Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 105128, 13 pages
http://dx.doi.org/10.1155/2015/105128

Research Article

Chebyshev Similarity Match between Uncertain Time Series

Wei Wang,[1] Guohua Liu,[2] and Dingjia Liu[3]

[1] *School of Information Science and Technology, Donghua University, Shanghai 201620, China*
[2] *School of Computer Science and Technology, Donghua University, Shanghai 201620, China*
[3] *School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China*

Correspondence should be addressed to Wei Wang; rid3008@qq.com

Received 27 April 2015; Revised 11 June 2015; Accepted 25 June 2015

Academic Editor: Hamed O. Ghaffari

In real application scenarios, the inherent impreciseness of sensor readings, the intentional perturbation of privacy-preserving transformations, and error-prone mining algorithms cause much uncertainty of time series data. The uncertainty brings serious challenges for the similarity measurement of time series. In this paper, we first propose a model of uncertain time series inspired by Chebyshev inequality. It estimates possible sample value range and central tendency range in terms of sample estimation interval and central tendency estimation interval, respectively, at each time slot. In comparison with traditional models adopting repeated measurements and random variable, Chebyshev model reduces overall computational cost and requires no prior knowledge. We convert Chebyshev uncertain time series into certain time series matrix; therefore noise reduction and dimensionality reduction are available for uncertain time series. Secondly, we propose a new similarity matching method based on Chebyshev model. It depends on overlaps between two sample estimation intervals and overlaps between central tendency estimation intervals from different uncertain time series. At the end of this paper, we conduct an extensive experiment and analyze the results by comparing with prior works.

1. Introduction

Over the past decade, a large amount of continuous sensor data was collected in many applications, such as logistics management, traffic flow management, astronomy, and remote sensing. In most cases, these applications organize the sequential sensor readings into time series, that is, sequences of data points ordered by temporal dimension. The problem of processing and mining time series with incomplete, imprecise, and even error-prone measurements is of major concern in recent studies [1–6]. Typically, uncertainty occurs due to the impreciseness of equipment and methods during physical data collection period. For example, the inaccuracy of a wireless temperature sensor follows a certain error distribution. In addition, intentional deviation brought by privacy-preserving transformation also causes much uncertainty. For example, the real time location information of some VIP may be perturbed [7, 8].

Managing and processing uncertain data were studied in the traditional database area during the 80s [9] and have been borrowed in the investigation of uncertain time series in recent years. Two widely adopted methods are introduced in modeling uncertain time series. First, a probability density function (pdf) over the uncertain values represented by a random variable is estimated in accord with a priori knowledge, among which the hypotheses of Normal distribution are ubiquitous [10–12]; however, the hypotheses of Normal distribution are quite limited in many applications; the uncertain time series data with Uniform or Exponential distribution is frequently found in some other applications, for example, Monte Carlo simulation of power load and evaluation of reliability of electronic components [13, 14]. Second, the unknown data distribution is summarized by repeated measurements (i.e., sample or observations) [15]; the accurate estimation of data distribution is obtained by large amount of repeated measurements; however, it causes high computational cost and more storage space.

In this paper, we propose a new model for uncertain time series by combining the two methods above and use descriptive statistics (i.e., central tendency) to resolve the uncertainty. On this basis, we present an effective matching method to measure the similarity between two uncertain time series, which is adaptive to distinct error distributions. Our model estimates the sample value range and the central tendency range derived from Chebyshev inequality, extracting the sample estimation interval and central tendency estimation interval drawn from repetitive measurements at each time slot. Unlike traditional similarity matching methods of uncertain time series based on the measurement of distance, we adopt the overlap between sample estimation intervals and that between central tendency estimation intervals to evaluate similarity. If both estimation intervals from two uncertain time series at corresponding time slot have a chance of being equal, the extent of similarity is larger as compared to the case in which they never be the same.

The rest of this paper is organized as follows. In Section 3 we propose the model of Chebyshev uncertain time series. Section 4 is on the preprocessing of uncertain time series based on Chebyshev model. Section 5 describes the process of similarity match with new method. Section 6 addresses the experiments. At last, Section 7 draws a conclusion.

To sum up, we list our contributions as follows:

(i) We propose a new model of uncertain time series based on sample estimation interval and central tendency estimation interval derived from Chebyshev inequality and convert Chebyshev uncertain time series into certain time series matrix for dimensionality reduction and noise reduction.

(ii) We present an effective method to measure the similarity between two uncertain time series within distinct error distributions without a priori knowledge.

(iii) We conduct extensive experiments and demonstrate the effectiveness and efficiency of our new method in similarity matching between two uncertain time series.

2. Related Work

The problem of similarity matching for certain time series has been extensively studied over the past decade; however the similar problem arises for uncertain time series. Aßfalg et al. first propose a probabilistic bounded range query (PBRQ) [15]. Formally, let D be a set of uncertain time series and let T_u be an uncertain time series as query input; let ϵ be a distance bound and let τ be a probability threshold. The $\mathrm{PBRQ}_{\epsilon,\tau}(T_u, D)$ is given by

$$\mathrm{PBRQ}_{\epsilon,\tau}(T_u, D) = \left\{T'_u \in D \mid P_r\left(\mathrm{DIST}\left(T_u, T'_u\right) \le \epsilon\right) \ge \tau\right\}. \quad (1)$$

Dallachiesa et al. proposed the method called MUNICH [16]; the uncertainty is represented by means of repeated observations at each time slot [15]. An uncertain time series is a set of certain time series in which each certain time series is constructed by choosing one sample observation for each time slot. The distance between two uncertain time series is defined as the set of distances between all combinations from one certain time series set to the other. Notice that the distance measures adopted by MUNICH are based on L_p-norm and DTW distances; if $p = 2$, the L_p-norm is Euclidean distance; the naive computation of the result set is not practical. Large result space causes exponential computational cost.

PROUD [12] processes similarity queries over uncertain time streams. It employs the Euclidean distance and models the similarity measurement as the sum of the differences of time series random variables. Each random variable represents the uncertainty of the value of corresponding time slot. The standard deviation of the uncertainty and a single observation for each time slot are prerequisites for modeling uncertain time series. Sarangi and Murthy propose a new distance measurement DUST. It is derived from the Euclidean distance and under the assumption that all time series values follow some specific distribution [11]. If the error of the time series values at different time slot follows Normal distribution, DUST is equivalent to the weighted Euclidean distance. Compared to the MUNICH, it does not need multiple observations and thus is more efficient. Inspired by the moving average, Dallachiesa et al. propose a simple similarity measurement that previous studies had not considered; it adopts Uncertain Moving Average (UMA) and Uncertain Exponential Moving Average (UEMA) filters to solve the uncertainty from time series data [16]. Although the experimental results show that they outperform the sophisticated techniques that have been proposed above, a priori knowledge of the error standard deviation is indispensable.

Most of the above techniques are based on the assumption that the values of time series are independent of one another. Obviously, this assumption is a simplification. Adjacent values in time series are correlated to a certain extent. The effect of correlations is studied in [16] and the research shows that there is a great benefit if the correlations are taken into account. Likewise, we implicitly embed correlations into estimation intervals in terms of repetitive observation values, adopting the degree of overlap to evaluate the similarity of uncertain time series. Our approach reduces overall computational cost and outperforms the existing methods on accuracy; new model requires no prior knowledge and makes dimensionality reduction available for uncertain time series.

3. Chebyshev Uncertain Time Series Modeling

As shown in [15], let $T = (X_1, X_2, \ldots, X_n)$ be an uncertain time series of length n; $X_t \in T$ is a random variable represented by a set $X_t = \{v_{t,1}, v_{t,2}, \ldots, v_{t,s}\}$ of s measurements (i.e., random sample observations), $v_{t,i} \in \mathfrak{R}^d$. s is denoted as sample size of T. Distribution of the points in X_t is the uncertainty at time slot t. The larger sample size s is, the more accurate data distribution is estimated. However computational cost is prohibitive. To solve the problem, we present a new model for uncertain time series by considering Chebyshev's inequality below.

Lemma 1. *Let X (integrable) be a random variable with finite expected value $E(X) = \mu$ and finite nonzero variance*

$D(X) = \sigma^2$. *Then, for any real number* $\varepsilon > 0$,

$$P\{|X-\mu| \le \varepsilon\} \ge 1 - \frac{\sigma^2}{\varepsilon^2}. \tag{2}$$

Formula (2) (Chebyshev's inequality) [17] is the lower bound of probability of $\{|X - E(X)| \le \varepsilon\}$; on condition that μ and σ^2 are known, the distribution information need not be considered. Real number ε has an important influence on the determination of the lower bound. For an appropriate ε, the probability of possible values of random variable falling in the boundaries satisfies desired threshold. The estimation of possible value range is as follows.

Theorem 2. *Given a random variable X with the finite expected value $E(X) = \mu$ and finite nonzero variance $D(X) = \sigma^2$, if the ε in inequality (2) equals 4σ, then*

$$P\{X \in [\mu - 4\sigma, \mu + 4\sigma]\} \ge 0.9375 \tag{3}$$

no matter which probability distribution X obeys.

