

DOI: 10.11766/trxb202008240312

周洋, 赵小敏, 郭熙. 基于多源辅助变量和随机森林模型的表层土壤全氮分布预测[J]. 土壤学报, 2022, 59 (2): 451–460.

ZHOU Yang, ZHAO Xiaomin, GUO Xi. Prediction of Total Nitrogen Distribution in Surface Soil Based on Multi-source Auxiliary Variables and Random Forest Approach[J]. Acta Pedologica Sinica, 2022, 59 (2): 451–460.

基于多源辅助变量和随机森林模型的表层土壤全氮分布预测*

周洋^{1, 2}, 赵小敏^{1, 2†}, 郭熙^{1, 2}

(1. 江西农业大学国土资源与环境学院, 南昌 330045; 2. 江西省鄱阳湖流域农业资源与生态重点实验室, 南昌 330045)

摘要: 土壤全氮与土壤质量和肥力密切相关, 准确掌握土壤全氮的空间分布特征对精准农业管理措施的实施具有重要意义。以寻乌县为研究区域, 利用随机森林 (RF) 和随机森林残差克里格 (RFRK) 方法, 结合地形因子、地理坐标、遥感因子、气候因子、距离因子和土壤理化因子等多源辅助变量, 对寻乌县表层土壤全氮的空间分布进行预测和制图, 并在迭代 100 次模型后对比了两种模型的预测精度。结果表明, 在选择 4 种模型精度指标中, RF 模型的决定系数均值 ($R^2=0.6291$) 和一致性相关系数均值 ($CCC=0.7613$) 均高于 RFRK 模型 ($R^2=0.5719$, $CCC=0.6881$), 而 RF 模型的平均相对误差均值 ($MAE=0.1570 \text{ g}\cdot\text{kg}^{-1}$) 和均方根误差均值 ($RMSE=0.2108 \text{ g}\cdot\text{kg}^{-1}$) 均小于 RFRK 模型 ($MAE=0.1682 \text{ g}\cdot\text{kg}^{-1}$, $RMSE=0.2267 \text{ g}\cdot\text{kg}^{-1}$)。将残差作为误差项加入 RF 模型并未提高其预测精度, 因此, RF 模型可以作为土壤属性预测的一种有效方法, 为农业管理措施的实施提供技术支撑。

关键词: 土壤全氮; 空间分布; 随机森林; 数字土壤制图

中图分类号: S158.9 **文献标志码:** A

Prediction of Total Nitrogen Distribution in Surface Soil Based on Multi-source Auxiliary Variables and Random Forest Approach

ZHOU Yang^{1, 2}, ZHAO Xiaomin^{1, 2†}, GUO Xi^{1, 2}

(1. Academy of Land Resource and Environment, Jiangxi Agricultural University, Nanchang 330045, China; 2. Key Laboratory of Poyang Lake Watershed Agricultural Resources and Ecology of Jiangxi Province, Nanchang 330045, China)

Abstract: [Objective] Soil total nitrogen is closely related to soil quality and fertility. It is of great significance to know the spatial distribution characteristics of soil total nitrogen for the implementation of precision agriculture management. [Method] The spatial distribution of total nitrogen in the surface soil of Xunwu County was predicted and mapped by using two methods: random forest and random forest plus residuals kriging. These methods were combined with multi-source auxiliary variables such as (i) terrain factors, (ii) geographical coordinate, (iii) remote sensing factors, (iv) climate factors, (v) distance factors, and (vi)

* 国家重点研发计划项目 (2017YFD0301603) 和江西省赣鄱英才“555”领军人才项目 (201295) 资助 Supported by the National Key R & D Program of China (No. 2017YFD0301603), and the Gan Po “555” Talent Research Funds of Jiangxi Province (No. 201295)

† 通讯作者 Corresponding author, E-mail: zhaoxm889@126.com

作者简介: 周洋 (1997—) 男, 江西宜春人, 硕士研究生, 主要从事土地资源利用与保护研究。E-mail: zyang325@163.com

收稿日期: 2020-08-24; 收到修改稿日期: 2021-04-03; 网络首发日期 (www.cnki.net): 2021-12-17

soil physical or chemical factors. Also, the prediction accuracy of the two models was compared after 100 times of repeated operation. 【Result】 Our results show that the mean values of the decision coefficient ($R^2=0.629$ 1) and concordance correlation coefficient ($CCC=0.761$ 3) of the random forest model were higher than those of the random forest plus residual kriging method ($R^2=0.571$ 9, $CCC=0.688$ 1). Also, the mean values of the mean absolute error ($MAE=0.157$ 0 $\text{g}\cdot\text{kg}^{-1}$) and root mean squared error ($RMSE = 0.210$ 8 $\text{g}\cdot\text{kg}^{-1}$) were lower than those of the random forest plus residual kriging method ($MAE = 0.168$ 2 $\text{g}\cdot\text{kg}^{-1}$, $RMSE = 0.226$ 7 $\text{g}\cdot\text{kg}^{-1}$). 【Conclusion】 Importantly, adding residual to the random forest model did not improve its accuracy. These results suggest that the random forest model can be used as a new method for predicting soil properties, and it provides technical support for the implementation of agricultural management.

Key words: Soil total nitrogen; Spatial distribution; Random forest; Digital soil mapping

作为植物生长期所必需的养分元素, 土壤氮素是决定土壤质量和肥力的主要因素和指标^[1-2], 但是, 土壤中的氮素如果过量则会导致水体富营养化等农业面源污染问题^[3], 且氮素能形成温室气体进入大气^[4]。因此, 出于精准农业和环境质量管理的需要, 了解和掌握土壤氮素含量的空间分布显得尤为重要。

