Support information

1 The detail information of sample collection

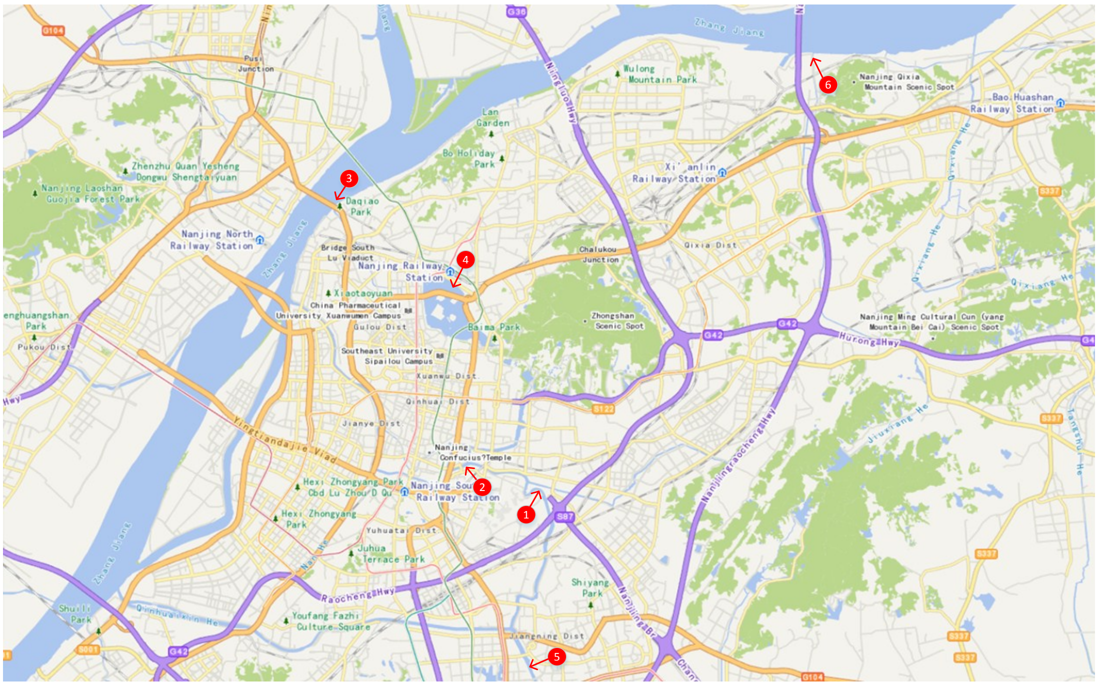
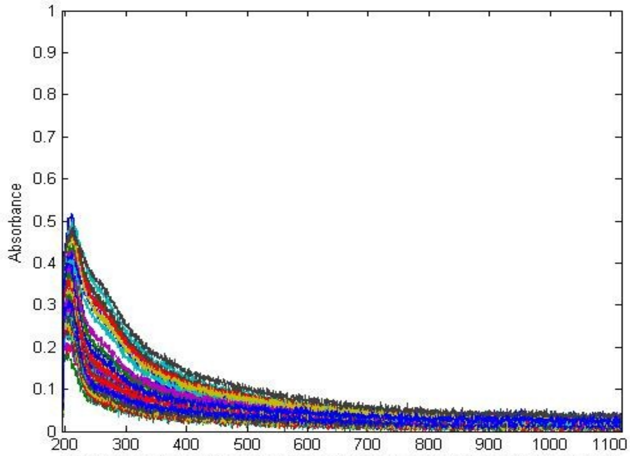


Figure 1 Sampling sites in Nanjing

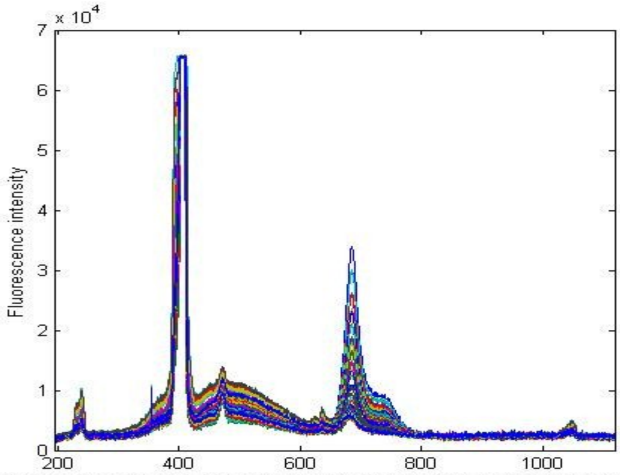
Table 1 The detail information of collected water samples

