

Comparative Analysis of Multi-fractal Data Missing Processing Methods

Lai Simin¹, Wan Li^{1,2,*}, Zeng Xiangjian¹

¹School of Mathematics and Information Science, Guangzhou University, Guangzhou, China

²Key Laboratory of Mathematics and Interdisciplinary Sciences of Guangdong Higher Education Institutes, Guangzhou University, Guangzhou, China

Email address:

1994650450@qq.com (Lai Simin), wanli03100@21cn.com (Wan Li), 1747587914@qq.com (Zeng Xiangjian)

*Corresponding author

To cite this article:

Lai Simin, Wan Li, Zeng Xiangjian. Comparative Analysis of Multi-fractal Data Missing Processing Methods. *Applied and Computational Mathematics*. Vol. 8, No. 2, 2019, pp. 44-49. doi: 10.11648/j.acm.20190802.14

Received: July 4, 2019; **Accepted:** July 24, 2019; **Published:** July 29, 2019

Abstract: Data missing often affects the characteristics of the sequence. Using appropriate methods to process the missing data is the premise and guarantee to obtain high quality information. In this study, a fractal interpolation method is proposed to fill the missing data with self-similar feature sequences. Two sets of binomial multifractal sequences with parameters of 0.25 and 0.35 are taken as the research objects, and the Hurst index value of the sequence after filling processing is calculated by MF-DMA, which verifies the practicability of the fractal interpolation filling method. At the same time, the method is applied to multi-fractal sequences with missing rates of 10%, 15% and 20% respectively, and compared with the filling effects of deletion method and random filling method, then, the applicability of the three methods is obtained. The results show that, for binomial multifractal sequences with different missing ratios, the Hurst index of the sequence processed by fractal interpolation has the highest degree of fitting with the theoretical value, its effect of repairing the fractal sequence is better than the other two methods, and has a good application prospect.

Keywords: Multi-fractal, Fractal Interpolation Filling, MF-DMA, Hurst Index

1. Introduction

In the course of investigations and research, there are often cases of data loss. Missing data often increases the difficulty of experimental analysis and even leads to unreliable experimental results. In this regard, many scholars at home and abroad have discussed the treatment of missing data. In 1978, Rubin proposed a multi-interpolation method for dealing with missing discrete data. This method replaces the missing data by creating an interpolation vector, which improves the shortcomings of the single interpolation [1]; Mao Qunxia in paper [2] discussed the application of multi-interpolation method in the case of missing data in survey data by data simulation technology, pointing out that multi-interpolation is limited to the filling of discrete data and can not be used to repair time series with correlation; In 2013, Tan et al. [3] proposed an attribution method based on tensor decomposition to repair the lost traffic flow data.

Although this method can effectively restore the time dependence of traffic flow itself, it is limited to processing the sequence with less missing data. In 2015, Dohoo compared the multiple attribution method with the complete case analysis method based on missing disease control data, and obtained the applicability and limitations of the multiple attribution method [4]. In 2018, Xia Zhiling and others filled the missing data with the optimal hybrid substitution method, solved the problem of missing data in water environment, and achieved good restoration results. In summary, these methods are mainly used for data processing in specific areas, without considering the non-linear characteristics of the data sequence itself. With the establishment and development of fractal theory, a large number of studies have shown that many data series in real life have self-similar characteristics, such as financial investment, commodity prices, network traffic and traffic flow time series [6-9]. Research on the methods and techniques that can effectively extract the information hidden behind these data is of great practical

significance for revealing the intrinsic characteristics and development laws of things and explaining the relevant principles of complex phenomena. At present, deletion method and single interpolation method are the main methods to deal with the missing segment data of non-stationary self-similar sequence. Single interpolation method includes mean interpolation, random interpolation and regression interpolation. It has changed the disadvantage of neglecting missing values in traditional methods, so that all kinds of statistical analysis can be completed after interpolation. The data set is expanded, but the disadvantage of single interpolation method is that it will change the distribution characteristics of sample data and lead to the deviation of statistical results. Therefore, the research on the processing method of non-linear missing sequence needs to be further improved.

This paper takes the typical binomial multifractal sequence as the research object. On the premise of random missing of original data, fractal interpolation method is proposed to repair missing fractal sequences. The Hurst index of the sequence is calculated by MF-DMA to verify the validity of the fractal interpolation method, and Compared with traditional deletion method and random filling method, the advantages, disadvantages and applicability of the three data processing methods under different missing rates of binomial multiple sequence are analyzed, so the most effective processing method is obtained. It provides valuable reference for the further study of fractal sequences.

