Stock market price prediction model based on grey prediction and ARIMA

Ruoqi Yang#, Yichao Deng#, Qiao Zhong#, Shouxin Zong*, #

Beijing Sport University, School of Sports Engineering, Beijing

*Corresponding author: zsx@bsu.edu.cn

#These authors contributed equally.

Abstract. Nowadays more and more people like to invest in volatile assets, and it is the goal of every market trader to maximize the total return by developing a reasonable investment strategy. We first predicted the daily value of gold and bitcoin for five years based on known data, we built two models, one is Improved Metabolic Gray Model (Abbreviated as IGM), the other is Time Series Model ARIMA. The application of the model helps investors make investment decisions and improve economic returns.

Keywords: Gray Model; ARIMA, stock market price; Investment forecast.

1. Introduction

Market traders buy and sell volatile assets frequently, with a goal to maximize their total return. There is usually a commission for each purchase and sale. Two such assets are gold and bitcoin.Gold, as a general equivalent, is a constant in asset allocation. Gold has been widely used throughout the world as money,[1] for efficient indirect exchange (versus barter), and to store wealth in hoards. For exchange purposes, mints produce standardized gold bullion coins, bars and other units of fixed weight and purity.

Bitcoin is a decentralization-based, peer-to-peer network and consensus initiative, open source code, with blockchain as the underlying technology cryptocurrency[2]. Bitcoin is an innovative payment network and a new kind of money. And it is the first decentralized digital currency. They are transferred directly from person to person via net without going through a bank or clearing house. The transaction fees are much lower and you can use them in every country.From its initial unpopularity to its worldwide recognition, Bitcoin's high returns come with high risks. They have a high volatility [3-4].

2. Improved Metabolic Gray Prediction

2.1 GM (1,1)

GM (1, 1), pronounced as “Grey Model First Order One Variable.”, is a time series forecasting model, which is able to make accurate predictions for forecasting of the monotonous type of processes. Details are as follows:

First set the original sequence

$$\boldsymbol{x}^{(0)}=\{\boldsymbol{x}^{(0)}(1),\boldsymbol{x}^{(0)}(2),...,\boldsymbol{x}^{(0)}(\boldsymbol{N})\} \tag{1}$$

then calculate

$$\boldsymbol{x}^{(1)}(\boldsymbol{i})=\left\{\sum\boldsymbol{x}^{(0)}(\boldsymbol{j})\ \mid\ \boldsymbol{i}=1,2,...,\boldsymbol{N}\right\} \tag{2}$$

get

$$\boldsymbol{x}^{(1)}=\{\boldsymbol{x}^{(1)}(1),\boldsymbol{x}^{(1)}(2),...,\boldsymbol{x}^{(1)}(\boldsymbol{N})\} \tag{3}$$

let $x^{(1)}$ satisfying first-order homogeneous differential equation

$$\frac{dx^{(1)}}{dt}+ax^{(1)}=u \tag{4}$$

the satisfaction condition of this differential equation is

$$\text{when } t=t_0,\ x^{(1)}=x^{(1)}(t_0) \tag{5}$$

the solution of the equation is

$$x^{(1)}(t)\left[x^{(1)}(t_0)-\frac{u}{a}\right]e^{-a(t-t_0)}+\frac{u}{a} \tag{6}$$

the discrete values sampled for equal intervals are

$$x^{(1)}(k+1)=\left[x^{(1)}(t_0)-\frac{u}{a}\right]e^{-ak}+\frac{u}{a} \tag{7}$$

when $\Delta t=1$,

$$x^{(1)}(t)-x^{(1)}(t-1)=\frac{\Delta x^{(1)}(t)}{\Delta t}=x^{(0)}(t) \tag{8}$$

the first-order homogeneous differential equation can be transformed into

$$x^{(0)}(t)=-ax^{(1)}(t)+u \tag{9}$$

that is

$$\begin{cases} x^{(0)}(2)=[-x^{(1)}(2),1]\begin{bmatrix}a\\u\end{bmatrix} \\ x^{(0)}(3)=[-x^{(1)}(3),1]\begin{bmatrix}a\\u\end{bmatrix} \\ \qquad \dots \\ x^{(0)}(N)=[-x^{(1)}(N),1]\begin{bmatrix}a\\u\end{bmatrix} \end{cases} \tag{10}$$

