# 基于 Pearson 卡方检验算法评价指标优选的波密–墨脱地区泥石流易发性评价

李群[1]，徐红剑[2]，杨金[1]，王林康[2]，孙靖宜[2]，章广成[2*]

1. 中交第二公路勘察设计研究院有限公司，武汉 430058；
2. 中国地质大学（武汉）工程学院，武汉 430074

第一作者，E-mail：35832308@qq.com

* 通信作者，E-mail：zhangguangc@cug.edu.cn

2024-03-13 收稿；2024-05-30 修回；2024-05-30 接受

基金项目：扎墨公路沿线地质灾害发育规律与工程影响评价研究

中图分类号：P642.23　　文章编号：2096-8523(2025)04-0316-14　　doi: 10.19509/j.cnki.dzkq.tb20240091

**摘　要：**西藏地区地貌单元复杂、地质构造活跃，为该地区泥石流提供了良好的孕育环境，也对人类生命财产构成了极大威胁，开展泥石流易发性评价可为地区防灾减灾明确重点区域。以西藏自治区波密县–墨脱县为研究区域，利用 Pearson 卡方检验算法优选出高程、坡度、地层岩性、降雨量等 12 个对泥石流影响较高的因素作为评价指标，以研究区 282 个泥石流点和非泥石流点为样本数据库，基于 ArcGIS 平台利用信息量法和机器学习方法建立了 4 种易发性评价模型，并引入接受者操作特性(ROC)曲线和 *AUC*(ROC 曲线下方的积分面积)指标对泥石流易发性精度进行评估，得到了该地区的泥石流易发性评价分区。研究表明：(1)考虑不同维度泥石流类型主控因素不同，采用纬度和气温相融合的归一化系数作为泥石流易发性评价指标，在一定程度上消除了低海拔地区泥石流对温度的过度响应。(2)气温、距水系距离、距道路距离、地层岩性、高程是研究区泥石流发生的主控因素；植被覆盖率、地形湿度、坡度等因素也发挥着重要作用。(3)考虑泥石流灾害点和影响因子分级属性关系，对影响因子各分级属性赋分，作为输入特征进行训练，机器学习模型预测效果较好，平均 *AUC* 值为 0.980，整体优于传统的信息量模型。(4)SVM 模型的 *AUC* 值高达 0.987，高易发区频率比率(*FR*)值为 41.13 且预测面积占比最小，具有在大尺度区域内进行高精度预测的能力。

**关键词：**卡方检验；评价指标优选；泥石流；易发性评价；高精度；信息量；机器学习；波密–墨脱地区

泥石流具有时间短、运动快、爆发突然、破坏性大等特点，是我国西部最重要的地质灾害类型，给国家及人民的生命财产带来了巨大损失[1]。西藏地区地形地貌多样、地质构造活动强烈，地质环境复杂，为泥石流灾害提供了良好的孕育环境。据统计，截至 2020 年，西藏自治区现存泥石流灾害点 4 593 处(数据来源：地理遥感生态网(http://www.gisrs.cn/))，区域内频发的泥石流对基础设施及人类活动造成了极大威胁，因此摸清区域泥石流的空间分布特征，并开展易发性评价工作至关重要，能够在一定程度上为泥石流灾害的防灾减灾工作提供参考，对保障人民生命财产安全、减少地区经济损失、促进人与自然和谐共处具有重要意义。

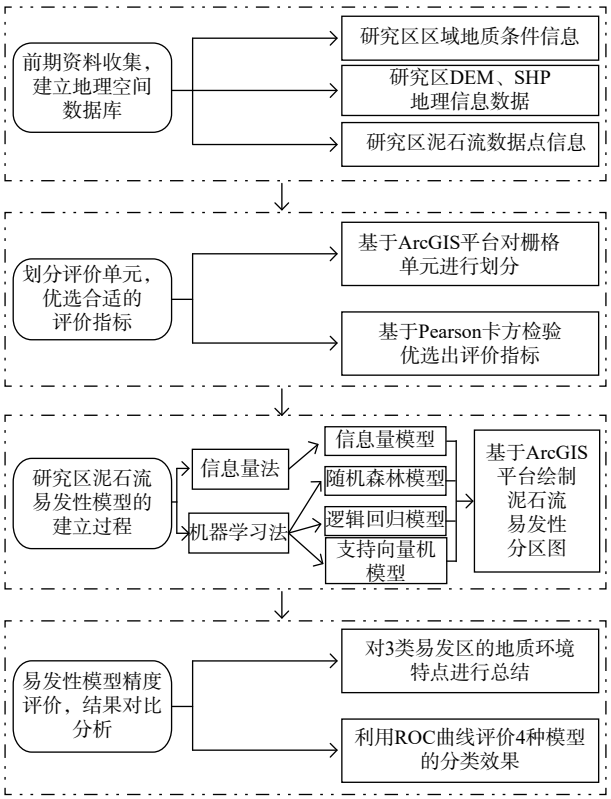
地质灾害易发性评价起源于 20 世纪 70 年代，BRABB 等[2] 以岩性和坡度 2 个因子为基础，依靠自身经验对 San-Mateo 地区滑坡易发性分区进行了评价；80 年代谭炳炎[3] 基于对大量泥石流案例的研究，建立了包含 3 项主要影响因素的反应矩阵，实现了

**著录信息：**李群，徐红剑，杨金，等. 基于 Pearson 卡方检验算法评价指标优选的波密–墨脱地区泥石流易发性评价 [J]. 地质科技通报，2025，44(4)：316-329.
LI Qun, XU Hongjian, YANG Jin, et al. Evaluation of debris flow susceptibility in Bomi-Motuo area using Pearson Chi-square test algorithm based indicator optimization[J]. Bulletin of Geological Science and Technology, 2025, 44(4): 316-329.

© Editorial Office of Bulletin of Geological Science and Technology. This is an open access article under the CC BY-NC-ND license.　https://dzkjtb.cug.edu.cn

对泥石流危害程度的综合定量评判。21 世纪以来，计算机技术和以 GIS(地理信息系统)、遥感为代表的空间信息技术广泛应用于地质灾害易发性评价中。XU 等[4] 基于 GIS 平台，通过计算高程、坡度等 7 个影响因子的信息量值实现了对四川省的泥石流易发性评价。XIONG 等[5] 基于 RS(遥感)和 GIS 对地质环境数据进行处理，采用多种机器学习算法对四川省内的泥石流进行了易发性评估。ZHU 等[6] 基于 SBAS-InSAR(小基线集雷达干涉测量)技术，通过构建相应的权变矩阵对已训练出的模型进行分类误差校正，得到了更为可靠的滑坡易发性模型。在地质灾害易发性评价研究工作中，评价因子的选取对评价结果至关重要。WU 等[7] 结合遥感影像、DEM(数字高程模型)数据、地质图等资料综合分析，选择了与区域泥石流相关性较大的 10 个影响因子作为评价指标。刘永垚等[8] 以模型训练后的 *AUC*(接受者操作特性曲线下与坐标轴围成的面积)变化为依据，基于随机森林筛选出贡献率较高的因素作为评价指标。评价模型选取方面，常见的方法主要有信息量法[9-10]、层次分析法[11]、模糊数学法[12]、聚类和最大似然法[13]、多元回归分析[14]、人工神经网络法[15-19] 等，这些方法各有优势和缺陷，将 2 种方法耦合后进行地质灾害易发性评价能够发挥所长并弥补缺陷[20-22]，在一定程度上提高评价精度。

笔者以西藏自治区林芝市波密和墨脱 2 个县级行政区所辖范围为研究区域，在收集前期资料的基础上，依据研究区泥石流灾害形成条件和影响因素，结合 Pearson 卡方检验算法优选出高程、坡度、地层岩性、温度、降雨量等 12 个因素作为泥石流易发性影响因子，并进行分级定量化；然后基于信息量法和机器学习方法，构建出 4 种泥石流易发性模型，对研究区泥石流易发性特点进行综合评价。技术路线框架如图 1 所示。




