土壤有机质含量可见-近红外光谱反演模型校正集优选方法

陈奕云[1,2,3,4,5]，齐天赐[1,6]，黄颖菁[1]，万 远[7※]，
赵瑞瑛[1,8]，亓 林[1,9]，张 超[1]，费 腾[1,3]

（1.武汉大学资源与环境科学学院，武汉 430079；2.土壤与农业可持续发展国家重点实验室，南京 210008；3.武汉大学苏州研究院，苏州 215123；4.武汉大学地球空间信息技术协同创新中心，武汉 430079；5.武汉大学教育部地理信息系统重点实验室，武汉 430079； 6.湖泊与环境国家重点实验室（中国科学院南京地理与湖泊研究所），南京 210008；7.湖北师范大学，城市与环境学院，黄石 435002；8.浙江大学农业遥感与信息技术应用研究所，杭州 310058；9.中国科学院地理科学与资源研究所，北京 100101）

摘 要：土壤有机质含量可见-近红外光谱反演过程中校正集的构建策略对模型的预测精度有重要影响。以江汉平原洪湖地区水稻土为研究对象，采用 Kennard-Stone（KS）法，Rank-KS（RKS）和 Sample set Partitioning based on joint X-Y distance（SPXY）法，构建样本数占总校正集不同比例的子校正集，通过偏最小二乘回归，建立土壤有机质含量的可见—近红外光谱反演模型。结果表明：KS 法无法提高模型预测精度，但可以在保证标准差与预测均方根误差比（ratio of performance to standard deviation，RPD）>2.0 的前提下减少 30%的校正样本；基于 SPXY 法的模型，当子校正集样本比例为总校正集的 50%时达到最佳的模型预测精度，RPD 为 2.557；RKS 法能够在保证预测精度的情况下（RPD>2.0），最多减少总校正集 70%的样本，对应模型 RPD 为 2.212。当校正集与验证集的有机质含量分布相近时，能够以较少的建模样本达到与总校正集相近甚至更高的模型预测精度，提升土壤有机质光谱反演模型的实用性。

关键词：土壤；模型；土壤有机质；可见-近红外反射光谱；偏最小二乘回归；校正集优选
doi：10.11975/j.issn.1002-6819.2017.06.014
中图分类号：S151.9 **文献标志码**：A **文章编号**：1002-6819(2017)-06-0107-08

陈奕云，齐天赐，黄颖菁，万 远，赵瑞瑛，亓 林，张 超，费 腾. 土壤有机质含量可见-近红外光谱反演模型校正集优选方法[J]. 农业工程学报，2017，33(6)：107－114. doi：10.11975/j.issn.1002-6819.2017.06.014 http://www.tcsae.org

Chen Yiyun, Qi Tianci, Huang Yingjing, Wan Yuan, Zhao Ruiying, Qi Lin, Zhang Chao, Fei Teng. Optimization method of calibration dataset for VIS-NIR spectral inversion model of soil organic matter content[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(6): 107－114. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.2017.06.014 http://www.tcsae.org

0 引 言

土壤有机碳库是全球碳库中最为活跃的碳库，在全球碳循环研究和农业生产实践中具有重要地位[1]。获取土壤有机质（soil organic matter，SOM）空间分布信息，是研究土壤有机碳库时空动态变化的基础[2]。近年来，采用可见-近红外光谱技术获取土壤中有机质等土壤组分含量信息已成为土壤遥感与地理信息科学领域的重要研究方向[3-5]。该方法相比传统的化学分析方法具有操作简单、快速、无污染、成本低等优点，但仍存在模型预测精度相对较低、实用性不强等问题。

为进一步提高土壤有机质光谱估算模型的预测精度和实用性，国内外学者选取了不同研究区域、获取不同类型的土壤样本，从光谱波段与土壤有机质的相关关系入手，尝试了多种光谱预处理和回归建模方法，取得了一定成果[6-8]。然而有关校正集构建的研究开展较少。如何构建“合适”的校正集，即挑选出足以揭示土壤光谱对土壤有机质响应关系的样本组成校正集并用于回归模型构建[3]，以及如何在保证一定模型预测精度的前提下降低建模成本是当前研究的重要方向[9-10]。

Liu 等[11]研究发现，含有多种土地利用类型土壤样本组成的校正集所建立的回归模型可以较好地预测其中某类用地土壤的有机质含量，并且具有浓度代表性或光谱代表性的校正集所建模型有更好的预测精度。刘艳芳等[3]在对样本集进行地类分层的基础上结合浓度梯度法、Kennard-Stone（KS）法与 C-KS 法等方法构建校正集，发现考虑多层次土壤信息代表性的校正集构建方法能够有效提高土壤有机质光谱估算模型的适用性。因此，为了得到更加稳健的模型，不但需要从光谱预处理和回归建模方法入手，还需在构建校正集的时候，尽可能综合考虑土壤光谱、土壤组分含量信息乃至成土要素等可能影响土壤光谱与组分含量关系的各种因素及其变异，进而构建具有多元代表性的校正集[3]。此外，当前少有研究关注校正集样本量对模型预测精度的影响，“大样本光谱库对于局域土壤组分含量估算是否是必须的”这一问

收稿日期：2016-09-30 修订日期：2017-02-25
基金项目：国家自然科学基金项目（41501444）；苏州市应用基础农业项目（SYN201422, SYN201309）
作者简介：陈奕云，男，福建泉州人，副教授，博士，主要从事土壤遥感与地理信息科学研究。武汉 武汉大学资源与环境科学学院，430079。Email：chenyy@whu.edu.cn
※通信作者：万远，男，湖北鄂州人，博士，主要从事土地管理与地理信息的研究。黄石 湖北师范大学城市与环境学院，435002。
Email：wanyuan14@163.com

