A “Roof” on Prices: Forecasting the Prices of Houses in San Diego

•••

Tanisha Dighe
Math 111A
Prof. David Meyer

Time series analysis

- Way of analyzing a sequence of data points collected over an interval of time.
- Understands the underlying causes of trends or systemic patterns over time.
- SARIMA - Seasonal Autoregressive Integrated Moving Average
- SARIMAX - Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors

Understanding the Model

Differencing (*d*)

- *d* is the number of differencing required to make the time series stationary
  - I.e the data needs to not have trends or seasonality
- Done by subtracting the previous observation from the current observation

AR model (*p*)

- In an AR (Auto-regressive) model the model predicts the next data point by looking at previous data points and using a mathematical formula similar to linear regression
- The order *p* determines how many previous data points will be used.

MA model (*q*)

- An MA (Moving Average) model performs calculations based on noise in the data along with the data's slope
- The order *q* determines how many previous data points will be used.

SARIMA: (SARIMA(p,d,q)(P,D,Q,m))

$$\Phi_p(L)\phi_P(L^s)\Delta^d\Delta_s^D y_t = \Theta_q(L)\theta_Q(L^s)\epsilon_t$$

$\Phi_p(L)$ = Non-seasonal autoregressive lag polynomial
$\phi_P(L^s)$ = Seasonal autoregressive lag polynomial
$\Delta^d\Delta_s^D y_t$= times series, differenced d times and seasonally differenced D times
$\Theta_q(L)$ = non-seasonal moving average lag polynomial
$\theta_Q(L^s)$ = seasonal moving average lag polynomial
ϵ_t = the noise at time t

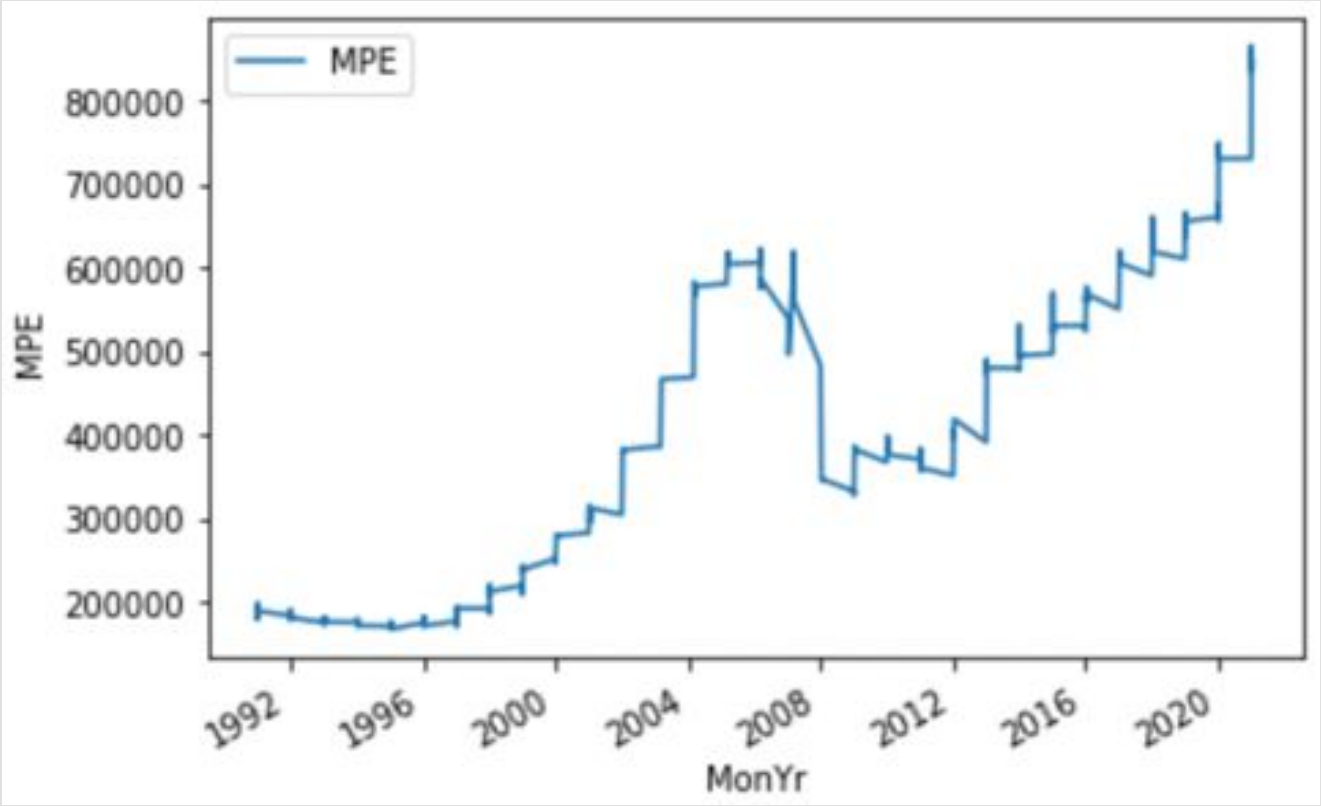
SARIMAX: (SARIMAX(p,d,q)(P,D,Q,m))

$$\Phi_p(L)\phi_P(L^s)\Delta^d\Delta_s^D y_t = \mathrm{A(t)} + \Theta_q(L)\theta_Q(L^s)\epsilon_t$$

