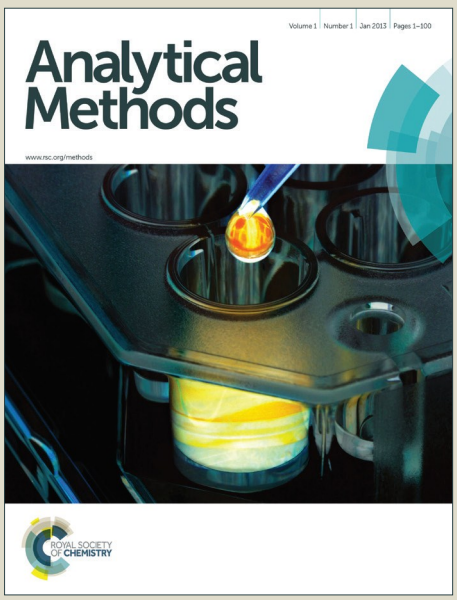
Analytical Methods

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Characteristics Region Extraction of Time Series Three-dimensional Fluorescence Spectroscopy

Shaohui Yu[1*], Xue Xiao[2], Nanjing Zhao[2], Jisheng Yang[1]

[1] *School of Mathematics and Statistics, Hefei Normal University, Hefei 230061, China*

[2] *Key Laboratory of Environmental Optics & Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei 230031, China*

Abstract: Three-dimensional fluorescence spectroscopy of online monitoring, as a time series three-dimensional fluorescence spectroscopy, contains rich information including not only the usual organic matter as monitored by spectroscopy but also the unusual organic matters as the outliers. In order to analyze the organic matter and the outliers, we proposed a method to extract the characteristics regions of time series three-dimensional fluorescence spectroscopy, which is achieved along the time mode and the spectral mode. From the time mode, some time series change obviously because of the presence of outliers. So the characteristics region of three-dimensional fluorescence spectroscopy, caused by unexpected events, can be characterized by the outliers region. From the spectral mode, there are significant differences between the fluorescence intensities of the noise and the organic matter, and therefore the characteristics region of organic matter can be extracted by clustering analysis. Time series three-dimensional fluorescence spectroscopy of tryptophan in water samples with catechol as the outlier is tested. Numerical experiments show that the proposed method can effectively extract the organic matter fluorescence region of tryptophan and the outlier fluorescence region of catechol.

Keywords: time series; characteristics region; cluster analysis; three-dimensional

* E-mail:yushaohui2005@163.com

fluorescence spectroscopy

1. Introduction

At present, the water pollution is serious and water pollution incidents occasionally occur. So it is urgent to develop fast and accurate methods to timely detect and analyze the water pollution. Three-dimensional fluorescence spectroscopy, as a fingerprint with the excitation and emission spectra, has been widely used in recent years. Because it has played an important role for quick access to the information on water pollution [1], three-dimensional fluorescence spectroscopy of online monitoring has attracted more and more scholars' attentions. Carstea [2] ever analyzed the monitoring data of two consecutive weeks by the self organizing maps and humic acids are analyzed from several obvious fluorescence peaks. Stedmon[3] discussed the organic matter features of the drinking water by PARAFAC method. Murphy [4] investigated the EEM of recycled water treatment plants by PARAFAC method. And Singh[5] analyzed the different fluorescence peaks of reverse osmosis that permeates from water recycling plants. Hao[6] monitored two major disinfection by-product precursors of reclaimed water by the fluorescence peaks and regression models. Peiris[7] assessed the membrane constituents in drinking water treatment system by principal component analysis and three-dimensional fluorescence spectroscopy. Bridgeman [8] and Elinatan [9] considered the water quality at different processing stages of wastewater treatment plant by three-dimensional fluorescence spectroscopy. It is not difficult to find that the research of online monitoring fluorescence spectroscopy focuses primarily on several specific fluorescence regions of dissolved organic matter[10]. The specific fluorescence regions are also important to the computation and assessment of some algorithms[11, 12]. However, all wavelength points are usually taken into in some qualitative and quantitative algorithms, which will increase the amount of computation and computational complexity. Furthermore, the involvement of unrelated fluorescence region will decrease the accuracy of the analysis results. So it has great significance to effectively extract the characteristics region of online monitoring three-dimensional fluorescence spectroscopy before the qualitative and quantitative analysis.

