Data Science in Finance and Economics

DSFE, 3 (3): 251–266.
DOI: 10.3934/DSFE.2023015
Received: 29 May 2023
Revised: 13 July 2023
Accepted: 28 July 2023
Published: 08 August 2023

http://www.aimspress.com/journal/dsfe

Research article

Research on crude oil price forecasting based on computational intelligence

Ming Li, Ying Li*

Zhongnan University of Economics and Law, Wuhan, 430073, China

* **Correspondence:** Email: z0004364@zuel.edu.cn.

Abstract: The crude oil market, as a complex evolutionary nonlinear driving system, is by nature a highly noisy, nonlinear and deterministic chaotic series of price series. In this paper, a computational intelligence-based portfolio model is constructed to forecast crude oil prices using weekly price data of West Texas intermediate crude oil (WTI) crude oil futures from 2011 to 2021. First, the WTI crude oil price series are decomposed using the ensemble empirical modal decomposition method (EEMD) and the set of component series is reconstructed using the cluster analysis method. Second, the reconstructed series are modeled and predicted using neural network models such as time-delay neural network (TDNN), extreme learning machine (ELM), multilayer perceptron (MLP) and the GM (1,1) gray prediction algorithm and the output of the model with the best prediction effect for each component is integrated. Finally, the EGARCH model is used to further optimize the predictive power of the combined model and output the final predicted values. The results show that the combined model based on computational intelligence has higher forecasting accuracy than single models such as GM (1,1), ARIMA, MLP and the combined EEMD-ELM model for forecasting crude oil futures prices.

Keywords: crude oil price; neural network model; gray forecasting algorithm; ensemble empirical modal decomposition

JEL Codes: C63, E37

1. Introduction

Commodities can be divided into three major categories: energy commodities, basic raw materials commodities and agricultural and sideline products, which are related to the lifeline of economic and social development. Oil, as one of the most important strategic substances, affects the security strategy of every sovereign country in the world. In the context of the financialization of commodities, crude oil futures have become the largest trading variety in the commodity futures trading market. Due to the continuous increase in China' s demand for crude oil and the frequent changes in international crude oil prices this year, the accuracy of international crude oil price forecasts is commonly valued by domestic investors and has become an important basis for decision-making bodies and government departments to effectively control crude oil market risks. For this reason, accurate forecasting of the long-term trend of international crude oil prices has become a hot topic being studied by experts and scholars at home and abroad. As a complex evolutionary nonlinear driving system, the futures market price series is in fact a highly noisy, nonlinear, unsteady and deterministic chaotic series. Model fitting and model prediction of such dynamic unsteady series are of strong guidance for investors to make investment decisions. So, researchers have proposed methods to model crude oil prices from different perspectives in recent decades. However, such forecasting methods are generally based on the assumption that the time series is linear and therefore does not perform as well in the complex and volatile futures market. Since the vast majority of the time series and the driving problems caused by them are considered to be a nonlinear, nonstationary process in practice and many traditional analytical methods can only be performed under the premise that these data are smooth signals. This directly results in the loss of some local information in the series. To extract the local structure of the signal in a more precise and detailed way, it is necessary to adopt an effective processing method suitable for nonstationary time series.

Computational intelligence is a collective term for algorithms created and developed inspired by the wisdom of nature. With the development of science and economy, the difficulties encountered in the study of computer technology have become increasingly complex and the application of traditional computational models to solve these difficulties has faced problems such as high computational complexity and long computational durations. For this reason, mathematicians and computational theorists have developed a series of computational intelligence algorithms with heuristic features such as multilayer perceptrons (MLP), time-delayed neural networks (TDNN), extreme learning machines (ELM) and ensemble empirical modal decomposition (EEMD). Some of these algorithms mimic the evolutionary process of natural organisms and some mimic the thinking process of human brains and researchers hope to achieve problem solving and optimization in natural sciences and economic and social fields by simulating nature and human intelligence. This paper focuses on the problem of forecasting futures price series in financial markets and mainly uses a combinatorial model based on computational intelligence to model and analyze futures price time series. From the evaluation of the prediction results, the combinatorial model proposed in this paper achieves better prediction results.

2. Current status of crude oil price research

In the mid to late 1970s, due to the frequent outbreak of wars and other events, crude oil prices entered a phase of ultrahigh oil prices, which had a huge impact on the economies of industrial countries and led academics to start in-depth research on crude oil prices. Academic research on crude oil prices mainly focuses on two aspects: oil price influencing factors and oil price forecasting.

2.1. Crude oil price influencing factors study

The demand for crude oil as an important industrial raw material for economic development is directly dependent on economic dynamics. During economic booms, the demand for crude oil is high and prices rise, while during recessions, the demand for crude oil is low and prices fall (Baumeister et al. 2015). During a study of the decline in crude oil prices since 2014 it was found that some of these predictable price declines were due to the decline in demand for crude oil as a result of the global economic slowdown (Baumeister et al., 2014). The results of modeling the cointegration relationship between the spot price of North American West Texas crude oil (WTI) and OPEC crude oil supply (OPEC), World OECD member countries' oil consumption demand (OECDD), and World OECD member countries' oil inventories (OECDS) indicate that the long-term trend of oil prices is still determined by economic factors such as supply and demand at the most fundamental level and that the short-term oil price movements are mainly derived from sudden non-economic factors. (Li and Yang, 2007). In the empirical analysis, it is found that whether in the long term or short term, the impact of financial factors on crude oil prices is stronger than that of the fundamentals of demand (Tian and Tan, 2015). As crude oil has the political attribute of *oil diplomacy*, geopolitical risk directly affects the supply of crude oil. A geopolitical risk index (Caldara, 2022) is constructed from text data of newspapers and it is found that geopolitical actions have a more significant impact on crude oil prices than geopolitical threats (Bouoiyour, 2019). The inclusion of terrorist attacks as an input variable for predicting oil prices improves the predictive power of the model and finds that terrorist attacks have a positive impact on oil prices and that this impact is mainly concentrated on terrorist attacks that occur in oil-producing countries. (Phan, 2021)

