第42卷，第11期
2022年11月

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

Vol. 42, No. 11, pp3423-3427
November, 2022

地表水总有机碳含量紫外-可见光谱检测方法

李庆波[1]，毕智棋[1]，崔厚欣[2]，郎嘉晔[2]，申中凯[2]

1. 北京航空航天大学仪器科学与光电工程学院，精密光机电一体化技术教育部重点实验室，北京 100191
2. 河北先河环保科技股份有限公司，河北 石家庄 050035

摘 要 总有机碳是以碳含量评价水质有机污染的指标，可以反映水体受污染程度。目前地表水总有机碳检测多采用现场取样后实验室分析检测方法，该方法存在费时费力、操作复杂、二次化学污染等缺点。紫外-可见光谱法具有环保、操作简便、可实时在线原位检测等优点，在地表水总有机碳检测中具有很好的应用前景。针对总有机碳检测问题，采用了一种基于自适应增强学习的区间偏最小二乘回归方法，该方法将总有机碳吸收光谱波段分为若干子区间，初始化训练样本权重，依次在各子区间建立偏最小二乘回归模型，根据子区间模型预测误差率计算该子区间预测结果的权重系数，并更新下一子区间训练样本权重，最后将各子区间模型预测结果线性加权得到总有机碳的检测结果。实验配制总有机碳标准溶液浓度25～150 mg·L^{-1}共43个样品，第一时间段采集35个总有机碳标准样品光谱分为训练集和测试集，建立并验证总有机碳检测算法模型。为评价算法模型鲁棒性，在另一时间段采集剩余的8个标准样品光谱进行反测验证。实验结果表明，采用基于自适应增强学习的区间偏最小二乘回归法建立的总有机碳定量模型具有较高的精度和鲁棒性，分组验证和反测验证的预测均方根误差分别为1.304和1.533 mg·L^{-1}，均优于偏最小二乘回归和极限学习机方法。为进一步验证该方法的有效性，使用该建模方法预测生活污水的总有机碳含量。实际地表水样本取样于河北石家庄藁城污水处理厂排污口污水及河北先河公司园区的生活污水，经稀释后共获得50组地表水样本，采用SPXY方法分为训练集33组水样，测试集17组水样。在实际水样检测中，采用净信号分析方法进行光谱预处理，降低总有机碳与其他水质参数间的交叉干扰；分组验证预测均方根误差为3.26 mg·L^{-1}，平均绝对值百分比误差为3.46%。综上所述，基于自适应增强学习的区间偏最小二乘回归方法，可以快速准确地对地表水中总有机碳进行检测，为在线水质总有机碳检测提供了方法支撑。

关键词 紫外-可见光谱；自适应增强学习；区间偏最小二乘法；总有机碳检测；地表水

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)11-3423-05

引 言

水资源是人类生存发展最重要的战略资源，保护地表水资源安全对社会可持续性发展具有重大意义。为保护水资源安全，需要采用有效方法对水质进行评价。总有机碳是反映水中含碳有机污染物的指标，可以作为评价地表水质的重要依据。国内外对总有机碳检测进行了很多方法的尝试，现行的国家标准为2009年制定的燃烧氧化-非分散红外吸收法，将试样通过高温燃烧管高温催化氧化获得总碳转化的二氧化碳，经低温反应管酸化测得无机碳转化的二氧化碳，经非分散红外检测器检测，总碳与无机碳差值即为总有机碳。在2017年，Ma等采用臭氧氧化化学发光信号进行在线海水总有机碳含量检测[1]，取得了较好的测量结果。2018年Shin-Ichi Ohira等研制出以水洗脱液为基础的高效液相色谱的总有机碳检测器[2]，将分离的分析物在线氧化为二氧化碳，收集到超纯水中，然后通过电导率检测总有机碳含量。2020年，Luo等采用比色传感器，在高通量过程中与水样反应产生特征模式，采用机器学习建立传感器与总有机碳含量的模型[3]。上述方法均需要进行复杂的前处理，近年来，紫外可见光谱法因具有无需化学前处理、可在线原位检测、快速响应等优点在水质检测中被广泛应用[4-6]。本工作采用浸入式的紫外-可见光谱仪器采集水样光谱，采用基于自适应增强学习的区间偏最小二乘回归方法建立光谱与总有机碳含量的定量分析模型，实现地表水总有机碳的定量分析。采用净信号分析降低地表水中因其他物质对总有机碳检测产生的干

