DOI: 10.13758/j.cnki.tr.2024.06.023

王桐, 宋洁, 王鑫, 等. 基于特征变量多重扩增与筛选的区域土壤容重随机森林预测. 土壤, 2024, 56(6): 1347–1357.

基于特征变量多重扩增与筛选的区域土壤容重随机森林预测①

王 桐[1,2], 宋 洁[1,2], 王 鑫[1,2], 于东升[1,2*], 马利霞[1,2], 樊剑波[3], 刘 明[1,3]

(1 土壤与农业可持续发展重点实验室(中国科学院), 南京 211135; 2 中国科学院大学, 北京 100049; 3 鹰潭农田生态系统国家野外科学观测研究站(中国科学院南京土壤研究所), 江西鹰潭 335000)

摘 要: 以江西省鹰潭市为研究区, 调查不同土地利用及土壤类型的 131 个样点表层(0 ~ 20 cm)土壤容重, 结合多源环境大数据提取地形、遥感和气候等环境因子的 66 个变量构成原始特征集, 创建特征变量多重扩增与筛选方法, 即针对原始特征集及随机森林(RF)模型, 依次开展基于主成分分析(PCA)的主成分提取–主成分扩增–交叉验证递归(RFECV)筛选–特征多项式扩增(PFE) –交叉验证递归(RFECV)再次筛选, 最终获得了 3 个变量组合的最优特征集。基于最优特征集的 RF 土壤容重空间预测精度 R^2 达 0.469, 比原始特征集的预测精度(R^2=0.315) 提升了 34%, 且特征维度降低了 95%, 显著提升了空间预测效果及效率。

关键词: 土壤容重; 随机森林模型; 特征集; 数字土壤制图

中图分类号: S159.9 **文献标志码:** A

Combining Multiple Feature Expansion and Screening for Predicting Regional Distribution of Soil Bulk Density in Random Forest Algorithm

WANG Tong[1,2], SONG Jie[1,2], WANG Xin[1,2] ,YU Dongsheng[1,2*], MA Lixia[1,2], FAN Jianbo[3], LIU Ming[1,3]

(1 *Key Laboratory of Soil and Sustainable Agriculture*, *Chinese Academy of Sciences*, *Nanjing* 211135, *China*; 2 *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*; 3 *Farmland Ecosystem National Field Observation and Research Station* (*Institute of Soil Science*, *Chinese Academy of Sciences*), *Yingtan*, *Jiangxi* 335000, *China*)

Abstract: Taking Yingtan City, Jiangxi Province as the study area, the topsoil (0–20cm) bulk densities in 131 sampling sites under different land uses and soil types were investigated, 66 environmental factors such as topography, remote sensing, climate and so on were extracted to form the original feature set by combining the multi-source environmental big data. And a method of multiple amplification and screening of the feature variables was created, that was, for the original feature set and random forest (RF) model, PCA component extraction - component augmentation - recursive feature elimination with cross validation (RFECV) - polynomial feature expansion (PFE) - recursive feature elimination with cross validation (RFECV) re-screening being carried out in order. In the end, the optimal set of features for the combination of 3 variables was obtained. The spatial prediction accuracy of RF for soil bulk density based on the optimal feature set reached R^2 of 0.469, which was 34% higher than that of the original feature set (R^2=0.315), and the feature dimensionality was reduced by 95%, which significantly improved the spatial prediction effect and efficiency.

Key words: Soil bulk density; Random forest model; Feature set; Digital soil mapping

土壤容重指在自然状态下单位体积原状土壤的烘干质量, 表征土壤的密实程度[1]。土壤容重空间分布是土壤性状改良、土壤养分储量及污染物环境容量等研究的重要基础数据, 对耕地地力提升、土壤养分管理及环境治理具有重要的意义[2], 但其空间分布受多重结构因素及随机因素的影响[3], 预测精度严重受限。

土壤容重空间预测以往多采用基于样点监测数据的地统计方法[4]。其中, 普通克里格方法是基于给定样本得到区域化变量的结构信息, 联系待推测点有限邻域内的样本数据, 对待推测点进行的无偏最优估计, 并给出推测方差[5-6]。此外, 为进一步提升空间预测精度, 也有方法充分考虑了土壤容重与其他属性

①基金项目: 国家重点研发计划专项(2021YFC1809104, 2022YFB3903302)和国家农业重大科研项目(NK2022180104)资助。

* 通讯作者(dshyu@issas.ac.cn)

作者简介: 王桐(1998—), 男, 山东烟台人, 硕士研究生, 主要从事数字土壤制图研究。E-mail: wangtong@issas.ac.cn

之间的相关性，通过建立交叉协方差函数进行局部估计，如协同克里格方法[7]。但地统计方法受样点均一性及样点密度影响较大[8]，仅考虑各土壤属性之间线性关系，权重强弱取舍存在缺陷[9]，空间变异性是否符合内蕴假设尚不清楚[10]，且预测精度仍需提升[11]。

目前，机器学习因具有优秀的非线性关系捕捉及高维特征处理能力，成为土壤容重空间预测的热门方法[12]。李民赞等[13]基于集成模型，利用土壤属性纹理特征对土壤容重进行实时预测，得到的容重平均绝对误差(Mean average error, MAE)为 0.04 g/cm³，满足田块尺度上精准、快速空间预测要求。在区域尺度上，卢宏亮等[14]基于随机森林(Random forest, RF)算法对安徽省土壤容重的预测精度 R^2 为 0.22，狄晓双[15]基于 BP 神经网络对新疆草地土壤容重的预测精度 R^2 为 0.59，表明机器学习方法对土壤容重空间预测行之有效，但因研究区域及选取环境特征指标差异，导致预测精度存在明显差异。Hateffard 等[16]还对比了 RF、多元线性回归(Multiple linear regression, MLR)、支持向量机(Support vector machine, SVM)、人工神经网络(Articicial neural network, ANN)4 种方法对土壤容重的预测效果，表明 RF 预测方法效果最好，特别在样点数量严重受限条件下。Hengl 等[17]基于公开数据库对非洲土壤属性进行空间预测时发现，RF 比 MLR 方法具有更高预测精度，而 RF 建模成功关键在于输入特征变量的优选，因为一些看似相关的变量对模型预测精度会产生负面影响[18]。

特征变量筛选方法常分为过滤、封装、嵌入和集合 4 类[19]。其中，过滤法和集合法，常依据特征变量对目标变量的影响程度进行筛选，如灰色关联理论[20]、相关性选择[21]、重要性评价[22-23]等。封装法，如递归特征消除法 (Recursive feature elimination, RFE)，是一种最大间隔原理的序列后向选择算法，每次剔除一个排序准则分数最小变量，直到获得符合规则要求的特征变量集[24]。张柄华等[25]使用 RFE 算法对多源遥感影像特征进行筛选，特征数量减少 34.5%，提升运算效率的同时也使藏东南土地覆被 RF 分类整体精度略有提升。常见的特征变量变换嵌入方法，如通过主成分分析(Principal component analysis, PCA)获得新的正交特征[26]、一阶微分变换去除噪声[27]等。但目前将特征筛选与变换嵌入相结合，关注特征之间非线性信息提取的研究极少。姚凯丰等[28]则使用多项式特征扩增方法(Polynomial feature expansion, PFE)对特征变量进行变换，根据 SVM 的留一错误率，剔除权重较小的特征，有效提高了分类器泛化能力，使四川观音场油气井分类预测误差降低 50%。但这种特征变换扩增与筛选相结合的方法，对土壤容重空间预测精度提升是否行之有效，如何以此进一步提升其预测精度，有待进一步研究。