题近期亦引起一些学者的关注与讨论[12-14]。通常在模型预测精度达到实用预期的前提下，更少的建模样本意味着更低的建模成本。因此，在当前大样本光谱库建设与使用存在一定限制的背景下，开展校正集样本量与模型预测精度关系的研究对于提升土壤组分可见-近红外光谱模型实用性具有重要意义。

本文选取湖北省洪湖地区 100 份水稻土样本作为研究对象，研究目标及内容包括：1）光谱数据预处理后，构建基于不同校正样本挑选方法的 26 个子校正集，通过比较使用 26 个子校正集分别建立的估算土壤有机质含量的 PLSR 模型及用验证样本对各个模型进行检验的结果，以确定预测精度最高的子校正集样本比例与挑选策略。2）通过比较不同子校正集的土壤有机质数据分布特征及与其对应模型的预测结果，探究两者的对应关系，以期归纳得到预测效果较好的子校正集的共性，得出校正集优化的一般原则，进而为减少土壤有机质光谱估算模型建模成本，提高预测精度和实用性提供参考。

1 材料和方法

1.1 样本采集与化学分析

研究所使用的 100 份水稻土样本于 2014 年 7 月采集自洪湖市滨湖农田地区，样点空间分布及研究区土地利用类型如图 1 所示。采样时，在每个采样点约 10 m^2 的范围内采集表层土壤样本（0～15 cm）5 份均匀混合在一起，取不少于 500 g 的土壤样本装入自封袋后带回实验室做进一步分析。使用手持 GPS 记录采样点地理坐标。取回的土壤样本在实验室内进行自然风干，研磨并过 0.25 mm（60 目）土壤筛后，最终得到用于量测反射光谱及有机质含量的样本。

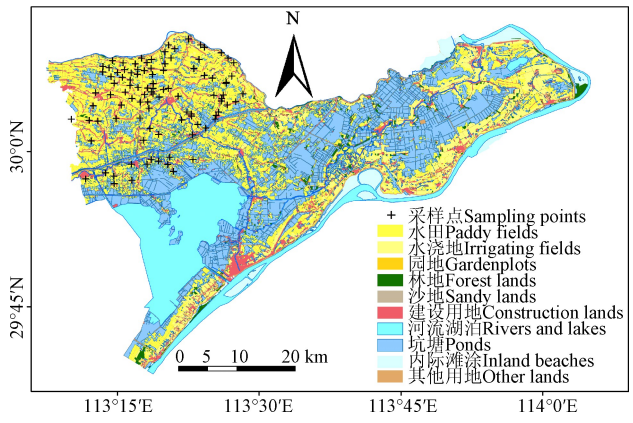


图 1 研究区土地利用类型及采样点分布图

Fig.1 Land use types of study area and spatial distribution of soil samples

参照《土壤有机碳的测定重铬酸钾氧化-分光光度法(HJ 615-2011)》测定土壤样本有机碳含量，结果乘以 van Bemmelen 因数 1.724[15]得到样本土壤有机质含量。

1.2 土壤样本的实验室光谱量测与预处理

采用美国 Analytical Spectral Devices 公司生产的 FieldSpec 3 地物光谱仪进行土壤样本反射率光谱量测。该仪器波长范围 350～2 500 nm，输出波段数 2 151，重采样间隔 1 nm。在干燥的暗室环境中，将处理好的土壤样本置于盛样皿中，以卤素灯为唯一入射光源。光源入射角 45°，距土壤样本表面 30 cm，光谱仪探头接 10°视场角镜头，位于土样垂直上方 15 cm 处。土壤样本光谱观测几何模拟了野外土壤光谱量测，同时避免野外观测过程中由于太阳辐射、大气水汽变化而产生的观测不确定性。每个样品量测 10 次反射率光谱后取算术平均，量测过程中每 10 份样本进行 1 次标准白板校正。

对试验量测得到的样本反射率光谱进行预处理，去除随机波动较大的边缘光谱波段，保留 400～2 450 nm 波段，采用 Savitzky-Golay 平滑、对数变换、多元散射校正与均值中心化处理[15]。

1.3 模型建立与验证

1.3.1 校正集与验证集的构建

采用箱型图剔除有机质含量异常的土壤样本，使用主成分分析法剔除光谱异常的土壤样本。以采集的 100 个土壤样本进行异常样本剔除后得到的 97 个样本作为试验样本。

为了使验证集具备一定的代表性和独立性，本文根据浓度梯度法挑选出 20 个样本组成验证集，剩下的样本作为总校正集，并使用 KS 法、Sample set Partitioning based on joint X-Y distance（SPXY）法与 Rank-KS（RKS）法[10]按照 10%、20%、…、90%的比例从中挑选样本组成多个子校正集用于后续的回归建模。其中因为样本数量有限，无法使用 RKS 法挑选 10%比例的样本组成子校正集，因此略去。

1.3.2 建立偏最小二乘回归模型

偏最小二乘回归是 1983 年由 Wold 等[17]首次提出的一种多元统计数据分析方法，该方法适用于处理存在多重共线性的数据，尤其在解决样本容量小、解释变量个数多、变量间存在多重相关性问题方面具有独特的优势。本文采用舍一交叉验证法（leave-one-out cross validation）确定最佳主因子数。

1.3.3 模型检验与评价指标

采用模型决定系数 R^2、均方根误差（root mean square error，RMSE）和标准差与预测均方根误差比（ratio of performance to standard deviation，RPD）这 3 个指标来检验与评价模型的定标效果和预测能力。以上评价指标中，外部检验决定系数 R_p^2越大，预测均方根误差（root mean square error of prediction, RMSEP）越小，说明模型预测效果越好。此外，一般认为当 RPD<1.4 时，模型较差无法对样本进行预测；1.4≤RPD<2 时，模型较好，可以用来进行 SOM 的粗略估算；2.0≤RPD<2.5 代表模型质量很好，可以用于 SOM 的定量预测；RPD≥2.5 代表模型具有极好的预测能力[18]。

