第40卷，第4期
2020年4月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 40, No. 4, pp1082-1086
April, 2020

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

吕美蓉[1]，任国兴[1,2]，李雪莹[1]，范萍萍[1]，孙中梁[1]，侯广利[1]，刘　岩[1*]

1. 齐鲁工业大学(山东省科学院)海洋仪器仪表研究所，山东省海洋监测仪器装备技术重点实验室，国家海洋监测设备工程技术研究中心，山东 青岛　266100
2. 中国海洋大学信息科学与工程学院，山东 青岛　266100

摘　要　可见-近红外光谱已被证明是一种快速、有效的有机碳(TOC)含量预测方法。但是，当前利用光谱预测TOC含量的研究对象主要为土壤或湖泊沉积物，还未见潮间带海洋沉积物的研究报道。为了快速准确预测潮间带沉积物TOC含量，通过异常样本剔除、光谱特征变换、特征波长提取相结合，构建TOC预测模型，即，采集潮间带沉积物样品光谱，采用马氏距离、标准杠杆值和学生残差联合分析的方法剔除异常样本，利用多元散射校正(MSC)、平滑＋微分进行光谱变换，利用遗传算法(GA)提取特征波长，采用偏最小二乘法(PLS)、最小二乘支持向量机(LSSVM)和BP神经网络(BPNN)对沉积物TOC含量进行建模和预测，通过决定系数(R^2)和剩余估计偏差(PRD)来评价模型精度。结果表明，剔除异常样本有助于提升模型精度，BPNN模型的检验R^2和PRD分别提升了28%和39%。MSC光谱变换效果优于平滑＋微分，基于MSC光谱变换的PLS，LSSVM和BPNN模型检验R^2分别为0.81，0.86和0.78，PRD分别为2.25，2.59和2.07，比平滑＋微分提升了9%～20%(R^2)和11%～22%(PRD)，意味着MSC具有较强的TOC信息提取能力。GA不利于增加预测模型精度，基于GA特征波长的模型预测R^2降低了9%～36%，PRD降低了18%～33%，可能与GA提取的特征波长数量偏少有关。BPNN模型的预测精度最低，可能与其容易陷入局部极小点有关。PLS模型精度较高，可以很好的预测潮间带沉积物TOC含量。基于异常样本剔除和MSC光谱变换，PLS模型的建模R^2为0.98，检验R^2为0.81，RPD为2.25。LSSVM模型精度更优于PLS，LSSVM模型建模R^2为0.99，检验R^2和RPD分别为0.86和2.59，显示极好的TOC定量预测能力。总之，针对潮间带沉积物TOC含量预测，可以将剔除异常样本、MSC光谱变换、LSSVM建模结合起来，以获得可靠、稳定的预测模型。

关键词　潮间带沉积物；可见-近红外光谱；预测模型；有机碳含量

中图分类号：O657.3　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2020)04-1082-05

引　言

可见-近红外光谱速测是利用物质在近红外光谱区内的光学特性快速反演物质组成和化学成分含量的一种方法。当前国内外已开展了不少的沉积物/土壤碳光谱速测技术研究。Alaoui等采用偏最小二乘法(PLS)建立沉积物光谱原数据和碳含量的对应关系模型，可以较好的反演沉积物碳含量[1]。章海亮等采用遗传算法结合连续投影算法提取特征波长，应用偏最小二乘回归方法建立土壤有机质模型，预测R^2为0.83[2]。申艳等采用多元散射校正和多元线性回归法建立了土壤有机碳光谱模型，预测值与实测值的相关系数为0.82[3]。

潮间带是海陆相互作用的一个重要界面，沉积物中的碳含量是海洋污染程度的标志之一[4]。尽管采用可见-近红外光谱快速预测土壤/湖泊沉积物TOC含量已成为热点，但未见使用可见-近红外光谱预测潮间带沉积物TOC含量。潮间带沉积物和湖泊沉积物在粒度、有机碳含量、盐含量等方面都存在很大差异，这些都会对预测模型产生很大的影响。此外，在建模方面，以往多采用PLS、多元线性回归等方法，

