

doi: 10.6046/zrzyyg.2022294

引用格式: 钟骁勇,李洪义,郭冬艳,等. 基于多源环境变量和随机森林模型的江西省耕地土壤 pH 值空间预测[J]. 自然资源遥感,2023,35(4):178-185. (Zhong X Y, Li H Y, Guo D Y, et al. Spatial distribution prediction of soil pH in arable land of Jiangxi Province based on multi-source environmental variables and the random forest model[J]. Remote Sensing for Natural Resources, 2023,35(4):178-185.)

基于多源环境变量和随机森林模型的 江西省耕地土壤 pH 值空间预测

钟骁勇^{1,2}, 李洪义^{3,4}, 郭冬艳², 谢模典^{3,4}, 赵婉如^{3,4}, 胡碧峰^{3,4}

(1. 江西财经大学财税与公共管理学院, 南昌 330013; 2. 中国自然资源经济研究院, 北京 101149; 3. 江西财经大学旅游与城市管理学院, 南昌 330013; 4. 江西财经大学财经数据科学重点实验室, 南昌 330013)

摘要: 研究多源环境变量条件下随机森林(random forest, RF)模型和普通克里格(ordinary Kriging, OK)模型对大尺度耕地土壤 pH 值空间预测的性能差异, 研究分析 RF 模型对提升土壤 pH 值预测精度的参考价值。以中国江西省为研究区, 基于气候、地形和植被等环境协变量信息, 结合土壤属性和耕地利用条件, 利用 RF 模型对江西省耕地土壤 pH 值进行空间预测, 识别土壤 pH 值空间变异的影响因素, 并与 OK 模型计算精度进行对比。结果表明, 增加土壤属性和耕地利用条件作为环境变量的 RF-A 模型预测耕地土壤 pH 值的精度优于以地形、气候、植被属性作为环境变量的 RF-B 模型和 OK 模型的预测结果, 气候因素是决定土壤 pH 值的最重要因素, 地形地貌因子和人为因素对 pH 值变异也有重要影响。研究结果表明该方法对提升大尺度耕地土壤 pH 值预测制图精度具有一定的理论和现实意义。

关键词: 耕地; 土壤 pH 值; 随机森林; 江西省; 影响因素

中图法分类号: TP 79 **文献标志码:** A **文章编号:** 2097-034X(2023)04-0178-08

0 引言

耕地资源是人类赖以生存的重要物质基础和生产资料, 土壤 pH 值作为耕地土壤关键属性之一, 不仅是影响农作物生长发育的重要因素, 也是影响土壤养分有效性和土壤重金属元素迁移转换的关键因子^[1]。但是, 随着人类社会工业化进程加快, 酸性气体(SO₂)、颗粒污染物排放以及氮肥的大量施用等, 使得土壤酸化在许多地方成为一个普遍的趋势^[2-4]。张福锁院士团队在 Science 上发表的研究结果表明, 从 20 世纪 80 年代—21 世纪初, 中国主要农作物产区土壤 pH 值明显下降^[5]。土壤酸化会增加土壤重金属的活性, 不利于农作物健康生长, 进而威胁中国的粮食生产安全, 这一问题在中国南方

地区尤为突出^[6-7]。因此, 获得准确的土壤 pH 值空间分布信息对于遏制土壤酸化趋势、保护土壤环境质量和保障粮食生产安全具有十分重要的意义和价值。

国内外许多研究学者使用地统计学等空间插值方法用于土壤属性空间信息的预测, 主要有普通克里格插值法(ordinary Kriging, OK)、反距离权重插值法(inverse distance weighted, IDW)、普通最小二乘法(ordinary least squares, OLS)、地理加权回归(geographically weighted regression, GWR)等^[6,8-12]。近年来, 随着计算机技术、卫星遥感和土壤近地传感器等技术的快速发展, 机器学习算法应用于数字土壤制图(digital soil mapping, DSM)成为一个新的研究热点^[13-21]。张甘霖等^[22]指出: DSM 未来有从地统计学走向机器学习为主导模型的趋势。随机森林

收稿日期: 2022-07-27; 修订日期: 2023-03-02

基金项目: 国家自然科学基金项目“新疆滴灌棉田水盐四维时空变异性与管理风险评价研究”(编号: 42071068)、江西省高校人文社会科学研究项目“江西省自然资源生态系统服务价值评估”(编号: GL21217)、江西省教育厅科技项目“近 20 年江西省耕地生态系统服务价值时空变化及驱动因子分析”(编号: GJJ210541)和江西省社会科学基金项目“江西省耕地资源资产价值核算及时空变异特征研究”(编号: 21YJ43D)共同资助。

第一作者: 钟骁勇(1990-), 男, 博士, 副研究员, 主要从事自然资源权益管理和土地资源核算研究。Email: zhongxy0509@126.com。

通信作者: 胡碧峰(1992-), 男, 博士, 副教授, 主要从事土地生态与环境、资源环境遥感与信息技术、时空统计分析与建模研究。Email: hbfddmm297@163.com。

