第42卷，第9期
2022年9月

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

Vol. 42, No. 9, pp2924-2930
September, 2022

砂壤潮土有机质含量可见-近红外光谱预测

钟翔君[1,2]，杨 丽[1,2*]，张东兴[1,2]，崔 涛[1,2]，和贤桃[1,2]，杜兆辉[1,2]

1. 中国农业大学工学院，北京 100083
2. 农业农村部土壤-机器-植物系统技术重点实验室，北京 100083

摘 要 土壤有机质(SOM)是影响播量的土壤关键参数，根据SOM信息对播量进行实时调控，投入最优化的种子量，充分利用地力资源挖掘产量潜力，节约良种，实现种植收益最大化，是目前播种领域最前沿的研究方向。以玉米主产区之一的华北平原为研究区域，对该区域砂壤潮土进行了可见-近红外(300～2 500 nm)光谱采集。采用蒙特卡罗交叉验证剔除了异常样本，结合Savitzky-Golay卷积平滑法对光谱数据进行平滑去噪处理。分别通过竞争性自适应重加权算法(CARS)、连续投影算法(SPA)、竞争性自适应重加权-连续投影(CARS-SPA)、无信息变量消除(UVE)及变量组合集群分析法(VCPA)等波长筛选方法提取有效变量，并结合偏最小二乘回归(PLSR)分别建立了全波长和特征波长的SOM含量预测模型。结果表明，不同方法筛选的波长数目及波长位置存在显著差异，CARS和SPA算法选择的光谱特征在整个光谱范围都有分布，UVE和VCPA筛选的波段较为集中，且基于CARS-SPA方法可以进一步优选特征变量，其特征波长仅为全波长数量的15%。通过对比不同模型的建模及预测效果，除UVE和VCPA算法外，其余算法构建的模型均能实现SOM含量的有效预测，其RPD值均大于2.0。基于CARS-SPA构建的PLSR模型效果最好，其R_P^2和RPD分别0.901和3.188，均高于其他方法，不仅降低了无效信息对预测效果的干扰，且模型的运算效率得到了明显的提高，可以很好地实现该地区SOM含量的可靠预测。该研究可以为SOM含量快速预测及仪器设计提供方法参考。

关键词 土壤有机质；播种；可见-近红外；砂壤潮土；竞争性自适应重加权-连续投影算法

中图分类号：S153.6 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)09-2924-07

引 言

土壤有机质(soil organic matter，SOM)是影响播量的土壤关键参数，对作物的生长发育起至关重要的作用[1-2]。根据田间SOM信息对播量进行实时调控，可以充分挖掘土壤潜力、节约良种用量，对作物的提质增效具有重要意义[3-6]。传统SOM信息的获取多以实验室化学分析为主[7]，虽然应用较广，但分析过程繁琐、时效性差、成本高且采样的密度难以满足大面积检测需求。近年来可见-近红外光谱分析因其具有操作方便、采样速率快等优势，还可提供高分辨率和丰富的土壤光谱信息，成为SOM快速获取的热门途径。

国内外许多学者对SOM含量的光谱预测已开展了大量研究[8-12]，其中，光谱特征筛选方法[13]有效解决光谱信息量大、数据冗杂等造成预测模型效率低的问题，是光谱分析过程的重要环节。Vohland等[14]对德国不同类型的土样进行光谱分析，通过竞争性自适应重加权算法(competitive adaptive reweighted sampling，CARS)结合偏最小二乘回归(partial least squares regression，PLSR)建立了SOM含量预测模型。Viscarra Rossel等[15]对澳大利亚不同类型土壤有机碳的组成进行研究，基于决策树算法推导传递函数并构建预测模型来预测土壤总有机碳组分。Shi等[16]对不同省份的土类数据进行可见-近红外光谱分析，通过空间约束局部-偏最小二乘方法建立了SOM预测模型。张智涛等[17]基于分数阶微分结合支持向量机分类-随机森林构建荒漠土SOM含量预测模型。张娟娟等[18]分析了5种砂姜黑土样本的光谱特征，通过遗传算法筛选特征波长并结合支持向量机建立了预测模型。于雷等[19]采用不同变量筛选方法对汉江平原土样进行特征提取，并构建了SOM含量预测模型。Hong等[20]通过分数阶微分结合不同的变量筛选方法，分析了华中地区土样的光

收稿日期：2021-08-04，**修订日期**：2021-11-19

基金项目：国家自然科学基金项目(32071915)资助

作者简介：钟翔君，1995年生，中国农业大学工学院博士研究生 e-mail：xjzhong1004@163.com

*通讯作者 e-mail：yangli@cau.edu.cn

谱特征并构建了 SOM 含量预测模型。综上可以看出，利用特征变量筛选方法可以有效优化模型，但是不同类型土壤差异较大，构建的模型大多仅针对某种特定类型的土壤，对不同土壤类型的估测精度和适用性难以估测[21]。

