

三维荧光结合自组织映射神经网络考察自来水厂有机物去除效果

杜尔登^{1,2}, 郭迎庆², 孙悦², 高乃云^{1*}, 王利平²

1. 同济大学污染控制与资源化研究国家重点实验室, 上海 200092
2. 常州大学环境与安全工程学院, 江苏 常州 213164

摘要 三维荧光光谱在水体监测和水处理领域日益引起广大研究者的关注。自组织映射神经网络(SOM网络)作为一种非监督、自学习的神经网络,具有自稳定性高、抗噪声能力强等特点。使用SOM网络对某自来水厂处理流程中水样的荧光光谱进行解析,可以将三维荧光光谱聚类成三类,分别对应为络氨酸类蛋白有机物、色氨酸类蛋白有机物、紫外富里酸类物质。整个自来水厂处理工艺能够有效的去除水体中的有机物,其中络氨酸类、色氨酸类、紫外富里酸类物质的去除率分别为84.6%, 79.9%, 69.1%。研究结果表明,SOM网络可以作为一种有效的水体荧光光谱分析工具,有助于优化水处理工艺参数,提高水处理工艺性能、以及自来水厂的监测和管理。

关键词 自来水处理; 三维荧光(3D-EEM); 自组织映射神经网络(SOM); 有机物去除

中图分类号: X830.2 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2012)07-1846-06

引言

自来水厂原水中有机污染物浓度非常低,常见的水质指标,如高锰酸盐指数、UV₂₅₄等无法有效对水处理过程中有机物的去除情况进行充分的评估。三维荧光光谱(excitation-emission matrix, EEM)技术由于其高度的灵敏性,不破坏样品结构,在水体监测和水处理领域日益引起研究者的关注^[1]。三维荧光光谱中包含有极为丰富的荧光信息,常见的荧光光谱分析方法有寻峰法(peak-picking)、平行因子分析法(PARAFAC)^[2,3]、主成分分析法(PCA)^[4]、体积积分法(FRI)^[5,6]等。寻峰法是较为普遍的光谱分析方法,但是寻峰法只考虑三维光谱中的特定峰值,大量荧光数据并没有充分得到使用^[7]。平行因子分析法处理荧光数据时收敛速度慢,对噪声或模型偏差较为敏感。因此对大量荧光光谱信息进行处理,构建稳健的光谱模型,发展稳健的三维荧光光谱解析方法依旧是亟待解决的问题。

自组织映射神经网络(self-organizing map, SOM),也称为Kohonen神经网络,属于非监督、自学习的神经网络,广泛应用于模式识别、特征提取、数据压缩等领域。已有研究者将SOM网络用于环境水体检测、水质评价等方面^[8,9],但

少有将SOM网络用于水体三维荧光光谱的解析。本研究采用SOM网络,结合K-means算法,对自来水厂工艺流程中不同水样的三维荧光光谱进行解析,提取有效的荧光光谱特征,结合其他水质指标,对水处理过程中有机物的去除情况进行综合评估,以优化水处理工艺参数,提高水处理工艺性能。

1 实验部分

1.1 水样采集

水样来自江苏南部太湖地区某自来水厂,水源为太湖。由于近年来太湖水域一直存在水体富营养化问题,此自来水厂在传统水处理工艺基础上又增加了臭氧、生物活性炭等深度处理工序,以强化对水体有机污染物的去除,水厂处理流程见图1。

水样采集时间为2011年4月,从每个水处理单元出水中取3个平行样,折板絮凝和平流沉淀池为一个单体,记为从沉淀池取水。总共采集18个水样,将水样采集后放入干净聚四氟乙烯塑料瓶内,迅速带回实验室,0.45 μm滤膜过滤后置于棕色玻璃瓶中,在4℃冰柜中保存待用。所有水样分析在4d内完成。

收稿日期: 2012-02-12, 修订日期: 2012-04-10

基金项目: 国家水体污染控制与治理科技重大专项(2008ZX07421-002), 国家自然科学基金项目(41101233)和常州大学自主科研项目(ZMF10020085, 102002)资助

