https://doi.org/10.3799/dqkx.2021.042

不同环境因子联接和预测模型的滑坡易发性建模不确定性

李文彬[1],范宣梅[2],黄发明[1*],武雪玲[3],殷坤龙[4],常志璐[1]

1. 南昌大学建筑工程学院,江西南昌 330031
2. 成都理工大学地质灾害防治与地质环境保护国家重点实验室,四川成都 610059
3. 中国地质大学地球物理与空间信息学院,湖北武汉 430074
4. 中国地质大学工程学院,湖北武汉 430074

摘　要: 拟深入探讨滑坡与其环境因子间的非线性联接计算以及不同数据驱动模型等因素,对滑坡易发性预测建模不确定性的影响规律.以江西省瑞金市为例共获取370处滑坡和10种环境因子,通过概率统计(probability statistics, PS)、频率比(frequency ratio, FR)、信息量(information value, IV)、熵指数(index of entropy, IOE)和证据权(weight of evidence, WOE)等5种联接方法分别耦合逻辑回归(logistic regression, LR)、BP神经网络(BP neural networks, BPNN)、支持向量机(support vector machines, SVM)和随机森林(random forest, RF)模型共构建出20种耦合模型,同时构建无联接方法直接将原始数据作为输入变量的4种单独LR、BPNN、SVM和RF模型,预测出总计24种工况下的滑坡易发性;最后分别使用ROC曲线、均值、标准差和差异显著性等指标分析上述24种工况下易发性结果的不确定性.结果表明:(1)基于WOE的耦合模型预测滑坡易发性的平均精度最高且不确定性较低,基于PS的耦合模型预测精度最低且不确定性最高,基于FR、IV和IOE的耦合模型介于两者之间;(2)单独数据驱动模型易发性预测精度略低于耦合模型,且未能计算出环境因子各子区间对滑坡发育的影响规律,但其建模效率高于耦合模型;(3)RF模型预测精度最高且不确定性较低,其次分别为SVM、BPNN和LR模型.总之WOE是更优秀的联接法且RF模型预测性能最优,WOE-RF模型预测的滑坡易发性不确定性较低且更符合实际滑坡概率分布特征.

关键词: 滑坡易发性预测;不确定性分析;联接方法;数据驱动;证据权;随机森林;工程地质学.

中图分类号: P642.22　　**文章编号:** 1000－2383(2021)10－3777－19　　**收稿日期:** 2020－11－28

Uncertainties of Landslide Susceptibility Modeling under Different Environmental Factor Connections and Prediction Models

Li Wenbin[1], Fan Xuanmei[2], Huang Faming[1*], Wu Xueling[3], Yin Kunlong[4], Chang Zhilu[1]

1. *School of Civil Engineering and Architecture, Nanchang University, Nanchang* 330031, *China*
2. *State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu* 610059, *China*
3. *Institute of Geophysics & Geomatics, China University of Geosciences, Wuhan* 430074, *China*
4. *Faculty of Engineering, China University of Geosciences, Wuhan* 430074, *China*

基金项目: 国家自然科学基金项目(Nos. 41807285, 41762020, 51879127, 51769014);江西省自然科学基金项目(Nos. 20192BAB216034, 20192ACB2102, 20192ACB20020);中国博士后面上基金项目(Nos.2019M652287, 2020T130274);江西省博士后基金项目(No.2019KY08).

作者简介: 李文彬(1986－),女,博士研究生,研究方向为滑坡易发性预测建模. ORCID: 0000-0001-7831-4120. E-mail: 351113619004@email.ncu.edu.cn

* **通讯作者:** 黄发明,ORCID:0000-0001-9307-9085. E-mail:faminghuang@ncu.edu.cn

引用格式: 李文彬,范宣梅,黄发明,等,2021.不同环境因子联接和预测模型的滑坡易发性建模不确定性.地球科学,46(10):3777－3795.

Abstract: This study aims to explore the influences of some modeling factors including the non-linear correlation calculation between landslides and environmental factors and the different data-based models on the uncertainty law of landslide susceptibility prediction (LSP) modeling. The Ruijin City of Jiangxi Province in China with investigated 370 landslides and 10 environmental factors is used as study case. Accordingly, a total of 20 types of different coupling modeling conditions are proposed for LSP with five different connection methods(probability statistics (PS), frequency ratio (FR), information value (IV), index of entropy (IOE) and weight of evidence (WOE)) and four different data-based models including logistic regression (LR), back propagation neural networks (BPNN), support vector machines (SVM) and random forest (RF). Meanwhile, four single LR, BPNN, SVM and RF models with the original data as input variables are also proposed, as a whole, a total of 24 types of modeling conditions for LSP are obtained based on the above 20 types of coupling conditions and 4 types of single models. Finally, the uncertainty characteristics in the LSP modeling are assessed using the area under the receiver operation curve (ROC), mean value, standard deviation and significance test, respectively. Results show follows. (1) WOE-based models have the highest LSP accuracy and low uncertainty while PS-based models have the lowest LSP accuracy and the highest uncertainty, and the FR, IV and IOE-based models are in between. (2) The single data-based models have slightly lower LSP accuracies than those of the coupling models on the whole and cannot calculate the influence law of each sub-interval of environmental factors on landslide evolution, however, the single data-based models have higher modeling efficiency than those of the coupling models. (3) Among all the data-based models, RF model has the highest LSP accuracy and relatively low uncertainty, followed by the SVM, BPNN and LR models, respectively. It is concluded that the WOE is a very excellent correlation method and the RF model predicts the optimal LSP performance, the LSP results of WOE-RF model have the lowest uncertainties and the predicted landslide susceptibility indexes are more consistent with the actual landslides distribution characteristics.

Key words: landslide susceptibility prediction; uncertainty analysis; nonlinear connection method; data-based model; weight of evidence; random forest; engineering geology.

0 引言

滑坡作为一种非常复杂的自然现象和破坏性的地质事件,对人类的生命财产和社会环境构成严重威胁,因此如何有效开展滑坡易发性制图已成为现阶段滑坡研究的热点(邱海军等, 2020).通过研究影响滑坡易发性建模的相关不确定性因素以便提高易发性预测的可靠性,对滑坡高发地区的防灾减灾工作具有重要意义(吴益平等, 2014;冯杭建等, 2016).

滑坡易发性可定义为特定地点在环境因子非线性综合作用下发生滑坡的空间概率.基于地理相似性规律即"地理环境越相似,地理特征越相近"可知,通过分析过去滑坡发生的环境因子来建立预测模型即可预测将来可能发生滑坡的位置(朱阿兴等, 2020).很明显,从滑坡样本点中确定滑坡易发性与其环境因子的关系是易发性预测的关键所在,因此选择用以获取输入变量的滑坡－环境因子联接方法非常重要,同时也应考虑各联接方法与数据驱动模型耦合建模所产生的诸多不确定性(Huang *et al.*, 2021).随着滑坡编录和环境因子获取技术的快速发展,空间数据质量得到了较大提升(Jacobs *et al.*, 2020; 许强等, 2019).一般而言,具体研究区内的滑坡环境因子类型可通过相关文献综述和研究区的自然地质条件来确定(邱海军等, 2014; 于宪煜等, 2016).本文重点关注滑坡与其环境因子的非线性联接以及基于数据驱动的易发性预测模型这两个建模核心环节.

通常使用数据驱动模型如逻辑回归(logistic regression, LR)(马思远等, 2019)、BP神经网络(back propagation neural networks, BPNN)(郭子正等, 2019)、支持向量机(support vector machines, SVM)(黄发明等, 2019)、C5.0决策树和随机森林(random forest, RF)(张书豪和吴光, 2019; Sun *et al.*, 2020)等预测滑坡易发性.数据驱动模型通过从训练样本中学习滑坡与其环境因子间的非线性关系,能够仅利用输入－输出变量来计算大范围内的滑坡易发性指数(Huang *et al.*, 2021).然而对于哪种模型最适合滑坡易发性预测还没有一致的意见,且即使滑坡易发性精度稍有提高也可能对易发性分区产生显著影响(Huang *et al.*, 2020a; Sun *et al.*, 2020).因此,本文拟分别利用LR、BPNN、SVM和RF等数据驱动模型开展易发性建模对比分析,探索不同数据驱动模型预测易发性的不确定性规律.

各类联接法是将滑坡易发性指数与其环境因

子(不考虑诱发因子)联系起来的重要纽带,其联接性能对数据驱动模型的成功与否至关重要(Hong *et al.*, 2017).目前常用联接法包括证据权(weight of evidence, WOE)(郭子正等, 2019)、信息量(information value, IV)(邱海军等, 2014)、概率法(probability statistics, PS)(张俊等, 2016)、熵指数(index of entropy, IOE)(徐胜华等, 2020)和频率比(frequency ratio, FR)(Huang *et al.*, 2020a)等,但具体选择哪一种联接法没有具体的论据与评估.如文献显示研究人员在使用WOE或者FR时并没有合理解释选用该种联接方法的原因(Chen *et al.*, 2015; Hong *et al.*, 2017; Sun *et al.*, 2020).而且不同联接法会给数据驱动模型预测滑坡易发性带来较大不确定性,联接方法太粗糙会导致信息丢失以至降低模型预测精度;而优秀的联接法能获取更合理的模型输入变量,提高建模的准确性.因此探讨基于不同联接方法的数据驱动模型预测滑坡易发性的不确定性规律具有重要意义.

文献显示有些学者采用不同联接方法和数据驱动模型开展易发性预测建模,如Chen *et al.*(2015)等采用FR、统计指数等联接法分别在喜马拉雅山脉中部地区和宝鸡市宝中地区进行滑坡易发性制图;Xu *et al.*(2020)将IOE融入SVM对陕西省滑坡易发性开展预测建模;张俊等(2016)构建IV-LR模型预测万州区滑坡易发性等.但是大多数情况下,现有研究使用特定的联接方法或数据驱动模型开展易发性建模,而较少提供可信的依据和合理的解释,且较少深入探讨这两种不确定性因素对建模的影响.其实通过分析不同联接方法和数据驱动模型工况下的易发性结果不确定性特征,能更深入地理解滑坡易发性建模的可行性和可靠性,以便降低这些不确定性因素对建模的影响.

综上所述,本文采用PS、FR、IV、IOE和WOE等5种非线性联接值和原始环境因子数据作为LR、BPNN、SVM和RF等4类数据驱动模型的输入变量,以此形成24种不同的建模工况.再以江西省瑞金市为例分别在各建模工况下开展滑坡易发性预测的各种不确定性分析,具体包括建模精度评价、易发性指数差异显著性分析及指数分布规律等方面.

1　滑坡易发性建模分析

建立环境因子联接方法和数据驱动模型耦合工况下的滑坡易发性预测建模具体流程如下(图1):(1)获取研究区滑坡编录及相关环境因子的数据源,构建滑坡易发性预测建模的空间数据集;(2)将上述5种联接法和原始环境因子作为4种模型的输入变量来构成24种建模工况;(3)在每种建模工况基础上分别进行滑坡易发性预测,绘制易发性图并开展建模不确定性分析;(4)采用ROC曲线下面积(area under ROC, AUC)(Huang *et al.*, 2021)对其易发性结果进行精度评估;(5)采用Kendall协同系数检验法在0.05显著性水平下分析各种工况下的易发性指数分布的差异显著性;(6)对24种建模工况下预测的滑坡易发性指数分布特征从均值和标准差的角度进行分析;(7)通过对比得到最佳联接法和数据驱动模型组成的耦合工况,为滑坡易发性预测提供理论指导.

