

# Equiprobable Symbolization Pattern Entropy for Time Series Complexity Measurement

Fuyi Wang (✉ [wangfuyi@deakin.edu.au](mailto:wangfuyi@deakin.edu.au))

Deakin University <https://orcid.org/0000-0002-8216-3238>

Leo Yu Zhang

Deakin University

---

## Research Article

**Keywords:** Equiprobable symbolization, Multiscale sample entropy, Time series complexity, Wind speed signals, Logistic system

**Posted Date:** August 13th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-800663/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# Equiprobable symbolization pattern entropy for time series complexity measurement

Fuyi Wang · Leo Yu Zhang

Received: date / Accepted: date

**Abstract** In order to more effectively mine the structural features in time series, while simplifying the complexity of time series analysis, equiprobable symbolization pattern entropy (EPSPE) based on time series symbolization combined with sliding window technology is proposed in this paper. Firstly, time series are implemented symbolic procession according to the equal probability distribution of the original data, which greatly simplifies the difficulty of analyzing the signal on the premise of small loss of precision to the original signal. Then, sliding window technique is used to obtain a finite number of different symbolic patterns, and the pattern pairs are determined by calculating the conversion between the symbolic patterns. Next, the conversion frequency between symbolized patterns is counted to calculate the probability of the pattern pairs, thus estimating the complexity measurement of complex signals. The results of test using the Logistic system with different parameters show that compared with multiscale sample entropy (MSE), EPSPE can more concisely and intuitively reflect the structural characteristics of time series. Finally, EPSPE is used to investigate the natural wind field signals collected at an outdoor space in which nine high precision two-dimensional (2D) ultrasonic anemometers are deployed in line with 1m in-

terval. The values of EPSPE show consistent increase or decrease trend with the spatial regular arrangement of the nine anemometers. While the results of MSE are irregular, and cannot accurately predict the spatial deployment relationship of nine 2D ultrasonic anemometers.

**Keywords** Equiprobable symbolization · Multiscale sample entropy · Time series complexity · Wind speed signals · Logistic system

## 1 Introduction

Time series is the external manifestation of complex systems in nature and contains rich information about the system structure and dynamic evolution rule, usually showing nonlinear and non-stationary characteristics. The complexity measurement of time series provides valuable clues to further reveal the dynamic evolution of complex systems, thus attracting many scholars to study this problem, and many complexity measurement methods have been proposed. As a measure of the complexity and determinism of real systems, entropy plays a pivotal role in the characterization of the nonlinear dynamics of systems. The classical methods include Kolmogorov-Sinai entropy [9], permutation entropy [1], approximate entropy [14] and sample entropy [15], etc. Kolmogorov-Sinai entropy mainly reflects the rate of increase of sub-sequences with time series length. Permutation entropy method is based on the principle

---

F. Wang · L. Zhang  
School of Information Technology, Deakin University, VIC  
3216, Australia  
E-mail: wangfuyi@deakin.edu.au, leo.zhang@deakin.edu.au

of phase space reconstruction, and measures the relative frequency by counting the occurrence of various alignment patterns in time series. Sample entropy measurement method is an improvement of approximate entropy method, which is different from approximate entropy in that it calculates the logarithm of the probability sum and has no matching process of its own. At present, sample entropy method is widely used in time series complexity analysis and the measurement results are more reliable and robust to the problem of missing data. Based on the sample entropy, Costa proposed multiscale sample entropy (MSE) from a multiscale perspective by means of coarse-graining, which is a method to measure the complexity of finite long time series [3]. Research shows that MSE is more comprehensive than sample entropy in revealing the differences between disease and health status. MSE is widely used, which can comprehensively reveal the intrinsic dynamics of complex systems in such fields as metal materials, environment and physiology.

With the continuous expansion of time series complexity measurement applications, it has been widely used in many fields, such as stocks [2], traffic flow [25], meteorology [24] [20], and biomedicine [19] [13] [21]. A large number of test results show that there are certain problems in the accuracy and universality of the existing methods in complexity measurement, and some methods, such as the approximate entropy, are better for some specific types of signals, but have large deviations and poor stability when they are used for other types of signals [15]. Permutation entropy is based on the alignment pattern confirmed by the data size relationship within the time series nearest neighbor value or behavior vector to extract important information features of the time series, but it has scale effect to deal with small sample sequences and cannot accurately estimate the complexity of small sample sequences [26]. In addition, the time series with noise in the real world usually jump within a certain range and there are many jump values, making it difficult to find the intrinsic patterns contained in the original values of the time series [23]. Symbolization can mitigate the impact of noise, reduce data consumption on memory, and accelerate signal processing. Therefore, symbolization is of great practical value for natural signal processing. Studies have shown that when the symbolization method is selected appropriately, the symbol sequence can ef-

fectively retain the dynamic nature of time series [18]. Bian et. al. [2] introduce a new two-index entropy by analyzing the maximization of entropy for the case of equal probabilities of basic elements extracted from the system and analysis the complexities of financial markets. At the same time, we note that different symbolic methods have been used for physiological time series and have obtained rich research results [22] [4] [5]. Shang et al. [16] adopt symbolic method to propose topological entropy and geometric entropy, which examines the correlations and complexities of different time series and can effectively distinguish between the different time series. A proper symbolic processing of time series will therefore help to uncover the inherent modalities in them, while greatly simplifying the complexity of the analysis.

