农业环境科学学报
Journal of Agro-Environment Science
ISSN 1672-2043,CN 12-1347/S

《农业环境科学学报》网络首发论文

题目：基于 RF-XGBoost 的土壤镉污染影响因子及空间分布研究
作者：冯锋，王育红，左雨芳
收稿日期：2022-09-02
网络首发日期：2022-10-27
引用格式：冯锋，王育红，左雨芳．基于 RF-XGBoost 的土壤镉污染影响因子及空间分布研究[J/OL]．农业环境科学学报.
https://kns.cnki.net/kcms/detail/12.1347.S.20221027.1016.006.html

网络首发时间：2022-10-27 11:22:50
网络首发地址：https://kns.cnki.net/kcms/detail/12.1347.S.20221027.1016.006.html

基于 RF-XGBoost 的土壤镉污染影响因子及空间分布研究

冯 锋，王育红*，左雨芳

（江苏师范大学地理测绘与城乡规划学院，江苏 徐州 221000）

摘要：为降低土壤重金属研究中人工采样成本，宏观掌握大尺度研究区内土壤重金属污染的空间分布特征，本文以贵阳市、遵义市和毕节市为研究区，镉（Cadmium, Cd）元素为研究对象，提出利用随机森林（Random Forest, RF）分析评估影响因子的贡献率，并根据贡献率进行影响因子筛选后构建极端梯度提升（eXtreme Gradient Boosting, XGBoost）模型，即 RF-XGBoost 模型，用以预测研究区内土壤 Cd 污染的空间分布特征。结果表明，研究区内土壤 Cd 含量平均值仅比贵州省背景值高出 0.023 mg·kg^{-1}，污染程度较低，变异系数为 125.37%，属于强变异；研究区内对土壤 Cd 污染贡献率最高的影响因子为土壤侵蚀程度、高程和年平均气温，贡献率分别为 0.100、0.088 和 0.084，说明在大尺度研究区中自然环境对土壤内 Cd 富集影响最大；RF-XGBoost 模型的精度和稳定性高于 RF 和 XGBoost 模型，准确率提升了 4.66%，Kappa 系数分别提升了 46.34%、4.21%，F1_score 分别提升了 61.42%、69.61%；研究区内土壤 Cd 污染整体程度较低，但在毕节市西南部出现多个中度-中强污染带。研究表明，RF-XGBoost 模型可准确预测大尺度范围的土壤 Cd 污染空间分布，有助于宏观掌握土壤 Cd 污染的空间分布特征，为污染治理修复提供参考。

关键词：土壤重金属污染；随机森林；极端梯度提升；影响因子；空间分布

doi:10.11654/jaes.2022-0884

A study on factors that influence the spatial distribution of soil cadmium pollution based on RF-XGBoost

FENG Feng, WANG Yuhong*, ZUO Yufang

(School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, 221000, China)

Abstract: To reduce the cost of manual sampling in the study of soil heavy metals and understand the spatial distribution characteristics of soil heavy metal pollution over a large-scale study area, our research examines Cadmium (Cd) levels in the Guiyang, Zunyi, and Bijie regions. We proposed the use of Random Forest (RF) analysis to evaluate the contribution rate of influencing factors. The eXtreme Gradient Boosting (XGBoost) model, namely the RF-XGBoost model, was constructed after screening influencing factors according to the contribution rate to predict the spatial distribution characteristics of soil Cd pollution in the study area. The results showed that the average soil Cd content in the study area was only 0.023 mg/kg higher than the background value in Guizhou province. The results also showed that the pollution degree was low, and the coefficient of variation was a strong 125.37%. The highest contributing factors to soil Cd pollution in the study area were soil erosion degree, elevation, and annual average temperature, with contribution rates of 0.100, 0.088, and 0.084, respectively. This indicates that the natural environment has the greatest impact on soil Cd enrichment in the study area. The accuracy and performance of the RF-XGBoost model were higher than that of the RF and XGBoost models, with accuracy increased by 4.66%, Kappa coefficient increased by 46.34% and 4.21%, and F1_score increased by 61.42% and

收稿日期：2022-09-02
作者简介：冯锋（1995-），男，江苏宿迁人，硕士研究生，从事基于 GIS 的土壤重金属空间分析。E-mail:1553930559@qq.com
***通信作者：**王育红 E-mail: wyhhyk@126.com
基金项目：国家自然科学基金项目（U1304401）；江苏省研究生科研与创新计划项目（KYCX21_2572）
Project supported：The National Natural Science Foundation of China (U1304401); Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_2572)

69.61%, respectively. The overall degree of soil Cd pollution in the study area is low, but there are several moderate to moderately strong pollution zones in the southwest of Bijie City. The results show that the RF-XGBoost model can accurately forecast the spatial distribution of soil Cd pollution at a large scale. This aids our understanding of the spatial distribution characteristics of soil Cd pollution at a macro level and provides a reference for pollution control and remediation.

