE* and *RMSE* of the models decreased, while the overall ACC increased. All models achieved *AUC* value greater than 0.8, indicating excellent predictive performance. When the sample proportion was less than 1∶3, the increasing distance from the known hazard points on model error and accuracy became less pronounced, stabilizing the results. The most suitable model for the study area was found to have a sample ratio of 1∶10 and a distance of 1 000 m from known hazard points. (2) High and very high susceptibility areas were primarily located in the central and northern regions, adjacent to roads and rivers, making them key areas for hazard prevention and reduction in Liulin County. (3) Differences in sample selection led to varying susceptibility results mainly due to changes in the RF model's data feature collection and judgment during the modeling process, as well as the representativeness of the samples. These research findings hold significant implications for the implementation of hazard prevention and reduction measures.

Keywords: non-geological hazard；GIS；random forest；susceptibility；error；confusion matrix；ROC

0　引言

近年来，地质灾害频繁发生[1]，严重影响了人类生命财产安全、社会稳定和可持续发展。因此，对地质灾害进行易发性分析和评价，提高预测准确度，明确地质灾害易发区，对当地地质灾害的防治具有十分重要的意义[2－3]。

世界各地的研究人员已将各种方法应用于地质灾害易发性评价中，这些方法可以是定性的，也可以是定量的，定量方法涉及数值分析和计算[4]。常用到的模型有层次分析法（analytic hierarchy process, AHP）[5]、模糊逻辑（Fuzzy Logic, FL）[6]、信息量（Information Value, IV）[7]、证据权（Weight of Evidence, WOE）[8]、频率比（frequency response, FR）[9]、逻辑回归（logistic regression, LR）[10]、人工神经网络（artificial neural network, ANN）[11]、支持向量机（support vector machine, SVM）[12]和随机森林（random forests, RF）[13]等。其中机器学习模型因处理各因子间非线性关系较强被广泛应用于地质灾害易发性分析中。FR 属于机器学习模型，Ge 等[14]的研究表明 RF 模型在易用性、稳定性和时间花费等方面具有较大的优势，且预测精度普遍较高。RF 模型最明显的优点是重复随机选取样本进行建模，最终在众多模型中选取最优模型，该方法不易出现过拟合现象，许多研究也表明 RF 模型的预测效果要高于其它模型。吉日伍呷等[15]采用逻辑回归、K 近邻、朴素贝叶斯和随机森林算法以鲁甸地震为例进行易发性评价，结果表明 RF 模型要优于其他三种。李坤等[16]采用 RF 和 SVM 模型对东川泥石流进行易发性评价，结果表明 RF 模型要优于 SVM。邱维蓉等[17]通过 RF、SVM、LR 和 BP 神经网络对甘肃省灵台县进行滑坡易发性评价，经 ROC 曲线验证 RF 模型预测精度最高。RF 模型在建模过程中涉及到正负样本比例的设定，但以往的研究中对地质灾害样本比例的研究较少。

在地质灾害易发性研究中，非地质灾害样本的合理选择极为重要。目前，对于非地质灾害样本的选择无统一的规范[18]，不同学者对于其选取比例存在明显的差异。地质灾害与非地质灾害比例（后文称样本比例）通常有 1∶1、1∶1.5、1∶3、1∶5 和 1∶10[19－20]，距已知灾害点距离（后文称影响距离）通常有 100, 500, 800, 1 000 m[21]。如果选取不当的条件进行易发性分析，会直接影响研究结果的准确性，不利于精准进行防灾减灾工作[22]。因此对样本比例和影响距离进行地质灾害易发性研究显得十分重要。

综上所述，以柳林县（山西）为研究区。通过 GIS 技术，基于 RF 模型对样本比例按 1∶1、1∶1.5、1∶3、1∶5 和 1∶10 划分，影响距离按 100, 500, 800, 1 000 m 划分并交叉组合进行易发性分析（当选取更小的样本比例，即样本更多时，非灾害点样本的选取将超出研究区，因此 1∶10 为研究区比例选取的极限）。判断最适宜的非地质灾害选取条件，对不同条件的易发性分区图进行分析，评价模型产生差异的原因，明确研究区灾害防治重点区域，为当地进行灾害治理和防治提供科学依据。

1　研究区概况

柳林县地处吕梁山西麓，黄河东岸，是山西的西

大门，离石区、临县、中阳县和石楼县围绕其四周（图 1）。辖 8 镇、7 乡、197 个行政村，总人口 34.6 万（截至 2022 年 2 月 15 日）。该县属暖温带大陆性季风气候，平均降水量为 456.3mm，历年降雨量极值为 632 mm 和 37.44 mm。柳林县属西北黄土高原的丘陵沟壑区，其海拔高度由东向西递减。地质构造运动使吕梁山构造隆起，岩层西倾，黄河河道下切，形成了本县境内的基岩东北高西南低。地表覆盖的第四纪黄土层，久经雨水侵蚀剥蚀，被逐渐切割成梁峁起伏、沟壑纵横、山丘交错、支离破碎的复杂地貌单元。根据柳林县地质灾害调查与区划结果显示，结合中国科学院资源环境科学与数据中心共查明已发生的地质灾害及隐患点 151 处（包括崩塌、滑坡、泥石流和不稳定斜坡），地质灾害及隐患点分布如图 1 所示。基于研究区域范围，创建柳林县栅格数据集。