Interestingly, we propose a complexity measurement method based on time series symbolization combined with the sliding window technique, which is called equiprobable symbolization pattern entropy (EPSPE). Firstly, time series are symbolized according to the equiprobable segmentation. Then combines the sliding window technique to determine the symbolic patterns at different moments, and two adjacent patterns are regarded as a pattern pair. Finally, the number of occurrences of each pattern pair is counted and the corresponding probability is calculated to define the entropy value. The calculation results of the complexity of the simulation time series and the natural wind speed time series show that compared with the classic MSE, this novel method can be more concise and visually show the small internal differences of time series. Altogether, we make the following contributions:

- The symbolic processing of complex jumping data can greatly simplify the complexity of signal analysis while ensuring the accuracy of analysis.
- The pattern pairs and their probabilities are determined through the frequency and direction of different symbolic patterns conversions, which more accurately reflect the degree of interconnection and directionality of the internal primitives of the complex system.
- Since a finite number of symbolic pattern pairs are used as the events, the method is less affected by the length of time series analysis, which is suitable for big data time series analysis. The feasibility and advancement of this method are verified by testing

the classical Logistic signals and natural wind field signals.

The remainder of the paper is organized as follows. In the following section, we introduce some preliminaries. Sect. 3 presents our design. We provide complexity analysis experiment based on simulation time series in Sect. 4. Furthermore, the method proposed in this paper is used to analyze the natural wind speed time series signal in Sect. 5. Finally, Sect. 6 concludes this paper.

## 2 Preliminaries

### 2.1 Equiprobable symbolization

The method of equiprobable symbolization was first proposed by Lin et al. [11] in 2003 and has been widely applied in the field of data mining. The equiprobable symbolization method refers to the following: assuming that time series meets the Gaussian distribution, the critical points can be determined according to the Gaussian distribution to obtain a number of equal probability regions, and then combined with the segmental aggregation approximation principle for symbolic aggregation representation [12]. Its process is briefly described as follows.

For the time series  $x_i : 1 \leq i \leq N$ , firstly, they are sorted by amplitude. When the number of symbols  $n$  is given,  $n-1$  equal probability values are found (denoted as  $t_1, t_2, \dots, t_{n-1}$ ) As a threshold for equiprobable segmentation. According to the rules:

$$s_i = \begin{cases} 0 & x_i \leq, t_1, \\ 1 & t_1 < x_i \leq, t_2, \\ \vdots & \vdots \\ n-2 & t_{n-2} < x_i \leq, t_{n-1}, \\ n-1 & t_{n-1} \leq, x_i. \end{cases} \quad (1)$$

The original time series are converted into discrete symbol sequences  $s_i : 1 \leq i \leq N$ .

The equiprobable symbolization has three main benefits. Firstly, after symbolization, the influence of the probability distribution of the original sequence is eliminated. The symbol sequence fully reflects the relationship of the original time series, thus solving the problem that the entropy is affected by the probability distribution. Secondly, the symbolization result is not affected by extreme values, so it can well combat non-stationary

mutation interference. Thirdly, a variable resolution in the amplitude domain is realized, that is, more symbols are used in areas with dense amplitude distribution to improve resolution, and fewer symbols are used in the sparse area to reduce redundancy, thus improving the utilization of symbols. At the same time, the variable resolution also breaks through the linear constraints of traditional uniform symbolization.

### 2.2 Multiscale sample entropy

Generally, for a one-dimensional discrete time series, the calculation process of MSE is as follows.

a) Construct the coarse-grained time series  $y^\tau$  :

$$y_j^\tau = \frac{1}{\tau} \sum_{(j-1)\tau+1}^{j\tau} x_i, 1 \leq j \leq N/\tau, \quad (2)$$

where  $\tau$  is the scale factor and the length of the sequence is  $M = \text{int}(N/\tau)$ .

b) Construct a set of  $m$  dimensional vectors  $X^m(i) = \{y_{i+k}^\tau, 0 \leq k \leq m-1\}$ , where  $m$  is the embedded dimension.

c) For an arbitrary vector  $X^m(i)$ , calculate the distance to all the remaining vectors  $X^m(j)$ ,

$$d[X^m(i), X^m(j)] = \max_{k=0, \dots, m-1} (|y_{i+k}^\tau - y_{j+k}^\tau|). \quad (3)$$

d) Set a given tolerance  $r$ , then count the number  $B^m(i)$  of  $d[X^m(i), X^m(j)] < r$ , ( $i, j = 1 \sim M-m+1$ ;  $i \neq j$ ) and calculate the ratio to the total number  $C_i^m(r)$ :

$$C_i^m(r) = \frac{B^m(i)}{M-m+1}. \quad (4)$$

e) Calculate the mean value of  $C_i^m(r)$  as follows:

$$C^m(r) = \frac{1}{M-m} \sum_{i=1}^{M-m} C_i^m(r). \quad (5)$$

f) When  $m \rightarrow m+1$ , repeat steps b) ~ e) to obtain  $C^{m+1}(r)$ .

When the coarse-grained time series has a finite length, the entropy value  $\text{SampEn}(m, r, M)$  of the  $\tau$  scale is obtained as described above,

$$\text{SampEn}(m, r, M) = -\ln \left[ \frac{C^{m+1}(r)}{C^m(r)} \right]. \quad (6)$$

We repeat the above steps to calculate the entropies at different scales to obtain the curve of MSE.

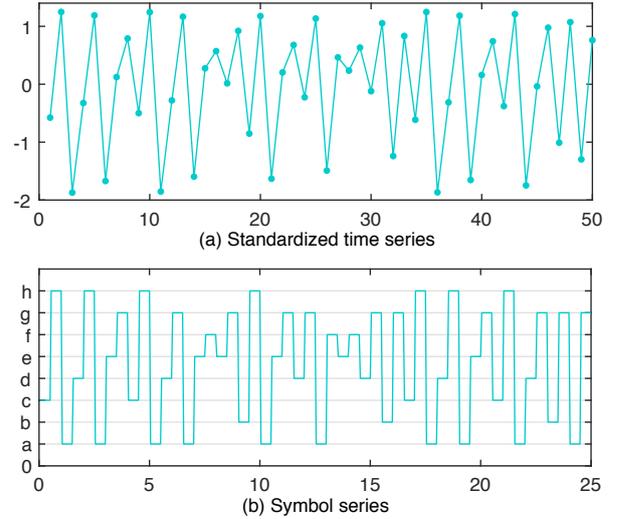
The larger the value of MSE, the more complex the time series is. The smaller the value of MSE, the more self-similar the time series is. It is noteworthy that the tolerance limitation of MSE is  $r = 0.1 \sim 0.25std(X)$ , so the MSE of time series at one scale is not the same as its sample entropy at the same scale, and therefore MSE can reflect the complexity of the system at multiple scales.

