US Home and Rental Values Analysis and Prediction

Yein Kim
University of California, San Diego
y5kim@ucsd.edu

1. Introduction

In this project, we process and analyze the US housing data to understand the trend of home and rental values. We apply statistical metrics that we learned in class, including average, median and Pearson correlation coefficient to get a snapshot of the data. We also go beyond what we learned in class by developing ARIMA models to forecast home and rental values. With these statistical tools and models, we shed light upon three problems: how national income, home and rental values are related, what is the regional pattern of home and rental values and how accurately we can project home and rental values.

2. Data Description

We leverage two data sources in this project: housing data from Zillow Research[3] and income data from US Census Bureau[1]. In total, we use three datasets including home value, rental value and income data.

- Home value: monthly raw values of homes in the 35th to 65th percentile range from January 1995 to November 2020; provided at both national and regional levels.
- Rental value: typically observed market rate rent tracked on a monthly basis from January 2014 to November 2020; provided at both national and regional levels.
- Income: yearly median income provided at national and state levels from 1984 to 2019

Zillow Research provides smoothed and seasonally adjusted versions of home and rental values. However, we use the raw and not adjusted versions in this project.

3. Problem Statement

The aim of this project is to investigate three problems - two exploratory and one predictive. First, we try to explore the relationship among national income, home and rental values. In particular, how is home value correlated with income or rental value? One would assume that income and home value are driven by the national economy and are likely to be positively correlated. We also hypothesize that the rental value is under the macro-economic influence and follow a similar trend as home value. Between income and rental value, can we examine which one is more strongly correlated with home value? In a similar vein, what portion does the rent take out of a person's total income and how does the portion evolve throughout the time? Along with this national data exploration, we examine which regions have the highest home or rental values. There are some regions that we would think of as "expensive" areas, such as New York City or San Francisco Bay Area. Do they indeed have high average home and rental values? Is there a discrepancy in the regional rankings between home and rental values?

Then we move on to a predictive problem of whether we can build a time series model to capture the trend and forecast the US home and rental values. Especially since we are living in an "abnormal" period of so called COVID era, we develop two different models to predict 2019 values (pre-COVID) and to predict 2020 values (during COVID). We examine how the model performance differs between the two periods. The problems that we study can be summarized in threefold as follows:

1. Exploration: relationship analysis among income, home and rental values at national level
2. Exploration: regional analysis of home and rental values
3. Prediction: time series model predictions of home and rental values pre-COVID and during COVID

4. Data Exploration

4.1. Data preprocessing

We first convert the downloaded data frames into time series. We note that the frequency and time range of housing and income data differ - housing data are collected monthly while income data yearly. Hence, when performing the relationship analysis with income, we aggregate housing data time series by taking either the median or average per year. Then we inner join the housing and income time series on year to only consider the time range with both data available.

4.2. Methodology

Correlation. To study the relationship among income, home and rental values, we utilize the Pearson correlation coefficient. Given two time series at comparison $X = \{x_1, \dots, x_T\}$ and $Y = \{y_1, \dots, y_T\}$, we compute

$$\frac{\sum_t (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_t (x_t - \bar{x})^2 (y_t - \bar{y})^2}}$$

where $\bar{x}$ is the sample mean of the time series X and $\bar{y}$ is the sample mean of Y. X is home value time series while Y is either rental value or income time series. We use the aggregated yearly home value time series when Y is income so that the length and frequency of the time series are consistent. For home and rental value comparison, we also compute the correlation coefficient of each year as well as the overall correlation coefficient, since we have monthly data points.

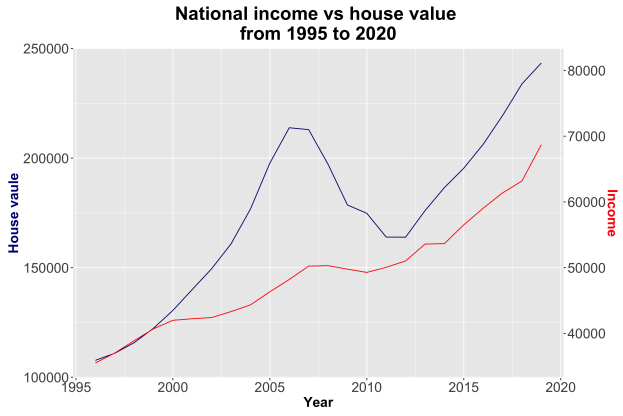
Rent income ratio. We compute the ratio between median rent value and median income each year to understand how much one usually pays for the rent out of their income.