作者简介: 杜尔登, 1978年生, 同济大学污染控制与资源化研究国家重点实验室博士后 e-mail: duerdeng@gmail.com

*通讯联系人 e-mail: gaonaiyun@sina.com

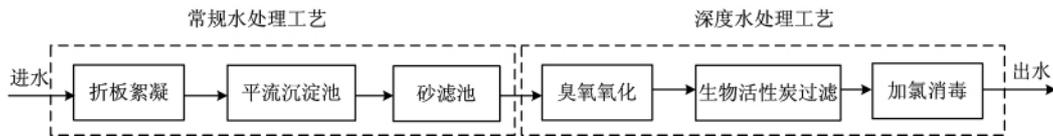


Fig 1 Process flow of water treatment plant

1.2 三维荧光光谱测量、水质指标和数据预处理

水样三维荧光光谱由荧光分光光度计(Cary Eclipse, 美国安捷伦)测量和采集。测量波长范围: 激发波长(λ_{ex})220~400 nm, 增量 5 nm; 发射波长(λ_{em})280~500 nm, 增量 2 nm; 狭缝宽度 5 nm, PMT 电压 600 V, 扫描速度 1 200 nm·min⁻¹, 水样在 1 cm 石英荧光比色皿中测量。实验空白水为 Milli-Q 超纯水(Millipore, 18.3 M Ω ·cm)。

在对荧光光谱进行解析前, 首先需要对荧光数据进行预处理, 以消除瑞利和拉曼散射的影响, 提高 EEM 光谱解析效率。将瑞利散射上方光谱数据置零, 以去除瑞利散射的影响。此外, 以空白水样做参比, 扣除空白水样光谱数据, 以消除拉曼散射的影响。

1.3 SOM 网络和解析荧光光谱过程

SOM 网络是由芬兰学者 Kohonen 于 1981 年提出的一种无监督学习的神经网络模型, 分成上、下两层: 下层为输入层, 上层为输出层(或映射层)^[10]。输出层的每个神经元同它周围的其他神经元侧向连接, 排列成棋盘状平面; 输入层为单层神经元排列(见图 2)。

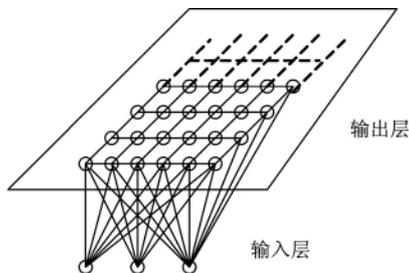


Fig 2 Schematic diagram of SOM model

使用 SOM 网络对三维荧光光谱进行解析, 结合研究中使用的荧光光谱数据, 主要解析过程如下。

第一步, 荧光数据降维。在使用 SOM 网络进行分析之前, 需要把已经预处理过的 18 个水样的三维荧光光谱展开, 转换成二维向量。展开后, 产生维度为 2 109×18 的矩阵, 其中: 列表示展开的激发-发射波长数据组, 行表示所要处理的水样个数。

第二步, 数据标准化。将二维向量进行标准化处理, 保证标准化后的数据平均值为 0, 方差为 1, 以避免数量级不同带来的对训练结果的影响。数据准备完成后, 数据样本被转化成一个标准化的 SOM 数据结构, 这就是训练网络的输入数据。

第三步: SOM 网络初始化、训练。初始化包括权值向量、相应训练参数的初始化。训练采用高斯函数批量训练方式, 分粗调和精调两个阶段。经过学习和训练, 输入的每一类荧光数据都会在神经网络上有特定的映射, 最终获得荧光

数据的映射神经元。

第四步: SOM 网络聚类分析。利用 K-means 算法对 SOM 网络的竞争层神经元的权值进行分类, 以 DBI 值(davies-bouldin index)自动选择聚类数。通过计算各神经元之间的欧式距离, 获得的最小欧式距离为每一类神经元的中心区域, 然后联合每类中的多个竞争层神经元权值作为每类的代表性特征向量集, 从而间接表征了荧光光谱所含组分的相对浓度。

三维荧光 SOM 网络的构建、验证, 以及分析结果的可视化在 Matlab 7.0 和 SOM Toolbox 2.0 软件平台上完成。SOM 网络输入层中的输入向量包含 18 个样本的 2109 个激发-发射波长对。根据输入向量最大两个特征值确定输出层的神经元为 225 个(地图大小[25, 9])。

