**DOI：10.11766/trxb202208090441**

张晓婷，黄魏，傅佩红，孟可，王苏放. 基于特征筛选算法的数字土壤制图研究[J]. 土壤学报，2024，61（3）：635–647.

ZHANG Xiaoting，HUANG Wei，FU Peihong，MENG Ke，WANG Sufang. Research on Digital Soil Mapping Based on Feature Selection Algorithm[J]. Acta Pedologica Sinica，2024，61（3）：635–647.

# 基于特征筛选算法的数字土壤制图研究*

张晓婷，黄　魏†，傅佩红，孟　可，王苏放

（华中农业大学资源与环境学院，武汉 430070）

**摘　要：**平缓地带数字土壤制图中，环境协变量的选择是提高制图精度的关键。已有研究证明遥感影像可作为推理制图的辅助因子，而如何确定环境因子推理制图时各自的权重已成为现阶段研究的重点。选取湖北省麻城市乘马岗镇为研究区，采用 3 种特征筛选方法进行有效环境变量筛选，探索参与平原-丘陵混合区域制图的因子并确定其重要性，依据选择的相对稳定的指标，进一步探索提高土壤类型制图准确性的途径。根据 141 个野外独立样点的检验结果表明：在推理制图中，遥感因子在平原区域的重要性程度高于丘陵区域，且遥感因子中归一化植被指数（NDVI）和均值（Mean）较为稳定；基于递归特征算法的按地形推理制图精度最高为 75.89%，分别高于 ReliefF 算法和基于 Tree 的特征筛选算法 13.48%和 4.97%；此外 3 种特征筛选算法制图结果中，按地形因子分区制图的精度均高于整体区域制图。因此，遥感因子作为辅助手段参与推理过程可有效提高制图精度。本研究采用的特征挖掘与机器学习算法对提升土壤制图精度具有一定的理论意义。

**关键词：**土壤-环境知识获取；特征筛选；数字土壤制图；贝叶斯优化；梯度提升树

**中图分类号：**S159　　**文献标志码：**A

# Research on Digital Soil Mapping Based on Feature Selection Algorithm

ZHANG Xiaoting, HUANG Wei†, FU Peihong, MENG Ke, WANG Sufang

(*College of Resource and Environment*, *Huazhong Agricultural University*, *Wuhan* 430070, *China*)

**Abstract:** 【Objective】Traditional digital soil mapping methods are unable to produce detailed soil maps within a reasonable cost and time. Digital soil mapping is a powerful technique, which is popular and widely used by scholars coupled with environmental covariates to map soil types or properties. The selection of environmental covariates is the key to ensuring the accuracy of mapping. Previous studies have proven that remote-sensing images can be used as auxiliary factors for reasoning mapping. Remote sensing data can provide rich soil landscape information, which is consistent with the core idea of using grids to express spatial changes of soil features in digital soil mapping. Moreover, remote sensing technology can obtain real-time information quickly. However, there are few relevant studies on how principal components and texture information of remote sensing factors contribute to the reasoning process. Thus, determining the weight of remote sensing factors in the reasoning process is the key content of this study, which is tested by the reliability of testing mapping results. 【Method】 Chengmagang Town, Macheng City,

* 国家自然科学基金项目（42171056，41877001）资助 Supported by the National Natural Science Foundation of China（Nos. 42171056，41877001）

† 通讯作者 Corresponding author，E-mail：ccan@mail.hzau.edu.cn

作者简介：张晓婷（1995—），女，山西阳泉人，硕士研究生，主要从事精细数字土壤制图研究。E-mail：1045442835@qq.com

收稿日期：2022–08–09；收到修改稿日期：2023–05–12；网络首发日期（www.cnki.net）：2023–10–24

Hubei Province was selected as the study area. Using Chinese soil classification and soil type map with a spacing of 10 meters, which were extracted from the contour data and remote sensing image using a variety of feature selection algorithms to effective screening of variables, this study conducted the soil digital mapping by reasoning machine learning algorithms. Specifically, the recursive feature elimination screening algorithm, ReliefF algorithm and tree-based feature screening algorithm were used to rank all environmental factors in the whole area, plain and hilly areas of the study area, respectively. Then, it screened the effective environmental variables of environmental factors and analyzed the weight of remote sensing factors in the reasoning process. The factors involved in plant-hill region mapping were explored and their importance was determined. According to the selected relatively stable indicators, the gradient boosting decision tree model after parameter tuning of the Bayesian optimization algorithm based on TPE was used for modeling. Also, the mapping accuracy results after different feature screening algorithms were compared between the whole region and the terrain region to further explore ways to improve the accuracy of soil type mapping.【Result】The soil type inference map was verified by 141 independent field sampling sites. The results showed that the importance of remote sensing factors in the plain area was higher than that in the hilly area and the NDVI and Mean values of the remote sensing factors were relatively stable. The highest accuracy of topographical inference mapping based on the recursive feature algorithm was 75.89%, which was higher than the 13.48% and 4.97% of the ReliefF algorithm and tree-based feature screening algorithm, respectively. In addition, among the mapping results of the three feature screening algorithms, the accuracy of the mapping based on terrain factors was higher than that of the overall region mapping. It suggests that remote sensing factors as an auxiliary means to participate in the reasoning process can effectively improve mapping accuracy. 【Conclusion】This study uses a feature selection algorithm to select features with a strong correlation with soil types as auxiliary variables in the machine learning model. The method is efficient and cost-effective for soil type prediction. Compared to other methods, the soil type mapping method based on machine learning is advantageous and the feature mining and machine learning algorithms have theoretical significance and practical value.

**Key words:** Soil-environment knowledge acquisition; Feature selecting; Digital soil mapping; Bayesian optimization; Gradient Boosting Decision Tree

数字土壤制图（Digital soil mapping）以土壤—景观模型为理论基础，利用具备协同空间变化的土壤类型和地理环境数据，借助于多种空间分析统计手段，推测土壤的空间分布，进而实现土壤类型的快速成图，因而被公认为是一种高效、实用的土壤数据更新方法[1-2]，该方法是基于土壤成因理论和土壤-环境知识理论，核心是土壤空间差异与环境因子的协同关系[3-5]。然而，将具有协同变化的地形因子作为环境变量参与成图推理过程时，在丘陵区域研究效果比较显著，而在环境变异较小的平原区域，难以有效地反映土壤的空间变异，探索和验证平缓地带的辅助推理因子成为目前本领域的研究瓶颈。已有研究分析和评估了遥感数据源与常规数据集组合，对研究典型土壤分布空间差异进行了探索，证明了遥感因子也可以作为平缓地区土壤推理制图的重要知识源[6-7]。因此，获取平缓区域土壤空间变异信息的环境协变量成为数字土壤制图的研究重点。

Xu 等[8]探索 TN（Total Nitrogen）与从 Landsat 8 、Sentinel 2 和 WorldView-2 图像得出的各种光谱指数之间的相互关系，基于不同遥感图像的 TN 多空间预测模型，评估不同遥感传感器预测总氮的能力；Zhang 等[9]综合分析训练样本和降水数据的像素日期对基于时间序列 MOD09A1 图像的区域 SOM 预测精度的影响；Duan 和 Zhang[10]利用 Landsat8 遥感影像证明了添加纹理特征能够提高土壤类型解译的准确性；在研究方法上，利用决策树、随机森林、支持向量机等方法进行数据挖掘推理土壤属性进行制图研究较多，但在对于参数较多的集成算法环节，目前使用较多的是运用自动调参的方法-网格搜索（Grid Search），但此方法只适用于较小的数据量级[11]。此外，影响土壤属性的环境因子较多，如何在较多的因子中选取最优的适用于平原和丘陵的因子、科学有效地进行数字土壤制图也是需要解决的问题。综上所述，虽有研究将纹理特征作为参与遥感图像地表识别的指标，但很少考虑将纹理特征与地形因子重要性程度用于土壤分类。

