DOI:10.13869/j.cnki.rswc.2024.05.041.
袁可，张晨，赵建林，等. 有限样本下土壤有机碳密度空间分布预测模型对比分析[J].水土保持研究,2024,31(5):173-181,191.
Yuan Ke, Zhang Chen, Zhao Jianlin, et al. Comparative Analysis on Models for Predicting the Spatial Distribution of Soil Organic Carbon Density with Limited Samples[J]. Research of Soil and Water Conservation,2024,31(5):173-181,191.

有限样本下土壤有机碳密度空间分布预测模型对比分析

袁 可[1]，张 晨[1]，赵建林[1]，汪珍亮[1]，杨 节[1]，许中胜[2]
(1.长安大学 地质工程与测绘学院，西安 710054；2.安徽省第二测绘院，合肥 230031)

摘 要:[目的]探讨不同机器学习模型在有限样本条件下预测表层土壤有机碳密度(SOCD)空间分布的精度和适用性,为黄土高原流域尺度碳库研究提供参考。[方法]基于延河子流域有限样本,对比多元线性逐步回归(SR)、随机森林(RF)、极端梯度提升(XGB)、支持向量机(SVM)这 4 种机器学习模型对表层土壤(0—20 cm)SOCD 的预测精度和稳定性。[结果](1) 在有限样本条件下,4 种机器学习模型均可以较好地预测流域尺度 SOCD 空间分布,其中 SVM 模型精度最优,其 50 次预测的 RMSE,R^2,MAE 平均值分别为 0.74,0.43,0.64;(2) 不同土地利用类型的 SOCD 均值估算结果大小一致且具有显著差异,均为灌木林>林地>草地>耕地,研究区总有机碳储量(0—20 cm)为 2.39×10^6 t;(3) SOCD 空间分布预测因子重要性评价结果表明地形因子、$NDVI_{max}$、近红外波段地表反射率(B5)以及 *K-T* 变化中的 Brightness 因子对模型预测精度具有显著贡献。[结论]研究表明在有限样本条件下机器学习模型结合相关变量因子可有效应用于黄土高原流域尺度表层 SOCD 空间分布反演及碳库研究。

关键词:土壤有机碳；机器学习；控制因子；黄土高原
中图分类号:S153.621 **文献标识码**:A **文章编号**:1005-3409(2024)05-0173-09

Comparative Analysis on Models for Predicting the Spatial Distribution of Soil Organic Carbon Density with Limited Samples

Yuan Ke[1], Zhang Chen[1], Zhao Jianlin[1], Wang Zhenliang[1], Yang Jie[1], Xu Zhongsheng[2]
(1.*College of Geological Engineering and Geomatics, Chang'an University, Xi'an* 710054, *China*;
2.*The Second Surveying and Mapping Institute of Anhui Province, Hefei* 230031, *China*)

Abstract: [Objective] The aim of this study is to explore the accuracy and applicability of different machine learning models for predicting the spatial distribution of surface soil organic carbon density (SOCD) with limited samples, which can provide a references for the study of watershed scale carbon pool in the Chinese Loess Plateau.[Methods] In this study, we compared the accuracy and stability of the predicted SOCD in topsoil (0—20 cm) by four machine learning models, namely Multiple Linear Stepwise Regression (SR), Random Forest (RF), Extreme Gradient Boosting (XGB) and Support Vector Machine (SVM), based on the limited measured samples in a sub-watershed of Yanhe River in the Chinese Loess Plateau. [Results] (1) Under the condition of limited samples, all models successfully and appropriately predicte the spatial distribution of SOCD, among which the SVM model has the best model performance, and the average RMSE, R^2 and MAE of 50 predictions is 0.74, 0.43 and 0.64, respectively. (2) The average SOCD of different land use types are consistent between measured and predicted values but shows significant difference among land use types. SOCD decreases in the order: shrubland > forestland > grassland > cropland. The total organic carbon of cultivated land in the study area is 2.39×10^6 t (0—20 cm). (3) The evaluation of feature importance shows

收稿日期:2023-11-03 **修回日期**:2023-11-17
资助项目:陕西林业科技创新重点专项(SXLK2023-02-15);国家自然科学基金(41907048);中央高校基本科研费专项资金(300102260206)
第一作者:袁可(1998—),男,四川省南部人,硕士研究生,研究方向为土壤有机碳遥感反演与制图。E-mail:2021126063@chd.edu.cn
通信作者:赵建林(1988—),男,陕西省镇巴人,副教授,主要从事土壤侵蚀与区域生态评价工作。E-mail:jianlin.zhao@chd.edu.cn

that terrain factors, $NDVI_{max}$, near-infrared surface reflectance (B5) and Brightness index have significant contributions to the accuracy of predictions. [Conclusion] Under the condition of limited samples, the machine learning model combined with controlling features can be effectively applied to the prediction of the spatial distribution of topsoil SOCD at the watershed scale in the Chinese Loess Plateau.

Keywords: soil organic carbon; machine learning model; controlling factor; Chinese Loess Plateau

土壤碳库是陆地生态系统中最大的碳库,约占全球碳储量的50%～80%,是大气和生物圈碳含量的2～3倍[1]。土壤有机碳是衡量土壤质量的重要指标,在碳固存、保水能力以及全球气候影响等生态服务功能上具有非常重要的意义[2]。基于土壤碳库含量以及对大气二氧化碳浓度贡献,可将其定义为全球二氧化碳循环的"汇"(从大气中净吸收二氧化碳)、"源"(往大气净排放二氧化碳)[3],因此土壤有机碳储量的微小变化将显著影响全球碳循环和土壤的物理、化学和生物特性[4]。土壤有机碳密度(Soil organic carbon density,SOCD)作为定量评价土壤有机碳储量的重要参数,精确、快速地获得其空间分布及其动态对陆地生态系统碳通量动态变化、全球气候变化以及我国实现"双碳"目标具有重要意义[5]。

基于点位实测数据在区域尺度上精确地预测表层土壤有机碳含量,并揭示其空间变异规律,一直以来是数字化土壤制图(Digital soil mapping,DSM)的重要内容。现阶段常用的估算方法主要分为样地清查和模型估算,其中样地清查的方法主观性较强,对地形复杂和土壤属性空间变异较大的地区,估算结果具有较大的不确定性。模型估算方法主要分为过程模型与地统计模型,广泛使用的过程模型包括Century,Roth C及DNDC模型等[6]。常用的地统计模型为克里金(Kriging)插值法,但由于土壤本身的复杂性和外界环境的多变性,该方法对样本点的数量要求较高,同时较难准确分析影响SOCD空间分布的特征变量因子[7]。近年来,随着机器学习、数据挖掘算法以及遥感技术的不断发展和完善,基于多种相关因子结合机器学习方法预测SOCD成为了数字化土壤制图的新兴方法[8]。同时机器学习算法可以对相关因子的重要性进行评估,建立更加精准的预测模型。当前机器学习中常用的SOCD估算模型有随机森林、支持向量机以及XGBoost等模型[9]。不同模型在处理样本及特征维数上所表现的模型性能有所差异,而相关研究报道较少。目前SOCD估算精度以及空间分辨率主要受制于特征变量因子的分辨率以及训练样本的数量和空间分布。对于地形复杂的黄土高原区域,获取足够样本具有一定挑战,因此评价在有限样本条件下不同机器学习模型的精度、稳定性以及适用性具有重要的研究价值。

