

DOI:10.16136/j.joel.2025.10.0642

基于荧光发射光谱的水质化学需氧量的检测

孙连升¹, 王 诺¹, 牛鑫濛¹, 周昆鹏^{1,2*}, 周思薇¹

(1. 内蒙古民族大学 物理与电子信息学院, 内蒙古 通辽 020800; 2. 内蒙古自治区核与辐射探测联合实验室, 内蒙古 通辽 020800)

摘要: 基于荧光发射光谱检测水质化学需氧量(chemical oxygen demand, COD), 对光谱数据分别进行多元散射校正(multiplicative scatter correction, MSC)、一阶微分、标准正态变换(standard normal variate transformation, SNV)、最大最小归一化及 Savitzky-Golay(SG)平滑等预处理, 运用后向区间偏最小二乘法(backward interval partial least squares, BiPLS)和联合区间偏最小二乘法(synergy interval partial least squares, SiPLS)筛选关键特征波段, 再用偏最小二乘法(partial least square, PLS)构建预测模型, 以提升光谱处理效果与模型预测精度。实验结果显示, 在对荧光发射光谱数据预处理时, SG 卷积平滑效果最佳, BiPLS 特征提取选择性更好。当 $E_x=310$ nm 时, 经 SG 卷积平滑与 BiPLS 提取特征波段后建立的 PLS 模型各项指标最优, 检验集相关系数 r_p 达 0.9191, 检验均方根误差(root mean square error of prediction, RMSEP) 3.3488 mg/L, 检验偏差 $Bias$ 为 -0.2835 mg/L。本文方法为水质 COD 的快速检测提供了一种实用方案。

关键词: 荧光发射光谱; 化学需氧量(COD); 数据预处理; 特征提取; 偏最小二乘法(PLS)

中图分类号: O657.3 文献标识码: A 文章编号: 1005-0086(2025)10-1034-11

Detection of chemical oxygen demand in water based on fluorescence emission spectroscopy

SUN Liansheng¹, WANG Nuo¹, NIU Xinmeng¹, ZHOU Kunpeng^{1,2*}, ZHOU Siwei¹

(1. School of Physics and Electronic Information, Inner Mongolia Minzu University, Tongliao, Inner Mongolia 020800, China; 2. Inner Mongolia Autonomous Region Joint Laboratory of Nuclear and Radiation Detection, Tongliao, Inner Mongolia 028000, China)

Abstract: This study proposes a method for detecting chemical oxygen demand (COD) in water using fluorescence emission spectroscopy. The spectral data were preprocessed using multiple techniques, including multiplicative scatter correction (MSC), first-order derivative, standard normal variate transformation (SNV), max-min normalization, and Savitzky-Golay (SG) smoothing. Key feature bands were selected using backward interval partial least squares (BiPLS) and synergy interval partial least squares (SiPLS). A prediction model was then developed using partial least square (PLS) to improve spectral processing performance and prediction accuracy. Experimental results demonstrated that SG smoothing provided the best preprocessing performance, while BiPLS showed superior selectivity for feature extraction. At an excitation wavelength (E_x) of 310 nm, the PLS model, optimized by combining SG smoothing and BiPLS feature extraction, achieved optimal performance, with the validation set correlation coefficient (r_p) of 0.9191, the root mean square error of prediction (RMSEP) of 3.3488 mg/L, and the prediction $Bias$ of -0.2835 mg/L. This method offers a practical approach for rapid COD detection in water quality assessment.

Key words: fluorescence emission spectroscopy; chemical oxygen demand (COD); data preprocessing; feature extraction; partial least square (PLS)

* E-mail: kunpeng032@imn. edu. cn

收稿日期: 2024-12-12 修订日期: 2024-02-08

基金项目: 国家自然科学基金(62463023, 61963031)和内蒙古自治区自然科学基金(2023LHMS06019)资助项目

0 引言

水资源在人类社会生活中不可或缺,其质量直接关系到人们的身体健康。水质检测分析方法有生物传感法、分子光谱法、化学分析法、色谱分离法等。水质化学需氧量(chemical oxygen demand, COD)是进行水质检测最重要的指标之一,它表示在一定的条件下,采用一定的强氧化剂处理水样时,所消耗的氧化剂量^[1]。目前检测 COD 的化学方法主要是高锰酸盐指数法和重铬酸钾回流法,但两者均存在测量耗费时间、需要加入化学试剂、存在对水体的二次污染、使用时维护成本较高等不足。光谱检测技术具有绿色无污染、检测速度快、重复性好的特点^[2],广泛应用于多种领域。如:在食品检测领域,张微微等^[3]利用同步荧光光谱技术结合支持向量机(support vector machine, SVM),成功实现了对掺杂牛奶的智能判别,为食品掺假检测提供了高效方法;在医学应用与公共安全领域,于欣冉等^[4]基于近红外和遗传算法优化的支持向量回归模型实现了对人体血糖浓度的无创检测,验证了启发式智能算法对于近红外无创检测的可行性;庄园等^[5]利用高光谱成像技术对血痕种属进行了鉴别研究,为法医学领域提供了新的技术途径;南迪娜等^[6]基于拉曼光谱技术实现了对危险液体的快速识别,为公共安全提供了高效检测手段;在文化遗产保护与材料分析领域,徐军平等^[7]通过对会理出土的战汉时期铜矛进行铅料来源分析,为古代冶金技术研究提供了重要数据支持;DONG 等^[8]利用深度学习辅助的光谱技术对等离子体纳米结构进行了表征,展示了人工智能在光谱分析中的潜力;在农业与环境监测领域,李智缘等^[9]基于光谱指数对土壤重金属 Zn 含量进行了定量预测与空间分布研究,为土壤污染治理提供了科学依据;叶彬强等^[10]提出了一种多源光谱融合的水样 COD 检测方法,显著提高了检测精度;LAN 等^[11]通过紫外-可见光谱和三维荧光光谱对长江三角洲农村生活污水中的 DOM 进行了追踪与预测,为水环境管理提供了技术支持;ZEESHAN 等^[12]研究了季节性变化对模拟河岸过滤系统中溶解性有机物(dissolved organic matter, DOM)浓度和组成的影响,揭示了 DOM 的动态变化规律;GEETHA 等^[13]则采用卷积神经网络-双向长短期记忆模型对 Kaveri 河水质进行了监测,为河流水质管理提供了智能化解决方案。

综上所述,采用光谱法检测水质 COD 具有分析速度快、无化学试剂污染、操作维护简单、运行费用低等显著优点,受到了研究人员的普遍关注且发展迅

速^[14]。荧光光谱检测水质 COD 时所需样品量少且无需对水样进行前处理,因此该方法具有检测简单、效率高、误差小等优点^[15,16]。通过采集水样的荧光光谱,能很方便地对水质 COD 进行定量分析。

