

Research Article

Forecasting the Direction of Short-Term Crude Oil Price Changes with Genetic-Fuzzy Information Distribution

Xinyu Wang , Kegui Chen , and Xueping Tan

School of Management, China University of Mining and Technology, China

Correspondence should be addressed to Kegui Chen; guikechen@sina.com

Received 22 June 2018; Revised 29 October 2018; Accepted 21 November 2018; Published 5 December 2018

Academic Editor: Juan P. Amezcua-Sanchez

Copyright © 2018 Xinyu Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper proposes a novel approach to the directional forecasting problem of short-term oil price changes. In this approach, the short-term oil price series is associated with incomplete fuzzy information, and a new fused genetic-fuzzy information distribution method is developed to process such a fuzzy incomplete information set; then a feasible coding method of multidimensional information controlling points is adopted to fit genetic-fuzzy information distribution to time series forecasting. Using the crude oil spot prices of West Texas Intermediate (WTI) and Brent as sample data, the empirical analysis results demonstrate that the novel fused genetic-fuzzy information distribution method statistically outperforms the benchmark of logistic regression model in prediction accuracy. The results indicate that this new approach is effective in direction accuracy.

1. Introduction

It is well documented that the oil price has strong connection with the business cycle, macroeconomics variables, global economic conditions, and policy uncertainty [1–3]. In addition, oil price contributes a crucial risk factor to explain cross-sectional asset prices in equity market and derivative market [4]. Therefore, government and financial institutions such as US Energy Information Administration, Bank of Canada, and Deutsche Bank regularly publish the short-term forecasts of the crude oil price; and oil forecasting price has attracted a growing large number of literature among economists (see [5]), statistics (see [3, 6–8]), econometricians (see [9–11]), engineers (see [12–15]), policy makers (see [1]), and market players (see [16]).

The oil price, or equivalently, the return of oil price can be decomposed as a product of two components: the direction of price change (or the sign of log return) and change magnitude (the absolute value of the return). While the forecasting models of variance have been widely developed in statistics and econometrics like Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models, however, the magnitude of directional change is less understood and less developed in literature except for a few remarkable

exceptions. Several methods have been utilized in literature for forecasting of oil price; for example, Ghaffari developed a soft computing approach to predict the daily variation of the West Texas Intermediate (WTI) crude oil price, and adopted the direction prediction accuracy ratio to prove its effectiveness (see [14]). Murat investigated whether there was a causal relationship between the crack spread futures and the spot oil markets in a vector error correction framework and found that the crack spread futures could be a good predictor of oil price movements (see [17]). Shin applied semisupervised learning to forecast the crude oil price prediction of WTI from January 1992 to June 2008 (see [18]). Tang et al. [19] proposed an ensemble learning paradigm coupling complementary ensemble empirical mode decomposition (CEEMD) and extended extreme learning machine (EELM) to enhance the prediction accuracy for crude oil price, and indicated the model's superiority. Yu et al. [20] built a decomposition-and-ensemble forecasting model with extended extreme learning machine for crude oil price prediction and showed high accuracy, time saving, and robustness. Wang et al. [16] used a Markov switching multifractal (MSM) volatility model to forecast crude oil return volatility. Similar to stock market, options market and short-term load forecasting problem (see [21, 22]), the world crude oil market is also an important

commodity trading market with a large number of investors participating in trading. There is considerable amount of literature devoted to forecasting crude oil price, such as nonlinear models (see [23–25]), econometric models (see [9–11]), and directional forecasting models (see [26–31]).

Several other researchers have analyzed the directional forecasting problem in other contexts of economics. The directional accuracy as a reasonable utility-based measure of forecasting performance was also advocated by Engel and Hamilton [32]. In equity market, Pierdzioch et al. [33] found that an asymmetric loss function often (but not always) makes forecasts look rational. Christoffersen and Diebold [26] found the asset return signs and asset return volatilities were closely interrelated, and volatility dependence was entirely consistent with the direction of the market. And lastly, Merino and Albacete [5] constructed a congruent econometric model with financial and fundamental variables and analyzed the relative weight of the variables to explain the short-term oil price forecast. Thomakos and Wang [34] studied the directional predictions of financial returns and the probability of returns exceeding a threshold. Ahn et al. [35] employed an intelligent approach for directional forecasting in the options market, and showed the reasonably strong performance with empirical study. Rulke and Pierdzioch [36] used ROC-techniques (receiver operating characteristic) to study the directional accuracy of forecasts with respect to directional changes of exchange rates. This paper focuses on the short-term oil price change directional forecasting as an application of the fuzzy information distribution theory.

Through the literature review, most of the studies originated from modeling the oil price series as a complete information structure, and less studies have perceived the fuzziness of oil price information. This paper contributes to literature by presenting a new approach to forecast the directional change of oil price. In our approach, the crude oil price series is associated with an incomplete and inaccurate data set due to possible missing samples or noise information. To deal with incomplete data, a fuzzy framework for incomplete data is developed to explore the true hidden data generation process. Huang (see [37–39]) gave an introduction to fuzzy information distribution method. This paper provides some important extensions for methodical perfection especially for directional forecasting. The fuzzy information distribution method emphasizes that there exists certain transition trend of price series from incompleteness to completeness, and that each sample point has the trend to be evolved into multiple points; thus, the oil price time series corresponds to an incomplete fuzzy information set. Each oil price sample should not be regarded as an independent point. Instead, it should be regarded as the fuzzy information that has certain influence area around it but with variable degrees. Since the long-term oil price distribution and fluctuation patterns may have high instability or structural change, it is intended to implement fuzzy modeling, fuzzy inference and fuzzy pattern recognition on the fluctuation of the short-term oil price series and to reveal the intrinsic nonlinearity of short-term oil price series.

In this paper, a new fused genetic-fuzzy information distribution method is proposed to predict the direction of

the short-term crude oil price changes in a fuzzy incomplete information setting. The adjustable weighted sum of the reciprocal of directional accuracy and root of mean square error (RMSE) are set as the fitness function to identify the solution with the least RMSE under the same directional accuracy. The numbers of information controlling points and lagged order of returns in fuzzy information distribution are optimized by the presented genetic algorithm. A coding algorithm of multidimensional information controlling points are further introduced to fit genetic-fuzzy information distribution to time series forecasting, with the crude oil spot prices of WTI and Brent as sample data, an empirical analysis on the oil price changes are adopted for following the presented approach. We select logistic regression as a benchmark model to compare the directional forecasting accuracy.

The rest of the paper is organized as follows. In Section 2, the new genetic-fuzzy information distribution model is proposed and analyzed; the empirical analysis is presented in Section 3. Finally the conclusions are given in Section 4.

2. A Genetic-Fuzzy Information Distribution Model

The fuzzy information distribution theory has been considerably successful in processing the fuzziness of information, especially when observed information is incomplete or inaccurate. To understand its essential characteristics, the genetic-fuzzy information distribution model is explained next in details.

In fuzzy information distribution method, the objective is to construct a fuzzy inference from X to Y , where X contains a set of independent variables and Y is a dependent variable. The range of each variable is discretized into some information controlling points, so the space of $X \times Y$ is discretized into a multidimensional grid. Each sample supplies a unit of information. Because of the characteristics of incomplete samples, each sample has the latent trend to be evolved into other unobserved neighbour samples. Therefore, an information distribution function is adopted to distribute one unit of sample information over its neighbour information controlling points. Eventually, the information of all samples is rearranged over the crossing points in the multidimensional grid using multidimensional information distribution function. In the end, the obtained information distribution structure in the multidimensional grid reflects the empirical knowledge learned from samples and to form a fuzzy inference mechanism based on the empirical knowledge. However, as far as the authors are concerned, obtaining the optimal empirical knowledge is not resolved.

For the purpose of directional forecasting, the fuzzy information distribution theory is extended in several important theoretical aspects. (1) The fuzzy information distribution is fused and a new fuzzy forecasting model is developed, a genetic algorithm by using the weighted sum of the reciprocal of direction accuracy and root of mean square error (RMSE) are developed as the fitness function in the genetic algorithm. The role of genetic algorithm is to search the “optimal

parameters” for enhancing the quality of fuzzy reasoning, which has been missed in the field of fuzzy information distribution theory and its applications. (2) It demonstrates that there is no sample information loss throughout the multi-dimension linear information process. (3) In order to fit fuzzy information distribution to the oil price time series analysis, a coding algorithm of multidimensional information controlling points is adopted, which can maintain the temporal structure of time series data particularly for large sample applications.