本文以黄土高原延河典型子流域为研究区,基于有限样本条件,探索不同机器学习模型结合相关特征变量因子估算表层SOCD空间分布及其精度的差异性,同时对模型的超参数调优和变量因子的重要性进行分析,以期为黄土高原地区流域尺度表层土壤有机碳空间分布模拟和碳库储量估算研究提供方法探索。

1 数据和方法

1.1 样本及研究区域

本研究使用的SOCD样本来自于中国陆地生态系统碳储量数据库[10],该数据库发布了近年来有关中国陆地生态系统土层深度为0—20 cm,0—100 cm碳储量的实测数据。本研究主要关注土壤有机碳的表层土壤(0—20 cm),通过水文分析将延河流域划分为不同子流域,选取样本点覆盖最密集的两个子流域作为研究区域。两个子流域位于108°58′15″—109°34′16″E和36°22′39″—36°58′57″N。流域总面积为1302.2 km^2,海拔473～1 800 m,平均海拔1 220 m,地形起伏较大,属于半干旱大陆季风气候区。流域内梁峁起伏、沟壑纵横,土壤类型主要为黄绵土、冲积土等,是典型的黄土丘陵沟壑区(图1)。

研究区总样本数共计99个,从样本点的空间分布上看,样本总体覆盖研究区的主要土地利用类型(表1)和土壤类型,空间分布相对均匀。样本数量与研究区土地利用类型、土壤类型、地形地貌空间分布特征相对应,总体上具有较好的代表性。

1.2 特征因子选取和数据来源

相关研究表明,影响土壤有机碳空间分布的因子主要包括:地形、土地利用、土层深度、植被指数和气候条件等[11-12]。基于数据的可获得性和研究区域地貌特征,本文共选取14个土壤有机碳分布控制因子,主要包括地形因子、遥感反演指数、影像波段地表反射率三大类。其中地形因子包括:高程(Elevation)、地形位置指数(TPI)、地形坚固性指数(TRI)、坡度(Slope);遥感反演指数用以表征研究区植被覆盖、生物量以及半干旱气候区的水分空间特征,主要包括:归一化差异水指数(NDWI)、植被指数年最大值

($NDVI_{max}$)、增强型植被指数(EVI)以及 K-T 变化的前 3 个分量(亮度、绿度、湿度);以及影像波段的地表反射率,主要包括:B4(红波段)、B5(近红外波段)、B6(短波红外 1)、B7(短波红外 2),各影像波段地表反射率波长范围为 0.630～2.300 μm,可较好地表征裸露土壤的特征。对于小尺度流域,现阶段气候因子数据分辨率无法有效表征空间上的气候变化特征以满足研究需要,因此本研究暂未考虑气候因子。

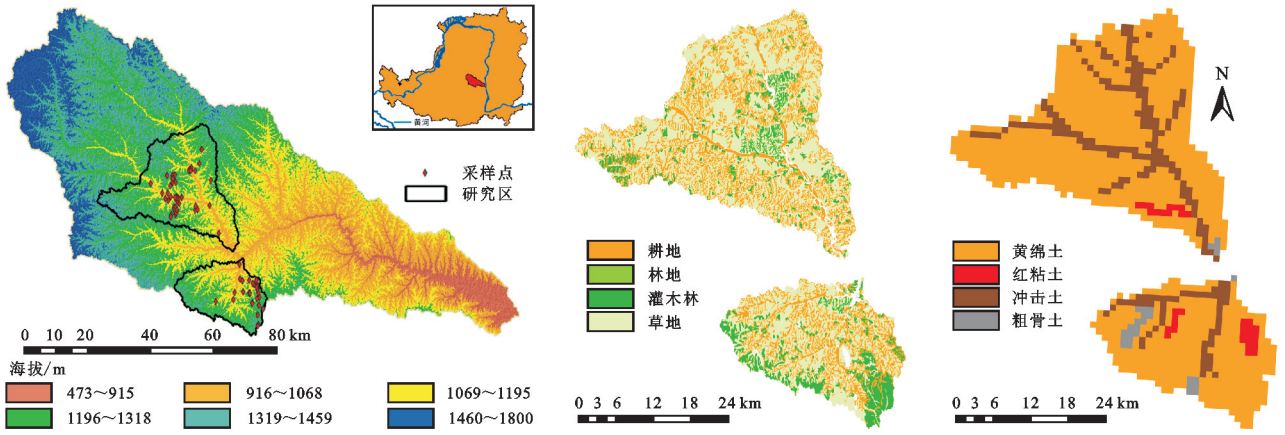

图 1　研究区及样本点空间分布示意图

Fig. 1　Schematic diagram of study area and spatial distribution of samples

表 1　小流域土壤有机碳密度样本信息

Table 1　Information of soil organic carbon density samples

土地利用类型	最小值/(kg·m^{-2})	最大值/(kg·m^{-2})	平均碳密度/(kg·m^{-2})	样本个数
草地	0.7200	3.2830	1.5494±0.1604	45
耕地	0.6161	1.4700	1.1624±0.1253	20
林地	0.4275	4.5914	1.7749±0.1815	17
灌木林	0.6800	6.1275	2.8749±0.3486	17

上述因子的计算方法和数据来源如表 2 所示。其中,遥感影像数据为地理空间数据云(http://www.gscloud.cn)提供的 Landsat 8 OLI 数据,空间分辨率为 30 m。为了避免影像获取时间差异和外界环境的干扰,选择云量小于 5%且相同时间段完全覆盖两个小流域的两幅影像作为影像数据源,并进行波段融合、辐射定标、大气校正、裁剪等一系列预处理。后续分析所使用的土地利用数据和土壤类型数据来源于中国科学院资源环境数据中心(https://www.resdc.cn)。其中土地利用数据空间分辨率为 30 m,选取耕地、林地、灌木林、草地等一级地类作为研究对象,其余地类均归类为其他类型并作掩膜处理。土壤类型数据空间分辨率为 1 km,采用了传统的"土壤发生分类系统",本研究主要关注土类属性特征。

1.3　机器学习模型

1.3.1　*多元线性逐步回归*　多元线性回归模型常被用于研究多个变量间相互依赖的关系,由于各因子变量可能存在一定的共线性,因此本研究采用多元线性回归模型中的逐步回归(Stepwise Regression,SR)模型建立最终的预测方程。SR 模型基于一定阈值排序建立"最优"的多元线性回归方程,因此不仅可以保证与 SOCD 显著相关的自变量进入回归模型,而且可以去除自变量间的共线性[13]。