(random forest, RF) 作为一种以数据为驱动的机器学习方法^[23],具有训练速度快、对多元线性不敏感、易于处理缺失数据等优点,还可避免对模型的过度拟合^[24-26]。RF 能够获取和表达土壤属性空间变化与环境变量的复杂非线性关系,并根据这种关系推测土壤属性空间分布,还能以相对较少的样本量取得较好的预测精度^[27]。目前已有研究将 RF 模型用于土壤有机碳、土壤类型和土壤质地等属性空间预测和制图^[28-31,26]。但是目前将 RF 应用于土壤 pH 值空间预测制图的研究还较为少见,尤其是国内运用 RF 模型开展大尺度耕地 pH 值空间预测分析的研究鲜有报道。

因此,本文以土壤酸化现象显著的江西省为例,使用 RF 算法和环境协变量信息开展耕地土壤 pH 值的预测研究。同时将预测结果与使用 OK 方法插值结果进行对比,验证使用 RF 算法开展大尺度耕地土壤 pH 值预测研究的可行性和可靠性,为江西省土壤酸化问题的治理和农业生产管理提供数据基础和依据。

1 研究区概况及数据源

1.1 研究区概况

江西省位于中国东南部,长江中下游南岸,位于 E113°34'36"~118°28'58",N24°29'14"~30°04'41"之间(图 1),属于亚热带季风性湿润气候,是我国重要的粮食、油料、棉花、蔬菜和水果生产基地。江西地貌类型以山地、丘陵为主,山地占全省面积的 36%,丘陵占 42%,平原占 12%,水域占 10%。主要土壤类型为偏酸性的红壤,面积达 93 111.32 km²,约占全省总面积的 56%,具有较为明显的地带性和地域分布规律。

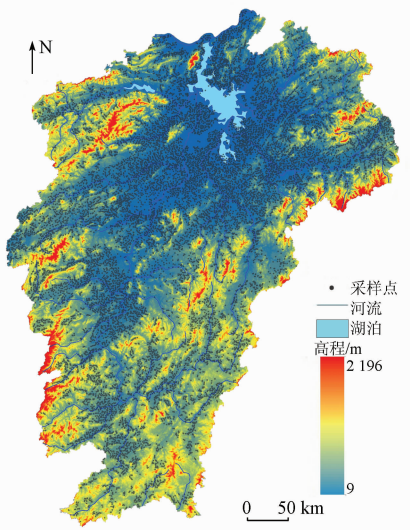


图 1 研究区土壤采样点分布

Fig.1 Sampling locations in study area

1.2 数据源及预处理

采集土壤样点 16 582 个,同时选取样点的土壤属性以及植被指数、地形变量、气候变量和耕地利用条件作为耕地土壤 pH 值空间预测的环境变量。其中土壤属性包括土壤类型、有机质、有效磷、速效钾、全氮、全磷、全钾、阳离子交换量、成土母质和耕层质地,数据来源于 2018 年江西省耕地质量等别数据库、1980 年和 2010 年江西省农业测土配方项目数据。归一化植被指数、地形变量的地貌类型和高程以及气象因子的年均气温和年降水量均通过资源环境科学数据中心(<http://www.resdc.cn/>)网站下载获取,影像空间分辨率均为 30 m。为了更好地反映出江西省地形条件对土壤属性变化的影响,选择高程、坡度、坡向、剖面曲率、平面曲率、沟谷深度、地形湿度指数、垂直到沟谷距离和多分辨率谷底平坦度等地形因子,分别借助 ArcGIS 10.2 和 SAGA 7.6.2 软件从数字高程模型(digital elevation model, DEM)数据中提取。耕地利用条件包括灌溉保证率、排水条件、氮肥用量、磷肥 P₂O₅用量、钾肥 K₂O 用量、秸秆还田比例和秸秆还田量,数据来源于 2018 年江西省耕地质量等别数据库(表 1)。

表 1 土壤 pH 值空间预测的环境变量及数据来源

Tab.1 Data sources of environmental covariates used for predicting pH		
变量类别	具体指标	数据来源
地形变量	地貌类型、高程、坡度、坡向、剖面曲率、平面曲率、沟谷深度、地形湿度指数、垂直到沟谷距离和多分辨率谷底平坦度	中国科学院资源环境科学数据中心 (http://www.resdc.cn/)
气候变量	年均气温、年降水量	中国科学院资源环境科学数据中心 (http://www.resdc.cn/)
植被指数	归一化植被指数	中国科学院资源环境科学数据中心 (http://www.resdc.cn/)
土壤属性	土壤类型、有机质、有效磷、速效钾、全氮、全磷、全钾、阳离子交换量、成土母质和耕层质地	2018 年江西省耕地质量等别数据库、1980 年和 2010 年江西省农业测土配方项目实测数据
耕地利用条件	灌溉保证率、排水条件、氮肥用量、磷肥 P ₂ O ₅ 用量、钾肥 K ₂ O 用量、秸秆还田比例和秸秆还田量	2018 年江西省耕地质量等别数据库

2 研究方法

2.1 环境变量的组合和筛选

为全面分析土壤 pH 值空间变异规律以及探索不同环境变量对模型预测精度的贡献程度,将表 1

中 30 个环境变量分成 2 种不同的组合：①以地形变量、气候变量和植被指数等遥感数据为基础,增加土壤属性和耕地利用条件等实测数据进行模型训练(RF-A);②采用传统的地形变量、气候变量和植被指数等遥感数据训练模型(RF-B)。