DEM. 数字高程模型; SHP. 地理信息系统(GIS)中常用的矢量数据格式; 下同

图 1　技术路线图

Fig. 1　Technical roadmap

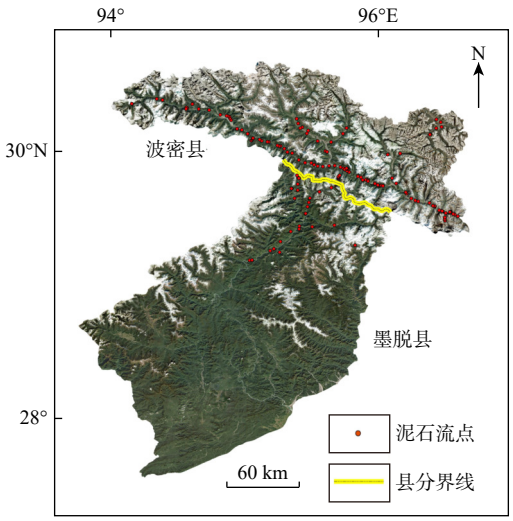



图 2　研究区范围及泥石流灾害点分布示意图

Fig. 2　Scope of the study area and distribution of debris flow disaster points

## 1　研究区概况及数据来源

### 1.1　研究区概况

研究区范围包含西藏自治区波密、墨脱全县区域，如图 2 所示，其中波密县位于西藏自治区东南部，帕隆藏布河北岸，地处 94°00′E ~ 96°40′E, 29°21′N ~ 30°40′N 之间，全县总面积约 $1.65\times10^4$ km$^2$，南与察隅县、巴宜区、墨脱县交界。县域范围内地势总体北高南低，群山连绵，县境内海拔 2 001 ~ 6 648 m，相对高差大，主要地貌类型为河流堆积、高山峡谷以及冰川等地貌[23]。县城区多年平均气温约为 8.5℃，历史极端高温 31℃，极端低温−20.3℃，昼夜温差较大，多年平均降水量为 977 mm，其中 3−10 月降水占全年 90% 以上。波密地区地层岩性种类繁多，年代分布较为广泛，主要包括侏罗纪、白垩纪、古近纪和新近纪时期形成的中酸性岩浆岩；从前寒武纪一直延续到二叠纪的沉积岩；以及寒武纪、奥陶纪、泥盆纪、

石炭纪生成的波密杂岩[24]。该地区地质构造较为复杂，位于雅江缝合带与怒江−班公错缝合带之间的冈底斯岛弧地块[25]，区域地质构造作用强烈。墨脱县位于西藏自治区东南部、雅鲁藏布江下游，地处青藏高原−喜马拉雅山脉东端南麓，山势险峻，地理位置介于 93°46′E ~ 96°05′E，27°34′N ~ 29°56′N 之间，全县总面积约 $3.11\times10^4$ km$^2$，海拔 665 ~ 6 648 m，山顶与河谷之间的相对高差一般在 2 000 m 以上，主要地貌类型有河谷堆积、冰川、高山峡谷和中−高山峡谷 4 类地貌，如图 3 所示；县境内地层岩性种类多样，主要包含震旦系蛇绿混杂岩[26]；石炭系砂岩、板岩、玄武岩、大理岩、变安山岩；白垩系花岗岩；新近系花岗岩、闪长岩；以及第四系堆积物等；全县年平均气温为 16℃，冬季温暖，夏无酷暑，空气湿度较高，降雨主要集中在 5−9 月，年平均降水量约 2 358 mm。墨脱县处于欧亚板块前锋地带，板块构造运动强烈活动断裂十分发育，多个深大断裂例如通麦−嘎隆寺断裂带、马尼翁断裂带等穿切该区域，加剧了区域内泥石流等各类地质灾害的发生。

### 1.2　数据来源

依据研究区泥石流灾害形成条件和影响因素，初步选取 16 个泥石流影响因子进行研究，数据来源见表 1。

## 2　研究方法

### 2.1　评价指标筛选方法

Pearson 卡方检验是一种强大的统计算法，该算法计算相对简单，易于理解，在实际应用中对数据是否符合正态分布并不做严格要求，适用于大样本分类问题的分析处理。当需要对分类问题中不同特征的重要性进行排序时，可以用该方法计算样本的实际例数与期望例数之间的偏离程度来判断特征和响




图 3　研究区主要地貌类型

Fig. 3　Primary landform types in the study area

应变量之间的关联性[27]，在对所有特征和响应变量依次进行分析结束后，按照重要性大小进行排序，以此完成特征筛选工作。对于单个特征：

$$R_{\text{all}}=\sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij},\quad P_i=\frac{\sum_{j=1}^{m}R_{ij}}{R_{\text{all}}},\tag{1}$$

式中：$n$ 为特征的类别数量；$m$ 为响应变量的类别数量；$R_{ij}$ 代表响应变量属于第 $j$ 个类别且特征属于第 $i$ 个类别的样本数量；$R_{\text{all}}$ 为样本总数量；$P_i$ 为第 $i$ 类特征对应样本的期望概率。

卡方检验统计量 $\chi^2$ 和自由度 $F$ 表达式为：

$$\chi^2=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(R_{ij}-P_i\sum_{k=1}^{n}R_{kj}\right)^2}{P_i\sum_{k=1}^{n}R_{kj}},\quad F=(n-1)\times(m-1)。\tag{2}$$

式中：$k$ 为第 $k$ 行；$R_{kj}$ 为响应变量属于第 $j$ 个类别且

**表 1　数据类型及来源**

Table 1　Data types and source

| 泥石流影响因子 | 数据获取 | 数据来源 |
|---|---|---|
| 泥石流点 | 西藏自治区地质灾害分布数据 | 地理遥感生态网（http://www.gisrs.cn/） |
| 高程、坡度、坡向、曲率、起伏度、地形湿度指数 | 西藏自治区30 m精度DEM数字高程数据 | 地理空间数据云（https://www.gscloud.cn/） |
| 土地利用类型 | 林芝市30 m精度土地覆盖数据 | 资源环境科学与数据中心（https://www.resdc.cn/） |
| 气温、降雨量 | 林芝市2011—2020年气象数据集 | 国家青藏高原科学数据中心（https://data.tpdc.ac.cn/home） |
| 地层岩性 | 西藏自治区地层岩性数据 | 地质云（https://geocloud.cgs.gov.cn/） |
| 距道路距离、距水系距离 | 西藏自治区1∶100万公众版地形数据 | 全国地理信息资源目录服务系统（https://www.webmap.cn） |
| 距断层距离 | 西藏自治区断层数据 | 中国地震局 |
| 人口密度 | 西藏自治区1 km人口分布栅格数据集 | LandScan全球人口动态数据（https://landscan.ornl.gov/） |
| 归一化植被指数 | 林芝市MOD13A3数据集 | 美国国家航空航天局（https://search.earthdata.nasa.gov/search） |

特征属于第 $k$ 个类别的样本数量。

在利用 Pearson 卡方检验解决分类问题时，首先假设特征和响应变量之间不相关，通过计算不同类别下样本实际例数与期望例数之间的偏离程度来确定卡方检验统计量(需要注意的是，Pearson 卡方检验对样本量的要求较为严格，对于任意属于第 $j$ 个类别且特征属于第 $i$ 个类别的样本，其期望例数的数量一般要大于 5，否则计算结果将不可靠)，结合自由度计算结果查阅卡方表，确定某特征和响应变量有关系的可能性。当检验统计量的值越大时，代表某特征不同类别的响应差异性越显著，即特征的重要性越高，最后通过对不同特征的重要性进行排序来实现特征的筛选。

## 2.2 易发性评价方法

(1)信息量法

信息量模型以信息论为理论基础，它是将现有的描述信息进行量化从而表征其对研究对象的影响的一种统计评价方法[28]。信息量的计算表达式为：

$$I(b,a_1,a_2,a_3,\cdots,a_n)=\mathrm{In}\frac{P(b,a_1,a_2,a_3,\cdots,a_n)}{P(b)}, \quad (3)$$

式中：$I(b, a_1, a_2, a_3, \cdots, a_n)$为因素 $a_1, a_2, a_3, \cdots, a_n$ 影响下泥石流的信息量值；$P(b, a_1, a_2, a_3, \cdots, a_n)$为因素 $a_1, a_2, a_3, \cdots, a_n$ 组合下发生泥石流的概率；$P(b)$为发生泥石流的概率。

在计算单因素对泥石流发生的影响时常使用频率对条件概率进行估算，通过计算每个因素的信息量值最后叠加得到所有因素组合条件下对泥石流提供的信息量值，其具体公式如下：

$$I=\sum_{i=1}^{n}I(i)=\sum_{i=1}^{n}\mathrm{In}\frac{E_i/E}{Z_i/Z}, \quad (4)$$

式中：$I$ 为总信息量值；$I(i)$ 为单因素信息量值；$E_i$ 为研究区内第 $i$ 因素所包含的泥石流数；$E$ 为研究区泥石流总数；$Z_i$ 为研究区内第 $i$ 因素所占评价单元数；$Z$ 为研究区评价单元总数。

(2)随机森林

随机森林(random forest，简称 RF)是一种强大的多分类器集成学习算法，由多棵决策树组成[29]，每棵决策树都是独立训练。在构建每棵决策树时，若原始数据集 $D$ 中共有 $w$ 个样本，随机森林算法会对训练集进行 Bootstrap 采样，从总训练集中抽取 $z(z<w)$个样本，即有放回地随机抽取样本来生成不同的训练数据集。在单个决策树的训练过程中，随机森林通常采用 *Gini* 系数进行节点的不纯度评价，如式(5)所示：

$$Gini_r^j=1-\sum_{i=1}^{k}P_{ij}^2, \quad (5)$$

式中：$Gini_r^j$ 为第 $r$ 颗决策树中的第 $j$ 个决策节点；$k$ 为数据的总类别数量；$P_{ij}$ 为第 $j$ 个决策节点的分类结果中，类别 $i$ 的数量占总类别的比例。*Gini* 系数越小代表该决策节点分出的样本纯度越高，分类效果越好。

