DOI: 10.13758/j.cnki.tr.2016.01.000

母质与土地利用类型对土壤光谱反演模型的影响①

邬登巍[1,2]，张甘霖[1,2*]

(1 土壤与农业可持续发展国家重点实验室(中国科学院南京土壤研究所)，南京 210008；2 中国科学院大学，北京 100049)

摘 要：用可见光近红外(Vis-Near Infrared, VNIR)光谱建立的土壤反演模型可以快速高效测定土壤某些属性，但不考虑土壤自身特点的纯统计模型反演的精度会受到制约。本文研究了母质和土地利用类型对土壤光谱反演模型建立的影响。研究所用集合为 SF(安徽宣城的林地样品)、SP(安徽宣城的水田样品)、DP(安徽定远的水田样品)，结果显示：母质和土地利用类型的差异会显著影响异地模型的适应性，一个地区建立的反演模型不可随便用于母质和土地利用类型不同的其他地区；当异地模型不适用于反演时，可考虑采用精度稍低的全局模型进行预测。

关键词：母质；土地利用类型；光谱；可见光近红外

中图分类号：P237；S-3

可见光–近红外光谱能够快速、高效、无损测定土壤某些属性，但建立精度符合要求的土壤光谱反演模型一直是具有积极现实意义的挑战性工作。一些研究考虑到土壤类型会影响到其光谱特征而将土壤分类的概念引入到土壤光谱的分析研究[1−5]，并从光谱的角度对土壤进行分类或归纳[4−7]，其中有的是纯粹从光谱的差异来进行土壤分类[6]，有的则加入了其他的协变量，如气候变量、地形变量[8]。

光谱模型的建立涉及建模回归算法、土壤属性以及不同研究区域等[1, 10−12]，研究表明本地光谱模型的预测精度总高于异地模型。有研究认为母质会显著影响土壤光谱特征，考虑了土样背景的光谱反演模型将更具解释性[13]。但除母质外，影响土壤属性和土壤光谱特征的因素还有土地利用、生物和时间等[9]，目前综合考虑母质和土地利用类型对光谱模型影响的报道尚甚少。为此，本研究尝试以土壤有机质(SOM)为目标属性[11,14−16]，选择母质和土地利用类型不同的样品，分析母质和土地利用类型对光谱反演模型建立的影响。

1 材料与方法

1.1 样品信息

本研究采用控制变量的思想，挑选 3 个目标集合为研究对象，样品信息如图 1 所示。3 个集合中 SF、SD 和 DP 分别为安徽宣城的林地样品、安徽宣城的水田样品和安徽定远的水田样品，SF 的母质主要是第四纪红黏土以及岩类风化形成的坡积物或残积物，SP 的母质为上述物质经过河流搬运形成的冲积物和河湖相沉积物，DP 的母质则多是黄土类物质的河流冲积物。

为尽可能避免样品数量对建模结果的影响，选择相近个数的样品为目标，最终 3 个集合中分别包含了 25、22 和 22 个土壤剖面，实际调查剖面位置是依据宣城和定远的第二次土壤普查资料[13]中土种的典型剖面位置确定，挖掘土壤剖面(宽 1.2 m × 深 1.2 ~ 1.5 m × 长 3 ~ 3.5 m)，划分发生层，按发生层进行采样，合计分别包含了 106、107 和 108 个发生层样品。

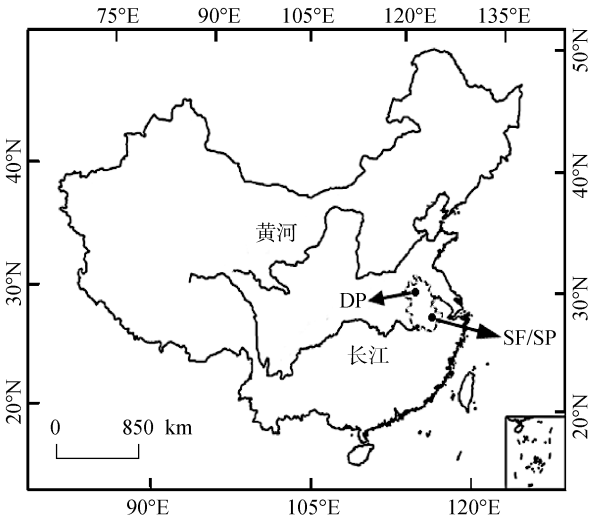
1.2 理化属性和光谱数据测定

土样经室内风干、去杂、研磨过 0.15 mm 筛后用于测定，其中土壤有机质(SOM)测定采用重铬酸钾–硫酸硝化法[17]。使用 Cary 5000 分光光度计采集光谱数据，波段范围为 350 ~ 2 500 nm，步长为 1 nm。测量前，将过 0.15 mm 筛的样品于 45℃烘箱中烘干 24 h，然后置于干燥器中待测。用 Cary 5000 配套样品池进行制样，即将适量样品放在样品池中即可采集光谱数据。经独立实验重复检验，由于样品池法的测量精度高，重复测量数据稳定，因此本研究采用样品