收稿日期：2021-11-01，**修订日期**：2022-01-20

基金项目：国家自然科学基金项目(61575015)和国家重点研发计划“制造基础技术与关键部件”重点专项(2020YFB2009000)资助

作者简介：李庆波，女，1975年生，北京航空航天大学仪器科学与光电工程学院副教授 e-mail：qbleebuaa@buaa.edu.cn

扰，提高总有机碳检测方法在不同地表水环境的鲁棒性。

1 实验部分

1.1 样本

根据国标法采用分析纯邻苯二甲酸氢钾配置总有机碳标准溶液共 43 个样品，浓度范围为 25.0～150.0 mg·L^{-1}。选取 25 个样本作为建模训练集，10 个样本作为测试样品集，8 个样本作为第二时间段的反测样本集。

实测样本为现场采集藁城污水厂排污口污水及河北先河公司园区的生活污水，进行等梯度稀释共得到 50 组水样，总有机碳浓度范围为 7.2～272.0 mg·L^{-1}，选取 33 个样品进行建模，17 个样品作为测试集验证，实际水样总有机碳含量采用国标法经实验室化验得到。

1.2 仪器

采用河北先河环保科技股份有限公司研发的浸入式在线水质分析仪。该设备光源为氙灯，光程长为 2 mm，采集光谱范围为 188～722 nm，共 256 个波段，每个水样光谱连续扫描 10 次，每次间隔 15 s，取平均光谱作为该样品的对应光谱。

1.3 性能评价指标

使用预测均方根误差(RMSEP)和平均绝对值百分比误差(MAPE)作为模型预测测试集样品浓度的精度评价指标，其计算方法如式(1)和式(2)

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

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{y_i}\times 100\% \tag{2}$$

其中，n 为测试集样品数，y_i 为测试集总有机碳实际浓度，$\hat{y}_i$ 为对应水样的预测浓度。

1.4 建模方法

针对总有机碳定量分析问题，采用基于自适应增强学习[7-8]的区间偏最小二乘回归法[9](Adaboost interval partial least squares regression，Ada-iPLSR)。将总有机碳吸收光谱波段分为若干子区间，初始化训练样本权重，依次在各子区间建立偏最小二乘回归模型，根据子区间模型预测误差率计算该子区间预测结果的权重系数，并更新下一子区间训练样本权重，最后将各子区间模型预测结果线性加权组合得到总有机碳的检测结果。具体算法过程如下：

首先将水质某参数的特征吸收峰光谱区间分为互不重叠的 n 个子区间，训练集样本数为 m，然后进行初始化权重 $W_1=(w_{11},w_{12},\cdots,w_{1m})$，$w_{1i}=\frac{1}{m}$，$i=1,2,\cdots,m$；计算当前子区间偏最小二乘回归法训练集上的最大误差

$$E_n=\max|y_i-G_n(x_i)|,\ i=1,2,\cdots,m \tag{3}$$

式(3)中，x_i 为训练集第 i 个样本子区间波长吸光度值，y_i 为训练集第 i 个样本水质参数真值，$G_n(x)$ 为第 n 个子区间的定量模型函数。然后计算每个训练集水样样本参数的相对误差

