第42卷，第5期
2022年5月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 5, pp1535-1540
May, 2022

X射线荧光光谱结合CARS变量筛选选择方法用于土壤中铅砷含量的测定

江晓宇[1,2]，李福生[2*]，王清亚[1,2]，罗 杰[3]，郝 军[1,2]，徐木强[1,2]

1. 东华理工大学核技术应用教育部工程研究中心，江西 南昌 330013
2. 东华理工大学核资源与环境国家重点实验室，江西 南昌 330013
3. 长江大学，湖北 武汉 430000

摘 要 X射线荧光光谱分析作为一种以化学计量学为基础的定量分析技术，所建立模型优劣对结果的预测准确性显得十分重要。竞争性自适应重加权算法(CARS)采用自适应重加权采样技术，利用交互验证选出互验证均方根误差(RMSECV)值最低原则，寻出最优变量组合。为了进一步提高PLS模型的解释和预测能力，将竞争性自适应重加权算法(CARS)与X射线荧光光谱分析技术相结合，对土壤中重金属元素铅和砷进行特征波长变量筛选后建立偏最小二乘(PLS)模型。首先，利用CARS算法对铅含量密切相关的波长变量进行筛选，当采样次数为26次时，筛选出60个有效波长点；对砷含量密切相关的波长变量进行筛选，当采样次数为34次时，筛选出19个有效波长点；然后对优选出的波长点利用PLS方法分别建立土壤中铅和砷含量定量分析模型，并与经连续投影算法(SPA)及蒙特卡罗无信息变量消除(MC-UVE)方法波长变量筛选后所建立的PLS模型进行比较。结果显示：铅的CARS-PLS模型的预测集决定系数(R^2)、交互验证均方根误差(RMSECV)、预测均方根误差(RMSEP)和相对预测误差(RPD)分别为0.995 5，2.598 6，3.228和9.401 1，砷的CARS-PLS模型的预测集R^2，RMSECV，RMSEP和RPD分别为0.989 9，3.013 2，2.737 1和8.211 6；两元素的CARS-PLS模型性能均优于全波段PLS，SPA-PLS和MC-UVE-PLS模型。基于CARS-PLS的算法可以有效筛选出X射线荧光光谱特征波长点，在简化了建模复杂程度的同时，提高了模型的准确性和稳健性。

关键词 竞争性自适应重加权算法(CARS)；偏最小二乘(PLS)；波长变量选择；X射线荧光光谱

中图分类号：TH741.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)05-1535-06

引 言

能量色散X射线荧光(EDXRF)光谱仪因其在多元素检测中具有无损、快速的特点，相比传统检测方法，在土壤重金属分析中具有先天的优势。另外，EDXRF因其较小的体积、较轻的重量、更快的分析速度以及较高的准确度，广泛应用于野外现场分析。近几年来，EDXRF越来越受环保领域的欢迎，成为土壤修复行业和环境监管部门的首选仪器。然而，X射线荧光光谱易受噪声、变量维度高和多重共线性等问题的干扰，特别是在测土壤样品时，因其样品来源广泛，基体成分复杂，采用偏最小二乘(PLS)直接建模的话会导致模型复杂，并且降低了模型的预测能力和鲁棒性。因此，如何选择合适的变量显得尤为重要。近年来，科学技术的飞速发展，IT和计算机技术快速应用，特征变量筛选方法被大量提出，如基于统计学方面的变量选择方法[1]、基于单一指标的变量选择方法[2-3]以及群体智能优化算法[4-5]等。

竞争性自适应重加权算法(competitive adaptive reweighted algorithm，CARS)是利用蒙特卡罗(MC)的优势进行采样和PLS回归系数为指标的一种特征波长变量选择方法[6]。其核心是利用自适应重加权采样(ARS)技术，然后在构建的模型中只保留权重显著(回归系数绝对值大)的波长点，最后按照均方根误差值最小的原则选择最优组合子集变量。此外，在对大多文献调研过程中发现，很少有对土壤样品X射线荧光光谱波长变量进行筛选。但X射线荧光光谱往往也存在维度过高，变量数大于建模样本数问题，建立的

