ISYE 6402 Group 45

December 4, 2022

Final Report: Analyzing Homeownership in the United States

Summary

In this study, we analyze the important topic of homeownership in the United States. As a fundamental factor of personal and financial well-being, there are impactful insights that can be gathered from studying homeownership. Homeownership has been a hot topic during the pandemic, with dramatic fluctuations in homeownership rates, rising prices, and increased demand.

Our goal in this study is to investigate three questions of interest around homeownership in the United States. We begin by analyzing characteristics of the univariate US Census dataset of seasonally adjusted homeownership rates from 1980 to 2021. While the raw data comes seasonally adjusted, trend and seasonality are evaluated, both separately and jointly, for their contributions to the predictability of homeownership rates. Ultimately, trend is determined to be most impactful. ARIMA modeling is recommended based on the characteristics of the time series and provide a good fit. The model parameters are fine tuned and different variations of the model are compared. It is ultimately determined that ARIMA modeling provides accurate predictions for homeownership rates, with the best predictions occurring over shorter timeframes. Despite encouraging prediction accuracy, recent homeownership trends during the pandemic are a bit of anomaly that are not anticipated by time series modeling. ARMA-GARCH is explored to predict conditional variance and handle potential heteroskedasticity in homeownership rates in recent years. Unlike ARIMA, ARMA-GARCH predictions are more accurate over a longer timeframe. This may be due to the volatility of homeownership rates in recent years.

There are many factors that are closely related to homeownership, including the health of the economy. As such, a multivariate time series analysis is performed to study homeownership rates alongside GDP, median home sales prices, and interest rates for potential relationships with endogenous factors. While there are significant relationships between these variables, only median home sales prices is found to Granger-cause homeownership rates.

Introduction

Homeownership is a goal for most Americans, with 87% saying that it is a part of the “American Dream[1],” providing means for stability and the accumulation of wealth. According to Federal Reserve data[2], a typical homeowner’s net worth in 2013 was $195,400, while that of a renter was only $5,400. Homeownership contributes to economic growth and job creation in the real estate and development industries. Public policy and legislation often operate with homeownership in mind as a key topic. Yet, in recent years, there have been many challenges to homeownership, including rising prices and mortgage rates, high down payments, changes in mobility patterns and challenges in the labor market.

According to a 2021 study from the Pew Research Center, 49% of Americans say that the availability of affordable housing in their community is a major issue[3]. This marks a 10% increase from early 2018. During the COVID-19 pandemic, the housing market came to a halt in 2020, before homeownership picked back up farther away from cities. A US Census study determined that many renters became homeowners as a result of increased telework opportunities

[1] Yun, Lawrence. “Why Homeownership Matters.” Forbes.com. 2016, Aug 12.

[2] Duca, John V, and Anthony Murphy. *Why House Prices Surged as the COVID-19 Pandemic Took Hold*. Federal Reserve Bank of Dallas, 28 Dec. 2021, https://www.dallasfed.org/research/economics/2021/1228.aspx.

[3] Schaeffer, Katherine. Pew Research Center. “A growing share of Americans say affordable housing is a major problem where they live.” 2022, Jan 18. https://www.pewresearch.org/fact-tank/2022/01/18/a-growing-share-of-americans-say-affordable-housing-is-a-major-problem-where-they-live/

farther away from metropolitan areas afforded by the pandemic[4]. Housing prices rose substantially during the pandemic due to an increased demand for housing but a limited supply. There was a peak increase of 19.3 percent in housing prices year-over-year in July 2021, compared to a moderate 5 percent average year-over-year increase from 2013 to 2020, per the Federal Reserve Bank of Dallas.

To understand the importance of homeownership trends, one needs to look no further than the Great Recession in 2006, where the subprime mortgage crisis saw the creation of a housing pricing bubble. This ultimately led to high foreclosure rates and a prolonged recession with severe economic consequences.

In this paper, we analyze trends in quarterly homeownership rates from 1980-2021. We conduct univariate time series analysis, examining homeownership rates alone over this time period. We also conduct multivariate time series analysis, examining homeownership rates in conjunction with three external time series of interest in GDP, interest rates and median home sales price.