Proof. Consider

$$\begin{aligned} P\{|X-\mu| \le \varepsilon\} &\ge 1 - \frac{\sigma^2}{\varepsilon^2} \\ &\Longrightarrow P\{|X-\mu| \le 4\sigma\} \ge 1 - \frac{\sigma^2}{(4\sigma)^2} \\ &\Longrightarrow P\{|X-\mu| \le 4\sigma\} \ge 1 - \frac{1}{16} \\ &= 0.9375 \\ &\Longrightarrow P\{X \in [\mu - 4\sigma, \mu + 4\sigma]\} \\ &\ge 0.9375. \end{aligned} \tag{4}$$

□

The above proof shows that when ε equals 4σ, the probability of X within interval $[\mu - 4\sigma, \mu + 4\sigma]$ exceeds 0.9; nearly all possible measurements fall in the interval. We substitute the random variable X with $[\mu - 4\sigma, \mu + 4\sigma]$ to express the uncertainty.

According to the probability distribution of X, possible value range description of uncertainty is insufficient; a central or typical value is another feature for a probability distribution; it indicates a center or location of the distribution, called central tendency [18]. The most common measure of central tendency is arithmetic mean (mean for short), so the central tendency of a random sample set C in form of mean $\overline{X}_C$ is defined below.

Given a random sample set C drawn from X with μ and σ^2, $C = \{X_{c1}, X_{c2}, \ldots, X_{cm}\}$, each sample satisfies *i.i.d.* hypothesis; then

$$\overline{X}_C = \frac{1}{m}\sum_{i=1}^{m} X_{ci}. \tag{5}$$

As a random variable, the expectation $E(\overline{X}_C)$ and variance $D(\overline{X}_C)$ are evaluated below:

$$E\left(\overline{X}_C\right) = E\left(\frac{1}{m}\sum_{i=1}^{m} X_{ci}\right) = \frac{1}{m}\cdot m\mu = \mu, \tag{6}$$

$$\begin{aligned} D\left(\overline{X}_C\right) &= D\left(\frac{1}{m}\sum_{i=1}^{m} X_{ci}\right) = \frac{1}{m^2} D\left(\sum_{i=1}^{m} X_{ci}\right) \\ &= \frac{1}{m^2}\sum_{i=1}^{m} D\left(X_{ci}\right) = \frac{1}{m^2}\cdot m\sigma^2 = \frac{\sigma^2}{m}. \end{aligned} \tag{7}$$

Analogously, for central tendency variable $\overline{X}_C$, in accord with Lemma 1, the corresponding estimation interval can be obtained.

Theorem 3. *Given a random variable X with μ and σ^2, a random sample set $C = \{X_{c1}, X_{c2}, \ldots, X_{cm}\}$ drawn from the population of X, for the variable $\overline{X}_C$ with μ and σ^2/m, if the ε in inequality (2) equals $4\sigma/\sqrt{m}$, then*

$$P\left\{\overline{X}_C \in \left[\mu - \frac{4\sigma}{\sqrt{m}}, \mu + \frac{4\sigma}{\sqrt{m}}\right]\right\} \ge 0.9375. \tag{8}$$

Proof. Consider

$$\begin{aligned} P\{|\overline{X}_C - \mu| \le \varepsilon\} &\ge 1 - \frac{\sigma^2/m}{\varepsilon^2} \\ &\Longrightarrow P\left\{|\overline{X}_C - \mu| \le \frac{4\sigma}{\sqrt{m}}\right\} \\ &\ge 1 - \frac{\sigma^2/m}{(4\sigma/\sqrt{m})^2} \\ &\Longrightarrow P\left\{|\overline{X}_C - \mu| \le \frac{4\sigma}{\sqrt{m}}\right\} \ge 1 - \frac{1}{16} \\ &= 0.9375 \\ &\Longrightarrow P\left\{\overline{X}_C \in \left[\mu - \frac{4\sigma}{\sqrt{m}}, \mu + \frac{4\sigma}{\sqrt{m}}\right]\right\} \\ &\ge 0.9375. \end{aligned} \tag{9}$$

□

In summary, the sample estimation interval $[\mu - 4\sigma, \mu + 4\sigma]$ of X is the range of possible measurements and central tendency estimation interval $[\mu - 4\sigma/\sqrt{m}, \mu + 4\sigma/\sqrt{m}]$ is the range of central tendency of X. The uncertainty of X is represented by a combination of the two intervals at each time slot. Uncertain time series can be defined below.

Definition 4. For an uncertain time series $T = (X_1, X_2, \ldots, X_n)$ of length n, each element X_t is a random variable with μ_t and ${\sigma_t}^2$, $\overline{X}_{C_t}$ is the central tendency of random sample set C_t from the population corresponding to X_t, and an Chebyshev uncertain time series T_{Che} is defined below:

$$\begin{aligned} T_{\text{Che}} = \Bigg(\Bigg(&[\mu_1 - 4\sigma_1, \mu_1 + 4\sigma_1], \\ &\left[\mu_1 - \frac{4\sigma_1}{\sqrt{m}}, \mu_1 + \frac{4\sigma_1}{\sqrt{m}}\right], t_1\Bigg), \end{aligned}$$

$$\left([\mu_2 - 4\sigma_2, \mu_2 + 4\sigma_2], \left[\mu_2 - \frac{4\sigma_2}{\sqrt{m}}, \mu_2 + \frac{4\sigma_2}{\sqrt{m}}\right], t_2\right), \dots, \left([\mu_n - 4\sigma_n, \mu_n + 4\sigma_n], \left[\mu_n - \frac{4\sigma_n}{\sqrt{m}}, \mu_n + \frac{4\sigma_n}{\sqrt{m}}\right], t_n\right)\Bigg), \tag{10}$$

where m is the cardinality of random sample set C. Consider the Chebyshev uncertain time series above; μ_t and σ_t are difficult to be obtained because of the unidentified distribution of population. We choose two statistics to estimate the μ_t and σ_t; one is the arithmetic mean of C, mentioned in (5); the other is the sample standard deviation S_C, calculated by the following equation:

$$S_C = \sqrt{{S_C}^2} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(X_{ci} - \overline{X}_C\right)^2}, \tag{11}$$

$$\begin{aligned} E\left({S_C}^2\right) &= E\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(X_{ci} - \overline{X}_C\right)^2\right] \\ &= E\left[\frac{1}{m-1}\left(\sum_{i=1}^{m} X_{ci}^2 - m{\overline{X}_C}^2\right)\right] \\ &= \frac{1}{m-1}\left[\sum_{i=1}^{m} E\left(X_{ci}^2\right) - mE\left({\overline{X}_C}^2\right)\right] \\ &= \frac{1}{m-1}\left[\sum_{i=1}^{m}\left(\mu^2 + \sigma^2\right) - m\left(\frac{\sigma^2}{m} + \mu^2\right)\right] \\ &= \sigma^2. \end{aligned} \tag{12}$$

Equations (12) and (6) show that $\overline{X}_C$ and S_C are unbiased estimator for μ and σ. μ_t and σ_t in Definition 4 can be replaced with $\overline{X}_{C_t}$ and S_{C_t}; T_{Che} is rewritten as follows.

Definition 5. Given a sample set $C_t = \{X_{c1}^t, X_{c2}^t, \dots, X_{cm}^t\}$ at time slot t, T_{Che} is represented as follows:

$$\begin{aligned} &\Bigg(\left(\left[\overline{X}_{C_1} - 4S_{C_1}, \overline{X}_{C_1} + 4S_{C_1}\right], \left[\overline{X}_{C_1} - \frac{4S_{C_1}}{\sqrt{m}}, \overline{X}_{C_1} + \frac{4S_{C_1}}{\sqrt{m}}\right], t_1\right), \\ &\left(\left[\overline{X}_{C_2} - 4S_{C_2}, \overline{X}_{C_2} + 4S_{C_2}\right], \left[\overline{X}_{C_2} - \frac{4S_{C_2}}{\sqrt{m}}, \overline{X}_{C_2} + \frac{4S_{C_2}}{\sqrt{m}}\right], t_2\right), \dots, \\ &\left(\left[\overline{X}_{C_n} - 4S_{C_n}, \overline{X}_{C_n} + 4S_{C_n}\right], \left[\overline{X}_{C_n} - \frac{4S_{C_n}}{\sqrt{m}}, \overline{X}_{C_n} + \frac{4S_{C_n}}{\sqrt{m}}\right], t_n\right)\Bigg). \end{aligned} \tag{13}$$

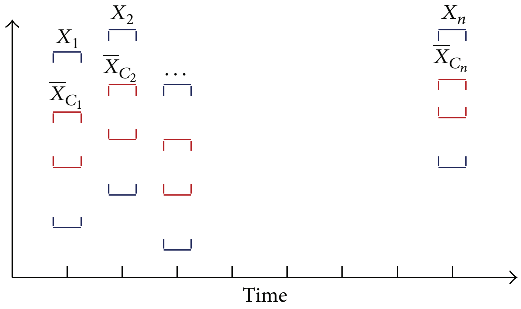


Figure 1: The Chebyshev uncertain time series model.