①基金项目：国家自然科学基金项目(41130530)资助。

* 通讯作者(glzhang@issas.ac.cn)

作者简介：邬登巍(1984—)，女，湖北通山人，博士研究生，主要从事土壤光谱研究。E-mail: wdw@issas.ac.cn



(DF：定远水田样品，SF：宣城林地样品，SP：宣城水田样品)

图 1 采样区位置示意图

Fig. 1 Sampling regions

池法进行土壤光谱采集。

1.3 建模方法及模型评估方法

1.3.1 光谱吸收强度提取 1 400、1 900、2 200 nm 为羟基吸收峰，是矿物的主要吸收峰，也是土壤光谱在可见光近红外(Vis-Near Infrared, VNIR)波段最主要、最明显的吸收波段，其信息对光谱解析和建模具有重要意义[18]。提取 1 400、1 900 及 2 200 nm 波段的吸收强度值的具体步骤包括：①将光谱反射率通过 log(1/*R*)转化变为光谱吸收率。②通过对数据的观察后确定，在 1 350～1 450，1 850～2 050，2 130～2 250 nm 波段范围内作光谱基线校正。校正类型为线性消去两端数值，即以两个端点为依据拟合线性的基线，然后将原吸收率减去基线值，于是两端的数值为零。③在基线校正后，以各波段吸收峰处的最大值为波段的吸收强度值。

1.3.2 数据分析 主成分分析 PCR(Principal Component Analysis)和偏最小二乘回归 PLSR(Partial Least Squares Regression)多用于分析光谱数据。本研究用 PCR 计算光谱数据主成分，得到每个样品的光谱数据对应的 PC_SCORE；用 PLSR 对光谱数据和 SOM 数据间建立回归模型。使用 R^2 和 *RMSE* 为衡量统计模型的统计量。数据运算在 UMSCRAMBLER 中完成。

2 结果与分析

2.1 有机质含量分析

3 个样品集合的 SOM 含量统计信息如图 2 所示。结果显示：①SF 为林地样品，由于多年枯枝落叶和根系腐解的影响，其 SOM 总体上比 SP(水田样品)高。②SP 和 DP 虽同为水田，但 SP 经历了长期的双季稻轮作，DP 则是长期的小麦–晚稻轮作，总体上 SP 的秸秆和根系还田生物量相对较高，土壤湿度也相对较高，均有利于 SOM 的积累，导致 SP 的 SOM 含量相对高于 DP。

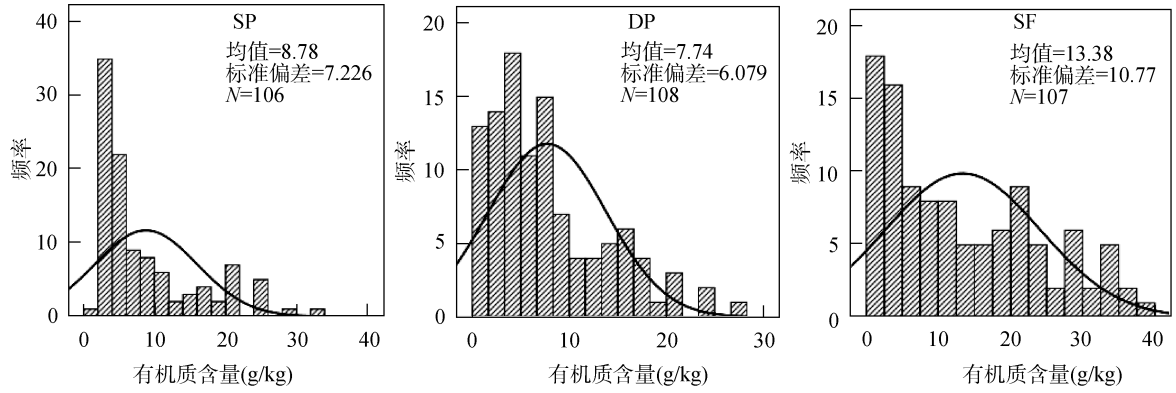


图 2 土壤有机质含量

Fig. 2 Descriptions of SOM contents of studied soil samples

2.2 光谱特性分析

对 3 个样品集合的光谱曲线在 1 400、1 900 和 2 200 nm 附近的吸收峰强度进行提取，由于 3 个波段吸收峰强度信息维度为 1，即相关性很大，所以在结果展示中只选用了 1 400 nm 和 2 200 nm 两个波段的结果，如图 3 所示。在 1 400、2 200 nm 波段的吸收强度比值有两个特点：一是每个集合中的点都沿着一条直线分布；二是 SF 和 SP 集合所沿直线的斜率相近，而 DP 则与两者的斜率相差较大。