选取变量投影重要性（variable importance in the projection，VIP）评估波段变量对模型的重要性，VIP 值大于 1 的波段变量较 VIP 值小于 1 的波段变量对 SOM 变异的解释具有更加重要的作用[19]。

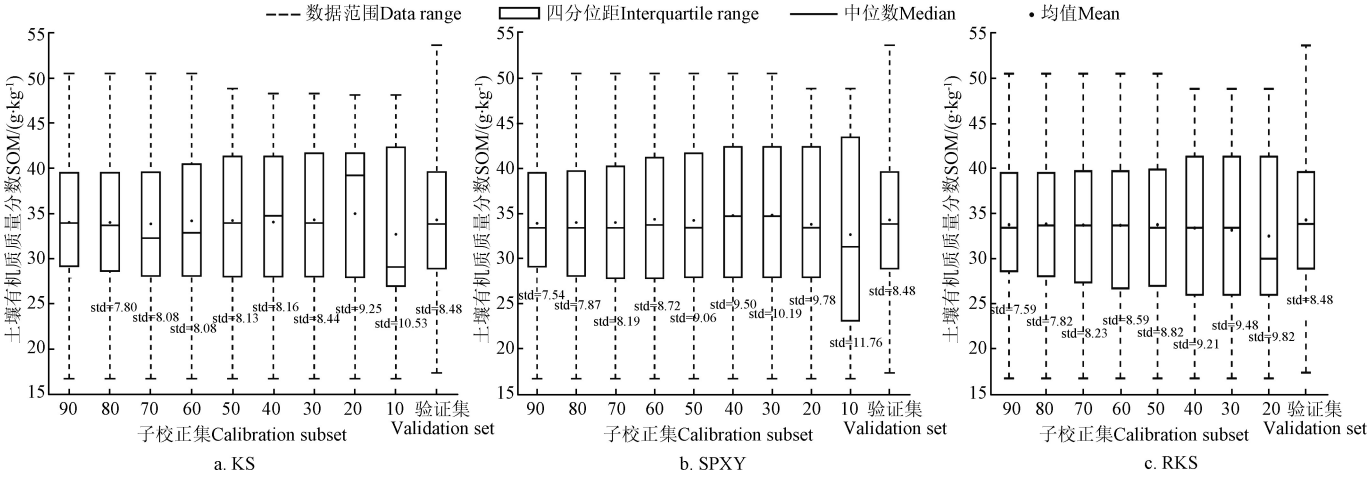
本文中土壤样本的反射光谱与有机质含量数据的预

处理、偏最小二乘回归及模型评价指标计算使用 MathWorks 公司的分析建模软件 MATLAB 及基于 MATLAB 的 PLS toolbox（Eigenvector Research 公司，8.0 版本）实现；KS 法、RS 法与 SPXY 法的实现使用基于 MATLAB 的 SPA_GUI 完成[20-22]。

2　结果与分析

2.1　样本集的统计特征

全部样本、总校正集、验证集以及不同方法构建的子校正集 SOM 含量的统计特征如图 2、图 3 与表 1 所示。

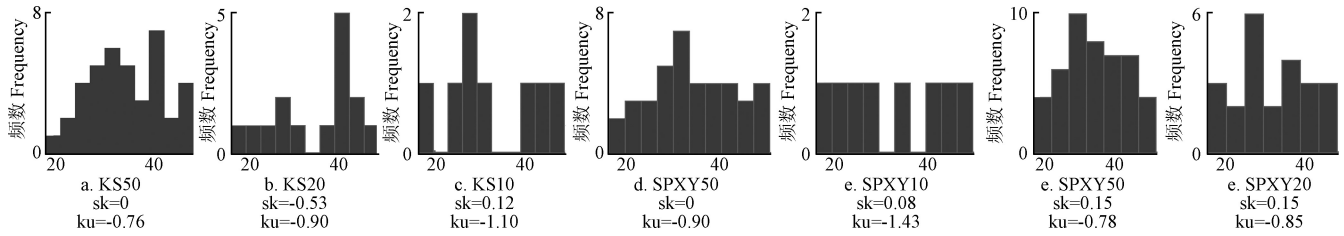


注：KS 为 Kennard-Stone（KS）法，SPXY 为 sample set partitioning based on joint X-Y distance（SPXY）法，RKS 为 Rank-KS 法；10、20、……90 表示以 10%、20%……90%的比例从总校正集中挑选样本组成的子校正集，其样本数量分别为：8、15、23、31、39、46、54、62、69。下同。

Note: KS is Kennard-Stone (KS), SPXY is sample set partitioning based on joint X-Y distance (SPXY), RKS is Rank-KS; 10, 20…90 indicate that 10%, 20%…90% of samples were selected from total calibration set, forming different calibration subsets, and the samples number of calibration subsets was 8, 15, 23, 31, 39, 46, 54, 69, respectively. The same as below.

图 2　子校正集土壤有机质含量的箱形图

Fig.2　Box plots of SOM content from calibration subsets



注：sk 为偏度系数，ku 为峰度系数。　　Note: sk is skewness coefficient, ku is kurtosis coefficient

图 3 部分子校正集土壤有机质含量的频数直方图

Fig.3　Histograms of SOM content from part of calibration subsets

表 1　土壤有机质含量的描述性统计

Table 1　Descriptive statistics of SOM content

样本集 Sample set	样本数 Number	土壤有机质质量分数 SOM content(g·kg^{-1})				偏度系数 Skewness coefficient	峰度系数 Kurtosis coefficient
		最大值 Max value	最小值 Min value	均值 Mean value	标准差. Standard Deviation		
总校正集 Total calibration set	77	50.50	16.70	33.92	7.30	0.09	-0.32
验证集 Validation set	20	53.63	17.34	34.28	8.48	0.29	0.04
全部样本 Whole sample set	97	53.63	16.70	34.00	7.51	0.15	-0.17

