基于随机变异—Kennard—Stone和偏最小二乘法的土壤重金属镉含量反演

——以雄安新区西南部为例

www.
geojournals.cn/georev

黄照强,倪斌

中国冶金地质总局矿产资源研究院,北京, 101300

内容提要:土壤重金属污染对人类身体健康造成了严重威胁,作为快捷、高效、无损监测分析土壤重金属含量的方法之一,高光谱遥感反演土壤重金属含量的方法正逐步得到发展。本文以雄安新区西南部作为研究区,针对潜在生态风险最大且活性最强并易被植物吸收的重金属镉,开展高光谱反演研究。将采集来的426件土壤样品除杂、风干、过筛在实验室测得重金属含量,并用SVC便携式地面光谱仪测量样品350~2500 nm范围光谱。采用Savitzky—Golay卷积平滑方法进行光谱降噪平滑处理。由于粒径大小而不是化学成分差异可能会导致基线效应和漂移现象,为了增强光谱差异和光谱曲线形状,将数据进行标准正态变量变换(SNV)、一阶微分(FD)、二阶微分(SD)、多元散射校正(MSC)等数学变换,并分析变换后光谱与Cd含量的相关性。本文提出一种集成随机变异法—Kennard—Stone法—偏最小二乘回归的方法。针对变换后的光谱集和Cd含量集,第一步采用随机变异法—Kennard—Stone法将样本集分为70%训练集和30%验证集,使样本数随性质分布均匀并覆盖整个样本空间;第二步用偏最小二乘回归法结合交叉验证建立回归模型,用确定系数R^2、均方根误差*RMSE*、偏差比值*RPD*、误差范围比值*RER*等参数开展模型评价,如果没有达到预期效果,则回到第一步,迭代反复选择,直至达到最优效果。结果表明,适用的最优反演重金属镉含量的技术方法是采用FD变换后,不断迭代集成RM—KS样本选择和PLS偏最小二乘法回归建立的模型,其验证综合效果最好,建模主成分数为11个,确定系数R^2达到0.909,*RMSE*为0.604,*RPD*为2.696,*RER*达为15.516。成果可为类似区域快速、无损的土壤重金属Cd含量反演提供技术支撑。

关键词:土壤重金属污染;光谱变换;RM—KS样本集选择;偏最小二乘回归;高光谱

土地资源是人类赖以生存的最基本自然资源(朱永官等,2015)。随着全球经济的迅猛发展,工业化、农业现代化进程中所引起的工业污染和农田施肥农药造成了污染越来越严重,特别是重金属污染物越来越多的进入到农业土壤中。农业土壤中的重金属离子较易被农作物吸收并形成富集,会对人类身体健康产生威胁。因此,了解土壤中的重金属浓度将是非常重要的,以便有效进行农用地土壤风险筛查和管制。土壤重金属含量和农田土壤环境质量评价及治理一直是国内外研究的难点和热点问题,是环境科学的重要研究领域之一(Pkasads, 1995;陈玉娟等,2005;崔邢涛等,2014)。全球工业化导致大量潜在有毒化合物释放到生物圈中,其中包括微量元素,如镉、汞、铅、砷、锌和镍等,这些元素通常被称为重金属(Lasat et al., 1998)。在过去的50 a里,全球重金属释放量达到了22 kt Cd、939 kt Cu、783 kt Pb和1350 kt Zn(Singh et al., 2003)。2014年全国土壤污染状况调查公报结果显示,我国耕地土壤环境质量堪忧,耕地土壤点位超标率达19.4%,主要以镉、镍、铜、砷、汞、铅等重金属污染为主,其中镉(Cd)污染最为严重,点位超标率达7.0%(环境保护部和国土资源部,2014)。农田土壤Cd主要来源为大气沉降、畜禽粪便(有机肥)、化肥/农药、污灌及其他(磷肥、饼肥等)等,全国范围农田土壤中Cd的年输入通量约为0.725g/(hm^2·a),河北省农田年输入通量则达到2.153 g/(hm^2·a),为全国农田土壤中Cd年输入通量的2.97倍(陈世宝等,2019)。近年来,国内外众多学者对土壤重金属污染的状况和成因进行了研究,通过多目标地球化学调查,系统分析土壤重金属污染特征,采用内梅

注:本文为中国冶金地质总局科技创新项目(编号:CMGB202001),国家重点研发计划项目(编号:2016YFC0600210)和国家自然科学基金资助项目(编号:41272366)的成果。

收稿日期:2021-03-09;改回日期:2021-07-02;网络首发:2021-07-20;责任编辑:章雨旭。Doi:10.16509/j.georeview.2021.07.035

作者简介:黄照强,1973年生,博士、教授级高工,主要从事遥感地质、岩石光谱研究;Email:hzhaoq@126.com。

罗综合污染指数法、富集因子、地累积指数、潜在生态危害指数和健康风险评估模型对土壤重金属污染程度、潜在生态风险和健康风险进行评估(樊燕等,2008;崔邢涛等,2014;张素荣等,2020;郭志娟等,2021)。然而,传统的获取土壤重金属含量和空间分布的方法是采用网格化采样和费时的实验室分析,这具有破坏性,然后再采用地统计学插值得到结果(Leenaers et al.,1990; Kemper and Sommer, 2002)。反射光谱法是一种快速无损表征不同材料的方法(Ben-Dor et al.,1994, 2009; Clark, 1999)。随着高光谱遥感技术的发展,其快捷、高效、无损、易获取等特点将为土壤遥感监测提供有力支撑。可见光—短波红外分析方法获得的精度有助于快速土壤表征和遥感应用(Ben-Dor and Banin, 1994)。金属在近红外区域没有吸收特征,但可被检测,因为它们与有机物络合并与可检测的部分如氢氧化物、硫化物、碳酸盐或氧化物相关联,或被具有光谱吸收特征的黏土吸附(Malley and Williams, 1997)。Kemper 等基于土壤光谱与重金属间的相关关系,采用多元线性逐步回归和人工神经网络方法建立反演模型高精度的预测出矿区土壤中 As、Fe、Hg、Pb、S、Sb 等的含量(Kemper and Sommer,2002)。Kooistra 等采用偏最小二乘回归和反射光谱建立了莱茵河流域土壤中的镉、锌含量的预测模型(Kooistra et al., 2001)。Sun Wei 等用反射率偏态分布校正和偏最小二乘法估算矿区重金属 Cu 浓度(Sun Wei et al., 2019)。Wu Yunzhao 等利用反射光谱法研究了农业土壤中的汞污染,通过地面采集的光谱数据做反演模型研究,发现汞浓度分别与针铁矿和粘土在 496 nm 和 2210 nm 处的吸收特征相关(Wu Yunzhao et al., 2005)。Chen Wei 等利用紫外—可见/荧光光谱结合相关分析,研究了 SOM 与 Pb^{2+} 和 Cd^{2+} 相互作用过程中的相干结构变化(Chen Wei et al., 2020)。众多研究者都是利用多种光谱变换如多元散射校正、一阶微分、二阶微分、连续统去除等用于提取可识别的光谱敏感波段和/或光谱转换指数,然后采用各种线性/非线性的分析方法如逐步多元线性回归、偏最小二乘回归和神经网络等从这些变换后的光谱来建模预测土壤重金属含量(Liu Meiling et al.,2012; Wang Fenghe et al.,2018; Riedel et al., 2018; 邹滨等,2019)。逐步多元线性回归方法是选择几个特征波段进行建模且特征波段数不能大于样本数,而提取高光谱数据中的几个特征波段对于高光谱数据丰富的光谱信息是极其有限,其重要信息易丢失(Lee et al.,2004; 刘克等,2012),故在分析中需尽量保持高光谱数据的丰富内涵。偏最小二乘回归则特别适合当变量比观测样本值更多,以及变量中具有强共线相关性、噪声等问题的情况(Wold et al., 2001)。

