# A New Symbolization and Distance Measure based Anomaly Mining Approach for Hydrological Time Series

Pengcheng Zhang[1+], Yan Xiao[1], Yuelong Zhu[1], Dingsheng Wan[1], Wenrui Li[1,2], Hareton Leung[3]

1. College of Computer and Information, Hohai University, Nanjing, China
2. School of Mathematics & Information Technology, Nanjing Xiaozhuang University, China
3. Department of Computing, Hong Kong Polytechnic University, HongKong, China
pchzhang@hhu.edu.cn; hhu_xiaoyan@163.com; wenrui_li@163.com
+ Corresponding Author

**ABSTRACT:**

Most of the time series data mining tasks attempt to discover data patterns that appear frequently. Abnormal data is often ignored as noise. There are some data mining techniques based on time series to extract anomaly. However, most of these techniques cannot suit big unstable data existing in various fields. Their key problems are high fitting error after dimension reduction and low accuracy of mining results.

This paper studies an approach of mining time series abnormal patterns in the hydrological field. We propose a new idea to solve the problem of hydrological anomaly mining based on time series. We propose Feature Points Symbolic Aggregate Approximation (FP_SAX) to improve the selection of feature points, and then measures the distance of strings by Symbol Distance based Dynamic Time Warping (SD_DTW). Finally, the distances generated are sorted. A set of dedicated experiments are performed to validate our approach. The results show that our approach has lower fitting error and higher accuracy compared to other approaches.

**KEY WORDS:**
*Hydrological Time Series; Data Mining; Pattern Representation; Distance Measure*

# INTRODUCTION

With the advance development of technology, the data which need to be dealt with is becoming various and complicated. Furthermore, the scale of the data is huge, the form of the data is more diverse and the speed of processing data is lacking. So how to gain valuable information and meaningful knowledge quickly from the numerous and complex big data has become a key challenge.

Data Mining (Lin & Chen, 2011) extracts potentially useful information which people is interested in and which is unknown in advance. There is a variety of knowledge representation, such as concepts, rules, regularities and patterns. Time series (Box, Jenkins et al. 2013), which reflect the characteristic of the attribute value related with time, have large data scale, high dimension and frequent update.

As an important branch in the field of data mining research, time series data mining is the process of pattern discovery and knowledge extraction from the time sequence. The basic task includes similarity searching, periodical pattern mining, analysis and prediction, clustering and classification, visualization and abnormal data mining of time series. For a long time, the aim is finding the data pattern which appears frequently. People want to summarize some rules from the

data pattern. Currently, almost all abnormal data are considered as noisy data and then ignored. In some cases, compared with the normal data, the frequency of abnormal data is not high, but it may hide some important information. And finding these abnormal data and the corresponding hidden information may provide valuable information and more enlightening knowledge.

Abnormal data are considered intuitively to be those which are associated with the data model or the data objects but do not conform to the general distribution. Nowadays there is not a definition generally accepted for abnormal data. It changes with the specific application. The research for abnormal time sequence mining is derived relatively late, but recently it is attracting more interest. For instance, this technology can monitor the crime hidden in the electronic commerce. And on the other hand, it can be used to detect possible intrusions by Hackers in the daily management of the Internet. In these applications, abnormal data ofen represent more useful meaning than other normal data. Although the probability of the abnormal data is very small, their importance cannot be ignored. Therefore capturing these small probability events in the abnormal cases is important for some applications.

Hydrological data are discrete records for hydrological process, such as flow, rainfall, water level and so on (Ping, 2003). They are huge, noisy, unstable and have poor correlation. Nowadays, hydrological modernization is popular, which contains information collection, extraction, analysis and so on. During the hydrological process, it is indispensable to acquire valuable information and meaningful knowledge quickly from large amount of data. The rapid development of data mining provides a new approach for water resource management, hydrology and hydroinformatics research (Kozerski, 2010).

Hydrological time series data mining (Kozerski, 2010) can be used to extract the unknown process which contains important information from hydrological data, which is valuable for hydrological forecasting and hydrological data analysis. Abnormal hydrological time series are the data objects which are obviously inconsistent with the universal rule of the hydrological phenomenon. In the field of hydrological data mining, there are three branches: similarity search, sequence pattern mining and cycle analysis. The research of abnormal hydrological time series data mining is still at the starting stage.

There are many approaches for abnormal time series data mining. Most of them have clear problems. For example, the approach based on immunology (Chaovalit, Gangopadhyay et al. 2011) cannot apply to diverse data. Computational efficiency of Support Vector Machine (SVM) (Verdejo, Garc á et al. 2011) is high, but its theory and modeling process are very complex which can only be adopted by experts. The accuracy of TSA-tree (Zhang, Meratnia et al. 2010) is low. To solve these problems, this paper puts forward a new approach which is based on Extended Symbolic Aggregate Approximation (ESAX) (Lkhagva, Suzuki et al. 2006) and Dynamic Time Warping (DTW) (Müller, 2007).

In summary, the contributions of this paper are as follows:

- Approaches of selecting feature points (extreme points, minimum or maximum) are added into ESAX in pattern representation to reduce dimension, which is called FP_SAX. FP_SAX looks for new feature points which are more representative to replace the maximum and minimum proposed in ESAX. In FP_SAX, feature points consist of the following three parts: the beginning and ending points, the extreme feature points and the piecewise average feature points.
- This paper achieves a first combination of symbolic process and SD_DTW which is based on the distance between each symbol and DTW. A good mining result is acquired by this combination.

- A set of dedicated experiments have been conducted to validate our approach. Many approaches are studied in the experiment, such as Lagrange interpolation, data compression ratio and so on. Experimental results show low fitting error, acceptable time complexity and high accuracy of our approach.

The rest of this paper is organized as follows. Section 2 reviews related work and discusses some background materials about time series data mining. In Section 3, the anomaly mining approach is proposed. Firstly, based on ESAX, we improve the approach of selecting feature points in symbolization, and then FP_SAX is proposed. Secondly, we measure the distance of strings according to the distance of DTW, which is called SD_SAX. The experimental validation of the proposed approach is performed in Section 4. Finally, Section 5 offers some discussion and suggestions for future work.

# Related Work

Because of the large amount and high dimensionality of time series data, using the original data set pays abundant time and space cost. Dimensionality reduction techniques, also called sequence feature extraction, can translate big data to small data. After dimensionality reduction, we measure the distance of strings. Finally we are able to identify the main feature from the results of distance measure. In the following, we review related work about pattern representation and similarity measurement of time series.

### Pattern Representation

So far there have been four basic pattern representation approaches to extract sequence feature. These approaches are as follows: Piecewise Linear Representation (PLR) (Fu, 2011), Frequency Domain Representation (FDR) (Bigioi, Ciuc et al. 2009), Singular Value Decomposition (SVD) (Henry & Hofrichter, 1992), and Symbolic Aggregate Approximation (SAX) (Lin, Keogh et al. 2007).

Keogh, Chakrabarti et al. (2001) proposed Piecewise Aggregate Approximation (PAA). This approach divides the original time series into several segments with equal length. The mean value of each segment is used as the feature of the segment. Then the original time series is represented by the feature of the segment. So the dimensionality of data set is reduced. After that, Keogh proposed PLR (Fu, 2011). In PLR, after being divided into several segments, time series is represented by end-to-end segments. The number of segments impacts on the level of compression. It should not be too high or too low. But there is no criterion for the choice of this number. At the same time, PLR is not applicable to nonlinear sequence.

Frequency Domain Representation changes the original time series from the time domain to frequency domain. There are two typical FDR approaches: Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT). DFT appears in the field of digital signal processing early. Then it is proposed by Agrawal, Faloutsos et al. (1993) again to apply to similarity searching. DFT allows a good dimension reduction, and measures the distance between each point in k-dimensional space by Euclidean distance. But some important extreme points are missed. In DWT (Starck, Murtagh et al. 2010), the time series is analyzed by translation transformation and stretching transformation. The dimensionality is reduced without losing important points. However, both approaches cannot apply to weighted Euclidean Distance.

SVD is a significant matrix distributing approach in linear algebra. It is mainly applied to image processing, text indexing, signal processing and statistics. SVD (Henry & Hofrichter, 1992) transfers a group of given correlated variables into another group of uncorrelated variables whose variances are in a descending order, then generates a coordinate axis by mathematical manipulation. Due to the difference of the variance of each axis, the original time series can be represented by several coordinate coefficients whose variances are at the top. Then the purpose of reducing dimensions is accomplished. However, the original time series loses basic physical significance after SVD (Esling & Agon, 2012). In addition, the process for reducing dimension involves global transformation of all data, which is relatively complex.

