Distribution Agreement

Yue Gao April 04, 2022

Predicting Housing Price in Beijing Using ARIMA Models

by

Yue Gao

Ruoxuan Xiong
Adviser

Department of Quantitative Theory and Methods

Ruoxuan Xiong

Adviser

Jeremy Jacobson

Committee Member

Mi Luo

Committee Member

2022

Predicting Housing Price in Beijing Using ARIMA Models

By

Yue Gao

Ruoxuan Xiong

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Quantitative Theory and Methods

2022

Abstract

Predicting Housing Price in Beijing Using ARIMA Models

By Yue Gao

With the rapid increase in housing demand, more and more homebuyers in Beijing tend to purchase housing properties without collecting enough information. To provide a solution for information asymmetry and high information acquiring cost in the housing market, this study focusses on the housing price forecasting. Reliable forecasts could provide valuable information for homebuyers and sellers and help them better understand the local housing market. This paper seeks to predict the housing price in Beijing using ARIMA models due to the model's demonstrated outperformance in predicting time series data accuracy. In addition, district-by-district housing price prediction is also performed.

Index Terms—ARIMA models, housing price data, time series analysis

Predicting Housing Price in Beijing Using ARIMA Models

By

Yue Gao

Ruoxuan Xiong

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Quantitative Theory and Methods

2022

Acknowledgements

Foremost, I would like to express my sincere gratitude to my faculty advisor, Dr. Ruoxuan Xiong, for her continuous support and guidance of my research, for her patience, motivation, and enthusiasm. I could not have imagined having a better advisor and mentor. I also want to thank my committee members Dr. Mi Luo and Dr. Jeremy Jacobson, for catching up with me about my thesis progress, for their encouragement, and valuable comments along the way.

Table of Contents

1. Introduction

Information asymmetry in the housing market is a common problem home buyers encounter. For example, home sellers may have more information about the local housing markets than homebuyers; and local buyers can acquire more comprehensive housing information than non-local buyers. According to Qiu et al. (2019), "homebuyers from outside a geographical area (non-locals) typically pay more for a given property compared to local buyers" due to more information search costs and weaker bargaining power. The problem of asymmetric information worsens in global cities like Beijing. The modifications of the hukou (household registration) system in China encourage population mobility and create the chance for vast inter-regional migration. (Shen, 2013) Beijing, as the political, cultural, and economic center of China, has also attracted enormous internal migrants from other cities in China or global migrants from foreign countries. According to the 2020 Chinese census, inter-regional migrants to Beijing are about 8.42 million, making up 38.5% of permanent residents.

A significant number of migrants lead to a massive non-local homebuyers' group in Beijing. Therefore, a potential solution to the problem of information asymmetry in the housing market benefits non-local homebuyers by providing them applicable information of the local housing market and bringing down their time and financial cost when acquiring housing information.

Researchers have done many studies to predict housing prices with different approaches. Wu and Brynjolfsson (2009) collected housing search index data from search engines like Google to predict future housing market activities and forecasted housing price trends based on searching frequencies. Research by Park and Bae (2015) proposed a housing price prediction model based on machine learning algorithms including C4.5, RIPPER, Naïve Bayesian, and AdaBoost to analyze the housing data in Fairfax County, VA. Wang et al. (2019) suggested a housing price forecasting model based on deep learning and the ARIMA model to capture the relationship between determinant factors and housing price. In this study, ARIMA models are adopted to forecast the housing price in Beijing because it is one of the most used models in estimating housing prices and other commodity prices.

For example, in the study by Contreras et al. (2003), ARIMA models have been applied to forecast the next-day electricity price. ARIMA models have also been used to estimate the gold price and show satisfying performance in predicting the short-run gold price, but they failed to capture the sudden changes in the gold price (Bandyopadhyay and Guha, 2016). ARIMA models are also being applied to forecast Pu'er tea price and showed relatively low prediction errors, especially with shorter forecasting periods (Dou et al. 2021). Although ARIMA models generated relatively low prediction error in predicting these commodity prices, we are curious if ARIMA models can produce reliable prediction results for Beijing's housing price. Since housing prices are influenced by factors such as property characteristics, surrounding infrastructures, government housing policies, etc., housing prices could exhibit more volatility and are relatively unstable. In this study, we will focus on the performances of ARIMA models in predicting the housing price level in Beijing. In addition, the housing price changing pattern differs from district to district in Beijing because of the districts' distinctive location, economic status, and more. Therefore, we will also investigate the different housing price changing patterns and ARIMA models' prediction accuracy based on districts.

2. Research Methodology

This part of the article discusses the basic principles of the ARIMA model used for prediction, elaborates the modeling processes to fit ARIMA models with the time series data, and introduces the success criteria adopted to evaluate the forecasting performances.

2.1. Background: ARIMA Model Approach

This section explains the background knowledge of the time series model used in this article. ARIMA is an acronym for Autoregressive Integrated Moving Average. After Box and Jenkins (1976) developed a systematic ARIMA modeling approach, this linear approach has become the standard and widely applied to perform time series forecasting on future quantities or prices based on historical data.

ARIMA (p, d, q) is a composite time series model that incorporates differencing with autoregression and moving average approach. Since the ARIMA model involves differencing, it can be viewed as a generalized form of a simple ARMA model. The acronym AR, I, and MA represent Autoregressive, Integrated, and Moving Average. Autoregression stands for regressing the variable against itself. The autoregression model, AR (p), studies the dependencies between the variable of interests and several past values of the variable with lag order p. We can write the formula of the AR (p) as:

$$y_t = \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t + c,$$

where y_t is the stationary variable, predictors y_{t-i} are lagged values of y_t, $\emptyset_i$ is the autocorrelation coefficient at lag i, ε_t is the normally distributed white noise with mean zero and variance one, and c is the constant.

Integration corresponds to a differencing step which is one way to make non-stationary data stationary by calculating the differences of two consecutive observations. A time series $\{y_t\}$ is stationary if, for all s, the distribution of $(y_t, y_{t+1}, \dots, y_{t+s})$ does not depend on t. In other words, to apply the autoregression model and moving average approach, we are looking for a time series without trend and periodic fluctuations (seasonality). This is because trend and seasonality influence the value of times series over time. Differencing can be applied one or more times to stabilize the mean and eliminate rising or decreasing trends. When seasonality presents in the time series, seasonal differencing, also referred to as "lag-m" differences, can remove the seasonal component by taking the difference between an observation and another observation from the previous season. The parameter d, in I (d), stands for the degree of differencing, which is the times of differencing.

MA (q) refers to a moving average model with order q. Moving average model is very similar to the autoregressive model. But instead of regress on its own past values, MA model forecasts the variable of interest based on past forecast errors. The formula of the MA (q) model can be written as:

$$y_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon_t + \mu,$$

where y_t is the stationary variable, predictors $\varepsilon_t, \varepsilon_{t-i}$ are white noise error terms, θ_i is the moving average coefficient at lag i, and μ is the mean of the series (usually assumed to be zero because of the stationarity).

We can add these models together and form the function for an ARIMA (p, d, q) model:

$$y'_t = \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{p-1} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon_t + c,$$

where y'_t indicates the differenced time series.

2.2. Workflow Chart

The process of fitting an ARIMA model to the time series housing data to forecast the housing price is summarized below in Figure 1.

