基于 CARS 算法的不同类型土壤有机质高光谱预测

唐海涛[1]，孟祥添[1]，苏循新[2]，马 涛[3]，刘焕军[1,4]，鲍依临[1]，张美薇[1]，张新乐[1※]，霍海志[3]

（1. 东北农业大学公共管理与法学院，哈尔滨 150030；2. 黑龙江省地质资料档案馆，哈尔滨 150030；3. 黑龙江省第五地质勘察院，哈尔滨 150030；4. 中国科学院东北地理与农业生态研究所，长春 130012）

摘 要：不同土壤类型的理化性质和光谱性质存在差异，以往研究多以高光谱反射率或光谱吸收特征建立模型，输入变量类型结构单一，往往导致土壤有机质（Soil Organic Matter，SOM）预测模型的精度不高。为提高 SOM 高光谱预测模型精度，该研究以黑龙江省海伦市为研究区，将不同类型土壤分别以竞争自适应重加权采样（Competitive Adaptive Reweighted Sampling，CARS）筛选的特征波段、数字高程模型（Digital Elevation Model，DEM）数据和光谱指数作为输入变量，结合随机森林（Random Forest，RF）算法建立 SOM 预测模型。结果表明：1）通过 CARS 算法筛选后，各土壤类型特征波段压缩至全波段数目的 16%以下，在很大程度上降低土壤高光谱变量维度和计算复杂程度，从而提高了模型的预测能力，说明 CARS 算法在提取特征关键波段变量、优化模型结构方面起到重要作用；2）不同类型土壤的 SOM 预测精度存在差异，沼泽土的预测精度最高为 0.768，性能与四分位间隔距离的比率（Ratio of Performance to InterQuartile distance，RPIQ）为 3.568；黑土次之，草甸土的预测精度最低，仅 0.674，RPIQ 为 1.848。3 类土壤的 RPIQ 均达到 1.8 以上，模型具有较好的预测能力；3）局部回归预测精度最优，验证集的调整后决定系数为 0.777，均方根误差（Root Mean Square Error，RMSE）为 0.581%，模型验证 RPIQ 为 2.689，模型稳定性高。该试验筛选的预测因子通过 RF 模型可实现 SOM 含量的快速预测，简化了传统复杂的程序，可为中尺度区域不同类型土壤的 SOM 预测提供依据，为输入量的选择提供参考。

关键词：遥感；土壤；有机质；光谱指数；地形；特征波段筛选；随机森林
doi：10.11975/j.issn.1002-6819.2021.2.013
中图分类号：S153.621；O433.4 **文献标志码**：A **文章编号**：1002-6819(2021)-2-0105-09

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. doi：10.11975/j.issn.1002-6819.2021.2.013 http://www.tcsae.org
Tang Haitao, Meng Xiangtian, Su Xunxin, et al. Hyperspectral prediction on soil organic matter of different types using CARS algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(2): 105-113. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2021.2.013 http://www.tcsae.org

0 引 言

土壤有机质（Soil Organic Matter，SOM）可以通过生物合成和分解，改善土壤的物理、化学和生物特性[1]，在控制土壤功能和质量、抵消温室气体排放、完善全球碳循环系统信息等方面发挥着重要作用[2]。高光谱预测模型为实现 SOM 等土壤属性速测与遥感反演以及表层碳库估算等提供数据信息[3]，并为 SOM 速测仪器研制、土壤制图与退化监测、精准农业实施等提供数据与技术支持[4]。高光谱技术具有精细的光谱分辨率，可获取地物纳米级的连续光谱信息，SOM 具有多种官能团（如羟基、羧基等），分别在红外光谱区域有特征性吸收，且不同波段的吸收强度与该物质的分子结构及浓度存在对应关系，因此，红外光谱可以反映 SOM 含量，为其定量估算提供了一种有效的手段，为预测 SOM 提供了可能[5]。黑龙江省海伦市位于世界三大黑土地分布区之一的松嫩平原东北端，土壤类型多样，其中黑土面积达到全市面积 1/2 以上，且是中国重要的商品粮基地，了解其 SOM 的分布情况、空间变化规律，有利于科学评价土壤的质量情况并对农场合理施肥提供指导，对耕地资源的可持续利用具有十分重要的实际意义，可为海伦市耕地的可持续利用和土壤质量保护监测提供技术支持，为将来海伦市土地管理建立完整的空间土壤信息系统提供框架。

以往室内高光谱对于SOM的输入变量研究多停留在以全波段反射率或对应的数学变换上，选取相关系数较大的波段进行建模，该方法仅考虑了 SOM 与光谱间的关系，并没有考虑光谱间的重叠吸收或相互影响[6]。利用光谱指数技术预测 SOM 的研究成为当前热点，光谱指数是由几个窄波段或宽波段组合而成，可通过分析特定波段间的相互作用，提高对待测属性的敏感程度[7]，有助于挖掘波段间的隐晦信号[8]。SOM 空间分布特征受到高程、坡度、坡向等地形因子不同程度的影响，地形条件影响

收稿日期：2020-11-04 修订日期：2021-01-04
基金项目：国家自然科学基金（41671438），东北农业大学“学术骨干”项目（54935112）资助
作者简介：唐海涛，主要研究方向为农业遥感。Email：tht0918@yeah.net
※通信作者：张新乐，博士，副教授，主要研究方向为生态遥感。
Email：zhangxinle@gmail.com

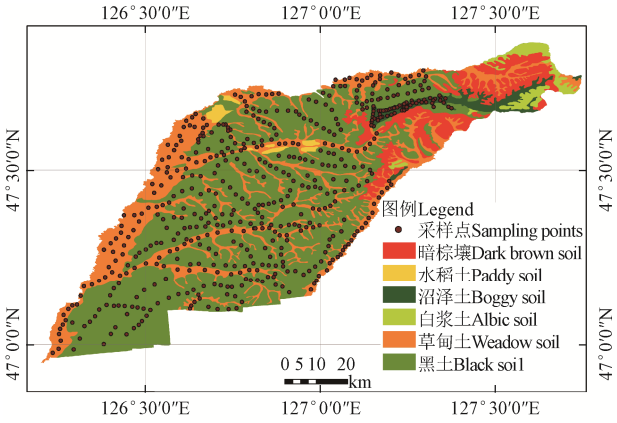
其物质循环过程和强度[9]，通过数字高程模型（Digital Elevation Model，DEM）提取高程作为模型辅助变量参与建模。同时特征波段选择是进行 SOM 含量预测的一个重要方面，已经引起了越来越多学者的关注。土壤光谱反射数据通过竞争自适应重加权采样（Competitive Adaptive Reweighted Sampling，CARS）筛选出的特征波段不仅将输入波段压缩至全波段数目的一半以下，同时提升了模型估测精度，降低了变量维度和模型复杂度[10]，Vohland 等[11]发现，在 60 个农业样品的土壤属性预测中，CARS 算法减少了建模时间，且能够合理、精确、有效的确定特征波段在全波段中的位置。以往的学者多以一种类型的土壤为对象，进行 SOM 高光谱响应特性研究，但是由于土壤的光谱反射率是土壤内在理化特性光谱行为的综合反应，不同类型土壤的光谱特征不同[12]，因此模型普适性较弱。卢艳丽等[13]利用不同土壤类型分组试验分析了东北平原土壤光谱反射率曲线形状变化，确定了 8 种不同类型土壤与原始光谱反射率的相关敏感波段并建立了同质性 SOM 预测线性模型，从而达到简化 SOM 预测模型的目的。Bao 等[14]对比了多种土壤分组策略下 SOM 的预测精度，同时引入竞争自适应重加权采样方法进行模型输入量的筛选，证实了土壤分类的优势与多输入量降维的必要性。因此，不同类型土壤分别提取输入变量进行高光谱 SOM 预测将有利于分析各类土壤的理化性质，从而提高预测精度。

