Inflation Rate Prediction based on SARIMA-RF

Guangsheng Liu*

School of Mathematics, South China Normal University, Guangzhou, China, 510631

*Corresponding author: 20202231008@m.scnu.edu.cn

Abstract. Accurate prediction of the inflation rate can help governments prevent social problems caused by inflation, maintain social stability, and is crucial for the economy and society. Considering the nonlinear characteristics of China's inflation rate series, this paper uses the month-on-month Consumer Price Index (CPI) data from January 2011 to March 2023 to measure the inflation rate in China. Firstly, the monthly inflation rate series from January 2011 to December 2021 is analyzed to identify significant and complex seasonal effects. Therefore, suitable SARIMA models and Random Forest (RF) models are separately constructed. In order to better capture the nonlinear features of the inflation rate series, this paper proposes a SARIMA-RF fusion model based on simple linear weighting. Using monthly data from December 2022 to March 2023, the mean absolute error and root mean square error are used as evaluation metrics for forecasting errors. The results indicate that the prediction error of the SARIMA-RF fusion model is significantly lower than that of the SARIMA model and the RF model. The proposed model can more effectively extract the nonlinear features of China's inflation rate series and accurately predict its changes. It has valuable implications for the government, society, and individuals in China.

Keywords: Prediction of the Inflation Rate, SARIMA, Random Forest, SARIMA-RF.

1. Introduction

The inflation rate refers to the rate at which the general price level of goods and services in a specific country or region increases over a certain period of time. It is typically expressed as a percentage and reflects the extent of price increases caused by an increase in the money supply or a situation where demand exceeds supply. It is an important indicator for monitoring the macroeconomic performance of a specific area. High inflation rates can weaken the purchasing power of currency, increase costs for businesses and individuals, and harm economic growth and investment environment [1]. Therefore, controlling inflation rate contributes to maintaining economic stability and sustainable growth. Additionally, high inflation rates can lead to skyrocketing prices and increased cost of living, introducing instability into society. This may trigger social discontent and exacerbate inequality, negatively affecting social stability. It is evident that forward-looking prediction of the inflation rate is necessary. Accurate inflation rate forecasting helps address inflationary pressures and uncertainties in the economic environment, enabling better achievement of economic goals and maintenance of economic and social stability.

There are two main methods for predicting the inflation rate: the first involves survey analysis through statistical approaches to forecast future inflation. This method typically involves distributing questionnaires or conducting telephone surveys to collect inflation rate data from a predefined sample. The second method involves using econometric models, including time series methods, simultaneous equation models, and structural models, to forecast the inflation rate. Considering that the first method requires significant human and financial resources, resulting in higher costs, and the second method, particularly the use of simultaneous equation and structural models, may encounter situations where certain variables causing inflation rate changes are not significant or unobservable, rendering the models ineffective. Therefore, this study focuses on conducting research on inflation rate prediction using time series methods.

Early literature on inflation rate prediction using time series methods primarily focused on linear forecasting models. However, with the work of Liu Jinquan (2002) [2], who utilized the ARMA(mn)-GARCH(p, q) model to fit China's inflation rate and discovered significant conditional heteroscedasticity in the inflation rate series, researchers began to pay attention to the nonlinear

characteristics of inflation series. For instance, Wang Shaoping and Peng Fangping (2006) [3] conducted research on China's inflation rate using the SETAR model based on the 1952-2004 retail price index, ultimately finding that the SETAR model outperformed the AR(p) model. Similarly, Zhang Jian (2011) [4] used the SARIMA model to forecast the inflation rate based on China's month-on-month and year-on-year CPI data, and the results showed that the SARIMA model had superior performance in both fitting and forecasting compared to the ARIMA model. Furthermore, Chen Yanbin et al. (2020) [5] selected monthly data from 2004 to 2014 in China and established LSTM and BVAR models based on machine learning algorithms to predict the inflation rate. The results demonstrated that the nonlinear LSTM model outperformed the linear BVAR model in predicting the inflation rate.

Considering that the nonlinear changes in the inflation series are not solely caused by seasonal effects but may also be influenced by macroeconomic conditions and government policies, this study will use the month-on-month CPI data from January 2011 to March 2023 to measure China's inflation rate. Based on the monthly inflation series data from January 2011 to December 2021, suitable SARIMA models and Random Forest (RF) models will be constructed according to their sequence characteristics. To better capture the nonlinear features of the inflation series, a simple linear weighting approach will be used to combine these two models into the SARIMA-RF fusion model. Using monthly data from December 2022 to March 2023, the predictive errors of the fusion model, SARIMA model, and RF model will be compared to identify the relatively optimal model that can effectively extract the nonlinear characteristics of China's inflation rate series and achieve more accurate predictions.