不同母质样本的吸收峰强度信息沿不同斜率的直线分布，证明不同母质发育成的土壤由于其母质中矿物类型及其组成的差异，会使其形成的土壤在 1 400、1 900 和 2 200 nm 处展现出不同的吸收特征，而这一特征在经过吸收强度提取后，可以通过如图 3

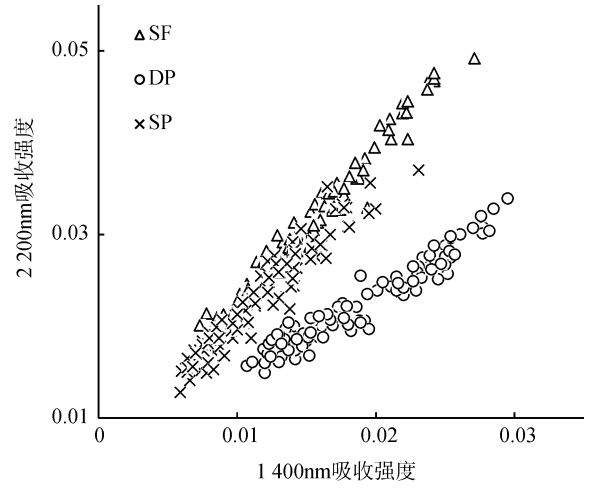


图 3　1 400 nm 和 2 200 nm 附近吸收强度提取结果
Fig. 3　Extraction results of absorption near 1 400 nm and 2 200 nm

的散点图的形式来表现。

除了特征吸收峰的吸收强度分析，常见的对土壤光谱的分析还有 PCA。在对原始光谱进行重采样预处理后，将其进行 PCA 分析，主成份 1 和主成份 2 的分数图见图 4。PCA 结果有两个特点：一是 SF 和 DP 的点有较明显的界线，而 SP 则与 SF 和 DP 集合都有交集；二是图中粗黑线框点 SP-AVG、SF-AVG、DP-AVG 分别代表 SP、SF 和 DP 对应集合中所有点的平均值，SP 与 DP 的均值点非常接近，SF 的均值点与 SP、DP 均值点距离较远。

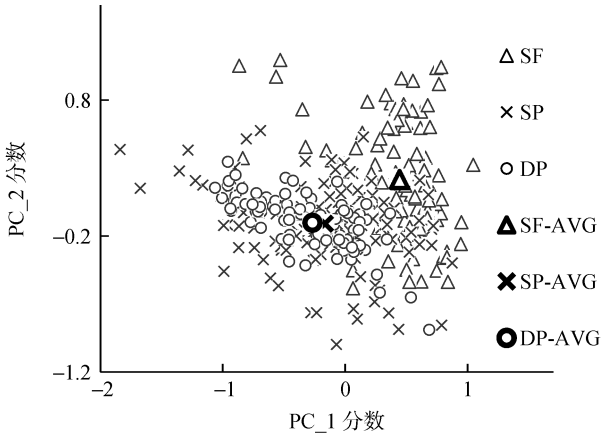


图 4　光谱数据 PCA 结果
Fig. 4　PCA results of spectral data

主成分是以在原变量基础上通过线性组合得来的新变量，目的是降维，以尽可能少的变量来表达尽可能多的变量信息。通常前两个变量能解释原变量 80% 以上的信息。由此，主成份 1 表达的是最共性、最概括的信息，主成份 2 次之。土壤 VNIR 光谱的整体反射率与 SOM 含量相关，而 PC_1、PC_2 代表样本最共性的信息，所以 PC_1、PC_2 携带很多的关于 SOM 的信息。

SP 与 DP 均是水田样品且 SOM 均值相近，无论是有机质的组成形式和含量都比 SF 更接近，所以在 PCA 分析中，SP 和 DP 分布比较接近。

由吸收强度和 PCA 的分析结果可知，土壤光谱对矿物和 SOM 的响应可以不同的形式得到显示，当集合中存在不同母质或不同土地利用类型的土壤时，可能会能以某种形式在光谱中显示出其分异，在建立反演模型时应考虑相应的策略。

2.3　光谱反演模型分析

对于光谱数据的建模算法有过不少研究，即对于同一组光谱数据和属性数据，用不同的算法进行建模，在统一对比参数的基础上(通常是 R^2、*RMSE* 等统计参数)将各种算法得到的结果进行对比，如 MLR、PLSR、PCR、MARS、SVM、RF、BT 和 ANN 等，或者两种算法的联用等[1, 10−12]。研究表明，PLSR 的综合运算能力最高，且没有一种算法能够在大多数的光谱反演运算中超过 PLSR，其虽然是一种经典、不复杂的算法，但在土壤光谱反演中是最稳定和可靠的，故本研究仍选用 PLSR。

