

DOI:10.13870/j.cnki.stbcbx.2026.03.021 CSTR:32310.14.stbcbx.2026.03.021

郑海清,张美玲,伊兰.基于机器学习模型的张掖地区土壤有机碳与全氮估算[J].水土保持学报,2026,40(3):412-423.

ZHENG Haiqing, ZHANG Meiling, YI Lan. Estimation of soil organic carbon and total nitrogen in Zhangye region based on machine learning models [J]. Journal of Soil and Water Conservation, 2026, 40(3): 412-423.

## 基于机器学习模型的张掖地区土壤有机碳与全氮估算

郑海清,张美玲,伊兰

(甘肃农业大学理学院,数量生物学研究中心,兰州730070)

**摘要:** [目的] 为提高张掖地区0~30 cm土层土壤有机碳(SOC)与全氮(TN)的空间估算精度,并识别关键驱动因子及其作用方向。[方法] 整合气候、地形、植被与土壤理化性质等多源协变量,结合实测样点( $N=979$ )构建特征集;系统比较9类机器学习模型,并筛选XGBoost、梯度提升回归树(GBRT)与随机森林(RF)为代表性树基模型;进一步构建自适应加权(Auto-Weighted)、混合法(Blending)与自助聚合(Bagging)3类融合模型,采用10折交叉验证与独立测试集综合评估,并以SHAP量化变量贡献及作用方向。[结果] 单模型测试集上,SOC估算中XGBoost、GBRT、RF的 $R^2$ 分别为0.768、0.773、0.729, RMSE分别为0.472、6.421、7.011, MAE分别为0.314、4.319、4.641; TN估算中三者 $R^2$ 分别为0.636、0.645、0.629, RMSE分别为0.602、5.381、5.498, MAE分别为0.360、3.225、3.265。融合模型整体优于单模型,其中Auto-Weighted最优(SOC:  $R^2_{\text{test}}=0.879$ , RMSE=4.607 9, MAE=2.948 2; TN:  $R^2_{\text{test}}=0.775$  8, RMSE=4.330 8, MAE=2.418 6), Blending次之(SOC:  $R^2_{\text{test}}=0.848$  7, RMSE=5.107 2, MAE=2.991 9; TN:  $R^2_{\text{test}}=0.734$  6, RMSE=4.576 2, MAE=2.337 3)。SHAP解释表明,降水与植被指数呈稳定正向贡献,最低气温等温度因子呈负向贡献,阳离子交换量(CEC)及黏粒、粉粒比例等土壤性质贡献突出;空间上,高值区主要分布于水分相对充足、温度较低且植被覆盖较好的山地,低值区集中于绿洲边缘与荒漠过渡带。[结论] 树基模型融合与SHAP解释相结合可实现SOC与TN高精度、稳健且可解释的区域化制图,为干旱-半干旱过渡带碳氮评估与分区管理提供定量依据。

**关键词:** 土壤有机碳; 全氮; 机器学习; 融合模型; SHAP

中图分类号: S153.6

文献标识码: A

文章编号: 1009-2242(2026)03-0412-12

## Estimation of Soil Organic Carbon and Total Nitrogen in Zhangye Region Based on Machine Learning Models

ZHENG Haiqing, ZHANG Meiling, YI Lan

(Center for Quantitative Biology, College of Science, Gansu Agricultural University, Lanzhou 730070, China)

**Abstract:** [Objective] This study aims to enhance the spatial estimation accuracy of soil organic carbon (SOC) and total nitrogen (TN) in the 0–30 cm soil layer in the Zhangye region and to identify the key driving factors and their directional effects. [Methods] Multi-source covariates, including climate, topography, vegetation, and soil physicochemical properties, were integrated to construct a feature set based on field sampling data ( $N=979$ ). Nine types of machine learning models were systematically compared, from which XGBoost, gradient boosting regression trees (GBRT), and random forest (RF) were selected as representative tree-based models. Three ensemble strategies were adopted to establish ensemble models, including auto-weighted, blending, and bagging. Their performance was comprehensively evaluated using 10-fold cross-validation and an independent test set. SHapley Additive exPlanations (SHAP) values were applied to quantify variable contributions and their directional effect. [Results] For single models on the test set, SOC estimation yielded  $R^2$  values of 0.768, 0.773, and 0.729, root mean square error (RMSE) values of 0.472, 6.421, and 7.011, and mean absolute error (MAE)

收稿日期: 2025-11-13

修回日期: 2025-12-30

录用日期: 2026-01-21

网络首发日期(www.cnki.net): 2026-03-18

资助项目: 国家自然科学基金项目(32260353); 甘肃省重点研发计划项目(21YF5WA096); 科技部高端外国专家引进计划项目(G2022042009L); 甘肃省自然科学基金项目(1606RJZA077, 1308RJZA262)

第一作者: 郑海清(1997—), 男, 硕士研究生, 主要从事土壤有机碳、全氮估算研究。E-mail: 2876018939@qq.com

通信作者: 张美玲(1978—), 女, 博士, 教授, 博士生导师, 主要从事碳氮模型模拟和气候变化研究。E-mail: zhangmeiling@gsau.edu.cn

http://stbcbx.alljournal.com.cn

values of 0.314, 4.319, and 4.641 for XGBoost, GBRT, and RF, respectively. For TN estimation, the single models produced  $R^2$  values of 0.636, 0.645, and 0.629, RMSE values of 0.602, 5.381, and 5.498, and MAE values of 0.360, 3.225, and 3.265, respectively. Overall, the ensemble models outperformed the single models. The auto-weighted demonstrated the best performance (SOC:  $R^2_{\text{test}}=0.887\ 9$ , RMSE=4.607 9, MAE=2.948 2; TN:  $R^2_{\text{test}}=0.775\ 8$ , RMSE=4.330 8, MAE=2.418 6), followed by blending (SOC:  $R^2_{\text{test}}=0.848\ 7$ , RMSE=5.107 2, MAE=2.991 9; TN:  $R^2_{\text{test}}=0.734\ 6$ , RMSE=4.576 2, MAE=2.337 3). SHAP analysis revealed stable positive contributions from precipitation and vegetation indices, negative contributions from temperature factors such as minimum temperature, and prominent contributions from soil properties including cation exchange capacity (CEC) and the proportions of clay and silt. Spatially, areas with high SOC and TN values were mainly distributed in mountainous regions characterized by relatively sufficient moisture, lower temperatures, and better vegetation coverage. In contrast, low-value areas were concentrated at oasis margins and desert transition zones. [Conclusion] The integration of tree-based ensemble modeling and SHAP interpretation enables highly accurate, robust, and interpretable regional mapping of SOC and TN. This approach provides a quantitative basis for carbon and nitrogen assessment and zonal management in arid-semiarid transition zones.

**Keywords:** soil organic carbon; total nitrogen; machine learning; ensemble model; SHAP

**Received:** 2025-11-13

**Revised:** 2025-12-30

**Accepted:** 2026-01-21

**Online**(www.cnki.net): 2026-03-18

土壤有机碳(soil organic carbon, SOC)与全氮(total nitrogen, TN)是评价土壤质量与生态系统服务功能的核心要素,也是陆地碳氮循环的物质基础与关键“调节器”。SOC直接关系土壤肥力、作物产量与生态系统生产力,并在全球陆地碳库中占据重要地位;TN通过影响植被生长与微生物活性,进而调控土壤养分循环与生态系统结构稳定性。二者的协同变化对维系土壤健康与提升生产力具有基础意义。全球气候变化背景下,SOC作为重要陆地碳汇,对调节大气CO<sub>2</sub>浓度与缓解变暖具有不可替代作用;与此同时,极端气候事件加剧、水热格局重塑与土地利用变迁,使土壤碳氮库的时空动态及其对气候系统的反馈更趋复杂与不确定。

西北干旱与半干旱区是全球气候变化敏感区域之一,生态环境脆弱,对水热变化响应显著。张掖地处河西走廊中部,属典型干旱-半干旱过渡区,地貌类型多样、农牧复合生态系统典型,草地与耕地资源对区域水源涵养、防风固沙与生态安全屏障建设具有重要意义。近年来,基于机器学习的SOC与TN空间估算研究快速发展。WADOUX等<sup>[1]</sup>研究指出,集成树模型因其对非线性、交互项与高维协变量的鲁棒性,已成为土壤属性预测的重要主力算法;POGGIO等<sup>[2]</sup>在SoilGrids 2.0的全球制图中进一步表明,基于大样本观测并结合多源环境协变量(气候、植被、地形等)可显著提升土壤属性空间制图能力。鉴于该类协变量通常呈现明显的非线性响应与多尺度交互特征,相较于线性回归等需要预设函数形式的传统方法,随机

森林、梯度提升与XGBoost等树基集成模型能够在较少先验假设下自动学习复杂的非线性关系与高阶交互,因此在数字土壤制图(DSM)与土壤属性空间预测研究中往往表现出更高的预测精度与更强的适用性。

