文章编号:1000-7601(2025)04-0203-08 **doi**:10.7606/j.issn.1000-7601.2025.04.21

基于 CR-CARS-RF 的黑土区土壤有机质含量高光谱估算

李雨鸿[1,2,3,4], 冯　锐[1,2], 纪瑞鹏[1,2], 张　霞[5],
武晋雯[1,2], 于文颖[1,2], 王　婷[3,4], 李　晶[3,4]

(1. 中国气象局沈阳大气环境研究所, 辽宁 沈阳 110166; 2. 辽宁省农业气象灾害重点实验室, 辽宁 沈阳 110166;
3. 辽宁省生态气象和卫星遥感中心, 辽宁 沈阳 110166; 4. 盘锦国家气候观象台,辽宁 沈阳 110166;
5. 中国科学院空天信息创新研究院,北京 100101)

摘　要:以辽宁昌图和苏家屯区 149 个黑土样本为基础,结合地面高光谱数据与实验室土壤有机质(SOM)含量化学测定,系统分析了不同高光谱预处理与建模方法对 SOM 反演精度的影响,旨在构建快速且精准的 SOM 含量高光谱反演模型。通过 Savitzky-Golay 平滑法(SG)对光谱数据进行去噪处理后,对比分析了倒数(1/R)、倒数的对数(log(1/R))、一阶导数(FDR)、标准化(SNV)及连续统去除(CR)五种光谱变换方法,并采用皮尔逊相关分析(PCC)和竞争自适应重加权(CARS)算法筛选特征波段,结合偏最小二乘(PLSR)、多元线性回归(MLR)和随机森林(RF)三种建模方法,构建了 12 种反演模型,对各模型的预测精度进行对比分析。结果表明:(1)CR 变换显著增强了光谱与 SOM 的相关性,2 166 nm 之后的 128 个波段相关系数绝对值超过 0.5,最高达 0.75;(2)CARS 算法有效压缩了特征波段数量,将其控制在全波段的 6%以下,与 PCC-PLSR 和 PPC-RF 相比,CARS-PLSR 和 CARS-RF 模型预测精度显著提升,其决定系数(R^2)分别提高 13.4%和 14.5%,均方根误差(*RMSE*)分别降低 12.8%和 11.9%;(3)非线性 RF 模型的预测精度最优,与 MLR 和 PLSR 相比,其 R^2 分别提升 32.1%和 3.5%,*RMSE* 分别降低 34.9%和 4.4%;(4)在 12 种预测模型中,CR-CARS-RF 模型表现最佳,其建模 R^2 为 0.91,*RMSE* 为 1.76 g·kg^{-1};预测 R^2 为 0.79,*RMSE* 为 2.49 g·kg^{-1},表明 CR-CARS-RF 模型具有较高的预测精度和可靠性,可为辽宁黑土区土壤有机质含量的高效精准监测提供有力的技术支撑。

关键词:黑土;土壤有机质;高光谱;竞争自适应重加权算法;随机森林模型

中图分类号:S153.6^{+}21; S127　　**文献标志码**:A

Hyper-spectral estimation of soil organic matter content in black soil based on CR-CARS-RF

LI Yuhong[1,2,3,4], FENG Rui[1,2], JI Ruipeng[1,2], ZHANG Xia[5],
WU Jinwen[1,2], YU Wenying[1,2], WANG Ting[3,4], LI Jing[3,4]

(1. *Institute of Atmospheric Environment*, *China Meteorological Administration*, *Liaoning*, *Shenyang* 110166, *China*;
2. *Key Laboratory of Agro-Meteorological Disasters*, *Liaoning*, *Shenyang* 110166, *China*;
3. *Liaoning Ecological Meteorology and Satellite Remote Sensing Center*, *Liaoning*, *Shenyang* 110166, *China*;
4. *National Climate Observatory in Panjin*, *Liaoning*, *Shenyang* 110166, *China*;
5. *Aerospace Information Research Institute*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*)

Abstract: This study, based on 149 black soil samples from Changtu and Sujiatun districts in Liaoning Province, combined ground hyperspectral data with laboratory chemical determination of soil organic matter (SOM) content, systematically analyzed the impacts of different hyperspectral preprocessing and modeling methods on SOM inversion accuracy, and aimed to construct a rapid and precise hyperspectral inversion model for SOM content.

收稿日期:2024-09-04　　**修回日期**:2024-10-15

基金项目:辽宁省农业气象灾害重点实验室开放基金(2023SYIAEKFZD06);中国气象局创新发展专项(CXFZ2023J059);辽宁省自然科学基金(2023-MS-042);中国气象局重点创新团队建设项目(CMA2024ZD02)

