基于PCA的土壤Cd含量高光谱反演模型对比研究

郭 飞[1,2]，许 镇[3*]，马宏宏[1,2]，刘秀金[1,2]，杨 峥[1,2]，唐世琪[1,2]

1. 中国地质科学院地球物理地球化学勘查研究所，河北 廊坊 065000
2. 中国地质调查局土地质量地球化学调查评价研究中心，河北 廊坊 065000
3. 中国科学院空天信息创新研究院，北京 100101

摘 要 土壤重金属污染对人类健康造成了极大的威胁，如何快速摸清土壤污染情况尤为重要。高光谱遥感具备光谱分辨率高，快速无损等优势，使其在土壤组分反演方面具有巨大的潜力。针对高光谱信息冗余及光谱变换对土壤镉(Cd)含量估算的影响进行分析，并利用变换前后的高光谱数据对比研究了不同高光谱模型对土壤Cd含量反演的性能。首先利用等离子体质谱法和FieldSpec4地物光谱仪收集了56组土壤样品的Cd含量和对应的高光谱曲线(350～2 500 nm)；为了弱化光谱测定中光亮变化和土壤表面凹凸对实验结果的影响，研究对高光谱数据进行倒数对数预处理；考虑到高光谱数据中存在大量的信息冗余，研究采用了主成分分析(PCA)对高光谱数据进行降维处理并最终保留了前12个主成分量作为特征变量。针对高光谱反演模型，研究选择了偏最小二乘(PLSR)、支持向量机(SVM)、人工神经网络(ANN)和随机森林(RF)四种回归模型建立PCA主成分与Cd含量之间的关系；最后，研究选取了决定系数(R^2)、均方根误差(RMSE)和RPD三种精度评估指标评估回归模型的拟合精度，结果表明针对光谱采用PCA波段降维的方法处理后，选取的12个主成分对变化前后的光谱累计贡献率均达到99.99%，作为模型的输入变量，四种模型均具有一定的预测能力。无论光谱变换与否，PCA-RF反演模型的预测能力均为最好(R^2 分别为0.856和0.855，RPD均高达3.39)。利用PCA对高光谱数据降维处理可以有效降低高光谱数据冗余，有力的保证模型的预测能力。以PCA筛选出的主成分量可以作为模型极好的输入变量，以RF为基础的高光谱反演模型在反演土壤Cd含量时具有最佳效果，可为该区域及类似地区的土壤重金属污染物反演提供新的方法支撑。

关键词 Cd含量；高光谱；主成分分析；反演模型对比

中图分类号：TP79 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)05-1625-06

引 言

镉(Cadmium，Cd)是一种有毒重金属[1]，它不仅会降低土壤微生物的生物活性，还易通过在植物可食部位的累积，进入食物链危害人体[2]。矿产资源的开采是造成其周边农用地土壤污染的重要原因之一[3]。如何快速有效测定土壤中Cd的含量及空间分布已成为目前亟待解决的问题。高光谱遥感由于光谱分辨率高、波段连续，能快速高效获取精细的光谱信息等优势，成为快速查明土壤重金属污染状况的新技术方法之一[4]。

利用可见-近红外光谱对土壤重金属含量进行定量反演已成为国内外热点研究问题。Kemper等[5]利用可见-近红外光谱，基于线性模型MLR建模预测As和Cd等重金属含量，认为土壤重金属含量与铁、铁氧化物相关；有研究采用SMLR，PLSR等线性建模方法建立了土壤重金属含量反演模型；有报道基于逐步回归和相关系数方法，筛选出对重金属敏感的特征波段，将它们组合成综合特征变量对研究区Cu元素进行了反演。尽管国内外关于土壤重金属含量估算相关研究逐渐增多，但是仍存在一些问题。例如，针对高光谱数据波段信息冗余的问题，多数研究选择丢掉大量的波段，仅利用相关系数以及逐步回归法筛选出了部分特征波段，损失了大量有用的信息。事实上，土壤中重金属含量与光谱曲线之间的关系很难用几个波段解释。因此，选择一种

收稿日期：2020-05-26，**修订日期**：2020-08-31

基金项目：中国地质科学院地球物理地球化学勘查研究所所长基金项目(AS2019J02)，国家自然科学基金项目(41503024)，中国地质调查局地质调查项目(DD20190518)资助

