

## 南亚热带典型岩溶区耕地土壤有机质空间分布预测的方法比较<sup>①</sup>

任 艳, 尹秋月\*, 尹晋磊, 王 钰, 李 鑫

(广西壮族自治区国土测绘院, 南宁 530023)

**摘 要:** 本研究以隆安县测土配方施肥采集的土壤有机质含量数据为基础, 并划分岩溶区和受岩溶区影响的非岩溶区, 使用 SPSS 的相关性分析和 R 语言的随机森林对环境变量进行筛选, 构建了普通克里格(OK)、反距离加权(IDW)、线性回归(LR)、回归克里格(RK)、地理加权回归(GWR)、地理加权回归克里格(GWRK)、随机森林(RF)、随机森林克里格(RFK)等 8 种数字土壤制图模型, 对比不同模型对土壤有机质空间预测的性能差异, 探讨该区域土壤有机质空间分布预测方法。结果表明: ①分区域预测后, 岩溶区模型精度明显高于非岩溶区; ②通过环境变量重要性排序, 发现土壤类型、土地利用类型、有机肥施用量、排水能力、高程、多年降水、灌溉能力等是影响岩溶区土壤有机质空间分布的主要环境变量; ③岩溶区预测模型整体精度由高到低排序为 RFK>RF>RK>GWRK>GWR>IDW>LR>OK, RFK 预测模型的精度最高,  $R^2$  为 0.572; 非岩溶区预测模型整体精度由高到低排序为 RK>RFK>RF>LR>OK>GWRK>GWR>IDW, RK 预测模型的精度最高,  $R^2$  为 0.439; ④预测模型 LR、GWR 和 RF 的残差进行普通克里格插值后, 模型精度得到一定程度提高, 表明土壤有机质在空间上存在一定的空间相关性。研究结果可为南亚热带岩溶区土壤有机质高精度制图提供技术参考。

**关键词:** 土壤有机质; 南亚热带岩溶区; 地统计学; 随机森林; 空间分布

**中图分类号:** S158.9 **文献标志码:** A

## Comparison of Predictive Methods for Spatial Distribution of SOM in Cultivated Land in the South Subtropical Karst Region

REN Yan, YIN Qiuyue\*, YIN Jinlei, WANG Yu, LI Xin

(Guangxi Bureau of Surveying and Mapping, Nanning 530023, China)

**Abstract:** In this study, the performances of different models were compared in spatial prediction of soil organic matter (SOM) in a karst region of south subtropical China. SOM content data were collected from the soil testing and formula fertilization in Longan County, and divided into karst areas and non-karst areas affected by karst areas. SPSS correlation analysis and R language random forest were used to screen environmental variables, eight kinds of digital soil mapping models were constructed and used for SOM prediction, including ordinary kriging (OK), inverse distance weighting (IDW), linear regression (LR), regression kriging (RK), geographical weighted regression (GWR), geographical weighted regression kriging (GWRK), random forest and random forest kriging (RFK). The results showed that: 1) After the partition of the karst and no-karst areas, the prediction accuracy of karst region was obviously higher than that of non-karst region; 2) By ranking the importance of environmental variables, it was found that soil type, land use type, organic fertilizer dosage, drainage capacity, long-term precipitation, elevation, and irrigation capacity were the main environmental variable factors affecting the spatial distribution of SOM in karst areas; 3) The overall prediction accuracy from high to low was ranked as RFK>RF>RK>GWRK>GWR>IDW>LR>OK for the karst areas, with the  $R^2=0.572$  for RFK; while as RK>RFK>RF>LR>OK>GWRK>GWR>IDW of the non-karst areas, with the  $R^2=0.439$  for RK; 4) The prediction accuracies of LR, GWR and RF were improved to some extent after ordinary kriging interpolation of the residuals, indicating that SOM had a certain spatial correlation in space. The above results can provide technical reference for the high precision mapping of SOM in karst area of south subtropical regions.

**Key words:** Soil organic matter; Subtropical karst region; Geostatistics; Random forest; Spatial distribution

①基金项目: 广西耕地质量等级成果补充完善项目(GXKJ20133073-E)资助。

\* 通信作者(yue1018@126.com)

作者简介: 任艳(1988—), 女, 四川广元人, 硕士, 高级工程师, 主要从事土地调查评价和数字土壤制图。E-mail: 370119606@qq.com

土壤有机质是维持土壤肥力和提高农业生产力的重要组成部分,对于农业可持续生产有着重要意义<sup>[1-2]</sup>。精确的有机质含量分布图是精细农业施肥、耕地质量建设、农田土壤固碳减排核算等工作的基础<sup>[3]</sup>。传统的土壤调查只能获得离散点位上的有机质含量,需要通过数字土壤制图的方式根据离散点位的有机质含量信息,预测未采样点的土壤有机质含量,获得有机质含量空间分布图,以指导农业生产<sup>[4]</sup>。

近几十年来,地统计学方法被普遍用来描述和预测土壤数据的空间分布,已有的研究表明地统计学是研究土壤空间分布的有效方法<sup>[5]</sup>,普通克里格(OK)因其操作简便而成为最广泛使用的地统计学方法之一,但该方法对样点数量、样点分布要求严格,数据需满足地统计学相关假设,且未考虑到辅助变量对土壤属性的影响,其成图只能描述土壤属性的整体空间分布规律,对局部信息描述不够清晰<sup>[6-11]</sup>。反距离权重法(IDW)是基于相近相似原理<sup>[12]</sup>,使用预测位置周围的测量值预测未采样位置的值,该方法同样未考虑辅助变量对土壤属性的影响。基于此,利用辅助变量进行土壤属性预测的方法逐渐发展。线性回归(LR)、回归克里格(RK)等综合多个变量对土壤属性进行空间插值,有效提高了预测精度,但土壤属性的强空间变异性使得全局模型无法很好地体现土壤属性的局部特征<sup>[13]</sup>。地理加权回归(GWR)作为一种局部回归模型,能充分解释土壤属性的局部空间关系和空间异质性<sup>[14]</sup>。地理加权回归克里格法(GWRK)将 GWR 与 RK 相结合,不但能体现土壤属性的局部特征,还有效反映土壤属性的空间自相关性,目前应用广泛<sup>[9,15-16]</sup>。然而上述方法均是对变量的线性估计,土壤属性与环境变量间存在复杂的非线性关系<sup>[17]</sup>,随机森林(RF)作为常用机器学习方法可有效解决土壤属性与环境变量之间的非线性问题,逐渐应用到土壤属性预测中,但受到环境变量选择的限制,对土壤属性精细的差异不能很好体现<sup>[18-21]</sup>。

