·农业信息与电气技术·

新疆艾比湖湿地土壤有机碳含量的光谱测定方法对比

杨爱霞，丁建丽※

（1. 新疆大学资源与环境科学学院，乌鲁木齐 830046； 2.新疆大学绿洲生态教育部重点实验室，乌鲁木齐 830046）

摘 要：干旱半干旱地区湿地土壤中的有机碳是影响土壤质量，制约植物生长的重要因素之一，其含量的变化会影响生态系统的安全和稳定。为快速估测湿地土壤有机碳含量，在新疆艾比湖湿地保护区采集 140 个荒漠土壤样品，利用土壤可见/近红外光谱数据以及化学分析获取的土壤有机碳数据，在对土壤原始光谱反射率进行卷积平滑的基础上，获取了一阶微分、倒数对数一阶微分 2 种光谱预处理指标，采用蚁群-区间偏最小二乘法、基于支持向量机的回归特征消去法，选择土壤有机碳含量近红外光谱特征波长，在此基础上构建土壤有机碳含量偏最小二乘回归、支持向量回归模型。结果表明：1）利用原始一阶微分建立的模型，预测能力优于倒数对数一阶微分建立的模型。2）4 种建模结果比较显示，利用原始一阶微分经基于支持向量机的回归特征消去法进行特征变量选择后建立的土壤有机碳含量模型，预测精度最高。训练集的相关系数以及均方根误差分别为 0.9687、0.158%；测试集的相关系数和均方根误差分别为 0.9091 以及 0.268%。因此，经过卷积平滑以及一阶微分预处理、并利用基于支持向量机的回归特征消去法建立的模型具有较高的预测精度和较好的稳健性，可以作为有效手段估算荒漠湿地土壤有机碳含量。

关键词：土壤；遥感；回归；艾比湖湿地

doi：10.11975/j.issn.1002-6819.2015.18.023

中图分类号：S127 **文献标志码**：A **文章编号**：1002-6819(2015)-18-0162-07

杨爱霞，丁建丽．新疆艾比湖湿地土壤有机碳含量的光谱测定方法对比[J]．农业工程学报，2015，31(18)：162－168.
doi：10.11975/j.issn.1002-6819.2015.18.023 http://www.tcsae.org

Yang Aixia, Ding Jianli. Comparative assessment of two methods for estimation of soil organic carbon content by Vis-NIR spectra in Xinjiang Ebinur Lake Wetland[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(18): 162－168. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2015.18.023 http://www.tcsae.org

0 引 言

湿地是陆地与水体间相互作用形成的独特生态系统，在碳的储存中起着重要作用[1]。尽管全球湿地面积仅占地球表面积的 6%，但其碳储量却占陆地生物圈碳储量的 35%，并且其单位面积碳储量在陆地各生态系统中居于首位[2]。湿地土壤碳储量的微小变动很可能会在很大程度上对大气 CO_2 浓度以及全球碳收支平衡产生影响[3]。受干旱环境背景控制的干旱区湿地，与荒漠基质有密切的生态过程联系，对人类活动的扰动非常敏感，一旦破坏很难恢复[4]。近 50 年来，人类活动日益频繁，改变了干旱区湿地的格局，对干旱区湿地碳收支平衡产生重要影响，进而影响到整个生态系统的健康运行和区域社会经济可持续发展。因此，及时掌握干旱区湿地土壤有机碳含量对干旱区湿地保护有重要意义，因而如何大范围、快速获取土壤有机碳的含量，成为一个重要的课题[5-6]。

传统获取土壤有机碳含量常用的化学分析方法主要有 $K_2M_nO_4$ 或 $K_2C_{r2}O_7$ 氧化法[7]和酸溶液提取法[8]，此种方法虽检测结果较准确，但测试样品的价格昂贵、前期样品预处理耗费时间，并且容易产生化学废料污染环境。可见光/近红外反射光谱替代传统化学分析方法监测和评估土壤质量，被认为是一个可行的方法[9]。可见光/近红外反射光谱是一种间接分析方法，通过经验模型的建立，利用复杂光谱数据实现土壤成分浓度分析。它的优势在于测量样品便捷快速、需要制备的东西少以及不需要使用危险化学品等。

近年来，可见光/近红外光谱分析方法越来越多地应用于预测土壤有机碳含量的研究。但在土壤类型方面，大多集中于有机碳含量较高的淋溶土、黑土[10-11]，对西北干旱、半干旱地区，土壤有机碳水平较低的湿地荒漠土壤的研究较少；另外，建模方面，多集中于多元线性回归、偏最小二乘回归等方法[12-13]，这些方法存在自相关、非线性以及过拟合现象[14]。近年来，机器学习方法较好地克服了这些问题，在土壤有机碳含量建模研究方面取得了不错的预测精度[15-16]。

基于此，本文以干旱区艾比湖湿地为研究区，对采集自湿地的 140 个土壤样品进行化验分析、光谱测量和处理，利用蚁群-区间偏最小二乘回归法、基于支持向量的回归特征消去法提取光谱特征变量后，建立艾比湖湿地土壤有机碳含量的回归预测模型，并对比 2 种模型的预测精度，以期找到适合干旱区湿地土壤有机碳的最优预测模型。

