Shanghai Stock Exchange Composite Index Prediction Using Macro Variables Based on ARIMAX Model

Liyu Zheng[1,a,*]

[1]*Department of Architecture, Tsinghua University, Beijing, China*
a. zheng-ly19@mails.tsinghua.edu.cn
**corresponding author*

Abstract: Stock market prediction has always been a prevailing topic among investors and researchers. Among numerous market index, the Shanghai Stock Exchange Composite Index (SSE Index) is recognized as one of the most indicative stock indexes in China's A-share market. As it is discovered that the fluctuations of SSE index and exchange rate (USD/CNY) displays highly similar pattern since the subprime mortgage crisis, this study aims to use macro variables (including exchange rate) to predict the SSE index based on ARIMAX model. The data are collected since 2006 and based on which ARIMA (14,1,4) model is generated. Distinct macro variables are added in this ARIMA model respectively and it concludes that the incorporation of exchange rate with certain lags can significantly increase the fitting degree. Cross validations on different lengths of validation sets are implemented, which shows that the model can make accurate forecast in relatively short term. The result manifests increasing accuracy when incorporating certain explanatory regressive factors, which might provide valuable and enlightening information in short term for researchers or investors.

Keywords: SSE Index, China’s a-share market, ARIMAX model, macroeconomic variables, exchange rate

1. Introduction

One of the major stock indexes in China's A-share market is the Shanghai Stock Exchange Composite Index (SSE Index). It covers companies with large market capitalization in the Shanghai Stock Exchange, so it is widely used as a representative of the overall performance of China's stock market. Therefore, it becomes an essential topic for investors to make prediction on the movement of the SSE Index.

Based on autoregressive integrated moving average (ARIMA) model, this study aims to fit and make forecast of the SSE Index. Additionally, macro variables (such as exchange rate etc.) are expected to be taken as explanatory factors. The macro variables involved in this research are exchange rate, interest rate, inflation rate. Thus, integrated ARIMAX model is used to make the prediction more accurately.

The SSE Index rose continuously from 2006 to 2007 and reached a peak in 2007(in Figure 1). Since then, affected by the subprime mortgage crisis, the SSE Index began to decline gradually. In the subsequent trend, the SSE Index fluctuated up and down, and reached a historical summit at the end of 2015, which shows similarity with the fluctuations of the CNY exchange rate. USD/CNY exchange rate is used as exchange rate in this study.

Decreasing interest rate might prompt people to invest in stock market. This study tries to discover whether the change of interest rate can have a certain impact on the SSE Index. 1 year's Shibor rate is used to represent interest rate. As China's benchmark interest rate, Shibor rate reflects borrowing costs and liquidity in the financial system.

Besides, CPI growth rate reveals consumers' purchasing power and the overall price level, which have an impact on the investment behaviors of the consumers. CPI growth rate on month-on-month basis is recognized as inflation rate in this study.

A combination of these factors and their macroeconomic implications is probably enlightening in predicting the direction of the SSE Index. Incorporating these macro variables into SSE index prediction may bring out more precise prediction, thus providing valuable information for investors and researchers.

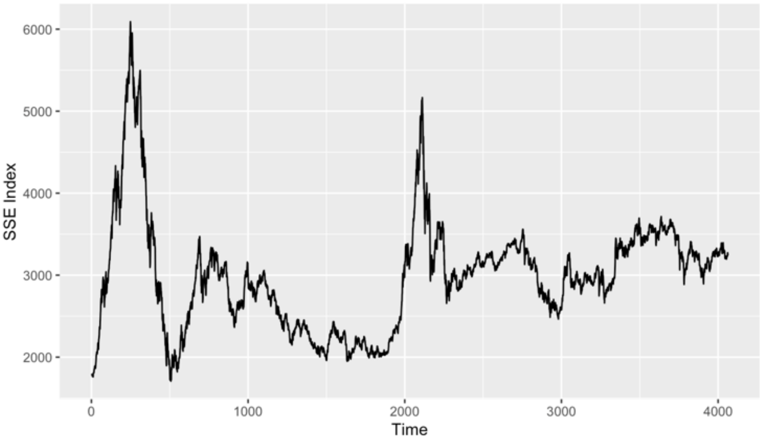

Figure 1: SSE Index from 2006/10/09 to 2023/06/20.