各种数字制图模型各具特点,对于亚热带岩溶区,高温高湿,地理条件特殊,如何选择模型及环境辅助变量来提高有机质含量预测精度是一个值得研究的课题。本文以典型亚热带岩溶区为例,通过探索该区域不同类型环境变量,利用 OK、RK、随机森林克里格(RFK)等 8 种预测模型对研究区土壤有机质含量进行空间分布预测制图,比较不同模型预测精度及差异,为亚热带岩溶区精细制图方法的选取提供参考。同时,探索研究区域土壤有机质含量空间分布特征,为土壤属性空间变异研究提供依据。

## 1 材料与方法

### 1.1 研究区概况

研究区广西隆安县位于北回归线以南(107°18'E ~ 108°4'E, 22°51'N ~ 23°21'N),根据隆安县母岩母质图中石灰岩分布范围和 LandSat 遥感影像将隆安县划分为岩溶区和非岩溶区(图 1)。该研究区地势由东北、西和西南部向中部右江河谷逐渐倾斜;东北部为砂页岩丘陵和峰林石山组成,海拔为 200 ~ 700 m;西南面以石灰岩峰丛洼地、峰丛谷地为主,海拔高度为 300 ~ 700 m;中部为右江河谷,海拔高度多在 200 m 以下,属亚热带季风气候的典型桂西南岩溶丘陵区。年平均降水量 1 301 mm,年内降水量分配不均匀,季节性较强,降水多集中在夏季。多年平均气温 21.8 °C,多年平均最高月气温 28.4 °C,最低月平均气温 13.2 °C。

### 1.2 数据获取

**1.2.1 样点数据获取** 研究区耕地表层土壤有机质含量数据来自在研究区实施的“国家测土配方补贴资金项目”,2007—2009 年共完成采样任务 5 441 个,采样深度 0 ~ 20 cm,采集土壤样本的同时,记录采样点经纬度、高程、土壤类型、地貌类型等信息。剔除异常值及明显错误信息后,研究区有效样点 4 013 个,通过 GIS 软件对样点数据集进行均匀随机划分,获得训练集数据 3 211 个(80%),其中岩溶区样点 1 582 个,非岩溶区样点 1 629 个;验证集数据 802 个(20%),其中,岩溶区样点 395 个,非岩溶区样点 407 个。

**1.2.2 环境变量数据获取** 为全面分析亚热带岩溶区土壤有机质空间分布的环境变量影响情况。初步选择与土壤形成和有机质积累相关的 22 个因子作为研究区有机质含量空间预测的环境变量,本研究拟选取的环境变量信息及来源见表 1。

### 1.3 主要模型方法

**1.3.1 普通克里格** 普通克里格(OK)是一种随机性局部插值法,用已知点的样本估计未知点的值,考虑变量的随机性。OK 模型着重于空间自相关因素,用拟合的半变异直接进行插值。可表达为:

$$z = \sum_{i=1}^n Z_i W_i \quad (1)$$

式中:  $z$  为未知点的估计;  $Z_i$  为  $i$  点的已知值;  $W_i$  为  $i$  点的权重;  $n$  为用于估算的已知点数。权重不仅与估算点和已知点之间的半变异有关,还与已知点之间的半变异有关。

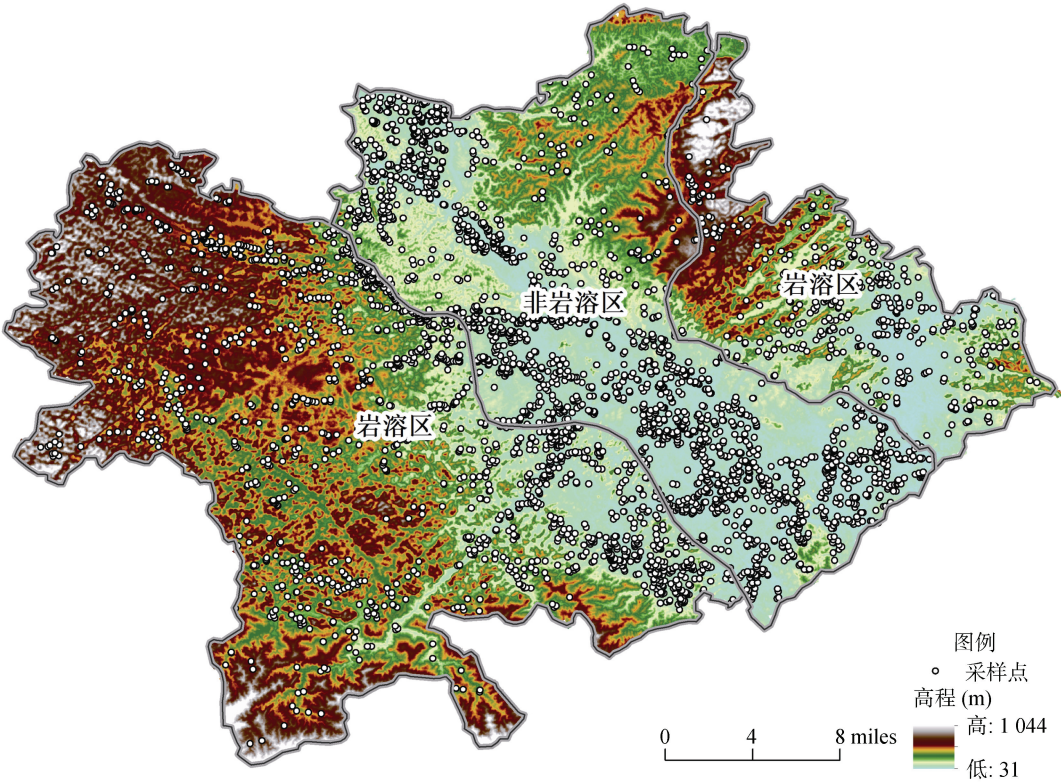


图 1 研究区高程与样点分布  
Fig. 1 DEM and sampling sites in study area

表 1 环境变量数据及来源  
Table 1 Data and sources of environmental variables selected

环境变量		来源
土壤因素	土壤类型	第二次全国土壤普查
气候因素	多年气温、多年降水	中国气象数据网
地形因素	高程、坡度、坡向、剖面曲率、平面曲率、曲率、地形湿度指数(TWI)、侵蚀模数	地理空间数据云平台
生物因素	归一化植被指数(NDVI)、比值植被指数(RVI)、增强植被指数(EVI)、差值指标指数(DVI)	TM 30 m 影像提取, 计算 2007—2009 3 年 4 个季度的总体平均值
人为因素	土地利用类型、灌溉能力、排水能力、氮肥施用量、化肥施用量、有机肥施用量	自然资源部门、测土配方数据
位置因素	与居民点距离	根据居民地(点)数据分析生成