While most existing literature has used the year-on-year Consumer Price Index (CPI) series to measure China's inflation rate at different time periods, the year-on-year data may be influenced by base effects. Furthermore, studies have shown that seasonally adjusted month-on-month CPI data is more suitable for real-time macroeconomic monitoring compared to year-on-year data [6]. Moreover, this paper proposes a novel approach, the SARIMA-RF fusion model based on simple linear weighting, to further extract the nonlinear features of the inflation series.

2. Data Description

This study utilizes the month-on-month CPI data from January 2011 to March 2023 to measure the inflation rate in China and conducts an investigation based on this data. The relevant data is sourced from the China Economic Prosperity Monthly Report. Initially, the monthly data from January 2011 to December 2021 is selected as the dataset for model fitting and subsequent analysis.

2.1. Stability Analysis of the Original Series

In this study, the time series plot of the month-on-month CPI data from January 2011 to December 2021 is first generated. Additionally, the autocorrelation function (ACF) plot of the series is examined to conduct an initial analysis of its stationarity.

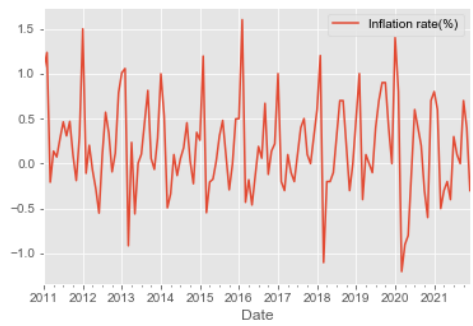


Figure 1. Time Series Plot of the Original Series

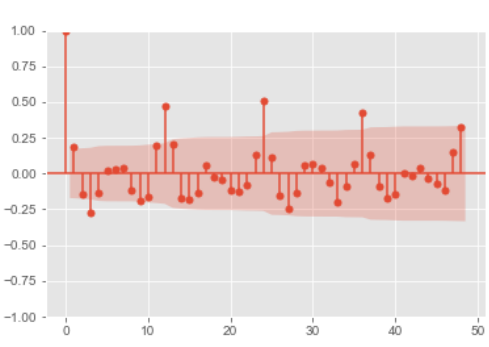


Figure 2. Autocorrelation Coefficient Plot of the Original Series

According to Figure 1, it can be observed that the original inflation rate series does not exhibit a clear trend, and its fluctuations mainly range between -1.5 and 2. The magnitude of fluctuations does not appear to change over time. Therefore, it can be concluded that the series does not exhibit heteroscedasticity. However, upon examining Figure 2, the autocorrelation function (ACF) plot of the original series, it is noticed that the autocorrelation coefficients at lag 12, 24, and 36 are all outside the confidence interval of two standard deviations. Hence, the original series is considered non-stationary, indicating the presence of significant seasonal effects.

2.2. Stability Analysis of the Difference Series

Taking into account the seasonal effect observed in the autocorrelation coefficient plot of the original series, this study applies a differencing technique with a lag of 12 months to the original inflation rate series. The differenced series is then analyzed for stationarity through the examination of its time series plot, the Augmented Dickey-Fuller (ADF) test, and the autocorrelation coefficients.

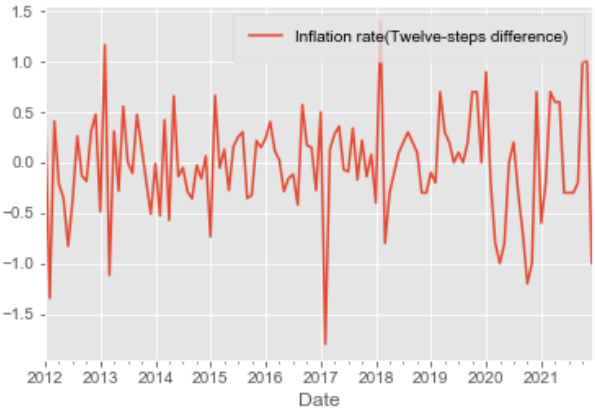


Figure 3. Time Series Plot of the Differenced Series

Table.1. ADF Test Results of the Differenced Series

Types	Lag Orders	P-value
Autoregressive Structure without a Drift Term	0	<0.0001
	1	<0.0001
	2	<0.0001
Autoregressive Structure with a Drift Term	0	<0.0001
	1	<0.0001
	2	<0.0001
Autoregressive Structure with Trend Regression	0	<0.0001
	1	<0.0001
	2	<0.0001