### 3 Equiprobable symbolization pattern entropy

Symbolization is the conversion of a complex time series into a discrete sequence consisting of a fixed number of numerical values or special symbols. This process allows the efficient and rapid capture of valuable quantitative information from complex systems, and can mitigate the effects of noise and reduce the complexity of calculations. In the real world we often encounter a class of complex signals (e.g., wind field signals), which are typically characterized by a jumping pattern of amplitude rather than the usual gradual variation. In view of the characteristics of this kind of signal, the multi-region amplitude equiprobable segmentation processing can greatly simplify the difficulty of analyzing the signal on the premise of guaranteeing a small loss of precision to the original signal.

Considering the intrinsic patterns of signals, this paper proposes equiprobable symbolization pattern entropy (EPSPE), a new complexity measurement method based on equiprobable symbolization combined with sliding window technique. This method symbolizes time series based on the equal probability distribution of jump data, and then uses sliding window technique to obtain a finite number of different symbolization patterns. The pattern pairs are determined by calculating the transition direction of the symbolization patterns and the frequency of the occurrence of pattern pairs determines the probability of a pattern pair. The final implementation is complexity measurement of complex signals. The method is simple to implement, has low computational complexity, strong robustness and is not affected by the length of the time series. The detailed steps are as follows.

a) The original time series  $\{x_i, i = 1, \dots, n\}$  can be Z-standardized using Eq. 7 to obtain a new standard-



**Fig. 1** Time series symbolic process.

ized time series  $\{y_i, i = 1, \dots, n\}$ . The processed time series satisfies a Gaussian distribution with a mean of 0 and a variance of 1.

Obtain  $y_i$  as

$$y_i = \frac{x_i - \bar{x}}{\alpha}, \quad (7)$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (8)$$

$$\alpha = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad (9)$$

$\bar{x}$  is the mean of the original time series and  $\alpha$  is the standard deviation of the original time series.

b) The time series is divided into a number of intervals according to the principle of equal probability, and each interval boundary point is determined. The value falling in each interval is represented by a letter symbol to realize the symbolic conversion of time series. The symbolic sequence  $S$ :

$$S = \{s_1 s_2 s_3 \dots s_k \dots s_n\}, s_k \in \{a, b, c, d, \dots\}. \quad (10)$$

As can be seen in Fig. 1, the symbolized sequences preserve the jump behavior of the original sequence, and these symbols describe the dynamics of the system in the simplest way possible.

c) A sliding window with a fixed length  $L$  is set. Starting with the first symbol of the symbolic sequence, slide from left to right to get a set of symbolic subsegment of length  $L$ . If we set  $L$  to be 4, we can get  $M : chad, haeg, hadg, aege, gbha, \dots$ , where each subsegment is a symbolic pattern, thus enabling symbolic pattern characterization of the original time series.

d) The pattern pairs are determined by the conversion relationship between the different symbolization patterns of length  $L$  in the symbolization subsegment set  $M$ , where a pattern pair is an event. For example,  $chad \rightarrow haeg$  is a pattern pair. The conversion frequency between the symbolization patterns (the number of occurrences of the event) is counted to calculate probability of transition between symbolic patterns. The total number of events is  $num$  and the number of event  $i$  is  $n_i$ , then the probability that event  $i$  is

$$P(i) = n_i / \sum_{i=1}^{Num} n_i. \quad (11)$$

e) The novel entropy  $E$  of the time series can be defined as:

$$E = - \sum_{i=1}^{Num} P(i) \ln P(i), \quad (12)$$

The advantages of this method are mainly reflected in the following aspects: 1) the symbolic processing of complex hopping data, which can greatly simplify the complexity of signal analysis while guaranteeing analysis accuracy; 2) through the conversion direction and frequency of different symbolic patterns to determine the events and their probabilities, to more accurately reflect the degree of correlation and directionality of the intrinsic patterns from the complex system more effectively; 3) due to the use of limited symbolic patterns, so this method is less affected by the length of the time series analysis, which can be applicable to big data time series analysis. The feasibility and advancement of the method in this paper are verified by testing the classical Logistic signal and the natural wind field signal.

#### 4 Analysis of simulated data

Chaos has captured the fancy of many scholars and researchers. The attractiveness of chaotic dynamics is

its ability to generate large movements which appear to be random, with greater frequency than linear models [7]. The simplest chaotic mapping operator, which was brought to the attention of scientists in 1976, is the Logistic map [6]. In order to verify the effectiveness and advancement of this method, the simulation time series of the Logistic mapping system is taken as an example. MSE and EPSPE proposed in this paper are used to analyze Logistic time series with different behaviors.

