International Journal of World Economic Research
ISSN: 3080-9886 (Print), ISSN: 3080-9894 (Online) | Vol. 2, No. 2, 2026

Stock Analysis Based on the ARIMA Model

Sichen Zhou

School of Nanjing University of Science and Technology 210000, China

Abstract: This study applies the ARIMA model to analyze and forecast the daily opening price data of Industrial and Commercial Bank of China (stock code: 601398.SH), spanning from January 2019 to December 2024, with a total of 1,456 observations. The Augmented Dickey-Fuller (ADF) test confirms that the original series is non-stationary but becomes stationary after first-order differencing, establishing it as an I(1) process. The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced series exhibit a clear cutoff at lag 1, guiding model identification. Among 16 candidate ARIMA models evaluated using AIC, BIC, and HQIC criteria, the ARIMA (1, 1, 2) model is selected as optimal, achieving the lowest values across all three information criteria. Diagnostic checks, including the Ljung-Box Q-test at various lag orders (5, 10, 15, 18), indicate that residuals are white noise, confirming adequate model specification. The model demonstrates strong in-sample fit with an R-squared of 0.9145. Forecast evaluations over short-term (10 days), medium-term (50 days), and long-term (116 days) horizons yield MAE/RMSE values of 0.07/0.09, 0.05/0.06, and 0.08/0.11, respectively, indicating robust predictive performance across different horizons. The findings suggest that the ARIMA (1, 1, 2) model effectively captures the underlying dynamics of stock opening prices and provides reliable forecasts, offering valuable decision-making support for investors in volatile financial markets.

Keywords: ARIMA model, Time series forecasting, Stock price prediction, Ljung-Box test.

1. Introduction

In a society with rapid economic development, most business and financial activities tend to use regression analysis and other methods to make future predictions. Admittedly, the results are often unsatisfactory, as large errors and uncertain factors cast a veil over the prediction outcomes of regression analysis. The ARIMA model we have selected, however, possesses the characteristic of exploring time series, enabling it to analyze economic phenomena while considering the fluctuating effects of time series. In this way, errors can be significantly reduced, providing valuable analytical reference for the volatile stock market.

Currently, an increasing number of people choose to invest in stocks to obtain economic benefits. To achieve substantial financial returns, investors study selected stocks or predict their trends before making decisions. Stock price data represent a typical type of financial time series, and using historical price data to establish prediction models for forecasting future stock prices has become common practice [1]. Numerous models and methods have emerged in the study of financial time series, yielding reliable conclusions that provide references for investors' decision-making. Wang Ying [2] took the stock of Bank of China as a case study, applied the ARIMA model to conduct short-term analysis and forecasting of stock opening prices, and found that the predicted values were close to the actual values with relatively small errors. Liu Jie [3] employed the ARMA model to forecast the stock price of Gree Electric Appliances, and the model demonstrated favorable performance in short-term forecasting. Wu Yuxia et al. [4] conducted short-term predictions of the closing price of "Huatai Securities" based on the ARIMA model and concluded that the ARIMA model exhibits strong performance in both short-term dynamic and static forecasting.

As more people, including students, gain knowledge of financial planning, there is a growing demand for selecting financial products such as stocks, bonds, and funds. However, the financial industry is a relatively complex system where even a slight operational mistake can easily lead to significant losses. At this point, we can leverage big data analysis techniques to conduct analyses. With technological advancement, the term "big data" has come increasingly closer to us. There are also many data mining software tools such as R and Python, all of which help us perform better data analysis. In this report, Python is selected for data analysis.

As a crucial component of financial markets that is complex and influenced by various interacting uncertain factors, the analysis and prediction of stock prices are of paramount importance. Although stock price fluctuations are often affected by macroeconomic information, non-recurring gains and losses, human factors, and others, time factor serves as the most fundamental influencing variable. Therefore, exploring suitable time series fitting models is also very important. Thus, this paper selects the Industrial and Commercial Bank of China (ICBC) stock, which embodies comprehensive value, for reasonable analysis and subsequent prediction. The data were obtained from the Wind database with time period 41, stock code "601398.SH" (ICBC). A total of 1,456 daily opening price data points from January 2019 to December 2024 are used for time series regression fitting.