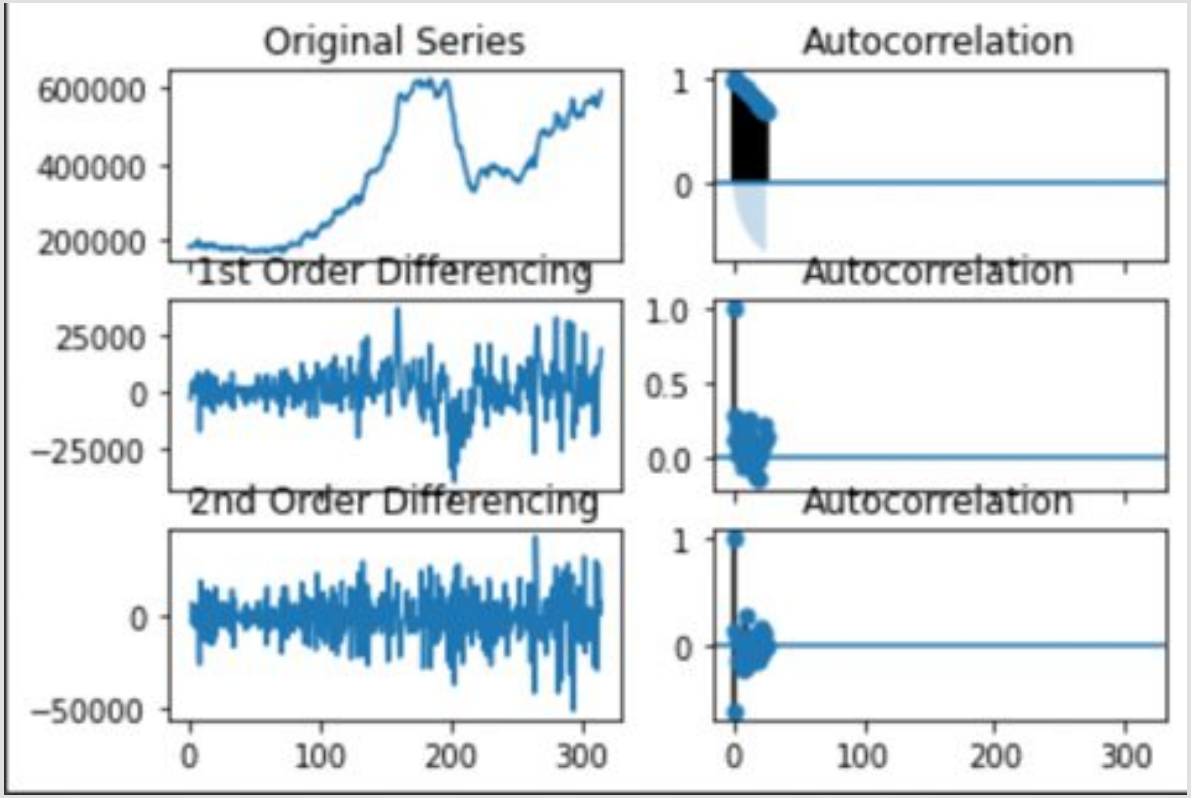
$\Phi_p(L)$ = Non-seasonal autoregressive lag polynomial
$\phi_P(L^s)$ = Seasonal autoregressive lag polynomial
$\Delta^d\Delta_s^D y_t$= times series, differenced d times and seasonally differenced D times
$\Theta_q(L)$ = non-seasonal moving average lag polynomial
$\theta_Q(L^s)$ = seasonal moving average lag polynomial
ϵ_t = the noise at time t
A(t) = the trend polynomial

Excerpt of data:

	A	B	C	D
1	MonYr	MPE	MTM	PCS
2	01/01/90	180484	57	
3	02/01/90	180714	61.8	
4	03/01/90	183701	59.9	
5	04/01/90	181567	58.3	
6	05/01/90	180794	58.3	
7	06/01/90	186733	65.6	
8	07/01/90	185861	67.2	
9	08/01/90	185639	70.6	
10	09/01/90	186272	72.8	
11	10/01/90	179444	66	
12	11/01/90	176980	71.5	
13	12/01/90	181470	80.1	
14	01/01/91	182000	80.9	-0.266
15	02/01/91	178888	88.9	-0.342
16	03/01/91	183111	81.4	-0.344
17	04/01/91	183114	73.7	-0.077
18	05/01/91	188732	74.7	0.022
19	06/01/91	188509	74	0.117

