基于机器学习算法的湖滨绿洲土壤电导率高光谱估算模型

孟 珊[1, 2], 李新国[1, 2*], 焦 黎[1, 2]
（1. 新疆师范大学地理科学与旅游学院，乌鲁木齐 830054；2. 新疆干旱区湖泊环境与资源实验室，乌鲁木齐 830054）

摘 要：【目的】为湖滨绿洲土壤高光谱估算土壤电导率值提供方法支持，实现区域土壤盐分快速估测。**【方法】**利用实测的土壤电导率值与土壤高光谱数据联合分析，采用竞争自适应重加权采样（CARS）、连续投影算法（SPA）、遗传算法（GA）筛选土壤电导率的特征波段，并基于全波段及特征波段构建 BP 神经网络（BPNN）、支持向量机（SVM）、极限学习机（ELM）三种机器学习算法模型，引入偏最小二乘模型（PLSR）进行对照，比较其模型精度。**【结果】**研究区土壤电导率值变化范围 0.02 ~ 17.22 mS cm^{-1}，平均值为 2.61 mS cm^{-1}，变异系数为 134.87%，呈现强变异性；CARS、SPA、GA 算法筛选的特征波段将建模输入量分别压缩至全波段数量的 0.87%、1.68%、0.70%，减少建模输入量，提升建模速率，变量方法的选择 CARS > SPA > GA；三种机器学习算法模型均优于 PLSR 模型，决定系数（R^2）平均增加 20.57%，相对分析误差（RPD）平均增加 17.84%，土壤电导率高光谱估算模型以 CARS-SVM 最优，训练集与验证集 R^2 分别为 0.76 和 0.75，RMSE 分别为 1.79 和 1.68 mS cm^{-1}，RPD 分别为 2.04 和 2.00。土层深度 20 ~ 30 cm 的土壤电导率高光谱估算模型精度最高，训练集与验证集 R^2 分别为 0.83 和 0.84，RMSE 分别 1.37 和 1.77 mS cm^{-1}，RPD 分别为 2.41 和 2.50。**【结论】**基于 CARS-SVM 的土壤电导率高光谱估算模型精度高，估算能力最优，可以为湖滨绿洲土壤电导率估算提供科学参考。

关 键 词：土壤电导率值；竞争自适应重加权采样；连续投影算法；遗传算法；机器学习算法；高光谱估算模型
中图分类号：S151.9 **文献标识码：**A **文章编号：**0564−3945(2023)02−0286−09

DOI: 10.19336/j.cnki.trtb.2022011003
孟 珊, 李新国, 焦 黎. 基于机器学习算法的湖滨绿洲土壤电导率高光谱估算模型 [J]. 土壤通报, 2023, 54(2): 286 − 294
MENG Shan, LI Xin-guo, JIAO Li. Hyperspectral Estimation Model of Soil Conductivity in the Lakeside Oasis Based on Machine Learning Algorithm[J]. Chinese Journal of Soil Science, 2023, 54(2): 286 − 294

【研究意义】土壤电导率是衡量土壤盐渍化的重要指标，土壤电导率值的准确估算对于掌握区域土壤的盐渍化程度，开展区域盐渍化防治与调控，生态环境的保护以及精细农业的可持续发展都具有重要作用[1−2]。高光谱遥感技术可以快速、准确、高效、充分的挖掘光谱信息实现动态监测，土壤的多种属性信息均可用连续的光谱曲线进行综合反映，以此构建高精度的土壤属性光谱模型[3−5]。机器学习算法在土壤属性定量模拟中，可提高建模速率，提升建模精度，相较于传统的统计回归模型，模型性能更为优越[6−7]。**【前人研究进展】**目前，有关土壤属性的高光谱定量估算，学者们已进行多方面研究。王涛等[8]研究发现去包络线处理结合连续投影算法（SPA）筛选光谱特征波段可以实现土壤电导率快速检测，SPA 算法具有较强的特征波长选择能力，且能够最大程度避免光谱波段信息的重叠。唐海涛等[9]应用竞争自适应重加权采样算法（CARS）对不同类型的土壤有机质进行特征波段筛选，极大程度减少建模输入量，降低计算的复杂程度及变量维度，有效选择与土壤属性相关的最优波长组合。于雷等[10]探究土壤有机质的高光谱波长变量筛选，研究发现单个特征变量筛选方法 CARS 算法优于 SPA 算法。乔天等[11]提出遗传算法（GA）可减少信息冗余及处理共线性问题，GA-PLS 模型能有效去除光谱数据的冗余信息，减少建模所用的变量数目，有效提高模型精度。亚森江·喀哈尔等[12]基于分数阶微分方法对光谱指数进行优化，构建土壤电导率偏最小二乘模型（PLSR）高光谱估算模型。赵慧等[13]在分数阶微分方法的基础上，利用 PLSR 和支持向量机（SVM）分别构建土壤电导率高光谱估算模型，结果表明 SVM 估算模型效果更好。王雪梅等[14]对干旱区绿洲耕层重金属进行高光谱估算发现 PLSR 模型精度低

收稿日期：2022−01−10；**修订日期：**2022−04−20
基金项目：新疆维吾尔自治区自然科学基金项目（2022D01A214）；新疆维吾尔自治区重点实验室开放课题（2018D04026）
作者简介：孟 珊（1997−），女，安徽固镇，硕士研究生，主要从事土壤资源变化及其遥感应用研究。Email: mengshan1997@163.com
*通讯作者：Email: onlinelxg@163.com.

