

基于GWRF模型的县域土壤有机质数字制图方法 ——以鲁中山区为例

张树琳¹, 崔丽芹³, 刘健¹, 张灿婷¹, 王洪佳¹, 张婷婷¹, 王瑗玲^{1,2*}

(1. 山东农业大学资源与环境学院, 山东泰安 271018, 中国; 2. 山东农业大学土肥高效利用国家工程研究中心, 山东泰安 271018, 中国; 3. 沂源县农业农村局, 山东淄博 256100, 中国)

摘要: [目的/意义] 土壤有机质 (Soil Organic Matter, SOM) 是土壤肥力和质量的重要指标。为解决山区复杂地形条件下 SOM 空间变异性强、传统数字土壤制图方法预测精度受限的问题, 本研究旨在引入并验证地理加权随机森林 (Geographically Weighted Random Forest, GWRF) 模型在县域尺度复杂地形区 SOM 数字制图中的适用性与精度优势, 并实现 SOM 含量的高精度空间预测, 对合理利用和科学管理土壤资源具有重要意义。[方法] 以山东省典型山地农区沂源县为研究区, 利用实测 SOM 样点数据, 考虑气候、地形、土壤、植被和土地利用等环境变量, 通过构建融合空间局部建模与非线性建模能力的 GWRF 模型对研究区 SOM 预测、数字制图, 并与普通克里格 (Ordinary Kriging, OK)、多元线性回归 (Multiple Linear Regression, MLR)、GWR 和随机森林 (Random Forest, RF) 等模型进行精度对比。[结果和讨论] GWRF 模型表现最优, 决定系数 (R^2) 为 0.48、均方根误差 (Root Mean Square Error, RMSE) 为 5.12 g/kg, 较 OK、MLR、GWR 和 RF 模型 R^2 分别提升 0.24、0.16、0.13 和 0.07, RMSE 降低了 1.06、0.73、0.59 和 0.36 g/kg。研究区 SOM 含量整体偏低, 呈中部高、西南和东北低的空间分布特征, 主要受土壤类型、年蒸散量、坡度和砂粒含量等环境因素影响。[结论] 研究结果证明 GWRF 模型在复杂地形区的县域 SOM 预测中具有显著优势, 可为第三次全国土壤普查背景下的高精度 SOM 数字制图提供技术支撑。

关键词: 土壤有机质; 数字土壤制图; 山区; 地理加权回归; 随机森林

中图分类号: S159; TP3-0

文献标识码: A

文章编号: SA202508020

引用格式: 张树琳, 崔丽芹, 刘健, 张灿婷, 王洪佳, 张婷婷, 王瑗玲. 基于 GWRF 模型的县域土壤有机质数字制图方法——以鲁中山区为例[J]. 智慧农业(中英文), 2026, 8(2): 72-85. DOI: 10.12133/j.smartag.SA202508020

ZHANG Shulin, CUI Liqin, LIU Jian, ZHANG Canting, WANG Hongjia, ZHANG Tingting, WANG Ailing. Geographically Weighted Random Forest for County-scale Digital Mapping of Soil Organic Matter: A Case Study in the Central Shandong Mountains[J]. Smart Agriculture, 2026, 8(2): 72-85. DOI: 10.12133/j.smartag.SA202508020 (in Chinese with English abstract)

0 引言

土壤有机质 (Soil Organic Matter, SOM) 是土壤中最重要营养成分之一, 是衡量土壤肥力和质量的核心指标^[1]。精准预测区域 SOM 含量并揭示其空间分布特征, 对科学评价土壤质量、合理利用土壤资源和有效促进农业可持续发展等具有重要作用^[2,3]。数字土壤制图 (Digital Soil Mapping, DSM) 是一种高效、精准表达土壤及其属性空间分布的技

术^[4], 根据已知点的土壤信息, 以土壤-景观模型为理论基础, 以空间分析和数学方法为技术手段, 生成数字格式的土壤属性空间分布图^[5]。中国正全面推进第三次全国土壤普查工作, 亟须研究数字土壤制图方法^[6], 实现对土壤属性由点到面的精准预测。县域作为第三次全国土壤普查的基本单元, 是实现高精度土壤属性制图的关键空间尺度。SOM 含量受多种因素共同影响, 山区地形起伏大、环境变化显著, 较小区域或邻近田块的 SOM 含量差异

收稿日期: 2025-08-21

基金项目: 国家自然科学基金 (42171378); 山东省自然科学基金 (ZR2021MD018); 泰山学者青年专家项目 (tsqnz20231205)

作者简介: 张树琳, 硕士研究生, 研究方向为数字土壤制图。E-mail: 2023120274@sda.edu.cn

*通信作者: 王瑗玲, 博士, 教授, 研究方向为土地利用与信息技术。E-mail: ailingwang@sda.edu.cn

copyright©2026 by the authors

(登录 www.smartag.net.cn 免费获取电子版全文)

较大, 导致 SOM 数字制图精度较低、绘制难度大^[7]。因此, 开展山区县域 SOM 数字制图方法研究、提升制图精度成为土壤普查的重要科学问题, 对于提升普查成果的可用性、推进土壤普查工作科学开展具有重要意义。

常见的数字制图方法包括地统计学 (Geostatistics)、回归分析法和机器学习等^[8]。其中, 普通克里格 (Ordinary Kriging, OK) 模型作为地统计学方法典型代表, 因操作简单被广泛使用^[9], 但未考虑环境变量对 SOM 的影响, 不能很好地体现其空间局部特征^[10]。近年来考虑环境变量的土壤属性数字制图方法应用广泛, 如多元线性回归 (Multiple Linear Regression, MLR) 模型, 但 SOM 较强空间异质性使得全局模型难以进行较好的局部预测^[11]。地理加权回归 (Geographically Weighted Regression, GWR) 模型基于空间变化关系构建局部线性回归模型, 能够较好地刻画 SOM 的局部空间分布, 在环境复杂区可有效提高预测精度^[12]。OK、MLR 和 GWR 等模型均是对 SOM 进行线性估计, 然而环境变量和 SOM 存在复杂的非线性关系^[13]。近年来, 相关学者将机器学习模型广泛运用到 SOM 预测^[14, 15], 其中随机森林 (Random Forest, RF) 模型表现出色。但此类模型依靠统一的全局参数, 没有考虑 SOM 与环境变量之间关系的空间异质性, 导致制图结果的准确性受到一定程度的限制^[16]。如何解决上述两类制图模型的缺陷, 在土壤属性制图中既能捕捉土壤属性与环境变量间非线性关系, 又能解释土壤属性的局部效应是需要解决的问题。地理加权随机森林 (Geographically Weighted Random Forest, GWRF) 模型是在 RF 非线性建模的基础上结合 GWR 空间局部建模的机器学习模型^[17]。已有研究表明, GWRF 对土壤碳储量^[18]和土壤含盐量^[19]的预测精度较 GWR 和 RF 模型有明显的提升。山区因其复杂的环境条件导致 SOM 预测和制图难度较大, GWRF 应用于县域山地丘陵区 SOM 预测效果如何, 能否有效提高制图精度, 亟须加强研究。

基于上述问题, 本研究选取典型山区农业县——山东省沂源县为研究区, 根据 SOM 样点实测数据, 考虑气候、土壤、地形地貌、植被和土地利用等环境变量, 构建 GWRF 模型对 SOM 含量进行预测和数字制图, 并与 OK、MLR、GWR 和 RF 模型比较预测精度与制图表现, 旨在探讨复杂地形区高精度 SOM 制图方法, 为第三次全国土壤普查背

景下县域尺度 SOM 数字制图提供技术支撑。

1 数据与方法

1.1 研究区概况

沂源县地处山东省中部 (东经 117°43'00"~118°14'00", 北纬 35°59'30"~36°23'00"), 总面积约 1 636 km², 辖 2 个街道、10 个镇。地势东南低、西北高, 地形起伏显著。地貌类型主要为中、低山和丘陵, 占总面积的 99.3%。属暖温带半湿润大陆性季风气候, 年均气温 12.6 °C, 年均降水量 753.3 mm。土壤母质主要为花岗岩、石灰岩风化物及第四纪沉积物, 土壤类型以褐土和潮土为主。农业土地利用以园地和林地为主, 分别占总面积的 33.92% 和 33.04%, 此外, 耕地和草地分别占总面积的 14.58% 和 5.59%, 是典型的山区农业县。研究区地形与行政区划如图 1 所示。