1.1　滑坡与环境因子的联接方法

1.1.1　概率统计法　*PS*法定义为某一环境因子属性区间内滑坡发生的面积与滑坡总面积的比值.*PS*值可作为环境因子对滑坡易发性的贡献依据,越接近1说明该分类对滑坡发育影响越大(公式(1)).式中S_{ij}^{z}为第i类环境因子中第j个状态中的滑坡面积,λ_i为第i类环境因子下的状态数.

$$PS_{ij}=\frac{S_{ij}^{z}}{\sum_{j=1}^{\lambda_i}S_{ij}^{z}}. \tag{1}$$

1.1.2　频率比法　滑坡环境因子采用*FR*作为区间分类依据,表征环境因子各属性区间对滑坡发生的相对影响程度(公式(2)).公式(2)中N_j为具有环境因子在区间中出现滑坡面积,N为研究区内已知滑坡总面积,S_j为环境因子j属性区间的面积,S为研究区总面积.

$$FR=\frac{N_j/N}{S_j/S}. \tag{2}$$

1.1.3　信息量　信息量(*IV*)模型所考虑的是一定地质环境下的最佳滑坡环境因子组合,包括基本因子的数量和基本状态.对于某一具体单元而言,*IV*模型所考虑的是一定区域内所获取的与滑坡相关的所有信息的数量和质量.在具体计算过程中,为计算方便通常将总体概率改用样本频率进行估算,于是*IV*公式可转换为:

$$IV=\ln\frac{N_j/N}{S_j/S}. \tag{3}$$

式中:*IV*为环境因子在j状态下滑坡发生的*IV*,N_j为具有环境因子的区间中出现滑坡的栅格数,N为

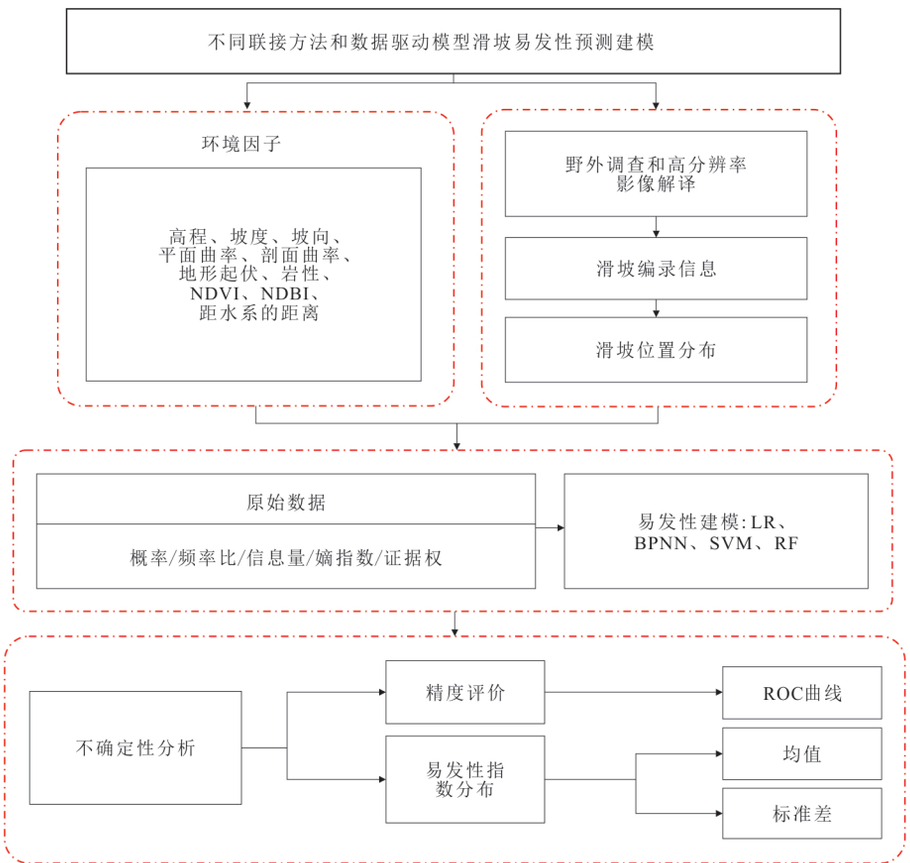

图 1 建模流程

Fig.1 Modeling flow chart

研究区内已知滑坡栅格总数,S_j为环境因子的单元数而S为研究区栅格总数.$IV>0$说明环境因子在状态j下可提供滑坡发生的信息,IV越大则滑坡发生概率越大;$IV<0$说明环境因子在状态j下不利于滑坡发育;$IV=0$说明环境因子在状态j不提供滑坡发育信息.

1.1.4 熵指数 在滑坡易发性预测中,熵代表不同环境因子对滑坡发育的影响程度.通过IOE计算每个输入变量的权重,首先基于FR计算概率密度(P_{ij})如公式(4),P_{ij}为各环境因子分级的FR,S_j表示状态分级数量,i和j分别表示环境因子分级的序列号.

$$(P_{ij})=\frac{P_{ij}}{\sum_{j=1}^{S_j}P_{ij}}. \tag{4}$$

其次将环境因子状态分级下的概率密度(P_{ij})得到各参数熵值,再通过熵值得到该环境因子信息系数I_j如公式(5).最后由I_j与滑坡失效概率耦合以计算参数的最终权重W_j如公式(6).

$$I_j=\frac{\log_2 S_j-\sum_{i=1}^{S_j}\left(P_{ij}\right)*\log_2\left(P_{ij}\right)}{\log_2 S_j},I=(0,1),$$
$$j=1,2,\cdots,n, \tag{5}$$

$$W_j=\frac{I_j}{S_j}\sum_{j=1}^{S_j}P_{ij}. \tag{6}$$

1.1.5 证据权 证据权(WOE)是一种基于贝叶斯准则综合各种证据层来预测某种事件发生概率的定量方法.其将已有滑坡和各个环境因子进行空间联接,得到滑坡处各环境因子的分布情况.权重因子W^+和W^-在每一环境因子分级中的计算如公式(7)和公式(8)所示.式中W^+和W^-分别为环境因子存在区和不存在区的权重值,对于原始数据缺失的区域其权重值为0;B和D分别为环境因子存在区的滑坡和非滑坡单元数,$\overline{B}$和$\overline{D}$分别为环境因子不存在区的滑坡和非滑坡单元数.证据层和滑坡点正相关表示为$W^+>0$和$W^-<0$,而负相关可表示为$W^+<0$和$W^->0$,在不相关或数据缺失时权重为0.相对系数$C=W^+-W^-$用来度量证据图层和滑坡点之间的相关性大小.

$$W^{+}=\ln\left(\frac{B/\left(B+\overline{B}\right)}{D/\left(D+\overline{D}\right)}\right), \tag{7}$$

$$W^{-}=\ln\left(\frac{\overline{B}/\left(B+\overline{B}\right)}{\overline{D}/\left(D+\overline{D}\right)}\right). \tag{8}$$

1.2　数据驱动模型简介

1.2.1　LR 模型　LR模型是用线性回归预测结果去逼近真实标记的对数概率(Zhu *et al.*, 2020).对于滑坡事件来说,LR模型可对分类可能性进行建模而无需事先假设数据分布,其直接得到滑坡发生概率如下公式(9)和公式(10)所示.式中Z为滑坡事件的有效函数,P为滑坡发生的概率,$P\in[0,1]$,B_0为截距,B_i为LR系数,X_i为滑坡环境因子.在滑坡易发性建模中LR模型的作用就是寻找最优的拟合函数来描述滑坡发生与否和一组独立的指标如高度、地层岩性等之间的关系.

$$Z=\mathrm{Logit}(P)=\ln(P/(1-P))=B_0+\sum_{i=1}^{n}B_iX_i, \tag{9}$$

$$P=\frac{\exp(Z)}{1+\exp(Z)}. \tag{10}$$

1.2.2　BPNN 模型　BPNN模型是机器学习中应用最早和最广泛的非线性映射架构之一,其主要由输入层、隐藏层和输出层组成(王智伟等, 2020).研究表明该模型结构可以精确拟合任意非线性函数,已广泛应用于不同的学科领域来解决一些复杂的非线性问题(李云良等, 2015).BPNN模型的每一层由一定数量的神经元组成,神经元通过权值将输入层、隐藏层和输出层连接起来.一般采用误差反向传播算法来确定这些权值,它的基本思想是利用梯度下降算法,使网络的实际输出值和期望输出值之间误差的均方差达到最小.

1.2.3　支持向量机　SVM是为了寻找使类别间距最大化的最佳超平面,并利用超平面上支持向量来构建模型(Huang and Zhao, 2018).对于非线性数据则是通过核函数将其变换到n维特征空间,实现输入变量线性可分.对于一组线性可分训练向量$m_i(i=1,2,\cdots,n)$,m_i表示各环境因子.相应输出类别$y_i=\pm1$,分别表示滑坡和非滑坡.通过确定n维超平面的最大间距来对滑坡进行分类,其最大间距为$\frac{1}{2}\|s\|^2$.对于线性不可分离数据,利用松弛变量ξ_i来控制分类误差,相应的约束条件为$y_i(s\cdot m_i+b)\geqslant1-\xi_i$.此外通过引入$\nu(0,1)$来考虑错误的分类,超平面的距离如公式11所示.其中$\|s\|$为正常超平面的范数;b是常数;λ_i为拉格朗日乘数.

$$L=\frac{1}{2}\|s\|^2-\frac{1}{\nu n}\sum_{i=1}^{n}\xi_i. \tag{11}$$

另外,SVM的核函数主要包括线性、多项式、径向基函数(RBF)和Sigmoid共4种类型,且*RBF*核函数在滑坡易发性预测中有着很好的运用(Xu *et al.*, 2012).本文将广泛应用的RBF核函数作为SVM的非线性映射函数(公式12).

$$f(x)=\omega\cdot x+b=\sum_{i=1}^{n}(\alpha_i-\alpha_i^{*})x_i\cdot x+b. \tag{12}$$

1.2.4　随机森林　RF是由多个决策树组成的集成分类模型,每个决策树都通过投票来选择最优的分类结果(Wang *et al.*, 2018).其原理是首先利用bootstrap抽样从原始训练集中有放回地抽取K个样本,且各样本的特征数都与原始训练集相同;再分别对K个样本建立决策树模型,得到K种分类结果;各样本中随机选取$n(n\leqslant m)$个特征作为分裂特征集,从中选择最优特征对节点进行生长,当$n<m$时每一棵决策树之间又存在差异性.最后形成RF且根据K种分类结果进行投票表决以决定其最终分类.RF模型在训练集的随机性与节点分裂最优属性的两处随机性共同作用下,防止模型的过拟合且增加其稳定性.本文在R studio中通过训练数据集开发了RF模型计算出研究区滑坡易发性值.

1.3　不确定性分析方法

1.3.1　ROC 曲线精度分析　受试者工作特征曲线(receiver operating characteristic, ROC)是一种基于量化指标的评估建模整体性能的指标,可采用ROC曲线下面积AUC值来量化具体的预测精度值.ROC思路是首先计算出滑坡易发性指数值并对测试集中各样例进行排序,然后按此顺序依次选择不同的截断点以便逐个把样例作为正例进行预测,最后将每次计算出当前分类器的“真阳率”和“假阳率”作为ROC曲线的纵轴和横轴绘图.AUC值等于随机挑选的正样本的排名高于随机挑选的负样本的概率,AUC值越大则模型预测性能越好(公式(13)).式中n_0为负样本数量,n_1为正样本数量,r_i表示第i个负样本在整个测试样本中的排序.

$$AUC=\frac{\sum_{i=1}^{n_0}r_i-n_0\times\left(n_0+1\right)/2}{n_0\times n_1}. \tag{13}$$

1.3.2　易发性指数统计规律分析　均值(mean value)反映了滑坡易发性指数分布的平均水平,标准

差表明了易发性指数的离散程度.本文采用均值和标准差从整体上分析易发性指数值的分布特征,揭示不同联接方法和数据驱动模型耦合工况下的预测性能,为易发性预测研究提供理论指导.

