土壤重金属 Cu 含量遥感反演的波段选择与最佳光谱分辨率研究

黄长平[1,2]，刘 波[1,2]，张 霞[1]，童庆禧[1,3]

（1. 中国科学院遥感应用研究所遥感科学国家重点实验室，北京 100101；
2. 中国科学院研究生院，北京 100049；3. 北京大学遥感与地理信息系统研究所，北京 100871）

摘要：高光谱数据以其高光谱分辨率和多而连续的光谱波段为预测土壤重金属污染提供了有力工具，但波段选择方法与光谱分辨率的影响不容忽视。利用实验室测定的 181 个土壤光谱样本数据，利用逐步回归法进行土壤 Cu 含量反演的波段选择，进而利用偏最小二乘方回归 PLSR 方法建模，分析了波段数对 Cu 含量反演的影响；此外，采用高斯响应函数重采样方法，探讨了光谱分辨率降低对反演精度的影响。实验表明，预测重金属元素 Cu 含量的最佳波段数为 10 个，模型可决系数 $R^2=0.7523$，拟合均方根误差 RMSE=0.4699；预测 Cu 含量的最佳光谱采样间隔为 32 nm，$R^2=0.7028$，RMSE=0.5147。该结果可能为将来设计低廉实用的高光谱卫星传感器提供指标论证，为模拟卫星传感器波段预测土壤重金属含量提供理论依据。

关 键 词：Cu 含量遥感预测；高光谱数据；光谱重采样；PLSR；波段选择

中图分类号：TP 79 **文献标志码**：A **文章编号**：1004-0323(2010)03-0353-05

1 引 言

伴随工业化的快速进程，土壤重金属污染不断加重。土壤在农业生态系统中扮演重要角色，然而由于无节制的矿产开采，废水、废渣和废气的不合理排放，导致矿床周边土壤退化，甚至遭受严重污染[1]。铜(Cu)作为一种重要的重金属，其对土壤乃至整个生态系统的危害已引起国内外科学工作者和政府的高度重视[2]，土壤重金属污染监测及修复工作迫在眉睫。

随着高光谱遥感技术的发展，其以光谱分辨率高、波段多且连续性强等特点将逐步取代传统以化学分析为主的监测方法，高光谱遥感可以获得地物的精细光谱信息，为定量预测分析土壤重金属污染提供了强有力的工具。Kemper[3] 等利用 FieldSpec Pro FR 分光辐射光谱仪测得土壤光谱反射率，并基于多线性回归 MLR 建模以预测 As、Cd、Cu、Fe、Hg、Pb、S、Sb 和 Zn 等 9 种土壤重金属含量，其中有 6 种重金属含量的预测精度超过 70%（$R^2>0.7$），但 Cd、Cu 和 Zn 的预测效果不明显。吴昀昭[4] 等用 PCR 法建立了室内土壤光谱与 Hg 含量的预测模型，二者相关系数 $R=0.69$，预测均方根误差 RMSE=0.15。

然而，在对土壤光谱分析时，国内外学者往往忽略了波段的选择和光谱分辨率对土壤重金属预测的重要性，进而导致预测精度不能进一步提高。此外，在建模算法选择方面，由于高光谱数据波段众多且冗余性高，利用 MLR 和 PCR 法建模反演重金属含量都不是最佳选择，尤其是基于 MLR 的预测模型易出现过拟合现象，而 PLSR 法能够在波段个数相对较多且严重自相关的情况下进行回归建模，适用于对连续光谱的分析[5]。

本文对在南京城郊采集的 181 个土壤样本数据进行实验室 Cu 含量化验分析和室内光谱测量，研究了在不同入选波段数、不同光谱采样间隔下，基于 PLSR 法构建土壤高光谱反射特征与重金属元素 Cu 含量的预测模型，并提出了预测重金属 Cu 的最佳入选波段数和最佳光谱采样间隔，以期为土壤重金属的高光谱遥感预测提供新思路，同时可能为将

收稿日期：2009-12-16；**修订日期**：2010-03-15

基金项目：国家自然科学基金(40971205)、国家科技支撑计划项目(2007BAH15B01)和国防科工委民用航天空间应用项目“新一代环境监测高光谱卫星指标论证”。

