doi: 10.6046/zrzyyg.2020378
引用格式: 肖烨辉,宋妮迪,孟盼盼,等. 模型集群分析策略联合 ELM 的土壤重金属 Pb 含量预测[J]. 自然资源遥感,2021,33(4):143 - 152. (Xiao Y H, Song N D, Meng P P, et al. Prediction of lead content in soil based on model population analysis coupled with ELM algorithm[J]. Remote Sensing for Natural Resources, 2021, 33(4): 143 - 152.)

模型集群分析策略联合 ELM 的土壤重金属 Pb 含量预测

肖烨辉[1], 宋妮迪[2], 孟盼盼[2], 王培俊[2], 范胜龙[2]
(1. 福建农林大学资源与环境学院,福州 350007; 2. 福建农林大学公共与管理学院,福州 350007)

摘要: 为探寻区域土壤重金属含量最佳反演模型,以龙海市为研究区,对土壤原始光谱数据分别进行 SG 平滑、小波变换、高斯滤波和多元散射校正 4 种光谱预处理,运用基于模型集群分析(model population analysis, MPA)策略开发的波长选择算法: 竞争适应性重加权采样算法(competitive adaptive reweighted sampling, CARS)、变量空间迭代收缩算法(variable iterative space shrinkage approach, VISSA)、迭代变量子集优化算法(iteratively variable subset optimization, IVSO)和区间组合优化算法(interval combination optimization, ICO)剔除干扰与无信息波长变量,采用线性模型偏最小二乘回归(partial least squares regression, PLSR)、非线性模型支持向量机(support vector machine, SVM)及神经网络模型极限学习机(extreme learning machine, ELM)进行土壤重金属铅(Pb)含量回归预测。结果表明: 经过多种预处理方法建立的 Pb 含量反演模型中,基于小波变换第七层重构后的光谱数据构建的模型预测精度最优,其验证集 $R^2 = 0.736$, $RMSE = 5.426$, $RPD = 1.976$, $RPIQ = 2.560$。基于 MPA 策略开发的 CARS, VISSA, IVSO 和 ICO 都能显著提升模型解释性与泛化性能,并且提高建模效率。3 种回归模型总体的预测表现排序: ELM > PLSR > SVM。其中 ICO - ELM 预测精度最高,其验证集 $R^2 = 0.863$, $RMSE = 3.953$, $RPD = 2.712$, $RPIQ = 3.514$。所建最优模型可为区域土地质量和生态指标快速准确监测提供新的理论参考。

关键词: 模型集群分析策略; 小波变换; 区间组合优化; 极限学习机
中图法分类号: TP 79 **文献标志码**: A **文章编号**: 2097 - 034X(2021)04 - 0143 - 10

0 引言

伴随着城市化以及矿产、电镀、冶金等产业的不断发展,因工业和生活污水灌溉、涉重金属企业“三废”排放、矿产开发等原因导致的土壤重金属污染问题已经被广泛关注[1]。重金属因其不易被土壤微生物分解、易于积累的特性,在土壤中不断富集。当土壤中重金属含量超过界限值时,会影响作物对氮、磷、钾营养元素的吸收从而抑制作物的生长,甚至会通过食物链在人体内蓄积,间接危害人体的健康[2]。探寻高效准确地获取大区域土壤重金属含量信息的方法对进行污染风险评价与土壤修复工作具有重要意义。相比于传统实验室理化分析方法,高光谱技术以其快速、无损、无污染、低成本等优势,近年来被广泛的应用于森林生态系统结构参数估算[3]、食品品质检测[4]、煤矿分类[5]和土壤理化性质预测[6-7]等领域中。

然而利用高光谱技术中的可见光—近红外光谱数据建立目标成分定量分析模型时,由于光谱波段数众多,常常面临着建模样本量远远小于光谱变量数而导致模型预测结果不佳的问题。此外由于近红外波段受到 C - H, C - O 和 N - H 等官能团的合频和倍频吸收峰相互重叠的影响,波长变量之间高度相关,多重共线性显著[8]。上述问题连同波段中大量存在的无信息与干扰变量将会严重影响模型的可解释性、计算复杂性、稳定性与泛化性能。因此,通过波长选择算法从可见光—近红外全谱数据中优选出最佳波长组合提高预测模型性能显得尤为重要。基于 Li 等[9]提出的模型集群分析(model population analysis, MPA)策略开发的波长选择算法近年来逐渐展露头角,开始被应用于各类高光谱反演模型之

收稿日期: 2020 - 12 - 01; **修订日期**: 2021 - 03 - 03
基金项目: 福建省自然科学基金面上项目“生物炭和脱硫石膏改良滨海滩涂新围垦耕地的耦合效应及其机制”(编号: 2019J01397)资助。
第一作者: 肖烨辉(1997 -),男,硕士,主要研究方向为农业环境保护。Email: 260939662@qq.com。
通信作者: 范胜龙(1976 -),男,博士,教授,主要研究方向为土地资源可持续利用。Email: fsl@fafu.edu.cn。

中。于雷等[10]利用迭代和保留信息变量法(iteratively retains informative variables,IRIV)算法对大豆叶片叶绿素含量进行了预测,结果表明 IRIV 算法能有效保留光谱中与叶绿素有关的弱信息变量,提升模型估测能力;刘贵珊等[11]对比区间变量迭代空间收缩算法(interval variable iterative space shrinkage approach,IVISSA)、变量组合集群分析法(variable combination population analysis,VCPA)、竞争适应性重加权采样算法(competitive adaptive reweighted sampling,CARS)和连续投影算法(successive projection algorithm,SPA)4 种算法估算冷鲜滩羊肉嫩度,得出 IVISSA 算法的预测精度与模型鲁棒性要高于其他算法。基于 MPA 的波长选择算法打破了以往基于单一模型或目标函数进行变量筛选的思路,通过大量建立子模型,从模型集总体来分析特定评价参数,并根据参数值赋予各波长高低不同的采样权重,从而逐步优化收缩变量空间,达到特征波长筛选目的。能够充分挖掘出小样本数据的有用信息,减小偶然性误差,提升模型稳定性与泛化性能[12]。

