中国农业信息
第 31 卷第 1 期 2019 年 2 月　　2019，31（1）：58-71
China Agricultural Informatics
Vol.31，No.1 Feb.，2019

基于 SVR 和 PLSR 的土壤有机质高光谱估测模型研究*

沈兰芝[1]，高懋芳[2※]，闫敬文[1]，姚艳敏[2]
（1. 汕头大学工学院，广东汕头 515063；2. 中国农业科学院农业资源与农业区划研究所 / 农业农村部农业遥感重点实验室，北京 100081）

摘要：【目的】探讨高光谱遥感数据不同预处理及不同估测算法下土壤有机质估测模型的优劣，为提高土壤有机质估测精度奠定基础。**【方法】**使用高光谱仪在室内条件下对土壤样品进行光谱测量，对光谱数据进行 4 种去噪处理（无去噪处理、Savitzky-Golay（S-G）平滑滤波去噪、小波包去噪以及 S-G 平滑与小波包结合去噪），然后对去噪后的光谱数据进行 8 种数据变换（原始光谱数据 R、倒数 1/R、对数 log（R）、倒数对数 log（1/R）、一阶导数 R′、倒数一阶导数（1/R）′、对数一阶导数（log（R））′、倒数对数一阶导数（log（1/R））′），接着对变化后的光谱数据进行 3 种降维处理（无降维处理、敏感波段降维和主成分分析降维），最后运用支持向量回归法和偏最小二乘回归法分别建立 SOM 含量估测模型。**【结果】**研究中所涉及的各种数据预处理和估测算法中，小波包去噪、PCA 降维、反射率倒数一阶导数（1/R）′光谱数据变换处理条件下，使用 PLSR 方法的估测模型精度最高、模型最稳定，可以较精确地估测吉林省伊通县 SOM 含量。**【结论】**合适的数据预处理，尤其是小波包去噪和 PCA 降维相结合，可有效改善光谱数据质量，提高 SOM 含量估测模型精度及稳定性。

关键词：土壤有机质；支持向量回归；偏最小二乘回归；小波包去噪；PCA 降维；高光谱
DOI：10.12105/j.issn.1672-0423.20190106

0　引言

土壤有机质（Soil Organic Matter，SOM）是土壤肥力的重要指标，不仅能为作物提供养分，改善土壤物理性质，还具有保水和保肥的作用[1]。因此，SOM 的快速、准确估测对粮食产量提高、农业可持续发展具有重要意义。传统的 SOM 估测方法一般成本高，比较耗时、费力，并且估测结果还具有一定滞后性，很难满足当前生产管理的需要。高光谱分析技术的发展，给土壤研究带来了许多新的方法。很多国内外研究者利用土壤光谱信息进行土壤属性的反演研究，越来越多的建模方法被用于 SOM 高光谱建模中，且模型精度较高[3-12]。土壤光谱信息不仅与 SOM 含量、氧化铁含量等土壤化学组分以及土壤含水量有关，而且与土壤的颗粒大小、形状、密度等物理性质有关[2]。司海清等[3-4]通过

收稿日期：2019-01-10
第一作者简介：沈兰芝（1994—），硕士研究生。研究方向：机器学习。Email：1663230519@qq.com
※ **通信作者简介：**高懋芳（1980—），博士、副研究员。研究方向：农业定量遥感。Email：gaomaofang@caas.cn
* **基金项目：**国家自然科学基金项目“耦合遥感与作物生长模型的农业干旱预警研究”（41871282）；中国地质调查工作项目（DD20160068）；高分辨率对地观测系统国家科技重大专项（09-Y30B03-9001-13/15）

对不同颗粒大小土样及不同含水率土样进行光谱数据测量，对光谱数据进行平滑滤波去噪，并对平滑后的数据进行 3 种光谱数据变换：反射率 R、反射率一阶导数 R′ 和反射率倒数对数 log（1/R），然后运用偏最小二乘回归（Partial Least Square Regression，PLSR）等方法建立 SOM 含量估测模型，表明土壤颗粒大小对土壤反射率有着十分明显的影响，且不同建模方法对模型的结果有明显影响。孙小香等[5]采用全波段原始光谱进行对数、倒数对数、一阶导数、二阶导数变换数据结合 3 种建模方法：PLSR、BP 神经网络和支持向量机（Support Vector Machine，SVM）构建不同的山地红壤全氮含量高光谱估测模型，表明全波段建立的土壤高光谱全氮含量估测模型中，精度由高到低依次为 SVM>BP 神经网络 >PLSR。卢艳丽等[6]采用样本光谱数据敏感波段建立线性回归来估测 SOM 含量。王永敏等[7]采用小波分析法去除土壤样本光谱数据部分噪音并通过光谱一阶导数变换结合回归分析法建立 SOM 估测模型，研究表明与未采用小波分析法建立的模型相比，模型的判定系数提高了 0.207。张锐、李兆富等[8]通过小波包和局部最相关算法建立 SOM 估测模型，决定系数可达 0.781。徐夕博等[9]通过 PCA（Principal Component Analysis）将实测高光谱数据降维为 6 个主成分，结合多元逐步线性回归（MLR）和 BP 神经网络，对建立的 SOM 估测模型进行了进一步分析。乔娟峰等[10]运用土壤原始反射率 R、倒数对数 log（1/R）、去包络线（CR）等 5 种光谱变换数据基于全波段和显著性波段利用 PLSR 建立 SOM 含量估测模型，研究表明选择显著性波段 CR 模型作为所研究区域的 SOM 含量估测模型更简洁、科学。

由于测量仪器、测量方法及测量环境等影响，土壤的光谱反射率数据必然存在噪声。并且，高光谱数据具有波段多、数据量大，数据冗余的特点，增加了数据处理与建模的工作量和复杂度。因此，在建模前选择合适的操作对数据进行预处理，如去噪、降维、数据形式变换等，对模型精度的提高至关重要。前述研究在建模前或多或少都进行了一些数据预处理。但在实验室进行土壤样品光谱测量反演 SOM 时，不同研究者对土壤样本处理方式不尽相同，且在测得土壤样本的光谱数据后，对光谱数据的预处理方式也不尽相同，使研究结果缺少可比性。除此之外，不同算法的模型对模型的估测结果影响也很大。

