DOI: 10.13758/j.cnki.tr.2024.04.020
刘雅璇, 于慧, 罗勇. 基于辅助变量的紫色土耕地土壤有机质空间预测. 土壤, 2024, 56(4): 857–865.

基于辅助变量的紫色土耕地土壤有机质空间预测①

刘雅璇[1,2], 于 慧[1*], 罗 勇[3]

(1 中国科学院·水利部成都山地灾害与环境研究所, 成都 610299; 2 昆明理工大学国土资源工程学院, 昆明 650093; 3 成都理工大学地理与规划学院, 成都 610059)

摘 要: 研究收集了川中丘陵区域紫色土耕地共 135 个土壤样本, 基于 GEE(Google Earth Engine)云平台调用高分辨率 Sentinel-2A 数据、SRTMGL1v3.0 高程数据、SoilGrids 土壤属性数据, 并创新性地加入了纹理特征作为辅助变量, 分别通过梯度提升决策树(GBDT)和随机森林(RF)构建两种预测模型反演研究区土壤有机质。结果表明: 研究区内紫色土耕地土壤有机质含量偏低, 养分级别为二 ~ 六级; GBDT 算法构建的模型相比于 RF 算法预测精度更高, R^2、r、RMSE 分别为 0.687、0.829、5.668 g/kg 和 0.514、0.717、6.765 g/kg; 加入纹理特征的模型 R^2 分别增加了 6.80% 和 1.70%, 为土壤有机质预测研究提供了新的思路。

关键词: 土壤有机质; 机器学习; 紫色土; GEE

中图分类号: S127 **文献标志码:** A

Soil Organic Matter Prediction of Purple Soil Based on Auxiliary Variables

LIU Yaxuan[1,2], YU Hui[1*], LUO Yong[3]

(1 *Institute of Mountain Hazards and Environment* (*IMHE*), *Chinese Academy of Sciences*, *Chengdu* 610299, *China*; 2 *Faculty of Land Resources Engineering*, *Kunming University of Science and Technology*, *Kunming* 650093, *China*; 3 *Faculty of Geography and Planning*, *Chengdu University of Technology*, *Chengdu* 610059, *China*)

Abstract: This study collected a total of 135 samples from purple soil farmlands in the hilly region of central Sichuan. Based on the GEE cloud platform, high-resolution Sentinel-2A data, SRTMGL1v3.0 elevation data, and SoilGrids soil attribute data were invoked, and texture feature data was innovatively added. Two prediction models were constructed by using gradient enhancement decision tree (GBDT) and random forest (RF) to invert SOM. The results showed that SOM content of purple soil farmlands in the study area was relatively low, with the level ranging from 2 to 6 levels. The models constructed by GBDT algorithm had higher prediction accuracy (R^2=0.687, r=0.829, RMSE=5.668 g/kg) compared to RF algorithm (R^2=0.514, r=0.717, RMSE=6.765 g/kg). The R^2 with texture features increased by 6.80% and 1.70%, respectively. TGIS study can provide a new scientific approach for SOM prediction.

Key words: Soil organic matter; Machine learning; Purple soil; GEE

紫色土是我国一种重要且独特的土壤资源。紫色土具有土壤养分丰富、耕性好、自然肥力高、生产力高等特点, 因此盛产各种粮食作物和经济作物[1]。据全国第二次土壤普查(1979—1994 年), 含重庆在内的原四川省是我国紫色土分布面积最广的省份, 紫色土面积约为全国的 51%, 土壤生产力高, 用作耕地适宜各种农作物种植, 是四川省重要的一类农业土壤资源[2]。然而, 紫色土同时也是土层薄、易退化流失的土壤。紫色土区域属于亚热带, 气候温暖湿润, 物质循环强烈[3], 加之长期不合理的土壤管理, 紫色土退化问题突出[2]。因此及时了解和监测紫色土土壤质量有助于紫色土管理和耕地管理, 为农业发展提供基础参考。耕地紫色土质量与土壤属性等有关, 土壤有机质是土壤质量的重要组成部分, 是土壤普查中衡量土壤质量的核心指标之一。据第二次土壤普查报告, 四川省紫色土有机质含量整体较低[2]。李韦亨等[4]测得川中丘陵区紫色土有机质含量平均为 10.25 g/kg, 与其他典型土壤类型东北黑土区[5]相比有

①基金项目: 国家自然科学基金项目(41971273)和四川省地质调查研究院财政资金项目(SCIGS-CZDXM-2024014)资助。

* 通讯作者(yuhui05@126.com)

作者简介: 刘雅璇(1999—), 女, 四川绵阳人, 硕士研究生, 主要研究方向为土壤数字制图。E-mail: liuyaxuan@stu.kust.edu.cnm

机质含量较少。鉴于紫色土土壤有机质含量较低且易流失，监测土壤有机质含量对了解土壤质量和土壤肥力具有重要意义。

耕地土壤有机质预测研究在不断进步和完善，主要的预测方式包括传统地统计学方法和基于遥感技术的土壤有机质预测。地统计学方法包括地理加权回归[6]、普通克里格[7]、偏最小二乘回归[8]、指数递减函数拟合[9]等。遥感卫星的应用发展为基于遥感技术的土壤有机质预测提供了新数据和新思路。基于实测土壤有机质数据以及遥感影像波段数据可进行土壤有机质制图，目前国内专家已从不同方面取得了进展。早期刘焕军等[10]基于 Landsat TM 遥感影像的红、绿、近红外波段反演黑土区土壤有机质，并借用光谱指数提高了土壤有机质反演精度。对于田块小尺度的东北黑土区[11]和较大范围的亚热带区域水稻土[12]等典型土壤类型均开展了土壤有机质空间预测。也有研究从时空格局方面进行分析，如从长时序和三维空间角度研究土壤有机碳或土壤有机质[13-14]。模型变量的加入也可提高模型预测的准确度，同时变量选择需视区域实际情况来定。郭静等[15]研究了红边波段对模型预测能力的影响，表明红边波段有效提高了土壤有机质的预测精度；优化光谱输入量相较于直接使用光谱反射率建模的预测精度更高[16-17]；也有学者选取特殊特征波长建立预测效果更好的模型[18]；农业活动因子的轮作模式的加入也可提高耕地土壤有机碳预测的精度[19]。预测模型的选择与对比也是土壤有机质预测的重要研究方面，近年来学者们应用较为广泛的模型包括随机森林算法(Random forest，RF)、支持向量机(Support vector machine，SVM)、决策树(Decision tree，DT)以及梯度提升决策树(Gradient boosting decision tree，GBDT)等，其中多个研究表明 GBDT 和 RF 算法预测精度相对较高[15]。