Keywords: soil heavy metal pollution; Random Forest; eXtreme Gradient Boosting; influencing factors; spatial distribution

镉（Cadmium, Cd）是一种重金属元素，具有非生物降解性，在土壤中富集不但改变土壤理化性质，直接降低农作物产量，同时造成农作物 Cd 含量超标，在经食物链进入人体后，会导致人体骨骼严重软化，并对肝脏、肾脏造成损害[1-2]。据调查显示，我国约有16.7%的农田土壤遭到重金属污染，其中 40%以上为 Cd 污染[3]。因此为有效防范治理土壤重金属污染，我国相继制定出台了《土壤污染防治行动计划》《中华人民共和国土壤污染防治法》等一系列政策、计划和法律法规，以期“守护土壤健康，助力高质发展”。

土壤 Cd 污染具有极强的空间变异性，即使是相邻区域，由于各种因素影响，污染也不尽相同，因此当前对于土壤 Cd 污染的研究多集中于小尺度的研究区，并应用经过实地布点采样、实验室分析获取不同点位的 Cd 含量进行污染特征及分布评价、污染来源解析以及影响因子分析的模式，对于中大尺度研究区，由于区域面积过大，受限于人力、物力、财力，往往难以实现[4]。但大尺度的研究分析往往能揭示土壤 Cd 污染宏观上的空间分布规律，同时在大尺度上进行污染的影响因子分析对污染的防治具有重要的决策支撑作用。传统的土壤重金属污染空间分布预测方法包括反距离加权、径向基函数、克里格插值等，但上述方法仅根据未知点到样本点的距离估算污染状况，而重金属在土壤内的富集受多种因子影响，包括众多自然因子（成土母质、成土过程、土壤质地等）和人为因子（交通污染、工业污染、农业施肥灌溉等），因此地统计学等插值方法预测结果往往精度较低[5-6]。近年来，随着机器学习的成熟和普及，支持向量机、随机森林、神经网络等算法被应用于土壤重金属污染的空间分布研究中，如胡梦珺[7]等采用随机森林模型预测了兰州市主城区校园地表灰尘中重金属污染程度，并与传统的空间插值结果进行了对比；范俊楠等[8]利用 BP 神经网络对湖北省重点区域行业企业周边的土壤重金属污染进行预测；Omondi 等[9]使用随机森林模型对内罗毕和特里尔卡河汇流区的土壤重金属进行了预测，并计算了不同影响因子的贡献率。

尽管上述方法相较于传统的预测方法极大提升了准确度，但由于不同研究区内的重要影响因子具有较大差异，在构建预测模型时将影响力极小的影响因子输入模型，将导致预测精度降低。基于此，本文面向贵州省的贵阳、遵义和毕节三市的土壤 Cd 污染，初选 20 个原始影响因子，创新性地利用随机森林（Random Forest, RF）对原始影响因子进行筛选，再利用极端梯度提升（eXtreme Gradient Boosting, XGBoost）算法对筛选后的影响因子进行训练，提高预测精度，以充分了解三市土壤 Cd 污染空间分布状况。

1 材料与方法

1.1 研究区概况

贵阳市、遵义市和毕节市位于贵州西部，地势南高北低，喀斯特地貌面积占一半以上。三市矿产资源丰富，煤、铁、磷等矿产资源储量在全国名列前茅。长期以来，三市的经济发展依托于矿产资源的开采，是受到土壤 Cd 污染的典型研究区。

1.2 研究方法

1.2.1 地累积指数法

地累积指数又名 Muller 指数，于 1969 年提出后，被广泛应用于水、土壤、沉积物等

环境中的重金属污染评价中[10]。其计算公式为：

$$I_{geo}=\log_2\left(\frac{C_i}{1.5B_i}\right) \tag{1}$$

式中：I_{geo}为地累积指数；C_i为重金属 i 的实测浓度，mg·kg^{-1}；B_i为重金属 i 的地球化学背景值，mg·kg^{-1}；1.5 为考虑环境差异造成的背景值浮动而加入的修正常数。根据 I_{geo}得分将土壤重金属污染划分为 7 个等级，标准如表 1 所示。

表 1 地累积指数与污染程度等级划分

Table 1 The geo-accumulation index and classification of pollution degree

等级 Pollution grade	地累积指数 I_{geo}	污染程度 Pollution degree
I	$I_{geo}\leq 0$	无污染
II	$0<I_{geo}\leq 1$	轻微污染
III	$1<I_{geo}\leq 2$	中度污染
IV	$2<I_{geo}\leq 3$	中强污染
V	$3<I_{geo}\leq 4$	强污染
VI	$4<I_{geo}\leq 5$	较强污染
VII	$I_{geo}>5$	极强污染

1.2.2 RF 算法

RF 是 2001 年由 Breiman 等[11]在贝尔实验室创立的随机决策森林方法基础上提出的一种集成学习算法。该算法通过自助法（bootstrap）重采样从原始的训练数据内抽取若干样本构建出多个弱学习器——分类回归树（Classification and Regression Tree，CART），经组合汇总后生成强学习器以解决分类或回归两类预测问题。

除了用于预测外，RF 也被广泛用于评估影响因子的特征重要性，即贡献率。特征重要性求解主要基于决策树构建时选择节点的指标 Gini 指数，通过计算每个特征在每一棵决策树上进行节点分割时的 Gini 指数差值，汇总取平均值后，再计算其与所有特征 Gini 指数的总变化值的百分比即为特征重要性[12]。

假设一个训练好的 RF 模型由 T 颗 CART 树组成，RF 所用训练样本共包含 F 个特征自变量$(X_1,X_2,X_3,\ldots,X_F)$，其因变量共有 K 个类别取值。对于 RF 中任意一颗编号为 t（$1\leq t\leq T$）的 CART 树，假设其共包含 N 个节点，则该树任意节点 n（$1\leq n\leq N$）的 Gini 指数 GI_{tn}可用式（2）加以计算。式中，p_{nk}表示节点 n 上样本属于第 k 类（$1\leq k\leq K$）的经验概率值。