因此本文以雄安新区西南部作为研究区,针对潜在生态风险最大且活性最强并易被植物吸收的重金属镉 Cd(张素荣等,2020),开展土壤中重金属 Cd 的含量和空间分布研究,通过采用多种光谱曲线变换方法,消除噪声等不利因素,提高变换后光谱特征与重金属镉 Cd 含量之间的相关性,然后提出一种集成随机变异—Kennard—Stone —偏最小二乘回归(Random Mutation — Kennard—Stone — Partial Least Squares,RM—KS—PLS)的方法,不断循环迭代选择合适样本训练集和测试集,使得样本能均匀覆盖整个样品空间,建立预测模型,并开展精度对比研究,探索适用的最优反演重金属镉含量的预测模型。

1 研究区与实验分析

1.1 研究区概况

研究区位于雄安新区西南部,行政区划归属于安新县,包括何桥乡老河头镇和芦庄乡,属暖温带季风型大陆性半湿润半干旱气候(朱江等,2017)。区内交通十分便利,分布有 331 省道、336 国道等交通要道。该区地势相对平坦,地形地貌为冲积洼地平原,地势低平,土壤母质主要为黄土母质和洪冲积物母质为主,主要为潮土、及少量褐土和沼泽土,唐河和清水河流经该区域。农业主产玉米、小麦等大宗农作物。区内冶炼、有色金属等工业产业发达。在一段时间内存在数以百计规模大小不等的多金属冶炼厂和无库尾矿渣堆,大气沉降、降雨淋溶、大风吹扬等导致大量有毒有害物质进入土壤耕作层,造成部分耕地土壤环境 Cd、Cu、Zn、Pb 等重金属元素不同程度超标(崔邢涛等,2014;张素荣等,2020;郭志娟等,2021)。

1.2 样品采集

采样点主要在研究区的农田,少数在建筑用地和尾矿渣堆(图 1)。在研究区采集 0~20 cm 耕作层土壤样品,为提高样品的代表性,在每个采样点 50 m 范围内采用梅花形布点法布设 5 处子采样点,去除石块、杂草、树根等杂质,然后采用四分法合成一个混合样,样品重量大于 1 kg,样品装在聚乙烯样品袋中,并获取 WGS84 坐标系统的 GPS 经纬度定

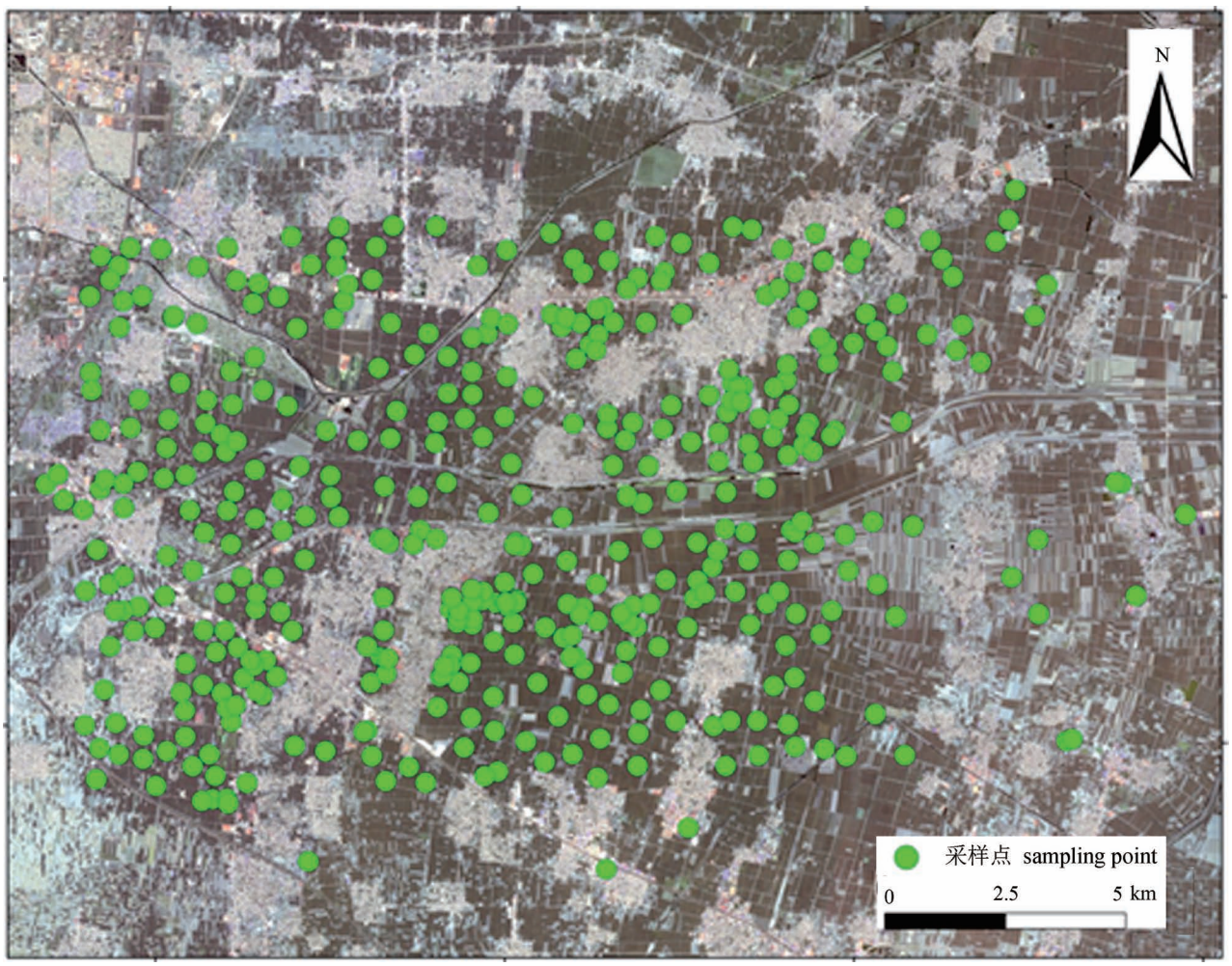


图 1 雄安新区西南部土壤重金属采样点分布图

Fig. 1 Distribution of soil heavy metal sampling sites in southwest of Xiong'an new district

位坐标。同时,实地调查和详细记录了采样点土壤性状和周边环境。

1.3 样品测试

样品的测试分析由中国冶金地质总局地球物理勘查院测试中心完成。样品带回测试中心后,经风干、研磨、过筛(200 目),然后采用电感耦合等离子质谱仪 NexIoN 350x 型测定重金属的含量。根据国家标准《硅酸盐岩石化学分析方法 第 30 部分:44 个元素量测定》(GB/T 14506.30-2010)测定了土壤重金属 Cd 的含量,中国土壤背景值 Cd 为 0.097 μg/g(中国环境监测总站,1990),冀中平原背景值 Cd 为 0.152 μg/g(张素荣等,2020),研究区的土壤 pH 平均值约为 8.0,偏碱性,因此其农田土壤污染风险筛选值 Cd 为 0.6 μg/g,管控值为 4 μg/g(生态环境部和国家市场监督管理总局,2018)。

1.4 样品光谱测量

众多研究学者在研究类似问题时,都是首先采用实验室测量和建模,在模型成熟后,再应用到实际航空高光谱和卫星遥感数据中。本文也是采取此思路开展模型研究。土壤样品光谱测量采用已经过风干、研磨和过筛(200 目)的样品,用美国 SVC Hr1024 便携式地面光谱仪进行测量,卤素灯作为光谱测量光源,光谱采样范围是 350 ~ 2500 nm(VNIR),共 1024 个波段,其中 350~1000 nm 范围的光谱分辨率为 3.5 nm,1000~1850 nm 范围的光谱分辨率为 9.5 nm,1850~2500 nm 范围的光谱分辨率为 6.5 nm。测量前用白色参考板辐射定标,每个样品测量 5 次,然后取其算术平均值。受环境影响,小于 400 nm 紫外光谱的噪声较大,因此去除小于 400 nm 的波段,只采用 400~2500 nm 范围的光谱,共 924 个波段。