1.3.2　*随机森林*　随机森林(Random Forest,RF)模型是一种基于分类回归树的机器学习算法,采用 Bagging 思想,由多棵相互没有关联的决策树组成的集成决策树[14],可以有效避免特征间的多元共线性问题,其对缺失或非平衡的数据表现出较好的稳定性,同时能对解释变量进行重要性评估[15]。本研究利用节点纯度增量(IncNodePurity)指标来定性表征特征变量重要性[16]。RF 模型的构建需要考虑到两个关键的超参数:决策树的数量(ntree)以及分割节点处的随机变量个数(mtry)。本研究采用网格搜索结合 5 折交叉验证的方法确定最优 ntree 和 mtry 参数的选取。

1.3.3　*极端梯度提升*　极端梯度提升(Extreme Gradient Boosting,XGB)算法是基于 GBDT(Gradient Boosting Decision Tree)的一种改进机器学习算法,通过构建多个弱学习器的输出值集成给出最终学习结果。XGB 模型可以高效地处理大规模数据集和高维稀疏特征,对样本缺失值和异常值具有较强的容错能力[17]。同时 XGB 提供了丰富的模型解释和可视化功能,可以输出特征变量的重要性、分裂贡献等信息,以提高模型的可解释性和可用性。XGB 模型的参数主要包括学习率(eta)、决策树分裂参数(gamma)、决策树分裂深度(max_depth)、分裂变量的选择(bytree)、迭代次数(nrounds)以及模型中止策略等[18]。本研究采取试错法结合经验法确定相关参数的最优值。

表2 相关因子详细信息

Table 2 Details of relevant factor

因子	缩写	空间分辨率/m	计算方法	数据描述
高程	Elevation	30		高程数据类型为ASTER GDEM。各地形因子基于ArcGIS软件计算，式中：Z_0表示中心点高程；R表示预设领域；Z_i表示领域内高程；n表示领域内高程点数。其中TPI，TRI可以表达区域范围内的地形起伏特征，反映地形表面形状、凹凸变化
地形位置指数	TPI	30	$TPI=Z_0-\frac{1}{n_R}\sum_{i\in R}Z_i$	
地形坚固性指数	TRI	30	$TRI=\sum_{i=1}^{8}\frac{\lvert Z_0-Z_i\rvert}{8}$	
坡度	Slope	30		
归一化差异水指数	NDWI	30	$NDWI=\frac{(Green-NIR)}{(Green+NIR)}$	式中：Green表示绿波段；Red表示红波段；NIR表示近红外波段；SWIR表示短波红外波段；Blue表示蓝波段。其中NDWI可以凸显水体信息，植被指数可以对地表植被状况进行度量。K-T变化中，亮度分量反映了地物总体的亮度变化；绿度分量与植被覆盖、生物量等相关；湿度分量反映地面水分条件，特别是土壤的湿度条件
植被指数年最大值	$NDVI_{max}$	30	$NDVI=\frac{(NIR-Red)}{(NIR+Red)}$	
增强型植被指数	EVI	30	$EVI=2.5\times\frac{NIR-Red}{NIR+6\times Red-7.5\times Blue+1}$	
湿度	Wetness	30	缨帽变换（K-T变换）第三分量	
绿度	Greeness	30	缨帽变换（K-T变换）第二分量	
亮度	Brightness	30	缨帽变换（K-T变换）第一分量	
短波红外2反射率	B7	30		影像波段地表反射率经Landsta8遥感影像辐射定标、大气校正等预处理后得到
短波红外1反射率	B6	30		
近红外波段反射率	B5	30		
红波段反射率	B4	30		

1.3.4 支持向量机 支持向量机（Support Vector Machine，SVM）模型是机器学习领域的一种经典算法，SVM模型可以较好地解决小样本、非线性、维数灾难、过学习和局部极小等问题[19]。本研究选取适用性最强的径向基函数（radial basis function，RBF）进行建模。对于核函数为RBF的SVM，影响模型性能的主要参数是惩罚参数（cost）和核函数参数（gamma）[20]。本研究采用网格搜索结合5折交叉验证的方法确定最优cost和gamma参数的选取。

1.4 模型精度指标

基于样本数据情况，采用等比、随机的方法进行数据抽样，其中训练集和测试集样本点个数比为3∶1。RF，XGB，SVM模型的精度评价采用均方误差（RMSE）、决定系数（R^2）、平均绝对误差（MAE）3种指标进行衡量，各指标计算方法如公式（1），（2），（3）所示。为体现模型的稳定性和泛化能力，将各模型50次预测的均值作为最终预测的输出值，同时计算50次模型预测的标准差作为模型的估算误差。

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

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\widetilde{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y}_i)^2} \tag{2}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\lvert(y_i-\widetilde{y}_i)\rvert \tag{3}$$

式中：y_i为实测值；$\widetilde{y}_i$为预测值；$\overline{y}_i$为实测值的平均值，n为样本数量。

SR模型的精度和稳定性使用Jack-Knife程序[21]进行评价：总体思路是将总样本依次剔除（总数为n），利用剩余的$n-1$个样本构建回归模型，将剔除的样本的因子参数作为输入量预测该样本的SOCD，使用模型效率（ME）对所得模型的性能进行评估，ME计算方法见公式（4）。同时计算SR模型的RMSE，MAE两种精度指标，对比分析并评价各模型的精度差异。

$$ME=1-\frac{\sum_{i=1}^{n}(M_i-P_i)^2}{\sum_{i=1}^{n}(M_i-\overline{M})^2} \tag{4}$$

式中：M_i和P_i分别为第i个样本SOCD的实测值和预测值；$\overline{M}$为所有样本SOCD平均实测值。

1.5 不同地类碳密度和总有机碳储量估算

利用各机器学习模型结合土地利用数据，估算研究区不同地类的土壤有机碳密度均值（mean soil organic carbon density，MSOCD）和总土壤有机碳量（total soil organic carbon，TSOC）。对比分析不同机器学习模型、不同地类的MSOCD和TSOC大小分布特征。其中基于样本数据估算TSOC见公式（5），基于机器学习模型估算TSOC见公式（6）。

$$TSOC_{(SC)}=\sum(SOCD_{\overline{M}}\times S)/1000 \tag{5}$$

$$TSOC_{(ML)}=\sum_{i=1}^{n}(SOCD_i\times S_i)/1000 \quad (6)$$

式中:$TSOC_{(sc)}$ 为样本数据估算的总有机碳量(t);$SOCD_{\overline{M}}$ 为不同地类样本的 SOCD 均值(kg/m²);S 为不同地类的面积(m²);$TSOC_{(ML)}$ 为各机器学习模型估算的总有机碳量(t);$SOCD_i$ 为第 i 个土壤像元的有机碳密度(kg/m²);S_i 为第 i 个土壤像元的面积(m²);n 为图斑个数。相关的建模和统计计算在 R 平台进行,空间分析在 ArcGIS 10.6 平台进行。