2. Literature Review

Macroeconomic variables imperceptibly influence the stock market. A large number of scholars use empirical research methods to focus on how macroeconomic factors affect stock price fluctuations. This article mainly focuses on exchange rate, interest rate and inflation rate.

As for exchange rate, in 1980, Dornbusch and Fischer analyzed the relationship between exchange rates and stock prices, believing that the former can affect a company's stock price through its economic activities, resulting in a one-way impact relationship [1]. Subsequently, Granger et al. studied the data of Asian countries and found that exchange rate is the Granger causality of stock prices in 2000 [2]; Ba Shusong et al., Wang Bo et al. found a unidirectional price guidance relationship between the foreign exchange market and the stock market in China [3,4]. Liu Jiming also found that the CNY exchange rate has a unidirectional mean and volatility spillover effect on the Chinese stock market [5]. Li Xindan et al. found that in the short term, China's exchange rates and stock prices are mutually causal too [6].

When it comes to inflation rate and interest rate, in 1992, Fama studied the securities market in the United States and demonstrated an equilibrium relationship between macro indicators [7]. The research results of Masih, Dickinson and Wong also respectively show that stock prices can be influenced by macroeconomic factors include money supply, interest rate, etc. [8-10].

Therefore, based on the former studies, this research proposes to use macro variables as explanatory factors to make prediction on the SSE Index based on ARIMAX model.

3. Methodology

3.1. Model Introduction

ARIMA model is a classical time series auto regressive model which is widely used in finance and economics to capture changing patterns of time series data and to make further forecasting. By integrating regressive time series explanatory factors in ARIMA model gives out ARIMAX model. Therefore, macro variables can be incorporated to increase the accuracy of prediction.

3.2. Data Collection

The study mainly focuses on whether exchange rate, interest rate and inflation rate can contribute to the forecasting of SSE index. USD to CNY rate, Shibor rate, CPI growth rate (month on month basis) are used to represent them respectively.

The data collected ranges from 2006/10/09 (before the arise of subprime mortgage crisis) to 2023/06/20, total of 4063 trading days on Choice Financial Terminal Software.

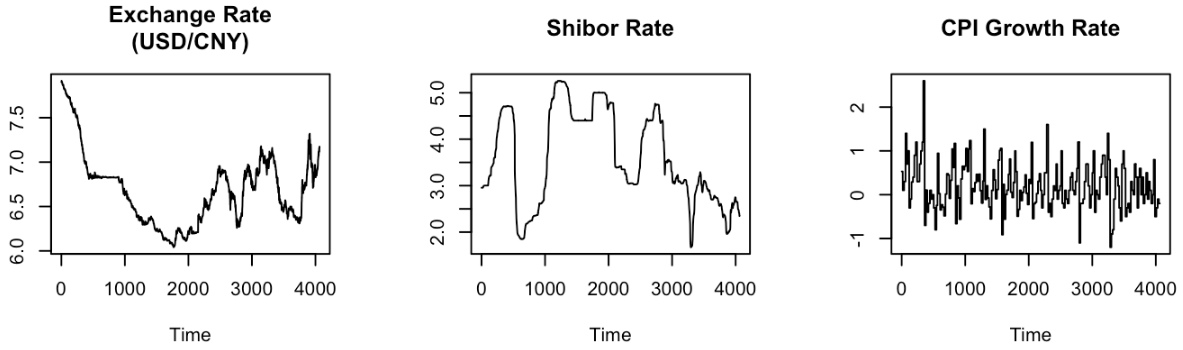

Figure 2: Data visualization.

3.3. Data Transformation

3.3.1. Stationarity and ADF Test

A significant prerequisite of ARIMA model fitting is the stationarity of data. Therefore, all the time series data (including the regressive explanatory factors) are required to be transformed into stationary time series data.