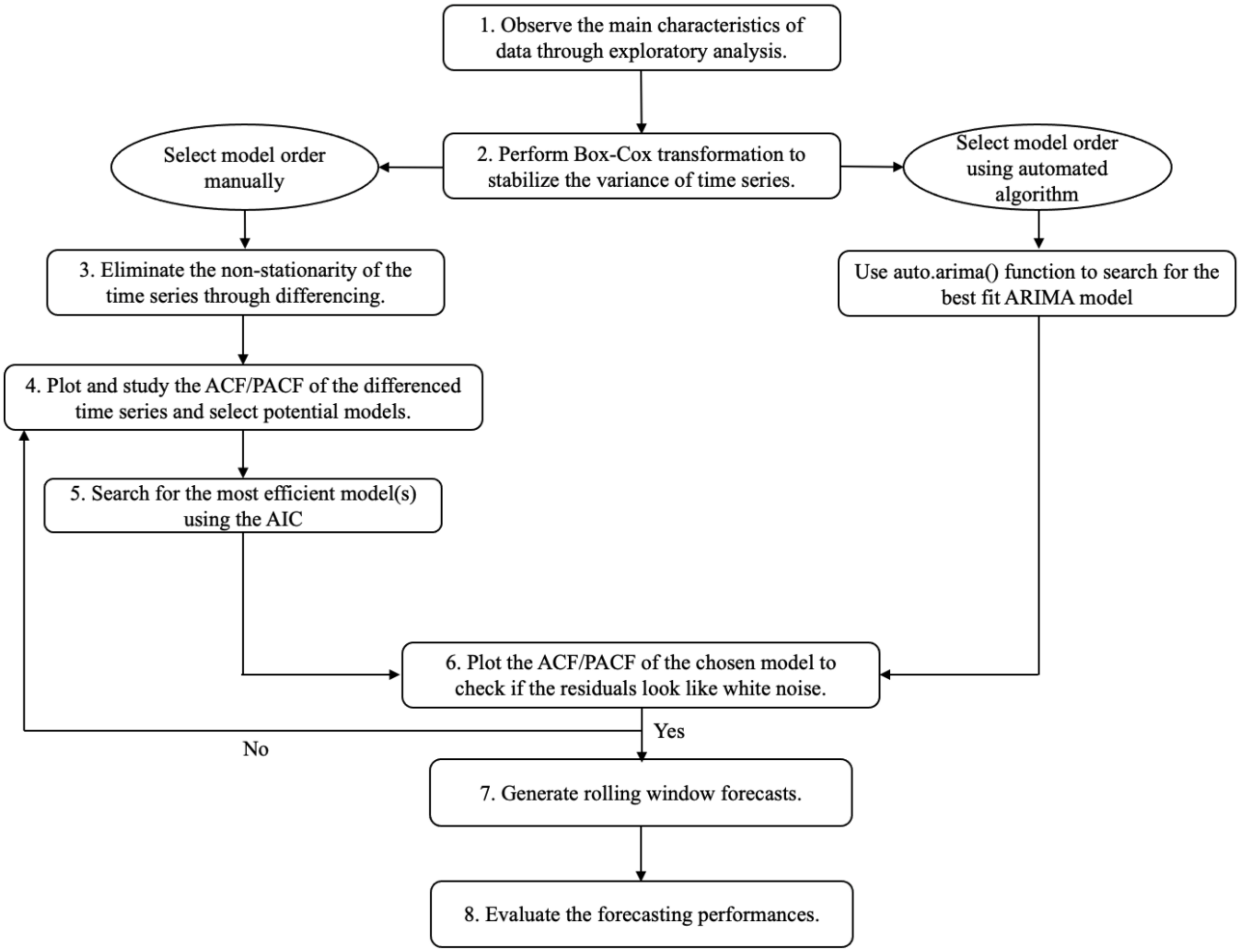

1. Observe the main characteristics of data through exploratory analysis.
2. Perform Box-Cox transformation to stabilize the variance of time series.
Select model order manually
Select model order using automated algorithm
3. Eliminate the non-stationarity of the time series through differencing.
Use auto.arima() function to search for the best fit ARIMA model
4. Plot and study the ACF/PACF of the differenced time series and select potential models.
5. Search for the most efficient model(s) using the AIC
6. Plot the ACF/PACF of the chosen model to check if the residuals look like white noise.
Yes
No
7. Generate rolling window forecasts.
8. Evaluate the forecasting performances.

Figure 1. ARIMA Model Workflow Chart

Note: the auto.arima () function is developed based on the Hyndman-Khandakar algorithm (Hyndman & Khandar, 2008).

2.3. Evaluation Criteria of ARIMA models

Future values are always unknown. No matter what orders an ARIMA model adopts, it is impossible for this model to yield 100% accurate results. In other words, a certain level of error that exists in the prediction needs to be measured and studied. According to Sarı (2016), the most important criteria to evaluate the predictive success is the accuracy of the prediction, which is measured by analyzing the predicted errors. Therefore, in step 8, we adopt the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) to measure the predictive success of different ARIMA models. The formulas to calculate the RMSE, MAE, and MAPE are shown below in Table 1. (Sallehuddin et al. 2009). In the below formulas, y_t represents the actual value, $\hat{y}_t$ represents predicted value, n represents the number of predicted periods. ARIMA models that produce the lowest RMSE, MAE, MAPE should be selected as the best fit model.

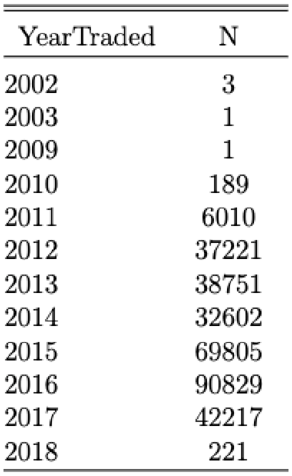
Evaluation Criterion	Formula
Root Mean Square Error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$
Mean Absolute Error (MAE)	$\mathrm{MAE} = (\sum_{t=1}^{n}\left\|\frac{y_t - \hat{y}_t}{n}\right\|)$
Mean Absolute Percentage Error (MAPE)	$\mathrm{MAPE} = (\sum_{t=1}^{n}\left\|\frac{y_t - \hat{y}_t}{y_t}\right\|)\frac{100}{n}$

Table 1. Evaluation Criteria Formulas

3. Data and Data Pre-Processing

3.1. Data

The data used in this study are daily housing transaction data in Beijing. It is fetched from the website Lianjia, an online Chinese real-estate transaction platform with open access. The total sampling period studied is from January 2010 to December 2017. The table below shows that although data from a more extended period are available, the number of transactions recorded is unevenly distributed across the years.



YearTraded	N
2002	3
2003	1
2009	1
2010	189
2011	6010
2012	37221
2013	38751
2014	32602
2015	69805
2016	90829
2017	42217
2018	221

Table 2. Number of Transactions Recorded by Year

The table below shows that the number of transactions recorded before 2010 is too limited to perform a thorough analysis. Therefore, we select the data from 2010 to 2017 to ensure the study's statistical power. After further examination of the data, four variables: Date (the transaction date of each property), Total Price (the total price of the property in 10,000 yuan), Price (the price per square meter in yuan), and District (housing located districts in Beijing) are selected to build the model and answer the research questions. The "District" variable includes 12 districts: ChangPing, ChaoYang, DaXing, DongCheng, Fangshan, FengTai, HaiDian, MenTouGou, ShiJingShan, ShunYi, TongZhou, and XiCheng districts.

3.2. Data Pre-Processing

During data pre-processing, we first examine the distribution of two key variables, unit housing price and total housing price, to get an overview of Beijing's housing price level. From the histogram of unit price in Figure 2, we observe that majority of the observations fall into the range of 25,000 to 65,000 yuan per square meter, with the lowest unit price around zero yuan and the highest around 150,000 yuan. And the observation frequency flattens with the increasing unit price. Overall, the distribution of the unit price skewed to the right with a long right tail. For the total price histogram, observations between the 2000,000-to-5000,000-yuan range are the most common. The lowest total price is around 500,000 yuan, and the highest total price is closing to 20,000,000 yuan. This histogram also exhibits a right-skewed distribution with a long right tail. The lowest unit and total price, as well as the long tails of these two histograms, indicate the potential risk of erroneous data. According to the data, the actual average housing price in 2010 was around 1.35 million yuan, and the average unit price was around 15,574 yuan. In addition, it is unrealistic to sell housing properties for free. This means that the data entries with 0-yuan unit prices are likely to be inaccurate or erroneous. To mitigate the negative effect caused by mistaken data entries, we cleaned the dataset by removing the missing values and entries with 0-yuan unit prices.