2　结果与分析

2.1　模型参数寻优

SR 模型的最优模型见公式(7),其中具有显著相关性的特征变量为 3 个地形因子(Elevation,TRI,TPI)和 K-T 变化中反映地物总体亮度变化的 Brightness 因子。RF 模型最优参数见图 2A,对比不同参数组合,结果表明当 mtry 值为 2,ntree 值为 500 时 RF 模型误差最小。XGB 模型中,本研究的 eta 设置为 0.3,gamma 为 0.001,max_depth 为 2,nrounds 为 1 000,为防止过拟合,bytree 值为 40%;并设置提前终止策略为每迭代 100 次显示结果,再迭代 200 次后无误差降低即中止模型训练。SVM 模型的最优参数见下图 2B 所示,对比不同参数组合,结果显示当 cost 值为 2,gamma 值为 0.5 时 SVM 模型误差最小。

$$SOCD=3.4265+0.0027\times Elevation^{**}-3.5465\times TRI^{*}-0.1147\times TPI^{*}-3.2269\times Brightness^{**} \quad (7)$$

式中:* 表示 $p<0.05$,** 表示 $p<0.01$。

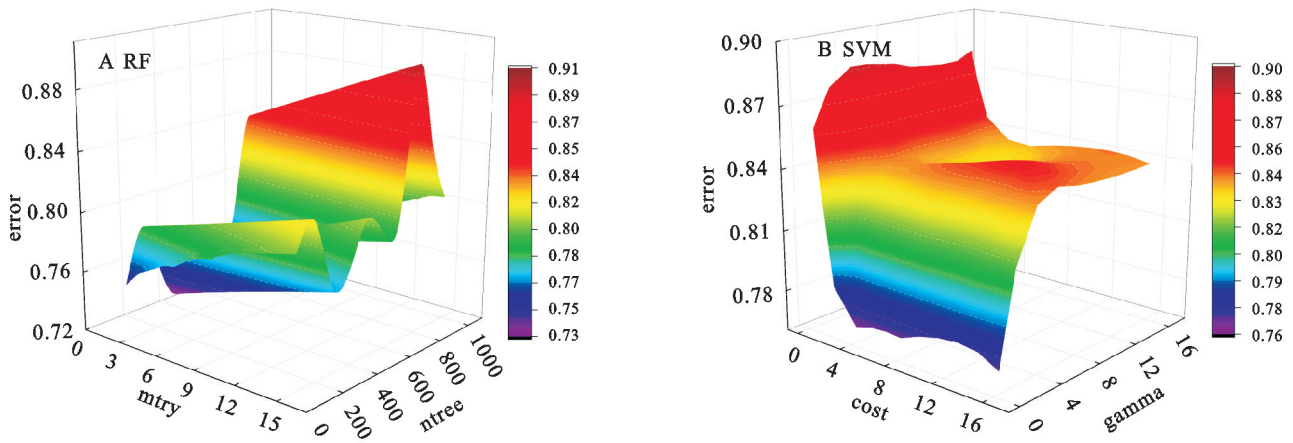

图 2　机器学习模型参数寻优

Fig. 2　Parameter optimization of machine learning model

2.2　模型估算精度

基于 50 次 RF,XGB 和 SVM 模型预测的平均精度指标结果如图 3A 所示,其中 RF 模型、XGB 模型和 SVM 模型的 RMSE,R^2,MAE 的平均值依次为0.75,0.38,0.68;0.77,0.39,0.66;0.74,0.43,0.64。SR 模型的预测精度如图 3B 所示,模型精度评价指标 RMSE,ME,MAE 的值分别为 0.83,0.40,0.61,上述结果表明 SVM 模型的精度和稳定性均优于 RF,XGB 和 SR 模型。

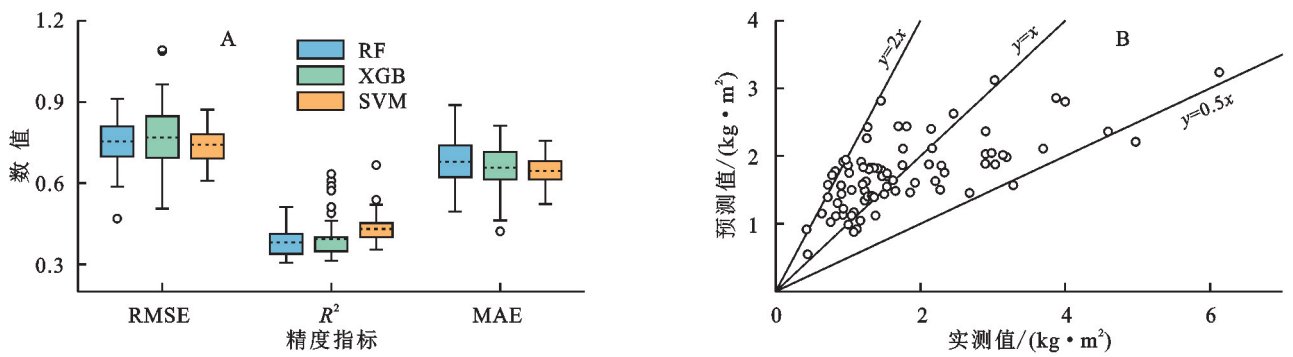

图 3　机器学习模型精度

Fig. 3　Accuracy of machine learning model

2.3　土壤有机碳密度空间分布

各模型预测的表层 SOCD 空间分布如图 4 中的 RF_Mean,SVM_Mean,XGB_Mean 和 SR 所示。总体来看,各种机器学习模型预测的 SOCD 空间分布趋势大致相同,均呈现西北和南部区域值高、中部偏低。预测结果表明林地和灌木林 SOCD 较高,而耕地 SOCD 相对较小(图 5)。从 SOCD 数值区间上来看,SVM 模型的预测值与样本值区间相符,SR 模型

下的预测值会出现一些极小值，而 RF 模型和 XGB 模型预测的值偏低。

各模型预测的标准差空间分布情况如图 4 中的 RF_SD，SVM_SD 和 XGB_SD 所示，总体上 SOCD 相对较大的区域的标准差也会相对较大。数值上来看，SVM 模型和 RF 模型的标准差均小于 XGB 模型，而整体上 SVM 模型的标准差值更小，模型相对更稳定。