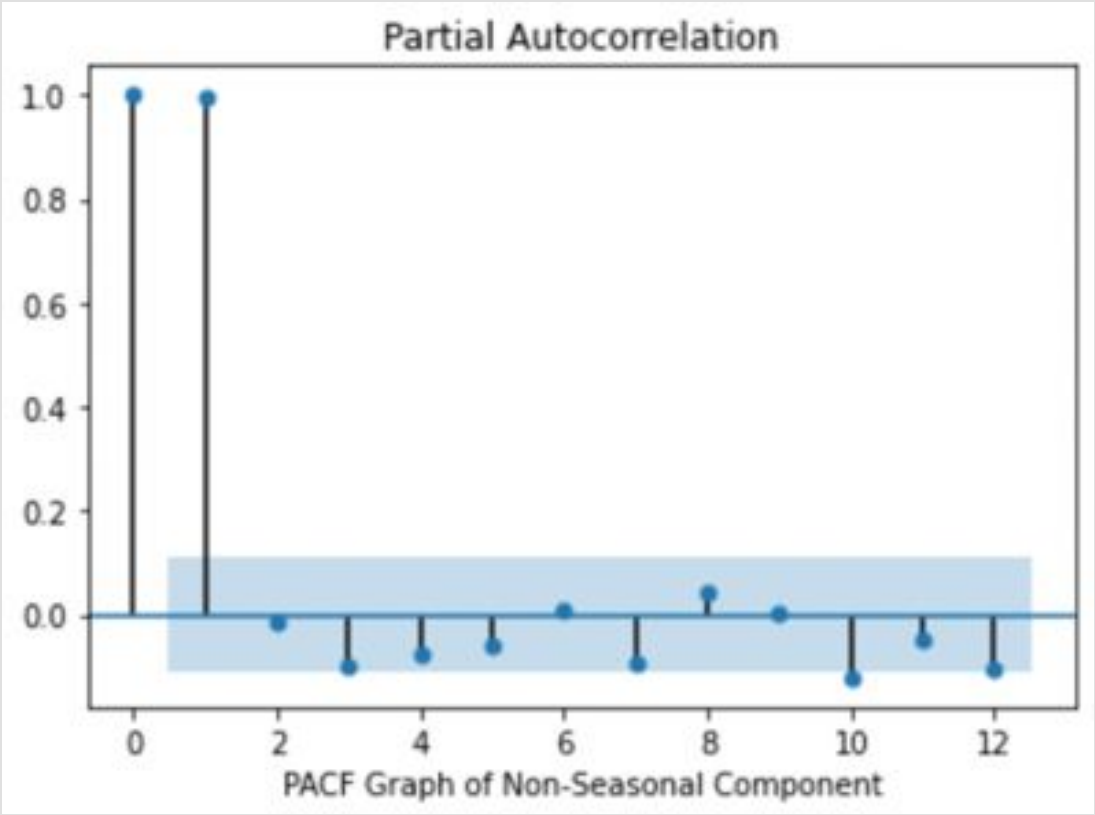

MPE
800000
700000
600000
500000
400000
300000
200000
MPE
1992
1996
2000
2004
2008
2012
2016
2020
MonYr

Finding d:


Original Series
600000
400000
200000
Autocorrelation
1
0
1st Order Differencing
25000
0
−25000
Autocorrelation
1.0
0.5
0.0
2nd Order Differencing
0
−50000
Autocorrelation
1
0
0
100
200
300
0
100
200
300

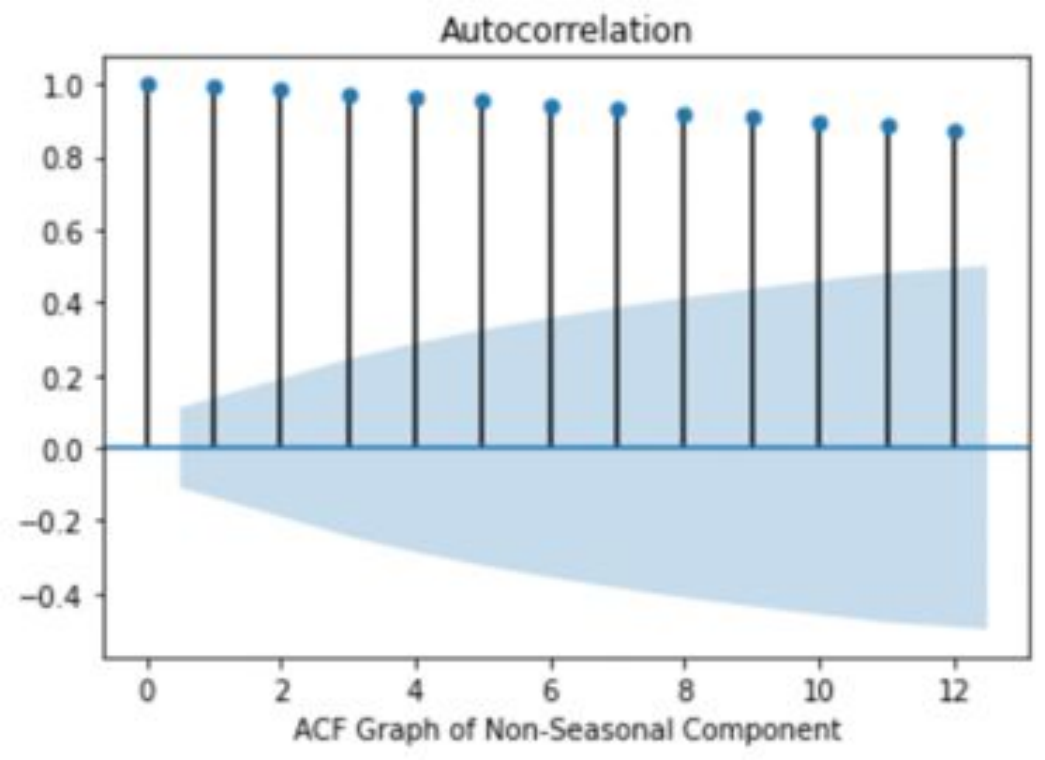
$d = 2$

Finding p:


Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
PACF Graph of Non-Seasonal Component