2.3.1　模型建立　在 PLSR 的运算过程中，主成分个数根据变量的特点会有所不同。本研究首先对各个数据集在不同主成分个数下的模型精度参数进行了统计(图 5)。结果显示，3 个独立集合与一个 3 合 1 的混合集合，分别在 1 ~ 20 个主成分的设置下进行回归运算，在 10 ~ 15 个主成分的时候 4 个集合都达到最佳回归结果。按照回归精度排序依次为 SF>SP>SF+SP+DP>DP，建模 R^2 分别为 0.94、0.89、0.87、0.84，其中 SF 集合得出的模型回归精度最高，DP 集合的最低。

2.3.2　独立模型验证　为验证各个独立模型的预测精度，用以下两种方法进行计算：一是留一交叉验证(leave-one-out validation)，另一种是用异地模型对 SOM 含量进行反演预测。留一交叉验证的结果可表示在母质、土地利用类型等属性与建模集样本最大程度上相似的情况下得到的预测结果，也可理解为在最理想的采样情况下所能得到的预测精度；异地模型验证是用一个地方的模型对另一地方的样品进行反演预测。图 6 是留一交叉验证的结果，与建模时的精度排序一样，留一交叉验证的结果排序为 SF>SP>SF+SP+DP>DP，其中 R^2 分别为 0.90、0.82、0.81、0.75，*RMSE*(g/kg)分别为 2.35、4.54、3.75、3.03。如果能在光谱库中依照母质、土地利用类型对应的样品形成样本数相当的样品集，则最终最优的结果就是如此。

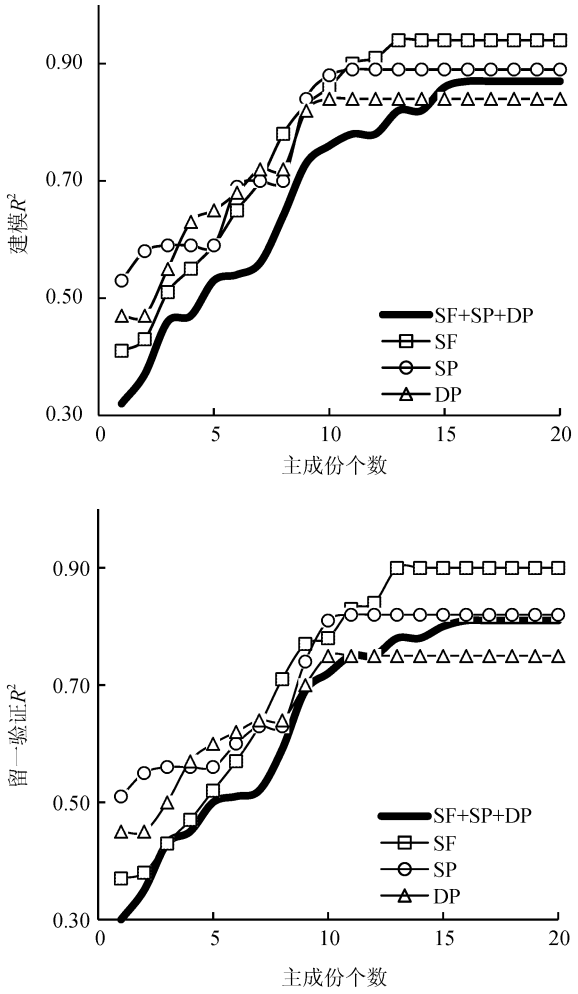


图 5 主成份个数对模型精度的影响
Fig. 5 Effects of principle component numbers on model accuracies

图 7 为异地模型预测结果，用 SF 模型对 SP 的反演结果最好(图 7A，*RMSE* 为 5.54 g/kg)，DP 模型对 SP 的反演结果次之(*RMSE* 为 8.38 g/kg)。其中图 7b、7d、7e 分别出现了不同程度的坐标偏移，图 7C 中低 SOM 含量的样本值被严重高估。从异地模型预测的结果看，无论是母质还是土地类型，都会对反演结果造成很大影响，再一次证明异地模型应用的风险非常大。

另外，图 7A 和 7C 为 SF 和 SP 互作为异地反演模型的预测结果，其结果分别为 6 个模型中最好和最差。这一结果与常规思维所得的结果不同。在计算得出这一系列结果之前，原本的设想中 SF 和 SP 作为两个独立集合，在预测和被预测的关系中应该处于对称的位置，即 SF 模型对 SP 的预测结果会与 SP 模型对 SF 的预测结果相差不大。但最终预测结果中一个 *RMSE* 为 5.54 g/kg，表现尚可；另一个却为 28.42 g/kg，表现极差。若加上土地利用类型来表述，即为林地 SOM 模型可用来反演农田 SOM 含量，而农田 SOM 模型却不适用于林地。对此的解释是，结合 SOM 含量分布频率图，林地样本中 SOM 含量高于 10 g/kg 以上的样品显著多于农地，而农地除了表层 SOM 含