As a third-order tensor, three-dimensional fluorescence spectroscopy at each sampling

time is a second-order tensor and the third order tensor is constituted by the sampling times. As the computational complexity is very high and the calculation is large for the qualitative and quantitative from the tensor, it can be viewed as a special time series for the online monitoring three-dimensional fluorescence spectroscopy along the time mode. In order to compress data, reduce the dimensions and computational complexity before statistical analysis and modeling, the outliers along the time mode and the characteristics regions along the spectral mode can be detected and extracted by the advantage of the time series and the characteristics of three-dimensional fluorescence spectroscopy.

2. Theory

2.1 Outliers extraction along the time mode

Time series three-dimensional fluorescence spectroscopy is generated from three-dimensional fluorescence spectroscopy collected in accordance with different sampling time. The index t denotes the sampling time and the time series three-dimensional fluorescence spectroscopy is denoted by $X_1, X_2, \cdots, X_t, \cdots$, where $X_t \in R^{J\times K}, t = 1,2,\cdots T$. J and K respectively represent the emission wavelength number and the excitation wavelength number. All X_t array along the time axis and a three-way data matrix $\underline{\mathbf{X}} \in R^{T\times J\times K}$ is formed, where T denotes the sampling time number. In view of the continuous variation of organic matter concentration in water, the two adjacent samplings X_t, X_{t+1} usually have certain similarity. However, unexpected events (outliers) may result in obvious difference in the fluorescence region of three-dimensional fluorescence spectroscopy at specific sampling time.

Extract $J\times K$ time series from $\underline{\mathbf{X}}$

$$\mathbf{x_i} = \left(x_{1jk}, x_{2jk}, \cdots . x_{Tjk}\right), j = 1,2,\cdots, J, k = 1,2,\cdots, K, i = 1,2,\cdots,(J\times K). \quad (1)$$

In order to eliminate the impact of the noise and improve the computation accuracy, each time series $\mathbf{x_i}$ is firstly denoised by the wavelet transform method.

Then, compute the variance of $\mathbf{x_i}$

$$v_i = \frac{\sum_{t=1}^{T}\left(x_{tjk} - \overline{\mathbf{x_i}}\right)^2}{T},\ i = 1,2,\cdots,(J \times K),\ j = 1,2,\cdots,J,\ k = 1,2,\cdots,K \quad (2)$$

Where $\overline{\mathbf{x_i}}$ denotes the average value of the time series $\mathbf{x_i}$. The value of v_i implies the deviation from the average fluorescence intensity. Usually, the larger the fluorescence intensity is, the bigger the value of v_i has. This is mainly because the index v_i is an absolute index. As the outliers may occur at any wavelengths including the fluorescence region of the noise, it is not sufficient to detect the outliers only by the value of v_i. So the outlier index m which is a relative index is introduced and the outlier matrix M is formed, which can reduce the influence of fluorescence intensity.

$$m_i = \frac{C_{\mathrm{x_i}}{}^2}{v_i} \quad (3)$$

Here $C_{\mathrm{x_i}} = \max\left(\left|x_{1jk} - \overline{\mathbf{x_i}}\right|, \left|x_{2jk} - \overline{\mathbf{x_i}}\right|, \cdots, \left|x_{Tjk} - \overline{\mathbf{x_i}}\right|\right)$ is the absolute deviation and $\overline{\mathbf{x_i}}$ denotes the average value of $\mathbf{x_i}$. $\mathbf{x_i}$ denotes the $i-th$ time series. All m_i array in the matrix $M \in R^{J \times K}$ and the value m_i in matrix M reflects the relative possibility for every wavelength, which may be outliers. The greater the outlier value m_i is, the more likely the wavelength that contains the unexpected events is. The outlier region can be, ultimately, extracted according to the threshold of the outlier values m_i. Compared with Grubbs test, the threshold of the outliers value m_i here can be set to 16 as the G value of the Grubbs test is 3.6055 with the detection level $\alpha = 0.05$ and the number of test points $n = 200$.

2.2 Characteristics region extraction along the spectral mode

According to the variation difference of fluorescence intensity between the noise and the organic matter, all time series $\mathbf{x_i}$（$i = 1,2,\cdots,(J \times K)$）are clustered by clustering analysis and the characteristics region along the spectral mode is extracted.

