利用可见光近红外的尾矿区农田土壤 Cu 含量反演

吕 杰，郝宁燕，崔晓临
（西安科技大学测绘科学与技术学院，西安 710054）

摘 要：矿山开采普遍存在土壤重金属污染问题，有效的进行尾矿区农田土壤重金属含量估算迫在眉睫。以陕西金堆城矿区尾矿为研究区，采集土壤样本，测量土壤可见光近红外光谱，测试分析土壤铜元素含量。将 Isomap（Isometrio mapping）和 LLE（locally linear embedding）流形学习方法应用于土壤高光谱降维，基于随机森林构建估算模型，反演土壤铜含量。结果表明：降维后的高光谱数据反演精度更高，Isomap 降维后模型预测结果均方根误差为 30.50，R^2=0.76，优于 LLE 降维结果。研究为尾矿区土壤 Cu 元素含量的快速反演估算提供了理论依据。

关键词：土壤；光谱测定；重金属；铜；流形学习；随机森林
doi：10.11975/j.issn.1002-6819.2015.09.040
中图分类号：P575.4 **文献标志码**：A **文章编号**：1002-6819(2015)-09-0265-06

吕 杰，郝宁燕，崔晓临. 利用可见光近红外的尾矿区农田土壤 Cu 含量反演[J]. 农业工程学报，2015，31(9)：265－270.
Lü Jie, Hao Ningyan, Cui Xiaolin. Inversion model for copper content in farmland of tailing area based on visible-near infrared reflectance spectroscopy[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(9): 265－270. (in Chinese with English abstract)

0 引 言

在过去的 20 a 里，采矿、运输、污水处理和施肥等人类活动，已经对土壤健康构成一个持续的威胁[1]。尤其在对矿产资源的开发和利用过程中，更是不可避免的带来许多环境和灾害问题。其中，尾矿农田土壤重金属污染是矿区生态环境最严重的问题之一。重金属离子进入农田土壤，影响土壤的正常功能，污染了生长在土壤中的农作物（如水稻、玉米和大豆）。这样不仅大大提高了食品的安全隐患，也时刻威胁着人体健康。因此，尾矿农田土壤重金属的测定是监控土壤健康状况和预防土壤污染的必要手段。

有效防治土壤重金属污染的关键问题是如何快速准确地获取其含量及分布信息[2]。传统土壤重金属污染监测和识别方法是野外采样带回实验室进行化学分析[3-5]。虽然测量精度高，准确性强，但在大尺度监测土壤重金属含量时不仅成本高，而且费时费力。因此，将高光谱遥感引入尾矿土壤重金属含量的快速估算具有一定的研究价值和实用意义。高光谱遥感因其波段多、光谱响应范围广、数据丰富被广泛应用于土壤生态环境、矿产资源调查等领域[6-9]。利用高光谱遥感对土壤重金属污染进行探测，具有高效、便捷、无损等优势[10]。但同时也因其数据量大、数据冗余度高和 Hughes 现象给数据处理带来极大的挑战[11]。因此高维数据处理一直是高光谱遥感信息获取地表参数的难题[12]。

流形学习（manifold learning）是模式识别、机器学习和数据挖掘研究领域中的热点，它能够利用数据集的局部几何结构来揭示其内在的流形结构，以达到高效维数降维的目的[13]。近年来已被广泛用于遥感影像的降维与特征提取[12,14-16]。重金属元素在污染土壤中的含量甚微，光谱中的重金属元素响应信号也较为微弱，因此有必要对土壤光谱进行处理以降低信息的相关性和冗余。将流形学习引入到重金属污染土壤的高光谱数据降维，是一个有意义的尝试。本文以陕西金堆城矿区尾矿为研究区，利用 Isomap 和 LLE 流形学习方法对样点实测土壤高光谱进行降维处理，基于随机森林构建土壤 Cu 含量高光谱估算模型，分析土壤 Cu 含量估算模型的机理，以期实现高光谱遥感大尺度估算矿区尾矿土壤重金属含量状况。

1 材料与方法

1.1 研究区与采样

研究区位于陕西渭南著名的西岳华山南麓华县金堆城（图 1），是中国大型的露天矿山和钼业生产基地。属暖温带半湿润气候区，年平均气温 13.4℃，降水量 583.4 mm。该矿区面积约 4.5 km²，精矿产量居全国之冠，已探明钼资源量 1 011 461.22 t。该区域农作物以玉米、小麦、豆类为主，农业生产占主要地位。在研究区分别建立 4 个采样区，按照梅花采样法在每个采样区采取 0～20 cm 土层土壤样本 15 个（图 1），覆盖的土地类型包括小麦、玉米、大豆和赤裸的土地。各采样点的位置由精度为±5 m 的手持全球定位系统（GPS）获得。每个土壤样品单独保存在一个标记的样品袋中。