2.2 RF 模型构建

基于采集到的 16 582 个土壤样点数据,按照 8:2,7:3和 6:4的比例分为训练集和验证集。RF 模型采用 bootstrap 方法对样本进行放回抽样,没有被抽取的样本会自动生成一个对照集,因此不需要进行交叉验证。RF 模型中 2 个可调参数决策树数量 *ntree* 和节点分裂次数 *mtry* 决定了模型的配置。

2.3 精度分析与比较

RF 模型精度评价选用皮尔逊相关系数 *r*、平均误差 *ME*、绝对平均误差 *MAE* 和均方根误差 *RMSE* 等 4 个统计指标,其计算公式分别为:

$$r = \frac{\sum_{i=1}^n (p_i - \bar{p_i}) (o_i - \bar{o_i})}{\sqrt{\sum_{i=1}^n (p_i - \bar{p_i})^2} \sqrt{\sum_{i=1}^n (o_i - \bar{o_i})^2}}, \quad (1)$$

$$ME = \frac{1}{n} \sum_{i=1}^n (p_i - o_i), \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |p_i - o_i|, \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2}, \quad (4)$$

式中: *p_i* 和 *o_i* 分别为样点的预测值和观测值; *p̄_i* 和 *ō_i* 分别为样点预测值和观测值的平均值; *n* 为样本总数。并将 RF 预测结果与常用的地统计插值方法 OK 模型结果进行对比,模型输出结果的 *ME*,*MAE* 和 *RMASE* 越接近 0,*r* 越接近 1,表明模型的预测精度越高。

3 结果与分析

3.1 土壤 pH 值的描述性统计分析

通过江西省耕地土壤 pH 值描述性统计结果可知,土壤 pH 值范围在 4.00~8.30 之间,平均值为 5.20,中位数为 5.10,标准差为 4.91,变异系数为 9.43%,表明江西省耕地土壤 pH 值呈弱变异。参考通用标准将江西省土壤 pH 值划分为 7 个等级:极强酸性(pH<4.5)、强酸性[4.5,5.5)、酸性[5.5,6.5)、中性[6.5,7.5)、碱性[7.5,8.5)、强碱性[8.5,9.5)和极强碱性(pH≥9.5)。结果显示,江西省耕地土壤 pH 值样点中,极强酸性样本数 67 个,占比 0.43%;强酸性样本数 12 718 个,占比

76.69%;酸性样本数 3 301 个,占比 19.90%;中性样本数 435 个,占比 2.62%;碱性样本数 61 个,占比 0.36%;没有强碱性样本。综上可知,江西省耕地土壤总体呈现明显的酸性特征。

3.2 RF 模型最优参数分析

通过逐次试验,确定 RF 模型中 *mtry* 和 *ntree* 参数的最优值。固定 *mtry* (分别设为 2,4,6 和 8),逐次调整 *ntree* (分别设为 300,600,900,1 200 和 1 500),进行 4 组 20 次试验。为避免过拟合问题,通过比较训练集和验证集的相关系数(*r* 值),根据两者最为接近的结果确定最优预测模型的参数。结果表明(表 2):当 *mtry* = 8,*ntree* = 900 时,pH 值预测模型的训练集和验证集 *r* 值最为接近,表明此时的模型稳定性最佳。

表 2 RF 模型中节点分裂次数和决策树数量的筛选
Tab.2 Optimization of critical parameters for RF model

组别	<i>mtry</i>	<i>ntree</i>	训练集 <i>r</i>	验证集 <i>r</i>
1 组	2	300	0.969	0.551
		600	0.970	0.558
		900	0.970	0.556
		1 200	0.970	0.559
		1 500	0.970	0.555
2 组	4	300	0.970	0.567
		600	0.970	0.574
		900	0.970	0.573
		1 200	0.971	0.573
		1 500	0.971	0.572
3 组	6	300	0.969	0.579
		600	0.969	0.578
		900	0.970	0.579
		1 200	0.970	0.580
		1 500	0.970	0.580
4 组	8	300	0.969	0.582
		600	0.970	0.582
		900	0.970	0.583
		1 200	0.970	0.582
		1 500	0.969	0.581

3.3 不同训练集的 RF 和 OK 模型预测精度比较

基于不同训练集运行模型所获得的预测结果有所不同(表 3)。从相关系数 *r* 看,不同训练集的 RF-A 和 RF-B 模型预测的土壤 pH 值与实测值的 *r* 较高(均大于 0.96)且差别不大,而 OK 模型 *r* 处于 0.65 左右,预测精度明显低于 RF 模型;从误差看,RF-A 模型的 *ME*,*MAE* 和 *RMSE* 值最小,而 OK 模型的误差明显更高。3 种模型的拟合能力随训练样本量呈现不同的变化趋势。总体上看,RF-A 模型略微优于 RF-B 模型,OK 模型明显弱于前 2 个模型,并且训练集和验证集抽样比在 7:3 时 RF-A 模型精度最佳。在验证数据集中,除了抽样比 7:3 时,OK 模型比 RF-A 模型精度高外,其他情况下 RF-