Firstly, each time series $\mathbf{x_i}$ is normalized

$$\left(\mathbf{x_i} - \overline{\mathbf{x_i}}\right) \Big/ \sqrt{v_i} \qquad (4)$$

Where v_i is the same as the variance in formula (2).

Then, compute the Euclidean distance between all time series $\mathbf{x_i}$

$$d_{rs}^2 = \left(\mathbf{x_r} - \mathbf{x_s}\right)\left(\mathbf{x_r} - \mathbf{x_s}\right)', \quad r,s = 1,2,\cdots,(J \times K) \qquad (5)$$

Where $\left(\mathbf{x_r} - \mathbf{x_s}\right)'$ is the transpose of $\left(\mathbf{x_r} - \mathbf{x_s}\right)$. Finally, clustering analysis of all time series $\mathbf{x_i}$ is performed according to the minimum distance and the cluster matrix D along the spectral mode is formed.

The characteristics region of time series three-dimensional fluorescence spectroscopy has rich information. It includes not only the outliers regions caused by unexpected organic matter but also the regions with high fluorescence intensity of monitored organic matter, which is different from the noise region. So, the characteristics region of online monitoring three-dimensional fluorescence spectroscopy can be extracted by the outliers' matrix M along the time mode and the cluster matrix D along the spectral mode.

3. Experiment results and discussions

In the experiments, tryptophan is tested as the monitored organic matter in water samples and catechol is considered as the outlier organic matter. Tryptophan solutions with different concentration (0.08mg/L, 0.16mg/L, 0.24mg/L, etc.)and catechol solutions（0.8mg/L,1.2mg/L）are prepared in the laboratory. Each solution is scanned by Hitachi-7000

fluorescence spectrophotometer with the excitation wavelength 230nm -320nm and the emission wavelength 250nm-500nm. The scan interval for the excitation wavelength and the emission wavelength is respectively 4nm and 2nm. The Rayleigh scattering and Raman scattering are eliminated by the interpolation method (Fig.1) [13-14].

In order to demonstrate the validity of the method, we test two groups. One group includes the monitored tryptophan with the higher concentration and the outlier catechol with the lower concentration, and the other is vise versa. In the first test group, average three-dimensional fluorescence spectroscopy of tryptophan with two different concentrations (0.16mg/L, 0.24mg/L) is considered as the generating basis of time series three-dimensional fluorescence spectroscopy and three-dimensional fluorescence spectroscopy of catechol with concentration 0.8 mg/L is the generating basis of the outlier interferences for time series three-dimensional fluorescence spectroscopy of tryptophan. Different from the first test group, three-dimensional fluorescence spectroscopy of tryptophan with concentration 0.16mg/L and that of catechol with concentration 1.2mg/L are respectively considered as the generating basis of monitored tryptophan and the outlier catechol in the second test group.

In order to accurately simulate the real concentration variation of tryptophan with time in water samples, we choose the polynomial curves as the variation curves of the tryptophan concentration. 200 continuous samples acquired from the polynomial curve are simulated as the concentrations of tryptophan in water samples. According to the Beer-Lambert law, time series three-dimensional fluorescence spectroscopy is simulated by the three-dimensional fluorescence spectroscopy of tryptophan obtained from laboratory (the generating basis) and the simulated concentrations obtained from the polynomial curves. And this time series

three-dimensional fluorescence spectroscopy is tested as the time series of monitored tryptophan in water samples. Further, this time series three-dimensional fluorescence spectroscopy is disturbed by the three-dimensional fluorescence spectroscopy of outlier catechol at specific points in time, such as T=101,102,103, which results in the outliers of time series three-dimensional fluorescence spectroscopy of tryptophan. Thus a time series three-dimensional fluorescence spectroscopy $X_1, X_2, \cdots, X_t, \cdots (X_t \in R^{J\times K}, t = 1, 2, \cdots 200$, $J = 126$, $K = 23$) is formed.

