Multi-Scale Prediction Model of Chloride Content at Huayuankou Station Based on EEMD

Zhanfeng Hou¹, yunpeng Zhang²

¹Henan Provincial Chemi-Industries Research Station Co., Ltd Zhengzhou 450000, China; ²Institute of electric power, North China University of Water Resources and Electric Power, Zhengzhou, Henan 450011, China

Abstract

Chloride is a common inorganic anion in water, whose variation is nonlinear and non-stationary. This paper first performs ensemble empirical mode decomposition (EEMD) of chloride content sequence to acquire a series of intrinsic mode functions (IMFs) which reflect the evolution of chloride sequence, and then performs the Hilbert transform of IMFs, and build a non-linear multi-scale prediction model with the amplitude and frequency of each IMF component. This paper takes the monthly data of chloride content in Huayuankou in the Yellow River as an example and draws the following conclusions: (1) the cycles for chlorides in Huayuankou in the Yellow River include 9, 12, 19, 25, 37 and 68 months, especially 9 and 12 months; (2) the fitting accuracy, significance and residuals of the proposed multi-scale non-linear model all conform to the requirements; (3) predictions are made on chloride content for the next 6 months, with a relative error of 1.1% - 17.1%.

Key words: multi-scale analysis, EEMD, prediction, chloride content

1.INTRODUCTION

Chloride is a common inorganic anion in water. Almost all natural water contains chloride ions, to different extents. The content of chlorine in uncontaminated river water ranges from 10 to 20 mg/L. For example, the content of chloride is only 1.77 mg/L in rivers of the Sanjiang Plain in Northeast China. In human activities, chloride is an important material for physiological and industrial use. Therefore, when the concentration of chloride is rapidly increased, we should consider whether the water is contaminated. The chloride content change is a very complex dynamic process, which is a result of interactions among various complicated factors and is nonlinear and non-stationary (Duan et al.,2016). The prediction of chloride content is an important subject for water quality researchers. The current prediction methods are mostly based on a single scale; however, the evolution of chloride content in runoff is multi-scale, with inter-monthly, inter-annual and even intergenerational variations. Empirical Mode Decomposition (EMD) (Huang et al.,1998; Huang et al.,1999) is decomposition based on local characteristics of time series. It decomposes complex sequences into the sum of finite intrinsic mode functions. EMD is a time-varying adaptive process, which is useful for processing multi-scale non-linear, non-stationary complex sequences. This method has been widely used in signal processing, singularity detection (Xu et al., 2013) and other aspects (Lu,2016). EEMD(Flandrin et al.,2004) is a noise-aided data analysis method proposed to make up the shortfalls of EMD. In this paper, we first perform EEMD of the chloride content sequence in Huayuankou Station of the Yellow River, and then carry out Hilbert transformation, and build a non-linear multi-scale model based on the amplitude and frequency of variations in the sequence extracted through Hilbert transformation.

2.METHODS AND PRINCIPLES

2.1 Empirical mode decomposition (EMD)

The EMD algorithm(Wu and Huang,2009; Ajibade et al.,2016) is defined by the algorithm process, which involves multi-step processing of the time series and gradually decomposing them into IMFs. Each IMF should satisfy two conditions(Mauricio et al.,2016): (1) the number of extreme points (maximum value/minimum value) and the number of zero-crossing points should be equal or at most differ by one; (2) at any time the average value of the upper and lower packets determined by local extreme points (maximum value/minimum
value) is zero. These two conditions are the principles for EMD decomposition and convergence. By EMD, any complex signal can be expressed by a finite number of IMFs. The decomposition process is as follows:

1. Calculate all local maximum value points and minimum value points (in the algorithm, an extreme point refers to a point whose first derivative is zero) in the sequence \( x(t) \). Fit out the upper envelope \( u(t) \) of \( x(t) \) from the maximum value points, and the lower envelope \( v(t) \) from the minimum value points. The average value of the upper and lower envelopes is: \( m(t) = \frac{u(t) + v(t)}{2} \).