变量空间迭代收缩算法(variable iterative space shrinkage approach,VISSA)[13]、迭代变量子集优化算法(iteratively variable subset optimization,IVSO)[14]、区间组合优化算法(interval combination optimization,ICO)[15]和 CARS 是 4 种基于 MPA 策略的代表性波长选择算法。相比于 CARS,另外 3 种算法在土壤成分高光谱预测研究中鲜有文献报道。因此,本研究以国控重金属污染重点防控区龙海市为研究区,选择重金属 Pb 为研究对象,综合对比这 4 种波长优选算法,并联合线性模型偏最小二乘回归(partial least squares regression,PLSR)、非线性模型支持向量机(support vector machine,SVM)和神经网络模型极限学习机(extreme learning machine,ELM),研究各波长选择算法耦合各类型回归模型在土壤重金属反演领域的预测表现,探寻研究区内土壤重金属 Pb 含量的最佳反演模型,为区域重金属污染高效监测提供理论和技术支持。

1 研究区及其数据源

1.1 研究区概况

龙海市(E117°29′~118°14′,N24°11′~24°36′)地处福建省东南部,位于九龙江出海口处。中部为冲积平原,北部、西部、南部三面环山,东南部临海,总占地面积约 1 128 km^2。该地属于亚热带季风气候,年均气温为 21.5 ℃,年均降雨量为 1 563.2 mm。主要土壤类型包括红壤、水稻土、滨海砂泥及潮土等。龙海市涉重金属企业分布较多,长期以来的污染排放以及农业不合理的耕作施肥等原因,导致龙海市土壤中以 Pb 为代表的部分重金属含量超出了环境背景值[16]。

1.2 土壤样品采集和理化分析

利用卫星影像图对龙海市进行样本初步布设,共设代表性样本点数 79 个,涉及土地利用类型耕地、林地、园地等。根据初始布点坐标实地采样时,避开田埂、沟渠及马路边缘等特殊区域,利用蛇形采样法对每个采样点挖取 0~20 cm 表层土壤 5 份,每份土样挖取时用竹片刮去与金属制铁铲接触面,均匀混合后用四分法取 500 g 装进密封袋内作为一份样本。同时利用手持 GPS 校正样本真实经纬度坐标,最终采集的样本分布见图 1。

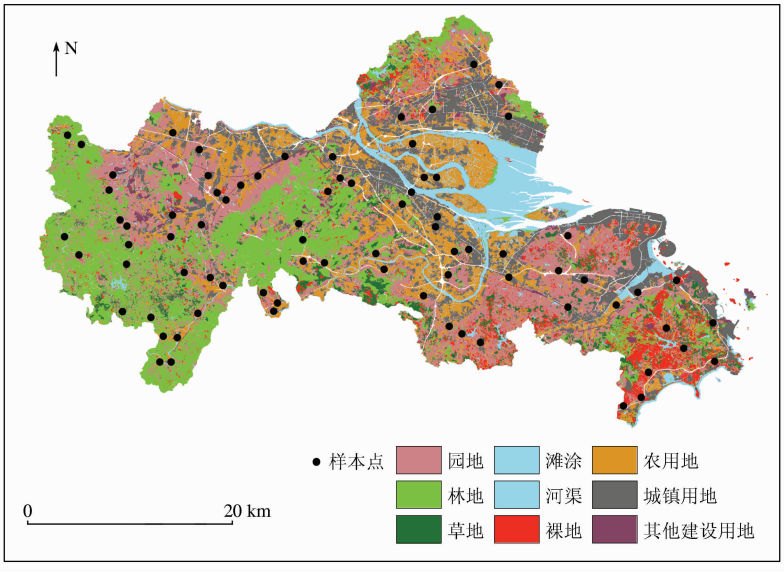


图 1 研究区位置及采样点分布

Fig. 1 Location of the study area and distribution of soil sampling sites

采回的样本放置室内风干数日,之后去除植物残体、小石块等杂质,用木棒研磨至过 100 目尼龙筛。研磨后的土样按样本号分成 2 份,分别用于光谱数据及 Pb 含量的测定。检测 Pb 含量时,将土样经氢氟酸 - 高氯酸 - 硝酸反复消煮等处理后由电感耦合等离子体质谱仪(ICP - MS)最终测定。

1.3　光谱数据采集

土壤光谱数据的采集使用美国 ASD(Analytical Spectral Device)公司生产的 ASD FieldSpec 3 Pro 地物光谱仪。光谱波段范围为 350 ~ 2 500 nm,其中 350 ~ 1 000 nm 和 1 000 ~ 2 500 nm 之间采样间隔分别为 1.4 nm 与 2 nm,经光谱重采样后的间隔为 1 nm。为了减少非环境光源对光谱测试的影响,实验在一个封闭的暗室内进行。测定前先将光谱仪开机预热 15 min 以上,之后进行暗电流校准以及对参考板进行优化。将待测土样放置于 60 mm 规格的玻璃培养皿中,样品浸没深度大于 0.5 cm,以免光源透射对光谱测定精度产生影响。选用 1 000 W 的卤素灯作为光源,卤素灯天顶角为 15°,光谱仪探头距离样品 10 cm。探头垂直于土壤表面进行测量,每个土壤样品共采集 10 条光谱曲线,取平均值作为原始光谱数据。由于光谱边缘波段产生了大量不稳定的噪声,所以最终去除边缘波段后取 400 ~ 2 400 nm 范围内共 2 001 个波段数据进行研究。

2　研究方法

2.1　建模集样本划分与光谱数据预处理

为防止因异常样本的存在对模型结果预测精度以及稳定性产生的影响,利用基于蒙特卡洛交叉验证的预测残差法[17]识别并剔除了 3 个异常样本。对剔除后共计 76 个样本采用光谱理化值共生距离法划分样本,将土壤样本空间划分为建模集(占比 75%)与验证集(占比 25%),结果见表 1。该算法通过样本间欧式距离划分样本,同时考虑了光谱自变量空间与因变量空间的距离差异性,使得建模集与验证集样本与总体样本拥有近似的统计分布特征,有助于提升模型的泛化能力。上述方法均基于 MATLAB2017a 软件实现。

表 1　土壤 Pb 含量统计特征

Tab. 1　Statistical characteristics of Pb of soil samples

重金属类型	样本集	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	全样本	15.88	95.01	50.98	15.83	0.31
	建模集	18.87	84.39	50.80	15.46	0.30
	验证集	31.99	65.48	49.50	10.72	0.22

在进行光谱测量时,实验环境、机器自身误差、土壤基质与背景干扰等因素引起的随机噪声会将数据中有用信息遮蔽,从而影响模型回归结果。故本文分别利用 Savizky Golay(SG)平滑、小波变换(wavelet transform, WT)、高斯滤波(Gaussian filter, GF)和多元散射校正(multiple scatter correction, MSC)4 种预处理方法对原始光谱数据进行处理,以此提升数据信噪比。其中 WT 变换利用 MATLAB2017a 软件小波工具箱实现,其余预处理方法采用 The Unscrambler X 软件完成。

