Los Angeles Home Values — An Areal Analysis of Zillow Data

Abstract

A Zestimate's home valuation is Zillow's estimated market value. It is used as a starting point to determine a home's value. In this study, we are interested in looking at the spatial distribution of Zestimate's over-estimation in the greater Los Angeles area. The primary purpose of this study was to investigate the spatial pattern of Zillow price estimation error in each zip code area. A secondary objective was to evaluate the spatial relationships between median house age in a postal zip code area vs. various house-hold related features.

The data come from two sources. Kaggle's Zillow competition dataset (https://www.kaggle.com/competitions/zillow-prize-1) is a household-based sample of Zillow estimation error, provide the point data sample which contains the longitude and latitude of each sampling location (n=383). The ZIP Code Tabulation Areas data provided by the U.S. Census Bureau contains the shapefile for all zip code areas in Greater LA area. We created a shapefile by joining those two datasets by zip code. Spatial autocorrelation in key study variables was calculated with the Global Moran's I statistic. Local Moran's I and Getis G* were used for local cluster detection. We fit and compared ordinary least squares (OLS) regression, a spatial simultaneous autoregressive error (SAR-error) model, and a conditional autoregressive (CAR) model.

We found significant positive spatial autocorrelation in the median house age (Global Moran's I most = 0.08; p = 0.045) but not in proportion of overestimation (Global Moran's I = -0.034, p = 0.745). Median size of house in an area was associated with an older median house age in an area (regression coefficient: -2.26, p= 0.01), and proportion of house of good quality was associated with a younger median house age in an area (regression coefficient = -37.83, p<0.0001). Fitting a simple linear model with no spatial autocorrelation, we see there is still spatial autocorrelation in the residuals. Fitting a SAR model or CAR model, the residuals are no longer spatially auto-correlated.

Introduction

Zillow is an online real estate database with data on homes across the United States. One of Zillow's most popular features is a proprietary property value prediction algorithm: the Zestimate. A Zestimate is Zillow's estimated home market value. It is often used as a starting point to determine a home's actual value. Zillow is constantly trying to improve its Zestimate. To help advance its accuracy even further, it launched a Kaggle competition with a $1.2 million prize. The objective of the Kaggle competition was to predict the difference between the Zestimate and the actual sales price of homes, i.e. the log-error between the predicted log-error and the actual log-error. The log-error is defined as:

$$\text{log-error} = \log(\text{Zestimate}) - \log(\text{actual sale price})$$

A log-error greater than 0 means that Zillow overestimate the sale price while log error smaller than 0 means that Zillow under-estimate the sale price. For the purpose of this study, we only use this dataset for spatial analysis purpose, and we might be interested in predicting the log error in the future. The data provided by Kaggle contains real estate data from three counties in Greater LA area: Los Angeles, Orange and Ventura, California. The data has the actual logerror and feature information for 90,725 properties. Each observation had 56 features.

Figure 1, Figure 2 and Figure 3 shows the sampling locations and the spatial distribution of log error, overestimation, and house age.