本文结合荧光发射光谱对实际水样中的 COD 浓度进行检测。对不同激发波长的荧光发射光谱进行不同的预处理和特征波长提取后,使用偏最小二乘法(partial least square, PLS)建立 PLS 回归模型检测水质 COD 浓度,并取得了良好的检测效果,为测定水质中的 COD 提供了一种绿色、简单的思路。

1 实验部分

1.1 实验样本和仪器

本研究所用到的实验水样包括公园湖水、市内河水、生活污水等,共采集水样 97 份,采集后在实验室环境下静置 30 min,取实验水样的上层液体进行水质 COD 理化值的检测和荧光光谱的采集。为了方便后续建模分析,将 97 组数据随机分成 64 组校正集和 33 组检验集。

实验样本的荧光光谱采用日立 F-7000 型荧光光谱仪采集得到,实验样本的 COD 理化值采用快速消解分光光度法检测得到。采集荧光光谱时,设置激发光谱波长(excitation wavelength, E_x)范围为 200—330 nm,步长 5 nm,发射光谱波长(emission wavelength, E_m)范围为 250—500 nm,步长 2 nm;快速消解分光光度法检测样本理化值时设置消解时间 120 min,消解温度设置为 150 °C。采集某实际水样所得到的去散射峰后的三维荧光光谱如图 1 所示。图 1 中,底部坐标 E_x 、 E_m 分别为激发波长和发射波长,纵向坐标为荧光强度。

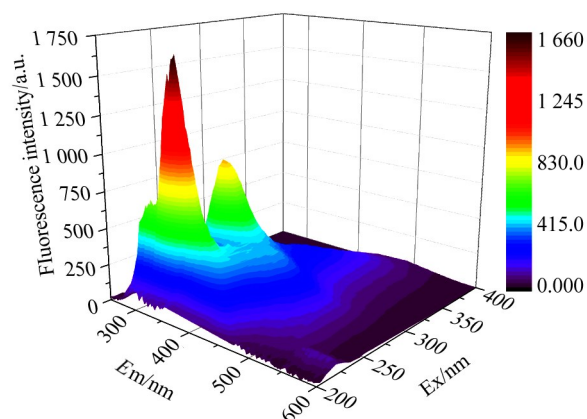


图 1 某实际水样去散射后的三维荧光光谱图

Fig. 1 EEMs of a real water sample after removal of scattering

本文将三维荧光光谱按照激发波长 E_x 进行展开,分别得到 $E_x = 200\text{ nm}, 205\text{ nm}, \dots, 325\text{ nm}, 330\text{ nm}$ 下的荧光发射光谱,发射波长 E_m 范围为 $250\text{—}500\text{ nm}$ 。以 $E_x = 255\text{ nm}$ 下展开的荧光发射光谱图为例,如图 2(a)所示,每条谱线代表一个实验水样,各条谱线的颜色、幅度等均不相同,表明各水样的 COD 值各不相同。

1.2 模型评价指标

根据光谱分析常用指标:相关系数 (r)、校正集交叉验证均方根误差 (root mean square error of cross-validation, $RMSECV$)、检验集外部检验均方根误差 (root mean square error of prediction, $RMSEP$)、整体平均偏差 ($Bias$)。对本文而言, $Bias$ 是检验样品的测量值与真值之间的整体平均偏差。

各评价指标表达式如式(1)~式(4)所示:

$$r = \frac{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (1)$$

$$RMSECV = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - y_i)^2}, \quad (2)$$

$$RMSEP = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2}, \quad (3)$$

$$Bias = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i), \quad (4)$$

式中, \hat{y}_i 为基于光谱建模的预测结果, \bar{y} 为标准方法测定值的平均值, y_i 为标准方法测定得到的结果, n 为校正集样本数, m 为检验集样本数。比较和分析不同激发波长下不同模型的预测效果, r 越接近 1,且 $RMSECV$ 、 $RMSEP$ 和 $Bias$ 越小,说明建立的模型预测结果越接近真实测量结果,反之效果越差。由此可以选出最优的光谱建模算法。

2 结果与讨论

2.1 荧光光谱数据的预处理及 PLS 建模

荧光发射光谱数据预处理是通过减少光谱中的噪声以提高信噪比、校正数据中的基线漂移、将数据缩放放到一定的范围等方法,去消除干扰、降低数据的复杂性,达到提高光谱数据质量和准确性的目的。预处理后的光谱数据可提高后续所建模型的可靠性和鲁棒性。在水质检测中,PLS 是最常用的建模方法之一。为突出数据预处理对建模效果的影响,本文将采用 5 种不同的方法来对原始光谱数据进行预处理,并利用处理后的数据在整个光谱范围内进行

PLS 建模,对各模型的性能指标进行对比分析,旨在筛选出最佳的预处理技术。

2.1.1 荧光发射光谱的预处理

本文采用 Savitzky-Golay 卷积平滑(SG 平滑)、多元散射校正 (multiplicative scatter correction, MSC)、标准正态变换 (standard normal variate transformation, SNV)、最大最小归一化、一阶微分的预处理方法对原始光谱数据进行预处理。在 $E_x = 255\text{ nm}$ 下展开的荧光发射光谱在预处理前后的光谱数据如图 2(b)~图 2(f)所示。其中,图(b)为 SG 平滑处理后的光谱图,图(c)为多元散射校正(MSC)处理后的光谱图,图(d)为标准正态变换(SNV)处理后的光谱图,图(e)为最大最小归一化处理后的光谱图,图(f)为一阶微分处理后的光谱图。由于上述预处理方法的目的、原理和应用效果上存在差异,因此造成图 2(b)~图 2(f)的纵坐标名称和数值刻度上有所不同。SG 平滑是为了减少光谱数据中的噪声,提高数据的平滑度,不改变光谱的物理意义;MSC 是为了消除散射效应和颗粒大小差异导致的光谱基线漂移和幅度变化,校正后的光谱更能反映样品的真实化学信息;SNV 是为了使光谱数据更加符合正态分布,增强数据的稳定性,在处理后的纵坐标在物理意义上仍与原始光谱相关,但从数据特征角度可理解为经过标准化后的相对值;最大最小归一化处理是为了消除量纲的影响,使数据处于同一量纲上,便于比较和分析,数值刻度被严格限制在 $0\text{—}1$ 之间,突出了光谱数据在相对位置上的分布情况,忽略了原始的绝对数值;一阶微分处理是计算光谱数据的一阶导数,用于突出光谱的变化率信息,能够增强光谱的特征,与原始光谱的数值刻度没有直接的对应关系,得到的是光谱数据的变化率。

