基于 GIS 和随机森林模型的泥石流敏感性分析
——以吉林省洮南市北部山区为例

扈秀宇[1], 秦胜伍[1], 窦 强[1], 刘 飞[1], 乔双双[1], 董 冬[2]
(1. 吉林大学 建设工程学院, 吉林 长春 130000; 2. 吉林省地质环境监测总站, 吉林 长春 130000)

摘 要: [目的] 对区域性泥石流敏感性进行分析,为吉林省洮南市泥石流灾害预测研究提出一种高效快捷的分析模型。[方法] 针对现行大多数概率统计模型预测率较低的不足,利用人工智能算法中效果明显的随机森林算法,以吉林省洮南市西北部山区为研究区域,选用高程、坡度、坡向、平面曲率、剖面曲率、河流、归一化差分植被指数、地形湿度指数、土地利用及岩性 10 个评价因子构建了频率比和随机森林泥石流敏感性评价模型进行对比验证。模型准确性的验证方法采用受试者特征曲线(ROC 曲线)及累积频率曲线下面积(area under curve,AUC)。[结果] 随机森林对研究区泥石流敏感性进行分析,并通过 GIS 将敏感性图分为 5 个敏感性区域,位于高敏感性区以上的灾害点占 82.3%。验证模型成功率及预测率分别为 88.4%与 90.4%,相较于频率比的成功率及预测率(86.4%和 75.1%)效果良好。[结论] 在洮南市北部进行泥石流敏感性分析中,采用随机森林方法进行建模,并利用频率比方法进行对比,结果显示随机森林法结果可靠准确。

关键词: 泥石流敏感性分析; GIS; 随机森林模型; 频率比模型; 洮南北部

文献标识码: B **文章编号**: 1000-288X(2019)05-0204-07 **中图分类号**: P642.23

文献参数: 扈秀宇, 秦胜伍, 窦强, 等. 基于 GIS 和随机森林模型的泥石流敏感性分析[J]. 水土保持通报, 2019,39(5):204-210. DOI:10.13961/j.cnki.stbctb.2019.05.028; Hu Xiuyu, Qin Shengwu, Dou Qiang, et al. Susceptibility analysis of debris flow based on GIS and random forest[J]. Bulletin of Soil and Water Conservation, 2019,39(5):204-210.

Susceptibility Analysis of Debris Flow Based on GIS and Random Forest
—A Case Study of a Mountainous Area in Northern Taonan City, Jilin Province

Hu Xiuyu[1], Qin Shengwu[1], Dou Qiang[1], Liu Fei[1], Qiao Shuangshuang[1], Dong dong[2]
(1. *College of Construction Engineering, Jilin University, Changchun, Jilin* 130000, *China*; 2. *Jilin Institute of Geo-environment Monitoring, Changchun, Jilin* 130000, *China*)

Abstract: [Objective] The susceptibility of regional debris flow was analyzed and an efficient and rapid analysis model proposed for the debris flow disaster prediction research Taonan area of Jilin Province. [Methods] Owing to the current shortcomings of most probabilistic statistical models that have low prediction rates, the random forest algorithm with obvious effect built on an artificial intelligence algorithm was used to study the northwestern mountainous area of Taonan City, Jilin Province. The elevation, slope, aspect, plane curvature, profile curvature, river, normalized difference vegetation index, topographic humidity index, land use, and lithology were selected. A random forest debris flow susceptibility assessment model for the study area was constructed using these factors. The frequency ratio method was used to model and compare with the random forest model. The model effect was verificated using a receiver operating characteristic curve and an area under the curve. [Results] Random forests were used to analyze the sensitivity of debris flow in the

收稿日期:2019-03-26 **修回日期**:2019-04-12
资助项目:吉林省科技厅科技发展计划项目"长白山火山泥石流时空演化机理及动态预警技术研究"(20190303103SF),"长白山天池火山地质结构演变与外因触发次生灾害风险评估"(20170101001JC)
第一作者:扈秀宇(1995—),男(回族),安徽省蚌埠市人,硕士研究生,主要从事地质灾害与防治研究。E-mail:huxiuyu95@163.com。
通讯作者:秦胜伍(1980—),男(汉族),吉林省长春市人,博士,副教授,博士生导师,主要从事地质工程、地质灾害治理研究。E-mail:qinsw@jlu.edu.cn。

study area and were divided into five sensitivity areas by GIS. The disaster points above the high sensitivity area accounted for 82.3%. The success rate and prediction rate of the verification model were 88.4% and 90.4%, respectively, which was better than the success rate and prediction rate of frequency ratio (86.4% and 75.1%). [Conclusion] The sensitivity analysis of debris flow in the Northern Taonan City was performed using the random forest method and compared with the frequency ratio method. The random forest results were reliable and accurate.