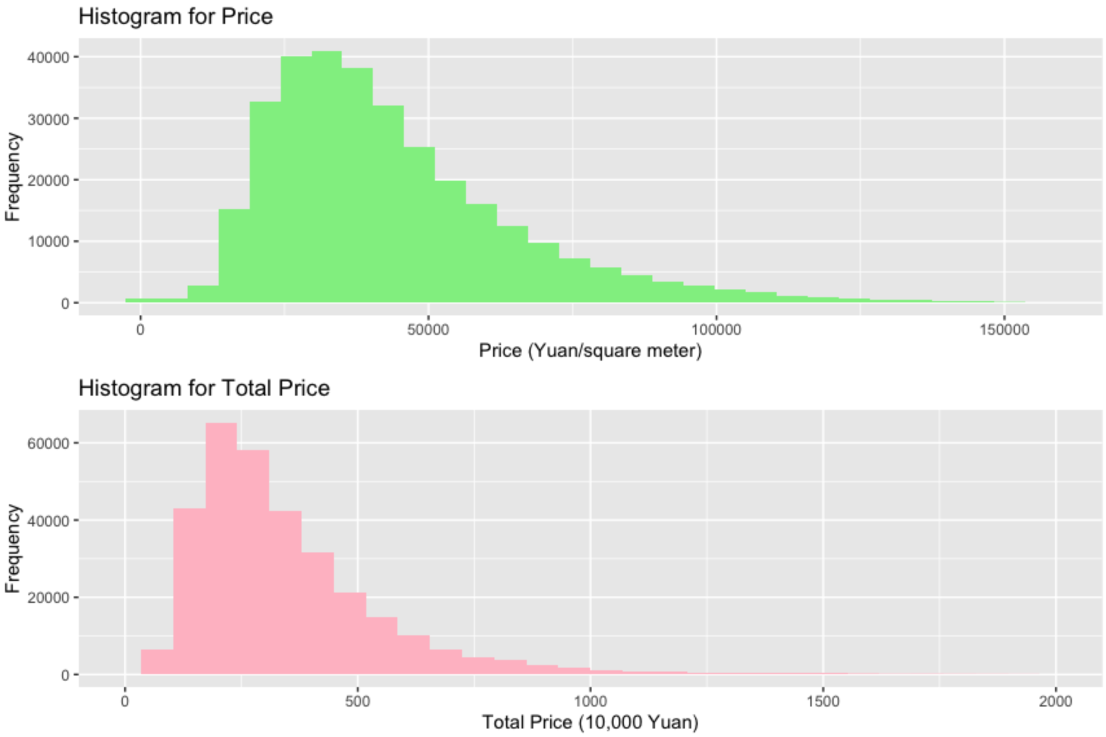


Figure 2. Total Price and Unit Price Distribution

To perform the ARIMA model, we create two time series objects: total price monthly average over time and unit price monthly average. The trends of the changing unit price and total price are plotted in Figure 3. The overall trends for the average unit price and average total price are very similar. They both show a significant increasing trend from 2010 to 2017. However, they differ in specific periods. For example, there was a sharp increase in the average total price around the first and second quarters of 2010, but the average unit price did not rocket until the last quarter of 2010. And the unit price soon increased again during 2011 when the total price remained at a steadily increasing rate. Since the total price is highly influenced by the size of the property and the size of houses in Beijing can range from under 20 square meters to over 500 square meters, we choose the unit price variable instead of the total price as the indicator for price level in Beijing in this study.

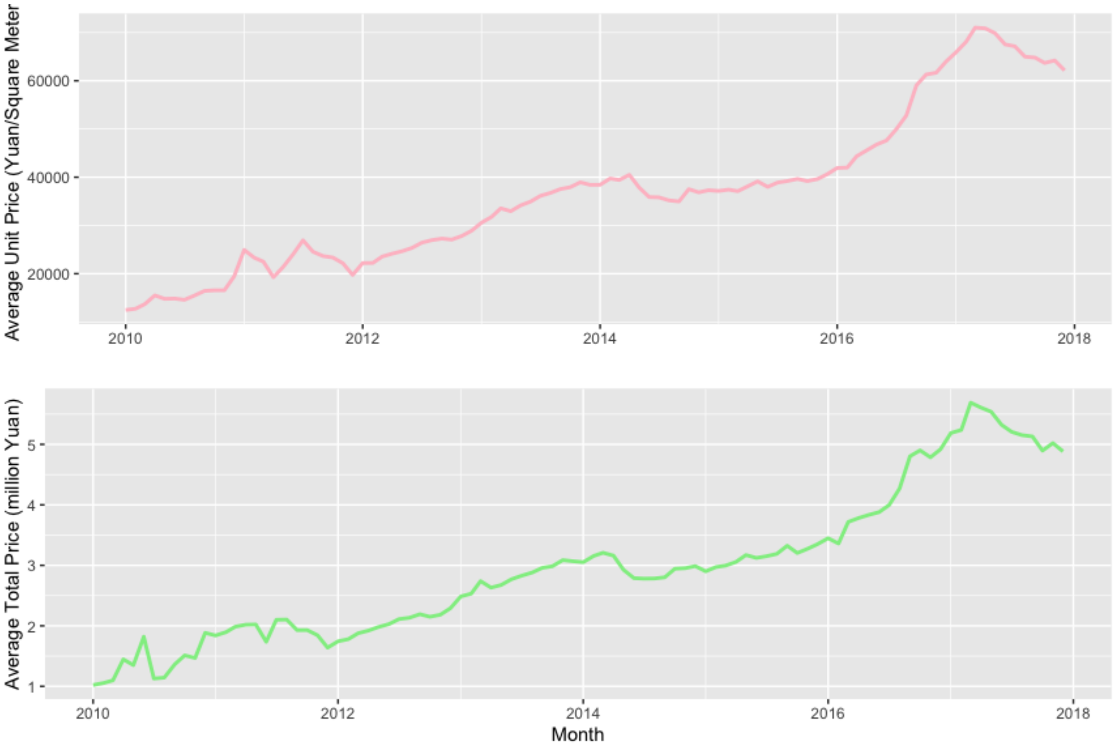


Figure 3. Total Price and Unit Price Trend

The study by Ngai and Tenreyro (2014) suggests that "each year a housing boom of considerable magnitude takes place in the second and third quarters of the calendar year (the "hot season"), followed by a bust in the fourth and first quarters (the "cold season")." Since the housing market tends to have hot and winter seasons, it is necessary to consider seasonality before fitting prediction models. We draw the seasonality plot based on the time series data to examine if there is an apparent periodic pattern in the Beijing housing price. The seasonality plot in Figure 4 has a circular time axis. One complete circle represents one year. Suppose the plot is in the shape of a regular circle, the housing price increases at a stable rate from month to month. If the plot is in the form of an ellipse or other irregular shapes, these can be the signs for potential periodic patterns. From Figure 4, we can see a sign of seasonality in 2010 and 2011 because the average housing price was higher around January and July. However, the pattern did not continue through the following years. Starting from 2012, the average price level increases yearly without a significant seasonal pattern. Since ARIMA models can handle certain types of seasonality and the periodic pattern in our time series data is not significant, we do not decompose the seasonality from the time series object before fitting ARIMA models.

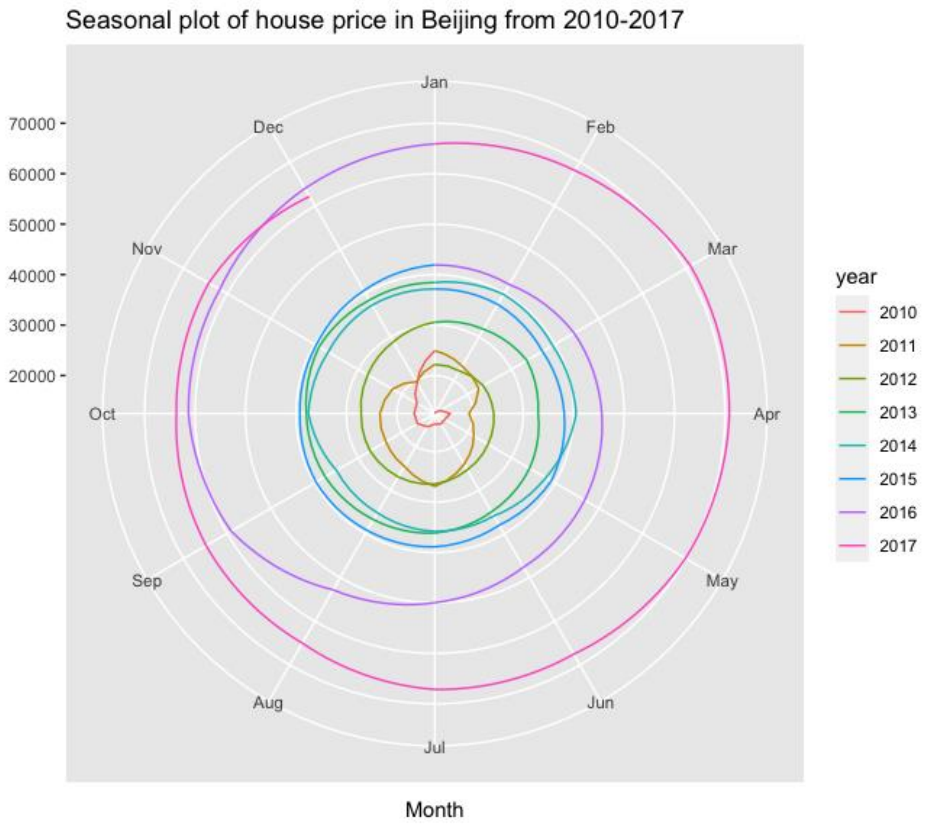


Figure 4. Seasonality Plot of Housing Price in Beijing

Before fitting ARIMA models, it is crucial to check the data’s stationarity and differencing the data if necessary. The ACF plot of the time series in Figure 5 can help identify the stationarity. The ACF will quickly drop to zero if the time series is stationary. In Figure 5, the ACF decreases slowly, showing the training series was non-stationary. Also, we can use the Augmented Dickey-Fuller (ADF) test to examine if the time series is stationary. ADF test is a robust unit root test. The null hypothesis is that a unit root is present in the time series sample, and the alternative hypothesis is that the sample is stationary. In our case, the p-value is 0.4078. Since the p-value is greater than 0.05, there is no significant evidence to reject the null.

