Global Economic Perspectives Vol 2 Issue 2 2024

Original Research Article

Open Access

Omniscient Pte.Ltd.

Forecasting Fund Purchase Amount Using Arima Model

Rui He*

Department of Sydney Smart Technology College, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei Province, 066004, China

***Correspondence to:** Rui He, Department of Sydney Smart Technology College, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei Province, 066004, China, E-mail: rui.he205@outlook.com

Abstract: Predicting fund purchase amounts is a significant challenge for economists and statisticians. Accurate predictions can lead to favourable outcomes for both companies and individuals. This paper proposes an ARIMA model to extract insights from the data and provide a method for forecasting fund purchase amounts over the coming days. We begin by examining the characteristics and structure of the dataset, calculating various statistical indices. Subsequently, we implement data transformation strategies to address outliers and missing values. Using R software, we analyse the data and develop several candidate ARIMA models. The most suitable model is selected based on the Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF), and the Akaike Information Criterion (AIC). Finally, we evaluate the chosen model using Mean Squared Error (MSE) and various residual analysis plots. The residuals from our model indicate strong performance in predicting fund purchase amounts.

Keywords: ARIMA, data analysis , fund purchase amount, prediction.

1. Introduction

Fund purchase amount prediction is a widely challenging problem for statistics and statisticians. The fund purchase amount refers to the total amount of funds that investors need to pay when purchasing a fund, that is, the amount of funds that investors decide to invest in the fund. The purchase amount of a fund directly reflects the investor's investment willingness, risk tolerance, and financial condition, and is one of the important decisions in the fund investment process. Predicting the amount of fund purchases is of great significance to investors. It not only helps guide investment decisions and optimize asset allocation, but also helps investors reduce investment risks, increase risk awareness, and ultimately enhance long-term return potential. Many economists and statisticians put forward a plenty of methods to deal this problem, and which is also a hot topic in this era. The capital asset pricing model (CAPM) can be used to predict the purchase amount of funds by treating the fund as an asset and estimating its expected return rate, thereby inferring investors' purchase intention and amount[1]. In the prediction of fund purchase amount using regression analysis models, factors that affect purchase amount (such as

market index, investor sentiment, etc.) can be used as independent variables, and purchase amount can be used as the dependent variable for regression analysis. Machine learning models can use neural network models to learn historical data, market data, investor behaviour data, and predict future purchase amounts in fund purchase amount prediction[2]. The random walk model can be used to estimate future purchase amounts in predicting highly random and difficult to predict fund purchase amounts[3]. However, achieving precise predictions and establishing confidence intervals for fund purchase amounts remain significant challenges.

Numerous studies have been conducted to predict fund purchase amounts using various time series prediction algorithms, including Neural Networks, statistical models, and Autoregressive Integrated Moving Average (ARIMA) models[4,5]. One proposed approach involves combining ARIMA with Generalized Autoregressive Conditional Heteroscedasticity (ARIMA-GARCH) for short-term time series that exhibit stationary characteristics. However, this approach did not demonstrate significant improvements over the standard ARIMA model[6,7].

Time series analysis is widely employed for a variety of purposes, including forecasting, event detection, and decision-making. Notably, time series forecasting is regarded as a critical research area within econometrics and operations research[8]. A key factor influencing the outcomes of time series forecasting is the time horizon (or lag), which captures patterns within the series and dictates its autocorrelation with itself[9]. Additionally, the trend component is an important consideration, as it reflects whether the time series displays low, medium, or high frequencies[10,11].

Numerous statistical techniques have been devised for predicting time series data, with the cornerstone being Box and Jenkins' Autoregressive Integrated Moving Average (ARIMA) model[7,12]. ARIMA has been employed across diverse domains to construct forecasting models, including Guha and Bandyopadhyay's prediction of gold price[13], and Fan's forecast of monetary fund. To achieve acceptable accuracy, ARIMA necessitates training on large datasets; otherwise, predictive accuracy may be unacceptably low. The preponderance of research found in the literature focuses on large-scale time series analysis, where abundant data enhances the model learning process, thereby improving forecasting outcomes[14]. Consequently, traditional forecasting methodologies consider small-scale data unsuitable for modelling[5,15]. Moreover, time series forecasting models often assume stationarity, meaning that the standard deviation and mean remain constant over time[16]. In other words, the data does not exhibit seasonal patterns. Our hypothesis is that analyzing non-stationary short-term time series data with the ARIMA model is valuable, as it effectively transforms non-stationary series into stationary ones by determining the necessary number of differencing operations (e.g., $d = 0,1,2,\ldots$) to achieve a stationary sequence. The fund purchase amount data exhibits short-term increasing and decreasing trends, along with several prominent peaks. Therefore, we should apply differencing techniques to convert the data into a stationary format.

