第42卷，第5期
2022年5月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 5, pp1614-1619
May, 2022

基于土壤光谱库和光谱相异度的局部模型构建

彭青青[1]，陈颂超[2]，周明华[3]，李 硕[1*]

1. 华中师范大学地理过程分析与模拟湖北省重点实验室，湖北 武汉 430079
2. 浙江大学杭州国际科创中心，浙江 杭州 311200
3. 中国科学院、水利部成都山地灾害与环境研究所山地表生过程与生态调控重点实验室，四川 成都 610041

摘 要 掌握土壤在空间和时间上的表征至关重要。土壤可见-近红外(Vis-NIR)光谱可以估算土壤有机碳(SOC)等属性，与传统的实验室理化分析相比，光谱技术能有效实现土壤信息的快速获取。土壤光谱库为建立经验模型提供了大量具有丰富变异性和多样性的样本作数据基础。但受限于库中土壤样本的异质性和模型的适应性，通常区域或局部尺度模型的稳健性欠佳。已有的研究主要通过目标样本部分入库的方式改善库的性能，但影响了光谱技术的低成本优势。该研究在不入库的前提下基于土壤光谱的相异度，探究经典距离算法结合土壤光谱库构建局部预测模型的可行性，并比较分析局部模型样本容量对预测精度的响应。基于全球土壤光谱库(GSSL)的677个土柱，从每个国家随机取十分之一的土柱(97个)组成局部目标测试集(Test)，其余580个作土壤光谱库(SSL)。分别采用欧氏距离(ED)、马氏距离(MD)、和光谱角(SAM)来分别度量Test与SSL间的光谱相异度并生成距离矩阵。按距离矩阵的前0.04%，0.05%，0.1%，0.2%，0.3%，0.4%，0.5%，1%和5%从SSL中提取与Test最相似的光谱样本构建共计9个容量的局部建模集(Local)，使用偏最小二乘回归(PLSR)建立Vis-NIR和SOC含量的预测模型并通过Test验证模型精度，通过光谱的主成分空间考察并解释各种距离算法下Local的“容量-精度”变化。结果表明，在待测样本不入库的情况下，三种距离算法构建的Local模型相较于全局模型的预测精度均有一定提升，但三者的“容量-精度”的拐点存在显著差异。SAM兼顾了光谱的波形和幅度因此较MD、ED更具优势；其前0.2%比例的Local不仅预测精度最优，且用于建模所需的样本容量最少。因此认为，SAM法更适用于从土壤光谱库中构建局部模型，距离矩阵的前0.2%可作为局部模型的容量参考。

关键词 光谱库；相异度；距离矩阵；容量；偏最小二乘

中图分类号：TP79 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)05-1614-06

引 言

联合国粮农组织(FAO)于2020年9月牵头发起了全球土壤实验室网络(GLOSOLAN)的倡议，旨在帮助各国及地区充分利用光谱分析技术，高效、经济地获得更为详实的土壤信息，最终实现土壤资源的可持续利用和管理[1]。土壤可见-近红外(visible near-infrared，Vis-NIR)光谱中包含了大量的分子及化学组分信息，不同组分的跃迁能级差不同令土壤吸收光谱曲线存在不同的吸收特征，以此来定量分析如土壤有机碳(soil organic carbon，SOC)等属性[2-4]。拥有丰富且多样化信息的土壤光谱库有助于解决土壤数据短缺的问题，以服务于评估和监测从区域到全球的日益增长的土壤信息需求。全球(如GSSL)[5]、洲[6](欧洲的LUCAS)和国家尺度[7-8](中国的CSSL，巴西BSSL)的光谱库相继建立，使Vis-NIR光谱的预测能力得到了更全面的评价，并着眼于克服大量土壤样本带来的异质性以改善经验模型从全局预测局部样本的稳健性。

当前主要有局部入库和优化子集两种策略。局部入库策略将一定数量的局部样本的光谱及相关理化属性入库并重新校正再去预测的“Spiking”及加权优化[9]，或约束入库数量以平衡预测精度与成本同时提升的“RS-Local”[10]等方法相继改善了库在局部应用的效果。但上述策略仍需对一定数量的样本进行实验室理化分析，削弱了光谱技术的优势及光谱库

