Proceedings of the 2023 International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/22/20230320

A Prediction of the Construction Area of Residential Real Estate in Shenzhen Based on ARIMA Model

Zixuan Guan[1,a,*]

[1]*Department of Mathematics with Computer Science, Guangdong Technion - Israel Institute of Technology (GTIIT), Shantou, China*
a. guan09236@gtiit.edu.cn
**corresponding author*

Abstract: The real estate markets are an important component of Chinese economy structure which is worth to study and predict. This paper uses the autoregressive integrated moving average (ARIMA) models to predict the construction area of residential real estate in Shenzhen. The data is given by a convinced and reliable company CRIC. CRIC focuses on the Chinese real estate market for years. With the dataset from CRIC and based on the ARIMA methodology, the author gives a roughly prediction which shows that the construction area of residential real estate in Shenzhen will continuously rise in a short term. Serious of tests are invoked, such as white noise test and unit-root test etc. The author then draws a conclusion by those tests, which implies that the construction area of residential real estate in Shenzhen is worth to predict statically. And in the real-life dimension this prediction can provide a valuable reference to politicians and investors. For politicians, the prediction can help them make policy. And for investors, the prediction can help them maximize their benefits.

Keywords: ARIMA, real estate, Shenzhen

1. Introduction

The real estate markets in China are becoming superheating. Under this situation, it is very important to analysis the markets. As one can see, the area provided in the future is based on the present construction area. And the popular of the real estate markets mostly depends on the residential real estate. Therefore, the construction area of residential real estate is worth analyzed and predicted. Shenzhen, as one of the first-tier cities in China, is worth studying with no doubt.

The prices of residential real estate in Chinese have soared [1]. However, the papers of Chinese real estate markets are abundant, especially the construction area of Chinese real estate markets. And this paper roughly predicts the construction area of residential real estate in Shenzhen in a short time, which is based on the ARIMA model. The ARIMA model is reliable. Many different prices have been predicted, such as stock [2], bitcoin [3], or house [4]. And it is well performed when doing a short-term prediction [5].

This paper shows the construction area of residential real estate in Shenzhen from 1991 to 2020 with the unit 10,000/m2 in line chart, which construct a time series. And various of data processing methods are provided in order to evaluate whether the time series is worth to predict. The result is the time series can be predicted roughly. Then the author identifies the model and provides a prediction

DOI: 10.54254/2754-1169/22/20230320

of the construction area of residential real estate in Shenzhen with the unit 10,000/m2. This prediction is based on the ARIMA model. The area is predicted to rise which means that the supply of residential real estate will increase. Since the construction area will come into the market eventually.

2. Methodology

2.1. Source of Data

A dataset, from CRIC, is used to describe the construction area of residential real estate in Shenzhen. The dataset consists of construction area of residential real estate in Shenzhen from 1991 to 2020. The independent variable is time, and the dependent variable is the construction area of residential real estate in Shenzhen. For each year, the dataset is performed as a numerical number and the unit is 10,000/m2.

2.2. Data Processing and Models

2.2.1. Turning the Series into Stationary

Step 1: Natural logarithm (NL): One can use the nature logarithm function to reduce variance [6].

$$\boldsymbol{X = ln(X)} \tag{1}$$

Step 2: Differentiation: One can use the differentiation to remove trend signals and also to reduce the variance [7].

$$\boldsymbol{x_k = x_0 + kh, (k = 0, 1, \dots, n)} \tag{2}$$

$$\boldsymbol{\Delta f(x_k) = f(x_{k+1}) - f(x_k)} \tag{3}$$

2.2.2. Stability Test

Step 1: Augmented Dickey-Fuller test (ADF): One can proved that the time series is stable if and only if the solution x_i of

$$\boldsymbol{1 - \Phi_1 x - \Phi_2 x^2 - \dots - \Phi_k x^k} \tag{4}$$

satisfied that $\forall i, |x_i| > 1$.

If the ADF shows that the time series is stable, then one can calculate the ACF and PACF to confirm the order of the ARIMA model [8].