华北平原是全国重要的粮食和经济作物区，同时是我国玉米主产区之一，通过研究该区域 SOM 信息指导播种、施肥及其他土壤改良作业，可有效降低生产投入、提高肥料利用率。基于此，以该区域北部的砂壤潮土为研究对象，以高灵敏度微型可见-近红外光谱仪采集并分析 300～2 500 nm 波长范围的光谱反射率，以多种波长选择方法筛选出特征波长，在对不同特征波长进行建模分析的基础上，找出反演 SOM 的优选方法，为该区域 SOM 的快速获取设备的设计方法和模型选择提供参考。

1　实验部分

1.1　土样采集与处理

研究区位于河北省廊坊市(39°19′N，116°17′E)中部平原地带，地处华北平原北部，是我国玉米生产主区之一。地势平坦，土壤类型以砂壤质为主，占土壤总面积 90%以上，光照充足，温差较大，这些独特的土壤及气候条件，使得该地区以种植玉米、花生、甘薯等作物为主。

在常年耕作的地块上以五点采样法采集 0～20 cm 耕作层的土壤样本，采集时去除地表残茬及砾石，并将采集的土样密封带回实验室进行处理。共采集了 60 份土样，每份大约 3 kg。为不破坏其内部成分，将取回的土样分别置于恒温干燥箱(DHG-9123A 型，上海)并在 40 ℃下烘干 24 h 至恒重，然后将烘干后的土壤研磨并过 1 mm 筛网后备用，分别供实验室分析及光谱测试用。

1.2　土样实验室检测

样品的 SOM 含量采用 TOC 元素分析仪(Elementar vario TOC cube，德国)进行测定。首先分别用万分之一电子天平(FA324 型，上海)称取研磨后的土样 15～20 mg，并置于准备好的直径 4 mm、高 6 mm 开口银囊中，随后在每个银囊滴入 1 mol·L^{-1} HCl 将土样完全浸润，静置 30 min 后转移至恒温干燥箱中干燥至恒重。将烘干后的银囊封口并用锡纸包裹、压实，随后依次投放于 TOC 元素分析仪中测量其 SOM 含量。为保证数据的有效性，每个样本准备 5 个重复并求均值，得到 SOM 含量统计结果如表 1 所示。

表 1　SOM 含量统计

Table 1　Statistics of SOM content

样本数量	最大值/(g·kg^{-1})	最小值/(g·kg^{-1})	平均值/(g·kg^{-1})	标准差/(g·kg^{-1})	变异系数/%
60	42.41	15.1	24.98	7.40	29.63

1.3　光谱数据采集

土壤样品的光谱数据用美国海洋光学公司的 QE Pro 高性能光谱仪及 NIR Quest 系列近红外光谱仪同步采集。其中，NIR Quest512-2.5 近红外光谱仪采用稳定性高的滨松铟镓砷化物(InGaAs)阵列探测器，可测量 900～2 500 nm 波长范围的光谱数据，光学分辨率为 9.0 nm。QE Pro 高性能光纤光谱仪采用低噪音的电子部分与 18 位 A/D 转换器，同时配备高容量的板存缓冲区，具有高灵敏度与宽动态范围特性，可大大提高光谱检测的准确度，同时具有很高的信噪比(大于 1 000∶1)和稳定性，可测量 185～1 100 nm 可见-近红外波长范围的光谱数据，满足高速及宽浓度范围的快速高精度的光谱测量，光学分辨率为 1.7 nm。

图 1 为光谱采集装置实物图，其中，光源为 5W HL-2000-FHSA 型卤钨灯光源(Ocean Optics，Inc.，美国)，其内部集成风扇冷却、快门和手动衰减器功能，可以保证持续稳定的光源输出；光源配合实验试级 QR200-12-MIXED 型全光谱一分三光纤(Ocean Optics，Inc.，美国)进行试验，该光纤主要包括 1 个入射光纤、2 个反射光纤(UV-Vis 和 Vis-NIR)和光纤探头组成；光纤探头固定在 Stage-RTL-T 型多功能检测台(Ocean Optics，Inc.，美国)光具座上，装有土样的培养皿置于检测台下方的样品支座上；通过笔记本电脑的 Ocean View 软件采集样本的反射光谱。

为降低环境及仪器噪声的影响，获取高精度的光谱反射率数据，样品测量前用美国海洋光学公司 99%漫反射标准白板进行校正，分别获取开启光源及关闭光源后得到的亮、暗光谱数据后，根据式(1)运算得到校正后的反射率数据。