Kendall协同系数检验是一个无参数假设检验,其通过计算系数 W 检验 K 个预测结果是否一致.本文使用Kendall检验来分析不同模型预测的滑坡易发性指数分布的差异,原假设是不同模型的预测结果一致.Kendall秩相关系数 W 公式(14),m 是预测模型数,n 是样本数,R_i 第 i 个样本的秩和,$W\in[0,1]$.

$$W=\frac{12}{m^2\left(n^3-n\right)}\sum_{i=1}^{n}\left(R_i-\frac{m\left(n+1\right)}{2}\right)^2. \quad (14)$$

α=0.05的显著水平下当不同模型预测结果一致时 W 等于1;当 W 值小于1时Kendall协同系数应拒绝原假设(原假设的结果差异不显著).当样本容量趋于无穷大时可用公式(15)进行显著性检验.另外通过Kendall秩计算各工况预测易发性指数的平均秩可实现建模性能的排序,若平均秩越靠前则模型性能越好.

$$m\left(n-1\right)W=\frac{12}{mn\left(n+1\right)}\rightarrow\chi_{\alpha}^{2}\left(n-1\right). \quad (15)$$

2 瑞金市简介及环境因子分析

2.1 瑞金简介及滑坡编录

瑞金市位于江西省东南部,介于东经115°42′~116°22′和北纬25°30′~26°20′之间,总面积2 441.4 km^2.该地属亚热带季风湿润型气候,年均降雨量780 mm.区内高度范围为139~1 117 m,地层岩性主要由变质岩、岩浆岩、碎屑岩和碳酸盐构成.根据瑞金市自然资源局地灾资料显示截至2014年底累计发生滑坡、崩塌、泥石流等灾害414起,其中滑坡发生370起,占地灾总数的89%,以中小型滑坡为主(图2).相关文献综述表明区域滑坡分布主要受地形地貌、岩土体类型、距河流距离、地表覆被等因素影响(黄发明等, 2019;Chang *et al.*, 2020a).

2.2 环境因子分析

2.2.1 数据源 根据瑞金滑坡发育特点及相关文献,基于地质图、遥感影像和GIS平台,从数据源中提取10个环境因子(Huang *et al.*, 2020a, 2020b).主要数据源如表1所示,滑坡编录和环境因子均采用30 m分辨率栅格单元进行制图表达.代表滑坡环境因子的数据层有:(1)地形地貌,包括高度、坡度、坡向、剖面曲率、平面曲率和地形起伏度(图3);(2)地表覆被,包括归一化植被指数(normalized difference vegetation index, NDVI)和归一化建筑物指数(normalized difference built-up index, NDBI);(3)水文因子,如距河流距离等;(4)岩土体类型.

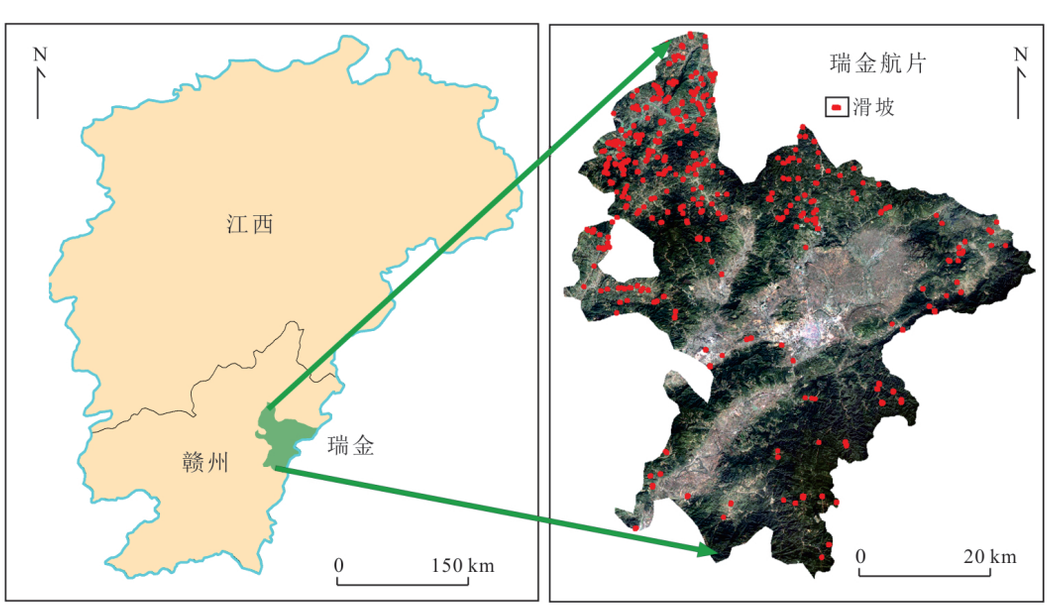

图2 瑞金市地理位置和滑坡编录图

Fig.2 Location of the study area and landslide inventory map

表 1　瑞金市滑坡易发性预测数据源

Table 1　Ruijin landslide susceptibility prediction data source

数据集	空间分辨率	时间	数据用途	数据来源
滑坡编录数据库		2014－12－30	瑞金市滑坡分布	江西省自然资源厅
DEM	30 m	2016－06－06	地形因子	来源于网站 http://solargis.cn/imaps/
Landsat 8 TM	多光谱 30 m	2013－10－15	NDVI, MNDWI, NDBI	中科院对地观测中心 http://ids.ceode.ac.cn/index.aspx
地层岩性分布图	1:50 000	2014－12－30	岩土类型	江西省自然资源厅

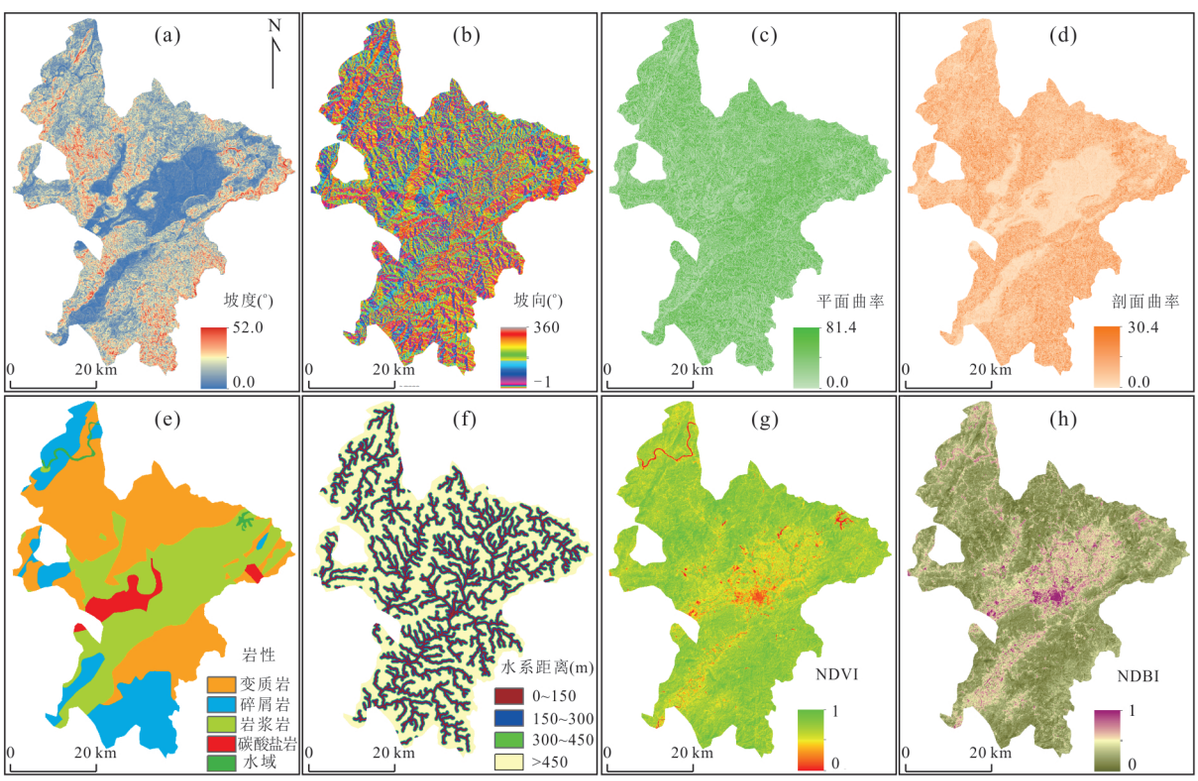

图 3　环境因子(a～h)(高度和地形起伏度未列入)

Fig.3　The environmental factors (a－h) (elevation and topographic relief are not included)

2.2.2　环境因子及其联接分析　获取研究区滑坡编录及其环境因子之后,本文将连续型环境因子按自然间断点法分成 8 个状态级或子区间(Huang *et al.*, 2020b, 2021).其中平地的坡向取值为－1,单独分为一类;离散型岩土体类型和距河流距离按实际确定状态分级.

(1) 地形地貌.地形地貌因子均从 DEM 中提取(图 3).以坡度为例(表 2),滑坡发生概率与坡度成正态分布形式且在 14°时滑坡概率接近峰值;FR 在坡度位于 7°～26.8°的范围内均大于 1.更进一步在 7°～14°的坡度范围内,滑坡发生概率均大于 0.25,FR 分别为 1.651 和 1.592,IV 和 WOE 都显示出较强正相关性,IOE 显示坡度具有仅次于岩性和距离水系距离的权重且其值为 0.09.可见各种联接方法在表达滑坡与坡度的非线性联接性时具有较为一致的趋势和计算效果.

(2) 水文环境和基础地质.由于河流对边坡的浸润和侵蚀作用,越靠近河流的边坡的土壤含水量越高,导致斜坡岩土体的稳定性越差(Li *et al.*, 2017; Chang *et al.*, 2020b; Huang *et al.*, 2020c).本文利用距水系距离来表征水文环境对滑坡发育的影响(Guo *et al.*, 2019).通过统计计算,距水系的距离小于 300 m 的区域滑坡最为集中达到 67%,IOE 显示距离水系距离因子的权重值为 0.15.