为此，本研究以江西省鹰潭市为研究区，基于样点表层土壤容重调查数据，通过提取影响土壤容重的 4 类环境因子，包括土地利用[29]、地形因子[30]、气候因子[31]和遥感光谱及指数[32]，尝试创建并运用特征变量多重扩增与筛选方法，构建最优特征集，建立 RF 高精度空间预测模型，揭示研究区土壤容重的空间分布特征，以为充分发挥机器学习方法处理大数据优势[33]、提升数字土壤制图精度及质量提供新思路。

1 材料与方法

1.1 研究区概况

鹰潭市(116°41′E ~ 117°30′E, 27°35′N ~ 28°41′N)位于江西省东北部，信江中下游，地处武夷山脉与鄱阳湖平原过渡的交接地带，北部连接怀玉山脉，南部连接武夷山脉，中部为信江盆地。区域内最高点海拔 1 483 m，最低点 –84 m，平均 40 m。气候类型为亚热带湿润季风气候，年均温 18 ℃，年均降水量 1 750 mm，全年无霜期 262 d。鹰潭市现有余江区、月湖区和贵溪市 3 个辖区，总面积约 3 520 km²。土地利用类型包括耕地、林地、园地、草地、工矿仓储用地、住宅用地、水域及水利设施用地、交通用地及其他土地。其中，林地面积约占 54.6%，主要分布在鹰潭市南北两端；耕地面积约占 25.6%，主要分布于中部信江盆地，沿信江向南北方向辐射，是江西省重要商品粮基地。市内矿产资源丰富，也是全国最大铜冶炼、铜加工基地及重要的铜消费区[34]。

1.2 样点布设及容重数据获取

2021 年 12 月，考虑土地利用及土壤类型空间分布特征基础上，布设 131 个调查样点(图 1)。其中，耕地、林地、草地、工矿仓储用地、园地的样点数分别约占 48.1%、33.6%、7.6%、4.6%、2.3%，其他土地包括裸土地、空闲地和设施农用地，样点数共占 3.8%。每个采样点使用环刀法采集 0 ~ 20 cm 土壤容重平行样品 3 个，以均值作为该样点土壤容重数值。采用环刀法测定土壤容重[35]。

1.3 环境协变量及数据来源

以 2020 年 9 月至 2021 年 6 月空间分辨率为 10 m、云量小于 0.05% 的 Sentinel-2 Level-2A 遥感影像作为环境协变量数据源(https://sentinel.esa.int/web/

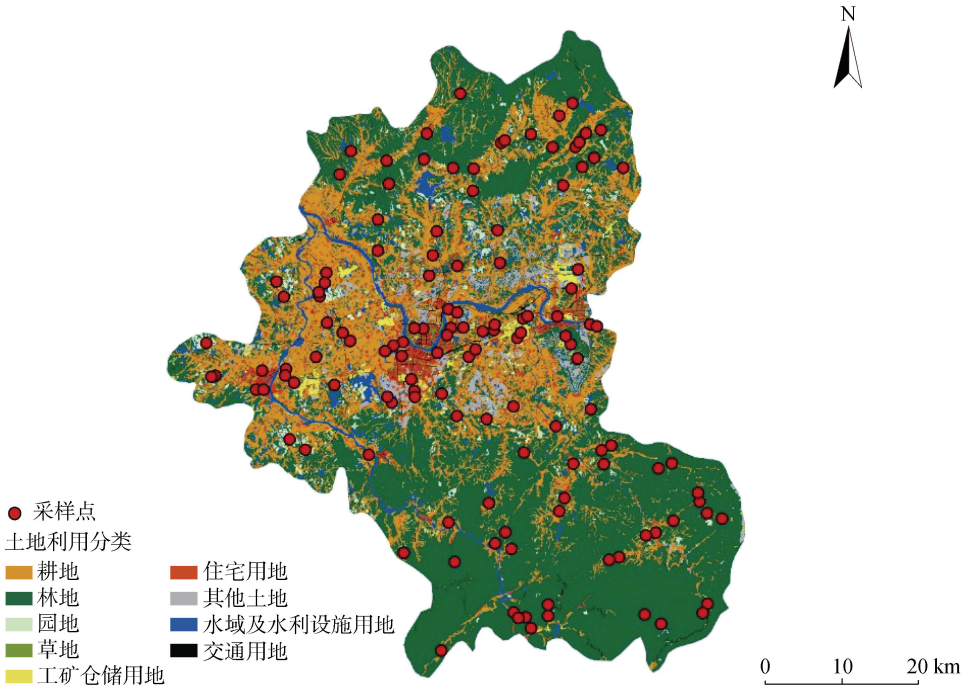


图 1　研究区土地利用与土壤样点分布

Fig. 1　Distribution of land use types and soil sample sites in study area

sentinel/user-guides/sentinel-2-msi/processing-levels/level-2)。

首先，参照 GB/T 21010—2017《土地利用现状分类》标准[36]进行目视解译，获得土地利用类型矢量数据。利用 ArcGIS 空间连接功能获得样点处的土地利用类型及邻接指定区域面积变量，指定区域包括住宅用地、工矿仓储用地和空闲地；采用近邻分析工具获得样点距建筑、城市建筑、农村建筑距离和距河流、公路距离共 5 个距离变量。

其次，利用 Google Earth Engine(GEE)工具提取单波段数据(B2、B3、B4、B5、B6、B7、B8A、B8、B11、B12)和 15 种遥感指数数据，包括亮度指数(BI)、颜色指数(CI)、差值植被指数(DVI)、绿光归一化差值植被指数(GNDV)、红外植被百分比指数(IPVI)、改进红边叶绿素指数(IRECI)、修正的叶绿素吸收反射指数(MCAR)、陆地叶绿素指数(MTCI)、归一化差异指数(NDI)、植物衰老反射率指数(PSSR)、红边位置指数(REIPI)、红光指数(RI)、比值植被指数(RVI)、土壤背景指数(SBI)、转换归一化植被指数(TNDVI)；再利用 SNAP 9.0.0 工具计算 10 种纹理指数，包括最大概率指数(MAX)、角二阶矩指数(ASM)、对比度指数(Contrast)、不相似性指数(Dissimilar)、能量指数(Energy)、熵指数(Entropy)、相关性指数(GLCMCorrel)、均值指数(GLCMMean)、方差指数(GLCMVarian)、均匀性指数(Homogeneit)。