已有 SOM 高光谱预测研究常基于一种土壤类型建立模型或者多种土壤类型进行全局回归建模，且输入变量的类型较为单一，存在 SOM 预测精度不高的情况[15]。为了充分考虑土壤光谱信息及影响因素，本研究以海伦市为研究区域，根据全国第二次土壤普查结果及对采样点的地理位置对土样进行分类。在土壤分类的前提下，以土壤光谱反射率数据、DEM 数据以及光谱指数作为输入变量，建立基于随机森林算法（Random Forest，RF）的分类高光谱 SOM 预测模型。为了降低输入量之间的共线性，引入 CARS 算法筛选特征波段，提高不同类型 SOM 预测的精度，以期实现动态快速预测 SOM 含量。

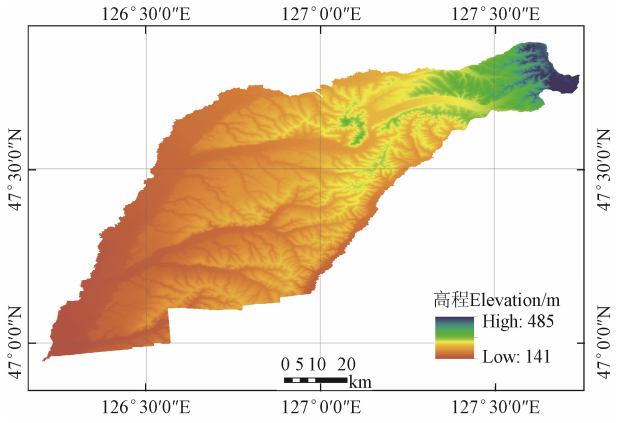
1 材料与方法

1.1 研究区概况

海伦市位于松嫩平原的中心地带，地理位置在 46°58'N～47°52'N，126°14'E～127°45'E 之间，属温带大陆性季风气候，地势平坦，土质肥沃，耕地面积广阔，是国家重要的商品粮基地。其土壤类型主要为黑土、草甸土和沼泽土，在该研究区内还有少量的水稻土、暗棕壤及白浆土。黑土土层深厚，结构良好，富含 SOM 和腐殖质，自然肥力高。沼泽土所处的地势大都比较低洼，SOM 累积明显。由于该区地形高程差较大，加上耕地的长期粗放利用导致土壤侵蚀严重，降水将地势较高的土壤冲积到地势较低的草甸土表面，导致表层草甸土性质较为复杂多样[16]。海伦市主要土壤类型（全国第二次土壤普查结果）及采样点分布图和海伦市 30 m 空间分辨率的 DEM 数据见图 1。



a. 海伦市土壤类型及采样点分布
a. Distribution of soil types and sampling points in Hailun City



b. 海伦市 DEM
b. DEM of Hailun City

图 1 研究区土壤类型、采样点分布及 DEM

Fig.1 Distribution and DEM , soil types and sampling points in the study area

1.2 土样样品

2019 年 5 月 15—20 日，于作物出苗前，沿主要乡级以上道路，在黑龙江省海伦市全市进行样本采集。选择土壤裸露的地区作为样区，考虑土地利用类型和土壤类型采集 0～20 cm 耕层土壤。为保证采样点的有机质含量能够代表采样点附近一定空间内的 SOM 水平，采用四分法收集样品，同时利用 GPS 记录采样点经纬度，总共采集土壤样本 548 个。采集的样品经过风干，研磨，过 2 mm 筛。每个样品分 2 份，一份用于光谱测量；一份用于 SOM 含量分析。SOM 含量用高温外热重铬酸钾氧化容量法测定[17]。

1.3 光谱测量及数据预处理

采用 ASD FieldSpec®3 便携式光谱仪在暗室内对风干土进行光谱测试。光谱测试流程详见文献[18]。由于反射率波谱在 400～430 和 2 400～2 500 nm 范围内噪声较为强烈，为减少高频噪声的干扰，本文选取光谱反射率波谱范围为 430～2 400 nm，并对其进行 9 点平滑、10 nm 重采样处理，此过程分别在 EXCEL 和 ENVI 5.3 中实现。

不考虑土壤空间差异性，将整个土壤样本作为全局回归预测数据集。同时，土壤样本根据全国第二次土壤普查图，利用 ArcGIS 10.1 中的工具箱提取每个土壤样本的土壤类型，将土壤样本划分为不同土壤类型，同一种土壤具有相同光谱表现特征的土壤样本集。根据中国土

壤分类，土壤类型可分为黑土、草甸土、沼泽土，然后针对不同分类样本进行局部回归预测建模。

1.4　输入量提取

国内外学者进行 SOM 高光谱估测时，输入量多选择为高光谱反射率或光谱吸收特征建立模型，输入变量类型结构单一，容易忽略土壤高光谱反射率之间的高度共线性[19]。本研究通过 CARS 算法挑选的特征变量、光谱指数结合 DEM 数据作为模型输入变量。

1.4.1　光谱指数

在高光谱数据预测 SOM 的研究中，为了确定敏感的波段，必须从 SOM 含量信息中获取深度信号，因此光谱指数常作为一个重要指标[20]。本文探讨归一化指数（Normalized Difference Index，NDI）、再归一化指数（Renormalized Difference Vegetation Index，RDVI）、比值指数（Ratio Index，RI）与 SOM 含量之间的关系。

表 1　光谱指数及公式

Table 1　Spectral indices and formula

光谱指数 Spectral index	公式 Formula	参考文献 Reference
归一化指数 Normalized Difference Index (NDI)	$\mathrm{NDI}(R_i,R_j)=\frac{R_i-R_j}{R_i+R_j}$	[20-21]
再归一化指数 Renormalized Difference Vegetation Index (RDVI)	$\mathrm{RDVI}(R_i,R_j)=\frac{R_i-R_j}{\sqrt{R_i+R_j}}$	[22-23]
比值指数 Ratio Index (RI)	$\mathrm{RI}(R_i,R_j)=\frac{R_i}{R_j}$	[24]

注：式中 R_i、R_j 属于 430～2 400 nm 中任意两波段，且 $R_i \neq R_j$。

Note: R_i and R_j in the formula belong to any two wavelengths in 430-2 400 nm, and $R_i \neq R_j$.