2.1.2 PLS

PLS 是一种多元统计分析方法,常用于处理具有多重共线性和高维特征的数据。它通过建立水质参数与光谱数据之间的关联模型,从而实现对水质参数的预测和分析。

PLS 将光谱矩阵 X 和水质浓度矩阵 Y 同时进行主成分分解,即:

$$X = TP + E, \quad (5)$$

$$Y = UQ + F, \quad (6)$$

式中, T 是 X 的得分矩阵, U 是 Y 的得分矩阵, P 是 X 的载荷矩阵, Q 是 Y 的载荷矩阵, E 是 PLS 模型去拟合 X 时所引进的误差, F 是 PLS 模型去拟合 Y 时所引进的误差。PLS 方法要求对 T 和 U 的相关性最大,因此,将 T 和 U 作线性回归得到式(7)和式(8):

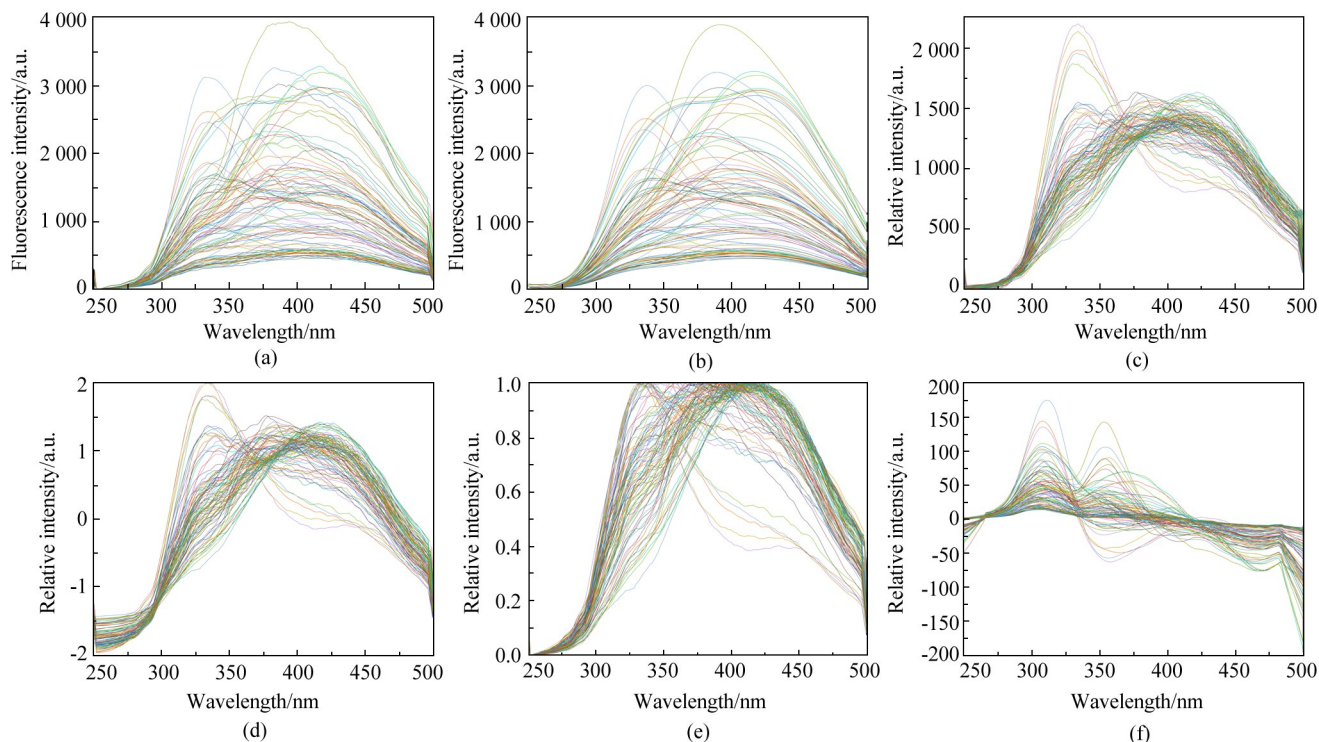


图 2 某实际水样在 $E_x=255$ nm 处的荧光发射光谱图:(a) 原始光谱;

(b) SG 平滑处理后的光谱;

(c) 多元散射校正(MSC)处理后的光谱;

(d) 标准正态变换(SNV)处理后的光谱;

(e) 最大最小归一化处理后的光谱;

(f) 一阶微分处理后的光谱

Fig. 2 Fluorescence emission spectra of a real water sample at $E_x=255$ nm:(a) Original spectra; (b) SG smoothed spectra; (c) Spectra processed by MSC; (d) Spectra processed by SNV; (e) Spectra processed by max-min normalization; (f) Spectra processed by first-order derivative

$$U = TB, \quad (7)$$

$$B = (T'T)^{-1}T'U, \quad (8)$$

在未知水样水质参数预测时,由未知水样的光谱矩阵 X 与通过训练校正得到的 P 即可求出未知水样的 T 矩阵,然后即可得到未知水样的 Y 矩阵,如式(9)所示:

$$Y_{\text{未知}} = T_{\text{未知}}BQ. \quad (9)$$

2.1.3 不同光谱预处理效果对比

本文根据按激发波长 E_x 展开后发射光谱曲线效果的优劣,择优选取了在特定激发波长($E_x=200$ nm, 210 nm, 235 nm, 255 nm, 270 nm, 300 nm, 330 nm)下获得的荧光发射光谱信息,并对这些数据执行了以上 5 种不同的预处理步骤。在此基础上,分别针对经过不同预处理步骤的光谱数据以及未作任何处理的原始光谱数据,建立了 PLS 回归模型,并对模型进行了对比分析以评估其性能差异。各模型的评价指标对比如表 1 所示,其中 r_c 和 r_p 分别为校正集和检验集的相关系数。

通过对比分析表 1 可知,当 $E_x=200$ nm 时,经过 SG 平滑处理建模后相关系数 r_c 略高于原始数据所得结果,经过 MSC 处理建模后的相关性则是 5 种