2. Deduct the average value of the upper and lower envelopes from the original sequence. Let \( h_1(t) = x(t) - m(t) \). Check if \( h_1(t) \) meets the two conditions for convergence. If it does, \( h_1(t) \) is the IMF. If it does not, repeat (1) for until \( h_k(t) \) meets the two conditions for convergence, and \( h_k(t) \) is the component with the highest frequency in the sequence; deduct the first IMF component from the original sequence, and the remaining portion is \( r_1(t) = x(t) - c_1(t) \).

3. Usually \( r_1(t) \) is a non-stationary sequence. For the remaining part \( r_1(t) \), repeat the steps (1) ~ (2) until the remaining part is a monotonic function, and the decomposition is completed.

Usually, the decomposition is completed when the standard deviation (SD) between the two \( h(t) \)’s

\[
SD = \sum_{i=0}^{T} \frac{(h_{i-1}(t) - h_i(t))^2}{h_{i-1}^2(t)} \quad (1)
\]

When the SD value is between 0.2 and 0.3, the IMF will have clear physical significance.

2.2 Ensemble empirical mode decomposition (EEMD)

In the EMD method, the ability to obtain a proper IMF depends on the distribution of the signal extreme points. If the signal extreme points are not evenly distributed, mode mixing will occur. For this reason, white noise, having evenly distributed frequency spectrums, is added to the signals to be decomposed. Signals on different time scales will be automatically distributed to the appropriate reference scale. Due to the characteristics of white noise, after being averaged for multiple times, the noise will cancel each other, and the resulting average value will be the final output. The decomposition principle for EEMD: when the additional white noise is evenly distributed in the entire time-frequency space, such time-frequency space consists of different scale components divided by the filter bank. When uniformly distributed white noise background is added to the signals, signal zones on different scales will be automatically mapped to the appropriate scales related to the background white noise (Adeniyi et al., 2015).

Steps of EEMD are as follows:

1. Add white noise that obeys normal distribution into the original sequence;
2. Decompose the sequence with white noise into several IMF components through EMD;
3. Repeat steps (1) and (2) and add a new sequence with white noise every time;
4. Take the average value of the IMF obtained each time as the final result.

2.3 Hilbert transformation

Carry out Hilbert transformation of all IMFs decomposed. Hilbert transformation is defined in the following formula (Shi et al., 2015):

\[
y_i(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{c_i(t)}{t-t} dt \quad (2)
\]
In Formula (2), \( P \) is the Cauchy principal value, and \( c_i(t) \) and \( y_i(t) \) form the complex conjugate to produce the analytic function \( C_i(t) \):

\[
C_i(t) = c_i(t) + iy_i(t) = a_i(t)e^{\theta_i(t)} \quad (3)
\]

In the expression of \( C_i(t) \), \( a_i(t) \) is amplitude, and \( \theta_i(t) \) is the phase, i.e.:

\[
a_i(t) = |c_i(t) + iy_i(t)| = \sqrt{c_i^2(t) + y_i^2(t)} \quad (4)
\]

\[
\theta_i(t) = \arctan\left(\frac{y_i(t)}{c_i(t)}\right) \quad (5)
\]

The corresponding frequency is:

\[
\omega_i(t) = \frac{d\theta_i(t)}{dt} \quad (6)
\]

Hilbert transformation can be regarded as a convolution computation of \( c_i(t) \) and \( 1/t \). As the instantaneous frequency is a monotropic function of time, \( \omega_i(t) \) has actual physical significance (Franzke, 2010).

