ISSN: 1813-4890

Predict the closing price of Zhongyuan Expressway based on ARIMA model

Zhiqing Zhang [a], Xuanjing Ti [b]

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233000, China.

[a]2894665925@qq.com, [b]1398695962@qq.com

Abstract

With the rapid development of social economy and the continuous advancement of financial globalization, stock price, as an important indicator of the stock market, its analysis and prediction has always been one of the hot spots in finance. Based on the closing price of Zhongyuan Expressway from September 1, 2022 to September 1, 2023, this article uses R language to conduct a time series prediction analysis of the closing price of Zhongyuan Expressway in the next three days. First, the original time series is tested for stationarity and white noise, and it is found that the first difference is a stationary and non-white noise sequence. When finally selecting the model, the auto.arima function is used to determine the arima(1,1,0) model, and finally the forecast function is used to predict the closing price of Zhongyuan Expressway in the next three days, and a comparative analysis is conducted with the actual value. It is concluded that ARIMA has a better effect in short-term stock price prediction and provides a reference for stock price prediction.

Keywords

ARIMA model, time series, stock price forecast, R language.

1. Introduction

With the rapid development of social economy and the continuous advancement of financial globalization, my country's stock market has formed a characteristic path that is consistent with my country's economic development. The scale has continued to expand, the number of listed companies has continued to increase, and the system construction has become increasingly perfect. Investors' awareness and enthusiasm for financial management are also constantly improving, and they expect to gain profits from trading in the stock market. However, financial management is risky and investment needs to be cautious. There are great risks and uncertainties in the stock market.

As an important indicator of the stock market, stock price analysis and prediction have always been one of the hot topics in finance. Many scholars have also done a lot of research on stock price prediction. Regarding the prediction of stock prices, Liu Song used the ARIMA model to model the closing price of Southwest Securities as historical data, using the currently popular Python as a modeling tool. Forecasting stock prices in the next four days has proven to be effective in short-term stock price predictions. Zhang Ni used the Long Short-Term Memory (LSTM) method to predict the stock price of Kweichow Moutai, providing strong support for investors in terms of the rationality, effectiveness and scientificity of the method used. Shi Linfeng took the closing price of Vanke Enterprise Co., Ltd.'s stock as the research object. Through empirical research, she used the ARIMA model to fit the closing price sequence of Vanke Company and conducted error analysis to provide decision-making reference for investors and government departments. Min Lin used the statistical software R to obtain the

ISSN: 1813-4890

data length of the CSI 300 Index for half a year, processed it effectively, and then used the ARIMA model and the support vector machine to build models respectively, and finally predicted and compared the CSI 300 Index. In summary, scholars have used various models to predict stock prices, and have achieved good results.

2. Theoretical Analysis and Model Introduction

Predicting the future values of time series data is a basic human activity, and the study of time series data also has wide applications in the real world. Economists try to understand and predict financial markets through time series analysis; urban planners predict future transportation needs based on time series data; climatologists use time series data to predict global climate change. Time series data is a data sequence that is arranged in chronological order, changes over time, and is interrelated. By studying the changing trends of historical data, we can evaluate and predict future data. Two models commonly used for forecasting financial time series data are ARMA and ARIMA models.

2.1. ARMA and ARIMA models

2.1.1 Autoregressive process

If a linear process that eliminates the mean and deterministic components can be expressed as

$$Y_t = \mu + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_t \quad (1)$$

where$\beta_i, i = 1,2,3,\cdots,p$ is an autoregressive parameter and ε_t is a white noise process, which y_t is called an p order autoregressive process and is represented by $AR(p)$. Y_t is the weighted sum of its p lag variables and the addition of ε_t.

Autoregressive models have many limitations:

①The autoregressive model uses its own data to make predictions

②Time series data must be stationary

③Autoregression is only suitable for predicting phenomena related to its own previous period (autocorrelation of time series)

2.1.2 Moving average process

If a linear random process that eliminates the mean and deterministic components can be expressed by the following formula

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-q} \quad (2)$$

Among them, $\theta_1, \theta_2 \cdots, \theta_q$ is the moving average parameter and ε_t is a white noise process. The above equation is called an $q$order moving average process, recorded as $MR(q)$. The reason why it is called "moving average" is that y_t is constructed from the weighted sum of ε_t and lag terms of ε_t. "Move" refers to the change of t, and "average" refers to the weighted sum. It should be pointed out that the influence of historical white noise in the AR model indirectly affects the current predicted value (by affecting the historical time series value).

