江苏农业学报(*Jiangsu J.of Agr.Sci.*),2024,40(1):112-120
http://jsnyxb.jaas.ac.cn

张永亮,汪　泓,肖玖军,等. 基于高光谱的山区耕地土壤有机质含量估测[J]. 江苏农业学报,2024,40(1):112-120.
doi:10.3969/j.issn.1000-4440.2024.01.012

基于高光谱的山区耕地土壤有机质含量估测

张永亮[1]，汪　泓[1]，肖玖军[2,3]，李可相[2,3]，王　宇[1]，邢　丹[4]
(1.贵州大学矿业学院,贵州 贵阳 550025；2.贵州省山地资源研究所,贵州 贵阳 550001；3.贵州省土地绿色整治工程研究中心,贵州 贵阳 550001；4.贵州省农业科学院辣椒研究所,贵州 贵阳 550009)

摘要： 以贵州省典型山区耕地土壤高光谱数据为研究对象,基于光谱变换法和机器学习原理构建贵州省山区耕地土壤有机质(SOM)含量估算模型。于2020年8月至2021年3月在贵州省13个县(区、市)采集了120个土壤样品,检测土壤可见光-近红外波段光谱信息,利用5种光谱数据变换(原始光谱、一阶微分、二阶微分、倒数对数的一阶微分、连续统去除)和4类模型(偏最小二乘回归、支持向量机、随机森林和BP神经网络)组合出不同土壤有机质含量的预测模型,按照3∶1选择训练样本和测试样本以估算山区SOM含量。结果表明,一阶微分数据变换与山区SOM含量的相关性较高,相关系数最高达到-0.635;反演模型中,基于一阶微分光谱变换构建的BP神经网络模型精度最高,训练集、测试集的决定系数(R^2)分别为0.845、0.838,测试集均方根误差(*RMSE*)为3.452,相对分析误差(*RPD*)达到2.470,其次是RF、PLSR模型的*RPD*较高,SVM模型的*RPD*最低。光谱数据变换中一阶微分法能极大程度提取出山区耕地的SOM含量信息,BP神经网络模型是估算山区SOM含量的最优模型,本研究结果可为贵州省山区耕地土壤肥力的监测以及农业生产提供理论参考。

关键词： 土壤有机质；高光谱；山区耕地；一阶微分；BP神经网络
中图分类号： S153　　**文献标识码：** A　　**文章编号：** 1000-4440(2024)01-0112-09

Estimation of soil organic matter content in mountain farmland based on hyperspectral data

ZHANG Yong-liang[1], WANG Hong[1], XIAO Jiu-jun[2,3], LI Ke-xiang[2,3], WANG Yu[1], XING Dan[4]
(1.*Mining College of Guizhou University*, *Guiyang* 550025, *China*; 2.*Guizhou Provincial Institute of Mountain Resources*, *Guiyang* 550001, *China*; 3.*Guizhou Province Land Green Remediation Engineering Research Center*, *Guiyang* 550001, *China*; 4.*Chili Research Institute*, *Guizhou Academy of Agricultural Sciences*, *Guiyang* 550009, *China*)

Abstract: Taking the hyperspectral data of cultivated land in typical mountainous areas of Guizhou province as the research object, a model for estimating soil organic matter (SOM) content in mountainous areas of Guizhou province was established by using spectral transformation method and machine learning. From August 2020 to March 2021, 120 soil samples were collected from 13 counties and cities of Guizhou province, and the visible near-infrared spectral information of soil was detected. Five spectral data transformations (original spectra, first-order differential, second-order differential, first-order differential of reciprocal logarithm, continuum removal) and four types of models (partial least squares regression, support vector machine,

收稿日期:2022-11-18
基金项目:贵州科学院青年基金项目[黔科院J字(2018)25号];贵州省科技支撑计划项目[黔科合支撑(2020)1Y172号];贵州省科技支撑计划项目[黔科合支撑(2021)一般496号];国家重点研发计划项目(2022YDF1100307);贵州省基础研究计划项目[黔科合基础-ZK(2021)一般100号、黔科合基础-ZK(2022)一般276号]
作者简介:张永亮(1995-),男,贵州都匀人,硕士研究生,主要从事土壤高光谱遥感方面的研究。(E-mail)17864159359@163.com
通讯作者:汪　泓,(E-mail)7653606@qq.com

random forest and BP neural network) were used to combine different soil organic matter content prediction models. The training samples and test samples were selected according to the ratio of 3 ∶ 1 to estimate the SOM content in mountain area. The correlation between the first-order differential data transformation and the SOM content in mountain area was high, and the highest correlation coefficient was −0.635. In the inversion model, the BP neural network model based on the first-order differential spectral transformation had the highest accuracy. The determination coefficients (R^2) of the training set and the test set were 0.845 and 0.838, respectively. The root mean square error (*RMSE*) of the test set was 3.452. The relative analysis error (*RPD*) reached 2.470, followed by RF, PLSR and SVM. The first-order differential method in spectral data transformation could greatly extract the SOM content information of mountain cultivated land. The BP neural network model was the optimal model for estimating the SOM content in mountain areas. The results of this study can provide theoretical reference for the monitoring of soil fertility and agricultural production in mountainous areas of Guizhou province.

