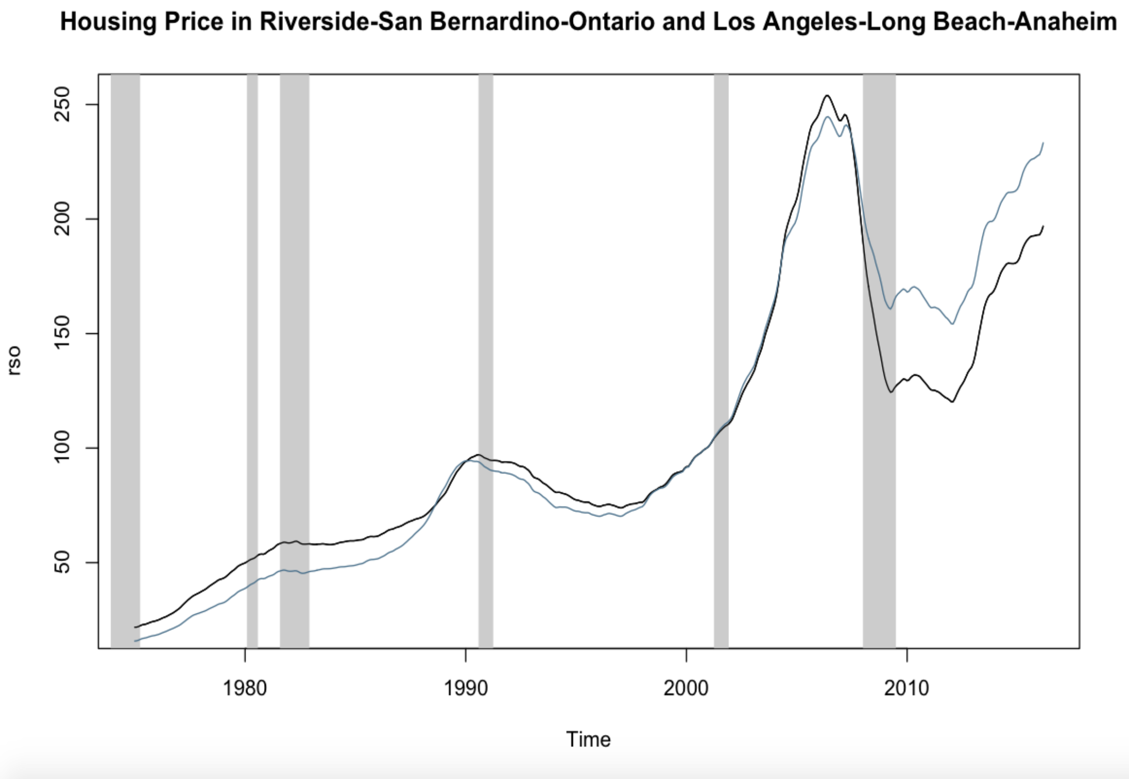
Final Report

April 2017
Author: Cong Cao

I. Introduction

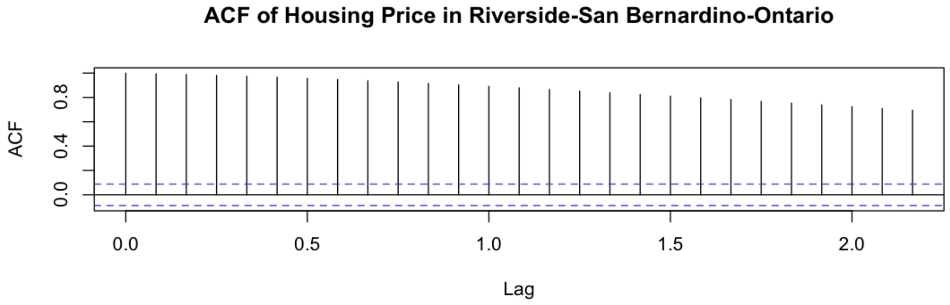
In this project, we examined two time series, the housing prices of Riverside-San Bernardino-Ontario district and the housing prices of Los Angeles-Long Beach-Anaheim district. In real world, future housing prices are always difficult to forecast since they change over time and may relate to the housing prices of other different areas. Thus, we are interested in whether the two time series present some relationships. The data spans from January 1975 to March 2016, and it is collected monthly. It is reasonable to conclude that the data is able to capture the full characteristics of the housing prices in these two districts since the data covers a long time period. We analyzed the data through multiple statistical tools such as ACF and PACF, and we will be able to conclude a fairly complete relationships and models for the time series. The further details are presented in the report.

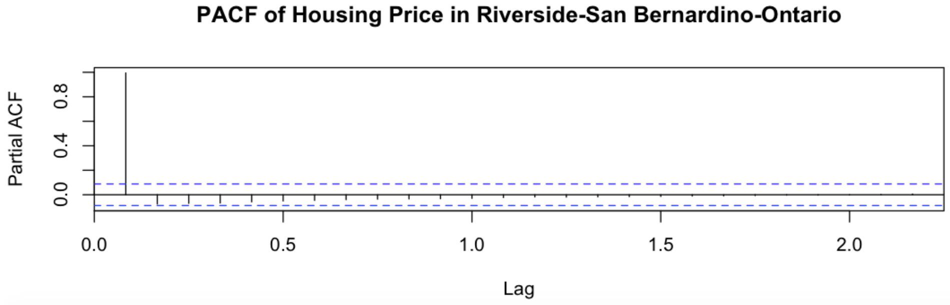
II. Results

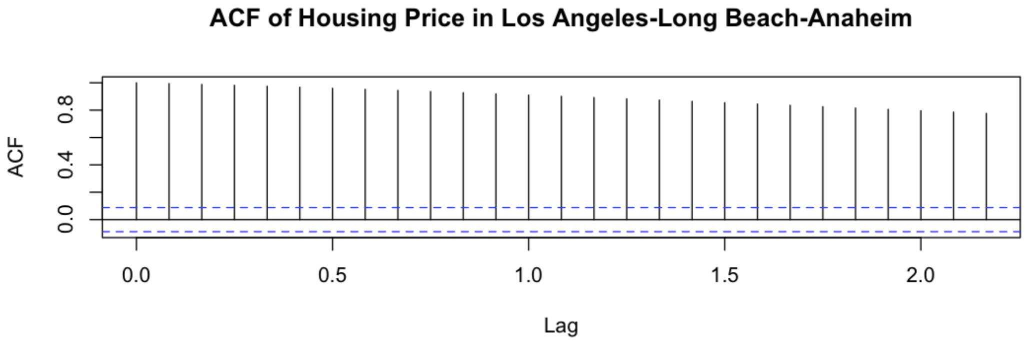

Housing Price in Riverside-San Bernardino-Ontario and Los Angeles-Long Beach-Anaheim

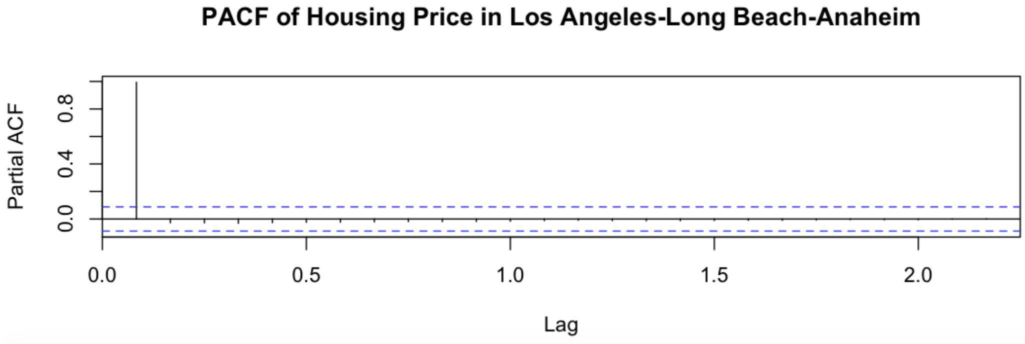
Discussion:

The graph above shows the housing prices of RSO (Riverside-San Bernardino-Ontario) area (Black line) and LLA (Los Angeles-Long Beach-Anaheim) area (Blue line); the housing prices of both areas increase slowly from year 1975 to 1990. Then the housing prices suddenly increase from $90 to $250 from the year 2000 to 2008; the housing prices dropped dramatically from 2008 to 2010, which was consistent with the financial crisis and broken housing bubble during this time period. Moreover, since the year 2010, the housing price in LLA turned to be a little bit higher than it in RSO. In addition, we can see from the graph above, that the housing prices for these two areas move in the same pattern, so maybe we could expect an inter-effect of them.



