DOI: 10.20103/j.stxb.202301130090

凌晓丹,王罗其,赵科理,傅伟军,叶正钱,丁立忠.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.生态学报,2024,44(2): 662-675.

Ling X D, Wang L Q, Zhao K L, Fu W J, Ye Z Q, Ding L Z.Spatial distribution characteristics of soil available nutrients in hickory plantation based on random forest method. Acta Ecologica Sinica, 2024, 44(2):662-675.

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

凌晓丹^{1,2},王罗其^{1,2},赵科理^{1,2,*},傅伟军^{1,3},叶正钱^{1,2},丁立忠⁴

1浙江农林大学环境与资源学院,杭州 311300

2浙江农林大学,浙江省土壤污染生物修复重点实验室,杭州 311300

3 浙江农林大学,浙江省森林生态系统碳循环与固碳减排重点实验室,杭州 311300

4 临安区农林技术推广中心,杭州 311300

摘要:土壤氮磷钾是土壤肥力管理的重要元素,是植物生长的必要养分元素。对土壤氮磷钾的空间分布进行特征解译,可为精 准管理临安山核桃产区林地土壤肥力,促进山核桃林产业可持续发展提供理论依据。研究以临安山核桃主产区为研究区域,利 用随机森林(RF)、普通克里格(OK)和 Shapley 加性解释(SHAP)方法,结合地形因子、气候因子、土壤因子、遥感因子等环境变 量,对山核桃林地土壤碱解氮(AN)、有效磷(AP)、速效钾(AK)的空间分布特征进行分析。研究结果表明:相比于 OK 模型,基 于环境协变量所构建的 RF 模型对 AN、AP 和 AK 含量空间分布预测表现最佳,*R*²分别为0.68、0.60 和 0.64,均方根误差(RMSE) 分别为 20.005、10.287 和 22.426,平均绝对误差(MAE)分别为 15.425、7.709 和 21.628。RF 模型 SHAP 分析显示,AN 和 AK 含量 分布主要受土壤有机质(SOM)的影响,并且 SOM 与 AN 和 AK 存在正相关性;AP 主要受 pH 的影响,其次为色调指数,AP 与 pH 和色调指数均具有负相关性;AK 和 AP 同时受到海拔和坡向的影响。两种模型预测的氮磷钾空间分布趋势总体相似,不同速 效养分存在明显的空间异质性。碱解氮高值区域主要分布于研究区东部;有效磷高值区域主要分布于研究区西部,但分散度 高;速效钾高值区域则主要分布于研究区中部。总体而言,基于随机森林模型可以高精度模拟山核桃林地土壤氮磷钾含量空间 分布特征,并依据主要环境协变量对土壤氮磷钾的影响关系,提出相应改良措施。在有效磷含量低值区域可以施用石灰来缓解 土壤酸化,同时补追磷肥;碱解氮含量高值区域可以合理碱少氮肥施用;速效钾含量低值区域合理施加钾肥;对于海拔较高及迎 风坡多降雨的区域,可以构建林下高效水土保持植被,碱轻水土流失;在林地施用有机肥料,改善土壤理化性质,增加土壤养分 含量。

关键词:山核桃林地;随机森林;Shapley 加性解释;速效养分

Spatial distribution characteristics of soil available nutrients in hickory plantation based on random forest method

LING Xiaodan^{1,2}, WANG Luoqi^{1,2}, ZHAO Keli^{1,2,*}, FU Weijun^{1,3}, YE Zhengqian^{1,2}, DING Lizhong⁴

1 College of Environmental and Resource Sciences, Zhejiang Agriculture and Forestry University, Hangzhou 311300, China

2 Key Laboratory of Soil Contamination Bioremediation of Zhejiang Province, Zhejiang Agriculture and Forestry University, Hangzhou 311300, China

3 Key Laboratory of Carbon Cycling in Forest Ecosystems and Carbon Sequestration of Zhejiang Province, Zhejiang Agriculture and Forestry University, Hangzhou 311300, China

4 Agricultural and Forestry Technology Extension Center of Lin'an, Hangzhou 311300, China

* 通讯作者 Corresponding author.E-mail: kelizhao@zafu.edu.cn

基金项目:浙江省基础公益研究计划项目(LY20C160004);杭州市临安区《"临安山核桃"亮牌战略三年攻坚行动方案》项目 收稿日期:2023-01-13; 网络出版日期:2023-10-17

Abstract: Soil nitrogen, phosphorus, and potassium are important elements for soil fertility management as well as vital nutrient elements for plant development. Understanding the geographical distribution features of soil nitrogen, phosphorus, and potassium can offer a theoretical foundation for properly controlling the soil fertility of forest land in the Lin'an hickory plantation and boosting the hictory forest industry's sustainable growth. In this study, the main production area of hickory plantation Lin'an was taken as the study area, and the random forest (RF), ordinary kriging (OK), and shapley additive interpretation (SHAP) methods were used to analyze the spatial distribution characteristics of soil alkali-hydrolyzable nitrogen (AN), available phosphorus (AP), and available potassium (AK) in hickory forest land in combination with environmental variables such as terrain factors, climate factors, soil factors, and remote sensing factors. According to the findings, the environmental covariates-based RF model outperformed the OK model in terms of accurately predicting the geographic distribution of AN, AP, and AK content. And the R^2 of AN, AP, and AK in the RF model were 0.68, 0.60, and 0.64, respectively, as well as the root mean square error (RMSE) of 20.005, 10.287, and 22.426, and the mean absolute error (MAE) of 15.425, 7.709, and 21.628. SHAP analysis of the RF model showed that the content distribution of AN and AK was mainly affected by soil organic matter (SOM), while SOM was positively correlated with AN and AK. AP was mainly affected by pH, followed by hue index. AP was negatively correlated with pH and hue index, and AK and AP were affected by elevation and aspect at the same time. The spatial distribution trends of AN, AK, and AP predicted by the two models were generally similar, but the spatial patterns of different available nutrients were different. The high value of AN was mainly distributed in the eastern part of the study area. The region with high AP value was mainly distributed in the west of the study area, but the dispersion was high. The high value of AK was mainly distributed in the middle of the study area. In general, based on the RF model, the spatial distribution characteristics of AN, AP and AK content in hickory forest can be simulated with high accuracy, and the corresponding management scheme can be put forward according to the relationship between AN, AP, AK, and main environmental covariates. It is proposed that lime can be applied to the areas with low AP content to alleviate soil acidification, and phosphate fertilizer can be supplemented at the same time, and nitrogen fertilizer application can be reasonably reduced in the areas with high AN content, potassium fertilizer should be applied reasonably in areas with low AK content. For the areas with high altitude and more rainfall on the windward slope, efficient soil and water conservation vegetation can be constructed under the forest to reduce soil erosion, and organic fertilizer can be applied to the forest land to improve soil physical and chemical properties and increase soil nutrient content.