2.2. Crude oil price forecasting model study

The main methods used in traditional econometric modeling to forecast commodity futures prices are autoregressive moving average model (Wang, 2016) and autoregressive conditional heteroskedasticity model (Morana, 2001). After linear regression modeling for oil price prediction, it was found that the technical indicators showed a stronger ability in price prediction compared to the economic indicators. (Yin, 2016) and other literature has similarly considered the use of linear regression models to forecast crude oil prices (He et al., 2021). However, due to the nonlinear nature of the commodity futures market (Plourde and Watkins, 1998), traditional econometric models are often difficult to capture the characteristics of price fluctuations and their application is somewhat limited. Therefore, in recent years, machine learning methods such as artificial neural networks (ANN) have been widely used in the prediction of futures prices. After the prediction of gold futures prices by using ANN and ARIMA models respectively, the empirical results show that the prediction effect of ANN is significantly better than that of ARIMA (Chen et al., 2016).To alleviate the problem of high volatility of time series, radial basis function (RBF) neural network model based on empirical modal decomposition (EMD) is used to forecast stock market futures, which improves the accuracy of forecasting by decomposing the original series with long correlation into multiple short-correlated subsequences (Lu, 2017). In the use of variational modal decomposition and extreme learning machine to predict wind power, it was found that the machine learning model was found to be able to predict the trend of the sub-sequences more accurately by decomposing the raw time series data when compared with other hybrid decomposition models and a single model (Abdoos, 2016).

3. Data sources and model setting

This chapter introduces the data sources, decomposition methods and model construction for empirical analysis, forecasting process design and forecasting effect evaluation indicators.

3.1. Data selection and sources

When analyzing price fluctuations in financial time series, the shortest possible time interval should be chosen. In this paper, weekly prices of WTI crude oil futures are selected as the dataset for this experiment, and the data are taken from the website of Yingwei Financial Intelligence (https://cn.investing.com/). The data in the training set established in this paper are 553 weekly data points from May 2010 to December 2020. The data in the test set is from 20 weekly data points from January 2021 to May 2021. The main reasons for using weekly crude oil prices in this paper are as follows: (1) the literature studying crude oil prices often uses daily data and less often involves weekly data, (2) the medium- and long-term forecasting of weekly prices is highly flexible and a four-period forecast can show the trend of the coming month. The dataset is divided into a training set and a test set in a ratio of 9:1, where the training set is used to train the model and the test set is used to test the model prediction accuracy.

3.2. Decomposition method

Empirical mode decomposition (EMD) is an adaptive decomposition technique based on the local characteristics of the data and is commonly used for time series forecasting analysis with high frequency fluctuations. EMD analyzes and processes the complex and various crude oil price trend series fluctuations in a smooth and quantitative manner to facilitate the timely extraction of the continuous fluctuations of crude oil price and its price trend components at different scales, thus dividing the complex crude oil price fluctuation series information into a number of different scales of intrinsic mode functions (IMFs). The IMFs have the following characteristics: (1) the number of extreme values is equal to or at most one different from the number of crossing zeros, (2) at any moment, the average value of the upper and lower two and the average value of the envelope must remain zero at any moment. The specific decomposition process is:

1. Determine all the extreme and extreme minima of the WTI crude oil futures price series $p(t)$ and fit the upper and lower envelopes using the three-sample combination function and the difference between the mean values m_1 of the series $p(t)$ and the upper and lower envelopes is recorded as h_1.
2. Considering h_1 as a new sequence, the above process is repeated, and as long as h_1 satisfies the above two conditions of the eigenmode function, it is selected as the first IMF component c_1 filtered from the original time fluctuation sequence representing the highest frequency component of the original sequence. The other remaining quantities can be expressed as $r_1 = p(t) - c_1$.
3. The above stated decomposition is continued for r_1 until the decomposition process proceeds to the n-th stage when the residual sequence shows monotonicity or its numerical magnitude has been less than a pregiven value, at which point the decomposition ends.

However, empirical mode decomposition is prone to fluctuation sequence mode mixing, i.e., an IMF component contains part of the signal with relatively large differences in fluctuation frequency scales. Based on the above shortcomings, an algorithm is proposed to improve the EMD model, ensemble empirical mode decomposition (EEMD). First, a set of white noise sequences with a low signal-to-noise ratio is added to the time fluctuation series, and then the reconstructed sequences are subjected to the abovementioned empirical mode decomposition process. After several calculations, the combined evaluation is performed to cancel the added white noise information with each other. This processing method not only retains the basic signal information of the original data but also overcomes the mode confusion problem encountered in EMD (Wu and Huang, 2009). Therefore, in the subsequent analysis of this paper, the EEMD technique will be used to decompose the WTI crude oil futures price series, adding a white noise series with an amplitude of 20% of the standard deviation of the original series each time when processing the data and performing 250 ensemble evaluations.

3.3. Selection of prediction methods

With the increasing complexity of problems in economic and social development, using some traditional computational techniques and methods to address them has to solve a series of difficulties, such as the complexity of data computation and long computation time. For this reason, in this paper, we choose multilayer perceptron (MLP), time-delayed neural network (TDNN), extreme learning machine (ELM) which are computational intelligence algorithms with heuristic features and GM (1,1) which is a traditional forecasting method, to predict WTI crude oil prices.

1. GM (1,1)

In this paper, we use the GM (1,1) model in the gray forecasting model for time series forecasting of crude oil futures data. GM (1,1) is essentially a *one-time accumulation generation* to form a new series with logical correlation. Based on the new data series, a prediction model with differential, difference and approximate exponential law compatibility is built and then the predicted values of the original data series are obtained through the *one-time cumulative generation* process.