Keywords: debris flow susceptibility analysis; GIS; random forest model; frequency ratio model; northern Taonan

泥石流是由土、碎石及水组成并在重力作用下以>10 m/s 速度[1]前行的混合物,是山区最具破坏性的自然灾害之一。泥石流令其物源区水土流失,流通区的沟谷遭受侵蚀;其在堆积区淹没住房、农田和林地,堵塞河流。研究区洮南北部山区于 2014 年夏季由于持续性强降雨,形成 126 处泥石流灾害,使人民财产安全受到威胁。为减少泥石流灾害所带来的损失,利用影响泥石流发育的因子及已发生的泥石流灾害点进行训练建模,从而能够对泥石流敏感性分级评价,评价结果能够有效而直观地预测泥石流敏感区域。

对于泥石流敏感性研究,国外起步较早。1976 年,联合国委托国际工程地质联合会开展泥石流敏感性研究,此后逐渐成为泥石流地质灾害防治评价的重要内容。20 世纪 80 年代初,美国学者 Kovacs 等利用定性评价方法对泥石流敏感性建立了评价模型,为泥石流敏感性评价提供了新的方向[2]。此后利用因子叠加评估地质灾害危险度的方法被很多学者使用。而国内起步稍晚,1986 年,谭炳炎[3]第一次对泥石流的易发性进行了数量化综合评价。1988 年,刘希林[4]首次使用多因子叠加方法对泥石流危险性进行研究。1994 年,唐川等[5]利用数值模拟对泥石流堆积扇进行了危险度评价。之后,众多学者利用定性和定量分析方法对泥石流进行敏感性分析,如层次分析法[6],模糊数学法[7-8],信息量法[9],回归分析[10],频率比法[11],人工神经网络[12]等。这些方法由定性到半定量再到定量过渡并且已应用较长时间,使用范围广泛。但不能对因子进行权重分析,而且在灾害点和评价因子较多的情况下,上述方法并不能在运算复杂大数据上取得优势。而现在兴起的人工智能算法对大数据能够进行高速处理并进行自组织学习。

随机森林算法是以简单高效著称的人工智能算法[13],开始是在 2001 年由 Breiman 提出的一种人工智能算法[14],用于市场营销、医学领域,预测客户的保留与流失以及预测疾病风险与病患的易感性[15]。它将决策树作为单元进行集成学习的树形分类器组合算法。其相比于决策树及其他算法,它集成了 Bagging 算法和随机选择特征分裂特点,准确率较高,能够处理大数据并评估每个因子在分类上的重要性。本文在研究区使用随机森林进行泥石流敏感性建模,然后与概率统计方法中使用广泛且简单高效的频率比法进行比较验证。以期能够客观验证随机森林模型的效果,又能对洮南北部山区泥石流灾害预警提供直观有效的参考。

1　研究区概况

研究区域位于吉林省洮南市北部,其面积约 1 181 km^2。洮南地处东三省与内蒙古交界中心。地理坐标为东经 122°45′—122°48′,北纬 45°19′—45°21′。洮南市总体上属北温带大陆性季风气候区,春季干燥多风,夏季温热多雨,秋季凉爽多旱,冬季寒冷少雪,平均年降雨量 391.3 mm。洮南地处大兴安岭东麓,松嫩平原西部,北部丘陵高低起伏,嫩江支流发源于此。研究区多属于低山丘陵,低山海拔多在 500～550 m,坡度多在 25°以下,主要由花岗闪长岩、凝灰质砂岩和花岗岩等组成。丘陵海拔 250～500 m,相对高度在百米以下,以低丘陵为主。多由火山碎屑岩、花岗岩等组成。山坡残积,坡积物发育。研究区泥石流频发,大部分位于洮南市西北角,包括胡力吐乡、万宝乡及部分万宝镇、东升乡、那金镇。泥石流多发生在低山丘陵区,山峰呈北东～南西方向展布,山坡坡度为 15°～25°,坡面冲沟发育,植被覆盖率较低,出露岩石主要是侏罗系碎屑岩和燕山期岩浆岩。发生泥石流的坡面由于过度开垦、放牧,植被覆盖率低,裸露岩石极易受到风化,坡面抗冲刷能力较差,致使泥石流多沿河流发育,多威胁农田,屋舍和道路。

研究区经调查发现 126 处泥石流灾害,对此 126 处灾害点随机分为训练集灾害点和验证集灾害点。训练集灾害点用于数据的训练和模型的建立;验证集灾害点用于验证模型的可靠程度。运用随机森林算法进行数据分析时,传统方法是将数据按照 7∶3 进行训练集和验证集的划分。所以本文将训练集灾害点划分为 85 个(70%)和验证集 41 个(30%)用于模型建立和验证。

2 研究方法

2.1 评价因子选取

影响泥石流发育的因子十分复杂,因子之间相互联系,相互影响,本文泥石流评价因子主要依据研究区泥石流灾害的分布及发育现状,在充分考虑研究区资料获取难易程度以及研究尺度后,选取地表高程、坡向、坡度、平面曲率、剖面曲率、到河流距离、土地利用、岩性、归一化差分植被指数(NDVI)及地形湿度指数(TWI)作为评价因子,利用 ArcGIS 进行分类后如图 1 及附图 1 所示。数字高程数据(DEM)直接反映研究区的地形起伏变化,在一定程度上反映了沟谷,植被变化及堆积物的状态。DEM 栅格大小为 30 m×30 m,高程范围为 163～650 m,将其按照自然断点法分为 5 类,分别为<250,250～300,300～350,350～410,410～650 m。

