Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/50/20230559

# *China's Urban and Rural CPI Prediction Based on ARIMA Model*

**Yuzhi Liu**[1,a,*]

[1]*Khoury School of Computer Science, Northeastern University, Boston, MA, U.S.*
*a. liu.yuzhi@northeastern.edu*
**corresponding author*

***Abstract:*** Inflation represents the continuous rise of the overall price level of a country. In severe cases, it may cause an imbalance between social supply and demand and lead to a crisis of currency confidence. Therefore, it is necessary to measure and predict the level of inflation. The CPI index is an important indicator to measure the level of inflation, which can largely reflect the national economic situation in a certain period. This paper conducts research by selecting the urban and rural CPI data of the National Bureau of Statistics from January 2007 to June 2023, a total of 198 months. After data processing and inspection, this paper use ARIMA model to forecast. The experimental research results show that the ARIMA (12,0,1) model and the ARIMA (12,0,0) model have good predictive effects on the CPI index of cities and villages respectively. In the short term, the ARIMA model can accurately predict the changing trend of the CPI index, with an error rate of less than 0.5%. The model predicts that China's urban and rural inflation from June 2023 to June 2024 will be stable and improving overall.

***Keywords:*** Time-Series analysis, CPI prediction, ARIMA model

## 1. Introduction

### 1.1. Research Background

After the outbreak of the epidemic in 2020, the world economy has encountered severe challenges. Consumption, investment, and import and export have been severely impacted. Consumer prices in residents' lives also showed a trend of polarization. Due to the backlog of goods, the price level of some commodities also dropped sharply. However, the demand for some commodities has soared during the epidemic. Because of short supply, so that the price has also soared. Since 2023, the situation of epidemic prevention and control has continued to improve, economic activities have resumed expansion after the continuous recession, commodity prices have continued to rise, and inflation has resurfaced. In such a background, it is important and useful to study and forecast China's inflation rate. Inflation refers to the depreciation of a country's currency due to the overall and continuous rise in prices. The direct cause of this phenomenon is usually that the actual amount of currency issued by a country exceeds the amount of money it needs. The Consumer Price index (CPI) is generally used to measure the degree of inflation. This index reflects the changes in the price level of consumer goods and services generally purchased by urban and rural households [1]. If the CPI index rises continuously and comprehensively for a period, it indicates that inflation has occurred.

DOI: 10.54254/2754-1169/50/20230559

China divides the country into two parts, urban and rural, in the division of regions, and the economic levels of these two different regions are different. In the context of the Covid-19 epidemic, the lives of urban and rural residents have changed again. Different incomes and prices between urban and rural areas may cause the gap in living standards to widen again, causing social unrest [2,3]. Therefore, the accurate prediction of urban CPI index and rural CPI index can enable the government to make timely and accurate judgments on the economic situation. It also helps the government to check if the existing policies cannot cope with future development, so as to promote economic flourish and high-speed development.

### 1.2. Literature Review

Chinese academic circles have plenty research results on inflation forecasting and ARIMA model forecasting. In terms of inflation rate prediction, Jiayue Liu et al. proposed a very complex non-linear autoregressive neural network forecasting [4]. In terms of the use of the ARIMA model, Lingxiao Zhao et al. used this model to forecast the PM2.5 content in Beijing [5]. Hongye Cai and Wenxuan Qiu use ARIMA (2,2,3) model to predict Shenzhen GDP in the next 5 years and the relative error between the actual data and the prediction results is less than 3 percent [6]. Lu Bai, Ke Lu et al. combined ARIMA (0, 1, 1) (0, 1, 0) $^{[12]}$ model with the covariates of PM2.5, SO2, and CO to form the SARIMAX model, which also has an error of less than 3 percent [7]. In 2021, Yao Ma and others once analyzed and predicted the CPI of China, the United States, and Germany. Its model is extremely accurate, and the errors of the three models are all less than 1 percent [8]. Yurou Chen made a 2-month prediction of China's core CPI by using the ARIMA model with coefficients (1,1,1), but did not classify the CPI data [9]. Xiaodan Sun used the ARIMA (13,0,0) model to make a CPI index prediction of China and concluded that the CPI index has a long lagging term [10]. Since there are few studies on the distinction between urban and rural CPI trends, this paper will make a short-term prediction of China's urban and rural CPI respectively and give corresponding suggestions.