作者简介：黄长平(1986－)，男，硕士研究生，研究方向为高光谱遥感。E-mail：hcp2qq@163.com。

来设计低廉实用的高光谱卫星传感器提供指标论证和理论依据。

2 数据获取与研究方法

2.1 数据源获取与预处理

实验数据由南京大学提供，数据采用南京城郊受Cu污染的江宁区和八卦洲区采集的181个土壤样本，其中江宁区120个样本，八卦洲区61个样本。

野外土样采样密度为一个样2 km^2，采样深度为0～20 cm，在每个采样位上采取蛇形采样法[6]。重金属Cu含量经化验分析而得，所用方法为电感耦合等离子光谱法(ICP-AES)[7-8]，实测得Cu含量均值38.2205 mg/kg，大于南京地区土壤重金属Cu的背景值31.69 mg/kg，采用Muller提出的地积累指数法($Igeo=\log_2(Cn/Bn)$，其中Cn为研究元素实测含量，Bn为土壤背景值)对土壤重金属含量进行衡量可知$Igeo>0$，故本实验区土壤属于Cu中度污染[6]；土壤反射率光谱测量采用的仪器为Lambda900光谱仪，其光谱分辨率为2 nm，光谱测量范围400～2 500 nm。测得的土壤样本反射光谱在2 300～2 500 nm波段因数据信噪比低，在840～900 nm处由于仪器换灯致使数据存在噪声，故舍弃这两波段范围。因此有效波段范围为400～840 nm和900～2 300 nm，共922个波段，图1为室内所测江宁区10号土壤样本光谱曲线。

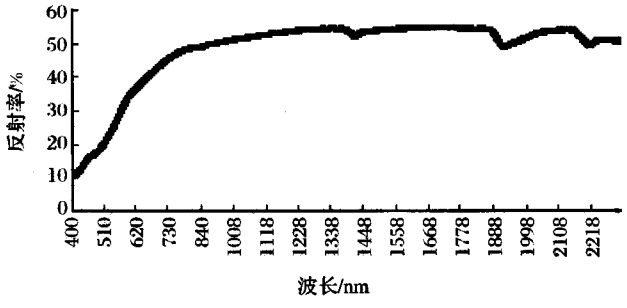


图1 江宁区10号土壤样本的光谱曲线

Fig. 1 Soil spectral curve of Jiangning 10

为校正样品间因散射而引起的光谱的误差，对所有实验数据用公式 $X'=\frac{X-\bar{X}}{\delta}$ 进行标准化处理，式中X'为标准化后的反射率值，X为原始反射率值，$\bar{X}$为均值，δ为标准差。

2.2 研究方法

2.2.1 波段选择

常用波段选择法有前向选择、后向剔除和逐步回归等方法[9-12]。本文采用逐步回归分析法，依据RMSE最小、R^2最大原则，确定入选波段数和波长位置，并以F概率作为引入和剔除变量的依据，F统计量的概率P值Entry设为0.05，Removal值设为0.10[12]。逐步回归选择波段实质是基于多线性回归的方法，入选的波段并不一定是最优波段组合，因此本文进一步利用PLSR法进行建模验证并筛选波段。

2.2.2 光谱分辨率重采样

为分析光谱分辨率对Cu反演精度的影响，给出高光谱遥感监测土壤重金属Cu含量的最佳光谱分辨率，利用高斯响应函数对土壤反射率光谱进行重采样，获取不同的光谱分辨率下的数据集，进而建立不同样本集的PLSR模型，计算相应的建模与验证精度，分析预测精度随光谱分辨率的变化。

具体而言，对181个土样的原始光谱数据进行高斯函数重采样，采样间隔依次为：4 nm、8 nm、16 nm、32 nm、64 nm、128 nm和256 nm，对应波段数分别为：466、236、120、61、31、16和9。高斯函数的表达式如公式(1)，式中μ为期望值，代表每个函数的中心波长；σ为标准差，用来计算每一个高斯函数的两端到中心的距离t。高斯函数的个数为采样后的波段数，采样后的光谱值为每个高斯函数在区间$[\mu-t,\mu+t]$内的积分。高斯函数光谱重采样算法在IDL6.4中编程实现。

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

2.2.3 PLSR模型及其优化

PLSR的基本思路是一种逐步回归，即逐步提取光谱数据中的成分，逐步增加变量，逐步检验模型的显著性，一旦满足要求即停止计算[5]，其本质上还是建立重金属Cu的含量矩阵(因变量)与181个土壤样本入选波长点处的光谱反射率矩阵(自变量)的线性回归模型，但在形式上引入了一个潜在函数，并用自变量的线性组合来表示这个潜在函数，进而再建立因变量对潜在函数的一元线性回归模型以实现由自变量预测因变量(土壤样本光谱反射率预测Cu含量)的目的。