2. Organization of the Text

2.1. Descriptive Statistics

Table 1. Descriptive statistics

Obs	Mean	Std.Dev.	Min	Max
1456	5.1107	0.5410	4.05	6.94

Based on these descriptive statistics results, the following conclusions can be drawn: Among the 1,456 observation samples, the data has a mean of 5.1107 and a standard deviation of 0.5410, indicating that the data generally fluctuates around 5.11, with relatively small dispersion and a relatively concentrated distribution. The data has a minimum value of 4.05 and a maximum value of 6.94, with a range of

2.89, indicating that the data covers a range from approximately 4.05 to 6.94, with no extreme outliers. From the perspective of the relationship between the mean and the extreme values, the difference between the minimum and the mean is approximately 1.06, while the difference between the maximum and the mean is approximately 1.83. The data distribution is generally symmetric, but the deviation of the maximum is slightly larger than that of the minimum, suggesting a possible slight right-skewed trend. Overall, this set of data exhibits good central tendency, moderate volatility, and a large sample size, resulting in high stability and representativeness of the statistical findings.

2.2. Stability Test and Pure Randomness Test

Table 2. Stability Test

Difference order	origin	Frist Difference
P-value	0.873235	0.000000

The Augmented Dickey-Fuller (ADF) test was conducted to examine the stationarity of the original time series and its first-differenced series. The test results show that the p-value for the original series is 0.873235, which is significantly greater than the 0.05 significance level. Therefore, the null hypothesis of the existence of a unit root cannot be rejected, indicating that the original series is non-stationary. In contrast, the p-value for the first-differenced series is 0.000000, which is less than 0.05, leading to the rejection of the unit root hypothesis. This confirms that the first-differenced series is stationary. This finding justifies the use of first-order differencing in the ARIMA modeling process to achieve stationarity.

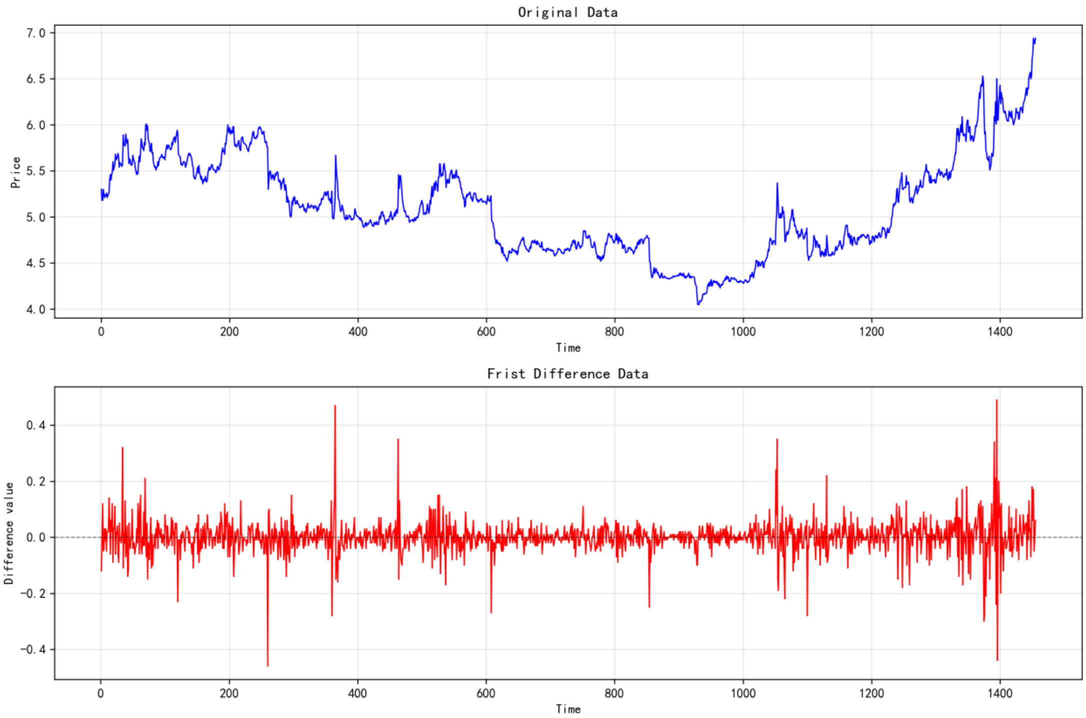

Figure 1. Opening Price and First-Difference Opening Price Time Series Chart

As shown in the original data time series plot (Figure 1), the data sequence exhibits a clear nonlinear downward trend. In the first half (approximately observations 0-400), the data values slowly decline from about 5.2 to around 5.0, with relatively small fluctuations. Entering the middle section (approximately observations 400-800), the decline accelerates noticeably, with data values rapidly falling from about 5.0 to 4.2. In the latter half (approximately observations 800-1400), the downward trend gradually slows and eventually stabilizes at around 4.0-4.1.

