doi:10.6041/j.issn.1000-1298.2022.08.024

融合光谱和空间特征的土壤重金属含量极端随机树估算

于海洋[1,2]　谢赛飞[1,2]　郭灵辉[1]　刘　鹏[3,4]　张　平[1,2]

(1. 河南理工大学测绘与国土信息工程学院, 焦作 454003;
2. 河南理工大学自然资源部矿山时空信息与生态修复重点实验室, 焦作 454003;
3. 河南省自然资源科学研究院河南省国土资源动态监测重点实验室, 郑州 450053;
4. 河南省航空物探遥感中心遥感技术研究院, 郑州 450053)

摘要: 针对高光谱遥感土壤重金属含量估算研究中光谱特征信息弱、模型反演鲁棒性差的问题,提出构建污染源-汇空间特征量化污染物扩散与汇聚空间影响因子,融合光谱特征建立基于极端随机树(Extremely randomized trees, ERT)的土壤重金属含量估算模型。以济源市耕地土壤为研究区,布设采集土壤样本 249 个,分析了光谱特征、地形特征和污染源空间特征在土壤重金属铅(Pb)、铬(Cd)含量反演中的有效性及影响机理,采用置换重要性指数优选多源特征,通过与多种回归模型对比,评价 ERT 模型的预测精度。研究表明,变换后的土壤光谱特征构建 ERT 模型引入地形特征和污染源空间特征后精度提升显著,尤其是污染源空间特征优势更为明显,Pb 的 ERT 模型均方根误差由 43.185 mg/kg 下降到 22.301 mg/kg,下降了 48.36%。Cd 的 ERT 模型均方根误差由 0.738 mg/kg 下降到 0.371 mg/kg,下降了 49.73%,充分说明引入污染扩散空间特征的有效性。与其他回归模型对比,ERT 估算模型在各项指标评价中优势明显,其中 Pb 的 ERT 模型的测试集 R^2 达 0.964,Cd 的 ERT 模型 R^2 为 0.923。

关键词: 土壤; 重金属; 高光谱遥感; 空间特征; 极端随机树; 置换重要性

中图分类号: S127　　**文献标识码**: A　　**文章编号**: 1000-1298(2022)08-0231-09　　**OSID**:

Extremely Randomized Trees Estimation of Soil Heavy Metal Content by Fusing Spectra and Spatial Features

YU Haiyang[1,2]　XIE Saifei[1,2]　GUO Linghui[1]　LIU Peng[3,4]　ZHANG Ping[1,2]

(1. *School of Surveying and Land Information Engineering, Henan Technology University, Jiaozuo* 454003, *China*
2. *Key Laboratory of Mine Spatio-temporal Information and Ecological Restoration, Ministry of Natural Resources, Henan Technology University, Jiaozuo* 454003, *China*
3. *Henan Key Laboratory for Dynamic Monitoring of Land Resources, Henan Academy of Natural Resources Sciences, Zhengzhou* 450053, *China*
4. *Institute of Remote Sensing Technology, Henan Aero Geophysical Survey and Remote Sensing Center, Zhengzhou* 450053, *China*)

Abstract: Aiming at the problems of weak spectral characteristic information and poor robustness of model inversion in the estimation of soil heavy metal content by hyperspectral remote sensing, it was proposed to construct spatial features of pollution source and sink to quantify the spatial influence factors of pollutant diffusion and aggregation, and integrate the spectral features to establish the estimation model of soil heavy metal content based on extremely randomized trees (ERT). Taking the cultivated soil of Jiyuan City as the study area, totally 249 soil samples were collected. The effectiveness and influence mechanism of spectral features, topographic features and spatial features of pollution sources in the inversion of soil heavy metal Pb and Cd were analyzed. The multi-source characteristics were optimized by permutation importance index, and the prediction accuracy of ERT model was evaluated by comparing with various regression models. The research showed that the ERT model constructed from the transformed soil spectral features can achieve a certain inversion accuracy, and the accuracy was significantly improved after the introduction of topographic features and spatial features of pollution

收稿日期: 2022-03-03　修回日期: 2022-05-26

基金项目: 国家自然科学基金项目(U1304402、41977284)和河南省自然资源厅自然科技项目(2019-378-16)

作者简介: 于海洋(1978—),男,副教授,博士,主要从事遥感地学应用研究,E-mail: yuhaiyang@hpu.edu.cn

通信作者: 郭灵辉(1983—),男,副教授,博士,主要从事土地系统研究,E-mail: guolinghui@hpu.edu.cn

sources. In particular, the advantage of the spatial features of pollution sources was more obvious, the RMSE of Pb ERT model was decreased from 43.185 mg/kg to 22.301 mg/kg, with decrease of 48.36%, the RMSE of Cd ERT model was decreased from 0.738 mg/kg to 0.371 mg/kg, with down of 49.73%, which fully demonstrated the effectiveness of the pollution diffusion spatial features. The results of multi-feature combination modeling experiments showed that the features with the high permutation importance index were the spatial features of the pollution source, followed by the spectral features. In the research, the estimation model established by using the selected features of the permutation importance index was very close to the optimal modeling accuracy when all the features were used, which showed the effectiveness of the feature screening method based on the permutation importance index. Compared with regression models such as MLR, SVM, RF, and GBDT, the ERT estimation model had obvious advantages in the evaluation of various indicators. The R^2 value of the Pb ERT model in the test set reached 0.964, and the R^2 value of the Cd ERT model was 0.923. The experimental results showed that the introduction of the pollutant diffusion spatial features and the fusion of spectral features to construct ERT model to estimate soil heavy metal content had high accuracy and certain popularization and application value.