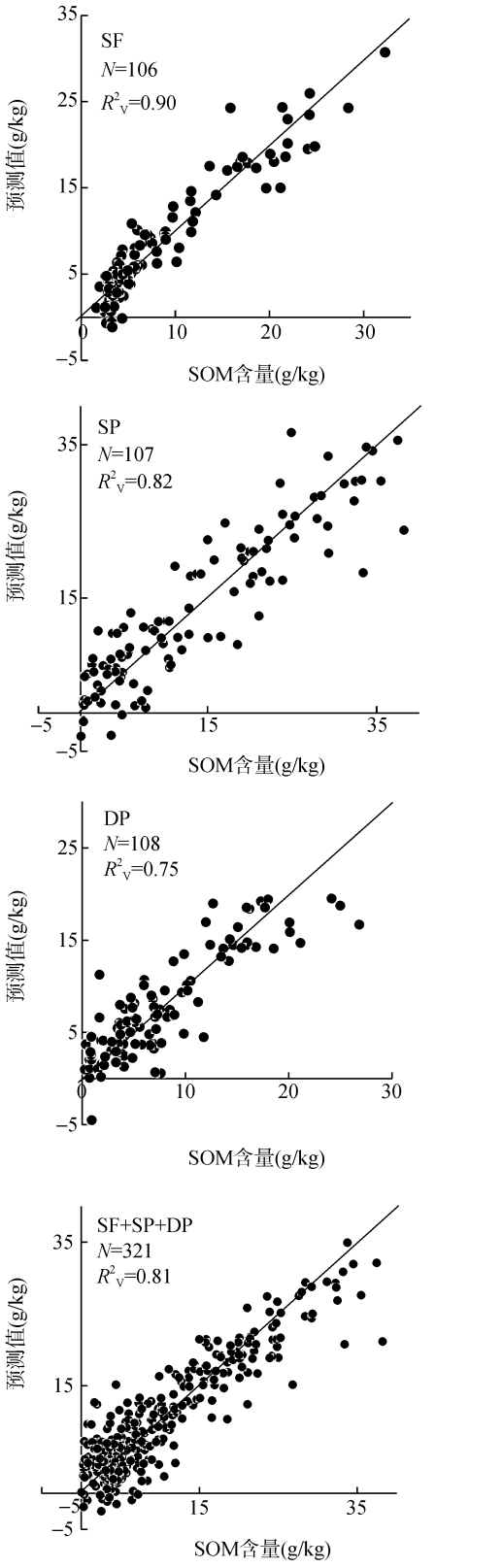
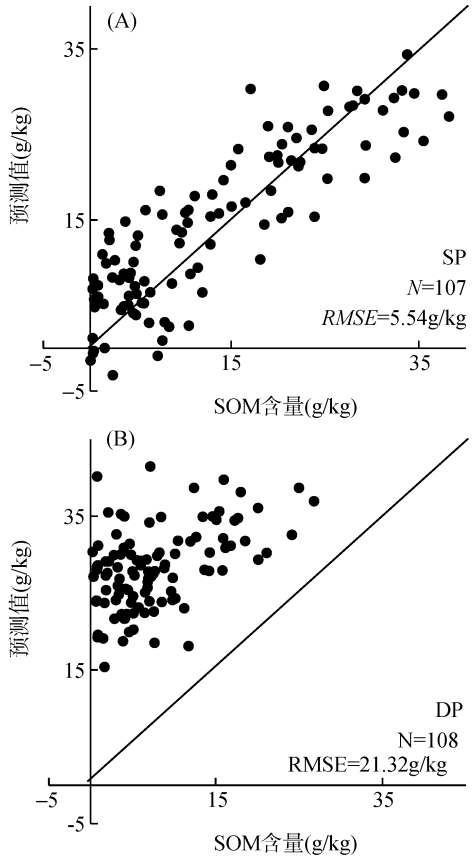
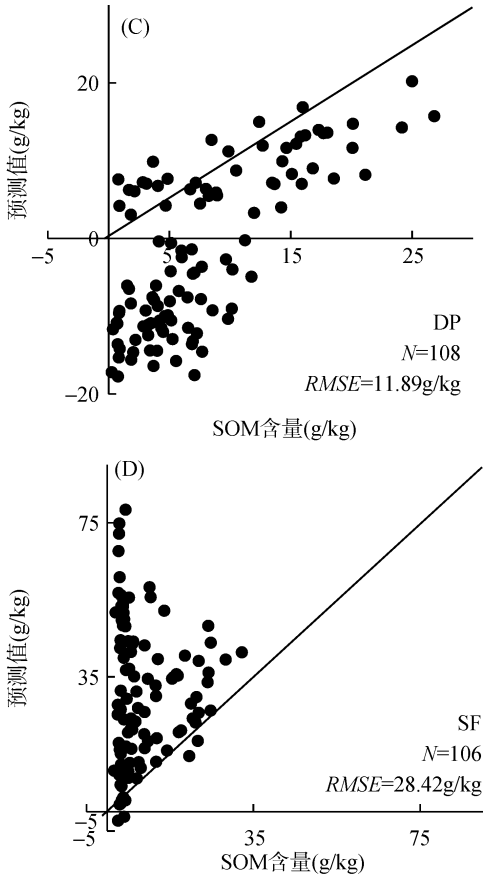