本研究的主要目的是综合利用地形因子和遥感影像作为推理制图的知识源，分别运用递归特征消除算法、ReliefF 算法和基于 Tree 的特征选择算法探究环境协变量在土壤推理制图中的重要性，利用经基于 TPE（Tree Parzen Estimator）的贝叶斯优化（Bayesian optimization）算法参数调优后的梯度提升树（Gradient Boosting Decision Tree）模型进行建模，同时比较不同特征筛选算法后整体区域与按照地形区域推理制图精度结果，为提高平缓区域制图精度奠定方法基础。

## 1　材料与方法

### 1.1　研究区概况

研究区为湖北省麻城市乘马岗镇，地势北高南低，属于典型的平原向丘陵过渡的地貌类型，经纬度范围为 114°54′—115°03′E，31°21′—31°30′N，面积约 28.14km$^2$。成土母质有“花岗岩片麻岩坡积或残积物”、“白云石花岗岩片麻岩坡积或残积物”和“近代河流冲积物”3 种类型，研究区如图 1 所示。

研究区属亚热带湿润季风气候，具有冬冷夏热、四季分明、降水充沛、光照充足、无霜期长、雨热同季的特征。农业集约化程度高。该区资源比较丰富，农业耕地历史悠久，具有优越的自然条件。

### 1.2　数据来源

**1.2.1**　地形数据与遥感数据　　本研究以中国土壤发生分类土壤类型图、等间距为 10m 的等高线数据（来源于湖北省测绘局）和遥感影像（选取植被覆盖度适中的 2020 年 12 月 15 日的 Sentinel-2 遥感影像）为知识源。传统土壤类型图（1∶50 000）来源于第二次全国土壤普查成果，研究区内共有 9 种土壤类型，如表 1 所示。

**1.2.2**　样本数据　　本文将从传统土壤图中选取训练样本，采用均匀网格与随机抽样集合的方式确定了 1973 个训练样点集（表 1），再利用数据挖掘算法概括样本中所蕴含的土壤学家“隐形知识”[12-15]并对其可视化，训练样点及野外独立验证样点 141 个空间分布（图 2）。

为探究环境因子在按地形推理制图与整体区域制图的差异，本研究依据地形起伏度分布图和土壤志等历史资料，以海拔 150m 为平原和丘陵区域的临界值，得到平原和丘陵区域的训练样点集。

**1.2.3**　土壤-环境知识获取　　利用等高线数据生成高程（Elevation），从而使用 ArcGIS10.7 提取坡度（Slope）、坡向（Aspect）、水平曲率（Horizontal curvature）、平面曲率（Plan curvature）、剖面曲率

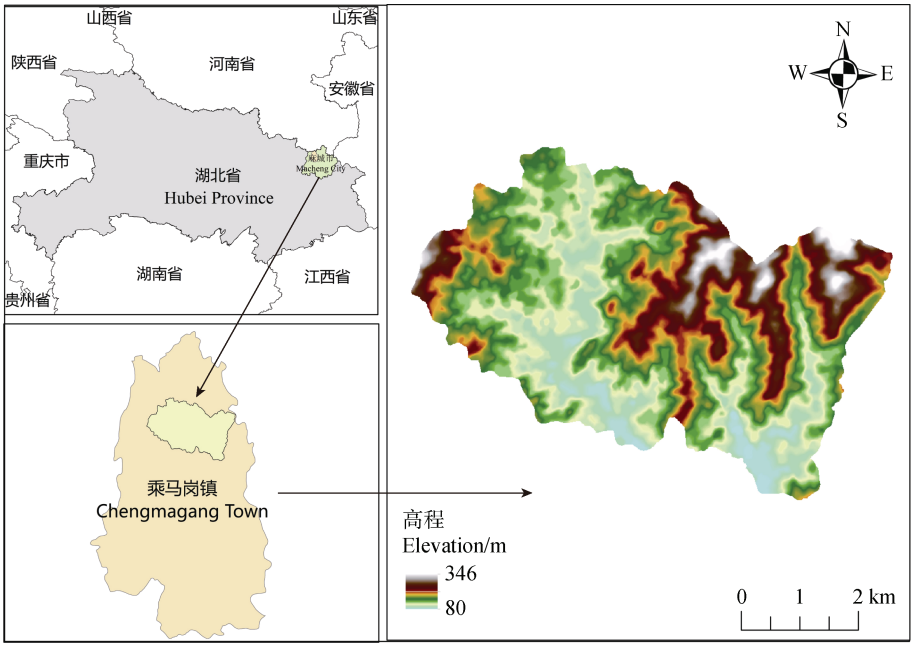


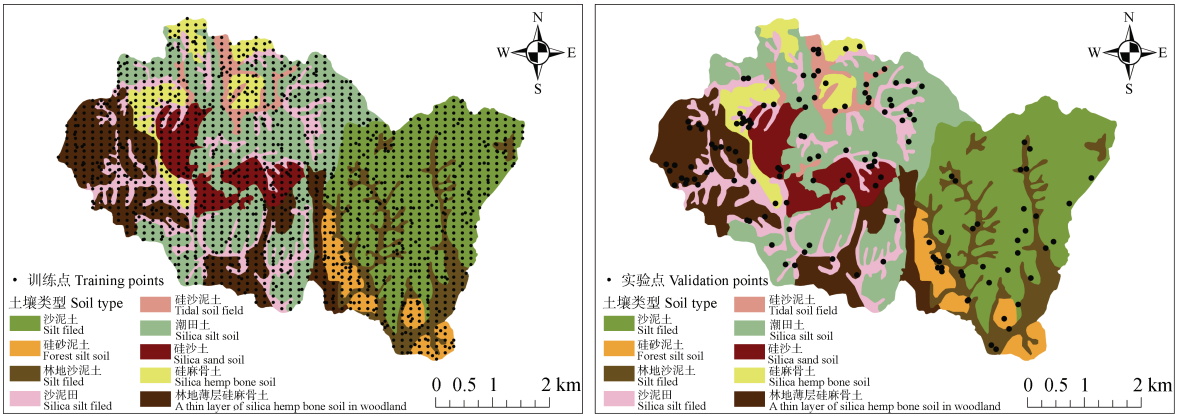

图 1　研究区示意图

Fig. 1　Location map of the study area

**表 1 土壤类型面积与训练点、验证点信息**

**Table 1** Soil type area and information of training points and validation points

| 土壤类型 Soil type | 土壤代码 Soil code | 母质 Parent material | 面积 Area /km$^2$ | 面积占比 Proportion of area/% | 训练点 Training points | 验证点 Validation points |
|---|---|---|---|---|---|---|
| 沙泥土 Silt soil | 1 | 花岗岩片麻岩的坡积或残积物 | 6.99 | 24.84 | 490 | 16 |
| 硅沙泥土 Silica silt soil | 4 | 白云石英片岩花岗片麻岩坡积或残积物 | 6.63 | 23.56 | 469 | 25 |
| 林地泥沙土 Forest silt soil | 1-7 | 花岗岩片麻岩的坡积或残积物 | 1.12 | 3.98 | 90 | 9 |
| 沙泥田 Silt field | 22 | 花岗岩片麻岩的坡积或残积物 | 2.40 | 8.53 | 164 | 14 |
| 硅沙泥田 Silica silt field | 24 | 白云石英片岩花岗片麻岩坡积或残积物 | 3.18 | 11.30 | 216 | 19 |
| 潮土田 Tidal soil field | 30 | 近代河流冲积物 | 0.73 | 2.59 | 46 | 13 |
| 硅沙土 Silica sand soil | 4-4 | 白云石英片岩花岗片麻岩坡积或残积物 | 1.74 | 6.18 | 122 | 12 |
| 硅麻骨土 Silica hemp bone soil | 4-5 | 白云石英片岩花岗片麻岩坡积或残积物 | 1.40 | 4.98 | 96 | 11 |
| 林地薄层硅麻骨土 A thin layer of silica hemp bone soil in woodland | 4-6 | 白云石英片岩花岗片麻岩坡积或残积物 | 3.95 | 14.04 | 280 | 22 |
| | | 合计 | 28.14 | 100 | 1 973 | 141 |