中最差的。当 $E_x=210$ nm 时,原始荧光光谱数据经过一阶求导预处理后建模效果最好,数据在 SG 平滑后所建立的模型效果次之,而 SNV 预处理后的模型效果最差。当 $E_x=235$ nm 时,经过 SG 平滑预处理建立的模型的相关性最好,但其与一阶求导预处理后的校正集数据拟合程度均不如原始数据;而比较检验集建模的相关性可知,一阶求导后模型预测最佳, MSC 后 PLS 模型的相关性拟合最差。当 $E_x=255$ nm 时,经 SG 平滑处理后模型的相关性最佳,经 SNV 处理后建模效果最差。当 $E_x=270$ nm 时,经过 SG 平滑处理后的校正集建模效果最好,但由检验集模型的评价指标可知,其预测效果不如一阶微分预处理后的 PLS 模型效果,由于一阶微分预处理后建模的 r_c 和 r_p 的差值更小,因此一阶微分预处理后所建模型的效果更好;最大最小归一化预处理后建模效果最差。当 $E_x=300$ nm 时,经过 SG 平滑预处理后的模型拟合效果最好,最大最小归一化预处理后的建模效果最差。当 $E_x=330$ nm 时,经过 SG 平滑预处理后模型的预测性能最好且评价指标 r_c 和 r_p 的差值更小,因此 SG 平滑预处理建模效果更好; SNV 预处理建模效果最差。

表 1 基于不同预处理方法的 PLS 模型评价指标
Tab. 1 PLS model evaluation indexes based on different preprocessing methods

Ex/nm	Pretreatment method	Calibration set		Validation set	
		r_c	RMSECV/(mg/L)	r_p	RMSEP/(mg/L)
200	Original data	0.7916	8.3189	0.2211	8.3790
	First-order derivative	0.7783	8.5374	0.1419	8.5406
	SG smoothing	0.7917	8.3171	0.2223	8.3747
	Max-min normalization	0.7006	9.7124	0.3677	8.3002
	SNV	0.5898	11.2200	0.1086	10.8649
	MSC	0.5546	11.5930	0.2041	10.4215
210	Original data	0.9614	3.7742	0.6430	6.5560
	First-order derivative	0.9693	3.3536	0.6755	6.2792
	SG smoothing	0.9548	4.0548	0.6183	6.7338
	Max-min normalization	0.7659	8.9483	0.6041	7.2304
	SNV	0.6943	9.9800	0.5316	0.5316
	MSC	0.7541	9.0572	0.5995	7.2060
235	Original data	0.9664	3.5139	0.6901	6.2474
	First-order derivative	0.9447	4.4677	0.8027	5.0639
	SG smoothing	0.9538	4.0828	0.7955	5.1461
	Max-min normalization	0.7488	9.1333	0.0924	11.6225
	SNV	0.7348	9.2508	0.1184	11.6906
	MSC	0.7271	9.3862	0.0831	10.4448
255	Original data	0.9495	4.2805	0.8551	4.4185
	First-order derivative	0.9499	4.2575	0.8382	4.6563
	SG smoothing	0.9521	4.1717	0.8471	4.5358
	Max-min normalization	0.7224	9.4604	0.2442	9.1042
	SNV	0.7207	9.4885	0.3029	8.9184
	MSC	0.7359	9.3363	0.3250	9.1251
270	Original data	0.9538	4.0912	0.7790	5.4111
	First-order derivative	0.9559	4.0009	0.7895	5.2686
	SG smoothing	0.9563	3.9860	0.7787	5.3832
	Max-min normalization	0.7621	8.9514	0.0971	10.9341
	SNV	0.7894	8.4102	0.0687	10.8396
	MSC	0.7670	8.8031	0.1236	10.3164
300	Original data	0.9341	4.8715	0.8172	4.8979
	First-order derivative	0.9333	4.9175	0.8458	4.5404
	SG smoothing	0.9368	4.7722	0.8328	4.6922
	Max-min normalization	0.7939	8.4156	0.1588	10.7439
	SNV	0.8083	8.0748	0.1464	9.9379
	MSC	0.8074	8.1512	0.2532	9.0727
330	Original data	0.9331	4.9435	0.6546	6.4466
	First-order derivative	0.9371	4.7878	0.7472	5.6584
	SG smoothing	0.9397	4.6797	0.6628	6.3967
	Max-min normalization	0.7928	8.3362	0.0112	11.0299
	SNV	0.7854	8.4931	0.0204	11.9823
	MSC	0.7897	8.4096	0.0053	11.7086

由于 r 越接近 1 且 RMSEP 和 Bias 的值越小建模预测越精准,经综合考虑,由 SG 平滑预处理后的 PLS 回归模型(SG-PLS)效果最好,一阶微分预处理

后的 PLS 建模效果仅次于 SG-PLS 模型效果。下文将以由 SG 平滑预处理后的荧光发射光谱数据为研究对象,进行特征提取算法和建模效果的对比研究。

2.2 光谱特征波段提取与 PLS 建模

光谱特征波段的选取是为了去除光谱矩阵中的无关信息和冗余数据,提升检测结果的准确性以及预测模型的性能,方便判断光谱和样品中需要检测成分的相关性。本文将 64 组校正集和 33 组检验集的荧光发射光谱数据分别使用后向区间偏最小二乘法(backward interval partial least squares, BiPLS)和联合区间偏最小二乘法(synergy interval least squares, SiPLS)提取特征波段,对不同激发波长下的特征发射光谱数据进行 PLS 建模,得出最优模型。

2.2.1 BiPLS 提取特征

BiPLS 算法是一种基于后向选择的区间 PLS 方法。它通过逐步剔除对模型贡献较小的区间,保留对模型预测最有用的区间,从而优化模型性能。光谱特征的提取过程如下:首先将自变量(X)的整个光谱区域划分为若干等宽区间,之后对所有区间建立

PLS 模型,计算模型的性能指标(如 $RMSECV$),然后进行后向选择:每次剔除一个区间并重新建立 PLS 模型,计算模型的性能指标,比较剔除不同区间后模型的性能;重复以上过程,遍历所有可能的区间组合,逐步剔除对模型贡献最小的区间,直到满足停止条件;最后,保留下的区间组合所包含的波长变量就被认为是与目标变量最相关、最能有效解释目标变量变化的特征变量。

在建立 PLS 回归模型中使用的潜在变量又称主成分数(partial least squares correlation, PLSC),它决定了在建立 PLS 模型时需要考虑的信息量和模型的复杂性。本文将最大 PLSC 设置为 10,对荧光发射光谱进行特征提取和建模,选择区间间隔数、PLSC、所选区间组合和模型性能评价指标如表 2 所示。