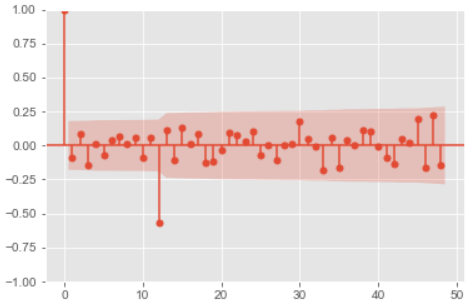


Figure 4. Autocorrelation Coefficient Plot of the Differenced Series

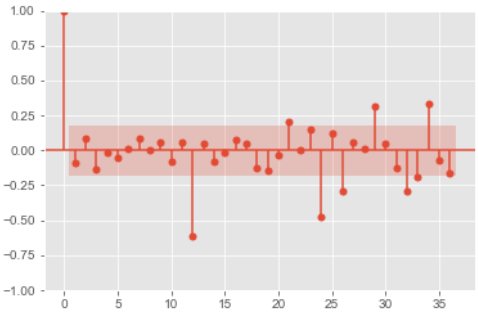


Figure 5. Partial Autocorrelation Coefficient Plot of the Differenced Series

According to the time series plot of the differenced series (Figure 3), it can be observed that the differenced series continues to fluctuate within the range of -1.5 to 2, with no significant change in the magnitude of fluctuations over time. There is no apparent trend effect or heteroscedasticity. Furthermore, considering the ADF test results presented in Table 1, where the p-values for different types are all less than 0.0001, it can be concluded that the differenced series has no trend effect. However, as shown in Figure 4, the autocorrelation coefficients of the differenced series still significantly exceed the confidence interval at a lag of 12. Therefore, it can be concluded that differencing alone is not sufficient to effectively eliminate the seasonal effect from the series.

2.3. Test of Randomness for the Differenced Series

In order to ensure that the seasonal effect observed in the differenced inflation rate series is not a random occurrence, and considering the limited sample size, this study conducts a test of randomness using the Ljung-Box test statistic on the differenced series. The test results are presented in the following table:

Table.2. Test of Randomness for the Differenced Series

Lag Orders	P-value
2	0.02258
4	0.00050
6	0.00252

According to the results presented in Table 2, this study finds that at lag orders 2, 4, and 6, the p-values are all less than 0.05. This indicates that, at a significance level of 5%, the null hypothesis of the differenced series being white noise is rejected. Therefore, it can be concluded that the series is not a white noise series. Taking into account the previous analysis, it is suggested that there may exist complex relationships between the seasonal effect and short-term correlations in the original inflation rate series. Consequently, in the subsequent analysis, this study will attempt to fit the series using both the SARIMA model, specifically the ARIMA multiplication model, and the RF (Random Forest) model.

3. Model Construction

3.1. Construction of the SARIMA Model

SARIMA (Seasonal AutoRegressive Integrated Moving Average) is a statistical model used for time series analysis and forecasting. It is an extension of the ARIMA model specifically designed to handle data with noticeable seasonal patterns. The SARIMA model consists of four main components: seasonal component, autoregressive component, differencing component, and moving average component. It combines these components to capture and forecast the seasonal variations and trends in the time series data. The seasonal component models repetitive seasonal patterns in the time series data, such as yearly, quarterly, or monthly seasonal changes. The autoregressive component considers the relationship between the current observation and past observations. The differencing component is used to handle non-stationary time series, while the moving average component captures the relationships among error terms [7][8][9]. The structure of the SARIMA model, after d-order differencing for trend and D-order seasonal differencing with a periodicity of S, is as follows:

$$\nabla^{d}\nabla_{S}^{D}x_{t}=\frac{\Theta(B)\Theta_{S}(B)}{\Phi(B)\Phi_{S}(B)}\varepsilon_{t} \tag{1}$$

Where $\Theta(B)=1-\theta_{1}B-\ldots-\theta_{q}B^{q},\Phi(B)=1-\emptyset_{1}B-\ldots-\emptyset_{p}B^{p},\Theta_{S}(B)=1-\theta_{1,S}B^{S}-\ldots-\theta_{Q,S}B^{QS},\Phi_{S}(B)=1-\emptyset_{1,S}B^{S}-\ldots-\emptyset_{P,S}B^{PS}$, B represents the lag operator, $\{\varepsilon_{t}\}$represents a white noise series, the model can be abbreviated as ARIMA(p, d, q)×$(P,D,Q)_{S}$.