收稿日期：2015-07-08 修订日期：2015-08-28

基金项目：国家自然科学基金项目（U1303381，41261090，41130531）；新疆大学优秀博士研究生创新项目（XJUBSCX-2012026）。

作者简介：杨爱霞，女，河北赤城人，博士生，主要从事遥感应用研究。乌鲁木齐 新疆大学资源与环境科学学院，830046。

Email：yangaixia0310@126.com

※通信作者：丁建丽，男，山东成武人，教授，博士生导师，主要从事干旱区资源环境及遥感应用研究工作。乌鲁木齐 新疆大学资源与环境科学学院，830046。Email：watarid@xju.edu.cn

1　材料与方法

1.1　研究区概况

新疆艾比湖湿地是中国内陆干旱区湖泊湿地的典型，湿地地处天山北麓，准噶尔盆地的西南部，其南、西、北三面环山，东部与木特塔尔沙漠相连，地理坐标为 82°36′～83°50′E，44°30′～45°09′N，保护区总面积 2 670.85 km^2。该区年均降水量 90.9 mm，年均潜在蒸发量 3 400 mm，为典型的中温带大陆性干旱气候，独特的自然地理因素决定了其生态环境极其脆弱、对气候变化和人类活动的响应较为敏感。

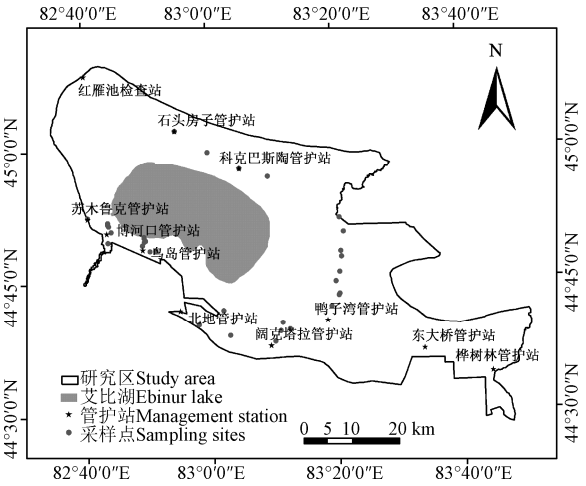


图 1　研究区位置以及采样点分布

Fig.1　Location of study area and distribution of sampling points

1.2　样本采集与制备

本研究的土壤样品采集自新疆艾比湖湿地保护区，土壤采集时间为 2012 年 10 月 1 日至 7 日，对 35 个样点的 4 层剖面土壤（0～5、>5～20、>20～40、>40～60 cm）进行五点呈梅花状采集，混合均匀后作为该样点的样品，一共采集样品 140 个。样品带回室内，自然风干，挑拣出根系等其他杂质，研磨过 2 mm 筛处理。磨碎后的每个样品分为两份，一份用作有机碳含量的化学分析，一份用作室内光谱分析。有机碳测定采用重铬酸钾容量—外加热法[17]。

1.3　光谱测定及预处理

采用美国 ASD 公司的 ASD FieldSpec®3 HR 光谱仪获取土壤光谱反射率数据，波段范围为 350～2 500 nm，光谱仪将数据重采样为 1 nm，因此每条光谱曲线包含有 2151 个波长变量。光谱测量在光源为 50 W 卤化灯，探头视场角为 25°的暗室内进行，入射角度 15°，探头距离测量样品表面的距离 10 cm，光源距离样品表面 50 cm[18]。将 140 个土壤样品分别装入直径 12 cm 和深 1.8 cm 的容器中，装满后将土样表面刮平。测量之前进行白板标定，每个样品重复测量 10 次，取光谱数据的均值作为土壤样品反射率光谱值，最终获取 140 个土壤样品原始反射率数据。

在采集土壤样品光谱时，常有高频随机噪声、基线漂移、光散射等噪声带入光谱值中，从而干扰光谱与样品化学组分的真实关系，进而影响定量估测模型的可靠性与准确性。Savitaky-Golay 平滑结合一阶微分预处理的光谱，可以消除基线漂移以及干扰产生的高频随机噪声的影响，也可以提高信噪比[19]。光谱的倒数对数变换可降低非线性和散射效应[20-21]。因此，本文首先对 140 个原始土壤光谱曲线进行 Savitaky-Golay 平滑（2 次多项式，5 个点）处理（处理后的光谱曲线如图 2 所示），之后进行一阶微分（A′）以及倒数对数一阶微分[lg(1/A)]′处理。光谱数据使用 Origin 9.0 软件进行处理，其余方法的建模和验证以及作图在 Matlab R2013a 和 Microsoft Visio 中进行。