In statistics, the existence of unit root can be tested by augmented Dickey-Fuller test (ADF test), which is equivalent to the stationarity. Consequently, all the data must pass ADF test before being integrated into ARIMA model.

3.3.2. Box-Cox Transformation

In 1964, Box and Cox proposed a generalized power transformation method, Box-Cox transformation, which is commonly used in statistical modeling. Box-Cox transformation decreases the correlation between variables to a certain degree.

The best lambda of Box-Cox transformation on SSE index data evaluated by R is given by -0.2891, fairly close to 0. Therefore, logarithm (lambda = 0) is used to normalize the data. Nevertheless, the logarithm of SSE index does not pass the ADF test, which indicates logarithm of SSE index is not stationary time series data.

3.3.3. First-Order Differencing

Differencing data helps to stabilize the mean. By implementing first-order differencing to the logarithm of SSE index, the ACF (sample autocorrelation function) of which gradually falls to insignificant level.

The p-value of ADF test on the first-order differenced data is smaller than 0.01, indicating the first-order differencing of the logarithm of the index data is stationary time series data (in Figure 3).

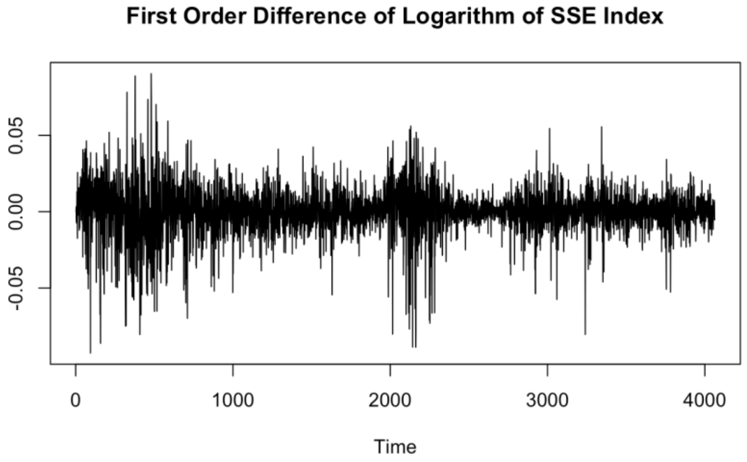

Figure 3: First order difference of logarithm of SSE index.

3.3.4. Transformation on Macro Variables

Analogous transformations are added on the 3 macro variables data. Due to the magnitude of these variables are not large enough, it is unnecessary to use logarithm transformation. The inflation (growth) rate itself is identified as stationary data by ADF test. The first-order differences of other two variables pass the ADF test (with all p-values are smaller than 0.01), which implies that their first-order differences are stationary time series data (in Figure 4).

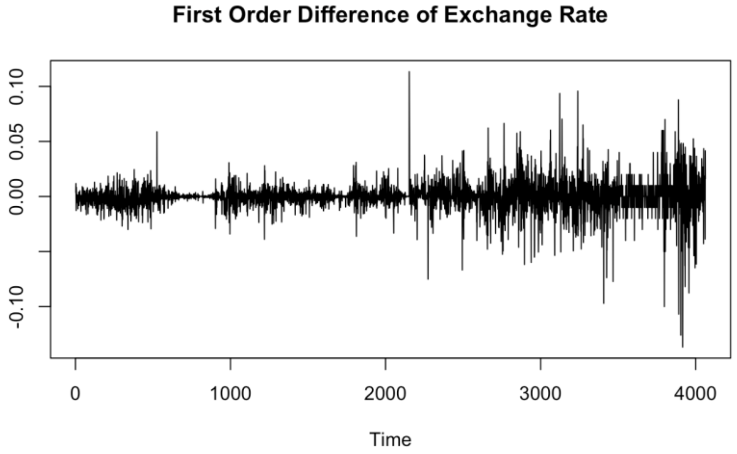

Figure 4: First order difference of exchange rate (USD/CNY).

3.4. Model Fitting

3.4.1. ARIMA Parameters Selection

In ARIMA (p, d, q) model, p refers to order of the autoregressive part AR(p). d refers to degree of first differencing involved and q refers to order of the moving average part MA(q) respectively.