同时，以采样前后 3 个月内同时相 Sentinel-1 SAR GRD 为数据源(https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/)，利用 GEE 工具提取后向散射系数(VV 和 HV)；以同时相 USGS Landsat 8 level 2、collection2、Tier1 为数据源(https://www. usgs.gov/landsat-missions/landsat-collection-2-level-2-science-products)，提取最大归一化植被指数(maxNDVI)、最大增强植被指数(maxEVI)和土壤调节植被指数(MSAVI)以及数字高程模型(DEM)；进而利用 SAGA GIS 8.3.0 计算 14 种地形因子，如坡度(Slope)、坡向(Aspect)、坡长(SL)、平面曲率(Plan_cur)、剖面曲率(Prof_cur)、地形位置指数(TPI)、地形湿度指数(TWI)、地表粗糙指数(TRI)、谷深(VallyDepth)、流域面积(CatchmentA)、山体阴影(Hillshader)、水流强度指数(SPI)、多尺度山谷平坦指数(MRVBF)、多尺度山顶平坦指数(MRRTF)。

另外，气候因子来自国家地球系统科学数据中心(http://www.geodata.cn)，利用 ArcGIS 工具提取 2015—2020 年年均温(MAT)和年均降水量(MAP)数据。

最后，利用 ArcGIS 10.8 将所有环境协变量图层投影转换至同一大地坐标系，并以 30 m 栅格分辨率进行数据重采样。

1.4　特征变量多重扩增与筛选路径

首先，将提取的 66 个原始环境协变量标记为特征集 S1；其次，利用 SPSS 26.0 软件对特征集 S1 进

行 PCA 分析[37]，以特征值大于 1 为标准，提取主成分特征，标记为特征集 S2，并将其与 S1 的组合标记为特征集 S3；接着，对特征集 S3(79 个特征变量)进行 Pearson 相关分析[38]，根据显著性值 P<0.05，筛选出高相关性特征变量，标记为特征集 S4。

交叉验证递归筛选(Recursive feature elimination with cross-validation，RFECV)是一种结合递归特征消除和交叉验证思想、自动选择特征变量的 RFE 方法，可给出体现模型性能最佳的特征组合和数量[39]。RFECV 通常利用指定模型(如 SVM、RF)对当前特征子集进行变量重要性评价，同时利用指定方法(如十折交叉验证、留一交叉验证)进行模型性能评价，删除当前特征子集中最不重要特征变量，从而获得新特征子集；通过循环不断评价、选择，直到获得最优特征子集，使得模型性能达到最优。本研究使用 Python 中的 RFECV 函数[40]对特征集 S3 进行首次 RFECV 筛选，并获得特征集 S5。

多项式特征扩增(Polynomial feature expansion，PFE)是一种对数据进行高阶组合的方法，通过对特征进行高阶多项式映射扩增特征变量，从而捕捉特征之间更多非线性关系，例如二维特征$(X_1, X_2)^T$的二阶多项式扩展为$\left(1, X_1, X_2, X_1X_2, X_1^2, X_2^2\right)^T$，三阶多项式扩展为$\left(1, X_1, X_2, X_1X_2, X_1^2, X_2^2, X_1^3, X_2^3, X_1^2X_2, X_1X_2^2\right)^T$。本研究使用 Python 中的 PolynomialFeatures 函数，对特征集 S5 进行 PFE 特征扩增。通过尝试不同阶数多项式，比较分析不同阶数下的模型精度变化，对最优阶数下的特征进行扩增，获得新特征集 S6。

最后，对特征集 S6 再次进行 RFECV 筛选，获得最终参与建模的最优特征变量集合，标记为 S7。

特征变量多重扩增与筛选路径如图 2 所示。

1.5 随机森林预测模型

随机森林模型(Random forest, RF)是由 Breiman[41]提出的基于决策树的集成学习技术，每棵决策树代表一个分类器，通过对多个分类器的输出结果进行投票或取平均值，实现分类或回归预测。RF 实现回归预测步骤为：①采用重采样(bootstrap)有放回的方法选择 n 个样本作为训练集，构建 n 棵回归树(ntree)，每次未被选中的袋外样本(OOB)构成测试集，用于验证模型的泛化能力；②从特征变量中抽取 m 个解释变量(mtry)，按照 OOB 预测误差最小原则确定 mtry 的值，由此决定每棵决策树使用的特征数量；③将所有决策树的预测值取平均，得到 RF 最终回归预测结果。本研究使用 Python 中的 RandomizedSearch 与 GridSearchCV 函数确定 RF 参数[42]。

1.6 模型验证与精度评价

使用 Python 中留一交叉验证(LeaveOneOut)函数对模型精度进行评价。已知原始数据有 N(N=131)个样点，选择(N–1)个样点作为训练样本，将每个样点视为独立的验证样本，经过 N 次模型运行后获得 N 个预测结果，对应于样本真实值，计算均方根误差(RMSE)及决定系数(R^2)，RMSE 越小、R^2 越接近 1，表明预测精度越高[43]。