2 结果与讨论

2.1 水处理流程中各水样三维荧光光谱的变化

进水、以及各处理单元出水的三维荧光光谱见图 3。根据传统寻峰法, 采用研究者广泛采用的 Cobel 分类标准^[11], 从图 3(a)可以看出, 原水荧光光谱中有 2 个非常明显的荧光峰: 峰 T1(ex/em 280 nm/305 nm), T2(ex/em 230 nm/330 nm), 这两个峰均属于类蛋白荧光峰, 分别对应络氨酸类、色氨酸类蛋白有机物。此外, 原水荧光光谱中还有一个面积较广、峰值不突出的峰 A(ex/em 215~240 nm/400~420 nm), 通常认为这是紫外富里酸类有机物所表现出来的荧光特征。

从图 3 中可以看出, 原水中有机物的构成主要是以络氨酸类、色氨酸类蛋白有机物为主, 紫外富里酸类有机物的含量比较低。这和饮用水原水中通常以富里酸类、腐植酸类物质为主的有机物特征是不一样的^[12]。络氨酸类、色氨酸类蛋白质与芳环氨基酸结构有关, 主要是细菌分解过程中产生的酶或者生物残骸中含有的大量蛋白质。此自来水厂原水来自太湖, 太湖承接了周围区域污水厂的尾水排放, 以及附近污水的直接排放。污水厂生化处理过程中产生微生物代谢产物, 太湖周边城市大量工业废水和生物污水的排放, 携带了大量有机物进入太湖, 从而导致太湖水体中呈现出强烈的类蛋白有机物特征。这和 Song^[13]和 Wang^[14]等对太湖水体荧光光谱特征的考察基本是一致的。

此外, 从图 3(a)~(f)中可以看出, 随着水处理流程的进行, 水样的三维荧光光谱的特征发生明显的变化, 峰值强度逐渐减小, 荧光峰甚至消失。当原水经过混凝沉淀池[图 3(b)]、砂滤池[图 3(c)]传统处理工艺后, 3 个峰的荧光强度略有降低, 这说明传统水处理工艺对水体中有机物的去除能力有限, 无法有效去除水体中的有机物。在深度处理流程

中, 当水体经过臭氧氧化工艺[图 3(d)]后, 3 个荧光峰**的强度明显降低**, 尤其是经过生物活性炭过滤[图 3(f)]后, 蛋白荧光峰和紫外富里酸荧光峰基本消失, 这说明深度处理工艺