收稿日期：2014-12-23 修订日期：2015-03-26
基金项目：国家自然科学基金（51104116）；江西省数字国土重点实验室开放研究基金（DLLJ201305）；农业部农业信息技术重点实验室开放课题（2013006）；地理空间信息湖南省工程实验室开放研究基金（2013GSIJJ002）
作者简介：吕 杰，男，山东蓬莱人，博士，讲师，主要从事高光谱遥感研究。西安 西安科技大学测绘科学与技术学院，710054。Email：rsxust@163.com

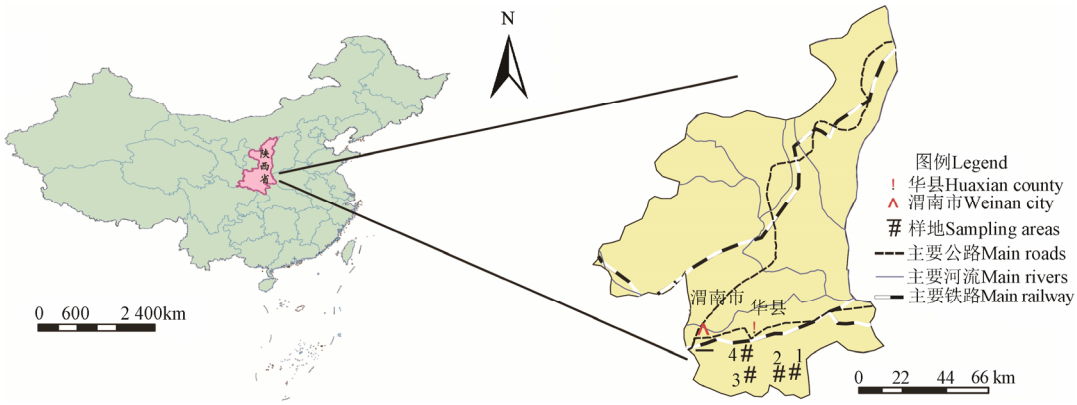

图 1 陕西省渭南市华县金堆城钼矿采集的 4 个土壤样区分布图

Fig.1 Spatial distribution of 4 soil sampling areas at Huaxian, Shaanxi Province, China

1.2 土壤样品 Cu 含量化学分析

将采集的土壤样品均匀风干，在风干过程中碾碎后过 2 mm 的聚乙烯筛，以筛除碎石、卵石以及植物残骸，研磨后过 0.15 mm 的聚乙烯筛，将最终筛过的样品分成 2 份存储于磨口广口瓶中，分别用于 Cu 元素含量的测定和土壤光谱的测量。

土壤 Cu 元素含量采用火焰原子吸收分光光度法（GB/T17138-1997）测定，测量前用研磨机将筛选的土样粉碎。

1.3 土壤样品的光谱测量

土壤样品的光谱反射率采用美国 ASD（Analytical Spectral Devices，ASD）便携式光谱仪测量，它覆盖的光谱范围为 350~2500 nm，每次可进行 10 组数据采集。为了控制光照条件和减少杂散光的影响，光谱测量在室内进行。测量前用标准白板进行校正，设定视场角为 8°，探头到土壤样本表面距离为 1.35 m。每个土壤样本采集 10 组土壤光谱，取均值作为土壤的反射率光谱。

在 ASD 光谱仪采集、获取以及传输光谱信号的过程中，会产生一些噪声。因为噪声光谱偏离了真实状况，必然影响到估算土壤 Cu 金属含量，因此有必要进行光谱降噪处理，本研究采用 db6 小波函数对土壤原始高光谱运行小波变换，得到变换后的 60 个土壤反射率光谱。（如图 2）。