$$GI_{tn}=\sum_{k=1}^{k}p_{nk}(1-p_{nk}) \tag{2}$$

则第 i 个特征在节点 n 上分裂前后的 Gini 指数差值为：

$$VIM_{in}=GI_n-GI_l-GI_r \tag{3}$$

式中：VIM 表示特征 i 在节点 n 上的 Gini 指数变化量；GI_l表示节点 n 分裂后的左节点；GI_r表示节点 n 分裂后的右节点。

假设特征 i 在决策树 t 中出现的节点集合为 N，则可求得特征 i 在决策树 t 中的 Gini 指数变化量：

$$VIM_{ti}=\sum_{n\in N}VIM_{in} \tag{4}$$

同理可得在所有决策树 T 中，特征 i 的 Gini 指数变化量总和：

$$VIM_i=\sum_{t=1}^{T}VIM_{ti} \tag{5}$$

最后对所求得的第 i 个特征的 Gini 指数变化量进行归一化可知其特征重要性得分为：

$$VIM'_j=\frac{VIM_j}{\sum_{f=1}^{F}VIM_f} \tag{6}$$

由于 RF 采用有放回的采样并按照预定数目随机选择参与决策树的特征，对异常值和噪声具有较高的容忍度，不容易出现过拟合，整体上具有较强的泛化能力、数据挖掘能力、很高的预测准确率，曾被誉为代表集成学习技术水平的最好算法之一[13]。

1.2.3 XGBoost 算法

XGBoost 是一种优化的分布式梯度提升树算法，其基本思想是通过不断增加新的决策树参与训练以拟合预测值与真实值之间的残差，并利用集成思想获得最终的预测值[14]。其计算公式如下：

$$y_i = \sum_{k=1}^{K} f_k(x_i) \tag{7}$$

式中：f_k为第 k 个基学习器；y_i 为第 i 个样本的预测值。

XGBoost 算法为提高预测精度，构建损失函数 L 代表模型的偏差，同时与梯度提升树不同的是，XGBoost 算法引入正则项 Ω 以抑制模型复杂度，故最终得到的目标函数为：

$$Obj = L + \sum_{k=1}^{K} \Omega(f_k) \tag{8}$$

$$L = \sum_{i=1}^{n} l(y_{i,} y_i) \tag{9}$$

式中：$\Omega(f_k)$ 为第 k 棵决策树的正则项；$l(y_{i,} y_i)$ 为第 i 个样本的预测误差。

假设 XGBoost 训练完成后共生成了 K 棵决策树，则将每个样本结构映射到每棵决策树的叶节点上的分数相加即可得到预测值，公式如下：

$$F = \{f(x) = \omega_{q(x)}\}(q: R^f \to T, \omega \in R^T) \tag{10}$$

式中：F 为预测值；$f(x)$为某一棵决策树的模型；$\omega_{q(x)}$ 为决策树 q 的所有叶节点分数组成的集合；T 为决策树 q 的叶节点数量。

1.3 数据描述

本文土壤 Cd 含量数据来源于文献[15]提供的附件资料，该附件以经纬度及数值方式收集了我国土壤 As、Cu、Pb、Cr、Zn 和 Cd 含量数据。该文献经筛选后收集我国 2006—2016 年内公开发表的 2450 篇文献内土壤重金属含量数据，收集的文献中的样本点布设方法大多基于随机采样，土壤样品数据采集深度为 0~20 cm，土壤颗粒大小集中于 0.5~4 mm，Cd 含量采用原子吸收分光光度法测定。提取附件内贵阳市、遵义市和毕节市的土壤 Cd 含量数据进行预处理，根据 Z-score 剔除离群值点，最终得到 170 个 Cd 样本点数据，如图 1 所示。

