文章编号： 1672-8785(2015)02-0042-07

基于可见－近红外光谱特征波长选择的土壤有机质快速检测研究

杨海清 * 祝 旻

(1. 浙江工业大学信息工程学院，浙江 杭州 310023)

摘 要：选择光谱特征波长进行建模可以减少冗余波长的干扰，提高模型的预测精度。采用小波阈值消噪法对采集的 104 个土壤样本光谱数据进行了预处理，并通过间隔偏最小二乘法、无信息变量消除、连续投影算法和群智能算法等 9 种方法筛选了建模波长。结果表明，小波阈值消噪法能有效降低光谱中的噪声。利用波长选择方法筛选建模波长不仅能减少建模变量的个数，而且还能提高模型的预测精度，特别是离散粒子群优化算法利用 26 个波长进行建模，预测决定系数达到了 0.81，预测的相对标准误差为 2.31。实验结果证明，通过对光谱波长进行选择不但可以降低模型的复杂度，还能有效预测土壤有机质达的含量。

关键词：可见－近红外光谱；土壤有机质；小波消噪；波长选择；群智能算法

中图分类号： S123 **文献标志码**： A **DOI**: 10.3969/j.issn.1672-8785.2015.02.008

Study of Rapid Detection of Soil Organic Matter Based on Characteristic Wavelength Selection of Visible-near Infrared Spectra

YANG Hai-qing *, ZHU Min

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310032, China)

Abstract: Selecting the characteristic wavelength in spectra for modeling can reduce the interference by redundant wavelengths and improve modeling accuracy. The spectral data of 104 soil samples collected are preprocessed by a wavelet threshold de-noising method. The wavelengths are selected for modeling by 9 wavelength selection methods including interval partial least squares, uninformative variable elimination, successive projection algorithm and swarm intelligence algorithm. The results show that the wavelet threshold de-noising method can reduce the noise in spectra effectively. Using wavelength selection methods to select the wavelengths for modeling not only can reduce modeling variables, but also can improve the prediction accuracy of the model. Particularly, since the discrete particle swarm optimization algorithm uses 26 wavelengths for modeling, its prediction determination coefficient reaches 0.81 and its relative standard prediction error is 2.31. The selection of spectral wavelengths not only can reduce the complexity of the model, but also can effectively predict the organic matter content in soil.

Key words: VIS-NIR spectroscopy; soil organic matter; wavelet denoising; wavelength selection; swarm intelligence algorithm

收稿日期：2014–12–31

基金项目：国家自然科学基金项目 (41271234)；浙江省自然科学基金项目 (LY13F010008)；浙江省人力资源厅 2012 年留学人员科技择优资助项目

作者简介：杨海清 (1971-)，男，浙江温岭人，副教授，主要从事先进传感技术和传感器研究。

* 通讯作者 E-mail: yanghq@zjut.edu.cn

0 引言

土壤有机质含量 (Soil Organic Matter, SOM) 与土壤质量密切相关，是衡量土壤肥力的重要指标。如何快速实时测定 SOM 含量已经成为发展精准农业 [1] 的关键问题之一。可见 - 近红外 (Vis–NIR) 光谱技术克服了传统化学检测方法操作复杂、消耗时间长的缺点，能够取得较高的测量精度，为 SOM 含量的快速测定提供了一种有效的方法。 Conforti 等 [2] 对采集自不同地区的 215 个土壤样本进行了 PLS 建模， SOM 预测结果 R^2 达到了 0.84 ；胡林等 [3] 对取自果园的 81 个土壤样本的 SOM 含量进行了测定，相关系数达到了 0.95 ，取得了良好的效果。选择特定的 Vis–NIR 光谱波长进行了建模，可以剔除冗余波长，减少其他组分信息波长的干扰，提高模型的预测精度。 Yang 等 [4] 对采自农场的 122 个土壤样本的有机碳含量进行了测定，并用提取的 5 个特征波长进行建模，决定系数 R^2 达到了 0.91 。