文章尝试对同一批土壤样本不同光谱数据预处理下的模型进行模型估测效果比对，所包含的数据预处理有数据降维、数据去噪和光谱数据形式变换，并且选用 PLSR 和支持向量回归（Support Vector Regression，SVR）作为模型研究对象，来比对不同数据预处理对这两种模型的影响，以期为高光谱 SOM 估测的相关研究提供技术支撑。

1 研究区与数据

1.1 研究区域与土壤采样

该文研究区域为吉林省伊通县，位于吉林省中南部，东经 124°49′~125°46′、北纬 43°3′~43°38′。土样采样时间为 2017 年 4 月 21—23 日。土壤采样点按照 1 km × 1 km 网格点布设，采样深度 0~5 km，涉及的土地利用类型为玉米耕地。调查区属于黑土区，土

壤类型包括草甸土、黑土、白浆土、水稻土 4 种。使用采集器在样点处垂直采集 1 个直径 10 cm、深 5 cm 的原状土，放入大铝盒中，尽量保持原状土样，用于室内土样光谱测量。每个样点用采集器各采集 3 个土样，3 个土样位置成三角形，相距 10 m 左右，深度为 0~5 cm，各放入直径 5.5 cm、高 3.5 cm 的小铝盒中，用于测定样点的 SOM 含量及其他土壤参数。

1.2 光谱数据测量

野外土样采集当天，采用 ASD FieldSpec 4 High-RS 高光谱仪，对放置在直径 10 cm、高 5 cm 铝盒中的原状土样进行室内光谱测量（图 1）。ASD 的波长范围为 350~2 500 nm，光谱分辨率为 3 nm@700 nm、8 nm@1 400/2 100 nm。测量时 50 W 的卤素灯放置于土样一侧，光源入射角度为 60（天顶角 30），距离土样为 30 cm。探头垂直距离土样 15 cm。每个土样测量 4 次光谱（每次测量前将大铝盒转动 90 度，共转动 3 次），每次测量自动采集 10 条光谱曲线，算术平均后作为该次的光谱曲线。每次测量前进行标准白板校正。

图 1 土样室内光谱测量工作图

Fig.1 Chart of indoor Soil Sample Spectral measurement

采用 ViewSpecPro 软件对室内原始光谱数据进行断点修正（GAP 窗口取值 5×5）和光谱平均，设置光谱分辨率为 1 nm，共 2151 个波段。去除 350~400 nm 因设备不稳定引起的噪声。再将室内光谱数据重采样成与 Hymap 机载高光谱影像（2017 年 4 月 30 日至 5 月 1 日获取）光谱分辨率相同（400~905 nm 光谱分辨率为 15 nm，880~2 500 nm 光谱分辨率为 18 nm），共获得 135 个波段，213 条室内光谱曲线。

1.3 建模样本确定

研究表明，SOM 含量增加会使土壤光谱反射率降低[13-14]。考虑到样本质量，该文根据上述研究结论删除数据较异常的 15 个样本，剩余的 198 个样本按 4∶1 的比例用于模型的建立与验证。将 198 个样本按 SOM 含量从小到大排列，从第 5 个开始，每隔 4 个样本挑选 1 个，总计 40 个作为验证样本，其他 158 个作为建模训练样本。各样本集 SOM 含量统计信息见表 1。

表 1 样本 SOM 含量统计

Table 1 Sample SOM content statistics

样本集	样本数量（个）	SOM（%）			
		最大值	最小值	平均值	标准差
总样本	198	4.25	1.15	1.81	0.50
建模样本	158	4.25	1.46	2.21	0.50
验证样本	40	3.59	1.15	2.17	0.50

2 实验与方法

2.1 实验设计

该研究共设有 192 组实验。原始土壤光谱数据经不同预处理（即经不同去噪处理，不同形式数据变化和不同降维处理）得到 SOM 含量估测模型的建模数据，之后经不同建模方法进行建模。其中，去噪处理有 4 种：无去噪处理、S–G 平滑滤波去噪、小波包去噪以及 S–G 平滑与小波包结合去噪。数据变化有 8 种：原始光谱数据 R、倒数 1/R、对数 log（R）、倒数对数 log（1/R）、一阶导数 R′、倒数一阶导数（1/R）′、对数一阶导数（log（R））′、倒数对数一阶导数（log（1/R））′。降维处理有 3 种：无降维处理、敏感波段降维和 PCA 降维。建模方法有 2 种：SVR 和 PLSR。不同预处理及建模方法的组合方式共 4×8×3×2=192 种，每种组合设为一组实验。最终从 SVR 和 PLSR 两种模型中分别选出几种估测结果较具代表性的组合操作，分析其预处理组合对估测模型的影响。整个实验中涉及的所有算法都通过 Python 编程语言在 Python3.7 软件上编程实现。

2.2 光谱数据预处理

高光谱仪器采集数据时受环境的影响，采集的数据一般会包含噪声。此外，高光谱仪器采集光谱数据的波长范围大、波段数据多，如果将采集到的所有波段作为模型的输入，数据量大，计算速度慢，因此有必要采取合适的操作对模型输入数据进行去噪及降维处理等预处理。

