基于 NMF-PLS 对含水量影响下土壤重金属含量反演模型研究

吴希军[1]，张 杰[1]，肖春艳[2]，赵学亮[1,3]，李 康[3]，庞丽丽[3]，史彦新[3]，李少华[4]

1. 燕山大学电气工程学院河北省测试计量技术及仪器重点实验室，河北 秦皇岛 066004
2. 河南理工大学资源与环境学院，河南 焦作 454000
3. 中国地质调查局水文地质环境地质调查中心，自然资源部地质环境监测工程技术创新中心，河北 保定 071051
4. 河北先河环保科技股份有限公司，河北 石家庄 050000

摘 要 土壤中过高的重金属含量危害巨大，不仅造成了严重的环境污染，而且通过食物链进入人体对人体健康造成严重威胁，所以对重金属检测十分重要。X射线荧光光谱法具有检测时间短、无损检测、检测成本低等特点被广泛使用，然而检测的光谱数据因受到土壤含水量因素的严重干扰，导致直接对土壤重金属含量估算精度较低。以河北省保定市满城区土样为研究对象，对采集的土样进行除杂、过筛、烘干后加入一定量重金属溶液制备不同含水量不同重金属的样本进行检测。对实验中异常数据计算了马氏距离和进行NJW聚类予以剔除，分析了土壤含水量对土壤重金属光谱的影响，结果表明不同含水量间光谱重复性差，随着土壤含水量的增加光谱强度呈非线性降低。采用Savitzky-Golay卷积平滑去噪法和线性本底法对光谱进行预处理，以解决因环境、仪器本身带来的噪声和基线漂移等问题。然后针对于土壤含水量这一主要干扰，采用非负矩阵分解算法进行处理，并使用峰值信噪比这一评价模型确定端元数目，结果表明当端元数目增至10时峰值信噪比趋于稳定波动很小，非负矩阵分解处理后相同重金属含量不同含水量间光谱重复性好、相似性好，并计算了光谱间的相关系数进一步证明了光谱间的相似性。去除含水量对于光谱干扰后建立了偏最小二乘法预测模型，为了验证预测模型的精度，建立了未去除含水量的偏最小二乘法预测模型和使用外部参数正交化法去除含水量建立的偏最小二乘法预测模型，并使用评价参数决定系数(R^2)、交叉验证均方根误差(RMSECV)、平均绝对误差(MAE)和相对分析误差(RPD)进行评价。验证结果表明，相比较未去除含水量建立的模型，使用非负矩阵分解去除含水量建立的偏最小二乘法模型 R^2 和 RPD 分别提高了0.019 7和1.029 2，RMSECV 和 MAE 分别降低了2.386 3和1.439 6；相对于外部参数正交化法建立的偏最小二乘法模型，R^2 和 RPD 分别提高了0.009 9和0.108 1，RMSECV 和 MAE 分别降低了0.244 7和0.356 6，说明了经过非负矩阵分解去噪后建立的模型有效提高了预测的精度和鲁棒性。非负矩阵分解可以有效消除土壤含水量对光谱的影响，在此基础上建立的偏最小二乘法模型实现了土壤重金属含量的反演，为重金属定量检测提供了一定的技术支持。

关键词 土壤重金属；X射线荧光光谱；非负矩阵分解；偏最小二乘法

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)01-0271-07

引 言

土壤中的重金属会对人体健康产生严重影响，对于重金属的检测显得尤为重要[1]。现有的土壤重金属检测技术有原子吸收光谱法、电感耦合等离子体法、原子荧光光谱法等[2]，这些检测技术存在检测时间长、前期处理复杂、可能造成二次污染和成本较高的缺点，而X射线荧光光谱法具有检测时间短、无污染且对多元素同时检测等特点，更加满足对土壤重金属的检测要求。

在实际的检测中，土壤含水量、温度、土壤粒径大小都会对光谱的采集产生影响，其中土壤含水量会对光谱产生严重的干扰使得直接对土壤重金属进行估计精度较低。现有很多学者对去除含水量影响进行了研究，Ji等[3]使用直接标准

收稿日期：2019-12-17，**修订日期**：2020-04-29

基金项目：国家重点研发计划项目(2016YFC1400601-3，2018YFC1800903)，河北省教育厅高等学校科技计划青年基金项目(QN2018071)，河北省专家出国培训项目资助