$$e_{ni}=|y_i-G_n(x_i)|/E_n \tag{4}$$

得到第 n 个子区间偏最小二乘回归模型的预测误差率

$$e_n=\sum_{i=1}^{m}w_{ni}e_{ni} \tag{5}$$

由此得到该子区间预测模型的权重系数

$$a_n=e_n/(1-e_n) \tag{6}$$

样本权重更新公式为

$$w_{n+1,i}=\frac{w_{ni}}{Z_n}a_n^{1-e_{ni}} \tag{7}$$

其中 Z_n 为规范化因子

$$Z_n=\sum_{i=1}^{m}w_{ni}a_n^{1-e_{ni}} \tag{8}$$

最后将各子预测模型结果加权得到自适应增强学习后的预测结果

$$y=\sum_{n=1}^{N}\left(\ln\frac{1}{a_n}\right)G_n(x) \tag{9}$$

1.5 预处理方法

针对实际地表水基质对总有机碳光谱检测造成交叉干扰问题，采用净信号分析方法[9]提取总有机碳净信号光谱信息。具体计算过程如下：

首先将样品原始光谱 X 向浓度矩阵 y 进行正交投影得到 X_{-k}，即得到除被分析参数以外其他成分的张成空间，得

$$\begin{aligned}X_{-k}&=X-\alpha y^{*}\overline{X}\\ \alpha&=\frac{1}{\overline{X}X^{+}y^{*}}\\ y^{*}&=XX^{+}y\end{aligned} \tag{10}$$

式(10)中，X^{+} 为 X 奇异值分解取前 f 个主成分得到的逆矩阵。然后对 X_{-k} 进行奇异值分解，取前 $f-1$ 个主成分得到 X_{-k}^{+}。将 X 向 X_{-k} 进行正交投影，得

$$X_k^{*}=X[I-(X_{-k}^{+})^{\mathrm{T}}X_{-k}] \tag{11}$$

最后对未知样品进行变换

$$x_k^{*}=x[I-(X_{-k}^{+})^{\mathrm{T}}X_{-k}] \tag{12}$$

2 结果与讨论

2.1 水质参数光谱特征曲线

图 1 为第一时间段实验室配制总有机碳标准溶液光谱，

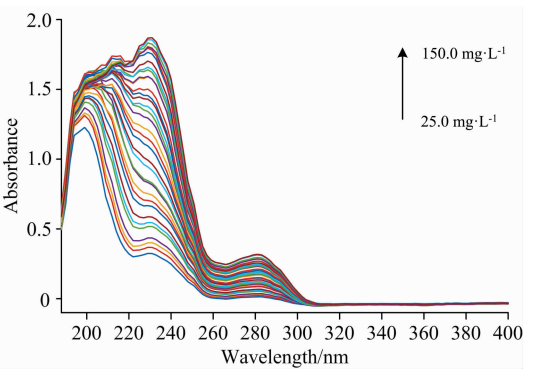

图 1 第一时间段总有机碳标准溶液光谱

Fig. 1 Spectra of total organic carbon standard solution in the first period

总有机碳含量范围为 25.0～150.0 mg·L^{-1}，共 35 个不同浓度的标准总有机碳溶液。从图中可以看出，标准溶液光谱在 230～260 和 260～300 nm 有两个吸收峰，为减少与其他水质参数吸收峰重叠，选择在 230～260 nm 波段进行光谱与总有机碳的定量建模。图 2 为另一时间段采集剩余的 8 个标准样品光谱，总有机碳含量范围为 37.0～145.0 mg·L^{-1}。图 3 为实际地表水进行梯度稀释后的共 50 个水样样本光谱。