All the possible (p, q) selections can be given by the ACF and PACF plots of the transformed data. For the SSE Index, specifically, significant values of p are 4, 6, 13, 14, etc. and significant values of q are 4, 6, 7, 13, 14, etc. In practice, the orders of p, q chosen are usually no more than 10 to avoid the overfitting problem brought by excessive parameters. ACF and PACF of first order difference of logarithm of SSE index are shown in Figure 5.

The optimal ARIMA model is selected by means of AIC rules. Lower AIC manifests higher accuracy of model fitting. By enumerating all the possible (p, q) combinations it gives out the lowest AIC order for SSE Index is (p, q) = (6,2).

As for the macro variables series, by conducting the same process, the ARIMA model selected for exchange rate, interest rate and inflation rate are identified as ARIMA (2,2,1), ARIMA (4,1,2) and ARIMA (0,1,0). Noticeably, the series of inflation (growth) rate are identified as random walk.

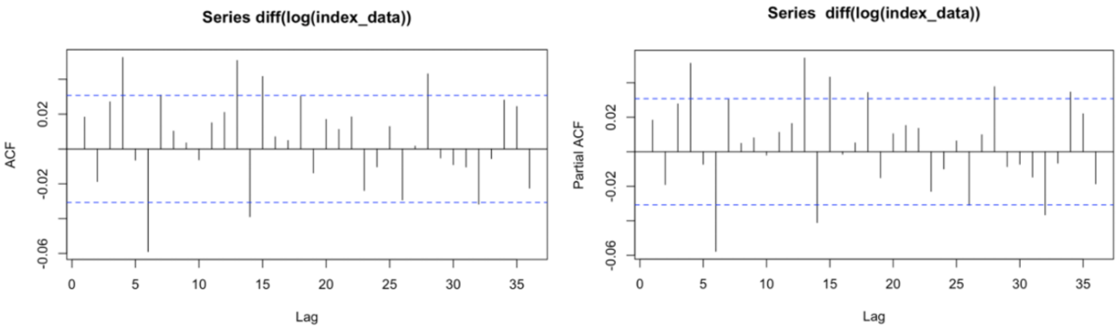

Figure 5: ACF and PACF of first order difference of logarithm of SSE index.

3.4.2. ARIMAX Selection

In addition to regressing the index data independently on ARIMA model, more macro variables are expected to added as explanatory factors. All the regressive explanatory factors are stationary time series data after differencing transformation.

The transformed regressive data are added as regressive explanatory factors respectively in ARIMA (6,1,2) to observe whether the AIC will decrease. If the AIC of integrated ARIMAX model is significantly lower than that of ARIMA model, then the integrated macro variable is reckoned as effective explanatory factor to forecast SSE index.

3.4.3. Ljung-Box Test on Residuals

To ensure that the characteristics of SSE index is exploited sufficiently and the model is effective it is essential to confirm that the residual of model fitting are white noises.

Ljung-Box test is a widely-used method to verify whether the residuals are i.i.d. white noises. It constructs a statistical measure from Chi-squared distribution based on the autocorrelation coefficients and the numbers of sampling. The null hypothesis of Ljung-Box test is that the residuals are i.i.d. white noises. If the p-value of Ljung-Box test given by R is greater than 0.05, then the residuals are identified as i.i.d. white noises, which also implies the effectiveness of model fitting.

3.5. Cross Validation

Time-series cross validation is used to measure the accuracy of our model forecasting. The 4063 days of data are divided into 4063 – x days of training set and x days of validation set, where x = 5, 10, 20,50, 100. The first 4063 - x days of data are used to generate ARIMAX (6,1,2) model, and the subsequent x days of data are used to testify whether the ARIMAX (6,1,2) model can make accurate prediction. As there may exists potential hysteresis or under-reaction of index to the change of macro variables, different lags of regressive series are also set for verification. Since each week has 5 trading days, the lags are set no more than one week (lags = 0, 1, 2, 3, 4, 5). Besides, predicted regressive explanatory series by ARIMA model and real regressive explanatory series are used respectively to make forecast for the index value of x days validation set. The forecast error is given by the percentage of the difference between real index data and forecasting index data. Average forecasting error, MAE and RMSE of each validation are calculated respectively.