作者简介：吴希军，1979年生，燕山大学电气工程学院副教授 e-mail：wuxijun@ysu.edu.cn

化法(direct standardization，DS)对湿土光谱进行校正，使用校正后的光谱预测土壤属性；安晓飞等提出使用水分吸收指数对不同含水量样本进行分类并给与修正系数，来消除含水量的干扰；胡晓艳等[4]提出了水分修正系数法(moisture determination index，MDI)有效降低了含水量的干扰，提高了利用干土土样有机质定量预测湿土有机质的精度；Minasny等[5]使用外部参数正交化法(external parameter orthogonalization，EPO)去除近红外光谱中土壤含水量的影响，对土壤有机碳含量进行校正。这些算法具有易造成过度校正、忽视待估土壤属性浓度与受含水量影响光谱之间关系等不足，且是对于去除含水量对土壤属性预测精度提升的研究，对去除含水量影响提高土壤重金属含量预测精度研究较少。

本研究采用非负矩阵分解算法(nonnegative matrix factorization，NMF)，对原始能谱矩阵进行分段式分解，在保留与土壤重金属有关的信息下，滤除能谱矩阵中因含水量引起的噪声偏移，处理后建立基于偏最小二乘法预测(partial least squares regression，PLSR)的土壤重金属含量反演模型，并对其进行定量评估。

1 实验部分

1.1 仪器

CIT-3000SYB 能量色散 X 射线荧光分析仪，用于光谱检测；瑞绅葆 PrepP-01 100T XRF 大吨位压片机，用于制作土壤样品压片；金坛大地电动搅拌器，用于水土充分搅拌；DHG-9141A 电热恒温干燥箱，用于烘干原土壤中水分；艾泽拉小型磨粉机，用于打磨土壤等。

1.2 试剂

土壤成分分析标准物质 GSS-4(GBW07404)、GSS-5(GBW07405)、GSS-7(GBW07407)、GSS-8(GBW07408)、GSS-15(GBW07429)和 GSS-28(GBW07457)，购自地球物理地球化学勘查研究所；硝酸镉分析纯、硝酸铬分析纯、重金属标准溶液(Pb，Zn，Cr，Cd，As)，购自北京世纪奥科生物技术有限公司等。

1.3 样品制备及测试

土壤样本采样于河北省保定市满城区北庄村附近，采集耕层(0～20 cm)的土壤样本。清除土样中的杂物带到实验室进行烘干、用玛瑙研钵碾压后过 10 目(1 500 μm)的筛子；再将土壤倒入研磨机进行研磨，之后过 200 目(75 μm)的筛子，能过筛子的土壤倒入保鲜袋中备用。称量不同质量分析纯和重金属标准溶液，配制浓度为 100，200，400，600 和 800 mg·kg^{-1}不同含水量的重金属溶液。华北平原的土壤含水量在 15%左右，在土壤中加入配制好的重金属溶液，制作出含水量 10%～25%之间等步长十个含水梯度的土壤样品，将麦拉膜与样品杯嵌套固定，装入处理后的土壤并压紧。

测定前先启动能量色散 X 荧光分析仪预热一段时间，将标准样放入仪器检测台的样品腔中对仪器进行校准，使仪器处于最佳的工作状态；再将分析仪的探测窗对准土壤样品杯进行检测，检测时间为 200 s。

2 结果与讨论

能量色散 X 射线荧光光谱检测法原理是建立重金属浓度与荧光强度之间的关系。本研究以重金属铅为例，基于莫塞来(Moseley)定律和普朗克方程可以通过 Pb 的 X 射线激发能量找到 Pb 的特征峰位置，以特征峰的净峰面积为荧光强度。原始光谱经过平滑去噪、本底扣除、基于 NMF 算法去除含水量影响后，对特征峰求积分得到净峰面积，再使用偏最小二乘法建立重金属含量和净峰面积之间的关系。

2.1 异常样本数据剔除

在实验过程中存在着人为因素例如仪器操作不当、土壤样品搅拌程度不够等因素，导致小部分实验数据不准确，使用这些偏差较大的数据进行分析时会使得分析结果精度低，所以应予以剔除。原始矩阵图是在二维空间的坐标系中以样本点的形式体现数据的差异，计算了马氏距离和使用了 NJW 聚类分析的方式，使原始矩阵图切割成几个子图，使得几个子图间相似度最弱而每个子图里样本数据相似度强。图 1 是相同含水量下的原始数据矩阵经过 NJW 聚类分析结果图。