现有SOC、TN估算研究普遍采用实测样点和多源环境协变量的建模范式,其协变量配置并非任意堆叠,而是与成土因子及碳氮形成和周转过程存在明确对应关系。YU等<sup>[3]</sup>在SOC估算中发现,温度与降水长期位居关键驱动因子。其原因在于,气候因子通过改变区域水热条件与有效水分供给,进而影响植被生产力与微生物分解,最终共同调节SOC,并在很大程度上影响TN的积累。植被遥感指标可用于表征地表碳输入潜力。归一化植被指数(NDVI)反映植被覆盖度与生长活力的空间差异,可间接指示凋落物与根系周转等有机质输入强度的相对变化,因此常作为SOC数字制图的重要生态协变量<sup>[4]</sup>。植被净初级生产力(NPP)直接表征植被净生产力及其空间差异,可用于表征区域尺度生物量积累与潜在有机质输入规模,为SOC与TN的估算提供过程指向的协变量支撑<sup>[5]</sup>。WERE等<sup>[6]</sup>以南非地区为例开展SOC储量预测,将Mg、Ca、P、全氮与pH等土壤属性与遥感、地形及其衍生因子联合纳入模型,研究结果显示,土壤理化部分因子有助于提升SOC空间估算的稳定性与解释性。在具体应用层面,我国学者也在耕地、林地等不同地表类型上验证多源协变量与树模型的有效性,如周洋等<sup>[7]</sup>利用多源辅助变量与随机森林开展表层土壤全氮空间预测;刘尊方等<sup>[8]</sup>基于XGBoost实现对耕地土壤全氮

的遥感反演并获得较高精度,二者均从应用层面支持上述“过程指向”的协变量组织方式。

基于此,本文以张掖地区 0~30 cm 土层为研究对象,整合气候(Tmin、Tmax、Tmean、Pre)、地形(ALT、SLP、ASP、TRI等)、植被(NDVI、NPP)与土壤理化属性(质地、BD、pH、CEC等)等多源特征,系统评估 9 类常用机器学习模型对 SOC 与 TN 的估算能力;在此基础上构建自适应加权(auto-weighted)融合方案以提升泛化性能与稳健性,并引入基于 SHAP 的可解释性分析识别主控因子及其作用方向,阐明气候、植被生产力、土壤性质与地形过程对 SOC 与 TN 空间格局的共同影响,为干旱与半干旱区碳氮评估与区域生态管理提供方法支撑。

## 1 材料与方法

### 1.1 研究区概况

张掖位于甘肃省河西走廊中段(37°28′~39°57′N, 97°20′~102°12′E),行政区划包括甘州区、临泽县、高台县、山丹县、民乐县及肃南裕固族自治县。区域自南向北依次呈现“祁连山高山-山前冲洪积扇与绿洲平原-荒漠与戈壁”的地貌序列,属高原大陆性气候,降水随海拔升高而增加并主要集中于夏季。土壤类型以高山草甸土、栗钙土、灰钙土及盐渍化土为主;植被沿海拔梯度呈高山草甸-草原-荒漠的带谱更替,黑河为区域主要水系。在显著的地形-气候-植被-土壤性质梯度背景下,SOC 与 TN 表现出高度的时空异质性,适宜借助机器学习模型开展高分辨率制图并识别主导驱动因子。研究区域见图 1。

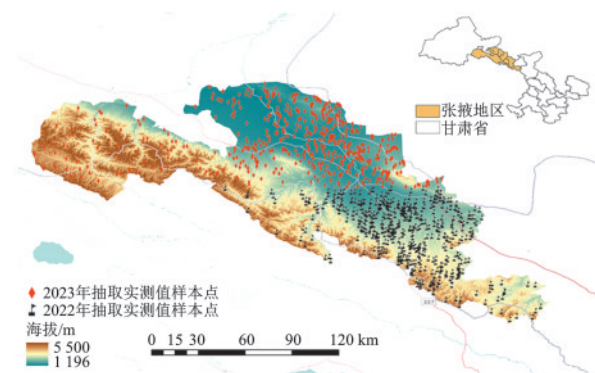


图 1 研究区域地理位置

Fig.1 Geographical location of study area

### 1.2 数据来源

1.2.1 土壤数据来源 本研究 SOC 与 TN 实测值来源于 2022—2023 年野外实地采样并经实验室分析获得。为覆盖研究区主要环境梯度,在各气象站点周边采用典型样点布设与分层取样相结合的方式开展野外采样。野外采样前,基于 ArcGIS 预先设置候选

采样点;外业阶段使用手持 GPS 导航到达样点,并记录样点的经纬度与海拔等信息。野外采样结束后,将坐标导入 ArcGIS 10.8 软件进行空间核验与统一管理,确保采样点位置与后续多源栅格因子匹配一致。采样点主要涵盖山地草原、绿洲农田及荒漠-戈壁裸地等。每个采样点设置 3 个标准样方以提高代表性,采用专用土钻分层取样,土层深度为 0~10、10~20、20~30 cm。为保证测量精度,每份土样采集量不少于 200 g;采集后置于密封袋并编号。样品运回实验室后自然风干,随后经 0.15 mm 筛网过筛以去除大颗粒杂质并保证均匀性;筛后样品质量均 >100 g,以满足后续检测需求。SOC 与 TN 含量采用全自动氮-形态碳分析仪(SKALAR Primacs SNC 100-IC-E)测定;试验前将仪器预热至 1 100 °C 以稳定测量条件;测定时称取每份样品 500 mg(称量误差控制在 ±1 mg),置于专用坩埚中进行高温燃烧分析(单样约 8 min),获得 SOC 与 TN 含量数据。

土壤理化性质数据主要来源于国家青藏高原科学数据中心(TPDC, <http://data.tpdc.ac.cn>)。本研究从 TPDC 获取研究区尺度可用的土壤属性栅格数据产品,包括 pH、阳离子交换量 CEC、土壤体积质量 BD、质地组分等定量理化参数;并对其进行统一投影、分辨率匹配与研究区裁剪。随后,基于采样点坐标在 ArcGIS 10.8 软件中采用栅格采样提取各点位对应的土壤属性值,实现与 SOC、TN 样点的一致空间匹配。所用土壤属性数据产品相关数据经过规范化的采集、处理与质量控制,可作为区域尺度土壤背景协变量用于提升 SOC、TN 估算的稳定性与解释性。

1.2.2 气象数据 本研究所用气象数据来源于国家青藏高原科学数据中心(TPDC, <https://data.tpdc.ac.cn>)。该数据集基于 CRU 全球 0.5° 气候数据集与 WorldClim 高分辨率气候数据,采用 Delta 空间降尺度方案在中国区域生成,并使用 496 个地面气象站观测数据开展独立验证。本研究提取 2022—2023 年逐月最高气温、最低气温、平均气温及降水,并在 ArcGIS 10.8 软件中完成研究区范围裁剪与栅格格式转换,以确保与研究区样点数据的空间精确匹配。

1.2.3 高程数据 研究区的高程数据来源于 NASA Earth data 平台下阿拉斯加卫星设施数据中心(ASFDAAC; <https://search.asf.alaska.edu/#/>)分发的 ALOS 数字高程模型(DEM,空间分辨率 12.5 m)。基于 ArcGIS 10.8 的 Spatial Analyst 模块,首先由 DEM 派生坡度(slope)、坡向(aspect)与地形粗糙度等地形因子,其中地形粗糙度以移动窗口内高程标准差(standard deviation, STD)表征。鉴于本文建模

与制图以1 km目标网格作为统一支持尺度,上述地形衍生因子在12.5 m计算后进一步进行尺度聚合至1 km(与其他协变量保持一致),随后采用栅格采样提取至SOC、TN采样点用于建模。

1.2.4 植被指标数据 植被指数(NDVI)基于Sentinel-2卫星影像数据计算获得,本研究通过USGS EarthExplorer (<https://earthexplorer.usgs.gov/>)平台获取研究区2022—2023年影像。为保证数据质量,筛选云量低于10%的影像,并进行必要的质量控制与预处理。依托ArcGIS 10.8软件对影像数据依次开展镶嵌与裁剪等处理,统一坐标系并获取研究区地表反射率,分别计算2022年与2023年NDVI以表征植被绿度,计算公式为:

$$NDVI = \frac{NIR - R}{NIR + R} \quad (1)$$

表1 数据汇总

Table 1 Data summary

类别	变量名称
土壤变量	土壤有机碳(SOC)、全氮(TN)、砂粒占比(SAND)、黏粒占比(CLAY)、粉粒占比(SILT)、土壤体积质量(BD)、土壤酸碱度(PH)、速效钾(AK)、交换性铝(AL)、碱解氮(AN)、速效磷(AP)、交换性钙(CA)、阳离子交换量(CEC)、黏土含量(CL)、土壤结持性比例(CW1)、土壤结持性(C1)、砾石含量(GRAV)、可交换的H <sup>+</sup> 离子(H)、可交换的K <sup>+</sup> 离子(K)、土壤颜色-彩度干土(Dc)、土壤颜色-色相干土(Dh)、数据字典定义的K <sub>1</sub> 指标(K <sub>1</sub> )、交换性镁(MG)、交换性钠(NA)、孔隙度(POR)、降雨侵蚀力因子(R)、土壤结构类型比例(SW1)、土壤结构类型(S1)、土壤颜色湿土(Unh)、土壤色调湿土(Wh)、洼地深度(PDEP1)、局地洼地深度(LDEP1)、局地洼地数量(LNUM1)、地形位置指数(SI)
地形变量	坡度(SLP)、坡向(ASP)、地形粗糙度(TRI)、海拔(ALT)
气候变量	年平均最低气温(T <sub>min</sub> )、年平均最高气温(T <sub>max</sub> )、年平均气温(T <sub>mean</sub> )、年降水量(Pre)
植被变量	归一化植被指数(NDVI)、植被净初级生产力(NPP)