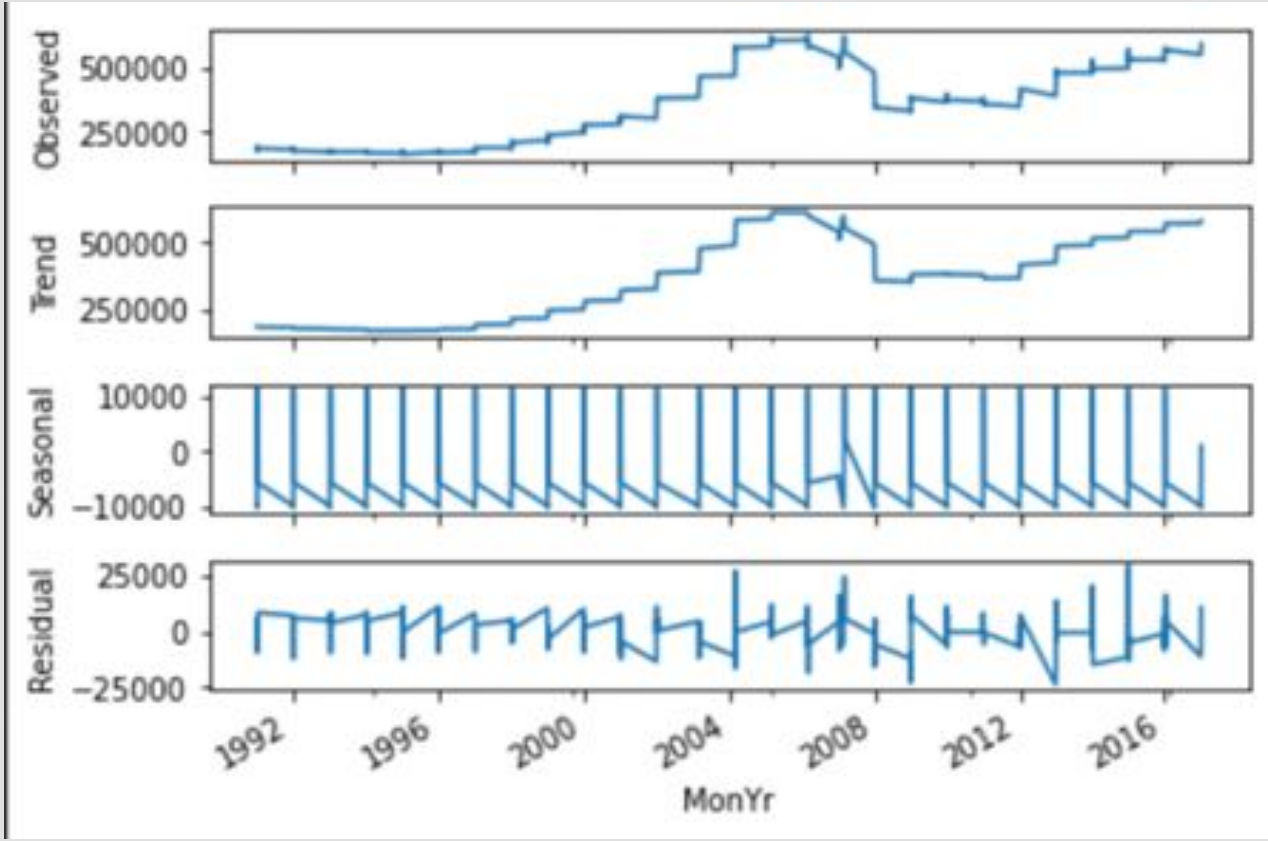
$p = 2$

Finding q:

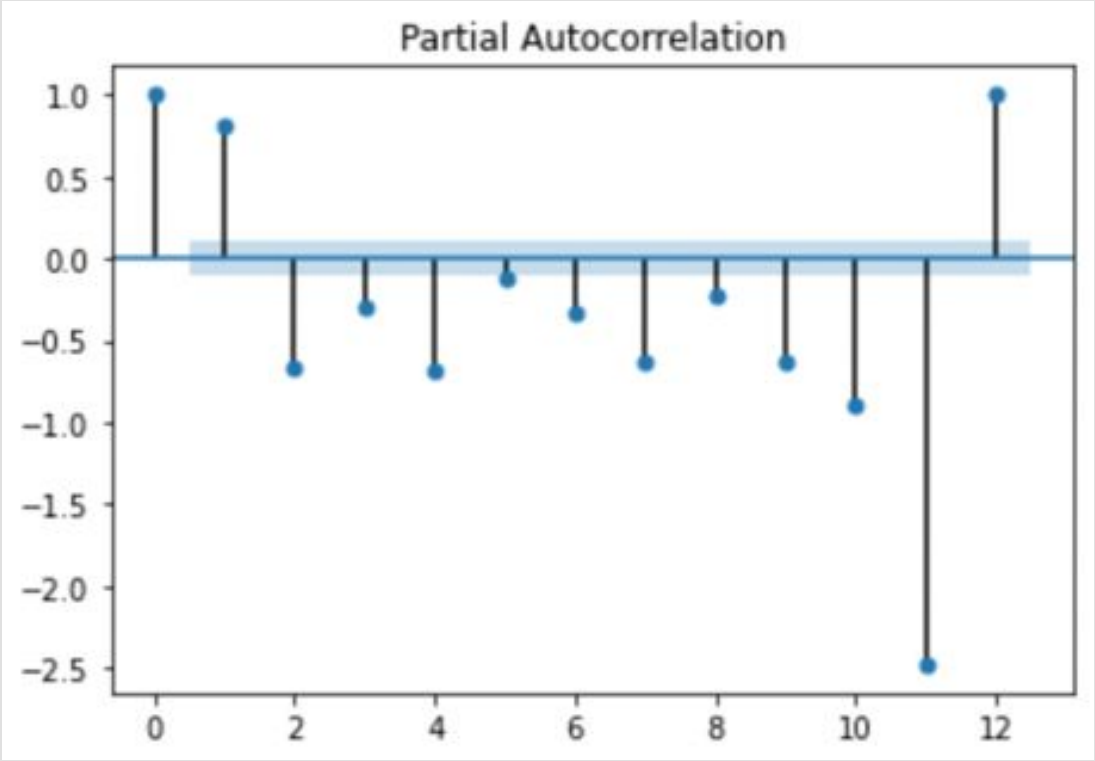


$q = 12$

Seasonal Decomposition of Median Housing Prices

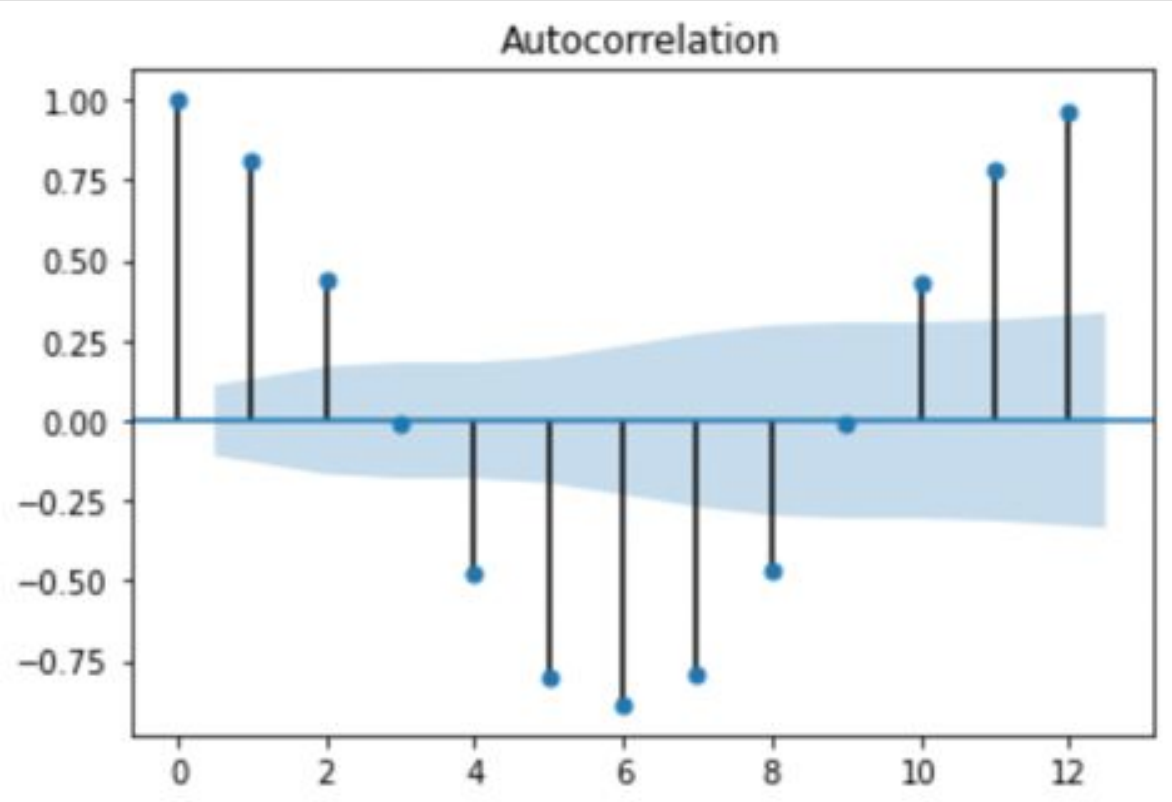


Finding *P*:

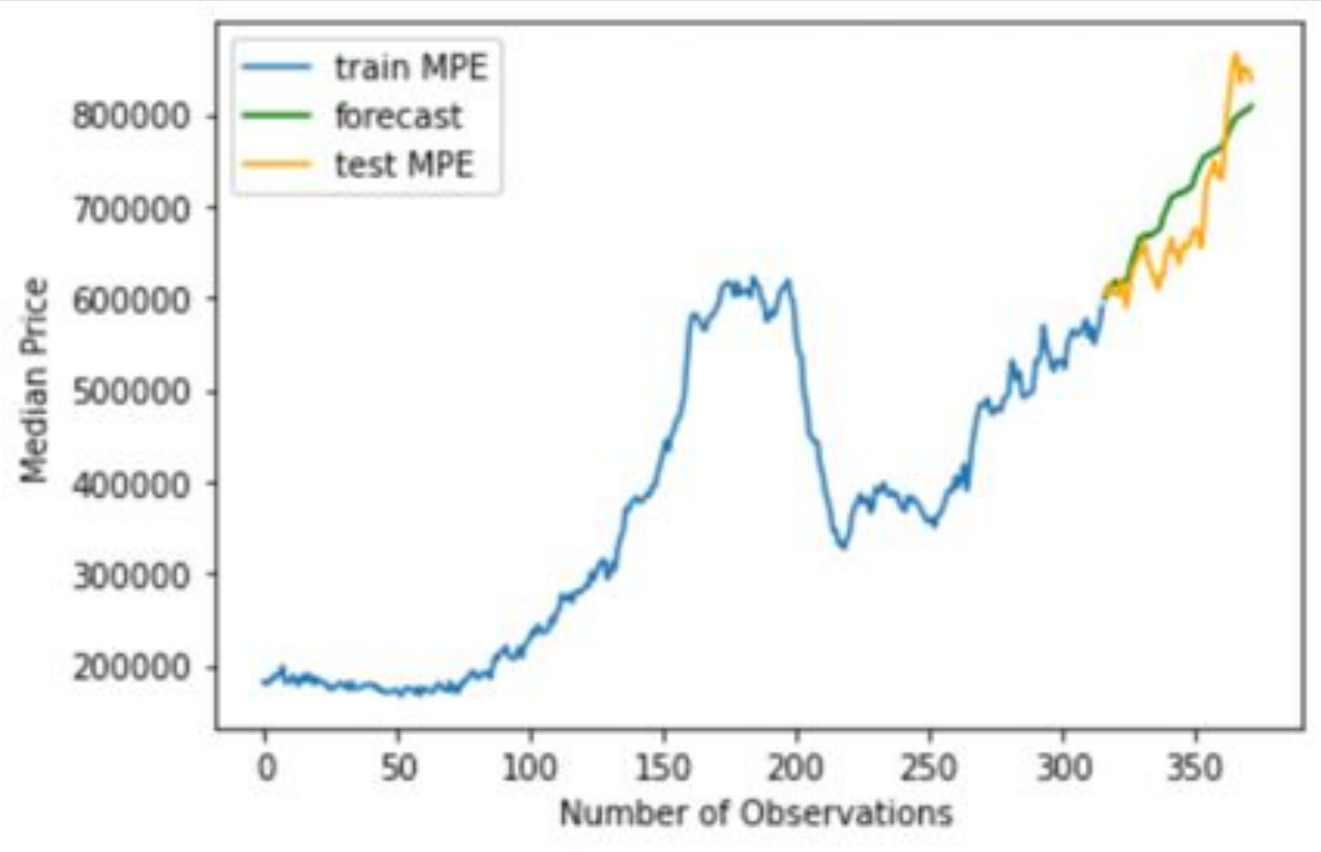


$P = 4$

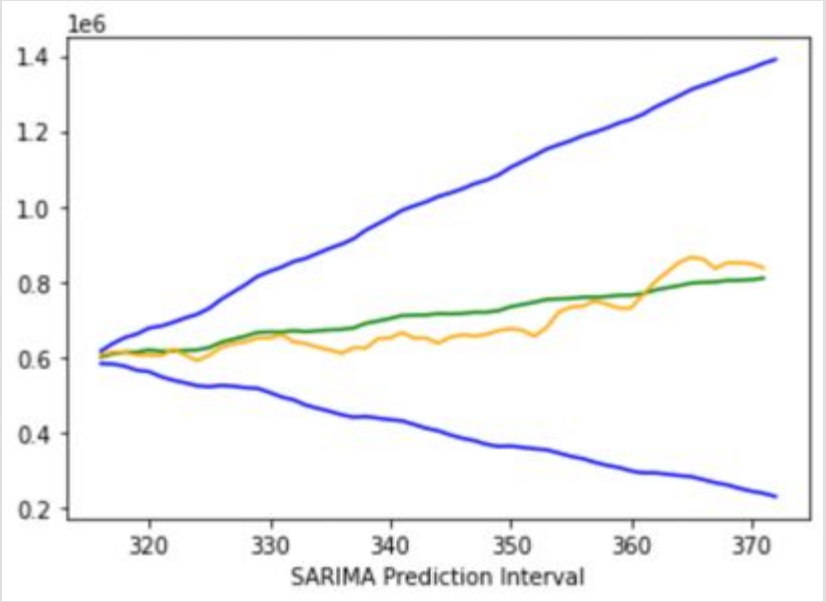
Finding Q:

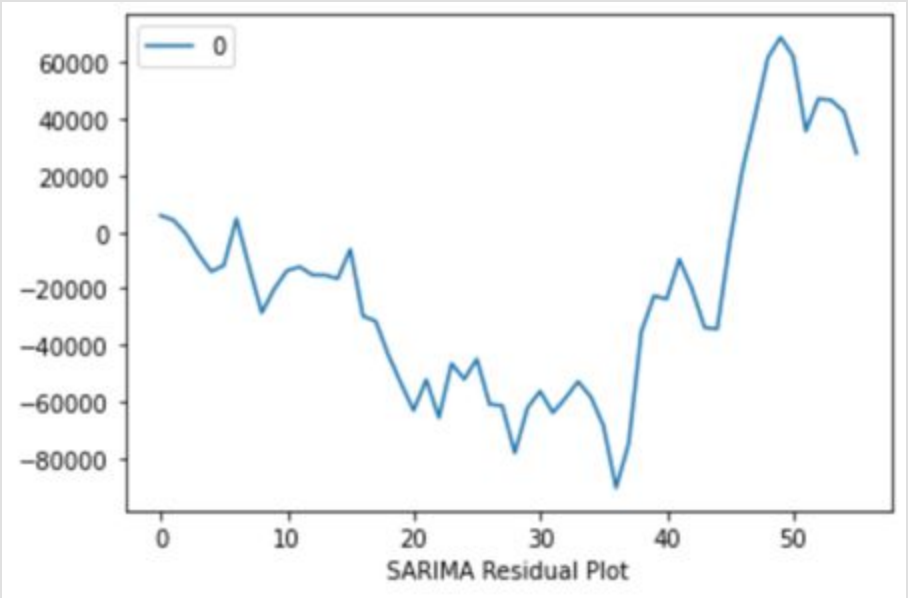


Q = 7

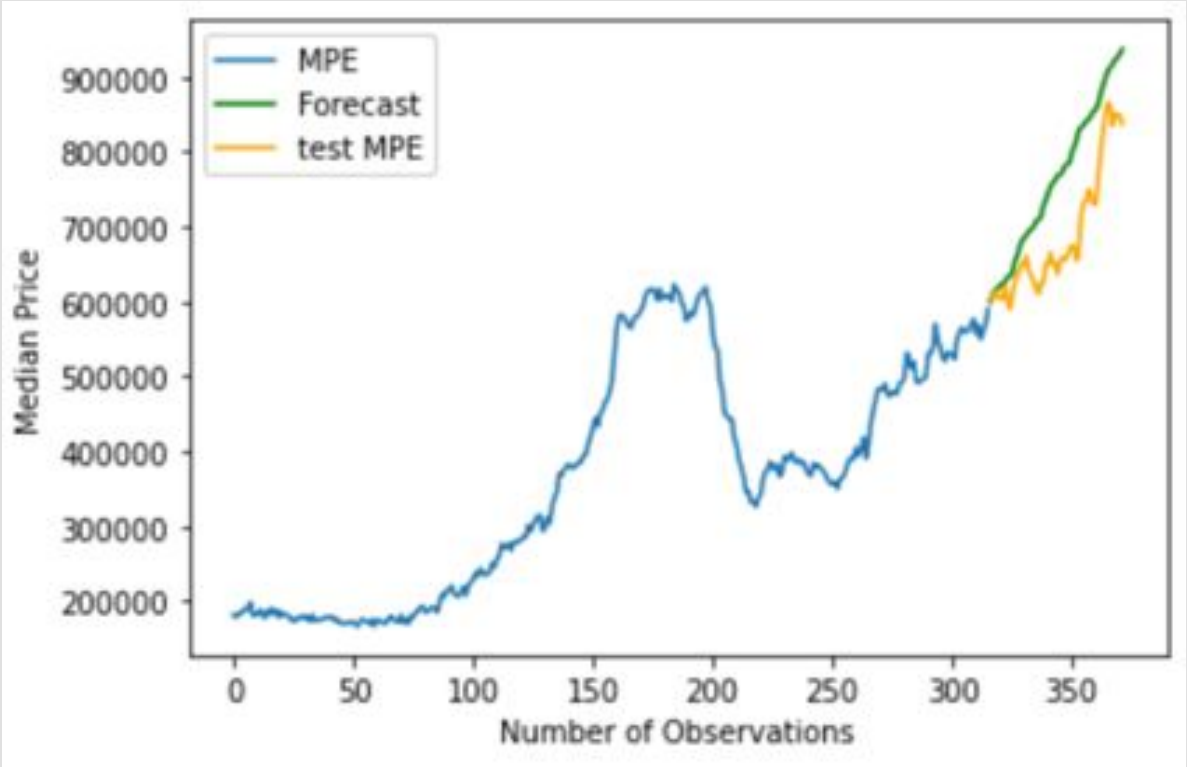


SARIMA(2,2,12)(1,0,2,12)

