水文地质工程地质
HYDROGEOLOGY & ENGINEERING GEOLOGY
中文核心期刊 CSCD核心期刊
中科双效期刊 中国科技核心期刊
Caj-cd规范获奖期刊

滑坡易发性评价中样本不均衡问题处理研究

田　尤，高　波，殷　红，李元灵，张佳佳，陈　龙，李洪梁

Handling imbalanced samples in landslide susceptibility evaluation

TIAN You, GAO Bo, YIN Hong, LI Yuanling, ZHANG Jiajia, CHEN Long, and LI Hongliang

在线阅读 View online: https://doi.org/10.16030/j.cnki.issn.1000-3665.202307002

DOI: 10.16030/j.cnki.issn.1000-3665.202307002

田尤, 高波, 殷红, 等. 滑坡易发性评价中样本不均衡问题处理研究 [J]. 水文地质工程地质, 2024, 51(6): 171-181.
TIAN You, GAO Bo, YIN Hong, et al. Handling imbalanced samples in landslide susceptibility evaluation[J]. Hydrogeology & Engineering Geology, 2024, 51(6): 171-181.

滑坡易发性评价中样本不均衡问题处理研究

田 尤[1,2], 高 波[1,2], 殷 红[3,4], 李元灵[1,2], 张佳佳[1,2], 陈 龙[1,2], 李洪梁[1,2]
（1. 中国地质科学院探矿工艺研究所, 四川 成都 611734; 2. 自然资源部地质灾害风险防控工程技术创新中心, 四川 成都 611734; 3. 四川省地质灾害防治工程技术研究中心, 四川 成都 610081; 4. 四川省地质环境调查研究中心, 四川 成都 610081）

摘要： 滑坡易发性评价中, 样本不均衡问题的不同处理方案通常会带来评价结果的大量不确定性。针对这一问题, 以藏东昌都市部分县（区）为研究区, 构建滑坡/非滑坡样本不均衡数据集, 采用不处理、下采样和合成少数类过采样（synthetic minority oversampling technique, SMOTE）3 种处置方案, 运用逻辑回归方法分别构建滑坡易发性评价模型。基于 ROC 曲线、准确度、精确率、召回率、漏检率等评价指标, 采用综合评价指标 F_1'同数对模型分类的精度进行验证。结果表明: 数据处理成均衡数据集（过采样/下采样）建立的模型效果较不处理数据建立的模型效果有了大幅提升, F_1'同数的值最大提高了 53.17%; 在下采样、过采样两种数据处理方案中, 过采样方法比下采样方法 F_1'分数的值提高了 16.30%, 表明过采样方法对处理样本不均衡数据问题方面具有较好效果。研究成果可为滑坡预测和地质灾害预测前的数据集处理提供参考, 为进一步提高区域防灾减灾水平提供理论与技术支持。

关键词： 滑坡易发性; 合成少数类过采样技术; 评价模型; 昌都市; 样本不均衡数据
中图分类号： P642.22 **文献标志码：** A **文章编号：** 1000-3665(2024)06-0171-11

Handling imbalanced samples in landslide susceptibility evaluation

TIAN You[1,2], GAO Bo[1,2], YIN Hong[3,4], LI Yuanling[1,2], ZHANG Jiajia[1,2], CHEN Long[1,2], LI Hongliang[1,2]
（1. *Institute of Exploration Technology, Chinese Academy of Geological Sciences, Chengdu, Sichuan* 611734, *China*; 2. *Technology Innovation Center for Risk Prevention and Mitigation of Geohazard, Ministry of Natural Resources, Chengdu, Sichuan* 611734, *China*; 3. *Sichuan Province Engineering Technology Research Center of Geohazard Prevention, Chengdu, Sichuan* 610081, *China*; 4. *Sichuan Geological Environment Survey and Research Center, Chengdu, Sichuan* 610081, *China*）

Abstract: In landslide susceptibility assessment, different approaches to handling sample imbalance can introduce significant uncertainty in evaluation outcomes. To address this issue, this study focused on the Changdu area of eastern Tibet and constructed the landslide susceptibility evaluation model using a dataset with imbalanced landslide and non-landslide samples. Three disposal schemes were applied: no treatment, downsampling, and SMOTE oversampling. The logistic regression method was used to construct the landslide susceptibility evaluation model. Based on ROC curve, accuracy, precision, recall, missed detection rate, and other evaluation indicators, the comprehensive evaluation index of F_1' score was used to verify the accuracy of model classification. The results show that the modeling effect of landslide susceptibility obtained by data processing into equilibrium data

收稿日期： 2023-07-02; **修订日期：** 2023-11-07 **投稿网址：** www.swdzgcdz.com
基金项目： 中国地质调查局地质调查项目（DD20230449; DD20190644）; 第二次青藏高原综合科学考察研究项目（2019QZKK0902）
第一作者： 田尤（1991—）, 男, 硕士, 工程师, 主要从事地质灾害调查与评价研究工作。E-mail: tianyou2013@yeah.net

(downsampling/oversampling) is greatly improved compared with that obtained without processing data. Specifically, the value of the F_1'score of the comprehensive index was increased by 53.17%. In the two schemes for processing data (downsampling and oversampling), the oversampling method increased the value of the composite index F_1' score by 16.30% compared with the downsampling method, indicating that the oversampling method has effectiveness in handling unbalanced data. This study can provide basic information for processing of data sets before landslide prediction and geological disaster prediction, and provide theoretical and technical support for further improving regional disaster prevention and mitigation.

Keywords: Landslide susceptibility; SMOTE ; evaluation model; Changdu; unbalanced data

滑坡易发性评价是开展滑坡危险性、风险性评价的基础,强调静态的地质灾害易发条件和灾害发生的空间概率问题[1],评价方法可以分为定性评价方法和定量评价方法[2]。定性评价方法主要包括滑坡编录法和知识驱动法[3-4],定量评价方法主要包括数据驱动法和物理驱动法[5-7]。早期滑坡易发性评价方法以定性评价为主,评价精度高度依赖专家的主观经验[8-10],缺乏定量表达,无法对评价结果进行对比分析[4]。随着计算机技术和地理信息技术的快速发展,基于统计分析和机器学习的定量评价被广泛运用在滑坡的易发性评价中[11-14]。特别是机器学习模型,由于可以处理评价因子间的非线性关系,在滑坡的易发性评价中具有较好的效果,主要包括逻辑回归模型[15-16]、随机森林模型[17]、旋转森林模型[18]、支持向量机模型[19]、神经网络模型[20-21] 等。

机器学习方法是基于数据驱动的,进行滑坡易发性评价需要构建滑坡\非滑坡样本库。从两类样本在空间上的分布数目上看,滑坡样本数一般远小于非滑坡样本数,数据具有显著不均衡性。针对样本不均衡的数据集,目前主要采用不处理、下采样和过采样 3 种处置方案。