2.1. One-Dimensional Linear Information Distribution. Let $\Omega = \{(r_t, z_t), t = 1, \dots, T\}$ denote a set of observations with m -dimension input indexes $r_t = (r_{1t}, r_{2t}, \dots, r_{mt})$ and one-dimension output index z_t . Let U_i be the universe of the i th input index, and V be the universe of z_i , that is, $(r_{1t}, r_{2t}, \dots, r_{mt}, z_t) \in \Omega \subset (\prod_{i=1}^m U_i) \times V$.

Without loss of generality, the output variable z_t is taken as an example to introduce one-dimensional linear information distribution. Let $V^d = \{v_1, v_2, \dots, v_n\} \subset V$, where V represents a discrete universe and V^d consists of the important nodes of the output index z . These nodes are called *information controlling points*, which can be characterized according to expert experiences, while a simple approach is to choose information controlling points in the universe of equidistance. According to the transition property of sample points, the information distribution process is a fuzzy partition process of sample information on V^d .

Let $V = [a, b]$, $v_1 = a$, and $v_n = b$, and the rest of information controlling points are chosen in V equidistantly, that is,

$$v_k = v_1 + (k-1) \frac{b-a}{n-1}, \quad k = 2, \dots, n. \quad (1)$$

For any one sample z_t , each information controlling point v_k leads to an information gain by an one-dimensional linear information distribution function,

$$\mu_{m+1}(z_t, v_k) = \begin{cases} 1 - \frac{|z_t - v_k|}{\Delta_{m+1}}, & |z_t - v_k| < \Delta_{m+1} \\ 0, & \text{others} \end{cases} \quad (2)$$

$$\forall t, \forall k,$$

where $\Delta_{m+1} = (b-a)/(n-1)$ is called step length of controlling points.

Similarly, for input variables, let $U_i^d = \{u_{i1}, u_{i2}, \dots, u_{ih^i}\} \subset U_i$, $i = 1, \dots, m$, be the discrete universe of U_i with equal step length, and h^i is the number of information controlling points of U_i^d . Then the one-dimension linear information distribution of the input index r_{it} is

$$\mu_i(r_{it}, u_{ij}) = \begin{cases} 1 - \frac{|r_{it} - u_{ij}|}{\Delta_i}, & |r_{it} - u_{ij}| < \Delta_i \\ 0, & \text{others} \end{cases} \quad (3)$$

$$\forall i, \forall t, \forall j,$$

where Δ_i is also evenly spaced.

2.2. Multidimensional Information Distribution Matrix. A multidimensional sample, $(r_{1t}, r_{2t}, \dots, r_{mt}, z_t) \in (\prod_{i=1}^m U_i) \times V$, $t = 1, 2, \dots, T$, offers one unit of information. The multidimensional linear distribution functions are used to distribute it into the multidimensional information controlling points, which is denoted as $(u_{1j_1}, u_{2j_2}, \dots, u_{mj_m}, v_k) \in (\prod_{i=1}^m U_i^d) \times V^d$.

The multidimensional linear distribution function is equal to the product of all one-dimensional linear distribution functions:

$$\begin{aligned} \mu_M[(r_{1t}, \dots, r_{mt}, z_t), (u_{1j_1}, \dots, u_{mj_m}, v_k)] \\ = \left(1 - \frac{|z_t - v_k|}{\Delta_{m+1}}\right) \prod_{i=1}^m \left(1 - \frac{|r_{it} - u_{ij_i}|}{\Delta_i}\right) \end{aligned} \quad (4)$$

as long as $|z_t - v_k| < \Delta_{m+1}$, $|r_{it} - u_{ij_i}| < \Delta_i$, $\forall t \in \{1, \dots, T\}$, $\forall j_i \in \{1, \dots, h^i\}$, $i = 1, \dots, m$, $\forall k \in \{1, \dots, n\}$; otherwise, $\mu_M[(r_{1t}, \dots, r_{mt}, z_t), (u_{1j_1}, \dots, u_{mj_m}, v_k)] = 0$.

In practical application of fuzzy information distribution, lots of multidimensional information controlling points like $(u_{1j_1}, \dots, u_{mj_m})$, are generated. The information gains on all multidimensional information control points, obtained from each sample $(r_{1t}, r_{2t}, \dots, r_{mt}, z_t)$, $t = 1, \dots, T$, should be accumulated. After doing it, the information structure stored in the multidimensional information controlling points is one kind of empirical knowledge learned from samples, and serves as the basis of fuzzy inference. Therefore, a valid encoding rule is provided to identify all multidimensional information controlling points next.

To maintain the temporal structure of time series data, a coding method with m digit number in base H is proposed to record the multidimensional information controlling points, in which $H = \max_{i=1, \dots, m} \{h^i\} + 1$. That is to say, the multidimensional information controlling point $(u_{1j_1}, \dots, u_{mj_m})$ is coded as $j_1 \dots j_m$, a m digit number in base H . In practical computation process, $j_1 \dots j_m$ need to be converted into a decimal number:

$$f(j_1 \dots j_m) = \sum_{i=1}^m j_i H^{i-1}, \quad 1 \leq j_i \leq h_i, \quad i = 1, \dots, m. \quad (5)$$

By using this method, dynamic non-repetitive coding according to the number of information controlling points of each index are conveniently realized. The coding method enables us to keep the logical temporal structure of time series data, and therefore, facilitate the application of fuzzy information distribution in time series analysis.

Through multidimensional information distribution, all information controlling points with $h^1 \times h^2 \times \dots \times h^m \times n$ dimensions are generated, but it cannot be directly used to make fuzzy inference. In order to clearly express the two-dimensional information conversion relationship from the inputs to the output, the coding method is adopted to integrate all inputs into one new input. More specifically, let $X = \{x\} = \prod_{i=1}^m U_i^d$, $Y = \{y\} = V^d$, the multidimensional information controlling point set X can be encoded as

$$\begin{aligned}
X &= \prod_{i=1}^m U_i^d \\
&= \{f(j_1, \dots, j_m) \mid j_i \in \{1, \dots, h^i\}, i = 1, \dots, m\}.
\end{aligned} \tag{6}$$

Finally, the two-dimensional information distribution matrix from X to Y is formed under all observed samples, which is written as

$$\begin{aligned}
Q(X, Y) &= \{q(x, y)\} \\
&= \left\{ \sum_{t=1}^T \mu_M \left[(r_{1t}, \dots, r_{mt}, z_t), (u_{1j_1}, \dots, u_{mj_m}, v_k) \right] \mid j_i \right. \\
&\quad \left. \in \{1, \dots, h^i\}, i = 1, \dots, m, k \in \{1, \dots, n\} \right\}.
\end{aligned} \tag{7}$$

As shown by the next result, the whole information gains on all multidimensional information controlling points from sample information are preserved; in other words, there is no sample information loss through the multidimensional linear information process.

2.3. The Fuzzy Relationship Matrix R . The distribution matrix $Q(X, Y)$ is the basis and starting point of fuzzy inference. Two types of fuzzy relationship matrixes R can be generated from $Q(X, Y)$. The first one is to construct a fuzzy relationship matrix based on fuzzy concept. According to the theory of the factor space proposed by Huang [37] and Jun and Kang [40], when $X \times Y$ is a general factor space, an element corresponds to a fuzzy concept. Supposing that Y is a set of state space with fuzzy concept, any element y in Y is a fuzzy concept. X can be regarded as Y 's domain; hence the fuzzy membership of X about Y can be constructed from $Q(X, Y)$, yielding a fuzzy relation matrix R_f , that is,

$$\begin{aligned}
R_f &= \left\{ \frac{q(x, y)}{s(y)} \right\} \\
s(y) &= \max_{x \in X} \{q(x, y)\}.
\end{aligned} \tag{8}$$

Alternatively, the set-valued statistics and conditional falling shadow formula based the falling shadow theory are used by Jun and Kang [40], to generate another fuzzy relation matrix R_s . For discrete $X \times Y$, the conditional falling shadows of Y on X are obtained,

$$\xi(y_j \mid x_i) = \frac{q(x_i, y_j)}{\sum_{j=1}^n q(x_i, y_j)}. \tag{9}$$

The fuzzy relation matrix is

$$R_s = \{r_{ij}\} = \{\xi(y_j \mid x_i)\}. \tag{10}$$

Both these two approaches will be used in our model below. We now move on to discuss fuzzy inference process and use it to deal with incomplete information data set.