对 $r$ 颗决策树重复上述步骤，将训练好的决策树结合起来即生成随机森林，其输出结果如式(6)所示。

$$VOTE(x)=\frac{\sum_{r'=1}^{r}H(x)_{r'}}{r}, \quad (6)$$

式中：$x$ 为输入的测试数据；$H(x)_{r'}$为第 $r'$颗决策树的分类结果；$VOTE(x)$ 为所有决策树的平均投票结果。

这种随机性能够增加不同决策树彼此之间的差异性，避免了 RF 训练过程中的模型过拟合问题。

(3)逻辑回归

逻辑回归(logistic regression，简称 LR)是一种用于解决二分类或多分类问题的经典线性回归分析方法[30]，通过建立响应变量和特征变量之间的关系，并输出一个 [0, 1] 的概率值，完成对某一事件发生概率的预测。由于在泥石流易发性评价中泥石流是否发生为二分类问题，因此本研究所述的逻辑回归分析均指二元逻辑回归分析。

LR 的基本假设是，输入特征向量 $\boldsymbol{A}$ 和输出标签 $\boldsymbol{B}$ 之间存在线性关系。假设有训练样本有 $h$ 个特征，则模型可以表示为：

$$\boldsymbol{B}=\theta^{\mathrm{T}}\boldsymbol{A}, \quad (7)$$

式中：$\theta$ 为模型参数(权重)；T 为列向量；$\theta^{\mathrm{T}}$ 为其转置；$\boldsymbol{A}=(1, A_1, A_2, A_3, \ldots, A_h)$；$\theta=(\theta_0, \theta_1, \theta_2, \theta_3, \ldots, \theta_h)$，$\theta_0$ 为常数项。

在此基础上引入 Sigmoid 函数 $H_\theta(\boldsymbol{A})$，这样就能使得输出标签的值一定会落在 0 和 1 之间。然后再定义一个损失函数来评估模型的预测结果是否准确可靠，逻辑回归通常使用对数损失函数来衡量模型预测响应值与实际响应值之间的差异。对于单个样本对数损失函数的表达式如下：

$$J(\theta)=\frac{1}{v}\left[\sum_{i=1}^{v}-\boldsymbol{B}^i\ln(H_\theta(\boldsymbol{A}^i))-(1-\boldsymbol{B}^i)\ln(1-H_\theta(\boldsymbol{A}^i))\right], \quad (8)$$

式中：$J(\theta)$ 为对数损失函数；$v$ 为训练样本数量；$\boldsymbol{A}$ 为输入特征向量；$\boldsymbol{B}$ 为输出标签。

最后使用梯度下降法不断调整 $\theta$，直至收敛到

局部最优解，达到最小化损失函数、得到最佳模型的目的。

（4）支持向量机

支持向量机（support vector machine，简称 SVM）是一种可用于解决分类或回归问题的监督学习算法，其主要目标是通过构建最优超平面来实现对数据的最大间隔分类，从而提高分类准确性。

SVM 的核心思想是利用维度转化函数将原维度数据进行升维转换，再根据高维度下的决策边界超平面来对需要预测的数据进行分类和判断。具体来说，对于线性不可分的数据，SVM 通过引入合适的核函数对其进行非线性变换处理，使得低维度空间中难以分类的数据在高维空间中变得线性可分。在高维度下求解分割超平面时，问题即转化为在约束条件下目标函数的优化问题，通过引入惩罚因子和松弛变量来平衡分类间隔和误分类点的影响[31]，从而得到一个较好的分类模型。判别函数的表达式为：

$$F(x)=sgn\left(\sum_{i=1}^{f}L_iy_iK(g,g_i)+q\right), \tag{9}$$

式中：$L_i$ 为支持向量对应的拉格朗日乘子；$y_i$ 为第 $i$ 个训练样本的真实标签；$K(g,g_i)$ 为核函数；$q$ 为偏置项。

核函数中的参数对模型的影响尤为重要，不同类型的核函数直接影响到 SVM 模型的准确性和性能。

### 2.3　模型精度评价方法

（1）ROC 曲线法

ROC 曲线是一种用于评估二分类模型性能的常用方法，将灵敏度与特异性以图示方法结合，可以直观地体现预测精度。该曲线展示了分类模型在不同阈值条件下的真正例率（*TPR*）与假正例率（*FPR*）之间的关系，*TPR* 和 *FPR* 的计算公式如式（10）、（11）所示：

$$TPR=\frac{TP}{TP+FN}, \tag{10}$$

$$FPR=\frac{FP}{FP+TN}, \tag{11}$$

式中：*TP* 为真正例，即实际为正例，被预测为正例的数量；*FN* 为假负例，即实际为正例，被预测为负例的数量；*FP* 为假正例，即实际为负例，被预测为正例的数量；*TN* 为真负例，即实际为负例，被预测为负例的数量。

*AUC* 代表 ROC 曲线下方的积分面积，当 *AUC* 的值越接近 1 时，表示模型性能越好；越接近 0.5 时，表示模型的性能越一般[32]。通过观察 ROC 曲线及其下的 *AUC* 值，可以评估模型在不同阈值条件下的特异性和敏感度，以及模型分类效果的好坏。

（2）频率比法

频率比法基于地质灾害分布与诱灾因子之间的关系，通过数理统计方法表示地质灾害与易发性评价之间的相关程度，频率比值 *FR* 的定义式为：

$$FR_t=\frac{E_t/E}{S_t/S}, \tag{12}$$

式中：$FR_t$ 表示第 $t$ 种易发性等级的频率比值；$E_t$ 表示第 $t$ 种易发性等级的滑坡数量；$E$ 表示研究区滑坡总数；$S_t$ 表示第 $t$ 种易发性等级的栅格数量；$S$ 表示研究区栅格总数。*FR* 以 1 为界限，其值越大地质灾害与易发性等级的相关性越强，易发性评价精度越高。

## 3　泥石流易发性模型构建

### 3.1　评价单元划分

研究区评价单元是进行数据处理分析和各类图层制作的最基本单元，其划分方式和划分大小是区域泥石流易发性评价的基础。目前评价单元选取方式较多，其中区域泥石流评价单元主要有栅格单元、地貌单元、流域单元、行政单元等[33]。栅格单元的优点在于其可以将地域快速划分为规则的方形网格，对每个网格赋予不同的值，将数据以矩阵的形式保存，在此基础上通过计算机进行运算处理时十分高效快捷[34]；此外，张以晨等[35]、李益敏等[36] 等提出在易发性评价工作中栅格单元更适用于研究范围较大且比例尺较小的区域，考虑到本研究区总面积达 $4.76\times10^4$ km$^2$，研究范围较大，综合考虑后采用栅格单元作为研究区的基本评价单元，基于 WGS1984 投影坐标系，利用 ArcGIS 软件将各要素图层转化为 30 m×30 m 大小的栅格数据。

### 3.2　评价指标选取

泥石流的成因较为复杂，通过参考泥石流易发性相关文献 [37-39]，以研究区泥石流孕育机制、发育现状、分布情况为基础，结合泥石流发育的 3 个必备条件（充足的水源、陡峭的地形、丰富的物质来源），初步选取了高程、坡度、坡向、起伏度、平面曲率、剖面曲率、土地利用类型、地层岩性、距水系距离、距道路距离、距断层距离、归一化植被系数、地形湿度指数、人口密度、气温、降雨量共 16 个要素作为研究区泥石流评价指标因素。

考虑到研究区内绝大多数泥石流类型为冰川型泥石流，因此需要对温度这一评价指标进行额外分

析和处理。据研究区 2011—2020 年共 10 a 内的气温数据显示，每年 11–次年 2 月份月平均最低气温均在 0℃ 以下，此阶段冰雪大量堆积，而到了 3 月，温度逐步上升引起冰雪融化，融水将浸润岩土体并形成地表径流，当地形和物源条件合适时，极易诱发冰川型泥石流。因此本研究主要就 3 月份气温对泥石流的显著影响进行研究，统计了研究区近 10 a 的 3 月份气温栅格数据，通过 ArcGIS 地图代数功能得到近 10 a 的 3 月平均气温，并将其分为 4 个级别。