In this study, the following research questions and initial hypotheses will be explored:

- Are there specific characteristics of the time series representing homeownership over time that contribute most to the predictability of the time series?
    - It is hypothesized that trend will most contribute to the predictability of the time series.
    - The raw data provided to us is already seasonally adjusted by the US Census, so there is likely a seasonality component as well that cannot be estimated (since the data have already been adjusted).
- Can the time series be used for short-term or also long-term predictions? Is the recent trend in homeownership an anomaly or expected?
    - It is hypothesized that short-term predictions would be more accurate and that the recent trend in homeownership is an anomaly driven by COVID-19.
    - It is hypothesized that ARIMA modeling would provide the most accurate univariate predictions.
    - For multivariate modeling, it is also hypothesized that short-term predictions will be most accurate.
- Are there external or exogenous factors that help in predicting homeownership?
    - It is hypothesized that each of the three exogenous factors of GDP, interest rates, and median home sales prices will contribute to homeownership rates. It is hypothesized that as GDP increases, homeownership would increase, signaling a strong economy. It is hypothesized that interest rates would have an inverse relationship with homeownership, as interest rate hikes are a form of contractionary monetary policy and signal weaker economic conditions. Finally, it is hypothesized that median home sales prices would have a positive relationship with homeownership, as higher prices signal stronger demand. It is also hypothesized that these relationships may be bi-directional in nature.
    - It is hypothesized that multivariate time series predictions using VAR would be more accurate than univariate time series predictions.

Analysis

Research Question 1: Are there specific characteristics of the time series representing homeownership over time that contribute most to the predictability of the time series?

To investigate this question, we begin by evaluating the characteristics of the univariate time series of homeownership. The data that we were provided with consists of quarterly seasonally adjusted homeownership rates from 1980-2021, representing 42 complete years of data and 168 data points. As seen in Figure 1 below, homeownership rates fluctuated between a high of 69.4% in 2004 and a low of 63.1% in 2016.

[4] Dowell, Earlene K.P. "Remote Working, Commuting Time, Life Events All Affect Home Buyers' Decisions." United States Census Bureau, US Census, 4 Oct. 2021, https://www.census.gov/library/stories/2021/10/zillow-and-census-bureau-data-show-pandemics-impact-on-housing-market.html.

There are clear trends in the data, as seen in Figure 1. There is a gradual and moderate decreasing trend from 1980 to 1995. There is then a sharp and substantial increasing trend from 1995 to 2005. There is a sharp decreasing trend from 2006, coinciding with the start of the Great Recession, that ends in 2016. Finally, there is another sharp increasing trend from 2017 to 2019. Around the onset of the COVID-19 pandemic, this increasing trend continues a few more quarters before shifting to a decreasing trend from the end of 2020 onward.

There does not appear to be a clear seasonality pattern present, however, we will come back to this, as the economic lifecycle is generally thought to exhibit recurring patterns. We should note that the raw data comes seasonally adjusted already. According to the US Census[5], research has shown that seasonality for homeownership rates is present. This is supported by external research as well, which indicates that homeownership rates are generally higher in the spring and summer months.

Next, we evaluate the data for stationarity using the ACF plots shown in Figures 2 and 3. Both ACF plots show violations of the stationarity assumption, as the autocorrelation values for the time series are outside of the confidence bands. There appears to be a linear trend in the data. When evaluating the time series over a longer timeframe of 40 lags (representing 10 years of quarterly data), as seen in Figure 3, there may be some seasonality as well; although the data has already been seasonally adjusted. In summary, there are definite violations of the stationarity assumption with respect to a nonconstant mean and a nonconstant autocorrelation function.

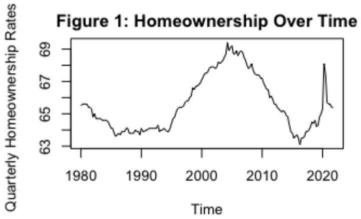


Figure 1: Homeownership Over Time

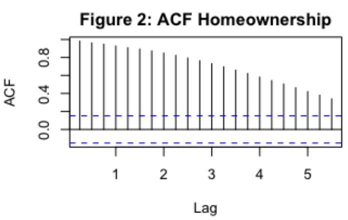


Figure 2: ACF Homeownership

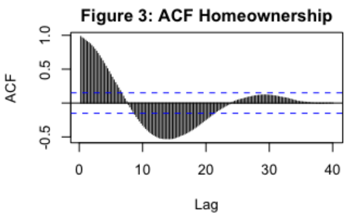
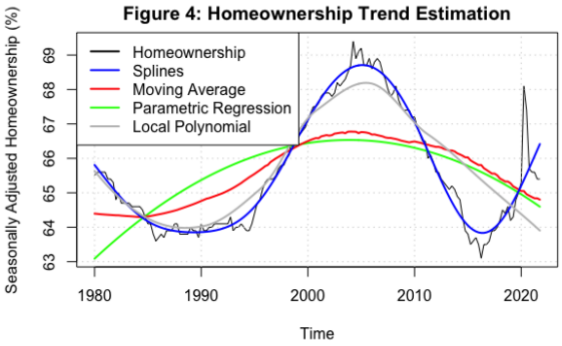


