DOI：10.11766/trxb202003010082

李宏达，李德成，曾荣. 基于光谱相似性匹配的土壤有机碳估算[J]. 土壤学报，2021，58（5）：1224–1233.

LI Hongda，LI Decheng，ZENG Rong. Estimation of Soil Organic Carbon Based on Spectral Similarity Matching[J]. Acta Pedologica Sinica，2021，58（5）：1224–1233.

基于光谱相似性匹配的土壤有机碳估算*

李宏达[1]，李德成[2]，曾　荣[1†]

（1. 南京信息工程大学地理科学学院，南京 210044；2. 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）

摘　要：全球土壤光谱库的建立，为利用可见-近红外光谱预测土壤属性提供了参考集，如何从光谱库中挑选合适的建模集以实现对局部地区土壤有机碳的高精度预测，是一个值得研究的问题。本研究基于欧氏距离、马氏距离和光谱角三种光谱相似性指数，探索利用全球光谱库预测局部地区土壤有机碳的有效策略，并比较了不同光谱相似性指数、不同建模集数量及不同建模方法对预测精度的影响。研究表明：（1）三种相似性算法较全局模型均极大提升了预测精度，其中光谱角预测精度稍高，最佳预测精度为 R^2=0.75，RPD=1.73；（2）建模数量对建模精度有较大影响，三种算法的最佳建模集数量范围在本研究中约为 400～500（0.71<R^2<0.75，1.56<RPD<1.73）；（3）针对光谱角算法，建模集数量较少（<50）时，赋值预测精度较高（R^2>0.6，RPD>1.4）；建模集数量较多（>50）时，PLSR 建模预测精度较高（R^2>0.6，RPD>1.4）。

关键词：全球土壤光谱库；光谱匹配；光谱角匹配；偏最小二乘回归；土壤有机碳

中图分类号：S151.9　　**文献标志码：**A

Estimation of Soil Organic Carbon Based on Spectral Similarity Matching

LI Hongda[1], LI Decheng[2], ZENG Rong[1†]

(1. *School of Geographical Sciences*, *Nanjing University of Information Science & Technology, Nanjing* 210044, *China*; 2. *State Key Laboratory of Soil and Sustainable Agriculture*, *Institute of Soil Science*, *Chinese Academy of Sciences*, *Nanjing* 210008, *China*)

Abstract: 【Objective】 The objective of this study is to explore effective strategies for predicting soil organic carbon (SOC) in local areas with high precision based on the spectral similarity indexes in the global spectral library. It goes specifically as follows: (1) to compare different spectral similarity indexes obtained separately with three different similarity matching algorithms (Euclidean Distance, Mahalanobis Distance and Spectral Angle Mapper) in effect on prediction accuracy; (2) to compare calibration sets different in size in effect on prediction accuracy; and (3) to compare different modeling strategies (PLSR modeling and the assignment strategy) in effect on prediction accuracy. 【Method】From the global spectra library a total of 245 China samples were selected to form a prediction set and the remaining 3 537 samples a reference set. From each spectrum in the prediction set, different numbers of similar spectra (5, 10, 20, 30, 40, 50, 100, 150, 200, 250, 300, 400 and 500) were picked out

* 国家重点研发计划项目（2017YFC0803807）、中国烟草总公司贵州省公司科技项目（201910）、南京信息工程大学人才启动经费共同资助 Supported by the National Key Research and Development Program of China（No.2017YFC0803807），China Tobacco Corporation Guizhou Provincial Company Science and Technology Project（201910） and the Startup Foundation for Introducing Talent of NUIST

† 通讯作者 Corresponding author，E-mail：rzeng@ nuist.edu.cn

作者简介：李宏达（1997—），男，河北承德人，主要从事土壤近地遥感研究。E-mail：hongdalll@126.com

收稿日期：2020–03–01；收到修改稿日期：2020–06–23；网络首发日期（www.cnki.net）：2020–10–19

separately with the three different similarity matching algorithms for comparison between the spectra selected by the different similarity algorithms. Based on the reference sets of different sizes selected by the different algorithms, PLSR models were built to predict SOC contents, and effects of the similarity matching algorithms and size of the modeling set on prediction accuracy were evaluated using R^2, RMSE and RPD. Then a similarity matching algorithm with the highest prediction accuracy was selected and on such a basis, comparison was performed between the different modeling strategies in effect on prediction accuracy: Ⅰ, Prediction with PLSR modeling; Ⅱ, Prediction with direct assignment. 【Result】Compared with the overall model, the three similarity algorithms greatly improved prediction accuracy. Among the three, the SAM model was a bit higher than the other two in prediction accuracy (R^2=0.75, RPD=1.73). The low prediction accuracy might be attributed to the wide distribution of the soil samples in the global soil spectral library that caused marked variation. Size of the modeling sets did have a great impact on modeling accuracy, and the optimal size for the three similarity algorithms varied in the range of 400~500 ($0.71<R^2<0.75$, $1.56<RPD<1.73$). The SAM model needed a smaller-sized modeling set (<50), but remained higher in accuracy using the assignment prediction strategy ($R^2>0.6$, RPD>1.4). However, when the modeling set was big in size (>50), the PLSR modeling strategy was higher in prediction accuracy ($R^2>0.6$, RPD>1.4).【Conclusion】Compared with the global model, the models based on the three spectral similarity indices all significantly improve SOC prediction accuracy. In general, the spectral angle algorithm is slightly better than Euclidean distance and Mahalanobis distance; type of the similarity algorithm, size of the modeling set and method of the modeling all have a great impact on precision of the SOC prediction.