收稿日期：2019-03-18，**修订日期**：2019-07-04

基金项目：国家重点研发计划项目(2016YFC1400800，2017YFC1403700)，山东省自然科学基金项目(ZR201801300002，ZR20182B0523，ZR2018LD007，ZR2016DM17，ZR2014YL006)资助

作者简介：吕美蓉，女，1983年生，齐鲁工业大学(山东省科学院)海洋仪器仪表研究所助理研究员　　e-mail：444868063@qq.com

*通讯联系人　　e-mail：sdqdliuyan@126.com

存在自相关、过适应性等问题。最小二乘支持向量机(LSSVM)是基于结构风险最小化原理和学习理论的一种方法，通过不断优化调整，找到最优函数。BP 神经网络(BPNN)是一种多层前馈神经网络，根据预测误差调整网络权值和阈值，使 BPNN 预测输出不断逼近期望输出。从原理上看，LSSVM 和 BPNN 都可以保证在在测试中能够达到非常高的精度，但是，这两种方法在 TOC 预测上应用相对较少。

为此，对潮间带海洋沉积物样品进行光谱测量，采用马氏距离、标准杠杆值＋学生残差联合分析的方法剔除异常样本，用平滑＋微分、多元散射校正(MSC)进行光谱预处理、遗传算法(GA)提取特征波长、KS 方法进行样本分类，采用 PLS、LSSVM 和 BPNN 对沉积物 TOC 进行建模和预测，以期找到适合的光谱变换方法和特征波长，优化光谱模型，实现潮间带沉积物有机碳的快速、定量检测。

1　实验部分

1.1　沉积物样品采集

于青岛海洋潮间带采集了 197 份沉积物样品，采样深度为 0～10 cm。每个采样点之间至少间隔 10 m。潮间带沉积物主要由黏土和粉砂组成。将采集的样品自然风干，研磨、过 60 目筛，混匀，分成两份，分别用于光谱采集和化学分析。

1.2　分析测定方法

沉积物碳含量采用重铬酸钾氧化法测定，以此作为光谱建模的标准值。光谱反射率测定采用海洋光学 QE65000 光谱仪，光谱采样间隔为 1 nm，积分时间 600 ms，谱区范围 200～1 100 nm。取 3～5 g 沉积物样品放在自制样品盒中，轻轻刮平，用 45°视场角光纤探头采集光谱。每个土壤样品采集 5 次光谱反射率，取平均值。为减少噪声影响，剔除信噪比较低的边缘波段，保留 230～970 nm 的光谱数据。

1.3　数据处理

1.3.1　光谱预处理

采用马氏距离、标准杠杆值和学生残差联合分析的方法剔除异常样本。分别用多元散射校正(MSC)、平滑＋一阶微分对沉积物光谱进行变换。采用遗传算法(GA)进行特征波长提取。

1.3.2　模型建立

采用 KS 方法对 197 个沉积物样品进行分类。采用偏最小二乘算法(PLS)、最小二乘支持向量机法(LSSVM)和 BP 神经网络法(BPNN)进行建模。偏最小二乘回归算法是将相关分析、多元线性回归和主成分的优点集合在一起，在计算过程中同时考虑自变量(光谱数据)和因变量矩阵(化学参考值)对建模效果的影响，能够较好地处理数据多重共线性、因子结果不确定性和数据非正态分布等问题[5]。最小二乘支持向量机是通过非线性映射函数建立回归模型，将输入变量映射到高维特征空间；然后将优化问题改成等式约束条件[6]。BP 神经网络法是是一种按误差逆传播算法训练的多层前馈网络，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小[7]。

1.3.3　模型检验

模型检验采用决定系数(R^2)和剩余估计偏差(RPD)为评价参数。当 $R^2>0.90$ 表示预测结果出色，$0.81<R^2<0.90$ 表示预测结果很好，$0.66<R^2<0.80$ 表示预测结果一般，$R^2<0.66$ 表示预测结果很差。此外，当 RPD＜1.0 时，表明模型预测能力很差，模型不可靠；当 1.0＜RPD＜1.4 时，表明模型预测能力较差；当 1.4＜RPD＜1.8 时，表明模型预测能力较好，可以对样本进行估测；当 1.8＜RPD＜2.0 时，表明模型预测能力好，可对样本进行定量估测，当 2.0＜RPD＜2.5 时，表明模型具有很好的定量预测能力；当 RPD＜2.5 时，模型具有极好的预测能力[7]。

