

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

咸 阳^{1,2} 宋江辉^{1,2} 王金刚^{1,2} 李维弟^{1,2} 张文旭^{1,2} 王海江^{1,2}

(1. 石河子大学农学院, 石河子 832003; 2. 新疆生产建设兵团绿洲生态农业重点实验室, 石河子 832003)

摘要: 为了提高农田土壤养分含量空间插值精度, 准确掌握土壤养分的空间分布特征, 以新疆玛纳斯河流域绿洲为研究区域, 测定土壤有机质含量、全氮含量、有效磷含量、速效钾含量、pH值和盐分含量, 协同经度、纬度、地形、气象和植被指数因子作为环境变量, 经过皮尔逊相关系数(Person correlation coefficient, PCC)、方差膨胀系数(Variance inflation factor, VIF)和极端梯度提升(Extreme gradient boosting, XGBoost)算法进行变量筛选, 采用决策树(Decision tree, DT)、随机森林(Random forest, RF)、径向基函数神经网络(Radial basis function, RBF)和长短期记忆网络(Long short-term memory, LSTM)4种机器学习模型与普通克里格(Ordinary Kriging, OK)方法, 对研究区农田土壤有机质、全氮、有效磷和速效钾含量进行空间插值。结果表明: 研究区土壤有机质、全氮、有效磷、速效钾含量分别为0.226~32.275 g/kg、0.117~1.272 g/kg、3.159~53.884 mg/kg和81.510~488.422 mg/kg, 变异系数为30.636%~43.648%, 均属于中等程度变异。PCC、VIF和XGBoost变量筛选均表明, 土壤有机质、全氮、有效磷和速效钾间具有一定的关联性, 可用于目标属性空间插值的环境变量, 但不同变量筛选方法对经度、纬度、地形、气象和植被指数因子筛选结果具有一定的差异性。XGBoost方法可以更有效地筛选出对空间插值结果重要的环境变量, 且基于此方法筛选变量后建立的模型精度明显优于PCC和VIF筛选变量后建立的模型精度, 而且协同环境变量的机器学习模型精度普遍优于未加入环境变量的OK模型精度, 同一土壤养分含量空间插值模型精度从大到小依次为RF、LSTM、RBF、DT、OK, 其中基于XGBoost筛选出的变量对土壤有机质、全氮、有效磷和速效钾含量构建的RF空间插值模型精度相较于未加入环境变量的OK模型有显著提高, 决定系数分别提高43.02%、101.00%、86.04%和137.89%, 均方根误差分别降低27.39%、42.78%、13.12%和28.39%, 平均绝对误差分别降低29.01%、43.84%、11.20%和29.62%。利用RF模型对研究区农田土壤养分进行反演得到土壤有机质和全氮含量具有较强的空间分布一致性, 含量较高的主要集中在研究区南部和东部区域, 有效磷和速效钾含量具有一定的空间相似性, 东南部、中北部区域含量较低。综上, XGBoost变量筛选方法结合RF模型可以更好地实现土壤养分空间插值, 可作为土壤养分空间插值的有效方法。

关键词: 土壤养分; 变量筛选; 机器学习; 空间分布; 空间插值模型

中图分类号: S158.9

文献标识码: A

文章编号: 1000-1298(2024)10-0379-13

OSID:



Spatial Interpolation of Soil Nutrients Content Based on Environmental Variables Screening and Machine Learning

XIAN Yang^{1,2} SONG Jianghui^{1,2} WANG Jin'gang^{1,2} LI Weidi^{1,2} ZHANG Wenxu^{1,2} WANG Haijiang^{1,2}

(1. Agricultural College, Shihezi University, Shihezi 832003, China

2. The Key Laboratory of Oasis Eco-agriculture, Xinjiang Production and Construction Group, Shihezi 832003, China)

Abstract: In order to improve the accuracy of spatial interpolation of soil nutrients in farmland and accurately grasp the spatial distribution characteristics of soil nutrients, variable screening were performed by using Pearson correlation coefficient, variance inflation factor and extreme gradient boosting algorithms. Then, decision tree, random forest, radial basis function and long short-term memory were used with ordinary Kriging to interpolation the content of soil nutrients in the farmland. The results showed that the soil organic matter, total nitrogen, available phosphorus, and available potassium

收稿日期: 2023-12-20 修回日期: 2024-02-03

基金项目: 国家自然科学基金项目(42161042)、兵团科技创新领军人才项目(2023CB008-10)和兵团农业核心攻关项目(2023AA601)

作者简介: 咸阳(1998—), 女, 博士生, 主要从事农田生态环境研究, E-mail: 2720127530@qq.com

通信作者: 王海江(1980—), 男, 教授, 博士生导师, 主要从事绿洲水土资源利用研究, E-mail: wanghaijiang@shzu.edu.cn

contents in the study area ranged from 0.226 g/kg to 32.275 g/kg, 0.117 g/kg to 1.272 g/kg, 3.159 mg/kg to 53.884 mg/kg, and 81.510 mg/kg to 488.422 mg/kg, respectively, with moderate variability. PCC, VIF and XGBoost variable screening all showed that soil organic matter, total nitrogen, available phosphorus and available potassium had some correlation among them and can be used as environmental variables for the spatial interpolation of target attributes. XGBoost method can more effectively screen out the environmental variables that were important to the spatial interpolation results, and the accuracy of the model built after screening variables based on this method was significantly better than the accuracy of the model built after screening variables by PCC and VIF. Moreover, the accuracy of the machine learning model with the synergistic environmental variables was generally better than the accuracy of the OK model without environmental variables, and the accuracy of the spatial interpolation model for the same soil nutrient content showed the following order: RF > LSTM > RBF > DT > OK. Using the RF model to invert soil nutrients in the study area, it was found that the soil organic matter and total nitrogen higher content was mainly concentrated in the southern and eastern regions of the study area, the available phosphorus and available potassium lower content in the southeastern and north-central regions. In summary, the XGBoost variable screening method combined with RF model can better realize the spatial interpolation of soil nutrients, and can be used as an effective method for the spatial interpolation of soil nutrients.

Key words: soil nutrients; variable screening; machine learning; spatial distribution; spatial interpolation model

0 引言

农田土壤养分是评价土壤肥力的重要指标,是作物生长的基础,也是作物产量和品质的重要保障^[1]。土壤养分的形成和转化受到自然因素(气候、母质、地形、生物等)和人为因素(灌溉、耕作、施肥等)的共同作用,在时间和空间上均表现出了普遍的变异性^[2]。长期以来,农户通过盲目的增加化学肥料投入来提高作物产量,导致了土壤酸碱失衡、土壤环境恶化及农产品品质下降等问题,因此,对土壤养分空间分布精准插值与制图是实现土壤养分精准管理、减少农业生产中的盲目投入、保护和改善耕地土壤环境质量的必要前提。

传统土壤调查与制图主要是通过实地采样、室内化学分析和土壤地理信息系统制图,未考虑环境变量影响,单一的空间插值制图结果精度较低^[3]。3S技术和机器学习的发展与应用,促进了数字土壤制图的发展。目前,统计学方法^[4]、地统计学方法^[5]和机器学习方法^[6-7]等模型被广泛应用于数字土壤制图领域,其中统计学和地统计学方法并不能准确表达土壤属性与环境变量之间复杂的非线性关系,而机器学习方法由于其学习能力强、精度高、抗过拟合能力强、对处理复杂的非线性数据的优势被越来越多地应用于土壤属性空间插值^[8]。近年来,随机森林、决策树、支持向量回归、人工神经网络和长短期记忆网络等多种机器学习方法被用于估测土壤属性空间分布,其有效性和可靠性在多项研究中得到证实^[9-11]。但机器学习方法在不同区域,对不同土壤属性的估测性能表现不同,由于机器学习方

法在土壤属性空间插值中的应用起步较晚,为特定土壤属性选择最佳的空间插值模型还需进一步研究。

土壤属性的空间插值方法是精确反映土壤属性空间分异的基础,单一利用土壤属性估测区域土壤空间分布信息,往往只考虑样点之间距离的空间关联性,忽略了环境变量对土壤属性形成的局部差异性^[12]。现有研究表明,土壤属性空间插值中,植被指数、盐分指数、高程、坡度、坡向、气温、降水量和土地利用类型等环境因子是最常用的辅助变量,与土壤属性之间关系密切且容易获取,能够从不同角度表征土壤属性变化^[13-15]。然而不同辅助变量对土壤属性空间分布影响程度不同,不同辅助变量之间可能存在冗余信息,为了提高模型插值精度和效率,选择合适的变量筛选方法尤为重要^[16]。在对土壤属性进行空间插值建模变量筛选时,部分学者采用皮尔逊相关法、极端梯度提升、灰色关联分析法和循环迭代计算变量重要性等方法对特征参数进行优化后大大提高了模型插值精度^[17-19]。变量筛选结合机器学习方法已经在土壤属性空间插值上得到了应用,但多集中在土壤含水率、盐分含量等方面的研究,不同变量筛选方法结合不同机器学习方法估测不同土壤养分含量的精度还需进一步验证。