由高程和气温数据可知，研究区海拔与纬度呈正相关，与气温呈负相关。气温与高海拔地区泥石流具有很强的相关性，在相同的气温（例如 0 ~ 1℃）条件下，海拔较高的波密地区实际存在大量的泥石流灾害点，这是因为冰川型泥石流对温度极其敏感，在此温度下高寒地区冰雪开始融化；而在海拔较低的墨脱地区，此区域内受温度变化诱发的泥石流灾害点较少，温度不是该地区泥石流发生的主控因素，因此不能直接将气温作为整个研究区的评价指标。为此本研究基于海拔、气温与纬度三者之间的关系，将气温和纬度栅格数据做归一化计算处理，处理过程如下：通过 ArcGIS 软件获取研究区各栅格单元的气温数值与纬度数值，采用 Min-Max 数据标准化方法分别对气温数据与纬度数据进行归一化处理，气温数值由高到低为 0 ~ 1，纬度数值由低到高为 0 ~ 1，将栅格内 2 数值相乘获得了气温与纬度耦合的影响系数，称该系数为气温/纬度归一化系数，当气温较低且纬度较高时，气温/纬度归一化系数越大。用该系数作为后续泥石流易发性评价指标，将其分为 6 个级别，旨在突出高海拔地区泥石流对温度的敏感型，修正了低海拔地区泥石流易发性评价对温度的过度响应，气温处理前后对比见图 4。

在泥石流易发性特征集合的建立过程中，许多学者往往通过经验和自身判断选取相应的影响因子，这样的做法不仅主观性较强，且未必能在后续的模型训练过程中取得较为满意的效果。因此本研究利用 Pearson 卡方检验算法，对初步选取的 16 个影响因子进行重要性评估，将重要性系数较低的影响因子进行剔除，以提升后续模型训练中输入特征和输出标签之间的相关性。基于 Pearson 卡方检验算法的泥石流影响因子重要性排序如图 5 所示，其中气温、距道路距离、距水系距离、地层岩性、高程的重要性分数相对较高，是研究区泥石流发生的主控因素。

气温与高寒地区冰川型泥石流的爆发紧密关联，每年 3–4 月份高寒地区冰雪消融，融水将直接浸润岩土体并形成地表径流，为泥石流的爆发提供了必要的水源条件。由图 5 可知，将气温与纬度归一化处理后其重要性分数得到大幅度提升，突显了高海拔地区泥石流对温度的敏感型，消除了低海拔地区泥石流对温度的过度响应。

道路属于人类工程活动范畴，在道路修建过程中大量岩土体被挖走，改变了坡体内部的应力分布状态，在不利条件下坡体容易失稳，为泥石流的爆发提供了动力和物源条件。距道路距离越近，发生泥石流的可能性越高。

水系通常存在于地势低洼的河谷地带，附近地表松散物质较多，极易受到水流的冲刷和侵蚀，同时

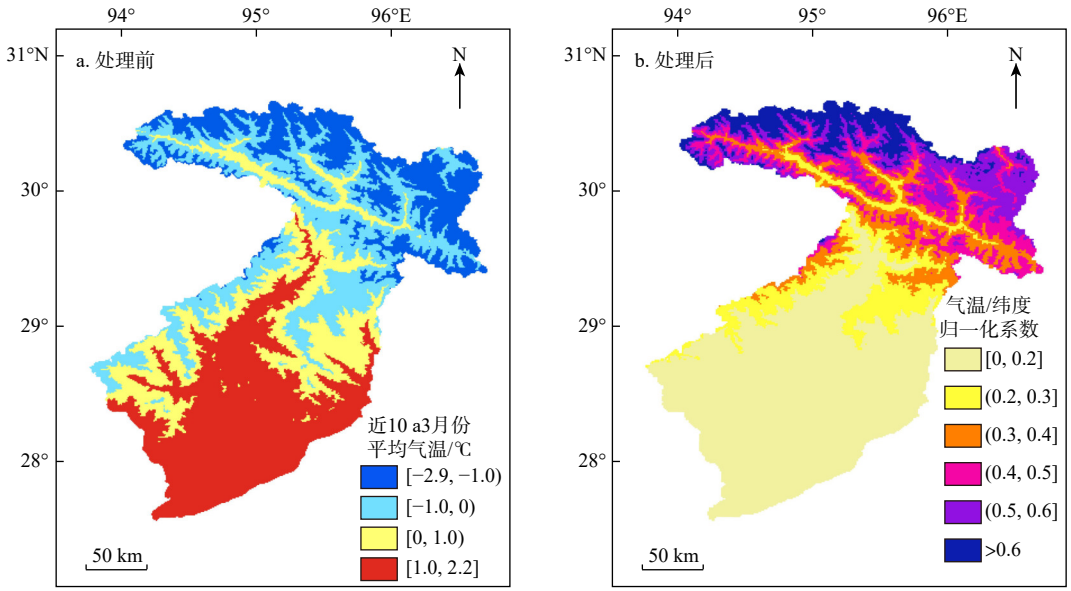



图 4　气温与纬度归一化处理

Fig. 4　Normalization of temperature and latitude

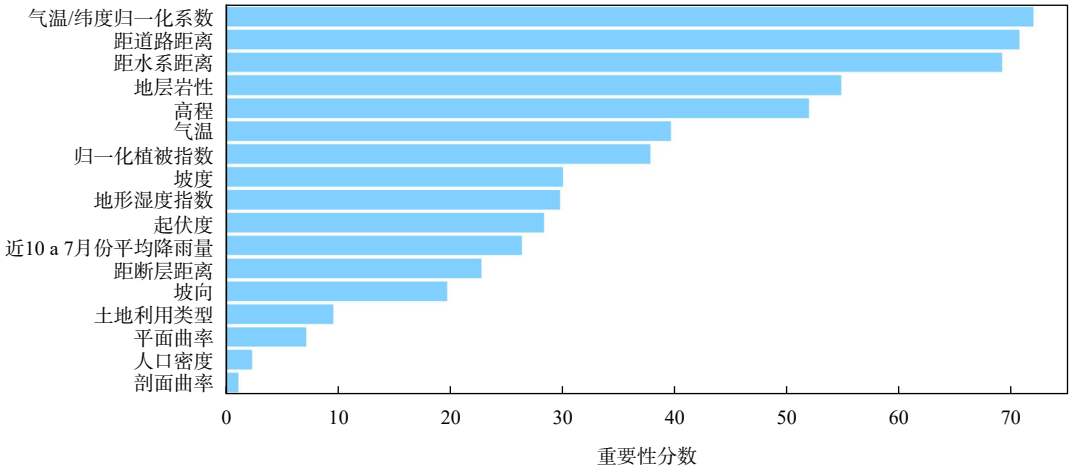



图 5　基于 Pearson 卡方检验算法的泥石流影响因子重要性排序

Fig. 5　Importance ranking of debris flow impact factors based on Pearson Chi-square test algorithm

水系通道的存在为泥石流补充了充足的水源条件。距水系距离越近，发生泥石流的可能性越高。

地层岩性的差异决定了泥石流物源条件的充足与否。通常来说软岩、极软岩其岩体本身强度较差，更容易遭受风化侵蚀而松散破碎，岩土体的松散破碎程度决定了泥石流物源是否充足，进而影响泥石流灾害的规模。

高程对泥石流的爆发有重要影响，研究区位于喜马拉雅山东端，山势险峻，最高海拔与最低海拔高差可达 5 000 m，极大的落差为泥石流提供了巨大的潜在势能。随着高程增加，岩体的风化程度加剧，为崩滑体的形成提供了更为有利的条件。而崩塌、滑坡等地质灾害的发生恰巧为泥石流提供了良好的物源条件，还能在一定条件下转化为泥石流。同时不同的海拔高度具有不同的降雨量、植被覆盖率，进而对泥石流的发育产生影响。

结合易发性模型预构建效果，最终将重要性分数最低的 4 种影响因子剔除，优选出了高程、坡度、坡向、起伏度、归一化植被指数、地层岩性、距水系距离、距道路距离、距断层距离、地形湿度指数、气温/纬度归一化系数和近 10 a 7 月平均降雨量共 12 个影响因子作为研究区泥石流评价指标，各评价指标状态分级如图 6 所示。

### 3.3　基于信息量法的泥石流易发性模型

评价指标体系建立完成后，根据信息量模型基本原理，首先对各评价指标进行单因素信息量值计算，后将各指标值进行叠加处理。通过地理遥感生态网获取西藏自治区地质灾害分布数据，从中提取出研究区泥石流灾害点共 141 个，选取泥石流沟口中心处作为泥石流样本点位置；基于 ArcGIS 空间分析运算，使用“提取分析”工具将 12 个指标因素以分级状态为标准提取至 141 个泥石流灾害点数据表中，获取研究区内分布在各类因素的泥石流单元数；研究区内分布在各类因素的评价单元数及研究区评价单元总数可根据各指标状态分级栅格获取；研究区泥石流单元总数为 141 个，根据信息量模型公式可计算各指标因素的信息量值（表 2），通过栅格叠加分析计算 12 个指标因素的信息量总值，将计算结果依据统计学方法分为高、中、低共 3 类易发区。

