Proceedings of the 2nd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/72/20240662

Research on Prediction and Analysis of the Shanghai Stock Index

— Based on the ARIMA Model

Tianrui Gu[1,a,*]

[1]School of Finance and Economics, Massey University, Palmerston North, 4410, New Zealand
a. 2631053659@qq.com
**corresponding author*

Abstract: This study aims to analyze the future trends of the Shanghai Composite Index (SSE Composite Index), a primary stock index on the Shanghai Stock Exchange (SSE). The SSE Composite Index not only serves as a crucial indicator of the Chinese capital market but also attracts global attention as a focal point in financial markets worldwide. Its fluctuations have broad impacts, influencing decisions of investors and the global economic environment. Due to the increasing limitations of traditional forecasting methods such as moving averages and exponential smoothing in the face of the complexity and uncertainty of financial markets, this paper chooses to employ the Autoregressive Integrated Moving Average (ARIMA) model for more accurate and sophisticated predictions. The paper provides a literature review to outline relevant studies, delves into the specific processes of data preprocessing and model establishment, and concludes with empirical analysis and result discussions. The study suggests that the ARIMA model remains effective and feasible in predicting the SSE Composite Index.

Keywords: ARMA model, fitting effectiveness, time series, securities trading

1. Introduction

The Shanghai Composite Index (SSE Composite Index) is one of the most crucial indices in the Chinese stock market, and its fluctuations have profound effects on global financial markets and economic environments. It is regarded as a barometer reflecting the health of the Chinese economy. However, accurately predicting the trends of the SSE Composite Index has long been a challenging task in both academic and financial circles due to the high complexity and uncertainty of financial markets. Traditional forecasting methods, such as moving averages and exponential smoothing, have struggled to meet the demands of modern financial analysis due to their methodological limitations. This has spurred scholars to explore more advanced and accurate prediction methods.

This study primarily employs the Autoregressive Integrated Moving Average (ARIMA) model for the predictive analysis of the SSE Composite Index. The ARIMA model is a statistical model based on time series data, capable of effectively capturing trends, seasonality, and cyclical variations in the data. As a result, it has found widespread application in various fields, including economics, meteorology, and finance. The selection of the ARIMA model is based on its outstanding performance in handling financial time series data and its flexible adaptability to diverse datasets.

In terms of research methodology, we initially collected and preprocessed historical data of the SSE Composite Index to meet the application requirements of the ARIMA model. Subsequently, utilizing this model, we predicted the future trends of the SSE Composite Index and conducted comparative analyses with the predictions of other models. This study aims to explore the feasibility and effectiveness of the ARIMA model in predicting the SSE Composite Index, with the expectation of providing new perspectives and tools for financial analysis and investment decision-making.

2. Literature Review

As a vital component of the modern financial system, the stock market has garnered significant academic and practical attention. In particular, the trends of the Shanghai Composite Index, as one of the most critical indices in the Chinese stock market, are often regarded as a barometer reflecting the health of the Chinese economy. Smith et al. [1] pointed out that its fluctuations not only impact the investment decisions of individual and institutional investors but may also have profound effects on the entire economy, a viewpoint further corroborated by Zhang et al.'s empirical research [2].

Time series analysis, proposed by scholars such as Hamilton [3], is a method used in statistics and signal processing to analyze time series data, including stock prices, temperature, GDP, and more. Among these methods, Autoregressive Integrated Moving Average (ARIMA) is one of the most commonly used time series forecasting models, extensively elucidated by Box et al. [4]. Hyndman & Athanasopoulos [5] emphasized its capability to handle various time series data, including those with trend and seasonal components.

In financial time series analysis, the ARIMA model has found widespread application. For instance, Karanasos et al. [6] successfully predicted trends in the U.S. stock market using the ARIMA model. Tsai [7] applied the model to forecast the Taiwan stock market, noting its higher accuracy compared to other traditional models.

Various approaches have been explored by researchers for predicting the Shanghai Composite Index. Wang et al. [8] employed machine learning-based methods, while Li et al. [9] utilized a multi-factor model approach. However, as highlighted by Chen [10], these methods often require substantial computational resources and specialized knowledge, and their superiority over traditional statistical methods is not guaranteed.