1.5 光谱处理

预测质量在很大程度上取决于反射光谱的预处理步骤和方法(Kemper and Sommer,2002)。测量的光谱曲线常存在噪声,为了有效地消除光谱噪声,同时保持光谱表达的化学信息,采用 Savitzky—Golay 卷积平滑方法(S—G,5 个点平滑)进行光谱平滑处理,该方法是一种在时域内基于局域多项式最小二乘拟合的滤波方法(Savitzky and Golay, 1964)。另外由于粒径大小而不是化学成分差异可能会导致基线效应,因此为了独立于强度标度的数据比较,增强吸收特性和光谱曲线形状(Mackin et al. ,1991),将数据进行标准正态变量(SNV)、一阶微分(FD)、二阶微分(SD)、多元散射校正(MSC)等变换。

标准正态变量 (Standard Normal Variate,SNV)变换是一种面向行的变换,通过对每条光谱进行集中和缩放来消除因固体颗粒大小、表面散射及光程变化引起的光谱散射效应。SNV 使用来自该光谱的数据标准化每个光谱。数据 X 行中的每个值 x_k 根据以下公式进行转换:

$$x_{k,\mathrm{SNV}} = \frac{x_k - Mean(X)}{SDev(X)} \tag{1}$$

其中 $x_{k,\mathrm{SNV}}$ 表示转换后的值,$Mean(X)$ 表示数据 X 行的平均值,$SDev(X)$ 表示数据 X 行的标准差。

微分(Differential)是光谱应用中广泛使用的一种经典技术,用于校正光谱中的基线效应,以消除非化学效应并建立稳健的校正模型。一阶或二阶微分处理后,“隐藏”在光谱中的一些信息可能更容易被揭示。光谱的一阶微分(First Differential,FD)只是光谱曲线在每个点的斜率的一个度量。曲线的斜率不受光谱中纯加性基线偏移的影响,因此一阶导数能最小化反射光谱的背景干扰和基线漂移,是消除此类偏移的非常有效的方法。二阶微分(Second Differential,SD)是曲线斜率变化的量度。除了去除纯加性偏移外,它不受数据中可能存在的任何线性“倾斜”的影响,因此是从光谱中去除基线偏移和斜率的非常有效的方法,二阶微分有助于分辨附近的峰和锐化光谱特征,但对噪声更敏感。

一阶微分公式为:

$$y'(i) = \frac{y(i+1) - y(i-1)}{2\Delta y} \tag{2}$$

$y'(i)$ 指波段 i 处的一阶微分,$y(i+1)$ 为下一个采样波段的反射率,$y(i-1)$ 是上一个采样波段的反射率,Δy 是采样间隔。

二阶微分公式为:

$$y''(i) = \frac{y'(i+1) - y'(i-1)}{2\Delta y} \tag{3}$$

$y''(i)$ 指波段 i 处的二阶微分,$y'(i+1)$ 为下一个采样波段的一阶微分,$y'(i-1)$ 是上一个采样波段的一阶微分,Δy 是采样间隔。

多元散射校正(Multiplicative Scatter Correction,MSC)是一种用于补偿光谱数据中加性和/或乘性效应的变换方法,用来消除样本间的基线平移和漂移现象,增强光谱差异性。首先计算所有样本的平均光谱,然后将平均光谱作为标准光谱,进行一元线性回归求得各光谱相对于标准光谱的线性平移量(回归常数)和倾斜偏移量(回归系数),最后在每个样品原始光谱中减去线性平移量同时除以回归系数,得到修正后光谱,提高光谱信噪比。多元散射校正公式:

$$A_{i(\mathrm{MSC})} = \frac{A_i - b_i}{m_i} \tag{4}$$

其中 A_i 是样品光谱,m_i 和 b_i 分别表示各样品光谱与平均光谱进行一元线性回归得到的相对回归系数和平移量。

1.6 模型校准和验证

在建模阶段,需要将化学定量数据与光谱曲线组合来建立化学定量模型。在模型构建前,样本被分成校准集(训练集,70%)和验证集(测试集,30%)。利用随机变异法和 Kennard—Stone 法(RM—KS)进行训练集和验证集样本的挑选,可明显改善样本数随性质分布的均匀性并覆盖整个样本空间,并在生物信息领域得到应用(Kennard and Stone,1969; Morais et al. ,2019b)。KS 算法是基于欧几里德距离的计算,首先选取与所有其它样本距离最大的样本,接着选取离所选样本尽可能远的样本,直至达到所选样本数。然后在 KS 结果中使用随机变异因子 10%,将训练集中的一些样本转移到测试集中,将测试集中的一些样本转移到训练集中,其中随机变异因子设为 10%是一个常用阈值,用于在随机度和模型收敛性之间保持平衡(Morais et al. ,2019a,b)。

偏最小二乘回归(Partial Least Squares Regression ,PLSR)(Wold et al. ,2001)是一种通过线性多元模型将两个数据矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 关联起来的方法,能同时对 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 进行建模,以找到 $\boldsymbol{X}$ 中最能预测 $\boldsymbol{Y}$ 中潜在变量的隐藏变量,特别适合当预测矩阵比观测值有更多 $\boldsymbol{X}$ 的变量,以及 $\boldsymbol{X}$ 的值中具有强共线相关性、噪声。其核心思想是在 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 中提取

出成分 $t1$ 和 $u1$,满足 $t1$ 和 $u1$ 应尽可能大地携带他们各自数据表中的变异信息,以及 $t1$ 和 $u1$ 与的相关程度能够达到最大,则此成分确定;不断反复迭代,选出相应的成分,直至达到满意的精度。第一个潜在变量可以通过最大化预测变量和响应变量的协方差 $cov(t_1,u_1)$ 来获得,公式如下:

$$\max[cov(t_1,u_1)]=\max_{||w_1||=1,||c_1||=1} cov(Xw_1,Yc_1) \tag{5}$$

其中载荷权重 w_1,c_1 分别是对称矩阵 $X'YY'X$,$Y'XX'Y$ 的特征向量,X'、Y' 分别是 X、Y 的转置矩阵。$||w_1||=1$ 表示 w_1 的 1 范数,也就是绝对值的和是 1。第一个潜变量定义为:$t_1=Xw_1$。利用最小二乘法对 X 和 Y 进行压缩。根据主成分回归思想可以把 X 和 Y 分别对它们的主成分 t_1 和 u_1 进行回归建模:

$$\begin{cases} X=t_1p'_1+X^* \\ Y=u_1q'_1+y_1^*=t_1r'_1+y^* \end{cases} \tag{6}$$

其中 p'_1、q'_1、r'_1 是回归系数向量,X^*、y_1^*、y^* 是残差矩阵。若回归方程精度达到要求,则结束,否则利用由样本计算的相关矩阵与由因素模型推出的相关矩阵之差即残差矩阵 X^* 和 y^* 计算下一个主成分,重新建立回归方程,直至满足精度为止。

而本研究中作为自变量的光谱曲线具有几百乃至上千个波段,往往超过了观测样本数,因此正可以利用偏最小二乘回归法的适合于当预测矩阵比观测值有更多 X 变量的特点,以变换后光谱训练集与重金属镉 Cd 含量之间建立校准模型,用验证集进行验证以便获取最佳预测模型。本研究集成一种 RM—KS—PLS 方法,用 Matlab 软件开发功能模块,进行不断循环迭代以便获取最佳的预测模型。模型建立流程见图 2。