作者简介:李雨鸿(1987-),女,辽宁鞍山人,高级工程师,主要从事生态与农业气象研究。E-mail:liyh0609@163.com

通信作者:冯锐(1972-),女,辽宁凌海人,研究员级高级工程师,主要从事农业气象与卫星遥感研究。E-mail:fengrui_k@126.com

After denoising spectral data using the Savitzky-Golay smoothing method, five spectral transformation methods were compared: reciprocal (1/R), logarithmic reciprocal (log(1/R)), first derivative (FDR), standard normal variate (SNV), and continuum removal (CR). Pearson correlation coefficient (PCC) analysis and competitive adaptive reweighted sampling (CARS) algorithm were used for feature band selection. Three modeling methods—partial least squares regression (PLSR), multiple linear regression (MLR), and random forest (RF)—were combined to construct 12 inversion models, with their prediction accuracies compared and analyzed. The results showed that: (1) CR transformation significantly enhanced the correlation between spectra and SOM, with the absolute values of correlation coefficients exceeding 0.5 for 128 bands after 2 166 nm, reaching up to 0.75. (2) The CARS algorithm effectively reduced the number of feature bands to below 6% of the full bands, and compared with PCC-PLSR and PCC-RF, the prediction accuracies of CARS-PLSR and CARS-RF models were significantly improved, with their coefficients of determination (R^2) increasing by 13.4% and 14.5%, and root mean square errors (*RMSE*) decreasing by 12.8% and 11.9%, respectively. (3) The nonlinear RF model achieved the best prediction accuracy, with R^2 increasing by 32.1% and 3.5%, and *RMSE* decreasing by 34.9% and 4.4% compared to MLR and PLSR, respectively. (4) Among the 12 prediction models, the CR-CARS-RF model performed optimally, with a modeling R^2 of 0.91 and *RMSE* of 1.76 $g \cdot kg^{-1}$, and a prediction R^2 of 0.79 and *RMSE* of 2.49 $g \cdot kg^{-1}$. These results indicate that the model has high prediction accuracy and reliability, providing a strong technical support for efficient and precise monitoring of SOM in black soil region of Liaoning.

Keywords: black soil; soil organic matter; hyper-spectral; competitive adaptive re-weighting algorithm; random forest model

土壤有机质(soil organic matter, SOM)是土壤固体部分的重要组成成分,其含量是评价土壤肥力的关键指标,直接影响作物生长状况。东北黑土区是我国重要的玉米、粳稻等商品粮供应基地,但随着黑土地利用强度增加和垦殖率提升,黑土区有机质含量正在以平均每年0.1%的速率降低。因此,快速、有效地评价黑土区土壤有机质含量,对于土地资源利用和区域农业保护发展具有重要意义。传统的SOM含量测定方法主要通过野外采集后进行化学分析,需要大量的野外样本点,耗时耗力且成本高,较难实现大范围、周期性监测,无法满足现代农业发展需求。高光谱遥感作为一种通过窄而连续的光谱通道实现对地物快速无损动态监测的技术手段,因其操作简单、光谱信息丰富、高效且成本低等特点,被广泛应用于SOM的定量反演[1-3]。

已有研究表明,光谱数据采集、数据处理和建模方式都会影响SOM反演精度[4-5]。早期SOM的反演多采用相关分析筛选敏感波段建立线性回归模型,针对东北黑土和北疆绿洲农田灰漠土,建立多元线性回归(MLR)SOM反演模型,精度可达到0.8以上[6-7];针对江西红壤和辽河三角洲滨海湿地建立偏最小二乘(PLSR)SOM估算模型,取得了较高的预测精度,决定系数R^2分别达到0.96和0.87[8-9]。由于土壤反射光谱的复杂性,并非简单的线性机理,基于线性相关所建模型在特定区域内可能表现较高精度,但应用于不同环境条件和区域时,其适用性和准确性会受到限制。随着机器学习方法的发展,支持向量机回归(SVM)、神经网络(NNP)和随机森林(RF)等非线性模型被广泛应用到SOM的估算研究中。沈强等[10]、郄欣等[11]发现,采用SVM建立的区域SOM含量高光谱预测模型反演精度高于线性回归和PLSR。王永敏等[12]、刘杰亚等[13]发现,经过小波分析或主成分分析实现光谱数据降维后,采用BP神经网络和卷积神经网络建模,其反演SOM精度优于传统回归模型。随机森林作为一种处理高维数据的机器学习方法,是由决策树发展而来的一种自然的非线性建模工具,其计算效率高,性能稳定,克服了其他机器学习模型中过渡拟合和不稳定性问题,可以更准确地解释变量之间的非线性关系。杨煜岑等[14]、韩杏杏等[15]、刘恬琳等[16]在反演陕西、河南、山东省局部地区的SOM含量时发现,随机森林模型可以精确地反映SOM与光谱和自然环境等变量之间的复杂非线性关系,预测精度高于线性回归和SVM。以上研究针对不同区域建立了SOM的最佳估算模型,但如何剔除高光谱波段间的冗余信息,降低数据维度仍是高光谱反演的重点和难点。白婷等[17]、唐海涛等[18]引进竞争适应重加权(CARS)方法,将特征波段压缩至全波段数据的16%,通过优选特征波段简化了RF模型复杂程度,可以实现SOM的快速、高效预测。

东北黑土地主要分布在黑龙江省和吉林省境内,辽宁省覆盖面积较小,目前关于东北黑土土壤有机质含量的反演研究多集中在黑龙江省[19]和吉林省[20],对面积较小且分布较为分散的辽宁黑土土壤有机质含量反演研究较少。基于此,本工作在结合不同光谱数据变换的基础上,利用 CARS 结合 RF 模型,并与传统的特征波段筛选方式及建模方法结果对比,分析 CARS-RF 在辽宁黑土区 SOM 反演的适用程度,筛选辽宁省黑土地 SOM 含量的快速反演方法,以期为辽宁农业生产和高标准农田建设提供基础数据支撑。