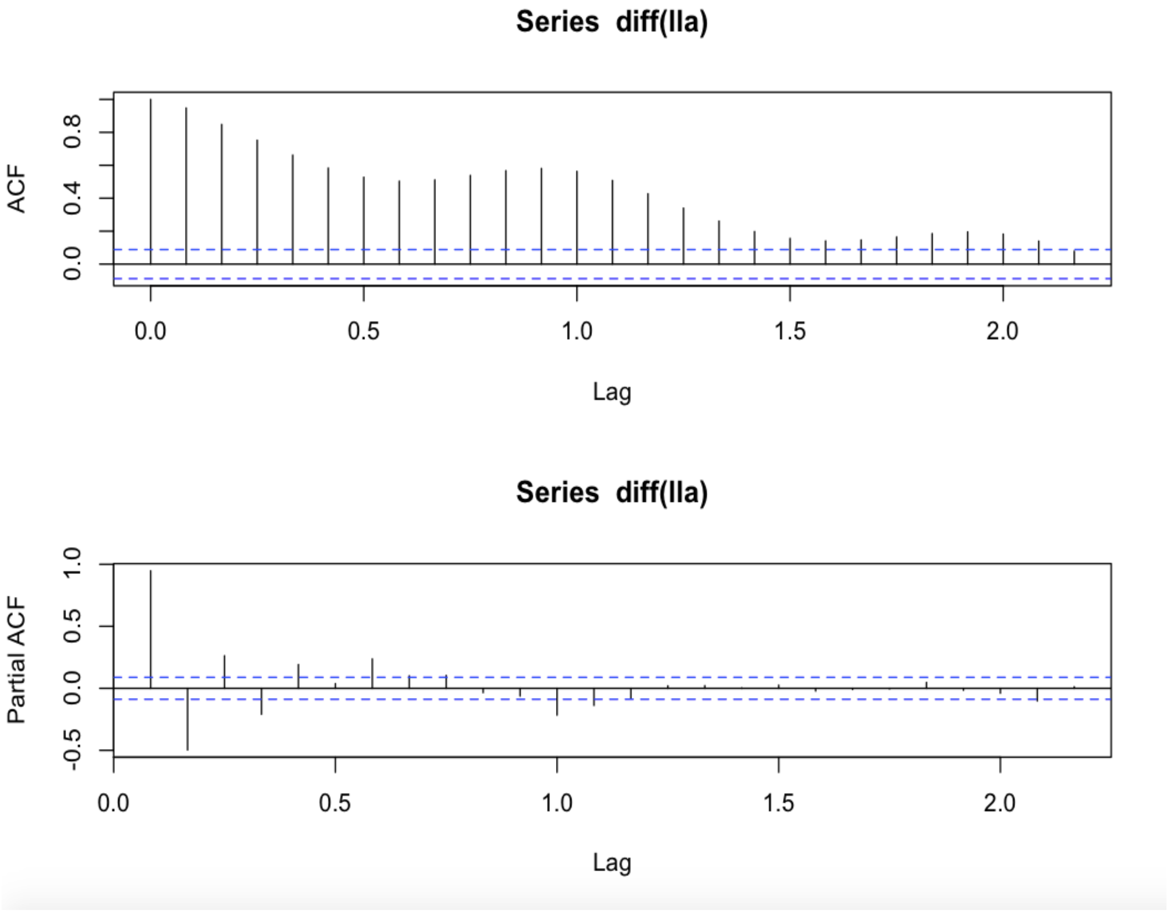


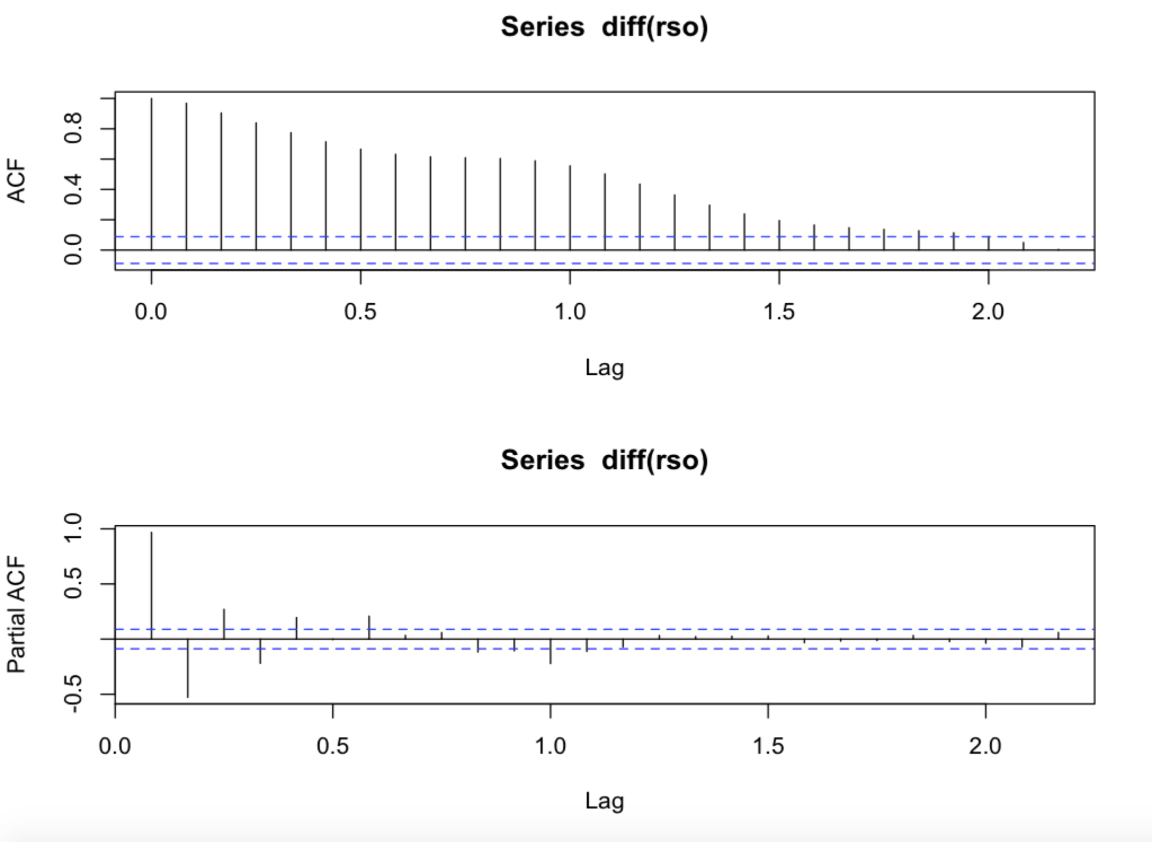




Discussion:

Then we plot the ACF and PACF of housing prices. Both of the ACF graphs show very strong persistence; and both PACF graphs show statistically significant spike at order one. It seems that there is a strong dependence on how housing prices change overtime, and the housing prices of previous month are strongly correlated to the housing prices of the next month.



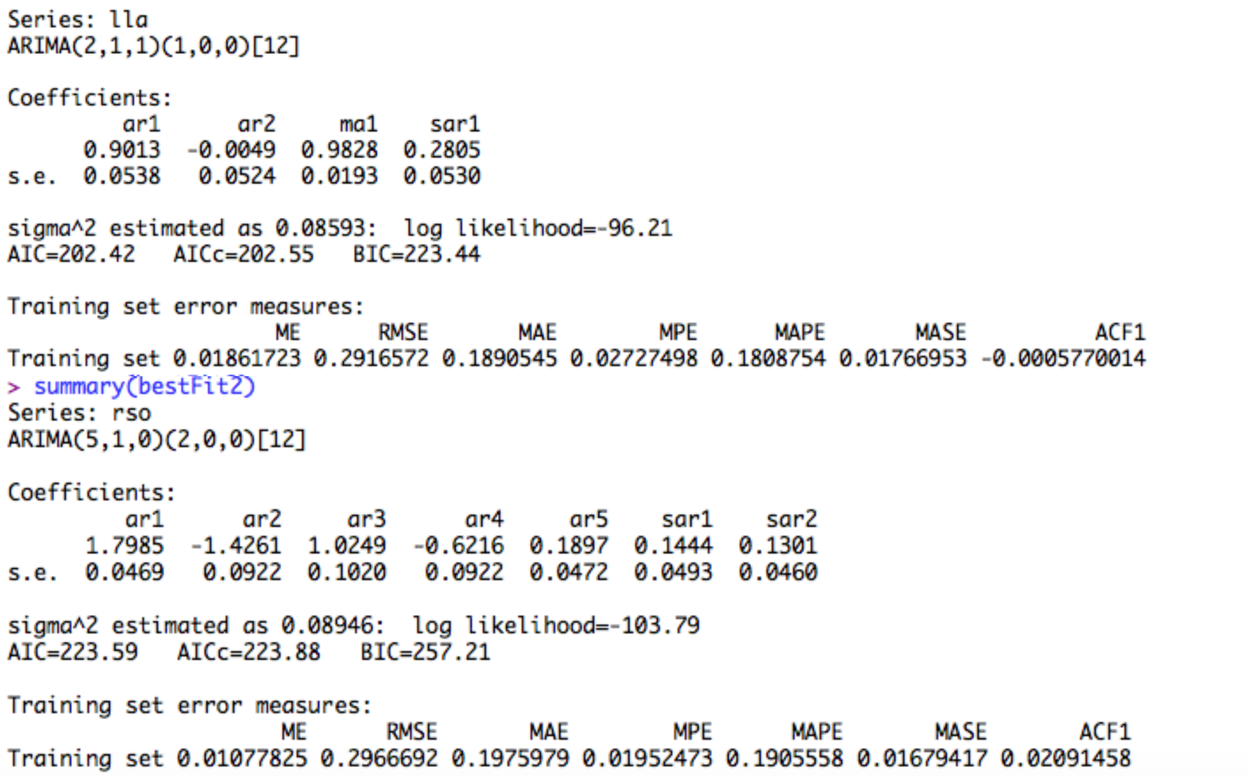


Discussion:

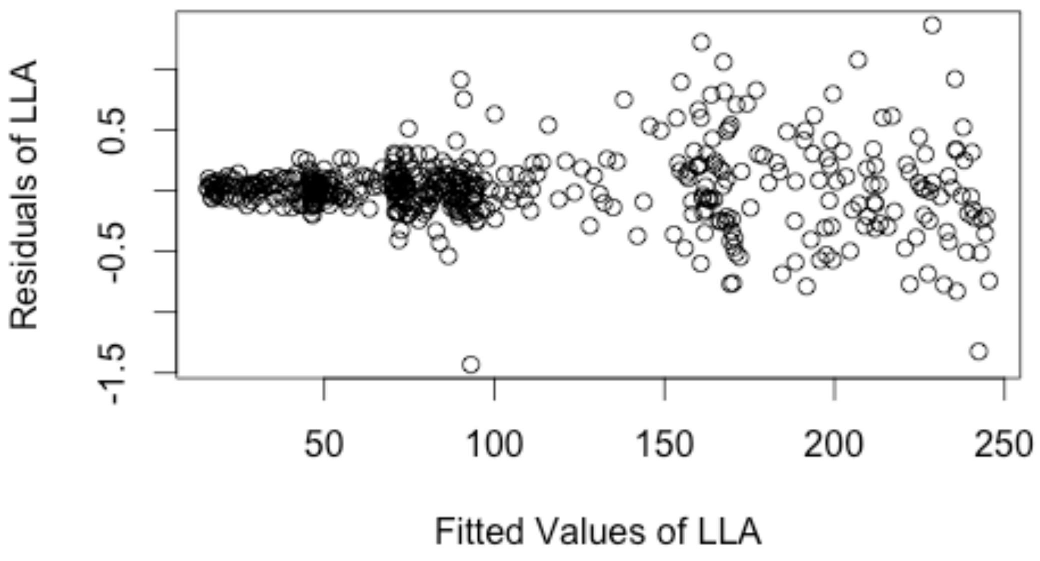
Since the ACF and PACF for the original data seems hardly tell how we can address them with a proper model, so we take first difference of our original data.