于 BP 神经网络（BPNN）模型。田安红等[15] 对阜康市盐渍土的 Na^+ 含量进行高光谱估算，模型估算能力 BPNN 优于 PLSR 优于逐步多元线性回归（SMLR）。Rocha Neto[16] 利用极限学习机（ELM）、普通最小二乘（OLS）、PLSR 与多层感知器（MLP）对巴西半干旱区土壤电导率进行估算，评价土壤盐渍化状况，线性模型和 ELM 的估算能力优于 MLP 的能力。Bao[17] 等对 PLS 与 PLS-SVM 两种建模方法进行比较，结果表明非线性模型来估算光谱与土壤养分含量的精度要高于线性模型。蔡亮红等[18] 对于土壤含水量的研究认为 ELM 模型对非线性问题具有较强解析能力，并且模型的稳健性更好。曹肖奕等[19] 基于光谱指数构建机器学习算法估算土壤电导率，发现机器学习算法建模精度更高，模型估算精度 ELM > SVM > BPNN，可以较好处理非线性问题。

【本研究切入点】PLSR 方法在土壤光谱方面的研究已非常普遍，PLSR 能够解决变量间多重共线性问题，然而只能对一些特定的土壤属性与其光谱之间的线性关系进行模拟，且土壤性质也并非标准正态分布[20]。光谱数据测试的环境以及测试仪器的状态变化都会引起光谱的非线性变化，则以 PLSR 的线性回归方法可能并不适用于处理非线性问题，机器学习算法构建的模型能够得出更理想的估算结果，在估算能力上相较于 PLSR 模型可以表现出更好的稳定性，具有较强的泛化能力，选择非线性的机器学习算法估算土壤电导率，有望得到更好的估算结果[21−22]。高光谱数据的全波段数目较多，易造成光谱信息冗余，应用波长变量筛选方法可以有效减少建模输入量，降低计算量，提高建模速率[10]。现阶段利用不同波长变量筛选方法并结合机器学习算法构建土壤电导率估算模型鲜有报道。**【拟解决的问题】**以博斯腾湖西岸湖滨绿洲为研究区，应用 CARS、SPA、GA 三种算法筛选光谱特征波段，并基于全波段及特征波段建立 BPNN、SVM、ELM 三种土壤电导率估算模型，此外，PLSR 模型也应用于土壤电导率的估算，作为三种机器学习算法模型的对照，比较模型精度差异，为湖滨绿洲土壤高光谱估算土壤电导率值提供方法支持，实现区域土壤盐分的快速估算。

1　材料与方法

1.1　研究区概况

博斯腾湖西岸湖滨绿洲位于焉耆盆地东南部，处于 86°15′ ~ 86°55′ E，41°45′ ~ 42°10′ N，为山前湖滨绿洲，总面积约为 1360.0 km^2；年均降水量 47.7 ~ 68.1 mm，年均蒸发量 1880.0 ~ 2785.8 mm，年平均气温 8.2 ~ 11.5 ℃，蒸降比大于 40∶1，无霜期 176 ~ 211 d；地下水埋深 1.0 ~ 2.5 m，矿化度为 0.1 ~ 10.0 g L^{-1}；属于大陆性荒漠气候；主要自然植被类型有芦苇、柽柳、胡杨和梭梭等；主要土壤类型有盐土、沼泽土、草甸土、风沙土、灌耕潮土等；研究区土壤盐分平均含量约为 2.84 g kg^{-1}[23]。采样点分布如图 1 所示。

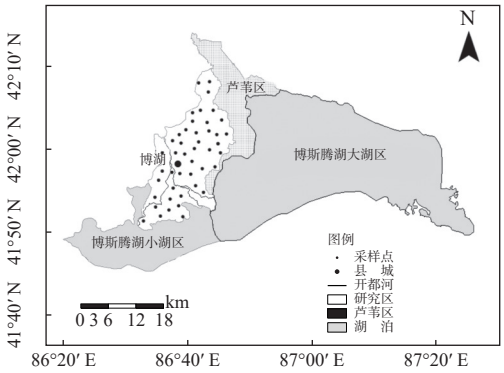


图 1　研究区及采样点分布示意图

Fig.1　Sampling point and location of the study area

1.2　土壤样品测定及分析

依据研究区土壤现状，针对土壤类型、土地利用类型以及植被类型状况等因素，于 2020 年 9 月 24 日 ~ 9 月 28 日进行土壤样品采集，采样点需涵盖研究区主要土地利用类型，以“S”型路线随机布设样点，共计选取 47 个样点，土壤采集深度 0 ~ 50 cm，每 10 cm 为一层采集土壤样品，处理土壤中杂物后以四分法混合均匀，选取约 200 g 土壤样品装袋，共计采集土壤样品 235 份；将土壤样品自然风干后研磨、过筛，重新封装，用作土壤高光谱数据的测试以及土壤有机碳含量的测定。

土壤样品的高光谱数据利用 ASD FieldSpec3 地物光谱仪进行室外测定，选择无风或是风力小于 3 级的晴朗天气，云量低于 5.0%，在光照最佳的时间段北京时间 12:00 ~ 14:00 进行采集；光纤探头垂直于土壤样品表面 15 cm 高处放置且视场角小于 25°；选用 5 点梅花采样法进行高光谱数据采集，选取 5 个位置分别采集 3 次光谱信息，每组数据共计采集 15 次；土壤样品每测一组需重新采集暗电流，同时进行白板优化校正，以减小误差[24]。考虑环境因素的影响，去除高光谱数据 2450 ~ 2500 nm 噪声较大的尾部波段，去除 1350 ~ 1450 nm 与 1800 ~ 1950 nm 水汽影响波段，提高信噪比，减少高频噪音对光谱

数据的影响[25]。应用 Savitzky-Golay 滤波方法进行光谱曲线的平滑处理，去除干扰波段的光谱曲线。235 份土壤样品中剔除 2 个异常值，共计可用 233 份样品。