全部 97 个样本的土壤有机质含量在 16.70～53.63 g/kg 之间，均值为 34.00 g/kg，与河南伊川县城关镇的水稻土有机质含量（35.7 g/kg）相差不大，略低于四川省邛崃市回龙镇柏杨村的水稻土的有机质含量（41.8 g/kg）[23]。由图 2、图 3 可知，KS 法、RS 法与 SPXY 法按照不同比例从总校正集中挑选出的子校正集样本有机质含量均值、中位数、标准差、偏度系数、峰度系数以及四分位距等统计指标存在差异，表明不同子校正集构建策略会影响校正样本有机质含量分布特征。对于 KS 算法，当子校正集样本比例为总校正集的 20%和 10%的时候，中位数与均值存在较大的偏离，前者的直方图还呈现明显的负偏（偏度系数 sk=−0.53）；对于 SPXY 算法，当子校正集样本比例为 10%时，直方图呈现平峰的特征，四分位距也明显大于其他子校正集。在回归分析

中，对于分布中间多两端少的样本集，常常导致模型预测结果偏离真实值而趋向于“均值化”[10]，同样若是样本集分布呈较明显的正偏或负偏也可能导致模型预测向着一个固定方向偏移。

2.2 基于反射率光谱的土壤有机质含量估算模型

分别对基于不同校正集构建策略得到的子校正集进行 PLSR 建模，各个子校正集最佳模型对应的评价指标如表 2 所示。

表 2 各子校正集 PLSR 模型结果

Table 2 Performance of PLSR models calibrated from different calibration subsets

样本集 Sample set	建模均方根误差 RMSEC/(g·kg^{-1})	建模决定系数 R_c^2	预测均方根误差 RMSEP/(g·kg^{-1})	预测决定系数 R_p^2	标准差与预测均方根误差比 RPD	主因子数 Principal factor number
总校正集 Total calibration set	3.586	0.755	3.926	0.813	2.184	6
KS90	3.563	0.769	3.888	0.825	2.170	6
KS80	3.513	0.794	4.000	0.807	2.120	6
KS70	3.604	0.797	4.188	0.797	2.097	6
KS60	3.794	0.775	4.291	0.757	1.994	5
KS50	3.965	0.756	4.427	0.731	1.886	5
KS40	4.272	0.716	4.649	0.714	1.791	5
KS30	4.086	0.755	5.337	0.588	1.555	4
KS20	6.089	0.536	6.617	0.417	1.259	2
KS10	5.923	0.639	7.069	0.433	1.285	2
SPXY90	3.322	0.803	4.064	0.787	2.151	8
SPXY80	3.336	0.818	4.209	0.772	2.086	8
SPXY70	3.179	0.847	4.264	0.779	2.111	8
SPXY60	2.718	0.901	4.008	0.799	2.229	8
SPXY50	2.499	0.922	3.555	0.848	2.557	8
SPXY40	2.116	0.949	4.312	0.765	2.042	8
SPXY30	4.603	0.787	5.219	0.609	1.598	4
SPXY20	3.716	0.845	6.168	0.521	1.435	4
SPXY10	3.695	0.887	8.180	0.333	1.088	3
RKS90	4.171	0.694	4.202	0.777	2.033	5
RKS80	4.187	0.709	4.174	0.780	2.038	5
RKS70	3.158	0.850	4.140	0.780	2.091	7
RKS60	2.972	0.878	4.283	0.776	2.088	7
RKS50	2.766	0.899	4.273	0.779	2.120	7
RKS40	3.539	0.847	4.117	0.775	2.107	5
RKS30	3.313	0.872	4.219	0.802	2.212	5
RKS20	2.025	0.954	6.145	0.554	1.482	5

对于 KS 法构建的子校正集，随着样本数目的减少，建模均方根误差（root mean square error of calibration，RMSEC）不断增加，R_c^2 稳定在 0.770 左右，只有当样本数少于总校正集的 20%后才呈现出明显的下降，表明 KS 法挑选出的子校正集所建模型对数据自身拟合效果良好且稳定。使用基于 KS 法构建的子校正集建立的偏最小二乘回归模型对验证集样本 SOM 含量进行估算时发现：当子校正集样本数占总校正集比例高于 60%的时候，各子校正集所建立的模型预测能力相当，模型质量很好（RPD>2.0）。当比例低于 60%的时候，随着校正集样本数的减少，RPD 开始逐渐降低，表明 KS 法在校正样本数量下降到一定比例后容易丢失对模型预测精度有明显贡献的样本；当子校正集样本数占总校正集的 20%及以下时，RPD<1.4，已经无法用于 SOM 含量的估算[18]。

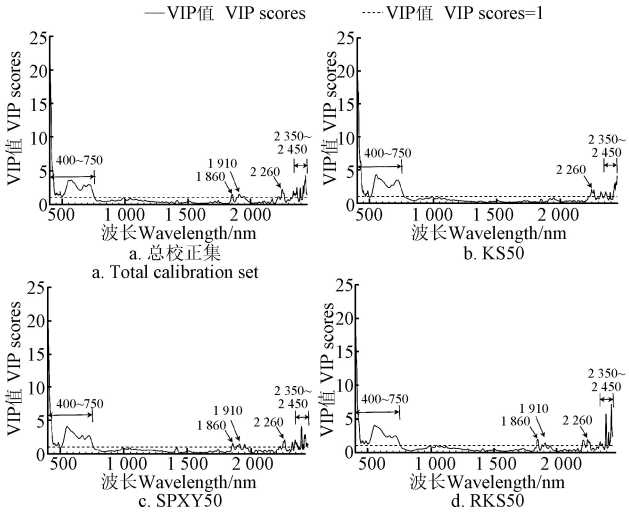
对于 SPXY 法构建的子校正集，其 R_c^2 普遍高于 KS 法构建的子校正集，除了子校正集 SPXY30 之外均在 0.8 以上，当子校正集样本数为总校正集 40%时（子校正集 SPXY40），R_c^2 达到了最高的 0.949，这表明综合覆盖光谱空间与理化性质空间的校正集样本挑选方法较仅均匀覆盖光谱空间的校正集样本挑选方法具有潜在的优越性。在模型验证方面，当子校正集样本数占总校正集 50%的时候（子校正集 SPXY50），所建立的模型 RPD 达到最大值 2.557，高于总校正集的 2.184；当样本数少于总校正集 50%的时候，RPD 开始下降。