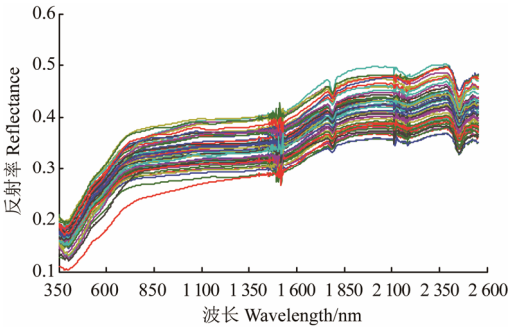

图 2 小波变换降噪的 60 个样本近红外光谱图

Fig.2 NIR spectra of 60 samples denoised by wavelet transform

2 数据分析

2.1 Isomap 算法

等距特征映射（Isometrio mapping，Isomap）算法是 Tenenbaum 等学者提出的一种典型的全局特性保持方法[17]。通过计算样本向量之间的测地距离来代替欧氏距离，挖掘出样本向量在空间中内在的几何关系。测地距离计算的时候需要确定本征维数 d 和邻域大小 k。但是，Isomap 要求所学习的流形必须是非凸的，因为当流形曲率较大或流形上有“洞”时，嵌入结果会产生较大的变形。

Isomap 算法的主要思想是[17]：

构造领域图 G：计算每个样本点和其余样本点之间的欧氏距离。如果样本点 x_i 和 x_j 的欧氏距离 $d_x(i,j)$小于给定阈值 ε 或者 x_i 是 x_j 的第 k 个邻近点，则规定 x_i 和 x_j 是相邻的。即图 G 有边，并设边的权重为 $d_x(i,j)$。

计算最短路径：当图 G 有边时，初始化最短路径 $d_G(i,j)=d_x(i,j)$，否则 $d_G(i,j)=\infty$。根据迪杰斯特拉（Dijkstra）算法求出任意 2 个样本点之间的最短路径距离 $d_G(i,j)=\min\{d_G(i,j),d_G(i,k)+d_G(k,j)\}$（$k$=1,2,...,$n$；$n$ 为样本数），得到最短路径距离矩阵 $\boldsymbol{D_G}=\{d_G(i,j)\}$。

计算 e 维嵌入：令 λ_p 表示矩阵 $\boldsymbol{\tau(D_G)}$的第 p 个特征值（降序），v^i_p 表示第 p 个特征值的第 i 个组分。然后设置 e 维坐标矢量 y_i 的第 p 个组分的值为 $\sqrt{\lambda_p}v^i_p$。运用 Isomap 对 60 个土壤样本高光谱数据进行降维。

2.2 LLE 算法

局部线性嵌入（local linear embedding，LLE）算法是 Sanl 和 Roweis 提出的与 Isomap 相似的一种局部非线性降维算法[18]。该算法在保持原始数据性质不变的情况下，将高维空间的信号映射到低维空间上。降维过程需要先确定局部邻域参数 k 和嵌入维数 d。

Isomap 试图保留数据点的全局特性，LLE 只尝试保留数据点的局部性质，这使得 LLE 能够解决“洞”的问题。LLE 算法的具体思想为[18]：

利用欧氏距离计算出每个样本点 $\vec{x}_i$ 的 k 个近邻点。由每个样本点 $\vec{x}_i$ 的近邻点计算出该样本点的最佳局部重建权值 w_{ij}，即最小化：

$$\varepsilon(w)=\sum_i\left|\vec{x}_i-\sum_j w_{ij}\vec{x}_j\right|^2 \quad (1)$$

式中：$\sum_j w_{ij}=1$，如果 $\vec{x}_j(j=1,2,...,n)$ 不是 $\vec{x}_i(i=1,2,...,n)$ 的近邻，则 w_{ij}=0。

由该样本点的局部重建权值矩阵计算低维嵌入矢量

$\vec{y}_i$，由于在低维空间中尽量保持高维空间中的局部线性结构，而 w_{ij} 表示局部信息，所以固定 w_{ij}，使下面的损失函数最小化，从而计算出该样本点的输出值。

$$\phi(y)=\sum_i\left|\vec{y}_i-\sum_j w_{ij}\vec{y}_j\right| \tag{2}$$

式中：$\sum_i y_i=0$，$\frac{1}{n}\sum_i \vec{y}_i\vec{y}_i^{\,T}=1$，以使 $\phi(y)$ 对平移、旋转和伸缩变化都具有不变性。

2.3　随机森林

随机森林算法（random forests，RF）是 Leo Breiman 提出的一种分类器集合算法[19]。随机森林通过自助法（bootstrap）重采样技术，从原始训练样本集 A 中有放回的，重复的随机抽取 K 个样本生成新的训练样本集合，然后根据自助样本集生成 K 个分类树组成随机森林，待测数据的分类结果按分类树投票所得到的分数而定[20]。随机森林算法的实质是对决策树算法的一种改进。