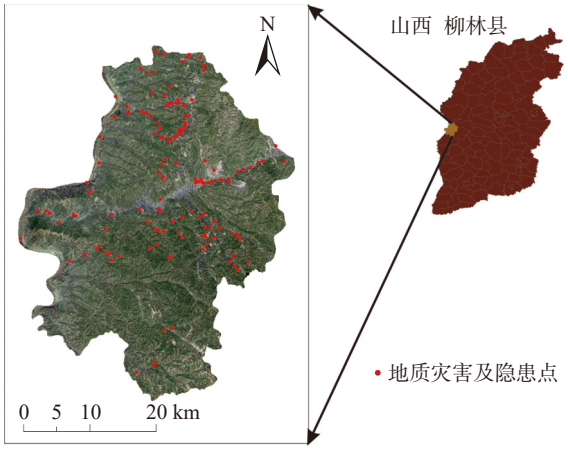


图 1 研究区位置

Fig. 1 Location of the research area

2 研究方法

2.1 随机森林（RF）

随机森林最初由 Breiman[23]开发，是一种把多个决策树（DT）结合在一起进行分类和预测的集合模型。随机森林算法包括以下步骤：①获取原始训练数据，并多次重新采样；②在每次重新取样时，选择一组随机特征；③给定重新采样和一组随机特征，估计决策树；④聚集该组估计的决策树，以便得到单个决策树。不同于其它机器学习模型，随机森林样本、特征的随机抽样减少了分类过程中对数据噪声和异常值的敏感性，有效地避免了过度拟合。随机森林特性之一是可以给出易发性评价因子的重要性，基尼系数的降低用于计算每个因子对易发性分类结果的重要程度，公式如下：

$$P_r = \frac{\sum_{i=1}^{k}\sum_{j=1}^{t} D_{Grij}}{\sum_{r=1}^{m}\sum_{i=1}^{k}\sum_{j=1}^{t} D_{Grij}} \tag{1}$$

式中：P_r——第 r 个评价因子在所有评价因子中的重要程度；

m、k、t——评价因子总数、分类树棵数和单棵树节点数；

D_{Grij}——第 i 个评价因子在第 r 棵树的第 j 个节点的基尼系数减少值。

2.2 模型验证方法

2.2.1 误差统计指标

平均绝对误差（*MAE*）可以避免误差相互抵消的问题，因而可准确反映实际预测误差的大小。均方根误差（*RMSE*）是预测值与真实值差的平方与观测次数 n 比值的平方根，可以用来衡量观测值同真值之间的偏差[24]。

$$MAE = \frac{\sum_{i=1}^{n}|x_i - y_i|}{n} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \tag{3}$$

式中：n——样本总量；

x——真实值；

y——预测值。

2.2.2 混淆矩阵

由于地质灾害与非地质灾害样本的极不平衡性，仅用统计方法判断模型预测精度适用性较差[25]。因此本文结合混淆矩阵对 RF 模型进行精度评估，其中常用的是准确率（*ACC*），比率越高，模型准确率越高。

$$ACC = \frac{TP + TN}{P + N} \times 100\% \tag{4}$$

式中：*TP*——真阳性；

TN——真阴性；

P——所有阳性样本；

N——所有阴性样本。

2.2.3 ROC 曲线

以往的研究[26－27]多采用工作特征曲线（ROC）及曲线下面积（*AUC*）来综合检验和评价模型效果。ROC 曲线反映了真阳性率（*TPR*）和假阳性率（*FPR*）之间的关系。本次研究通过 python 语言，可得到曲线图和 *AUC* 值。

$$TPR = \frac{TP}{TP+FN} \quad (5)$$

$$FPR = \frac{FP}{FP+TN} \quad (6)$$

式中：*FN*——假阴性；

FP——假阳性。

AUC 取值范围为 0.5 ~ 1，值越高表明模型精度越高。

以往的研究[26]多采用工作特征曲线（*ROC*）及曲线下面积（*AUC*）来综合检验和评价模型效果。*ROC* 曲线反映了“*TPR*”和“*FPR*”之间的关系。本次研究通过 python 语言，可得到曲线图和 *AUC* 值。

2.3　评价因子

地质灾害易发性分析的一个重要阶段是确定影响研究区灾害发生的致灾因子。Vakhshoori 等[27]指出，没有关于致灾因子选择的标准规则，可根据研究区域的规模、地形地质环境条件、研究区域内的地质灾害发生机制和数据可用性等因素来选择。考虑到这些标准，根据研究区地质灾害发育条件，结合以往研究者对本地区的经验[28]，本研究共使用 9 个影响因子，即高程、坡度、曲率、岩性、距道路距离、距水系距离、降雨、归一化植被指数和地形湿度指数，并基于 ArcGIS 进行分级（图 2）。

通过地理空间数据云网站（https://www.gscloud.cn）获取柳林县 30 m×30 m 高程（DEM）数据（可生成坡度、曲率和水系数据，坡度可生成地形湿度指数数据）、道路矢量数据和遥感数据（可生成归一化植被覆盖指数）；通过中国科学院资源环境科学与数据中心获取历年的平均降雨数据；通过吕梁市 1∶100 000 地质图获取岩性数据。