土壤全氮 (Soil total nitrogen, STN) 是指土壤中各种形式氮素的总和, 由于受到人为活动^[5]、土壤成土因素^[6]和地形因子^[7]等因素的影响, STN 往往表现出较强的变异程度, 准确估计 STN 的空间分布及其受到的主控因子可以为精准施肥等农业管理措施的实施提供理论依据和指导^[8]。随着“3S 技术”的发展, 地形、遥感、气候等环境因子变得易于获取, 因此有学者提出了数字土壤制图方法来实现土壤属性在区域范围内的精准预测^[9]。在数字土壤制图中, 多元线性回归^[10-11]、普通克里格法^[12]、回归克里格法^[13]、地理加权回归^[14]等方法被广泛用于预测 STN 等土壤属性的空间分布, 但土壤与环境变量往往表现为复杂的非线性关系, 多元线性回归和基于地统计的克里格方法不能很好地体现这种复杂关系, 因此有学者将人工神经网络^[15]、回归树^[16]和随机森林^[17]等机器学习模型应用在土壤属性预测中, 其中随机森林模型 (RF) 可以防止过度学习和过拟合, 相比人工神经网络计算更加简便, 且预测效果稳定, 被证明是一种有效的土壤属性预测方法^[18-19], 还有研究指出, 将 RF 模型的预测残差与预测结果相结合的随机森林残差克里格法 (RFRK) 能提高随机森林模型的预测精度^[20-21]。Guo 等^[21]对海南岛橡胶人工林的土壤有机质进行了预测, 研究表明, 相比逐步线性回归 (SLR) 和 RF, RFRK 表现出了更好的预测效果。Hengl 等^[22]对比了普通克里格 (OK)、

多元线性回归 (MLR)、RF、梯度提升树 (GBT) 和 3 种混合模型对希腊北部地区土壤有机质的预测效果, 结果表明 RFRK 的预测精度要优于其他方法。Scarpone 等^[23]使用了广义线性模型 (GLM)、RF 及其与地统计学相结合的混合方法对加拿大哥伦比亚省中南部的土壤厚度进行了建模预测, 发现混合方法的预测精度要优于普通的机器学习算法。作为一种有效的土壤属性预测方法, RFRK 得到了越来越广泛的应用, 但在国内的土壤属性预测研究中较少采纳。寻乌县位于赣南地区, 地形以山地丘陵为主, 属于典型的南方丘陵区, 果园产业发达, 而 STN 对土壤肥力和土壤碳循环均有着很大的影响, 因此, 很有必要对寻乌县 STN 的空间分布进行分析和预测, 为此本研究基于地形因子、遥感因子等多源辅助变量, 使用了随机森林 (RF) 和随机森林残差克里格 (RFRK) 两种预测模型对寻乌县表层 STN 的空间分布进行了预测, 为寻乌县的土壤肥力评估和精准施肥等农业措施提供理论指导和数据支撑。

1 材料与方 法

1.1 研究区概况

寻乌县位于江西省南部, 为赣州市所辖县, 位于闽粤赣三省相连处, 地理坐标介于 $24^{\circ}30'40''$ — $25^{\circ}12'10''\text{N}$ 和 $115^{\circ}21'22''$ — $115^{\circ}54'25''\text{E}$ 之间, 面积为 $2\,352\text{ km}^2$ 。该地区属于亚热带季风气候, 阳光充足, 雨量充沛, 年平均气温为 $18.9\text{ }^{\circ}\text{C}$, 在 7 月达到最大值 $27.2\text{ }^{\circ}\text{C}$, 1 月最低, 为 $9.1\text{ }^{\circ}\text{C}$, 而年降水量则介于 $1\,651\text{ mm}$ 到 $1\,750\text{ mm}$ 之间, 其中 4—6 月降水量最多。寻乌东北、西北与东南部地势高, 且向西南部倾斜, 海拔高度介于 $132\sim 1\,468\text{ m}$ 之间, 地貌类型以山地丘陵为主 (图 1a)。县内植被覆盖率

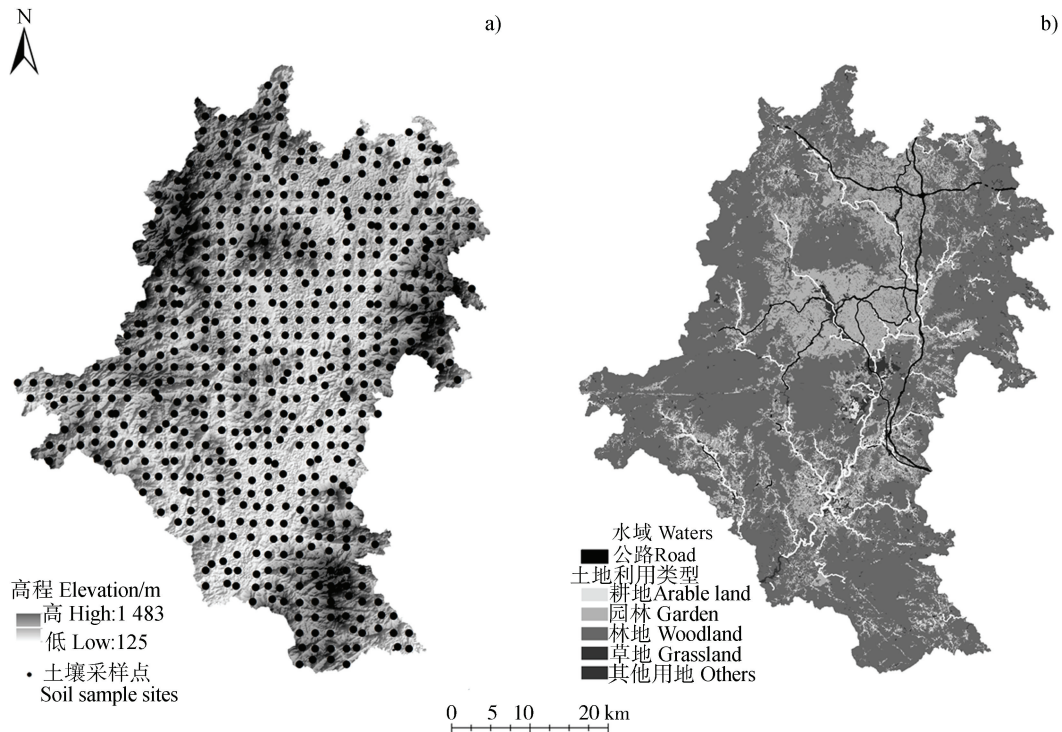


图 1 研究区 DEM、土壤样点分布及土地利用类型

Fig.1 Digital Elevation Model, distribution of soil sampling points and land use types in the study area

高，主要植被类型为亚热带常绿阔叶林和低山针叶林。寻乌属于南方亚热带红壤区的南部，土壤肥力较高，pH 普遍呈酸性，土壤类型主要为红壤、黄壤和水稻土；母质类型主要有酸性结晶岩类风化物 and 泥质岩类风化物，还有少量的石英岩类、泥质岩类和红砂岩类风化物；土地利用类型则以林地、园地和耕地为主，其中林地占土地总面积的 83.8% (图 1b)。