Figure 3: ACF Homeownership

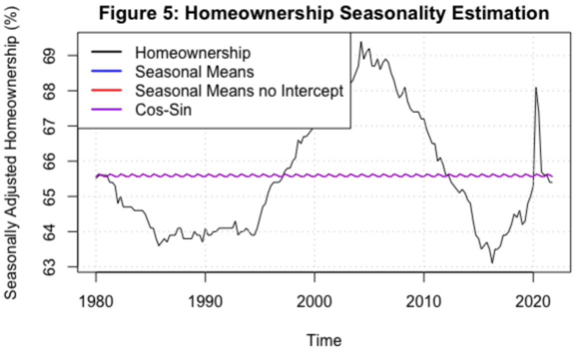
To determine characteristics of the homeownership time series that contribute to predictability over time, we first estimated trend. As seen in Figure 4 below, the trend characteristic does appear to contribute to predictability significantly. Of the different trend estimation approaches, it was notable that splines regression, shown with the blue line, performed by far the best at capturing trend, with a mean absolute percentage error (MAPE) of 0.003162. The local polynomial approach performed well, while other trend estimation approaches were not as successful at capturing predictability.

[5] US Census. QUARTERLY RESIDENTIAL VACANCIES AND HOMEOWNERSHIP, THIRD QUARTER 2022. https://www.census.gov/housing/hvs/files/currenthvspress.pdf



Trend Estimation Technique	Mean Absolute Percentage Error (MAPE)
Splines	0.003162
Moving Average	0.015605
Parametric Regression	0.019226
Local Polynomial	0.007989

Next, the homeownership data was estimated for seasonality. Although the data was already seasonally adjusted, seasonality estimation techniques were executed on the time series given the seasonal nature of the economic lifecycle. The seasonality coefficients were not statistically significant. In Figure 5 below, all seasonality estimation techniques produced similar results and have an overlapping trendline. From visually inspecting the graph, we can tell that estimating for seasonality was not as effective. This is confirmed by MAPE values around 0.02 for each seasonality method applied, much higher than the MAPE values when estimating for trend alone. The lack of effectiveness of seasonality estimation can likely be attributed to the fact that the underlying data has already been seasonally adjusted.



We also attempted to estimate trend and seasonality jointly, via both a parametric and nonparametric approach. The non-parametric approach had a MAPE almost identical to that of Splines regression trend estimation, around 0.003. Autoregressive and moving average components of the time series will be explored in question 2 below.

In conclusion, our null hypothesis is supported that trend contributes to the predictability of homeownership the most. The most effective trend estimation was accomplished through Splines regression. Since we are already working with seasonally adjusted data, there is not much of a need to estimate for seasonality. When estimating trend and seasonality jointly, the predictive accuracy was almost identical to that of trend estimation alone.

Research Question 2: Can the time series be used for short-term or also long-term predictions? Is the recent trend in homeownership an anomaly or is it expected?

Prior to training time series models, it is important to address the nonconstant mean and the violation of stationarity. This could be achieved through differencing the data. In Figure 6, it is apparent that the first order differencing of the data has removed the linear trend and resolved the non-constant mean, for the most part. There is however, a notable spike and subsequent dip in the differenced data that occurs at the onset of the COVID-19 pandemic. In this instance, since homeownership trends during COVID-19 are of interest to be studied, we decided against removing the outliers. Figure 7 shows that after applying differencing, the data lie within the confidence bands and thus could be considered stationary.