Key words: Global Soil Spectral Library; Spectral matching; Spectral angle mapper; PLSR; Soil organic carbon

土壤是地球关键带的核心之一[1]，定量研究土壤发生学、土壤属性和功能、土壤发生的时空变化，是当今全球变化背景下土壤科学研究面临的挑战[2-3]。土壤有机碳含量是评价土壤质量演变的重要指标之一，它直接影响土壤肥力和作物产量，作为衡量土壤肥力的重要指标[4]，其定量化快速监测成为精准农业研究的热点[5]。传统的测定方法费时、费力、费钱，且污染环境，可见-近红外光谱的发展为该问题提供了解决方法[6]，土壤可见-近红外光谱是土壤理化参数的综合反映，结合化学计量模型已实现了对有机碳等诸多土壤属性的快速估算。目前，建立了局地、国家、洲际乃至全球等不同尺度的土壤光谱库，但前人的研究表明，基于光谱库采用全局建模预测的效果并不理想[7]。如何从光谱库中挑选合适的建模集以实现对任一局部地区土壤有机碳的高精度预测，是一个值得研究的问题。

前人的研究表明，利用光谱相似性匹配挑选建模集可以提高预测精度[8]，光谱相似性匹配是依据某一特定的光谱相似性指数来判定未知光谱与参考光谱之间的相似度[9]。该方法提高预测精度存在如下假设：即相似的光谱具有相似的理化属性，依据光谱相似性匹配可以挑选光谱维度上的局部建模集（local model）。光谱相似性指数在遥感影像分类中的应用较多，但是利用光谱相似性匹配来预测土壤属性的研究较少[10]。光谱相似性匹配算法有多种，常用的有欧氏距离（Euclidean distance，ED）、马氏距离（Mahalanobis distance，MaD）、光谱角匹配（Spectral angle mapper，SAM）等[11]，不同匹配算法所挑选的光谱会存在差异，进而也会影响模型预测精度，前人研究较少关注不同相似性匹配算法对局地土壤有机碳预测精度的影响。如魏昌龙等[10]采用光谱角匹配（SAM)、偏最小二乘回归（PLSR）和 SAM-PLSR 三种方法预测土壤有机质（SOM）和阳离子交换量（CEC），其研究表明光谱角匹配结合 PLSR 算法极大地提高了 SOM（R^2=0.89，RPD=3.00）和 CEC（R^2=0.91，RPD=3.06）的预测精度，也大大降低了建模样本的数量，但该研究基于土壤类型单一、土壤属性变异不大的小区域，对于土壤类型复杂、土壤属性变异大的区域是否适用有待研究。

在利用光谱相似性匹配算法挑选建模集时，建模集数量的确定也尤为重要。以往研究多采用简单数值比例来确定预测集与建模集的大小，比如 1∶1、1∶3，这种划分缺乏一定的科学依据[12]。前人的研究表明，通过合适的建模集挑选，仅使用少量的样本数量，或建模集较小时，也能取得较高的预测精度。如 Peng 等[13]利用丹麦土壤光谱库预测农田土壤有机碳，仅从土壤光谱库中提取了 30 个样本，也得

到了较高的预测精度，但该研究并未对最优建模集做具体探究。Zeng 等[14]探索利用区域土壤光谱库预测局地土壤有机质的有效策略，也并未研究建模集最优数量。

利用相似性匹配算法挑选建模集后，可以采取两种不同的策略建模：（1）建模集结合化学计量模型对未知样本进行预测；（2）将挑选的相似性样本属性的均值赋值给未知样本。魏昌龙等[15]的研究结果表明,光谱相似的土壤，其部分理化属性也具有一定的相似性（如 SOM 和 CEC），在该研究中两种建模策略均取得了较高的预测精度。但针对大区域、大尺度的光谱库，何种研究策略更优还有待探究。

综上所述，前人的研究主要存在如下可改进之处：（1）主要针对较小的研究区、较小的光谱库[13]；（2）缺少不同光谱相似性指数影响模型预测精度的比较研究[16]；（3）相似性光谱组成的建模集大小有待进一步探索。

因此，本研究旨在基于光谱相似性指数，探索利用全球光谱库[17]预测局部地区土壤有机碳的有效策略，以实现对局部地区有机碳的高精度预测。研究目标如下：（1）比较不同的光谱相似性指数对建模精度的影响；（2）比较不同建模集数量对建模精度的影响；（3）比较不同建模策略（建模预测与赋值预测）对建模精度的影响。

1 材料与方法

1.1 全球土壤光谱库

全球土壤光谱库由国际土壤参比与信息中心（ISRIC）建立，覆盖来自非洲、亚洲、欧洲、北美洲和南美洲等 58 个国家的土壤样本，挑选了其中测定了光谱及有机碳的样本，共计 3 782 个[17]。其中 245 个中国区域的样本作为预测集，3 537 个非中国区域样本作为参考光谱库，用于建模集挑选及后续的预测研究。光谱测量采用 FieldSpec FR（Analytical Spectral Devices, Boulder, CO）光谱仪记录，波长范围为 350～2 500 nm，采样间隔为 1 nm。光谱采集在暗室中进行，将约 20 g 风干研磨过 2 mm 筛的土样放入直径 7.4 cm 的玻璃培养皿内，样品高度约为 1 cm，采用 4.5 W 卤素灯作为光源。为了减少数据冗余，在 350～2 500 nm 范围内，每 10 nm 求取平均值，每条光谱共 216 个波段数据。

1.2 光谱预处理

首先将光谱反射率转换为吸收率，通过公式 Lg（1/R）进行转换，其中 R 为光谱反射率。然后对吸收率采用 Savitzky-Golay（SG）卷积平滑一阶求导方法[18]进行平滑去噪，其光谱预处理在软件 R 3.2.5 中完成。