2　结果与讨论

2.1　潮间带沉积物光谱特征

从图 1 中可以看出，光谱曲线在 230～600 nm 范围内呈快速上升态势，然后在 600～970 nm 范围内趋于平缓，与王哲等报道的湖泊沉积物的反射光谱在 650～700 nm 波段有明显的波谷[8]显然是不同的。一般认为，有机碳含量与光谱反射率成负相关，有机碳含量越高，光谱反射率越低。此外，基线漂移也是影响光谱曲线的一个重要因素，图中上部和下部的光谱出现了明显的分离，也有可能是沉积物样品颗粒大小差异等造成了光谱基线漂移。

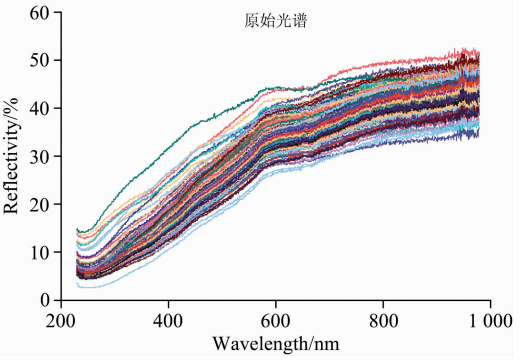

图 1　潮间带沉积物反射光谱

Fig. 1　The reflection spectra of intertidal sediment

2.2　异常样本剔除

采用马氏距离、标准杠杆值和学生残差联合分析方法辨别异常值，剔除了 8 个异常样本。基于剔除异常样本后的光谱数据，采用 PLS，LSSVM 和 BPNN 方法建模，结果发现(表 1)，剔除异常样本对 PLS 和 LSSVM 模型精度影响较小，但能够增加 BPNN 模型精度，检验 R^2 从 0.57 增加到 0.73，RPD 从 1.39 增加到 1.93，即，剔除异常样本很好的提升了 BPNN 模型精度。

2.3　光谱数据变换

在剔除异常样本的基础上，采用 MSC 和平滑＋微分方法进行光谱变换，然后对比光谱变换后的模型精度(表 2)。结果发现，MSC 增加了模型精度，PLS 模型的预测 R^2 从 0.74 上升到 0.81，RPD 从 1.93 上升到 2.25；LSSVM 模型的预测 R^2 从 0.74 上升到 0.86，RPD 从 1.92 上升到 2.59。而平滑＋微分预处理对模型精度影响较小。光谱变换是提升

模型精度的重要手段[9]。因此，选择适当的方法进行潮间带沉积物光谱变换很重要。

表 1 异常样本剔除对模型精度的影响

Table 1 Effect of abnormal sample elimination on model accuracy

异常样本剔除	建模方法	检验集	
		R^2	RPD
无	PLS	0.77	2.04
无	LSSVM	0.74	1.95
无	BPNN	0.57	1.39
剔除	PLS	0.74	1.93
剔除	LSSVM	0.74	1.92
剔除	BPNN	0.73	1.93

表 2 光谱变换对模型精度的影响

Table 2 Effect of spectral transformation on model accuracy

预处理	建模方法	检验集	
		R^2	RPD
无	PLS	0.74	1.93
MSC	PLS	0.81	2.25
平滑＋微分	PLS	0.76	2.02
无	LSSVM	0.74	1.92
MSC	LSSVM	0.86	2.59
平滑＋微分	LSSVM	0.79	2.18
无	BPNN	0.73	1.93
MSC	BPNN	0.78	2.07
平滑＋微分	BPNN	0.65	1.69