(1)样本不处理,即不考虑或弱考虑样本不均衡带来的影响,在滑坡的易发性评价中通常参考滑坡样本,选取数倍于滑坡灾害数目的非滑坡样本,组成数据集,如穆科等[22] 通过选取铜川市耀州区 71 个正样本与 213 个负样本共同组成数据集,采用 LR-RF 模型评价了其易发性;刘坚等[23] 选取 2 倍于滑坡灾害样本的非滑坡样本点,组成数据集,使用优化后的随机森林模型评价三峡库区沙镇溪镇—泄滩乡滑坡易发性。

(2)下采样法是通过对样本多的一侧采用随机抽样的方法,使得样本量多的一侧的样本量减少,达到数据均衡的目的。如 Hu 等[24] 通过从低边坡区、无滑坡区和极低易感区随机选取 3 组与滑坡数量相等的“非滑坡”样本,讨论不同“非滑坡”负样本的选择对易发性评价结果的影响;Hu 等[6] 通过从非滑坡区随机选取与滑坡数量相同的“非滑坡”样本,使用集成技术和基础学习器,构建 5 个集成模型,研究各种模型的拟合度、泛化能力和稳健性;黄发明等[25] 通过构建滑坡、“非滑坡”均衡数据集,讨论不同空间分辨率和训练测试集比例下的滑坡易发性;王毅等[26] 以铅山县滑坡为研究对象,通过从非滑坡区随机选择等量的“非滑坡”样本,提出了 3 种卷积神经网络模型的滑坡易发性分析处理框架并进行验证;杜国梁等[27] 通过自然随机生成、结合遥感解译判别的方法选取等量的非滑坡点,运用逻辑回归-信息量模型评价川藏交通廊道滑坡的易发性;陈涛等[28] 以三峡库区秭归县滑坡为研究对象,通过随机的方法构建等量的滑坡、非滑坡栅格数据集,运用深度信念网络模型对研究区滑坡易发性进行评价;杨强等[29] 通过自然随机选取等量的非滑坡样本,使用多种概率统计及组合模型对陇南白龙江流域中游及其岷江支流段滑坡易发性进行评价;郭子正等[30] 以三峡库区万州区滑坡为研究对象,通过 GIS 随机选取 10 000 个滑坡、非滑坡栅格样本,运用证据权法-神经网络模型评价了滑坡易发性;贾雨霏等[31] 通过 GIS 随机生成等量的非滑坡样本,运用自组织映射神经网络-信息量模型-支持向量机耦合模型,评价十堰市茅箭区滑坡的易发性。

(3)过采样法是运用一定的数据扩充方法,使样本量少的一侧的样本量增多,以达到数据均衡的目的,组成数据集;如武雪玲等 [32] 、李坤等[33]分别使用合成少数类过采样技术(SMOTE)扩充滑坡、泥石流样本数,组成数据集并评价研究区滑坡、泥石流的易发性;赵占鳌等 [34]通过引入数据增强处理,将有限的滑坡正样本与对应因子图像通过水平、垂直翻转扩充数据集,评价西藏色东普沟滑坡的易发性。

总体而言,3 种样本处置方法中,下采样法是现有研究使用频率最高的方法。哪种方法对滑坡的易发性评价更合适,从已有文献检索情况来看,鲜有讨

论。为了研究 3 种样本处置方法对评价结果的影响，以地质灾害较为发育的西藏东部昌都市为研究区，搜集已有资料，通过构建不均衡数据集，运用逻辑回归模型方法，分析滑坡易发性评价中样本不均衡数据的不同处理方法对评价结果带来的不确定性，讨论模型精度的分类指标。研究成果可为高山峡谷区滑坡预测和地质灾害预测提供参考。

1　模型方法

1.1　预测框架

1.1.1　构建数据集

准确的样本数据是模型学习的前提，选用点提取栅格的方式构建数据集。在预处理阶段，确定栅格大小为 90 m×90 m，将研究区 1 868 个滑坡编目图进行二值化，结果为滑坡与非滑坡两种类型，以供数据筛选。滑坡点所处栅格被标记为滑坡，为 1 868 个栅格，其他栅格标记为非滑坡。为了充分模拟样本不均衡问题，选择 200 000 个非滑坡样本，数量远大于滑坡样本数。为保证所选取的非滑坡样本尽量为真的“非滑坡”，先选用频率比模型对研究区滑坡易发性进行初评价，按照自然断点法进行分区，在低易发分区内随机选择非滑坡样本。最终，滑坡与非滑坡样本构成整体不均衡数据集。

1.1.2　对比研究

为验证不均衡数据集在不同处置方案的建模预测效果，设立 3 种方案：①不处理，使用原始的不均衡数据集；②下采样，即将非滑坡样本随机向下采样，使得非滑坡样本数量与滑坡样本数量相同，组成数据集；③过采样，使用合成少数类过采样技术（synthetic minority oversampling technique, SMOTE）进行样本扩充，使得滑坡样本数量向上扩充到与非滑坡样本数量相同，组成数据集。分别将数据集打乱顺序后随机按比例划分为训练集、验证集和测试集。训练集用于模型学习，初步建立模型参数；验证集用于检验模型训练状态，优化模型参数；测试集则用于检验模型精度。下采样/过采样均衡数据集选用的模型参数与原始不均衡数据集模型参数完全相同，以便比较各方案的优劣。

1.1.3　构建模型

使用逻辑回归模型进行建模预测。逻辑回归模型是一种广义的线性回归分析模型，适用于二分类问题的建模。其核心原理是一个因变量与多个自变量（x_1, x_2, …, x_n）之间形成多元回归的关系，Logistic 函数表达式为：

$$\mathrm{Logistic}(P)=\ln\left(\frac{P}{1-P}\right)=a+b_1x_1+b_2x_2+\cdots+b_nx_n \quad (1)$$

$$P=\frac{1}{1+\mathrm{e}^{-y}} \quad (2)$$

$$y=(a+b_1x_1+b_2x_2\cdots+b_nx_n) \quad (3)$$

式中：P——滑坡发生的概率；

a——Logistic 回归计算出的常数项；

$x_1, x_2, \cdots, x_n$——评价因子；

$b_1, b_2, \cdots, b_n$——对应因子的逻辑回归系数，采用最大拟然估计方法求解。

通过对回归概率值指定一个分类阈值，实现滑坡和非滑坡的分类。损失函数的惩罚项选用 L1 正则化，正则化强度倒数 c 取 0.01，概率阈值取 0.5，算法在 python 中实现，研究技术框架见图 1。