研究区影响因子特征具体如下：（1）高程变化是一个地区地形地貌最直接的表现，频繁被用在地质灾害易发性分析中。研究区高程范围在 597 ~ 1 467 m，地形呈东高西低地貌。在 ArcGIS 中按自然间断法分为 7 级，见图 2（a），地质灾害在 840 ~ 922 m 分布最多，约占总数的 42.38%。（2）坡度是边坡稳定性的关键因素之一，也是地质灾害易发性分析中最常用的参数之一。研究区坡度范围在 0 ~ 56°，按自然间断法分为 7 级，见图 2（b），地质灾害在 0 ~ 17°集中分布，约占总数的 78.81%。（3）曲率是与滑坡发生相关的因素之一，代表了斜坡形状和地形形态。研究区曲率分为 4 级，见图 2（c），地质灾害在 −2 ~ 0 之间分布最多，约占总数的 66.89%。（4）岩性决定了岩体的物理和化学性质，是地质灾害形成和演化的重要因素。根据岩土的力学和其它性质，将研究区分为 4 类，见图 2（d），其中地质灾害在黄土中分布最多，约占总数的 57.62%，在岩浆岩类中无分布。（5）人类工程活动因素反映了人类对自然环境影响的范围和强度，频繁的干扰可能会导致边坡失稳。地质灾害大多发生在人口密度较高或距离公路 1 km 以内，随着距离的增加，地质灾害数量迅速减少。距道路距离按自然间断法分为 7 级，见图 2（e），其中地质灾害在距道路 0 ~ 437 m 分布最多，约占总数的 78.81%。（6）水系对边坡表层岩石和土壤的侵蚀是地质灾害发生的重要原因之一。在 ArcGIS 中进行距离分析并分为 7 级，见图 2（f），其中距水系 391 m 内的地质灾害的分布频率最高，约占总数的 59.6%。（7）降雨是诱发地质灾害发生的最重要的因素之一，降雨会降低岩石和土壤的物理性质，使边坡的稳定性降低。通常降雨量越大灾害发生频率越高。通过研究区年平均降雨数据，基于 ArcGIS 进行插值分析并按自然间断法分为 7 级，见图 2（g），地质灾害在 470 ~ 478 mm 分布最多，约占总数的 42.38%。植被通过吸收水分，加固岩土体等方式提高边坡的稳定性，通常来讲，植被覆盖率越高边坡稳定性越强。（8）归一化植被覆盖指数是地表植被疏密程度的数字化体现，取值范围在 −1 ~ 1，值越接近 1 说明该地植被覆盖越茂盛。将遥感影像分析处理得到植被覆盖指数。按自然间断法分为 7 级，见图 2（h），地质灾害在 −0.17 ~ 0.02 分布最多，约占总数的 60.93%。（9）地形湿度指数是区域地形对径流流向和积水影响的一个指标，它表明了土壤水分的变化。研究区 *TWI* 分为 7 级，见图 2（i），地质灾害在 8.1 ~ 10 分布最多，约占总数的 48.34%。

3　结果

3.1　模型构建

将交叉组合好的样本和研究区栅格数据基于 ArcGIS 提取高程、坡度、岩性等属性值并转换为 Excel 表格，将每组样本的属性值作为输入变量带入 RF 模型进行建模，每组样本按 7∶3[29]划分为训练集和测试集。将研究区栅格数据转为 Excel 表依次带入 20 组训练好的模型中，最终可输出每个栅格内地质灾害易发性的概率。通过 ArcGIS 插值分析，按自然间断法将研究区易发性分为 5 级，即极低、低、中、高和极高，得到研究区不同样本比例和距离灾害点不同距离的地质灾害易发性分区图（图 3）。

3.2　模型检验

3.2.1　误差统计指标

平均绝对误差（*MAE*）表示数据集中实际值和预测值之间的绝对差异的平均值，它衡量的是数据集中残差的平均值。均方根误差（*RMSE*）是均方误差（*MSE*）的平