### 3.4　基于机器学习方法的泥石流易发性模型

结合前文的各评价指标状态分级图（图 6）、信息量统计结果、泥石流样本数据点的分布情况及野外工程地质调查情况，对所选用的 12 个指标因素进行分级赋分评估，最终完成对所有指标赋予 1 ~ 4 分的分级得分[40]，见表 2。

通过前期的资料搜集工作可知，研究区目前共有 141 个泥石流点，为遵循等量原则，在 ArcGIS 平台中利用“创建随机点”工具在研究区范围内随机生成 141 个非泥石流点，使其均匀分布于研究区，二者组成样本数据库。接着使用“多值提取至点”工具将每个栅格点对应的 12 个影响因子对应的分级得分提取出来用于模型训练。在实际训练过程中，可将泥石流点对应的输出标签定为 1，将非泥石流点的输出标签定为 0，将 12 个影响因子作为模型训练的输入变量；随后将 282×13 列的数据导入 Matlab 平台，选取其中 80% 的数据作为训练集进行模型训练，剩余 20% 的数据作为测试集用于模型的测试。以逻辑回归算法为例，将训练集数据导入后，设置合适的学习率和迭代次数，在每次迭代过程中，首先计算损失函数关于模型参数的梯度（即偏导数），然后根据

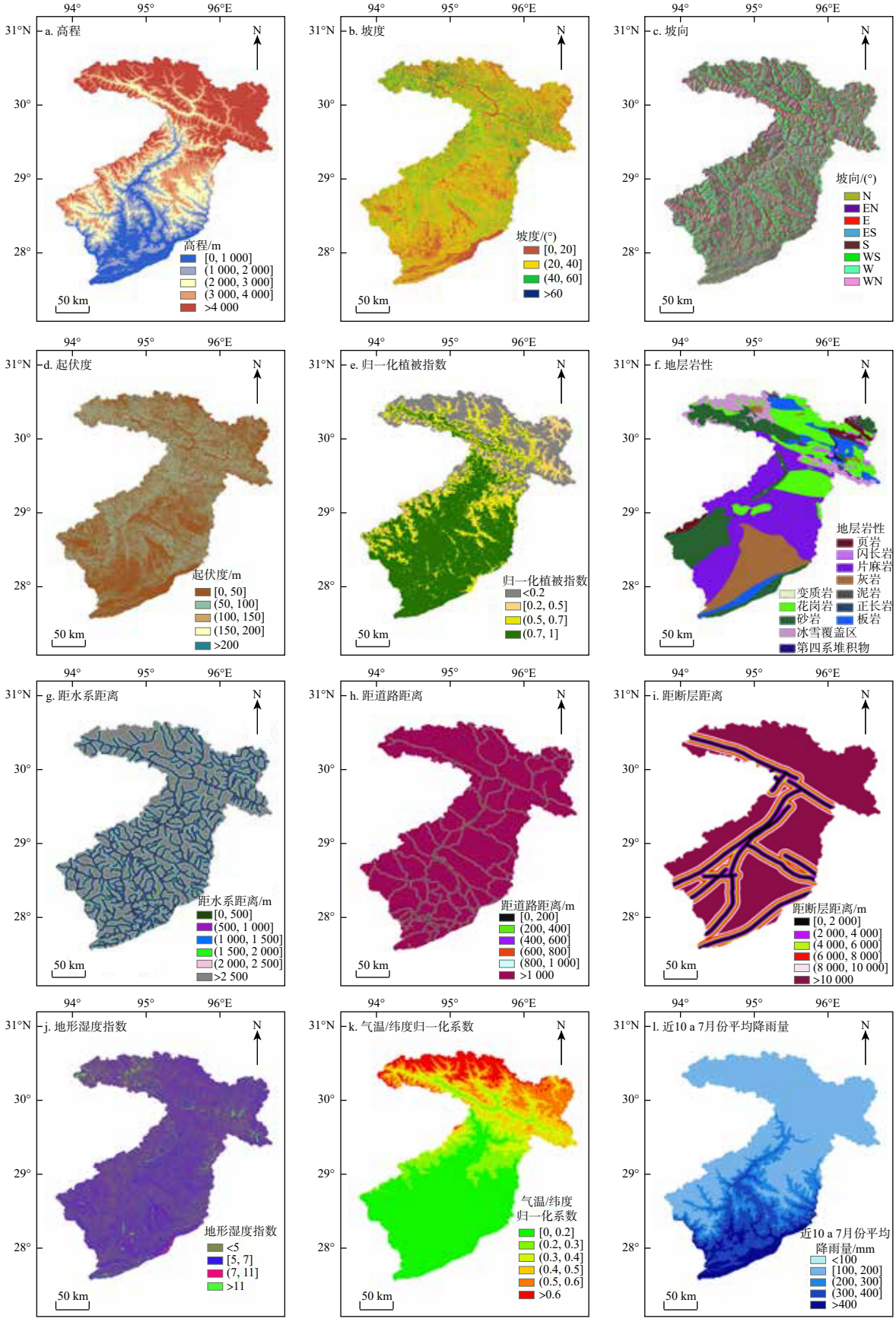



图 6　各评价指标状态分级图

Fig. 6　State classification diagram of each evaluation index

**表2　各指标因素信息量值与分级得分汇总**

Table 2　Information value and graded scores of each indicator factor

| 指标因素 | 分级 | 信息量值 | 分级得分 | 指标因素 | 分级 | 信息量值 | 分级得分 |
|---|---|---|---|---|---|---|---|
| 高程/m | [0，1000) | −1.212 | 1 | 坡度/(°) | [0，20] | 0.854 | 4 |
| | (1000，2000] | −0.483 | 2 | | (20，40] | −0.593 | 2 |
| | (2000，3000] | 1.015 | 4 | | (40，60] | −1.485 | 1 |
| | (3000，4000] | 0.485 | 3 | | >60 | 0.766 | 4 |
| | >4000 | −2.656 | 1 | 地层岩性 | 页岩 | 0.000 | 1 |
| 坡向/(°) | N[0，22.5)，[337.5，360) | −0.520 | 1 | | 闪长岩 | 0.183 | 1 |
| | EN[22.5，67.5) | 0.021 | 2 | | 花岗岩 | 0.850 | 1 |
| | E[67.5，112.5) | −0.151 | 1 | | 第四系堆积物 | 2.596 | 4 |
| | ES[112.5，157.5) | −0.069 | 2 | | 砂岩 | 0.043 | 1 |
| | S[157.5，202.5) | 0.262 | 3 | | 片麻岩 | −0.442 | 2 |
| | WS[202.5，247.5) | 0.647 | 4 | | 灰岩 | 0.000 | 1 |
| | W[247.5，292.5) | −0.458 | 1 | | 泥岩 | 1.160 | 3 |
| | WN[292.5，337.5) | −0.433 | 1 | | 正长岩 | 0.000 | 1 |
| 起伏度/m | [0，50] | 0.298 | 3 | | 板岩 | 0.425 | 1 |
| | (50，100] | −0.966 | 1 | | 变质岩 | 0.000 | 1 |
| | (100，150] | −1.356 | 1 | | 冰雪覆盖区 | −1.560 | 1 |
| | (150，200] | 0.000 | 1 | 气温/纬度归一化系数 | [0，0.2] | −1.217 | 2 |
| | >200 | 1.253 | 4 | | (0.2，0.3] | 1.294 | 4 |
| 近10 a 7月份平均降雨量/mm | <100 | 0.000 | 3 | | (0.3，0.4] | 1.253 | 4 |
| | [100，200] | 0.334 | 4 | | (0.4，0.5] | −0.395 | 3 |
| | (200，300] | −0.471 | 2 | | (0.5，0.6] | 0.000 | 1 |
| | (300，400] | −1.660 | 1 | | >0.6 | 0.000 | 1 |
| | >400 | 0.000 | 1 | 地形湿度指数 | <5 | −0.905 | 1 |
| 距水系距离/m | [0，500] | 1.532 | 4 | | [5，7] | −0.332 | 1 |
| | (500，1000] | 0.300 | 3 | | (7，11] | 0.436 | 2 |
| | (1000，1500] | −0.628 | 2 | | >11 | 2.109 | 4 |
| | (1500，2000] | −1.742 | 1 | 归一化植被指数 | <0.2 | 0.000 | 1 |
| | (2000，2500] | −2.717 | 1 | | [0.2～0.5] | 0.928 | 3 |
| | >2500 | 0.000 | 1 | | (0.5～0.7] | 1.048 | 4 |
| 距断层距离/m | [0，2000] | 0.915 | 4 | | (0.7，1] | −0.702 | 1 |
| | (2000，4000] | 0.100 | 2 | 距道路距离/m | [0，200] | 2.810 | 4 |
| | (4000，6000] | −0.117 | 1 | | (200，400] | 1.553 | 4 |
| | (6000，8000] | −0.355 | 1 | | (400，600] | 0.357 | 2 |
| | (8000，10000] | −0.379 | 1 | | (600，800] | 0.716 | 3 |
| | >10000 | −0.191 | 1 | | (800，1000] | −0.653 | 1 |
| | | | | | >1000 | −1.567 | 1 |