2.2　光谱特征波长选择方法

2.2.1　VISSA

VISSA 首次引入了加权二进制采样方法(weighted binary matrix sampling, WBMS),每轮采样根据上轮不同变量在表现最佳的模型集里出现的频率赋予各个变量新的权重,以达到不断优化收缩变量空间的目的,最终获取最优的变量组合。主要步骤如下:

1)利用 WBMS 对光谱变量空间进行随机采样,每个变量初始采样权重设为 0.5,随机采样 N 次,共生成 N 个波长变量子集。

2)基于每个波长子集分别建立 PLSR 模型,计算每个模型的交叉验证均方根误差(*RMSECV*)。选出 *RMSECV* 最低的一部分模型作为最优模型集,选取比例记为 α。之后计算每个变量在最优模型集的出现频率,作为下次采样该变量的权重。同时记录最优模型集的 *RMSECV* 均值。第 k 个变量权重计算公式为:

$$w_k = \frac{f_k}{N_{\text{best}}}, \tag{1}$$

式中: w 为权重; f 为变量出现在最优模型集的频次; N_{best} 为最优模型集的数量。

3)重复执行 1)—2)步骤,其中 WBMS 采样权重每轮根据式(1)进行更新,直至新一轮 *RMSECV* 均值无法进一步改善。此时根据前一轮采样权重,将权重为 1 的光谱波段记为信息波段,权重为 0 的记为干扰波段,其余波段记为弱信息波段。

4)将信息波段、干扰波段、弱信息波段按权重值进行降序排序,从大到小逐个加入变量集建立 PLSR 模型计算 *RMSECV*,取 *RMSECV* 最小值对应的模型变量组合做为最终的特征波长集。

2.2.2　IVSO

IVSO 以不同模型变量的回归系数为研究对象进行统计分析,以此决定每一轮迭代不同波长变量的权重,从而逐步优选出最佳波长子集。IVSO 的实

现步骤为：

1）使用 WBMS 生成 N 个波长变量子集，记录该轮 WBMS 选中的变量数为 L_1。

2）基于每个波长子集分别建立 PLSR 模型，建立回归系数矩阵 $\boldsymbol{B}$，即

$$\boldsymbol{B} = [b_1, b_2, \cdots, b_N]^{\mathrm{T}} \quad , \tag{2}$$

式中 $b_j(1<j<N)$ 代表第 j 个波长子集建立 PLSR 模型每个变量的回归系数绝对值（*absRC*）。接着将 b_j 中每个变量的 *absRC* 进行归一化处理，即

$$c_{ij} = \frac{b_{ij}}{\max(b_j)} \quad , \tag{3}$$

式中：b_{ij} 为 b_j 中第 i 个变量的 *absRC*；$\max(b_j)$ 表示 b_j 中所有 *absRC* 的最大值；c_{ij} 为经处理后的 b_{ij}，所有 c_{ij} 组合成归一化回归系数矩阵 $\boldsymbol{C}$。将新矩阵 $\boldsymbol{C}$ 的每一列进行求和得到行向量 $\boldsymbol{s}$ 为 $[s_1, s_2, \cdots, s_N]$，s_i 表示为第 i 个变量在 N 个子模型中经归一化处理后 *absRC* 的总和。将 L_1 个变量按照 s_i 值从大到小进行重要性排序，值越大代表对回归方程贡献越大。按顺序逐个将 L_1 个变量加入变量集建立 L_1 个 PLSR 模型，记录 L_1 个模型 *RMSECV* 最小值 min*RMSECV*，以及对应模型的变量数 L_2。

3）在下一次迭代中，变量采样权重被定义为：

$$w_i = \frac{s_i}{\max(\boldsymbol{s})}, \ i = 1, 2, \cdots, N \quad , \tag{4}$$

式中：$\max(\boldsymbol{s})$ 为行向量 $\boldsymbol{s}$ 中的最大值；w_i 为第 i 个变量的采样权重。

4）重复步骤 1）—3），直至 $L_1 = L_2$。比较每轮迭代生成的 *minRMSECV*，最小值对应的波长组合即为最终的特征波长集。

2.2.3 ICO

ICO 区别于本文其他波长选择算法，以波长区间代替波长点作为优化对象，利用加权自举采样（weighted bootstrap sampling，WBS）逐步收缩优化波长区间组合，最后结合局部搜索策略进一步优化各个波长区间边缘波段。ICO 的具体步骤为：

1）将光谱变量空间均匀分成 M 个等长的波长子区间。

2）利用 WBS 从 M 个波长区间里采样出 N 个不同波长区间组合，每个波长区间初始权重设置为 1。根据权重决定各个波长区间被选中概率，即

$$p_k = \frac{w_k}{\sum_{1}^{n} w_k} \quad , \tag{5}$$

式中：波长数量为 n；p_k 表示第 k 条波长被选择的概率；w_k 表示第 k 条波长的采样权重。

3）基于 N 个波长区间组合分别建立 PLSR 模型，计算各个模型的 *RMSECV*，取 *RMSECV* 最小值的部分模型集为最优模型集，选取比例记为 α。计算每个波长区间在最优模型集内出现的频率，做为下次采样的权重，同公式（1）。同时记录这轮采样最优模型集的 *RMSECV* 均值。

4）重复步骤 2）—3），直至此轮 *RMSECV* 均值高于上一轮。记录上一轮迭代产生的所有模型中 *RMSECV* 最小值的模型对应的波长区间组合。

5）基于上一步产生的最优波长区间组合，采用局部搜索策略优化每个波长区间边缘波段，将波长区间边缘相邻的波长点分别纳入或剔除进行建模，通过 *RMSECV* 变化来判断该波长点的选入或排除。重复进行数次，直至 *RMSECV* 不受波长点的加入或剔除所影响，此时所选中的波长区间组合即为最优特征波长集。

2.2.4 CARS

不同于上述算法从变量空间进行随机采样生成模型子集，CARS 利用蒙特卡洛采样法从样本空间随机选取样本建立 PLSR 模型，之后以变量回归系数大小作为权重判别指标，结合指数衰减函数（exponentially decreasing function，EDF）与适应性重加权采样法（adaptive reweighted sampling，ARS）对低权重样本进行剔除达到变量优选目的。最后通过迭代建立大量子模型确定最优特征波长集。

