

基于机器学习方法的花生镉富集系数预测

毕卫冬, 丁昌峰, 周志高, 王兴祥

引用本文:

毕卫冬, 丁昌峰, 周志高, 王兴祥. 基于机器学习方法的花生镉富集系数预测[J]. 农业环境科学学报, 2024, 43(6): 1230–1238.

在线阅读 View online: <https://doi.org/10.11654/jaes.2023-1084>

您可能感兴趣的其他文章

Articles you may be interested in

土壤镉污染北方小麦生产阈值及产区划分初探

管伟豆, 肖然, 李荣华, 刘翔宇, 潘君廷, 黄永春, 张增强, 郭堤

农业环境科学学报. 2021, 40(5): 969–977 <https://doi.org/10.11654/jaes.2020-1151>

东北设施叶菜类蔬菜镉铅污染安全生产分区研究

李想, 龙振华, 朱彦彦, 杨昶, 李明堂

农业环境科学学报. 2020, 39(10): 2239–2248 <https://doi.org/10.11654/jaes.2020-0315>

辽宁某冶炼厂周边农田土壤与农产品重金属污染特征及风险评价

都雪利, 李波, 崔杰华, 李国琛, 王颜红

农业环境科学学报. 2020, 39(10): 2249–2258 <https://doi.org/10.11654/jaes.2020-0434>

基于空间聚类与随机森林的稻米富集镉影响因素筛选研究

郭新蕾, 赵玉杰, 刘潇威, 周其文, 王夏晖, 李志涛, 朱智伟, 张铁亮, 王祖光, 张璠, 孙扬

农业环境科学学报. 2019, 38(8): 1794–1801 <https://doi.org/10.11654/jaes.2019-0096>

基于Cubist多元混合回归的稻米富集Cd模型构建研究

刘佳凤, 田娜娜, 赵玉杰, 周其文, 刘潇威, 袁旭, 郭新蕾

农业环境科学学报. 2018, 37(6): 1059–1065 <https://doi.org/10.11654/jaes.2017-1740>



关注微信公众号, 获得更多资讯信息

毕卫冬, 丁昌峰, 周志高, 等. 基于机器学习方法的花生镉富集系数预测[J]. 农业环境科学学报, 2024, 43(6): 1230–1238.

BI W D, DING C F, ZHOU Z G, et al. Prediction of cadmium bioconcentration factor for peanuts based on machine-learning methods[J].

Journal of Agro-Environment Science, 2024, 43(6): 1230–1238.



开放科学 OSID

基于机器学习方法的花生镉富集系数预测

毕卫冬^{1,2}, 丁昌峰^{1,2*}, 周志高¹, 王兴祥^{1,2,3}

(1. 土壤与农业可持续发展国家重点实验室(中国科学院南京土壤研究所), 南京 210008; 2. 中国科学院大学, 北京 100049; 3. 中国科学院红壤生态实验站, 江西 鹰潭 335211)

摘要:通过采集我国14个省份花生主产区的100组土壤-花生样品,分析土壤-花生系统镉(Cd)污染特征及土壤理化性质,基于机器学习方法建立花生Cd富集系数预测模型并识别花生富集Cd的重要影响因素。结果表明,采集的全国范围土壤样品总体以酸性为主,有60%的土壤pH<6.50,花生籽粒Cd含量均值为0.27 mg·kg⁻¹,富集系数均值达到2.42。利用全国以及南北方产区分组数据构建的随机森林模型($R^2=0.930\sim0.966$)对花生Cd富集系数的预测性能均明显优于相应的多元线性回归模型($R^2=0.471\sim0.657$)。随机森林模型分析结果表明不同区域相对重要性较高的特征变量有所差异,影响北方产区花生Cd富集系数预测最重要的特征变量为土壤游离锰氧化物、游离铁氧化物含量和pH,而影响南方产区花生Cd富集系数预测最重要的特征变量为土壤游离锰氧化物、黏粒、游离铁氧化物和有机质含量。研究表明,相较于传统多元线性回归模型,随机森林模型对花生Cd富集系数的预测性能更为优越,为田间大尺度下土壤-花生系统中Cd的迁移预测提供了新的视角和解决方案。

关键词:土壤;花生;镉;随机森林;预测模型

中图分类号:X53 文献标志码:A 文章编号:1672-2043(2024)06-1230-09 doi:10.11654/jaes.2023-1084

Prediction of cadmium bioconcentration factor for peanuts based on machine-learning methods

BI Weidong^{1,2}, DING Changfeng^{1,2*}, ZHOU Zhigao¹, WANG Xingxiang^{1,2,3}

(1. State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China; 3. Experimental Station of Red Soil, Chinese Academy of Sciences, Yingtan 335211, China)

Abstract: In this study, 100 pairs of soil and peanut samples were collected from 14 provinces in China. The soil-peanut cadmium (Cd) contamination characteristics and soil physicochemical properties were analyzed. Prediction models of the Cd bioconcentration in peanuts were established based on machine-learning methods and the important factors influencing Cd enrichment in peanuts were identified. The results showed that the soil samples collected were mainly acidic, with 60% of the soils being pH<6.5. The average Cd content in peanut kernels was 0.27 mg·kg⁻¹ and the average bioconcentration factor was 2.42. The prediction performance was significantly better for the random forest models ($R^2=0.930\sim0.966$), based on the data for the whole country, and the grouped northern and southern producing areas, than for the corresponding multiple linear regression models ($R^2=0.471\sim0.657$). The results of random forest model analysis showed that the characteristic variables with high relative importance were different in different regions. The most important variables affecting the prediction of Cd bioconcentration in northern producing areas were the free manganese oxide content, free iron oxide content, and pH of the soil, while the most important variables affecting the Cd bioconcentration in southern producing areas were the free manganese oxide, clay, free iron oxide, and organic matter contents of the soil. The results revealed that, compared with the traditional multiple linear regression

收稿日期:2023-12-18 录用日期:2024-03-06

作者简介:毕卫冬(1997—),男,安徽六安人,硕士研究生,研究方向为农产品产地污染风险管控。E-mail:biweidong@issas.ac.cn

*通信作者:丁昌峰 E-mail:cfding@issas.ac.cn

基金项目:国家现代农业产业技术体系项目(CARS-13);国家自然科学基金项目(42077151)

Project supported: The Earmarked Fund for China Agriculture Research System(CARS-13); National Natural Science Foundation of China(42077151)

models, random forest models had better performance at predicting the Cd bioconcentration of peanuts. This provides a new perspective and solution for predicting Cd transfer in soil-peanut systems at a large scale in the field.