2.2.1 S–G 平滑滤波去噪

平滑滤波是光谱分析中常用的数据预处理方法之一。S–G 滤波是一种在时域内基于局域多项式最小二乘拟合的滤波方法，也是一种移动窗口的加权平均算法，但是其加权系数不是简单的常数窗口，而是通过在滑动窗口内对给定高阶多项式的最小二乘拟合得出[15]。其设计思想是通过反复迭代处理，使得重建后的曲线逐渐逼近原始曲线的上包络线[16]。用 S–G 方法进行平滑滤波去噪，可以提高光谱的平滑性，并较低噪声的干扰。S–G 滤波表达式可表示为：

$$Y_j^*=\frac{\sum_{i=-m}^{m} C_i Y_j}{N} \quad (1)$$

式（1）中，Y_j^* 为重建后的光谱数据，Y_j 为原始光谱数据，C_i 为滤波系数，N 为滑动窗口内的数据个数（N=2m+1），其中 2m+1 为窗口宽度。在实际应用中，S–G 滤波需要设置两个参数，即滤波窗口宽度和平滑拟合多项式阶次。滤波窗口宽度能够影响平滑结果，其窗口宽度越大，结果越平滑。平滑拟合多项式阶次影响滤波细节，阶次越高，细节纹理越清晰[17]。该研究最终确定滤波窗口大小为 101，拟合多项式阶次为 5 作为最终参数。

2.2.2 小波包去噪

Daubechies 等研究表明小波包可以同时顾及信号的高频和低频成分，并能实现各个频段有用信息的有效提取，去噪效果好[18–19]。采用小波包去噪时，小波基函数和信号的

分层数选择都尤为重要。小波包去噪将原始信号分解为高频信号和低频信号，高频信号反应噪声细节部分，低频信号反应原始信号的近似。该文选用 db2 小波基函数进行两层小波包分解，通过软阈值函数对信号分解后叶子层的高频信号节点 d 进行阈值去噪，然后对阈值去噪后的信号进行信号重构。阈值确定公式为：

$$tar=\sigma\sqrt{2\log_e N} \quad (2)$$

式（2）中，σ 为高频信号 d 中所有系数绝对值的中位数除 0.6745，N 为 d 中数据个数。该阈值由 Donoho 提出，是噪声系数的最大值。该文选取 tar/2 做为噪声信号的去噪阈值。

2.2.3 S–G 平滑与小波包结合去噪

研究中所用的 S–G 平滑与小波包结合去噪是将光谱数据经 S–G 平滑滤波后再进行小波包去噪。所用到的对应参数设置同前述设置，即 S–G 平滑滤波的滤波窗口大小设置为 101，拟合多项式阶次设置为 5。小波包去噪仍然选用 db2 小波基函数进行两层小波包分解，去噪阈值为 tar/2。

2.2.4 光谱数据变换

共 8 种光谱变换数据，R、1/R、log（R）、log（1/R）、R′、（1/R）′、（log（R））′ 和（log（1/R））′。由于光谱仪采集的是离散型数据，故用如下公式近似计算一阶导数光谱数据：

$$R'(\lambda_i)=\frac{R(\lambda_{i+1})-R(\lambda_i)}{\lambda_{i+1}-\lambda_i} \quad (3)$$

式（3）中，$R'(\lambda_i)$ 为波长 λ_i 处的反射率一阶导数值，$R(\lambda_{i+1})$ 为波长 λ_{i+1} 处的反射率，$R(\lambda_i)$ 为波长 λ_i 处的反射率。

2.2.5 敏感波段降维

将经去噪处理且经形式变换的光谱数据与对应 SOM 含量做相关性分析，选出决定系数 R^2（即相关系数 R 的平方值）大于等于 0.25 的波段作为每条光谱曲线的敏感波段。相关系数 r 的计算公式：

$$r_i=\frac{\mathrm{cov}(x,y)}{\sqrt{D(x)}\sqrt{D(y)}}=\frac{\sum_{i=1}^{N}(x_{ni}-\bar{x}_i)(y_n-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_{ni}-\bar{x}_i)^2\sum_{i=1}^{N}(y_n-\bar{y})^2}} \quad (4)$$

式（4）中，r_i 为第 i 个波段的光谱数据与土壤 SOM 的相关系数，x_{ni} 为第 n 个样本的第 i 个波段所对应的光谱数据值，$\bar{x}_i$为第 i 个波段所对应的光谱数据的平均值，y_n 为第 n 个样本的 SOM 含量，$\bar{y}$为所有样本 SOM 含量的平均值。

2.2.6 PCA 降维

PCA 是一种较常使用的降维方法，已广泛应用于高光谱遥感领域。PCA 变换的目的是通过线性变换，找到一组最优的单位正交向量基（即主成分），用线性组合来重构与原样本均方差的误差最小的一种变换方法[20]。在 PCA 中，数据从原来的坐标系转换到了新的坐标系，新坐标的选择是由数据本身决定的。第一个新坐标轴选择的是原始数据中方差最大的方向，第二个新坐标轴的选择与第一个坐标轴正交且具有最大方差的方向。以

此类推依次选择坐标轴来组成最优单位正交向量基。大部分方差都包含在最前面的几个新坐标轴中。该文实验中每条光谱曲线共有波段 135 个，选最终估测精度最高所对应的维数 25 作为最优 PCA 降维数。

2.3 建模与精度评定方法

2.3.1 SVR

支持向量机 SVM 是 20 世纪 90 年代中期发展起来的基于统计学习理论的一种机器学习方法，通过使用非线性映射算法，将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分；也通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。SVM 是一种监督学习算法，通常用于模式识别、分类以及回归分析。SVR 即将 SVM 用于回归分析。

2.3.2 PLSR

偏最小二乘回归法 PLSR 是一种新型的多元统计数据分析方法，它主要研究的是多因变量对多自变量的回归建模，特别当各变量内部高度线性相关时，用偏最小二乘回归法更有效。另外，偏最小二乘回归较好地解决了样本个数少于变量个数等问题。偏最小二乘法集主成分分析、典型相关分析和多元线性回归分析 3 种分析方法的优点于一身。它与主成分分析法都试图提取出反映数据变异的最大信息，但主成分分析法只考虑一个自变量矩阵，而偏最小二乘法还有一个“响应”矩阵，因此具有估测功能。

