House Price Index.tsd

Stat 451 Nonseasonal Project

Sijie Huang

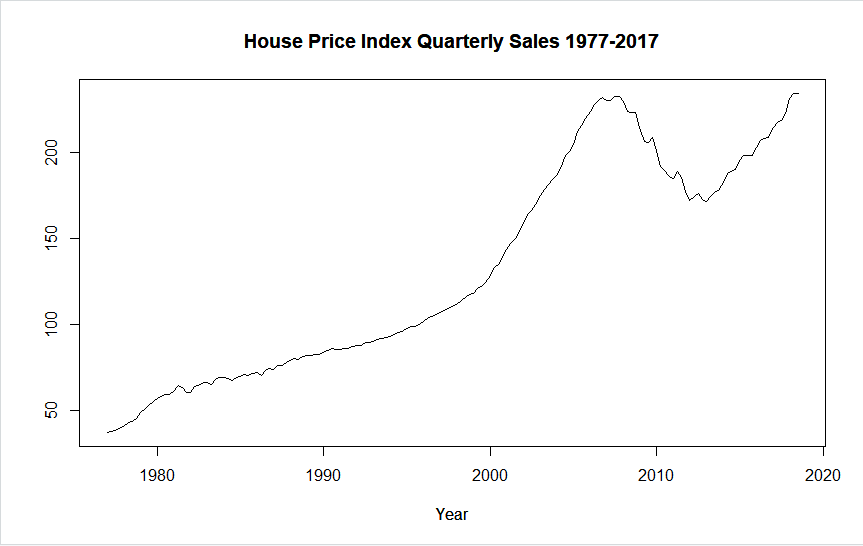
Tianchen Wang

Shangkun Zuo

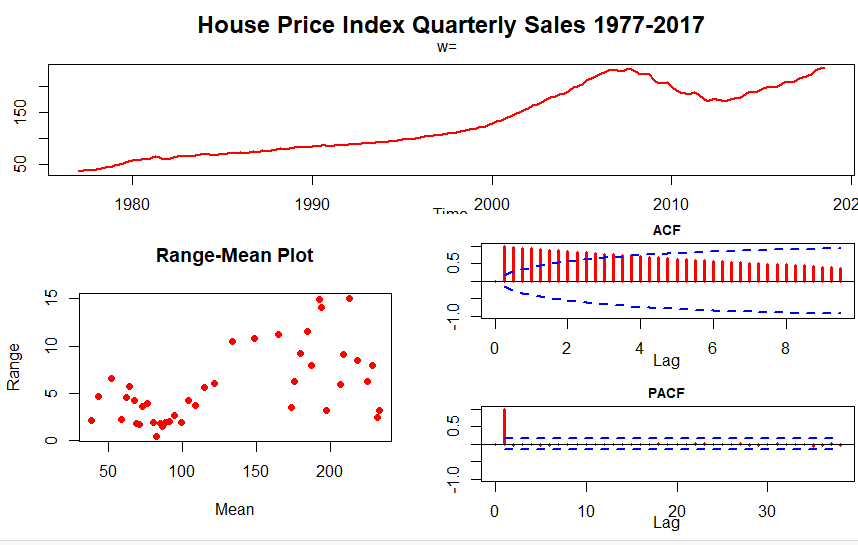
Executive Summary

Since we want the data more professional, we use the home of the U.S Government’s open data to find data and the publisher is federal housing finance agency. Because the house price is unlikely be affected by seasonal reasons, we decided to use House price indexes of Minneapolis-St.Paul-bloomington from 1977-2017. The URL for the original data is <https://catalog.data.gov/dataset/fhfa-house-price-indexes-hpis>, because we are interested in the house price index change based on year change, and the house price index is a weighted and repeat-sales index, so it’s not big, and we don’t need to rescale. The current scale is very easy and directly to see the house average price changes in repeat sales on the same properties over time. However, because the entire data is so big, we want to shorten the data. We simplified the data by filtering “Minneapolis-St. Paul-Bloomington, MN-WI. Since the original data at first quarter in 1976 is missing, we cut the whole year off because we want to have the same length of period in each year. The period we choose is quarterly, the total years from 1977-2017 is not that long, therefore we assume the current length of the realization we use is great. There are no any missing observations in our dataset. Our purpose of this time series model is to analyze the house index change from 1977 to 2017, and trying to predict the future trend. Since we want to choose a better fit model, we need to use model ARIMA(3,1,1) and the transformation of gamma=0.333.