作者简介：郭 飞，1991年生，中国地质科学院地球物理地球化学勘查研究所助理工程师 e-mail：guofei@igge.cn

*通讯作者 e-mail：xuzhen@radi.ac.cn

既可以保证波段主要信息量，又能减少输入变量的特征参数尤为重要。此外，关于土壤重金属含量估算模型的问题，绝大部分的研究主要采用线性回归模型，如 SMLR 和 PLSR 等；而非线性回归模型考虑较少。客观上讲，土壤中重金属含量在光谱曲线上的响应会受多种因素影响，二者之间关系非常复杂；而简单线性回归模型很难处理非线性、随机性等复杂的问题。因此，在高光谱模型的选择上应对非线性模型加以考虑。

选择湖北省黄石市矿山周边农用地土壤为研究对象，针对于高光谱反演中波段信息冗余等问题，提出了基于 PCA 的降维方法，结合多种高光谱反演模型，验证 PCA 筛选主成分量可实现土壤重金属含量的高精度反演，并通过不同高光谱模型的对比，确定了适合该研究区域 Cd 含量的最佳预测模型，从而实现了土壤 Cd 含量的快速、精确光谱检测，为土壤重金属反演提供新的思路。

1 实验部分

1.1 研究区与土壤采样

研究区位于湖北省东南部的黄石市(114°30′—115°30′E，29°30′—30°20′N)，地处长江中下游，具有典型的大陆性季风气候。地势南高北低，东西平，海拔高度为 120～200 m。研究区内矿产资源丰富，有多个大中型矿床，矿山开采、冶炼生产对周边土壤造成一定的重金属污染。在研究区共采集 0～20 cm 表层土壤 56 件，采样点(图 1)位于矿山周边的农用地，采集表层土样初始质量大于 1 kg，样品经室内自然风干、研磨后过 10 目(孔径 2 mm)的尼龙筛，利用四分法分成两份，分别用于室内光谱测试和实验室化学分析。

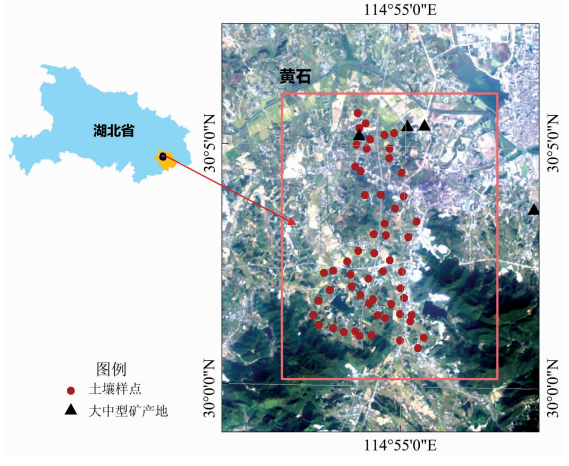

图 1 研究区采样点分布示意图

Fig. 1 The distribution of sampling point in the study area

1.2 光谱测定

土壤光谱数据获取采用美国 ASD 公司的 FieldSpec4 地物光谱仪(光谱波段范围 350～2 500 nm)，利用卤素光源和标准白板完成测量。该光谱仪采样间隔为 1.4 nm(350～1 000 nm)和 2 nm(1 000～2 500 nm)，经光谱重采样后(间隔 1 nm)，共输出 2 151 个波段。测试在暗室进行，选择一稳固平台，将土样放入直径 90 mm，高 19 mm 的透明玻璃器皿，使其表面尽量平整，以 50 W 的卤素灯为光源，光源与样品保持 50 cm 距离，光源探头位于样本正上方 7 cm 高，光线与样品保持 15°的照射角度，保证测量时无阴影遮挡。开机预热 30 min 后对仪器进行调整和校准并开始测量。每个土壤样本采集 10 条光谱曲线，取光谱反射率的平均值作为样本的反射率光谱值，剔除 350～399 和 2 450～2 500 nm 信噪比低、噪声大的边缘波段，共获得 2 050 个波段数据。