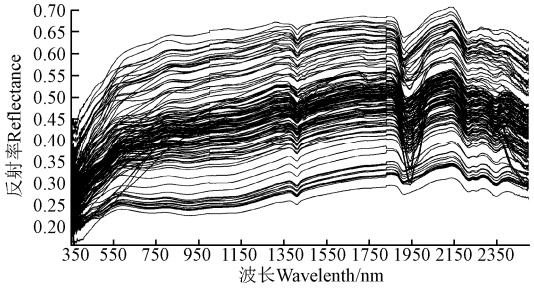


图 2　平滑后的土壤样本光谱反射率

Fig.2　Spectral reflectance of soil samples after smoothing

1.4　变量选择和预测模型

本研究用到 2 种模型，分别是蚁群-区间偏最小二乘回归法（ant colony optimization-interval partial least square, ACO-iPLS）、基于支持向量机的回归特征消去法（recursive feature elimination based on support vector machine, SVM-RFE）模型。

1.4.1　ACO-iPLS 变量选择与建模

PLS 算法借鉴了相关分析和主成分分析的思想，很好地解决了自变量间的多重共线性问题；另外，也较好解决了样本数少于变量数的问题，但此方法不能对变量进行选择。而 ACO 算法是一种模仿蚂蚁觅食的优化算法，蚂蚁觅食会在路径上留下信息素，蚁群会朝着信息素含量大的方向进行寻址，从而增强路径的选择，蚂蚁最终依据信息素最大的路径进行觅食。结合 ACO 算法和 PLS 算法可以有效地获取最佳的光谱区间，具体算法过程详见文献[22]。在 Matlab R2013a 编程环境中，进行 ACO-iPLS 算法的设计和验证，ACO 算法的参数设置为：初始群体大小为 50，最大迭代次数为 50 次，最大循环次数 20 次，变量选择概率阈值为 0.3，显著性因子 Q=0.01。信息素衰减系数设定为 0.53，关于 PLS 模型，最大变量数初始设定为 15。

1.4.2　SVM-RFE 变量选择与 SVM 建模

SVM 是一种机器学习模型，将非线性可分样本数据通过核函数映射到高维线性可分空间，之后用优化法求解超平面，确定决策函数参数，使其结构风险最小化。SVM-RFE 方法利用排序策略进行变量选择，属于贪心算法。RFE 方法从全集开始，设定一个特征排序规则，逐步消除相关性最差的特征，从而得到特征的排序。根据 RFE 算法，相关性低的特征，最先被消去，因此排在列表的最

后面，相反，相关性最高的特征，最后被消除，排在列表最前面。其算法详见参考文献[23]。在 Matlab R2013a 编程环境中，使用 libsvm-3.1-[FarutoUltimate3.1Mcode]工具包，然后利用二维网格搜索算法进行参数优选，通过多次训练，最终确定采用 epsilon-SVR 模型，核函数类型选用 Sigmoid，模型的参数 Gamma 值为 0.0039，Eps 值为 0.01，C 值为 1。

1.5 模型检验

2 种算法都利用十折交叉验证法，计算交叉验证相关系数（correlation coefficient of cross validation，R_{cv}）、交叉验证均方根误差（root mean squares error of cross validation，RMSECV）来优化建模参数。模型的检验通过相关系数（correlation coefficient，R）、均方根误差（root mean squared error，RMSE）及相对分析误差来（relative prediction deviation，RPD）衡量。当 RPD≥2 时，模型预测的精度极佳；当 1.4≤RPD<2 时，预测的精度尚可；当 RPD<1.4 时，预测精度极差[24]。另外，1∶1 线指由到 y 轴、x 轴距离相等的点组成的对角线，当模型检验指标差异较小时，可通过实测值、估算值构成的点所偏离 1∶1 线的程度来估测模型精度[25]。

2 结果与分析

2.1 土壤有机碳描述性统计分析

采用 Kennard-Stone（K-S）算法划分训练集和测试集，选取 70 个样本作为训练集，用来构建模型，70 个样本作为测试集，用来检测所建模型的预测效果。通过计算各个样品土壤有机碳含量值之间的欧氏距离，选择样品集中最具代表性的样品作为训练集。

经 K-S 划分出的训练集和测试集的土壤有机碳含量统计结果如表 1 所示。从表 1 中可以看出，训练集中有机碳质量分数最大值是 2.9723%，最小值是 0.0241%，平均值为 0.5050%，标准差 0.6421%；测试集中有机碳质量分数最大值是 3.4283%，最小值是 0.0094%，平均值为 0.4008%，标准差 0.6461%，总体表明训练集和测试集的抽取随机性好，所建模型普适性好。

表 1 土壤有机碳含量描述性统计分析

Table 1 Descriptive statistics of soil organic matter content

模型 Models	样本数 Sample number	最小值 Min./%	最大值 Max./%	均值 Mean/%	标准差 Standard deviation/%
训练集 Training sets	70	0.0241	2.9723	0.5050	0.6421
测试集 Testing sets	70	0.0094	3.4283	0.4008	0.6461

2.2 ACO-iPLS 和 SVM-RFE 变量选择及建模结果

由于土壤高光谱数据一般拥有数百乃至上千变量，这当中存有一部分变量，包含与观测样品无关信息，利用这些变量建模，不但干扰模型建立，还影响模型精度及可靠性。因此，对土壤有机碳含量分析前，首先利用 ACO-iPLS 和 SVM-RFE 两种方法分别对原始一阶微分、倒数对数一阶微分进行特征变量的选取，在此基础上再建立预测模型。