(臭氧、生物活性炭)能够有效的去除水体当中的污染物, 从而保障了饮用水安全。

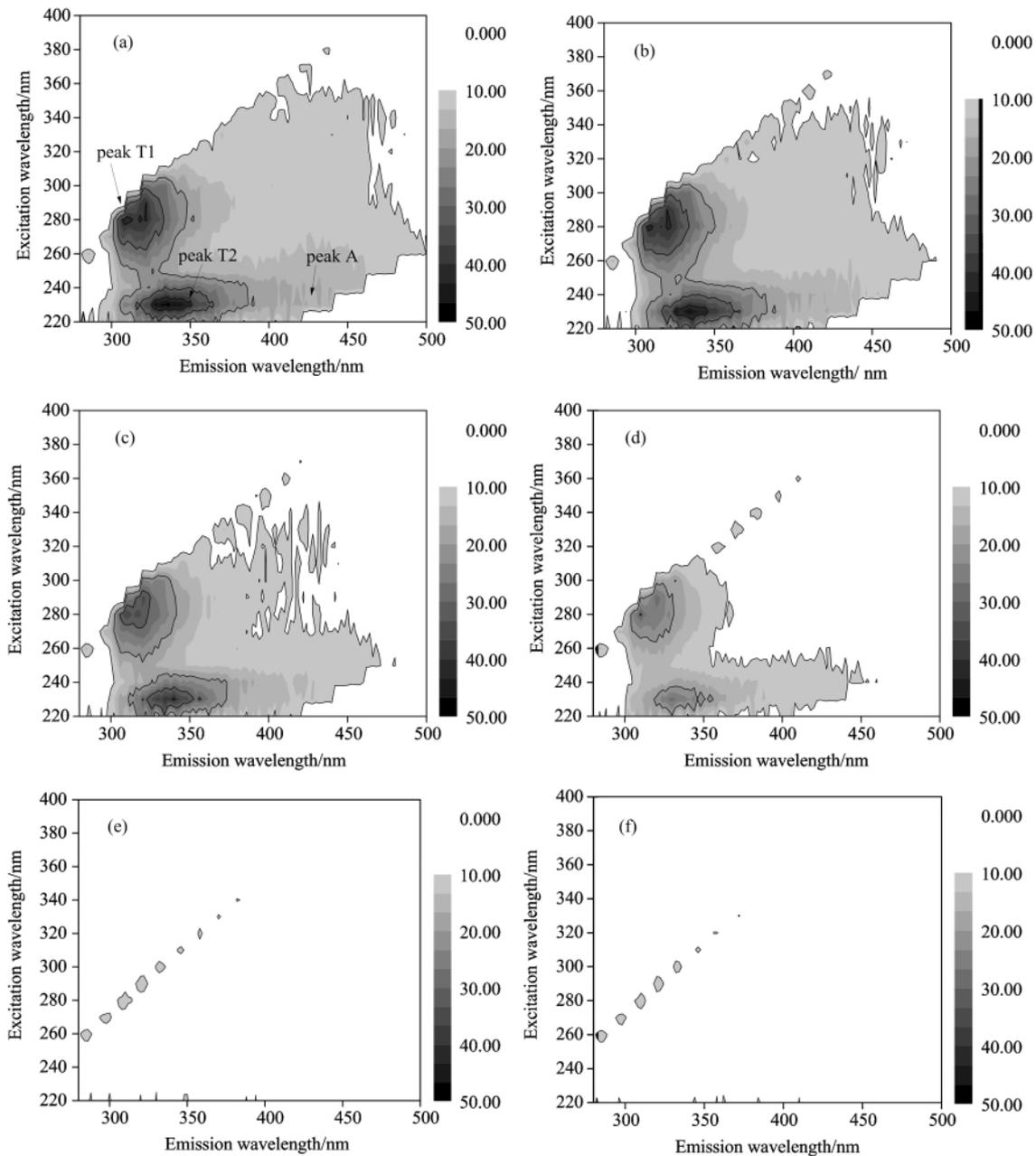


Fig. 3 EEM spectra of water samples from different water treatment units

(a): Raw water; (b): Sedimentation effluent; (c): Sand filtration effluent; (d): Ozonation effluent; (e): Biological activated carbon (BAC) filter effluent; (f): Cl_2 disinfection effluent (Final outlet)

2.2 SOM 网络模型的构建、训练和聚类过程

基于 Matlab 7.0 以及 SOM Toolbox 2.0 来实现荧光光谱 SOM 网络的初始化、训练和可视化输出。SOM 网络的训练是一个无监督的过程, 在这个过程中不断对输入层、输出层神经元的连接权值进行修正和调整, 输出层经过优化, 包含 225 个神经元(地图大小 [25, 9]), 最终量化误差和最终图形误差分别为 0.526 和 0.07。

图 4 显示了经过训练的 SOM 网络的部分图形, 包括 U 矩阵[图 4(a)]、不同荧光样本的组面[图 4(b)–(f)]。 U 矩阵是一种常见的 SOM 网络图形结构的表达方式, 通过计算 SOM 学习后获得的神经元特征向量和邻近神经元的特征向量之间的欧氏距离, 并通过颜色的深浅差异将群聚结构可视化。图形中距离越小表示输入单元之间的差距越小, 表明彼此之间同质性越高; 神经元与神经元之间颜色的深浅表明

距离之远近,颜色越淡表明距离越近,性质越接近^[15, 16]。在图 4(a)中,淡灰色表明神经元的距离和差异很小,属于聚类中心,颜色越深,表明聚类中心与其他神经元之间的距离较远,差距较大,这有助于从淡灰色统一区域中确定聚类边

界。从图 4(a)中可以看出, U 矩阵将荧光光谱分为不同的区域类型,在图 4(a)的右上角,有一条斜的颜色较深的聚类边界区域带,但区域划分较为粗糙,无法获得明确的聚类数。