3. CASE STUDY

3.1 Chloride Content Prediction Process

We perform EEMD of the monthly data of chloride content at Huayuankou Station, and the results are shown in Figure 1:

![Figure 1. EEMD of Chloride Content at Huayuankou](image)
We perform a sinusoid curve fitting of each intrinsic mode function (IMF) and a simple linear fitting of the trend term, i.e.:

\[ \text{IMF}_i(t) = A_i \sin \left( \frac{2\pi}{T_i} t + \varphi_i \right) = A_i \sin (\omega_i t + \varphi_i), i = 1, 2, \cdots, 5 \]

Residual(t) = \beta_0 + \beta_1 t

Through Hilbert transformation, we can calculate the instantaneous frequency of each IMF and the average frequency and further obtain the average cycle \( T_i \). The Fourier transform is defined as below:

\[
Y(\omega) = \int_{-\infty}^{\infty} \text{IMF}_i(t) e^{-j\omega t} dt = \frac{A_i \sin(\omega - \omega_i) \frac{T_i}{2}}{(\omega - \omega_i)} e^{-j\omega_i \frac{\pi}{2} + j(\omega - \omega_i) \frac{\pi}{2}}
\]

(7)

When \( \omega = \omega_i \), \( Y(\omega) = \frac{A_i T_i}{2} e^{-j\omega_i \frac{\pi}{2}} \)

Based on \( Y(\omega) \), we can obtain the estimated value \( A_i \) of amplitude and phase.

Where, \( A_i = \frac{2}{T_i} \left| Y(\omega_i) \right|, \varphi_i = \frac{\pi}{2} - \angle Y(\omega_i) \)

We fit all IMF components and obtain \( \text{IMF}_i(t) = A_i \sin(\omega_i t + \varphi_i) \).

Figure 2. Instantaneous Frequency of Chloride Content IMF at Huayuankou

After EEMD, the original sequence of chloride content can be decomposed into six IMFs with different frequency components and a trend term. IMF1–IMF6 represent the 6 decomposed IMFs with different frequencies and the last one is the trend term of the sequence. We perform Hilbert transformation of all IMF components, and calculate the instantaneous frequency, as shown in Figure 2. We obtain the average amplitude from the instantaneous amplitudes of the components and the average frequency from instantaneous
frequencies. The reciprocal of the average frequency is the average cycle. Through Hilbert transform and the Fourier transform, we can obtain the phase. The average cycle, average amplitude and phase are shown in Table 1.

Table 1  Statistical Characteristic Values of IMF Components of Annual Runoff

<table>
<thead>
<tr>
<th>Mode</th>
<th>Average cycle (month)</th>
<th>Average amplitude (mg/L)</th>
<th>Phase (rad)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMF1</td>
<td>9.29</td>
<td>1.31</td>
<td>-0.05</td>
</tr>
<tr>
<td>IMF2</td>
<td>12.16</td>
<td>1.68</td>
<td>-0.03</td>
</tr>
<tr>
<td>IMF3</td>
<td>18.91</td>
<td>1.26</td>
<td>0.27</td>
</tr>
<tr>
<td>IMF4</td>
<td>25.67</td>
<td>0.38</td>
<td>0.14</td>
</tr>
<tr>
<td>IMF5</td>
<td>37.53</td>
<td>0.97</td>
<td>-0.18</td>
</tr>
<tr>
<td>IMF6</td>
<td>67.85</td>
<td>0.63</td>
<td>-0.08</td>
</tr>
</tbody>
</table>

From the results of EEMD shown in Figure 1 and Table 1, we can see that the first component contains the information on the oscillation with the highest frequency, which indicates that the annual runoff sequence in Huayunkou has an oscillation cycle of about 9 months; the frequency from IMF1 to IMF6 gradually decreases. Besides the 9-month cycle, other cycles for chloride content also include 12, 19, 26, 37 and 68 months, with 9 and 12 months as the dominant cycles.

3.2 Multi-scale modeling process

According to the results of EEMD and Hilbert transformation, the sequence of chloride content at Huayunkou is a non-linear sequence with multi-scale changes. By using 6 approximate cycle terms and a linear term, we build a model as follows:

\[ Y = \beta_0 + \beta_1 t + \sum_{i=1}^{n} A_i \sin\left(\frac{2\pi}{T_i} t + \varphi_i\right) + \varepsilon \]  (8)

Where, \( A_i \) is the average amplitude, \( T_i \) is the average cycle for IMF, \( \beta_0 + \beta_1 t \) is the trend term, and \( n \) is the number of components, and in this paper, \( n=6 \), \( \varphi_i \) is the phase and \( \varepsilon \) is the residual term.