A 模型的 r 更高。这表明环境变量增加耕地土壤实测数据有利于提高模型预测精度。因此在土壤样点数据充足的情况下,应优先选择 RF - A 模型用于土壤 pH 值的建模预测制图。

表 3 使用不同训练集时 RF 和 OK 模型预测精度比较

Tab.3 Comparison of model performance with different training and validation dataset ratios									
预测模型	抽样比	训练集				验证集			
		r	ME	MAE	$RMSE$	r	ME	MAE	$RMSE$
RF - A	8:2	0.970	-0.004	0.121	0.167	0.599	-0.009	0.291	0.392
	7:3	0.970	-0.004	0.121	0.167	0.567	-0.014	0.299	0.401
	6:4	0.970	-0.004	0.122	0.168	0.566	-0.018	0.302	0.406
RF - B	8:2	0.966	-0.005	0.127	0.175	0.548	-0.023	0.307	0.413
	7:3	0.967	-0.005	0.127	0.175	0.525	-0.019	0.308	0.418
	6:4	0.967	-0.005	0.127	0.174	0.504	-0.014	0.314	0.425
OK	8:2	0.679	0.002	0.261	0.364	0.559	0.009	0.311	0.409
	7:3	0.683	0.001	0.259	0.357	0.588	0.007	0.289	0.394
	6:4	0.652	0.002	0.278	0.379	0.541	0.010	0.328	0.414

3.4 环境变量重要性分析

使用 RF 模型定量计算出不同变量对于土壤 pH 值预测的重要性(图 2)。由图 2 可知,气候因素如年均气温和年降水量是决定土壤 pH 值的最重要因素,这一结果与 Chen 等^[32]、卢宏亮等^[33]和王世航等^[34]的研究结果基本一致。此外,地形地貌因子

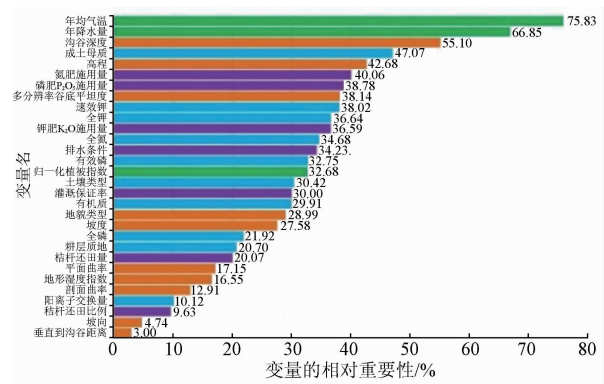


图 2 RF - A 模型变量相对重要性
Fig.2 Relative importance of covariates for RF - A model

如沟谷深度、高程和多分辨率谷底平坦度也是决定土壤 pH 值的重要因子。说明江西省土壤 pH 值分布与地形地貌特征高度相关。土壤理化性质对于土壤 pH 值也有着不可忽视的影响。人为因素如化肥施用量、排水和灌溉条件等也对土壤 pH 值有较大影响。但是部分因子如土壤类型、秸秆还田比例的重要性低于预期,需要后续进一步研究加以分析。

3.5 耕地土壤 pH 值空间预测制图

研究区耕地土壤 pH 值空间分布如图 3 所示,整体上使用 OK 模型、RF - A 和 RF - B 模型获得的研究区土壤 pH 值空间分布特征大体相同,其中土壤 pH 低值区主要分布在江西省中部鄱阳湖平原的南昌市、抚州市和吉安市等地区,高值区域主要分布于江西省西部和东部山地丘陵区域,如九江市、鹰潭市、萍乡市和上饶市等地区。结合图 1 高程信息可知,这一分布特征与江西省地形地势较为相似,表明江西省土壤 pH 值分布与地形地貌因素存在较大的