收稿日期：2021-03-17，**修订日期**：2021-06-06

基金项目：2019年江西省“双千计划”引进项目(2120800003)，国家自然科学基金项目(21876014)资助

作者简介：江晓宇，1988年生，东华理工大学核技术应用教育部工程研究中心博士研究生 e-mail：jjxxyy1988@126.com

*通讯作者 e-mail：lifusheng@ecit.cn

模型容易过拟合，模型稳定性变差。

先利用能量色散 X 射线荧光光谱仪对土壤中的铅和砷进行分析获取原始光谱信息，然后利用 CARS 算法先对所获取的原始光谱进行波长变量选择，最后利用 PLS 分别建立土壤中铅、砷的定量分析模型。为了评估建模的有效性，一般采用预测集决定系数(determination coefficient，R^2)、模型交互验证均方根误差(root mean square error of cross validation，RMSECV)、模型预测均方根误差(root mean square error of prediction，RMSEP)和模型相对预测误差(relative prediction deviation，RPD)等为模型评价指标，并与全波段、SPA 和 MC-UVE 等变量选择算法所建立的定量分析模型进行比较。

1 实验部分

1.1 材料与仪器

主要仪器：TS-XH4000 型便携式 X 射线荧光光谱仪，浙江泰克松德能源科技有限公司；SDD 探测器，能量分辨率为 125 eV，美国 Amptek 公司；球磨仪，江苏宜兴丁蜀浩强机械设备有限公司；样品杯(聚乙烯)，尺寸为 Φ3 cm×1 cm，单开口，带固定麦拉膜的颈圈；麦拉膜，厚度为 3.6 μm，宽 7.6 cm，美国 Chemplex 公司。

1.2 土壤样品采集

本试验中，共计样品 139 个，其中野外采集土壤样品 80 个(江西鄱阳湖地区)，另外 59 个为国家土壤标准样品(GSD 和 GSS 系列)。样品采集和制备方法必须严格按照《土壤环境质量标准》(GB15618—2018)的技术规范执行。将采集到的所有土样铺开自然风干，去除土样中明显的沙子、草屑等杂物，使用四分法取其 2 份，1 份用于实验分析，1 份留作备用。将国家土壤标准样品和实验分析的土壤样品均匀填入玛瑙钵体中，用球磨机研磨 5 min，然后过 200 目筛子。将处理后的土壤样品使用 TS-XH4000 便携式 XRF 分析仪在管压 35 keV、电流 40 μA 和时间 90 s 下，采集土壤 X 射线荧光光谱原始数据，每个样本测量 3 次，移动不同位置 3 次，最后取平均值作为光谱数据，共获取样品在 0～45 keV 范围内共 2 048 个通道数的光谱信息。

1.3 竞争性自适应重加权算法(CARS)

1.3.1 CARS 算法原理

CARS 算法是模拟生物进化论中的“适者生存”的法则，每次通过 ARS 技术和 PLS 回归系数的绝对值对变量进行筛选，保留 PLS 回归系数中的绝对值大的点，去掉绝对值较小的点，得到一系列最优子集[7]。然后使用交叉验证(CV)方法选择模型 RMSECV 最小值的子集，并最终将子集确定为与测量元素相关的最佳波长组合。

1.3.2 CARS 算法步骤

假设 $\boldsymbol{Y}$ 表示为 $m\times1$ 样本目标属性矩阵，$\boldsymbol{X}$ 为 $m\times n$ 样本光谱矩阵，其中 m 为样本数，n 为变量数，α 表示组合系数；$\boldsymbol{T}$ 为 $\boldsymbol{X}$ 与 α 的线性组合，是 $\boldsymbol{X}$ 的分矩阵；θ 是 $\boldsymbol{Y}$ 和 $\boldsymbol{T}$ 所建 PLS 模型的回归系数向量；其中，β 和 ε 分别表示为 n 维的回归系数向量和样本预测残差。假设式(1)和式(2)成立。

$$\boldsymbol{T}=\alpha\boldsymbol{X} \tag{1}$$

$$\boldsymbol{Y}=\theta\boldsymbol{T}+\varepsilon=\theta\alpha\boldsymbol{X}+\varepsilon=\beta\boldsymbol{T}+\varepsilon \tag{2}$$