According to the descriptions above, the expression at each time slot can be transformed into a vector. It consists of four elements (except time value), namely, $\overline{X}_{C_t} - 4S_{C_t}$, $\overline{X}_{C_t} - (4S_{C_t}/\sqrt{m})$, $\overline{X}_{C_t} + (4S_{C_t}/\sqrt{m})$, and $\overline{X}_{C_t} + 4S_{C_t}$, in ascending order, denoted as $\mathbf{v}_t$; consider

$$\mathbf{v}_t = \left(\overline{X}_{C_t} - 4S_{C_t}, \overline{X}_{C_t} - \frac{4S_{C_t}}{\sqrt{m}}, \overline{X}_{C_t} + \frac{4S_{C_t}}{\sqrt{m}}, \overline{X}_{C_t} + 4S_{C_t}\right)^T. \tag{14}$$

Definition 6. An uncertain time series T_{Che} of length n can be rewritten in terms of matrix with the following formula:

$$\mathbf{V}_{\text{Che}} = \left[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\right]. \tag{15}$$

Additionally, it can be expanded as follows:

$$\begin{bmatrix} TS_{X,l} \\ TS_{\overline{X},l} \\ TS_{\overline{X},u} \\ TS_{X,u} \end{bmatrix} = \begin{bmatrix} \left(\overline{X}_{C_1} - 4S_{C_1}, \overline{X}_{C_2} - 4S_{C_2}, \dots, \overline{X}_{C_n} - 4S_{C_n}\right) \\ \left(\overline{X}_{C_1} - \frac{4S_{C_1}}{\sqrt{m}}, \overline{X}_{C_2} - \frac{4S_{C_2}}{\sqrt{m}}, \dots, \overline{X}_{C_n} - \frac{4S_{C_n}}{\sqrt{m}}\right) \\ \left(\overline{X}_{C_1} + \frac{4S_{C_1}}{\sqrt{m}}, \overline{X}_{C_2} + \frac{4S_{C_2}}{\sqrt{m}}, \dots, \overline{X}_{C_n} + \frac{4S_{C_n}}{\sqrt{m}}\right) \\ \left(\overline{X}_{C_1} + 4S_{C_1}, \overline{X}_{C_2} + 4S_{C_2}, \dots, \overline{X}_{C_n} + 4S_{C_n}\right) \end{bmatrix}, \tag{16}$$

where $TS_{X,l}$ is the lower bound sequence of random variable X_t composed of $\overline{X}_{C_t} - 4S_{C_t}$, $TS_{\overline{X},l}$ is referred to as lower bound sequence of variable $\overline{X}_t$, $TS_{\overline{X},u}$ is named $\overline{X}_t$ upper bound sequence, and the upper bound sequence of X_t is denoted as $TS_{X,u}$, illustrated in Figure 1. Four certain time series constitute an uncertain time series based on Chebyshev model.

4. Uncertain Time Series Preprocessing

4.1. Outlier Elimination from Sample Set. In the process of the sample collection, the occurrence of outliers is inevitable. As an abnormal observation value, it is distant from others

[19]. This may be ascribed to undesirable variability in the measurement or experimental errors. Outliers can occur in any distribution; naive interpretation of statistics such as sample mean and sample variance derived from sample set that include outliers may be misleading. Excluding outliers from sample set enhances the effectiveness of statistics. The definition of an outlier O can be formalized below.

Definition 7. Given a sample set $C_t = \{X_{c1}^t, X_{c2}^t, \ldots, X_{cm}^t\}$ at time slot t, C_t is sorted in ascending order. The sorted elements constitute a sample sequence, denoted as $(V_1, V_2, \ldots, V_m)$. Q_1 and Q_3 are the lower and upper quartiles, respectively; then we could define an outlier to be any sample outside the range:

$$[Q_1 - k(Q_3 - Q_1), Q_1 + k(Q_3 - Q_1)] \tag{17}$$

for a nonnegative constant k, which adjusts the granularity of excluding outliers.

4.2. Exponential Smoothing for Noise Reduction. In the area of signal processing, noise is a general term of unwanted (and, in general, unknown) modifications during signal capture, storage, transmission, processing, or conversion. To recover the original data from the noise-corrupted signal, the filters applied to noise reduction are ubiquitous in the design of signal processing systems. An Exponential smoothing filter assigns exponentially decreasing weights to the sample in time order and is effective [20–22]. In this subsection, we use exponential smoothing to process the noise in time series data. Given an certain time series Y, $Y(t-1)$ is the observation at time slot $t-1$, ES is a smoothed sequence associated with Y, and $\mathrm{ES}(t)$ is the smoothed value at time slot t. If the first sample is chosen in raw time series as initial value and an appropriate smoothing factor is picked, all values composed of smoothed sequence ES are available iteratively. The single form of exponential smoothing is given in formula

$$\begin{aligned} \mathrm{ES}(0) &= Y(0), \\ \mathrm{ES}(t) &= \alpha Y(t-1) + (1-\alpha) S(t-1). \end{aligned} \tag{18}$$

The raw time series begins at time $t = 0$; smoothing factor α falls in interval $[0,1]$. On the basis of the equation, the exponential smoothing of an uncertain time series modeled in Chebyshev matrix (Definition 6) is defined as follows:

$$\mathrm{ES}_{\mathrm{Che}}(0) = \begin{bmatrix} TS_{X,l}(0) \\ TS_{\overline{X},l}(0) \\ TS_{\overline{X},u}(0) \\ TS_{X,u}(0) \end{bmatrix}, \tag{19}$$

$$\mathrm{ES}_{\mathrm{Che}}(t) = \alpha \begin{bmatrix} TS_{X,l}(t-1) \\ TS_{\overline{X},l}(t-1) \\ TS_{\overline{X},u}(t-1) \\ TS_{X,u}(t-1) \end{bmatrix} + (1-\alpha)\,\mathrm{ES}_{\mathrm{Che}}(t-1). \tag{20}$$

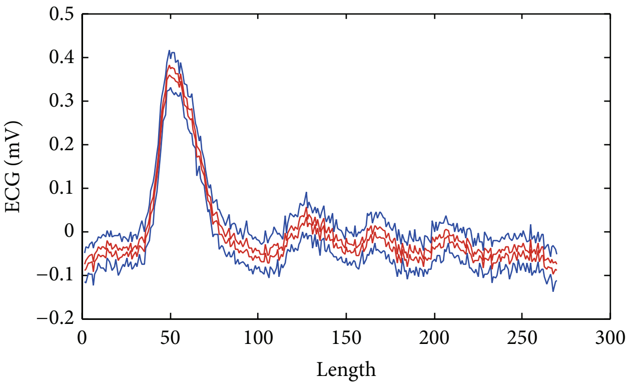


Figure 2: Illustration of Chebyshev uncertain time series before smoothing.

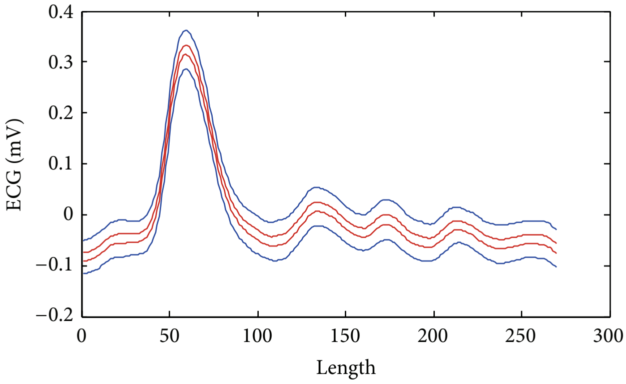


Figure 3: Illustration of Chebyshev uncertain time series smoothed.

For example, a raw time series is chosen from the ECG200 dataset in UCR time series collection [23]; after the disturbance by standard deviation 0.2, it is modeled as Chebyshev uncertain time series illustrated in Figure 2; tiny fluctuations around four lower and upper bound sequences reflect the existence of noise. We perform the exponential smoothing against the uncertain time series, choosing the first sample of each bound sequence as initial value and setting the smoothing factor α to 0.3. Note that higher value of α actually reduces the level of smoothing; in the limiting case with $\alpha = 1$ the output series is just the same as the original series. After triple exponential smoothing, the uncertain time series become clearer, because triple exponential smoothing takes into account seasonal changes as well as trends, illustrated in Figure 3.

4.3. Dimensionality Reduction Using Wavelets. In the process of analysis and organization of high-dimensional data, the difficulty is the problem of "curse of dimensionality" coined by Bellman and Dreyfus [24]. When the dimensions of the data space increase, data size soars, and thus the available data becomes sparse. Extracting these valid sparse data as feature vectors in lower dimension feature space is the essence of dimensionality reduction. Time series, as the special high-dimensional data, is under the influence of curse of dimensionality as well. We adopt wavelets frequently used in dimension reduction to deal with the time series data [25–27].

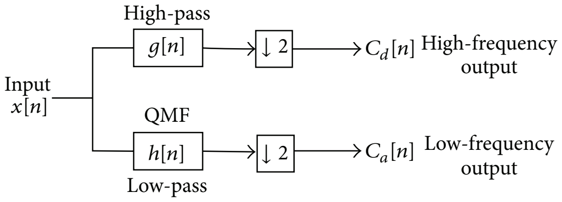


Figure 4: QMF wavelet transform for dimensionality reduction.