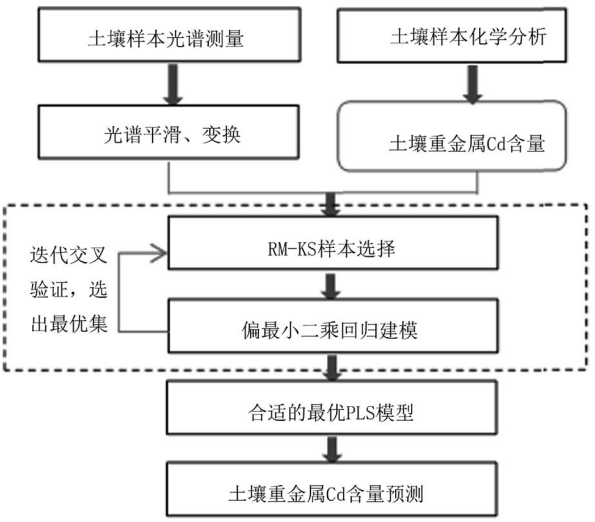


图 2 模型建立流程

Fig. 2 Flow of Modeling

2　结果与讨论

2.1　土壤样本重金属 Cd 含量分析

通过对研究区土壤样本初步分析,剔除异常值数据,则选择了 426 个样本划分为模型训练集和测试集,按照 KS 算法+随机变异因子法,得到训练集样本 297 个和测试集样本 129 个。通过上述的土壤样品采集和实验室分析测试得到了研究区土壤样本重金属 Cd 含量值,其基本特征统计见表 1,重金属 Cd 含量平均值为 1.31 μg/g 高于冀中平原背景值和土壤污染风险筛选值,同时根据单因子污染指数评价法(张素荣等,2020)对重金属 Cd 累计污染程度进行评价,计算公式为:

$$p_i=\frac{C_i}{S_i} \tag{5}$$

其中 C_i 是污染物的实测值,S_i 指重金属 Cd 的风险管理目标值即风险筛查值,p_i 指污染指数,$p_i>1$ 表明土壤受到污染,越大则重金属累积污染程度越大,其中污染指数在 1~3 为轻度污染,3~6.7 为中度污染,大于 6.7 为重度污染。样本分析结果显示 198 个点无污染,轻度污染有 158 个点,中度有 35 个点,重度污染有 35 个点(图 3)。故表明研究区存在土壤重金属 Cd 污染。而且数据涵盖了从清洁到重度污染,为建模提供了可靠的基础数据和验证集。

表 1 雄安新区西南部土壤重金属 Cd 含量统计表

Table 1 Heavy metal Cd concentration (Range, Average) of soil in southwestern Xiongan area

土壤重金属	最大值(μg/g)	最小值(μg/g)	均值(μg/g)	标准差(μg/g)	变异系数(%)	中国土壤背景值(μg/g)	冀中平原背景值(μg/g)	土壤污染风险筛选值(μg/g)
Cd	12.48	0.01	1.31	2.0	1.53	0.097	0.152	0.6

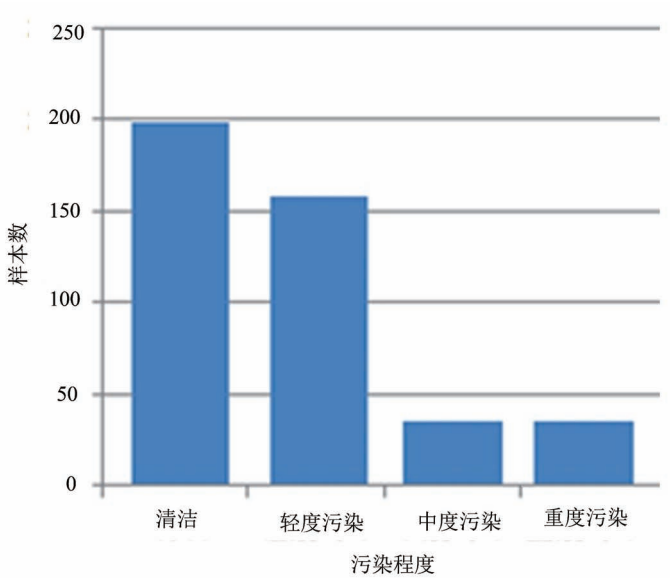


图 3 土壤样品 Cd 污染情况分类图

Fig. 3 Soil Samples Cd Contamination classification

2.2 土壤光谱分析

晶体场中金属离子(如 Fe^{2+}、Fe^{3+})的电子迁移分子团 H—O—H、OH^-、CO_3^{2-} 及 Si—O 键等的振动在可见—短波红外—中红外光谱区域形成一系列的吸收特征,这些诊断性的吸收特征构成光谱识别岩石矿物的理论基础(Hunt,1977,1979;Clark,1999)。土壤样品的 SVC 地面光谱仪测量的光谱曲线(图 4)显示样品的光谱形态特征基本一致,只是由于土壤组分的差异导致反射率的差异。光谱的吸收特征往往反映的是 Fe^{2+}、Fe^{3+} 等金属离子的电子跃迁和 OH^-、金属—OH、H_2O 等离子团振动组合产生的倍频和谐频。700 和 1000 nm 附近表现 Fe^{2+} 的吸收带,450、860 和 900 nm 附近表现 Fe^{3+} 的吸收带,1400 nm 附近表现 OH^- 吸收带,2200 和 2300 nm 附近表现金属氢氧化物(Al—OH、Mg—OH、Fe—OH 等)吸收带,2300 nm 附近表现金属碳酸盐(—CO_3^{2-})吸收带,1900 nm 附近表现 H_2O 吸收带,有机质在 1150