1.4.2　地形因素

地表微气候、土壤中的水分运动以及物质的重新分配进程，都受到地形的影响[25]。在美国地质勘探局网站（http://www.usgs.gov/）下载 DEM 数据，其空间分辨率为 30 m。在 ArcGIS 10.1 中，利用 Spatial Analyst Tools 中的 Extract Multi Values to Points 工具，提取出每个采样点的 DEM，将 DEM 作为模型的输入变量。

1.4.3　竞争性自适应加权算法

土壤高光谱数据量大、存在光谱信息冗余和重叠现象，通过 CARS 算法挑选特征变量可以降低光谱波段之间的高度共线性问题，从而提高预测模型的精度及速度。CARS 算法将各波段变量作为单一个体，在进行个体选择的过程中，保留具有较强适应能力的个体。其具体步骤为：首先，随机抽取固定比率的样本作为校正集建立 PLS 模型，计算回归系数的绝对值和每个波段点对应的权重，然后利用指数衰减函数（Exponentially Decreasing Function，EDP）和自适应重加权采样法（Adaptive Reweighted Sampling，ARS）对变量进行选择，通过交叉验证的方法计算交叉验证均方根误差（Root Mean Square Error of Cross-Validation，RMSECV），N 次蒙特卡罗采样后选择 N 个子集，得到 N 个 RMSECV，选择 RMSECV 最小的波段子集，该子集所包含的变量即为最优变量组合[14,26]。本次试验在 MATLAB 2014a 软件环境中运行 CARS 算法。由蒙特卡罗交叉验证法选择最优潜在波段变量，其中将蒙特卡罗采样次数设定为 100，对采样次数进行反复迭代，通过对比各次采样的 RMSECV 值，当其值最小时，相应采样次数的变量被筛选为最优变量子集。

1.5　模型构建与验证

RF 是基于决策树分类集成算法，其中每一棵树都依赖于一个随机向量，通过对数据集的列变量和行变量观测进行随机化，生成多个分类树，最终将分类树结果进行汇总。RF 对于非线性问题有很好的解释能力，降低了运算量的同时也提高了预测精度[27]。本试验在 R 语言中，利用‘Random Forest’工具包进行预测，在进行拟合前，分别对需要生成树的数量（*ntree*）参数设定为 500，每个节点用于分割节点的预测变量树（*mtry*）参数设定为 1/3 总变量数[28]。

模型构建按照建模集与验证集 2∶1 的比例选取样本。以 CARS 筛选后土壤高光谱反射率数据、DEM 以及光谱指数为自变量，SOM 含量作为因变量，运用 RF，构建 SOM 预测模型。使用调整后决定系数（R^2_{adj}）、均方根误差（RMSE）以及性能与四分位间隔距离的比率（Ratio of Performance to Interquartile distance，RPIQ）为精度评价指标。R^2_{adj} 越大、表明模型越稳定；RMSE 越小、表明模型精度越高；RPIQ 同时考虑了预测误差和观测值的变化，提供了一个更客观、更容易在模型验证研究中进行比较的模型有效性度量。RPIQ 越大，模型的预测能力越强。与残差预测偏差不同，RPIQ 对观测值的分布没有任何假设[29]，其公式如下：

$$\mathrm{RPIQ}=\frac{\mathrm{IQ}}{\mathrm{RMSE}} \tag{1}$$

式中 IQ 是第三和第一个四分位数之间的差值。

2　结果与分析

2.1　SOM 描述统计

土壤样本 SOM 含量统计特征见表 2，质量分数最大值为 11.38%，最小值为 0.98%，土壤样品 SOM 差异较大，这为全面解析 SOM 反射光谱特性研究提供了较完整的样本数据。根据土壤样本 SOM 描述统计表的偏度和峰度值可以判断 SOM 含量数据呈现非正态分布。在 SOM 相关的研究中可知 SOM 质量分数达到 2%以上，对土壤光谱特征起主导作用[30]，SOM 质量分数小于 2%的土壤，其光谱曲线特征易受其他母质等成分的影响，而本次研究中 SOM 平均含量（质量分数）4.5%以上，能够充分说明 SOM 的含量决定了土壤光谱的特征。

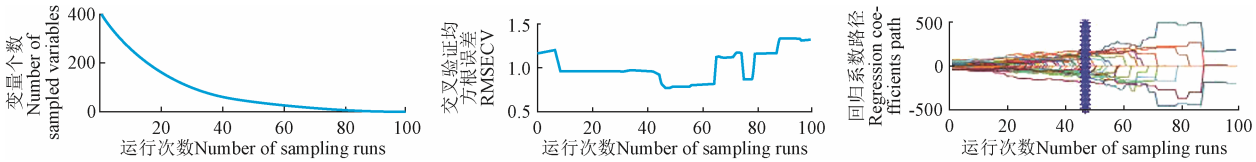
表 2　土壤样本有机质含量统计结果

Table 2　Statistical results of organic matter content in soil samples

土壤类型 Soil type	样本数量 Sample	平均值 Mean/%	最大值 Max/%	最小值 Min/%	标准差 Std/%	峰度 Kurtosis	偏度 Skewness
黑土 Black soil	268	4.55	10.80	0.98	1.38	1.71	0.92
草甸土 Meadow soil	213	4.70	11.27	1.14	1.48	2.24	0.95
沼泽土 Boggy soil	67	5.76	11.38	1.84	2.56	-0.75	0.51
整体 Total	548	4.75	11.38	0.98	1.65	2.06	1.15

2.2 CARS 算法筛选特征波段

3 种土壤类型以及未分类整体在指数衰减函数的作用下，优选变量的数量均随迭代次数的增加呈指数减少，其 RMSECV 值整体均呈现先减后升的趋势。以黑土为例（图 2），从图 2a 可以看出，随着运行次数增加，被优选出的波段变量数逐渐减少，前 5 次采样过程有明显递减，此后逐渐平稳。图 2b 整体上在 1～47 次采样中，RMSECV 值不断降低，表明筛选过程中剔除的变量与 SOM 去除量无关，而 47 次采样迭代以后，RMSECV 值呈回升趋势，表明反射率光谱中与 SOM 无关的大量信息或噪声被添加，从而导致 RMSECV 值上升。图 2c 为所有变量在每次采样过程中的回归系数路径变化图，图中各线表示随着运行次数的增加各波段变量回归系数的变化趋势。结合图 2b 分析发现当采样次数为第 47 次时，RMSECV 值最小即所选择的光谱变量子集最优。草甸土、沼泽土以及未分类整体的 RMSECV 最小值、相应运行次数及特征波段见表 3。



a. 波段变量个数的变化
a. Changes in the number of waveband variables

b. 交叉验证均方根误差的变化
b. Variation of RMSECV

c. 变量回归系数路径
c. Path of variable regression coefficients

注：*所对应点即为 RMSECV 值最低点。
Note: *denotes the optimal point where Root Mean Square Error of Cross-Validation (RMSECV) values achieve the lowest.