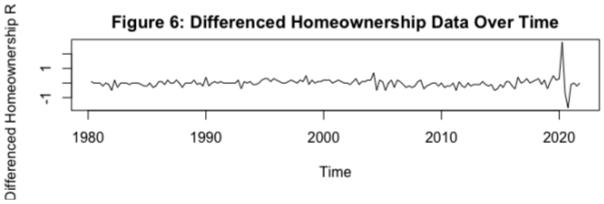


Figure 6: Differenced Homeownership Data Over Time

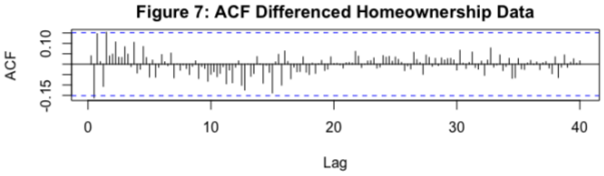


Figure 7: ACF Differenced Homeownership Data

Since using differenced data removed concerns of a non-constant mean and non-stationarity, ARIMA was the first time series model used. ARIMA is a good model to use, with differencing accomplished through the parameter, d. Note that this would be equivalent to building an ARMA model on the first-order differenced data. In tuning model parameters, all models were compared with orders up to 8 for p and q and orders up to 2 for d. The final orders selected were p (AR order) = 3, d (Differencing order) = 1, and q (MA order) = 3. Thus, we conclude that there are autoregressive and moving average effects found in the data. These orders were found to be optimal as measured by the lowest AIC. The diagnostic plots in figures 8-11 show a proper model fit under ARIMA.

The model was applied to both 2-year predictions (2020-2021) and 5-year predictions (2017-2021). While both prediction intervals performed well, the shorter-term 2-year prediction had a lower error rate, confirming our hypothesis. This is particularly notable, given that the homeownership data in 2020-2021 was impacted by COVID-19 trends.

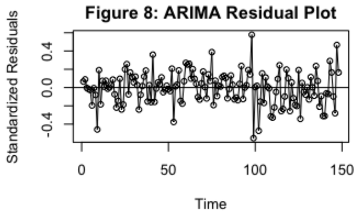


Figure 8: ARIMA Residual Plot

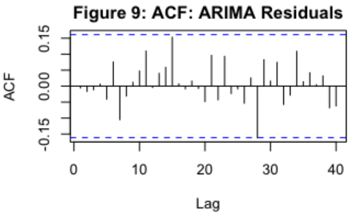


Figure 9: ACF: ARIMA Residuals

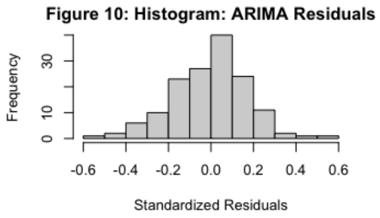


Figure 10: Histogram: ARIMA Residuals

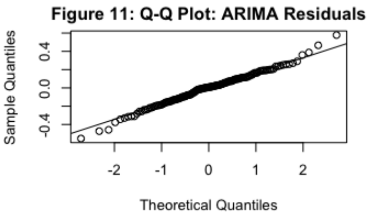


Figure 11: Q-Q Plot: ARIMA Residuals

Given our initial discussion into seasonality, seasonal ARIMA was also evaluated on the data for fit. The optimal seasonal orders were selected as 3, 0, and 3, respectively. The predictions from the seasonal ARIMA model had a bit higher of an error rate than traditional ARIMA. Once again, the model performed better on the two-year predictions than the five-year predictions.

We also explored a log transformation on the data, given the larger variance observed in recent years. The log differenced data was evaluated through ARIMA modeling. While modeling the log differenced data, we did find a slight improvement in prediction accuracy. The tradeoff, however, is that the model becomes more difficult to interpret and the transformation does not appear to be needed for ARIMA modeling, as the diagnostic plots of the residuals using differenced data did not raise any immediate concerns, other than possible heteroskedasticity in recent years. Some of the seasonal ARIMA models found correlated residuals, while the traditional ARIMA models did not find correlated residuals.

Because of potential volatility in the homeownership data, especially during the pandemic, we also explored ARMA-GARCH models on the original and differenced data. We used ARMA-GARCH to account for the conditional mean as well as the conditional variance and to check for volatility. GARCH assumes that the time series exhibits heteroskedasticity, or nonconstant variance, which is a possibility given the ARIMA residual plot above (Figure 8).