### 1.3 研究方法

1.3.1 机器学习算法 以张掖地区作为研究对象,选取生态与土壤研究中常用的九类机器学习模型用于SOC与TN的估算:线性回归(linear regression, LR)、支持向量回归(support vector regression, SVR)、多层感知机(multilayer perceptron, MLP)、随机森林(random forest, RF)、梯度提升回归树(gradient boosting regression tree, GBRT)、极端梯度提升(XGBoost)、light gradient boosting machine(LightGBM)、CatBoost(CatBoost)和AdaBoost(AdaBoost)。在模型训练前,对连续特征在训练集上进行标准化处理,并将相同参数应用于验证和测试数据。为降低多重共线性干扰,结合相关性阈值与方差膨胀因子(VIF)对输入变量进行筛查。

1.3.2 变量筛选与共线性诊断 采用皮尔逊相关分析,探讨SOC和TN与主要环境因子的关系。整合气候、植被、地形和土壤属性等数据为特征变量,对SOC、TN显著相关的因子( $p < 0.05$ )纳入候选集;在

式中:NIR为近红外波段反射率;R为红光波段反射率。

净初级生产力(NPP)数据来源于NASA Earth data(LP DAAC)发布的MODIS年尺度NPP产品MOD17A3HGF v061。该产品由给定年份内8 d尺度总初级生产力(GPP)产品MOD17A2H的累积(总和)计算得到。本研究在ArcGIS 10.8软件中对NPP数据进行研究区裁剪、投影统一与其他环境因子一致处理,并与样点坐标叠加提取样点对应的2022年与2023年NPP值,作为模型输入变量之一。

1.2.5 数据汇总表格 综合气候(年最高气温、最低气温、平均气温及降水)、地形(海拔、坡度、坡向、地形粗糙度)、植被(NDVI、NPP)与土壤理化性质等要素,按统一空间分辨率与时间尺度整合形成特征集见表1。

此基础上进行方差膨胀因子(VIF, 阈值10)共线性诊断,迭代剔除高共线变量,最终筛选出的特征变量用于建模估算SOC与TN含量。

1.3.3 模型优化 在单模型基础上进一步提升泛化性能,依据10折交叉验证5次重复的结果,从表现较优的LR、CatBoost、LightGBM、XGBoost、RF、SVR、GBRT、MLP等基学习器中遴选预测精度高的M个作为候选基模型。记第m个基模型在样本i上的预测为 $\hat{y}_{m,i}$ 。

1)自适应加权(auto-weighted)

在训练数据的折外预测上,采用带约束的最小二乘或岭回归求解权重 $\omega = (\omega_1, \dots, \omega_m)^\top$ ,并施加非负与和为1的约束以防过拟合:

$$\min_{\omega} \sum_i y_i \left( y_i - \sum_{m=1}^M \omega_m \hat{y}_{m,i} \right)^2 + \lambda \|\omega\|_2^2, \text{ s.t. } \omega_m \geq 0, \sum_m \omega_m = 1 \quad (2)$$

得到的 $\omega$ 固定后,最终融合预测为:

$$\hat{y}_i^{(AW)} = \sum_m \omega_m \hat{y}_{m,i} \quad (3)$$

### 2) 混合法 (blending)

将训练集按比例划分为基学习器训练集和保留集。先在训练子集上拟合各基模型,并在保留集上产生预测  $\hat{y}_{m,i}$ ;再以保留集的这些预测为自变量、真实值为因变量,训练一个线性或岭回归元学习器得到融合系数,形成最终预测  $\hat{y}^{(BL)}$ 。

### 3) 自助聚合 (bagging)

以验证表现最优或方差较大的单一基模型为基底,进行  $B$  次自助采样训练,最终预测取平均值:

$$\hat{y}_i^{(BG)} = \frac{1}{B} \sum_{b=1}^B \hat{y}_i^{(b)} \quad (4)$$

**1.3.4 模型验证与评估** 将数据集按 8:2 的比例随机划分为训练集与测试集,其中 80% 的样本用于模型训练与参数优化,剩余 20% 的样本作为独立测试集用于最终性能验证。为降低一次随机划分带来的偶然性并获得更稳健的模型比较结果,在训练集内部采用  $K$  折交叉验证进行参数优化与模型选择:将训练集划分为  $K$  份,每次用其中  $K-1$  份训练、1 份验证,循环  $K$  次,使每份数据均作为验证集 1 次。交叉验证阶段的评价指标在每一折验证集上计算,并对各折结果取平均作为训练阶段模型性能的总体评价;最终模型的泛化能力以独立测试集上的指标结果为准。

模型性能评价采用平均绝对误差 (MAE)、均方根误差 (RMSE) 和判定系数 ( $R^2$ )。设在验证集或测试集中共有  $n$  个样本,观测值为  $y_i$ ,模型预测值为  $\hat{y}_i$  ( $i=1, 2, \dots, n$ ),观测平均值为  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ 。各指标计算公式为:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

式中:MAE 与 RMSE 的量纲与被估算变量一致,分别反映平均误差幅度与对大误差的惩罚强度; $R^2$  为无量纲指标,衡量模型对观测值方差的解释程度, $R^2$  越接近 1 表明拟合效果越好;当模型预测效果劣于以观测平均值  $\bar{y}$  作为基准预测时, $R^2$  可能  $< 0$ 。综合比较各模型在训练集交叉验证与独立测试集上的表现,以 RMSE、MAE 较低且  $R^2$  较高作为最优估算模型的选择依据,并在测试集上报告其最终泛化性能。

## 2 结果与分析

### 2.1 SOC、TN 含量描述性分析及与环境变量的研究分析

张掖地区 SOC、TN 的描述性统计见表 2。张掖地区 0~30 cm 土层的 SOC 质量分数为 5.97~73.36 g/kg,平均值为 25.35 g/kg,标准差为 13.75 g/kg,整体上呈现较大的空间差异性。TN 质量分数为 0.13~45.07 g/kg,平均值为 7.24 g/kg,标准差为 9.05 g/kg,变异程度同样较高。由方差结果可知,SOC 的空间变异幅度明显大于 TN,说明 SOC 的异质性更为显著。总体而言,二者均表现出较高的变异系数,提示研究区 SOC 与 TN 的分布受多种环境因子共同驱动,存在较强的空间非均质性。

依据相关性热力图(图 2),通过检验 SOC、TN 与各环境变量的相关关系。结果表明,不同因子与 SOC、TN 的相关方向与强度存在显著差异;其中部分变量与 SOC、TN 的相关性不显著 ( $p \geq 0.05$ ),故在后续建模中予以剔除。同时,为避免不必要的模型复杂度与冗余信息干扰,对仅包含单一数值或缺乏统计学意义的变量亦不予保留。为提高稳健性,后续分析仅保留通过相关性检验且具有明确物理含义的变量集。

表 2 SOC、TN 描述性统计

Table 2 Descriptive statistics of SOC and TN

指标	样本数	最小值/(g·kg <sup>-1</sup> )	最大值/(g·kg <sup>-1</sup> )	平均值/(g·kg <sup>-1</sup> )	标准差/(g·kg <sup>-1</sup> )	方差/(g·kg <sup>-1</sup> )
SOC	979	5.97	73.36	25.35	13.75	189.07
TN	979	0.13	45.07	7.24	9.05	81.85

依据相关性检验 ( $p < 0.05$ ) 及图 2 热力图, SOC 与 TN 均与多类环境因子呈显著相关,但方向与强度存在差异: SOC 在地形与气候方面,与海拔 (ALT)、降水量 (Pre)、坡度及地形起伏度 (K、K<sub>1</sub>、SLP、TRI) 及地形位置指数 (SI) 呈正相关;在土壤与植被方面,与阳离子交换量 (CEC)、黏粒占比

(CLAY)、孔隙度 (POR)、铵态氮 (AN)、速效磷 (AL)、钙含量 (Ca)、镁含量 (Mg)、钠含量 (Na) 及归一化植被指数 (NDVI) 呈正相关;而与土壤体积质量 (BD)、pH、砂粒含量 (SAND)、交换性氢 (H)、速效钾 (AK)、年最低气温 ( $T_{\min}$ )、年平均气温 ( $T_{\text{mean}}$ )、年最高气温 ( $T_{\max}$ )、坡向 (ASP)、局地地形参数