图 2 CARS 算法筛选变量
Fig.2 Key variables selected by Competitive Adaptive Reweighted Sampling (CARS) algorithm

表 3 CARS 下基于不同土壤类型的特征波段，运行次数和最小交叉验证均方根误差
Table 3 Characteristic wavebands, number of sampling runs and minimal RMSECV of different soil types under CARS

土壤类型 Soil type	运行次数 Number of sampling runs	最小 RMSECV Minimal RMSECV	特征波段 Characteristic waveband/nm
黑土 Black soil	47	0.770	1 280、1 380、1 450、1 460、1 470、1 480、1 620、1 650、1 690、1 700、1 910、1 940、1 980、2 010、2 050、2 080、2 090、2 100、2 150、2 160、2 170、2 190、2 230
草甸土 Meadow soil	42	0.770	440、530、550、670、1 100、1 140、1 150、1 170、1 240、1 340、1 350、1 370、1 380、1 410、1 420、1 450、1 470、1 650、1 660、1 700、1 720、1 740、1 770、1 790、1 800、1 900、1 920、1 970、2 070、2 150
沼泽土 Boggy soil	49	0.599	430、1 300、1 320、1 430、1 440、1 450、1 460、1 470、1 610、1 620、1 810、1 910、1 930、1 940、1 990、2 010、2 050、2 130、2 150、2 220、2 310
整体 Whole	67	0.763	1 470、1 790、1 800、1 990、2 150、2 170、2 220、2 230、2 280

从表 3 可知，通过 CARS 算法，黑土、草甸土、沼泽土以及整体未分类分别筛选出 23、30、21 和 9 个特征波段，输入波段压缩至全波段数目的 16%以下。黑土特征波段的分布主要在 1 280～2 230 nm 近红外光谱区域，这是由于受到 NH，CH 和 CO 等基团的分子振动的倍频与合频吸收影响[31]，草甸土在可见光-近红外光谱区域均有波段选中，其中 1 700～1 790 nm 处 SOM 响应可能是由氧化铝影响的光谱变化引起的。沼泽土筛选的特征波段在 1 300～2 000 nm 比较均匀分布，这主要是由于沼泽土中的大量三氧化物被还原。值得注意的是，波段 1 450、1 470、2 150 nm 在 3 种土壤类型中均被选择，这是由于 SOM 在 1 400 nm 附近受到土壤黏土矿物质中所含羟基的影响，2 220 nm 附近存在一个与 SOM 相关的烷烃特征峰和存在的氢氧化铝黏土矿物吸收带影响[32]。沼泽土、草甸土筛选的 430、440、530、550、670 nm 少量特征波段位于可见光波段，这是由于受到了土壤发色团和 SOM 本身黑色的影响，可见光波段存在较宽的吸收波段。