Key words: soil; heavy metal; hyperspectral remote sensing; spatial features; extremely randomized trees; permutation importance

0 引言

矿业活动、冶金以及工业生产中产生的无机污染物通过大气降尘和污水排放等途径进入土壤,不断积聚造成较为严重的土壤重金属污染,这些污染物渗透进入土壤后移动性较差,残留时间长,且不易被微生物降解,严重威胁生态安全。据统计,我国耕地土壤污染超标率高达 19.4%,其中以重金属污染最为严峻[1]。因此,高效、快速获取土壤重金属含量及空间分布,对于重金属污染防治、农业生产和生态安全具有重要意义。

传统土壤重金属污染调查采用实地采集土壤样品和化学分析的方法,需耗费大量的人力物力资源。遥感光谱反演方法为快速、高效的土壤重金属污染信息获取提供了可选方案,已有文献对采用高光谱遥感技术监测土壤重金属污染进行了有益的尝试[2]。主要针对光谱信息增强变换与特征选取、反演算法[3]等进行了分析,使用的光谱信息增强方法包括光谱微分(一阶、二阶等)[4]、连续统去除、高斯卷积平滑、多元散射校正[5]等,常用反演模型包括多元线性回归(Multivariable linear regression, MLR)、偏最小二乘回归[6-8]、支持向量机(Support vector machine, SVM)[9-10]、极限学习机[11]、人工神经网络、随机森林(Random forest, RF)[12]等。已有研究进行反演建模时仅考虑了土壤光谱特征,由于土壤光谱影响因素众多,重金属元素含量低,光谱信息弱,导致反演模型鲁棒性差,泛化能力不强[13]。

土壤重金属污染物主要来源于矿业活动、冶金以及工业生产中形成的大气降尘、污水排灌等,潜在污染源长期存在,一般较为明确。因此,考虑构建污染源-汇空间特征量化污染物扩散与汇集空间影响因子,融合光谱特征建立土壤重金属含量估算模型。近年来,极端随机树(Extremely randomized trees, ERT)集成学习方法在其他领域机器学习建模研究中表现优异,具有鲁棒性高、泛化能力强的特点[14],本文将该算法引入土壤重金属含量反演,以期进一步提升模型预测精度和泛化能力。

1 研究区概况

研究区位于河南省济源市市区周边的山前平原区(图 1),分布于 112.46°~112.68°E,35.01°~35.19°N 之间,东、南分别以二广高速、菏宝高速为界,西、北以山麓地形为界,面积 263.5 km^2。济源市区周边分布的部分金属冶炼厂导致其耕地土壤形成了多种重金属元素浓度异常分布区,其主要污染元素为铅(Pb)、铬(Cd)等,表层土壤铅异常浓集中心位于铅锌冶炼厂附近,高值区 Pb 平均含量(质量比)为 611 mg/kg,达到三级污染标准值的 1.22 倍。

2 实验数据获取与预处理

2.1 土壤光谱数据测定及预处理

野外土壤样品采集时间为 2019 年 10—11 月,按照放射状为主、环形补充的网状方法布设采样点(图 1),采用 GNSS RTK 定位后采集表层土壤(0~20 cm)作为样本,采集数量为 249 个。

土壤光谱数据采用美国 ASD 公司的 ASD Fieldspec3 型光谱仪进行测量,光谱波长范围为 350~2 500 nm。在实验室内将土样干燥、研磨后过 100 目筛,使用高密度反射探头测量土壤样本光谱反射率,每个样本连续测定 10 次,取其平均值作为最终的光

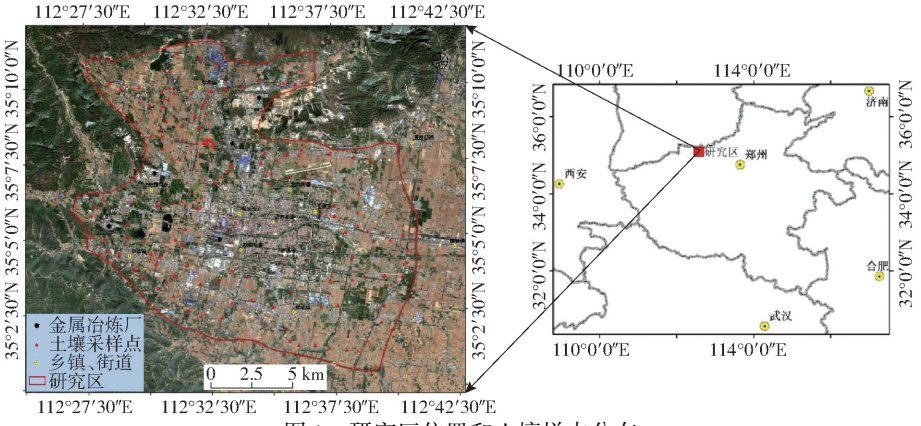

图 1　研究区位置和土壤样本分布

Fig. 1　Study area and distribution of soil samples

谱反射率(图 2)。

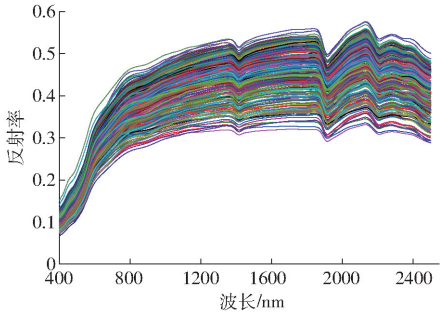

图 2　土壤光谱曲线

Fig. 2　Soil spectral curves

因受外界测量条件和传感器本身的影响,在土壤样品光谱采集过程中存在不同程度的噪声,故有必要对原始光谱进行降噪处理。采用 Savizky - Golay(SG)卷积平滑法对光谱曲线进行平滑去噪。原始光谱反射率测量值与土壤重金属含量之间的相关性较低,对滤波后的光谱反射率进行光谱变换增强光谱信息。研究选取的光谱变换方法包括:一阶微分(First order differential reflectance,FD)、二阶微分(Second order differential reflectance,SD)、多元散射校正(Multiplicative scatter correction,MSC)、标准正态变量变换(Standard normal variate,SNV)、连续统去除(Continuum removal,CR)、倒数一阶微分(Reciprocal first derivative, RFD)、倒数二阶微分(Reciprocal second derivative, RSD)。其中多元散射校正可以有效消除样品散射的影响[15];标准正态变量变换通过加权平均化消除固体的颗粒不均对光谱的影响;光谱微分可以有效消除基线和减弱背景干扰,分辨混合光谱,增强光谱特征[16];连续统去除可以有效突出光谱曲线的吸收和反射特征[17]。