Key Words: Chinese hickory plantation; random forest model; Shapley additive explanations; available nutrient

土壤氮、磷和钾不仅是土壤肥力的重要指标,也是植物生长期间所必备的养分元素^[1-2]。因此,农业土壤 中氮、磷和钾含量状况以及含量调控一直受到农业管理部门的重视^[3]。基于精准化农业管理需求,了解并掌 握土壤氮、磷和钾空间分布状况及其驱动力显得尤为重要。

土壤碱解氮(AN)、速效磷(AP)和速效钾(AK)指的是土壤中可直接被植物吸收利用的速效养分;由于受到人为活动、地形因素、土壤理化性质等因素的影响,AN、AP和AK往往具有较强的空间变异性^[4]。而基于实地调查获得的AN、AP和AK数据,评估仅限于特定范围,并且无法将土壤与环境变量之间的非线性关系联系起来,不能有效精准地反映整个区域的土壤状况^[5]。目前,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的数字土壤制图方法已被国内外学者使用,来实现区域内土壤属性的精准预测^[6-7]。在数字土壤制图中,多元线性回归、普通克里格法、地理加权回归、神经网络、偏最小二乘法和支持向量机等方法被广泛用于AN、AP和AK等土壤属性的预测^[8-9]。而作为机器学习与数据挖掘典型方法的随机森林,以其防止过拟合,计算量小但预测精度高的优点被众多学者使用。Wang等^[10]对天津东丽农业生产区的土壤重金属与潜在指标之间的复杂关系,并且有更好地预测结果。Fathizad等^[11]使用了随机森林对1986—2030年伊朗中部沙漠土

壤质量进行时空变化分析,为当地土壤退化状况提供了科学依据。

然而,尽管 RF 模型对土壤属性进行了准确预测,但建模结果很难解释。因此,长期以来的观点认为机器 学习是黑盒模型,无法对底层物理过程进行机械洞察^[12]。迄今为止,先前研究中进行的模型解释仅限于特征 重要性或者部分相关性分析,这很大程度上无法深入了解复杂模型的机制,甚至可能导致有偏差的解 释^[13-14]。为对模型预测进行可解释性分析,可以将 Shapley 加性解释(SHAP)方法与随机森林结合,从多角 度提供影响因素对模型结果的可视化解释。SHAP 是一种加性特征归因方法,不仅可以解决多重共线性问 题,还考虑了变量之间可能的协同效应^[15]。这些优点使其成为一个强大的模型解释器,并已在交通、基因工 程、环境等领域研究中使用^[16-18]。而这种模型构建搭载解释方法的应用在国内土壤属性预测中较少使用。

山核桃(*Carya cathayensis* Sarg.)作为中国经济林作物的典型代表之一,其林地土壤肥力状况及养分供给能力备受关注^[19]。AN、AP和AK作为土壤肥力的重要指标,能直观反映林地土壤状况^[20],其含量高低影响树木的生长发育及果实产量。临安拥有中国 50%以上的山核桃种植面积,主要位于湍口镇、清凉峰镇、河桥镇、昌化镇、龙岗镇、岛石镇、太阳镇等 7 个乡镇,也是该地区林农主要经济收入来源^[21]。目前,由于临安山核桃林地的快速发展和过度经营,林地正面临着土壤退化、病虫害频发、生态环境破坏等问题。这些问题对山核桃本地的快速发展和过度经营,林地正面临着土壤退化、病虫害频发、生态环境破坏等问题。这些问题对山核桃产业的健康和可持续发展构成了严重威胁^[22]。为此,本研究基于遥感因子、地形因子、土壤因子及气候因子组成的环境协变量构建随机森林预测模型,并使用 SHAP 解释方法,对山核桃林地产区 AN、AP和AK的空间分布影响因素进行可视化分析,为临安山核桃产区林地土壤肥力评估和精准施肥管理等农业措施提供科学依据。

1 材料与方法

1.1 研究区概况

本研究以山核桃主产区——临安区为研究区域,其位于浙江省西北部天目山区,为杭州市所辖区,地理坐标介于118°51′—119°21′E和29°56′—30°23′N之间(图1),海拔19m至1768m之间,平均温度为16.4℃,年平均降水量为1613.9mm^[23]。根据世界粮农组织(FAO)分类系统,研究区土壤被划分为粘土、壤土、砂粘壤土、粉砂壤土、沙土和粘壤土6个土壤质地单元,土壤类型主要包括潮土、潜育土、硬土、人为土、寒武土、淋溶土、黑土、瘦土、表土、盐土和高铁土11种^[24]。

1.2 土壤样品采集及测定

本研究土壤样品主要来源于临安区山核桃主产区域的7个乡镇。在主产区域内,按照每1km²布设一个 样点,结合山核桃分布情况,实际布设188个样点,在山核桃林地施肥前(3、4月份),使用差分全球定位系统 (GPS)野外采样表层土,均匀混合成1kg的混合样。土样在经过风干、研磨、过筛后测定土壤AN、AP、AK、pH 和土壤有机质(SOM)元素的含量。其中AN含量采用碱解扩散法测定;AP含量采用盐酸-氟化铵浸提-钼锑 抗比色法测定;AK含量采用醋酸铵浸提-火焰光度计测定;pH采用电极法测定;SOM采用重铬酸钾氧化-外加 热法测定。