correct x by taking the mean value of the cumulative result

$$x^{(1)}(i)=\frac{1}{2}[x^1(i)+x^1(i-1)]\,(i=2,3,\dots,N) \tag{11}$$

and then get

$$\begin{bmatrix}x^{(0)}(2)\\x^{(0)}(3)\\\dots\\x^{(0)}(N)\end{bmatrix}=\begin{bmatrix}-\frac{1}{2}[x^{(1)}(2)+x^{(1)}(1)] & 1\\ -\frac{1}{2}[x^{(1)}(3)+x^{(1)}(2)] & 1\\ \dots & \dots\\ -\frac{1}{2}[x^{(1)}(N)+x^{(1)}(N-1)] & 1\end{bmatrix}\begin{bmatrix}a\\u\end{bmatrix} \tag{12}$$

the matrix forms is

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{U} \tag{13}$$

estimate by least-square method and then get

$$\hat{\boldsymbol{U}} = \begin{bmatrix} \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{y} \tag{14}$$

bring the estimated values into the differential equation gives yields

$$\hat{\boldsymbol{x}}^{(1)}(\boldsymbol{k}+1) = \left[\boldsymbol{x}^{(1)}(1) - \frac{\hat{\boldsymbol{u}}}{\hat{\boldsymbol{a}}}\right]\boldsymbol{e}^{-\hat{a}k} + \frac{\hat{\boldsymbol{u}}}{\hat{\boldsymbol{a}}} \tag{15}$$

when $\boldsymbol{k} = 1, 2, \ldots \boldsymbol{N} - 1$, the solution is fitted value, when $\boldsymbol{k} \geqslant \boldsymbol{N}$, the solution is predicted value.

2.2 Pretest

Whether a a given sequence $\boldsymbol{x}^{(0)}$ can built GM(1,1) model with high accuracy is generally determined by the class ratio $\boldsymbol{\sigma}^{(0)}(\boldsymbol{k})$ of $\boldsymbol{x}^{(0)}$. The class ratio need to meet certain intervals.

$$\boldsymbol{x}^{(0)} = \left(^{(0)}(1), \boldsymbol{x}^{(0)}(2), \ldots, \boldsymbol{x}^{(0)}(n)\right) \tag{16}$$

The definition of class ration is

$$\boldsymbol{\sigma}^{(0)}(\boldsymbol{k}) = \frac{\boldsymbol{x}^{(0)}(\boldsymbol{k}-1)}{\boldsymbol{x}^{(0)}(\boldsymbol{k})} \tag{17}$$

If

$$\boldsymbol{\sigma}^{(0)}(\boldsymbol{k}) \in \left(\boldsymbol{e}^{-\frac{2}{n+1}}, \boldsymbol{e}^{\frac{2}{n+1}}\right) \tag{18}$$

then $\boldsymbol{x}^{(0)}$ can built GM(1,1) mode.

Considering the volatility of asset values and the small amount of data just starting to forecast, we let $n = 4$, and we find $\boldsymbol{\sigma}^{(0)}(\boldsymbol{k}) \in (0.67, 1.49)$. Conditions are met to the model.

2.3 Process and result

We realize the above modeling process by programming. We improve the model by carefully considering the cycles of asset volatility,we bring $\boldsymbol{\Delta t} = 1$, $\boldsymbol{\Delta t} = 2$, $\boldsymbol{\Delta t} = 3$, $\boldsymbol{\Delta t} = 4$ into the equation separately and use GM(1,1) to predict the data for five years[5]. We find that the best result is bring $\boldsymbol{\Delta t} = 1$, $\boldsymbol{\Delta t} = 2$ into the equation separately and average the two sets of predicted values obtained. As shown in figure1 ,the prediction is very accurate[6].