I. Data Versus Time

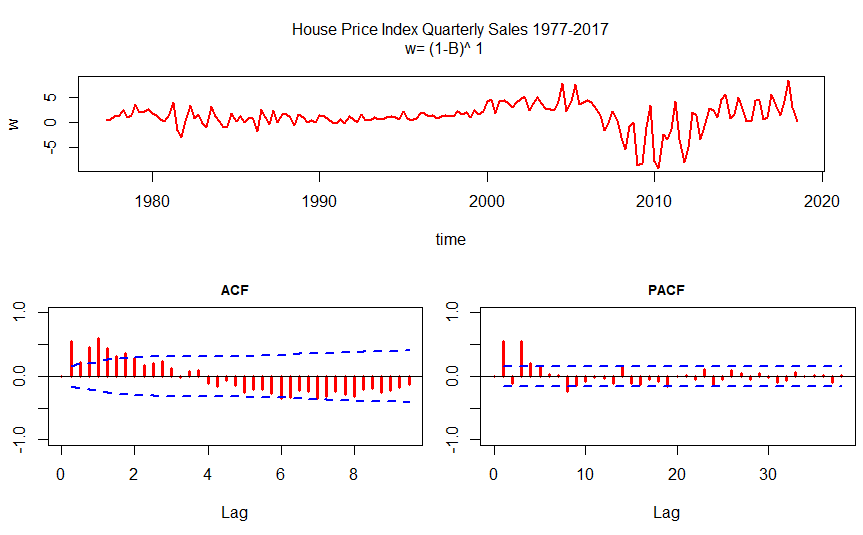


Overall, for our time series plot, there is a very obviously upward trend non-seasonal model, but a valley occurred in 2010 maybe because the inflation, and president Obama focus on the nation’s economy.