**1.3.2 回归克里格** 回归克里格(RK)模型是线性回归模型(LR)与 OK 模型的结合体,前者用来模拟解释变量与因变量的回归关系,即趋势项分布;后者用来探测观察值与预测值之差,即残差项的空间变异<sup>[15]</sup>。模型有两个特点,一是综合了各自变量对因

变量的解释作用,二是在残差存在空间自相关的情况下对随机性变异进行解释。基本思路是:通过建立因变量与自变量之间的“最优”线性回归关系,得到代表确定性部分的趋势项与代表随机性部分的残差项,在残差存在空间自相关的情况下使用 OK 法对残差插值,最后将二者相加得到回归克里格插值结果<sup>[15]</sup>。

**1.3.3 地理加权回归克里格** 地理加权回归克里格法(GWRK)是在地理加权回归(GWR)基础上扩展得到。GWR 是基于普通最小二乘法原理的一种面向地理空间的建模方法,其特点在于将空间位置嵌入回归参数中,它在研究区域的每一处产生一个描述局部关系的回归模型。GWRK 是对局部模型 GWR 拟合后得到的残差进行 OK 插值,该方法同时考虑了目标变量的空间异质性和空间自相关性<sup>[15-16]</sup>。基本思路是:通过对目标变量和辅助变量进行 GWR 拟合,得到局部回归的残差项,然后使用 OK 对所得残差项进行插值,将 GWR 预测值和 OK 插值误差相加得到预测结果。

**1.3.4 随机森林** 随机森林(RF)是 Breiman<sup>[22]</sup>提出的基于决策树的分类回归算法,它利用 bootstrap 重抽样方法从原始数据集中随机抽取多个样本,对每

个样本数据进行决策树建模,然后组合成多棵决策树的预测模型,综合所有决策树的投票结果得到模型的预测值。基本思路是:从总样本中有放回地随机选出  $n$  个样本构建回归树,余下的样本作为袋外数据,袋外数据用来评价模型性能和自变量的重要性,通过投票或取平均值的方法得到回归树最终预测结果<sup>[23]</sup>。

**1.3.5 随机森林克里格** 随机森林克里格(RFK)是一种结合了 RF 与 OK 的混合地统计方法,它考虑了变量间的非线性关系及目标变量的空间自相关性,提高了预测精度<sup>[6,21]</sup>。通过 RF 算法得到的预测值与观测值之间存在残差,如果这些残差存在空间自相关性,可将残差当作随机变量,通过 OK 插值提高对未知点的预测精度。基本思路是:首先使用 RF 得到研究区有机质的预测值,计算 RF 预测残差,然后对预测残差进行 OK 插值,最后将 RF 预测值和 OK 插值误差相加得到预测结果。

1.4 环境变量筛选

全局或局部线性回归模型,采用 SPSS 22 的相关性分析筛选出与土壤有机质显著相关的影响因子,使用逐步线性回归和方差膨胀因子(VIF)对环境变量之间的共线性进行检验。若  $VIF>7.5$ ,则移除相应的环境变量。随机森林模型,采用变量重要性函数 importance()对解释变量进行筛选,将岩溶区重要性较弱的多年气温、曲率、平面曲率、坡向等变量剔除,非岩溶区重要性较弱的多年气温、EVI、RVI、NDVI 等变量剔除。

1.5 预测模型精度评价

本研究选用决定系数( $R^2$ )、均方根误差(RMSE)

和平均绝对误差(MAE)3 个指标对研究区耕地表层有机质空间预测输出结果的精度以及所选 8 种模型的性能表现进行评价。

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

(2)

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

(3)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

(4)

式中:  $n$  代表采样点数量;  $\hat{y}_i$  代表采样点  $i$  的预测值;  $y_i$  代表采样点  $i$  的实测值,  $\bar{y}$  代表土壤样点实测值的平均值。

2 结果与分析

2.1 土壤有机质含量基本统计

由土壤有机质含量的基本统计特征可知(表 2),全部数据集的有机质含量均值为 35.94 g/kg,范围为 7.6~101.9 g/kg,最大和最小值间差异显著。按照变异系数的划分等级:变异系数<10% 为弱变异,10%~100% 为中等变异性、>100% 为强变异性<sup>[24]</sup>,岩溶区和非岩溶区均属中等变异。说明土壤有机质在该研究区范围内具有一定的变异性,适合空间局部估计。岩溶区训练集变异系数为 41.82%,Moran I 指数为 0.65;非岩溶区训练集变异系数为 34.72%,Moran I 指数为 0.23。空间相关性明显,岩溶区变异程度和空间相关性均强于非岩溶区。

表 2 土壤有机质含量的基本统计特征  
Table 2 Basic statistical characteristics of SOM content

数据集	样本数量	平均值±标准差 (g/kg)	最小值 (g/kg)	最大值 (g/kg)	变异系数 (%)	莫兰指数 Moran I
全部数据	4 013	35.94 ± 13.90	7.6	101.9	38.68	0.51
岩溶区训练集	1 582	35.70 ± 14.93	7.6	101.9	41.82	0.65
岩溶区验证集	395	36.18 ± 15.77	9.1	94.6	43.59	0.62
非岩溶区训练集	1 629	36.17 ± 12.56	8.4	88.90	34.72	0.23
非岩溶区验证集	407	35.68 ± 12.91	9.4	82.6	36.18	0.15

2.2 土壤有机质含量与环境变量的相关性

根据土壤有机质与环境变量的相关性分析结果,Pearson 相关性越大表示越相关。如表 3 所示,岩溶区,与土壤有机质相关性显著的有土壤类型、土地利用类型、排水能力、灌溉能力、有机肥施用量、与居民点距离、坡度、化肥施用总量、剖面曲率、侵蚀模数、DVI、RVI、TWI、EVI、高程、多年降水、氮肥