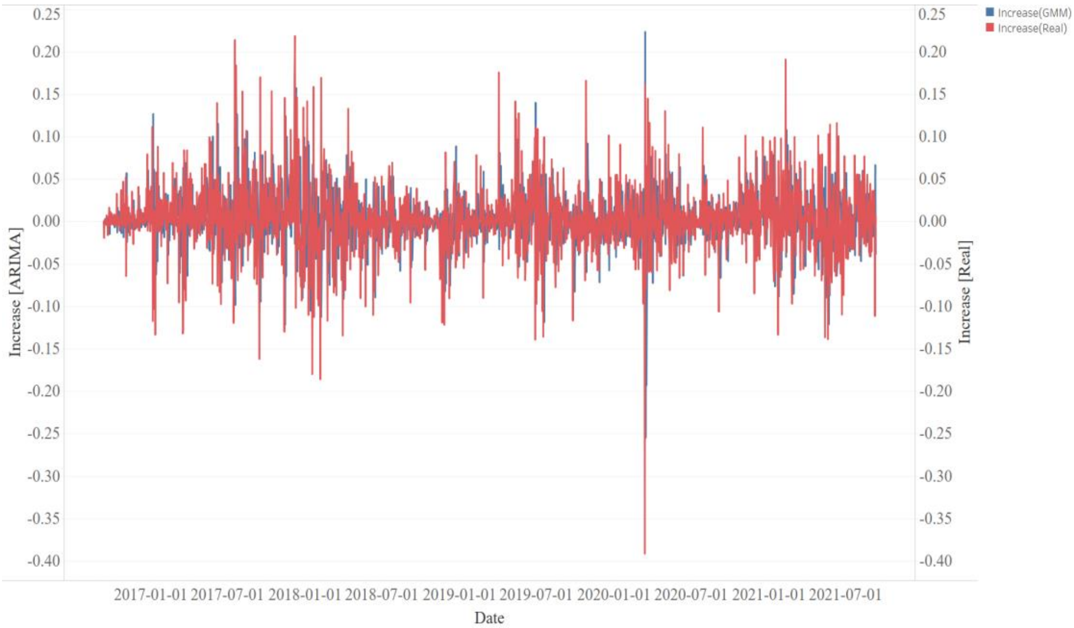

Figure 1. Increase

2.4 Post-test and model evaluation

We test the accuracy of magic by residual test.

First calculate $\hat{x}^{(1)}(i)$

Then do cumulative reduction calculation

$$\begin{cases} \hat{x}^{(0)}(i) = \hat{x}^{(1)}(i) - \hat{x}^{(1)}(i-1)\ (i=1,2,\ldots,n) \\ \hat{x}^{(0)}(1) = \hat{x}^{(1)}(1) \end{cases} \tag{19}$$

And then calculate absolute error (A.E.) and relative Error (R.E.)

$$\varepsilon^{(0)}(i) = x^{(0)}(i) - \hat{x}^{(0)}(i) \quad (i=1,2,\ldots,n) \text{ A.E.} \tag{20}$$

$$\Omega^{(0)}(i) = [\varepsilon^{(0)}(i)/x^{(0)}(i)] \times 100\% \quad (i=1,2,\ldots,n) \text{ R.E.} \tag{21}$$

We find that almost all of the $|\Omega|$ are less than 20%, which are reasonable, except for some values. We call them ‘Abnormal Values’ and list them in the following table. All the abnormal values come from bitcoin predictions[7].

Table 1. Abnormal Values of bitcoin

Date	Real_Value	Pre_Value	$\lvert\Omega\rvert$
2020/3/13	4830.21	7694.2	64.02%
2017/12/9	15142.83415	19960.82	26.37%
2018/2/7	8099.958333	5827.29	26.19%
2017/9/16	3763.62604	3093.97	25.63%
2019/6/28	11132.85	13277.83	25.36%
2020/3/15	5166.26	4041.67	23.78%
2018/1/17	11116.94667	13517.49	21.55%
2017/7/18	2320.12225	1827.18	21.46%
2018/2/5	6838.816667	7793.67	20.38%

By observing the table, we find that there is a very large error in the predicted value, on March 13, 2020 bitcoin plummeted from 7936.65 on the 12th to 4830.21, an increase of -39.1404434%. By checking the information, we know the plunge on that day was due to the outbreak of the global new coronavirus epidemic . In addition, many investors' disappointment in the market accelerated after US President Trump announced on Wednesday night that travelers from 26 "Schengen Convention countries" would be restricted from entering the United States for 30 days from Friday (13th), in addition to the United Kingdom and Ireland[8].The plunge on the 13th also had an impact on the forecast value on the 15th, which was low.

We know that bitcoin is very volatile and various factors can affect its value, the influencing factors are twofold: market and government macro-regulation. The outbreak of the black swan event of the new crown epidemic, the U.S. government's implementation of entry controls, and the greatly reduced market demand caused bitcoin to plummet. There are also some other reasons, such as the February 5, 2018, Global regulators banning, UK and US bankers banning the use of credit cards to buy bitcoin, etc. have caused a lot of concern in the market. The same lead to a high forecast[9].