### 1.3. Research Contents and Framework

This article will rely on the R language and the ARIMA model to carry out model fitting and forecasting of China's urban and rural comprehensive consumer price index from January 2007 to the present. The specific research process will be divided into three parts. The first part is data collection and data processing. The second part is model fitting and model selection. The third part is the solution of the model and the test of the results. In the first stage, this research will select the urban CPI index and rural CPI index provided by the National Bureau of Statistics of China. After data collection, all data are analyzed and processed. First, the research makes up the insufficient data, and then carry out the ADF stationarity test on the made-up data. According to the test results, further processing is performed on the data, such as differencing, Box-Cox transformation, etc. In the second part, this paper will use the improved data to generate mathematical model, and at the same time determine the possible values of (p, q) in the ARIMA (p, d, q) model based on the ACF image and PACF image. Next, the study will test all possible values one by one and record the result that passed the Ljung-Box test. After comparing the AIC and BIC values of all results, this study will take the coefficient (p, q) corresponding to the best result as the optimal model coefficient. In the third part, the research uses the previous model to make 15-month forecasts, compare the forecast results with the actual data and calculate the error rate under different time lengths, and finally evaluate the model based on the error rate.

DOI: 10.54254/2754-1169/50/20230559

## 2. Methodology

### 2.1. Data Collection

The research first needs to collect data. On the official website of the National Bureau of Statistics of China (http://data.stat.gov.cn), there are detailed statistics on the data of Urban CPI and Rural CPI, which are continuously updated. The data on the official website are uniformly provided by the National Bureau of Statistics, with high reliability and accuracy, so these data are suitable for analysis and research. When selecting different types of statistical data, the 100% statistical method of the previous month is the most suitable for the modeling needs of time series, and it is convenient for subsequent forecasting activities, so it was selected. Since the National Bureau of Statistics of China changed the statistical classification of the Consumer Price Index (CPI) in 2016, it is necessary to integrate old and new data when collecting urban and rural data, and to discard data with changed classification. Finally, the urban and rural data from December 2006 to June 2023 were obtained. There were no significant deviations from the data. Both urban and rural data were missing in October 2007, August 2008, March 2009, and June 2009, and were temporarily vacant during the collection stage without further processing (in Figure 1).