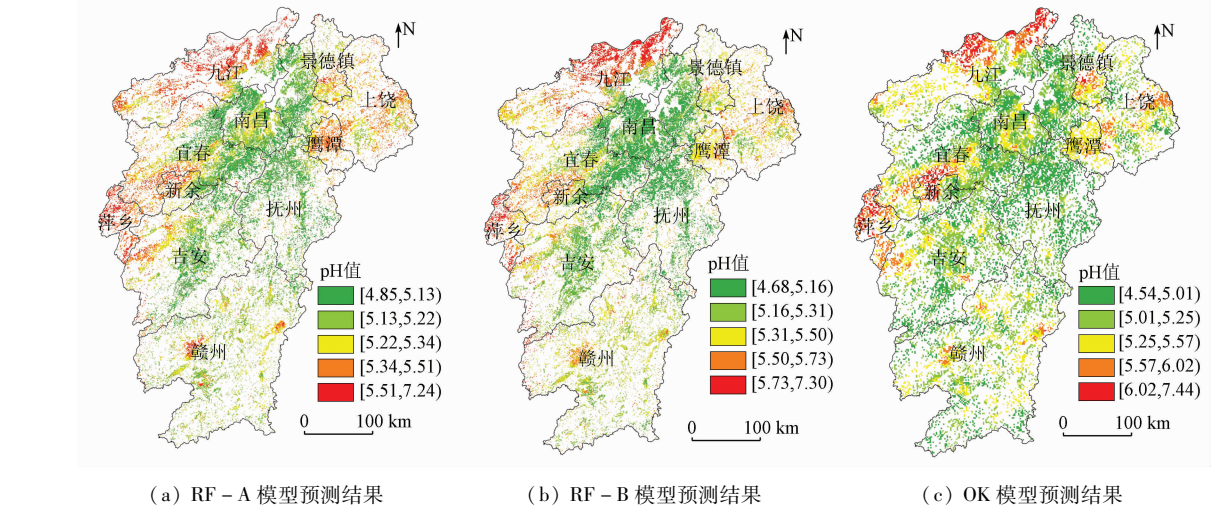


图 3 基于 RF - A,RF - B 和 OK 模型的耕地 pH 值空间分布
Fig.3 Distribution of soil pH produced by RF - A, RF - B and OK model

关联性,这一结果与图 2 结果相互印证。本文预测结果与周宏冀^[35]采用 OK 模型、协同克里格和 BP 神经网络所获得的预测结果也基本一致。

4 讨论

4.1 模型精度分析

根据表 3 评价结果所示,当训练集和验证集划分比例为 8:2 时,RF-A 模型对于 pH 值的预测结果精度最高,其 r 为 0.599, $RMSE$ 为 0.392,其预测值与实测值散点图见图 4。模型精度略低于齐雁冰等^[36]的研究结果。主要原因可能是江西省地貌类型复杂,山地丘陵地貌区所占比重较大,山区耕地分布面积小而破碎,土壤 pH 值预测难度相比地势平坦、地形地貌类型单一的小区域难度更大^[37]。虽然本文使用 RF 模型预测精度仍需进一步提高,但是预测结果基本符合预期,仍然可以为江西省耕地资源管理提供宝贵的数据基础。

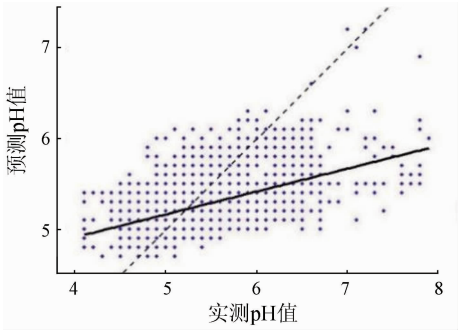


图 4 RF-A 模型预测值与实测值散点图
Fig.4 Scatter plot of soil pH predicted and measured by RF-A model

4.2 pH 值空间变异的影响因素分析

通过运用 R Studio 软件开展 pH 值空间变异的影响因素分析,结果显示江西省耕地土壤 pH 值与年均气温和年降水量关系密切(图 2),这一结论与 Chen 等^[32]、卢宏亮等^[33]和王世航等^[34]的研究结果基本一致。此外,沟谷深度、多分辨率谷底平坦度、高程、坡度和坡向等地形地貌因子也可以通过影响温度、光照、水分和热量的再分配以及母岩形成分布等影响土壤的发育,从而间接影响土壤 pH 值^[38]。因此,地形因子对于预测土壤 pH 值也具有很高的重要性(图 2)。土壤 pH 值与土壤有机质含量呈现极其显著负相关关系,而土壤 pH 值与阳离子交换量呈现极其显著的正相关关系,这一结果与戴万宏等^[39]、许亚琪^[40]、康婷等^[41]和王晖等^[42]的研究结果一致。耕层质地和土壤类型对于土壤 pH 值也有明显的影响,轻壤 pH 值明显高于其他耕层质地类型,黏土次之,不同土壤类型 pH 值差异也十分显著,其中草洲沙泥土、鳝泥棕红壤和棕色石灰土土壤 pH 值相对最高,这一结果与前人研究基本一致^[32]。需要注意的是,排水条件和灌溉保证率对于耕地土壤 pH 值也有不可忽视的影响,本研究结果表明,耕地排水能力不满足的情况下耕地土壤 pH 值最高,在排水能力充分满足和基本满足的耕地土壤 pH 值相对较低。同时灌溉能力不满足的耕地土壤 pH 值也高于灌溉能力基本满足和充分满足的耕地土壤,这一结果与朱丹妹等^[43]和 Ding 等^[44]的研究结果较为相符(图 5)。