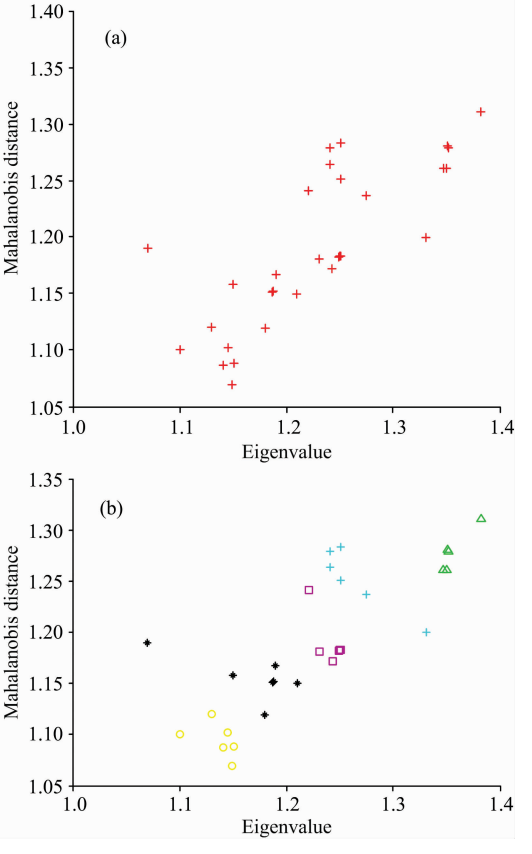

图 1 NJW 聚类分析结果

(a)：原始样本点；(b)：聚类后样本点

Fig. 1 NJW cluster analysis result

(a)：Original sample point；

(b)：Sample point after clustering

从聚类切割图结果中发现了 600 mg·kg^{-1}中有一个数据点被划分到 400 mg·kg^{-1}中不具有重复性，予以剔除。

2.2　土壤含水量对测定结果的影响

土壤中的水分会影响所制土壤样品的均一性，而且水分会对源初级 X 射线和特征 X 射线产生吸收作用，加大 X 射线的散射效应。故土壤中的水分的存在，降低了仪器对土壤中重金属的特征峰强度，从而增加了仪器的检测误差。

每一含水量的测定光谱都是由相同重金属含量同一含水量的多次测试结果进行主成分分析(principal component analysis，PCA)加权处理得到的，可以准确地反映这一含水量的真实光谱。图 2 表示在重金属含量 800 mg·kg^{-1}下重金属 Pb 不同含水量的测定光谱，清楚显示了水分对测试结果的影响，在重金属特征吸收峰附近不同含水量之间的光谱相似性差、重叠程度低，随着土壤含水量的增加检测精度明显下降，重金属 Pb 的相对偏差从 7.41%升高到 27.78%。

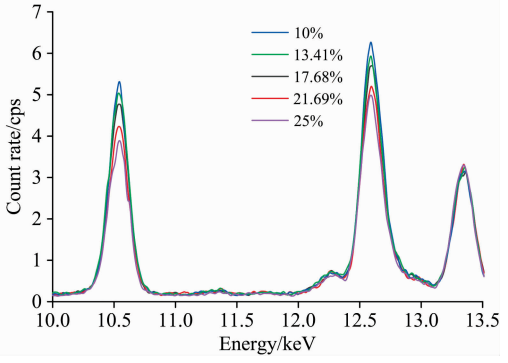

图 2　不同含水量下浓度 800 mg·kg^{-1}重金属 Pb 测定光谱

Fig. 2　Spectral determination spectra of 800 mg·kg^{-1} heavy metal Pb under different water contents

2.3　测定光谱预处理

在对样本进行测定获取光谱的过程中由于外界环境、测定条件的差异会导致原始光谱中存在噪声信号和基线漂移问题。光谱预处理的方法比较多，针对产生的噪声信号，本研究使用了 5 点 2 次 Savitzky-Golay 卷积平滑去噪法，能够非常有效地提高谱图的信噪比；针对仪器本身带来的影响本研究采用了线性本底法扣除本底。

图 3 是原始光谱经过预处理的结果图，(a)表示实际重金属含量为 400 和 600 mg·kg^{-1}光谱经过平滑后的效果图，图中可以看出光谱区间的噪声明显降低并且光谱的形状和宽度没有发生变化，说明这种方法可以去除噪声，提高光谱的信噪比；(b)表示使用线性本底法进行本底扣除，确定特征峰左右边界计算本底面积，扣除本底面积。

2.4　基于 NMF 算法土样含水量因素的去除

2.4.1　NMF 算法

预处理后光谱的主要干扰是土壤含水量这一因素，常见的去除含水量影响算法有直接标准化法、外部参数正交化法，直接标准化法是对全谱端进行整体处理易过拟合，外部参数正交化法忽视了变量前后之间的相互影响，本研究采用非负矩阵分解算法用于去除含水量的影响。

