AIMS Mathematics, 8(6): 12778–12792.
DOI: 10.3934/math.2023643
Received: 26 December 2022
Revised: 17 March 2023
Accepted: 20 March 2023
Published: 30 March 2023

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

Research article

Forecasting stock prices based on multivariable fuzzy time series

Zhi Liu*

Department of Basic, Shenyang University of Technology, Liaoyang, Liaoning, China

* **Correspondence:** Email: liuzhi2099@sut.edu.cn; Tel: +8618741982223.

Abstract: With the development of the stock market, the proportion of the stock assets in the asset structure of the residents increases rapidly. Therefore, the research on the prediction of stocks has great theoretical significance and application potential. A key point of researching stock prices is how to pick out the main factors. In this study, principal component analysis (PCA) is applied to find out the main factors which mainly affect the stock price. Then an improved cluster analysis algorithm is proposed to fuzzy the data, and a qualitative analysis method is given to find the most suitable prediction set from the multiple fuzzy sets corresponding to the current fuzzy set. We also extend the inverse fuzzy number formula to a more general form to get the predicted value. Finally, Xishan Coal and Electricity Power (XSCE) and Taiwan Futures Exchange (TAIFEX) time series are predicted, using the proposed multivariate fuzzy time series method. The results show that the prediction error is lower than that of the previous models. The proposed method produces better forecasting performance.

Keywords: quantitative analysis; qualitative analysis; inverse fuzzy number; stock prices; fuzzy time series
Mathematics Subject Classification: 62P20

1. Introduction

Time series analysis, which uses curve fitting methods to describe the time series data, was proposed by Yule in 1927 [27]. Time series analysis models include Auto Regression Model (AR), Moving Average Model (MA), Auto Regression Moving Average Model (ARMA), Autoregressive Integrated Moving Average Model (ARIMA), Autoregressive Conditional Heteroskedasticity Model (ARCH), Generalized Autoregressive Conditional Heteroskedasticity Model (GARCH), etc. After decades of development, time series analysis has been widely applied to data analysis of practical problems. A key point of time series analysis is to establish the appropriate model based on numerical data. However, the data of some practical problems are non-numerical. For example, the temperature of a day can be described as "very cold", "cold", "worm", "hot", and "very hot". These descriptions are all fuzzy

without numerical definitions, and are more suitable to be analyzed by fuzzy theory. The concept of fuzzy sets was proposed by Zadeh in 1965 [28].

Qiang and Brad combined the theory of time series analysis and fuzzy sets together, and proposed the fuzzy time series in 1993 [23]. Since then, the theory of fuzzy time series has attracted more and more attention as the increasing need to solve complex problems. Many related methods have been proposed and applied into practical data analysis [2, 7, 9, 18, 29]. Fuzzy time series have been widely used in various fields of social life [4, 5, 16, 20]. There are many ways to classify the fuzzy time series: According to whether the fuzzy relations are fixed or not, it can be classified into the time-variant fuzzy time series and the time-invariant fuzzy time series; according to the span of fuzzy mapping, it can be divided into the first-order fuzzy time series and the high-order fuzzy time series; according to the number of the current state, it can be distinguished into one-factor fuzzy time series and multi-factors fuzzy time series [13, 25].

The stock market is an important part of the securities market and is one of the most active markets, which is a big concern. Stock prices are influenced by many factors, and it is always difficult to do a scientific calculation and evaluation. The reason is that there are some factors that are hard to quantify. Many scholars have conducted in depth research on it [17, 21]. The current stock price time series forecasting methods can be divided into three categories: the traditional statistical method, the computing intelligence method, and the combined forecasting method. This paper uses the combination of qualitative and quantitative methods. We apply the theory of fuzzy time series to analyze quantitative factors as well as consider the qualitative factor to forecast the stock price. First, we use the method of PCA to determine the main factors that affect stock prices. Then we use the automatic clustering analysis to divide the universe of discourse, and establish the n-factors first-order Markov transition matrix. Finally, we consider the qualitative factors correlating to stock price, and use the inverse fuzzy number to forecast the stock prices.

There are three contributions of this paper:

(1) The theory of PCA is used to abstract the main factors affecting stock price, which reduces the research difficulty and improves the accuracy.

(2) The cluster analysis algorithm is improved to divide the discussion domain and the method of inverse fuzzy number is extended. And the convergence theorem of inverse fuzzy numbers is proved. The effectiveness of this method is proved theoretically

(3) The qualitative factors are considered in the fuzzy time series to remove the incorrect prediction. The prediction accuracy is therefore improved.

The rest of this paper is organized as follows. In Section 2, some basic concepts of the fuzzy time series used in the paper are introduced. In Section 3, a hybrid fuzzy time series model is introduced in detail. In Section 4, many implementations of the proposed method and some other methods are performed to make comparisons. In Section 5, some conclusions and remarks are discussed.

2. Basic concepts

In this section, some basic concepts of the fuzzy time series are introduced.

Definition 2.1. [22] Let $\mathbf{Y}(t)(t = 1, 2, \cdots)$ be the universe of discourse on which fuzzy sets $f_i(t)$, $(i = 1, 2, \cdots)$ are defined, and $\mathbf{F}(t)$ is the collection of $f_i(t)$, then $\mathbf{F}(t)$ is called a fuzzy time series on $\mathbf{Y}(t)$.