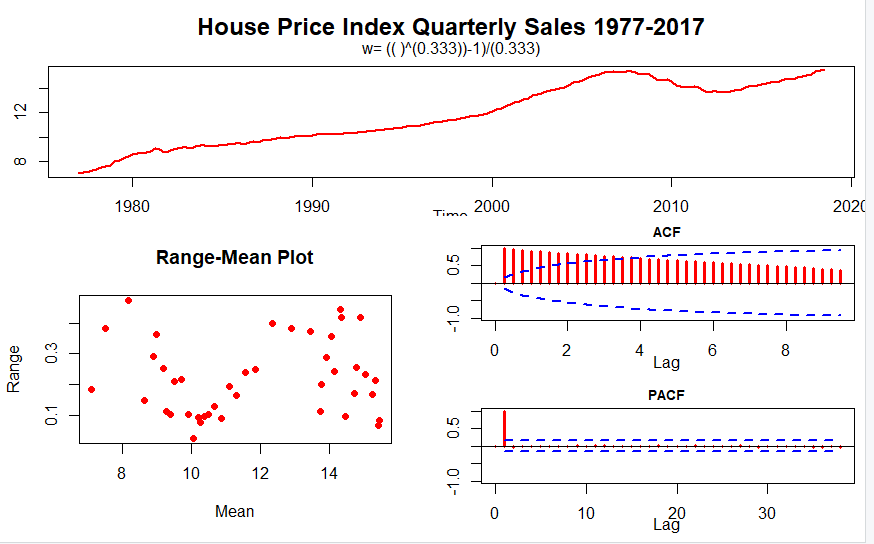
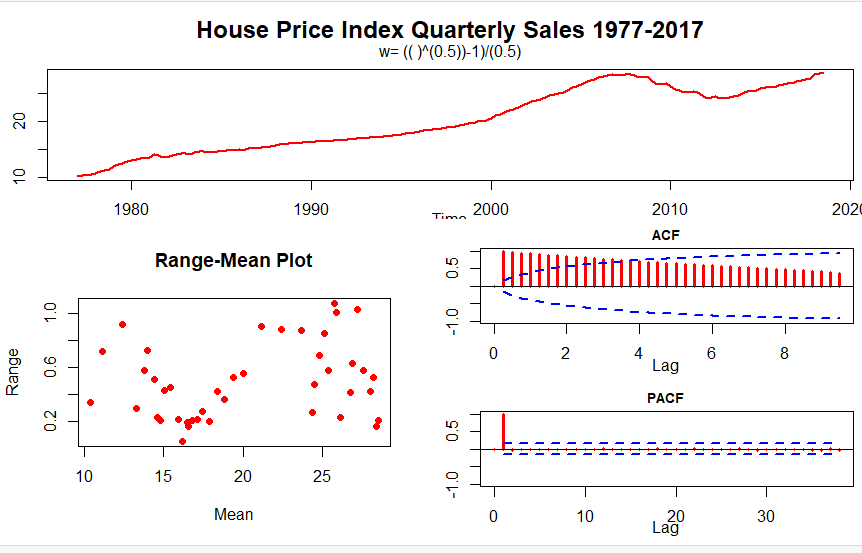
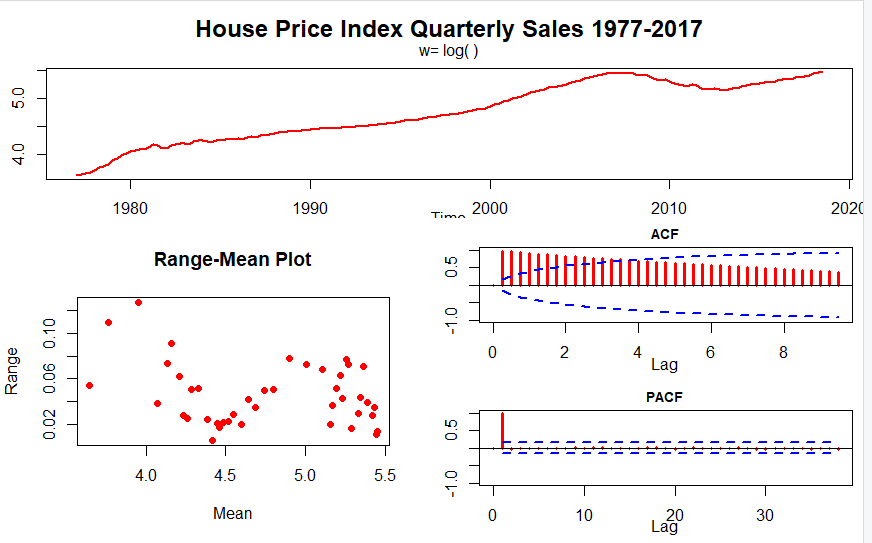
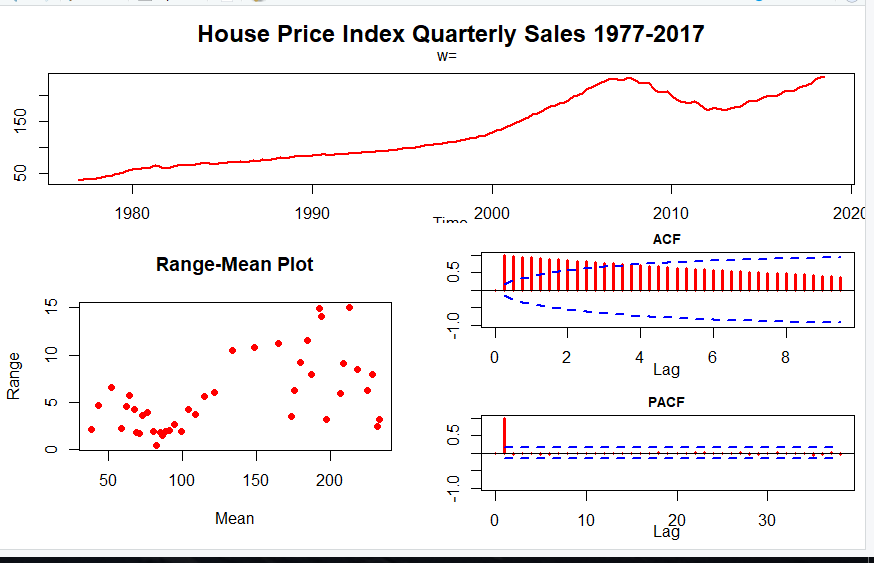