由表 2 可知,对于校正集而言,在激发波长 $E_x =$

表 2 基于 BiPLS 算法提取特征的 PLS 模型评价指标

Tab. 2 PLS model evaluation index based on BiPLS algorithm

E_x/nm	Number of intervals n	PLSC	Selected interval	Calibration set		Validation set	
				r_c	$RMSECV/(mg/L)$	r_p	$RMSEP/(mg/L)$
200	12	3	[1 8 11 12]	0.8183	7.8089	0.3650	7.9913
205	10	5	[3 9]	0.8713	6.6915	0.7461	5.6464
210	10	10	[1 2 5 6 8 9 10]	0.9792	3.3006	0.8158	4.9923
215	9	6	[1 3 4 6 9]	0.9717	3.2089	0.5975	6.9460
220	9	7	[2 4 5 6 8]	0.9766	2.9316	0.6617	6.5177
225	12	5	[1 2 3 5 8 10 11]	0.9779	2.8411	0.6927	6.3292
230	9	5	[2 3 6 8]	0.9761	2.9517	0.7455	5.6786
235	9	10	[2 3 5 6 7]	0.9705	3.2906	0.6352	6.9372
240	12	4	[2 3 9 10]	0.9649	3.5711	0.7283	5.8355
245	10	7	[2 3 8]	0.9602	3.7968	0.7644	5.5405
250	10	5	[1 3 8]	0.9589	3.8562	0.7340	5.7800
255	8	6	[1 3 5 6 7]	0.9596	3.8251	0.8246	4.8924
260	10	4	[1 2 9]	0.9533	4.1077	0.8070	5.0327
265	11	6	[3 4 5 6 8 10]	0.9588	3.8643	0.8271	4.8153
270	10	6	[4 5 8]	0.9622	3.7136	0.8232	4.8266
275	12	4	[5 6 9 11]	0.9537	4.0948	0.8288	4.7421
280	11	3	[2 9 10]	0.9520	4.1584	0.7387	5.7389
285	9	7	[4 5 7]	0.9578	3.9124	0.8487	4.4890
290	9	10	[2 3 4 5 7]	0.9585	3.9784	0.7431	5.8317
295	11	6	[3 7 8 9]	0.9469	4.3820	0.8289	4.7765
300	12	9	[1 3 7 8 10 12]	0.9533	4.1358	0.7557	5.8326
305	12	3	[3 10 11]	0.9448	4.4549	0.7301	5.7970
310	10	9	[4 5 6 8 9 10]	0.9480	4.3564	0.9191	3.3488
315	9	7	[3 5 7]	0.9539	4.0817	0.8659	4.2434
320	10	7	[1 3 6 8]	0.9485	4.3077	0.8158	4.9085
325	12	3	[5 8 9 10 12]	0.9412	4.6068	0.7302	5.8010
330	12	7	[4 8]	0.9450	4.4614	0.7660	5.4520

210 nm, 225 nm 时,利用 BiPLS 提取发射光谱特征后所建立的 PLS 模型效果较好。其中,当 $E_x = 210$ nm 时,其校正集的均方根误差 $RMSECV$ 为

3.3006 mg/L,校正集相关系数 r_c 为 0.9792;当激发波长 $E_x = 225$ nm 时, $RMSECV$ 为 2.8411 mg/L, r_c 为 0.9779。观察检验集可知,在 $E_x = 310$ nm 下的

荧光发射光谱数据模型对应的检验集相关系数 r_p 和检验集均方根误差 $RMSEP$ 为所有模型中最好的, 此时的光谱区间个数 $n=10$, 选出 6 个子区间, 分别为第 4、5、6、8、9、10 区间, 对应发射光谱范围分别为 325—350 nm、350—375 nm、375—400 nm、425—450 nm、450—475 nm、475—500 nm。PLS 建模时, 变量个数为 75, PLSC 为 9。对激发波长为 310 nm 处的荧光发射光谱数据进行 PLS 建模时计算的 PLSC 如图 3 所示, 所选特征光谱区间如图 4 所示。

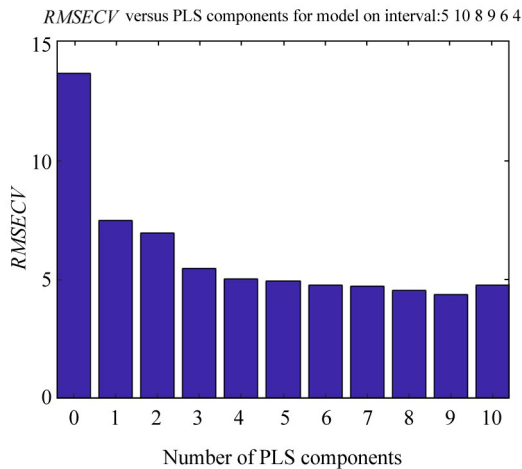


图 3 $E_x=310$ nm 处的校正均方根误差和主成分数
Fig. 3 $RMSECV$ and $PLSC$ at $E_x=310$ nm

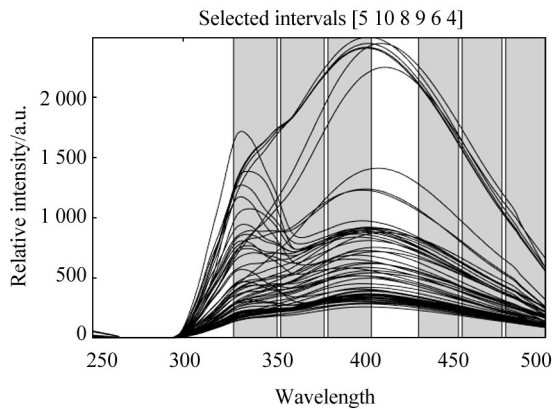


图 4 $E_x=310$ nm 处 BiPLS 提取的特征发射光谱
Fig. 4 Feature emission spectral range extracted by BiPLS at $E_x=310$ nm

经 BiPLS 提取特征后建立的 PLS 模型的校正集相关系数 r_c 为 0.9480, 校正集均方根误差 $RMSECV$ 为 4.3564 mg/L, 系统偏差为 0.0267 mg/L; 检验相关系数 r_p 为 0.9191, 检验均方根误差 $RMSEP$ 为 3.3488 mg/L, 系统偏差为 -0.2835 mg/L, 说明该模型能较为精准地反应实际水样的水质情况。结果表明在 $E_x=310$ nm 的激发波长下, 根据 BiPLS 算

法提取的特征数据所建立的 PLS 回归模型具有很好的预测能力和很小的预测误差。

2.2.2 SiPLS 提取特征

SiPLS 是一种基于协同效应的区间 PLS 方法。该方法将光谱数据的波长范围划分为 2—4 个等间隔的区间, 然后通过组合不同数量的区间来构建 PLS 模型, 选择校正集均方根误差作为模型评价指标, 寻找能够使模型性能最佳的区间组合, 从而确定与目标变量最相关的光谱区间, 实现特征选择和数据降维。光谱特征的提取过程如下: 首先将自变量 (X) 的整个光谱区域划分为若干个等宽的区间, 并生成所有可能的区间组合 (本文取 2 个区间进行组合), 对每一种区间组合建立 PLS 模型并计算不同区间组合模型的性能指标 (如 $RMSECV$), 选择使模型性能最优的区间组合方式; 根据确定的区间数目和区间光谱范围得到敏感特征的光谱矩阵, 然后进一步从这些区间中选择出与目标变量相关性最强、共线性最小的变量, 作为最终的特征变量用于建立模型。