1.2 数据来源与预处理

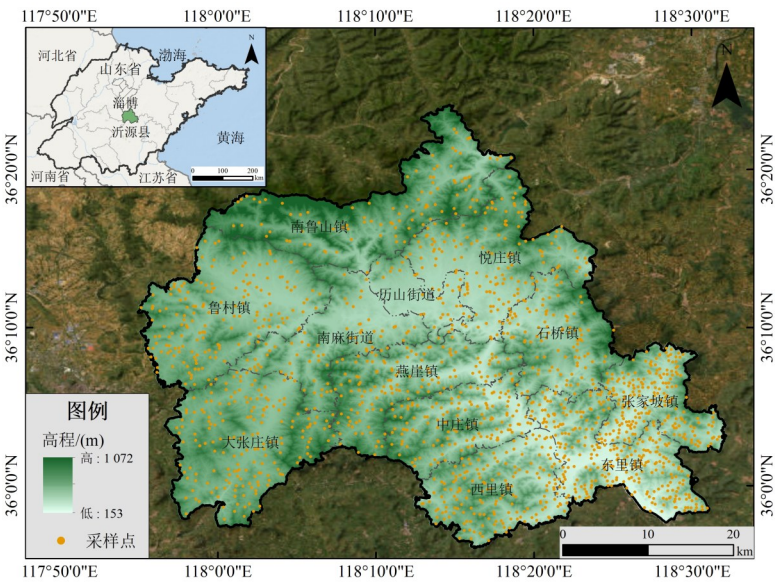
1.2.1 土壤样品采集与预处理

SOM 含量数据来源于沂源县农业农村局测土配方施肥项目土壤实测样点数据。综合考虑研究区地形、土地利用方式和土壤类型等特征布设采样点, 2020—2023 年共采集 1 574 个土壤样品, 采样深度为 0~20 cm。土壤样品经风干、去除杂质、研磨、过筛后, 采用重铬酸钾氧化-滴定法测定 SOM 含量。为避免样点采集与化学分析过程的不确定性, 对化验分析结果进行统计、检验。将每个样点的测试值与总体及其邻近 8 个样点的均值和标准差进行比较, 如果样点值与总体均值的差值超过 5 倍总体标准差, 且样点值与近邻均值的差值超过 3 倍近邻标准差^[20], 则认为该样点值异常, 需剔除。经检验, 共剔除异常值 9 个, 得到 1 565 个有效样点。

1.2.2 环境变量选取与预处理

基于土壤发生学理论, SOM 含量与成土条件、环境变量等因素关系密切^[21]。本研究结合研究区实际和数据的可获取性, 初步选取气候、土壤、地形地貌、植被和土地利用 5 类 19 个变量作为预测 SOM 含量的环境变量。环境变量信息及来源见表 1。

本研究中对于类别变量, 通过算术平均值变换将其转为数值型^[22]。同时, 为消除不同变量量纲影响, 对所有变量进行极差标准化处理。为方便后续进入制图模型, 所有环境变量的空间分辨率均重



注:该图基于自然资源部标准地图服务网站下载的审图号为GS(2019)1822号标准地图制作,底图无修改。

图1 山东省沂源县地形与采样点分布

Fig.1 Topography and distribution of sampling points in Yiyuan county, Shandong province

表1 SOM制图模型的环境变量及来源

Table 1 Environmental variables and data sources for the SOM prediction model

类别	名称	来源	分辨率/m
气候	年均温 (Mean Annual Temperature, MAT)	《1901—2023年中国1 km分辨率逐月平均气温、降水量、潜在蒸发散量数据集》(国家地球系统科学数据中心, https://www.geodata.cn/)	1 000
	年降水 (Mean Annual Precipitation, MAP)		
	年蒸散 (Mean Annual Evapotranspiration, MAE)		
土壤	土壤类型 (Soil Type, ST)	全国第二次土壤普查土壤类型图	—
	黏粒含量 (Clay)	《中国高分辨率国家土壤信息网格基本属性数据集_90米土壤砂粒、粉粒、黏粒含量》(国家地球系统科学数据中心, https://www.geodata.cn/)	90
	粉粒含量 (Silt)		
	砂粒含量 (Sand)		
	高程 (DEM)	ASTER GDEM 30 m 数据 (地理空间数据云, https://www.gscloud.cn/)	30
	坡度 (Slope)		
	坡向 (Aspect)		
地形	平面曲率 (Plan)		
	剖面曲率 (Profile)		
	地形位置指数 (Topographic Position Index, TPI)	《Landsat 8-9 OLI/TIRS C2 L2》 (地理空间数据云, https://www.gscloud.cn/)	30
	地形湿度指数 (Topographic Wetness Index, TWI)		
	径流强度指数 (Stream Power Index, SPI)		
植被	归一化植被指数 (Normalized Difference Vegetation Index, NDVI)	《Landsat 8-9 OLI/TIRS C2 L2》 (地理空间数据云, https://www.gscloud.cn/)	30
	增强型植被指数 (Enhanced Vegetation Index, EVI)		
	土壤调整植被指数 (Soil-Adjusted Vegetation Index, SAVI)		
土地利用	土地利用方式 (Land Use, LU)	国土变更调查数据	—

采样至 30 m，统一坐标系为 2000 大地坐标系。

1.3 研究方法

1.3.1 制图模型构建

GWRF 模型是在 RF 模型的非线性建模基础上，引入 GWR 的空间局部建模思想，通过构建空间加权的局部 RF 子模型，实现了对非线性关系与空间异质性的同步建模^[23-25]。GWRF 模型以目标预测点为中心，定义其邻域范围，并依据空间距离计算邻域内样本的权重，进而构建预测点的局部 RF 子模型^[26, 27]。

样本权重基于预测点与样本点的空间距离采用高斯核函数计算得出，如公式 (1) 所示。

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{bw}\right)^2\right) \quad (1)$$

式中： w_{ij} 为样本点 j 对预测点 i 的空间权重； d_{ij}

为二者间的欧氏距离； bw 为带宽，控制邻域范围大小。

GWRF 模型的表达式如公式 (2) 所示。

$$Y_i(\mu_i, v_i) = f(x_i, (\mu_i, v_i)) + \varepsilon_i \quad (2)$$

式中： Y_i 为位置 i 的预测值； x_i 为该位置的环境变量集； $f(\cdot)$ 为局部随机森林模型； (μ_i, v_i) 为该点坐标； ε_i 为误差项。

本研究中 GWRF 模型在 R 语言环境下使用“SpatialML”包实现，具体流程包括：1) 提取 SOM 样点数据与空间位置，叠加多源环境变量数据；2) 在每个预测位置计算其与训练样点的空间距离，并通过高斯核函数确定样本权重；3) 基于加权样本数据，在每个预测位置构建局部 RF 模型；4) 输出每个预测点的预测值并最终生成研究区 SOM 含量空间分布图。GWRF 模型的建模流程如图 2 所示。

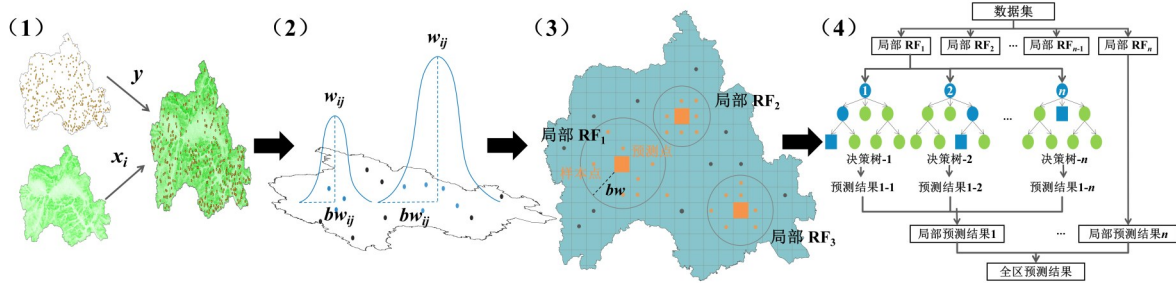


图 2 GWRF 建模流程

Fig. 2 Modeling workflow of the GWRF model

GWRF 模型中，需要调整的参数包括：带宽和局部 RF 模型的“ntree”和“mtry”。其中，GWRF 模型带宽的设定方式有固定带宽和自适应带宽两类，前者使用固定半径定义邻域，后者根据设定的邻近样本数量动态调整邻域大小，更适用于样本空间分布不均的区域^[28]。鉴于研究区地形起伏大、样点分布不均，本研究采用自适应带宽方式。ntree 代表每个局部 RF 中决策树的数量，默认设置为 500；mtry 代表每棵决策树节点分裂时选用的随机特征变量的数量，通过网格搜索结合交叉验证的方法确定最优 mtry。

选取 OK、MLR、GWR 和 RF 模型与 GWRF 模型进行对比分析，比较不同制图模型的预测精度与空间表现。OK 模型基于空间自相关性原理和变异函数理论，利用空间邻近性实现无偏线性估计^[29]，通过 ArcGIS 软件的统计模块构建 OK 模型。MLR 模型通过构建 SOM 与环境变量间的线性关系进行建模，利用 SPSS 软件的线性回归构建回归方程。

GWR 模型引入空间权重，在每个地理位置上构建局部回归模型。在 ArcGIS 中选用改进的赤池信息量准则 (corrected Akaike Information Criterion, AICc) 确定最优带宽^[30]，构建局部回归模型。RF 模型作为一种集成学习算法，构建多个决策树进行集成预测，每个决策树基于训练数据集中随机样本进行训练，决策树节点依赖随机重采样技术 bootstrap 生成；集成多棵回归树形成随机森林，并以所有决策树预测结果的平均值作为最终预测结果^[31]。本研究 RF 模型在 R 环境中的“randomForest”包中实现。

1.3.2 环境变量筛选

环境变量的准确选取是高效预测、提高模型精度的关键。利用不同模型对 SOM 含量进行预测建模时，过多的环境变量不仅不利于提高模型的预测精度，还可能降低模型的计算效率。因此，对环境变量进行筛选，以消除冗余变量，提升制图模型的预测性能。采用相关性分析筛选出与 SOM 相关的

环境变量，并对环境变量之间的共线性进行检验，剔除存在共线性的变量。具体流程如下。

1) 相关变量筛选。采用皮尔逊相关系数 (Pearson Correlation Coefficient, PCC) 分析 SOM 含量与各环境变量的相关性，相关系数的取值范围为 -1~1，绝对值越大表示相关性越强^[32]。本研究在 SPSS 软件中计算皮尔逊相关系数，保留与 SOM 呈显著相关的变量。