Overall, the original data exhibits obvious non-stationary characteristics, with its mean and variance changing over time and demonstrating time dependence, indicating that the series does not satisfy the stationarity assumption and requires differencing to achieve stationarity. In contrast, the first-difference data plot presents a stark contrast to the original data plot. The first-difference values fluctuate around zero, with the vast majority falling within the interval of -0.2 to 0.2, showing no distinct trend or cyclical patterns. The fluctuation range of the difference values remains relatively consistent across different time periods, indicating that the variance of the series is stable. It is worth noting that in the interval of approximately observations 400-800 (corresponding to the fastest decline phase of the original data), negative difference values appear more concentrated, but this does not undermine the stationary characteristics of the series as a whole.

Furthermore, the evident symmetric distribution around the zero line suggests that the series does not have a significant autocorrelation structure. A comparative analysis of the two plots leads to the conclusion that the original data series is a first-order integrated process, meaning it has been successfully transformed into a stationary series after first-order differencing. The first-differenced data have eliminated the trend component of the original series, presenting random fluctuation characteristics, which provides the necessary prerequisite for subsequent analysis and forecasting using time series models such as ARIMA.

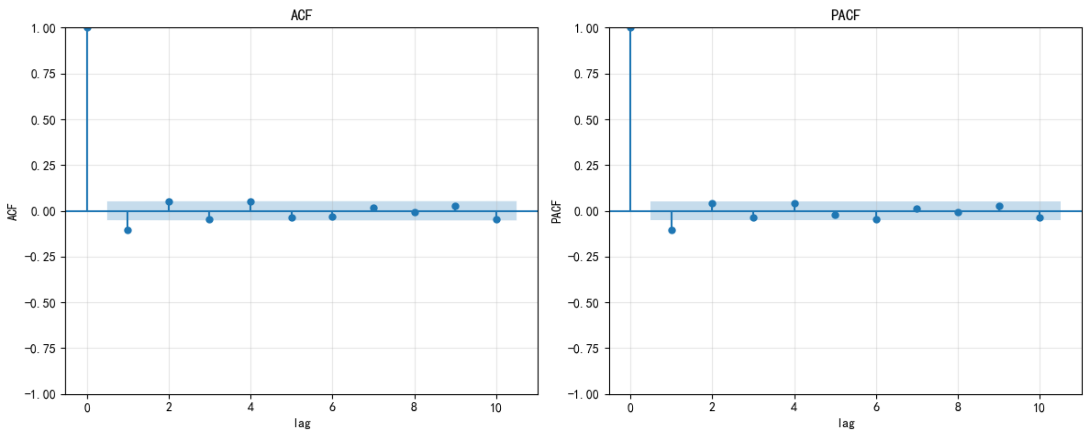

Figure 2. ACF and PACF plots of first-order differenced data

As shown in the autocorrelation function (ACF) plot in Figure 2, the first-differenced series exhibits an autocorrelation coefficient of -0.15 at lag 1, indicating a slight negative correlation with a relatively small absolute value. From lag 2 through lag 10, all ACF values are 0.00. This indicates that the autocorrelation of the series is only weakly present at lag 1 and then rapidly decays to zero, presenting a typical first-order truncation characteristic—that is, the ACF falls within the confidence interval after lag 1. At the same time, none of the ACF values at any lag exhibit periodic fluctuations or slow decay trends, suggesting that the series lacks obvious seasonality or long-memory characteristics.

As for the partial autocorrelation function (PACF), its values are completely consistent with those of the ACF: -0.15 at lag 1, and 0.00 from lag 2 through lag 10. This also exhibits a first-order truncation characteristic, meaning the PACF rapidly drops to zero after lag 1. The rapid decay of the PACF further confirms that the autocorrelation structure of the series is of finite order, with no higher-order partial correlations present.