2. Time delay neural network (TDNN)

TDNN is a multilayer feedforward neural network model with the help of introducing time delay neurons. Between the hidden and output layers of the model, time delay neurons (TDN) are present between all neurons and the output of the previous layer. The TDNN consists of several parallel time delay units, which are generally larger in the TDN the further down the layer.

The output of a TDN is a weighted sum of its inputs at moments $t - d_1(t$ is the current moment and $d_1 = 1,2, \ldots, D_1$), and its excitation function is f(x) = 1/(1-e$^{(-x)}$). If the input is $x_1, x_2, \ldots, x_n$, then the output of each neuron in its hidden layer at moment t is $Z_m(t)$, (m = 1,2,…,L)

$$\mathrm{Z_m(t)} = \sum_{i=1}^{n} \sum_{d_1=0}^{D_1} W_{id_1m} \cdot x_i(t - d_1) - a_m \tag{1}$$

where D_1 is the number of hidden layer delay steps, W_{id_1m} represents the weight between the i-th neuron in the input layer at moment $t - d_1$ and the m-th neuron in the hidden layer, x_i $(t - d_1)$ represents the value of the i-th input value at moment $t - d_1$ and a_m represents the bias to which the m-th neuron in the hidden layer of the model belongs. The output result $O_r(t)$ $(r = 1,2,\cdots,\mathrm{R})$ for each neuron unit in the output layer of the model at time t is:

$$\mathrm{O_r(t)} = \sum_{m=1}^{L} \sum_{d_2=0}^{D_2} W_{md_2r} \cdot Z_m(t - d_2) - b_r \tag{2}$$

where D_2 is the number of delay steps in the output layer, W_{id_2m} is denoted as the weight between the m-th neuron in the hidden layer at moment $t - d_2$ and the r-th neuron in the output layer, Z_m $(t - d_2)$ denotes the value of the output value of the m-th neuron in the hidden layer at moment $t - d_2$ and b_r is the bias of the r-th neuron in the hidden layer.

3. Multilayer perceptron (MLP)

The feedforward signal network received by a multilayer feedforward perceptron is composed of a multilayered unidirectional feedforward neural network, with each layer consisting of one or several neuron connections. The first of these layers is called the entry input control layer, the last layer is called the entry output control layer and the middle is the hidden layer. MLPs are often used to solve nonlinear problems.

Taking a single hidden layer feedforward neural network as an example, we set the number of unit neurons in the initial input layer, hidden layer and output layer as n, L and M respectively and the activation function f(x) = 1/(1-e$^{(-x)}$). When the input x = (x_1, x_2,..., x_n), then the output Z_l (l = 1,2,...,L) of each neuron in the hidden layer is:

$$Z_l = f(\sum_{i=1}^{n} W_{il} \cdot x_i + b_l) \tag{3}$$

where W_{il} denotes the weight between the i-th neuron in the input layer and the l-th neuron in the hidden layer and d_l denotes the bias of the lth neuron in the hidden layer.

Then, the output O_m (m = 1,2,...,M) of each neuron in the output layer is:

$$O_m = f(\sum_{l=1}^{L} W_{lm} \cdot Z_l + b_m) \tag{4}$$

where W_{lm} denotes the weight between the l-th neuron in the input layer and the m-th neuron in the hidden layer, and b_m denotes the bias of the m-th neuron in the hidden layer.

4. Extreme learning machine (ELM)

The extreme learning machine algorithm is an algorithm specifically studied for application to single hidden layer feedforward neural network (SLFN) training. Unlike some other traditional minimum gradient-based typical SLFN feedforward training network algorithms, the weights of each training input layer and the bias between each hidden layer of the ELM algorithm are determined randomly and the occupation weights of the model output layers are determined by solving for the minimum value of the training error of the model. Compared with several other traditional neural networks, the ELM algorithm has three major features, less learning training time parameters, faster learning response time and better generalization learning performance, based on fully guaranteeing the learning training accuracy.

Suppose the training set is $\{x_i,\ t_i | x_i \in R^n,\ t_i \in R^m,\ i = 1,2,\dots,N\}$, where x_i denotes the i-th training data and t_i denotes the token corresponding to the i-th training data. The number of neurons in the hidden layer is L. Then, the network structure of ELM is shown in Figure 1.

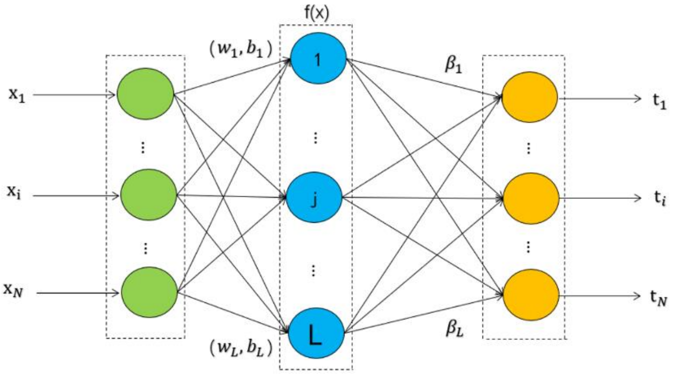

Figure 1. Extreme learning machine network structure.