2.3 回归模型

基于特征波长选择算法优选出的波长子集，选择线性回归模型 PLSR、非线性模型 SVM 和神经网络模型 ELM 进行土壤重金属 Pb 含量的预测。其中 PLSR 与 SVM 因其应对小样本多变量数据的出色能力，而被广泛应用于高光谱数据的研究之中；ELM 是一种前向传播神经网络，由输入层、隐含层及输出层 3 层网络结构组成。在运行前只需设置隐含层神经元的个数，并且可随机生成输入层到隐含层的权重以及隐含层各节点阈值，省去了反复训练调参的时间，因此具有训练速度快、泛化能力强等优点[18]。

模型评价指标选用决定系数（R^2）、均方根误差（*RMSE*）、相对分析误差（*RPD*）、四分位相对预测误差（*RPIQ*）。其中 *RPIQ* 为验证集四分位距（第三四分位数与第一四分位数的差值）与 *RMSE* 的比值，由于土壤重金属含量分布并不对称，不符合正态分布规律，*RPIQ* 被认为比 *RPD* 更适合评价土壤重金属反演模型[19]。R^2 越接近 1，*RMSE* 越小以及 *RPD* 和 *RPIQ* 越高则模型回归表现越佳。本研究各波段选择算法及回归模型均基于 MATLAB2017a 软件实现。

3　结果与讨论

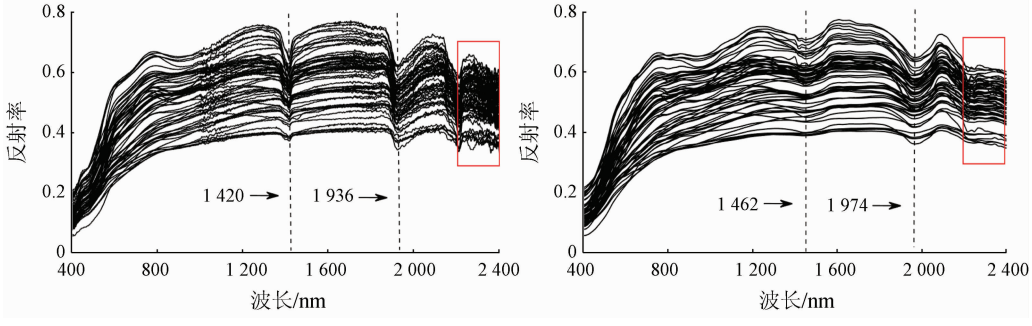
3.1　光谱数据预处理结果

基于不同光谱预处理方法分别建立 PLSR 模型，结果如表 2 所示。WT 中不同基函数的选择，会造成模型结果的差异，通过多次试验比较，最终选用 db4 作为基小波。光谱数据经 WT 多尺度分解后，最后通过逆变换得到 8 层特征光谱，各层重构光谱用 WT_1 ~ WT_8 进行表征。由表 2 可知，除 MSC 外，其他 3 种预处理方法模型精度都得到了不同程度的提升，这可能是因为本次实验土壤经研磨后颗粒大小相对均匀，颗粒间散射影响较小造成。其中基于 WT_7 构建的模型回归效果最优，验证集 R^2 相比原始光谱提升了 4.69%，*RPIQ* 提高了 7.02%。WT_7 与原始光谱如图 2 所示。相比于原始光谱图像，重构后的光谱曲线 2 200 nm 附近因水分子吸收形成的波谷与 590 nm 处小波峰的特征被淡化，1 420 nm 和 1 936 nm 附近的水分吸收深度也都不同程度的降低，并且波谷向右发生了些微偏移。但同时也平滑了大量的“毛刺”噪声，尤其是 2 200 ~ 2 400 nm 之间光谱曲线变得光滑，使得模型总体信噪比提升，回归精度得到了有效改善，因此后续研究将基于 WT_7 展开。

表 2　基于 PLSR 模型的不同预处理方法

Tab. 2　Different pre – processing methods based on PLSR models

预处理方法	建模集		验证集			
	R^2	*RMSE*	R^2	*RMSE*	*RPD*	*RPIQ*
None	0.659	8.580	0.703	5.807	1.846	2.392
WT_1	0.659	8.583	0.714	5.655	1.896	2.456
WT_2	0.658	8.589	0.714	5.654	1.896	2.457
WT_3	0.658	8.598	0.714	5.651	1.897	2.458
WT_4	0.657	8.611	0.714	5.655	1.896	2.456
WT_5	0.656	8.623	0.717	5.625	1.906	2.469
WT_6	0.670	8.439	0.703	5.759	1.862	2.412
WT_7	0.665	8.502	0.736	5.426	1.976	2.560
WT_8	0.664	8.524	0.681	6.018	1.781	2.308
SG	0.667	8.458	0.723	5.604	1.913	2.479
GF	0.660	8.564	0.718	5.615	1.909	2.474
MSC	0.659	8.582	0.702	5.818	1.843	2.387



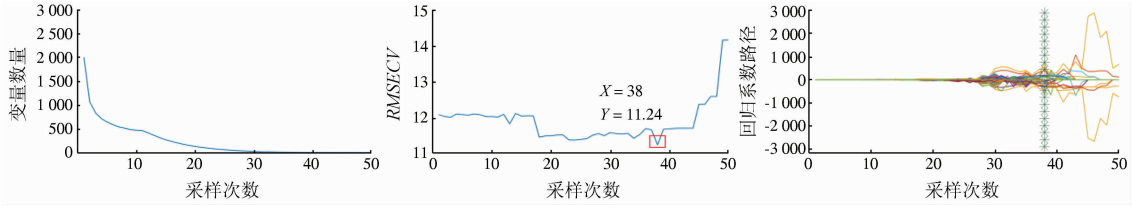
（a）原始光谱　　（b）WT_7 重构光谱

图 2　土壤样本光谱曲线

Fig. 2　Spectra of soil samples

3.2　特征波长优选结果

图 3 为基于 CARS 的特征波段筛选过程。其中图 3（a）为光谱变量数在迭代过程中的变化。在 EDF 函数强制筛选波长的作用下，变量数呈现指数形式衰减，前期迭代时，下降速度较快，之后趋于平缓，可大致分为粗选和精选 2 个波段筛选过程。图 3（b）是每轮迭代 *RMSECV* 的变化，可以看出，在粗选阶段，虽然剔除了大量变量，*RMSECV* 并没有呈现出明显下降趋势，总体变化起伏较小，可能此时去除的绝大部分波段为无信息变量，对模型预测精度影响较小。综合图 3（c）所有波长变量的回归系数路径变化趋势，选择第 38 次迭代时模型建模变量为最佳特征波长子集，此时 *RMSECV* 取得最小值 11.24。