(LNUM1、LDEP1,以及SI的部分层级)、土壤颜色变量(Dh、Wh、Unh)及净初级生产力(NPP)呈负相关。TN的相关模式与SOC大体相似,但在个别因子上存在差异:TN与地形因子(TRI、SLP、ALT、K、K<sub>1</sub>)、降水量(Pre)、土壤理化属性(CEC、AN、AL、Ca、Mg、Na、CLAY)及NDVI呈显著正相关;与土壤体积质量(BD)、pH、砂粒占比(SAND)、年最低气温( $T_{min}$ )、年平均气温( $T_{mean}$ )、年最高气温( $T_{max}$ )、局地地形参数(LDEP1、PDEP1)及部分结构与颜色因子(C1、S1、Dh、Wh、Unh)呈显著负相关。

为提高后续建模的稳健性,仅将相关性显著且具明确物理意义的变量纳入模型,不显著或冗余变量不再进入分析。剔除不显著变量后,SOC与TN分别保留39个与32个候选因子;为进一步降低自变量之间的冗余性,对上述变量开展多重共线性诊断,结果显示部分因子存在较强共线性(方差膨胀因子 $VIF>10$ )。通过迭代剔除共线性较高的因子,最终获得SOC 30个、TN 25个显著且互不共线的相关特征变量(图3),并以此作为输入特征开展SOC与TN的建模估算。

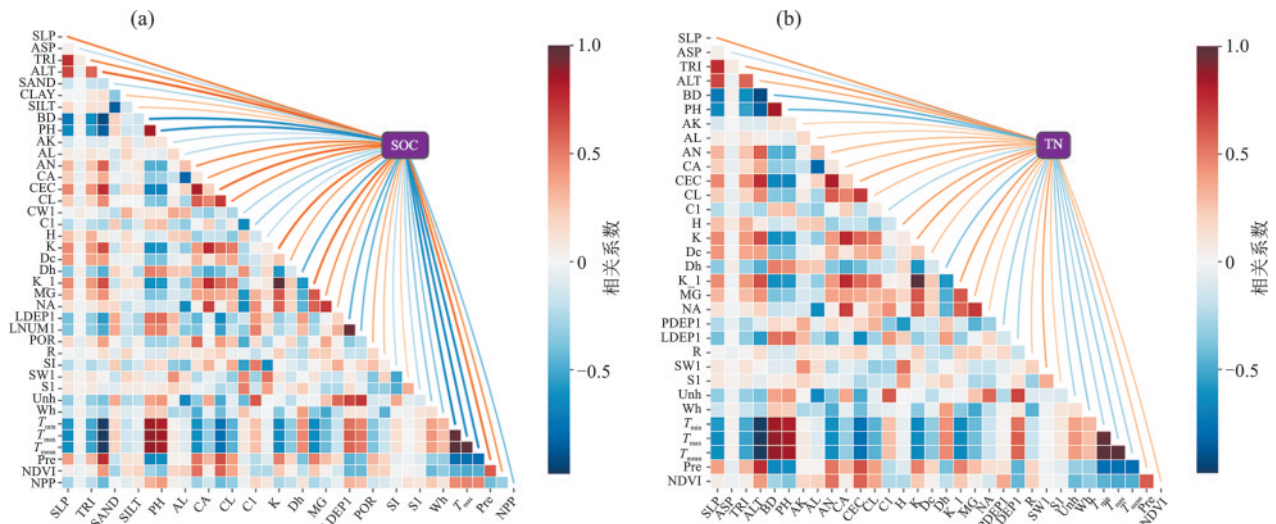


图2 相关性热力图

Fig.2 Correlation heatmap

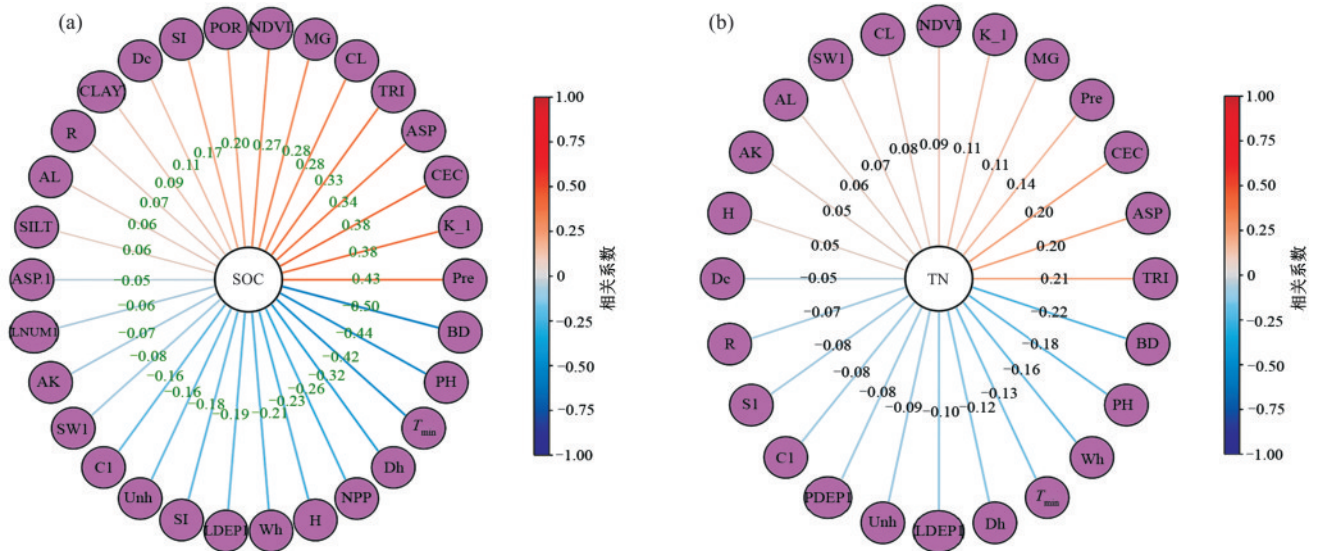


图3 剔除共线性剩余特征变量

Fig.3 Remaining feature variables after removing multicollinearity

### 2.2 机器学习模型选取及影响分析

2.2.1 机器学习模型预测精度评估与验证 在完成相关性筛选与多重共线性诊断后,将保留的显著环境因子作为输入特征,分别构建并评估9类模型:LR、CatBoost、LightGBM、XGBoost、RF、SVR、

AdaBoost、GBRT和MLP。所有模型在相同的8:2训练集与测试集划分及统一随机种子条件下训练;连续特征在训练集上z-score标准化(再映射至测试集),超参数仅在训练集内寻优。模型性能以MAE、RMSE与 $R^2$ 在训练集与测试集上统一报告(表3)。

