DOI:10.13870/j.cnki.stbcxb.2024.03.008
李潼亮,赵梓鉴,李斌斌,等. 基于梯度提升树模型的坡耕地土壤水蚀模拟与分析[J].水土保持学报,2024,38(3):54-63.
LI Tongliang, ZHAO Zijian, LI Binbin, et al. Simulation and analysis of hydraulic erosion in sloping farmland based on gradient Boosting decition tree mode[J]. Journal of Soil and Water Conservation,2024,38(3):54-63.

基于梯度提升树模型的坡耕地土壤水蚀模拟与分析

李潼亮[1],赵梓鉴[1],李斌斌[2],张风宝[1,3],郭 正[1],何琪琳[3,4],何 庆[1],杨明义[1,3]
(1.西北农林科技大学水土保持科学与工程学院黄土高原土壤侵蚀与旱地农业国家重点实验室,712100,陕西 杨凌;
2.水利部水土保持监测中心,北京 100053;3.中国科学院水利部水土保持研究所,陕西 杨凌 712100;
4.中国科学院大学,北京 100049)

摘 要:[**目的**]针对黄土高原坡耕地土壤侵蚀过程复杂、人为干扰强烈且难以量化的特点,利用机器学习定量解析主要影响因素对坡耕地土壤水蚀的作用与贡献,模拟分析坡耕地土壤水蚀特征并探究其机理,为坡耕地土壤侵蚀的预报提供基础支撑。[**方法**]基于黄土高原子洲试验站坡耕地小区 1959—1969 年产流产沙观测数据,精细化表征其影响因子,运用梯度提升树模型对侵蚀量和径流深的变化及其影响因素的贡献进行分析。[**结果**]数据集中次降雨侵蚀量(0~122.72 t/km^2)、径流深(0.02~17.20 mm)、降雨历时(2~1 410 min)及平均雨强(0.02~4.63 mm)属强变异,变异系数均>1,且多数变量呈右偏态;在相同训练集和测试集划分情况下,对侵蚀量模型预测精度(R^2=0.81)略优于径流深模型(R^2=0.80),但侵蚀量模型的层数(8 层)大于径流深模型(5 层),表明侵蚀机理相较径流机理更为复杂;通过梯度提升树模型与 SHAP 算法对自变量重要性进行排序发现,影响侵蚀模型与径流模型的自变量重要性不同。[**结论**]受特征提取的限制,在侵蚀量与径流深较小时预测结果不理想,未来研究应当通过引入更多自变量组合方式寻找更多相关变量以提高对侵蚀事件的预测。产流和产沙的主要影响因素存在差异,降水本身特征对产流过程起主要作用,侵蚀产沙过程中主要受到降水与地形相关自变量的共同影响。基于数据驱动,为揭示黄土高原坡耕地侵蚀机理提供参考,并为区域坡耕地土壤侵蚀防治提供科学依据。

关键词: 预报模型; 梯度提升树模型; 坡耕地; 黄土坡面
中图分类号:S157.1 **文献标识码:**A **文章编号:**1009-2242-(2024)03-0054-10

Simulation and Analysis of Hydraulic Erosion in Sloping Farmland Based on Gradient Boosting Decition Tree Mode

LI Tongliang[1], ZHAO Zijian[1], LI Binbin[2], ZHANG Fengbao[1,3], GUO Zheng[1], HE Qilin[3,4], HE Qing[1], YANG Mingyi[1,3]
(1.*State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, College of Water and Soil Conservation Science and Engineering, Northwest A&F University, Yangling, Shaanxi* 712100, *China*; 2. *Water and Soil Conservation Monitoring Center, Ministry of Water Resources, Beijing* 100053, *China*; 3. *Institute of Water and Soil Conservation, Ministry of Water Resources, Chinese Academy of Sciences, Yangling, Shaanxi* 712100, *China*;
4. *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: [**Objective**] This article employs machine learning to quantitatively analyze soil water erosion in Loess Plateau slope farmland, addressing its complexity and quantification challenges due to human interference. We aim to simulate erosion characteristics, explore its mechanisms, and support erosion prediction. [**Methods**] Using 1959 - 1969 data from Zizhou Experimental Station, we characterized the influencing factors and analyzed erosion and runoff depth changes with a gradient boosting decision tree. [**Results**] The dataset showed significant variability in secondary rainfall erosion (0~122.72 t/km^2), runoff

收稿日期:2023-09-05 **修回日期:**2023-10-18 **录用日期:**2023-12-10 **网络首发日期(**www.cnki.net**):**2024-04-10
资助项目:国家自然科学基金项目(42077071,42177338,41830758);国家重点研发计划项目(2022YFF1300805);中央高校基本科研业务费专项资金项目(2023HHZX001)
第一作者:李潼亮(1999—),男,硕士研究生,主要从事土壤侵蚀预报与机理研究。E-mail:tlleeeee99@foxmail.com
通信作者:张风宝(1980—),男,博士,教授,博士生导师,主要从事坡面土壤侵蚀过程及其环境效应研究。E-mail:fbzhang@nwsuaf.edu.cn

depth (0.02～17.20 mm), rainfall duration (2～1 410 min), and average intensity (0.02～4.63 mm), often right-skewed. The erosion model (R^2=0.81) slightly outperformed the runoff depth model (R^2=0.80), despite its greater complexity (8 layers vs. 5). Using the gradient boosting tree model and SHAP algorithm, we found differing key factors for erosion and runoff. [**Conclusion**] Limitations in feature extraction lead to less accurate predictions for small erosion and runoff depths. Future research should explore more independent variable combinations to enhance predictions. Main influencing factors differ between runoff and sediment production. Precipitation mainly influences runoff, while erosion and sediment production depend on precipitation and terrain-related variables. In summary, this data-driven study illuminates slope farmland erosion mechanisms on the Loess Plateau, providing a scientific basis for erosion control in the region.