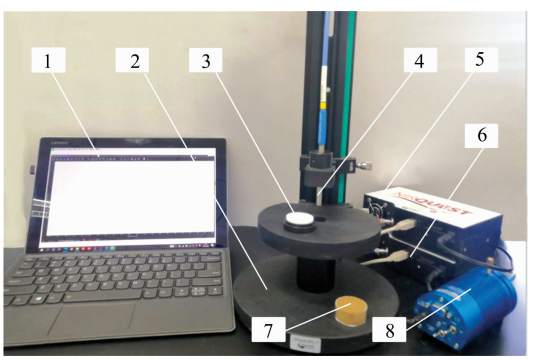

图 1　光谱采集装置图

1：电脑；2：多功能检测台；3：标准白板；4：光纤探头；5：NIR-Quest；6：QE-Pro；7：土壤样本；8：光源

Fig. 1　Spectral data acquisition device

1：PC；2：Multifunctional testing platform；3：Standard white plate；4：Fiber probe；5：NIR Quest；6：QE Pro；7：Soil samples；8：Light source

$$R_f = \frac{R_S - D_S}{W_S - D_S} \tag{1}$$

式(1)中：W_S 为开启光源得到的校正亮光谱，D_S 为关闭光源的得到的校正暗光谱，R_S 为样品初始反射率光谱，R_f 为校正后的样品反射率光谱。

经白板校准后，将不同 SOM 含量的土壤样本置于直径 3.5 mm 的培养皿中，通过调节检测台滑轨使光纤探头位于样品上表面，试验时每采集 5 个样本，用标准白板校正 1 次。其中，试验时光纤探头距标准白板及样品的上表面高度均为 3 mm。采用五点法选取样本 5 个位置采集光谱，每个位置连续采集 5 次的均值作为该位置的反射光谱，每个土壤样本准备 3 个重复，试验共得到 900 条光谱数据。

1.4 光谱数据处理

由于低于 380 nm 和高于 2 400 nm 波长的数据噪声较大，因此将上述波段从每组光谱数据中去除，只保留 380～2 400 nm 范围的光谱数据用于后续分析。为降低因仪器噪声、测量环境及土样表面粗糙度等因素对采样的影响，采用蒙特卡洛交叉验证法(Monte Carlo cross validation，MCCV)筛选异常数据并剔除。对剔除异常样本后的光谱数据采用 Savitzky-Golay(SG)平滑法进行预处理，并用作后续分析。

1.5 SOM 含量特征筛选方法

1.5.1 CARS 算法

CARS 方法首先抽取部分样本作为校正集，利用 MCCV 方法及 PLSR 构建模型，以模型中回归系数绝对值权重作为基准，保留模型中权重值大的特征波长并建立新的模型，经过多次计算，结合交叉验证确定交叉验证均方根误差(root mean square error of cross validation，RMSECV)小的波长集合为最优特征组合[19]。该方法可以降低冗余数据的干扰，从而选出优化后的变量组合，提高模型的稳定性及预测效果。

1.5.2 连续投影算法

连续投影算法(successive projections algorithm，SPA)首先将校正集波长矩阵投影到其他波长上，计算出每个波长点对应的投影值，以投影值为基准，筛选并保留最大投影值所在的波长，通过不断计算筛选出最优的波长组合。通过 SPA 方法选择的是冗余信息低及共线性少的变量组合，可以在一定程度上避免光谱信息重叠，有利于简化模型结构、提高运算效率。

1.5.3 其他特征提取算法

无信息变量消除(uninformative variables elimination，UVE)方法通过噪声信息加入到光谱数据中，通过交叉验证剔除无效信息变量并建立 PLSR 模型，通过对比系数矩阵的绝对值大小，确定出特征变量组合。变量组合集群分析法(variable combination population analysis，VCPA)采用二进制矩阵采样策略，利用指数衰减函数筛选无效变量，并依据交叉验证均方根误差最终选择出特征变量组合。

1.6 模型构建及检验

利用光谱-理化值共生距离法(sample set partitioning based on joint x-y distance，SPXY)将样本集按 7∶3 划分为建模集和预测集。分别以全波长及 CARS，SPA，UVE，VCPA 及 CARS-SPA 等不同方法筛选的特征波长为自变量，SOM 含量为因变量，基于 PLSR 结合交叉验证构建 SOM 含量预测模型。分别以决定系数(R^2)、校正均方根误差(root mean square error of calibration，RMSEC)、预测均方根误差(root mean square error of prediction，RMSEP)及剩余预测偏差(residual prediction deviation，RPD)等作为模型的评价指标[16]。其中，RPD 越大、R^2 越接近 1、RMSEC 与 RMSEP 越小表明模型效果越好。