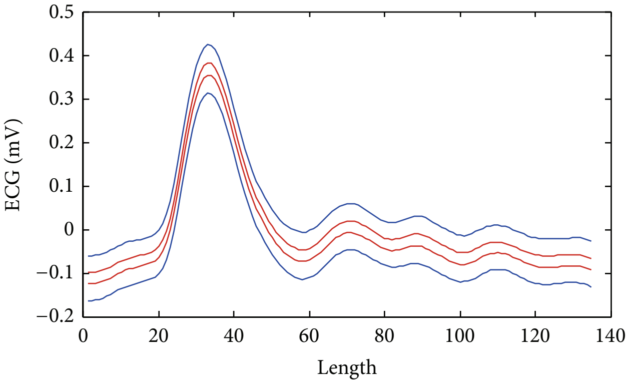


Figure 5: Illustration of smoothed Chebyshev uncertain time series after wavelet dimensionality reduction.

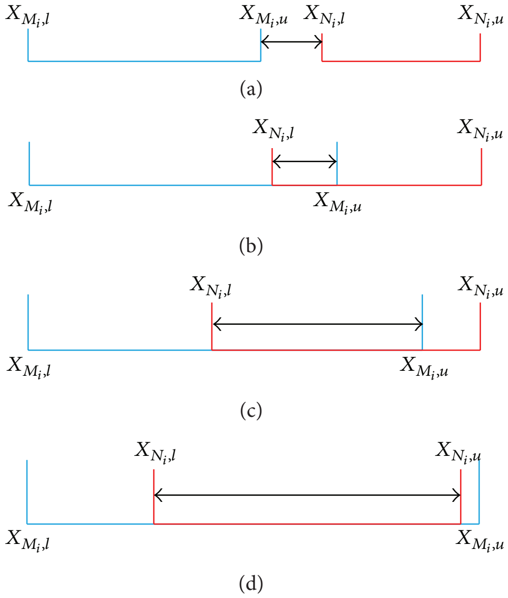


Figure 6: The illustration of similarity degrees.

Daubechies [28] finds that wavelet transforms can be implemented using a pair of Finite Impulse Response (FIR) filters, called a Quadrature Mirror Filter (QMF) pair. These filters are often used in the area of signal processing as they lend themselves to efficient implementation. Each filter is represented as a sequence of numbers. The filter lends this the length of this sequence. The output of a QMF pair consists of two separate components: a high-pass and a low-pass filter, which correspond to high-frequency and low-frequency output, respectively. Wavelet transforms are considered to be hierarchical since they operate stepwise. The input on each step is passed through the QMF pair. Both high-pass and low-pass component of the QMF output are in half of the length of the input. The high-pass component is naturally associated with details while the low-pass component concentrates most of the energy or information of the data. The low-pass component is used as further input; hence the length of the input is reduced by a factor of 2 at each step. The single step is illustrated in Figure 4, where n refers to the length of signal sequence in general, not some concrete value.

For example, as shown in Figure 3, we choose Haar wavelet to build QMF pair; the low-pass output is a dimension-reduced uncertain time series whose length shortens from 270 to 135, illustrated in Figure 5; the sequence of QMF pair based on Haar wavelet is defined as follows:

$$\begin{aligned} g[n] &= \left[\frac{-\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right], \\ h[n] &= \left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]. \end{aligned} \tag{21}$$

Note that the low-pass output is obtained through the convolution of $h[n]$ and the uncertain time series to be reduced in dimension; in the same manner, the convolution of $g[n]$ and the uncertain time series is the high-pass output.

5. Similarity Match Processing

We present a new matching method based on Chebyshev uncertain time series. As shown in Definition 5, without loss of generality, we utilize two variables M_i, N_i from different uncertain time series M and N at time slot i to specify the matching procedure. Let $[X_{M_i,l}, X_{M_i,u}]$ and $[X_{N_i,l}, X_{N_i,u}]$ be the sample estimation interval from M and N at time slot i in Figure 6(a). If the two intervals overlapped as shown in Figure 6(b), M_i and N_i have possibility of taking identical value from the overlap intersection set; with the increasing of overlap in Figures 6(c) and 6(d) (expressed by the double arrow solid lines), the possibility increases gradually. Thus, M_i and N_i become more similar in terms of the range of samples. The above analysis outlines the similarity measure based on the overlap of sample estimation intervals qualitatively; then we analyze the process quantitatively. The lengths of two sample estimation intervals at identical time slot are different. As shown in Figure 6, let L_{M_i} and L_{N_i} be the length of sample estimation intervals of M_i and N_i, respectively:

$$L^X_{M_i} = X_{M_i,u} - X_{M_i,l}, \tag{22}$$

$$L^X_{N_i} = X_{N_i,u} - X_{N_i,l}. \tag{23}$$

L^X_{opi} denote the length of overlap between M_i and N_i illustrated in Figures 6(b) and 6(c):

$$L^X_{\text{opi}} = \left| X_{M_i,u} - X_{N_i,l} \right|. \tag{24}$$

In Figure 6(d), L^X_{opi} equals

$$L^X_{\text{opi}} = \left| X_{N_i,u} - X_{N_i,l} \right|. \tag{25}$$

If the two observations intervals are not overlapped in Figure 6(a), the problem arises. In fact, it should be marked; we put a negative symbol into formula like this

$$L^X_{\text{opi}} = -\left| X_{N_i,l} - X_{M_i,u} \right|. \tag{26}$$

If $L^X_{\text{opi}} \leq 0$, the two observation intervals have no overlap, and the lower L^X_{opi} is, the farther two intervals become. Let *Overlap Ratio* be the ratio of the length of overlap to length of observation intervals to quantify the degree of overlap, denoted as rop; thus,

$$\begin{aligned} \text{rop}^X_{M_i} &= \frac{L^X_{\text{opi}}}{L^X_{M_i}}, \\ \text{rop}^X_{N_i} &= \frac{L^X_{\text{opi}}}{L^X_{N_i}}, \end{aligned} \tag{27}$$

where each of them falls in $(-\infty, 1]$ (only when the length of overlap equals the length of observations interval, rop^X equals 1 in Figure 6(d), $\text{rop}^X_{N_i} = L^X_{\text{opi}}/L^X_{N_i} = 1$).

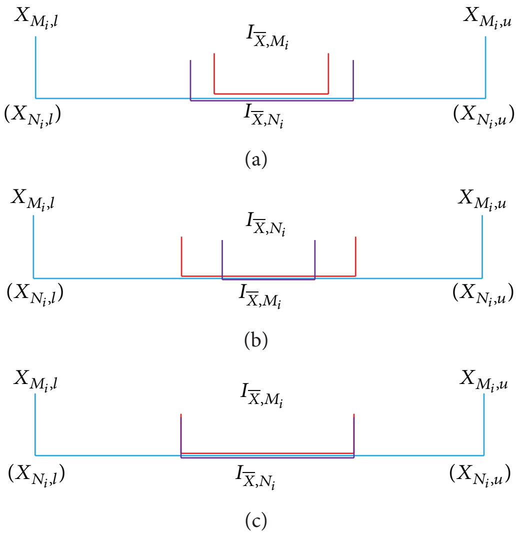
We combine $\text{rop}^X_{M_i}$ and $\text{rop}^X_{N_i}$ and construct a single quantity called *Overlap Degree* of sample estimation interval, denoted as dop^X, so that it measures the overlaps linearly. Here is the definition

$$\text{dop}^X_{M_i,N_i} = \frac{2\text{rop}^X_{M_i} \cdot \text{rop}^X_{N_i}}{\text{rop}^X_{M_i} + \text{rop}^X_{N_i}}, \tag{28}$$

where $\text{dop}^X_{M_i,N_i}$ also belongs to $(-\infty, 1]$. The sum of $\text{dop}^X_{M_i,N_i}$ denotes the degree of overlap between the two uncertain time series M and N such that

$$\text{DOP}^X_{M,N} = \sum_{i=1}^{n} \text{dop}^X_{M_i,N_i}. \tag{29}$$

We will further discuss the similarity between M_i and N_i. As illustrated in Figure 7, even if two sample estimation intervals at time i are entirely overlapped, it is difficult to determine whether the two variables have similarity or not to a certain degree, because of a variety of overlapping between central tendency estimation intervals $I_{\overline{X},M_i}$ and $I_{\overline{X},N_i}$. In other words, the degree of overlap between $I_{\overline{X},M_i}$ and $I_{\overline{X},N_i}$ determines the degree of similarity between M_i and N_i on condition of identical sample estimation intervals. As shown in Figure 7(c), the two variables, compared to the case in Figures 7(a) and 7(b), are more similar obviously; the larger overlapping is, the more similar two variables are. If central tendency estimation intervals have no overlap or a little and sample estimation intervals overlap to some extent, the estimation of similarity cannot be obtained. With regard to the above cases, only $\text{DOP}^X_{M,N}$ is not sufficient to measure the similarity; we need further to measure the similarity between two variables with central tendency estimation intervals.