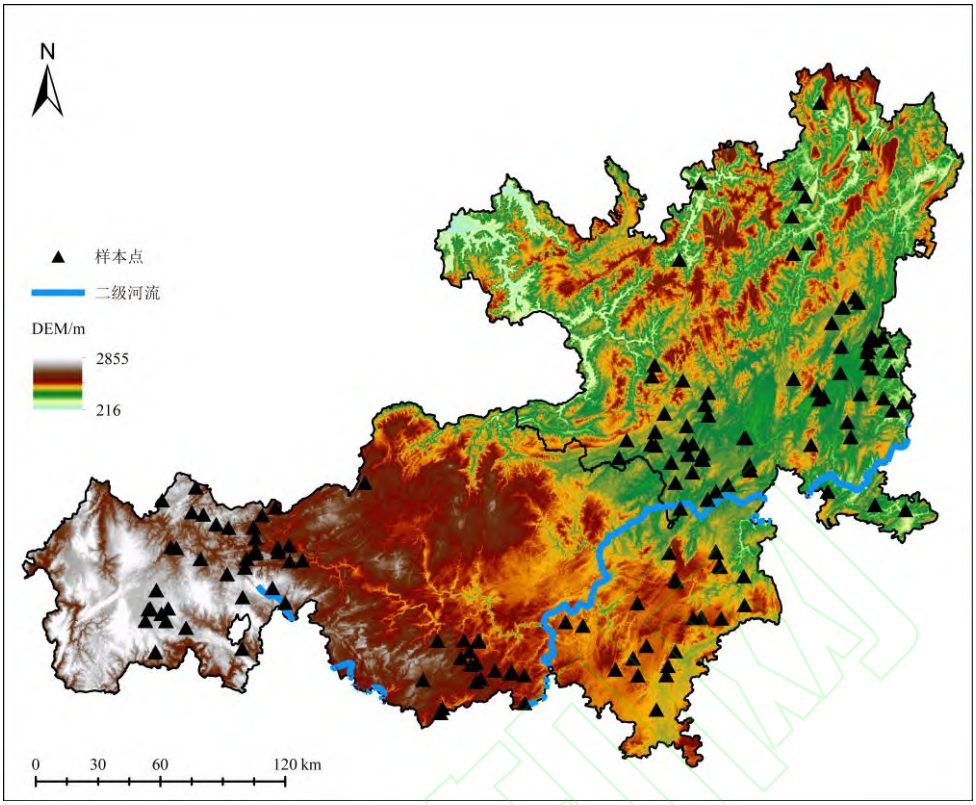


图 1 样本点分布图

Figure 1 Sample points distribution map

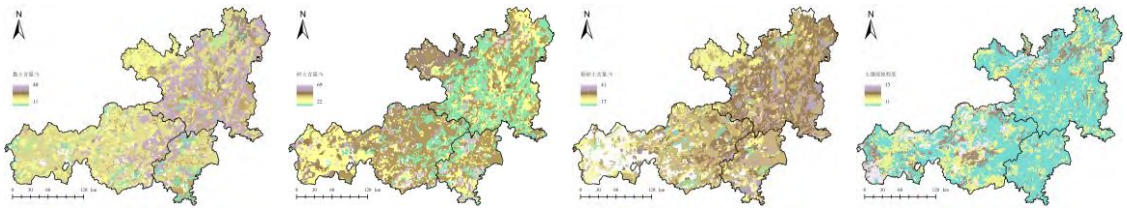
通过阅读文献[16-20]以及考虑到数据获取的难易程度，笔者选择土壤质地、年平均气温、植被指数、高程、河流、土地利用类型、国民生产总值、人口密度、道路等影响因子，为便于数据的查找和使用，将数据组织成如下格式：

表 2 样本点数据基本结构与形式

Table 2 Basic structure and form of sample points data

OID	CON	F01	F02	…	F10	F11	F12	…	F20
1	0.02	43	24		317.701	12.996	170708		341
…	…	…	…	…	…	…	…	…	…
170	4.1	12	62		110.344	2.786	123958		303

表 2 中，OID 为每个样本点的唯一 ID 编号。CON 为 Cd 含量，单位为：$mg \cdot kg^{-1}$。F01-F13 为重金属含量自然影响因子；F14-F20 为人为影响因子。F01 为黏土含量；F02 为砂土含量；F03 为粉砂土含量；F04 为土壤侵蚀度；F05 为年均气温；F06 为年均湿润指数；F07 为年均干燥度指数；F08 为归一化植被指数；F09 为高程；F10 为坡向；F11 为坡度；F12 为与一级河流距离；F13 为与二级河流距离；F14 为土地利用类型；F15 为人口密度；F16 与主干道距离；F17 为与次干道距离；F18 为与高速距离；F19 为与铁路距离；F20 为人均 GDP。各影响因子空间分布特征如图 2 所示。



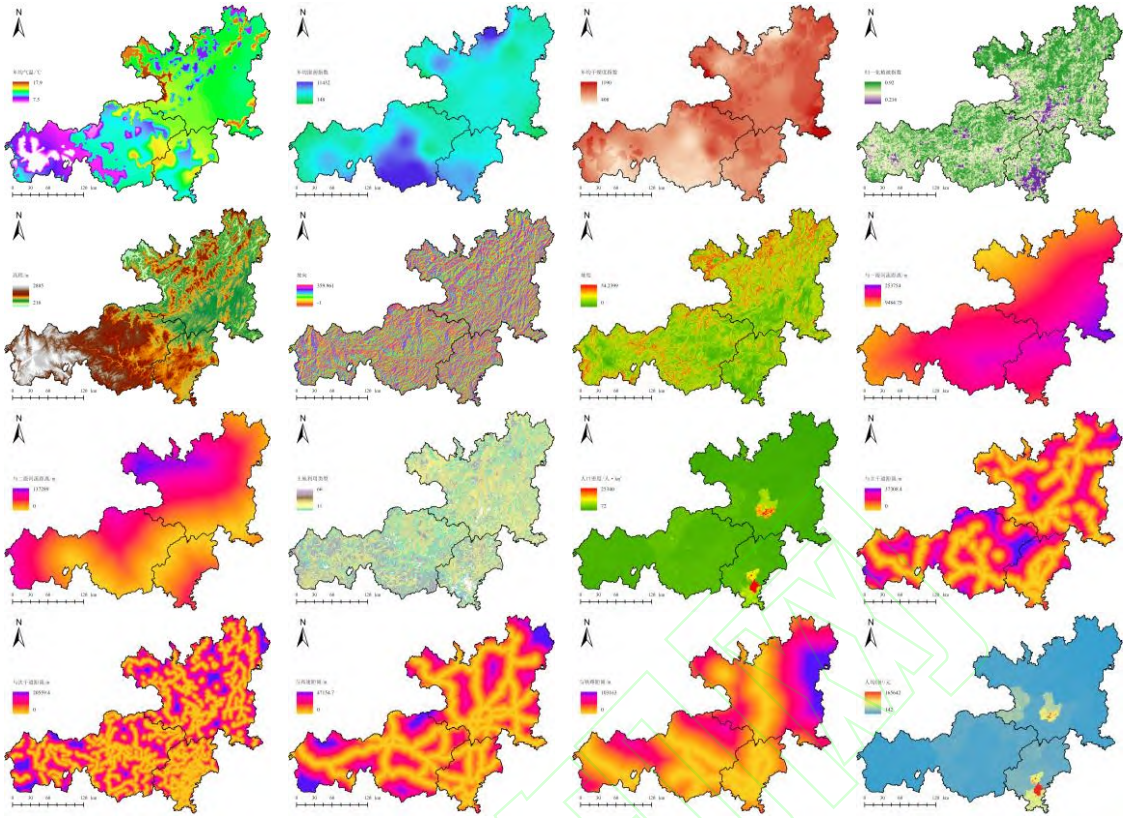


图 2 影响因子分布图

Figure 2 Influence factors distribution map

2 结果与讨论

2.1 描述性统计分析

研究区内 Cd 的最小值为 0.02 mg·kg^{-1}，最大值为 4.10 mg·kg^{-1}，平均值为 0.682 mg·kg^{-1}，标准差为 0.855，变异系数为 125.37%。贵州省的土壤 Cd 背景值为 0.659 mg·kg^{-1}，贵阳、遵义和毕节三市的平均值仅比背景值高出 0.023 mg·kg^{-1}，表明整体上研究区内土壤 Cd 污染程度不高。变异系数反应样本分布的空间均衡性，研究区内 Cd 的变异系数为 125.37%，说明 Cd 的分布极为不均衡，受到较大的外源影响，表明了研究区内的土壤 Cd 富集受到人类活动影响。