综合来看，一方面近年来紫色土研究较少，构建基于遥感的紫色土耕地土壤有机质监测模型可为加强紫色土耕地管控和建设提供理论依据[20]；另一方面在目前已有研究中，缺乏纹理特征在土壤有机质预测研究中的探索，纹理特征常用于作物分类提取等[21-23]，而少用于土壤有机质预测。因此构建川中丘陵区紫色土有机质预测模型，探究纹理特征对土壤有机质预测的影响，可为基于遥感技术的土壤有机质预测提供新思路。本文基于 Google Earth Engine 云平台(https://earthengine.google.com/)，以 Sentinel-2A 影像数据为主要数据源，以光谱指数、地形因子、土壤属性以及纹理特征为辅助变量，使用 GBDT 和 RF 算法两种主要预测模型，构建川中丘陵区紫色土有机质预测模型，探究纹理特征对川中丘陵区紫色土有机质预测模型精度的影响，以期为紫色土有机质预测和空间制图提供新思路。

1 数据和方法

1.1 研究区概况

研究区地处四川省盐亭县，位于我国四川盆地中部，属于典型的紫色土丘陵区，位于 31°14′50″N ~ 31°16′30″N，105°24′40″E ~ 105°27′40″E。该区为亚热带湿润季风气候，气候温和，热量充沛，全年平均气温约 17 ℃。盐亭县以丘陵地貌为主，是典型的丘陵区农业大县，农业资源丰富。研究区内耕地类型主要为旱地、水田、水旱轮作等，种植作物主要有水稻、油菜、小麦、玉米等。水旱轮作地轮作方式主要为夏水稻–冬油菜，旱地作物轮作模式通常为冬小麦–夏玉米，水田种植作物为夏水稻。研究区高程范围约为 372 ~ 636 m，地势东高西低，起伏较大。通过 GlobeLand30 平台(http://www.globallandcover.com/)下载研究区 2020 年土地利用数据，通过耕地矢量数据裁剪出研究区耕地范围(图 1)，面积约 1 010 hm^2。

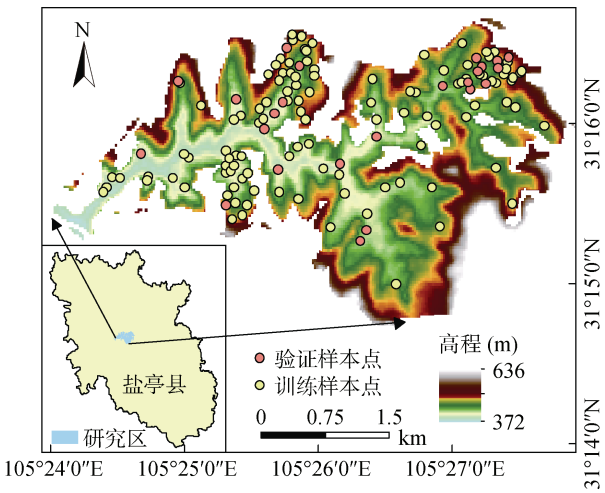

图 1 研究区位置图及地面样本示意图
Fig. 1 Location of study area and sampling sites

1.2 土壤样本实地采集及其有机质含量测定

2022 年 9 月为研究区裸土期，基本完成水稻、玉米等作物收割，于 9 月 26 日至 9 月 28 日采集耕地土壤样本共 135 个，包含水田土壤样本 12 个、水旱轮作土壤样本 35 个以及旱地土壤样本 88 个，并通过手持 GPS 记录采样点经纬度坐标。将土壤样品带回实验室，放置于阴凉通风处风干，保证不接受太阳直射。剔除杂质后采用四分法选取约 50 g 土样，研磨

过 0.25 mm 的塑料筛，通过精密称重仪称取约 0.2 g 样品，采用重铬酸钾容量法测定土壤有机质含量[24]。

1.3　预测模型所用变量数据

1.3.1　遥感波段及其光谱指数　　使用的 Sentinel-2A 数据由 Google Earth Engine 云平台(https://code.earthengine.google.com/)提供。哨兵 2 号携带一枚多光谱成像仪(MSI)，分为了 2A 和 2B 两颗卫星，高度为 786 km，幅宽达 290 km，共包含了 13 个光谱波段(443 ~ 2 190 nm)，空间分辨率包括 10、20 和 60 m。选取 3 个可见光波段(B2、B3、B4)，4 个红边波段(B5、B6、B7、B8A)，一个近红外波段(B8)以及两个短波红外波段(B11、B12)(表 1)。首先调用已经过辐射定标、几何校正和大气校正等预处理的 Sentinel-2A 影像数据“COPERNICUS/ S2_SR”，筛选拍摄日期与采样时间最相近的单期影像(9 月 28 日)；简单去云处理后，对于分辨率为 20 m 的红边波段和短波红外波段，重采样为 10 m 分辨率。

表 1　模型中的辅助变量
Table 1　Auxiliary variables in models

变量	波段/参数
光谱波段	B2 (蓝波段)、B3 (绿波段)、B4 (红波段)、B5 (红边波段 1)、B6 (红边波段 2)、B7 (红边波段 3)、B8A (红边波段 4)、B11 (短波红外波段 1)、B12 (短波红外波段 2)
光谱指数	NDVI、EVI、SATVI、RI
地形因子	DEM、SLOPE、ASPECT
土壤属性	pH、SAND、CLAY、CFVO、SILT
纹理特征	IMCORR1、SHADE

影像光谱指数可有效提高构建模型的精度。经多次试验，选取 NDVI(归一化植被指数)、EVI(增强植被指数)、RI(红度指数)以及 SATVI(土壤调整总植被指数)4 种光谱指数用于土壤有机质预测模型构建(表 1)，其定义及计算公式如表 2。NDVI、EVI 以及 SATVI 均可反映一定的植被信息和土壤信息；RI 则用于遥感图像处理，可分析土壤、植被等地表特征等。以下光谱指数的计算波段分别为 B2、B3、B4、B8、B11、B12。

1.3.2　地形因子　　SRTM 全名航天飞机雷达地形测绘任务，由美国航空航天局(NASA)和美国国防部国家测绘局(NIMA)以及德国和意大利的航天机构合作完成。在 GEE 云平台上传研究区矢量范围，调用 SRTMGL1v3.0 高程数据和裁剪函数得到研究区范围的 DEM 数据，通过 GEE 计算得到坡度和坡向。通过 GEE 云平台的最邻近重采样函数获得 10 m 分辨率的高程(DEM)、坡度(SLOPE)和坡向(ASPECT)，作为地形因子(表 1)参与模型构建。