This paper investigates and analyses a time series dataset comprising 448 observations from 2021-1-4 to 2022-11-9. We pre-process and prepare the dataset for modelling by employing the "3-sigma rule" to remove 13 outliers. Linear interpolation is utilized to fill in the missing values, and the data is scaled using Student's transformation. Next, we use R software to build the ARIMA model and estimate the unknown parameters. For the model selection procedure, we employ ACF and PACF charts to determine the appropriate orders for the ARIMA model. Additionally, we adopt an automated procedure from the forecast package. Finally, we evaluate the selected model using Mean Squared Error (MSE) and conduct several residual analysis plots.

The organization of the subsequent sections of this paper is outlined as follows. Initially, a concise explanation of the methodologies are provided in Section 2. This is followed by Section 3, which describes and analyses the data. Section 4 focuses on the analysis and identification of the model, including a comparison and determination of the model parameters. Finally, the paper concludes with a comprehensive summary in Section 5.

2. Methodology

In this paper, we utilize the ARIMA model for the task of prediction, as the fund purchase amount represents a financial time series, and ARIMA is a classical statistical tool for modeling such series. This model is grounded in robust mathematical and

statistical theories that facilitate the generation of prediction intervals. The ARIMA model consists of three components: Autoregression (AR), differencing (I), and Moving Average (MA). We denote the model as ARIMA(p, d, q), where these three parameters represent autoregression order, differencing order, and the moving average window size, respectively. To specify these parameters, we first apply one or more lag-1 differencing operations to account for trends or seasonal differences, followed by fitting the ARIMA model to the resulting sequences. We use Equation (1) to determine the ARMA(p, q) components:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

These parameters can be calculated using a typical method that involves visual inspection of the time series to identify trends, as well as examining correlation and partial correlation plots. A challenge in implementing the model on non-normally distributed data is that the value of the dependent fund purchase amount at time t is influenced by past purchase values; thus, neglecting lags can adversely affect the estimation of the standard error of the estimated coefficients. Common information criteria such as AIC and BIC can also be employed to determine the unknown parameters and select the best-fitting model among various candidates. The best model is as simple as possible and minimizes certain criteria, namely AIC, BIC,variance and maximum likelihood[17,18,19]. Additionally, a straightforward approach to this task is to utilize computer programs, such as auto.arima from the forecast package in R and the Econometrics Toolbox in MATLAB.

3. Data description

To study the prediction of fund purchase amounts, we analyse a dataset containing fund purchase prices. This time series, recorded from January 4, 2021, to November 9, 2022, consists of 448 data points, which can be downloaded from the Alibaba website. The mean of the dataset is 0.4595, and the standard error is 0.6481. There are 227 missing values in the dataset. After filling these missing values using linear interpolation, we obtain a total of 675 data points. The data exhibits a clear trend, as illustrated in **Figure 1**.