Keywords: soil; peanut; cadmium; random forest; prediction model

花生是我国重要的油料和经济作物,是我国居民食用植物油和蛋白质的重要来源。我国是世界花生的生产和消费大国,花生是我国为数不多的具有明显竞争优势的出口创汇农产品^[1]。随着花生直接食用和食品加工需求不断增加,国际上对花生重金属超标问题越来越关注。近年来花生重金属特别是镉(Cd)含量超标等食品安全问题,已经成为我国花生出口贸易的主要制约因素之一^[2]。调查发现花生对土壤中的Cd具有较强的富集能力,按照我国食品安全限量标准(GB 2762—2022, $Cd \leq 0.5 \text{ mg} \cdot \text{kg}^{-1}$),我国花生Cd超标现象在局部地区较为严重。Dai等^[3]于2009—2014年在我国15个花生主产省的调查发现,8 698份花生样品籽粒Cd平均含量为 $0.17 \text{ mg} \cdot \text{kg}^{-1}$,有2.8%的样品Cd含量超出食品安全国家标准的限量值。王珊珊等^[4]在辽宁省调查发现,在土壤Cd含量低于筛选值的花生产区,花生籽粒Cd含量达到了 $0.21 \sim 0.75 \text{ mg} \cdot \text{kg}^{-1}$,超标比例高达31%。如果按照FAO/WHO的国际标准(CXS 193—1995, $Cd \leq 0.10 \text{ mg} \cdot \text{kg}^{-1}$)或欧盟标准(EC 18861/2006, $Cd \leq 0.20 \text{ mg} \cdot \text{kg}^{-1}$),我国局部地区花生Cd超标问题将更为突出。

目前,关于Cd在土壤-作物系统中的迁移预测模型研究主要集中在水稻、小麦、蔬菜等大宗作物^[5],而土壤-花生体系中Cd的迁移预测研究较少。Yang等^[6]基于田间调查采样,建立了广西地质高背景区花生富集Cd的多元线性回归模型。由于大尺度田间环境下污染物-土壤-作物之间复杂的相互作用,传统的多元线性回归模型难以准确预测污染物在土壤-作物系统中的非线性关系。通过大样本数据训练的机器学习方法可实现对目标变量的高精度预测,在面对复杂系统中自变量和因变量的非线性关系以及处理多维特征数据集时均表现出较大潜力。Huang等^[7]通过机器学习,发现基于土壤铁锰氧化物结合态Cd、土壤pH、土壤可还原态锰含量和水稻灌浆成熟后期的土壤含水率,能很好地预测区域尺度稻米Cd含量($R^2=0.81$)。牟力言等^[8]在全国大尺度空间范围内,基于随机森林等机器学习算法识别了不同区域稻米富集Cd、As的重要影响因素。然而机器学习方法在土壤-花生系统中Cd的迁移预测方面还未见报道,我国不同区域花生富集Cd的重要影响因素还不明确。

富集系数(BCF,作物Cd含量与土壤Cd含量的商)可以量化Cd在土壤-作物中的转移过程,表征作物吸收积累Cd的能力^[9]。本研究结合全国范围调查的土壤-花生系统采样数据,利用随机森林模型对全国以及南、北方花生产区花生Cd富集系数进行预测,进一步利用随机森林模型分析特征变量的相对重要性,解析影响Cd在土壤-花生系统中迁移的主要因素,以期为我国花生产地土壤污染风险管控与修复技术研发提供理论依据。

1 材料与方法

1.1 数据获取

综合考虑我国花生主产区分布、潜在土壤污染情况等因素,在我国北方花生主产区吉林(16个采样点)、山东(18个采样点)、山西(9个采样点)、河南(7个采样点)、辽宁(7个采样点)、河北(3个采样点)、新疆(2个采样点)以及南方花生主产区江苏(13个采样点)、安徽(6个采样点)、江西(6个采样点)、广西(5个采样点)、四川(5个采样点)、湖北(2个采样点)、广东(1个采样点)等地共点对点采集土壤-花生样品100组。利用五点取样法分别采集相应花生和表层土壤(0~20 cm)样品,采集后的土壤样品密封存于自封袋中,经风干过筛后保存备用。花生荚果样品经自来水洗净后脱壳,籽粒105℃杀青30 min,60℃烘干至质量恒定,研磨粉碎。土壤pH、有机质、阳离子交换量、黏粒、交换性钙、游离铁氧化物、游离铝氧化物、游离锰氧化物等土壤基本理化性质的分析测试方法详见文献^[10]。土壤全量Cd采用 $\text{HNO}_3 + \text{H}_2\text{O}_2 + \text{HF}$ 高压密封消解^[11],电感耦合等离子体质谱(ICP-MS)(Agilent 7700x,美国)测定,采用国家土壤标准物质(GBW07405a)进行质量控制,测定值均在标准值范围($0.16 \pm 0.03 \text{ mg} \cdot \text{kg}^{-1}$)内。土壤有效态Cd含量采用 $0.005 \text{ mol} \cdot \text{L}^{-1}$ 的DTPA提取^[12],ICP-MS测定。花生籽粒Cd含量采用 $\text{HNO}_3 + \text{H}_2\text{O}_2$ 高压密封消解^[11],ICP-MS测定,采用大米标准物质(GSB-23a)进行质量控制,测定值均在标准值范围($0.32 \pm 0.04 \text{ mg} \cdot \text{kg}^{-1}$)内。

1.2 预测模型建立

多元线性回归模型(Multiple Linear Regression, MLR)是利用多个自变量解释因变量变化的一种常用

模型,通过优化自变量的组合来解释与因变量的线性关系,并且该模型还可以避免自变量多重共线性^[13]。基于土壤理化性质数据,模型由式(1)建立:

$$y=a_1x_1+a_2x_2+\cdots+a_nx_n+b \quad (1)$$

式中: y 是因变量,本研究中代表花生Cd富集系数; x_i ($i=1,2,3,\cdots,n$)是自变量,即土壤理化性质参数; a_i ($i=1,2,3,\cdots,n$)是变量 x_i 的回归系数,反映 x_i 和 y 的关系; b 为常数项。

随机森林模型(Random Forest, RF)是一种基于分类和回归的集成机器学习算法,可用于分析自变量与因变量之间复杂的非线性关系,其特点是利用Bootstrap抽样法,从样本数据中随机抽取 n 个样本数据作为训练集,基于这些数据建立决策树,且重复抽样和训练步骤建立大量决策树,从而最终形成随机森林。由于数据选取具有随机性和决策树参数的可调节性,并且该模型不用对数据的正态分布和独立性等假设条件进行检验,不用考虑变量的共线性问题,因此在处理分类和回归的问题上,RF具有良好的表现^[13]。