式(2)中，回归系数向量 $\beta=\alpha\theta=[\beta_1,\beta_2,\cdots,\beta_n]$，第 i 个波长变量对 $\boldsymbol{Y}$ 贡献，那么所有波长对 $\boldsymbol{Y}$ 的总贡献用第 i 个元素绝对值 $|\beta_i|$ $(1\leqslant i\leqslant p)$ 来表示，$f=\sum_{i=1}^{p}|b_i|$。使用权重 w_i 作为变量优选指标来评估每个波长的重要性，其中 w_i 为 $|\beta_i|$ 对总贡献所占的比例，如果 w_i 值越大，则该变量重要性越明显，如式(3)所示

$$w_i=|\beta_i|\,i/f \tag{3}$$

式(3)中，每计算一次 w_i 的过程实际上就是波长变量重要性评估的过程。将每次计算的 $|\beta_i|$ 值较大波长变量保留，然后采用 ARS 技术从中重新组合新的变量，在此基础上利用 PLS 建模，计算其 RMSECV 值。其中，采样次数设为 N，重复 N 次，直到采样结束，我们将得到最优变量子集集合，即一系列 RMSECV 值最小的变量子集。

最后，CARS，PLS，SPA 和 MC-UVE 的算法编写通过 Matlab R2016b 实现，而图表绘制由 Origin9.0 软件完成。

2 结果与讨论

2.1 光谱预处理

X 射线荧光光谱为特征谱，其中铅元素的 Lα 和 Lβ 特征峰分别在 10.549 和 12.61 keV 附近；砷元素的 Kα 和 Kβ 特征峰在 10.532 和 11.729 keV 附近。X 射线荧光光谱采集会产生大量的高频随机噪声、基线漂移和散射等噪声信息干扰，使 X 射线荧光光谱与元素含量之间的相关性变差，导致所建模型的准确性和稳定性会受到影响。为消除噪声和基线的影响，尽可能完整保留土壤样品中原始 X 射线荧光光谱的特征峰，去噪选用小波变换(sym4 小波基)，而校正基线采用