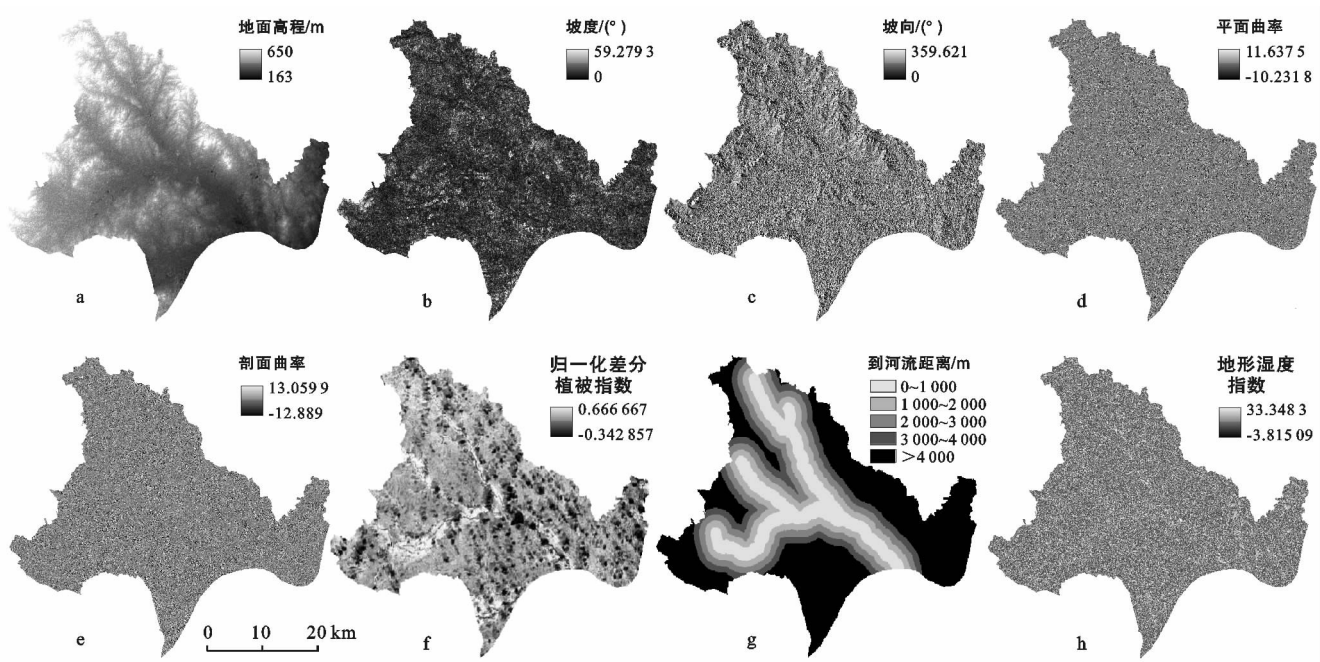


图 1 泥石流敏感性分析的评价因子

坡向对泥石流发育主要影响山坡岩体风化程度以及植被生长发育情况。ArcGIS 计算坡向是按照每个栅格与相邻栅格值变化最大的下坡方向计算,故会产生-1°坡向,即不具有下坡方向的平坦区域。研究区坡向分类按照自然间断点分为 10 类,分别为-1°,0°～22.5°,22.5°～67.5°,67.5°～112.5°,112.5°～157.5°,157.5°～202.5°,202.5°～247.5°,247.5°～292.5°,292.5°～337.5°,337.5°～360°。

坡度对于泥石流的发育是十分重要的一个因素,常被直接用作泥石流敏感性制图。研究区坡度变化范围为 0°～59.28°,其分为 6 类,分别为 0°～4°,4°～7°,7°～11°,11°～16°,16°～24°,24°～59.28°。

平面曲率及剖面曲率代表地形表面的凹凸程度,其间接影响泥石流发育范围。平面曲率值为正值时,表明地形表面为向上凸,负值表明地形向下凹,值为 0 时,表面为水平。剖面曲率与平面曲率相反,正值表明地面向下凹,负值表明向上凸。平面曲率划分为 5 类,分别为-10.23～-1.14,-1.14～-0.28,-0.28～0.32,0.32～1.17,1.17～11.64。剖面曲率划分为 5 类,分别为-12.89～-1.39,-1.39～-0.47,-0.47～0.24,0.24～1.15,1.15～13.06。

泥石流常发生于山区沟壑,经常堵塞河流,而河流沟谷也为泥石流提供流动通道,使其流通距离加长,规模变大。本文在 ArcGIS 平台上对河流进行缓冲区划分,分为 5 个缓冲区,分别为 0～1 000,1 000～2 000,2 000～3 000,3 000～4 000,>4 000 m。

地形湿度指数(TWI)表征地形与土壤随空间的水分分布,是一种对径流长度及产流面积的定量描述[16]。其直接反映了土壤含水量随地形及空间变化规律,以及间接反映研究区的水土流失状况。研究区地形湿度指数分为 5 类,分别为<3,3～6,6～9,9～12,12～33.2。