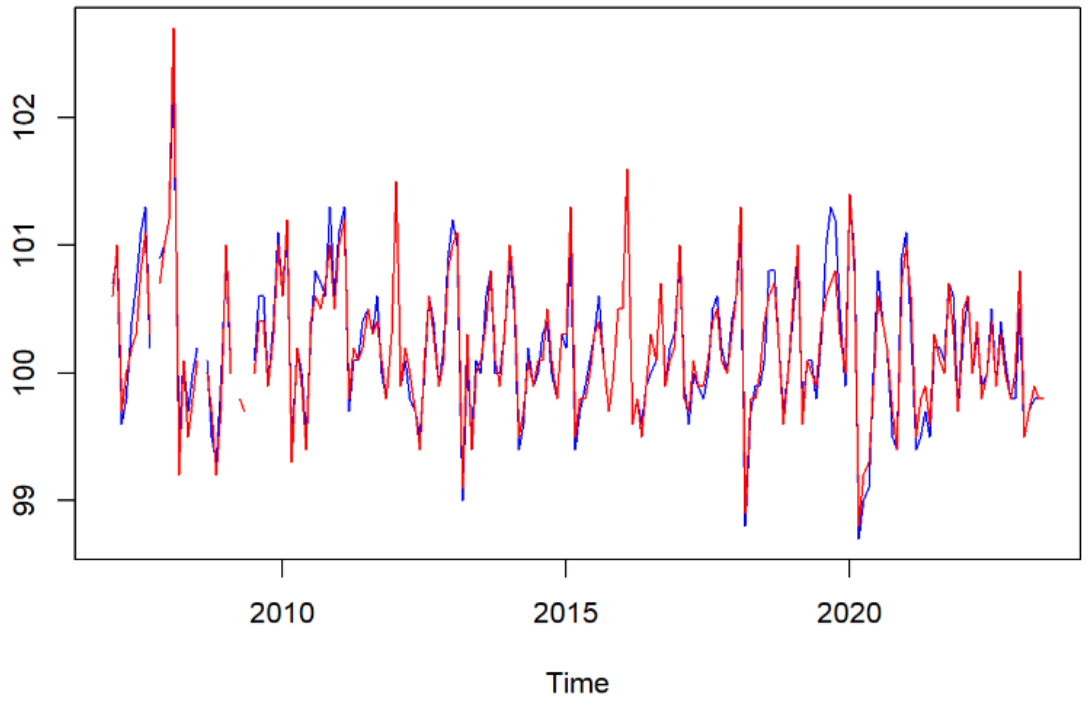


Figure 1: Urban and rural data visualization (red: urban, blue: rural) (Photo credited: original).

### 2.2. Model Selection

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. ARIMA model has plenty of advantages. Its model is very simple, only endogenous variables are needed and No need to resort to other exogenous variables. At the same time, the ARIMA model has stable performance and accurate prediction, which is very suitable for data with linear assumptions. Therefore, this paper will choose ARIMA model for research. However, the ARIMA model also has

DOI: 10.54254/2754-1169/50/20230559

some shortcomings. The prediction of this model has high data requirements, and the model requires complete time series data and no missing values. In practical applications, data is often missing and incomplete, which affects the effect of the model.

## 2.3. Data Completion

For the ARIMA model, one of the conditions for its use is that the data cannot be missing. However, there are four discontinuities in the data in this paper. Therefore, in the data processing stage, the missing data must be filled first. The program filled missing values by averaging the existing values before and after the value (in Figures 2-3).

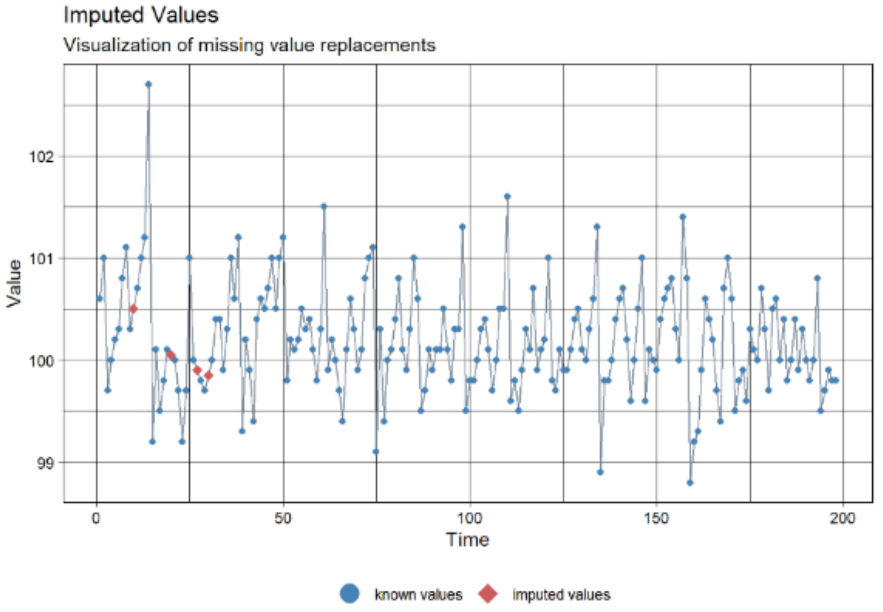


Figure 2: Complete data (urban) (Photo credited: original).