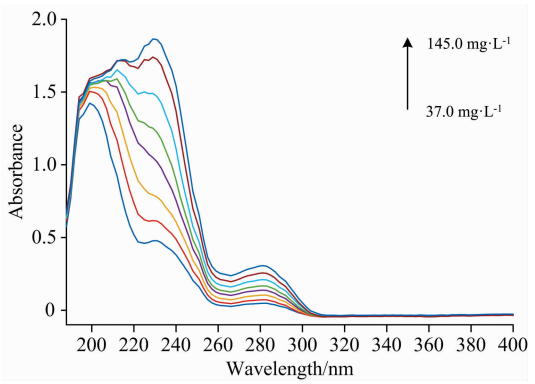

图 2　第二时间段总有机碳标准溶液光谱

Fig. 2　Spectra of total organic carbon standard solution in the second period

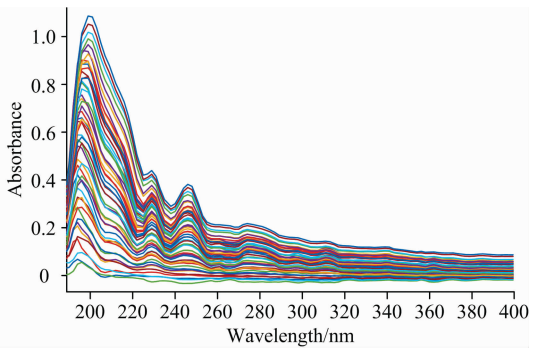

图 3　实际地表水水样光谱

Fig. 3　Spectra of actual surface water samples

2.2　定量模型分析结果

2.2.1　总有机碳标准溶液分组验证及反测验证结果

首先采用 SPXY 算法[11]选出 25 个浓度总有机碳溶液作为训练集，10 个浓度总有机碳溶液作为测试集。另配制 8 个浓度总有机碳样品，作为第二时间段反测样品，用来检验仪器状态变化时模型预测准确性及鲁棒性。

由表 1 结果可知，由于仪器状态的变化，在第二时间段进行的反测验证实验中同一模型总有机碳预测的均方根误差要大于分组验证实验。采用 Ada-iPLSR 算法回归模型在分组验证和反测验证中均方根误差为 1.304 和 1.533 mg·L^{-1}，均为最小结果，具有最好的定量分析精度，且具有很好的鲁棒性，和偏最小二乘回归方法和极限学习机方法比较，反测实验定量精度分别提高了 27.33%和 3.72%。

2.2.2　实际水样总有机碳预测结果

实际水样验证实验，分别于河北石家庄藁城污水处理厂排污口和河北先河公司园区采集生活污水，通过蒸馏水对污水进行稀释共得到 50 个水样样本，经实验室国标法化验得到总有机碳实际浓度。采用 SPXY 算法选择 33 个样本作为训练集，17 个样本作为测试集，建模方法采用偏最小二乘回归法(PLSR)、自适应增强学习区间偏最小二乘回归法(Ada-iPLSR)、净信号分析偏最小二乘回归法(Nas-PLSR)以及净信号分析自适应增强学习区间偏最小二乘回归法(Nas-Ada-iPLSR)进行对比，评价指标采用预测均方根误差和相对误差绝对值的平均值，结果如表 2 和表 3 所示。

表 1　总有机碳标准溶液浓度预测结果

Table 1　The prediction results of total organic carbon concentration in standard solution

实验类型	建模方法	RMSEP/(mg·L^{-1})
分组验证	PLSR	1.306
	ELM	1.371
	Ada-iPLSR	1.304
反测验证	PLSR	1.952
	ELM	1.590
	Ada-iPLSR	1.533

表 2　实际地表水总有机碳浓度预测结果

Table 2　The prediction results of total organic carbon concentration in surface water

建模方法	RMSEP/(mg·L^{-1})	MAPE/%
PLSR	4.68	5.18
Ada-iPLSR	3.67	3.70
Nas-PLSR	4.40	5.04
Nas-Ada-iPLSR	3.26	3.46

表 3　实际地表水测试集样本预测结果

Table 3　The prediction results of actual surface water samples in test set

实际浓度/(mg·L^{-1})	预测浓度/(mg·L^{-1})	相对误差的绝对值/%
21.7	25.3	16.45
32.6	31.5	3.57
43.5	42.3	2.75
50.7	54.3	7.14
54.4	50.8	6.62
65.2	64.0	1.74
76.2	74.7	1.92
86.9	85.0	2.22
101.4	101.6	0.22
115.8	113.3	2.22
130.3	128.3	1.58
144.8	145.5	0.48
152.0	151.5	0.34
166.5	173.6	4.27
173.8	180.0	3.60
185.0	179.9	2.76
217.6	219.7	0.96