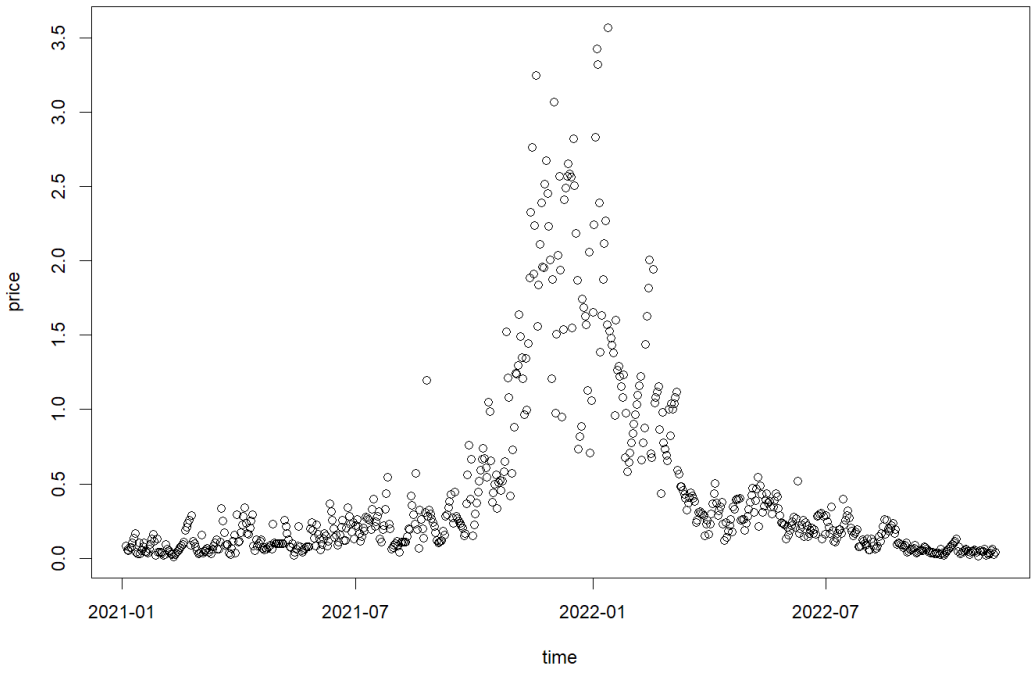


Figure 1. The points plot of a fund product.

From **Figure 1**, it is evident that there is a significant upward trend followed by a decline in the middle of the time series. Therefore, before constructing the ARIMA model, we should eliminate this trend through a differencing operation. The result after differencing is illustrated in **Figure 2**.

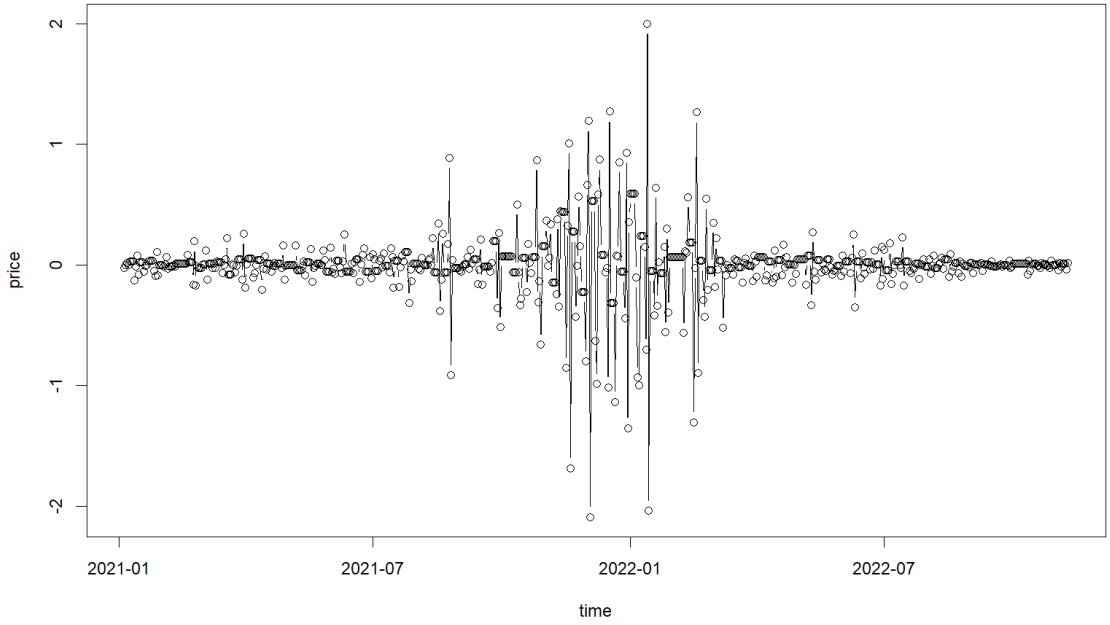


Figure 2. The time series after the differencing operation.

The new time series, as shown in **Figure 2**, oscillates around zero, indicating that there is no apparent trend in the plot. To ensure scientific rigor, we will conduct several statistical tests, including the QQ plot test, Ljung-Box test, and ADF test, to determine whether the differenced data behaves like white noise.