Time series is generally composed of continuous real values. Symbolic representation makes the original time series discrete, mapping to the limited symbol table, then expressed as an ordered set of limited symbols -- the string sequence. This method is firstly applied to similarity search of time series, which uses a symbol to represent a sequence, so as to achieve the goal of time series dimension reduction. Then some existing string matching methods can be used to implement the similarity search. SAX (Lin, Keogh et al. 2007), which is the most representative symbolic approach, is proposed by Keogh. It is based on PAA for dimensionality reduction that minimizes dimensionality by the mean values of equal sized frames (Lkhagva, Suzuki et al. 2006). Then the results are turned into SAX symbols. However, SAX often loses some points that may contain important information, such as the max/min points and extreme points. Finally, Lkhagva (Lkhagva, Suzuki et al. 2006) put forward ESAX (Extended SAX) in 2006. ESAX overcomes the problem of missing important points from SAX, which has a good performance in economic time series data mining. However, some of the feature points are still missed by both techniques.

## Similarity measurement

After reducing dimensionality, it is time to do similarity measurement. In many fields, distance measurement is similar with similarity measurement. Furthermore, distance between two objects is easy to compute. Therefore, similarity measurement is often replaced by distance measurement. There are two classical means: Euclidean Distance (Russ & Woods, 1995) and Dynamic Time Warping (DTW) (Müller, 2007).

Euclidean Distance (Russ & Woods, 1995), as the most widely used distance measurement, is simple, intuitive and easy to calculate. However, it is applicable to the time series with equal length only. And it is susceptible by noise and short time fluctuation. When some deformation occurs in time series, the interference is unable to be eliminated using Euclidean Distance, which lead to a bad result. The calculation is shown in (1). Weighted Euclidean Distance (Keogh & Pazzani, 1998) is an improvement of Euclidean Distance, whose calculation is shown in (2). It can eliminate the influence of linear drift factors. For instance, in hydrology, the weight of peaks and troughs can be increased to gain more accurate similarity results.

$$L_p(X,Y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{1}$$

$$D_w(X,Y) = \sqrt{\sum_{i=1}^{n} w_i |x_i - y_i|^2} \tag{2}$$

Comparing with Euclidean Distance, DTW (Müller, 2007) is able to measure the distance without point-to-point correspondence when the timeline of time series has compand and bend. It also applies to the time series with different lengths.

After symbolization, the time series become character strings composed of discrete, relatively abstract symbols. The theoretical achievements in text mining can be used to extract character strings after symbolization. There are three basic distance measure approaches in text mining: Levenshtein Distance, Longest Common String, and vector space method. However, they need the source string for comparison, which is impossible in the field of hydrology.

# Preliminaries

Hydrological data is nonlinear, huge and has poor correlation. Hydrological process is complex, stochastic and unstable. The existing approaches cannot meet the demand of hydrological time series anomaly mining. In order to solve these problems, this paper proposes a new approach which combines FP_SAX and SD_DTW to give a good dimension reduction, low fitting error and relatively high accuracy.

## ESAX

ESAX is proposed in 2006 to overcome the problem of losing critical and extreme points in SAX (Lin, Keogh et al. 2007). In ESAX, the sequence must conform to a standard normal distribution. But hydrological data set is random and does not conform to the standard normal distribution. So before symbolization, the hydrological sequence should be standardized. The specific steps of ESAX are as follows (Lkhagva, Suzuki et al. 2006):

a) Standardize original sequence C to C'. u and v respectively represent the mean value and standard deviation of this sequence:

$$c_i' = \frac{c' - u}{v} \tag{3}$$

b) Translate C' whose length is n to X whose length is k by PAA:

$$x_i = \frac{k}{n} \sum_{j=\frac{n}{k}(i-1)+1}^{\frac{n}{k}i} c_j' \tag{4}$$

c) Divide *X* into *a* equiprobable spaces. Based on the values of $x_1, x_2, \ldots, x_k$, the sequential values in the same probability space are represented by one symbol. The total number of symbols is *a*. Then we get a symbol string whose length is *k*. Table I shows the division of equal probability interval (Lin, Keogh et al. 2003).

*Table I The division of equal probability interval based on the number of symbols (from 3 to 10)*

| $a$ / $\beta_i$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| $\beta_1$ | -0.43 | -0.67 | -0.84 | -0.97 | -1.07 | -1.15 | -1.22 | -1.28 |
| $\beta_2$ | 0.43 | 0 | -0.25 | -0.43 | -0.57 | -0.67 | -0.76 | -0.84 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\beta_3$ | | 0.67 | 0.25 | 0 | -0.18 | -0.32 | -0.43 | -0.52 |
| $\beta_4$ | | | 0.84 | 0.43 | 0.18 | 0 | -0.14 | -0.25 |
| $\beta_5$ | | | | 0.97 | 0.57 | 0.32 | 0.14 | 0 |
| $\beta_6$ | | | | | 1.07 | 0.67 | 0.43 | 0.25 |
| $\beta_7$ | | | | | | 1.15 | 0.76 | 0.52 |
| $\beta_8$ | | | | | | | 1.22 | 0.84 |
| $\beta_9$ | | | | | | | | 1.28 |

d) Find the maximum and minimum of each PAA subsection $x_i$. Translate them separately into symbols $S_{max}$ and $S_{min}$. Save their positions $P_{max}$ and $P_{min}$ at the same time. Then calculate the mean value of each subsection whose middle position is as follows (calculated from both the beginning position $S_k$, and the ending position $E_k$ on the time axis):

$$P_{mid} = \frac{S_k + E_k}{2} \tag{5}$$

e) The three symbols of each subsection $x_i$ can be represented by the following equation:

$$< S_1, S_2, S_3 >= \begin{cases} < S_{max}, S_{mid}, S_{min} > & if\ P_{max} < P_{mid} < P_{min} \\ < S_{min}, S_{mid}, S_{max} > & if\ P_{min} < P_{mid} < P_{max} \\ < S_{min}, S_{max}, S_{mid} > & if\ P_{min} < P_{max} < P_{mid} \\ < S_{max}, S_{min}, S_{mid} > & if\ P_{max} < P_{min} < P_{mid} \\ < S_{mid}, S_{max}, S_{min} > & if\ P_{mid} < P_{max} < P_{min} \\ < S_{mid}, S_{min}, S_{max} > & otherwise \end{cases} \tag{6}$$

In ESAX, the time series are divided equally. All changes are considered in the same way whether they are slow or not. In fact, these two situations represent different significance in time series, especially in hydrological phenomena. If merely using maxima and minima of every segment as feature points, we are unable to make sure that every point which has effect on the series is included. On the other hand, if dividing the original time series into pieces of segments with different lengths based on the change state, expanding the range interval when changing slowly, narrowing the range interval when changing acutely, and remaining all feature points which can represent the trend of time series by specific methods, we can find more suitable feature points to represent the original series. Based on these feature points, the symbolic process will be the best-fit for the original series. At the same time, it can keep important features of the original series. Therefore, in this paper, a new improved algorithm FP_SAX is put forward.

## FP_SAX

*1)Dimension reduction approaches based on extreme points*

In hydrological time series data mining, the peaks and valleys in a time series are often particularly important to analyze the transformation of hydrological phenomena for a certain time period. So peaks and valleys are more special than other sequence points, and they must be retained as much as possible. Therefore, to keep extreme points such as peaks and valleys, dimension reduction approaches based on extreme points are indispensable.

**Definition 1: Extreme points.** For time series $X = (x_1, x_2, \ldots, x_m)$, if one of the following conditions is met, we call $x_i$ the extreme point:

$$(1) x_i \geq x_{i-1} \; and \; x_i \geq x_{i+1};$$
$$(2) x_i \leq x_{i-1} \; and \; x_i \leq x_{i+1}.$$

This approach can retain all extreme points of the series. But its ability to reduce dimension depends on changes of the series, because it just simply considers the relative size between the previous sequence and the later one, and has no macroscopic consideration of the entire sequence. If the sequence changes gently, this dimension reduction approach based on extreme points is good to choose important sequence point to represent the series. If the sequence changes acutely, such as zigzag sequence, the compression ratio of this dimension reduction will be very low, because it is not able to smooth small amplitude vibration, which may lead to too much noise in the result of the dimension reduction.

Another widely used dimension reduction approach – PAA (Papapetrou, Athitsos et al. 2011), which is simple and easy to implement, gives good results on the sequence which changes gently. Its core idea is to fix the width of window to keep the consistency. Then it calculates piecewise average of each window. The width of each window is demanded to be appropriate. If the width of the window is too wide, it can't reflect changes of sequences in time. If too narrow, the data compression is small relatively. It cannot reach the purpose of dimension reduction. Therefore, improper selection of the window will have a big impact on subsequent mining. For example, if putting a series of data which changes greatly into one window, or using average values when the sequence changes fast, it is easy to smooth sequence fluctuation. The resulting average values cannot represent the feature of this sequence well.