Possible Models for our Time Series

From iden function, we can see clearly that acf is dying down, while pacf cut off after p=1, this is an Autoregressive model, which can also be identified as ARIMA(1,0,0). Since there is a strong autocorrelation in original acf function, we used the first difference trying to stabilize the data. The result from the first difference still shows autocorrelation although it looks much smaller than the original. Eventually, we took the second difference and the output looks much more reasonable. As we look into the second different acf and pacf plot, it seems like both of them are dying down slowly, which can be classified as ARIMA model. Therefore, ARIMA and IMA are identified as possible models in this time series plot.



II. Check for Transformation

We use iden ( ) to do a range mean analysis to get an indication of whether a transformation is needed or not. 

With the log transformation, there’s a clear pattern in the range-mean plot. With the square root transformation, there’s a similar pattern as reciprocal cube-root, which is more random distributed and without exact pattern. But the range of reciprocal cube is smaller than the square root transformation. Compare to no transformation and the reciprocal cube transformation, using reciprocal cube transformation(gamma=.333) is better.

III. Table to Compare the Models

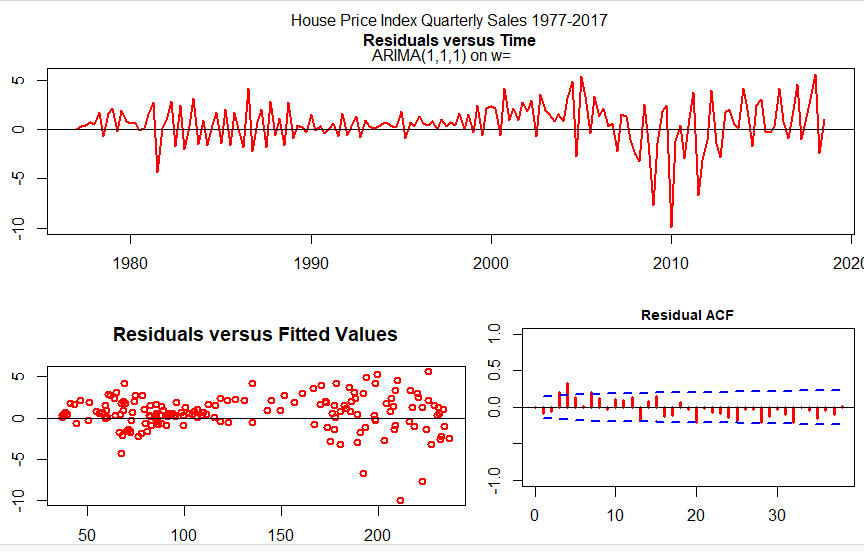
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | IMA(1,1) | ARIMA(1,1,1) | ARIMA(1,2,1) | ARIMA(3,1,1) |
| d | 1 | 1 | 2 | 1 |
| Ф1 | 0 | 0.2394 | 0.1907 | 0.9493 |
| θ1 | -0.7268 | -0.6099 | 0.7813 | -0.6120 |
| Sig. pk(a) | 1 | 1 | 2 | 0 |
| S | 2.2634 | 2.23 | 2.22 | 1.83 |
| AIC | 733.62 | 730.51 | 725.11 | 671.46 |
| -2log(Likelihood) | 729.62 | 724.51 | 719.11 | 661.46 |
| Ljung-Box x26 | 59.55 | 41.00 | 65.50 | 12.97 |

There are four models we have fit. They are IMA(1,1), ARIMA(1,1,1), ARIMA(1,2,1), ARIMA(3,1,1).