两种方法入选波段以及建模结果如表 2，从表 2 中可以看出：利用 ACO-iPLS 方法通过特征变量选取之后建立的原始一阶微分模型最优，R_{cv} 达到了 0.8647，RMSECV 仅为 0.329%，预测 R_p 达到了 0.8297，RMSEP 仅为 0.396%，RPD 为 1.63，模型能够较好地对土壤样品进行预估；倒数对数一阶微分的 RPD 为 1.10，模型不能够对样品进行预估。利用 SVM-RFE 方法通过特征波长选取之后建立的原始一阶微分模型最优，R_{cv} 达到了 0.9687，RMSECV 仅为 0.158%，预测 R_p 达到了 0.9091，RMSEP 仅为 0.268 %，RPD 为 2.41，模型能够对样品很好地进行预估；倒数对数一阶微分的 RPD 为 1.44，模型能够对样品进行预估。

表 2 基于 ACO-iPLS 和 SVM-RFE 的波段选取建模及验证

Table 2 Selected feature wavelengths and training sets and testing sets results by ACO-iPLS and SVM-RFE methods

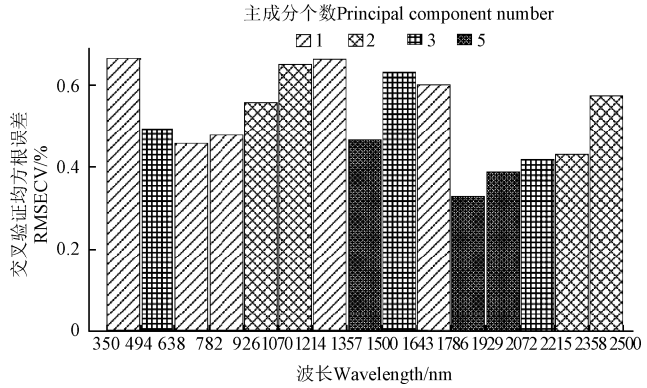
建模方法 Modeling methods	预处理 Pre-Processing	入选波长 Selected wavelengths /nm	训练集 Training sets		测试集 Testing sets		
			R_{cv}	RMSECV/%	R_p	RMSEP/%	RPD
ACO-iPLS	A′	1786～1929	0.8647	0.329	0.8297	0.396	1.63
	[lg(1/A)]′	1786～1929	0.7293	0.496	0.8243	0.586	1.10
SVM-RFE	A′	780,1911,783,779,768,759,793,794, 2254,910,1677,1912,2089,745,825, 2088,746,2090,1913,1751	0.9687	0.158	0.9091	0.268	2.41
	[lg(1/A)]′	706,736,731,1943,779,721,413,510, 704,397,732,1944,1085,2091,2347, 881,2422,1966,2257,2111	0.9989	0.033	0.8111	0.448	1.44

注：A′为原始一阶微分，[lg(1/A)]′为倒数对数一阶微分，R_{cv} 为交叉验证相关系数，RMSECV 为交叉验证均方根误差，R_p 为测试集相关系数，RMSECP 为测试集均方根误差，RPD 为相对分析误差，下同。

Note: A′ represents first derivative with reflectance, [lg(1/A)]′ represents logarithm of inversed first derivative, R_{cv} represents correlation coefficient of cross validation, RMSECV represents root mean squares error of cross validation, R_p represents correlation coefficient of the testing set, RMSECP represents root mean square error of prediction, RPD represents relative prediction deviation, the same below.

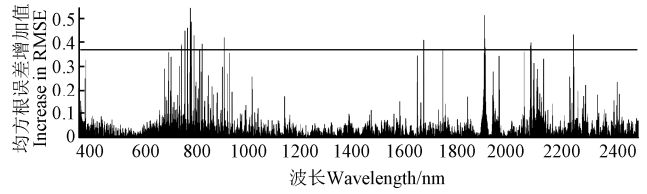
2.3 不同方法变量选择及建模结果对比

结合前文分析所得结果，原始一阶微分变换的数据建模效果最好，因此只对原始一阶微分数据进行 2 种方法特征波长选取以及建模方法的对比分析。图 3、图 4 分别为原始一阶微分变换下不同算法的特征波长选取结果。表 3 是不同算法在原始一阶微分变换下的建模结果比较。图 5 是不同算法的原始一阶微分训练集和测试集的拟合结果。



注：RMSECV 为交叉验证均方根误差，下同。
Note: RMSECV represents root mean squares error of cross validation, the same below.

图 3　基于原始一阶微分的 ACO-iPLS 波长区间的选择
Fig.3　Selected spectral interval by ACO-iPLS with first derivative spectra



注：横线以上为选取的特征波长.
Note: selected wavelengths above the line.