PLSR分析将所有参与建模的波段压缩成一些相互独立的主成分，但PLSR建模并不需要所有的主成分[1,5-6]。当增加主成分个数时，会降低模型拟合误差，同时提高模型的预测精度。但是当主成分数过多时，又会发生过拟合。因此建立一个最佳的PLSR回归模型，最关键的问题之一就是确定抽出主成分数PCs。李民赞[5]等提出预测误差最低点所对应的成分数即为最佳成分数，如图2所示为抽出主成分数与拟合误差和预测误差之间的关系，图中竖线对应的成分数是PLSR模型所需的最佳主成分数。

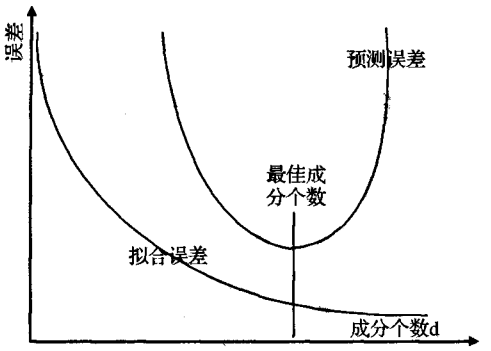


图 2　抽出主成分数与均方根误差 RMSE 的关系

Fig. 2　Relationship between PCs and RMSE

本文用交叉验证法中的外部证实算法来确定 PLSR 算法中抽出的主成分数。具体做法是：将采集的 181 个土壤样本分成建模样本和验证样本，随机选取 72 个样本作为验证样本，其余样本用来建立回归模型。然后利用建模样本建立1个主成分的PLSR模型，将验证样本的光谱反射率代入该模型，计算出预测值 y_p，进而计算模型的预测均方根误差 RMSE。接着，利用建模样本建立 2 个成分的 PLSR 模型，计算相应的 RMSE，然后不断增加主成分数，重复以上步骤，直到新增加到模型中的主成分使得预测 RMSE 的减小不超过 2%为止，此时的成分数为最佳主成分数，相应的 PLSR 模型为最优回归模型[5,13]。

3　结果与讨论

3.1　基于最优波段的 PLSR 预测模型

利用逐步回归法得到入选预测 Cu 含量的波段共 11 个，按变量进入回归模型先后顺序分别为：1 836、1 886、2 284、536、1 764、1 648、1 434、1 512、2 094、530 和 1 894 nm。对不同的入选波段数分别基于 PLSR 建立重金属 Cu 含量的预测模型。表 1 是基于 3～11 个入选波段数和全波段利用 PLSR 建模的结果。

表 1　不同入选波段数的 PLSR 建模与预测结果

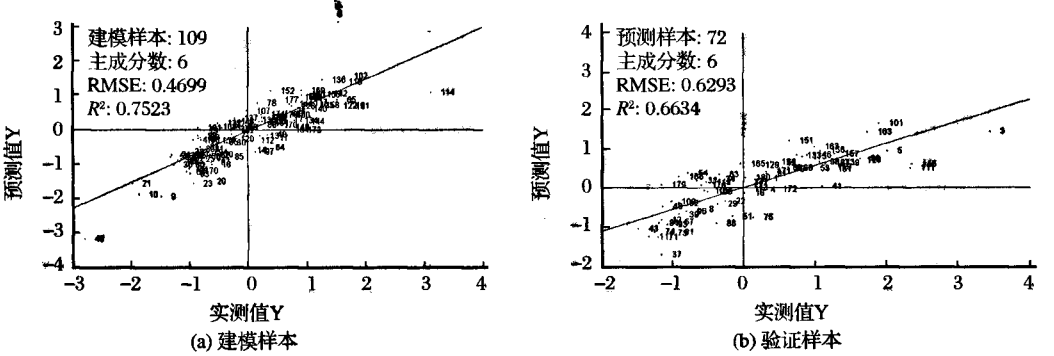
Table 1　Result of PLSR model using different number wavelengths

波段	3	4	5	6	7	8	9	10	11	922
建模 RMSE	0.5315	0.5299	0.5347	0.5329	0.5276	0.4590	0.4844	0.4699	0.4792	0.5276
R^2	0.6831	0.6850	0.6793	0.6814	0.6877	0.7636	0.7367	0.7523	0.7424	0.6878
预测 RMSE	0.7116	0.6956	0.7031	0.7036	0.6751	0.6460	0.6332	0.6293	0.6372	0.7057
R^2	0.5696	0.5887	0.5798	0.5792	0.6126	0.6453	0.6592	0.6634	0.6549	0.5767