In order to solve the problems existed in extreme points and PAA dimension reduction approaches, we improve the traditional technology of extreme points and ESAX Symbolic approaches, then propose FP_SAX which can meet the needs of dimension reduction for hydrological time series.

*2)The selection of feature points*

FP_SAX is still based on SAX. Besides the maximum and minimum proposed in the ESAX, we look for new feature points which are more representative to replace the maximum and minimum, so that the important information of the sequence won't be lost. The symbolic process according to these new feature points is able to preserve important feature of the original sequence.

In FP_SAX, feature points consist of the following three parts: the beginning and ending points, the extreme feature points which keep the extreme time period and the piecewise average feature points containing certain number of extreme points.

**Definition 2: Extreme feature points.** (Xiao & Hu, 2005) For time series $X = (x_1, x_2, \ldots, x_m)$, $x_i$ is extreme feature points, if it meets the following two conditions:
(1) $x_i$ must be extreme value point of the sequence;

(2) $x_i$ keeps extreme time period (the distance between the previous extreme point and the later one). The ratio, which is obtained through dividing the extreme time period by the length of this sequence, must not be less than the threshold value *B*. According to the length of the original time series and domain knowledge, the value of *B* is usually between 0.001 and 0.1.

**Definition 3: Average feature points.** Time series $X = (x_1, x_2, \ldots, x_m)$ has $k$ extreme feature points. Divide the sequence into subsequences to make sure that each subsequence contains *N* extreme feature points. The mean value of each subsequence is called the average feature point.

The minimum value of *N* is 1, which means that this subsequence only contains one extreme point. The maximum value of *N* is *k* which is the number of extreme feature points of this sequence. In this case, the sequence has only one average feature point that is the mean value of all sequence points.

Average feature points break the requirement that PAA dimension reduction approaches must use fixed window size. The width of window is decided by the change of sequence. If a subsequence contains many feature points, frequent fluctuations of this subsequence are illustrated and this subsequence has an effect on the whole sequence. So it must be recorded in detail. Then we should narrow the width of the window. If a sequence contains less feature points, which shows that this sequence changes gently, we should enlarge the width of window, in order to improve the data compression ratio. Average feature points can achieve both the purpose of dimension reduction, and capture important characteristics of the sequence.

After selecting suitable feature points, according to abscissa ascending order size of feature points, the total number of symbols can be determined. According to table I (Lin, Keogh et al. 2003) which provides the dividing principle of normal distribution equal probability interval, each feature point is mapped to the matched symbol interval. Then symbols are obtained, and the original sequence is translated into a string.

The algorithm of FP_SAX is specified in Algorithm 1. Lines 1 to 7 standardize the original sequence. Lines 8 to 12 keep extreme points. Lines 14-17 determine the extreme feature points and save them. Finally, feature points are transformed to symbols according to Table I and equation (6).

Algorithm 1：Dimensionality Deduction of Time Series based on FP_SAX

Input： Original time series $X = (x_1, x_2, \ldots, x_i, \ldots, x_n)$, threshold value *w*, number of extreme feature points *k*;
Output： The string $S = (s_1, s_2, \ldots, s_i, \ldots, s_m)$;
(1) **for** i = 0 to n **do**
(2) xi→a[i]； // Calculate the mean value and standard deviation *u*;
(3) **end for**
(4) **for** i = 0 to n **do**
(5) **if** ((a[i]>a[i+1]) and (a[i]>a[i-1])) or ((a[i]<a[i+1]) and (a[i]<a[i-1]))
(6) {a[*i*]→b[*j*]; *j*++;}
(7) **end if**
(8) **end for**
(9) **for** i = 1 to j-1 **do**
(10) {if (b[i+1]—b[i-1]>w*n)
(11) { b[i]→c[p].value;
(12) i→c[p].location; p++}

```
(13)  end if
(14)  if (p = = k)
(15)    Calculate the mean value and coordinate;
(16)    continue；}
(17)  end if
(18) end for
(19) for i=0 to m do
(20)  s[i]=corresponding symbols of mean and extreme feature points;
(21)  cout<<s[i];}
(22) end for
```

## SD_DTW

*1) SD_DTW*

For similarity measurement, DTW is more accurate than Euclidean distance and can be applied to the compand of the time shaft, so some distance measurement has been developed from it. Based on the idea of DTW, combining the distance between the FP_SAX symbols, SD_DTW is proposed to solve the problem of distance measurement.

SD_DTW describes the distance between each symbol through a matrix, of which *i* and *j* respectively represent rows and columns. The element of matrix is as follows (Enright, Van et al. 2002):

$$\text{dis[i][j]} = \begin{cases} 0, & if\ |i-j| \le 1 \\ \beta_{\max(i,j)-1} - \beta_{\min(i,j)}, & otherwise \end{cases} \tag{7}$$

The value of $\beta_n$ is in the reference Table I (Lin, Keogh et al. 2003).

For example, when the total number of symbols *a* is 5, A, B, C, D, E are used to represent the original time series. Then the distance between each symbol is shown in Table II.

*Table II The distance between each symbol when a=5*

| | *A* | *B* | *C* | *D* | *E* |
|---|---|---|---|---|---|
| *A* | 0 | 0 | 0.59 | 1.09 | 1.68 |
| *B* | 0 | 0 | 0 | 0.5 | 1.09 |
| *C* | 0.59 | 0 | 0 | 0 | 0.59 |
| *D* | 1.09 | 0.5 | 0 | 0 | 0 |
| *E* | 1.68 | 1.09 | 0.59 | 0 | 0 |

**Definition 4: Symbol Distance based Dynamic Time Warping (SD_DTW)**. Two strings $S = (s_1, s_2, \dots, s_{n1})$ and $T = (t_1, t_2, \dots, t_{n2})$, whose length are respectively $n_1$ and $n_2$, are arranged by time. An *m*n* matrix *A* is constructed to represent DTW distance of the two strings.

The element $a_{ij}$ in matrix *A* is the distance between $s_i$ and $t_j$, $\mathrm{d}(s_i, t_j)$, the equation is shown as (7).

There is a set $\mathrm{W} = \{\mathrm{w}_1, \ldots, \mathrm{w}_\mathrm{k}, \ldots, \mathrm{w}_\mathrm{K}\}(\mathrm{w}_\mathrm{k} = \mathrm{d}(\mathrm{x}_\mathrm{i}, \mathrm{y}_\mathrm{j})$ containing a group continuing matrix elements. *W* is the bent lane of S and T, which complies with the following principles (Ouyang, Ren et al. 2010):

(i) *Boundary*: both beginning point and ending point of the bent lane are located in the back-diagonal of the relational matrix and respectively the beginning/ending point of two time series are $w_1 = \mathrm{d}(x_1, y_1), w_k = \mathrm{d}(x_m, y_n)$.

(ii) *Continuity*: any two points of the bent lane must be adjacent elements or diagonal adjacent elements of relational matrix. If $w_k = \mathrm{d}(x_a, y_b), w_{k-1} = \mathrm{d}(x_{a'}, y_{b'})$, then a-a'≤1,b-b'≤1.

(iii) *Monotonicity*: all points in the bent lane should satisfy monotonicity. If $w_k = \mathrm{d}(x_a, y_b), w_{k-1} = \mathrm{d}(x_{a'}, y_{b'})$, then a-a'≥0,b-b'≥0.

Following is the equation of DTW distance (Ouyang, Ren et al. 2010):

$$\begin{gathered} \mathrm{D}(<>,<>) = 0; \\ \mathrm{D}(\mathrm{x},<>) = \mathrm{D}(<>, y) = 0; \\ \mathrm{D}(1,1) = \mathrm{d}(x_1, y_1); \\ \mathrm{D}(\mathrm{i},\mathrm{j}) = \mathrm{d}(x_i, y_j) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\} \end{gathered} \tag{8}$$

D(i,j) is cumulative distance, which is the sum of minimum value of this point and minimum bent path in the upper left. This is a kind of dynamic planning approach based on the cumulative distance matrix to calculate DTW distance.

The algorithm of SD_DTW is shown in Algorithm 2. In lines 2-4 and 5-7, the algorithm separately gives the distance of first line and column. Lines 8-11 calculate cumulative distances.