Step 2: Autocorrelation function (ACF): The autocorrelation function of stationary sequence is independent of any particular t time and is a function of time interval h.

$$\boldsymbol{\rho_h = \rho(X_t, X_{t+h}) = \frac{cov(X_t, X_{t+h})}{\sigma_t \sigma_{t+h}}} \tag{5}$$

Step 3: Partial autocorrelation function (PACF): The partial autocorrelation function can describe the random process structure. The Φ_{kj} represents the number j coefficient of the autocorrelation equation of order k. Then the autoregressive model (AR) is shown as

$$\boldsymbol{x_t = \Phi_{k1} x_{t-1} + \Phi_{k2} x_{t-2} + \dots + \Phi_{kk} x_{t-k} + u_t} \tag{6}$$

Notice that Φ_{kk} is the last coefficient. And Φ_{kk} can be recognized as a function of delayed k. Then $\Phi_{kk}, k = 1,2, \dots$ is the partial autocorrelation function.

DOI: 10.54254/2754-1169/22/20230320

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t - \sum_{j=1}^{q} \theta_{t-j} \epsilon_j, (\phi_p \neq 0, \theta_q \neq 0) \tag{7}$$

$$E(\epsilon_t) = 0, Var(\epsilon_t) = \delta_\epsilon^2, E(\epsilon_t \epsilon_s) = 0, s \neq t, (E(X_s \epsilon_t) = 0, \forall s < t) \tag{8}$$

2.2.3. White Noise Test

Ljung-box test [9]

$$Q(m) = T(T+2) \sum_{i=1}^{m} \frac{\hat{\rho}_i(a_t^2)}{T-i} \tag{9}$$

where T is sample size, m is a constant, a_t is residuals and $\hat{\rho}_i(a_t^2)$ is the ACF of order i with respect to a_t^2.

2.2.4. ARIMA Model Identification

Bayesian Information Criterions:

$$BIC = ln(n)k - 2ln(L) \tag{10}$$

where L is the prediction function, n is the sample size and K is the number of coefficients.
It is helpful to identify the ARIMA model [10].

2.2.5. Model Checking

QQ-plot: One can use QQ-plot to check that if the residuals follow the Gaussian distribution $N(\mu, \sigma^2)$.

3. Results and Discussion

3.1. Data Visualization

Fig. 5 shows the numerical features in the dataset. The ordinate represents area, and the abscissa represents year. As one can see, the area is on the rise. Therefore, one can assume that the time series is worth predicted. However, it should be illustrated more further by rigorous approach.

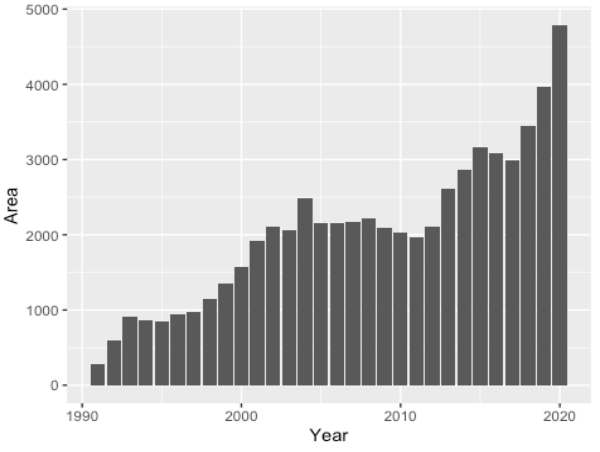

Figure 1: Construction area of residential real estate in Shenzhen (Photo credit: Original).

DOI: 10.54254/2754-1169/22/20230320

Fig. 2 shows the construction area of residential real estate in Shenzhen in line chart. The increment of the construction area of residential real estate in Shenzhen is given by the difference of first order.