1.3 数据处理

1.3.1 光谱预处理

土壤样品光谱数据测定过程中，由于光线亮度变化和土壤表面凹凸不平会对实验结果产生影响，采用取光谱反射率倒数对数的方法来避免此影响。倒数对数[6]计算公式为

$$\log\left(\frac{1}{R(\lambda_i)}\right)=-\log[R(\lambda_i)] \tag{1}$$

其中 λ_i 为光谱波长值，$R(\lambda_i)$为对应光谱波段的反射率。

主成分分析(principal components analysis，PCA)是由 Pearson 于 1901 年提出的一种分析、简化数据集的方法[7]。该方法的优势在于降低数据集维数，同时保证信息量最大，对于拥有大量波段信息的高光谱数据，通过一系列的矩阵变化，在测量空间寻找几组正交向量，保留数据方差最大、信息量最多的组分，从而达到高光谱数据降维的目的。主要步骤如下：

(1)将波段数据组合成为矩阵，设随机变量 X_1，X_2，…，X_P；其样本均数为 X_1，X_2，…，X_P；样本标准差记为 S_1，S_2，…，S_P。首先进行标准化变换

$$x_i=\frac{X_i-\overline{X}}{S_i} \tag{2}$$

(2)计算标准化后样本矩阵的协方差矩阵，若 $C_1=a_{11}x_1+a_{12}x_2+\cdots+a_{1p}x_p$，$a_{11}^2+a_{12}^2+\cdots+a_{1p}^2=1$，且令 Var($C_1$)最大，则称 C_1 为第一主成分；若 $C_2=a_{21}x_1+a_{22}x_2+\cdots+a_{2p}x_p$，$a_{21}^2+a_{22}^2+\cdots+a_{2p}^2=1$，($a_{21}$，$a_{22}$，…，$a_{2p}$)垂直于 a_{11}，a_{12}，…，a_{1p}，且令 Var(C_{12})最大，则称 C_2 为第二主成分；

以此类推求得第三，第四，…，第 p 个主成分。保留主成分个数取决于累积方差在总方差所占百分比(贡献率)。

1.3.2 反演模型

利用 PCA 对原始光谱以及倒数对数光谱进行波段降维，将累积贡献率达到 99.99%的主成分作为特征变量，选择线性模型 PLSR，以及非线性模型 SVM，ANN 和 RF 分别建立土壤 Cd 含量估算模型。PLSR 是一种常用于高光谱反演土壤元素含量的新型多元统计方法[8]，它能够很好地解决自变量间多重共线问题。SVM 是以内核统计学理论为基础理论，它的优势主要体现在解决小样本、非线性以及高维模式的识别[9]。ANN 由一组相互连接的人工神经元组成，利用大量神经元之间的链接结构进行分布式并行信息处理的数学模型，该模型基本架构由输入层、输出层和隐藏层三部分构成[10]。RF 是一个组合分类器算法[11]，由一系列决策树组成，利用自助法重采样技术，在初始样本数据集上生成多个自助样本集，每个自助样本集是每棵分类树的全部训练数据，然后根据自主样本集生成多个分类树组成随机森林。

1.3.3　精度评估

采用 R^2、RMSE 和 RPD 评价指标对估算模型的反演精度进行评估。R^2 和 RPD 越大，RMSE 越小，说明预测效果越好，通常认为 R^2 越趋近于 1，模型的预测效果越佳。当 RPD＞2 时，模型具极好的预测能力；当 1.4＜RPD＜2 时，模型有一定的预测能力；当 RPD＜1.4 时，模型不具备预测能力。

2　结果与讨论

2.1　土壤重金属含量统计分析

利用等离子体质谱法测定 Cd 含量，Cd 元素的描述性统计结果如表 1 所示，Cd 均值为 0.64 mg·kg^{-1}。根据土壤环境质量标准(GB15618—2018)，该区域的 Cd 含量高于农用地污染风险筛选值，而低于管制值，存在一定的土壤污染风险。从空间分布看，其变异系数介于 0.5～0.75 之间，属于中等变异，说明 Cd 在土壤中分布不均，空间变异较为显著。将 56 个样本数据按照 7∶3 比率随机分割，训练样本 39 个，用于筛选模型输入变量。验证样本 17 个，用于对高光谱模型的评估。

2.2　光谱变换与 PCA 降维

所有土壤样本原始光谱反射率曲线[图 2(a)]趋势大致相同，在可见光区域反射率呈明显上升，超过 800 nm 后光谱曲线趋于平缓。在 1 400，1 900 和 2 200 nm 附近有三个明显凹陷的吸收峰，为土壤黏土矿物的吸收特征。经倒数对数变化后[图 2(b)]的光谱曲线与原始曲线的变化趋势基本相反。

表 1　土壤 Cd 含量描述统计分析(mg·kg^{-1})