1.3 环境协变量数据

本研究基于已有的研究,选择了 21 种环境协变量对山核桃林地 AN、AP 和 AK 进行预测,包括地形因子、 气候因子、土壤因子和遥感因子。使用 ArcGIS 10.2 软件及 ENVI 5.3 软件平台将山核桃林地 2013 年环境协 变量的栅格数据进行处理。地形因子作为最广泛运用的土壤预测环境因子^[25],能反映出研究区域的地形地 貌状况。通过数字高程模型(DEM)获得空间分布率为 30 m 的数据,并在 ArcGIS 10.2 软件中通过 Spatial Analyst 模块计算得到坡度、坡向。遥感信息来自 Landset 8,空间分辨率为 30 m,选取 7 月到 10 月林地生长茂 盛期且晴天云量少于 10%的卫星遥感数据,在 ENVI 5.3 软件平台进行大气校正、镶嵌和裁剪等前期处理,根 据波段计算公式提取波段信息,反映植被生长状况及覆盖度^[26]。土壤理化因子选择了 pH、有机质和土壤质 地状况。pH 和有机质使用样点实测值,对其进行空间插值以及栅格数据重采样,以获得研究区域内 pH 和有



Fig.1 The spatial distribution of soil samples in the study area

机质 30 m 分辨率数据。土壤质地类型数据信息来自世界土壤数据库(HWSD),原始空间分辨率为 1 km,将 数据库中浙江省的数据导出至 ArcGIS 10.2 软件后,使用插值分析把栅格数据重采样为 30 m 分辨率。选择 WGS_1984 地理坐标系定义投影,通过栅格属性及提取分析获取每个样点的土壤质地类型数据。年均温及年 降水量信息来自英国国家大气科学中心(CRU TS),数据为覆盖陆地表面的 0.5°分辨率的月度数据,使用 ArcGIS 10.2 通过插值分析中的反距离权重法将栅格数据设置为 30 m 分辨率,并进行裁剪、投影及插值处理, 并对获得的波段进行年均温及年降水量计算。总结见表 1。

1.4 模型构建及精度判断

1.4.1 随机森林构建

随机森林(RF)是基于树的算法的一种非常有代表性的袋装集成算法^[35]。其优势在于灵活性,能够处理 连续、离散、有序和定性预测的组合;此外,它们不需要数据正态性假设^[36-37]。本研究以 7:3 的比例随机划分 训练集及测试集,使用 Python 3.9 版本,加载 Scikit-Learn 库进行建模,并且通过交叉性验证进行调参,模型运 行重复迭代 100 次,以最优值作为模型的最终结果。

1.4.2 普通克里格插值

普通克里格(OK)基于 AN、AP 和 AK 的空间自相关,进行内插或外插以达到预测的目的,其实质是对 AN、AP 和 AK 实测值进行线性无偏差最优估测^[38]。OK 基于 AN、AK 和 AP 实测样点在 ArcGIS 10.2 中通过 Geostatistical Analysis 模块进行。

1.5 模型精度判断

本研究选用平均绝对误差(MAE)、均方根误差(RMSE)和绝对系数(R²)三个模型指标评价指标来评价 模型预测精度。以每次随机挑选的 30%数据作为验证集,进行 100 次模型迭代后,对预测 AN、AP 和 AK 的 OK 模型和 RF 模型实行精度检验。计算公式如下:

Table 1 Brief description of Environment covariates							
 类型 Type	协变量 Covariates	定义 Definition	来源 Beference				
遥感因子 Remote sensing factor	植被归一化指数(NDVI) ^[11, 27] 增强型植被指数(EVI) ^[11, 27] 绿色归一化差值植被指数 (GNDVI) ^[11, 27] 绿土指数(BI) ^[28-29] 绿色叶绿素指数(CI) ^[28-29] 每色叶绿素指数(CI) ^[28-29] 鱼调指数(HI) ^[28-29] 红色指数(RI) ^[28-29] 红色指数(SI) ^[28-29]	$(\text{Red-NIR})/(\text{Red+NIR})$ $2.5(\text{NIR-Red})/(\text{NIR+6}\times\text{Rde}-7.5\times\text{Blue})+1$ $(\text{NIR-Green})/(\text{NIR+Green})$ $^{0.5}\sqrt{((\text{Red}^2+\text{Green}^2+\text{Blue}^2)/3)}$ $(\text{Red-Green})/((\text{Red}+\text{Green})$ $(2\times\text{Red}-\text{Green}-\text{Blue})/((\text{Green}-\text{Blue}))$ $\text{Red}^2/((\text{Blue}\times\text{Green}^3))$ $((\text{Red}-\text{Blue})/((\text{Red}+\text{Blue}))$	地理数据空间云(http://www. gscloud.cn) Landsat 8 OLI-TIRS 数字 产品 30m 分辨率				
地形因子 Topographical factor	坡度 (SLOPE) ^[30-31] 海拔 (ELEVATION) ^[30-31] 坡向 (ASPECT) ^[30-31]	两点的高程差与其水平距离的百 分比 以海平面为基准面计算高度 以度为单位按逆时针方向进行测量	ASTER GDEM V3 30 m 分辨率 (http://www.gscloud.cn)				
土壤因子 Soil factor	 土壤类别(SU_SYM90)^[24, 32] 土壤粘粒含量(CLAY)^[24, 32] 土壤含砂量(SAND)^[24, 32] 土壤容重(BULK DENSITY)^[24, 32] 土壤含泥量(SILT)^[24, 32] 砾石体积比(GRAVEL)^[24, 32] 	粮农组织-90 土壤分类 土层中粘粒含量占比 土层中含砂量占比 参考体积密度 土层中淤泥含量占比 土层中砾石体积的占比	世界土壤数据库 (https://www.fao.org/soils-portal/soil- survey/soil-maps-and-databases/ harmonized-world-soil-database- v12/en/)				
气候因子 Climatic factor	土壤 pH ^[33] 土壤有机质 (SOM) ^[33] 年均温 (MAT) ^[27, 34] 年降水量 (AAR) ^[27, 34]	土壤酸碱度 土壤中含碳有机化合物 研究区年均温 研究区年降水量	研究区样点实测值 英国国家大气科学中心(https:// crudata.uea.ac.uk/cru/data/hrg/)				

