第42卷，第7期 光 谱 学 与 光 谱 分 析 Vol. 42，No. 7，pp2113-2119
2022年7月 Spectroscopy and Spectral Analysis July，2022

分散式农村污水基于三维荧光光谱和紫外-可见全波段吸收光谱的“聚类-回归”COD预测模型

周铭睿[1,2]，曲江北[2]，李 彭[1,2*]，何义亮[1,2]

1. 上海交通大学中英国际低碳学院，上海 201306
2. 上海交通大学环境科学与工程学院，上海 200240

摘 要 基于三维荧光光谱与有机物特征荧光峰之间的关系，提出利用三维荧光光谱进行聚类，再针对不同类的水样利用紫外-可见全波段吸收光谱数据建立COD预测模型的技术路线。比较分析了平行因子分析(PARAFAC)算法和荧光体积积分(FRI)算法两种不同的光谱分析方法，再使用模糊c-均值(FCM)算法进行聚类，并完成了不同类水样的COD预测模型的建立。研究的水样采集于江苏省常熟市周边的农村区域，样品均来自不同的分散式农村生活污水处理装置出水，共100个实验水样；将测得的水样三维荧光光谱数据经过去散射预处理后利用PARAFAC算法和FRI算法分别提取荧光特征数据；之后，利用FCM聚类算法进行相似性聚类；最后，利用偏最小二乘(PLS)算法建立水样的紫外-可见全波段吸收光谱和COD之间的回归和预测模型，并使用决定系数和均方根误差对模型的预测精度进行评价。研究结果表明：未分类、使用FRI、使用PARAFAC算法提取荧光特征信息后再预测的模型的平均决定系数 R^2 分别为0.632，0.819和0.906；平均均方根误差RMSE分别为27.857，23.621和13.071。聚类后的回归和预测精度均得到显著提升，且使用PARAFAC算法提取荧光特征信息后再建模具有最高的预测精度，相比于未分类预测模型的 R^2 提高了0.274。本研究提出的基于三维荧光光谱联合紫外可见全波段吸收光谱，采用“PARAFAC-FCM-PLS”组合算法构建的COD预测模型，可以有效的提高COD的预测精度，为高精度的水质在线监测提供了一种新的思路。

关键词 全光谱；化学需氧量；平行因子分析；模糊c-均值聚类；偏最小二乘法

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)07-2113-07

引 言

近年来，随着国家对环境保护政策的相继提出以及人民的环境保护意识的不断提高，水环境保护问题愈来愈成为人们关注的对象，水环境中的水质在线监测也越来越成为关注的焦点。在环境大数据，环境智能化的趋势下，在线监测设备需要有成本低、监测实时连续、易维护、无污染等特点。无论是城市还是农村，污水中溶解性有机物含量一直是一项重点控制的指标[1]。化学需氧量(COD)作为一种表征污水中有机物含量总体水平的重要指标，有检测精度很高的传统化学法，但传统化学法的检测时间长、维护成本高、所采用的化学试剂具有二次污染等不足之处[2]，无法满足在线监测的需求，且难以大量布设，无法获得实时的数据。尤其是对于农村污水，其采用分散式处理的模式，污水处理设施规模小、设置位置分散、数量多[3]。所以需要寻求一种快速、精确、高效的实时在线监测方法及模式来满足水质的在线监测。

目前，光谱法应用于水质COD的在线监测已具有很多鲜明的优势。与传统化学法相比，光谱法，尤其是紫外-可见光谱法在监测COD方面具有操作简单、检测速度较快、无二次污染、可实现实时连续测量等优势[4]，使得紫外-可见光谱法在COD的监测领域得到了广泛研究。但是，对于组分复杂且种类不同的污水来说，仅仅使用紫外-可见光谱法来预测水质中的COD，其预测精确度和稳定性仍待提高。现有的研究多采用实验室配水来进行光谱法模型的校准和预测，