1.3 模型方法与精度检验

以 CARS[9−10]、SPA[10,26]、GA[11] 变量筛选方法筛选特征波段，以全波段为对照，对比分析三种筛选方法所选取的变量个数；并将不同算法筛选的特征波段结合 BPNN[14,27]、SVM[28−29]、ELM[18] 机器学习算法构建土壤电导率模型，以 PLSR 模型为对照，对比分析四种模型对土壤电导率的估算效果；选取最优模型对研究区不同土层深度的土壤电导率进行估算，并比较其精度差异。土壤样本依据浓度梯度法按照 3:1 的比例对训练集与验证集进行划分[4,30]。

评价模型精度，模型的稳定性与估算能力是其主要的两种表现形式，其中决定系数（Determination Coefficients，R^2）用来检验模型稳定性，均方根误差（Root Mean of Squared Error，RMSE）用来检验模型估算能力。R^2 的值域为 0 ~ 1.0，R^2 的值越大，模型的稳定性越高。RMSE 值越小，模型的估算能力越好。相对分析误差（Relative Percent Deviation，RPD），RPD < 1.40 模型估算能力差；1.40 ≤ RPD < 2.00 模型估算能力提高；RPD ≥ 2.00 模型具有较好的估算能力[25]。

2 结果与分析

2.1 土壤电导率统计特征值

由表 1 可知，样本总集土壤电导率范围在 0.02 ~ 17.22 mS cm^{-1}，平均值为 2.61 mS cm^{-1}，标准差为 3.52 mS cm^{-1}；训练集土壤电导率范围在 0.02 ~ 17.22 mS cm^{-1}，平均值为 2.61 mS cm^{-1}，标准差为 3.52 mS cm^{-1}；验证集土壤电导率范围在 0.05 ~ 14.50 mS cm^{-1}，平均值为 2.64 mS cm^{-1}，标准差为 3.54 mS cm^{-1}；不同土层深度以 0 ~ 10 cm 土壤电导率均值与标准差最高，分别为 2.94 mS cm^{-1}、4.10 mS cm^{-1}；以 40 ~ 50 cm 土壤电导率均值与标准差最低，分别为 2.14 mS cm^{-1}、2.95 mS cm^{-1}。变异系数 CV 值表示离散程度，CV ≥ 100%，为强变异性；10% < CV < 100%，为中等变异性；CV ≤ 10%，为弱变异性[31]。样本总集、训练集、验证集以及不同土层深度的土壤电导率值均呈现强变异性，变异系数 CV ≥ 100%，说明数据具有离散性。验证集与训练集的平均值、标准差与样本总集的平均值、标准差基本一致，具有建模的可行性[32]。

表 1 土壤电导率的统计特征

Table 1 Statistical characteristics of soil electrical conductivity

样本类型 Type of sample	样本数 Number	最小值 Minimum (mS cm^{-1})	最大值 Maximum (mS cm^{-1})	平均数 Average (mS cm^{-1})	标准差 SD	变异系数 CV(%)
样本总集	233	0.02	17.22	2.61	3.52	134.87%
训练集	175	0.02	17.22	2.61	3.52	134.87%
验证集	58	0.05	14.50	2.64	3.54	134.09%
0 ~ 10 cm	46	0.06	17.22	2.94	4.10	139.46%
10 ~ 20 cm	47	0.02	14.48	2.74	3.49	127.37%
20 ~ 30 cm	47	0.05	14.00	2.53	3.41	134.78%
30 ~ 40 cm	46	0.03	14.50	2.75	3.67	133.45%
40 ~ 50 cm	47	0.06	14.00	2.14	2.95	137.85%

2.2 土壤电导率与土壤高光谱特征分析

采用 K-均值（K-means）聚类分析方法将土壤电导率划分 < 0.82 mS cm^{-1}、0.82 ~ 5.57 mS cm^{-1}、5.57 ~ 11.91 mS cm^{-1}、 > 11.91 mS cm^{-1} 四类，图 2 为 4 种不同土壤电导率值的平均光谱曲线进行 Savitzky-Golay 平滑后的效果图。由图 2 可知，四类光谱曲线形状变化基本一致，土壤电导率值越高，土壤光谱反射率越低；当土壤电导率值 < 0.82 mS cm^{-1} 时，反射率均值为 0.37；当土壤电导率值为 0.82 ~ 5.57 mS cm^{-1} 时，反射率均值为 0.35；当土壤电导率值为 5.57 ~ 11.91 mS cm^{-1} 时，反射率均值为 0.32；当土壤电导率值 > 11.91 mS cm^{-1} 时，反射率均值为 0.31；在 350 ~ 600 nm 光谱反射率变化趋势呈现出不断增加，600 ~ 1350 nm、1450 ~ 1800 nm 光谱反射率变化趋势表现为逐渐趋于平缓，1950~ 2450 nm 之间光谱反射率变化波动较大，在 2120 ~ 2150 nm、2380 ~ 2400 nm 波长存在反射峰、2200 ~ 2220 nm、2330 ~ 2350 nm 波长存在吸收谷。

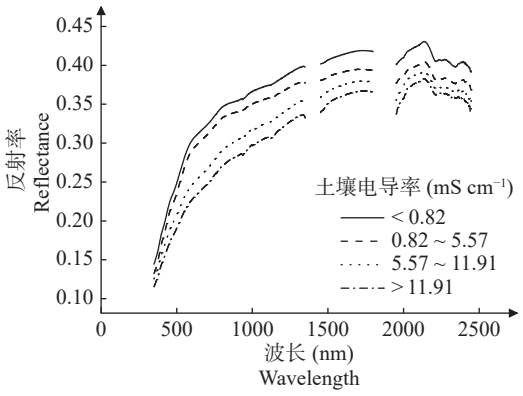


图 2　土壤电导率与土壤高光谱反射率关系

Fig.2　The relationship between soil electrical conductivity and soil hyperspectral reflectance

