ISSN: 2688-8645

Analysis and Forecast of Currency Based on ARIMA Model and Random Forest Regression Model

Wenjun Yang

Department of statistics, Jinan University, Guangzhou, 510632, China

Abstract

This paper collects the monthly data of money supply from January 1996 to March 2020, which includes eight indicators: money and quasi-money, money and quasi-money year-on-year growth, currency and its year-on-year growth, cash in circulation and its year-on-year growth, current deposit and its year-on-year growth. ARIMA model and random forest regression model are used to model and analyze these data respectively. The data used in ARIMA model only involves money and quasi-money, while the random forest regression model takes money and quasi-money as dependent variables and others as independent variables. Although the data involved in the two models are different, the effect of prediction of thses two models are similar.

Keywords

Money and quasi-money; ARIMA model; Random forest regression.

1. Introduction

Money supply, also known as money stock, means the sum of cash and deposits in circulation in certain time. Money supply is one of the main economic statistical indicators, and it is compiled and issued by the central banks in various countries. Predicting the growth and changes of money stock is basic for a country to formulate monetary policy. Broad money supply (M2) refers to the cash outside the banking system and other deposits. It is an important indicator to reflect the money supply. It includes all forms of money reflecting purchasing power, and it measures changes in social aggregate demand and the future inflation.

Narrow measure of money supply (M1) is the symmetry of "broad money supply". It means the sum of cash in circulation and demand deposits of commercial banks. The current deposit is the currency of deposit in the narrow currency. Generally speaking, both M1 and M2 tend to increase overall, and M1 is always lower than M2. Although the two are generally on the rise, there are some small fluctuations. The year-on-year growth of M1 and M2 both show irregular changes, sometimes rising and falling, but the year-on-year growth trend of M1 and M2 is roughly the same. This paper collects the monthly data of money supply from January 1996 to March 2020, which includes eight indicators: money and quasi-money, money and quasi-money year-on-year growth, currency and its year-on-year growth, cash in circulation and its year-on-year growth, current deposit and its year-on-year growth. The ARIMA model and the random forest regression model are adopted to analyze and predict M2.

2. Methodology

2.1. ARIMA Model

ARIMA model[1][2], also known as Autoregressive Integrated Moving Average model. In ARIMA(p, d, q), AR is "autoregression", and p is the number of autoregressive items; MA is "moving average", and q is the number of moving average items ; d is the number of differences. If the time series y_t satisfies:

ISSN: 2688-8645

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \text{B} + \beta_p Y_{t-p} + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \text{B} \ \alpha_q \varepsilon_{t-q} \quad (1)$$

which means y_t obeys the (p, q) order autoregressive moving average mixed model.

A model with the following structure is called a differential integrated moving average autoregressive model, abbreviated as ARIMA(p, d, q):

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1-L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\varepsilon_t \quad (2)$$

where L is the lag operator, $d > 0$.

Denote ∇ as the difference operator, then we have

$$y_{t-p} = B^p y_t, \forall p \geq 1 \quad (3)$$

If there is an order homogeneous non-stationary time series y_t, and $\nabla^d y_t$ is a stationary time series. So it can be set as the ARMA(p, q) model, that is,

$$\lambda(B)(\nabla^d y_t) = \theta(B)\varepsilon_t \quad (4)$$

where $\lambda(B) = 1 - \lambda_1 B - \lambda_2 B^2 - \text{B} \ - \lambda_p B^p$, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \text{B} \ - \theta_p B^p$ are the autoregressive coefficient polynomial and the moving average coefficient polynomial respectively. ε_t is white noise sequence. Because the order difference is made for the sequence , the model set can be called the integrated moving average autoregressive model, which is also abbreviated as ARIMA(p, d, q).