1.2 样点数据获取与测定

采用公里网格布设土壤采样点，于 2018 年 7 月至 9 月完成采集，采样深度为 0~20 cm，共采集土壤样品 572 个 (图 1a)。土壤样品经风干，研磨和过筛后测定全氮 (TN)、硫 (S)、pH 等元素和指标含量。其中 TN 采用半微量凯氏定氮法测定；S 采用燃烧碘量法测定；pH 采用电极法测定。

1.3 环境协变量及其数据来源

初步选择了 22 种环境协变量对研究区 STN 进行预测，包括地理坐标 (x 、 y 坐标)、地形因子、距离因子、气候因子、遥感因子和土壤理化因子。利用 ArcGIS10.2 软件平台，将所有的协变量栅格数据重采样为 100 m 分辨率。地形因子是土壤属性预测中使用最广泛的环境因子之一^[24]。本研究使用数字

高程模型 (Digital elevation model, DEM) 来得到各种地形因子。从地理空间数据云平台下载研究区 30 m 分辨率 DEM 数据，使用 ArcGIS 10.2 软件平台处理后得到高程、坡度、坡向、曲率、地形起伏度 5 个地形因子，使用 SAGA GIS 2.3.2 软件平台处理后得到地表粗糙度 TRI、地形湿度指数 TWI 和多尺度谷底平坦度 MrVBF 这 3 个地形因子。距离因子包括距河流距离 (DTR) 和距公路距离 (DTH)，利用 ArcGIS 10.2 软件平台的距离分析工具计算得到。气候因子使用从中国科学院资源环境科学数据中心下载的 1 km×1 km 分辨率的年平均气温 (MAT) 和年平均降水量 (MAP) 栅格数据集 (2006 年至 2015 年均值)。遥感因子从地理空间数据云平台上下下载的 2017 年 11 月 1 日的寻乌县 Landsat 8 OLI 遥感影像中提取，利用 ENVI 5.3 软件平台进行辐射定标和大气校正后，提取归一化植被指数 (NDVI)、绿带反射率 (B3)、裸土指数 (BI)、绿色叶绿素指数 (CIg)、绿度指数 (GI) 和对波段进行主成分分析后的第一主成分 (PCA1)。其中 NDVI 和绿度指数是反应植被生长和覆盖情况的常用指标^[25]，裸土指数体现了地表的裸露情况，而绿色叶绿素指数则反映了植物冠层氮的情况。土壤理化因子选择了 pH 和硫，采

用普通克里格法进行空间插值后,提取 100 m 分辨率的栅格数据。为了避免预测变量之间的多重共线性影响模型预测效果,在将协变量代入模型之前,计算了各环境变量的方差膨胀因子 VIF,将 VIF > 10 的变量剔除^[26]后再代入模型进行计算。

1.4 建模方法

随机森林 (RF) 是一种基于多个决策树的数据挖掘算法,通过在训练数据集中应用 bootstrap 方法随机抽取样本构建单独的树模型,生成大量的树后组成随机森林,与 CART 决策树模型相比,RF 模型对过度拟合不敏感^[27]。决策树的数量 ntree 和每个分隔节点的预测变量数 mtry 是 RF 模型中的两个关键参数^[28]。本研究在随机划分 100 组训练集和验证集数据后,使用 R3.6.1 中的“caret”包对每个 RF 模型分别进行五折交叉验证调参,之后使用“randomForest”包,代入不同训练集对应的最优参数来构建随机森林模型。随机森林残差克里格 (RFRK) 是 RF 模型的一种拓展,利用普通克里格法,对 RF 模型的预测残差进行空间插值后将其与 RF 模型的预测结果相加,从而提高模型预测精度。RFRK 模型的公式如下:

$$P_{\text{RFRK}}(x_0) = P_{\text{RF}}(x_0) + \varepsilon_{\text{OK}}(x_0) \quad (1)$$

式中, $P_{\text{RFRK}}(x_0)$ 为 RFRK 在 x_0 位置的 STN 预测值, $P_{\text{RF}}(x_0)$ 为 RF 模型拟合的趋势项, ε_{OK} 则为普通克里格法拟合的误差项。

1.5 模型精度与预测不确定性评估

为了获取稳定的模型预测值,对 RF 模型重复迭代了 100 次,每次均随机选择 70% 的土壤样点 (400 个) 作为训练集参与建模,其余 30% 的样点 (172 个) 作为验证集,最后取 100 次迭代结果的平均值作为 RF 模型的预测结果。选取了平均相对误差 (MAE)、均方根误差 (RMSE)、决定系数 (R^2) 和一致性相关系数 (CCC) 4 个模型精度指标评价模型预测结果,计算公式如下:

$$\text{CCC} = 2r\delta_o\delta_p / \left[\delta_o^2 + \delta_p^2 + (\bar{O} - \bar{P})^2 \right] \quad (2)$$

式中, \bar{O} 为平均测量值, \bar{P} 为平均预测值, δ_o 为测量值方差, δ_p 为预测值方差, r 为测量值和验证值的 Pearson 相关性系数。通过计算重复迭代 100 次

的 RF 模型预测值的标准偏差 (Standard deviation, SD), 并生成 STN 的 SD 图来评估模型预测的不确定性^[29-30]。SD 表示预测值与测量值之间的离散程度,可能由指标计算误差或建模误差引起。

2 结果

2.1 土壤全氮描述性统计特征

寻乌县 STN 的描述性统计结果表明,土壤 TN 含量范围为 0.42~2.52 g·kg⁻¹, 均值为 1.21 g·kg⁻¹, 低于全国的平均水平^[31]; 标准差为 0.35 g·kg⁻¹, 偏度系数为 1.10, 峰度系数为 3.39, 变异系数为 26.88%, 属于中等程度的空间变异。

不同土壤类型、土地利用类型和成土母质下样点 STN 含量的分布情况如图 2 所示。不同土壤类型下 STN 的分布差异不大,黄壤的整体 STN 含量分布要高于红壤和水稻土,不同土地利用类型 STN 含量中位数大小依次为旱地>水田>林地>园地,5 种成土母质的 STN 分布有较明显的差异,紫色岩类风化物母质的 STN 含量要低于其他 4 种母质。

2.2 环境协变量的选择

为避免变量共线性对模型预测精度的影响,计算了所有环境协变量的方差膨胀因子 VIF,并去除了 VIF>10 的变量 (x 坐标、地形起伏度、年平均气温、归一化植被指数、绿带反射率和裸土指数),之后计算了 STN 与剩余变量之间的 Pearson 相关系数。结果表明,STN 与土壤 S ($r=0.72$)、高程 ($r=0.32$)、坡度 ($r=0.19$)、地形粗糙度 ($r=0.17$) 绿色叶绿素指数 ($r=0.12$) 等变量表现出正相关关系,与 y 坐标 ($r=-0.17$)、年平均降水 ($r=-0.12$) 等因子则有着负相关关系。