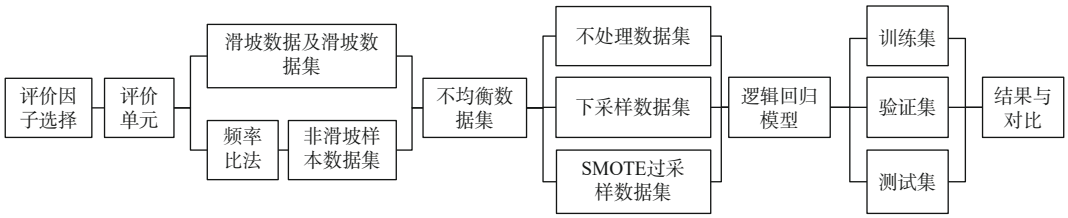


图 1　研究技术框架图

Fig. 1　The technical framework

1.2　评价方法

1.2.1　样本提取和建立

（1）频率比模型

频率比模型是一种数理统计方法，通过分析滑坡的分布与各分级因子之间的空间关系，计算每个因子在不同分级条件下滑坡发生概率。频率比模型见式（4）。

$$FR=\frac{N_{ij}/N}{A_{ij}/A} \quad (4)$$

式中：FR——频率比值；

N_{ij}——第 i 个因子第 j 级内的滑坡栅格数；

N——滑坡栅格总数；

A_{ij}——第 i 个因子第 j 级栅格总数；

A——研究区栅格总数。

(2)SMOTE 过采样

SMOTE 过采样方法基于随机过采样算法，以每个样本点为依据，随机选择 k 个近邻点进行插值，乘上一个范围为 [0,1] 的随机值，以达到新生成样本的目的。主要步骤为：①对于每一个少数类样本 x_i，计算它到少数类所有样本的欧式距离，得到其 k 近邻；②根据样本不平衡比例，确定采样倍率，对于每一个少数类样本 x_i，从其 k 近邻中随机选择 n 个样本；③对于选择出的 n 个样本，分别在与 x_i 间进行随机插值，构建新的样本。

1.2.2 精度分析

混淆矩阵是机器学习中总结模型分类预测结果的常用方法[27]，利用混淆矩阵可以直观地显示模型预测结果与真实滑坡结果，从而形成对模型学习结果的评价。混淆矩阵形式如图 2 所示。

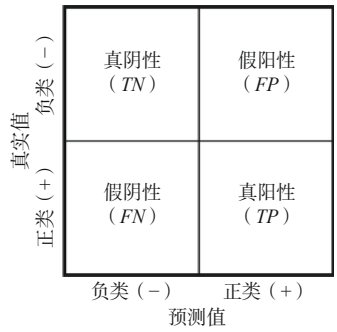


图 2 混淆矩阵

Fig. 2 confusion matrix

其中 TP 为正确的正类(真滑坡数)，FP 为错误的正类(假滑坡数)，TN 为正确的负类(真非滑坡数)，FN 为错误的负类(假非滑坡数)。

目前，评价模型性能的主要手段是测试模型在测试集数据上的检测效果，常用的技术指标有准确度(A)、精确率(P)、召回率(R)和漏检率(M)。各指标计算见式(5)—(8)。

$$A=\frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$P=\frac{TP}{TP+FP} \tag{6}$$

$$R=\frac{TP}{TP+FN} \tag{7}$$

$$M=\frac{FN}{TP+FN} \tag{8}$$

2 研究区概况及评价因子选取

2.1 研究区概况

研究区位于西藏昌都市，行政区划上包括江达县、贡觉县、察雅县、卡若区、类乌齐县和洛隆县，面积约 59 960 km^2。大地构造位置位于青藏高原东南部，三江地区北段。金沙江、澜沧江、怒江自北向南穿过，区内支流众多，水资源丰富。研究区气候类型为高原温带半干旱季风型气候。境内多年平均降雨量为 488 mm。地形以高山峡谷地貌为主，西南部地势整体较高，三大河流地势较低，海拔范围为 2 005 ~ 5 782 m。根据县市地质灾害风险调查成果，结合遥感解译及野外调查结果，确定研究区共发育滑坡 1 868 处。各滑坡空间上分布不均匀，主要集中分布在怒江、澜沧江、金沙江及支流两侧斜坡(图 3)。

2.2 滑坡影响因子选取

滑坡的形成主要受地形地貌、地质构造和岩性等因素的影响。选取地形地貌(坡度、高程、曲率和坡向)、地质条件(距断层距离和工程地质岩组)、水文条件(距水系距离)和植被条件(植被指数)共 8 个因子反映滑坡特征。

其中坡度、高程、曲率、坡向因子基于数字高程模型(DEM)通过 ArcGIS 平台提取，DEM 来源为地理空间数据云(http://www.gscloud.cn)，分辨率为 90 m；距断层距离因子根据区域 1∶25 万地质资料进行缓冲区分析；工程地质岩组因子基于 1∶25 万区域地层资料整合获得；距水系距离因子根据对明显水系做缓冲区分析获得；植被指数因子根据自然资源部 2020 年数据获得。所有因子均转化为 90 m×90 m 栅格单元，其中连续型因子按照自然间断法分割成若干个区间，离散型因子按照属性值进行分级(表 1)，得到分级后的各因子图层(图 4)。

3 结果与分析

3.1 频率比法滑坡易发性评价结果及评价数据集

通过式(1)对研究区 8 个因子进行统计，得到每个因子分级后的频率比值(表 1)。利用 ArcGIS 栅格计算器将各因子频率比值进行叠加，得到滑坡易发性初步评价图，按照自然断点法将易发性评价结果划分为极高易发区 [7.856, 4.334]，高易发区(4.334, 3.341]，中易发区(3.341, 2.498] 和低易发区(2.498, 0.180] 4 级(图 5)。