In constructing the model, the excitation function f(x) = 1/(1-e$^{(-x)}$) is taken and the mathematical model expression of its single hidden layer feedforward neural network is:

$$\sum_{i=1}^{L} \beta_i f(W_i \cdot x_j + b_i) = o_j, j = 1,2, \dots, N \tag{5}$$

where β_i denotes the connection weight between the i-th hidden layer neuron and the output layer, W_i denotes the connection weight between the i-th hidden layer neuron and the input layer, $W_i \cdot x_j$ is the inner product of W_i and x_j, b_i is the bias of the i-th hidden layer neuron and o_j is the output value of the jth input value. The learning goal of the single hidden layer feedforward neural network is to make the output o_j approximate the training sample with zero error, which can be expressed as

$$\sum_{j=1}^{N} \| o_j - t_j \| = 0 \tag{6}$$

Then, there exists β_i, W_i and b_i such that

$$\sum_{i=1}^{L} \beta_i f(W_i \cdot x_j + b_i) = t_j, j = 1,2, \dots, N \tag{7}$$

The matrix form of the above equation is Hβ=T, where H is the output matrix of the hidden layer,

$$H = (W_1, \dots, W_L, b_1, .., b_L, x_1, .., x_L) = \begin{pmatrix} f(W_1 \cdot x_1 + b_1) & \cdots & f(W_L \cdot x_1 + b_L) \\ \vdots & \ddots & \vdots \\ f(W_1 \cdot x_N + b_1) & \cdots & f(W_L \cdot x_N + b_L) \end{pmatrix}_{N \times L} \tag{8}$$

$$\beta = \begin{pmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{pmatrix}_{L \times m}, \quad \mathrm{T} = \begin{pmatrix} t_1^T \\ \vdots \\ t_N^T \end{pmatrix}_{N \times m} \tag{9}$$

Because W_i $b_i$in the output matrix of the hidden layer have been randomly selected during training, solving the above matrix form, the solution is

$$\hat{\beta} = H^{+} T \tag{10}$$

where H^+ is the generalized inverse matrix of the output matrix H of the hidden layer and the solution is solved with a minimum and unique parametrization.

3.4. Prediction process

In this paper, we first decompose the WTI crude oil futures data by the EEMD ensemble empirical modal decomposition method to obtain the eigenmode function (IMF) and its residual components. Then, we perform cluster analysis on the correlation coefficients between the eigenmode functions and obtain the high-, low- and medium-frequency volatility series by reconstructing the eigenmode functions according to the analysis results. Then, we use ELM, GM (1,1) and TDNN. The best prediction results of each fluctuation term are combined and the GARCH model is used to extract further information from the residuals of the combined model to obtain the final prediction results. The forecasting process is shown in Figure 2.

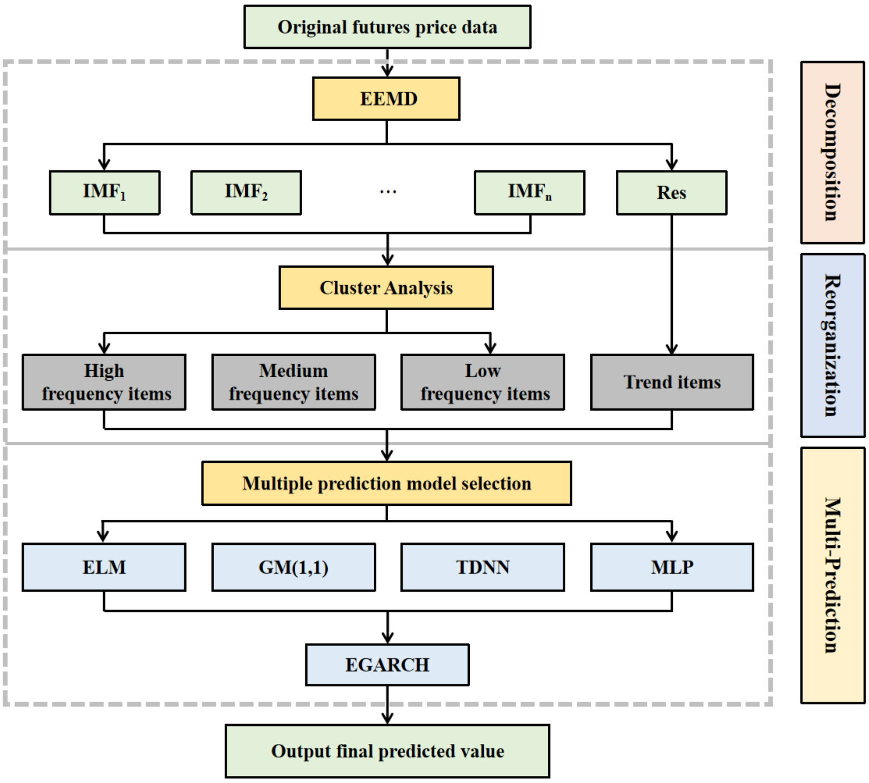

Figure 2. Combined model flow chart.

3.5. Selection of evaluation indicators

In this paper, the mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE) are selected to evaluate the effectiveness of each forecasting method in predicting oil prices.

1. Mean absolute error (MAE), which indicates the absolute mean of the relative error between the average estimate of statistical forecasts and the average estimate of real statistics. The

larger the value of the evaluation index is, the lower the prediction accuracy.

$$MAE = \frac{1}{n}\left(\sum_{t=1}^{n} |y(t) - \hat{y}(t)|\right) \tag{11}$$

2. The mean absolute percentage error (MAPE) indicates the mean value of the mean absolute percentage error between the model prediction estimates and the real actual measurements, and the smaller the value of the evaluation index is, the higher the accuracy of the prediction.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y(t)-\hat{y}(t)}{y(t)}\right| \tag{12}$$

3. The root mean squared error (RMSE) is used to express the deviation of the mean square of the sum of squares of the distance between the predicted estimate and the true set value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\left(y(t) - \hat{y}(t)\right)^2\right.} \tag{13}$$

4. Empirical analysis

4.1. Decomposition and reconstruction of WTI crude oil futures prices

The commodity futures price data used in this paper is the daily data of West Texas Intermediate (WTI) crude oil futures price from May 30, 2010 to May 16, 2021 on the New York Stock Exchange. The EEMD decomposition of the weekly price index of U.S. West Texas light crude oil futures from May 30, 2010, to May 16, 2021, were analyzed using R i386 4.0.5 software and according to the output of the EEMD decomposition results of crude oil prices in Figure 3, IMF1 to IMF8 are the eight eigenmode functions with high to low volatility frequencies, where R is the residual component (long-term trend term).