2) 共线性分析。通过逐步线性回归和方差膨胀因子 (Variance Inflation Factor, VIF) 检验环境变量间的多重共线性。VIF 值越大表明共线性越严重。本研究通过构建 SOM 与环境变量的最小二乘法 (Ordinary Least Square, OLS) 模型计算 VIF，逐一剔除 VIF>7.5 的环境变量，以减弱多重共线性对模型稳定性的影响。

1.3.3 制图精度检验

本研究对 SOM 制图模型精度采用独立验证点进行检验。按照 4 : 1 的比例划分训练集和验证集：随机选取 80% 的样点作为训练集，用于训练制图模型；选取 20% 的样点作为验证集，用于验证制图精度。考虑单次随机划分可能导致精度评价的不确定性，共进行 20 次随机划分与建模实验，分别计算每次的预测精度评估指标，并取 20 次结果的平均值来评估模型性能。选用决定系数 (Coefficient of Determination, R^2) 和均方根误差 (Root Mean Square Error, RMSE) 两个指标来评价各模型预测精度。 R^2 用以衡量回归模型中自变量对因变量变异解释程度，其值为 0~1，值越大表示模型对数据的拟合效果越好。RMSE 是实测值和预测值之间误差

的平方和的算术平方根，其数值越小，说明模型的精度越高。各指标计算如公式 (3) 和公式 (4) 所示。

$$R^2 = 1 - \frac{\sum_1^n \hat{Z}(x_i) - Z(x_i)^2}{\sum_1^n \hat{Z}(x_i) - \bar{Z}(x_i)^2}$$

(3)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n [\hat{Z}(x_i) - Z(x_i)]^2}{n}}$$

(4)

式中： n 为验证集中的样点个数； $\hat{Z}(x_i)$ 为预测值； $Z(x_i)$ 为实测值； $\bar{Z}(x_i)$ 为实测值的平均值。

2 结果与分析

2.1 描述性统计分析

研究区 SOM 样本描述性统计特征，包含最大最小值、平均值、标准差、变异系数、偏度和峰度 (表 2)。描述性统计结果显示，研究区 1 565 个总体样本中，SOM 含量范围为 2.0~51.1 g/kg，平均值 15.62 g/kg；变异系数 47.70%，表明研究区 SOM 含量具有中等程度变异性，存在空间异质性，可以对其进行空间局部预测；从偏度和峰度来看，研究区的 SOM 含量近似符合正态分布。

根据《全国九大农区及省级耕地质量监测指标分级标准》中的山东省耕地质量监测指标分级标准 (表 3)，研究区内 SOM 平均含量属于中等级别，且总体样本 SOM 含量主要分布在中、较低和低级别，分别占样点总数的 23.71%、31.94% 和 22.43%，整体来看，研究区内 SOM 含量属于中低水平。

表2 山东省沂源县 SOM 描述性统计
Table 2 Descriptive statistics of SOM in Yiyuan county, Shandong province

数量	最小值/(g/kg)	最大值/(g/kg)	平均值/(g/kg)	标准差/(g/kg)	变异系数/%	偏度	峰度
1 565	2.0	51.1	15.62	7.45	47.70	0.53	-0.17

表3 山东省耕地质量监测 SOM 指标分级标准
Table 3 Classification standards for SOM indicators in cultivated land quality monitoring of Shandong province

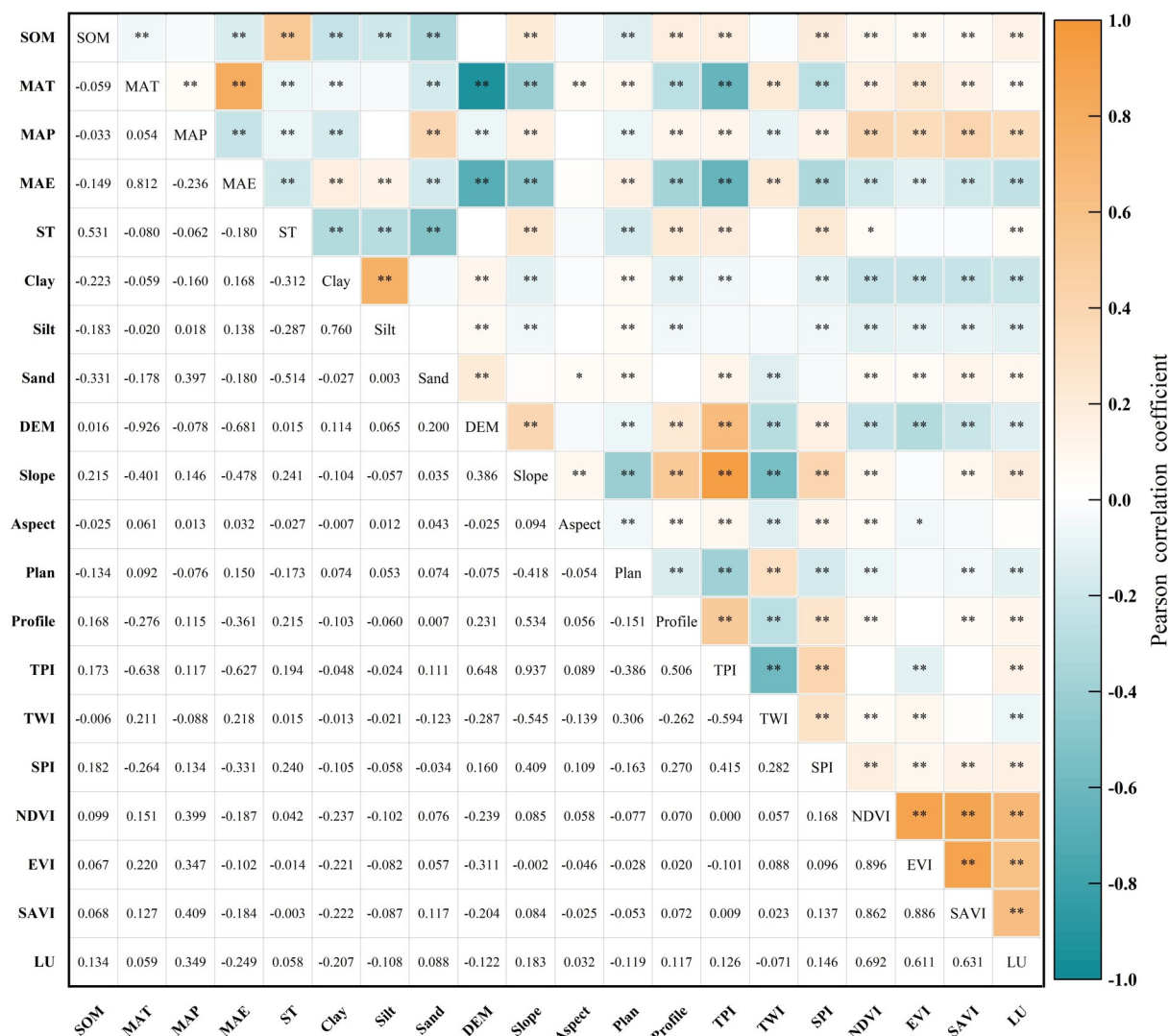
级别/(g/kg)	I级(高)>25.0	II级(较高)(20.0,25.0]	III级(中)(15.0,20.0]	IV级(较低)(10.0,15.0]	V级(低)≤10.0
样点数量	170	173	371	501	350
样点占比/%	10.86	11.05	23.71	32.01	22.37

2.2 环境变量筛选

皮尔逊相关系数能够表示 SOM 与环境变量之间的相关关系，通过皮尔逊相关系数筛选出显著相

关的环境变量。结果如图 3 所示，研究区 SOM 与 MAT (-0.06)、MAE (-0.15)、Clay (-0.22)、Silt (-0.18)、Sand (-0.33)、Plan (-0.13) 这 6 个变量呈显著负相关 ($P<0.01$)，与 ST (0.53)、

Slope (0.22)、Profile (0.17)、TPI (0.17)、SPI (0.07)、LU (0.13) 这9个变量呈显著正相关 ($P < 0.01$)、NDVI (0.10)、EVI (0.07)、SAVI (0.18)。



注: *表示相关性达到 $P < 0.05$ 显著水平, **表示相关性达到 $P < 0.01$ 显著水平。

图3 SOM与环境变量的相关性

Fig. 3 Correlation heatmap between SOM and environmental

计算与SOM显著相关的环境变量的VIF值并设置筛选环境变量阈值为7.5, 筛选后的结果如图4所示, 其中Plan、TPI、EVI的VIF值均大于7.5, 进行剔除。将环境变量MAT、MAE、ST、Clay、Silt、Sand、Slope、Profile、SPI、NDVI、SAVI、LU共12个作为参与SOM制图模型的环境变量。

2.3 制图模型构建

OK模型在ArcGIS软件中构建, 通过半方差函数确定SOM是否具有空间自相关性; 块金效应表示随机因素对SOM空间自相关性的影响程度, 块金效应越大, 表示空间自相关性越弱, 多次重复建

