Prediction of Electric Load Neural Network Prediction Model for Big Data

Xinhao He *

School of economics and management, Southeast University, Nanjing, Jiangsu, 211189

* Corresponding Author Email: 936107954@qq.com

Abstract. The characteristics of non-stationarity, non-linearity, and long-memory of the stock index series make it challenging to forecast. In order to improve the prediction accuracy of the existing models, this paper proposes a new ensemble prediction model of CSI 300 index returns by integrating variational mode technique and long short-term memory, which consists of particle swarm optimization (PSO), variational mode decomposition (VMD), sample entropy (SE) and long and short-term memory networks (LSTM). The stock index prices of the CSI 300 index during three consecutive months from October 2021 to December 2021 are selected as the research sample, and 48 sets of data are obtained in the 5-minute set every day, totaling more than 2,900 trading data as the modeling object, and the volume-weighted average price (VWAP) index is introduced to portray investor behavior. The experimental comparison shows that this method gives the smallest root mean square error and the best prediction fit, which significantly outperforms the existing portfolio model and has significant prediction advantages.

Keywords: Stock Index Prediction; Particle Swarm Optimization; Variational Mode Decomposition; Long Short-Term Memory

1. Introduction

With the development of the economy, more and more enterprises and individuals have joined the stock market, promoting the balance between supply and demand. The fluctuations in the market are closely related to the interests of both parties. Since the stock market information is intricate and the situation is rapidly changing [1]. Modeling and analyzing every stock would inevitably be redundant, time-consuming, and laborious. Therefore, many financial institutions and stock exchanges have developed stock indices based on multiple representative stocks [2].

Stock market prediction methods can be divided into two categories: linear and non-linear. Linear prediction models include the autoregressive integrated moving average (ARIMA) model, and the generalized autoregressive conditional heteroscedasticity (GARCH) model [3]. However, because the original stock index data is a single time series with limited embedded information and characteristics of non-linearity and non-smoothness, and also contains a large amount of noise and useless information, it is difficult for the traditional time series prediction models to make effective forecasts of stock market and asset price trends.

In recent years, neural network models in machine learning have made breakthroughs in various fields. This also provides a new tool for the financial investment industry based on big data. After a few years of development, neural network models have been widely used in stock predictions. Bayogl compared the prediction effectiveness of Bayesian models and neural network models, and the results proved that the latter had a better performance [4]. A distinctive feature of a neural network model is nonlinearity; therefore, it belongs to nonlinear models. In the early stage of machine learning, the main prediction methods are support vector machines (SVMs), which include BP neural network and MP neural network. However, the prediction models have some potential problems, such as overfitting and inability to generalize; local optimum rather than global optimum results in inaccurate predictions. Hammda [5] used a multilayer BP neural network model to make predictions of index price advantage in the Jordanian stock market. However, he did not solve the problem that BP neural networks tend to fall into a local minimum, and the results were not convincing enough. Considering the shortcomings of the above models for time series predictions, more efficient and comprehensive

deep neural networks (DNN) models have started to emerge in the field of financial forecasting, which include convolutional neural networks (CNN), recurrent neural networks (RNN) [6] and long short-term memory neural networks (LSTM) [7]. Deep neural network models have multiple advantages. They do not restrict the types of input variables, and any relevant information can be an input variable. At the same time, they can effectively deal with the complex nonlinear relationships between variables, improve the degree of aggregation, reduce the number of neuron weights, and avoid overfitting [8]. Deng Fengxin used the LSTM model to predict individual stocks such as Amazon, and the results once again proved that LSTM has higher accuracy in financial forecasting [9].

The prediction method of the ensemble model has become popular now because it overcomes some of the limitations in neural networks. In an ensemble model, the time series are firstly decomposed using signal processing methods, and then the decomposed individual series are predicted using neural networks. The idea of combination makes up for the shortcomings of a single neural network and significantly improves the prediction accuracy. Guo Jinlu further decomposed the decomposed residual series to obtain more information, and he proposed a VMD-EEMD-LSTM model with better prediction accuracy than the separate ensemble model [10].

Signal processing methods are a key part of ensemble model prediction and directly affect predictions' accuracy. Empirical mode decomposition (EMD) [11] is the most typical signal decomposition method. However, Xiong Tao discovered that there was a drawback of the EMD method in the prediction process, that is, mode mixing [12]. To prevent the shortcoming of EMD from bringing a large impact on the prediction results, Dragomiretskiy proposed variational mode decomposition (VMD) [13]. VMD showed better noise robustness in the signal decomposition process. Lahmiri first applied VMD to the financial forecasting and achieved relatively good results [14].