为尽量保证所选取的非滑坡样本具有代表性，在

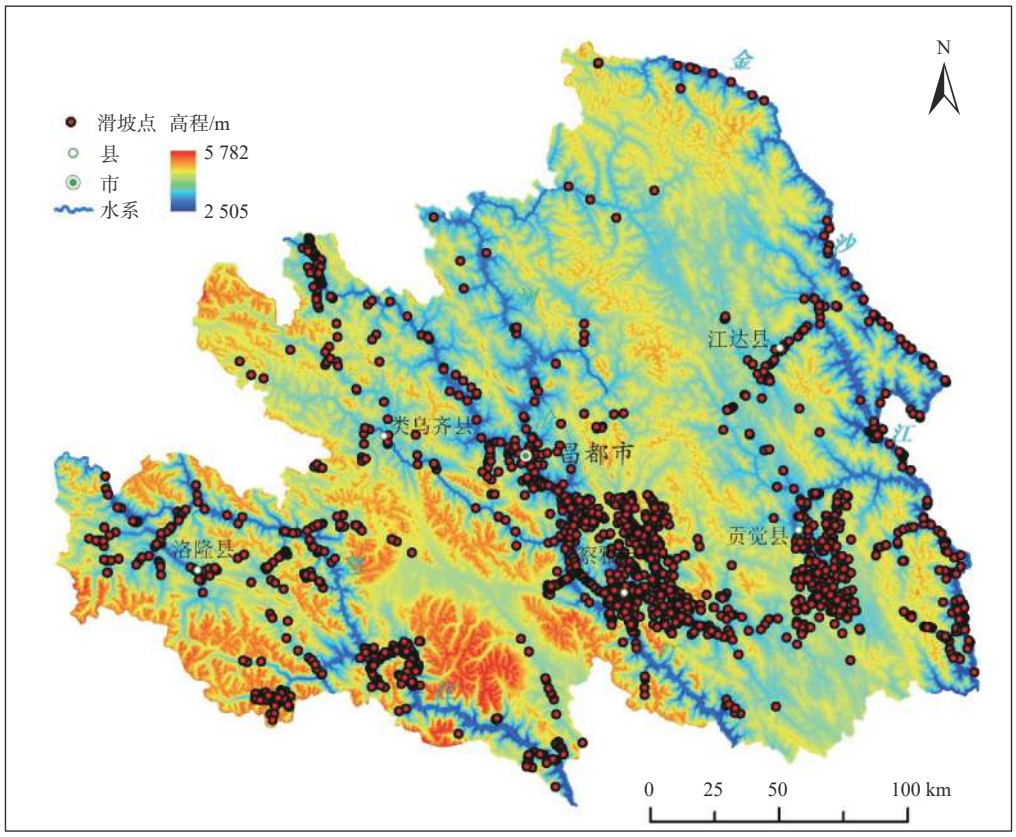


图 3　研究区位置及滑坡分布图

Fig. 3　Study area location and landslide distribution

滑坡低易发区中随机生成 200 000 个点数据作为非滑坡点, 非滑坡数据点数量远大于滑坡点数量。运用 ArcGIS 平台点提取栅格属性功能, 提取滑坡点与非滑坡点处评价图层的属性值, 组成不均衡数据集, 其中滑坡点标记为正样本, 非滑坡点标记为负样本。共获得样本 201 836 个, 其中正样本 1 836 处, 负样本 200 000 处, 每个样本均包含 8 个因子的属性值及标签值。

3.2　结果验证及分析

分别采用不处理、下采样和 SMOTE 过采样法对原始不均衡数据集进行处理, 得到 3 份数据集。其中不处理的数据集标记为数据集 A, 数据量样本共 201 836个, 其中正样本 1 836 处, 负样本 200 000 处, 为不均衡数据集; 下采样数据集标记为数据集 B, 通过在所有负样本中随机选择 1 836 处样本, 与 1 836 处滑坡正样本组成均衡数据集, 数据量样本共 3 672 个; 过采样数据集标记为数据集 C, 使用 SMOTE 样本生成策略将滑坡样本扩充至 199 996 个, 与 200 000 个非滑坡样本组成数据集, 数据量样本共 399 996 个, 基本为均衡数据集。将得到的 3 组数据集 A、B、C, 分别取 70% 数据, 运用逻辑回归方法进行训练建模, 取剩余 30% 数据用于模型测试, 得到测试集验证结果。

在滑坡易发性评价领域, 常用受试者工作特征曲线(receiver operating characteristic curve, 简称 ROC 曲线)来检验模型的精度, 主要表征样本正确预测比例(敏感度)与样本错误预测比例(特异性)的综合指标, 纵坐标为敏感度, 横坐标为 1-特异性。分别对 3 组模型测试数据集结果构建 ROC 曲线, 结果如图 6 所示。

从 ROC 曲线及曲线下面积(*AUC*)结果来看, 3 种方案中以过采样方案得到的 *AUC* 值最高, 为 0.97, 效果最好; 不处理方案次之, 为 0.95; 下采样方案得到的结果最低, 为 0.91。

由于 *AUC* 值是样本正确预测比例(敏感度)与样本错误预测比例综合指标, 当样本不均衡时, 有存在结果向样本多一侧倾斜的风险。为更清晰直观地分析模型预测结果, 选用 0.5 作为概率分类阈值, 对 3 组测试集数据按照预测结果与实际结果(0 代表非滑坡; 1 代表滑坡)进行统计, 构建真实值与预测值的混淆矩阵结果见图 7。

选用准确度(*A*)、精确率(*P*)、召回率(*R*)和漏检率(*M*)4 个单一评价指标, 用于检验 3 种处置方式得

表 1 各评价因子分级及频率比值

Table 1 Frequency ratio of each evaluation factor

指标因子	指标分级	占滑坡栅格比/%	占总栅格比/%	频率比	归一化值
坡度/(°)	[0, 10)	0.05	0.09	0.55	0
	[10, 17)	0.09	0.13	0.71	0.07
	[17, 23)	0.13	0.16	0.81	0.12
	[23, 28)	0.18	0.18	1.03	0.22
	[28, 33)	0.18	0.18	0.98	0.19
	[33, 39)	0.17	0.15	1.08	0.24
	[39, 47)	0.15	0.09	1.66	0.50
	[47, 81]	0.06	0.02	2.79	1.00
高程/m	[2 496, 3 435)	0.34	0.03	10.53	1.00
	[3 435, 3 787)	0.26	0.07	3.58	0.34
	[3 787, 4 055)	0.21	0.12	1.73	0.16
	[4 055, 4 278)	0.12	0.17	0.71	0.07
	[4 278, 4 487)	0.05	0.21	0.24	0.02
	[4 487, 4 705)	0.02	0.19	0.12	0.01
	[4 705, 4 966)	0	0.14	0.02	0
	[4 966, 5 784]	0	0.07	0.02	0
曲率	[−9.23, −1.25)	0.07	0.03	2.55	1.00
	[−1.25, −0.65)	0.16	0.10	1.60	0.46
	[−0.65, −0.28)	0.21	0.18	1.13	0.20
	[−0.28, 0.09)	0.19	0.24	0.78	0
	[0.09, 0.46)	0.18	0.22	0.86	0.04
	[0.46, 0.90)	0.12	0.14	0.82	0.02
	[0.90, 1.57)	0.06	0.08	0.81	0.02
	[1.57, 9.63]	0.02	0.02	0.98	0.11
坡向	平面	0	0	0.50	0
	北	0.10	0.12	0.83	0.49
	东北	0.17	0.15	1.17	0.98
	东	0.14	0.13	1.06	0.83
	东南	0.13	0.11	1.18	1.00
	南	0.13	0.11	1.18	1.00
坡向	西南	0.13	0.14	0.95	0.66
	西	0.13	0.12	1.00	0.74
	西北	0.07	0.11	0.60	0.15
距断层距离/m	[0, 500)	0.15	0.14	1.04	0.55
	[500, 1 000)	0.12	0.13	0.90	0.12
	[1 000, 2 000)	0.22	0.22	0.99	0.41
	[2 000, 4 000)	0.22	0.26	0.86	0
	≥4 000	0.29	0.25	1.18	1.00
工程地质岩组	较坚硬层状碎屑岩组	0.13	0.19	0.68	0.25
	较坚硬层状碳酸盐岩组	0.10	0.12	0.89	0.52
	软硬相间互层状碎屑岩组	0.48	0.41	1.16	0.86
	坚硬块状侵入岩组	0.15	0.13	1.16	0.85
	较软弱薄层浅变质岩组	0.06	0.08	0.73	0.31
	坚硬厚层-块状深变质岩组	0.07	0.05	1.27	1.00
	第四系松散岩组	0.01	0.02	0.49	0
距河流距离/m	[0, 500)	0.72	0.23	3.09	1.00
	[500, 1 000)	0.13	0.21	0.62	0.18
	[1 000, 1 500)	0.08	0.19	0.42	0.11
	[1 500, 2 000)	0.04	0.16	0.28	0.06
	[2 000, 2 500)	0.02	0.12	0.19	0.03
	≥2 500	0.01	0.10	0.10	0
植被指数	[0, 0.10)	0.01	0.09	0.07	0
	[0.10, 0.21)	0.06	0.07	0.85	0.45
	[0.21, 0.30)	0.23	0.13	1.79	1.00
	[0.30, 0.37)	0.30	0.22	1.35	0.74
	[0.37, 0.43)	0.21	0.23	0.90	0.48
	[0.43, 0.51)	0.11	0.15	0.71	0.37
	[0.51, 0.62)	0.06	0.08	0.73	0.38
	[0.62, 1]	0.03	0.03	1.14	0.62