1　研究区与数据

1.1　研究区概况

选取辽宁省沈阳市苏家屯区和铁岭市昌图县为研究区。这两个区域地处辽北低丘平原,地势相对平坦,为温带大陆性季风气候,年平均气温为 8℃左右,其中 7 月平均最高气温达 28~30℃,1 月平均最低气温为-16~-14℃,四季分明;年平均降水量为 633.5 mm 左右,雨热同季;年平均日照时数约为 2 651 h,日照丰富。其中,苏家屯全区耕地土壤质量较好,以棕壤、草甸土和水稻土为主,分布着少量黑土,主要农作物为水稻、玉米和蔬菜。昌图县地处辽河上游,黑土面积广阔,处于北纬 42°的“黄金玉米带”,素有“辽宁粮仓”之称。

1.2　研究数据

数据采集时间为 2023 年 3 月 26—30 日,地面高光谱数据测定采用 ASD FieldSpec Pro 光谱仪,波段范围为 350~2 500 nm,光谱采样间隔为 1 nm,野外采样时每个观测点重复采集 10 次,利用 ViewSpecPro 处理后取平均值。光谱测定时天气状况良好,在 10∶00—14∶00 之间共采样 149 个,采样点分布在研究区黑土集中区域,且无农作物生长,土壤处于裸露状态,可有效避免植被光谱对土壤光谱信息的干扰。土壤有机质采样点与光谱数据采集为同一地点,土壤样本风干后,基于外加热条件,用确定标准浓度的重铬酸钾-硫酸溶液进行氧化和滴定,从而测定土壤有机质含量。土样有机质含量见表 1,数值范围为 13.30~42.56 g·kg^{-1},其中 20 g·kg^{-1}以下土样占比为 67.8%(101 个),20~30 g·kg^{-1}土样占比为 24.8%(37 个),30 g·kg^{-1}以上土样占比 7.3%(11 个)。将所有样本按照 SOM 含量高低排序,分层抽取样本点作为预测集,将 149 个样本划分为建模集(119 个)和预测集(30 个)。从表 1 可以看出,研究区总样本变异系数为 29.6%,为中等变异强度,介于建模样本和预测样本之间,说明按照 SOM 含量高低分层抽取样本建立建模集和预测集可以反映研究区的 SOM 现状。

1.3　光谱数据预处理

由于仪器测量边缘部分受外界噪声影响较大,以及部分波段受水汽影响明显等原因,本研究剔除 1 811~1 910、2 401~2 500 nm 等波段数据,选择 350~1 810 nm 和 1 911~2 400 nm 波段作为本研究的光谱波长范围。为增强 SOM 含量对光谱信息的敏感程度,对光谱数据进行 Savizky-Golay(SG)平滑处理后,进一步开展光谱倒数(1/R)、倒数的对数(log(1/R))、一阶导数(FDR)、标准化(SNV)和连续统去除(CR)等变换。

2　研究方法

2.1　特征波段筛选

相关性分析(PCC)是对两个独立变量之间关联程度的检验,通过计算 SOM 含量与全光谱波段之间的相关系数,设定入选阈值,选取相关性较高的光谱波段作为输入变量,降低数据“维度”。

竞争自适应重加权采样法(CARS)是通过多次迭代统计和交叉验证筛选特征变量的方法。通过随机抽取固定比例样本建立 PLS 模型,保留 PLS 模型中回归系数绝对值权重较大的波长作为新的子集,应用指数衰减和自适应重加权采样筛选变量,交叉验证计算均方根误差 *RMSE*,经过多次循环迭代选择不同子集,选取 *RMSE* 最小子集和对应迭代次数所含变量作为最优波段组合。本次研究在 Matlab 2018a 环境下运行 CARS 算法,固定比例为 80%,迭代次数为 50 次。

表 1　样本土壤有机质含量统计

Table 1　Statistic of soil organic matter content in sample soils

数据集 Data set	样本数 Number of samples	平均值 Average value /(g·kg^{-1})	最大值 Max value /(g·kg^{-1})	最小值 Min value /(g·kg^{-1})	变异系数 Coefficient of variation/%
总样本 Total samples	149	19.78	42.56	13.30	29.6
建模集 Modeling set	119	19.84	42.56	13.38	30.1
预测集 Prediction set	30	19.51	38.71	13.30	28.3

2.2 建模方法

多元线性回归(MLR)模型构建:MLR 模型用于解释变量与其他多个变量之间的线性关系。通过筛选与 SOM 含量关联程度较强的光谱波段,实现从全变量到特优变量的“降维”。本研究借助 SPSS 软件运行 MLR 模型,将筛选出的有效特征波段作为自变量,实测 SOM 含量作为因变量,从而构建 MLR 模型,分析特征波段对 SOM 含量的线性影响。

偏最小二乘法(PLSR)模型构建:PLSR 是基于最小化误差平方和,筛选出函数最优匹配的一种多元线性回归方法,可以集成主成分分析和多元线性回归及数据之间的典型相关关系。其优点是解决了自变量之间的多重线性相关问题,可以剔除无解释作用的信息,强化变量的解释能力。建模过程中,首先选取经过特定光谱变换后与 SOM 的相关系数大于 0.5 的波段,然后分别利用 PCC 和 CARS 筛选特征波段,将筛选后的波段作为输入变量,实测 SOM 含量作为输出变量,在 Matlab 2018a 环境下运行 PLSR 模型,构建出不同的 PLSR 模型,进而实现对 SOM 含量的精准估算。