收稿日期：2021-07-06，**修订日期**：2021-10-11

基金项目：国家重点研发计划项目(2019YFD1100200)资助

作者简介：周铭睿，1997年生，上海交通大学中英国际低碳学院硕士研究生 e-mail：mingruizhou@163.com

*通讯作者 e-mail：lipeng2016@sjtu.edu.cn

配水多为固定的有机物组成，但实际水体中的有机物成分复杂且不固定，所以许多研究缺乏不同有机物组成样本的研究，导致使用光谱法在不同水体环境中应用时存在监测精度较低，难于推广应用的问题。而三维荧光光谱法(excitation-emission matrix，EEM)用来描绘荧光有机物的荧光信息，具有数据量大且完整等特点[5]。三维荧光光谱的荧光数据解析得到激发光谱矩阵、发射光谱矩阵以及得分矩阵，其中得分矩阵在一定条件下正比于荧光物质浓度，可进行半定量表征[6]。且不同类型的有机物在三维荧光光谱上的峰位置有显著差异，因此可以利用不同水样的三维荧光光谱，按照有机物组成的近似度划分类别。三维荧光体积积分法对特定的荧光区域进行标准体积积分可以间接表示不同组分的相对浓度[7-8]。高连敬等[9]以三维荧光光谱技术为手段，结合荧光区域积分(FRI)方法，证明其可以有效监测和分析水体中低浓度有机物的去除情况，可以作为一种有效的技术手段，用于净水厂的日常运行和水质监测。孔德明[10]等利用平行因子分析(PARAFAC)方法分解去散射后的三维荧光光谱后的数据，实现了对污染物的快速、有效的检测。但是，目前没有公认的用于三维荧光光谱数据特征分析处理的方法，也没有将污水水样分类再建立预测模型的尝试。我们的研究尝试将这两种方法进行对比，观察对 COD 预测效果的影响。

以实际生活污水为研究对象，对水样的三维荧光光谱分别使用荧光体积积分(FRI)算法、平行因子分析(PARAFAC)算法，提取水样的荧光特征信息再使用 FCM 算法进行水样的聚类。对聚类后不同类别水样的紫外-可见全波段光谱和 COD 数据进行偏最小二乘法(PLS)模型的回归及预测，从而建立一种全新的“聚类-回归”COD 预测模型，具体过程如图 1 所示。

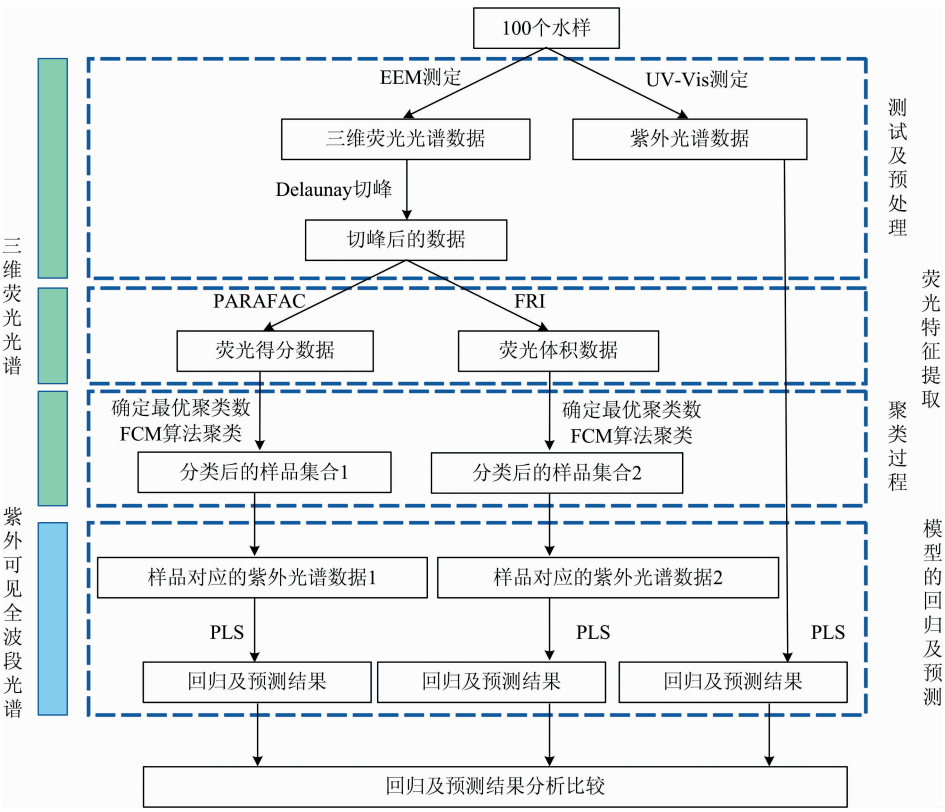

图 1 模型设计流程图

Fig. 1 Flow chart of model design

1 实验部分

1.1 样本采集与光谱检测

水样采集地点为江苏省常熟市，采集地点为常熟市周边的农村区域，采集时间为 2019 年 3 月 10 日。为满足样品有机物组成的多样性，采集时选取 100 个分散式农村生活污水处理装置出水作为采集点，每个采集点采 1 个水样，具体的采集信息见图 2。采集后的水样使用 250 mL 聚乙烯瓶在 4 ℃低温条件下贮存，样品的 COD 浓度使用国标法测定。

水样的紫外可见光谱数据由 HACH DR/6000 光谱仪扫描得到，扫描范围为 200～1 000 nm，间隔为 1 nm。水样的三维荧光光谱数据由日立 F-7000 荧光分光光度计扫描得到，激发波长的扫描范围为 200～500 nm，间隔为 5 nm；发射波长的扫描范围为 250～550 nm，间隔为 5 nm。为了避免仪器本身的散射对三维荧光光谱测试的影响，设置初始的发射波长滞后于初始的激发波长 50 nm。