归一化差分植被指数(NDVI)通常与植物蒸腾作用、光合作用有关,是指示植被生长状态及植被覆盖

度的最佳指示因子[17]。其优点是可以消除太阳高度角、卫星观测角、地形、云影及大气相关的辐射影响。其计算公式为:

$$NDVI=\frac{NIR-R}{NIR+R} \tag{1}$$

式中:NIR——近红外波段;R——红波段处的反射率值。NDVI 取值在−1 与 1 之间,负值表示地面覆盖为对可见光高反射的物体,如云、水、冰、雪等,0 表示近红外波段与红波段近似,代表裸地或岩石。正值代表植被覆盖,其值越大,植被覆盖度越大。研究区 NDVI 范围−0.34～0.67,分为 5 类,分别为−0.34～0,0～0.22,0.22～0.32,0.32～0.43,0.43～0.67。

土地利用则分为采矿用地、城市建设用地、道路、湿地、农林用地及河流湖泊。岩性分为火山碎屑岩、凝灰质砂岩、花岗岩、砂岩、全新世沉积物及更新世沉积物。

评价因子数据来源如表 1 所示。

表 1　评价因子数据来源

评价因子	数据来源
高 程	地理空间数据云下载
坡 向	
坡 度	
平面曲率	Arcgis 高程提取
剖面曲率	
到河流距离	
土地利用	全国地表覆盖专题图 30 m×30 m
岩 性	1:5 万地质图
NDVI	Landsat8 遥感影像
TWI	Arcgis 高程提取

注:NDVI 为归一化植被指数;TWI 为地形湿度指数。下同。

2.2　随机森林模型

随机森林由很多决策树模型组成的分类模型。它与 Bagging 算法相同,计算过程如图 2 所示:①采用 Bootstrap 方法随机放回地选择 i 个训练数据集中的数据 $D_1,D_2,D_3,\cdots,D_i$;②构造 i 个决策树进行分类训练,得到 i 个弱分类器 $C_1,C_2,C_3,\cdots,C_i$。③对得到的弱分类器按照投票组合成强分类器。由于组成随机森林算法(RF)的决策树模型不进行剪枝,可得到偏差较小的分类树,确保 RF 对测试数据分类训练的准确性。弱分类器之间能够互补差异,把单个分类器的误差降到最小,从而增强整体的准确率。

与其他算法相比,随机森林算法还可以对分类特征(即评价因子)进行重要性分析,得出每个评价因子的重要性。计算重要性的方法为重要性评分(VIM),其主要分为两种计算方法,分别为基尼指数计算和 OOB 误差率置换计算法。本文的模型使用基尼指数计算节点的不纯度来衡量特征的重要性,计算过程如下。

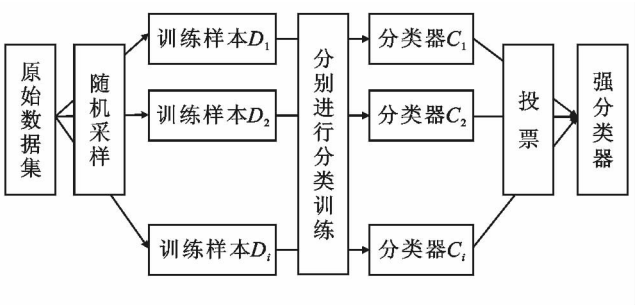


图 2　随机森林方法流程

假设有 M 个特征为 $X_1,X_2,\cdots,X_m$,基尼指数计算公式为:

$$GI_c=1-\sum_{k=1}^{|K|}P_{ck}^2 \tag{2}$$

式中:K——K 个类别;P_{ck}——c 节点中类别 k 所占的比例。

$$VIM_{jc}^{gini}=GI_c-GI_l-GI_r \tag{3}$$

式中:VIM_{jc}——第 j 个因子在节点 c 的重要性;GI_l,GI_r——向下分枝后两个新节点处的基尼值。

$$VIM_{ij}^{gini}=\sum VIM_{jc}^{gini} \tag{4}$$

式中:VIM_{ij}——第 j 个因子在第 i 个树的重要性。

假设有 n 个决策树,则第 j 个因子的重要性为 VIM_j。

$$VIM_j=\frac{\sum_{i=1}^{n}VIM_{ji}^{gini}}{\sum_{i=1}^{m}VIM_i} \tag{5}$$

根据随机森林算法原理,结合 GIS 进行随机森林评价的过程为[18-20]:

(1) 对泥石流灾害点数进行分类,70%(85 个)灾害点作为训练集进行随机森林训练计算,其余 30%(41 个)作为验证数据集对结果进行预测率验证。

(2) 根据研究区泥石流发育特点,选取 10 个评价因子。

(3) 从训练数据 N 中使用 Bootstrap 重抽样方法随机有放回地选出 K 个不同的样本,进行 n 次采样,得到 n 个训练集。

(4) 对 n 个训练集,分别训练 n 个决策树模型。

(5) 对于单个决策树模型,假设训练样本特征的个数为 M,那么每次分裂时都从 M 中抽取 m 个特征作为当前节点的分裂特征集。

(6) 每棵树都一直这样分裂下去,直到该节点的所有训练样例都属于同一类。在决策树的分裂过程中不需要剪枝。

(7) 利用生成的各个决策树组成随机森林,其预测结果取 K 棵决策树的平均值,从而得到最终预测结果。

(8) 根据模型预测结果进行敏感性分级,利用 GIS 生成研究区泥石流敏感性评价图,并得到评价因子重要性评估值。

2.3 频率比模型

利用概率统计的方法,对 10 个评价因子进行频率比方法的统计计算,然后对评价因子进行叠加,得到泥石流敏感性评价图。首先利用随机森林划分的训练集(70%)与验证集(30%)对灾害点分类。其次对每个泥石流评价因子的 FR 计算。利用 ArcGIS 将泥石流训练集灾害点与每个评价因子图叠加提取所需数据,并且使用公式(6)计算每个类别的频率比值。这一步对每个评价因子进行计算(如表 2 所示)。最后,通过将常规 FR 方法中的 FR 值(式 7)求和来创建泥石流敏感性指数并进行敏感性制图。

$$FR_{ij}=\frac{N_{ij}/A_{ij}}{N_T/A_T} \tag{6}$$

式中:N_{ij}——因子 i 的第 j 个子类中的泥石流灾害点数;A_{ij}——相应子类所占栅格数;N_T——泥石流灾害点总数;A_T——调查区域所占总栅格数。

$$DSI=\sum_{i=1}^{n}FR_i^j \tag{7}$$

式中:DSI——泥石流敏感性指数,敏感性指数越大,泥石流发生的概率越大;FR——每个因子级别的计算频率比;N——因子的数量。

3 结果及分析

3.1 模型结果

随机森林算法模型(RF)通过 SPSS Modeler 18.0 软件构建,参数设置为 500 个决策树,节点个数 m 为评价因子个数(10 个)的平方根,故将 m 取值为 3。样本数据使用训练集进行训练,使用自然断点法对滑坡敏感性图进行分类,共分为 5 个敏感区,分别为极低敏感区、低敏感区、中敏感区、高敏感区和极高敏感区,最后得到敏感性结果见附图 2,由附图 2 可知,大约有 14.6%区域对泥石流敏感性极低,其余 22.2%,30.3%,17.6%,15.3%的区域分别表现为低、中、高和极高敏感性。位于极低、低、中、高、极高敏感性区的训练集灾害点分别有 2,5,8,30 个及 40 个,位于高敏感性区以上的灾害点占 82.3%。

随机森林算法模型模拟出的高敏感和极高敏感区主要集中在洮南市那金镇东北部地区,胡力吐蒙古族乡北部地区,万宝乡大部分地区和万宝镇西部和北部地区。高敏感区共分布已发生泥石流 61 处,占训练集的 71%。高程大部分位于 250 m 至 400 m,坡度分布为 10°至 30°,此段区域主要为低山至河流过渡区域,高差较大,坡度适中,能够为泥石流提供势能,同时并不影响农业开垦;NDVI 分布为 0 至 0.3,地表主要为裸地和农作物,土地类型中农业用地占 90.4%,致使水土保持能力变差,土壤疏松,岩性主要为易风化的火山碎屑岩和凝灰质砂岩,为泥石流形成提供充足的物质来源;高敏感区地表沟谷发育,泥石流灾害点距河流距离分布大多大于 3 000 m,泥石流发育多沿沟谷流通,而沟谷前的平原地区又为泥石流提供堆积场所,致使大型河流对研究区的影响较低。

随机森林算法模型根据决策树各个节点的平均基尼值的减少量来计算评价因子的重要性,并对所计算的因子权重进行归一化处理。如图 3 所示,坡度、高程、TWI 和坡向是影响研究区泥石流发育十分重要的 4 个因子,重要性均超过 10%,占全部因子的 52.57%。平面曲率、到河流距离、剖面曲率和 NDVI,重要性占比为 37.24%,对泥石流发育的影响较为重要。岩性和土地利用重要性占比为 10.2%,对泥石流发育影响较轻。

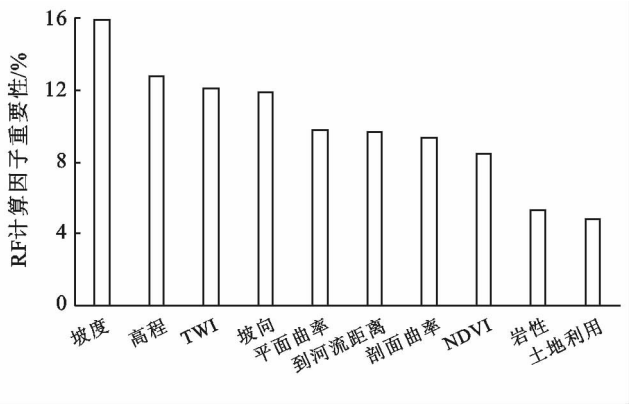


图 3 吉林省洮南地区随机森林评价因子重要性

频率比敏感性评价中,按照公式(2)计算 10 个评价因子频率比模型,然后根据公式(3)在 ArcGIS 中进行叠加计算,由此得出泥石流频率比敏感性图,如附图 3 所示。按照自然断点法对其进行分类,与随机森林模型相同,共分为 5 个敏感区,极低敏感区 12.4%,低敏感区占 19.1%,中敏感区占 23.7%,高敏感区占 26.9%,极高敏感区占 17.9%。其中训练集中位于极低、低、中、高极高敏感区的灾害点分别有 0,10,16,27,32 个,位于高敏感性区域以上灾害点占训练集的 69%。