Keywords: prediction model; gradient lifting tree model; sloping farmland; loess slope surface

Received:2023-09-05　　**Revised**:2023-10-18　　**Accepted**:2023-12-10　　**Online**(www.cnki.net):2024-04-10

土壤侵蚀作为全球性重大环境问题之一,与气象、水文、地形地貌、土壤类型、人类活动等诸多因素均有极其复杂的关系,严重威胁人类生存与社会的可持续性发展[1]。坡耕地作为水土流失的重要策源地,具有影响因子复杂、人为干预强烈且难以定量表达等特点,对其开展系统研究对于深入了解土壤侵蚀规律及构建土壤侵蚀模型具有重要意义。土壤侵蚀模型作为土壤侵蚀过程定量化研究的有效手段[2],备受广大研究者关注。针对坡耕地土壤侵蚀,大量研究[3-4]基于监测资料,构建经验模型或基于过程的物理模型,但现有模型存在物理模型预测精度有限,经验模型可解释性不强的缺点[5],虽然在建模时考虑相关因子,但在因子贡献量分析时,又因复杂性而不得不忽略此类交互变量影响的不足[6]。因此,高效、准确、考虑全面的坡耕地土壤侵蚀模型的构建是研究的难点,也是热点。

随着计算机技术的发展,机器学习以其独特的优势在各个领域发挥重要作用,国内外研究者[7-19]纷纷将机器学习方法引入到土壤侵蚀模型构建及预测预报中。与传统模型相比,机器学习模型具有更好的持续性与预测性,在建模过程中可引入更多影响因素并探究其之间存在的非线性关系。国内机器学习在土壤侵蚀方面的应用最早可以追溯到 1997 年,洪伟等[7]基于径流试验场观测数据,首次提出应用人工神经网络方法对土壤侵蚀实现预测预报并发现,其预测精度高于通用土壤流失方程(USLE)。随后诸多研究者开始利用 BP 神经网络,对诸如坡面产沙量[8]、坡面入渗[9]等侵蚀相关变量进行建模预报,段军彪等[10]尝试使用遗传算法对 BP 神经网络进行优化,并基于杨家沟小流域多年观测数据进行模型训练与模拟发现,改进模型取得较高的预测精度与较快的收敛速度;黄俊等[11]针对广东红壤可蚀性构建 BP 神经网络预测模型与逐步回归分析模型发现,BP 神经网络预测精度更高。随后,学者们又尝试使用偏最小二乘法[12]、人工蜂群算法[13]等方法对模型继续改进优化。胡亚萍等[14]在利用 BP 神经网络对小流域次降雨进行侵蚀产沙模拟外,还使用支持回归向量机模型进行模拟表明,支持回归向量机模型结果相较 BP 神经网络预测精度更高,稳定性更强。在国际上,DE 等[15]建立决策树与神经网络的混合模型,用于预测西班牙南部安达卢西亚地区的侵蚀事件,模型表现良好;LICZNAR 等[16]利用神经网络与径流小区数据建立侵蚀敏感性地图发现,神经网络比 WEPP 预测更加准确。近年来,利用机器学习方法结合地理信息系统,在较大尺度上分析侵蚀敏感性是国际研究热点,ARABAMERI 等[17-19]使用元自适应回归样条(MARS)、逻辑回归(LR)和证据信仰函数(EBF)、神经网络(ANN)等多种机器学习模型对侵蚀敏感性进行判别,并尝试对土壤沟蚀敏感性影响因素进行探究。

目前,在土壤侵蚀研究中,主要应用机器学习模型是神经网络模型或以神经网络为基础的复合模型,以此对土壤侵蚀量、土壤侵蚀敏感性进行预测[20],其特点是预测精准度和准确度高[13]。复合模型预测原理为模拟神经元网络,结构和参数依赖于大量测试,费时费力,容易产生局部最小[14]及模型结果解释性不强[21]等问题。梯度提升树模型(GBDT)是以树模型为基础的集成模型,具有较强的泛化能力[22],适用于因变量与自变量间存在非线性关系的情况,并能防止自变量之间的多重共线性[23]。具有预测精准、运算结果可解释性强的特点。相较神经网络模型,梯度提升树模型构建过程更清晰,能直观探究各个自变量因子对土壤侵蚀结果影响的大小,更适合运用于土壤侵蚀复杂过程的预测与驱动因子影响力分析。同时,曹玉茹等[24]研究指出,梯度提升树模型与 SHAP 算法(shapley additive explanations)结合,对于特征重

要性的解释具有极大的提升。基于此，通过使用机器学习算法中的梯度提升树模型，以黄土高原坡耕地小区水沙资料为基础，对次降雨侵蚀量、径流深与其影响因素间的关系进行建模分析，利用梯度提升树模型的特征重要性排序功能结合 SHAP 算法，对坡耕地小区产流产沙驱动因子影响程度进行定量分析，为全面定量化探究坡耕地侵蚀过程、完善坡面侵蚀机制提供科学依据与技术支撑。

1 材料与方法

1.1 研究区概况

研究区位于陕北子洲县，隶属于陕西省榆林市（37°15′—38°50′ E，109°29′—110°07′ N，图 1）。子洲县地跨暖温带与中温带。境内沟壑纵横，梁峁起伏，地势西高东低。黄河流域子洲径流试验站水文资料数据集[25]（1959—1969 年）中指出，该区域侵蚀严重，1959 年最大实测含沙量为 1 050 kg/m^3，1954—1958 年，年平均侵蚀模数为 15 780 t/km^2，最大为 23 670 t/km^2。子洲县海拔最高 1 357 m，最低 870 m。境内 95%为山区，5%为川区。日照充足，光能较强。年平均气温 8 ℃，最低温度－27 ℃，无霜期 145 天，降水变率较大，旱涝频繁，降水量 428.1 mm，主要集中在 6—9 月[26]。春夏季多行偏南风，秋冬季春盛行偏北风，为大陆性干旱半干旱气候。

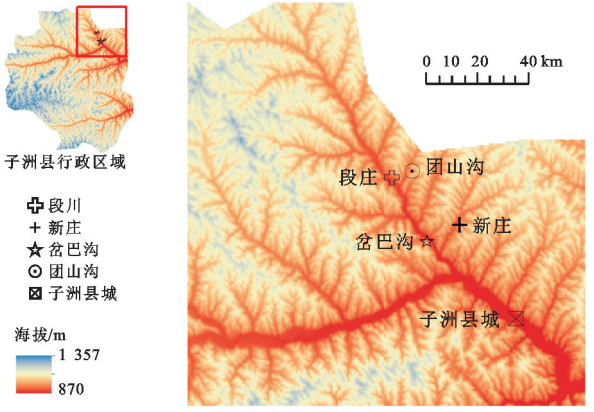


图 1 研究区域

Fig. 1 The study area

1.2 数据来源

数据来源于国家科技基础条件平台—国家地球系统科学数据中心—黄土高原分中心，使用黄河流域子洲径流试验站（以下简称“子洲站”）水文试验资料数据集（1959—1969 年）。1959—1969 年，子洲站在子洲县岔巴沟、段川、团山沟等区域设立多个不同条件的径流场观测并记录自然降雨条件下的侵蚀情况。研究数据包括 1959—1969 年次降雨 245 场，监测指标包括次降雨量、降雨历时、最大雨强、侵蚀量、径流深等数据，以上数据均为次降雨坡耕地径流场数据。径流场具体情况见表 1。