1.2 数据处理

1.2.1 平行因子分析算法(PARAFAC)

平行因子分析(PARAFAC)方法是一种基于三线性模型

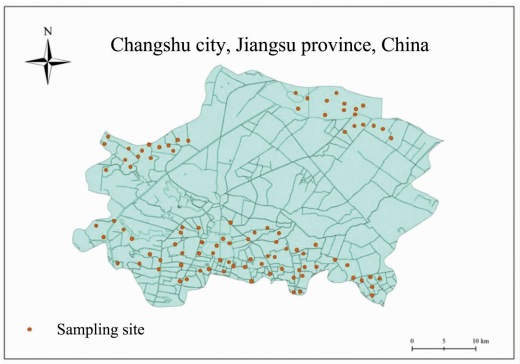

图 2　常熟市采样分布图

Fig. 2　The map of sampling sites in Changshu

实现多维数据矩阵分解的经典迭代算法。传统的三维荧光光谱数据通常采用寻峰法进行特征荧光团的识别，但对于多组分水样通常会有峰重叠的现象，造成荧光峰被全部或部分掩盖的情况，导致检测结果误差偏大[11]。使用平行因子分析(PARAFAC)方法首先需要建立一个三维矩阵 $\boldsymbol{X}$，矩阵类型为 $I\times J\times K$。其中 I 和 J 分别是三维荧光光谱的激发波长和发射波长的扫描个数。三线性模型分解过程可以表示为

$$x_{ijk}=\sum_{m=1}^{M}a_{im}b_{jm}c_{km}+e_{ijk} \tag{1}$$

式(1)中，$i=1,2,\cdots,I$；$j=1,2,\cdots,J$；$k=1,2,\cdots,K$；x_{ijk} 为三维荧光光谱矩阵 $\boldsymbol{X}$ 中的元素；a_{im} 为相对激发光谱矩阵中的任一元素；b_{im} 为相对发射光谱矩阵中的任一元素；c_{im} 为相对浓度矩阵中的任一元素；e_{ijk} 为残差矩阵中的任一元素；M 为得分矩阵、负荷矩阵的列数。

1.2.2　荧光体积积分算法(FRI)

将三维荧光光谱的等高线图分为 5 个连续的区域Ⅰ，Ⅱ，Ⅲ，Ⅳ和Ⅴ。王聪颖等在研究中指出，荧光光谱的特定区域可以间接反映水体中部分可溶性有机物，区域Ⅰ(Ex<250 nm，Em<330 nm)，区域Ⅱ(Ex<250 nm，330 nm<Em<380 nm)，区域Ⅲ(Ex<250 nm，380 nm<Em<550 nm)，区域Ⅳ(Ex>250 nm，Em<380 nm)，区域Ⅴ(Ex>250 nm，380 nm<Em<550 nm)[7]。三维荧光物质区域分布见图 3，各个区域的区域积分表达式可以表示为

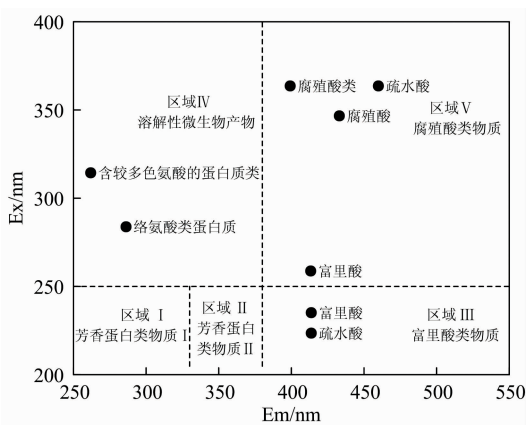

图 3　三维荧光物质区域分布

Fig. 3　Three-dimensional map of the regional distribution of fluorescent substances

$$\varphi_i=\sum_{Ex_{\min}}^{Ex_{\max}}\sum_{Em_{\min}}^{Em_{\max}}E(\lambda_{Ex}\lambda_{Em})\Delta d\lambda_{Ex}\Delta d\lambda_{Em} \tag{2}$$

式(2)中，$i=1,2,\cdots,5$；φ_i 是区域 i 的区域积分和；$E(\lambda_{Ex}\lambda_{Em})$为三维荧光光谱在激发波长 λ_{Ex} 和发射波长 λ_{Em} 处的强度值；$\Delta d\lambda_{Ex}$ $\Delta d\lambda_{Em}$ 分别为激发波长 λ_{Ex} 和发射波长 λ_{Em} 的积分增量。