The values of alpha and K of VMD directly affect the decomposition results, and the artificial way of determining them is defective. Before using VMD, particle swarm optimization (PSO) is used to determine the optimal parameters. Sample entropy (SE) [15] is used to merge and recombine similar time series to obtain time subseries with more significant complexity. LSTM prediction model is built by adding influencing factors for each recombined series, and finally, the individual prediction values are superimposed to obtain the final prediction results.

The stock index volatility is affected by many factors, so we consider the VMD decomposition of index series data, which can divide the economic operation into the superposition of several different fluctuations, which is more interpretable in the short-term fluctuation and long-term trends.

Considering that the parameter selection of VMD decomposition will significantly affect the prediction accuracy of LSTM, we add a PSO particle swarm optimization algorithm to select the best economic parameters by simulating birds' predation in nature. The significance of this paper is as follows:

(1) The PSO optimization decomposition parameter method is proposed to solve the important parameter selection problem in the combined neural network model, which significantly improves the model's prediction accuracy. (2) The SE method is used to recombine the sequence after correlation decomposition, making the complexity distinction between the sequences more obvious and explanatory. (3) Taking investors' performance into account in neural network prediction enriches the explanatory dimension of the financial forecast

The remaining chapters of this paper are set as follows: Chapter 2 is methodology. Firstly, the data decomposition methods and neural network theory used in this paper are introduced, mainly including VMD, SE, PSO and LSTM; Chapter 3 is a case study. A prediction of CSI 300 index based on PSO-VMD-SE-LSTM is provided in this chapter. Firstly, the optimized VMD method is used to decompose and restructure the price data, followed by LSTM prediction and finally, comparisons with several other models; Chapter 4 is the conclusion.

2. Methodology

2.1. Combined decomposition strategies PSO-VMD

The core principle of VMD is to decompose the original input signals into k mode components u_k with center frequency and limited bandwidth using an adaptive and quasi-orthogonal decomposition method, and to minimize the sum of the bandwidth estimates of all the modes. The process of VMD signals decomposition is also the process of solving constrained variational problems.

PSO is an adaptive optimization method based on swarm intelligence, simulating the predatory behavior of a flock of birds.

In order to determine the number of IMF components that need to be decomposed after VMD processes the original data, this paper uses PSO to make the selection.

BP neural network is back-propagating, mainly composed of three parts: input layer, middle layer and output layer. The number of nodes in the input and output layers is relatively easy to determine. However, determining the number of nodes in the hidden layer is a significant and complex problem.

2.2. Sample Entropy

Sample entropy (SE) is a new method proposed by Richman et al. in 2000, which can measure the complexity of time series. It is a modification of approximate entropy (AE), reduces the dependence on the length of time series, and can effectively reduce the error of approximate entropy in the calculation process.

2.3. LSTM

Long short-term memory (LSTM) is a special recurrent neural network. By carefully designing structures called gates, it controls the degree of the correlation between current data and historical data. It avoids the vanishing and exploding gradient problems caused by traditional recurrent neural networks, which can effectively learn long-term dependencies. Therefore, the LSTM model with memory function shows stronger advantages in classifying and making predictions based on time series data [16]. The basic model of LSTM is shown in Figure 1.