Algorithm 2：SD_DTW

```
Input：  String X and String Y, whose lengths are separately m and n;
Output：  Distance of String X and String Y based on DTW
(1) v[0][0]=dis(x[0],y[0]);
(2) for i = 0 to m do
(3)   v[i][0]=v[i-1][0]+dis(x[i],y[0]);
(4) end for
(5) for j = 0 to n do
(6)   v[0][j]=v[0][j-1]+dis(x[0],y[j]);
(7) end for
(8) for j = 0 to n do
(9)   for i = 0 to m do
(10)    v[i][j]=dis(x[i],y[j])+min{v[i-1][j],v[i][j-1],v[i-1][j-1]};
(11)  end for
(12) end for
(13) return v[m-1][n-1]
```

The time complexity of DTW is $\mathrm{O}(\mathrm{n}_1 * \mathrm{n}_2)$, where $\mathrm{n}_1$ and $\mathrm{n}_2$ are the lengths of two strings.

For instance, we want to calculate DTW distance between S=ABDC and T=DCADBE. The total number of symbols *a* is 5. Then DTW distance between S and T is 1.68, and the grey plaid entries in Table III give the best winding path.

*Table III The DTW distance between S and T*

| T \ S | *A* | *B* | *D* | *C* |
|---|---|---|---|---|
| *D* | 1.09 | 1.59 | 1.59 | 1.59 |
| *C* | 1.68 | 1.09 | 1.09 | 1.09 |
| *A* | 1.68 | 1.09 | 2.18 | 1.68 |
| *D* | 2.77 | 1.59 | 1.09 | 1.09 |
| *B* | 2.77 | 1.59 | 1.59 | 1.09 |
| *E* | 4.45 | 2.58 | 1.59 | 1.68 |

*2)Parallel computing*

When dealing with large amount of data, it is time consuming. In this case, parallel computing is a good choice to minimize time. Parallel computing refers to the process of the simultaneous use of multiple computing resources to solve computational problems, which is an effective means to improve the processing power. Its basic idea is to use multiple processors to solve the same problem collaboratively. The problem will be broken down into several parts. Each part will be computed in parallel by a separate processor. Parallel computing can solve problems that contain the following features well. 1) The problem can be divided into discrete parts, which is helpful to be solved simultaneously. 2) Multiple program instructions can be executed concurrently. 3) Consuming time under multiple computing resources is less than a single computing resource. When calculating many groups of DTW distance between strings, parallel computing can be used to reduce time and improve efficiency.

# Experimental evaluation

In this section, we conduct a set of experiments to show the accuracy and usability of our new approach by comparing it with ESAX (Liu, Zhu et al. 2012). The experiments are designed to answer the following three research questions:

- REQ 1: What is the difference of fitting error between the two approaches under different data compression ratios?
- REQ 2: How much execution time is required?
- REQ 3: What is the accuracy of our approach?

**Experimental setup:** To address REQ 1, fitting error is estimated by interpolation approaches based on daily water level data from Xiaomeikou gauge station in 2006. Various data compression ratios are used. We compare fitting errors under different data compression ratios. For REQ 2, we separately select water level data in the same period of 2 months (July and

August), 3 months (June to August), 4 months (June to September), 5 months (June to October), 6 months (May to October) and 7 months (April to October) from 1956 to 2005. Then the average execution time of each subsequence is calculated by two approaches respectively. For REQ 3, three experiments with water level data in May and June, July and August, and June to October from 1956 to 2005, are conducted to investigate the results of our approach. In each experiment, we identify the first six largest distances by the two approaches, which represent the abnormal patterns.

Software and hardware environment are presented in Table IV.

*Table IV Software and hardware environment*

| Num | Property | Parameters |
|---|---|---|
| 1 | Processor | Intel（R）Core 2 |
| 2 | CPU | 2.40GHz |
| 3 | Memory | 2G |
| 4 | OS | Windows Vista |
| 5 | Software | VC++6.0 |

**Experimental process and results:**

*REQ 1: What is the difference of fitting error between the two methods under different data compression ratios?*

A dimensionality reduction approach can be judged by the fitting error between the original sequence and the one after reducing dimension. Under the same data compression ratio, the smaller fitting error means that this dimensionality reduction approach is better.

To address this question, the following definitions are needed:

Data compression ratio: Translate time series $X = (x_1, x_2, \ldots, x_N)$ to a new series $X' = (x'_1, x'_2, \ldots, x'_n)$ (n<N) by reducing dimension, which means that the dimensionality is changed from N to n. Then the data compression ratio is computed by the following formula:

$$\frac{N-n}{N} * 100\% \tag{9}$$

Fitting error: Translate time series $X = (x_1, x_2, \ldots, x_N)$ to a new series $X' = (x'_1, x'_2, \ldots, x'_n)$ (n<N) by reducing dimension. Then restore X' into another sequence $Y' = (y'_1, y'_2, \ldots, y'_N)$ which has the same dimension as X by interpolation. The fitting error between X' andX is defined as follows:

$$\delta = \sum_{i=1}^{N} |x_i - y_i| \tag{10}$$

There are three basic interpolation approaches: linear interpolation, least square interpolation approach and Lagrange interpolation. Linear interpolation is applied to linear data or polynomial function. Hydrological data has big fluctuation and is nonlinear as a whole. So the approach of linear interpolation is not a good choice. The function fitted by least square interpolation is not

necessarily for passing sample points. But hydrological data needs to keep the information of sample points when fitting, which means that this approach should not be applied to hydrological data. Lagrange interpolation requires new function to pass sample points, and has a better effect on nonlinear function than other approaches. Therefore, we use Lagrange interpolation to determine fitting error which is then used to evaluate the dimensionality reduction approach, whose theory is as follows:

T between $T_i$ and $T_{i+1}$ can be calculated by the first three points $T_{i-1}$, $T_i$, $T_{i+1}$, and also by the latter three points $T_i$, $T_{i+1}$, $T_{i+2}$. The interpolation formula for the front three points is shown as equation 11, and the one of latter three points is shown as equation 12:

$$T = \frac{(t-t_i)(t-t_{i+1})}{(t_{i-1}-t_i)(t_{i-1}-t_{i+1})}T_{i-1} + \frac{(t-t_{i-1})(t-t_{i+1})}{(t_i-t_{i-1})(t_i-t_{i+1})}T_i + \frac{(t-t_i)(t-t_{i-1})}{(t_{i+1}-t_i)(t_{i+1}-t_{i-1})}T_{i+1} \quad (11)$$

$$T = \frac{(t-t_{i+1})(t-t_{i+2})}{(t_i-t_{i+1})(t_i-t_{i+2})}T_i + \frac{(t-t_i)(t-t_{i+2})}{(t_{i+1}-t_i)(t_{i+1}-t_{i+2})}T_{i+1} + \frac{(t-t_i)(t-t_{i+1})}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}T_{i+2} \quad (12)$$

To improve the reliability of results, the mean value of interpolating points respectively from the front and latter three points is used for the final interpolating point.

We want to compare the fitting error of the two approaches (ESAX in (Liu, Zhu et al. 2012) and the approach in this paper) by Lagrange interpolation under the same data compression ratio. The following are the specific experimental steps.

*Step 1*: Select the daily water level data of Year 2006.

*Step 2*: Change the parameter in the algorithm of ESAX (Liu, Zhu et al. 2012): the number of segments in PAA. Then we can get feature points under different data compression ratios. Calculate the fitting error between the new sequence based on Lagrange interpolation and the original sequence.

*Step 3*: Change parameters in the algorithm of the new approach: The threshold value *B* which keeps the extreme time period and the number of extreme feature points in a subsequence *N*. Then we can get feature points under different data compression ratios. Calculate the fitting error between the new sequence based on Lagrange interpolation and the original sequence.

*Step 4*: Compare the performance of the two approaches.