衣 I 环境协定里信息	表 1	环境协变量信息
-------------	-----	---------

 Table 1
 Brief description of Environment covariates

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |f_i - y_i|$$
 (1)

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

$$R^{2} = \frac{\sum_{i=1}^{N} (f_{i} - \bar{y}_{i})^{2}}{\sum_{i=1}^{N} (y_{i} - \bar{y}_{i})^{2}}$$
(3)

式中, N 是样本数, i 是每个数据样本, f_i 是模型对样本点 i 进行的预测值, y_i 是样本点 i 的实际值。精度运算通 过 Python 3.9 版本运行 Scikit-Learn 库中的 metrics 包进行, 同时加载 matplotlib 库, 绘制 AN、AP 和 AK 不同模 型下预测值和真实值的拟合图。

1.6 SHAP 模型解释法

SHAP 是由 Lundberg 和 Lee 提出的一种基于博弈论来描述机器学习模型性能的方法,它使用了夏普利值 (shapley 值)来估计每个特征贡献值是多少^[39]。根据博弈论,数据集中的每一个特征变量可看作成员人员, 用该数据集去训练模型得到预测的结果,可以视为所有成员合作完成一个项目的收益,而 shapley 值则是通过 考虑各个成员作出贡献,来公平的分配合作的收益。

假设使用 F 组(具有 n 个特征)来预测输出的 RF 模型。在 SHAP 中,每个特征对模型输出 f(f)的贡献是 基于其边际贡献进行分配。Shapley 值通过以下公式确定:

$$\phi_{i} = \sum_{S \subseteq F[i]} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_{S}(x_{S})]$$
(4)

式中, ϕ_i 是*i* 特征的 Shapley 值,*F* 是所有特征的集合,*S* 是移除*i* 特征后从*F* 生产的所有特征子集的集合。 |S|! (|F|-|S|-1)! |F|! 是指在特征置换和组合之后导出的*S* 的概率权重。 $f_{S\cup\{i\}}$ 和 f_s 分别表示*S* 特征子集 的集合。输入后模型的*i* 特征和预测值,并将其预测与当前输入的 $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_s(x_s)$ 进行比较,其中表示 集合 *S* 中输入特征的值。

在本研究中,进行了 SHAP 分析,以确定影响山核桃林地土壤碱解氮、有效磷、速效钾的主要环境特征,并 探讨速效养分和环境变量之间的相互作用。SHAP 平均值用于显示重要特征,SHAP 相关性图也用于评估选 定环境特征的影响。SHAP 分析通过在 Python 3.9 版本中加载 shap 库运行树解释器构建^[40],作图使用 shap 库中的 summary_plot 和 dependence_plot 包。

2 结果与分析

2.1 山核桃林地土壤养分含量描述性统计分析

临安山核桃林地土壤 AN、AP、AK、SOM 和 pH 均值分别为 155.53 mg/kg、14.06 mg/kg、85.66 mg/kg、31.61g/kg和5.27,标准差分别为 42.27 mg/kg、19.85 mg/kg、39.41 mg/kg、11.86 g/kg 和 0.41。除了 AP 的变异 系数高达 141%,属于高度变异外,AN、AK、SOM 和 pH 的变异系数是 27%、46%、34%和 12%,属于中等变异 (表 2)。

Table 2 Descriptive statistics analysis result of soil sample in the study area						
分析指标	平均值	最小值	最大值	标准差	变异系数	
Analysis indicators	Mean	Min	Max	SD	CV/%	
碱解氮 AN/(mg/kg)	155.53	56.61	268.97	42.27	27	
有效磷 AP/(mg/kg)	14.06	1.37	143.28	19.85	141	
速效钾 AK/(mg/kg)	85.66	18.90	255.70	39.41	46	
土壤有机质 SOM/(g/kg)	31.61	9.70	67.70	11.86	34	
pH	5.27	4.19	7.52	0.41	12	

表 2 研究区土壤养分含量描述性统计

Mean:平均值 mean value; Min:最小值 minimum value; Max:最大值 maximum value; SD:标准差 standard deviation; CV:变异系数 coefficient of variation; AN:碱解氮 alkali-hydrolyzed nitrogen; AP:有效磷 available phosphorus; AK:速效钾 available potassium; SOM:土壤有机质 soil organic matter

根据表 3 浙江森林土壤的分级标准(共分 4 级,一级最差,四级最优),临安山核桃林地土壤 AN 以三级为 主,三级占比 76.60%,林地土壤 AN 含量呈现出中高水平;AP 一级占比最多,一级和二级共占 63.83%,林地土 壤 AP 含量呈现出低水平;AK 二级和三级共占 64.89%,林地土壤 AK 呈现出中水平,而一级占据了约 20%,表 明当地有部分区域 AK 含量为低水平;SOM 以四级为主,占比 53.72%,林地 SOM 含量呈现出高水平;pH 一级 和二级共占比 87.16%,林地土壤 pH 呈现出低水平,表明当地土壤酸化严重。

Table 3	Classification of a	vailable nu	trients, pH and	SOM bas	ed on Zhejiang	forest soil	classification s	tandard	
<u>/////////////////////////////////////</u>	碱解氮		有效磷		速效钾		土壤有机质		
守奴 Louolo	AN⁄	%	AP/	%	AK/	%	SOM/	%	$_{\rm pH}$
Levels	(mg/kg)		(mg/kg)		(mg/kg)		(g/kg)		
—级 First grade	<80	2.13	<5	40.43	<50	20.21	<10	0.53	<5
二级 Second grade	80—100	6.38	5—10	23.40	50—80	31.38	10—20	11.70	5—6
三级 Third grade	100—200	76.60	10—20	14.36	80—120	33.51	20—30	34.04	6—7
四级 Fourth grade	>200	14.89	>20	21.81	>120	14.89	>30	53.72	>7