非负矩阵分解是由 Lee 和 Seung 独立提出后，在图像处理、人脸识别中得到广泛的使用，较少应用于 X 射线光谱分解和融合中。NMF 算法是把一个非负矩阵分解成两个非负矩阵因子的乘积，是一种有效的光谱分解的方法。

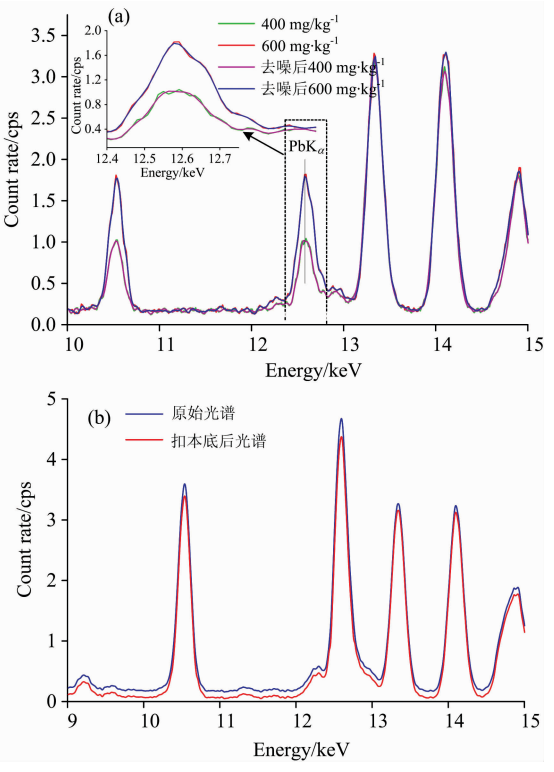

图 3　光谱预处理结果

(a)：S-G 平滑去噪；(b)：线性本底扣除本底

Fig. 3　Spectral preprocessing result

(a)：S-G smooth denoising;

(b)：Linear background minus background

理论上在相同重金属浓度下光谱强度应是几乎一致的，由于不同含水量的影响，导致光谱强度存在明显的差异。光谱数据的灰度值为非负的，所以可以建立一个初始非负矩阵用于光谱分解。将处理后的相同重金属浓度不同含水量 X 射线荧光矩阵看作 n 个样本 m 个能量段的非负数矩阵 $\boldsymbol{A}_{n\times m}$，非负矩阵的行表示样本数据，列表示能量端。在非负矩阵分解时应在保证有效信息最大化的前提下进行的，利用欧几里得距离作为成本函数，用于量化分解精度即

$$\| \boldsymbol{A}-\boldsymbol{WH} \|^2 = \sum_{n,m}(\boldsymbol{A}_{nm}-(\boldsymbol{WH})_{nm})^2 \tag{1}$$

先预设置端元数目 r，其值小于行值、列值，对非负数矩阵 $\boldsymbol{A}_{n\times m}$ 按照分解的数学表达式进行分解

$$\begin{cases} \boldsymbol{A}_{n\times m} \approx \boldsymbol{W}_{n\times r}\boldsymbol{H}_{r\times m} = \tilde{\boldsymbol{A}}_{n\times m} \\ \text{s. t. } \boldsymbol{W}_{n\times r} \geqslant 0 \\ \text{s. t. } \boldsymbol{H}_{r\times m} \geqslant 0 \end{cases} \tag{2}$$

其中，$\boldsymbol{W}_{n\times r}$是端元光谱矩阵，其每一列代表着一个端元光谱，$\boldsymbol{H}_{r\times m}$是端元丰度矩阵，其每一列代表一个样本不同端元的丰度值。

非负矩阵分解(NMF)数据分析具体操作过程如下：

(1)随机生成端元光谱矩阵 $\boldsymbol{W}_{n\times r}$；

(2)固定 $\boldsymbol{W}_{n\times r}$，按照成本函数下的乘性更新法则 $\boldsymbol{H}_{rm} \leftarrow \boldsymbol{H}_{rm}\frac{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V})_{rm}}{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{H})_{rm}}$更新到 $\boldsymbol{H}_{r\times m}$直到收敛；

(3)固定 $\boldsymbol{H}_{r\times m}$，按照第二条乘性更新法则 $\boldsymbol{W}_{nr} \leftarrow \boldsymbol{W}_{nr}\frac{(\boldsymbol{V}\boldsymbol{H}^{\mathrm{T}})_{nr}}{(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})_{nr}}$更新到 $\boldsymbol{W}_{n\times r}$直到收敛；