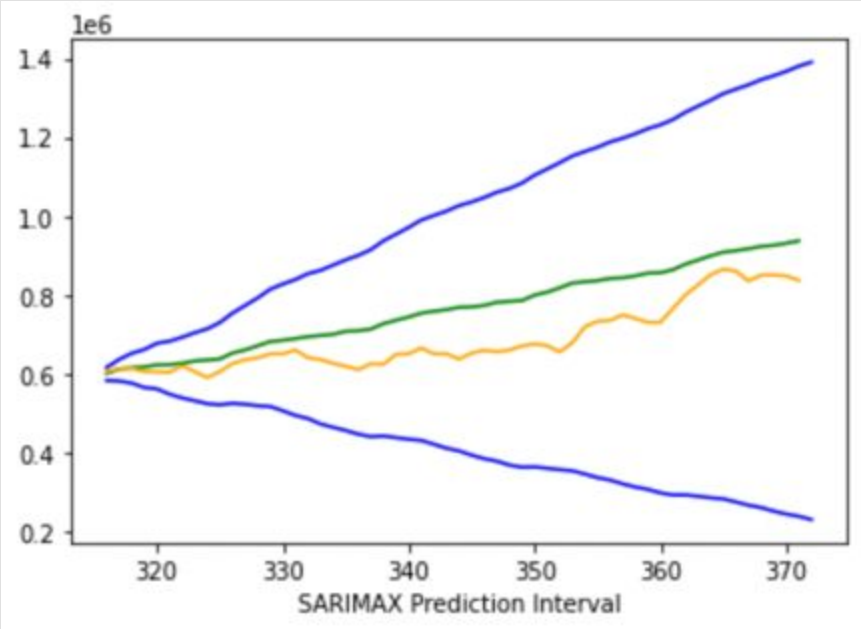

1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
320
330
340
350
360
370
SARIMA Prediction Interval

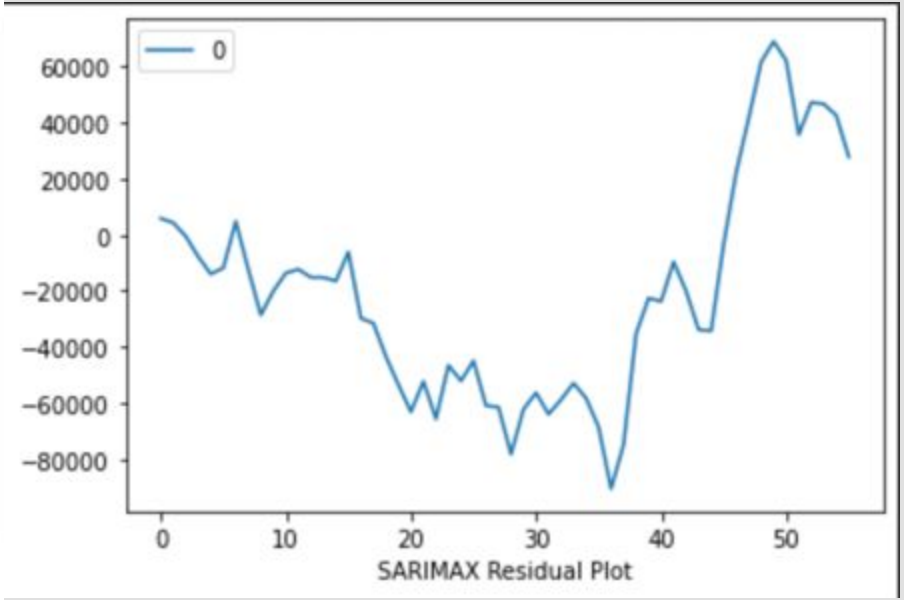

0
60000
40000
20000
0
−20000
−40000
−60000
−80000
0
10
20
30
40
50
SARIMA Residual Plot

RMSE = 0.1599



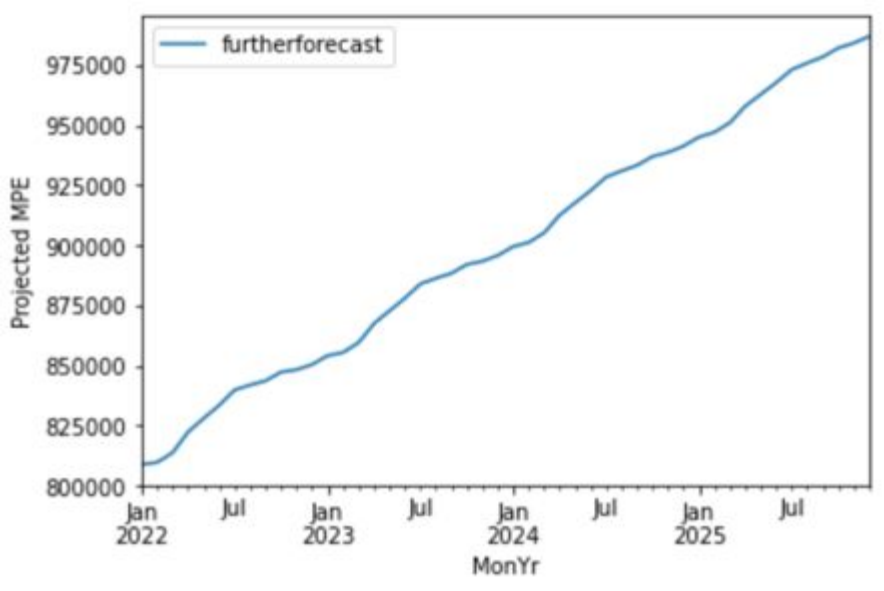
SARIMAX(0,2,1)(2,0,1,12)