2.2 基于 RF 的影响因子贡献率分析

对收集的 170 个样本点计算地累积指数，分级结果如表 3 所示。研究区内 77.65%的样本点无污染，轻微污染、中度污染和中强污染分别占 14.71%、5.88%、1.76%。

表 3 研究区土壤 Cd 污染地累积指数分级

Table 3 Classificathion of soil Cd based on geo-accumulation index in study area

等级 Pollution grade	地累积指数范围 I_{geo} range	各级样本数量 Number of samples at all levels	所占比例/% Proportion/%
I	-5.627 - -0.201	132	77.65
II	0.002 - 0.958	25	14.71
III	1.017 - 1.811	10	5.88
V	2.017 - 2.052	3	1.76

依据 I_{geo} 等级为样本点设置标签，I、II、III、V 级的标签分别为 0、1、2、3。对 170 个样本点，选择 70%作为训练集，余下的 30%样本点作为测试集导入 RF 模型内，模型的主要参数如下：*n_estimators*=100，*criterion*='*gini*'，*max_features*='*sqrt*'，*random_state*=2022，*n_jobs*=-1，基于 python 语言利用 PyCharm 编译器实现模型。得到的各影响因子的贡献率如图 3 所示。

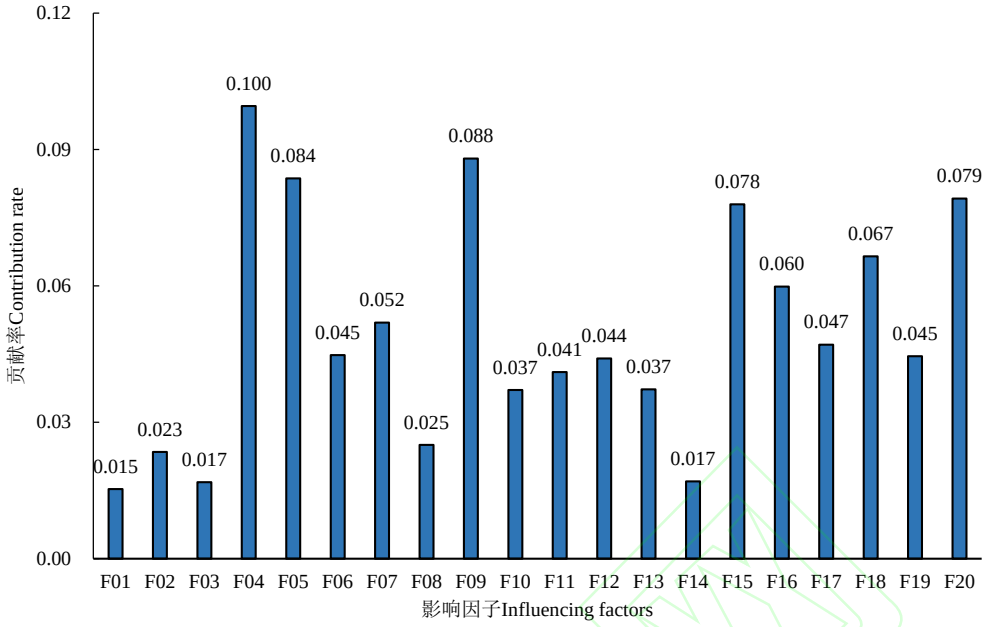


图 3 各影响因子贡献率

Figure 3 Contribution rate of each influencing factor

由图 3 可知：

（1）贡献率排名前三的影响因子分别为土壤侵蚀程度（0.100）、高程（0.088）和年平均气温（0.084）。土壤侵蚀程度对土壤 Cd 污染的贡献率最高，表明土壤侵蚀程度与研究区土壤 Cd 污染的分布最为密切，分析原因在于研究区内多为喀斯特地貌，具有土层浅薄、土被连续性差的特点，大大降低了土壤对于 Cd 元素的吸收、吸附能力，在雨水作用下，含 Cd 的颗粒悬浮物易随泥沙和地表径流进行迁移；高程对土壤 Cd 污染的贡献率位居第二，主要原因在于研究区内山地较多，地势起伏变动大，而相对较高的地形将对 Cd 的迁移扩散起到阻隔作用；年平均气温对土壤 Cd 污染贡献率位居第三，分析原因在于贵州境内气温较高地区降水丰富，地表径流量大，将导致 Cd 扩散。

（2）除上述三项影响因子外，贡献率超过 0.5 的影响因子还有人均 GDP（0.079）、人口密度（0.078）、与高速距离（0.067）、与主干道距离（0.060）以及年均干燥度指数（0.052）。其中人为影响因子包括 4 项，说明研究区内土壤 Cd 污染受到人为影响较为严重，印证了研究区内 Cd 含量变异系数大，受到外源性影响大，且交通为主要污染源之一。年均干燥度指数是蒸发量与降水量的比值，用于衡量地区的气候干燥程度，环境越干燥，表明蒸发量越大，则土壤内的水分越少，将抑制 Cd 随水分扩散。