(4)设置迭代次数，重复(2)和(3)的步骤直到成本函数不变或变化很小，迭代算法的收敛性在文献[7]中已经给出证明。

2.4.2 非负矩阵端元数目的确定

如果 r 过小会使得部分真实信息丢失，过大会达不到去除含水量这一因素影响的效果[8]，对于 r 的确定采用峰值信噪比(peak signal-to-noise ratio，PSNR)作为评价标准，峰值信噪比使用均方误差(MSE)来定义[9]

$$\begin{cases}\mathrm{MSE}_k = \dfrac{1}{N}\sum_{i=1}^{N}(A_{ik} - (WH)_{ik})^2 \\ \mathrm{PSNR}_k = 10\log_{10}\dfrac{\max_k}{\mathrm{MSE}_k}\end{cases} \tag{3}$$

其中，N 表示样本个数，下标(i, k)表示第 i 个样本第 k 个能量段表示的像素点，$\max_k$ 表示第 k 个能量段的最大像素值。

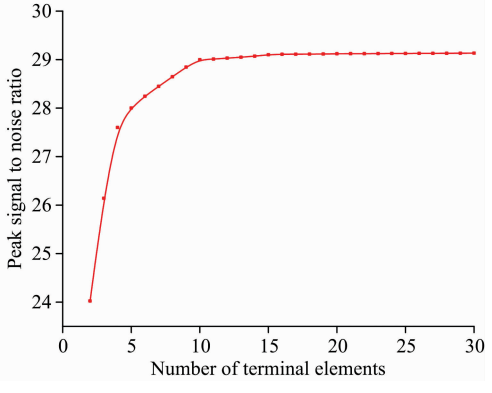

图 4 r 对峰值信噪比的影响

Fig. 4 Effect of r on peak signal to noise ratio

PSNR 的值越大表示端元光谱矩阵 $\boldsymbol{W}_{n\times r}$和端元丰度矩阵 $\boldsymbol{H}_{r\times m}$融合成新的光谱矩阵的准确度越高，在非负矩阵分解时端元数目设置从 2 到 30，其间隔为 1，图 4 中可以清楚看出当端元数目至 10 时趋于稳定，后面数据波动很小，所以端元数目设置为 10。

2.4.3 NMF 模型构建

NMF 算法的目的是获取有用的光谱，去除含水量影响的干扰光谱。由非负矩阵分解得到端元光谱矩阵 $\boldsymbol{W}_{n\times r}$和端元丰度矩阵 $\boldsymbol{H}_{r\times m}$融合成新的光谱矩阵，新的光谱矩阵与原始光谱矩阵的差异矩阵为光谱差异矩阵。图 5 表示了同一重金属含量不同含水量下 NMF 算法处理前后的光谱。

当 r 等于 5 时，除了 21.69%和 25%含水量较大的样本其他样本间光谱计数率曲线变化一致，光谱间的差异较小；当 r 等于 10 时，不同含水量样本光谱间的差异微小，呈现出较好的相似性，重金属特征峰附近差异也明显消失，有效地降低了含水量对光谱的影响。

计算了样本处理前含水量梯度两两之间的光谱数据相关系数，之间的相关系数相差较大，相关系数随着含水量的增加而不断减小，其中含水量在 10%和 25%之间的相关系数最小为 0.72；使用 NMF 算法处理后，样本间的相关系数变化较小且接近于 1，故进一步说明了 NMF 算法消除了含水量对光谱的影响。