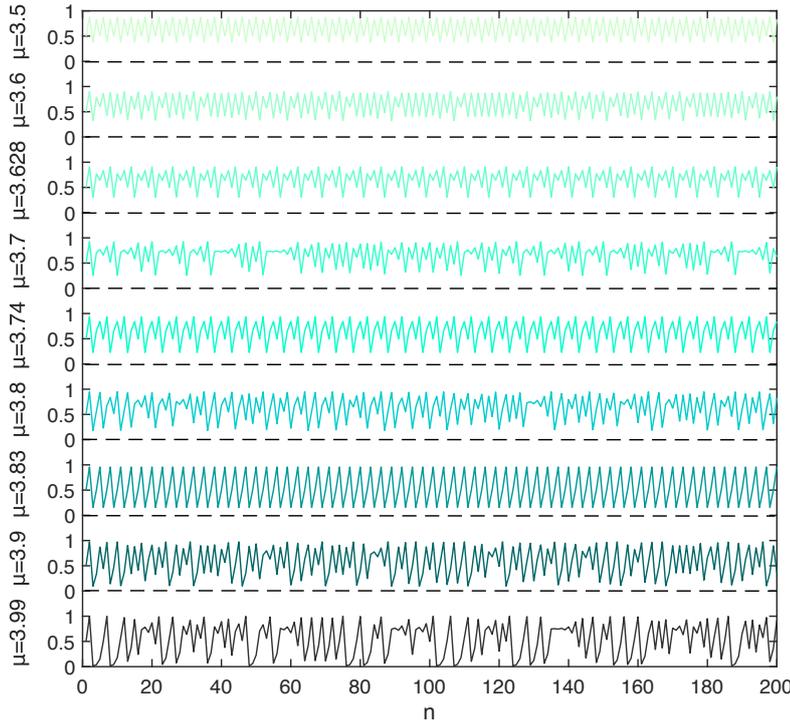
##### 4.1 Logistic mapping system

The Logistic mapping system is a classical model for generating chaotic time series that enter chaos through doubly-periodic bifurcation. The Logistic mapping system model equations are as follows.

$$x_{n+1} = \mu x_n (1 - x_n), \quad (13)$$

where  $x_n$  is the  $n$ -th chaotic number,  $n$  denotes the iteration number and  $\mu$  is the main parameter that affects the behavior of the system. Logistic mapping system includes all the properties of chaotic systems, such as self-similarity, ergodicity, semi-random motion, and sensitivity to initial conditions. A detailed explanation about chaotic properties can be found in [17]. On the other hand, the Logistic mapping system can provide more diversity than randomly selected initial solutions [8].

In this experiment, we set the initial value of the Logistic mapping system as 0.5 and the step size as 0.01. The parameter  $\mu$  is taken as 3.5, 3.6, 3.628, 3.7, 3.74, 3.8, 3.83, 3.9 and 3.99 respectively, so that we can get the Logistic time series of different behaviors with the length of 1500. Due to the large amount of data, only the first 200 data of each time series with different behaviors is plotted in Fig. 2 in order to show the time series characteristics of the different behaviors more clearly. Tab. 1 gives the behaviors of the Logistic mapping system for different parameters. It can be seen that when the parameter  $\mu$  is 3.5, 3.628, 3.74 and 3.83, the corresponding time series present typical periodical behaviors with fixed number of solutions, which are 4, 6, 5, and 3, respectively. While when the parameter  $\mu$  takes 3.6, 3.7, 3.8, 3.9, 3.99, the number of solutions of the system is large and the jumps of the time series are



**Fig. 2** Time series of Logistic system with different parameters.

**Table 1** The behavior of Logistic system with different parameters.

Logistic system parameter	3.5	3.6	3.628	3.7	3.74	3.8	3.83	3.9	3.99
behavior	periodicity	chaos	periodicity	chaos	periodicity	chaos	periodicity	chaos	chaos
Number of solutions	4	/	6	/	5	/	3	/	/

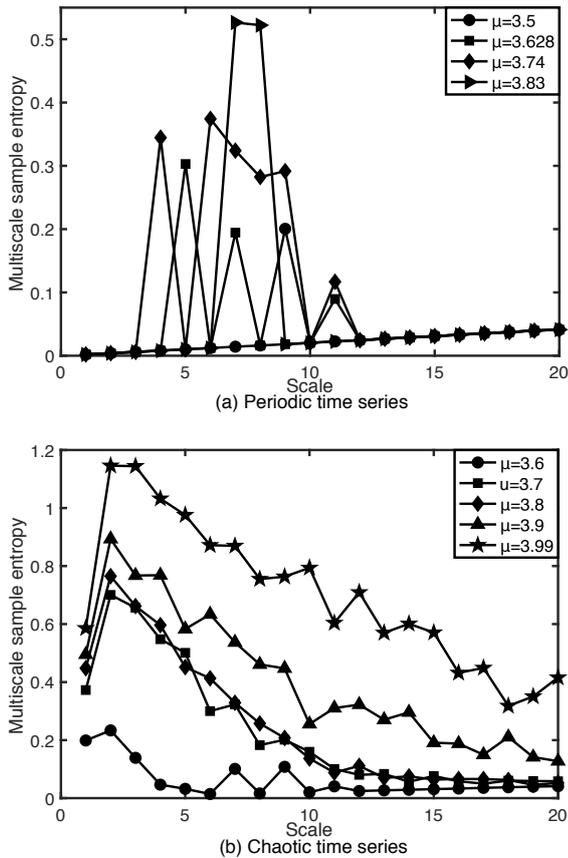
irregular, presenting typical chaotic behaviors, and the degree of chaos increases with increasing parameter.