光谱特征中含有较多的冗余和共线性变量,在进行定量分析前,需筛选重金属元素的特征波段,提高建模效率。将相关性分析中满足 $P = 0.01$ 假设性检验的波段集合作为特征波段区域,然后采用连续投影算法[18]从特征波段区域内提取重金属元素的特征波段。

2.2　潜在污染源空间特征量化

矿山开采、金属冶炼过程中产生的无机污染物扩散以点源污染为主,通常以开采、冶炼厂为中心向四周扩散,同时受到风向、风速、地形以及降水等自然因子的影响,这些污染物扩散空间影响因子对于土壤污染浓度的分布产生直接影响,因此考虑引入适当的污染源-汇空间特征对上述污染扩散模型进行量化。

研究区内主要分布有较大的铅锌、钢铁等金属冶炼厂 5 家,位置分布如图 1 所示,这些冶炼厂是造成研究区土壤重金属污染的潜在污染源。污染源污染物向四周扩散,样本点与这些潜在污染源的距离以及方位关系是影响样本点位置污染物累积量的重要因子,因此,主要选取了污染源与采样点的空间距离和方位角作为空间特征因子。针对每个污染源分别计算以下 2 个空间特征:

(1)污染源与采样点的空间距离

空间上距离污染源越近,一般污染物累积越多,因此引入空间距离特征描述距离因子对于污染源污染物向四周扩散的影响。根据采样点与每个污染源的平面坐标采用欧氏距离公式进行计算,距离污染源越远,受到污染源的影响越小,因此将该特征量化为距离的倒数进行建模。

(2)污染源与采样点连线的方位角

在偏离每个污染源的不同方位,由于风向、风速的差异,污染物扩散会出现明显变化,如果采样点位于污染源主导风向的下风向,其污染物积聚浓度更高,因此,采样点与污染源的方位角信息可以模拟在污染源不同方向上大气扩散条件的差异。方位角空间特征是指以污染源 O 为起点、样本点 A 为终点的连线与正北方向的夹角,描述了样本采样点与污染源的空间方位关系。方位角 α 具体计算步骤如下:

首先计算潜在污染源 $O(x_0,y_0)$ 与样本 $M(x_i,y_i)$ 之间的象限角 β,计算公式为

$$\beta = \arctan\frac{\Delta y}{\Delta x} \tag{1}$$

其中 $\Delta x = x_i - x_0 \quad \Delta y = y_i - y_0$

如图 3 所示,根据计算得到 Δx、Δy 来判断象限角 β 位于第几象限,并以此来计算方位角:①当 $\Delta x > 0, \Delta y > 0$ 时,角 β 位于第Ⅰ象限,方位角 $\alpha = \beta$。②当 $\Delta x < 0, \Delta y > 0$ 时,角 β 位于第Ⅱ象限,方位角 $\alpha = 180° - \beta$。③当 $\Delta x < 0, \Delta y < 0$ 时,角 β 位于第Ⅲ象限,方位角 $\alpha = 180° + \beta$。④当 $\Delta x > 0, \Delta y < 0$ 时,角 β 位于第Ⅳ象限,方位角 $\alpha = 360° - \beta$。

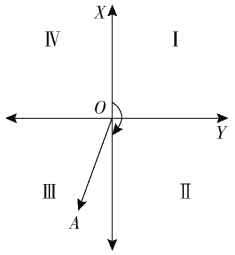

图 3 方位角计算

Fig. 3 Azimuth calculation

在具体计算时,针对 5 个潜在污染源分别计算距离倒数因子和方位角因子 2 个空间特征,因此,每个样本计算 10 个污染源空间特征。

2.3 污染扩散地形影响因子选取

地形地势决定了水流流向、流速,对局部气流风向、风速等也具有一定控制作用,从而对重金属污染物的扩散和汇集产生影响,引入高程(Elevation)、坡度(Slope)、坡向(Aspect)、坡长因子(LS factor, LSF)[19]、形态特征(Morphometric features, MF)[20]、广义表面指数(Generalized surface index, GSI)[20]、风效指数(Wind exposition, WE)[20]、地形湿度指数(Topographic wetness index, TWI)[21],用于分析地形因子对于土壤重金属污染物浓度的影响。以上 8 个地形因子均基于 DEM 数据计算,并通过插值提取相应样本点位置特征值,DEM 数据采用 SRTM 数据,网格尺寸为 30 m。

2.4 土壤样本重金属元素含量测定

研究区土壤重金属污染以 Pb 和 Cd 污染为主,重点针对这 2 种元素进行分析。土壤重金属 Pb、Cd 含量采用 XSERIES - 2 型电感耦合等离子体质谱仪测定,样本的统计信息如表 1 所示。Cd 含量与 Pb 含量之间相关系数达到 0.825,两者具有较高的同源性,发生协同作用的可能性较大。

统计数据显示重金属 Pb 达到重污染的样本占全部样本的 19.3%,重金属 Cd 达到重污染的样本占全部样本的 75.5%,研究区内存在较为严重的 Pb、Cd 污染,需全面加强对该地区土壤重金属污染的监测。

表 1 研究区重金属 Pb、Cd 含量数据基本统计信息

Tab. 1 Basic content statistics of heavy metal Pb and Cd of study area

样本	样本数	最小值/(mg·kg^{-1})	最大值/(mg·kg^{-1})	均值/(mg·kg^{-1})	标准差/(mg·kg^{-1})	变异系数/%
Pb	249	26.54	545.00	145.53	114.86	79
Cd	249	0.27	9.34	2.01	1.42	70