2.3　基于机器学习算法的土壤电导率估算模型

由表 2、表 3 可知，CARS、SPA、GA 算法将输入波段分别压缩至全波段数目的 0.87%、1.68%、0.70%。BPNN 模型能力表现为 SPA-BPNN > CARS-BPNN > Full-spectral-BPNN > GA-BPNN。SVM 模型能力表现为 CARS-SVM > SPA-SVM > Full-spectral-SVM > GA-SVM。ELM 模型能力表现为 CARS-ELM > SPA-ELM > GA-ELM > Full-spectral-ELM。PLSR 模型能力表现为 CARS-PLSR > SPA-PLSR > Full-spectral-PLSR > GA-PLSR。

综合分析三种算法在构建模型时简化模型的能力及 R^2、RPD、RMSE 三种模型评价指标，研究区变量方法的筛选 CARS > SPA > GA。

16 种模型中以 CARS 算法构建的 SVM 模型精度最高，训练集与验证集 R^2 分别为 0.76、0.75，RPD 分别为 2.04、2.00，RMSE 分别为 1.79 mS cm^{-1}、1.68 mS cm^{-1}。以 PLSR 模型为对照，基于 CARS 算法构建的 BPNN、SVM、ELM 模型，训练集与验证集 RPD 分别平均提高 21.80%、22.92%，R^2 分别平均提高 20.22%、21.31%；基于 SPA 算法构建的

表 2　特征波段筛选结果

Table 2　Feature band screening results

筛选方法 Screening method	变量数量 Number of variable	特征波段(nm) Characteristic band
CARS	16	1486、1487、1519、1520、1951、1984、2061、2348、2350、2386、2387、2395、2396、2419、2427、2447
SPA	31	946、1001、1494、1731、1951、1957、1963、1978、2011、2063、2226、2309、2323、2344、2348、2352、2358、2365、2370、2392、2396、2403、2410、2412、2417、2423、2437、2440、2442、2446、2447
GA	13	355、956、1972、1973、2104、2153、2260、2344、2347、2362、2373、2390、2426
Full- spectral	1848	350 ~ 1349、1451 ~ 1799、1951 ~ 2449

表 3　基于机器学习算法的土壤电导率估算结果

Table 3　Estimation results of soil electrical conductivity based on machine learning algorithm

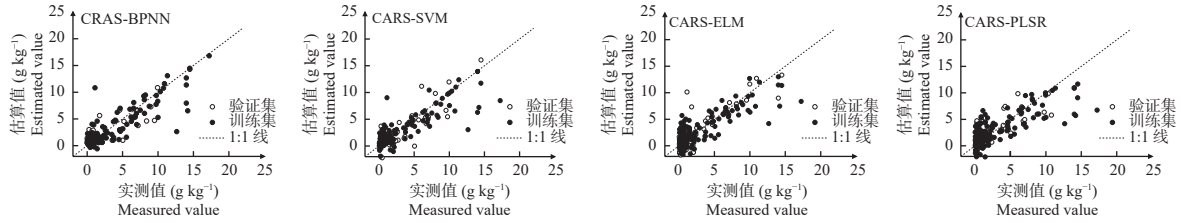
模型 Model	筛选方法 Screening method	训练集 Training set			验证集 Verification set		
		R^2	RMSE	RPD	R^2	RMSE	RPD
BPNN	CARS	0.73	1.90	1.92	0.75	1.57	2.01
	SPA	0.75	1.85	1.99	0.76	1.78	2.03
	GA	0.57	2.36	1.52	0.53	2.40	1.46
	Full-spectral	0.72	1.82	1.87	0.73	2.62	1.94
SVM	CARS	0.76	1.79	2.04	0.75	1.68	2.00
	SPA	0.72	2.04	1.89	0.73	1.34	1.94
	GA	0.63	2.36	1.64	0.64	1.91	1.67
	Full- spectral	0.70	2.07	1.82	0.66	2.55	1.72
ELM	CARS	0.71	1.91	1.85	0.72	1.95	1.89
	SPA	0.67	2.10	1.73	0.67	1.91	1.73
	GA	0.59	2.24	1.57	0.60	2.36	1.59
	Full- spectral	0.57	2.37	1.52	0.59	2.18	1.57
PLSR	CARS	0.61	2.29	1.59	0.61	2.01	1.60
	SPA	0.57	2.38	1.53	0.57	2.14	1.53
	GA	0.49	2.46	1.40	0.52	2.57	1.44
	Full- spectral	0.56	2.40	1.52	0.56	2.11	1.51

BPNN、SVM、ELM 模型，训练集与验证集 RPD 分别平均提高 22.22%、24.18%，R^2 分别平均提高 25.15%、26.32%；基于 GA 算法构建的 BPNN、SVM、ELM 模型，训练集与验证集 RPD 分别平均提高 12.62%、9.26%，R^2 分别平均提高 21.77%、13.46%；全波段构建的 BPNN、SVM、ELM 模型，训练集与验证集 RPD 分别平均提高 14.25%、15.45%，R^2 分别平均提高 18.45%、17.86%。综上所述，机器学习算法估算能力及模型稳定性会因为输入量不同有所改变，但不同输入量应用机器算法构建的土壤电导率模型精度均优于线性模型。