2.3.3 精度评定方法

该文中采用的模型精度评价参数包括训练集决定系数 R^2_t、验证集决定系数 R^2_v、训练集均方根误差（Root-Mean-Square Error of Training Set，$RMSE_T$）、验证集均方根误差（Root-Mean-Square Error of Verification Set，$RMSE_V$）和相对分析误差（Residual Prediction Deviation，RPD）。R^2 越大，模型的相关性越高。$RMSE_T$ 和 $RMSE_V$ 的值应尽量小，二者越接近，模型的估测精度越高、稳定性越高。RPD 一般分为 3 类：当 RPD ≥ 2.0 时，说明该模型适合于利用高光谱数据估测土壤有机质含量；1.4 < RPD < 2.0 时，认为可以通过别的建模方法来提高模型的可靠性；RPD ≤ 1.4 时，说明该模型不可靠[21]。

式（5）给出训练集均方根误差 $RMSE_T$ 的计算公式，验证集均方根误差 $RMSE_V$ 计算公式与 $RMSE_T$ 计算公式一致。

$$RMSE_T=\sqrt{\frac{\sum_{i=1}^{n}(y_{it}-y_{ip})^2}{n}} \quad (5)$$

式（5）中，y_{it} 为第 i 个样本的 SOM 含量真实值，y_{ip} 为第 i 个样本的 SOM 含量估测值，n 为训练集中样本个数。

式（6）给出相对分析误差 RPD 的计算公式：

$$RPD=SD/RMSE_V \quad (6)$$

式（6）中，SD 为验证集样本 SOM 含量标准差。

3 结果与分析

3.1 SVR 土壤有机质高光谱估测模型

基于 SVR 的 SOM 估测模型中，预处理去噪处理为小波包去噪、降维处理为 PCA 降维、光谱数据变换为 R′ 的模型建模效果相对最好。表 2 给出无预处理和预处理为小波包去噪、PCA 降维且光谱数据变换为 R′ 的基于 SVR 的 SOM 估测结果精度。图 2 给出基于 SVR 的无预处理 SOM 含量估测结果散点图。图 3 给出基于 SVR 的预处理为光谱数据 R′ 小波包去噪 PCA 降维的 SOM 含量估测结果散点图。图 3 所对应 $RMSE_V$、R^2_V 和 RPD 的值分别为 0.359、0.475 和 1.337，而不经任何预处理（即无去噪处理、无降维处理，无光谱数据变换处理）的 SVR 模型所对应的 $RMSE_V$、R^2_V 和 RPD 的值分别为 0.439、0.264 和 1.091。相比不经预处理的 SVR 估测模型而言，经预处理为小波包去噪、PCA 降维且数据变换为 R′ 的 SVR 估测模型的 R^2_V 提高了 0.211，RPD 提高了 0.246。

表 2 基于 SVR 的无预处理估测结果与预处理下结果最优的 SOM 含量估测结果

Table 2 SOM content estimation model results of without pretreatment and the optimal results in pretreatment based on SVR

预处理	$RMSE_T$	$RMSE_V$	R^2_T	R^2_V	RPD
无去噪 无降维 光谱数据 R	0.462	0.439	0.233	0.264	1.091
小波包去噪 PCA 降维 光谱数据 R′	0.372	0.359	0.464	0.475	1.337