Nas-Ada-iPLSR 模型在四种建模方法中均方根误差和相对误差绝对值的平均值均为最小，分别为 3.26 mg·L^{-1}和 3.46%。Nas-Ada-iPLSR 模型与偏最小二乘回归法、自适应增强学习区间偏最小二乘回归法、净信号分析偏最小二乘回归法相比，均方根误差分别提高了 43.56%，12.58%，34.97%，具有了较好的预测精度和适应性，能够对实际地表水样中的总有机碳含量进行准确预测。

3 结 论

总有机碳是依据碳含量评价水质有机物污染的关键指标，采用紫外-可见光谱技术能够对地表水中总有机碳进行在线快速准确检测。实验结果表明，与传统的定量分析方法相比，本文提出的基于自适应增强学习的区间偏最小二乘回归方法获得更好的水质总有机碳预测结果，分组验证和反测验证的预测均方根误差分别为 1.304 和 1.533 mg·L^{-1}。经净信号预处理后的光谱降低了地表水基质对总有机碳检测的影响，提升了预测精度。分组验证中均方根误差为 3.36 mg·L^{-1}，平均绝对值百分比误差为 3.46%，具有较好的预测精度，验证了模型的有效性和鲁棒性，为地表水总有机碳检测提供了方法支撑。

References

[1] Ma R, Xie Z X, Chu D Z, et al. IOP Conference Series: Earth and Environmental Science, 2017, 82: 012086.
[2] Ohira S I, Kaneda K, Matsuzaki T, et al. Analytical Chemistry, 2018, 90(11): 6461.
[3] Luo R, Ma G, Bi S, et al. Analyst, 2020, 145(6): 2197.
[4] Guo Y, Liu C, Ye R, et al. Applied Sciences, 2020, 10(19): 6874.
[5] LIN Chun-wei, GUO Yong-hong, HE Jin-long(林春伟，郭永洪，何金龙). China Measurement & Test(中国测试), 2019, 45(5): 79.
[6] CHEN Ying, HE Lei, CUI Xing-ning, et al(陈 颖，何 磊，崔行宁，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(5): 1489.
[7] Koduri S B, Gunisetti L, Ramesh C R, et al. Journal of Physics: Conference Series, 2019, 1228: 012005.
[8] Wang J, Xue W, Shi X, et al. Sensors, 2021, 21(18): 6260.
[9] Mishra P, Woltering E, Harchioui N E. Infrared Physics and Technology, 2020, 110: 103459.
[10] Alessandro Z, Lucia M, Giuliano G, et al. European Journal of Pharmaceutical Sciences, 2019, 130: 36.
[11] Yang Zhenfa, Xiao Hang, Zhang Lei, et al. Analytical Methods, 2019, 11(31): 3936.

Detection of Total Organic Carbon in Surface Water Based on UV-Vis Spectroscopy

LI Qing-bo[1], BI Zhi-qi[1], CUI Hou-xin[2], LANG Jia-ye[2], SHEN Zhong-kai[2]

1. Key Laboratory of Precision Opto-Mechatronics Technology, Ministry of Education, School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China
2. Hebei Sailhero Environmental Protection Hi-Tech Co., Ltd., Shijiazhuang 050035, China