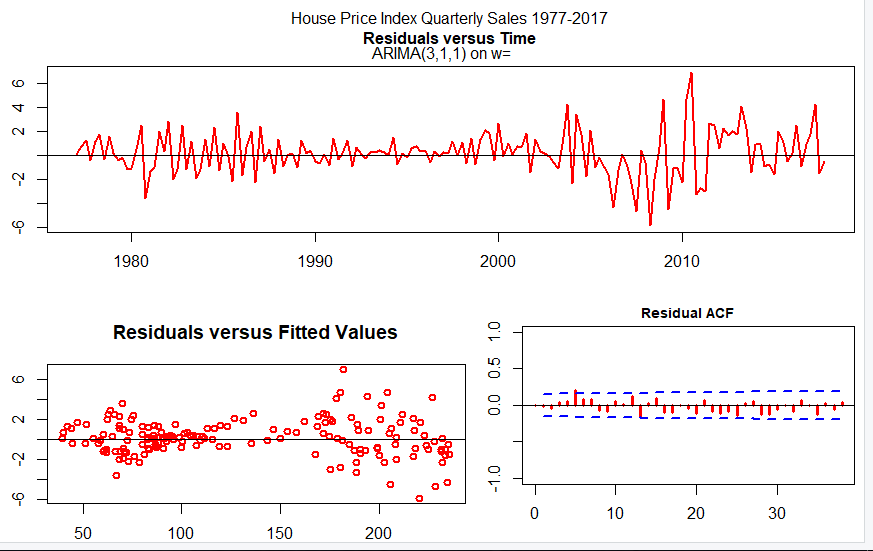
By comparing the four models above, there are two spikes in ARIMA(1,2,1) model which indicates we might over-differencing our model. IMA(1,1) has the largest AIC as well as the S value. We ended up choosing between ARIMA(1,1,1) and ARIMA(3,1,1), ARIMA(3,1,1) has a comparatively lower S and AIC value, not to mention that Ljung-Box value is significant. Moreover, there is no spike in residual ACF plot. Therefore, we consider ARIMA(3,1,1) model fits best to our data.

IV. esti() Graphical Output for the Best Two Models

ARIMA(1,1,1)



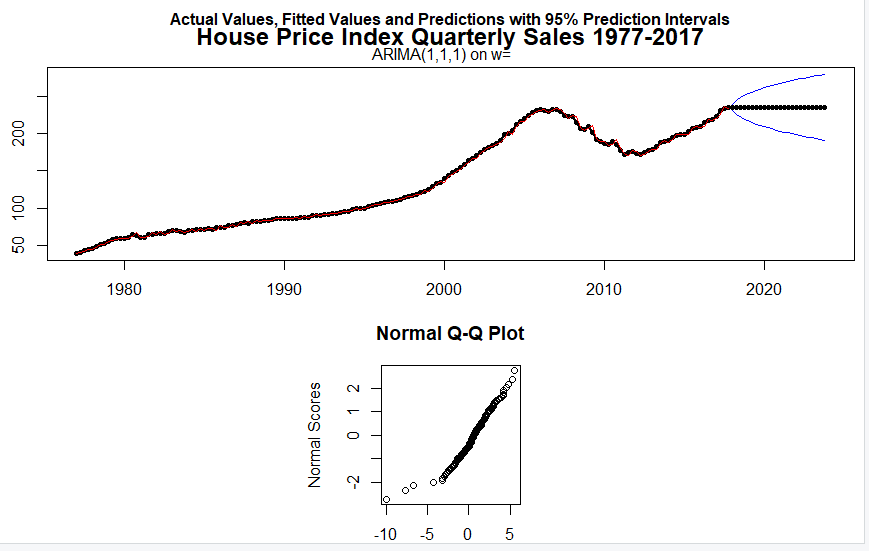
ARIMA(3,1,1)