表 2　Sentinel-2A 的光谱指数定义及公式
Table 2　Spectral index definitions and formulas of Sentinel-2A

光谱指数	定义	公式
NDVI	$\frac{NIR-Red}{NIR+Red}$	$\frac{B8-B4}{B8+B4}$
EVI	$\frac{2.5(NIR-Red)}{NIR-6Red-7.5Blue+1}$	$\frac{2.5(B8-B4)}{B8-6B4-7.5B2+1}$
RI	$\frac{Red\cdot Red}{Green\cdot Green\cdot Green}$	$\frac{B4\cdot B4}{B3\cdot B3\cdot B3}$
SATVI	$\frac{2(SWIR1-Red)}{SWIR1+Red+1}-\frac{SWIR2}{2}$	$\frac{2(B11-B4)}{B11-B4+1}-\frac{B12}{2}$

1.3.3　土壤属性　　SoilGrids 是基于机器学习算法和 230 000 个来自 WoSIS database(https://www.isric.org/explore/wosis)的土壤剖面数据绘制的全球数字土壤地图[25]，空间分辨率为 250 m。SoilGrids 包含土壤堆积密度、黏土含量、粗粒含量等土壤理化属性。如表 1 所示，通过 GEE 云平台选取土壤 pH、黏粒含量(CLAY)、砂粒含量(SAND)、粗粒含量(CFVO)、粉粒含量(SILT)，重采样为 10 m 分辨率，作为土壤理化性质参与模型构建。

1.3.4　纹理特征　　纹理特征是一种不依赖于颜色或亮度而反映图像中同质现象的视觉特征，包含事物表面的结构排列信息以及与周围事物的联系。纹理是基于灰度分布在空间位置上反复出现而形成的。遥感图像中纹理信息丰富，纹理特征可反映遥感影像中的结构信息。灰度共生矩阵(Gray-level co-occurrence matrix, GLCM)是提取纹理特征的常用方法，由计算机科学家 Haralick 等[26]于 1973 年提出。首先基于光谱波段 B3、B4 和 B8 计算出灰度图层，并调用.glcmTexture()函数计算灰度共生矩阵，得到包含 18 个纹理特征波段的影像。GEE 云平台的 18 个纹理特征包含了 Haralick 所提出的基于 GLCM 的 14 个指标，以及 Conners 等[27]于 1984 年提出的 4 个指标。加入变量过多会造成数据重叠和冗余，导致模型预测精度不增反减。因此经多次试验选取 IMCORR1 和 SHADE 两个指标作为纹理特征变量(表 1)加入模型。IMCORR1 为相关信息测度，可表示灰度图像中行方向或者列方向上的线性相关程度，值的大小代表局部灰度相关性。SHADE 为聚类荫，代表图像中的一致性特征，当一定规则映射到土壤表面时，表面纹理会呈现不同的属性。纹理特征具有旋转不变性，抗噪声能力较强，土壤中具有的粗细、疏密差异能够体现在

纹理特征方面。

1.4 回归模型方法与模型构建

1.4.1 RF 回归 随机森林(Random forest，RF)是一种通过集成学习的 Bagging 思想将多棵树集成的算法，它的基本单元是决策树[28]。随机森林的关键在于“随机”和“森林”，“森林”中的每一棵决策树都是一个分类器，N 棵树对于同一个样本会有 N 个分类结果，随机森林集成了所有的分类结果，并将出现次数最多的结果指定为最终输出。随机森林能够处理高维度的特征数据，具有良好的适用性和鲁棒性，目前已经广泛应用于各种分类和预测问题。在 GEE 中使用随机森林算法，直接调用 ee.Classifier.smileRandomForest()函数，需要设置的参数包括 ntree 和 Mtry，即生成树的数量和每个节点处用于分割节点的反演变量数，ntree 设置为 400。其他参数均可保持默认值。

根据 Bagging 思想，随机森林的实现过程如下：①从 N 个训练集原始样本中进行 M 次有放回的抽样试验，产生 M 个训练集 $T_1, T_2, \cdots T_m$；②每一个训练集对应一棵决策树 $f_1, f_2, \cdots f_m$，每一棵决策树各自分裂出最佳属性；③多棵决策树结果集成随机森林，最终结果为取决策树平均，公式如下：

$$f(x)=\frac{1}{M}\sum_{m=1}^{M}f_m(x) \tag{1}$$

1.4.2 GBDT 回归 梯度提升决策树(Gradient boosting decision tree，GBDT)由 Friedman[29]提出，通过集成学习的 Boosting 算法进行回归预测。GBDT 同样以决策树为基础的弱学习器，通过不断迭代与当前模型线性组合得到新模型，此过程中模型的残差在梯度方向上减少，当决策树数量达到指定的值，迭代停止并得到最终的强学习器。GBDT 算法具有强大的预测能力，但相较于 RF，在土壤有机质预测方面的应用相对较少。在 GEE 平台使用 GBDT 算法需要调参，ntree 设置为 30，shrinkage 设置为 0.07，其余参数默认，其相关计算公式如下：

$$f_0(x)=\arg\min\sum_{i=1}^{N}L(y_i,c) \tag{2}$$

$$f_0(x)=c \tag{3}$$

$$r_{mi}=\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{4}$$

$$c_{mj}=\arg\min\sum_{x_i\in R_{mj}}L(y_i,f_{m-1}(x_i)+c) \tag{5}$$

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}c_{mj}I(x\in R_{mj}) \tag{6}$$

$$f(x)=f_M(x)=f_0(x)+\sum_{m=1}^{M}\sum_{j=1}^{J}c_{mj}I(x\in R_{mj}) \tag{7}$$

首先是初始化弱学习器如式(2)，初始化时，c 取值为所有训练样本标签值的均值，y_i 为样本真实值，因此得到初始学习器如式(3)。然后通过迭代训练 m=1,2,3⋯,M 棵树，如式(4)对 i=1,2,3,⋯N 的每个样本计算残差，$f(x_i)$表示训练样本根据函数计算出的拟合值；根据式(5)计算 j=1,2,3⋯,J 个叶子结点的最佳拟合值，并更新强学习器如式(6)，最终学习器如式(7)。