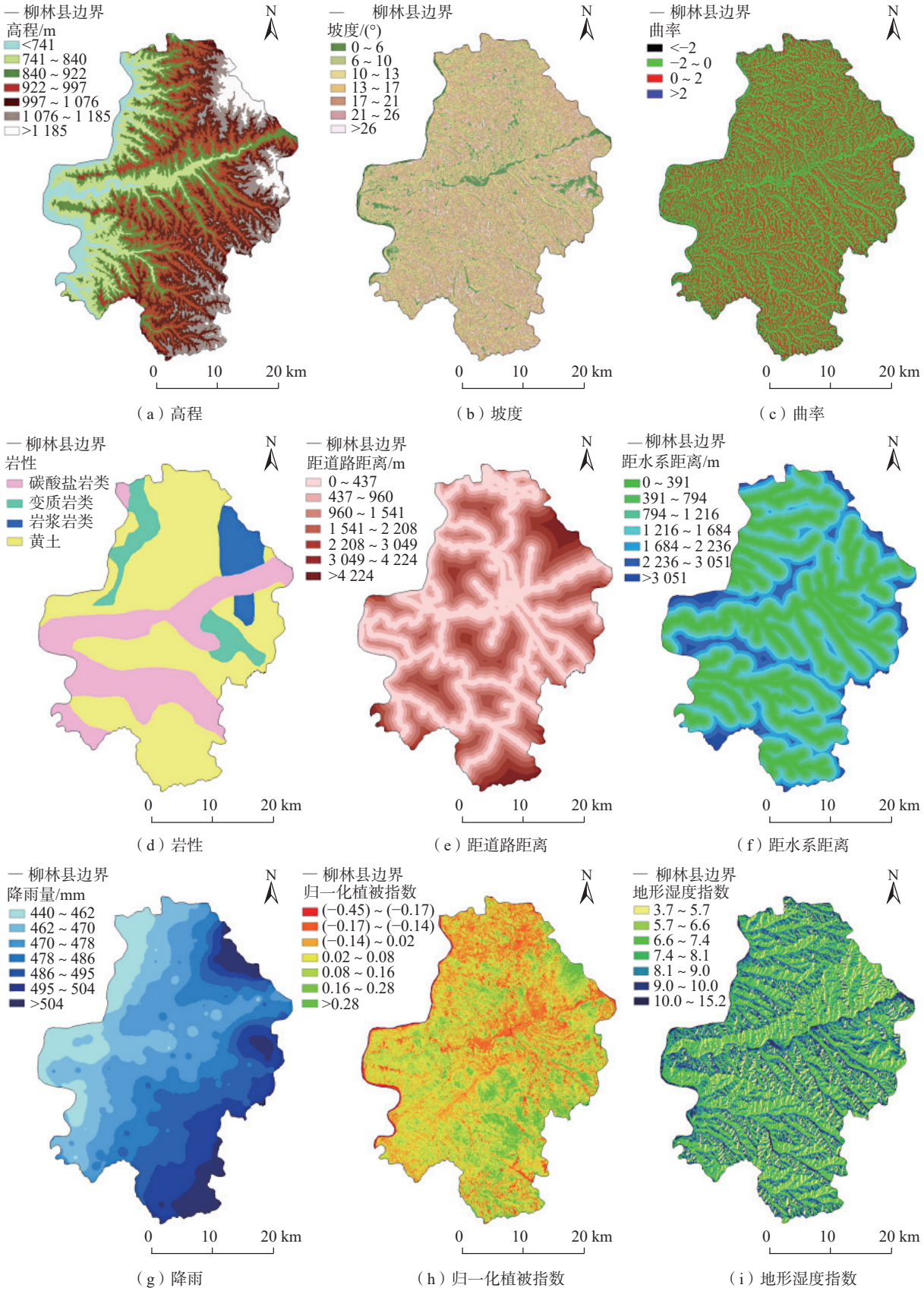


图 2　因子分级图

Fig. 2　Factor grading diagram

图 3　易发性分区图

Fig. 3　Susceptibility zoning maps

方根，它衡量的是残差的标准偏差。*RMSE* 广泛用于评估回归模型与其他随机模型的性能。*MAE* 和 *RMSE* 的值越低，意味着回归模型的准确性越高。根据公式(2)(3)，基于 Python 计算各模型 *MAE* 和 *RMSE* 值，计算结果如表 1 所示。

3.2.2　混淆矩阵

混淆矩阵用于观察模型在各个类别上的表现，可以计算模型对应各个类别的准确率，使得类别更有区分性。基于 Python 计算各比例模型的 TP、TN、FP 和 FN（表 2），根据式(4)，各模型 *ACC* 值被计算（表 3）。

3.2.3　ROC 曲线

ROC 曲线简单、直观，是对模型预测能力评价的一种典型方法。ROC 曲线下面积 *AUC* 值大于 0.8 时，模型预测效果较好，大于 0.9 时则表示预测效果极好。通过图 4 可知，各比例模型的 *AUC* 值都高于 0.8，皆有较好的预测性能，其中 1∶10(距已知灾害点 100 m)模型 *AUC* 值最高，为 0.91，模型预测能力最强；1∶1(距已知灾害点 800 m)模型 *AUC* 值最低，为 0.827，较其它比例模型预测能力最差。

4　讨论

4.1　模型准确率评价

根据表 1 和图 5，随着样本比例的变小，非地质灾害样本的增加，*MAE* 值在样本比例为 1∶1.5 和 1∶3 略

表 1 *MAE* 和 *RMSE* 值

Table 1 *MAE* and *RMSE* values

	距离	误差统计指标	1∶1	1∶1.5	1∶3	1∶5	1∶10
计算数据	距已知灾害点100 m	*MAE*	0.279	0.285	0.275	0.196	0.136
		RMSE	0.373	0.367	0.408	0.315	0.267
	距已知灾害点500 m	*MAE*	0.332	0.304	0.270	0.205	0.130
		RMSE	0.410	0.401	0.393	0.322	0.260
	距已知灾害点800 m	*MAE*	0.323	0.279	0.264	0.181	0.135
		RMSE	0.414	0.361	0.385	0.280	0.269
	距已知灾害点1 000 m	*MAE*	0.281	0.254	0.267	0.188	0.129
		RMSE	0.368	0.337	0.388	0.302	0.258

表 2 混淆矩阵

Table 2 Summary table of confusion matrix

地质灾害与非地质灾害样本比例1∶1							
距已知灾害点100 m		真实值/个		距已知灾害点500 m		真实值/个	
		地质灾害	非地质灾害			地质灾害	非地质灾害
预测值	地质灾害	34	7	预测值	地质灾害	26	15
	非地质灾害	11	39		非地质灾害	11	39
距已知灾害点800 m		真实值/个		距已知灾害点1 000 m		真实值/个	
		地质灾害	非地质灾害			地质灾害	非地质灾害
预测值	地质灾害	31	10	预测值	地质灾害	35	6
	非地质灾害	13	37		非地质灾害	10	40
.....							
地质灾害与非地质灾害样本比例1∶10							
距已知灾害点800 m		真实值/个		距已知灾害点1 000 m		真实值/个	
		地质灾害	非地质灾害			地质灾害	非地质灾害
预测值	地质灾害	435	13	预测值	地质灾害	442	6
	非地质灾害	38	13		非地质灾害	41	10

表 3 *ACC* 值表

Table 3 Summary table of *ACC* values

距离	1∶1 占比/%	1∶1.5 占比/%	1∶3 占比/%	1∶5 占比/%	1∶10 占比/%
距已知灾害点100 m	80.2	79.8	78.0	86.0	90.1
距已知灾害点500 m	71.4	77.2	78.0	86.0	91.4
距已知灾害点800 m	74.7	80.7	79.7	87.9	89.8
距已知灾害点1 000 m	82.4	82.5	80.8	89.3	92.3