From **Figure 3**, it is observed that to achieve the same probability, the sample quantiles are lower than those of the theoretical normal distribution in the left tail. In contrast, the sample quantiles are larger than those of the theoretical normal distribution for the same probabilities in the right tail. This indicates that the empirical distribution exhibits characteristics of a sharp peak and heavy tail while maintaining a symmetric shape. The results of the Ljung-Box test are presented in **Table 1**.

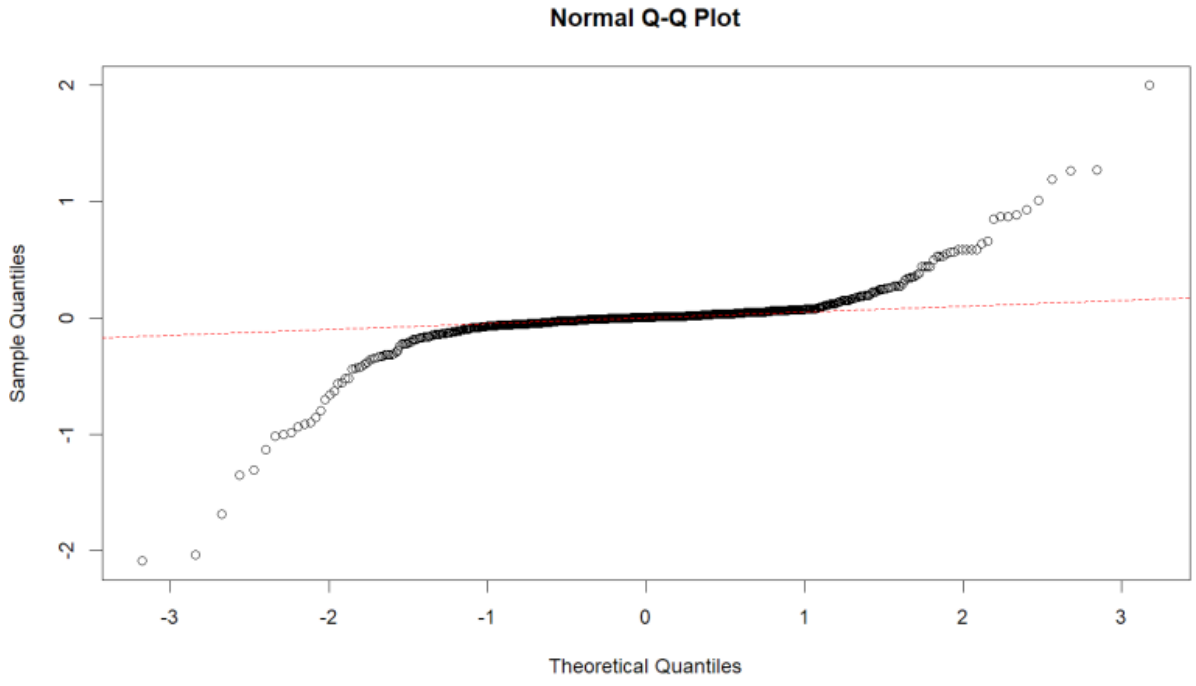


Figure 3. The normal QQ plot.

Table 1. The results of Ljung-Box test.

Lags	2	4	6
χ^2	67.991	68.439	80.797
p-value	1.77×10^{-15}	4.852×10^{-14}	2.442×10^{-15}

From **Table 1**, we observe that all p-values are less than 0.05. Therefore, we conclude that there is a strong relationship among the data points, indicating that this series is not white noise. The ADF test is employed to assess whether the differenced series is stationary. We utilize the *adf.test* function from the *tseries* package in R. The Dickey-Fuller statistic is −12.097 with a lag order of 8, and the corresponding p-value is 0.01. Consequently, we reject the null hypothesis and conclude that the differenced series is stationary.

4. Model building

4.1 Identification of model

Generally, the parameters p and q in ARMA(p, q) is determined by the ACF and PACF plots. Now, the results are shown in **Figure 4**.