2.3. Model Testing

Table 3. Model Recognition Results

Model	AIC	BIC	HQIC
ARIMA (1, 1, 2)	-4031.841955	-4010.71366	-4023.958629
ARIMA (2, 1, 3)	-4031.297157	-3999.604715	-4019.472169
ARIMA (3, 1, 3)	-4030.169037	-3993.194522	-4016.373218
ARIMA (2, 1, 2)	-4029.894983	-4003.484615	-4020.040826
ARIMA (3, 1, 1)	-4029.255646	-4002.845277	-4019.401489
ARIMA (2, 1, 1)	-4028.96107	-4007.832775	-4021.077745
ARIMA (1, 1, 1)	-4028.473407	-4012.627186	-4022.560913
ARIMA (0, 1, 3)	-4027.830157	-4006.701863	-4019.946832
ARIMA (0, 1, 2)	-4027.191786	-4011.345565	-4021.279292
ARIMA (1, 1, 3)	-4026.693339	-4000.28297	-4016.839182
ARIMA (3, 1, 2)	-4025.610686	-3993.918244	-4013.785698
ARIMA (0, 1, 1)	-4014.279929	-4003.715782	-4010.338266
ARIMA (3, 1, 0)	-3720.063456	-3698.935161	-3712.18013
Model	AIC	BIC	HQIC
ARIMA (2, 1, 0)	-3589.231684	-3573.385463	-3583.31919
ARIMA (1, 1, 0)	-3444.378094	-3433.813947	-3440.436432
ARIMA (0, 1, 0)	-2873.390736	-2868.108662	-2871.419905

In time series analysis, the order determination of ARIMA models is typically based on a comprehensive evaluation using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Hannan-Quinn Information Criterion (HQIC). These three criteria are all grounded in the principle of maximum likelihood estimation and avoid overfitting by balancing model goodness-of-fit against parameter complexity. Their common core idea is to seek the optimal balance between fitting accuracy and model parsimony. Among them, the AIC criterion is suitable for prediction-oriented model selection, with its calculation formula given by AIC = -2ln(L) + 2k, where L is the likelihood function value and k is the number of parameters. The BIC criterion adds a penalty term for sample size based on AIC, with the formula BIC = -2ln(L) + kln(n), where n is the sample size. When the sample size is large, BIC tends to favor more parsimonious models. The HQIC falls between the two, with the formula HQIC = -2ln(L) + 2kln(ln(n)). Generally speaking, the smaller the value of an information criterion, the better the balance the model achieves between fitting performance and complexity.

A comparative analysis of the 16 ARIMA models described above shows that the ARIMA (1, 1, 2) model achieves the minimum values across all three information criteria: AIC is -4031.84, BIC is -4010.71, and HQIC is -4023.96, ranking first in all three metrics. The closely following ARIMA (2, 1, 3) model, although having an AIC value (-4031.30) very close to the former with a difference of only 0.54, has a BIC value (-3999.60) that is approximately 11.11 higher than that of ARIMA (1, 1, 2), indicating that the latter has an obvious advantage in model parsimony. In terms of the number of parameters, ARIMA (1, 1, 2) contains one autoregressive parameter and two moving average parameters, totaling three parameters to be estimated; whereas ARIMA (2, 1, 3) contains two autoregressive parameters and three moving average parameters, totaling five parameters to be estimated. According to the penalty characteristic of the BIC criterion regarding the number of parameters (with a penalty term of k·ln(n)), when the sample size is large, BIC tends to select models with fewer parameters. The significant advantage of ARIMA (1, 1, 2) in BIC further verifies that this model is less prone to overfitting. In addition, HQIC also confirms the optimality of ARIMA (1, 1, 2), enhancing the robustness of the model selection.

In summary, based on the comprehensive evaluation of the three information criteria—AIC, BIC, and HQIC—it can be concluded that ARIMA(1,1,2) is the optimal model among all candidate models. This model not only ranks first across all three criteria but also possesses a relatively parsimonious parameter structure, achieving the best balance between fitting accuracy and model complexity. Therefore, ARIMA (1, 1, 2) is adopted as the final model for subsequent time series forecasting analysis.