表 3 浙江森林土壤分级标准及土壤养分分级结果

2期

2.2 环境协变量的筛选

为避免多重共线性及模型过拟合对预测精度的影响,需要对环境协变量进行筛选,去除冗余和不相关的 协变量。首先使用相关性分析(图2),选择相关系数小于0.7的变量作为预测因子^[41],当两个协变量大于0.7 时,保留与 AN、AP 和 AK 相关性较高的变量。相关分析筛选后,AN 初步保留了除绿色叶绿素指数(CI)和土 壤粘粒含量(CLAY)的全部变量;AP 和 AK 则保留了除 CI 和土壤容重(BULK DENSITY)的全部变量。为进 一步明确环境协变量之间的相关性和依赖性以及对预测 AN、AP 和 AK 含量的影响程度,使用 SHAP 分析进 行特征重要性排序,根据由低到高的排序结果,排除约 30% SHAP 值较低的特征(AN 和 AK 筛除 SHAP 值<1 的协变量,AP 筛除 SHAP 值<0.5 的协变量,以全部协变量为标准)^[42]。经过筛选分析,AN、AP 和 AK 预测模 型的山核桃林地环境协变量由最初的 21 个变为 14 个,AN、AP 和 AK 预测使用的环境协变量如表 4 所示。



2.3 不同模型预测精度及精度检验

OK 模型和 RF 模型对预测 AN、AP 和 AK 的精度评价参数见表 5。决定系数 R²数值表明, RF 模型预测的 山核桃林地土壤表层 AN、AP 和 AK 含量变异可解释性均在 0.60 及以上, 高于 OK 模型预测的可解释性。预 测 AN 的 RF 模型和 OK 模型平均绝对误差 MAE 值分别为 15.425 和 26.530, 两者之间相差 11.105, 表明 RF 模

型预测值误差远小于 OK 模型;预测 AP 和 AK 的 RF 模型和 OK 模型平均绝对误差 MAE 值则分别为 7.709、 7.653和 21.628、20.495,尽管 RF 模型下 AP 和 AK 的预测值误差高于 OK 模型的预测值误差,但两者数值相差 较少,对比性弱^[43]。均方根误差 RMSE 数值表明,RF 模型下的 AN、AP 和 AK 预测值与真实值的偏差均小于 OK 模型,分别为 20.005、10.287 和 22.426,说明 RF 模型相比于 OK 模型预测误差少。同时 AN、AP 和 AK 不同 模型下预测值与真实值均显著相关(*P*<0.01),随机森林模型预测值与真实值的 Pearson 相关系数分别为0.895、0.710和 0.921,OK 模型下则分别为 0.805、0.669 和 0.723,证明 RF 模型预测数据相比于 OK 模型更接近实际值。 因此,结合三个精度参数及相关系数,作为集成算法的随机森林在预测精度上优于传统空间预测方法。

Table 4 Statistical table of model prediction environment covariates					
类型	协变量	碱解氮 AN	有效磷 AP	速效钾 AK	
Types	Covariates	WANT SKI THI		MAN AR	
遥感因子	植被归一化指数 (NDVI)				
Remote sensing factors	增强型植被指数 (EVI)	\checkmark	\checkmark	\checkmark	
	绿色归一化差值植被指数 (GNDVI)	\checkmark	\checkmark	\checkmark	
	裸土指数 (BI)		\checkmark		
	绿色叶绿素指数 (CI)				
	色调指数 (HI)	\checkmark	\checkmark	\checkmark	
	红色指数 (RI)	\checkmark	\checkmark	\checkmark	
	饱和度指数 (SI)		\checkmark		
地形因子	坡度 (SLPOE)	\checkmark	\checkmark	\checkmark	
Topographical factors	海拔 (ELEVATION)	\checkmark	\checkmark	\checkmark	
	坡向 (ASPECT)	\checkmark	\checkmark	\checkmark	
土壤因子	土壤类别 (SU_SYM90)	\checkmark			
Soil factors	土壤粘粒含量 (CLAY)		\checkmark	\checkmark	
	土壤含砂量 (SAND)	\checkmark		\checkmark	
	土壤容重 (BULK DENSITY)				
	土壤含泥量 (SILT)		\checkmark	\checkmark	
	砾石体积比 (GRAVEL)	\checkmark			
	土壤 pH	\checkmark	\checkmark	\checkmark	
	土壤有机质 (SOM)	\checkmark	\checkmark	\checkmark	
气候因子	年均温 (MAT)			\checkmark	
Climatic factors	年降水量 (AAR)	\checkmark			

表 4	AN、AP、AK 模型预测所使用的环境协变量统计表

√表示该变量被选中作为预测变量

由图 3 可直观看出, RF 模型预测输出的 AN、AP 和 AK 含量与验证集真实值的相关度高于 OK 模型, 但两种模型都存在 AN、AP 和 AK 含量在高值的区域被低估、在低的区域被高估的现象。

表 5 AN、AP 和 AK 不同预测模型的精度参数							
Table 5 Accuracy parameters of different prediction models of AN, AP and AK							
模型对象 Model objects	模型 Models	决定系数 R^2	平均绝对误差 MAE	均方根误差 RMSE			
碱解氮 AN	随机森林 RF	0.68	15.425	20.005			
	普通克里格 OK	0.42	26.530	32.621			
有效磷 AP	随机森林 RF	0.60	7.709	10.287			
	普通克里格 OK	0.26	7.653	11.998			
速效钾 AK	随机森林 RF	0.64	21.628	22.426			
	普通克里格 OK	0.35	20.495	24.581			

R²:决定系数 coefficient of determination; MAE:平均绝对误差 mean absolute error; RMSE:均方根误差 root mean square error; RF:随机森林 random forest; OK:普通克里格 ordinary kriging

真实值





图 3 AN、AP 和 AK 不同模型预测值和真实值的散点图 Fig. 3 Fitted plots of true and predicted values for different models of AN, AP, and AK AN:碱解氮; AP:有效磷; AK:速效钾; RF:随机森林; OK:普通克里格