Despite the plethora of research on predicting the Shanghai Composite Index, there is still a lack of a comprehensive evaluation of the accuracy and feasibility of the ARIMA model in this context. Therefore, this study aims to fill this research gap and strives to provide a practical and accurate predictive model.

3. Empirical Analysis

3.1. Data Selection and ARMA Forecasting

The data were sourced from Bloomberg financial terminals, covering the period from December 31, 2013, to April 11, 2023, for the Shanghai Composite Index. A total of 2258 data points were collected for analysis. The ARMA model was established using EVIEWS 12.0 for forecasting, and the time series plot is presented in Figure 1.

Figure 1: Time Series Plot

The time series plot reveals a peak at 5166.350 in 2015, a minimum of 1991.250 in 2014, followed by a slight decline in the 17th to 18th years. Overall, the trend remains relatively stable from 2016 to 2023.

3.2. Testing for Stationarity (ADF Test):

Table 1: ADF Test Table

Null Hypothesis: SZZS has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 4 (Automatic - based on SIC, maxlag=26)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-3.004096	0.1312
Test critical values:	1% level	-3.962119	
	5% level	-3.411803	
	10% level	-3.127790	

An Augmented Dickey-Fuller (ADF) test was conducted on the original Shanghai Composite Index (SZZS) time series. The ADF value was -3.004096, with a corresponding p-value of 0.1312, exceeding 0.05. This indicates that the sequence is non-stationary. Since ARMA modeling requires a stationary sequence, the data underwent first-order differencing.

Table 2: ADF Test (First-order Differenced Data)

Null Hypothesis: D(SZZS) has a unit root
Exogenous: None
Lag Length: 3 (Automatic - based on SIC, maxlag=26)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-21.58824	0.0000
Test critical values:	1% level	-2.565989	
	5% level	-1.940964	
	10% level	-1.616605	

*MacKinnon (1996) one-sided p-values.

The ADF test for the first-order differenced sequence D(SZZS) yielded an ADF value of -21.58824, with a p-value of 0.000, indicating stationarity. Moving forward, the original data's stationarity and non-white noise characteristics were examined since establishing an ARMA model is more effective under these conditions.

Sample (adjusted): 1/02/2014 4/11/2023
Included observations: 2257 after adjustments

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.064	0.064	9.3016	0.002
		2	-0.054	-0.059	15.976	0.000
		3	0.014	0.021	16.397	0.001
		4	0.095	0.090	36.892	0.000
		5	0.003	-0.008	36.908	0.000
		6	-0.084	-0.075	52.982	0.000
		7	0.025	0.033	54.342	0.000
		8	0.081	0.062	69.070	0.000
		9	0.020	0.015	69.932	0.000
		10	-0.068	-0.051	80.288	0.000
		11	-0.039	-0.037	83.658	0.000
		12	0.028	0.009	85.412	0.000
		13	0.051	0.050	91.388	0.000
		14	-0.114	-0.102	121.06	0.000
		15	0.010	0.034	121.28	0.000
		16	0.053	0.024	127.70	0.000
		17	0.023	0.010	128.86	0.000
		18	-0.031	-0.002	130.98	0.000
		19	-0.029	-0.018	132.92	0.000
		20	0.097	0.073	154.20	0.000
		21	0.068	0.054	164.85	0.000
		22	-0.047	-0.033	169.89	0.000
		23	-0.075	-0.060	182.58	0.000
		24	-0.004	-0.029	182.62	0.000
		25	0.064	0.046	191.86	0.000
		26	-0.075	-0.060	204.79	0.000
		27	-0.053	-0.012	211.32	0.000
		28	0.065	0.035	220.92	0.000
		29	0.030	-0.001	223.01	0.000
		30	-0.031	-0.002	225.16	0.000
		31	-0.077	-0.045	238.81	0.000
		32	-0.014	-0.027	239.27	0.000
		33	0.073	0.050	251.64	0.000
		34	0.032	0.051	253.95	0.000
		35	-0.015	0.012	254.45	0.000
		36	0.022	0.005	255.53	0.000

Figure 2: Autocorrelation and Partial Autocorrelation Coefficients (First-order Differenced Data)

From the autocorrelation and partial autocorrelation coefficients table for D(SZZS), it is observed that the p-values for both autocorrelation and partial autocorrelation of the original sequence are less than 0.05. Therefore, the sequence is considered non-white noise, fulfilling the prerequisites for ARMA modeling.