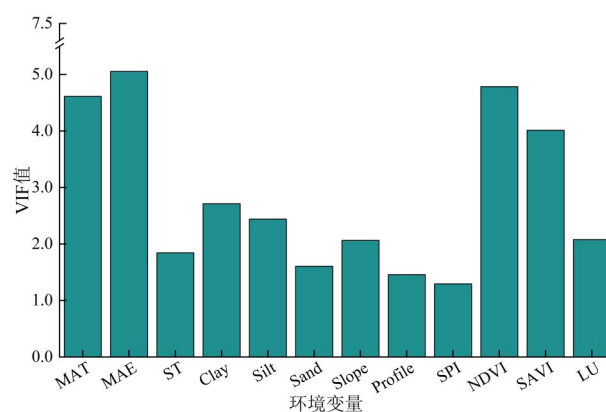


图4 环境变量VIF筛选结果

Fig. 4 Screening results of environmental variable VIF

模研究区 SOM 的块金效应在 38.68%~41.07%，存在中等程度空间自相关性，可以进行 OK 建模。

MLR 和 GWR 模型以 SOM 为因变量，筛选的环境变量为自变量进行建模，模型诊断结果如表 4 所示，包括残差平方和（Residual Sum of Squares, RSS）、AICc、 R^2 和调整后的 R^2 （Adjusted R^2 ），GWR 模型 RSS 和 AICc 值均小于 MLR， R^2 和 Adjusted R^2 均高于 MLR，表明本研究区中 GWR 模型优于 MLR 模型。各环境变量回归系数结果如表 5 所示，MLR 模型中 MAE、ST、Clay、Sand 和 LU 等环境变量在 0.01 的水平上显著，此外，联合 F 统计量（54.28*）通过显著性检验，表明回归模型具有整体显著性。GWR 模型中各环境变量系数的平均值与 MLR 模型的系数基本对应，但存在差距，这与研究区内环境变量复杂、空间异质性较强有关。

表 4 MLR 模型与 GWR 模型 SOM 数字制图结果
Table 4 Results of the MLR model and the GWR model for SOM digital mapping

模型	RSS	AICc	R^2	Adjusted R^2
MLR	37 851.59	9 319.46	0.38	0.32
GWR	29 578.75	8 711.93	0.46	0.37

表 5 MLR 模型与 GWR 模型的回归系数
Table 5 Regression coefficients of the MLR model and the GWR model

环境变量	MLR 模型 回归系数	GWR 模型回归系数			
		最小值	中位数	最大值	平均值
截距	7.29*	-18.03	8.47	51.12	12.23*
MAT	-1.65	-59.31	-0.84	21.06	-8.71*
MAE	-1.56*	-24.90	-2.68	13.62	-3.13*
ST	16.11*	10.06	15.01	22.00	15.13*
Clay	-4.12*	-10.61	-0.47	12.68	-0.46
Silt	0.57	-10.95	-2.68	13.46	-1.80
Sand	5.79*	-16.09	-3.87	11.37	-3.93*
Slope	1.75	-15.59	-0.05	25.21	1.35
Profile	0.61	-22.67	-0.08	22.36	-1.18
SPI	2.80	-13.66	2.22	16.88	2.22
NDVI	1.95	-35.79	-1.73	26.47	-2.17
SAVI	-1.30	-30.65	0.17	35.21	0.92
LU	3.88*	-18.11	1.75	13.11	1.93

注：*表示在 0.01 的水平下差异显著。

RF 和 GWR 模型均使用 R 语言构建，2 个模型中决策树的数量（ntree）和每棵决策树节点分裂时选用的变量数量（mtry）是两个关键参数。本研究

RF 和 GWR 模型 ntree 均设置为 500，采用交叉验证方法分别验证不同 mtry 对不同模型精度的影响，选取在带外数据上令 R^2 最高，同时 RMSE 最低的值设为最佳 mtry（图 5），RF 模型最佳 mtry 为 3，GWR 最佳 mtry 为 4。在 GWR 模型构建中，还需确定重要参数“带宽”，本研究选取“自适应”核函数，根据训练样本数量定义模型带宽测试范围，多次测试带外数据的交叉验证结果，最终确定 GWR 模型最佳带宽为 500（图 6），基于以上参数构建 RF 和 GWR 模型。

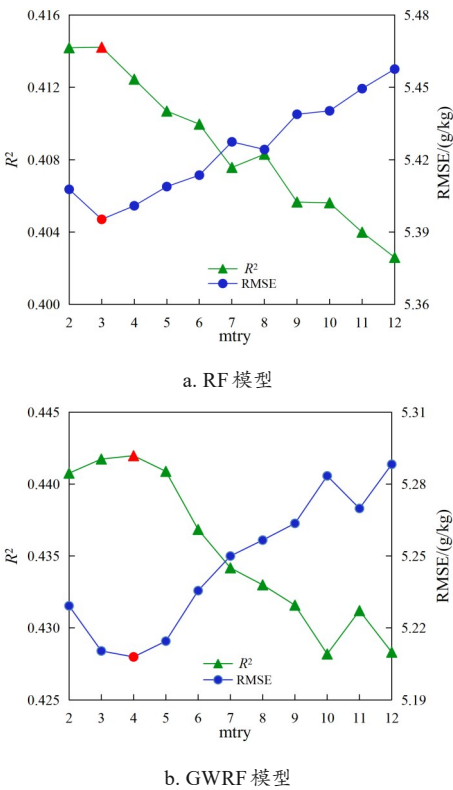


图 5 RF 与 GWR 模型的 mtry 确定结果
Fig. 5 Results of mtry determination for RF and GWR models

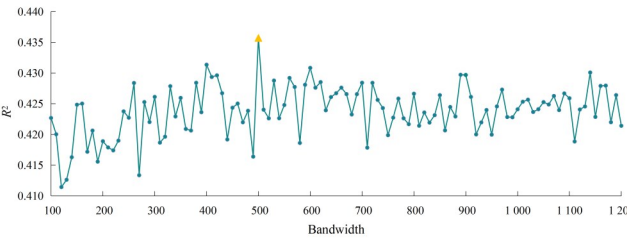


图 6 GWR 模型带宽测试结果
Fig. 6 Bandwidth test results of GWR model

2.4 模型性能对比

基于 20 组随机独立验证集，5 种模型对研究区 SOM 预测精度评价结果见表 6。不同模型对

SOM 预测精度表现存在显著差异，其中GWRF 模型 ($R^2=0.48$, $RMSE=5.12$ g/kg) 预测精度最高；其次是 RF 模型 ($R^2=0.41$, $RMSE=5.48$ g/kg)，也具有较高的预测精度；GWR 模型 ($R^2=0.35$, $RMSE=5.71$ g/kg) 和 MLR 模型 ($R^2=0.32$, $RMSE=5.85$ g/kg) 预测精度较低，OK 模型 ($R^2=0.24$, $RMSE=6.18$ g/kg) 预测精度最低，在山地丘陵为主的地区难以对 SOM 进行精确预测。整体而言，GWRF 模型与 OK、MLR、GWR 和 RF 模型相比， R^2 分别提升 0.24、0.16、0.13 和 0.07，RMSE 分别降低 1.06、0.73、0.59 和 0.36 g/kg，精度明显提升。为进一步验证 GWRF 模型精度提升的可靠性，采用配对样本 t 检验对 GWRF 模型与其他模型的预测精度差异进行显著性检验（表 7）。结果显示，GWRF 模型相较于其他模型的 R^2 提高和 RMSE 降低均显著。研究区 20 组验证集中随机 1 组的 SOM 实测值与 5 种模型预测值关系散点图如图 7 所示，OK、MLR 和 GWR 模型的离散程度较为明显，拟合精度较低；RF 和 GWRF 表现出较强的相关性，拟合趋势线较为接近 1:1 对角线，表明预测值与实测值

拟合度更高，整体预测效果好。

表 6 SOM 制图模型精度对比

Table 6 Accuracy comparison of SOM mapping models					
制图模型	OK	MLR	GWR	RF	GWRF
R^2	0.24	0.32	0.35	0.41	0.48
RMSE/(g/kg)	6.18	5.85	5.71	5.48	5.12

表 7 GWRF 与其他模型预测精度差异显著性检验

Table 7 Significance test of prediction accuracy differences between GWRF and other models					
模型对比	指标	平均差值	t 值	自由度	显著性(P 值)
GWRF-OK	R^2	0.24	29.79	19	<0.01
	RMSE	-1.06	-35.55	19	<0.01
GWRF-MLR	R^2	0.16	19.03	19	<0.01
	RMSE	-0.73	-23.26	19	<0.01
GWRF-GWR	R^2	0.13	13.91	19	<0.01
	RMSE	-0.59	-18.03	19	<0.01
GWRF-RF	R^2	0.07	4.32	19	<0.01
	RMSE	-0.36	-5.56	19	<0.01