2.4. Fuzzy Inference. A fuzzy inference process is, in essence, a fuzzy transform from an input fuzzy set \underline{A} defined on X to an output fuzzy set \underline{B} defined on Y . Given a fuzzy relation matrix R , \underline{B} can be obtained by fuzzy operating like,

$$\underline{B} = \underline{A} \circ R. \tag{11}$$

Suppose that (r_{1t}, \dots, r_{mt}) is the input vector of a new sample; our aim is to infer out the possible output. The sample information on the space of $\prod_{i=1}^m U_i$ are distributed, and then the information gains on controlling points are equivalent to the fuzzy membership of (r_{1t}, \dots, r_{mt}) on multidimensional information controlling points, which constitutes the fuzzy set \underline{A} on X ,

$$\begin{aligned}
\underline{A} &= \{p(x)\} \\
&= \left\{ \mu_M \left[(r_{1t}, \dots, r_{mt}), (u_{1j_1}, \dots, u_{mj_m}) \right] \mid j_i \right. \\
&\quad \left. \in \{1, \dots, h^i\}, i = 1, \dots, m \right\}.
\end{aligned} \tag{12}$$

For fuzzy inference on R_f in last section, the classical max-min inference method are applied,

$$\underline{B}(y) = \sup_{x \in X} \{\underline{A}(x) \wedge R(x, y)\}, \quad \forall y \in Y. \tag{13}$$

For inference for R_s , we know that R_s is formed by the conditional falling shadow formula, so the inference using total falling shadow formula is proper; we obtain

$$\underline{B}(y_j) = \frac{\sum_i \underline{A}(x_i) R(x_i, y_j)}{\sum_i \underline{A}(x_i)}, \quad \forall y_j \in Y. \tag{14}$$

Moreover, two *defuzzification methods* are adopted to transform the fuzzy set \underline{B} into clear forecast values. The first one is the maximum membership principle, and we set the information controlling point with the maximum membership degree as the final recognition result,

$$\hat{y} = \arg \max \{\underline{B}(y_j) \mid j = 1, \dots, n\}. \tag{15}$$

Another one is the weighted average of the information controlling points with the membership degree as weights, that is,

$$\hat{y} = \frac{\sum_{j=1}^n y_j \underline{B}(y_j)}{\sum_{j=1}^n \underline{B}(y_j)}. \tag{16}$$

At last, to simplify notations of *four* fuzzy inference and defuzzification methods stated in (13)-(16), the following notations are used: $R_s - max$ is R_s inference with the defuzzification of maximum membership degree; $R_s - avr$ is R_s inference with the defuzzification of weighted average, $R_f - max$ is R_f inference with the defuzzification of maximum membership degree; $R_f - avr$ is R_f inference with the defuzzification of weighted average. In empirical applications, the forecasting accuracy usually relies on the proper fuzzy inference and defuzzification method, which is described in next section.

2.5. Parameters Estimation

2.5.1. Error Evaluation. In fuzzy information distribution process, the key optimization parameters are: the numbers of input indexes m , the number of input information controlling points h and the number of output information controlling points n . How to select parameter appropriately affects the prediction precision.

The genetic algorithm is applied to optimize the parameter combinations. For this purpose, D_0^s and D_0^f are defined as the in-sample directional prediction accuracy with the maximum membership degree inference for R_s and R_f . Let D_1^s and D_1^f be the out-of-sample directional prediction accuracy with the maximum membership degree inference for R_s and R_f , respectively. D_2^s, D_2^f are the out-of-sample directional prediction accuracy with the weighted average inference for R_s and R_f . Similarly, $RMSE_1^s$ and $RMSE_1^f$ denote the out-of-sample root mean squared errors (RMSE) with the maximum membership degree inference for R_s and R_f respectively; $RMSE_2^s, RMSE_2^f$ are the out-of-sample RMSE with the weighted average inference for R_s and R_f .

Since the directional change of the crude oil spot price is investigated, the logarithmic returns are used as $r_t = \ln(P_t/P_{t-1})$, $t = 1, \dots, T$, where P_t is the close prices at day t . Let \hat{r}_t , $t = 1, \dots, T$, be the forecasted returns and the out-of-sample predicted direction accuracy is

$$D_{1,2}^{s,f} = \frac{1}{T} \sum_{t=1}^T I(\hat{r}_t r_t), \quad I(\hat{r}_t r_t) = \begin{cases} 1, & \hat{r}_t r_t > 0 \\ 0, & \text{others} \end{cases} \quad (17)$$

Evidently, $\hat{r}_t \cdot r_t > 0$ mean that r_t and \hat{r}_t have the same sign. The prediction error of RMSE is defined by:

$$RMSE_{1,2}^{s,f} = \sqrt{\frac{1}{T} \sum_{t=1}^T (\hat{r}_t - r_t)^2}. \quad (18)$$

2.5.2. Fitness Function in Genetic Algorithm. In the fuzzy information distribution theory, the numbers of information controlling points of input and output indexes are the vital parameters of information distribution model, which directly affect the precision of the model. To reduce the number of estimated parameter, we consider a special case where each input index has the same number of controlling points, i.e., $h^i = h$, $i = 1, \dots, m$. The lagged orders of explanatory variables m in time series data applications are also an important parameter to be optimized. Generally, the input index number m , the number of input information points h , and the number of output information point n are three core parameters; however, there exist no literatures to study the trade-off for the best inference. Then, a fusion approach of combining fuzzy information distribution with genetic algorithm is proposed to obtain the optimal parameters. Two fuzzy relation matrixes R_s and R_f and two defuzzification modes comprise four fuzzy inferences modes. In the construction of the fitness function of genetic algorithm, the

reciprocal of out-of-sample directional accuracy and out-of-sample RMSEs are synthesized under four inference methods by weighted average method,

$$f(\lambda, m, h, n) = \lambda \frac{4}{\sum_{i=1}^2 (D_i^s + D_i^f)} + (1 - \lambda) \frac{\sum_{i=1}^2 (RMSE_i^s + RMSE_i^f)}{4} \quad (19)$$

where $(\lambda, 1 - \lambda)$, $0 < \lambda < 1$, is the weight parameter to reflect a preference between directional accuracy and RMSE. One advantage of the fitness function is to integrate the predicted effects of four fuzzy inference modes, and the optimal parameters smooth out the differences of four predicted effects.

2.5.3. Pareto Solutions and Forecasting. There are three steps in finding Pareto solution.

First, an empirical weight set such as $\{(\lambda_l, 1 - \lambda_l), 0 \leq \lambda_l \leq 1, l = 1, \dots, L\}$ are constructed, which reflects the latent preference of decision-makers. Then the optimal solution set, written as $S_a = \{(m^*, h^*, n^*), l = 1, \dots, L\}$ are derived, by a genetic algorithm.

Second, for each $(m^*, h^*, n^*)_l \in S_a$ at given λ_l , in order to distinguish the best fuzzy inference and defuzzification method from $\{R_s - max, R_s - avr, R_f - max, R_f - avr\}$, w the fitness values are decomposed into four predicted combinations of directional accuracy and RMSEs, so $4L$ predicted combinations of directional accuracy and RMSEs denoting as $\mathbf{PCDAR} = \{(D_i^s, RMSE_i^s), (D_i^f, RMSE_i^f)\}_l$, $i = 1, 2, l = 1, \dots, L\}$ are obtained. Because of its nonlinearity feature of the fitness function, some combinations in \mathbf{PCDAR} might not be Pareto optimal towards two goals of the direction accuracy and RMSE; then, we have to find out the Pareto optimal ones in step three below.