2.3 不同模型预测精度及不确定性评价

RFRK 需要对 RF 模型的预测残差进行普通克里格插值,因此,对 STN 的 RF 模型 100 次预测结果的残差分别进行半方差分析,得到理论最优半方差函数模型及参数后,再使用普通克里格法对其进行空间插值,得到 RF 模型 100 次预测结果残差的空间分布。

在每次随机抽取 30% 的样点数据作为验证集,并进行 100 次模型迭代后,对 RF 模型和 RFRK 模型的预测精度进行了评估,结果表明 RF 模型的预测精度指标的均值 (MAE=0.1570 g·kg⁻¹, RMSE=

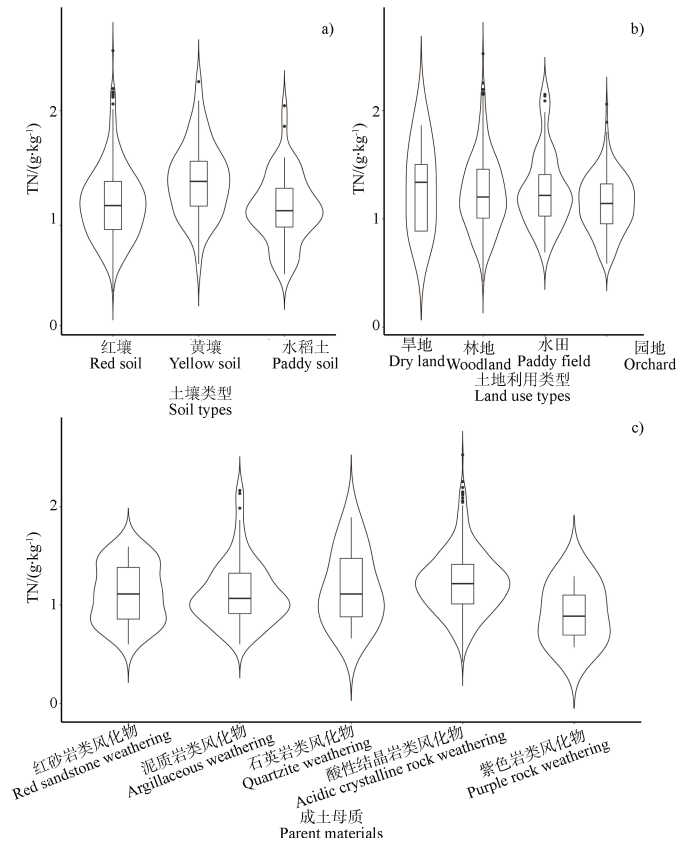


图 2 不同土壤类型、土地利用类型和成土母质下土壤全氮含量小提琴图

Fig. 2 Violin plot of soil TN content in different soil types, land use types and parent materials

0.210 8 $\text{g}\cdot\text{kg}^{-1}$, $\text{CCC}=0.761\ 3$, $R^2=0.629\ 1$) 均优于 RFRK 模型 ($\text{MAE}=0.168\ 2\ \text{g}\cdot\text{kg}^{-1}$, $\text{RMSE}=0.226\ 7\ \text{g}\cdot\text{kg}^{-1}$, $\text{CCC}=0.688\ 1$, $R^2=0.571\ 9$), 这表明将预测残差作为误差项加入 RF 模型并没有改善其预测效果, 反而使 RF 模型的预测精度下降。

使用 RF 模型迭代 100 次的预测值标准差来评估模型预测的不确定性。RF 模型预测值标准差的范围为 $0.01\sim 0.22\ \text{g}\cdot\text{kg}^{-1}$, 均值为 $0.04\ \text{g}\cdot\text{kg}^{-1}$, 总体上模型的预测不确定性较低, 仅在研究区的西北部、西南部、东部以及南部等海拔较大的区域表现出较大的预测不确定性, 这可能是因为环境协变量在海拔较大区域有着较强的空间变异性。

2.4 环境协变量重要性评估

对 RF 模型中的环境协变量计算了其对 STN 预测结果的相对重要性 (Relative importance, RI), 并将其转化为百分比 (图 3)。结果表明, 土壤硫, 高程, 年平均降水, 绿色叶绿素指数, y 坐标, 多尺度谷底平坦度、坡度、地形粗糙度这些因子是影响研究区 STN 含量的主要协变量, 占 RI 的 87.44%;

此外, 在所有类别的协变量中, 土壤理化因子的解释能力最高 (占 RI 的 48.52%), 其次为地形因子 (24.06%)、遥感因子 (9.76%)、气候因子 (6.88%) 和地理坐标 (5.45%)。

2.5 土壤全氮对各环境协变量的响应变化

为探究 STN 与环境协变量之间的响应变化关系, 随机选择了 100 次迭代模型其中的 1 个 RF 模型, 绘制了各环境协变量对 STN 的响应变化曲线 (图 4), 该曲线代表了 STN 与每个环境变量之间的变化关系, 可以看出 STN 与各变量的响应关系差异较大, 且大致趋势与相关性分析结果一致, 其中, STN 含量随土壤硫、高程、绿色叶绿素指数、坡度、地形粗糙度、距河流距离、pH 和波段第一主成分等因子值的增大而表现出非线性增加, 与年平均降水则表现为相反趋势的非线性变化, 而与 y 坐标、多尺度谷底平坦度、距公路距离、绿度指数等因子表现为波动变化趋势。在影响 STN 含量主要协变量的响应变化曲线中, 高程、坡度、地形粗糙度、多尺度谷底平坦度和 pH 增大到一定值后, STN 含量不再响应其变化。