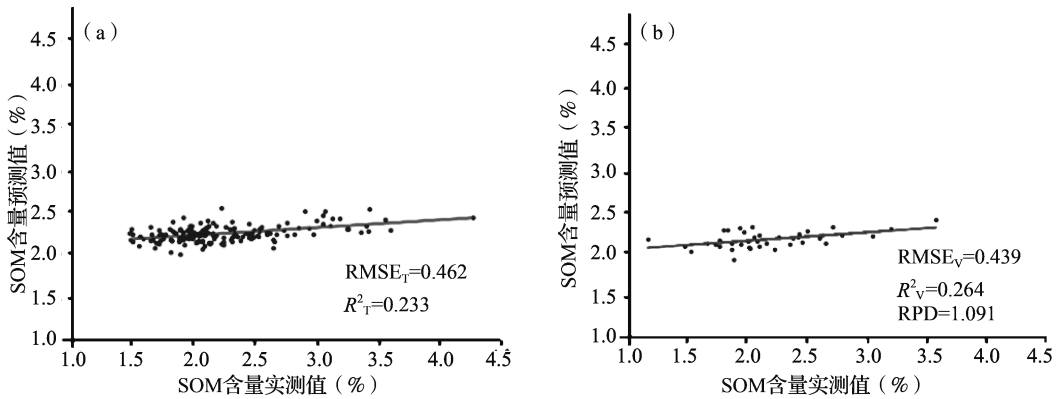

图 2 基于 SVR 的无预处理 SOM 含量估测结果

（a）为训练集估测结果散点图，（b）为验证集估测结果散点图

Fig.2 SOM content estimation results without pretreatment based on SVR

研究表明，同种去噪处理且同种光谱变换数据下，PCA 降维效果大多略优于无降维效果，而敏感波段降维效果基本都略差于无降维效果。由于该研究中各预处理下的 SVR

估测模型精度都不算太高且其他预处理下的 SVR 模型估测结果未显示出明显规律，此处不再列举。

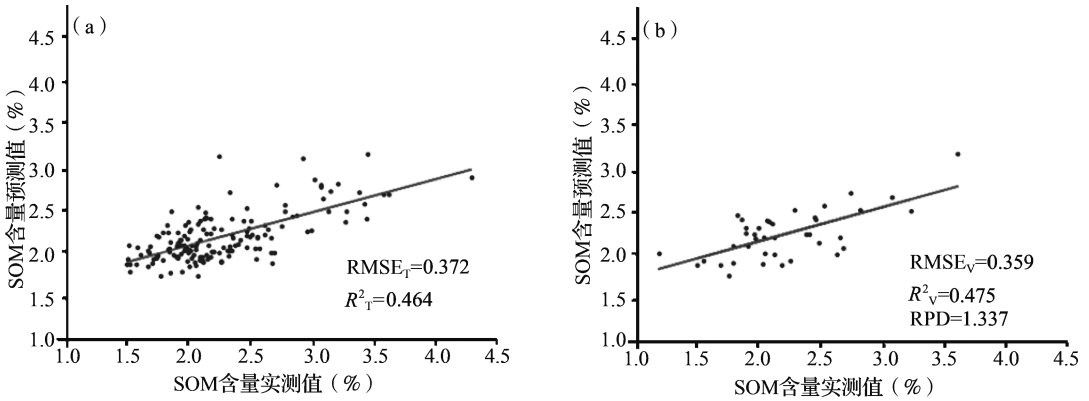

图 3　基于 SVR 的预处理为光谱数据 R′ 小波包去噪 PCA 降维的 SOM 含量估测结果

（a）为训练集估测结果散点图，（b）为验证集估测结果散点图

Fig.3　SOM content estimation results in spectral data R′ wavelet packet denoising and PCA dimensionality reduction based on SVR

3.2　PLSR 土壤有机质高光谱估测模型

基于 PLSR 的 SOM 估测模型中，预处理去噪处理为小波包去噪、降维处理为 PCA 降维、光谱数据变换为（1/R）′ 的模型建模效果相对最好。表 3 给出无预处理和经预处理小波包去噪、PCA 降维且光谱数据变换为（1/R）′ 的基于 PLSR 的 SOM 估测结果精度。图 4 给出无预处理下基于 PLSR 的 SOM 含量估测结果散点图。图 5 给出预处理下基于 PLSR 的光谱数据（1/R）′ 小波包去噪 PCA 降维 SOM 含量估测模型估测结果散点图。图 5 所对应 RMSE$_V$、R^2_V 和 RPD 的值分别为 0.280、0.713 和 0.712，而不经任何预处理的 PLSR 模型所对应的 RMSE$_V$、R^2_V 和 RPD 的值分别为 1.200、0.007 和 0.400。相比不经数据预处理的 PLSR 估测模型而言，经小波包去噪、PCA 降维且数据变换为（1/R）′ 的 PLSR 估测模型的 R^2_V 提高了 0.706，RPD 提高了 0.312。不经数据预处理的 PLSR 模型，训练集和验证集估测结果相差较大。

表 3　基于 PLSR 的无预处理与预处理下结果最优 SOM 含量估测模型结果

Table 3　SOM content estimation model results of without pretreatment and the optimal results in pretreatment based on PLSR

预处理	RMSE$_T$	RMSE$_V$	R^2_T	R^2_V	RPD
无去噪 无降维 光谱数据 R	0.150	1.200	0.921	0.007	0.400
小波包去噪 PCA 降维 光谱数据（1/R）′	0.241	0.280	0.775	0.713	1.712

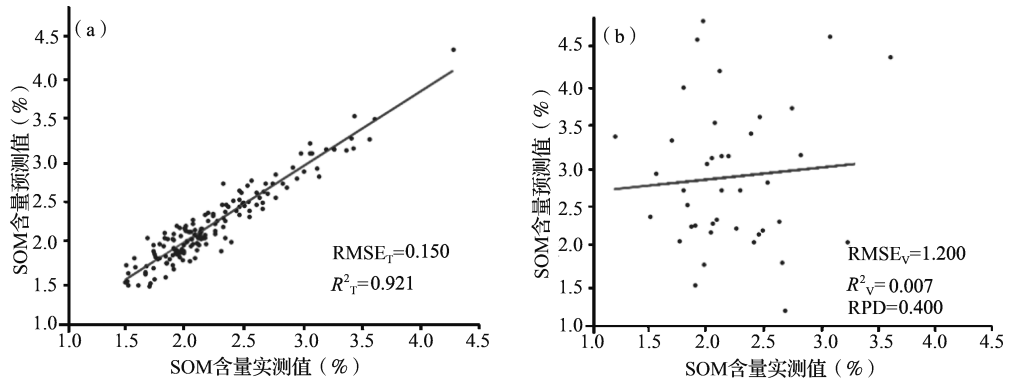

图 4　基于 PLSR 的无预处理 SOM 含量估测结果

（a）为训练集估测结果散点图，（b）为验证集估测结果散点图

Fig.4　SOM content estimation results without pretreatment based on PLSR

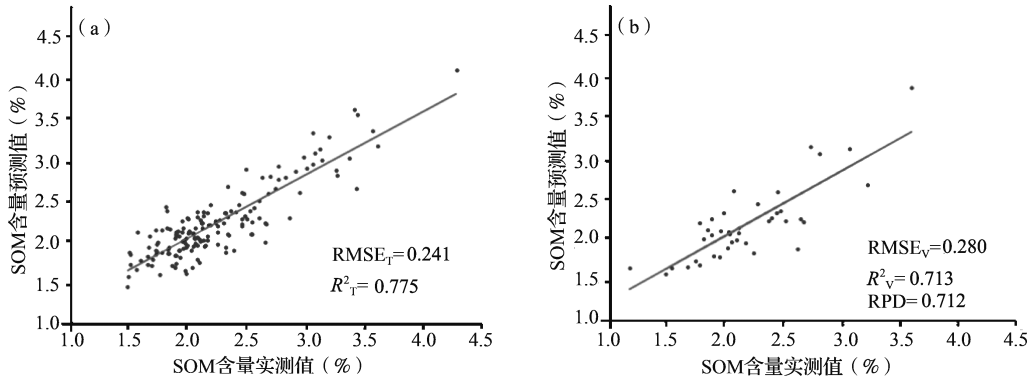

图 5　基于 PLSR 的预处理为光谱数据（1/R）′小波包去噪 PCA 降维的 SOM 含量估测结果

（a）为训练集估测结果散点图，（b）为验证集估测结果散点图

Fig.5　SOM content estimation results in spectral data（1/R）′ wavelet packet denoising and PCA dimensionality reduction based on PLSR.

