Supplemental Material

Understanding organic non-point source pollution in watersheds via pollutant indicators, disinfection byproduct precursor predictors, and composition of dissolved organic matter

Yixiang Zhang a, Xinqiang Liang a,b,*

a Department of Environmental Engineering, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China

b Zhejiang Provincial Key Laboratory of Water Pollution Control and Environmental Safety, Zhejiang University, Hangzhou 310058, China

* Corresponding author. Tel.: +86-571-88982018; E-mail address: liang410@zju.edu.cn.

Number of pages: 22
Number of Figures: 8
Number of Tables: 1
Table of Contents

Section S1
Description and comparison of peak-peaking, FRI and PARAFAC.

Section S2
More details about SVM.

Fig. S1
Schematic diagrams of SVC and SVR predictions of pollution state and DBP FPs.

Fig. S2
The fluorescence classification method of peak-picking (Coble, 1996).

Fig. S3
The fluorescence classification method of peak-picking (Coble, 2007).

Fig. S4
The fluorescence classification method of FRI (Chen et al., 2003).

Fig. S5
The PARAFAC components in other literatures.

Fig. S6
Parameter optimization of SVR for HAA FP, HAN FP and CH FP using a two-step grid search with cross-validation.

Fig. S7
Parameter optimization of SVR for HAA FP, HAN FP and CH FP using genetic algorithm and particle swarm optimization.

Fig. S8
Graphical abstract of pollution indication and DBP FP prediction by means of SVM.

Table S1
MSE and corresponding parameter $C$ and $\gamma$ derived from the parameter optimization processes for HAA FP, HAN FP and CH FP through grid search, GA and PSO on the basis of three kinds of input independent variables.
Section S1. Description and comparison of peak-peaking, FRI and PARAFAC

The first introduced method is peak-picking proposed by (Coble, 2007), through which she tried to explain marine optical biogeochemistry. The method is quite simple just to pick the fluorescence intensities from the peaks of the EEM landscape. However, since the DOM is source dependent, the location of specific organic component will shift with different sampling sources. The second introduced method is FRI approach proposed by (Chen et al., 2003), through which they tried to quantify spectra for DOM. As an empirical method, depending on its merits of simpleness and comprehensibility, FRI solved DOM quantification problems in many environmental researches. However, integration technique could not improve intensity/interference (signal/noise, S/N) because fluorescence at the curve peak would be the ideal signal rather than at the curve slope. The third introduced method is PARAFAC analysis which is the dominant DOM composition extraction method currently. Despite complex algorithm, PARAFAC solves the problem of FRI that the location of DOM composition must be defined and the problem of peak-picking that the peaks with low intensity is difficult to identify. The fundamental principle of PARAFAC is an alternating least-squares algorithm decomposing data into a set of trilinear terms and a residual array, and the detailed description can be found in many studies (Murphy et al., 2013).
Section S2. More details about SVM

Before the SVM computing processes, besides label and property groups (SVC) or dependent and independent variables (SVR), several parameters should be input to define the function and properties of SVM including SVM type, kernel function type, specific parameters for each kernel function and specific parameters for each SVM type. In this study, we chose C-SVC for SVC and epsilon-SVR for SVR, and RBF as kernel function. Among specific parameters for kernel functions and SVM types, $C$ and $\gamma$ with significant leverage on training results should be emphasized. $C$ is a specific parameter for both C-SVC and epsilon-SVR. The $C$ is a penalty parameter that controls the trade-off between maximizing the geometric margin and minimizing the training error. Insufficient fitting will be placed on training data for a low value of $C$ while the algorithm will overfit for too large value of $C$. Therefore, $C$ should be properly chosen to train a robust model. $\gamma$ is a specific parameter for RBF kernel function. The $\gamma$ is a kernel parameter that controls the amplitude of the kernel. Similar to $C$, improper choice of $\gamma$ will lead to overfitting and underfitting in prediction. Therefore, both $C$ and $\gamma$ were carefully chosen via trial and error approach as well as heuristic algorithms.
Fig. S1 Schematic diagrams of SVC and SVR predictions of pollution state and DBP FPs. The figure was created using GVEdit, graphviz, BiocInstaller, Rgraphviz.
Fig. S2 The fluorescence classification method of peak-picking (Coble, 1996). The figure was created using Canopy Command Prompt, Enthought Canopy, VisPy, Hornil StylePix.
Fig. S3 The fluorescence classification method of peak-picking (Coble, 2007). The figure was created using xarray, HoloViews, Wing 101, Sketsa SVG Editor 7.1.1.
Fig. S4 The fluorescence classification method of FRI (Chen et al., 2003). The figure was created using Python 3.6.2, Anaconda Navigator 1.4.3, Jupyter QtConsole 4.2.1, ggplot, pymysql, DBeaver Version 4.3.5, GIMP 2.8.22.
Fig. S5 The PARAFAC components in other literatures.