环境协变量重要性评估 2.4

选取 RF 最优模型,绘制了各环境协变量对 AN、AK 和 AP 含量的重要性排序柱图,并使用 SHAP 平均值 作为量化值,表达了环境协变量的模型贡献度,即对 AN、AK 和 AP 含量的影响程度。根据图 4 可知, AN 含量 和 AK 含量主要受 SOM 的影响; AP 含量则主要受 pH 和色调指数的影响。图 5 显示了环境变量按照重要性 排序的 SHAP 值, SHAP 值的正负则代表协变量对预测目标含量产生了积极的或消极的影响;每个点代表一 个样本,点的颜色代表一个协变量的值,红色表示高值,蓝色表示低值。因此,丰富的 SOM 能促进 AN 含量和 AK含量。对于AP,碱性土更利于其积累,同时色调指数越低,AP含量越高。

山核桃林地土壤速效养分空间分布特征 2.5

采用 RF 模型对研究区 AN、AP 和 AK 进行空间分布预测,同时使用普通克里格进行插值。两种空间预测

250 225



Fig.4 Ranking of the importance of environmental covariates on the content of AN, AP, and AK SHAP: Shapley 加性解释

方法下的空间分布图如图 6 所示。RF 模型和 OK 模型两种方法预测的 AN、AP 和 AK 空间分布总体一样,AN 高值区域多分布于研究区东部的太阳镇和湍口镇以及中部的昌化镇和西北部的岛石镇,AP 高值区域则是呈 现出零散的分布状况,而 AK 高值区域多分布于清凉峰镇、昌化镇和河桥镇。然而 OK 模型模拟的空间分布 结果在局部高值以及高值向低值过渡区域表现出明显的平滑效应,不能较好地体现出速效养分的空间异质 性。RF 模能体现出受到环境协变量影响后出现的数据波动性,模拟出符合实际的土壤 AN、AP 和 AK 含量空 间分布趋势,可以充分表达出速效养分的异质性。

3 讨论

3.1 山核桃林地土壤速效养分含量状况

根据对山核桃林地土壤速效养分的调查及依据浙江森林土壤分级标准,目前临安山核桃林地土壤 AN 含量充足,AK 含量有部分地区存在不足,而 AP 含量存在严重不足的现象。先前邬奇峰等^[44]对临安山核桃林 地氮磷钾现状调查也证实了山核桃林地缺磷缺钾较为明显。AN 含量丰富可能与当地林农为追求山核桃高 产量,减少"大小年"现象的发生而过度施用氮肥有关^[45]。此外,不合理的施肥带来了土壤酸化、养分平衡失 调的问题,致使山核桃林地出现 AN 含量高,AP 和 AK 含量低的现象。

3.2 环境协变量对土壤速效养分的影响

在所有的环境协变量中, AN 与 AK 主要受 SOM 的影响。这与有机质和 AN、AK 的耦合性有关, SOM 是土 壤氮素的主要来源, 丰富的 SOM 可以提高碱解氮含量, 同时有机质可以改善土壤供钾状况^[46]。但通过图 4 和图 5 可知, SOM 对 AN 的影响要强于 AK, 并且 SOM 主导着 AN 含量变化, 这与 SOM 和 AN 具有极强的正相







关性有关^[47]。土壤中的氮素绝大多数是贮藏在土壤有机质中的有机态化合物,土壤全氮量的消长取决于有 机质含量的变化,而其它因素如 pH、海拔、土壤含砂量等对氮素供应的影响,是通过影响土壤有机质的积累和 分解而起作用,同时全氮量与碱解氮含量呈极紧密的正相关^[48]。因此,相比于其他环境协变量,SOM 是 AN 的主导变量。而 AK 除了 SOM,还相应受到海拔、坡向等其他环境协变量的影响。根据图 5,AK 与海拔和坡 向具有负相关性,这可能与山地海拔升高,土层厚度变薄,土壤养分少有关^[49-50],同时临安受山地气候影响, 多降雨,水力侵蚀严重,促使 AK 流失。因此,为减少水土及养分的流失,可以在林地间实行套作,增加灌木、 草本植物的种植,固定土壤^[51]。AP 含量则深受 pH 的影响,这结果与之前的研究一致^[52]。这是由于土壤的 酸化会使土壤磷固定为磷酸铁或磷酸铝,导致磷的有效性降低^[53]。根据本研究采用的遥感数据,发现色调指 数与 AP 呈现较强的负相关性,吻合叶片反射率与养分浓度之间具有相关性且 P 浓度与叶片反射率之间存在 负相关的现象^[54]。同时 AP 与 AK 一致,在一定程度上受到海拔、坡向等地形影响。

3.3 模型预测精度对比及影响预测精度因素

本研究中,根据精度参数以及预测值和真实值的相关系数,随机森林模型对临安山核桃林地主产区土壤 AN、AP和AK的空间预测要优于普通克里格插值。这是由于以变异函数为基础的OK,依靠采样点数据对空 间加以插值,没有考虑环境协变量,而RF能拟合AN、AP和AK与环境协变量之间的复杂关系^[55]。因此,基 于RF模型进行预测能有效提高山核桃林地土壤AN、AP和AK含量空间分布趋势的精确性。然而,在RF预 测模型中,AP的*R*²只有0.60,拟合解释度低于AN和AK的预测。这可能与环境协变量有关。尽管本文已对 环境变量进行筛选,大大降低数据冗余,但环境协变量本身存在精细化不足及协变量选取不周全的问题。本 文统一采用的是30m的分辨率栅格数据,这导致在处理DEM、遥感、气候等数据时会出现精度下降现象。同



图 6 山核桃林地土壤 AN、AP 和 AK 含量空间分布预测图

Fig.6 Prediction map of spatial distribution of AN AP and AK in hickory plantation produce area

时本文对于人为因素的考虑过少,虽然遥感数据所获得的植被因子能在一定程度上反映人为活动,但若添加 肥料施加量、山核桃种植制度等变量可能会在一定程度上提高模型预测的精度。

4 结论

本文利用随机森林、普通克里格和 Shapley 加性解释方法,结合地形因子、气候因子、土壤因子、遥感因子 等环境变量,对山核桃林地土壤碱解氮、有效磷和速效钾的空间分布特征进行分析,主要结论如下:

(1)山核桃林地土壤速效养分存在不平衡、不充分现象,林地缺磷缺钾较为明显。

(2) RF 模型预测山核桃林地土壤 AN、AP 和 AK 含量的变异可解释性均在 0.60 及以上,模型预测值与真 实值的偏差均小于 OK 模型,相关系数均高于 OK 模型。与 OK 模型相比, RF 模型在预测精度上和空间制度 的准确度上更具优势。

(3)不同速效养分所受到的主要环境协变量影响各不相同。AN和AK主要受到SOM的影响,并且SOM与AN和AK存在正相关性,AP主要受到pH的影响并具有负相关性。AP和AK在受到主要协变量影响的基础上还相应受到海拔和坡向的影响。

(4)不同速效养分具有不同空间分布格局。AN含量高值区域主要分布于研究区东部;AP含量高值区域 主要分布于研究区西部,但分散度高;AK含量高值区域则主要分布于研究区中部。

2期

基于以上研究结果,为促进山核桃林地的科学化、精准化管理,在综合考虑林地土壤酸碱度、有机质含量 以及地质条件等现有情况下,结合速效养分的分布状况和变化因素,制定相应的林地土壤养分管理方案。在 山核桃林地管理中,对 AP 含量低的研究区域可以施用石灰来缓解土壤酸化,同时施加磷肥;对 AN 高值主要 集中区域如昌化镇、太阳镇、湍口镇合理减少氮肥;对 AK 含量低的研究区域合理施加钾肥;针对海拔较高土 层较薄及迎风坡多降雨的区域,种植适宜的林下植被,减轻水土养分的流失;适当增加有机肥料的施用,改善 土壤理化性质。

参考文献(References):

- [1] 李云祥,胡福平,冯玉磊,卢军帅,王兴富,白斌,刘谦.甘肃中部灌区春小麦土壤氮磷钾养分丰缺指标及施肥指标研究.中国土壤与肥料,2021(5):34-41.
- [2] 马鹏,陈沛,李冰,王昌全,陈玉蓝,王勇.烟类生物质材料对土壤氮磷钾有效性的影响.西昌学院学报:自然科学版,2021,35(3): 17-21.
- [3] 中共中央、国务院.中共中央国务院关于抓好"三农"领域重点工作确保如期实现全面小康的意[EB/OL].(2020-02-05).http://www.gov. cn/zhengce/2020-02/05/content_5474884.html.
- [4] 赵影星, 宋彤, 陈源泉, 王彪, 刘晴, 隋鹏. 华北平原麦-玉年际轮作的土壤氮磷钾分布及淋洗研究. 中国农业大学学报, 2022, 27(2): 1-14.
- [5] 袁紫倩,叶正钱,李皓,董建华,胡俊靖,赵伟明.影响山核桃林地土壤生产性能的主要肥力因子及其临界区间.植物营养与肥料学报, 2020,26(1):163-171.
- [6] Pahlavan-Rad M R, Akbarimoghaddam A. Spatial variability of soil texture fractions and pH in a flood plain (case study from eastern Iran). CATENA, 2018, 160: 275-281.
- [7] Teng H F, Viscarra Rossel R A, Shi Z, Behrens T. Updating a national soil classification with spectroscopic predictions and digital soil mapping. CATENA, 2018, 164: 125-134.
- [8] 郝辰恺,孙孝林,王会利.广义线性地统计模型在典型亚热带丘陵区数字土壤制图中的应用.土壤学报,2022,60(4):1-16.
- [9] Peng J, Biswas A, Jiang Q S, Zhao R Y, Hu J, Hu B F, Shi Z. Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province, China. Geoderma, 2019, 337: 1309-1319.
- [10] Wang H Z, Yilihamu Q, Yuan M N, Bai H T, Xu H, Wu J. Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: a comparison of regression and random forest. Ecological Indicators, 2020, 119: 106801.
- [11] Fathizad H, Ali Hakimzadeh Ardakani M, Heung B, Sodaiezadeh H, Rahmani A, Fathabadi A, Scholten T, Taghizadeh-Mehrjardi R. Spatiotemporal dynamic of soil quality in the central Iranian Desert modeled with machine learning and digital soil assessment techniques. Ecological Indicators, 2020, 118: 106736.
- [12] Lipton Z C. The mythos of model interpretability. Communications of the ACM, 2018, 61(10): 36-43.
- [13] De Clercq D, Wen Z G, Fei F, Caicedo L, Yuan K, Shang R X. Interpretable machine learning for predicting biomethane production in industrialscale anaerobic co-digestion. The Science of the Total Environment, 2020, 712: 134574.
- [14] Mangalathu S, Hwang S H, Jeon J S. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Engineering Structures, 2020, 219: 110927.
- [15] Guo Y H, Ma W, Li J J, Liu W, Qi P Z, Ye Y Y, Guo B Y, Zhang J S, Qu C K. Effects of microplastics on growth, phenanthrene stress, and lipid accumulation in a diatom, *Phaeodactylum tricornutum*. Environmental Pollution, 2020, 257: 113628.
- [16] Dong S, Khattak A, Ullah I, Zhou J B, Hussain A. Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with SHAPley additive exPlanations. International Journal of Environmental Research and Public Health, 2022, 19(5): 2925.
- [17] Vega García M, Aznarte J L. Shapley additive explanations for NO₂ forecasting. Ecological Informatics, 2020, 56: 101039.
- [18] Zhong S, Zhang K, Wang D, Zhang H. Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds. Chemical Engineering Journal, 2021, 405:126627.
- [19] 王罗其,傅伟军,叶正钱,马闪闪,丁立忠,赵科理.基于地理信息系统和遥感技术的山核桃产区林地土壤质量退化评价.浙江大学学报:农业与生命科学版,2022,48(2):216-226.
- [20] 倪幸, 窦春英, 丁立忠, 石红静, 马闪闪, 赵伟明, 赵科理, 叶正钱. 有机物料对山核桃林地土壤的培肥改良效果. 植物营养与肥料学报, 2018, 24(5): 1266-1275.
- [21] 冯亚枝, 胡彦蓉, 刘洪久. 临安山核桃产量的 Lasso-灰色预测模型研究. 林业资源管理, 2021(1): 94-102.
- [22] 周军,吴丽娟,刘波.临安区"三举措"加快山核桃产业高质量发展.浙江林业,2020(9):10-11.
- [23] 胡颖槟,金锦,童志鹏,吴家森.山核桃人工林土壤肥力研究进展.浙江农林大学学报,2021,38(5):1066-1075.
- [24] Micheli E, Dent D. World reference base for soil resources: a framework for international classification, correlation, and communication. 2006 ed. Rome: Food and Agriculture Organization of the United Nations, 2006.
- [25] Moore I D, Gessler P E, Nielsen G A, Peterson G A. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal,

1993, 57(2): NP.