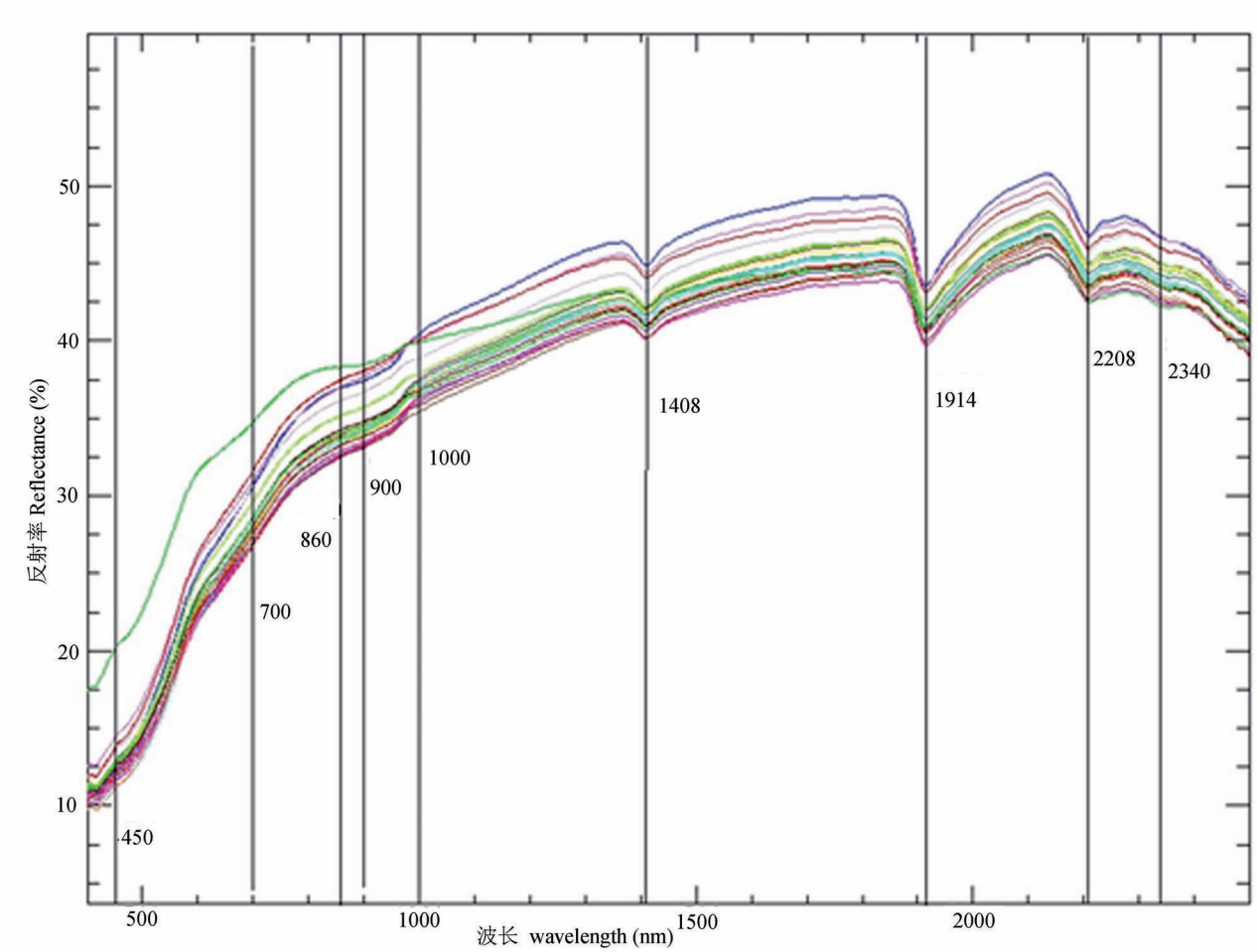


图 4 土壤样品的 SVC 光谱曲线图

Fig. 4 SVC Spectral Curve of Soil Samples

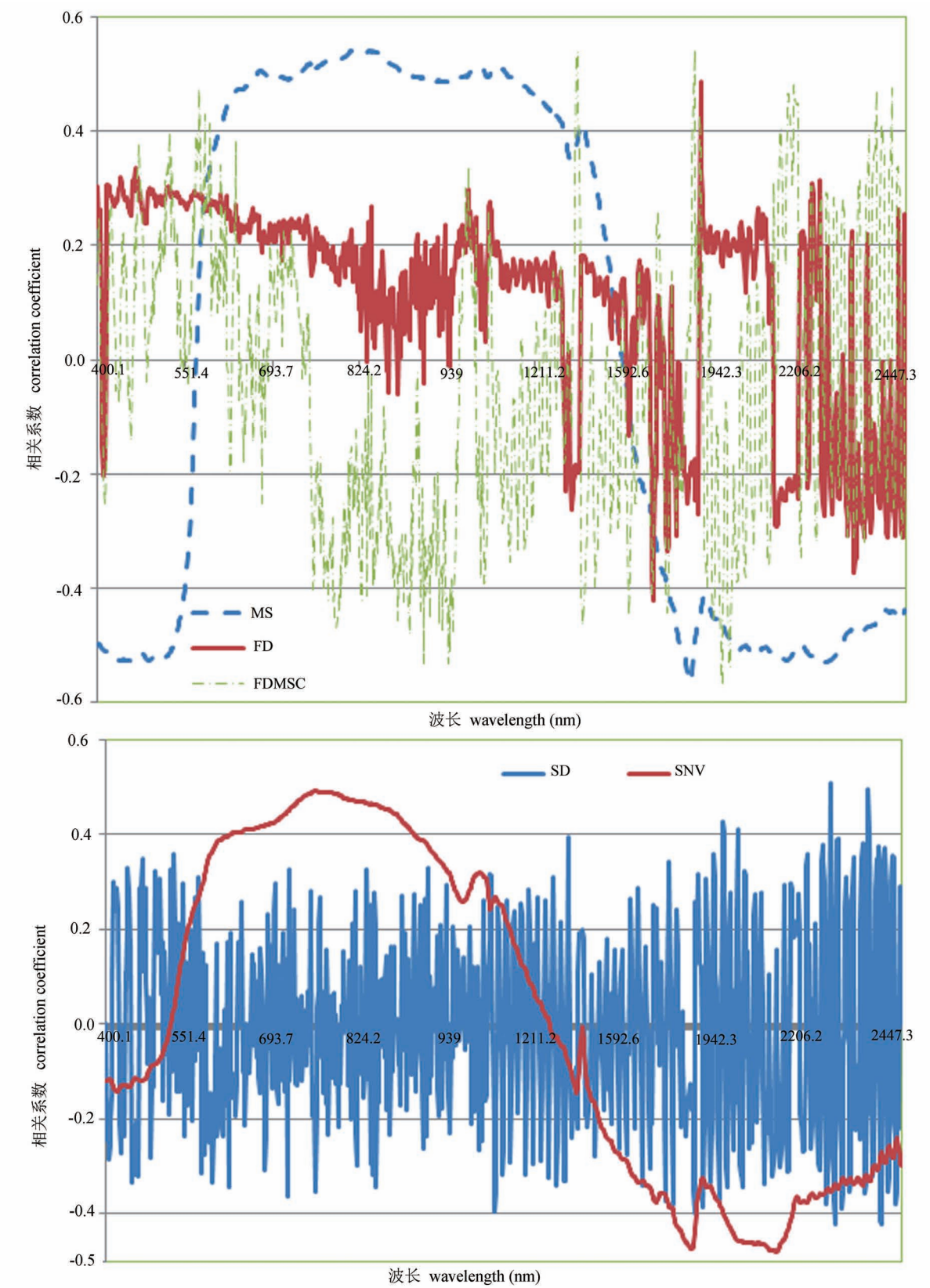


图 5 FD、MSC、FD+MSC、SD、SNV 变换光谱波长与 Cd 含量相关关系

Fig. 5 Correlation Coefficient between wavelength of FD, MSC, FD+MSC, SD, SNV transformed Spectrum and Cd concentration

和 1700 nm 附近有吸收(Clark,1999)。重金属元素一般吸附或赋存于有机质、铁锰氧化物、硅酸盐以及碳酸盐等具有明显光谱特征的主要土壤组分中,重金属与这些组分之间的相关关系是利用反射光谱对其进行定量分析的机理。

SVC 光谱反射率数据经过 S—G 平滑后,采用 FD、MSC、FD+MSC、SD、SNV 五种方法进行光谱变换,得到的结果与 Cd 含量进行相关分析。FD 变换后光谱与 Cd 含量的相关系数 R 较大且显著相关的有 R 为 0.48 在 1911 nm 波长(水吸收),-0.42 在 1721 nm 波长(有机质吸收),-0.37 在 2365 nm 波长(碳酸盐吸收),0.34 在 466 nm 波长(铁锰氧化物吸收);MSC 和 FD+MSC 变换后光谱与 Cd 含量相关系数绝对值较大的且显著相关在 0.5~0.6 之间,在 570(氧化铁吸收特征)、2200 和 2300 nm(黏土矿物吸收特征)高度相关;SNV 和 SD 变换后光谱与 Cd 含量相关系数绝对值较大且显著相关的在 0.4~0.5 之间(图 5)。显著相关指示了变换光谱与 Cd 含量之间关系建模过程中,这些光谱波长具有很大的贡献。

2.3 PLS 模型预测

各种光谱变换后的曲线,采用 PLS 建模和验证得到的结果见表 2 所示。将预测值与化学分析的参考值进行相关分析,常采用拟合优度(Goodness of Fit)来表示回归直线对观测值的拟合程度。度量拟合优度的统计量是确定系数 R^2,其取值范围是[0,1],越接近 1,说明回归直线对观测值的拟合程度越好,反之则越差。微量元素在自然背景环境中含量很低,多呈对数正态分布,而在人为污染情况下可能大量富集,造成含量的概率分布向高浓度方向偏斜(Zhang and Selinus,1998)。因此,需要计算另外的参数值共同来更好的评价模型精度,即均方根误差 *RMSE*(越小越好)、偏差比值 *RPD*(Ratio of percent deviation,为参考值标准差与 *RMSE* 的比值,*RPD*>1.4 效果较好)、误差范围比值 *RER*(Ratio of error range,为预测值的取值范围与 *RMSE* 的比值,*RER*>10 效果较好)(Malley and Williams,1997)等。*RPD*<1.0 表示非常差的模型,不建议使用它们;*RPD* 在 1.0 和 1.4 之间表示差的模型,其中只有高值和低值可区分;*RPD* 在 1.4 和 1.8 之间表示恰当的能用于评估和相关分析的模型/预测;*RPD* 值在 1.8 和 2.0 之间表示模型良好,可以进行定量预测;*RPD* 值>2.0 之间表示非常好的定量模型/预测(Viscarra et al.,2006)。