到的逻辑回归模型预测结果，计算见式(5)—(8)，结果如表 2 所示。

准确度(A)表征的是样本预测正确数占样本总数的比值，既包括滑坡样本预测正确数，也包括非滑坡样本预测正确数。从研究区预测结果来看，准确度以不处理方案最高，过采样方案次之，下采样方案最低。原因是在藏东不均衡样本数据集中，非滑坡数据远大于滑坡数据，预测结果会向数据多一侧倾斜，极多的非滑坡数据被预测正确，所以显示不处理的精确度更高。对于处理后的均衡数据，过采样方案比下采样方案的准确度增高了 11.35%。

精确率(P)表征的是正确预测为正的样本占全部预测为正的样本的比例。从预测结果看，精确率以下采样方案最高，不处理方案次之，过采样方案最低。由于精确率更强调滑坡预测的正确率，希望预测结果尽可能不出错，忽略滑坡的检漏率，可能导致滑坡的漏检。从数据量上看，下采样方案最少，不处理方案居中，过采样方案数据量最多。随着数据量的增大，预测为正的比例增大，但是实际为正的样本数是固定的，所以表现为随着数据量增大，准确率呈现降低的趋势。

召回率(R)表征的是正确预测为正样本占全部正样本的比例。从研究区预测结果来看，召回率以过采样方案最高，下采样方案次之，不处理方案最低。召回率指标强调的尽可能多地找到滑坡样本，分子分母均不考虑非滑坡数，这意味着高的召回率可能会存在更多的误检；漏检率(M)表征的是错误预测为负占全部正样本的比例，与召回率趋势相反，召回率越高，漏检率越低，两者之和为 1。从研究区预测结果来看，漏检率以不处理方案最高，下采样方案次之，过采样方案最低。

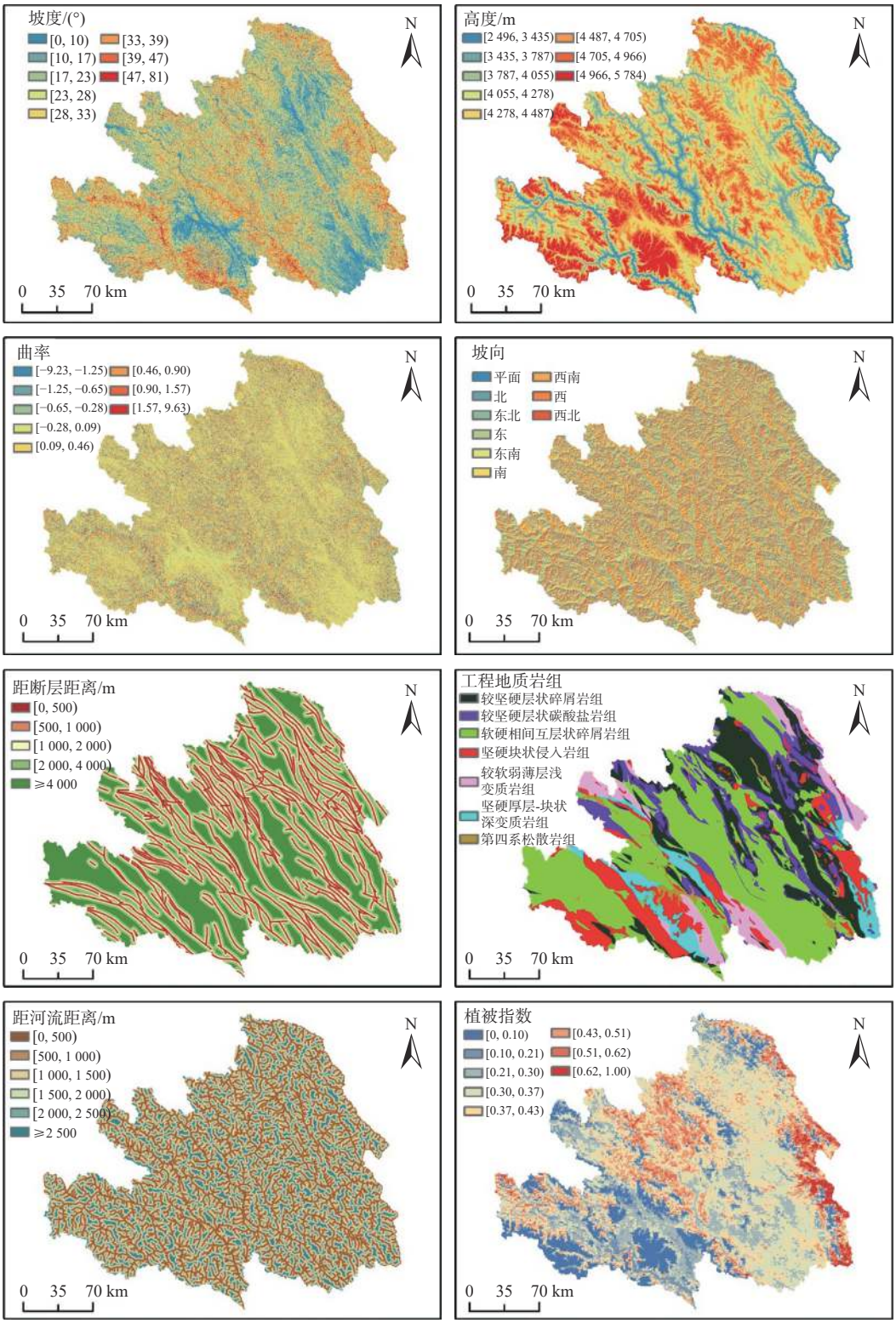


图 4　评价因子

Fig. 4　Evaluation Factors

4　讨论

ROC 曲线作为滑坡预测评价的一个常用综合指标, 得到的结果是过采样方案模型效果最好, *AUC* 值为 0.97。但是值得注意的是, 不处理方案也得到了相对高的评价结果, *AUC* 值为 0.95, 该方案的滑坡召回率却是较低的。对于模型验证而言, ROC 曲线更多的