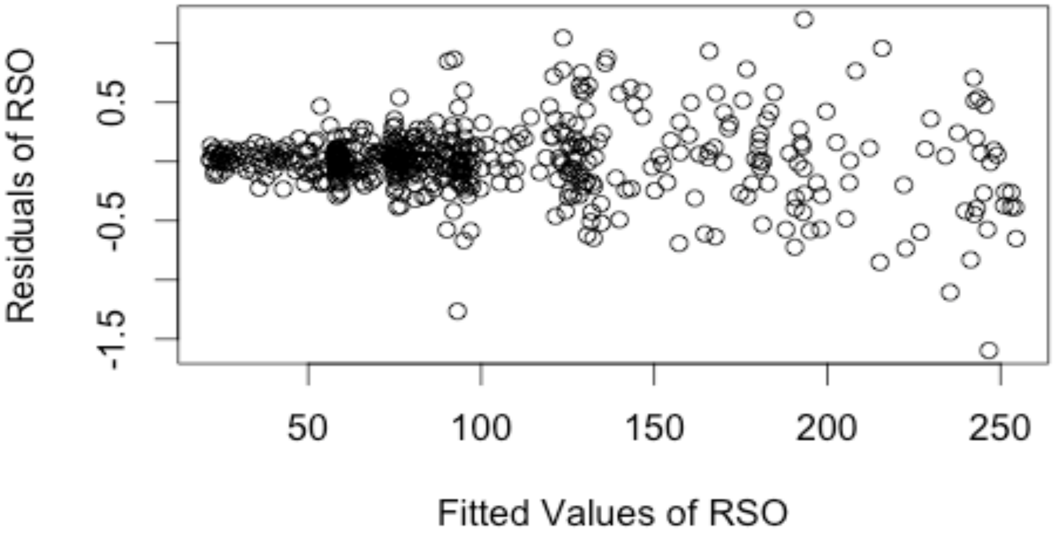
Now, we could apparently see 2 spikes in LLA's PACF and 5 spikes in RSO's PACF; a wavy pattern in the ACF of LLA and RSO. The wavy pattern suggests us there is a seasonal factor need to be considered.

Hence, we conclude a ARIMA (2,1,1)(1,0,0) model for LLA housing price, and a ARIMA(5,1,0)(2,0,0) for RSO housing price.



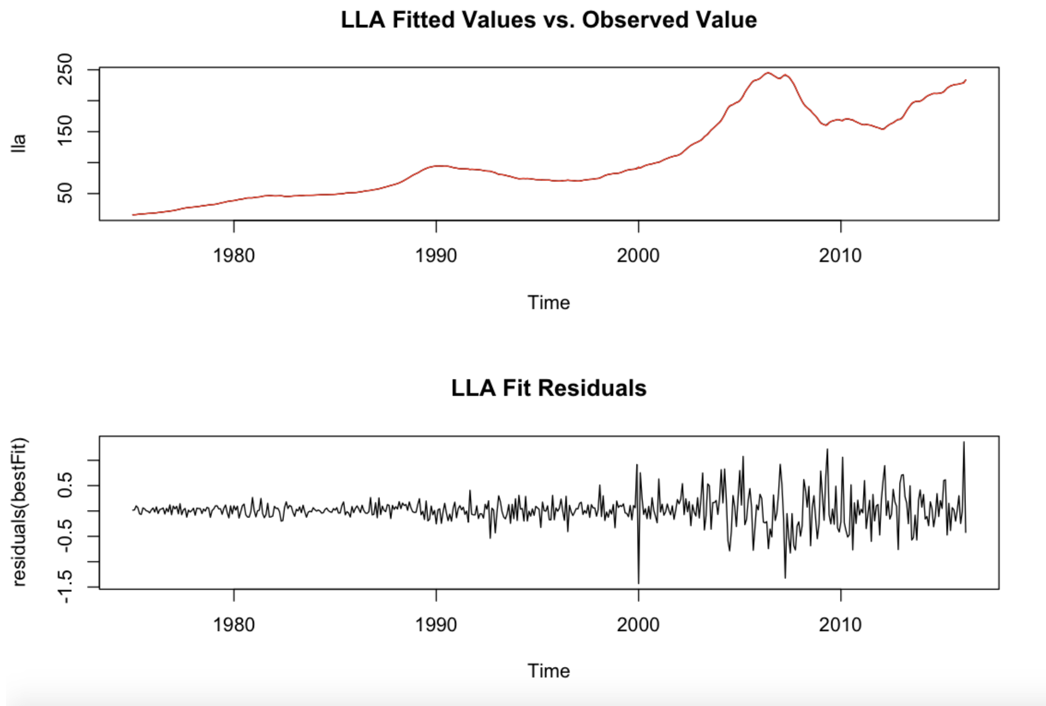
```
Series: lla
ARIMA(2,1,1)(1,0,0)[12]

Coefficients:
         ar1      ar2     ma1    sar1
      0.9013  -0.0049  0.9828  0.2805
s.e.  0.0538   0.0524  0.0193  0.0530

sigma^2 estimated as 0.08593:  log likelihood=-96.21
AIC=202.42   AICc=202.55   BIC=223.44

Training set error measures:
                     ME      RMSE       MAE        MPE      MAPE       MASE          ACF1
Training set 0.01861723 0.2916572 0.1890545 0.02727498 0.1808754 0.01766953 -0.0005770014
> summary(bestFit2)
Series: rso
ARIMA(5,1,0)(2,0,0)[12]

Coefficients:
         ar1      ar2     ar3      ar4     ar5    sar1    sar2
      1.7985  -1.4261  1.0249  -0.6216  0.1897  0.1444  0.1301
s.e.  0.0469   0.0922  0.1020   0.0922  0.0472  0.0493  0.0460

sigma^2 estimated as 0.08946:  log likelihood=-103.79
AIC=223.59   AICc=223.88   BIC=257.21

Training set error measures:
                     ME      RMSE       MAE        MPE      MAPE       MASE       ACF1
Training set 0.01077825 0.2966692 0.1975979 0.01952473 0.1905558 0.01679417 0.02091458
```