1.3 光谱相似性指数

在进行光谱相似性匹配计算时，分别使用欧氏距离、马氏距离[19]和光谱角算法[20]挑选相似样本，为探究不同数量建模集对建模精度的影响，本文将建模集大小（即所挑选的相似光谱的数量）设定为 13 个等级：5、10、20、30、40、50、100、150、200、250、300、400 和 500，便于探究不同建模数量对有机碳预测精度的影响。

1.4 建模方法及模型验证

本研究采用的建模方法为经典的偏最小二乘回归法。偏最小二乘回归法（Partial least squares regression，PLSR）是一种基于因子分析的多元统计数据分析方法，1965 年由 Herman Wold 首次提出[21]。本研究对 PLSR 模型采用五折交叉验证进行精度评定，交叉验证的方法可在所有样品参与建模的基础上，较好地评估模型精度。

PLSR 因子数量的选择根据五折交叉验证最小的均方根误差来确定，中国区域数据作为预测集，非中国区域数据作为建模集。模型预测精度的评价指标包括决定系数（R^2）、预测均方根误差（RMSEp）和相对分析误差（Relative percent deviation，RPD）。

R^2 越大，RMSEp 越小，RPD 越大，预测效果越佳。R^2 越接近 1 时，说明拟合效果越好；根据 Chang 等[22]对 RPD 的划分：当 1.0<RPD<1.4 时，模型预测能力较差；当 1.4<RPD<1.8 时，模型能够用来做相关性评估；当 1.8<RPD<2.0 时，模型可以用于定量预测；当 2.0<RPD<2.5 时，可进行较好的定量预测；当 RPD>2.5 时，说明模型预测效果较佳。

1.5 建模策略

本研究采用两种建模策略：第一种是 PLSR 建模预测，针对欧氏距离、马氏距离和光谱角算法挑选出来的相似光谱，利用 PLSR 建模来估算预测集中的土壤有机碳含量；第二种策略是赋值预测，利用相似性匹配算法挑选出相似样本，将相似样本土

壤有机碳含量的平均值直接赋值给预测集中的未知样本。赋值预测建立于如下假设：光谱相似的土壤，其土壤有机碳的含量也较为相似。

2　结果与讨论

2.1　土壤光谱库概况及光谱曲线特征

全球土壤光谱库中土壤有机碳的相关统计特征如表 1 所示。建模集土壤有机碳含量范围为 0～60 g·kg^{-1}，平均值约为 1.19 g·kg^{-1}，含量分布极其不均衡。这是由于光谱库样本较多且分布涉及全球，造成土壤有机碳含量标准差以及变异系数较大。预测集土壤有机碳含量分布范围为 0～6.03 g·kg^{-1}，标准差和变异系数相对建模集而言较小，但变异系数也较高（119%），这是因为中国国土面积辽阔，土壤类型多样，造成土壤有机碳含量差异较大。

本研究为探索土壤有机碳含量与光谱曲线形态的关系，对预测集有机碳含量由小到大进行四等分[23]，四个等级的有机碳含量平均值分别为 0.10 g·kg^{-1}、0.27 g·kg^{-1}、0.58 g·kg^{-1} 和 1.71 g·kg^{-1}，其对应的光谱曲线平均值如图 1 所示。

有机碳含量越高，其整体反射率越低，可见光波段反射率曲线分异较小，随着波长增加，分异逐渐变大。虽然有机质含量为 0.27 g·kg^{-1} 和 0.58 g·kg^{-1} 的光谱反射率差异不大，但 0.58 g·kg^{-1} 光谱平均反射率仍然低于 0.27 g·kg^{-1} 光谱平均反射率。在近红外光谱波段的 1 400 nm，1 900 nm 和 2 200 nm 处存在明显吸收峰，通常被认为与黏土矿物中所含的水分子和羟基有关[24]。

2.2　土壤光谱库全局建模

为了便于建模精度的比较分析，在用光谱匹配方法选取合适建模集之前，采用参考光谱库中的所有光谱运用 PLSR 方法对预测集中的土壤有机碳进行全局建模预测，并评估其预测精度。全局建模预测精度低（R^2=0.3，RPD=0.77），并不能较好地预测土壤有机碳含量。主要原因可能是建模集样本数量较多，且全球土壤光谱库样本来自于全球各个区域，参考集样本与预测集样本在地理位置、理化性质等方面均存在较大差异，光谱库中光谱曲线以及有机碳含量差异较大，才导致建模精度不高。这与邬登巍和张甘霖[6]的研究结果相似，通常情况下全局建模的预测精度不高。因此，准确预测局部地区土壤有机碳含量需要筛选合适的建模集，下文中将采用欧氏距离、马氏距离和光谱角算法来挑选合适的建模集。

表 1　全球光谱库土壤有机碳统计特征

Table 1　Soil organic carbon statistical features of the global soil spectral library

样本集 Sample set	样本数量 Sample size	最小值 Min /（g·kg^{-1}）	最大值 Max /（g·kg^{-1}）	平均值 Mean /（g·kg^{-1}）	标准差 Std /(g·kg^{-1}）	变异系数 CV/%
预测集 Prediction set	245	0	6.03	0.67	0.80	119
参考光谱库 Reference set	3 537	0	60.00	1.19	2.62	220