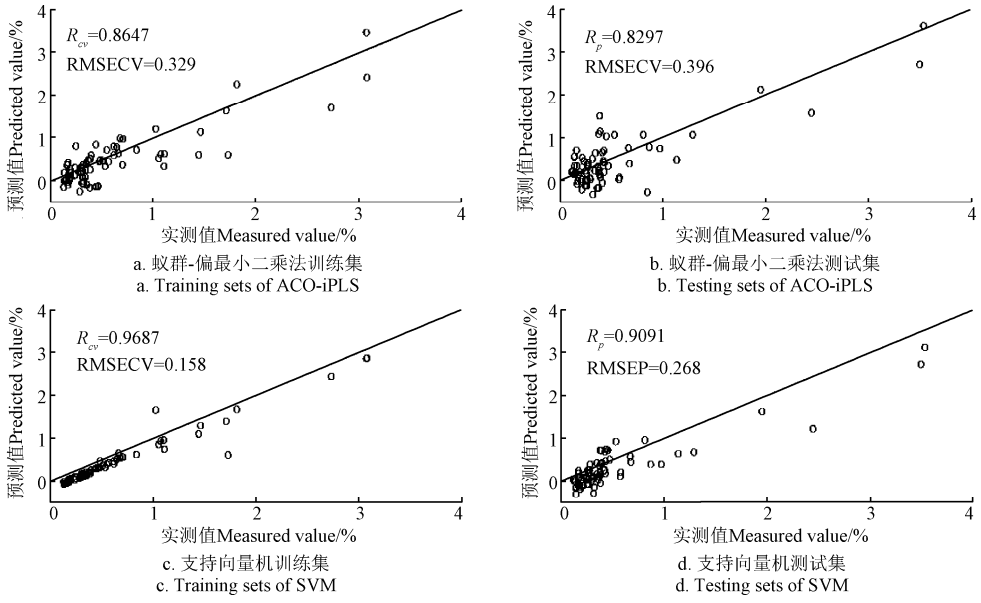
图 4　基于原始一阶微分的 SVM-RFE 波长选取
Fig.4　Selected wavelengths by SVE-RFE first derivative spectra

表 3　不同算法建立的土壤有机碳含量估测模型及验证
Table 3　Comparison of results by different models

建模方法 Modeling methods	训练集 Training sets		测试集 Testing sets		相对分析误差 RPD
	交叉验证相关系数 R_{cv}	交叉验证均方根误差 RMSECV/%	相关系数 R_p	均方根误差 RMSEP/%	
AOC-iPLS	0.8647	0.329	0.8297	0.396	1.63
SVM-RFE	0.9687	0.158	0.9091	0.268	2.41

从图 3 中可以看出，ACO-iPLS 方法把整个波长区间（350～2 500 nm）划分为 15 个子区间，其中第 11 个子区间的交叉验证均方根误差 RMSECV 最小，因此最佳的选择区间位于：1 786～1 929 nm。从图 4 中可以看出，利用 SVM-RFE 进行特征变量的选择，通过均方根误差增加值选取 20 个特征波长（尝试选取 30 个或更多波长，也大致位于这些区间），这些波长主要集中在 745～910 nm，1 677 nm，1 755 nm，1 911～2 254 nm 附近，因此最佳特征位于这些区间。

从表 3 中可以看出，两种建模方法中，基于支持向量机的回归特征消去法模型的 RPD 超过了 2.0，高于蚁群-区间偏最小二乘法。此外，图 5 是利用两种方法对原始一阶微分数据的校正和验证的拟合结果，从图中可以看出支持向量机的样本点较为均匀地分布在 1∶1 线的两侧，预测效果较好，这与上面的结果一致。



注：R_{cv} 为交叉验证相关系数，R_p 为测试集相关系数，RMSECP 为测试集均方根误差。
Note: R_{cv} represents correlation coefficient of cross validation, R_p represents correlation coefficient of the testing set, RMSECP represents root mean square error of prediction.

图 5　不同算法的原始一阶微分训练集和测试集的土壤有机碳含量
Fig.5　Comparison of measured and estimated soil organic matter content by different models

3　讨　论

建立土壤有机碳含量的定量估测模型是近红外光谱分析的关键技术，但选取合适的预处理方法和建模方法至关重要。本文在采集 140 个土壤样品的基础上，利用 2 种光谱变换形式，对蚁群-区间偏最小二乘回归法、基于支持向量机的回归特征消去法进行特征变量选取并建模。

3.1　不同预处理的建模效果对比

光谱预处理能较好地消除土壤类型及所处环境等因素的影响，从而突出光谱反射率与有机碳含量之间的相关性。不同变换方法对建模结果有较大影响。在基于光谱的土壤有机碳含量估测研究中，大多研究表明，通过

预处理方法可以提高预测的精度，如一阶微分、倒数对数一阶微分、对数一阶微分等。本研究结果显示，在两种建模方法中，原始一阶微分的精度都是最高的，高于倒数对数一阶微分。Zornoza 等[26]在研究中也显示，采用原始一阶微分建立的有机碳模型，取得较好的结果，决定系数分别达到 0.95 和 0.98，这与本文研究结论一致。