1.2.3　最优聚类数与 FCM 算法

聚类通常是指运用特定标准(如距离标准)将数据集分割成为不同的类或者簇，使得簇内的数据相似度尽可能高，簇间的数据相似度尽可能小，最终使得特征高度相似的数据相聚成簇。聚类与分类不同，聚类是一种无监督学习模式，即不需要给定特定的数据划分特征，在聚类过程中即可自聚类成簇。对于成分复杂的水样来说，其特征是不明确的，因此使用聚类方法可以对不同特征的样品进行区分。

最优聚类数的确定也是依据距离标准，通过簇内离差矩阵来描述数据的紧密度，通过簇间离差矩阵来描述数据的分离度。簇内簇间离差度比值指标 D 的定义为

$$D(i)=\frac{\mathrm{tr}A(i)/(i-1)}{\mathrm{tr}B(i)/(n-i)} \tag{3}$$

式(3)中，n 表示聚类的数目；i 表示当前所运算的类；$\mathrm{tr}A(i)$ 表示簇内离差矩阵的迹；$\mathrm{tr}B(i)$表示簇间离差矩阵的迹。

FCM 算法又称模糊 C-均值聚类算法，是基于目标函数最优的聚类算法。通过隶属度函数来确定数据间的相似度，算法的目标函数和约束条件可以描述为

$$\min J=\sum_{i=1}^{m}\sum_{j=1}^{n}u_{ij}^{k}\parallel x_j-c_i\parallel^2 \tag{4}$$

$$\sum_{i=1}^{m}u_{ij}=1,\ j=1,2,\cdots,n \tag{5}$$

式中，m 为聚类数目，即最佳聚类数；n 为数据总数；u_{ij} 为每个样本 j 属于某一类 i 的隶属度。

2　结果与讨论

2.1　光谱预处理

对荧光光谱影响较大的拉曼散射和瑞利散射在使用光谱仪对水样进行测定时无法被直接去除，两种散射的存在可能会导致使用荧光体积积分(FRI)算法和平行因子分析(PARAFAC)算法进行分析时有效光谱信息被掩盖，导致分析结果产生严重偏差，所以在进行光谱信息分析前需要去除散射的干扰。分析图 4(a)和(b)：通过 MATLAB R2018b，使用 Delaunay 三角形内插值方法可以有效去除两种散射对荧光光谱的影响，使本身的荧光信息更加明显。

2.2　基于荧光体积积分和平行因子分析算法

采用荧光体积积分(FRI)算法对预处理后得到的三维荧光矩阵 $\boldsymbol{X}$ 进行分析，矩阵 X 结构为 100×61×61(100 为样品数量，61 为激发波长数量，61 为发射波长数量)。荧光积分区域依据可被荧光所反映的水体中的溶解性有机物质分为 5 个区域，分别为芳香蛋白类物质Ⅰ区域、芳香蛋白类物质Ⅱ

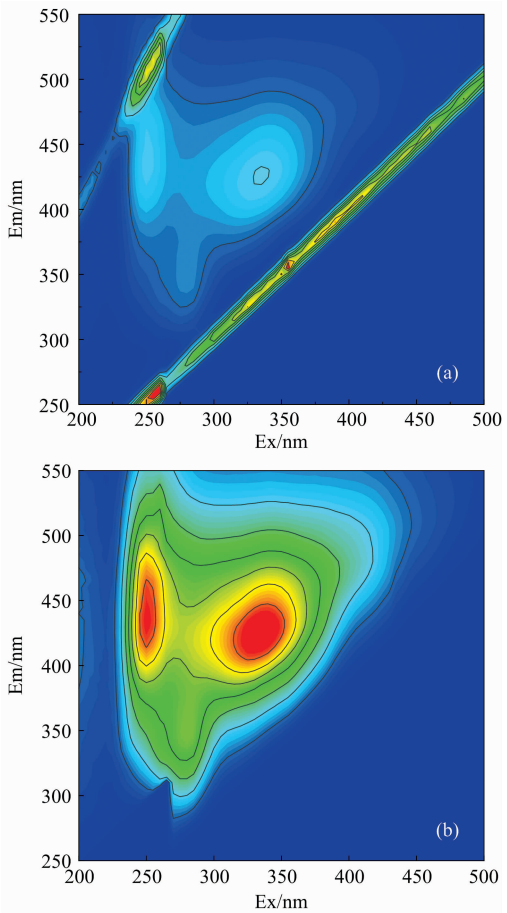

图 4 水样的荧光光谱图

(a)：去除散射前的三维荧光光谱；(b)：去除散射后的三维荧光光谱

Fig. 4 Fluorescence spectra of water samples

(a)：Three-dimensional fluorescence spectra before removal of scattering；(b) Three-dimensional fluorescence spectra after scattering removal