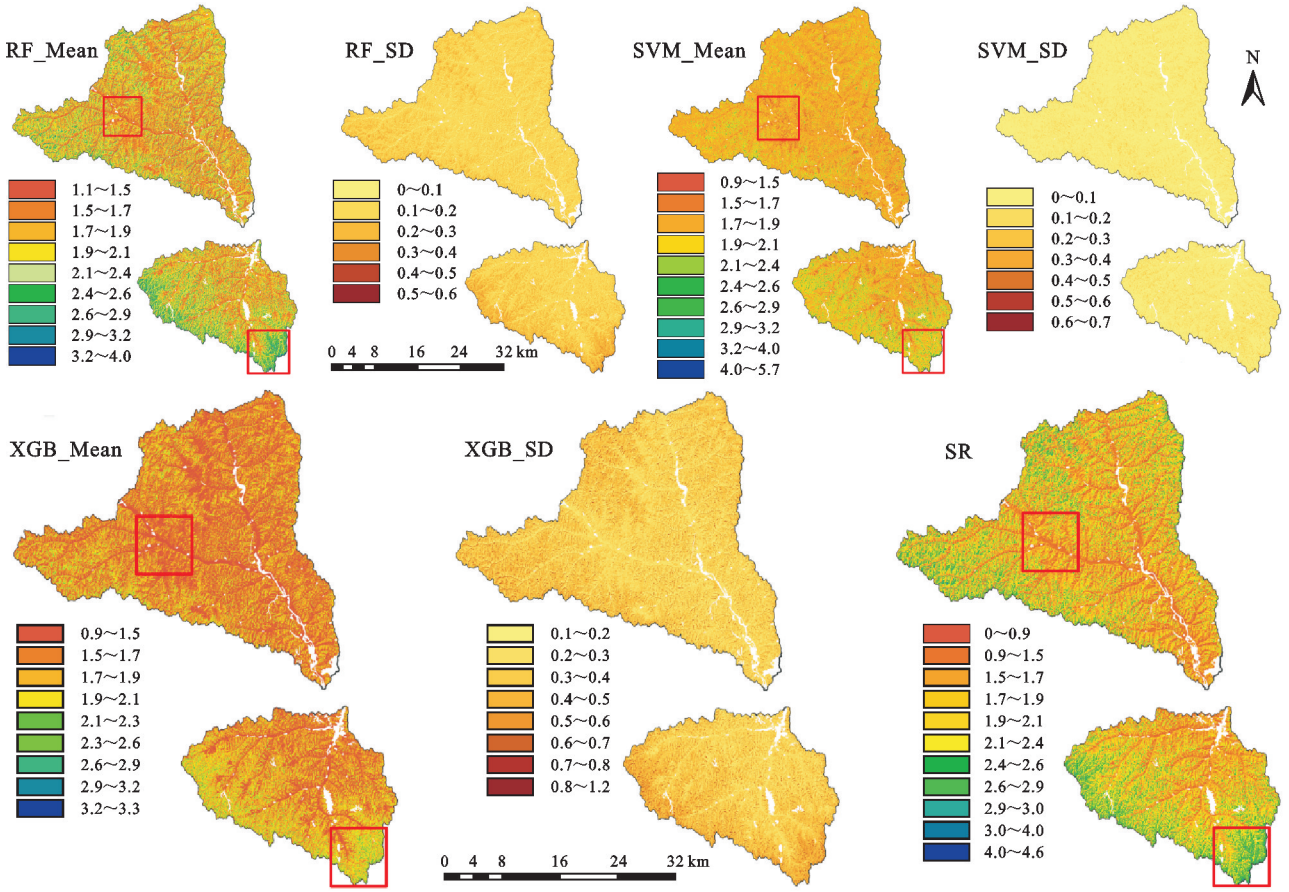

图 4 SOCD 空间分布和标准差

Fig. 4 Spatial distribution of soil organic carbon density and standard deviation

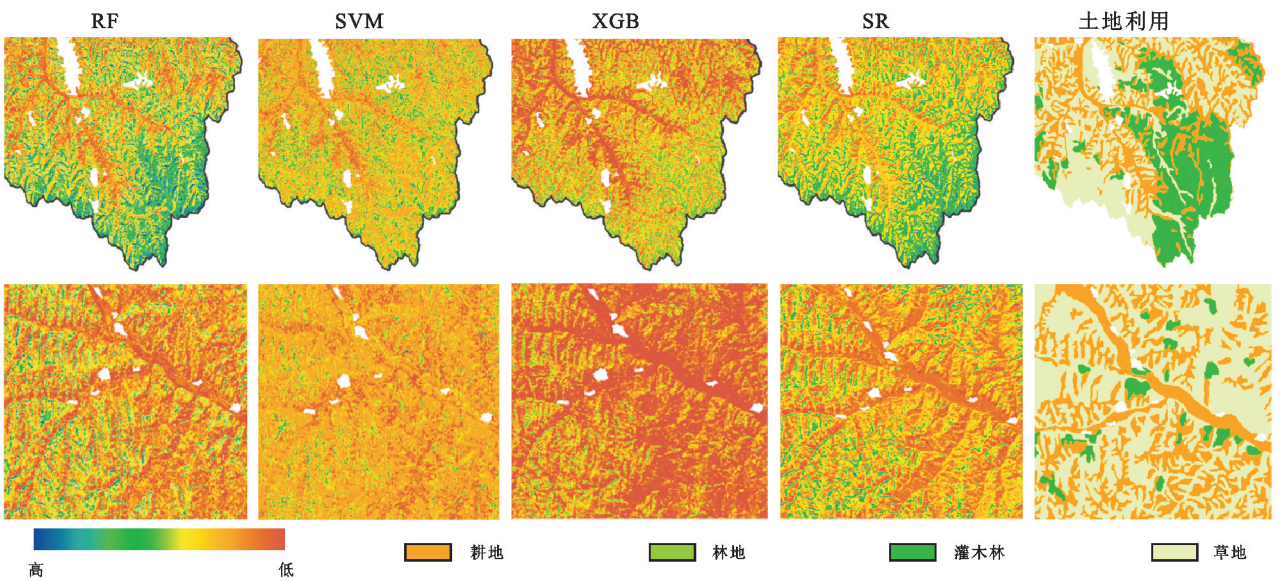

图 5 SOCD 空间分布和土地利用类型对比细节

Fig. 5 Spatial distribution of soil organic carbon density and land-use type

2.4 流域土壤有机碳总量估算

图 6 表示各模型预测不同土地利用类型 SOCD 均值和总有机碳含量(TSOC)。研究结果表明在样本数据估算和机器学习模型预测上，不同土地利用类型下的

SOCD 均值含量大小顺序一致,均为:灌木林>林地>草地>耕地;研究区各地类的 TSOC 含量大小顺序均为:草地>耕地>林地>灌木林。以本研究最优机器学习模型(SVM)估算,研究区草地 TSOC 为 1.33×10^6 t,耕地 TSOC 为 7.6×10^5 t,林地 TSOC 为 1.9×10^5 t,灌木林 TSOC 为 1.2×10^5 t,研究区 TSOC 为 2.39×10^6 t。