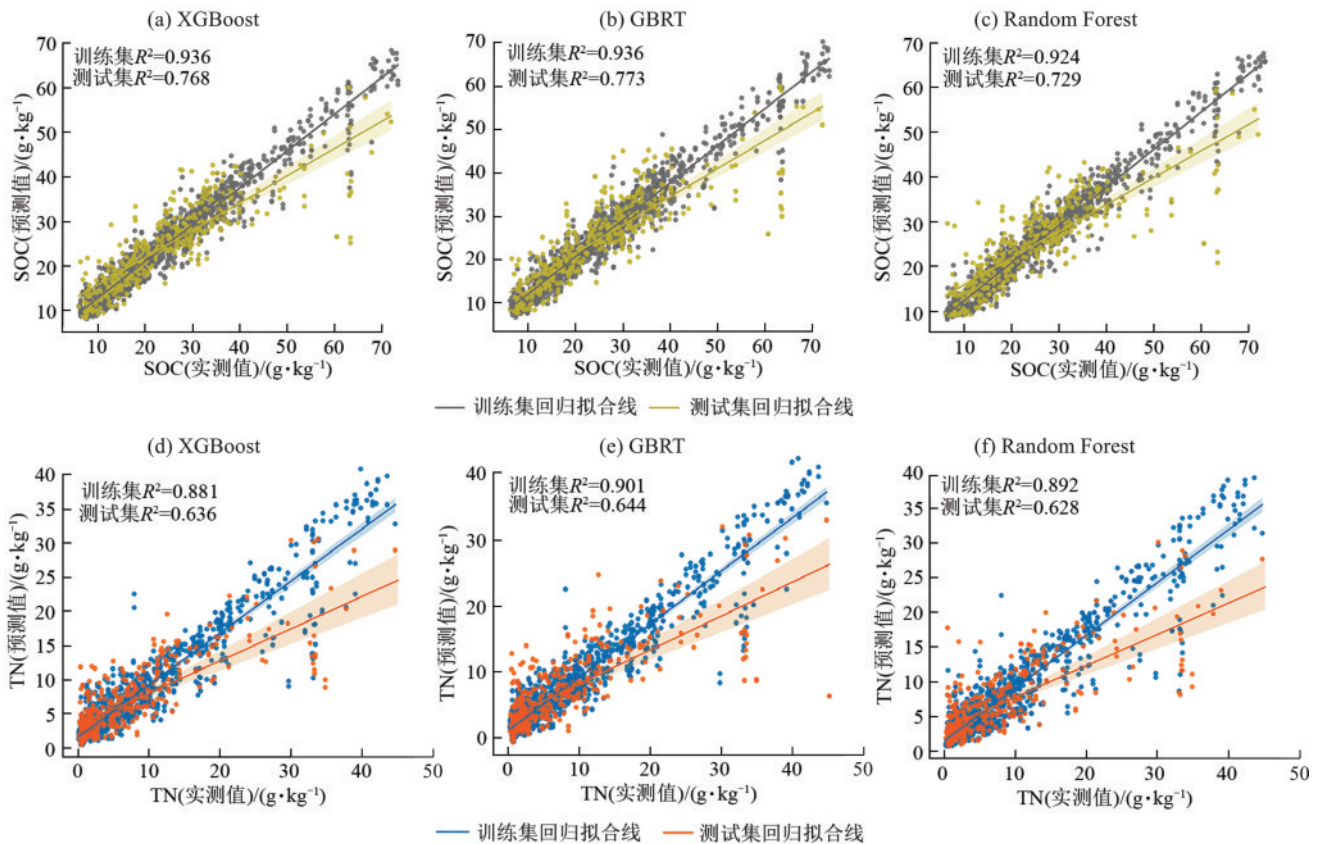
表 3 9 种机器学习模型拟合程度

**Table 3 Goodness of fit of nine machine learning models**

模型	目标变量	训练集			测试集		
		$R^2$	RMSE	MAE	$R^2$	RMSE	MAE
LR	SOC	0.355	11.063	8.191	0.296	11.331	8.363
	TN	0.194	8.106	8.433	0.128	8.433	5.798
CatBoost	SOC	0.682	7.768	5.571	0.557	8.873	6.302
	TN	0.630	5.497	3.639	0.469	6.579	4.204
LightGBM	SOC	0.826	5.745	4.048	0.662	7.837	5.420
	TN	0.747	4.545	2.916	0.539	6.129	3.798
XGBoost	SOC	0.936	0.255	0.183	0.768	0.472	0.314
	TN	0.881	0.344	0.225	0.636	0.602	0.360
RF	SOC	0.924	3.788	2.684	0.729	7.011	4.641
	TN	0.892	2.972	1.991	0.629	5.498	3.265
SVR	SOC	0.525	9.499	5.314	0.422	10.260	6.161
	TN	0.326	7.414	3.487	0.201	8.076	4.123
AdaBoost	SOC	0.479	9.950	8.062	0.385	10.583	8.420
	TN	0.341	7.334	6.045	0.239	7.874	6.259
GBRT	SOC	0.936	3.485	2.553	0.773	6.421	4.319
	TN	0.901	2.848	1.891	0.645	5.381	3.225
MLP	SOC	0.972	2.290	1.349	0.736	6.891	4.185
	TN	0.888	3.015	1.811	0.512	6.077	3.556

9 种机器学习模型拟合程度对比发现,基于树的集成学习方法整体表现优于线性回归与核方法。其中,XGBoost、GBRT 和 RF 在 SOC 与 TN 的估算中均表现出较低的 RMSE 和 MAE,同时具有较高的  $R^2$ ,在 9 个候选模型中表现最为稳健。相比之下,线性回归与 SVR 的拟合精度明显偏低,而 MLP 虽在部分指标上接近优选模型,但稳定性略显不足。

图 4 进一步展示 XGBoost、GBRT 与 RF 3 种优选模型在训练集和测试集上的观测值与预测值、拟合线及置信带,用于直观呈现模型拟合形态与泛化表现。理想状态下,散点应沿  $y=x$  参考线分布,且置信带越窄表示预测不确定性越小。结果表明,在 SOC 与 TN 2 类目标变量上,3 种模型的测试集点云分布相对紧凑,拟合线与  $y=x$  参考趋势一致性较高且置信带较窄,显示出良好的拟合一致性与泛化能力。结合表 3 的交叉验证量化结果 RMSE、MAE 与  $R^2$ ,因而最终确定 XGBoost、GBRT 与 RF 为后续 SOC 与 TN 估算及融合建模的核心基学习器。



注:图 a、图 b、图 c 分别为 XGBoost、GBRT 和 Random Forest 在 SOC 上的预测值与实际值对比;图 d、图 e、图 f 分别为 XGBoost、GBRT 和 Random Forest 在 TN 上的预测值与实际值对比。

图 4 机器学习模型 SOC、TN 预测结果与实际值对比

Fig.4 Comparison between predicted and observed values of SOC and TN by machine learning models

2.2.2 机器学习模型变量重要程度 对入选的 3 种最优模型(XGBoost、GBRT、RF)进行后验解释。按各因子的重要性排序并结合蜂群图展示各因子具体影响程度(图 5)。3 种模型在 SOC 上呈高度一致的变量结构:BD(土壤体积质量)与  $T_{min}$ (年最低气温)贡献居前且总体表现为负向影响,Pre(降水量)与 NDVI(植被指数)稳定呈正向影响,SILT、CLAY(细粒组分)、CEC(阳离子交换量)与 POR(孔隙度)多为正向贡献,pH 多为负向贡献。地形因子(SLP、TRI、ASP、LNUM1)贡献度居中,方向随局地条件略有差异,模型间仅存在细微差异:XGBoost、

GBRT 对 Pre 与 NDVI 的灵敏度相对更高,RF 对粒度与地形的权重相对靠前;在 TN 上 3 个模型呈现出 BD(土壤体积质量)与  $T_{min}$ (年最低气温)贡献居前且总体为负向影响;其后依次为 CEC(阳离子交换量)、Pre(降水量)、pH、NDVI(植被指数)及细粒组分(CL/SILT),其中 Pre、NDVI、CEC、CL/SILT 多表现为正向作用,pH 为负向作用。地形因子(SLP、ASP、PDEP1、LDEP1、TRI)贡献度居中。XGBoost、GBRT 对 Pre、NDVI 的敏感性相对更高;RF 在粒度与地形上的权重略靠前,但三者对主导因子的识别基本一致。

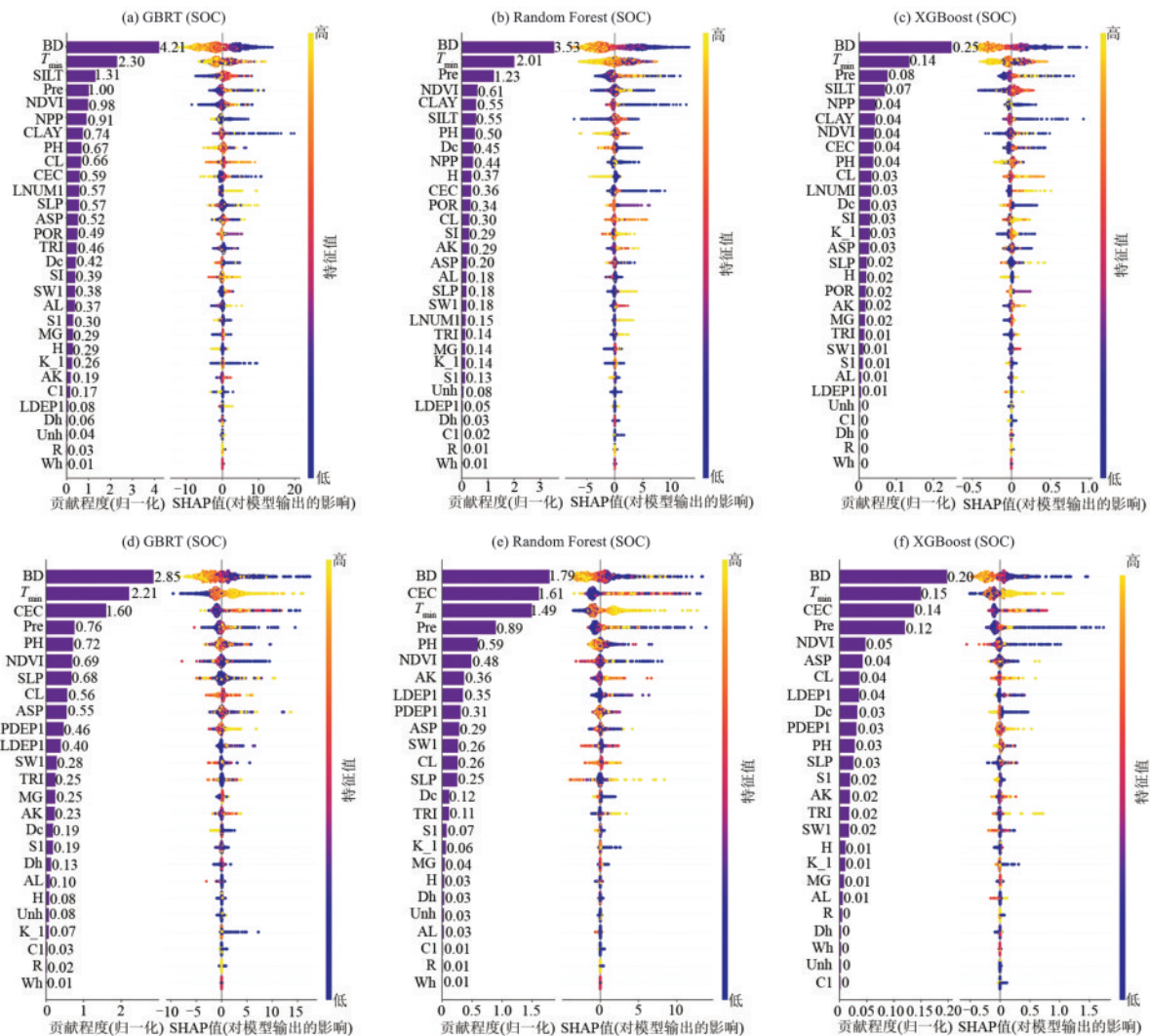


图 5 最优机器学习模型因子贡献程度及 SHAP 值

Fig.5 Factor contributions and SHAP values of optimal machine learning model

2.2.3 模型融合及 SOC、TN 空间估算 在与单模型相同的数据划分与预处理条件下,3 种融合策略在测集上的精度整体优于单模型(图 6)。在 SOC 预测中 Auto-Weighted 的综合表现最优,测试集的拟合优度  $R^2_{test}$  为 0.887 9,误差指标 RMSE 为 4.607 9、MAE 为 2.948 2。Blending 的测试集预测表现次之,  $R^2_{test}$  为 0.848 7, RMSE 为 5.107 2, MAE 为 2.991 9。