Based on the speculation that there may exist complex relationships between the seasonal effect and short-term correlations in the original inflation rate series, this study will proceed to determine the order of the SARIMA model based on the characteristics observed in the autocorrelation and partial autocorrelation plots of the differenced series.

Based on the characteristics of the autocorrelation and partial autocorrelation coefficients within lag 12 of the differenced series in Figures 4 and 5, this study suggests that fitting the series using $\mathrm{ARIMA}(1,0,1)\times(0,0,1)_{12}$ is possible. However, considering the limitations of determining the order solely based on the autocorrelation and partial autocorrelation plots, this study also considers $\mathrm{ARIMA}(1,0,1)\times(1,0,1)_{12}$ to capture its nonlinear features. Finally, taking into account the limited sample size, the study selects the relatively optimal model based on the AIC criterion.

Table.3. AIC Values for Different SARIMA Models

Models	AIC
$\mathrm{ARIMA}(1,0,1)\times(0,0,1)_{12}$	205.975
$\mathrm{ARIMA}(1,0,1)\times(1,0,1)_{12}$	145.515

The results in Table 3 show that the AIC value of $\mathrm{ARIMA}(1,0,1)\times(1,0,1)_{12}$ is significantly lower than that of $\mathrm{ARIMA}(1,0,1)\times(0,0,1)_{12}$. Therefore, this study ultimately selects $\mathrm{ARIMA}(1,0,1)\times(1,0,1)_{12}$ to establish the time series model for the inflation rate series, capturing the short-term autocorrelation features and seasonal effects in the series.

After determining the order of the model, it is necessary to estimate the parameters of the model. In this study, considering that the least squares estimation method has relatively good estimation accuracy compared to moment estimation and maximum likelihood estimation methods, and it does not require prior assumptions about the distribution of the series, the least squares estimation method is employed to estimate the parameters.

Table.4. Model Parameter Results

Parameters	Parameter-values	P-value
$\emptyset_1$	0.8219	<0.001
θ_1	-0.8121	<0.001
$\emptyset_{1,12}$	0.9996	<0.001
$\theta_{1,12}$	-0.9659	0.019

The results in Table 4 show that the p-values of the significance tests for each parameter are significantly less than 0.05, indicating that all the parameters of the model pass the significance tests at a 5% significance level. Based on these results, this study further conducted the model validity test using the LB statistic. The test results are presented in Table 5.

Table.5. Model Validity Test Results

Lag Orders	P-value
6	0.99719
12	0.83491
18	0.90175

The results in Table 5 indicate that the p-values for the residual series at lag orders 6, 12, and 18 are all significantly greater than 0.05. This means that at a 5% significance level, we cannot reject the null hypothesis that the residual series is white noise. Therefore, it can be concluded that the residual series of the model exhibits characteristics of white noise, indicating that the model effectively captures the information in the series. Based on the validity test, the final model specification is as follows:

$$\nabla_{12}x_t=\frac{(1+0.8121B)(1+0.9659B^{12})}{(1-0.8219B)(1+0.9996B^{12})}\varepsilon_t,\mathrm{Var}(\varepsilon_t)=0.1158 \quad (2)$$

Additionally, in order to further investigate the heteroscedasticity of the series, this study plotted the residual plot of the model as follows:

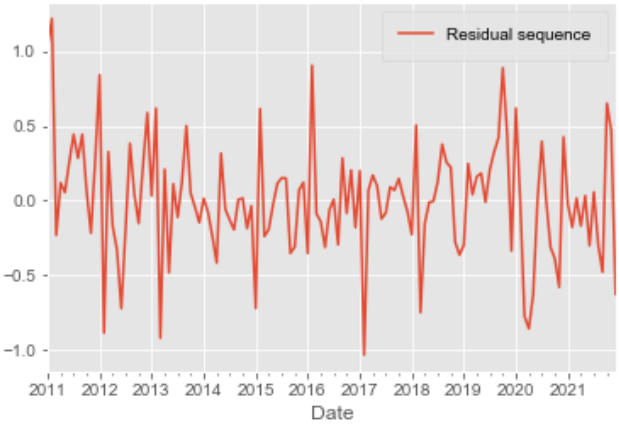


Figure 6. Residual Plot

Figure 6 shows that the residual series of the model fluctuates mainly within the range of [-1.5, 1.5] without any apparent trend. This indicates that the assumption of approximate homoscedasticity holds, and there is no need to apply the GARCH model to adjust the model.