MLR和RF模型的性能评价指标选取决定系数(R^2)和均方根误差(RMSE)。特征重要性利用随机森林解释器中节点不纯度基尼系数(Gini)计算,特征变量 x_i 的重要性为决策树节点分支前后基尼指数变化量的归一化数值^[14]。随机森林决策树数量为500,最大深度为8。

1.3 数据分析

数据分析通过SPSS 26.0实现,MLR通过SPSS 26.0中的线性回归模型进行拟合,南、北花生生产区污染特征以及土壤理化性质显著性差异由SPSS 26.0中的独立样本 t 检验分析。RF通过Python 3.9 sklearn包中的函数搭建,可视化过程通过matplotlib包和OriginPro 2019b实现。

2 结果与讨论

2.1 土壤性质、土壤Cd含量与花生Cd含量特征分析

采集的全国100组样品的土壤理化性质、土壤Cd含量、花生Cd含量以及富集系数数据特征见表1,南、北方花生生产区土壤理化性质以及土壤和花生Cd污染特征见图1、图2。全国100组样品的土壤pH在4.03~9.10之间,平均值6.45,其中5%呈极强酸性($\text{pH}<4.5$)、28%呈强酸性($\text{pH} 4.5\sim 5.5$)、27%呈酸性($\text{pH} 5.5\sim 6.5$)、11%呈中性($\text{pH} 6.5\sim 7.5$)、19%呈碱性($\text{pH} 7.5\sim 8.5$)、10%呈强碱性($\text{pH}>8.5$),酸性土壤占60%。南方花生生产区土壤pH平均值(5.96)显著低于北方产

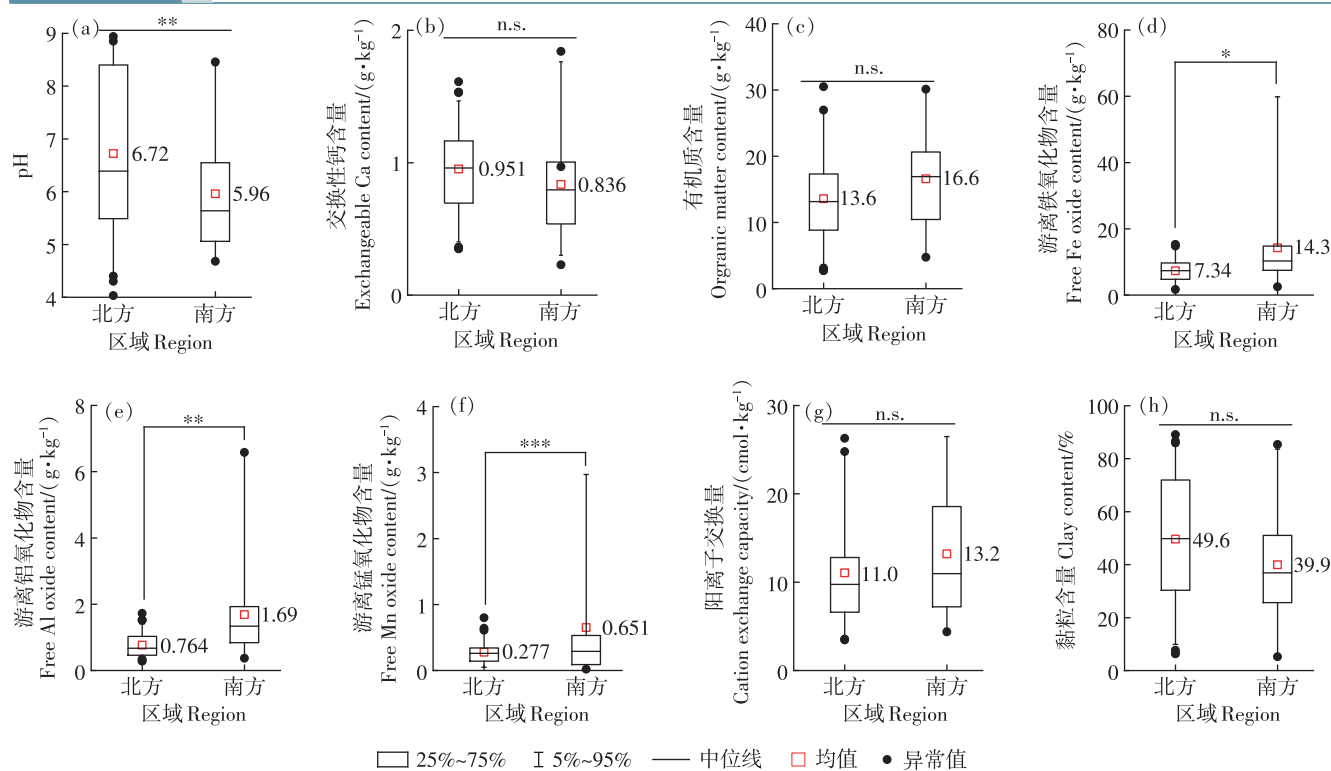
区pH平均值(6.72)。

全国100组样品的土壤阳离子交换量、黏粒、交换性钙、有机质、游离铁氧化物、游离铝氧化物、游离锰氧化物含量的平均值分别为 $11.97 \text{ cmol}\cdot\text{kg}^{-1}$ 、45.93%、 $0.92 \text{ g}\cdot\text{kg}^{-1}$ 、 $14.72 \text{ g}\cdot\text{kg}^{-1}$ 、 $9.89 \text{ g}\cdot\text{kg}^{-1}$ 、 $1.10 \text{ g}\cdot\text{kg}^{-1}$ 、 $0.43 \text{ g}\cdot\text{kg}^{-1}$ 。进一步分析发现,南、北方花生生产区的游离铁氧化物、游离铝氧化物和游离锰氧化物含量存在显著差异,而阳离子交换量、土壤有机质、黏粒含量、交换性钙含量无显著差异。

全国100组样品的土壤全量Cd含量为0.03~1.19 $\text{mg}\cdot\text{kg}^{-1}$,有14%的样品超过我国《土壤环境质量 农用地土壤污染风险管控标准(试行)》(GB 15618—2018)中的风险筛选值。利用 $0.005 \text{ mol}\cdot\text{L}^{-1}$ DTPA提取的土壤有效态Cd含量为0.02~0.20 $\text{mg}\cdot\text{kg}^{-1}$,平均值为0.04 $\text{mg}\cdot\text{kg}^{-1}$ 。全国100组样品的花生籽粒Cd含量为0.01~1.43 $\text{mg}\cdot\text{kg}^{-1}$,有15个采样点位的花生籽粒Cd含量超过我国食品安全限量标准($0.5 \text{ mg}\cdot\text{kg}^{-1}$)。花生籽粒Cd超标的点位中仅有3个土壤Cd含量超标,其余均未超标,这意味着我国现行的土壤污染风险筛选值可能并不能准确评估我国花生生产地Cd污染风险。对比南、北方花生生产区Cd污染特征,发现南方产