*Table V The fitting error of ESAX and new approach*

| Num | ESAX | | | New approach | | |
|---|---|---|---|---|---|---|
| | Dimension after reducing | Data compression ratio | Error of fitting | Dimension after reducing | Data compression ratio | Error of fitting |
| 1 | 274 | 24.9% | 11.767 | 270 | 26.0% | 8.908 |
| 2 | 157 | 57.0% | 13.882 | 158 | 56.7% | 16.096 |
| 3 | 85 | 76.7% | 30.399 | 84 | 76.9% | 23.107 |
| 4 | 42 | 88.5% | 116.457 | 42 | 88.5% | 43.284 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 5 | 36 | 90.1% | 85.822 | 38 | 89.6% | 29.079 |
| 6 | 30 | 91.8% | 109.275 | 29 | 92.0% | 46.876 |

*Figure 1 The comparison of fitting error of ESAX and new approach*

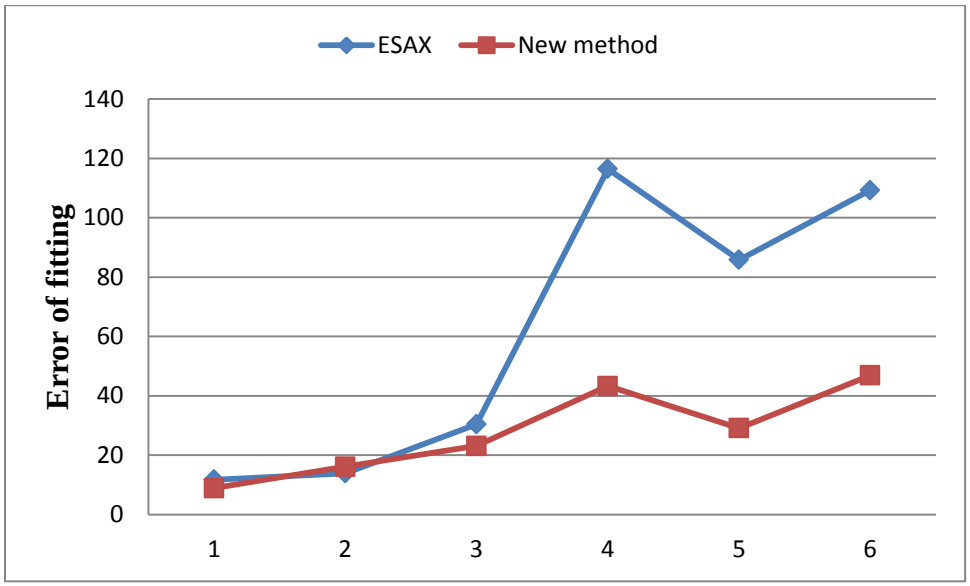



Table V and Figure 1 show that the two fitting errors are almost same when the data compression ratio is relatively small. With the increasing of compression ratios, the fitting error also increases. However, the fitting error of our new approach is always lower than that of ESAX, especially when the data compression ratio is relatively large. Furthermore, the growth rate of fitting error of ESAX is higher than our approach. Therefore, when extracting the feature of time series, our approach is a better choice to obtain important feature and extract key feature points of the original sequence.

*REQ 2: How much execution time is required?*

The execution time is a key factor in selecting a particular algorithm. In this section, we compare the execution time of ESAX and the new approach with different lengths of sequences. The experimental data is the daily water level data of Xiaomeikou gauge station from 1956 to 2005 in Taihu Lake.

The following are the specific experimental steps.

*Step 1*: Select the daily water level data of all the year from 1956 to 2006.

*Step 2*: Choose different lengths of subsequences from the fifty years as experimental data, such as two months, three months, four months and so on.

*Step 3*: Record the execution time of ESAX and our new approach under different lengths of subsequences.

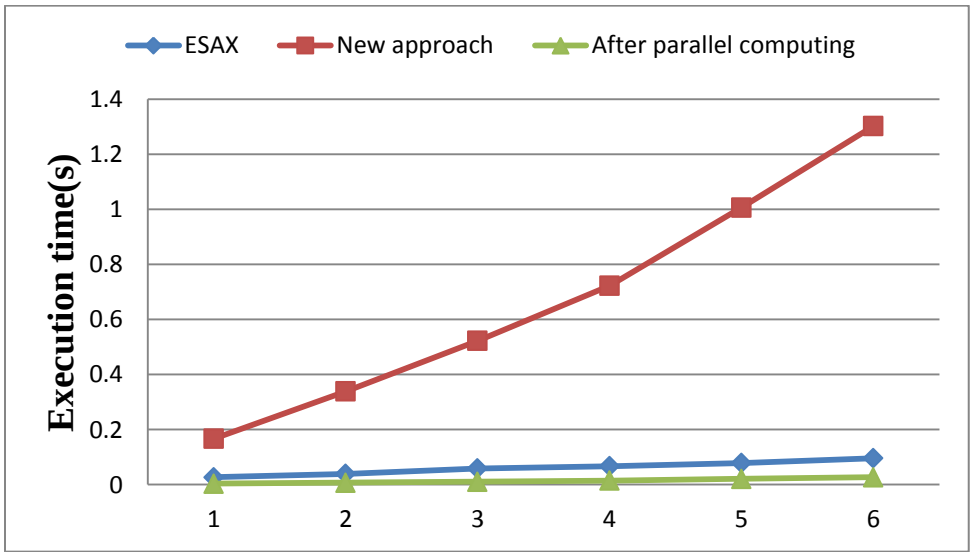
*Step 4*: Repeat Step 2 and 3 for five times, and record the average time of each approach.

*Step 5*: Compare the performance of the two approaches.

*Table VI The execution time of the two approaches*

| Num | Month | Length of subsequence | Total search length | ESAX | New approach | After parallel computing |
|---|---|---|---|---|---|---|
| 1 | 7-8 | 62 | 3100 | 0.027s | 0.167s | 0.004s |
| 2 | 6-8 | 92 | 4600 | 0.039s | 0.339s | 0.007s |
| 3 | 6-9 | 122 | 6100 | 0.059s | 0.523s | 0.011s |
| 4 | 6-10 | 153 | 7650 | 0.067s | 0.723s | 0.015s |
| 5 | 5-10 | 184 | 9200 | 0.079s | 1.007s | 0.021s |
| 6 | 4-10 | 214 | 10700 | 0.096s | 1.303s | 0.027s |

*Figure 2 The comparison of runtime between ESAX, new approach and new approach after parallel computing*




From Table VI and Figure 2, we can see that the execution time of the new approach is obviously higher than ESAX. The main reason is that ESAX adopts Euclidean distance which needs only one for loop to get the distance between strings. However, in our approach, the length of strings after symbolization is different. Consequently we adopt the approach based on DTW to measure precisely. The time complexity of our approach is $O(m * n)$ (*m* and *n* respectively represent the length of two strings X and Y).

To enhance the efficiency, parallel computing is introduced to this scenario to reduce the computational time. Specifically, one resource handles the task that compares the strings of 1956 with the remains, another resource compares 1957 with it's remains, and so on. So, under parallel computing, multiple parts begin to work at the same time, which will save a lot of time in calculating DTW distance. As we can see in Table VI and Figure 2, the execution time after parallel computing is even lower than ESAX.

*REQ 3: What is the accuracy of our approach?*

We use the daily water level data of Xiaomeikou gauge station in Taihu Lake. The data are obtained from 1956 to 2005. We select the parameter values as follows. The subsequence length of first two experiments is about 60. Consequently the threshold value B is designated as 0.05. And we set B to 0.02 in the third experiment whose subsequence length is 150. The number N, which is the number of extreme feature points in subsequence, is usually 4 or 5 and has little effect on the experimental results. The total number of symbols *a* is 5. The top six largest distances (Top_5) are chosen as the results of abnormal mining.

The following are the specific experimental steps.

*Step 1:* Select the daily water level data from 1956 to 2006.

*Step 2*: Choose different lengths of subsequences from the fifty years. In this experiment, we use water level data in May and June, July and August, and June to October from 1956 to 2005.