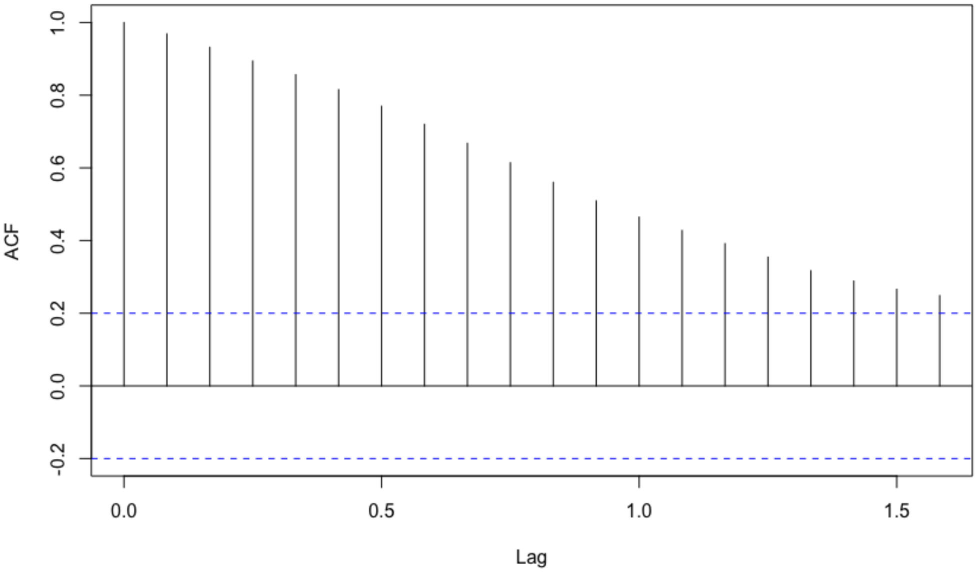


Figure 5. ACF Plot for Time Series Sample

To make the non-stationary time series sample stationary, we take the first difference between consecutive observations and plot the ACF of the differenced series in Figure 6. After the first differencing, we draw the ACF of the differenced time series and apply the ADF test. Although the ACF of the differenced series quickly decreases to zero, the p-value is 0.1917 indicating there is no sufficient evidence to reject the null. Because the time series sample is still non-stationary after the first differencing, we take the second difference to check the ACF and ADF statistics again. With the ACF quickly dropping to 0 and a p-value of 0.01, there is enough evidence to reject the null hypothesis, indicating the differenced time series is stationary.

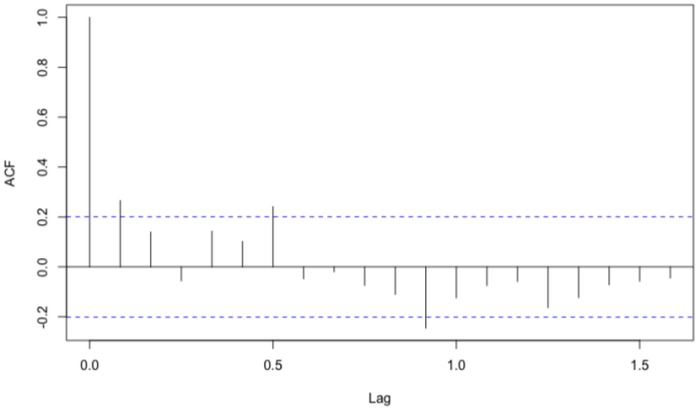


Figure 6. ACF with 1st Difference

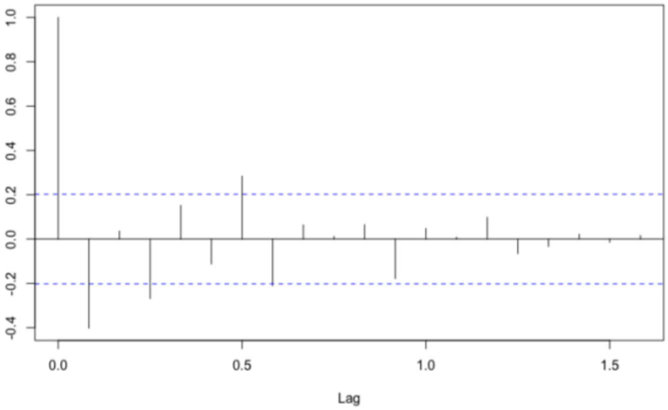
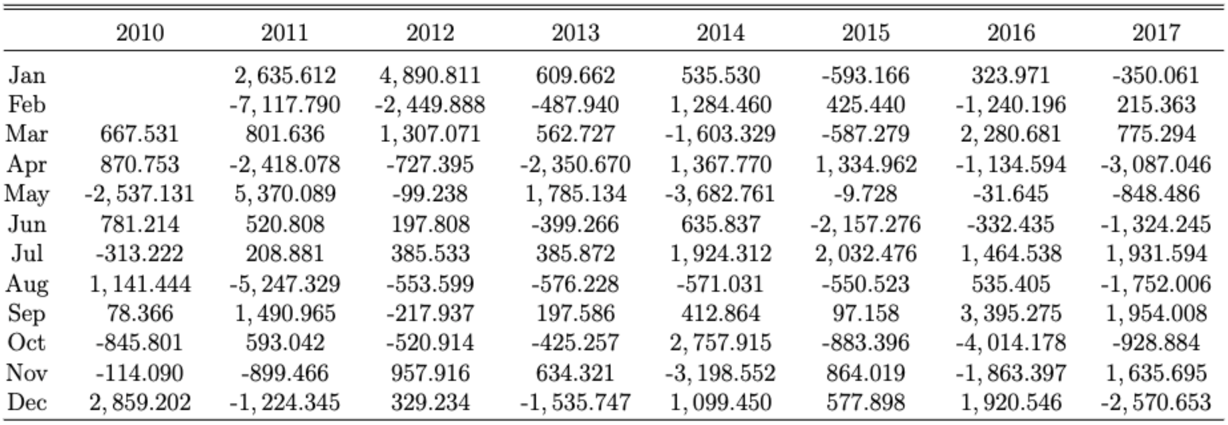


Figure 7. ACF with 2nd Difference

Finally, the time series is split into the training and test data set with the 87% and 13% ratios to test the forecasting performances. The final time series are shown in Table 3. It contains the monthly average unit housing price from 2010, March to 2017, December with 94 entries in total (first 2-month data lost due to differencing). The monthly average unit prices from 2010 to 2016 are treated as the in-sample data. The monthly average unit prices in 2017 are treated as the hold-out sample for testing purposes.



	2010	2011	2012	2013	2014	2015	2016	2017
Jan		2,635.612	4,890.811	609.662	535.530	-593.166	323.971	-350.061
Feb		-7,117.790	-2,449.888	-487.940	1,284.460	425.440	-1,240.196	215.363
Mar	667.531	801.636	1,307.071	562.727	-1,603.329	-587.279	2,280.681	775.294
Apr	870.753	-2,418.078	-727.395	-2,350.670	1,367.770	1,334.962	-1,134.594	-3,087.046
May	-2,537.131	5,370.089	-99.238	1,785.134	-3,682.761	-9.728	-31.645	-848.486
Jun	781.214	520.808	197.808	-399.266	635.837	-2,157.276	-332.435	-1,324.245
Jul	-313.222	208.881	385.533	385.872	1,924.312	2,032.476	1,464.538	1,931.594
Aug	1,141.444	-5,247.329	-553.599	-576.228	-571.031	-550.523	535.405	-1,752.006
Sep	78.366	1,490.965	-217.937	197.586	412.864	97.158	3,395.275	1,954.008
Oct	-845.801	593.042	-520.914	-425.257	2,757.915	-883.396	-4,014.178	-928.884
Nov	-114.090	-899.466	957.916	634.321	-3,198.552	864.019	-1,863.397	1,635.695
Dec	2,859.202	-1,224.345	329.234	-1,535.747	1,099.450	577.898	1,920.546	-2,570.653