a. 土壤类型与训练点分布图
a. Map of soil type and training points

b. 土壤类型与验证点分布图
b. Map of soil type and validation points

图 2 研究区训练点及验证点分布图

Fig. 2 Distribution of training points and validation points in the study area

（Profile curvature）、地形湿度指数（Topographic wetness index，TWI）、地表粗糙度（Roughness）和地形起伏度（Relief）即 8 个地形因子。同时，使用 ENVI5.3 对预处理后的遥感影像的多个波段进行主成分分析（Principal component analysis，PCA），进而基于主成分分析提取纹理特征（Textural feature）。光谱信息的主成分（PC）波段是原始波段的线性合成，主成分分析可以隔离噪声和较少数据维数，并提供更广泛的空间信息。因此本文提取归一化植被指数（Normalized differential vegetation index，NDVI）、第一主成分（First principal component，FPC）和 8 个纹理特征包括均值（Mean）、方差（Variance）、相异性（Dissimilarity）、信息熵（Entropy）、对比度（Contrast）、协同性（Homogeneity）、二阶矩（Second Moment）、相关性（Correlation）为后续筛选推理制图的遥感因子。

### 1.3　特征筛选方法

**1.3.1**　递归特征消除算法　　本文使用 PyCharm2021 中的 RFE 和 RandomForestClassifier 包，通过十折交叉验证的递归特征消除来确定所选特征的最佳数量。递归特征消除算法（Recursive feature elimination，RFE）是一种贪心算法，用于找到最优特征子集并降低数据维度以获得最高精度的模型。算法的基本思想：首先构建底层模型并对初始特征集进行训练，计算每个特征的重要性并赋予权重，然后去掉重要性低的权重最小的特征，将剩余特征组合成新的特征子集，并进行训练。该过程递归重复，直到达到所需的最终特征数目[16]。在本研究中，RFE 使用随机森林（Random forest，RF）作为底层模型，选取默认参数，并选择训练结果精度最大的特征子集作为构建模型的特征组合。RF 为由多棵决策树构成的集成模型，模型的最终输出结果由森林中的每棵决策树共同决定，最终输出结果为每棵决策树输出的均值。RF 算法的具体过程如下：在原始训练特征集中用 Bootstrap 抽样方法获得 $n$ 个特征子集；对每个特征子集选择 $m$ 个特征，并对每个训练特征子集构建决策树，得到 $n$ 个决策树模型，建立起随机森林；计算每棵决策树的结果，将 $n$ 棵决策树输出结果的均值作为最终结果。

**1.3.2**　ReliefF 算法　　Relief 算法只能处理两类类别的数据。Kononenko[17]提出的 ReliefF 算法是在 Relief 算法的基础上进行改进的，它不仅能够处理有噪声的多类及回归问题，还可以处理不完整的数据。在解决多类及回归问题时，ReliefF 算法会从训练样本集中随机选取一个样本 $R$，并从与 $R$ 同类的样本集中找出 $R$ 的 $k$ 个近邻样本（Near hits），从与 $R$ 不同类的样本集中找出 $k$ 个近邻样本（Ness misses），然后更新每个特征的权重，该过程会重复 $m$ 次[18]。

ReliefF 具有较高的评估效率、对数据的类型没有限制要求等优点，可以很好去除无关特征，但不能去除冗余特征，是公认的效果较好的过滤式的特征评估算法，具备过滤式算法的典型特征，即省去了对特征子集的分类器进行训练的步骤，减少了计算量，且简单高效。

**1.3.3**　基于 Tree 的特征选择　　Python 中 SelectFromModel 包是一个 meta-transformer（元转换器），它可以用来处理任何带有 coef_ 或者 feature_importance 属性的训练之后的评估器。如果相关的 coef_或者 feature_importance 属性值低于预先设置的阈值，这些特征将会被认为不重要并且移除掉。除了指定数值上的阈值之外，还可以通过给定字符串来使用内置的启发式方法找到一个合格的阈值，可以使用的启发式方法有 mean、median 以及使用浮点数据乘以阈值。

本文加权模式使用梯度提升树来提取所有特征的重要性。

### 1.4　土壤推理制图

本文运用梯度提升树（Gradient Boosting Decision Tree，GBDT）模型进行推理制图。GBDT 集成梯度提升（Gradient Boosting）与决策树（Decision Tree）两种算法思想，其中梯度提升属于 boosting 算法，其每次建立的单个学习器是在之前建立的模型损失函数梯度下降方向。损失函数（loss function）越大，模型越容易出错，如果损失函数持续下降，则意味着模型在不断地自我改进，而最好的改进方式就是损失函数在其梯度（Gradient）方向下降。基于梯度提升算法的学习器叫作梯度提升机（Gradient Boosting Machine）。理论上，梯度提升机可以选择各种不同的学习算法作为基础学习器。GBDT 采用决策树作为基础学习器，通过构造一颗决策树，然后在已有模型和实际样本输出的残差上构造另一颗树来迭代训练。最终将所有决策树的结果作为输出。在每轮迭代中，GBDT 会产生一个弱

分类器，该分类器在上一轮分类器的残差基础上进行训练。为了避免过拟合，弱分类器应该足够简单，满足低方差和高偏差[19-20]。在训练过程中，降低偏差可以提高分类器精度，因为模型越简单越容易产生欠拟合。为了提高数据的有效性，训练集采用随机抽取，以最大限度保证分布情况的一致性。当然，选择合适的超参数和建立有效的预测模型也是至关重要的。

贝叶斯优化算法采用的是一种逼近思想，当算法的计算过程非常复杂、需要较高的迭代次数时通常能够起到很好的效果，大多被用于机器学习算法模型的超参数取值确定。实现贝叶斯优化的库有很多种，本文选用基于 TPE 贝叶斯对 GBDT 进行优化。

对 GBDT 算法模型进行参数优化的基本描述是：首先设置超参数取值区间组合，接着利用贝叶斯最优化算法的相关理论，通过不断训练 GBDT 模型并使用评价函数对每组参数组合得到的分类预测结果进行评价，以此寻求最优的参数取值组合。最后，将最优参数取值组合代入 GBDT 算法模型中，提高预测准确度并降低算法运行耗时。

### 1.5 精度验证

本研究采用混淆矩阵以及 Kappa 系数对推理结果的精度进行评价。混淆矩阵是用矩阵形式来评价土壤分类精度的一种方式，用矩阵中的总体精度、用户精度以及生产精度三个评价指标来比较推理结果与实际采样之间的差异。