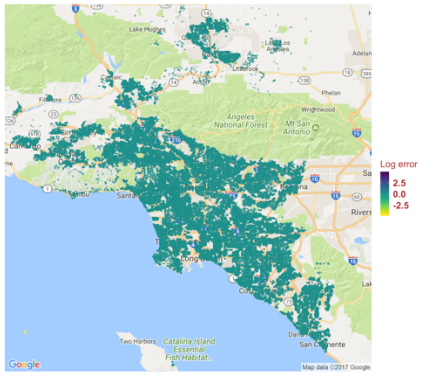

Figure 1

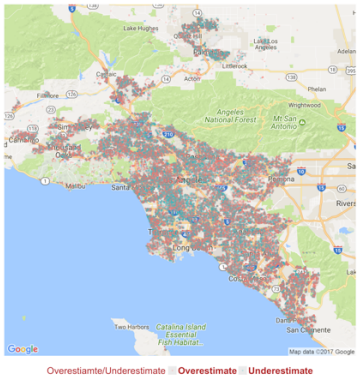

Figure 2

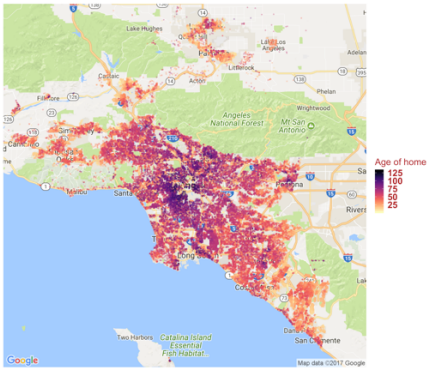

Figure 3

Data processing

Since the Zillow dataset contains point data, we created an area data based on zip code. To do that, we first downloaded the zip code tabulation areas shapefile from the U.S. Census Bureau (nhgis.org). We then used the revgeocode() function in ggmap to extract the zip code for each observation and match that with the shapefile by zip code. For the last step calculate features based on zip code area including: the proportion of Zestimate over-estimation, median house age, mean house size (square feet), mean tax amount, proportion of house of good quality and proportion of apartment.

Data Visualization

We conducted exploratory spatial data analysis. Using ggmap() in R, we are able to map features of houses. This facilitated an initial inspection of potential spatial patterns. We constructed maps to show the spatial distribution of proportion of over-estimation, median house age, median house size and median tax amount the among the sample.

Figure 4 shows the proportion of over-estimation across each zip code area. From Figure 4, we see that the proportion of over-estimation seems to be higher in central LA area and lower in northeast area. There is no apparent spatial pattern from this figure. Figure 5 shows the median house in each zip code area. In general, the houses in central LA area and northwest is older and house in northeast area is newer. Similarly, Figure 6 and Figure 7 shows the median house size and median

tax amount for each zip code area. The median house size is smaller for central and western LA area and larger for eastern LA area.