Third, given an element $A \in \mathbf{PCDAR}$, if there is no $B \in \mathbf{PCDAR}$ which is better than A in both directional accuracy and RMSE, then A is a pareto optimal case. According to the above definition of Pareto optimization, we screen out the bad combinations from \mathbf{PCDAR} , and identify the Pareto ones inside:

$$\mathbf{PCDAR}_p = \left\{ (D_i^s, RMSE_i^s)_{p_i^s}, (D_i^f, RMSE_i^f)_{p_i^f}, i = 1, 2, p_i^s = 1, \dots, P_i^s, p_i^f = 1, \dots, P_i^f \right\}. \quad (20)$$

Clearly, each combination in \mathbf{PCDAR}_p corresponds to a special parameter setting including the weight λ_p , the fuzzy inference and defuzzification method and the optimal (m, h, n) . Therefore, when decision-makers determine the final element from \mathbf{PCDAR}_p with their preference towards directional accuracy and RMSE, the corresponding parameter setting is also made simultaneously, including $(m, h, n)_f \in S_a, \lambda_f$, fuzzy inference, and defuzzification method. Next, the final parameter setting is used to perform forecasting of new samples.

3. Empirical Analysis

We obtained crude oil daily spot prices from US Department of Energy: Energy Information Administration, unit: Dollars per Barrel. Brent crude oil spot prices (Brent-Europe) and West Texas Intermediate crude oil spot prices (WTI-Cushing, Oklahoma) are extracted from Nov. 13, 2017, to Sep. 28, 2018, with a total of 220 samples. The first 200 samples are used for modeling, while the latter 20 samples are for testing the prediction accuracy. The descriptive statistics of samples are shown in Table 1. The statistics of skewness and kurtosis clearly show that return distributions are not normal, leading to some difficulties in classical econometric modeling. However, as a nonparametric approach, the genetic-fuzzy information distribution method needs not consider such a constraint of return distribution.

3.1. Data Preprocessing. To apply the method of fuzzy information distribution, the first step is to determine the number of input indexes and the number of information controlling points of each index. As is known, the crude oil market assimilates and reacts to the new information with a time delay. The time length of the delay process depends on the maturity degree of crude oil market. So in the model the lagged returns are set as the input indexes (explanatory variables) and the returns of the following day as the output index (dependent variables). Matlab codes are written to carry out the fuzzy inference, and the genetic optimization is done by the GA(x) function in Matlab genetic algorithm toolbox, allowing for integer optimization. In crude oil spot market, the historical short-term price information in about one week is of particularly importance to affect the current oil price; and it also indicate that a larger m does not necessarily enhance the performance substantially. Therefore the initial range of m is set to be $\{4, 5, 6\}$.

The intervals of information controlling points are set as 1-1.1 times the range of sample return, and it can be dynamically adjusted after future oil price data accumulation. In order to simplify the calculation, the information controlling points with equal steps are used. However, if there is enough information, a more flexible approach can be taken to arrange information controlling points in line with experts' knowledge, such as placing more information controlling points over the interested range of return. Another rule is that the number of controlling points should be appropriate. We aim to determine the numbers of controlling points objectively by a genetic optimization algorithm. The initial ranges of h and n are both set to be a wide range of $\{10, 11, \dots, 30\}$.

3.2. Fuzzy Inference and Forecasting

3.2.1. Generate Alternative Solutions. In order to achieve the best prediction effect, several weights ($\lambda_i, 1 - \lambda_i$) are chosen in the interval $[0, 1]$ with a step of 0.05 and then apply the genetic optimization program to determine the optimal parameters under each group of weights. Integer programming with genetic algorithm involves special creation, crossover, and mutation functions that enforce variables to be integers. Taking $\lambda = 0.8$ as an example, the evolving fitness values are

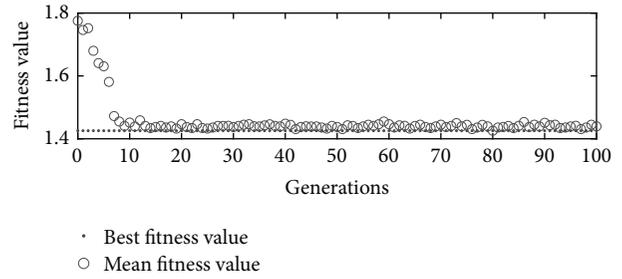


FIGURE 1: The fitness values during optimization for forecasting WTI oil price ($\lambda = 0.8$). The setting of genetic algorithm is as follows: the size of the population is 10; the fraction of the population at the next generation is 0.85; the fraction of individuals in each subpopulation that migrates to a different subpopulation is 0.2; and the optimization terminates when average change in the fitness value is less than $1e-6$. The optimization process converged at the 100th generation. The estimated (m, h, n) is $(5, 10, 27)$ with the best fitness value is 1.426.

shown in Figure 1. The estimated results for Brent and WTI crude oil spot price series are shown in Table 2.

3.2.2. Pareto Solutions. The parameter optimization problem under directional accuracy and RMSE is similar to a multi-objective programming. Because of the complexity of fitness function of fuzzy information model, however, the linear weighted average of two objectives method does not necessarily produce pareto solutions. Since the fitness function of fuzzy information model is the synthesis of four fuzzy inference modes, and each solution includes four sub-solutions, there are 84 sub-solutions in Table 2. All Pareto solutions are identified and the corresponding fuzzy inference methods are explained in Table 3, where RMSE and D are out-of-sample forecasting error and directional accuracy.

The solutions among pareto optimal solutions with much preference to directional accuracy are characterized. For Brent crude oil price series, $(5, 13, 26)$ is the final solution of (m, h, n) by $R_f - max$ inference. For WTI crude oil price series, $(5, 10, 27)$ is the final solution of (m, h, n) by $R_s - avr$ inference. To present spatial structure of inference knowledge, three-dimensional stereograms of information distribution matrix at the final solutions are plotted in Figure 2.

3.2.3. Forecasting. As shown in Table 2, the fuzzy information distribution approach gives a high in-sample forecasted directional accuracy over 0.8. Additionally, under the optimal parameters for Brent and WTI crude oil price series, the out-of-sample one-step-ahead forecasted returns are calculated and then converted into the out-of-sample one-step-ahead forecasted prices. The oil spot prices and forecasting prices are shown in Figure 3.

Clearly, the in-sample predictions are better than the out-of-sample ones. The out-of-sample forecasted values and the direction consistency comparison are presented in Table 4. Besides reaching a good directional accuracy, the forecasted

TABLE 1: The descriptive statistics of oil returns.

| Series | Mean | Maximum | Minimum | Standard deviation | Skewness | Kurtosis |
|--------|----------|----------|-----------|--------------------|-----------|----------|
| Brent | 0.001359 | 0.045343 | -0.044136 | 0.015835 | -0.282877 | 3.24081 |
| WTI | 0.001153 | 0.073341 | -0.052511 | 0.016796 | -0.052855 | 4.89266 |