1.4.3 模型构建 为构建紫色土土壤有机质评价体系，并探究不同方法的预测能力以及土壤属性、纹理特征对土壤有机质预测的影响，将遥感波段和光谱指数作为不变的数据构建基本模型，依次加入地形因子数据、土壤属性数据和纹理特征数据作为模型的可变数据，如表 3 所示。在 GEE 云平台上的模型构建过程如下：①选取接近采样日期的单张 Sentinel-2A 影像，提取 10 个遥感波段，并将光谱指数波段提取至上述包含 10 个遥感波段的影像中；②为验证模型精度以及预测结果与样本点的相关性，调用.randomColumn()函数把样本点随机分为训练数据集和验证数据集，分别占样本数量的 80% 和 20%。训练数据集加入模型训练得到预测结果，验证模型用于模型精度验证；③将包含遥感波段、光谱指数的波段加入训练数据集，然后分别加入地形因子、土壤属性和纹理特征数据，调用 ee.Classifier.smileGradient TreeBoost()函数和 ee.Classifier.smileRandomForest()函数进行回归预测，两种预测方法构建的模型分别为模型 GBDT 和模型 RF。为有效评估土壤有机质预测结果精度，采用决定系数 R^2、皮尔逊相关系数 r 和均方根误差 RMSE 评价模型精度，3 个指标通过 Origin 2018 计算得出。R^2 越大表明模型越稳定，r 越大表明预测值越接近真实值，RMSE 越小表明精度越高。

表 3 两种模型的构建方法

Table 3 Two methods for constructing models

模型	模型构建变量
GBDT-1	遥感波段+光谱指数
GBDT-2	遥感波段+光谱指数+地形因子
GBDT-3	遥感波段+光谱指数+地形因子+土壤属性
GBDT-4	遥感波段+光谱指数+地形因子+土壤属性+纹理特征
RF-1	遥感波段+光谱指数
RF-2	遥感波段+光谱指数+地形因子
RF-3	遥感波段+光谱指数+地形因子+土壤属性
RF-4	遥感波段+光谱指数+地形因子+土壤属性+纹理特征

2　结果与分析

2.1　研究区紫色土土壤有机质统计特征

对所有样本进行统计分析，研究区紫色土土壤有机质最小值为 2.68 g/kg，最大值为 33.02 g/kg，土壤有机质含量总体差异较大，平均值为 14.04 g/kg，中位数为 12.85 g/kg，与平均值相近，标准差为 7.56 g/kg。根据全国第二次土壤普查养分分级标准，研究区紫色土土壤有机质级别位于二 ~ 六级(图 2)。样本的变异系数(CV)为 54%，属于中等变异水平。研究区紫色土样本土壤有机质含量呈现出差异变化较大的特征，整体级别为中等偏下水平。

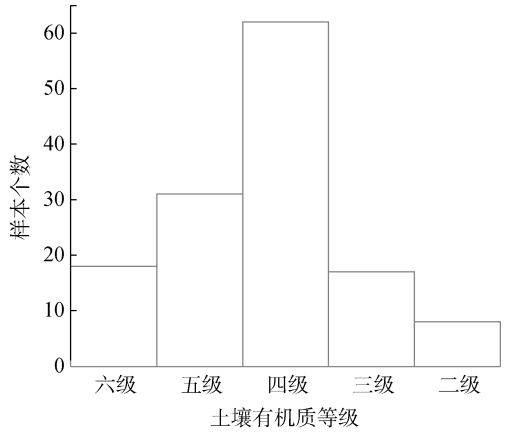

图 2　研究区土壤有机质统计特征
Fig. 2　Statistics of SOM in study area

2.2　土壤有机质预测结果的空间分布

图 3 为 GBDT 和 RF 构建的模型预测的研究区土壤有机质空间分布图。从整体来看，两种模型的预测结果相似，均为中部土壤有机质值较高，周围区域较低。GBDT 算法预测的土壤有机质含量最小值为 5.01 g/kg，最大值为 27.71 g/kg；RF 算法预测的土壤有机质含量最小值为 8.13 g/kg，最大值为 25.44 g/kg。根据所有样本统计土壤有机质含量为 2.68 ~ 33.02 g/kg，因此 GBDT 算法下所预测的土壤有机质含量区间相较于 RF 算法更接近研究区真实结果。

2.3　不同预测模型对比

为提高模型预测精度并探究不同变量对于模型预测的影响，研究采用 3 个评价指标对 6 种模型中的验证集进行评价，比较其结果与真实值之间的接近程度。如表 4 所示，从每个模型内部来看，在模型 GBDT-1 和 RF-1 的基础上加入地形因子后，R^2 分别增加了 17.30% 和 22.10%，预测精度明显提高。模型 GBDT-3 和 RF-3 中，土壤属性的加入使得 R^2 分别从 0.278 和 0.232 提高到 0.619 和 0.497，相关系数 r 也对应增大，RMSE 和 MAE 相应减小。加入纹理特征后，GBDT-4 和 RF-4 的 R^2 相对于模型 GBDT-3 和 RF-3 分别增加了 6.80% 和 1.70%，相关系数 r 由 0.787、0.705 提高到 0.829、0.717，RMSE 由 6.011、6.814 g/kg 减小到 5.668、6.765 g/kg，MAE 由 5.059、5.868 g/kg 减小到 4.757、5.800 g/kg。从模型 GBDT-1 和 RF-1 到 GBDT-4 和 RF-4，R^2 和 r 不断提高，RMSE 和 MAE 呈不断减少趋势，说明了辅助变量的加入有助于模型预测能力的提高。

从 GBDT 和 RF 算法模型预测结果来看，无论是否加入辅助变量，GBDT 算法所构建的模型的预测结果始终比 RF 算法的预测结果更好。GBDT 模型的预测结果区间比 RF 模型更接近于真实值区间，GBDT 模型对边缘的最大值和最小值预测越接近真实值。两种模型预测精度有差异的原因可能是 RF 模型在部分回归预测问题中容易过拟合，会受噪声点的影响；GBDT 模型对于异常值较敏感，对于靠近边缘数量较少的最大值和最小值预测精度较高。