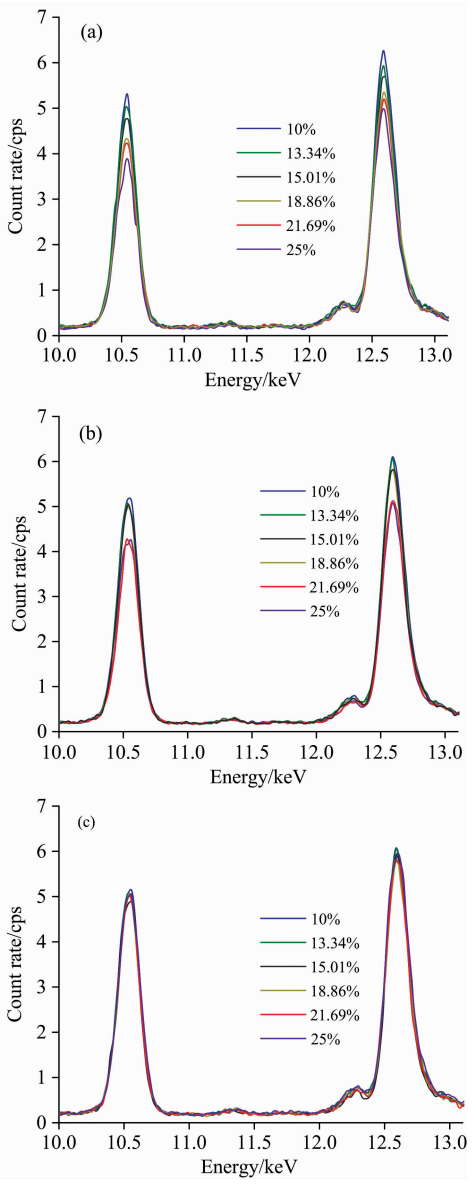

图 5 NMF 算法处理前后的光谱

(a)：NMF 处理前的光谱图；

(b)：NMF 在端元数目 $r=5$ 处理后的光谱图；

(c)：NMF 在端元数目 $r=10$ 处理后的光谱图

Fig. 5 Spectrum before and after NMF algorithm processing

(a)：Spectrum before NMF processing；(b)：The spectrum of NMF processed by the number of end elements $r=5$；(c)：The spectrum of NMF after the number of end elements $r=10$

2.5 PLSR，EPO-PLSR 和 NMF-PLSR 模型的建立与对比

偏最小二乘法预测(PLSR)是一种多变量统计分析方法，

同时考虑了分解自变量光谱矩阵和目标因变量重金属含量矩阵的影响，将数据压缩与回归拟合相结合，使建立的模型具有更好的稳健性，并能有效地解决自变量之间的多重共线性。本研究中使用了留一法交叉验证防止过拟合，主成分数是由交叉验证决定的 $n=2$。

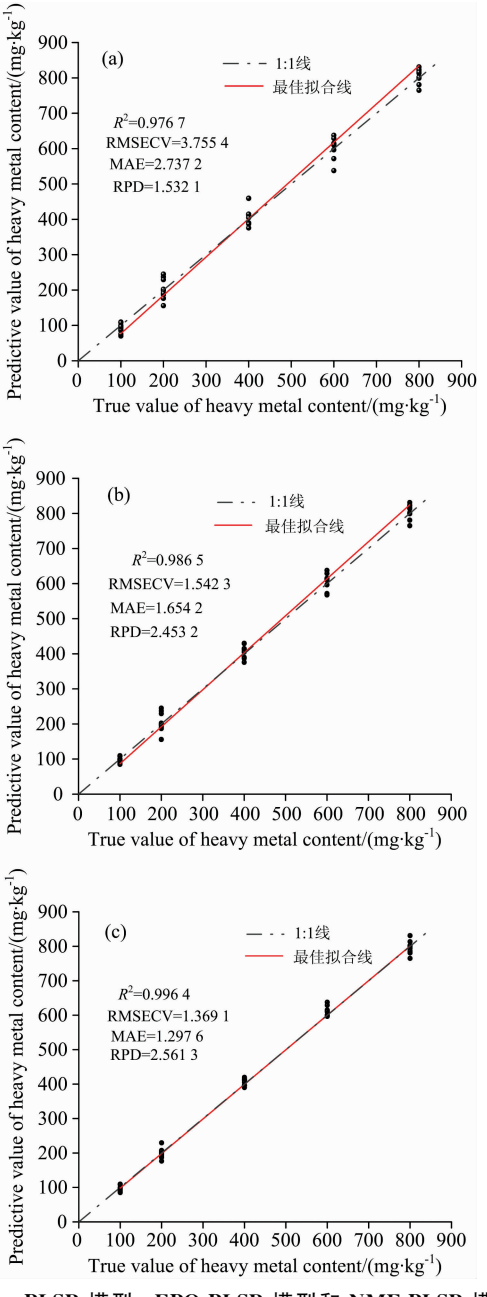

图 6　PLSR 模型、EPO-PLSR 模型和 NMF-PLSR 模型实测值和预测值对比

(a)：PLSR 模型；(b)：EPO-PLSR 模型；(c)：NMF-PLSR 模型

Fig. 6　Comparison of measured and predicted values of PLSR model, EPO-PLSR model and NMF-PLSR model