随机森林相对于其他集成学习算法的最主要优势在于对结果的可解释性。随机森林结果的可解释性在于变量重要性的测算。即计算每棵树 OOB（out-of-bag）误差和挑选自变量序列后每棵树的 OOB 误差的差（式（3）、式（4））[21]。OOB 误差是无偏估计，近似于交叉验证得到的误差。

$$FI^{(t)}(f)=\sum_{x_i\in\beta^{c(t)}} I\left(I_j=c_i^{(t)}\right)\Big/\left|\beta^{c(t)}\right|-\sum_{x_i\in\beta^{c(t)}} I\left(I_j=c_{i,\pi f}^{(t)}\right)\Big/\left|\beta^{c(t)}\right| \tag{3}$$

式中：$\beta^{c(t)}$ 与第 t 棵树的 OOB 样本相关，其中 $t=\{1,...,T\}$。$c_i^{(t)}$ 和 $c_{i,\pi f}{}^{(t)}$ 预测的样本 x_i 在转换特征 f 前后的预测类别。需要指出的是，如果特征 f 不在第 t 棵树中时，$FI^{(t)}(f)=0$。特征 f 作为全部树的变量重要性计算如下：

$$FI(f)=\sum_T FI^{(t)}(f)\Big/T \tag{4}$$

式中：T 为树的数目。

2.4　模型精度检验

选用决定系数 R^2 和均方根误差（root mean squared error，RMSE）作为流形学习降维和随机森林构建的尾矿农田 Cu 含量高光谱反演模型的评价依据。决定系数是用以反映变量之间相关关系密切程度的统计指标，由式（5）计算得到。均方误差是实际数据与预测数据平均化的方差，是衡量“平均误差”的一种较方便的方法，可以评价数据的变化程度：（式（6））。

$$R^2=1-\sum_{i=1}^{T}\left(F(x_i)-y_i\right)^2\Big/\sum_{i=1}^{T}\left(y_i-\bar{y}\right)^2 \tag{5}$$

$$\mathrm{MSE}=\sum_{i=1}^{T}\left(F(x_i)-y_i\right)^2\Big/N \tag{6}$$

式中：N 验证样本数，$\bar{y}=\sum_{i=1}^{T}y_i/N$ 是验证样本的平均值。

3　结果分析与讨论

3.1　结果分析

随机森林构建土壤 Cu 金属含量高光谱反演模型时的变量重要性结果如图 3 所示。重要性值越大，说明变量特征越重要。由图 3 可以看出，最重要的光谱特征值位于波长 674、1 486、1 639 nm 处，其次重要的特征位于 359、448、770、808、911、1 404、1 710、1 827、2 202 nm 处，其余变量特征也对模型有影响，但对随机森林构建土壤 Cu 金属含量高光谱反演模型影响甚微。

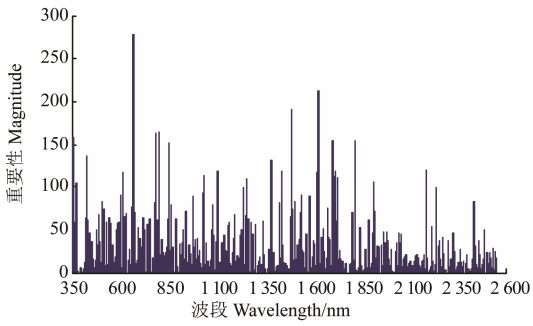

图 3　随机森林变量重要性

Fig.3　Variable importance of random forests

对采集的 60 个原始土壤高光谱样本（60×881）分别进行 Isomap 和 LLE 流形学习降维。在降维过程中，不同算法的参数设置会产生不同的结果。研究分别选取 k 值范围从 10～50，e 值范围从 8～15 对土壤高光谱数据分别进行 Isomap 和 LLE 流形学习降维处理。降维后随机选择 42 个样本用作随机森林模型的构建，而其余的 18 个样本用做模型的预测和检验。图 4 分别为原始数据、LLE 降维、Isomap 降维的土壤 Cu 含量反演图。当 e=12，k=45 时，Isomap 降维具有最小的 MSE 值 30.50，R^2 为 0.76；当 e=12，k=12 时，LLE 降维具有最小的 MSE 值 32.82，R^2 为 0.67（如表 1）。