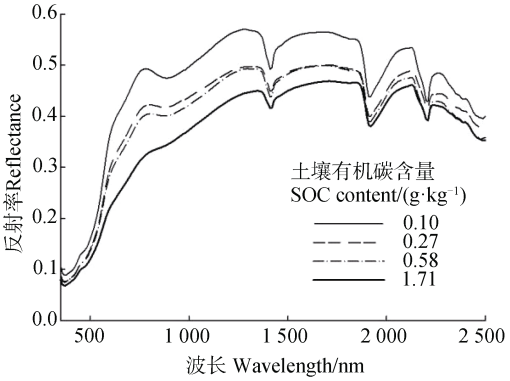

图 1　预测集四个等级有机碳含量的平均光谱

Fig. 1　Average reflectance of the soil samples of four levels of SOC content in the prediction set

2.3　不同光谱匹配方法预测土壤有机碳含量

将以上三种光谱匹配方法挑选出的 13 个数量等级的建模样本分别作为建模集进行 PLSR 建模预测土壤有机碳含量，其预测精度如下：

表 2 展示了基于三种相似性光谱匹配算法及不同建模集数量下土壤有机碳的预测精度。基于欧氏距离匹配所建立的模型，决定系数 R^2_p 范围介于 0.40～0.75，RPD 范围介于 1.27～1.73；基于马氏距离匹配所建立的模型，决定系数 R^2_p 范围介于 0.38～0.72，RPD 范围介于 1.23～1.67；基于光谱角匹配所建立的模型，决定系数 R^2_p 范围介于 0.35～0.75，RPD 范围介于 1.21～1.73。使用上述三种匹配方法，

表 2 不同相似性匹配算法及不同建模集数量下土壤有机碳的预测精度

Table 2 SOC prediction accuracy relative to similarity matching algorithm and size of the modeling sets

建模集数量 Size of modeling sets	欧氏距离 ED			马氏距离 MaD			光谱角算法 SAM		
	R^2_p	RMSEp /（g·kg^{-1}）	RPD	R^2_p	RMSEp/（g·kg^{-1}）	RPD	R^2_p	RMSEp/（g·kg^{-1}）	RPD
5	0.40	0.63	1.27	0.38	0.65	1.23	0.35	0.65	1.21
10	0.50	0.57	1.40	0.49	0.58	1.38	0.32	0.67	1.19
20	0.46	0.60	1.34	0.52	0.57	1.39	0.47	0.61	1.30
30	0.53	0.60	1.33	0.48	0.64	1.25	0.46	0.62	1.28
40	0.57	0.63	1.26	0.46	0.66	1.20	0.56	0.59	1.35
50	0.51	0.65	1.23	0.55	0.60	1.32	0.59	0.57	1.38
100	0.56	0.63	1.26	0.62	0.56	1.42	0.68	0.53	1.50
150	0.59	0.60	1.32	0.60	0.58	1.38	0.71	0.50	1.58
200	0.62	0.56	1.41	0.70	0.55	1.45	0.63	0.59	1.34
250	0.60	0.60	1.33	0.68	0.52	1.52	0.69	0.52	1.54
300	0.57	0.65	1.21	0.71	0.49	1.62	0.70	0.53	1.49
400	0.71	0.51	1.56	0.71	0.50	1.59	0.72	0.48	1.64
500	0.75	0.46	1.73	0.72	0.48	1.67	0.75	0.46	1.73

尽管只挑选极少量相似的光谱（例如 5 条），相比于全局建模（RPD=0.77），预测精度也得到了极大的提升，RPD 提升至 1.20 左右。

整体而言，在建模集数量较少时（<50），三种方法的预测精度均不高（RPD<1.4），而建模集数量较多时，光谱角算法略优于其他两种算法。从 R^2_p、RMSEp 和 RPD 随建模集数量的变异来看，建模数量大小对模型精度有较大的影响，有关建模集数量对建模精度的影响将在下文中深入讨论。

为探究不同相似性光谱指数所挑选的光谱有何差异，本研究在预测集中挑选出 3 条具有明显差异的土壤有机碳含量光谱曲线，其土壤有机碳含量分别为 0.02 g·kg^{-1}、0.65 g·kg^{-1} 和 6.03 g·kg^{-1}，提取出其所匹配的最相似的 5 条光谱进行对比分析，比较三种相似性匹配算法所挑选的光谱有何异同之处。如表 3 所示，无论是对低、中、高有机碳含量，欧氏距离与马氏距离挑选出来的相似样本重合率高达 80%～100%；而 SAM 算法所挑选的光谱与欧氏距离、马氏距离有较大差异，相似度仅为 10%～20%。这是由于算法的不同，欧氏距离和马氏距离的算法相似，均是计算两点间的空间距离，而 SAM 算法则是比较两光谱曲线的余弦夹角，故差异较大。

三种匹配方法所挑选出来的土壤光谱有机碳含量差异也较大，比如针对低土壤有机碳含量的样本（0.02 g·kg^{-1}），ED 算法所挑选出的 5 条相似性光谱中，有两条光谱有机碳含量（0.07 g·kg^{-1} 和 0.08 g·kg^{-1}）与目标样本相近，其他光谱有机碳的含量（0.12～0.38 g·kg^{-1}）远高于目标样本。针对中高有机碳含量样本的匹配，也存在类似的情况，这反映了光谱的相似性与土壤有机碳含量的相似性并不完全一致。这是因为光谱曲线的反射率高低、曲线形态由多种因素控制，不仅是土壤有机碳，还受其他土壤属性的影响，如颗粒组成、氧化铁含量、碳酸钙含量等[25]。造成这一结果的另一原因在于，所匹配的建模集样品来自于除中国外的全球各地，光谱本身的差异性大，与预测集样品在光谱、理化属性上均存在较大差异。