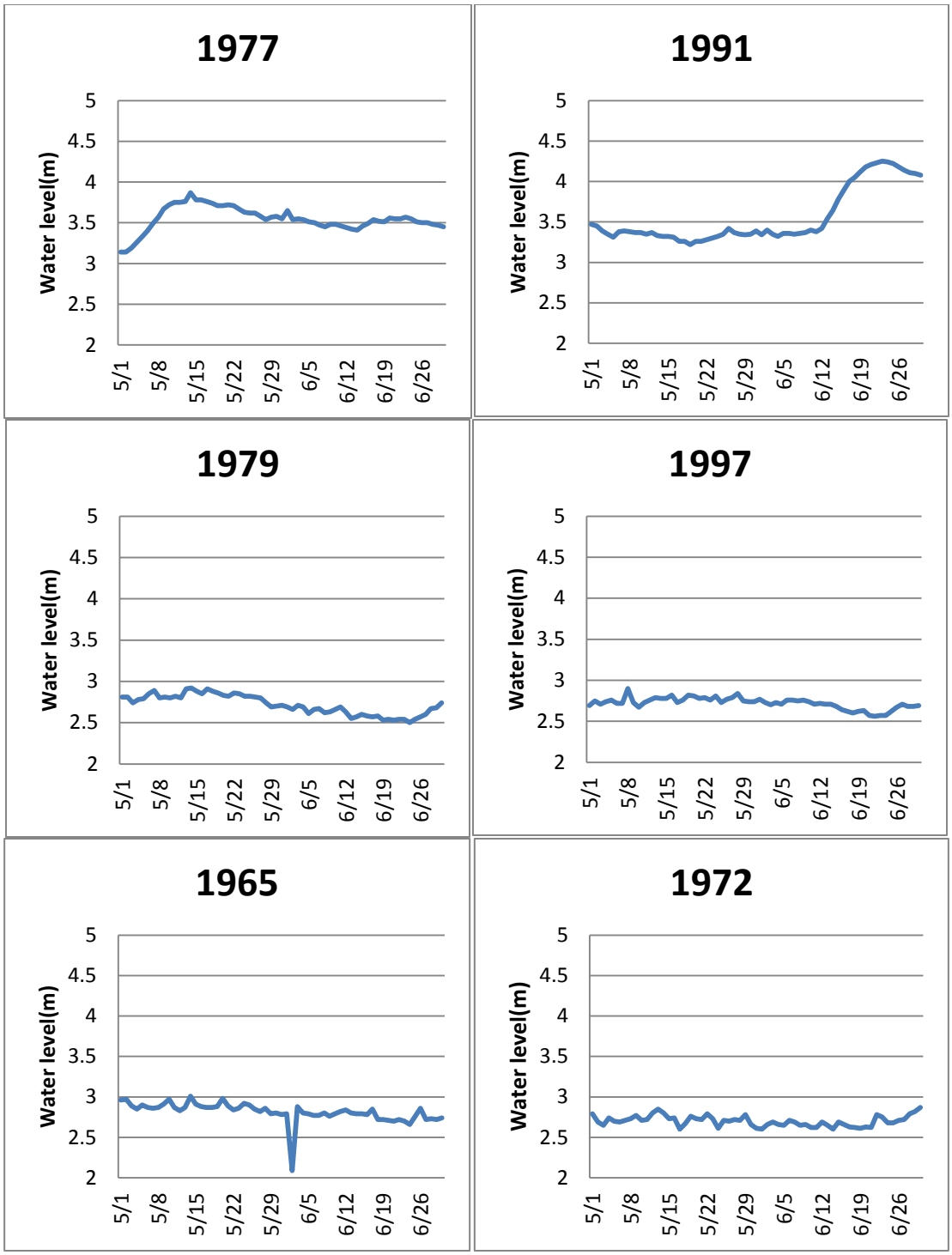
*Step 3*: Get the first six largest distances by the two approaches in each part, which represent the abnormal patterns.

*Step 4*: Compare results of the two approaches.

*Table VII The mining results of the subsequence in May and June*

| Results | ESAX | | New approach | |
|---|---|---|---|---|
| | Year | Distance | Year | Distance |
| 1 | 1977 | 602.91 | 1977 | 513.26 |
| 2 | 1991 | 588.20 | 1991 | 473.11 |
| 3 | 2002 | 560.50 | 1979 | 462.45 |
| 4 | 1960 | 545.75 | 1997 | 458.18 |
| 5 | 1972 | 539.69 | 1965 | 450.29 |
| 6 | 1959 | 536.20 | 1973 | 443.77 |

*Figure 3 The water level in May and June*



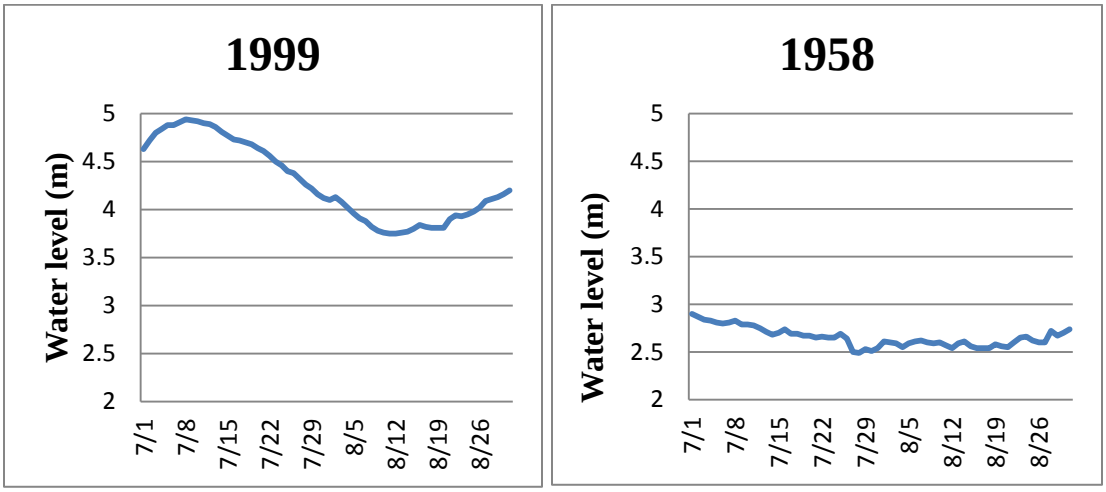

(1) Sequence abnormal pattern mining results of May and June are shown in Table VII. Firstly, according to hydrological characteristics and rules of season change in Taihu Lake, the water level in May is relatively stable and rises in June. As we can see in Figure 3, 1979, 1997, 1972 and so on do not conform to this law. These abnormal patterns are not very obvious, but we should not ignore them. However, there is an abnormal pattern in which the water level firstly rose then fell in May and June in 1977. Secondly, the water level of June in 1991 rose rapidly. Furthermore, the increasing range and speed is higher than any other year, which is also an

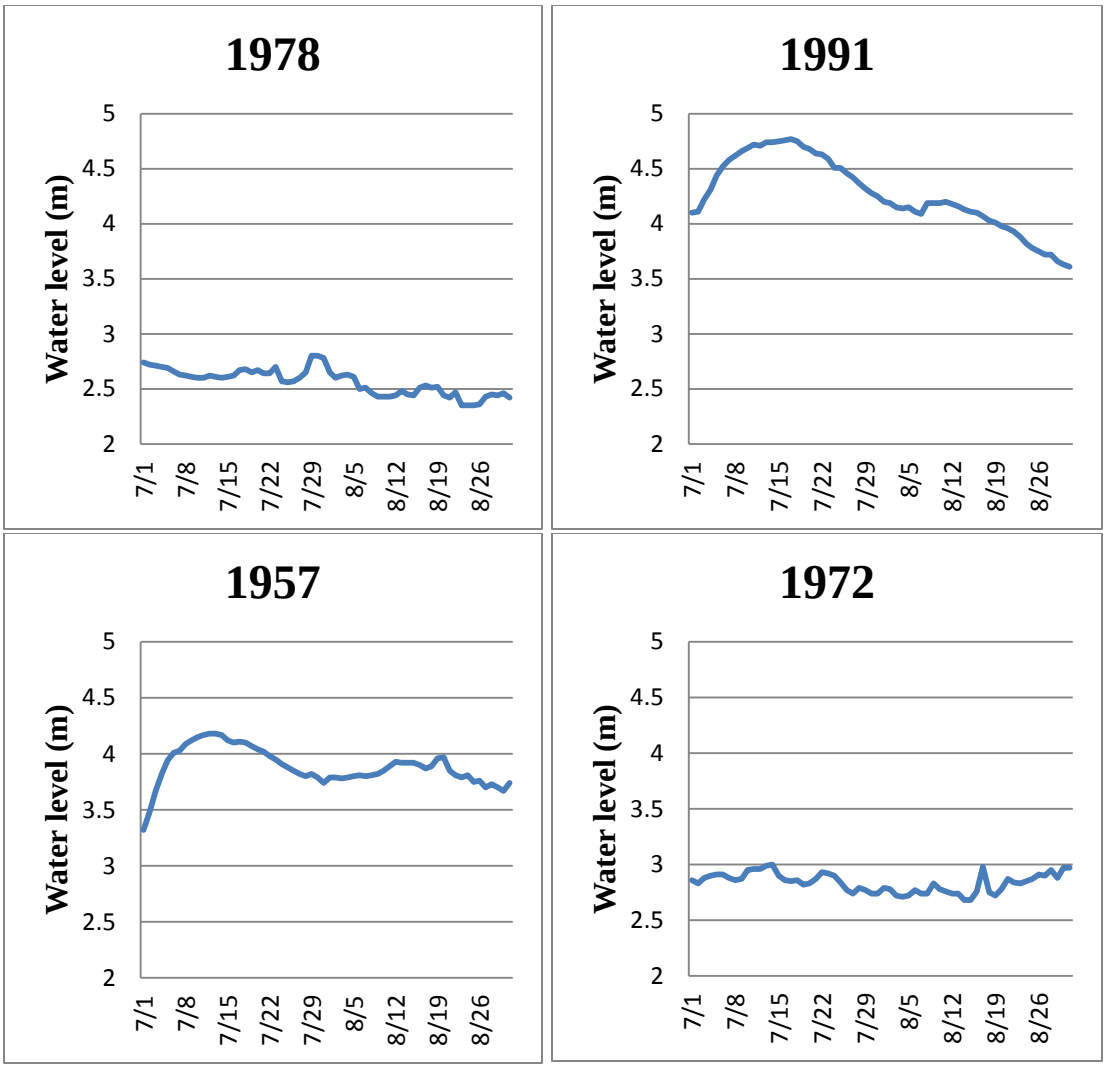
abnormal pattern. The above two abnormal patterns can be extracted by both approaches. Finally, there is an obvious outlier which is extracted by the approach proposed by us but ignored by ESAX. As we can see in Figure 3, the water level of May and June has a glaring trough in 1965: the height of water is 2.04 meters on the 2nd of June which has a big fall compared with adjacent days, which is a conspicuous outlier.