5 种光谱变换后建模和验证的结果见表 2,F 检验的 P 值都小于 0.01,具有显著相关性。光谱 FD 变换后,建模主成分数为 11 个,验证集(Validation)预测值与验证集化学实测含量之间的相关关系确定系数 R^2 达到 0.909,为这几种光谱变换方法中最大的,*RMSE* 最小为 0.604,*RPD* 最大为 2.696,*RER* 大于 10 为 15.516(表 2,图 6),表明模型验证很好,但有些高值被低估了。而光谱 FD+MSC 变换后,建模主成分数为 11 个,验证集预测值与化学实测值之间的确定系数 R^2 为 0.858,*RMSE* 为 0.622,*RPD* 为 2.298,*RER* 大于 10 为 21.177,模型验证良好程度次之。SD 变换后,验证集预测值与化学实测值之间的确定系数 R^2 为 0.831,*RMSE* 为 0.593,*RPD* 为 2.080,*RER* 为 12.350;MSC 变换后,验证集预测值与化学实测值之间的确定系数 R^2 为 0.775,*RMSE* 为 1.238,*RPD* 为 1.310,*RER* 为 17.922,有些值被高估了;这两类总体效果良好。如果剔除一些异常值,结果会更好。光谱经过 SNV 变换后,验证集预测值与验证集化学实测含量之间的确定系数 R^2 小于 0.680,效果较差些。五种光谱变换后的 PLS 校准集(训练集)预测值与参考值之间的一致性都较好,FD 中的部分高值被低估,部分低水平样本的预测值为负值;而 FD、MSC、SD、FD+MSC 变换后 PLS 验证集预测值与参考值之间的一致性较好,SNV 一致性相对偏低(图 6)。

总体效果是 FD 变换后,采用 RM—KS 方法进行样本集选择和 PLS 方法迭代循环建模得到的预测效果最好,效果排在其次的是 FD+MSC 变换、SD 和 MSC 变换,最后是 SNV 变换的效果。

表 2 RM-KS-PLS 建模和验证的结果

Table 2 Calibration and Validation Results For RM-KS-PLS

光谱变换方法	主成分数	Calibration				Validation			
		R^2	*RMSE*	*RPD*	*RER*	R^2	*RMSE*	*RPD*	*RER*
FD	11	0.807	0.933	2.278	12.391	0.909	0.604	2.696	15.516
SD	5	0.653	1.307	1.698	7.415	0.831	0.593	2.080	12.350
FD+MSC	11	0.798	0.979	2.225	12.567	0.858	0.622	2.298	21.177
MSC	23	0.761	1.041	2.045	12.883	0.775	1.238	1.310	17.922
SNV	20	0.737	1.079	1.950	11.939	0.680	1.114	1.549	12.566

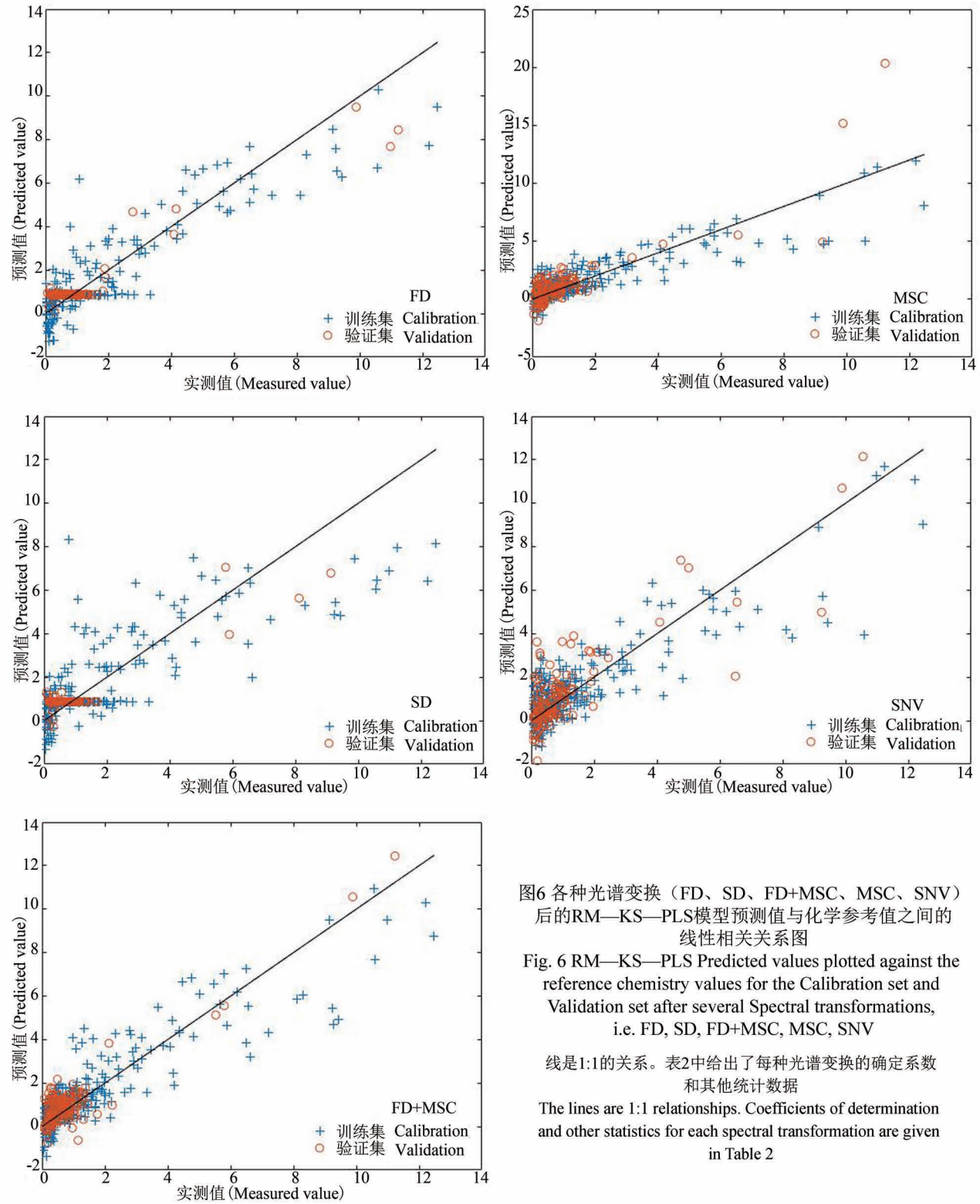


图6 各种光谱变换（FD、SD、FD+MSC、MSC、SNV）后的RM—KS—PLS模型预测值与化学参考值之间的线性相关关系图

Fig. 6 RM—KS—PLS Predicted values plotted against the reference chemistry values for the Calibration set and Validation set after several Spectral transformations, i.e. FD, SD, FD+MSC, MSC, SNV

线是1:1的关系。表2中给出了每种光谱变换的确定系数和其他统计数据

The lines are 1:1 relationships. Coefficients of determination and other statistics for each spectral transformation are given in Table 2