目前，实验室分析时使用的光谱仪价格昂贵，制约了光谱技术在农业生产中的推广。本文使用一种价格相对较低的光谱仪采集土壤光谱，并采用多种方法选择建模波长，从而对 SOM 进行预测。同时，在筛选土壤光谱建模波长时引入了离散蚁群优化算法和离散粒子群优化算法两种群智能优化算法，进一步提高了有机质含量预测模型的精度。

1 材料与方法

1.1 土壤光谱的获取

土壤样本由浙江大学农业遥感与信息技术应用所提供，共 104 个土壤样本，分别采自浙江省十个地区 0~20 cm 的土壤肥力层，土壤类型为水稻土和滨海盐土。待土壤样本自然风干后，研磨并过 2 mm 孔筛。将每个土壤样本用四分法分为两份：一份采用重铬酸钾容量法 - 外加热法测定土壤有机质的化学值，另一份用于光谱测试。土壤光谱采集使用美国海洋光学 (Ocean Optics, Inc) 生产的 USB4000 型 CCD 阵列光谱仪，其最大响应波段为 345.3 nm ~1046.1 nm ，光谱分辨率为 0.3 nm ~10.0 nm ；所使用的光源为海洋光学生产的 HL–2000 型卤钨灯光源，光谱范围为 360 nm~2000 nm ，输出功率为 7 W ，光谱仪和光源用双分叉光纤连接。在暗室中采集土壤光谱，将各个土壤放置于样本皿内，抹平土壤表面并去除植物根系等杂质，将光谱仪的光纤探头置于距土壤表面上方 15 cm 处，视场角为 25 ° 。对每个土壤样本采集 10 次光谱，然后取平均光谱。

1.2 波长选择方法

1.2.1 间隔偏最小二乘法

间隔偏最小二乘法 [5](Interval Partial Least Squares, iPLS) 是由 PLS 发展来的扩展方法。其主要步骤为：首先将光谱等分为若干个子区间，并在每个子区间内建立 PLS 回归模型；然后以留一交叉验均方根误差 (Root Mean Square Error, $RMSE_{CV}$) 为指标从所建立模型中选出精度最高的一个作为建模区间；最后，以该区间为中心向光谱两侧延展或者削减波长，得到一个建模子区间。

组合间隔偏最小二乘回归 (Synergy interval Partial Least Squares, SiPLS) 是 iPLS 的一种改进方法。SiPLS 对不同区间进行组合，克服了 iPLS 只选取单一区间用于建模的缺点，它是通过联合几个预测精度较高的区间来共同预测组分的含量的。向后间隔偏最小二乘法 (Barkward interval Partial Least Squares, BiPLS) 将全谱段分为若干个子区间，并对每个子区间进行 PLS 建模，将各个子区间模型按照交叉验证均方根误差的大小排序，去除最大子区间，利用剩下的区间再次进行 PLS 建模校验。不断去除 $RMSE_{CV}$ 最大的子区间，当 $RMSE_{CV}$ 最小时，所对应的子区间即为最佳建模子区间。移动窗口偏最小二乘法 (Moving Window Partial Least Squares, MWPLS) 首先确定光谱窗口宽度为 w ，选取 w 个波长； n 个波长的光谱中，从第一个波长开始一直移动到最后，在光谱上共选取 $n-w+1$ 个宽度为 w 的窗口。然后对这些子区间 PLS 建模，以 $RMSE_{CV}$ 作为判定条件，从中找出一个或若干个波长子区间。