The virtual currency exchanges FireCoin and OKCoin Coin announced that "only RMB trading business will be stopped, but the rest of the business will not be affected" on September15,2017. Some people see this as a way to continue coin trading on the platform or to provide information aggregation of individual-to-individual virtual currency trading business, so the market demand is somewhat higher and the forecast is low[10].

Large deviations are due to the objective reasons, which cause market changes or government regulation, and bitcoin plunged after a period of growth or rebounded after a period of decline. In addition, the abnormal value only accounted for 0.4939627%, which shows that the GM(1,1) model we built is relatively accurate in its forecast.

3. Time Series Model

We then use Time Series Model ARIMA[5] to predict the five years' daily prices of the two assets.

3.1 ARIMA

ARIMA models are applied in some cases where data show evidence of non-stationarity in the sense of mean (but not variance/autocovariance), where an initial differential step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationary of the mean function (i.e., the trend).[6]The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past.[7] The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differential process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.Our modeling process is as follows.

3.2 Process

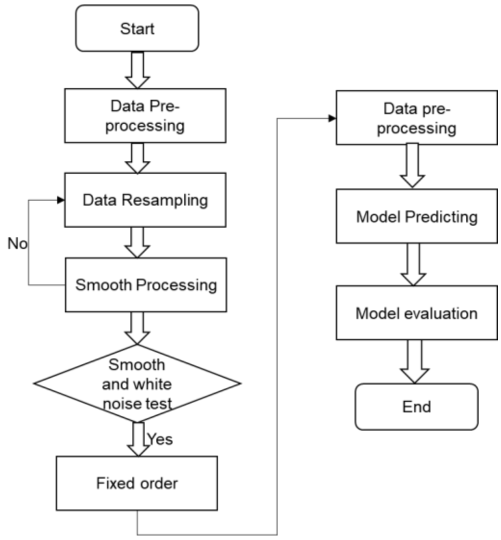

Figure 2. ARIMA algorithm

Here are the ARIMA forecasts for gold. For the value of gold on non-trading dates, we choose to skip.Firstly we perform the first-order difference operation and the second-order difference operation.The difference between the first-order difference result and the second-order is not much, and if we use the second-order difference there will be over fitting, so we choose the result of the first-order difference processing.

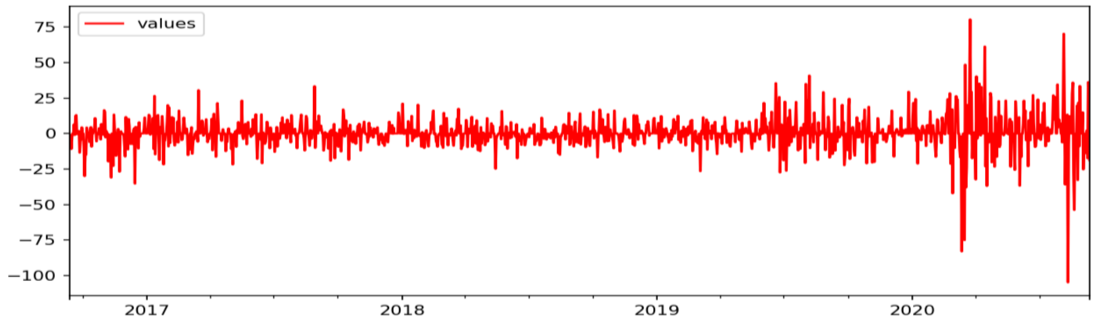

Figure 3. First order differential results

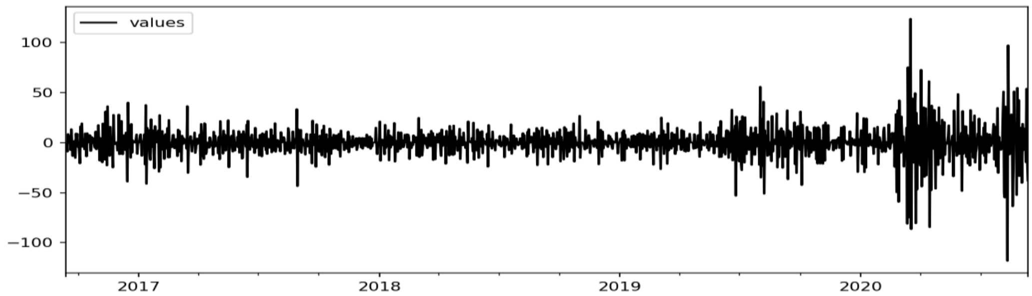

Figure 4. Second order differential results

Then we use smoothing method to process the data, but the results are not good.