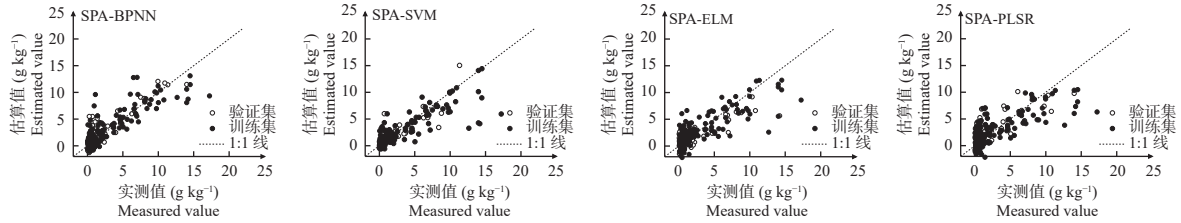
由图 3 可知，采用 CARS 算法筛选的特征波段建立的 CARS-SVM 模型与采用 SPA 算法筛选的特征波段建立的 SPA-BPNN 模型，实测值与估算值样点较为均匀的分布在 1∶1 线两侧，模型估算效果好，精度高于其余 14 种模型组合；采用 GA 算法筛选的特征波段建立的 GA-BPNN、GA-SVM、GA-ELM、GA-PLSR 模型估算效果略差，存在明显偏离 1∶1 线的分布点，且大多位于 1∶1 线下方，说明估算值较实测值偏低，存在低估现象；采用全波段光谱数据建立的 Full spectral-BPNN、Full spectral-SVM、Full spectral-ELM、Full spectral-PLSR 模型样本的估算值小于实测值，但相较于 GA 所建立的 4 类模型效果稍好，其中 Full spectral-BPNN 模型精度高于全波段光谱数据建立的其余 3 种模型组合。

2.4 基于不同土层深度的土壤电导率估算结果

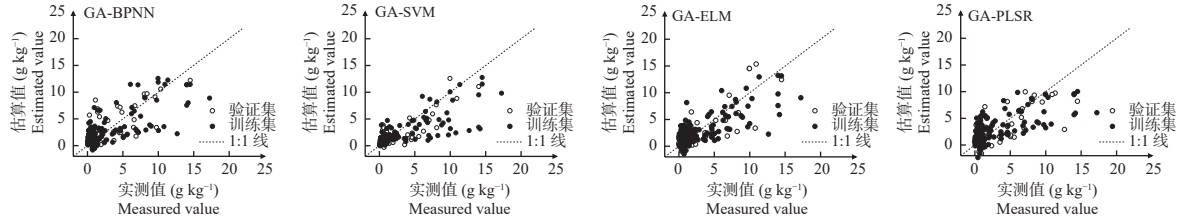
选取表 3 中 16 种土壤电导率高光谱估算模型中精度最高的 CARS-SVM 模型对不同土层深度的土壤电导率进行估算。由表 4 可知，不同土层深度 CARS 算法所筛选的响应波段不同，特征波段较多集中于



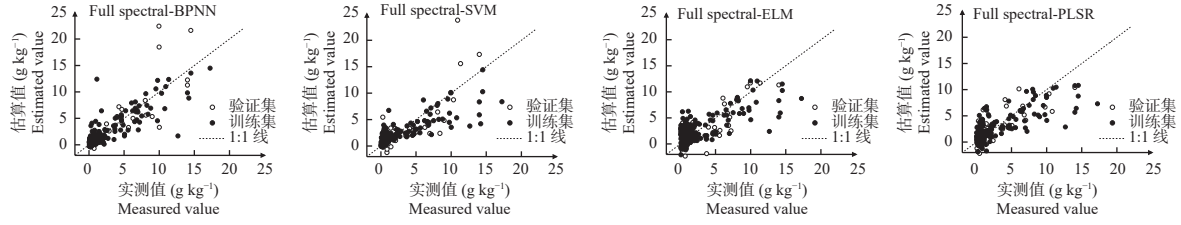
(a) 基于 CARS 算法的模型精度



(b) 基于 SPA 算法的模型精度



(c) 基于 GA 算法的模型精度



(d) 基于全波段光谱数据的模型精度

图 3 土壤电导率高光谱模型的精度比较

Fig.3 Comparison of the accuracy of soil electrical conductivity high spectroscopy model

1500 ~ 2500 nm 近红外长波波段。亚森江·喀哈尔等[12] 构建的土壤电导率光谱敏感波段为 2011、1890，2011、1891 nm。曹肖奕等[19] 研究表明光谱 350~880 nm 附近、1500 ~ 2100 nm 附近以及 2200 ~ 2450 nm 附近与土壤电导率具有较高的相关性。CARS 算法所筛选的特征波段与上述波段范围多有重合，说明特征波段的筛选具有合理性。

表 4　CARS 算法筛选特征波段
Table 4　CARS algorithm to screen characteristic bands

土层深度（cm） Soil Depth	变量数量 Number of variables	特征波段(nm) Characteristic band
0 ~ 10	7	1959、2275、2285、2307、2395、2417、2447
10 ~ 20	16	938、1153、1154、1690、1692、1750、1982、1983、2165、2168、2351、2352、2384、2393、2394、2445
20 ~ 30	10	1962、1975、1977、2020、2021、2350、2395、2404、2413、2439
30 ~ 40	7	1968、1969、2037、2350、2391、2396、2415
40 ~ 50	26	670、672、837、838、839、840、841、842、935、963、1951、1962、1963、2009、2233、2325、2327、2387、2393、2394、2397、2404、2408、2418、2428、2437