Table 3. Differenced Time Series

4. Result

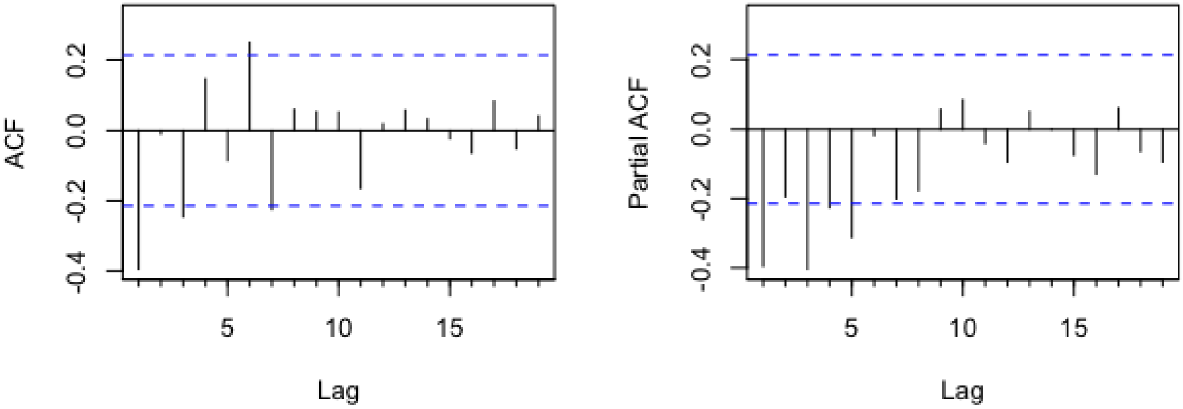
4.1. RQ1: What is the performance of ARIMA in predicting housing price in Beijing?

To answer the first research question, 21 different ARIMA (p, d, q) models were selected and tested to forecast the average housing price in Beijing. The various combinations of the parameters are first determined by the automated function in R. The complete ARIMA models and their corresponding AIC scores are recorded in the Appendix. Four models are selected among the 21 ARIMA models based on the AIC score. AIC is short for Akaike Information Criterion, one of the most widely used model selection criteria. AIC estimates the relative quality of each model by rewarding goodness of fit and penalizing overfitting. The AIC value of a model is calculated by:

$$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\hat{\mathrm{L}}),$$

where k is the number of estimated parameters and is the maximum value of the likelihood function for the model. Among a set of candidates ARIMA models, ones with smaller AIC values are preferred. Apart from the AIC value, the ACF and PACF plot of the final time series sample in Figure 8 also provides useful information in choosing the AR and MA orders in the ARIMA model. We use the PACF to determine the terms used in the AR model because the AR model examines the dependencies between two time spots, and the PACF measures the real correlation between two time spots by taking out the indirect effect brought by other time spots. And we use the ACF factor to evaluate the MA order. When examining the PACF plot, only the significant PACF values will be chosen to determine the order of the AR model. And we can apply the same rule to the ACF and AR order. In figure 8, there are horizontal blue dash lines representing significant thresholds and vertical solid lines representing the ACF and PACF values at each time spot. Only the vertical lines that exceed the horizontal dash lines are considered significant. We have three significant ACF values and three significant PACF values in this case. The significant ACF values are at lag 1, lag 3, and lag 6. The significant PACF values are at lag 1, lag 3, and lag 5. Therefore, possible combinations for AR and MA orders include ARMA (1, 1), ARMA (3, 1), ARMA (3,3), and so on. Between these models, models with minimal error terms are preferred and simpler models with fewer independent variables are preferred to keep our model concise.

Figure 8. ACF/PACF



After examining the AIC value and the ACF/PACF plot, four ARIMA models are selected and their error terms are further examined: ARIMA (3, 0, 1) which is the ARIMA model with the minimum AIC value, ARIMA (1, 0, 1), ARIMA (1, 0, 3) and ARIMA (5, 0, 1). The results of the evaluation criteria are given in Table 4. From Table 4, we noticed that ARIMA (1, 0, 3) has the minimum error terms.

ARIMA Model Comparison			
ARIMA Model	RMSE	MAE	MAPE
(3, 0, 1)	1427.716	1017.948	209.7144
(1, 0, 1)	1472.094	1035.227	180.8662
(1, 0, 3)	1416.856	1000.855	162.9572
(5, 0, 1)	1417.866	1009.412	189.5479

Table 4. Evaluation Criteria Statistics

In addition to the error terms, we also double-check the residuals plots of the four ARIMA models in Figure 9. From the four ACF plots in the below figure, we discover almost all ACF values are within the significant threshold, which ensures the robustness of the four ARIMA models. In addition, ARIMA (1, 0, 3) has the smallest residual range, which is consistent with the evaluation criteria in Table 4. After examining the evaluation criteria statistics and the residuals plot, ARIMA (1, 0, 3) is selected to forecast the housing price level.

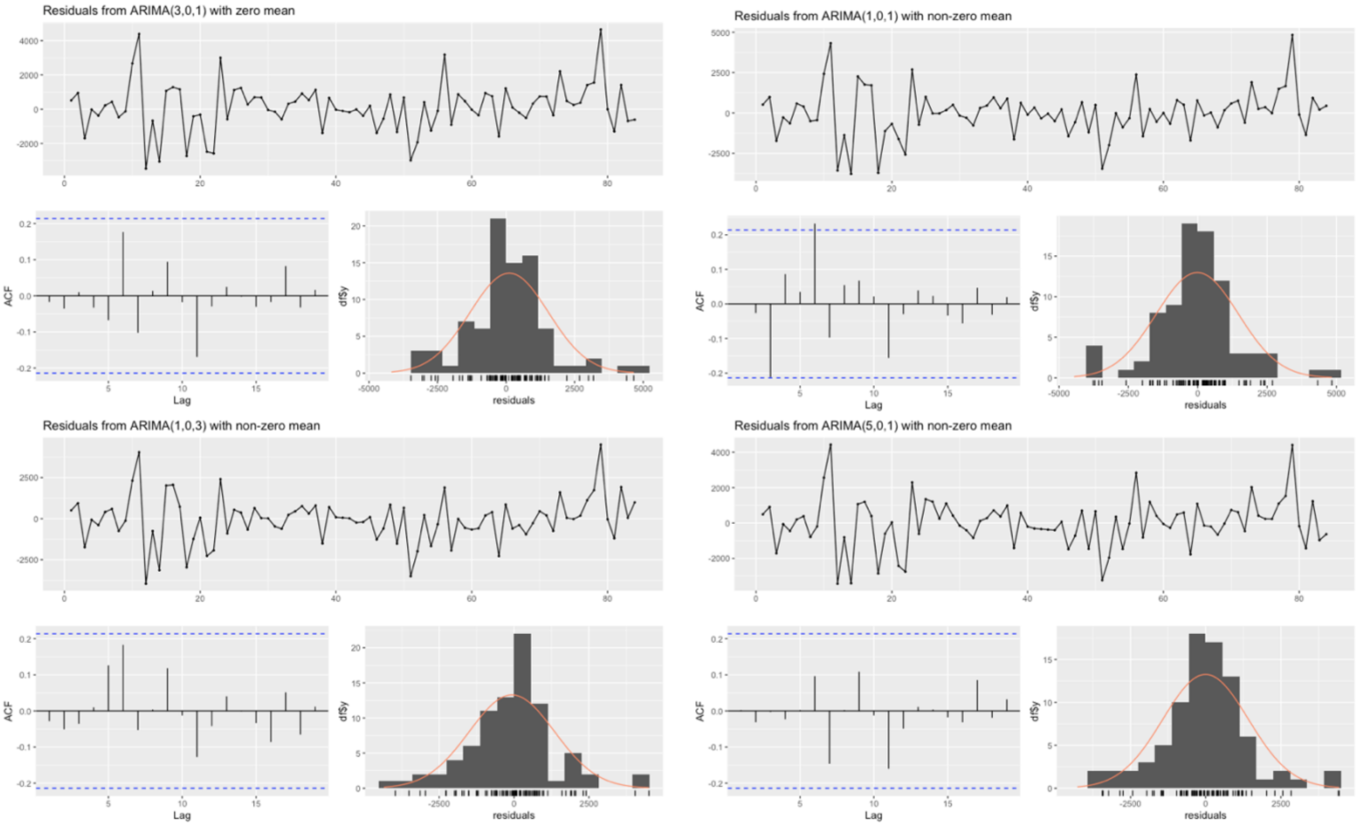


Figure 9. Residuals Plot from 4 ARIMA Models

We adopted rolling forecasting and calculated the rolling window cross-validated error to estimate the forecasting performances in the forecasting process. Rolling forecast is a commonly used measure in literature and application when forecasting time series data. The basic idea is to always roll the training sample forward and add new data points to the training sample when available. Compared with the traditional multi-step forecasting, where all of the training data are used to forecast the next several months simultaneously, the rolling forecast measure continuously incorporates new information when predicting the future values. The multi-step forecast and rolling forecast results are shown in Figure 10 below. There is a solid black line representing the actual data in the plot, a blue line indicating the prediction results based on the multi-step forecast, and a red line denoting the forecasting result generated by the rolling forecast. They both used the training sample to predict the housing price in the next ten months. While the multi-step forecasting is a flatter horizontal line, the rolling forecasted results are closer to the real data.