将采集的 249 个样本按照约 3∶1 的比例随机划分为建模集和测试集,其中建模集 186 个样本,测试集 63 个样本。

3 研究方法

3.1 极端随机树

极端随机树类似于随机森林方法,是一种由多棵决策树构成的集成学习方法。随机森林采用随机采样来选择样本集作为每个决策树的训练集[22],该方法不能保证所有样本能被充分利用,并且各决策树之间可能存在相似性。基于以上考虑,GEURTS 等[23]提出极端随机树模型。

在极端随机树中,每棵决策树均采用全部训练集,训练样本的利用率高,能在一定程度上减少最终预测偏差;为了保证每棵决策树间的结构差异,极端随机树在节点拆分时引入了更大的随机性:从子数据集中随机选取每个特征的判断阈值,并选择拆分效果最好的特征作为最优判断属性。由于节点拆分判断阈值的随机性,极端随机树的泛化能力一般会优于随机森林方法。

一般以节点的不纯度作为最优判断属性的选取依据[24],回归类问题衡量节点不纯度的函数一般选择均方误差(MSE)或平均绝对误差(MAE)。选用 MSE 作为函数的节点不纯度(G),计算公式为

$$G(u_i, v_{ij}) = \frac{1}{N_S}\left[\sum_{y_i \in X_{\text{left}}}(y_i - \bar{y}_{\text{left}})^2 + \sum_{y_j \in X_{\text{right}}}(y_j - \bar{y}_{\text{right}})^2\right] \tag{2}$$

式中 u_i——某一个节点判断属性

v_{ij}——判断属性的取值

N_S——当前节点所有训练样本个数

X_{left}、X_{right}——左、右子节点的训练样本集合,$y_i \in X_{\text{left}}, y_j \in X_{\text{right}}$

$\bar{y}_{\text{left}}$、$\bar{y}_{\text{right}}$——当前节点左、右子节点的样本目标变量平均值

经过 k 轮训练得到 k 棵结构不同的决策树,最后通过投票或取平均的方式集合不同决策树的预测结果 $h_i(x)$,得到模型的最终结果 $H(x)$。在回归类

问题中，常采用平均的方式计算模型的最终结果，即

$$H(x)=\frac{1}{k}\sum_{i=1}^{k}h_i(x) \tag{3}$$

3.2　置换重要性指数

随机森林和极端随机树算法可以基于不纯度测量每个特征对模型预测的相对重要性，这种基于不纯度计算特征重要性倾向于夸大连续特征或高基数属性特征的重要性，另一种特征重要性计算方法是置换重要性（Permutation importance，PI）[22]，该指数是通过观察每个预测属性的随机重排对模型预测精度的影响来直接衡量特征的重要性。

该方法计算过程为：首先训练基线模型，并通过验证集记录 R^2 得分为基准评分 S。然后选定数据集中的一个特征要素 F_j，打乱顺序重新排列其属性值为 $F_{m,j}$（m 表示 M 次打乱数据中某一次），利用修改后数据集重新建立预测模型，通过验证集计算 R^2 得分 $S_{m,j}$。特征重要性 P_{F_j} 是基准评分 S 与属性值重新排列后数据集构建模型的评分之间的差异。公式为

$$P_{F_j}=S-\frac{1}{M}\sum_{m=1}^{M}S_{m,j} \tag{4}$$

PI 指数计算方便，特征重要性评价准确，可解释性较好。

3.3　反演模型精度评价

模型的精度评价采用验证集的决定系数（R^2）、均方根误差（Root mean square error，RMSE）、相对分析误差（Relative percent deviation，RPD）以及训练集交叉验证得分（Cross validate score，CVS）等。

相对分析误差 E_{RPD} 的计算公式为[25]

$$E_{RPD}=\frac{\sigma}{e} \tag{5}$$

式中　σ——验证集样本的标准差

　　e——均方根误差

一般当 $E_{RPD}<1.4$ 时，模型无法对样品进行预测；当 $1.4\leqslant E_{RPD}<2.0$ 时，模型精度一般，具有粗略评估样品的能力；当 $E_{RPD}\geqslant 2.0$ 时，模型具有较好的预测能力。

4　实验与结果分析

针对光谱变换特征、污染源空间特征以及地形特征进行组合实验，分别采用光谱、光谱与地形、光谱与空间以及光谱、空间与地形的特征组合进行实验，分析不同建模特征的置换重要性，评价光谱特征以及污染扩散空间影响因子选取的有效性。反演模型同时选取了多元线性回归、支持向量机、随机森林、梯度提升决策树（Gradient boosting decision tree，GBDT）等回归模型作为参考，评估极端随机树估算模型的有效性和先进性。

4.1　特征重要性分析

4.1.1　光谱特征分析

原始光谱反射率测量值与土壤重金属含量相关系数较低，计算后的变换特征相关性显著提高，其中 CR1765（CR 表示连续统去除光谱变换，1765 表示波长位置为 1 765 nm，其他编号含义相同）特征与土壤 Pb 含量最大相关系数达到 −0.70。分别计算光谱变换特征与土壤 Pb、Cd 含量的相关性，对相关系数进行 $P=0.01$ 水平上的假设性检验，将通过假设性检验的波段集合作为特征波段区域。基于 SPA 算法在特征波段区域内分别筛选不同元素共线性最小的有效特征波段组合，其中与 Pb 相关的筛选光谱特征为 72 个，与 Cd 相关的筛选光谱特征为 65 个。

图 4 为单独使用光谱变换特征构建 Pb、Cd 元素 ERT 估算模型时分析得到的置换重要性计算结果。每个特征重新排列计算 10 次，然后对计算所得 PI 值的均值和方差进行排序，在此展示的是 PI 均值最高的 15 个特征及其统计值。其中 Pb 元素重要性评价最高的特征为 CR2262，其次为 CR2174 和 SD1345，这 3 个特征 PI 值明显高于其他特征。Cd 元素重要性评价最高的特征为 FD1802，该特征 PI 值明显高于其他特征，剩余特征差异较小，其中 PI 值较高的包括 MSC1799 和 SNV1729 等。由统计结果可以看到，PI 值较高的波段多为近红外波段，与已有研究成果[2]较为吻合。