2 结果与讨论

2.1 预处理结果分析

采用 MCCV 方法分别对不同样本的反射率数据进行异常筛选，其中每个样本的光谱数据作为一个独立的数据点，分别以样本的标准偏差作为 y 轴，平均预测误差为 x 轴，对所有样本光谱数据(数据点)进行筛选，不同样本的数据集分布结果如图 2 所示。从图中可以看出，不同土样光谱的数据集离散程度不一样，但大部分数据点在某范围内呈现集中分布。将远离大部分数据集分布的数据点(即平均误差和标准偏差越大)视为异常样本并予以剔除，留下的样本数据作为有效数据，用于后续分析与运算。经过异常值的筛选剔除，最终共保留了 809 个有效数据。

对剔除异常数据后的光谱进行 SG 平滑，得到平滑后的光谱曲线如图 3 所示。从图中可以看出，不同 SOM 含量的光谱反射率曲线总体变化趋势类似，随着波长的增加，光谱反射率呈现先增加后减小的趋势。同时，所有光谱曲线均在 1 410，1 910 和 2 200 nm 附近出现明显的水分吸收谷，这与

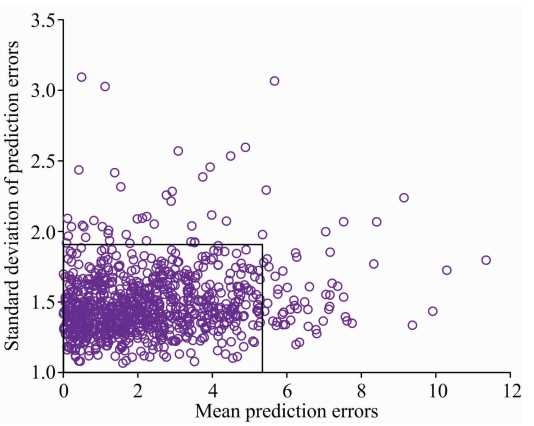

图 2 MCCV 异常值筛选结果

Fig. 2 Outlier filtering results of MCCV method

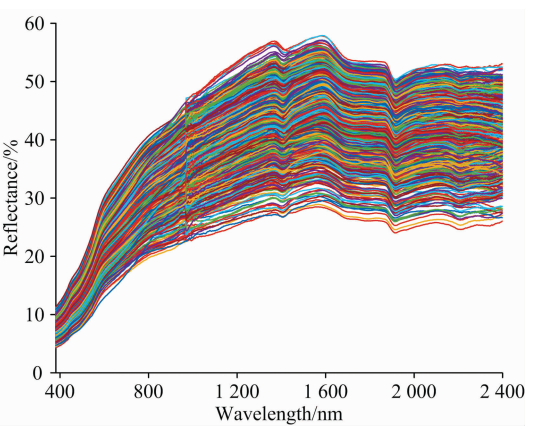

图 3 光谱反射率曲线

Fig. 3 Spectral reflectance curves

Laamrani 等[13]得到的光谱曲线特征结论类似。另外，由于两台光谱仪在 Ocean View 软件中进行拼接，所以在 970 nm 附近的反射率出现明显波动。

2.2　SOM 含量特征变量筛选

经过 CARS，SPA，CARS-SPA，UVE 及 VCPA 方法筛选变量结果如图 4 所示，从图中可以看出，不同筛选方法筛选出的波长数目及波长所在位置存在显著差异。从图 4(a)中可以看出，CARS 算法在采样次数增加至 200 次的过程中，特征变量的个数逐渐减少，其趋势由快速下降逐渐变为平缓，而 RMSECV 的值呈现先减小后增加的趋势，这与 Hong 等[20]对汉江平原土样进行光谱数据处理得到的结论类似。如图 4(a)中黑色竖直线标注，当采样次数为 46 次时 RMSECV 取得最小值，该采样次数对应筛选出的特征波长个数为 288 个，使得波段数目压缩至全波段数目的 23.4%，波长的分布如图 4(b)所示。将基于 SPXY 方法划分好的建模集和预测集数据通过 SPA 算法进行计算，结合图 4(c)可以看出，随着变量个数的增加，RMSECV 的值大致呈现快速减小然后趋于稳定的趋势，而当变量个数为 138 个时，其值达到最小，筛选出的特征变量分布如图 4(d)所示，波段数目压缩至全波段的 11.2%。相较于 CARS 方法，SPA 法筛选的变量共