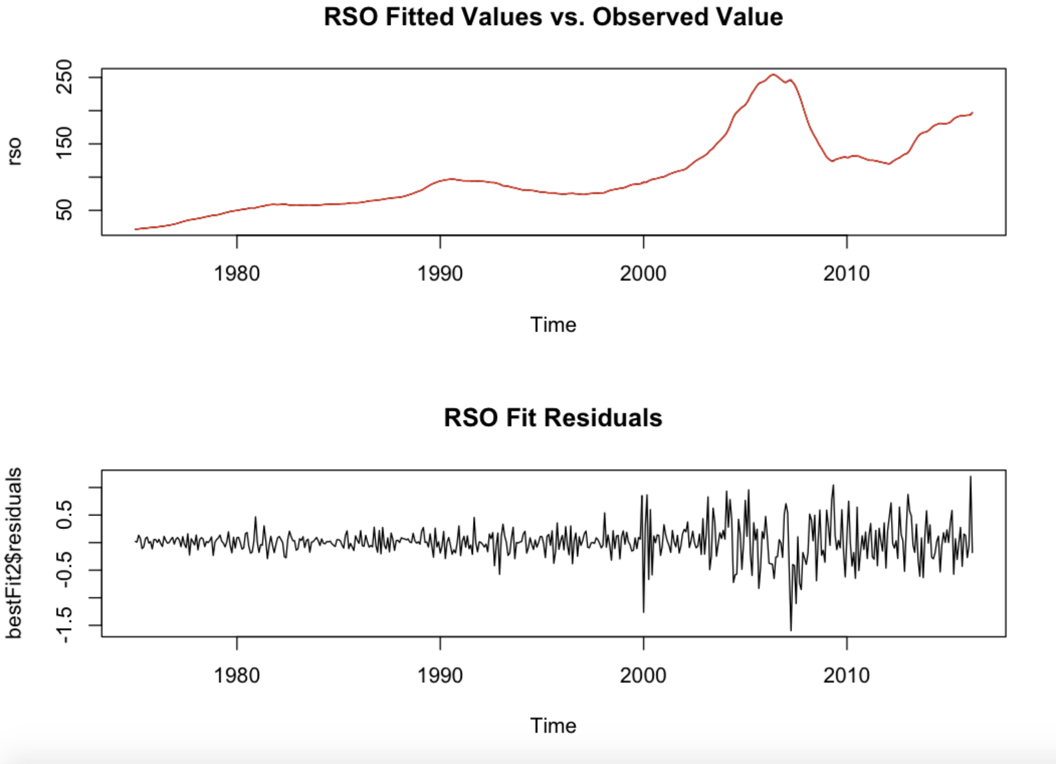


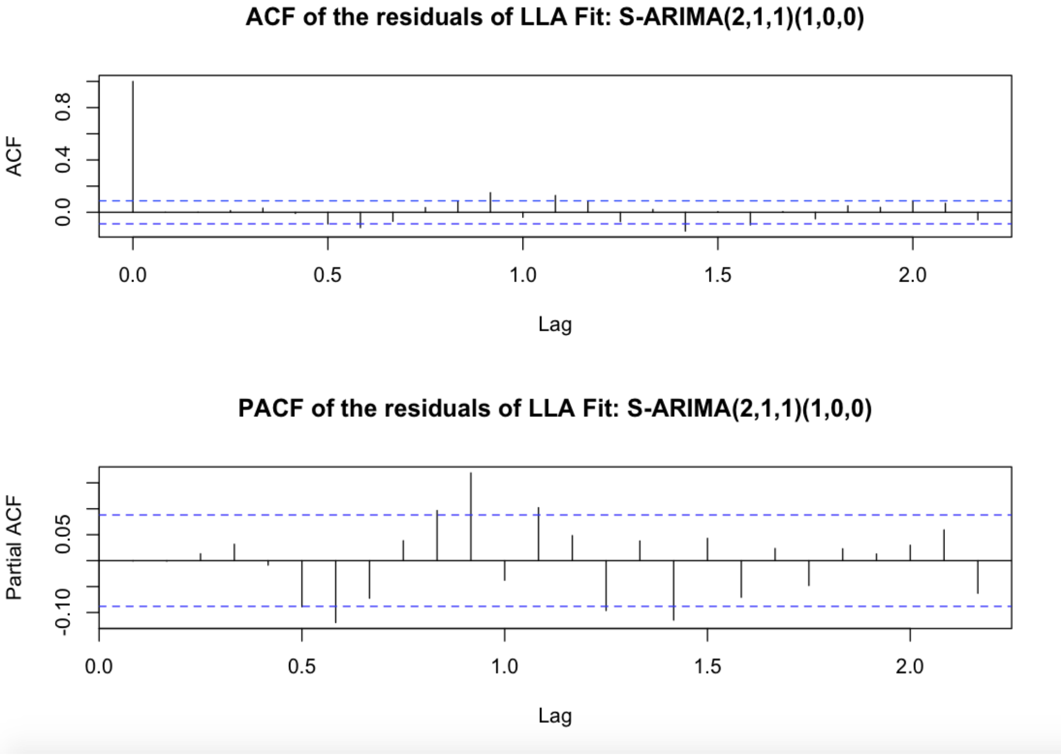
Discussion:

From these two plots, we can see that residuals of two models are scattered around zero. But there seems a pattern between fitted values and residuals. Residuals seem to be more scattered as fitted values become larger, which may indicate as heteroscedasticity.



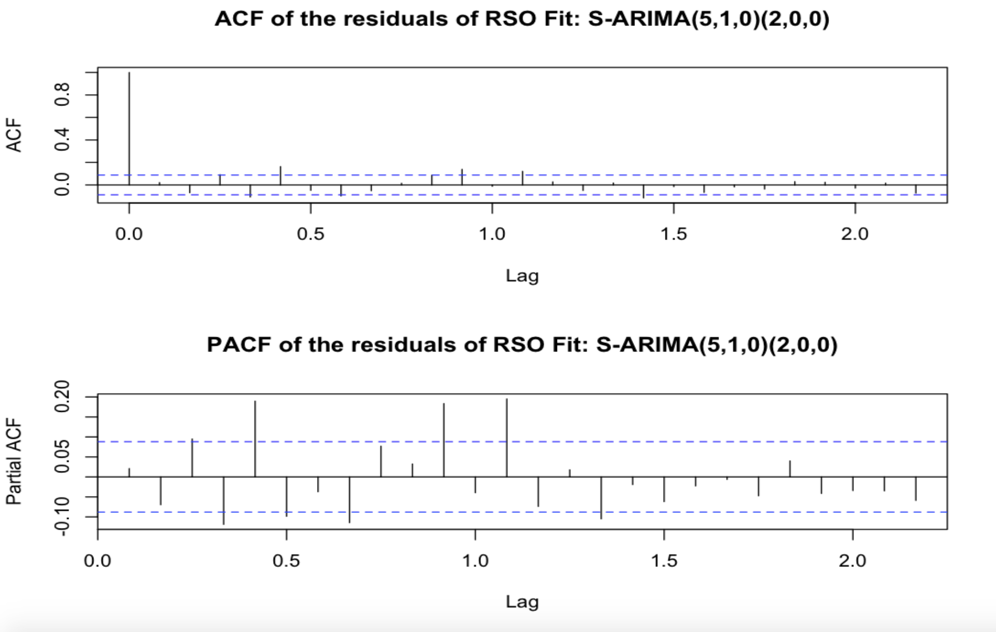
the graph above shows the LLA line of LLA





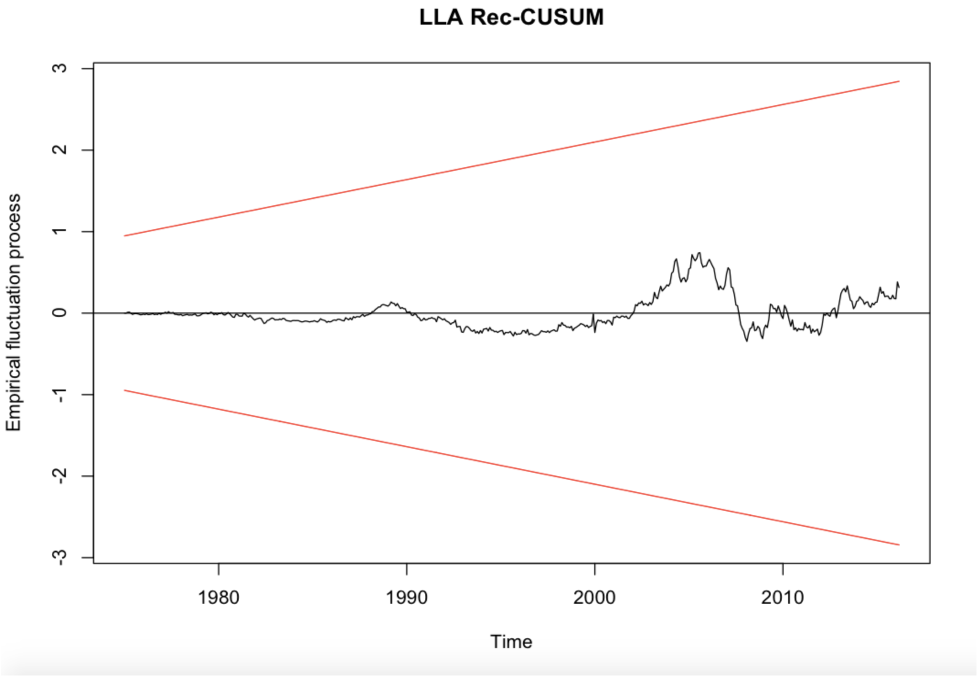
Discussion:

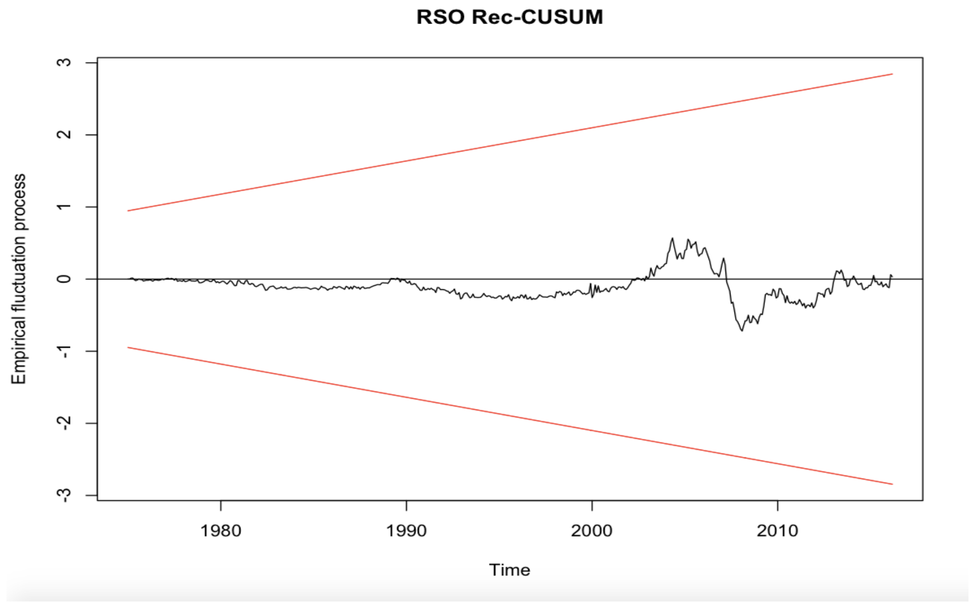
From the ACF plot, most values are not statistically significant except lag=0, which is an indication of a good fit. But in the PACF plot, there seems still some dynamics in the residuals. Fortunately, most are not quite statistically significant.



Discussion:

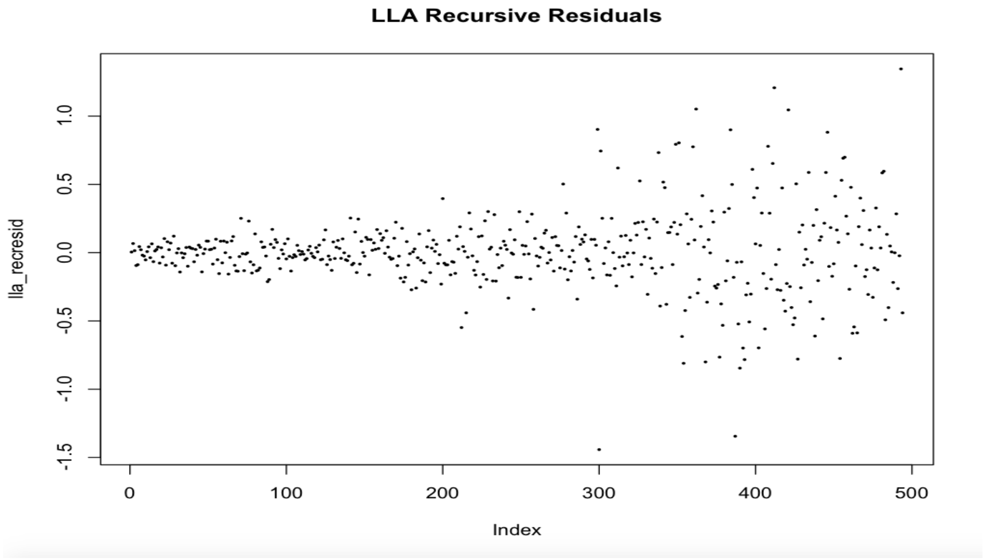
Similarly, in the PACF, we can see some dynamics. But those values are not quite statistically significant.