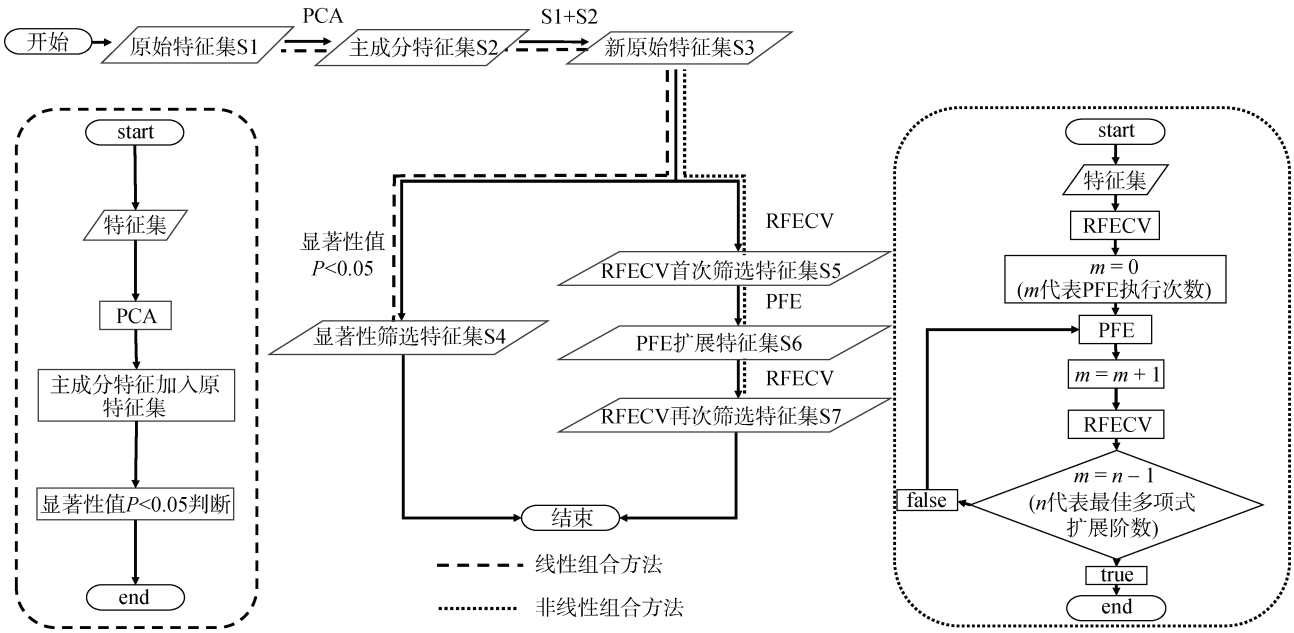


图 2 特征变量多重扩增与筛选路径

Fig. 2 Roadmap for feature expanding and screening

1.7　土壤容重空间分布制图

使用 Python 中 gdal 函数结合上述 RF 建模过程，对研究区土壤容重空间分布进行回归预测制图。为比较和展现基于本文特征变量多重扩展与筛选的 RF 方法预测优势，利用 ArcGIS 10.8 中 Geostatistical Wizard 模块和最优变量集，进行土壤容重协同克里格插值(COK) [44]制图。空间分辨率均为 30 m×30 m。

2　结果与讨论

2.1　土壤容重样点统计特征

研究区样点土壤容重变化范围为 0.67 ~ 1.72 g/cm^3，平均值为 1.22 g/cm^3，标准差为 0.22 g/cm^3，变异系数为 0.18，数值相对集中且变化范围较小，数据相对离散程度较低；偏度为 0.32，数据分布向右偏，峰度为 –0.51，数据分布有平坦峰度(表 1)。

不同土地利用类型中，人为活动频繁的工矿仓储用地土壤容重平均值最高为 1.45 g/cm^3，原因是该类土地存在改造、挖掘填埋等情况，特别是机械碾压导致土壤紧实，通气透水性减弱，土壤容重偏高[45]；耕地土壤容重平均值最低，为 1.14 g/cm^3，这可能是翻耕等农业活动减轻土壤压实，增加土壤通气性，同时施肥和秸秆还田等增加土壤有机质，有利于土壤团粒结构形成，改善土壤质地，增加土壤孔隙度。

2.2　多重扩增与筛选特征集变化特征

2.2.1　主成分变量扩增　　原始特征变量集 S1 的 PCA 分析表明，前 13 个主成分变量总解释方差达 84.43%，而前 14 个主成分总解释方差达 85.92%(表 2)，但第 14 个主成分起始特征值小于 1，该因子包含信息不足以证明其应该保留[46]。主成分碎石图(图 3)显示，当提取的主成分为 13 个时，折线由陡峭变得平稳。经综合考虑，本研究提取 13 个主成分变量，分别标记为 PCA1 ~ PCA13，在此首次实现线性过程的变量扩增。

表 1　不同土地利用类型土壤容重基本统计特征

Table 1　Statistical characteristics of soil bulk densities under different land use types

样本	样本数	最小值(g/cm^3)	最大值(g/cm^3)	均值(g/cm^3)	标准差(g/cm^3)	变异系数	偏度	峰度
总样本	131	0.67	1.72	1.22	0.22	0.18	0.32	–0.51
耕地	63	0.86	1.69	1.14	0.20	0.18	0.94	0.67
林地	44	0.67	1.64	1.24	0.22	0.18	–0.05	–0.31
草地	10	1.18	1.65	1.38	0.17	0.12	0.59	–1.26
工矿仓储用地	6	1.17	1.72	1.45	0.22	0.19	–0.15	–1.75
其他土地	5	1.13	1.58	1.33b	0.07	0.05	1.59	0.76
园地	3	1.34	1.54	1.40	0.10	0.07	1.29	–

表 2　特征变量总方差解释表

Table 2　Explanatory table for total variance of characteristic variables

成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积方差百分比(%)	总计	方差百分比	累积方差百分比(%)
1	20.95	31.74	31.74	20.95	31.74	31.74
2	8.02	12.15	43.88	8.02	12.15	43.88
3	5.58	8.45	52.33	5.58	8.45	52.33
4	5.22	7.91	60.24	5.22	7.91	60.24
5	2.80	4.25	64.49	2.80	4.25	64.49
6	2.54	3.85	68.34	2.54	3.85	68.34
7	2.04	3.08	71.42	2.04	3.08	71.42
8	1.96	2.98	74.40	1.96	2.98	74.40
9	1.58	2.40	76.80	1.58	2.40	76.80
10	1.46	2.21	79.01	1.46	2.21	79.01
11	1.31	1.99	81.00	1.31	1.99	81.00
12	1.19	1.81	82.81	1.19	1.81	82.81
13	1.07	1.62	84.43	1.07	1.62	84.43
14	0.99	1.49	85.92			
…	…	…	…			
66	8.42×10^{-17}	1.28×10^{-16}	100.00			

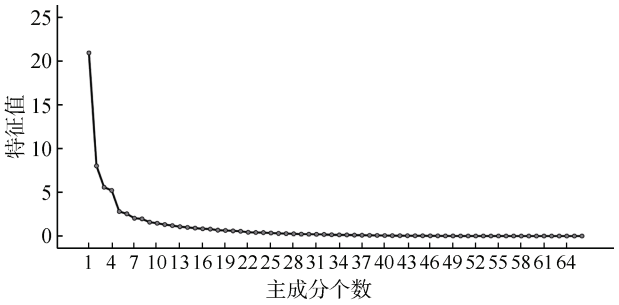


图 3 环境协变量主成分碎石图

Fig. 3 Principal components crushed stone figure of environmental covariates

包含 66 个原始特征的变量集 S1、包含 13 个 PCA 新增特征的变量集 S2 及其组合包含 79 个特征的变量集 S3 (PCA 扩增特征集)，分别参与 RF 土壤容重预测的精度 R^2 为 0.351、0.112 和 0.354，基于特征集 S3 的 RF 预测精度仅略有提高。

与原始特征变量集 S1 相比，基于 PCA 新增特征集 S2 的预测精度降低，原因在于 PCA 方法通过线性组合提取能够最大程度保留原始数据方差的新特征，将原始特征映射到主成分空间会导致信息损失，某些与目标变量相关但方差较小的特征被忽略[47]。而 PCA 扩增特征集 S3 维度比特征集 S1 增加约 19.7%，引入了与目标变量相关信息，预测精度提高约 0.85%。

2.2.2 线性筛选 通过新原始特征集(PCA 扩增特征集，S3)变量与土壤容重之间的 Pearson 线性相关性分析 (图 4)，对特征集 S3 进行降维，其中显著性 $P<0.05$ 的变量有 RVI、PSSR、经度、MRRTF、maxNDVI、SPI、PCA6、PCA7、距离建筑物的距离、SL、PCA4、REIPI、距离城市建筑的距离、Dem、邻接区域面积和 MAT 共 16 个变量，由此，获得新特征集 S4。基于特征集 S4，RF 对土壤容重的预测精度 R^2 为 0.344，虽然较特征集 S3 减少了约 79.7% 输入变量，但预测精度 R^2 却低于特征集 S3 (R^2=0.351)，降低约 2.8%，与提升空间预测精度目标背道而驰。