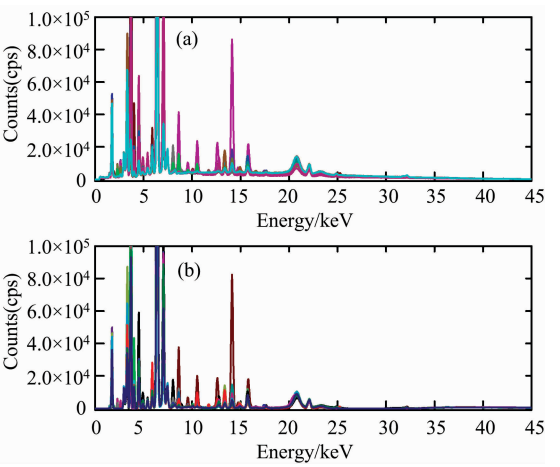

图 1 土壤样品光谱的噪声和基线校正结果

(a)：139 个土壤样品的原始光谱；

(b)：139 个土壤样品的去噪声和基线校正光谱

Fig. 1 Noise and baseline correction results for the spectra of soil samples

(a)：Raw spectra of 139 soil samples；

(b)：Denoising and baseline corrected spectra of 139 soil samples

适应迭代重加权惩罚最小二乘(airPLS)法[8]，处理结果如图 1 所示。最后，选择处理后的 X 射线荧光光谱数据进行特征变量选择。

2.2　校正集与验证集的划分

采用 Kennard-Stone(K-S)算法[9]对 139 个土壤样本进行校正集与验证集的划分。K-S 算法的原理：(1)计算样本两两之间的距离，选择样本间距离最大的两个作为选中的集合样本，其余为未选中的集合样本；(2) 对于剩余样本，分别计算其与选中的两个样本之间的距离；(3)然后选择最短距离与所选样本之间相对最长的距离对应的样本，作为所选样本集；(4)重复步骤(3)，直到所选样本数等于之前确定的数量，例如 10 个或 20 个。本实验选取的样本集为校正集，约 70%的铅和砷样品转入校正集，共 97 个样品，剩余 42 个样本归为预测集。表 1 列出了被测土壤中铅和砷实测值的变化范围和平均值(Mean)等统计量。K-S 算法也是通过 Matlab R2016b 软件完成。

表 1　土壤铅和砷含量实测值的统计结果

Table 1　Descriptive statistics of measured soil Lead and Arsenic content

元素	样本集	样本数	范围/(mg·kg^{-1})	均值/(mg·kg^{-1})
Pb	校正集	97	10.2～2 690.2	400.9
	预测集	42	13.4～636.3	359.6
As	校正集	97	1.7～706.9	274.4
	预测集	42	4.4～416.7	235.8

2.3　特征波长选择

2.3.1　土壤中铅特征波长选择

先以铅 X 射线荧光光谱全部的 2 048 个波数点作为选择对象，采用 CARS 算法筛选样本光谱中与铅相关的光谱波长变量，筛选结果如图 2 所示。从图 2(a)中，我们看到选择的波长变量的数量随着采样次数的增加而减少，趋势是先快后缓，说明波长变量先经历了一个粗略的选择过程后再进行精选过程；图 2(b)中，随着采样次数的增加，RMSECV 值先减后增，即所选波长变量的个数逐渐减少，RMSECV 值也在减小，说明与铅无关的冗余波长变量在 CARS 变量筛选时优选剔除掉，而后 RMSECV 值上升，说明是剔除了与铅相关的波长变量引起的；图 2(c)中红色“*”处的 MC 采样次数为 26，此时 RMSECV 值最小，经过 CARS 筛选后，共选择了 60 个波长变量，且所选择的波长变量组合最优。

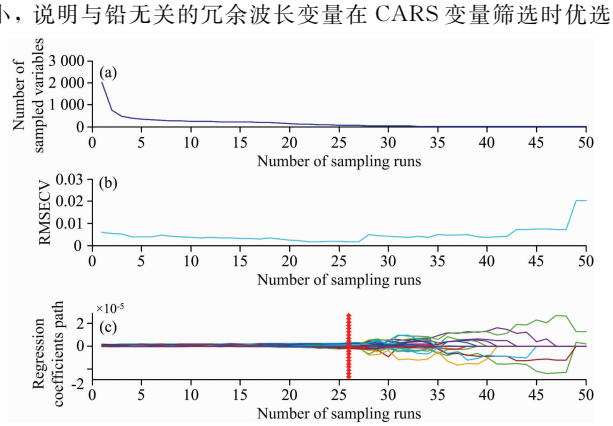

图 2　土壤中铅的 CARS 变量筛选结果

Fig. 2　Plots of CARS variable selection for Lead in soil

2.3.2　土壤中砷特征波长选择

以砷的 X 射线荧光光谱全部的 2 048 个波数点作为选择对象，采用 CARS 算法筛选样本光谱中与砷相关的光谱波长变量，筛选结果如图 3 所示。类似于上述铅的情况，从图 3(a)中我们可以看到随着采样数增加，被优选波长变量的数量迅速减少。在图 3(b)中，在 1～34 次采样期间，RMSECV 值不断减小，表明变量筛选时去除了与砷含量相关的变量，但在 34 个样品后，RMSECV 值再次开始上升，这表明与砷含量相关的重要变量被去除。在采样为 34 次时，即图 3(c)中“*”的位置，出现 RMSECV 值最小，共选择了 19 个波长变量，所对应的光谱变量子集最优。