对于 RKS 法构建的子校正集，当样本数减至总校正集 70%及以下时，其 R_c^2 稳定在 0.84 以上，表明模型对建模数据的拟合效果良好。在模型验证方面，使用 RKS 挑选的子校正集所建立的模型 RPD 值稳定在 2.1 左右。当子校正集样本数为总校正集 30%的时候（子校正集 RKS30），建立的模型表现出了最佳的预测效果，其 RPD 为 2.212，R_p^2 为 0.802，优于 KS30 与 SPXY30 对应的 RPD 与 R_p^2 值。

由表 2 可知，KS 法挑选子校正集的最优挑选比例是 70%（即子校正集 KS70），对应的 RPD 为 2.097，尽管低于 KS90 和 KS80 的 2.170 和 2.120，但可以在保证模型预测精度达到“很好”标准（RPD>2.0）的前提下，仅使用总校正集 70%的样本就可以达到与总校正集所建立模型相近的预测精度；SPXY 法的最佳挑选比例是 50%（SPXY50），在提高了模型预测精度的同时（RPD 由 2.184 提升至 2.557），相比总校正集减少了 50%的建模样本；RKS 法挑选的子校正集所建模型虽然在预测精度上没有显著的提升，但是能够在校正集样本数较少的情况下保证模型预测精度与全样本模型相当，就试验结果来看最多可以减少最多 70%的建模样本，即仅需使用总校正集 30%的样本（RKS30）。

2.3 预测模型重要波段分析

对总校正集、KS50 校正集、SPXY50 校正集与 RKS50 校正集进行 PLSR 建模，并做出 VIP 曲线图，如图 4 所示。



注：VIP 为变量投影重要性。

Note: VIP is variable importance in the projection.

图 4 不同校正集 PLSR 模型 VIP 曲线图

Fig.4 VIP scores of PLSR models from different calibration subsets

综合比较 4 个校正集对应的 VIP 曲线可知，共同的重要波段主要为 400～750、2 260、2 350～2 450 nm；KS50 对应曲线相较于其他 3 种曲线缺少了 1 860 和 1 910 nm 附近的重要波段；RKS50 对应曲线中 2 350～2 450 nm 波段的相对重要程度大于其他三者。其中，400～750 nm 波段主要与有机质和铁氧化物相关[23-24]，波段 1 450、1 860、1 910 nm 与土壤及矿物中水分有关[26]，波段 2 260 与 2 440 nm 主要与羟基伸缩振动以及 Al-OH 和 Mg-OH 弯曲振动的合谱带有关[26-29]。

结合 VIP 曲线可知，不同挑选方法构建的子校正集 PLSR 模型都能较好地利用土壤有机质、铁氧化物以及土壤矿物的光谱响应波段来描述或辅助描述土壤有机质的变异。然而与总校正集构建的模型相比，KS50 所建模型表现为对土壤及矿物水分信息的压抑；SPXY50 所建模型表现为对土壤水分信息的增强和土壤矿物信息的压抑；RKS50 所建模型表现为对土壤矿物及水分信息的增强。

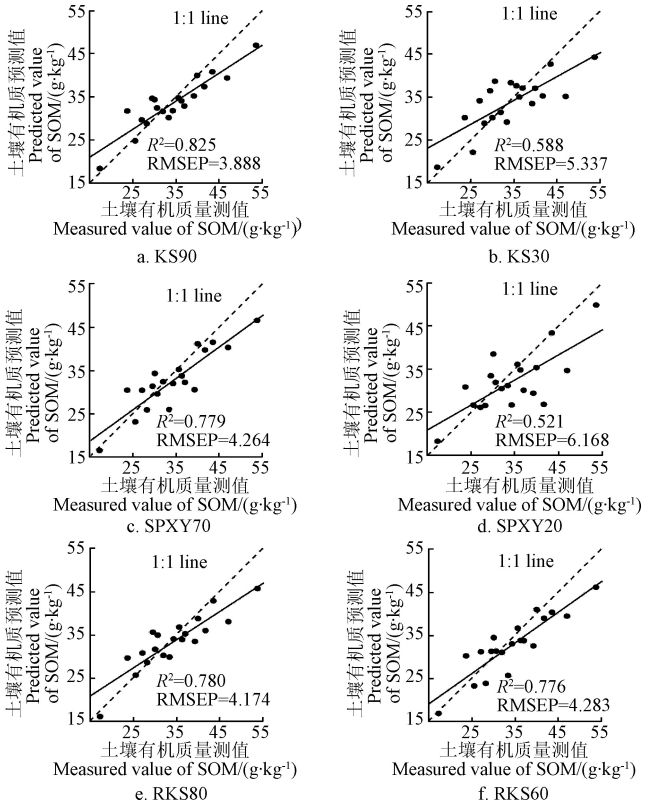
3　讨　论

结合预测值与实测值的部分散点图（图 5），有机质含量分布特征（图 2、图 3）以及预测统计指标（表 2）对研究结果进行进一步探讨：

KS 法构建的子校正集所建模型部分预测值和实测值散点图见图 5a～图 5b，随着样本数量的减少，预测值的分布趋于中间密集，导致模型预测结果的“均值化”现象越来越明显。KS20 与 KS10 均值与中位数偏离较远（图 2），对比表 2 中的预测结果，其 RPD 均小于 1.4，说明当子校正集样本量小于 30%时，使用 KS 法挑选出来的样本很可能不够具有代表性，导致模型预测精度明显降低。KS 法与 SPXY 和 RKS 法相比，优点在于无需使用总校正集的 SOM 含量作为先验信息，校正集样本量的减少即意味着建模成本的降低。本研究中，KS 法可以减少总校正集最多达 70%的校正样本（KS30 的 RPD>1.4，见表 2）。