收稿日期：2021-08-25，**修订日期**：2021-12-11

基金项目：国家自然科学基金项目(41601370)和华中师范大学青年团队项目(CCNU19TD002)资助

作者简介：彭青青，1995年生，华中师范大学城市与环境科学学院硕士研究生 e-mail：1085873892@qq.com

*通讯作者 e-mail：shuoguoguo@zju.edu.cn

建设的必要性。优化子集策略侧重于衡量待测样本与库样本间的关系，如基于聚类思想的光谱学习算法[6]和基于样本距离的地理加权回归算法[7]等将全库优化为数个子集进行建模。虽然该策略也都提高了土壤光谱库预测局部样本的精度，但分析受黑盒效应未能深入解释，仅停留在光谱表征及精度指标的对比层面。

随着土壤光谱库建设经费的不断投入，土壤光谱库所涵盖的样本特征亦将更加丰富。当库已包含相对足够的局部样本特征信息时，能否直接以距离思想从库里构建局部模型从而改善预测精度？因此，本工作的目的是：(1)考察经典的距离算法从光谱库组建局部建模集的可行性；(2)比较局部模型的容量与预测精度的响应并从光谱维度解释。同时，以此工作作为响应 FAO 帮助提升国家及地区层面的光谱技术应用服务能力的倡议。

1　实验部分

1.1　供试数据

所用数据来自国际土壤参考和信息中心(International Soil Reference and Information Center)提供的全球土壤光谱库(Global Soil Spectral Library，GSSL)。数据包括土壤的 Vis-NIR 漫反射光谱以及常规实验室理化分析方法获得的 40 余种土壤属性。本研究从库中筛选出同时含有 Vis-NIR 及 SOC 含量共计 677 个土柱(3755 个土壤发生层样本)作为研究数据(图 1)。其中，Vis-NIR 光谱由≤2 mm 粒径的风干研磨土样测得，波段范围 350～2 500 nm，采样间隔 10 nm。Viscarra Rossel 等给出了该数据更丰富的细节[5]。

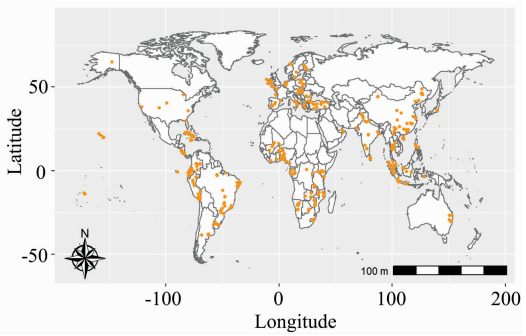

图 1　全球土壤光谱库中的 677 个样点分布

Fig. 1　Locations of 677 sites from the Global Soil Spectral Library

1.2　集合划分

使用 Kennard-Stone(KS)法[7]从每个国家挑选十分之一的土柱样本共 97 个组成一个光谱多样的“局部”目标作测试集(Test，588 个样)，即小样本容量的局部而非地理空间局部[11]；其余 580 个土柱留作包含相对足够局部样本特征信息的土壤光谱库(SSL，3 167 个样)。如表 1 所示，SSL 的 SOC 含量均值为 1.20%但跨度较大，且标准差及变异系数均较大，其与 SSL 中土壤样品的土地利用类型、母质、气候和地形的高度一致性密切相关。Test 的 SOC 含量均值(1.13%)与 SSL 接近。

表 1　Test 与 SSL 的 SOC 含量特征统计

Table 1　Statistics of SOC content in the Test and SSL datasets

Data set	Sample size	Minimum/%	Maximum/%	Mean/%	Standard deviation/%	Coefficient of variation/%
Test	588	0.01	14.28	1.13	1.78	158
SSL	3 167	0.00	60.00	1.20	2.78	232

1.3　光谱预处理

去除光谱首尾两端噪声较大波段之后保留 400～2 450 nm 范围作后续研究，即每条共计 206 个波段。采用 Savitzky-Golay(SG)卷积平滑法(2 阶 3 窗口)和一阶微分进行预处理。既可有效保留光谱的变化信息，又能将相互干扰或相互重叠的吸收峰分离，同时起到光谱基线校正以消除基线漂移。主成分分析(principal components analysis，PCA)可从海量数据中提取少数几个变量以代表样本关键的变异特征，在不损失信息的情况下达到降维目的。