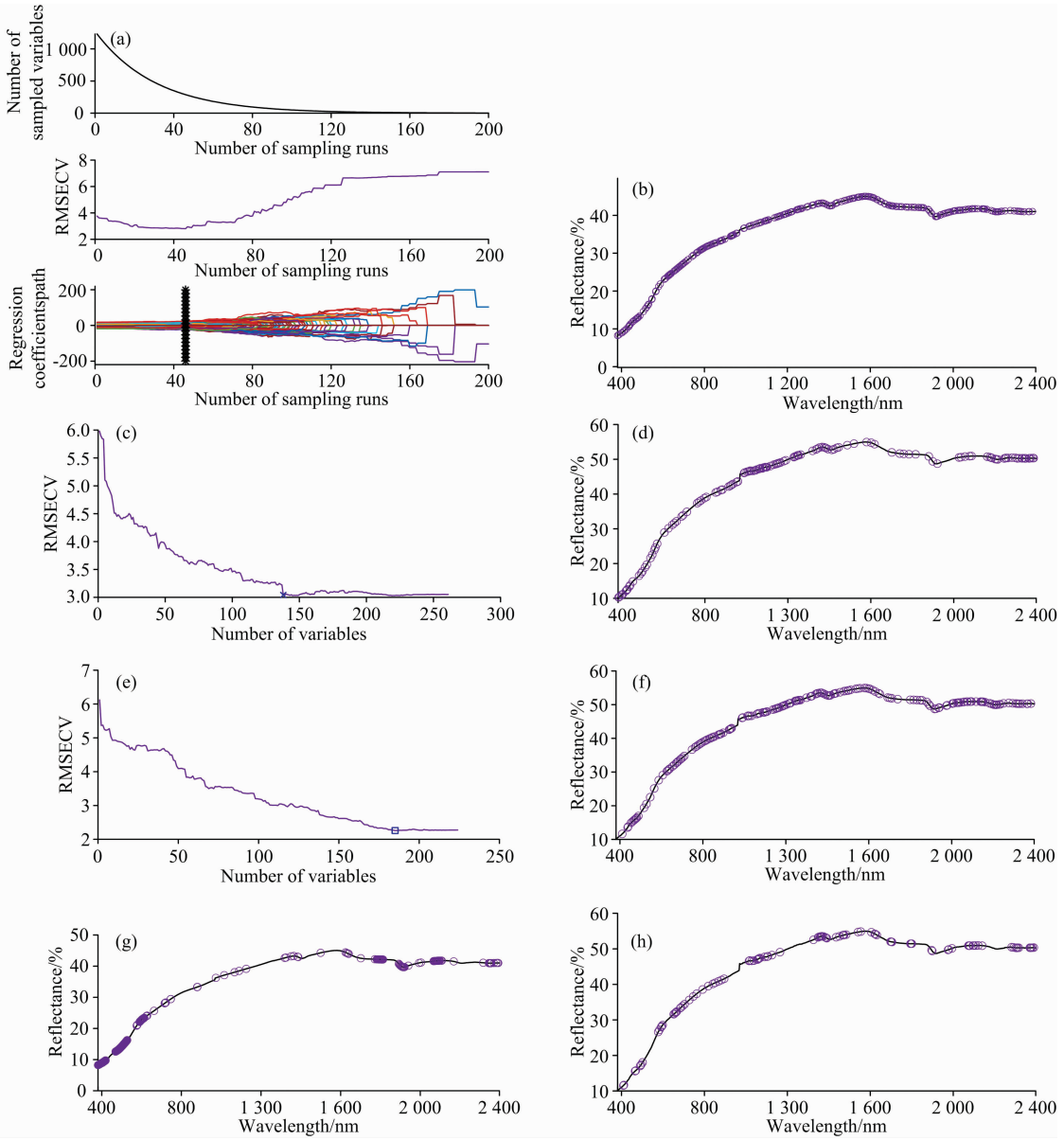

图 4　不同方法筛选特征变量结果

(a)：CARS 方法筛选变量；(b)CARS 筛选的变量分布；(c)：SPA 方法筛选变量；(d)：SPA 筛选的变量分布；(e)：CARS-SPA 方法筛选变量；(f)：CARS-SPA 筛选的变量分布；(g)：UVE 方法筛选的变量分布；(h)：VCPA 筛选的变量分布

Fig. 4　Results of different methods of screening characteristic variables

(a)：Variables selected by CARS method；(b)：The distribution of variables selected by CARS；(c)：Variables selected by SPA method；(d)：The distribution of variables selected by SPA；(e)：Variables selected by CARS-SPA method；(f)：The distribution of variables selected by CARS-SPA；(g)：The distribution of variables selected by UVE；(h)：The distribution of variables selected by VCPA