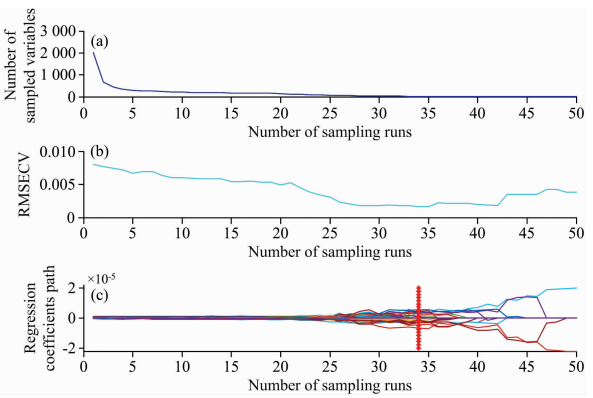

图 3　土壤中砷的 CARS 变量筛选结果

Fig. 3　Plots of CARS variable selection for Arsenic in soil

CARS 模型 RMSECV 值最小时，铅和砷对应的最优采样次数和最优变量子集中包含的变量个数如表 2 所示。

表 2　土样中铅和砷在 RMSECV 值最小时对应的采样次数及最优变量子集包含的变量个数

Table 2　Sampling frequency and variable number in optimal variables subset of Pb and As when RMSECV value is lowest in soil samples

元素	采样次数	变量个数
Pb	26	60
As	34	19

2.4　PLS 模型的建立与验证

提出采用 CARS 算法对原始光谱进行波长信息变量筛选，并与蒙特卡罗无信息变量消除(MC-UVE)和经连续投影算法(SPA)方法进行比较，然后分别采用偏最小二乘(PLS)方法建立土壤中铅和砷含量的定量检测模型，评价所建模型的建模效果。CARS 变量筛选方法，基于蒙特卡罗(MC)交叉验证确定成样次数设置为 50 次，可满足其可用的最大因子数。针对 SPA 变量选择方法，其利用向量投影分析原理，能有效地消除波长之间共线性问题，分别设置好最小最大波长

数，其最佳波长组合通过交叉验证建模实现，然后找到具有最小冗余信息的变量组，最终提高模型精度。MC-UVE 变量选择方法是基于 PLS 回归系数 b 的算法，重复 N 次，得到 N 个回归系数组成的矩阵，大大减少了最终 PLS 模型中所包含的变量数量，模型的复杂度和稳定性得到改善。其中 SPA 和 MC-UVE 变量选择方法的具体原理和步骤见文献[10-12]。

采用决定系数(R^2)、交互验证均方根误差(RMSECV)、预测均方根误差(RMSEP)和模型相对预测误差(RPD)等 4 个参数来评价 PLS 模型性能。其中，R^2 值越接近于 1，模型的拟合度和稳定性越好；RMSECV 和 RMSEP 值越小，模型预测能力越强；RPD 值等于样本标准偏差与均方根误差的比值。如果 RPD≥3，认为所建立的模型预测效果良好，具有良好应用价值；如果 2.25≤RPD<3，则认为所建立的模型预测效果较好，具有较好实际应用价值；如果 1.75≤RPD<2.25，则认为模型可用，模型对样本能进行粗略评估；如果 RPD<1.75，模型预测效果差，无法预测样本。

2.4.1 土壤中铅的 PLS 模型的建立与验证

经 CARS，SPA 及 MC-UVE 变量筛选后，采用 PLS 方法建立土壤中铅含量的定量检测模型，建模结果见表 3。从表 3 可以看出，经过 CARS 筛选后，CARS-PLS 模型铅的波长变量数从 2 048 减少到 60 个，模型最优，所得建模集的 R^2，RMSECV，RMSEP 和 RPD 分别为 0.997 3，2.610 1，3.322 1 和 9.351 8，预测集的 R^2，RMSECV，RMSEP 和 RPD 分别为 0.995 5，2.598 6，3.228 和 9.401 1；与 CARS-PLS 模型相比，虽然 SPA-PLS 和 MC-UVE-PLS 模型建模的波长变量更少，但建模集和预测集的 R^2，RMSECV，RMSEP 和 RPD 均劣于 CARS-PLS 模型。另外，从表 3 还发现，与全波段 PLS 模型相比，SPA-PLS 模型的预测集 R^2，RMSECV，RMSEP 和 RPD 分别 0.980 5，3.549 5，5.344 5 和 8.611 4，劣于全波段 PLS 模型，模型的稳定性不如 PLS，MC-UVE-PLS 和 CARS-PLS 模型。

