Forecasting Housing Price Index of the Puget Sound Region

OMSBA 5305, Fall Quarter 2021, Seattle University

Group 3: Eric Beekman, Jonathan Burns, Vishaal Diwan, Glen Lewis, and Andrew Nalundasan

December 07, 2021

# Forecasting Housing Prices in Puget Sound

In the Puget Sound region, housing price growth has exploded in recent months and house prices in general have increased from 1975 onward. It has become vital to lenders, individuals, and government officials to monitor changes in house prices over time to appropriately plan for home ownership and changes in housing affordability at scale. Given these circumstances, our team decided to center the focus of our project on the following research question:

How did the COVID-19 pandemic (“the pandemic”) impact the behavior of the housing price index and what is the appropriate scheme (fixed, recursive, or rolling) that will help us best forecast the housing price index values into the future, after the pandemic?

# Data Description

We used the Freddie Mac House Price Index (FMPHI) available at <http://www.freddiemac.com/research/indices/house-price-index.page>.

Per the Freddie Mac website “the FMHPI provides a measure of typical price inflation for houses within the United States. Values are calculated monthly and released at the end of the following month. For example, the FMHPI for March is published in late April.” The data includes seasonally and non-seasonally adjusted series which are available at three different geographical levels (metropolitan, state, and national) for each month, dating back to January 1975.

For this project, we utilized our entire monthly dataset, which ranged from January 1975 to September 2021. Figure 1 displays the seasonally adjusted housing index trends of our data as a timeseries. We see a steady increase in the trend line from 1975 up to 2008, which marks the beginning of the housing bubble burst. The trend decreases until about 2012, then the trend dramatically increases up to the present.

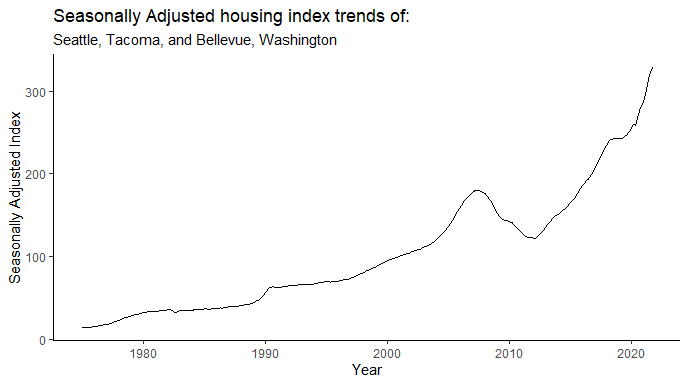


Figure 1: The plot does not display a traditional stationary process profile.

To make our data stationary, we applied the log difference as seen in Figure 2. We witness a similar drop in index between 2008-2012, then see the trend return to the mean. Taking the log difference reveals a spike in 1990 that did not seem significant in Figure 1. We also see dips in the early 80’s and late 10’s that do not appear to be present in the raw data plot. The plot appears to show a stationary profile.

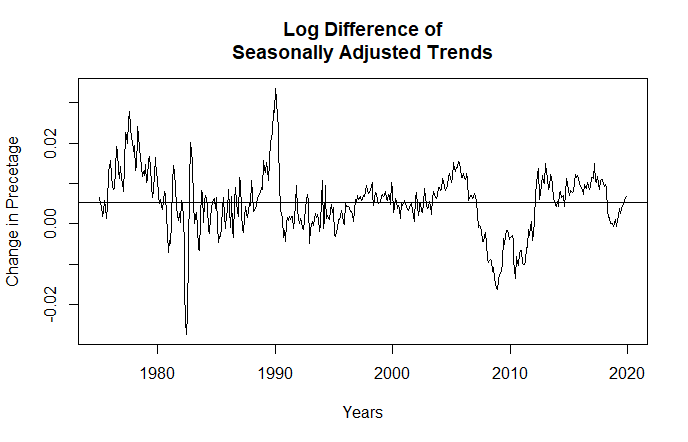


Figure 2: The log diff time series values are normally distributed with a mean of 0.0053 and a standard deviation of 0.0077.

Below are visuals of the Auto-Correlation Function distribution and the Partial-Auto-Correlation Function distribution for our entire dataset.