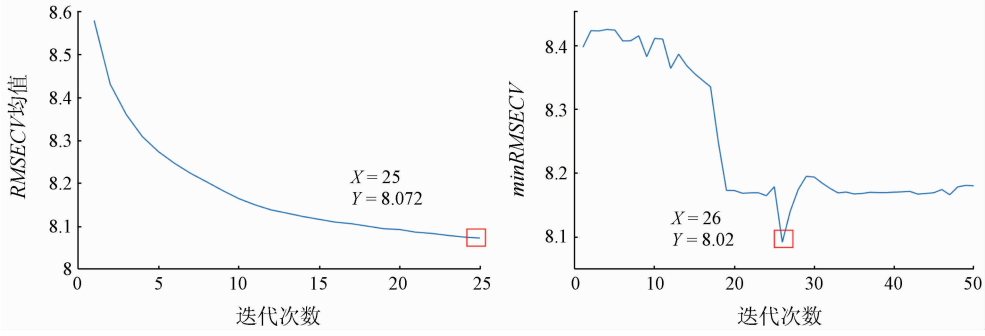


（a）变量数量变化趋势　　（b）*RMSECV* 变化趋势　　（c）每个变量回归系数变化

图 3　CARS 算法筛选变量

Fig. 3　Variables selected by CARS

VISSA 每轮 WBMS 采样次数 N 设为 5 000，最优模型集比例 α 设为 0.05。从图 4(a)看出 VISSA 迭代过程中 *RMSECV* 均值变化曲线相对光滑，下降速度由快至慢，共进行 25 轮迭代。IVSO 每轮采样次数 N 取 5 000，该算法以 *minRMSECV* 在迭代全过程中的最小值作为收敛条件，如图 4(b)所示，曲线一开始变化较慢，随后急剧下降，这可能是因为此时已经剔除了大部分与模型无关的干扰变量，使得模型精度显著提升。在下降至第 26 轮时达到最低值，min*RMSECV* = 8.02，之后缓慢上升至趋于平稳。因此选择第 26 次迭代中 min*RMSECV* 所对应模型的变量集作为最终的波长优选子集。

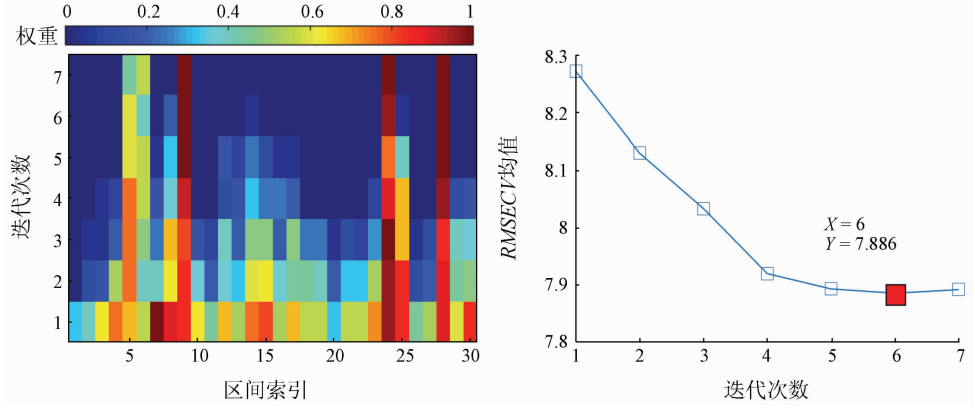


(a) 迭代过程中 *RMSECV* 均值变化　　(b) 迭代过程中 *minRMSECV* 变化

图 4　VISSA 和 IVSO 算法迭代过程

Fig. 4　Iterative process of VISSA and IVSO

基于 ICO 进行优选波段时，波长子区间等分数设为 30 个，WBS 每轮采样次数取 1 000，最优模型集比例 α 设为 0.05。图 5 为 ICO 筛选变量过程。其中图 5(a)为每轮采样各区间的权重变化，深蓝到深红的颜色渐变代表权重值随迭代过程进行的不断增大。ICO 利用的 WBS 采样方法相较于 WBMS 区别在于，即使某一波段因为偶然因素在前一轮采样中权重变为 1，在之后的迭代中仍有可能将其排除在外。可以看出第 7 个子区间采样权重初始为 1，但随着不断的迭代，其重要性逐渐降低，权重值在最后一次迭代中归零，因此未被选入最优子区间集，说明 ICO 在进行波长选择时的容错性较高。图 5(b)为每轮迭代 *RMSECV* 均值的变化趋势，仅通过 6 轮迭代便达到收敛条件。可能是因为 ICO 以 30 个波长区间为建模对象，相对于全波长，不同变量间的组合计算量要显著下降。



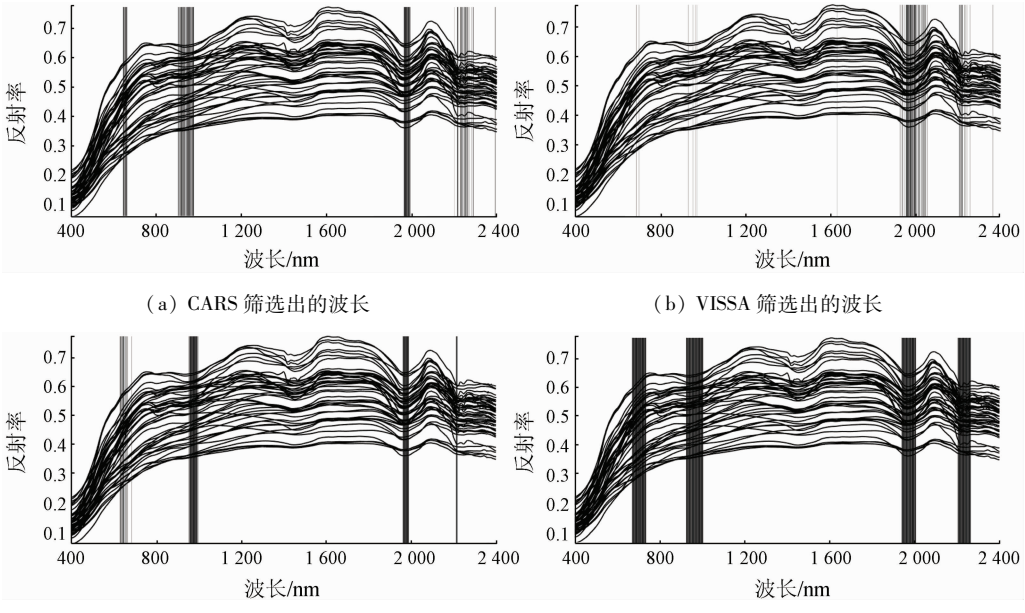
(a) 迭代过程中各波长区间权重变化　　(b) 每轮迭代 *RMSECV* 均值变化