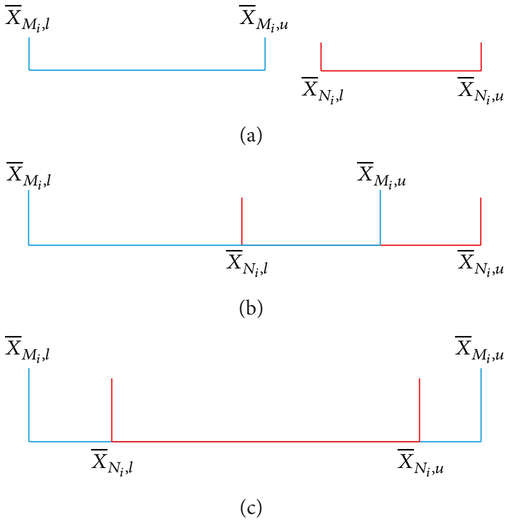


(a)

(b)

(c)

FIGURE 7: The situations of entire overlapping.



(a)

(b)

(c)

FIGURE 8: The illustration of overlap between $\overline{X}$ intervals.

As illustrated in Figure 8, there are three cases of overlapping. Let $L^{\overline{X}}_{\text{opi}}$ be the overlap between two central tendency estimation intervals. In Figure 8(a), for the estimation intervals $[\overline{X}_{M_i,l}, \overline{X}_{M_i,u}]$ and $[\overline{X}_{N_i,l}, \overline{X}_{N_i,u}]$, the lengths of estimation interval $L^{\overline{X}}_{M_i}$ and $L^{\overline{X}}_{N_i}$ are represented as

$$\begin{aligned} L^{\overline{X}}_{M_i} &= \overline{X}_{M_i,u} - \overline{X}_{M_i,l}, \\ L^{\overline{X}}_{N_i} &= \overline{X}_{N_i,u} - \overline{X}_{N_i,l}. \end{aligned} \tag{30}$$

With no overlapping between them, the $L^{\overline{X}}_{\text{opi}}$ is denoted as

$$L^{\overline{X}}_{\text{opi}} = -\left| \overline{X}_{N_i,l} - \overline{X}_{M_i,u} \right|. \tag{31}$$

In Figure 8(b), $[\overline{X}_{M_i,l}, \overline{X}_{M_i,u}]$ and $[\overline{X}_{N_i,l}, \overline{X}_{N_i,u}]$ have overlap as described below:

$$L_{\text{opi}}^{\overline{X}} = \left|\overline{X}_{M_i,u} - \overline{X}_{N_i,l}\right|. \tag{32}$$

In Figure 8(c), $[\overline{X}_{M_i,l}, \overline{X}_{M_i,u}]$ contains $[\overline{X}_{N_i,l}, \overline{X}_{N_i,u}]$; the overlap is represented as follows:

$$L_{\text{opi}}^{\overline{X}} = \left|\overline{X}_{N_i,u} - \overline{X}_{N_i,l}\right|. \tag{33}$$

Analogous to rop^{X}, the Overlap Ratio of $\overline{X}$ estimation interval between M_i and N_i is defined:

$$\begin{aligned} \text{rop}_{M_i}^{\overline{X}} &= \frac{L_{\text{opi}}^{\overline{X}}}{L_{M_i}^{\overline{X}}}, \\ \text{rop}_{N_i}^{\overline{X}} &= \frac{L_{\text{opi}}^{\overline{X}}}{L_{N_i}^{\overline{X}}}. \end{aligned} \tag{34}$$

The *Overlap Degree* of $\overline{X}$, namely, $\text{dop}^{\overline{X}}$ between M_i and N_i, is depicted below:

$$\text{dop}_{M_i,N_i}^{\overline{X}} = \frac{2\text{rop}_{M_i}^{\overline{X}} \cdot \text{rop}_{N_i}^{\overline{X}}}{\text{rop}_{M_i}^{\overline{X}} + \text{rop}_{N_i}^{\overline{X}}}. \tag{35}$$

We sum up rop_{M_i,N_i} of the two uncertain time series M and N in length of n; the sum indicated by $\text{DOP}_{M,N}^{\overline{X}}$ is represented as

$$\text{DOP}_{M,N}^{\overline{X}} = \sum_{i=1}^{n} \text{rop}_{M_i,N_i}^{\overline{X}}. \tag{36}$$

In conclusion, we combine the DOP^{X} and $\text{DOP}^{\overline{X}}$ to evaluate the degree of similarity between two uncertain time series, which is signified by DOS and expressed as follows:

$$\text{DOS}_{M,N} = \alpha \text{DOP}_{M,N}^{X} + (1-\alpha)\,\text{DOP}_{M,N}^{\overline{X}}. \tag{37}$$

α is the factor in the range of $[0, 1]$; in different applications, DOP^{X} and $\text{DOP}^{\overline{X}}$ refer to different weights; here set $\alpha = 1/2$. Consider $\text{DOS}_{M,N} \in (-\infty, n]$ (n is length of uncertain time series).

6. Experimental Validation

In this section, we examine the effectiveness and efficiency of the new method proposed in this paper. Firstly, we introduce the uncertain time series value generation and experimental datasets; then we analyse the results of the experiments. All the methods are implemented in MATLAB and C++, and the experiments are run on a PC with 3.1 GHz CPU and 4 GB of RAM.

6.1. Uncertainty Model and Assumption. As described in Definition 5, an uncertain time series T is a time series including sample estimation interval and central tendency estimation interval derived from a set of observations at each time slot. Given a time slot i, the value of uncertain time series modeled as

$$T_i = d_i + e_i, \tag{38}$$

where d_i is the true value and e_i is the error. In general, the error e_i could be drawn from distinct probability distribution; this is why we treat T_i as a random variable at the time i.

Table 1: Details of time series sets.

Dataset	Quantity	Length
50words	450	270
Adiac	390	176
Beef	470	30
CBF	500	128
Coffee	500	28
ECG200	199	96
Lighting2	121	637
SyncCtrl	120	300
Wafer	6164	152
FaceFour	112	350
FaceAll	560	131
Fish	349	463
Lighting7	318	73
GunPoint	199	150
OliveOil	570	30
OSULeaf	441	427
SwedLeaf	1125	128
Trace	200	270
Yoga	300	427

6.2. Experimental Setup. Inspired by [11, 12, 15], we use real time series datasets of exact values and subsequently introduce uncertainty with uncertainty model through perturbation. In our experiments we consider *Uniform*, *Normal*, and *Exponential* error distributions with zero mean and vary standard deviation within interval $[0.2, 2.0]$.

We selected 19 real datasets from the UCR classification dataset collection [23]; they represent a wide range of application areas: *50words*, *Adiac*, *Beef*, *CBF*, *Coffee*, *ECG200*, *Lighting2*, *SyncCtrl*, *Wafer*, *FaceFour*, *FaceAll*, *Fish*, *Lighting7*, *GunPoint*, *OliveOil*, *OSULeaf*, *SwedLeaf*, *Trace*, and *Yoga*. The training and testing sets are reconfigured, and we acquired the time series sets as Table 1.

6.3. Accuracy. On the purpose of evaluating the quality of the results, we use the two standard measures of recall and precision. Recall is defined as the percentage of the truly similar uncertain time series that are found by the algorithm. Precision is the percentage of similar uncertain time series identified by the algorithm, which are truly similar. Accuracy is measured in terms of the harmonic mean of recall and precision to facilitate the comparison. The accuracy is defined as follows:

$$\text{Accuracy} = \frac{2\text{recall} * \text{precision}}{\text{recall} + \text{precision}}. \tag{39}$$

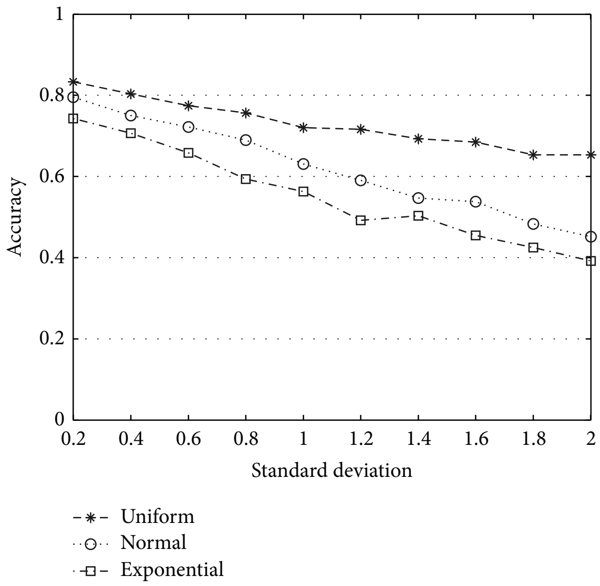


FIGURE 9: Accuracy with three error distributions averaged over all datasets.

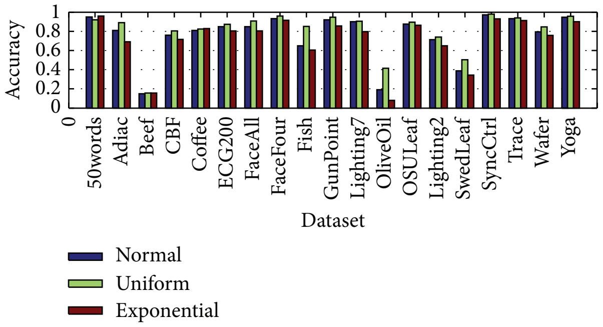


FIGURE 10: Accuracy of 19 datasets on three error distributions with accuracy of 0.4 and 1.0 mixed deviation.