Key words: soil organic matter; hyperspectral; mountainous farmland; first-order differential; BP neural network

贵州全省山地面积占比近 90%,是全国唯一一个没有平原支撑的省份[1]。山区耕地的零碎化分布,导致部分耕地利用率低,土壤质地分布不均,在一定程度上限制了农业生产。土壤有机质(Soil organic matter, SOM)是存在于土壤中的含碳有机化合物的总和,具有提供养分、保水保肥、促进土壤团粒结构形成及改善土壤理化性质等作用[2-3],其含量是衡量土壤肥力的重要指标,快速、准确地监测 SOM 含量对于山区耕地科学管理具有重要意义。传统的 SOM 含量测定主要通过田间取样、实验室化验分析,该方法使用成本较高且效果欠佳[4]。近年来,高光谱遥感技术以其时效高、信息量大且无污染的优势逐渐在 SOM 快速检测中得到应用[5]。

近年来,研究者针对不同地区的土壤性质,利用高光谱遥感技术从数据处理和模型算法等方面反演出契合当地的 SOM 预测模型。韩兆迎等[6]通过相关分析确定了 7 个特征波段,建立了 SOM 含量估测模型,发现用二次多项式逐步回归模型反演黄河三角洲土壤 SOM 含量的效果最优;勾宇轩等[7]用小波变换结合稳定性竞争自适应重加权采样(Stability competitive adaptive reweighted sampling, CARS)算法,较好地反演了东北旱作农田土壤类型的 SOM 含量;南锋等[8]使用偏最小二乘回归(Partial least squares regression, PLSR)分析方法,建立了能够很好地反演黄土高原煤矿区复垦农田 SOM 含量的模型。Nawar 等[9-10]利用便携式光谱仪(Analytical Spectral Devices FieldSpec4 Standard-Res, ASD)获取埃及四奈北部地区土壤光谱信息,对光谱数据采用 7 种预处理技术预处理之后构建线性 PLSR、非线性支持向量机回归(Support vector machine, SVM)和多元自适应回归样条(Multivariate adaptive regression splines, MARS)等 3 种模型进行盐渍土有机质含量的预测,交叉验证结果显示,MARS 模型的预测效果最佳。张娟娟等[11]将遗传算法与 SVM 回归结合进行砂姜黑土 SOM 含量的估测,发现决定系数(R^2)高达 0.95。张森等[12]用反向传播(Back propagation, BP)神经网络、SVM 模型对滨海湿地土壤有机质含量进行估算,结果显示,用 SVM 模型估测的 SOM 含量在精度方面明显更优。钟亮等[13]基于卷积神经网络(Convolutional neural network, CNN)模型,探讨不同网络结构对 SOM 含量预测的建模效果,经大量训练得出,小卷积核的 VGGNet-7 适用于红壤地区 SOM 含量的预测且 CNN 能够简化光谱预处理过程。

从以上研究结果可以看出,估测 SOM 含量的机器学习方法大多基于线性与非线性模型,各地区适用的模型均不相同,主要与土壤的光谱特性、数据处理和建模方法的选择有关。从目前的研究内容看,基于平原地区展开的研究相对偏多,这是由于其成土母质受到适宜的湿度、光照条件的影响,使得土壤理化性质良好,因此通过线性模型即可稳定高效地对该地区 SOM 含量进行反演,如武彦清等[14]分别采用多元线性逐步回归和 PLSR 2 种方法建立的模型均能满足松嫩平原 SOM 含量的速测要求,陆龙妹等[15]利用 PLSR 方法建立的 SOM 含量光谱预测模型能预测出淮北平原 SOM 含量,文锡梅等[16]利用 PLSR 模型定量反演出喀斯特地区 SOM 含量并获得较好的模型精度,但小范围研究区和单一种类土壤建模的模型通用性还较为欠缺。贵州省内地喀斯特地貌分布广泛,地形复杂且气候多变,土壤干旱、侵蚀现象较为严重,耕地分布零碎且土壤类型多样,在此地进行大范围土壤光谱监测容易造成光谱数据冗

余,反演出的模型在进行较大尺度 SOM 含量估算时精度欠佳。因此,运用合适的模型算法估测山区耕地 SOM 含量是当前亟待解决的问题之一。本研究拟以从贵州山区耕地采集的 120 个土壤样本为研究对象,通过 ASD 便携式地物光谱仪采集样品光谱,在 PLSR 基础上探讨非线性模型[SVM、随机森林(Random forest, RF)、BP]在山区耕地 SOM 含量反演中的结果,通过对比分析以获得精度最高的光谱变换和模型组合,以期为山区耕地 SOM 含量估测提供快速可靠的算法。

1 材料与方法

1.1 研究区概况

贵州省地处中国西南内陆腹地,在地形上属于中国西南高原山区,地势特点是自西向东低,由中向北、向东、向南倾斜。研究区选取贵州省贵阳市、遵义市、黔南州、黔东南州和毕节市等 5 个地区下辖的 13 个县(区、市),图 1 是研究区内部分采样点及其 13 个县(区、市)的边界范围,采样点耕地分布在山地、丘陵和沟谷等地域,在海拔620~1 580 m 内采集土样,土类以黄壤、黄色石灰土、水稻土和紫泥土为主。贵州省占比最高的土壤类型是黄壤,占全省土壤面积的 46.4%[17];黄色石灰土分布范围最广,各地均有分布,但相对集中分布在黔中地区,大泥土属于黄色石灰土的一类;水稻土是贵州省农业生产中极为重要的土壤资源,92.8% 水稻土分布在海拔 1 400 m 以下的区域;紫泥土的面积相对较少,主要分布在黔北、黔西北等高海拔地区。据前人记载,贵州省山区的 SOM 含量总体较高,但由于耕地分布零碎,导致其撂荒严重[18]。