Definition 2.2. [23] If $f_j(t)$ is caused only by $f_i(t-1)$, denoted as $f_i(t-1) \to f_j(t)$, then there exists a fuzzy relation $R_{ij}(t-1,t)$, such that $f_j(t) = f_i(t-1) \circ R_{ij}(t-1,t)$, where '$\circ$' is the composition operator. This relation is called a first-order model of $f(t)$.

Definition 2.3. [24] If $f_j(t)$ is caused by $f_{i_1}(t-1), f_{i_2}(t-2), \cdots$, and $f_{i_n}(t-n)$ simultaneously, denoted as

$$f_{i_1}(t-1) \cap f_{i_2}(t-2) \cap \cdots \cap f_{i_n}(t-n) \to f_j(t), \tag{2.1}$$

where '$\cap$' is the intersection operator, then the fuzzy relational equation is:

$$\mathbf{F}(t) = (\mathbf{F}(t-1) \times \mathbf{F}(t-2) \times \cdots \times \mathbf{F}(t-n)) \circ \mathbf{R}(t-n,t). \tag{2.2}$$

Definition 2.4. The forecasting formula uses the generalized inverse fuzzy number

$$\alpha_i = \frac{\mu_1 + \cdots + \mu_{i-1} + \mu_i + \mu_{i+1} + \cdots + \mu_n}{A_i} \tag{2.3}$$

where A_i is the membership function of a fuzzy set, μ_i is the grade of membership of U_i and α_i is the predicted value of i.

The formula (2.3) can deal with not only the triangular membership functions, but also any forms of the membership functions. It is a more generalized form of inverse fuzzy number formulain [8, 26].

Theorem 2.1. *Let $U = \{U_1, U_2, \cdots, U_n\}$ be the universe of discourse, μ_i represents the grade of membership of U_i in fuzzy set A_i. And α_i yields the predicted value of i. In the Definition 2.4:*

1) When $\mu_{i-1} \to 0$, and $\mu_{i+1} \to 0$, then $\alpha_i \to U_i$.

2) For any real number $\zeta > 0$, there exist $\xi > 0, \eta > 0$, such that the average forecast error rate $AER = \frac{|\alpha_i - U_i|}{U_i} < \zeta$, if $\mu_{i-1} < \xi$, $\mu_{i+1} < \eta$.

Proof. 1) When $\mu_{i-1} \to 0$, and $\mu_{i+1} \to 0$, it is obviously.

2) Following from 1). For any $\zeta > 0$, it exists $\eta_1 > 0$, such that $|\alpha_1 - U_1|/U_1 < \zeta$, if $\mu_2 < \eta_1$; it exists $\xi_2 > 0, \eta_2 > 0$, such that $|\alpha_2 - U_2|/U_2 < \zeta$, if $\mu_1 < \xi_2$ and $\mu_3 < \eta_2$; $\cdots$; it exists $\xi_i > 0, \eta_i > 0$, such that $|\alpha_i - U_i|/U_i < \zeta$, if $\mu_{i-1} < \xi_i$ and $\mu_{i+1} < \eta_i$; $\cdots$; it exists $\xi_n > 0$, such that $|\alpha_n - U_n|/U_n < \zeta$, if $\mu_{n-1} < \xi_n$. Put $\xi = min\{\xi_1, \xi_2, \cdots, \xi_n\} > 0$, $\eta = min\{\eta_1, \eta_2, \cdots, \eta_n\} > 0$, if $\mu_{i-1} < \xi, \mu_{i+1} < \eta$, then the average forecast error rate

$$AER = (\frac{|\alpha_1 - U_1|}{U_1} + \frac{|\alpha_2 - U_2|}{U_2} + \cdots + \frac{|\alpha_n - U_n|}{U_n})/n < \frac{1}{n}(\zeta + \zeta + \cdots + \zeta) = \zeta.$$

□

Apparently, U_i is a fuzzy set, while the inverse of fuzzy number α_i is a set of real numbers. The set of real numbers.

3. Multivariate fuzzy time series method

According to the feature of stock price, a multivariate fuzzy time series method is propose in this paper, which is explained in this section. First, the PCA algorithm is applied to select the factors which mainly affect the stock price. Second, an improved cluster analysis algorithm is used to define and

divide the universe of discourse. The Markov transition matrix is set up, and the prediction fuzzy sets according to the transition matrix is obtained. Finally, the qualitative factors are considered to increase the prediction accuracy. Inverse fuzzy number formula is used to inversely fuzzify the data and obtain the prediction results.

Step 1: Select the main factors.

The PCA is a multivariate statistical technique based on the statistical characteristics of multidimensional orthogonal linear transformation. The PCA is usually used to feature extraction and data dimension reduction. The concept of this technique was firstly proposed by Pearson in 1901 [19], and was developed by Hotelling [10], Jackson [11] and other researchers. Later, probability theory was applied to describe the PCA algorithm, which further developed the PCA method. Nowadays, a lot of related research has been carried out, and this method has been widely used in chemistry, pattern recognition, image processing and other fields [12]. The PCA is also named Karhunen-Loeve transform [14], Hotelling transform [10], subspace approach, eigen-structure approach, etc.

In order to reduce the number of variables, the PCA algorithm is applied. Do the following to process the data:

(1) Calculate the mean of the sample data,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.1}$$

(2) Centralize the sample data,

$$\tilde{\mathbf{X}} = \mathbf{X} - \bar{x}E. \tag{3.2}$$

(3) Construct the matrix covariance of the sample data,

$$\mathbf{V} = \frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T. \tag{3.3}$$

(4) Calculate the eigenvectors ω_i and the eigenvalues λ_i of the matrix covariance, and arrange the eigenvalues λ_i as the descending order.