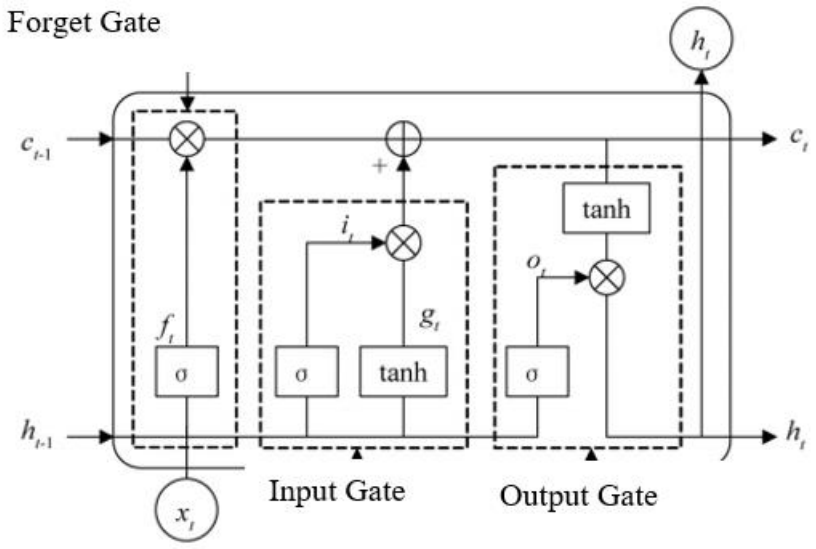


Figure 1 An LSTM Unit

2.4. Prediction Model of Stock Index Based on IPSO-VMD-SE-LSTM

The structure of the PSO-VMD-SE-LSTM model for CSI 300 Index forecasting is shown in Figure 2, and the steps are as follows:

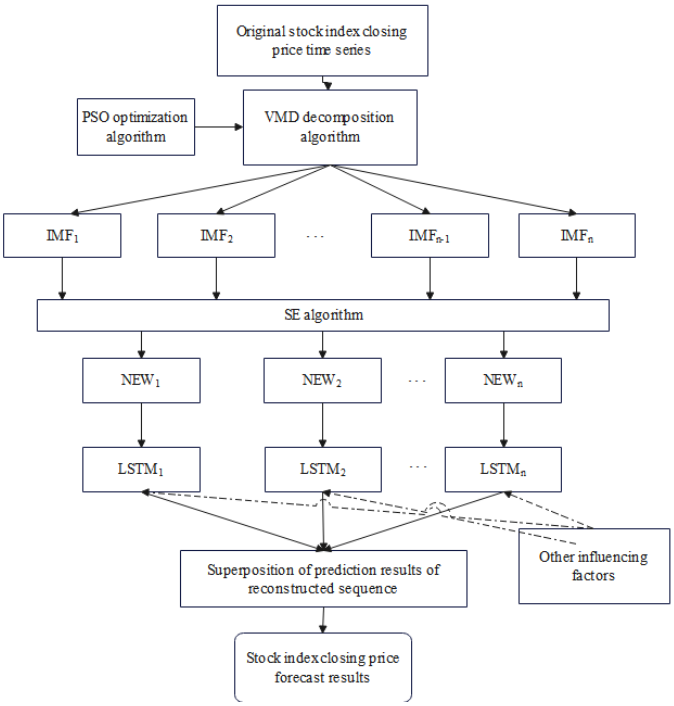


Figure 2 PSO-VDM-SE-LSTM Prediction Model

(1) Obtain modeling data containing historical stock index closing prices and other trading parameters.

(2) Decompose the original CSI 300 Index closing price series into several IMF components by using VMD optimized by PSO.

(3) Calculate the sample entropy values of each IMF component, combine and restructure the subseries with similar sample entropy values to obtain a new subseries with significant complexity differences.

(4) For each new subseries, add other transaction parameters to build the LSTM neural network prediction model with corresponding parameter spaces, and output the prediction value of each model.

(5) Sum up all the models and get the prediction result of stock index closing price.

(6) Compare the predictions with actual time series data and calculate the error indicators for error analysis.

3. Case Study

3.1. Data Collection and Processing

To evaluate the effectiveness and practicability of the model in stock index forecasting, this paper selects the CSI 300 Index as the research object, and the index prices for three consecutive months from October 2021 to December 2021 are chosen as the research sample. The sample contains more than 2,900 sets of data, and 48 sets are obtained in a 5-minute set daily. The variables include opening price, high price, low price, closing price, and volume-weighted average price (VWAP). VWAP is often used in high-frequency data to measure investors' performance [17], and the data are obtained from Wind. Eighty percent of the sample data set is taken as features to select the sample set, and the other twenty percent of the sample is used as the independent test set. The selected features are tested on the independent test set, and classification and prediction without feature selection are conducted as a comparison experiment. A feature set is constructed using the selected data.

Due to the differences in calculation methods of each indicator, the eigenvectors constructed based on technical indicators have different value ranges, and the differences are quite huge. This huge difference will make it challenging to find the best solution to the model parameters. And overfitting is likely to occur, adversely affecting the final prediction results. Therefore, this paper uses the

following equation to normalize the eigenvectors and convert the eigen components in every dimension to [-1, 1]. The equation is as follows:

$$\hat{x}_d = \frac{x_d - x_{d,\min}}{x_{d,\max} - x_{d,\min}} \tag{1}$$

where $x_{d,max}$ is the maximum value of d-dimensional eigen component, $x_{d,min}$ is the minimum value of d-dimensional eigen component. In order to maintain the practical meaning of the results, the prediction data needs to be denormalized using the following equation:

$$\hat{y}_d = MAE = \frac{1}{n} * \sum_{i=1}^{n} |y_i - \hat{y}_i| (x_{d,\max} - x_{d,\min}) + x_{d,\min} \tag{2}$$

where y_d denotes the normalized data from the neural network output and $\hat{y}_d$ denotes the predicted value after denormalization.