线性达到最小，极大地减少了建模所需的波长个数，而经过CARS方法筛选的变量个数虽然相较于全波长有所降低，但是波长数量仍然较多，在全波长范围内均有分布，所以采用SPA算法对CARS筛选后的变量进行二次筛选，进一步优化变量的结构，结果如图4(e)和(f)所示，共筛选出了185组特征波长，波段数目压缩至全波段的15.0%。通过比较UVE方法运算得到的系数矩阵，筛选出248组特征波长，波段数目压缩至全波段的20.1%，如图4(g)所示，该方法筛选出的波段较为集中。经过对比RMSECV的值，基于VCPA方法最终筛选出100组特征波长，波段数目压缩至全波段的8.1%，波长分布如图4(h)所示。

2.3 模型建立与检验

分别基于全波长及不同变量筛选方法得到的特征波长为自变量，SOM含量为因变量，采用SPXY法将光谱数据按7∶3分为建模集和预测集，结合留一法交叉验证，构建PLSR预测模型，得到不同模型的预测效果如图5所示。

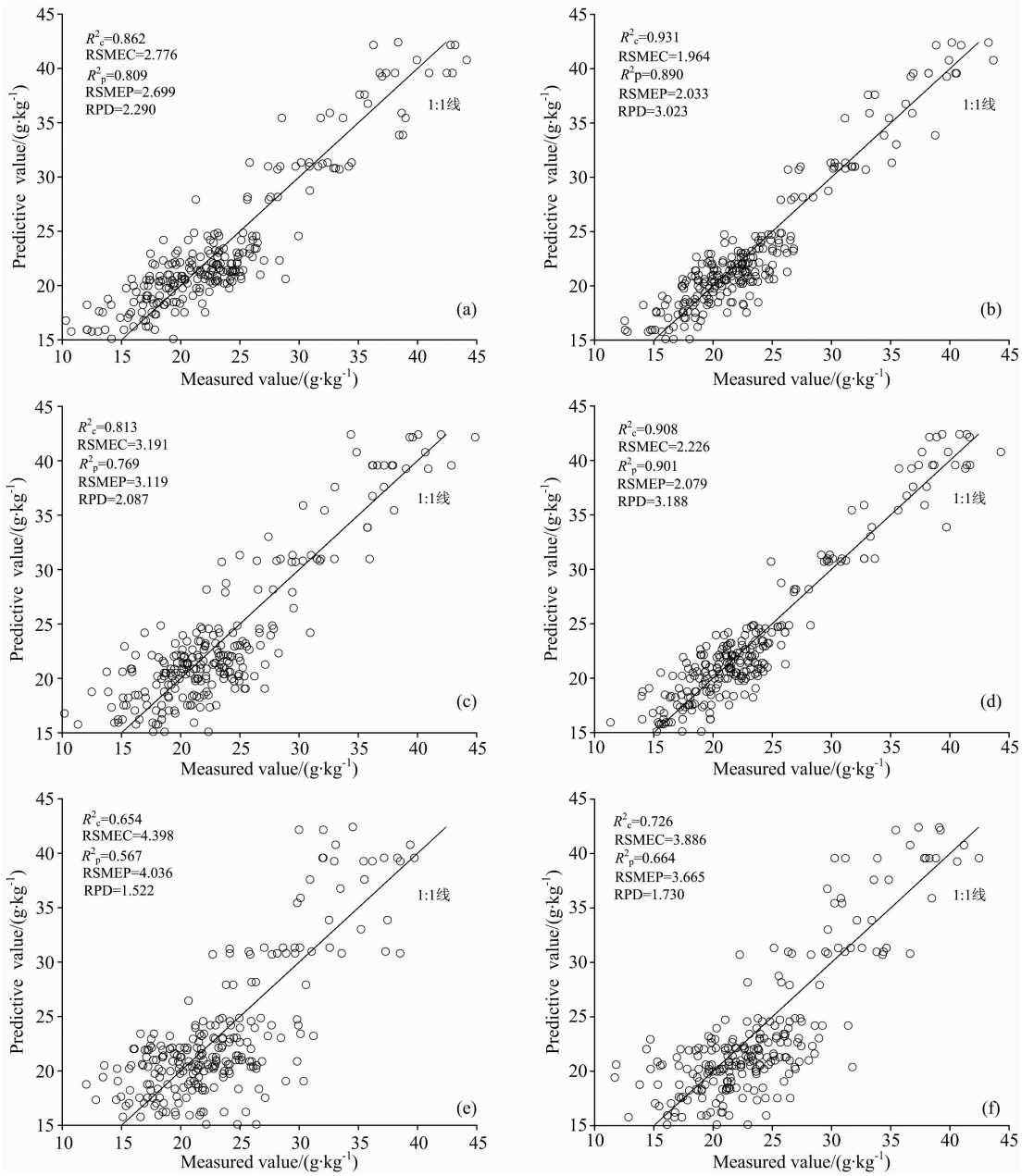

图5 不同波长PLSR建模结果

(a)：All-PLSR；(b)：CARS-PLSR；(c)：SPA-PLSR；(d)CARS-SPA-PLSR；(e)：UVE-PLSR；(f)：VCPA-PLSR

Fig. 5 Modeling results of PLSR using different variables

(a)：All-PLSR；(b)：CARS-PLSR；(c)：SPA-PLSR；(d)CARS-SPA-PLSR；(e)：UVE-PLSR；(f)：VCPA-PLSR

从图 5 中可以看出，除 UVE 和 VCPA 算法外，其余算法构建的模型均能实现 SOM 含量的有效预测，RPD 值均大于 2.0。其中，基于全波长建模的 R_C^2 和 R_P^2 分别为 0.862 和 0.809，RPD 为 2.290，可以进行 SOM 含量的预测，但是包含的波长数据信息冗杂，模型的运算效率较低。基于 SPA 方法构建的模型，虽然其建模的波长数量相较于全波长明显降低，但是模型的效果没有提高。而基于 UVE 和 VCPA 方法构建的模型，只能实现 SOM 含量粗略的估测，其 RPD 分别为 1.522 和 1.730。