由表 1 可知，随着入选波段数的不断增加，PLSR 回归模型的预测 RMSE 并不是一直在减小，当波段数为 11 时，RMSE 开始增大，尤其当全波段参与建模时，模型的预测 RMSE 增加到 0.7057。这说明波段选择可以提高 Cu 含量的预测精度；但当入选波段数增加到 11 个时，预测模型出现过拟合现象，反而降低了模型的预测精度。根据确定最佳预测模型的标准：即相对低的预测 RMSE，且建模 RMSE 与预测 RMSE 接近的准则[6]，确定入选波段数为 10 时的 PLSR 回归模型为预测重金属元素 Cu 的最佳模型。利用所得 6 个主成分与土壤重金属 Cu 含量做回归模型，所得计算结果如下：

$$\begin{aligned} y = &-7.32b_{1836} + 2.195b_{1886} - 6.009b_{2284} \\ &- 5.055b_{536} + 7.789b_{1764} - 5.797b_{1648} \\ &+ 7.181b_{1434} - 5.933b_{1512} + 7.66b_{2094} \\ &+ 4.548b_{530} - 0.02047 \end{aligned} \tag{2}$$

其中：y 为土壤重金属 Cu 的预测值，b_i 代表波长 i 处的光谱反射率。模型的建模 $R^2 = 0.7523$，说明上述 10 个波段处的光谱反射率与重金属 Cu 含量之间存在显著的线性关系。回归结果的散点图如图 3 所示。



(a) 建模样本　　(b) 验证样本

图 3　PLSR 模型的建模与验证结果比较

Fig. 3　Comparison between predicted and measured values of PLSR model

图 4 是 Cu^{2+} 和土壤的反射率曲线，在晶体场的影响下，Cu^{2+} 在 530 nm 附近有一反射峰值，因而 Cu^{2+} 呈绿色；在 1 470、1 836、2 000、2 280 nm 等附近，有盐的 OH 和水吸收峰[6,14]。与上述 10 个波段对照可知，这 10 个波段大都位于 Cu 的吸收特征谱段，是反演土壤 Cu 的敏感波段，具有很强的物理意义。

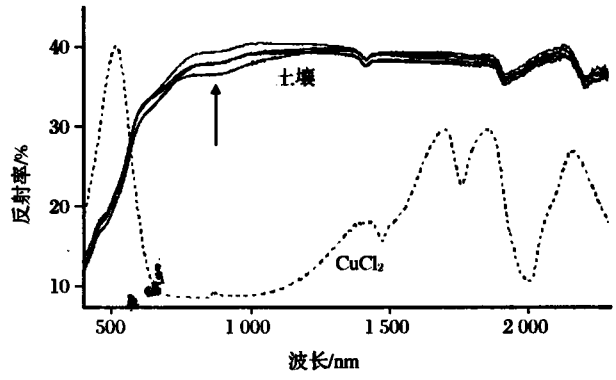


图 4 铜离子和土壤的反射率光谱

Fig. 4 Reflectance curves of Cu^{2+} and soil

3.2 基于最佳光谱分辨率的 PLSR 预测模型

为了研究预测土壤重金属 Cu 含量所需要的光谱分辨率，本文对上述基于高斯响应函数重采样的所有分辨率的样本集都计算了 PLSR 模型，PLSR 模型回归结果如表 2。

表 2 不同光谱分辨率下的 PLSR 建模结果

Table 2 Result of PLSR model based on different resolutions

光谱分辨率/nm	2	4	8	16	32	64	128	256
PCs	5	8	6	6	8	7	7	3
建模 RMSE	0.5276	0.5020	0.5271	0.5271	0.5147	0.5187	0.5226	0.5773
R^2	0.6878	0.7173	0.6884	0.6883	0.7028	0.6982	0.6936	0.6262
预测 RMSE	0.7057	0.6900	0.7043	0.7042	0.6895	0.6899	0.6930	0.8008
R^2	0.5767	0.5954	0.5784	0.5785	0.5959	0.5955	0.5918	0.4550