2.4 特征波长提取

在剔除异常样本和 MSC 光谱变换的基础上，采用 GA 方法进行特征波长提取，并基于该特征波长进行建模。结果发现(表 3)，GA 降低了模型精度，尤其是 BPNN 模型精度。基于全波长的 BPNN 模型可以很好地定量预测潮间带沉积物 TOC 含量($R^2=0.86$，RPD=2.59)，而基于 GA 特征波长的 BPNN 模型仅能对沉积物 TOC 进行粗略估测。这些暗示着 GA 可能不是潮间带沉积物有机碳特征波长的有效提取方法，这可能是由于 GA 提取的特征波长数量少(表 4)，所包含的有用信息少，不能很好地表征有机碳含量。

表 3 特征波长提取对模型精度的影响

Table 3 Effect of feature wavelength extraction on model accuracy

特征波长提取	建模方法	检验集	
		R^2	RPD
无	PLS	0.81	2.25
GA	PLS	0.72	1.85
无	LSSVM	0.86	2.59
GA	LSSVM	0.78	2.09
无	BPNN	0.78	2.07
GA	BPNN	0.50	1.39

表 4 提取的潮间带沉积物碳特征波长

Table 4 Extracted characteristic wavelengths of intertidal sediment carbon

提取方法	提取的波长					
GA	498.95	499.73	518.39	519.16	519.94	535.45
	536.23	537.00	537.77	566.38	567.15	577.95
	578.72	589.50	590.27	591.04	714.66	715.41
	716.17	720.71	721.47	722.22	725.25	726.00
	730.54	731.29	732.05	732.80	733.56	734.31
	741.10	741.86	742.61	744.12	744.87	745.63
	753.92	754.67	755.42	756.17	777.97	778.72
	779.47	896.35	897.08	914.78	915.52	916.26
	961.77	962.50	963.23			

2.5 建模

在剔除异常样本和 MSC 预处理的基础上，采用 PLS，LSSVM 和 BPNN 方法进行建模。结果表明(表 5)，LSSVM 模型具有高的建模集决定系数($R^2=0.99$)、检验集决定系数($R^2=0.86$)和剩余估计偏差(RPD=2.59)，指示着 LSSVM 模型是预测潮间带沉积物 TOC 含量的优势模型。PLS 模型效果次之，PLS 模型的建模集 R^2 为 0.98、检验集 R^2 为 0.81、RPD 为 2.25，这些指示着线性模型也具有较好的定量预测能力。而 BPNN 模型的建模效果和预测能力最差，建模集 R^2、检验集 R^2 以及 RPD 分别为 0.90，0.78 和 2.07，这可能是在训练过程中出现了过拟合现象。

表 5 PLS，LSSVM 和 BPNN 模型精度评价

Table 5 Accuracy evaluation of PLS，LSSVM and BPNN models

建模方法	建模集	检验集	
	R^2	R^2	RPD
PLS	0.98	0.81	2.25
LSSVM	0.99	0.86	2.59
BPNN	0.90	0.78	2.07

我们的研究结果表明，MSC 光谱变换提升了预测模型精度，这可能是 MSC 降低了光谱变量之间的信息冗余，突出光谱与 TOC 含量之间的关联。崔霞等认为微分能较好地消除母质等潜在因素对光谱的影响，使一些原本被遮蔽的 TOC 光谱特征显现出来[10]。但我们的结果表明，平滑＋微分对模型精度影响不大，暗示着该方法不适合海岸带沉积物光谱变换，这可能是微分处理在消除基线和其他背景干扰的同时扩大了噪声的作用[11]。

筛选特征波长可去除不相关的光谱信息，简化模型，提高预测精度与稳定性[12]。我们采用 GA 算法提取特征波长，共提取了 22 个特征波长，分布于 692～970 nm 之间(数据未列出)。这与纪文君等认为的有机碳含量敏感波段(600～800 nm)[13]仅部分重叠，这可能是研究对象不同所造成的，研究对象的成分组成、物理结构、颜色等都会影响到反射光谱。但是，我们基于该特征波长建模，发现模型精度不增反降。推测可能的原因是：(1)沉积物成分复杂，可能会掩盖部分