3.2 SVM-RFE 与 ACO-iPLS 特征变量选取

特征选择，是可见光-近红外光谱研究至关重要的一步，一方面可以简化模型，更主要的是不相关变量的剔除，得到预测能力强、稳健性好的校正模型。

从表 3 可知，基于特征选取 SVM-RFE 的 20 个波段建立的模型预测能力优于 ACO-iPLS 特征选取建立的模型，说明 SVM-RFE 可能是土壤有机碳光谱有效的提取波长方法。本研究利用 SVM-RFE 提取到的干旱区土壤有机碳的特征波段集中于 745～910 nm 和 1911～2254 nm 这两个区间，这与纪文君等[27]和彭杰等[28]选择的土壤有机碳含量的敏感区域（600～800 nm 和 570～630 nm）有一定差异，这可能和研究的土壤质地、采样时间等不同造成的土壤光谱反射率差异有关，还可能是变量选择方法不同导致选择的特征波段不同。从 SVM-RFE 特征选择的原理可知，该方法不但能选出相关特征还能去除冗余特征，而相关法选出的特征，特征之间可能存在冗余。

3.3 SVM-RFE 与 ACO-iPLS 模型精度的比较

从表 3 可以看出，两种建模方法中，基于支持向量机的回归特征消去法模型建模结果最好，高于蚁群-区间偏最小二乘回归法。主要原因是蚁群区间偏最小二乘法中，蚁群算法是生物智能算法，模仿生物的行为建立优化算法，这种算法在执行过程中容易出现停滞现象、并且当问题规模较大时存在陷入局部最优的可能性，虽然本文使用 ACO-iPLS 选择了最佳的子区间波长，但是有可能在其他区间也存在少量的波长适合建模，因此这种算法的主要缺陷是会遗漏部分次优的波长，假设增加子区间的划分数目（本文尝试将 15 个子区间改为为 20，甚至 30 个），虽然可以部分解决这个问题，但是不能从根本上解决。

SVM 算法属于结构经验风险模型，用优化方法求解划分超平面，即确定决策函数的参数。由于这种算法，在训练集中为了保证结果风险最小化，允许存在部分误差，并且对这部分误差进行一定的惩罚，来保证这种算法在测试集中能够达到非常高的精度，本文也验证了该结论，其预测精度 RPD 高达 2.41，高于 ACO-iPLS 的 1.63。

本研究运用线性模型偏最小二乘和非线性模型支持向量机在实验室内对干旱区湿地土壤有机碳含量的测定，取得了较为满意的结果。但影响土壤反射率的因素除了有机碳含量以外还有土壤颜色、土壤粒径、粗糙度等[29]，此外，还需考虑土壤水分含量等因素。因此，在未来的工作中，需将研究重点由室内控制试验转为野外原位测定，进一步开展干旱区湿地土壤野外反射光谱的测量以及考虑更多外界因素对其影响。采集更多地区和不同类型的样品进行研究，以优化干旱区湿地土壤有机碳含量预测模型的精度。

4 结 论

本研究以干旱区艾比湖湿地为研究区，利用采集的 140 个土壤样品的室内实测光谱和土壤有机碳为数据源，基于原始一阶微分和倒数对数一阶微分 2 种光谱预处理指标，通过蚁群-区间偏最小二乘法（ant colony optimization-interval partial least square，ACO-iPLS）和支持向量机的回归特征消去法（recursive feature elimination based on support vector machine，SVM-RFE）两种算法提取特征变量，建立了 4 种预测模型，分别预测了土壤有机碳含量，得出以下主要结论：

1）原始一阶微分和倒数对数一阶微分的建模结果对比发现，经原始一阶微分预处理建立的偏最小二乘和支持向量机模型，建模精度均表现出高于倒数对数一阶微分预处理，原始一阶微分建立的偏最小二乘和支持向量机模型的相对分析误差分别为 1.63 和 2.41，倒数对数一阶微分建立的偏最小二乘和支持向量机模型的相对分析误差分别为 1.10 和 1.44。

2）对比 2 种特征变量选择方法，利用原始一阶微分和倒数对数一阶微分 2 种预处理方式，基于支持向量机的回归特征消去法（SVM-RFE）选择的 20 个变量建立的模型，建模精度优于蚁群-区间偏最小二乘法（ACO-iPLS），以原始一阶微分的预测精度最为突出，其相对分析误差为 2.41，其反演结果可以很好地估算该区域土壤有机碳含量。因此，采用 SVM-RFE 方法代替相关分析、逐步回归提取特征变量，并建立预测模型的方法，为干旱区湿地土壤有机碳光谱特征选择提供了新思路。

3）利用 2 种特征变量选择的原始一阶微分建立的两种模型的相对分析误差都高于 1.4，均可以用来进行干旱区湿地土壤有机碳含量的预测，但预测精度不同。支持向量回归（support vector regression, SVR）模型的相对分析误差精度高于偏最小二乘回归（partial least squares regression, PLSR）。SVR 模型测试集的相关系数达到了 0.9091，测试集的均方根误差仅为 0.268%，相对分析误差为 2.41；PLSR 模型预测集的相关系数达到了 0.8297，测试集的均方根误差仅为 0.396%，相对分析误差为 1.63。可以看出，与相对复杂的机器学习方法相比，传统的线性模型也取得了不错的建模结果。

[参 考 文 献]

[1] 徐欢欢，曾从盛，王维奇，等. 艾比湖湿地土壤有机碳垂直分布特征及其影响因子分析[J]. 福建师范大学学报：自然科学版，2010，26(5)：86－91.
Xu Huanhuan, Zeng Congsheng, Wang Weiqi, et al. Study on vertical distribution and the influencing factors of soil organic carbon in Ebinur Lake Wetland[J]. Journal of Fujian Normal University: Natural Science Edition, 2010, 26(5): 86－91. (in Chinese with English abstract)

[2] IPCC (International Panel of Climate Change). Land use, land-use change, and forestry[R]. Cambridge: Cambridge University Press, 2000.