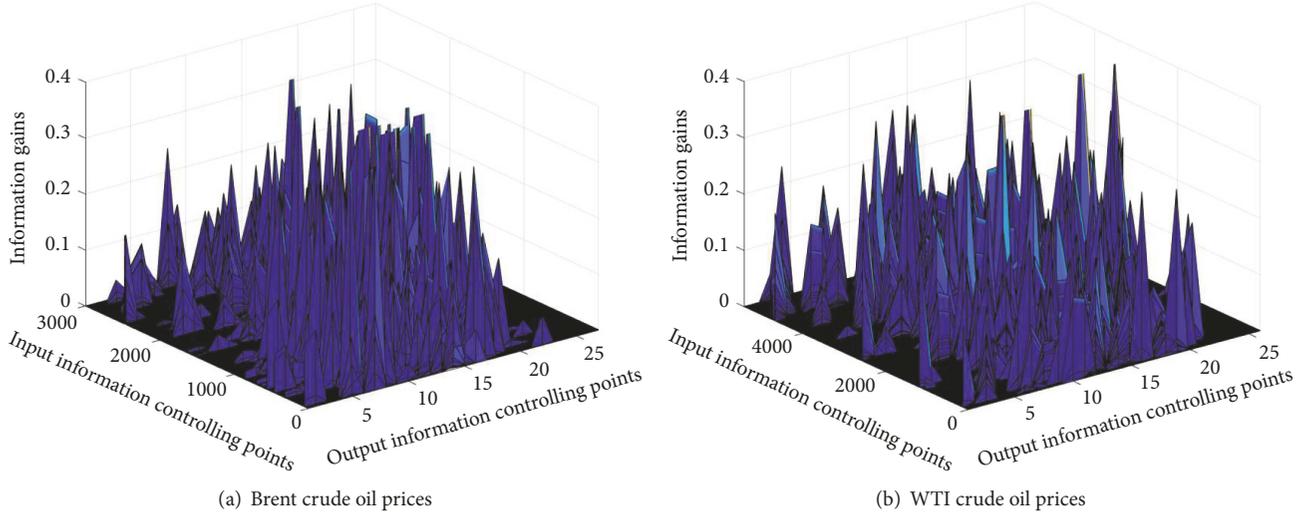


FIGURE 2: The three-dimensional stereograms of information distribution matrix at the final solutions.

prices are also very close to the actual values, because the proper structure of fitness function in (19) can supply an efficient optimization rules to reduce the RMSEs.

By fuzzy inference, a fuzzy set $B(y)$ is defined on the universe of crude oil return by (11). Therefore, the membership function of fuzzy set $B(y)$ does provide a fuzzy probability distribution about oil returns. For example, the out-of-sample probability distributions of WTI oil returns are plotted in Figure 4. Although the forecasted fuzzy possibility curve is not very smooth, it supplies useful information about the uncertainty of oil prices.

3.2.4. Directional Forecasting Comparison. In order to test the effectiveness of the model in directional forecasting, the classical logistic regression and random walk with drift are taken as two benchmark models to compare the directional prediction accuracy. The results are promising.

First, we discuss logistic regression model. The number of lagged input indexes is the only one parameter for logistic regression model. The identified results of logistic regression model are shown in Table 5.

According to the principle of maximum D and maximum AUC in logistic regression, the optimal lagged order for Brent crude oil price series is $m = 5$, and the one for WTI crude oil price series is $m = 6$. For Brent crude oil prices, the optimal directional accuracy (D) of the genetic-fuzzy information distribution model is 0.75 that is higher than the optimal directional accuracy of the logistic regression ($D=0.7$). For WTI crude oil prices, the optimal directional accuracy

(D) of the genetic-fuzzy information distribution model is 0.75 that is also higher than that of the classical logistic regression ($D=0.7$). Because the logistic regression model can only predict the direction of price changes, the proposed genetic-fuzzy information distribution approach has obvious appealing of forecasting the direction, the magnitude and the fuzzy possibility distribution of price changes overall.

Second, we assume that the stock price P_t obeys a geometric random walk process with a dynamic drift as follows:

$$\ln P_t = \ln P_{t-1} + u_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_t^2), \quad (21)$$

where u_t represents a dynamic drift that is used to measure a trend in the price. ε_t is a white noise. The model is equivalent to $r_t = u_t + \varepsilon_t$. The drift u_t equals the moving average of previous m returns. The forecasting error RMSE and directional accuracy D are presented in Table 6. The optimal directional accuracy values for Brent oil prices and WTI oil prices are 0.55 and 0.50 respectively, but they are also lower than the corresponding directional accuracy values from genetic-fuzzy information distribution approach, which are 0.75 and 0.75 for Brent oil prices and WTI oil prices, respectively, in Table 3. Further, we find that genetic-fuzzy information distribution approach for WTI oil prices can reach a high precision with RMSE, D equal to 0.0147, 0.75 in Table 3, which has less RMSE and higher D than the optimal result of random walk with RMSE, D equal to 0.0157, 0.5 in Table 6.

TABLE 2: The estimated results of the genetic-fuzzy in formation distribution approach.

| m | h | n | λ | D_0^s | D_0^f | D_1^s | D_1^f | D_2^s | D_2^f | $RMSE_1^s$ | $RMSE_1^f$ | $RMSE_2^s$ | $RMSE_2^f$ | D^{avr} | $RMSE^{avr}$ |
|------------------------|----|----|-----------|---------|---------|---------|---------|---------|---------|------------|------------|------------|------------|-----------|--------------|
| Brent crude oil prices | | | | | | | | | | | | | | | |
| 5 | 13 | 25 | 1 | 0.9231 | 0.9231 | 0.65 | 0.7 | 0.55 | 0.55 | 0.0145 | 0.0151 | 0.0132 | 0.0131 | 0.6125 | 0.014 |
| 5 | 13 | 26 | 0.95 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 13 | 26 | 0.9 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 13 | 26 | 0.85 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 14 | 11 | 0.8 | 0.8615 | 0.8821 | 0.55 | 0.6 | 0.7 | 0.6 | 0.0133 | 0.0137 | 0.0132 | 0.0137 | 0.6125 | 0.0135 |
| 5 | 14 | 21 | 0.75 | 0.9128 | 0.9385 | 0.65 | 0.45 | 0.7 | 0.65 | 0.0139 | 0.0155 | 0.0132 | 0.0134 | 0.6125 | 0.014 |
| 5 | 13 | 16 | 0.7 | 0.9538 | 0.9538 | 0.6 | 0.65 | 0.55 | 0.6 | 0.0123 | 0.0145 | 0.0132 | 0.0131 | 0.6 | 0.0133 |
| 5 | 13 | 26 | 0.65 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 14 | 21 | 0.6 | 0.9128 | 0.9385 | 0.65 | 0.45 | 0.7 | 0.65 | 0.0139 | 0.0155 | 0.0132 | 0.0134 | 0.6125 | 0.014 |
| 5 | 13 | 26 | 0.55 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 14 | 11 | 0.5 | 0.8615 | 0.8821 | 0.55 | 0.6 | 0.7 | 0.6 | 0.0133 | 0.0137 | 0.0132 | 0.0137 | 0.6125 | 0.0135 |
| 5 | 13 | 26 | 0.45 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 14 | 11 | 0.4 | 0.8615 | 0.8821 | 0.55 | 0.6 | 0.7 | 0.6 | 0.0133 | 0.0137 | 0.0132 | 0.0137 | 0.6125 | 0.0135 |
| 5 | 14 | 11 | 0.35 | 0.8615 | 0.8821 | 0.55 | 0.6 | 0.7 | 0.6 | 0.0133 | 0.0137 | 0.0132 | 0.0137 | 0.6125 | 0.0135 |
| 5 | 14 | 11 | 0.3 | 0.8615 | 0.8821 | 0.55 | 0.6 | 0.7 | 0.6 | 0.0133 | 0.0137 | 0.0132 | 0.0137 | 0.6125 | 0.0135 |
| 5 | 13 | 26 | 0.25 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 13 | 26 | 0.2 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 13 | 26 | 0.15 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 13 | 26 | 0.1 | 0.9538 | 0.9487 | 0.55 | 0.75 | 0.55 | 0.6 | 0.0133 | 0.0146 | 0.0132 | 0.0133 | 0.6125 | 0.0136 |
| 5 | 14 | 11 | 0.05 | 0.8615 | 0.8821 | 0.55 | 0.6 | 0.7 | 0.6 | 0.0133 | 0.0137 | 0.0132 | 0.0137 | 0.6125 | 0.0135 |
| 6 | 13 | 11 | 0 | 0.8608 | 0.8814 | 0.55 | 0.55 | 0.55 | 0.45 | 0.0137 | 0.0137 | 0.0134 | 0.0134 | 0.525 | 0.0135 |
| WTI crude oil prices | | | | | | | | | | | | | | | |
| 5 | 10 | 27 | 1 | 0.9231 | 0.8462 | 0.5 | 0.45 | 0.75 | 0.55 | 0.0159 | 0.024 | 0.0147 | 0.0155 | 0.5625 | 0.0175 |
| 4 | 14 | 22 | 0.95 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 4 | 10 | 29 | 0.9 | 0.8622 | 0.7551 | 0.55 | 0.55 | 0.5 | 0.65 | 0.0194 | 0.0222 | 0.0152 | 0.0161 | 0.5625 | 0.0182 |
| 4 | 10 | 29 | 0.85 | 0.8622 | 0.7551 | 0.55 | 0.55 | 0.5 | 0.65 | 0.0194 | 0.0222 | 0.0152 | 0.0161 | 0.5625 | 0.0182 |
| 5 | 10 | 27 | 0.8 | 0.9231 | 0.8462 | 0.5 | 0.45 | 0.75 | 0.55 | 0.0159 | 0.024 | 0.0147 | 0.0155 | 0.5625 | 0.0175 |
| 4 | 14 | 22 | 0.75 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 4 | 14 | 22 | 0.7 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 4 | 14 | 25 | 0.65 | 0.9082 | 0.8214 | 0.45 | 0.5 | 0.7 | 0.55 | 0.0178 | 0.0235 | 0.0152 | 0.0168 | 0.55 | 0.0184 |
| 4 | 14 | 22 | 0.6 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 4 | 14 | 22 | 0.55 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 4 | 14 | 22 | 0.5 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 4 | 10 | 29 | 0.45 | 0.8622 | 0.7551 | 0.55 | 0.55 | 0.5 | 0.65 | 0.0194 | 0.0222 | 0.0152 | 0.0161 | 0.5625 | 0.0182 |
| 4 | 14 | 22 | 0.4 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 5 | 10 | 27 | 0.35 | 0.9231 | 0.8462 | 0.5 | 0.45 | 0.75 | 0.55 | 0.0159 | 0.024 | 0.0147 | 0.0155 | 0.5625 | 0.0175 |
| 4 | 14 | 22 | 0.3 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 4 | 14 | 22 | 0.25 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 6 | 10 | 13 | 0.2 | 0.7784 | 0.7577 | 0.6 | 0.5 | 0.5 | 0.6 | 0.018 | 0.0249 | 0.0147 | 0.016 | 0.55 | 0.0184 |
| 5 | 10 | 15 | 0.15 | 0.8718 | 0.8462 | 0.5 | 0.45 | 0.75 | 0.5 | 0.0162 | 0.0264 | 0.0147 | 0.0163 | 0.55 | 0.0184 |
| 5 | 10 | 27 | 0.1 | 0.9231 | 0.8462 | 0.5 | 0.45 | 0.75 | 0.55 | 0.0159 | 0.024 | 0.0147 | 0.0155 | 0.5625 | 0.0175 |
| 4 | 14 | 22 | 0.05 | 0.9184 | 0.8878 | 0.45 | 0.55 | 0.7 | 0.65 | 0.0201 | 0.0232 | 0.0152 | 0.0168 | 0.5875 | 0.0188 |
| 6 | 13 | 10 | 0 | 0.8969 | 0.9639 | 0.45 | 0.45 | 0.55 | 0.45 | 0.0143 | 0.0143 | 0.0148 | 0.0151 | 0.475 | 0.0147 |