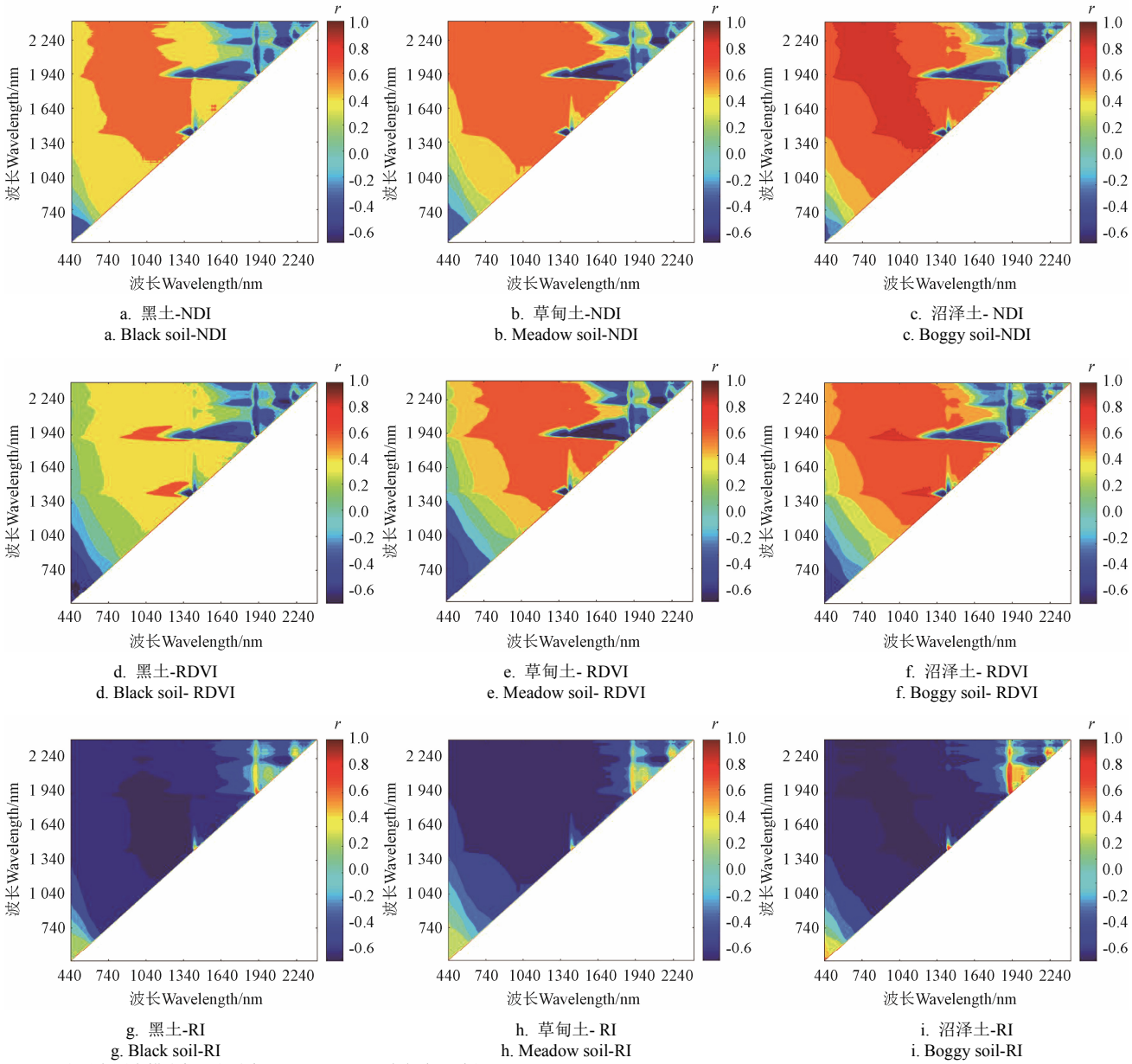
2.3 光谱指数的选取依据

光谱指数通过迭代运算，充分考虑波段之间的协同作用，同时最小化无关波段的影响[33]。研究选取的光谱指数是通过文献查阅，选择可用来估测 SOM 的一系列物理和化学参数的相关光谱指数，并结合本次实际采样点数据进行相关性计算得出。3 种土壤类型原始反射率数据与 SOM 之间的 NDI、RDVI、RI 指数的相关性均较高，且均通过了 P=0.01 水平上的极显著性检验（表 4）。黑土 RI 指数与 SOM 的相关性最高，相关系数为 0.757，草甸土 RDVI 指数与 SOM 的相关性最高，相关系数为−0.784，沼泽土 RDVI 指数与 SOM 的相关性最高，相关系数为 0.922。图 3 是不同土壤类型的 3 种光谱指数与 SOM 含量的二维相关系数矩阵图。3 种土壤类型的 SOM 敏感波段区域主要集中于短波红外部分，主要集中在 1 000、1 900 和 2 200 nm 附近。

表 4 土壤有机质含量与最佳光谱指数的关系
Table 4 Relationship between soil organic matter content and optimal spectral index

土壤类型 Soil type	归一化指数 NDI		再归一化指数 RDVI		比值指数 RI	
	公式 Formula	r	公式 Formula	r	公式 Formula	r
黑土 Black soil	$\frac{R_{1910}-R_{1030}}{R_{1910}+R_{1030}}$	0.714**	$\frac{R_{1950}-R_{1910}}{\sqrt{R_{1950}+R_{1910}}}$	−0.737**	$\frac{R_{1910}}{R_{1940}}$	0.757**
草甸土 Meadow soil	$\frac{R_{2230}-R_{2150}}{R_{2230}+R_{2150}}$	−0.776**	$\frac{R_{2230}-R_{2150}}{\sqrt{R_{2230}+R_{2150}}}$	−0.784**	$\frac{R_{1340}}{R_{1420}}$	−0.779**
沼泽土 Boggy soil	$\frac{R_{1930}-R_{1920}}{R_{1930}+R_{1920}}$	−0.910**	$\frac{R_{1930}-R_{1920}}{\sqrt{R_{1930}+R_{1920}}}$	−0.922**	$\frac{R_{1930}}{R_{1920}}$	0.910**

注：**，相关性在 0.01 水平上显著（双尾）。r 为相关系数。
Note: ** indicates that correlations are significant at 0. 01 level (two-tail). r is the correlation coefficient.



a. 黑土-NDI
a. Black soil-NDI

b. 草甸土-NDI
b. Meadow soil-NDI

c. 沼泽土- NDI
c. Boggy soil-NDI

d. 黑土-RDVI
d. Black soil- RDVI

e. 草甸土- RDVI
e. Meadow soil- RDVI

f. 沼泽土- RDVI
f. Boggy soil- RDVI

g. 黑土-RI
g. Black soil-RI

h. 草甸土- RI
h. Meadow soil-RI

i. 沼泽土-RI
i. Boggy soil-RI

注：R_i、R_j 分别为图中横坐标、纵坐标 430～2 400 nm 中任意一波长。
Note: R_i and R_j are respectively any wavelength in the abscissa and ordinate of 430-2 400 nm.

图 3　不同土壤类型的 3 种光谱指数与 SOM 含量的二维相关系数矩阵

Fig.3　Two dimensional correlation coefficient matrix of three spectral indices and SOM content in different soil types

2.4　SOM 光谱预测模型

由表 5 可知，黑土、草甸土、沼泽土的验证集调整后决定系数依次为 0.678、0.674、0.768，其中沼泽土精度最高，草甸土精度最低，这是由于沼泽土在积水条件下，空气隔绝，微生物活动受到强烈抑制，植物残体不能充分分解，而以粗 SOM 和半腐烂 SOM 的形式积累于地表。全局回归模型 R^2_{adj} 达到 0.742，局部回归模型 R^2_{adj} 达到 0.777。通过局部回归，在一定程度上提高了 SOM 的预测精度。无论是单一土壤类型，还是整体 SOM 预测，其 R^2_{adj} 均达到 0.67 以上，RPIQ 均大于 1.8，表明该模型能较好实现 SOM 预测。

在高光谱 SOM 预测研究中，波段筛选是一个关键方面。本研究通过 CARS 算法筛选波段与已往学者利用相关分析取相关系数大于 0.65 筛选出的波段[34]进行建模比较，研究发现 CARS 算法不仅极大地降低土壤高光谱变量维度和计算复杂程度，验证集 R^2_{adj} 提高了 0.167，精度有一定程度的提升。

表 5　不同土壤类型随机森林预测模型精度

Table 5　Prediction model accuracy of random forest for different soil types

土壤类型 Soil type	建模集 Calibration set		验证集 Validation set		
	R^2	RMSE/%	R^2_{adj}	RMSE/%	RPIQ
黑土 Black soil	0.842	0.560	0.678	0.706	1.936
草甸土 Meadow soil	0.839	0.593	0.674	0.767	1.848
沼泽土 Boggy soil	0.956	0.541	0.768	0.933	3.568

注：R^2_{adj} 为调整后决定系数，RPIQ 为性能与四分位间隔距离的比率。
Note: R^2_{adj} is the adjusted coefficient of determination, and RPIQ is ratio of performance to interquartile distance.