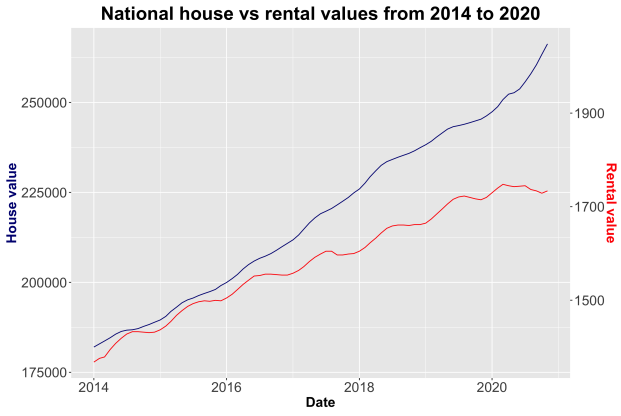
4.3. Results and insights

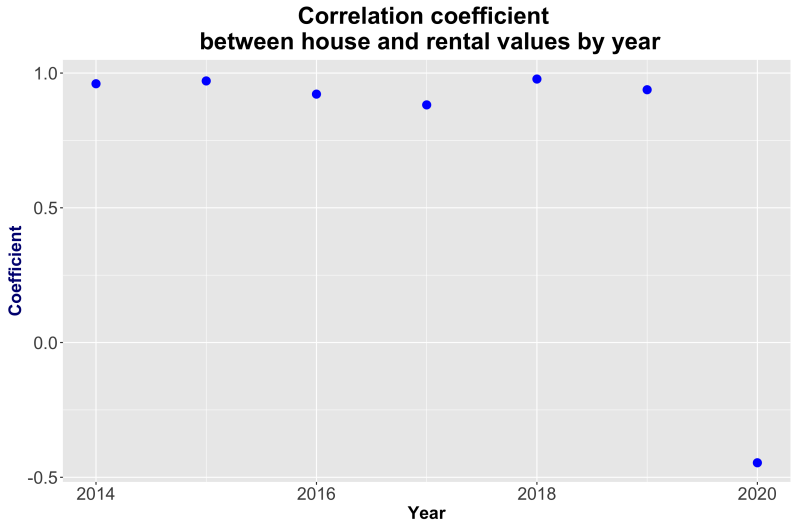
4.3.1 Relationship analysis

By plotting the time series of national home value against national income and rental value, we observe that they are increasing in general. One thing to note is that the comparisons are performed on different time ranges. Especially, rental and home values are compared since 2014, excluding the data between 2007 and 2012 when the house market collapsed. The home values which were increasing almost exponentially peaked in 2006 and continually declined until 2012. Although national income decreased between 2008 and 2010, the drop was not as huge as that of the home value and the median income began to increase again after 2010. The overall Pearson correlation coefficient between national income and home value is 0.89, still showing a strong positive correlation between the two.