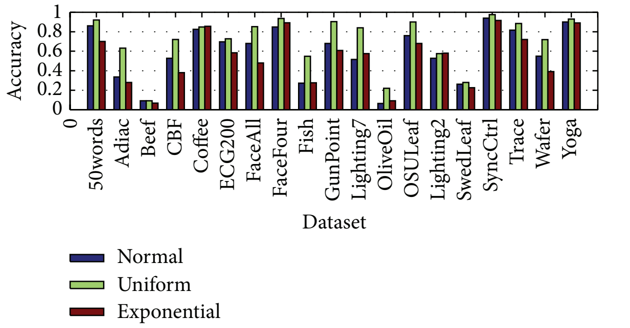


FIGURE 11: Accuracy of 19 datasets on three error distributions with accuracy of 1.4 and 2.0 mixed deviation.

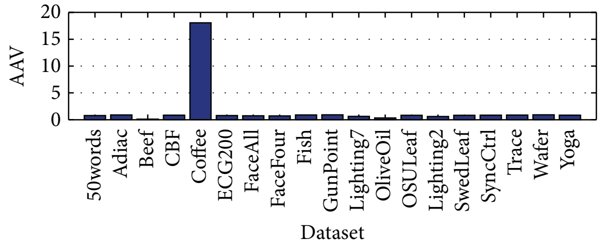


FIGURE 12: Average absolute value (AAV) of each dataset disturbed data.

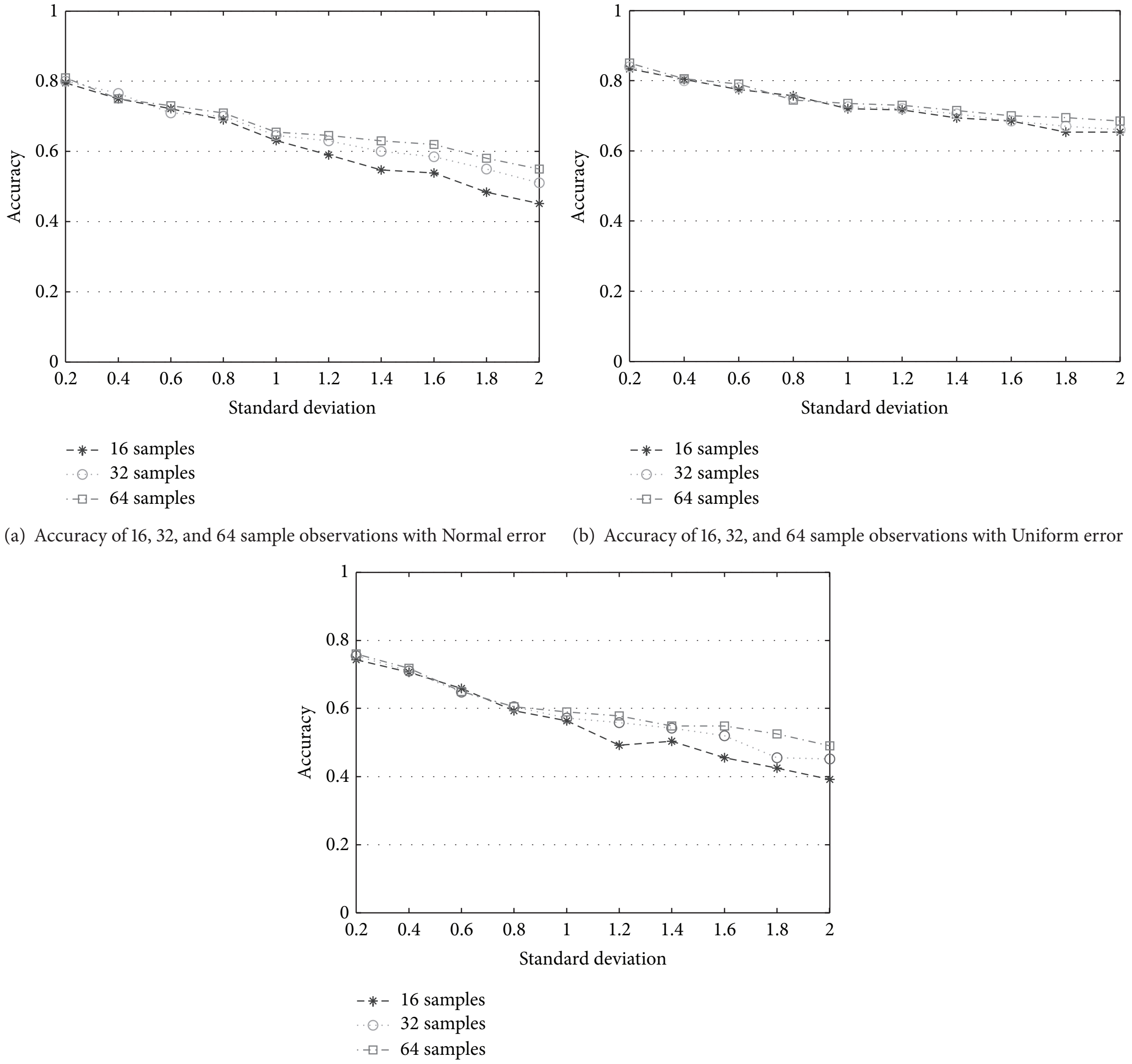
As mentioned in [11], an effective similarity measure on uncertain data allows us to reason about the original data without uncertainty. For the sake of validating new method, we conduct experiments from different aspects.

In the first experiment, we examine the effectiveness of our approach for different error standard deviations and error distributions. In Figure 9, the results from different error distributions are averaged over all datasets and shown at various error standard deviations. The accuracy decreases linearly with increasing error standard deviation from 0.2 to 2 and the performance with Uniform distribution is better than the other two distribution performances. Bigger standard deviations produce more uncertainty to time series data.

Next, we verify the effectiveness for different datasets. In Figure 10, each time series from each dataset is perturbed with different error, that is, Normal, Uniform, and Exponential; combining 20% accuracy of the match in standard deviation 1 with 80% accuracy of the match in standard deviation 0.4 as the accuracy of relative small standard deviations on each dataset, most of datasets perform well (accuracy reaches 80% or so, some come to 90%), with *SyncCtrl* being the best performer (accuracy = 96%), except *Beef*, *OliveOil*, and *SwedLeaf*, which will be explained below. Similarly, the trend is verified also with Uniform and Exponential error distributions.

Figure 11 summarizes the performance of each dataset in relative big standard deviations of error, integrating the 20% accuracy of match in standard deviation 2 with 80% accuracy in standard deviation 1.4. As with the increasing of standard deviation, the accuracy of all datasets decreases. With Normal error, the accuracy of *Adiac* drops the most fast, nearly 50% (from 81% to 33%), and the tendency is also held with Exponential error distribution. *Coffee*, *FaceFour*, *SyncCtrl*, and *yoga* are exceptions; the increasing standard deviations have no significant impact on their accuracy. With Uniform error, the accuracy of *Fish* drops the most fast, up to 30.4%, the accuracy of *Adiac* drops 25.8%, and ECG200 decreases 14.4%; the accuracy of other datasets falls lightly. With Exponential error, most datasets drop fast and the most fast dataset is *Adiac*, up to 41%. In conclusion, the Uniform error impacts all datasets lightly with the increasing standard deviation, compared to the Normal and Exponential error.

As mentioned above, the datasets *Beef*, *OliveOil*, and *SwedLeaf* have poor performance, but *Coffee*, *FaceFour*, *syncCtrl*, and *yoga* perform well in Figures 10 and 11. We find that all of these are partially related to the average absolute value of respective datasets which are disturbed. As shown in Figure 12, we compute the average absolute values of all disturbed datasets; the AAVs (average absolute values) of *Beef* and *OliveOil* are 0.0956 and 0.3337, respectively, smaller than others. The AAV of disturbed *Coffee* is 18.0541, which is the