玛纳斯河流域绿洲区是北疆地区最大的农业耕作区,长期的人为开垦导致农田土壤养分存在较强的空间异质性^[20],因此利用机器学习对玛纳斯河流域绿洲农田土壤养分进行空间插值是实现精准施肥的重要途径之一。本文以玛纳斯河流域绿洲为研究区域,选取土壤属性、经纬度、地形、气象和植被指数

因子等辅助变量并通过皮尔逊相关系数 (PCC)、方差膨胀系数 (VIF) 和极端梯度提升 (XGBoost) 算法筛选最优变量子集作为模型输入, 构建决策树 (DT)、随机森林 (RF)、径向基函数神经网络 (RBF) 和长短期记忆网络 (LSTM) 4 种机器学习模型, 对农田土壤有机质、全氮、有效磷和速效钾 4 种土壤养分含量进行空间插值, 并与普通克里格 (OK) 插值结果进行对比, 筛选出估测不同土壤养分含量的最佳变量筛选方法和最优模型, 绘制土壤养分的空间分布图, 以期为准、快速地估测农田土壤养分含量提供理论依据。

1 材料与方法

1.1 研究区概况

玛纳斯河流域位于新疆维吾尔自治区准噶尔盆地西南部 (43°21' ~ 45°20'N, 84°43' ~ 86°35'E), 流域面积约 $3.35 \times 10^4 \text{ km}^2$, 主要分布在玛纳斯县、石河子市、沙湾市等地区^[21]。该地区常年干旱少雨, 年平均降水量 110 ~ 200 mm, 年平均蒸发量 1 700 ~ 2 200 mm, 属于典型的温带大陆性气候^[22]。流域内地势南高北低, 依次形成冲积扇、泉水溢出带、冲积平原、三角洲、湖滨平原等地貌, 为典型的山地-绿洲-盆地结构^[23]。研究区位于玛纳斯河流域绿洲区, 是新疆维吾尔自治区天山北坡经济带的重要发展区域^[24], 包括下野地灌区、安集海灌区、金沟河灌区、石河子灌区和莫索湾灌区以及新湖总场灌区 (图 1), 主要农作物为棉花、小麦和玉米。

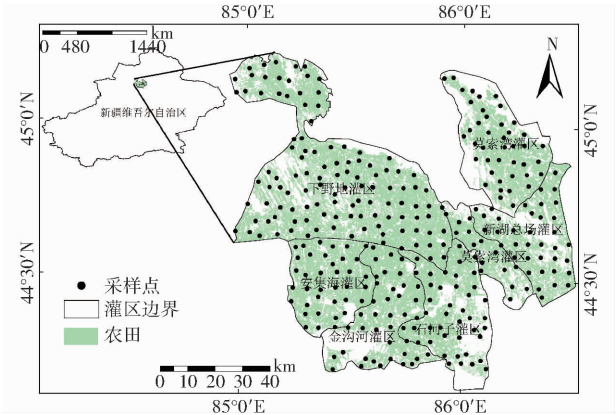


图 1 研究区地理位置及采样点分布

Fig. 1 Geographical location and sampling points distribution of study area

1.2 数据来源与处理

本研究于 2021—2022 年秋季作物收获后进行野外土壤样品采集, 考虑土壤样本采集的代表性和可获得性原则, 将研究区划分为 $5 \text{ km} \times 5 \text{ km}$ 的网格, 以网格中心作为预设采样点, 并根据地貌类型、植被类型、交通可达性等实际情况调整采样点位置,

共采集土壤样本 327 个。采样深度为 0 ~ 30 cm 的耕层, 每个采样点周围取 5 个土样混合均匀, 为该采样点待测样品, 并用 GPS 记录每个实际采样点的经纬度坐标。土样采集后带回实验室去除石子、地膜、动植物残骸等杂质, 经自然风干, 研磨后过 0.15 mm 筛。其中有机质含量采用重铬酸钾容量法-外加热法进行测定, 全氮含量采用全自动凯氏定氮仪进行测定, 有效磷含量采用碳酸氢钠提取-钼锑抗比色法进行测定, 速效钾含量采用乙酸铵提取-火焰光度法进行测定, 土壤 pH 值采用电位法测定, 盐分含量采用电导法进行测定^[25]。

环境变量由土壤属性、位置、地形、气象和植被指数因子构成 (表 1)。其中土壤类型数据来源于国家地球系统科学数据中心 (<http://soil.geodata.cn>); 地形因子数据来源于地理空间数据云网站 (<http://www.gscloud.cn/>), 是基于 DEM 数字高程数据由 ArcGIS 提取得到; 气象因子是获取研究区 4 个气象站点 2021 年每日的气温和降水数据并通过 ArcGIS 进行插值后提取得到; 植被指数是从地理空间数据云网站下载 2021 年 8 月 9 日并且云覆盖度小的 Landsat 8 OLI-TIRS 卫星影像数据, 在 ENVI 中经辐射定标、大气校正和指数提取得到, 用来反映植被

表 1 环境变量构成

Tab. 1 Composition of environment variables

变量类别	指标
土壤属性	有机质 (Soil organic matter, SOM) 含量
	全氮 (Total nitrogen, TN) 含量
	有效磷 (Available phosphorus, AP) 含量
	速效钾 (Available potassium, AK) 含量
	土壤 pH 值
	土壤盐分 (Soil salt content, SSC) 含量
	土壤类型 (Soil type, ST)
位置因子	经度 (Longitude, Long)
	纬度 (Latitude, Lat)
地形因子	高程 (Digital elevation model, DEM)
	坡度 (Slope, SLO)
	坡向 (Aspect, ASP)
气象因子	年均气温 (Annual mean temperature, AMT)
	年降水量 (Annual rainfall, AR)
	年最高气温 (Annual maximum temperature, AMAXT)
	年最低气温 (Annual minimum temperature, AMINT)
植被指数	归一化植被指数 (Normalized difference vegetation index, NDVI)
	土壤调节植被指数 (Soil adjusted vegetation index, SAVI)
	比值植被指数 (Ratio vegetation index, RVI)
	差值植被指数 (Difference vegetation index, DVI)
	绿度植被指数 (Green normalized difference vegetation index, GNDVI)

相对丰度和活性情况,从而间接反映土壤养分的供应情况。本文中所使用数据的空间分辨率均为 30 m,投影坐标系均为 WGS_1984_UTM_Zone_45N。

1.3 建模方法

1.3.1 变量筛选

影响土壤养分空间分布的环境变量有很多,并不是所有的环境变量参与机器学习建模都能取得很好的空间插值精度,因此在训练机器学习模型前,需要进行变量筛选^[26]。本文选取皮尔逊相关系数(PCC)、方差膨胀系数(VIF)和极端梯度提升(XGBoost)算法对环境变量进行筛选,其中 PCC 通过 IBM SPSS 26 软件完成,VIF 和 XGBoost 分别通过 Python 3.9.12 中的 variance_inflation_factor 函数和 XGBoost 库实现。

PCC 可以用来反映不同变量之间关系的密切程度和相关方向,相关系数取值范围为 -1~1。

VIF 是用来表征变量之间的多重共线性严重程度的一种度量,VIF 值越大,表明共线性越严重。有研究表明当 VIF 值大于 0 小于 10,不存在多重共线性;当 VIF 值大于 10 时存在高度相关^[27],对模型的贡献就会降低,需要考虑删除该变量。

XGBoost^[28]是一个经过优化的分布式梯度提升库,具有灵活、高效和可移植的优点,其自带的 feature_importances_属性可以用来衡量变量在模型中的贡献度,找出对空间插值结果最重要的变量,从而提高模型的准确性和解释性。在 XGBoost 实现过程中均采用默认参数,以减少人为因素带来的影响,并且当变量重要性指数小于 0.01 时无统计学意义^[8]。

1.3.2 空间插值模型

本研究使用 ArcGIS 实现 OK 插值,并通过 Matlab 2020a 中的 fitrtree、TreeBagger、newrbe 和 trainNetwork 函数分别对 DT、RF、RBF 和 LSTM 进行构建和训练,根据多次训练结果调整合适的模型超参数,以得到更加符合实际情况的插值结果。

普通克里格法(OK)^[29]是目前应用使用最广的克里格插值法。其原理是通过给待估测点附近的已知点赋予不同的权重,进而估计出待估测点的估测值。

决策树(DT)是一种以树形数据结构来表示决策规则和回归结果的模型,由一个根节点、一组内部子节点、一组叶子节点和多条有向边组成^[30]。DT 的构建过程是先从训练样本矩阵中选择一个变量特征进行子表划分,然后在每一个子表中选择下一个变量特征按照同样的规则继续划分更小的子表,并不断递归直到所有的变量特征全部使用完毕,此时