表 6 不同波段筛选方式随机森林预测模型精度

Table 6 Accuracy of random forest prediction model with different band screening methods

方式 Methods	建模集 Calibration set		验证集 Validation set		
	R^2	RMSE/%	R^2_{adj}	RMSE/%	RPIQ
CARS（全局回归） CARS (Global regression)	0.788	0.752	0.742	0.836	1.922
CARS（局部回归） CARS (Local regression)	0.897	0.326	0.777	0.581	2.689
相关分析法（r>0.65） Correlation analysis（r>0.65）	0.834	0.665	0.610	0.839	1.914

3 讨 论

本研究将不同土壤类型（黑土、草甸土、沼泽土）分别进行 SOM 的预测，取得了较高精度。通过土壤分类进行 SOM 预测，消除了不同土壤类型由于“向邻性”导致的反射光谱曲线相似的影响，从而有利于提高预测精度。由于不同类型土壤中矿物成分与 SOM 含量的差异，造成反射光谱间存在显著的区别，通过土壤分类，将有利于提取不同类型土壤光谱参数进行 SOM 预测。陆龙妹等[35]通过全局回归与局部回归进行 SOM 预测比较，依照传统土壤类型建立各自的有机质光谱预测模型精度并不好，这是由于砂姜黑土和黄褐土 2 种土壤类型的黏土矿物都存在蒙脱石且含量较高，SOM 含量接近，所以 2 种土壤类型之间光谱曲线特征相似，造成 SOM 全局回归精度低。而黑土、草甸土、沼泽土之间黏土矿物存在着较大的差异，因此通过全局回归与局部回归比较，全局回归能够提高有效信息的获取程度提高模型精度。其沼泽土的预测精度高于草甸土，这是由于沼泽土土壤湿、土层紧且富有弹性，有机质含量丰富、土体酸碱度从微酸到碱性、土壤颜色深，而草甸土土壤表层砂砾化、有浮沙覆盖、有机质含量较低、土体呈碱性、质地较粗、细粒物质少、土壤色泽浅有一定的关系。

以往许多学者们采用相关分析法研究SOM与土壤光谱反射率（或其不同数学变换形式）的关系，将相关系数高的波段作为 SOM 敏感波段。而后，越来越多的学者采用 CARS 变量优选方法，从全波段中滤除无效变量或冗余变量，优选出敏感波段。本研究基于 CARS 算法，黑土、草甸土、沼泽土分别选择 23、30、21 个特征变量，占全波段数目的 11.6%、15.2%、10.6%，极大地缩减了波段信息，解决了 SOM 预测研究中波段数目多，计算任务繁重的问题。结果表明，CARS 筛选的最优子集存在一定的规律性，波段主要集中在 1 100～2 400 nm 之间，这主要由于受到羰基、酰胺和羟基等基团的分子振动的倍频与合频吸收影响。其中，黑土筛选的特征波段少位于 1 000 nm 以下，这是由于 CARS 是通过利用线性模型偏最小二乘法作为适应度函数，及交叉验证不断优化计算，最终选择出最优子集而不是常用的相关性分析确定特征波段。已有的相关研究表明：SOM 在整个 NIR-SWIR 范围比较敏感，李稳冠等[26]将栗钙土、黑钙土、灰钙土、山地草甸土等土壤光谱曲线通过CARS挑选的特征波段，变量主要分布在 1 900～2 400 nm 的近红外光谱区域，在可见-近红外光谱区域均有分布。CARS 对原始光谱进行特征变量筛选，在保证模型精度的同时显著减少构建模型的变量数。Bao 等[14]对黑土、黑钙土、风沙土、草甸土 4 种土壤类型通过 CARS 算法筛选最优变量子集，其波段大多位于 1 350～2 400 nm 范围内，少量位于 400～1 200 nm。因此，通过 CARS 算法筛选的特征波段，与已有研究 SOM 的反射光谱响应波段相吻合。不同土壤类型通过 CARS 筛选的最优子集也存在差异，其选择的特征变量具有不稳定性。

通过耦合敏感波段的反射率数值进行数学变换所计算得到的光谱指数，避免了由于原始反射率作为输入量所造成的数据冗余，以及产生的共线性问题。黑土筛选出的波段主要为 1 030、1 910、1 940、1 950 nm，草甸土在 1 420、1 340、2 150、2 230 nm，沼泽土集中在 1 920 和 1 930 nm。3 种土壤类型的筛选的波段都位于 NIR-SWIR 范围，这是由于羰基基团的基频振动和其在 NIR-SWIR 范围所对应的酰胺、羟基等基团倍频和合频吸收影响，也与以往的研究一致[36]。因此通过将不同类型土壤分别，以 CARS 筛选的特征波段、DEM 数据和光谱指数作为数据源，建立的 RF 模型能够有效实现 SOM 预测，使精度有着显著的提升。然而，本次研究仍存在不足之处：土壤的光谱反射率还会受到土壤的成土母质、矿物成分、土壤表面粗糙度、粒径、水分等因素的影响，因此，后续研究在原土室外光谱的基础上，将考虑更多的影响因素，加强原土室外光谱 SOM 的估测模型研究，以提升 SOM 的预测精度。

4 结 论

为了解决不同类型土壤预测有机质（Soil Organic Matter，SOM）输入量类型单一造成精度偏低的问题，本文以海伦市 3 种土壤类型（黑土、草甸土、沼泽土）的室内光谱反射率为研究对象，结合数字高程模型（Digital Elevation Model，DEM）以及光谱指数作为输入量，运用随机森林算法（Random Forest，RF）进行 SOM 预测，得出以下结论：

1）通过竞争自适应重加权采样（Competitive Adaptive Reweighted Sampling，CARS）算法，筛选出的特征波段不仅将输入波段压缩至全波段数目的 16%以下，而且能够在很大程度上降低土壤高光谱变量维度和计算复杂程度，从而提高了模型的预测能力。光谱变量经 CARS 算法筛选后模型调整后决定系数提高 0.167，估测效果更好。说明 CARS 算法在提取特征关键波段变量、优化模型结构方面起到关键作用。

2）通过土壤分类进行 SOM 预测，不同土壤类型的 SOM 调整后决定系数存在差异，沼泽土的调整后决定系数最高为 0.768，黑土次之，草甸土的预测精度最低，只有 0.674，运用 RF 对 3 类土壤的 SOM 预测性能与四分位间隔距离的比率均大于 1.8，说明无论是黑土、草甸土还是沼泽土，该模型都有一定的可信度，具有较好的预测能力。

3）通过将 CARS 筛选的特征波段、DEM 以及光谱

指数作为输入量，运用 RF 模型，SOM 的局部回归模型验证集精度最优，调整后决定系数为 0.777，且 RPIQ 达到 2.689，与全局回归模型相比，模型的验证精度提高了 0.035。研究表明，3 种类型的输入量，进行单一土壤类型分别建模和全局回归建模，其均具有较好的预测能力，在一定程度上可为以后不同土壤类型 SOM 预测时输入量的选择提供帮助，从而促进区域不同类型土壤进行 SOM 预测研究的进展，为农业和环境领域 SOM 的动态监测和建模提供理论支撑。

[参 考 文 献]

[1] Gu Xiaohe, Wang Yancang, Sun Qian, et al. Hyperspectral inversion of soil organic matter content in cultivated land based on wavelet transform[J]. Computers and Electronics in Agriculture, 2019, 167: 105053.

[2] Nowkandeh S N, Noroozi A A, Homaee. Estimating soil organic matter content from Hyperion reflectance images using PLSR, PCR, MinR and SWR models in semi-arid regions of Iran[J]. Environmental Development, 2018, 25: 23-32.