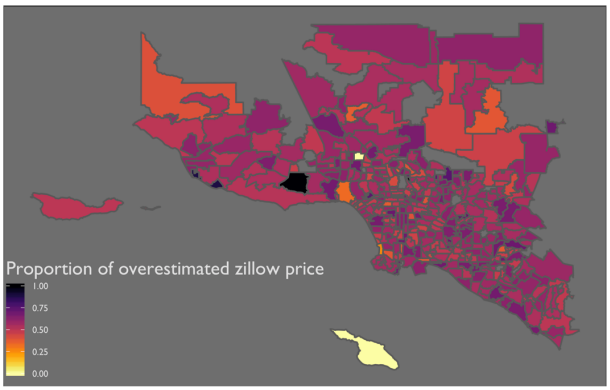

Figure 4 Spatial distribution of over-estimation proportion

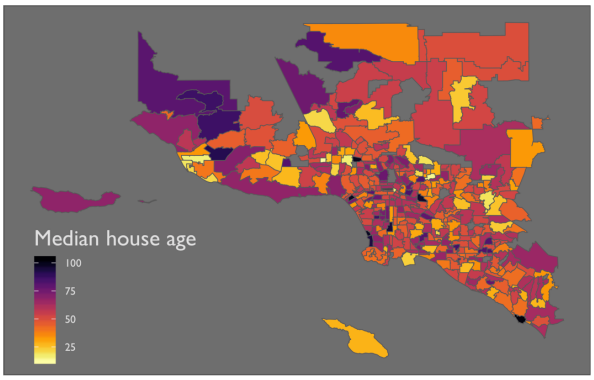

Figure 5 Spatial distribution of Median house age

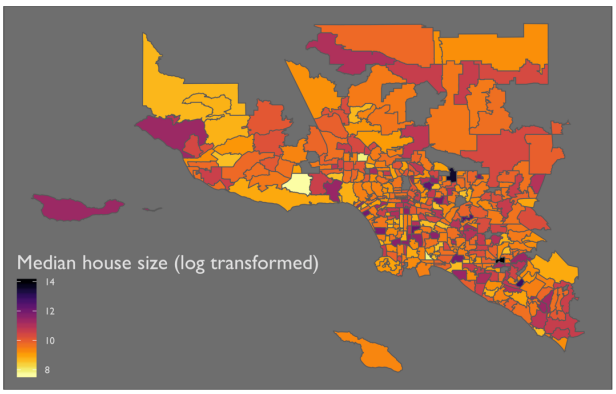

Figure 6 Spatial distribution of Median house size

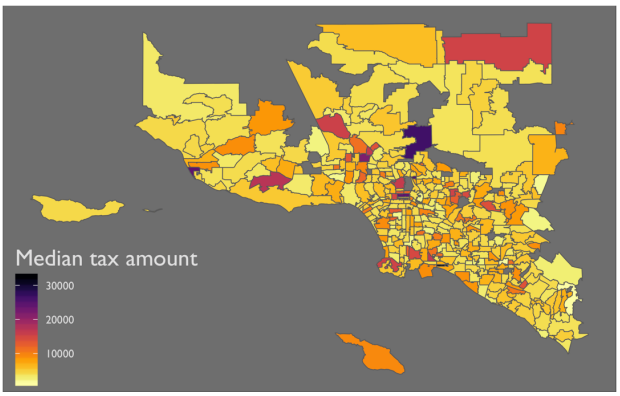

Figure 7 Spatial distribution of Median tax amount (per house)

Spatial analysis — global cluster detection

We assessed whether there is spatial pattern for proportion of overestimation and median house age with the Global Moran's I statistic. For the Global Moran's I calculations and all subsequent spatial regression models, we specified a k nearest neighbor (KNN) spatial weights matrix with k=2. Contiguity based neighbors including queen and rook neighbors were also explored. Figure 8 and Figure 9 shows the connectivity of queen and rook. The connectivity of those two neighborhoods criterion are very similar. Figure 10 to Figure 12 shows the distance based neighbors KNN including K=1, K=2 and K=3. KNN-2 was chosen as the structure for spatial relationships because: (a) we wanted all respondents to have an equal number of neighbors; (b) there is one isolated island in the data which causes some error when using queen or rook. We used row standardization weight matrix to assign weights to the areas that are linked.

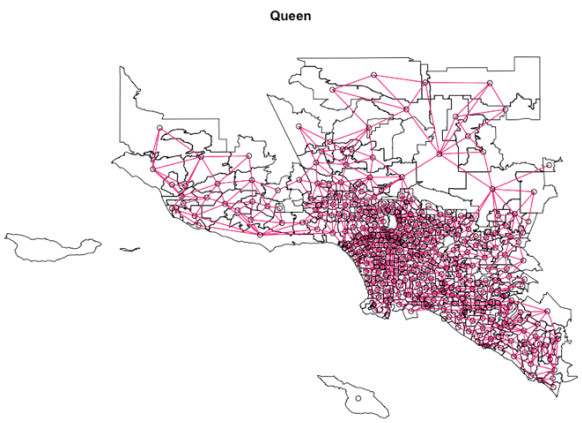

Figure 8

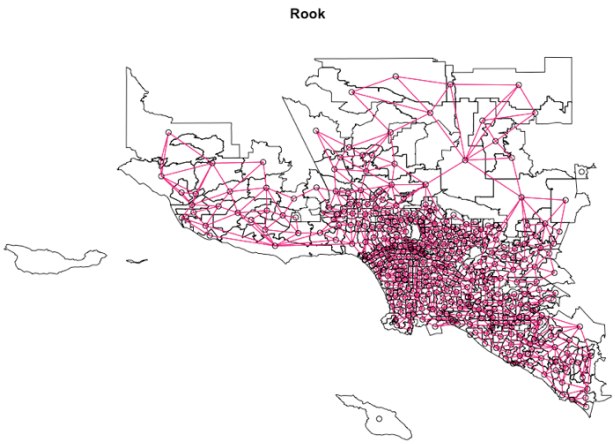

Figure 9

Moran's I is an index that shows the spatial autocorrelation with similarity between areal units. The null hypothesis is that there is no spatial pattern and a Moran's I value near 0 indicates a lack of spatial pattern, i.e. values observed at one location do not depend on values observed at neighboring locations. Positive Moran's I statistic indicates that neighboring regions have similar values. Negative Moran's I statistic reflects that neighboring regions have different values.