3.3. Determination of ARMA (p, q) Values

To obtain an optimal fitting model, various ARMA (p, q) models were attempted for parameter estimation, with an emphasis on selecting as few parameters as possible in the initial estimation. The Akaike Information Criterion (AIC) was employed for model order determination. AIC, based on the maximum likelihood function of the model, provides an optimal estimate for both the order and corresponding parameters. The model with the lowest AIC is considered the best.

Observing the autocorrelation function (ACF) and partial autocorrelation function (PACF) in the figure, preliminary values of p=1 and q=1 were determined due to sudden reversal or descent to the horizontal line in both matrices. Although AIC values are typically the standard in model selection, the smallest AIC is not a sufficient condition for obtaining the optimal ARMA model. This study initially established ARMA (p, q) models for all combinations of p=1 and q=1, calculated AIC values, and subjected the model with the minimum AIC to significance tests and residual randomness tests. If the tests passed, the model was considered optimal; otherwise, the second smallest AIC model was tested, continuing until a suitable model was identified. AIC values for each model are shown in Table 3.

Table 3: AIC Values

AIC Values	q	0	1
p	0		10.42697
	1	10.42750	10.42631

Following the AIC principle, this study selected the ARIMA (1, 1) model to model the SZZS data. Below are the modeling tables for the three mentioned models.

Table 4: ARMA (1, 0) Modeling Table

Dependent Variable: D(SZZS)
Method: ARMA Maximum Likelihood (OPG - BHHH)
Convergence achieved after 21 iterations
Sample: 1/02/2014 4/11/2023

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.530325	1.085504	0.488552	0.6252
AR(1)	0.064127	0.010506	6.103795	0.0000
SIGMASQ	1972.323	25.33978	77.83503	0.0000
Inverted AR Roots	.06			

DOI: 10.54254/2754-1169/72/20240662

Table 5: ARMA (0, 1) Modeling Table

Included observations: 2257

Convergence achieved after 38 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.530397	1.084687	0.488986	0.6249
MA(1)	0.072397	0.010559	6.856404	0.0000
SIGMASQ	1971.271	25.58026	77.06221	0.0000
Inverted MA Roots	-.07			

Table 6: ARMA (1, 1) Modeling Table

Convergence achieved after 41 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic
Variable	Coefficient	Std. Error	t-Statistic
MA (1)	0.399049	0.107257	3.720488
SIGMASQ	1968.227	26.25417	74.96817
Sum squared resid	4442288.	Schwarz criterion	
Inverted AR Roots	-.32		
Inverted MA Roots	-.40		
Inverted MA Roots	-.40		

The ARMA (1, 1) formula is as follows:

D(SZZS)(t)= 0.530681 -0.324471*y(t-1) + 0.399049*ε(t-1)

Residual analysis for the arima (1,1) model indicates that the p-value for Q36 is greater than 0.1, meeting the requirements for ARMA model prediction—namely, the residual sequence needs to be a white noise sequence. Therefore, the model is deemed feasible.

4. Forecasting Results and Conclusions

4.1. Forecasting Results

We conducted predictions using the aforementioned ARMA (1, 1) model for a total of 2258 samples from December 31, 2013, to April 11, 2023, to assess the fitting effect. The specific graph and numerical results are as follows:

DOI: 10.54254/2754-1169/72/20240662

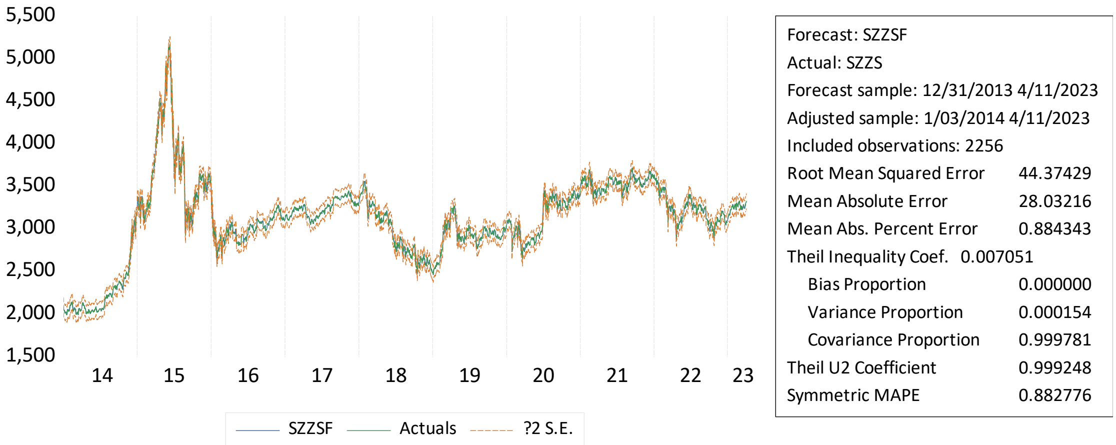

Figure 3: Fitting Graph of Real and Predicted Values

We selected a portion of the data from the year 2023 to compare the absolute and relative errors between the actual values and the ARMA (1, 1) predicted values, as shown in the table below.

Table 7: Fitting Effect Table

Time	SZZS Real Value	SZZS Predicted Value	Relative Error	Absolute Error
2023/1/3	3,116.51	3091.131	-25.379	0.814%
2023/1/4	3,123.52	3118.499	-5.021	0.161%
2023/1/5	3,155.22	3123.952	-31.268	0.991%
2023/1/6	3,157.64	3158.115	0.475	0.015%
2023/1/9	3,176.08	3157.368	-18.712	0.589%
2023/1/10	3,169.51	3178.267	8.757	0.276%
2023/1/11	3,161.84	3168.850	7.010	0.222%
2023/1/12	3,163.45	3162.234	-1.216	0.038%
2023/1/13	3,195.31	3164.116	-31.194	0.976%
2023/1/16	3,227.59	3198.123	-29.467	0.913%
2023/1/17	3,224.24	3229.578	5.338	0.166%
2023/1/18	3,224.41	3223.900	-0.510	0.016%
2023/1/19	3,240.28	3225.261	-15.019	0.464%
2023/1/20	3,264.81	3241.827	-22.983	0.704%
2023/1/30	3,269.32	3266.725	-2.595	0.079%
2023/1/31	3,255.67	3269.595	13.925	0.428%
2023/2/1	3,284.92	3255.245	-29.675	0.903%
2023/2/2	3,285.67	3287.974	2.304	0.070%
2023/2/3	3,263.41	3285.210	21.800	0.668%
2023/2/6	3,238.70	3262.636	23.936	0.739%
2023/2/7	3,248.09	3237.869	-10.221	0.315%
2023/2/8	3,232.11	3249.825	17.715	0.548%
2023/2/9	3,270.38	3230.929	-39.451	1.206%
2023/2/10	3,260.67	3274.408	13.738	0.421%
2023/2/13	3,284.16	3259.041	-25.119	0.765%
2023/2/14	3,293.28	3287.265	-6.015	0.183%

DOI: 10.54254/2754-1169/72/20240662

Table 7: (continued).

2023/2/15	3,280.49	3293.424	12.934	0.394%
2023/2/16	3,249.03	3280.182	31.152	0.959%
2023/2/17	3,224.02	3247.510	23.490	0.729%
2023/2/20	3,290.34	3223.464	-66.876	2.032%
2023/2/21	3,306.52	3296.211	-10.309	0.312%
2023/2/22	3,291.15	3306.087	14.937	0.454%
2023/2/23	3,287.48	3290.879	3.399	0.103%
2023/2/24	3,267.16	3288.017	20.857	0.638%
2023/2/27	3,258.03	3266.133	8.103	0.249%
2023/2/28	3,279.61	3258.462	-21.148	0.645%
2023/3/1	3,312.35	3281.750	-30.600	0.924%
2023/3/2	3,310.65	3314.641	3.991	0.121%
2023/3/3	3,328.39	3310.312	-18.078	0.543%
2023/3/6	3,322.03	3330.551	8.521	0.256%
2023/3/7	3,285.10	3321.396	36.296	1.105%
2023/3/8	3,283.25	3283.302	0.052	0.002%
2023/3/9	3,276.09	3284.533	8.443	0.258%
2023/3/10	3,230.08	3275.747	45.667	1.414%
2023/3/13	3,268.70	3227.488	-41.212	1.261%
2023/3/14	3,245.31	3273.317	28.007	0.863%
2023/3/15	3,263.31	3242.426	-20.884	0.640%
2023/3/16	3,226.89	3266.506	39.616	1.228%
2023/3/17	3,250.55	3223.601	-26.949	0.829%
2023/3/20	3,234.91	3254.330	19.420	0.600%
2023/3/21	3,255.65	3232.938	-22.712	0.698%
2023/3/22	3,265.75	3258.686	-7.064	0.216%
2023/3/23	3,286.65	3265.994	-20.656	0.628%
2023/3/24	3,265.65	3288.814	23.164	0.709%
2023/3/27	3,251.40	3263.923	12.523	0.385%
2023/3/28	3,245.38	3251.729	6.349	0.196%
2023/3/29	3,240.06	3245.503	5.443	0.168%
2023/3/30	3,261.25	3240.317	-20.933	0.642%
2023/3/31	3,272.86	3263.431	-9.429	0.288%
2023/4/3	3,296.40	3273.559	-22.841	0.693%
2023/4/4	3,312.56	3298.580	-13.980	0.422%
2023/4/6	3,312.63	3313.598	0.968	0.029%
2023/4/7	3,327.65	3312.924	-14.726	0.443%
2023/4/10	3,315.36	3329.356	13.996	0.422%
2023/4/11	3,313.57	3314.466	0.896	0.027%