1.3 因子选择与预处理

坡耕地水力侵蚀驱动因子众多，只有初步遴选出合适的自变量才有可能准确有效地预测侵蚀过程并分析自变量对侵蚀产沙作用及贡献。考虑以往研究[27]成果与数据集的基本情况，以次降雨量、最大雨强、降雨历时、平均雨强、坡度、坡长、径流场面积、种植作物种类、距离锄草天数、种植密度、作物生长天数等 11 个指标作为坡耕地土壤水蚀的自变量。对于作物类型变量，使用独热编码对其进行编码。独热编码（one－hot encoding）可将类别形式数据转换为机器学习模型使用的形式，能够扩充数据特征维度。使用独热编码将作物种类转化为二进制输入形式，处理后作物种类由 1 维扩充为 7 维，并作为 0，1 矩阵输入模型，具体输入变量处见表 2。

1.4 研究方法

1.4.1 梯度提升决策树模型与 SHAP 算法 以次降雨量、最大雨强、降雨历时、平均雨强、坡度、坡长、径流场面积、种植作物种类、距离除草天数、种植密度、作物生长天数等 11 个指标为梯度提升树模型的输入项，以侵蚀量、径流深作为输出项，建立模型；并通过梯度提升树模型提供的特征重要性确定各驱动因素对产流产沙的相对贡献。超参数是在训练机器学习模型时需要手动设定的参数，它们的大小直接影响着模型的性能。不同的超参数设定可导致同一个模型展现出不同的预测性能。对于梯度提升树模型而言，有 4 个重要的超参数需要设定，包括弱分类器的数量、学习率、树的最大深度以及分裂节点所需的最小样本数量。当弱分类器的数量越多、学习率越大、树的深度越深及分裂节点所需的数据越少时，模型就能更好地吸收训练集中的细节信息，反之亦然。根据过去的研究[28]经验和反复测试优化，设定弱分类器的数量为 1 500 个；学习率为 0.008（用于侵蚀量模型）和 0.01（用于径流深模型）；树的最大深度分别为 7（侵蚀量模型）和 4（径流深模型）；分裂节点所需的数据分别为 4 个（侵蚀量模型）和 6 个（径流深模型）。

SHAP 算法是被机器学习广泛使用的一种基于博弈论理论的解释性算法[29]，其原理是根据博弈论计算每个变量的预期边际贡献（Shapley 值），通过贡献值可直观表示在模型预测过程中各个变量所产生影响的方式，包括该自变量对模型的正负影响。预期边际贡献计算公式为[30]：

$$\varphi_i(v)=\frac{1}{|K|!}\sum_R[v(S_I^R\cup\{i\}-v(S_i^R)] \quad (1)$$

式中：φ_i 为 Shapely 值；S_I^R 为一组具有顺序的变量；K 为特征的数量；$v(S_i^R)$ 为特征在预测目标特征中的贡献。

表 1　径流场基本情况

Table 1　List of basic conditions of radial flow field

年份/a	径流场号	河流名称	位置	土壤情况	坡向	坡度/‰	尺寸		附注
							长/m	面积/m^2	
1959	岔巴沟 1 号	岔巴沟	中游右岸	黄土	西南	467.0	20.00	200	
1959	岔巴沟 2 号	岔巴沟	中游右岸	黄土	西南	467.0	20.00	200	
1959	岔巴沟 3 号	岔巴沟	中游分水岭	黄土	西北	140.0	20.00	200	
1959	岔巴沟 4 号	岔巴沟	中游左岸	黄土	东北	577.0	40.00	400	
1959	岔巴沟 5 号	岔巴沟	中游左岸	黄土	东北	577.0	20.00	200	
1959	岔巴沟 6 号	岔巴沟	中游左岸	黄土	东北	577.0	30.00	300	
1959	岔巴沟 7 号	岔巴沟	中游左岸	黄土	东北	577.0	30.00	300	
1960—1962	新庄 1 号	新庄沟	右岸山坡	黄土	西南	466.0	18.60	186	
1960—1961	新庄 2 号	新庄沟	右岸山坡	黄土	西南	466.0	18.70	187	
1960—1962	新庄 3 号	新庄沟	糜山渠与新庄沟分水岭	黄土	西北	141.0	20.02	202	
1960—1962	新庄 4 号	新庄沟	左岸山坡	黄土	东北	577.0	34.60	346	
1960—1961	新庄 5 号	新庄沟	左岸山坡	黄土	东北	577.0	17.30	173	
1960—1961	新庄 6 号	新庄沟	左岸山坡	黄土	东北	577.0	26.00	520	
1961	新庄 7 号	新庄沟	左岸山坡	黄土	东北	577.0	43.30	866	
1960—1964，1967	团山沟 1 号	蛇家沟	左岸卯坡	黄土	西北	158.0	20.00	150	1960—1962 年坡度为 162‰，后续降为 158‰
1960—1964，1967	团山沟 2 号	蛇家沟	左岸卯坡	黄土	东北	404.0	40.00	600	
1960—1964，1967—1969	团山沟 3 号	蛇家沟	左岸卯坡	黄土	东北	404.0	60.00	900	
1960—1964，1967	团山沟 4 号	蛇家沟	左岸卯坡	黄土	东北	404.0	20.00	300	
1960—1964，1967	团山沟 5 号	蛇家沟	左岸卯坡	黄土	东北	603.0	20.00	300	1960—1962 年坡度曾为 603‰，后续降为 601‰
1960—1964，1967	团山沟 6 号	蛇家沟	左岸沟坡	黄土	东北	827.0	20.00	1 160	
1960—1964，1967—1969	团山沟 7 号	蛇家沟	左岸沟坡	黄土	西北	1 730.0	100.00	4 084	
1964，1967—1969	团山沟 9 号	蛇家沟	右岸沟掌浅凹地	黄绵土	西南	8.0	20.00	17 200	
1967	团山沟 10 号	蛇家沟	右岸峁坡	黄绵土	东南	62.5	30.00	300	
1967	团山沟 11 号	蛇家沟	右岸峁坡	黄绵土	东南	625.0	15.00	150	
1967	段川 1 号	岔巴沟	水旺沟右岸无名沟左岸 16°阴坡上	黄沙壤土	西北	287.0	7.57	30.3	
1967	段川 2 号	岔巴沟	水旺沟右岸无名沟左岸 16°阴坡上	黄沙壤土	西北	287.0	20.08	200	

注:数据集中部分径流小区坡度与坡长数据来源于数据集文本描述[25]。

表 2 输入自变量编码

Table 2 Input independent variable encoding table

序号	特征名称	编码
1	次降雨量	X(0)
2	最大雨强	X(1)
3	降雨历时	X(2)
4	平均雨强	X(3)
5	坡度	X(4)
6	坡长	X(5)
7	径流场面积	X(6)
8	距离锄草天数	X(7)
9	种植密度	X(8)
10	作物生长天数	X(9)
11	作物种类—种植麦子	X(10－1)
12	作物种类—种植洋芋	X(10－2)
13	作物种类—种植谷子间绿豆	X(10－3)
14	作物种类—种植谷子间绿豆	X(10－4)
15	作物种类—种植糜子间绿豆	X(10－5)