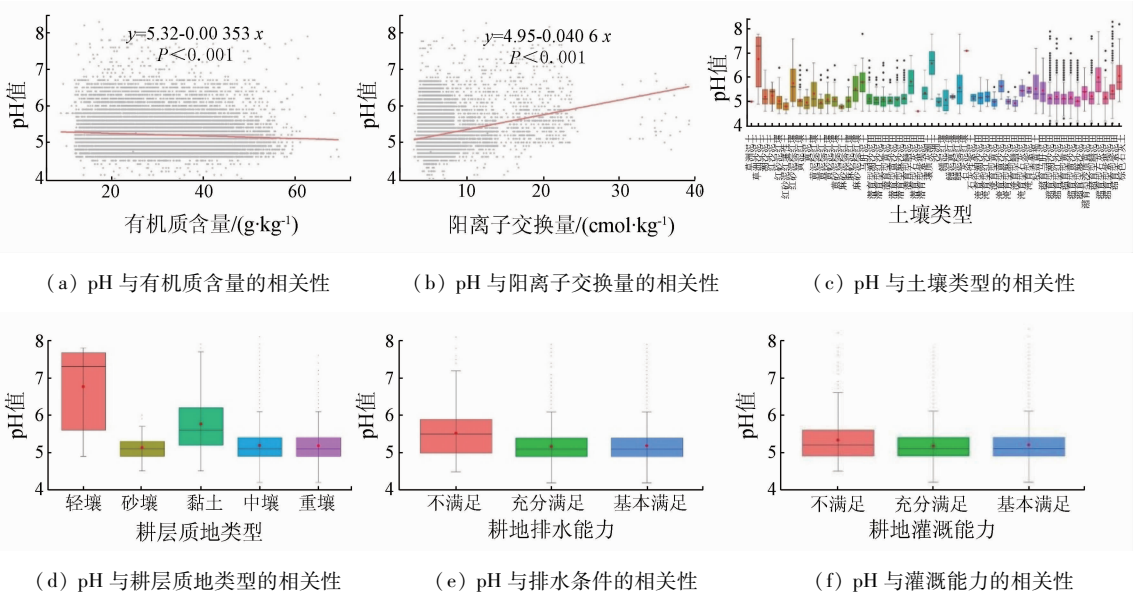


图 5 pH 值空间变异的影响因素关系
Fig.5 Relevance of pH and influencing factors

5 结论

综合使用地形、气候、植被指数、土壤属性和耕地利用条件等自然环境协变量和人为管理因素数据,基于大量土壤实测样点,分别训练使用 2 种不同变量组合、不同模型参数设置和不同训练抽样比数据集的随机森林模型,并预测江西省耕地土壤 pH 值,再将预测结果与普通克里格方法进行对比。得到如下结论:

1)增加土壤属性和耕地利用条件作为环境变量的 RF - A 模型预测耕地土壤 pH 值的精度优于以地形、气候、植被属性作为环境变量的 RF - B 模型和普通克里格的预测结果。

2)环境变量重要性分析结果显示,气候因素如年均气温和年降水量是决定土壤 pH 的最重要因素,地形因子、化肥施用量、排水和灌溉条件对土壤 pH 值变异的影响次之,分析结果与前人研究基本一致,表明相关结论较为可靠,对开展大尺度耕地土壤 pH 值预测具有一定的理论借鉴和实际应用价值。

但是,对于土壤 pH 值空间预测精度仍有待提高,今后将考虑增加环境协变量因子,提高基础数据空间分辨率等,进一步提高随机森林模型预测土壤 pH 值的精度。

参考文献 (References):

[1] 毛伟,李文西,高晖,等. 扬州市耕地土壤 pH 值 30 年演变及其驱动因子[J]. 植物营养与肥料学报,2017,23(4):883 - 893.
Mao W, Li W X, Gao H, et al. pH variation and the driving factors of farmlands in Yangzhou for 30 years[J]. Journal of Plant Nutrition and Fertilizers, 2017, 23(4): 883 - 893.

[2] Oh N H, Richter D D. Soil acidification induced by elevated atmospheric CO₂ [J]. Global Change Biology, 2004, 10(11): 1936 - 1946.

[3] Kirk G J D, Bellamy P H, Lark R M. Changes in soil pH across England and Wales in response to decreased acid deposition[J]. Global Change Biology, 2010, 16(11): 3111 - 3119.

[4] Tian D, Niu S. A global analysis of soil acidification caused by nitrogen addition[J]. Environmental Research Letters, 2015, 10(2): 024019.

[5] Guo J H, Liu X J, Zhang Y, et al. Significant acidification in major Chinese croplands[J]. Science, 2010, 327(5968): 1008 - 1010.

[6] Hu B F, Jia X L, Hu J, et al. Assessment of heavy metal pollution and health risks in the soil - plant - human system in the Yangtze River Delta, China[J]. International Journal of Environmental Research and Public Health, 2017, 14(9): 1042.

[7] Hu B F, Shao S, Fu Z, et al. Identifying heavy metal pollution hot spots in soil - rice systems: A case study in South of Yangtze River Delta, China[J]. Science of the Total Environment, 2019, 658: 614

- 625.

[8] 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 geostatistics[J]. Chinese Journal of Applied Ecology, 2018, 29(6): 1893 - 1901.

[9] Xia F, Hu B, Zhu Y, et al. Improved mapping of potentially toxic elements in soil via integration of multiple data sources and various geostatistical methods[J]. Remote Sensing, 2020, 12(22): 3775.

[10] Hu B, Zhou Y, Jiang Y, et al. Spatio - temporal variation and source changes of potentially toxic elements in soil on a typical plain of the Yangtze River Delta, China (2002 - 2012) [J]. Journal of Environmental Management, 2020, 271: 110943.

[11] 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.