SPXY 法构建的子校正集所建的部分模型预测值和实测值散点图见图 5c～图 5d，当子校正集样本数目大于总样本数的 70%时，SPXY 子校正集模型预测表现与 KS 子校正集相近（RPD 在 2.0～2.2 之间）；当子校正集样本数目小于 30%时，预测效果明显下降，散点图中各点呈散乱分布。且在图 5d 中，有机质含量较低与较高的样本预测误差较小，而中值区间的样本误差较大，很可能与对应子校正集样本呈现出的两极样本数与中间值样本数相近有关（当子校正集样本数目小于 30%时，其峰度系数均在 2.0 以下；而当子校正集样本数目大于 30%时，其峰度系数均在 2.0 以上）。根据 SPXY 法距离度量的原理，越是靠近理化性质空间两端的样本，其理化性质空间距离就越大，也就更容易被选入校正集中。然而从模型构建角度来说，中间的样本也具有校正意义，所以当使用 SPXY 法选取样本比例较小的时候，构建的子校正集样本有机质含量趋于平峰分布，进而导致模型预测精度的下降，对于土壤有机质含量这种普遍呈正态分布的土壤属性影响更为明显。

RKS 法构建的子校正集所建的部分模型预测值和实测值散点图见图 5e～图 5f，结合表 2 可以发现随着子校正集样本数目的减少，模型预测结果拟合优度保持稳定且较好（RPD>2.0）。就散点的分布来看，在子校正集样本数占总校正集样本数 60%以上（图 5e）时，RKS 法与其他 2 种方法模型预测效果并无明显差别，RPD 均在 2.0 以上（表 2）。但是当样本数继续减少，RKS 各子校正集有机质含量均值及中位数与验证集相近，偏度系数接近于 0 或呈正偏（即偏度系数大于 0，RKS20 除外），而 KS 校正集样本有机质含量趋于负偏（即偏度系数小于 0，KS10 除外）、SPXY 校正集样本有机质含量趋于平峰分布（图 2、图 3），RKS 校正集模型的预测结果优于另外 2 种，这说明近似于验证集的校正集能够带来更加稳定的模型。



注：RMSEP 为预测均方根误差。
Note: RMSEP is root mean square error of prediction.

图 5　不同子校正集模型土壤有机质量测值与预测值散点图
Fig.5　Scatter plots of measured and predicted value of SOM from models with different calibration subsets

综合比较 3 种方法挑选的校正集所建模型预测结果，要提升线性模型预测精度，需要保证校正集样本尽可能广地覆盖光谱空间[30-31]，同时较近似于验证集地分布在 SOM 含量浓度区间[11]，如 SPXY50 及 RKS50。

结合不同挑选方法 50%子校正集所建模型的重要波段（图 4）及其预测结果（表 2）可以看出，有机质含量分布较为近似于验证集的子校正集（图 2、图 3），可以保证建模过程中土壤有机质与羟基的特征波段能够在因变量解释中占有较大比例。对比 SPXY 子校正集与 RKS 子校正集可知，虽然 2400 nm 波段附近的土壤矿物信息

对因变量解释也有一定作用，但当其比例过大时可能对模型预测产生负面影响。

本文仅以 77 个土壤样本作为总校正集，比较各百分比子校正集的不同，总样本数略少，且每 10%相差的样本只有 7 个，因此可能存在某些特征样本对模型的影响，特别是在小样本量样本集中表现尤为明显，从而可能对试验结果造成一定干扰。

4 结 论

KS 法、SPXY 法及 RKS 法均能在保证模型预测精度的前提下（例如 RPD>2.0）降低建模成本即使用更少的校正样本，同时由于加权效应的影响，当校正集样本有机质含量分布与预测样本集数据分布特征相近时（即相近的均值、中位数、偏度系数和峰度系数），所建立的模型能够获得更好的预测结果。

KS 法挑选策略仅考虑样本光谱特征，当样本量过少的时候，校正集样本有机质含量分布多呈负偏，与验证集差异较大，故需要较多的校正样本，综合考虑样本量和预测精度的最佳挑选比例为 70%，模型对应 RPD 为 2.097，R_p^2 为 0.779，与全样本模型预测精度相当。SPXY 法挑选策略导致校正集样本理化性质分布呈双峰，因此也需保证一定的样本数目，其最佳挑选比例为 50%，模型对应 RPD 为 2.557，R_c^2 为 0.922，R_p^2 为 0.848，优于全样本模型。RKS 法保证了校正集样本有机质含量的均匀分布，更加适合线性模型，在样本数目足够的情况下各模型有着稳定的表现，本研究中的最佳挑选比例为 30%，对应模型的 RPD 为 2.212，R_c^2 为 0.872，R_p^2 为 0.802，与全样本模型预测精度相当，但是极大地降低了建模所需的样本数量。

土壤反射率光谱是土壤内在组分和外在成土要素的综合体现，未来研究可以在土壤光谱与组分含量信息的基础上增加土地利用、景观环境等其他可能影响土壤光谱与组分关系的辅助信息，提升校正集样本的代表性，以提高土壤组分光谱反演模型的实用性。

[参 考 文 献]

[1] 潘根兴，赵其国. 我国农田土壤碳库演变研究：全球变化和国家粮食安全[J]. 地球科学进展，2005，20(4)：384－393.
Pan Genxing, Zhao Qiguo. Study on evolution of organic carbon stock in agricultural soils of China: Facing the challenge of global chance and food security[J]. Advances in Earth Science, 2005, 20(4): 384－393. (in Chinese with English abstract)

[2] Liu Yaolin, Guo Long, Jiang Qinghu, et al. Comparing geospatial techniques to predict SOC stocks[J]. Soil & Tillage Research, 2015, 148: 46－58.