TOC 光谱信息，导致这部分有用的光谱波段在特征波长提取过程被滤掉。(2)沉积物中 TOC 成分复杂，难以仅用某些特征波长的光谱信息来表征。(3)GA 可能对反馈信息利用不充分，当求解到一定范围时，做了大量冗余迭代[14]。

PLS 是目前比较常用的一种线性光谱模型建立方法。卢延年等认为 PLS 保证了主成分与 TOC 相关，是全谱在 TOC 方向上的投影，能够很好的解决光谱之间的多重共线性问题[15]。我们的研究显示，PLS 能够很好的预测潮间带沉积物 TOC 含量，即预测 R^2 为 0.81；PRD 为 2.25。但是，LSSVM 更有优势，这暗示着非线性模型更适合海岸带沉积物 TOC 预测。这可能是因为沉积物有机碳组成复杂，且受到外在环境的干扰，与光谱反射率之间呈非线性关系。此外，LSSVM 泛化能力强，有助于精确预测 TOC 含量。而 BPNN 效果最差，可能是神经网络容易陷入局部极小点。

3　结　论

光谱定量快速监测潮间带沉积物碳含量具有重要的意义。光谱测量结果表明，潮间带沉积物成分含量与湖泊沉积物不同，因而预测 TOC 含量的模型也有所差异。采用剔除异常样本＋MSC 光谱变换＋LSSVM 建模，建模集 R^2 达到 0.99，检验集 R^2 为 0.86，RPD 为 2.59；该方法可以很好的预测潮间带沉积物碳含量。

References

[1] Alaoui G, Léger M N, Gagné J, et al. Chemical Geology, 2011, 286: 290.

[2] ZHANG Hai-liang, LUO Wei, LIU Xue-mei, et al(章海亮, 罗　微, 刘雪梅, 等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2017, 37(2): 584.

[3] SHEN Yan, ZHANG Xiao-ping, LIANG Ai-zhen, et al(申　艳, 张晓平, 梁爱珍, 等). System Sciemces and Comprehensive Studies in Agriculture(农业系统科学与综合研究), 2010, 2: 174.

[4] YIN Sheng-le, LIU Xiao-shou, YUAN Chao, et al(尹盛乐, 刘晓收, 袁　超, 等). Transactions of Oceanology and Limnology(海洋湖沼通报), 2012, 2: 97.

[5] GAO Sheng, WANG Qiao-hua, LI Qing-xu, et al(高　升, 王巧华, 李庆旭, 等). Chinese Journal of Analytical Chemistry (分析化学), 2019, 47(6): 941.

[6] LIU Xue-mei, ZHANG Hai-liang(刘雪梅, 章海亮). Journal of Irrigation and Drainage(灌溉排水学报), 2013, 32: 138.

[7] DONG Zhe, YANG Wu-de, ZHU Hong-fen, et al(董　哲, 杨武德, 朱洪芬, 等). Journal of Shanxi Agricultural Sciences(山西农业科学), 2019, 47(5): 751.

[8] WANG Zhe, NIE Ya-guang, CHEN Qian-qian, et al(王　哲, 聂亚光, 陈倩倩, 等). Polar Science(极地研究), 2016, 28: 317.

[9] ZHANG Rui, LI Zhao-fu, PAN Jian-jun(张　锐, 李兆富, 潘剑君). Transactions of the Chinese Society of Agricultural Engineering (农业工程学报), 2017, 33: 175.

[10] CUI Xia, SONG Qing-jie, ZHANG Yao-yao, et al(崔　霞, 宋清洁, 张瑶瑶, 等). Acta Prataculturae Sinica(草业学报), 2017, 10: 20.

[11] SHEN Yan, ZHANG Xiao-ping, YANG Xue-ming, et al(申　艳, 张晓平, 杨学明, 等). Acta Pedologica Sinica(土壤学报), 2010, 5: 1006.

[12] YANG Hong-fei, ZHENG Li-ming, HAO Zhong-yao, et al(杨红飞, 郑黎明, 郝中要, 等). Journal of Anhui Agricultural University(安徽农业大学学报), 2018, 45(1): 101.