Abstract Total organic carbon is an index to evaluate the organic pollution of water quality based on carbon content, which can reflect the degree of water pollution. Currently, the detection of total organic carbon in surface water mostly adopts the laboratory analysis method after field sampling. This method has the disadvantages of being time-consuming and laborious, complex operation, secondary chemical pollution, etc. UV-Vis spectroscopy has the advantages of environmental protection, simple operation and real-time on-line in-situ detection. It has a good application prospect in detecting total organic carbon in surface water. The interval partial least squares regression method based on the adaboost algorithm (Ada-iPLSR) is adopted. In this method, the total organic carbon absorption spectrum band is divided into several sub-intervals. The training sample weight is initialized. The partial least squares regression model is established in each sub-interval in turn, the weight coefficient of the prediction result of the sub-interval is calculated according to the prediction error rate of the sub-interval model, and the training sample weight of the next sub-interval is updated. Finally, the prediction results of each sub-interval model are linearly weighted to obtain the detection results of total organic carbon. 43 total organic carbon standard solution samples concentrations of 25～150 mg·L^{-1} were prepared in the experiment. 35 total organic carbon standard samples were collected in the first period, and the spectra were divided into training and test sets. The total organic carbon detection algorithm model was established and verified. In order to evaluate the robustness of the algorithm model, the spectra of the remaining 8 standard samples were

collected in another period for test verification. The experimental results show that the total organic carbon quantitative model established by Ada-iPLSR has high accuracy and robustness. The root means square errors of group verification and test verification are 1.304 and 1.533 mg • L^{-1} respectively, which are better than partial least squares regression and Extreme Learning Machine methods. In order to further verify the effectiveness of this method, this modeling method is used to predict the total organic carbon content of domestic sewage. The actual surface water samples were taken from the sewage at the sewage outlet of Gaocheng sewage treatment plant in Shijiazhuang, Hebei and the domestic sewage in the park of Hebei Xianhe company. After dilution, 50 surface water samples were obtained. SPXY method was used to divide them into 33 water samples in the training set and 17 water samples in the test set. In the actual water sample detection, the net signal analysis method is used for spectral pretreatment to reduce the interference of other substances in surface water on the detection of total organic carbon. The root means square error of group verification prediction is 3.26 mg • L^{-1}, and the average absolute value percentage error is 3.46%. To sum up, the Ada-iPLSR method can quickly and accurately detect the total organic carbon in surface water, providing a method support for the on-line detection of total organic carbon in water quality.

Keywords UV-Vis spectroscopy; Adaboost algorithm; Interval partial least squares regression; Total organic carbon detection; Surface water

(Received Nov. 1, 2021; accepted Jan. 20, 2022)

《光谱学与光谱分析》期刊社决定采用 ScholarOne Manuscripts 在线投稿审稿系统

《光谱学与光谱分析》期刊社与汤森路透集团签约，自 2010 年 12 月 1 日起《光谱学与光谱分析》决定采用 Thomson Reuters旗下的 ScholarOne Manuscripts 在线投稿审稿系统。

• ScholarOne Manuscripts，该系统不仅能轻松处理稿件，而且能提速科技交流。

• 全球已有 360 多家学会和出版社的 3 800 多种期刊选用了 ScholarOne Manuscripts 系统作为在线投稿、审稿平台，全球拥有超过 1 350 万的注册用户，代表着全球学术期刊在线投审稿的一流水平。

• ScholarOne Manuscripts 与 EndNote，Web of Science 无缝链接和整合；使科研探索、论文评阅和信息传播效率大为提高。

• ScholarOne Manuscripts 是汤森路透科技集团的一个业务部门，拥有丰富的学术期刊业务经验，为学术期刊提供综合管理工作流程系统，使期刊更有效管理投稿、同行评审、加工和发表过程，提高作者心中的专业形象，缩短论文发表时间，削减管理成本，帮助期刊提高科研绩效和实现学术创新。

《光谱学与光谱分析》采用“全球学术期刊首选的在线投稿审稿系统—ScholarOne Manuscripts”，势必对 2010 年 11 月 30 日以前向本刊投稿的作者在查阅稿件信息时，会带来某些不便，在此深表歉意！为了推进本刊的网络化、数字化、国际化进程，以实现与国际先进出版系统对接；为了不断提高期刊质量，加快网络化、数字化建设，加快与国际接轨的进程，希望能得到广大作者、读者们的支持与理解，对您的理解和配合深表感激。这是一件新事物，肯定有不周全、不完善的地方，让我们共同努力，不断改进和完善起来。

《光谱学与光谱分析》期刊社

2010 年 12 月 1 日