图 5 是不同分辨率 PLSR 模型的性能比较，直观地说明光谱分辨率对基于 PLSR 模型的土壤重金属 Cu 预测精度的影响。

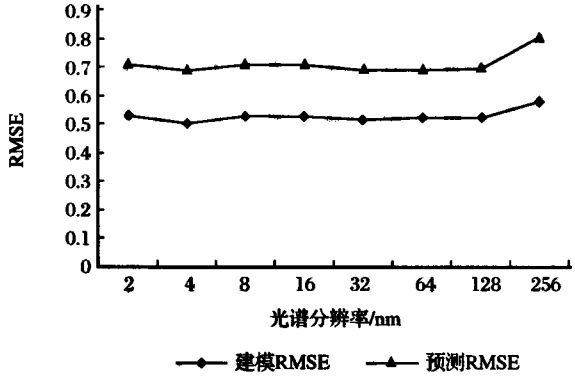


图 5 不同光谱分辨率的 PLSR 模型的 RMSE 变化

Fig. 5 The performance of PLSR model using different spectral resolutions

由图 5 可知，随着光谱分辨率的降低，PLSR 模型的预测 RMSE 先减小后增大，到光谱分辨率为 256 nm 时，RMSE 最大，达 0.8008，对应的 R^2 最小为 0.4550。按照确定最佳预测模型的标准，确定光谱分辨率为 32 nm 时的 PLSR 模型为预测土壤重金属 Cu 含量的最佳模型。最佳模型的主成分数为 8 个，各项指标分别为：建模 $R^2=0.7028$，RMSE=0.5147；预测 $R^2=0.5959$，RMSE=0.6895。由此看来并不是光谱分辨率越高，重金属元素 Cu 含量的预测精度就越高，相反，相对低的光谱采样间隔(32 nm)可以确保达到最佳预测精度。这是因为重金属元素光谱特征较宽，不需要尖锐的吸收峰，而且相对较低的光谱分辨率可能增强了光谱信噪比，从而提高了预测精度。这一结论说明将来设计监测土壤重金属污染的高光谱卫星传感时，并不需要追求过高的光谱分辨率。

4 结 语

本研究主要对室内土壤光谱进行分析与处理，分别探讨了基于不同入选波段数、不同光谱采样间隔 PLSR 建模预测南京城郊江宁和八卦洲土壤重金属 Cu 含量的精度问题。结果表明，并不是波段数越多，光谱采样间隔越窄，预测效果越好，对于本文 Cu 含量的预测，基于 10 个入选波段建立的 PLSR 模型和基于 32 nm 光谱采样间隔建立的 PLSR 模型都具有理想的拟合效果和可靠的预测能力。此研究不仅适用重金属 Cu 的预测，同时为其他土壤重金属的预测提供了思路，也可能为模拟卫星传感器波段预测土壤重金属含量提供理论依据，为将来设计低廉实用的高光谱卫星传感器提供指标论证。

文中土壤光谱反射率为室内测得，条件理想，若应用于大范围的室外土壤重金属污染的监测，还需

考虑很多因素，如光照、大气、地形、土壤表面粗糙度等。因此利用遥感技术快速、大范围、经济地制图土壤重金属污染还有待进一步探讨，也是今后主要的研究工作。

参考文献：

[1] Ren H Y, Zhuang D F. Estimation of As and Cu Contamination in Agricultural Soils around a Mining Area by Reflectance Spectroscopy: A Case Study [J]. Pedospere, 2009, 19 (2): 719-726.

[2] Liu Suhong, Liu Xinhui, Hou Juan, *et al*. Study on the Spectral Response of Brassica Campestris to the Cu Pollution[J]. Science in China Series E, 2007, 37(5): 693-699. [刘素红，刘新会，侯娟，等. 植物光谱应用于白菜铜胁迫响应研究[J]. 中国科学-E 辑，2007，37(5)：693-699.]

[3] Kemper T, Sommer S. Estimate of Heavy Metal Contamination in Soils after a Mining Accident Using Reflectance Spectroscopy[J]. Environmental Science and Technology, 2002, 36 (12): 2742-2747.

[4] Wu Y Z, Chen J, Ji J F, *et al*. Feasibility of Reflectance Spectroscopy for the Assessment of Soil Mercury Contamination [J]. Environmental Science and Technology, 2005, 39 (3): 873-878.

[5] Li Minzan. Spectral Analysis Technology and Its Application [M]. Beijing: Science Press, 2006. [李民赞. 光谱分析技术及其应用[M]. 北京：科学出版社，2006.]

[6] Wu Yunzhao. Heavy Metal Pollution in Suburban Soils of the Nanjing Area——A Feasibility Study of Remote Sensing Geochemistry[D]. Nanjing: Nanjing University, 2005. [吴昀昭. 南京城郊农业土壤重金属污染的遥感地球化学基础研究[D]. 南京：南京大学，2005.]