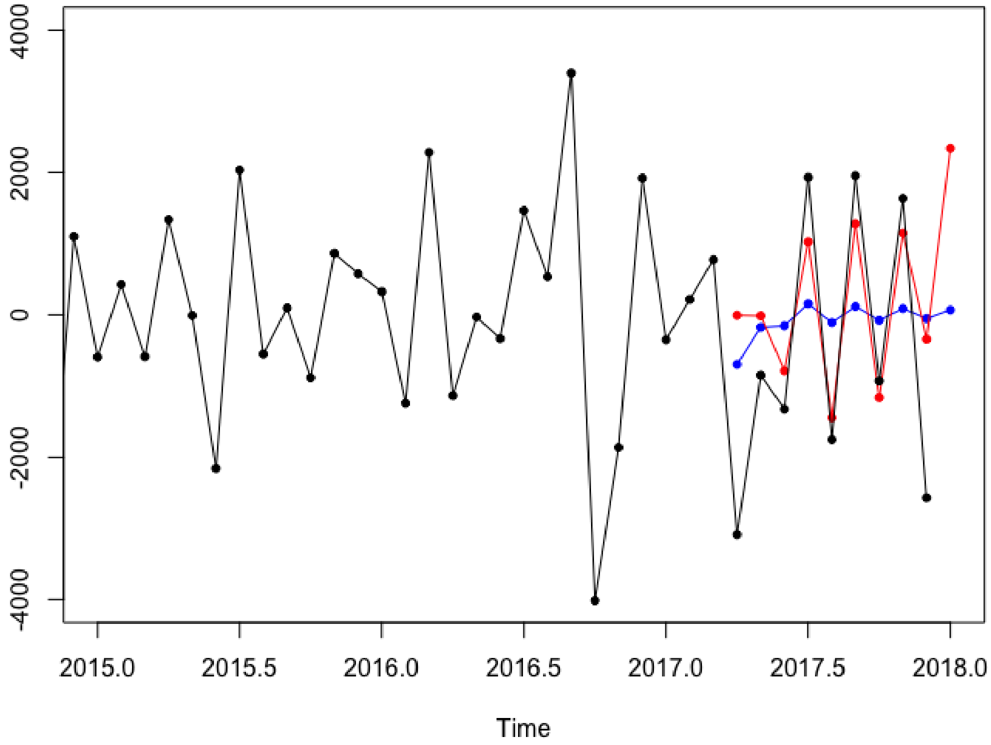


Figure 10. Multi-step and Rolling Forecasts

To further examine the prediction error of rolling forecast, we use the training sample to perform one-step, two-steps, three-steps, and six-step forward estimates, which forecast the next month, two months, three months, and six months housing price average, respectively. We then calculated the root mean square forecasting error (RMSE). The forecasting error are summarized in the table below. From Table 5, we noticed the RMSE is larger for longer forecasting periods. One possible explanation for this observation is outside influencing factors, such as housing policy, are more likely to take part and affect the housing price average when trying to predict the prices several months ahead.

Rolling Forecast Error (Cross-Validated)	
Forecasting Period	RMSE
1 month	1289.098
2 months	1346.923
3 months	1361.159
6 months	1428.687

Table 5. Rolling Forecast Error

4.2. RQ2: Will the forecasting performance differ from district to district?

To answer this research question, we apply ARIMA models to the data collected in each district respectively. The error terms of the most suitable ARIMA model for each district are recorded in Table 6. Since the model with the minimum RMSE does not necessarily have the lowest MAE and MAPE, we use the Root Mean Squared Error (RMSE) as the main evaluating standard. The reason to choose RMSE instead of MAE and MAPE is that RMSE is more widely used as a predictive ARIMA model that minimizes the RMSE lead to forecasts of the mean. By comparing RMSE values for 12 different districts, we noticed that the performances of ARIMA models vary with districts. ARIMA models have better performances when predicting housing price change in MenTouGou, DaXing, TongZhou, and FengTai. However, ARIMA models have more significant forecasting errors when predicting housing price change in ChaoYang, ShiJingShan, DongCheng, and Xicheng.

District ARIMA Model Comparison				
District	ARIMA	RMSE	MAE	MAPE
ChangPing	(2,1,2)	1730.915	1124.315	5.256555
ChaoYang	(0,1,3)	4476.518	2543.045	7.173551
DaXing	(1,1,0)	1095.464	832.7196	2.943391
DongCheng	(0,1,2)	2561.954	1895.088	3.46419
FangShan	(0,1,0)	1771.014	1222.542	5.082396
FengTai	(0,1,1)	1205.705	870.2298	2.697913
HaiDian	(0,1,1)	1667.434	1218.876	2.352648
MenTouGou	(2,1,1)	1034.501	783.9738	3.564443
ShiJingShan	(2,1,1)	2713.665	1597.873	4.851847
ShunYi	(0,1,0)	1473.631	1051.865	4.337018
TongZhou	(1,1,0)	1157.394	830.2064	3.139349
XiCheng	(1,1,0)	2084.676	1533.075	2.778483

Table 6. Evaluation Criteria Statistics by District

5. Discussion

5.1. Limitations of Study

One limitation of this study is insufficient Data. The data used is the housing transaction data fetched from Lianjia. It is inadequate in three aspects. First, the data covered the housing transaction data from 2010 to 2017, giving 96 monthly average data. There is still potential to improve model accuracy if ARIMA models are fitted with more training data. In addition, the dataset only covered the transaction data in 12 districts of Beijing, with four other districts' housing prices left out. Although it covered most districts in Beijing, it failed to collect the housing price information from 4 additional rural districts in Beijing. The housing price data may not be a throughout representation of the housing price level in Beijing because it includes more data from the urban districts than the rural districts. And the housing price level differs significantly from district to district. Finally, the data is not guaranteed accurate in every entry since the data is fetched from a website with open access. Many individuals have the right to adjust the relevant data. Also, home buyers and home sellers who choose to transact through Lianjia instead of other platforms may be a group with specific characteristics. For example, sellers may be unwilling to sell luxury properties through Lianjia. They may have access to channels, such as private agencies, to help them target potential buyers more efficiently. With that being said, the housing price data from Lianjia is under the risk of selection bias because the data is collected from a selected group instead of a randomized group. As a result, the limitations of the data may restrict the ability to conduct a thorough analysis of the housing price forecasting in Beijing. Potential improvements are collecting housing price data randomly from multiple platforms within a more extended time range.

The second limitation is methodological limitations. Due to the time limit, the ARIMA models used in this study are univariate regression that only adopts single input— time. Since the housing price is highly influenced by other influencing factors such as location, build year, etc., adding these factors as additional input variables and adopting the multivariate approach may improve prediction accuracy. In addition, in this study, only ARIMA models are adopted to predict the housing price in Beijing. Although ARIMA is one of the most adopted time series models, it has been preferred as a linear model. ARIMA models are unable to capture the non-linear portion of the data. One potential improvement is to adopt a linear and non-linear hybrid model to improve the prediction accuracy. In 2011, Khashei and Bijari proposed a novel hybridization of ARIMA and Artificial Neural Network (ANN) to improve the time series prediction accuracy. The study by Temür et al. (2019) also adopted the ARIMA and LSTM Hybrid model to predict the housing sales in Turkey and indicated the performance of the hybrid model is better than either ARIMA or LSTM on its own. Therefore, the hybrid approach may have the potential to improve the prediction performance in the case of Beijing as well.

5.2. ARIMA Model Performance

From the results in section 4.1, we can see that the prediction error of ARIMA models increases with the increasing prediction window. This is because the variance and uncertainty increase with the expanding prediction window. Also, when the prediction window increases, the possibility of having additional factors influencing the housing price increases. Policy change is one of the possible factors, and ARIMA models cannot estimate the effect of the sudden changes on housing prices, which decreases the prediction accuracy.