因此总结来说，地形因子、土壤属性和纹理特征的加入使得两种预测模型的精度提高，GBDT 算法最终

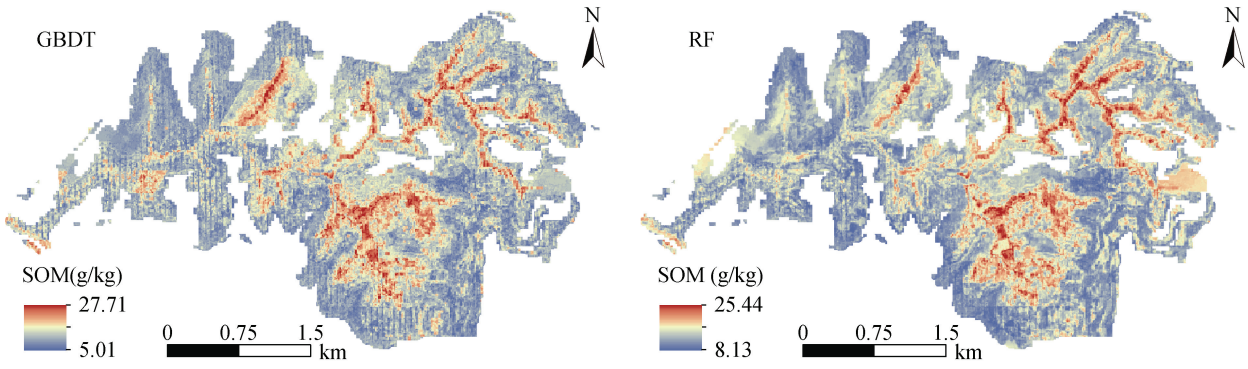

图 3　两种模型的土壤有机质预测结果
Fig. 3　SOM prediction results of two models

表 4 两种模型的土壤有机质预测精度
Table 4 SOM prediction accuracies of two models

指标	GBDT 模型				RF 模型			
	GBDT-1	GBDT-2	GBDT-3	GBDT-4	RF-1	RF-2	RF-3	RF-4
R^2	0.105	0.278	0.619	0.687	0.011	0.232	0.497	0.514
r	0.324	0.527	0.787	0.829	0.106	0.482	0.705	0.717
RMSE (g/kg)	8.050	7.277	6.011	5.668	8.672	7.501	6.814	6.765
MAE (g/kg)	6.758	5.936	5.059	4.757	7.108	6.155	5.868	5.800

构建的模型预测精度为 R^2=0.687，r=0.829，RMSE=5.668 g/kg，MAE=4.757 g/kg；RF 算法的预测结果为 R^2=0.514，r=0.717，RMSE=6.765 g/kg，MAE=5.800 g/kg。分别将模型训练集和验证集的结果与真实值比较，图 4 和图 5 为样本点的土壤有机质真实值与预测值所构成的散点图。从验证集来看，随着模型预测精度不断提高，从模型 GBDT-1 到模型 GBDT-3，样本点分布呈现出较分散到逐渐集聚的过程，样本趋势线偏离 1∶1 线的幅度明显变小；从模型 GBDT-3 到模型 GBDT-4，样本点集聚程度继续增加，趋势线偏离幅度继续变小。模型 RF-1 到模型 RF-4 的散点图同样表现为样本点不断集聚趋势，以及样本趋势线偏离距离变小。

2.4 变量的重要性分析

图 6 为 GBDT 和 RF 算法构建的模型中特征变量的重要性得分情况，得分越高表明变量对预测结果的影响和贡献越大。在 GBDT 算法模型中，纹理特征 IMCORR1、土壤属性 SAND 重要性得分分别排名第一、第二，地形因子 SLOPE、DEM 和 ASPECT 分别排名第三、四、五。前 4 个变量对预测结果有着较大的贡献率，从 ASPECT 开始，重要性得分开始明显下降。纹理特征 IMCORR1 和 SHADE 均排名靠前，体现出较大的贡献率。3 个地形因子均得分较高，贡献较大。加入的土壤属性指标中，SAND、CFVO 以及 SILT 重要性排名位于前半部分，而 CLAY 和 pH 相对属性排名靠后，表明在模型中贡献率较小。短波红外波段 B12、蓝波段 B2 和绿波段 B3 以及红边波段 B5 与其他光谱波段相比重要性相对更高。

在 RF 模型中，重要性得分靠前的是坡度 SLOPE、土壤属性 SAND 以及高程 DEM。在 GBDT 模型中排名第一的纹理特征 IMCORR1 下降到了第四名，SHADE 也由第八名下降到第十七名，土壤属性中 CLAY 以及 SILT 重要性较高。光谱波段中重要性前三的由 GBDT 模型中的短波红外波段 B12、蓝波段 B2、绿波段 B3 变为红波段 B4、绿波段 B3 以及红边波段 B7。地形因子 SLOPE 重要性升至第一，DEM 升为第三名，ASPECT 保持不变。

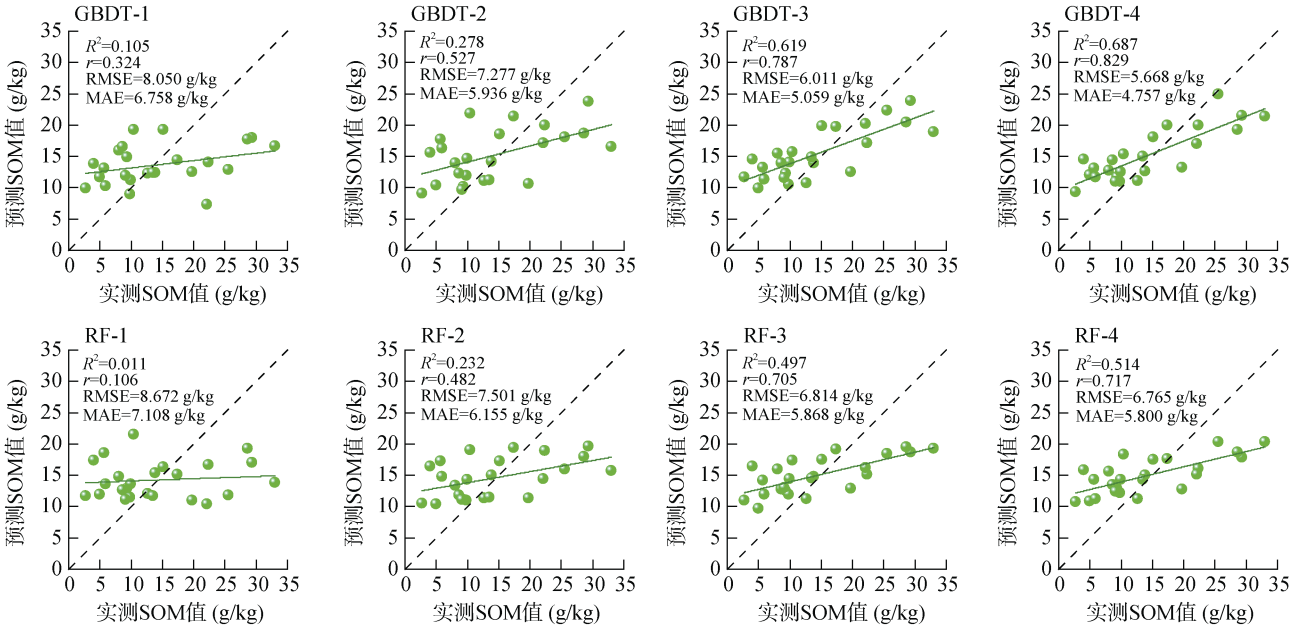

图 4 两种模型的土壤有机质预测精度及散点图
Fig. 4 SOM prediction accuracies and scatter plots of two models