Table 1　Descriptive statistics analysis of soil Cd content

元素	样本数	平均值	最大值	最小值	标准差	偏度	峰度	变异系数
训练集	39	0.65	2.11	0.04	0.40	1.20	2.26	0.62
验证集	17	0.55	1.36	0.09	0.32	0.7	1.31	0.58
总样本集	56	0.64	2.11	0.04	0.40	1.20	2.26	0.62

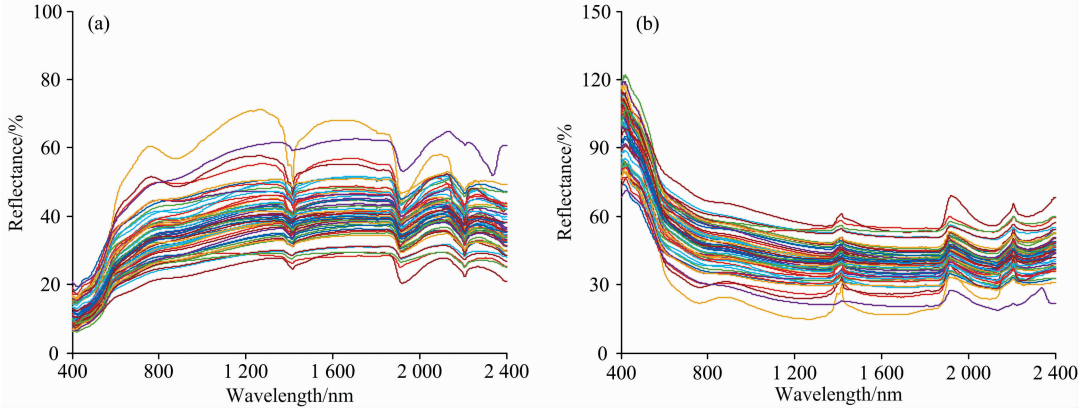

图 2　(a)土壤样本原始光谱反射率曲线；(b)土壤样本倒数对数光谱反射率曲线

Fig. 2　(a) The original spectral reflectance curve of soil samples; (b) The reciprocal logarithmic spectral reflectance curve of soil samples

利用 PCA 算法，对原始光谱曲线以及变换后的倒数对数光谱曲线的 2 050 个波段进行降维，原始光谱曲线和倒数对数光谱曲线各主成分的贡献率和累计贡献率值如表 2 所示。选取经 PCA 之后，累计贡献率达到 99.99%的主成分个数作为模型的输入变量，其中，原始光谱累积贡献率达到 99.99%的主成分个数为 12 个，光谱变换之后累积贡献率达到 99.99%的主成分个数也为 12 个。将 PCA 降维选取的组分作为四种模型的输入变量。

2.3　土壤重金属含量反演模型建立与检验

将 PCA 降维选取的主成分作为模型的自变量(X)，土壤 Cd 含量为因变量(Y)，采用线性回归 PLSR 模型，以及非线性回归 SVM，ANN 和 RF 模型分别建模比对，验证基于 PCA 筛选的特征变量对不同模型预测能力的影响，以及优选出研究区 Cd 含量的最佳预测模型。

2.3.1　基于 PCA 原始光谱建模

利用 PCA 对原始光谱数据降维，选择累计贡献率达 99.99%的 12 个主成分量作为模型输入变量，运用 PLSR，SVM，ANN 和 RF 方法建模，土壤 Cd 含量的反演模型[图 3(a)]的精度评价如表 3 所示，根据图 3(a)与表 3 可知：PCA-RF 的决定系数(R^2＝0.856)最高，RPD 高达 3.39，表明 PCA-RF 模型具有极好的预测能力，是预测土壤 Cd 含量的优势模型；PCA-ANN 和 PCA-SVM 的 RPD 都高于 2，其决定系数(R^2)分别为 0.621 和 0.581，两种模型同样具有好的预测能力；而 PCA-PLSR 的 R^2 和 RPD 分别仅为 0.484 和 1.8，该模型的预测能力一般。经 PCA 降维选取的特征波段，使得模型均具有一定的预测能力。