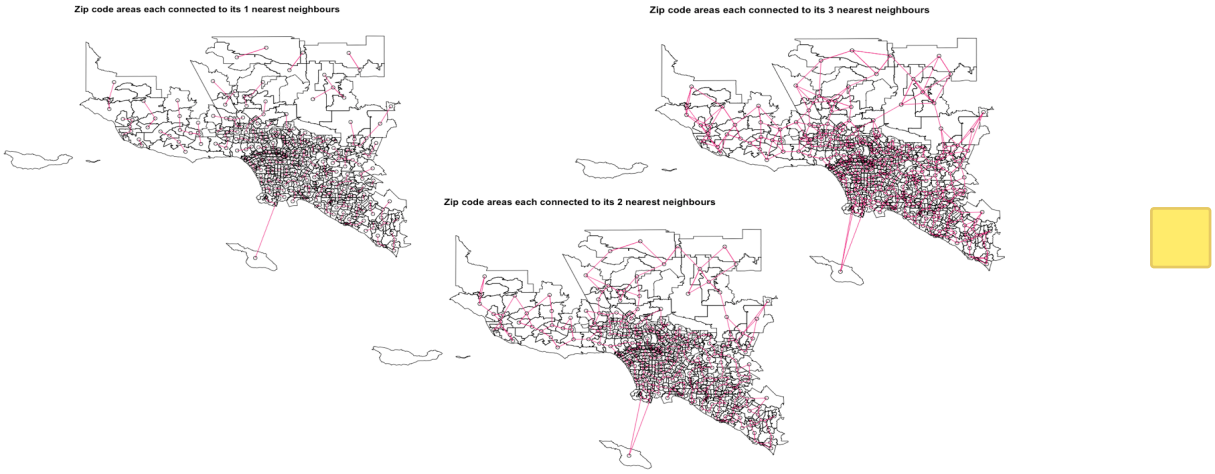

Figure 10 to Figure 12

Table 1 shows the Moran's I statistic and p-value for global cluster detection for proportion of overestimation and median house age. Figure 13 shows a Monte Carlo simulation consisting of 999 random replications to calculate Global Moran's I. From Table 1 and Figure 13, we see that the spatial pattern for proportion of overestimation is not statistically significant. However, there is evidence for spatial pattern for median house age.

Figure 13. Permutation test for Moran's I

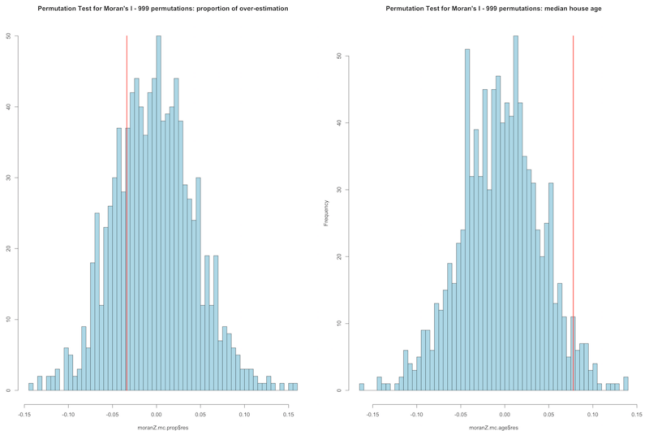

Table 1. Moran's I test

Variable	Statistic	P value
Proportion of over-estimation	-0.034	0.745
Median house age	0.078	0.045

Figure 14 shows the Moran scatter plot for proportion of overestimation and median house age. The slope in the Moran scatter plot is the same as the global Moran's I statistic. From those two plots, we see several outliers including area with zip code 93033 and 90277.