2. Theory and Method

2.1. Algorithm of MF-DMA

MF-DMA is an effective method for analyzing the multifractal characteristics of signal sequences [10-12]. Its main algorithm steps are as follows.

In the first step, assuming the signal sequence is $x(i), t=1,2,\dots,N$ (N is the length of the signal sequence), calculate the accumulated sequence by

$$y(t) = \sum_{i=1}^t x(i), \quad t=1,2,3,\dots,N \quad (1)$$

Taking scale n as the size of sliding window, the moving average function is calculated by

$$\tilde{y}(t) = \frac{1}{n} \sum_{k=\lfloor (n-1)\theta \rfloor}^{\lceil (n-1)(1-\theta) \rceil} y(t-k) \quad (2)$$

Here, $t \in (n - \lfloor (n-1)\theta \rfloor, N - \lfloor (n-1)\theta \rfloor)$, $\lfloor x \rfloor$ represents the largest integer not greater than x , $\lceil x \rceil$ represents the smallest integer not less than x . $\theta \in [0,1]$ is the position parameter of sliding window. Gu and Zhou in paper [10] have proved that $\theta=0$ is a good estimation of moving average function. Therefore, $\theta=0$ was applied in this research.

In the third step, the residual sequence is obtained by subtracting the moving average function from the cumulative

sum sequence.

$$\varepsilon(t) = y(t) - \tilde{y}(t), \quad t \in (n - \lfloor (n-1)\theta \rfloor, N - \lfloor (n-1)\theta \rfloor) \quad (3)$$

In the fourth step, the sequence is divided into N_n non-overlapping interval segments by an interval of n , and $N_n = \lfloor N/n \rfloor$. Each interval is expressed as $\varepsilon(i) = \varepsilon((v-1)n + i), i=1,2,3,\dots,n$. Then, by calculating the square mean of the entire interval sequence to obtained the mean square function of the v th interval segment

$$F_v^2(n) = \frac{1}{n} \sum_{i=1}^n \varepsilon_v^2(i) \quad (4)$$

The fifth step is to calculate the q -order fluctuation function.

$$F_q(n) = \left\{ \frac{1}{N_n} \sum_{v=1}^{N_n} F_v^q(n) \right\}^{1/q}, \quad q \neq 0 \quad (5)$$

$$\ln[F_q(n)] = \frac{1}{N_n} \sum_{v=1}^{N_n} \ln[F_v(n)], \quad q \neq 0 \quad (6)$$

In the sixth step, the scaling behavior of fluctuation function is investigated by changing the value of window scale n . If the data sequence is long-range power-law-dependent, there are

$$F_q(n) \propto n^{H(q)} \quad (7)$$

Drawing logarithmic graphs of $F_q(n)$ and n , the slope fitted by least square method is the generalized Hurst exponent of the multifractal series. It is an important index to measure whether the data sequence has fractal characteristics and its long-range correlation. When $H(q)$ changes with q value, the sequence is multifractal, and the larger the variation range, the stronger the autocorrelation trend of the sequence; if $H(q)$ is a fixed constant independent of q value, it means the sequence is single fractal.

2.2. Brief Introduction of Data Missing Processing Method

The delete method is to cancel the missing data segment from the original sequence directly, and get a new sequence whose length is less than the original sequence.

Random filling method is based on the idea of random sampling, a corresponding number of data are randomly selected from the existing data of the sequence to fill the missing sequence segments, and a complete new sequence is obtained.

For the fractal interpolation method, the concept of fractal interpolation was proposed by Barnsley in 1986, assuming that the data set consisting of a series of interpolation points is $\{(x_i, y_i), i=0,1,2,\dots,N\} \in R^2$, here $x_0 < x_1 < \dots < x_N$. The fractal interpolation function is a continuous function that is interpolated at these points, and its graph is an attraction of the Iterated Function System (IFS). The specific construction process [13-15] is as follow.