2.1.3 Autoregressive moving average process

The stochastic process composed of autoregressive and moving average parts is called an autoregressive moving average process, recorded as where represents the maximum order of the autoregressive and moving average parts respectively. The general expression of is

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (3)$$

This formula shows:

①A random time series can be represented by an autoregressive moving average model, that is, the series can be explained by its own past or lagged values and random disturbance terms.

② If the sequence is stationary, that is, its behavior does not change over time, then we can predict the future through the past behavior of the sequence.

ISSN: 1813-4890

2.1.4 Differential autoregressive moving average process

Combining the autoregressive model (AR), the moving average model (MA) and the difference method, we get the differential autoregressive moving average model(ARIMA(p, d, q)), where d is the order that needs to be differentiated on the data.

2.2. ARIMA Modeling Steps

The modeling of ARIMA model can include the following steps:

First: Ensure that the time series is stationary. To do this, we need to perform differential operations or other operations to complete the stationary processing, and then complete the stationary test through the unit root test.

Second: Find a reasonable model, that is, select possible p values and q values. You can draw autocorrelation diagrams and partial autocorrelation diagrams to select alternative models, or you can find the optimal model through the functions that come with the R language model.

Third: Fit the model and evaluate the model. Generally speaking, if a model is suitable, the residuals of the model should satisfy the independent normal distribution. We can use the Q-Q plot to judge whether the residual terms comply with the normal distribution and conduct a white noise test. Determine whether the residual sequence is white noise.

Fourth: Use data to predict and analyze the model.

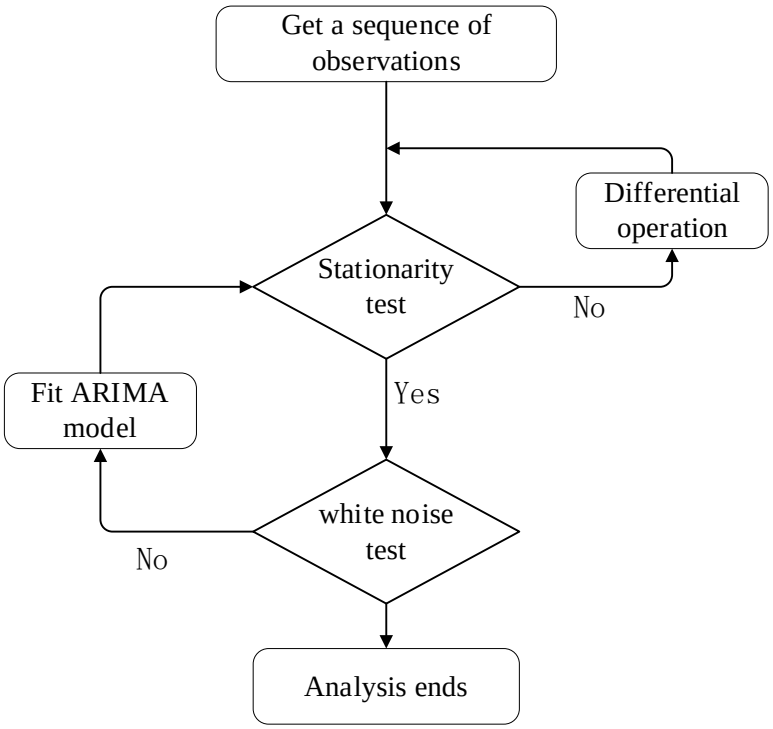

Fig 1: Flow chart of ARIMA modeling steps

3. Data Sources and Data Processing

3.1. Data Selection and Sources

On August 8, 2003, 280 million RMB ordinary shares (A shares) publicly issued by Zhongyuan Expressway were listed and traded on the Shanghai Stock Exchange under the stock code 600020, which is the only listed company in the transportation industry of Henan Province at present. The closing price is the price commonly recognized by market participants, which is the price accepted by everyone during the day, and investors generally choose to trade at the end. In this paper, the closing price of Zhongyuan Expressway (stock code 600020) on the stock exchange is selected as the research object, and the research period is from September 1, 2022

ISSN: 1813-4890

to September 1, 2023, with a total of 243 sets of sample data, and the first 240 data will be used for modeling and analysis, and the last 3 data will be used for forecasting and comparison to test the accuracy of the model. Data from Choice Financial Terminal.