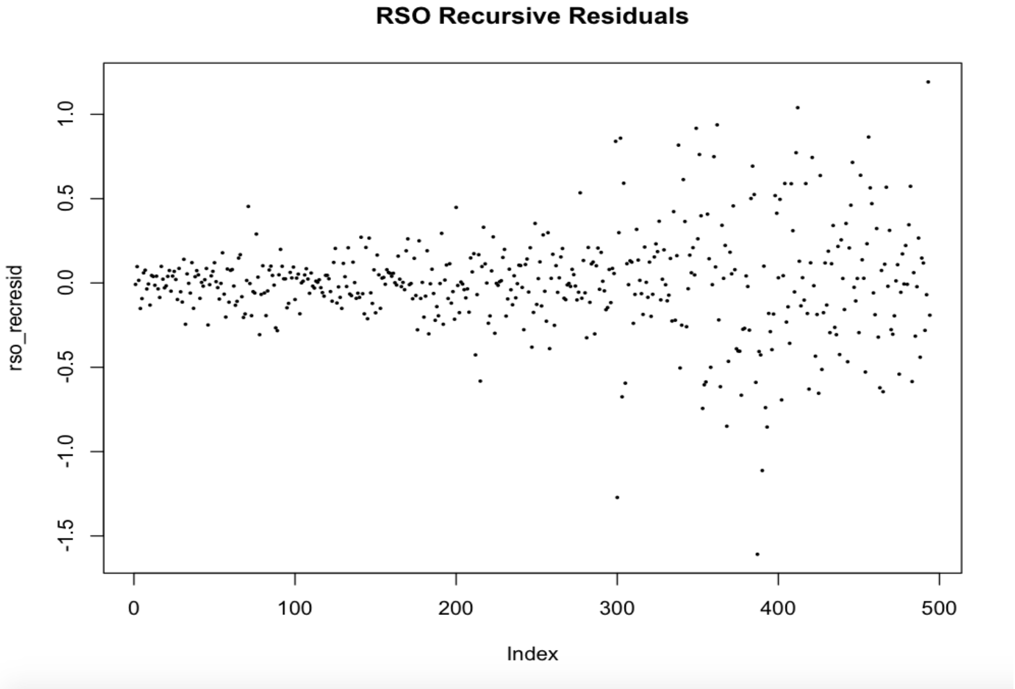




Discussion:

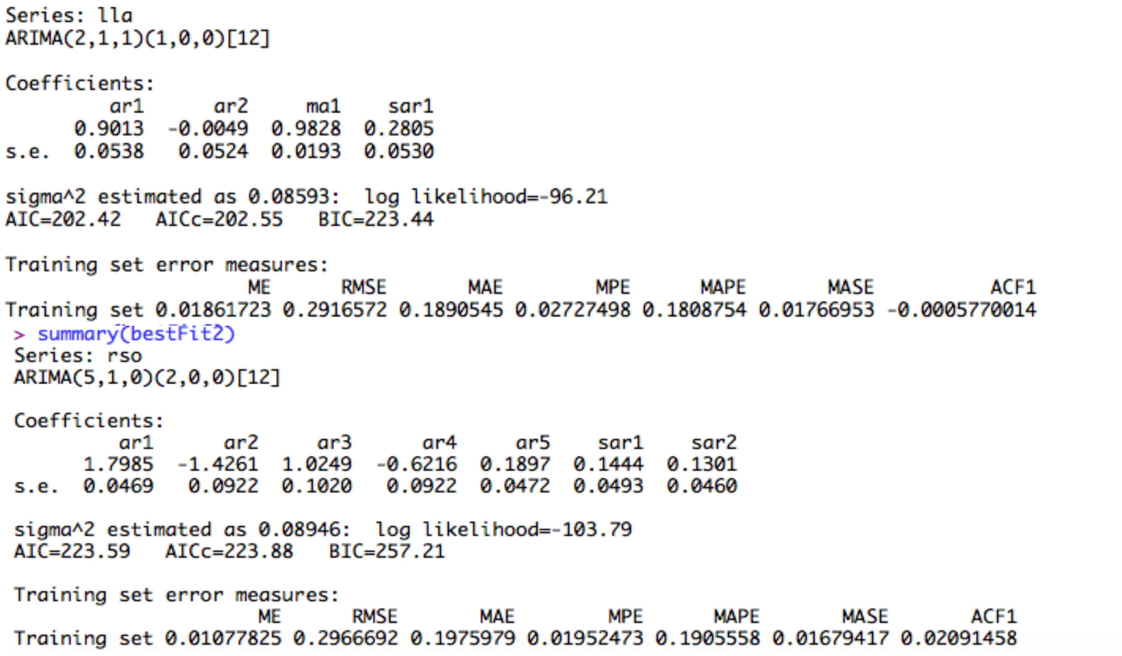
From these two plots, we can see that the estimated model parameters are quite stable over time. Although there might be some fluctuation during 2002-2010, they all stays within the confidence band. Thus, the models do not break in between and we do not need to change to difference models for different time period.





Discussion:

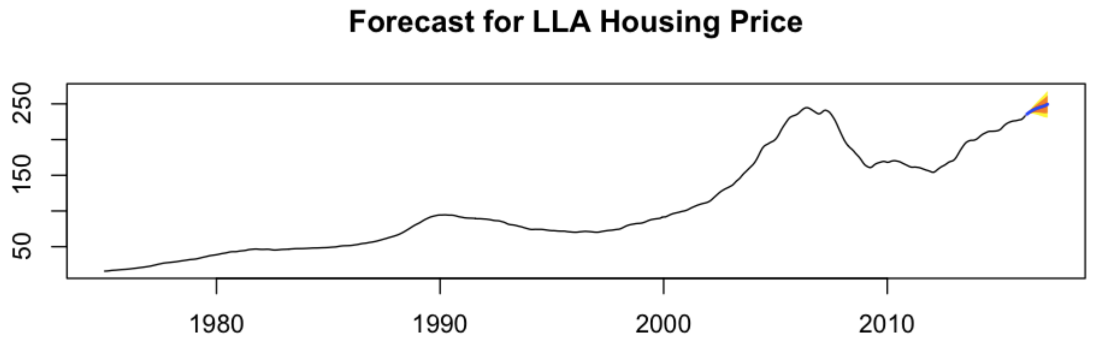
From these two plots, we can see that Recursive Residuals become more spread out as index increase. So the variance of Recursive Residuals is larger when index is larger. This might indicate that the model fits better to the data before 1995.

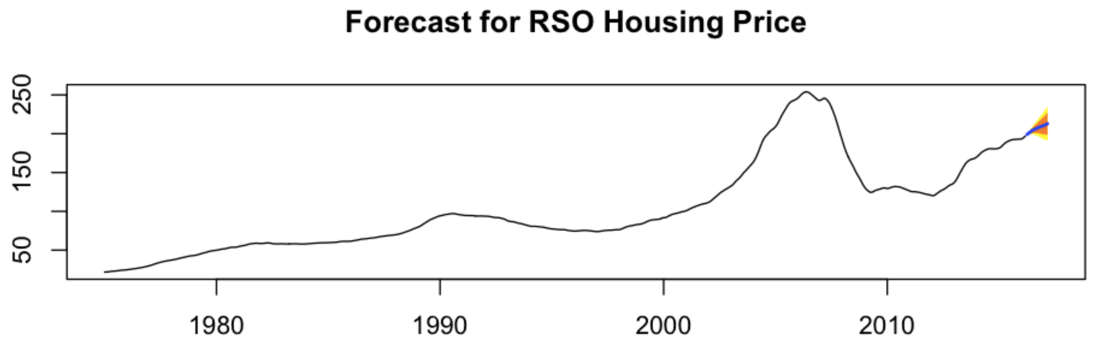

```
Series: lla
ARIMA(2,1,1)(1,0,0)[12]

Coefficients:
         ar1      ar2     ma1    sar1
      0.9013  -0.0049  0.9828  0.2805
s.e.  0.0538   0.0524  0.0193  0.0530

sigma^2 estimated as 0.08593:  log likelihood=-96.21
AIC=202.42   AICc=202.55   BIC=223.44

Training set error measures:
                     ME      RMSE       MAE        MPE      MAPE       MASE          ACF1
Training set 0.01861723 0.2916572 0.1890545 0.02727498 0.1808754 0.01766953 -0.0005770014
> summary(bestFit2)
Series: rso
ARIMA(5,1,0)(2,0,0)[12]

Coefficients:
         ar1      ar2     ar3      ar4     ar5    sar1    sar2
      1.7985  -1.4261  1.0249  -0.6216  0.1897  0.1444  0.1301
s.e.  0.0469   0.0922  0.1020   0.0922  0.0472  0.0493  0.0460

sigma^2 estimated as 0.08946:  log likelihood=-103.79
AIC=223.59   AICc=223.88   BIC=257.21

Training set error measures:
                     ME      RMSE       MAE        MPE      MAPE       MASE       ACF1
Training set 0.01077825 0.2966692 0.1975979 0.01952473 0.1905558 0.01679417 0.02091458
```

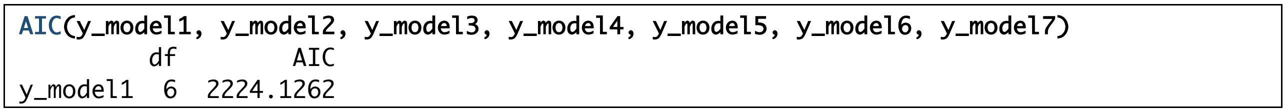
Discussion:

From LLA model we choose, the coefficient of ar1 and ma1 is quite close to 1, which indicates that the data more rely on the previous 1-period statistics. The s-AR estimate is around 0.28 which is not that significant, but still has a 0.28 influence; it consistent with the ACF wavy pattern.