*Table VIII The mining results of the subsequence in July and August*

| Results | ESAX | | New approach | |
|---|---|---|---|---|
| | Year | Distance | Year | Distance |
| 1 | 1991 | 1044.74 | 1999 | 631.31 |
| 2 | 1999 | 1044.74 | 1958 | 539.58 |
| 3 | 1957 | 1015.58 | 1978 | 539.58 |
| 4 | 1987 | 919.04 | 1991 | 479.45 |
| 5 | 1958 | 914.77 | 1957 | 464.03 |
| 6 | 1978 | 914.77 | 1972 | 406.71 |

*Figure 4 The water level in July and August*





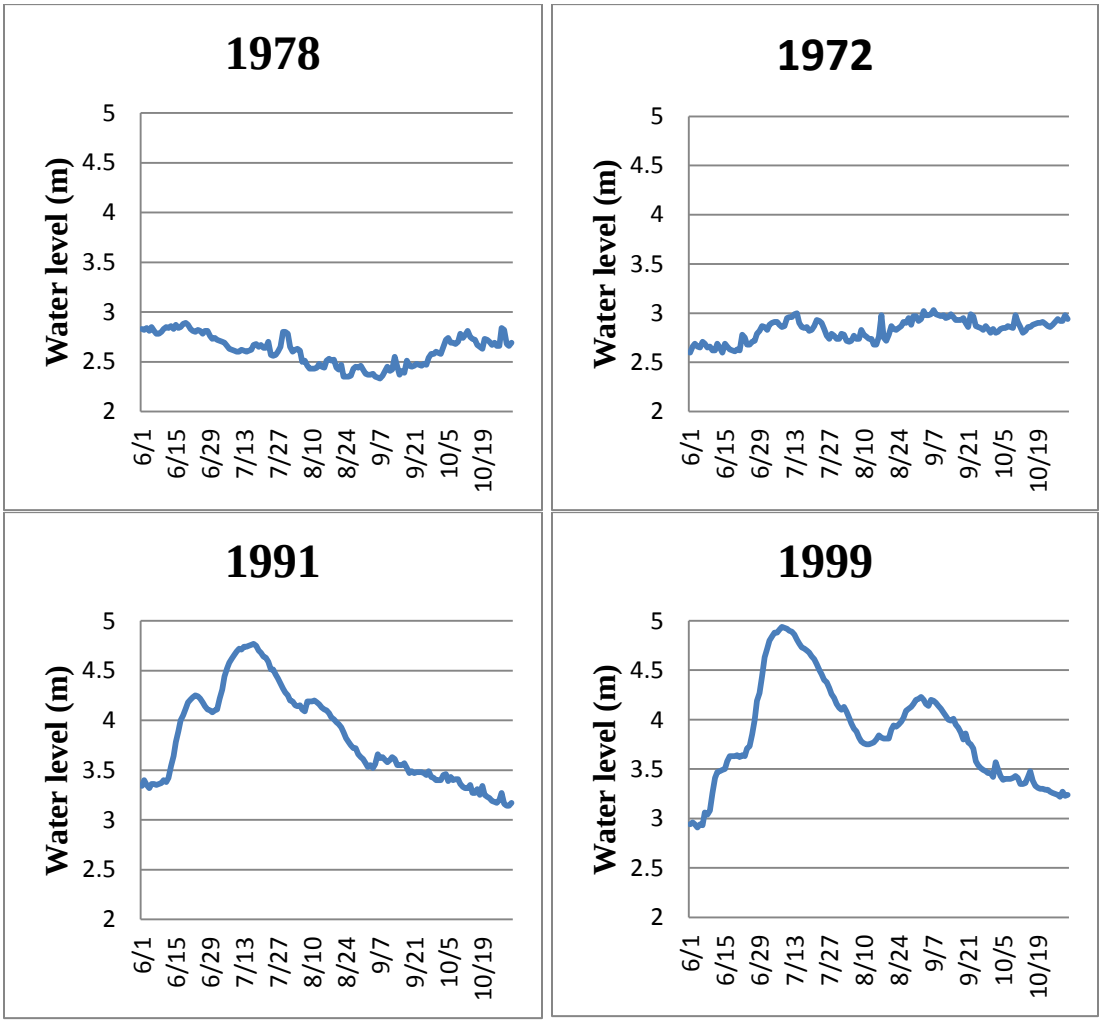

(2) The results of July and August are shown in Table VIII. In 1991 and 1999, the water levels of July are nearly five meters and the average water level in July and August is relatively high. But the theoretical average water level of July and August in the flood season is 3.5 meters. The Taihu Lake basin really suffered a once-in-a-century flood respectively in 1991 and 1999. While the flood of 1957 is less violent than 1991 and 1999, the water level in July and August is higher than the average level, sometimes even higher than 4 meters. Besides the flood, the lake basin also suffered a serious drought in 1978. The water level of July and August during the flood season is no more than 3 meters this year. And the water level even rises but not falls in the short time just as in Figure 4, which is obviously abnormal. In addition, the water level of 1958 and 1972 remains below 3 meters. There is a drop of the water level from 1th to 29th in July in 1958, which is topically abnormal. As we can see, the results of both approaches are almost the same besides the different ranking. And both results are accurate.

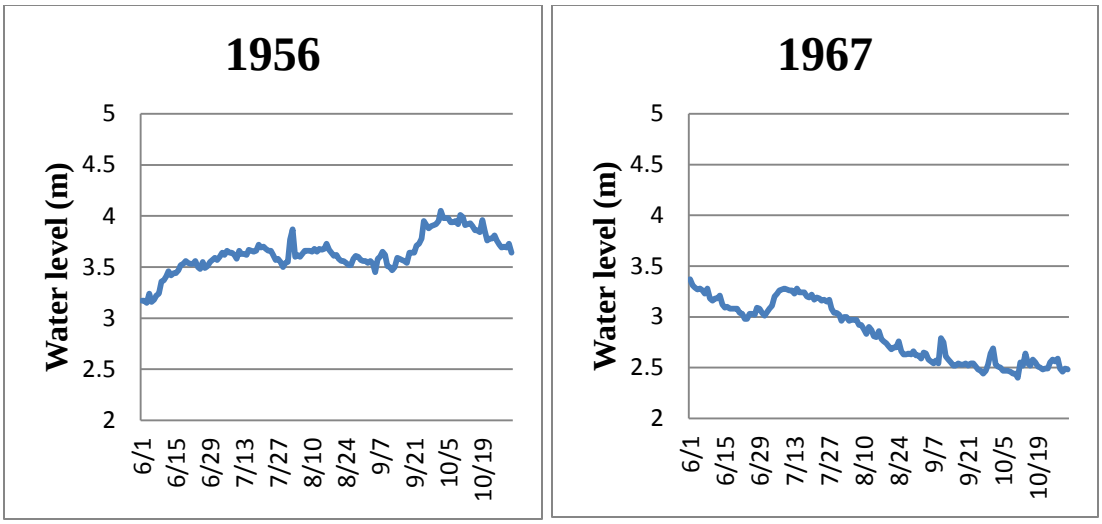
*Table IX The mining results of the subsequence from June to October*

| Results | ESAX | | New approach | |
|---|---|---|---|---|
| | Year | Distance | Year | Distance |

| | | | | |
|---|---|---|---|---|
| 1 | 1978 | 2324.62 | 1978 | 1256.45 |
| 2 | 1972 | 2023.10 | 1972 | 654.65 |
| 3 | 1999 | 1947.34 | 1991 | 590.28 |
| 4 | 1991 | 1931.17 | 1999 | 581.36 |
| 5 | 1956 | 1869.95 | 1956 | 569.49 |
| 6 | 1967 | 1812.86 | 1967 | 478.08 |

*Figure 5 The water level from June to October*






(3) The results of June to October are shown in Table IX. The subsequence in this part is the longest. The result shows that the two approaches both have good performance for the long subsequence. Firstly, in Figure 5, the trend of water level is relatively flat in 1978 and 1972. The water level is always lower than 3 meters. There are not obvious ups and downs even in the flood season of July and August, which is an abnormal pattern. Second is the special situation of water level in 1967. The water level around 11th of July falls constantly, which should rise according to history. Lastly, the water levels of 1999 and 1991 grow rapidly. Extraordinary flood happened in these two years. Except the difference of the ranking of 1991 and 1999, the results of ESAX and our approach are same. However, in our approach, the distance of 1978 is longer than any other years, which may represent significant knowledge. Perhaps it can be explained by some hydrological experts.