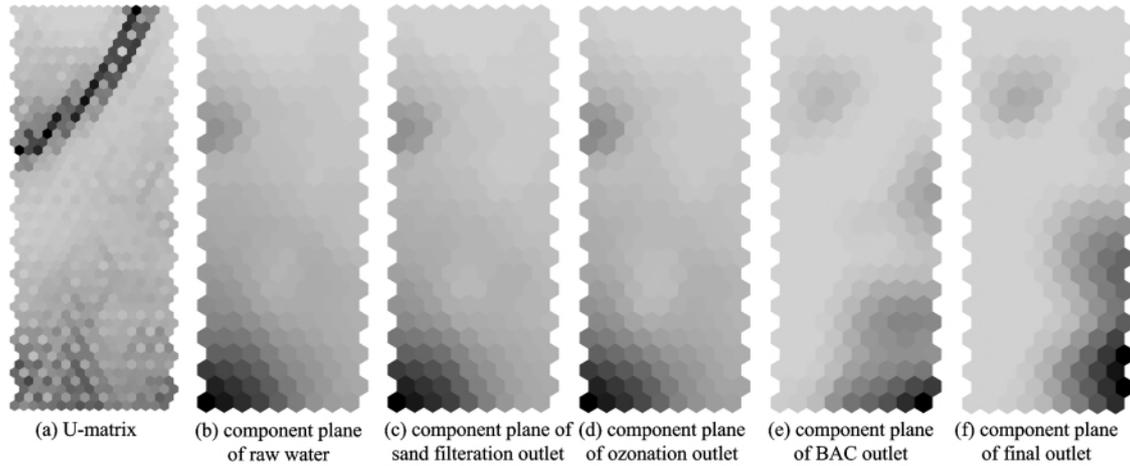


Fig 4 Visualization of the SOM map for fluorescence data of water samples

图 4(b)–(f)为不同水样的组分面,各组分面的坐标轴大小并未统一,描述了不同水样荧光光谱原型向量的相对大小变化。组分面的水平轴和垂直轴分别对应荧光光谱中的发射和激发波长。从图 4(b)和图 4(c)可以发现,原水和砂滤出水的组分面特征基本类似,而经过生物活性炭过滤后[图 4(e)],组分面特征发生较大的变化,原水组分面右下角的峰值已经消失,这部分主要是对应荧光光谱的蛋白峰,说明蛋白类有机物得到了去除,在最终出水[图 4(f)]中紫外富里酸类物质还没有完全去除。

U 矩阵将荧光光谱大致分为不同的区域类型,最终聚类数可以通过 K-means 算法,采用聚类数和平均方差之间的相关性,以及 DBI 指数来确定最佳聚类数^[17, 18]。当不同聚类数之间的平均方差低于 5%时, DBI 指数最低,此时所对应的聚类数目为最佳聚类结果。

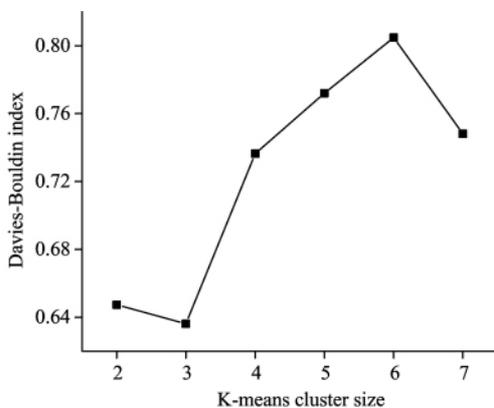


Fig 5 Davies-Bouldin index for different K-means clusters

因此,将训练好的神经元节点的权值输入并进行 K-means 聚类分析,以 DBI 值为指标选取聚类数,并进行分类,结果见图 5 和图 6。从图 5 可以看出,当聚类数为 3 时, DBI

指数最低,因此荧光光谱可以分成 3 类区域,分别为聚类 1、聚类 2、聚类 3,见图 6。每个聚类包含相似的有机物特征,聚类结果和前面单纯寻峰法发现荧光光谱有 3 个峰的结果相对应,表明水体中有效的荧光组分是 3 个。SOM 网络对荧光光谱进行解析是基于激发波长的变化,因此根据不同聚类的位置,结合寻峰法所观察到的 3 个峰,可以认为聚类 1 所对应的是络氨酸类蛋白有机物,聚类 2 所对应的是色氨酸类蛋白有机物,聚类 3 所对应的是紫外富里酸类有机物。

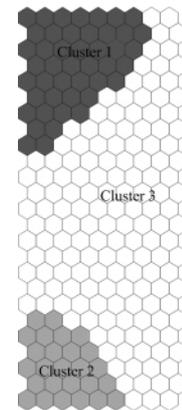


Fig 6 SOM cluster distribution

Rhee^[16]基于 SOM 网络,对不同条件下大肠杆菌发酵过程中的荧光光谱进行了分析,分别提取出五种、七种有效的组分,这些组分在发酵过程中权重浓度的变化和其他发酵过程中的指标(OUR、CPR 等)具有较好的对应关系。此外, Bieroza^[19]针对英国 16 个水厂的原水、出水水样的荧光光谱,分别采用平行因子(PARAFAC)和 SOM 网络进行了解析,提取出紫外腐植酸类、富里酸类、蛋白类有机物组分,并对平行因子和 SOM 网络进行了比较,平行因子和 SOM 网络的主要区别在于前者是有监督的算法,而后者是无监督的过