梯度方向更新模型参数，使得损失函数逐渐减小。这样反复迭代，直到损失函数收敛或达到最大迭代次数，就得到了逻辑回归模型的最优参数。

本研究其余机器学习算法（RF、SVM）皆采用相似的流程，首先使用训练数据，通过不同的算法学习如何将输入的数据精确地映射到目标标签，然后通过参数优化来最小化预测误差。模型训练完成后将研究区的总栅格点导入训练好的模型中进行预测，得到研究区总栅格点的泥石流易发性指数，在ArcGIS平台中利用“点转栅格工具”绘制泥石流易发性图层，并将计算结果依据统计学方法分为高、中、低共3类易发区。

## 4　易发性评价结果分析与精度评价

### 4.1　结果分析

本研究基于4种不同方法建立了对应的泥石流易发性模型，并将研究区所有栅格的信息导入训练

好的模型中进行预测，得到每一个栅格的泥石流易发性指数，绘制出不同模型下泥石流易发性分区图，如图 7 所示。

由图 7 可知：①泥石流高易发区主要集中在地形极为陡峭、靠近河谷中心的地区，大体以“一横一纵”的形态分布于研究区的北部，其中嘉黎断裂、察隅断裂、墨脱断裂等主要断裂带穿切该区域，使得区域内地形切割深度较大，相对高差可达 1 500 ~ 2 500 m；区内岩性以砂岩、板岩、花岗岩为主，在强烈的构造作用影响下，岩体结构面发育，极易遭受风化侵蚀而松散破碎，为泥石流的发生提供了充足的物源条件；同时该区域途径易贡藏布、波都藏布、帕隆藏布、嘎弄曲等雅江主要支流，大量的水系通道为泥石流的形成提供了充足的水源条件；加之北部地区气温普遍较低，冰雪大量堆积，当气温回升冰雪融化时，极易诱发泥石流。在多种因素的综合作用下使得该区域泥石流爆发极为频繁和强烈。②研究区泥石流中易发区主要分布于上述河谷外围的山区地带，相对高差在 500 m 左右，其中在墨脱县格当乡的分布最为集中。区域内水系较为发育，有多条雅江支流流经；由遥感影像资料可知，区域内冰雪覆盖面积较大，在合适的条件下，冰雪融水会浸润岩土体，形成地表冰雪融水径流，可能诱发冰川型泥石流。同时该区域内还分布着少量滑坡、崩塌等其他不良地质现象，在特定条件下可能转化为泥石流灾害。③研究区泥石流低易发区主要集中在墨脱县南部低海拔区域以及波密县北部高海拔区域，区域内地势平缓开阔、地表植被覆盖率较高，地层岩性以片麻岩、灰岩、砂岩、花岗岩为主，相较于高易发区，整体上岩性更为稳定，难以风化侵蚀，物源条件相对匮乏，同时历史上有泥石流发生的记载极少，总体来说，此区域内相对缺乏泥石流孕育的基础环境条件。

### 4.2 模型精度评价

在前文基础上将 282 个泥石流和非泥石流数据点作为评估数据库，利用 ROC 曲线法对每个模型的二分类性能进行评估，绘制出 ROC 曲线如图 8 所示。

在 ArcGIS 平台中提取各易发性分区的栅格数量，对不同模型下不同泥石流易发等级的面积占比进行统计并计算各分区频率比值，以频率比值表示泥石流发生与易发性分区的相关程度，统计结果如表 3 所示。

通过对比图 8 中 ROC 曲线及表 3 中频率比值可知，4 种泥石流灾害易发性模型均展现出较好的效果（*AUC* 值均大于 0.9，*FR* 值均大于 1），其中机器学习模型的整体训练效果更好，相比之下，传统的信息量模型预测结果则相对较差，结合图 7 可知，信息量模型并未在泥石流频发的扎墨公路段展现出高易发性，并且在墨脱县南部预测出了较多的中易发区，这与事实有所偏差，原因可能与信息量法自身的计算原理有关，由于大量的泥石流灾害点聚集于波密县，同时两地的地形地貌、气候条件等部分影响因子差别较大，导致计算出的部分墨脱县的影响因子对应的分级指标信息量值较小，使得模型在预测时会有部分误差。

在 4 种模型中分类效果最好的为 RF 模型、LR 模型以及 SVM 模型（*AUC* 值均大于 0.97），虽然三者的分类性能指标极为优异且较为接近，但是结合对不同模型下泥石流易发等级面积占比及频率比值统计结果可知，SVM 模型在高易发区 *FR* 值为 41.13，高于 LR、RF 及 IM 模型，且 SVM 模型对高易发区的预测面积占比最小，这也表明该类模型具有在大尺度区域范围内进行高精度预测的能力，原因在于 SVM 模型能够更好处理线性不可分的数据，通过引入合适的核函数对其进行非线性变换处理，使得低维度空间中难以分类的数据在高维空间中变得线性可分，从而提高对泥石流易发性的预测能力。因此，此类模型能够在今后为工程勘察人员在大尺度区域

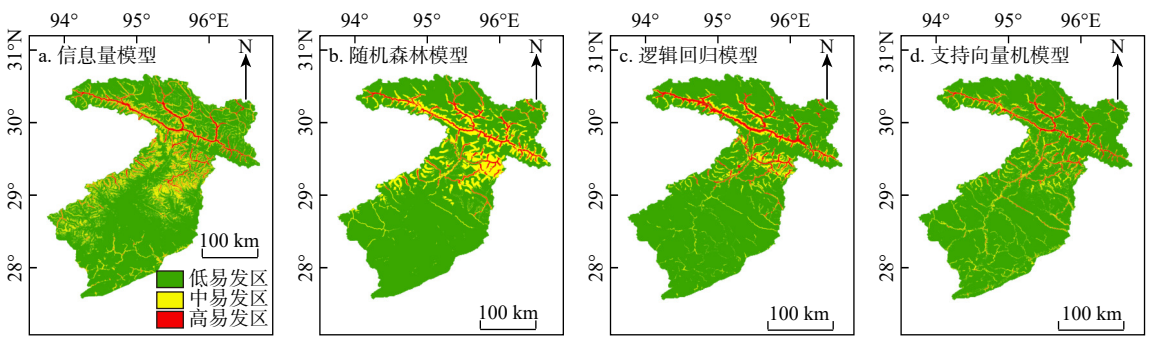



图 7　4 种易发性模型分区图

Fig. 7　Partition diagram of four susceptibility models

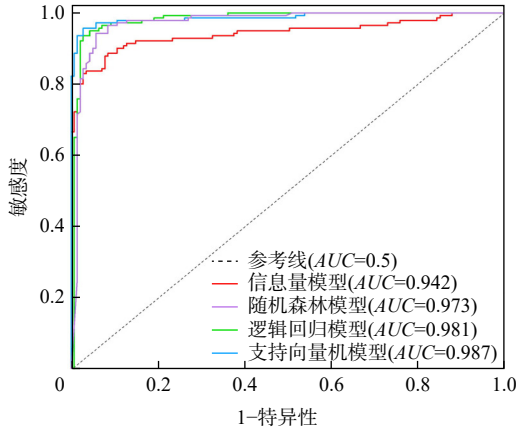



图 8　不同泥石流易发性模型下的 ROC 曲线

Fig. 8　ROC curves under different debris flow susceptibility models

**表 3　不同模型下泥石流易发等级面积占比及频率比值统计结果**

Table 3　Statistical results of area proportion and frequency ratio of debris flow prone grade under different models

| 统计类型 | 模型方法 | 易发性等级 | | |
|---|---|---|---|---|
| | | 低 | 中 | 高 |
| 面积占比/% | 信息量(IM) | 77 | 18 | 5 |
| | 随机森林(RF) | 82 | 15 | 3 |
| | 逻辑回归(LR) | 87 | 9 | 4 |
| | 支持向量机(SVM) | 89 | 9 | 2 |
| 频率比值（*FR*） | 信息量(IM) | 0.10 | 0.91 | 15.18 |
| | 随机森林(RF) | 0.04 | 1.13 | 26.48 |
| | 逻辑回归(LR) | 0.05 | 1.65 | 20.21 |
| | 支持向量机(SVM) | 0.07 | 1.26 | 41.13 |

内开展高精度泥石流易发性和风险评估工作提供更为有效的参考。

## 5　结　论

(1)考虑到研究区内的泥石流灾害类型以冰川型泥石流为主，存在少量的降雨型泥石流，因此在模型训练前对温度这一影响因子进行了特殊处理。用纬度和气温共同决定的归一化系数作为后续泥石流易发性评价指标，旨在突出高海拔地区泥石流对温度的敏感性，极大程度上消除了低海拔地区泥石流对温度的过度响应。