Let each function in the iteration function system be an

affine transformation, written as $W_i: R^2 \rightarrow R^2, i=1, 2, \dots, N$.

$$\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow W_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_i & 0 \\ c_i & d_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix}, \quad i=1, 2, \dots, N \quad (8)$$

The transformation satisfies:

$$W_i \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} x_{i-1} \\ y_{i-1} \end{pmatrix}, \quad W_i \begin{pmatrix} x_N \\ y_N \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix}, \quad i=1, 2, \dots, N \quad (9)$$

According to equation (9) and equation (10), for each i , the five constants of function W_i satisfy the following equation.

$$\begin{cases} a_i x_0 + e_i = x_{i-1} \\ a_i x_N + e_i = x_i \\ c_i x_0 + d_i y_0 + f_i = y_{i-1} \\ c_i x_N + d_i y_N + f_i = y_i \end{cases} \quad (10)$$

d_i in the formula is a free variable, which satisfies $0 < d_i < 1$ ($i=1, 2, \dots, N$). It is called the vertical scale factor. The value of d_i in the experiment is 0.5. The coefficients a_i, e_i, c_i, f_i ($i=1, 2, \dots, N$) are completely determined by the interpolation data $\{x_n | n=0, 1, 2, \dots, N\}$.

Fractal interpolation filling method is a method of filling the missing values of the original sequence by generating a series of new points between the interpolated data using the fractal function generated by the recursion of the upper iteration.

3. Analysis of Experimental Process and Results

3.1. Model Construction and Parameter Selection

Binomial multifractal sequence is a typical multifractal time series, which is the representative of non-stationary signal sequence [16-18]. The multifractal sequence with length of (n to 10) can be constructed by the following formula.

$$x(i) = a^{n-n(i-1)} (1-a)^{n(i-1)}, \quad i=1, 2, 3, \dots, N \quad (11)$$

Here, a is the main parameter to generate the sequence, $n(i)$ denotes the number of values i in binary 1.

Figure 1 shows the multi-fractal sequence images with parameter a of 0.2, 0.25, 0.35, 0.4 and lengths of 1024 generated in MATLAB 7.0 environment.

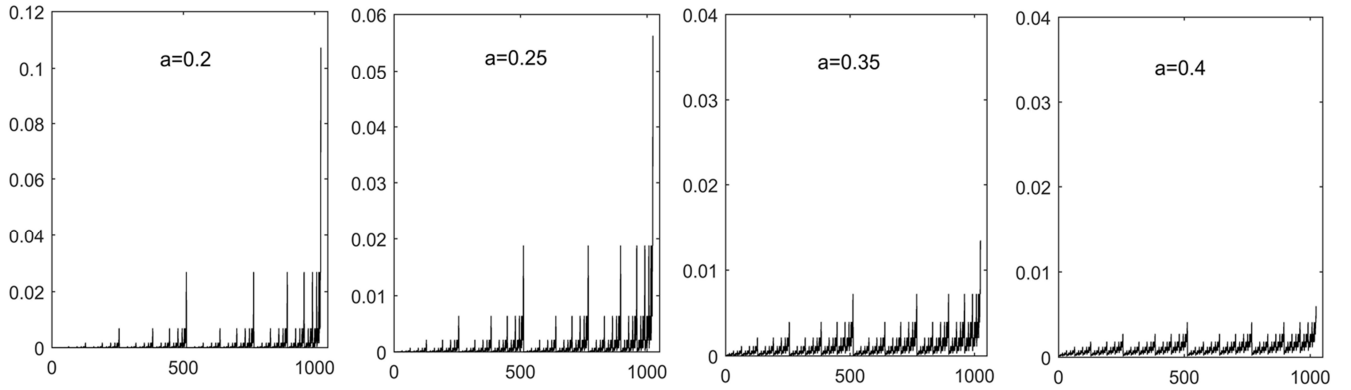


Figure 1. Binomial multifractal sequences generated by different parameters.

According to Figure 1, with the increase of parameters, the difference between the two ends of the sequence decreases and the fluctuation decreases gradually. Because the fluctuation range of the data is too large or too small, it is not conducive to analysis and research, so this paper chooses parameters and binomial multifractal series of length 1024 to carry out experiments.

3.2. Results and Analysis

According to the steps introduced in 2.2, let the parameters

$a=0.25$ and $a=0.35$. Two sets of multifractal sequences with 1024 lengths are obtained by self-programming with MATLAB. The proportion of continuous missing data in the sequences is 10%, 15% and 20% respectively. Based on MF-DMA, the missing data are processed by direct removal method, random sampling filling method and fractal interpolation filling method. The Hurst index of the sequence is obtained as follows:

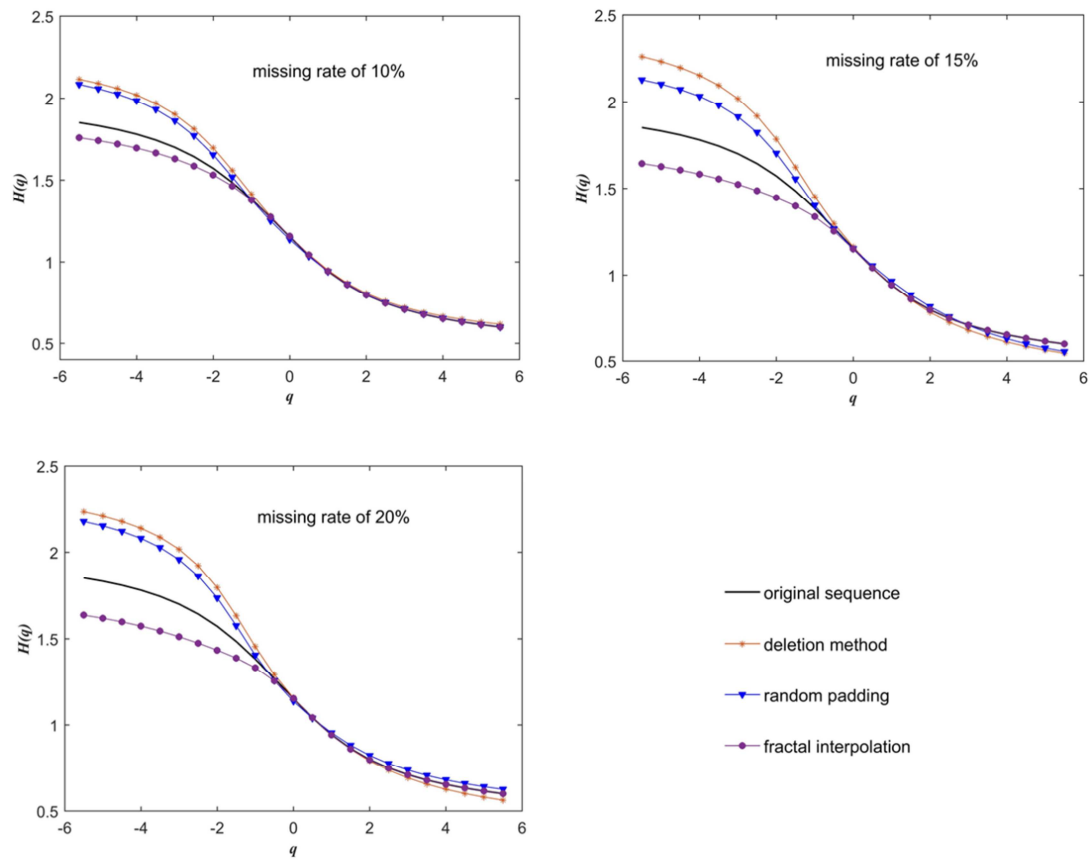


Figure 2. The sequence with parameter 0.25 corresponds to the Hurst exponents obtained by different processing methods.