表 2 主成分贡献率

Table 2 Principal component contribution rate

成分	原始光谱		成分	倒数对数光谱	
	贡献率/%	累积贡献率/%		贡献率/%	累积贡献率/%
1	92.341	92.341	1	92.913	92.913
2	3.890	96.231	2	3.063	95.977
3	2.332	98.563	3	2.479	98.456
4	0.827	99.390	4	0.919	99.375
5	0.374	99.764	5	0.405	99.780
6	0.076	99.840	6	0.083	99.864
7	0.055	99.895	7	0.052	99.916
8	0.048	99.943	8	0.033	99.948
9	0.027	99.970	9	0.024	99.973
10	0.008	99.979	10	0.008	99.981
11	0.006	99.985	11	0.005	99.986
12	0.004	99.989	12	0.004	99.990

2.3.2 基于 PCA 倒数对数光谱建模

运用四种方法对 PCA 降维后的倒数对数光谱进行建模[图 3(b)]，其反演精度如表 3。由图 3(b)和表 3 可知：PCA-RF 模型的预测能力在光谱变换后仍为最佳，其 R^2 为 0.855，RPD 为 3.39，表明模型仍具有极好的预测能力；PCA-ANN 次之，其 R^2 为 0.623，RPD 为 2.12，模型同样具有好的预测能力；PCA-SVM 的 R^2 为 0.607，RPD 为 2.00，模型也具有好的预测能力，而 PCA-PLSR 的 R^2 为 0.535，RPD 仅为 1.89，模型预测能力一般。

2.3.3 基于 PCA 原始光谱-倒数对数模型对比分析

四种模型的预测能力顺序在光谱变换前后未发生改变(图 4)，光谱变换对于各模型的预测能力有所提升，其中提升效果最为显著的是 PCA-PLSR 模型，该模型的 R^2 提升了 10.5%，RPD 提升了 5.0%，其次为 PCA-SVM 模型，该模型的 R^2 提升了 4.5%，RPD 提升了 2.5%，PCA-ANN 模型，R^2 和 RPD 分别提升了 1.8%和 1.4%，而 PCA-RF 模型无明显改变。

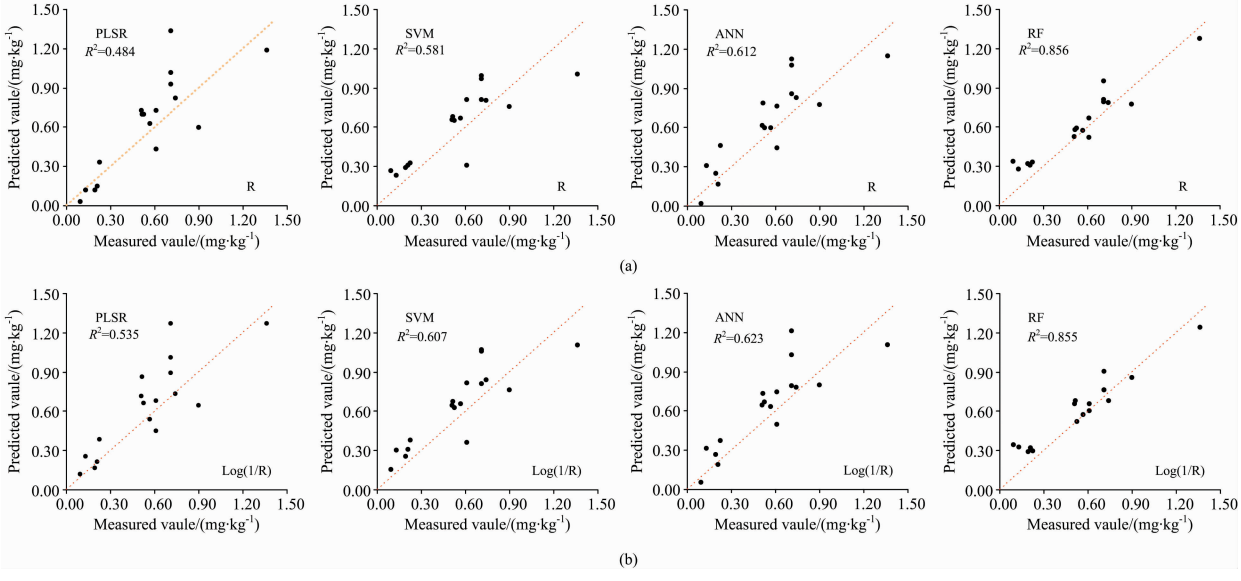

图 3 (a) 原始光谱不同预测模型散点图；(b) 倒数对数光谱不同预测模型散点图

Fig. 3 (a) Scatterplots of different prediction models based on original spectral data; (b) Scatter plots between different prediction models based on reciprocal logarithmic spectral