Sequence Number	Address	Range of COD (mg/L)	Number of samples
1	Seven Bridge Wen	0~6	56
2	Jiezhizha	0~5	49
3	Nanjingchangjiang Bridge	0~5	42
4	Xuanwu Lake	0~8	58
5	Yangqiao	0~8	71
6	Jiuxiang estuary	0~5	47

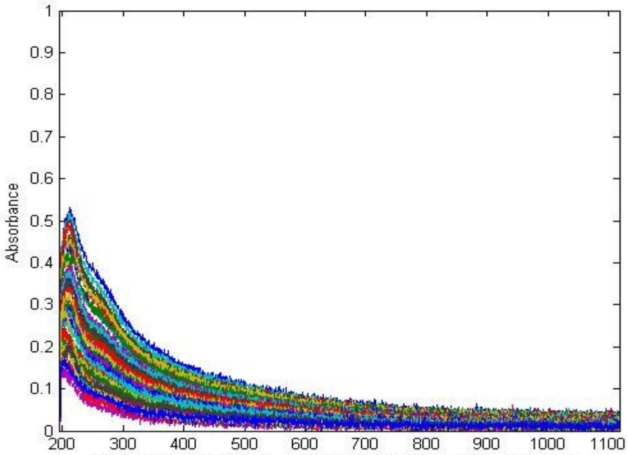
The spectral data of water samples at each sampling sites are shown below.



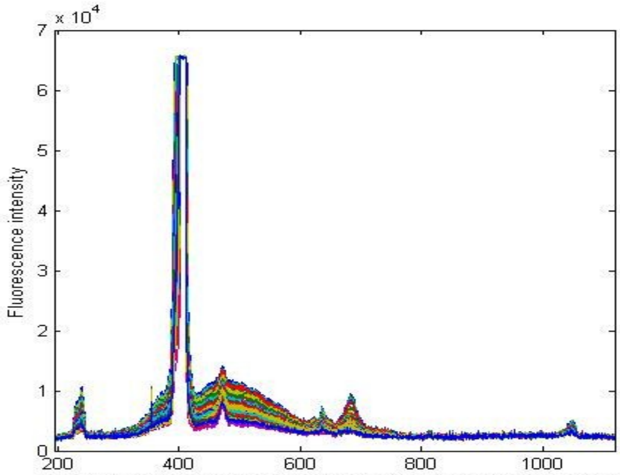
Raw UV-Vis absorbance spectra of 56 water samples in Seven Bridge Wen



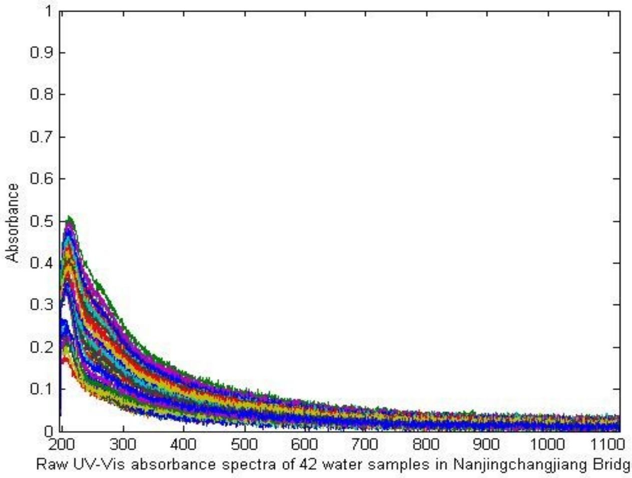
Raw fluorescence emission spectra of 56 water samples in Seven Bridge Wen