(a) Accuracy of 16, 32, and 64 sample observations with Normal error

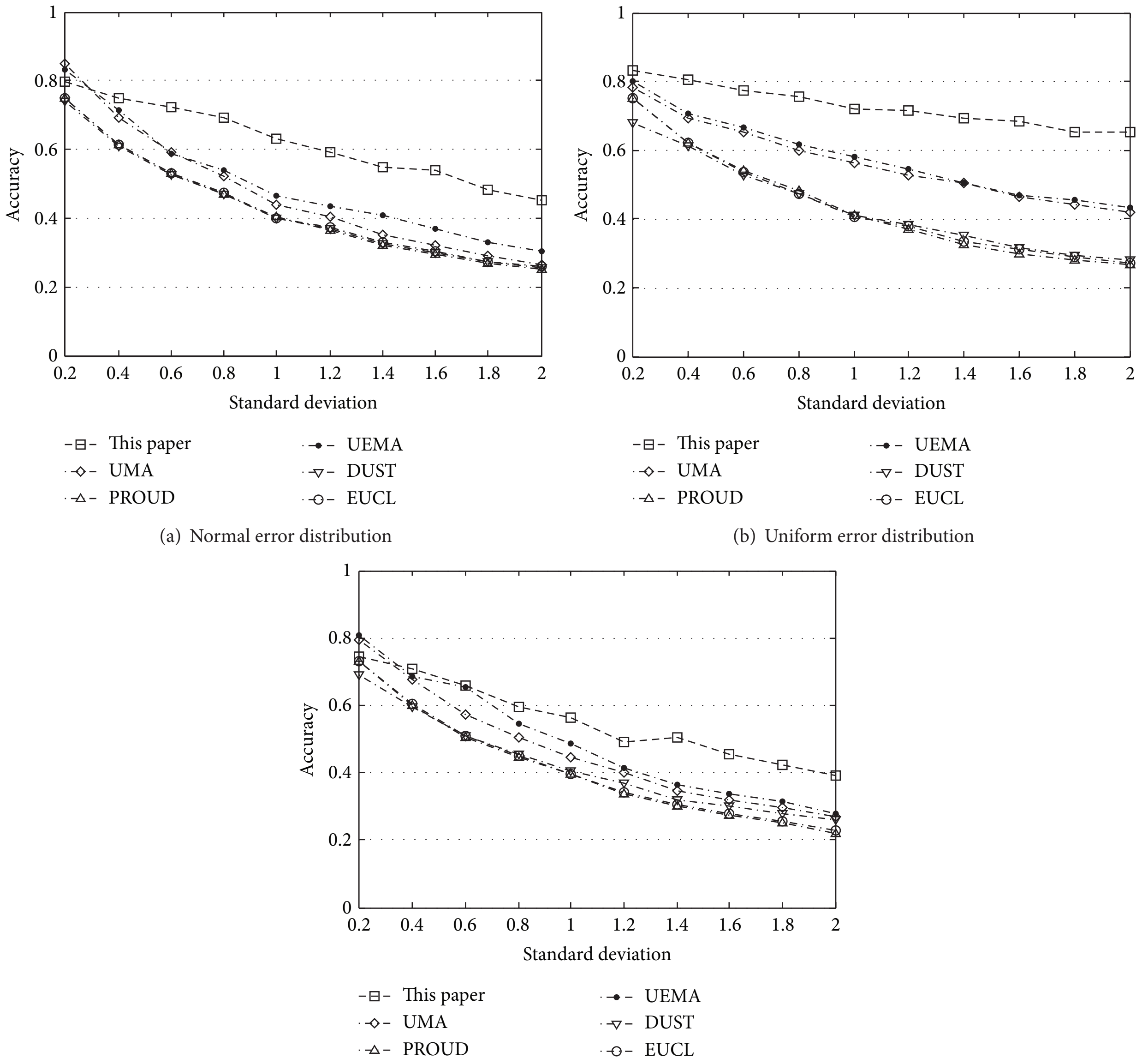
(b) Accuracy of 16, 32, and 64 sample observations with Uniform error

(c) Accuracy of 16, 32, and 64 sample observations with Exponential error

Figure 13: Comparison of accuracy with different sample size.

biggest among all datasets; the other three datasets are also big ones. In other words, for large AVVs it is difficult to be impacted with small uncertainty even though standard deviation of error comes to 2. On the contrary, *Beef* and *OliveOil* are easier to be impacted even if standard deviation of error is 0.2. However, *SwedLeaf* is different; it may be ascribed to the wave form, which we will explore in future research. Considering the impact of the size of observation samples, it is important for two kinds of estimation intervals which stem from observation samples. As described above, all experiments results are based on $m = 16$ observation samples. We describe how the results come to be if the size of observation sample gets large. In Figure 13(a), with the Normal error, the accuracy of three sizes of observation sample is shown at various standard deviations. The result of 64 samples is the best; 32 samples result is better than 16 samples. At relative small standard deviations (0.2–0.8), the results of three sizes are of little difference; with the deviation growing, the differences gradually become more observable. The results of Uniform and Exponential distributions are similar to Normal and are reported in Figures 13(b) and 13(c). The differences with Uniform error among three sizes are smaller than the other two distributions.

In Figure 14(a) we compare our approach with other techniques under Normal error distribution, namely, PROUD, DUST, Euclidean distance, UMA, and UEMA, referring to the methodology proposed in [16]. The results demonstrate that our approach is more effective than other techniques with



(a) Normal error distribution

(b) Uniform error distribution

(c) Exponential error distribution

Figure 14: Comparison of accuracy with existing methods.

three distribution errors. With 0.2 error deviations, UEMA and UMA outperform others; PROUD performs slightly better than DUST and Euclidean, but with larger error standard deviation its accuracy drops slightly below DUST and Euclidean. This trend is also kept with Uniform and Exponential distribution, illustrated in Figures 14(b) and 14(c).

We also compare the performance of execution time for our approach with other techniques mentioned above. Because the results of three distributions are analogous, the Normal distribution is drawn as an example to show the trend of the results. Figure 15 shows the CPU time per query for Normal error distribution with varying error standard deviation from 0.2 to 2. It shows that the varying standard deviations for error do not impact the performance of these techniques basically. The performance of our approach is slightly better than DUST, UMA, and UEMA. The best time performer is Euclidean. Note that we do not apply PROUD to wavelet synopses; this may be the reason why it does not perform well.

In Figure 16, we describe the CPU time per query for Normal error distribution with varying time series length between 50 and 1000. The time series of different length are obtained by reconstitution of raw datasets. The figure shows that the execution time increases linearly to the time series length. The results of our approach are better than DUST and PROUD; Euclidean gets the best performance.

7. Conclusion

In this paper, we propose a new model of uncertain time series and a new approach that measures the similarity between uncertain time series. It outperforms the

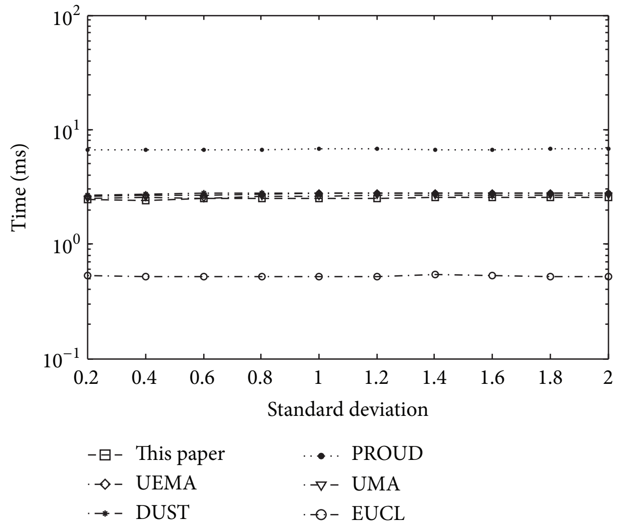


Figure 15: Average CPU time per query for Normal error distribution with varying deviation.

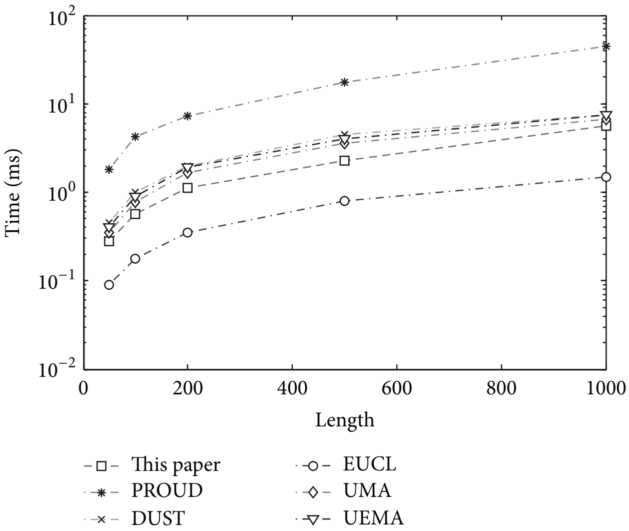


Figure 16: Average CPU time per query for Normal error distribution with varying length.

state-of-the-art techniques, most of which employ the distance measure to evaluate the similarity.

We validate the new approach with three kinds of error distributions and the standard deviations of error span the range from 0.2 to 2; meanwhile, we compare the new approach with the techniques previously proposed in the literature. Our experiments were based on 19 authentic datasets. The results demonstrate that overlap measuring, based on observations interval and central tendency, outperforms the other complex alternatives. If the expected value of the error in the experiments is considered to be zero, the average of these samples may be a good estimate for unknown values at each time slot; it characterizes the center of data distribution.

In the future, we will make a deeper exploration of the modeling of uncertain time series data when the expected value of the error is zero. We will extend our work to index technique about uncertain time series. We will explore the influence of wave characteristics of time series data and the management of volume uncertain time series.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] M. Orang and N. Shiri, "An experimental evaluation of similarity measures for uncertain time series," in *Proceedings of the 18th International Database Engineering and Applications Symposium (IDEAS '14)*, pp. 261–264, ACM, July 2014.

[2] M. Ceriotti, M. Corrà, L. D'Orazio et al., "Is there light at the ends of the tunnel? Wireless sensor networks for adaptive lighting in road tunnels," in *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pp. 187–198, April 2011.