施用量;非岩溶区,与土壤有机质相关性显著的有土壤类型、氮肥施用量、土地利用类型、化肥施用总量、排水能力、灌溉能力、与居民点距离、有机肥施用量、高程、DVI、TWI、EVI。

2.3 预测模型构建及模型精度分析

OK 预测模型在 ArcGIS 10.2 软件中构建,因 OK 模型要求数据符合正态分布,通过剔除部分数据和对



表 3 土壤有机质含量与环境变量相关性统计  
Table 3 Correlation coefficients between SOM content and environmental variables

全域		岩溶区		非岩溶区	
环境变量	Pearson 相关性	环境变量	Pearson 相关性	环境变量	Pearson 相关性
土壤类型	0.461**	土壤类型	0.553**	土壤类型	0.344**
土地利用类型	0.356**	土地利用类型	0.457**	氮肥施用量	0.291**
排水能力	-0.283**	排水能力	-0.372**	土地利用类型	0.284**
灌溉能力	0.258**	灌溉能力	0.305**	化肥施用总量	0.250**
有机肥施用量	-0.209**	有机肥施用量	-0.265**	排水能力	-0.199**
化肥施用总量	0.175**	与居民点距离	-0.193**	灌溉能力	0.190**
与居民点距离	-0.173**	坡度	-0.136**	与居民点距离	-0.142**
氮肥施用量	0.141**	化肥施用总量	0.133**	有机肥施用量	-0.137**
坡度	-0.107**	剖面曲率	-0.107**	高程	0.082**
DVI	-0.083**	侵蚀模数	0.090**	DVI	-0.073**
剖面曲率	-0.080**	DVI	-0.089**	TWI	0.073**
TWI	0.073**	RVI	-0.088**	EVI	-0.053*
EVI	-0.061**	TWI	0.073**	坡向	0.023
RVI	-0.060**	EVI	-0.065**	平面曲率	0.007
高程	0.046**	高程	0.063*	多年降水	0.006
多年降水	0.046**	多年降水	0.061*	曲率	0.001
侵蚀模数	0.041*	氮肥施用量	0.061*	NDVI	-0.009
平面曲率	0.028	平面曲率	0.042	剖面曲率	-0.017
多年气温	0.026	多年气温	0.039	RVI	-0.023
坡向	0.019	坡向	0.016	多年气温	-0.027
曲率	-0.006	曲率	-0.011	侵蚀模数	-0.034
NDVI	-0.031	NDVI	-0.045	坡度	-0.047

注：\*、\*\*分别表示相关性达  $P<0.05$ 、 $P<0.01$  显著水平。

数变换后，数据符合正态分布，岩溶区和非岩溶区的块金效应分别为 38.8% 和 48.9%，存在中等程度空间相关性，可进行 OK 建模。

LR 预测模型中，由于土壤类型和土地利用类型为类别变量，不能直接用于回归分析，通过算术平均值变换<sup>[25]</sup>将其转为数值型，算术平均值变换是以类别自变量各水平条件下定量因变量的算术平均值来代替类别自变量的一种方法。通过 SPSS 的逐步线性回归构建回归方程，使用 ArcGIS 10.2 的地图代数和克里格插值工具实现 LR 预测。LR 模型以土壤类型、土地利用类型、高程等变量作为输入指标，岩溶区方程为： $SOM = -105.038 + \text{土壤类型} \times 0.723 + \text{土地利用类型} \times 0.537 + \text{高程} \times 0.024 - \text{有机肥施用量} \times 0.005 + \text{灌溉能力} \times 0.172 + \text{多年降水} \times 0.057 + \text{TWI} \times 0.183$ ；非岩溶区方程为： $SOM = -12.731 + \text{土壤类型} \times 0.564 + \text{氮肥施用量} \times 0.566 + \text{土地利用类型} \times 0.367 - \text{有机肥施用量} \times 0.005 + \text{高程} \times 0.024$ 。

使用 ArcGIS 10.2 的地理加权回归工具实现

GWR 模型预测，GWR 需去除局部共线性，通过去除  $VIF>2.0$  的变量剔除模型的局部共线性。使用 R 语言构建 RF 模型，RF 树节点的变量个数(mtry)和树的个数(ntree)是 RF 的两个重要参数，通过对不同个数辅助变量逐次计算袋外误差(OOB)确定出最优的 mtry 和 ntree 值，选择使用最优参数组合进行预测，岩溶区 ntree 为 700、mtry 为 4；非岩溶区 ntree 为 750，mtry 为 4。参与 RF 建模的解释变量相对重要性排序如图 2 所示，其中 IncNodePurity 越大，环境变量越重要。从图 2 可以发现，岩溶区的土壤类型、土地利用类型、有机肥施用量、排水能力、高程、多年降水、灌溉能力等是影响土壤有机质含量的主导因素。

根据 LR、GWR、RF 的真实值与预测值之差计算残差，经分析，3 个模型的残差均存在空间自相关性，对 3 个模型的残差进行 OK 插值，将插值结果与各模型预测值相加分别得到 RK、GWRK、RFK 的预测结果。不同预测模型的精度见表 4。从总体来看，岩溶区精度明显高于全域精度；全域和非岩溶区均为