From the results in section 4.2, we found that the performance of ARIMA models differs by the district. By looking at the trend and the range of housing prices in each district (Figure 11

& 12), we discovered that ARIMA models perform better when predicting the housing prices of districts with a smaller housing price range and smoother housing price increasing trend. This observation may be the result of many factors. For example, the district's location can influence the housing price range. Urban districts located closer to the center of Beijing experienced faster developments in the past years and thus experienced more rapid and significant price increases. On the other hand, districts located farther away from the economy center may experience fewer economic and political changes. The housing price in these districts may go through a slower and more stable increase in the price level. And for the housing price increase trend, more centered districts, like Chaoyang, tend to experience more sudden changes in their housing prices. These sudden changes create more variance and volatility in the data and weaken ARIMA model's performances.

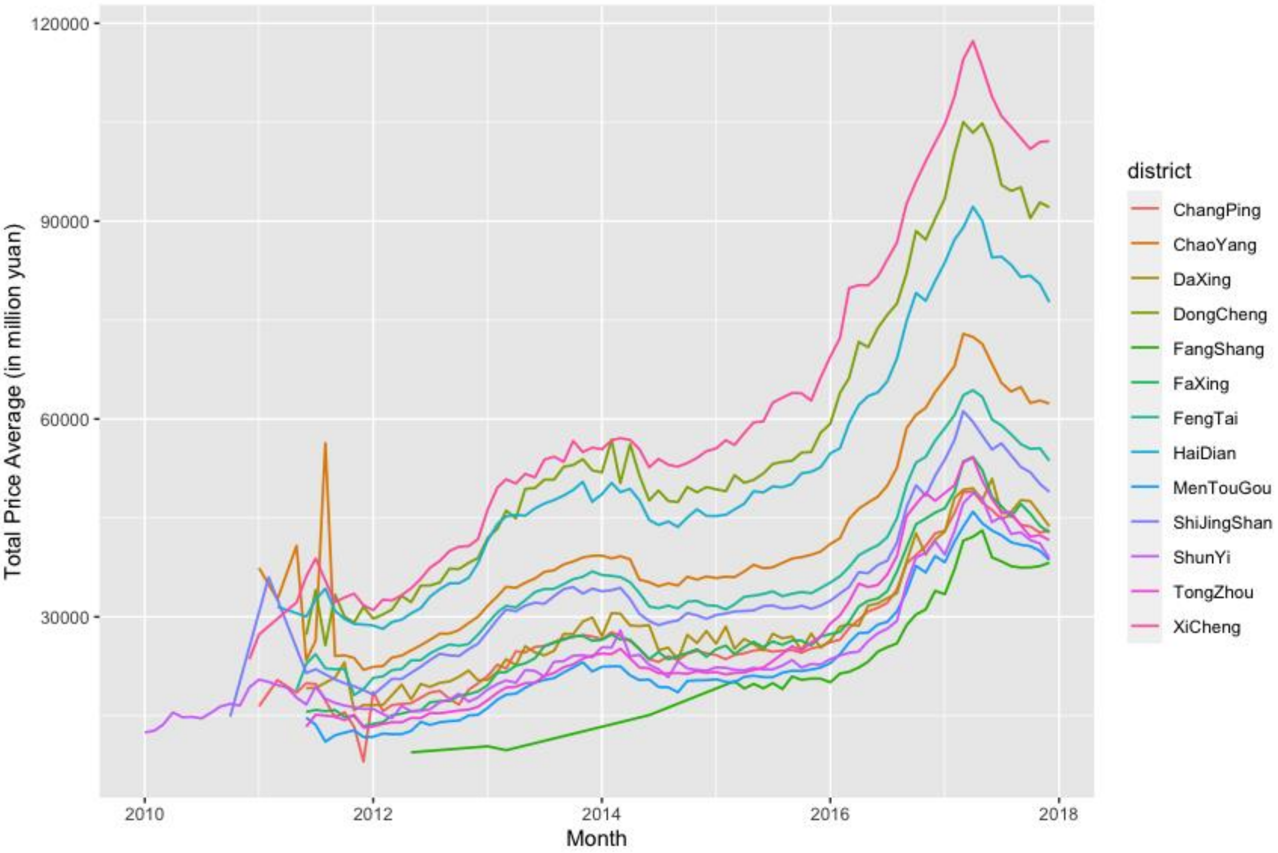


Figure 11. Housing Price Trend by District

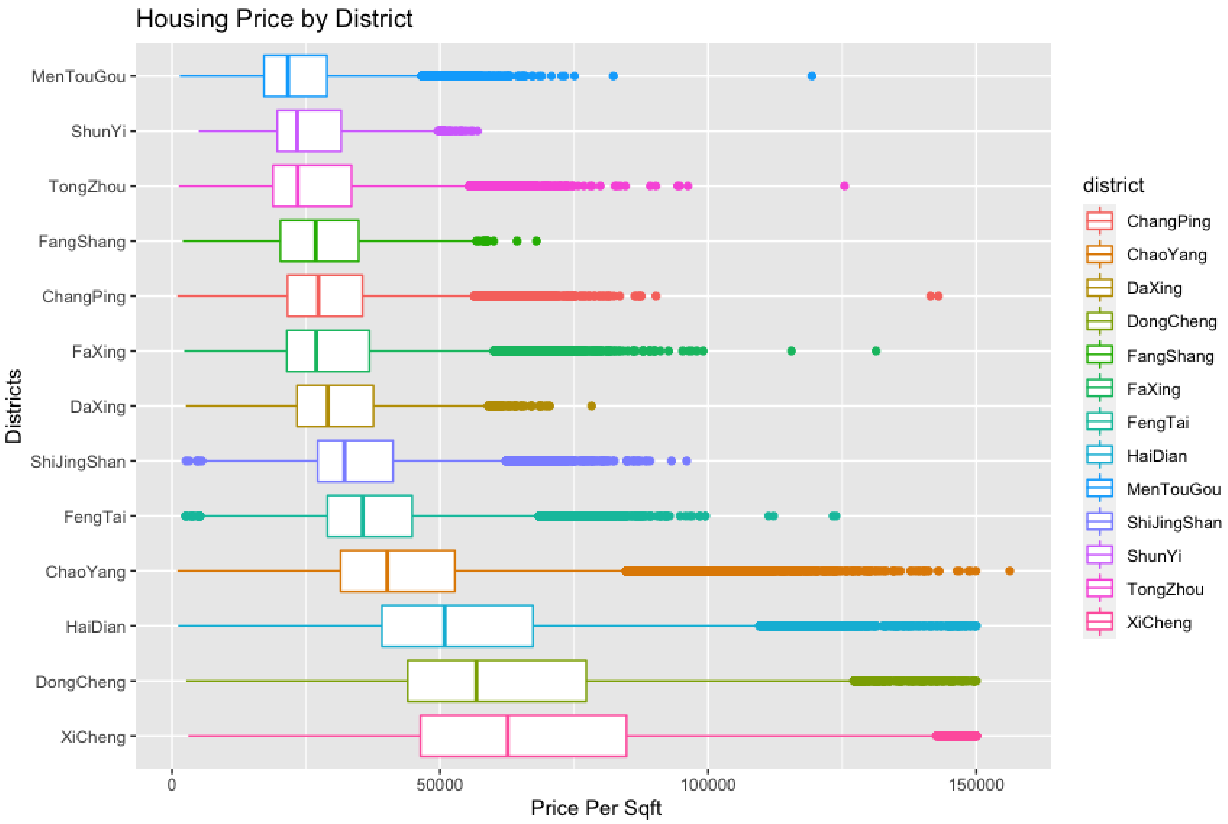


Figure 12. Housing Price Range by District

6. Conclusion

This study proposes and investigates two research questions: the performance of ARIMA models in forecasting housing prices in Beijing and the performance differences in predicting housing prices of different districts in Beijing. Answering these two research questions will help identify whether ARIMA is a suitable time series model in predicting housing prices in Beijing. It also provides valuable information on which districts' housing prices can be predicted more accurately with ARIMA models. To answer these two questions, 12 different ARIMA models are tested, and one best model, ARIMA (1, 0, 3), is selected based on the evaluation criteria. Also, the housing price dataset is divided based on the districts. The best fit ARIMA model is selected for each district, and the prediction accuracies are compared. Furthermore, this study shows that ARIMA models tend to perform better when the housing price have less variance.

Future studies can adopt other nonlinear time series models, such as Long-Term Short Memory (LSTM) or other hybrid models to predict housing prices in Beijing. Also, we can continue to analyze the reasons behind the performance differences when predicting the housing prices of different districts. Furthermore, it is also interesting to investigate ARIMA models' performance in predicting housing prices of other Chinese cities.