3.2. Decomposition and Recombination of Time Series Based on PSO-VMD-SE

To conduct a more accurate local analysis of the CSI 300 Index time series, PSO is used to determine the core parameters. Then use VMD to decompose the time series data and obtain 7 IMF components, as shown in Figure 3.

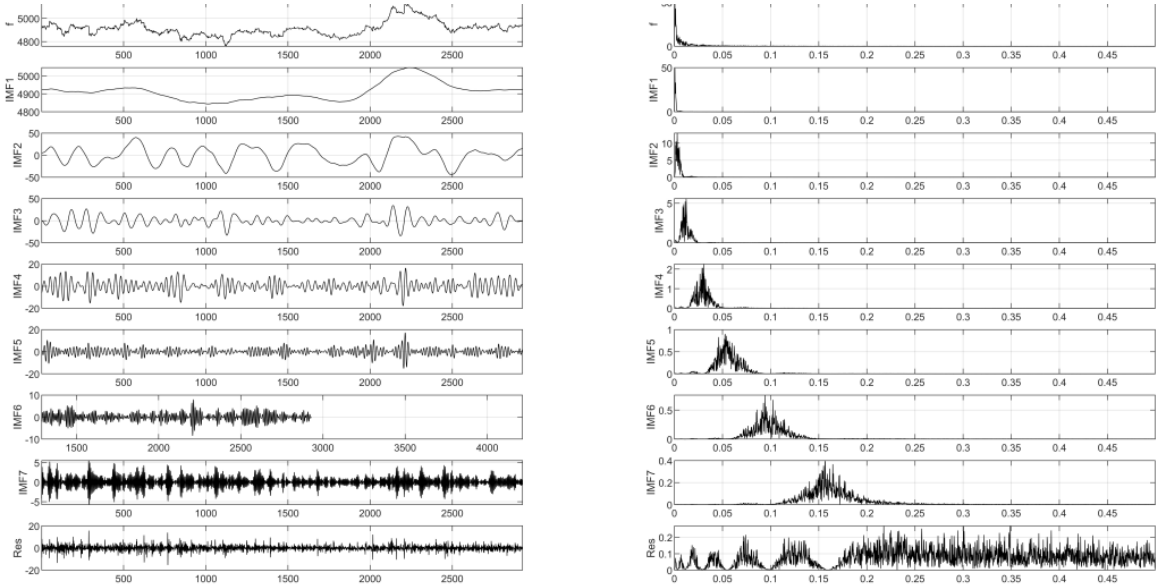


Figure 3 Decomposition of CSI 300 Index Closing Prices Time Series Using VMD

To reduce the computational scale for more effective prediction of the CSI 300 Index time series, sample entropy theory is used to evaluate the complexity of each IMF component obtained from the VMD decomposition, and the sample entropy values of each remaining component are calculated. The distribution of the sample entropy values of each IMF component is shown in Table 1. To improve the operation's efficiency and accuracy, four sets of recombination sequences are obtained by comprehensively comparing the differences before and after each IMF component. The detailed results are shown in Table 1. The new subseries after the combination is shown in Figure 4:

Table 1 New subsequences after the merger of IMF

IMF_n	H_{SE}(n)	Combined component	New IMF component
IMF_1	0.038467	IMF_1、IMF_2	NEW_1
IMF_2	0.10535		
IMF_3	0.32753	IMF_3、IMF_4、IMF_5	NEW_2
IMF_4	0.68793	IMF_4	NEW_3
IMF_5	0.97143	IMF_5	NEW_4