1.4　光谱相异度度量

使用欧式距离(euclidean distances，ED)、马氏距离(mahalanobis distances，MD)和余弦距离来衡量样本之间光谱维的相异度。ED 和 MD 距离值越小表示越相似，反之越相异；余弦距离也称光谱角(spectral angle mapper，SAM)，夹角余弦值越接近最大值 1 说明两者越相似，值越小则越相异。

度量过程如图 2 所示：Test(容量 n)中每一待测样都与 SSL(容量 m)经光谱相异度计算得到共计 n 列 m 行个距离值，

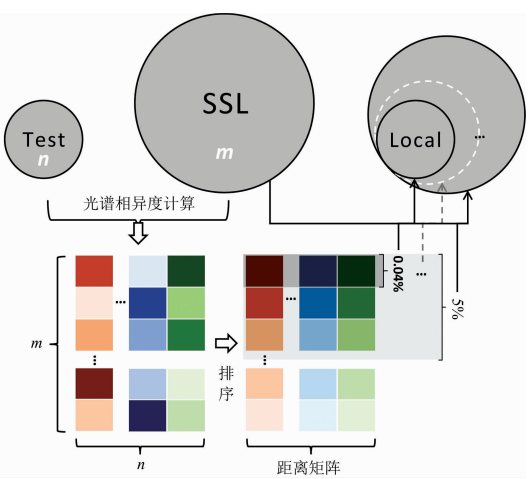

图 2　光谱相异度计算流程

Fig. 2　The scheme of calculating the spectral dissimilarity

每列按值大小排序后组成距离矩阵($n\times m$)。以矩阵第一行也即 Test 容量为 Local 容量起点(约占矩阵前 0.04%)，提取

SSL 中的相应样本并剔除重复后构建 Local 集；矩阵前 5% 几近 SSL 容量，作为 Local 容量的终点。此外，还增加了 0.05%，0.1%，0.2%，0.3%，0.4%，0.5%和 1%以充分考察局部建模集的容量与预测精度之间变化的关系。本研究的 m 为 3 167，n 为 588，ED 或 MD 的距离矩阵按值升序排，而 SAM 则按降序排。

1.5 建模方法及评价指标

使用偏最小二乘回归(partial least-squares regression，PLSR)构建预测 SOC 含量的光谱模型。为避免过度拟合，采用留一法交叉检验确定模型最优潜变量数。模型精度评价指标采用决定系数(coefficient of determination，R^2)和均方根误差(root mean square errors，RMSE)。通常，优良的模型具有较大的 R^2 和以及较小的 RMSE 值。当精度差别不显著时，更少的 Local 容量则更优。因 R^2 和残差预测偏差(residual prediction deviation，RPD)存在依赖性[12]，故二者可选其一。所有的数据处理和建模预测均在软件 R 中实现。

2 结果与讨论

2.1 光谱分析

图 3 显示 SSL 和 Test 的光谱吸光度在 400～800 nm 范围随波长增加而急速下降，之后趋于平缓；在 1 400，1 900 和 2 200 nm 三处吸收峰主要由 O－H 官能基团的伸缩振动或转角振动所致。谱线主要表征了 SSL 和 Test 的特征共性，掩盖了两者特征差异，不论波形还是变化范围都未见显著差异，这是因 KS 法使子集的特征保持最大的变异[13]。SSL 和 Test 光谱的前三个主成分共同解释了超 99%的总体变异，PC1，PC2 和 PC3 分别代表具有土壤中富含赤铁矿、有机物以及赤铁矿、伊利石和赤铁矿[14]。两者的 PC1 特征向量曲线不论是波形还是峰或谷的起伏都明显不同，相应的 PC2 或 PC3 则波形更相似，只在峰谷高低上呈现不同。因此，PC2 和 PC3 包含了 SSL 和 Test 相似的特征信息。