[3] Wang D D, Chakraborty S, Weindorf D C, et al. Synthesized use of VisNIR DRS and PXRF for soil characterization:

Total carbon and total nitrogen[J]. Geoderma, 2015(243/244): 157－167.

[4] Zhao Ruifeng, Chen Yaning, Zhou Huarong, et al. Assessment of wetland frag-mentation in the Tarim River basin, western China[J]. Environmental Geology, 2009, 57(2): 455－464.

[5] Stevens A, Udelhoven T, Denis A, et al. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy[J]. Geoderma, 2010, 158(1/2): 32－45.

[6] Vohland M, Besold J, Hill J, et al. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy[J]. Geoderma, 2011, 166(1): 198－205.

[7] Logninow W, Wisniewski W, Strony W, et al. Fractionation of organic carbon based on susceptibility to oxidation[J]. Polish Journal of Soil Science, 1987, 20(1): 47－52.

[8] Polglase P J, Jokela E J, Comerford N B. Phosphorus, nitrogen, and carbon fractions in litter and soil of southern Pine Plantations[J]. Soilence Society of America Journal, 1992, 56(2): 566－572.

[9] Kinoshita R, Moebius-Clune B N, van Es H M, et al. Strategies for soil quality assessment using visible and near-infrared reflectance spectroscopy in a western Kenya chronosequence[J]. Soil Science Society of America Journal, 2012, 76(5): 1776－1788.

[10] Summers D, Lewis M, Ostendorf B, et al. Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties[J]. Ecological Indicators, 2011, 11(1): 123－131.

[11] Araújo S R, Söderström M, Eriksson J, et al. Determining soil properties in Amazonian Dark Earths by reflectance spectroscopy[J]. Geoderma, 2015(237/238): 308－317.

[12] Liu Huanjun, Zhang Yuanzhi, Zhang Bai, et al. Novel hyperspectral reflectance models for estimating black-soil organic matter in Northeast China. Environmental Monitoring and Assessment[J]. Environmental Monitoring and Assessment, 2009, 154(1/2/3/4): 147－154.

[13] Luce M S, Ziadi N, Zebarth B J, et al. Rapid determination of soil organic matter quality indicators using visible near infrared reflectance spectroscopy[J]. Geoderma, 2014(232/233/234): 449－458.

[14] Drake J M, Randin C, Guisan A. Modelling ecological niches with support vector machines[J]. Journal of Applied Ecology, 2006, 43(3): 424－432.

[15] Peng Xiaoting, Shi Tiezhu, Song Aihong, et al. Estimating soil organic carbon using Vis/NIR spectroscopy with SVMR and SPA methods[J]. Remote Sensing, 2014, 6(4): 2699－2717.

[16] Were K, Bui D T, Dick Ø B, et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape[J]. Ecological Indicators, 2015, 52: 394－403.

[17] 鲍士旦. 土壤农化分析. 第 3 版. 北京：中国农业出版社，1999.

[18] 张东，塔西甫拉提·特依拜，张飞，等. 分数阶微分在盐渍土高光谱数据预处理中的应用[J]. 农业工程学报，2014，30(24)：151－160.
Zhang Dong, Tashpolat Tiyip, Zhang Fei, et al. Application of fractional differential in preprocessing hyperspectral data of saline soil[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2014, 30(24): 151－160. (in Chinese with English abstract)

[19] Cao Pu, Pan Tao, Chen Xingdan. Choice of wave band in design of mini type near infrared corn protein content analyzer[J]. Optics and Precision Engineering, 2007, 15(12): 1953－1957.

[20] Kemper T, Sommer S. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy[J]. Environ. Sci. Technol. 2002, 36(12): 2742－2747.

[21] Gomez C, Lagacherie P, Coulouma G. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements[J]. Geoderma, 2008, 148(2): 141－148.

[22] Huang Xiaowei, Zou Xiaobo, Zhao Jiewen, et al. Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with ant colony optimization models[J]. Food chemistry, 2014, 164: 536－543.

[23] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine learning, 2002, 46(1/2/3): 389－422.

[24] 栾福明，张小雷，熊黑钢，等. 基于不同模型的土壤有机质含量高光谱反演比较分析[J]. 光谱学与光谱分析，2013，33(1)：196－200.
Luan Fuming, Zhang Xiaolei, Xiong Heigang, et al. Comparative analysis of soil organic matter content based on different hyperspectral inversion models[J]. Spectroscopy and Spectral Analysis, 2013, 33(1): 196－200. (in Chinese with English abstract)

[25] 高志海，白黎娜，王琣瑜，等. 荒漠化土地土壤有机质含量的实测光谱估算[J]. 林业科学，2011，47(6)：9－16.
Gao Zhihai, Bai Lina, Wang Bengyu, et al. Estimation of soil organic matter content in desertified lands using measured soil spectral data[J]. Scientia silvae sinicae, 2011, 47(6): 9－16. (in Chinese with English abstract)

[26] Zornoza R, Guerrero C, Mataix-Solera J, et al. Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils[J]. Soil Biology and Biochemistry, 2008, 40(7): 1923－1930.