图 2 分别展示了低（0.02 g·kg^{-1}）、中（0.65 g·kg^{-1}）、高（6.03 g·kg^{-1}）有机碳含量下三种相似性匹配算法所挑选的光谱曲线。

针对低土壤有机碳含量的相似光谱，光谱角算法所匹配光谱的反射率均高于目标光谱，虽然光谱

表 3　不同相似性算法匹配光谱的比对

Table 3　Comparison between similarity algorithms in number of matching spectra

土壤有机碳含量 SOC contents / (g·kg^{-1})	欧氏距离 ED		马氏距离 MaD		光谱角匹配 SAM	
	相似光谱序号 Sample ID of similar spectra	相似样本 SOC 含量 SOC contents of similar samples/ (g·kg^{-1})	相似光谱序号 Sample ID of similar spectra	相似样本 SOC 含量 SOC contents of similar samples/ (g·kg^{-1})	相似光谱序号 Sample ID of similar spectra	相似样本 SOC 含量 SOC contents of similar samples/ (g·kg^{-1})
0.02	839	0.12	846	0.18	2 019	0.08
	970	0.07	970	0.07	2 604	0.08
	971	0.08	971	0.08	2 668	0.23
	2 042	0.38	2 042	0.38	3 040	0.18
	3 045	0.22	3 045	0.22	3 045	0.22
0.65	77	0.18	77	0.18	77	0.18
	1 468	1.19	1 421	0.26	2 526	0.65
	2 483	0.27	2 483	0.27	2 544	1.1
	2 526	0.65	2 526	0.65	2 873	1.81
	3 141	1.28	3 141	1.28	3 141	1.28
6.03	456	2.25	456	2.25	474	9.11
	788	3.97	788	3.97	2 527	0.67
	909	4.72	909	4.72	2 610	0.11
	1 590	0.23	1 590	0.23	2 729	13.86
	3 103	3.82	3 103	3.82	3 103	3.82

反射率存在较大差异，但光谱曲线形态较为相似。欧氏距离和马氏距离所匹配光谱中，有两条光谱曲线反射率较为接近，但曲线形态差异较大。

针对中土壤有机碳含量的相似光谱，通过欧氏距离和马氏距离方法匹配的光谱几乎完全相同。整体来看，三种算法所匹配光谱的反射率均低于目标光谱。光谱角算法所匹配的一条相似性光谱，其整体反射率明显低于其他光谱，这也反映了光谱角算法的特质，着重于形态差异而非反射率高低。

针对高土壤有机碳含量的相似光谱，欧氏距离和马氏距离算法匹配的光谱完全一致，其中三条匹配曲线的反射率高于参考光谱，两条匹配曲线的反射率低于参考光谱，无论是从反射率大小还是形态来看，已知光谱与匹配光谱间均存在较大差异。光谱角匹配的光谱从形态特征上与参考光谱更加相似，这也与光谱角计算算法的内涵相一致，光谱角所挑选的相似光谱其反射率均高于目标光谱。

无论是低、中、高有机碳含量的光谱，相比于目标光谱，匹配光谱的相似程度均不是太高，因而基于此的模型预测精度也比较低(1.21<RPD<1.27)。魏昌龙[15] 研究中使用欧氏距离和光谱角算法所挑选的光谱曲线与参考光谱拟合度很高，是因为其土壤样本来自同一研究区域 (宣城市), 样本本身的差异性很小；而本研究中之所以相似度较低，是因为土壤光谱库中土壤样本的分布广，涉及全球土壤光谱样本，光谱本身分异较大。