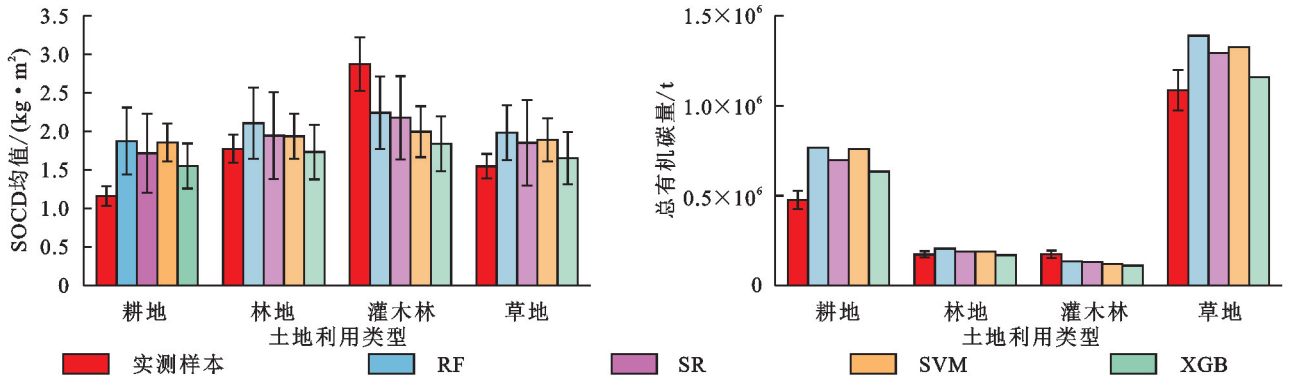

图 6　不同机器学习模型下的各地类 SOCD 和 TSOC 比较

Fig. 6　SOCD and TSOC comparison of land-use type under different machine learning model

2.5　模型因子重要性分析

黄土高原具有复杂的地形地貌和环境特征,基于流域尺度和有限样本数据,不同因子在不同模型预测的重要性不同。其中基于多因子综合分析(公式 7)表明,地形因子(Elevation,TRI 和 TPI)和反映地物总体亮度变化的 Brightness 指数是 SR 模型预测的主要因子。而 RF 模型和 XGB 模型中排名前 5 的因子均为:地形因子(TPI,Elevation)、反映地物总体亮度变化的 Brightness 指数、近红外波段地表反射率(B5)以及植被指数年最大值($NDVI_{max}$)(图 7)。上述结果为黄土高原流域尺度预测 SOCD 特征变量因子选择提供了一定的参考。

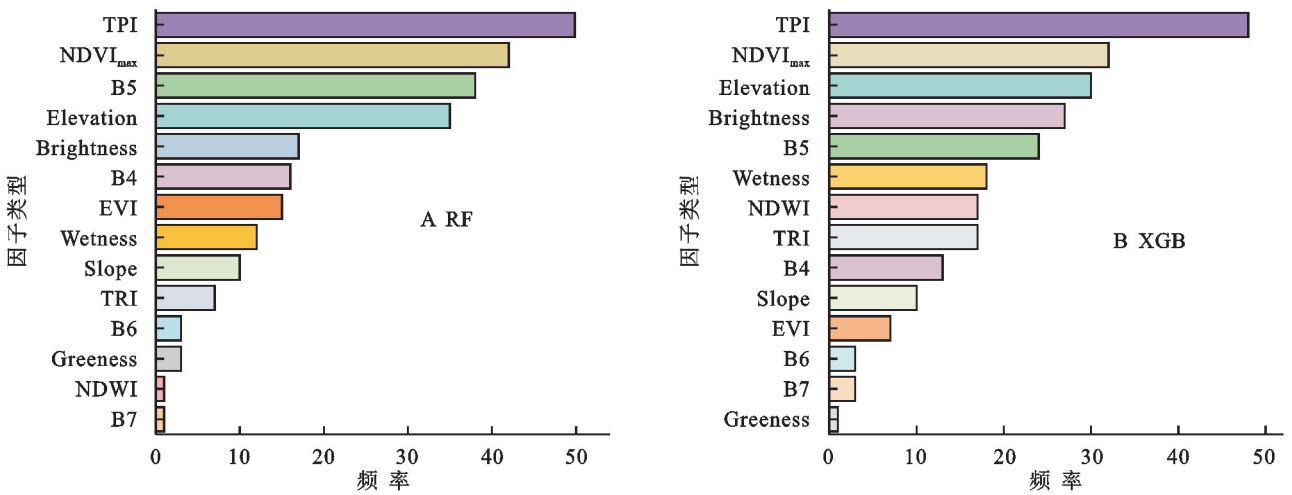

图 7　机器学习模型因子重要性

Fig. 7　Factor importance of machine learning model

3　讨 论

高分辨率和高精度区域 SOCD 空间分布估算是当前陆地生态系统碳循环以及全球气候变化研究的热点和难点,由于相关数据的缺乏和方法模型的不稳定,使得相关研究结果存在较大的不确定性,特别是在模型的选择以及相关特征变量因子的选取和量化上[22]。本研究基于表层土壤有限样本数据(0—20 cm),对比分析了目前常用的 4 种机器学习模型(即 SR,RF,XGB 和 SVM)在黄土高原流域尺度上对 SOCD 空间分布估算的精度和稳定性,并对机器学习模型的超参数寻优和相关特征变量因子的重要性进行了探讨,成功获取了目前精度相对较高(30 m)的流域尺度 SOCD 空间分布图,同时估算了流域各地类的平均碳密度和总有机碳储量。研究结果表明,在有限样本条件下采用机器学习模型方法能够快速获取和准确预测流域尺度精度可接受的 SOCD 空间分布数据。在本研究中虽然各机器学习模型的估算结果在空间分布表现上一致,但在数值上与样本值相比各模型仍存在一定的差异,例如 SR 模型存在极小值

问题,这可能与相关特征变量因子数值非同一量级有关;而 RF 模型和 XGB 模型的预测值过于保守从而忽略了极值样本的影响,这可能与模型的性能以及特征变量的维数有关,RF 模型和 XGB 模型更适用于处理高维稠密性数据,而相对于小样本、有限特征数据,SVM 模型在本研究中体现出更优的性能。

本研究中各机器学习模型预测的不同土地利用类型 SOCD 均值含量大小顺序与样本值一致,均为:灌木林>林地>草地>耕地,且存在显著差异,这说明各机器学习模型在预测研究区 SOCD 时具有一定的可靠性。不同土地利用类型下的 SOCD 均值大小差异显著,说明土地利用是造成区域土壤有机碳空间分布差异的重要因素之一,这与国佳欣、李龙等学者[23-24]研究结果一致。因此在反演目标区域 SOCD 时应当考虑到不同土地利用类型的影响,同时可以考虑利用局部回归的方法提升模型的精度[25]。