In modeling ARMA-GARCH for the original and differenced data, we used the resulting AR and MA orders from the ARIMA model (p = 3, q = 3) as the starting point for the ARMA-GARCH model. From there, we used order selection to select the GARCH model with the lowest AIC value. Table A below details the order selection process for each variation of the ARMA-GARCH model. We found that, of the ARMA-GARCH models, the model with the best MAPE was the 5-Year Prediction using the original time series. This is an interesting finding and contrasts the original hypothesis that short-term predictions would be most accurate.

The differenced time series produced high MAPE values, so we decided to only pursue prediction on the original time series. In both the 5-year and 2-year predictions on the original time series, the Box-Ljung test was performed on the residuals and squared residuals, both of which yielded a p-value of less than 0.05, indicating that the residuals exhibit serial autocorrelation.

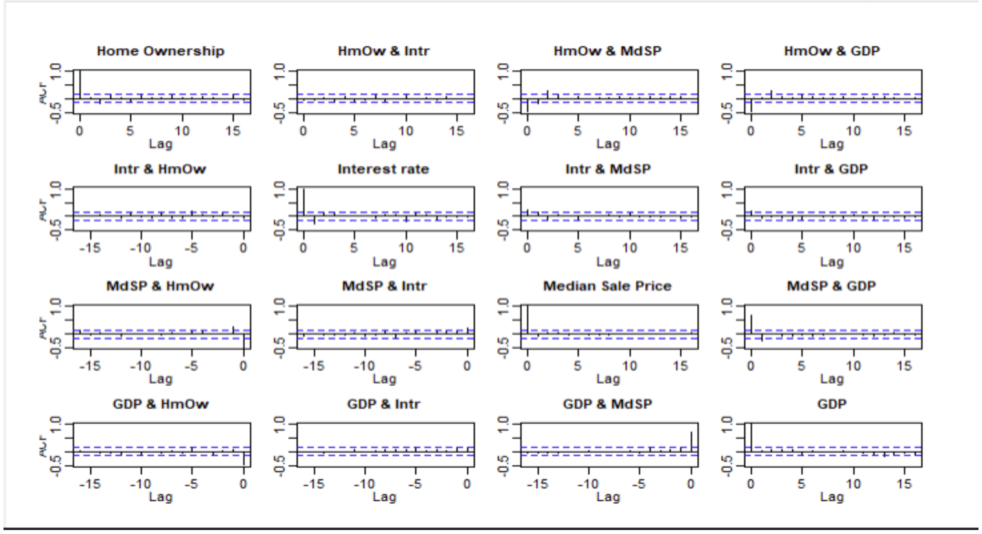
Table A

	5-Year Prediction		2-Year Prediction	
	Original	Differenced	Original	Differenced
Final ARMA Order and AIC	p = 3, q = 5, AIC = -0.6946477	p = 2, q = 5, AIC = -0.5761489	p = 4, q = 4, AIC = -0.6164826	p = 5, q = 5, AIC = -0.6311813
Final GARCH Order and AIC	m = 1, n = 1, AIC = -0.6946477	m = 1, n = 1, AIC = -0.5780169	m = 0, n = 1, AIC = -0.6164826	m = 0, n = 1, AIC = -0.6538681
Model Order Selected	ARMA(3,5) + GARCH(1,1)	ARMA(2,5) + GARCH(1,1)	ARMA(4,4) + GARCH(0,1)	ARMA(5,5) + GARCH(0,1)

As many factors can influence homeownership, short-term and long-term predictions were also made using multivariate VAR time series models including variables of GDP, median home sales prices and interest rates alongside homeownership rates. This will largely be the focus of research question 3, but ties into the predictability analysis of research question 2.

When modeling VAR, we used logged difference data, because like the univariate models, stationarity is required. Before applying the transformations to the data, there were violations of the stationarity assumption. Below in Array A are the auto correlation and cross correlation ACF plots for the logged differenced time series, with the auto-correlation functions being on the diagonal. After applying this transformation to the data, we ensured that the diagonal ACF's resembled a white noise.

Array A: Home Ownership Auto-Correlation and Cross Correlation Functions (Log Differenced Data)



Prior to building the VAR model, we needed to select the right order of P lagged values to use. Using VAR select we found that using an order of 7 resulted in the lowest AIC score while still avoiding higher order complexity. Using non-logged differences, an order of 3 resulted in the lowest AIC score, but this model resulted in a higher MAPE, as well as a non-stationary ACF plot, so it was not selected.