研究表明，同种去噪处理且同种光谱变换数据下，敏感波段降维效果优于无降维效果，PCA 降维效果优于敏感波段降维效果。表 4 给出基于 PLSR 的小波包去噪处理下不同降维处理下光谱数据 1/R 的 SOM 含量估测结果比对。图 6 给出基于 PLSR 的预处理为光谱数据 1/R 小波包去噪无降维处理的 SOM 含量估测结果散点图。图 7 给出基于 PLSR 的预处理为光谱数据 1/R 小波包去噪敏感波段降维的 SOM 含量估测结果散点图。图 8 给出基于 PLSR 的预处理为光谱数据 1/R 小波包去噪 PCA 降维的 SOM 含量估测结果散点图。实验结果表明，在无降维操作时，模型训练结果出现严重的过拟合问题，训练集与验证集的均方根误差值 RMSE 相差 0.742；在敏感波段降维下，过拟合现象有所缓解，训练集与验证集的均方根误差值 RMSE 相差 0.113；在 PCA 降维下，过拟合现象基本消除，训练集与验证集的均方根误差值 RMSE 相差 0.051。说明预处理中的降维操作能有效改善模型过拟合现象，提高模型估测精度及稳定性。其他去噪处理或光谱变换数据下的结果类似，此处不再一一列出。

表 4　基于 PLSR 的预处理为小波包去噪不同降维处理下光谱数据 1/R 的 SOM 含量估测结果比对

Table 4　The comparison of SOM content estimation results in spectral data 1/R different dimensionality reduction and wavelet packet denoising based on PLSR

预处理	$RMSE_T$	$RMSE_V$	R^2_T	R^2_V	RPD
小波包去噪无降维光谱数据 1/R	0.114	0.856	0.960	0.129	0.560
小波包去噪敏感波段降维光谱数据 1/R	0.291	0.404	0.666	0.509	1.186
小波包去噪 PCA 降维光谱数据 1/R	0.236	0.287	0.785	0.690	1.668

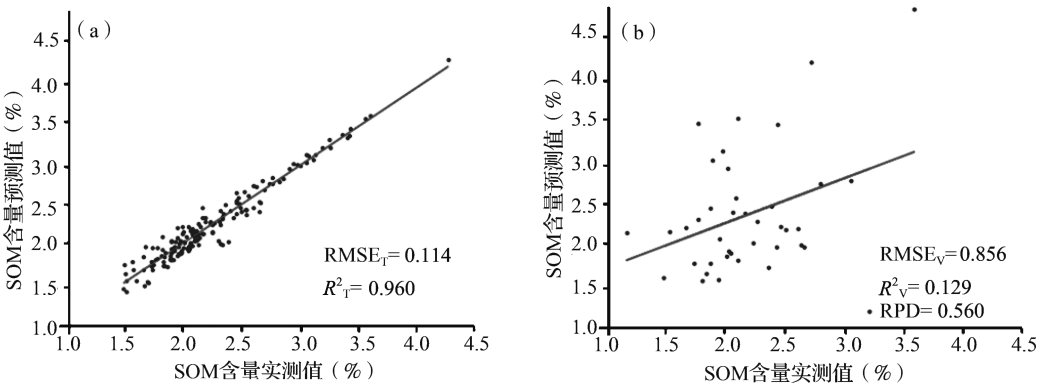

图 6　基于 PLSR 的预处理为光谱数据 1/R 小波包去噪无降维操作的 SOM 含量估测结果

（a）为训练集估测结果散点图，（b）为验证集估测结果散点图

Fig.6　SOM content estimation results in spectral data 1/R wavelet packet denoising and without dimensionality reduction based on PLSR

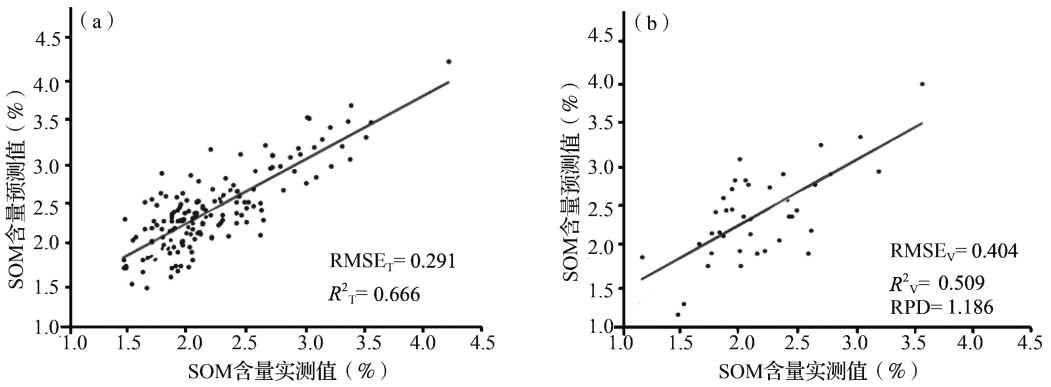

图 7　基于 PLSR 的预处理为光谱数据 1/R 小波包去噪敏感波段降维的 SOM 含量估测结果

（a）为训练集估测结果散点图，（b）为验证集估测结果散点图

Fig.7　SOM content estimation results in spectral data 1/R wavelet packet denoising and sensitive band reduction based on PLSR