1.4.2 产流产沙特征分析与模拟结果检验 对数据集基本统计特征进行描述，以便整体上了解数据的基本情况和后续构建模型，基本统计特征包括平均值、标准差、总和、偏度、峰度、变异系数、中位数与大小极值。

在模型训练前，将输入数据按照随机原则划分为互斥的 2 个子集，245 场降雨中的 94 场作为测试集，151 场作为训练集。使用训练集对模型进行训练，基于所得模型对测试集侵蚀量、径流深分别进行预测。采用决定系数(R^2)、纳什系数(NSE)、赤池信息量准则(AIC)和均方根误差(RMSE)为评价模型的统计量。R^2 与 NSE 值越接近 1，表明拟合效果越好。RMSE 反映预测值与实际值之间的差异程度，RMSE 越小差异越小。AIC 是衡量统计模型拟合结果，权衡模型复杂程度的标准，AIC 越小越好。具体计算方式为：

$$R^2=\frac{\left[\sum_{i=1}^{n}(y_i-\bar{y}_i)(\hat{y}_{i-}\dot{y}_i)\right]^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2\sum_{i=1}^{n}((\hat{y}_{i-}\bar{y}i)^2} \tag{2}$$

$$\mathrm{NSE}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \tag{3}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{4}$$

$$\mathrm{AIC}=2k-2\ln(L) \tag{5}$$

式中：n 为样本数量；y_i 为实测值；$\hat{y}_i$ 为预测值；$\bar{y}_i$ 为样本均值；$\bar{y}_i$ 为预测均值；k 为参数参量；L 为似然函数。

使用 Python 平台实现梯度提升树模型建模与 SHAP 算法的构建，数据的统计、整理、不同输入因子的相关性分析及作图在 Origin 2021 软件中完成。

2 结果与分析

2.1 降雨、产流产沙基本统计特征分析

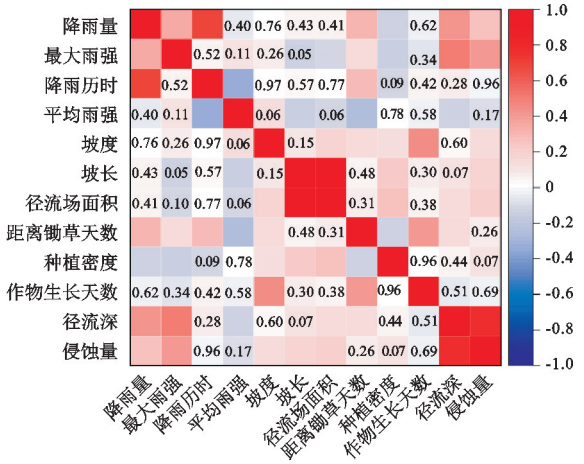
次降雨数据、径流深与侵蚀量描述性统计分析见表 3。不同场次降雨的降雨量、最大雨强、降雨历时、平均雨强、距离耕作时间等变量变异性较大，变异系数为 0.23～1.92，基本均处于中等变异(0.1～1)和强变异(>1)，尤其是降雨历时与平均雨强，变异系数分别为 1.92，1.21。径流深与侵蚀量属强变异，径流深为 0.9～17.2 mm，平均值为(2.32±3.23) mm，变异系数 1.39，偏度系数 2.28，属于右偏态，说明径流深小于平均值的样本占多数；侵蚀量为 0～122.72 t/km^{2}时，平均值为(7.78±17.58) t/km^2，变异系数达2.25，偏度系数为 3.84，属于右偏态，说明侵蚀量小于平均值的样本占多数。

表 3 数值型变量描述统计

Table 3 Statistical table for numerical variable description

变量	单位	总数(N)	平均值	标准差	总和	偏度	峰度	变异系数	最小值	中位数	最大值
降雨量	mm	245	14.77	12.22	3 619.30	2.01	5.35	0.83	0.70	10.40	80.80
最大雨强	mm		0.78	0.70	191.06	2.10	5.70	0.89	0.06	0.55	4.63
降雨历时	min		100.69	193.58	24 668.60	3.78	18.18	1.92	2.00	28.00	1 410.00
平均雨强	mm		0.64	0.77	155.85	2.55	7.93	1.21	0.02	0.40	4.63
坡度	%		40.00	13.73	9 799.20	−0.10	−0.54	0.34	14.00	40.40	62.50
坡长	m		25.96	14.30	6 360.72	1.33	0.85	0.55	7.57	20.00	60.00
径流场面积	m^2		335.46	245.87	82 187.00	1.21	0.45	0.73	30.30	300.00	900.00
距离锄草天数	Days		18.41	15.00	4 511.00	1.66	2.50	0.81	0	14.00	64.00
种植密度	株/m^2		7.64	2.23	1 871.00	0.06	−1.64	0.29	4.00	6.00	11.00
作物生长天数	Days		107.56	24.71	26 353.00	0.59	−0.33	0.23	61.00	109.00	186.00
径流深	mm		2.32	3.23	568.46	2.28	5.41	1.39	0.02	0.90	17.20
侵蚀量	t/hm^2		7.80	17.58	1 910.42	3.84	17.06	2.25	0	0.98	122.72

因变量与自变量间进行相关性分析(图 2),降雨量、最大雨强、坡长、径流场面积、距离锄草时间与径流深均呈显著正相关,与平均雨强呈负相关。侵蚀量则与降雨量、最大雨强、坡度、坡长、径流场面积、种植密度和径流深呈显著正相关。同时,径流深与侵蚀量间呈显著正相关性。



注:图中颜色代表相关性,红色为正相关,蓝色为负相关,图中数字为 p 值。

图 2　数据集变量相关性分析

Fig. 2　Correlation analysis

2.2　产流产沙模拟分析

表 4 为次降雨径流深模型(简称径流深模型)和侵蚀量模型在预测最优情况下的模型结构。综合考虑节点数量、层数和分裂变量,径流深模型相较侵蚀量模型更为精简,侵蚀量模型使用的分裂变量较径流深模型更多。径流深模型使用降雨量、最大雨强、降雨历时、平均雨强和距离除草时间作为分裂变量,而侵蚀量模型还在径流深模型的基础上增加使用坡度、径流场面积和种植密度作为分裂变量。

表 4　模型结构描述统计

Table 4　Model structure description statistical table

模型名称	模型层数	有向边数	分裂变量
径流深模型	4	12	降雨量、最大雨强、降雨历时、平均雨强、距离除草时间
侵蚀量模型	7	21	降雨量、最大雨强、降雨历时、平均雨强、距离除草时间、坡度、径流场面积、种植密度