## 4.2 Analysis and result

### 4.2.1 MSE analysis

According to the steps of the MSE algorithm in Sect. 2.2, the values of MSE of the Logistic time series with different behaviors are calculated with the embedding dimension  $m$  of 2, as shown in Fig. 3. Fig. 3(a) shows the MSE curves of the periodic time series from the Logistic system with different parameters. It can be seen that the curves are messy at multiple scales and cannot measure the complexity difference among the periodic time series. Fig. 3(b) shows the MSE curves of the chaotic

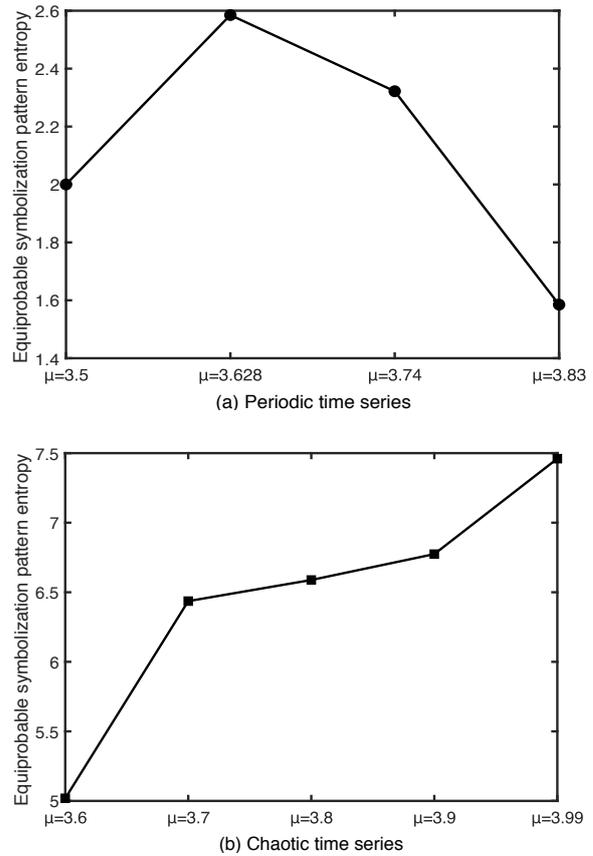
time series of the Logistic system with different parameters. When the parameter  $\mu = 3.6, 3.7, 3.8, 3.9$  and  $3.99$ , the chaotic intensity of the corresponding time series increases. It can be seen that at different scales, and the entropies of different time series show a decreasing trend. But when  $\mu = 3.7$  and  $3.8$ , the entropy values of the two corresponding time series cross or even coincide at multiple scales, which indicates that MSE cannot determine the complexity difference of Logistic time series with these two parameters. In short, MSE can roughly characterize the complexity differences between the chaotic signals, but it is impossible to identify the complexity differences between the signals with small differences. It has no discriminating ability to measure the complexity of the periodic signals.



**Fig. 3** Multiscale sample entropy of simulation time series with different parameter  $\mu =$  through the Logistic map. The left is the result of periodic time series. The right is the result of chaotic time series.

#### 4.2.2 EPSPE analysis

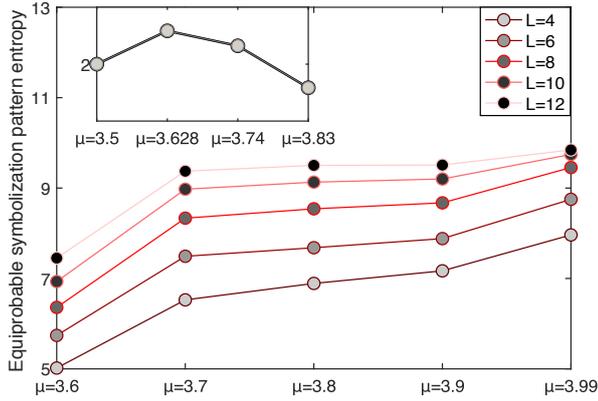
In accordance with the complexity measurement procedure described in Sect. 3, Logistic time series of different behaviors with a length of 1500 are mapped to symbolic time series. Here we set the length of the sliding window to 4, which can also be increased according to actual needs. In the following chapters, we will further analyze the influence of sliding window length on entropy. When the number of symbols is small, the distortion of the signal is large, and when the number of symbols is large, the number of possible symbolization patterns will increase sharply, which affects the analysis efficiency. In consideration of the complexity and accuracy of signal analysis, we passed a lot of tests



**Fig. 4** Equiprobable symbolization pattern entropy of simulation time series with different parameter  $\mu =$  through the Logistic map. The left is the result of periodic time series. The right is the result of chaotic time series.

and finally determined that the preferred symbolization level (the number of symbols) is 12. Fig. 4 shows the complexity measurement results of Logistic time series with different behaviors when the symbolization level is 12. Different symbol number settings do not change the relative relationship of the entropies of simulation time series, indicating that the method in this paper has good consistency.

When the parameter  $\mu$  is 3.5, 3.628, 3.74 and 3.83, the description in Tab. 1 is consistent with the complexity measurement trend of the time series given in Fig. 4(a). As we know, when the parameter  $\mu$  is taken as 3.6, 3.7, 3.8, 3.9, and 3.99, the corresponding time series are chaotic intensity increasing time series. The corresponding complexity measurement in Fig. 4(b) also



**Fig. 5** Equiprobable symbolization pattern entropy of simulation time series with different sliding window length  $L$ .

shows a gradual increasing trend, and is much larger than the complexity measurement of the periodic signal witch. It can be seen that the complexity measurement based on time series symbolization with sliding window technique can not only distinguish the periodic signals from the chaotic signals, but also has a good sensitivity to the small changes of the periodic and chaotic sequences. In addition, Fig. 4 shows that the analyzed time series length is 1500, it can also analyze time series with large amount of data and completely independent of the data length, thus making the analysis results more reliable and convincing.

#### 4.2.3 The length of sliding window analysis

When measuring the complexity of time series using a complexity measurement based on time series symbolization combined with a sliding window technique, the sliding window length  $L$  is a parameter that directly affects the size of the entropy value. Therefore, here we specifically analyze the effect of the sliding window length on the fineness of the entropy value and determine how the length of the sliding window is usually chosen, as shown in Fig. 5.