便得到叶级子表,估测样本根据 DT 不同层级的变量特征值进行查找,选择对应的叶级子表,用该叶级子表的输出作为待估测样本的输出。并且在构建 DT 模型时可以通过剪枝操作降低模型过拟合的风险,提高模型的泛化能力。

随机森林(RF)是利用多棵决策树对样本进行训练并估测的一种 bagging 算法,利用 bootstrap 抽样方法有放回地随机从原始训练数据集内抽取部分组成其中的一个子集用于构建决策树,重复此操作构建 NumTrees 棵决策树,将生成的多棵决策树组成随机森林,并将所有结果的平均值作为估测值输出^[31]。通过多次对比分析模型稳定性和精度,最终确定决策树数量(NumTrees)为 800。

径向基函数(RBF)神经网络是一种非线性三层向前神经网络。第 1 层是由信号源节点组成的输入层,仅向隐含层传输信号;第 2 层是由神经元组成的隐含层,对激活函数(径向基函数)的参数进行调整并与输入层建立非线性连接,将低维空间的输入通过非线性函数映射到高维空间并进行曲线拟合;第 3 层是输出层,对隐含层神经元的输出进行线性组合^[32]。在 RBF 神经网络模型中径向基函数的扩展速度(rbf_spread)是一个重要参数,其值越大,函数拟合越平滑,计算也越复杂;值越小,网络性能则会越差,愈会出现过拟合现象,需要设定合适的大小。

长短期记忆网络(LSTM)是为解决循环神经网络(Recurrent neural networks,RNN)长期依赖问题被提出的一种算法,具有重复模块链式结构。在 LSTM 中有一个可以判断信息是否有用的处理器,其结构被称为 cell,每一个 cell 包括输入门、遗忘门和输出门 3 部分,当一个信息进入 LSTM,只有符合算法认证的信息会被保留,不符合的信息会被遗忘门去除^[33]。在 LSTM 模型训练过程中优化器选择 Adam,设置最小批量为 50,最大迭代次数为 800,初始学习率为 0.01,学习率下降因子为 0.5。

1.3.3 模型精度验证

从 327 个样本数据中随机选取 80% (261 个)作为建模集,20% (66 个)作为验证集,通过计算验证集决定系数 R^2 、均方根误差(RMSE)和平均绝对误差(MAE)对模型插值精度进行验证。

2 结果与分析

2.1 土壤养分描述性统计分析

由表 2 可知,玛纳斯河流域绿洲农田土壤 SOM、TN、AP 和 AK 含量(质量比)分别为 0.226 ~ 32.275 g/kg、0.117 ~ 1.272 g/kg、3.159 ~ 53.884 mg/kg 和 81.510 ~ 488.422 mg/kg;根据均值

与全国第二次土壤普查养分分级标准^[34]对比分析,可以发现 SOM 含量处于中等(四级)水平、TN 含量处于低等(五级)水平、AP 含量处于高等(二级)水平、AK 含量则处于极高(一级)水平;4 个养分指标的空间变异程度由大到小依次为 AP 含量(43.648%)、SOM 含量(34.337%)、TN 含量(32.075%)、AK 含量(30.636%),均处于中等变异水平。

表 2 农田土壤养分含量的描述性统计特征
Tab.2 Descriptive statistical characteristics of farmland soil nutrients content

养分指标	最大值	最小值	均值	标准差
SOM 含量/(g·kg ⁻¹)	32.275	0.226	14.442	4.959
TN 含量/(g·kg ⁻¹)	1.272	0.117	0.689	0.221
AP 含量/(mg·kg ⁻¹)	53.884	3.159	26.221	11.445
AK 含量/(mg·kg ⁻¹)	488.422	81.510	303.834	93.083

2.2 变量筛选

2.2.1 皮尔逊相关性分析筛选

皮尔逊相关系数(PCC)能够表示土壤养分与环境变量之间的相关关系,利用 PCC 筛选出显著相关的变量作为模型输入变量。结果如图 2 所示

示,SOM 含量与 TN 含量、AP 含量、AK 含量、pH 值、Long、Lat、DEM、AMT、AR、AMAXT 和 AMINT 共 11 个环境变量通过了显著性检验;TN 含量与 SOM 含量、AP 含量、AK 含量、pH 值、SSC 含量、ST、Lat、DEM、AMT、AR、AMAXT、AMINT、NDVI、SAVI、RVI 和 GNDVI 共 16 个环境变量通过了显著性检验;AP 含量与 SOM 含量、TN 含量、AK 含量、SSC 含量、NDVI、SAVI、RVI、DVI 和 GNDVI 共 9 个环境变量通过了显著性检验;AK 含量则与 SOM 含量、TN 含量、AP 含量和 RVI 共 4 个环境变量通过了显著性检验。

2.2.2 方差膨胀系数筛选

计算各环境变量的 VIF 值并设置筛选环境变量的阈值为 10^[32],筛选后的结果如图 3 所示,由于 Long、Lat、AMT、AR、AMAXT、AMINT、NDVI、SAVI、DVI 和 GNDVI 的 VIF 值均大于 10,因此不参与建模。估测 SOM 含量的环境变量有 TN 含量、AP 含量、AK 含量、pH 值、SSC 含量、ST、DEM、SLO、ASP 和 RVI 共 10 个作为模型输入,其中 DEM 的 VIF 值为 6.09,其次是 RVI(5.63),而 SLO 的 VIF 值最小(1.10)。估测 TN 含量的环境变量有 SOM 含量、AP