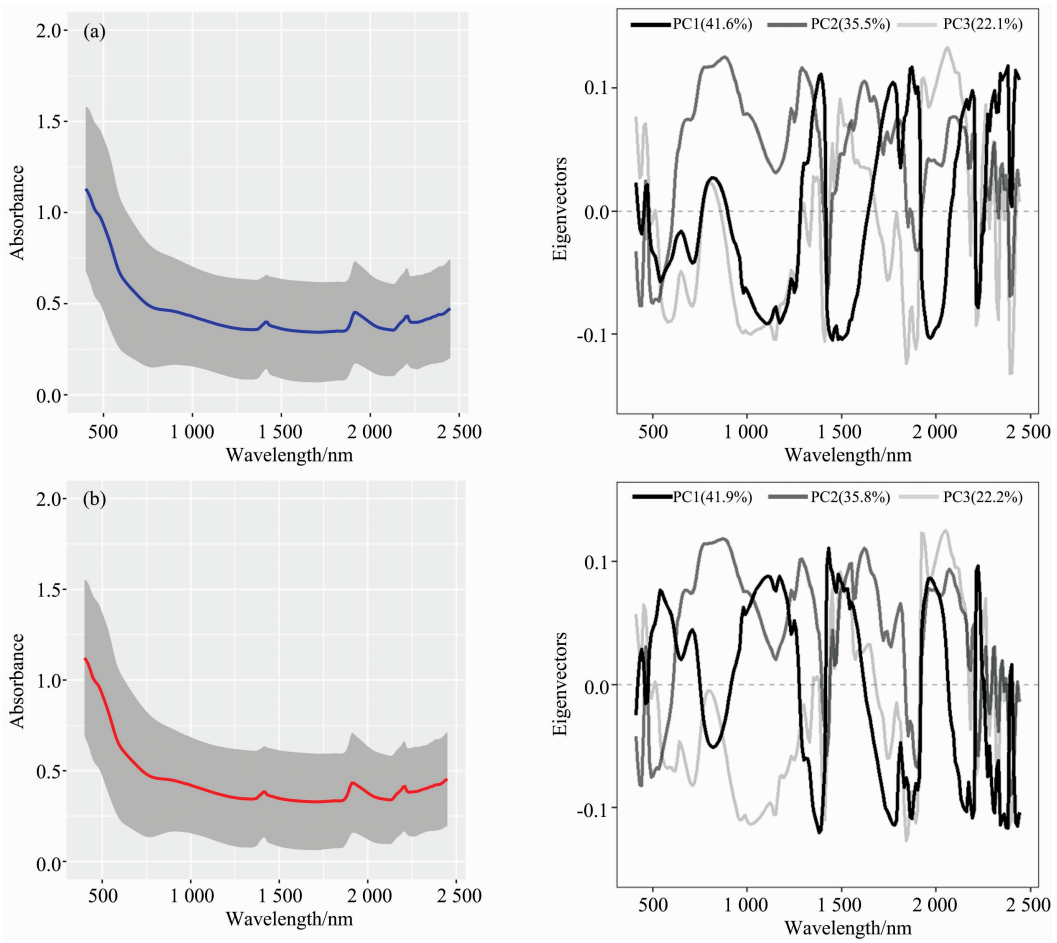

图 3 SSL(a)和 Test(b)平均光谱和标准差范围以及前三个主成分特征向量

Fig. 3 Averaged absorbance with standard deviation of SSL (a) and Test (b) and their eigenvectors of the first three PCs

2.2 预测精度对比

以 SSL 建立全局模型的预测精度做基线(R^2 = 0.47，RMSE=1.43%)，ED，MD 和 SAM 分别构建 Local 预测 Test 的 SOC 精度如图 4(a)，(c)和(e)所示。每种比例所对应的样本数即为 Local 容量。随着容量的增大，三者的精度变化规律均呈现出逐渐变好到逐渐变差直至与 SSL 精度趋同的趋势。ED 和 MD 在前 0.1%时精度达到最优，且 ED 略优于 MD；SAM 在前 0.1%时就已呈现出较 ED 和 MD 更优的