2.4. Model Fitting

Table 4. Model Statistics Table

Model	R2	Ljung-Box Q-statistic (lag 18)			lag5	Lag10	Lag15
		statistic	DF	P-value			
Open Price Model	0.9145	3.1047	15	0.999964	0.937620	0.999307	0.999948

Based on the model statistical results reported in Table 4, a systematic assessment of the fitting performance of the ARIMA (1, 1, 2) model can be undertaken. With respect to model goodness-of-fit, the R^2value of the opening price model is 0.9145, indicating that the model explains approximately 91.45% of the variability in the original series. This demonstrates a relatively ideal fitting performance and strong explanatory power of the model concerning the data. Regarding residual diagnostics, the Ljung-Box Q test is applied to examine whether the model residuals satisfy the white noise assumption. The results indicate that the Q-statistic at lag 18 is 3.1047 with 15 degrees of freedom, and the corresponding p-value is as high as 0.999964, which far exceeds the 0.05 significance level. Consequently, the null hypothesis of white noise residuals cannot be rejected. Further examination of the test results across different lag orders reveals that the p-value is 0.937620 at lag 5, 0.999930 at lag 10, and 0.999948 at lag 15. All p-values at various lag orders are substantially greater than 0.05. These findings suggest the absence of significant autocorrelation in the model residuals at different lag orders, and the residual series exhibits typical white noise characteristics. In conclusion, the ARIMA (1, 1, 2) model not only achieves a high degree of goodness-of-fit but also passes rigorous residual diagnostic tests, confirming that the model has adequately extracted the effective information embedded in the time series, leaving no further pattern to be captured in the residuals. Therefore, this model possesses sound statistical properties and reliable predictive capability, rendering it suitable for subsequent forecasting analysis.

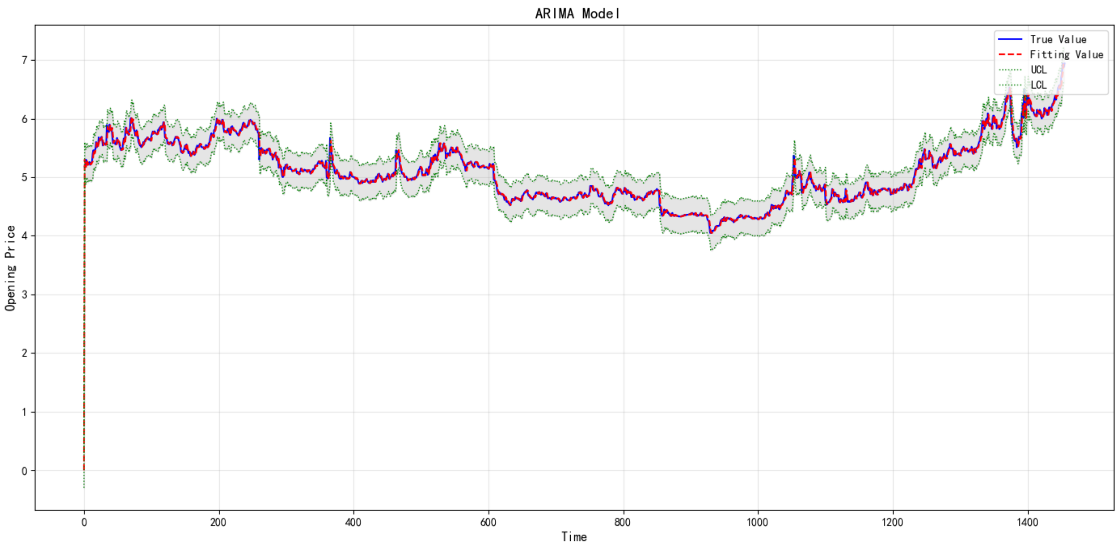

Figure 3. Stock Opening Price Fitting Chart Based on ARIMA Model

Figure 3 presents the fitting performance of the ARIMA (1, 1, 2) model for the opening price data of the Industrial and Commercial Bank of China (ICBC). In this figure, the solid blue line represents the actual observed values, the dashed red line represents the model fitted values, the dotted green lines respectively represent the upper confidence limit (UCL) and the lower confidence limit (LCL) of the 95% confidence interval, and the shaded gray area represents the confidence interval. Through a systematic analysis of this figure, an intuitive and comprehensive evaluation of the model's fitting performance can be conducted.