随机森林(RF)模型构建:RF 是由多个不同的决策树组成,每个决策树均不相同,其核心思想是对训练数据进行自主采样,形成多个训练集,随机选取一部分样本进行训练,每一次训练生成一棵决策树,所有决策树形成随机森林,进而对样本进行随机训练并预测的机器学习方法。在本研究中,先运用 PCC 和 CARS 方法筛选出特征光谱波段,再将这些波段作为输入,实测 SOM 含量作为输出,在 Matlab 2018a 环境下运行 RF 模型,构建出基于 RF 的 SOM 含量估算模型,利用其强大的非线性建模能力,挖掘光谱数据与 SOM 含量之间复杂的非线性关系,提高模型预测的准确性。

2.3 模型精度评价

决定系数(R^2)是对模型评估的重要评价指标,拟合优度越大,自变量对因变量的解释程度越高,模型精度越高,其计算公式为:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\tilde{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{1}$$

均方根误差($RMSE$)表示实测值与模型预测值之间的偏差程度,其值越小,表明模型预测能力越强,是评估预测模型效果的常用指标,其计算公式为:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{n}} \tag{2}$$

式(1)和(2)中,y_i 表示 SOM 含量观测值;$\tilde{y}_i$ 表示 SOM 含量预测值;$\bar{y}_i$ 表示 SOM 含量观测值的平均;n 为对应样本数。

3 结果与分析

3.1 特征波段选取结果分析

3.1.1 相关性分析(PCC) 为确定不同光谱变换形式下的 SOM 光谱反射率敏感波段,对 149 个土样的 SOM 值与 S-G 平滑及其各种变换形式光谱反射率进行相关性分析(表 2)。从表 2 可以看出,S-G 平滑光谱数据与 SOM 含量最大相关系数为 0.40,光谱倒数、倒数的对数和标准化变换后的光谱反射率与 SOM 含量最大相关系数介于 0.42~0.48 之间,没有明显改善,为无效变换,后续不再展开分析;而一阶求导和连续统去除变换后,相关性明显提高,为有效变换。经连续统去除变换后的光谱与 SOM 相关效果最好,最大相关系数达到 0.75,相关系数绝对值高于 0.5 的波段达到 128 个,主要分布在2 166 nm 波段之后。

表 2 不同光谱数据变换形式相关性统计结果

Table 2 The correlation statistics of different transformed spectral data

变换方式 Transform	R 绝对值>0.5 波段数 The band number of $\|R\|>0.5$	R 绝对值>0.5 最大相关波段 The band of maximum correlation of $\|R\|>0.5$	最大相关系数 Maximum correlation coefficient
S-G	0	0	0.40
1/R	0	0	-0.43
log(1/R)	0	0	0.42
SNV	0	0	0.48
FDR	8	2315	0.60
CR	128	2207	0.75

3.1.2 竞争自适应重加权采样(CARS) 将 SOM 含量实测值和经 CR、FDR 变换的全光谱波段分别进行蒙特卡洛采样,每次随机选取 80%样本作为建模集,20%样本作为预测集输入 PLSR 模型,经过 50 次迭代运算,最终选取 $RMSE$ 最小值和对应的波段子集为特征波长。由图 1 可知,经 CR 和 FDR 变化后光谱数据在 CARS 运算过程中,$RMSE$ 随运行次数增加先减小再增大,其中,CR 变换光谱数据在迭代次数为 22 次时对应 $RMSE$ 最小,为 2.69 $g \cdot kg^{-1}$,对应特征波段为 102 个,所选特征波段数为全波段数的 5.2%;FDR 变换光谱数据在迭代次数为 25 次时对应 $RMSE$ 最小,为 2.54 $g \cdot kg^{-1}$,对应特征波段为 67 个,所选特征波段数为全波段数的 3.4%,说明基于 CARS 筛选特征波段达到了很好的降维效果。

3.2　土壤有机质估算模型

将通过 PCC 和 CARS 筛选的有效特征波段作为输入量,以实测的 SOM 含量作为输出量,根据建模集和预测集的划定,分别利用偏最小二乘回归(PLSR)和多元线性回归(MLR)线性模型和随机森林(RF)方法进行土壤有机质含量反演。

3.2.1　基于 MLR 的 SOM 反演模型　将通过 PCC 和 CARS 筛选的特征波段分别进行逐步线性回归,建立 MLR 模型。由表 3 可知,经 FDR 和 CR 变换后建立的 FDR-PCC-MLR 和 CR-PCC-MLR 模型建模集和预测集 R^2 均达 0.6 以上,预测集均方根误差 *RMSE* 分别为 3.24 g · kg^{-1} 和 3.03 g · kg^{-1},说明经过 FDR 和 CR 变换后,基于 PCC 筛选特征波段建立 SOM 含量多元线性回归模型预测效果较好,其中 CR-PCC-MLR 模型的预测效果更好。通过 CARS 筛选特征波段建立 MLR 模型预测集 R^2 变小,*RMSE* 明显变大,说明经 CARS 选取特征波段后建立 MLR 回归模型效果较差。