m is the number of input variables and $h(n)$ are the numbers of input (output) information points. $D_0^{s(f)}$ are the in-sample directional prediction accuracy with the maximum membership degree inference for $R_{s(f)}$. $D_1^{s(f)}$ and $D_2^{s(f)}$ are the out-of-sample directional prediction accuracy for $R_{s(f)}$ with the maximum membership degree inference and the weighted average inference, respectively. $RMSE_1^{s(f)}$ and $RMSE_2^{s(f)}$ are the out-of-sample roots of mean square errors for $R_{s(f)}$ with the maximum membership degree inference and the weighted average inference, respectively. D^{avr} and $RMSE^{avr}$ are the means of out-of-sample directional accuracy and RMSE.

TABLE 3: The Pareto solutions and fuzzy inference methods.

| Series | Inference Method | m | h | n | RMSE | D |
|--------------|------------------|----------|-----------|-----------|---------------|-------------|
| Brent | Rf-max | 5 | 13 | 26 | 0.0146 | 0.75 |
| Brent | Rs-avr | 5 | 14 | 11 | 0.0132 | 0.7 |
| Brent | Rs-max | 5 | 13 | 16 | 0.0123 | 0.6 |
| WTI | Rs-avr | 5 | 10 | 27 | 0.0147 | 0.75 |

The bold estimators are optimal.

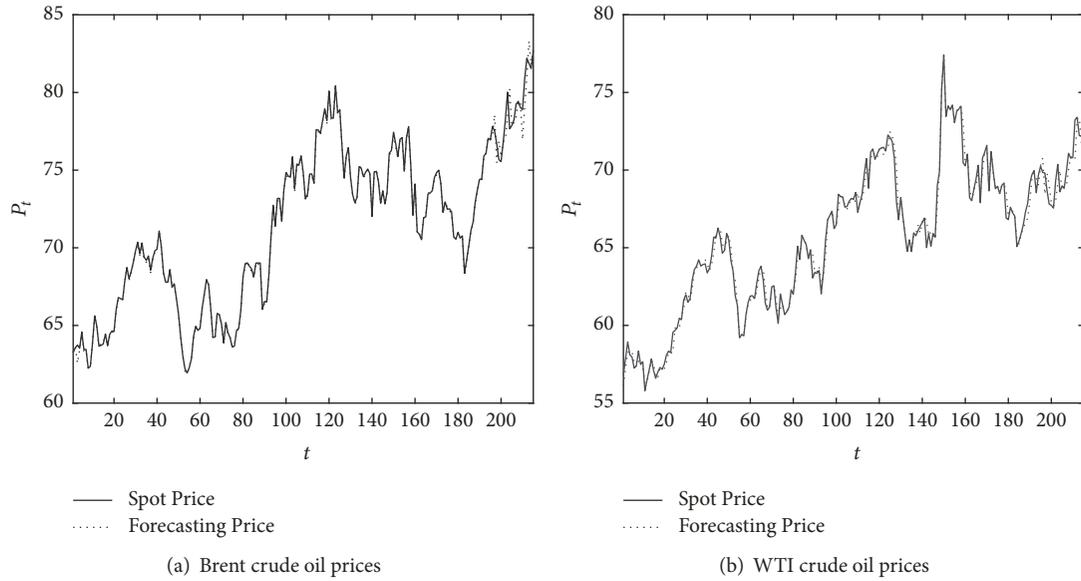


FIGURE 3: The oil spot prices (solid line) and forecasting prices (dotted line).

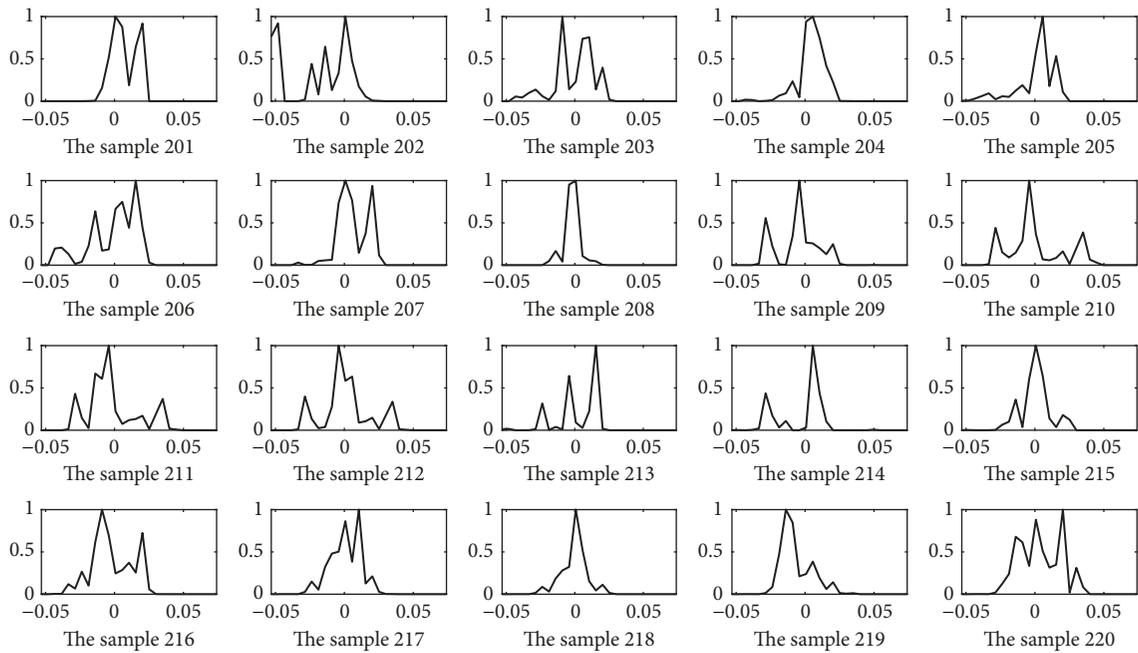


FIGURE 4: The fuzzy possibility distributions of the out-of-sample WTI oil returns.