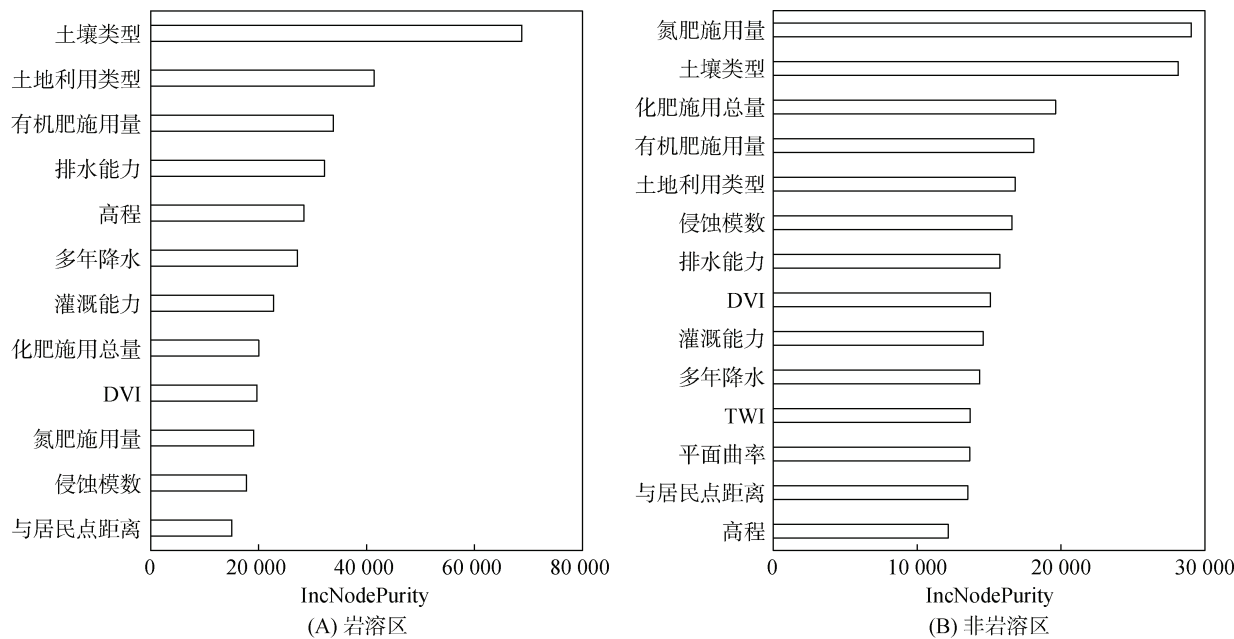


图 2 RF 预测模型中环境变量的重要性  
Fig. 2 Importance of environment variables in random forest prediction model

表 4 土壤有机质含量不同预测模型的精度				
Table 4 Accuracies of different prediction models on SOM content				
分区	模型	平均绝对误差 MAE	均方根误差 RMSE	决定系数 $R^2$
岩溶区	OK	8.365	11.298	0.382
	IDW	8.400	11.467	0.363
	LR	8.851	11.508	0.359
	RK	7.804	10.318	0.485
	GWR	8.836	11.698	0.337
	GWRK	7.978	10.946	0.420
	RF	7.523	10.052	0.434
	RFK	7.302	9.778	0.472
	OK	8.500	12.16	0.404
	IDW	8.375	11.516	0.465
	LR	8.908	12.017	0.418
	RK	7.904	10.723	0.536
	GWR	8.239	11.154	0.498
	GWRK	7.817	10.832	0.527
	RF	7.393	10.127	0.553
	RFK	7.213	9.921	0.572
非岩溶区	OK	8.250	10.68	0.315
	IDW	8.441	11.414	0.216
	LR	8.393	10.612	0.323
	RK	7.426	9.657	0.439
	GWR	8.782	11.249	0.239
	GWRK	8.385	10.919	0.283
	RF	8.341	10.409	0.380
	RFK	8.112	10.231	0.400

RK 模型精度最高,  $R^2$  分别为 0.485 和 0.439, 从面上看全域预测精度更高, 这是由于全域预测中岩溶区域拟合程度较好, 拉高了整体精度, 单看全域预测中非岩溶区域精度  $R^2$  仅为 0.402, 仍低于分区后非岩溶区精度。由此说明分区预测后, 精度得到明显提高。

岩溶区预测模型整体精度由高到低排序为 RFK>RF>RK>GWRK>GWR>IDW>LR>OK, RFK 精度最高,  $R^2$  为 0.572, 由此说明地理环境复杂的岩溶区, 更适合使用 RFK 模型。非岩溶区预测模型整体精度由高到低排序为 RK>RFK>RF>LR>OK>GWRK>GWR>IDW, RK 精度较高,  $R^2$  为 0.439, GWR 和 IDW 的  $R^2$  均小于 0.300。该非岩溶区, 样点主要集中在地势较低的河谷区域, 虽地势平坦, 但周边大部分为岩溶区, 因特殊的地理条件限制, 不建议使用 IDW 及 OK 预测模型。

### 2.4 不同预测模型绘制空间分布图

不同预测模型绘制的土壤有机质空间分布如图 3 所示, 土壤有机质分布的总体趋势基本一致, 整体上非岩溶区有机质含量水平高于岩溶区; 土壤有机质含量分布集中在 20 ~ 40 g/kg, 属于中等水平, 高值主要集中在分布在中北部的岩溶区与非岩溶区交界位置, 低值主要集中在西南部。

岩溶区, 精度最低的 OK 预测模型绘制的空间分布图出现边界明显的块状区域, 对于空间变异的细节信息描述不够详细; 精度较高的 RFK、RF、RK 及 GWRK 预测模型在局部细节描述上更清晰, 但对于