3.2. Data processing

After downloading the stock data from the choice financial data terminal, the highest price, lowest price, opening price and other data are deleted, while the security name, security code, trading time and closing price data are retained and saved in CSV format.

4. Data Sources and Data Processing

4.1. Generate timing objects

The prerequisite for analyzing time series in R is to convert the analyzed object into a time series object, that is, a structure that includes observation values, starting time, ending time, and period (such as month, quarter, or year). Only after the data is converted into a time series object can it be analyzed, modeled and plotted using various time series methods. Since there is no trading on Saturdays, Sundays and holidays in stock trading, the abscissa trading time is changed to the number of transactions, and the ordinate corresponds to the company's closing price, see Fig. 2.

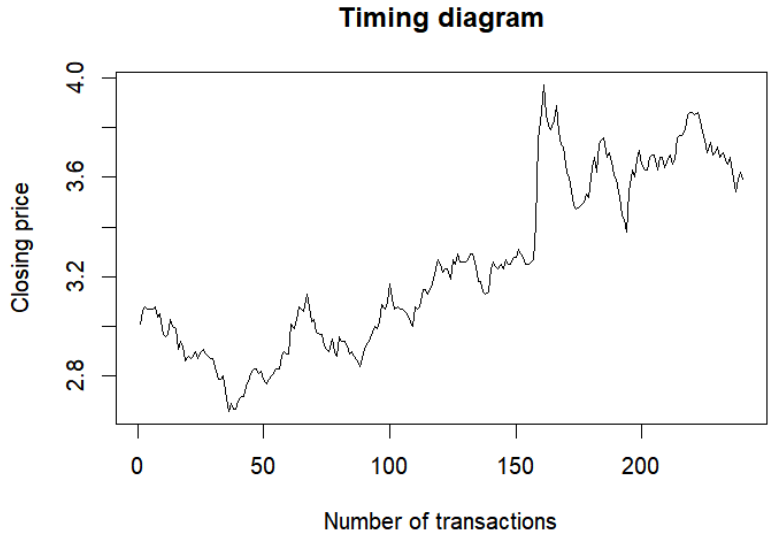

Fig 2. Zhongyuan Expressway closing price

4.2. Data preprocessing and stationarity testing

As can be seen from Figure 2, the variance between the number of observations seems to be stable, so we do not need to transform the data, but there may be some trend in the data. The time series of most stocks are basically non-stationary, and the original data need to be smoothed to become a stationary series.

Hypothesis testing method can be used to test the stationarity of time series. Judgment is made by constructing test statistics. The current mainstream method is the unit root test method, which tests whether there is a unit root in the sequence. If it exists, the time series is a non-stationary time series. This method is more scientific and accurate than the graphical analysis method.

The closing price of Zhongyuan Expressway has fluctuated greatly over time. Based on the graph, it is preliminarily judged that the time series is non-stationary, and relevant operations need to be performed to make it stationary. R language uses a function ndiff() function to solve

ISSN: 1813-4890

this problem - this function returns the number of differences required for a time series to become a stationary series. Therefore, for a time series x, k=ndiffs(x);

(a) If k=0, then it is a stationary sequence;

(b) If k>0, then x will become a stationary sequence after k-order difference.

For this example, the number of differences required for the sequence to become a stationary sequence is first order, so the first order difference operation is performed on the original data, as shown in Fig. 3. It can be seen that the sequence after first order difference is more stable and basically revolves around If the value of 0 fluctuates up and down in a small range, the time series can be judged to be a stationary time series.