表 3 土样中铅定量检测的 PLS 建模结果

Table 3 PLS modeling results of quantitative determination of Lead in soil samples

方法	变量个数	校正集				预测集			
		R^2	RMSECV	RMSEP	RPD	R^2	RMSECV	RMSEP	RPD
PLS	2 048	0.987 8	3.621 2	5.381 2	8.123 2	0.983 5	3.533 2	5.290 8	8.690 8
CARS-PLS	60	0.997 3	2.610 1	3.322 1	9.351 8	0.995 5	2.598 6	3.228 0	9.401 1
SPA-PLS	27	0.989 8	3.643 9	5.391 8	8.177 4	0.980 5	3.549 5	5.344 5	8.611 4
MC-UVE-PLS	49	0.995 5	3.153 1	3.553 2	8.997 8	0.995 7	3.058 5	3.685 0	9.036 6

2.4.2 土壤中砷的 PLS 模型的建立与验证

经 CARS，SPA 及 MC-UVE 变量筛选后，采用 PLS 方法建立土壤中砷含量的定量检测模型，建模结果见表 4。从表 4 可以看出，砷 CARS-PLS 模型的波长变量数由 2 048 个减少到 19 个，与全波段 PLS，SPA-PLS 和 MC-UVE-PLS 模型相比，砷的 CARS-PLS 模型建模集和预测集的 R^2，RMSECV，RMSEP 和 RPD 值均最优，所建模型效果最好。与其他三个模型相比，虽然 SPA-PLS 模型的波长变量最少，但建模集和预测集的 R^2，RMSECV，RMSEP 和 RPD 均劣于 CARS-PLS 和 MC-UVE-PLS 模型，仅优于全波段 PLS 模型。

从以上结果可以看出，CARS-PLS 模型定量检测土壤中的铅和砷要优于全波段 PLS，SPA-PLS 及 MC-UVE-PLS 模型，表明 CARS 方法在 X 射线荧光光谱的波长变量选择方面具有较明显优势，可以筛选出有用的波长信息变量并去除多余的波长变量，来提高模型的准确性和稳定性。

表 4 土样中砷定量检测的 PLS 建模结果

Table 4 PLS modeling results of quantitative determination of Arsenic in soil samples

方法	变量个数	校正集				预测集			
		R^2	RMSECV	RMSEP	RPD	R^2	RMSECV	RMSEP	RPD
PLS	2 048	0.975 9	4.623 7	4.784 9	7.623 9	0.975 1	4.520 3	4.564 3	7.520 1
CARS-PLS	19	0.990 6	2.923 7	2.933 1	8.156 6	0.989 9	3.013 2	2.737 1	8.211 6
SPA-PLS	9	0.984 3	4.317 1	4.358 4	7.643 3	0.988 1	4.334 5	4.288 5	7.593 0
MC-UVE-PLS	30	0.985 0	3.329 2	3.613 5	7.678 6	0.989 5	3.310 2	3.419 0	7.896 5

2.5 PLS 模型预测

图 4 显示了四种模型的预测值与传统化学方法测定值之间的相关关系。CARS-PLS 模型铅砷预测值与其实验室分析值或标准值最为接近，线性最好。这进一步说明 CARS 算法可以有效筛选波长变量，且用更少的变量建立更好的铅砷定量分析模型。