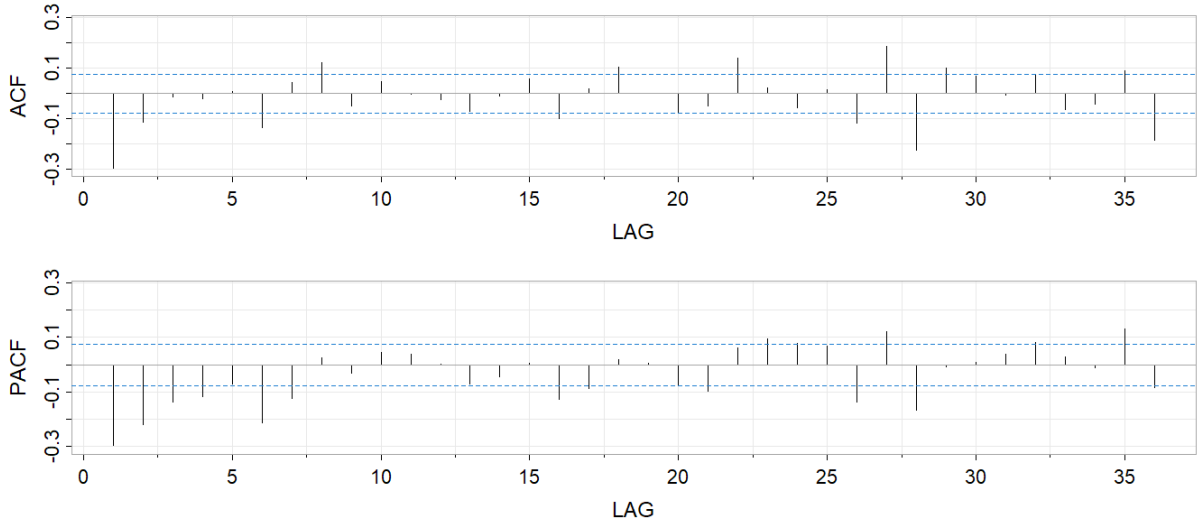


Figure 4. The ACF and PACF plots.

It is evident that neither of the two plots exhibits the phenomenon of truncation, indicating that a single AR(p) or MA(q) model is not appropriate for this series. From PACF, we guest the proper p is less than 5. Similarly, we suggest the proper q is than 3 from ACF. Then, we can try many candidate models such as: ARMA(1, 1), ARMA(2, 1), … , ARMA(5, 1), ARMA(1, 2), ARMA(2, 2), …, ARMA(5, 2), … , ARMA(6, 3) . In this paper, we utilize the AIC criterion to pick up the best working model among these candidate models.

Table 2. The AIC and BIC scores of candidate models.

Model	ARIMA(1, 1, 1)	ARIMA(2, 1, 1)	ARIMA(3, 1, 1)	ARIMA(4, 1, 1)	ARIMA(5, 1, 1)
AIC	44.722	46.700	48.509	50.120	51.339
Model	ARIMA(1, 1, 2)	ARIMA(2, 1, 2)	ARIMA(3, 1, 2)	ARIMA(4, 1, 2)	ARIMA(5, 1, 2)
AIC	46.698	48.372	49.901	51.787	53.786
Model	ARIMA(1, 1, 3)	ARIMA(2, 1, 3)	ARIMA(3, 1, 3)	ARIMA(4, 1, 3)	ARIMA(5, 1, 3)
AIC	48.598	49.959	50.100	53.784	48.374

After calculating the AIC scores of all candidate models, we can pick up the best model as ARIMA (1, 1, 1) with a smallest AIC score. To very our choice, we also use the function *auto.arima* from *forecast* package in R to find a good working model automatically. The result of this function tells

ARIMA(1, 1, 1) is the best one when $p \leqslant 6$ and $q \leqslant 4$. Calculating again, we obtain the parameter estimates of model ARIMA(1, 1, 1):

Table 3. The parameter estimates of ARIMA(1, 1, 1).

Parameter	ϕ_1	θ_1
Estimate	0.3746	-0.8236
Standard Error	0.0555	0.0343

Hence, as shown in **Table 3**, we can write the best candidate model as:

$$z_t = 0.3746 z_{t-1} - 0.8236 \epsilon_{t-1} + \epsilon_t \quad (2)$$

$$z_t = y_t - y_{t-1} \quad (3)$$

4.2 Prediction

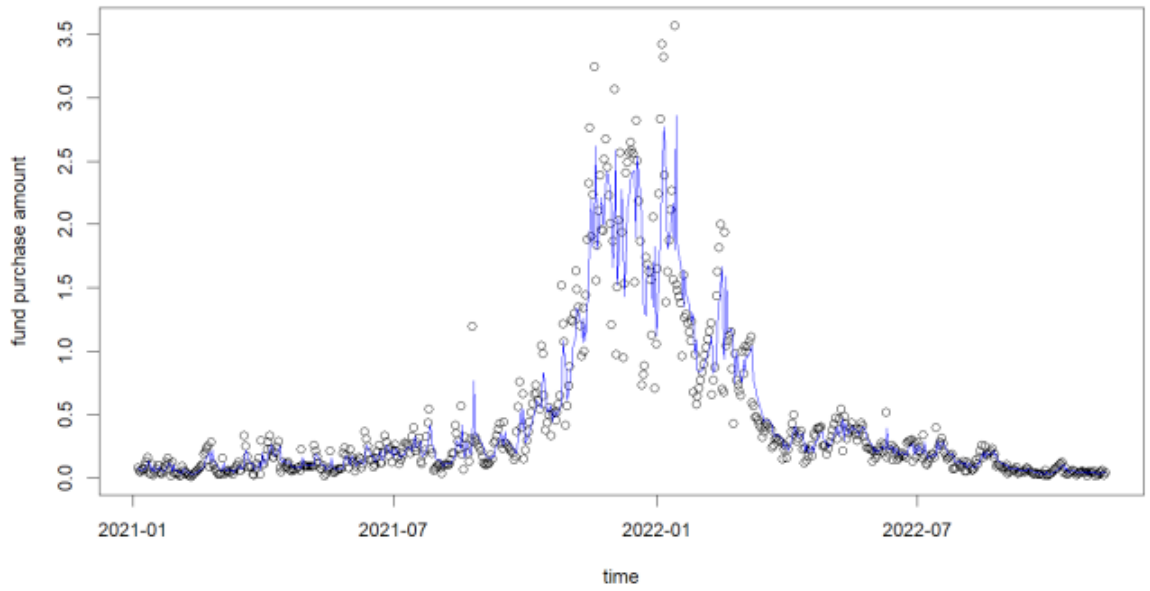
After selecting the best candidate model, we can predict the fund purchase amounts for the upcoming time points using the ARIMA(1, 1, 1) model. However, a point estimate alone is insufficient for making informed economic decisions. Therefore, we also consider interval predictions at a significance level of 0.05. The results of these predictions are presented in **Table 4**.

Table 4. The results of prediction.

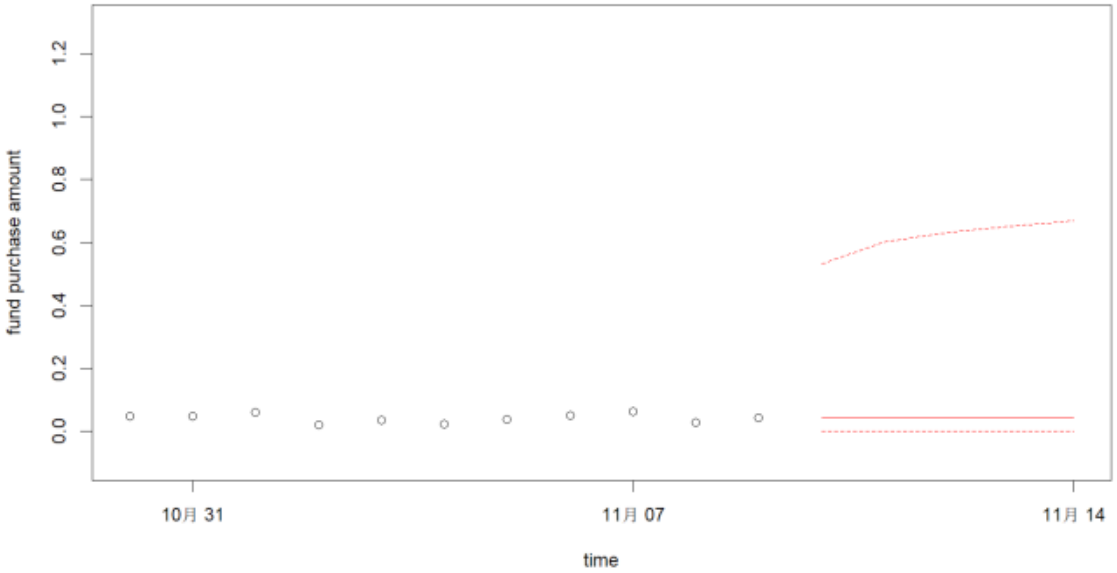
Time	2022-11-10	2022-11-11	2022-11-12	2022-11-13	2022-11-14
Prediction	0.0422	0.0419	0.0418	0.0418	0.0418
Prediction Interval	[0, 0.5315]	[0, 0.6006]	[0, 0.6310]	[0, 0.6514]	[0, 0.6684]