机器学习模型超参数的调优是模型精度的关键,本研究基于网格搜索结合交叉验证法和经验结合试错法对机器学习模型的超参数设置进行了一定的探讨,研究结果表明当调整至最优超参数时,模型精度将显著提高。SR 模型、RF 模型以及 XGB 模型自带的特征变量重要性解释可为反演 SOCD 空间分布的特征变量因子选择提供一定的参考。本研究中模型因子重要性表明地形因子具有显著贡献,研究区位于黄土高原腹地的延河流域,千沟万壑、地形破碎、地貌特征复杂多变,形成了典型的沟壑侵蚀地貌。复杂的地形对土壤的物质、能量过程、理化性质、土地生产力、土壤水分状况、微气候等均会产生影响[26]。因此在地貌起伏变化较大、地形特征相对复杂的区域,相关地形因子应该作为重要的特征因子,这与张祎、郭治兴等学者[27-28]研究结果一致。同时,光谱因子,特别是近红外波段(B5)对模型预测也具有显著贡献,这可能是土壤有机质中包含大量的氢基团,而近红外波段(B5)可以很好地捕捉到有机质中含氢基团的变化[29],因此能够反映 SOCD 的空间分布。*K-T* 变换是通过光谱线性变换以达到遥感图像光谱增强的方法,增强后的图像与植被生长和土壤有密切的关系,其变化后的第一分量即为反映地物总体亮度的指数(Brightness),本研究中所有变量解释模型均表明 Brightness 指数为反演 SOCD 的重要变量。植被是土壤中 SOC 的重要来源,它与表土中 SOC 的空间格局高度相关[30],NDVI 年最大值合成($NDVI_{max}$)是当前常用的 NDVI 合成方法,该方法可以进一步消除云、大气、太阳高度角等部分干扰,其反映了研究区植被覆盖的大小,本研究中 $NDVI_{max}$ 在 RF 和 XGB 模型中也表现出了较高的重要性。

然而本研究还存在一定的不足和局限性,本文探求获取小流域尺度 30 m 分辨率的 SOCD 空间分布,而目前缺乏同等精度或者更高精度的相关气候因子空间分布数据,故本文在相关特征因子选取时未考虑,造成一定模型精度损失。下一步应加强在机器学习模型中土壤有机碳样本采集策略优化以及更高精度和更具代表性的特征因子获取等领域的研究,从而提高 SOCD 的预测精度。

4 结 论

本研究以黄土高原小流域为研究对象,对比分析了 4 种机器学习模型在有限样本条件下开展流域尺度 SOCD 空间分布预测研究,结果表明:

(1) 在流域尺度上利用有限样本结合机器学习模型也可获得精度可接受的 SOCD 空间分布,并可以有效地估算研究区不同地类的 SOCD 和总土壤有机碳储量。

(2) 小样本数据集和有限维数特征变量条件下,不同机器学习模型性能将有所差异,本研究中 SVM 模型是精度和稳定性最优的模型,同时模型精度随超参数调优将显著提高。

(3) 对于不同的土地利用类型,其 SOCD 含量大小将有显著差异,同时在特殊地貌类型下影响 SOCD 分布的核心因子也将有所差异,机器学习能够对相关因子的重要性进行评估,从而进一步为模型建立和预测提供更加精准的指导。

参考文献(References):

[1] Li X H, Ding J L, Liu J, et al. Digital mapping of soil organic carbon using sentinel series data: A case study of the ebinur lake watershed in Xinjiang[J]. Remote Sensing, 2021,13(4):769.

[2] 马海丽.黄河源区土壤有机碳影响因子作用机制及模拟模型研究[D].兰州:兰州大学,2021.
Ma H L. Study on the Mechanism and Simulation Model of Soil Organic Carbon Influencing Factors in the Source Area of the Yellow River[D]. Lanzhou: Lanzhou University, 2021.

[3] Piao S L, Fang J Y, Ciais P, et al. The carbon balance of terrestrial ecosystems in China[J]. Nature, 2009, 458:1009-1013.

[4] Lamichhane S, Kumar L, Wilson B. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review[J]. Geoderma, 2019,352:395-413.

[5] 刘艳芳,宋玉玲,郭龙,等.结合高光谱信息的土壤有机碳密

度地统计模型[J].农业工程学报,2017,33(2):183-191.
Liu Y F, Song Y L, Guo L, et al. Geostatistical models of soil organic carbon density prediction based on soil hyperspectral reflectance[J]. Transactions of the Chinese Society of Agricultural Engineering, 2017,33(2):183-191.

[6] 李妙宇.黄土高原生态系统碳储量现状及固碳潜力评估[D].北京:中国科学院大学(中国科学院教育部水土保持与生态环境研究中心),2021.
Li M Y. Carbon Storages and Carbon Sequestration Potentials of the Terrestrial Ecosystems on the Loess Plateau[D].Beijing: Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, 2021.

[7] 丁亚鹏,张俊华,刘玉寒,等.基于 GWR 模型的伊河流域土壤有机碳空间分布特征及影响因素分析[J].生态学报,2021,41(12):4876-4885.
Ding Y P, Zhang J H, Liu Y H, et al. Spatial distribution characteristics and influencing factors of soil organic carbon in Yihe River Basin based on GWR model[J]. Acta Ecologica Sinica, 2021,41(12):4876-4885.

[8] 魏宇宸,卢晓丽,朱昌达,等.基于地形与遥感辅助信息的小流域尺度高分辨率有机碳空间分布预测研究[J].土壤学报,2023,60(1):63-76.
Wei Y C, Lu X L, Zhu C D, et al. High-resolution digital mapping of soil organic carbon at small watershed scale using landform element classification and assisted remote sensing information[J]. Acta Pedologica Sinica, 2023,60(1):63-76.

[9] Odebiri O, Mutanga O, Odindi J. Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data[J]. Geoderma, 2022,411:115695.

[10] Xu L, Yu G R, He N P, et al. Carbon storage in China′s terrestrial ecosystems: A synthesis[J]. Scientific Reports, 2018,8:2806.

[11] Sothe C, Gonsamo A, Arabian J, et al. Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations[J]. Geoderma, 2022,405:115402.

[12] 杜雪,王海燕.中国森林土壤有机碳活性组分及其影响因素[J].世界林业研究,2022,35(1):76-81.
Du X, Wang H Y. Active components of forest soil organic carbon and its influencing factors in China[J]. World Forestry Research, 2022,35(1):76-81.

[13] 赵青,刘爽,陈凯,等.武夷山自然保护区不同海拔甜槠天然林土壤有机碳变化特征及影响因素[J].生态学报,2021,41(13):5328-5339.
Zhao Q, Liu S, Chen K, et al. Change characteristics and influencing factors of soil organic carbon in *Castanopsis eyrei* natural forests at different altitudes in Wuyishan Nature Reserve[J]. Acta Ecologica Sinica, 2021,41(13):5328-5339.

[14] Breiman L. Random forests[J]. Machine Learning, 2001,45(1):5-32.

[15] 韩杏杏,陈杰,王海洋,等.基于随机森林模型的耕地表层土壤有机质含量空间预测①:以河南省辉县市为例[J].土壤,2019,51(1):152-159.
Han X X, Chen J, Wang H Y, et al. Spatial prediction of SOM content in topsoil based on random forest algorithm: A case study of Huixian city, Henan Province[J]. Soils, 2019,51(1):152-159.