表1 全国样品土壤性质、土壤Cd含量与花生Cd含量统计结果
Table 1 Statistical results of soil properties, soil Cd content and peanut Cd content for the whole samples

指标 Parameter	最小值 Minimum	最大值 Maximum	平均值 Mean	标准差 Standard deviation	变异系数 Coefficient of variation
土壤pH	4.03	9.10	6.45	1.43	0.22
土壤有机质/($\text{g}\cdot\text{kg}^{-1}$)	2.73	43.2	14.7	7.33	0.50
土壤阳离子交换量/ ($\text{cmol}\cdot\text{kg}^{-1}$)	3.45	30.7	11.97	6.76	0.56
土壤黏粒/%	5.24	89.1	45.9	24.6	0.54
土壤交换性钙/ ($\text{g}\cdot\text{kg}^{-1}$)	0.23	1.84	0.92	0.35	0.38
土壤游离铁氧化物/ ($\text{g}\cdot\text{kg}^{-1}$)	1.65	87.0	9.89	10.5	1.06
土壤游离铝氧化物/ ($\text{g}\cdot\text{kg}^{-1}$)	0.28	6.58	1.10	1.03	0.94
土壤游离锰氧化物/ ($\text{g}\cdot\text{kg}^{-1}$)	0.02	7.74	0.43	0.85	1.99
土壤全量Cd/ ($\text{mg}\cdot\text{kg}^{-1}$)	0.03	1.19	0.19	0.16	0.84
土壤有效态Cd/ ($\text{mg}\cdot\text{kg}^{-1}$)	0.02	0.20	0.04	0.03	0.82
花生籽粒Cd/ ($\text{mg}\cdot\text{kg}^{-1}$)	0.01	1.43	0.27	0.27	1.00
BCF	0.05	13.2	2.42	2.84	1.18



* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, n.s. 无显著差异 No significant difference, 下同。The same below.

图1 南、北方花生产区土壤理化性质差异

Figure 1 Differences of soil physicochemical properties in peanut producing areas of south and north China

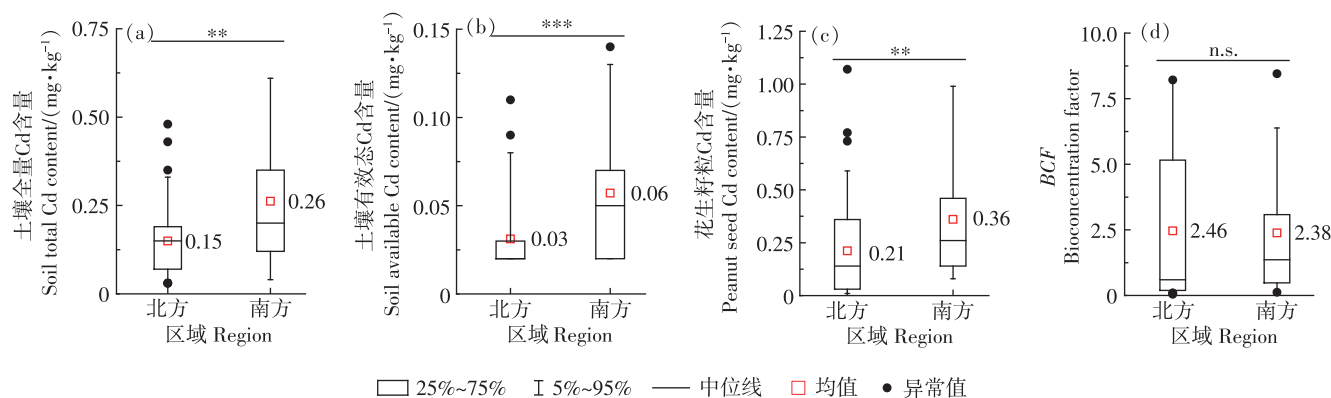


图2 南、北方花生产区Cd污染特征差异

Figure 2 Differences of pollution characteristics in peanut producing areas of south and north China

区土壤全量Cd、有效态Cd以及花生籽粒Cd含量均显著高于北方。

采集的全国100组花生样品中BCF范围为0.05~13.23,平均值为2.42。南、北方花生产区花生BCF无明显差异。Wang等^[2]通过田间试验发现,在pH值6.03、全量Cd含量 $0.31 \text{ mg} \cdot \text{kg}^{-1}$ 的土壤上,种植的19个花生品种Cd富集系数为0.83~3.26,均值为2.08。王飞等^[15]发现在湖南湘江流域酸性Cd污染稻田

($0.70 \text{ mg} \cdot \text{kg}^{-1}$)上种植的5个花生品种Cd平均富集系数达到1.10。Gu等^[16]在广西南宁地区Cd污染土壤($0.33 \text{ mg} \cdot \text{kg}^{-1}$)调查发现,花生对Cd的平均富集系数为1.65。以上文献报道的花生BCF均在本研究的BCF范围内,但本研究的BCF平均值要更高,其原因在于本研究采集的花生产区土壤样品全量Cd含量($0.19 \text{ mg} \cdot \text{kg}^{-1}$)相对更低,由于作物Cd含量并非随着土壤Cd含量的升高而线性升高,导致BCF往往随着

土壤Cd含量的升高而降低^[17]。从本研究结果和以往文献报道都可以看出,花生对Cd的富集能力非常强,这加剧了花生的食品安全风险。

2.2 花生Cd富集系数预测模型

花生籽粒Cd含量、富集系数与土壤Cd含量及理化参数之间的相关性见图3。结果发现土壤有效态Cd含量与pH呈显著负相关,与有机质、游离铁氧化物、游离铝氧化物、游离锰氧化物、阳离子交换量、黏粒含量呈显著正相关关系。花生籽粒Cd含量与土壤有效态Cd含量、游离铝氧化物含量、阳离子交换量、黏粒含量呈显著正相关,与pH、交换性钙含量呈显著负相关。土壤重金属有效态含量往往比总量更能反映重金属的生物富集效应,Liu等^[18]在珠三角区域调查发现,白菜Cd含量与土壤有效态Cd含量的相关性($r=0.673$)明显高于土壤全量Cd($r=0.485$)。花生Cd富集系数BCF则与土壤pH、交换性钙、有机质、游离铁氧化物、游离锰氧化物含量呈显著负相关关系。

对数据采用逐步回归分析的方法建立最优的多元线性回归模型,结果见表2。通过建立的预测方程可知,对于全国100组样品,利用土壤pH和游离铁氧化物含量建立的预测方程回归系数 R^2 为0.471,表明可以解释花生Cd富集系数47.1%的变化。对于北方