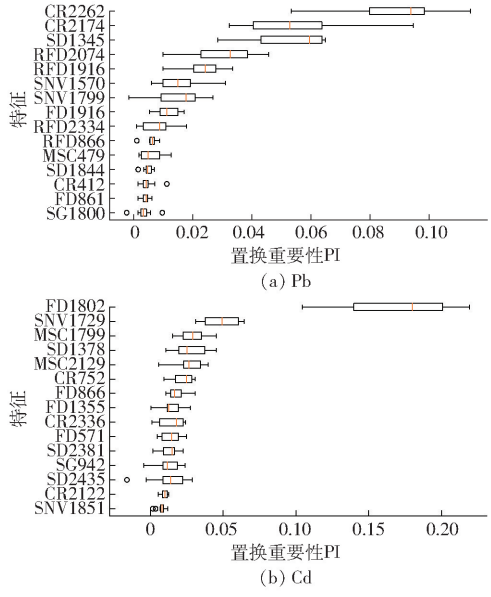

图 4　建模特征为光谱时 ERT 模型计算特征 PI 统计结果

Fig. 4　Statistical results of features PI calculated by ERT model when modeling features were spectrum

表 2 统计了不同回归模型和建模特征土壤重金

表 2 不同回归模型和建模特征土壤重金属 Pb、Cd 含量反演精度对比

Tab. 2 Comparison of precision of soil heavy metal Pb and Cd inversion with different regression models and modeling features

建模特征	反演模型	Pb				Cd			
		R^2	RMSE/(mg·kg^{-1})	RPD	CVS	R^2	RMSE/(mg·kg^{-1})	RPD	CVS
光谱	MLR	0.679	65.751	1.764	0.634	0.448	0.938	1.346	0.257
	SVM	0.655	68.112	1.703	0.395	0.694	0.794	1.808	0.529
	RF	0.838	46.663	2.486	0.664	0.636	0.866	1.658	0.413
	GBDT	0.833	47.339	2.451	0.636	0.591	0.918	1.564	0.415
	ERT	0.861	43.185	2.686	0.738	0.736	0.738	1.945	0.523
光谱 + 地形	MLR	0.671	66.525	1.744	0.522	0.477	0.839	1.382	0.192
	SVM	0.684	65.189	1.780	0.379	0.754	0.626	2.017	0.534
	RF	0.851	44.770	2.591	0.718	0.796	0.570	2.216	0.487
	GBDT	0.910	34.871	3.327	0.765	0.783	0.588	2.148	0.485
	ERT	0.912	34.338	3.378	0.798	0.800	0.564	2.238	0.589
光谱 + 空间	MLR	0.887	39.054	2.970	0.667	0.557	0.840	1.503	0.312
	SVM	0.783	54.089	2.145	0.519	0.866	0.462	2.734	0.600
	RF	0.954	24.936	4.652	0.863	0.893	0.414	3.052	0.665
	GBDT	0.956	24.292	4.775	0.875	0.885	0.429	2.943	0.683
	ERT	0.963	22.301	5.202	0.880	0.914	0.371	3.403	0.727
光谱 + 空间 + 地形	MLR	0.860	43.462	2.669	0.534	0.483	0.908	1.390	0.087
	SVM	0.746	58.433	1.985	0.488	0.816	0.541	2.334	0.581
	RF	0.952	25.398	4.568	0.871	0.919	0.360	3.507	0.661
	GBDT	0.954	24.886	4.661	0.878	0.885	0.428	2.947	0.693
	ERT	0.964	22.081	5.254	0.888	0.928	0.340	3.715	0.720
置换重要性指数筛选后特征	MLR	0.860	43.462	2.669	0.534	0.483	0.908	1.390	0.087
	SVM	0.746	58.434	1.985	0.488	0.816	0.541	2.334	0.581
	RF	0.951	25.739	4.507	0.878	0.906	0.386	3.270	0.680
	GBDT	0.957	24.041	4.825	0.875	0.880	0.437	2.891	0.694
	ERT	0.964	21.912	5.294	0.895	0.923	0.349	3.613	0.718

属 Pb、Cd 含量反演测试集精度,其中仅使用光谱特征时 Pb 元素的 ERT 模型 R^2 可达 0.861,RPD 为 2.686,具有较高的定量反演精度,Cd 元素的 ERT 模型 R^2 可达 0.736,RPD 为 1.945,具有粗略的预测能力,说明对于土壤重金属元素含量反演,光谱特征虽然为弱信息,但由于污染物扩散中形成的落尘等改变了土壤组分和性状,从而在土壤光谱特征中表现出来,变换后的土壤光谱特征能够在一定程度上反映这种污染程度。

4.1.2 污染扩散地形影响因子分析

地形特征对于污染物的扩散和汇聚能够产生一定的影响,从表 2 可以看到,当建模特征中加入 8 个地形特征后,Pb 和 Cd 的建模精度均有明显提升,Pb 的 ERT 模型 R^2 由 0.861 提升至 0.912,Cd 的 ERT 模型 R^2 由 0.736 提升至 0.800,其他统计值也得到了有效提升,说明了地形特征的有效性。

建模特征为光谱和地形组合时 ERT 模型计算特征置换重要性统计结果显示,除光谱特征外地形广义表面指数(DTM_GS)、高程(DTM_E)、坡度(DTM_S)和风效指数(DTM_WE)因子具有较高的 PI 值,说明这些地形特征能够较好地反映地形对污染物的扩散和汇聚产生的影响。其中广义表面指数、高程、坡度等因子与水流流向、流速以及土壤侵蚀与堆积等相关性较强,影响土壤污染物的运移与扩散。风效指数能够在一定程度上反映地形对大气污染物扩散产生的影响。