2.4　不同建模集数量对土壤有机碳含量预测精度影响

在使用欧氏距离、马氏距离和光谱角算法挑选出建模样本之后，建模样本数量的差异极大地影响

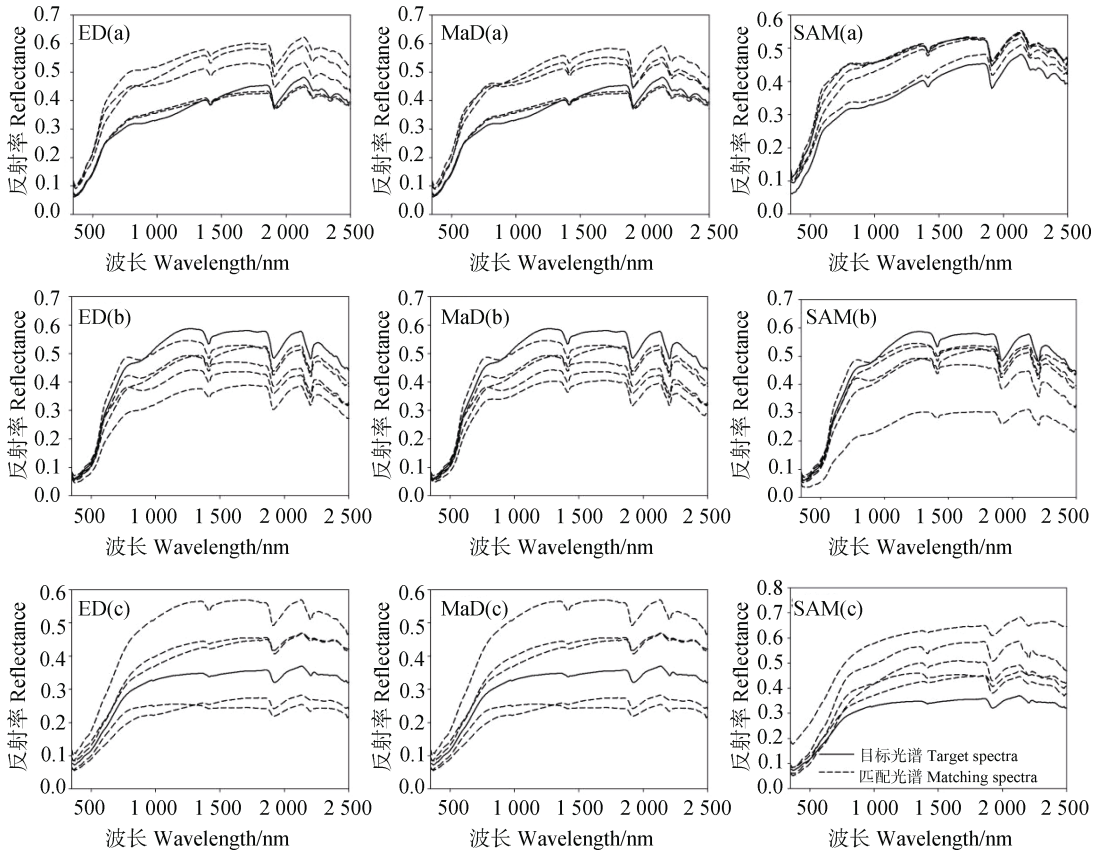

图 2 低（a）、中（b）、高（c）土壤有机碳含量样本所匹配的最相似的 5 条光谱

Fig. 2 Five most similar spectra of the soil samples in matching relative to SOC content, low（a）, medium（b）and high（c）

了模型的预测精度，图 3 直观展示了三种匹配算法下 13 个等级建模数量对有机碳预测精度的影响。

如图 3 所示，建模集数量的差异对建模精度有较大的影响。从整体来看，随着建模集数量的增多，三种匹配方法的预测精度均呈现上升趋势，并且相比于全局建模，预测精度提升显著。

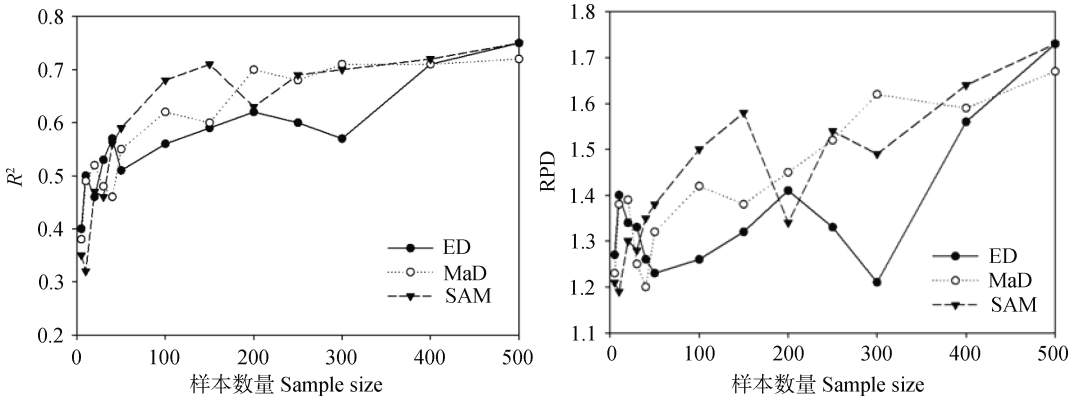

图 3 不同匹配方法不同建模集数量下的土壤有机碳预测精度

Fig. 3 SOC prediction accuracy of the model relative to matching strategy and size of the modeling set

针对欧氏距离，建模集数量小于 100 时，R^2 呈上升趋势，但 R^2<0.6、RPD<1.4，此时所建立的模型并不能很好地预测土壤有机碳；当建模集数量大于 100，小于 300 时，R^2 基本处于 0.6 附近，但 RPD<1.4，模型预测能力较差；当建模集数量大于 300，小于 500 时，R^2>0.6、RPD>1.4,可以较好地预

测有机碳含量，该方法下最优建模集数量范围为400～500。但由图上趋势来看，还需进一步探究建模集数量为 600、700 甚至更多的情形以及更细致的等级划分。

针对马氏距离，建模集数量小于 100 时，与欧氏距离的预测结果相似，R^2 呈上升趋势，但 RPD 和 R^2 均较低，土壤有机碳预测精度不高；当建模集数量大于 100，小于 500 时，R^2>0.6、RPD>1.4，有机碳含量预测精度较高，由图 3 可知马氏距离的整体预测精度均优于欧氏距离，其最优建模集数量范围也为 400～500。

针对光谱角算法，在建模集数量小于 50 时，R^2 和 RPD 均较低，所建立的模型并不能很好地预测有机碳；当建模集数量在 50～100 时，预测精度高于欧氏距离和马氏距离所匹配的模型；当建模集数量大于 50，小于 500 时，除去样本数为 200 的情况下，R^2>0.6、RPD>1.4，可以较好地预测土壤有机碳含量，由图 3 可知，除样本数为 200 以外其余建模集预测精度均优于欧氏距离和马氏距离，该方法下最优建模集数量范围同样为 400～500。

建模集数量较少（样本数<50）时，三种方法预测精度均不高，但随建模集数量的增多，预测精度皆有提高，三者中预测精度较高的是光谱角匹配，且最优建模集数量范围为 400～500。

本研究结果与预期有一些偏差，预期当建模数量样本较大时，越接近于全局建模，预测精度会降低，但三种匹配的最优建模集数量范围均为 400～500。原因可能如下：（1）由于光谱运算计算量较大，本文仅采用了 13 个等级，对有限的建模集数量（5～500）进行了比较，在今后的研究中，有必要对建模集数量做更细致的等级划分；（2）相比于全局建模，相似性样本匹配的建模策略存在差异，全局建模是仅利用一个模型对整个预测集做估算；而相似性匹配是针对每一个预测样本均进行建模预测，这可能是样本数量为 500 时预测精度依然较高的原因。