2.2. Random Forest Regression

Random Forest (RF) consists of multiple decision trees, and the final output is jointly determined by each decision tree. When dealing with regression problems, the average of the output of each decision tree is the final result. For training stage, RF uses bootstrap to collect multiple different sub-training datasets from the training dataset to train multiple different decision trees; for prediction, RF obtains final results by averaging the prediction results of every decision tree. The RF regression (RFR) has the following steps: model training; model data prediction; computing feature importance. The random forest algorithm is an improvement of the Bagging Algorithm [3][4].

By calculating the feature importance, we can know which indicator has a large influence on the predictor. The importance of a node is

$$n_k = w_k G_k - w_{left} G_{left} - w_{right} G_{right} \quad (5)$$

where w_k, w_{left}, w_{right} are the ratio of the samples in left and right to the total samples. G_k, G_{left}, G_{right} are the impurity of left and right child nodes. By calculating (6), the importance of a feature can be calculated by (6):

$$f_i = \frac{\sum_{j \in nodes \ split \ on \ feature \ i} n_j}{\sum_{k \in all \ nodes} n_k} \quad (7)$$

ISSN: 2688-8645

In order to make the importance of all features add up to 1, we need to normalize them by the following formula

$$f_{ni} = \frac{f_i}{\sum_{j \in all\ features} f_j} \tag{8}$$

3. Results and Discussion

3.1. Results of the ARIMA Model

To apply ARIMA model[5][6], the ADF unit root test is needed to test the stationarity of the data. For non-stationary time series, logarithm or difference is generally performed to make them become stationary. The first-order difference and second-order difference are performed on the sequence respectively, and the results show that the second-order difference is more stable. The unit root test is used to test the stationarity of the first-order difference sequence and the second-order difference sequence. The results are shown in Table 1 and Table 2. Since the P value of the first-order difference series is 0.69, and at the 10% significance level, the value of the statistic (-1.17) is larger than the critical level (-2.57), so it can be considered that the series has a unit root and is not stationary. The P value of the second-order difference sequence is very small and the value of the statistic (-13.68) is smaller than the critical value, so the second-order difference sequence is considered to be stationary. Based on this result, we use the M2 second-order difference sequence to build ARIMA model.

Table 1. Stationarity of first-order difference sequence

Statistics	-1.17
P value	0.69
Lag order	12
1% significance level	-3.45
5% significance level	-2.87
10% significance level	-2.57

Table 2. Stationarity of second-order difference sequence

Statistics	-13.68
P value	1.39E-25
Lag order	11
1% significance level	-3.45
5% significance level	-2.87
10% significance level	-2.57

Figure 1. Shows the prediction results of M2 from 1996.01 to 2014.07 compared with the actual value. The total error is 51.52.

ISSN: 2688-8645

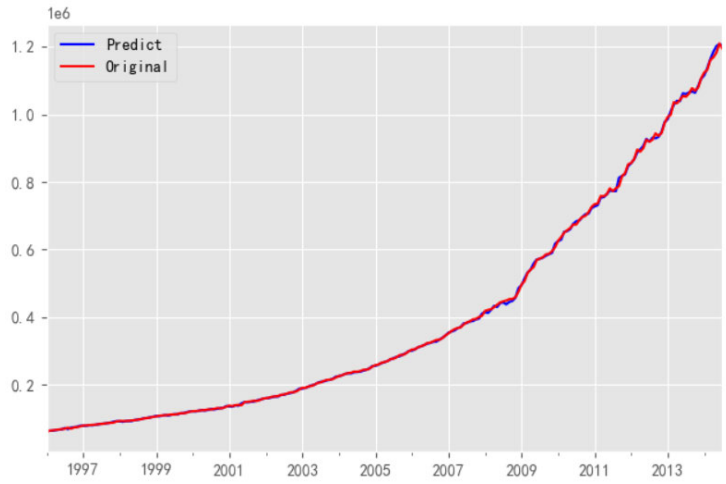

Figure 1. Comparison of the actual and predicted values

The specific order of the model is determined by the AIC or BIC. The AIC value of the ARMA (7, 8) model is the smallest, which is 4557.77. And the BIC value of the ARMA (4, 5) model is the smallest. By executing residual autocorrelation test and white noise test, the ARMA (7, 8) model is better than the ARMA (4, 5) model. Table 3 shows the parameter estimates of ARMA (7, 8) model. Besides the constant, the estimated values of other coefficients are significant.