程。同寻峰法比较,平行因子和 SOM 网络均能更方便、更高效的处理大量复杂荧光数据,获取更多的荧光信息。

2.3 不同类型有机物在水处理中的去除效果

根据 SOM 网络的聚类结果,对每一聚类,其权重可以作为水样中所含有机物的相对浓度,从而得到在水处理过程中,水样所含的三类有机物的权重(即相对浓度)的变化,见图 7。

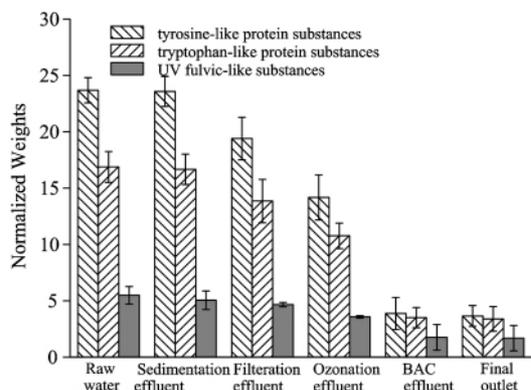


Fig 7 The relative concentration of different components in water samples presented by normalized weights from SOM network

从图 7 可以看出,原水中所含有的有机物主要是络氨酸类、色氨酸类蛋白有机物,而紫外富里酸类有机物的含量很少。随着水处理工艺流程的进行,络氨酸类、色氨酸类蛋白有机物、紫外富里酸类有机物的相对浓度持续下降,最终出水中有机物的权值非常低,络氨酸类、色氨酸、紫外富里酸类有机物总的去除率分别为 84.6%, 79.9% 和 69.1%,从而达到了较好的水处理效果。

3 结 论

采用 SOM 网络对太湖流域某自来水厂水样的三维荧光光谱进行解析,结合 k-means 算法,提取了三种有效的有机物成分,分别为络氨酸类蛋白有机物、色氨酸类蛋白有机物、紫外富里酸类有机物。在整个水处理工艺流程中,络氨酸类、色氨酸、紫外富里酸类有机物总的去除率分别为 84.6%, 79.9% 和 69.1%,出水水质较好。

采用水体三维荧光光谱,结合 SOM 网络,能够获得水处理过程中有机物的光谱特征,有机物的组成和变化规律。研究表明,作为一种有效的光谱解析工具,SOM 网络和三维荧光光谱的结合,有助于对水处理的处理效果进行综合评估,从而提高自来水厂水处理的运行、监测和管理。

References

- [1] WU Jing, CUI Shuo, XIE Chao-bo, et al(吴静,崔硕,谢超波,等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2011, 31(12): 3302.
- [2] Holbrook R D, Breidenich J, Derose P C. Environmental Science & Technology, 2005, 39(17): 6453.
- [3] Murphy K R, Hambly A, Singh S, et al. Environmental Science & Technology, 2011, 45(7): 2909.
- [4] Guimet F, Ferr J, Boqu R, et al. Analytica Chimica Acta, 2004, 515(1): 75.
- [5] Chu W H, Gao N Y, Eeng Y, et al. Environmental Science & Technology, 2010, 44(10): 3908.
- [6] Johnstone D W, Miller C M. Environmental Engineering Science, 2009, 26(7): 1163.
- [7] GUO Wei-dong, HUANG Jian-ping, HONG Hua-sheng, et al(郭卫东,黄建平,洪华生,等). Environmental Science(环境科学), 2010, 31(6): 1419.
- [8] Astel A, Tsakovski S, Barbieri P, et al. Water Research, 2007, 41(19): 4566.
- [9] Lee B H, Scholz M. Water Research, 2006, 40(18): 3367.
- [10] Seo J, Kang J, Seung-Won L, et al. Water Research, 2011, 45(14): 4183.
- [11] Coble P G. Marine Chemistry, 1996, 51(4): 325.
- [12] Bridgeman J, Bieroza M, Baker A. Reviews in Environmental Science and Biotechnology, 2011, 10(3): 277.
- [13] SONG Xiao-na, YU Tao, ZHANG Yuan, et al(宋晓娜,于涛,张远,等). Acta Scientiae Circumstantiae(环境科学学报), 2010, 30(11): 2321.
- [14] Wang Z K, Liu W Q, Zhao N J, et al. Journal of Environmental Sciences, 2007, 19(7): 787.
- [15] Bieroza M, Baker A, Bridgeman J. Advances in Engineering Software, 2011, 44(1): 126.
- [16] Rhee J I, Lee K, Kim C, et al. Biochemical Engineering Journal, 2005, 22(2): 135.
- [17] Bieroza M, Baker A, Bridgeman J. Environmetrics, 2011, 22(3): 256.
- [18] Bieroza M, Baker A, Bridgeman J. Journal of Geophysical Research-Biogeosciences, 2009, 114(7): 1.