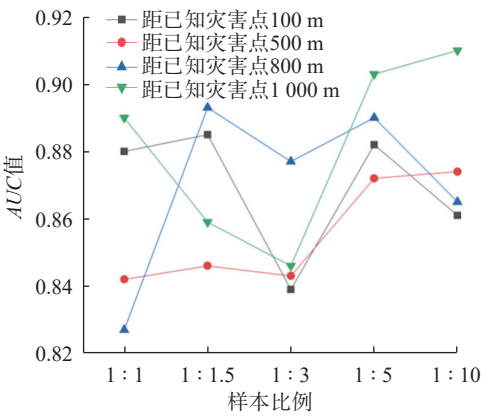
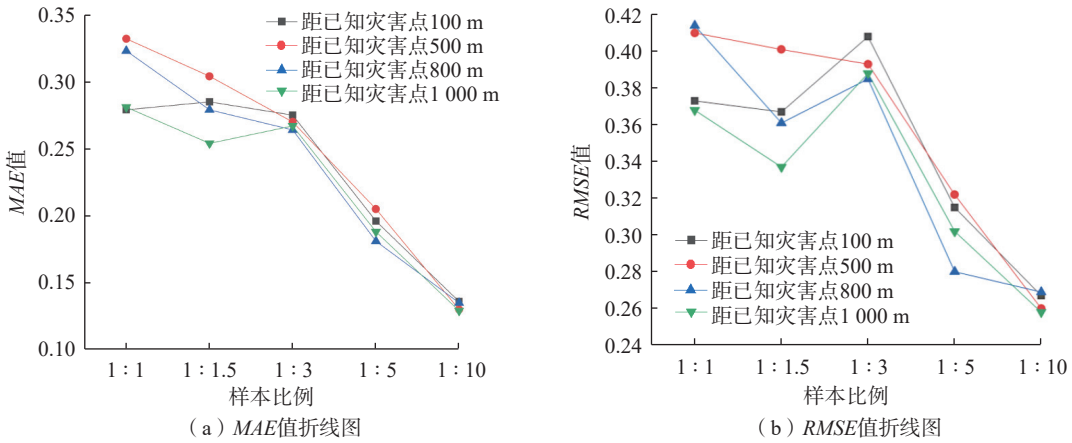


图 4 *AUC* 值折线图

Fig. 4 Line chart of *AUC* values

有上升，*RMSE* 值在样本比例为 1∶3 时有明显上升，但是 *MAE* 和 *RMSE* 值总体呈下降的趋势，均在样本比例为 1∶10 时达到最小值，此时误差最小。随着影响距离的增加，*MAE* 和 *RMSE* 值波动变化，整体先升高再降低。当样本比例为 1∶1 和 1∶1.5 时，误差值波动较大，差值最大分别为 0.051 和 0.064。最大差值随着样本比例的变小趋于稳定，在样本比例为 1∶10 时差值最小，*MAE* 和 *RMSE* 差值分别 0.007 为 0.011。通过误差指标分析，样本比例变化对易发性模型的准确度有极大的影响，当样本比例较大时，影响距离对易发性模型的准确度影响较为明显，当样本比例较小时，影响距离对易发性模型的准确度影响较小。整体来看，当样本比例为 1∶10，影响距离为 1 000 m 时，进行易发性建模准确度最高，误差最小(表 2)。

根据图 6 和表 3，随着样本比例的变小，非地质灾害样本的增加，样本比例为 1∶1.5 或 1∶3 时，*ACC* 有下降趋势，下降程度极小，总体上 *ACC* 呈上升趋势，均



（a）*MAE*值折线图　　（b）*RMSE*值折线图

图 5　*MAE* 值折线图和 *RMSE* 值折线图

Fig. 5　Line chart of *AME* value and *RMSE* value

在样本比例为 1∶10 时达到最高。随着影响距离的增加，*ACC* 波动变化，但最终都是以上升趋势结束。当样本比例为 1∶1 和 1∶1.5 时，*ACC* 波动较大，最大差值分别为 11% 和 5.3%，最大差值随着样本比例的变小趋于稳定，在样本比例为 1∶10 时差值最小，为 0.9%。整体来看，在影响距离 1 000 m 外创建非地质灾害样本构建模型准确率最高。通过 *ACC* 分析，样本比例变化对易发性模型的准确度有极大的影响，样本比例越小。模型准确率越高。当样本比例较大时，影响距离对易发性模型的准确影响较为明显，当样本比例较小时，影响距离对易发性模型的准确度影响较小。整体来看，当样本比例为 1∶10，距已知灾害点 1 000 m 时，进行易发性建模准确率最高。

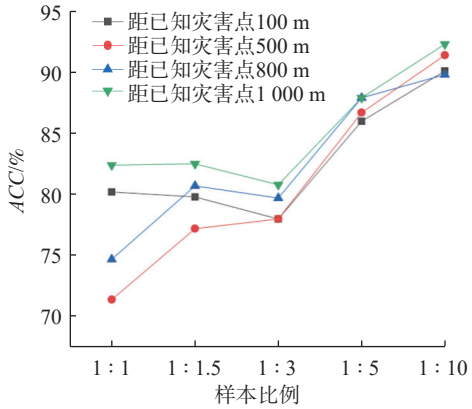


图 6　*ACC* 折线图

Fig. 6　*ACC* line chart

根据表 3 和图 7，随着样本比例的变小，非地质灾害样本的增加，*AUC* 值整体呈先下降再上升的趋势，在样本比例为 1∶8 或 1∶10 时达到最大值。随着影响距离的增加，*AUC* 值整体呈先下降再上升的趋势，在距已知灾害点 800 m 或 1 000 m 时达到最大值。通过 ROC 曲线分析，样本比例和距已知灾害点变化对易发性模型的准确率影响均较大。