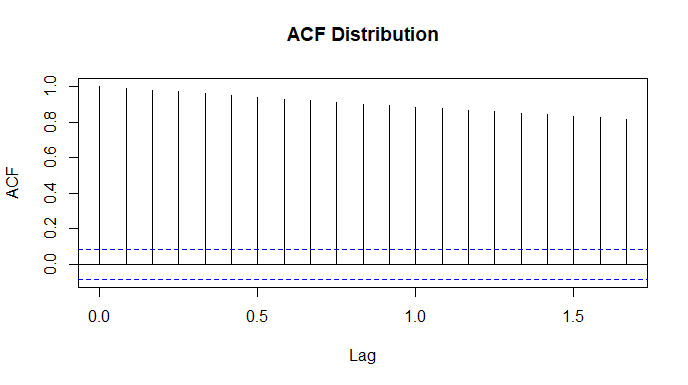


Figure 3: The ACF indications that the process is not stationary.

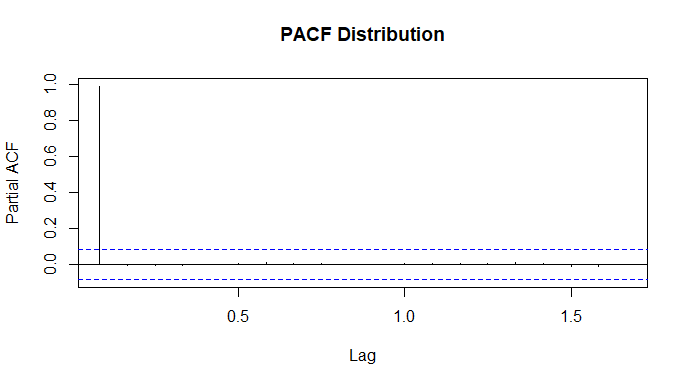


Figure 4: PACF distribution of our entire dataset

An Augmented Dickey-Fuller Test (adf test) was done to verify that our timeseries was indeed stationary. The results of our adf test is found below:

##   
## Augmented Dickey-Fuller Test  
##   
## data: log\_diff\_sa\_ts  
## Dickey-Fuller = -3.2757, Lag order = 8, p-value = 0.07488  
## alternative hypothesis: stationary

Figure #: Augmented Dickey-Fuller test results

From these results, we concluded that the series is stationary. The Dickey-Fuller critical value is -2.87 (~500 observations, 5% without trend, with trend is -2.41). The order 1 diff time series’ stat was -3.2757, well below the critical value and within the null rejection area.

# Selecting a model

Now that we confirmed that we were working with a stationary series after applying the log difference, the ACF and PACF were both taken with the updated timeseries to determine the model that would best forecast our data.

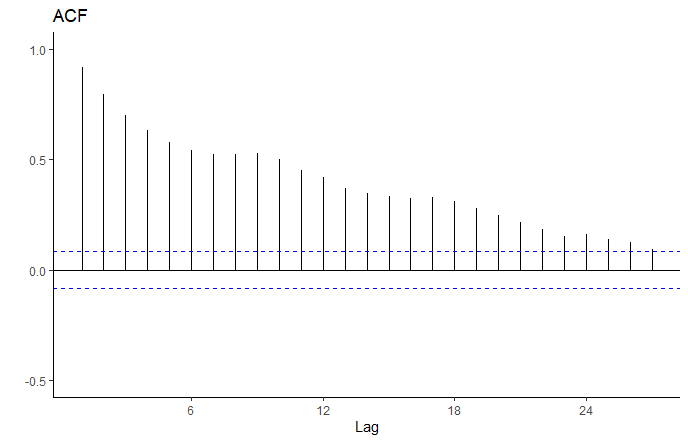


Figure #: The ACF has a geometrically decaying correlation indicative of a AR model.

Because the ACF above has a geometrically decaying collation, it was concluded that the use of an Auto Regressive (AR) model would be most appropriate.

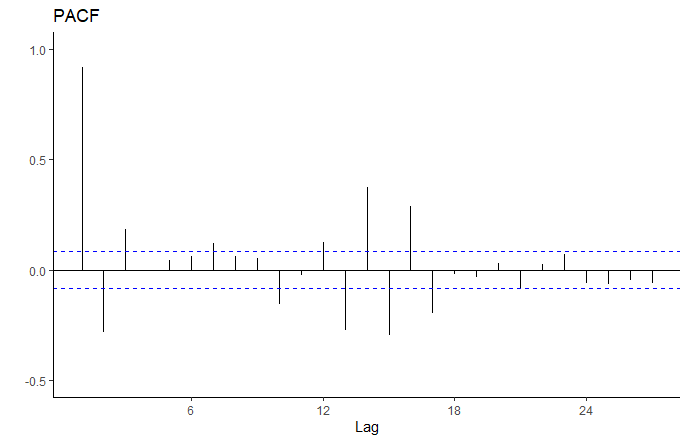


Figure #: The PACF has a significant drop off at lag 3, then another group of significant spikes between lags 12 and 17.