岩土类型表征滑坡体的物质基础,表 2 显示在变质岩条件下滑坡发生概率高达 59.3%,*FR* 为 1.338;而碎屑岩条件下滑坡发生概率为 29.1% 且 *FR* 为 1.587,且在这两种岩性下滑坡的 *IV* 和 *WOE*

表 2 各环境因子的联接值计算

Table 2 The connection values of environmental factors

环境因子	变量值	全区栅格数	滑坡栅格数	*PS*	*FR*	*IV*	*WOE*	*IOE*
高度(m)(连续性)	139～293	730 572	1 939	0.354	1.332	0.124	0.414	0.035
	293～308	647 032	1 563	0.285	1.212	0.084	0.260	
	308～373	558 257	964	0.176	0.866	−0.062	−0.178	
	373～446	369 863	587	0.107	0.796	−0.099	−0.260	
	446～534	231 817	254	0.046	0.550	−0.260	−0.640	
	534～642	121 414	98	0.018	0.405	−0.393	−0.932	
	642～780	66 004	44	0.008	0.334	−0.476	−1.113	
	780～1118	25 732	33	0.006	0.643	−0.191	−0.445	
坡度(°)(连续性)	0～3.6	569 695	51	0.009	0.045	−1.348	−3.328	0.083
	3.6～7.0	490 091	537	0.098	0.550	−0.260	−0.693	
	7.0～10.6	532 865	1 396	0.255	1.315	0.119	0.352	
	10.6～14.0	438 190	1 442	0.263	1.651	0.218	0.634	
	14.0～17.6	338 424	1 074	0.196	1.592	0.202	0.553	
	17.6～21.6	221 097	613	0.112	1.391	0.143	0.366	
	21.6～26.8	121 534	300	0.055	1.239	0.093	0.225	
	26.8～52.0	38 795	69	0.013	0.892	−0.049	−0.116	
坡向(连续性)	−1.0	499	0	0	0	0	0	0.054
	0～22.5 337.5～360.0	324 822	668	0.122	1.032	0.014	0.035	
	22.5～67.5	297 924	585	0.107	0.985	−0.006	−0.017	
	67.5～112.5	354 479	943	0.172	1.335	0.125	0.340	
	112.5～157.5	359 791	816	0.149	1.138	0.056	0.150	
	157.5～202.5	332 830	695	0.127	1.048	0.020	0.053	
	202.5～247.5	332 143	620	0.113	0.937	−0.028	−0.075	
	247.5～292.5	378 011	655	0.119	0.869	−0.061	−0.161	
	292.5～337.5	370 192	500	0.091	0.678	−0.169	−0.439	
平面曲率(连续性)	0～10.2	448 550	1 544	0.282	1.727	0.237	0.700	0.07
	10.2～18.8	523 511	1 430	0.261	1.371	0.137	0.407	
	18.8～27.8	429 580	1 002	0.183	1.170	0.068	0.189	
	27.8～37.7	347 255	632	0.115	0.913	−0.039	−0.104	
	37.7～48.2	272 547	343	0.063	0.631	−0.200	−0.500	
	48.2～59.1	223 692	126	0.023	0.283	−0.549	−1.327	
	59.1～70.6	205 059	121	0.022	0.296	−0.529	−1.274	
	70.6～81.4	300 497	284	0.052	0.474	−0.324	−0.810	
剖面曲率(连续性)	0～1.5	671 767	858	0.157	0.641	−0.193	−0.556	0.009
	1.5～3.2	695 534	1 628	0.297	1.174	0.070	0.221	
	3.2～4.8	534 972	1 244	0.227	1.167	0.067	0.195	
	4.8～6.6	378 519	827	0.151	1.096	0.040	0.107	
	6.6～8.7	243 529	507	0.092	1.045	0.019	0.048	
	8.7～11.0	138 110	258	0.047	0.937	−0.028	−0.068	
	11.0～14.4	68 229	134	0.024	0.985	−0.006	−0.015	
	14.4～30.4	20 031	26	0.005	0.651	−0.186	−0.432	
地形起伏度(°)(连续性)	0～5.5	588 993	151	0.028	0.129	−0.891	−2.266	0.061
	5.5～11.0	611 553	1 116	0.204	0.916	−0.038	−0.113	
	11.0～16.1	562 924	1 572	0.287	1.401	0.147	0.447	
	16.1～21.3	428 304	1 289	0.235	1.510	0.179	0.512	
	21.3～26.8	287 508	718	0.131	1.253	0.098	0.256	

续表2

环境因子	变量值	全区栅格数	滑坡栅格数	*PS*	*FR*	*IV*	*WOE*	*IOE*
	26.8～33.4	170 550	411	0.075	1.209	0.082	0.204	
	33.4～42.6	80 923	204	0.037	1.265	0.102	0.244	
	42.6～93.8	19 936	21	0.004	0.529	−0.277	−0.642	
地层岩性(离散型)	变质岩	1 218 584	3 249	0.593	1.338	0.126	0.603	0.259
	碎屑岩	503 748	1 593	0.291	1.587	0.201	0.603	
	岩浆岩	899 363	359	0.065	0.200	−0.698	−1.939	
	碳酸盐	107 442	136	0.025	0.635	−0.197	−0.469	
	水域	21 554	145	0.026	3.376	0.528	1.240	
NDVI(连续性)	0～0.014	15 215	13	0.002	0.429	−0.368	−0.851	0.029
	0.014～0.120	51 765	38	0.007	0.368	−0.434	−1.012	
	0.120～0.190	104 953	144	0.026	0.688	−0.162	−0.386	
	0.190～0.243	233 124	546	0.100	1.175	0.070	0.178	
	0.243～0.284	487 817	1 149	0.210	1.182	0.073	0.207	
	0.284～0.322	723 461	1 508	0.275	1.046	0.019	0.061	
	0.322～0.363	734 588	1 503	0.274	1.027	0.011	0.035	
	0.363～1	399 768	581	0.106	0.729	−0.137	−0.362	
NDBI(连续性)	＜0	473 698	651	0.119	0.690	−0.161	−0.435	0.009
	0～0.061	820 803	1 550	0.283	0.948	−0.023	−0.077	
	0.061～0.126	616 979	1 541	0.281	1.253	0.098	0.302	
	0.126～0.201	339 466	873	0.159	1.290	0.111	0.297	
	0.201～0.286	225 388	496	0.090	1.104	0.043	0.109	
	0.286～0.374	145 844	262	0.048	0.901	−0.045	−0.110	
	0.374～0.482	90 590	95	0.017	0.526	−0.279	−0.659	
	0.482～1	37 923	14	0.003	0.185	−0.732	−1.699	
距水系距离(m)(离散型)	＜150	508 453	2 135	0.389	2.107	0.324	1.036	0.150
	150～300	464 069	1 573	0.287	1.701	0.231	0.685	
	300～450	420 058	480	0.088	0.573	−0.242	−0.632	
	＞450	1 358 111	1 294	0.236	0.478	−0.320	−1.152	

都表现正相关性(Liu *et al.*, 2019).其余岩土体类型地区滑坡分布较少,占全区最多的岩浆岩反而滑坡发生概率较小.IOE显示岩性具有最高权重值为0.259,总之滑坡在变质岩和碎屑岩等地区相对高发而岩浆岩地区相对低发.另外本区碳酸盐岩分布较少且其*FR*值为0.6.

(3)地表覆被.选择NDVI和NDBI作为地表覆被因子(图3g, 3h).NDBI反映研究区域内的工程建筑和植被对滑坡发育的影响.从结果看研究区内NDBI取值位于0.126～0.201之间时与滑坡有较强的关系,其*PS*值为0.159、*FR*值为1.290、*IV*值为0.111,且*WOE*值为0.297.NDVI可定量估计植被生长和生物量.在本研究中NDVI值位于0.190～0.0.284范围时滑坡发生的概率较大.

3　瑞金滑坡易发性预测建模

3.1　数据准备

本文采用30 m分辨率栅格作为滑坡预测单元,整个研究区被划分为2 750 691个栅格单元.通过上述5种不同联接方法给10个环境因子重新赋值,作为数据驱动模型输入变量;另外对于单独数据驱动模型的建模,也将原始环境因子值作为模型输入变量.已发生的370处滑坡被划分为5 482个栅格单元(赋值为1),同时随机挑选与滑坡栅格相同数量的非滑坡栅格(赋值为0),并作为模型输出变量.在滑坡和非滑坡栅格中按7:3随机划分得到模型训练集和测试集.最后将整个研究区栅格单元的5种赋值结果代入训练好的模型中以便预测研究区滑坡易发性指数,并将其按自然间断点法划分为5个易发性级别(Huang *et al.*, 2021).

表 3 各工况下 LR 系数和常数项

Table 3 Logistic regression coefficients and constant terms

环境因子	单独 LR	PS-LR	FR-LR	IV-LR	IOE-LR	WOE-LR
高度	−0.005	4.103	1.534	3.093	11.803	1.166
坡度	0.139	5.187	1.238	1.843	13.144	0.724
坡向	−0.001	5.366	1.074	2.384	10.451	0.911
剖面曲率	−0.015	4.798	0.775	1.553	6.533	0.588
平面曲率	0.017	1.124	0.663	0.872	6.062	0.319
地形起伏度	−0.019	0.031	0.299	0.288	2.850	0.105
地层岩性	−0.434	1.977	1.135	1.848	10.465	0.616
NDVI	0.807	−0.767	0.170	−0.224	1.259	−0.085
NDBI	2.579	1.723	1.160	2.554	9.973	0.987
水系距离	0	5.191	0.794	2.012	5.474	0.594
常数	2.598	−6.018	−9.805	−0.003	−10.337	−0.178

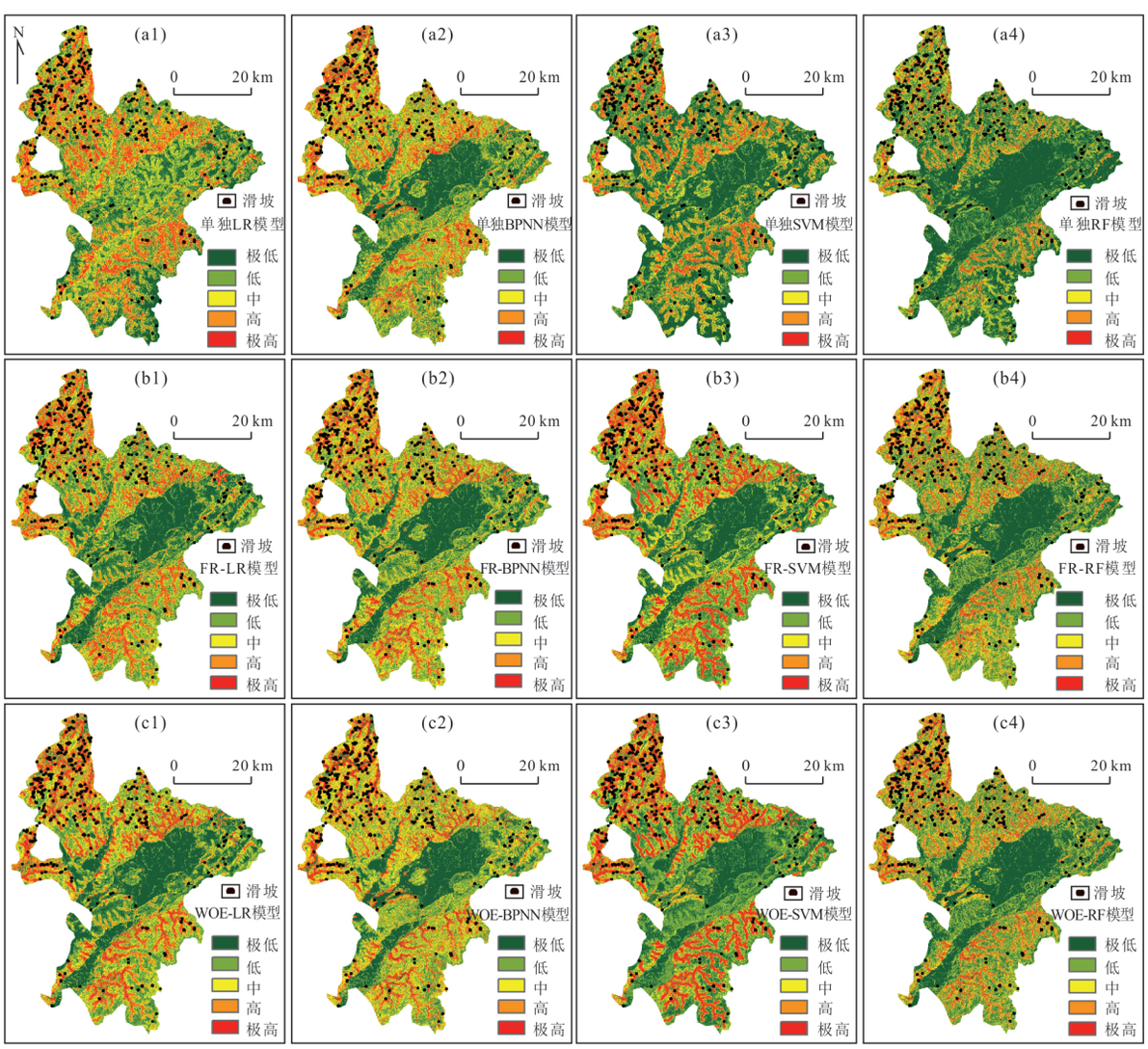

图 4 基于数据驱动模型的滑坡易发性制图

Fig.4 Landslide susceptibility maps of data-based models

a1.单独 LR 模型;a2.单独 BPNN 模型;a3.单独 SVM 模型;a4.单独 RF 模型;b1～b4.FR-based 模型;c1～c4.WOE-based 模型

3.2　滑坡易发性预测结果

3.2.1　LR 模型预测易发性　分别利用训练集中 8 251 个滑坡－非滑坡样本和测试集中 2 713 个滑坡－非滑坡样本进行 LR 建模,得到每个环境因子的回归系数.回归系数越大表明相应的环境因子对滑坡发育的贡献度越高.不同建模工况下的 LR 系数如下表 3 所示,将得到的回归系数和联接值代入公式(9)和公式(10)中,即可预测出每个栅格单元的滑坡易发性指数.