[3] 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[J]. 光谱学与光谱分析，2018，38(8)：2556-2562.
Xu Xibo, Lv Jianshu, Wu Quanyuan, et al. Prediction of soil organic matter hyperspectral based on PCA-MLR and PCA-BPN in the coastal plain south of Laizhou Bay[J]. Spectroscopy and Spectral Analysis, 2018, 38(8): 2556-2562. (in Chinese with English abstract)

[4] Meng Xiagtian, Bao Yilin, Liu Jiangui, et al. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data[J]. International Journal of Applied Earth Observations and Geoinformation, 2020, 89: 102111.

[5] Liu Shangshi, Shen Haihua, Chen Songchao, et al. Estimating forest soil organic carbon content using vis-NIR spectroscopy: Implications for large-scale soil carbon spectroscopic assessment[J]. Geoderma, 2019, 348: 37-44.

[6] Wang Xiaoping, Zhang Fei, Ding Jianli, et al. Estimation of soil salt content (SSC) in the Ebinur Lake Wetland National Nature Reserve (ELWNNR), Northwest China, based on a Bootstrap-BP neural network model and optimal spectral indices[J]. Science of the Total Environment, 2018, 615: 918-930.

[7] 朱传梅，王宏卫，谢霞，等. 基于光谱指数和机器学习的土壤有机质含量反演[J]. 江苏农业科学，2020，48(22)：233-241.
Zhu Chuanmei, Wang Hongwei, Xie Xia, et al. Retrieval of soil organic matter content based on spectral index and machine learning[J]. Jiangsu Agricultural Sciences, 2020, 48(22): 233-241. (in Chinese with English abstract)

[8] Wei Lifei, Yuan Ziran, Wang Zhengxiang, et al. Hyperspectral inversion of soil organic matter content based on a combined spectral index model[J]. Sensors, 2020, 20(10): 2777.

[9] 高凤杰，马泉来，韩文文，等. 黑土丘陵区小流域土壤有机质空间变异及分布格局[J]. 环境科学，2016，37(5)：325-332.
Gao Fengjie, Ma Quanlai, Han Wenwen, et al. Spatial variability and distribution pattern of soil organic matter in small watershed of black soil hilly region[J]. Environmental Science, 2016, 37(5): 325-332. (in Chinese with English abstract)

[10] Hong Yongsheng, Chen Songchao, Chen Yiyun, et al. Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoil organic carbon: Feature selection coupled with random forest[J]. Soil & Tillage Research, 2020, 199: 104589.

[11] Vohland M, Ludwig M, Thiele-Bruhn S, et al. Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection[J]. Geoderma, 2014, (223/225): 88-96.

[12] 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]. Journal of Infrared and Millimeter Waves, 2012, 31(3): 277-282. (in Chinese with English abstract)

[13] 卢艳丽，白由路，杨俐苹, 等. 东北平原不同类型土壤有机质含量高光谱反演模型同质性研究[J]. 植物营养与肥料学报，2011，17(2)：456-463.
Lu Yanli, Bai Youlu, Yang Liping, et al. Homogeneity of retrieval models for soil organic matter of different soil types in Northeast Plain using hyperspectral data[J]. Journal of Plant Nutrition and Fertilizers, 2011, 17(2): 456-463. (in Chinese with English abstract)

[14] Bao Yilin, Meng Xiangtian, Ustin Susan, et al. Vis-SWIR spectral prediction model for soil organic matter with different grouping strategies[J]. Catena, 2020, 195: 104703.

[15] 吴才武，夏建新，段峥嵘. 土壤有机质测定方法述评与展望[J]. 土壤，2015，47(3)：453-460.
Wu Caiwu, Xia Jianxin，Duan Zhengrong. Review and prospect of methods for determination of soil organic matter[J]. Soils, 2015, 47(3): 453-460. (in Chinese with English abstract)

[16] 刘焕军，孟祥添，王翔，等. 反射光谱特征的土壤分类模型[J]. 光谱学与光谱分析，2019，39(8)：2481-2485.
Liu Huanjun, Meng Xiangtian, Wang Xiang, et al. Soil classification model based on the characteristics of soil reflectance spectrum[J]. Spectroscopy and Spectral Analysis, 2019, 39(8): 2481-2485. (in Chinese with English abstract)

[17] 中国土壤学会. 土壤农业化学分析方法[M]. 北京：中国农业科技出版社，2000.

[18] Zhang Xiaokang, Liu Huanjun, Zhang Xinle, et al. Allocate soil individuals to soil classes with topsoil spectral characteristics and decision trees[J]. Geoderma, 2018, 320: 12-22.

[19] Vasques G M, Grunwald S, Sickman J O. Comparison of multivariate methods for inferential modeling of soil carbon

using visible/near- infrared spectra[J]. Geoderma, 2008, 146(1): 14-25.

[20] Jin Xiuliang, Du Jia, Liu Huanjun, et al. Remote estimation of soil organic matter content in the Sanjiang Plain, Northest China: The optimal band algorithm versus the GRA-ANN model[J]. Agricultural and Forest Meteorology, 2016, 218/219: 250-260.

[21] Hong Yongsheng, Chen Songchao, Zhang Yong, et al. Rapid identification of soil organic matter level via visible and near-infrared spectroscopy: Effects of two-dimensional correlation coefficient and extreme learning machine[J]. Science of the Total Environment, 2018, 644: 1232-1243.

[22] Ihuoma S O, Madramootoo C A. Narrow-band reflectance indices for mapping the combined effects of water and nitrogen stress in field grown tomato crops[J]. Biosystems Engineering, 2020, 192: 133-143.

[23] 郑曼迪，熊黑钢，乔娟峰，等. 基于宽波段与窄波段综合光谱指数的土壤有机质遥感反演[J]. 激光与光电子学进展，2018，55(7)：457-465.
Zhen Mandi, Xiong Heigang, Qiao Juanfeng, et al. Remote sensing inversion of soil organic matter based on broad band and narrow band comprehensive spectral index[J]. Laser & Optoelectronics Progress, 2018, 55(7): 457-465. (in Chinese with English abstract).

[24] Jin Xiuliang, Song Kaishan, Du Jia, et al. Comparison of different satellite bands and vegetation indices for estimation of soil organic matter based on simulated spectral configuration[J]. Agricultural and Forest Meteorology, 2017, (244/245): 57-71.