The entropy values of the Logistic mapping system with different parameters are shown in Fig. 5 for sliding window lengths of 4, 6, 8, 10, and 12, respectively. When the parameter  $\mu$  is 3.5, 3.628, 3.74, and 3.83, the corresponding time series are periodical time series with fixed number of solutions, and the entropy values is constant and does not change with the sliding window length. The entropy values are identical with

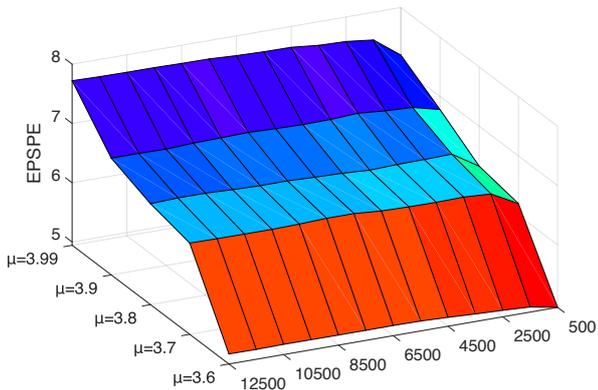
the number of solutions. When the parameter  $\mu$  is 3.6, 3.7, 3.8, 3.9 and 3.99, the entropy values of the corresponding chaotic time series increase with the increase of the sliding window length and is much larger than the entropy values of the periodic signals. For chaotic signals, as the sliding window length increases, the range of variation of the entropy values between the signals becomes smaller, i.e., the difference between the signals decreases. For example, when the sliding window is 12 and the parameter  $\mu$  is 3.7, 3.8 and 3.9, the entropy differences between the signals are negligible. Therefore, in order to finely characterize the differences between the complex signals and measure their complexity, the sliding window length is set to 4 in the subsequent analysis in this paper.

#### 4.2.4 The length of time series analysis

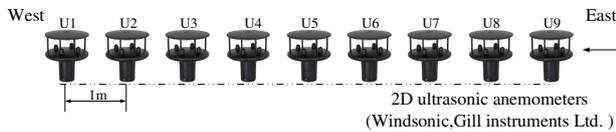
In order to analysis the influence of the length of the time series on EPSPE, it is performed on the chaotic time series of the Logistic system with different lengths. Fig. 6 shows the three-dimensional diagram of the EP-SPE of the chaotic signals at different lengths. First of all, it can be clearly observed that the EPSPE of the time series at different lengths has the same trend in all lengths, i.e., the values of EPSPE increase as the parameter  $\mu$  increases, which indicates that the complexity of the system is increasing. For each parameter, there is an obvious increase in the EPSPE between the time series length of 500 and 1500. When the time series length is greater than 1500, the EPSPE tends to be stable with the increase of the time series length, which shows that the proposed method has good robustness to the time series length. Moreover, in order to describe the complexity of the time series more comprehensively and accurately, it is suggested in this paper that the selected time series length is not less than 1500.

## 5 Application to natural wind field

In the previous section, it has been demonstrated that it is easy and intuitive to distinguish periodic signals from the chaotic signals. It is possible to distinguish the subtle differences among different chaotic signals. In order to further investigate its ability to analyze complex nonlinear and non-stationary natural signals, this sec-



**Fig. 6** Equiprobable symbolization pattern entropy of simulation data with different lengths .



**Fig. 7** Wind speed time series of nine two-dimensional ultrasonic anemometers.

**Table 2** The means of nine wind speed time series.

	U1	U2	U3	U4	U5	U6	U7	U8	U9
Mean value	2.48	2.57	2.52	2.53	2.59	2.61	2.65	2.68	2.71

tion will discuss the fineness of the proposed method in analyzing complex outdoor natural wind field signals.

### 5.1 Acquisition of natural wind field data

The main objective of this experiment is to test the ability of innovative complexity estimation methods to quantify and analyze natural time series signals. For this purpose, a set of natural wind field signals with a specific spatial position relationship was collected. Nine two-dimensional ultrasonic anemometers are arranged in a row with an interval of 1m adjacent to each other (as shown in Fig. 7), and the height above the ground is 0.6m. The data collection time is 1 hour, and the sampling frequency is 4Hz. The collected wind speed time series are shown in Fig. 8, and the mean value of each wind speed time series is listed in Tab. 2.

## 5.2 Complexity analysis of natural wind field data

Using the method proposed in this paper, the experimentally collected wind speed time series are measured for complexity. We quantify to analyze and mine the inherent patterns of the wind speed time series, and compare with the results of classical MSE analysis.

### 5.2.1 MSE analysis

The embedding dimension is set to 2. The MSE of wind speed time series at different spatial positions is calculated according to MSE algorithm in Sect. 2.2, as shown in Fig. 9.

Fig. 9 shows the MSE analysis results of the wind speed signals at different locations. For the experimental environment where the wind speed time series are collected in this paper, it has been shown that although the gaseous flow field changes continuously with time, the variation of wind speed in a certain region (usually within 10m) is approximately the same [10]. As shown in Fig. 8, the large scale fluctuations of the nine wind speed time series are basically the same, but there are some differences in the small scale fluctuations. The closer the collection locations are, the higher the similarities of the fluctuations are. It can be seen from Fig. 9 that the entropy values of the wind speed time series at different locations shows an increasing trend as the scale increases, i.e., the complexity of the wind speed time series increases as the scale increases. However, the entropy values of the time series at different locations are very close to each other at any scales, and the entropy curves of MSE are haphazardly intertwined. This means that the MSE values are not directly related to the spatial location of the anemometers. In addition, the computational complexity of the MSE is large, and there is a limit to the length of the time series.

### 5.2.2 EPSPE analysis

According to the complexity measurement procedure described in Sect. 3, we set the symbolization level to 12 and the sliding window length to 4. A set of outdoor natural wind field signals with a specific spatial position relationship are analyzed. The complexity measurement of the wind speed time series collected by nine 2D anemometers is shown in Fig. 10.