由表 5 可知，估算结果发现土层深度 20 ~ 30 cm 土壤电导率估算模型精度最高，R^2 均大于 0.80，RPD 均大于 2.00，RMSE 均小于 2.00 mS cm^{-1}，该土层深度与样本总集的平均值、标准差、变异系数均相差较小，相较于其余不同土层深度土壤电导率估算模型稳定性更高，训练集与验证集 R^2 接近，泛化能力好[33]；土层深度 30 ~ 40 cm 土壤电导率估算模型精度仅次于 20 ~ 30 cm，模型估算能力较好。土层深度 10 ~ 20 cm 与 40 ~ 50 cm 构建的土壤电导率估算模型中 40 ~ 50 cm 模型精度略高。土层深度 0 ~ 10 cm 构建的土壤电导率估算模型精度最差，模型验证集 RMSE 大于 5.00 mS cm^{-1}，且训练集与验证集 R^2 相差 0.12，相较于其余不同土层深度土壤电导率估算模型泛化能力及稳定性较差。

表 5　基于 CARS-SVM 的土壤电导率估算结果
Table 5　Estimation results of soil electrical conductivity based on CARS-SVM

模型 Model	土层深度 Soil depth	训练集 Training set				验证集 Verification set			
		R^2	RMSE	RPD	线性方程 Linear equation	R^2	RMSE	RPD	线性方程 Linear equatiom
CARS-SVM	0 ~ 10	0.54	3.03	1.47	$y = 0.42x + 1.01$	0.66	5.85	1.72	$y = 1.85x - 0.42$
	10 ~ 20	0.56	2.54	1.50	$y = 0.41x + 0.68$	0.59	2.40	1.56	$y = 0.49x + 0.65$
	20 ~ 30	0.83	1.37	2.41	$y = 0.77x + 0.20$	0.84	1.77	2.50	$y = 0.80x - 0.19$
	30 ~ 40	0.78	1.89	2.11	$y = 0.72x + 0.71$	0.75	1.50	1.99	$y = 1.15x + 0.16$
	40 ~ 50	0.69	1.37	1.79	$y = 0.49x + 0.63$	0.66	3.44	1.71	$y = 0.31x + 0.87$

3　讨论

土壤电导率的快速估算可以为土壤盐渍化提供一定的理论依据和模型参考，土壤高光谱数据的全波段数目较多，尽管包含丰富信息，却易造成信息冗余，本文利用三种变量筛选方法优选的特征波段作为建模输入量，以简化计算过程，提高模型的建模效率及估算能力[34]。研究结果表明变量方法的选择 CARS > SPA > GA，这与李冠稳[35] 简化模型能力 CARS > GA > SPA 结果不一致，可能是由于研究的土壤属性不同，且 GA 提取的特征波长数量相对于 CARS 与 SPA 较少，所包含的有用信息少，GA 算法多是解决共线性问题，不能将光谱信息较好地表达出来，损失一部分信息[36]。基于筛选的特征波段构建的 BPNN、SVM、ELM 三种机器学习算法估算模型，将 PLSR 建模结果作对照，机器学习算法估算模型的精度明显提高，这与葛翔宇等[21] 结果基本一致，机器学习模型不仅在统计结果上优于 PLSR，在估算能力上也表现出更好的稳健性和泛化能力，相比 PLSR 线性模型，不同建模输入量所构建的 BPNN、SVM、ELM 模型训练集与验证集 R^2 分别平均增加 21.40%、19.74%，RPD 分别平均增加 17.72%、17.95%。三种机器学习算法的建模效果随建模输入量不同而有所变化，其中以 GA 算法筛选的特征波段作为输入量构建的模型与曹肖奕[19]ELM > SVM > BPNN 结果基本

一致；以全波段数据作为输入量构建的模型 BPNN > SVM > ELM；以 CARS 算法筛选的特征波段作为输入量构建的模型 SVM > BPNN > ELM；以 SPA 算法筛选的特征波段作为输入量构建的模型 BPNN > SVM > ELM 等结果不一致。这可能是由于不同特征波段筛选方法所筛选的响应波段不同所导致的，其机理有待于进一步研究。CARS 算法筛选不同土层深度的特征波段也有所不同，响应波段的差异会导致模型精度差异，以 CARS-SVM 模型对不同土层深度的土壤电导率值估算，0 ~ 10 cm 的土壤电导率估算模型精度最差，这可能是由于 0 ~ 10 cm 土壤电导率值相较于其余土层变异系数值最高，模型估算能力差； 20 ~ 30 cm 的土壤电导率估算模型精度最好，此土层与样本总集的土壤电导率统计特征值较为相近，对比其它土层土壤电导率估算模型稳定性更高。

4 结论

研究区土壤电导率值变化范围为 0.02 ~ 17.22 mS cm^{-1}，平均值为 2.61 mS cm^{-1}，变异系数为 134.87%，呈现强变异性；以土层深度 0 ~ 10 cm 的土壤电导率平均值最高、变异性最强，平均值为 2.94 mS cm^{-1}、变异系数为 139.46%。土壤电导率值越高，土壤光谱反射率越低；当土壤电导率值 < 0.82 mS cm^{-1} 时，反射率均值为 0.37；当土壤电导率值为 0.82 ~ 5.57 mS cm^{-1} 时，反射率均值为 0.35；当土壤电导率值为 5.57 ~ 11.91 mS cm^{-1} 时，反射率均值为 0.32；当土壤电导率值 > 11.91 mS cm^{-1} 时，反射率均值为 0.31。

CARS、SPA、GA 算法筛选的光谱特征波段数量分别为 16、31、13，去除水汽及噪声过大的全波段数量为 1848，将建模输入波段分别压缩至全波段数目的 0.87%、1.68%、0.70%，3 种变量筛选方法简化建模输入量能力为 GA > CARS > SPA，结合 BPNN、SVM、ELM 及 PLSR 模型评价指标，研究区变量方法的选择 CARS > SPA > GA。