和南方产区的分组数据,分别构建出以土壤pH、游离铁氧化物和交换性钙含量为变量的最优预测模型($R^2=0.545, n=62$)和以土壤pH、游离锰氧化物、有机质和黏粒含量为变量的最优预测模型($R^2=0.657, n=38$)。

利用土壤pH、交换性钙、有机质、游离铁氧化物、游离铝氧化物、游离锰氧化物、阳离子交换量、黏粒含量这些特征变量($n=8$)构建随机森林模型对花生Cd富集系数进行预测。结果显示在全国100组样品的花生Cd富集系数预测方面,与最优多元线性回归模型($R^2=0.471$)相比,RF模型($R^2=0.935$)取得了明显更好的预测效果(图2a、图2b),同时RF模型($RMSE=0.789$)的均方根误差也低于多元线性回归模型($RMSE=2.345$)。同时在北方和南方花生产区花生Cd富集系数预测方面,与相应的最优多元线性回归模型相比,RF模型均取得了更好的预测效果和更低的均方根误差(图2c、图2d、图2e、图2f)。牛硕等^[14]基于区域调查,构建了随机森林($R^2=0.583$)和神经网络($R^2=0.490$)的机器学习模型,对小麦Cd富集系数进行预测,发现二者预测性能均优于多元线性回归模型($R^2=0.410$)。Xie等^[19]建立了土壤-水稻系统中Cr、Cu、Zn、Ni的生物富集预测模型,发现随机森林模型($R^2=0.58\sim0.79$)对4种元素的预测效果均优于多元线性回

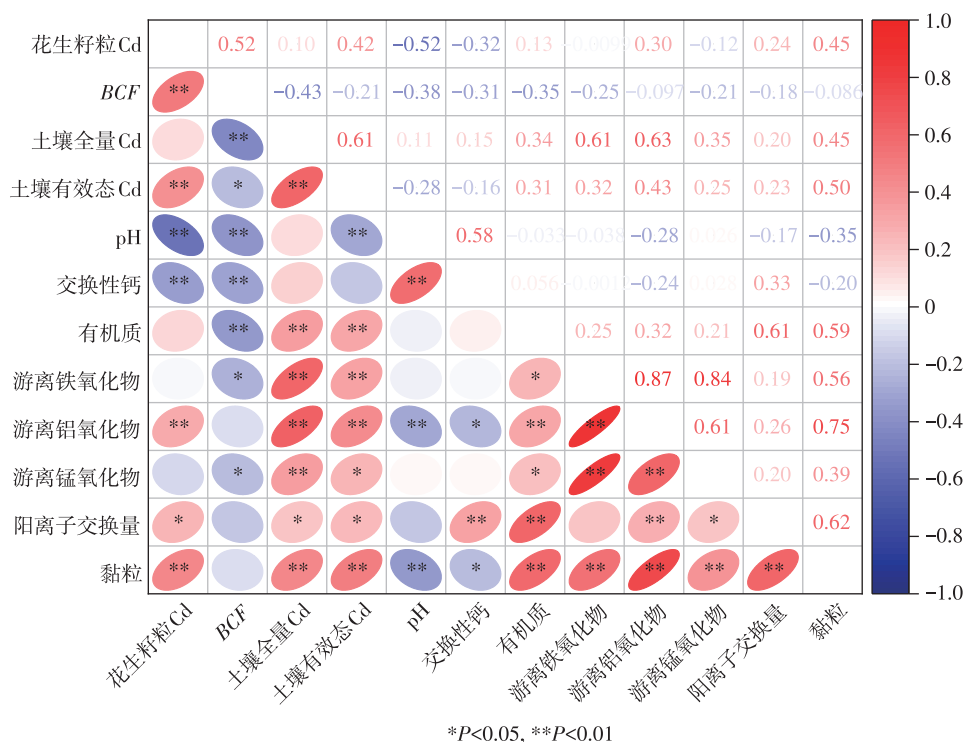


图3 特征变量相关性

Figure 3 Correlation of feature variables

表2 花生Cd富集系数预测模型

地区Region 编号No.		预测模型Prediction model	n	R ²	P	RMSE
全国	2	$\lg BCF=2.51-0.259 \text{ pH}-0.942 \lg[\text{游离铁氧化物}]$	100	0.471	<0.001	2.345
	3	随机森林	100	0.935	<0.001	0.789
北方	4	$\lg BCF=-0.305-0.344 \text{ pH}-1.60 \lg[\text{游离铁氧化物}]+1.30 \lg[\text{交换性钙}]$	62	0.545	<0.001	2.525
	5	随机森林	62	0.930	<0.001	0.930
南方	6	$\lg BCF=2.89-0.185 \text{ pH}-0.419 \lg[\text{游离锰氧化物}]-1.02 \lg[\text{有机质}]-0.461 \lg[\text{黏粒}]$	38	0.657	<0.001	1.533
	7	随机森林	38	0.966	<0.001	0.492

