Time Series Analysis of California Zillow Dataset

Marti Heit, Zhiyi Ren, William Guenneugues

I. Dataset

The data investigated in this report consists of monthly records of median prices of houses sold on Zillow, unemployment rate, and median mortgage rate in California from February 2008 to December 2016. For the purposes of reporting final model performance, we reserved the last year's data (January 2016 - December 2016). We then reserved one more year's data for validation and model selection purposes (January 2015 - December 2015). The goal of this investigation was to forecast the monthly median price of houses sold.

II. Methodology

Time series data is unique in the way it is modeled and the ways in which models are evaluated. Most model families assume time series come with three components - trend, seasonality, and stationary noise. Trend refers to the overall change in response over time, seasonality refers to repeated, cyclic changes in response, and stationary noise refers to the remaining noises which do not have trend or seasonality. We evaluated models from four families - univariate SARIMA, univariate ETS, multivariate VAR, and multivariate SARIMAX. We chose not to investigate Prophet models because the data are not daily, nor pertaining to business, meaning performance is unlikely to be strong. The terminology used in time series modeling can become confusing, so we have included an appendix which explains the models in more detail.

Due to the sequential nature of time series data, traditional cross validation is not applicable. Thus, we use one step rolling forward cross validation on our training data to tune

parameters. In other words, we split off 20% of the training data to predict, used a model fit on the other 80% to forecast one step out of sample, rolled the training and predictions sets forward one step, and repeated. We then compared the out of sample forecasts generated by the model to the original 20% prediction set. Once we generate the out of sample forecasts, we can calculate different metrics - such as root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) - which use the distance between actual and predicted values to evaluate the model's performance. We chose RMSE as our criterion of choice because it penalizes large errors more strongly than other metrics. There are also model selection metrics which use the concept of maximum likelihood, rather than the residual. One such metric used in this analysis is Akaike Information Criterion, which also penalizes models based on complexity. For our analysis, we used one step rolling forward cross validation with an 80% split and RMSE, a stepwise search to minimize the AIC, or some combination of both to select candidate models within each family. Once we had candidates from each model family, we evaluated forecasting performance on our validation set (January 2016 - December 2016) with RMSE to select a final model.