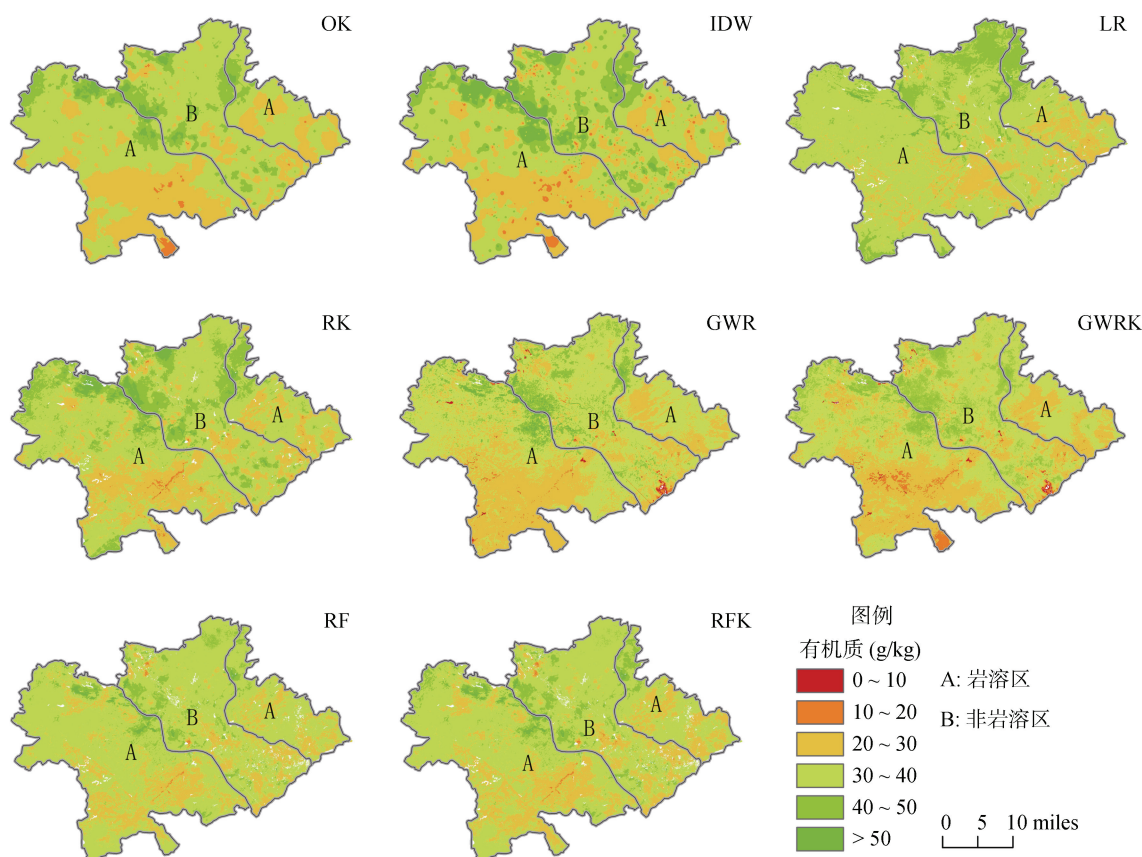


图3 不同模型预测的土壤有机质空间分布图

Fig. 3 Spatial distributions of SOM content predicted by different models

土壤有机质的低值区间,因样点少预测结果体现不明显,由此说明 RF 和 GWRK 等预测模型在采样点有限的区域表现较差。非岩溶区,精度最低 IDW 模型也出现明显的块状和点状区域。综上所述,岩溶区 RKF 预测精度最优,非岩溶区 RK 预测精度最优。将岩溶区的 RKF 预测结果与非岩溶区 RK 预测结果使用镶嵌工具合并形成土壤有机质空间分布图(图 4)。

### 3 讨论

#### 3.1 岩溶区解释变量重要性分析

根据随机森林相对重要性排序,岩溶区的土壤类型、土地利用类型、有机肥施用量、排水能力、多年降水、高程、灌溉能力等是影响土壤有机质含量的主导因素。这是由于不同成土过程形成的土壤类型以及不同的土地利用方式都会导致土壤理化性质差异较大;岩溶区域不同高程对土壤所处的气候环境具有调控作用,土壤有机质的分解和积累影响明显<sup>[26]</sup>;同时,岩溶区域受多年降水、灌溉及排水等影响,都会加剧岩溶区的溶蚀作用,导致土壤有机质在坡地流失而在低洼区域积累<sup>[27-29]</sup>。因此,相比非岩溶区,岩溶地区土壤有机质分布具有较强的空间异质性。

#### 3.2 土壤有机质空间分布及其影响因素

从土壤有机质空间总体分布来看,研究区土壤有机质含量呈现中北部高、西南部低的空间分异特征。从分区来看,非岩溶区有机质含量整体高于岩溶区。非岩溶区土壤有机质含量集中在 30 ~ 50 g/kg,北部为土山,中部和东南部为河谷地带,地理条件较好,为隆安县主要的粮食产区,受人为灌溉、施肥、秸秆还田等影响,土壤有机质含量较高,土壤较肥沃。岩溶区域,土壤有机质含量大部分区域集中在 20 ~ 40 g/kg,属于中等水平;该区域海拔相对较高,多为坡地,特别是西南部和东部区域,受到基岩出露及降雨等影响,植被覆盖度低,存在一定的水土流失和迁移,导致养分流失,因此在岩溶地区土壤有机质含量相对较低。

通过比较土壤类型、土地利用类型、排水能力、多年降水、灌溉能力及海拔等环境变量与土壤有机质的关系可知(图 5),水稻土的土壤有机质含量最高,为 39.72 g/kg,主要集中分布在中部河谷区;潮土和紫色土的土壤有机质含量较低,潮土主要为潮砂土,该土壤类型易漏水漏肥;紫色土面积不大,主要分布在非岩溶区北部,属于易风化易侵蚀的一种土壤类

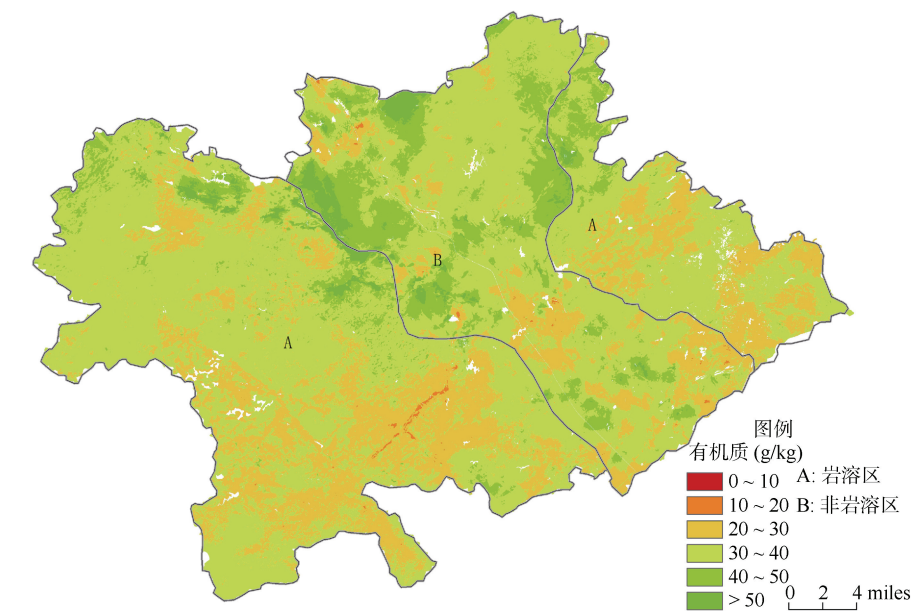


图 4 土壤有机质含量预测结果镶嵌图(岩溶区 RKF 预测, 非岩溶区 RK 预测)  
Fig. 4 Mosaic map of predicted SOM content (RKF for karst region, RK for non-karst region)