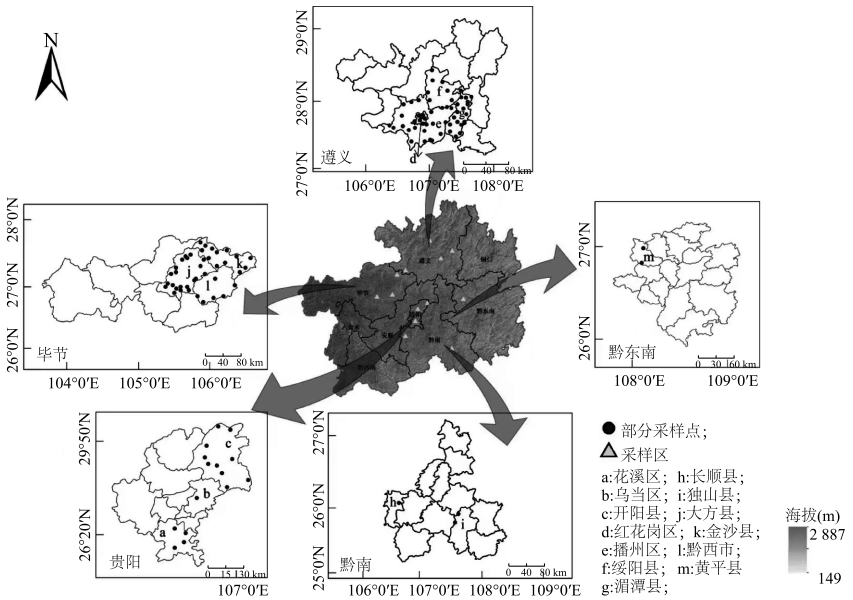


图 1 研究区域的部分采样点分布

Fig.1 Distribution of some sampling points in the study area

1.2 土壤样品采集与处理

根据贵州省土壤空间分布特征,于 2020 年 8 月至 2021 年 3 月在研究区开展土壤取样,共计 120 个土壤样品,研究区域土壤概况见表 1。在研究区内采集耕地表层 20 cm 以内的土壤作为样本,用手持全球定位系统(GPS)定位并随时记录相关信息。经实验室风干、去杂、研磨后过 2 mm 筛,分成 2 份,分别用于光谱采集和有机质含量的测定。土壤有机质含量的测定采用重铬酸钾-硫酸硝化法[19]。

1.3 土壤高光谱数据的测定

土壤原始光谱反射率的采集使用 ASD 便携式地物波谱仪,光谱范围为350~2 500 nm,光谱重采样间隔为 1 nm,在暗室条件下测定,将土样放入直径 15 cm、深度 2 cm 的硼硅玻璃皿中,用尺子刮平表面。将高密度探头贴近土壤样品,使探头视野充满土壤样品,固定探头后垂直对准被测物体。光源使用高密度探头自带光源,成功进行初始白板校正后,确保高密度探头和土壤样品的相对位置保持不变。

在获得原始光谱反射率后,再将土壤样品旋转 3 次,旋转角度为 90°,分别采集 10 条光谱曲线,土壤样品的光谱反射数据为 10 条光谱曲线的均值[20]。每间隔15~20 min 重新进行优化。

1.4 数据预处理

由于土壤样品光谱曲线边缘的350~400 nm、2 400~2 500 nm 部分受到外界噪声的影响较大,因此将其去除以减少干扰。在 OriginPro 2021 软件中用 Savitzky-Golay(SG)[21-22]滤波进行 9 点平滑去噪处理,该滤波方法是一种在时域内基于局域多项式最小二乘法拟合的滤波方法,其最大特点是在滤除噪声的同时可以保持信号的形状和宽度不变。为更有效筛选山区土壤光谱的特征波段,对平滑后的原始光谱反射率(R)进行一阶微分(First derivative, FD)、二阶微分(Second derivative, SD)、倒数对数的一阶微分(First derivative of reciprocal logarithm, LRD)、连续统去除(Continuum removal, CR)等 4 种变换处理。光谱一阶微分处理可在消除背景噪声干扰的同时提高光谱分辨率、降低相关波段的寻找难度[23-25],倒数对数变换法可减少乘数因子对光照条件变化的影响[26],连续统去除法有利于突出光谱曲线的吸收、反射特征,分类识别提取敏感波段[27-28]。上述过程中的 FD、SD、LRD 处理在 Matlab R 2016b、The Unscrambler X10.4 软件中完成,CR 处理在 ENVI 5.3 中完成。

1.5 模型的构建与精度验证

PLSR 集结了主成分分析、典型相关分析和线性回归分析的特点,在同时包含多个变量的情况下能实现多对多的模型构建,并在一定程度上解决自变量之间共线的问题[29],因此采用 The Unscrambler X10.4 软件的 PLSR 模块完成 SOM 含量反演模型。

SVM 可将数据从低维空间映射到高维空间中,然后在此高维空间中进行线性回归,从而取得在原空间非线性回归的效果[30-31]。SVM 模型构建在 Matlab 中完成,SVM 模型参数设定如下:类型选择 C-SVC,核函数类型为 RBF,惩罚因子($Cost$)为 1,核函数系数($Gamma$)为 0.001,损失函数的 P 值为 0.01,收敛精度(Eps)为 0.001[32]。

RF 属于机器模型,它通过随机方式形成了由多个决策树组成的一片森林,当新样本作为数据变量输入到构建好的森林中时,森林中的每棵决策树就会分别判断并识别这个样本所属的类别[33],再统计哪个类别被判定得最多,进而预测该样本所属的类别。RF 可产生高准确度的分类器,处理大量的数据变量,在判断类别时还能考虑变量的重要性,且训练速度快。RF 模型构建在 Matlab 中完成,RF 的参数设置如下:决策数目($Ntree$)为 200,训练节点变量数($Mtry$)为 2。

表 1 研究区域的土壤概况

Table 1 Soil profile of study area

研究区	海拔(m)	土壤类型	样本数	平均有机质含量(g/kg)
花溪区	1 030~1 060	黄色石灰土	4	12.22
乌当区	1 315~1 320	黄壤、水稻土	11	32.88
开阳县	620	黄壤	1	20.80
红花岗区	830~1 070	黄色石灰土、大泥土	12	27.75
播州区	844~1 160	黄壤	20	28.77
绥阳县	890~930	黄色石灰土	7	22.66
湄潭县	760~930	水稻土、黄泥土	19	30.55
独山县	910~940	黄壤	2	29.05
长顺县	680	水稻土	1	32.90
黄平县	1 050	水稻土	2	45.60
大方县	1 310~1 410	大泥土、紫泥土	22	28.89
金沙县	930~1 210	紫泥土	13	27.25
黔西市	1 420~1 580	大泥土	6	36.80