表 2　泥石流敏感性频率比值

评价因子	分 类	子类所占栅格数	子类所占比例/%	子类泥石流灾害点数	子类泥石流灾害点所占比例/%	频率比
高 程	＜250	381 120	0.221	2	0.02	0.11
	250～300	521 421	0.302	19	0.22	0.74
	300～350	442 120	0.256	30	0.35	1.38
	350～410	269 875	0.157	31	0.36	2.33
	410～650	109 749	0.064	3	0.04	0.55
坡 向	－1	2 211	0.001	0	0	0
	0～22.5	113 051	0.066	5	0.06	0.90
	22.5～67.5	208 971	0.121	11	0.13	1.07
	67.5～112.5	231 898	0.134	7	0.08	0.61
	112.5～157.5	221 755	0.129	9	0.11	0.82
	157.5～202.5	236 499	0.137	14	0.16	1.20
	202.5～247.5	212 036	0.123	11	0.13	1.05
	247.5～292.5	211 887	0.123	8	0.09	0.77
	292.5～337.5	191 630	0.111	12	0.14	1.27
	337.5～360	94 347	0.055	8	0.09	1.72
坡 度	0～4	459 055	0.266	30	0.35	1.33
	4～7	497 626	0.289	24	0.28	0.98
	7～11	425 639	0.247	20	0.24	0.95
	11～16	227 026	0.132	9	0.11	0.80
	16～24	93 677	0.054	2	0.02	0.43
	24～62	21 262	0.012	0	0	0
平面曲率	－10.23～ －1.14	100 875	0.059	6	0.07	1.21
	－1.14～ －0.28	426 928	0.248	19	0.22	0.90
	－0.28～0.32	721 567	0.418	39	0.46	1.10
	0.32～1.17	376 120	0.218	14	0.16	0.76
	1.17～11.64	98 795	0.057	7	0.08	1.44
剖面曲率	－12.89～－1.39	88 666	0.051	6	0.06	1.23
	－1.39～－0.47	325 163	0.189	12	0.13	0.67
	－0.47～0.24	675 223	0.392	34	0.36	0.91
	0.24～1.15	498 995	0.289	35	0.37	1.27
	1.15～13.06	136 238	0.079	8	0.08	1.07
到河流距离	0～1 000	259 037	0.150	3	0.04	0.24
	1 000～2 000	235 786	0.136	10	0.12	0.86
	2 000～3 000	226 252	0.131	19	0.22	1.71
	3 000～4 000	207 050	0.120	13	0.15	1.28
	＞4 000	800 010	0.463	40	0.47	1.02
TWI	＜3	762 373	0.442	33	0.39	0.88
	3～6	295 267	0.171	14	0.16	0.96
	6～9	459 662	0.267	20	0.24	0.88
	9～12	154 064	0.089	10	0.12	1.32
	12～33.2	52 918	0.031	8	0.09	3.07
NDVI	－0.34～0	173 995	0.543	7	0.30	0.56
	0～0.22	320 722	0.722	17	0.71	0.98
	0.22～0.32	444 201	0.883	23	1.64	1.86
	0.32～0.43	502 795	1.780	24	0.28	0.16
	0.43～0.67	282 474	0.164	14	0.16	1.01
土地利用	采矿用地	2 046	0.001	0	0	0
	城市建设用地	72 323	0.042	7	0.08	1.97
	湿 地	24 456	0.014	1	0.01	0.83
	农林用地	1 579 752	0.915	74	0.87	0.95
	道 路	18 992	0.011	1	0.01	1.07
	河流及湖泊	29 394	0.017	2	0.02	1.38
岩 性	火山碎屑岩	488 402	0.283	37	0.44	1.54
	凝灰质砂岩	158 109	0.092	7	0.08	0.90
	花岗岩	292 148	0.169	4	0.05	0.28
	砂 岩	380 550	0.221	22	0.26	1.17
	全新世沉积物	258 829	0.150	10	0.12	0.78
	更新世沉积物	147 401	0.085	5	0.06	0.69

3.2 ROC 曲线验证

ROC 曲线即称受试者特征曲线(receiver operating characteristic curve),其可以广泛应用于医学检验及疾病预测[21-22]。每个灾害点及其评价因子组成的个体相当于医学检验中的受试者,再将发生泥石流及未发生泥石流作为二分类(正类及负类)进行分析,得到每个模型所对应的 ROC 曲线及其曲线下面积(AUC)[23]。AUC 是判断所用随机森林及频率比模型的优劣标准,当 AUC=0.5 时,说明其模型结果不具有参考价值,而小于 0.5 则说明模型不符合真实情况。AUC 大于 0.5 时,其值越靠近 1,说明模型效果越准确。