1.2.2 无信息变量消除－连续投影算法

无信息变量消除 (Uninformative variables elimination, UVE) 用于消除光谱中对建模结果影响较小甚至有干扰的波长，是一种基于 PLS 回归系数的波长选择方法。 PLS 模型中的回归系数向量 b_j 与光谱波长相对应，采用公式 $S_j = mean(b_j)/STD(b_j)$ 衡量每个波长 j 的重要性。当均值 $mean(b_j)$ 越大而标准差 $STD(b_j)$ 越小时，如果 S_j 越大，就说明该波长越重要。因此可以根据 S_j 设定一个阈值，以去除光谱中无信息的波长。为了得到合适的阈值，可在光谱矩阵中引入一组人工噪声矩阵，由于引入的随机变量的稳定性应小于光谱波长，因此可得到消除的阈值为 $S_{max} = max(abs(S_{noise}))$ 。

连续投影算法 (Successive Projections Algorithm, SPA) 可以从光谱中寻找含有最低冗余信息限度的变量组合，因而可使多元线性回归中变量的共线性最小化 。作为一种前向循环选择方法， SPA 首先从一个波长开始计算其在选入波长中的投影，然后将投影向量最大的波长引入到组合中，如此循环 N 次，从而提取 N 个波长组合。 SPA 的选取规则为每次新选中的波长都与前一个的线性关系最小。由于 SPA 的计算量较大，且选中波长的信噪比有可能较低，因此人们往往把 UVE 和 SPA 联合起来使用，先使用 UVE 剔除无信息的变量，再用 SPA 提取用于建立定量分析模型的波长 [6] 。

1.2.3 群智能算法

群智能算法 (Swarm Intelligence Algorithm, SIA) 是一种概率搜索算法，其基本思想是通过模拟自然界生物的群体行为来构造随机优化算法。二进制蚁群算法 (Binary Ant Colony Optimization, BACO) 和二进制粒子群算法 (Binary Particle Swarm Optimization, BPSO) 是蚁群算法和粒子群算法的离散二进制版本，可用于组合优化领域。BACO 和 BPSO 将搜索优化过程模拟成个体觅食过程，用搜索空间中的点即光谱的波长点模拟自然界中的个体，将求解问题的目标适应度函数度量成个体对环境的适应能力，将个体的觅食过程类比为使用较好的可行解替代较差的可行解的迭代搜索过程 [7] 。 BACO 和 BPSO 的具体算法过程见文献 [8,9] 。本文采用 $F = (1+RMSE_{CV}$ 作为适应度函数，其中 R^2_{CV} 为交叉验证决定系数。

1.3 结果评价指标

将 PLS 用于光谱建模。建模效果用建模决定系数、预测决定系数、校正均方根误差 RMSEC 、预测均方根误差 RMSEP 及相对分析误差 RPD 等指标评价。决定系数越大，均方根误差越小，说明所建模型的精度越高。当 RPD 大于 2.5 时，模型具有良好的预测能力。此外，$1.4 < RPD < 2$ 表明模型可对样品作粗略估测，而 $RPD < 1.4$ 则表明模型无法对样品预测 [10] 。本实验的数据处理是在 Matlab R2012a 中完成的。

2 结果与讨论

2.1 光谱预处理

利用公式 $A = \log(1/R)$ 将土壤漫反射光谱 R 转换为吸光度光谱。图 1 为 10 个不同地区土壤样本的典型吸光度曲线，光谱在 475 nm 波长左右出现一个吸收峰，然后呈下降趋势，其中 475 nm~525 nm 波段的光谱下降较快，之后下降趋势较为平缓。同时可以看出，各条曲线在测量范围内抖动比较明显，特别是在光谱两端抖动十分剧烈，仪器产生的噪声较大。因此去除两端噪声较大的部分，留下 500 nm~950 nm 波段中共 2342 个波长点用于分析。