BP 神经网络是一种非线性映射模型,具有完整的数学算法,理论上能够无限逼近任意复杂的非线性函数[34],对于样品较多的机器学习问题,传统的线性回归会存在欠拟合或过拟合现象,神经网络可以让它们不断训练以达到最好效果。BP 神经网络模型在 Matlab 中完成,训练参数设置如下:迭代次数为1 000次,训练均方根误差小于 10^{-5},神经元设置为 5 个,学习率为 0.05,最大失败次数为 5 次,经过试凑法最终确定 BP 神经网络模型隐含层节点数依次为 3 个、6 个、9 个[35]。

模型精度测试用如下 3 个参数进行评估:决定系数(Determination coefficient, R^2)、均方根误差(Root mean square error, $RMSE$)和相对分析误差(Residual predictive deviation, RPD)。其中,R^2 用于测量模型的稳定性,R^2>0.6 表明模型能够粗略预测 SOM 含量;R^2>0.8,表明模型的稳定性较强[36]。$RMSE$ 用来检验模型的预测能力,$RMSE$ 越小,表明

模型的精度越高。*RPD* 用来评价测试模型的预测能力，当*RPD*>2.0 时，说明模型的预测效果较好；当1.4≤*RPD*≤2.0 时，说明模型具有基本预测能力，经过改进后预测效果更好；当*RPD*<1.4 时，模型预测能力较弱。相关公式如下：

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

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

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}} \quad (3)$$

$$RPD = \frac{SD}{RMSE} \quad (4)$$

式中：y_i、$\hat{y}_i$ 分别代表样品实测值、预测值；$\bar{y}_i$ 为样品均值；*SD* 为标准差；*n* 为样本数。

2 结果与分析

2.1 土壤有机质含量的统计分析

将 120 个土壤样本按有机质含量从大到小排序，根据有机质含量梯度，按照3∶1 的比例选取训练样本和测试样本，最终确定 90 个训练样本、30 个测试样本。由表 2 可以看出，土壤有机质含量为 11.40~48.60 g/kg，均值为 28.91 g/kg，标准差为 8.31 g/kg，训练样本、测试样本的标准差分别为 8.24 g/kg、8.53 g/kg，总体变异系数偏中等。

表 2 贵州山区土壤有机质含量的统计分析结果

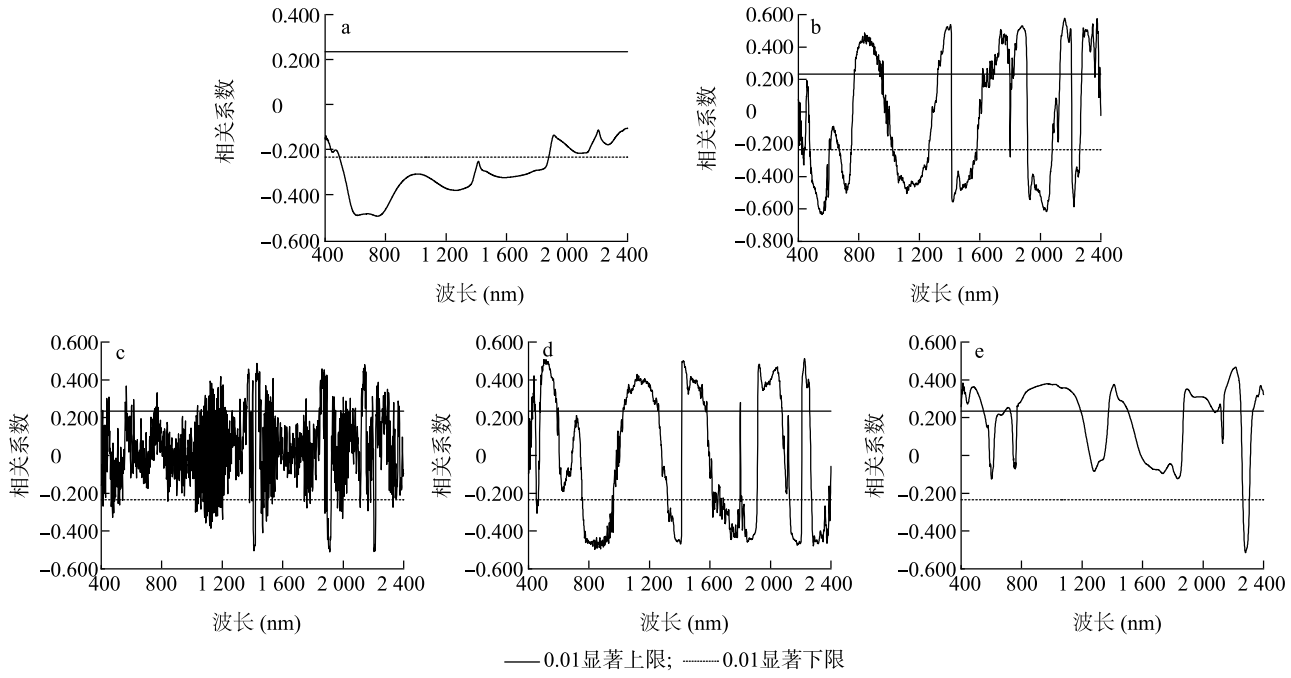
Table 2 Statistical analysis for soil organic matter content in Guizhou mountainous areas

样本类型	样本数	有机质含量(g/kg)				变异系数(%)
		最大值	最小值	平均值	标准差	
总体样本	120	48.60	11.40	28.91	8.31	28.76
训练样本	90	48.60	11.40	28.94	8.24	28.48
测试样本	30	48.00	11.40	28.80	8.53	29.62