由于将样本灾害点分为训练集及验证集,所以利用训练集数据可以得到模型准确率 ROC 曲线,其曲线下面积即模型的准确率。验证集数据得到预测率 ROC 曲线,用于验证模型是否预测准确,其曲线下面积为预测率。随机森林及频率比模型准确率及预测率如表 3 所示。

表 3 模型准确率及预测率结果

项 目	随机森法	频率比法
准确率/%	88.4	86.4
预测率/%	90.4	75.1

随机森林与频率比的准确率为 88.4%与 86.4%,说明随机森林模型训练效果较频率比结果准确。而预测率两者相差较大,随机森林预测率达到 90.4%,其模型预测结果准确性高。频率比预测率为 75.1%,和随机森林模型相比有一定差距,模型预测效果一般。准确率和预测率之间的差值可以提现模型的稳定性,随机森林准确率和预测率相差 2%,而频率比相差 11.3%,由此可知随机森林模型更稳定可靠。

4 结 论

吉林省洮南市北部山区是泥石流多发区域,根据该地区的地形因素、地质因素、植被因素及人为因素,选择了 10 个评价因子及随机森林模型进行泥石流敏感性评价,并用频率比模型与其对比验证。得到以下结论:

(1) 随机森林算法应用于泥石流敏感性分析无需提前设置因子权重,计算过程简化,运行平台丰富,可在可视化软件 SPSS Modeler 上运行,使用简便。

(2) 随机森林模型可以对评价因子进行重要性分析,便于分析各评价因子对泥石流发育的影响。

(3) 随机森林模型结果将研究区域划分为 5 个泥石流敏感性区域,分别为极低、低、中、高和极高敏感区,分别占研究区的面积的 14.6%,22.2%,30.3%,17.6%,15.3%。位于各分区的泥石流灾害数占比为 2.35%,5.88%,9.41%,35.29%和 47.06%。泥石流发生数随敏感性等级增大而增大,敏感性分级符合实际野外调查结果。

(4) 随机森林方法的准确率与预测率分别为比频率比方法高,准确率与预测率相差较小,模型稳定。随机森林模型可以应用于研究区的泥石流敏感性分析且效果良好。

[参 考 文 献]

[1] Hungr O, McDougall S, Wise M, et al. Magnitude-frequency relationships of debris flows and debris avalanches in relation to slope relief[J]. Geomorphology, 2008,96(3/4):355-365.

[2] 许冲,戴福初,徐锡伟.汶川地震滑坡灾害研究综述[J].地质论评,2010,56(6):860-874.

[3] 谭炳炎.泥石流沟严重程度的数量化综合评判[J].铁道学报,1986,8(2):74-82.

[4] 刘希林.泥石流危险度判定的研究[J].灾害学,1988,3(3):10-15.

[5] 唐川,刘琼招.中国泥石流灾害强度划分与危险区划探讨[J].中国地质灾害与防治学报,1994,5(S1):30-35,52.

[6] 郭瑞,马富存,郭一兵,等.基于层次分析法的泥石流易发性评价[J].东华理工大学学报:自然科学版,2016,39(4):347-351.

[7] 鲍春,宋丹青.模糊数学在库区泥石流易发性评价中的应用[J].水资源与水工程学报,2015,26(6):132-136.

[8] 蒋庆丰.山坡泥石流的风险性分析与评价研究[D].重庆:西南师范大学,2002.

[9] 张以晨,秦胜伍,翟健健,等.基于信息量的长白山地区泥石流易发性评价[J].水文地质工程地质,2018,45(2):150-158.

[10] 孙正超.基于 GIS 和 Logistic 回归模型的区域泥石流危险性评价[D].成都:西南石油大学, 2016.

[11] 齐信,黄波林,刘广宁,等.基于 GIS 技术和频率比模型的三峡地区秭归向斜盆地滑坡敏感性评价[J].地质力学学报,2017,23(1):97-104.

[12] 王国峰.基于遗传神经网络的矿山泥石流危险性评价[D].西安:西安科技大学,2008.

[13] 王奕森,夏树涛.集成学习之随机森林算法综述[J].信息通信技术,2018,12(1):49-55.

[14] Breiman L. Random Forests[J]. Machine Learning, 2001,45(1):5-32.

[15] 方匡南,吴见彬,朱建平,等.随机森林方法研究综述[J].统计与信息论坛,2011,26(3):32-38.