（3）剩余 12 项影响因子贡献率均低于 0.05，涉及到植被、土壤质地、河流等，说明在研究区内，这些因子与土壤 Cd 污染的空间分布关联性较低。

（4）研究区内对土壤 Cd 污染贡献率排名前三的均为自然因子，且贡献率超过 0.5 的影响因子中，自然影响因子（0.324）高于人为影响因子（0.284），说明在大尺度研究区内，自然因子对土壤 Cd 污染分布起决定性作用。

2.3 影响因子筛选

准确率（Accuracy，ACC）是指测试样本中模型预测“正确”的样本所占比重，其值在[0,1]范围内，值越大，模型的分类结果越准确；Kappa 系数用于判断分类模型分类结果的好坏，是衡量模型分类精度的重要指标，其值[-1,1]范围内，值越接近 1，分类结果越好。将两者结合进行影响因子的筛选有助于降低评估的误差。其计算公式如下：

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{11}$$

$$Kappa = \frac{ACC - p_e}{1 - p_e} \tag{12}$$

$$p_e = \frac{(TP + FP) \times (TP + FN) + (TN + FN) \times (TN + FP)}{(TP + TN + FP + FN)^2} \tag{13}$$

式中：*TP*、*TN*、*FP*、*FN* 分别为真正、真负、假正、假负的样本数。

将所有影响因子的贡献率按从低到高进行排序，从贡献率最低的因子进行筛选，每筛选掉一个影响因子则重新构建一个新的 RF 模型。以上文 170 个样本点作为基础，在保证训练集、测试集和模型各参数不变的前提下，迭代构建 RF 模型，并记录下每一个模型的 ACC 和 Kappa 系数，探索最优的影响因子集合。实验记录的 ACC 和 Kappa 系数如表 4 所示。

表 4 实验过程记录

Table 4 Record of experimental process

实验编号 Experiment Number	删除影响因子 Delete the influencing factor	准确率 ACC	卡帕系数 Kappa
1	无	0.8235	0.502169
2	黏土含量	0.8039	0.392133
3	黏土含量、粉砂土含量	0.8039	0.427609
4	黏土含量、粉砂土含量、土地利用类型	0.8431	0.562232
5	黏土含量、粉砂土含量、土地利用类型、砂土含量	0.8235	0.501087
6	黏土含量、粉砂土含量、土地利用类型、砂土含量、归一化植被指数	0.8039	0.462025
7	黏土含量、粉砂土含量、土地利用类型、砂土含量、归一化植被指数、坡度	0.8235	0.501087
8	黏土含量、粉砂土含量、土地利用类型、砂土含量、归一化植被指数、坡度、与二级河流距离	0.8235	0.500675

由表 4 可知，从实验 1 到实验 8，随着影响因子输入的减少，ACC 和 Kappa 系数呈现上下波动，具体表现为从实验 1 到实验 4，Kappa 系数总体呈现逐步升高的趋势，从实验 5 到实验 8，Kappa 系数呈现稳定趋势，受限于样本量，从实验中无法看出 ACC 的变化规律，但在实验 4 时，ACC 和 Kappa 系数同时达到最大值，说明随着影响因子的筛选，输入模型内的干扰信息逐步减少，输入的数据集得到优化，即删除掉黏土含量、粉砂土含量、土地利用类型 3 项影响因子后，模型的准确度最高，一致性越好。

2.4 XGBoost 结果分析

以筛选优化后得到的 17 个影响因子为基础，引入 XGBoost 模型进一步预测贵阳、遵义和毕节三市土壤 Cd 污染的分级。经反复测试后模型的主要参数设置如下：*booster*='*gbtree*'，*n_estimators*=100，*num_class*=4，*max_depth*=5，*subsample*=0.6，*min_child_weight*=1，*learning_rate*=0.1，*seed*=12。为验证 RF- XGBoost 模型的性能，除选取 ACC、Kappa 系数两个指标衡量模型精度外，再次引入 F1_score 指标同经影响因子筛选后的 RF 以及未经影响因子筛选的 XGBoost 模型进行稳定性对比。F1_score 的计算公式如下：

$$F1\_score = \frac{2 \times precision \times recall}{precision + recall} \tag{14}$$

式中：*precision* 为精确率；*recall* 为召回率。3 种模型的性能指标对比结果如表 5 所示。

表 5 三种模型的性能指标对比

Table 5 Comparison of performance indexes of the three models

模型 Model	准确率 ACC	卡帕系数 Kappa	平衡 F 分数 F1_score

RF	0.8431	0.3774	0.4077
XGBoost	0.8431	0.5300	0.3880
RF-XGBoost	0.8824	0.5523	0.6581

由表 4 可知，RF-XGBoost 模型的 ACC、Kappa 系数和 F1_score 均高于另外两种模型，其中 ACC 提升了 4.66%，Kappa 系数分别提升了 46.34%、4.21%，F1_score 分别提升了 61.42%、69.61%，表明了 RF-XGBoost 模型在土壤 Cd 污染分级的预测上具有较好的准确性和稳定性。