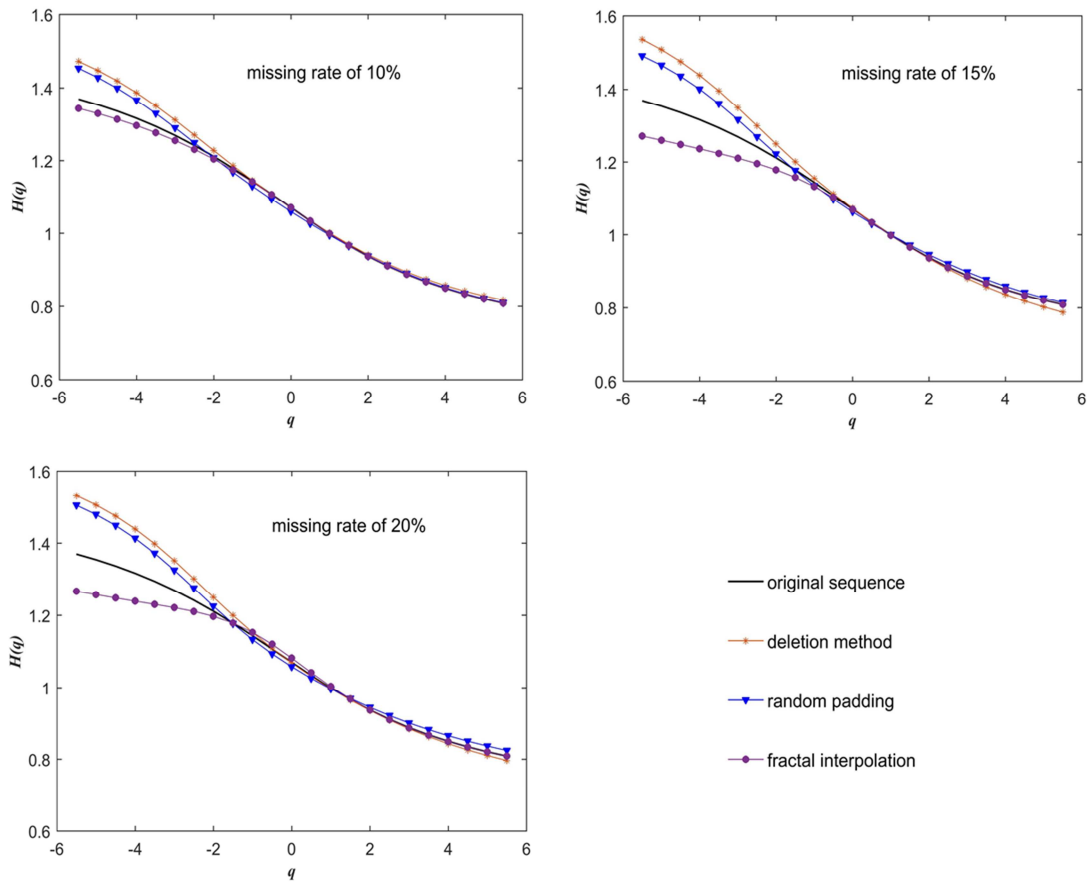


Figure 3. The sequence with parameter 0.35 corresponds to the Hurst exponents obtained by different processing methods.

The trend of Hurst exponent in the graph reflects that the sequence repaired by the above three methods still has multifractal characteristics, and the fitting degree of Hurst values of each repaired sequence is different from that of the theoretical value. Among them, the sequence repaired by fractal interpolation is similar to the original sequence's Hurst values, regardless of the missing ratio of the data being 10%, 15% or 20%. In addition, the results show that the Hurst values of the fractal interpolation sequence are lower than the theoretical values, which indicates that the fractal interpolation method can deal with the missing values of fractal sequence. Eliminate the role of trends.

By comparing Figures 2 and 3, it is found that the Hurst value of binomial multifractal sequence generated by parameter $\alpha=0.35$ changes more smoothly, which indicates

that the autocorrelation trend intensity of the multifractal sequence is lower. At this time, fractal interpolation filling method, random sampling filling method and direct removal method are used to repair the missing items of the multifractal sequence. The Hurst index of F-DMA is close to that of the original sequence.

In order to further compare the effectiveness of the three methods introduced in this paper for missing sequence processing, we calculated the mean square error (MSE) of Hurst values of each repair sequence corresponding to the missing ratio of 10%, 15% and 20% for the two groups of polynomial multifractal sequences with parameters $\alpha=0.25$ and $\alpha=0.35$, respectively. The variance value reflecting the stability of error variation. The results are shown in the following Tables.

Table 1. $\alpha=0.25$, the error between the obtained Hurst index value and the theoretical value.

methods	missing rate of 10%	missing rate of 15%	missing rate of 20%	standard deviation
Deletion method	0.0174	0.0434	0.0405	0.00014
Random padding	0.0122	0.0193	0.0274	0.00004
Fractal interpolation	0.0021	0.0123	0.0136	0.00002

Table 2. $\alpha=0.35$, the error between the obtained Hurst index value and the theoretical value.

methods	missing rate of 10%	missing rate of 15%	missing rate of 20%	standard deviation
Deletion method	0.0016	0.0048	0.0046	0.000002
Random padding	0.0009	0.0023	0.0030	0.0000008
Random padding	0.0001	0.0020	0.0018	0.0000007

Datas in Table 1 and Table 2 clearly reflect the fitting errors between the repair series and the original sequence. It can be seen that, on the one hand, the deviation between Hurst value and theoretical value obtained by fractal interpolation is the smallest under various missing ratios, which further shows that fractal interpolation fills the missing multifractal sequence with the best repair effect compared with the other two filling methods; on the other hand, the standard deviation in Table 2 is obviously smaller than the standard deviation in Table 1, indicating that the missing value has the best repair effect on the two kinds of fractal sequences. The sequence model generated by parameters has different influence degrees. For the multifractal sequence with $\alpha = 0.35$ parameter, it is advisable to use the above three methods to repair the missing values of the sequence. The magnitude of variance shows that the fractal interpolation method has better stability when dealing with missing values of fractal sequence.