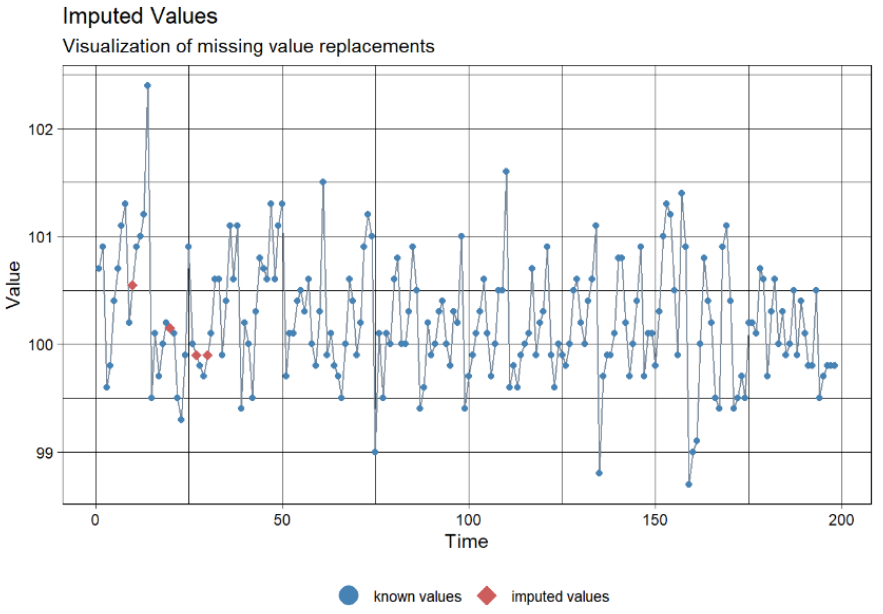


Figure 3: Complete data (rural) (Photo credited: original).

DOI: 10.54254/2754-1169/50/20230559

## 2.4. ADF Test

A significant prerequisite of ARIMA model fitting is the stationarity of the time series data. Therefore, all the time series data are required to be transformed into stationary time series data. In statistics, Augmented Dickey-Fuller test (ADF test) can test whether an autoregressive model has unit root, which is equivalent to the stationarity of time series data. Consequently, all the data must pass ADF test before being integrated into ARIMA model. After proceeded ADF test on both data set, the p-value given by both are smaller than 0.01, thus this paper rejected the null hypothesis and concluded that the rural CPI data and the urban CPI data are stationary time series data. The two sets of data rural CPI and urban CPI used in this article have passed the ADF test.

## 2.5. BOX-COX

Box-Cox transformation is a data transformation commonly used in statistical modeling. It is used when continuous response variables do not meet the Normal distribution. After the transformation, unobservable errors and the correlation between predicted variables can be reduced to a certain extent. The best lambda on rural CPI data evaluated by R is given by -0.99, close to -1. Therefore, in subsequent predictions we confirm the value of lambda as -1.

## 2.6. Model Fitting

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

Typically, the study should select AR(p) where there is a significant spike at lag p in PACF plot and should select MA(q) where there is a significant spike at lag q in ACF plot. Therefore, all the possible (p, q) selections can be given by the ACF and PACF plots of the transformed data. Specifically, significant values in rural CPI of p are 1, 11, 12(in Figure 4). and significant values of q are 1, 9, 10, 11, 12, 15(in Figure 4). Meanwhile, significant values in urban CPI of p are 1, 11, 12, 14(in Figure 4). Significant values of q are 1, 9, 10, 11, 12(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 the research obtained the lowest AIC order is (p, q) = (12,1) for rural CPI and (p, q) = (12,0) for urban CPI.

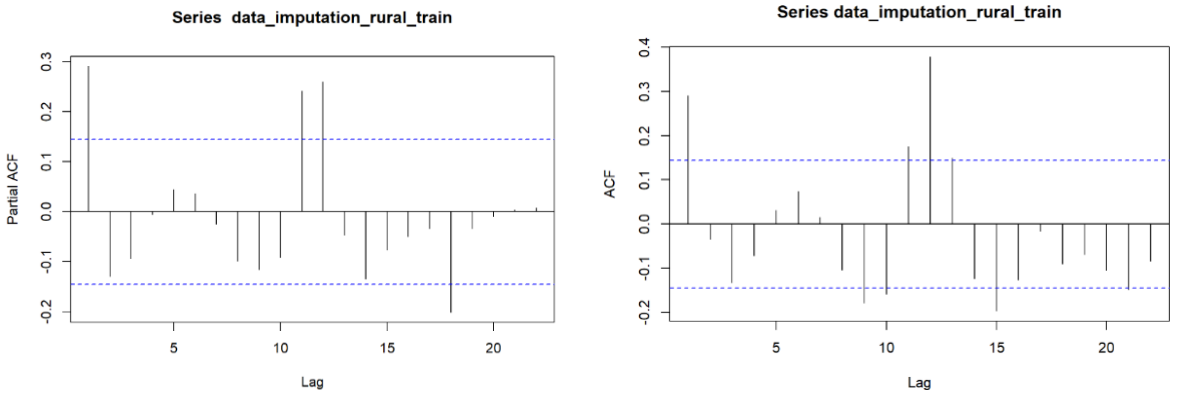


Figure 4: Rural PACF and ACF (left: PACF, right: ACF) (Photo credited: original).