[16] Liaw A, Wiener M. Classification and regression by RandomForests[J]. R News, 2002,2(3):18-22.

[17] 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.

[18] 叶淼,朱琳,刘旭东,等.基于连续小波变换、SHAP 和 XGBoost 的土壤有机质含量高光谱反演[J/OL].环境科学:1-19[2024-02-18].
Ye M, Zhu L, Liu X D, et al. Hyperspectral inversion of soil organic matter content based on continuous wavelet transform, SHAP, and XGBoost[J/OL]. Environmental Science:1-19[2024-02-18].

[19] 傅贵,韩国强,逯峰,等.基于支持向量机回归的短时交通流预测模型[J].华南理工大学学报:自然科学版,2013,41(9):71-76.
Fu G, Han G Q, Lu F, et al. Short-term traffic flow forecasting model based on support vector machine regression[J]. Journal of South China University of Technology (Natural Science Edition), 2013,41(9):71-76.

[20] 沈飞龙.基于机器学习与光谱信息的土壤铁氧化物估算模型研究[D].武汉:华中农业大学,2022.
Shen F L. A Model for Soil Iron Oxides Estimation Based on Machine Learning and Spectral Information[D].Wuhan: Huazhong Agricultural University, 2022.

[21] Zhao J L, Vanmaercke M, Chen L Q, et al. Vegetation cover and topography rather than human disturbance control gully density and sediment production on the Chinese Loess Plateau[J]. Geomorphology, 2016, 274:92-105.

[22] 丁倩,张弛.基于地理探测器的中国陆地生态系统土壤有机碳空间异质性影响因子分析[J].生态环境学报,2021,30(1):19-28.
Ding Q, Zhang C. Influential factors analysis for spatial heterogeneity of soil organic carbon in Chinese terrestrial ecosystem with geographical detector[J]. Ecology and Environmental Sciences, 2021,30(1):19-28.

(下转第 191 页)

[24] Enoki T, Kawaguchi H, Iwatsubo G. Topographic variations of soil properties and stand structure in a *Pinus thunbergii* plantation[J]. Ecological Research, 1996,11(3):299-309.

[25] 马风云,李新荣,张景光,等.沙坡头人工固沙植被土壤水分空间异质性[J].应用生态学报,2006,17(5):789-795.
Ma F Y, Li X R, Zhang J G, et al. Spatial heterogeneity of soil moisture in Shapotou sand-fixing artificial vegetation area[J]. Chinese Journal of Applied Ecology, 2006,17(5):789-795.

[26] 龙利群,李新荣.微生物结皮对两种一年生植物种子萌发和出苗的影响[J].中国沙漠,2002,22(6):581-585.
Long L Q, Li X R. Effect of soil microbiotic crust on seed germination and emergence of two annual herb species: *Bassia dasyphlla* and *Eragrostics poaeoides*[J]. Journal of Desert Research, 2002,22(6):581-585.

[27] 贾海坤,刘颖慧,徐霞,等.皇甫川流域柠条林地水分动态模拟:坡度、坡向、植被密度与土壤水分的关系[J].植物生态学报,2005,29(6):910-917.
Jia H K, Liu Y H, Xu X, et al. Simulation of soil water dynamics in a caragana intermedia woodland in huangfuchuan watershed: Relationships among slope, aspect, plant density and soil water content[J]. Chinese Journal of Plant Ecology, 2005,29(6):910-917.

[28] 张雪皎,高贤明,吉成均,等.中国北方 5 种栎属树木多度分布及其对未来气候变化的响应[J].植物生态学报,2019,43(9):774-782.
Zhang X J, Gao X M, Ji C J, et al. Response of abundance distribution of five species of *Quercus* to climate change in Northern China[J]. Chinese Journal of Plant Ecology, 2019,43(9):774-782.

[29] Wang A, Goslee S C, Miller D A, et al. Topographic variables improve climatic models of forage species abundance in the northeastern United States[J]. Applied Vegetation Science, 2017,20(1):84-93.

[30] Sainani K L. Multivariate regression: The pitfalls of automated variable selection[J]. Pm & R,2013,5(9):791-794.

(上接第 181 页)

[23] 国佳欣,朱青,赵小敏,等.不同土地利用类型下土壤有机碳含量的高光谱反演[J].应用生态学报,2020,31(3):863-871.
Guo J X, Zhu Q, Zhao X M, et al. Hyper-spectral inversion of soil organic carbon content under different land use types[J]. Chinese Journal of Applied Ecology, 2020,31(3):863-871.

[24] 李龙,秦富仓,姜丽娜,等.土地利用方式和地形对半干旱区土壤有机碳含量的影响[J].土壤,2019,51(2):406-412.
Li L, Qin F C, Jiang L N, et al. Effects of land use type and terrain on soil organic carbon (SOC) content in semi-arid region[J]. Soils, 2019,51(2):406-412.

[25] 唐海涛,孟祥添,苏循新,等.基于 CARS 算法的不同类型土壤有机质高光谱预测[J].农业工程学报,2021,37(2):105-113.
Tang H T, Meng X T, Su X X, et al. Hyperspectral prediction on soil organic matter of different types using CARS algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021,37(2):105-113.

[26] 王富华,黄容,高明,等.生物质炭与秸秆配施对紫色土团聚体中有机碳含量的影响[J].土壤学报,2019,56(4):929-939.
Wang F H, Huang R, Gao M, et al. Effect of combined application of biochar and straw on organic carbon content in purple soil aggregates[J]. Acta Pedologica Sinica, 2019,56(4):929-939.

[27] 张祎,李鹏,肖列,等.黄土高原丘陵区地形和土地利用对土壤有机碳的影响[J].土壤学报,2019,56(5):1140-1150.
Zhang Y, Li P, Xiao L, et al. Effects of topography and land use on soil organic carbon in hilly region of Loess Plateau[J]. Acta Pedologica Sinica, 2019,56(5):1140-1150.

[28] 郭治兴,袁宇志,郭颖,等.基于地形因子的土壤有机碳最优估算模型[J].土壤学报,2017,54(2):331-343.
Guo Z X, Yuan Y Z, Guo Y, et al. Optimal estimation model of soil organic carbon based on the terrain factor[J]. Acta Pedologica Sinica, 2017,54(2):331-343.

[29] 王绍强,周成虎.中国陆地土壤有机碳库的估算[J].地理研究,1999,18(4):349-356.
Wang S Q, Zhou C H. Estimating soil carbon reservior of terrestrial ecosystem in China[J]. Geographical Research, 1999,18(4):349-356.

[30] 周国模,姜培坤.不同植被恢复对侵蚀型红壤活性碳库的影响[J].水土保持学报,2004,18(6):68-70,83.
Zhou G M, Jiang P K. Changes in active organic carbon of erosion red soil by vegetation recovery[J]. Journal of Soil Water Conservation, 2004,18(6):68-70,83.