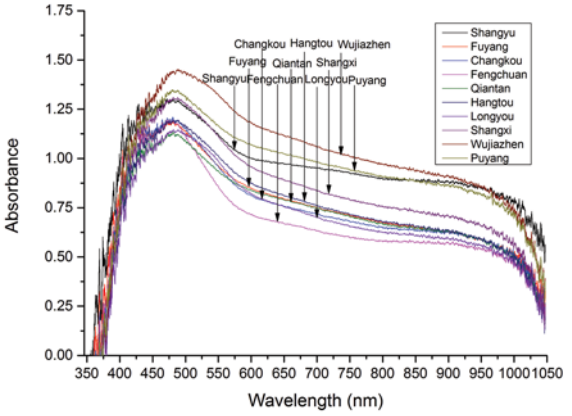
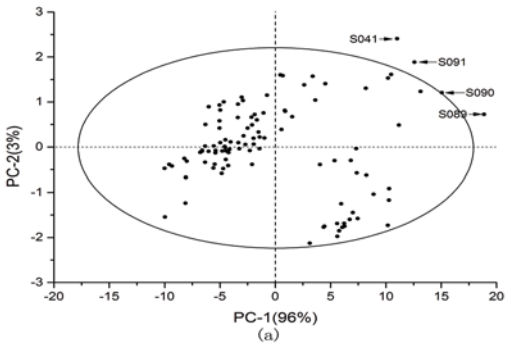


图 1 10 个不同地区典型样本土壤的吸光度光谱

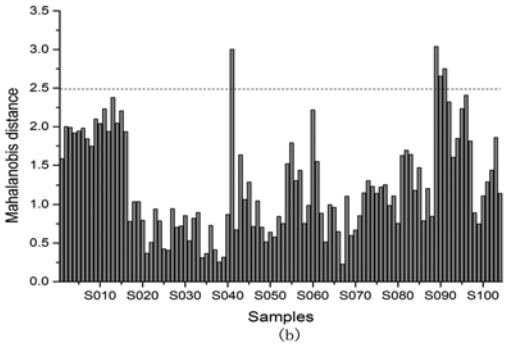
采用 PCA 结合马氏距离法剔除异常样本。在对土壤光谱的 PCA 进行降维后，前两个主成分 PC-1 和 PC–2 的累积贡献率 (该部分的累积方差在总方差中所占的百分比) 达到了 99% 。图 2(a) 为 PC–1 和 PC–2 的分布图，有 4 个样本在 Hotelling T2 椭圆之外，因此将这 4 个土壤样本初步判定为异常样本。图 2(b) 为 104 个样本的马氏距离分布图，样本 S041 、S089 、S090 和 S091 的马氏距离超过了所有样本的马氏距离平均值即剔除阈值的两倍，一般认为这 4 个样本为异常样本，结果与通过 PCA 判定的异常样本相同。

将 4 个异常样本剔除后还剩 100 个土壤样本。按有机质含量从低到高的顺序对它们进行排列，采用 2:1 的比例将其划分为建模样本集和预测样本集。所划分的建模样本集的组分含量应当涵盖预测样本集的组分含量，这样建模集所建立的模型就能够有效描述预测样本集中有机质和光谱之间的关系 [11] 。表 1 为划分建模集和预测集的有机质含量统计表，建模集中的有机质含量范围为 6.94~60.5g/Kg ，覆盖了预测集中的有机质含量范围 7.01~60.5 g/Kg 。

用小波阈值消噪法 [12] 进行预处理以消除光谱中的噪声。对不同小波母函数及分解尺度进行组合，对消噪后的光谱建立 PLS 模型，并对预测集进行预测。最终， sym6 小波 7 层分解对光谱信号的消噪效果最好。图 3(a) 、图 3(b) 为 500 nm~950 nm 波段原始光谱与消噪后的光谱对比。小波消噪后的光谱除了 850 nm~950 nm 之间略有抖动外，整条谱线变得比较平滑。这与原始光谱在该波段的噪声较大有关。