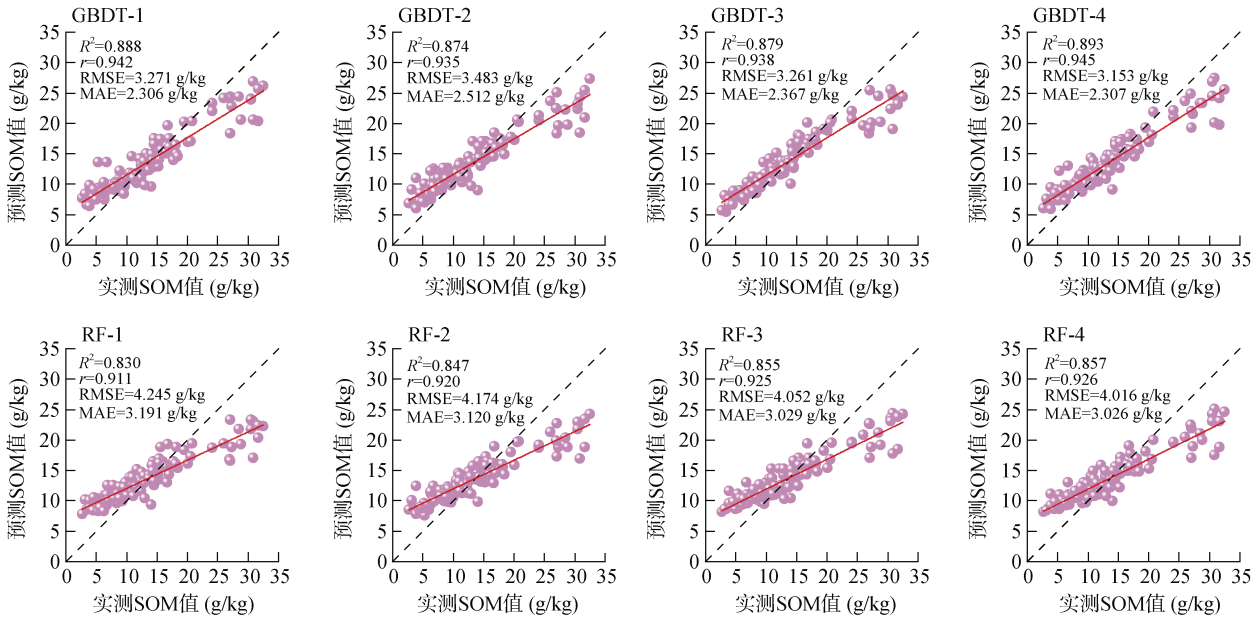

图 5　两种模型的训练集散点图

Fig. 5　Training scatter plots for two models

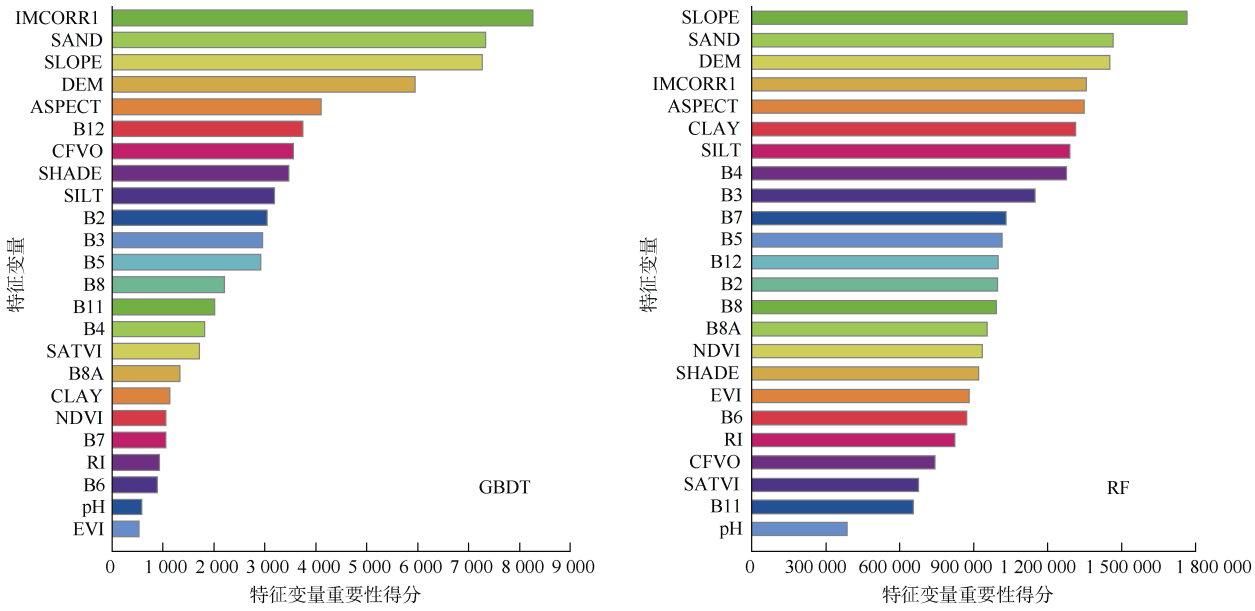

图 6　两种预测模型中特征变量的重要性得分

Fig. 6　Importance scores of feature variables in two prediction models

总的来说，3 个地形因子均表现出较大的贡献率，而纹理特征的加入可为土壤有机质预测提供新的研究思路，土壤属性以及地形因子对土壤有机质的影响作用较大，为紫色土耕地管理提供了科学依据。

3　讨论

3.1　辅助变量对土壤有机质预测的影响

地形因子是常见的预测模型构建的变量，应用于 GBDT 和 RF 模型中 R^2 分别提高 17.30% 和 22.10%，对两种模型的精度提升均起到较大作用。本研究区面积较小，而地形起伏大，最大高程与最小高程间相差 264 m，具有川中丘陵区的典型地形特征。高程、坡度、坡向不同的土壤中土壤有机质存在明显差异，地形因子的加入有利于提高模型的预测精度。可见在模型构建中，对于地形起伏大，高程差异较大的研究区域，地形因子作为辅助变量不可忽视。