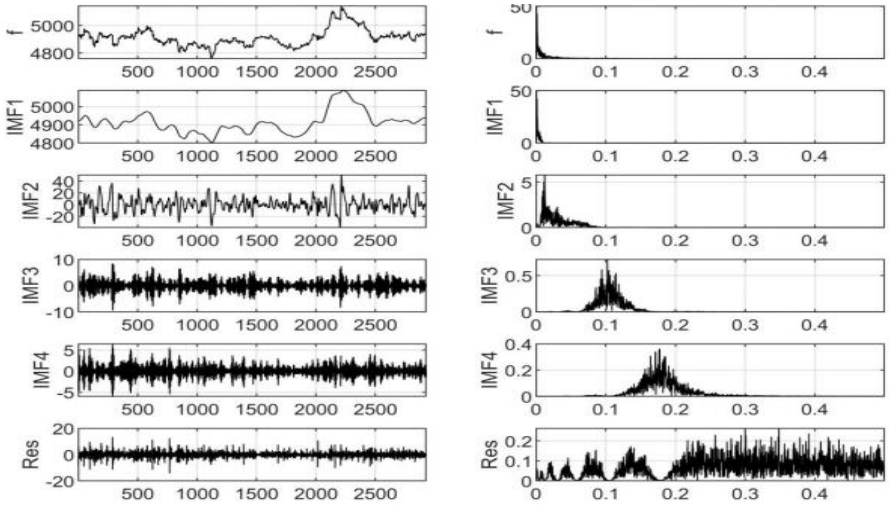


Figure 4 CSI 400 Index Closing Prices Subseries Processed by VMD-SE

where n is the number of test data, y_i and $\hat{y}_i$ are the predicted value and actual value of the CSI 300 Index, respectively.

3.3. Prediction and Comparative Analysis

This paper uses a multi-step forecasting model to predict the series of closing prices. The first 50 data are used as input variables to predict the closing price at time 51, then iteratively use the closing prices from time 2 to time 51 to predict the closing price at time 52, and so on.

Use the Adam optimizer and set the learning rate to 0.01. To avoid overfitting in the experiment, dropout is used by randomly dropping out some units. Set the dropout rate at 0.3.

To verify the superiority of the PSO-VMD-SE-LSTM model, PSO-VMD-SE-LSTM, LSTM, VMD-LSTM, and BP models are constructed in the same computing environment to compare the prediction results. The evaluation indicators of the prediction results are shown in Table 2.

Table 2 Evaluation index of each prediction model

Model	MAE	RMSE	MAPE	R^2
LSTM	3.4232	4.3113	0.069411%	0.9852
BP	4.1539	6.5871	0.083612%	0.9841
VMD-LSTM	2.5937	3.2659	0.052577%	0.9915
PSO-VMD-LSTM	2.3668	2.9526	0.047989%	0.9936

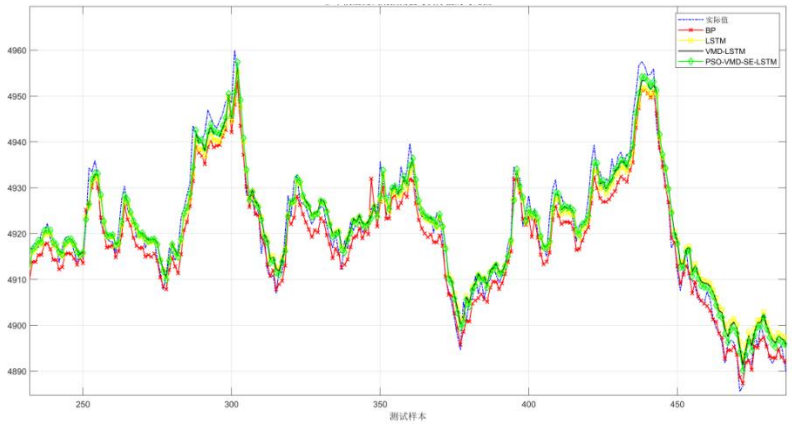


Figure 5 Prediction Results of Each Model

By comparing the values of MAE, RMSE, MAPE, and R^2 of the four models in Table 2 and the prediction results in Figure 5, it can be seen that the PSO-VMD-SE-LSTM model has the smallest MAE, RMSE, MAPE and the largest R^2, which proves that this model has the best performance in

prediction. The scale of the data fluctuation has some influence on the prediction effect of the model, which also reflects the good prediction effect of VMD on the data with large fluctuations.

We collect the data of Shanghai Composite Index in the same period and forecast it. As the second example, by comparing the error index of each prediction model (as shown in the table 3), the prediction model proposed in this paper also has a good prediction effect for Shanghai Composite Index, and the error index is obviously lower than other contrast models.