图 3 为径流深模型与侵蚀量模型在测试数据集上的评估结果。相较于侵蚀量模型,径流深模型预测结果略差,侵蚀量模型预测值与实际值更为接近,由 R^2 和 NSE 可得出,侵蚀量模型在测试集上预测效果更好。但 RMSE 与 AIC 结果显示,径流深模型泛化性更优,具有更好的鲁棒性。综上,本研究中,侵蚀量模型预测更为精准,而径流深模型在模型结构上更简洁,在应用新数据时适应性更强。

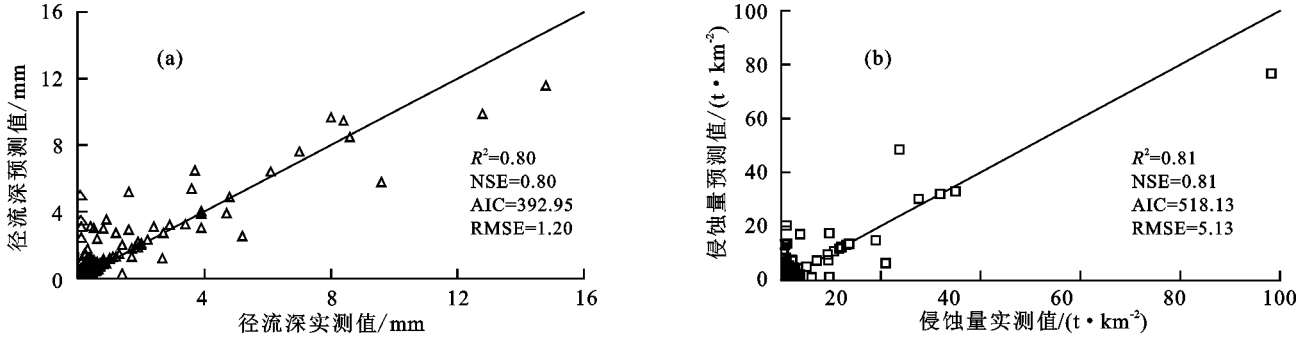


图 3　预测模型结果检验

Fig. 3　Prediction Model Result Verification

对训练所得模型在测试集上的预测结果进行误差分析(图 4)。径流深绝对误差为(−6.09～3.26) mm,平均值为(0.14±1.20) mm,相对误差为−8.84%～32.47%,平均值为 1.07%±4.07%;侵蚀量绝对误差为−23.58～25.07 t/km^2,平均值为(5.04±98.89) t/km^2,相对误差为 26.27%～1 655.12%,平均值为26.27%±176.89%。径流深的相对误差极值远小于侵蚀量。在实测值较低时,侵蚀量和径流深容易出现较大的相对误差,且预测值大多数高于实测值。

2.3　影响因子定量化解析

利用梯度提升树模型对自变量影响产流产沙的重要性进行排序(图 5),在径流深模型中,次降雨量、平均雨强和作物生长天数是主要的影响因素,其重要性比例分别为 0.31,0.24,0.11。而在侵蚀量模型中,主要影响因素为最大雨强、径流场面积和坡度,其重要性比例分别为 0.29,0.18,0.12。次降雨量、平均雨强、作物生长天数对径流深模型的影响高于对侵蚀量模型,而最大雨强、坡度、坡长、径流场面积对于侵蚀量模型的影响高于对径流深模型,说明影响产流和产沙的主要因子存在差异。

图 6 为基于 SHAP 算法分析自变量对产流产沙影响作用的大小。SHAP 越大,表示越有利于出现更高的预测结果,即随 SHAP 增大而增加的自变量特征在模型预测过程中与因变量呈正相关,其值的增加

导致侵蚀量/径流深的增加。依照 SHAP 大小自上而下进行排序，自变量越靠上，说明其对预测过程的影响越大。SHAP 排序结果与梯度提升树模型的排序结果有一定的差异，侵蚀量模型中的次降雨量、坡度、平均雨强等在 SHAP 算法中的排序更高，而最大雨强、径流场面积等因素排序更低。径流深模型中的最大雨强、种植密度在 SHAP 算法排序中更高，而作物生长时长、作物种类等因素排序更低。SHAP 算法排序结果和梯度提升树模型排序结果均表明，最大雨强和坡度对侵蚀产沙的影响强于对产流的影响；而平均雨强、降雨历时、作物生长天数和距离除草天数对产流的影响强于对产沙的影响。