The multi-scale runoff model is as follows:

\[
\text{Runoff} = 7.7152 + 0.0043t + 1.31\cos\left(\frac{2\pi}{9.29} t - 0.05\right) + 1.68\cos\left(\frac{2\pi}{12.16} t - 0.03\right) \\
+ 1.26\cos\left(\frac{2\pi}{18.91} t + 0.27\right) + 0.38\cos\left(\frac{2\pi}{25.67} t + 0.14\right) + 0.97\cos\left(\frac{2\pi}{37.53} t - 0.18\right) \\
+ 0.63\cos\left(\frac{2\pi}{67.85} t - 0.08\right) + \varepsilon \quad t \in [1, 240]
\]

The fitting result is shown in Figure 3.
4. RESIDUAL ANALYSIS

The residual (Li et al., 2014) is defined as $\varepsilon_i = y_i - \hat{y}_i$. It is the sequence formed by the estimated value of $\varepsilon$ in the model, called the residual sequence. When the fitting equation can well reflect the characteristics and variation of the dependent variable, there should not be obvious regularity and trend in the residual sequence. For a time sequence with a definite interval, correlation test needs to be performed on the residual sequence (Li and Chang, 2012; Zhang et al., 2015). The residuals are shown in Figure 4.

![Residual Plot](image)

Figure 4. Residual Plot

It can be seen from the residual plot that the residuals are random, i.e. there is no obvious trend and regularity.

5. MODEL PREDICTION

We use the model to predict the data in the 241st-245th month, and the results are listed in Table 2:
The EEMD approach decomposes the data according to the time-scale characteristics, without any pre-set basis functions. This is essentially different from the Fourier decomposition and the wavelet decomposition based on respectively harmonic basis function and wavelet basis function. Fourier's harmonic basis function assumes that in case of limited data, when the characteristic frequency in the data is not consistent with the harmonic frequency of the basis function, it can easily lead to serious distortion of signals. Due to its difference from the Fourier and wavelet decomposition methods, EMD has higher prediction accuracy.

6. CONCLUSIONS

In this paper, we use EEMD, Hilbert transform and Fourier transform to simulate the chloride content at Huayuankou for 240 months, and the chloride content shows obvious cycles of 9 and 12 months. The modeling method presented in this paper can reflect the physical meaning of the evolution of the runoff sequence and the variation characteristics of time scale. The orthogonality of EEMD makes the multi-scale model predict better and more accurate. Based on the predicted dissolved oxygen content over the next 6 months, the relative error of the model is proved to be between 1.1% and 17.1%.

REFERENCES


Tabla 2 Relative Errors between Measured Values and Simulated Values

<table>
<thead>
<tr>
<th>Time (month)</th>
<th>Measured value (mg/L)</th>
<th>Simulated value (mg/L)</th>
<th>Relative error</th>
</tr>
</thead>
<tbody>
<tr>
<td>241</td>
<td>7.92</td>
<td>9.27</td>
<td>17.1%</td>
</tr>
<tr>
<td>242</td>
<td>10.39</td>
<td>10.28</td>
<td>1.1%</td>
</tr>
<tr>
<td>243</td>
<td>11.68</td>
<td>10.35</td>
<td>11.4%</td>
</tr>
<tr>
<td>244</td>
<td>9.12</td>
<td>9.59</td>
<td>5.1%</td>
</tr>
<tr>
<td>245</td>
<td>8.03</td>
<td>8.38</td>
<td>4.4%</td>
</tr>
<tr>
<td>246</td>
<td>7.87</td>
<td>7.18</td>
<td>8.8%</td>
</tr>
</tbody>
</table>