预测精度和更少的 Local 容量，并在前 2%时精度达到最优（$R^2=0.59$，RMSE＝1.14%），而此时的 ED（$R^2=0.46$，RMSE＝1.48%）或 MD（$R^2=0.50$，RMSE＝1.39%）均不可与 SAM 相比且精度都显著变差。ED 或 MD 表现不如 SAM 的情况也出现在相关文献中[15-16]。

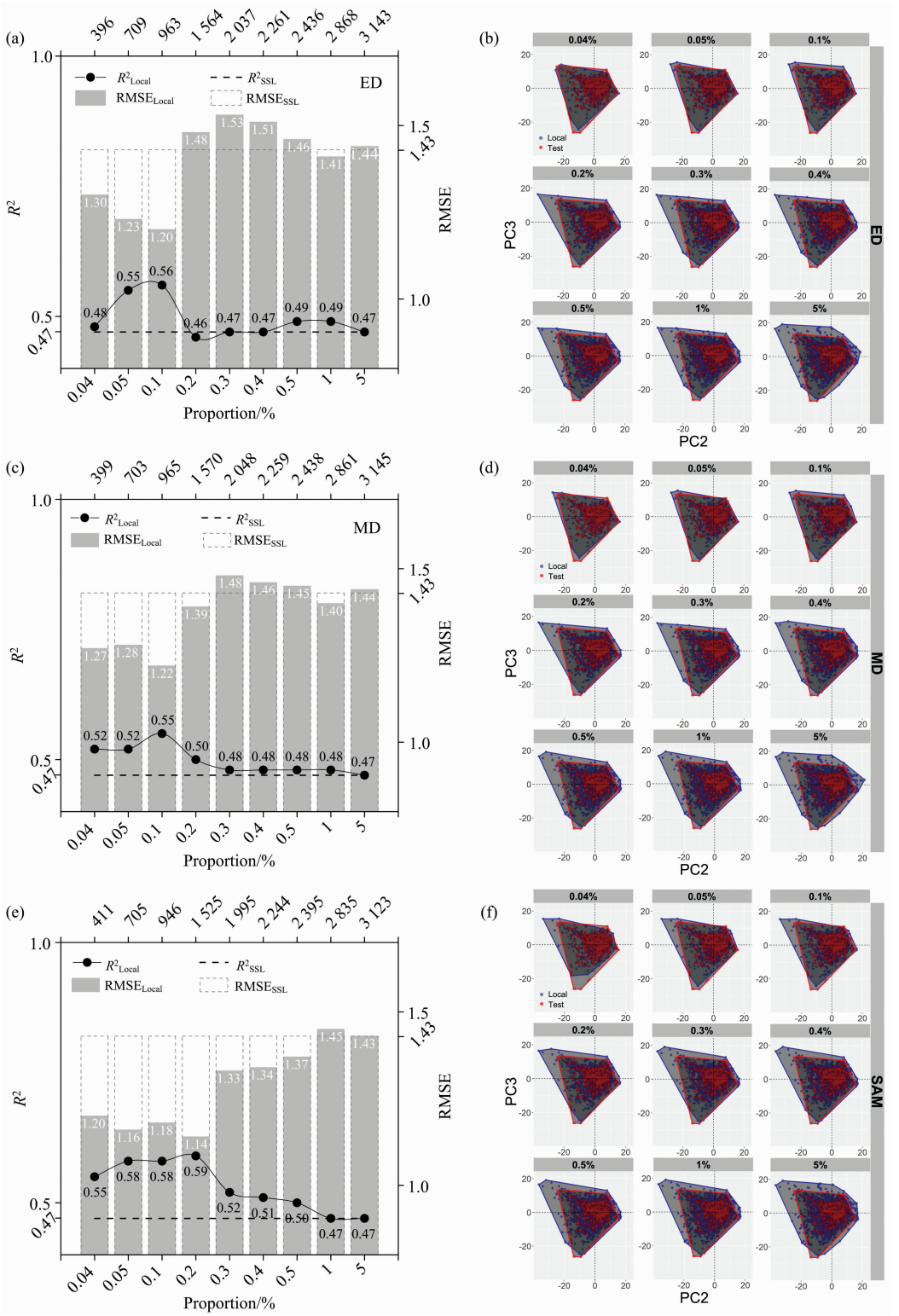

图 4　ED，MD 和 SAM 构建九种容量 Local 建模集的预测精度结果及主成分分析图

Fig. 4　Prediction accuracies of 9 sample sizes of Local datasets and PCA plots developed by ED, MD, and SAM