图 5　ICO 算法迭代过程

Fig. 5　Iterative process of ICO

全光谱 2 001 个波段经过 CARS，VISSA，IVSO 和 ICO 这 4 种波段选择算法优选后变量大大减少，4 种算法选出的波段数分别为 119，78，98 和 276 个。图 6 给出了各种波段选择方法优选出的波段分布位置。可以看出波段总体分布范围相似，主要集中在近红外光谱区域 904 ~ 1 003 nm，1 927 ~ 2 055 nm，2 201 ~ 2 395 nm 及可见光光谱区域 627 ~ 734 nm 这 4 个区间之内。其中 600 ~ 800 nm 与土壤中有机质相关；900 nm 与 1 900 nm 附近光谱特性主要受土壤铁氧化物、黏土矿物和有机质的综合影响；2 250 nm 附近除受水分子吸收影响外还与黏土矿物中的 O - H 基团伸缩振动有关。优选的波段主要分布于这 4

个区间内可能是因为重金属 Pb 离子常常吸附于铁氧化物、黏土矿物和有机质这 3 种物质上,所以与其呈现出较强的相关性。此外,因为光谱采样间隔较短,相邻波段往往蕴含着相似的信息特征,所以各算法选出的波段呈现出一定的连续性。



(a) CARS 筛选出的波长　　(b) VISSA 筛选出的波长

(c) IVSO 筛选出的波长　　(d) ICO 筛选出的波长

图 6　不同波长选择算法优选出的波段

Fig. 6　Selected wavelengths of different wavelength selection methods

3.3　回归模型选择

利用 CARS,VISSA,IVSO 和 ICO 优选出的波段分别建立 PLSR,SVM 和 ELM 这 3 种回归模型,并将 PLSR 结合全波段建模结果进行对比,以此探寻波段优选算法的实质表现。其中 ELM 和 SVM 模型参数设置如下: SVM 核函数选择高斯核; ELM 最佳隐含层神经元个数设为 20,激活函数选择 ReLU 函数,结果如表 3 所示。

表 3　3 种回归模型联合各波段选择算法预测结果

Tab. 3　Prediction results of three regression models based on different wavelength selection methods

模型	波长选择方法	变量数量	建模集		验证集			
			R^2	*RMSE*	R^2	*RMSE*	*RPD*	*RPIQ*
PLSR	全波段	2 001	0.702	8.450	0.736	5.426	1.976	2.560
	CARS	119	0.726	6.958	0.759	5.235	2.048	2.653
	VISSA	78	0.720	7.033	0.780	5.003	2.143	2.777
	IVSO	98	0.718	7.050	0.802	4.748	2.258	2.926
	ICO	276	0.720	7.023	0.813	4.610	2.325	3.013
SVM	CARS	119	0.643	7.939	0.735	5.485	1.954	2.532
	VISSA	78	0.645	7.915	0.745	5.385	1.991	2.579
	IVSO	98	0.630	8.075	0.757	5.252	2.041	2.645
	ICO	276	0.631	8.069	0.770	5.116	2.095	2.715
ELM	CARS	119	0.814	6.338	0.806	4.703	2.279	2.953
	VISSA	78	0.861	4.948	0.837	4.304	2.491	3.227
	IVSO	98	0.899	4.229	0.858	4.013	2.671	3.461
	ICO	276	0.877	4.653	0.863	3.953	2.712	3.514

相比于以全波段为自变量进行建模,PLSR 各模型结果都得到了明显的改善,说明 4 种波段选择算法通过优选变量,不但能减少模型的计算复杂度,还能有效提升模型的预测性能。对比 3 种回归模型里 4 种波长选择算法的表现,发现 CARS 回归精度普遍低于 VISSA,IVSO 和 ICO 这 3 种方法。可能是因为 CARS 每轮采样后利用 EDF 函数对波段进行了强制性剔除,从而误将部分对模型回归有用的弱信息波段也排除在外,导致模型拟合效果下降。相比于 CARS,另外 3 种算法则采用了一种柔性收缩变量空间的策略[20]。每轮采样并不会硬性剔除波长变量,而是根据不同权重赋值方法,给予各个波长或

波长区间高低不同的权重值,使得即使在前一轮因偶然因素表现欠佳的部分信息变量仍然有机会在下一轮被选中。VISSA 的精度低于 IVSO 和 ICO 的原因可能是其意外选出了 4 个相关区间外的波长,导致模型混入无信息或干扰变量。总体而言,基于 MPA 策略的波长选择算法,通过迭代建立大量子模型,具有较强的稳定性与泛化性能。但由于模型数量庞大,VISSA 与 IVSO 以单一波长点为建模变量,计算量较大。CARS 迭代模型数量较少,且利用 EDF 函数每轮强制剔除一定数量变量后,计算复杂度降低,但泛化表现低于其他算法。4 种方法中 ICO 在 3 种回归模型中的表现最优。具有优秀泛化能力的同时,其以波长区间作为建模变量,不但能提升模型的解释能力和运算效率,因参与变量过多而容易引发的过拟合概率也大大下降。

ELM 各模型回归总体表现显著优于 SVM,CARS – ELM,VISSA – ELM,IVSO – ELM 和 ICO – ELM 相较于 CARS – SVM,VISSA – SVM,IVSO – SVM 和 ICO – SVM,验证集 R^2 分别提升了 9.66%,12.35%,13.34% 和 12.08%,*RPIQ* 提升了 16.63%,25.13%,30.85% 和 29.43%。基于特征波长建立的 SVM 模型在土壤重金属 Pb 含量预测研究中相比另外 2 种模型表现一般,3 种模型按回归总体表现排序为:ELM > PLSR > SVM。PLSR 模型的精度高于 SVM 可能是因为 SVM 很大程度上依赖核函数与参数的选择,加上研究所用波长选择算法是基于 PLSR 建立的大量子模型,不能较好地适用于 SVM。而神经网络模型 ELM 因其处理非线性数据泛化能力强、不易陷入局部最优解等优点,能很好地弥补 PLSR 面对非线性的高光谱数据的不足。回归模型与波长选择算法的各种耦合中,ICO – ELM 表现最为优异,其验证集 R^2 和 *RPIQ* 达到了 0.863 及 3.514。ICO – ELM 模型验证集样本预测值与实测值如图 7 所示。