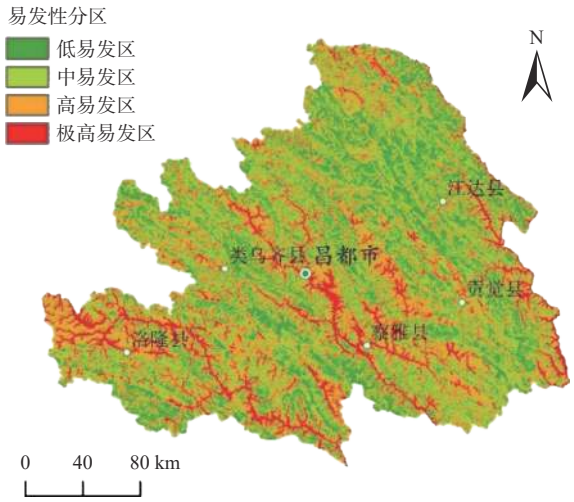


图 5 基于频率比法的研究区易发性评价结果

Fig. 5 Evaluation results of susceptibility in the study area based on frequency ratio method

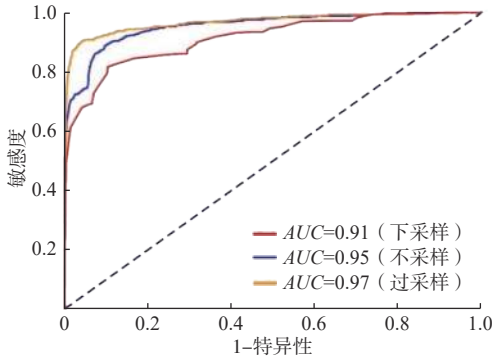


图 6 三种处置方案的 ROC 曲线对比

Fig. 6 Comparison of ROC curves of three disposal schemes

检验滑坡与非滑坡的综合预测能力，由于不处理方案样本是不均衡的，大量的非滑坡样本被预测正确，即使滑坡样本的召回率较低，仍然获得较高的 *AUC* 值。

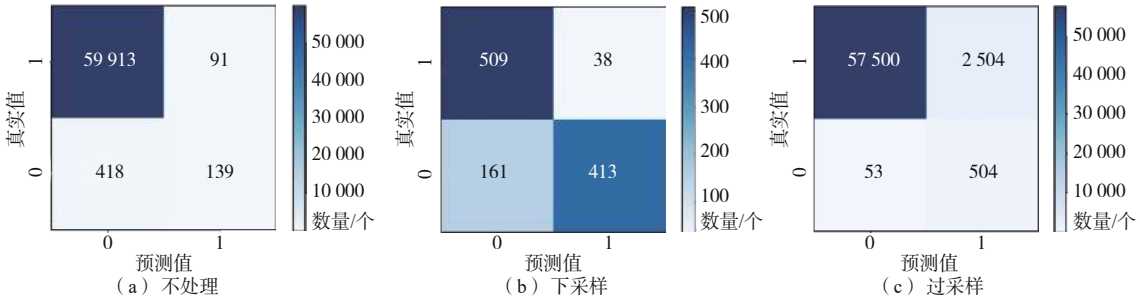


图 7 各处理方法验证结果混淆矩阵

Fig. 7 confusion matrix of verification results of each processing method

表 2 不同处理方式得到的逻辑回归模型预测结果评价

Table 2 Evaluation of logistic regression model prediction results obtained by different processing methods

评价指标 处理方式	准确度/%	精确率/%	召回率/%	漏检率/%
不处理	99.16	60.43	24.96	75.04
下采样	82.25	91.57	71.95	28.05
过采样	95.78	16.76	90.48	9.52

可见，对于滑坡预测的不均衡数据，ROC曲线作为综合指标检验模型的能力存在一定缺陷。

从综合防灾角度出发，为最大限度保证人民的生命财产安全，希望能最大程度地找到滑坡样本，降低漏检率，那么召回率是一个比较合适的评价指标。但是，在最大程度找到所有滑坡样本的同时，又希望能让防治的范围尽量小，模型精度能高一些，这样才能让防治更有针对性，仅使用召回率指标评估模型的优劣存在一定的片面性。ROC 曲线作为滑坡预测模型评价常用的指标，但是对检验不均衡数据模型也存在一定局限性。为达到上述目的，参考 F_1 分数指标计算方法，提出一种适用于地质灾害领域模型评价的综合精度指标 F_1' 分数，表征准确率 A 与召回率 R 的调和平均数（式 9）。指标同时兼顾分类模型的召回率和精确率，对于滑坡模型预测结果评价是比较理想的指标。

$$F_1' = 2 \times \frac{A \times R}{A + R} \tag{9}$$

从 3 种处置方案得到的结果来看，不处理方案的 F_1' 指标值为 39.88%；下采样方案的 F_1' 分数指标值为 76.76%；过采样方案的 F_1' 分数指标值为 93.05%。表明对于样本不均衡数据集，处理数据（下采样、过采样）方案比不处理数据方案所建立的模型效果更好，综合指标 F_1' 分数的值最大提高了 53.17%；在两种处理数据的方案中，过采样方案比下采样方案更适用一些，综合指标 F_1' 分数的值提高了 16.30%。分析其原因，主要是因为不处理方案的数据正、负样本严重不均衡，所建立的模型会更多的偏向样本数量多的一侧，造成精度指标高，但对滑坡的预测意义不大；下采样方案通过样本数量向下削减的方式，使得大量的负样本信息

没有得到有效利用，样本利用率不高是导致所建立的模型精度偏低的主要原因；过采样方案通过相近邻算法扩充样本数，更多地利用非滑坡样本信息，可以获得更好的模型精度。

5　结论

（1）研究提出一种适用于滑坡易发性预测模型精度的评价指标F_1'分数，同时兼顾分类模型的召回率和精确率，不受样本是否均衡的影响，是滑坡空间预测比较理想的评价指标。

（2）对于样本不均衡数据集，数据处理成均衡数据集（过采样/下采样）建立的模型效果较不处理数据建立的模型效果有了大幅提升，综合指标F_1'分数的值最大提高了 53.17%。

（3）在两种处理数据（过采样/下采样）的方案中，过采样方案比下采样方案综合指标F_1'分数的值提高了 16.30%，ROC 曲线下面积 *AUC* 的值提高了 6%。表明过采样方案对处理样本不均衡数据方面具有较好的有效性。

参考文献（References）：

［1］ 吴树仁，石菊松，张春山，等. 地质灾害风险评估技术指南初论［J］. 地质通报，2009，28（8）：995－1005.［WU Shuren，SHI Jusong，ZHANG Chunshan，et al. Preliminary discussion on technical guideline for geohazard risk assessment［J］. Geological Bulletin of China，2009，28（8）：995－1005.（in Chinese with English abstract）］

［2］ ADITIAN A，KUBOTA T，SHINOHARA Y. Comparison of GIS-based landslide susceptibility models using frequency ratio，logistic regression，and artificial neural network in a tertiary region of Ambon，Indonesia［J］. Geomorphology，2018，318：101－111.