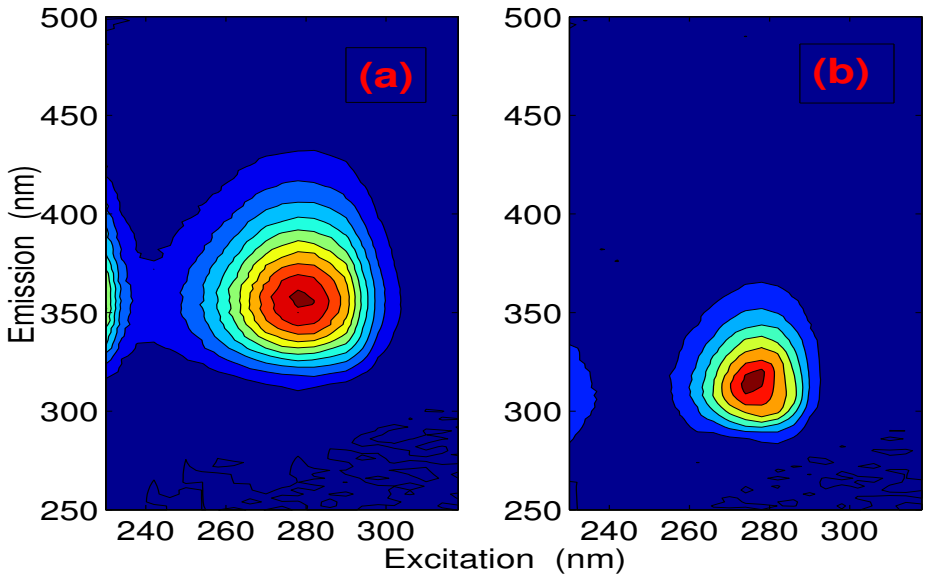


Fig.1. Three dimensional fluorescence spectroscopy: (a) tryptophan; (b) catechol

Several time series $\mathbf{x_i}$ of specific wavelength around the fluorescence peak (excitation: 282nm, emission: 352nm) are shown in Fig.2. Obviously, the fluorescence intensity of tryptophan varies continuously which is well fitted by the polynomial curves. In order to eliminate the impact of the noise, each time series $\mathbf{x_i}$ is denoised by the wavelet transform method.

The outlier value m_i of all time series $\mathbf{x_i}$ are computed by the formula (3) and the

outlier matrix M generates (Fig.3). As can be seen from Fig.3, the region with excitation wavelength 265-285nm and emission wavelength 290-310nm has larger outlier values than the other region. By comparison with Fig.1, it is not difficult to find that this region is just the outlier values of catechol appeared in the $101-th$, $102-th$ and $103-th$ samples of time series $X_1, X_2, \cdots, X_t, \cdots$. Most importantly, the outliers region is more obvious if the concentrations of the outlier catechol are higher compared with the concentrations of monitored tryptophan (Fig.3 (b)). Here, if the threshold of the outliers' value m_i is set to 16, the outliers region can be effectively extracted by the formula (3) (Fig.3).

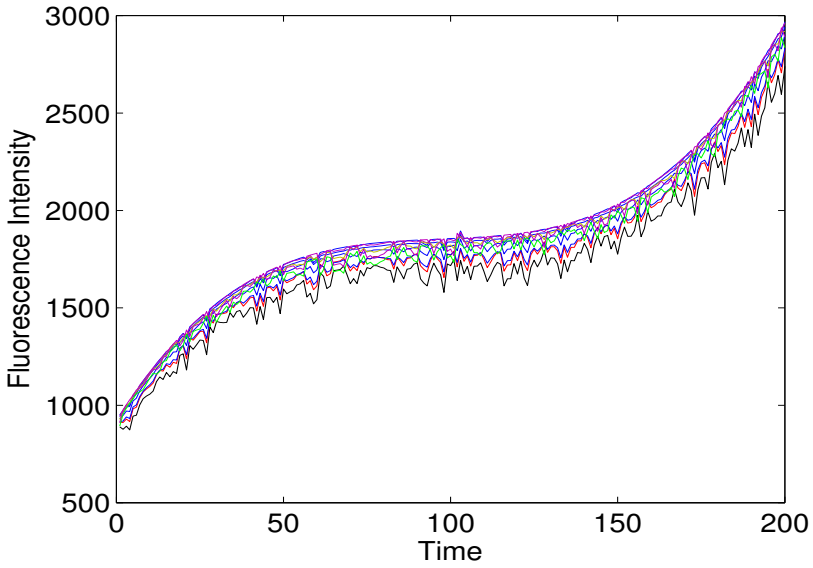


Fig.2. Time series x_i of specific wavelength around the fluorescence peak (excitation: 282nm, emission: 352nm)