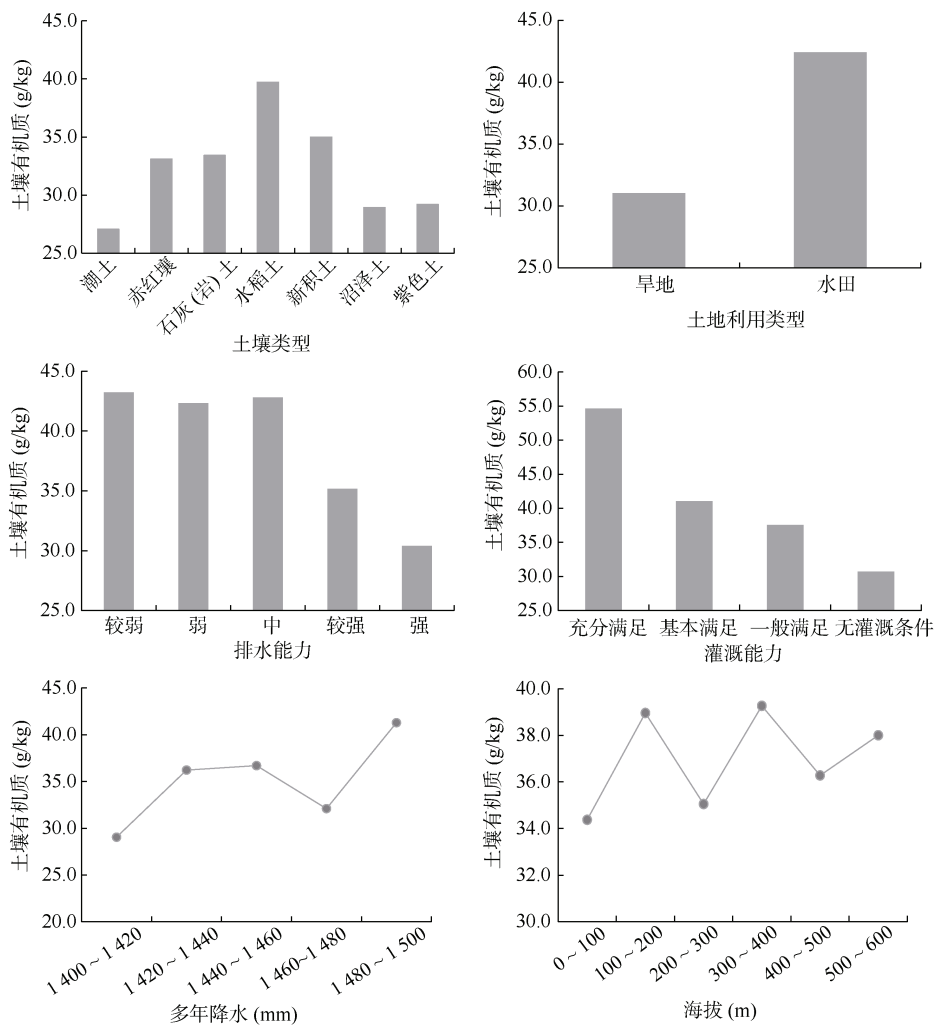


图 5 主导环境变量下土壤有机质含量  
Fig. 5 SOM contents under dominant environmental variables

型,其表土层薄、土质疏松结构性差,土壤养分易流失。水田的土壤有机质含量明显高于旱地,这主要是受稻田淹水影响,产生还原环境,有利于缓解土壤有机质矿化分解,增加土壤有机质积累<sup>[30]</sup>。土壤有机质与多年降水、灌溉能力趋于正相关,与排水能力趋于负相关,由此说明灌溉条件越好、降水量充沛、适中的排水条件,更有利于土壤有机质积累。从海拔来看,土壤有机质与海拔关系呈现M型,海拔0~100 m主要为中部的河谷区,受人为耕作影响,有机质积累与流失同步进行;海拔100~200 m为岩溶区的低洼区域以及北部岩溶区与非岩溶区交界地带,海拔300~400 m为岩溶区西北部的低洼区域,土壤有机质相对较高,这是由于岩石溶蚀过程形成的可溶性碳酸盐与土壤腐殖质结合、凝聚,形成稳定的腐殖酸钙,极有利于土壤有机质的累积<sup>[22-23]</sup>,导致土壤有机质在这些低洼区域积累;海拔200~300 m和400~500 m均为岩溶区的坡地,受岩溶区基岩出露及植被覆盖度低等特殊地理环境影响,土壤发育缓慢,形成的土层较为浅薄,土壤总量较少,相应的土壤有机质含量也较少<sup>[24]</sup>。

### 3.3 土壤有机质空间分布预测方法比较

研究认为非岩溶区虽然成土环境相对单一,但不推荐使用OK模型,这与李梦佳等<sup>[6]</sup>在不同模型预测土壤有机质含量空间分布对比分析中提到的观点不一致,一般认为成土环境相对单一区域,OK模型预测结果精度较高。由于本研究的非岩溶区大部分位于河谷,水系发达,在成土过程中受到周边岩溶区影响较大,仅依靠采样点数据进行插值的OK模型,将忽略重要环境变量对土壤有机质的影响,不推荐直接使用OK模型预测,推荐使用RK等环境变量与地统计相结合的模型预测土壤有机质空间分布。岩溶区地理环境特殊,特别是受到土壤类型、土地利用类型以及水分影响,单纯的地统计模型已不适用,推荐使用RFK与地统计相结合的模型进行预测。

从模型精度来看,岩溶区域整体预测精度明显高于非岩溶区,岩溶区最优模型RFK的 $R^2$ 为0.572,非岩溶区最优模型RK的 $R^2$ 为0.439,这可能与环境变量的选择有关。尽管本研究已收集了各类环境变量,并对环境变量进行筛选,大大降低数据冗余,但环境变量的选取还存在考虑不周全的问题。非岩溶区域作为主要的粮食产区,受人为干扰明显大于岩溶地区,如秸秆还田、种植绿肥、耕作制度等都会影响土壤有机质含量,但这类人为因素在本文中考虑较少,因此,对于非岩溶区与土壤有机质相关的环境变量还需要继续挖掘。

## 4 结论

1) 岩溶区土壤有机质的变异程度和空间相关性均强于非岩溶区。岩溶区土壤有机质主要受土壤类型、土地利用类型、有机肥施用量、排水能力、多年降水、高程、灌溉能力等因素影响。