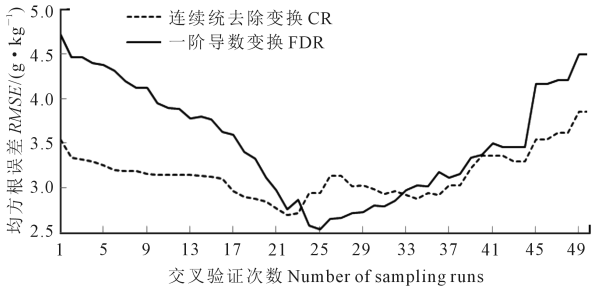

图 1　交叉验证的均方根误差
Fig.1　*RMSE* of cross-validation

3.2.2　基于 PLSR 的 SOM 反演模型　利用 PLSR 能有效提取具有最强解释能力的综合变量,去除无用信息。首先选取经过 FDR 和 CR 变换后的与 SOM 相关系数大于 0.5 的波段,与 SOM 进行 PLSR 建模分析,通过 PCC 筛选特征波段,建立的 FDR-PCC-PLSR 和 CR-PCC-PLSR 模型预测集 R^2 分别为 0.62 和 0.72,预测效果与多元线性回归建模效果相差不大(表 3)。

将通过 CARS 筛选的特征波段与 SOM 含量进行 PLSR 模型预测发现,建立的 FDR-CARS-PLSR 和 CR-CARS-PLSR 模型建模集和预测集 R^2 均在 0.7以上,与 MLR 模型相比,两种变换方式的 *RMSE* 均明显降低,且经 FDR 变换后通过 CARS 选取特征波段后建模效果最好。说明经过 CR 和 FDR 变换,通过 CARS 选取特征波段后建立 PLSR 模型,对土壤有机质含量的预测能力增强。

3.2.3　基于 RF 的反演模型　将经过 PCC 和 CARS 选取的特征光谱波段分别利用随机森林模型进行 SOM 含量预测发现,基于 RF 建立的 SOM 建模集 R^2 均达 0.8 以上,预测集 R^2 稳定在 0.69 以上,建模效果优于 MLR 和 PLSR 模型,预测稳定度和精确度均有提升;其中经 CR 变换后,利用 CARS 算法选取特征波段,建立 RF 预测模型,建模精度和预测精度 R^2 达到最高,分别为 0.91 和 0.79,*RMSE* 最小,分别为 1.76 g · kg^{-1} 和 2.49 g · kg^{-1},说明利用 CR-CARS-RF 建模对 SOM 的预测效果最好,为最优模型(表 3)。

3.3　不同组合模型精度对比

本研究选取 2 种有效的光谱变换方式、2 种敏感波段选取方法和 3 种建模方式,组合后共 12 种建模方法,各方法的观测值与预测值散点图如图 2 所示。从图 2 可以看出,采用 CR-CARS-RF 建模反演 SOM 含量的 R^2 最高、*RMSE* 最小,且散点图中实测值与预测值拟合效果最优。通过表 3 发现,多元线性回归模型(MLR)预测 SOM 含量的 4 种组合方式平均 R^2 为 0.56,*RMSE* 为 4.41 g · kg^{-1},其中,基于 PCC 筛选特征波段后 2 种建模预测结果平均 R^2 为 0.67,*RMSE* 为 3.13 g · kg^{-1},建模精度明显优于 CARS-MLR,其 R^2 增加48.9%,*RMSE* 减小 44.9%,这是因为多元线性回归模型的基础是通过筛选关联性较强的因子实现数据降维,所以采用 PCC 筛选具有高相关性的特征波段与 SOM 含量关联程度明显高于 CARS 算法。

偏最小二乘法(PLSR)预测 SOM 含量的 4 种组合方式平均 R^2 为 0.71,*RMSE* 为 3.00 g · kg^{-1},其中,基于 CARS 筛选特征波段后建模预测结果平均 R^2 为 0.76,*RMSE* 为 2.80 g · kg^{-1},建模精度明显优于 PCC-PLSR,较其 R^2 增加 13.4%,*RMSE* 减小 12.8%,说明引进 CARS 结合 PLSR 预测精度优于 PCC-PLSR,这是因为 CARS 在运算过程中结合 PLS 进行交叉验证,通过多次迭代优选最小 *RMSE* 对应的波段作为 PLSR 模型的输入变量,对优化 PLSR 建模精度更为有效。

随机森林(RF)预测 SOM 含量的 4 种组合方式平均 R^2 为 0.74,*RMSE* 为 2.87 g · kg^{-1},R^2 较 MLR 和 RLSR 方法分别平均增大 32.1%和 3.5%,*RMSE* 分别减小34.9%和 4.4%。其中,基于 CARS 筛选特征波段后建模预测精度平均 R^2 为 0.79,*RMSE* 为 2.69 g · kg^{-1};与 PCC-RF 相比,CARS-RF 的 R^2 增加 14.5%,*RMSE* 减小11.9%。由此可见,基于随机森林方法反演辽宁黑土区 SOM 含量效果优于多元线性回归和偏最小二乘回归,引进 CARS 筛选特征波段与 PLSR 和 RF 组合估算 SOM 含量后,模型预测精度明显提升,且基于 CARS 算法结合 RF 估算精度最高,预测效果最好。

表 3 不同组合模型 SOM 预测精度对比

Table 3 Comparison of the prediction of different combinations of models for SOM

模型 Model	波段筛选 Bands filtering	光谱变换 Spectral transformation	建模集 R^2 Modeling set R^2	建模集 *RMSE* Modeling set *RMSE*/(g·kg^{-1})	预测集 R^2 Prediction set R^2	预测集 *RMSE* Prediction set *RMSE*/(g·kg^{-1})
MLR	PCC	FDR	0.65	3.64	0.65	3.24
		CR	0.74	3.16	0.69	3.03
	CARS	FDR	0.75	5.47	0.28	4.88
		CR	0.85	6.33	0.62	6.51
PLSR	PCC	FDR	0.64	3.86	0.62	3.45
		CR	0.63	3.61	0.72	2.97
	CARS	FDR	0.89	1.95	0.77	2.68
		CR	0.88	2.05	0.75	2.92
RF	PCC	FDR	0.83	2.45	0.69	3.06
		CR	0.88	2.02	0.69	3.05
	CARS	FDR	0.89	1.98	0.79	2.89
		CR	0.91	1.76	0.79	2.49