2.2.3 首次非线性筛选 在指定 RF 模型和留一交叉验证方法前提下，RFECV 方法从特征集 S4 中初步筛选出的新特征集 S5 包含 4 个特征变量，分别为 REIPI、MAT、NDI 和经度，RF 对土壤容重的预测精度 R^2 为 0.435，与先前特征集(S1、S2、S3、S4)比较，预测精度均得到明显提升(表 3)。特征集 S5 代表了 RFECV 首次筛选结果，对比特征集 S3，特征维度降低约 94.9%，预测精度提高约 22.9%。与线性扩增和线性筛选相比，RFECV 方法在 RF 模型下在特征降维和精度提升两方面均有明显优势[48]。

2.2.4 多项式扩增 在特征集 S5 基础上，尝试 1 ~ 6 阶扩增的 RF 预测结果显示，基于 2 阶的 PEF 扩增特征集 S6 的 RF 预测精度 R^2 最高，为 0.442(表 4)，表明组合迭代 1 次就能够使原始特征变换到最佳状态，且比先前所有特征集的预测精度又有了明显提升(表 3)。但 PFE 通过原始特征间相乘和幂运算提高了特征维度，引入了更多非线性关系，特征维度提升约为特征集 S5 的 4 倍，这会导致模型计算开销的增高[49]，对 RFECV 与 PFE 组合迭代次数也有重要影响，因此需选择最佳多项式阶数。

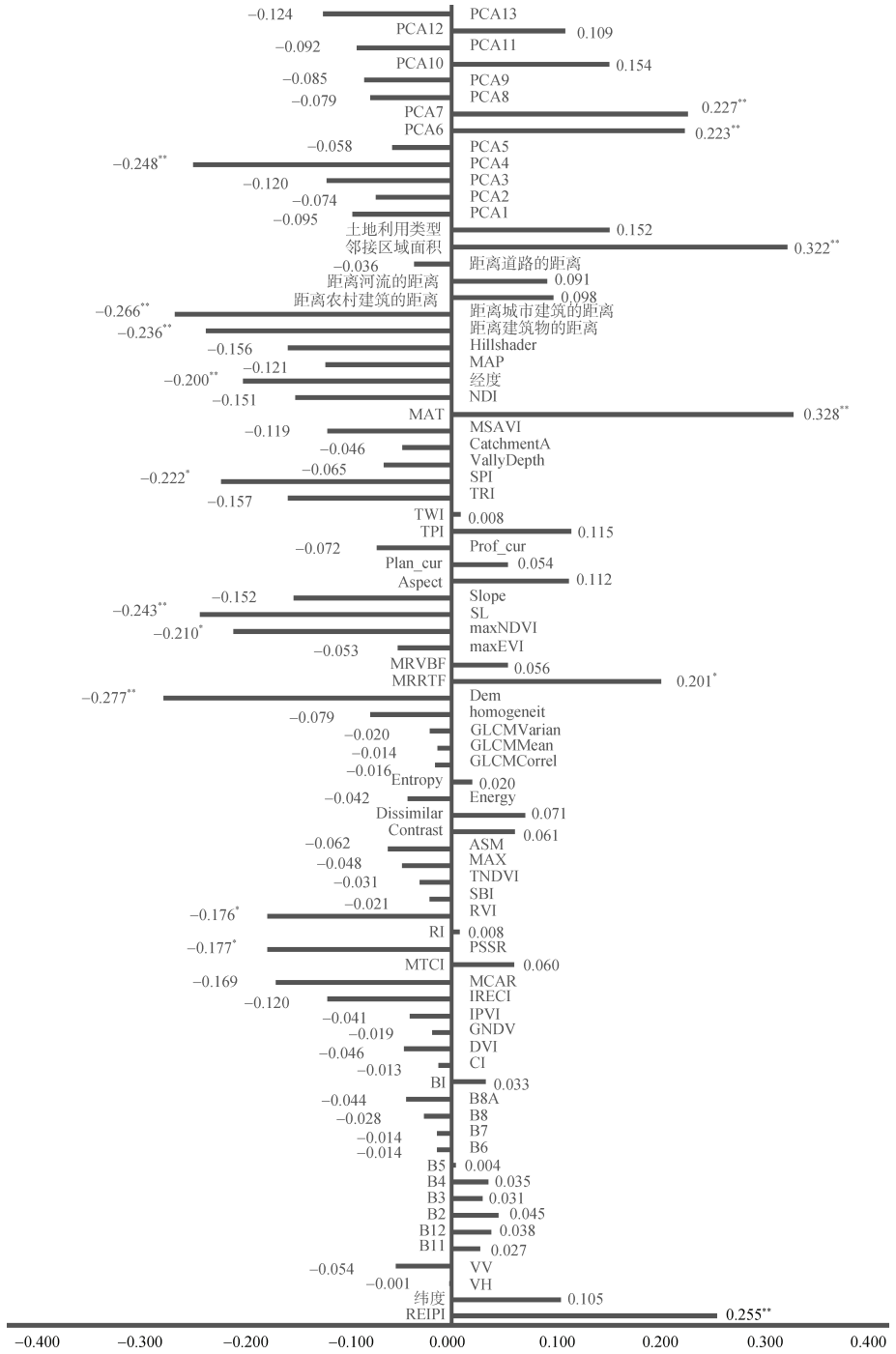
2.2.5 再次非线性筛选 在 2 阶 PFE 扩展特征集 S6 的基础上再进行 RFECV 筛选，筛选出 3 个变量构成新的特征集 S7，包括 REIPI、REIPI 与 MAT 的乘积、NDI 与经度的乘积，其参与 RF 预测的精度 R^2 最高，为 0.469，超出了先前所有特征集的预测精度(表 3)。S6 特征集经 RFECV 再次筛选得出的特征集 S7，特征维度降低 80.0%，不仅预测精度提高约 6.1%，也抑制了特征多项式扩增后的计算成本上升。

总之，通过本研究创建的特征变量多重扩增与筛选路径，对比 S7 和 S3 特征集，经非线性扩增和筛选使原始特征维度降低约 96.2%，土壤容重 RF 空间预测精度提高约 32.5%，达到预设研究目标。

2.3 土壤容重空间预测分布特征

将通过多重扩增与筛选方法获得的特征集 S7 变量(REIPI、REIPI 与 MAT 的乘积、NDI 与经度的乘积)分别输入 RF 模型和 COK 模型进行土壤容重的空间预测，结果表明，两种方法所展示的土壤容重空间变化趋势具有相似性，即由南北山地到中部盆地土壤容重逐渐升高(图 5)，与样点数据空间分布特征一致，原因在于中部地区集中有住宅用地和工矿仓储用地，土壤容重高值点相对集中。此外，邻接域面积代表人为扰动范围和强度，与土壤容重的相关系数为 0.322**，有显著正相关性(图 4)，也间接验证了研究区中部余江、月湖和贵溪市区一带土壤容重较高的现象。