区域、富里酸类物质区域、溶解性微生物代谢产物区域、腐殖酸类区域。所以由荧光体积积分(FRI)算法得到的荧光特征信息矩阵为二维矩阵 $\boldsymbol{X}_1$(100×5)。但是使用荧光体积积分(FRI)算法就默认了每个水样均有 5 个荧光特征区域，且重叠的荧光信息无法分开，可能会造成提取的荧光特征信息出现部分冗余，对之后的聚类过程造成一定的影响。

采用平行因子分析(PARAFAC)算法对预处理后得到的三维荧光矩阵 X 进行分析，对 100 个样本进行荧光数据的杠杆验证时发现 33，34 和 49 号样品的验证杠杆值明显偏离其他样品，结果如图 5(a)所示，应当剔除此三种样品。对剔除异常样品的 97 组数据进行平行因子分析，利用对半分析验证不同组分情况下的模型稳定性，分别验证了 2～7 组分下模型的稳定性，由计算结果得出只有 3 组分情况下模型是稳定的。所以由平行因子分析(PARAFAC)算法得到的荧光特征信息矩阵为二维矩阵 $\boldsymbol{X}_2$(97×3)。由图 4(b)和(c)可知三个特征荧光峰的位置：第一个特征荧光峰激发/发射波长为 335/420 nm；第二个特征荧光峰激发/发射波长为 255/470 nm；第三个特征荧光峰激发/发射波长为 280/350 nm。使用平行因子(PARAFAC)算法对三维荧光矩阵进行处理时，可以去除与其他水样荧光信息有明显差异的异常水样，并应用对半分析方法模型对选取的特征组分数进行稳定性的验证，保证了选取的特征组分数为最优组分数，使荧光信息的特征更加的明显。此外，平行因子分析(PARAFAC)算法还能将重叠的荧光特征峰进行数据层面的分离，保证了特征信息不出现冗余的情况，使之后的聚类效果更优。

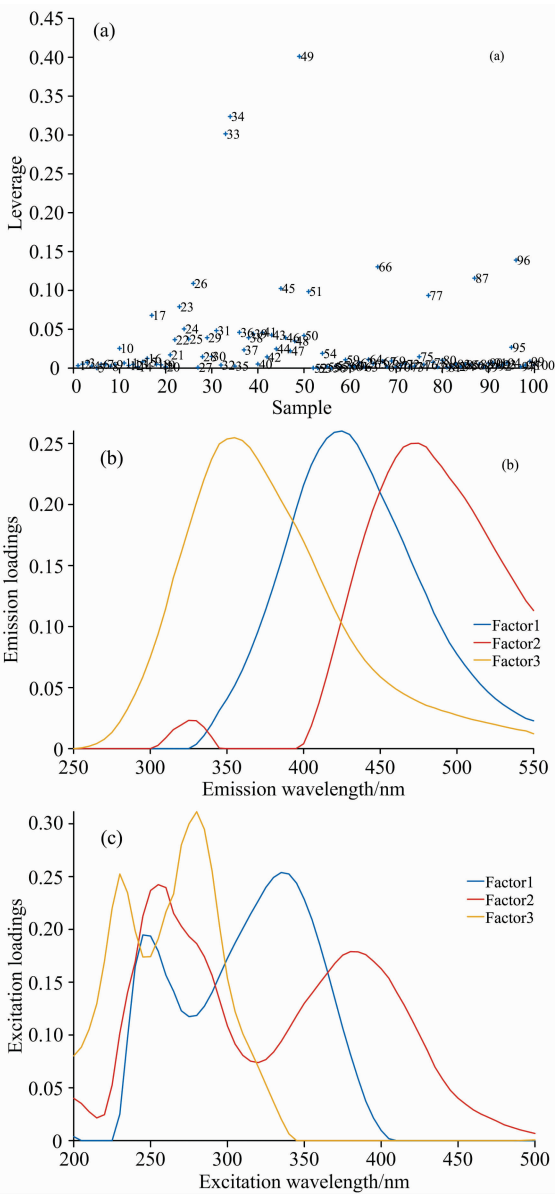

图 5 水样的平行因子分析

(a)：样品的杠杆分析；(b)：组分数为 3 的发射波长对半分析；(c)：组分数为 3 的激发波长对半分析

Fig. 5 PARAFAC of water samples

(a)：Leverage analysis of the sample；(b)：The split half analysis of emission wavelength with factors of 3；(c)：The split half analysis of excitation wavelength with factors of 3