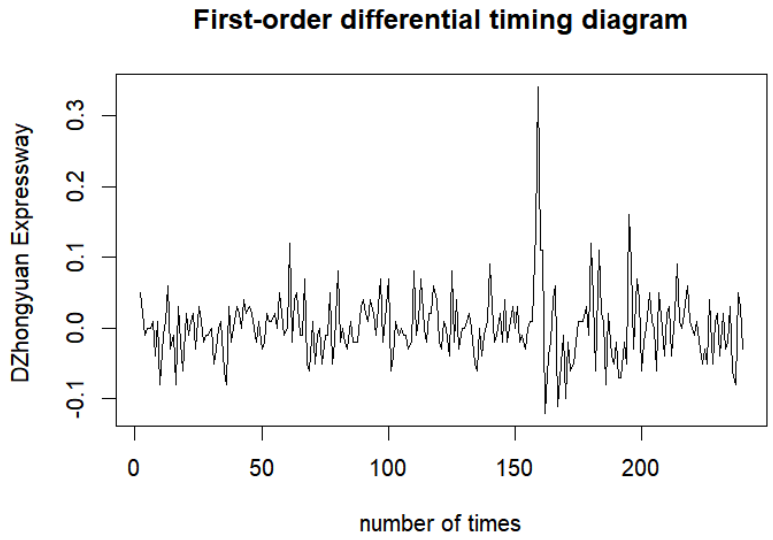

Fig 3. Closing price chart of Zhongyuan Expressway after first-order difference

The unit root test is used to test whether the time series after the first difference is stationary. The ADF (Augmented Dickey-Fuller) test can be used to verify the stationarity assumption of a time series. The test results are shown in Table 1. The test results show that the Dickey-Fuller statistic is -12.7797, which is far less than the Dickey-Fuller critical value at the 5% significance level. Therefore, the null hypothesis is rejected, indicating that the time series is a stationary time series.

Table 1. Unit root test results

Dickey-Fuller	-12.7797
Lag order	0
P-value	0.00
Critical Values	1%: -3.4580 5%: -2.8737 10%: -2.5733

Through purely random testing, also called white noise testing, the time series after the first difference is tested. If the sequence is white noise, it means there is no serial correlation and there is no need to model it. The Ljung-Box test can be performed, and the null hypothesis is that there is no correlation between sequence values with a lag period of less than or equal to m periods. In R, white noise testing can be implemented using the Box.test() function. The test results show that the X-squared statistic is 8.1018 and the p-value is 0.004<0.05. The null hypothesis is rejected, indicating that the time series is a non-white noise sequence.

Table 2. White noise test results

X-squared	df	P-value
8.1018	1	0.004422

ISSN: 1813-4890

4.3. Model identification and ordering

From the above analysis, we can see that the original time series data obtains a stationary time series through first-order difference, and the data is non-white noise and has serial correlation, so we choose ARIMA (p, d, q) to fit the time series model. . For determining the order of the model, we determine the values of p, d, and q. Since we perform a first-order difference operation on the original data, d=1 is selected.

Autocorrelation measures the correlation between individual observations in a time series. ACk is the correlation between a series of observations (Yt) and the observation before k period (Yt-k). The graph composed of correlations in different periods is the AutoCorrelation Function plot (ACF). The ACF plot can be used to select appropriate parameters for the ARIMA model and evaluate the fit of the final model. Partial AutoCorrelation Function plot (PACF) is the correlation between the two sequences when the effects of all values (Yt-1, Yt-2,...Yt-k+1) between the sequences Yt and Yt-k are removed. Partial autocorrelation plots are drawn based on different k values. The partial autocorrelation plot can also be used to find the most suitable parameters of the ARIMA model. The autocorrelation plots and partial autocorrelation plots drawn in R language are shown in Fig. 4 and 5.