[3] L. Krishnamurthy, R. Adler, P. Buonadonna et al., "Design and deployment of industrial sensor networks: experiences from a semiconductor plant and the North Sea," in *Proceedings of the 3rd ACM International Conference on Embedded Networked Sensor Systems (SenSys '05)*, pp. 64–75, ACM, New York, NY, USA, November 2005.

[4] M. S. Mit, J. B. Slac, D. D. Microsoft et al., "Requirements for science data bases and SciDB," in *Proceedings of the 4th Biennial Conference on Innovative Data Systems Research (CIDR '09)*, Asilomar, Calif, USA, January 2009.

[5] S. Dan, B. Howe, and A. Connolly, "Embracing uncertainty in large-scale computational astrophysics," in *Proceedings of the 3rd VLDB Workshop on Management of Uncertain Data (MUD '09)*, pp. 63–77, Lyon, France, August 2009.

[6] T. T. L. Tran, L. Peng, B. Li, Y. Diao, and A. Liu, "PODS: a new model and processing algorithms for uncertain data streams," in *Proceedings of the International Conference on Management of Data (SIGMOD '10)*, pp. 159–170, June 2010.

[7] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," *Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 107–144, 2007.

[8] Q. Chen, L. Chen, X. Lian et al., "Indexable PLA for efficient similarity search," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB Endowment, September 2007.

[9] C. C. Aggarwal, *Managing and Mining Uncertain Data*, Advances in Database Systems, Springer, 2011.

[10] Y. Zhao, C. Aggarwal, and P. S. Yu, "On wavelet decomposition of uncertain time series data sets," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*, pp. 129–138, Toronto, Canada, October 2010.

[11] S. R. Sarangi and K. Murthy, "DUST: a generalized notion of similarity between uncertain time series," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pp. 383–392, July 2010.

[12] M.-Y. Yeh, K.-L. Wu, P. S. Yu, and M.-S. Chen, "PROUD: a probabilistic approach to processing similarity queries over uncertain data streams," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '09)*, pp. 684–695, ACM, March 2009.

[13] E. J. Thalassinakis and E. N. Dialynas, "A Monte-Carlo simulation method for setting the underfrequency load shedding relays and selecting the spinning reserve policy in autonomous power systems," *IEEE Transactions on Power Systems*, vol. 19, no. 4, pp. 2044–2052, 2004.

[14] L.-I. Tong, K. S. Chen, and H. T. Chen, "Statistical testing for assessing the performance of lifetime index of electronic components with exponential distribution," *International Journal of Quality and Reliability Management*, vol. 19, no. 7, pp. 812–824, 2002.

[15] K. Aßfalg, H.-P. Kriegel, P. Kröger, and M. Renz, "Probabilistic similarity search for uncertain time series," in *Scientific and Statistical Database Management*, vol. 5566 of *Lecture Notes in Computer Science*, pp. 435–443, Springer, Berlin, Germany, 2009.

[16] M. Dallachiesa, B. Nushi, K. Mirylenka, and T. Palpanas, "Uncertain time-series similarity: return to the basics," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1662–1673, 2012.

[17] J. D. Storey, "False discovery rates," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed., p. 239, Springer, 1st edition, 2011.

[18] H. F. Weisberg, "Central tendency and variability," *Learning and Individual Differences*, vol. 21, no. 5, pp. 549–554, 1992.

[19] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 2012.

[20] R. H. Jones, "Exponential smoothing for multivariate time series," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 28, no. 1, pp. 241–251, 1966.

[21] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, Springer Texts in Statistics, Springer, 1996.

[22] C. Chatfield, *The Analysis of Time Series: An Introduction*, CRC Press, 2013.

[23] X. Keogh, "The UCR Time Series Classification/Clustering," 2006, http://www.cs.ucr.edu/~eamonn/time_series_data/.

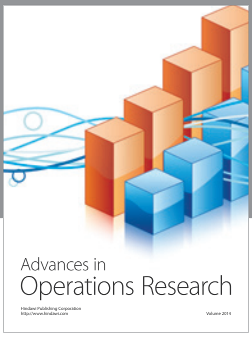
[24] R. E. Bellman and S. E. Dreyfus, "Applied dynamic programming," *Journal of the American Statistical Association*, vol. 59, no. 305, p. 293, 1964.

[25] I. Popivanov and R. J. Miller, "Similarity search over time-series data using wavelets," in *Proceedings of the 18th IEEE International Conference on Data Engineering*, pp. 212–221, IEEE Computer Society, San Jose, Calif, USA, 2002.

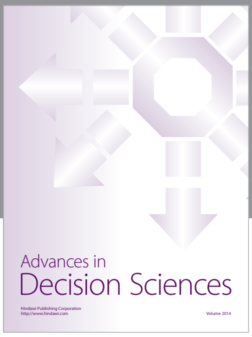
[26] Z. R. Struzik and A. Siebes, "The haar wavelet transform in the time series similarity paradigm," in *Principles of Data Mining and Knowledge Discovery*, vol. 1704 of *Lecture Notes in Computer Science*, pp. 12–22, 1999.

[27] K.-P. Chan and A. W.-C. Fu, "Efficient time series matching by wavelets," in *Proceedings of the 15th International Conference on Data Engineering (ICDE '99)*, pp. 126–133, IEEE, Sydney, Australia, March 1999.

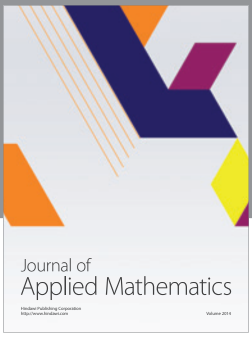
[28] I. Daubechies, "Ten lectures on wavelets," *Acoustical Society of America Journal*, vol. 93, no. 3, p. 1671, 1992.



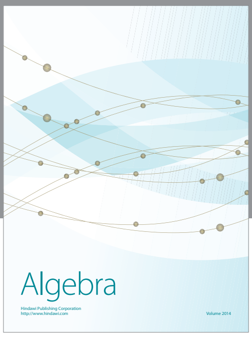
Advances in Operations Research



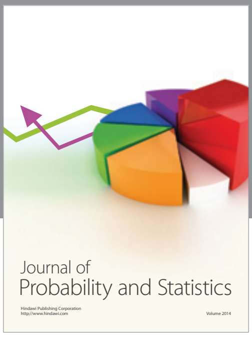
Advances in Decision Sciences



Journal of Applied Mathematics

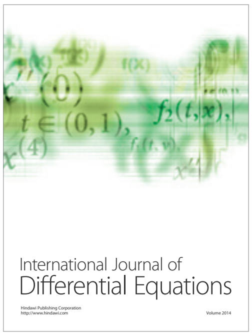


Algebra



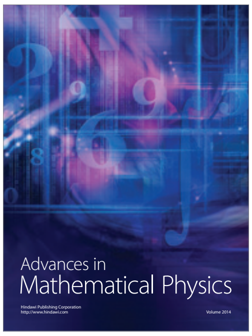
Journal of Probability and Statistics

The Scientific World Journal

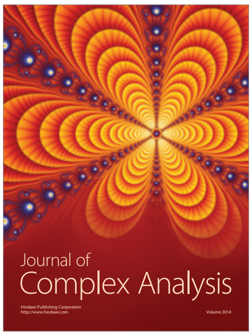


International Journal of Differential Equations

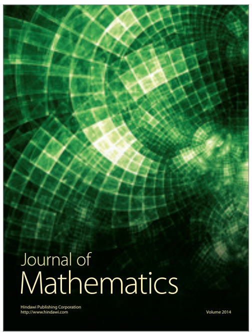
International Journal of Combinatorics



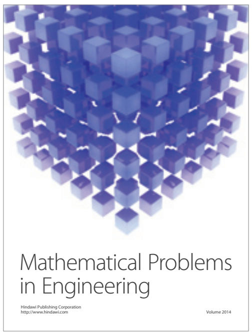
Advances in Mathematical Physics



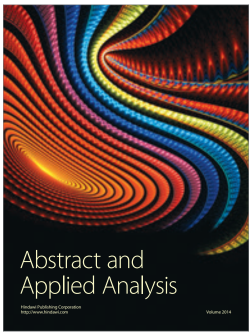
Journal of Complex Analysis



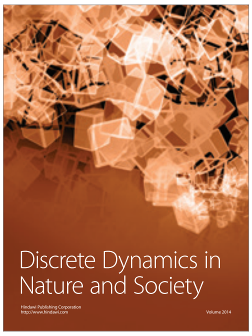
Journal of Mathematics



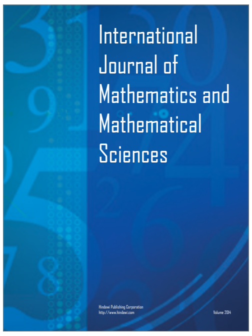
Mathematical Problems in Engineering



Abstract and Applied Analysis



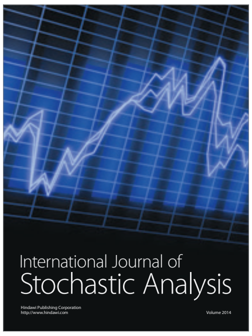
Discrete Dynamics in Nature and Society



International Journal of Mathematics and Mathematical Sciences

Journal of Discrete Mathematics

Journal of Function Spaces



International Journal of Stochastic Analysis

Journal of Optimization