2) 相较于全域模型,分区建模显著提升了土壤有机质的预测精度,其中岩溶区模型精度明显高于非岩溶区。对比8种模型,岩溶区RFK为最优模型, $R^2$ 为0.572;非岩溶区最优模型为RK, $R^2$ 为0.439。研究发现预测模型LR、GWR和RF的残差进行OK插值后,模型精度得到提高,表明在预测建模时需集成土壤有机质的空间自相关特征以提高制图精度。

3) 单纯的地统计模型已不适用于岩溶区,使用RFK与地统计相结合的模型可有效提高土壤有机质的预测精度;非岩溶区虽成土环境相对单一,但受周边岩溶区影响,仅使用OK模型无法很好表征土壤有机质空间分布情况,推荐使用RK与地统计相结合的模型预测。

## 参考文献:

- [1] 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.
- [2] 李丹,王京文,李凤根,等. 1982—2012年来杭州市郊水稻田有机质变化特征及其调控措施[J]. 水土保持通报, 2016, 36(3): 131-135.
- [3] Wang Y Q, Zhang X C, Zhang J L, et al. Spatial variability of soil organic carbon in a watershed on the Loess Plateau[J]. Pedosphere, 2009, 19(4): 486-495.
- [4] 王雨雪,杨柯,高秉博,等. 基于两点机器学习方法的土壤有机质空间分布预测[J]. 农业工程学报, 2022, 38(12): 65-73.
- [5] 孟博,周一帆,杨林生,等. 勐海县甘蔗土壤养分空间分布特征及肥力评价[J]. 土壤, 2022, 54(2): 277-284.
- [6] 李梦佳,王磊,刘洪斌,等. 不同模型预测土壤有机质含量空间分布对比分析[J]. 西南农业学报, 2021, 34(3): 610-617.
- [7] 陆访仪,赵永存,黄标,等. 海伦市耕层土壤有机质含量空间预测方法研究[J]. 土壤通报, 2012, 43(3): 662-667.
- [8] 石光辉,毛伟,曾洪玉,等. 扬州市江都区耕地土壤有机质35年变化特征及其影响因素分析[J]. 扬州大学学报(农业与生命科学版), 2022, 43(6): 44-50.
- [9] 杜佩颖,张海涛,郭龙,等. 平原丘陵过渡区土壤有机质空间变异及其影响因素[J]. 土壤学报, 2018, 55(5): 1286-1295.
- [10] 张欢,高小红. 复杂地形区土壤有机质空间变异性分析及制图[J]. 水土保持研究, 2020, 27(5): 93-100.
- [11] 姜赛平,张怀志,张认连,等. 基于三种空间预测模型的海南岛土壤有机质空间分布研究[J]. 土壤学报, 2018, 55(4): 1007-1017.



- [12] 马利芳, 熊黑钢, 孙迪, 等. 不同干扰程度下土壤有机质空间最优插值法研究[J]. 生态学报, 2019, 39(19): 7153–7160.
- [13] 姚雪玲, 傅伯杰, 吕一河, 等. 基于 GIS 和统计模型的黄土丘陵沟壑区土壤水分插值方法[J]. 水土保持学报, 2013, 27(6): 93–96, 102.
- [14] 王幼奇, 张兴, 赵云鹏, 等. 基于 GIS 和地理加权回归的砂田土壤阳离子交换量空间预测[J]. 土壤, 2020, 52(2): 421–426.
- [15] 杨顺华, 张海涛, 郭龙, 等. 基于回归和地理加权回归 Kriging 的土壤有机质空间插值[J]. 应用生态学报, 2015, 26(6): 1649–1656.
- [16] 李运龙, 熊立华, 闫磊. 基于地理加权回归克里金的降水数据融合及其在水文预报中的应用[J]. 长江流域资源与环境, 2017, 26(9): 1359–1368.
- [17] 李豪, 刘涛, 徐精文. 基于混合地理加权回归与克里格的区域降水量空间插值方法[J]. 中国农业气象, 2018, 39(10): 674–684.
- [18] 任丽, 杨联安, 王辉, 等. 基于随机森林的苹果区土壤有机质空间预测[J]. 干旱区资源与环境, 2018, 32(8): 141–146.
- [19] 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——以河南省辉县市为例[J]. 土壤, 2019, 51(1): 152–159.
- [20] 凌晓丹, 王罗其, 赵科理, 等. 基于随机森林法的山核桃林地土壤速效养分含量空间分布特征研究[J]. 生态学报, 2024, 44(2): 662–675.
- [21] 马重阳, 孙越琦, 巫振富, 等. 基于不同模型的区域尺度耕地表层土壤有机质空间分布预测[J]. 土壤通报, 2021, 52(6): 1261–1272.
- [22] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5–32.
- [23] Cutler A, Cutler D R, Stevens J R. Random forests[M]// Ensemble machine learning. New York: Springer, 2012: 157–175.
- [24] 秦静, 孔祥斌, 姜广辉, 等. 北京典型边缘区 25 年来土壤有机质的时空变异特征[J]. 农业工程学报, 2008, 24(3): 124–129, 314.
- [25] 胡良平. 提高回归模型拟合优度的策略(II)——算术均值变换与其他变量变换[J]. 四川精神卫生, 2019, 32(1): 9–15.
- [26] 李洪义, 贺任彬, 谢模典, 等. 自然和人为因素对耕地土壤有机质含量影响的研究[J]. 土壤通报, 2023, 54(5): 1050–1059.
- [27] 欧阳资文, 彭晚霞, 宋同清, 等. 喀斯特峰丛洼地土壤有机质的空间变化及其对干扰的响应[J]. 应用生态学报, 2009, 20(6): 1329–1336.
- [28] 邸欣月, 安显金, 董慧, 等. 贵州喀斯特区域土壤有机质的分布与演化特征[J]. 地球与环境, 2015, 43(6): 697–708.
- [29] 文鑫, 王艺惠, 钟聪, 等. 贵州表层土壤有机质空间变异特征及其影响因素分析[J]. 水土保持学报, 2023, 37(3): 218–224.
- [30] 李冬初, 黄晶, 马常宝, 等. 中国稻田土壤有机质时空变化及其驱动因素[J]. 中国农业科学, 2020, 53(12): 2410–2422.