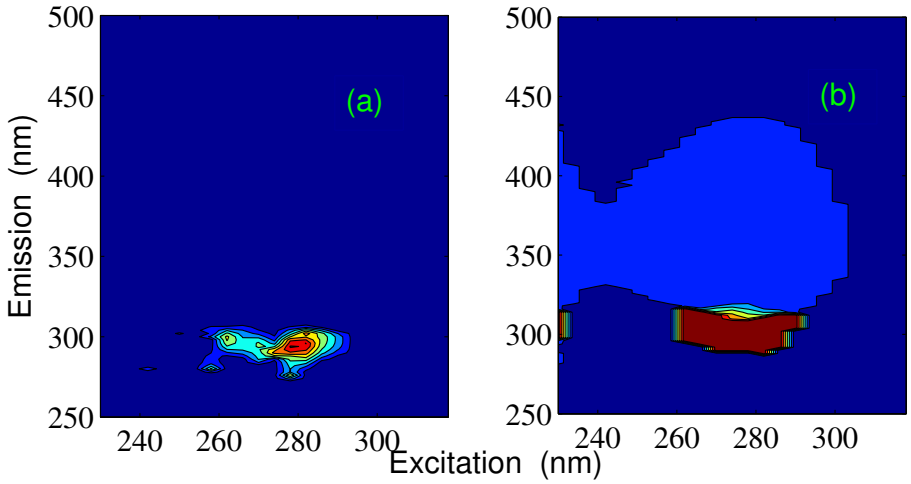


Fig.3. The outliers region of time series three-dimensional fluorescence spectroscopy of monitored tryptophan with the outlier catechol in the 101-th, 102-th and 103-th samples: (a) the first test group;(b)the second test group.

In addition, the fluorescence values of the noises compared with that of the organic matter are usually small and there are no obvious changes along the time mode (Fig.4 (b)). Different from the fluorescence regions of the noise, the fluorescence regions of organic matter generally have obvious variation trends (Fig.4 (a)). By comparing the Fig.4 with Fig.2, we can also find that there is the similar tendency of fluorescence intensity for the same kind of organic matter at different wavelengths.

So, the characteristics regions of organic matter can be extracted by the cluster analysis. Each $X_t(t=1,2,\cdots 200)$ includes 2898 wavelength points (23 for excitation wavelength and 126 for emission wavelength). Every wavelength along the time mode corresponds to a time series $\mathbf{x_i}\ (i=1,2,\cdots 2898)$ with 200 sample points. So there are 2898 time series for time series three-dimensional fluorescence spectroscopy $X_1, X_2, \cdots, X_t, \cdots$ ($X_t \in R^{J\times K}$, $t=1,2,\cdots 200$, $J=126$, $K=23$). Obviously, there are differences between 2898 time

series for the noise and the organic matters (tryptophan and catechol). So the characteristics regions of organic matter are extracted successfully by the cluster analysis of all time series $\mathbf{x_i}$ ($i = 1, \cdots, 2898$).

The cluster results of all time series $\mathbf{x_i}$ are shown in Fig.5 according to the formula (4) and (5) (Here, we choose 6 as the number of the clusters.). It can be seen from Fig.5 the characteristics regions of tryptophan (class 1-3, 5-6) and the noise (class 4) belong to different clusters.