注：RMSE 平均差值单位为 g/kg； $P<0.01$ 表示两指标在 0.01 水平下呈现显著差异。

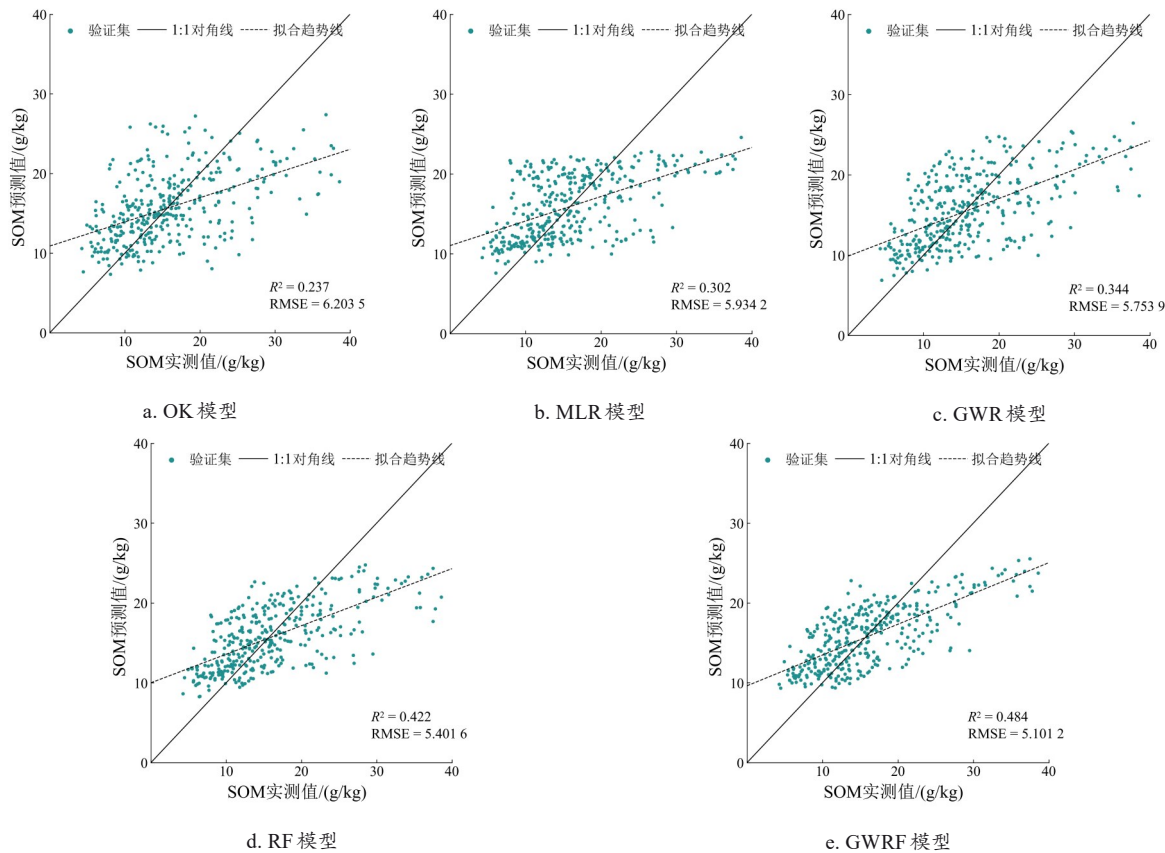


图 7 沂源县 SOM 不同模型预测散点图

Fig. 7 Scatter plots of SOM prediction by different models in Yiyuan county

对不同模型 20 次随机建模中表现最优的 SOM 空间预测结果进行制图, 制图表现如图 8 所示。5 种模型预测的 SOM 含量均表现出中部高、西南和东北低的总体趋势, 但在数值、局部变化等细节上存在差异。OK 模型生成 SOM 含量空间分布图呈现出整体平滑的趋势, 难以反映局部突变和细微变

化。MLR 和 GWR 模型能够反映环境变量对 SOM 分布的影响, 局部变化优于 OK 模型, 但受线性假设限制, 高值或低值会接近总体均值, 出现高值低估、低值高估现象。RF 和 GWRF 模型通过有效建模多变量间的非线性关系, 精细地捕捉整体趋势与局部变化, 提供精细的 SOM 含量空间分布图。

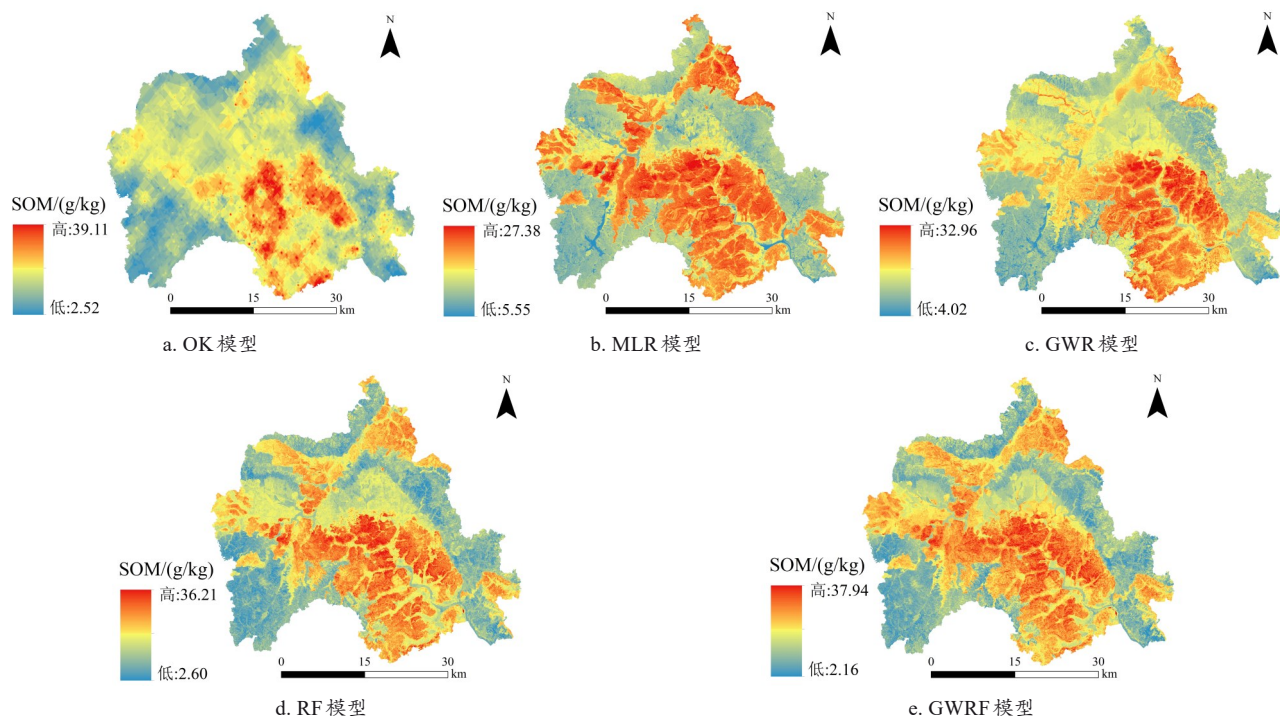


图 8 不同模型 SOM 数字制图表现对比

Fig. 8 Comparison of mapping performance of different models for SOM digital mapping

此外, 为全面评估各模型的计算效率, 本研究在相同硬件环境下记录了模型的平均运行时间。所有模型均在 Windows 10 操作系统、AMD Ryzen 74800H with Radeon Graphics (2.90 GHz)、32 GB RAM 的硬件环境中运行。模型运行时间包括训练与预测两个阶段, 均以单线程运行时间为准。结果表明, 不同模型的计算效率差异较大, 其中 OK 和 MLR 模型运行时间最短, 训练和预测均在数秒到十几秒内运行完成; GWR 模型训练需约 10 min, 预测环节在数秒内完成; RF 模型训练和预测总用时在 2~3 min; GWRF 模型计算量最大, 运行时间在 1 h 左右。

2.5 SOM 含量空间分布

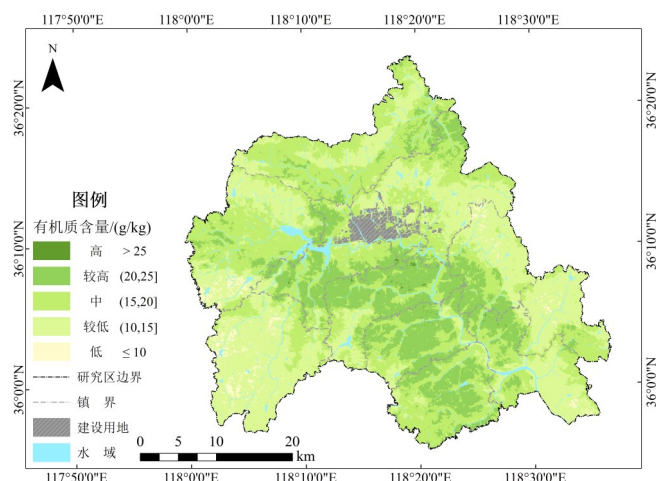
基于 GWRF 模型, 预测研究区内 SOM 含量, 并依据山东省耕地质量监测指标分级标准制作 SOM 含量空间分布图 (图 9)。从整体上看, 研究区 SOM 含量呈现中部高、西南和东北低的趋势,

其中 SOM 含量中等和较低的土壤面积最大, SOM 含量高的土壤面积最小, 这与土壤样点级别数量占比情况较为相符。SOM 含量高和较高土壤主要分布在研究区中南部的燕崖镇、中庄镇和西里镇; SOM 含量低和较低土壤主要分布在研究区西南部的大张庄镇和东南部的张家坡镇和东里镇; SOM 含量中等的土壤则在研究区北部的南鲁山等镇分布较多。

3 讨论

3.1 制图模型比较分析

本研究构建了 SOM 的数字制图 GWRF 模型, 并与其他 4 种模型进行比较, 发现各模型在预测精度上存在显著差异, 预测精度由高到低排序为: GWRF>RF>GWR>MLR>OK。其中, OK 模型仅依靠样点属性的空间自相关性进行预测, 忽略了研究区内复杂的环境变量对 SOM 的影响, 预测精度最