[3] 刘艳芳，卢延年，郭龙，等. 基于地类分层的土壤有机质光谱反演校正样本集的构建[J]. 土壤学报，2016，53(2)：332－341.
Liu Yanfang, Lu Yannian, Guo Long, et al. Construction of calibration set based on the land use types in visible and near-infrared(VIS-NIR) model for soil organic matter estimation[J]. Acta Pedologica Sinica, 2016, 53(2): 332－341. (in Chinese with English abstract)

[4] Shi Tiezhu, Cui Lijuan, Wang Junjie, et al. Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy[J]. Plant and Soil, 2013, 366(1/2): 363－375.

[5] Soriano-Disla J M, Janik L J, Viscarra Rossel R A, et al. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical chemical, and biological properties[J]. Applied Spectroscopy Reviews, 2014, 49(2): 139－186.

[6] Viscarra Rossel R A, Behrens T. Using data mining to model and interpret soil diffuse reflectance spectra[J]. Geoderma, 2010, 158(1/2): 46－54.

[7] Stevens A, Nocita M, Tóth G, et al. Prediction of soil organic carbon at the european scale by visible and near infraRed reflectance spectroscopy [J]. Plos One, 2013, 8(6): e66409.

[8] 陈奕云，漆锟，刘耀林，等. 顾及土壤湿度的土壤有机质高光谱预测模型传递研究[J]. 光谱学与光谱分析，2015，35(6)：1705－1708.
Chen Yiyun, Qi Kun, Liu Yaolin, et al. Transferability of hyperspectral model for estimating soil organic matter concerned with soil moisture[J]. Spectroscopy and Spectral Analysis, 2015, 35(6): 1705－1708. (in Chinese with English abstract)

[9] Kuang B, Mouazen A M, Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale[J]. European Journal of Soil Science, 2012, 63(3): 421－429.

[10] 刘伟，赵众，袁洪福，等. 光谱多元分析校正集和验证集样本分布优选方法研究[J]. 光谱学与光谱分析，2014，34(4)：947－951.
Liu Wei, Zhao Zong, Yuan Hongfu, et al. An optimal selection method of sample of calibration set and validation set for spectral multivariate analysis[J]. Spectroscopy and Spectral Analysis, 2014, 34(4): 947－951. (in Chinese with English abstract)

[11] Liu Yaolin, Jiang Qinghu, Fei Teng, et al. Transferability of a visible and near-Infrared model for soil organic matter estimation in riparian landscapes[J]. Remote Sensing, 2014, 6(5): 4305－4322.

[12] Guerrero C, Wetterlind J, Stenberg B, et al. Do we really need large spectral libraries for local scale SOC assessment

with NIR spectroscopy?[J]. Soil & Tillage Research, 2016, 155: 501－509.

[13] Guerrero C, Stenberg B, Wetterlind J, et al. Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset[J]. European Journal of Soil Science, 2014, 65(2): 248－263.

[14] 刘会增，石铁柱，王俊杰，等.利用区域土壤光谱库研究土壤有机碳反演模型传递性[J]. 武汉大学学报：信息科学版，2016，41(7)：1－7.
Liu Huizeng, Shi Tiezhu, Wang Junjie, et al. Transferability of retrieval models for estimating soil organic carbon contents based on regional soil spectral libraries[J]. Geomatics and Information Science of Wuhan University, 2016, 41(7): 889－895. (in Chinese with English abstract)

[15] 曾志远. 土壤肥力的卫星遥感探测[J]. 土壤，1987(2)：73－78，72.

[16] Rinnan A, van den Berg F, Engelsen S B. Review of the most common pre-processing techniques for near-infrared spectra[J]. Trac-Trends in Analytical Chemistry, 2009, 28(10): 1201－1222.

[17] Wold S, Albano C, Dunn WJ III, et al. Pattern recognition: finding and using regularities in multivariate data[J]. Food Research and Data Analysis, 1983: 147－188.

[18] 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.

[19] Chong I G, Jun C H. Performance of some variable selection methods when multicollinearity is present[J]. Chemometrics and Intelligent Laboratory Systems, 2005, 78(1/2): 103－112.

[20] Araújo M C U, Saldanha T C B, Galvão R K H, et al. The successive projections algorithm for variable selection in spectroscopic multicomponent[J]. Chemometrics and Intelligent Laboratory Systems, 2001, 57(2): 65－73.

[21] GalvãoR K H, Araújo M C U, Fragoso W D, et al. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm[J]. Chemometrics and Intelligent Laboratory Systems, 2008, 92(1): 83－91.

[22] Galvão R K H, Araújo M C U, Jose G E, et a1. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67: 736－740.

[23] 孙丽蓉，王旭刚，高翔. 有机质和铁氧化物对水稻土吸附 Cd^{2+}的贡献[J]. 河南农业科学，2010(4)：57－61.
Sun Lirong, Wang Xugang, Gao Xiang. Contribution of organic matter and iron oxides to adsorption of Cd^{2+} on paddy soils[J]. Journal of Henan Agricultural Sciences, 2010(4): 57－61. (in Chinese with English abstract)

[24] Viscarra Rossel R A, Bui E N, de Caritat P, et al. Mapping iron oxides and the color of Australian soil using visible–near-infrared reflectance spectra[J]. Journal of Geophysical Research Atmospheres, 2010.