3.2.2　BPNN, SVM 和 RF 预测易发性　对于 BPNN 和 SVM 模型,首先将训练集和测试集数据导入 SPSS modeler 18.0 软件中.BPNN 使用 boosting 算法创建一个整体并由其生成模型序列以增强模型准确度,采用单隐藏层 BPNN 预测滑坡易发性.WOE 联接值作为训练数据,得到 WOE-BPNN 最佳隐藏层神经元个数为 12,用于 boosting 的模型数量为 10,采用梯度下降算法进行优化,其他参数采用默认值,其他联接方法参数设置较为区别不大,在此未列出.

同时利用 SPSS modeler 18.0 软件中的建模节点训练和测试 SVM 模型,选用径向基函数 RBF 作为核函数,采用交叉验证法得到 WOE-SVM 模型参数 C_0、ε 和 γ 分别为 10、0.1 和 0.5,其他联接方法下 SVM 模型参数波动不大,在此未一一列出.然后将训练好的 SVM 模型预测所有栅格单元的易发性.

对于 RF 模型,利用 R 语言循环迭代计算不同 RF 袋外误差,袋外误差越小则对应模型预测的精度越高.经过交叉验证分析得到 WOE-RF 下最优的随机特征数为 3,RF 决策树数目为 800,其他连接方法耦合 RF 模型参数与 WOE-RF 较为一致.最后同样用训练测试好的 RF 模型进行滑坡易发性预测.

3.3　滑坡易发性制图表达

本文在 24 种建模工况下分两步开展滑坡易发性预测.首先将各建模工况下分别预测出的滑坡易发性指数导入 ArcGIS 10.3 软件中,然后根据自然间断点法将研究区划分为 5 个易发性等级区间.在 24 种建模工况中展示部分典型易发性制图结果,其

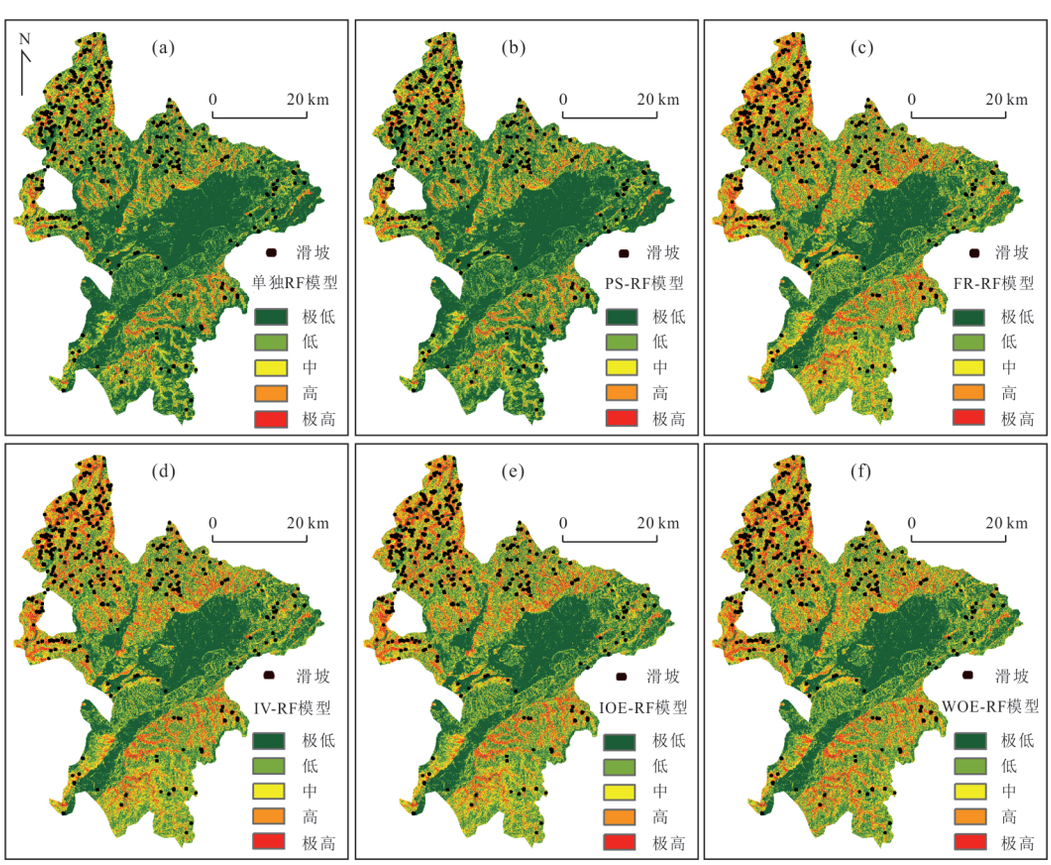

图 5　基于 RF 模型的易发性制图

Fig.5　Landslide susceptibility maps

a.单独 RF 模型;b. PS-RF 模型;c.FR-RF 模型;d.IV -RF 模型;e.IOE-RF 模型;f.WOE-RF 模型

中 FR-based 模型、WOE-based 模型和单独的 4 种数据驱动模型预测的易发性结果如图 4 所示，原始数据、PS、FR、IV、IOE 和 WOE 与 RF 模型预测的易发性结果如图 5 所示.

从图 4 和图 5 可知，瑞金市大部分地区处于低和极低易发区域，且滑坡高易发区主要位于距离水系 300 m 范围之内以及坡度和高度中等的山地丘陵等地区，这与野外调查结果相符.图 5 中可见，同一种数据驱动模型下 5 种不同联接方法和原始数据得到的滑坡易发性级别差异显著，且低和极低易发区面积差异更大，尽管各种建模工况下预测易发性的 AUC 精度差异不大.

4 建模结果不确定性分析

4.1 ROC 精度评价

采用测试集 AUC 值作为具体指标量化不同预测模型的预测性能，AUC 值越大，意味着预测模型整体预测性能越好(Li *et al.*, 2020).24 种建模工况下得到的测试结果的 ROC 曲线如图 6 所示，从图中可知，FR、IV、IOE 和 WOE 在同一种数据驱动模型中结果较为一致且相对稳定；单独 BPNN、SVM 和 RF 模型的易发性预测精度比单独 LR 模型高，另外单独模型与基于联接方法的耦合模型的易发性预测结果较为接近.单独 RF 模型效果最好且较其他耦合模型精度提高 2%，其建模效率也较高.基于 PS 的耦合模型预测效果最差，如表 4 所示.同时所有滑坡易发性预测结果对比再次证明 RF 模型较其他数据模型有更好的预测效果，且机器学习模型比传统统计模型 LR 的 AUC 精度提高了 10%(图 7).

表 4 基于不同联接方法和数据驱动模型的 AUC 值

Table 4 AUC values of different connection methods and original value under different data-based models

预测模型	AUC 值				
	RF 模型	SVM 模型	BPNN 模型	LR 模型	平均 AUC
无联接	0.922	0.809	0.838	0.781	0.838
PS	0.906	0.817	0.806	0.779	0.827
FR	0.905	0.836	0.840	0.832	0.853
IV	0.907	0.838	0.838	0.838	0.855
IOE	0.905	0.837	0.839	0.833	0.854
WOE	0.896	0.839	0.843	0.838	0.857