3　结论

通过对雄安新区西南部受到不同程度重金属污染的区域采集土壤样品,开展重金属污染监测,分析光谱与 Cd 含量的相关关系,采用集成的随机变异法—Kennard—Stone 法（RM—KS）—偏最小二乘（PLS）方法不断迭代交叉验证优选训练集和验证集以使样本分布均匀并覆盖整个样本空间,建立 PLS 模型实现土壤重金属含量反演,探索适用的最优反演重金属镉含量的预测模型。得出结论如下:

(1)样品 SVC 光谱反射率数据经过 S—G 平滑后,采用 FD、MSC、FD+MSC、SD、SNV 五种方法进行光谱变换,消除因粒径大小导致基线效应,增强了光谱响应吸收特征和光谱曲线形状。

(2)集成随机变异法—Kennard—Stone 法(RM—KS)—PLS 方法,不断迭代交叉验证,优选训练集(70%)和验证集(30%),改善样本选取的均匀性和覆盖全域,有利于提高模型效果;通过确定系数 R^2、均方根误差 *RMSE*、偏差比值 *RPD*、误差范围比值 *RER* 等参数比较,表明采用 FD 变换后,RM—KS—PLS 方法建模效果最好,为将来利用光谱方法快速、无损实现土壤中重金属 Cd 含量的反演提供支撑。

参 考 文 献 / References

(The literature whose publishing year followed by a "&" is in Chinese with English abstract; The literature whose publishing year followed by a "#" is in Chinese without English abstract)

陈世宝,王萌,李杉杉,郑涵,雷小琴,孙晓艺,王立夫. 2019. 中国农田土壤重金属污染防治现状与问题思考. 地学前缘,26(6):35~41.

陈玉娟,温琰茂,柴世伟. 2005. 珠江三角洲农业土壤重金属含量特征研究. 环境科学研究,18(3):75~77.

崔邢涛,秦振宇,栾文楼,宋泽峰. 2014. 河北省保定市平原区土壤重金属污染及潜在生态危害评价. 现代地质,28(3):523~530.

樊燕 , 刘洪斌, 武伟. 2008. 土壤重金属污染现状评价及其合理采样数的研究. 土壤通报, 39(2):369~374.

郭志娟,周亚龙,王乔林,王成文,宋云涛,刘飞,杨峥,孔牧. 2021. 雄安新区土壤重金属污染特征及健康风险. 中国环境科学,41(1):431~441.

环境保护部和国土资源部. 2014. 全国土壤污染状况调查公报. 中国环保产业, 5:10~11.

刘克, 赵文吉, 郭逍宇, 王翊虹, 孙永华, 苗茜. 2012. 基于地面实测光谱的湿地植物全氮含量估算研究. 光谱学与光谱分析, 32(2): 465~471.

生态环境部, 国家市场监督管理总局. 2018. 土壤环境质量农用地土壤污染风险管控标准(GB15618-2018). 北京:中国环境科学出版社: 1~7.

张素荣,王昌宇,刘继红,邢怡,刘庆学,杨俊泉,贺福清. 2020. 雄安新区西南部土壤重金属污染特征及生态风险评价. 地学前缘. https://doi. org/10. 13745/j. esf. sf. 2020. 7. 1

中国环境监测总站. 1990. 中国土壤元素背景值. 北京:中国环境科学出版社.

朱江,马柱国,严中伟,袁星, 符淙斌. 2017. 气候变化背景下雄安新区发展中面临的问题. 中国科学院院刊, 32(11):1231~1236.

朱永官,李刚,张甘霖,傅伯杰. 2015. 土壤安全:从地球关键带到生态系统服务. 地理学报, 70(12):1859~1869.

邹滨,涂宇龙,姜晓璐,陶超,周茉,熊立伟. 2019. 土壤 Cd 含量实验室与野外 DS 光谱联合反演. 光谱学与光谱分析, 39(10):3223~3231.

Ben-Dor E, Banin A. 1994. Visible and near-infrared (0.4~ 1.1 μm) analysis of arid and semiarid soils. Remote Sensing of Environment, 48(3): 261~274.

Ben-Dor E, Chabrillat S, Demattê J A M, Taylor G R, Hill J, Whiting M L, Sommer S. 2009. Using imaging spectroscopy to study soil properties. Remote Sensing of Environment, 113(S1): S38~S55.

Chen Shibao, Wang Meng, Li Shanshan, Zheng Han, Lei Xiaoqin, Sun Xiaoyi, Wang Lifu. 2019&. Current status and discussion on farmland heavy metal pollution prevention in China. Earth Science Frontiers, 26(6): 35~41.

Chen Wei, Peng Li, Hu Keren, Zhang Zhang, Peng Changhong, Teng Chunying, Zhou Kanggen. 2020. Spectroscopic response of soil organic matter in mining area to Pb/Cd heavy metal interaction: A mirror of coherent structural variation. Journal of Hazardous Materials, 393: 122425.

Chen Yujuan, Wen Yanmao, Chai Shiwei. 2005&. The heavy metal content character of agricultural soil in the Pearl River Delta. Research of Environmental Sciences, 18(3):75~77.

China National Environmental Monitoring Centre. 1990#. Background values of soil elements in China. Beijing: China Environmental Press.

Clark R N. 1999. Chapter 1: Spectroscopy of rocks and minerals, and principles of spectroscopy. In: Rencz A N. ed. Manual of Remote Sensing, Volume 3, Remote Sensing for the Earth Sciences. New York: John Wiley and Sons: 3~58.

Cui Xingtao, Qin Zhenyu, Luan Wenlou, Song Zefeng. 2014&. Assessment of the heavy metal pollution and the potential ecological hazard in soil of plain area of Baoding City of Hebei Province. Geoscience, 28(3):523~530.

Fan Yan, Liu Hongbin, Wu Wei. 2008&. Evaluation to soil heavy metals' pollution and the study of reasonable sampling number. Chinese Journal of Soil Science, 39(2):369~374.

Guo Zhijuan, Zhou Yalong, Wang Qiaolin, Wang Chengwen, Song Yuntao, Liu Fei, Yang Zheng, Kong Mu. 2021&. Characteristics of soil heavy metal pollution and health risk in Xiong'an New District. China Environmental Science,41(1):431~441.

Hunt G R. 1977. Spectral signatures of particulate minerals in the visible and near-infrared. Geophysics, 42:273~288.

Hunt G R. 1979. Near-infrared (1.3~2.4 μm) spectral of alteration minerals-potential for use in remote sensing. Geophysics, 44(12): 1974~1986.

Kemper T, Sommer S. 2002. Estimate of heavy metal contamination in soil after a mining accident using reflectance spectroscopy. Environmental Science and Technology,36(12):2742~2747.

Kennard R W, Stone L A. 1969. Computer aided design of experiments. Technometrics, 11:137~148.

Kooistra L, Wehrens R, Wleuven R S E, Buydens L M C. 2001. Possibilities of visible-near-infrared spectroscopy for the assessment of soil contamination in river floodplains. Analytica Chimica Acta, 446(1~2):97~105

Lasat M M, Fuhrnmann M, Ebbs S D, Cornish J E, Kochian L V. 1998. Phytoremediation of a radiocesium contaminated soil: evaluation of cesium-137. J. Environ. Qual., 27:165~169.

Lee K S, Cohen W B, Kennedy R E, Maiersperger T K, Gower S T. 2004. Hyperspectral versus multispectral data for estimating leaf area index in four different biomes. Remote Sens. Environ., 91(3~4), 508~520.

Leenaers H, Okx J P, Burrough P A. 1990. Employing elevation data for efficient mapping of soil pollution on floodplains. Soil Use and Management, 6(3):105~114.

Liu Ke, Zhao Wenji, Guo Xiaoyu, Wang Yihong, Sun Yonghua, Miao

Qian. 2012&. Estimating total nitrogen content in wetland vegetation based on measured reflectance spectra. Spectroscopy and Spectral analysis, 32(2): 465~471.