2.3 主成分空间分异

将 Local 的光谱在 PCA 空间的分布投影到 Test 的相应 PCA 空间以供对比，图 4(b)，(d)和(f)分别为 ED，MD 和 SAM 在九种比例下 PC2 和 PC3 的空间分布。其中，蓝色和红色框分别代表 Local 和 Test 的边界，灰色阴影代表分布范围。Test 的红点位置和范围恒不变。随着比例逐渐增大，代表 Local 的蓝点逐渐增多表明容量随之增大，且受点位分布影响，范围也随之形变且变大。

如前 0.04%的 ED，此时其 Local 只有 396 个样本，因此其阴影范围小于 Test；同样的情形也出现在前 0.04%的 MD(399)和 SAM(411)。ED 和 MD 都从 0.1%[图 4(b)和(d)第 1，2，4 象限]开始超越 Test，且两者的 Local 范围以及与 Test 的重合情况都极相似，这不仅同两者相似的容量吻合还同两者在 0.1%相似的精度结果吻合。SAM 则是从 0.2%开始超越 Test[图 4(f)第 1 和 4 象限]。总之，以上规律和图 4 中三者所呈现的“容量-精度”拐点相一致。

Local 容量的拐点过后，同 Test 更相异的样本也更多地被加入到 Local 中，是预测精度降低的主要原因[9]。尤其，前 0.2%下的 SAM 同 ED 和 MD 相比，在第 1 和 4 象限的范围更贴合 Test，而后两者则更加远离。鉴于前 0.2%下的 SAM 精度优于 ED 和 MD，表明 1 和 4 象限的样本对模型精度的影响更显著。前 0.1%下 ED 和 MD 的范围结合拐点也进一步证实了这一点。

值得注意的是不论 Local 的容量如何增大，三种方法的 Local 范围都未能覆盖第 3 象限的 Test。这表明，虽然能通过光谱相异度从 SSL 中筛选出同 Test 最相似的 Local，但其对 Test 全部特征的可达性是受限的。为此，我们将在今后的研究中考察如异构多源信息融合、多层随机验证策略等同 Local 精度提升间的响应。

3 结 论

通过度量 Vis-NIR 的光谱相异度考察了从土壤光谱库构建局部模型预测 SOC 含量的能力。基于全球光谱库数据先创建了一个 Test 和一个包含其主要特征的 SSL。ED，MD 和 SAM 用于度量 Test 与 SSL 相异度并分别构建了九种容量的 Local，再使用 PLSR 建立 SOC 预测模型并评价。研究表明，三种算法在“容量-精度”的拐点存在显著差异，SAM 法较 MD，ED 相比具有明显优势，其前 0.2%比例下的 Local 不仅预测精度最优，且用于建模的样本容量也最少。SAM 兼顾了光谱的波形和幅度，与计算欧氏空间距离的 ED 和计算协方差距离的 MD 相比，更适合用于从光谱库中构建局部模型以改善全局模型的精度。

References

[1] Food and Agriculture Organization of the United Nations (FAO). 2020. http://www.fao.org/global-soil-partnership/glosolan/soil-analysis/dry-chemistry-spectroscopy/en/.

[2] Yang M, Chen S, Li H, et al. Land Degradation & Development, 2021, 32(3): 1301.

[3] SHI Zhou, XU Dong-yun, TENG Hong-fen, et al(史 舟，徐冬云，滕洪芬，等). Progress in Geography(地理科学进展), 2018, 37(1): 79.

[4] LI Shuo, LI Chun-lian, CHEN Song-chao, et al(李 硕，李春莲，陈颂超，等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2021, 41(4): 1234.

[5] Viscarra Rossel R A, Behrens T, Ben-Dor E, et al. Earth-Science Reviews, 2016, 155: 198.

[6] Chen S, Xu H, Xu D, et al. Geoderma, 2021, 400: 115159.