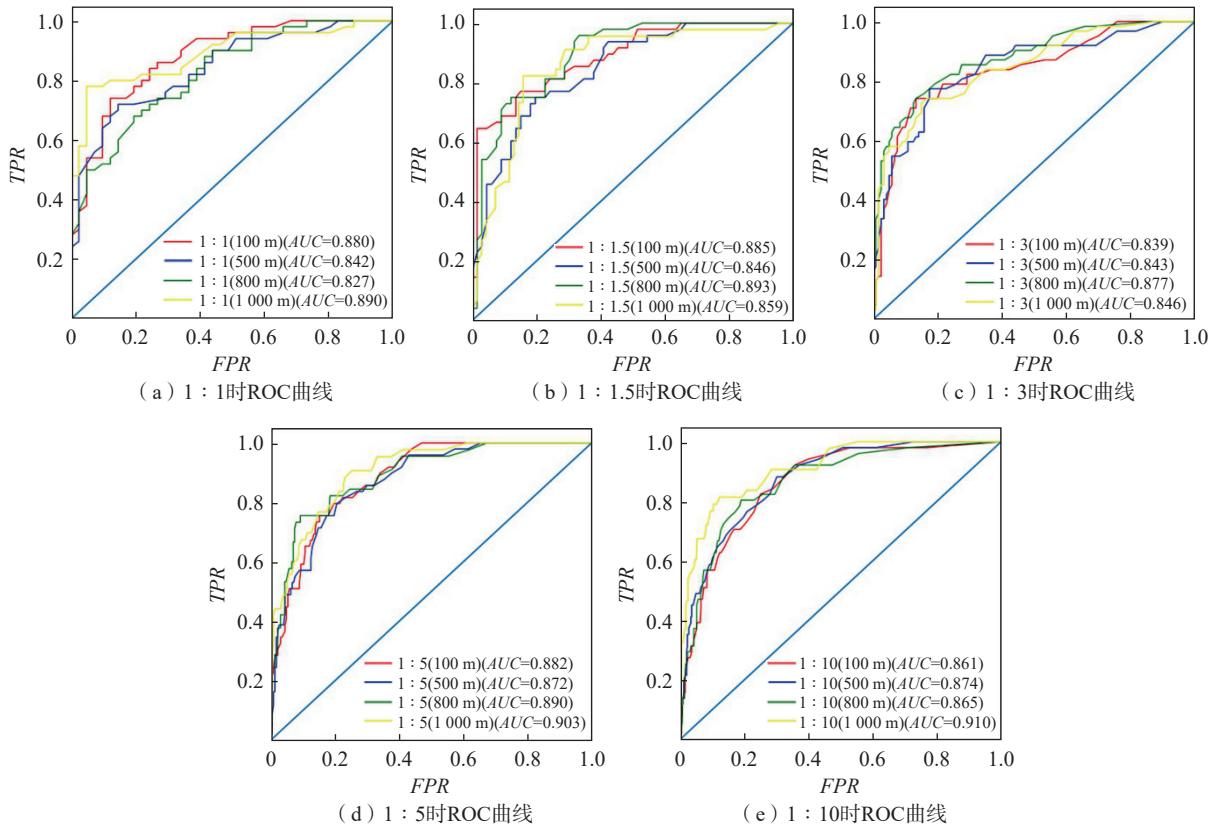
通过模型检验得知，样本比例和影响距离变化对地质灾害易发性评价结果有直接影响，合适的样本选取条件可以提高预测准确度。通过对不同选取条件进行对比，当样本比例为 1∶10 时，*MAE* 和 *RMSE* 平均比其它比例低 0.13 和 0.1；*ACC* 和 *AUC* 值平均比其它模型高 11% 和 0.01。影响距离为 1 000 m 时，*MAE* 和 *RMSE* 平均比其它比例均低 0.02；*ACC* 和 *AUC* 值平均比其它模型高 3.4% 和 0.01。因此正负样本比例为 1∶10、影响距离为 1 000 m 的模型预测准度最高，是最适合研究区的样本选取条件。同时预测准确度也高于以往的研究(图 7)。

4.2　易发性分区评价

从图 3 可以看出，随着样本比例的变小，高和极高易发区明显的减少，低和极低易发区明显地增加，反映出样本比例变化对易发性分区结果有明显的影响。随着影响距离的增加，中及中以上易发区会逐渐变多，也印证了 4.1 节所述，影响距离变化对易发性建模有一定影响。各模型易发性分区空间分布整体保持一致，极低和低易发区主要分布在东北部、东南部和西南部；中易发区分布在高和低易发区之间的区域；高和极高易发区主要分布在中部和北部沿道路和河流两侧的地区，这是柳林县防灾减灾的关键区。

4.3　易发性评价结果差异分析

根据上文，各模型的易发性分区、*MAE*、*RMSE*、*ACC* 和 ROC 曲线结果均有明显的差异，表明样本比例和影响距离选取的差异会直接影响易发性模型评价结果。分析其原因，从模型本身来讲，RF 模型是众多决策树分类和综合评价的结果，当训练集样本数量变化时，每个



（a）1 : 1时ROC曲线　（b）1 : 1.5时ROC曲线　（c）1 : 3时ROC曲线

（d）1 : 5时ROC曲线　（e）1 : 10时ROC曲线

图 7　ROC 曲线

Fig. 7　ROC curve

决策树选取特征产生的结果就会变化，当众多树聚类时，模型的判断则会产生差异，即每个栅格内发生地质灾害的概率会发生相应变化。当比例发生变化时，随机选取的非地质灾害数据特征将发生变化，包括数量的增减和空间位置的改变，其栅格内各因子信息值就会改变，这些变化会影响非地质灾害数据是否具有代表性，从而影响模型预测能力，反映在易发性分区图上，则是相同地区的易发性级别可能不同。本文是样本比例不断变小，随机选取的非地质灾害样本不断变多的过程。当负样本足够多时，模型训练过程中对负样本的判断就越合理准确，因此随着样本比例变小，极低和低易发区变多，模型准确率会提高。当样本比例较大，影响距离较近时，随机选取的非地质灾害样本很可能靠近已发生的地质灾害点，这时负样本各因子属性就与正样本较为相似，那么负样本就不具有代表性，样本数量本就不多，这样使得模型准确率就会下降。随着距离变远，正负样本属性值相似这种情况就会迅速降低，因此准确率会上升。随着样本比例变小，距离变化对易发性评价的影响将趋于稳定，这是因为样本比例变小，非地质灾害样本变多，当距离较近时虽然会有正负样本属性值相似的情况发生，但较远的负样本足够多，模型构建时仍能准确地判断负样本样本的特征。Tsangaratos 等[30]认为训练样本的增加会提升模型预测能力。Shirzadi 等[31]则认为模型预测精度会随训练样本的增加而增加。本研究随着样本数量的增加，模型预测效果和精度不断提高，表明研究区内正负样本比例越低模型预测性能越好，与上诉学者的研究结果保持一致。