[12] 张阳,屠乃美,谢会雅,等. 茶陵县烟稻复种区土壤养分时空变异特征[J]. 土壤通报, 2021, 52(2): 297 - 305.
Zhang Y, Tu N M, Xie H Y, et al. Spatial and temporal variation of soil nutrients in tobacco - rice multiple - cropping area of Chaling [J]. Chinese Journal of Soil Science, 2021, 52(2): 297 - 305.

[13] McBratney A B, Santos M L M, Minasny B. On digital soil mapping [J]. Geoderma, 2003, 117(1 - 2): 3 - 52.

[14] Teng H, Rossel R A V, Shi Z, et al. Updating a national soil classification with spectroscopic predictions and digital soil mapping [J]. Catena, 2018, 164: 125 - 134.

[15] Peng J, Biswas A, Jiang Q, et al. Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province, China [J]. Geoderma, 2019, 337: 1309 - 1319.

[16] 庞龙辉,刘峰,赵霞,等. 青海省表层土壤属性数字制图[J]. 土壤通报, 2019, 50(3): 505 - 513.
Pang L H, Liu F, Zhao X, et al. Digital mapping of topsoil attributes in Qinghai Province[J]. Chinese Journal of Soil Science, 2019, 50(3): 505 - 513.

[17] Arrouays D, Poggio L, Guerrero O A S, et al. Digital soil mapping and GlobalSoilMap. Main advances and ways forward[J]. Geoderma Regional, 2020, 21: e00265.

[18] Li H, Liu X, Hu B, et al. Field - scale characterization of spatio - temporal variability of soil salinity in three dimensions[J]. Remote Sensing, 2020, 12(24): 4043.

[19] Yan F, Shang G W, Zhang J, et al. Depth - to - bedrock map of China at a spatial resolution of 100 meters [J]. Scientific Data, 2020, 7(1): 1 - 13.

[20] Fu T, Zhao R, Hu B, et al. Novel framework for modelling the cadmium balance and accumulation in farmland soil in Zhejiang Province, east China: Sensitivity analysis, parameter optimisation, and forecast for 2050 [J]. Journal of Cleaner Production, 2021, 279: 123674.