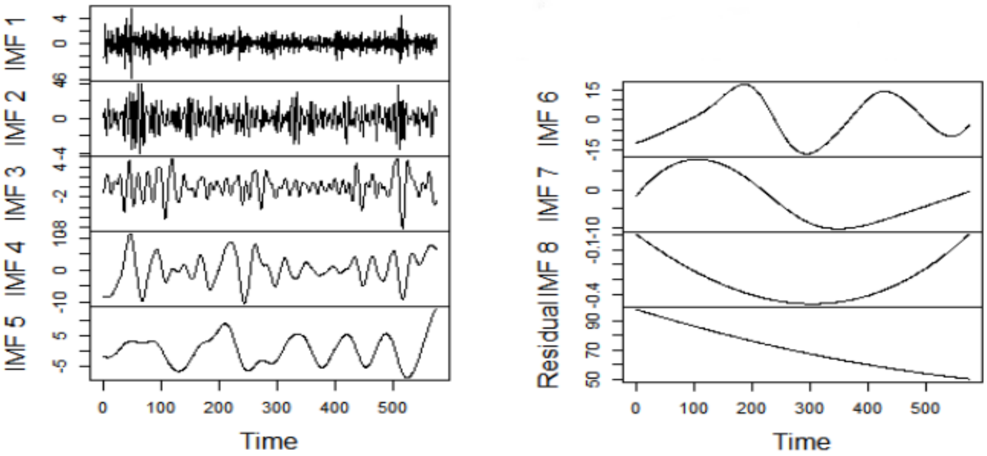

Figure 3. Graph of EEMD decomposition results for crude oil prices.

Then, the correlation between the intrinsic mode functions IMF1 to IMF8 was analyzed using SPSS 26.0 software and the output Pearson correlation coefficient values are shown in Table 1. As seen from the table, the correlation coefficients between IMF1 and IMF2, IMF2 and IMF3, IMF3 and IMF4, IMF5 and IMF6, IMF6 and IMF7 and IMF7 and IMF8 are significant at the 1% significance level. Among them, IMF7 and IMF8 have the largest Pearson correlation coefficient value of 0.49, followed by IMF4 and IMF5 with a correlation coefficient value of 0.34 and again IMF3 and IMF4 with a correlation coefficient value of 0.29. The above correlation output shows the possibility of cluster analysis for each eigenmode function.

Table 1. Pearson correlation coefficient table.

	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8
IMF1	1							data
IMF2	0.29***	1						
IMF3	0.02	0.23***	1					
IMF4	0.02	0.03	0.29***	1				
IMF5	0.01	0.02	−0.03	0.34***	1			
IMF6	0.00	0.01	−0.04	0.04	0.2***	1		
IMF7	0.00	0.00	0.01	0.01	0.05	0.23***	1	
IMF8	−0.01	0.00	0.03	0.00	0.07	−0.2***	0.49***	1

Note: (*** indicates significant correlation at the 1% (two-tailed) level.)

The clustering analysis of IMF1 to IMF8 components was performed by the rect.hclust function of the R software and three classes were obtained according to the output results (Figure 4), which were IMF1 classified as a separate class, i.e., high-frequency terms; IMF2 classified as a separate class, i.e., medium-frequency terms and IMF3 to IMF8 classified as a class, i.e., low-frequency terms.

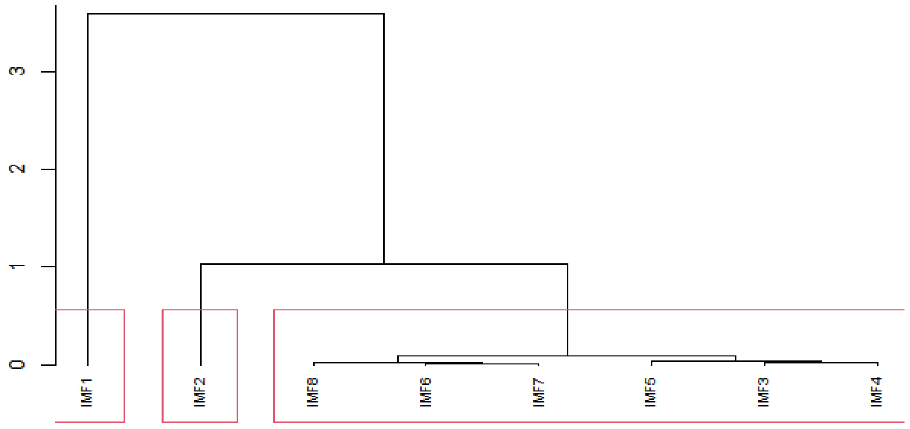

Figure 4. Clustering diagram based on the correlation coefficient.

Based on the output results of the cluster analysis, the original mode function was reconstructed and divided into high-frequency, medium-frequency, low-frequency and trend terms and the time series trends of different volatility frequencies, trend terms and WTI crude oil prices were plotted. The low-frequency series in the graph showed characteristics other than periodicity and trend. To further analyze the data characteristics of each subitem, the basic situation of the period of each subitem, the Pearson correlation coefficient of each item with the original data series and the contribution rate to the variance of the original WTI crude oil futures price series were calculated by Python software (as shown in Table 2).

Table 2. Subseries period and variance contribution rate.