4. Results and Discussion

Before model fitting, transformation of logarithm and first-order differencing are used to convert all the time series data into stationary data. Then, ARIMA model with different parameters are enumerated to fit the index data. The lowest AIC of -22331.60 is reached in ARIMA (6,1,2) and the p-value of Ljung-Box test on residuals of ARIMA (6,1,2) is checked simultaneously, which gives a highly significant result of 0.7456. Consequently, (6,1,2) is identified as the best fitted order of ARIMA model on the SSE index data.

Moreover, three different macro variables are respectively added as regressive explanatory factors in ARIMAX model. Among the three updated ARIMAX (6,1,2) models, only the ARIMAX model with exchange rate as explanatory factor can significantly lower the AIC, which drops the AIC from -22331.60 to -22377.95. Nevertheless, by including the other two macro variables into the ARIMA model, the AIC increases inversely. At the same time, it is beneficial to discover that the p-value of Ljung-Box test on residual increases from 0.7456 to 0.7737 when incorporating the exchange rate as an explanatory factor. Consequently, ARIMAX (6,1,2) with one regressive explanatory factor (exchange rate, USD/CNY) is selected as the optimal fitted model.

Table 1: Regression with ARIMAX (6,1,2) errors.

Coefficients	ar1	ar2	ar3	ar4	ar5	ar6	ma1	ma2	xreg
	-0.3806	-0.1683	0.0244	0.0580	0.0200	-0.0541	0.4069	0.1644	0.0914
s.e.	0.2440	0.1951	0.0176	0.0179	0.0244	0.0207	0.2443	0.1978	0.0123
sigma^2 = 0.0002362 log likelihood = 11198.98									
AIC = -22377.95 AICc = -22377.9 BIC = -22314.86									

Subsequent to the model fitting, the author proceeded cross validation on different durations. The numbers of validation dataset are 5, 10, 20, 50, 100 respectively. Besides, prediction made by predicted regressive series and real regressive series are both conducted. After generating ARIMAX model from the training set, the average forecasting error (given by the percentage of the biases with real index data), MAE, RMSE (all are compared with the initial data without taking logarithm and first-order differencing) of different validation sets are calculated.

First of all, discovered from Table 2 and Table 3, the forecast result between prediction made by predicted regressive series and real regressive series are considerably slight, which implies the forecasting accuracy of regressive series. Furthermore, when it comes to the forecast made by predicted regressive series, in short term, when x = 5, lags = 0, the average forecasting error given by percentage is 0.5167%, the MAE is 16.8449, the RMSE is 20.9241, significantly less than when lags is larger than zero. Yet when x = 10, there is little difference between the overall forecast accuracy when lags are positive, all the accuracies are higher than that when lag is zero. In relatively long term, when x = 20, the forecasting errors are larger than those of when x = 50, 100, indicating there exists intensive fluctuations of SSE index that are not captured by the model. Moreover, there is also slight difference between the overall forecast accuracy in relative long term (x = 20, 50, 100) whatever the lags are.

Table 2: Average forecasting errors, MAE, RMSE of forecast on predicted regressive series.

x\lags	0	1	2	3	4	5
5	0.0051669 16.84485 20.92412	0.0057666 18.80327 23.43748	0.0057596 18.78041 23.39914	0.0057670 18.80484 23.43571	0.0057622 18.78893 23.41006	0.0059269 19.32635 24.08182
10	0.0069837 22.50434 26.16495	0.0056986 18.41177 23.40379	0.0057210 18.48191 23.41581	0.0057097 18.44616 23.40474	0.0057210 18.4822 23.42225	0.0056823 18.36037 23.39049
20	0.0225119 72.45996 75.33656	0.0220319 70.91142 73.8715	0.0221626 71.33295 74.27698	0.0220272 70.89627 73.85737	0.0220695 71.03283 73.98813	0.0220272 70.89647 73.85807
50	0.0159675 52.007 63.41346	0.0163294 53.15712 64.91562	0.0163069 53.08525 64.83151	0.0163235 53.13819 64.89796	0.0163059 53.08201 64.82666	0.0162990 53.06022 64.79428
100	0.0138598 45.82445 58.45894	0.0157124 51.9533 65.14716	0.0157851 52.19304 65.38484	0.0157448 52.06011 65.25338	0.0157781 52.17002 65.36361	0.0158166 52.29682 65.48676