4.3 Evaluation

In this section, we aim to evaluate the performance of the ARIMA(1, 1, 1) model from several perspectives. First, the fit for historical data is illustrated in **Figure 5(a)**, while the prediction performance is depicted in **Figure 5(b)**. In this figure, the blue solid line represents the fitted values for the historical data, and the red solid line denotes the forecast for the next five days. Additionally, the two red dotted lines indicate the upper and lower bounds of the prediction interval at a significance level of 0.05.



(a). The performance of fitness.



(b). The performance of prediction.

Figure 5. The fitness and prediction interval of ARIMA(1,1,1).

Next, to evaluate the performance of the working model, Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are two widely used indicators. It is an empirical rule that a good prediction is expected if the fit for historical data is also satisfactory. Specifically, MAE and MAPE are defined as follows:

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T}\left|\widehat{y_i} - y_i\right| \tag{4}$$

$$\mathrm{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\widehat{y_i} - y_i}{y_i}\right| \times 100 \tag{5}$$

The results of MAE and MAPE for the top three models of **Table 2** are shown in **Table 5**, from which the picked model owns smallest MAPE 37.62% and the penultimate MAE 0.1226.

Table 5. Measure accuracy errors.

Model	ARIMA(1, 1, 1)	ARIMA(1, 1, 2)	ARIMA(5, 1, 3)
MAE	0.1226	0.1227	0.1221
MAPE	37.62%	37.62%	37.88%

Table 6. The Ljung-Box test for residuals.

Lags	2	4	6
χ^2	0.0052	0.3890	7.995
p-value	0.9974	0.9834	0.2385

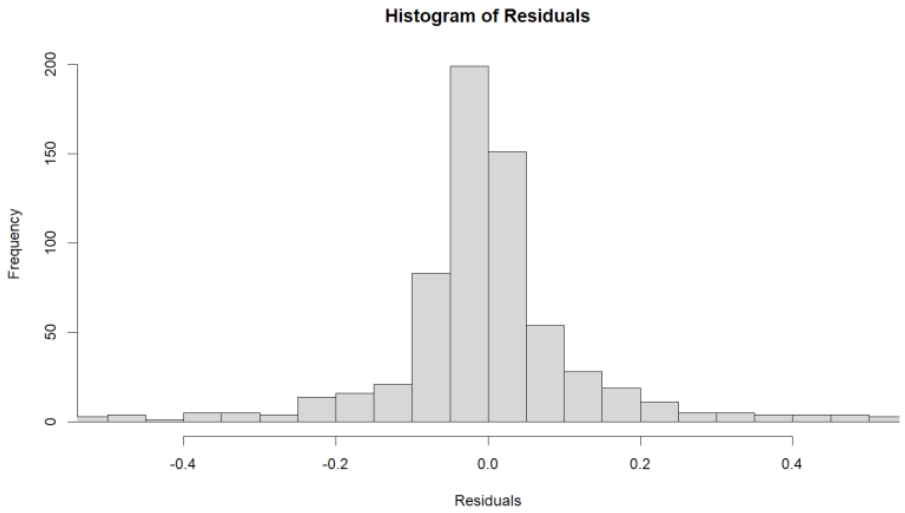


Figure 6. The histogram of residuals.

Residual analysis is a crucial tool for determining whether there is still information within the data that can be further explored. The histogram of the residuals is illustrated in **Figure 6**, where we observe a clear central tendency towards zero, resembling the bell curve of a normal distribution. Additionally, we perform the Ljung-Box test on the residuals, with the results presented in **Table 6**. Compared to **Table 1**, the results in **Table 6** indicate that all tests are not significant, as the *p*-values are less than 0.05. This suggests that there is no mutual effect among the series after modelling. Finally, we also conduct the ADF test on these residuals, yielding a Dickey-Fuller statistic of −7.8527 with a lag order of 8, and a corresponding *p*-value of 0.01. Therefore, we reject the null hypothesis and conclude that the series after differencing is stationary, indicating that there is no need to investigate other potential trends within the series.