图 6 预测值与测量值的散点图(验证方法为留一交叉验证)
Fig. 6 Scatter plots of predicted and reference values (leave-one-out validation)



SP集合所建立的模型为反演预测模型



DP集合所建立的模型为反演预测模型

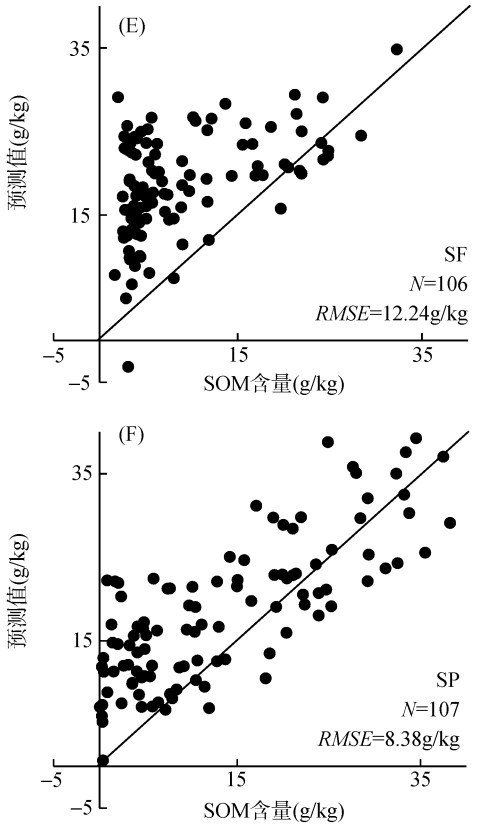


图 7　预测值与测量值的散点图(验证方法为异地模型预测)
Fig. 7　Scatter plots of predicted and measured values (using models developed by other calibration data)

量较高，其他层次基本都低于 10 g/kg，特别是 SOM 含量 2 ~ 4 g/kg 的样本数量约为 35 个，为整个建模集规模的 1/3；另外，农地的 SOM 均值水平也与林地差异较大，农地的为 7.2 g/kg，林地为 10.8 g/kg，农地为林地的 2/3。所以，在 SF 的林地模型中，各个含量的样本的分布较 SP 农田的平均，模型对各个含量都有不错的预测能力；而在 SP 的农田模型中，低含量的样本太多，且整个模型所涉及的阈值范围较窄，所以对 SF 林地集合中预测目标值范围较大的样本进行运算时表现出极度的不适应。再将交叉验证和异地验证的结果进行对比，结果如表 1 所示。唯有当建模集和预测集母质、土地利用类型等条件相同时，才能达到比较理想的预测结果。这表明：一个地区建立的反演模型不可随便应用于母质和土地利用类型不相同的其他区域；当某地的模型能适用于另一个区域时，并不代表反之也能行得通，这不仅与母质、土地利用类型有关，还跟目标属性的分布情况有关。

2.3.3　全局模型验证　　除各个集合建立的独立模型，还可将各种土壤样品放在一起作为一个笼统的建模集而建成全局反演模型。全局模型通常尽可能多的包含各种类型的土壤样本，使各种类型的数据特征都能在最终模型中有所体现。将 3 个集合所有样本放在

一个集合中作为建模集，然后将 3 个独立集合的样本分别代入并反演得出的预测结果如图 8A 及表 1 所示。从 *RMSE* 的结果可得，全局模型的结果比交叉验证的精度高。交叉验证的结果显示，只有预测集中的样本的母质、土地利用类型等属性与建模集中样本一致时才能得到相同精度的交叉验证结果，但是此时全局模型的精度已超过原独立模型，其可能性只有一个——建模样本数量较大从而提高了模型精度，因为独立模型的样本数平均为 107，而全局模型建模集的规模为 321。为了验证这一猜想，将原样本集中的样本以剖面为单位，随机选择 1/3 作为新的全局模型建模集。其验证反演结果如图 8B 及表 1。当样本数统一到同一水平，交叉验证的结果明显好于 1/3 全局模型的反演结果。证明，样本数量确实能在一定程度上显著影响反演模型的精度。