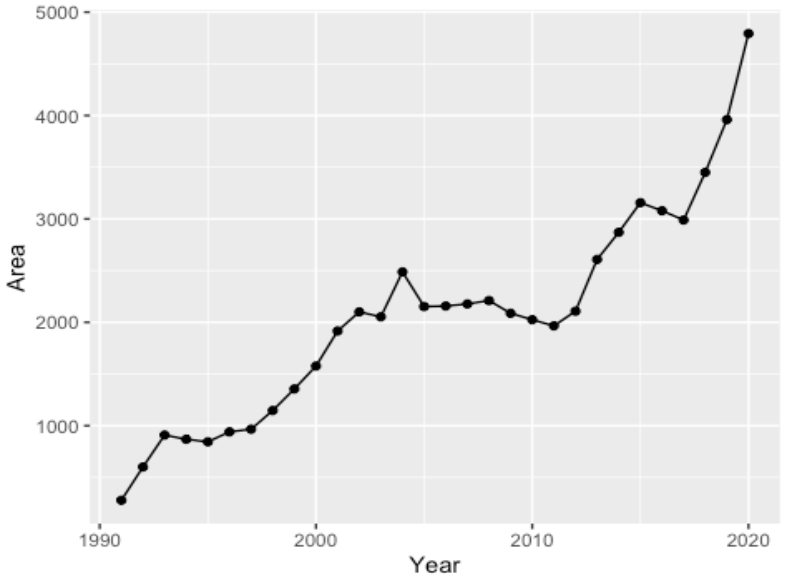

Figure 2: Construction area of residential real estate in Shenzhen (Photo credit: Original).

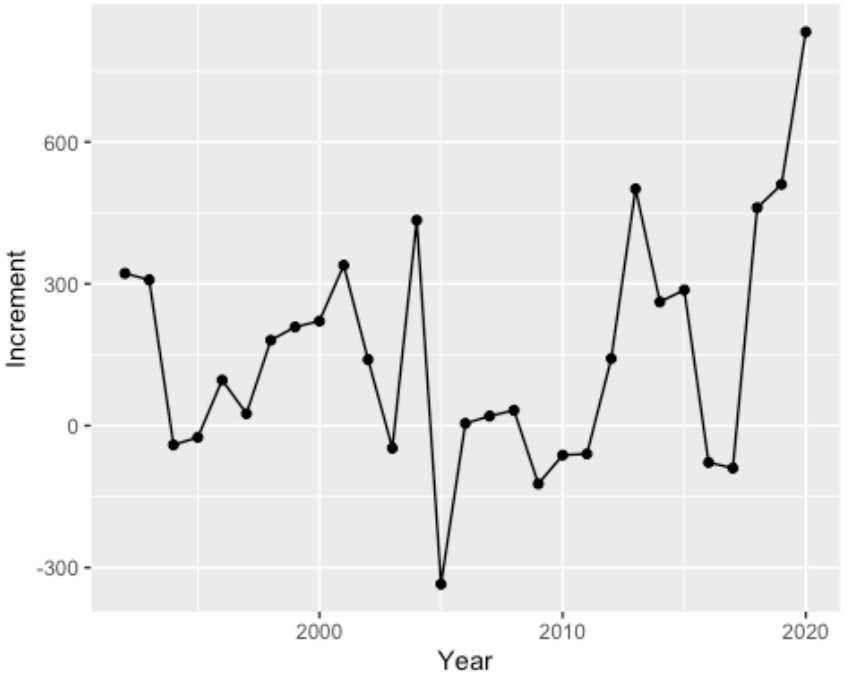

Figure 3: Construction area of residential real estate in Shenzhen after difference (Photo credit: Original).

The increment fluctuates periodically over time and the growth rate of the construction area of residential real estate in Shenzhen has increased (see Fig. 3). However, it is not a stationary time series, and it cannot be used for effective model prediction and analysis.

DOI: 10.54254/2754-1169/22/20230320

3.2. Statistical Analysis

3.2.1. Natural Logarithm

The author takes nature logarithm to the original data. Fig. 4 shows the data after the logarithm. The time series after logarithm is denoted as “szl”.