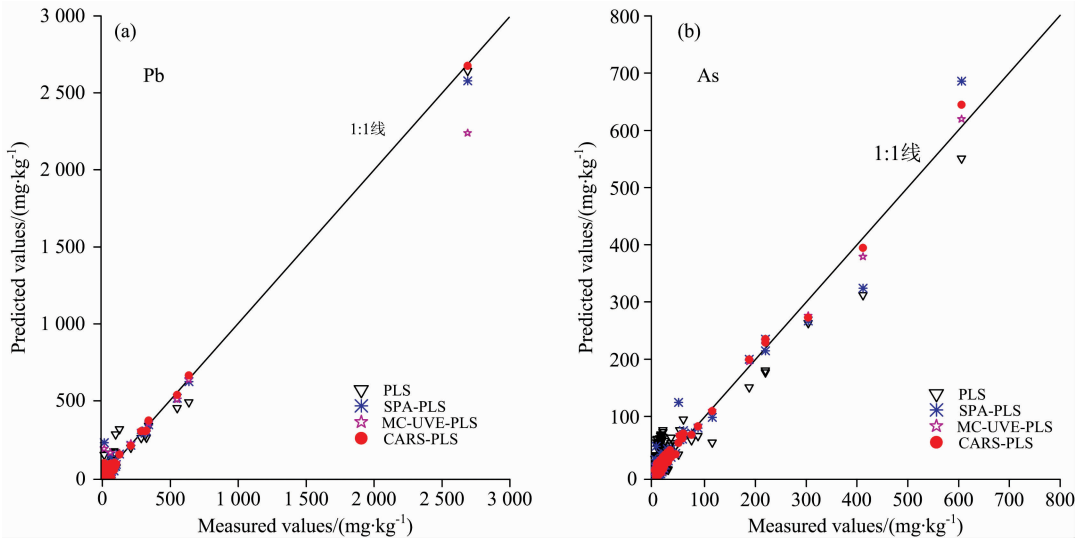

图 4　各模型铅、砷校正集真实值与预测值对比

Fig. 4　Comparison of measured values and predicted values of Lead and Arsenic corrected Set of each model

3　结　论

采用 CARS 波长变量筛选算法，建立了土壤中 X 射线荧光光谱定量分析重金属铅和砷含量检测模型(CARS-PLS)，筛选出具有较高适用性的波长变量子集组合，实现了铅和砷含量的准确预测。具体结论如下：

(1)通过对土壤中铅和砷的 X 射线荧光光谱进行建模，结果表明 CARS 方法是一种有效的波长变量选择方法，在降低模型的维数同时还剔除了多余的干扰信息，使模型的计算效率和稳健性得到提升。

(2)采用 CARS 方法对土壤中铅和砷的波长信息变量进行筛选，分别筛选得到 60 和 19 个波长变量作为预测铅和砷的优选变量集。

(3)与全波段 PLS，SPA-PLS 和 MC-UVE-PLS 模型相比，采用 CARS-PLS 所建模型具有最优的预测精度和预测能力，同时有效减少了波长变量。

由于此次试验采用的土壤样品经过晾干、筛分等物理前处理过程，消除了土壤含水率、粒径等因素对检测结果的影响，所建立的铅砷的定量分析模型在现场的准确性如何是下一步研究的重点。另外，在应对极低浓度元素时会受到一定噪声影响，在做波长变量筛选时，会影响建模的结果，这也是我们下一步需要优化的地方。

References

[1] WEN Zhen-cai, SUN Tong, XU Peng, et al(温珍才，孙　通，许　朋，等). Journal of Jiangsu University · Natural Science Edition(江苏大学学报 · 自然科学版), 2015, 36(6): 673.

[2] REN Shun, ZHANG Xiong, REN Dong, et al(任　顺，张　雄，任　东，等). J. Instrum. Anal.(分析测试学报), 2020, 39(7): 829.

[3] BIN Jun, FAN Wei, ZHOU Ji-heng, et al(宾　俊，范　伟，周冀衡，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(1): 95.