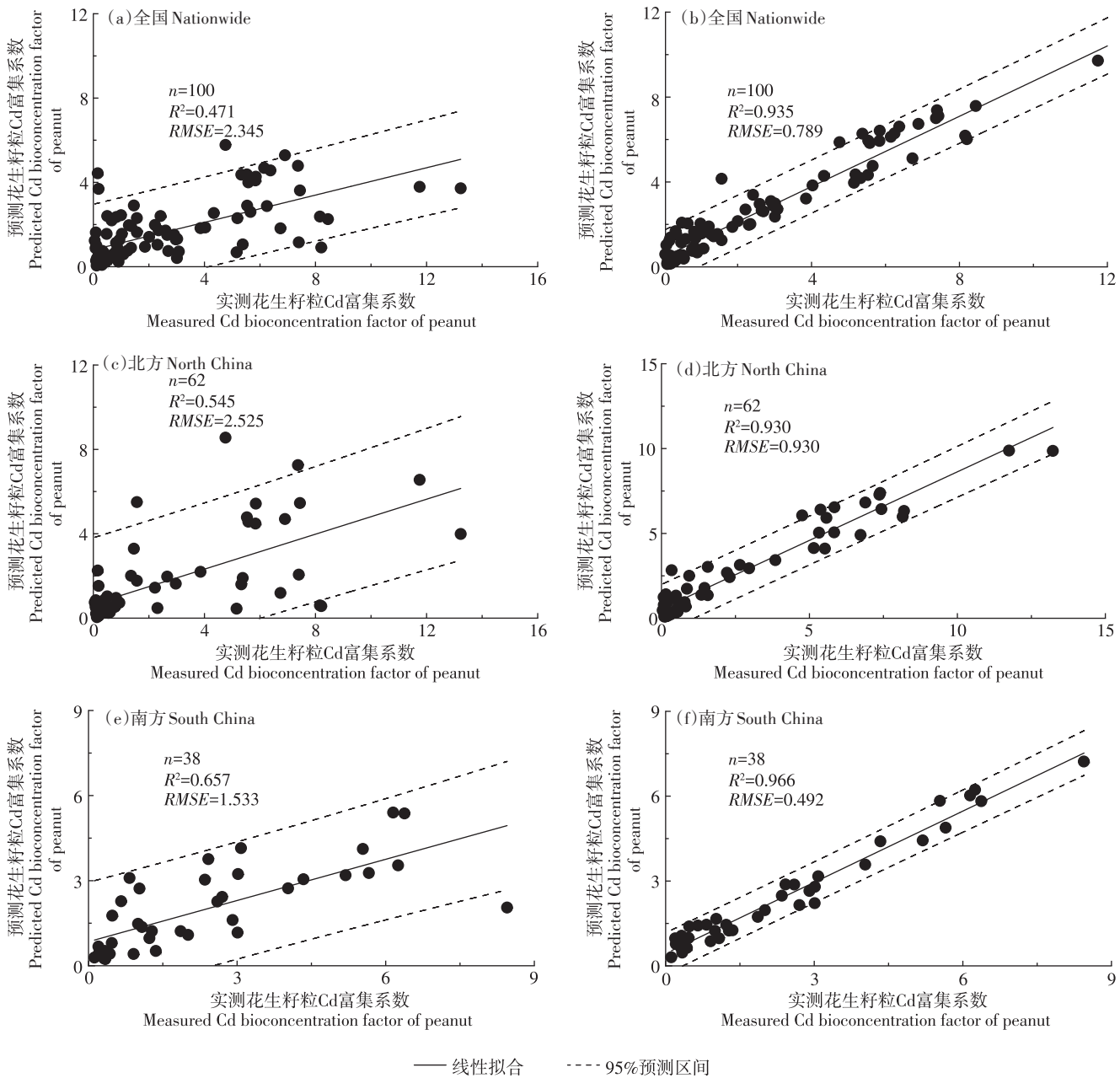


图4 多元线性回归(a,c,e)和随机森林(b,d,f)模型对花生Cd富集系数的实测值与预测值

Figure 4 Measured and predicted Cd bioconcentration factor of peanut using multiple linear regression (a,c,e) and random forest(b,d,f) models

归模型($R^2=0.46\sim0.67$)。这些结果与本研究相似,均表明与传统的多元线性回归模型相比,基于随机森林等算法的机器学习模型表现更为可靠,有较大应用潜力。其原因可能是土壤重金属生物有效性与环境因子之间存在复杂的非线性关系,MLR 仅能构建变量之间简单的线性关系,导致模型预测精度受到限制。而基于集成机器学习算法的RF,具有较强的非线性处理能力,RF 模型预测不需要对数据的正态分布和独立性等假设条件进行检验,也不需要考虑变量的共线性问题^[13],即共线性问题不会对RF模型的预测性产生影响。

2.3 随机森林模型特征变量重要性

应用RF解释器对特征变量重要性进行排序,结果如图5所示。土壤游离铁氧化物含量和pH是影响全国产区花生Cd富集系数预测最重要的2个特征变量,其相对重要性分别为51.5%和16.6%,其余特征变量的相对重要性均低于10%。影响北方产区花生Cd富集系数预测最重要的3个特征变量为土壤游离锰氧化物含量、游离铁氧化物含量和pH,其相对重要性分别为30.0%、29.9%、11.7%,其余特征变量的相对重要性均低于10%。影响南方产区花生Cd富集系数预测最重要的特征变量为土壤游离锰氧化物、黏粒、游离铁氧化物和有机质含量,其相对重要性分别为29.7%、20.8%、14.0%、13.0%,其余特征变量的相对重要性均低于10%。

由于pH可以显著影响土壤中Cd的沉淀/溶解、吸附/解吸等过程,土壤pH通常被认为是所有理化性质中影响土壤Cd有效性最重要的因子之一。低pH条件下由于 Cd^{2+} 与正电荷的相斥作用、Cd形态间的相互转化作用以及Cd与其他金属及非金属离子的竞争作用等,使得有效态Cd的含量迅速增加^[20-21]。Zhao等^[22]通过对江苏省稻米Cd含量的机器学习预测发现,除土壤全量Cd外,土壤pH是影响稻米Cd含量预测的最重要变量,其相对重要性为17.2%。Cd化学形态模型预测表明,当土壤 $\text{pH}<6.5$ 时,Cd主要与土壤有机质结合;当 $\text{pH}>6.5$ 时,铁锰氧化物成为主要吸附位点,土壤中无定形锰氧化物与Cd的结合能力强于铁铝氧化物或氢氧化物。因此,锰氧化物含量高的土壤通常Cd的活性也较低^[23]。土壤中黏粒含量越高,对Cd的吸附固定能力越强,导致作物体内Cd含量降低。陈宏坪等^[24]基于8种水稻土的盆栽试验发现,土壤pH和黏粒含量是影响水稻土Cd临界值的两个最主要因素。我国花生种植的最适宜区土壤黏粒含量

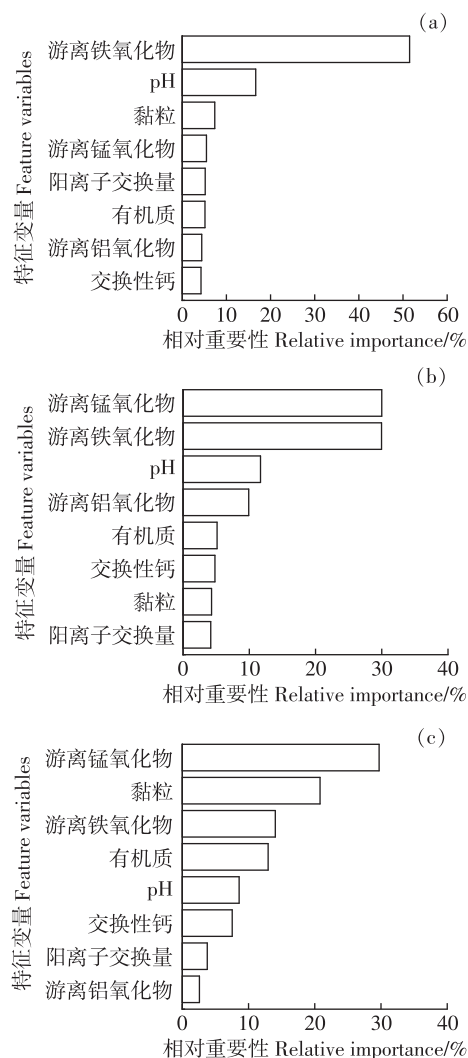


图5 特征变量对预测全国(a)、北方(b)、南方(c)产区花生Cd富集系数的相对重要性

Figure 5 Relative importance of characteristic variables in predicting Cd bioconcentration factor of peanut in nationwide(a), north(b), and south(c) China producing areas