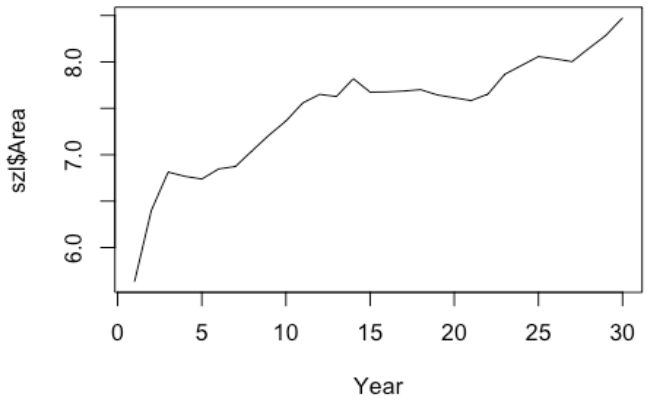

Figure 4: Construction area of residential real estate in Shenzhen after logarithm (Photo credit: Original).

3.2.2. Stability Test before Difference

From ACF (see Fig. 5) one can see that the function trails and waves like cosine function.

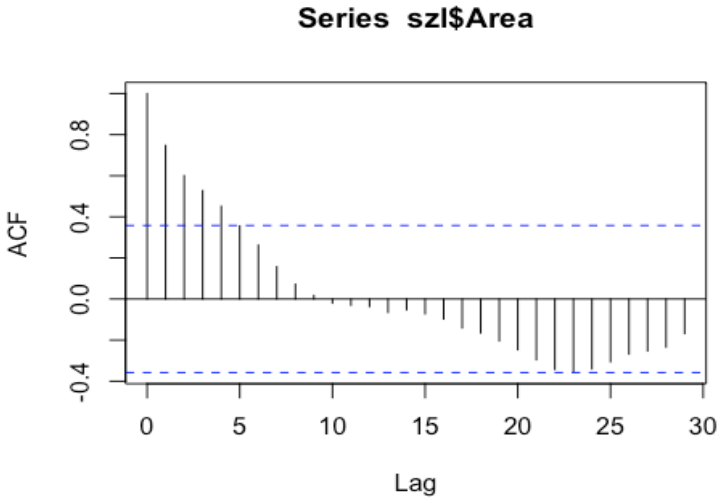

Figure 5: ACF of the time series after logarithm (Photo credit: Original).

And the PACF (see Fig. 6) truncates. However, the value of ACF is out of confidence interval until the 6th order lag. If a time series is stable for a specific probability, the ACF value will be in the confidence interval after the 1st order lag under the specific probability. In author’s case, the probability of the time series being unstable is 95%.

DOI: 10.54254/2754-1169/22/20230320

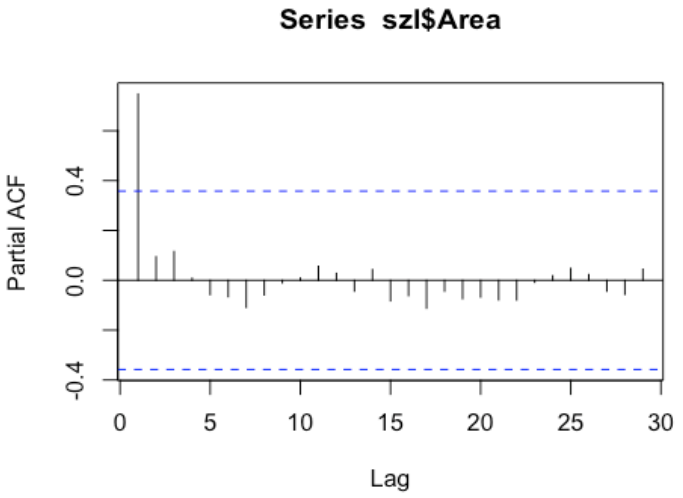

Figure 6: PACF of the time series after logarithm (Photo credit: Original).

What's more the Augmented Dick-Fuller Test shows that the p-value is 0.9842. In another word, the probability such that the time series is unstable is 98.42%. in the author's case, the p-value show be less than 0.05. Therefore, the author differences the data in order to get a stable time series.