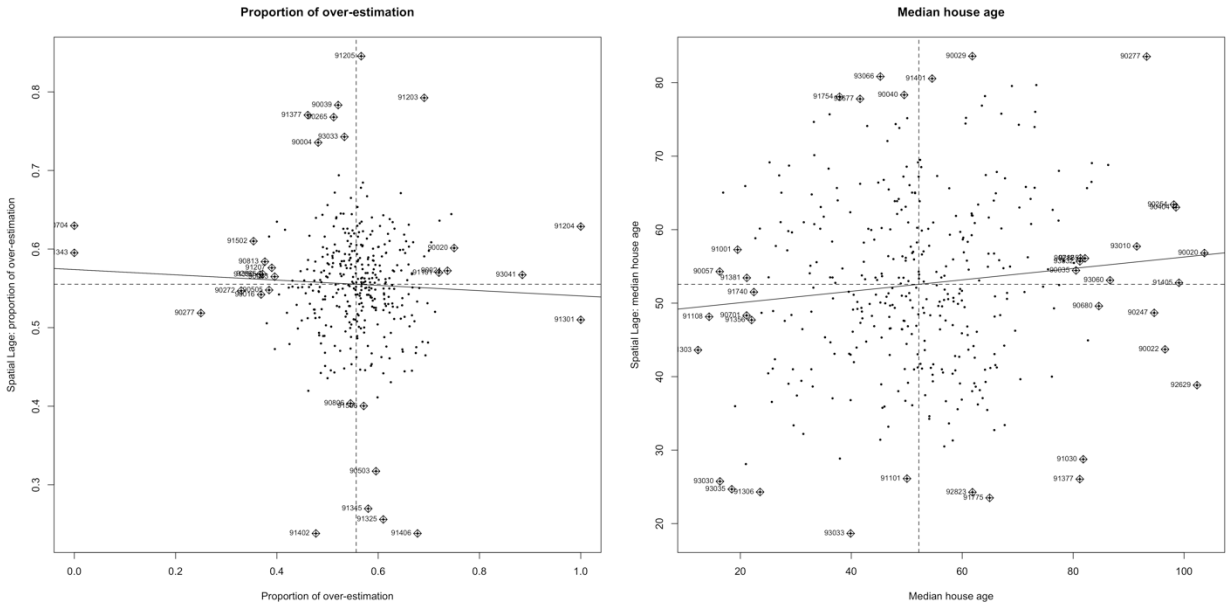

Figure 14

Spatial pattern analysis — local cluster detection

Global Moran's I suggests that there is clustering for median house age but does not identify areas of particular clusters. We used Local Moran's I and Getis Ord G* to detect local areas of similar value. Local Moran's I and Getis Ord G* are local cluster test to identify the location and the statistical significance of local clusters.

Figure 15 shows the Local Moran's I z-score. Figure 16 shows statistically significant local Moran's I as blue dots. Figure 17 shoes the Getis-Ord G* statistic for median house age. The three plots show a similar pattern. From Figure 17, there are some hotspots in central LA area and some

cold spots in north-east area.

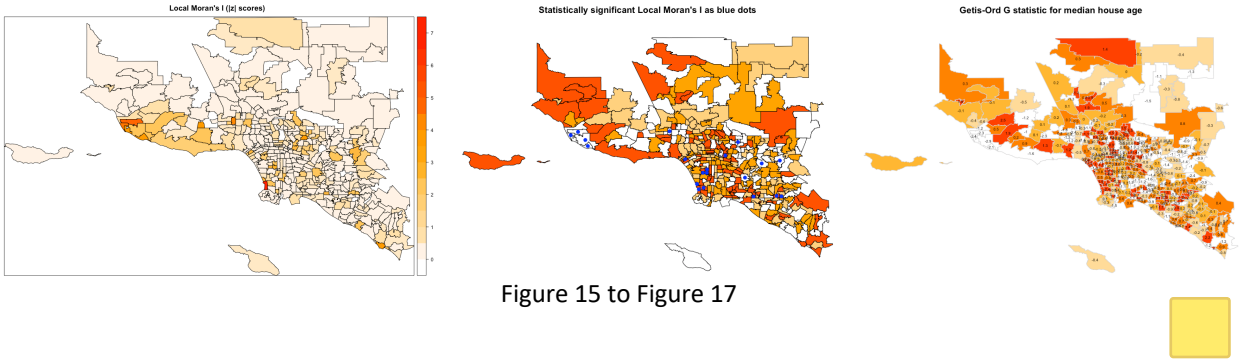

Figure 15 to Figure 17

Spatial regression

To evaluate the spatial relationships between median house age in a postal zip code area vs. various household related features, we first fit ordinary least squares (OLS) regression models. Four predictors were included: mean house size (square feet), tax amount, proportion of house of good quality and proportion of apartment. The residual was plot to show if there is spatial pattern. Global Moran's I test was also applied to the residuals to assess if the OLS regression residuals had significant spatial autocorrelation. If there is still spatial autocorrelation in the residuals, simultaneous autoregressive modeling (SAR) and conditional autoregressive modeling (CAR) were applied. Model fit was also compared using the R-square and Akaike Information Criterion (AIC).