(a)104 个土壤样本的吸收光谱 PCA 得分图

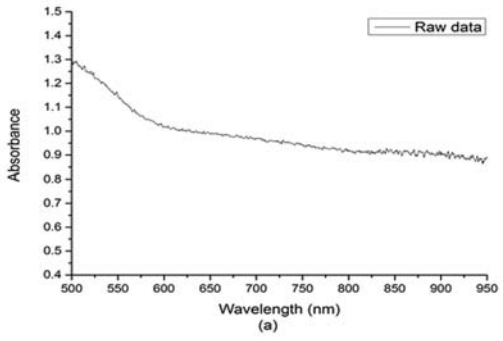


(b)104 个土壤样本的吸收光谱马氏距离分布图

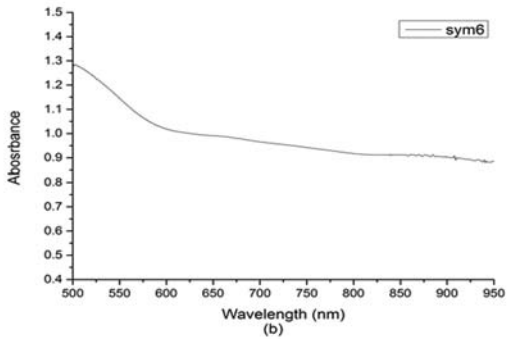
图 2　异常样本判别

表 1　土壤样本有机质含量的统计情况 (g/Kg)

样本集	样本个数	最小值	最大值	平均值	标准偏差
建模集	66	6.94	60.5	28.56	14.13
预测集	34	7.01	60.5	28.80	13.33



(a) 原始光谱



(b) sym6 小波 7 层分解消噪后的光谱

图 3　小波阈值消噪前后光谱对比

2.2 波长选择

首先采用 4 种间隔 PLS 方法选择建模波段，结果见表 2 。除 iPLS 外，其余 3 种方法对 SOM 的预测效果都有所提升，其中 BiPLS 的预测效果最佳，其选出的建模波段也最少。采用 iPLS 方法时，由于只选一个波段用于建模，间隔太多，导致选择的波长数量较少，不足以解释有机质含量。BiPLS 将光谱分段后每次剔除一个 RMSEcv 最小的波段，直到剩下最后一个波段，因此 BiPLS 可以将整条光谱划分为更多的波长间隔区间，从而减少所选择光谱区间无关波长对建模效果的干扰。间隔 PLS 波长选择方法基于等间隔波段选择， UVE 则考虑了每个波长的重要性，它先将低于该重要性阈值的波长予以去除，然后将得到的 1128 个波长用于建模。

以上几种波长选择方法选出的 PLS 建模波长均为一个或多个连续波段，选出的 PLS 建模波长依然很多。由于相邻的波长高度相关，包含了大量冗余信息。同时由于连续波段中一些与有机质相关性较弱的波长会被用于建模，或者部分相关性较高的波长包含在其他波段中而未被选中，因此采用上述几种波长选择方法进行 PLS 建模的预测效果提高不明显。 SPA 选择建模波长过程时，以 α=0.25 为标准 F 检验， $RMSE$ 波形开始随着所选变量数的增加迅速下降，随后趋于水平 [13] 。使用 SPA 分别对全谱段和 UVE 选出的波段进行筛选后，均只需用 5 个波长进行建模，而且将这两种方法选出的波长用于有机质 PLS 建模的预测效果均优于全谱建模的预测效果。

群智能算法属于随机优化算法，它能够在合理的时间范围内逼近问题的最优解，同时能以较大的概率找出问题的最优解。BACO 和 BPSO 在搜索过程中可能陷入局部最优，因此对 BACO 和 BPSO 多次运行选出一组波长组合，以用于土壤有机质含量预测。本研究中， BACO 和 BPSO 迭代数设为 200 ，分别运行 50 次，选取一组最好的波长组合。算法运行时，适应度函数首先随着迭代次数的增加而减小，随后迭代至 120 次左右时趋于稳定不变，如图 4 所示。对 BACO 进行初始设定时，选择了 5~20 个波长，其中选中 14 个波长时的建模效果最好。BPSO 的初始波长设定为 10 个，随着迭代次数的增加， BPSO 有可能会增加或者减少选中的波长数量。最终， BPSO 选中 26 个波长用于 PLS 建模时的效果最佳。