3.2.3. Difference

Fig. 7 shows the time series after difference. The time series is denoted as "diffa". One can suppose that the time series is fluctuated around a constant. In order to further verify this assumption, the author shows the ACF and PACF as follows.

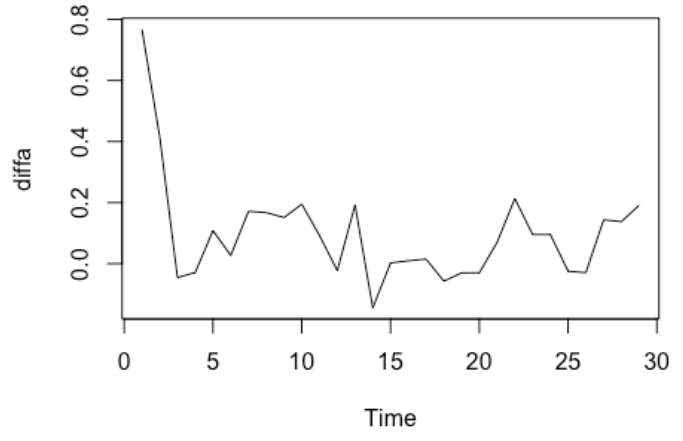

Figure 7: The time series after logarithm and difference (Photo credit: Original).

3.2.4. Stability Test after Difference

As one can see, the ACF value is in the confidence interval after the 1st order lag (see Fig. 8), and the PACF value truncates (see Fig. 9). And the p-value of unit root test is 0.00006582 which is less than 0.05. Therefore, the unit root test passed, i.e., the time series is stable. Namely, it's worth to predict.

DOI: 10.54254/2754-1169/22/20230320

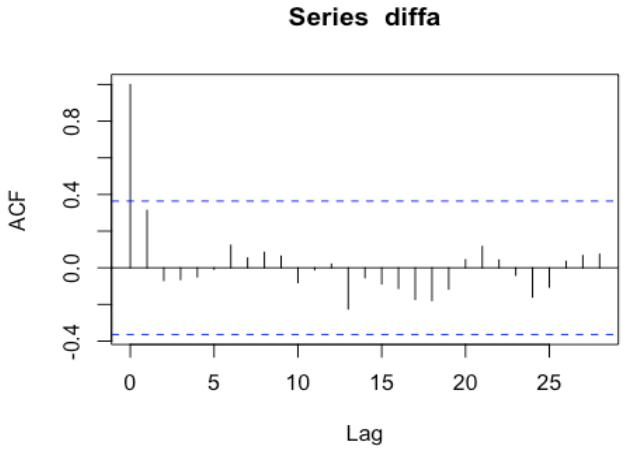

Figure 8: ACF of the time series after logarithm and difference (Photo credit: Original)

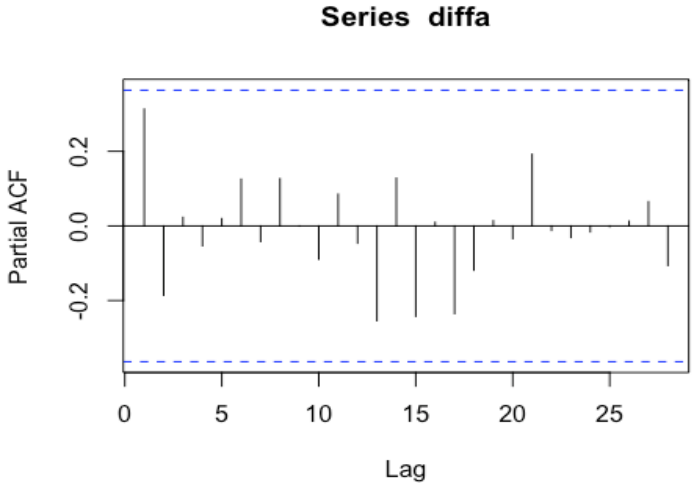

Figure 9: PACF of the time series after logarithm and difference (Photo credit: Original)

3.2.5. White Noise Test

White noise test is a test to evaluate if the time series is white noise, namely, worth predicted. In the author's case, the p-value of white noise test is required bigger than 0.05. That is, the probability of time series being white noise is less than 5%. The p-value of white noise test is 0.5579 which is bigger than 0.05. Therefore, the residuals pass the white noise test.