TABLE 4: The out-of-sample forecasted and actual values.

| The forecasted samples | Brent | | | | | WTI | | | | |
|------------------------|---------|------------------|---------|-------------------|-----------------------|-------|------------------|---------|-------------------|-----------------------|
| | Price | Price forecasted | Return | Return forecasted | Direction consistency | Price | Price forecasted | Return | Return forecasted | Direction consistency |
| 201 | 77.8100 | 77.1249 | 0.0112 | 0.0024 | Y | 69.84 | 70.7789 | -0.0059 | 0.0075 | N |
| 202 | 77.5100 | 78.5606 | -0.0039 | 0.0096 | N | 69.82 | 68.4365 | -0.0003 | -0.0203 | Y |
| 203 | 76.6800 | 75.4830 | -0.0108 | -0.0265 | Y | 68.69 | 69.7991 | -0.0163 | -0.0003 | Y |
| 204 | 75.6700 | 76.8643 | -0.0133 | 0.0024 | N | 67.81 | 69.0274 | -0.0129 | 0.0049 | N |
| 205 | 75.5500 | 75.8518 | -0.0016 | 0.0024 | N | 67.73 | 67.8982 | -0.0012 | 0.0013 | N |
| 206 | 76.7700 | 76.8376 | 0.0160 | 0.0169 | Y | 67.55 | 67.6894 | -0.0027 | -0.0006 | Y |
| 207 | 78.2200 | 76.9545 | 0.0187 | 0.0024 | Y | 69.29 | 67.9905 | 0.0254 | 0.0065 | Y |
| 208 | 80.0200 | 78.4080 | 0.0228 | 0.0024 | Y | 70.37 | 69.1516 | 0.0155 | -0.002 | N |
| 209 | 77.6600 | 80.2123 | -0.0299 | 0.0024 | N | 68.60 | 69.9840 | -0.0255 | -0.0055 | Y |
| 210 | 77.8700 | 77.8466 | 0.0027 | 0.0024 | Y | 68.98 | 68.6549 | 0.0055 | 0.0008 | Y |
| 211 | 78.2200 | 78.0571 | 0.0045 | 0.0024 | Y | 68.86 | 68.8491 | -0.0017 | -0.0019 | Y |
| 212 | 79.2500 | 78.4080 | 0.0131 | 0.0024 | Y | 69.87 | 68.9496 | 0.0146 | 0.0013 | Y |
| 213 | 79.4300 | 79.4404 | 0.0023 | 0.0024 | Y | 71.08 | 70.0309 | 0.0172 | 0.0023 | Y |
| 214 | 79.0300 | 79.3347 | -0.0050 | -0.0012 | Y | 70.77 | 70.8529 | -0.0044 | -0.0032 | Y |
| 215 | 78.9000 | 76.9632 | -0.0016 | -0.0265 | Y | 70.80 | 70.7983 | 0.0004 | 0.0004 | Y |
| 216 | 80.8900 | 79.0896 | 0.0249 | 0.0024 | Y | 73.23 | 70.6656 | 0.0337 | -0.0019 | N |
| 217 | 82.2100 | 81.0844 | 0.0162 | 0.0024 | Y | 73.40 | 73.3033 | 0.0023 | 0.001 | Y |
| 218 | 81.8700 | 83.3107 | -0.0041 | 0.0133 | N | 72.22 | 73.3780 | -0.0162 | -0.0003 | Y |
| 219 | 81.5400 | 81.7718 | -0.0040 | -0.0012 | Y | 72.18 | 71.7377 | -0.0006 | -0.0067 | Y |
| 220 | 82.7200 | 82.9297 | 0.0144 | 0.0169 | Y | 73.16 | 72.4548 | 0.0135 | 0.0038 | Y |

Y represents the consistent forecasting results and N is for inconsistent cases.

TABLE 5: The forecasting result of logistic regression.

| Series | m | D | TN | TP | AUC | Optimal threshold |
|--------------|----------|---------------|---------------|---------------|---------------|-------------------|
| Brent | 4 | 0.7000 | 0.8889 | 0.5455 | 0.6364 | 0.5584 |
| Brent | 5 | 0.7000 | 0.8889 | 0.5455 | 0.6566 | 0.5557 |
| Brent | 6 | 0.6500 | 0.7778 | 0.5455 | 0.6061 | 0.5574 |
| WTI | 4 | 0.6000 | 1.0000 | 0.1111 | 0.4444 | 0.6748 |
| WTI | 5 | 0.7000 | 1.0000 | 0.3333 | 0.5253 | 0.6436 |
| WTI | 6 | 0.7000 | 0.7273 | 0.6667 | 0.6465 | 0.5958 |

The bold estimators are optimal. *D* is the directional accuracy, TN is true negative ratio, and TP is true positive ratio. The ROC (receiver operating characteristic) curve plots the true positive rate as a function of false positive rate for differing classification thresholds. AUC is used to measure the prediction performance that is the area under the ROC curve.

TABLE 6: The forecasting result of the geometric random walk with a dynamic drift.

| Series | m | RMSE | D |
|--------------|----------|---------------|-------------|
| Brent | 4 | 0.016 | 0.45 |
| Brent | 5 | 0.0149 | 0.55 |
| Brent | 6 | 0.0145 | 0.55 |
| WTI | 4 | 0.0167 | 0.5 |
| WTI | 5 | 0.0157 | 0.5 |
| WTI | 6 | 0.0153 | 0.45 |

The bold estimators are optimal.

4. Conclusions

In this paper, a new approach is proposed to enhance the directional forecast accuracy of short-term oil price. It is distinct from previous studies that the crude oil price series is viewed as an incomplete data set with inaccurate fuzzy information and then it is modeled by fuzzy information distribution theory. A new fused approach of genetic-fuzzy information distribution is adopted to forecast the direction of oil price changes. The genetic algorithm is used to optimize the assignment of fuzzy information controlling points, and a coding algorithm of multidimensional information controlling points is applied to make genetic-fuzzy information distribution approach feasible in its time series applications. With the crude oil spot prices of WTI and Brent as sample data, the empirical analysis demonstrates that the novel approach offers high accuracy on the directional forecasting of short-term oil prices, compared with the logistic regression and random walk.

The genetic-fuzzy information distribution can distribute the information of the high level oil price into two neighbour information points; thus it is a new method to construct the oil price distribution structure and then form a fuzzy knowledge-based inference to predict oil prices. It particularly helps to describe the true oil price behaviour hidden in oil price bubbles when oil prices present an unsustainable fast rise. Besides, we demonstrate an effective time series prediction in this paper, but it is worthy of selecting economic or political factors as input variables to enhance the prediction ability and performing the medium and long-term oil price forecasting in future.

Data Availability

The data used to support the findings of this study can be found on the website of US Department of Energy: Energy Information Administration, which is free for academic use.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research described in this paper was substantially supported by the National Natural Science Foundation of China under Grant no. 71871215; the Humanity and Social Science Youth Foundation of Ministry of Education of China under Grant no. 17YJC630012; the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant no. KYCX17_1501.

References

[1] W. Q. Tang, L. B. Wu, and Z. X. Zhang, "Oil price shocks and their short- and long-term effects on the Chinese economy," *Energy Economics*, vol. 32, no. 1, pp. S3–S14, 2009.