Kappa 系数用于衡量分类的效果，计算结果通常介于 0～1 之间，0～0.2 为极低一致性，0.2～0.4 为一般一致性，0.4～0.6 为中等一致性，0.6～0.8 为高度一致性，0.8～1 为几乎完全一致。

## 2 结果与讨论

### 2.1 特征筛选结果

**2.1.1** 基于递归特征消除算法的特征筛选 随着递归特征消除反复构建随机森林模型，精度达到最高值，分类效果最佳。此外，RF 算法提供了总和为 1 的特征重要性接口，用来比较各个特征的重要性。经过参数优选和多次训练，确定 RF 的主要参数决策树数量为 1000。

RF-RFE 算法使用精度（Accuracy）值作为特征选择过程中筛选的原则。算法首先将删除重要性最小的特征，如下图 3 所示，删除第 20 个特征，然后 RF-RFE 重新计算剩余特征的精度值，再次删除重要性最小的特征，依次循环，不断迭代删除特征[21]。图 3 表示的是每一次迭代时采用十折交叉验证所删除特征的精度值，迭代结束后，根据精度的大小选取最优特征集合。

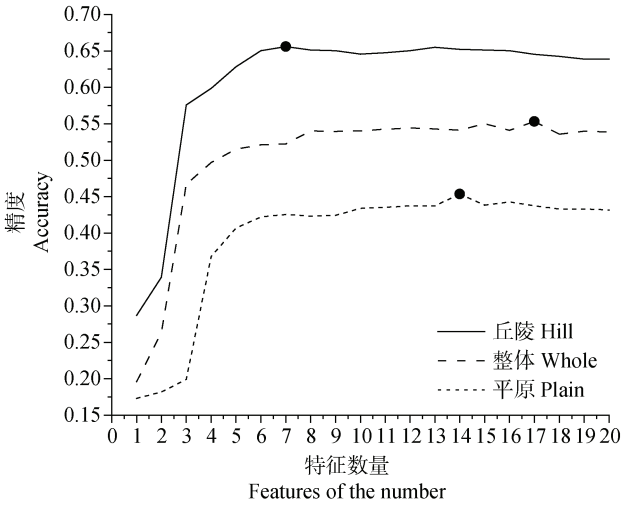



图 3 特征筛选过程中精度的变化

Fig. 3 Changes in accuracy during feature selecting

从下表 2 可以得出，在 20 个环境因子中，母质在整体区域（Whole）、平原（Plain）和丘陵区域（Hill）均是重要性第一的因子，但其重要性不同，在丘陵区域重要性最大，重要性值为 0.253，平原区域重要性较低，重要性值为 0.170。遥感因子中的第一主成分、均值和相关性在平原区域的重要性评分明显高于整体和丘陵区域。协同性、对比度和相异性纹理特征在整体区域、平原区域和丘陵区域重要性较低，均不参与后续建模。

由图 3 可知，通过十折交叉验证，得出丘陵的最佳特征数量为 7，对应精度最高为 0.655，整体区域特征数量达到 17 时精度最高，精度为 0.553，当特征数量达到 14 时平原区域精度为 0.453，且精度最高。在每次筛选得到的精度结果中，整体区域精度和丘陵区域精度高于平原区域。表 4 显示丘陵区域选取特征排名前 7 的因子、整体区域选取特征排名前 17 的因子以及平原区域选取特征排名前 14 的因子。

**2.1.2** 基于 ReliefF 算法的特征筛选 母质相比其他因子重要性较大，基于 ReliefF 算法的特征权重

**表 2　因子重要性评分排序**

**Table 2**　Factor importance rating sort

| 整体区域 While | 评分 Score | 平原区域 Plain | 评分 Score | 丘陵区域 Hill | 评分 Score |
|---|---|---|---|---|---|
| 母质 Parent material | 0.206 | 母质 | 0.170 | 母质 | 0.253 |
| 高程 Elevation | 0.081 | 归一化植被指数 | 0.062 | 归一化植被指数 | 0.072 |
| 归一化植被指数 NDVI | 0.066 | 高程 | 0.061 | 坡向 | 0.069 |
| 坡向 Aspect | 0.058 | 地形湿度指数 | 0.056 | 高程 | 0.069 |
| 地形湿度指数 TWI | 0.049 | 第一主成分 | 0.053 | 地形湿度指数 | 0.042 |
| 第一主成分 FPC | 0.047 | 相关性 | 0.049 | 第一主成分 | 0.041 |
| 坡度 Slope | 0.044 | 剖面曲率 | 0.048 | 坡度 | 0.041 |
| 地形起伏度 Relief | 0.043 | 均值 | 0.047 | 地形起伏度 | 0.041 |
| 地表粗糙度 Roughness | 0.041 | 坡向 | 0.047 | 地表粗糙度 | 0.041 |
| 水平曲率 Horizontal curvature | 0.041 | 水平曲率 | 0.046 | 水平曲率 | 0.040 |
| 平面曲率 Plan curvature | 0.041 | 平面曲率 | 0.045 | 平面曲率 | 0.039 |
| 相关性 Correlation | 0.040 | 坡度 | 0.043 | 剖面曲率 | 0.038 |
| 剖面曲率 Profile curvature | 0.040 | 地表粗糙度 | 0.040 | 相关性 | 0.035 |
| 均值 Mean | 0.040 | 地形起伏度 | 0.039 | 均值 | 0.033 |
| 信息熵 Entropy | 0.031 | 信息熵 | 0.036 | 方差 | 0.026 |
| 方差 Variance | 0.029 | 方差 | 0.036 | 信息熵 | 0.026 |
| 二阶矩 Second moment | 0.028 | 协同性 | 0.033 | 二阶矩 | 0.025 |
| 协同性 Homogeneity | 0.027 | 二阶矩 | 0.033 | 协同性 | 0.024 |
| 对比度 Contrast | 0.026 | 对比度 | 0.032 | 对比度 | 0.022 |
| 相异性 Dissimilarity | 0.021 | 相异性 | 0.026 | 相异性 | 0.020 |

图中使得其他因子重要性显示不明显，因此，本文将得到去除母质的特征权重（如图 4 所示），特征权重越大，表示该特征对样本的区分效果越好。本文计算所有特征权重均值作为阈值选择最优特征子集[22]。整体区域中，特征权重的均值为 0.07。高程、归一化植被指数、坡向、坡度、地形起伏度、地形湿度指数的权重超过 7%，其余特征的权重低于 7%。因此，权重高于均值的特征包括高程、归一化植被指数、坡向、坡度、地形起伏度、地形湿度指数共 6 个特征，加上母质共计 7 个用于构建特征子集。平原区域中，特征权重的均值为 0.031。高程、坡向、归一化植被指数的权重超过 8%，地形湿度指数、坡度、地形起伏度、均值的权重值在 3.1%～8%之间，其余特征的权重低于 3.1%。因此，权重高于均值的特征包括高程、坡向、归一化植被指数、地形湿度指数、坡度、地形起伏度、均值共 7 个特征，加上母质共计 8 个特征用于构建特征子集。丘陵区域中，特征权重的均值为 0.081。高程、坡向、归一化植被指数、坡度、地形起伏度超过 8%，其余特征的权重低于 8.1%。因此，权重高于均值的特征包括高程、坡向、归一化植被指数、坡度、地形起伏度共 5 个特征，加上母质共计 6 个特征用于构建特征子集。

综上，根据筛选结果表明，高程、坡向、归一化植被指数、坡度、地形起伏度为公共特征因子，遥感因子中的纹理特征均值只出现在平原区域中，可知均值在平原区域中重要性程度大于整体和丘陵区域，在土壤类型推理制图中均值在平原区域占有一定的比重。