From the perspective of the overall trend, the fitted values of the ARIMA (1, 1, 2) model exhibit a high degree of consistency with the actual observed values in terms of global movement. The two curves move synchronously throughout the entire observation period (observation numbers 0 to 1400), and the fitted values effectively capture the downward trend of the original data as well as its phase-specific variation characteristics. Specifically, in the first half of the series (approximately observations 0-400), the original data slowly declines from approximately 5.2 to approximately 5.0, and the fitted curve accurately follows this gradual downward trend, demonstrating a good degree of alignment between the two. In the middle segment of the series (approximately observations 400-800), the original data undergoes the most dramatic decline, rapidly falling from approximately 5.0 to approximately 4.2. The fitted curve also exhibits a distinctly accelerated downward trend, indicating that the model has a good ability to respond to abrupt trend changes. In the latter segment of the series (approximately observations 800-1400), the original data gradually stabilizes, fluctuating within a narrow range around 4.0-4.1. The fitted values likewise enter a stationary state, achieving the highest degree of alignment with the actual values.

From the perspective of local fitting accuracy, although the model performs excellently in terms of overall trend, certain deviations exist between the fitted values and the actual values in some locally volatile regions. This phenomenon is particularly evident in the rapidly declining phase between approximately observations 400 and 800, where the dashed red line does not completely coincide with the solid blue line at certain time points. However, it is worth noting that these deviations do not exhibit a systematic directional preference; that is, the fitted values are sometimes higher than the actual values and sometimes lower, indicating the absence of any obvious systematic prediction bias in the model. Such localized fitting errors primarily originate from the inherent stochastic fluctuation characteristics of the stock market itself, rather than from deficiencies in the model structure.

From the perspective of confidence interval analysis, the shaded gray area in the figure represents the 95% confidence interval, providing a quantitative reference for model prediction uncertainty. Observation reveals that the vast majority of actual observed values fall within the confidence interval, with only a very few points located near or slightly beyond the interval boundaries. This indicates that the model's estimation of fitting uncertainty is reasonably appropriate and well-calibrated. The width of the confidence interval remains relatively stable throughout the entire observation period, showing no obvious tendency to expand over time. This further suggests that the model variance exhibits homoscedasticity, satisfying the classical linear regression assumption.

In summary, the ARIMA (1, 1, 2) model demonstrates good performance in terms of overall trend fitting, phase-specific variation capture, and uncertainty quantification. Although certain fitting errors exist during locally volatile periods, these errors are within an acceptable range and exhibit no systematic bias. Combined with the statistical results presented in Table 4, where the R^2 reaches 0.9145 and the Ljung-Box test p-values are substantially greater than 0.05, it can be concluded that the ARIMA (1, 1, 2) model achieves an excellent fitting effect for the ICBC opening price data, effectively extracting the useful information embedded in the time series and providing a reliable model foundation for subsequent forecasting analysis.

Table 5. Predictive statistical indicators

Predict days	MAE	RMSE
10 days	0.07	0.09
50 days	0.05	0.06
116 days	0.08	0.11

Based on the error statistics of the ARIMA (1, 1, 2) model under different forecast horizons as presented in Table 5, a systematic evaluation of the model's forecasting performance can be conducted, leading to the following conclusions.

Overall, the ARIMA (1, 1, 2) model exhibits low forecast errors in short-term, medium-term, and long-term predictions. The MAE and RMSE values across all forecast horizons do not exceed 0.11, indicating that the model possesses high accuracy and stability in predicting the opening price of the Industrial and Commercial Bank of China (ICBC). Specifically, in the short-term forecast (10 days), the MAE is 0.07 and the RMSE is 0.09, demonstrating that the model can make relatively accurate predictions of recent price movements with errors controlled within a small range. In the medium-term forecast (50 days), the model achieves its lowest forecast errors, with an MAE of 0.05 and an RMSE of 0.06. This encouraging result indicates that the model not only maintains good predictive capability over the medium-term horizon but also that errors do not accumulate and amplify as the forecast step length increases. Instead, errors show a certain degree of optimization, which may be related to the tendency of the data series to become stationary and exhibit reduced volatility during the medium-term period. In the long-term forecast (116 days), the MAE is 0.08 and the RMSE is 0.11. Although slightly higher than the accuracy of the medium-term forecast, these errors remain at a low level, indicating that the model still possesses reliable predictive capability over a longer time span, with errors increasing moderately but remaining within an acceptable range.