From RSO model we choose, the data positively more rely on 1-period lag (1.8) and 3-period(1.02) lag data, and negatively rely on 2-period(-1.4) lag and 4-period(-0.62) lag data. The s-AR estimates are around 0.14 and 0.13 for 1-period lag and 2-period lag respectively; They not that significant, but still has somehow influence; it consistent with the ACF pattern.


Forecast for LLA Housing Price


Forecast for RSO Housing Price


```
AIC(y_model1, y_model2, y_model3, y_model4, y_model5, y_model6, y_model7)
         df        AIC
y_model1  6  2224.1262
```

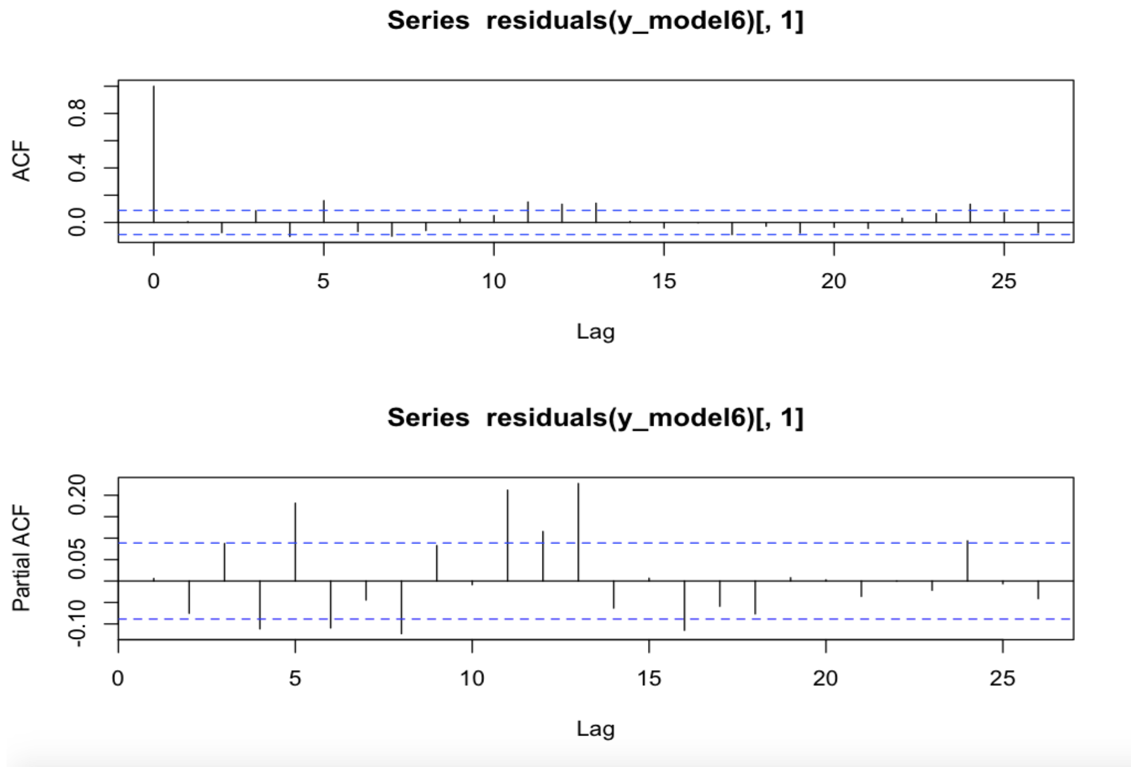
```
y_model2 10  -554.7218
y_model3 14 -1098.2483
y_model4 18 -1163.5161
y_model5 22 -1223.4858
y_model6 26 -1238.7392
y_model7 30 -1229.2035
BIC(y_model1, y_model2, y_model3, y_model4, y_model5, y_model6, y_model7)
         df        BIC
y_model1  6  2249.3414
y_model2 10  -512.7167
y_model3 14 -1039.4696
y_model4 18 -1087.9801
y_model5 22 -1131.2088
y_model6 26 -1129.7377
y_model7 30 -1103.4940
```

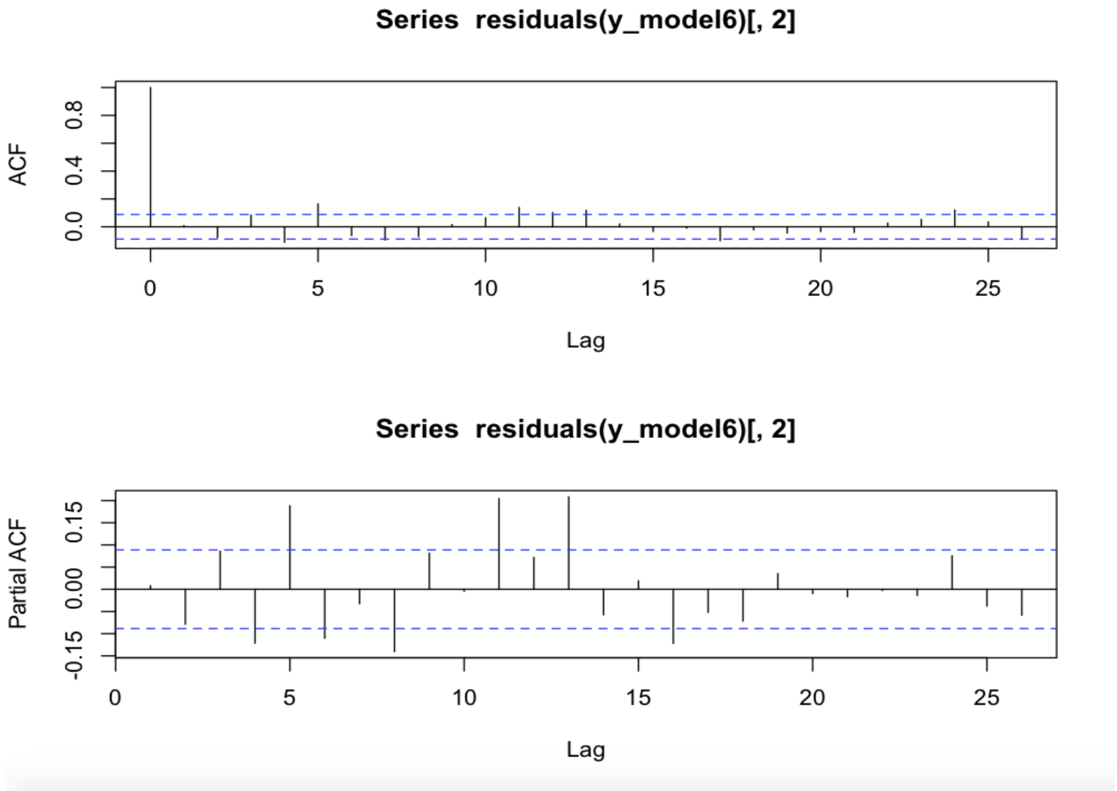
Discussion:

We may choose y_model6, which is VAR(6).

```
summary(y_model6)
VAR Estimation Results:
=========================
Endogenous variables: lla, rso
Deterministic variables: const
Sample size: 489
Log Likelihood: 645.37
Roots of the characteristic polynomial:
1.003 0.9793 0.9793 0.7817 0.7817 0.7344 0.7344 0.7251 0.7251 0.5455 0.5455 0.2842
Call:
VAR(y = y_tot, p = 6)

Estimation results for equation lla:
====================================
lla = lla.l1 + rso.l1 + lla.l2 + rso.l2 + lla.l3 + rso.l3 + lla.l4 + rso.l4 + lla.l5 +
rso.l5 + lla.l6 + rso.l6 + const

        Estimate Std. Error t value Pr(>|t|)
lla.l1  2.650564   0.274520   9.655  < 2e-16 ***
rso.l1  0.158446   0.291204   0.544  0.58662
lla.l2 -3.387692   0.831632  -4.074 5.42e-05 ***
rso.l2  0.094859   0.890626   0.107  0.91522
lla.l3  3.486131   1.196862   2.913  0.00375 **
rso.l3 -0.969832   1.284005  -0.755  0.45043
lla.l4 -2.851589   1.198931  -2.378  0.01778 *
rso.l4  1.149502   1.284844   0.895  0.37142
lla.l5  1.348422   0.827291   1.630  0.10378
rso.l5 -0.439463   0.885896  -0.496  0.62008
lla.l6 -0.244206   0.266663  -0.916  0.36024
rso.l6  0.004756   0.284429   0.017  0.98666
const   0.051944   0.034409   1.510  0.13181
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3194 on 476 degrees of freedom
Multiple R-Squared:     1, Adjusted R-squared:     1
F-statistic: 1.707e+06 on 12 and 476 DF, p-value: < 2.2e-16
Estimation results for equation rso:
====================================
rso = lla.l1 + rso.l1 + lla.l2 + rso.l2 + lla.l3 + rso.l3 + lla.l4 + rso.l4 + lla.l5 +
rso.l5 + lla.l6 + rso.l6 + const

       Estimate Std. Error t value Pr(>|t|)
lla.l1 -0.208582   0.258621  -0.807 0.420349
rso.l1  3.048178   0.274339  11.111  < 2e-16 ***
lla.l2 -0.109461   0.783467  -0.140 0.888946
rso.l2 -3.226597   0.839045  -3.846 0.000137 ***
lla.l3  1.263057   1.127545   1.120 0.263200
rso.l3  1.252725   1.209641   1.036 0.300907
lla.l4 -1.667579   1.129494  -1.476 0.140499
rso.l4 -0.003843   1.210432  -0.003 0.997468
lla.l5  0.902212   0.779378   1.158 0.247606
rso.l5 -0.018563   0.834589  -0.022 0.982265
lla.l6 -0.177636   0.251219  -0.707 0.479851
rso.l6 -0.054260   0.267956  -0.202 0.839614
const   0.064981   0.032416   2.005 0.045574 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3009 on 476 degrees of freedom
Multiple R-Squared:     1, Adjusted R-squared:     1
F-statistic: 1.429e+06 on 12 and 476 DF, p-value: < 2.2e-16

Covariance matrix of residuals:              Correlation matrix of residuals:
        lla     rso                                  lla    rso
lla 0.10204 0.09477                          lla 1.0000 0.9859
rso 0.09477 0.09056                          rso 0.9859 1.0000
```