5　结论和展望

（1）样本比例和距已知灾害点距离变化对地质灾害易发性评价结果有直接影响。随着样本比例变小，距已知灾害点距离增加，模型准确度会有不同程度的上升。综合各检验方法最终判断样本比例为 1 : 10，距已知灾害点距离 1 000 m 是最适合研究区的模型。

（2）根据最适宜研究区的模型，低和极低易发区主要分布在研究区东北、东南和西南部分；中易发区主要分布在研究区中部和北部地区；高和极高易发区主要分布在研究区中部和北部沿道路和河流两侧的地区，这是柳林县防灾减灾的关键地区。

（3）样本比例和距已知灾害点距离的变化会使易发性结果产生较大的影响，从 RF 模型本身来讲，每棵树选取样本特征的结果就会发生变化，多棵树聚类判断时

就会产生差异，模型训练的准确度就会有差异；最关键的原因是随机选取的非地质灾害样本是否有代表性发生变化，模型准确判定负样本的特征能力发生变化。

（4）本文基础数据精度存在缺陷，如果精度更高评价结果可能更准确；地质灾害易发性评价与研究区栅格大小可能也有关系，值得研究；可对模型误差的临界值进行更深入的研究。

参考文献（References）：

［1］ ARABAMERI A, YAMANI M, PRADHAN B, et al. Novel ensembles of COPRAS multi-criteria decision-making with logistic regression, boosted regression tree, and random forest for spatial prediction of gully erosion susceptibility［J］. Science of the Total Environment, 2019, 688: 903 − 916.

［2］ UITTO J I, SHAW R. Sustainable development and disaster risk reduction: Introduction［M］//Sustainable Development and Disaster Risk Reduction. Tokyo: Springer, 2016: 1 − 12.

［3］ JIANG Weiguo, RAO Pingzeng, CAO Ran, et al. Comparative evaluation of geological disaster susceptibility using multi-regression methods and spatial accuracy validation［J］. Journal of Geographical Sciences, 2017, 27(4): 439 − 462.

［4］ PAWLUSZEK-FILIPIAK K, OREŃCZAK N, PASTERNAK M. Investigating the effect of cross-modeling in landslide susceptibility mapping［J］. Applied Sciences, 2020, 10(18): 6335.

［5］ 申怀飞，董雨，杨梅，等. 基于 AHP 与信息量法的甘肃省滑坡易发性评估［J］. 水土保持研究，2021, 28(6): 412 − 419.［SHEN Huaifei, DONG Yu, YANG Mei, et al. Assessment on landslide susceptibility in Gansu Province based on AHP and information quantity method［J］. Research of Soil and Water Conservation, 2021, 28(6): 412 − 419.（in Chinese with English abstract）］

［6］ KANUNGO D P, ARORA M K, SARKAR S, et al. A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas［J］. Engineering Geology, 2006, 85(3/4): 347 − 366.

［7］ 孙滨，祝传兵，康晓波，等. 基于信息量模型的云南东川泥石流易发性评价［J］. 中国地质灾害与防治学报，2022, 33(5): 119 − 127.［SUN Bin, ZHU Chuanbing, KANG Xiaobo, et al. Susceptibility assessment of debris flows based on information model in Dongchuan, Yunnan Province［J］. The Chinese Journal of Geological Hazard and Control, 2022, 33(5): 119 − 127.（in Chinese with English abstract）］

［8］ 熊小辉，汪长林，白永健，等. 基于不同耦合模型的县域滑坡易发性评价对比分析——以四川普格县为例［J］. 中国地质灾害与防治学报，2022, 33(4): 114 − 124.［XIONG Xiaohui, WANG Changlin, BAI Yongjian, et al. Comparison of landslide susceptibility assessment based on multiple hybrid models at County level: A case study for Puge County, Sichuan Province［J］. The Chinese Journal of Geological Hazard and Control, 2022, 33(4): 114 − 124.（in Chinese with English abstract）］

［9］ 吴常润，角媛梅，王金亮，等. 基于频率比-逻辑回归耦合模型的双柏县滑坡易发性评价［J］. 自然灾害学报，2021, 30(4): 213 − 224.［WU Changrun, JIAO Yuanmei, WANG Jinliang, et al. Frequency ratio and logistic regression models based coupling analysis for susceptibility of landslide in Shuangbai County［J］. Journal of Natural Disasters, 2021, 30(4): 213 − 224.（in Chinese with English abstract）］

［10］ 杜国梁，杨志华，袁颖，等. 基于逻辑回归-信息量的川藏交通廊道滑坡易发性评价［J］. 水文地质工程地质，2021, 48(5): 102 − 111.［DU Guoliang, YANG Zhihua, YUAN Ying, et al. Landslide susceptibility mapping in the Sichuan-Tibet traffic corridor using logistic regression-information value method［J］. Hydrogeology & Engineering Geology, 2021, 48(5): 102 − 111.（in Chinese with English abstract）］

［11］ 郭飞，王秀娟，陈玺，等. 基于不同模型的赣南地区小型削方滑坡易发性评价对比分析［J］. 中国地质灾害与防治学报，2022, 33(6): 125 − 133.［GUO Fei, WANG Xiujuan, CHEN Xi, et al. Comparative analyses on susceptibility of cutting slope landslides in southern Jiangxi using different models［J］. The Chinese Journal of Geological Hazard and Control, 2022, 33(6): 125 − 133.（in Chinese with English abstract）］

［12］ 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, 41(2): 79 − 90.［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, 41(2): 79 − 90.（in Chinese with English abstract）］

［13］ 何书，鲜木斯艳·阿布迪克依木，胡萌，等. 基于自组织特征映射网络-随机森林模型的滑坡易发性评价——以江西大余县为例［J］. 中国地质灾害与防治学报，2022, 33(1): 132 − 140.［HE Shu, ABUDIKEYIMU XMSY, HU Meng, et al. Evaluation on landslide susceptibility based on self-organizing feature map network and random forest model: A case study of Dayu County of Jiangxi Province［J］. The Chinese Journal of Geological Hazard and Control, 2022, 33(1): 132 − 140.（in Chinese with English abstract）］

［14］ GE G, SHI Zhongjie, ZHU Yuanjun, et al. Land use/cover classification in an arid desert-oasis mosaic landscape of China using remote sensed imagery: Performance assessment of four

machine learning algorithms［J］. Global Ecology and Conservation, 2020, 22: e00971.