Figure 18

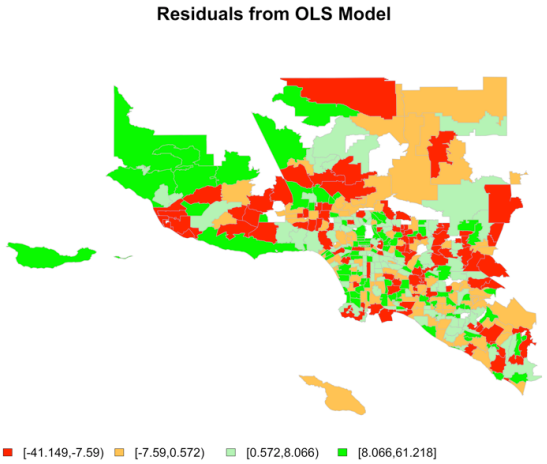

Table 2. Coefficient estimates from ordinary least square model

Variables	Estimate	P-value
Intercept	88.46	<0.0001
House size	-2.1	0.019
Tax amount	1.357e-05	0.945
Apartment proportion	4.65	0.144
Proportion of good quality	-38.7	<0.0001

Figure 18 shows the residuals from OLS model. We still see some spatial pattern from this figure. Table 2 shows the coefficient estimates from OLS regression. Both house size and proportion of house in good quality are negatively associated with median house age.

The Global Moran's I evaluating spatial autocorrelation in the OLS regression residuals for the association between predictors and median house size indicated that there was significant positive

spatial autocorrelation (Global Moran's I: 0.18, p value < 0.0001). Therefore, to take spatial autocorrelation into account, we utilized the SAR error model using the KNN-2 matrix. The residuals were no longer spatially auto-correlated after fitting the SAR model (Global Moran's I: -0.005, p value = 0.519). Figure 19 shows the residuals from SAR model. Table 3 shows the coefficient estimates from SAR model. We also fitted a CAR model for comparison. The residuals are also no longer spatially auto-correlated after fitting the CAR model (Global Moran's I: -0.176, p value = 1). Figure 20 shows the residuals from CAR model. Table 4 shows the coefficient estimates from CAR model. The AIC for SAR is 2889.1 and is 2989 for CAR model. We choose CAR model as our final model because it has a larger value for p-value, indicating it was better at addressing the spatial autocorrelation in the data. There is evidence that there is local spatial correlation since the residuals of the CAR model had a less statistically significant Moran's I.

Figure 19

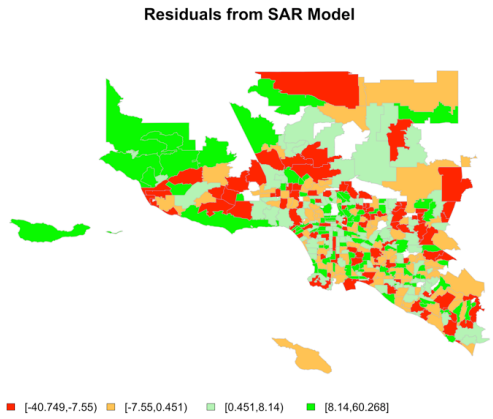

Table 3. Coefficient estimates from SAR model

Variable	Coefficient	P-value
Intercept	91.08	<0.0001
Size	-2.26	0.008
Proportion of good quality	-37.83	<0.0001

Figure 20

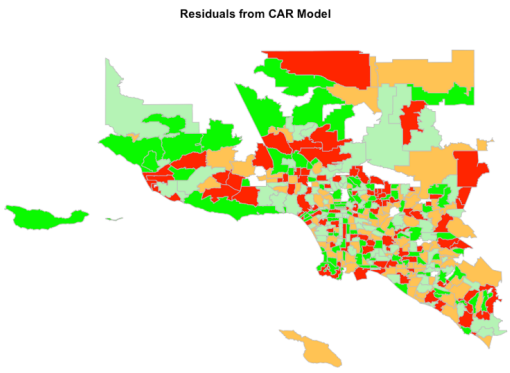

Table 4. Coefficient estimates from CAR model

Variable	Coefficient	P-value
Intercept	81.85	<0.0001
Size	-1.17	0.168
Proportion of good quality	-41.63	<0.0001

Limitations

First, the spatial analysis was based on 383 postal zip code areas in Greater LA area. In some zip code areas, there may not be enough points (i.e. an unbalanced number of houses in the zip codes leading to the modifiable areal unit problem). The result of our analysis might differ if we use a different level, say county level. Second, median house age might not be the right outcome variable for modeling. It might be more interesting to look at other variables such as actual sale price. We used median house age in this study because the predictors in the dataset was small.