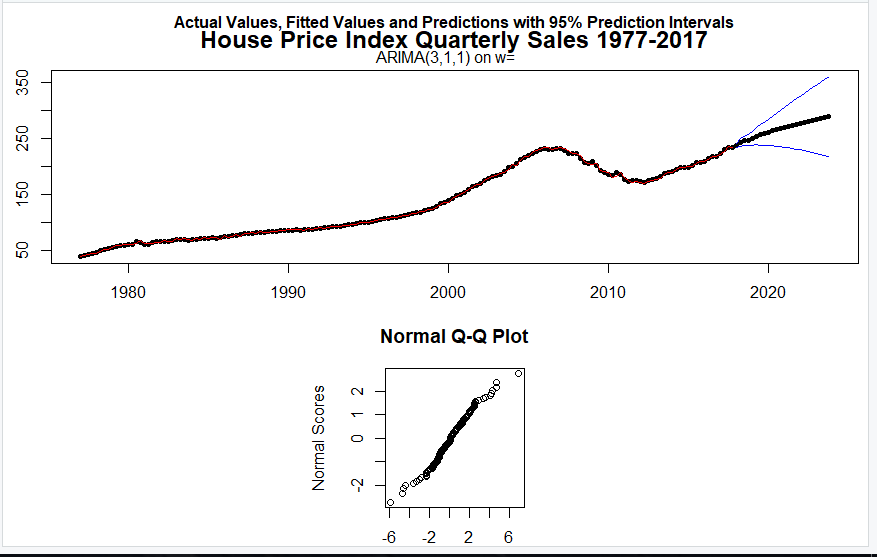


Since our data is highly non-stationary, we need to take differences before we fit a model. We used four different models, they are IMA(1,1), ARIMA(1,1,1), ARIMA(1,2,1),and ARIMA(3,1,1). We initially eliminate ARIMA(1,2,1) model, since it does not seem to improve the output by taking the second difference, IMA(1,1) and ARIMA(1,1,1) and ARIMA(3,1,1) are all look reasonable, we picked ARIMA(3,1,1) to be our final model is because there are no significant spike compared to the other two. Thus, we finally determined the best model that describes our data is ARIMA(3,1,1).

V. Forecasts for the Best Two Models

ARIMA(1,1,1)



ARIMA(3,1,1)

By looking at the forecasts for best two models we fit, there is not any curvature showed in the QQ plot, indicates the normality is met. The two prediction intervals behave completely different, ARIMA(1,1,1) is converge while ARIMA(3,1,1) diverges quickly. Furthermore, the forecast in ARIMA(3,1,1) follows the overall trend, on the other hand, the forecast in ARIMA(1,1,1) goes in a flat way and does not follow the trend. The reason is probably because ARIMA(3,1,1) predicts three steps ahead thus it has a longer realization than ARIMA(1,1,1).

VI. Conclusion

In a word, we do believe the ARIMA(3,1,1) is the best model. Compare to all iden() we made, we think that using reciprocal cube transformation(gamma=.333) is the best. Considering the esti(), ARIMA(3,1,1) to be our final model because there are no significant spike compared to the other models. By Comparing the table of our four models, ARIMA(3,1,1) has a comparatively lower S and AIC value, not to mention that Ljung-Box value is significant. Moreover, there is no spike in residual ACF plot. Therefore, ARIMA(3,1,1) model fits best to our data. Also, on the one hand, the forecast in ARIMA(3,1,1) follows the overall trend; on the other hand, the forecast in ARIMA(1,1,1) goes in a flat way and does not follow the trend because ARIMA(3,1,1) predicts three steps ahead thus it has a longer realization than ARIMA(1,1,1). Consequently, ARIMA(3,1,1) is our best model to do the time series analysis and forcasting.