3.2.6. ARIMA Model Identification

Notice that the author differences the time series of order 1. Therefore, the coefficient of d equal to 1. The ARIMA (p,d,q) model requires 3 coefficient. p and q need to be confirmed. By the figures of ACF and PACF, the author confirms that p=1, q=0. That is because from ACF (see Fig. 8) one can see that the function trails and waves like cosine function. And the PACF (see Fig. 9) truncates. More rigorous, the BIC matrix (see Fig. 10) shows that ARIMA (1,1,0) model is more precise than the others [7]. This is because the block is darkest under p=1, q=0. However, it is not very clear. Therefore, the author use auto.arima function, which also suggests to construct the model of ARIMA (1,1,0).

DOI: 10.54254/2754-1169/22/20230320

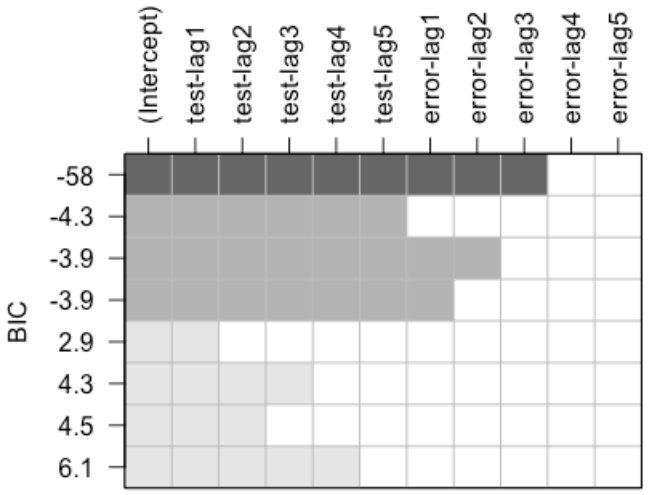

Figure 10: BIC matrix (Photo credit: Original).

3.2.7. Model Check

If the time series is worth predicted, the residuals of time series should follow the Gaussian distribution. One can use QQ-plot to visualize the residuals. More precisely, the value should lie in the red area and follows Gaussian distribution. In the author's case, the QQ-plot (see Fig. 11) shows that the data follows Gaussian distribution roughly. One can see that the middle part lies in the red area perfectly. Therefore, one can do a roughly prediction based on the analysis above.

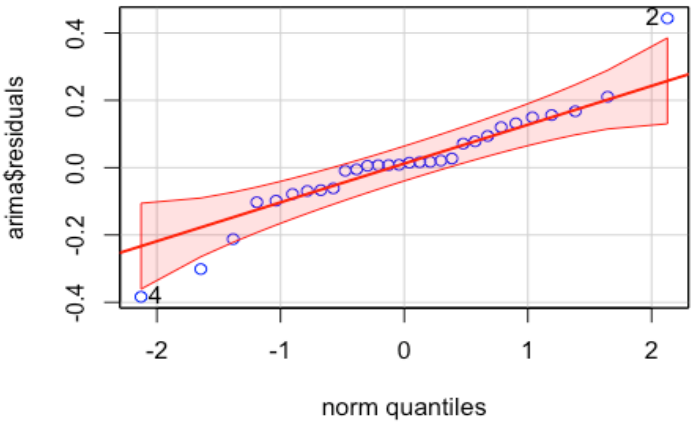

Figure 11: QQ-plot (Photo credit: Original).

3.2.8. Model Prediction

Fig. 12 shows the prediction of the model, which is differenced first-order autoregressive model after natural logarithm.