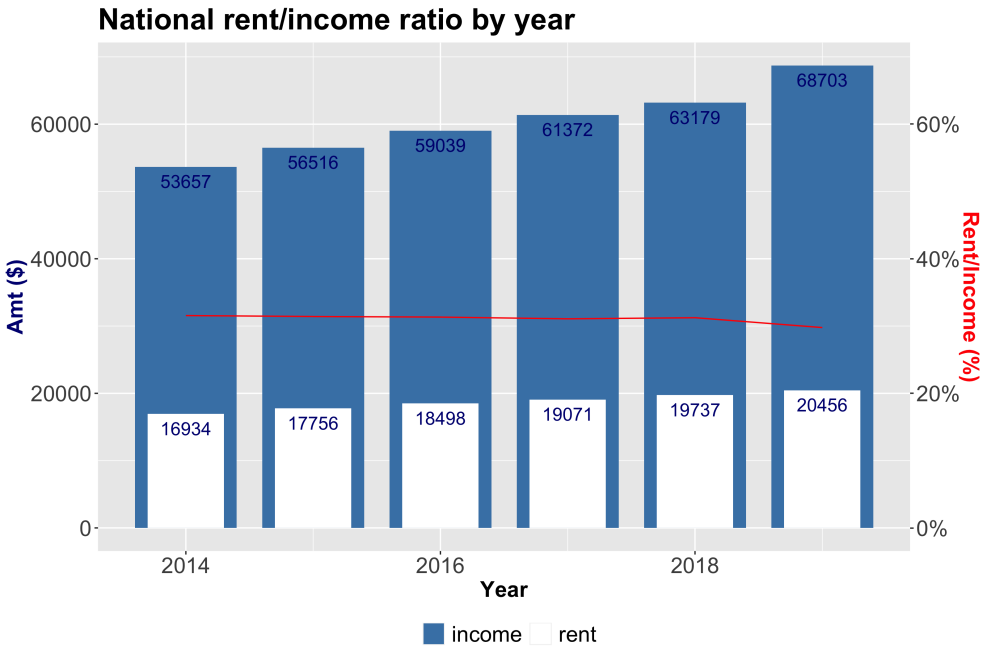
On the other hand, home and rental values have been increasing smoothly together since 2014 despite the seasonality and this is confirmed by the overall Pearson correlation coefficient of 0.99. This trend, however, breaks off in 2020 when the rental value decreases while home value increases sharply. This recent trend seems to make sense as many renters moved out of expensive areas to more affordable areas during COVID as the remote work policy took in place. Indeed, the yearly correlation coefficient dropped to a negative value in 2020.





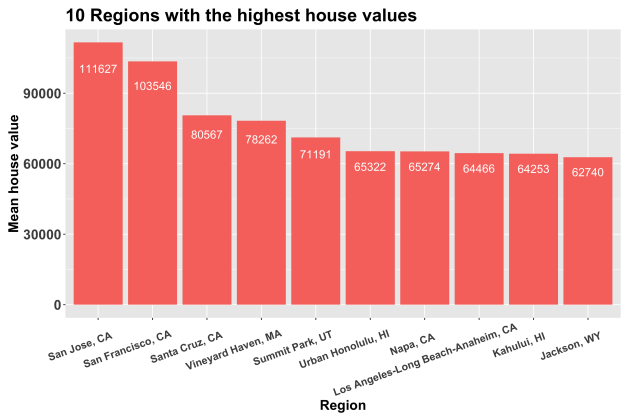


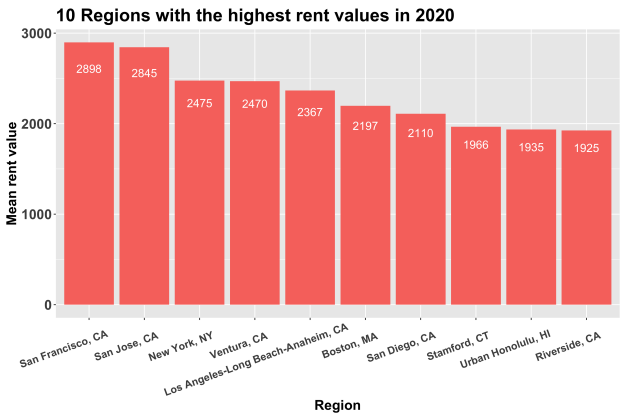
We also compute and visualize the national rent and income ratio from 2014 to 2020. From the bar plot, we can observe that the ratio has been consistent around 30%, implying that the income and rental value have been increasing at similar rates. There is a downtick in 2020 as rental values in expensive areas declined during COVID.