DOI: 10.54254/2754-1169/50/20230559

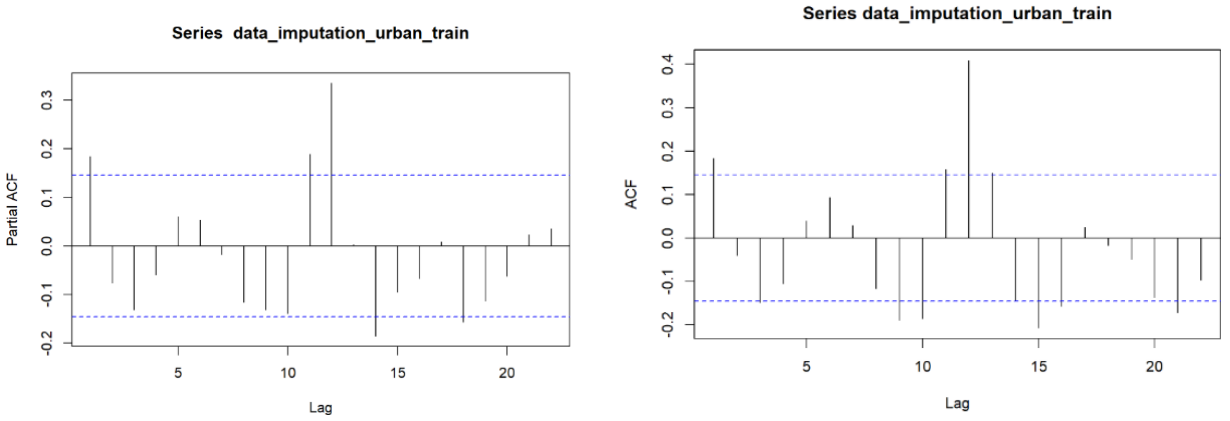


Figure 5: Urban PACF and ACF (left: PACF, right: ACF) (Photo credited: original).

## 2.7. Ljung-Box Test on Residuals

To ensure that the research exploit the characteristics of Both CPI data index sufficiently and our model is effective, what we need to do is 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 iid. white noises. The null hypothesis of Ljung-Box test is that the residuals are white noises. If the p-value of Ljung Box test given by R is greater than 0.05, then the research accepts the null hypothesis and recognize the residuals have passed the test, which also implies the effectiveness of model fitting.

Time-series cross validation is used to measure the accuracy of our model forecasting. The author divide the 198 days of data into 183:15 portions. The first 183 days of data are used to generate ARIMA (12,0,1) and ARIMA (12,0,0) model, and the subsequent 15 days of data are used to testify whether the model can make accurate prediction. The forecast error is given by the percentage of the difference between real CPI data and forecasting CPI data.

## 3. Results and Discussion

In the previous section, ARIMA model with different parameters are enumerated to fit the rural CPI data. The model reached the lowest AIC of 289.86 in ARIMA (12,0,1) and then simultaneously checked the p-value of Ljung-Box test on residuals of ARIMA (12,0,1). The result is 0.1988 which is a significant result. Consequently, the research can confirm that (12,0,1) is the best fitted order of ARIMA model on the rural CPI data. For urban CPI data, the research did the same thing. The model reached the lowest AIC of 273 in ARIMA (12,0,0) and then simultaneously checked the p-value of Ljung-Box test on residuals of ARIMA (12,0,0). The result is 0.09758 which is also a significant result (in Figures 6-7).