Note: In each blank, the line1,2,3 are average forecasting errors, MAE and RMSE respectively.

Table 3: Average forecasting errors, MAE, RMSE of forecast on real regressive series.

x\lags	0	1	2	3	4	5
5	0.0051364 16.73226 19.80457	0.0057469 18.73919 23.35317	0.0057482 18.74368 23.39958	0.0057549 18.76543 23.39585	0.0057713 18.81865 23.44669	0.0059334 19.34711 24.07219
10	0.0067684 21.80188 25.55354	0.0056832 18.36365 23.33282	0.0057160 18.46565 23.43359	0.0057049 18.43077 23.3752	0.0057227 18.48821 23.4425	0.0057028 18.42552 23.48377
20	0.0221290 71.22964 74.1769	0.0220258 70.89224 73.83856	0.0221446 71.275 74.21625	0.0220308 70.90816 73.86639	0.0220540 70.98288 73.94533	0.0220951 71.11457 74.10088

Table 3: (continued).

50	0.0155316 50.5936 62.38491	0.0163220 53.13327 64.88273	0.0162877 53.02253 64.77616	0.0163299 53.15893 64.91868	0.0162969 53.053 64.79676	0.0163370 53.18256 64.95762
100	0.0142267 47.0383 59.82499	0.0157258 51.99774 65.20324	0.015796 52.23022 65.41656	0.0157405 52.04604 65.24346	0.0157844 52.19069 65.38278	0.0157925 52.21603 65.36555

Note: In each blank, the line1,2,3 are average forecasting errors, MAE and RMSE respectively.

As the average forecasting error is around 0.5% ~0.6% in short period, it concludes that the model can fit the real data well especially in short term. The highest forecast accuracy is obtained when x = 5, lags = 0, which gives an average forecasting error by 0.5167%. In long run, the average forecasting error is over 1%, the cumulated error will be enlarged massively as time goes. Besides, both the MAE and RMSE increase first and subsequently drop, which implies that in relatively long term, the model cannot capture a peak or trough of the fluctuation of index, but it can give out a general trend of SSE Index movement.

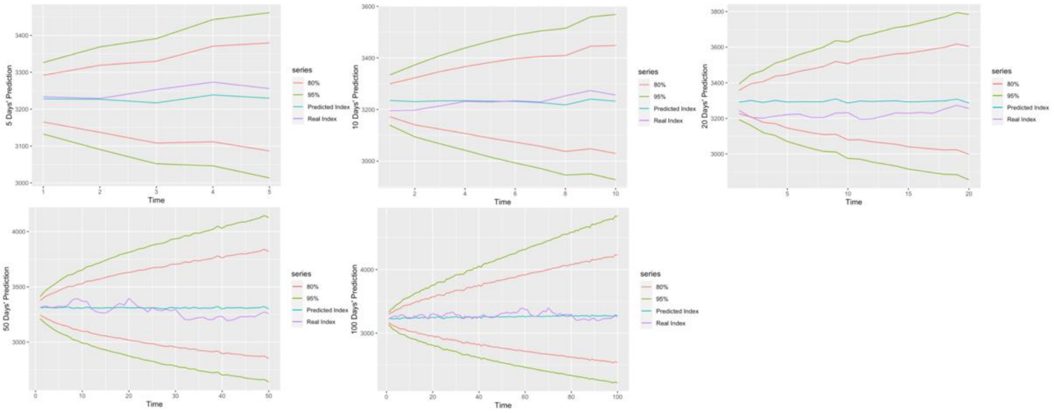

Figure 6: 5, 10, 20, 50, 100 days' forecasting of cross validation.