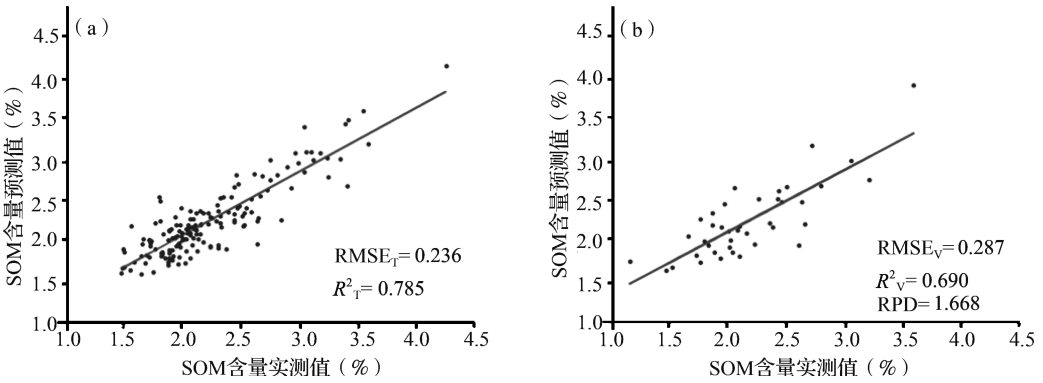

图 8　基于 PLSR 的预处理为光谱数据 1/R 小波包去噪 PCA 降维的 SOM 含量估测结果

（a）为训练集估测结果散点图，（b）为验证集估测结果散点图

Fig.8　SOM content estimation results in spectral data 1/R wavelet packet denoising and PCA dimensionality reduction based on PLSR

此外，部分光谱变换数据在同种降维操作下进行去噪处理后，模型估测精度也有所提高。比如，在不降维处理下，光谱数据 R 经 S-G 平滑滤波去噪后，模型 RPD 比无去噪处理提高了 0.186；光谱数据 R 经小波包去噪后，模型 RPD 比无去噪处理提高了 0.175；光谱数据 R 经 S-G 平滑滤波和小波包结合去噪后，模型 RPD 比无去噪处理提高了 0.114。而部分光谱变换数据在同种降维操作下进行去噪处理后，模型估测精度反而降低。比如，在敏感波段降维处理下，光谱数据 R 经 S-G 平滑滤波去噪后，模型 RPD 比无去噪处理降低了 0.011；光谱数据 R 经小波包去噪后，模型 RPD 与无去噪处理值一样；光谱数据 R 经 S-G 平滑滤波和小波包结合去噪后，模型 RPD 比无去噪处理降低了 0.110。

总体而言，各光谱变换数据在无降维操作和敏感波段降维操作下的估测精度都较低，RPD 值大都小于 1.4。相比而言，在 PCA 降维操作下的估测精度都有较大提高，RPD 值基本都大于 1.4，R^2_v 值可达 0.713，此时，PLSR 模型估测精度较高且模型最稳定。

3.3　基于 SVR 与基于 PLSR 的 SOM 含量高光谱估测模型结果对比确定性分析

表 5 给出基于 SVR 与基于 PLSR 的无预处理和预处理下结果最优的 SOM 含量高光谱估测模型结果对比。从表 5 可看出，在不经数据预处理时，基于 SVR 的 SOM 含量估测模型结果优于基于 PLSR 的 SOM 含量估测模型结果；而在经预处理后，基于 PLSR 的 SOM 含量估测模型结果反优于基于 SVR 的 SOM 估测模型结果。原因于 SVR 是机器学习中的典型分类回归算法，虽是一种非线性回归方法，在解决非线性问题上优于线性方法，但其对模型输入数据的要求也比较高，且样本数量也应尽可能地多。实验所用的高光谱数据获取时受较多外界因素影响，数据质量比较差，可能在训练模型时出现过拟合现象，所以导致模型即使在预处理后估测精度虽有提高但还是比较低。而 PLSR 是线性回归方法的代表之一，在样本数较少条件下优势发挥比较明显，再加上小波包去噪等提高了数据质量，所以在该实验中模型估测精度较 SVR 模型估测精度要高。

表 5　基于 SVR 与基于 PLSR 的 SOM 高光谱估测模型结果对比

Table 5　The comparison of SOM content estimation results based SVR and PLSR

SVR				PLSR			
预处理	$RMSE_V$	R^2_V	RPD	预处理	$RMSE_V$	R^2_V	RPD
无预处理	0.439	0.264	1.091	无预处理	1.200	0.007	0.400
小波包去噪 PCA 降维光谱数据 R′	0.359	0.475	1.337	小波包去噪 PCA 降维光谱数据（1/R）′	0.280	0.713	1.712

此外，在经预处理后，基于 SVR 的 SOM 含量估测模型结果和基于 PLSR 的 SOM 含量估测模型结果较不经预处理结果精度都有较明显的提升，说明合适的数据预处理可以显著提高高光谱 SOM 含量模型估测精度及稳定性。并且预处理中，都是在小波包去噪 PCA 降维下，基于 SVR 的光谱数据 R′ 的 SOM 含量估测精度最高，基于 PLSR 的光谱数据（1/R）′ 的 SOM 含量估测精度最高，说明小波包去噪和 PCA 降维结合可有效去除光谱数据部分噪声，提高土样光谱数据质量。

4　结论

以吉林省伊通县土样为研究对象，采集 213 份土壤样本，对土样的光谱数据和 SOM 含量进行测量和分析，采用 SVR 和 PLSR 方法建立不同数据预处理下的 SOM 含量的估测模型，并用验证样本对吉林伊通土样 SOM 含量高光谱估测模型进行验证，得到以下结论：（1）在该研究所涉及的数据预处理下，基于 SVR 的 SOM 含量估测模型中，预处理为小波包去噪、PCA 降维、R′ 光谱数据变换的建模效果最好。（2）基于 PLSR 的 SOM 含量估测模型中，预处理为小波包去噪、PCA 降维、（1/R）′ 光谱数据变换的建模效果最好。（3）基于 PLSR 的 SOM 含量估测模型下，同种去噪处理且同种光谱变换数据下，敏感波段降维估测效果优于无降维处理，PCA 降维估测效果优于敏感波段降维处理。（4）合适的数据预处理，尤其是小波包去噪和 PCA 降维相结合，可有效改善光谱数据质量、提高 SOM 含量估测模型精度及稳定性。（5）当光谱数据经过数据预处理小波包去噪、数据变换（1/R）′ 及 PCA 降维后，再使用 PLSR 方法可较高精度地估测吉林伊通地区 SOM 含量。