研究结果表明在 GBDT 和 RF 构建的两种模型中，如表 4 所示土壤属性因子的加入使 R^2 分别提高了 34.10% 和 26.50%。这可能是因为土壤砂粒主要成分为矿物颗粒，具有通透性好的特点，土壤黏粒和

粉粒颗粒较小、黏性大，而土壤有机质主要以胶膜形式包裹于矿物表面，形成具有黏结力的有机胶体[3]。土壤因子有助于提高土壤有机质吸附养料原子的能力和减少土壤有机质的分解，土壤因子含量越多，土壤有机质积累能力越强。

本研究将 IMCORR1 和 SHADE 作为纹理特征加入模型，R^2 分别提高了 6.80% 和 1.70%，GBDT 算法下的模型对纹理特征的响应效果更优于 RF。纹理特征广泛应用于农作物分类和提取，有研究[22]表明提取冬小麦和冬油菜作物时在其他变量的基础上加入部分纹理特征，可提高分类精度；在冬小麦种植结构的精细提取研究中，纹理特征对于提取结果的贡献率最大[24]，优于其他特征变量，且 GBDT 算法下的模型精度最高，而纹理特征应用于回归预测研究相对较少。本研究创新性地将纹理特征应用于土壤有机质预测，并证明纹理特征反映的地表信息同样可用于土壤有机质回归预测，为土壤有机质预测提供了新的思路。

总的来说，变量加入后构成最终模型预测精度相较于只加入光谱波段和光谱指数的模型精度有大幅提升。研究区内紫色土土壤有机质预测不可只依靠基础光谱波段，区域土壤属性与气候条件、自然地理环境等息息相关，早期有专家研究得出在典型岩溶流域土壤有机质含量及分布与地质地貌分布具有一致性[30]；陕西渭北农田的土壤属性、地形因子等分别对土壤有机质有不同程度的影响[8]。因此大部分土壤有机质空间预测研究中模型需加入辅助信息提高精度，尤其对于较大尺度范围的土壤有机质预测，加入容易获取的地形因子、气候因子等辅助信息，提高精度的同时也节省实地测量的人力物力。根据研究区自然环境等特征适当加入辅助变量对提升土壤有机质预测结果精度有所帮助。

3.2 土壤有机质预测模型变量的重要性得分

地形因子中的所有变量在预测模型中占据重要地位，其次是纹理特征中的 IMCORR1 和土壤属性中的 SAND(图 6)。IMCORR1 的贡献率较大表明了其可能在像素对的相关性识别时发挥了重要作用。坡度 SLOPE、高程 DEM 以及坡向 ASPECT 是常用的地形因子变量，在 GBDT 和 RF 的算法模型中重要性得分均较高，贡献率较大，其中坡度 SLOPE 的贡献率最大。由土壤有机质空间分布图(图 3)和研究区概况图(图 1)可知，研究区内土壤有机质较高区域坡度较低，地势平缓，这可能是因为坡度小的区域其抗侵蚀能力以及抗冲击性相对于坡度大的区域更强，土壤经过长时间的外界侵蚀后，低坡度地区可更好集聚土壤有机质，反之高坡度土壤因缺乏抗侵蚀能力，导致土壤有机质流失。地势起伏较大、地形复杂的研究区与地形因子相关性更高[31]。有研究表明平缓地势如江汉平原[32]对于土壤属性预测结果的影响较小，这也反向印证了本文的研究结果。目前许多地区的耕地也采用了坡改梯等方式改善土壤生态条件，殷庆元等[33]的研究也表明坡地改梯田的改良方式可提升土壤抗侵蚀能力，利于土壤有机质增加。除此之外，在 GBDT 模型中 B12 在光谱波段中排名最高，这可能是因为短波红外波段对于含水量较为敏感，而土壤有机质具有保水功能，土壤有机质含量高则土壤含水量较高，有助于土壤有机质含量预测。

3.3 不足与展望

本研究为进行对照，选择同一组变量进行试验，并比较模型精度。由图 6 可以看出，在 GBDT 模型中重要性得分靠前的纹理特征中的 SHADE、土壤属性 CFVO 等，在 RF 算法下的模型中的贡献率排名相对靠后；而在 GBDT 模型中贡献较小的土壤属性 CLAY 在 RF 模型中排名靠前，这表明不同变量在不同模型中的贡献率不是唯一。文中采用相同的变量进行试验比较，表明在此套模型变量下，GBDT 算法下的模型预测精度高于 RF 算法。未来需继续尝试不同模型对于变量组合的响应试验，提高预测精度。此外，由于官方耕地地块数据的缺乏，本文使用 Globeland30 土地利用数据提取耕地矢量面作为研究区范围，此范围所提取研究区耕地的精准程度还有待提高，未来还需依据更高精度的耕地数据，实现更为精细的紫色土耕地土壤有机质空间制图。本文研究结果表明，基于辅助变量和实测数据的紫色土耕地土壤有机质预测模型具有可实施性，以及创新性地加入纹理特征可在一定程度上提高土壤有机质预测精度。由于目前纹理特征作为辅助信息较少应用于土壤有机质预测研究，本研究旨在为土壤有机质研究提供新思路，更多区域土壤有机质预测还需根据具体情况选取合适的纹理特征因子。该方法通过辅助因子提高模型精度，但同时也受辅助因子数据空间分辨率等精度影响。除此之外，由于土壤有机质含量或受其他环境因素影响，不同区域的土壤有机质的相关因素有差异，除本文所用的辅助因子外，也可结合实际和试验情况选取其他合适的土壤因子或其他因素，如比值光谱指数[34]、碱解氮[35]等有助于土壤有机质模型预测精度提高。本研究只针对小范围研究区进行预测，对于大尺度以及其他区域的推广性仍需基于更多的实测数据进行探索。

4　结论

1)研究区内紫色土土壤有机质含量普遍偏低，含量约为 2.68～33.02 g/kg，养分分级级别为二 ～ 六级。

2)在 GBDT 和 RF 算法的 6 个模型中，GBDT 相比于 RF 算法均具有更高精度的预测能力，其中最优预测结果为 R^2=0.687，r=0.829，RMSE=5.668 g/kg。

3)本研究实现了基于云平台的紫色土土壤有机质预测，GBDT 算法模型加入了地形因子、土壤属性和纹理特征后 R^2 分别提高了 17.30%、34.10% 和 6.80%，RF 算法中 R^2 分别提高了 22.10%、26.50% 和 1.70%。试验创新性地加入了纹理特征，并验证其有助于提高紫色土土壤有机质的预测精度，为土壤有机质预测提供了可靠的研究思路。

参考文献：

[1] 中国科学院成都分院土壤研究室. 中国紫色土–上篇[M]. 北京: 科学出版社, 1991.