Bagging 的测试集表现最低,  $R^2_{test}$  为 0.770 1, RMSE 为 6.599 3, MAE 为 4.343 8。TN 测试集预测中,Auto-Weighted 的表现仍然最好,  $R^2_{test}$  为 0.775 8, RMSE 为 4.330 8, MAE 为 2.418 6。Blending 的测试集表现次之,  $R^2_{test}$  为 0.734 6, RMSE 为 4.576 2, MAE 为 2.337 3。Bagging 的测试集预测最差,  $R^2_{test}$  为 0.655 6, RMSE 为 5.367 7, MAE 为 3.157 6。

总体来看,3种融合方法均实现对SOC与TN预测性能的提升。综合比较拟合优度 $R^2_{test}$ 及误差指标RMSE与MAE,本文最终选取Auto-Weighted作

为SOC与TN的最优融合方案。基于Auto-Weighted模型对张掖地区SOC与TN进行估算,并绘制其空间分布图(图7)。

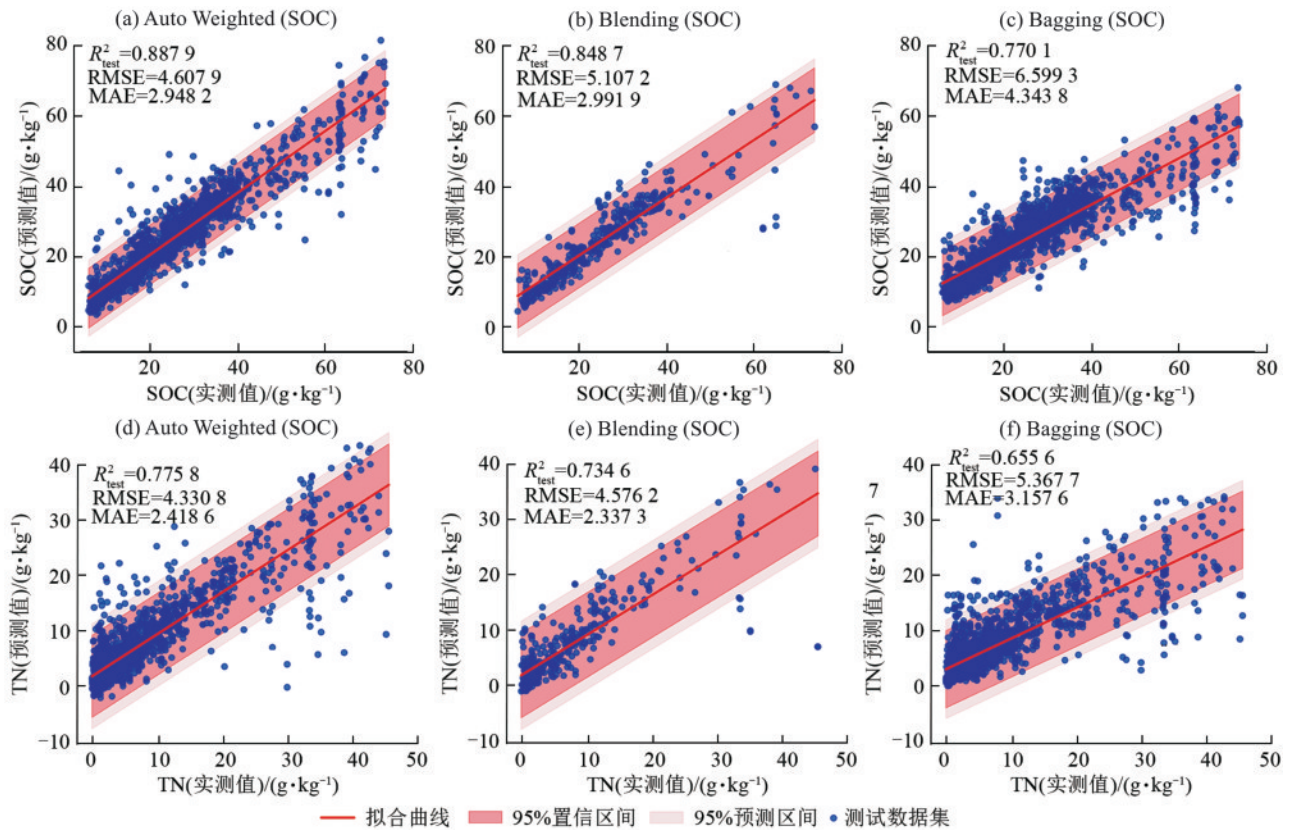


图6 融合模型预测精度对比

Fig.6 Comparison of prediction accuracy among ensemble models

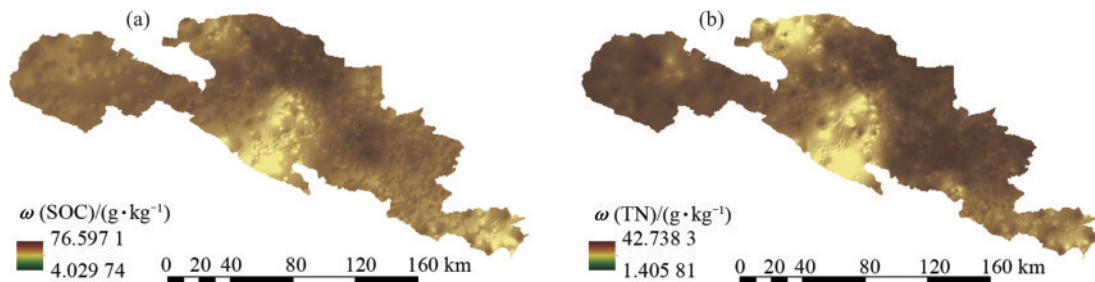


图7 Auto-Weighted 融合模型对SOC与TN的空间分布估算

Fig.7 Spatial distribution of SOC and TN estimated by Auto-Weighted ensemble model

### 3 讨论

本研究表明,以决策树为基学习器的集成学习方法XGBoost、GBRT和RF在SOC与TN的估算中整体优于线性回归(LR)与支持向量回归(SVR,核方法),而MLP的精度虽在部分指标上接近优选模型,但稳定性相对不足,与曹脊翔等<sup>[9]</sup>关于集成算法优于传统机器学习算法的结论一致。其算法层面的原因在于,张掖地区0~30 cm土层SOC、TN与气候、地形、植被及土壤属性等环境因子之间往往存在显著的非线性关系。线性模型在函数形式设定方面天然受限;而树模型能够通过递归分裂实现分段非线性逼近,无需预

设响应函数即可捕捉复杂关系。进一步地,RF通过Bagging与特征随机抽样降低方差、提升抗噪稳健性,GBRT、XGBoost通过逐步拟合残差并结合学习率、子采样与正则化机制在降低偏差的同时抑制过拟合,因此更适用于本研究的SOC、TN估算场景。

在此基础上,引入Auto-Weighted融合策略后,SOC的测试集性能能达到 $R^2_{test}=0.888$ , $RMSE=4.61$ , $MAE=2.95$ ;TN的测试集性能能达到 $R^2_{test}=0.776$ , $RMSE=4.33$ , $MAE=2.42$ ,体现出稳定增益。表明多基学习器之间的互补性能能够在一定程度上降低单一模型的偏差与方差,并提升测试集上的泛化

能力<sup>[10]</sup>。该结果与廖燕等<sup>[11]</sup>集成优于单模的研究结论相吻合;相比之下,简单算术平均容易受到表现较弱子模型拖累,传统加权在测试集泛化有可能不足,而更具自适应性的融合策略,也进一步提升精度与稳定性<sup>[12]</sup>。类似地,王宵宇等<sup>[13]</sup>在页岩总有机碳含量评价中指出,改良的集成框架相较常见集成策略在 $R^2$ 、MSE与MAE等指标上更具优势,其中“组合策略”的设计是决定集成效果的关键,此观点与本研究“以多基模型为基础、通过自适应权重实现融合增益”的思路一致。