4 讨 论

高光谱技术庞大的数据量在丰富土壤光谱信息的同时,也带来了噪声干扰、波段共线性和信息冗余等问题,数据变换和特征波段筛选可以有效消除噪声,降低波段维度,突出土壤有效信息。本研究发现,光谱数据经过 SG 平滑后,采用 FDR 和 CR 变换后与 SOM 含量的相关系数明显提升,这与周伟等[21]和丁启东等[22]的研究结果一致。采用 PCC 相关性分析是一种最为广泛的特征波段筛选方法,可以明确土壤与光谱数据间的相关关系。本研究在采用 PCC 的基础上,利用 CARS 再次进行波段筛选,筛选结果分别与 MLR、PLSR 和 RF 结合后建模,结果表明,基于 CARS 筛选特征波段后与 PLSR 和 RF 结合,建模精度更为精确、合理,降维效果更为明显,且与 RF 结合后建模效果明显优于 MLR 和 PLSR,这也与唐海涛[18]的研究结果一致。

本研究通过不同的光谱数据变换、特征波段筛选构建了研究区共 12 种 SOM 含量估算模型,结果表明,CR-CARS-RF 是一种快速、有效的 SOM 含量估算方法,预测效果较好,但预测精度仍有提升空间。一方面可能是因为研究区 SOM 含量差异略大,没有将 SOM 含量按照高、低水平区分建模,可能会致使模型存在一定的偏差;另一方面,受模型输入变量的影响,因单一波段光谱反射率存在一定的片面性和地域性,后续仍需对模型输入变量进行调整和优化,尝试引进双光谱指数或三光谱指数,通过结合光谱指数的数据变换和最优光谱指数组合筛选,进一步提升模型的稳定性和普适性。同时,因野外实测光谱存在空间不连续性,无法实现区域有机质含量反演与空间制图。2019 年国产高分五号卫星投入使用,其高光谱传感器具有 330 个波段,具有高度详细的光谱信息,因此,后续将围绕星-地光谱结合,基于实测高光谱数据对高分五号卫星遥感影像进行地面校验,最终实现基于高分五号卫星的黑土区有机质含量反演与制图,以期增强反演模型的适用性和实用性。

5 结 论

1)高光谱数据经过平滑去噪,采用 FDR 和 CR 数学变换后与 SOM 含量之间的相关性明显提高,其中经 CR 变换后相关性增幅提升明显。

2)特征波段筛选方式影响 SOM 反演精度,基于 MLR 建模时采用 PCC 筛选特征波段后预测精度较高;引进 CARS 算法与 PLSR 和 RF 结合后,与 PCC-PLSR 和 PCC-RF 预测精度相比,其平均 R^2 分别增加 13.4%和 14.5%,平均 *RMSE* 分别减小 12.8%和 11.9%,模型预测精度明显提升,且基于 CARS 算法结合 RF 估算精度最高,预测效果最好。

3)基于随机森林方法反演辽宁省黑土区 SOM 含量效果优于多元线性回归和偏最小二乘回归,其预测精度 R^2 较 MLR 和 PLSR 分别增加 32.1%和 3.5%,*RMSE* 分别减小 34.9%和 4.4%。其中,高光谱数据经连续统去除变换后,采用 CARS 算法结合 RF 建模精度和预测精度达到最高,R^2 分别为 0.91 和 0.79,*RMSE* 最小,分别为 1.76 g·kg^{-1} 和 2.49 g·kg^{-1},说明 CR-CARS-RF 是一种高效、快速估算辽宁黑土区土壤有机质含量的建模方法,预测效果最好。