5. Conclusion

This paper addresses the issue of predicting fund purchase amounts by employing the ARIMA model to extract information from the time series and provide interval predictions for the upcoming days. Various data preparation strategies are utilized to handle outliers and missing values. Additionally, we use R software to analyze and construct the ARIMA model. Specifically, we evaluate the selected model using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and other residual analyses, which indicate that the ARIMA(1, 1, 1) model demonstrates a good fit for the historical data and offers reasonable interval predictions.

In the future, we plan to extend our analysis of fund purchase amounts by developing models that account for change points and the interactive influences among other fund purchase series, such as GARCH, VAR, and other more complex models. Regarding interval prediction, we aim to explore the conformal prediction method, which does not rely on assumptions about the working model.

References

[1] Ross, S. A. (1978) The current status of the capital asset pricing model (CAPM). The Journal of Finance, 33(3): 885-901.

[2] Mishra, A. K., Bansal, R., Maurya, P. K., Kar, S. K., & Bakshi, P. K. (2023). Predicting the antecedents of consumers' intention toward purchase of mutual funds: A hybrid PLS-SEM-neural network approach. *International Journal of Consumer Studies*, 47(2), 563-587.

[3] Fama, E. F. (1995). Random walks in stock market prices. *Financial analysts journal*, 51(1), 75-80.

[4] K. Li, C. Zhai, and J. Xu. (2017) Short-term traffic flow prediction using a methodology based on ARIMA and RBF-ANN. In *2017 Chinese Automation Congress (CAC)*, 2804–2807.

[5] H. Dong, L. Jia, X. Sun, C. Li, and Y. Qin. (2009) Road Traffic Flow Prediction with a Time-Oriented ARIMA Model. In *2009 Fifth International Joint Conference on INC, IMS and IDC*, 1649–1652.

[6] B. Zhou, D. He, and Z. Sun. (2006) Traffic predictability based on ARIMA/GARCH model," in *2006 2nd Conference on Next Generation Internet Design and Engineering*, 206-207.

[7] C. Chen, J. Hu, Q. Meng, and Y. Zhang. (2011) Short-time traffic flow prediction with ARIMA-GARCH model. In 2011 IEEE Intelligent Vehicles Symposium (IV), 607–612.

[8] 25 years of time series forecasting - ScienceDirect. [Online]. Available: https://www-sciencedirect-com.uproxy.library.dcuoit.ca/science/article/pii/S0169207006000021.

[9] C. W. Ostrom. (1990) Time series analysis: regression techniques. 2nd ed. Newbury Park, Calif, Sage Publications.

[10] Trend Modeling for Traffic Time Series Analysis: An Integrated Study - IEEE Journals & Magazine. [Online]. Available: https://ieeexplore-ieee-org.uproxy.library.dcuoit.ca/document/7180371. [Accessed: 20-Mar-2019].

[11] Trend analysis of climate time series: A review of methods ScienceDirect. [Online]. Available: https://wwwsciencedirect-com.uproxy.library.dcuoit.ca/science/article/pii/S0012825218303726.

[12] Comparison of two new ARIMA-ANN and ARIMA-Kalman hybrid methods for wind speed prediction - ScienceDirect.[Online]. Available: https://www.ciencedirectcom.uproxy.library.dcuoit.ca/science/article/pii/S0306261912002875.

[13] Guha, B., & Bandyopadhyay, G. (2016). Gold price forecasting using ARIMA model. Journal of Advanced Management Science, 4(2):117-121.

[14] T. Fu. (2011) A review on time series data mining. Eng. Appl. ArtifIntell., 24(1): 164–181.

[15] K. Duangnate and J. W. Mjelde. (2017) Comparison of data-rich and small-scale data time series models generating probabilistic forecasts: An application to U.S. natural gas gross withdrawals. Energy Econ., 65: 411–423.

[16] A. Adib, M. M. K. Kalaee, M. M. Shoushtari, and K. Khalili. (2017) Using of gene expression programming and climatic data for forecasting flow discharge by considering trend, normality, and stationarity analysis. Arab. J. Geosci., 10(9): 207-208.

[17] Box G and Jenkins G. (1994)Time series analysis, forecasting and control, 3rd ed. San Francisco: Holden-Day.

[18] Brockwell PJ and Davis RA. (1987)Time series: theory and method. Berlin: Springer-Verlag.

[19] Hamilton JD. (1994)Time series analysis. Princeton: Princeton University Press.