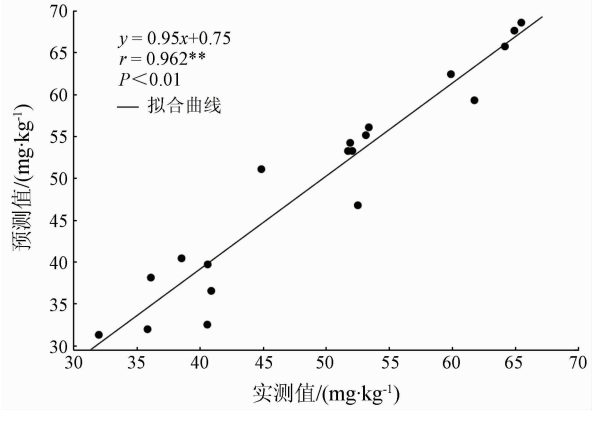


图 7 ICO – ELM 模型验证集实测值与预测值关系

Fig. 7 Relationship between measured and predicted values of validation set based on ICO – ELM model

其相关系数 *r* 通过了 $P = 0.01$ 水平上的显著性检验,表现出很强的相关性。说明 ICO – ELM 具有出色的预测能力,能有效反演土壤重金属 Pb 的含量,为土壤其他重金属成分乃至其他理化属性成分含量反演提供了一种快速高效的耦合模型参考。

4 结论

以龙海市为研究区,对数据进行前期处理后引入基于模型集群分析策略的波长选择算法 CARS,VISSA,IVSO 和 ICO 分别对光谱变量进行优选,并分别建立 PLSR,SVM 及 ELM 这 3 种回归模型对龙海市土壤重金属 Pb 含量进行预测,主要结论如下:

1)对比小波变换、SG 平滑、高斯滤波和多元散射校正 4 种光谱预处理方法,基于小波变换第七层重构后光谱数据建立的模型表现最佳。其验证集 $R^2 = 0.736$,$RMSE = 5.426$,$RPD = 1.976$,$RPIQ = 2.560$。

2)CARS,VISSA,IVSO 和 ICO 各自耦合 PLSR 的回归结果表明,相比于全波段建模,剔除无信息与干扰变量后,模型的预测精度、可解释性、计算复杂度等都得到了明显改善,证明基于模型集群分析策略的波长选择算法能有效应用于土壤重金属 Pb 含量的回归预测之中。总体表现排序为:ICO > IVSO > VISSA > CARS。

3)相较于 SVM 及 PLSR 这 2 种回归模型,各波段选择算法耦合 ELM 后,预测精度得到了较高的提升。其中 ICO – ELM 模型表现最佳,其验证集 $R^2 = 0.863$,$RMSE = 3.953$,$RPD = 2.712$,$RPIQ = 3.514$,进一步提升了模型的泛化性能。

考虑到不同类型波长选择算法对变量筛选的原理不同,本文筛选变量时仅用到单一算法,且 ICO 挑选出的特征波长数量是否可以继续精简有待考证。为了充分发挥算法之间的互补性,在进一步的研究中,可以将例如智能优化算法、连续投影算法等方法与本文方法进行双重耦合甚至三重耦合,以此探讨最佳的算法组合。

参考文献(References):

[1] 宋 伟,陈百明,刘 琳. 中国耕地土壤重金属污染概况[J]. 水土保持研究,2013,20(2):293 – 298.
Song W,Chen B M,Liu L. Soil heavy metal pollution of cultivated land in China[J]. Research of Soil and Water Conservation,2013,20(2):293 – 298.

[2] 贺军亮,韩超山,韦 锐,等. 基于偏最小二乘的土壤重金属镉

间接反演模型[J]. 国土资源遥感,2019,31(4):96 – 103. doi:10.6046/gtzyyg.2019.04.13.

He J L,Han C S,Wei R,et al. Research on indirect hyperspectral estimating model of heavy metal Cd based on partial least squares regression[J]. Remote Sensing for Land and Resources,2019,31(4):96 – 103. doi:10.6046/gtzyyg.2019.04.13.

[3] 高凯旋,焦海明,王新闯. 融合影像纹理、光谱与地形特征的森林冠顶高反演模型[J]. 国土资源遥感,2020,32(3):63 – 70. doi:10.6046/gtzyyg.2020.03.09.

Gao K X,Jiao H M,Wang X C. Inversion model of forest canopy height based on image texture,spectral and topographic features [J]. Remote Sensing for Land and Resources,2020,32(3):63 – 70. doi:10.6046/gtzyyg.2020.03.09.

[4] 段宏伟,朱荣光,许卫东,等. 基于 GA 和 CARS 的真空包装冷却羊肉细菌菌落总数高光谱检测[J]. 光谱学与光谱分析,2017,37(3):847 – 852.

Duan H W,Zhu R G,Xu W D,et al. Hyperspectral imaging detection of total viable count from vacuum packing cooling mutton based on GA and CARS algorithms[J]. Spectroscopy and Spectral Analysis,2017,37(3):847 – 852.

[5] Le B T,肖　冬,毛亚纯,等. 可见、近红外光谱和深度学习 CNN – ELM 算法的煤炭分类[J]. 光谱学与光谱分析,2018,38(7):2107 – 2112.

Le B T,Xiao D,Mao Y C,et al. Coal classification based on visible,near – infrared spectroscopy and CNN – ELM algorithm[J]. Spectroscopy and Spectral Analysis,2018,38(7):2107 – 2112.

[6] 汪六三,鲁翠萍,王儒敬,等. 土壤碱解氮含量可见/近红外光谱预测模型优化[J]. 发光学报,2018,39(7):1016 – 1023.

Wang L S,Lu C P,Wang R J,et al. Optimization for Vis/NIRS prediction model of soil available nitrogen content[J]. Chinese Journal of Luminescence,2018,39(7):1016 – 1023.