相对较低^[25],这是由于砂性土壤有利于花生荚果膨大和干物质积累,而黏土土壤致密,透气性差,不利于花生根系深扎和荚果干物质积累^[26]。但砂土中黏粒含量较低导致Cd的有效性往往较高,这也可能是引起花生Cd富集能力较强的原因之一。 Ca^{2+} 和 Cd^{2+} 具有同样的化合价,且离子半径极度接近(Ca^{2+} 为 $9.9\times10^{-2}\text{ nm}$, Cd^{2+} 为 $9.7\times10^{-2}\text{ nm}$),它们能够竞争土壤中黏土矿物、氧化物及有机质上的阳离子吸附交换位点,因此,外源Ca可以通过影响土壤中Cd的生物有效性从而调控植物对Cd的吸收^[27]。宋正国等^[28]通过盆栽实验发现施用硝酸钙后,土壤溶液中 $\text{Ca}^{2+}/\text{Cd}^{2+}$ 比值与小油菜体内Cd浓度呈显著正相关,说明土壤中钙浓度升

高会促进Cd的解吸,进而影响小油菜对Cd的吸收。Zhang等^[29]通过水培试验发现硝酸钙增强了水稻体内Cd从根向地上部的转运,表明Ca还可能会促进Cd在作物体内的转运。

比较南方、北方花生产区的MLR模型和RF模型的特征变量,发现不同区域相对重要性较高的特征变量有所差异,可能是由于不同区域土壤理化性质及污染特征差异造成的。本研究发现南方花生产区预测模型中土壤黏粒和有机质含量相对重要性更高,原因可能在于南方花生产区土壤黏粒含量的变异系数(0.59)要高于北方花生产区的(0.49),同时南方花生产区土壤有机质含量的变异系数(0.43)也明显高于北方花生产区的(0.39)。全国以及南、北方花生产区土壤的游离铁氧化物含量变异系数均远高于土壤pH的变异系数,这可能是导致全国以及南、北方花生产区预测模型中游离铁氧化物含量相对重要性均更高的原因之一。此外,据褚宏欣等^[30]在我国17个省份开展的旱地土壤微量元素调查发现,北方土壤有效铁含量相较于南方明显偏低,缺铁现象主要集中在北方,而最近Xu等^[31]发现缺铁会促进双子叶植物根系铁相关转运基因IRT1介导的Cd转运增强。

3 结论

(1)采集的全国花生样品Cd富集系数平均值为2.42,花生籽粒Cd含量超标的15个点位中仅有3个土壤Cd含量超标,说明花生对Cd的富集能力较强,我国现行的土壤污染风险管控标准可能并不能准确评估我国花生产地Cd污染风险。

(2)利用全国以及南北方产区分组数据构建的随机森林模型($R^2=0.930\sim0.966$)对花生Cd富集系数的预测性能均明显优于相应的多元线性回归模型($R^2=0.471\sim0.657$),为田间大尺度下土壤-花生系统中Cd的迁移预测提供了新的视角和解决方案。

(3)随机森林模型分析结果表明影响北方产区花生Cd富集系数预测最重要的特征变量为土壤游离锰氧化物含量、游离铁氧化物含量和pH,而影响南方产区花生Cd富集系数预测最重要的特征变量为土壤游离锰氧化物、黏粒、游离铁氧化物和有机质含量。结果为有针对性地研发区域花生产地土壤污染风险管控与修复技术提供了科学依据。

参考文献:

[1] 万书波,张佳蕾.中国花生产业降本增效新途径探讨[J].中国油料

- 作物学报, 2019, 41(5): 657-662. WANG S B, ZHANG J L. Discussion on new ways to reduce cost and increase efficiency of peanut industry in China[J]. *Chinese Journal of Oil Crop Sciences*, 2019, 41(5): 657-662.
- [2] WANG S H, LI G. Assessment of cadmium bioaccumulation and distribution in the kernels of peanut widely cultivated in China[J]. *Ecotoxicology and Environmental Safety*, 2014, 108: 23-28.
- [3] DAI X H, BAI Y Z, JIANG J, et al. Cadmium in Chinese postharvest peanuts and dietary exposure assessment in associated population[J]. *Journal of Agricultural and Food Chemistry*, 2016, 64(41): 7849-7855.
- [4] 王姗姗,王颜红,张红.污染花生籽实中镉的分布特征及其对膳食健康的风险分析[J].农业环境科学学报, 2007, 26(增刊): 12-16. WANG S S, WANG Y H, ZHANG H. Cd-contaminating peanut seeds: distribution characteristics of cadmium and risk assessment on dietary health[J]. *Journal of Agro-Environment Science*, 2007, 26(Suppl): 12-16.
- [5] SHAHID M, DUMAT C, KHALID S, et al. Cadmium bioavailability, uptake, toxicity and detoxification in soil-plant system[J]. *Reviews of Environmental Contamination and Toxicology*, 2016, 241: 73-137.
- [6] YANG Q, YANG Z F, ZHANG Q Z, et al. Transferability of heavy metal (loid)s from karstic soils with high geochemical background to peanut seeds[J]. *Environmental Pollution*, 2022, 299: 118819.
- [7] HUANG B Y, LV Q X, TANG Z X, et al. Machine learning methods to predict cadmium(Cd) concentration in rice grain and support soil management at a regional scale[J]. *Fundamental Research*, 2023, doi: 10.1016/j.fmre.2023.02.016.
- [8] 牟力言,刘春湘,陈敏,等.基于机器学习识别中微量元素与我国七大片区产地稻米镉、砷的富集规律[J].农业环境科学学报, 2023, 42(10): 2165-2174. MOU L Y, LIU C X, CHEN M, et al. Enrichment patterns of cadmium and arsenic in rice from seven major regions in China based on machine learning recognition of minor and trace elements[J]. *Journal of Agro-Environment Science*, 2023, 42(10): 2165-2174.
- [9] MCBRIDE M B, SHAYLER H A, SPLIETHOFF H M, et al. Concentrations of lead, cadmium and barium in urban garden-grown vegetables: the impact of soil variables[J]. *Environmental Pollution*, 2014, 194: 254-261.
- [10] 鲁如坤.土壤农业化学分析方法[M].北京:中国农业科技出版社, 2000. LU R K. Soil and agro-chemistry analytical methods[M]. Beijing: China Agricultural Science and Technology Press, 2000.
- [11] GUO F Y, DING C F, ZHOU Z G, et al. Effects of combined amendments on crop yield and cadmium uptake in two cadmium contaminated soils under rice-wheat rotation[J]. *Ecotoxicology and Environmental Safety*, 2018, 148: 303-310.
- [12] ME E, DU LAING G, UNAMUNO V. Comparison of cadmium extractability from soils by commonly used single extraction protocols[J]. *Geoderma*, 2007, 141: 247-259.
- [13] 潘泳兴,陈盟,王橧橧,等.基于RF和MLR的土壤重金属影响因素分析及生物有效性预测[J].农业环境科学学报, doi: 10.11654/jaes.2023-0368. PAN Y X, CHEN M, WANG X T, et al. Analysis