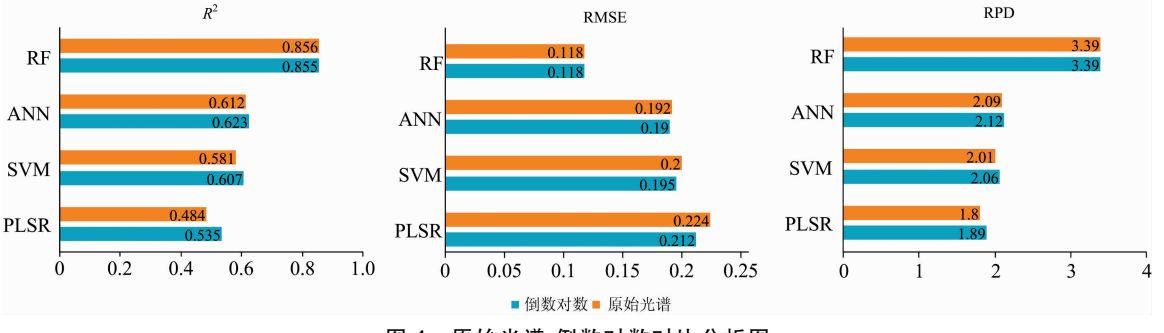

图 4 原始光谱-倒数对数对比分析图

Fig. 4 The contrast analysis diagram between original spectral and reciprocal logarithmic

表 3　基于原始光谱-倒数对数光谱不同模型精度评价

Table 3　Accuracy evaluation of different models based on original spectral-reciprocal logarithmic spectral

	模型	原始光谱			倒数对数光谱		
		R^2	RMSE	RPD	R^2	RMSE	RPD
线性模型	PCA-PLSR	0.484	0.224	1.80	0.535	0.212	1.89
非线性模型	PCA-SVM	0.581	0.200	2.01	0.607	0.195	2.06
	PCA-ANN	0.612	0.192	2.09	0.623	0.190	2.12
	PCA-RF	0.856	0.118	3.39	0.855	0.118	3.39

通过对比光谱变化前后各模型的预测精度可得，非线性模型的预测能力优于线性模型，倒数对数光谱变换对于模型的预测能力有所提升，可弱化光谱数据测定时光线亮度和土壤表面凹凸产生影响。

3　结　论

以湖北省黄石市矿区周边农用地土壤为研究对象，利用 PCA 方法对光谱变化前后数据进行降维，选取特征变量，在此基础上对比分析了不同反演模型对土壤 Cd 含量测定的反演精度，得出如下结论：

(1)经倒数对数变换后的光谱，预测能力有所提升，PCA-PLSR 模型的提升效果最为明显，PCA-SVM 和 PCA-ANN 稍有提高，倒数对数变换可弱化光谱测定中光强度变化和土壤表面凹凸的影响。

(2)利用 PCA 方法进行降维处理可以有效降低高光谱数据量，选取的 12 个主成分量对变化前后的光谱累计贡献率可达 99.99%，四种模型均具有一定的预测能力，保证模型具有极好的输入变量。

(3)不同模型的反演精度顺序为：PCA-RF＞PCA-ANN＞PCA-SVM＞PCA-PLSR，非线性模型 PCA-RF，PCA-ANN 和 PCA-SVM 的 RPD 均大于 2，具有极好的预测能力，其中 PCA-RF 模型的 RPD 超过 3，说明模型具有较高稳定性和预测精度。

本研究主要采用 PCA 对光谱数据进行降维，对比分析不同模型的反演能力，PCA-RF 模型可为土壤重金属含量反演提供很好的参考依据。PCA 对高光谱数据特征变量选取具有显著效果，但仍存在其他的降维方法，需要进一步深入研究。

References

[1] Liao M, Xie X, Ma A, et al. Journal of Soil & Sediments, 2010, 10(5): 818.

[2] Gu Y W, Li S, Gao W, et al. Acta Ecologica Sinica, 2015, 35(13): 4445.

[3] Zhang X, Sun W, Cen Y, et al. The Science of the Total Environment, 2019, 650(PT. 1(1-834)): 321.

[4] JIANG Zhen-lan, YANG Yu-sheng, SHA Jin-ming(江振蓝, 杨玉盛, 沙晋明). Journal of Geographical(地理学报), 2017, 72(3): 533.