表 1　反演模型在不同数据集上精度比较

Table 1　Accuracy comparison of inversion model on different data sets

数据 Dataset	均方根误差 RMSE	R^2
原始数据集 Original dataset	36.89	0.50
LLE 降维数据集 Reduced dimentional dataset by LLE	32.82	0.67
Isomap 降维数据集 Reduced dimentional dataset by Isomap	30.50	0.76

为了更清晰展示 Isomap 流形学习降维对高光谱数据集有效信息的提取结果，绘制 Isomap 降维后构建的随机森林模型估算的 Cu 含量和实际测量的 Cu 元素含量之间的散点图（如图 5）。

结果表明：流形学习降维后的高光谱数据对尾矿土壤 Cu 含量的反演精度明显高于原始土壤高光谱数据。说明高达 881 维的原始高光谱数据拥有严重的信息冗余，且可能存在一定的非线性流形结构，而 Isomap 和 LLE 等流形学习方法可以有效提取这一信息，找到高维空间中的低维流形，并挖掘其相应的嵌入映射，以实现数据的有效降维。

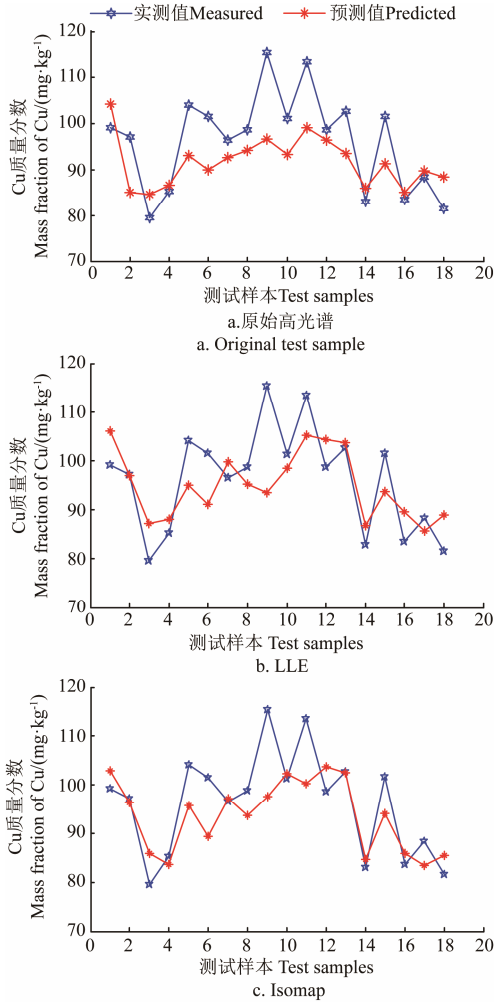

图 4 不同方法的 Cu 含量反演结果

Fig.4 Inversion of Cu on test sample with different methods

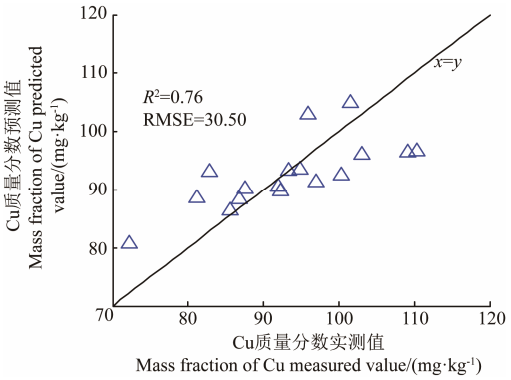

图 5 Isomap-随机森林模型估算 Cu 含量与实测值散点图

Fig.5 Scatter plot of measured Cu content values versus estimated Cu content by Isomap-RF model

Isomap 流形学习降维后构建的随机森林反演 Cu 含量模型相比 LLE 流形学习降维后构建的随机森林反演 Cu 含量模型的预测值与实测值更接近。说明 Isomap 流形学习降维方法更适合于本研究的尾矿土壤高光谱降维，它提取了隐含在土壤高光谱中重金属的有效信息，同时也说明试验数据不存在曲率较大和“洞”的现象。

从 Isomap 降维后构建的随机森林模型估算的 Cu 含量和实际测量的 Cu 元素含量之间的散点图可以看出：大多数样本实测与反演值集中在 1:1 线附近，说明反演结果比较准确。