BPNN、SVM、ELM 估算模型优于 PLSR 估算模型，对照 PLSR 模型，不同建模输入量所构建的 BPNN、SVM、ELM 模型 R^2 平均增加 20.57%，RPD 平均增加 17.84%。利用全波段及 3 种变量筛选方法分别构建 4 种估算模型，最优模型为 CARS-SVM，训练集与验证集 R^2 分别为 0.76、0.75，RMSE 分别为 1.79 mS cm^{-1}、1.68 mS cm^{-1}，RPD 分别为 2.04、2.00。

不同土层深度的土壤电导率估算模型，以土层深度 0 ~ 10 cm 土壤电导率估算模型精度最差，训练集与验证集 R^2 分别为 0.54、0.66，RMSE 分别为 3.03 mS cm^{-1}、5.85 mS cm^{-1}，RPD 分别为 1.47、1.72；以土层深度 20 ~ 30 cm 土壤电导率估算模型精度最高，训练集与验证集 R^2 分别为 0.83、0.84，RMSE 分别为 1.37 mS cm^{-1}、1.77 mS cm^{-1}，RPD 分别为 2.41、2.50。

参考文献：

[1] 曹肖奕, 丁建丽, 葛翔宇, 等. 基于不同卫星光谱模拟的土壤电导率估算研究[J]. 干旱区地理, 2020, 43(1): 172 − 181.

[2] 李 相, 丁建丽, 侯艳军, 等. 干旱半干旱区土壤含盐量和电导率高光谱估算[J]. 冰川冻土, 2015, 37(4): 1050 − 1058.

[3] Chen K, Li C, Tang R N. Estimation of the nitrogen concentration of rubber tree using fractional calculus augmented NIR spectra[J]. Industrial Crops & Products, 2017, 108: 832 − 839.

[4] 赵 慧, 李新国, 靳万贵, 等. 基于地理加权回归模型的博斯腾湖湖滨绿洲土壤盐分离子含量高光谱估算[J]. 土壤, 2021, 53(3): 646 − 653.

[5] 罗德芳, 冯春晖, 吴家林, 等. 基于电磁感应协同野外原位光谱的土壤盐分反演研究[J]. 中国土壤与肥料, 2020, (6): 107 − 113.

[6] Zhao W, Sánchez N, Lu H, et al. A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression[J]. Journal of Hydrology, 2018, 563: 1009 − 1024.

[7] 杨丽萍, 侯成磊, 苏志强, 等. 基于机器学习和全极化雷达数据的干旱区土壤湿度反演[J]. 农业工程学报, 2021, 37(13): 74 − 82.

[8] 王 涛, 喻彩丽, 张楠楠, 等. 基于去包络线和连续投影算法的枣园土壤电导率光谱检测研究[J]. 干旱地区农业研究, 2019, 37(5): 193 − 199 + 217.

[9] 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.

[10] 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]. 农业工程学报, 2016, 32(13): 95 − 102.

[11] 乔 天, 吕成文, 肖文凭, 等. 基于遗传算法的土壤质地高光谱预测模型研究[J]. 土壤通报, 2018, 49(4): 773 − 778.

[12] 亚森江·喀哈尔, 杨胜天, 尼格拉·塔什甫拉提, 等. 基于分数阶微分优化光谱指数的土壤电导率高光谱估算[J]. 生态学报, 2019, 39(19): 7237 − 7248.

[13] 赵 慧, 李新国, 靳万贵, 等. 基于分数阶微分的博斯腾湖湖滨绿洲土壤电导率高光谱估算[J]. 甘肃农业大学学报, 2021, 56(1): 118 − 125.

[14] 王雪梅, 玉米提·买明, 毛东雷, 等. 干旱区绿洲耕层土壤重金属铬含量的高光谱估测[J]. 生态环境学报, 2021, 30(10): 2076 − 2084.

[15] 田安红, 付承彪, 熊黑钢, 等. BPNN对不同人为活动区域的盐渍土Na^+高光谱估测[J]. 水土保持研究, 2020, 27(2): 364 − 369.

[16] Rocha Neto O, Teixeira A, Leão R, et al. Hyperspectral Remote Sensing for Detecting Soil Salinization Using ProSpec TIR-VS Aerial Imagery and Sensor Simulation[J]. Remote Sensing, 2017, 9(1): 1 − 16.

[17] Bao N S, Wu L X, Ye B Y, et al. Assessing soil organic matter of reclaimed soil from a large surface coal mine using a field spectroradiometer in laboratory[J]. Geoderma, 2017, 288: 47 − 55.

[18] 蔡亮红, 丁建丽. 基于变量优选和ELM算法的土壤含水量预测研究[J]. 光谱学与光谱分析, 2018, 38(7): 2209 − 2214.

[19] 曹肖奕, 丁建丽, 葛翔宇, 等. 基于光谱指数与机器学习算法的土壤电导率估算研究[J]. 土壤学报, 2020, 57(04): 867 − 877.

[20] Xiang Y, Liu Q, Wang Y B, et al. Evaluation of MLSR and PLSR for estimating soil element contents using visible/near-infrared spectroscopy in apple orchards on the Jiaodong peninsula[J]. Catena, 2016, 137: 340 − 349.

[21] 葛翔宇, 丁建丽, 王敬哲, 等. 基于竞争适应重加权采样算法耦合机器学习的土壤含水量估算[J]. 光学学报, 2018, 38(10): 393 − 400.

[22] 曾　胤, 陆宇振, 杜昌文, 等. 应用红外光声光谱技术及支持向量机模型测定土壤有机质含量[J]. 土壤学报, 2014, 51(6): 1262 − 1269.