a) The PARAFAC components (Nie et al., 2016).
b) The PARAFAC components in the Global model (Yu et al., 2015).
c) The modelled PARAFAC components (Heinz et al., 2015).
d) The PARAFAC components (Schartup et al., 2015).
e) The components of the PARAFAC model (Gabor et al., 2015).
f) The PARAFAC components (Sanchez et al., 2014).
g) The PARAFAC components (Li et al., 2014).
h) The PARAFAC components (Bianchi et al., 2014).
i) The PARAFAC components (Hosen et al., 2014).
j) The PARAFAC components (Chen et al., 2014).
k) The PARAFAC components (Stubbins et al., 2014).
l) The PARAFAC components of Elliot HA (Orsetti et al., 2013).
m) The PARAFAC components of Fluka HA (Orsetti et al., 2013).
n) The PARAFAC components (Yu et al., 2012).
o) The PARAFAC components (Osburn et al., 2012).
p) The PARAFAC components (Beggs and Summers, 2011).
q) The PARAFAC components (Liu et al., 2018).
r) The PARAFAC components (Bai et al., 2017).
s) The PARAFAC components (Zhang et al., 2017).
t) The PARAFAC components (Zeng et al., 2017).
u) The PARAFAC components (Cui et al., 2016).
**Fig. S6a** Parameter optimization of SVR for HAA FP using a two-step grid search with cross-validation.
**Fig. S6b** Parameter optimization of SVR for HAN FP using a two-step grid search with cross-validation.
**Fig. S6c** Parameter optimization of SVR for CH FP using a two-step grid search with cross-validation.
Fig. S7a Parameter optimization of SVR for HAA FP using genetic algorithm and particle swarm optimization.
Fig. S7b Parameter optimization of SVR for HAN FP using genetic algorithm and particle swarm optimization.
Fig. S7c Parameter optimization of SVR for CH FP using genetic algorithm and particle swarm optimization.
Fig. S8 Graphical abstract of pollution indication and DBP FP prediction by means of SVM.
Table S1

MSE and corresponding parameter $C$ and $\gamma$ derived from the parameter optimization processes for HAA FP, HAN FP and CH FP through grid search, GA and PSO on the basis of three kinds of input independent variables. The combination in bold is the adopted parameter optimization method and input independent variables.

<table>
<thead>
<tr>
<th></th>
<th>HAA FP</th>
<th>Grid search</th>
<th>GA</th>
<th>PSO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C$</td>
<td>1</td>
<td>0.87919</td>
<td>2.5623</td>
</tr>
<tr>
<td>PARAFAC</td>
<td>$\gamma$</td>
<td>16</td>
<td>15.5993</td>
<td>7.9208</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.13279</td>
<td>0.13237</td>
<td>0.13607</td>
</tr>
<tr>
<td></td>
<td>$C$</td>
<td>0.25</td>
<td>1.0912</td>
<td>0.1</td>
</tr>
<tr>
<td>FRI</td>
<td>$\gamma$</td>
<td>2.8284</td>
<td>16.4099</td>
<td>3.7634</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.15126</td>
<td>0.15744</td>
<td>0.15971</td>
</tr>
<tr>
<td></td>
<td>$C$</td>
<td>0.70711</td>
<td>0.75607</td>
<td>0.84546</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.14125</td>
<td>0.14077</td>
<td>0.14527</td>
</tr>
<tr>
<td>Peak-picking</td>
<td>$\gamma$</td>
<td>11.3137</td>
<td>11.6091</td>
<td>7.9208</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.14125</td>
<td>0.14077</td>
<td>0.14527</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>HAN FP</th>
<th>Grid search</th>
<th>GA</th>
<th>PSO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C$</td>
<td>0.35355</td>
<td>2.4124</td>
<td>28.3147</td>
</tr>
<tr>
<td>PARAFAC</td>
<td>$\gamma$</td>
<td>256</td>
<td>322.7723</td>
<td>284.4743</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.00032015</td>
<td>0.00031842</td>
<td>0.00032048</td>
</tr>
<tr>
<td></td>
<td>$C$</td>
<td>0.044194</td>
<td>0.23584</td>
<td>0.1</td>
</tr>
<tr>
<td>FRI</td>
<td>$\gamma$</td>
<td>256</td>
<td>312.6786</td>
<td>312.5197</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.00033817</td>
<td>0.00033838</td>
<td>0.00033838</td>
</tr>
<tr>
<td></td>
<td>$C$</td>
<td>0.044194</td>
<td>26.1429</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0003276</td>
<td>0.00032579</td>
<td>0.00032594</td>
</tr>
<tr>
<td>Peak-picking</td>
<td>$\gamma$</td>
<td>256</td>
<td>321.9608</td>
<td>348.8044</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0003276</td>
<td>0.00032579</td>
<td>0.00032594</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CH FP</th>
<th>Grid search</th>
<th>GA</th>
<th>PSO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C$</td>
<td>0.0625</td>
<td>6.3532</td>
<td>0.1</td>
</tr>
<tr>
<td>PARAFAC</td>
<td>$\gamma$</td>
<td>256</td>
<td>814.8745</td>
<td>217.6479</td>
</tr>
</tbody>
</table>

17
<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>C</th>
<th>γ</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRI</td>
<td>0.00041364</td>
<td>0.0055243</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>0.00042722</td>
<td>77.5693</td>
<td>62.3147</td>
</tr>
<tr>
<td></td>
<td>0.00041311</td>
<td>380.2017</td>
<td>284.4743</td>
</tr>
<tr>
<td>Peak-picking</td>
<td>0.00045456</td>
<td>0.0625</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>0.00043697</td>
<td>20.2316</td>
<td>291.2019</td>
</tr>
<tr>
<td></td>
<td>0.00043904</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.00041723</td>
<td>0.00041693</td>
<td>0.00041704</td>
</tr>
</tbody>
</table>
References