为充分了解研究区内的 Cd 污染空间分布，基于《土壤环境监测技术规范》中规定，以 2.5 km 的精度在研究区内均匀布设 41950 个预测点，利用训练好的 RF-XGBoost 模型预测各点位的污染分级后，使用地统计学普通克里格插值获取研究区内的污染分布情况，由于普通克里格插值的估算数值为连续数值，故按照[0,0.5)、[0.5,1.5)、[1.5,2.5)、[2.5,3)的间隔使用重分类功能对土壤 Cd 污染空间分布进行调整，最终得到研究区内 Cd 污染的分布图，如图 4 所示。由图可知研究区内的 Cd 污染主要呈现斑块状，但总体污染程度较低，污染主要集中于毕节市及遵义市东北方向。其中毕节市出现较大范围的中度-中强污染带，分析原因在于毕节市矿产资源丰富，矿业经济为该市的支柱经济，并且毕节市的矿产开采历史长，导致 Cd 长期富集，形成较强污染。此外，毕节市土壤侵蚀程度较高，Cd 易随地表径流进行迁移扩散，又因该市的地势起伏较大，山脉较多，对 Cd 迁移扩散有阻隔作用，导致 Cd 富集，从而形成中度-中强污染带。

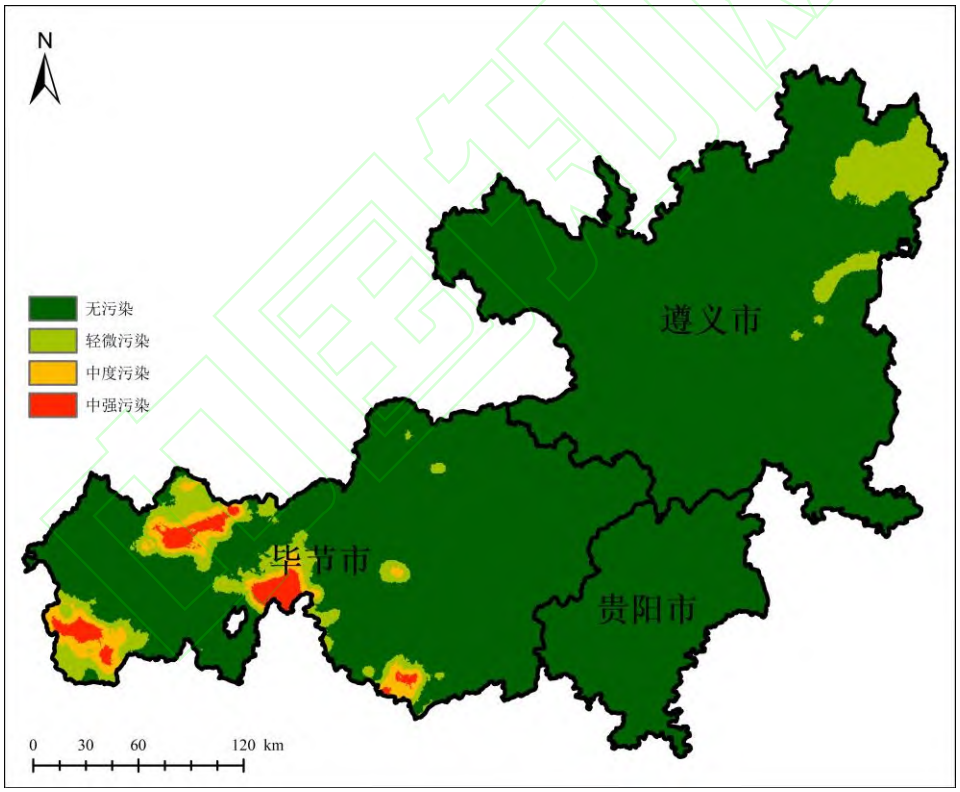


图 4 研究区 Cd 污染分布图

Figure 4 Distribution map of Cd pollution in study area

3 结论

（1）研究区的土壤 Cd 含量平均值略高于贵州省的背景值，整体污染程度较低，但土壤 Cd 污染分布极不均衡，受到较大的外源性影响。

（2）土壤侵蚀程度、高程和年平均气温 3 项自然影响因子对研究区土壤 Cd 污染贡献率最高，揭示在较大尺度研究区内自然环境是造成土壤 Cd 富集的主要影响因素；人均 GDP、人口密度、与高速距离、与主干道距离 4 项人为影响因子对土壤 Cd 污染贡献率较高，印证研究区内的土壤 Cd 含量受到较大外源性影响，并提示在研究区内，人类活动是

造成土壤 Cd 污染的重要来源，其中交通污染源需重点防治。

（3）RF-XGBoost 模型的精度和稳定性显著高于 RF、XGBoost 模型，结合地统计学可准确客观地预测土壤 Cd 污染空间分布特征，该模型在土壤重金属污染空间分布研究中具有可推广性，为宏观掌握大尺度范围内土壤重金属污染分布提供参考。

参考文献：

[1] 黄志伟,李文静,李伟杰,等.东江流域土壤重金属污染特征及潜在风险评价[J].农业环境科学学报,2022,41(3):504-515.
HUANG Z W, LI W J, LI W J, et al. Evaluation of heavy metal pollution characteristics and the potentical risks of soil in the Dongjiang River basin[J]. *Journal of Agro-Environment Science*, 2022,41(3):504-515.

[2] 崔姗姗,李占彬,朱平,等.贵州遵义地区镉大气沉降通量与表层土壤分布特征[J].环境化学,2022,41(4):1324-1334.
CUI S S, LI Z B, ZHU P, et al. Atmospheric deposition flux of cadmium and distribution characteristics of surface soil in Zunyi, Guizhou [J]. *Environmental Chemistry*,2022,41(4):1324-1334.

[3] 全国土壤污染状况调查公报 [J].中国环保产业,2014(5):10-11.
Report on the national general survey of soil contamination[J]. *China Environmental Protection Industry*,2014(5):10-11.

[4] 宋金茜, 朱权, 姜小三, 等. 基于 GIS 的农业土壤重金属风险评价研究——以南京市八卦洲为例[J]. 土壤学报, 2017, 54(1):81-91.
SONG J Q, ZHU Q, JIANG X S, et al. GIS-Based heavy metals risk assessment of agricultural soils: A case study of Baguazhou, Nanjing[J]. *Acta Pedologica Sinica*,2017,54(1):81-91.