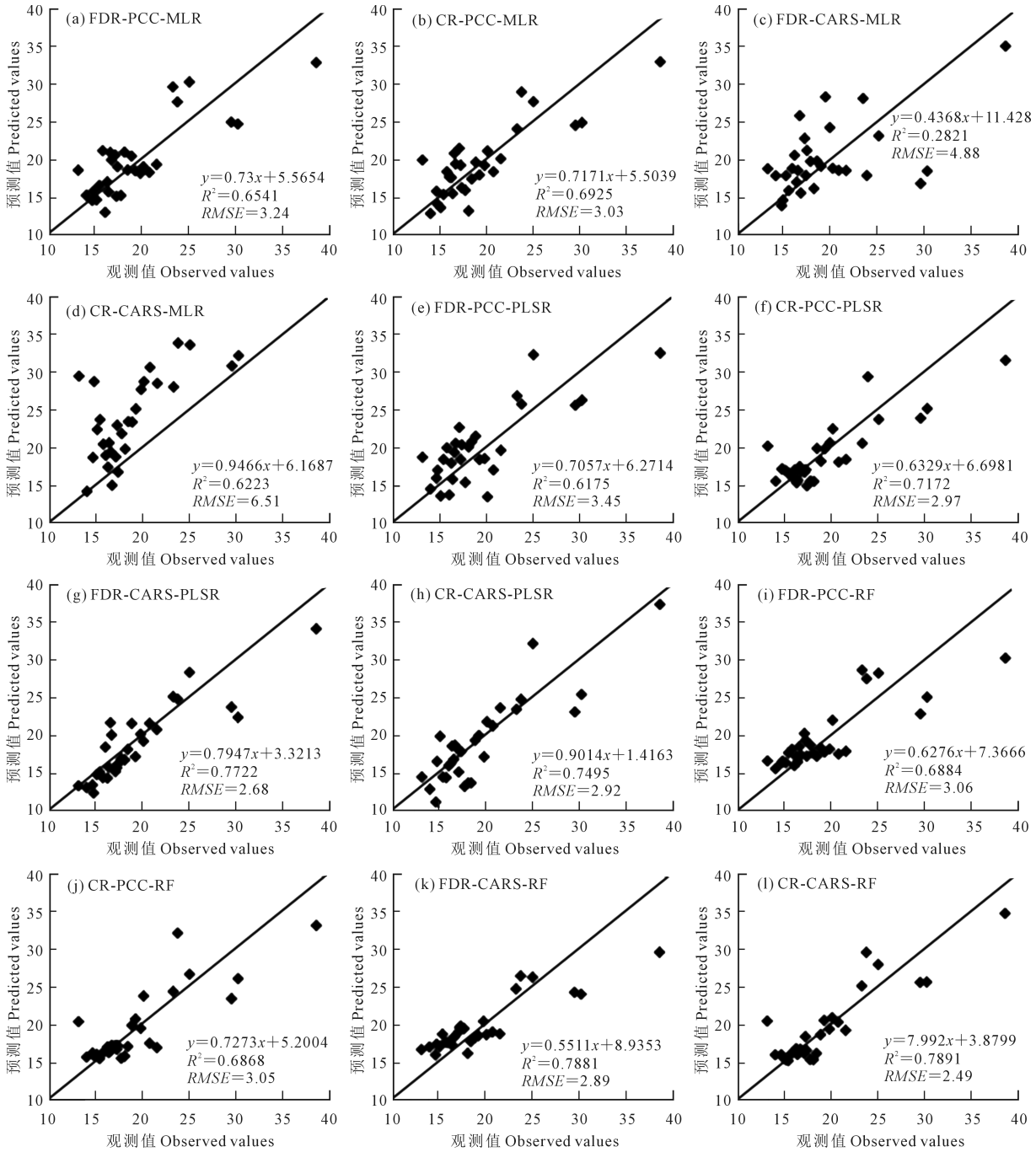

图 2　不同预测方法观测值与预测值散点图

Fig.2　The scatterplot of observed and predicted values by multiple models

参 考 文 献:

[1] 叶勤, 姜雪芹, 李西灿, 等. 基于高光谱数据的土壤有机质含量反演模型比较[J]. 农业机械学报, 2017, 48(3): 164-172.
YE Q, JIANG X Q, LI X C, et al. Comparison on inversion model of soil organic matter content based on hyperspectral data [J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(3): 164-172.

[2] 王丽萍, 刘焕军, 郑树峰, 等. 东北农牧交错带耕地土壤有机质遥感反演研究[J]. 土壤, 2022, 54(1): 184-190.
WANG L P, LIU H J, ZHENG S F, et al. Soil organic matter inversion inagro-pastoral ecotone of northeast China[J]. Soils, 2022, 54(1): 184-190.

[3] MIAO C H, LI X C, LU J H, et al. The modified model of soil organic matter grey relation estimating pattern based on hyper-spectral data [J]. The Journal of Grey System, 2019, 31(2): 51-64.

[4] 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.
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 Sinic A, 2020, 57(3): 636-645.

[5] 张智韬, 劳聪聪, 王海峰, 等. 基于 FOD 和 SVMDA-RF 的土壤有机质含量高光谱预测[J]. 农业机械学报, 2020, 51(1): 156-167.

ZHANG Z T, LAO C C, WANG H F, et al. Estimation of desert soil organic matter through hyperspectra based on fractional-order derivatives and SVMDA-RF[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(1): 156-167.

[6] 刘焕军, 张柏, 赵军, 等. 黑土有机质含量高光谱模型研究[J]. 土壤学报, 2007, 44(1): 27-32.
LIU H J, ZHANG B, ZHANG B, et al. Spectral models for prediction of organic matter in black soil[J]. Acta Pedologica Sinica, 2007, 44(1): 27-32.

[7] 单海斌, 蒋平安, 颜安, 等. 基于高光谱数据的北疆绿洲农田灰漠土有机质反演[J]. 农业资源与环境学报, 2018, 35(3): 276-282.
SHAN H B, JIANG P A, YAN A, et al. Inversion of organic matter content in grey desert soil of Northern Xinjiang oasis farmland based on hyper-spectral data[J]. Journal of Agricultural Resources and Environment, 2018, 35(3): 276-282.

[8] 李媛媛, 李微, 刘远, 等. 基于高光谱遥感土壤有机质含量预测研究[J]. 土壤通报, 2014, 45(6): 1313-1318.
LI Y Y, LI W, LIU Y, et al. Study on the prediction of soil organic matter content based on hyperspectral remote sensing[J]. Chinese Journal of Soil Science, 2014, 45(6): 1313-1318.

[9] 刘磊, 沈润平, 丁国香. 基于高光谱的土壤有机质含量估算研究[J]. 光谱学与光谱分析, 2011, 31(3): 762-766.
LIU L, SHEN R P, DING G X. Studies on the estimation of soil organic matter content based on hyper-spectrum[J]. Spectroscopy and Spectral Analysis, 2011, 31(3): 762-766.