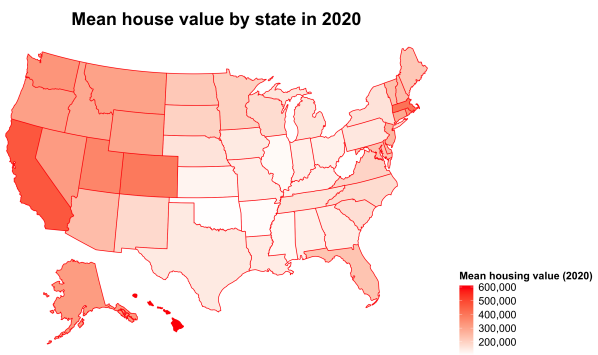
4.3.2 Regional analysis

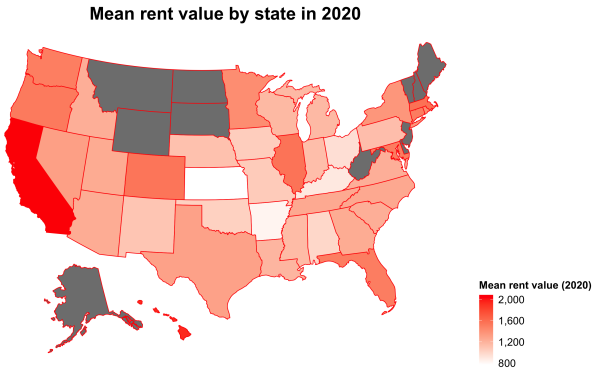
We visualize the home and rental values of 2020 across different regions and states. We first identify ten regions with the highest home and rental values. Unsurprisingly, regions in California dominate the top ten lists. New York City, which has the third highest rental value, did not make it to the top ten in home values while the ranking of Hawaii is higher in home value.





This pattern is confirmed in the state maps below: California and Hawaii have the highest home and rental values. Overall, western and coastal states tend to have higher home values while this discrepancy is not as clear in rent values. For example, Texas and Florida have relatively higher rental values than home values. We can hypothesize that rental values are somewhat equalized during COVID despite the dominance of California.



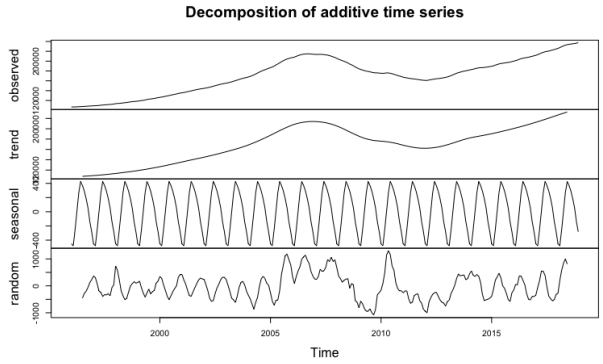


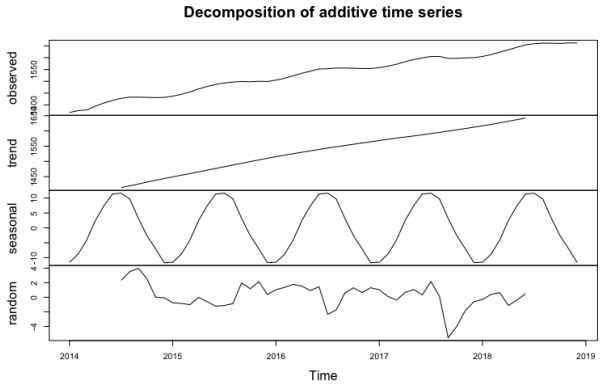
5. Prediction

In this section, we build ARIMA[2] models to capture the temporal trend of home and rental values and to forecast one-year values. ARIMA, short for ”Auto Regressive Integrated Moving Average”, is a class of models that explains time series based on its own lags and lagged forecast errors.