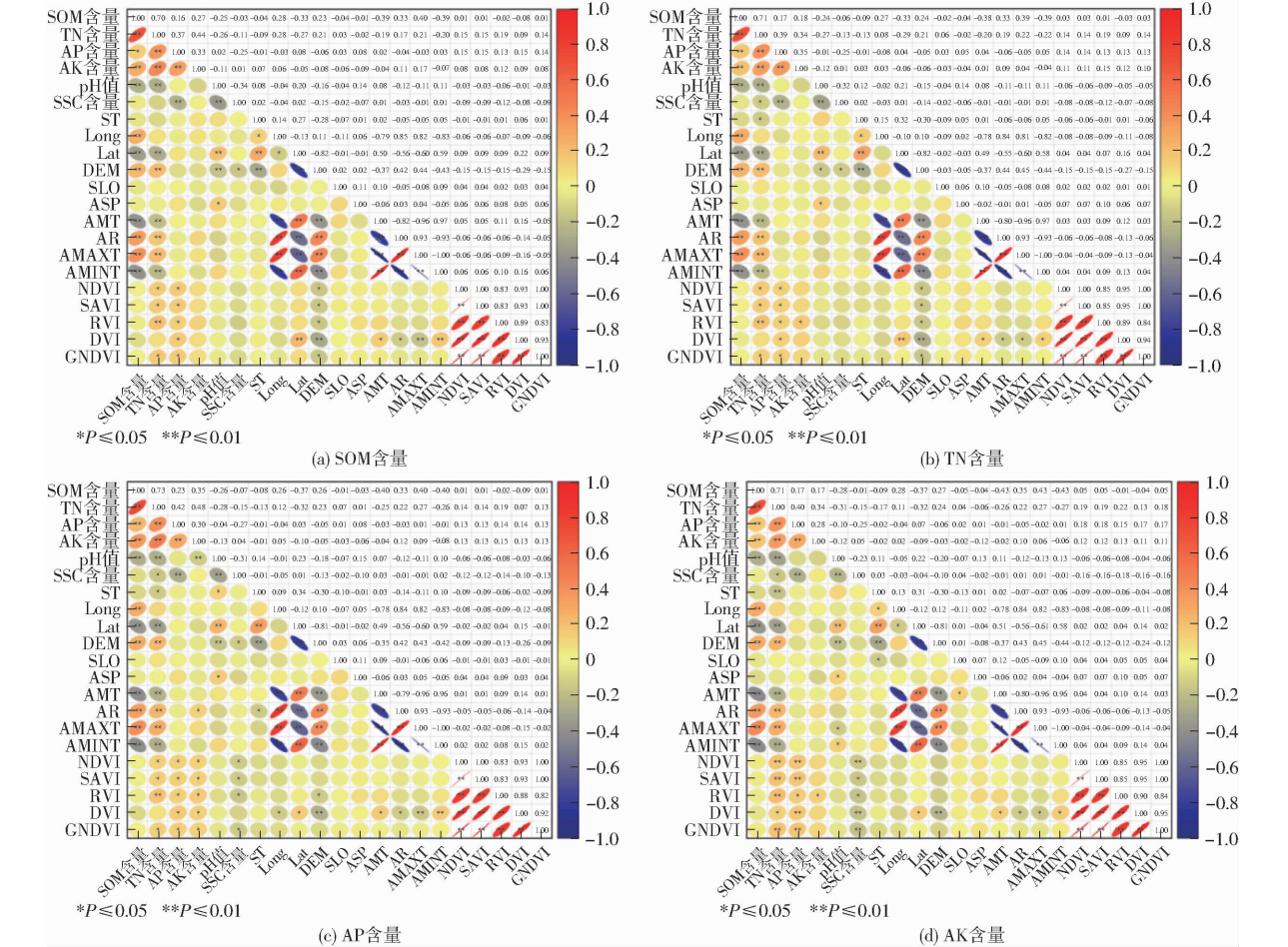


图 2 Person 相关关系矩阵图

Fig.2 Person correlation coefficient matrices

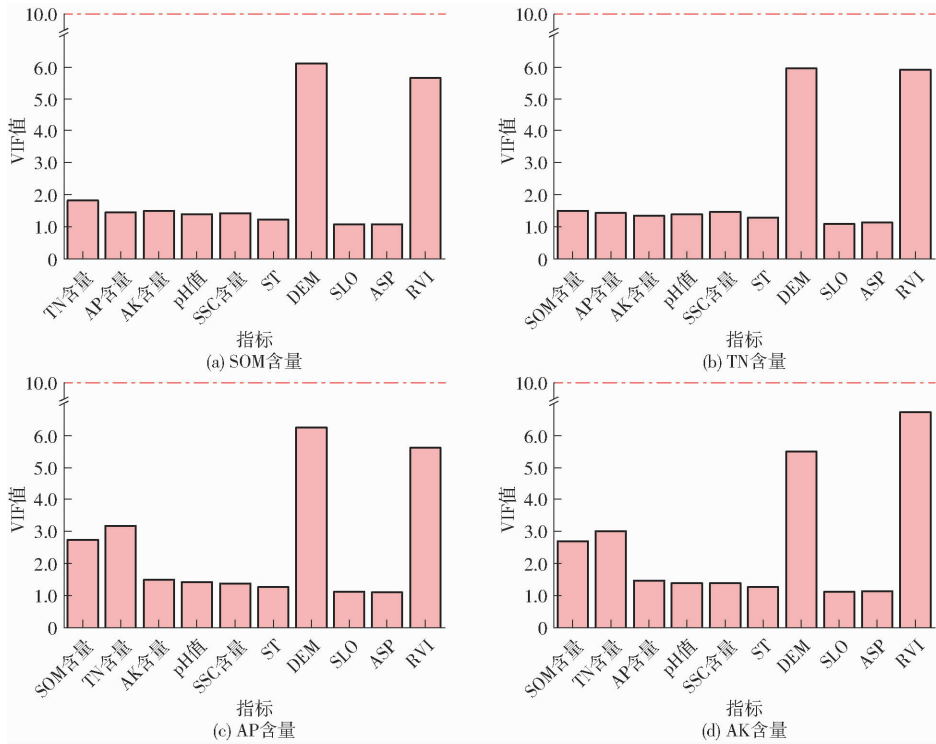


图 3 VIF 变量筛选结果
Fig. 3 VIF screening results

含量、AK 含量、pH 值、SSC 含量、ST、DEM、SLO、ASP 和 RVI 共 10 个作为模型输入,其中 DEM 的 VIF 值为 5.93,其次是 RVI(5.89),而 SLO 的 VIF 值最小(1.08)。估测 AP 含量的环境变量有 SOM 含量、TN 含量、AK 含量、pH 值、SSC 含量、ST、DEM、SLO、ASP 和 RVI 共 10 个作为模型输入,其中 DEM 的 VIF 值为 6.24,其次是 RVI(5.61),而 ASP 的 VIF 值最小(1.10)。估测 AK 含量的环境变量有 SOM 含量、TN 含量、AP 含量、pH 值、SSC 含量、ST、DEM、SLO、ASP 和 RVI 共 10 个作为模型输入,其中 RVI 的 VIF 值为 6.73,其次是 DEM(5.51),而 SLO 的 VIF 值最小(1.11)。

2.2.3 变量重要性筛选

通过十折交叉验证法训练 XGBoost 算法计算变量重要性,将筛选阈值设置为 0.01,在对插值 SOM 含量的变量进行筛选时共有 15 个变量被筛选出参与建模,其中 TN 含量重要性指数为 0.44,其次是 AMAXT(重要性指数为 0.16),而 ST、APS、SAVI、RVI 和 GNDVI 重要性指数均小于 0.01,因此不参与建模。在对插值 TN 含量的变量进行筛选时共有 14 个变量被筛选出参与建模,其中 SOM 含量重要性指数为 0.54,其次是 AP 含量(重要性指数为 0.11),而 ST、Long、DEM、AMINT、SAVI 和 RVI 的重要性指数均小于 0.01,全部剔除。在对插值处理 AP 含量的变量进行筛选时共有 18 个变量被筛选出作为模型输入,其中 TN 含量重要性指数为 0.23,其次是

AK 含量(重要性指数为 0.08),而 SAVI 和 RVI 重要性指数小于 0.01,因此剔除这两个变量。在对插值处理 AK 含量的变量进行筛选时共有 16 个变量被筛选出作为模型输入,其中重要性指数最大的变量是 AR(重要性指数为 0.19),其次是 TN 含量(重要性指数为 0.18),而 ST、AMINT、SAVI 和 RVI 的重要性均小于 0.01,因此不参与建模。筛选后结果如图 4 所示。

2.3 模型插值精度评价

通过 PCC、VIF 和 XGBoost 筛选变量方法建立土壤养分空间插值变量集,对所选变量集分别构建 DT、RF、RBF 和 LSTM 模型,然后对插值结果进行精度验证,并与 OK 模型插值精度进行对比(表 3),从 R^2 、RMSE 和 MAE 指标来看,基于 XGBoost 算法筛选变量建立的模型精度都比基于 PCC 和 VIF 方法筛选变量建立的模型精度高;不同土壤养分插值模型中 RF 模型的插值精度都明显高于其他 4 个模型,而 OK 的插值精度最低;在基于 XGBoost 筛选变量的 RF 模型插值中, SOM、TN 含量 R^2 分别为 0.748、0.802, RMSE 分别为 2.680、0.107 g/kg, MAE 分别为 2.073、0.082 g/kg,具有较高的空间插值精度,这是因为 SOM 和 TN 之间较强的正相关性会使对目标变量的插值结果更加准确可靠^[35],而对于 AP、AK 含量插值中精度不高, R^2 分别为 0.413、0.540, RMSE 分别为 8.996、68.291 mg/kg, MAE 分别为 7.365、56.100 mg/kg,但相比于其他模型,