Because the PACF has a significant drop off at lag 3 we decided to use an AR(3) model. We also considered an AR(16) model but assumed that model would most likely be a case of over fitting. The spikes at lags 6 and 9 are not overly significant and were considered random variation.

# Assessing the models

The AR(3) model was explored first. The summary statistics of the AR(3) model can be found in the figure below:

##   
## Time series regression with "ts" data:  
## Start = 1975(5), End = 2019(12)  
##   
## Call:  
## dynlm(formula = log\_diff\_sa\_ts ~ stats::lag(log\_diff\_sa\_ts, -1) +   
## stats::lag(log\_diff\_sa\_ts, -2) + stats::lag(log\_diff\_sa\_ts,   
## -3), start = c(1975, 1), end = c(2019, 12))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0116405 -0.0015019 -0.0001353 0.0013940 0.0114984   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0004668 0.0001559 2.995 0.00288 \*\*   
## stats::lag(log\_diff\_sa\_ts, -1) 1.2287061 0.0426064 28.839 < 2e-16 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -2) -0.5012145 0.0646545 -7.752 4.62e-14 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -3) 0.1850975 0.0425998 4.345 1.67e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.002915 on 532 degrees of freedom  
## Multiple R-squared: 0.8577, Adjusted R-squared: 0.8569   
## F-statistic: 1069 on 3 and 532 DF, p-value: < 2.2e-16

Figure #: Summary statistics of the AR(3) model

The plots of the error residuals of the AR(3) model were analyzed and can be found in the figures below.

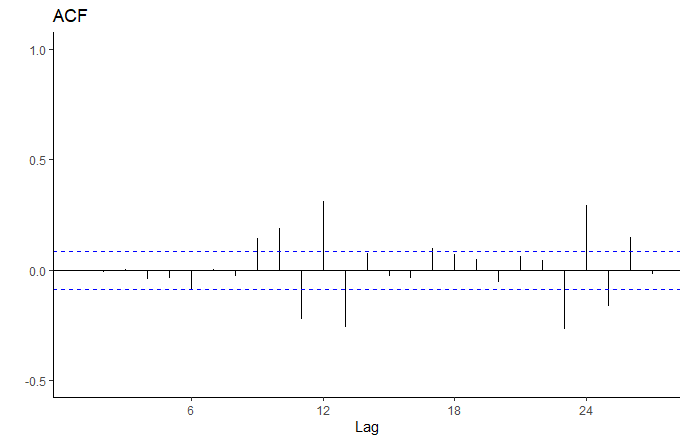


Figure #: ACF AR(3) residuals plot

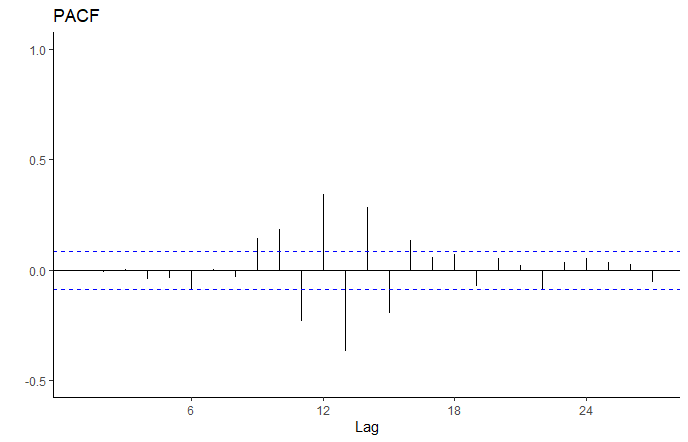


Figure #: PACF (AR3) residuals plot

Both the ACF and PACF have numerous spikes that are of significance especially lag 8 and onward. It was determined that this was not the best model so we decided to run the same analysis for the AR(16) model.