[2] 何毓蓉. 中国紫色土–下篇[M]. 北京: 科学出版社, 2003.

[3] 黄兴成. 四川盆地紫色土养分肥力现状及炭基调理剂培肥效应研究[D]. 重庆: 西南大学, 2016.

[4] 李亨伟, 胡玉福, 邓良基, 等. 川中丘陵区紫色土微地形下有机质空间变异特征[J]. 土壤通报, 2009, 40(3): 552–554.

[5] 刘焕军, 张美薇, 杨昊轩, 等. 多光谱遥感结合随机森林算法反演耕作土壤有机质含量[J]. 农业工程学报, 2020, 36(10): 134–140.

[6] 罗梅, 郭龙, 张海涛, 等. 基于环境变量的中国土壤有机碳空间分布特征[J]. 土壤学报, 2020, 57(1): 48–59.

[7] 高浩然, 周勇, 王丽, 等. 基于 Geodetector 模型的鄂北岗地土壤有机质空间格局及影响因素分析——以枣阳市为例[J]. 长江流域资源与环境, 2022, 31(1): 166–178.

[8] 尉芳, 刘京, 夏利恒, 等. 陕西渭北旱塬区农田土壤有机质空间预测方法[J]. 环境科学, 2022, 43(2): 1097–1107.

[9] 李珊, 李启权, 王昌全, 等. 成都平原区水稻土有机碳剖面分布特征及影响因素[J]. 环境科学, 2018, 39(7): 3365–3372.

[10] 刘焕军, 赵春江, 王纪华, 等. 黑土典型区土壤有机质遥感反演[J]. 农业工程学报, 2011, 27(8): 211–215.

[11] 刘焕军, 潘越, 窦欣, 等. 黑土区田块尺度土壤有机质含量遥感反演模型[J]. 农业工程学报, 2018, 34(1): 127–133.

[12] 任必武, 陈瀚阅, 张黎明, 等. 机器学习用于耕地土壤有机碳空间预测对比研究——以亚热带复杂地貌区为例[J]. 中国生态农业学报(中英文), 2021, 29(6): 1042–1050.

[13] 赵彦锋, 李怡欣, 马盼盼, 等. 近 30 年河南省耕地土壤有机碳的三维变化与关键因素研究[J]. 土壤学报, 2023, 60(5): 1409–1420.

[14] 李莹莹, 赵正勇, 杨旗, 等. 基于 GF-1 遥感数据预测区域森林土壤有机质含量[J]. 土壤, 2022, 54(1): 191–197.

[15] 郭静, 龙慧灵, 何津, 等. 基于 Google Earth Engine 和机器学习的耕地土壤有机质含量预测[J]. 农业工程学报, 2022, 38(18): 130–137.

[16] 张笑寒, 孟祥添, 唐海涛, 等. 优化光谱输入量的土壤有机质随机森林预测模型[J]. 农业工程学报, 2023, 39(2): 90–99.

[17] 尼加提·卡斯木, 茹克亚·萨吾提, 师庆东, 等. 基于优化光谱指数的土壤有机质含量估算[J]. 农业机械学报, 2018, 49(11): 155–163.

[18] 曹永研, 杨玮, 王懂, 等. 基于水分和粒度的土壤有机质特征波长提取与预测模型[J]. 农业机械学报, 2022, 53(S1): 241–248.

[19] 聂祥琴, 陈瀚阅, 牛铮, 等. 基于时序影像的农业活动因子提取与闽西耕地 SOC 数字制图[J]. 地球信息科学学报, 2022, 24(9): 1835–1852.

[20] 张超, 高璐璐, 郧文聚, 等. 遥感技术获取耕地质量评价指标的研究进展分析[J]. 农业机械学报, 2022, 53(1): 1–13.

[21] 王朝阳, 师银芳, 侯诚. 基于 Sentinel-2A 影像的枸杞种植区域识别[J]. 生态学杂志, 2022, 41(5): 1033–1040.

[22] 何昭欣, 张森, 吴炳方, 等. Google Earth Engine 支持下的江苏省夏收作物遥感提取[J]. 地球信息科学学报, 2019, 21(5): 752–766.

[23] 张海洋, 张瑶, 田泽众, 等. 基于 GBDT 和 Google Earth Engine 的冬小麦种植结构提取[J]. 光谱学与光谱分析, 2023, 43(2): 597–607.

[24] 张甘霖, 龚子同. 土壤调查实验室分析方法[M]. 北京: 科学出版社, 2012.

[25] Hengl T, Mendes de Jesus J, Heuvelink G B M, et al. SoilGrids250m: Global gridded soil information based on machine learning[J]. PLoS One, 2017, 12(2): e0169748.

[26] Haralick R M, Shanmugam K, Dinstein I. Textural features for image classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1973, SMC-3(6): 610–621.

[27] Conners R W, Trivedi M M, Harlow C A. Segmentation of a high-resolution urban scene using texture operators[J]. Computer Vision, Graphics, and Image Processing, 1984, 25(3): 273–310.

[28] Breiman L. Random forests[J]. Machine Learning, 2001, 45: 5–32.

[29] Friedman J H. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189–1232.

[30] 蒋勇军, 袁道先, 谢世友, 等. 典型岩溶流域土壤有机质空间变异——以云南小江流域为例[J]. 生态学报, 2007, 27(5): 2040–2047.

[31] 袁玉琦, 陈瀚阅, 张黎明, 等. 基于多变量与 RF 算法的耕地土壤有机碳空间预测研究 ——以福建亚热带复杂地貌区为例[J]. 土壤学报, 2021, 58(4): 887–899.

[32] 沈佳丽, 陈颂超, 胡碧峰, 等. 基于机器学习的江汉平原土壤有机碳预测及制图[J]. 农业资源与环境学报, 2023, 40(3): 644–650.

[33] 殷庆元, 王章文, 谭琼, 等. 金沙江干热河谷坡改梯及生物地埂对土壤可蚀性的影响[J]. 水土保持学报, 2015, 29(1): 41–47.

[34] 王欣怡, 王昌昆, 马海艺, 等. 基于双时相卫星遥感光谱指数估算土壤有机质含量[J]. 土壤, 2023, 55(5): 1106–1113.

[35] 张一扬, 粟深河, 林北森, 等. 靖西市植烟土壤有机质含量的时空变异特征[J]. 土壤, 2020, 52(1): 202–206.