4.1.3 污染源空间特征分析

污染源空间特征能够较好地表征污染物扩散浓度分布,从表 2 可以看出,当建模特征中加入 10 个污染源空间特征后,Pb 和 Cd 的建模精度均有极大提升,Pb 的 ERT 模型 R^2 由 0.861 提升至 0.963,RMSE 由 43.185 mg/kg 下降到 22.301 mg/kg,下降了 48.36%。Cd 的 ERT 模型 R^2 由 0.736 提升至 0.914,RMSE 由 0.738 mg/kg 下降到 0.371 mg/kg,下降了 49.73%,其他统计值也提升明显,充分说明了引入污染源空间特征的有效性。

建模特征为光谱和污染源空间特征组合时 ERT 模型计算特征置换重要性统计结果显示,10 个污染源空间特征均在 PI 值最高的前 15 个特征中,且 Pb 模型的前 7 个特征、Cd 模型的前 9 个特征均

为污染源空间特征，从侧面说明引入空间特征的有效性。该 PI 值也在一定程度上反映了不同污染源对于土壤重金属污染的贡献程度，研究区 SB 和 WY 2 个潜在污染源（图 5 中 SB、WY 表示某冶炼厂，DIST 表示空间距离倒数特征，ANGLE 表示空间方位角特征）的贡献度较高。

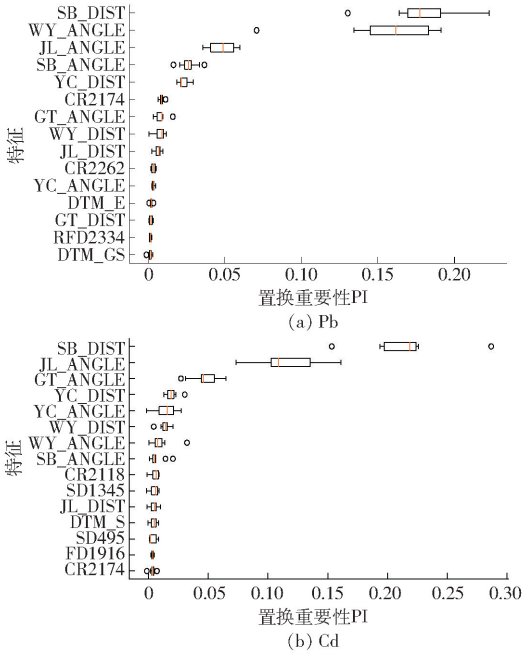

图 5　建模特征为光谱、污染源空间特征和地形特征组合时 ERT 模型计算特征 PI 统计结果

Fig. 5　Statistical results of features PI calculated by ERT model when modeling features were combination of spectrum, spatial features of pollution sources and topographic features

4.1.4　多特征组合分析

当建模特征同时加入光谱、地形和污染源空间特征时，表 2 统计结果表明，建模精度与使用光谱和污染源空间特征组合时基本相同，变化较小，说明污染源空间特征优势更为明显，地形因子与污染源空间因子有一定重叠。图 5 多特征建模置换重要性统计结果也显示出同样的特点，整体 PI 值最高的特征为污染源空间特征，光谱特征次之，地形特征整体 PI 值最低，重要性最弱。

利用光谱、空间和地形特征组合建模时筛选出的 PI 值大于 0.002 的特征进行建模实验，其中 Pb 选取特征 15 个，Cd 选取特征 14 个。表 2 统计结果表明，筛选特征建模精度与使用全部特征时精度极为接近，Pb 的 ERT 模型 R^2 均为 0.964，Cd 的 ERT 模型 R^2 分别为 0.923 和 0.928，说明利用置换重要性进行特征筛选是有效的。

4.2　极端随机树模型估算精度分析

为评价土壤重金属污染极端随机树 ERT 估算模型的先进性，选取 MLR、SVM、RF、GBDT 等回归模型作为对比，表 2 测试集反演模型精度评价统计结果显示，ERT 模型在不同重金属元素、不同特征集的反演建模中均取得了最优精度，Pb 的 ERT 模型的测试集 R^2 达 0.964，Cd 的 ERT 模型 R^2 为 0.923，模型稳定性最佳。整体上 MLR 模型反演精度最低，增加污染扩散空间特征时，Pb 的 MLR 模型精度提升较明显，Cd 元素的 MLR 模型精度提升不大，MLR 模型鲁棒性较差。SVM 模型与 MLR 模型相反，Pb 的 MLR 模型精度提升不大，Cd 的 MLR 模型精度提升明显，但是模型稳定性弱。RF 和 GBDT 反演模型精度接近，优于 MLR 和 SVM 模型，增加地形、空间特征时，反演模型精度均得到较大提升，模型鲁棒性较高。

训练集交叉验证得分 CVS 对模型的泛化能力有较好的评价，在表 2 CVS 统计中 ERT 模型的优势也较为明显，均优于 RF 和 GBDT 模型。图 6 给出了测试集实测真实值与 ERT 模型预测值的序列分布情况，结果表明重金属 Pb 和 Cd 含量预测值与真实值的吻合度较高。

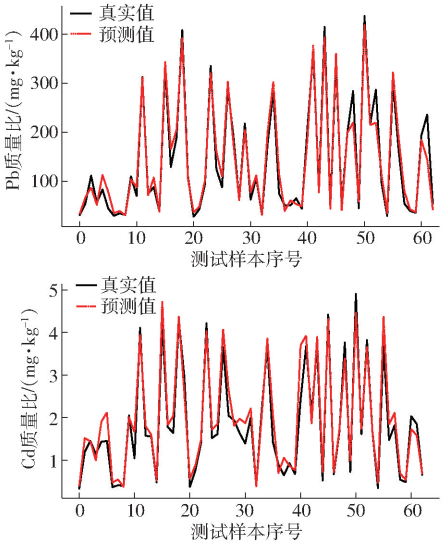

图 6　测试集实测真实值与 ERT 模型预测值序列分布

Fig. 6　Sequence distributions of measured true value of test set and predicted value of ERT model