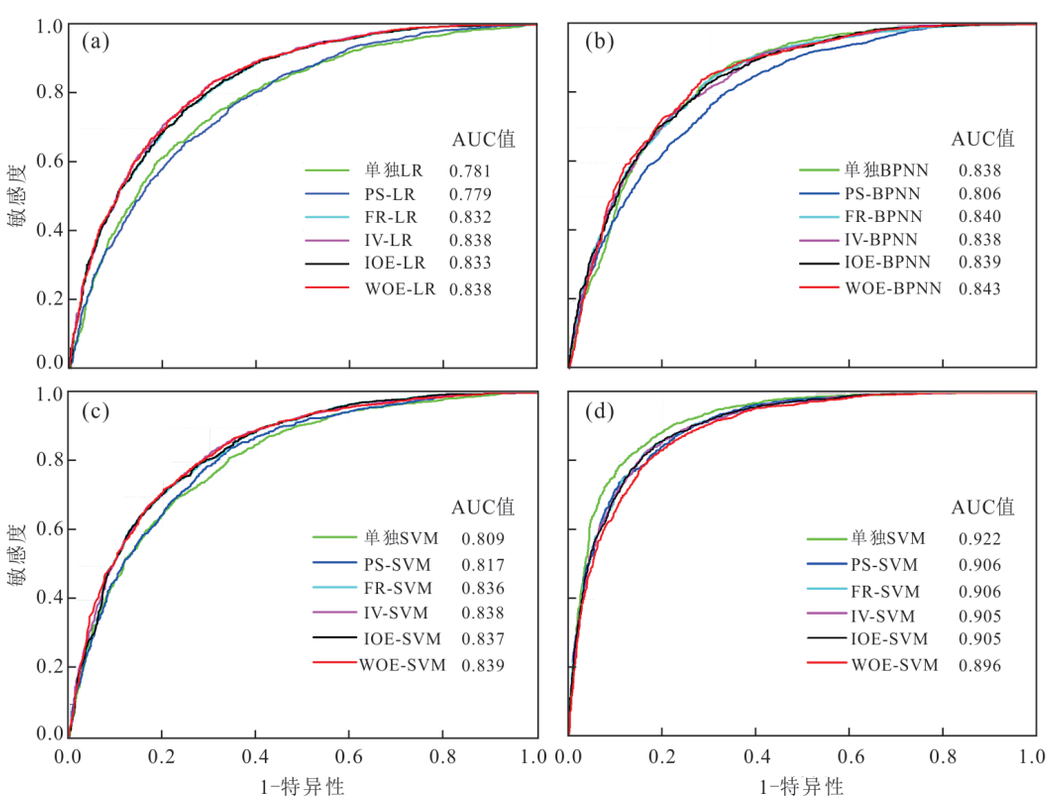

图 6 不同组合工况下的滑坡易发性建模 ROC 曲线

Fig.6 ROC curves of LSP under different conditions

a.LR；b.BPNN；c.SVM；d.RF models

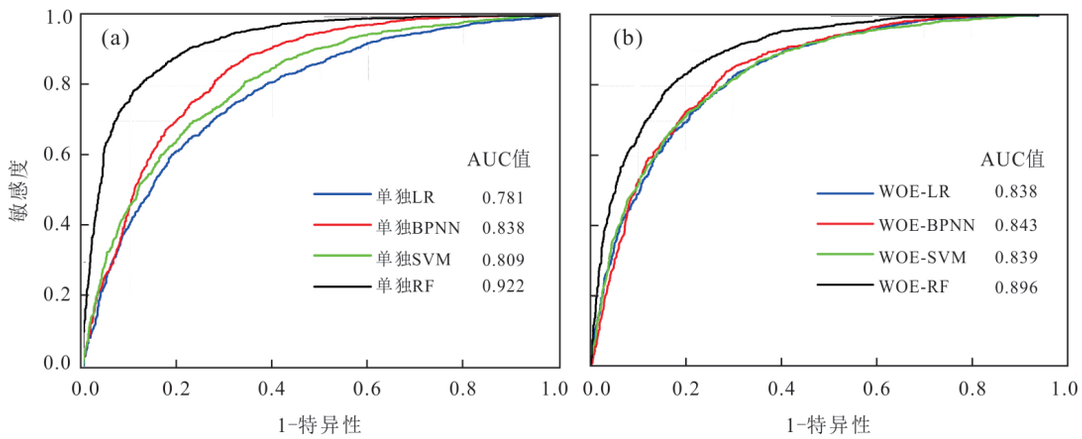

图 7　基于数据驱动模型和基于 WOE-based 模型的 ROC 曲线

Fig.7　ROC curves of data-based and WOE-based models

a.Data-based 模型;b.WOE-based 模型

4.2　滑坡易发性指数分布规律

本文采用均值(mean value)和标准差(standard value)分别反映滑坡易发性指数分布的平均水平和离散程度,并以此分析建模工况下的易发性预测不确定性.

(1)各耦合模型预测的易发性指数分布的不确定性规律较为一致,WOE 耦合模型的易发性指数按均值大小排名为:$Mean_{(WOE\text{-}BPNN)}>Mean_{(WOE\text{-}SVM)}>Mean_{(WOE\text{-}LR)}>Mean_{(WOE\text{-}RF)}$(图 8 和表 5).其中 SVM 和 BPNN 模型的易发性指数分布规律较为一致,这两个模型预测的易发性指数普遍偏大且预测效果低于 LR 和 RF 模型.结合易发性预测的 AUC 精度值,可见这两个模型识别滑坡的能力较低.另外 RF 和 LR 模型的易发性指数分布规律较相似,在极低和低易发区分布较集中,而在高和极高易发区分布逐渐减少.另外这 4 个模型的离散程度正好与其均值大小相反($SD_{(WOE\text{-}RF)}>SD_{(WOE\text{-}LR)}>SD_{(WOE\text{-}SVM)}>SD_{(WOE\text{-}LR)}$),表明 WOE 耦合 RF 和 LR 模型对研究区滑坡易发性的区分度较好,能很好地反映不同栅格单元易发性指数的差异,且用较少的高易发性指数反映尽可能多的滑坡编录信息.

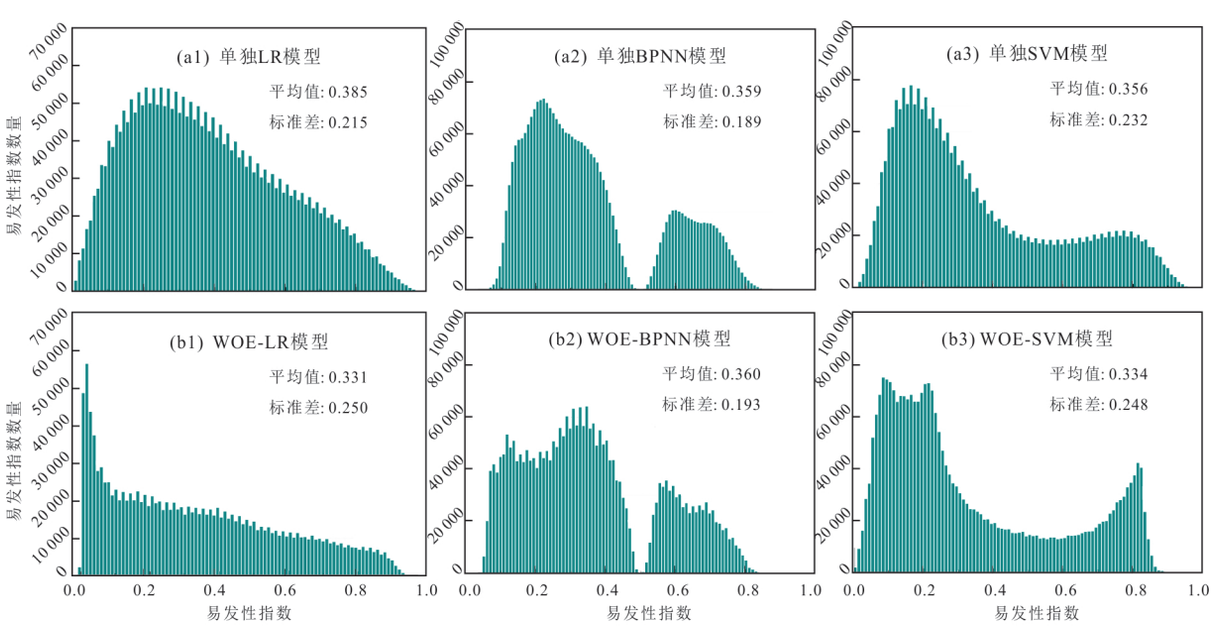

图 8　滑坡易发性分布

Fig.8　Landslide susceptibility index distributions

a1～a3.单独模型;b1～b3.WOE-based 模型

表5 基于不同连接方式和不同数据模型下的平均值和标准差

Table 5 Mean and standard deviation of different connection methods and original value under data-based models

预测模型	RF模型		SVM模型		BPNN模型		LR模型	
	平均值	标准差	平均值	标准差	平均值	标准差	平均值	标准差
原始	0.263	0.240	0.355	0.233	0.358	0.189	0.385	0.215
PS	0.279	0.250	0.344	0.247	0.398	0.161	0.383	0.211
FR	0.278	0.254	0.331	0.252	0.376	0.173	0.337	0.242
IV	0.278	0.255	0.335	0.249	0.367	0.182	0.331	0.251
IOE	0.278	0.254	0.330	0.250	0.367	0.189	0.336	0.246
WOE	0.283	0.261	0.334	0.249	0.359	0.193	0.331	0.251

(2)以RF模型为例分析单独模型和耦合模型预测出的滑坡易发性指数分布规律,如表5和图9所示.不同联接方法预测的易发性指数均值大小排名为:$Mean_{(WOE\text{-}RF)}>Mean_{(IOE\text{-}RF)}>Mean_{(FR\text{-}RF)}>Mean_{(IV\text{-}RF)}>Mean_{(PS\text{-}RF)}>Mean_{(单独RF)}$;其标准差大小排名为:$Standard_{(WOE\text{-}RF)}>Standard_{(IOE\text{-}RF)}>Standard_{(FR\text{-}RF)}>Standard_{(IV\text{-}RF)}>Standard_{(PS\text{-}RF)}>Standard_{(单独RF)}$.上述对比可知,单独模型预测的易发性指数的均值和标准差均最小;WOE-RF的易发性指数均值较小,标准差较大;IOE-RF、FR-RF和IV-RF模型的结果较为一致,而PS-RF效果最差,其均值最大且标准差最小.LR、BPNN和SVM均出现与RF模型类似的易发性指数规律.单独BPNN、SVM和RF模型预测的易发性指数分布规律与WOE耦合模型较为一致.但是单独LR模型预测的易发性指数区分度较差,其规律性与WOE耦合模型差异较大.

4.3 各建模工况下易发性指数的差异性

采用Kendall协同系数检验法,对任意两组不同建模工况下预测的易发性指数进行差异显著性检验.若Kendall秩相关系数W小于1及检验结果的显著性小于0.05,说明这两组工况下易发性指数的差异是显著的,拒绝原假设.本文通过成对因子显著性检验发现W值为0.139,且P值均小于0.05,可见各建模工况下的易发性指数间差异显著.

同时计算各建模工况下预测的易发性指数的平均秩,以便对易发性模型性能排序.平均秩越小则模型性能越好,最终模型比较结果如表6.单独RF模型预测的易发性指数的平均秩最小,且WOE-based模型预测易发性指数的平均秩在同一种数据驱动模型中均比较小.基于FR、IV和IOE的数据驱动模型预测易发性指数平均秩较为一致,PS-BPNN模型预测的平均秩最大.显著性差异水平和平均秩显示出各建模工况的易发性预测存在不确定性,利用这些检验结果来规避建模不确定性具有重要意义.

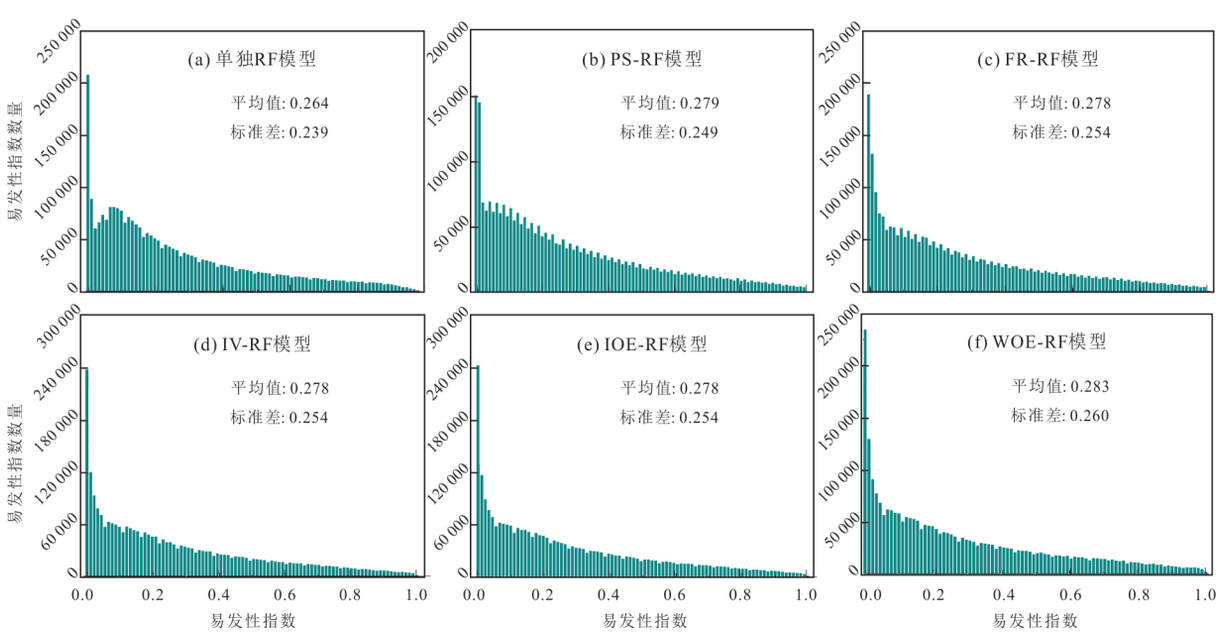

图9 基于RF的滑坡易发指数分布

Fig.9 Susceptibility index distributions

a.单独RF;b.PS-RF;c.FR-RF;d.IV-RF;e.IOE-RF;f.WOE-RF

表 6　各建模工况下易发性指数的平均秩

Table 6　Mean rank of different connection methods under different data-based models

预测模型	平均秩			
	RF 模型	SVM 模型	BPNN 模型	LR 模型
PS	8.77	13.12	16.87	15.82
FR	8.69	11.87	16.08	12.58
IV	8.64	12.38	15.39	11.90
IOE	8.64	11.85	15.30	12.43
WOE	8.97	12.48	14.79	12.06
原始数据	8.08	13.74	14.65	14.90

5　讨论

5.1　不同联接方法的建模不确定性

环境因子各属性区间对滑坡易发性的空间影响可通过联接方法进行定量统计,在诸多研究中将其结果作为数据驱动模型的输入变量(Devkota *et al.*, 2013; Zhu *et al.*, 2021).不同联接方法下的环境因子分级中,WOE 比另外 4 种联接法更能反映环境因子内部影响滑坡发育的空间信息的差异,具有更优的平均预测精度(AUC=0.857);FR 相较于 IV 和 IOE 法更加直观,在保证预测精度的同时有效避免太复杂的统计计算;PS 法反映滑坡对各环境因子子区间的贡献率,但未能充分体现滑坡与各环境因子子区间的空间相关性.Regmi *et al.* (2014)和 Hong *et al.* (2017)等学者对上述部分联接方法反映滑坡与其环境因子空间联接的性能进行了对比分析,所得结果与本文结果较为一致.由上述分析可知,环境因子与滑坡间的空间信息的联接性表达越充分则易发性指数的区分度越大,进一步的易发性预测效果就越佳.