[7] Xie Sujing, Xie Shulian, Xie Baomei. Analysis of Ca, Mg, Fe, Mn, Cu and Zn in Algae[J]. Spectroscopy and Spectral Analysis, 2003, 23(3): 615-616. [谢苏婧，谢树莲，谢宝妹. 藻类植物中钙、镁、铁、锰、铜和锌含量分析[J]. 光谱学与光谱分析，2003，23(3)：615-616.]

[8] Chen Sining, Liu Xinhui, Hou Juan, *et al*. Study on the Spectral Response of Brassica Campestris Leaf to the Zinc Pollution[J]. Spectroscopy and Spectral Analysis, 2007, 27(9): 1797-1801. [陈思宁，刘新会，侯娟，等. 重金属锌胁迫的白菜叶片光谱响应研究[J]. 光谱学与光谱分析，2007，27(9)：1797-1801.]

[9] Li Yunmei, Ni Shaoxiang, Wang Xiuzhen. The Robustness of Linear Regression Model in Rice Leaf Chlorophyll Concentration Prediction[J]. Journal of Remote Sensing, 2003, 7(5): 364-370. [李云梅，倪绍祥，王秀珍. 线性回归模型估算水稻叶片叶绿素含量的适宜性分析[J]. 遥感学报，2003，7(5)：364-370.]

[10] Peng Y K, Wang W, Huang H, *et al*. Prediction of Chlorophyll Content of Winter Wheat Using Leaf-level Hyperspectral Imaging Data[Z]. ASABE, 2009.

[11] Gao Y H, Chen L F, Zhou X, *et al*. Optimal Bands for Estimation of Mixed Canopy Chlorophyll Content[J]. Journal of Remote Sensing, 2009, 13(4): 616-622.

[12] Liu Dahai, Li Ning, Chao Yang. SPSS 15.0 Statistical Analysis[M]. Beijing: Tsinghua University Press, 2008. [刘大海，李宁，晁阳. SPSS 15.0 统计分析[M]. 北京：清华大学出版社，2008.]

[13] Kooistra L, Salas E A L, Clevers J G P W, *et al*. Exploring Field Vegetation Reflectance as an Indicator of Soil Contamination in River Floodplains [J]. Environmental Pollution, 2003, 127(2004): 281-290.

[14] Tong Qingxi, Zhang Bing, Zheng Lan fen. Hyperspectral Remote Sensing[M]. Beijing: Higher Education Press, 2006. [童庆禧，张兵，郑兰芬. 高光谱遥感——原理、技术与应用[M]. 北京：高等教育出版社，2006.]

Study on Band Selection and Optimal Spectral Resolution for Prediction of Cu Contamination in Soils

HUANG Chang-ping[1,2], LIU Bo[1,2], ZHANG Xia[1], TONG Qing-xi[1,3]

(1. *The State Key Laboratory of Remote Sensing Sciences, Institute of Remote Sensing Application, Chinese Academy of Sciences, Beijing* 100101, *China*;
2. *Graduate University of Chinese Academy of Sciences, Beijing* 100049, *China*;
3. *Institute of Remote Sensing and GIS, Peking University, Beijing* 100871, *China*)

Abstract: Hyper-spectral data offers a powerful tool for predicting soil heavy metal contamination due to its high spectral resolution and many continuous bands. Band selection and spectral resolution, however, are the prerequisite of heavy metal inversion by hyper-spectral data. In this study, soil reflectance spectra and

下转 341 页

上接 357 页

their Cu contents were measured for 181 soil samples in the laboratory. Based on these dataset, band selection was conducted to inverse Cu content using stepwise regression approach, and prediction accuracies of Cu based on partial least-squares regression (PLSR) model with different selected bands were analyzed. In addition, the influences of spectral resolution on prediction results of Cu were discussed by a Gaussian re-sampling function. It demonstrated that the optimal band number was 10 for Cu inversion and the corresponding model prediction accuracy was R^2 0. 7523 and RMSE of 0. 4699. The optimal spectral resolution was 32 nm and the model on this basis had an accuracy of $R^2=0.7028$ and RMSE=0. 5147. Results of this paper may provide scientific verification for designing low-cost and practical hyper-spectral space-borne sensors and provide theoretical bases for simulating space-borne sensors to predict soil heavy metals content in the future.

Key words: Remote sensing prediction of Cu; Hyper-spectral data; Spectral re-sampling; PLSR; Band selection