［3］ CHEN Wei，LI Wenping，CHAI Huichan，et al. GIS-based landslide susceptibility mapping using analytical hierarchy process（AHP）and certainty factor（CF）models for the Baozhong region of Baoji City，China［J］. Environmental Earth Sciences，2015，75（1）：63.

［4］ 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.

［5］ LARI S，FRATTINI P，CROSTA G B. A probabilistic approach for landslide hazard analysis［J］. Engineering Geology，2014，182：3－14.

［6］ HU Xudong，MEI Hongbo，ZHANG Han，et al. Performance evaluation of ensemble learning techniques for landslide susceptibility mapping at the Jinping county，Southwest China［J］. Natural Hazards，2021，105（2）：1663－1689.

［7］ 王世宝，庄建琦，樊宏宇. 基于频率比与集成学习的滑坡易发性评价——以金沙江上游巴塘—德格河段为例［J］. 工程地质学报，2022，30（3）：817－828.［WANG Shibao，ZHUANG Jianqi，FAN Hongyu，et al. Evaluation of landslide susceptibility based on frequency ratio and ensemble learning—Taking the Batang-Dege section in the upstream of Jinsha River as an example［J］. Journal of Engineering Geology，2022，30（3）：817－828.（in Chinese with English abstract）］

［8］ KAYASTHA P，DHITAL M R，DE SMEDT F. Application of the analytical hierarchy process（AHP）for landslide susceptibility mapping：a case study from the tinau watershed，West Nepal［J］. Computers & Geosciences，2013，52：389－408.

［9］ MANDAL B，MANDAL S. Analytical hierarchy process（AHP）based landslide susceptibility mapping of Lish river basin of eastern Darjeeling Himalaya，India［J］. Advances in Space Research，2018，62（11）：3114－3132.

［10］ AKGUN A，DAG S，BULUT F. Landslide susceptibility mapping for a landslide-prone area（Findikli，NE of Turkey）by likelihood-frequency ratio and weighted linear combination models［J］. Environmental Geology，2008，54（6）：1127－1143.

［11］ AZIZ K，SARKAR S，SAHU P. Comparative analysis of frequency ratio，information value，and analytical hierarchy process statistical models for landslide susceptibility mapping in Kashmir Himalayas［J］. Arabian Journal of Geosciences，2024，17（1）：36.

［12］ BILGILIOĞLU H. A comparison of different machine learning models for landslide susceptibility mapping in Rize（Türkiye）［J］. Baltica，2023，36（2）：115－132.

［13］ 张钟远，邓明国，徐世光，等. 镇康县滑坡易发性评价模型对比研究［J］. 岩石力学与工程学报，2022，41（1）：157－171.［ZHANG Zhongyuan，DENG Mingguo，XU Shiguang，et al. Comparison of landslide susceptibility assessment models in Zhenkang County，Yunnan Province，China［J］. Chinese Journal of Rock Mechanics and Engineering，2022，41（1）：157－171.（in Chinese with English abstract）］

［14］ MA Yanbin，LI Hongrui，WANG Lin，et al. Machine learning algorithms and techniques for landslide susceptibility investigation：A literature review［J］. Journal

of Civil and Environmental Engineering, 2022, 44(1): 53 − 67.

[15] CHEN W, ZHAO X, SHAHABI H, et al. Spatial prediction of landslide susceptibility by combining evidential belief function, logistic regression and logistic model tree[J]. Geocarto International, 2019, 34(11): 1177 − 1201.

[16] OH H J, KADAVI P R, LEE C W, et al. Evaluation of landslide susceptibility mapping by evidential belief function, logistic regression and support vector machine models[J]. Geomatics, Natural Hazards and Risk, 2018, 9(1): 1053 − 1070.

[17] CHEN Wei, XIE Xiaoshen, PENG Jianbing, et al. GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method[J]. CATENA, 2018, 164: 135 − 149.

[18] 刘渊博, 牛瑞卿, 于宪煜, 等. 旋转森林模型在滑坡易发性评价中的应用研究 [J]. 武汉大学学报(信息科学版), 2018, 43(6): 595 − 964. [LIU Yuanbo, NIU Ruiqing, YU Xianyu, et al. Application of the rotation forest model in landslide susceptibility assessment[J]. Geomatics and Information Science of Wuhan University, 2018, 43(6): 959 − 946.(in Chinese with English abstract)]

[19] 王卫东, 刘攀, 龚陆. 基于支持向量机模型的四川省滑坡灾害易发性区划 [J]. 铁道科学与工程学报, 2019, 16(5): 1194 − 1200. [WANG Weidong, LIU Pan, GONG Lu. Landslide susceptibility mapping of Sichuan province based on support vector machine[J]. Journal of Railway Science and Engineering, 2019, 16(5): 1194 − 1200.(in Chinese with English abstract)]

[20] 牟家琦, 庄建琦, 王世宝, 等. 基于深度神经网络模型的雅安市滑坡易发性评价 [J]. 中国地质灾害与防治学报, 2023, 34(3): 157 − 168. [MU Jiaqi, ZHUANG Jianqi, WANG Shibao, et al. Evaluation of landslide susceptibility in Ya'an City based on depth neural network model[J]. The Chinese Journal of Geological Hazard and Control, 2023, 34(3): 157 − 168. (in Chinese with English abstract)]

[21] PAVEL M, NELSON J D, JONATHAN FANNIN R. An analysis of landslide susceptibility zonation using a subjective geomorphic mapping and existing landslides[J]. Computers & Geosciences, 2011, 37(4): 554 − 566.

[22] 穆柯, 谢婉丽, 刘琦琦, 等. 基于 LR-RF 模型的滑坡易发性评价——以铜川市耀州区为例 [J]. 灾害学, 2022, 37(3): 212 − 218. [MU Ke, XIE Wanli, LIU Qiqi, et al. Research on landslide susceptibility evaluation based on logistic regression and LR coupling model[J]. Journal of Catastrophology, 2022, 37(3): 212 − 218. (in Chinese with English abstract)]

[23] 刘坚, 李树林, 陈涛. 基于优化随机森林模型的滑坡易发性评价 [J]. 武汉大学学报(信息科学版), 2018, 43(7): 1085 − 1091. [LIU Jian, LI Shulin, CHEN Tao. Landslide susceptibility assesment based on optimized random forest model[J]. Geomatics and Information Science of Wuhan University, 2018, 43(7): 1085 − 1091. (in Chinese with English abstract)]

[24] HU Qiao, ZHOU Yi, WANG Shixing, et al. Machine learning and fractal theory models for landslide susceptibility mapping: Case study from the Jinsha River Basin[J]. Geomorphology, 2020, 351: 106975.