表 1 交叉验证和全局模型验证的 RMSE 结果(g/kg)
Table 1 RMSE results of cross-validation and global model validation

验证方法	SF	SP	DP
交叉验证	2.35	4.54	3.03
全局模型验证	2.38	3.74	2.98
1/3 全局模型验证	3.31	4.85	3.46

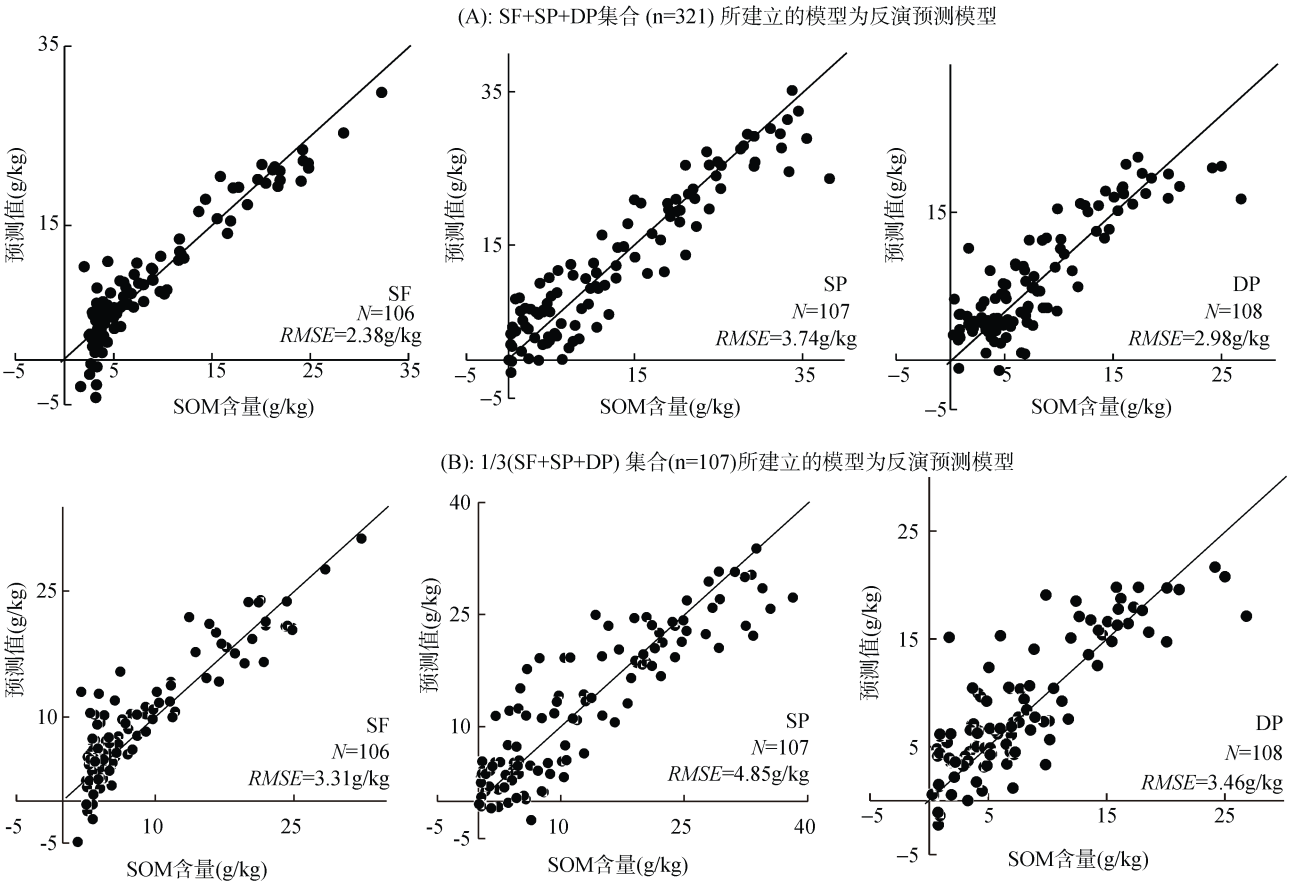


图 8 预测值与测量值的散点图(全局模型预测)
Fig. 8 Scatter plots of predicted and measured values (using global model)

3 结论

本文研究了母质和土地利用类型对光谱反演模型建立的影响，所用的 3 个集合中，SF 和 SP 有共同的母质，SP 和 DP 有相同土壤利用类型。研究结果表明：母质和土地利用类型的差异会显著影响异地模型的适应性，一个地区建立的反演模型不一定适用于母质和土地利用类型不同的其他地区；某类土地利用类型建立的模型也不一定适用于另外一类土地利用类型；当异地模型不适用于反演时，可考虑采用精度稍低的全局模型进行预测，在预测精度上的排序为本地模型>全局模型>异地模型。