(5) By trying different thresholds, it is finally found that when the cumulative contribution rate is between 85% and 95%, the selected variables are both representative and can represent the original variables to a greater extent extract. The top m eigenvalues $\Lambda = diag[\lambda_1, \lambda_2, \cdots, \lambda_m]$ and the corresponding eigenvectors $\mathbf{W}_m = [\omega_1, \omega_2, \cdots, \omega_m]$ as the base of subspaces. Then we can obtain the m main factors.

Step 2: Preprocess data.

The dimension of fuzzy time series is determined by the number of principal components. Suppose two variables x and y are selected as pivot entries, and each variable has $m+1$ and $n+1$ observed samples, respectively. Then process each set of data as follows:

(1) Calculate the gradient of the original data by the following formulas.

$$a_i = \frac{x_{i+1} - x_i}{x_i} \times 100\%, \ (i = 1, 2, \cdots, m), \tag{3.4}$$

$$b_j = \frac{y_{j+1} - y_j}{y_j} \times 100\%, \ (j = 1, 2, \cdots, n). \tag{3.5}$$

(2) Arrange the data in an ascending sequence excluding duplicate data. The gradient sequences are written as follows:

$$\mathbf{a} = \{a_1, a_2, \cdots, a_m\}, \ \ \mathbf{b} = \{b_1, b_2, \cdots, b_n\}.$$

Step 3: Define and divide the universe of discourse.

The preprocessed sequences are divided into the universes of discourse, using the improved cluster analysis algorithm.

Let X_i be the real data points, the jth cluster center C_j can be calculated according to Eq (3.6).

$$C_j = \frac{\sum_{i=1}^{n} 4\mu_{ij}^4 X_i}{\sum_{i=1}^{n} 4\mu_{ij}^4}. \tag{3.6}$$

Where $\mu_{ij} \in [0,1]$ is the degree of membership of X_i in jth cluster, which is calculated as follows:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{m} \sqrt{2}\left(\frac{X_i - C_j}{X_i - C_k}\right)}, \quad \text{s.t.} \begin{cases} \mu_{ij} \in [0,1] \\ \sum_{j=1}^{c} 2{\mu_{ij}}^2 = 1 \end{cases}.$$

Where m represents the number of clusters, $i = 1, 2, \cdots, n$; $j = 1, 2 \cdots, m$; $k = 1, 2 \cdots, m$.

The objective function J is given as:

$$J = \sum_{i}^{n} \sum_{j}^{m} \mu_{ij} d(X_i, C_j). \tag{3.7}$$

Where $d(X_i, C_j)$ is the Euclidean distance, and $d(X_i, C_j) = \sqrt{(X_i - C_j)^2}$.

Then each cluster center is taken as the midpoint of the universe of discourse. The minimum value of half of the distance between it and the cluster center on both sides is taken as the radius to get the universe of discourse. Then the empty interval in the middle is added into a new domain to get the final result.

Step 4: Set up Markov transition matrix.

If $(\mu_1(t), \mu_2(t), \cdots, \mu_n(t)), (\mu_1(t+1), \mu_2(t+1), \cdots, \ \mu_n(t+1))$ represent the memberships of the observed value of the given fuzzy sets $\mathbf{F}(t), \mathbf{F}(t+1)$, respectively, and $\mu_i(t), \mu_j(t+1)$ correspond to the given fuzzy sets $A_i(t), A_j(t+1)$, then we can obtain the logical relation matrix $\mathbf{R} = \left[\mu_{ij}\right]$ (Markov transition matrix), which sets up the relation of two principal components and the prediction variable.

$$\begin{array}{c} \\ A_1(t-1) \\ A_2(t-1) \\ \vdots \\ A_n(t-1) \end{array} \begin{array}{c} \begin{array}{cccc} A_1(t) & A_2(t) & \cdots & A_n(t) \end{array} \\ \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n1} & \mu_{n2} & \cdots & \mu_{nn} \end{bmatrix} \end{array}.$$

Step 5: Optimize predictive fuzzy sets.

If there are more than one $A_j(t+1)$ corresponding to $A_i(t)$, the qualitative factors are considered. For example, whether there is a negative or a positive policy, the market environment changes, or the

country's economic policy changes (such as monetary policy, fiscal, etc.) to determine the prediction set. Assuming the current set is $A_i(t)$, if there are positive factors, we will take the fuzzy sets next time with subscript greater than i as the prediction set; if there are negative factors, we will take the fuzzy sets with subscript less than i as the prediction set; if the policy is stable, we do not change the fuzzy set of next time, and take it as the prediction set.

Step 6: Calculate the prediction values.

According to Step 5, the predictive fuzzy set is calculated by applying the generalized inverse fuzzy number formula (2.3). Then, the midpoints coordinates of it and its adjacent fuzzy sets are brought into the formula (3.8) to obtain the final prediction value.

$$p_i = x_{i-1}(1 + \alpha_i\%). \tag{3.8}$$

Where p_i represents the predicted value at the i-th time, α_i represents the predicted change rate at the i-th time, and x_{i-1} represents the time series value at the i-1st time.

The flow diagram of the algorithm proposed in this paper is shown on Figure 1.