此外 PS、FR、IV、IOE 和 WOE 等 5 种联接方法耦合数据驱动模型后预测出的易发性指数的平均值逐渐减小而标准差逐渐增大,且他们的平均秩也逐渐减小,可见这 5 种联接方法的滑坡易发性建模效果依次更优.本文也对单独数据驱动模型的预测精度和易发性指数统计规律进行分析,结果显示单独 LR 模型的易发性预测效果最差,而在更先进的数据驱动模型(如 RF)中,其精度逐渐提高.

5.2　不同数据驱动模型的建模不确定性

对于同一种联接方法与不同数据驱动模型耦合工况下(图 8),各模型预测精度均表现出一致性的规律:$AUC_{RF}>AUC_{BPNN}>AUC_{SVM}>AUC_{LR}$,该结果反映出机器学习模型预测精度高于常规的回归模型.从易发性指数特征分析:RF 模型预测的滑坡易发性成指数分布规律(图 9),易发性均值处于极低和低分布区间过渡带,且滑坡编录大部分落在极高和高易发性区间;BPNN 预测的滑坡易发性指数出现驼峰状,其精度虽高但稳定性较差,且非常依赖滑坡样本数量;而 SVM 预测的滑坡易发性指数趋向于偏正态分布且主要处于低易发区间(图 8), SVM 预测出的易发性指数相较于 BPNN 模型更集中分布在极低和低易发性区间.总体而言,越先进的机器学习模型其易发性预测的可靠性越高,不确定性越低.

5.3　建模不确定性的综合分析

从耦合模型角度看,WOE-RF 模型预测效果最好,而 PS-LR 模型预测效果最差.另外 PS-RF 模型同样也能达到良好预测精度(AUC=0.906),表明与统计模型相比机器学习能充分高效挖掘不完备信息且易发性建模训练测试效果优秀(Pham *et al.*, 2017).更进一步,RF 模型具有较强稳定性而 LR 模型依赖于联接方法,联接方法的统计规律越明显则 LR 模型预测的效果越好.

RF 是一种基于决策树的有监督集成学习算法,可以处理非线性数据和高维数据,不需要进行特征选择.由于袋外数据的存在,RF 模型在树生成过程中得到了真实误差的无偏估计,训练数据不会丢失.随着样本和特征随机性的引入,RF 在测试过程中具有一定的抗噪声和抗过拟合能力.同时 RF 既能处理离散数据又能处理连续型数据,可见其对数据集适应性强,很适合用于同时包含离散型和连续型变量的易发性建模过程.

采用 Kendall 协同系数检验法验证各工况下的易发性预测效果的差异显著性.基于 WOE 的耦合模型得到的滑坡易发性指数与基于其他联接方法的耦合模型相比具有显著差异.RF 模型与 5 种联接方法耦合相比,其他数据驱动模型预测性能差异显著,单独 RF 模型预测结果与其他预测结果相比差异最为明显,平均秩最小且预测效果最好.

5.4　研究的普适性及存在的问题

本研究中的 24 种建模工况都只是应用于江西瑞金市的滑坡易发性预测,由此得到的关于不同联接方法和数据驱动模型的研究结论是否适用于其他地区?是一个值得深入探索的问题.对此问题以文献综述的方式开展进一步研究,将本文研究结果与重点文献中的结果进行对比分析.如 Chen *et al.*

(2018)、Regmi *et al.* (2014)、徐胜华等(2020)和张玘恺等(2020)等学者分别在陕西省太白县、尼泊尔喜马拉雅地区、三峡库区万州区、陕西省和九寨沟地区采用IOE-RF、IOE-SVM、IOE-LR、WOE-BPNN和IV-LR等不同耦合模型开展滑坡易发性预测建模研究,结果显示,在上述研究区中耦合模型的滑坡易发性预测精度均高于单独模型;另外Chen *et al.*(2017)、Huang *et al.*(2020a,2021)、吴润泽等(2021)、Merghadi *et al.*(2020)和Saha *et al.*(2020)等学者在陕西省陇县、江西省石城县、上犹县、三峡库区湖北段、米拉盆地和印度古马尼河流域等地区,用数据驱动模型预测的滑坡易发性结果显示RF模型性能优于LR, BP和SVM等更传统的数据驱动模型.可见本文在瑞金市的滑坡易发性建模研究结果与诸多学者在其他研究区的结果整体上是一致的,本研究结果可为其他研究区内的滑坡联接方法和数据驱动模型的耦合方式提供指导和借鉴,以降低易发性预测建模的不确定性.

一般而言,性能优异的数据驱动模型预测出的滑坡易发性ROC精度会较高而不确定性会较低.本文只有在少部分工况下出现滑坡易发性预测的ROC精度高而不确定性也高的情况,这种情况可能与预测出的滑坡易发性指数空间分布不合理有关.在实际应用中出现这种情况时可认为该滑坡易发性预测结果的可靠性有待提高,有必要考虑该建模过程中是否还存在其他不确定性因素,以便从整体上改善滑坡易发性建模.

6 结论

本文研究结论显示WOE耦合数据驱动模型的易发性预测结果可靠性最高,预测出的滑坡易发性指数的不确定性较低且更符合实际的滑坡概率分布特征.具体结论分析如下:

(1)在各耦合模型中基于WOE的数据驱动模型预测滑坡易发性的平均精度最高,均值和平均秩较小且标准差较大;基于FR、IV和IOE的数据驱动模型预测滑易发性的平均精度相似但低于WOE-based模型,这些耦合模型的均值和平均秩较大且标准差较小;而基于PS的数据驱动模型预测易发性效果最差.可见WOE具有比其他4类联接法更优秀的非线性联接性能.

(2)将原始环境因子数据直接用作输入变量的单独数据驱动模型预测的滑坡易发性精度整体而言略低于耦合模型,为了提高易发性建模效率可直接使用单独数据驱动模型.但要体现滑坡与其环境因子的空间联接性或分析环境因子各子区间对滑坡发育的影响规律,则考虑联接方法的耦合模型优势显著.

(3)在本文数据驱动模型中,RF模型预测滑坡易发性的不确定性最小,其次分别为SVM、BPNN和LR模型.LR预测精度依赖于联接方法的好坏,其与WOE耦合时能达到和SVM和BPNN一样的效果.可见先进的机器学习可有效降低滑坡易发性预测的不确定性.

(4)在联接方法和数据驱动模型耦合工况下,基于WOE的数据驱动模型预测精度更为优秀,均值和平均秩较小且标准差较大;PS-LR模型预测精度最低,均值和平均秩较大且标准差较小;其余耦合模型的易发性预测性能介于两者之间.

References

Chang, Z. L., Du, Z., Zhang, F., et al., 2020a. Landslide Susceptibility Prediction Based on Remote Sensing Images and GIS: Comparisons of Supervised and Unsupervised Machine Learning Models. *Remote Sensing*, 12(3): 502. https://doi.org/10.3390/rs12030502

Chang, Z. L., Gao, H. X., Huang, F. M., et al., 2020b. Study on the Creep Behaviours and the Improved Burgers Model of a Loess Landslide Considering Matric Suction. *Natural Hazards*, 103(1): 1479−1497. https://doi.org/10.1007/s11069-020-04046-0

Chen, W., Li, W. P., Hou, E. K., et al., 2015. Application of Frequency Ratio, Statistical Index, and Index of Entropy Models and Their Comparison in Landslide Susceptibility Mapping for the Baozhong Region of Baoji, China. *Arabian Journal of Geosciences*, 8(4): 1829−1841. https://doi.org/10.1007/s12517-014-1554-0

Chen, W., Xie, X. S., Peng, J. B., et al., 2018. GIS-Based Landslide Susceptibility Evaluation Using a Novel Hybrid Integration Approach of Bivariate Statistical Based Random Forest Method. *CATENA*, 164: 135−149. https://doi.org/10.1016/j.catena.2018.01.012

Chen, W., Xie, X. S., Wang, J. L., et al., 2017. A Comparative Study of Logistic Model Tree, Random Forest, and Classification and Regression Tree Models for Spatial Prediction of Landslide Susceptibility. *CATENA*, 151: 147−160. https://doi.org/10.1016/j.catena.2016.11.032

Devkota, K. C., Regmi, A. D., Pourghasemi, H. R., et al.,

2013. Landslide Susceptibility Mapping Using Certainty Factor, Index of Entropy and Logistic Regression Models in GIS and Their Comparison at Mugling-Narayanghat Road Section in Nepal Himalaya. *Natural Hazards*, 65(1): 135－165. https://doi.org/10.1007/s11069-012-0347-6

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

Guo, Z.Z., Yin, K.L., Fu, S., et al., 2019. Evaluation of Landslide Susceptibility Based on GIS and WOE-BP Model. *Earth Science*, 44(12): 4299－4312(in Chinese with English abstract).

Guo, Z. Z., Yin, K. L., Gui, L., et al., 2019. Regional Rainfall Warning System for Landslides with Creep Deformation in Three Gorges Using a Statistical Black Box Model. *Scientific Reports*, 9: 8962. https://doi.org/10.1038/s41598-019-45403-9

Hong, H. Y., Chen, W., Xu, C., et al., 2017. Rainfall-Induced Landslide Susceptibility Assessment at the Chongren Area (China) Using Frequency Ratio, Certainty Factor, and Index of Entropy. *Geocarto International*, 32(2): 139－154. https://doi.org/10.1080/10106049.2015.1130086

Huang, F. M., Cao, Z. S., Guo, J. F., et al., 2020a. Comparisons of Heuristic, General Statistical and Machine Learning Models for Landslide Susceptibility Prediction and Mapping. *CATENA*, 191: 104580. https://doi.org/10.1016/j.catena.2020.104580

Huang, F. M., Cao, Z. S., Jiang, S. H., et al., 2020b. Landslide Susceptibility Prediction Based on a Semi-Supervised Multiple-Layer Perceptron Model. *Landslides*, 17(12): 2919－2930. https://doi.org/10.1007/s10346-020-01473-9

Huang, F. M., Chen, J. W., Du, Z., et al., 2020c. Landslide Susceptibility Prediction Considering Regional Soil Erosion Based on Machine-Learning Models. *ISPRS International Journal of Geo-Information*, 9(6): 377. https://doi.org/10.3390/ijgi9060377

Huang, F.M., Wang, Y., Dong, Z.L., et al., 2019. Regional Landslide Susceptibility Mapping Based on Grey Relational Degree Model. *Earth Science*, 44(2): 664－676(in Chinese with English abstract).