[5] Kemper T, Sommer S. Environmental Science & Technology, 2002, 36(12): 2742.

[6] ZHANG Fang, XIONG Hei-gang, LUAN Fu-ming, et al(张　芳, 熊黑钢, 栾福明, 等). Journal of Infrared and Millimeter Waves(红外与毫米波学报), 2011, (1): 57.

[7] LI Yuan-bo, CAO Han(李远博, 曹　菡). Computer Technology and Development(计算机技术与发展), 2016, (2): 26.

[8] WANG Shi-dong, SHI Pu-jie, ZHANG He-bing, et al(王世东, 石朴杰, 张合兵, 等). Chinese Journal of Ecology(生态学杂志), 2019, 38(1): 300.

[9] DONG Cheng-wei, RUI Xiao-ping, DENG Yu, et al(董承玮, 芮小平, 邓　羽, 等). Geography and Geo-Information Science(地理学与地理信息科学), 2014, (4): 36.

[10] Darwishe H, El Khattabi J, Chaaban F, et al. Environmental Earth Sciences, 2017, 76(19): 649.1.

[11] FANG Kuang-nan, WU Jian-bin, ZHU Jian-ping, et al(方匡南, 吴见彬, 朱建平, 等). Statistics & Information Forum(统计与信息论坛), 2011, 26(3): 32.

A Comparative Study of the Hyperspectral Inversion Models Based on the PCA for Retrieving the Cd Content in the Soil

GUO Fei[1,2], XU Zhen[3*], MA Hong-hong[1,2], LIU Xiu-jin[1,2], YANG Zheng[1,2], TANG Shi-qi[1,2]

1. Institute of Geophysical & Geochemical Exploration, Chinese Academy of Geological Sciences, Langfang 065000, China
2. Research Center of Geochemical Survey and Assessment on Land Quality, China Geological Survey, Langfang 065000, China
3. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China

Abstract The soil heavy metal pollution poses a great threat to the human health, thus, it is quite important make out the contamination in the soil. There are a series of advantages in the hyperspectral remote sensing technology, such as the high spectral resolution, rapid response, non-destructive, etc., making it a well- suited in retrieving the soil's components. In this study, the impacts of the information redundancy in the spectral and spectral transformation on the inversion of Cd content in the soil are investigated. Further, based on the hyperspectral data before and after spectral transformation, the performance comparations of hyperspectral models are carried out in this paper, as well. By so doing, the Cd contents and the corresponding lab spectrum (350～2 500 nm) of 56 soil samples are measured by the ICP-MS and ASD Fieldspec4. Then, the reciprocal and logarithm changes are performed to weaken the impacts of the light variation and soil surface roughness on the experimental results. Due to the fact that there is much redundant information in the obtained data, the Principal Component Analysis (PCA) is carried out to reduce the dimensionality of the spectral bands in the data. After this processing, only 12 principal components are selected as the input variables of the model. Regarding the hyperspectral models, the Partial Least-Squares Regression (PLSR), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Random Forest (RF) are chosen to establish the relationship between the Cd content and PCA components. Finally, for evaluating the prediction capabilities of the regression models, three precision evaluation indexes are preferred to assess the accuracy of regression models in this study, they are the correlation coefficient (R^2), Root Mean Squared Error (RMSE) and Residual Predictive Deviation (RPD). Analysis results show that the cumulative contribution rate of 12 principal components of the original data after processed by the PCA can be up to 99.99%. Using principal components as the inputs, all four hyperspectral models show excellent performances in predicting the Cd content in the soil. The PCA-RF, in particular, has the most accurate prediction capability regardless of whether the spectral transformation is performed or not (whose R^2 before and after spectral transformation are 0.856 and 0.855, respectively, while the RPD under both conditions are 3.39). In conclusion, the PCA is used to reduce hyperspectral data's dimensionality, this processing can effectively reduce the redundancy of hyperspectral data and guarantee the predictive capability of hyperspectral models. Also, the principal component selected by the PCA method could be excellent input variables of the hyperspectral models. Further, the hyperspectral model based on the PCA-RF shows the most excellent performance for rapid detecting the Cd element in the soil within the study area and similar regions, which could be a new supplement for the inversion of heavy metals in the soil.

Keywords Cd content; Hyperspectral; PCA; Inversion model comparison

(Received May 26, 2020; accepted Aug. 31, 2020)

* Corresponding author