基于SHAP的主控因子分析显示,SOC与TN受降水量(Pre)、温度( $T_{min}$ )、植被(NDVI)、土壤理化性质与地形因子的共同控制。Pre与植被状况对SOC、TN的正向贡献主要体现在SHAP重要性排序及依赖关系中,揭示在张掖干旱与半干旱背景下,水分供给与植被生产力是限制有机质输入的重要环节,适宜降水可促进植被生长,从而增加凋落物与根系周转输入,并可能通过改善土壤团聚体结构与有机质稳定化过程促进SOC与TN的积累<sup>[14]</sup>。同时,降水对SOC、TN的影响具有情景依赖性,在强降水事件或坡地产流增强条件下,径流侵蚀与淋溶迁移可能导致SOC与TN的再分配甚至净流失,因此其效应并非在所有情景下均为单向正效应。年平均最低气温( $T_{min}$ )的负向贡献表明,当夜间温度整体升高时,土壤微生物代谢与有机质分解的基线水平可能提高,从而增强碳氮矿化并降低SOC、TN的净积累; $T_{min}$ 升高也可能通过提高蒸散需求、降低有效水分供给间接削弱植被碳输入与碳氮汇聚过程。较高的CEC与合理的CLAY、SILT比例及适宜的POR有助于改善保水与结构稳定性,从而促进SOC和TN的积累与固持<sup>[15]</sup>;pH与BD的负向影响可能与通气、保水条件变化以及微生物活性相关。坡度、地形粗糙度与坡向通过影响径流路径与物质再分配,塑造“汇-散”格局并调节碳氮空间异质性<sup>[16]</sup>。总体而言,可解释的集成建模与SHAP框架与陈晓文等<sup>[17]</sup>、唐湘博等<sup>[18]</sup>和宫航等<sup>[19]</sup>关于表层土壤性质贡献排序、阈值响应与交互效应的发现具有一致性,说明在复杂地表过程背景下,集成模型结合全局与局部解释工具可同时提供较高预测精度与可用的机理线索。

区域空间格局方面,SOC与TN均呈显著空间异质性。在降水与温度作为主要气候约束、植被与地形共同调节的背景下,水分供给相对充足、温度较低且植被状况良好的山地更易形成碳氮“热点”<sup>[20]</sup>;向绿洲边缘与荒漠过渡区则因水分供给与植被生产力双重受限而形成“冷点”<sup>[21]</sup>。此外,SOC与TN的空间分布

并非完全独立变化,二者在形成与周转过程中存在相互影响。TN的变化可通过影响植被生产力与有机质输入间接调控SOC<sup>[22]</sup>,在水分供给相对较好的山地或绿洲边缘局部地段,若氮素可利用性较高,植被生长与根系周转增强,凋落物与根系输入随之增加,SOC更易积累<sup>[23]</sup>,相反,在水分与养分同时受限的荒漠过渡区,植被输入不足会使SOC与TN更易同步偏低,表现为“冷点”;另一方面,SOC含量与稳定程度也反过来影响TN的保存与形态转化,SOC增加通常意味着有机质底物更充足,微生物同化过程可能增强并促进氮素暂时固定<sup>[24]</sup>;同时,当土壤质地(CLAY、SILT)、CEC与孔隙结构(POR)更有利于团聚体形成与有机质保护时,有机氮不易被快速矿化或迁移<sup>[25]</sup>,从而推动TN在空间上与SOC呈现更高的一致性。这种相互影响并不意味着SOC与TN处处同向,在坡面径流、侵蚀、沉积与淋溶过程较强的地貌单元,碳氮的侧向迁移与再分配可能造成局部SOC与TN的空间响应差异<sup>[26]</sup>,从而进一步强化空间异质性。

在SOC与TN估算中,相较于Blending与Bagging,Auto-Weighted融合策略取得更高的预测精度和更稳定的泛化表现。其优势在于依据机器学习器的误差结构与互补性进行自适应赋权<sup>[27]</sup>,从而在水热差异显著、地形起伏明显、植被与土壤属性空间变化较大的背景下,能够随环境梯度的改变动态调整权重并保持较稳定的外推表现。相对而言,Bagging主要降低方差、对共同偏差的修正作用有限<sup>[28]</sup>,Blending的固定或半固定权重难以及时响应分布变化,解释了三者在本区的表现差异。总体上,合理的融合策略不仅提升SOC与TN的总体估算精度与稳健性,也在空间上减弱高值与低值区域的系统误差,降低由非线性与变量之间交互引发的边缘失真<sup>[29]</sup>。

## 4 结论

1)张掖地区0~30 cm土层SOC与TN的关键驱动因子在以决策树为机器学习器的集成模型中高度一致,表明识别结果具有较强稳健性。总体上,土壤结构状态与温度因子构成最强约束,SOC的首要贡献因子为土壤体积质量(BD),其后依次为年平均最低气温( $T_{min}$ )与降水量(Pre);TN亦主要受BD、 $T_{min}$ 、阳离子交换量(CEC)与Pre共同控制。方向性上,适宜的降水与较高植被指标总体呈正效应,而BD与 $T_{min}$ 在高值区间更偏负效应,提示在干旱与半干旱过渡带,SOC与TN的空间格局主要由水分和植被碳输入的驱动,同时受到温度、土壤结构状态的调制。

2)在模型表现方面,以决策树为机器学习器的集成学习方法(XGBoost、GBRT、RF)在SOC与TN估

算中整体优于线性回归(LR)与支持向量回归(SVR);多层感知机(MLP)在部分评价指标上接近优选模型,但整体稳定性相对不足。进一步地,在多模型融合策略中,自适应加权融合(Auto-Weighted)取得最佳综合性能(SOC: $R^2_{\text{test}}=0.8879$ ,RMSE=4.6079,MAE=2.9482;TN: $R^2_{\text{test}}=0.7758$ ,RMSE=4.3308,MAE=2.4186),体现出更好的泛化能力与稳健性。

3)本研究的不足之处,时间尺度错配与滞后效应,SOC与TN为“慢变量”,对近年尺度气候与植被指标(如Pre、 $T_{\min}$ 、 $T_{\max}$ 、NDVI)可能存在滞后响应;样本的空间自相关可能导致训练与测试划分出现空间泄露,从而高估泛化性能;放牧、灌溉、施肥与土地利用史等管理因子以及微地貌的量化不足,残差中仍可能包含未观测异质性。

4)后续研究可以在验证上采用块状或缓冲区的空间交叉验证,配合外部样点复核与跨区迁移检验;引入自助抽样或蒙特卡洛输入扰动,给出预测区间并系统表征不确定性。补充管理与干预变量及其代理量,细化微地形指标;构建时滞与多尺度特征,识别阈值与关键交互区间。

#### 参考文献:

- [1] WADOUX A M J C, MINASNY B, MCBRATNEY A B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions [J]. *Earth-Science Reviews*, 2020, 210: e103359.
- [2] POGGIO L, DE SOUSA L M, BATJES N H, et al. Soil-Grids 2.0: Producing soil information for the globe with quantified spatial uncertainty [J]. *Soil*, 2021, 7(1): 217-240.
- [3] YU W P, ZHOU W, WANG T, et al. Significant improvement in soil organic carbon estimation using data-driven machine learning based on habitat patches [J]. *Remote Sensing*, 2024, 16(4): e688.
- [4] 闫坤, 杨慧敏, 王德彩. 基于NDVI时间序列特征的森林土壤有机碳数字制图: 以济源南山林场为例 [J]. *土壤通报*, 2024, 55(4): 921-931.  
YAN K, YANG H M, WANG D C. Digital mapping of forest soil organic carbon based on NDVI time series features: A case study of Jiyuan Nanshan forest farm [J]. *Chinese Journal of Soil Science*, 2024, 55(4): 921-931.
- [5] RUNNING S W, NEMANI R R, HEINSCH F A, et al. A continuous satellite-derived measure of global terrestrial primary production [J]. *BioScience*, 2004, 54(6): 547-560.
- [6] WERE K, BUI T D, DICK B Ø, et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape [J]. *Ecological Indicators*, 2015, 52: 394-403.
- [7] 周洋, 赵小敏, 郭熙. 基于多源辅助变量和随机森林模型的表层土壤全氮分布预测 [J]. *土壤学报*, 2022, 59(2): 451-460.  
ZHOU Y, ZHAO X M, GUO X. Prediction of total nitrogen distribution in surface soil based on multi-source auxiliary variables and random forest approach [J]. *Acta Pedologica Sinica*, 2022, 59(2): 451-460.
- [8] 刘尊方, 雷浩川, 盛海彦. 基于XGBoost模型的湟水流域耕地土壤养分遥感反演 [J]. *干旱区地理*, 2023, 46(10): 1643-1653.  
LIU Z F, LEI H C, SHENG H Y. Remote sensing inversion of soil nutrient on farmland in Huangshui River basin based on XGBoost model [J]. *Arid Land Geography*, 2023, 46(10): 1643-1653.
- [9] 曹脊翔, 陈思源, 肖柏夷, 等. 基于机器学习的致密砂岩储层成岩相测井识别: 以川中地区沙溪庙组一段为例 [J]. *科学技术与工程*, 2025, 25(21): 8858-8870.  
CAO J X, CHEN S Y, XIAO B Y, et al. Machine learning based diagenetic facies logging identification: A case of Shaximiao Formation in central Sichuan basin [J]. *Science Technology and Engineering*, 2025, 25(21): 8858-8870.
- [10] 王智宇, 王达, 邱昆峰, 等. 机器学习在矿物岩石地球化学大数据挖掘中的应用与展望 [J]. *成都理工大学学报(自然科学版)*, 2025, 52(5): 844-858.  
WANG Z Y, WANG D, QIU K F, et al. Applications and perspectives of machine learning in geochemical big data mining of minerals and rocks [J]. *Journal of Chengdu University of Technology (Science and Technology Edition)*, 2025, 52(5): 844-858.
- [11] 廖燕, 闫建平, 廖茂杰, 等. 融合传统机器学习与深度学习的页岩气储集层岩相预测新方法: 以四川盆地寒武统筇竹寺组为例 [J]. *古地理学报*, 2026, 28(1): 353-368.  
LIAO Y, YAN J P, LIAO M J, et al. A novel method for lithofacies prediction in shale gas reservoirs that integrates traditional machine learning and deep learning techniques: A case study of the Lower Cambrian Qiongzhusi Formation in Sichuan basin [J]. *Journal of Palaeogeography*, 2026, 28(1): 353-368.
- [12] 田丽洁, 刘秋红. 基于深度学习的计算机网络流量预测模型研究 [J]. *河北能源职业技术学院学报*, 2025, 25(3): 60-63.  
TIAN L J, LIU Q H. Research on computer network traffic prediction model based on deep learning [J]. *Journal of Hebei Energy Institute of Vocation and Technology*, 2025, 25(3): 60-63.
- [13] 王宵宇, 廖广志, 黄文松, 等. 基于机器学习的页岩总有机碳含量评价方法 [J]. *石油科学通报*, 2025, 10(2): 392-403.  
WANG X Y, LIAO G Z, HUANG W S, et al. Evaluation method of total organic carbon content in shale based on machine learning [J]. *Petroleum Science Bulletin*, 2025, 10(2): 392-403.
- [14] 崔雨微, 李森, 常莹莹, 等. 多情景绿洲化对河西内陆河流域碳储量的影响 [J]. *环境科学与技术*, 2025, 48(9): 221-236.  
CUI Y W, LI S, CHANG Y Y, et al. Research on impacts of multi-scenario oasisification on carbon reserves