对不同激发波长下荧光发射光谱进行 SiPLS 建模, 所选区间间隔数 (本文取间隔数为 2)、PLSC、所选区间组合和模型性能评价指标如表 3 所示。

由表 3 可知, 对于校正集而言, 在激发波长 $E_x=235$ nm 下的荧光发射光谱数据经 SiPLS 提取特征后的 PLS 校正模型的效果最优。该模型的 $RMSECV=3.0650$ mg/L, $r_c=0.9743$, 但此时检验效果不佳; 对于检验集而言, 当 $E_x=325$ nm 时, 经 SiPLS 提取特征后所建模型的检验效果最优。综合考虑, 本文选用在 $E_x=325$ nm 下的荧光发射光谱数据所建的 PLS 模型。此时, 其校正集的光谱区间数 $n=25$, 联合第 7、19 区间 (对应的荧光发射光谱范围分别为 310—320 nm、430—440 nm), PLSC 为 5。对激发波长为 325 nm 的荧光发射光谱进行 PLS 建模时确定的 PLSC 如图 5 所示, 所选特征光谱区间如图 6 所示。

建模结果显示, 校正集数据在经过 SiPLS 特征提取后进行 PLS 建模所得到的相关系数 r_c 为 0.9494, 均方根误差 $RMSECV$ 为 4.2699 mg/L, 系统偏差为 -0.0932 mg/L, 说明该模型对实际水样的水质情况能有较为精准的反应; 该模型的检验相关系数 r_p 为 0.8716, 均方根误差 $RMSEP$ 为 4.1645 mg/L, 系统偏差为 -0.1758 mg/L。以上数据说明在该激发波长下, 根据校正集在光谱的特征波段区间上建立的模型具有较好的预测性, 但预测精度仍有待进一步提升。

表 3 基于 SiPLS 算法提取特征的 PLS 模型评价指标

Tab.3 PLS model evaluation index based on SiPLS algorithm

Ex/nm	Number of intervals <i>n</i>	PLSC	Selected interval	Calibration set		Validation set	
				<i>r_c</i>	RMSECV/(mg/L)	<i>r_p</i>	RMSEP/(mg/L)
200	13	6	[1 3]	0.8514	7.1592	0.0444	10.5999
205	24	6	[5 19]	0.9335	4.8782	0.7259	5.8459
210	16	7	[5 13]	0.9592	3.8484	0.6608	6.4241
215	17	10	[4 14]	0.9671	3.4619	0.6948	6.2248
220	13	5	[3 10]	0.9699	3.3118	0.6616	6.3958
225	21	5	[5 16]	0.9698	3.3187	0.6950	6.1605
230	11	5	[3 8]	0.9713	3.2460	0.6153	6.7785
235	12	8	[4 12]	0.9743	3.0650	0.7552	5.6461
240	15	5	[2 15]	0.9672	3.4509	0.7671	5.4389
245	18	7	[4 18]	0.9634	3.6450	0.7631	5.4919
250	15	6	[3 15]	0.9634	3.6450	0.7238	5.9338
255	16	8	[6 15]	0.9628	3.6820	0.8166	4.9643
265	19	8	[2 17]	0.9651	3.5790	0.7387	5.7994
270	11	5	[5 9]	0.9645	3.5925	0.8147	4.9250
275	13	4	[6 11]	0.9551	4.0324	0.8179	4.8773
280	19	5	[8 16]	0.9581	3.8970	0.8385	4.6160
285	19	7	[8 17]	0.9614	3.7391	0.7708	5.5896
290	22	6	[6 17]	0.9546	4.0509	0.7168	6.1089
295	19	5	[5 15]	0.9490	4.2880	0.6883	6.3779
300	17	7	[4 13]	0.9519	4.1663	0.6765	6.6752
305	19	5	[4 16]	0.9502	4.2489	0.7044	6.1065
310	18	4	[4 15]	0.9431	4.5205	0.7042	6.0765
315	18	5	[4 15]	0.9452	4.4424	0.7557	5.5613
320	13	5	[4 11]	0.9482	4.3167	0.7780	5.3550
325	25	5	[7 19]	0.9494	4.2699	0.8716	4.1645
330	16	5	[5 11]	0.9505	4.2319	0.7048	6.0383

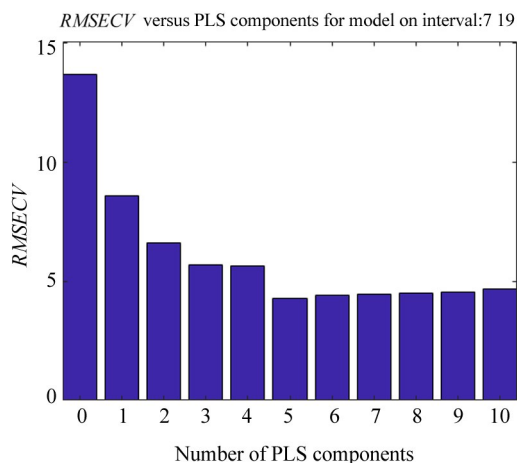


图 5 Ex=325 nm 处的校正均方根误差和主成分数

Fig.5 RMSECV and PLSC at Ex=325nm

2.2.3 特征提取后的建模效对比

本文对不同激发波长下的荧光发射光谱分别采用 SG 平滑、MSC、一阶微分、SNV 和最大最小归一化 5 种方法进行预处理后建立 PLS 模型。由 2.1.3

节可知,经过 SG 平滑预处理的效果略强于一阶微分预处理。为了方便比较不同的特征提取算法的效果优劣,本文分别将表 2 和表 3 中校正集的预测效果和检验集的验证效果表现优秀的模型指标进行挑选,得到较优模型的评价指标集合,如表 4 所示。

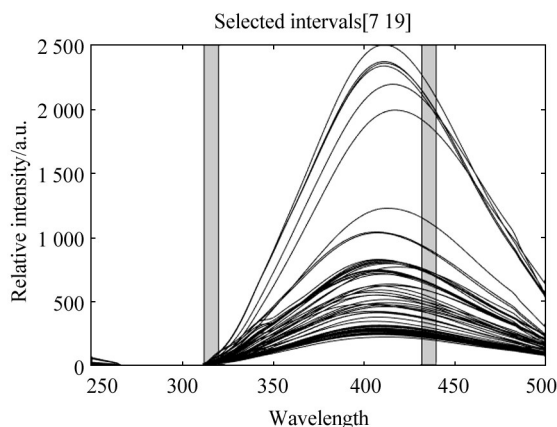


图 6 Ex=325 nm 处 SiPLS 提取的特征发射光谱

Fig.6 Feature emission spectral range extracted by SiPLS at Ex=325 nm

表 4 不同激发波长下采用不同特征提取算法的 PLS 模型评价指标

Tab. 4 The evaluation indexes of PLS model obtained by different feature extraction algorithms under different excitation wavelengths