**2.1.3**　基于Tree的特征选择结果　　利用PyCharm2021 软件中GradientBoostingClassifier和

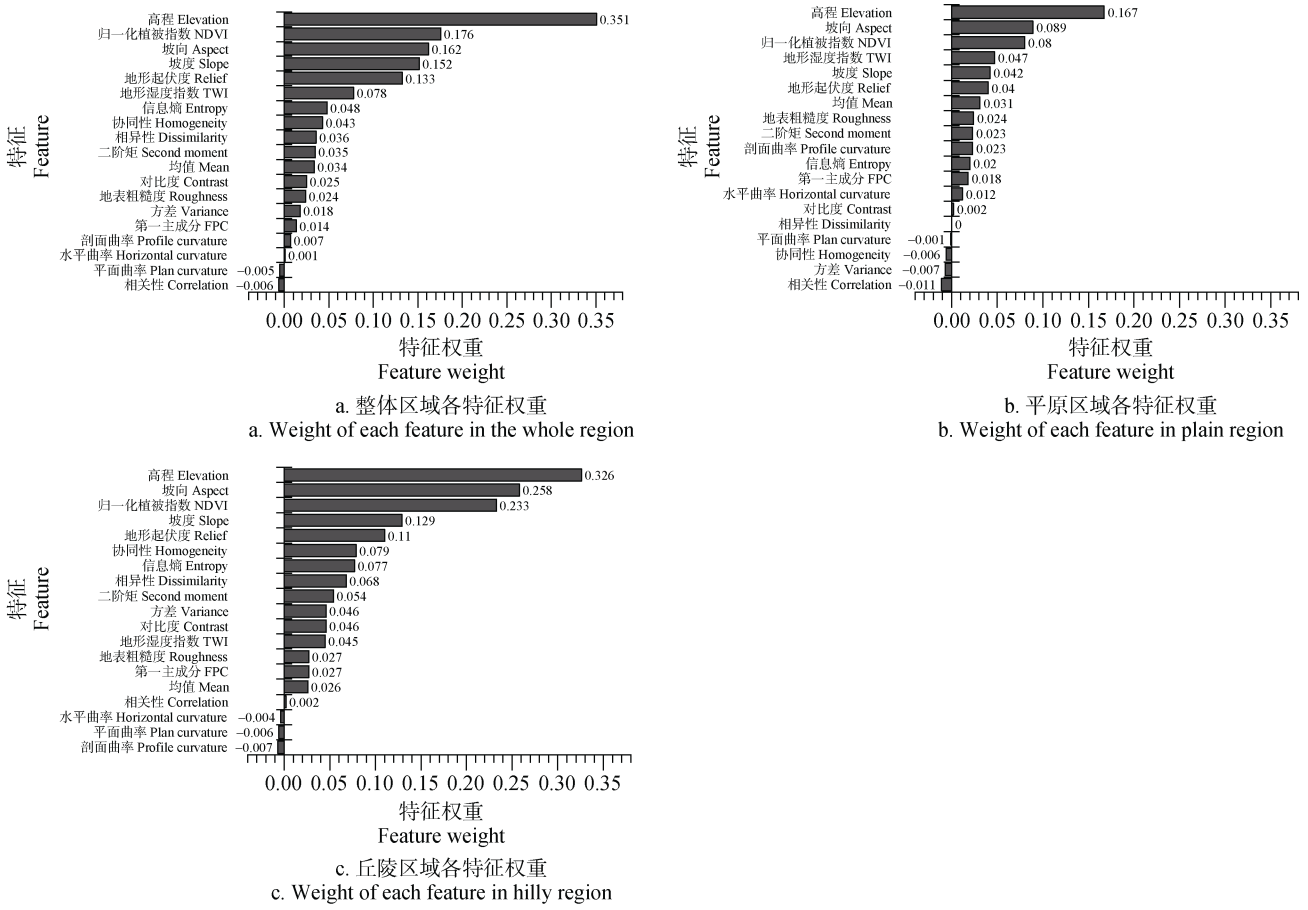



图 4 基于 ReliefF 算法的各区域特征权重

Fig. 4 Feature weight of each region based on ReliefF algorithm

SelectFromModel 包，对特征因子进行重要性评分，结果如下表 3 所示，通过给定内置阈值分别去除不重要因子，得出整体、平原和丘陵区域最优特征数量分别为 11、15 和 15。

依据整体区域重要性将环境因子从高到低进行排序，结果如表 3 所示，在整体区域中被筛选出的遥感因子占所有因子的 5%，而在平原和丘陵区域中被筛选出的遥感因子所占比重均为 25%，但其因子不同，平原区域重要性较高的遥感因子为归一化植被指数、第一主成分、相关性、均值和信息熵，而丘陵区域重要性较高的遥感因子为归一化植被指数、第一主成分、相关性、均值和方差，纹理特征中协同性、对比度、二阶矩和相异性均为不重要的因子。

## 2.2 特征筛选算法分析

本文基于递归特征消除算法、ReliefF 算法以及基于 Tree 的特征选择构建整体区域、平原以及丘陵区域的特征子集。

综上，由 3 种特征筛选算法筛选结果可以看出在平原区域纹理特征的重要性程度高于整体与丘陵区域，并且遥感因子中的归一化植被指数和均值相较于其他的遥感因子所占比重较大，原因是平原区域中采样点之前的地形因子差异比较小，平原区域地势起伏较小，难以有效地反映土壤的空间变异，所以需要遥感因子参与其中进而对平原区域土壤类型分类进行推理。

## 2.3 GBDT 参数优化

GBDT 作为一种集成算法，不同参数对算法结果的影响力不同。运用 Py Charm2021 中的 hyperopt 包，基于 TPE 对 GBDT 模型对如下参数进行优化。

以整体区域为例：

在进行 n_estimators 参数调优时，首先取值 25～200，步长为 25，得到最佳平均精度对应的值为 125。再通过逐步逼近法，取 n_estimators 为 75～150，步

**表 3　基于 Tree 的各区域特征权重**

**Table 3**　The feature weight of each region based on Tree

| 环境因子 Environmental factors | 整体 Whole | 平原 Plain | 丘陵 Hill | 环境因子 Environmental factors | 整体 Whole | 平原 Plain | 丘陵 Hill |
|---|---|---|---|---|---|---|---|
| 母质 Soil parent material | 0.435 | 0.351 | 0.463 | 地形起伏度 Relief | 0.022 | 0.022 | 0.017 |
| 高程 Elevation | 0.133 | 0.061 | 0.087 | 坡度 Slope | 0.021 | 0.03 | 0.023 |
| 归一化植被指数 NDVI | 0.071 | 0.068 | 0.085 | 均值 Mean | 0.02 | 0.03 | 0.02 |
| 坡向 Aspect | 0.066 | 0.046 | 0.089 | 地表粗糙度 Roughness | 0.012 | 0.018 | 0.015 |
| 地形湿度指数 TWI | 0.039 | 0.059 | 0.034 | 信息熵 Entropy | 0.012 | 0.021 | 0.007 |
| 平面曲率 Plan curvature | 0.032 | 0.043 | 0.033 | 方差 Variance | 0.011 | 0.014 | 0.014 |
| 第一主成分 FPC | 0.031 | 0.061 | 0.024 | 对比度 Contrast | 0.008 | 0.015 | 0.004 |
| 水平曲率 Horizontal curvature | 0.024 | 0.039 | 0.022 | 二阶矩 Second moment | 0.008 | 0.011 | 0.006 |
| 相关性 Correlation | 0.023 | 0.045 | 0.022 | 协同性 Homogeneity | 0.007 | 0.017 | 0.01 |
| 剖面曲率 Profile curvature | 0.022 | 0.046 | 0.022 | 相异性 Dissimilarity | 0.002 | 0.003 | 0.004 |