[21] 何林倩,刘倩,王德彩,等. 基于数字土壤制图技术的土壤有

- 机碳储量估算[J]. 应用生态学报, 2021, 32(2): 591–600.
- He L Q, Liu Q, Wang D C, et al. Estimation of soil organic carbon storage based on digital soil mapping technique[J]. Chinese Journal of Applied Ecology, 2021, 32(2): 591–600.
- [22] 张甘霖, 朱阿兴, 史舟, 等. 土壤地理学的进展与展望[J]. 地理科学进展, 2018, 37(1): 57–65.
- Zhang G L, Zhu A X, Shi Z, et al. Progress and future prospect of soil geography[J]. Progress in Geography, 2018, 37(1): 57–65.
- [23] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5–32.
- [24] 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.
- Ren L, Yang L A, Wang H, et al. Spatial prediction of soil organic matter in apple region based on random forest[J]. Journal of Arid Land Resources and Environment, 2018, 32(8): 141–146.
- [25] O'Brien R, Ishwaran H. A random forests quantile classifier for class imbalanced data[J]. Pattern Recognition, 2019, 90: 232–249.
- [26] Pouladi N, Möller A B, Tabatabai S, et al. Mapping soil organic matter contents at field level with Cubist, random forest and Kriging[J]. Geoderma, 2019, 342: 85–92.
- [27] Vaysses K, Lagacherie P. Using quantile regression forest to estimate uncertainty of digital soil mapping products[J]. Geoderma, 2017, 291: 55–64.
- [28] 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: 49–59.
- [29] da Silva Chagas C, de Carvalho W Jr, Bhering S B, et al. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions[J]. Catena, 2016, 139: 232–240.
- [30] Camera C, Zomeni Z, Noller J S, et al. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization[J]. Geoderma, 2017, 285: 35–49.
- [31] Amirian-Chakan A, Minasny B, Taghizadeh-Mehrjardi R, et al. Some practical aspects of predicting texture data in digital soil mapping[J]. Soil and Tillage Research, 2019, 194: 104289.
- [32] 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.
- [33] 卢宏亮, 赵明松, 刘斌寅, 等. 基于随机森林模型的安徽省土壤属性空间分布预测[J]. 土壤, 2019, 51(3): 602–608.
- Lu H L, Zhao M S, Liu B Y, et al. Spatial prediction of soil properties based on random forest model in Anhui Province[J]. Soils, 2019, 51(3): 602–608.
- [34] 王世航, 卢宏亮, 赵明松, 等. 基于不同特征挖掘方法结合广义提升回归模型估测安徽省土壤 pH[J]. 应用生态学报, 2020, 31(10): 3509–3517.
- Wang S H, Lu H L, Zhao M S, et al. Assessing soil pH in Anhui Province based on different features mining methods combined with generalized boosted regression models[J]. Chinese Journal of Applied Ecology, 2020, 31(10): 3509–3517.
- [35] 周宏冀. 江西省耕地土壤 pH 空间变异与管理分区研究[D]. 南昌: 江西财经大学, 2019.
- Zhou H J. Study on spatial variation and management zoning of soil pH value of cultivated land in Jiangxi Province[D]. Nanchang: Jiangxi University of Finance and Economics, 2019.
- [36] 齐雁冰, 王茵茵, 陈洋, 等. 基于遥感与随机森林算法的陕西省土壤有机质空间预测[J]. 自然资源学报, 2017, 32(6): 1074–1086.
- Qi Y B, Wang Y Y, Chen Y, et al. Soil organic matter prediction based on remote sensing data and random forest model in Shaanxi Province[J]. Journal of Natural Resources, 2017, 32(6): 1074–1086.
- [37] 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.
- [38] 王洪, 曹婧, 毋俊华, 等. 近 40 年来陕西省耕层土壤 pH 的时空变化特征[J]. 中国生态农业学报(中英文), 2021, 29(6): 1117–1126.
- Wang H, Cao J, Wu J H, et al. Spatial and temporal variability in soil pH of Shaanxi Province over the last 40 years[J]. Chinese Journal of Eco-Agriculture, 2021, 29(6): 1117–1126.
- [39] 戴万宏, 黄耀, 武丽, 等. 中国地带性土壤有机质含量与酸碱度的关系[J]. 土壤学报, 2009, 46(5): 851–860.
- Dai W H, Huang Y, Wu L, et al. Relationships between soil organic matter content(SOM) and pH in topsoil of zonal soils in China[J]. Acta Pedologica Sinica, 2009, 46(5): 851–860.
- [40] 许亚琪. 土壤阳离子交换量的分析结果研究[J]. 干旱环境监测, 2018, 32(1): 19–23.
- Xu Y Q. Study on the analytical results of soil cationic exchange capacity[J]. Arid Environmental Monitoring, 2018, 32(1): 19–23.
- [41] 康婷, 周春火, 魏宗强, 等. 江西省土壤阳离子交换量区域分布特征及其影响因素[J]. 中国农学通报, 2021, 37(14): 66–71.
- Kang T, Zhou C H, Wei Z Q, et al. Regional distribution characteristics and influencing factors of soil cation exchange capacity in Jiangxi[J]. Chinese Agricultural Science Bulletin, 2021, 37(14): 66–71.
- [42] 王晖, 邢小军, 许自成. 攀西烟区紫色土 pH 值与土壤养分的相关分析[J]. 中国土壤与肥料, 2007(6): 19–22.
- Wang H, Xing X J, Xu Z C. Relationship between pH value and available nutrients of purple soil in panxi tobacco-growing areas[J]. Soil and Fertilizer Sciences in China, 2007(6): 19–22.
- [43] 朱丹妹, 刘岩, 张丽, 等. 不同类型土壤淹水对 pH、Eh、Fe 及有效态 Cd 含量的影响[J]. 农业环境科学学报, 2017, 36(8): 1508–1517.
- Zhu D M, Liu Y, Zhang L, et al. Effects of pH, Eh, Fe, and flooded time on available-Cd content after flooding of different kinds of soil[J]. Journal of Agro-Environment Science, 2017, 36(8): 1508–1517.
- [44] Ding C, Du S, Ma Y, et al. Changes in the pH of paddy soils after

flooding and drainage: Modeling and validation [J]. Geoderma, 2019,337:511 – 513.

Spatial distribution prediction of soil pH in arable land of Jiangxi Province based on multi – source environmental variables and the random forest model

ZHONG Xiaoyong^{1,2}, LI Hongyi^{3,4}, GUO Dongyan², XIE Modian^{3,4}, ZHAO Wanru^{3,4}, HU Bifeng^{3,4}
(1. School of Finance and Public Administration, Jiangxi University of Finance and Economics, Nanchang 330013, China; 2. Chinese Academy of Natural Resources Economics, Beijing 101149, China; 3. School of Tourism and Urban Management, Jiangxi University of Finance and Economics, Nanchang 330013, China; 4. Key Laboratory of Data Science in Finance and Economics, Jiangxi University of Finance and Economics, Nanchang 330013, China)

Abstract: This study aims to compare the accuracy of random forest (RF) and ordinary kriging (OK) model for predicting spatial distribution of soil pH in arable land of Jiangxi Province using different covariates combination, and assess the feasibility and potential of RF method for improving the prediction accuracy of soil pH value. The RF algorithm is used to predict the pH value of cultivated soil in Jiangxi Province based on environmental covariate such as climate, topography and vegetation, combined with soil properties and cultivated land use conditions, identify the main influencing factors. The results produced by the RF was compared with the classical OK interpolation model. Our results showed that the accuracy of RF – A model with soil properties and cultivated land use conditions as environmental variables is better than that of RF – B model which only including terrain, climate and vegetation attributes as environmental variables. Climatic condition is the dominate factor which control the spatial variation of soil pH. the topographic factors and anthropogenic factors also have essential effect on spatial variability of soil pH. Thus, this study proved RF method has theoretical and practical significance for improving the accuracy of soil pH prediction at large – scale.

Keywords: arable land; soil pH; random forest; Jiangxi Province; influencing factor

(责任编辑: 陈 理)