	Correlation coefficient with the original series	p value	Periodicity	Variance contribution rate (%)
High Frequency	0.07	0.11	2.94	0.27
Medium Frequency	0.10	0.02	6.59	0.29
Low Frequency	0.81	0.00	20.46	47.28
Trend	0.73	0.00	-	34.69

As seen in Table 2, the high-frequency series (IMF1) and the medium-frequency series (IMF2) are short-term volatility series, which are WTI crude oil futures price series generated by uncertainty shocks with relatively short fluctuations in time. The volatility period of the high-frequency series is 2.938 weeks and the correlation coefficient with the original series is 0.07, which is statistically insignificant, the volatility period of the medium-frequency series is 6.586 weeks and the correlation coefficient with the original series is 0.1, which is insignificant at the level of 0.01. The degree of explanation of the original series by the high-frequency and medium-frequency series is 0.269% and 0.292% respectively, which shows that both the high-frequency and medium-frequency terms have little explanatory power for the WTI crude oil futures price series. The high-frequency and medium-frequency fluctuations contain the role of uncertainties such as market, nature and investor psychology, which reflect the influence of some short-term noise on WTI crude oil futures prices.

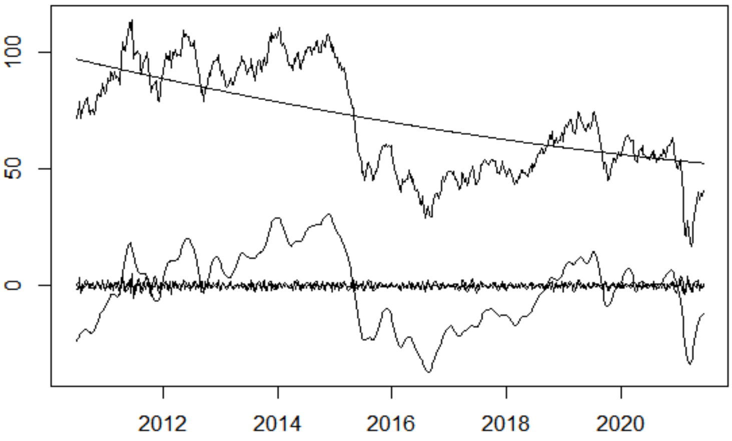

Figure 5. Crude oil price series with reconstructed series. The upper half of the graph is the original trend and trend term of crude oil futures prices, while the lower half is the decomposed and reconstructed high, medium and low frequency terms, where the high and medium frequencies fluctuate up and down around 0 and the low frequency term is basically the same as the original price trend but smoother.

The low-frequency series (IMF3 to IMF8), i.e., the major event impact series, has an average period of change of approximately 20.464 weeks and the correlation coefficient between the low-frequency volatility series and the original series is statistically significant and 0.81. The degree of explanation of the volatility of the original series accounts for 47.275% of the explanation of the WTI crude oil price. The low-frequency term is the cyclical cycle of WTI crude oil price fluctuations and is the most important component of crude oil price fluctuations. As seen in Figure 5, the low-frequency term is basically consistent with the fluctuation of crude oil prices and the interval of larger fluctuation of the low-frequency term reflects the fluctuation interval of crude oil prices affected by major events. For example, the war in Libya in 2011 and the Arab Spring campaign that lasted for several years afterwards caused international crude oil prices to reach \$124/barrel, \$122/barrel and \$117/barrel between 2011 and 2013, respectively and the attack on Syrian government forces in August 2013 caused international oil prices to rise. The Saudi King became seriously ill at the end of 2014 and investors were concerned about the change in the Saudi regime and the resulting policy changes, which led to short-term sustained strength in crude oil prices. 2015 saw the strong rise of U.S. shale oil and a shift in global crude oil supply and demand from a seller's market to a buyer's market. In 2017, the OPEC+ countries, led by Saudi Arabia and Russia, decided to sign a production limit offer agreement and began to implement the Joint Production Agreement. In December 2018, a new round of production cuts negotiated by OPEC+ countries came into effect, leading to a decline in international crude oil production and another change in the supply and demand relationship in the international crude oil market. The worldwide spread of the new coronavirus pneumonia epidemic in the second half of 2020 led to a decline in global demand for crude oil in various countries and according to the monthly report released by OPEC during this period, the expected growth rate of global crude oil demand in 2020 was revised from the original 920,000 barrels per day to 60,000 barrels per day. The downward revision was more than 90% and crude oil prices fell precipitously during this period.

The trend term occupies an important position among all components, with a correlation coefficient of 0.73 with the original series and a 34.694% strength of explanation of the original series, which explains the crude oil price. Trend prices are the main component of WTI crude oil prices and play a key role in the evolution of crude oil price movements. Despite the volatility of crude oil prices within a decade due to the influence of world politics and the unexpected world, crude oil prices will eventually gradually return to the vicinity of the trend price series due to the supply and demand relationship as time gradually passes.

4.2. WTI crude oil price forecast and comparison

According to the construction and prediction process of the combined model, different methods were used to predict each sequence after reconstruction, and the results are shown in Table 3.

As seen in Table 3, in the prediction of high-frequency sequences, ELM (extreme learning machine) and GM (1,1) gray prediction have the smallest MAPE value, while TDNN (time-delayed neural network) and MLP (multilayer perceptron) have MAPE values greater than 1. Meanwhile, ELM is smaller than the GM (1,1) model in MAE, MAPE and RMSE indexes, so the prediction of high-frequency sequences in the ELM neural network has the best prediction effect and the R software was used to determine the extreme learning machine with 4 layers of neural nodes in the input layer, 1 layer of neural nodes in the output layer and 100 layers of neural nodes in the hidden layer. For the medium-frequency sequences, the MAPE values of the four prediction models are all greater than 1 and the model prediction accuracy is not high. Considering that the MAE and RMSE values of the ELM model are not much different from those of GM (1,1), the GM (1,1) model is chosen for the prediction of the medium-frequency series. In the prediction of low-frequency terms, the TDNN has fewer prediction evaluation indexes than the other models, so a time-delayed neural network model with 7 input layers, 1 output layer and 4 hidden layers is constructed by the R software nnetar function for the prediction of low-frequency terms. For the prediction of trend terms, through the comparison of MAPE, RMSE and MAE, the MLP model is smaller than the TDNN, ELM and GM (1,1) models in all indicators, so the MLP neural network model with an input layer of 1, an output layer of 1 and a hidden layer of 5 is constructed for the prediction of trend terms by using the nnfor package of R software.