3.2. Construction of the RF Model

Random Forest [10][11], as a compositional supervised learning method, primarily operates by performing bootstrap sampling of the data set (sampling with replacement for rows) and feature sampling (sampling without replacement for columns). Following the principle of recursive splitting, it starts from the root node and progressively partitions the training set (where the root node continuously splits based on the minimum impurity principle, dividing the training data into left and right nodes, each containing a subset of the training data) until the branching stop criteria are met. This process constructs regression trees. By averaging the results of the dependent variable Y obtained from each regression tree, the final output result of the Random Forest can be obtained. As time series data often exhibit nonlinear trends and patterns, traditional linear models may not accurately capture and predict these nonlinear features. However, the RF (Random Forest) time series prediction model, by integrating multiple decision trees, allows each decision tree to adapt to the nonlinear relationships in the data, thereby better capturing and predicting the nonlinear features in the time series. Through the branching and combination of decision trees, the RF model can flexibly fit nonlinear trends, periodicities, and seasonality features in the data, providing accurate nonlinear time series predictions.

This study utilizes the RandomForestRegressor function from the sklearn.ensemble library in Python to establish the RF (Random Forest) model for the original inflation series.

3.3. Construction of the SARIMA-RF Model

To fully capture the nonlinear features in the original series and leverage the characteristics of both the SARIMA model and RF (Random Forest) model, this study combines the previously established SARIMA model and RF model using linear weighting [12][13], resulting in the SARIMA-RF model as follows:

$$X_t = \frac{S(t)+R(t)}{2} \tag{3}$$

Where S(T) and R(T) represent the fitted or predicted values of the previously established SARIMA model and RF model at time t. The fitting performance of the SARIMA-RF model for the period from January 2011 to December 2021 is shown in the following figure, which also presents the model's predicted values for the period from January 2022 to March 2023.

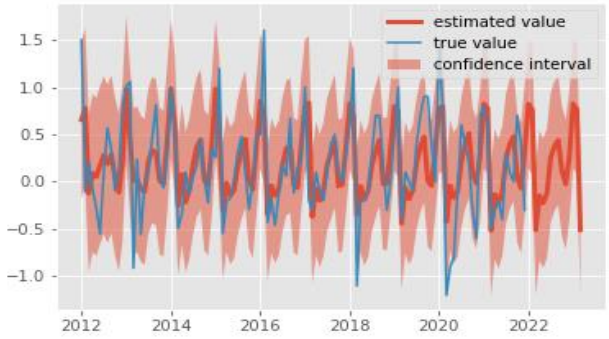


Figure 7. SARIMA-RF model fitting and prediction plot

From Figure 7, we can observe that the majority of the true values fall within the confidence interval, and the variations in the true values align closely with the variations in the fitted values. This indicates that the combined SARIMA-RF model is effective in capturing the non-linear features of the original inflation rate series. Moreover, the model predicts an overall upward trend followed by a downward trend in the monthly inflation rate in China from January 2022 to March 2023.

3.4. Model Comparison

To further demonstrate the superiority of the model, this study also utilizes the real data from January 2022 to March 2023. The evaluation is based on the mean absolute error (MAE) and root mean square error (RMSE), which are commonly used measures of error [14][15]. The formulas for calculating MAE and RMSE are as follows:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |\mathrm{pred}_i - \mathrm{actual}_i|}{n}, \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} (\mathrm{pred}_i - \mathrm{actual}_i)^2 / n}, \tag{5}$$

Where n represents the number of samples in the error calculation, pred_i represents the i-th predicted value, and actual_i represents the i-th true value. The prediction errors are calculated for the SARIMA model, RF model, and SARIMA-RF model, and the results are presented in the following table:

Table.6. Error Comparison of Different Models

Models	MAE	RMSE
SARIMA	0.3073	0.4347
RF	0.3245	0.4145
SARIMA-RF	0.2434	0.3287

The results in Table 6 demonstrate that both MAE and RMSE of the SARIMA-RF model are significantly smaller than those of the SARIMA and RF models. This indicates that the fusion model is effective in combining the strengths of both models and extracting the nonlinear features from the original inflation rate series.