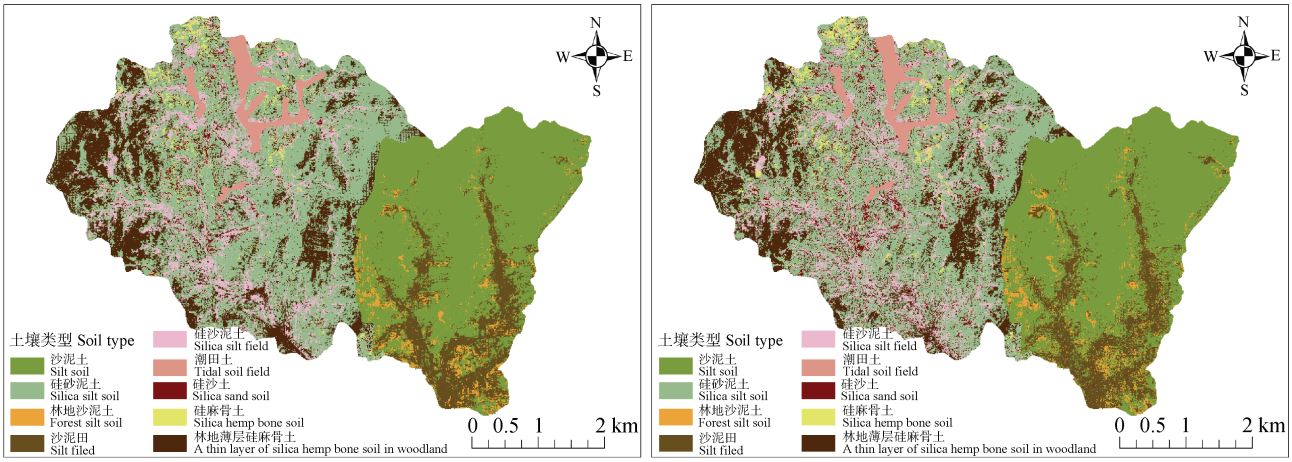
长为 5，再次得到最佳平均精度所对应的值为 80；进行多次尝试，将范围缩小，步长为 1，得到最优值为 71。在进行 learning_rate 参数调优时，首先取值 0.05～2.05，步长为 0.05，得到最佳平均测试分数为 0.35；再通过逐步逼近法，取 learning_rate 为 0.05～0.65，步长为 0.001，得到最优的值为 0.096。在进行 Subsample 参数调优时，首先取值 0.1～0.8，步长为 0.1，得到最佳平均测试分数为 0.6；再通过逐步逼近法,取 Subsample 为 0.4～0.8,步长为 0.05。进行多次尝试，将范围缩小，步长为 0.0025，得到最优的值为 0.782；进行 max_depth 调参，首先取值 2～30，步长为 2，得到最佳平均测试分数为 28；再通过逐步逼近法，取 max_depth：14～32，步长为 1，得到最优的值（18）。min_impurity_decrease 在进行参数调优时，此参数较不稳定，所以不再进行参数调整。

如果其中有的参数趋于稳定，将不进行继续调参，在目标函数中，直接写入固定值。通过实验，将基于递归特征消除算法、ReliefF 和基于 Tree 的特征选择的筛选结果对整体、平原及丘陵区域的 GBDT 算法的参数设置为当前场景下的最佳参数取值。

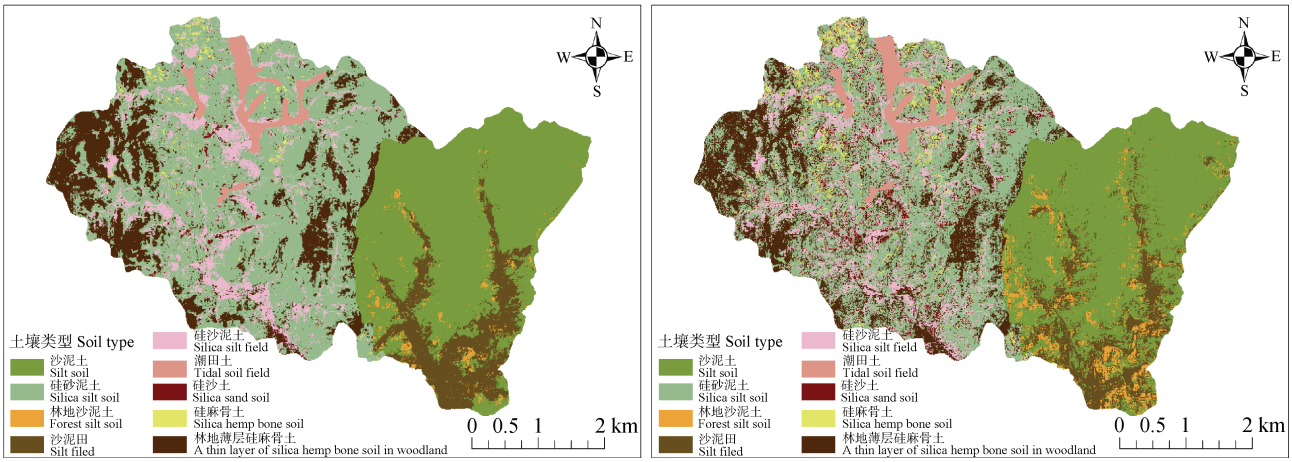
### 2.4　模型制图及精度验证

采用整体区域以及面向不同地形区域分别进行土壤类型预测，基于 3 种特征筛选算法的整体区域以及按照地形的推理结果如图 5 所示。对比基于 3 种特征筛选算法的整体区域以及分区域的推理图，可以发现整体上呈现出一致趋势，相比原始土壤图，整体区域制图以及按照地形推理图图斑更加破碎，沙泥土、潮土田、沙泥田以及林地泥沙土分布范围基本一致，沙泥土、沙泥田和林地泥沙土均由花岗岩片麻岩坡积物或残积物母质发育而来，沙泥土、沙泥田和林地泥沙土主要分布在东部地区，相比整体区域推理图，林地泥沙土按照地形推理效果更好。潮土田唯一由近代河流冲积物发育而来，较其他 8 种土壤类型而言推理效果最佳。硅沙泥田在平原区域分布，并与硅沙土相互夹杂，《麻城市土壤志》记载，这两种土壤类型的成土环境相似，土壤具有相似特征，因此这两种土壤往往相伴而生，虽如此，按照地形推理制图分类情况有所提升。局部细节中，相比于整体区域推理的土壤图，硅麻骨土和林地薄层硅麻骨土在按照地形的土壤图中图斑明显增多，空间细节更加丰富，由此可见土壤类型的空间分布与景观特征相吻合，这也进一步验证了土壤类型空间分布的准确性。