[5] 倪碧珩,陆胤,施维林.土壤重金属元素含量的预测方法仿真研究[J].计算机仿真,2022,39(5):234-237, 392.
NI B H, LU Y, SHI W L. Simulation study on prediction method of soil heavy metal content[J]. *Computer Simulation*,2022,39(5):234-237, 392.

[6] 沈洪艳,安冉,师华定,等.湖南省某典型流域农用地土壤重金属污染及影响因素[J].环境科学研究,2021,34(3):715-724.
SHEN H Y, AN R, SHI H D, et al. Heavy Metal pollutin and influencing factors of agricultural land in a typical watershed in Hunan Provine[J]. *Research of Environmental Sciences*,2021,34(3):715-724.

[7] 胡梦珺,王佳,张亚云,等.基于随机森林评价的兰州市主城区校园地表灰尘重金属污染[J].环境科学,2020,41(4):1838-1846.
HU M J, WANG J, ZHANG Y Y, et al. Assessment of Heavy Pollution in Surface Dust of Lanzhou schools based on Random Forests[J]. *Environmental Science*,2020,41(4):1838-1846.

[8] 范俊楠,张钰,贺小敏,等.基于 BP 神经网络的重点行业企业周边土壤重金属污染预测及评价[J].华中农业大学学报,2019,38(4):55-62.
FAN J N, ZHANG Y, HE X M, et al. BP neural network based prediction and evaluation of heavy metal pollution in soil around the enterprises in key areas of Hubei Province[J]. *Journal of Huazhong Agricultural University*,2019,38(4):55-62.

[9] OMONDI E, BOITT M. Modeling the spatial distribution of soil heavy metals using Random Forest model: A case study of Nairobi and Thirirka Rivers' Confluence[J]. *Journal of Geographic Information System*,2020,12(1):597-619.

[10] 胡永兴,宿虎,张斌,等.土壤重金属污染及其评价方法概述[J].江苏农业科学,2020,48(17):33-39.
HU Y X, SU H, ZHANG B, et al. Soil heavy metal pollution and its evaluation methods: a review[J]. *Jiangsu Agricultural Sciences*,2020,48(17):33-39.

[11] BREIMAN L. Random forests[J]. Machine Learning, 2001,45(1):5–32.

[12] 林奕晨,周鹏,潘悦,等.荆州市洪涝灾害影响因子探究及风险评估——基于随机森林和 XGBoost 算法[J].中国农村水利水电,2022(6):125-132.
LIN Y C, ZHOU P, PAN Y, et al. Influencing factor research and risk assessment of flood disasters in Jingzhou City: Based on Random Forest and XGBoost algorithm[J]. *China Rural Water and Hydropower*,2022(6):125-132.

[13] 赖成光,陈晓宏,赵仕威,等.基于随机森林的洪灾风险评价模型及其应用[J].水利学报,2015,46(1):58-66.
LAI C G, CHEN X H, ZHAO S W, et al. A flood risk assessment model based on Random Forest and its application[J]. *Journal of Hydraulic Engineering*,2015,46(1):58-66.

[14] SURAJ K B, TIYASHA T, SALIH M A, et al. Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models[J]. *Environmental Pollution*,2020,268(Pt B):115663.

[15] DUAN Q, LEE J, LIU Y, et al. Distribution of heavy metal pollution in surface soil samples in China: A graphical review[J]. *Bulletin of Environmental Contamination and Toxicology*,2016,97(3):303-309.

[16] 戴倩倩,徐梦洁,庄舜尧,等.基于地理探测器的封丘县农田土壤重金属分布影响因素研究[J].土壤,2022,54(3):564-571.
DAI Q Q, XU M J, ZHUANG S R, et al. Study on factors influencing heavy metal of farmland soils based on geographical detector in Fengqiu County[J]. *Soils*,2022,54(3):564-571.

[17] 杨其坡,武伟,刘洪斌.基于地形因子和随机森林的丘陵区农田土壤有效铁空间分布预测[J].中国生态农业学报,2018,26(3):422-431.
YANG Q P, WU W, LIU H B. Prediction of Spatial distribution of soil available iron in a typical hilly farmland using terrain attributes and random forest model[J]. *Chinese Journal of Eco-Agriculture*,2018,26(3):422-431.

[18] 秦元礼,张富贵,彭敏,等.云南省宣威市农耕区土壤重金属元素分布影响因素及生态风险评价[J].地质与勘探,2022,58(2):360-368.
QIN Y L, ZHANG F G, PENG M, et al. Influencing factors and ecological risk assessment of soil heavy metals in agricultural areas of Xuanwei City, Yunnan Province[J]. *Geology and Exploration*,2022,58(2):360-368.

[19] 齐杏杏,高秉博,潘瑜春,等.基于地理探测器的土壤重金属污染影响因素分析[J].农业环境科学学报,2019,38(11):2476-2486.
QI X X, GAO B B, PAN Y C, et al. Influence factor analysis of heavy metal pollution in large-scale soil based on the geographical detector[J]. *Journal of Agro-Environment Science*,2019,38(11):2476-2486.

[20] 宋运红,杨凤超,刘凯,等.三江平原耕地土壤重金属元素分布特征及影响因素的多元统计分析[J].物探与化探,2022,46(5):1064-1075.
SONG Y H, YANG F C, LIU K, et al.A multivariate statistical analysis of the distribution and influencing factors of heavy metal elements in the cultivated land of the Sanjiang Plain[J]. *Geophysical and Geochemical Exploration*,2022,46(5):1064-1075.