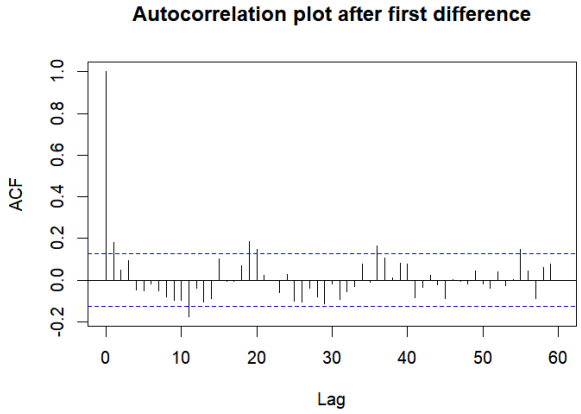

Fig.4 ACF

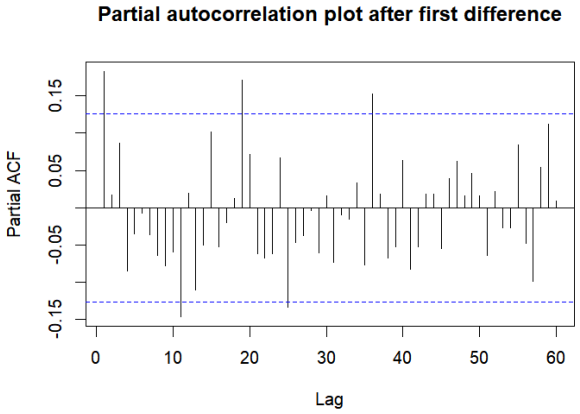

Fig.5 PACF

There is a certain degree of subjectivity in using graphics to artificially select values, which is biased. The auto.arima() function in the forecast package in R language can automatically select the optimal ARIMA model and obtain the model order more scientifically and effectively. The auto.arima function performs fixed-order identification of the model according to the BIC criterion, which is more scientific and accurate. The result of the function identification is the ARIMA (1,1,0) model, which is equivalent to performing first-order autoregression and first-order difference. Consistent with the d=1 obtained by the above graphical analysis method, in order to measure whether the model is accurate enough, this article uses the results of function identification as the criterion.

4.4. Model evaluation

Generally speaking, if a model is suitable, the residuals of the model should satisfy the normal distribution with a mean value of 0, and for any lag order, the autocorrelation coefficient of the residuals should be zero. In other words, the residuals of the model should satisfy independent normal distribution (that is, there is no correlation between the residuals). We can use the normal Q-Q plot and the Box.test() function to complete the relevant test, as shown in Fig. 6 and Table 3.

ISSN: 1813-4890

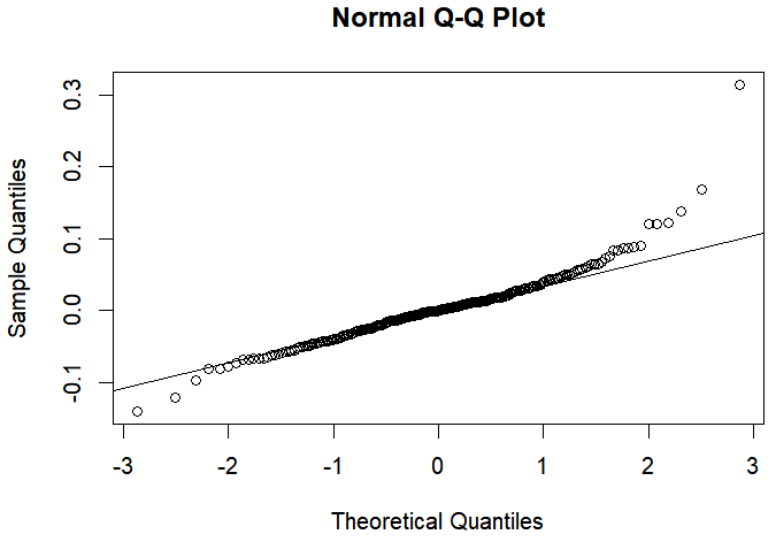

Figure 6. Normal Q-Q plot to determine whether sequence error satisfy the normality assumption

Table 3. Box-test error white noise test results

X-squared	df	P-value
0.005876	1	0.9389

The Q-Q plot is used to verify whether the residual term obeys the normal distribution. If the data satisfies the normal distribution, then the points in the data will fall on the line in the graph. Obviously, the results in this case are not bad. Before establishing the model, it is assumed that the residuals are uncorrelated. The residual white noise test result shows that the p value is 0.9389, which is greater than 0.05. The null hypothesis is accepted, that is, the effective information of the residual terms has been extracted, and the residual terms have no autocorrelation, belongs to the white noise sequence.

5. Data Sources and Data Processing

If the model residuals do not meet the assumption of normality or zero autocorrelation coefficient, If the model residuals do not meet the assumption of normality or zero autocorrelation coefficient, you need to adjust the model, add parameters, or change the number of differences. Obviously, the model selected above has better results and is used for prediction. Use the forecast function in R language to predict the closing price of Zhongyuan Expressway in the three days of August 30, 2023, August 31, 2023, and September 1, 2023. The results are as shown in Table 4.

Table 4. Closing price prediction results table

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
241	3.584441	3.523584	3.645297	3.491369	3.677512
242	3.58341	3.489035	3.677786	3.439075	3.727745
243	3.583219	3.463153	3.703286	3.399594	3.766845

The Plot() function can draw a prediction graph, in which the blue points represent the point estimates of the prediction points, and the light gray and dark gray areas represent the 80% and 95% confidence intervals respectively, as shown in Fig. 7.