- [2] J. D. Hamilton, "Nonlinearities and the macroeconomic effects of oil prices," *Macroeconomic Dynamics*, vol. 15, no. 3, pp. 364–378, 2011.
- [3] S. Bekiros, R. Gupta, and A. Paccagnini, "Oil price forecastability and economic uncertainty," *Economics Letters*, vol. 132, pp. 125–128, 2015.
- [4] S. B. Ramos and H. Veiga, "Risk factors in oil and gas industry returns: International evidence," *Energy Economics*, vol. 33, no. 3, pp. 525–542, 2011.
- [5] A. Merino and R. Albacete, "Econometric modelling for short-term oil price forecasting," *OPEC Energy Review*, vol. 34, no. 1, pp. 25–41, 2010.
- [6] C. Baumeister and L. Kilian, "Forecasting the real price of oil in a changing world: a forecast combination approach," *Journal of Business & Economic Statistics*, vol. 33, no. 3, pp. 338–351, 2015.
- [7] S. Reitz, J. C. Rülke, and G. Stadtmann, "Regressive oil price expectations toward more fundamental values of the oil price," *Jahrbücher für Nationalökonomie und Statistik*, vol. 230, no. 4, pp. 454–466, 2010.
- [8] H. Chiroma, S. Abdulkareem, and T. Herawan, "Evolutionary Neural Network model for West Texas Intermediate crude oil price prediction," *Applied Energy*, vol. 142, pp. 266–273, 2015.
- [9] T. A. Knetsch, "Forecasting the price of crude oil via convenience yield predictions," *Journal of Forecasting*, vol. 26, no. 7, pp. 527–549, 2007.
- [10] X. Zhang, K. K. Lai, and S.-Y. Wang, "A new approach for crude oil price analysis based on empirical mode decomposition," *Energy Economics*, vol. 30, no. 3, pp. 905–918, 2008.
- [11] C. Baumeister and L. Kilian, "Real-time forecasts of the real price of oil," *Journal of Business and Economic Statistics*, vol. 30, no. 2, pp. 326–336, 2012.
- [12] Y. Fan, Q. Liang, and Y.-M. Wei, "A generalized pattern matching approach for multi-step prediction of crude oil price," *Energy Economics*, vol. 30, no. 3, pp. 889–904, 2008.
- [13] L. Yu, S. Wang, and K. K. Lai, "Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm," *Energy Economics*, vol. 30, no. 5, pp. 2623–2635, 2008.
- [14] A. Ghaffari and S. Zare, "A novel algorithm for prediction of crude oil price variation based on soft computing," *Energy Economics*, vol. 31, no. 4, pp. 531–536, 2009.
- [15] R. Jammazi and C. Aloui, "Crude oil price forecasting: experimental evidence from wavelet decomposition and neural network modeling," *Energy Economics*, vol. 34, no. 3, pp. 828–841, 2012.
- [16] Y. Wang, C. Wu, and L. Yang, "Forecasting crude oil market volatility: A Markov switching multifractal volatility approach," *International Journal of Forecasting*, vol. 32, no. 1, pp. 1–9, 2016.
- [17] A. Murat and E. Tokat, "Forecasting oil price movements with crack spread futures," *Energy Economics*, vol. 31, no. 1, pp. 85–90, 2009.
- [18] H. Shin, T. Hou, K. Park, C.-K. Park, and S. Choi, "Prediction of movement direction in crude oil prices based on semi-supervised learning," *Decision Support Systems*, vol. 55, no. 1, pp. 348–358, 2013.
- [19] L. Tang, W. Dai, L. Yu, and S. Wang, "A novel CEEMD-based eelm ensemble learning paradigm for crude oil price forecasting," *International Journal of Information Technology & Decision Making*, vol. 14, no. 1, pp. 141–169, 2015.
- [20] L. Yu, W. Dai, and L. Tang, "A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting," *Engineering Applications of Artificial Intelligence*, vol. 47, pp. 110–121, 2015.

- [21] A. Kavousi-Fard, T. Niknam, and M. Golmaryami, "Short term load forecasting of distribution systems by a new hybrid modified FA-backpropagation method," *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, vol. 26, no. 1, pp. 517–522, 2014.
- [22] F. A. Oliveira, C. N. Nobre, and L. E. Zárate, "Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index - Case study of PETR4, Petrobras, Brazil," *Expert Systems with Applications*, vol. 40, no. 18, pp. 7596–7606, 2013.
- [23] H. Liu and D.-L. Zhang, "Analysis and prediction of hazard risks caused by tropical cyclones in Southern China with fuzzy mathematical and grey models," *Applied Mathematical Modelling: Simulation and Computation for Engineering and Environmental Systems*, vol. 36, no. 2, pp. 626–637, 2012.
- [24] S. Reitz, J.-C. Rülke, and G. Stadtmann, "Nonlinear expectations in speculative markets - Evidence from the ECB survey of professional forecasters," *Journal of Economic Dynamics and Control (JEDC)*, vol. 36, no. 9, pp. 1349–1363, 2012.
- [25] A. Azadeh, M. Moghaddam, M. Khakzad, and V. Ebrahimipour, "A flexible neural network-fuzzy mathematical programming algorithm for improvement of oil price estimation and forecasting," *Computers & Industrial Engineering*, vol. 62, no. 2, pp. 421–430, 2012.
- [26] P. F. Christoffersen and F. X. Diebold, "Financial asset returns, direction-of-change forecasting, and volatility dynamics," *Management Science*, vol. 52, no. 8, pp. 1273–1287, 2006.
- [27] D. Senol and M. Ozturan, "Stock Price Direction Prediction Using Artificial Neural Network Approach: The Case of Turkey," *Journal of Artificial Intelligence Research*, vol. 1, no. 2, pp. 70–77, 2010.
- [28] Y. Kara, M. Acar Boyacioglu, and Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5311–5319, 2011.
- [29] A. Skabar, "Direction-of-change financial time series forecasting using a similarity-based classification model," *Journal of Forecasting*, vol. 32, no. 5, pp. 409–422, 2013.
- [30] Y. Yuan, "Forecasting the movement direction of exchange rate with polynomial smooth support vector machine," *Mathematical and Computer Modelling*, vol. 57, no. 3–4, pp. 932–944, 2013.
- [31] C. Baumeister, P. Guérin, and L. Kilian, "Do high-frequency financial data help forecast oil prices? The MIDAS touch at work," *International Journal of Forecasting*, vol. 31, no. 2, pp. 238–252, 2013.
- [32] C. Engel and J. D. Hamilton, "Long swings in the dollar: are they in the data and do markets know it," *American Economic Review*, vol. 80, pp. 689–713, 1990.
- [33] C. Pierdzioch, J.-C. Rülke, and G. Stadtmann, "Oil price forecasting under asymmetric loss," *Applied Economics*, vol. 45, no. 17, pp. 2371–2379, 2013.
- [34] D. D. Thomakos and T. Wang, "'Optimal' probabilistic and directional predictions of financial returns," *Journal of Empirical Finance*, vol. 17, no. 1, pp. 102–119, 2010.
- [35] J. J. Ahn, D. H. Kim, K. J. Oh, and T. Y. Kim, "Applying option Greeks to directional forecasting of implied volatility in the options market: An intelligent approach," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9315–9322, 2012.
- [36] J. C. Rülke and C. Pierdzioch, "On the directional accuracy of forecasts of emerging market exchange rates," *International Review of Economics & Finance*, vol. 38, pp. 369–376, 2015.
- [37] C. Huang, "Information diffusion techniques and small-sample problem," *International Journal of Information Technology & Decision Making*, vol. 1, no. 2, pp. 229–249, 2002.
- [38] C. Huang and H. Inoue, "Soft risk maps of natural disasters and their applications to decision-making," *Information Sciences*, vol. 177, no. 7, pp. 1583–1592, 2007.
- [39] D.-F. Li, "A fuzzy closeness approach to fuzzy multi-attribute decision making," *Fuzzy Optimization and Decision Making*, vol. 6, no. 3, pp. 237–254, 2007.
- [40] Y. B. Jun and M. S. Kang, "Fuzzy positive implicative ideals of BCK-algebras based on the theory of falling shadows," *Computers & Mathematics with Applications*, vol. 61, no. 1, pp. 62–67, 2011.