3.2 讨论

流形学习降维方法在数据降维的过程中力求保持数据点的内在几何特性被广泛应用于各行各业，在以往的分类、反演研究中也取得了较好的精度[22-24]，是一种可替代的、高效的、很有前途的高光谱降维方法。因此，引入更多的降维方法对尾矿农田土壤高光谱数据进行信息提取具有一定的研究价值。研究发现 Isomap 流形学习方法提取了隐含在土壤高光谱中重金属的有效信息，并且反演土壤 Cu 含量的精度比 LLE 方法更高。

文中利用随机森林进行土壤 Cu 含量的高光谱建模，构建土壤 Cu 含量估算模型，估算模型估算土壤 Cu 含量获得了较高的精度。并且基于随机森林构建的估算模型能够解释输入模型的变量重要性，使得估算模型具有可解释性，这是之前的相关土壤 Cu 含量高光谱估算研究所未涉及的。

由于土壤重金属测试的限制，研究只对土壤 Cu 元素含量进行了反演预测，没有考虑其他重金属元素对土壤光谱的作用，未来的研究将开展尾矿土壤多种重金属元素含量的反演。

本研究没有考虑土壤的其他特性，事实上，土壤中的黏土、有机物与重金属（如 Fe、As、Pb、Zn 等）都是密切相互作用的[1]。不同的重金属、有机质对 Cu 的吸附强度也是不同的[2]。因此，今后将进一步研究黏土、有机物等土壤特性对重金属含量评估的影响。

4 结 论

本文以尾矿土壤 Cu 含量为研究对象，基于流形学习降维方法 Isomap 和 LLE 进行土壤 Cu 含量随机森林估算模型的尝试研究。结果表明：进行流形学习降维后的土壤光谱反演 Cu 含量精度高于原始土壤光谱，且 Isomap 流形学习方法能够有效提取土壤中的有效信息，其反演结果优于 LLE 流形学习方法，更适合研究区域土壤 Cu 含量的反演。Isomap-随机森林估算模型 Cu 含量和实测值结果集中在 1:1 线附近，R^2=0.76，RMSE=30.50。研究为利用可见光近红外光谱反演尾矿土壤重金属含量的研究提供参考，对矿区土壤污染监测和治理具有一定的指导意义。

[参 考 文 献]

[1] Wang Junjie, Cui Lijuan, Gao Wenxiu, et al. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy[J]. Geoderma, 2014, 216: 1－9.

[2] 贺军亮，蒋建军，孙中伟，等. 土壤重金属含量光谱估算模型的初步研究[J]. 农机化研究，2009，31(9)：22－25.
He Junliang, Jiang Jianjun, Sun Zhongwei, et al. Studying on retrieval of soil heavy metal content using the organic matter identification index[J]. Journal of Agricultural Mechanization Research, 2009, 31(9): 22－25. (in Chinese with English abstract)

[3] 陈翠华，倪师军，何彬彬，等. 基于污染指数法和 GIS 技术评价江西德兴矿区土壤重金属污染[J]. 吉林大学学报：

地球科学版，2008，38(1)：105－111.
Chen Cuihua, Ni Shjun, He Binbin, et al. Assessing heavy metals contamination of soils based on the pollution index and GIS methods in dexing mines, Jiangxi Province, China[J]. Journal of Jilin University: Earth Science Edition, 2008, 38(1): 105－111. (in Chinese with English abstract)

[4] 王英辉，陈学军，祁士华. 铅锌矿区土壤重金属污染与优势植物累积特征[J]. 中国矿业大学学报，2007，36(4)：447－454.
Wang Yinghui, Chen Xuejun, Qi Shihua. Heavy metal pollution in soils and plant accumulation in a restored Lead-Zinc Mineland in Guangxi, South China[J]. Journal of China University of Mining & Technology, 2007, 36(4): 447－454. (in Chinese with English abstract)

[5] 甘凤伟，方维萱，王训练，等. 锡矿尾矿库土壤-食用马铃薯和豌豆中重金属污染状况[J]. 生态环境，2008，17(5)：1847－1852.
Gan Fengwei, Fang Weixuan, Wang Xunlian, et al. The heavy metal contamination in soil-potato and pea of tin tailings[J]. Ecology and Environment, 2008, 17(5): 1847－1852. (in Chinese with English abstract)

[6] 徐明星，吴绍华，周生路，等. 重金属含量的高光谱建模反演:考古土壤中的应用[J]. 红外与毫米波学报，2011，30(2)：109－114.
Xu Mingxing, Wu Shaohua, Zhou Shenglu, et al. Hyperspectral reflectance models for retrieving heavy metal content: Application in the archaeological soil[J]. J Infrared Millim Waves, 2011, 30(2): 109－114. (in Chinese with English abstract)