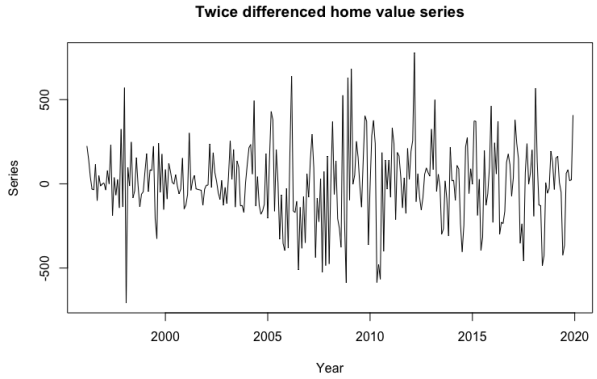
5.1. Data preprocessing

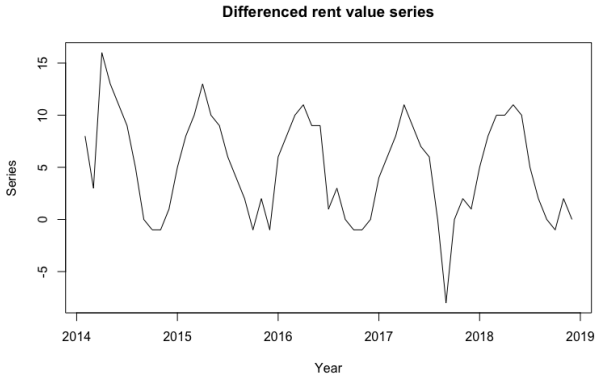
ARIMA model relies on an assumption that data is stationary, independent of the time when it is captured. Thus, we make home and rental value time series stationary by examining their time series components and taking differences. In the plots below, we can observe that home value series shows an increasing, decreasing and then increasing trend while rental value series is consistently increasing. Both series have season components so we adopt seasonal ARIMA models with frequency of 12.





We twice difference the home values and difference rental value series to obtain stationary series as below.





5.2. Methodology

After preprocessing the data, we fit four ARIMA models, two for each of the home and rental value series. Each ARIMA model is selected based on its standard error on training data (the lower, the better).

- Pre-Covid model: ARIMA model trained on series up to and including December 2018 and forecast values from January to December, 2019
- Covid model: ARIMA model trained on series up to and including December 2019 and forecast values from January to December, 2020

We adopt four metrics to evaluate the model performances on test data: MAE, MSE, RMSE and MAPE.

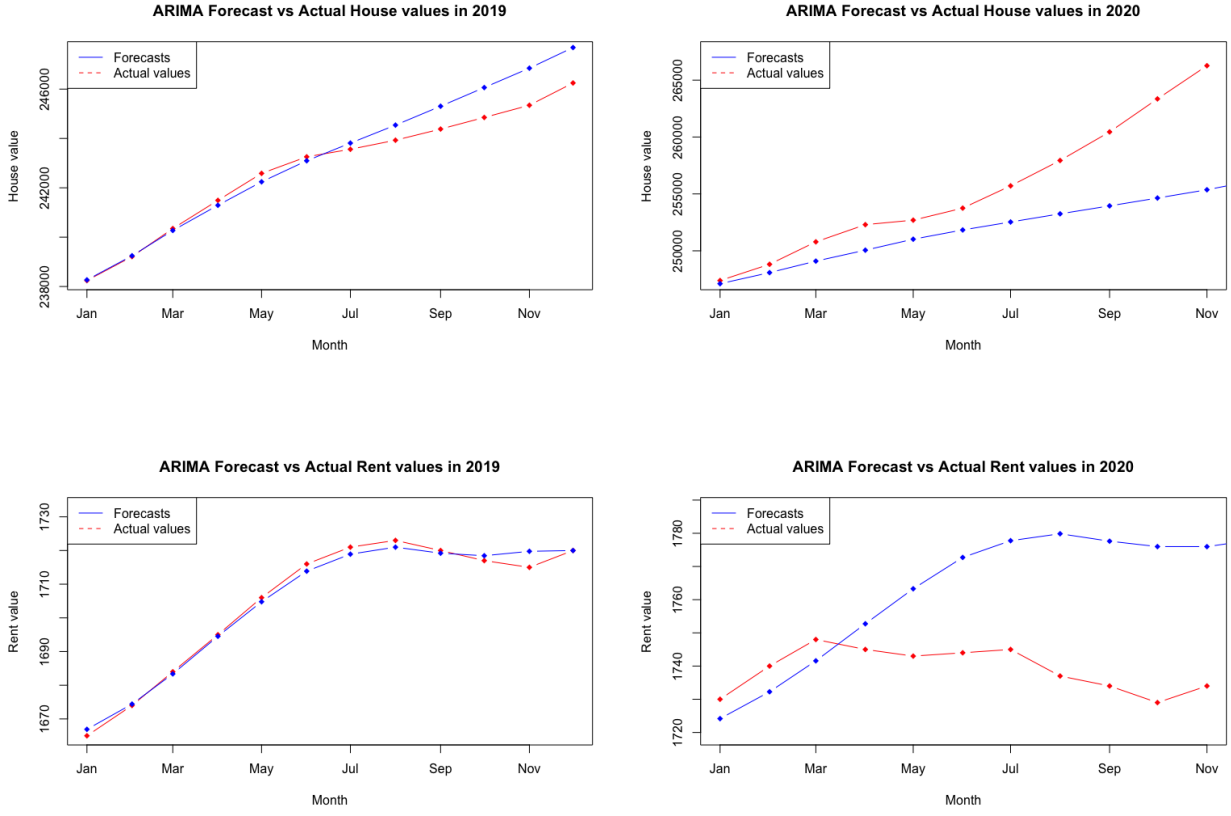
- Mean Absolute Error (MAE) $= \frac{\sum_{t=1}^{n} |\text{pred}_t - \text{actual}_t|}{n}$
- Mean Squared Error (MSE) $= \frac{\sum_{t=1}^{n} (\text{pred}_t - \text{actual}_t)^2}{n}$
- Root Mean Squared Error (RMSE) $= \sqrt{\frac{\sum_{t=1}^{n} (\text{pred}_t - \text{actual}_t)^2}{n}}$
- Mean Absolute Percentage Error (MAPE) $= \frac{1}{n} \left| \sum_{t=1}^{n} \frac{\text{pred}_t - \text{actual}_t}{\text{actual}_t} \right|$