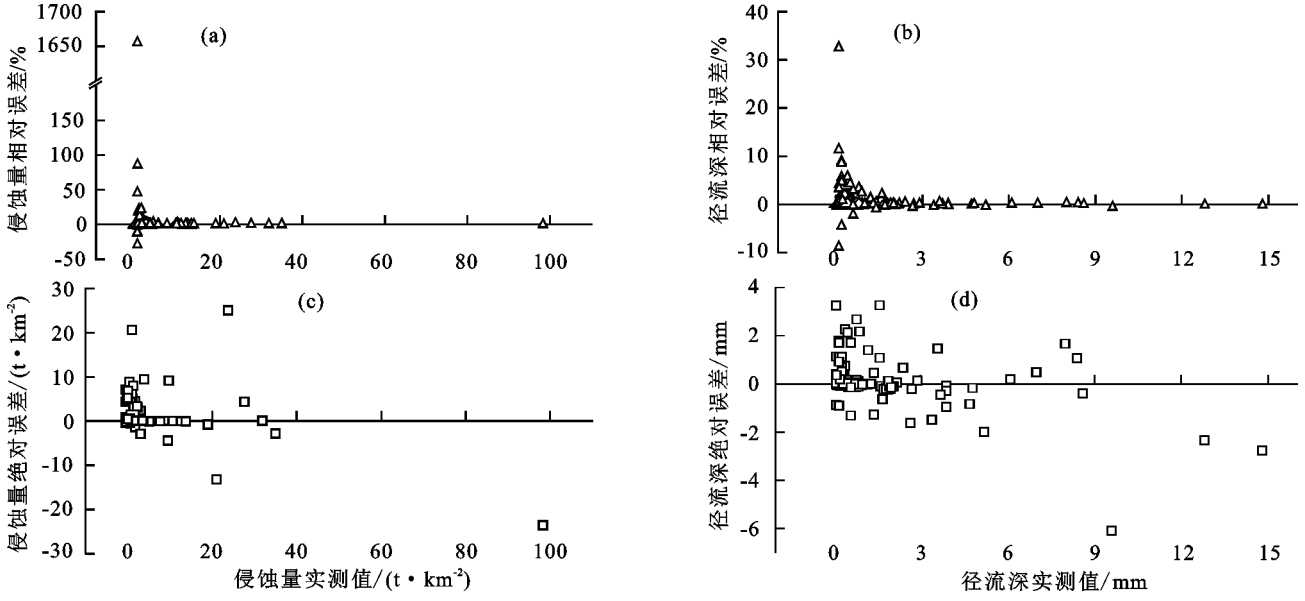


图 4 模型预测结果误差分析

Fig. 4 Error analysis of model prediction results

3 讨 论

本研究表明，基于相同训练集与测试集，在侵蚀量模型与径流深模型预测精度相近情况下，侵蚀量模型层数远大于径流深模型。树模型深度越大，越能够捕获更多数据细节与特征关系，说明产沙过程相较产流过程更加复杂。同时，从节点分裂指标来看，侵蚀量模型使用更多种类的自变量作节点分裂依据，梯度提升树通过迭代生成回归树的方式逐步降低计算结果的误差[22]，分裂方式则从侧面反映各变量间关系复杂程度，再次强调产沙过程的复杂性。由于径流深模型具有更为简洁的模型结构，具备更佳的适应性，径流深模型的 AIC 与 RMSE 表现优良于侵蚀模型。

模型模拟结果中侵蚀量的相对误差存在极大的离群值，可能是由于对应的实测值偏低，而相对误差计算公式为绝对误差与实测值的比，绝对误差相同情况下，实测值越小，相对误差越大。同时，该场次降雨为连续性降雨，前一场降雨产生的产沙间歇性的物理结皮对后一场次降雨产沙产生抑制[31]。因此，除与该次降雨密切相关的变量外，降雨前期情况也应当作为变量纳入模型进行考量。另外，侵蚀量和径流深较小时，存在一些微小的、难以捕捉的特征或模式，可能难以被模型识别和利用，进而导致较大的预测误差。在未来研究中，可通过引入更多自变量组合及自变量定量化的准确表达，在模型优化过程寻找被忽视的影响变量。

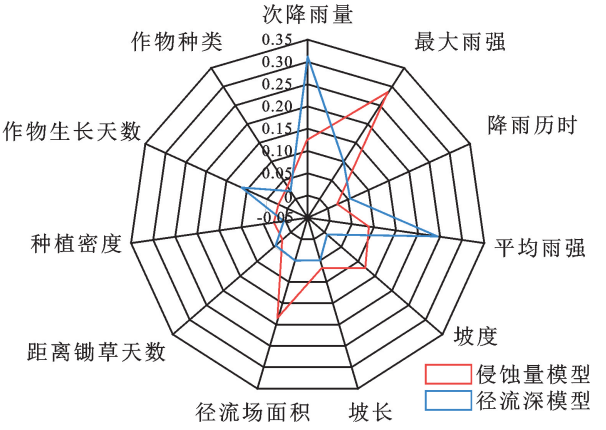


图 5 各预测模型自变量权重雷达图

Fig. 5 Radar chart of independent variable weights for each prediction model

SHAP 算法与梯度提升树模型在因子排序结果上存在差异，主要是由于自变量间存在显著相关性，最大雨强与次降雨量、降雨时长。在分裂增益排序时，对排序结果产生干扰；另一方面，二者排序结果表明，影响产流与产沙的主要自变量存在差异，产流过程主要受降雨本身特征影响，而降雨侵蚀力、区域地形相关因子对产沙过程起到主要影响作用。

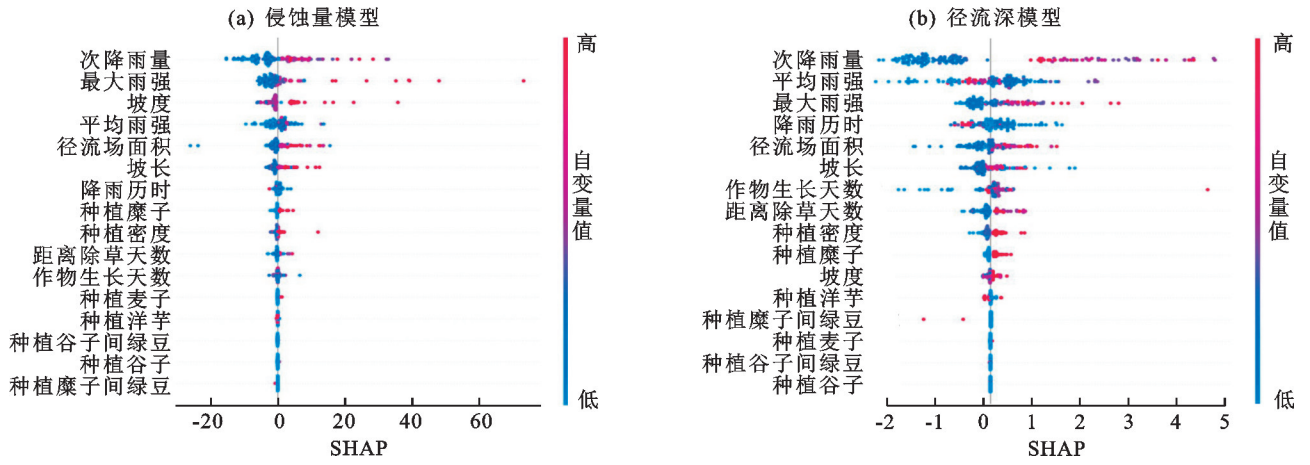


图 6　预测模型自变量 SHAP 值散点图

Fig. 6　Catter plots of independent variables SHAP values for each predictionmodel

降水是产流的物质来源,较大的降雨量增加土壤饱和度,超过土壤的渗透能力后形成产流,随着降雨强度的增大,雨滴的动能也相应提升,进而增强对土壤的侵蚀[32]。现有研究[33]表明,农作物的保水蓄土功能作用重大,作物对产流产沙的影响是地上部分拦蓄截留的作用与根系固结土壤的作用共同形成的[34],陈科兵等[35]研究发现,作物生长使得坡面产流更快趋于稳定,而不同生育期谷子拦蓄径流能力存在差异,与模型排序结果相同。因此,作物的生长可有效削弱径流,减少侵蚀。地形对产沙的影响在于调节水流的速度和侵蚀力。陡峭的地形加速水流的流动速度和能量,增强水流对土壤的侵蚀能力[36],从而促进产沙的形成。综上,尽管影响因素对产流产沙过程都有一定的影响,但各因子对产流产沙变化的贡献不同,说明产流产沙主要影响因素不一致。