Further comparison of forecasting performance across different horizons reveals that the model's forecast errors do not exhibit a monotonically increasing trend as the number of forecast days increases. The medium-term forecast errors are actually lower than the short-term forecast errors, a phenomenon suggesting that the ARIMA (1, 1, 2) model has a unique advantage in capturing medium-term price dynamics. In addition, across all forecast horizons, the ratio of RMSE to MAE is close to 1, indicating that the distribution of forecast errors is relatively symmetric with no extreme anomalous predictions, further verifying the robustness of the model.

In summary, the ARIMA (1, 1, 2) model demonstrates excellent performance in short-term, medium-term, and long-term forecasts, with forecast errors consistently maintained at low levels. The model effectively captures the time series characteristics of the ICBC opening price and provides investors with reliable price forecast references. Therefore, the ARIMA (1, 1, 2) model can serve as an effective tool for forecasting this stock price and has practical application value.

3. Summary

This study systematically applies the ARIMA model to analyze and forecast the daily opening price data of the Industrial and Commercial Bank of China (ICBC) from January 2019 to December 2024, comprising 1,456 observations. The empirical analysis proceeds through several stages. First, descriptive statistics reveal that the data exhibit good central tendency and moderate volatility, with a mean of 5.1107 and a standard deviation of 0.5410, and no extreme outliers are detected. Second, the Augmented Dickey-Fuller (ADF) test confirms that the original series is non-stationary (p-value = 0.873235) but becomes stationary after first-order differencing (p-value = 0.000000), establishing the series as an I(1) process. The time series plot of the original data shows a clear nonlinear downward trend, while the first-differenced series fluctuates randomly around zero, indicating the successful removal of trend components. Third, the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced series exhibit a distinct cutoff at lag 1, guiding the identification of candidate models. Fourth, among 16 candidate ARIMA models evaluated using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Hannan-Quinn Information Criterion (HQIC), the ARIMA (1, 1, 2) model is selected as optimal, achieving the lowest values across all three criteria. Fifth, diagnostic checks, including the Ljung-Box Q-test at lag orders 5, 10, 15, and 18, indicate that the residuals are white noise, with all p-values substantially greater than 0.05, confirming adequate model specification. The model demonstrates strong in-sample fit with an R-squared of 0.9145. The fitting plot shows a high degree of consistency between the fitted values and actual observations, with the vast majority of actual values falling within the 95% confidence interval, and no obvious systematic bias is observed. Finally, forecast evaluations over short-term (10 days), medium-term (50 days), and long-term (116 days) horizons yield MAE/RMSE values of 0.07/0.09, 0.05/0.06, and 0.08/0.11, respectively. Notably, the medium-term forecast errors are lower than the short-term errors, suggesting that the model has a unique advantage in capturing medium-term price dynamics. In conclusion, the ARIMA (1, 1, 2) model effectively captures the underlying dynamics of stock opening prices, exhibits robust predictive performance across different horizons, and can serve as an effective forecasting tool for investors in volatile financial markets.

References

[1] He, Y., Li, J. M., Ruan, S. M., et al. (2020). A hybrid model for financial time series forecasting—Integration of EWT, ARIMA with the improved ABC optimized ELM. IEEE Access, 8, 84501–84518. https://doi.org/10.1109/ACCESS.2020.2992075

[2] Wang, Y. (2021). Analysis and prediction of stock price based on ARMA model. Productivity Research, (9), 124–127.

[3] Liu, J. (2021). Application of ARMA model in stock price prediction—Taking Gree Electric Appliances as an example. China Management Informationization, 24(11), 153–155.

[4] Wu, Y. X., & Wen, X. (2016). Short-term stock price prediction based on ARIMA model. Statistics & Decision, (23), 83–86.

[5] Feng, P., & Cao, X. B. (2011). An empirical study on stock price analysis and prediction based on the ARIMA model. Mathematics in Practice and Theory, 41(22), 84–90.