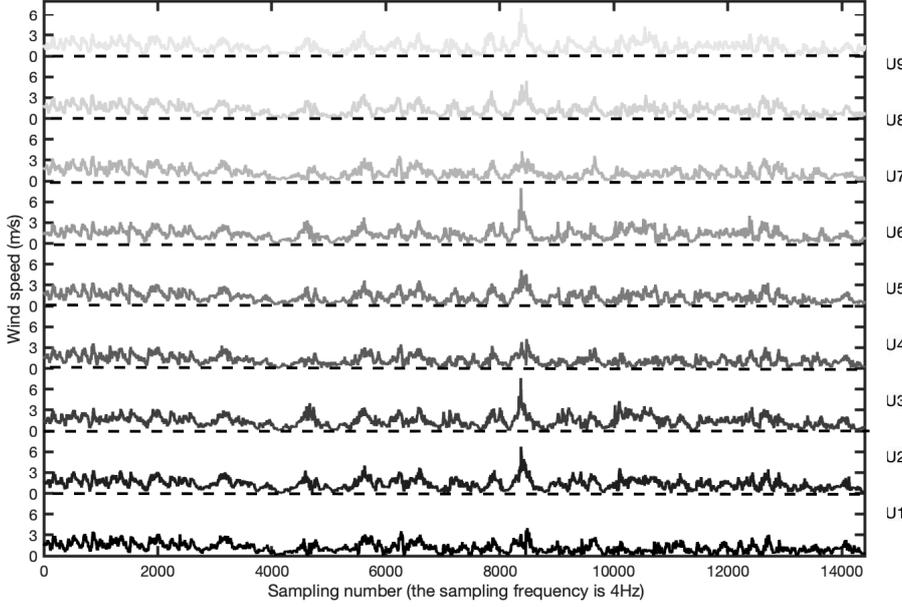


Fig. 8 Wind speed time series of nine two-dimensional ultrasonic anemometers.

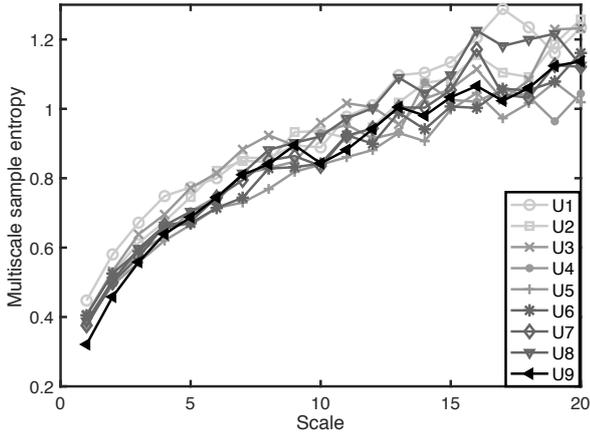


Fig. 9 Multiscale sample entropy of wind speed time series

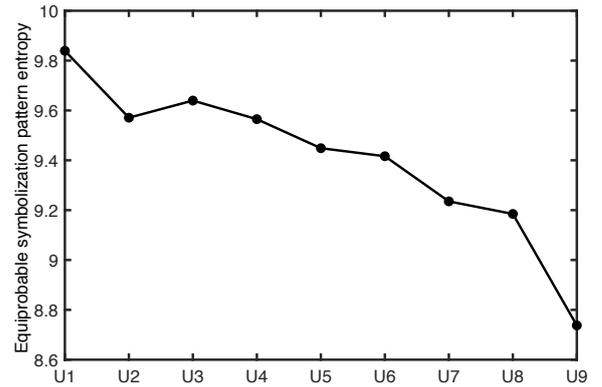


Fig. 10 Equiprobable symbolization pattern entropy of wind speed time series.

Fig. 10 shows the results of the complexity measurement analysis based on the time series symbolization combined with the sliding window technique proposed in this paper. The complexity measurement parameter, entropy, shows a consistent decay trend with the spatial locations of the arranged anemometers. The statistical results of the wind rose diagram from the nine acquisition locations show that the main wind direction of the test experiment is northeast. And the energy gradually

decreases with the direction of wind flow, i.e., the average wind speed of the wind time series collected by anemometer U9 is the largest and gradually decreases from east to west (with the exception of the U2 position). The research results of reference [10] show that the larger the average wind speed over a small local spatial range (usually less than 10m) the weaker the random fluctuations in the amplitude and frequency of the airflow, i.e. the less complexity the fluctuations are.

From the data presented in Fig. 10, it can be seen visually that the large-scale fluctuations of the U9 and U1 time series are almost the same, but the small-scale fluctuations of the U9 time series (maximum average wind speed) are significantly weaker than those of the U1 time series (minimum average wind speed). So the complexity of the U9 time series is smaller than that of the U1 time series. The entropy proposed in this paper reflects the number of symbolic patterns in the time series, and the more symbolic patterns indicate the more complex time series. While the entropy value of the time series from U1 is the largest due to the largest complexity of the fluctuation, and the entropy values of the other positions U3~U9 shows a consistent decaying trend due to the gradually decreasing complexity. Due to the abnormal increase of the average wind speed at U2, the abnormal entropy value is also reasonable. The above analysis shows that the new method proposed in this paper can accurately reflect the similarity and complexity differences of the analyzed signals. In applications, such as wind energy prediction in power systems and wind field zoning in meteorology, similarity and complexity analysis is often performed on wind field signals collected at different spatial locations. Since the natural wind field data is a kind of complex nonlinear and non-stationary time series, there is no effective solution to measure the similarity and complexity of these time series. The method proposed in this paper provides an effective tool to solve this problem.