[25] 黄发明, 陈佳武, 唐志鹏, 等. 不同空间分辨率和训练测试集比例下的滑坡易发性预测不确定性 [J]. 岩石力学与工程学报, 2021, 40(6): 1155 − 1169. [HUANG Faming, CHEN Jiawu, TANG Zhipeng, et al. Uncertainties of landslide susceptibility prediction due to different spatial resolutions and different proportions of training and testing datasets[J]. Chinese Journal of Rock Mechanics and Engineering, 2021, 40(6): 1155 − 1169. (in Chinese with English abstract)]

[26] 王毅, 方志策, 牛瑞卿, 等. 基于深度学习的滑坡灾害易发性分析 [J]. 地球信息科学学报, 2021, 23(12): 2244 − 2260. [WANG Yi, FANG Zhice, NIU Ruiqing, et al. Landslide susceptibility analysis based on deep learning[J]. Journal of Geo-Information Science, 2021, 23(12): 2244 − 2260. (in Chinese with English abstract)]

[27] 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)]

[28] 陈涛, 钟子颖, 牛瑞卿, 等. 利用深度信念网络进行滑坡易发性评价 [J]. 武汉大学学报(信息科学版), 2020, 45(11): 1809 − 1817. [CHEN Tao, ZHONG Ziying, NIU Ruiqing, et al. Mapping landslide susceptibility based on deep belief network[J]. Geomatics and Information Science of Wuhan University, 2020, 45(11): 1809 − 1817. (in Chinese with English abstract)]

［29］ 杨强，王高峰，丁伟翠，等. 多种组合模型的区域滑坡易发性及精度评价［J］. 自然灾害学报，2021，30(2)：36 − 51.［YANG Qiang，WANG Gaofeng，DING Weicui，et al. Susceptibility and accuracy evaluation of regional landsldie based on multiple hybrid models［J］. Journal of Natural Disasters，2021，30(2)：36 − 51.（in Chinese with English abstract）］

［30］ 郭子正，殷坤龙，付圣，等. 基于 GIS 与 WOE-BP 模型的滑坡易发性评价［J］. 地球科学，2019，44(12)：4299 − 4312.［GUO Zizheng，YIN Kunlong，FU Sheng，et al. Evaluation of landslide susceptibility based on GIS and WOE-BP model［J］. Earth Science，2019，44(12)：4299 − 4312.（in Chinese with English abstract）］

［31］ 贾雨霏，魏文豪，陈稳，等. 基于 SOM-I-SVM 耦合模型的滑坡易发性评价［J］. 水文地质工程地质，2023，50(3)：125 − 137.［JIA Yufei，WEI Wenhao，CHEN Wen，et al. Landslide susceptibility assessment based on the SOM-I-SVM model［J］. Hydrogeology & Engineering Geology，2023，50(3)：125 − 137.（in Chinese with English abstract）］

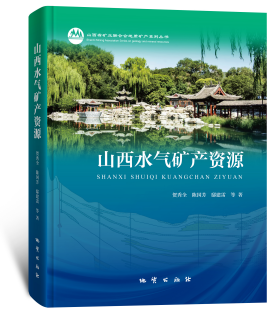
［32］ 武雪玲，杨经宇，牛瑞卿. 一种结合 SMOTE 和卷积神经网络的滑坡易发性评价方法［J］. 武汉大学学报(信息科学版)，2020，45(8)：1223 − 1232.［WU Xueling，YANG Jingyu，NIU Ruiqing. A landslide susceptibility assessment method using SMOTE and convolutional neural network［J］. Geomatics and Information Science of Wuhan University，2020，45(8)：1223 − 1232.（in Chinese with English abstract）］

［33］ 李坤，赵俊三，林伊琳，等. 基于 SMOTE 和多粒度级联森林的泥石流易发性评价［J］. 农业工程学报，2022，38(6)：113 − 121.［LI Kun，ZHAO Junsan，LIN Yilin，et al. Assessment of debris flow susceptibility based on SMOTE and multi-Grained Cascade Forest［J］. Transactions of the Chinese Society of Agricultural Engineering，2022，38(6)：113 − 121.（in Chinese with English abstract）］

［34］ 赵占骜，王继周，毛曦，等. 多维 CNN 耦合的滑坡易发性评价方法［J］. 武汉大学学报(信息科学版)，2024，49(8)：1466 − 1481.［ZHAO Zhan'ao，WANG Jizhou，MAO Xi，et al. A multi-dimensional CNN coupled landslide susceptibility assessment method［J］. Geomatics and Information Science of Wuhan University，2024，49(8)：1466 − 1481.（in Chinese with English abstract）］

编辑：王支农

• 新书介绍 •



《山西水气矿产资源》是由贺秀全等著，地质出版社 2024 年 10 月出版的专著，是山西省内第一部大型综合性关于水气矿产资源的集基础性、实用性于一体的工具书。该书由中国地质调查局水资源调查首席科学家李文鹏先生作序，并给予高度评价，李先生认为："该书是进一步开展山西水文地质勘查研究的基础资料数据库、是全面了解山西省水文地质条件的工具书、是一部全面反映山西省水文地质事业发展历程的地方志书。"

该书分 10 章 48 节，内容涉及基础地质、水文地质、地热资源、浅层地热能、干热岩、天然矿泉水、原生劣质地下水和气体矿产等专业领域，具有综合性、系统性、科学性和艺术性的特点，承载了丰富翔实的珍贵史料，展示了文化大省光辉灿烂的水文地质历史风貌，具有较高的学术价值和实用价值。

该书约 109.2 万字，插图 103 幅，表格 162 张，彩照 44 帧，装帧采用圆脊、精装、夹红色书签带，采用 105 克无光铜版纸彩色双面印刷，页眉图案设计为山西省版图，尽显山西特色及编辑精致；8 处页脚注记增强了本书的知识性；封面设计主色调为由蓝渐变绿，封面上方选用晋祠泉风景照片作为书的背景，体现山西岩溶大泉文化特色，下方以一幅隐含的水文地质剖面图作为背景，巧妙地融入了书的主题。该书的封底图片是晋祠泉的另一幅风景照片，照片下方以注释的形式引出李白咏晋祠泉之《忆旧游寄谯郡元参军》名句："晋祠流水如碧玉，百尺清潭写翠娥"，对本书文化品位的提升起到了画龙点睛的作用，使其更具诗情画意！