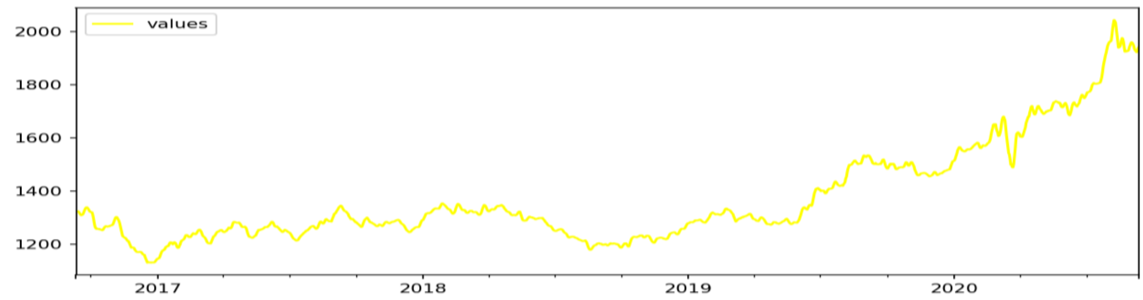

Figure 5. Rolling Mean

We use ADF test, the results indicate that the series is smooth. Next is the white noise test.

The white noise test result is(-8.518862480532741, 1.1147135643005494e-13, 18, 1442, {'1%': -3.4348929812602784, '5%': -2.8635464184851 67, '10%': - 2.5678382024888378}, 10801.463384932884), comparison of statistical values of 1%, %5, %10 different degrees of rejection of the original hypothesis and ADF Test result, ADF Test result less than 1%, 5%, 10% at the same time that means very good rejection of the hypothesis, in this data, the ADF result is -6.9, which is less than the statistical value of three levels.

The p-value of 0.184617 is less than the significance level of 0,05, so the hypothesis can be rejected with 95% confidence level and the series is considered as non-white noise series.

ACF,PACF are determined by using trailing and truncated tails.

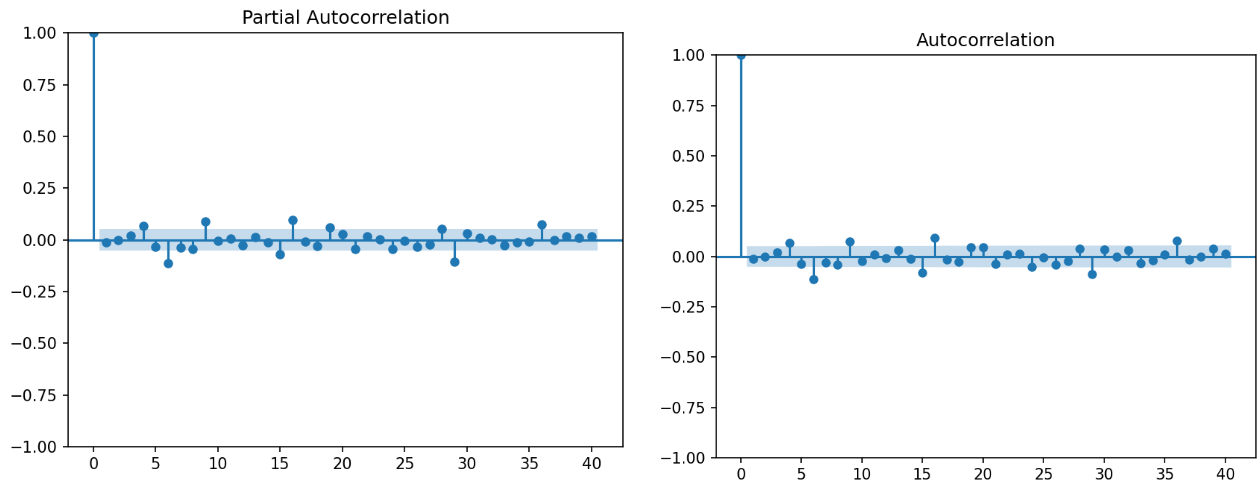

Figure 6. PACF and ACF

We use AIC to build ARMIA, and the parameters is (1,1,7).