COK 预测精度 R^2 为 0.139，远低于 RF 模型的预测精度(R^2=0.469)；COK 模型预测值和预测区间均偏小，预测区间仅为 0.80 ~ 1.34 g/cm^3，而 RF 模型预测区间拓宽为 0.86 ~ 1.57 g/cm^3；土壤容重空间预测分布图呈现的微域变异特征，RF 模型比 COK 模型预测效果更为精细。这表明应用多重扩增与筛选方法也需指定模型(如 RF)与验证方法(如留一交叉验证)对特征集变量重要性与模型性能进行评价，从而获得与指定模型相对应的最优特征集。基于 RF 模型进行的空间预测比基于 COK 模型的预测更具优势，是因为 RF 组合方法重在数据挖掘，充分考虑环境变量间的



(REIPI，红边位置指数；VH、VV，后向散射系数；B2、B3、B4、B5、B6、B7、B8、B8A、B11、B12，Sentinel-2 的 B2、B3、B4、B5、B6、B7、B8、B8A、B11、B12 单波段；BI，亮度指数；CI，颜色指数；DVI，差值植被指数；GNDV，绿光归一化差值植被指数；IPVI，红外植被百分比指数；IRECI，改进红边叶绿素指数；MCAR，修正的叶绿素吸收反射指数；MTCI，陆地叶绿素指数；PSSR，颜料特定的简单比率；RI，红光指数；RVI，比值植被指数；SBI，土壤背景指数；TNDVI 转换归一化植被指数；MAX，最大概率指数；ASM，角二阶矩指数；Contrast，对比指数；Dissimilar，不相似性指数；Energy，能量指数；Entropy，熵指数；GLCMCorrel，灰度共生矩阵的相关性指数；GLCMMean，灰度共生矩阵的均值指数；GLCMVarian，灰度共生矩阵的方差指数；homogeneit，均匀性指数；Dem，数字高程模型；MRRTF，多尺度山顶平坦指数；MRVBF，多尺度山谷平坦指数；maxEVI，最大增强植被指数；maxNDVI，最大归一化植被指数；SL 坡长；Slope，坡度；Aspect，坡向；Plan_cur，平面曲率；Prof_cur，剖面曲率；TPI，地形位置指数；TWI，地形湿度指数；TRI，地表粗糙指数；SPI，水流强度指数；Vally Depth，谷深；Catchment A，流域面积；MSAVI，土壤调节植被指数；MAT，年均温；NDI，归一化差异指数；MAP，年降水量；Hillshader，山体阴影；PCA1 ~ PCA13，主成分 1 到主成分 13。*、**分别表示相关性在 $P<0.05$、$P<0.01$ 水平显著)

图 4　环境协变量与土壤容重的相关性分析

Fig. 4　Correlation analysis of environmental covariates with soil bulk density

表 3 不同特征集的来源和 RF 预测精度对比

Table 3 Comparison of different feature sets and their RF prediction accuracies

特征集	特征来源	特征数量	MAE(g/cm^3)	RMSE(g/cm^3)	R^2
S1	原始环境协变量	66	0.142	0.179	0.351
S2	PCA 线性扩增 S1	13	0.166	0.209	0.112
S3	S1+S2	79	0.142	0.178	0.354
S4	P<0.05 线性筛选 S3	16	0.141	0.180	0.344
S5	RFECV 首次非线性筛选 S3	4	0.130	0.167	0.435
S6	PEF 非线性扩增 S5	15	0.129	0.166	0.442
S7	RFECV 再次非线性筛选 S6	3	0.123	0.162	0.469

表 4 多项式不同阶数特征扩增集的预测精度

Table 4 Prediction accuracies of polynomial feature expansion of different orders attempted

误差	多项式阶数					
	1	2	3	4	5	6
MAE(g/cm^3)	0.130	0.130	0.130	0.129	0.131	0.131
RMSE(g/cm^3)	0.167	0.166	0.167	0.167	0.167	0.166
R^2	0.435	0.442	0.432	0.435	0.436	0.438

非线性关系和交互影响，这再次彰显了机器学习处理大数据的强大能力。

比较各土地利用类型土壤容重的预测结果，发现 RF 模型对林地预测效果最好(表 5、表 1)，但对耕地存在高估，对园地、工矿仓储用地存在低估。一方面 RF 模型为全局最优预测模型，但预测精度 R^2 仅为 0.469，总体拟合精度有限；另一方面，耕地及工矿仓储用地集中分布在研究区中部，受人为活动强烈影响，空间变异性更强，且土地利用类型特征未能选取参与建模，局部拟合精度提升不足。尽管如此，各土地利用类型样点的平均预测值总体偏差均小于 10%(表 5)，预测结果仍可接受。

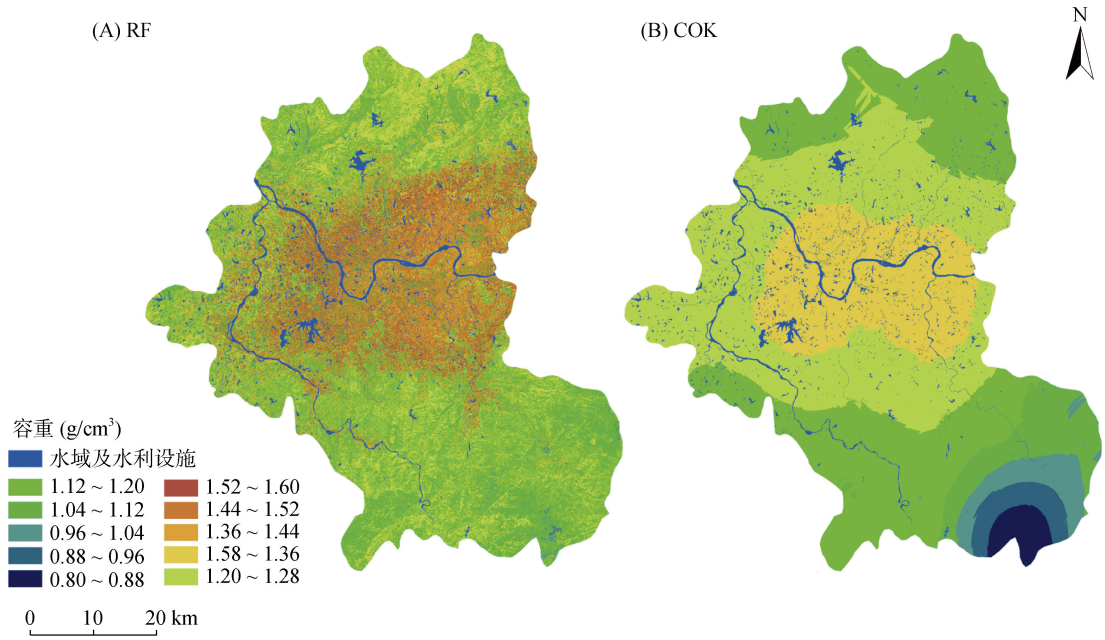


图 5 研究区表层土壤容重空间预测分布

Fig. 5 Results of spatial prediction of top soil bulk density in study area

2.4 最优特征集的合理性解释

基于多重扩增与筛选方法获得的最优特征集 S7，土壤容重 RF 空间预测取得了满意效果，与原始特征集 S1 比较，特征维度降低约 95%，RF 预测精度提升约 34%。这种特征多重扩增与筛选方法与 RF 模型组合，充分利用了非线性组合优势，最大限度在