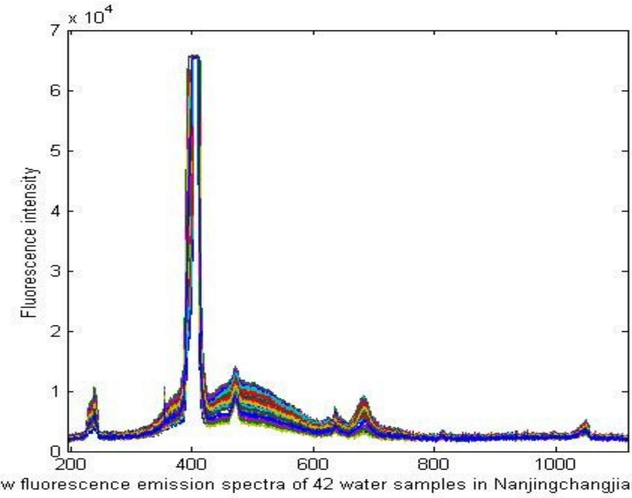
Raw UV-Vis absorbance spectra of 49 water samples in Jiezhizha



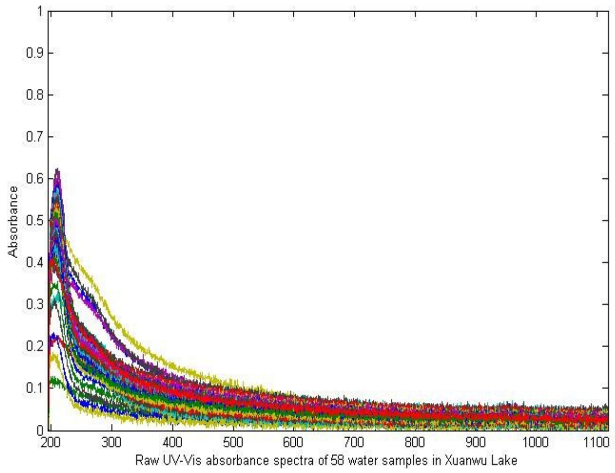
Raw fluorescence emission spectra of 49 water samples in Jiezhizha

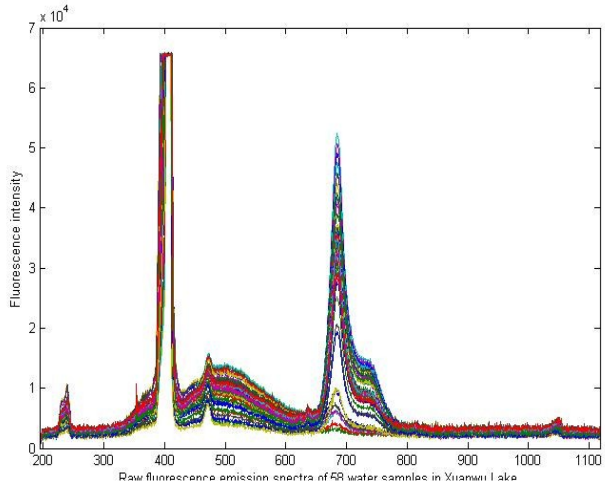


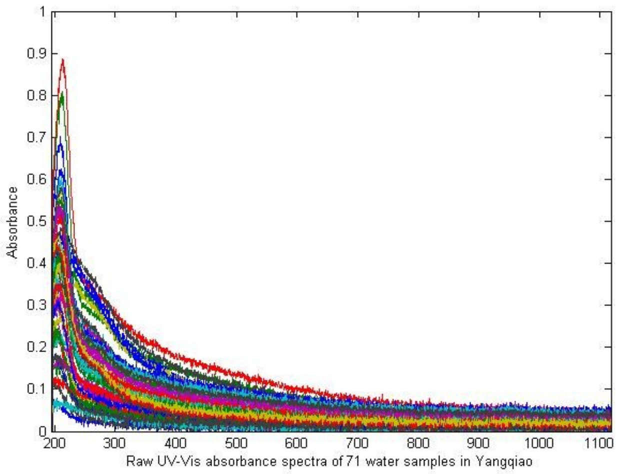
Raw UV-Vis absorbance spectra of 42 water samples in Nanjingchangjiang Bridge