- of influencing factors and bioavailability prediction of soil heavy metals based on RF and MLR[J]. *Journal of Agro-Environment Science*, doi:10.11654/jaes.2023-0368.
- [14] 牛硕, 李艳玲, 杨阳, 等. 基于机器学习方法的小麦镉富集因子预测[J]. *环境科学*, 2023, 44(6):3619-3626. NIU S, LI Y L, YANG Y, et al. Prediction of cadmium uptake factor in wheat based on machine learning[J]. *Environmental Science*, 2023, 44(6):3619-3626.
- [15] 王飞, 王建国, 刘登望, 等. 不同花生品种对稻田镉富集及转运的研究[J]. *中国油料作物学报*, 2019, 41(4):568-576. WANG F, WANG J G, LIU D W, et al. Cadmium concentration and translocation in paddy fields with different peanut varieties[J]. *Chinese Journal of Oil Crop Sciences*, 2019, 41(4):568-576.
- [16] GU Q B, YU T, YANG Z F, et al. Prediction and risk assessment of five heavy metals in maize and peanut: a case study of Guangxi, China[J]. *Environmental Toxicology and Pharmacology*, 2019, 70: 103199.
- [17] DING C F, MA Y B, LI X G, et al. Determination and validation of soil thresholds for cadmium based on food quality standard and health risk assessment[J]. *Science of the Total Environment*, 2018, 619-620: 700-706.
- [18] LIU C S, CHANG C Y, FEI Y H, et al. Cadmium accumulation in edible flowering cabbages in the Pearl River Delta, China: critical soil factors and enrichment models[J]. *Environmental Pollution*, 2018, 233:880-888.
- [19] XIE M, LI H, ZHU Y, et al. Predicting bioaccumulation of potentially toxic element in soil-rice systems using multi-source data and machine learning methods: a case study of an industrial city in southeast China[J]. *Land*, 2021, 10(6):558.
- [20] 窦韦强, 安毅, 秦莉, 等. 土壤pH对镉形态影响的研究进展[J]. *土壤*, 2020, 52(3):439-444. DOU W Q, AN Y, QIN L, et al. Advances in effects of soil pH on cadmium form[J]. *Soils*, 2020, 52(3):439-444.
- [21] MCLAUGHLIN M J, SMOLDERS E, ZHAO F J, et al. Managing cadmium in agricultural systems[M]//SPARKS D L. *Advances in agronomy*, 166. US: Academic Press, 2021: 1-129.
- [22] ZHAO B, ZHU W X, HAO S F, et al. Prediction heavy metals accumulation risk in rice using machine learning and mapping pollution risk[J]. *Journal of Hazardous Materials*, 2023, 448:130879.
- [23] 汪鹏, 赵方杰. 土壤-水稻系统中镉迁移与阻控[J]. *南京农业大学学报*, 2022, 45(5):990-1000. WANG P, ZHAO F J. The transfer and control of cadmium in the soil-systems[J]. *Journal of Nanjing Agricultural University*, 2022, 45(5):990-1000.
- [24] 陈宏坪, 戴碧川, 杨新萍, 等. 土壤与水稻籽粒镉含量相关性分析及水稻产地土壤镉临界值的研究[J]. *土壤*, 2018, 50(2):361-368. CHEN H P, DAI B C, YANG X P, et al. Cadmium(Cd) threshold values of paddy soils to brown rice as determined by Cd concentrations in soils and rice grains as well as soil properties[J]. *Soils*, 2018, 50(2):361-368.
- [25] 杨丽萍, 郭洪海, 朱振林, 等. 中国花生生态适宜性评价研究[J]. *中国油料作物学报*, 2019, 41(3):461-468. YANG L P, GUO H H, ZHU Z L, et al. Ecological suitability evaluation of peanut plantation in China[J]. *Chinese Journal of Oil Crop Sciences*, 2019, 41(3):461-468.
- [26] 贾立华, 赵长星, 王月福, 等. 不同质地土壤对花生根系生长、分布和产量的影响[J]. *植物生态学报*, 2013, 37(7):684-690. JIA L H, ZHAO C X, WANG Y F, et al. Effects of different soil textures on the growth and distribution of root system and yield in peanut[J]. *Chinese Journal of Plant Ecology*, 2013, 37(7):684-690.
- [27] 李涛, 李光照, 梅馨月, 等. 镉-钙交互作用对植物的生态学效应[J]. *中国农学通报*, 2017, 33(32):77-80. LI T, LI G Z, MEI X Y, et al. The interaction of cadmium and calcium and its ecological effects on plants[J]. *Chinese Agricultural Science Bulletin*, 2017, 33(32):77-80.
- [28] 宋正国, 徐明岗, 李菊梅, 等. 钙对土壤镉有效性的影响及其机理[J]. *应用生态学报*, 2009, 20(7):1705-1710. SONG Z G, XU M G, LI J M, et al. Effects of calcium on cadmium bioavailability in lateritic red soil and related mechanisms[J]. *Chinese Journal of Applied Ecology*, 2009, 20(7):1705-1710.
- [29] ZHANG S, LI Q, NAZIR M M, et al. Calcium plays a double-edged role in modulating cadmium uptake and translocation in rice[J]. *International Journal of Molecular Sciences*, 2020, 21(21):8058.
- [30] 褚宏欣, 党海燕, 王涛, 等. 我国主要麦区土壤有效铁锰铜锌丰缺状况评价及影响因素[J]. *土壤学报*, 2024, 61(1):129-139. CHU H X, DANG H Y, WANG T, et al. Evaluations and influencing factors of soil available Fe, Mn, Cu and Zn concentrations in major wheat production regions of China[J]. *Acta Pedologica Sinica*, 2024, 61(1): 129-139.
- [31] XU Z R, YOU T T, LIU W Y, et al. Mitigating cadmium accumulation in dicotyledonous vegetables by iron fertilizer through inhibiting Fe transporter IRT1-mediated Cd uptake[J]. *Chemosphere*, 2024, 346: 140559.

(责任编辑:叶飞)