ISSN: 1813-4890

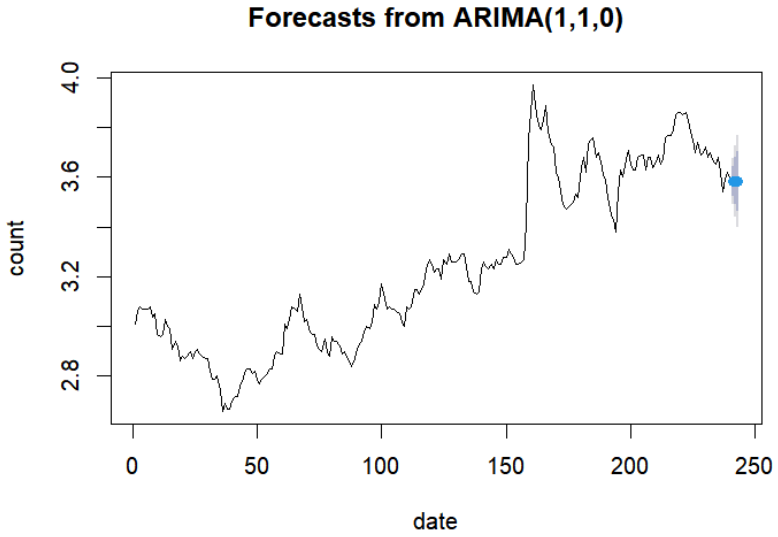

Fig 7. Stock prediction results and confidence interval chart

Comparing the real results with the predicted results, as shown in Table 5, it can be seen that the error between the predicted value and the real value is small, the error is controlled within 3%, and the three-day prediction results are within the 80% and 95% confidence intervals. . In summary, the model has higher prediction accuracy and better prediction effect, which reflects that the ARIMA model has better results in short-term prediction results.

Table 5. Comparison table between predicted values and actual values

Trading day	Predictive value	Actual closing price	Error
2022.11.24	3.5829	3.4800	0.030%
2022.11.25	3.5839	3.4700	0.033%
2022.11.28	3.5839	3.4900	0.027%

6. Conclusion and suggestions

As early as the 1920s, some scholars tried to use statistical principles to analyze and solve time series, instead of just discovering the changing patterns of random variables in the series from the descriptive level. To this end, British statisticians Yule and Walker successively proposed the autoregressive (Auto Regressive, abbreviated as AR) model, the moving average model (Moving Average, abbreviated as MA) and the ARMA model.

These three models laid the foundation for time series analysis and remain important even today.

This article uses R language to conduct time series prediction analysis on the closing price of Zhongyuan Expressway in the next three days. It conducts stationarity test and white noise test on the original time series, and concludes that the first-order difference sequence is a stationary non-white noise sequence. When finally selecting the model, use the function that comes with R language to select the arima(1,1,0) model. Finally, use the forecast function to predict the closing price of Zhongyuan Expressway in the next three days, and compare and analyze it with the actual value, achieved better results.

The analysis and prediction of stock prices have always been hot issues in the field of finance. For investors, they pay more attention to the rise and fall of stock prices, so accurately predicting the rise and fall trends of stock prices is more conducive to investors seeking advantages and avoiding disadvantages. There are risks in the stock market, so investment needs to be cautious. ARIMA has a better effect in short-term prediction of stock prices and hopes to provide some reference value to investors.

ISSN: 1813-4890

References

[1] Liu Song, Zhang Shuai. Empirical study on stock price prediction using ARIMA model [J]. Economic Research Guide, 2021(25):76-78.

[2] Zhang Ni. Application research on stock price prediction based on LSTM neural network [J]. Modern Business, 2021(16):116-118.

[3] Shi Linfeng. Research on the closing price of Vanke's stock based on ARIMA model [J]. Commercial Exhibition Economics, 2021(03):50-52.

[4] Min Lin. CSI 300 Index Forecast based on R language [J]. Modern Business, 2018(35):97-99.

[5] Wang Yan. Time series analysis - based on R [M]. Beijing: China Renmin University Press.

[6] Robert I, Kabacoff. R language practice [M]. 2nd edition. Beijing: People's Posts and Telecommunications Press, 2016.