Table B below summarizes the prediction results and model fit metrics for the different models. Prediction accuracy, measured by MAPE, was low for all models, providing encouraging news for prediction capabilities. For ARIMA modeling, short-term predictions were always more accurate than long-term predictions. In contrast, for ARMA-GARCH and VAR modeling, long-term predictions were more accurate. While the log transformation improves prediction accuracy a bit, we would argue that it is not needed for ARIMA modeling and using just differencing the original data affords a more interpretable model. We also found that while ARMA-GARCH helped predict the conditional variance of the homeownership time series, it did not significantly improve accuracy measures of predictions compared to ARIMA.

Table B

Model Prediction	MAPE	Correlated Residuals (Box-Ljung test)
ARIMA 5-Yr Prediction	0.02912	No (p = 0.0863)
ARIMA 2-Yr Prediction	0.01138504	No (p = 0.07806)
Log Transformed ARIMA 5-Yr Prediction	0.007121219	No (p = 0.08144)
Log Transformed ARIMA 2-Yr Prediction	0.002751335	No (p = 0.08013)
Seasonal ARIMA 5-Yr Prediction	0.0414973	No (p = 0.09141)
Seasonal ARIMA 2-Yr Prediction	0.01224573	Yes (p < 0.05)
Log Transformed Seasonal ARIMA 5-Yr Prediction	0.00541186	Yes (p < 0.05)
Log Transformed Seasonal ARIMA 2-Yr Prediction	0.002933699	Yes (p < 0.05)
ARMA-GARCH Log Transformed 5-Yr Prediction	0.00722482	Yes (p < 0.05)
ARMA-GARCH Log Transformed 2-Yr Prediction	0.01343855	Yes (p < 0.05)
VAR 5- Yr Log Transformed Prediction	0.006592377	N/A
VAR 2-Yr Log Transformed Prediction	0.03061297	N/A
VAR 5-Yr Prediction	0.02242129	N/A
VAR 2-YR Prediction	0.03934798	N/A

While we determined that short-term predictions were more accurate than long-term predictions for all ARIMA models that were trained and tested, we also took a closer look at the predictions themselves in recent years to assess the homeownership rates in light of the COVID-19 pandemic. As demonstrated in figures 12 and 13 below, whether doing a shorter-term prediction of 2 years, or a longer-term prediction of 5 years, a majority of the 2020 data was an anomaly that falls outside of the 95% confidence interval of our model predictions. This coincides with the dramatic increase in homeownership at the start of the pandemic, followed by a returning of homeownership levels to familiar level. For the 2017-2021 prediction model, even the 2021 data is at the upper bound of the confidence interval of our model predictions. The 2021 data is right in line with the predictions of the 2020-2021 ARIMA model. These facts support our hypotheses that short-term predictions are more accurate than long-term predictions for ARIMA modeling, while also confirming the hypothesis that recent trends in homeownership during the pandemic contained anomalies that could not be predicted.

On the other hand, we determined that short-term predictions were far more inaccurate for VAR models, as most of the data points were outside the 95% confidence band. For the 2-year prediction, the MAPE was much higher than the 5-year prediction. The poor predictions for the 2-year VAR model can likely be attributed to the sudden changes in predictors and home ownership rates right as the training data was cut off. Due to this, we would not suggest using VAR when trying to model the home ownership rate during the pandemic. While VAR can be used to model short-term home ownership, it is best used for long-term forecasting.