The Investigation of Organic Matter Removal in Water Treatment Plant by EEM Spectra Coupled with Self-Organizing Map

DU Er-deng^{1,2}, GUO Ying-qing², SUN Yue², GAO Nai-yun^{1*}, WANG Li-ping²

1. State Key Laboratory of Pollution Control and Resource Reuse, Tongji University, Shanghai 200092, China

2. School of Environmental and Safety Engineering, Changzhou University, Changzhou 213164, China

Abstract Three-dimensional excitation and emission matrix fluorescence spectra (3D-EEM) has attracted the increasing attention of researchers in water monitoring and water treatment areas. The self-organizing map (SOM) is a kind of non-supervised and self-learning neural network with the feature of high self-stability and noise tolerance. In the present paper, SOM technique was employed for the exploratory analysis of EEM spectra of water samples in a water treatment plant. The results showed that EEM spectra could be clustered into three classes, corresponding to tryptophan-like protein substances, tyrosine-like protein substances, and UV fulvic-like substances. The three components could be effectively removed during the whole water treatment process with the high removal of 84.6% (tyrosine-like), 79.9% (tryptophan-like), and 69.1% (UV fulvic-like). The results show that SOM technique can be used as an effective tool for EEM spectra analysis, which is helpful for the optimization of water treatment process parameters, the improvement of process performance, and the operation of water treatment plant.

Keywords Drinking water treatment; Three-Dimensional excitation and emission matrix fluorescence (3D-EEM); Self-organizing feature map (SOM); Organic matter removal

* Corresponding author

(Received Feb. 12, 2012; accepted Apr. 10, 2012)

庆祝《分析实验室》创刊 30 周年 暨分析测试技术学术交流会

为庆祝《分析实验室》创刊 30 周年, 回报长期以来关心、支持《分析实验室》杂志的广大分析测试工作者, 发挥《分析实验室》期刊作为专业媒体的作用, 拓宽和强化学术交流平台, 促进分析测试工作的交流与发展, 北京有色金属研究总院、中国分析测试协会联合《分析实验室》期刊, 定于 2012 年联合举办“分析测试技术学术交流会”。

会议时间: 2012 年 10 月

会议地点: 北京

会议将邀请专家针对分析测试技术在矿物、新型材料、环境保护、食品与药物、生物分析等领域的应用及最新进展作大会专题报告, 并以“《分析实验室》创刊 30 周年纪念专辑”的形式隆重出版大会论文集。现开始向全国各行业分析测试专业人士征稿, 并欢迎各界人士前往参加会议进行交流。

征文范围: 1. 矿物与金属材料分析测试研究论文; 2. 环境保护分析测试研究论文; 3. 有机物与药物、生化分析研究论文; 4. 关于各种仪器分析技术应用的最新进展的综合评述及研究论文等。

征文要求: 征文限 5000 字以内(包括文字、图表、中英文摘要及参考文献)。征文经会议学术委员会评审, 录用后即寄发录用通知, 凡前往参加会议交流并交纳版面费的论文收入《分析实验室》2012 年第 31 卷第 10 期(庆祝创刊 30 周年纪念专辑)或增刊正式出版。

征文截止日期: 2012 年 7 月 30 日

论文及参会联系: 100088 北京新外大街 2 号《分析实验室》编辑部会议筹备组;

电话/传真: 010-82013328; 联系人: 孙臣良; E-mail: ana-info@263.net

中国分析测试协会 北京有色金属研究总院
《分析实验室》期刊编辑部