Table 3. The parameter estimates of ARMA (7, 8) model

	Coefficient	Standard Error	Z-value	P-value	Lower CI	Upper CI
Constant	0.69	40.09	0.02	0.99	-77.90	79.27
AR(1)	-1.85	0.06	-33.58	0.00	-1.95	-1.74
AR(2)	-0.93	0.06	-16.44	0.00	-1.05	-0.82
AR(3)	0.85	0.01	84.55	0.00	0.83	0.87
AR(4)	0.81	0.05	15.78	0.00	0.71	0.92
AR(5)	-0.87	0.03	-33.88	0.00	-0.92	-0.82
AR(6)	-1.63	0.06	-29.74	0.00	-1.74	-1.53
AR(7)	-0.79	0.05	-17.10	0.00	-0.88	-0.70
MA(1)	0.83	0.07	11.22	0.00	0.69	0.98
MA(2)	-0.83	0.07	-12.07	0.00	-0.97	-0.70
MA(3)	-1.51	0.10	-15.10	0.00	-1.71	-1.32
MA(4)	0.21	0.07	3.09	0.00	0.08	0.35
MA(5)	1.56	0.08	18.67	0.00	1.39	1.72
MA(6)	0.68	0.09	7.76	0.00	0.51	0.85
MA(7)	-0.59	0.08	-7.75	0.00	-0.74	-0.44
MA(8)	-0.67	0.06	-11.15	0.00	-0.78	-0.55

3.2. Results of the Random Forest Regression

232 pieces of data are training samples and 59 pieces of data are testing samples. We set the number of decision trees as 5, 10, 20, 50, 100, 200 and set the maximum tree depth as 3, 5, 7 and set the maximum number of features as 0.6, 0.7, 0.8, 1. By using grid search to select the parameter with the best effect, we can specify the best model. After calculation, the random forest model with the best parameter combination is: 200, 7 and 0.8. Table 4 shows the importance of features (the sum of the importance of six features is one). The corresponding feature importance histogram is shown in Figure 2. It can be seen that the importance of m0,

ISSN: 2688-8645

m1, and saving are 0.441797, 0.377656 and 0.171556 respectively. So the influence of these three on M2 is relatively large, especially m0 and m1.

Table 4. Feature importance of different indicators

Features	Importance
Features m0	0.441797
Features m1	0.377656
Features saving	0.171556
Features m1_rise	0.008127
Features m0_rise	0.000829
Features saving_rise	0.000035