［15］ 吉日伍呷, 田宏岭, 韩继冲. 基于不同机器学习算法的地震滑坡易发性评价——以鲁甸地震为例［J］. 昆明理工大学学报(自然科学版), 2022, 47(2): 47－56. ［JI R, TIAN Hongling, HAN Jichong. Evaluation of the susceptibility of earthquake landslides based on different machine learning algorithms: Taking Ludian earthquake as an example［J］. Journal of Kunming University of Science and Technology (Natural Science), 2022, 47(2): 47 － 56. (in Chinese with English abstract)］

［16］ 李坤, 赵俊三, 林伊琳, 等. 基于RF和SVM模型的东川泥石流易发性评价研究［J］. 云南大学学报(自然科学版), 2022, 44(1): 107－115. ［LI Kun, ZHAO Junsan, LIN Yilin, et al. Assessment of debris flow susceptibility in Dongchuan based on RF and SVM models［J］. Journal of Yunnan University (Natural Sciences Edition), 2022, 44(1): 107－115. (in Chinese with English abstract)］

［17］ 邱维蓉, 吴帮玉, 潘学树, 等. 几种聚类优化的机器学习方法在灵台县滑坡易发性评价中的应用［J］. 西北地质, 2020, 53(1): 222－233. ［QIU Weirong, WU Bangyu, PAN Xueshu, et al. Application of several cluster-optimization-based machine learning methods in evaluation of landslide susceptibility in Lingtai County［J］. Northwestern Geology, 2020, 53(1): 222－233. (in Chinese with English abstract)］

［18］ DOU Jie, YUNUS A P, MERGHADI A, et al. Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning［J］. Science of the Total Environment, 2020, 720: 137320.

［19］ DOU Jie, YUNUS A P, TIEN BUI D, et al. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan［J］. Science of the Total Environment, 2019, 662: 332－346.

［20］ BEHNIA P, BLAIS-STEVENS A. Landslide susceptibility modelling using the quantitative random forest method along the northern portion of the Yukon Alaska Highway Corridor, Canada［J］. Natural Hazards, 2018, 90(3): 1407－1426.

［21］ WANG Yue, SUN Deliang, WEN Haijia, et al. Comparison of random forest model and frequency ratio model for landslide susceptibility mapping (LSM) in Yunyang County (Chongqing, China)［J］. International Journal of Environmental Research and Public Health, 2020, 17(12): 4206.

［22］ YI Yaning, ZHANG Zhijie, ZHANG Wanchang, et al. Landslide susceptibility mapping using multiscale sampling strategy and convolutional neural network: A case study in Jiuzhaigou region［J］. CATENA, 2020, 195: 104851.

［23］ BREIMAN L. Random forests［J］. Machine Language, 2001, 45(1): 5－32.

［24］ 於佳宁,刘凯,张冰玥,等.中国区域TanDEM-X 90 m DEM高程精度评价及其适用性分析[J].地球信息科学学报,2021,23(04):646-657. ［YU Jianing, LIU Kai, ZHANG Bingyue, et al. Vertical accuracy assessment and applicability analysis of TanDEM-X 90 m DEM in China［J］. Journal of Geo-information Science, 2021, 23(4): 646－657.(in Chinese with English abstract)］

［25］ OSNA T, SEZER E A, AKGUN A. GeoFIS: An integrated tool for the assessment of landslide susceptibility［J］. Computers & Geosciences, 2014, 66: 20－30.

［26］ 张玘恺, 凌斯祥, 李晓宁, 等. 九寨沟县滑坡灾害易发性快速评估模型对比研究［J］. 岩石力学与工程学报, 2020, 39(8): 1595 － 1610. ［ZHANG Qikai, LING Sixiang, LI Xiaoning, et al. Comparison of landslide susceptibility mapping rapid assessment models in Jiuzhaigou County, Sichuan Province, China［J］. Chinese Journal of Rock Mechanics and Engineering, 2020, 39(8): 1595 － 1610. (in Chinese with English abstract)］

［27］ VAKHSHOORI V, POURGHASEMI H R, ZARE M, et al. Landslide susceptibility mapping using GIS-based data mining algorithms［J］. Water, 2019, 11(11): 2292.

［28］ 段宇英, 汤军, 刘远刚, 等. 基于随机森林的山西省柳林县黄土滑坡空间敏感性评价［J］. 地理科学, 2022, 42(2): 343－351. ［DUAN Yuying, TANG Jun, LIU Yuangang, et al. Spatial sensitivity evaluation of loess landslide in Liulin County, Shanxi based on random forest［J］. Scientia Geographica Sinica, 2022, 42(2): 343 － 351. (in Chinese with English abstract)］

［29］ 吴润泽, 胡旭东, 梅红波, 等. 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［J］. 地球科学, 2021, 46(1): 321－330. ［WU Runze, HU Xudong, MEI Hongbo, et al. Spatial susceptibility assessment of landslides based on random forest: A case study from Hubei section in the Three Gorges Reservoir area［J］. Earth Science, 2021, 46(1): 321－330. (in Chinese with English abstract)］

［30］ TSANGARATOS P, ILIA I. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size［J］. CATENA, 2016, 145: 164－179.

［31］ SHIRZADI A, SOLAIMANI K, ROSHAN M H, et al. Uncertainties of prediction accuracy in shallow landslide modeling: Sample size and raster resolution［J］. CATENA, 2019, 178: 172－188.