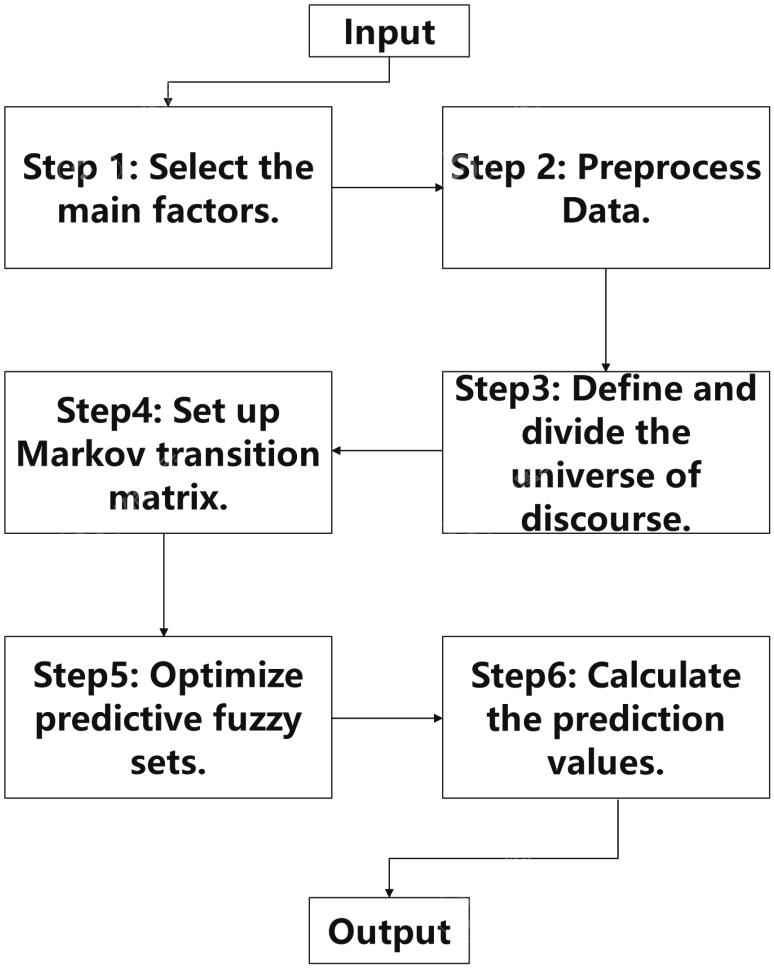

Figure 1. The flow diagram of the algorithm.

4. Simulation prediction examples

In this section, the stock prices of XSCE and TAIFEX time series are predicted to illustrate the effectiveness of the model presented in the paper.

4.1. Predicting stock prices of XSCE

Draw the opening price of XSCE from June 1st, 2015 to July 31st, 2015 (`https://finance.sina.com.cn/realstock/company/sz000983/nc.shtml`). It can be observed in Figure 2 that the stock price fluctuates wildly during the chosen period of time. It is hard to be analyzed by the classical time series methods or the other fuzzy time series methods. The multivariable fuzzy time series method proposed in this paper is used to analyze the stock price.

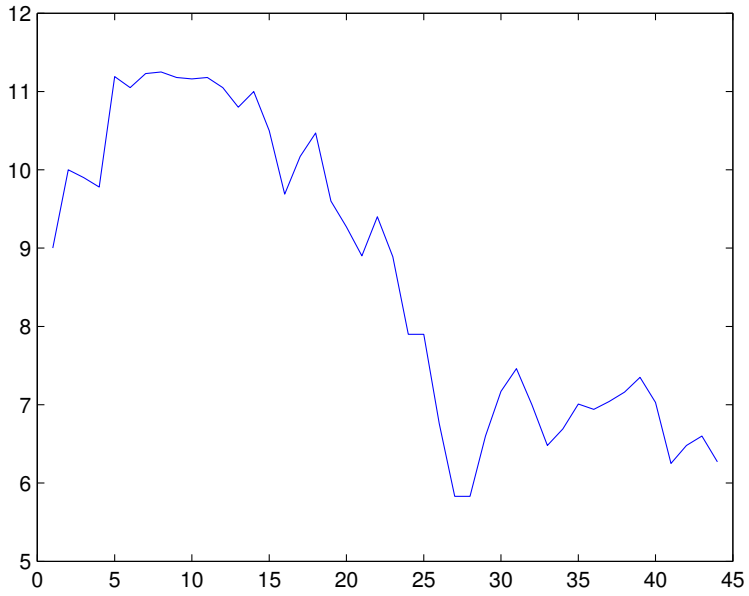

Figure 2. The opening price of stock price of Shanxi Coal and Electricity Power in China, from Jun. 1st, 2015 to Jul. 31st, 2015.

Step 1: Select the main factors.

The opening price (OP), the highest price (HP), the lowest price (LP), the closing price (CP), the range of daily fluctuations (RF), the trading volume (TV) and the turnover number (TN) are analyzed using the method of PCA. The correlation matrix is shown in Table 1. From Table 1, you can see that the OP and the HP are highly related to the LP and the CP, respectively. Clearly, there is an overlap of information between them. The PCA method uses the first m principal components whose eigenvalues are greater than 1. To some extent, the eigenvalues represent the influence of the principal components. If an eigenvalue is less than 1, it means that the explanation ability of the corresponding principal component is weaker than the average explanation ability of the original variables, so we only consider the principle components with eigenvalues greater than 1. According to the results in

Table 2, because the cumulative contribution of the first two factors reached 85.194%, two principal components are extracted. In other words, the OP and the HP are selected as the research variables, and abandon the rest variables. And then standardize their contributions, and obtain the weights of the considered variables: $0.7301, 0.2699$ of the OP and the HP, respectively.

Step 2: Preprocess data.