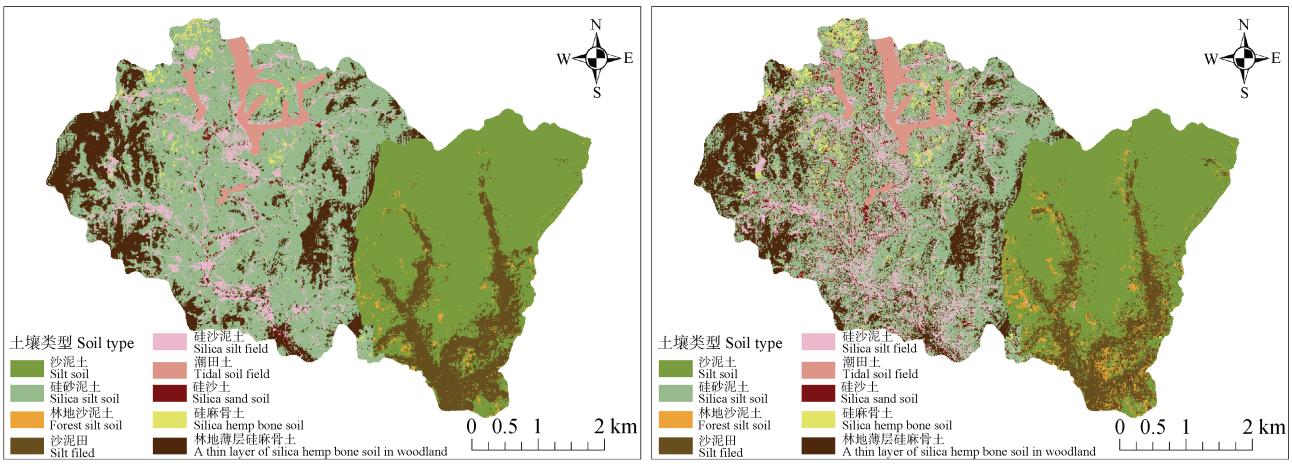
比较 3 种不同特征筛选算法的推理图，按照地形推理结果较整体区域推理结果更加破碎，利用野外实地采集的独立验证点对推理图进行精度评估，用于精度检验的野外实地样点共 141 个，对比基于递归特征消除算法、ReliefF 算法以及基于 Tree 的特征选择算法的整体区域以及按照地形区域推理图，按照地形区域的推理图正确分类的样点个数分别为




a. 基于递归特征消除算法土壤类型栅格图 (左图为整体区域推理图, 右图为按地形推理图)
a. Grid maps of soil type based on recursive feature elimination algorithm (The figure on the left is a whole regional inference map, and the figure on the right is an inference map by terrain)




b. 基于ReliefF算法土壤类型栅格图 (左图为整体区域推理图, 右图为按地形推理图)
b. Grid maps of soil type based on ReliefF algorithm (The figure on the left is a whole regional inference map, and the figure on the right is an inference map by terrain)




c. 基于Tree算法土壤类型栅格图 (左图为整体区域推理图, 右图为按地形推理图)
c. Grid maps of soil type based on Tree algorithm (The figure on the left is a whole regional inference map, and the figure on the right is an inference map by terrain)

图 5 土壤类型栅格图
Fig. 5 Grid maps of soil type

107 个、88 个和 100 个，总体精度分别为 75.89%、62.41%和 70.92%，分别高于整体区域推理图（64.54%、60.28%和 60.99%）；而 Kappa 系数分别为 0.727、0.565 和 0.666，分别高于整体制图的 0.137、0.022 和 0.116（整体区域分别为 0.589、0.543 和 0.55），能较好地反映研究区的土壤分布情况。

利用按照地形推理的原则建立混淆矩阵，以生产精度和用户精度为指标验证按照地形推理的土壤图精度。整体推理图和按照地形的推理图的生产精度和用户精度统计结果如下表 4 所示。生产精度是该土壤类型正确分类点个数与该土壤类型验证点个数之比，用户精度是正确分类点个数与被判定为该土壤类型的点数之比。土种级别的分类制图相较于土属级别，分类精度的提升从原理而言更加困难。

**表 4　基于三种算法的各土壤类型推理制图生产精度和用户精度**

**Table 4**　Production and user accuracy of inferential mapping for each soil type based on three algorithms（%）

| 土壤类型 Soil type | 基于递归特征消除算法的推理制图 Inferential mapping based on recursive feature elimination algorithm | | | | 基于 ReliefF 算法的推理制图 Inferential mapping based on ReliefF algorithm | | | | 基于 Tree 的特征选择算法的推理制图 Inferential mapping based on Tree feature selection algorithm | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 生产精度 Accuracy for production | | 用户精度 Accuracy for users | | 生产精度 Accuracy for production | | 用户精度 Accuracy for users | | 生产精度 Accuracy for production | | 用户精度 Accuracy for users | |
| | 整体 Whole | 按地形 By terrain | 整体 Whole | 按地形 By terrain | 整体 Whole | 按地形 By terrain | 整体 Whole | 按地形 By terrain | 整体 Whole | 按地形 By terrain | 整体 Whole | 按地形 By terrain |
| 沙泥土 Silt soil | 75 | 81 | 67 | 72 | 75 | 81 | 71 | 68 | 69 | 81 | 65 | 72 |
| 硅沙泥土 Silica silt soil | 72 | 80 | 40 | 61 | 64 | 64 | 42 | 36 | 64 | 72 | 39 | 51 |
| 林地泥沙土 Forest silt soil | 67 | 78 | 100 | 100 | 89 | 67 | 73 | 100 | 67 | 71 | 100 | 100 |
| 沙泥田 Silt field | 71 | 71 | 67 | 71 | 57 | 71 | 73 | 71 | 71 | 42 | 62 | 71 |
| 硅沙泥田 Silica silt field | 42 | 58 | 67 | 73 | 21 | 32 | 36 | 50 | 32 | 85 | 50 | 62 |
| 潮土田 Tidal soil field | 92 | 92 | 100 | 100 | 92 | 92 | 100 | 100 | 92 | 42 | 100 | 100 |
| 硅沙土 Silica sand soil | 33 | 58 | 57 | 78 | 42 | 33 | 62 | 57 | 25 | 58 | 43 | 64 |
| 硅麻骨土 Silica hemp bone soil | 18 | 55 | 100 | 86 | 18 | 18 | 33 | 67 | 27 | 55 | 100 | 86 |
| 林地薄层硅麻骨土 A thin layer of silica hemp bone soil in woodland | 86 | 95 | 79 | 81 | 82 | 86 | 67 | 79 | 86 | 91 | 70 | 80 |

从整体区域制图和按照地形推理制图结果得出，基于递归特征筛选算法的精度分别高于基于 ReliefF 算法和基于 Tree 的特征选择算法精度的 4.28%和 3.55%，按照地形推理中，在平原区域，由于基于 ReliefF 算法筛选出平原的因子较少，加入的遥感因子较少，导致总体精度低于基于递归特征消除算法以及基于 Tree 特征选择算法的精度。基于 3 种特征筛选算法推理制图中，按照地形推理制图精度高于整体制图精度。基于递归特征消除算法的制图中，9 种土壤类型生产精度和用户均高于整体区域制图，在基于 ReliefF 算法中，潮土田、硅沙泥田、硅麻骨土和林地薄层硅麻骨土的生产精度和用户精度高于整体推理图，基于 Tree 的特征筛选算法中，按照地形推理的沙泥土、硅沙泥土、林地泥沙田、硅沙土和林地薄层硅麻骨土的生产精度和用户精度均高于整体制图精度。