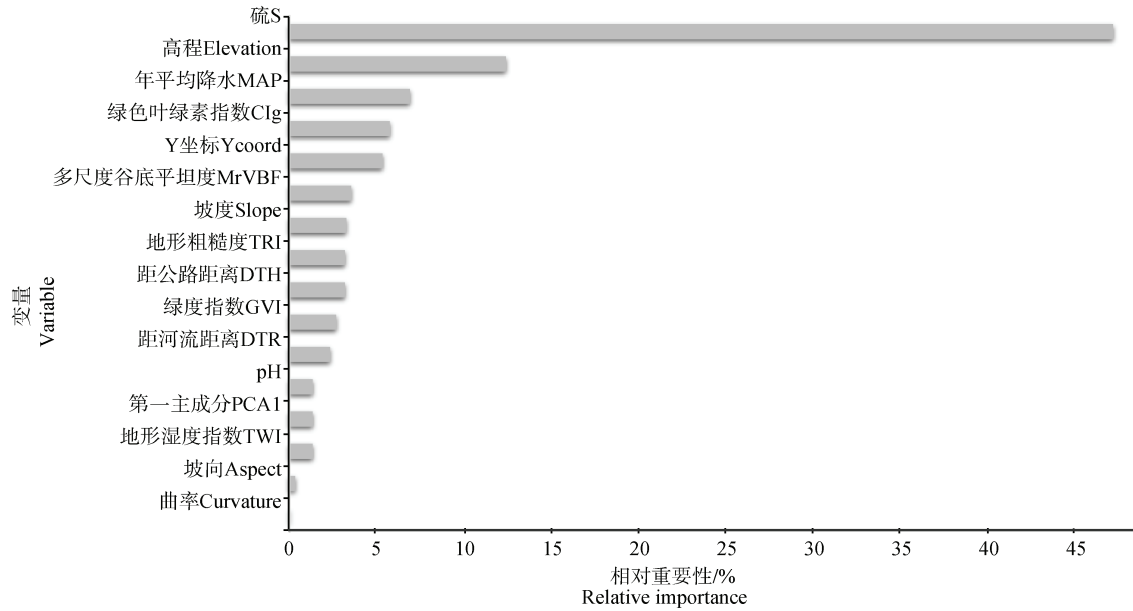
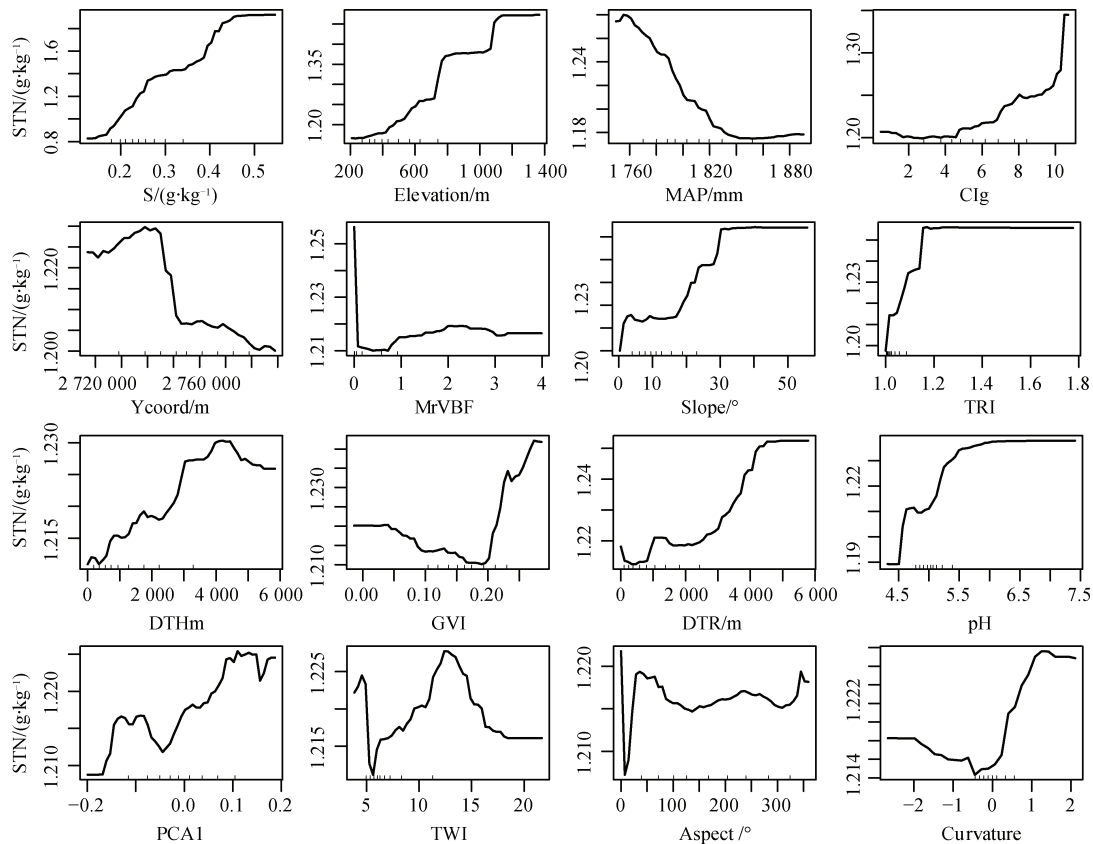


图 3 RF 模型各变量相对重要性

Fig. 3 Relative importance of variables in the random forest (RF) model



S: 硫 Sulfur; Elevation: 高程; MAP: 年平均降水 Mean annual precipitation; Cig: 绿色叶绿素指数 Green chlorophyll index; Ycoord: Y 坐标; MrVBF: 多尺度谷底平坦度 Multiresolution index of valley bottom flatness; Slope: 坡度; TRI: 地形粗糙度 Terrain ruggedness index; DTH: 距公路距离 Distance to highway; GVI: 绿度指数 Green Vegetation Index; DTR: 距河流距离 Distance to river; PCA1: 第一主成分 The first principal component; TWI: 地形湿度指数 Topographic wetness index; Aspect: 坡向; Curvature: 曲率

图 4 土壤全氮含量与各环境协变量的响应曲线

Fig. 4 Response of soil total nitrogen content to the variation of each environmental predictive factor

2.6 土壤全氮空间分布特征

研究区 STN 的空间分布表现出较强的空间异质性 (图 5), RF 模型和 RFRK 模型两种方法预测的 STN 空间分布的总体变化趋势基本一致, 均表现为西北部、东北部和南部高, 中部和中北部低的空间格局, 且与高程的变化特征大致相

符; RF 模型和 RFRK 模型迭代 100 次的 STN 预测值均值的变化范围分别为 $0.70 \sim 2.05 \text{ g}\cdot\text{kg}^{-1}$ 和 $0.69 \sim 2.10 \text{ g}\cdot\text{kg}^{-1}$, 均值分别为 $1.23 \text{ g}\cdot\text{kg}^{-1}$ 和 $1.24 \text{ g}\cdot\text{kg}^{-1}$, 从结果中可以看出, RF 模型与 RFRK 模型的 STN 预测值范围基本一致, 与土壤样点数据的统计范围较为接近。