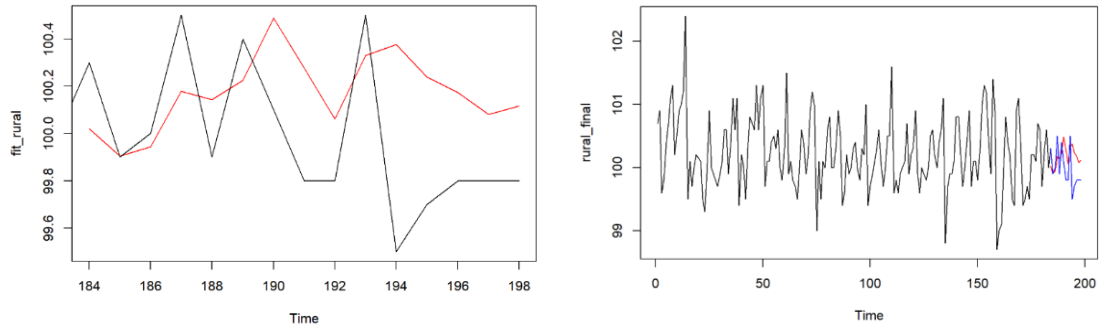


Figure 6: Rural CPI prediction result (left: detail, right: overall trend) (Photo credited: original).

DOI: 10.54254/2754-1169/50/20230559

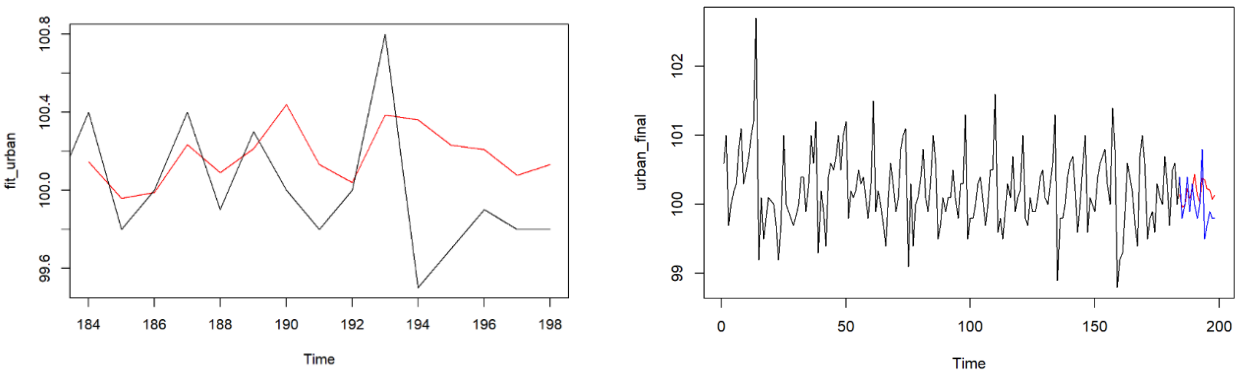


Figure 7: Urban CPI prediction result (left: detail, right: overall trend) (Photo credited: original).

In last part, the researcher proceeded cross validation on different durations. After generated ARIMA model from the training set, the author calculated the average forecasting error percentage of different lengths of validation set. In short term, when predicting 5 months CPI, the average forecasting error given by percentage are 0.1806% and 0.1559% for rural and urban data respectively; when predicting 10 months CPI, the average forecasting error given by percentage are 0.2372% and 0.2088%; when predicting 15 months CPI, the average forecasting error given by percentage are 0.3177% and 0.2937%

Table 1: Urban prediction error.

| Time Series: | | Start = 184 | | End = 198 | | Frequency = 1 | |
|---|---|---|---|---|---|---|---|
| [1] | 0.9974686 | 1.0015800 | 0.9998839 | 0.9983408 | 1.0019069 | 0.9991393 | 1.0044105 |
| [8] | 1.0033223 | 1.0003986 | 0.9959049 | 1.0086481 | 1.0053299 | 1.0030784 | 1.0027786 |
| [15] | 1.0033400 | | | | | | |

Table 2: Rural prediction error.

| Time Series: | | Start = 184 | | End = 198 | | Frequency = 1 | |
|---|---|---|---|---|---|---|---|
| [1] | 0.9972054 | 1.0000495 | 0.9994374 | 0.9968098 | 1.0024310 | 0.9982640 | 1.0038769 |
| [8] | 1.0047803 | 1.0026145 | 0.9983148 | 1.0088105 | 1.0053982 | 1.0037454 | 1.0028133 |
| [15] | 1.0031639 | | | | | | |

As the average forecasting error in 5 months is around 0.15%, in short period, this study conclude that the model can fit the real data well. In a longer period, the average forecasting error is getting larger, but the data for the next 15 months are still all within the 95% range. The forecasting error in 15 months is about 0.3% which is larger than the 5 months error. As a result, the research concludes that in relatively long term, the model can predict the CPI, but it cannot forecast a fluctuation of CPI index properly.