Table 3. Comparison of predicted effects.

		High Frequency	Medium Frequency	Low Frequency	Trend
TDNN	MAE	1.1546	0.6712	0.6783	2.1702
	MAPE	1.1003	1.3640	0.0328	0.0225
	RMSE	1.4580	0.9029	0.9180	2.4350
ELM	MAE	1.1207	0.9470	2.0885	0.5116
	MAPE	0.8676	1.1600	0.1037	0.0053
	RMSE	1.4293	1.1884	2.5533	0.7015
GM (1,1)	MAE	1.1907	1.0566	18.4303	0.7784
	MAPE	0.9464	1.0029	0.9540	0.0081
	RMSE	1.4740	1.2281	18.5840	0.7793
MLP	MAE	1.1091	0.8965	4.7841	0.0268
	MAPE	1.0278	1.2717	0.2352	0.0003
	RMSE	1.4220	1.1992	5.8658	0.0344

Through the above analysis to determine the high-frequency, medium-frequency, low-frequency and trend terms of their respective prediction of the best model respectively, modeling and prediction on their respective sequences and their prediction results are summed to obtain the final prediction value, the residual terms in the output results of the ARCH effect test and the output of the test shows that the ARCH effect of the residual series in the model constructed in this paper is statistically significant, i.e., there is a considerable degree of thick tail and volatility aggregation in this series, which will lead to an invisible amplification of the risk in the market. We therefore propose to use the EGARCH model as a variance model (for conditional variance) to model the residual series. The prediction results of the fitted GARCH model are integrated with those of the previous combined model and compared with the prediction results of the model without eGARCH. The prediction evaluation results show that after adding the GARCH model, the MAPE of the model decreases from 0.1933 to 0.1632, the RMSE value decreases from 15.030 to 14.026 and the MAE value decreases from 13.598 to 12.594. The values of all evaluation indexes have decreased, which shows that the prediction effect of the model has been further improved at this time.

In order to prove that the combination model in this paper has relative effectiveness, single model forecasting methods such as GM (1,1), ELM, MLP and TDNN and decomposition forecasting methods such as EEMD-GM (1,1), EEMD-ELM, EEMD-MLP and EEMD-TDNN are used to forecast the price of WTI crude oil futures and compare the forecasting effects and the specific results are shown in Figure 6. Among the decomposition prediction methods, the same clustering is carried out after decomposition to form high middle and low frequency and trend terms. A single ELM or MLP method is used to predict each sequence and the prediction results of each frequency are summed up to get the final prediction value and compared with the real value in the training set. Their prediction effect is given by using the prediction evaluation index.

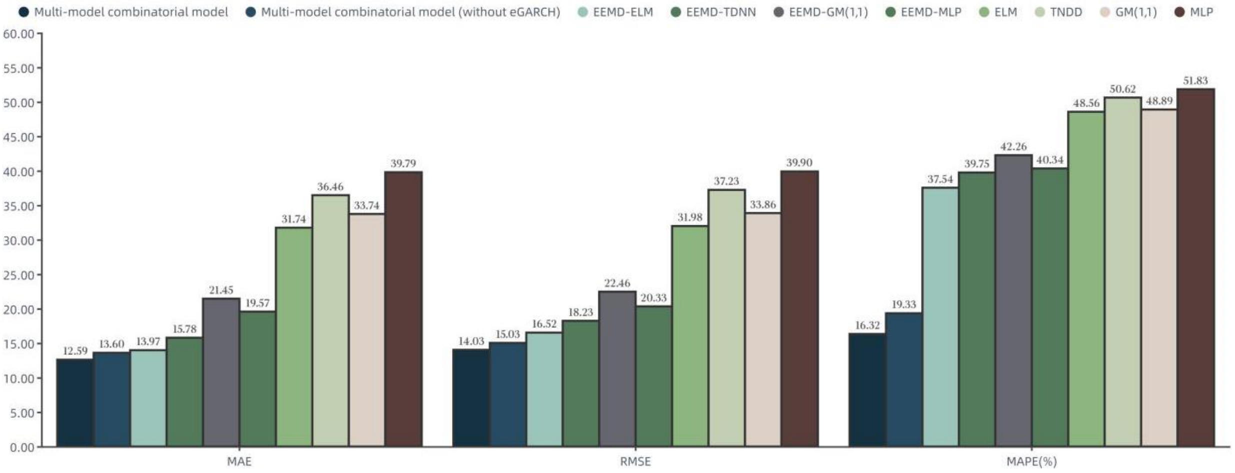

Figure 6. Effectiveness of different models on oil price forecasting.

As seen from Figure 6, in the forecasting of WIT crude oil prices with MAE, MAPE and RMSE as the evaluation indexes, the effect of the combined forecasting model is significantly better than that of the single model, and the values of MAE and RMSE in the evaluation indexes of the combined model are much smaller than those of the ordinary single model. In the combined model, the prediction evaluation indexes of the combined model established in this paper are smaller than those of the EEMD-ELM and the EEMD decomposition reconstruction of the WIT crude oil price series in this paper can give certain economic meanings to the reconstructed categories. Therefore, the combined model constructed in this paper is obviously better than other models in the prediction of WIT crude oil prices.