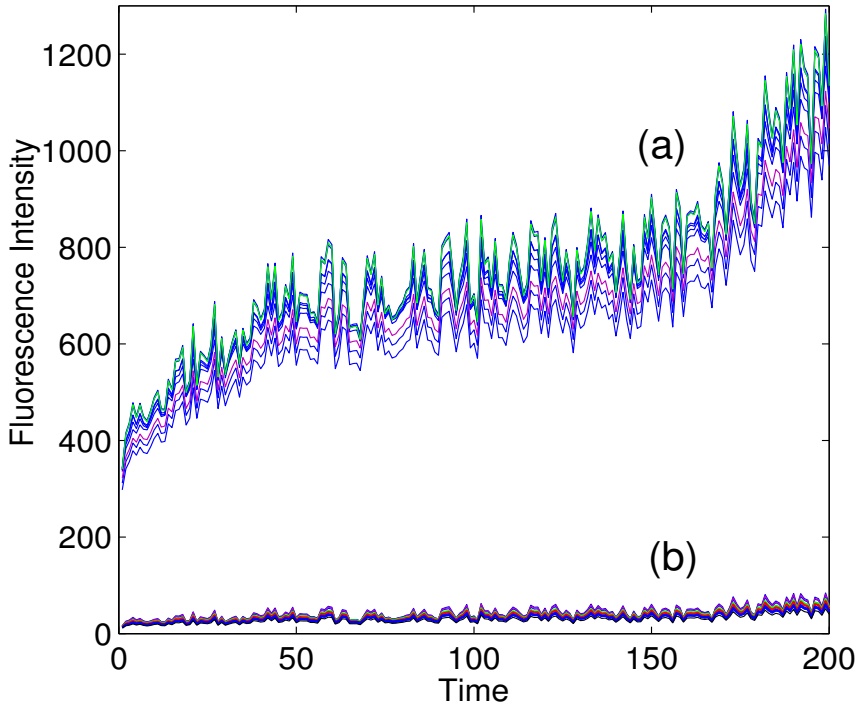


Fig.4. Comparisons of fluorescence spectroscopy between organic matter and noise along the time mode: (a) organic matter fluorescence (b) noise fluorescence

In a conclusion, the results shown in Fig.3 and Fig.5 confirm the effective of the method proposed in this paper. This method can successfully extract the characteristics region of the monitored tryptophan, which contains the outlier catechol at $T = 101,102,103$.

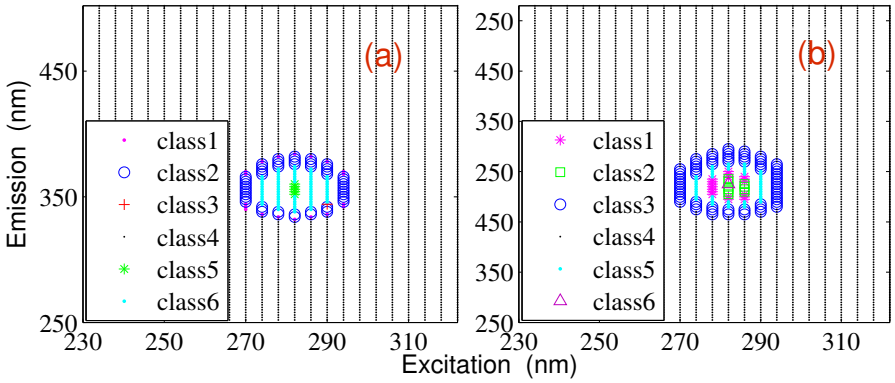


Fig.5. The characteristic region of time series three-dimensional fluorescence spectroscopy by the cluster analysis of all time series $\mathbf{x}_i (i = 1, \cdots, 2898)$:(a) the first test group ;(b) the second test group.

4. Conclusion

Data analysis of online monitoring three-dimensional fluorescence spectroscopy has been paid more and more attentions. Though the characteristics of high dimension and rich information can improve the accuracy of the analysis results, it also increases the computation and complexity. The extraction of characteristics regions may largely reduce the computation of the subsequent qualitative and quantitative analysis. In this paper, the characteristics region extraction of online monitoring three-dimensional fluorescence spectroscopy is achieved by the variance analysis, wavelet transform and cluster analysis. The extracted characteristics regions contain not only the outliers region of catechol (Fig.3), but also the fluorescence region of organic matter (tryptophan) (Fig.5).

Acknowledgements

The work was supported by Anhui Provincial Natural Science Foundation (1308085QF111), National Natural Science Foundation of China (61308063, 61378041), Projects of International Cooperation and Exchanges NSFC (61491240110) and Cultivation

Project for Outstanding Undergraduates Thesis of Hefei Normal University (2014lwpy12).

References

[1] B. Nuray. Water pollution. Croatia: Intechweb.org, 2012.

[2] E.M. Carstea, A. Baker, M. Bieroza, D.. Reynolds, Continuous fluorescence excitation-emission matrix monitoring of river organic matter, Water. Res. 18(2010) 5356-5366.

[3] C.A. Stedmon, B. Seredynska-Sobecka, R. Boe-Hansen, N. Le Tallec, C.K. Waul, E. Arvin, A potential approach for monitoring drinking water quality from groundwater systems using organic matter fluorescence as an early warning for contamination events，Water. Res. 18(2011)6030-6038.