综上,梯度提升树模型在众多领域的应用已较为成熟[22],表现出比其他模型更强的性能。在土壤侵蚀研究中,梯度提升树模型除进行预测预报外,也可单独作为侵蚀因素分析工具,由于精度相对较高,且同时能够对自变量影响程度进行快速排序,在未来侵蚀研究中具有广泛的应用空间。但本次研究仍有不足,首先,在侵蚀量实测值较低时,预报精度往往差强人意,在后续研究中,应当尽可能的丰富数据库自变量种类,如坡位、坡向、土壤可蚀性、土壤水分条件、相邻次降雨的间隔时间等,或者进行分级预测。其次,本研究仅考虑同一区域坡耕地状况下的次降雨侵蚀,并涉及区域及下垫面差异,在后续研究中,应当扩展数据集,探究不同区域、不同下垫面情况下,梯度提升树模型作为侵蚀预报与侵蚀因素分析工具的适用性。

4　结 论

基于坡耕地径流小区观测数据,结合梯度提升树模型与 SHAP 算法,构建径流和侵蚀预测模型,定量解析坡耕地水蚀驱动因子。数据集次降雨量、最大雨强、降雨历时、平均雨强、径流深、侵蚀量等变量变异性较大,属于中等变异和强变异。径流深与侵蚀量属右偏态分布,表明其大多数样本小于平均值。基于数据集对模型训练后,训练得到的侵蚀量模型测试精度($R^2=0.81$)略优于径流深模型($R^2=0.80$),但侵蚀量模型结构复杂,泛化性弱于径流深模型,表明侵蚀产沙过程比产流过程更为复杂。坡耕地产流产沙的主要影响因素存在差异,产流过程主要受降水本身特征影响,而产沙过程则除降水特征外,还受到地形相关因子的影响。梯度提升树模型能够较好的预测坡耕地次降雨产流产沙量,定量化解析影响因素的作用,可用于深入解析土壤侵蚀机制的研究。

本研究囿于原始数据,构建数据库时存在一定的局限性,可能存在缺失变量影响模型构建的准确性,后续研究将收集更多区域数据,尝试引入更多自变量组合与自变量表达方式,提升梯度提升树模型预报精度,探究不同区域、下垫面状况下梯度提升树模型是适用性情况。

参考文献:

[1] NGUYEN K A, CHEN W, LIN B S, et al. Using machine learning-based algorithms to analyze erosion rates of a Watershed in Northern Taiwan[J]. Sustainability, 2020,12(5):e2022.

[2] 江忠善, 郑粉莉. 坡面水蚀预报模型研究[J].水土保持学报,2004,18(1):66-69.
JIANG Z S, ZHENG F L. Water erosion prediction model at hillslope scale[J]. Journal of Soil Water Conservation,2004,18(1):66-69.

[3] 张玉斌,郑粉莉,贾媛媛.WEPP 模型概述[J].水土保持研究,2004,11(4):146-149.
ZHANG Y B, ZHENG F L, JIA Y Y. Overview of

WEPP model[J]. Research of Soil and Water Conservation,2004,11(4):146-149.

[4] 李凤,吴长文.RUSLE 侵蚀模型及其应用[J].水土保持研究,1997,4(1):109-112.
LI F, WU C W. RUSLE erosion model and its application[J]. Research of Soil and Water Conservation,1997,4(1):109-112.

[5] 郑海金,汤崇军,汪邦稳,等.基于人工神经网络的红壤坡面水土流失预测模型[J].亚热带水土保持,2009,21(4):10-13,33.
ZHENG H J, TANG C J, WANG B W, et al. Predicting model of soil and water loss based on manpower neural network on red-soil slope[J].Subtropical Soil and Water Conservation,2009,21(4):10-13,33.

[6] 陈劭锋,刘全友,陆中臣,等.黄土高原多沙粗沙区侵蚀产沙的多维临界[J].生态学报,2007,27(8):3277-3285.
CHEN S F, LIU Q Y, LU Z C, et al. The multi-dimensional thresholds of sediment yield in the area with abundant and coarse sediment on the Loess Plateau[J]. Acta Ecologica Sinica,2007,27(8):3277-3285.

[7] 洪伟,吴承祯.闽东南土壤流失人工神经网络预报研究[J].土壤侵蚀与水土保持学报,1997,11(3):53-58.
HONG W, WU C Z. Study on artificial neural network prediction of soil loss in southeast Fujian[J].Journal of Soil and Water Conservation,1997,11(3):53-58.

[8] 彭清娥,曹叔尤,刘兴年,等.坡面产沙 BP 神经网络模型研究[J].水土保持学报,2002,16(3):79-82.
PENG Q E, CAO S Y, LIU X N, et al. Research on BP neural network model for slope sediment yield[J].Journal of Soil and Water Conservation,2002,16(3):79-82.

[9] 赵西宁,王万忠,吴普特,等.坡面入渗的人工神经网络模型研究[J].农业工程学报,2004,20(3):48-50.
ZHAO X N, WANG W Z, WU P T, et al. Artificial neural network model for soil infiltration in slope farmland[J].Transactions of the Chinese Society of Agricultural Engineering,2004,20(3):48-50.

[10] 段军彪,景旭,上官周平.基于遗传算法的 BP 网络在小流域侵蚀量预测中的应用[J].西北农业学报,2008,17(2):317-320.
DUAN J B, JING X, SHANGGUAN Z P. Application of predicting soil erosion in the small watershed based on BP network model by genetic algorithm[J]. Acta Agriculturae Boreali-Occidentalis Sinica,2008,17(2):317-320.

[11] 黄俊,金平伟,向家平,等.基于 BP 网络和回归分析的红壤可蚀性预测[J].中国水土保持科学,2015,13(3):8-15.
HUANG J, JIN P W, XIANG J P, et al. Prediction of red soil erodibility based on BP neural network and regression analysis[J].Science of Soil and Water Conservation,2015,13(3):8-15.

[12] 李世欣,温建,邵孝侯,等.偏最小二乘法与人工神经网络耦合的小流域产沙模型[J].河海大学学报(自然科学版),2010,38(2):149-153.
LI S X, WEN J, SHAO X H, et al. Sediment yield model for small watersheds based on coupling of partial least square regression and artificial neural network[J]. Journal of Hohai University (Natural Sciences),2010,38(2):149-153.

[13] 王权威,唐莉.基于 ABC-BP 的土壤侵蚀量预报模型研究[J].水力发电,2017,43(9):1-4,44.
WANG Q W, TANG L. Study on soil erosion prediction model based on artificial bee colony algorithm and BP neural network[J].Water Power,2017,43(9):1-4,44.

[14] 胡亚萍,董贝贝,姜宏立,等.SVR 与 BP 神经网络对小流域次降雨侵蚀产沙预测结果的比较[J].北方环境,2013,25(1):114-117.
HU Y P, DONG B B, JIANG H L, et al. Comparison of the results of the small watershed rainfall erosion and sediment yield predicted by SVR and BP nerve network [J].Northern Environment,2013,25(1):114-117.