4. Conclusion

In order to effectively estimate the missing values of fractal data sequences and improve the efficiency of data analysis, this paper proposes a fractal interpolation method, which is applied to filling missing data of binomial multifractal sequences with different parameters and different missing ratios. Based on MF-DMA, the Hurst index of new method is compared with the deletion method and the random padding method, and the validity of the fractal interpolation padding is tested. The results show that the fractal interpolation filling sequence has a smaller error with the original sequence, and

its repair result is more stable and reliable, which provides a new idea and method for solving the problem of data missing from self-similar sequence.

Acknowledgements

Thanks to the editor for reviewing this paper and making detailed suggestions to help me improve the quality of the work. This work was supported by the National Natural Science Foundation of China (Grant No. 41872246); Innovation Research for the postgraduate of Guangzhou University (2018GDJC-D27).

References

- [1] Rubin D B. Multiple Imputation for Nonresponse in Surveys [M]. John Wiley, 1987.
- [2] Mao Qunxia, Li Xiaosong. Comparison of Multiple Filling Method and Ad Hoc Method for Processing Missing Values in Simulated Longitudinal Data Sets [J]. Modern Preventive Medicine, 2005 (04): 310-312.
- [3] Tan H, Feng G, Feng J, et al. A tensor-based method for missing traffic data completion [J]. Transportation Research Part C: Emerging Technologies, 2013, 28: 15-27.
- [4] Dohoo I R. Dealing with deficient and missing data.[J]. Preventive Veterinary Medicine, 2015, 122 (1-2): 221-228.
- [5] Xia Liling, Zhu Yuelong. Optimizing the method of water environment data missing processing [J]. Hydropower and Energy Science, 2018, 36 (04): 158-161+85.

- [6] Bouchaud J P, Potters M, Meyer M. Apparent multifractality in financial time series [J]. The European Physical Journal, 2000, 13 (3): 595-599.
- [7] Chen Peng, Zheng Manxian. Multifractal Characteristics and Conduction Effect of International Commodity Price Fluctuation [J]. Price Theory and Practice, 2018 (10): 81-84.
- [8] Zhang Yong, Guan Wei. Multifractal Analysis of Traffic Flow Time Series [J]. Computer Engineering and Applications, 2010, 46 (29): 23-25.
- [9] Chen Y, Xiang Z, Dong Y, et al. Multi-Fractal Characteristics of Mobile Node's Traffic in Wireless Mesh Network with AODV and DSDV Routing Protocols [J]. Wireless Personal Communications, 2011, 58 (4): 741-757.
- [10] Gu G F, Zhou W X. Detrending moving average algorithm for multifractals [J]. Physical Review E. 2010, 82 (1): 011136.
- [11] Mali P. Multifractal detrended moving average analysis of global temperature records [J]. Journal of Statistical Mechanics Theory & Experiment, 2016, 2016 (1): 013201.
- [12] Li Q, Cao G, Xu W. Relationship research between meteorological disasters and stock markets based on a multifractal detrending moving average algorithm [J]. International Journal of Modern Physics B, 2018, 32 (01): 1.
- [13] Amo E D, Carrillo M D, Sánchez J F. PCF self-similar sets and fractal interpolation [J]. Mathematics & Computers in Simulation, 2013, 92 (6): 28-39.
- [14] Jiang P, Liu F, Wang J, Song Y. Cuckoo search-designated fractal interpolation functions with winner combination for estimating missing values in time series [J]. Applied Mathematical Modelling, 2016, 40 (23-24): 9692-9718.
- [15] Fu Y, Zheng Z, Xiao R, Shi H. Comparison of two fractal interpolation methods [J]. Physica A: Statistical Mechanics and its Applications, 2017, 469.
- [16] Kantelhardt J W, Zschiegner S A, Koscielny-Bunde E, et al. Multifractal detrended fluctuation analysis of nonstationary time series [J]. Physica A: Statistical Mechanics and its Applications, 2002.
- [17] Mali P, Mukhopadhyay A. Multifractal characterization of gold market: A multifractal detrended fluctuation analysis [J]. Physica A: Statistical Mechanics and its Applications, 2014, 413: 361-372.
- [18] Xie W J, Han R Q, Jiang Z Q, et al. Analytic degree distributions of horizontal visibility graphs mapped from unrelated random series and multifractal binomial measures [J]. EPL (Europhysics Letters), 2017, 119 (4): 48008.