[4] WU Xin-yan, BIAN Xi-hui, YANG Sheng, et al(武新燕，卞希慧，杨　盛，等). J. Instrum. Anal.(分析测试学报), 2020, 39(10): 1288.

[5] YU Lei, ZHU Ya-xing, HONG Yong-sheng, et al(于　雷，朱亚星，洪永胜，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2016, 32(22): 138.

[6] JIANG Wei, FANG Jun-long, WANG Shu-wen, et al(姜　微，房俊龙，王树文，等). Journal of Northeast Agricultural University(东北农业大学学报), 2016, 47(2): 88.

[7] Li H D, Liang Y Z, Xu Q S, et al. Analytica Chimica Acta, 2009, 648: 77.

[8] Liang Y Z, Chen S, Zhang Z M. The Analyst, 2010, 135(5): 1138.

[9] Kennard R W, Stone L A. Technometrics, 1969, 11(1): 137.

[10] Centner V, Massart D L, de Noord O E, et al. Analytical Chemistry, 1996, 68(21): 3851.

[11] Araujo M C U, Saldanha T C B, Galvo R K H. Chemometrics and Intelligent Laboratory Systems, 2001, 57(2): 65.

[12] Galvao R K H, Araujo M C U, Fragoso W D. Chemometrics and Intelligent Laboratory Systems, 2008, 92(1): 83.

Determination of Lead and Arsenic in Soil Samples by X Fluorescence Spectrum Combined With CARS Variables Screening Method

JIANG Xiao-yu[1,2], LI Fu-sheng[2*], WANG Qing-ya[1,2], LUO Jie[3], HAO Jun[1,2], XU Mu-qiang[1,2]

1. Engineering Research Center of Nuclear Technology Application, Ministry of Education, East China University of Technology, Nanchang 330013, China
2. State Key Laboratory of Nuclear Resources and Environment, East China University of Technology, Nanchang 330013, China
3. Yangtze University, Wuhan 430000, China

Abstract As a quantitative analysis technique based on stoichiometry, X-ray fluorescence spectroscopy is very important to the prediction accuracy of the results. The competitive adaptive reweighted algorithm (CARS) adopted adaptive reweighted sampling technology and used interactive verification to select the lowest value square error (RMSECV) by interactive verification to find out the optimal combination of variables. To further improving the interpretation and prediction ability of PLS models, the competitive adaptive reweighted algorithm (CARS) was combined with X-ray fluorescence spectroscopy. A partial least square (PLS) model was established after screening the characteristic wavelength variables of lead and arsenic in the soil. Firstly, the CARS algorithm screened the wavelength variables closely related to lead content. When the sampling times were 26 times, 60 effective wavelength points were selected, and the wavelength variables closely related to arsenic content were screened. When the sampling times were 34 times, 19 effective wavelength points were selected. Then used the PLS method to establish the quantitative analysis model of lead and arsenic content in soil and compared it with the PLS model established by continuous projection algorithm (SPA) and Monte Carlo method. The results showed that the prediction sets Determination Coefficient (R^2), Root Mean Square Error of Cross-Validation (RMSECV), Root Mean Square Error of Prediction (RMSEP) and Relative Prediction Deviation (RPD) of the lead CARS-PLS model were 0.995 5, 2.598 6, 3.228 and 9.401 1, respectively. Moreover, the prediction sets R^2, RMSECV, RMSEP and RPD of arsenic CARS-PLS models were 0.999, 3.013 2, 2.737 1 and 8.211 6, respectively. The CARS-PLS model performance of the two elements is better than that of full-band PLS, SPA-PLS and MC-UVE-PLS model. The CARS-PLS algorithm based on the X fluorescence spectrum can effectively screen the characteristic wavelength, simplify the complexity of modeling, and improve the accuracy and robustness of the model.

Keywords Competitive adaptive reweighted algorithm (CARS); Partial least squares (PLS); Wavelength variable selection; X-ray fluorescence spectrum

(Received Mar. 17, 2021; accepted Jun. 6, 2021)

* Corresponding author