[7] 王维，沈润平，吉曹翔. 基于高光谱的土壤重金属铜的反演研究[J]. 遥感技术与应用，2011，26(3)：348－353.
Wang Wei, Shen Runping, Ji Caoxiang. Study on heavy metal Cu based on hyperspectral remote sensing[J]. Remote Sensing Technology and Application, 2011, 26(3): 348－353. (in Chinese with English abstract)

[8] 刘堂友，匡定波，尹球. 湖泊藻类叶绿素-α 和悬浮物浓度的高光谱定量遥感模型研究[J]. 红外与毫米波学报，2004，23(1)：11－15.
Liu Tangyou, Kuang Dingbo, Yi Qiu. Study on hyperspectral quantitative model of concentrations for chlorophyll-α of alga and suspended particles in Tailake[J]. J Infrared Millim Waves, 2004, 23(1): 11－15. (in Chinese with English abstract)

[9] 谭克龙，周日平，万余庆，等. 地下煤层燃烧的高光谱及高分辨率遥感监测方法[J]. 红外与毫米波学报，2007，26(5)：349－358.
Tan Kelong, Zhou Riping, Wan Yuqing, et al. Remote sensing monitoring method of hyperspectral and hige-resolution for underground coal bed combustion[J]. J. Infrared Millim Waves, 2007, 26(5): 349－358. (in Chinese with English abstract)

[10] 付馨，赵艳玲，李建华，等. 高光谱遥感土壤重金属污染研究综述[J]. 中国矿业，2013，22(1)：65－68.
Fu Xin, Zhao Yanling, Li Jianhua, et al. Research on hyper-spectral remote sensing in heavy metal pollution soil[J]. China Mining Magazine, 2013, 22(1): 65－68. (in Chinese with English abstract)

[11] Hughes G F. On the mean accuracy of statistical pattern recognizers[J]. IEEE Transactions on Information Theory, 1968, 14(1): 55－63. (in Chinese with English abstract)

[12] 杜培军，王小美，谭琨，等. 利用流形学习进行高光谱遥感影像的降维与特征提取[J]. 武汉大学学报：信息科学版，2011，36(2)：148－152.
Du Peijun, Wang Xiaomei, Tan Kun, et al. Dimensionality reduction and feature extraction from hyperspectral remote sensing imagery based on manifold learning[J]. Geomatics and Information Science of Wuhan University, 2011, 36(2): 148－152. (in Chinese with English abstract)

[13] 王自强，钱旭，孔敏. 流形学习算法综述[J]. 计算机工程与应用，2008，44(35)：9－12.
Wang Ziqiang, Qian Xu, Kong Min. Survey on manifold learning algorithms[J]. Computer Engineering and Applications, 2008, 44(35): 9－12. (in Chinese with English abstract)

[14] 陈宏达，普晗晔，王斌，等. 基于图像欧氏距离的高光谱图像流形降维算法[J]. 红外与毫米波学报，2013，32(5)：450－455.
Chen Hongda, Pu Hanye, Wang Bin, et al. Image Euclidean distance-based manifold dimensionality reduction algorithm for hyperspectral imagery[J]. J Infrared Millim Waves, 2013, 32(5): 450－455. (in Chinese with English abstract)

[15] 刘康，钱旭，王自强. 基于流形主动学习的遥感图像分类算法[J]. 计算机应用，2013，33(2)：326－328.
Liu Kang, Qian Xu, Wang Ziqiang. Remote sensing image classification based on active learning with manifold structure[J]. Journal of Computer Applications, 2013, 33(2): 326－328. (in Chinese with English abstract)

[16] 丁玲，唐娉，李宏益. 基于 ISOMAP 的高光谱遥感数据的降维与分类[J]. 红外与激光工程，2013，42(10)：2707－2711.
Ding Ling, Tang Ping, Li Hongyi. Dimensionality reduction and classification for hyperspectral remote sensing data using ISOMAP[J]. Infrared and Laser Engineering, 2013, 42(10): 2707－2711. (in Chinese with English abstract)

[17] Tenenbaum J B, De Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319－2323.

[18] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linea embedding[J]. Science, 2000, 290(5500): 2323－2326.

[19] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5－32.