5.3. Results and insights

In the table below, we can observe that ARIMA models forecast home and rental values for 2019 quite well with MAPE of below 1%. ARIMA model's forecasts on rental values are exceptionally accurate - we observed that the trend in rental values is more clear and straightforward. Nevertheless, ARIMA model performances degrade when forecasting 2020 data. Their MAPEs increase to 1.5%, which are more than 7 and 18 times larger than those on 2019 home and rental values respectively.

	Home value prediction				Rental value prediction			
	MAE	MSE	RMSE	MAPE (%)	MAE	MSE	RMSE	MAPE (%)
2019	564	606000	778	0.2	1.5	3.7	1.9	0.08
2020	3865	25823361	5081	1.5	26	929	30	1.5

In particular, we can observe that ARIMA model underestimates home values of 2020 while it overestimates rental values of 2020. This is consistent with our observations on the general trend of home and rental values - home values uptick while they downtick in 2020.



6. Conclusion

In this project, we study housing and income data provided by Zillow Research and US Census Bureau. Leveraging these data, we address three problems of how national income, home and rental values are related, which regions have highest home and rental values in 2020 and how we can forecast the values using ARIMA models. We observed that while home values are are positively correlated with income and rental values, there are periods when the trends of the series diverge, such as housing bubble crisis in 2018 and COVID in 2020. We also identified regions with highest home and rental values. As expected, California has the highest home and rental values although the rental value distribution is relatively equalized in 2020. We wrap up the project by building ARIMA models to forecast home and rental values of 2019 and 2020. While ARIMA models forecast accurately for 2019, their performances degrade when predicting 2020 data. ARIMA models underestimate home

values and overestimate rental values. These observations and projections confirm that 2020 is indeed an abnormal period where both home and rental values deviate from their general trends.

References

[1] US Census Bureau. Historical income tables: Households. https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html, 2020.

[2] Wikipedia. Autoregressive integrated moving average — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Autoregressive_integrated_moving_average&oldid=994736269, 2020.

[3] Zillow Research. Housing data. https://www.zillow.com/research/data/, 2020.