表 5　不同土地利用类型采样点土壤容重 RF 预测结果

Table 5　RF predictions for soil bulk density of sampling points under different land uses

土地利用类型	平均值(g/cm^3)	最大值(g/cm^3)	最小值(g/cm^3)	标准差(g/cm^3)	变异系数	偏度	峰度	平均值偏差(%)
总样本	1.26	1.56	0.92	0.18	0.14	0.05	–1.17	3.08
耕地	1.25	1.56	0.95	0.04	0.03	0.13	–1.36	9.41
林地	1.24	1.53	0.92	0.14	0.12	0.44	–0.21	0.11
草地	1.31	1.53	0.96	0.21	0.16	–0.70	–1.28	–5.29
工矿仓储用地	1.34	1.47	1.09	0.14	0.11	–1.12	0.67	–7.41
其他土地	1.29	1.45	1.07	0.14	0.11	0.34	–0.42	–2.95
园地	1.32	1.48	1.05	0.24	0.18	–1.67	–	–5.77

降低计算维度的同时提高了预测精度，成为一种经济、高效的特征优化方法。该组合方法重在数据挖掘，考虑环境变量间的非线性关系和交互影响，但变量对土壤容重影响的机理不明确，给特征变量的合理性解释带来了困难，是数据驱动预测的弱点[50]。尽管如此，本文依然尝试从土壤容重空间分布格局角度分析最优特征集 S7 变量(REIPI、REIPI 与 MAT 的乘积、NDI 与经度的乘积)的合理性。

REIPI 为红边位置指数，遥感红边指数是表征绿色植物生长状况的重要生化参数[51]；NDI 为归一化差异指数，其数值变化常用于监测植物生长情况[52]，二者的空间变化都可以反映植被覆盖差异。两个植被指数在林区数值较高，与冬季林区植被覆盖度高于农田和城镇的实际情况相符。RF 预测图中，河流沿岸土壤容重要高于附近山地土壤容重，是因为山地林区覆盖的枯枝落叶分解后增加了土壤有机物含量，良好的植被条件降低了土壤容重[53]。

MAT 为年均气温，该指数的空间分布反映了不同区域的温度差异，样点土壤容重与 MAT 相关系数为 0.328，相关性极显著(P<0.01)。MAT 高导致地区土壤水分蒸发量大，热量累积也会使土壤干燥收缩，进而导致土壤容重增大[54]。

经度代表了土壤容重东西走向的综合空间差异，样点容重与经度的相关系数为 –0.2，相关性显著(P<0.05)。根据相关系数分析，土壤容重预测值东部应该低于西部，但 RF 预测图中并没有明显的横向差异。经度与 NDI 指数的乘积成为提高 RF 预测精度的重要特征，反映二者对土壤容重有着非直观的交互影响，达到了本研究提取多个特征之间非线性信息的目的。

总之，本研究建立的最优特征集 S7，不是通过某个筛选方法能直接获取的，而是基于线性与非线性多种组合方式，通过对原始特征的多重扩增与筛选才能取得的。这种特征变量多重扩增与筛选方法，不仅有效降低了特征维度与运算成本，而且显著提升了机器学习方法(RF)的空间预测精度，为提升数字土壤制图精度及质量增添了新途径。

3　结论

本研究结合 PFE 和 RFECV 方法，基于 RF 模型对多源环境数据进行特征变量多重扩增与筛选，结果表明，该方法能经济、高效地优选出具有最佳预测精度的特征组合，相比传统线性特征筛选方法，进一步提高了研究区内表层土壤容重的预测精度，更充分挖掘了土壤与环境特征之间的非线性关系，为其他土壤属性高精度预测及特征前处理提供了新途径和成功案例。

参考文献：

[1] 邵明安, 王全九, 黄明斌. 土壤物理学[M]. 北京: 高等教育出版社, 2006.

[2] Kosmas C, Gerontidis S, Marathianou M. The effect of land use change on soils and vegetation over various lithological formations on Lesvos (Greece)[J]. CATENA, 2000, 40(1): 51–68.

[3] Li S, Li Q Q, Wang C Q, et al. Spatial variability of soil bulk density and its controlling factors in an agricultural intensive area of Chengdu Plain, Southwest China[J]. Journal of Integrative Agriculture, 2019, 18(2): 290–300.

[4] 巫振富, 赵彦锋, 程道全, 等. 样点数量与空间分布对县域尺度土壤属性空间预测效果的影响[J]. 土壤学报, 2019, 56(6): 1321–1335.

[5] 朱阿兴, 杨琳, 樊乃卿, 等. 数字土壤制图研究综述与展望[J]. 地理科学进展, 2018, 37(1): 66–78.

[6] 潘宗涛, 陈志强, 陈志彪. 朱溪流域土壤容重空间分异与地形和土地利用的关系[J]. 水土保持通报, 2018, 38(3): 263–268, 2.