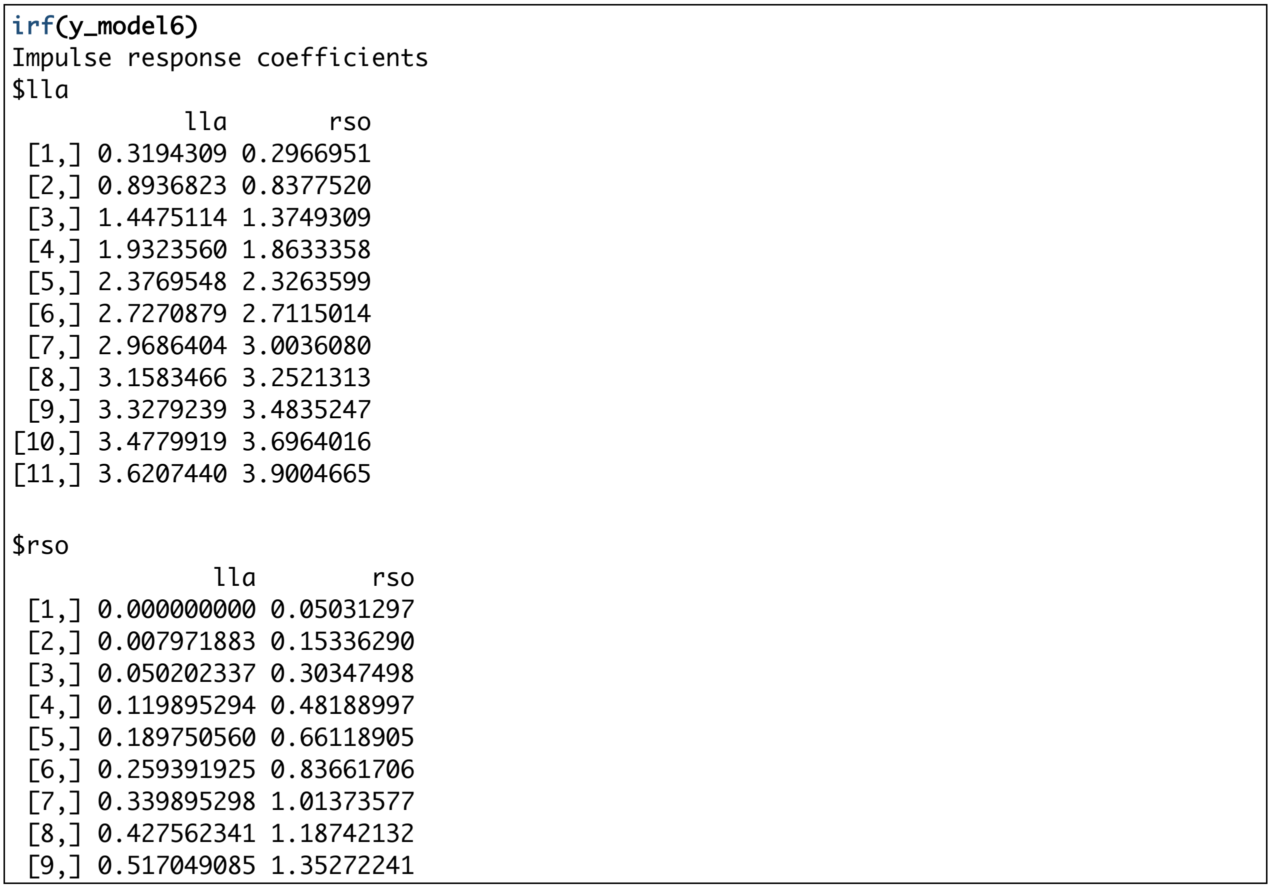

Series residuals(y_model6)[, 1]
ACF
0.0
0.4
0.8
0
5
10
15
20
25
Lag
Series residuals(y_model6)[, 1]
Partial ACF
-0.10
0.05
0.20
0
5
10
15
20
25
Lag


Series residuals(y_model6)[, 2]
ACF
0.0
0.4
0.8
0
5
10
15
20
25
Lag
Series residuals(y_model6)[, 2]
Partial ACF
-0.15
0.00
0.15
0
5
10
15
20
25
Lag

Discussion:

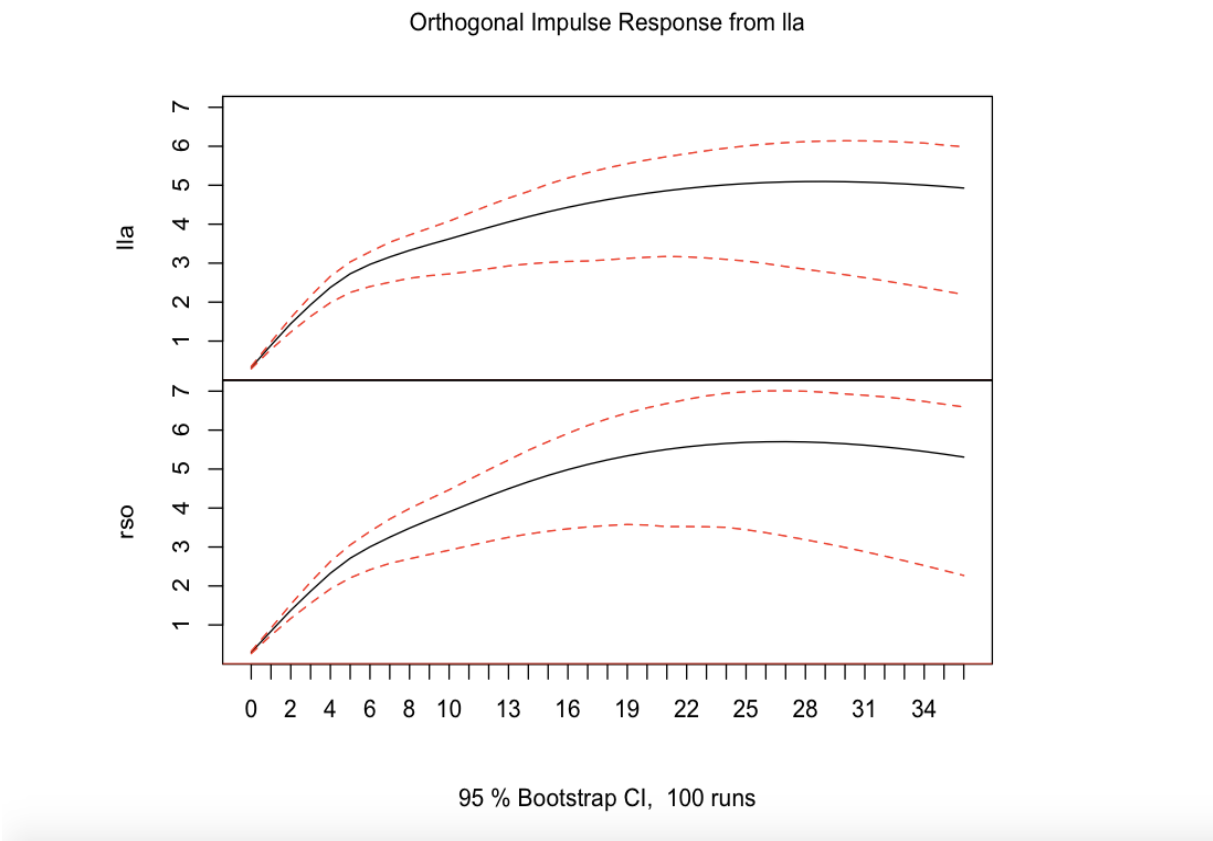
From the first part of this VAR(6) model we pick, we can see the LLA housing price is overall positively depending on its odd-period lag information, but negatively depends on its even-period lag information. The information provided by RSO seems require us to look more previous data, rso.l3=-0.97, rso.l4=1.15, rso.l5 = -0.44, the other estimates are really close to zero. **However,** looking at the p-value carefully, all the p-values are quite large, even close to zero! That implies that the data all these rso.lag estimates could be ignored. That is, the RSO does not cause LLA.