[25] Liu Yaolin, Chen Yiyun. Estimation of total iron content in floodplain soils using VNIR spectroscopy–a case study in the Le'an River floodplain, China[J]. International Journal of Remote Sensing, 2012, 33(18): 5954－5972.

[26] 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.

[27] Krishnan P, Alexander J D, Butler B J, et al. Reflectance technique for predicting soil organic matter[J]. Soil Science Society of America Journal, 1980, 44(6): 1282－1285.

[28] Bartholomeus H M. Schaepman M E, Kooistra L, et al. Spectral reflectance based indices for soil organic carbon quantification[J]. Geoderma, 2008, 145(2): 28－36.

[29] 纪文君，史舟，周清，等. 几种不同类型土壤的 VIS-NIR 光谱特性及有机质响应波段[J].红外与毫米波学报，2012，31(3)：277－282.
Ji Wenjun, Shi Zhou, Zhou Qing, et al. VIS-NIR reflectance spectroscopy of the organic matter in several types of soils[J]. Journal of Infrared and Millimeter Waves, 2012, 31(3): 277－282. (in Chinese with English abstract)

[30] 卢延年，刘艳芳，陈奕云，等. 江汉平原土壤有机碳含量高光谱预测模型优选[J]. 中国农学通报，2014(26)：127－133.
Lu Yannian, Liu Yanfang, Chen Yiyun, et al. Optimization of the hyperspectral prediction model of soil organic carbon contents of Jianghan Plain[J]. Chinese Agricultural Science Bulletin, 2014(26): 127－133. (in Chinese with English abstract)

[31] 于雷，洪永胜，耿雷，等. 基于偏最小二乘回归的土壤有机质含量高光谱估算[J]. 农业工程学报，2015，31(14)：103－109.
Yu Lei, Hong Yongsheng, Geng Lei, et al. Hyperspectral estimation of soil organic matter content based on partial least squares regression[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(14): 103－109. (in Chinese with English abstract)

Optimization method of calibration dataset for VIS-NIR spectral inversion model of soil organic matter content

Chen Yiyun[1,2,3,4,5], Qi Tianci[1,6], Huang Yingjing[1], Wan Yuan[7※], Zhao Ruiying[1,8], Qi Lin[1,9], Zhang Chao[1], Fei Teng[1,3]

(1. *School of Resource and Environment Science*, *Wuhan University*, *Wuhan* 430079, *China*; 2. *State Key Laboratory of Soil and Sustainable Agriculture*, *Nanjing* 210008, *China*; 3. *Suzhou Institute of Wuhan University*, *Suzhou*, *Jiangsu* 215123, *China*; 4. *Collaborative Innovation Center of Geospatial Technology*, *Wuhan University*, *Wuhan* 430079, *China*; 5. *Key Laboratory of Geographic Information System of Ministry of Education*, *Wuhan University*, *Wuhan* 430079, *China*; 6. *State Key Laboratory of Lake Science and Environment*, *Nanjing Institute of Geography and Limnology*, *Chinese Academy of Sciences*, *Nanjing* 210008, *China*; 7. *College of Urban and Environmental Sciences*, *Hubei Normal University*, *Huangshi* 435002, *China*; 8. *Institute of Agricultural Remote Sensing and Information Technology Application*, *Zhejiang University*, *Hangzhou* 310058, *China*; 9. *Institute of Geographic Sciences and Natural Resources Research*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*)

Abstract: Soil organic matter (SOM) is not only an important indicator of soil fertility but also an important source and sink of the global carbon cycle. Therefore, it is essential to acquire the information of SOM for soil management. Visible and near-infrared (VIS-NIR) reflectance spectroscopy, known as a novel, rapid, accurate, environment-friendly and efficient approach compared with conventional laboratory analyses, has proven to be promising in the acquisition of various soil properties. Construction of a calibration set is key to the VIS-NIR quantitative analysis in building up a prediction model of high quality. The aim of this paper was to explore how the sample selection method and the number of samples may affect the accuracy of VIS-NIR models for SOM estimation. A total of 100 paddy soil samples (0-15 cm) were collected from the Honghu City, which is located in the Jianghan Plain, China. After air drying, grinding and sieving (0.25 mm), reflectance of these pretreated samples was measured with FieldSpec3 (Analytical Spectral Devices Inc., America). Three samples were neglected after outlier detections of spectra and SOM content. Out of the remaining 97 samples, 20 samples were selected by means of concentration gradient, which then formed the validation sample set. The remaining 77 samples formed the total calibration set. With SOM content or soil spectral information as inputs, 3 sample selection methods, namely Kennard-Stone (KS), sample set partitioning based on joint X-Y distance (SPXY) and Rank-KS, were used in the construction of calibration subsets with different proportions of the samples in total calibration set, such as 10% and 20%. Based on the different calibration subsets, partial least squares regression (PLSR) was used for model calibrations. Results showed that the calibration set selected by KS approach could not improve model predictive capability compared with the total calibration set. The KS approach, however, could reduce as many as 30% samples of the total calibration set while the ratio of performance to standard deviation (RPD) was retained above 2.0. The SPXY approach performed the best when 50% samples of the total calibration set were selected in the model calibration. The determination coefficient for calibration (R_c^2) reached 0.922, the determination coefficient for prediction (R_p^2) was 0.848, and the RPD reached 2.557. This was because the SPXY approach took into account both SOM content and soil spectra in the sample selection process. With only 30% samples of the total calibration set selected by the Rank-KS method, it had the lowest cost of calibration with satisfactory performance (R_c^2=0.872, R_p^2=0.802 and RPD=2.212). Overall, such results indicate that it is possible to reduce the number of calibration samples while retaining or even improving the predictive capacity of VIS-NIR models for SOM estimation. All the 3 calibration selection approaches have been proven to be useful for the improvement of model practicability.

Keywords: soils; models; organic matter; visible and near-infrared reflectance spectrum; partial least squares regression; optimization of calibration set