Huang, F. M., Ye, Z., Jiang, S. H., et al., 2021. Uncertainty Study of Landslide Susceptibility Prediction Considering the Different Attribute Interval Numbers of Environmental Factors and Different Data-Based Models. *CATENA*, 202: 105250. https://doi.org/10.1016/j.catena.2021.105250

Huang, Y., Zhao, L., 2018. Review on Landslide Susceptibility Mapping Using Support Vector Machines. *CATENA*, 165: 520－529. https://doi.org/10.1016/j.catena.2018.03.003

Jacobs, L., Kervyn, M., Reichenbach, P., et al., 2020. Regional Susceptibility Assessments with Heterogeneous Landslide Information: Slope Unit- vs. Pixel-Based Approach. *Geomorphology*, 356: 107084. https://doi.org/10.1016/j.geomorph.2020.107084

Li, W. B., Fan, X. M., Huang, F. M., et al., 2020. Uncertainties Analysis of Collapse Susceptibility Prediction Based on Remote Sensing and GIS: Influences of Different Data-Based Models and Connections between Collapses and Environmental Factors. *Remote Sensing*, 12(24): 4134. https://doi.org/10.3390/rs12244134

Li, Y., Huang, J., Jiang, S. H., et al., 2017. A Web-Based GPS System for Displacement Monitoring and Failure Mechanism Analysis of Reservoir Landslide. *Scientific Reports*, 7(1): 17171. https://doi.org/10.1038/s41598-017-17507-7

Li, Y.L., Zhang, Q., Li, M., et al., 2015. Using BP Neural Networks for Water Level Simulation in Poyang Lake. *Resources and Environment in the Yangtze Basin*, 24(2): 233－240(in Chinese with English abstract).

Liu, W. P., Luo, X. Y., Huang, F. M., et al., 2019. Prediction of Soil Water Retention Curve Using Bayesian Updating from Limited Measurement Data. *Applied Mathematical Modelling*, 76: 380－395. https://doi.org/10.1016/j.apm.2019.06.028

Ma, S.Y., Xu, C., Tian, Y.Y., et al., 2019. Application of Logistic Regression Model for Hazard Assessment of Earthquake-Triggered Landslides: A Case Study of 2017 Jiuzhaigou (China) MS7.0 Event. *Seismology and Geology*, 41(1): 162－177 (in Chinese with English abstract).

Merghadi, A., Yunus, A. P., Dou, J., et al., 2020. Machine Learning Methods for Landslide Susceptibility Studies: A Comparative Overview of Algorithm Performance. *Earth-Science Reviews*, 207: 103225. https://doi.org/10.1016/j.earscirev.2020.103225

Pham, B. T., Tien Bui, D., Prakash, I., et al., 2017. Hybrid Integration of Multilayer Perceptron Neural Networks and Machine Learning Ensembles for Landslide Susceptibility Assessment at Himalayan Area (India) Using GIS. *CATENA*, 149: 52－63. https://doi.org/

10.1016/j.catena.2016.09.007

Qiu, H.J., Cao, M.M., Liu, W., et al., 2014. The Susceptibility Assessment of Landslide and Its Calibration of the Models Based on Three Different Models. *Scientia Geographica Sinica*, 34(1): 110－115(in Chinese with English abstract).

Qiu, H.J., Ma, S.Y., Cui, Y.F., et al., 2020. Reconsider the Role of Landslides. *Journal of Northwest University (Natural Science Edition)*, 50(3): 377－385(in Chinese with English abstract).

Regmi, A. D., Devkota, K. C., Yoshida, K., et al., 2014. Application of Frequency Ratio, Statistical Index, and Weights-of-Evidence Models and Their Comparison in Landslide Susceptibility Mapping in Central Nepal Himalaya. *Arabian Journal of Geosciences*, 7(2): 725－742. https://doi.org/10.1007/s12517-012-0807-z

Saha, S., Saha, M., Mukherjee, K., et al., 2020. Predicting the Deforestation Probability Using the Binary Logistic Regression, Random Forest, Ensemble Rotational Forest, REPTree: A Case Study at the Gumani River Basin, India. *Science of the Total Environment*, 730: 139197. https://doi. org/10.1016/j. scitotenv.2020.139197

Sun, D. L., Wen, H. J., Wang, D. Z., et al., 2020. A Random Forest Model of Landslide Susceptibility Mapping Based on Hyperparameter Optimization Using Bayes Algorithm. *Geomorphology*, 362: 107201. https://doi.org/10.1016/j.geomorph.2020.107201

Wang, P., Bai, X. Y., Wu, X. Q., et al., 2018. GIS-Based Random Forest Weight for Rainfall-Induced Landslide Susceptibility Assessment at a Humid Region in Southern China. *Water*, 10(8): 1019. https://doi. org/10.3390/w10081019

Wang, Z.W., Wang, L., Huang, G.W., et al., 2020. Research on Multi-Source Heterogeneous Data Fusion Algorithm of Landslide Monitoring Based on BP Neural Network. *Journal of Geomechanics*, 26(4): 575－582(in Chinese with English abstract).

Wu, R.Z., Hu, X.D., Mei, H.B., et al., 2021. Spatial Susceptibility Assessment of Landslides Based on Random Forest: A Case Study from Hubei Section in the Three Gorges Reservoir Area. *Earth Science*, 46(1): 321－330 (in Chinese with English abstract).

Wu, Y.P., Zhang, Q.X., Tang, H.M., et al., 2014. Landslide Hazard Warning Based on Effective Rainfall Intensity. *Earth Science*, 39(7): 889－895(in Chinese with English abstract).

Xu, C., Dai, F. C., Xu, X. W., et al., 2012. GIS-Based Support Vector Machine Modeling of Earthquake-Triggered Landslide Susceptibility in the Jianjiang River Watershed, China. *Geomorphology*, 145/146: 70－80. https://doi.org/10.1016/j.geomorph.2011.12.040

Xu, Q., Dong, X.J., Li, W.L., 2019. Integrated Space-Air-Ground Early Detection, Monitoring and Warning System for Potential Catastrophic Geohazards. *Geomatics and Information Science of Wuhan University*, 44 (7): 957－966(in Chinese with English abstract).

Xu, S.H., Liu, J.P., Wang, X.H., et al., 2020. Landslide Susceptibility Assessment Method Incorporating Index of Entropy Based on Support Vector Machine: A Case Study of Shaanxi Province. *Geomatics and Information Science of Wuhan University*, 45(8): 1214－1222(in Chinese with English abstract).

Yu, X.Y., Hu, Y.J., Niu, R.Q., 2016. Research on the Method to Select Landslide Susceptibility Evaluation Factors Based on RS-SVM Model. *Geography and Geo-Information Science*, 32(3): 23－28, 2(in Chinese with English abstract).

Zhang, J., Yin, K.L., Wang, J.J., et al., 2016. Evaluation of Landslide Susceptibility for Wanzhou District of Three Gorges Reservoir. *Chinese Journal of Rock Mechanics and Engineering*, 35(2): 284－296(in Chinese with English abstract).

Zhang, Q.K., Ling, S.X., Li, X.N., et al., 2020. Comparison of Landslide Susceptibility Mapping Rapid Assessment Models in Jiuzhaigou County, Sichuan Province, China. *Chinese Journal of Rock Mechanics and Engineering*, 39(8): 1595－1610(in Chinese with English abstract).

Zhang, S.H., Wu, G., 2019. Debris Flow Susceptibility and Its Reliability Based on Random Forest and GIS. *Earth Science*, 44(9): 3115－3134(in Chinese with English abstract).

Zhu, A.X., Lü, G.N., Zhou, C.H., et al., 2020. Geographic Similarity: Third Law of Geography? *Journal of Geo-Information Science*, 22(4): 673－679(in Chinese with English abstract).

Zhu, L., Huang, L. H., Fan, L. Y., et al., 2020. Landslide Susceptibility Prediction Modeling Based on Remote Sensing and a Novel Deep Learning Algorithm of a Cascade-Parallel Recurrent Neural Network. *Sensors*, 20(6): 1576. https://doi.org/10.3390/s20061576

Zhu, L., Wang, G. J., Huang, F. M., et al., 2021. Landslide Susceptibility Prediction Using Sparse Feature Extraction and Machine Learning Models Based on GIS and Remote Sensing. *IEEE Geoscience and Remote*

Sensing Letters, 1－5. https://doi. org/10.1109/LGRS.2021.3054029

附中文参考文献

冯杭建, 周爱国, 俞剑君, 等, 2016. 浙西梅雨滑坡易发性评价模型对比. 地球科学, 41(3): 403－415.

郭子正, 殷坤龙, 付圣, 等, 2019. 基于GIS与WOE-BP模型的滑坡易发性评价. 地球科学, 44(12): 4299－4312.

黄发明, 汪洋, 董志良, 等, 2019. 基于灰色关联度模型的区域滑坡敏感性评价. 地球科学, 44(2): 664－676.

李云良, 张奇, 李淼, 等, 2015. 基于BP神经网络的鄱阳湖水位模拟. 长江流域资源与环境, 24(2): 233－240.

马思远, 许冲, 田颖颖, 等, 2019. 基于逻辑回归模型的九寨沟地震滑坡危险性评估. 地震地质, 41(1): 162－177.

邱海军, 曹明明, 刘闻, 等, 2014. 基于三种不同模型的区域滑坡灾害敏感性评价及结果检验研究. 地理科学, 34(1): 110－115.

邱海军, 马舒悦, 崔一飞, 等, 2020. 重新认识滑坡作用. 西北大学学报(自然科学版), 50(3): 377－385.

王智伟, 王利, 黄观文, 等, 2020. 基于BP神经网络的滑坡监测多源异构数据融合算法研究. 地质力学学报, 26(4): 575－582.

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

吴益平, 张秋霞, 唐辉明, 等, 2014. 基于有效降雨强度的滑坡灾害危险性预警. 地球科学, 39(7): 889－895.

许强, 董秀军, 李为乐, 2019. 基于天－空－地一体化的重大地质灾害隐患早期识别与监测预警. 武汉大学学报·信息科学版, 44(7): 957－966.

徐胜华, 刘纪平, 王想红, 等, 2020. 熵指数融入支持向量机的滑坡灾害易发性评价方法: 以陕西省为例. 武汉大学学报·信息科学版, 45(8): 1214－1222.

于宪煜, 胡友健, 牛瑞卿, 2016. 基于RS-SVM模型的滑坡易发性评价因子选择方法研究. 地理与地理信息科学, 32(3): 23－28, 2.

张俊, 殷坤龙, 王佳佳, 等, 2016. 三峡库区万州区滑坡灾害易发性评价研究. 岩石力学与工程学报, 35(2): 284－296.

张玘恺, 凌斯祥, 李晓宁, 等, 2020. 九寨沟县滑坡灾害易发性快速评估模型对比研究. 岩石力学与工程学报, 39(8): 1595－1610.

张书豪, 吴光, 2019. 随机森林与GIS的泥石流易发性及可靠性. 地球科学, 44(9): 3115－3134.

朱阿兴, 闾国年, 周成虎, 等, 2020. 地理相似性:地理学的第三定律?. 地球信息科学学报, 22(4): 673－679.