2.2 土壤有机质含量的光谱相关分析

山区耕地的土壤高光谱在 5 种不同形式下与 SOM 含量之间的相关性分析结果见图 2。可以看出，原始光谱(*R*)与 SOM 含量整体呈负相关并在可见光部分相关系数达到极值(图 2a)；4 种光谱变换处理下，在可见光-近红外光范围内均有波段在正负值之间波动，并且有不少波段通过 0.01 水平的显著性检验；经 FD 变换提高了光谱与 SOM 含量在近红外范围内的相关性，敏感波段从可见光至近红外光之间呈均匀分配，有1 494个波段通过显著性检验，且与 SOM 含量呈极显著相关($P \leq 0.01$)，相关系数最高为−0.635(图 2b)；SD 处理后的光谱在近红外部分频繁出现吸收谷、反射峰，敏感波段范围也集中在此部分，统计有 461 个波段与 SOM 含量呈极显著相关($P \leq 0.01$)，相关系数极值为−0.561(图 2c)；LRD 与 FD 数据的变换相似，共有1 455个敏感波段，由于先经过倒数对数变换的原因，LRD 与 FD 的相关系数图类似于对称分布，相关系数极值为 0.512(图 2d)；通过 CR 变换，使得土壤光谱和有机质含量间大部分呈正相关，说明 CR 变换能增强山区土壤光谱的吸收特征，通过显著性检验的波段有 1 035个，相关系数极值为 0.514(图 2e)。基于相关系数绝对值、通过显著性波段数量，得出如下排序：FD>LRD>CR>SD，该排序说明，光谱数据经过 FD、LRD 变换后能提高山区耕地 SOM 含量与光谱波段之间的相关性，更有利于筛选特征波段。

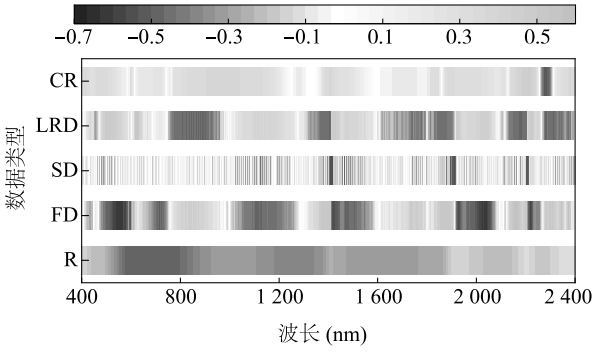
图 3 为 SOM 含量与不同形式光谱间相关性分析的特征波段范围，可以看出，原始光谱的特征波段大多集中在可见光部分，经光谱技术变换后，在不同程度上挖掘出近红外部分的特征光谱信息，说明可见光-近红外光范围都蕴含山区耕地土壤的特征波段，所以本研究通过显著性检验后，在不同变换形式的光谱中筛选出它们在可见光-近红外光部分相关系数较高的 60 个波段用于建模。这与于雷等[37]利用竞争性自适应重加权-连续投影算法筛选得到的建模波段以近红外光(1 800~2 400 nm)部分为主的结果有差异。推测其原因，可能由于研究区及土壤类型不同，前人以江汉平原的潮土、水稻田和黄棕壤为研究对象，该地区雨量充沛且土壤耕作程度高，其土壤含水量相对较高，而土壤含水量的光谱特性主要集中在近红外波段范围[38]，因此土壤含水量会影响土壤反射率和有机质含量相关的光谱信息。由于在1 400 nm、1 900 nm 和2 200 nm 等波段附近具有强烈的水分吸收谷，与黏土矿物中所含的 OH^- 有关[39]，因此最后获得的 5 种光谱的特征响应波段应将其剔除。



a:原始光谱;b:一阶微分;c:二阶微分;d:倒数对数的一阶微分;e:连续统去除。

图 2 山区土壤有机质含量与光谱变换间的相关系数

Fig.2 Correlation coefficient between soil organic matter content and spectral transformation in mountainous areas



CR:连续统去除;LRD:倒数对数的一阶微分;SD:二阶微分;FD:一阶微分;R:原始光谱反射率。

图 3 土壤有机质含量与不同形式光谱相关性分析的特征波段范围

Fig.3 The characteristic band range of correlation analysis between soil organic matter content and different forms of spectrum

2.3 不同估测模型 SOM 含量反演的结果

表 3 是山区耕地 SOM 含量反演的不同光谱变化和模型组合,可以看出,训练集的模型预测精度基本高于测试集,可能与二者样本数量差异有密切关系。基于偏最小二乘回归的模型中,除了 SD 光谱变换反演的模型外,其余 3 种光谱变换的模型均能达到预测土壤有机质含量的基本要求,表现最好的是 LRD,测试集的 R^2、*RPD* 分别为 0.717(>0.600)、1.894(>1.400),而 SD 的精度最低(其测试集的 R^2、*RPD* 分别为 0.457(<0.600)、1.357(<1.400),不适用于山区耕地 SOM 含量的反演。SVM 模型相对 PLSR 反演 SOM 含量的效果整体呈现略微下降的趋势,虽然总体 R^2 均大于 0.600,但模型拟合效果一般,不能作为山区耕地 SOM 含量估测的首选模型。在反演 SOM 含量的效果上,RF 模型相较于 PLSR 有明显提高,其中训练集的 R^2 均远高于同等变换的其他模型,R^2 最高达到 0.926,因此在训练集中,RF 可作为理想的估测模型使用,测试集中 FD、SD、CR 的 R^2 均大于 0.750,*RPD* 均大于 2.000,整体模型预测效果表现良好。4 类模型中,BP 神经网络预测模型的预测能力最高,测试集经过 FD、LRD 数据变换组合的 BP 神经网络模型 R^2 均在 0.800 以上,与它们对等的训练集间的差距进一步缩小,与 RF 相比具有明显优势,BP 建模效果最佳的模型组合是 FD-BP,训练集的 R^2 为 0.845,*RMSE* 最小,为 3.259,测试集的 R^2 为 0.838,*RMSE* 为 3.452,*RPD* 为 2.470,相对分析误差以 FD 最高,说明 BP 神经网络模型在预测山区耕地 SOM 含量方面具有较高的稳定性,可以进行有效预测。图 4 为 30 个测试样本