2.3 基于模糊 c-均值聚类算法的聚类结果分析

为了更好地将水样依据 FRI 算法和 PARAFAC 算法提

取出来的荧光特征信息进行聚类。首先应该选取最优聚类数，利用基于距离指标的最优聚类数选取方法，分别对使用了 FRI 算法和 PARAFAC 算法进行荧光特征提取的水样进行最优聚类数选取。如图 6(a)可知，使用 FRI 算法提取荧光特征的水样的最优聚类数为 3。如图 6(b)可知，使用 PARAFAC 算法提取荧光特征的水样的最优聚类数为 4。

其次在 MATLAB R2018b 中使用 FCM 算法分别对 FRI 算法和 PARAFAC 算法分析得到的荧光特征数据进行 3 类别和 4 类别的聚类，聚类结果由图 6(c)和(d)所示。其中，将 FRI 算法得到的荧光特征数据分为 3 类：第一类 57 个样品、第二类 34 个样品、第三类 9 个样品，总共 100 个样品。将 PARAFAC 算法得到的荧光特征数据分为 4 类：第一类 36 个样品、第二类 5 个样品、第三类 29 个样品、第四类 27 个样品，总共 97 个样品。具体的分类结果由表 1 所示，由表 1 中的结果可知，两种方法提取出的荧光特征数据主要特征较为相似，所以每一类的样品重合率均较高。但使用 FRI 算法提取特征信息再聚类后，每一类的样品在重复的样品之外出现了不少冗余样品，可能是 FRI 算法在处理荧光光谱数据时未剔除与特征荧光峰重叠的干扰信息造成的。

表 1　具体的聚类结果

Table 1　Specific clustering results

类别	样品数目		相同的样品数		重合率/%	
	PARAFAC	FRI	PARAFAC	FRI	PARAFAC	FRI
第一类	36	57	27	27	75	47.40
第二类	5	9	5	5	100	55.60
第三类	29	34	24	24	82.80	70.60
第四类	27	0	0	0	0	0

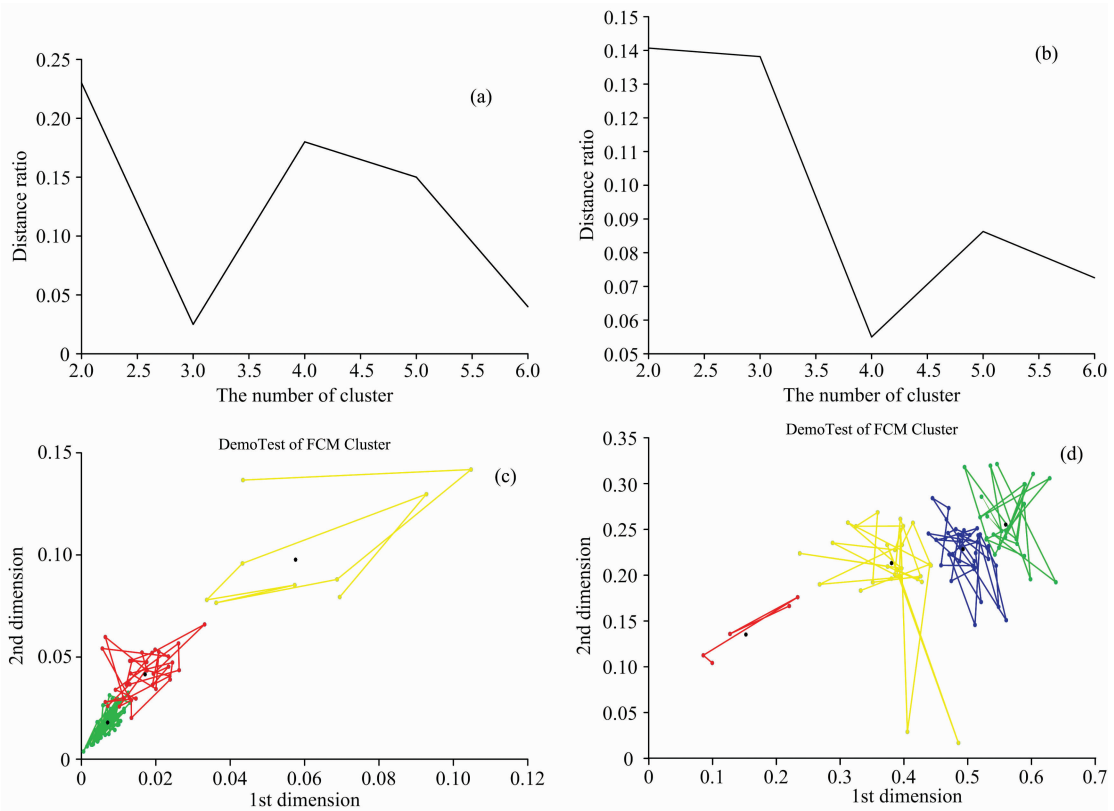

图 6　FCM 聚类分析结果

(a)：使用 FRI 后的最优聚类数选择；(b)：使用 PARAFAC 后的最优聚类数选择；(c)：使用 FRI 后的聚类结果；(d)：使用 PARAFAC 后的聚类结果