5. Conclusions and shortcomings

In this paper, the WTI crude oil price is decomposed into 8 IMF series and trend term series residuals through ensemble empirical modal decomposition (EEMD). Then, the decomposed IMF series are subjected to cluster analysis based on correlation coefficients and the volatility series of WTI crude oil prices with different frequencies are obtained from them. Then, the high- and medium-frequency volatility, low-frequency volatility and long-term trend term of WTI crude oil futures prices are constructed. The ELM, GM (1,1), TDNN and MLP models are constructed to forecast the high-frequency, medium-frequency, low-frequency and trend terms of crude oil price fluctuations respectively and the component model with the best prediction effect is selected through the evaluation and comparison of prediction accuracy. Then, the results of each component are integrated to obtain the total model. Finally, the prediction effect of the model is further optimized through the EGARCH model and the final prediction value is obtained. The final prediction value is obtained by the EGARCH model. The results of the empirical analysis yielded the following conclusions:

1. The combined model established in this paper has higher forecasting accuracy and better explanatory results. WTI crude oil prices are affected by a variety of factors such as international

politics, local conflicts, supply and demand and investor psychology, which lead to multiple volatility patterns. A single forecasting model only focuses on linear or nonlinear development trends, while the combined series after EEMD decomposition can take into account different volatility frequencies of crude oil prices. The combined model is often better than the single model in terms of the prediction effect. Meanwhile, the combined model constructed in this paper takes into account the combination of optimal forecasting models among components and the extraction of heteroskedasticity information, so it has a better forecasting effect than the simple combined model. For example, the forecasting evaluation of the combined model in this paper is obviously better than that of the EEMD-ELM model.

2. In this paper, the extracted series of crude oil price fluctuations can better discriminate the development direction of WTI crude oil prices. After decomposition and clustering reconstruction of WTI crude oil prices using ensemble empirical modal decomposition, the main development direction of crude oil price fluctuations is objectively described. The low-frequency term and the trend of the original series and its high Pearson correlation coefficient value of 0.81 show that the crude oil price series after stripping the high-frequency and medium-frequency fluctuations and the long-term trend retains the crude oil trend well, reflecting the main characteristics of the series after eliminating the fluctuations of short-term uncertainties and giving economic meaning to it.

3. For the prediction effect of different volatility frequencies of WTI crude oil prices, the medium frequency series has the worst effect and the trend term has a better prediction effect. For the prediction of medium-frequency fluctuations, the accuracy of all four models is indifferent and only the GM (1,1) method has a slightly better prediction evaluation effect compared with the other models. For the prediction of the trend term, except for TDNN which is relatively poor, the other three methods are more reasonable, among which the MLP neural network has the highest prediction accuracy.

Although the computational intelligence-based forecasting model developed in this paper alleviates the nonlinear series problem to a certain extent and presents a significant optimization effect in comparison with the traditional linear assumption-based model forecasting effect, it does not achieve a good treatment of the heteroskedasticity of the series. Other GARCH-derived models can be tested in future work to predict confidence intervals for the volatility of crude oil series.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

All authors declare no conflict of interest in this article.

References

Abdoos AA (2016) A new intelligent method based on combination of VMD and ELM for short term wind power forecasting. *Neurocomputing* 203: 111–120. https://doi.org/10.1016/j.neucom.2016.03.054

Akram QF (2020) Oil price drivers, geopolitical uncertainty and oil exporters' currencies. *Energ Econo* 89: 104801. https://doi.org/10.1016/j.eneco.2020.104801

Baumeister C, Kilian L (2015) Forecasting the real price of oil in a changing world: a forecast combination approach. *J Bu Econ Stat* 33: 338–351. https://doi.org/10.1080/07350015.2014.949342

Baumeister C, Kilian L (2016) Understanding the Decline in the Price of Oil since June 2014. *J Assoc Enviro Resour Econ* 3: 131–158. https://doi.org/10.1086/684160

Bouoiyour J, Selmi R, Hammoudeh S, et al. (2019) What are the categories of geopolitical risks that could drive oil prices higher? Acts or threats? *Energ Econ* 84: 104523. https://doi.org/10.1016/j.eneco.2019.104523

Caldara D, Iacoviello M (2022) Measuring geopolitical risk. *Am Econ Rev* 112: 1194–1225. https://doi.org/10.1257/aer.20191823

Chen HH, Chen M, Chiu C C (2016) The integration of artificial neural networks and text mining to forecast gold futures prices. *Commu Stat-Simul C* 45: 1213–1225. https://doi.org/10.1080/03610918.2013.786780

He M, Zhang Y, Wen D, et al. (2021) Forecasting crude oil prices: A scaled PCA approach. *Energ Econ* 97: 105189. https://doi.org/10.1016/j.eneco.2021.105189

Huifeng L (2017) Price forecasting of stock index futures based on a new hybrid EMD-RBF neural network model. *Agro Food Ind Hi-Tech* 28: 1744–1747.

Morana C (2001) A semiparametric approach to short-term oil price forecasting. *Energ Econ* 23: 325–338. https://doi.org/10.1016/S0140-9883(00)00075-X

Phan DHB, Narayan PK, Gong Q (2021) Terrorist attacks and oil prices: Hypothesis and empirical evidence. *Int Re Financ Anal* 74: 101669. https://doi.org/10.1016/j.irfa.2021.101669

Plourde A, Watkins GC (1998) Crude oil prices between 1985 and 1994: how volatile in relation to other commodities? *Resour Energy Econ* 20: 245–262. https://doi.org/10.1016/S0928-7655(97)00027-4

Tian LH, Tan DK (2015) Analysis of the factors influencing crude oil prices: financial speculation or Chinese demand? *Economics (Quarterly)* 14: 961–982.

Wang C (2016) Forecast on price of agricultural futures in China based on ARIMA model. *Asian Agric Res* 8: 9. https://doi.org/10.22004/ag.econ.253255

Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv Adapt Data Analysis* 1: 1–41. https://doi.org/10.1142/S1793536909000047

Yin L, Yang Q (2016) Predicting the oil prices: Do technical indicators help? *Energ Econ* 56: 338–350. https://doi.org/10.1016/j.eneco.2016.03.017