The two time series values of the OP and the HP are put into Eqs (3.4) and (3.5) respectively to calculate the gradient. After removing the same data, they are arranged into two ascending sequences of numbers. The results are shown as follows:

$$\begin{aligned}\mathbf{a} = \{&-14.4304, -13.7574, -11.1361, -11.0953,\\ &-8.3095, -7.7143, -7.4286, -6.1662,\\ &-5.4255, -5.0000, -4.5455, -4.3537,\\ &-3.9914, -3.4375, -2.2624, -1.2511,\\ &-1.2121, -1.1628, -1.0000, -0.9986,\\ &-0.6222, -0.1789, 0, 0.1781,\\ &0.1792, 1.4409, 1.6290, 1.7045,\\ &1.8519, 2.6536, 2.9499, 3.2407,\\ &3.6800, 4.0446, 4.7833, 4.9536,\\ &5.6180, 8.6364, 11.1111, 13.2075,\\ &14.4172\},\end{aligned}$$

$$\begin{aligned}\mathbf{b} = \{&-13.0380, -11.7904, -10.3563, -9.6639,\\ &-8.8803, -7.1521, -6.5341, -6.2630,\\ &-5.2685, -5.2259, -4.9533, -4.1543,\\ &-4.0034, -2.5424, -2.0851, -1.8771,\\ &-1.8634, -1.4047, -1.2431, -0.6585,\\ &-0.6098, -0.1753, 0, 0.1289,\\ &0.4193, 0.5343, 0.7474, 1.3986,\\ &1.4006, 1.7931, 1.8634, 2.0321,\\ &2.5278, 3.4783, 3.7209, 4.2005,\\ &4.2609, 4.5231, 5.7756, 6.4103,\\ &8.6191, 9.9844, 10.0709\}.\end{aligned}$$

Step 3: Define and divide the universe of discourse.

The improved clustering analysis algorithm proposed in Chapter 3 is used to cluster the sequences **a** and **b**, respectively. Each cluster center is taken as the midpoints of discussion domain, and half of the distance between adjacent cluster centers is taken as the clustering points to get the universe of discourse. So **a** is divided into universes U_i and **b** into universes V_i, respectively.

Fuzzy sets $A_1, A_2, \cdots, A_n$ can be defined on U by general triangular membership functions expressed

as below:

$$\begin{aligned} A_1 = &\ \frac{1}{U_1} + \frac{0.5}{U_2} + \frac{0}{U_3} + \cdots + \frac{0}{U_n}, \\ A_i = &\ \frac{0}{U_1} + \frac{0}{U_2} + \cdots + \frac{0}{U_{i-2}} + \frac{0.5}{U_{i-1}} \\ &+ \frac{1}{U_i} + \frac{0.5}{U_{i+1}} + \frac{0}{U_{i+2}} + \cdots + \frac{0}{U_n}, \\ &(i = 2, 3, \cdots, n-1), \\ A_n = &\ \frac{0}{U_1} + \frac{0}{U_2} + \cdots + \frac{0}{U_{n-2}} + \frac{0.5}{U_{n-1}} + \frac{1}{U_n}, \end{aligned} \tag{4.1}$$

where U_i is the sub-universe of discourse and the corresponding numerator represents the membership of U_i to A_i. By the same way, the fuzzy sets B_i which are defined on V_i can be get.

Step 4: Set up the relation matrix.

According to the membership maximum principle, the relation matrix between two principal components is set up. Two sets of fuzzy relationship chains are obtained ('#' means that there is no corresponding fuzzy set).

Table 1. Correlation matrix.

	OP	HP	LP	CP	RF	TV	TN
OP	1.000	0.994	0.978	0.972	-0.022	0.390	0.282
HP	0.994	1.000	0.972	0.987	0.063	0.442	0.321
LP	0.978	0.972	1.000	0.960	-0.005	0.365	0.254
CP	0.972	0.987	0.960	1.000	0.183	0.443	0.309
RF	-0.022	0.063	-0.005	0.183	1.000	0.320	0.185
TV	0.390	0.442	0.365	0.443	0.320	1.000	0.862
TN	0.282	0.321	0.254	0.309	0.185	0.862	1.000

Table 2. Total variance explained.

Component	Initial eigenvalues		
	Total	% of Variance	Cumulative %
1	4.354	62.196	62.196
2	1.610	22.998	85.194
3	0.880	12.566	97.759
4	0.117	1.669	99.428
5	0.031	0.441	99.869
6	0.007	0.105	99.974
7	0.002	0.026	100.000
Component	Extraction sums of squared loadings		
	Total	% of Variance	Cumulative %
1	4.354	62.196	62.196
2	1.610	22.998	85.194

Step 5: Optimize predictive fuzzy sets.

According to the above fuzzy relation chain, the predictive fuzzy set at the next time is obtained. When there are multiple fuzzy sets of next time corresponding to the current fuzzy set, the qualitative factors can be considered. If the news or policy is favorable, the fuzzy set with subscript greater than that of the current fuzzy set i is selected as the prediction set, according to expert evaluation or experience; otherwise, the fuzzy set with subscript less than i is selected.

Step 6: Make predictions.

Finally, using Eqs (2.3) and (3.8), the prediction fuzzy set is inversely fuzzy and the predicted value is obtained.