代入训练模型所得到实测值与预测值的散点图,通过分析 20 个模型各项精度指标,能够筛选出反演效果最好的 6 个模型组合。

表 3 不同组合形式土壤有机质含量的光谱反演模型精度

Table 3 Accuracy of spectral inversion model of soil organic matter content in different combinations

模型	数据类型	训练集		测试集		
		R^2	*RMSE*	R^2	*RMSE*	*RPD*
PLSR	R	0.712	4.421	0.635	5.196	1.657
	FD	0.781	3.858	0.693	4.840	1.807
	SD	0.445	6.139	0.457	6.543	1.357
	LRD	0.784	3.854	0.717	4.767	1.894
	CR	0.675	4.592	0.674	4.914	1.752
SVM	R	0.662	4.788	0.628	5.236	1.634
	FD	0.809	3.597	0.637	5.173	1.664
	SD	0.629	5.018	0.617	5.547	1.601
	LRD	0.729	4.269	0.681	4.883	1.759
	CR	0.717	4.281	0.598	5.563	1.569
RF	R	0.859	3.091	0.682	4.874	1.772
	FD	0.926	2.232	0.781	4.192	2.109
	SD	0.907	2.513	0.777	4.213	2.082
	LRD	0.872	2.989	0.663	5.034	1.725
	CR	0.910	2.412	0.760	4.361	2.017
BP	R	0.796	3.722	0.727	4.513	1.908
	FD	0.845	3.259	0.838	3.452	2.470
	SD	0.726	4.319	0.661	5.057	1.737
	LRD	0.840	3.323	0.803	3.876	2.251
	CR	0.832	3.382	0.694	4.863	1.816

PLSR:偏最小二乘回归;SVM:支持向量机;RF:随机森林;BP:反向传播。R:原始光谱反射率;FD:一阶微分;SD:二阶微分;LRD:倒数对数的一阶微分;CR:连续统去除;R^2:决定系数;*RMSE*:均方根误差,*RPD*:相对分析误差。

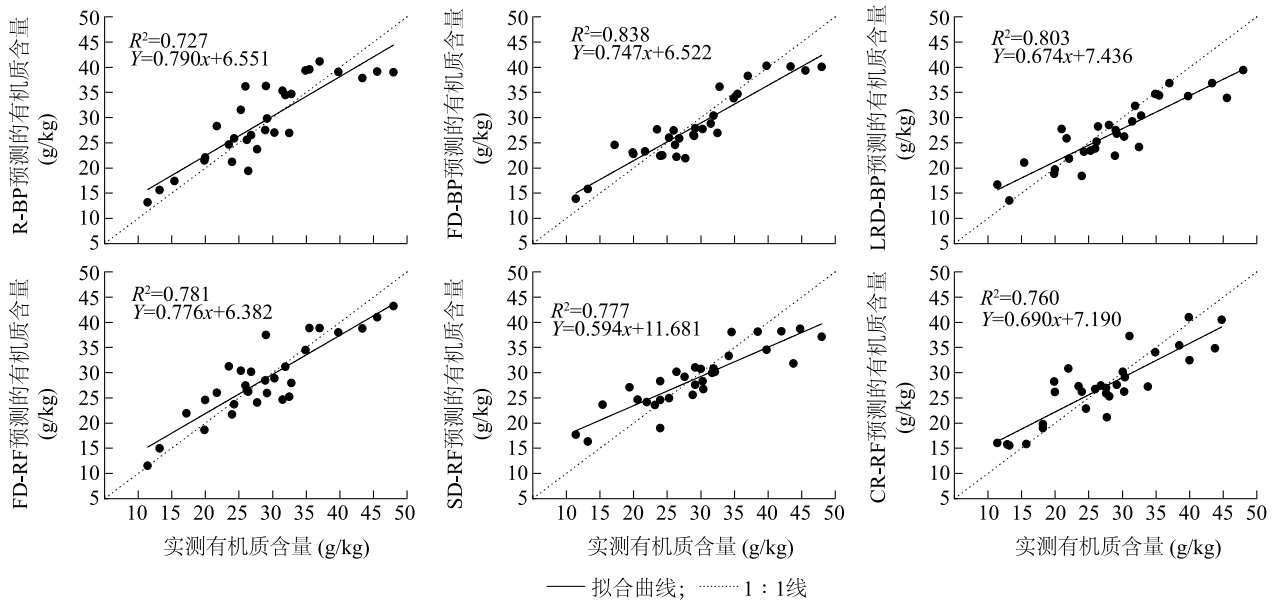
通过综合训练集和测试集各项验证指标分析发现,在不同变换的光谱数据与模型组合中,FD-BP 模型具有最稳定的估测能力,其次是 LRD-BP、R-BP、FD-RF、SD-RF 和 CR-RF 模型,有良好的预估能力。在数据变换方面,能提高模型预测 SOM 含量精度的光谱变换排序是FD>LRD>CR>SD。在模型选择方面,更适合山区 SOM 含量反演的模型依次为 BP、RF、PLSR、SVM。