[10] 沈强, 张世文, 夏沙沙, 等. 基于支持向量机的土壤有机质高光谱反演[J]. 安徽理工大学学报(自然科学版), 2019, 39(4): 39-45.
SHEN Q, ZHANG S W, XIA S S, et al. Hyperspectral inversion of soil organic matter based on support vector machine[J]. Journal of Anhui University of Science and Technology (Natural Science), 2019, 39(4): 39-45.

[11] 郄欣, 齐雁冰, 刘姣姣, 等. 基于室内高光谱数据的多种类型土壤有机质估算模型比较[J]. 干旱地区农业研究, 2021, 39(4): 109-116, 124.
XI X, QI Y B, LIU J J, et al. Comparison of multiple estimation models of soil organic matter based on laboratory hyperspectral reflectance in Shaanxi Province[J]. Agricultural Research in the Arid Areas, 2021, 39(4): 109-116, 124.

[12] 王永敏, 李西灿, 田林亚, 等. 土壤有机质含量地面高光谱估测模型对比分析[J]. 国土资源遥感, 2019, 31(1): 110-116.
WANG Y M, LI X C, TIAN L Y, et al. Comparison and analysis of estimation models of soil organic matter content established by hyperspectral on ground[J]. Remote Sensing for Land & Resources, 2019, 31(1): 110-116.

[13] 刘杰亚, 李西灿, 任文静, 等. 利用卷积神经网络的土壤有机质含量高光谱估测[J]. 河北农业大学学报, 2023, 46(4): 118-124.
LIU J Y, LI X C, REN W J, et al. Hyperspectral estimation of soil organic matter content using convolutional neural network[J]. Journal of Agricultural University of Hebei, 2023, 46(4): 118-124.

[14] 杨煜岑, 杨联安, 任丽, 等. 基于随机森林的农耕区土壤有机质空间分布预测[J]. 浙江农业学报, 2018, 30(7): 1211-1217.
YANG Y C, YANG L A, REN L, et al. Prediction for spatial distribution of soil organic matter based on random forest model in cultivated area[J]. Acta Agriculturae Zhejiangensis, 2018, 30(7): 1211-1217.

[15] 韩杏杏, 陈杰, 王海洋, 等. 基于随机森林模型的耕地表层土壤有机质含量空间预测①——以河南省辉县市为例[J]. 土壤, 2019, 51(1): 152-159.
HAN X X, CHEN J, WANG H Y, et al. Spatial prediction of SOM content in topsoil based on random forest algorithm: A case study of Huixian City, Henan Province[J]. Soils, 2019, 51(1): 152-159.

[16] 刘恬琳, 朱西存, 白雪源, 等. 土壤有机质含量高光谱估测模型构建及精度对比[J]. 智慧农业, 2020, 2(3): 129-138.
LIU T L, ZHU X C, BAI X Y, et al. Hyperspectral estimation model construction and accuracy comparison of soil organic matter content[J]. Smart Agriculture, 2020, 2(3): 129-138.

[17] 白婷, 丁建丽, 王敬哲. 基于机器学习算法的土壤有机质质量比估算[J]. 排灌机械工程学报, 2020, 38(8): 829-834.
BAI T, DING J L, WANG J Z. Estimation of soil organic matter content based on machine learning[J]. Journal of Drainage and Irrigation Machinery Engineering, 2020, 38(8): 829-834.

[18] 唐海涛, 孟祥添, 苏循新, 等. 基于 CARS 算法的不同类型土壤有机质高光谱预测[J]. 农业工程学报, 2021, 37(2): 105-113.
TANG H T, MENG X T, SU X X, et al. Hyperspectral prediction on soil organic matter of different types using CARS algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(2): 105-113.

[19] 李胜男, 曹萌萌, 李盛楠, 等. 黑土典型区有机质高光谱预测模型[J]. 国土与自然资源研究, 2016, (4): 73-76.
LI S N, CAO M M, LI S N, et al. Hyper spectral prediction model of organic matter in black soil region[J]. Territory & Natural Resources Study, 2016, (4): 73-76.

[20] 夏晨真, 姜艳艳, 张星宇, 等. 基于无人机高光谱影像的黑土区玉米农田土壤有机质估算[J]. 光谱学与光谱分析, 2023, 43(8): 2617-2626.
XIA C Z, JIANG Y Y, ZHANG X Y, et al. Estimation of soil organic matter in maize field of black soil area based on UAV hyperspectral image[J]. Spectroscopy and Spectral Analysis, 2023, 43(8): 2617-2626.

[21] 周伟, 谢利娟, 杨晗, 等. 基于高光谱的三江源区土壤有机质含量反演[J]. 土壤通报, 2021, 52(3): 564-574.
ZHOU W, XIE L J, YANG H, et al. Hyperspectral inversion of soil organic matter content in the three-rivers source region[J]. Chinese Journal of Soil Science, 2021, 52(3): 564-574.

[22] 丁启东, 王怡婧, 张俊华, 等. 基于高光谱信息的宁夏土壤有机质含量估算[J]. 生态学杂志, 2024, 43(8): 2531-2538.
DING Q D, WANG Y J, ZHANG J H, et al. Estimation of soil organic matter content in Ningxia based on hyperspectral information[J]. Chinese Journal of Ecology, 2024, 43(8): 2531-2538.