[7] Ji W, Li S, Chen S, et al. Soil and Tillage Research, 2016, 155: 492.

[8] Demattê J A M, Dotto A C, Paiva A F S, et al. Geoderma, 2019, 354: 113793.

[9] Guerrero C, Wetterlind J, Stenberg B, et al. Soil and Tillage Research, 2016, 155: 501.

[10] Lobsey C R, Viscarra Rossel R A, Roudier P, et al. European Journal of Soil Science, 2017, 68(6): 840.

[11] Seidel M, Hutengs C, Ludwig B, et al. Geoderma, 2019, 354: 113856.

[12] Minasny B, McBratney A B, Malone B P, et al. Advances in Agronomy, 2013, 118: 1.

[13] Xu D, Chen S, Xu H, et al. Environmental Pollution, 2020, 263: 114649.

[14] Viscarra Rossel R A, Webster R, Bui E N, et al. Global Change Biology, 2014, 20(9): 2953.

[15] LI Hong-da, LI De-cheng, ZENG Rong(李宏达，李德成，曾 荣). Acta Pedologica Sinica(土壤学报), 2021, 58(5): 1224.

[16] Zeng R, Zhang J P, Cai K, et al. PLOS ONE, 2021, 16(3): e0247028.

Developing of Local Model From Soil Spectral Library With Spectral Dissimilarity

PENG Qing-qing[1], CHEN Song-chao[2], ZHOU Ming-hua[3], LI Shuo[1*]

1. Key Laboratory for Geographical Process Analysis & Simulation of Hubei Province, Central China Normal University, Wuhan 430079, China
2. ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou 311200, China
3. Key Laboratory of Mountain Surface Processes and Ecological Regulation, Institute of Mountain Hazards and Environment, Chinese Academy of Sciences, Chengdu 610041, China

Abstract It is vital to understand the characteristics of soils and their distribution in space and over time. Spectroscopy in the visible-near-infrared (Vis-NIR) can estimate soil properties (e. g., SOC). Compared with traditional laboratory physical and chemical analysis, spectral technology enables the practical acquisition of soil information rapidly. The development of a soil spectral library (SSL) can provide large amounts of soil data with variability and diversity for empirical calibration. Calibrations derived with these SSLs, however, at the very least, help to improve the robustness of spectroscopic models at regional and local scales due to high soil heterogeneity and model adequateness. Previous studies usually put several target samples into SSL, called spiking; however, the cost-efficiency of spectral techniques was offset more or less. Without spiking samples, we aim to explore the feasibility of developing a local model by constraining the SSL with spectral dissimilarities using classical distance methods. The response between the capacity of the local model with prediction accuracy was also compared and analyzed. In this study, we built a local test set (Test) with the amount of spectral variation from 97 cores, divided by one-tenth of each country from the global soil spectral library (677 cores), and the remaining 580 cores were used as the SSL. We used Euclidean distance (ED), Mahalanobis distance (MD) and Spectral Angle Mapper (SAM) to measure the spectral dissimilarity between Test and SSL and to generate the distance matrix. For each method, nine Local subsets were selected and developed by selecting the spectra of SSL, which were considered similar to the Test. The selection based on the first 0.04%, 0.05%, 0.1%, 0.2%, 0.3%, 0.4%, 0.5%, 1% and 5% of the distance matrix. The statistical models were built to predict SOC concentrations from the spectra by partial least-squares regression. We decomposed the spectra using principal components analysis (PCA) to identify those variables of Local derived from ED, MD and SAM. Our results showed that all the Local models developed by the three distance algorithms without spiking samples still can improve the accuracy compared to the global one, but the inflection points of a sample size of Local with accuracy were significantly different. The SAM considers the waveform and amplitude of the spectrum, so it has more advantages than MD and ED. Its Local, with the first 0.2% ratio, performed the best prediction accuracy, also required the least samples for modeling. We conclude that SAM is more suitable for developing local models from SSL. The first 0.2% of the distance matrix can be used as a reference for the capacity of the local model.

Keywords Spectral library; Dissimilarity; Distance matrix; Sample size; PLSR

(Received Aug. 25, 2021; accepted Dec. 11, 2021)

* Corresponding author