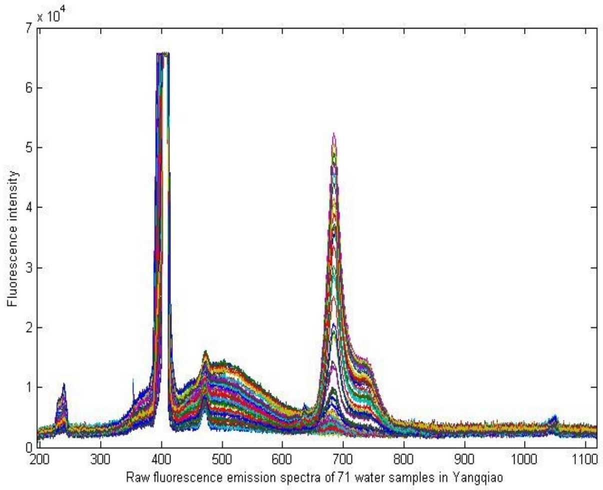


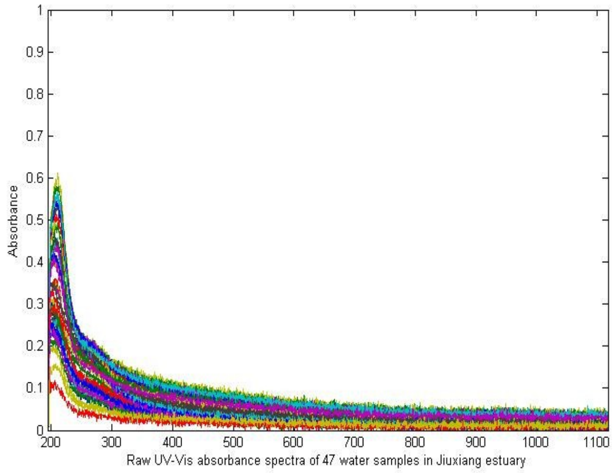
Raw fluorescence emission spectra of 42 water samples in Nanjingchangjiang Brid

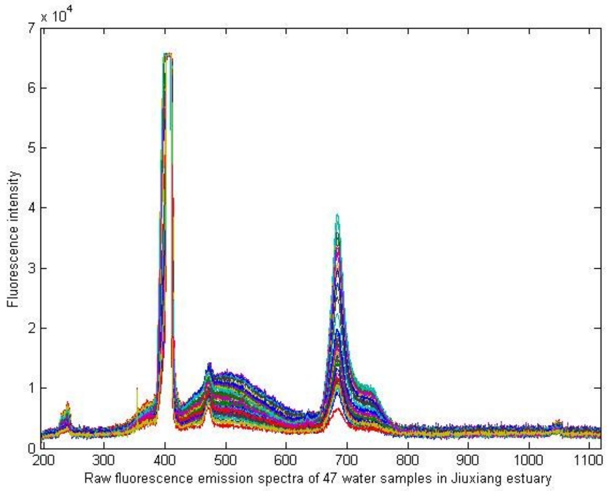












2 Mathematical algorithms

2.1 Joint x-y distance (SPXY) algorithm

The formula from (1) to (4) can express as main calculating steps of SPXY algorithm.

$$x = [x_{UV}, x_F] \qquad (1)$$

$$d_x(p,q) = \sqrt{\sum_{j=1}^{J}\left[x_p(j) - x_q(j)\right]^2} \qquad p,q \in N \tag{2}$$

$$d_y(p,q) = \sqrt{\left(y_p - y_q\right)^2} = \left|y_p - y_q\right| \qquad p,q \in N \tag{3}$$

$$d_{xy}(p,q) = \frac{d_x(p,q)}{max_{p,q \in [1,N]}d_x(p,q)} + \frac{d_y(p,q)}{max_{p,q \in [1,N]}d_y(p,q)} \qquad p,q \in N \tag{4}$$

Formula (1) represents merging each sample normalized UV-Vis absorbance spectral data (x_{UV}) and normalized Fluorescence emission spectral data (x_F) into one vector.