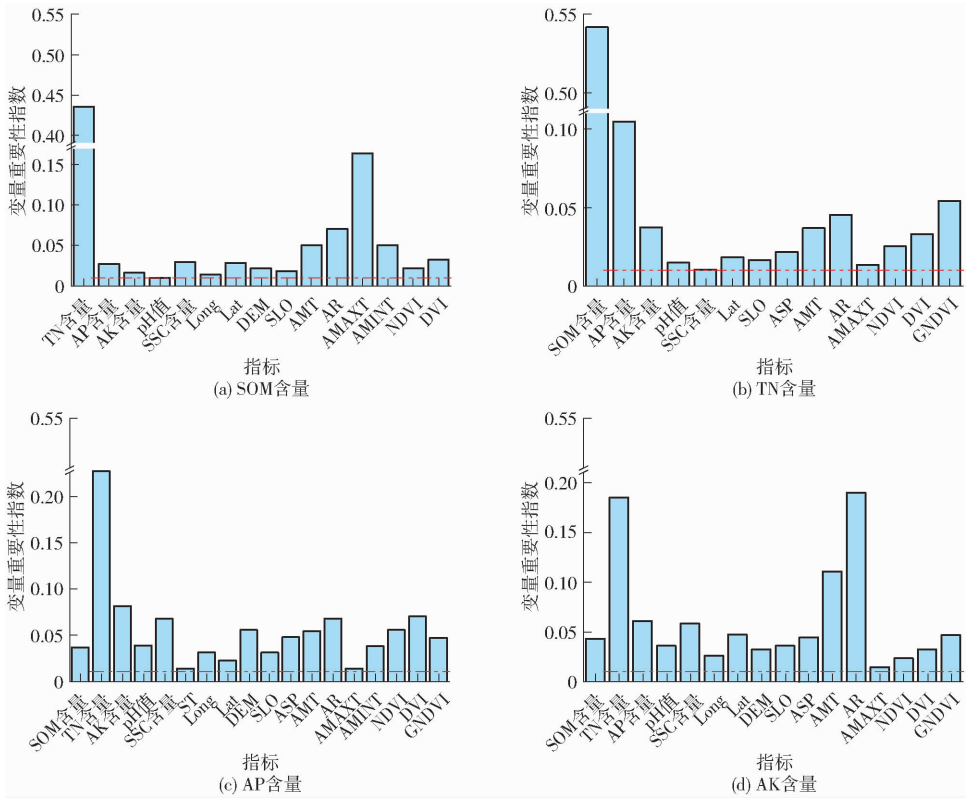


图 4 XGBoost 变量筛选结果
Fig. 4 XGBoost screening results

表 3 土壤养分含量空间插值结果精度验证

筛选方法		SOM 含量			TN 含量			AP 含量			AK 含量		
		R^2	RMSE/ ($\text{g}\cdot\text{kg}^{-1}$)	MAE/ ($\text{g}\cdot\text{kg}^{-1}$)	R^2	RMSE/ ($\text{g}\cdot\text{kg}^{-1}$)	MAE/ ($\text{g}\cdot\text{kg}^{-1}$)	R^2	RMSE/ ($\text{mg}\cdot\text{kg}^{-1}$)	MAE/ ($\text{mg}\cdot\text{kg}^{-1}$)	R^2	RMSE/ ($\text{mg}\cdot\text{kg}^{-1}$)	MAE/ ($\text{mg}\cdot\text{kg}^{-1}$)
PCC	DT	0.552	3.579	2.723	0.676	0.137	0.101	0.205	10.468	8.908	0.331	82.306	66.355
	RF	0.722	2.818	2.161	0.785	0.112	0.087	0.300	9.823	8.237	0.455	74.304	60.796
	RBF	0.657	3.130	2.266	0.735	0.124	0.103	0.272	10.020	8.118	0.358	80.673	62.383
	LSTM	0.687	2.990	2.314	0.770	0.116	0.092	0.281	9.957	8.241	0.422	76.510	64.067
VIF	DT	0.571	3.499	2.637	0.658	0.141	0.115	0.232	10.287	8.552	0.350	81.159	64.338
	RF	0.737	2.739	2.091	0.753	0.120	0.092	0.326	9.637	7.992	0.490	71.906	59.670
	RBF	0.673	3.054	2.416	0.688	0.134	0.109	0.301	9.817	8.094	0.381	79.186	64.413
	LSTM	0.713	2.865	2.061	0.735	0.124	0.099	0.315	9.719	8.047	0.476	72.844	60.114
XGBoost	DT	0.617	3.307	2.481	0.682	0.136	0.106	0.242	10.218	8.439	0.361	80.454	62.173
	RF	0.748	2.680	2.073	0.802	0.107	0.082	0.413	8.996	7.365	0.540	68.291	56.100
	RBF	0.680	3.021	2.236	0.752	0.120	0.088	0.361	9.384	7.642	0.404	77.711	66.153
	LSTM	0.735	2.749	2.174	0.781	0.113	0.090	0.405	9.059	7.352	0.509	70.556	55.571
OK		0.523	3.691	2.920	0.399	0.187	0.146	0.222	10.354	8.294	0.227	95.365	79.711

RF 模型的 R^2 仍然最大, RMSE 和 MAE 最小, 精度明显优于 OK、DT、RBF 和 LSTM 模型, 说明在流域尺度上 RF 模型对农田土壤养分具有不错的回归插值效果。

2.4 农田土壤养分空间分布特征

基于 XGBoost 算法筛选的变量集构建 RF 模型对玛纳斯河流域绿洲农田土壤养分进行空间插值并绘制空间分布图(图 5), 4 个土壤养分的空间分布

特征差异比较明显, 均呈斑块状分布, 其中 SOM 和 TN 空间分布规律较一致, SOM 含量估测值范围为 4.112 ~ 27.212 g/kg, TN 含量估测值范围为 0.210 ~ 1.174 g/kg, 均是从西北到东南养分含量呈现逐步增加的趋势, 含量较高区域主要分布在研究区东南部的石河子灌区范围内; AP 含量估测值范围为 13.160 ~ 41.978 mg/kg, 含量较高区域主要分布在研究区中部和东北部, 而东部新湖总场灌区含量

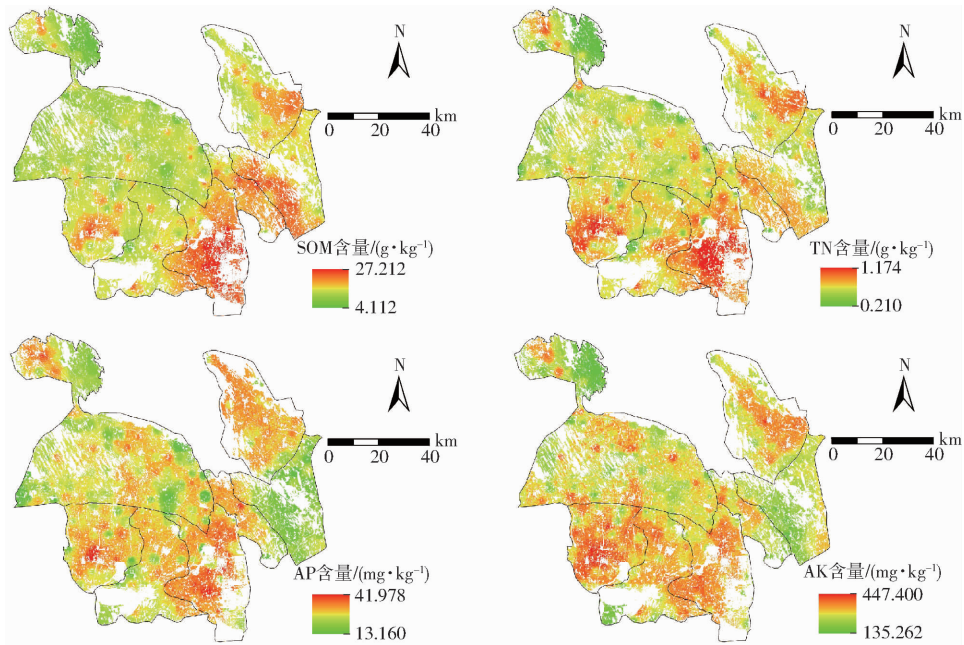


图 5 农田土壤养分含量的空间分布

Fig. 5 Spatial distribution of soil nutrients content in farmland

整体偏低; AK 含量估测值范围为 135.262 ~ 447.400 mg/kg, 含量中高值区域占研究区大部分, 主要分布在安集海灌区、金沟河灌区和石河子灌区, 而新湖总场灌区的含量偏低。根据土壤养分空间分布特征, 在研究区西北部区域可适度增施有机肥、深翻改土、作物轮作, 东部区域可适度增施磷肥和钾肥、秸秆还田, 提高农田生产力。

2.5 不确定性分析

通过最优插值模型迭代 50 次计算估测值标准

差(Standard deviation, SD)对模型插值不确定性进行分析, 结果如图 6 所示, SOM、TN、AP、AK 含量估测值标准差和变异系数均值分别为 0.114 g/kg 和 0.761%、0.005 g/kg 和 0.736%、0.312 mg/kg 和 1.168%、2.814 mg/kg 和 0.883%, 数据分布集中, 说明估测值波动较小, 模型稳定。总体上模型插值不确定性都比较低, 其中 TN 含量插值模型不确定性最低, 接着是 SOM、AK 含量, 而 AP 含量插值模型不确定性最高。SOM 和 TN 含量在研究区东南部表

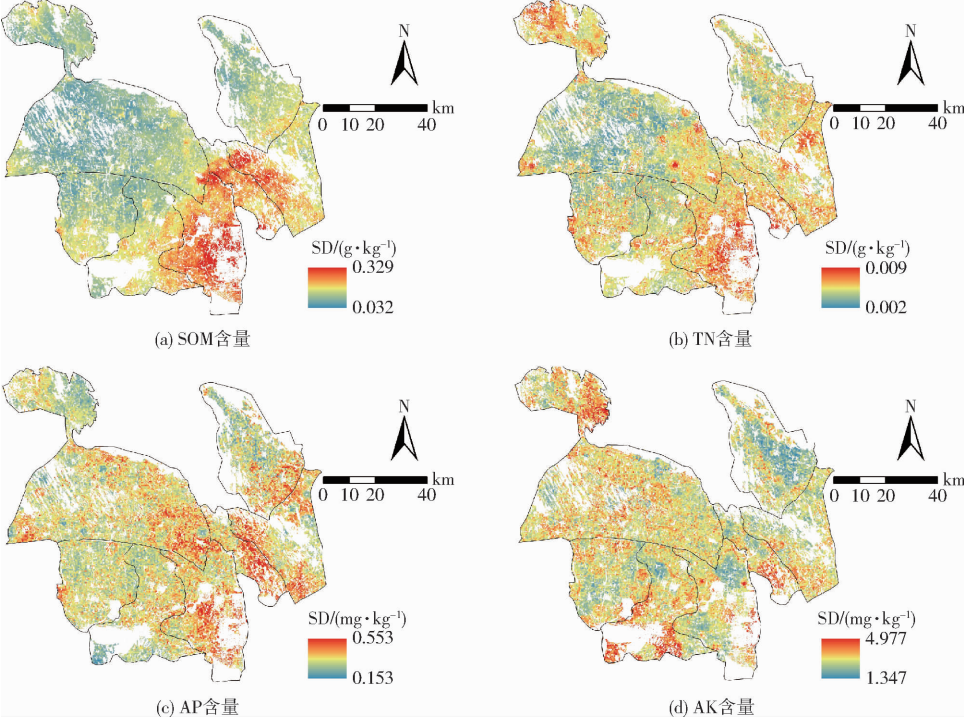


图 6 农田土壤养分估测标准差分布图

Fig. 6 Distribution maps of estimated standard deviation for soil nutrients in farmland