[15] DE LA ROSA D, MAYOL F, MORENO J A, et al. An expert system/neural network model (ImpelERO) for evaluating agricultural soil erosion in andalucia region, southern Spain[J]. Agriculture, Ecosystems and Environment,1999,73(3):211-226.

[16] LICZNAR P, NEARING M A. Artificial neural networks of soil erosion and runoff prediction at the plot scale[J].Catena,2003,51(2):89-114.

[17] ARABAMERI A, SANTOSH M, SAHA S, et al. Spatial prediction of shallow landslide: Application of novel rotational forest-based reduced error pruning tree[J]. Geomatics, Natural Hazards and Risk, 2021, 12(1): 1343-1370.

[18] ARABAMERI A, PRADHAN B, REZAEI K. Spatial prediction of gully erosion using ALOS PALSAR data and ensemble bivariate and data mining models[J].Geosciences Journal,2019,23(4):669-686.

[19] ARABAMERI A, PRADHAN B, REZAEI K, et al. Spatial modelling of gully erosion using evidential belief function, logistic regression, and a new ensemble of evidential belief function-logistic regression algorithm[J]. Land Degradation and Development,2018,29(11):4035-4049.

[20] HAO J Q, LIN Y, REN G X, et al. Comprehensive benefit evaluation of conservation tillage based on BP neural network in the Loess Plateau[J].Soil and Tillage Research,2021,205:e104784.

[21] 周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,40(6):1229-1251.

ZHOU F Y, JIN L P, DONG J. Review of convolutional neural network[J]. Chinese Journal of Computers, 2017,40(6):1229-1251.

[22] 孟然,沈蔚,栾奎峰,等.基于梯度提升决策树算法的水深反演研究[J].海洋湖沼通报,2023,45(1):45-50.
MENG R, SHEN W, LUAN K F, et al. Water depth retrieval based on gradient boosting decision tree algorithm[J]. Transactions of Oceanology and Limnology, 2023,45(1):45-50.

[23] ELITH J, LEATHWICK J R, HASTIE T. A working guide to boosted regression trees[J]. Journal of Animal Ecology,2008,77(4):802-813.

[24] 曹玉茹,高洋洋.基于 SHAP 值惩罚特征的集成分类方法研究[J].统计与决策,2023,39(6):21-26.
CAO Y R, GAO Y Y. An ensemble classification method based on SHAP value to penalize features[J]. Statistics and Decision,2023,39(6):21-26.

[25] 水利部黄河水利委员会. 黄河流域子洲径流实验站水文实验资料数据集(1959—1969 年)[S]. 地球系统科学数据共享平台-黄土高原科学数据共享平, 2012
Yellow River Conservancy Commission of the Ministry of Water Resources. Dataset of hydrological experiments at Zizhou station in the Yellow River Basin (1959—1969)[S]. Data Sharing Infrastructure of Earth System Science_Data Sharing Infrastructure of Loess Plateau, 2012.

[26] 张龙齐,贾国栋,吕相融,等.黄土高原典型地区不同植被覆盖下坡面土壤侵蚀阈值研究[J].水土保持学报,2023,37(2):187-198.
ZHANG L Q, JIA G D, LÜ X R, et al. Research of soil erosion thresholds on the lower slopes of different vegetation cover in typical areas of Loess Plateau[J]. Journal of Soil and Water Conservation,2023,37(2):187-198.

[27] 王尧,蔡运龙,潘懋.贵州省乌江流域土壤侵蚀模拟:基于 GIS、RUSLE 和 ANN 技术的研究[J].中国地质,2014,41(5):1735-1747.
WANG Y, CAI Y L, PAN M. Soil erosion simulation of the Wujiang river basin in Guizhou province based on GIS, RUSLE and ANN[J]. Geology in China, 2014, 41(5):1735-1747.

[28] QIU M L, VAN DE VOORDE T, LI T, et al. Spatio-temporal variation of agroecosystem service trade-offs and its driving factors across different climate zones[J]. Ecological Indicators,2021,130:e108154.

[29] HAMILTON J. Game theory: Analysis of conflict, by myerson, R. B., Cambridge: Harvard university press[J]. Managerial and Decision Economics,1992,13(4):369.

[30] EBRAHIMI-KHUSFI Z, TAGHIZADEH-MEHRJARDI R, ROUSTAEI F, et al. Determining the contribution of environmental factors in controlling dust pollution during cold and warm months of western Iran using different data mining algorithms and game theory[J]. Ecological Indicators,2021,132:e108287.

[31] 吴发启,范文波.土壤结皮对降雨入渗和产流产沙的影响[J].中国水土保持科学,2005,3(2):97-101.
WU F Q, FAN W B. Effects of soil encrustation on rainfall infiltration, runoff and sediment generation[J]. Science of Soil and Water Conservation,2005,3(2):97-101.

[32] 何绍浪,李凤英,何小武.水蚀预报中降雨侵蚀力研究进展[J].水土保持通报,2018,38(2):262-270.
HE S L, LI F Y, HE X W. Research progress of rainfall erosivity for water erosion prediction[J]. Bulletin of Soil and Water Conservation,2018,38(2):262-270.

[33] 李小辉,贾本有,范子武,等.典型作物对水土流失影响的小区试验研究[J].水利水电技术,2019,50(2):95-100.
LI X H, JIA B Y, FAN Z W, et al. Runoff plot observation study on the influences of typical crops on water and soil erosion[J]. Water Resources and Hydropower Engineering,2019,50(2):95-100.

[34] 杨晓芬,吴发启,马波,等.黄土坡耕地玉米作物的防蚀作用研究[J].西北农林科技大学学报(自然科学版),2012,40(10):97-102.
YANG X F, WU F Q, MA B, et al. Studies on anti-erosion effects of the maize in loess sloping fields[J]. Journal of Northwest A&F University (Natural Science Edition),2012,40(10):97-102.

[35] 陈科兵,吴发启,姚冲.黄土高原南部地区人工模拟暴雨条件下不同坡度谷子坡耕地产流产沙过程[J].水土保持学报,2021,35(3):90-95,103.
CHEN K B, WU F Q, YAO C. Runoff and sediment yield processes in millet cultivated land with different slopes under artificial simulated rainstorm in the southern Loess Plateau[J]. Journal of Soil and Water Conservation,2021,35(3):90-95,103.

[36] 崔钦凯,刘俊娥,陈浩,等.草被覆盖、雨强和坡度对黄土坡面径流含沙量的影响[J].水土保持学报,2023,37(5):40-47.
CUI Q K, LIU J E, CHEN H, et al The effects of grass cover, rain intensity, and slope on the sediment content of Loess Slope runoff[J]. Journal of Soil and Water Conservation,2023,37(5):40-47.