[23] 赵　慧, 李新国, 牛芳鹏, 等. 博斯腾湖湖滨绿洲土壤电导率高光谱估算模型[J]. 中国土壤与肥料, 2021, 2: 289 − 295.

[24] 牛芳鹏, 李新国, 麦麦提吐尔逊·艾则孜, 等. 基于连续投影算法的博斯腾湖西岸湖滨绿洲土壤有机碳含量的高光谱估算[J]. 浙江大学学报(农业与生命科学版), 2021, 47(5): 673 − 682.

[25] 张子鹏, 丁建丽, 王敬哲. 基于谐波分析算法的干旱区绿洲土壤光谱特性研究[J]. 光学学报, 2019, 39(2): 391 − 401.

[26] 吾木提·艾山江, 买买提·沙吾提, 马春玥. 基于分数阶微分和连续投影算法-反向传播神经网络的小麦叶片含水量高光谱估算[J]. 激光与光电子学进展, 2019, 15: 251 − 259.

[27] 董　哲, 杨武德, 朱洪芬, 等. 基于连续投影算法与BP神经网络的玉米叶片SPAD值高光谱估算[J]. 山西农业科学, 2019, 47(5): 751 − 755.

[28] 刘翠英, 张津瑞, 曾　涛, 等. 傅里叶变换红外光谱的土壤团聚体有机碳和全氮含量估测[J]. 光谱学与光谱分析, 2020, 40(12): 3818 − 3824.

[29] 孙亚楠, 李仙岳, 史海滨, 等. 河套灌区土壤水溶性盐基离子高光谱综合反演模型[J]. 农业机械学报, 2019, 50(05): 344 − 355.

[30] 肖云飞, 高小红, 李冠稳. 土壤有机质可见光—近红外光谱预测样本优化选择[J]. 土壤, 2020, 52(2): 404 − 413.

[31] 韩　宁, 陈蜀江, 朱　选, 等. 基于冗余分析的伊犁新垦绿洲不同农田土壤盐渍化特征研究[J]. 西南农业学报, 2019, 32(2): 366 − 372.

[32] 孙问娟, 李新举. 煤矿区土壤有机碳含量的高光谱预测模型[J]. 水土保持学报, 2018, 32(5): 346 − 351.

[33] 赵明松, 谢　毅, 陆龙妹, 等. 基于高光谱特征指数的土壤有机质含量建模[J]. 土壤学报, 2021, 58(1): 42 − 54.

[34] 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.

[35] 李冠稳, 高小红, 肖能文, 等. 基于sCARS-RF算法的高光谱估算土壤有机质含量[J]. 发光学报, 2019, 40(8): 1030 − 1039.

[36] 吕美蓉, 任国兴, 李雪莹, 等. 可见-近红外光谱的潮间带沉积物有机碳含量的几种模型预测方法[J]. 光谱学与光谱分析, 2020, 40(4): 1082 − 1086.

Hyperspectral Estimation Model of Soil Conductivity in the Lakeside Oasis Based on Machine Learning Algorithm

MENG Shan[1, 2], LI Xin-guo[1, 2*], JIAO Li[1, 2]

(1. *College of Geographic Sciences and Tourism*, *Xinjiang Normal University*, *Urumqi* 830054, *China*; 2. *Xinjiang Laboratory of Lake Environment and Resources in Arid Zone*, *Urumqi* 830054, *China*)

Abstract: [Objective] The paper aims to provide method for estimating the soil conductivity of lakeside oasis soil by hyperspectral, so as to realize the rapid estimation of regional soil salinity. **[Method]** Combined analysis of soil conductivity values and soil hyperspectral data, competitive adaptive reweighted sampling (CARS), successive projection algorithm (SPA) and genetic algorithm (GA) were used to screen the characteristic bands of soil conductivity. Based on the full band and characteristic band, three machine learning algorithm models, inlcuding BP neural network (BPNN), support vector machine (SVM) and extreme learning machine (ELM), were constructed, and the partial least squares model (PLSR) was introduced for comparing their accuracy. **[Result]** The soil conductivity ranged from 0.20 to 17.22 mS cm^{-1} in the study area, with an average value of 2.61 mS cm^{-1} and a coefficient of variation of 134.87%, showing strong variability; The characteristic bands screened by the CARS, SPA, and GA algorithms compress the modeling input to 0.87%, 1.68%, and 0.70% of the total number of bands, respectively, which reduced the amount of modeling input and increased the modeling speed. The choice of variable method CARS > SPA > GA; The three machine learning algorithm models were all better than PLSR model. The coefficient of determination (R^2) increased by 20.57% and the relative percent deviation (RPD) increased by 17.84% on average. The CARS-SVM was the best model for soil conductivity hyperspectral estimation, with R^2 of 0.76 and 0.75 for training set and validation set, respectively, RMSE of 1.79 mS cm^{-1} and 1.68 mS cm^{-1}, and RPD of 2.04 and 2.00, respectively; The soil conductivity hyperspectral estimation model with a soil depth of 20 ~ 30 cm has the highest accuracy, with R^2 of 0.83 and 0.84 for training set and validation set, respectively, RMSE of 1.37 mS cm^{-1} and 1.77 mS cm^{-1}, and RPD of 2.41 and 2.50, respectively. **[Conclusion]** The soil conductivity hyperspectral estimation model based on CARS-SVM has high accuracy and optimal estimation ability, which can provide a scientific reference for the estimation of soil conductivity in lakeside oasis.

Key words: Soil conductivity value; Competitive adaptive reweighted sampling; Successive projection algorithm; Genetic algorithm; Machine learning algorithm; Hyperspectral estimation model

[责任编辑：裴久渤]