The AR(16) model was explored next. The summary statistics of the AR(16) model can be found in the figure below:

##   
## Time series regression with "ts" data:  
## Start = 1976(6), End = 2019(12)  
##   
## Call:  
## dynlm(formula = log\_diff\_sa\_ts ~ stats::lag(log\_diff\_sa\_ts, -1) +   
## stats::lag(log\_diff\_sa\_ts, -2) + stats::lag(log\_diff\_sa\_ts,   
## -3) + stats::lag(log\_diff\_sa\_ts, -11) + stats::lag(log\_diff\_sa\_ts,   
## -12) + stats::lag(log\_diff\_sa\_ts, -13) + stats::lag(log\_diff\_sa\_ts,   
## -14) + stats::lag(log\_diff\_sa\_ts, -15) + stats::lag(log\_diff\_sa\_ts,   
## -16), start = c(1975, 1), end = c(2019, 12))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.0081927 -0.0013122 -0.0000845 0.0014206 0.0093798   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0002783 0.0001402 1.985 0.0477 \*   
## stats::lag(log\_diff\_sa\_ts, -1) 1.5136058 0.0408550 37.048 < 2e-16 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -2) -0.9966411 0.0649154 -15.353 < 2e-16 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -3) 0.4296467 0.0414985 10.353 < 2e-16 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -11) -0.2799476 0.0378526 -7.396 5.77e-13 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -12) 0.7521892 0.0609539 12.340 < 2e-16 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -13) -1.0248127 0.0688377 -14.887 < 2e-16 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -14) 0.8883466 0.0701656 12.661 < 2e-16 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -15) -0.5660211 0.0639148 -8.856 < 2e-16 \*\*\*  
## stats::lag(log\_diff\_sa\_ts, -16) 0.2289544 0.0376256 6.085 2.28e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.002419 on 513 degrees of freedom  
## Multiple R-squared: 0.9044, Adjusted R-squared: 0.9028   
## F-statistic: 539.4 on 9 and 513 DF, p-value: < 2.2e-16

Figure #: Summary statistics of the AR(16) model

The plots of the error residuals of the AR(16) model were analyzed and can be found in the figures below.

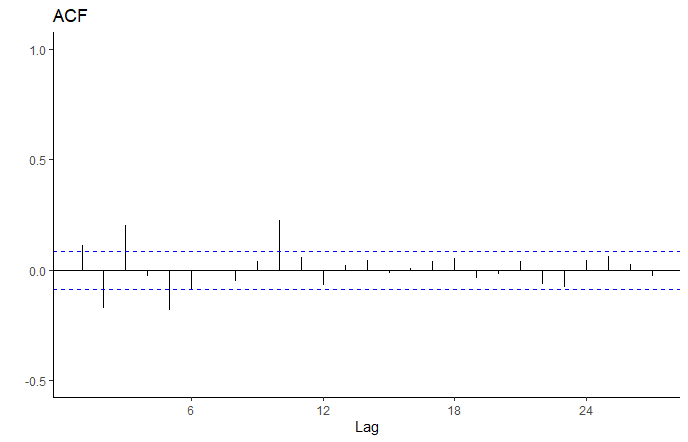


Figure #: ACF AR(16) residuals plot

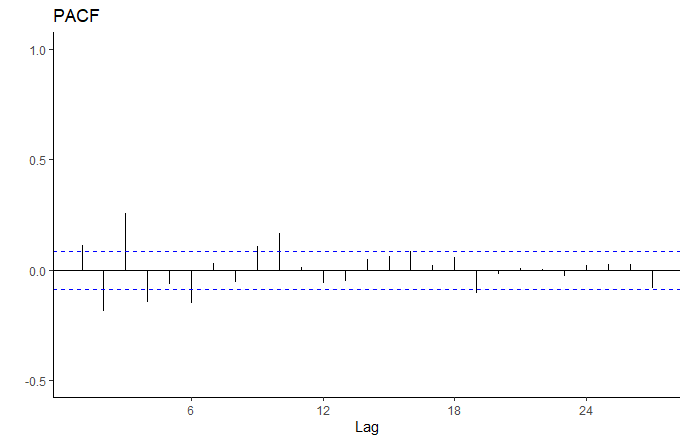


Figure #: PACF AR(16) residuals plot

The AR(16) model returned less lags that surpassed the significance threshold in both the ACF and PACF. Having consistent lags that do not surpass the significance threshold is an indicator that our process has become white noise. The AIC and BIC coefficients were calculated for both models to validate via calculation which model would perform better.



The model with lower AIC and BIC should be chosen, so we chose to move forward with the AR(16) model for the remainder of our analysis.

# Forecasting schemas

We then prepared AR(3) and AR(16) models using fixed, recursive, and rolling schemes to make the predictions. The figure below illustrates all three schemes included in the same plot.