参考文献：

[1] 窦森 . 土壤有机质 . 北京 : 科学出版社 , 2010.
[2] 刘雪梅 . 近红外漫反射光谱检测土壤有机质和速效 N 的研究 . 中国农机化学报 , 2013, (2): 84~88.
[3] 司海青 , 姚艳敏 , 王德营 , 等 . 不同颗粒大小对高光谱估测土壤有机质含量的影响 . 中国农学通报 , 2015, 31(18): 173~178.
[4] 司海青 , 姚艳敏 , 王德营 , 等 . 含水率对土壤有机质含量高光谱估测的影响 . 农业工程学报 , 2015, 31(9): 114~120.
[5] 孙小香 , 赵小敏 , 谢文 . 基于高光谱的山地红壤全氮含量估测模型对比研究 . 江苏农业科学 , 2018, 46(15): 287~291.
[6] 卢艳丽 , 白由路 , 杨俐苹 , 等 . 基于高光谱的土壤有机质含量预测模型的建立与评价 . 中国农业科学 , 2007, 40(9): 1989~1995.

[7] 王永敏，田林亚，李西灿，等 . 基于小波与包络线的土壤有机质高光谱估测 . 地理信息世界，2018, 25(4): 36~41.
[8] 张锐，李兆富，潘剑君 . 小波包—局部最相关算法提高土壤有机碳含量高光谱预测精度 . 农业工程学报，2017, 33(1): 175~181.
[9] 徐夕博，吕建树，吴泉源，等 . 基于 PCA-MLR 和 PCA-BPN 的莱州湾南岸滨海平原土壤有机质高光谱预测研究 . 光谱学与光谱分析，2018, 38(8): 2556~2562.
[10] 乔娟峰，熊黑钢，王小平，等 . 新疆阜康荒地土壤有机质高光谱特征及其反演模型研究 . 干旱地区农业研究，2018, 36(5): 207~214.
[11] 徐彬彬，戴昌达 . 南疆土壤光谱反射特性与有机质含量的相关分析 . 科学通报，1980(6): 282~284.
[12] Vasques G M, Grunwald S, Sickman J O. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma*, 2008, 146(1/2): 14~25.
[13] 冯云山，吴培祥，刘亚娟，等 . 土壤光谱反射特性的研究 . 吉林农业大学学报，1989, 11(2): 72~76.
[14] 彭杰，张杨珠，周清 . 去除有机质对土壤光谱特性的影响 . 土壤，2006, 38(4): 453~458.
[15] 黄耀欢，王建华，江东，等 . 利用 S-G 滤波进行 MODIS-EVI 时间序列数据重构 . 武汉大学学报 (信息科学版), 2009, 34(12): 1440~1443, 1513.
[16] Savitzky A, Golay M J E. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chenmistry*, 1964, 36: 1627~1639.
[17] 杨恒，沈润平，吴立叶，等 . 基于 S-G 滤波的江西省植被覆盖度时空变化遥感分析 . 科学技术与工程，2014, 14(22): 101~106.
[18] 孔玲军 . Matlab 小波分析超级学习手册 . 北京 : 人民邮电出版社，2014.
[19] Virmani J, Kumar V, Kalar N, et al. SVM-Based characterization of liver ultrasound images using wavele packet texture descriptors. Journal of Digital Imaging, 2013: 1~14.
[20] 董陈武 . 基于 PCA、SVM 算法实现人脸识别 . 广播电视信息，2018(10): 107~110.
[21] Chang C W, Laird A D, Mausbach M J, et al. Near infrared reflectance spectroscopy: Principal components regression analysis of soil properties. *Soil Science Society of America Journal*. 2001, 65(2): 480~490.

Estimation model of soil organic matter based on SVR and PLSR

Shen Lanzhi[1], Gao Maofang[2※], Yan Jingwen[1], Yao Yanmin[2]
(1. College of Technology, Shantou University, Guangdong Shantou 515063, China; 2. Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences/Key Laboratory of Agricultural Remote Sensing, Ministry of Agriculture, Beijing 100081, China)

Abstract: [**Purpose**] The advantages and disadvantages of Soil Organic Matter (SOM) estimation models under different pretreatment and different estimation algorithms for hyperspectral remote sensing data are discussed, which lays a foundation for improving soil organic matter estimation accuracy. [**Method**] In our work, the spectra of soil samples were measured in laboratory using a high spectral spectrometer. Four kinds of denoising methods (Non-denoising, Savitzky-Golay (S-G) smoothing filtering, wavelet packet denoising and S-G smoothing combined with wavelet packet denoising) were used to process the spectral data. Eight kinds of spectral data transformations (R, 1/R, log (R), log (1/R), R′, (1/R)′, (log (R))′ and (log (1/R))′) are performed on the denoised spectral data. And three kinds of dimensionality reduction processing (Non-dimensionality reduction, sensitive

band dimensionality reduction and Principal Component Analysis (PCA) dimensionality reduction) are carried out on the changed spectral data. Finally, the SOM content estimation model was established by Support Vector Regression (SVR) and Partial Least Square Regression (PLSR). [**Result**] Among the various data preprocessing and estimation algorithms involved in this work, wavelet packet denoising, PCA dimensionality reduction, and reflectance first derivative (1/R) ′ spectral data transformation has the highest accuracy and stability under the model that established by PLSR, which can accurately estimate the SOM content of Yitong County, Jilin Province. [**Conclusion**] Appropriate data preprocessing, especially the combination of wavelet packet denoising and PCA dimensionality reduction, can effectively improve the quality of spectral data and improve the accuracy and stability of SOM content estimation model.

Key words: SOM; SVR; PLSR; wavelet packet denoising; PCA dimension reduction; hyperspectral