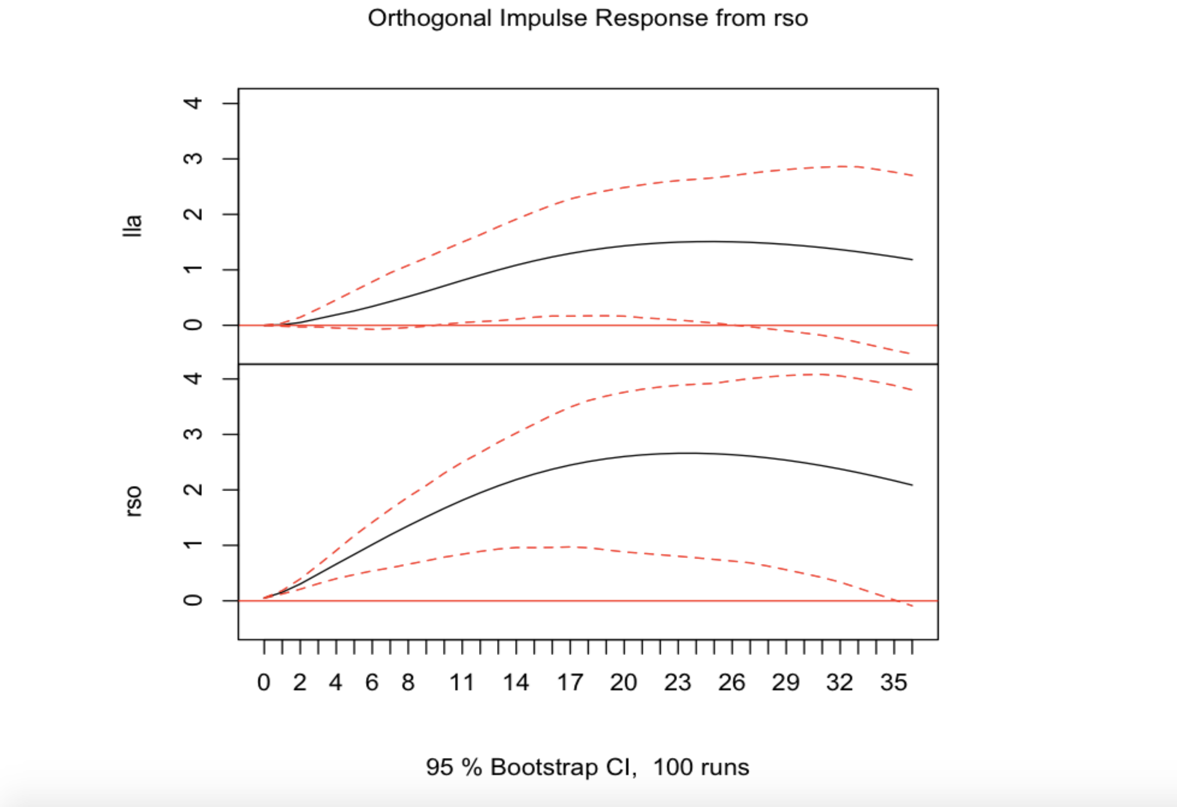
From the first part of this VAR(6) model we pick, we can see the RSO housing price is overall positively depending on its odd-period lag information, but negatively depends on its even-period lag information as LLA. lla.l3=1.26, lla.l4=-1.67, lla.l5 = 0.9, which seems place somehow impact on RSO, but the p-values are large as well.



```
irf(y_model6)
Impulse response coefficients
$lla
            lla       rso
 [1,] 0.3194309 0.2966951
 [2,] 0.8936823 0.8377520
 [3,] 1.4475114 1.3749309
 [4,] 1.9323560 1.8633358
 [5,] 2.3769548 2.3263599
 [6,] 2.7270879 2.7115014
 [7,] 2.9686404 3.0036080
 [8,] 3.1583466 3.2521313
 [9,] 3.3279239 3.4835247
[10,] 3.4779919 3.6964016
[11,] 3.6207440 3.9004665

$rso
              lla        rso
 [1,] 0.000000000 0.05031297
 [2,] 0.007971883 0.15336290
 [3,] 0.050202337 0.30347498
 [4,] 0.119895294 0.48188997
 [5,] 0.189750560 0.66118905
 [6,] 0.259391925 0.83661706
 [7,] 0.339895298 1.01373577
 [8,] 0.427562341 1.18742132
 [9,] 0.517049085 1.35272241
```

```
[10,] 0.611081931 1.51230479
[11,] 0.709948370 1.66679095

Lower Band, CI= 0.95
$lla
            lla       rso
 [1,] 0.2826397 0.2601058
 [2,] 0.7741213 0.7210469
 [3,] 1.2310802 1.1664274
 [4,] 1.6110762 1.5583904
 [5,] 1.9643183 1.9236653
 [6,] 2.2355231 2.2087752
 [7,] 2.4054499 2.4128408
 [8,] 2.5136048 2.5659946
 [9,] 2.5865325 2.6874804
[10,] 2.6340386 2.7807818
[11,] 2.6857015 2.8651296

$rso
              lla        rso
 [1,]  0.00000000 0.04531444
 [2,] -0.01716053 0.12200646
 [3,] -0.02644697 0.21375625
 [4,] -0.04403386 0.31143910
 [5,] -0.06276078 0.40808843
 [6,] -0.09389086 0.46855007
 [7,] -0.11225191 0.54952397
 [8,] -0.10378272 0.65748446
 [9,] -0.08514330 0.75663252
[10,] -0.05212396 0.80267661
[11,]  0.01218829 0.84913080

Upper Band, CI= 0.95
$lla
            lla       rso
 [1,] 0.3403911 0.3214197
 [2,] 0.9661119 0.9150376
 [3,] 1.5841211 1.5161037
 [4,] 2.1262443 2.0666388
 [5,] 2.6120414 2.5890614
 [6,] 3.0162184 3.0135409
 [7,] 3.3116829 3.3247745
 [8,] 3.5274346 3.6008202
 [9,] 3.7051412 3.8730027
[10,] 3.8959349 4.1590528
[11,] 4.0871198 4.4041246

$rso
             lla        rso
 [1,] 0.00000000 0.05361553
 [2,] 0.03133648 0.17883049
 [3,] 0.12360801 0.37421837
 [4,] 0.26434312 0.62499373
 [5,] 0.41210684 0.87179853
 [6,] 0.55589906 1.10805981
 [7,] 0.69484023 1.34876882
 [8,] 0.83254795 1.58747495
 [9,] 0.97458394 1.80901155
[10,] 1.11497989 2.04453818
[11,] 1.27458577 2.24431428
```





Discussion:

First plot: shocks in lla lead to little fluctuation in lla and rso at first. But its effect on lla and rso increases as time goes.

Second plot: shocks in rso lead to little fluctuation in lla and rso at first. But when lag is larger, its effect manifests.

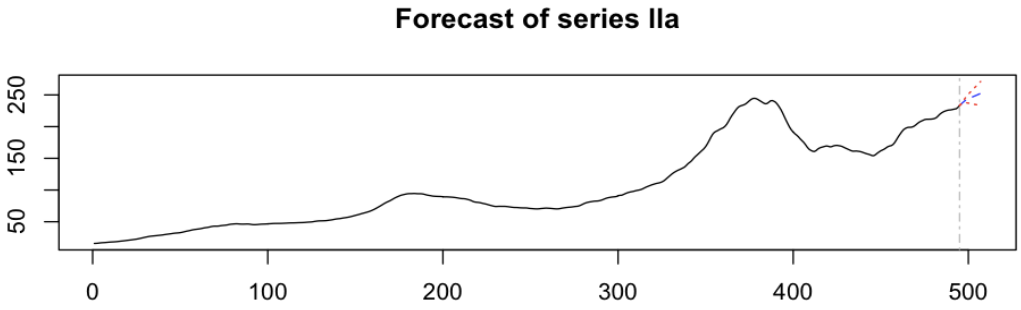
```
grangertest(lla ~ rso, order = 6)
Granger causality test
Model 1: lla ~ Lags(lla, 1:6) + Lags(rso, 1:6)
Model 2: lla ~ Lags(lla, 1:6)
 Res.Df Df      F Pr(>F)
1    476
2    482 -6 2.3361 0.0311 *
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

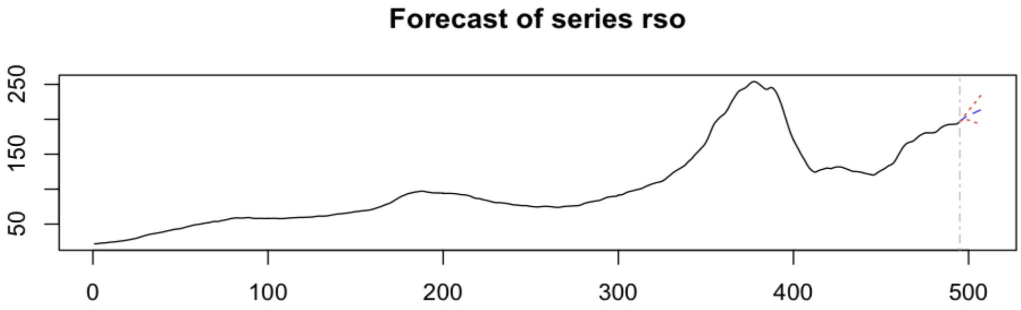
```
grangertest(rso ~ lla, order = 6)
Granger causality test
Model 1: rso ~ Lags(rso, 1:6) + Lags(lla, 1:6)
Model 2: rso ~ Lags(rso, 1:6)
Res.Df Df      F   Pr(>F)
1    476
2    482 -6 3.1199 0.005227 **
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion:

In the first test, we compare two time series on LLA. We found the statistic statistically significant. P-value = 0.03, so we fail to reject the null hypothesis, that means LLA is not affected by previous information provided by RSO within 99% confidence.

In the second test, we compare two time series on RSO. The result tells us that we fail to reject the null hypothesis at 99.9%, which means RSO is not affected by previous values of LLA within 99.9% confidence. But if we keep our confidence level at 99%, we would reject the null hypothesis, which means that the RSO is influenced by LLA at 99% level.





Discussion:

By the VAR(6) model, the forecast for future data will increase more aggressively than what we forecast using our ARMA model. And both forecasts for the next 12-step ahead are telling us that the housing in both LLA and RSO areas have an increasing trend.

III. Conclusions and Future Work

Housing price is one of the most important index to measure economy. Thus, it will be very useful for the investors if the housing price could be modeled and forecasted. After analyzing the the housing data from two districts, Riverside-San Bernardino-Ontario and Los Angeles-Long Beach-Anaheim, we observed that the two time-series are related to each other. First of all, we are able to model the two time-series individually by using ARIMA, and the model traces the data fairly well. Moreover, we realized that the two time serieses affect each other. Therefore, we are able to forecast one time-series by itself and the other one. According to the granger test, it is likely that the housing price in Los Angeles-Long Beach-Anaheim causes the housing price in Riverside-San Bernardino-Ontario to change. Although that we built the models for these two time serieses, and it seems like we are able to forecast the data using our models, we believe there is a room for improvement. According to the residual plots, there are still some structures left. Thus, we might be able to find a better model and reduce the residuals to white noise.

We learned a lot from this project. We obtained experience on dealing with real data. In the future, we will spend more time studying time series forecast since it is such a practical and powerful skill.

References

- House Price Index Archive: MSA(1975-Current)
  (Use Monthly Riverside-San Bernardino-Ontario, Los Angeles-Long Beach-Anaheim Housing Price Index)
  http://www.freddiemac.com/finance/fmhpi/archive.html