本研究提出的基于递归特征消除算法、ReliefF 算法以及基于 Tree 的特征筛选算法 3 种不同的特征挖掘算法对环境变量（包含地形因子和遥感因子）进行筛选，对比整体区域和按照地形的推理制图，分析遥感因子在推理过程中的权重，以选择稳定的遥感指标，探索进一步提高制图准确性的途径，取得了较好的效果，但仍存在不足需要进一步探讨。在小区域尺度下，能够反映土壤与环境之间关系环境因子还有很多，缺乏其他方面的因子参与筛选，比如：土地利用方式。因此，选择更多的环境因子进行科学的分析筛选并运用至推理制图中，使得因子在判别土壤类型更具准确性是今后重点研究方向之一。如今，新的机器学习方法层出不穷，选择更多更好的机器学习模型进行对比推理制图，使结论更具客观性和准确性也是今后研究者重点研究的课题之一。本文未采用 SoLIM 等模糊推理算法，因此不能进行推理结果的不确定性研究，但这应该是本研究今后工作的重要方向之一。

## 3 结 论

本研究提出了一种基于特征筛选算法进行土壤类型推理的数字土壤制图获取方法，利用递归特征消除算法、ReliefF 算法以及基于 Tree 的特征选择方法对环境协变量进行特征选择并分析，并且采用基于 TPE 贝叶斯优化的超参数优化后的 GBDT 进行建模，利用野外采样点进行独立验证。结果表明：环境因子在参与平原和丘陵区域的推理制图时具有不同的重要性，遥感因子在平原区域参与程度大于丘陵区域，其中遥感因子中的 NDVI 和 Mean 相较于其他遥感因子对土壤类型的分类重要性更大。3 种筛选算法对环境因子进行重要性评分，从而判断因子对土壤类型的影响程度，能够减少大量的冗余，提高分类精度及效率，为进一步环境变量与土壤类型之间的联系建立提供了新的研究思路。基于 3 种筛选算法按照地形推理的制图精度均高于整体推理制图精度，递归特征算法的按照地形推理制图精度最高为 75.89%，分别高于 ReliefF 算法和基于 Tree 的特征筛选算法。因此，本研究提出的不同特征筛选方法以及基于 TPE 贝叶斯优化对 GBDT 进行参数优化的方法可为土壤制图等相关研究提供理论参考。

**参考文献（References）**


[1] Carré F, McBratney A B, Mayr T, et al. Digital soil assessments: Beyond DSM[J]. Geoderma, 2007, 142(1/2): 69—79.

[2] Huang W, Xu W, Wang S Q, et al. Extraction of knowledge about soil-environment relationship based on an uncertainty model[J]. Acta Pedologica Sinica, 2018, 55(1): 54—63. [黄魏, 许伟, 汪善勤, 等. 基于不确定性模型的土壤——环境关系知识获取方法的研究[J]. 土壤学报, 2018, 55(1): 54—63.]

[3] Zhu A X, Li B L, Pei T, et al. Model and method of fine digital soil survey[M]. Beijing: Science Press, 2008: 21—26.[朱阿兴, 李宝林, 裴韬, 等.精细数字土壤普查模型与方法[M].北京: 科学出版社, 2008: 21—26]

[4] Huang W, Luo Y, Wang S Q, et al. Knowledge of soil-landscape model obtain from a soil map and mapping[J]. Acta Pedologica Sinica, 2016, 53(1): 72—80. [黄魏, 罗云, 汪善勤, 等. 基于传统土壤图的土壤—环境关系获取及推理制图研究[J]. 土壤学报, 2016, 53(1): 72—80.]

[5] Hao C K, Sun X L, Wang H L. Application of generalized linear statistical model to digital soil mapping in typical subtropical hilly region [J]. Journal of Soil Science, 2023, 60(4): 993—1006.[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[J].土壤学报, 2023, 60(4): 993—1006.]

[6] Liu F, Geng X Y, Zhu A X, et al. Soil texture mapping over low relief areas using land surface feedback dynamic patterns extracted from MODIS[J]. Geoderma, 2012, 171/172: 44—52.

[7] Guo S X, Zhu A X, Meng L K, et al. Unification of soil

feedback patterns under different evaporation conditions to improve soil differentiation over flat area[J]. International Journal of Applied Earth Observation and Geoinformation，2016，49：126—137.

[ 8 ] Xu Y M，Li B，Shen X B，et al. Digital soil mapping of soil total nitrogen based on Landsat 8，Sentinel 2，and WorldView-2 images in smallholder farms in Yellow River Basin，China[J]. Environmental Monitoring and Assessment，2022，194（4）：282.

[ 9 ] Zhang M W，Liu H J，Zhang M N，et al. Mapping soil organic matter and analyzing the prediction accuracy of typical cropland soil types on the northern Songnen plain[J]. Remote Sensing，2021，13（24）：5162.

[ 10 ] Duan M Q，Zhang X G. Using remote sensing to identify soil types based on multiscale image texture features[J]. Computers and Electronics in Agriculture，2021，187：106272.

[ 11 ] Zhang Z H，Ding J L，Wang J Z，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. [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.]

[ 12 ] Shi X，Long R，Dekett R，et al. Integrating different types of knowledge for digital soil mapping[J]. Soil Science Society of America Journal，2009，73（5）：1682—1692.

[ 13 ] Cheng W，Zhu A X，Qin C Z，et al. Updating conventional soil maps by mining soil-environment relationships from individual soil polygons[J]. Journal of Integrative Agriculture，2019，18（2）：265—278.

[ 14 ] Zhang S M，Xu M X，Zhang Z X，et al. Methods of sampling soil organic carbon in farmlands with different landform types on the loess plateau[J]. Journal of Natural Resources，2018，33（4）：634—643.[张圣民，许明祥，张志霞，等. 黄土高原不同地貌类型区农田土壤有机碳采样布点方法研究[J].自然资源学报，2018，33（4）：634—643.]

[ 15 ] Qi F. Knowledge discovery from area-class resource maps：Data preprocessing for noise reduction[J]. Transactions in GIS，2004，8（3）：297—308.

[ 16 ] Demarchi L，Kania A，Ciężkowski W，et al. Recursive feature elimination and random forest classification of natura 2000 grasslands in lowland river valleys of Poland based on airborne hyperspectral and LiDAR data fusion[J]. Remote Sensing，2020，12（11）：1842.

[ 17 ] Kononenko I. Estimating attributes：Analysis and extensions of RELIEF[C]//Machine Learning：ECML-94，1994：171—182.

[ 18 ] Ren J S，Wang R X，Liu G，et al. Partitioned relief-F method for dimensionality reduction of hyperspectral images[J]. Remote Sensing，2020，12（7）：1104.

[ 19 ] Wang Z G，Wang G C，Zhang Y G，et al. Quantification of the effect of soil erosion factors on soil nutrients at a small watershed in the Loess Plateau，Northwest China[J]. Journal of Soils and Sediments，2020，20（2）：745—755.

[ 20 ] Guo F，Xu Z，Ma H H，et al. Estimating chromium concentration in arable soil based on the optimal principal components by hyperspectral data[J]. Ecological Indicators，2021，133：108400.

[ 21 ] Wang T，Wang X F，Liu J Z. The prediction model of moisture content of young *Aquilaria sinensis* leaves based on RFE＿RF algorithm[J]. Journal of Nanjing Forestry University（Natural Sciences Edition），2022，46（4）：177—184. [王甜，王雪峰，刘嘉政. 基于 RFE＿RF 算法的幼龄沉香叶片含水率预估模型[J]. 南京林业大学学报（自然科学版），2022，46（4）：177—184.]

[ 22 ] Chen X Y，Yang K，Wang J S. Extraction of impervious surface in mountainous city combined with sentinel images and feature optimization[J]. Software Guide，2022，21（4）：214—219. [陈鑫亚，杨昆，王加胜. 结合 Sentinel 影像与特征优选的山地城市不透水面提取[J]. 软件导刊，2022，21（4）：214—219.]

（**责任编辑**：檀满枝）