对比所有模型，基于 CARS 和 CARS-SPA 特征变量构建的 SOM 含量模型的预测效果最好，其中 R_C^2 分别为 0.931 和 0.908，RMSEC 分别为 1.964 和 2.226 $g \cdot kg^{-1}$，其真实值与预测值相关性较高，其中 R_P^2 分别为 0.890 和 0.901，RMSEC 分别为 2.033 和 2.079 $g \cdot kg^{-1}$，RPD 分别为 3.023 和 3.188，表明这两种方式可以实现砂壤潮土 SOM 含量的可靠预测。而 CARS-SPA 在 CARS 筛选变量的基础上进行二次筛选，波长的数量进一步降低，且其 R_P^2 为 0.901，高于 CARS 的 0.890，RPD 为 3.188，高于 CARS 的 3.023，因此，该方法取得的建模及预测效果最优。相较于全波长，基于 CARS-SPA 方法筛选的波长仅为全波长数量的 15%，不仅降低了无效信息对预测效果的干扰，且模型的运算效率得到了明显的提高，可以很好地实现该地区 SOM 含量的可靠预测，也为该区域 SOM 含量快速预测及仪器设计提供方法参考。

利用光谱可以实现 SOM 的预测，但是光谱波段多、数据信息冗杂，且土壤光谱反射率易受土壤质地、颜色及外部工作环境等多种因素的影响，均为 SOM 的快速预测及仪器设计增加了难度。本研究针对玉米主产区之一华北平原地带的砂壤潮土进行一致的处理以后，对比不同的波长筛选方法提取有效变量，降低了无效信息对预测效果的干扰，实现 SOM 含量预测。在后续研究中，需要考虑其他影响因素如光照、温度、土壤类型等对预测效果的影响，优化数据处理及建模方法，以进一步提高 SOM 的预测精度，实现田间 SOM 快速高精度检测。

3　结　论

以玉米主产区之一华北平原为研究区域，对该区域砂壤潮土进行可见-近红外光谱采集，通过不同的波长筛选方法提取有效变量并进行 SOM 含量预测，得到主要结论如下：

(1)不同方法筛选的波长数目及波长位置存在显著差异，CARS 和 SPA 算法选择的光谱特征在整个光谱范围都有分布，UVE 和 VCPA 筛选的波段较为集中，且基于 CARS-SPA 方法可以进一步优选特征变量，其特征波长仅为全波长数量的 15%。

(2)通过对比不同模型的建模及预测效果，除 UVE 和 VCPA 算法外，其余算法构建的模型均能实现 SOM 含量的有效预测，其 RPD 值均大于 2.0。

(3)基于 CARS-SPA 构建的 PLSR 模型效果最好，其 R_P^2 和 RPD 分别 0.901 和 3.188，均高于其他方法，不仅降低了无效信息对预测效果的干扰，且模型的运算效率得到了明显的提高，可以很好地实现该地区 SOM 含量的可靠预测。