3 讨 论

由建模结果可以看出,经过 FD、LRD 的光谱数据变换组合的模型反演效果较好,表明 FD、LRD 变换不仅能消除光谱曲线周围噪声、提高光谱分辨率,还能突出可见光至近红外范围波段间的差异,提高光谱与 SOM 含量之间的相关性,增加敏感波段的数量,使建模精度得到保障,这与前人的研究结论相符[40];而 SD 变换让光谱与 SOM 含量相关性分析反应过甚,敏感波段间多重相关,无法有效提取山区土壤信息,导致建模效果不理想;CR 与原始光谱组合的模型效果对比没有明显变化,推测是因为相关系数偏低且敏感波段数量偏少,因此建模效果一般。

针对贵州省山区耕地分布零碎、土壤类型多样等问题,本研究的采样区域零星遍布于贵州省 13 个县(区、市),涵盖黄壤、黄色石灰土、水稻田和紫泥土等土壤。由于土样种类多且研究区域广,光谱在有效范围内蕴藏的土壤信息易出现重合、交叉等问题。与以往学者的研究结果相比,在筛选特征波段方面,本研究根据相关系数由大到小的原则,有间隔地挑选出敏感波段及波段范围,并且在原始及其他变换的光谱中选取同样数量的特征波段用于建模,既能降低高光谱数据冗余,又可保证后期山区耕地 SOM 含量反演模型的客观性。

本研究对线性模型(PLSR)和非线性模型(SVM、RF、BP)进行试验对比,探讨不同模型对贵州山区耕地 SOM 含量反演的实用性。研究得出,非线性建模效果更佳,这与周伟等[41]通过 PLSR、SVM、RF 对三江源地区土壤有机质含量反演得出的结论一致。经典线性模型 PLSR 反演出的模型效果仅能达到山区耕地 SOM 含量估测的基础水平,原因可能是 PLSR 没有充分提取出山区 SOM 含量的主成分信息,在未来的试验中,可事先采用主成分确定各类 SOM 含量相关性最大的波段范围,通过综合对比筛选出有效光谱波段进行建模以提高 PLSR 的精度模型。本研究与前人研究的不同之处在于,SVM 模型精度没有达到理想效果,推测与 SVM 所选用的 RBF 核函数有关,本研究中测试集的特征波段数量远大于样本数量,当核函数映射维度非常高时,计算量过大,导致 SVM 泛化能力变差,而可见光、近红外光 2 个部分含有大量山区 SOM 含量不可或缺的特征波段。可选择 Polynomial 或 Sigmoid 核函数再利用交叉验证方法寻找最佳参数以优化 SVM 的反演效果[42]。RF 测试集与训练集的精度相距甚远,离散程度大,造成测试集存在欠拟合的现象,可能与样



BP、RF、R、FD、SD、LRD、CR、R^2 见表 3 注。

图 4 贵州山区土壤有机质含量最优组合模型的测试集散点图

Fig.4 Scatter diagram of test set of soil organic matter content optimal combination model in Guizhou mountainous areas

本数量有很大关系,同时说明随机森林更适合对多变量的数据样本进行建模。综合 4 类模型比较发现,BP 神经网络模型可以更精准地反演山区 SOM 含量,这与 BP 神经网络可以实现以任意精度逼近任何连续函数有关[43],即它能随研究对象复杂程度的增加,通过调节隐含层节点数以提高模型精度。

另外,本研究重在讨论适用于贵州山区耕地 SOM 含量的估测模型,但受到贵州山区地形条件限制,部分研究区位于山高坡陡区域,耕作化程度低且交通不便,导致采样难度大、样本数量偏少。后续研究中我们将扩大采样范围和样本数量,优化分析以进一步提升模型的通用性。

4 结 论

贵州山区土壤的高光谱数据通过 SG 光谱预处理和 4 种光谱数据变换,在不同程度上提高了它们与 SOM 含量之间的相关性,其中一阶微分变换可充分挖掘山区土壤信息,通过显著性检验的波段数多达1 494个,相关系数最高达到-0.635。

与 SOM 含量进行相关性分析得出的敏感波段数量越多且范围(可见光-近红外)越宽,其构建的模型效果越好,说明通过相关系数由大到小的原则在光谱有效范围内均匀筛选的波段不仅能代表土样信息,还能在建模时减少自变量之间多重相关等问题。在估测山区耕地 SOM 含量方面,PLSR 具有粗略的估测能力;SVM 模型对山区耕地 SOM 含量的建模效果不佳;RF 优于前两者但测试模型精度一般;非线性模型中 BP 神经网络以其精度高、稳定性好等特点而适用于山区耕地 SOM 含量估测,以一阶微分-BP 神经网络预测效果最优(训练集:R^2 = 0.845, $RMSE$ = 3.259;测试集:R^2 = 0.838, $RMSE$ = 3.452, RPD = 2.470),对于贵州多地区 SOM 含量的监测更具备普适性。

参考文献:

[1] 贾文涛. 从土地整治向国土综合整治的转型发展[J]. 中国土地,2018(5):16-18.

[2] 张霄羽,姚艳敏,颜祥照. 中红外光谱土壤有机质含量估测研究进展[J]. 中国土壤与肥料,2021(4):327-336.

[3] 石朴杰,王世东,张合兵,等. 基于高光谱的复垦农田土壤有机质含量估测[J]. 土壤,2018,50(3):558-565.

[4] XIE R, XIAO H H. Application of remote sensing in the estimation of soil organic matter content[J]. Chemical Engineering Transactions,2018,66. DOI:10.3303/CET1866079.

[5] 向红英,柳维扬,彭 杰,等. 基于连续统去除法的南疆水稻土有机质含量预测[J]. 土壤,2016,48(2):389-394.

[6] 韩兆迎,朱西存,刘 庆,等. 黄河三角洲土壤有机质含量的高光谱反演[J]. 植物营养与肥料学报,2014,20(6):1545-1552.

[7] 勾宇轩,赵云泽,李 勇,等. 基于 CWT-sCARS 的东北旱作农田土壤有机质高光谱反演[J]. 农业机械学报,2022,53(3):331-337.