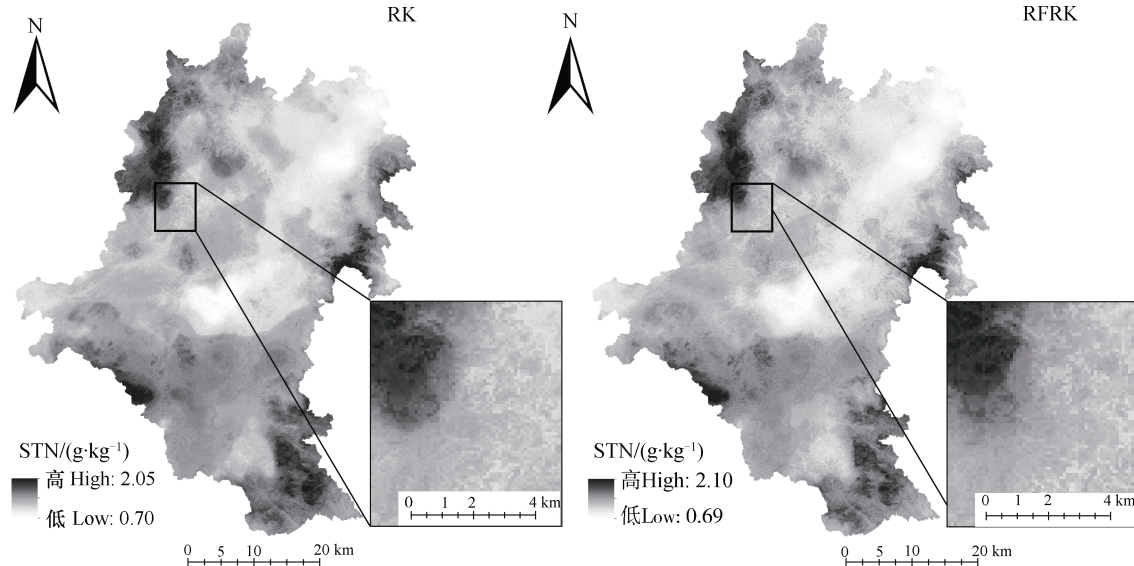


图 5 不同模型下研究区土壤全氮空间分布图

Fig. 5 Spatial distribution of soil total nitrogen (STN) by different model

3 讨论

3.1 影响土壤全氮分布的主要变量

在所有的环境协变量中, 土壤硫对研究区 STN 含量分布的影响最大, 这是因为土壤 S 元素和 N 元素的循环有着相互耦合性^[32], 从而直接影响 STN 的分布。从响应曲线 (图 5) 中可以看出, 土壤硫与 STN 表现出了正向耦合性。在施氮肥时, 含硫酸根离子的氮肥会使土壤中的 S 含量增多, 因此土壤硫和 STN 的空间分布呈现出相互影响的变化趋势。高程、坡度等地形因子对 STN 的影响仅次于土壤理化因子, 两者与 STN 含量的响应曲线变化的大致趋势也较为相似。有研究表明, 地形因子在土壤属性预测中有着重要的作用^[33-35], 其对 STN 的影响主要体现在影响地表水热交换^[36]和影响植被分布^[37]这两个方面。不同地形条件下, 表层土壤受到的风化淋溶作用存在差异, 从而影响 STN 的含量。而植被覆盖度较高处, 表层土壤有机质含量较高^[38], STN 等土壤养分元素易积累^[39], 且地形因子中的高程变化

对区域小气候特征也会造成一定的影响^[40]。随着海拔的上升, 气温会呈现出下降的趋势, 而 STN 在不同的土壤温度下, 分解速度也会存在差异^[41], 导致不同海拔条件下 STN 含量存在差异。遥感因子对研究区 STN 也有较大影响, 这与其他研究一致^[42-43]。有研究表明, 在植被覆盖度较高的区域, 植被与表层土壤中的氮含量较为相关^[44]。因此, 在植被茂密地区, 波段反射率和遥感指数可以作为预测 STN 的有效指标^[45]。本研究中, 绿色叶绿素指数和绿度指数等遥感因子对 STN 的影响较大, 但与其他研究相比, 影响程度较低^[19, 46], 这可能与本研究选取的遥感指数有关。由于共线性问题, 未将归一化植被指数和绿带反射率等遥感指数纳入到预测模型中; 气候因子中, 年平均降水对 STN 的影响主要体现在, 降水量的不同会导致地区的水文环境和土壤水分产生差异^[47], 影响土壤氮素的矿化速率, 从而影响 STN 的分布。地理坐标对 STN 也有一定的影响, 表明研究区 STN 的空间分布有着较强的空间异质性。Pearson 相关性分析结果表明, STN 与 y 坐标有着负

相关关系, 响应曲线中, STN 随着 y 坐标的增大表现出先增后减的变化趋势, 这与研究区特点有关。寻乌县属山地丘陵地形, 三面环山, 东北、西北和东南均有高山分布, 区域的高程差近 1 000 m, 因此研究区地理坐标的改变会使成土母质和植被产生较大的垂直性地带变化, 进而对 STN 产生较大的影响^[5]。

3.2 不同模型预测效果的对比

本研究中, RF 模型 100 次重复迭代的模型精度指标值 (MAE、RMSE、CCC、 R^2) 要优于 RFRK 模型, 这与其他研究结果有所不同^[17, 21], 原因在于, 虽然残差代表了 RF 模型在预测过程中无法解释的部分^[21], 但因为本研究中选取的环境协变量足够多, RF 模型的预测效果已较好, 预测残差的空间自相关性较弱, 使用普通克里格法对残差进行空间插值后并不能代表 RF 模型预测误差的空间分布趋势, 因此将残差的空间信息加入到 RF 模型的预测结果中并没有提高其预测精度, 反而使 RF 模型的预测精度有所下降。

3.3 预测精度的影响因素

在对 STN 进行建模预测时, 影响预测精度的因素主要有两点。第一, 部分协变量数据的精细化程度不高, 在使用的环境协变量中, 年平均降水因子使用的是 1 km × 1 km 分辨率的栅格数据, 而研究区为县域尺度, 该分辨率可能忽略了气温和降水的局部空间自相关性, 从而降低了 STN 的预测精度; 土壤理化因子中, 土壤硫和 pH 的栅格数据是利用点位数据, 采用普通克里格法进行空间插值得到, 而普通克里格法插值中的误差和不确定性也会对 STN 的预测效果产生影响, 且为了统一环境协变量的尺度而进行的栅格重采样也会导致协变量数据精度降低和信息丢失的问题。第二, 代表人为活动因素的协变量考虑的不够多, 所选取的环境协变量中, 除了距公路距离这个人为因素之外, 其他的多为自然因素, 虽然遥感植被因子能够在一定程度上代表人为活动, 但如果将更多的人为活动因素, 比如耕作制度、施肥方式等因素作为协变量纳入预测模型, 可能会在一定程度上提高模型预测精度和效果。