Fig. 6　FCM cluster analysis results

(a)：Selection of the optimal cluster number after using FRI；(b)：Selection of the optimal cluster number after using PARAFAC；(c)：Clustering results after using FRI；(d)：Clustering results after using PARAFAC

2.4　聚类后的模型拟合及预测

利用 The Unscrambler X 软件中的偏最小二乘法(PLS)对聚类后的各类水样的紫外-可见全波段光谱与对应的 COD 数据进行模型的回归及预测，回归及预测结果如表 2 所示。其中决定系数 R^2 表示因变量的变异部分可由自变量的变异来解释，R^2 越接近 1，模型的参考价值越高；均方根误差 RMSE 指预测值与真实值的偏离程度，RMSE 越小，模型的参考价值越高[12]。由表 2 结果可知，无论是运用 FRI 算法还是 PARAFAC 算法提取的荧光特征数据聚类后再建立紫外-可见全波段光谱和对应的 COD 数据之间的偏最小二乘(PLS)模型回归结果均优于未分类的结果，说明对水样进行聚类后再建模能有效提高模型的精度与稳定性；且由表 2 可

知，由平行因子分析(PARAFAC)算法提取的荧光特征数据进行聚类后再建立紫外-可见全波段光谱和对应的 COD 数据之间的偏最小二乘(PLS)预测模型的 COD 预测精度高于由荧光体积积分(FRI)算法提取的荧光特征数据进行聚类后再建立紫外-可见全波段光谱和对应的 COD 数据之间的偏最小二乘(PLS)预测模型的 COD 预测精度，说明平行因子分析(PARAFAC)算法提取出的荧光特征数据的特征性优于由荧光体积积分(FRI)算法提取出的荧光特征数据；由平行因子分析(PARAFAC)算法聚类后的每一类类内的水样特征相似度高于由荧光体积积分(FRI)算法聚类后的每一类类内的水样特征相似度。可能是平行因子分析(PARAFAC)算法将与荧光特征峰重叠的无效荧光信息分开，而荧光体积积分(FRI)算法直接进行荧光的积分，未考虑重叠的无效荧光信息的影响，使得使用荧光体积积分(FRI)算法提取的荧光特征数据存在信息冗余。使用平行因子分析(PARAFAC)算法结合 FCM 聚类后的聚类效果优于使用荧光体积积分(FRI)算法结合 FCM 聚类后的聚类效果，类内水样的相似度更高，使得预测精度更高。

表 2 模型的回归及预测结果

Table 2 The model fitting and prediction results

类别	回归						预测					
	未分类		FRI		PARAFAC		未分类		FRI		PARAFAC	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
第一类	—	—	0.86	17.169	0.923	10.315	—	—	0.852	17.618	0.954	14.252
第二类	—	—	0.796	26.177	0.999	1.122	—	—	0.787	38.484	—	—
第三类	—	—	0.819	13.19	0.918	10.962	—	—	0.819	14.761	0.902	5.937
第四类	—	—	—	—	0.921	13.623	—	—	—	—	0.863	19.025
均值	0.657	27.335	0.825	18.845	0.940	9.006	0.632	27.857	0.819	23.621	0.906	13.071

3 结 论

利用三维荧光光谱结合平行因子分析(PARAFAC)算法和 FCM 聚类算法对聚类后的水样的紫外-可见全波段光谱和相应的 COD 数据运用偏最小二乘(PLS)模型进行回归和预测。研究结果表明，使用 Delaunay 三角形内插值法去除拉曼和瑞利散射后，使用平行因子分析(PARAFAC)算法提取荧光特征信息并使用 FCM 算法聚类，使用聚类后各类水样的紫外-可见全波段光谱数据和相应 COD 数据进行偏最小二乘(PLS)模型的回归和预测，具有最佳的拟合和预测结果，回归平均 R^2 为 0.940，平均 RMSE 为 9.006，预测平均 R^2 为 0.906，平均 RMSE 为 13.071，相比于未分类的 PLS 模型预测平均 R^2 值 0.632，R^2 提高了 0.274。本研究提供了一种使用三维荧光光谱数据，利用平行因子分析(PARAFAC)算法提取荧光特征的先聚类再建模的方法，可为水样的快速检测提供一种新思路。但是，此方法由于需要测量样品的三维荧光光谱数据和紫外-可见全波段光谱数据，在实时在线监测的设备化方面仍需改进，后续可以进一步研究，以提高此方法的实际应用性。

References

[1] SHUAI Lei, LI Wei-hua, SHEN Hui-yan, et al(帅 磊, 李卫华, 申慧彦, 等). Chinese Journal of Environmental Engineering(环境工程学报), 2016, 10(4): 2127.