注：该图基于自然资源部标准地图服务网站下载的审图号为GS(2019)1822号标准地图制作，底图无修改。

图9 沂源县SOM含量空间分布

Fig. 9 Spatial distribution of SOM content in Yiyuan county

低，仅作为基准模型用于对比。这与袁玉琦等^[33]研究观点一致，OK模型在环境复杂或变量丰富区域的预测能力有限。MLR模型引入了环境变量作为解释因子，能够建立SOM与环境变量之间的全局线性关系，但未考虑空间位置对变量关系的影响，其建模预测能力仍较为有限，预测精度仅略高于OK模型。GWR模型在MLR基础上考虑环境变量对SOM影响的空间异质性，回归系数随空间位置变化^[34]，预测精度较MLR有所提高。但由于GWR本质上仍是局部线性模型，在面对变量间非线性关系时建模能力有限。RF通过引入非参数、集成学习机制，能较好地处理变量之间复杂的非线性关系，因此预测精度明显高于GWR、MLR和OK模型。GWRF模型在所有模型中表现最优。这与ZHOU等^[35]、李泽等^[19]的观点一致。

在本研究中，GWRF模型与OK、MLR、GWR和RF模型相比， R^2 分别提升0.24、0.16、0.13和0.07，RMSE分别降低1.06、0.73、0.59和0.36 g/kg。结果表明GWRF模型集合了GWR的局部建模思想与RF的非线性建模能力，在每个位置上利用加权的邻近样点构建不同的RF模型，能够根据空间位置变化对变量关系进行局部非线性建模，兼顾空间异质性与变量非线性特征的耦合建模能力，从而在环境复杂的山区县域SOM数字制图中更加精准、稳定，整体预测精度最高。然而，目前本研究GWRF模型中建立的每个局部RF均使用同样的环境变量和参数，未考虑局部模型的差异，因此在今后的研究中将考虑使用Boruta特征选择和

递归特征消除等方法实现局部参数调优和局部特征选择，进一步提升模型的适应性。

此外，不同模型的计算效率也是模型比较的重要方面。本研究在相同的硬件环境下对模型的运行时间进行统计，结果表明在精度上明显优于其他模型的GWRF模型，运行时间也相对更长，这主要是因为GWRF模型需要在每个预测位置构建局部RF模型，计算量远高于仅建立单一全局模型的RF等方法。尽管GWRF模型的运行时间较长，但考虑到其显著提升的预测精度，计算效率仍在可接受范围内。在研究区尺度较大的高分辨率制图任务中，可通过并行计算、提高硬件配置等方法提升GWRF模型的计算效率，以兼顾模型精度与运行效率。

3.2 SOM影响因素分析

土壤是由多种成土因素共同作用的结果，选取气候、土壤、地形、植被和土地利用5类影响因素，基于RF和GWRF模型环境变量重要性排序，分析以上因素对SOM含量分布的影响。环境变量重要性排序如图10，其中“%IncMSE”越大，表示环境变量越重要。RF模型和GWRF模型重要性最高的因素均是ST、MAE，最低则是Silt，而剩余环境变量重要性排序虽存在一定相似性，但也具有较明显差异。为深入分析SOM的影响因素，通过计算不同类别内环境变量重要性分值的平均值来代表不同类型环境因素对SOM的影响程度。RF和GWRF模型的环境变量类别重要性排序均为：土壤>地形>气候>土地利用>植被。土壤因素是影响研究区SOM含量的最重要因素，土壤类型决定土壤的基本矿物成分，土壤质地决定有机质的吸附能力与稳定性，因此土壤因素直接决定SOM的本底水平与富集能力^[36]，重要性最高。地形因素不仅反映区域环境，还控制土壤的水热条件，影响土壤物质的迁移与沉积，成为影响SOM含量的重要因素^[37]。此外，气候也是影响SOM含量的重要因素，其通过调控土壤水分状况、植被生长和微生物分解速率，影响SOM的积累与消耗^[38]。

GWRF模型不仅适用于整体影响因素的分析，同时能够揭示这些环境变量在地理空间上的分布差异，各环境变量重要性得分的空间分布如图11所示。各环境变量的重要性得分大致呈现东南高西北低的特征，燕崖镇、中庄镇和西里镇各环境变量的重要性得分整体较高，这与SOM含量的空间分布特征大致相同。此外，有学者指出，土壤属性不仅

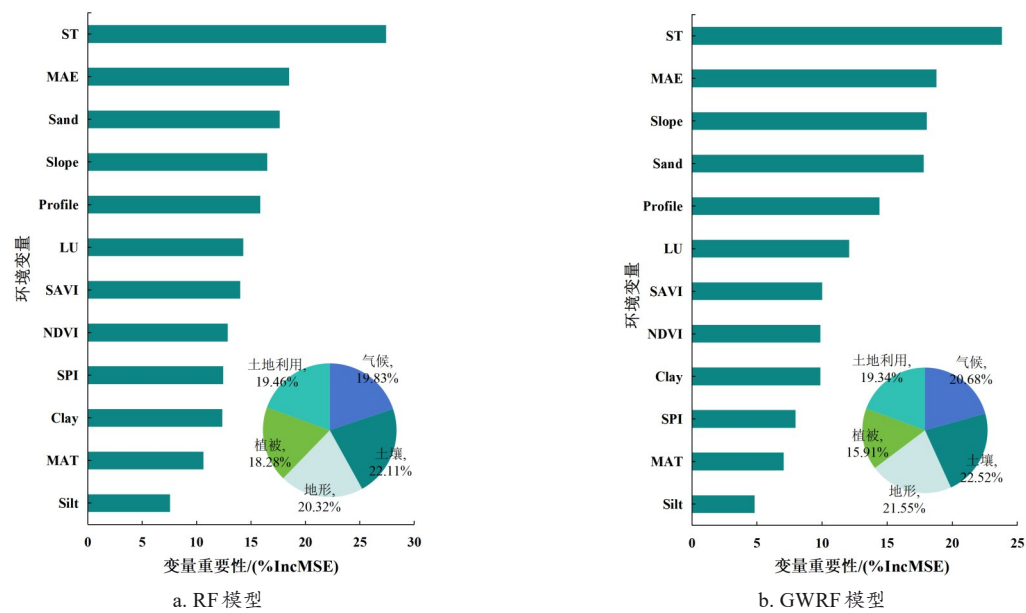


图 10 SOM 数字制图 RF 与 GWRf 模型环境变量重要性排序

Fig. 10 Ranking of environmental variable importance for RF and GWRf models for SOM digital mapping

受自然因素影响，而且受人为因素的影响也非常显著，如施肥、轮作、农田管理和距村庄距离等人类活动对 SOM 含量具有重要影响^[39]，因此后续的研究

将考虑纳入更多的人类活动等环境变量，以提升模型的预测精度。

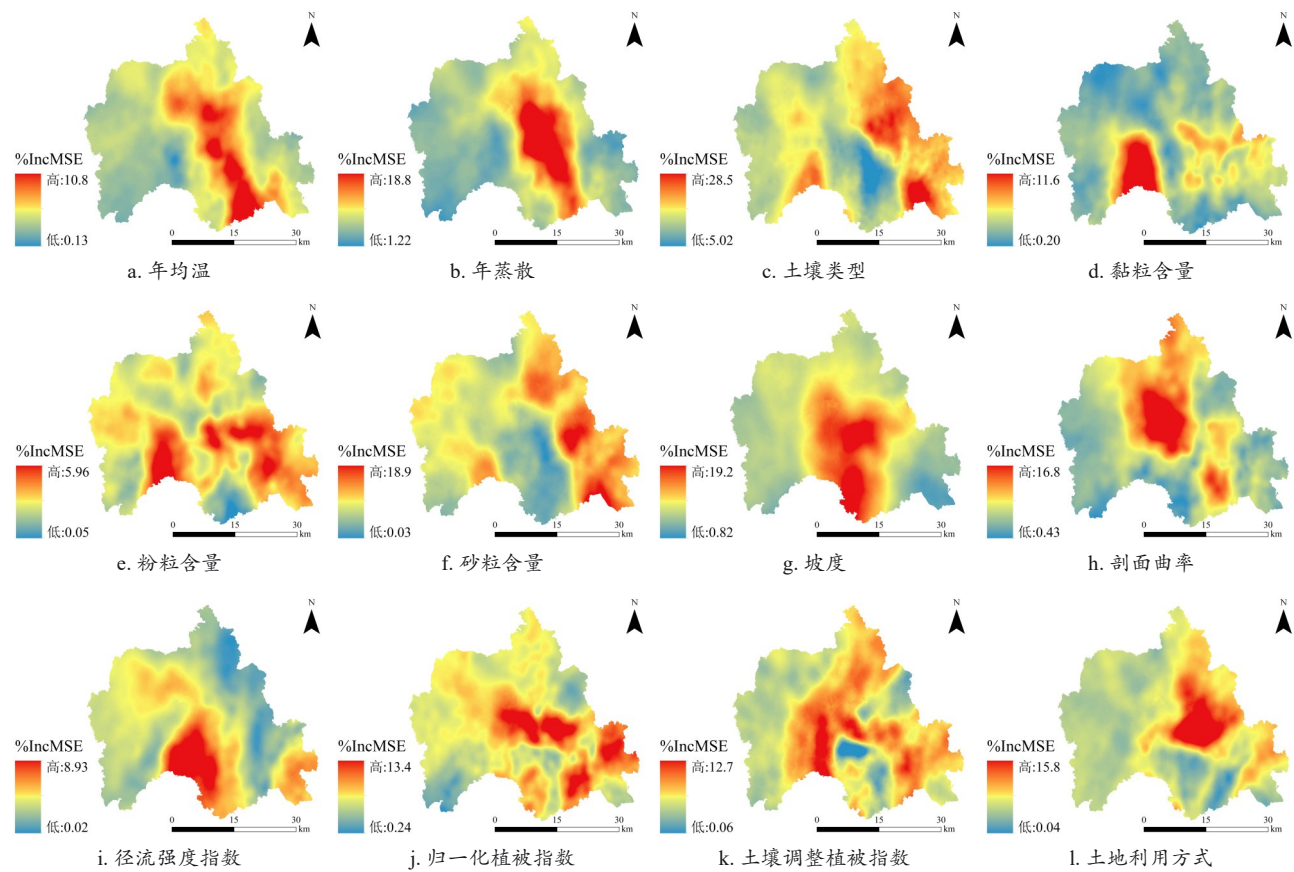


图 11 SOM 数字制图 GWRf 模型环境变量重要性得分空间分布

Fig. 11 Spatial distribution of environmental variable importance scores of the GWRf model for SOM digital mapping