2.5　不同建模策略对土壤有机碳含量预测精度影响

由以上研究得出光谱角算法的预测精度略高于欧氏距离和马氏距离，因而在建模策略比较上，本文将针对光谱角算法匹配所挑选的样本，探究不同建模策略对预测精度的影响。如表 4 所示，建模预测精度随相似光谱数量的增加呈上升趋势，R^2 从 0.35 升至 0.75，RPD 由 1.21 升至 1.73，此上升趋势

表 4　PLSR 建模预测和赋值预测精度比较

Table 4　Comparison between the PLSR modeling strategy and the assignment strategy in prediction accuracy

相似光谱数量 Number of similar spectra	PLSR 建模预测 PLSR modeling strategy			赋值预测 Assignment strategy		
	R^2	RMSEp/（g·kg^{-1}）	RPD	R^2	RMSEp/（g·kg^{-1}）	RPD
5	0.35	0.65	1.21	0.52	0.56	1.43
10	0.32	0.67	1.19	0.54	0.54	1.46
20	0.47	0.61	1.30	0.55	0.54	1.48
30	0.46	0.62	1.28	0.52	0.56	1.42
40	0.56	0.59	1.35	0.53	0.55	1.45
50	0.59	0.57	1.38	0.52	0.56	1.42
100	0.68	0.53	1.50	0.43	0.60	1.32
150	0.71	0.50	1.58	0.41	0.61	1.29
200	0.63	0.59	1.34	0.39	0.63	1.27
250	0.69	0.52	1.54	0.36	0.64	1.24
300	0.70	0.53	1.49	0.35	0.65	1.23
400	0.72	0.48	1.64	0.33	0.66	1.20
500	0.75	0.46	1.73	0.30	0.68	1.17

在 0～50 条光谱时尤为显著，50 条之后精度趋于平稳；赋值预测精度随相似光谱数量的增加呈下降趋势，该下降趋势在 50～500 条光谱时极其明显，R^2 从 0.52 降至 0.30，RPD 由 1.43 降至 1.17。因而，当建模集数量较少时（<50），宜选择赋值预测的建模策略（R^2>0.6、RPD>1.4）；而当建模集数量较大时（>50），宜选择 PLSR 建模预测（R^2>0.6、RPD>1.4）。

出现该结果的可能原因如下：当建模样本数量增多时，匹配的相似光谱与参考光谱的曲线形态相似程度越低，相似样本的有机碳含量差异就越大，其有机碳含量的平均值与实际偏差越大，精度越低。所以只有少量样本参与建模（<50）时，赋值预测精度才会较高。对于 PLSR 建模预测，当建模样本数量较多时（>50），模型的预测精度较高。建模策略的合理选取需参考光谱库本身的变异，以及参考光谱库与待预测样本光谱的差异。

3 结 论

相较全局模型，基于欧氏距离、马氏距离和光谱角三种光谱相似性指数选取样本所建立模型均显著提升了土壤有机碳的预测精度。总体而言，光谱角算法略优于欧氏距离和马氏距离。不过本研究中预测精度并不高，这可能是由于全球土壤光谱库中的光谱数据较多且土壤样本分布广泛，与预测集土壤样本属性差异较大有关。对于建模集大小的研究，当建模集数量较少时，三种方法预测精度均不高，但随建模集数量的增多，预测精度皆有提高，三者中预测精度较高的是光谱角匹配，且最优建模集数量范围为 400～500。在建模策略上，当建模集数量较少时，宜选择赋值预测的建模策略；而当建模集数量较大时，宜选择 PLSR 建模预测。本研究还有待在如下方面进行进一步探索：（1）相似性样本数量需要进一步细化，并探究建模集样本数量为 500 以上的预测精度，以进一步探究最优建模集数量的选取；（2）增加其他光谱相似性指数的比较，比如相关系数法、兰氏距离以及多重测度方法，探究最佳的土壤有机碳预测策略；（3）尝试将该方法应用于其他土壤属性的预测，比如交换性阳离子、全氮、全磷、全钾，探讨该方法对其他属性预测的适宜性。

参考文献（References）

[1] Zhang G L, Zhu Y G, Shao M A. Understanding sustainability of soil and water resources in a critical zone perspective[J] Science China: Earth Sciences, 2019, 49(12): 1945—1947. [张甘霖, 朱永官, 邵明安. 地球关键带过程与水土资源可持续利用的机理[J]. 中国科学：地球科学, 2019, 49(12): 1945—1947.]

[2] Zhang G L, Zhu A X, Shi Z, et al. Progress and future prospect of soil geography[J]. Progress in Geography, 2018, 37(1): 57—65. [张甘霖, 朱阿兴, 史舟, 等. 土壤地理学的进展与展望[J]. 地理科学进展, 2018, 37(1): 57—65.]

[3] Tang H J, Li H P, Chen W Y, et al. Research progress on soil organic carbon based on map of scientific knowledge[J]. Acta Pedologica Sinica, 2019, 56(3): 541—552. [唐浩竣, 李海萍, 陈文悦, 等. 基于科学知识图谱谈土壤有机碳研究进展[J]. 土壤学报, 2019, 56(3): 541—552.]

[4] Viscarra Rossel R A, Walvoort D J J, McBratney A B, et al. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties[J]. Geoderma, 2006, 131(1/2): 59—75.