For example, the rate change of OP on June 2nd, 2015 is 11.1111%, and the corresponding fuzzy set is A_{39}. The rate change of HP is 2.5278%, and the corresponding fuzzy set is B_{20}. The next fuzzy sets for them are both A_{19}. According to the result of dividing the universe in Step 3, the centers of U_{18}, U_{19} and U_{20} are $-1.1875, -1.0807$ and -0.8104. The prediction rate change of the opening price (PROP) is

$$\frac{0.5+1+0.5}{\frac{0.5}{-1.1875}+\frac{1}{-1.0807}+\frac{0.5}{-0.8104}} = -1.0187.$$

The prediction of the opening price (POP) on the next day is obtained as 9.8981. POP from June 1st, 2015 to July 31st, 2015 are shown in Table 3.

Table 3. Date of stock price from Jun. 1st, 2015 to Jul. 31st, 2015 snd the prediction results.

Date	OP(x)	HP	PROP	POP	AER %
6.1	9	9.89			
6.2	10	10.14			
6.3	9.9	10.14	-1.0187	9.8981	0.0192
6.4	9.78	10.79	-1.1691	9.7843	0.044
6.5	11.19	11.72	14.2098	11.1697	0.1814
6.8	11.05	11.5	-1.2902	11.0456	0.0398
6.9	11.23	11.99	1.6364	11.2308	0.0071
⋮	⋮	⋮	⋮	⋮	⋮
7.23	7.16	7.38	1.9436	7.1768	0.2346
7.24	7.35	7.69	2.9014	7.3677	0.2408
7.27	7.03	7.14	-4.1098	7.0479	0.2546
7.28	6.25	6.45	-9.4523	6.3655	1.848
7.29	6.48	6.69	3.8708	6.4919	0.1836
7.30	6.6	6.74	1.9436	6.6059	0.0894
7.31	6.27	6.46	-4.7869	6.2841	0.2249

The proposed method also can be used to estimate the unknown data. For example, on Mar. 26th, 2016, the OP is 8.160, and the HP is 8.310. On next day, the OP is 8.300, and the HP is 8.440. So the rate change of OP on Mar. 27th, 2016 is 1.72%, and the corresponding fuzzy set is A_{27}. The next fuzzy sets are A_{11},A_{10} and A_{29}. There are three fuzzy sets corresponding to the previous one, so the

qualitative factors are considered. Because the market was good without any negative information at that time, the fuzzy set with label greater than that of the current fuzzy set is selected as the prediction sets, namely, A_{29}. Then the trend-weighted method is applied to calculate the PROP, which is 1.935%, and the POP is $8.300 \times (1 + 1.935\%) = 8.461$. On Mar. 30th, 2016, the actual value of OP is 8.330. The absolute error (AER) is computed by Eq (4.2):

$$AER = \left| \frac{x_i - p_i}{x_i} \right| \cdot 100\%. \tag{4.2}$$

After calculation, the AER is 1.568%, far less than the that of literatures [23] 4.380% and [4] 3.117%.

Figure 3 shows the real data of OP and the predicted values using different methods. In the figure, it can be seen more intuitively that compared with the predicted values of particle swarm optimization (PSO) and BP neural network, the predicted values of the method proposed in this paper are closer to the real values. Equation (4.2) is used to calculate the AER. The results are shown in Table 3. From Table 3, we can see that the maximum of AER is 1.848%, and the minimum of AER is 0.007%. A more intuitive results are shown on Figure 4. The average absolute error rate (AAER) is calculated by Eq (4.3):

$$AAER = \frac{\sum_{i=1}^{n} \frac{|x_i - p_i|}{x_i}}{n} \cdot 100\%. \tag{4.3}$$

The AAER of the proposed method is 0.2284%, much smaller than that in literature [6], which is 1.5294%. By comparison, we can see that the method proposed in this paper is more effective.

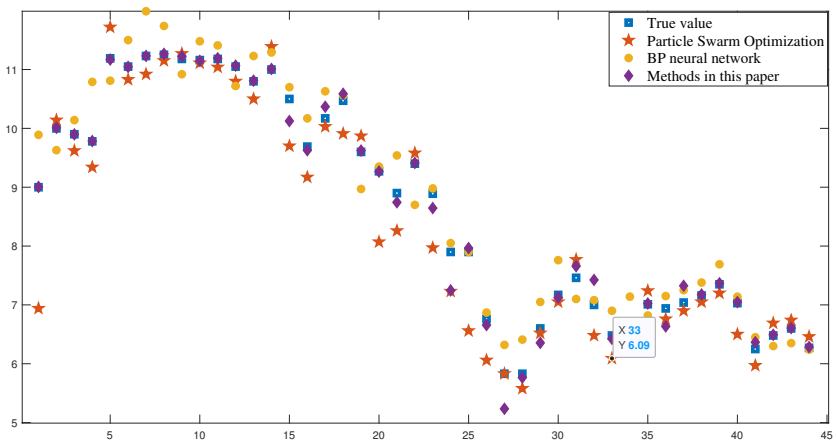

Figure 3. Opening price and the prediction using multivariate fuzzy time series method.