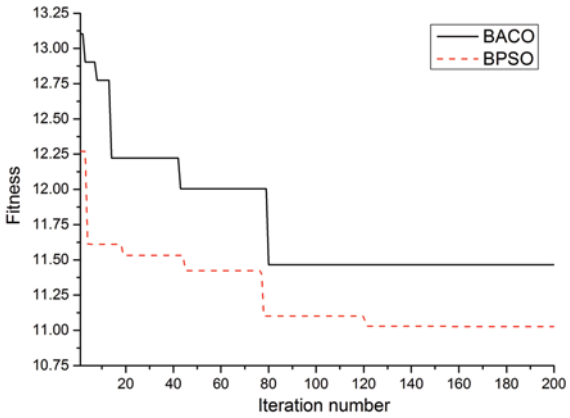


图 4 BACO 和 BPSO 波长选择过程

表 2 为采用 9 种波长选择方法进行土壤有机质 PLS 建模的结果对比，4 种间隔 PLS 选择方法选出的建模波长依然较多； SPA 、 UVE–SPA 、 BACO 和 BPSO 等 4 种方法选出波长的建模效果均优于 iPLS 、 SiPLS 、 BiPLS 和 UVE 等 5 种连续波段选择方法，同时也优于小波消噪预处理后的全谱建模效果。特别是 BPSO 选出的 26 个波长建模效果最好， R_p^2 由原始光谱的 0.73 提高到了 0.81 ， RPD 由 1.96 提高到了 2.31 。图 5 为上述 9 种方法选择出的建模波长对比。受本次测量光谱仪测量波长范围的限制，采集的土壤光谱主要集中在可见光波段。纪文君等 [14] 在总结了 7 组不同地区的土壤样本后，认为土壤有机质响应波段为 600 nm~800 nm ，彭杰等 [15] 也认为橙黄光波段为有机质敏感波段，并总结得出 800 nm~2400 nm 波段没有有机质引起的吸收峰，这也说明本实验中采用 500 nm~950 nm 波段进行光谱分析是可行的。在 9 种波长选择方法中， iPLS 选择的波长在 545 nm~590 nm 之间，因而预测结果最差。

表 2　不同波长选择方法 PLS 建模及预测结果对比

方法	选中波长数	因子数	建模集				预测集			
			R_C^2	$RMSE_C$	R_{CV}^2	$RMSE_{CV}$	R_P^2	$RMSE_P$	RPD	Bias
-	-	5	0.77	6.30	0.66	8.02	0.73	7.19	1.96	-0.17
iPLS	234	4	0.57	8.65	0.48	9.77	0.53	9.53	1.48	2.29
SiPLS	1171	5	0.76	6.42	0.66	7.58	0.74	7.09	1.99	-0.53
BiPLS	468	5	0.81	5.75	0.73	6.35	0.74	7.07	2.00	-1.40
MWPLS	1382	4	0.76	6.53	0.70	7.35	0.74	7.10	1.99	-0.67
UVE	1128	5	0.75	6.63	0.65	8.12	0.75	6.98	2.02	-0.30
SPA	5	5	0.76	6.46	0.67	7.57	0.77	6.71	2.11	0.29
UVE-SPA	5	5	0.78	6.43	0.68	7.46	0.78	6.55	2.16	-0.15
BACO	14	7	0.80	5.97	0.71	7.14	0.78	6.50	2.18	0.38
BPSO	**26**	**8**	**0.81**	**5.71**	**0.74**	**7.16**	**0.81**	**6.13**	**2.31**	**-1.01**