现出较大的插值不确定性,AP 含量在研究区中部和东南部表现出较大的插值不确定性,而 AK 含量在研究区西北部和南部表现出较大的插值不确定性,这可能是受到环境变量较强的空间变异性以及实验分析、数据集划分、插值模型等误差的影响。

3 讨论

3.1 影响土壤养分空间分布的主要环境变量

土壤养分空间分布受多种环境变量共同影响,本研究通过皮尔逊相关性分析表明土壤 AP 与 SOM 含量呈显著正相关关系,而其他土壤养分之间均呈极显著正相关关系,这可能是因为 SOM 与 TN、AP 和 AK 含量之间存在着一定的耦合作用^[36],伴随着有机质中腐殖质的释放,腐殖质携带的大量负电荷可以有效地吸收土壤速效养分,从而直接影响其空间分布。相关研究表明,随着土壤 pH 值的升高,土壤有机质和全氮含量显著降低^[37],本研究中,pH 值与 SOM 和 TN 含量之间呈极显著负相关关系,这或许是因为本研究区以碱性土壤为主,土壤 pH 值为 7.43~9.94,而 pH 值在 6.5~7.5 的中性环境是土壤微生物最佳的生存环境。在碱性土壤环境中,与有机质(如土生菌和诺卡氏菌)和全氮(如硝化和反硝化细菌)相关的土壤微生物丰度降低,活性被抑制,进而影响了有机质和氮素养分在土壤中的迁移、转化和供应^[38]。另外,土壤盐分含量也是影响土壤养分空间分布的主要环境变量,本研究中土壤盐分含量与 TN 含量呈显著负相关关系,与 AP 含量呈极显著负相关关系,而与 SOM 和 AK 含量的相关性不显著,可能是因为随着土壤盐渍化程度加深,土壤中大量的 Na⁺使土壤颗粒崩解、膨胀、分散,破坏了土壤的团粒结构,加剧了土壤养分的淋溶损失以及降低了土壤碳氮养分的周转效率^[39],进而对土壤养分的积累和矿化分解过程产生了消极的影响。不同的土壤类型,由于成土母质、风化类型及时间的因素,土壤养分的类型、含量及空间分布具有显著的差异性^[40]。在本研究中,皮尔逊相关性分析表明,土壤类型与 TN 含量呈显著负相关,而与 SOM、AP、AK 含量无显著相关性,这是因为研究区由灰漠土、盐土和风沙土等多种土壤类型构成,加之该区域干旱、缺水等复杂的气候环境,因此土壤养分含量在空间分布上表现出异质性。并且由于土壤属性数据获取困难,很少被作为环境变量用于土壤养分空间插值。

除了受土壤属性因子的影响,经、纬度和地形因子也是影响土壤养分空间分布的主要环境变量,本研究中,经度与 SOM 含量、纬度和高程与 SOM、TN

含量之间均呈极显著相关关系,而经度与 TN、AP 和 AK 含量,纬度和高程与 AP、AK 含量之间均无显著相关性,这是因为经度、纬度和高程可以通过影响区域土壤湿度和温度进而影响矿物分化和土壤微生物分解速率^[41],从而影响土壤养分的循环和积累。而坡度和坡向可以通过影响地表径流进而影响土壤养分运移过程,但本文中坡度和坡向与 SOM、TN、AP 和 AK 含量之间相关性不显著,可能是因为研究区农田地势相对平坦,限制了地表径流的产生,进而影响了 SOM、TN、AP 和 AK 的运移。此外,气象因子也是影响土壤养分空间分布的重要环境变量,年均气温、年降水量、年最高气温和年最低气温与 SOM、TN 含量之间呈极显著相关性,而与 AP 和 AK 含量之间相关性不显著,这可能是因为研究区属于温带大陆性气候,干燥少雨,土壤含水率较低,影响了对土壤养分的淋溶作用,并且磷和钾本身极易被固定,迁移能力较弱^[42],进而影响了土壤养分的释放和迁移过程。另外,植被指数因子对土壤养分空间分布也有重要影响,本研究中归一化植被指数与 TN 和 AP 含量、土壤调节植被指数与 TN 和 AP 含量、比值植被指数与 AP 和 AK 含量、差值植被指数与 AP 含量、绿度植被指数与 TN 和 AP 含量之间均呈显著正相关关系,比值植被指数与 TN 含量之间呈极显著正相关关系,这是因为植被覆盖度较高时,其植物凋落物和根系系统可以有效减轻土壤侵蚀,降低土壤养分流失从而有利于土壤养分的积累^[43]。

本研究也存在着不足之处,土壤养分空间分布受土壤质地、地形、气候和人为等因素共同影响,但是本文中并没有考虑人为活动因素的影响,在今后的研究中可以加入对土壤养分空间分布有影响的如施肥^[44]、灌溉^[45]、耕作方式^[46]等人为活动因素,从而更全面地了解土壤养分空间分布影响因素并且进一步提高模型空间插值精度和泛化能力。

3.2 不同变量筛选和建模方法对模型精度的影响

变量筛选是模型建立的前提,通过变量筛选可以剔除没有估测能力的变量,减少数据维度,降低模型复杂度,提高模型精度和泛化能力^[47-48]。本文引入皮尔逊相关系数(PCC)、方差膨胀系数(VIF)和极端梯度提升(XGBoost)算法对环境变量进行筛选,研究结果也表明通过变量筛选可以有效提高土壤养分空间插值模型的精度和稳定性,并且基于 XGBoost 筛选变量建立的模型精度最高。主要是 PCC 和 VIF 考虑的都是变量之间的线性关系,而土壤养分与环境变量之间的关系往往是非常复杂的,XGBoost 可以衡量变量在模型中的贡献,能更好地筛选出对土壤养分含量插值结果最重要的变量,从

而使模型得到更高的插值精度。

不同建模方法对同一目标变量的插值精度不同。本研究基于普通克里格(OK)、决策树(DT)、随机森林(RF)、径向基函数神经网络(RBF)和长短期记忆网络(LSTM)5种方法对SOM、TN、AP和AK含量构建空间插值模型,结果表明RF模型插值精度明显优于其他4个模型,这是因为RF是一种集成算法,相比于其他模型具有抗噪声能力和泛化能力强,不易过拟合的优势,可以更好地解释土壤养分和筛选出的环境变量之间复杂的非线性和层状关系,能够更好地揭示土壤养分的局部空间分布特征^[49]。OK模型仅从空间位置关联上对土壤养分进行插值,没有考虑环境变量的影响导致插值效果不理想。DT模型虽然可以解释土壤养分与环境变量间的非线性关系,但它只构建1棵决策树,即只进行1次递归过程,不像RF模型是通过构建多棵决策树(本研究中决策树的数量为800棵)并对结果求取均值作为最后的决策结果,因此其精度也明显低于RF模型。RBF和LSTM模型需要大量的数据进行训练,在本研究中可能训练数据不足导致模型泛化能力不强^[50],插值效果低于RF模型。同样,文献[51–55]的研究

结果也指出RF模型在SOM、TN、AP和AK含量插值中取得了不错的拟合效果, R^2 为0.59~0.79,与本文最优模型插值结果基本一致,这进一步证实RF模型对估测SOM、TN、AP和AK含量具有较好的有效性和适用性,可以作为一种估测农田土壤养分空间分布的新途径。

4 结 论

- (1)玛纳斯河流域绿洲农田土壤SOM、TN、AP和AK含量分别处于中等、低、高和极高水平,属于中等变异水平,其中AP含量的空间变异程度最大,为43.648%,土壤SOM、TN含量与部分土壤属性、经度、纬度、地形、气象、植被指数因子显著相关,AP、AK含量与部分土壤属性和植被指数显著相关。
- (2)PCC、VIF和XGBoost变量筛选方法可以减少数据维度,消除变量间的共线性,基于XGBoost筛选出的环境变量构建的模型空间插值效果最好,具有更高 R^2 和更小的RMSE。
- (3)协同环境变量的DT、RF、RBF和LSTM模型插值精度均比OK模型的精度高,RF插值效果最佳,空间插值不确定性较低。

参 考 文 献

[1] TANMOY S, CHANDRA G M, MAHUA B, et al. Prediction of the effect of nutrients on plant parameters of rice by artificial neural network[J]. Agronomy, 2022, 12(9): 2123.

[2] 林建平, 邓爱珍, 赵小敏, 等. 南方典型丘陵山区不同高程耕地土壤养分变化特征分析[J]. 农业机械学报, 2019, 50(5): 300–309.