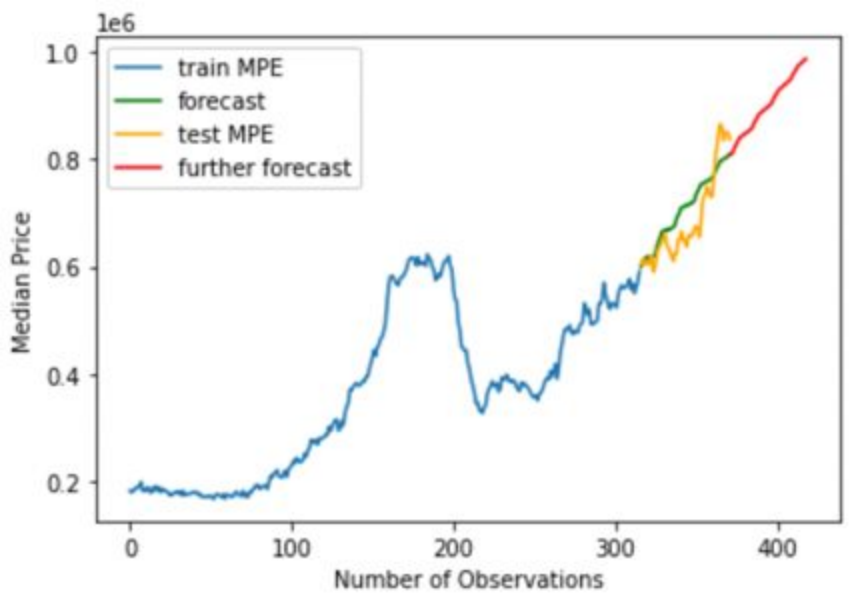

1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
320
330
340
350
360
370
SARIMAX Prediction Interval


0
60000
40000
20000
0
−20000
−40000
−60000
−80000
0
10
20
30
40
50
SARIMAX Residual Plot

RMSE = 0.3140

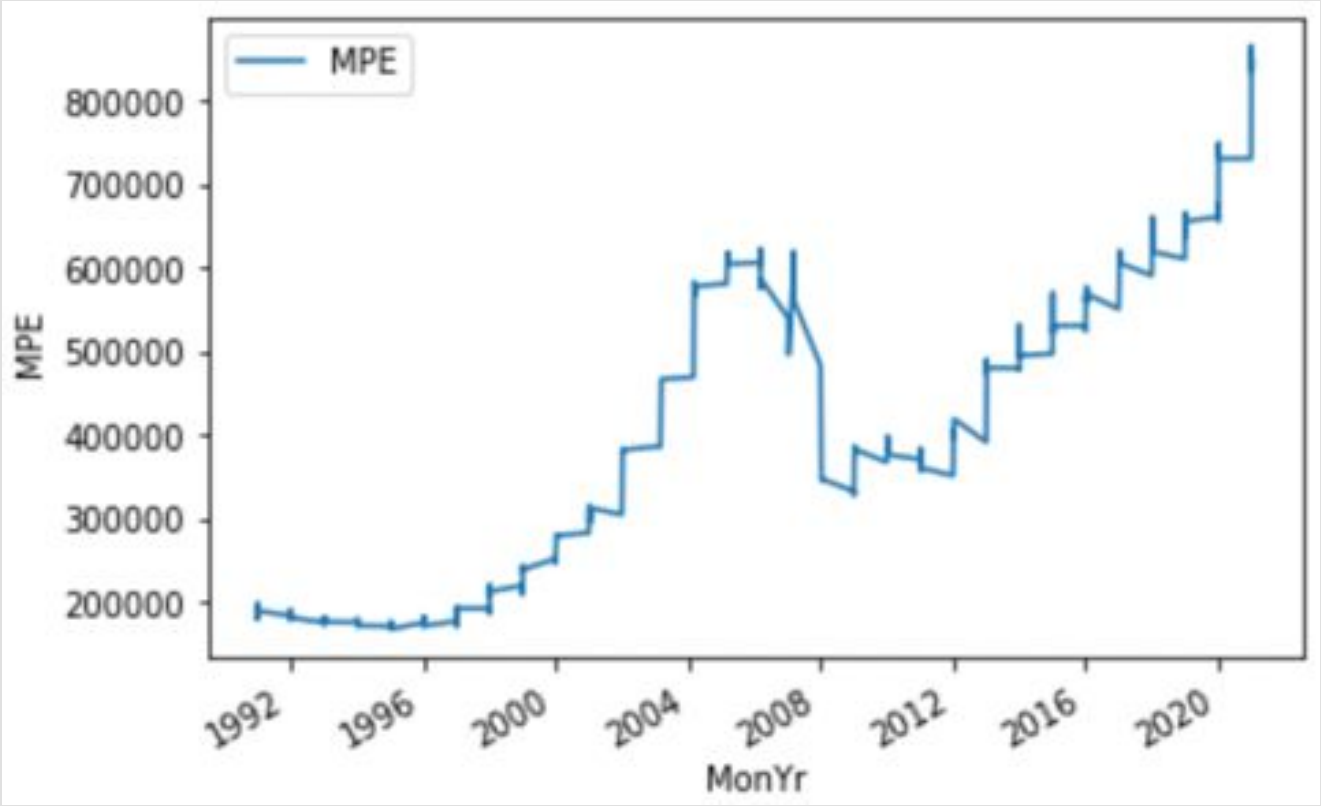
Train Set RMSE of SARIMA	Train Set RMSE of SARIMAX
0.034655	0.031277


furtherforecast
975000
950000
925000
900000
875000
850000
825000
800000
Projected MPE
Jan
2022
Jul
Jan
2023
Jul
Jan
2024
Jul
Jan
2025
Jul
MonYr


1e6
1.0
0.8
0.6
0.4
0.2
train MPE
forecast
test MPE
further forecast
Median Price
0
100
200
300
400
Number of Observations

Evaluating the Model

- Overfitting
  - More training data
  - Removing unnecessary layers/ extraneous variables
  - Weight regularization
- Lack of computing power
- “Acts of God”


MPE
800000
700000
600000
500000
400000
300000
200000
MPE
1992
1996
2000
2004
2008
2012
2016
2020
MonYr

Thank you!

Background

How are houses priced?

- Location
- Size and usable space
- Number of other properties on sale and number of buyers in the market
- The broader economy