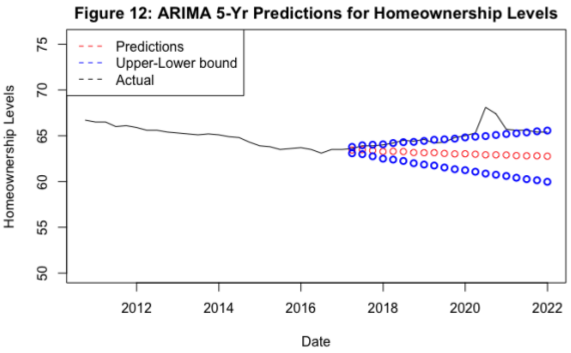


Figure 12: ARIMA 5-Yr Predictions for Homeownership Levels

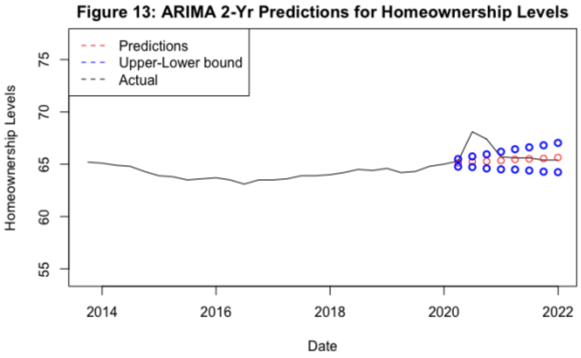


Figure 13: ARIMA 2-Yr Predictions for Homeownership Levels

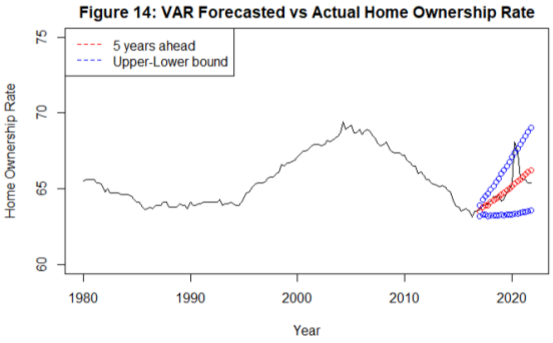


Figure 14: VAR Forecasted vs Actual Home Ownership Rate

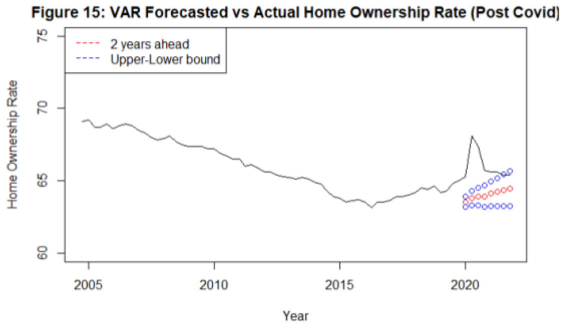


Figure 15: VAR Forecasted vs Actual Home Ownership Rate (Post Covid)

Based on the above models, we selected the model with the best accuracy measures among MAPE and PM to use for forecasting the time series for homeownership rates. For both short term and long term forecasts, we used 95% confidence bands to determine if the observed value is outside the predicted range.

Research Question 3: Are there external or exogenous factors that help in predicting homeownership?

As we continue to explore the relationship between exogenous factors and homeownership through VAR modeling, the variable coefficients for VAR modeling have large p-values. The results of this unrestricted VAR model suggest that there is no significant relationship between Interest rate, median sale price, and GDP, and home ownership rates. This is in contrast to our research hypothesis prior to conducting the analysis. We did, however, find a statistically significant relationship between the lag-1 median sale price and GDP, as well as lag 2 interest rate and median sale price.

After fitting and analyzing the order 3 model, it is necessary to perform a residual analysis. In this analysis, we are going to test for constant variance and normality. The results of the multivariate ARCH test and the JB test both had P values close to 0, indicating that we reject the null hypothesis of both normality and homoskedasticity. Both conditions were violated for the residuals of the VAR model.

We used a Granger causality test to find out if any of the 3 predictors implied forecasting causality. This Granger causality test implies forecasting ability, not real causality. The interpretation of these results is that median home sales price plausibly influences changes in the homeownership rate, while GDP and interest rates do not. We used a significance level of 0.1. The p-values after the Granger causality test were 0.55, 0.092, and 0.16 for interest rate, median sale price, and GDP, respectively. We can reject the null hypothesis that median sales price does not Granger-cause homeownership rates. Testing the opposite direction, homeownership was not found to imply forecasting causality for GDP, interest rates, or median home sales price.