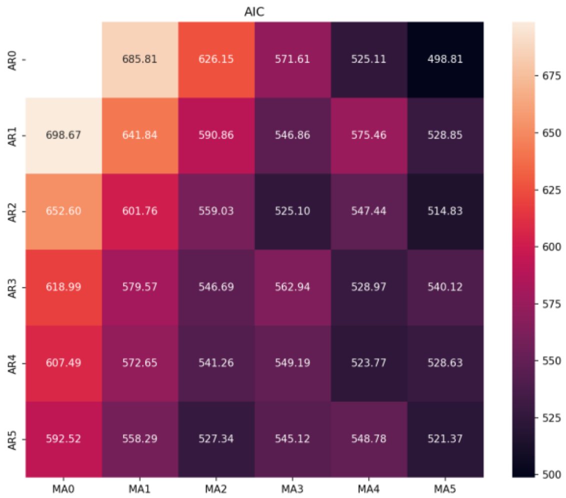

Figure 7. AIC

The prediction process for bitcoin is the same as for gold.We use the model ARIMA(1,1,7) to predict gold, use ARIMA(0,1,10) to predict bitcoin by programming. The results are shown in figure8.

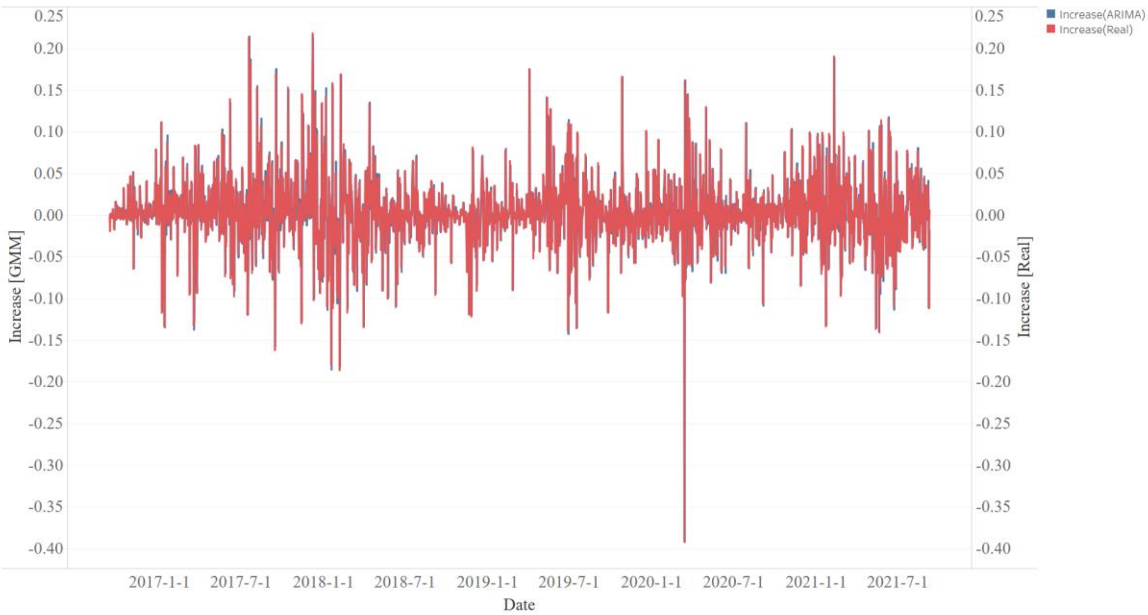

Figure 8. Increase Results

3.3 Model Comparison

Compared to Time Series Model, the Improved Metabolic Gray Model predicts more accurately by comparing the two models' abnormal value percentage.