[8] 南 锋,朱洪芬,毕如田. 黄土高原煤矿区复垦农田土壤有机质含量的高光谱预测[J]. 中国农业科学,2016,49(11):2126-2135.
[9] NAWAR S, BUDDENBAUM H, HILL J, et al. Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy[J]. Soil and Tillage Research,2016,155:510-522.
[10] 郑文博. 基于遥感数据的乐安河沿岸土壤有机质含量反演[D]. 淮南:安徽理工大学,2021.
[11] 张娟娟,席 磊,杨向阳,等. 砂姜黑土有机质含量高光谱估测模型构建[J]. 农业工程学报,2020,36(17):135-141.
[12] 张 森,卢 霞,聂格格,等. SVM 和 BP 检测滨海湿地土壤有机质[J]. 光谱学与光谱分析,2020,40(2):556-561.
[13] 钟 亮,郭 熙,国佳欣,等. 基于不同卷积神经网络模型的红壤有机质高光谱估算[J]. 农业工程学报,2021,37(1):203-212.
[14] 武彦清,张 柏,宋开山,等. 松嫩平原土壤有机质含量高光谱反演研究[J]. 中国科学院研究生院学报,2011,28(2):187-194.
[15] 陆龙妹,张 平,卢宏亮,等. 淮北平原土壤高光谱特征及有机质含量预测[J]. 土壤,2019,51(2):374-380.
[16] 文锡梅,兰安军,易兴松,等. 基于高光谱的喀斯特地区典型农田土壤有机质含量反演[J]. 西南农业学报,2018,31(8):1649-1654.
[17] 王 芳,刘林峰,冉秋霞,等. 施用秸秆炭对黄壤上小白菜生长及镉吸收的影响[J]. 绿色科技,2021,23(16):15-18.
[18] 邸欣月,安显金,董 慧,等. 贵州喀斯特区域土壤有机质的分布与演化特征[J]. 地球与环境,2015,43(6):697-708.
[19] 郝冠军,黄懿珍,赵晓艺,等. 重铬酸钾外加热法测定土壤有机质的不确定度评定[J]. 上海农业学报,2011,27(3):103-109.
[20] 肖 艳,辛洪波,王 斌,等. 基于小波变换和连续投影算法的黑土有机质含量高光谱估测[J]. 国土资源遥感,2021,33(2):33-39.
[21] 郭云鹏,张 弓,侯至丞,等. 采用 SG 平滑滤波的 Stewart 平台主从控制研究[J]. 自动化仪表,2019,40(2):30-33,38.
[22] XU P F, JIA Y J, JIANG M X. Blind audio source separation based on a new system model and the Savitzky-Golay filter[J]. Journal of Electrical Engineering,2021,72(3):208-212.
[23] ZHAO L, HU Y, ZHOU W, et al. Estimation methods for soil mercury content using hyperspectral remote sensing[J]. Sustainability, 2018, 10(7): 2474.
[24] 高 颖,王延仓,顾晓鹤,等. 基于微分变换定量反演土壤有机质及全氮含量[J]. 江苏农业科学,2020,48(24):220-225.
[25] ClOUTIS E A. Hyperspectral geological remote sensing: evaluation of analytical techniques[J]. International Journal of Remote Sensing,1996,17(12):2215-2242.
[26] 徐明星,周生路,丁 卫,等. 苏北沿海滩涂地区土壤有机质含量的高光谱预测[J]. 农业工程学报,2011,27(2):219-223.
[27] 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.
[28] ZHAO H Q, ZHAO X S. Nonlinear unmixing of minerals based on the log and continuum removal model[J]. European Journal of Remote Sensing,2019,52(1):277-293.
[29] MEVIK B H, WEHRENS R. The PLS package: principal component and partial least squares regression in R[J]. Journal of Statistical Software,2007,18(2):1-23.
[30] 郤 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.
[31] 孙玉婷,杨红云,王映龙,等. 基于支持向量机的水稻叶面积测定[J]. 江苏农业学报,2018,34(5):1027-1035.
[32] 于 欢,刘 健,刘亚秋,等. 丘陵区耕地土壤有机质含量高光谱估测研究[J]. 山东农业大学学报(自然科学版),2021,52(4):648-653.
[33] 孟亚琼. 改进的 Adaboost 算法在基因表达数据中的应用[D]. 杭州:中国计量大学,2018.
[34] 江叶枫,郭 熙,叶英聪,等. 应用集成 BP 神经网络模型预测土壤有机质空间分布[J]. 江苏农业学报,2017,33(5):1044-1050.
[35] 刘 清,关榆君. 电梯群控系统节能优化调度控制[J]. 计算机仿真,2018,35(10):340-344.
[36] 孙浩然,赵志根,赵佳星,等. 珠海一号高光谱遥感的表层土壤有机质含量反演方法[J]. 遥感信息,2020,35(4):40-46.
[37] 于 雷,洪永胜,周 勇,等. 高光谱估算土壤有机质含量的波长变量筛选方法[J]. 农业工程报,2016,32(13):95-102.
[38] 陈 祯. 基于近红外光谱分析的土壤水分信息的提取与处理[D]. 武汉:华中科技大学,2010.
[39] 韩 陈,唐 强,韦 杰. 紫色土和黄壤含水率的室内光谱反演[J]. 水土保持通报,2021,41(5):174-180,190.
[40] 玉米提·买明,王雪梅. 连续小波变换的土壤有机质含量高光谱估测[J]. 光谱学与光谱分析,2022,42(4):1278-1284.
[41] 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.
[42] 奉国和. SVM 分类核函数及参数选择比较[J]. 计算机工程与应用,2011,47(3):123-124,128.
[43] 卢志宏,刘辛瑶,常书娟,等. 基于 BP 神经网络的草原矿区表层土壤 N/P 高光谱反演模型[J]. 草业科学,2018,35(9):2127-2136.

(责任编辑:徐 艳)