LIN Jianping, DENG Aizhen, ZHAO Xiaomin, et al. Variation characteristics of soil nutrients of cultivated land in different elevation fields in typical hilly areas of southern mountains[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(5): 300–309. (in Chinese)

[3] 曹永研, 杨玮, 王懂, 等. 基于水分和粒度的土壤有机质特征波长提取与预测模型[J]. 农业机械学报, 2022, 53(增刊1): 241–248.

CAO Yongyan, YANG Wei, WANG Dong, et al. Soil organic matter characteristic wavelength extraction and prediction model based on moisture and particle size[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(Supp. 1): 241–248. (in Chinese)

[4] SASIRIN S, MASAYUKI K, PITAYAKON L. Digital mapping of soil chemical properties with limited data in the Thung Kula Ronghai region, Thailand[J]. Geoderma, 2021, 389: 114942.

[5] 李娜, 陈暑晃, 赖宁, 等. 基于GIS和地统计的伊犁州农田土壤养分空间变异特征分析[J]. 新疆农业科学, 2019, 56(11): 2096–2103.

LI Na, CHEN Shuhuang, LAI Ning, et al. Spatial variability analysis of soil nutrients in farmland of Yili Prefecture based on GIS and geostatistics[J]. Xinjiang Agricultural Science, 2019, 56(11): 2096–2103. (in Chinese)

[6] CHIRANJIT S, CHANDRA K S, SATIPRASAD S, et al. Prediction of soil nutrients through PLSR and SVMR models by VIs–NIR reflectance spectroscopy[J]. The Egyptian Journal of Remote Sensing and Space Sciences, 2023, 26(4): 901–918.

[7] 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/OL]. 生态学杂志, <http://kns.cnki.net/kcms/detail/21.1148.Q.20231120.1134.002.html>.

[8] 马国林, 丁建丽, 韩礼敬, 等. 基于变量优选与机器学习的干旱区湿地土壤盐渍化数字制图[J]. 农业工程学报, 2020, 36(19): 124–131.

MA Guolin, DING Jianli, HAN Lijing, et al. Digital mapping of soil salinization in arid area wetland based on variable optimized selection and machine learning[J]. Transactions of the CSAE, 2020, 36(19): 124–131. (in Chinese)

[9] FARZANEH P, AHMAD F F, ROHOLLAH S M, et al. Large-scale digital mapping of topsoil total nitrogen using machine

- learning models and associated uncertainty map[J]. *Environmental Monitoring and Assessment*, 2021, 193(4): 162.
- [10] HEECHAN H, CHANGHYUN C, JONGSUNG K, et al. Multiple-depth soil moisture estimates using artificial neural network and long short-term memory models[J]. *Water*, 2021, 13(18): 2584.
- [11] 唐茂淞, 张楠, 李国辉, 等. 基于机器学习算法的棉田土壤钾、钠离子量预测[J]. *灌溉排水学报*, 2023, 42(9): 32–39.
TANG Maosong, ZHANG Nan, LI Guohui, et al. Predicting soil K^+ and Na^+ contents in cotton field using machine learning algorithm[J]. *Journal of Irrigation and Drainage*, 2023, 42(9): 32–39. (in Chinese)
- [12] 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 Yapeng, ZHANG Junhua, LIU Yuhan, 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. (in Chinese)
- [13] 任必武, 陈瀚阅, 张黎明, 等. 机器学习用于耕地土壤有机碳空间预测对比研究——以亚热带复杂地貌区为例[J]. *中国生态农业学报(中英文)*, 2021, 29(6): 1042–1050.
REN Biwu, CHEN Hanyue, ZHANG Liming, et al. Comparison of machine learning for predicting and mapping soil organic carbon in cultivated land in a subtropical complex geomorphic region[J]. *Chinese Journal of Eco-Agriculture*, 2021, 29(6): 1042–1050. (in Chinese)
- [14] MOJTABA Z, YOUNES G, HAMID O R, et al. Improving the spatial prediction of soil organic carbon using environmental covariates selection: a comparison of a group of environmental covariates[J]. *Catena*, 2022, 208: 105723.
- [15] 张振华, 丁建丽, 王敬哲, 等. 集成土壤-环境关系与机器学习的干旱区土壤属性数字制图[J]. *中国农业科学*, 2020, 53(3): 563–573.
ZHANG Zhenhua, DING Jianli, WANG Jingzhe, et al. Digital soil properties mapping by ensembling soil–environment relationship and machine learning in arid regions[J]. *Scientia Agricultura Sinica*, 2020, 53(3): 563–573. (in Chinese)
- [16] 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 Yuqi, CHEN Hanyue, ZHANG Liming, 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. (in Chinese)
- [17] ZHAO Jianhui, ZHANG Chenyang, MIN Lin, et al. Retrieval of farmland surface soil moisture based on feature optimization and machine learning[J]. *Remote Sensing*, 2022, 14(20): 5102.
- [18] 韩文霆, 崔家伟, 崔欣, 等. 基于特征优选与机器学习的农田土壤含盐量估算研究[J]. *农业机械学报*, 2023, 54(3): 328–337.
HAN Wenting, CUI Jiawei, CUI Xin, et al. Estimation of farmland soil salinity content based on feature optimization and machine learning algorithms[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2023, 54(3): 328–337. (in Chinese)
- [19] 王飞, 杨胜天, 丁建丽, 等. 环境敏感变量优选及机器学习算法预测绿洲土壤盐分[J]. *农业工程学报*, 2018, 34(22): 102–110.
WANG Fei, YANG Shengtian, DING Jianli, et al. Environmental sensitive variable optimization and machine learning algorithm using in soil salt prediction at oasis[J]. *Transactions of the CSAE*, 2018, 34(22): 102–110. (in Chinese)
- [20] 朱磊, 杨爱民, 夏鑫鑫, 等. 基于空间自相关的 1975—2015 年玛纳斯河流域耕地时空特征变化分析[J]. *中国生态农业学报(中英文)*, 2020, 28(6): 887–899.
ZHU Lei, YANG Aimin, XIA Xinxin, et al. Spatial distribution pattern and change characteristics analysis of cultivated land in the Manas River basin from 1975 to 2015[J]. *Chinese Journal of Eco-Agriculture*, 2020, 28(6): 887–899. (in Chinese)
- [21] 杨小虎. 玛纳斯河流域绿洲土壤盐分反演及土壤质量评价[D]. 石河子: 石河子大学, 2021.
YANG Xiaohu. Soil salinity inversion and soil quality assessment in oasis of Manas River basin[D]. Shihezi: Shihezi University, 2021. (in Chinese)
- [22] 宋江辉, 史晓艳, 王海江, 等. 干旱区三种典型地貌下电磁感应式土壤盐分协同解译模型[J]. *农业工程学报*, 2021, 37(6): 81–90.
SONG Jianghui, SHI Xiaoyan, WANG Haijiang, et al. Synergistic interpretation model for soil salinity by electromagnetic induction under three typical landforms in arid areas[J]. *Transactions of the CSAE*, 2021, 37(6): 81–90. (in Chinese)
- [23] 衡通. 干旱区膜下滴灌棉田水盐与养分累积特征及优化调控[D]. 石河子: 石河子大学, 2022.
HENG Tong. Accumulation characteristics optimal regulation of water, salt and nutrient in film-mulched cotton (*Gossypium hirsutum* L.) field under drip irrigation in arid area[D]. Shihezi: Shihezi University, 2022. (in Chinese)
- [24] 陈东. 玛纳斯河流域绿洲驱动力分析及适宜规模研究[D]. 石河子: 石河子大学, 2019.
CHEN Dong. Study on the driving force analysis and suitable scale of oasis in the Manas River basin[D]. Shihezi: Shihezi University, 2019. (in Chinese)
- [25] 鲍士旦. 土壤农化分析[M]. 3 版. 北京: 中国农业出版社, 2000.
- [26] 赵文举, 段威成, 王银凤, 等. 基于敏感变量筛选的多光谱植被含水率反演模型研究[J]. *农业机械学报*, 2023,

54(9): 343–351, 385.

ZHAO Wenju, DUAN Weicheng, WANG Yinfeng, et al. Multispectral vegetation water content inversion model based on sensitive variable screening[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(9): 343–351, 385. (in Chinese)

[27] ZHANG Xianglin, CHEN Songchao, XUE Jie, et al. Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping[J]. Geoderma, 2023, 432: 116383.

[28] 刘尊方, 雷浩川, 盛海彦. 基于 XGBoost 模型的湟水流域耕地土壤养分遥感反演[J]. 干旱区地理, 2023, 46(10): 1643–1653.

LIU Zunfang, LEI Haochuan, SHENG Haiyan. Remote sensing inversion of soil nutrient on farmland in Huangshui River basin based on XGBoost model[J]. Arid Land Geography, 2023, 46(10): 1643–1653. (in Chinese)

[29] 万核洋, 齐泓玮, 尚松浩. 河套灌区土壤质地克里金插值与平滑效应校正[J]. 农业机械学报, 2023, 54(1): 339–350.