Feature extraction algorithm	E_x/nm	Number of intervals n	Selected interval	Calibration set		Validation set	
				r_c	$RMSECV/(mg/L)$	r_p	$RMSEP/(mg/L)$
BiPLS	210	10	[1 2 5 6 8 9 10]	0.979 2	3.300 6	0.815 8	4.992 3
	225	12	[1 2 3 5 8 10 11]	0.977 9	2.841 1	0.692 7	6.329 2
	310	10	[4 5 6 8 9 10]	0.948 0	4.356 4	0.919 1	3.348 8
SiPLS	235	12	[4 12]	0.974 3	3.065 0	0.755 2	5.646 1
	270	11	[5 9]	0.964 5	3.592 5	0.814 7	4.925 0
	280	19	[8 16]	0.958 1	3.897 0	0.838 5	4.616 0
	325	25	[7 19]	0.949 4	4.269 9	0.871 6	4.164 5

观察表4中校正集模型评价指标可知, $E_x = 210\text{ nm}$ 下的荧光发射光谱经过 SG 平滑预处理、Bi-PLS 提取特征波段后的预测模型最优; 观察检验集的验证指标可知, $E_x = 310\text{ nm}$ 下的荧光发射光谱经过 SG 平滑, BiPLS 提取特征波段后建立的 PLS 模型效果最优。

2.2.4 特征光谱提取的必要性

为了明确特征提取在光谱建模中的必要性, 研究特征提取的作用, 本文对全光谱数据也建立了 PLS 模型并与特征提取后建立的 PLS 模型进行了效果对比。由于 BiPLS 提取特征的效果优于 SiPLS, 故本文选取 BiPLS 作为最终的特征提取算法。由表 4 可知, 在激发波长 $E_x = 310\text{ nm}$ 时的荧光发射光谱经过 SG 平滑预处理、BiPLS 提取特征波段建立的 PLS 模型检验效果最优, 故本文对 $E_x = 310\text{ nm}$ 激发波长下的荧光发射光谱数据 ($E_m = 250\text{—}500\text{ nm}$) 进行 SG 平滑预处理后在全光谱内建立 PLS 模型, 观察模型的预测效果。结果发现当 PLSC 为 6 时具有最小的校正集均方根误差, $RMSECV$ 为 4.887 mg/L , 如图 7 所示。

此时的校正相关系数 $r_c = 0.9338$ 。检验集数据对全光谱校正模型预测效果的验证效果如图 8 所示。由图 8 可知, 检验集数据在校正模型下的相关系数 r_p 为 0.9074 , 均方根误差 $RMSEP$ 为 3.6050 mg/L , 系统偏差为 -0.2256 mg/L 。相对于全光谱数据直接建立的 PLS 模型效果而言, 经 Bi-PLS 提取特征光谱数据后建立的 PLS 模型效果更优。

激发波长为 310 nm 时的荧光发射光谱经过 SG 平滑预处理后直接建模和经过 BiPLS 特征提取后建模的结果对比如表 5 所示。

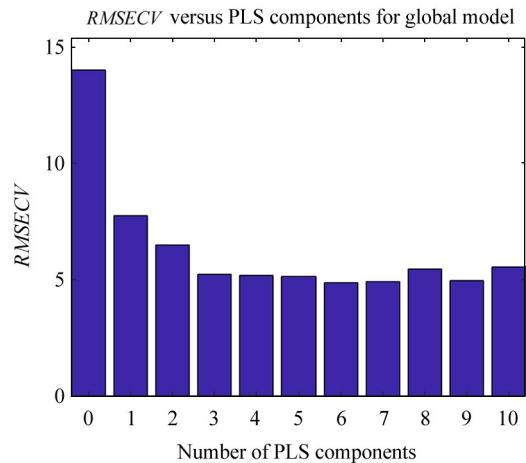


图 7 $E_x = 310\text{ nm}$ 处的全谱段 PLS 模型校正均方根误差和主成分数
Fig. 7 $RMSECV$ and PLSC of full-spectrum PLS model at $E_x = 310\text{ nm}$

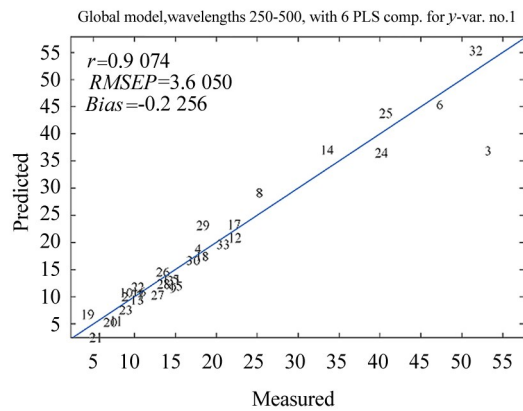


图 8 $E_x = 310\text{ nm}$ 处全谱 PLS 模型的检验效果
Fig. 8 Full-spectrum PLS model validation result at $E_x = 310\text{ nm}$

由表 5 可知, 无论是对于校正集还是对于检验集, 经 BiPLS 提取特征后建立的 PLS 模型效果均优

于全谱段无特征提取的 PLS 模型效果。对比结果表明,特征波段提取非常必要,这有利于减小数据的维度,简化计算,提高数据处理效率。提取的光谱数据

特征能在保留关键信息的基础上有效过滤了数据中的噪声和干扰,利于突出与水质 COD 浓度相关的光谱特征,提高模型的预测性能。

表 5 BiPLS 提取的特征光谱模型和全谱模型的评价指标对比

Tab. 5 Comparison of evaluation indexes between the characteristic spectral model extracted by BiPLS and the full spectral model

Ex/nm	Feature extraction	Characteristic wavelength/nm	Calibration set		Validation set	
			r_c	RMSECV/(mg/L)	r_p	RMSEP/(mg/L)
310	BiPLS	325—400 425—500	0.948 0	4.356 4	0.919 1	3.348 8
	—	250—500	0.933 8	4.887 4	0.907 4	-0.225 6

3 结 论

本文通过分析实际水样的荧光发射光谱数据实现了水质 COD 浓度的检测方法,采用 5 种方法对荧光发射光谱数据进行了预处理,通过对比得出了适用于荧光发射光谱数据的最优预处理方法,后对全部激发波长下的荧光发射光谱数据进行了特征提取并建立了 PLS 回归模型,对比模型效果得到了最优模型。本文得到以下结论:

1) 不同激发波长下的荧光发射光谱数据经不同的预处理算法处理后,建立的 PLS 模型效果不同。经 SG 平滑处理后的建模效果最优,采用该方法对荧光发射光谱进行预处理,可对实际测量数据中存在的噪声进行有效的优化,经一阶微分对荧光发射光谱的预处理效果次之。

2) 光谱数据的预处理和特征提取十分必要。对于本文数据,经 SG 平滑处理后进行特征提取,BiPLS 方法比 SiPLS 方法更加有效准确,BiPLS 特征提取算法的效果更优,且经 BiPLS 特征提取后所建 PLS 模型的效果最优,可精准地对水质 COD 进行定量分析。

3) 本文方法对基于光谱法检测水质 COD 的传感器研发具有一定的启发意义。本文数据均在实验室环境下测得,未考虑自然界中水体的环境参量(如温度、盐度、浊度、pH 等因素)对荧光光谱的影响。这些自然环境参量对荧光的影响在实际测量过程中应进一步考虑,以提高水质 COD 预测模型的泛化能力和鲁棒性。

参考文献:

[1] ZHOU K P, LIU Z Y, CONG M L, et al. Analysis on influencing factors of detecting chemical oxygen demand in water by three-dimensional spectroscopy[J]. Optoelectronics Letters, 2024, 20(1): 42-47.

[2] 周昆鹏,白旭芳,毕卫红.基于紫外-荧光多光谱融合的水质化学需氧量检测[J].激光与光电子学进展,2018,55(11):113003.

ZHOU K P, BAI X F, BI W H. Detection of chemical oxygen demand in water based on multi-spectral fusion of ultraviolet and fluorescence[J]. Laser & Optoelectronics Progress, 2018, 55(11): 113003.

[3] 张微微,璩怡,王强,等.同步荧光光谱技术结合支持向量机对掺杂牛奶智能判别研究[J].光谱学与光谱分析,2024,44(9):2428-2433.

ZHANG W W, QU Y, WANG Q, et al. Research on the synchronous fluorescence spectroscopy combined with support vector machines for intelligent discrimination of milk adulteration[J]. Spectroscopy and Spectral Analysis, 2024, 44(9): 2428-2433.

[4] 于欣冉,赵鹏,宦克为,等.基于 GA-SVR 的近红外无创检测智能算法研究[J].光谱学与光谱分析,2024,44(11):3020-3028.

YU X R, ZHAO P, HUAN K W, et al. Research on intelligent algorithm of near-infrared spectroscopy non-invasive detection based on GA-SVR method[J]. Spectroscopy and Spectral Analysis, 2024, 44(11): 3020-3028.

[5] 庄园,高树辉,谢菲,等.基于高光谱成像技术鉴别血痕种属的实验研究[J].激光与光电子学进展,2022,59(16):1630001.

ZHUANG Y, GAO S H, XIE F, et al. Identifying bloodstain species using hyperspectral imaging[J]. Laser & Optoelectronics Progress, 2022, 59(16): 1630001.

[6] 南迪娜,董力强,傅文翔,等.基于拉曼光谱的危险液体快速识别研究[J].光谱学与光谱分析,2021,41(6):1806-1810.

NAN D N, DONG L Q, FU W X, et al. Fast identification of hazardous liquids based on Raman spectroscopy[J]. Spectroscopy and Spectral Analysis, 2021, 41(6): 1806-1810.

[7] 徐军平,杨颖东,王晓婷,等.会理出土战国时期铜矛的铅料来源分析[J].光谱学与光谱分析,2024,44(2):

- 446-451.
XU J P, YANG Y D, WANG X T, et al. A study on lead sources of bronze spearheads from cultural relic administration center of Huili county, Sichuan Province by MC-ICP-MS[J]. Spectroscopy and Spectral Analysis, 2024, 44(2):446-451.
- [8] DONG Q A, WANG W Q, CAO X Y, et al. Plasmonic nanostructure characterized by deep-neural network-assisted spectroscopy[J]. Chinese Optics Letters, 2023, 21(1): 010004.
- [9] 李智缘, 田安红. 基于光谱指数的土壤重金属 Zn 的定量预测与空间分布研究[J]. 光谱学与光谱分析, 2024, 44(11):3287-3293.
LI Z Y, TIAN A H. Quantitative prediction and spatial distribution of soil heavy metal Zn based on spectral indices [J]. Spectroscopy and Spectral Analysis, 2024, 44(11): 3287-3293.
- [10] 叶彬强, 陈昶宏, 曹雪杰, 等. 一种多源光谱融合的水样 COD 实验检测方法 [J]. 光学学报, 2024, 44(12): 1230003.
YE B Q, CHEN C H, CAO X J, et al. COD Detection method of water quality based on multi-Source Spectral Fusion [J]. Acta Optica Sinica, 2024, 44(12): 1230003.
- [11] LAN J J, LIU L L, WANG X, et al. DOM tracking and prediction of rural domestic sewage with UV-vis and EEM in the Yangtze River Delta, China[J]. Environmental Science and Pollution Research, 2022, 29(49):74579-74590.
- [12] ZEESHAN M, ALI O, TABRAIZ S, et al. Seasonal variations in dissolved organic matter concentration and composition in an outdoor system for bank filtration simulation [J]. Journal of Environmental Sciences, 2024, 135: 252-261.
- [13] GEETHA T S, CHELLASWANMY C, RAJA E, et al. Deep learning for river water quality monitoring: a CNN-BiLSTM approach along the Kaveri River. Sustain[J]. Water Resources Management, 2024, 10:125.
- [14] 侯迪波, 张坚, 陈冷, 等. 基于紫外-可见光光谱的水质分析方法研究进展与应用[J]. 光谱学与光谱分析, 2013, 33(7):1839-1844.
HOU D B, ZHANG J, CHEN L, et al. Water quality analysis by UV-Vis spectroscopy: a review of methodology and application [J]. Spectroscopy and Spectral Analysis, 2013, 33(7):1839-1844.
- [15] CHRISTIAN E, BATISTA J R, GERRITY D. Use of COD, TOC, and fluorescence spectroscopy to estimate BOD in wastewater[J]. Water Environment Research, 2017, 89(2):168-177.
- [16] 熊秋燃, 沈鉴, 胡远, 等. 基于水质荧光指纹技术的复合污染河道溯源研究[J]. 光谱学与光谱分析, 2024, 44(6):1773-1780.
XIONG Q R, SHEN J, HU Y, et al. Study on pollution source identification of composite polluted river based on aqueous fluorescence fingerprint technology [J]. Spectroscopy and Spectral Analysis, 2024, 44(6): 1773-1780.

作者简介:

周昆鹏 (1983—), 男, 博士, 副教授, 硕士生导师, 主要从事光谱传感与光电检测方面的研究。