[13] JI Wen-jun, SHI Zhou, ZHOU Qing, et al(纪文君, 史　舟, 周　清, 等). Journal of Infrared and Millimeter Waves(红外与毫米波学报), 2012, 31(3): 277.

[14] QIAO Tian, LÜ Cheng-wen, XIAO Wen-ping, et al(乔　天, 吕成文, 肖文凭, 等). Chinese Journal of Soil Science(土壤通报), 2018, 49: 773.

[15] LU Yan-nian, LIU Yan-fang, CHEN Yi-yun, et al(卢延年, 刘艳芳, 陈奕云, 等). Chinese Agricultural Science Bulletin(中国农学通报), 2014, 30(26): 127.

Prediction of Organic Carbon Content of Intertidal Sediments Based on Visible-Near Infrared Spectroscopy

LÜ Mei-rong[1], REN Guo-xing[1,2], LI Xue-ying[1], FAN Ping-ping[1], SUN Zhong-liang[1], HOU Guang-li[1], LIU Yan[1*]

1. Institute of Oceanographic Instrmentation, Qilu University of Technology (Shandong Academy of Sciences), Shandong Provincial Key Laboratory of Marine Monitoring Instrument Equipment Technology, National Engineering and Technological Research Center of Marine Monitoring Equipment, Qingdao 266100, China
2. School of Information Science and Engineering, Ocean University of China, Qingdao 266100, China

Abstract Visible-near infrared spectroscopy has been shown to be a fast and effective tool for organic carbon (TOC) content prediction. However, the research target of using spectral prediction of TOC content is mainly soil or lake sediment, and there is little research on marine sediments in intertidal zone. In order to predict the content of TOC in intertidal sediments quickly and accurately, this study constructed TOC prediction model by combining abnormal sample elimination, spectral feature transformation and feature wavelength extraction, that is, collecting sediment spectra of samples in intertidal zone, using Markov distance, standard lever value and student residuals combined analysis method to remove abnormal samples, using multivariate scattering correction (MSC), smoothing + differential for spectral transformation, using genetic algorithm (GA) to extract characteristic wavelengths, using partial least squares method (PLS), least squares support vector machine (LSSVM) and BP Neural Network (BPNN) to model and predict sediment TOC content, using the decision coefficient (R^2) and residual estimation deviation (PRD) to evaluate model accuracy. The results showed that the elimination of abnormal samples improved model accuracy, and the test R^2 and PRD of the BPNN model increased by 28% and 39% respectively. The effect of MSC was better than that of smoothing+differential, and the test R^2 and PRD of PLS, LSSVM and BPNN models based on MSC were 0.81, 0.86, 0.78 and 2.25, 2.59, 2.07, respectively, which enhanced 9%～20% (R^2) and 11%～22% (PRD) than that based on smoothing+differential, suggesting that MSC has a strong ability to extract TOC information. GA is not conducive to increasing model accuracy, the test R^2 and PRD of models based on GA reducedby 9%～36% and 18%～33%, respectively. This may be related to the low number of characteristic wavelengths extracted by GA. The BPNN model has the lowest predictive accuracy and may be related to its vulnerability to local minimums. PLS model has high accuracy and can predict TOC content in intertidal zone. Basing on abnormal sample elimination and MSC, the modeling set R^2 of PLS model was 0.98, and the prediction set R^2 and RPD were 0.81 and 2.25 respectively. The accuracy of LSSVM model was better than that of PLS, the modeling set R^2 was 0.99, the test set R^2 and RPD were 0.86 and 2.59 respectively, implying excellent TOC quantitative prediction ability of LSSVM. In a word, for the prediction of TOC content in intertidal sediments, the combination of abnormal sample elimination, MSC spectral transformation and LSSVM modeling can obtain a reliable and stable prediction model.

Keywords Intertidal sediment; Visible-near infrared spectroscopy; Predictive model; Organic carbon content

(Received Mar. 18, 2019; accepted Jul. 4, 2019)

* Corresponding author