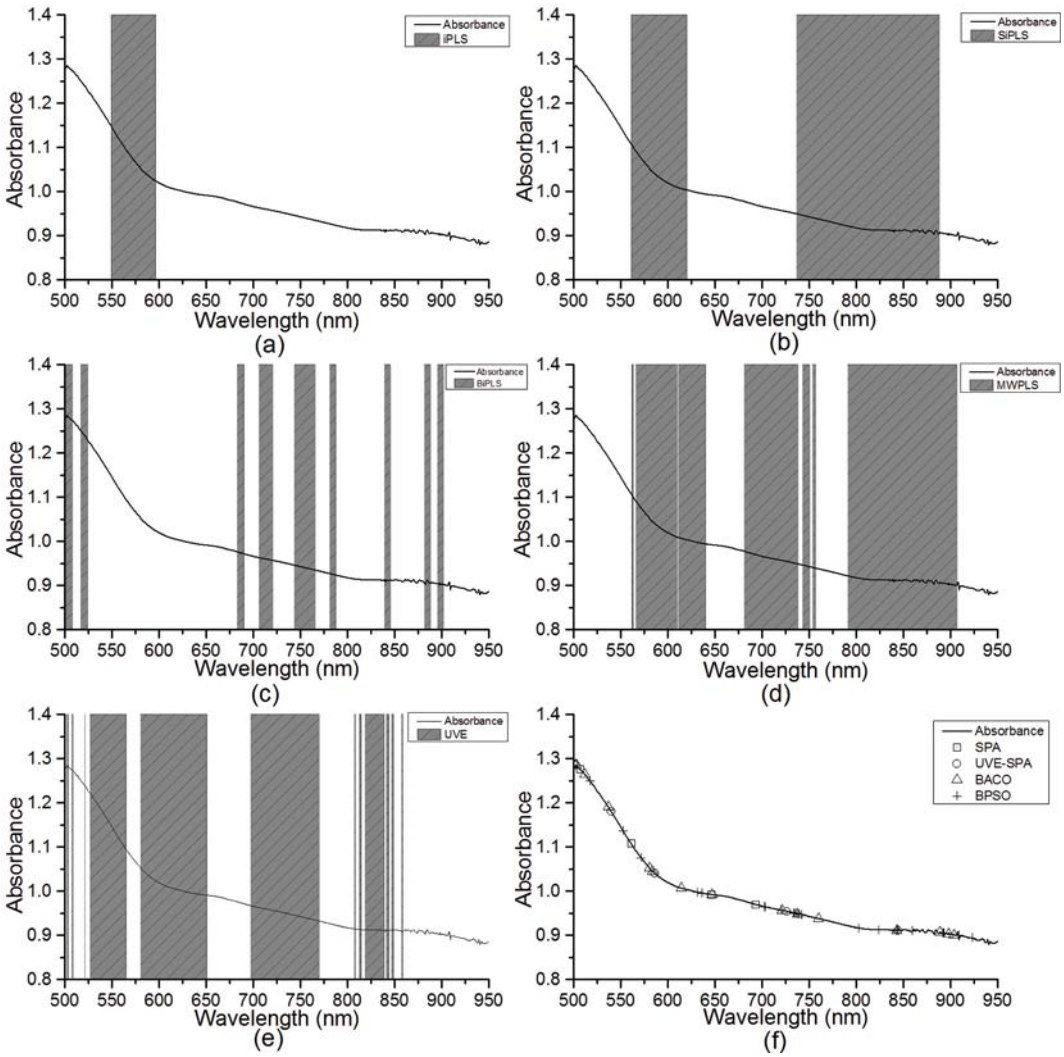


图 5 不同波长选择方法选中波长对比，其中 (a) iPLS (b)SiPLS (c)BiPLS (d)MWPLS (e)UVE (f)SPA 、 UVE-SPA 、 BACO 和 BPSO

3 结论

采用小波阈值消噪法对土壤光谱进行了预处理，并结合 9 种光谱波长选择方法建立了土壤有机质预测模型，取得了较好的预测效果，特别是 BPSO 方法，只需用不到 2% 的光谱波长进行建模就能有效预测土壤的有机质含量，降低了模型的复杂度和计算量。通过实验，我们得到了以下结论：