- in Hexi inland river basin [J]. *Environmental Science and Technology*, 2025, 48(9): 221-236.
- [15] 孙家敬, 贺玉晓, 陈雅兰, 等. 生物质炭的土壤固碳效应研究进展及展望 [J]. *土壤学报*, 2026, 63(1): 53-64.  
SUN J J, HE Y X, CHEN Y L, et al. Research progress and prospects of soil carbon sequestration effects of biochar [J]. *Acta Pedologica Sinica*, 2026, 63(1): 53-64.
- [16] 崔欣然, 张春来, 李庆, 等. 京津风沙源治理工程区土壤有机碳密度分布特征 [J]. *水土保持学报*, 2025, 39(6): 281-289.  
CUI X R, ZHANG C L, LI Q, et al. Distribution characteristics of soil organic carbon density in Beijing-Tianjin sandstorm source control project area [J]. *Journal of Soil and Water Conservation*, 2025, 39(6): 281-289.
- [17] 陈晓文, 苟佳常, 刘昌永, 等. 地貌敏感区土壤保持的长期演变及其气候时空响应机制: 以山西省为例 [J]. *湖北农业科学*, 2025, 64(9): 27-35.  
CHEN X W, XUN J C, LIU C Y, et al. Long-term evolution of soil conservation and its climatic spatio-temporal response mechanism in geomorphologically sensitive areas: A case of Shanxi Province [J]. *Hubei Agricultural Sciences*, 2025, 64(9): 27-35.
- [18] 唐湘博, 黄有为, 苏涵. 基于机器学习的中国农业净碳汇预测模型构建及驱动因素响应分析 [J/OL]. *环境科学*, 2025. <https://doi.org/10.13227/j.hjlx.202504072>.  
TANG X B, HUANG Y W, SU H. Construction and driving factors analysis of a machine learning-based prediction model for net carbon sink in Chinese agriculture [J/OL]. *Environmental Science*, 2025. <https://doi.org/10.13227/j.hjlx.202504072>
- [19] 宫航, 丁国栋, 刘昊, 等. 西藏地区 2000—2022 年植被覆盖度时空变化格局及驱动力分析 [J/OL]. *环境科学*, 2025. <https://doi.org/10.13227/j.hjlx.202505039>.  
GONG H, DING G D, LIU H, et al. Spatiotemporal patterns and driving factors of vegetation coverage in Tibet from 2000—2022 [J/OL]. *Environmental Science*, 2025. <https://doi.org/10.13227/j.hjlx.202505039>.
- [20] 王瑞, 汤占军. 基于多特征融合与集成学习的风机叶片缺陷检测方法 [J]. *计算机科学*, 2025, 52(S1): 458-465.  
WANG R, TANG Z J. Multi-feature fusion and ensemble learning-based wind turbine blade defect detection method [J]. *Computer Science*, 2025, 52(S1): 458-465.
- [21] 师垒, 申艳军, 彭建兵, 等. 山地土壤养分的双向补给特征及演化过程 [J]. *地球科学与环境学报*, 2025, 47(4): 780-793.  
SHI B L, SHEN Y J, PENG J B, et al. Bidirectional replenishment characteristics and evolution process of soil nutrients in mountainous areas [J]. *Journal of Earth Sciences and Environment*, 2025, 47(4): 780-793.
- [22] 刁二龙, 曹广超, 曹生奎, 等. 祁连山南坡不同土地利用方式下土壤碳氮含量及通径分析 [J]. *干旱区研究*, 2021, 38(5): 1346-1354.  
DIAO E L, CAO G C, CAO S K, et al. Soil carbon and nitrogen content and path analysis under different land use patterns on the southern slope of Qilian Mountains [J]. *Arid Zone Research*, 2021, 38(5): 1346-1354.
- [23] 童珊, 曹广超, 张卓, 等. 土壤全碳全氮空间异质性及影响因素分析: 以祁连山南坡黑河上游为例 [J]. *土壤*, 2022, 54(5): 1064-1072.  
TONG S, CAO G C, ZHANG Z, et al. Analysis of spatial heterogeneity and influencing factors of soil total carbon and nitrogen: Take the upper reaches of Heihe River on southern slope of Qilian Mountain as an example [J]. *Soils*, 2022, 54(5): 1064-1072.
- [24] 宋文婕, 梁誉正, 陶贞, 等. 微生物介导的土壤有机碳动态研究进展 [J]. *地球科学进展*, 2023, 38(12): 1213-1223.  
SONG W J, LIANG Y Z, TAO Z, et al. Advances on soil organic carbon dynamics mediated by microorganisms [J]. *Advances in Earth Science*, 2023, 38(12): 1213-1223.
- [25] 邓蕾, 李继伟, 瞿晴, 等. 退牧还草地土壤有机碳固持动态与驱动机制研究进展 [J]. *应用生态学报*, 2024, 35(11): 3208-3216.  
DENG L, LI J W, QU Q, et al. Dynamics and driving mechanisms of soil organic carbon sequestration in grasslands after grazing exclusion: A review [J]. *Chinese Journal of Applied Ecology*, 2024, 35(11): 3208-3216.
- [26] 齐瑜洁, 黄金权, 李威闻, 等. 基于水力侵蚀过程的土壤有机碳变化驱动机制研究进展 [J]. *土壤通报*, 2023, 54(5): 1196-1204.  
QI Y J, HUANG J Q, LI W W, et al. Research progress on driving mechanism of soil organic carbon change based on hydraulic erosion process [J]. *Chinese Journal of Soil Science*, 2023, 54(5): 1196-1204.
- [27] 张鹏, 杜英慧, 周博凡, 等. 融合特征选择与梯度提升决策树的典型侵财犯罪风险预测方法 [J]. *中国人民公安大学学报(自然科学版)*, 2025, 31(3): 38-50.  
ZHANG P, DU Y H, ZHOU B F, et al. Risk prediction method of typical property crime based on feature selection method and gradient boosting decision tree [J]. *Journal of People's Public Security University of China (Science and Technology)*, 2025, 31(3): 38-50.
- [28] 郑晓楠, 于洋, 潘虹, 等. 基于 Stacking 集成学习的水电机组负荷分配 [J]. *排灌机械工程学报*, 2025, 43(10): 1040-1048.  
ZHENG X N, YU Y, PAN H, et al. Load distribution for hydroelectric units based on Stacking ensemble learning [J]. *Journal of Drainage and Irrigation Machinery Engineering*, 2025, 43(10): 1040-1048.
- [29] 曾铃, 谢宇航, 章赛泽, 等. 基于堆叠集成算法的软岩填方路基沉降融合预测模型研究 [J]. *湖南大学学报(自然科学版)*, 2025, 52(9): 125-138.  
ZENG L, XIE Y H, ZHANG S Z, et al. Research on fusion prediction model of soft rock embankment subsidence based on stacked generalization integration algorithm [J]. *Journal of Hunan University (Natural Sciences)*, 2025, 52(9): 125-138.