[7] 石淑芹, 曹祺文, 李正国, 等. 区域尺度土壤养分的协

同克里格与普通克里格估值研究[J]. 干旱区资源与环境, 2014, 28(5): 109–114.
[8] 潘成忠, 上官周平. 土壤空间变异性研究评述[J]. 生态环境, 2003, 12(3): 371–375.
[9] 郭龙, 张海涛, 陈家赢, 等. 基于协同克里格插值和地理加权回归模型的土壤属性空间预测比较[J]. 土壤学报, 2012, 49(5): 1037–1042.
[10] 秦耀东. 土壤空间变异研究中的定量分析[J]. 地球科学进展, 1992, 7(1): 44–49.
[11] 连纲, 郭旭东, 傅伯杰, 等. 基于环境相关法和地统计学的土壤属性空间分布预测[J]. 农业工程学报, 2009, 25(7): 237–242.
[12] 刘付程, 郭衍游, 闫晓波. 土壤属性空间预测的广义回归神经网络方法研究[J]. 淮海工学院学报(自然科学版), 2008, 17(1): 68–71.
[13] 李民赞, 任新建, 杨玮, 等. 基于树莓派的农田表土层土壤容重检测系统研究[J]. 农业机械学报, 2021, 52(S1): 329–335, 376.
[14] 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.
[15] 狄晓双. 新疆主要草地土壤容重预测模型构建[D]. 乌鲁木齐: 新疆农业大学, 2021.
[16] Hateffard F, Szatmári G, Novák T J. Applicability of machine learning models for predicting soil organic carbon content and bulk density under different soil conditions[J]. Soil Science Annual, 2023, 74(1): 165879.
[17] Hengl T, Heuvelink G B M, Kempen B, et al. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions[J]. PLoS One, 2015, 10(6): e0125814.
[18] Archer K J, Kimes R V. Empirical characterization of random forest variable importance measures[J]. Computational Statistics & Data Analysis, 2008, 52(4): 2249–2260.
[19] Chen Y, Ma L X, Yu D S, et al. Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests[J]. Ecological Indicators, 2022, 135: 108545.
[20] 郭李娜, 樊贵盛. 基于灰色理论—BP 神经网络方法的表层土壤容重预测[J]. 节水灌溉, 2018(2): 93–97.
[21] 刘丽媛, 郑向群, 张春雪, 等. 稻田有机肥配施的土壤环境效应评价指标体系构建[J]. 农业资源与环境学报, 2022, 39(1): 129–138.
[22] 张东彦, 杨玉莹, 黄林生, 等. 结合 Sentinel-2 影像和特征优选模型提取大豆种植区[J]. 农业工程学报, 2021, 37(9): 110–119.
[23] 顾永昇, 丁建丽, 韩礼敬, 等. 基于多源环境变量的渭–库绿洲土壤颗粒含量预测研究[J]. 土壤, 2023, 55(2): 426–432.
[24] 林俊, 许露, 刘龙. 基于 SVM-RFE-BPSO 算法的特征选择方法[J]. 小型微型计算机系统, 2015, 36(8): 1865–1868.
[25] 张炳华, 张镱锂, 谷昌军, 等. 基于随机森林与特征选择的藏东南土地覆被分类方法及精度评价[J]. 地理科学, 2023, 43(3): 388–397.
[26] 聂红梅, 杨联安, 李新尧, 等. 基于 PCA-SVR 的冬小麦土壤水分预测[J]. 土壤, 2018, 50(4): 812–818.
[27] 何挺, 王静, 林宗坚, 等. 土壤有机质光谱特征研究[J]. 武汉大学学报(信息科学版), 2006, 31(11): 975–979.
[28] 姚凯丰, 陆文凯, 丁文龙, 等. 一种基于 SVM 特征选择的油气预测方法[J]. 天然气工业, 2004, 24(7): 36–38, 134.
[29] 邱倩倩, 张卓栋, 孙传龙, 等. 锡林郭勒草地景观系统土壤容重空间变异及其与风蚀的关系[J]. 水土保持通报, 2016, 36(6): 58–62, 66.
[30] 冯强, 段宝玲, 姜硕. 小流域尺度土壤容重及其影响因素的空间变异[J]. 山西农业大学学报(自然科学版), 2016, 36(1): 39–45.
[31] Aitkenhead M J, Coull M, Towers W, et al. Prediction of soil characteristics and colour using data from the National Soils Inventory of Scotland[J]. Geoderma, 2013, 200: 99–107.
[32] Pittman R, Hu B. Estimation of soil bulk density and carbon using multi-source remotely sensed data[J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2020, 3: 541–548.
[33] Salehi Hikouei I, Kim S S, Mishra D R. Machine-learning classification of soil bulk density in salt marsh environments[J]. Sensors, 2021, 21(13): 4408.
[34] 汪容基, 赵小敏, 郭熙, 等. “三生空间”视角下的土地利用转型与生态环境效应研究——以江西省鹰潭市为例[J]. 江西农业大学学报, 2021, 43(3): 681–693.
[35] 江胜国. 国内土壤容重测定方法综述[J]. 湖北农业科学, 2019, 58(S2): 82–86, 91.
[36] 国家质量监督检验检疫总局, 中国国家标准化管理委员会. 土地利用现状分类: GB/T 21010—2017[S]. 北京: 中国标准出版社, 2017.
[37] 徐夕博, 吕建树, 吴泉源, 等. 基于 PCA-MLR 和 PCA-BPN 的莱州湾南岸滨海平原土壤有机质高光谱预测研究[J]. 光谱学与光谱分析, 2018, 38(8): 2556–2562.
[38] 闫政旭, 秦超, 宋刚. 基于 Pearson 特征选择的随机森林模型股票价格预测[J]. 计算机工程与应用, 2021, 57(15): 286–296.
[39] 李安琪, 杨琳, 蔡言颜, 等. 基于递归特征消除-随机森林模型的江浙沪农田土壤肥力属性制图[J]. 地理科学, 2024, 44(1): 168–178.
[40] 李恒凯, 王利娟, 肖松松. 基于多源数据的南方丘陵山地土地利用随机森林分类[J]. 农业工程学报, 2021, 37(7): 244–251.
[41] Breiman L. Random forest[J]. Machine Learning, 1999, 45: 1–35.
[42] Ordoñez Palacios L E, Bucheli Guerrero V, Ordoñez H. Machine learning for solar resource assessment using satellite images[J]. Energies, 2022, 15(11): 3985.
[43] Chai T, Draxler R R. Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literature[J]. Geoscientific Model Development, 2014, 7(3): 1247–1250.

[44] Yang Q Y, Luo W Q, Jiang Z C, et al. Improve the prediction of soil bulk density by cokriging with predicted soil water content as auxiliary variable[J]. Journal of Soils and Sediments, 2016, 16(1): 77–84.
[45] 任婷婷, 王瑄, 孙雪彤, 等. 不同土地利用方式土壤物理性质特征分析[J]. 水土保持学报, 2014, 28(2): 123–126.
[46] 游家兴. 如何正确运用因子分析法进行综合评价[J]. 统计教育, 2003(5): 10–11.
[47] McCaskey S D, Tsui K L. Analysis of dynamic robust design experiments[J]. International Journal of Production Research, 1997, 35(6): 1561–1574.
[48] 杨珺雯, 张锦水, 朱秀芳, 等. 随机森林在高光谱遥感数据中降维与分类的应用[J]. 北京师范大学学报(自然科学版), 2015, 51(S1): 82–88.
[49] 王猛, 张新长, 王家耀, 等. 结合随机森林面向对象的森林资源分类[J]. 测绘学报, 2020, 49(2): 235–244.
[50] 杨倩倩, 靳才溢, 李同文, 等. 数据驱动的定量遥感研究进展与挑战[J]. 遥感学报, 2022, 26(2): 268–285.
[51] 方灿莹, 王琳, 徐涵秋. 不同植被红边指数在城市草地健康判别中的对比研究[J]. 地球信息科学学报, 2017, 19(10): 1382–1392.
[52] Amiri M, Pourghasemi H R. Mapping the NDVI and monitoring of its changes using Google Earth Engine and Sentinel-2 images[M].Computers in Earth and Environmental Sciences. Amsterdam: Elsevier, 2022: 127–136.
[53] 鲍文, 赖奕卡. 湘中红壤丘陵区不同土地利用类型对土壤特性的影响[J]. 中国水土保持, 2011(10): 47–50, 66.
[54] 乔宇鑫, 朱华忠, 钟华平, 等. 内蒙古地区草地表层土壤容重空间格局分析[J]. 草地学报, 2016, 24(4): 793–801.