4 结 论

1) GWRF模型集成非线性与空间异质性建模能力,表现出显著的SOM预测优势,其预测精度($R^2=0.48$, $RMSE=5.12$ g/kg)优于RF与传统线性模型。该模型适合应用于复杂地形地区的县域尺度SOM及其他土壤属性制图工作,研究结果为土壤三普背景下山地丘陵区县域SOM及其他土壤属性高精度数字制图提供了方法支撑。

2) 研究区SOM含量总体不高,空间异质性显著,SOM含量高和较高含量的土壤主要分布于研究区中南部地区;SOM含量中度土壤主要分布于研究区中部及北部地区;SOM含量低和较低土壤主要分布于研究区西南部和东南部地区。土壤类型、年蒸散发量、坡度和砂粒含量因素是影响研究区SOM含量最重要的因素,各影响因素在不同空间位置下的作用存在差异。

利益冲突声明: 本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献:

- [1] 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.
WANG Y X, YANG K, GAO B B, et al. Prediction of the spatial distribution of soil organic matter based on two-point machine learning method[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(12): 65-73.
- [2] CHEN Y X, WEI T X, REN K, et al. The coupling interaction of soil organic carbon stock and water storage after vegetation restoration on the Loess Plateau, China[J]. Journal of Environmental Management, 2022, 306: 114481.
- [3] 沈仁芳, 颜晓元, 张甘霖, 等. 新时期中国土壤科学发展现状与战略思考[J]. 土壤学报, 2020, 57(5): 1051-1059.
SHEN R F, YAN X Y, ZHANG G L, et al. Status quo of and strategic thinking for the development of soil science in China in the new era[J]. Acta Pedologica Sinica, 2020, 57(5): 1051-1059.
- [4] 朱阿兴, 杨琳, 樊乃卿, 等. 数字土壤制图研究综述与展望[J]. 地理科学进展, 2018, 37(1): 66-78.
ZHU A X, YANG L, FAN N Q, et al. The review and outlook of digital soil mapping[J]. Progress in Geography, 2018, 37(1): 66-78.
- [5] ARROUAYS D, MCBRATNEY A, BOUMA J, et al. Impressions of digital soil maps: The good, the not so good, and making them ever better[J]. Geoderma Regional, 2020, 20: e00255.
- [6] 刘顺国, 徐英德, 裴久渤, 等. 以土壤普查成果助推黑土地科学保护与利用[J]. 土壤通报, 2024, 55(4): 1185-1190.
LIU S G, XU Y D, PEI J B, et al. Promoting the scientific protection and utilization of black land with the results of soil census[J]. Chinese Journal of Soil Science, 2024, 55(4): 1185-1190.
- [7] 曹佳萍, 张黎明, 邱龙霞, 等. 基于稀疏样点的南方丘陵地区耕地土壤有效磷制图[J]. 中国生态农业学报(中英文), 2022, 30(2): 290-301.
CAO J P, ZHANG L M, QIU L X, et al. Mapping soil available phosphorus of cultivated land in hilly region of Southern China based on sparse samples[J]. Chinese Journal of Eco-Agriculture, 2022, 30(2): 290-301.
- [8] 王奇, 王世航, 陶勤, 等. 典型黑土区农场尺度土壤属性数字制图方法对比研究[J]. 土壤, 2025, 57(2): 430-444.
WANG Q, WANG S H, TAO Q, et al. A comparative study of farm-scale digital mapping methods for soil attributes in the typical black soil region[J]. Soils, 2025, 57(2): 430-444.
- [9] 陈琳, 任春颖, 王宗明, 等. 基于克里金插值的耕地表层土壤有机质空间预测[J]. 干旱区研究, 2017, 34(4): 798-805.
CHEN L, REN C Y, WANG Z M, et al. Prediction of spatial distribution of topsoil organic matter content in cultivated land using Kriging methods[J]. Arid Zone Research, 2017, 34(4): 798-805.
- [10] 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.
LI M J, WANG L, LIU H B, et al. Contrastive analysis of spatial distribution of soil organic matter content predicted by different models[J]. Southwest China Journal of Agricultural Sciences, 2021, 34(3): 610-617.
- [11] DUAN L X, LI Z W, XIE H X, et al. Large-scale spatial variability of eight soil chemical properties within paddy fields[J]. Cate-na, 2020, 188: 104350.
- [12] 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.
WANG Y Q, ZHANG X, ZHAO Y P, et al. Interpolation of soil CEC of sandy fields using GIS and geographically weighted regression-Kriging[J]. Soils, 2020, 52(2): 421-426.
- [13] 仇皓雷, 王海燕. 机器学习在土壤性质预测研究中的应用进展[J]. 生态学杂志, 2025, 44(1): 283-294.
QIU H L, WANG H Y. Application of machine learning to the prediction of soil properties: A review[J]. Chinese Journal of Ecology, 2025, 44(1): 283-294.
- [14] 薄延素, 李昊明, 王葛菲, 等. 松嫩平原西部草地土壤有机质含量预测[J]. 水土保持研究, 2025, 32(3): 63-71.
BO Y S, LI H M, WANG G F, et al. Prediction of soil organic matter content in the grassland of the western Songnen Plain[J]. Research of Soil and Water Conservation, 2025, 32(3): 63-71.
- [15] 胡志瑞, 赵万伏, 宋垠先, 等. 基于改进麻雀搜索算法优化BP神经网络的土壤有机质空间分布预测[J]. 环境科学, 2024, 45(5): 2859-2870.
HU Z R, ZHAO W F, SONG Y X, et al. Prediction spatial distribution of soil organic matter based on improved BP neural network with optimized sparrow search algorithm[J]. Environmental Science, 2024, 45(5): 2859-2870.
- [16] 李兰晖, 黄聪聪, 张懿铨, 等. 基于地理加权随机森林的青藏地区放牧强度时空格局模拟[J]. 地理科学, 2023, 43(3): 398-410.
LI L H, HUANG C C, ZHANG Y L, et al. Mapping the multi-temporal grazing intensity on the Qinghai-Tibet Plateau using geographically weighted random forest[J]. Scientia Geographica Sinica, 2023, 43(3): 398-410.
- [17] KHAN S N, LI D P, MAIMAITIJANG M. A geographically weighted random forest approach to predict corn yield in the US corn belt[J]. Remote Sensing, 2022, 14(12): 2843.
- [18] 卫格冉, 李明泽, 全迎, 等. 基于地理加权随机森林的黑龙江省森林碳储量遥感估测[J]. 中南林业科技大学学报, 2024, 44(7): 64-76.
WEI G R, LI M Z, QUAN Y, et al. Geographically weighted random forest approach to predict forest carbon storage by remote sensing in Heilongjiang[J]. Journal of Central South University of Forestry & Technology, 2024, 44(7): 64-76.
- [19] 李泽, 杜哲, 毕善婷, 等. 基于GWRF模型的滨海平原土壤含盐量预测及影响因素分析[J]. 环境科学, 2025, 46(8): 4982-4992.
LI Z, DU Z, BI S T, et al. Prediction of soil salinity and analysis of influencing factors in coastal Plains based on geographically weighted random forests[J]. Environmental Science, 2025, 46(8): 4982-4992.
- [20] 姚彩燕, 刘绍贵, 乔婷, 等. 基于时空变异的旱地土壤有机碳高效采样策略研究[J]. 土壤学报, 2021, 58(3): 638-648.
YAO C Y, LIU S G, QIAO T, et al. Strategy for efficient sampling of upland soil based on spatiotemporal variation of the soil[J]. Acta Pedologica Sinica, 2021, 58(3): 638-648.
- [21] 孟祥添. 基于深度学习模型东北耕作土壤有机碳时空分布及驱动因素研究[D]. 哈尔滨: 中国科学院大学(中国科学院东北地理与农业生态研究所), 2024.