5. Conclusion

This research aims to predict the SSE index, which is widely-used as a representative of the overall performance of China's stock market. By taking logarithm and first order differencing, the index data is transformed into stationary data and passes ADF test. Based on the AIC and BIC rules, the best fitted ARIMA (6,1,2) model was confirmed.

Due to the similar pattern of the fluctuations of SSE index and exchange rate around the subprime mortgage crisis, the author speculates that exchange rate might be a macro variable to predict the SSE index. Thus, more macro variables (interest rate, inflation rate, real interest rate) are added in as explanatory factors.

By integrating these macro variables into the ARIMA (6,1,2) model and establishing ARIMAX model respectively, it is found that the explanatory factor of exchange rate can significantly decrease the AIC of the ARIMAX model and enlarge the residuals' p-value of Ljung-Box test.

Cross validation is proceeded to test the prediction ability of the model. Besides, distinct lags of regressive series are also set. By comparing between the real index value from the validation set and the forecasting value of ARIMAX model generated from the training set and calculating the average forecast error, MAE and RMSE, it is found that the model can fit the real data well in short term. In relatively long run, the model predicts the trend of SSE Index, but it cannot forecast the fluctuation of index.

This result shows the effectiveness of incorporation of macro variables into the prediction of performance of stock market. In this research, the integration of exchange rate into ARIMA model increases the fitting degree and forecast accuracy. Therefore, for researchers and investors, it is wise to consider the expected exchange rate as an important factor to make prediction of stock market's future performance.

The innovation of this study is the discovery of the resembling patterns of fluctuations between SSE index and macro variables (especially exchange rate), and the introduction of macro variables as regressive explanatory factors into ARIMA model to obtain ARIMAX model. Nevertheless, there are also certain limitations of this research. This research only discusses the exchange rate of the CNY, the inflation rate and interest rate, without incorporating other influencing factors into the model. In the future, more other indicators can be introduced to analyze the impact of macro variables on the SSE Index and different models can be taken into consideration thus further deepening the research content.

References

[1] *Dornbusch R., & Fischer S. (1980). Exchange Rates and the Current Account. American Economic Review, 70(5), 960-971*

[2] *Granger Clive.W.J, Huangb Bwo-Nung, & Yang Chin-Wei. (2000). A Bivariate Causality between Stock Prices and Exchange Eates: Evidence from Recent Asian Flu. Quarterly Review of Economics and Finance, 40(3), 337-354*

[3] *Ba Shusong, &Yan Ming. (2009). The Dynamic Relationship between Stock Prices and Exchange Rates--Empirical Evidence. Nankai Economic Studies, 46-62*

[4] *Wang Bo, Liao Hui, & Ma Junlu. (2012). Rate, Interest Rate and the Stock Prices. Financial Economics Research: 35-46*

[5] *Liu Jiming. (2019). Analysis of CNY Exchange Rate and Stock Market Spillover Effect. Hebei Finance, 55-59+64*

[6] *Zhang Bin, Feng Sixian, Li Xindan, & Wang Huijian. (2008). Echange Rates and Stock Prices Interactions in China: An Empirical Studies after 2005 Exchange Rate Reform. Economic Research, 43(09), 70-81+135*

[7] *Fama E. F., & French K. R. (1992). The Cross-Section of Expected Stock Returns. The Journal of Finance, 47(2), 427-465.*

[8] *Masih A. M. M., & Masih R. (2001). Dynamic Modeling of Stock Market Interdependencies: An Empirical Investigation of Australia and the Asian NICs. Review of Pacific Basin Financial Markets and Policies, 04(02),235-264.*

[9] *Dickinson, & David G. (2000). Stock Market Integration and Macroeconomic Fundamentals: An Empirical Analysis, 1980-95. Applied Financial Economics, 10(3),261- 276.*

[10] *Wong W. K., Khan H., & Du J. (2006). Do Money and Interest Rates Matter for Stock Prices? An Economic Study of Singapore and USA. The Singapore Economic Review, 51(01),31-51.*