References

Bandyopadhyay, G., & Guha, B. (2016). Gold price forecasting using Arima model. Journal of Advanced Management Science, 117–121. https://doi.org/10.12720/joams.4.2.117-121

Box, G., & Jenkins, G. (1976). Time Series Analysis: Forecasting and Control. Holden-Day.

Contreras, J., Espinola, R., Nogales, F. J., & Conejo, A. J. (2003). Arima models to predict next-day electricity prices. IEEE Transactions on Power Systems, 18(3), 1014–1020. https://doi.org/10.1109/tpwrs.2002.804943

Dou, Z.-wu, Ji, M.-xin, Wang, M., & Shao, Y.-nan. (2021). Price prediction of pu’er tea based on Arima and BP Models. Neural Computing and Applications, 34(5), 3495–3511. https://doi.org/10.1007/s00521-021-05827-9

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: TheforecastPackage Forr. Journal of Statistical Software, 27(3). https://doi.org/10.18637/jss.v027.i03

Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and Arima models for time series forecasting. Applied Soft Computing, 11(2), 2664–2675. https://doi.org/10.1016/j.asoc.2010.10.015

Ngai, L. R., & Tenreyro, S. (2014). Hot and cold seasons in the housing market. American Economic Review, 104(12), 3991–4026. https://doi.org/10.1257/aer.104.12.3991

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia Housing Data. Expert Systems with Applications, 42(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

Qiu, L., Tu, Y., & Zhao, D. (2019). Information asymmetry and anchoring in the housing market: A stochastic frontier approach. Journal of Housing and the Built Environment, 35(2), 573–591. https://doi.org/10.1007/s10901-019-09701-y

Sallehuddin, R., & Hj. Shamsuddin, S. M. (2009). Hybrid grey relational artificial neural network and auto regressive integrated moving average model for forecasting time-series data. Applied Artificial Intelligence, 23(5), 443–486. https://doi.org/10.1080/08839510902879384

Shen, J. (2013). Increasing internal migration in China from 1985 to 2005: Institutional Versus Economic Drivers. Habitat International, 39, 1–7. https://doi.org/10.1016/j.habitatint.2012.10.004

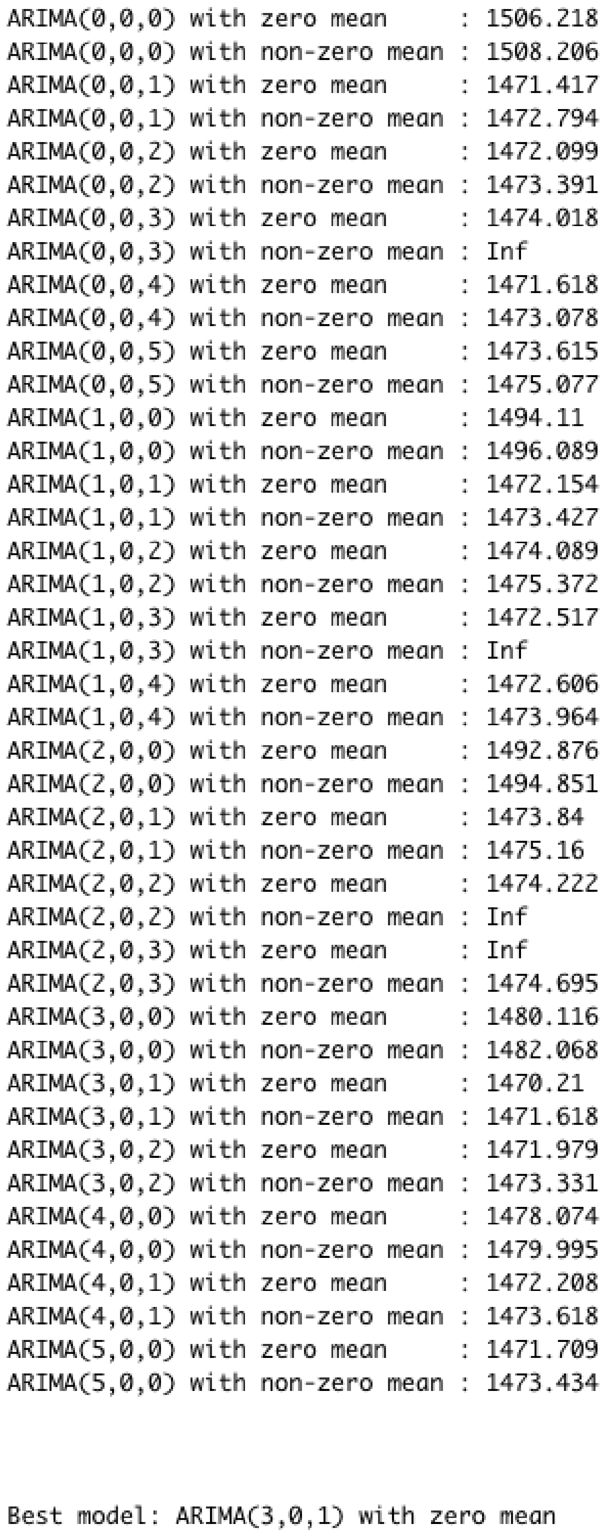
Temür, A. S., Akgün, M., & Temür, G. (2019). Predicting housing sales in Turkey using Arima, LSTM and hybrid models. Journal of Business Economics and Management, 20(5), 920–938. https://doi.org/10.3846/jbem.2019.10190

Wang, F., Zou, Y., Zhang, H., & Shi, H. (2019). House price prediction approach based on Deep Learning and Arima model. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). https://doi.org/10.1109/iccsnt47585.2019.8962443

Wu, L., & Brynjolfsson, E. (2009). The future of prediction: How google searches foreshadow housing prices and sales. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2022293

Appendix

Appendix 1. ARIMA Models AIC Score Summary



```
ARIMA(0,0,0) with zero mean     : 1506.218
ARIMA(0,0,0) with non-zero mean : 1508.206
ARIMA(0,0,1) with zero mean     : 1471.417
ARIMA(0,0,1) with non-zero mean : 1472.794
ARIMA(0,0,2) with zero mean     : 1472.099
ARIMA(0,0,2) with non-zero mean : 1473.391
ARIMA(0,0,3) with zero mean     : 1474.018
ARIMA(0,0,3) with non-zero mean : Inf
ARIMA(0,0,4) with zero mean     : 1471.618
ARIMA(0,0,4) with non-zero mean : 1473.078
ARIMA(0,0,5) with zero mean     : 1473.615
ARIMA(0,0,5) with non-zero mean : 1475.077
ARIMA(1,0,0) with zero mean     : 1494.11
ARIMA(1,0,0) with non-zero mean : 1496.089
ARIMA(1,0,1) with zero mean     : 1472.154
ARIMA(1,0,1) with non-zero mean : 1473.427
ARIMA(1,0,2) with zero mean     : 1474.089
ARIMA(1,0,2) with non-zero mean : 1475.372
ARIMA(1,0,3) with zero mean     : 1472.517
ARIMA(1,0,3) with non-zero mean : Inf
ARIMA(1,0,4) with zero mean     : 1472.606
ARIMA(1,0,4) with non-zero mean : 1473.964
ARIMA(2,0,0) with zero mean     : 1492.876
ARIMA(2,0,0) with non-zero mean : 1494.851
ARIMA(2,0,1) with zero mean     : 1473.84
ARIMA(2,0,1) with non-zero mean : 1475.16
ARIMA(2,0,2) with zero mean     : 1474.222
ARIMA(2,0,2) with non-zero mean : Inf
ARIMA(2,0,3) with zero mean     : Inf
ARIMA(2,0,3) with non-zero mean : 1474.695
ARIMA(3,0,0) with zero mean     : 1480.116
ARIMA(3,0,0) with non-zero mean : 1482.068
ARIMA(3,0,1) with zero mean     : 1470.21
ARIMA(3,0,1) with non-zero mean : 1471.618
ARIMA(3,0,2) with zero mean     : 1471.979
ARIMA(3,0,2) with non-zero mean : 1473.331
ARIMA(4,0,0) with zero mean     : 1478.074
ARIMA(4,0,0) with non-zero mean : 1479.995
ARIMA(4,0,1) with zero mean     : 1472.208
ARIMA(4,0,1) with non-zero mean : 1473.618
ARIMA(5,0,0) with zero mean     : 1471.709
ARIMA(5,0,0) with non-zero mean : 1473.434



Best model: ARIMA(3,0,1) with zero mean
```