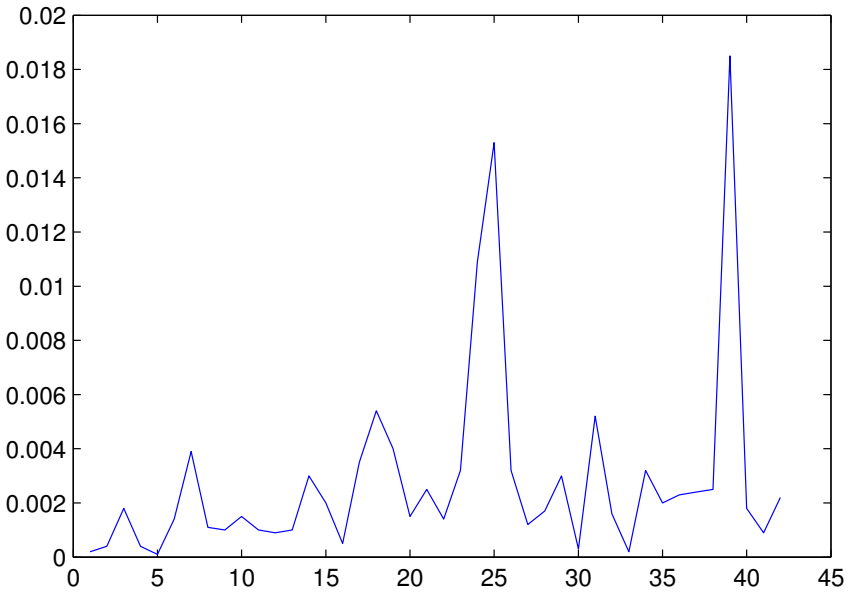

Figure 4. Absolute error rate.

4.2. Predicting stock price of TAIFEX

In order to further evaluate the performance of the proposed method. The method is applied to predict TAIFEX time series, whose observations are between Aug. 3rd, 1998 and Sep. 30th, 1998. The first 38 observations are used for training, and the last 9 data are used for testing.

Firstly, the PCA method is used to determine the main factors affecting the stock prices. The cumulative contribution rates of the CP and the TV are over 85%, so they are selected as research variables. Next, the two sets of observations are taken into Eqs (3.4) and (3.5), respectively, to calculate the gradient, and are arranged into two ascending sequences of numbers. Then, using the improved clustering analysis algorithm, they are divided into two groups of universes. According to the membership maximum principle, two sets of fuzzy relationship chains are obtained. The prediction fuzzy sets can be obtained from it. When there are multiple sets corresponding to the current fuzzy set, the optimal predictive fuzzy set is selected according to expert evaluation or experience, considering qualitative factors. Finally, using Eqs (2.3) and (3.8), the predicted values are obtained.

The test data of TAIFEX are also predicted using methods proposed by Lee et al. [15], Aladag et al. [1] and Bas et al. [3]. The *AAER* applying Eq (4.3) and the root mean square error (RMSE) applying Eq (4.4) are obtained.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (p_t - x_t)^2}{n}}. \tag{4.4}$$

A performance comparison is shown in Table 4.

It can be observed that the *AAER* and the RMSE of the proposed method is far smaller than that of Lee's method, Aladag's method and Bas's method. From Table 4, the experimental results show that the hybrid fuzzy time series model proposed in this paper obtains the minimum *AAER* and RMSE.

In other words, this model can reduce the prediction error more effectively and has better prediction performance than other stock price prediction methods.

Table 4. Comparison of the results of test set for TAIFEX time series.

Test data	Lee et al. [15]	Aladag et al. [1]	Bas et al. [3]	The proposed method
7039	6977	6850	6900	6993
6861	6875	6850	6900	6864
6926	7126	6850	6900	6935
6852	6863	6850	6900	6847
6890	6944	6850	6860	6925
6871	6832	6850	6900	6896
6840	6843	6850	6860	6838
6806	6859	6850	6860	6837
6787	6826	6750	6860	6809
RMSE	76.69	72.32	58.62	24.79
AAER	0.0076	0.0069	0.0074	0.0029

5. Conclusions

In this paper, the main factors among the multiple quantitative ones that affect the stock prices are selected by the PCA algorithm. This method reduces the dimension of the problem to be studied. Then, improved cluster analysis algorithm is used to divide the universe of discourse, which makes the division more reasonable. The fuzzy relations are set up according to Markov transition matrix, and consider the qualitative factors to remove the redundant prediction sets. This is a better way to start from the data and make more efficient use of the data's own attributes. Finally, the predicted value is calculated by using the inverse fuzzy number. The model is applied to different stock data. By comparison, it can be seen that the method effectively improves the prediction performance and is more suitable for dealing with complex nonlinear data. The application of multivariable fuzzy time series to stock price prediction effectively improves the prediction accuracy and is a future research direction.

Of course, there are a lot of methods to reduce the data dimension, the author will continue the study to find better dimensionality reduction methods and try to apply the research into other fields.

Acknowledgments

We would like to thank you for Scientific Research Fund Project of Education Department of Liaoning Province, No. LJKZ0164.

Conflict of interest

The author declares no conflict of interest.