(a): PLSR model; (b): EPO-PLSR model; (c): NMF-PLSR model

将计算得到的净峰面积和土壤含水量作为模型输入，土壤重金属 Pb 的含量作为输出。将得到的 150 组数据集使用留出法进行 7/3 分样分成两个互斥的集合，70% 的数据即 105 组数据作为训练集，30% 的数据即 45 组数据作为测试集。并采用 EPO 算法进行去除含水量，建立了 EPO-PLSR 模型用作对比。模型的精度检验及评价指标使用建模决定系数 R^2、交叉验证均方根误差 RMSECV、平均绝对误差 MAE 和相对分析误差 RPD，其中 RPD 是预测样本标准差和预测均方根误差 RMSEP 的比值，当 RPD＜1.4 时认为模型预测是不可用的；当 1.4≤RPD＜2 时认为模型预测比较粗糙，精度不够；当 RPD≥2 时认为模型的预测能力较好[10]。

由图 6(a)和(b)对比可以看出，与 PLSR 模型相比，EPO-PLSR 模型的 R^2 和 RPD 分别提高了 0.009 8 和 0.921 1，RMSECV 和 MAE 分别降低了 2.212 2 和 1.083，说明使用 EPO-PLSR 预测土壤重金属的结果明显优于 PLSR，相对于 PLSR 有了明显改善，且 PLSR 的 RPD 值为 1.53 预测不够准确，只能实现在不同含水量下对土壤重金属含量的粗略估算。经过 NMF-PLSR 模型处理后的数据相对于 PLSR 模型改善较大，R^2 和 RPD 分别提高了 0.019 7 和 1.029 2，RMSECV 和 MAE 分别降低了 2.386 3 和 1.439 6，相对于 EPO-PLSR 模型的 R^2 和 RPD 分别提高了 0.009 9 和 0.108 1，RMSECV 和 MAE 分别降低了 0.244 7 和 0.356 6，NMF-PLSR 最佳拟合线更加接近 1∶1 线，说明了 NMF-PLSR 模型比 EPO-PLSR 模型预测能力强、模型解释能力较好；观察预测值，NMF-PLSR 模型在相同重金属含量下预测值相差较小，进一步说明了去除含水量的能力及预测的稳定性。

3　结　论

通过实验的方法获取了在相同重金属含量条件下不同含水量土样的光谱，分析了土壤含水量对光谱的影响，然后采用了非负数分解算法去除含水量的影响，利用峰值信噪比评价标准确定端元数目，并建立了 NMF-PLSR 模型对土壤重金属含量进行反演，得到了以下结论：

(1)随着含水量的增加，土壤光谱反射率的非线性降低，相对偏差从 7.41% 升高到 27.78%，仪器的检测精度下降明显，含水量对基于能量色散 X 射线光谱法的土壤重金属含量反演影响显著。

(2)当端元数目为 10 时峰值信噪比趋于稳定，经 NMF 算法处理后土壤重金属特征峰重叠明显，光谱间的相似系数更加接近 1，光谱间的差异得到了明显去除。

(3)建立的 NMF-PLSR 模型相对于直接建立 PLSR 模型效果有显著改善，对比 EPO-PLSR 模型也有一定的改善，R^2 和 RPD 分别提高了 0.009 9 和 0.108 1，RMSECV 和 MAE 分别降低了 0.244 7 和 0.356 6，更加准确实现土壤重金属含量的反演，从而为土壤重金属含量检测提供了一定的技术支持。

References

[1] Martinuzzi Sebastián, Januchowski - Hartley Stephanie R, Pracheil B M, et al. Global Change Biology, 2014, 20(1): 113.

[2] Zheng R, Zhao J L, Zhou X, et al. Pedosphere, 2016, 26(1): 74.

[3] Ji W, Rossel R A V, Shi Z. European Journal of Soil Science, 2015, 66(3): 555.

[4] HU Xiao-yan, CUI Xu, HAN Xiao-ping, et al(胡晓艳, 崔 旭, 韩小平, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(4): 1059.

[5] Minasny B, Mcbratney A B, Bellon-Maurel V, et al. Geoderma, 2011, 167-168: 118.

[6] Maruyama R, Maeda K, Moroda H, et al. Neural Networks, 2014, 55(55c): 11.

[7] Lee D D, Seung H S. Proceedings of the 13th International Conference on Neural Information Processing Systems. MIT Press, 2000.

[8] Lin C J. IEEE Transactions on Neural Networks, 2007, 18(6): 1589.

[9] Sakhaii P, Bermel W. Journal of Magnetic Resonance, 2014, 242(3): 220.

[10] YU Lei, HONG Yong-sheng, ZHOU Yong, et al(于 雷, 洪永胜, 周 勇, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2016, 32(13): 95.