[25] 刘焕军，鲍依临，徐梦园，等. 基于 SOM 和 NDVI 的黑土区精准管理分区对比[J]. 农业工程学报，2019，35(13)：177-183.
Liu Huanjun, Bao Yilin, Xu Mengyuan, et al. Comparison of precision management zoning methods in black soil area based on SOM and NDVI[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2019, 35(13): 177-183. (in Chinese with English abstract)

[26] 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.
Li Wenguan, Gao Xiaohong, Xiao Nengwen, et al. Estimation soil organic matter contents with hyperspectra based on sCARS and RF algorithms[J]. Journal of luminescence, 2019, 40(8): 1030-1039. (in Chinese with English abstract)

[27] 包青岭，丁建丽，王敬哲，等. 基于随机森林算法的土壤有机质含量高光谱检测[J]. 干旱区地理，2019，42(6)：1404-1414.
Bao Qingling, Ding Jianli, Wang Jingzhe, et al. Hyperspectral detection of soil organic matter content based on random forest algorithm[J]. Arid Land Geography, 2019, 42(6): 1404-1414. (in Chinese with English abstract)

[28] 姜赛平，张怀志，张认连，等. 基于三种空间预测模型的海南岛土壤有机质空间分布研究[J]. 土壤学报，2018，55(4)：1007-1017.
Jiang Saiping, Zhang Huaizhi, Zhang Renlian, et al. Research on spatial distribution of soil organic matter in Hainan Island based on three spatial prediction models[J]. Acta Pedologica Sinica, 2018, 55(4): 1007-1017. (in Chinese with English abstract)

[29] Guo Long, Zhang Haitao, Shi Tiezhu, et al. Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images[J]. Geoderma, 2019, 337: 32-41.

[30] 刘焕军，赵春江，王纪华，等. 黑土典型区土壤有机质遥感反演[J]. 农业工程学报，2011，27(8)：211-215.
Liu Huanjun, Zhao Chunjiang, Wang Jihua, et al. Soil organic matter predicting with remote sensing image in typical blacksoil area of Northeast China[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2011, 27(8): 211-215. (in Chinese with English abstract)

[31] 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 & Tillage Research, 2016, 155: 510-522.

[32] 史舟，王乾龙，彭杰，等. 中国主要土壤高光谱反射特性分类与有机质光谱预测模型[J]. 中国科学：地球科学，2014，44(5)：978-988.
Shi Zhou, Wang Qianlong, Peng Jie, et al. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations[J]. Science China: Earth Sciences, 2014, 44(5): 978-988. (in Chinese with English abstract)

[33] 张子鹏，丁建丽，王敬哲，等. 利用三维光谱指数定量估算土壤有机质含量：以新疆艾比湖流域为例[J]. 光谱学与光谱分析，2020，40(5)：1514-1522.
Zhang Zipeng, Ding Jianli, Wang Jingzhe, et al. Quantitative estimation of soil organic matter content using three-dimensional spectral index: A case study of the Ebinur Lake Basin in Xinjiang[J]. Spectroscopy and Spectral Analysis, 2020, 40(5): 1514-1522. (in Chinese with English abstract)

[34] Abergaz A, Winowieckil A, Vagen T G, et al. Spatial and temporal dynamics of soil organic carbon in landscapes of the upper Blue Nile Basin of the Ethiopian Highlands[J]. Agriculture Ecosystems & Environment, 2016, 218: 190-208.

[35] 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.
Lu Longmei, Zhang Ping, Lu Hongliang, et al. Hyperspectral characteristics of soils in Huaibei Plain and estimation of SOM content[J]. Soils, 2019, 51(2): 374-380. (in Chinese with English abstract)

[36] 卢牧原，刘源，刘桂建. 基于组合模型的黑土区土壤有机质含量预测分析[J]. 浙江农业学报，2020，32(8)：1427-1436.

Lu Muyuan, Liu Yuan, Liu Guijian. Predictive analysis of soil organic matter content in black soil region based on combined model[J]. Acta Agriculturae Zhejiangensis, 2020, 32(8): 1427-1436. (in Chinese with English abstract)

Hyperspectral prediction on soil organic matter of different types using CARS algorithm

Tang Haitao[1], Meng Xiangtian[1], Su Xunxin[2], Ma Tao[3], Liu Huanjun[1,4], Bao Yilin[1], Zhang Meiwei[1], Zhang Xinle[1※], Huo Haizhi[3]

(1. *School of Public Administration and Law*, *Northeast Agricultural University*, *Harbin* 150030, *China*; 2. *Heilongjiang Provincial Geological Archive*, *Harbin* 150030, *China*; 3. *Heilongjiang Fifth Geological Survey Institute*, *Harbin* 150030, *China*; 4. *Northeast Institute of Geography and Agroecology*, *Chinese Academy of Sciences*, *Changchun* 130012, *China*)

Abstract: Soil organic matter (SOM) can improve the physical, chemical and biological properties of the soil through a variety of functions. An important role of SOM is performed on the soil function and quality, further to prevent the emission of greenhouse gas in global carbon circulation. Spectral characteristics of SOM depend mainly on types of soils, as well as different physical and chemical properties. Previous models constructed by the hyperspectral reflectance or spectral absorption characteristics often lead to the low accuracy in SOM prediction, due mainly to the input type structure was single. In order to improve the accuracy and speed of the prediction model, specific characteristic variables can be selected to reduce the high collinearity between spectral bands, where there is a large amount of hyperspectral data in the presence of redundancy and overlap. The spectral index is set to minimize the influence of independent wavelengths on iterative calculation. Furthermore, the topography significantly determines the surface microclimate, the movement of water on the surface and in the soil, as well as the process of material redistribution. In this study, taking the Hailun City, Heilongjiang Province as the research area, a SOM prediction random forest (RF) model was established for the different types of soil, in order to improve the accuracy of the SOM hyperspectral model. The characteristic bands were selected by a Competitive Adaptive Reweighted Sampling (CARS), while, the Digital Elevation Model (DEM) data and spectral index were data sources. The results showed that: 1) In CARS screening, the characteristic bands of each soil type were compressed to less than 16% of the total wavelength number, which greatly reduced the dimension of soil hyperspectral variables and computational complexity, thereby improving the prediction ability of the calibration model. The CARS was suitable for the extraction of characteristic key wavelength variables, further optimizing model structure. 2) Three types of input variables that extracted by the grouping experiment were then utilized for the prediction of different types of SOM. After grouping, the SOM prediction accuracy depended mainly on the type of soil. Specifically, the maximum prediction accuracy achieved in the Boggy soil of 0.768, where the Ratio of Performance to Interquartile distance (RPIQ) was 3.568. Black soil was the second most accurate. The prediction accuracy of meadow soil was the lowest, only 0.674, and RPIQ was 1.848. The RPIQ for the three types of soil was above 1.8, indicating the good prediction ability of the model. 3) Local regression was conducted to improve the prediction accuracy of SOM. The local regression prediction accuracy was the best. The adjusted coefficient of determination, RMSE and RPIQ of the validation set were 0.777, 0.581%, and 2.689, respectively, indicating the model was highly stable. The proposed prediction factors can be used to realize the rapid prediction of RF-SOM, where the traditional complex program can be simplified. The findings can provide a promising basis for the selection of input variables, thereby predicting the types of SOM in different regions.

Keywords: remote sensing; soils; organic matter; spectral index; terrain; characteristic band screening; random forest