Table 3 Evaluation index of each prediction model

Model	MAE	RMSE	MAPE	R^2
LSTM	2.1830	2.8159	0.060144%	0.9824
BP	2.1822	3.1935	0.060195%	0.9767
VMD-LSTM	2.1469	2.7739	0.05918%	0.9871
PSO-VMD-LSTM	2.0478	2.8724	0.05734%	0.9896

4. Conclusions

To improve the accuracy of stock index price prediction, this paper proposes an LSTM neural network model based on the VMD decomposition after PSO optimization to predict the stock index price trend. The model first uses PSO to optimize the two key parameters, the threshold and the number of decompositions of VMD, and then uses the optimized VMD to decompose the normalized stock index data into several IMF series. After that, using the SE algorithm to recombine the series with similar entropy values to obtain several NEW series with significant differences in volatility. Using the LSTM model to predict the reconstructed components separately, all series are summed up and denormalized to obtain the predicted values, which have a better accuracy than the benchmark model.

After experimental analysis, it is concluded that the PSD-VMD-SE-LSTM model has the following characteristics: (1) Stock index price movements reflect trading models with different frequencies. As an economic analysis tool, VMD decomposes the time series into a superposition of different frequencies. (2) VMD treats economic operations as a superposition of several fluctuations (including long-term trends and short-term fluctuations), making the predictions more interpretable. (3) Incorporating the excellent performance of LSTM model in dealing with high noise, high non-stationarity and high non-linearity, predictions using long trends and short-term fluctuation series filtered and decomposed by VMD all achieve high accuracy. (4) The test error of PSD-VMD-SE-LSTM is better than that of the benchmark model, indicating that the prediction accuracy and stability of PSD-VMD-SE-LSTM model are greatly improved.

References

[1] Zhao Y, Yang LY. Research on stock price prediction based on cca-bp neural network[J]. Statistics and Management, 2020,35(3): 66-69.

[2] Sun L,Liu JG. Research on the compilation of the subject matter index of China's stock index futures contract[J]. East China Economic Management, 2003(5): 105-106+62.

[3] Yang Q, Cao XB. Stock price analysis and prediction based on arma-garch model[J]. Journal of Mathematics in Practice and Theory, 2016, 46(6): 80-86.

[4] Ozbayoglu Murat A, Ismet Bahadir. Comparon of bayesian estimation and neural network model in stock market trading[J]. Intelligent Engineering Systemsthrough Artificial Neural Networks, 2011, (20): 74-81 .

[5] Hammad A A A, Ali S M A, Hall E L. Forecasting the jordanian stock prices using artificial neural networks [M [// Intelligent Engineering Systems Through Artificial Neural Networks, 2007: 502-505.

[6] Mikolov T,Karafiát M,Burget L, et al. Recurrent neural network based language model[C] . Eleventh annual conference of the international speech communication association. Chiba: ISCA, 2010: 1045-1048 .

[7] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation,1997,9(8): 1735-1780.

[8] Hu JW. Stock forecasting based on optimized LSTM model[J]. Scientific Journal of Computer Science, 2021, 48(S1): 151-157.

[9] Deng FX, Wang HL. Application of LSTM neural network in stock price trend prediction -- a study based on individual stock data in US and Hong Kong stock market[J]. Finance & Economy, 2018(14): 96-98.

[10] Guo JL. Research on prediction of CSI 300 index based on vmd-eemd-lstm model[J]. Modern Finance and Economics-Journal of Tianjin University of Finance and Economics, 2020, 40(8): 31-44.

[11] Huang N E. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of Royal Society of London, 1998, 454.

[12] Xiong T. Research on hybrid modeling technology of time series prediction based on EMD and its application[D]. Huazhong University of science and technology,2014.

[13] Dragomiretskiy K, Zosso D. Variational mode decomposition[J]. IEEE Transactions on Signal Processing, 2014, 62(3): 531-544 .

[14] Lahmiri. Long memory in international financial markets trends and short movements during 2008 financial crisis based on variational mode decomposition and detrended fluctuation analysis[J]. Physica A Statistical Mechanic& its Applications, 2015, 437: 130-138.

[15] RICHMAN JS, MOORMAN JR . Physiological time-series analysis using approximate entropy and sample entropy[J] . American Journal of Physiology Heart Circulatory Physiology, 2000, 278(6): 2039-2049.

[16] HOCHREITE S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[17] Yan RZ, Li P, Zeng Y. Research on price shock cost estimation and its influencing factors based on VWAP[J]. Journal of Industrial Engineering and Engineering Management, 2016, 30(3): 129-133.