(2)基于 Pearson 卡方检验算法对初步选取的 16 个影响因子进行了重要性评估，将重要性分数较低的影响因子进行剔除，提升了后续模型训练过程中输入特征和输出标签之间的相关性。其中气温、距道路距离、距水系距离、地层岩性、高程是研究区泥石流发生的主控因素；植被覆盖率、地形湿度、坡度等因素也发挥着重要作用。

(3)在机器学习模型训练过程中，没有采用常规的手段对每个影响因子进行分类后即作为输入特征，而是在考虑泥石流点和影响因子分级属性之间关系的基础下，对影响因子各个分级属性赋予相应得分，然后作为输入特征进行训练，在此基础上训练出的机器学习模型预测效果较好，且整体优于传统的信息量模型。

(4)4 种泥石流灾害易发性模型均展现出较好的预测效果(*AUC* 值均大于 0.9)，相较于其余模型，SVM 模型对高易发区的预测面积占比最小，*AUC* 值及高易发区 *FR* 值最大，对泥石流频发的扎墨公路段的预测效果也更为优异，这表明该类模型具有在大尺度区域范围内进行高精度预测的能力。因此，此类模型能够在今后为工程勘察人员在大尺度区域范围内开展高精度泥石流易发性评估工作提供更为有效的参考。

所有作者声明不存在利益冲突。

The authors declare that no competing interests exist.

**参考文献：**

[1] MY E A. Debris flow susceptibility mapping in a portion of the Andes and Preandes of San Juan, Argentina using frequency ratio and logistic regression models[J]. Earth Science Research Journal, 2013, 17(2): 159-167.

[2] BRABB E E, PAMPEYAN E H, BONILLA M G. Landslide susceptibility in San Mateo County, California[J]. Miscellaneous Field Studies Map, 1972.

[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.
TAN B Y. Quantified comprehensive evaluation for the scope and intensity of mud-rock flow gully activity[J]. Journal of the China Railway Society, 1986, 8(2): 74-82. (in Chinese with English abstract)

[4] XU W B, YU W J, JING S C, et al. Debris flow susceptibility assessment by GIS and information value model in a large-scale region, Sichuan Province (China)[J]. Natural Hazards, 2013, 65(3): 1379-1392.

[5] XIONG K, ADHIKARI B R, STAMATOPOULOS C A, et al. Comparison of different machine learning methods for debris flow susceptibility mapping: A case study in the Sichuan Province, China[J]. Remote Sensing, 2020, 12(2): 295.

[6] ZHU Z F, GAN S, YUAN X P, et al. Landslide susceptibility mapping with integrated SBAS-InSAR technique: A case study of Dongchuan district, Yunnan (China)[J]. Sensors, 2022, 22

(15)：5587.

[7] WU S E, CHEN J, ZHOU W, et al. A modified logit model for assessment and validation of debris-flow susceptibility[J]. Bulletin of Engineering Geology and the Environment, 2019, 78(6): 4421-4438.

[8] 刘永垚, 第宝锋, 詹宇, 等. 基于随机森林模型的泥石流易发性评价：以汶川地震重灾区为例 [J]. 山地学报, 2018, 36(5): 765-773.
LIU Y Y, DI B F, ZHAN Y, et al. Debris flows susceptibility assessment in Wenchuan earthquake areas based on random forest algorithm model[J]. Mountain Research, 2018, 36(5): 765-773. (in Chinese with English abstract)

[9] 温鑫, 范宣梅, 陈兰, 等. 基于信息量模型的地质灾害易发性评价：以川东南古蔺县为例 [J]. 地质科技通报, 2022, 41(2): 290-299.
WEN X, FAN X M, CHEN L, et al. Susceptibility assessment of geological disasters based on an information value model: A case of Gulin County in Southeast Sichuan[J]. Bulletin of Geological Science and Technology, 2022, 41(2): 290-299. (in Chinese with English abstract)

[10] 吴雅睿, 娄春辉, 侯龙君, 等. 基于信息量与机器学习耦合模型的滑坡易发性评价对比分析 [J]. 西安科技大学学报, 2024, 44(6): 1140-1153.
WU Y R, LOU C H, HOU L J, et al. Comparative analysis of landslide susceptibility evaluation based on the coupling model of information quantity and machine learning[J]. Journal of Xi'an University of Science and Technology, 2024, 44(6): 1140-1153. (in Chinese with English abstract)

[11] MYRONIDIS D, PAPAGEORGIOU C, THEOPHANOUS S. Landslide susceptibility mapping based on landslide history and analytic hierarchy process (AHP)[J]. Natural Hazards, 2016, 81(1): 245-263.

[12] YANG Z Q, ZHU Y Y, STEVE ZOU D H, et al. Activity degree evaluation of glacial debris flow along international Karakorum highway (KKH) based on fuzzy theory[J]. Advanced Materials Research, 2011, 261/262/263: 1167-1171.

[13] 胡凯衡, 崔鹏, 韩用顺, 等. 基于聚类和最大似然法的汶川灾区泥石流滑坡易发性评价 [J]. 中国水土保持科学, 2012, 10(1): 12-18.
HU K H, CUI P, HAN Y S, et al. Susceptibility mapping of landslides and debris flows in 2008 Wenchuan earthquake by using cluster analysis and maximum likelihood classification methods[J]. Science of Soil and Water Conservation, 2012, 10(1): 12-18. (in Chinese with English abstract)

[14] LEE S, MIN K. Statistical analysis of landslide susceptibility at Yongin, Korea[J]. Environmental Geology, 2001, 40(9): 1095-1113.

[15] CHEN W, POURGHASEMI H R, KORNEJADY A, et al. Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques[J]. Geoderma, 2017, 305: 314-327.

[16] 邬礼扬, 殷坤龙, 曾韬睿, 等. 不同栅格尺寸下输电线路地质灾害易发性评价 [J]. 地质科技通报, 2024, 43(1): 241-252.
WU L Y, YIN K L, ZENG T R, et al. Evaluation of geological disaster susceptibility of transmission lines under different grid resolutions[J]. Bulletin of Geological Science and Technology, 2024, 43(1): 241-252. (in Chinese with English abstract)

[17] 郑德凤, 潘美伊, 高敏, 等. 集中降雨影响下辽南仙人洞国家级自然保护区滑坡灾害多因子风险评价 [J]. 地质科技通报, 2025, 44(2): 48-58 .
ZHENG D F, PAN M Y, GAO M, et al. Multi-factor risk assessment of landslide disasters under concentrated rainfall in Xianrendong National Nature Reserve in southern Liaoning Province[J]. Bulletin of Geological Science and Technology, 2025, 44(2): 48-58. (in Chinese with English abstract)

[18] 陈航, 刘惠军, 王韬, 等. 基于频率比–深度神经网络耦合模型的滑坡易发性评价：以盐源县为例 [J]. 水文地质工程地质, 2024, 51(5): 161-171.
CHEN H, LIU H J, WANG T, et al. Landslide susceptibility evaluation based on FR-DNN coupling model: A case study on Yanyuan County[J]. Hydrogeology & Engineering Geology, 2024, 51(5): 161-171. (in Chinese with English abstract)

[19] 邓日朗, 张庆华, 刘伟, 等. 基于改进两步法采样策略和卷积神经网络的崩塌易发性评价 [J]. 地质科技通报, 2024, 43(2): 186-200.
DENG R L, ZHANG Q H, LIU W, et al. Collapse susceptibility evaluation based on an improved two-step sampling strategy and a convolutional neural network[J]. Bulletin of Geological Science and Technology, 2024, 43(2): 186-200. (in Chinese with English abstract)

[20] 王昌明, 黄健, 李桥, 等. 基于信息量模型与 Logistic 回归模型耦合的山西吕梁市地质灾害易发性评价研究 [J]. 水利水电技术, 2019, 50(3): 132-138.
WANG C M, HUANG J, LI Q, et al. Evaluation of geological hazard vulnerability in Lyuliang City in Shanxi Province based on coupling of information content model and Logistic regression model[J]. Water Resources and Hydropower Engineering, 2019, 50(3): 132-138. (in Chinese with English abstract)

[21] 陈朝亮, 彭树宏, 钱静, 等. 基于 AHP-Logistic 熵权模型的西南浅丘区地质灾害分布特征研究：以内江市为例 [J]. 长江科学院院报, 2020, 37(2): 55-61.
CHEN C L, PENG S H, QIAN J, et al. Distribution characteristics of geological hazards in southwestern shallow hill based on AHP-Logistic entropy combined weight model: A case study of Neijiang City[J]. Journal of Yangtze River Scientific Research Institute, 2020, 37(2): 55-61. (in Chinese with English abstract)

[22] TIEN BUI D, TUAN T A, KLEMPE H, et al. Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree[J]. Landslides, 2016, 13(2): 361-378.