[2] CAI Xin, WU Shao-feng, LI Dong-bo(蔡 鑫, 吴绍锋, 李东波). Machine Design and Manufacturing Engineering(机械设计与制造工程), 2019, 48(9): 95.

[3] ZHU Hui-feng(朱慧峰). Water Purification Technology(净水技术), 2018, 37(8): 39.

[4] LUO Ji-yang, WEI Biao, TANG Bin, et al(罗继阳, 魏 彪, 汤 斌, 等). Environmental Science and Technology(环境科学与技术), 2015, 38(S2): 246.

[5] Li Lingfei, Liu Ting, Dong Huiyu, et al. Chemosphere, 2021, 283: 131198.

[6] JIANG Dan-yang, PENG Wei-xuan, LIAO Si-ying, et al(蒋丹阳, 彭玮瑄, 廖思颖, 等). Acta Scientiae Circumstantiae(环境科学学报), 2021, 41(6): 2201.

[7] WANG Cong-ying, CHEN Wei, TAO Hui, et al(王聪颖, 陈 卫, 陶 辉, 等). Water Purification Technology(净水技术), 2019, 38(3): 56.

[8] QIAN Feng, WU Jie-yun, YU Hui-bin, et al(钱 锋, 吴婕赟, 于会彬, 等). Environmental Chemistry(环境化学), 2016, 35(10): 2016.

[9] GAO Lian-jing, DU Er-deng, CUI Xu-feng, et al(高连敬, 杜尔登, 崔旭峰, 等). Water & Wastewater Engineering(给水排水), 2012, 48(10): 51.

[10] KONG De-ming, SONG Le-le, CUI Yao-yao, et al(孔德明, 宋乐乐, 崔耀耀, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(9): 2798.

[11] HUANG Ting-lin, FANG Kai-kai, ZHANG Chun-hua, et al(黄廷林, 方开凯, 张春华, 等). Environmental Science(环境科学), 2016, 37(9): 3394.
[12] QU Jiang-bei, LI Peng, HE Yi-liang, et al(曲江北, 李　彭, 何义亮, 等). Water Purification Technology(净水技术), 2020, 39(7): 65.

The “Cluster-Regression” COD Prediction Model of Distributed Rural Sewage Based on Three-Dimensional Fluorescence Spectrum and Ultraviolet-Visible Absorption Spectrum

ZHOU Ming-rui[1,2], QU Jiang-bei[2], LI Peng[1,2*], HE Yi-liang[1,2]
1. China-UK Low Carbon College, Shanghai Jiaotong University, Shanghai 201306, China
2. School of Environmental Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China

Abstract Based on the relationship between the three-dimensional fluorescence spectrum and the characteristic fluorescence peaks of organic matter, this study proposed to use the three-dimensional fluorescence spectrum for clustering and then for different kinds of water samples, using UV-Vis full-band absorption spectrum data to establish the COD prediction model technical route. The parallel factor analysis (PARAFAC) algorithm and fluorescence volume integration (FRI) algorithm were compared and analyzed, and then the fuzzy c-means(FCM) algorithm was used for clustering, and the COD prediction model of different water samples was established. The water samples in this study were collected from the rural areas around Changshu City, Jiangsu Province, and 100 experimental water samples were collected from the effluent of different distributed rural domestic sewage treatment plants. The measured three-dimensional fluorescence spectrum of water samples was pretreated by de-scattering, and then the fluorescence characteristic data were extracted by the PARAFAC algorithm and FRI algorithm, respectively. Then, the FCM clustering algorithm was used for similarity clustering. Finally, the partial least squares (PLS) algorithm was used to establish the regression and prediction model between the UV-Vis full-band absorption spectrum and COD of water samples, and the prediction accuracy was evaluated by the coefficient of determination and the root mean square error (RMSE). The results showed that the prediction models' mean determination coefficients(R^2) were 0.632, 0.819 and 0.906, respectively, after the fluorescence feature information was extracted using FRI and PARAFAC algorithms. RMSE were 27.857, 23.621 and 13.071, respectively. The regression and prediction accuracy was significantly improved after clustering, and the modeling established after the extraction of fluorescence feature information using the PARAFAC algorithm had the highest prediction accuracy, which was 0.274 higher than theR^2 of the unclassified prediction model. The proposed COD prediction model based on a three-dimensional fluorescence spectrum combined with UV-Vis full-band absorption spectrum and using the combined algorithm of “PARAFAC-FCM-PLS” can effectively improve the prediction accuracy of COD and provide a new idea for high precision online monitoring of water quality.

Keywords Full spectral; Chemical oxygen demand; Parallel factor analysis; Fuzzy c-means classification; Partial least squares

(Received Jul. 6, 2021; accepted Oct. 11, 2021)

* Corresponding author