- MENG X T. Study on spatial and temporal distribution and driving factors of soil organic carbon in cultivated land in Northeast China based on deep learning model[D]. Harbin: Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, 2024.
- [22] 任艳, 尹秋月, 尹晋磊, 等. 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[J]. 土壤, 2025, 57(3): 673-682.
REN Y, YIN Q Y, YIN J L, et al. Comparison of predictive methods for spatial distribution of SOM in cultivated land in the south subtropical karst region[J]. Soils, 2025, 57(3): 673-682.
- [23] 陈艺敏, 苏漳文, 陈移萍, 等. 基于地理加权随机森林的长三角PM_{2.5}建模[J]. 中国环境科学, 2024, 44(8): 4240-4248.
CHEN Y M, SU Z W, CHEN Y P, et al. Modeling of PM_{2.5} in the Yangtze River Delta based on geographically weighted random forest[J]. China Environmental Science, 2024, 44(8): 4240-4248.
- [24] SANTOS F, GRAW V, BONILLA S. A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon[J]. PLoS One, 2019, 14(12): e0226224.
- [25] CHENG L W, YAN M Z, ZHANG W H, et al. Interpretable digital soil organic matter mapping based on geographical Gaussian process-generalized additive model (GGP-GAM)[J]. Agriculture, 2024, 14(9): 1578.
- [26] GEORGANOS S, GRIPPA T, NIANG GADIAGA A, et al. Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling[J]. Geocarto International, 2021, 36(2): 121-136.
- [27] DAI X L, ZHU Y Q, SUN K, et al. Examining the spatially varying relationships between landslide susceptibility and conditioning factors using a geographical random forest approach: A case study in Liangshan, China[J]. Remote Sensing, 2023, 15(6): 1513.
- [28] GEORGANOS S, KALOGIROU S. A forest of forests: A spatially weighted and computationally efficient formulation of geographical random forests[J]. ISPRS International Journal of Geo-Information, 2022, 11(9): 471.
- [29] 杨之江, 陈效民, 景峰, 等. 基于GIS和地统计学的稻田土壤养分与重金属空间变异[J]. 应用生态学报, 2018, 29(6): 1893-1901.
YANG Z J, CHEN X M, JING F, et al. Spatial variability of nutrients and heavy metals in paddy field soils based on GIS and Geo-statistics[J]. Chinese Journal of Applied Ecology, 2018, 29(6): 1893-1901.
- [30] 陈宣强, 赵明松, 卢宏亮, 等. 基于3种地理加权回归方法的安徽省土壤pH空间预测制图对比研究[J]. 地理科学, 2023, 43(1): 173-183.
CHEN X Q, ZHAO M S, LU H L, et al. Comparison and analysis of spatial prediction and variability of soil pH in Anhui province based on three kinds of geographically weighted regression[J]. Scientia Geographica Sinica, 2023, 43(1): 173-183.
- [31] LIU F, WU H Y, ZHAO Y G, et al. Mapping high resolution national soil information grids of China[J]. Science Bulletin, 2022, 67(3): 328-340.
- [32] 董雨昕, 韩文霆, 崔欣, 等. 基于无人机与Sentinel-2A遥感数据协同的裸土期土壤含盐量反演[J]. 农业机械学报, 2025, 56(6): 434-445.
DONG Y X, HAN W T, CUI X, et al. Soil salinity inversion during bare soil period based on collaboration of UAV and sentinel-2A remote sensing data[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(6): 434-445.
- [33] 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.
YUAN Y Q, CHEN H Y, ZHANG L M, et al. Prediction of spatial distribution of soil organic carbon in farmland based on multi-variables and random forest algorithm: A case study of a subtropical complex geomorphic region in Fujian as an example[J]. Acta Pedologica Sinica, 2021, 58(4): 887-899.
- [34] 卢宾宾, 葛咏, 秦昆, 等. 地理加权回归分析技术综述[J]. 武汉大学学报(信息科学版), 2020, 45(9): 1356-1366.
LU B B, GE Y, QIN K, et al. A review on geographically weighted regression[J]. Geomatics and Information Science of Wuhan University, 2020, 45(9): 1356-1366.
- [35] ZHOU Y, WEI G R, WANG Y, et al. Estimating regional forest carbon density using remote sensing and geographically weighted random forest models: A case study of mid- to high-latitude forests in China[J]. Forests, 2025, 16(1): 96.
- [36] 葛畅, 刘慧琳, 聂超甲, 等. 土壤肥力及其影响因素的尺度效应: 以北京市平谷区为例[J]. 资源科学, 2019, 41(4): 753-765.
GE C, LIU H L, NIE C J, et al. Scale effect of soil fertility spatial variability and its influencing factors[J]. Resources Science, 2019, 41(4): 753-765.
- [37] ZHANG W C, WAN H S, ZHOU M H, et al. Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques[J]. Ecological Indicators, 2022, 143: 109420.
- [38] 解文艳, 周怀平, 杨振兴, 等. 黄土高原东部渭河流域农田土壤有机质时空变异及影响因素[J]. 农业资源与环境学报, 2019, 36(1): 96-104.
- [39] XIE W Y, ZHOU H P, YANG Z X, et al. The spatial-temporal variation of soil organic matter and its influencing factors in Xiaohu River basin in eastern Loess Plateau, China[J]. Journal of Agricultural Resources and Environment, 2019, 36(1): 96-104.
- [39] 咸阳, 宋江辉, 王金刚, 等. 基于环境变量筛选与机器学习的土壤养分含量空间插值研究[J]. 农业机械学报, 2024, 55(10): 379-391.
XIAN Y, SONG J H, WANG J G, et al. Spatial interpolation of soil nutrients content based on environmental variables screening and machine learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(10): 379-391.

Geographically Weighted Random Forest for County-scale Digital Mapping of Soil Organic Matter: A Case Study in the Central Shandong Mountains

ZHANG Shulin¹, CUI Liqin³, LIU Jian¹, ZHANG Canting¹, WANG Hongjia¹,
ZHANG Tingting¹, WANG Ailing^{1,2*}

(1. College of Resources and Environment, Shandong Agricultural University, Tai'an 271018, China; 2. National Engineering Research Center for Efficient Utilization of Soil and Fertilizer, Shandong Agricultural University, Tai'an 271018, China; 3. Yiyuan County Agriculture and Rural Affairs Bureau, Zibo 256100, China)

Abstract: [Objective] Soil organic matter (SOM) is a fundamental indicator for evaluating soil fertility and soil quality. In mountainous counties characterized by complex terrain and pronounced environmental heterogeneity, SOM exhibits strong spatial variability even over short distances, which often results in limited prediction accuracy for conventional digital soil mapping (DSM) models. With the nationwide implementation of the Third National Soil Census, the demand for high-resolution and high-accuracy SOM mapping at the county scale has become increasingly urgent. Against this backdrop, Yiyuan county in Shandong province was selected as the study area to assess the applicability of the geographically weighted random forest (GWRF) model in SOM mapping within complex terrain regions. Furthermore, it sought to systematically compare the predictive performance of GWRF with several commonly used models, thereby providing technical support for soil resource surveys, census result compilation, and county-level land management. **[Methods]** The dataset consisting of 1 565 measured topsoil SOM samples was utilized, along with nineteen environmental variables representing five categories: topography, climate, vegetation, soil properties, and land use. Through correlation analysis and collinearity diagnostics, twelve key variables were retained for model construction. The GWRF model, which integrates localized spatial modeling with nonlinear machine-learning capability, was developed to generate high-resolution SOM predictions across the study area. An adaptive bandwidth strategy was employed, and the optimal bandwidth of 500 was determined. Grid search combined with cross-validation was used to identify the optimal mtry value of 4 for the random forest component. In addition to GWRF, four reference models were constructed for comparison: ordinary kriging (OK), multiple linear regression (MLR), geographically weighted regression (GWR), and random forest (RF). Model performance was evaluated using two commonly adopted accuracy metrics: the coefficient of determination (R^2) and root-mean-square error (RMSE). **[Results and Discussions]** Overall, SOM levels in Yiyuan county were relatively low, with a mean value of 15.62 g/kg. The spatial variation was moderate and exhibited a clear pattern: SOM values were higher in the central area and lower in the northeastern and southwestern areas. Considerable differences were observed in prediction accuracy among the five models. The GWRF model achieved the best overall performance, with an R^2 of 0.48 and an RMSE of 5.12 g/kg. This accuracy clearly surpassed that of RF ($R^2=0.41$) and GWR ($R^2=0.35$), and its advantage over MLR and OK was even more pronounced. A paired-sample t-test further confirmed that the accuracy improvements of GWRF over the other four models were statistically significant, supporting the robustness and reliability of the model's enhanced performance. According to the mapping results, the OK model produced an excessively smooth surface, making it difficult to reveal local details. While the MLR and GWR models could characterize certain environmental effects, they exhibited significant biases such as underestimation of high values and overestimation of low values. In contrast, the GWRF model performed prominently in capturing both global trends and local subtle variations. The analysis of variable importance showed that soil type, annual evapotranspiration, slope, and sand content were the most influential factors governing SOM distribution in the study area. Moreover, their spatially varying importance revealed notable heterogeneity. **[Conclusions]** This study demonstrated that the GWRF model possesses significant advantages in county-scale SOM digital mapping within mountainous areas. Its prediction accuracy markedly exceeded that of RF and conventional linear models, owing to its ability to simultaneously capture nonlinear environmental relationships and localized spatial variations. The enhanced mapping precision and improved representation of spatial details highlight the strong potential of GWRF for applications requiring high-accuracy soil information. GWRF is well-suited for SOM prediction under complex terrain conditions and can serve as an effective technical tool for county-level soil property estimation. Future research may incorporate human-activity-related variables, employ localized variable-selection strategies within the GWRF framework to further refine model performance, and explore the application potential of more advanced deep learning models in soil property mapping.

Keywords: soil organic matter; digital soil mapping; mountainous areas; geographically weighted regression; random forest

Foundation items: National Natural Science Foundation of China (42171378); The Natural Science Foundation of Shandong Province (ZR2021MD018); The Special Funds of Taishan Scholar of Shandong Province (tsqnz20231205)

First author: ZHANG Shulin, E-mail: 2023120274@sdau.edu.cn

***Corresponding author:** WANG Ailing, E-mail: ailingwang@sdau.edu.cn