Liu Meiling, Liu Xiangnan, Li Jonathan, Li Ting. 2012. Estimating regional heavy metal concentrations in rice by scaling up a field-scale heavy metal assessment model. International Journal of Applied Earth Observation and Geoinformation, 19:12~23.

Mackin S, Drake N, Settle J, Briggs S. 1991. Curve shape matching and end-member selection and mixture modelling of AVIRIS and GER data for mapping surface mineralogy and vegetation communities. Proceedings of the 3rd AVIRIS Workshop, 20~21 May 1991, JPL Publication 91~28, (Pasadena, CA: Jet Propulsion Laboratory): 158~162.

Malley D F, Williams P C. 1997. Use of near-infrared reflectance spectroscopy in prediction of heavy metals in freshwater sediment by their association with organic matter. Environmental Science & Technology, 31(12):3461~3467. doi:10.1021/es970214p.

Ministry of Ecology and Environment of the People's Republic of China, State Administration for Market Regulation. 2018 #. Soil Environmental Quality Risk Control Standard for Soil Contamination of Agricultural Land (GB15618-2018). Beijing: China Environmental Press: 1~7.

Ministry of Environmental Protection and Ministry of Land and Resources. 2014#. Bulletin of national soil pollution survey. China Environmental Protection Industry, (5): 10~11.

Morais C L M, Paraskevaidi M, Cui Li, Fullwood N J, Isabelle M, Lima K M G, Martin-Hirsch P L, Sreedhar H, Trevisan J, Walsh M J, Zhang Dayi, Zhu Yongguan, Martin F L. 2019a. Standardization of complex biologically derived spectrochemical datasets. Nat. Protoc., 14:1546~1577.

Morais C L M, Santos M C D, Lima K M G, Martin F L. 2019b. Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach. Bioinformatics, 35(24): 5257~5263; DOI:10.1093/bioinformatics/btz421

Pkasads Y. 1995. Characteristics of heavy metals in contaminated soil. J. Environ., 121(4):276~286.

Riedel F, Denk M, Müller I, Barth N, Gläßer C. 2018. Prediction of soil parameters using the spectral range between 350 and 15,000 nm: A case study based on the Permanent Soil Monitoring Program in Saxony, Germany. Geoderma, 315:188~198.

Savitzky A, Golay M J E. 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem., 36(8):1627~1639.

Singh O V, Labana S, Pandey G, Budhiraja R, Jain R K. 2003. Phytoremediation: an overview of metallic ion decontamination from soil. Applied Microbiology and Biotechnology, 61(5~6): 405~412.

Sun Wei, Skidmore A K, Wang Tiejun, Zhang Xia. 2019. Heavy metal pollution at mine sites estimated from reflectance spectroscopy following correction for skewed data. Environmental Pollution, 252: 1117~1124.

Viscarra R A, Mcglynn R N, Mcbratney A B. 2006. Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. Geoderma, 137:70~82.

Wang Fenghe, Gao J, Zha Yong. 2018. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. ISPRS Journal of Photogrammetry and Remote Sensing, 136:73~84.

Wold S, Sjostrom M, Eriksson L. 2001. PLS-regression: A basic tool of chemometrics. Chemom. Intell. Lab. Syst., 58(2):109~130.

Wu Yunzhao, Chen Jun, Ji Junfeng, Tian Qingjiu, Wu Xinmin. 2005. Feasibility of reflectance spectroscopy for the assessment of soil mercury contamination. Environmental Science Technology, 39:873~878.

Zhang C S, Selinus O. 1998. Statistics and GIS in environmental geochemistry-some problems and solutions. Journal of Geochemical Exploration, 64(1~3):339~354.

Zhang Surong, Wang Changyu, Liu Jihong, Xing Yi, Liu Qingxue, Yang Junquan, He Fuqing. 2020&. Assessment of heavy metal pollution and ecological risk in soils of the southwestern part of the Xiongan New Area. Earth Science Frontiers, https://doi.org/10.13745/j.esf.sf.2020.7.1

Zhu Jiang, Ma Zhuguo, Yan Zhongwei, Yuan Xing, Fu Congbin. 2017&. Problems faced by construction of Xiongan New Area under climate change. Bulletin of Chinese Academy of Sciences, 32(11): 1231~1236.

Zhu Yongguan, Li Gang, Zhang Ganlin, Fu Bojie. 2015&. Soil security: From Earth's critical zone to ecosystem services. Acta Geographic Sinica, 70(12): 1859~1869.

Zou Bin, Tu Yulong, Jiang Xiaolu, Tao Chao, Zhou Mo, Xiong Liwei. 2019&. Estimation of Cd content in soil using combined laboratory and field DS Spectroscopy. Spectroscopy and Spectral analysis, 39(10):3223~3231.

Retrieval of soil heavy metal Cadmium content based on Random Mutation, Kennard—Stone and partial least squares method: A case study of southwest of Xiong'an New District

HUANG Zhaoqiang, NI Bin

Institute of Mineral Resources, China Metallurgical Geology Bureau, Beijing, 101300

Abstract: Soil heavy metal pollution poses a serious threat to human health. As one of the fast, efficient and non-destructive methods for monitoring and analyzing soil heavy metal concentration data, hyperspectral remote sensing method is gradually developing. In this paper, the hyperspectral inversion study in the southwest of Xiong'an New District was carried out for cadmium, which has the greatest potential ecological risk and the strongest

activity and is easy to be absorbed by plants. 426 soil samples were collected for impurity removal, air drying and sieving. The heavy metal content was measured in the laboratory, and the spectrum of 350 ~ 2500 nm was measured by SVC portable ground spectrometer. Savitzky—Golay convolution smoothing method is used for spectral denoising and smoothing. Because the baseline effect and drift phenomenon may be caused by the difference of particle size rather than chemical composition, in order to enhance the spectral difference and the shape of spectral curve, the data were transformed by several mathematical transformations such as Standard Normal Variable Transformation (SNV), First-Order Differential (FD), Second-Order Differential (SD), Multiplicative Scattering Correction (MSC). And the correlations between the transformed spectrum and Cd content were analyzed. In this paper, a method of integrating Random Mutation (RM) — Kennard—Stone (KS) —Partial Least Squares Regression (PLSR) is proposed. For the transformed samples spectrum and Cd content set, the first step is to divide the samples set into 70% training set and 30% verification set by using RM — KS method, so that the number of samples is evenly distributed with the properties and covers the whole sample space; In the second step, the PLSR method combined with cross validation is used to establish the regression model, and the parameters such as determination coefficient R^2, root mean square error (*RMSE*), ratio of percent deviation (*RPD*), ratio of error range (*RER*) are used to carry out the model evaluation. If the expected situation is not achieved, return to the first step and iterate until the optimal effect is achieved. The results show that the optimal method for retrieving Cadmium concentration is the model established by using FD transformation, iterative integration of RM—KS sample selection and PLS regression. The comprehensive verification effect is the best, with 11 principal components, and the parameters R^2 of 0.909, *RMSE* of 0.604, *RPD* of 2.696 and *RER* of 15.516. The research results can provide technical support for rapid and non-destructive retrieval of soil heavy metal Cd content in similar areas.

Keywords: soil heavy metals contamination; Spectrum transformation; RM—KS Selection of sample sets; Partial Least Squares Regression; hyperspectral

Acknowledgements: This study was partially supported by the Science and Technology Innovation Program of China Metallurgical Geology Bureau (No. CMGB202001), the National Key R&D Project (No. 2016YFC0600210), and the National Natural Science Foundation of China (No. 41272366)

First author: HUANG Zhaoqiang, male, born in 1973, doctor, professor, focus on Remote Sensing of Geology and Spectroscopy of Petrology; Email: hzhaoq@126.com

Manuscript received on: 2021-03-09; Accepted on: 2021-07-02; Network published on: 2021-07-20

Doi: 10.16509/j.georeview.2021.07.035 **Edited by**: ZHANG Yuxu