Figure 16: Correlations of Homeownership

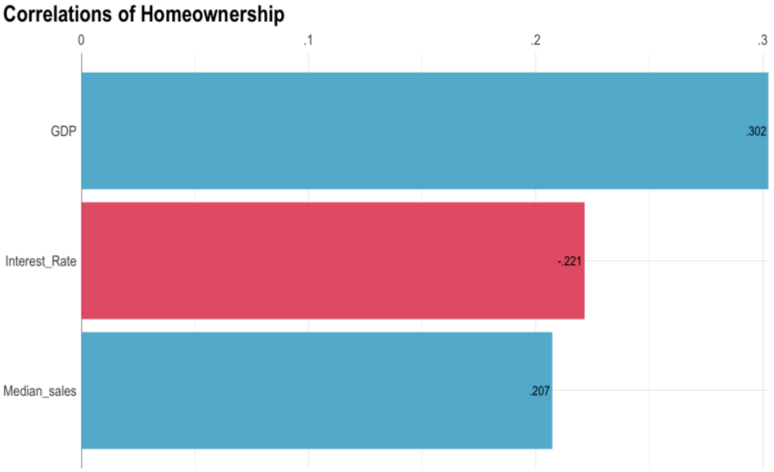


Figure 16 shows the correlation coefficients between the 3 prediction time series and home ownership and allows us to quantify their relationship. Due to the continuous nature of time series variables, it is likely there will always be correlation of some form. GDP had the strongest correlation with homeownership rate. This relationship is different than our findings from the Granger causality test. This is a friendly reminder that correlation is not causation. Overall, we were not able to determine any significant forecasting ability on home ownership rate for two of our time series; contrasting our research hypothesis. The Granger test allows us to rank the 3 factors based on forecasting implications. Those rankings based on Granger causality p-values are: 1. Median Sale price 2. GDP 3. Interest rates.

Conclusion

Through this study, we were able to use different methodologies to analyze homeownership in the United States. We began with univariate analysis, studying homeownership rates alone. We estimated the presence of trend in the time series, which proved to be the factor most responsible for the predictability of homeownership. Different trend techniques were applied, with splines regression providing the best fit. As the raw data was already seasonally adjusted, estimating seasonality was less effective than it would have been on un-adjusted data; although, estimating trend and seasonality together provided slightly more accurate estimates than estimating trend alone. The raw homeownership data also contained non-stationarity.

From there, univariate time series models were trained to homeownership data until 2016 and until 2019. We began with ARIMA modeling, which allowed for differencing the data through the d parameter to address non-stationarity. ARIMA parameters of 3, 1, and 3 were found to be optimal; indicating an autoregressive component, a differencing component and a moving average component. The ARIMA models were used to make 5-year predictions from 2017-2021 and to make 2-year predictions from 2020-2021. It was found that the ARIMA model had stronger predictive accuracy, measured by MAPE, for the 2-year predictions than the 5-year predictions, supporting our original hypothesis. It was also determined that the some of the quarterly homeownership rates between 2020 and 2021, which were heavily impacted by COVID-19 economic conditions, were found to be anomalies, as they were outside the 95% confidence interval of the ARIMA predictions (both 2-year and 5-year). Seasonal ARIMA was also tested, as was ARIMA modeling with log homeownership rates. ARIMA predictions for the log transformed data and seasonal ARIMA predictions were most accurate, but at the expense of interpretative ability.

One could argue that there was heteroskedasticity in homeownership rates in recent years, with homeownership fluctuating greatly, especially in 2020. To investigate further, we implemented ARMA-GARCH modeling on the differenced data and log differenced data. ARMA-GARCH provided more accurate predictions on the differenced data than ARIMA modeling. Interestingly, ARMA-GARCH modeling performed better on 5-year predictions than 2-year predictions. This may be driven by instability in recent years, as homeownership rates have moved greatly due to Covid-19.

There are many factors that influence homeownership rates, including GDP, median home sales prices, and interest rates. As such, we then trained a multivariate time series model, finding that the model would perform best with a lag value of 7 as its parameter, while using log differenced data. The model showed much higher prediction accuracy from the 5-year forecast than the 2-year forecast. The increased volatility in home ownership predictors due to the Covid–19 virus likely played a role in the struggles of our short-term forecasting from 2020-2022. These volatile predictors were not accounted for in the ARIMA model, which had a higher accuracy in short term forecasting. After the model we ran a Granger causality test and found that median home sales price was significant in forecasting implications, while GDP and interest rates were not.

From our findings we have learned lessons about our workflow process and ideas for future modeling. It may be beneficial to have more interaction between the 3 models. One example would be to feed the results from the splines smoothing into the ARIMA or VAR models, starting with adjusted data to magnify the impact of seasonality. Another idea is to incorporate foreign real estate data and analyze other countries that we are economically involved with to predict our domestic housing market. Another way we could expand on our analysis is to incorporate other real estate data aside from home ownership, such as rent prices and Air BNB rentals. The rental market interacts closely with the home ownership market, as real estate investors and a home buyer both factor these values into their decision making. We could also estimate the true seasonality if we began with unadjusted data.