[23] 张瀛. 波密至扎木 110 kV 输电线路工程沿线泥石流灾害分析与防治 [D]. 兰州: 兰州大学, 2021.
ZHANG Y. Analysis and prevention of debris flow disaster along the 110 kV power line from Bomi to Zhamu[D]. Lanzhou: Lanzhou University, 2021. (in Chinese with English abstract)

[24] 陈礼然. 波密杂岩锆石 U-Pb 年代学、地球化学和 Hf 同位素组成对拉萨地体前寒武纪演化的约束 [D]. 武汉: 中国地质大学(武汉), 2019.
CHEN L R. Constraints of zircon U-Pb chronology, geochemistry

and Hf isotopic composition of Bomi complex on Precambrian evolution of Lhasa terrane[D]. Wuhan: China University of Geosciences (Wuhan), 2019. (in Chinese with English abstract)

[25] 田海. 西藏波密地区冰川泥石流的成因机制分析 [D]. 成都: 成都理工大学, 2018.
TIAN H. Formation mechanism of glacial debris flow in Bomi, Tibet[D]. Chengdu: Chengdu University of Technology, 2018. (in Chinese with English abstract)

[26] 李翔. 雅鲁藏布江墨脱河段大型滑坡发育与新构造地貌演化关联性研究 [D]. 成都: 成都理工大学, 2019.
LI X. Study on the correlation between the development of large landslide and the evolution of neotectonic geomorphology in Medog reach of Yarlung Zangbo River[D]. Chengdu: Chengdu University of Technology, 2019. (in Chinese with English abstract)

[27] PHAM B T, VAN DAO D, ACHARYA T D, et al. Performance assessment of artificial neural network using Chi-square and backward elimination feature selection methods for landslide susceptibility analysis[J]. Environmental Earth Sciences, 2021, 80(20): 686.

[28] 陈前, 晏鄂川, 黄少平, 等. 基于样本与因子优化的黄冈南部地区地质灾害易发性评价 [J]. 地质科技通报, 2020, 39(2): 175-185.
CHEN Q, YAN E C, HUANG S P, et al. Susceptibility evaluation of geological disasters in southern Huanggang based on samples and factor optimization[J]. Bulletin of Geological Science and Technology, 2020, 39(2): 175-185. (in Chinese with English abstract)

[29] LING X, ZHU Y Q, MING D P, et al. Feature engineering of geohazard susceptibility analysis based on the random forest algorithm: Taking Tianshui City, Gansu Province, as an example[J]. Remote Sensing, 2022, 14(22): 5658.

[30] WANG H S, XU J, TAN S C, et al. Landslide susceptibility evaluation based on a coupled informative-logistic regression model: Shuangbai County as an example[J]. Sustainability, 2023, 15(16): 12449.

[31] QIN S W, QIAO S S, YAO J Y, et al. Establishing a GIS-based evaluation method considering spatial heterogeneity for debris flow susceptibility mapping at the regional scale[J]. Natural Hazards, 2022, 114(3): 2709-2738.

[32] STANCANELLI L M, PERES D J, CANCELLIERE A, et al. A combined triggering-propagation modeling approach for the assessment of rainfall induced debris flow susceptibility[J]. Journal of Hydrology, 2017, 550: 130-143.

[33] 邹强, 唐建喜, 李淑松, 等. 基于水文响应单元的泥石流灾害易发性分区方法 [J]. 山地学报, 2017, 35(4): 496-505.
ZOU Q, TANG J X, LI S S, et al. Susceptibility assessment method of debris flows based on hydrological response unit[J]. Mountain Research, 2017, 35(4): 496-505. (in Chinese with English abstract)

[34] 王向前, 丁明涛, 何雨枫. 基于 RF 模型的甘孜藏族自治州乡城县泥石流易发性评价 [J]. 灾害学, 2023, 38(3): 222-226.
WANG X Q, DING M T, HE Y F. Assessment of debris flow vulnerability in Xiangcheng County, Ganzi Tibetan Autonomous Prefecture based on RF model[J]. Journal of Catastrophology, 2023, 38(3): 222-226. (in Chinese with English abstract)

[35] 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.
ZHANG Y C, QIN S W, ZHAI J J, et al. Susceptibility assessment of debris flow based on GIS and weight information for the Changbai Mountain area[J]. Hydrogeology & Engineering Geology, 2018, 45(2): 150-158. (in Chinese with English abstract)

[36] 李益敏, 李驭豪, 赵志芳. 基于确定性系数模型的泸水市泥石流易发性评价 [J]. 水土保持研究, 2019, 26(4): 336-342.
LI Y M, LI Y H, ZHAO Z F. Assessment on susceptibility of debris flow in Lushui based on the certain factor model[J]. Research of Soil and Water Conservation, 2019, 26(4): 336-342. (in Chinese with English abstract)

[37] 饶姗姗, 冷小鹏. 基于 RSIV-RF 模型的凉山州泥石流易发性评价 [J]. 地质科技通报, 2024, 43(1): 275-287.
RAO S S, LENG X P. Debris flow susceptibility evaluation of Liangshan Prefecture based on the RSIV-RF model[J]. Bulletin of Geological Science and Technology, 2024, 43(1): 275-287. (in Chinese with English abstract)

[38] CASCINI L. Applicability of landslide susceptibility and hazard zoning at different scales[J]. Engineering Geology, 2008, 102(3/4): 164-177.

[39] PRADHAN B, LEE S. Delineation of landslide hazard areas on Penang Island, Malaysia, by using frequency ratio, logistic regression, and artificial neural network models[J]. Environmental Earth Sciences, 2010, 60(5): 1037-1054.

[40] 李泽群. 基于 BP 人工神经网络和 GIS 的湖南省鼎城区滑坡易发性评价 [D]. 石家庄: 河北地质大学, 2020.
LI Z Q. Evaluation of landslide susceptibility based on BP artificial neural network and GIS in Dingcheng District[D]. Shijiazhuang: Hebei GEO University, 2020.

Summary for “基于 Pearson 卡方检验算法评价指标优选的波密–墨脱地区泥石流易发性评价”

# Evaluation of debris flow susceptibility in Bomi-Motuo area using Pearson Chi-square test algorithm based indicator optimization

LI Qun[1]，XU Hongjian[2]，YANG Jin[1]，WANG Linkang[2]，SUN Jingyi[2]，ZHANG Guangcheng[2*]

1. China Communications Second Highway Survey and Design Research Institute Co., Ltd., Wuhan 430058, China；
2. Faculty of Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China

First author，E-mail：35832308@qq.com

* Corresponding author， E-mail：zhangguangc@cug.edu.cn

**Abstract**：Complex geomorphic units and active geological structures provide favorable conditions for debris flow in Tibet, which poses a great threats to human life and property. [**Objective**] The evaluation of debris flow susceptibility can identify key areas for disaster reduction and prevention in this region. [**Methods**] Taking Bomi and Motuo Counties of Tibet Autonomous Region as the study area, 12 factors with high influence on debris flow, including elevation, slope, stratigraphic lithology and rainfall, were selected by Pearson Chi-square test algorithm as evaluation indexes. Data collected from 282 sits with and without debris flows in the study area were taken as the sample database. Based on ArcGIS platform, four susceptibility evaluation models were established by using Information Value Method and Machine Learning Method. The ROC curve and AUC index were introduced to evaluate the accuracy of debris flow susceptibility obtained from the proposed methods. [**Results**] A debris flow susceptibility map for the study area was obtained. [**Conclusion**] The results indicate that: (1) Considering different types of debris flows in different dimensions and controlling factors, the normalization coefficients of latitude and temperature are used as the evaluation index of debris flow susceptibility, which can eliminate the excessive responses of debris flow to temperature in low altitude areas to a certain extent. (2) Air temperature, distance from water system, distance from road, formation lithology and elevation are the main factors of debris flow occurrence in the study area; Factors such as vegetation coverage, terrain humidity, and slope also play an important role. (3) Considering the relationship between the disaster points of debris flows and the classification attributes of the impact factors, the classification attributes of the impact factors are assigned scores and trained as input features. The machine learning model performs well, and its average *AUC* is 0.980, which was better than the traditional information models. (4) The *AUC* of SVM model is as high as 0.987, and the *FR* value of the highly prone region is 41.13. The prediction area of high-risk regions takes up the smallest proportion, demonstrating superior high-precision prediction capability in large-scale regions.

**Key words:** Chi-square test；indicator optimization；debris flow；susceptibility evaluation；high precision；information value；machine learning；Bomi-Motuo area

Received：2024-03-13；Revised：2024-05-30；Accepted：2024-05-30

doi：10.19509/j.cnki.dzkq.tb20240091