In above calculation (from formula (1) to formula (4)) , where 'J' represents the dimension of sample spectral data, 'N' represents the set of remaining sample, '$x_p(j)$' represents spectral data on sample p in j dimension, '$d_x(p,q)$' represents spectral data distance between $sample\ p\ and\ sample\ q$, '$d_y(p,q)$' represents COD distance between $sample\ p\ and\ sample\ q$, '$d_{xy}(p,q)$' represents comprehensive distance between $sample\ p\ and\ sample\ q$.

Each loop computation can acquire two samples with largest comprehensive distance, and grouped them into training set. The remaining samples are grouped as new 'N' to participate into next iteration. Through circular compute formula from (2) to (4) 130 times, a training set with 260 samples can be obtained. Meanwhile, remaining samples are placed into testing set.

2.2 Discrete wavelet transform

The DWT has been widely studied as a mathematical method that decompose a signal into diverse frequency groups, and provides a valid way for analyzing nonstationary signals. The DWT of signal $x(t)$ can be defined as:

$$DWT(j,k) = \frac{1}{\sqrt{2^j}}\int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t - k2^j}{2^j}\right)dt \tag{5}$$

In equation (5), where $\psi(t)$ is the mother wavelet, '$*$' represents the complex conjugate, 'j' and 'k' $(j,k \in R)$ are two scaling parameters. 'j' determines oscillator frequency and the length of wavelet, 'k' determines shifted position. From those parameters, it is possible to define a dyadic orthonormal wavelet transform and to provide the basis for multi-resolution analysis (MRA).

After determine the decomposition level, the signal $x(t)$ can be entirely decomposed by a set of successive filter bank, as expressed in equation (6).

$$x(t) = \sum_{k=0}^{2^{N-j}-1} a_{j,k}2^{-\frac{j}{2}}\phi\left(2^{-j}t - k\right) + \sum_{j=1}^{J}\sum_{k=0}^{2^{N-j}-1} d_{j,k}2^{-\frac{j}{2}}\psi(2^{-j}t - k \tag{6}$$

Where 'J'$(J \le N)$ is the number of decomposition level, 'N' is the maximum decomposition level, '$a_{j,k}$' is the approximate coefficients at level j, '$d_{j,k}$' is the detailed coefficients at level j.

In this study, a noisy spectrum of a water sample can be expressed in equation (7).

$$x(t) = f(t) + e(t) \tag{7}$$

Where '$x(t)$' is the measured spectral signal, '$f(t)$' is the pure signal, '$e(t)$' is the noise signal. The essence of de-noising signal is inhibit '$e(t)$' part to enhance '$f(t)$' in signal $x(t)$.

Generally, in practice '$f(t)$' always be stationary signal with low frequency. And '$e(t)$' can be thought of as Gaussian white noise with high frequency. Therefore, the noise signal is mostly included in detailed coefficients with higher frequency. Due to this, the following steps were used for de-noising process. First, the signal was decomposed by wavelet packet. Then threshold technique could be used on processing decomposed wavelet coefficients. At last, reconstructed signal to achieve the purpose of de-noising.

2.3 Successive projections algorithm (SPA)

SPA steps are described below, assuming that the first wavelength $k(0)$ and number N are given.

Step1. Before the first iteration$(n = 1)$, let $x_j = j$th column of X_{cal}; $j = 1,\cdots,J$.

Step2. Let S be the set of wavelengths which have not been selected yet. That is $S = \{j \ such \ that \ 1 \le j \le J \ and \ j \notin \{k(0),\cdots,k(n-1)\}\}$.

Step3. Calculate the projection of x_j on the subspace orthogonal to $x_{k(n-1)}$ as:

$$P_{X_j} = X_j - \left(X_j^T X_{k(n-1)}\right) X_{k(n-1)} \left(X_{k(n-1)}^T X_{k(n-1)}\right)^{-1} \tag{8}$$

For all $j \in S$, where P is the projection operator.

Step4. Let $k(n) = arg(max\left\|P_{X_j}\right\|,j \in S)$.

Step5. Let $X_j = P_{X_j}, j \in S$.

Step6. Let $n = n + 1$. If $n < N$ go back to step 1.

End : the resulting wavelengths are $\{k(n);n = 0,\cdots,N + 1\}$.