4 结 论

RF 和 RFRK 模型两种方法对寻乌县表层土壤的 STN 预测结果表明: 从空间分布来看, 两种方法

预测的 STN 空间分布大致相同, 总体上表现为西北部、东北部和南部高, 中部和中北部低的空间分布格局; 从预测精度来看, RF 模型的决定系数均值和一致性相关系数均值均高于 RFRK 模型, 而平均相对误差均值和均方根误差均值均小于 RFRK 模型, 表明 RF 模型的预测结果要优于 RFRK 模型; 在参与预测的环境协变量中, 土壤硫, 高程, 年平均降水, 绿色叶绿素指数, y 坐标, 多尺度谷底平坦度、坡度、地形粗糙度等因子是影响研究区 STN 分布的主要变量。

参考文献 (References)

- [1] Huang B, Sun W X, Zhao Y C, et al. Temporal and spatial variability of soil organic matter and total nitrogen in an agricultural ecosystem as affected by farming practices[J]. *Geoderma*, 2007, 139 (3/4): 336—345.
- [2] Reeves D W. The role of soil organic matter in maintaining soil quality in continuous cropping systems[J]. *Soil and Tillage Research*, 1997, 43 (1/2): 131—167.
- [3] Zhao Y. Temporal and spatial variations of nutrients in Baiyangdian Lake, North China[J]. *Journal of Environmental Informatics*, 2011, 17 (2): 102—108.
- [4] Peterjohn W T, Schlesinger W H. Nitrogen loss from deserts in the southwestern United States[J]. *Biogeochemistry*, 1990, 10 (1): 67—79.
- [5] Li L, Yao Y F, Qin F C, et al. Spatial distribution pattern of soil nitrogen in Huanghuadianzi watershed and related affecting factors[J]. *Chinese Journal of Applied Ecology*, 2015, 26 (5): 1306—1312. [李龙, 姚云峰, 秦富仓, 等. 黄花甸子流域土壤全氮含量空间分布及其影响因素[J]. *应用生态学报*, 2015, 26 (5): 1306—1312.]
- [6] Wang S Y, Lu P, Wang J L, et al. Spatial variability and distribution of soil organic matter and total nitrogen at different scales: A case study in Pinggu County, Beijing[J]. *Acta Ecologica Sinica*, 2008, 28 (10): 4957—4964. [王淑英, 路苹, 王建立, 等. 不同研究尺度下土壤有机质和全氮的空间变异特征——以北京市平谷区为例[J]. *生态学报*, 2008, 28 (10): 4957—4964.]
- [7] Xu G C, Li Z B, Li P, et al. Spatial distribution of soil total nitrogen in a typical watershed of the middle Danjiang river[J]. *Acta Geographica Sinica*, 2012, 67 (11): 1547—1555. [徐国策, 李占斌, 李鹏, 等. 丹江中游典型小流域土壤总氮的空间分布[J]. *地理学报*, 2012, 67 (11): 1547—1555.]
- [8] Guan F Y, Xia M P, Tang X L, et al. Spatial variability of soil nitrogen, phosphorus and potassium contents in Moso bamboo forests in Yong'an City, China[J]. *Catena*, 2017, 150: 161—172.

- [9] McBratney A B, Mendonça Santos M L, Minasny B. On digital soil mapping[J]. *Geoderma*, 2003, 117 (1/2): 3—52.
- [10] Ellili Y, Walter C, Michot D, et al. Mapping soil organic carbon stock change by soil monitoring and digital soil mapping at the landscape scale[J]. *Geoderma*, 2019, 351: 1—8.
- [11] Shen Q S, Wang Y, Wang X R, et al. Comparing interpolation methods to predict soil total phosphorus in the Mollisol area of Northeast China[J]. *Catena*, 2019, 174: 59—72.
- [12] Duan L X, Li Z W, Xie H X, et al. Large-scale spatial variability of eight soil chemical properties within paddy fields[J]. *Catena*, 2020, 188: 104350.
- [13] Owusu S, Yigini Y, Olmedo G F, et al. Spatial prediction of soil organic carbon stocks in Ghana using legacy data[J]. *Geoderma*, 2020, 360: 114008.
- [14] Wang D, Li X X, Zou D F, et al. Modeling soil organic carbon spatial distribution for a complex terrain based on geographically weighted regression in the eastern Qinghai-Tibetan Plateau[J]. *Catena*, 2020, 187: 104399.
- [15] Were K, Bui D T, Dick Ø B, et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afrotropical landscape[J]. *Ecological Indicators*, 2015, 52: 394—403.
- [16] Zeraatpisheh M, Ayoubi S, Jafari A, et al. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran[J]. *Geoderma*, 2019, 338: 445—452.
- [17] Chen S C, Liang Z Z, Webster R, et al. A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution[J]. *Science of the Total Environment*, 2019, 655: 273—283.
- [18] Jiang S P, Zhang H Z, Zhang R L, et al. Research on spatial distribution of soil organic matter in Hainan Island based on three spatial prediction models[J]. *Acta Pedologica Sinica*, 2018, 55 (4): 1007—1017. [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.]
- [19] Zhang Y, Sui B, Shen H O, et al. Mapping stocks of soil total nitrogen using remote sensing data: A comparison of random forest models with different predictors[J]. *Computers and Electronics in Agriculture*, 2019, 160: 23—30.
- [20] Wang L, Wu W, Liu H B. Digital mapping of topsoil pH by random forest with residual kriging (RFRK) in a hilly region[J]. *Soil Research*, 2019, 57 (4): 387.
- [21] Guo P T, Li M F, Luo W, et al. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach[J]. *Geoderma*, 2015, 237/238: 49—59.
- [22] Hengl T, Heuvelink G B M, Stein A. A generic framework for spatial prediction of soil variables based on regression-kriging[J]. *Geoderma*, 2004, 120 (1/2): 75—93.
- [23] Scarpone C, Schmidt M G, Bulmer C E, et al. Modelling soil thickness in the critical zone for Southern British Columbia[J]. *Geoderma*, 2016, 282: 59—69.
- [24] Moore I D, Gessler P E, Nielsen G A, et al. Soil attribute prediction using terrain analysis[J]. *Soil Science Society of America Journal*, 1993, 57 (2): 443—452.
- [25] Gao J B, Jiao K W, Wu S H. Revealing the climatic impacts on spatial heterogeneity of NDVI in China during 1982—2013[J]. *Acta Geographica Sinica*, 2019, 74 (3): 534—543. [高江波, 焦珂伟, 吴绍洪. 1982—2013 年中国植被 NDVI 空间异质性的气候影响分析[J]. *地理学报*, 2019, 74 (3): 534—543.]
- [26] Chatterjee S, Hadi A S. Regression analysis by example[M]//Oboken, N J, USA: John Wiley & Sons, Inc., 2006.
- [27] Breiman L. Random forests[J]. *Machine Learning*, 2001, 45 (1): 5—32.
- [28] Ließ M, Glaser B, Huwe B. Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models[J]. *Geoderma*, 2012, 170: 70—79.
- [29] Song X D, Wu H Y, Ju B, et al. Pedoclimatic zone-based three-dimensional soil organic carbon mapping in China[J]. *Geoderma*, 2020, 363: 114145.
- [30] Yang R M, Rossiter D G, Liu F, et al. Predictive mapping of topsoil organic carbon in an alpine environment aided by landsat TM[J]. *PLoS One*, 2015, 10 (10): e0139042.
- [31] Li Q Q, Yue T X, Fan Z M, et al. Spatial simulation of topsoil TN at the national scale in China[J]. *Geographical Research*, 2010, 29 (11): 1981—1992. [李启权, 岳天祥, 范泽孟, 等. 中国表层土壤全氮的空间模拟分析[J]. *地理研究*, 2010, 29 (11): 1981—1992.]
- [32] Ludwig J A, Whitford W G, Cornelius J M. Effects of water, nitrogen and sulfur amendments on cover, density and size of Chihuahuan Desert ephemerals[J]. *Journal of Arid Environments*, 1989, 16 (1): 35—42.
- [33] Xie E Z, Zhao Y C, Lu F Y, et al. Comparison analysis of methods for prediction of spatial distribution of soil organic matter contents in farmlands south Jiangsu, China[J]. *Acta Pedologica Sinica*, 2018, 55 (5): 1051—1061. [谢恩泽, 赵永存, 陆访仪, 等. 不同方法预测苏南农田土壤有机质空间分布对比研究[J]. *土壤学报*, 2018, 55 (5): 1051—1061.]
- [34] Yang Q P, Wu W, Liu H B. Prediction of spatial distribution of soil available iron in a typical hilly