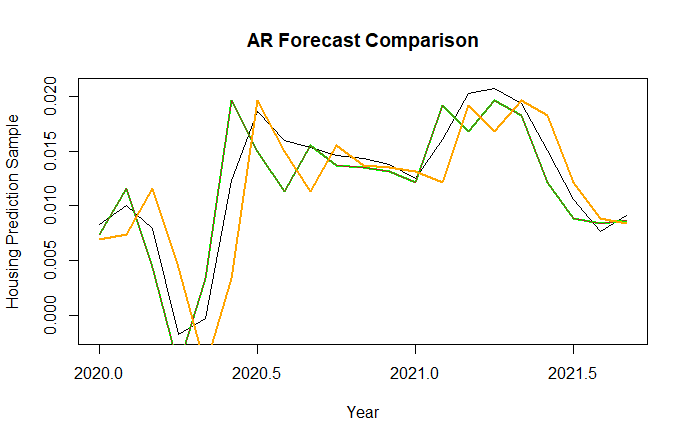


Figure #: AR Forecast comparison between Fixed, Recursive, and Rolling schemes

The AR(3) and AR(16) forecasts are very similar during the first half of 2020. Like all the models, the forecasts for the first half of 2020 lag the actual market behavior by about 30-45 days. We also note that the forecasts become more accurate during the second half of 2020. Based on the plots, after mid-2020, the AR(16) model becomes the better predictor of market behavior.

# Choosing a Loss Function

We think that we have a symmetrical loss function. Errors in predicting the house price index both in over predicting and under predicting do have similar consequences as the main parties impacted are either buyers or sellers and the trade-off in an estimation error in either direction proportionately impact both parties the same. We determined that there is a not a clear loss in either scenario as a gain for one party is a loss for another and vice versa.

# Assessing the forecasts

The traditional convention is to choose a simpler model (ie AR(3) in this scenario) because a simpler model is more likely to reflect the behaviors of a population when you build that model on observations of a sample group. A lower process number allows for randomness to exist in the model - which is realistic.

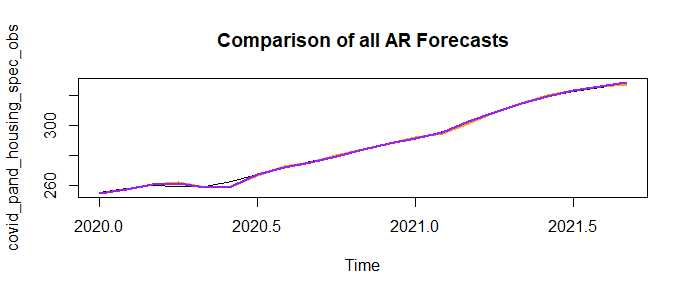


Figure #: Plot showing all AR Forecasts plotted together

Below is a plot of a simple rolling 3 month average of our data.

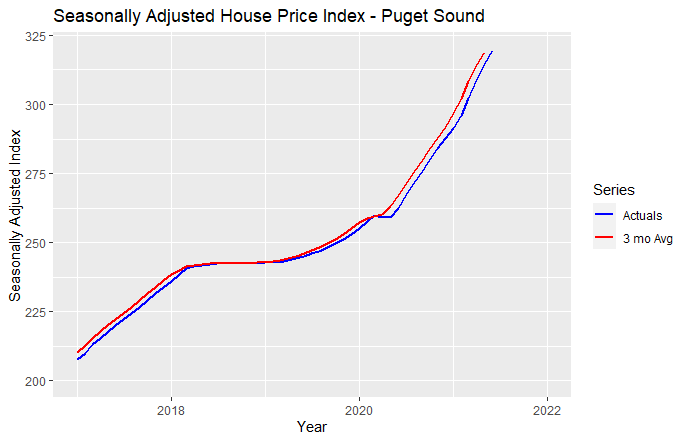


Figure 3: Plot showing 3 month rolling average of our data

# Results and future analysis

For the AR(3) model we completed a recursive, fixed, and rolling scheme. The fixed and rolling schemes are consistent with each other and the recursive model provided a forecast width of 0.1 in the index. We also used the fixed and rolling schemes in the AR(16) module and noted that those schemes also provide similar forecasts.

It appears that the AR(16) model is a better predictor of overall market behavior from mid 2020 and on vice being an over-fitted model is because lags 11 - 16 occurred during a time (2018) when the market index behaved in a fashion similar to the mid 2020- 2021 growth behavior. Because the behavior of lags 11-16 better mirror the 2020 -2021 behavior than the simpler lag 3 model (whose growth during the last 3 months of 2019 was significantly lower than previous lags). Including these additional lags refined the model by including the type of index behavior one could expect during a “hot” market time frame that is **not** captured in the simpler, short term, lag 3 model.