(下转第 217 页)

治理途径[J]. 人民黄河,2014,36(10):92-94,109.
[5] 毕慈芬,徐双民,李桂芬. 砒砂岩地区沟道沙棘植物"柔性坝"原型拦沙研究[J]. 国际沙棘研究与开发,2003,1(1):6-12.
[6] 肖培青,姚文艺,申震洲,等. 黄河中游砒砂岩区抗蚀促生技术研究[J]. 中国水土保持,2016(9):73-75.
[7] 姚文艺,刘慧,梁止水. 砒砂岩区生态恢复的新途径:一种环境友好型抗蚀促生技术[J]. 水土保持通报,2018,38(4):118-124.
[8] 刘丹丹,李敏,刘润进. 我国植物根围促生细菌研究进展[J]. 生态学杂志,2016,35(3):815-824.
[9] Burd G I, Dixon D G, Glick B R. Plant growth-promoting bacteria that decrease heavy metal toxicity in plants [J]. Canadian Journal of Microbiology, 2000, 46(3): 237-245.
[10] Yang Jungwook, Kloepper J W, Ryu C M. Rhizosphere bacteria help plants tolerate abiotic stress[J]. Trends in Plant Science, 2009,14(1):1-4.
[11] 余劲聪,何舒雅,曾润颖,等. 芽孢杆菌修复土壤重金属镉污染的研究进展[J]. 广东农业科学,2016,43(1):73-78.
[12] 杨丽,燕传明,贺卓,等. 重金属耐性芽孢杆菌的筛选及其对辣椒吸收镉铅的阻控效应[J]. 农业环境科学学报,2018,37(6):1086-1093.
[13] 庞海东,贺卓,燕传明,等. 耐重金属的植物促生芽孢杆菌筛选及其强化香蒲去除 Cd 的作用[J]. 农业环境科学学报,2017,36(11):2314-2321.
[14] Gordon S A, Weber R P. Colorimetric estimation of indoleacetic acid [J]. Plant Physiology, 1951,26(1): 192-195.
[15] Schwyn B, Neilands J B. Universal chemical assay for the detection and determination of siderophores [J]. Analytic Biochemistry, 1987,160(1):47-56.
[16] Xi Jun, He Linyan, Huang Zhi, et al. *Bacillus qingshengii* sp. nov., a rock-weathering bacterium isolated from weathered rock surface [J]. International Journal of Systematic & Evolutionary Microbiology, 2014,64: 2473-2479.
[17] 关松荫. 土壤酶及其研究法[M]. 北京:农业出版社,1986:274-276,294-297.
[18] 鲍士旦. 土壤农化分析 [M]. 3 版. 北京:中国农业出版社,2000.
[19] Vlamakis H, Chai Yunrong, Beauregard P, et al. Sticking together: building a biofilm the *Bacillus subtilis* way [J]. Nature Reviews Microbiology, 2013,11(3):157-168.
[20] Xu Zhihui, Zhang Ruifu, Wang Dandan, et al. Enhanced control of cucumber wilt disease by *Bacillus amyloliquefaciens* SQR9 by altering the regulation of its DegU phosphorylation [J]. Applied & Environmental Microbiology, 2014, 80(9):2941-50.
[21] 徐文静,靳晓东,杨秋生. 植物根际微生物的影响因素研究进展[J]. 河南农业科学,2014,43(5):6-12.
[22] 王浩,杨方社,李怀恩,等. 沙棘柔性坝对砒砂岩沟道泥沙粒径分布及有机质影响[J]. 水土保持学报,2017,31(5):158-163.
[23] 卢立娜,赵雨兴,胡莉芳,等. 沙棘(*Hippophae rhamnoides*)种植对鄂尔多斯砒砂岩地区土壤容重、孔隙度与贮水能力的影响[J]. 中国沙漠,2015,35(5):1171-1176.
[24] 刘龙,姚云峰,郭月峰,等. 准格尔旗砒砂岩区三种典型造林树种蒸腾耗水研究[J]. 中国农业科技导报,2018,20(3):124-131.
[25] 梁止水,吴智仁,杨才千,等. 基于 W-OH 的砒砂岩抗蚀促生机理研究[J]. 水利学报,2016,47(9):1160-1166.

(上接第 210 页)

[16] 张彩霞,杨勤科,李锐. 基于 DEM 的地形湿度指数及其应用研究进展[J]. 地理科学进展,2005,24(6):116-123.
[17] 尹超,王艳芳,张爱国. 基于 NDVI 的临汾市植被覆盖动态变化遥感监测研究[J]. 山西师范大学学报:自然科学版,2011,25(3):125-128.
[18] 王全喜,孙鹏举,刘学录,等. 基于随机森林算法的耕地面积预测及影响因素重要性分析:以甘肃省庆阳市为例[J]. 水土保持通报,2018,38(5):341-346.
[19] 甄亿位,郝敏,陆宝宏,等. 基于随机森林的中长期降水量预测模型研究[J]. 水电能源科学,2015,33(6):6-10.
[20] Shrestha S, Kang T S, Suwal M. An ensemble model for co-seismic landslide susceptibility using GIS and random forest method[J]. ISPRS International Journal of Geo-Information, 2017,6(11):365.
[21] 王运生,谢丙炎,万方浩,等. ROC 曲线分析在评价入侵物种分布模型中的应用[J]. 生物多样性,2007,15(4):365-372.
[22] 刘润幸. 使用 SPSS 作多变量观察值的 ROC 曲线分析[J]. 中国公共卫生,2003,19(9):1151-1152.
[23] 刘艺梁,殷坤龙,刘斌. 逻辑回归和人工神经网络模型在滑坡灾害空间预测中的应用[J]. 水文地质工程地质,2010,37(5):92-96.