(1) 小波阈值消噪法可以减少光谱中噪声较大的问题，通过小波阈值消噪预处理可以使光谱变得平滑，并能提高模型的精度；

(2) 利用 345.3 nm~1046.1 nm 光谱波段可以建立预测结果较好的土壤有机质检测模型；

(3) 利用群智能优化算法选择建模波长，由于其选择过程中具有一定的随机性，因此有可能陷入局部最优解。本研究中通过多次运行选择一组最优波长组合。

利用特征波长建立的模型降低了计算量，实验所选用的光谱仪具有较强的二次开发能力，为后续开发一种基于 ARM 的便携式模块化土壤有机质检测仪器奠定了基础。

参考文献

[1] 何东健，何勇，李明赞，等．精准农业中信息相关科学问题研究进展 [J]. **中国科学基金**, 2011, **1**(1): 10–16.

[2] Conforti M, Buttafuoco G, Leone A P, et al. Studying the Relationship between Water-induced Soil Erosion and Soil Organic Matter Using Vis-NIR Spectroscopy and Geomorphological Analysis: A Case Study in Southern Italy[J].*CATENA*, 2013, **110**(0): 44–58.

[3] 胡林，丘耘，周国民．基于可见 / 近红外光谱的土壤有机质快速测定方法的研究 [J].**安徽农业科学**, 2012, **40**(12): 7123–7124.

[4] Yang H, Kuang B, Mouazen A. Quantitative Analysis of Soil Nitrogen and Carbon at a Farm Scale Using Visible and Near Infrared Spectroscopy Coupled with Wavelength Reduction[J].*European Journal of Soil Science*, 2012, **63**(3): 410–420.

[5] Chen T, Li Z, Hu F R, et al. Quantitative Analysis of Mixtures Using Terahertz Time-Domain Spectroscopy and Different PLS Algorithms[J].*Advanced Materials Research*, 2013, **804**: 23–28.

[6] Balabin R M, Smirnov S V. Variable Selection in Near-infrared Spectroscopy: Benchmarking of Feature Selection Methods on Biodiesel Data[J].*Analytica Chimica Acta*, 2011,**692**(1): 63–72.

[7] 胡中功，李静．群智能算法的研究进展 [J].**自动化技术与应用**, 2008, 27(2): 13–15.

[8] Shamsipur M, Zare S V, Hemmateenejad B, et al. Ant Colony Optimisation: a Powerful Tool for Wavelength Selection[J].*Journal of Chemometrics*, 2006, **20**(3-4): 146–157.

[9] 夏阿林，叶华俊，周新奇，等．基于粒子群算法的波长选择方法用于苹果酸度的近红外光谱分析 [J].**分析试验室**, 2010, **29**(9): 12–15.

[10] 张娟娟，田永超，姚霞，等．同时估测土壤全氮、有机质和速效氮含量的光谱指数研究 [J].**土壤学报**, 2012, **49**(1): 50–59.

[11] Chen Q, Zhao J, Liu M,et al. Determination of Total Polyphenols Content in Green Tea Using FT-NIR Spectroscopy and Different PLS Algorithms[J]. *Journal of Pharmaceutical and Biomedical Analysis*, 2008,**46**(3): 568–573.

[12] Mappe-Fogaing I, Joly L, Durry G, et al. Wavelet Denoising for Infrared Laser Spectroscopy and Gas Detection[J].*Applied spectroscopy*, 2012,**66**(6): 700–710.

[13] Wu D, He Y, Nie P, et al. Hybrid Variable Selection in Visible and Near-infrared Spectral Analysis for Non-invasive Quality Determination of Grape Juice[J].*Analytica Chimica Acta*, 2010, **659**(1): 229–237.

[14] 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.

[15] 彭杰，周清，张杨珠，等．有机质对土壤光谱特性的影响研究 [J]. **土壤学报**, 2013,**50**(003): 517–524.