References

[1] Kgel-Knabner I, Amelung W. Geoderma, 2021, 384(82): 114785.
[2] Kane D A, Bradford M A, Fuller E, et al. Environmental Research Letters, 2021, 16(4): 044018.
[3] YANG Li, YAN Bing-xin, ZHANG Dong-xing, et al(杨　丽，颜丙新，张东兴，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2016, 47(11): 38.
[4] He X T, Ding Y Q, Zhang D X, et al. Computers and Electronics in Agriculture, 2019, 162(7): 318.
[5] He X T, Ding Y Q, Zhang D X, et al. Computers and Electronics in Agriculture, 2019, 162(7): 309.
[6] Ding Y, Yang L, Zhang D, et al. International Journal of Agricultural and Biological Engineering, 2021, 14(2): 151.
[7] Bakr N, El-Ashry S M. Communications in Soil Science & Plant Analysis, 2018, 49(20): 2587.
[8] Morellos A, Pantazi X E, Moshou D, et al. Biosystems Engineering, 2016, 152: 104.
[9] Said N, Mouazen A M. Computers and Electronics in Agriculture, 2018, 151: 469.
[10] Hong Y, Liu Y, Chen Y, et al. Geoderma, 2019, 337: 758.
[11] Michael S, Christopher H, Bernard L, et al. Geoderma. 2019, 354: 113856.
[12] Roemer C, Rodionov A, Behmann J, et al. Journal of Plant Nutrition and Soil Science, 2014, 177(6): 845.
[13] Laamrani A, Berg A, Voroney P, et al. Remote Sensing, 2019, 11(11): 1298.
[14] Vohland M, Ludwig M, Thiele-Bruhn S, et al. Geoderma, 2014, (223-225): 88.
[15] Viscarra Rossel R A, Hicks W S. European Journal of Soil Science, 2015, 66(3): 438.
[16] Shi Z, Ji W, Rossel R A Viscarra, et al. European Journal of Soil Science, 2015, 66(4): 679.
[17] ZHANG Zhi-tao, LAO Cong-cong, WANG Hai-feng, et al(张智韬，劳聪聪，王海峰，等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2020, 51(1): 156.
[18] ZHANG Juan-juan, XI Lei, YANG Xiang-yang, et al(张娟娟，席　磊，杨向阳，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2020, 36(17): 135.
[19] YU Lei, HONG Yong-sheng, ZHOU Yong, et al(于　雷，洪永胜，周　勇，等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2016, 32(13): 95.

[20] Hong Y S, Chen Y Y, Yu L, et al. Remote Sensing, 2018, 10(3): 479.
[21] Liu Y, Jiang Q, Fei T, et al. Remote Sensing, 2014, 6(5): 4305.

Prediction of Organic Matter Content in Sandy Fluvo-Aquic Soil by Visible-Near Infrared Spectroscopy

ZHONG Xiang-jun[1,2], YANG Li[1,2*], ZHANG Dong-xing[1,2], CUI Tao[1,2], HE Xian-tao[1,2], DU Zhao-hui[1,2]
1. College of Engineering, China Agricultural University, Beijing 100083, China
2. Key Laboratory of Soil-Machine-Plant System Technology of Ministry of Agriculture and Rural Affairs, Beijing 100083, China

Abstract Soil Organic Matter (SOM) is a crucial soil parameter that affects the sowing rate. Real-time control of the sowing rate based on SOM information is the cutting-edge research area of planting technology, which can make full use of land resources to tap the yield potential, accurately and adequately adjust the number of seeds to maximize the return. This article focuses on the North China Plain, one of the main corn-producing areas, as the study area, and the sandy loam soil in this area has been collected by visible-near infrared (300～2 500 nm) spectra. Monte Carlo cross-validation is used to eliminate abnormal samples, and the Savitzky-Golay convolution smoothing method is used to smooth and denoise the spectral data. Respectively through Competitive adaptive reweighted sampling (CARS), Successive projections algorithm (SPA), Competitive adaptive reweighted sampling-Successive projections algorithm (CARS-SPA), Uninformative variables elimination (UVE) and Variable Combination population Analysis (VCPA), and other wavelength screening methods to extract effective variables. Combined with Partial least squares regression (PLSR), the SOM content prediction models of full wavelength and characteristic wavelength were established respectively. The results showed significant differences in the number of wavelengths and wavelength positions screened by different methods. The spectral features selected by the CARS and SPA algorithms are distributed in the spectral range, while the bands selected by UVE and VCPA were concentrated. Moreover, the characteristic variables could be further optimized based on the CARS-SPA method, and the characteristic wavelength was only 15% of the total wavelength. By comparing the modeling and prediction effects of different models, except for the UVE and VCPA algorithms, the models constructed by the other algorithms can all effectively predict the SOM content, and their RPD values were all greater than 2.0. The PLSR model based on CARS-SPA has the best performance. Its R_P^2 and RPD were 0.901 and 3.188 respectively, higher than other methods. It reduces the interference of invalid information on the prediction effect, but the computational efficiency of the model is significantly improved, which can realize the reliable prediction of SOM content in this area. This research can provide method references for rapid prediction of SOM content and instrument design.

Keywords Soil organic matter; Seeding; Visible-near infrared; Sandy fluvo-aquic soil; CARS-SPA

(Received Aug. 4, 2021; accepted Nov. 19, 2021)

* Corresponding author