## 4. Conclusion

In these days, inflation has become a more and more serious problem in many countries. To evaluate the inflation rate, the CPI index data is the best choice. This research aims to predict the CPI index, which is widely used as a representative of the overall performance of a country's economics. Both data set passed the ADF test. Given the ACF and PACF of the transformed data we selected potential combination of (p, q). Based on the AIC and BIC rules, the best fitted ARIMA (12,0,0) model and ARIMA (12,0,1) model were confirmed respectively. The (12,0,0) model which was the optimal

DOI: 10.54254/2754-1169/50/20230559

choice of the urban CPI has an AIC of 273 and the residual passes the Ljung-Box Test with p-value of 0.09758. The (12,0,1) model which was the optimal choice of the urban CPI has an AIC of 289.86 and the residual also passes the Ljung-Box Test with p-value of 0.1988. After training the ARIMA models, the research divided the dataset into training set and validation set to test the prediction ability of the model. By comparison between the real index value from the validation set and the forecasting value of ARIMA model generated from the training set, we found that both models can fit the real data well in short term. In long run, those models do predict the trend of CPI index, but the error is getting larger than the short term. A more accurate long-term CPI forecasting method still needs further exploration. According to the results predicted by the model, we can see that China's future inflation rate is still within a controllable range. However, the government should still pay attention and actively take the following measures to prevent economic turmoil in the post-epidemic era. First, issue founds and increase subsidies. Second, carry out advanced infrastructure construction appropriately. Third, do a good job of reasonable guidance, increase the state's regulation and control, reduce market liquidity, and slow down the upward pressure on prices.

## References


[1] *Lei, P. (2014). Sequential analysis of CPI in China based on SARIMA model. Statistics and Application, 410, 32–34.*

[2] *Zheng, X., & Zhang, T. (2018). Analysis and prediction of urban-rural income gap based on ARIMA model -- Take Yunnan Province as an example. Rural Economy and Technology, 443, 139–141.*

[3] *Wu, Q. (2020). Prediction Analysis of Income Gap Between Urban and Rural Residents Based on ARIMA Model ——A Case Study of Hubei Province. JOURNAL OF HARBIN UNIVERSITY, 42, 31–34.*

[4] *Liu, J., Ye, J., & E, J. (2023). A multi-scale forecasting model for CPI based on Independent Component Analysis and non-linear autoregressive neural network. Physica A: Statistical Mechanics and Its Applications, 609.*

[5] *Zhao, L., Li, Z., & Qu, L. (2022). Forecasting of Beijing PM2.5 with a hybrid Arima model based on integrated AIC and improved GS fixed-order methods and seasonal decomposition. Heliyon, 8(12).*

[6] *Cai, H., & Qiu, W. (2022). Prediction analysis of Shenzhen GDP based on Arima model and implementation in R language. Academic Journal of Computing & Information Science, 5(10).*

[7] *Bai, L., Lu, K., Dong, Y., Wang, X., Gong, Y., Xia, Y., Wang, X., Chen, L., Yan, S., Tang, Z., & Li, C. (2023). Predicting Monthly Hospital outpatient visits based on meteorological environmental factors using the Arima model. Scientific Reports, 13(1).*

[8] *Ma, Y., Sun, Z., & Zou, Y. (2021). Analysis and forecast of CPI in China, the United States and Germany based on ARIMA model. ECONOMIC RESEARCH GUIDE, 466, 1–4.*

[9] *Chen, Y. (2020). Short-term forecast of China's core CPI based on ARIMA model. Time-Honored Brand Marketing, 37–38.*

[10] *Sun, X. (2021). Analysis and Forecast CPI Index of China Based on ARIMA Model. Circulation Economy, 15–17.*