[20] 郭颖. 森林地上生物量的非参数化遥感估测方法优化[D]. 北京：中国林业科学研究院，2011.
Guo Ying. Optimum Non-parametric Method for Forest Above Ground Biomass Estimation Based on Remote Sensing Data[D]. Beijing: Chinese Academy of Forestry, 2011. (in Chinese with English abstract)

[21] 吕杰. 基于机器学习和辐射传输模型的农作物叶绿素含量高光谱反演模型[D]. 北京：中国地质大学，2012.

Lu Jie. Hyperspectral Remote Sensing Inversion Models of Crop Chlorophyll Content Based on Machine Learning and Radiative Transfer Models[D]. Beijing: China University of Geosciences, 2012. (in Chinese with English abstract)

[22] 吴俊强，周激流，何坤，等. 基于 LLE 和 BP 神经网络的人脸识别[J]. 激光杂志：信息科学版，2006，27(5)：71－73.
Wu Junqiang, Zhou Jiliu, He Kun, et al. Face recognition based on LLE and SVM[J]. Journal of Jilin University: Information Science Edition, 2006, 27(5): 71－73. (in Chinese with English abstract)

[23] 袁远，季星来，孙之荣，等. Isomap 在基因表达谱数据聚类分析中的应用[J]. 清华大学学报：自然科学版，2004，44(9)：1286－1289.
Yuan Yuan, Ji Xinglai, Sun Zhirong, et al. Application of isomap for cluster analyses of gene expression data[J]. J T singhua Univ: Sci &Tech, 2004, 44(9): 1286－1289. (in Chinese with English abstract)

[24] 卜育德，潘景昌，陈福强. 基于 Isomap 算法的恒星光谱离群点挖掘[J]. 光谱学与光谱分析，2014，34(1)：267－273.
Bu Yude, Pan Jingchang, Chen Fuqiang. Stellar spectral outliers detection based on Isomap[J]. Spectroscopy and Spectral Analysis, 2014, 34(1): 267－273. (in Chinese with English abstract)

Inversion model for copper content in farmland of tailing area based on visible-near infrared reflectance spectroscopy

Lü Jie, Hao Ningyan, Cui Xiaolin

(*College of Geomatics, Xi'an University of Science and Technology, Xi'an,* 710054*, China*)

Abstract: Heavy metal pollution exists in many mining sites, and heavy metal in soils poses a great potential threat to the environment and human health. Therefore, it is urgent to estimate heavy metals in farmland in tailing areas of mining sites. The goal of this research was to estimate copper content in farmland of a tailing area based on visible-near infrared reflectance spectroscopy. This research took Jinduicheng mine tailings in Shaanxi as the study area. A total number of 288 soil samples were collected at the mining areas. The soil samples were divided into two groups, a training/calibration set (n=252) and an external validation set (n=36) for the Cu estimation model. The soil samples were air dried and passed through a 2 mm sieve. The Cu concentrations in soil were determined through chemical analysis in the laboratory by graphite furnace atomic absorption spectrometry (GB/T17141-1997). The visible-near infrared reflectance spectral measurements of soil Cu concentration were collected using an ASD field spectrometer for the solar reflective wavelengths (350-2500 nm) in the laboratory. The 8 angle probe was used, the distance from the contact probe to the surface of soil samples was set to 1.35 m in order to get the soil spectral in the range of 1 m^2, and each soil sample was achieved 10 spectral measurements. The original reflectance was transformed with a db6 wavelet analysis. The Isomap (Isometrio Mapping) and LLE (Locally Linear Embedding) manifold learning methods were applied to the hyperspectral data of soil for dimension reduction, parameter of k and d was 10 to 50 and 8–15, respectively. Copper concentration in the mine tailing soil was estimated by the method of random forests. The estimated results were compared with the original hyperspectral data and the dimension reduction spectral data. The results showed that the spectral characteristics of the most important values were at the wavelength of 475 802, and 868 nm. The estimation model had a better performance on dimension reduction spectral data set than that on the original spectral data set, and the estimation model achieved coefficient of determination R^2 of 0.7586 on the spectral data set after dimension reduced by Isomap, and the RMSE (root mean square error) was 30.50, the estimation accuracy was better than that on the dimension reduction by LLE, but the accuracy needed to be improved. The results provide a theoretical basis for rapid estimation copper content of farmland soil in the tailing area, and will provide theoretical basis and technological support for controls of mining tailings and mining wasteland and its ecological restoration and reconstruction.

Key words: soils; spectrometry; heavy metals; copper; hyperspectral; manifold learning; random forests