- farmland using terrain attributes and random forest model[J]. *Chinese Journal of Eco-Agriculture*, 2018, 26 (3): 422—431. [杨其坡, 武伟, 刘洪斌. 基于地形因子和随机森林的丘陵区农田土壤有效铁空间分布预测[J]. *中国生态农业学报*, 2018, 26 (3): 422—431.]
- [35] Guo P T, Li M F, Luo W, et al. Prediction of soil total nitrogen for rubber plantation at regional scale based on environmental variables and random forest approach[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2015, 31 (5): 194—202. [郭澎涛, 李茂芬, 罗微, 等. 基于多源环境变量和随机森林的橡胶园土壤全氮含量预测[J]. *农业工程学报*, 2015, 31 (5): 194—202.]
- [36] Cheng X F, Shi X Z. Relationship between total nitrogen and topographic, parent material in typical subtropical region—A case study in Xingguo County, Jiangxi Province[J]. *Journal of Soil Water Conservation*, 2004, 18 (2): 137—139. [程先富, 史学正. 亚热带典型地区土壤全氮和地形、母岩的关系研究——以江西省兴国县为例[J]. *水土保持学报*, 2004, 18 (2): 137—139.]
- [37] Yang R M, Zhang G L, Liu F, et al. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem[J]. *Ecological Indicators*, 2016, 60: 870—878.
- [38] Yang Q Y. *Jiangxi Natural Geography* [M]. Beijing: Fangzhi Publishing Press, 2003. [杨巧言. *江西省自然地理志*[M]. 北京: 方志出版社, 2003.]
- [39] Wei X R, Min D H, Shao M G, et al. Changes in soil properties and the availability of soil micronutrients after 18 years of cropping and fertilization[J]. *Soil and Tillage Research*, 2006, 91 (1/2): 120—130.
- [40] Baxter S J, Oliver M A. The spatial prediction of soil mineral N and potentially available N using elevation[J]. *Geoderma*, 2005, 128 (3/4): 325—339.
- [41] Wang S, Jin X X, Adhikari K, et al. Mapping total soil nitrogen from a site in northeastern China[J]. *Catena*, 2018, 166: 134—146.
- [42] Zhang S M, Wang Z M, Zhang B, et al. Prediction of spatial distribution of soil nutrients using terrain attributes and remote sensing data[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2010, 26 (5): 188—194. [张素梅, 王宗明, 张柏, 等. 利用地形和遥感数据预测土壤养分空间分布[J]. *农业工程学报*, 2010, 26 (5): 188—194.]
- [43] Sumfleth K, Duttman R. Prediction of soil property distribution in paddy soil landscapes using terrain data and satellite information as indicators[J]. *Ecological Indicators*, 2008, 8 (5): 485—501.
- [44] Yimer F, Ledin S, Abdelkadir A. Soil property variations in relation to topographic aspect and vegetation community in the south-eastern Highlands of Ethiopia[J]. *Forest Ecology and Management*, 2006, 232 (1/2/3): 90—99.
- [45] Bardgett R D, Streeter T C, Bol R. Soil microbes compete effectively with plants for organic-nitrogen inputs to temperate grasslands[J]. *Ecology*, 2003, 84 (5): 1277—1287.
- [46] Wang S, Zhuang Q L, Jin X X, et al. Predicting soil organic carbon and soil nitrogen stocks in topsoil of forest ecosystems in northeastern China using remote sensing data[J]. *Remote Sensing*, 2020, 12 (7): 1115.
- [47] Wang C, Wang S, Fu B J, et al. Precipitation gradient determines the tradeoff between soil moisture and soil organic carbon, total nitrogen, and species richness in the Loess Plateau, China[J]. *Science of the Total Environment*, 2017, 575: 1538—1545.

(责任编辑: 檀满枝)