参考文献：

[1] Tian Y C, Zhang J J, Yao X, et al. Laboratory assessment of three quantitative methods for estimating the organic matter content of soils in China based on visible/near-infrared reflectance spectra[J]. Geoderma, 2013(202/203): 161–170

[2] Morgan C L S, Waiser T H, Brown D J, et al. Simulated in

situ characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy [J]. Geoderma, 2009(151): 249–256

[3] Rawlins B G, Kemp S J, Milodowski A E. Relationships between particle size distribution and VNIR reflectance spectra are weaker for soils formed from bedrock compared to transported parent materials[J]. Geoderma. 2011, 166(1): 84–91

[4] Du C, Linker R, Shaviv A. Identification of agricultural Mediterranean soils using mid-infrared photoacoustic spectroscopy[J]. Geoderma, 2008, 143(1/2): 85–90

[5] Awiti A O, Walsh M G, Shepherd K D, et al. Soil condition classification using infrared spectroscopy: A proposition for assessment of soil condition along a tropical forest-cropland chronosequence[J]. Geoderma, 2008 (143): 73–84

[6] Ramirez-Lopez L, Behrens T, Schmidt K, et al. Distance and similarity-search metrics for use with soil vis–NIR spectra[J]. Geoderma, 2013, 199: 43–53

[7] Oliveira J F, Brossard M, Vendrame P R S. Soil discrimination using diffuse reflectance Vis–NIR spectroscopy in a local toposequence [J]. Comptes Rendus Geoscience. 2013, 345(11/12): 446–453

[8] Viscarra Rossel R A, Chen C. Digitally mapping the information content of visible–near infrared spectra of surficial Australian soils[J]. Remote Sensing of Environment, 2011, 115(6): 1443–1455

[9] 黄昌勇. 土壤学[M]. 北京: 中国农业出版社[M], 2000: 6–7

[10] Vohland M, Besold J, Hill J, et al. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy[J]. Geoderma. 2011, 166(1): 198–205

[11] Melendez-Pastor I, Navarro-Pedre O J, Gómez I, et al. Identifying optimal spectral bands to assess soil properties with VNIR radiometry in semi-arid soils[J]. Geoderma, 2008, 147(3/4): 126–132

[12] Summers D, Lewis M, Ostendorf B, et al. Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties [J]. Ecological Indicators, 2011, 11(1): 123–131

[13] Wu D W, Zhang G L. Study on paddy soil chronosequences based on visual-near infrared diffuse reflectance spectra (to be published in Spectroscopy and Spectral Analysis, 2015)

[14] Mcdowell M L, Bruland G L, Deenik J L, et al. Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy[J]. Geoderma, 2012, 189/190: 312–320

[15] Denis A, Stevens A, Van Wesemael B, et al. Soil organic carbon assessment by field and airborne spectrometry in bare croplands: accounting for soil surface roughness[J]. Geoderma, 2014, 226/227: 94–102

[16] AiChi H, Fouad Y, Walter C, et al. Regional predictions of soil organic carbon content from spectral reflectance measurements[J]. Biosystems Engineering, 2009, 104(3): 442–446

[17] 张甘霖, 龚子同. 土壤调查实验室分析方法[M]. 北京: 科学出版社, 2012

[18] 吴昀昭, 田庆久, 季峻峰, 等. 土壤光学遥感的理论、方法及应用[J]. 遥感应用, 2003(1): 40–47

Effects of Parent Materials and Land Use Types on Inversion Models by Using Soil Spectral Data

WU Deng-wei[1,2], ZHANG Gan-lin[1,2*]

(1 *State Key Laboratory of Soil and Sustainable Development*(*Institute of Soil Science Chinese Academy of Sciences*), *Nanjing* 210008, *China*; 2 *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: Using visible near-infrared (Vis-Near Infrared, VNIR) spectral data to build inversion model is a rapid and nondestructive potential method to measure soil properties. However, the models based on pure statistical methods are weak in explaining the information of soil properties. This paper studied the effects and strategies of parent materials and land use types on the established spectral inversion model. 3 sample sets of SF (forestry soils from Xuanchens), SP (paddy soils from Xuancheng) and DP (paddy soils from Dingyuan) were used, where, SF and SP with same parent material, SP and DP with same land use type. The results showed that the differences of parent materials and land use typse could affect the suitability of the off-site model significantly. It is possibly unreliable to use a model in a different region with different parent materials or land use types, however, the global model with expense of the prediction accuracy can be used when the off-site model does not work well.

Key words: Parent material; Land use type; Spectroscopy; Vis-Near Infrared