WAN Heyang, QI Hongwei, SHANG Songhao. Ordinary Kriging interpolation and smoothing effect correction for soil texture mapping in Hetao Irrigation District[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(1): 339–350. (in Chinese)

[30] NEMAI S, TIRTHANKAR G. GIS-based spatial prediction of recreational trail susceptibility in protected area of Sikkim Himalaya using logistic regression, decision tree and random forest model[J]. Ecological Informatics, 2021, 64: 101352.

[31] 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 Li, YANG Lian'an, WANG Hui, 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. (in Chinese)

[32] 谢梦姣, 王洋, 康营, 等. 人工神经网络与普通克里金插值法对土壤属性空间预测精度影响研究[J]. 生态与农村环境学报, 2021, 37(7): 934–942.

XIE Mengjiao, WANG Yang, KANG Ying, et al. Accuracy study of spatial predicting in soil attributes based on interpolations by artificial neural net work and ordinary Kriging[J]. Journal of Ecology and Rural Environment, 2021, 37(7): 934–942. (in Chinese)

[33] WANG Huan, ZHANG Lixin, ZHAO Jiawei, et al. Application of hyperspectral technology combined with genetic algorithm to optimize convolution long-and short-memory hybrid neural network model in soil moisture and organic matter[J]. Applied Sciences, 2022, 12(20): 10333.

[34] 全国土壤普查办公室. 中国土壤普查技术[M]. 北京: 农业出版社, 1992.

[35] 张迪, 姜柏志, 刘国辉, 等. 基于不同插值方法的三江平原白浆土磷素空间分布预测及其适用性分析[J]. 中国生态农业学报(中英文), 2021, 29(8): 1405–1416.

ZHANG Di, JIANG Bozhi, LIU Guohui, et al. Applicability of spatial interpolation methods to predict total phosphorus in the typical irrigated areas of the Sanjiang Plain[J]. Chinese Journal of Eco-Agriculture, 2021, 29(8): 1405–1416. (in Chinese)

[36] 何家莉, 宋怡珂, 王金牛, 等. 岷江源区高山林草交错带土壤碳、氮、磷生态化学计量关系的时空变化[J]. 应用与环境生物学报, 2021, 27(4): 869–877.

HE Jiali, SONG Yike, WANG Jinniu, et al. Spatial and temporal variations in soil C, N, and P ecological stoichiometry across the alpine forests and grassland ecotone at the Minjiang Headwaters[J]. Chinese Journal of Applied and Environmental Biology, 2021, 27(4): 869–877. (in Chinese)

[37] 刘靖宇, 尹芳, 刘磊, 等. 荒漠绿洲区不同土地利用方式土壤生态化学计量特征差异及驱动因素[J/OL]. 中国环境科学, <https://doi.org/10.19674/j.cnki.issn1000-6923.20230911.004>.

[38] 荣方旭, 吕十全, 胡树翔, 等. 南四湖流域农田土壤养分空间变异特征研究[J]. 山东农业科学, 2023, 55(9): 111–120.

RONG Fangxu, LÜ Shiquan, HU Shuxiang, et al. Spatial variation characteristics of farmland soil nutrients in the Nansi Lake basin[J]. Shandong Agricultural Sciences, 2023, 55(9): 111–120. (in Chinese)

[39] 窦旭, 史海滨, 李瑞平, 等. 盐渍化土壤剖面盐分与养分分布特征及盐分迁移估算[J]. 农业机械学报, 2022, 53(1): 279–290, 330.

DOU Xu, SHI Haibin, LI Ruiping, et al. Distribution characteristics of salinity and nutrients in salinized soil profile and estimation of salt migration[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(1): 279–290, 330. (in Chinese)

[40] 王琦, 常庆瑞, 落莉莉, 等. 陕西省农田土壤有机质时空变异与驱动因子定量研究[J]. 农业机械学报, 2022, 53(11): 349–359.

WANG Qi, CHANG Qingrui, LUO Lili, et al. Spatiotemporal variation and driving factors for cultivated soil organic matter in Shaanxi Province[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(11): 349–359. (in Chinese)

[41] 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. (in Chinese)
- [42] 陈桂香, 高灯州, 曾从盛, 等. 福州市农田土壤养分空间变异特征[J]. *地球信息科学学报*, 2017, 19(2): 216–224. CHEN Guixiang, GAO Dengzhou, ZENG Congsheng, et al. Characteristics of the spatial variation of soil nutrients in farmland of Fuzhou City[J]. *Journal of Geo-information Science*, 2017, 19(2): 216–224. (in Chinese)
- [43] 苟照君, 李英年, 刘峰贵, 等. 黄河上游高寒草地土壤全氮含量分布特征及其影响因素[J]. *生态学杂志*, 2019, 38(9): 2585–2594. GOU Zhaojun, LI Yingnian, LIU Fenggui, et al. Distribution and influencing factors of soil total nitrogen in alpine grassland of upper Yellow River[J]. *Chinese Journal of Ecology*, 2019, 38(9): 2585–2594. (in Chinese)
- [44] 尉芳, 刘京, 夏利恒, 等. 陕西渭北旱塬区农田土壤有机质空间预测方法[J]. *环境科学*, 2022, 43(2): 1097–1107. WEI Fang, LIU Jing, XIA Liheng, et al. Spatial prediction method of farmland soil organic matter in Weibei dryland of Shaanxi Province[J]. *Environmental Science*, 2022, 43(2): 1097–1107. (in Chinese)
- [45] WEN Dong, WU Tianjun, LUO Jiancheng, et al. Land parcel-based digital soil mapping of soil nutrient properties in an alluvial-diluvia plain agricultural area in China[J]. *Geoderma*, 2019, 340: 234–248.
- [46] 崔思远, 曹光乔, 朱新开. 耕作方式对稻麦轮作区土壤碳氮储量与层化率的影响[J]. *农业机械学报*, 2018, 49(11): 275–282. CUI Siyuan, CAO Guangqiao, ZHU Xinkai. Effects of tillage on stocks and stratification of soil carbon and nitrogen in rice-wheat system[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2018, 49(11): 275–282. (in Chinese)
- [47] LUO Chong, ZHANG Xinle, WANG Yihao, et al. Regional soil organic matter mapping models based on the optimal time window, feature selection algorithm and Google Earth Engine[J]. *Soil & Tillage Research*, 2022, 219: 105325.
- [48] 唐海涛, 孟祥添, 苏循新, 等. 基于 CARS 算法的不同类型土壤有机质高光谱预测[J]. *农业工程学报*, 2021, 37(2): 105–113. TANG Haitao, MENG Xiangtian, SU Xunxin, et al. Hyperspectral prediction on soil organic matter of different types using CARS algorithm[J]. *Transactions of the CSAE*, 2021, 37(2): 105–113. (in Chinese)
- [49] 张锡煜, 李思佳, 王翔, 等. 基于 Sentinel-2 卫星影像绥化市土壤全氮定量遥感反演[J/OL]. *农业工程学报*, <http://kns.cnki.net/kcms/detail/11.2047.S.20230908.1430.046.html>.
- [50] 赖雨晴, 孙孝林, 王会利. 人工神经网络及其与地统计的混合模型在小面积丘陵区土壤有机碳预测制图上的应用研究[J]. *土壤通报*, 2020, 51(6): 1313–1322. LAI Yuqing, SUN Xiaolin, WANG Huili. Mapping of soil organic carbon using neural network and its mixed model with geostatistics in a small area of typical hilly region[J]. *Chinese Journal of Soil Science*, 2020, 51(6): 1313–1322. (in Chinese)
- [51] 胡贵贵, 杨粉莉, 杨联安, 等. 基于主成分和机器学习的土壤有机质含量空间预测建模[J]. *干旱区地理*, 2021, 44(4): 1114–1124. HU Guigui, YANG Fenli, YANG Lian'an, et al. Spatial prediction modeling of soil organic matter content based on principal components and machine learning[J]. *Arid Land Geography*, 2021, 44(4): 1114–1124. (in Chinese)
- [52] ZHANG Meiwei, ZHANG Meinan, YANG Haoxuan, et al. Mapping regional soil organic matter based on Sentinel-2A and MODIS imagery using machine learning algorithms and Google Earth Engine[J]. *Remote Sensing*, 2021, 13(15): 2934.
- [53] 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 Longhui, LIU Feng, ZHAO Xia, et al. Digital mmapping of topsoil attributes in Qinghai Province[J]. *Chinese Journal of Soil Science*, 2019, 50(3): 505–513. (in Chinese)
- [54] 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 Jiaping, ZHANG Liming, QIU Longxia, 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. (in Chinese)
- [55] 凌晓丹, 王罗其, 赵科理, 等. 基于随机森林法的山核桃林地土壤速效养分含量空间分布特征研究[J/OL]. *生态学报*, <https://doi.org/10.20103/j.stxb.202301130090>.