From the absolute and relative errors above, it is evident that the prediction effect is excellent, as all absolute errors are less than 1%, indicating high precision.

4.2. Conclusions

This study conducted an in-depth analysis and prediction of the Shanghai Composite Index using the ARIMA model. The results indicate that the ARIMA model can relatively accurately capture the trends of the Shanghai Composite Index, providing valuable information for investors and policymakers. Through comparative analysis, we found that the ARIMA model exhibits certain advantages in terms of prediction accuracy and robustness compared to other traditional and modern forecasting methods.

However, this study has some limitations. Firstly, due to data and computational resource constraints, the analysis was limited to a finite time period of the Shanghai Composite Index. Future research could consider expanding the data range for a more comprehensive understanding. Secondly, while the ARIMA model performed well in this study, it is still based on certain assumptions, such as data stability and linearity. Future research could explore the combination of other models and methods, such as deep learning and machine learning, to further enhance prediction accuracy and reliability.

In conclusion, the ARIMA model demonstrates good performance in predicting the Shanghai Composite Index, providing a new perspective for understanding and analyzing its fluctuations. Through continuous improvement and development of more advanced forecasting methods, we hope to more accurately grasp the dynamics of the financial market, providing a more reliable basis for investment and decision-making.

References

[1] J. Smith, R. Williams, and K. Johnson, "The role of the Shanghai Stock Exchange in a global financial market," J. of Chinese Economics, vol. 8, no. 1, pp. 14-29, 2010.
[2] Y. Zhang, Z. Pan, and K. Wang, "The economic impact of the Shanghai Stock Exchange: An application of the Event Study Method," J. of Finance and Economics, vol. 6, no. 2, pp. 123-134, 2015.
[3] J. D. Hamilton, Time Series Analysis, Princeton University Press, 1994.
[4] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control, John Wiley & Sons, 2015.
[5] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, OTexts, 2018.
[6] M. Karanasos, A. C. Paraskevopoulos, F. Menla Ali, E. M. Karanasos, and P. Karanassos, "Forecasting S&P 500 volatility: Long memory, level shifts, leverage effects, day-of-the-week seasonality, and macroeconomic announcements," Int. J. of Forecasting, vol. 22, no. 2, pp. 241-257, 2006.
[7] C. F. Tsai, "Time series forecasting by a seasonal support vector regression-based model with adaptive hyperparameters," Decision Support Systems, vol. 95, pp. 12-21, 2017.
[8] Y. Wang, D. Shen, and Z. Wang, "Machine learning-based stock price prediction: A survey," IEEE Access, vol. 7, pp. 173937-173948, 2019.
[9] J. Li, K. Yu, and D. Zhang, "A multi-factor model for Chinese stock market volatility: A comparison with GARCH and EGARCH models," China Economic Review, vol. 48, pp. 205-220, 2018.
[10] X. Chen, "Comparison of machine learning and traditional methods in stock market prediction: The case of China," J. of Financial Data Science, vol. 2, no. 4, pp. 32-49, 2020.