[5] Liu H J, Zhang B, Yang L, et al. Review of soil optical remote sensing[J]. Chinese Journal of Soil Science, 2007, 38(6): 1196—1202. [刘焕军, 张柏, 杨立, 等. 土壤光学遥感研究进展[J]. 土壤通报, 2007, 38(6): 1196—1202.]

[6] Wu D W, Zhang G L. Effects of parent materials and land use types on inversion models by using soil spectral data[J]. Soils, 2016, 48(1): 173—179. [邬登巍, 张甘霖. 母质与土地利用类型对土壤光谱反演模型的影响[J]. 土壤, 2016, 48(1): 173—179.]

[7] Brown D J, Shepherd K D, Walsh M G, et al. Global soil characterization with VNIR diffuse reflectance spectroscopy[J]. Geoderma, 2006, 132(3/4): 273—290.

[8] Lobsey C R, Viscarra Rossel R A, Roudier P, et al. Rs-local data-mines information from spectral libraries to improve local calibrations[J]. European Journal of Soil Science, 2017, 68(6): 840—852.

[9] Shi Z, Xu D Y, Teng H F, et al. Soil information acquisition based on remote sensing and proximal soil sensing: Current status and prospect [J]. Progress in Geography, 2018, 37(1): 79—92. [史舟, 徐冬云, 滕洪芬, 等. 土壤星地传感技术现状与发展趋势[J]. 地理科学进展, 2018, 37(1): 79—92.]

[10] Wei C L, Zhao Y G, Li D C, et al. Prediction of soil organic matter and cation exchange capacity based on spectral similarity measuring[J]. Transactions of the Chinese Society of Agricultural Engineering, 2014, 30(1): 81—88. [魏昌龙, 赵玉国, 李德成, 等. 基于相

似光谱匹配预测土壤有机质和阳离子交换量[J]. 农业工程学报，2014，30（1）：81—88.]

[11] 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.

[12] Ding J J，Zhang S，Sun C，et al. Optimal quantity relationship between modeling group and prediction group in soil organic carbon prediction model based on PLSR[J]. Journal of Jianghan University，2018，46（5）：404—408. [丁建军，章盛，孙超，等. 基于 PLSR 的土壤有机碳预测模型中建模组与验证组最优数量关系[J]. 江汉大学学报（自然科学版），2018，46（5）：404—408.]

[13] Peng Y，Knadel M，Gislum R，et al. Predicting soil organic carbon at field scale using a national soil spectral library[J]. Journal of Near Infrared Spectroscopy，2013，21（3）：213—222.

[14] Zeng R，Zhao Y G，Li D C，et al. Selection of "local" models for prediction of soil organic matter using a regional soil vis-NIR spectral library[J]. Soil Science，2016，181（1）：13—19.

[15] Wei C L. Study on methods of predicting soil properties based on spectral mapping technique [D].University of Chinese Academy of Sciences，2013. [魏昌龙. 基于光谱匹配技术的土壤属性预测方法研究[D]. 中国科学院大学，2013.]

[16] Zeng R，Zhao Y G，Wu D W，et al. Comparison of different strategies for predicting soil organic matter of a local site from a regional vis-NIR soil spectral library[J]. Digital Soil Mapping Across Paradigms，Scales and Boundaries，2016：319-329. https：//doi.org/10.1007/978-981-10-0415-5_26.

[17] Garrity D，Bindraban P. A globally distributed soil spectral library visible near infrared diffuse reflectance spectra [EB/OL]. 2003 [2019-4-29]. http：//www.worldagroforestrycentre.org

[18] Luo J W，Ying K，He P，et al. Properties of Savitzky–Golay digital differentiators[J]. Digital Signal Processing，2005，15（2）：122—136.

[19] Zhong Y F，Lin X M，Zhang L P. A support vector conditional random fields classifier with a mahalanobis distance boundary constraint for high s patial resolution remote sensing imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，2014，7（4）：1314—1330.

[20] Kruse F A，Lefkoff A B，Boardman J W，et al. The spectral image processing system（SIPS）—interactive visualization and analysis of imaging spectrometer data[J]. Remote Sensing of Environment，1993，44（2/3）：145—163.

[21] Wold S，Algano C，Dunn M，et al. Pattern regression：Finding and using regularities in multivariate data[M]. London：Analysis Applied Science Publication，1983.

[22] Chang C W，Laird D A，Mausbach M J，et al. Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties[J]. Soil Science Society of America Journal，2001，65（2）：480—490.

[23] Guo J X，Zhao X M，Guo X，et al. Inversion of organic matter content in red soil based on PLSR-BP composite model[J]. Acta Pedologica Sinica，2020，57（3）：636—645. [国佳欣，赵小敏，郭熙，等. 基于 PLSR-BP 复合模型的红壤有机质含量反演研究[J]. 土壤学报，2020，57（3）：636—645.]

[24] Ji W J，Li X，Li C X，et al. Using different data mining algorithms to predict soil organic matter based on visible-near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis，2012，32（9）：2393—2398. [纪文君，李曦，李成学，等. 基于全谱数据挖掘技术的土壤有机质高光谱预测建模研究[J]. 光谱学与光谱分析，2012，32（9）：2393—2398.]

[25] Liu Y，Ding X，Liu H J，et al. Quantitative analysis of reflectance spectrum of black soil as affected by soil moisture for prediction of soil moisture in black soil[J]. Acta Pedologica Sinica，2014，51（5）：1021—1026. [刘洋，丁潇，刘焕军，等. 黑土土壤水分反射光谱特征定量分析与预测[J]. 土壤学报，2014，51（5）：1021—1026.]

（**责任编辑**：檀满枝）