As we can see from the above three experiments, in most cases, mining results of ESAX and our approach remain the same. When most points of the subsequence are in abnormal states, the abnormal patterns can be extracted by both approaches, such as 1991 and 1999 which suffered massive flooding. However, our approach extracts an obvious anomaly ignored by ESAX-- the water level on 2nd of June in 1965 which has been verified in the above experiment of May and June. Therefore, both approaches can mine abnormal patterns which account for more time of subsequences. For shorter time, the mining ability of our approach is better than ESAX. Our approach can dig up a variety of different types of abnormal data, which helps in the study of hydrological phenomenon.

**Threats to validity:** Although the experimental results reveal the accuracy and usability of our approach, there are still some threats to validity.

Firstly, the choice of *N*, which is the number of extreme feature points of each subsequence, and the threshold value *B* have a high impact on the accuracy of our approach. They should be based on abundant experiments. Consequently, a wrong choice may lead to the failure of approach.

Secondly, our approach can only mine anomaly currently. In other words, it is unable to distinguish the different types of anomaly (such as flood or drought).

# Conclusions and suggestions for future work

This paper studies the problem of the abnormal hydrological time sequence mining. Combining with the field of hydrology, we explore the effective and accurate approach of abnormal pattern mining. At present, the abnormal time pattern mining, especially in the field of hydrology, is based on distance measurement. These approaches need to calculate the distance between each pattern with high time complexity. In this work, we combine symbolization (FP_SAX) of time series with distance measurement (SD_DTW). Then a new approach that is suitable for feature extraction of hydrological time series and can identify abnormal model quickly is developed, which provides accurate and efficient mining results.

The study of time series anomaly mining and similarity search of hydrological data is still in the starting state. Although some efficient techniques have been proposed, a complete system of hydrology and hydroinformatics research hasn't been formed. Every approach has certain limitations. The implementation process of some approaches is very complicated. Even for those verified in theory, they have poor applicability. Therefore, more work remain in the future.

1. Symbolic approach has good results on the dimension reduction, which can also keep main features of the series. After symbolization, feature research can improve existing string matching approaches for subsequent excavation. Then text data mining can be used to mine time series anomaly.

2. In FP_SAX algorithm, the minimum value of N, which is the number of extreme feature points of each subsequence, is 1, and the maximum value is the number of total extreme feature points. This paper estimates the value of N based on the experience of previous experiments. In the future we should consider a more scientific approach to determine the optimal value of N.

3. How to make the classification result of subsequence with different length to be more suitable with feature extraction is also an interesting question. It is a good idea to classify time series based on some hydrological parameters and hydrology laws.

4. Due to the use of DTW, our approach has higher time complexity compared to ESAX. Although after parallel computing, it is obviously lower than ESAX. However, it requires more computer resouces. New techniques for similarity search are needed to ensure the accuracy of mining results and reduce the time complexity as well.

5. Cluster analysis can be used to distinguish the different types of anomaly based on the approach in this paper, which may make a better contribution to hydrological forecasting.

# ACKNOWLEGMENT

The work is supported by the National Natural Science Foundation of China under Grant (Nos. 61370091, 61202097 and 61202136), and Doctoral Fund of Ministry of Education of China (Grant No.20120094120009).

## REFERENCES

Agrawal, R., Faloutsos, C., & Swami, A. (1993). Efficient similarity search in sequence databases (pp. 69-84). Springer Berlin Heidelberg.

Bigioi, P., Ciuc, M., Ciurel, S., Corcoran, P., Prilutsky, Y., Steinberg, E., & Vertran, C. (2009). U.S. Patent No. 7,564,994. Washington, DC: U.S. Patent and Trademark Office.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2013). Time series analysis: forecasting and control. John Wiley & Sons.

Chaovalit, P., Gangopadhyay, A., Karabatis, G., & Chen, Z. (2011). Discrete wavelet transform-based time series analysis and mining. ACM Computing Surveys (CSUR), 43(2), 6.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic acids research, 30(7), 1575-1584.

Esling, P., & Agon, C. (2012). Time-series data mining. ACM Computing Surveys (CSUR), 45(1), 12.

Fu, T. C. (2011). A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1), 164-181.

Gómez-Verdejo, V., Arenas-García, J., Lázaro-Gredilla, M., & Navia-Vazquez, A. (2011). Adaptive one-class support vector machine. Signal Processing, IEEE Transactions on, 59(6), 2975-2981.

Henry, E. R., & Hofrichter, J. (1992). [8] Singular value decomposition: Application to analysis of experimental data. Methods in enzymology, 210, 129-192.

Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems, 3(3), 263-286.

Keogh, E. J., & Pazzani, M. J. (1998, August). An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback. In KDD (Vol. 98, pp. 239-243).

Kozerski, H. P. (2010). similarity search and pattern discovery in hydrological time series data mining. Hydrological Processes.

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003, June). A symbolic representation of time series, with implications for streaming algorithms. In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery (pp. 2-11). ACM.

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery, 15(2), 107-144.

Lin, K. P., & Chen, M. S. (2011). On the design and analysis of the privacy-preserving SVM classifier. Knowledge and Data Engineering, IEEE Transactions on, 23(11), 1704-1717.

Liu, Q., Zhu, Y. L., & Zhang, P. C. (2012). Extended symbolic aggregate approximation based anomaly mining of hydrological time series. Jisuanji Yingyong Yanjiu, 29(12).

Lkhagva, B., Suzuki, Y., & Kawagoe, K. (2006, April). New Time Series Data Representation ESAX for Financial Applications. In ICDE Workshops (p. 115).

Müller, M. (2007). Dynamic time warping. Information retrieval for music and motion, 69-84.

Ouyang, R. L., Ren, L. L., & Zhou, C. H. (2010). Similarity search in hydrological time series. Journal of Hohai University: Natural Sciences, 38(3), 242-245.

Papapetrou, P., Athitsos, V., Potamias, M., Kollios, G., & Gunopulos, D. (2011). Embedding-based subsequence matching in time-series databases. ACM Transactions on Database Systems (TODS), 36(3), 17.

Ping, A. (2003). Ni Weixin 21 (School of Computer&Information Engineering, Hohai University, Nanjing210098) 2 (Bureau of Hydrology/Water Resources Information Center, MWR, Beijing100053); Review and Preview of the Research on Hydrological Data Mining Technology in China [J]. Computer Engineering and Applications, 28.

Russ, J. C., & Woods, R. P. (1995). The image processing handbook. Journal of Computer Assisted Tomography, 19(6), 979-981.

Starck, J. L., Murtagh, F., & Fadili, J. M. (2010). Sparse image and signal processing: wavelets, curvelets, morphological diversity. Cambridge University Press.

Xiao, H., & Hu, Y. (2005). Data mining based on segmented time warping distance in time series database. Jisuanji Yanjiu yu Fazhan(Comput. Res. Dev.), 42(1), 72-78.

Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. Communications Surveys & Tutorials, IEEE, 12(2), 159-170.

## ABOUT THE AUTHOR(S)

**Pengcheng Zhang** received the Ph.D. degree in computer science from Southeast University in 2010. He is currently an associate professor in College of Computer and Information Engineering, Hohai University, Nanjing, China. His research interests include modeling, analysis, testing and verification of component based systems, software architectures, real-time and probabilistic systems, service-oriented systems.

**Xiao Yan** is a master in the College of Computer and Information, Hohai University, Nanjing, China. She received her bachelor degree in Electronic Information from Hohai University, Nanjing, China in 2013. Her current research interests include data mining and hydroinformatics.

**Wenrui Li** is an associate professor in the School of Mathematics & Information Technology at Nanjing Xiaozhuang University. She received Ph.D. degree in College of Computer and Information, Hohai University, in 2010. Her current research interests include service computing; cloud computing, software modeling and verification.

**Hareton Leung** joined Hong Kong Polytechnic University in 1994 and is now director of the Lab for Software Development and Management. He serves on the Editorial Board of Software Quality Journal. He is a fellow of Hong Kong Computer Society, chairperson of its Quality Management Division (QMSID) and chairperson of HKSPIN. He previously held team leader positions at BNR, Nortel, and GeneralSoft Ltd. He is also an accomplished industry consultant, giving advice on software testing, quality assurance, process and quality improvement, system development, and providing expert witness and litigation support.