4. Conclusion

Accurate inflation rate forecasting is essential in helping governments and individuals proactively respond to inflationary pressures and uncertainties in the economic environment. It enables better economic planning, maintenance of economic and social stability, and achievement of economic objectives. Through exploration, this study has identified the presence of significant seasonal effects in China's inflation rate series. However, considering that factors influencing the nonlinear changes in the inflation rate series extend beyond just seasonal effects, including government interventions and market adjustments, this study has employed a simple linear weighted fusion of the SARIMA

and RF models. The SARIMA-RF model has been established to fit and predict the inflation rate series, yielding results that demonstrate its effectiveness in extracting nonlinear features from the series and improving the accuracy of inflation rate predictions.

Indeed, the fusion approach used in this study, which is a simple linear weighted fusion, may not fully capture the interrelationships between the SARIMA and RF models. Future research could delve deeper into exploring more advanced fusion techniques that take into account the correlations and interactions between these models. By doing so, it would be possible to provide more valuable insights for inflation rate forecasting and macroeconomic policy-making in China.

References

[1] TAN Xiaofen, WANG Xinkang, YANG Yahan. Globalisation and inflation[J]. China Industrial Economy, 2023, No.422(05): 24-42.

[2] Liu Jinquan. Analysis of the symmetry and volatility of China's inflation path [J]. China Management Science, 2002, (6).

[3] Wang Shaoping, Peng Fangping. Fitting and forecasting of inflation rate in China using SETAR model [J]. Statistics and Decision Making, 2006(14):84-85.

[4] Jian Zhang. Application of SARIMA model in forecasting China's CPI [J]. Statistics and Decision Making, 2011, No.329(05): 28-30.

[5] CHEN Yanbin, LIU Lingjun, CHEN Xiaoliang. Inflation forecasting in China - A comparative analysis based on LSTM model and BVAR model [J]. Research on Financial Issues, 2021, No.451(06): 18-29.

[6] Guo Z. Is inflation persistence in China long memory or short memory [J]. Macroeconomic Research, 2023, No.290(01): 14-25+100.

[7] LI Tianquan, LI Hongmin, LI Yuna et al. Application of SARIMA model to assess the impact of a novel coronavirus infection outbreak on the epidemiological trend of hand-foot-mouth disease in Zhengzhou City[J/OL]. Modern Disease Prevention and Control, 2023(07): 500-503+542[2023-07-17].

[8] YANG Zhenhao, ZHANG Junbo, YANG Chenxing et al. Forecasting the consumer price index of aquatic products in China based on SARIMA model [J]. Marine Lake and Marsh Bulletin,2021,43(02):131-138.

[9] Deng Wenwen, Zeng Dewei, Wu Xiaohua et al. The role of seasonal autoregressive sliding average model (SARIMA) in the prediction of foodborne disease occurrence[C]//Chongqing Preventive Medical Association. Proceedings of the Fourth Council and Annual Meeting of Chongqing Preventive Medicine Society. [Publisher unknown],2022:13.

[10] Yujie Jiang. Price prediction of online short-term rentals [D]. Guizhou University of Finance and Economics, 2019.

[11] Wang Feng. Price prediction of treasury bond futures based on random forest regression [D]. South China Agricultural University, 2018.

[12] Qing Huang, Chen Xiaoxia, Liu Huaiyan. Network traffic prediction method based on SARIMA-LSTM combined model [J]. Jiangsu Communication, 2023, 39(02): 87-91.

[13] YUAN Yi, PANG Zhenlu, CUI Qingxia et al. Prediction of acute haemorrhagic conjunctivitis onset based on SARIMA model or SARIMA-GRNN combined model [J]. Journal of Zhengzhou University (Medical Edition), 2023,58(01):78-82.

[14] Cheng Yiting. Research on stock forecasting method based on investor attention [D]. Shanxi University of Finance and Economics,2023.

[15] Gao Bin. Research on the compensation technology of time-pressure dispensing system based on deep learning [D]. Shanghai Normal University, 2023.