[27] 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.
Ji Wenjun, Shi Zhou, Zhou Qing, et al. VIS-NIR reflectance spectroscopy of the organic matter in several types of soils[J]. J. Infrared Millim, Waves, 2012, 31(3): 277－282. (in Chinese with English abstract)

[28] 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(3)：517－524.
Peng Jie, Zhou Qing, Zhang Yangzhu, et al. Effect of soil organic matter on spectral characteristics of soil[J]. Acta Pedologica Sinica, 2013, 50(3): 517－524. (in Chinese with English abstract)

[29] 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算的影响[J]. 农业工程学报，2015，31(9)：114－120.
Si Haiqing, Yao Yanmin, Wang Deying, et al. Hyperspectral prediction of soil organic matter contents under different soil moisture contents[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(9): 141－120. (in Chinese with English abstract)

Comparative assessment of two methods for estimation of soil organic carbon content by Vis-NIR spectra in Xinjiang Ebinur Lake Wetland

Yang Aixia, Ding Jianli※

(1. *College of Resource and Environment Sciences*, *Xinjiang University*, *Urumqi* 830046, *China*;
2. *Key Laboratory of Oasis Ecology Ministry of Education*, *Xinjiang University*, *Urumqi* 830046, *China*)

Abstract: Soil organic carbon (SOC) is a critical soil property that has profound impact on soil quality and plant growth. It is involved in soil structural formation and atmospheric carbon sequestration. This is especially true in the arid and semi-arid regions. Accurately detecting SOC is an important issue. Traditionally, SOC is limited to laboratory determination using the techniques such as wet or dry combustion, ion sensing electrodes, loss on ignition, or via chemical assays. Yet those traditional approaches often involve expensive testing materials, time-consuming sample preparation and production of excessive environmental pollutants. An approach which can quantify SOC content with time and cost savings is needed. With 140 soil samples acquired from the Ebinur Lake wetland protection area in Xinjiang, China, this research attempts to apply 2 algorithms in hyperspectral data mining, namely, the ant colony optimization – interval partial least squares (ACO-iPLS) and recursive feature elimination – support vector machine (SVM-RFE) to improve the estimation accuracy of SOC content using the visible and near-infrared (VIS/NIR) spectroscopy of soils (350-2500 nm) in laboratory. After convolution smoothing (S-G), 2 common spectra pre-processing methods, namely, first order differential and first order differential of the logarithm of inverse, are applied in the hyperspectral data to extract the feature wavelengths. Results indicate that the feature wavelengths pertaining to SOC mainly are located within 1786-1929 nm with ACO-iPLS and 745-910, 1677, 1755, and 1911-2254 nm with SVM-RFE. With the extracted feature wavelengths, the ensuing models with the same 2 approaches are established with the half of the samples (70 soil samples) as training set and the other half (70 soil samples) as testing set. The results show that the spectra processed with the combination of the S-G and first order with reflectance perform much better than the logarithm of first order differential of the logarithm of inverse after the S-G. Compared to the linear model used commonly, i.e. ACO-iPLS, the nonlinear model SVM-RFE pre-processed with first order differential with reflectance produces the higher estimation accuracy. The root mean square error of cross validation (RMSECV) and the root mean square error of prediction (RMSEP) for the SVM-RFE approach are respectively 0.158% and 0.268% in the training and testing set. The correlation coefficient of cross validation (Rcv) and the correlation coefficient of prediction (Rp) are 0.9687 and 0.9091, respectively. The relative prediction deviation (RPD) of testing set is 2.41. The RMSECV and RMSEP for the ACO-iPLS approach are respectively 0.329% and 0.396% in the training and testing set. The Rcv and Rp are 0.8647 and 0.8297, respectively. The RPD of the testing set is 1.63. The SVM-RFE approach pre-processed with first order differential of the logarithm of inverse produces the higher estimation accuracy than the ACO-iPLS. The RMSECV and RMSEP for the SVM-RFE approach are 0.033% and 0.448%, respectively. The Rcv and Rp are 0.9989 and 0.8111, respectively. The RPD of testing set is 1.44. The RMSECV and RMSEP for the ACO-iPLS approach are 0.496% and 0.586%, respectively. The Rcv and Rp are 0.7293 and 0.586, respectively. The RPD of the testing set is 1.10. Over all, the good performance of the SVM model can be ascribed to its good capability of dealing with non-linear and hierarchical relationship between SOC and feature wavelengths. The results are fairly satisfactory. This practice provides an efficient, low-cost, potentially highly accurate approach to estimate SOC content and hence support better management and protection strategies for desert wetland ecosystems. The next step is to attempt to apply VIS/NIR spectroscopy technique in the field for further research.

Key words: soils; remote sensing; regression analysis; Ebinur Lake wetland