Table 2. Model comparision

abnormal value percentage	Time Series Model	Improved Metabolic Gray Model
gold	0	0
bitcoin	4.1642897889332%	0.4939627%

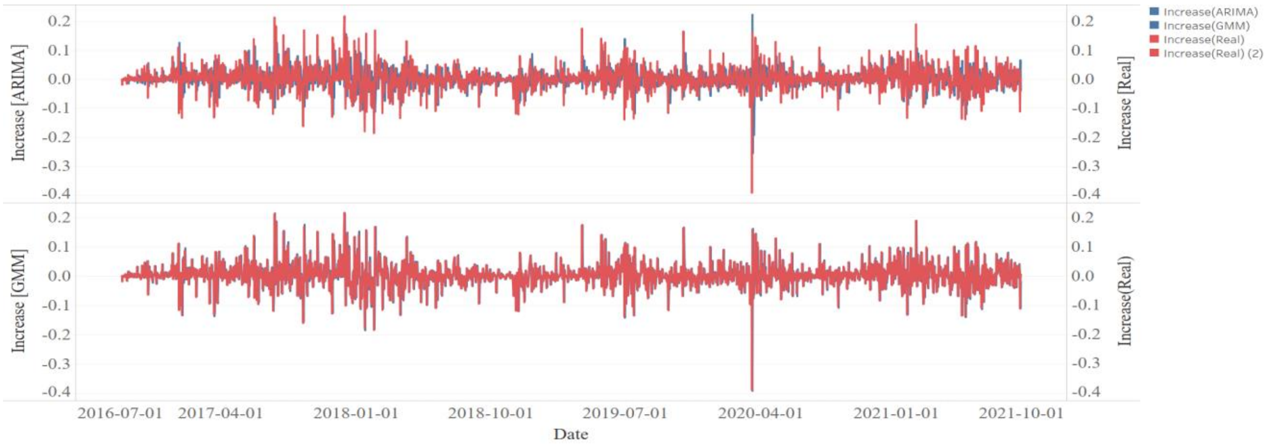

Figure 9. Comparison

4. Conclusion

Investment in volatile assets requires relevant practitioners to formulate reasonable investment strategies to maximize the total return. In view of the current needs of financial investors to expand their income, this paper establishes an improved metabolic grey model and ARIMA model based on the known data to predict the daily value of gold and bitcoin in five years. Compared to Time Series Model, the Improved Metabolic Gray Model predicts more accurately by comparing the two models' abnormal value percentage. The application of the model helps investors make correct investment decisions and improve economic benefits.

References

[1] Li Hui,Liu Yunmei,Luo Xilin,Duan Huiming. A novel nonlinear multivariable Verhulst grey prediction model: A case study of oil consumption forecasting in China[J]. Energy Reports,2022,8.

[2] Bouker, Sawsen,Mansouri, Faysal. Correction to: Sovereign contagion risk measure across financial markets in the eurozone: a bivariate copulas and Markov Regime Switching ARMA based approaches[J]. Review of World Economics,2022(prepublish).

[3] Zhang, Jinping,Xiao, Honglin,Fang, Hongyuan. Component-based Reconstruction Prediction of Runoff at Multi-time Scales in the Source Area of the Yellow River Based on the ARMA Model[J]. Water Resources Management,2022(prepublish).

[4] Zang Lizhuo,Xiao Baihui,Ma Jiayi,Liu Yufan,Tian Peiyu,Zhang Lingdi. Research on the cost effectiveness of carbon emission reduction in the full life cycle of electric vehicles based on grey prediction[J]. MATEC Web of Conferences,2022,355.

[5] Duan Huiming,Luo Xilin. A novel multivariable grey prediction model and its application in forecasting coal consumption[J]. ISA Transactions,2022,120.

[6] Javed Saad Ahmed,Cudjoe Dan. A novel grey forecasting of greenhouse gas emissions from four industries of China and India[J]. Sustainable Production and Consumption,2022,29.

[7] Yu Liu,Yingshuo Li. Analysis and prediction of GDP of three major industries based on grey prediction model[J]. Academic Journal of Business & Management,2021,3.0(10.0).

[8] Yaning Xiao,Dongchen Wu. Analysis on the Per Capita Disposable Income of Sichuan Residents Based on Grey Forecasting Extended Model[J]. Academic Journal of Business & Management,2021,3.0(10.0).

[9] Sansa Ines,Boussaada Zina,Bellaaj Najiba Mrabet. Solar Radiation Prediction Using a Novel Hybrid Model of ARMA and NARX[J]. Energies,2021,14(21).

[10] Ding Mingkuan,Shi Zhiyong,Du Binhan,Wang Huaiguang,Han Lanyi,Cheng Zejun. The method of MEMS gyroscope random error compensation based on ARMA[J]. Measurement Science and Technology,2021,32(12).