5　结论

（1）仅使用光谱特征构建的 Pb、Cd ERT 估算模型具有较高的 R^2 和 RPD，说明变换后的土壤光谱特征能够在一定程度上反映这种污染程度。其中 Pb 置换重要性评价最高的特征为 2 262 nm 连续统去除光谱变换特征，Cd 重要性评价最高的特征为 1 802 nm 一阶微分光谱变换特征。

（2）当建模特征中加入地形特征后，Pb 和 Cd 的建模精度均有明显提升，置换重要性统计结果显示地形广义表面指数、高程、坡度和风效指数等特征

具有较高的 PI 值,说明这些地形特征能够较好地反映地形对污染物的扩散和累积产生的影响。

(3)当建模特征中加入污染源空间特征后,Pb 和 Cd 的建模精度均有极大提升,各项统计值改善明显,充分说明了所提出构建污染扩散影响因子的有效性。污染源空间特征重要性分析也可以在一定程度上反映不同污染源对于土壤重金属污染的贡献度。

(4)光谱、地形和污染源空间特征组合建模结果表明 PI 值最高的特征为污染源空间特征,光谱特征次之,地形特征整体 PI 值最低。使用置换重要性指数优选特征建立的估测模型与使用全部特征时建模最优精度极为接近,说明了置换重要性指数特征筛选方法的有效性。

(5)与 MLR、SVM、RF、GBDT 等回归模型对比,ERT 估算模型在各项指标评价中优势明显,其中 Pb 的 ERT 模型的测试集 R^2 达 0.964,Cd 的 ERT 模型 R^2 为 0.923,ERT 土壤重金属估算模型估算精度较高,表明该方法反演土壤重金属含量具有较高的可行性。

(6)提出构建潜在污染源空间特征量化污染物扩散空间影响因子,该方法适用于污染物来源较为明确的点源、线源污染类型,一般土壤重金属污染物来源于矿业活动、冶金以及工业生产中形成的大气降尘、污水排灌等,污染源明确,因此本文提出方法具有较高的推广应用潜力。

参 考 文 献

[1] 环境保护部,国土资源部. 全国土壤污染状况调查公报[J]. 中国环保产业, 2014, 36(5): 10 - 11.
Ministry of Environmental Protection, Ministry of Land and Resources. National bulletin of soil pollution survey[J]. China Environmental Protection Industry, 2014, 36(5): 10 - 11. (in Chinese)

[2] 成永生,周瑶. 土壤重金属高光谱遥感定量监测研究进展与趋势[J]. 中国有色金属学报, 2021, 31(11): 3450 - 3467.
CHENG Yongsheng, ZHOU Yao. Research progress and trend of quantitative monitoring of hyperspectral remote sensing for heavy metals in soil[J]. The Chinese Journal of Nonferrous Metals, 2021, 31(11): 3450 - 3467. (in Chinese)

[3] 张秋霞,张合兵,张会娟,等. 粮食主产区耕地土壤重金属高光谱综合反演模型[J]. 农业机械学报, 2017, 48(3): 148 - 155.
ZHANG Qiuxia, ZHANG Hebing, ZHANG Huijuan, et al. Hybrid inversion model of heavy metals with hyperspectral reflectance in cultivated soils of main grain producing areas[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(3): 148 - 155. (in Chinese)

[4] BALESTRIERI C, COLONNA G, GIOVANE A, et al. Second-derivative spectroscopy of proteins. A method for the quantitative determination of aromatic amino acids in proteins[J]. The FEBS Journal, 2010, 90(3): 433 - 440.

[5] FEARN T, RICCIOLI C, GARRIDO-VARO A, et al. On the geometry of SNV and MSC[J]. Chemometrics and Intelligent Laboratory Systems, 2009, 96(1): 22 - 26.

[6] PANDIT C M, FILIPPELLI G M, LI L. Estimation of heavy-metal contamination in soil using reflectance spectroscopy and partial least-squares regression[J]. International Journal of Remote Sensing, 2010, 31(15):4111 - 4123.

[7] 陈元鹏,张世文,罗明,等. 基于高光谱反演的复垦区土壤重金属含量经验模型优选[J]. 农业机械学报, 2019, 50(1): 170 - 179.
CHEN Yuanpeng, ZHANG Shiwen, LUO Ming, et al. Empirical model optimization of hyperspectral inversion of heavy metal content in reclamation area[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(1): 170 - 179. (in Chinese)

[8] 张秋霞,张合兵,刘文锴,等. 高标准基本农田建设区域土壤重金属含量的高光谱反演[J]. 农业工程学报, 2017, 33(12): 230 - 239.
ZHANG Qiuxia, ZHANG Hebing, LIU Wenkai, et al. Inversion of heavy metals content with hyperspectral reflectance in soil of well-facilitied capital farmland construction areas[J]. Transactions of the CSAE, 2017, 33(12): 230 - 239. (in Chinese)

[9] 袁自然,魏立飞,张杨熙, 等. 优化 CARS 结合 PSO - SVM 算法农田土壤重金属砷含量高光谱反演分析[J]. 光谱学与光谱分析, 2020, 40(2): 567 - 573.
YUAN Ziran, WEI Lifei, ZHANG Yangxi, et al. Hyperspectral inversion and analysis of heavy metal arsenic content in farmland soil based on optimizing cars combined with PSO - SVM algorithm[J]. Spectroscopy and Spectral Analysis, 2020, 40(2): 567 - 573. (in Chinese)

[10] 郭云开,张思爱,谢晓峰, 等. 基于 GA - SVM 的耕地土壤重金属含量高光谱反演方法的研究[J]. 土壤通报, 2021, 52(4): 968 - 974.
GUO Yunkai, ZHANG Siai, XIE Xiaofeng, et al. The hyperspectral inversion method of heavy metal contents incultivated soils based on GA - SVM[J]. Chinese Journal of Soil Science, 2021, 52(4): 968 - 974. (in Chinese)