[7] 吴　倩,姜琦刚,史鹏飞,等. 基于高光谱的土壤碳酸钙含量估算模型研究[J]. 国土资源遥感,2021,33(1):138 – 144. doi:10.6046/gtzyyg.2020095.

Wu Q,Jiang Q G,Shi P F,et al. Estimation of soil calcium carbonate content based on hyperspectral data[J]. Remote Sensing for Land and Resources,2021,33(1):138 – 144. doi:10.6046/gtzyyg.2020095.

[8] 李　跑,周　骏,蒋立文,等. 窗口竞争性自适应重加权采样策略的近红外特征变量选择方法[J]. 光谱学与光谱分析,2019,39(5):1428 – 1432.

Li P,Zhou J,Jiang L W,et al. A variable selection approach of near infrared spectra based on window competitive adaptive reweighted sampling strategy[J]. Spectroscopy and Spectral Analysis,2019,39(5):1428 – 1432.

[9] Li H,Liang Y,Xu Q,et al. Model population analysis for variable selection[J]. Journal of Chemometrics,2010,24(7 – 8):418 – 423.

[10] 于　雷,章　涛,朱亚星,等. 基于 IRIV 算法优选大豆叶片高光谱特征波长变量估测 SPAD 值[J]. 农业工程学报,2018,34(16):148 – 154.

Yu L,Zhang T,Zhu Y X,et al. Determination of soybean leaf SPAD value using characteristic wavelength variables preferably selected by IRIV algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering,2018,34(16):148 – 154.

[11] 刘贵珊,张　翀,樊奈昀,等. IVISSA 算法冷鲜滩羊肉嫩度的高光谱模型优化[J]. 光谱学与光谱分析,2020,40(8):2558 – 2563.

Liu G S,Zhang C,Fan N Y,et al. Hyperspectral model optimization for tenderness of chilled tan – sheep mutton based on IVISSA[J]. Spectroscopy and Spectral Analysis,2020,40(8):2558 – 2563.

[12] 云永欢,邓百川,梁逸曾. 化学建模与模型集群分析[J]. 分析化学,2015,43(11):1638 – 1647.

Yun Y H,Deng B C,Liang Y Z. Progress of chemical modeling and model population analysis[J]. Chinese Journal of Analytical Chemistry,2015,43(11):1638 – 1647.

[13] Deng B C,Yun Y H,Liang Y Z,et al. A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling[J]. Analyst,2014,139(19):4836 – 4845.

[14] Wang W,Yun Y,Deng B,et al. Iteratively variable subset optimization for multivariate calibration[J]. RSC Advances,2015,5(116):95771 – 95780.

[15] Song X Z,Yue H,Hong Y,et al. A novel algorithm for spectral interval combination optimization[J]. Analytica Chimica Acta,2016,948:19 – 29.

[16] 孙　凯,孙彬彬,周国华,等. 福建龙海土壤重金属含量特征及影响因素研究[J]. 现代地质,2018,32(6):1302 – 1310.

Sun K,Sun B B,Zhou G H,et al. Study on concentration characteristics and influencing factors of heavy metals in soils in Longhai,Fujian Province[J]. Geoscience,2018,32(6):1302 – 1310.

[17] 刘智超,蔡文生,邵学广. 蒙特卡洛交叉验证用于近红外光谱奇异样本的识别[J]. 中国科学(B 辑:化学),2008,38(4):316 – 323.

Liu Z C,Cai W S,Shao X G. Identification of NIR outlier samples by MCCV[J]. Science in China Series B:Chemistry,2008,38(4):316 – 323.

[18] Huang G,Huang G,Song S,et al. Trends in extreme learning machines:A review[J]. Neural Networks,2015,61:32 – 48.

[19] Tan K,Wang H,Zhang Q,et al. An improved estimation model for soil heavy metal(loid) concentration retrieval in mining areas using reflectance spectroscopy[J]. Journal of Soils and Sediments,2018,18(5):2008 – 2022.

[20] 宋相中,唐　果,张录达,等. 近红外光谱分析中的变量选择算法研究进展[J]. 光谱学与光谱分析,2017,37(4):1048 – 1052.

Song X Z,Tang G,Zhang L D,et al. Research advance of variable selection algorithms in near infrared spectroscopy analysis [J]. Spectroscopy and Spectral Analysis,2017,37(4):1048 – 1052.

Prediction of lead content in soil based on model population analysis coupled with ELM algorithm

XIAO Yehui[1], SONG Nidi[2], MENG Panpan[2], WANG Peijun[2], FAN Shenglong[2]

(1. *College of Resource and Environment, Fujian Agriculture and Forestry University, Fuzhou* 350007, *China*;
2. *College of Public Management, Fujian Agriculture and Forestry University, Fuzhou* 350007, *China*)

Abstract: This paper aims to explore the optimal inversion model of regional heavy metal content in soil. With Longhai City taken as the study area, this study preprocessed the original spectral data of soil using the methods of Savizky Golay (SG), wavelet transform (WT), gaussian filter (GF), and multiple scatter correction (MSC) individually, then eliminated the interference and wavelength bearing no information using the wavelength selection algorithms developed based on model population analysis (MPA), including the competitive adaptive reweighted sampling (CARS), variable iterative space shrinkage approach (VISSA), iteratively variable subset optimization (IVSO), and interval combination optimization (ICO), and finally predicted the lead content in soil using the linear partial least squares regression (PLSR) model, nonlinear support vector machine (SVM) model, and extreme learning machine (ELM) based on neural network. The results are as follows. ① Among the inversion models of lead content in soil established using various preprocessing methods, the model built based on reconstructed spectral data of level 7th by wavelet transform had the most optimal prediction accuracy, with R^2 = 0.736, $RMSE$ = 5.426, RPD = 1.976, and $RPIQ$ = 2.560. ② The CARS, VISSA, IVSO, and ICO algorithms developed based on MPA significantly improved the performance of model interpretation and generalization and improved modeling efficiency. ③ In terms of overall prediction results, the three regression models were in the order of ELM > PLSR > SVM. Among them, the ICO – ELM had the highest prediction accuracy, with R^2 = 0.863, $RMSE$ = 3.953, RPD = 2.712, and $RPIQ$ = 3.514. Therefore, the optimal model established in this study can provide a new theoretical reference for the rapid monitoring of regional land quality and ecological indicators.

Keywords: model population analysis; wavelet transform; interval combination optimization; extreme learning machine

(责任编辑: 陈 理)