Study on Inversion Model of Soil Heavy Metal Content Based on NMF-PLS Water Content

WU Xi-jun[1], ZHANG Jie[1], XIAO Chun-yan[2], ZHAO Xue-liang[1,3], LI Kang[3], PANG Li-li[3], SHI Yan-xin[3], LI Shao-hua[4]

1. Hebei Province Key Laboratory of Test/Measurement Technology and Instrument, School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China
2. School of Resources and Environment, Henan University of Technology, Jiaozuo 454000, China
3. Center for Hydrogeology and Environmental Geology, China Geological Survey, Geological Environment Monitoring Engineering Technology Innovation Center of The Ministry of Natural Resources, Baoding 071051, China
4. Hebei Sailhero Environmental Protection Hi-Tech Co., Ltd., Shijiazhuang 050000, China

Abstract The excessively high content of heavy metals in the soil is hugely harmful, not only causing serious environmental pollution, but entering the human body through the food chain poses a serious threat to human health, so it is very important for heavy metal detection. X-ray fluorescence spectroscopy has been widely used because of its short detection time, non-destructive testing, and low testing costs. However, the detection of spectral data is severely disturbed by soil moisture factors, which leads to lower accuracy in estimating the heavy metal content in the soil directly. Taking the soil samples of Mancheng District, Baoding City, Hebei Province as the research object, the collected soil samples were cleaned, screened, dried, and then added with a certain amount of heavy metal solution to prepare samples with different water content and heavy metals for detection. The Mahalanobis distance and NJW clustering were calculated for the abnormal data in the experiment, and the influence of soil moisture content on the heavy metal spectrum was analyzed, the results show that the spectral repeatability of different water content is poor, and the spectral intensity decreases nonlinearly with the increase of soil water content. The Savitzky-Golay convolution smoothing denoising method and linear background method are used to preprocess the spectrum to solve the problems of noise and baseline drift caused by the environment and the instrument itself. A non-negative matrix factorization algorithm was used to deal with the peak signal-to-noise ratio evaluation model to determine the number of end elements. The results show that the peak signal-to-noise ratio tends to increase when the number of end elements increases to 10. The stable fluctuation is very small. After the non-negative matrix decomposition treatment, the spectrum repeatability and similarity are good among the same heavy metal content and different water content. The correlation coefficient between the spectra is calculated to prove the similarity between the spectra further. A partial least squares prediction model was established after removing the water content for spectral interference. In order to verify the accuracy of the prediction model, a partial least squares prediction model with no water content removed was established, and the partial water content was removed by orthogonalization with external parameters The least squares prediction model is evaluated using the evaluation parameter determination coefficient (R^2), cross-validated root mean square error (RMSECV), average absolute error (MAE), and relative analysis error (RPD). Validation results show that compared to models built without removing water content, non-negative moments are used partial least squares model estab-

lished by matrix decomposition and removal of water content R^2 and RPD increased by 0.019 7 and 1.029 2, RMSECV and MAE decreased by 2.386 3 and 1.439 6; Compared to the partial least squares model established by the external parameter orthogonalization method, the RPD and RPD increased by 0.009 9 and 0.108 1, and the RMSECV and MAE decreased by 0.244 7 and 0.356 6, it is shown that the model established after denoising by non-negative matrix decomposition can effectively improve the accuracy and robustness of prediction. Non-negative matrix factorization can effectively eliminate the effect of soil water content on the spectrum, and the partial least squares model established on this basis has realized the inversion of soil heavy gold content and provided certain technical support for quantitative detection of heavy metals.

Keywords Soil heavy metals; Energy dispersive X-ray fluorescence spectra; Non-negative matrix factorization; Partial least squares

(Received Dec. 17, 2019; accepted Apr. 29, 2020)

本 刊 声 明

近期以来，一些不法分子假冒《光谱学与光谱分析》期刊社名义，以虚假网站等形式欺骗广大作者、读者。这些虚假网站公然假冒《光谱学与光谱分析》期刊名义进行大肆的征稿并骗取作者的审稿费和版面费。经部分作者及读者举报，现有关部门已就此介入调查。我刊将通过法律途径向假冒者追究相应的责任，维护本刊权利。

本刊官方网站已正式开通，网址为

http://www.gpxygpfx.com/

在此郑重声明，本网址为《光谱学与光谱分析》期刊唯一开通运行的官方网站。本刊从未授权任何单位或个人以任何形式（包括网上网下）代理本刊征稿、审稿等项业务。

希望广大读者和作者切实维护好自身的合法权益，防止受骗上当。

《光谱学与光谱分析》期刊社

2019 年 3 月 15 日