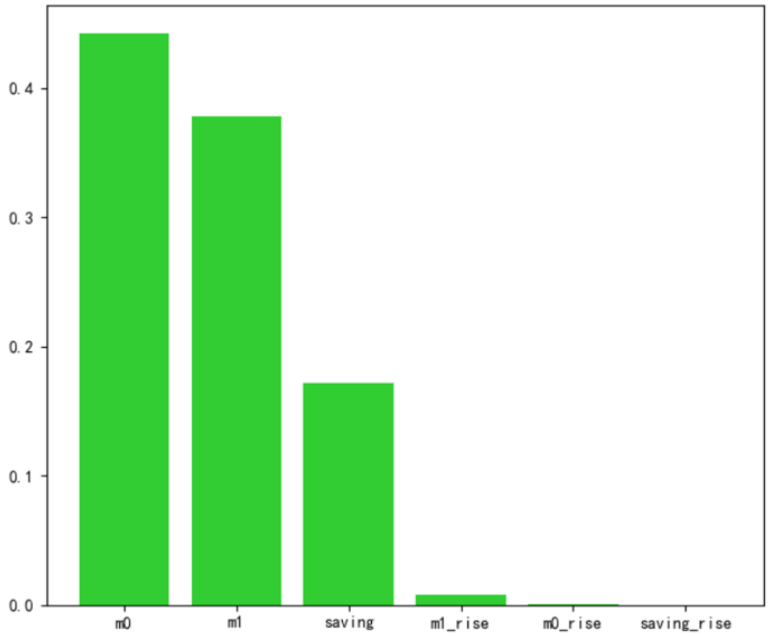

Figure 2. Importance histogram of different features

The model predicts 59 pieces of data in the test set. The actual and predicted values are shown in Figure 3. The horizontal axis represents the index of the test set, and the vertical axis is in 100 million. The error is 110.0115. The main reason for the large error is that the data itself is relatively large, and the difference between the actual values and the predicted values is in thousand.

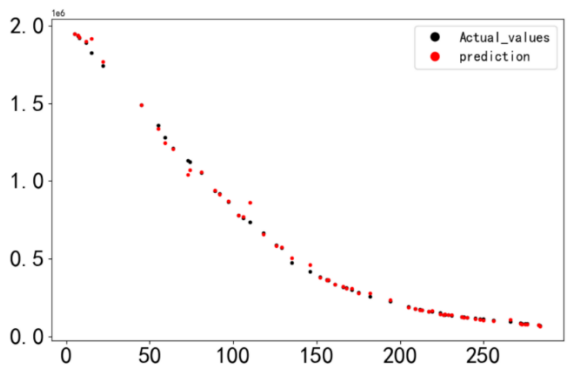

Figure 3. Comparison of the actual and predicted values

ISSN: 2688-8645

4. Conclusion

This paper collects eight indicators about money supply data from 1996.01 to 2020.03, and apply ARIMA model and RFR to analysis and predict these data. Firstly, the M2 sequence is tested for stationarity and processed by second-order difference. ARIMA model is built with the second-order difference sequence. The ARIMA (7, 2, 8) model with the lowest AIC value of 4557.77, and we use this model to predict the data with error being 51.52. This paper also uses RFR model to analyze and predict M2. RFR model with a maximum tree depth of 7, a maximum number of features of 0.8, and a number of decision trees of 200 is built. The importance of m0, m1, and saving are 0.441797, 0.377656, and 0.171556 respectively, which means they have a relatively large impact on M2. Finally, the model is used to predict the testing samples and the total error is 110.0115. Based on these two models, it can be seen that the prediction effect is good but the error is large. This is caused by the data itself, so the gap between the actual values and the predicted values is from the actual value itself.

Acknowledgments

This authors would like to thank the editor, the referees for their constructive comments.

References

[1] Hu Chengyu, Yang Zhengyuan, Liu Yaqing. Econometric Analysis of China's Money Supply Based on SARIMA Product Season Model[J]. Journal of Natural Science of Harbin Normal University,2020,36(02):19-25.

[2] Hou Tiantian, Yang Cong, Zhan Binghuan. Forecast of China's Fiscal Revenue Based on ARIMA and Markov Chain Model[J]. Journal of Pingdingshan University, 2020,35(02):6-11+42.

[3] Yang Xiaoni, Zhang Kaixuan, Yang Honggang, Yu Yuan. Multivariate regression and ARIMA model prediction of urban domestic waste generation in Xi'an[J]. Environmental Sanitation Engineering,2020,28(02):37-41.

[4] Wang Shuang, Wang Haifei. Prediction of Hainan's GDP based on ARIMA model [J]. Foreign Economic and Trade, 2020(04):44-46.

[5] Wei Qin, Chen Shijun, Huang Weibin, Ma Guangwen, Tao Chunhua. Prediction Method of Spot Market Clearing Price Using Random Forest Regression [J/OL]. Chinese Journal of Electrical Engineering: 1-10.

[6] Yuan Bo, Liu Shi, Jiang Lianxun, Wu Meixuan, Liang Junhua, Tong Xuming. Prediction Model of Housing Rent Based on Random Forest Regression Algorithm [J]. Computer Programming Skills and Maintenance, 2020(01):23-25.