[4] K.R. Murphy, A. Hambly, S. Singh, R.K. Henderson, A. Baker, R. Stuetz, S.J. Khan, Organic matter fluorescence in municipal water recycling schemes: toward a unified PARAFAC model，Environ. Sci. Technol. 7(2011)2909-2916.

[5] S. Singh, R.K. Henderson, A. Baker, R.M. Stuetz, S.J. Khan, Characterization of reverse osmosis permeates from municipal recycled water systems using fluorescence spectroscopy: Implications for integrity monitoring ，J. Membr. Sci. 421(2012)180-189.

[6] R.X. Hao, H.Q. Ren, J.B. Li, Z.Z. Ma, H.W. Wan, X.Y. Zheng, S.Y. Cheng, Use of three-dimensional excitation and emission matrix fluorescence spectroscopy for predicting the disinfection by-product formation potential of reclaimed water, Water.Res. 17(2012)5765-5776.

[7] R.H. Peirisa, M. Jaklewiczb, H. Budmana, R.L. Leggea, C. Moresoli，Assessing the role of feed water constituents in irreversible membrane fouling of pilot-scale ultrafiltration

drinking water treatment system，Water. Res. 10(2013)3364-3374.

[8] J. Bridgemana, A. Bakerb, C. Carliell-Marqueta , E. Carsteac, Determination of changes in wastewater quality through a treatment works using fluorescence spectroscopy, Environ. Technol. 23(2013)3069-3077.

[9] E. Cohen, G. J. Levy, M. Borisover, Fluorescent components of organic matter in wastewater: efficacy and selectivity of the water treatment, Water. Res. 55(2014)323-334.

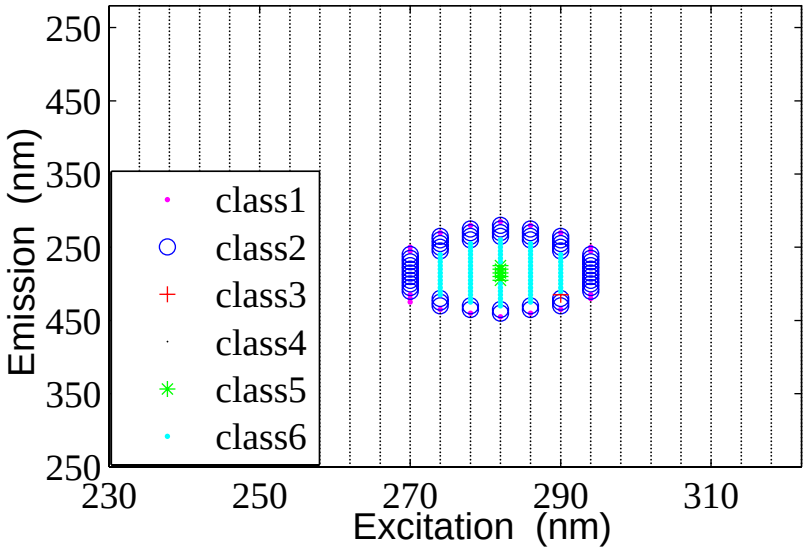
[10] P.G. Coble, Characterization of marine and terrestrial DOM in seawater using excitation emission matrix spectroscopy, Mar. Chem. 4(1996)325-346.

[11] W. Chen, P. Westerhoff, J.A. Leenheer et al. Fluorescence excitation - Emission matrix regional integration to quantify spectra for dissolved organic matter, Environ. Sci. Technol. 24(2003)5701-5710.

[12] J.A. Korak, A.D. Dotson, R. S. Summers, F.L. Rosario-Ortiz.. Critical analysis of commonly used fluorescence metrics to characterize dissolved organic matter, Water research, 49(2014)327-338.

[13] R.G. Zepp, W.M. Sheldon, M.A. Moran, Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Rayleigh and Raman scattering peaks in excitation-emission matrices. Mar. Chem. 89(2004)15-36.

[14] M. Bahram, R. Bro, C. Stedmon, A. Afkhami, Handing of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation, J. Chemom. 30(2006)99-105.



We propose a method to extract the characteristics region of time series three-dimensional fluorescence spectroscopy including the monitored organic matter and the outlier organic matter.