## 6 Conclusion

In this paper, a new method for complexity measurement based on time series symbolization combined with the sliding window technique is presented. The standardized time series is firstly processed with equal probability to transform the original time series into a more concise symbolic sequence. Then a series of symbolic patterns are extracted in combination with sliding window technique, and the conversion direction and frequency of the symbolic patterns are used as events and the probability of events, thus realizing the complexity estimation of complex signals. Time series analysis of the Logistic mapping system under different parameters shows that the complexity estimation method proposed in this paper can not only intuitively distinguish between periodic and chaotic signals, but also accurately

reflect the subtle changes in the periodic and chaotic time series. It is worth proposing that MSE exhibits significant deficiencies, whether it is to distinguish classical chaotic signals from periodic signals, or to measure the variability of natural signals. Finally, the analysis of the complexity measurement of the wind speed time series collected by the regularly arranged anemometers shows that EPSPE can predict the spatial location proximity of the anemometers more accurately. These interesting findings suggest that pattern pairs from EPSPE are potentially valuable characteristics of wind signals, which will have broad applications in researches such as wind power prediction, wind pattern classification and wind field dynamic analysis.

## References

1. Bandt, C., Pompe, B.: Permutation entropy: A natural complexity measure for time series. *Physical Review Letters* **88**(17), 174102 (2002)
2. Bian, S., Shang, P.: Refined two-index entropy and multiscale analysis for complex system. *Communications in Nonlinear Science Numerical Simulation* **39**(10), 233–247 (2016)
3. Costa, M., Goldberger, A.L., Peng, C.K.: Multiscale entropy analysis of complex physiologic time series. *Physical Review Letters* **89**(6), 705–708 (2007)
4. Cysarz, D., Bettermann, H., Leeuwen, P.V.: Entropies of short binary sequences in heart period dynamics. *American Journal of Physiology Heart Circulatory Physiology* **278**(6), H2163–72 (2000)
5. Cysarz, D., Porta, A., Montano, N., Leeuwen, P.V., Kurths, J., Wessel, N.: Quantifying heart rate dynamics using different approaches of symbolic dynamics. *The European Physical Journal Special Topics* **222**(2), 487–500 (2013)
6. D., G.: Chaos in economics and finance. *Annual Reviews in Control* **33**(1), 89–93 (2009)
7. Hsieh, D.A.: Chaos and nonlinear dynamics: Application to financial markets. *Journal of Finance* **46**(5), 1839–1877 (1991)
8. Kazem, A., Sharifi, E., Hussain, F.K., Saberi, M., Hussain, O.K.: Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing Journal* **13**(2), 947–958 (2013)
9. Kolmogorov, A.N.: A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. In: *Doklady Akademii Nauk*, vol. 119, pp. 861–864. Russian Academy of Sciences (1958)
10. Li, J.G., Meng, Q.H., Wang, Y., Zeng, M.: Odor source localization using a mobile robot in outdoor airflow environments with a particle filter algorithm. *Autonomous Robots* **30**(3), 281–292 (2011)

11. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: 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 (2003)
12. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: A novel symbolic representation of time series. *Data Mining Knowledge Discovery* **15**(2), 107–144 (2007)
13. Mizuno, T., Takahashi, T., Cho, R.Y., Kikuchi, M., Murata, T., Takahashi, K., Wada, Y.: Assessment of EEG dynamical complexity in Alzheimer’s disease using multiscale entropy. *Clinical Neurophysiology* **121**(9), 1438–1446 (2010)
14. Pincus, Steve: Approximate entropy (ApEn) as a complexity measure. *Chaos* **5**(1), 110–117 (1998)
15. Richman, J.S., Randall, M.J.: Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology* **278**(6), 2039–2049 (2000)
16. Rong, L., Shang, P.: Topological entropy and geometric entropy and their application to the horizontal visibility graph for financial time series. *Nonlinear Dynamics* **92**, 41–58 (2018)
17. Sauer, T., Yorke, J.A., Casdagli, M.: Embedology. *Journal of Statistical Physics* **65**(3-4), 579–616 (1991)
18. Song, A., Huang, X., Si, J., Ning, X.: Optimum parameters setting in symbolic dynamics of heart rate variability analysis. *Acta Physica Sinica* **60**(2), 120–127 (2011)
19. Takahashi, T., Cho, R.Y., Mizuno, T., Kikuchi, M., Murata, T., Takahashi, K., Wada, Y.: Antipsychotics reverse abnormal EEG complexity in drug-naïve schizophrenia: A multiscale entropy analysis. *Neuroimage* **51**(1), 173–182 (2010)
20. Vargas, M., Fuertes, G., Alfaro, M., Gatica, G., Gutierrez, S., Peralta, M.: The effect of entropy on the performance of modified genetic algorithm using earthquake and wind time series. *Complexity* **2018**, 1–13 (2018)
21. Wang, G., Liu, Z., Feng, Y., Li, J., Dong, H., Wang, D., Li, J., Yan, N., Liu, T., Yan, X.: Monitoring the depth of anesthesia through the use of cerebral hemodynamic measurements based on sample entropy algorithm. *IEEE Transactions on Biomedical Engineering* **67**(3), 807–816 (2020)
22. Yin, Y., Shang, P.: Weighted permutation entropy based on different symbolic approaches for financial time series. *Physica A Statistical Mechanics Its Applications* **443**(2016), 137–148 (2016)
23. Zeng, M., Wang, E., Zhao, M., Meng, Q.: Directed weighted complex networks based on time series symbolic pattern representation. *Acta Physica Sinica* **66**(21), 265–275 (2017)
24. Zhang, X., Liang, J.: Chaotic time series prediction model of wind power based on ensemble empirical mode decomposition-approximate entropy and reservoir. *Acta Physica Sinica* **62**(5), 50505–50505 (2013)
25. Zhang, Y., Shang, P.: The complexity–entropy causality plane based on multivariate multiscale distribution entropy of traffic time series. *Nonlinear Dynamics* **95**(1), 617–629 (2019)
26. Zhao, X., Shang, P., Huang, J.: Permutation complexity and dependence measures of time series. *Epl* **102**(4), 40005 (2013)

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [DataAvailabilityStatement.pdf](#)