DOI: 10.54254/2754-1169/22/20230320

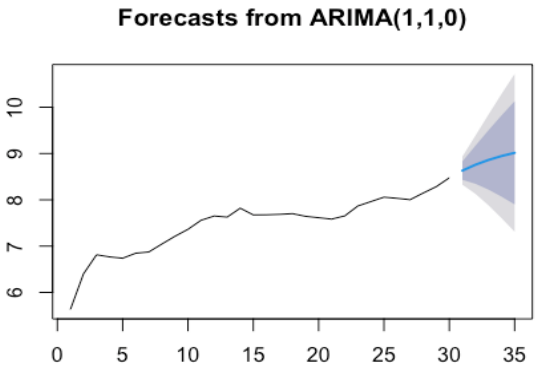

Figure 12: Prediction after logarithm (Photo credit: Original).

The above analysis results show that. The construction area of residential real estate in Shenzhen is predicted to rise. That is, the supply of residential real estate might be close to saturation.

4. Conclusion

The results of this paper indicate that the time series of the construction area of residential real estate of Shenzhen is unstable before natural logarithm. And the time series is unstable before difference. However, the time series is stable after difference of 1 order. And the time series is passed white noise test. Moreover, the residuals of the time series follow the Gaussian distribution. The ARIMA model of the time series is confirmed as (1,1,0) model. After that, the prediction of the time series shows that the statistics of the construction area of residential real estate in Shenzhen before natural logarithm will rise in a short term. Namely, the statistics of the construction area of residential real estate in Shenzhen rise in a short term. Since the natural logarithm function is an increasing function when variable is real.

The prediction can perform a roughly message, that is, the supply of residential real estate might be close to saturation. For investors and buyer, it is benefit for them to think about whether they should buy residential real estate or not. As one can see, more houses sold in the markets less the price is. However, the political issue is not under consideration, which is one of the limitations of this prediction.

What's more the qq-plot shows that this model can be more precise, since the end point doesn't lie in the red area. Also, it may be better to use high order difference instead of using nature logarithm. Since some mathematical structure may be destroyed after nature logarithm. What's more, the model will be more convincing, if the author uses T-test to finish the hypothesis testing.

References

[1] *Glaeser, E., Huang, W., Ma, Y., & Shleifer, A. (2017). A real estate boom with Chinese characteristics. Journal of Economic Perspectives, 31(1), 93-116.*

[2] *Khan, S., & Alghulaiakh, H. (2020). ARIMA model for accurate time series stocks forecasting. International Journal of Advanced Computer Science and Applications, 11(7).*

[3] *Poongodi, M., Vijayakumar, V., & Chilamkurti, N. (2020). Bitcoin price prediction using ARIMA model. International Journal of Internet Technology and Secured Transactions, 10(4), 396-406.*

[4] *Wang, F., Zou, Y., Zhang, H., & Shi, H. (2019, October). House price prediction approach based on deep learning and ARIMA model. In 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT) (pp. 303-307). IEEE.*

DOI: 10.54254/2754-1169/22/20230320

[5] *A. Meyler, G. Kenny and T. Quinn, "Forecasting Irish Inflation using ARIMA Models", Central Bank of Ireland Research Department, Technical Paper, 3/RT/1998.*

[6] *Bosse, N. I., Abbott, S., Cori, A., van Leeuwen, E., Bracher, J., & Funk, S. (2023). Transformation of forecasts for evaluating predictive performance in an epidemiological context. medRxiv, 2023-01.*

[7] *Wang, X. (2022). Research on the prediction of per capita coal consumption based on the ARIMA–BP combined model. Energy Reports, 8, 285-294.*

[8] *Sahoo, U. K., Chavan, R. V., & Bharati, S. V. (2022). Predictive analysis of coconut prices in Odisha: An ARIMA approach.*

[9] *Dare, J., Patrick, A. O., & Oyewola, D. O. (2022). Comparison of Stationarity on Ljung Box Test Statistics for Forecasting. Earthline Journal of Mathematical Sciences, 8(2), 325-336.*

[10] *Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. Structural equation modeling: a multidisciplinary journal, 21(1), 1-19.*