References

1. C. H. Aladag, Using multiplicative neuron model to establish fuzzy logic relationships, *Expert Syst. Appl.*, **40** (2913), 850–853. https://doi.org/10.1016/j.eswa.2012.05.039
2. S. N. Arslan, O. C. Yolcu, A hybrid sigma-pi neural network for combined intuitionistic fuzzy time series prediction model, *Neural Comput. Appl.*, **34** (2022), 12895–12917. https://doi.org/10.1007/s00521-022-07138-z
3. E. Bas, C. Grosan, E. Egrioglu, U. Yolcu, High order fuzzy time series method based on Pi-Sigma neural network, *Eng. Appl. Arti. Intel.*, **72** (2018), 350–356. https://doi.org/10.1016/j.engappai.2018.04.017
4. S. M. Chen, Forecasting enrollments based on fuzzy time series, *Fuzzy Sets Syst.*, **81** (1996), 311–319. https://doi.org/10.1016/0165-0114(95)00220-0
5. S. M. Chen, J. R. Hwang, Temperature prediction using fuzzy time series, *IEEE Trans. Syst. Man Cybern. B Cybern.*, **30** (2000), 263–275. https://doi.org/10.1109/3477.836375
6. S. M. Chen, Forecasting enrollments based on high-order fuzzy time series, *Cybern. Syst.*, **33** (2002), 1–16. https://doi.org/10.1080/019697202753306479
7. S. M. Chen, N. Y. Wang, J. S. Pan, Forecasting enrollments using automatic clustering techniques and fuzzy logical relationships, *Expert Syst. Appl.*, **36** (2009), 11070–11076. https://doi.org/10.1016/j.eswa.2009.02.085
8. M. Y. Chen, B. T. Chen, A hybrid fuzzy time series model based on granular computing for stock price forecasting, *Inf. Sci.*, **294** (2015), 227–241. https://doi.org/10.1016/j.ins.2014.09.038
9. J. Dombi, T. Jónás, Z. E. Tóth, Fuzzy time series models using pliant-and asymptotically pliant arithmetic-based inference, *Neural Process. Lett.*, **52** (2020), 21–55. https://doi.org/10.1007/s11063-018-9927-0
10. H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.*, **24** (1933), 498–520. https://doi.org/10.1037/h0070888
11. J. E. Jackson, G. S. Mudholkar, Control procedures for residuals associated with principal component analysis, *Technometrics*, **21** (1979), 341–349. https://doi.org/10.1080/00401706.1979.10489779
12. D. E. Johnson, *Applied multivariate methods for data analysis*, Duxbury Resource Center, 1998, 93–111.
13. I. Jolliffe, *Principal component analysis*, New York: Springer, 2002. https://doi.org/10.1007/b98835
14. K. Karhunen, *On linear methods in probability theory*, RAND Corporation, 1960, 16–28.
15. L. W. Lee, L. H. Wang, S. M. Chen, Temperature prediction and TAIFEX forecasting based on high-order fuzzy logical relationships and genetic simulated annealing techniques, *Expert Syst. Appl.*, **34** (2008), 328–336. http://dx.doi.org/10.1016/j.eswa.2006.09.007
16. W. J. Lee, J. Hong, A hybrid dynamic and fuzzy time series model for mid-term power load forecasting, International, *Int. J. Elec. Power Energy Syst.*, **64** (2015), 1057–1062. http://dx.doi.org/10.1016/j.ijepes.2014.08.006

17. Z. Liu, T. Zhang, A second-order fuzzy time series model for stock price analysis, *J. Appl. Stat.*, **46** (2019), 2514–2526. http://dx.doi.org/10.1080/02664763.2019.1601163

18. R. M. Pattanayak, S. Panigrahi, H. S. Behera, High-order fuzzy time series forecasting by using membership values along with data and support vector machine, *Arab. J. Sci. Eng.*, **45** (2020), 10311–10325. http://dx.doi.org/10.1007/s13369-020-04721-1

19. K. Pearson, On lines and planes of closest fit to systems of points in space, *Philos. Mag.*, **2** (1901), 559–572.

20. N. H. A. Rahman, M. H. Lee, Suhartono, M. T. Latif, Artificial neural networks and fuzzy time series forecasting: an application to air quality, *Qual. Quant.*, **49** (2015), 2633–2647. http://dx.doi.org/ 10.1007/s11135-014-0132-6

21. P. Saxena, K. Sharma, S. Easo, Forecasting enrollments based on fuzzy time series with higher forecast accuracy rate, *Int. J. Comput. Technol. Appl.*, **3** (2012), 957–961.

22. Q. Song, B. S. Chissom, Fuzzy time series and its models, *Fuzzy Sets Syst.*, **54** (1993), 269–277. https://doi.org/10.1016/0165-0114(93)90372-O

23. Q. Song, B. S. Chissom, Forecasting enrollments with fuzzy time series–part I, *Fuzzy Sets Syst.*, **54** (1993), 1–9. https://doi.org/10.1016/0165-0114(93)90355-L

24. Q. Song, B. S. Chissom, Forecasting enrollments with fuzzy time series–part II, *Fuzzy Sets Syst.*, **62** (1994), 1–8. https://doi.org/10.1016/0165-0114(94)90067-1

25. B. Q. Sun, H. F. Guo, H. R. Karimi, Y. J. Ge, S. Xiong, Prediction of stock index futures prices based on fuzzy sets and multivariate fuzzy time series, *Neurocomputing*, **151** (2015), 1528–1536. http://dx.doi.org/10.1016/j.neucom.2014.09.018

26. N. Y. Wang, S. M. Chen, Temperature prediction and TAIFEX forecasting based on automatic clustering techniques and two-factors high-order fuzzy time series, *Expert Syst. Appl.*, **36** (2009), 2143–2154. http://dx.doi.org/10.1016/j.eswa.2007.12.013

27. G. U. Yule, On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers, *Philosophical Transactions of the Royal Society of London, Series A*, London, **226** (1927), 267–298. https://doi.org/10.1098/rsta.1927.0007

28. L. A. Zadeh, Fuzzy sets, *Inf. Control*, **8** (1965), 338–353. http://dx.doi.org/ 10.1016/S0019-9958(65)90241-X

29. R. Zarei, M. Gh. Akbari, J. Chachi, Modeling autoregressive fuzzy time series data based on semi-parametri cmethods, *Soft. Comput.*, **24** (2020), 7295–7304. http://dx.doi.org/10.1007/s00500-019-04349-w

AIMS Press