2期

[26] 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. 地理学报, 2019, 74(3): 534-543.

- [27] 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.土壤学报,2022,59(2):451-460.
- [28] Mashaba-Munghemezulu Z, Chirima G J, Munghemezulu C. Modeling the spatial distribution of soil nitrogen content at smallholder maize farms using machine learning regression and sentinel-2 data. Sustainability, 2021, 13(21): 11591.
- [29] 郑友苗,王云南,岳春雷,王珺,李贺鹏,杨乐.山核桃林下不同植物配置模式水土保持功能的比较研究.浙江林业科技,2014,34(4): 72-75.
- [30] 马重阳, 孙越琦, 巫振富, 张靖一, 牛银霞, 侯占领, 陈杰. 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. 土壤通报, 2021, 52(6): 1261-1272.
- [31] 黄兴召,黄坚钦,陈丁红,吕健全,吴家森.不同垂直地带山核桃林地土壤理化性质比较.浙江林业科技,2010,30(6):23-27.
- [32] 祝小祥,徐祖祥,徐进,徐友成.临安山核桃主产区土壤理化性状变化的研究.农学学报,2014,4(6):32-35,40.
- [33] 陈謇,丁安娜,施加春.浙江省温岭市土壤 pH 及氮磷钾养分时空动态变化特征.浙江大学学报:农业与生命科学版,2021,47(4): 517-526.
- [34] 颜勇, 王惠丽, 刘强, 郑城, 陈国伟. 山核桃林地水土流失控制措施配置模式研究. 水土保持应用技术, 2018(5): 35-37.
- [35] Krzywinski M, Altman N. Classification and regression trees. Nature Methods, 2017, 14(8): 757-758.
- [36] Franklin J. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 2005, 27(2): 83-85.
- [37] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5-32.
- [38] 杨谦, 王晓晴, 孙孝林, 王会利. 基于 REML 的普通克里格和回归克里格在土壤属性空间预测中的比较. 土壤通报, 2018, 49(2): 283-292.
- [39] Lundberg S, Lee S I. A unified approach to interpreting model predictions. 2017: arXiv: 1705.07874. https://arxiv.org/abs/1705.07874
- [40] Wang F E, Wang Y X, Zhang K, Hu M, Weng Q, Zhang H C. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environmental Research, 2021, 202: 111660.
- [41] Lundberg S M, Erion G, Chen H, DeGrave A, Prutkin J M, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S I. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020, 2(1): 56-67.
- [42] Zhou Y N, Wu W, Wang H, Zhang X, Yang C, Liu H B. Identification of soil texture classes under vegetation cover based on sentinel-2 data with SVM and SHAP techniques. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 3758-3770.
- [43] Chen K Y, Chen H X, Zhou C L, Huang Y C, Qi X Y, Shen R Q, Liu F R, Zuo M, Zou X Y, Wang J F, Zhang Y, Chen D, Chen X G, Deng Y F, Ren H Q. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Research, 2020, 171: 115454.
- [44] 邬奇峰, 阮弋飞, 章林英, 黄仨仨, 应学兵, 俞俊. 临安市农地土壤氮磷钾现状与历史演变及施肥对策. 中国农学通报, 2017, 33(23): 42-50.
- [45] 邬奇峰, 章秀梅, 阮弋飞, 黄仨仨, 童小虎, 俞俊, 秦华. 临安市山核桃林地土壤肥力特征及其施肥对策. 浙江农业科学, 2017, 58(7): 1132-1135.
- [46] 林万树. 古田县果园土壤氮磷钾状况及其与有机质和 pH 的相关性. 上海农业学报, 2015, 31(1): 44-48.
- [47] 周启运,郑重谊, 荊永锋, 刘勇军, 彭曙光, 陈焘, 刘智炫, 胡瑞文, 周清明, 黎娟. 湘南稻作烟区不同土层土壤有机质含量与氮磷钾关系研究. 作物杂志, 2021(5): 114-119.
- [48] 王青山,何利平. 土壤有机质与氮素供应的相关关系. 山西林业科技, 2003, 32(S1): 25-27.
- [49] 张春苗,张有珍,姚芳,张圆圆,窦春英,叶正钱,胡国良,黄坚钦.临安山核桃主产区土壤 pH 值和有效养分的时空变化.浙江农林大学学报,2011,28(6):845-849.
- [50] 黄安香,周显勇,杨守禄,姬宁,朱亚艳,许杰,张彦雄.坡度对油茶林土壤氮磷钾含量与剖面分布特征的影响.中国农学通报,2020, 36(34):25-31.
- [51] 孙梦娇. 山核桃林地土壤养分的影响因素及其空间格局研究[D]. 合肥: 安徽农业大学, 2022.
- [52] 问青青. 基于 GIS 的山核桃生长适宜性评价研究[D]. 杭州:浙江农林大学, 2011.
- [53] 石红静,马闪闪,赵科理,叶立前,李皓,沈颖,赵伟明,叶正钱.有机物料对酸化山核桃林地土壤的改良作用.浙江农林大学学报, 2017,34(4):670-678.
- [54] de Oliveira L F R, Santana R C, de Oliveira M L R. Nondestructive estimation of leaf nutrient concentrations in eucalyptus plantations. CERNE, 2019, 25(2): 184-194.
- [55] 史广,刘庚,赵龙,苏迎庆,毕如田.基于多源环境数据和随机森林模型的农田土壤砷空间分布模拟.环境科学学报,2020,40(8): 2993-3000.