[11] 林楠,刘翰霖,孟祥发,等. 基于高光谱的黑土区土壤重金属含量估测[J]. 农业机械学报,2021,52(3):218 - 225.
LIN Nan, LIU Hanlin, MENG Xiangfa, et al. Hyperspectral estimation of heavy metal contents in black soil region[J]. Transactions of the Chinese Society for Agricultural Machinery,2021,52(3):218 - 225. (in Chinese)

[12] 吕杰,郝宁燕,崔晓临. 利用可见光近红外的尾矿区农田土壤 Cu 含量反演[J]. 农业工程学报, 2015, 31(9): 265 - 270.
LÜ Jie, HAO Ningyan, CUI Xiaolin. Inversion model for copper content in farmland of tailing area based on visible-near infrared reflectance spectroscopy[J]. Transactions of the CSAE, 2015,31(9): 265 - 270. (in Chinese)

[13] WANG Fenghe, GAO J, ZHA Yong. Hyperspectral sensing of heavy metals in soil and vegetation: feasibility and challenges [J]. ISPRS J Photogramm, 2018, 136: 73 - 84.
[14] 程渊,李玉霞,李凡, 等. 基于极端随机树的闪电河流域土壤水分反演[J]. 遥感学报, 2021, 25(4): 941 - 951.
CHENG Yuan,LI Yuxia,LI Fan, et al. Soil moisture retrieval using extremely randomized trees over the Shandian River Basin [J]. National Remote Sensing Bulletin, 2021, 25(4):941 - 951. (in Chinese)
[15] 王动民, 纪俊敏, 高洪智. 多元散射校正预处理波段对近红外光谱定标模型的影响[J]. 光谱学与光谱分析, 2014, 34(9): 2387 - 2390.
WANG Dongmin, JI Junmin, GAO Hongzhi. The effect of MSC spectral pretreament regions on near infrared spectroscopy calibration results[J]. Spectroscopy and Spectral Analysis, 2014, 34(9): 2387 - 2390. (in Chinese)
[16] 卓荦. 基于高光谱遥感的土壤重金属空间分布研究[D]. 武汉:武汉大学,2010.
ZHUO Luo. The research of estimating heavy metal spatial distribution of soil using hyperspectral data[D]. Wuhan: Wuhan University, 2010. (in Chinese)
[17] 王涛,喻彩丽,张楠楠, 等. 基于去包络线和连续投影算法的枣园土壤电导率光谱检测研究[J]. 干旱地区农业研究, 2019,37(5):193 - 199,217.
WANG Tao, YU Caili, ZHANG Nannan, et al. Spectral detection of electrical conductivity in jujube orchard soil based on continuum-removal and SPA[J]. Agricultural Research in the Arid Areas, 2019, 37(5): 193 - 199,217. (in Chinese)
[18] MÁRIO C U A, SALDANHA T C B, GALVO R K H, et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis[J]. Chemometrics & Intelligent Laboratory Systems, 2001, 57(2):65 - 73.
[19] DESMET G. A GIS procedure for automatically calculating the USLE LS factor on topographically complex landscape units[J]. Journal of Soil and Water Conservation, 1996,51(5):427 - 433.
[20] HENGL T, REUTER H. Geomorphometry-concepts, software, applications[M]. Amsterdam: Elsevier, 2009:195 - 226.
[21] 于海洋,罗玲,马慧慧, 等. SRTM(1″)DEM 在流域水文分析中的适用性研究[J]. 国土资源遥感, 2017, 29(2): 138 - 143.
YU Haiyang, LUO Ling, MA Huihui, et al. Application appraisal in catchment hydrological analysis based on SRTM 1 Arc-Second DEM[J]. Remote Sensing for Land and Resources, 2017, 29(2): 138 - 143. (in Chinese)
[22] BREIMAN L. Random forests[J]. Machine Learning, 2001,45(1): 5 - 32
[23] GEURTS P, ERNST D,WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3 - 42
[24] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
[25] SAEYS W, MOUAZEN A M, RAMON H. Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy[J]. Biosystems Engineering, 2005, 91(4):393 - 402.

(上接第 219 页)

[17] CAO W, QIAO Z, GAO Z, et al. Use of unmanned aerial vehicle imagery and a hybrid algorithm combining a watershed algorithm and adaptive threshold segmentation to extract wheat lodging[J]. Physics and Chemistry of the Earth, Parts A/B/C, 2021, 123: 103016.
[18] ZHANG D, DING Y, CHEN P, et al. Automatic extraction of wheat lodging area based on transfer learning method and DeepLab v3 + network[J]. Computers and Electronics in Agriculture, 2020, 179: 105845.
[19] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481 - 2495.
[20] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2881 - 2890.
[21] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 833 - 851.
[22] 杨蜀秦, 宋志双, 尹瀚平, 等. 基于深度语义分割的无人机多光谱遥感作物分类方法[J]. 农业机械学报, 2021, 52(3): 185 - 192.
YANG Shuqin, SONG Zhishuang, YIN Hanping, et al. Crop classification method of UVA multispectral remote sensing based on deep semantic segmentation[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(3): 185 - 192. (in Chinese)
[23] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint, 2017, arXiv:1706.05587.
[24] 宁纪锋, 倪静, 何宜家, 等. 基于卷积注意力的无人机多光谱遥感影像地膜农田识别[J]. 农业机械学报, 2021, 52(9): 213 - 220.
NING Jifeng, NI Jing, HE Yijia, et al. Attention based plastic mulching farmland identification via UAV multispectral remote sensing image[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(9): 213 - 220. (in Chinese)
[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//NIPS 2017, 2017: 30.
[26] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3 - 19.
[27] YANG L, ZHANG R Y, LI L, et al. SimAm: a simple, parameter-free attention module for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2021: 11863 - 11874.