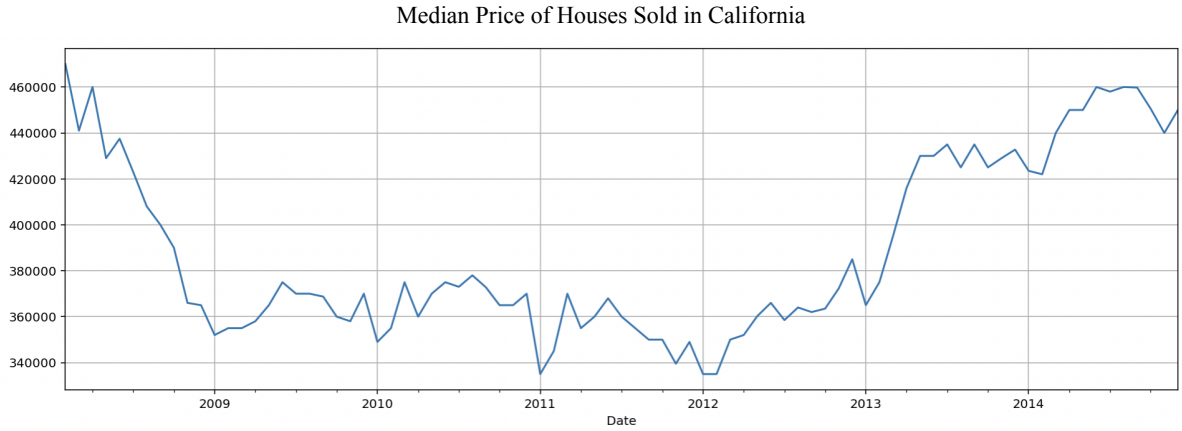


Figure 1

Classical Decomposition of Median House Price

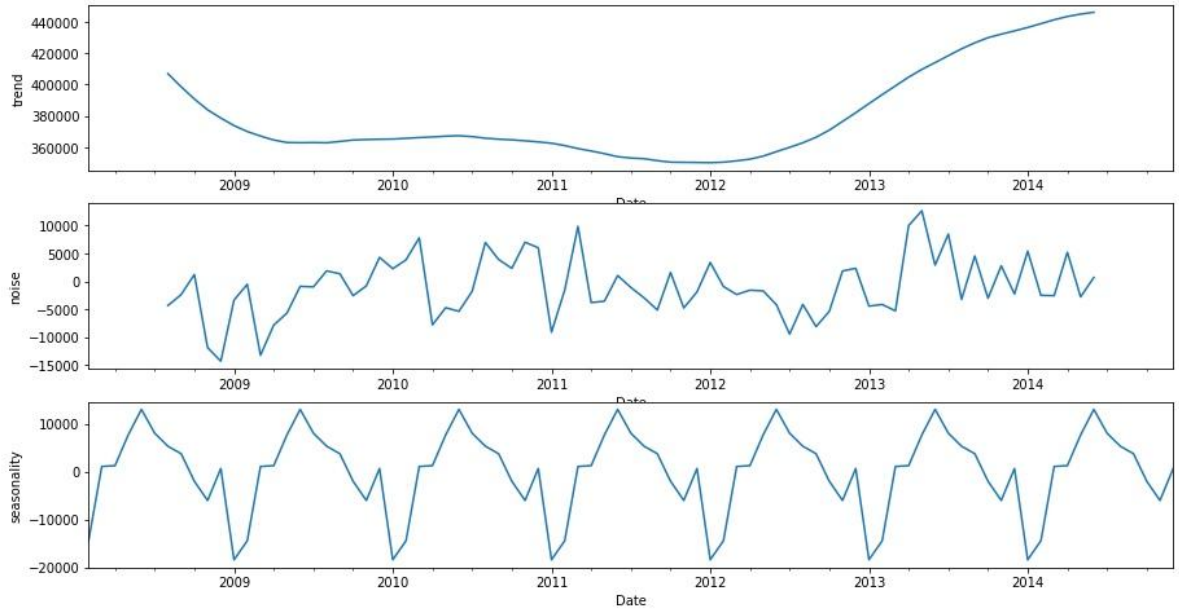


Figure 2

III. Model Families

SARIMA

Seasonal Auto Regressive Integrated Moving Average (SARIMA) models are used to model and forecast univariate time series data as a function of its own previous observations and errors. The initial time series plot of median house price from February 2008 to December 2015 (figure 1) and the classical decomposition (figure 2) showed a significant trend. This is consistent with the results of the Augmented Dickey Fuller (ADF) test for stationarity, which can be used to determine parameter 'd', or the number of times differencing is required to make the time series stationary. With a p-value of 0.86, which is much greater than the level of significance $\alpha = 0.05$, we decided to differentiate once. The time series plot still showed a trend, which was again confirmed by the ADF test, so we differentiated the data again. Then, using a

grid search, ARMA parameters p and q were chosen to minimize AIC. Based on AIC, we got the first raw candidate model which does not include seasonality: ARIMA (2, 2, 3).

Using the classical decomposition and prior knowledge about the data, we determined yearly seasonality was present. Our data were collected monthly, so the seasonal lag was set to 12. Then, we did a grid search combination of trend and seasonal order to find those which minimized RMSE: SARIMA (1,1,0), (2,1,1,12). We performed another stepwise search to minimize AIC, this time including seasonality and got our third SARIMA candidate model: SARIMA (3,1,5), (0,1,0,12). Finally, we used RMSE calculated from one-step cross-validation with an 80% split of the original training data to determine the best of these three candidates.

ETS

Exponential Smoothing Techniques (ETS) are a special type of SARIMA model in which the prediction is a weighted sum of the past observations, with exponentially decaying weights. ETS models are usually less general than ARIMA, but they don't require order selection in the same way ARIMA models do. To select the best ETS model, we exhaustively tested each parameter combination and selected the combination of parameters that gave us the lowest RMSE score. The parameters we tuned for were trend, which can be additive, multiplicative or None, and seasonality, which can be additive or multiplicative. To calculate the RMSE for each parameter combination, we performed rolling forward one step cross-validation with an initial split of 80% on the train set. Using this technique, we calculated the RMSE for each possible parameter combination, and found that the best ETS model has trend parameter set to None, and seasonality parameter set to additive.

VAR

Vector Auto Regressive (VAR) models are simply AR models which simultaneously model endogenous variables as a combination of their prior observations. We believe the feature of interest - median house price - could have a two-way relationship with unemployment rate and mortgage rate. Essentially, we believe the median price of houses sold might be affected by the unemployment rate and the mortgage rate; also the unemployment rate and mortgage rate might be affected by the price of houses sold. VAR models assume stationarity, so two rounds of differencing were necessary prior to fitting the model. Furthermore, because VAR models all variables simultaneously, we had to differentiate all variables together - even though the mortgage rate was initially stationary.

We fit a VAR model with house prices, mortgage rate, and unemployment rate all as endogenous variables. The VAR approach models all endogenous variables simultaneously, but for the purposes of this analysis, we were interested in only the forecasting of median house prices. AR, and thus, VAR models have only one parameter, lag order, which determines how many prior observations to consider. We chose to use AIC as the selection criterion, as RMSE can be prone to underfitting, and thus chose the lag order which minimized AIC.

SARIMAX

We also investigated the multivariate SARIMAX models, in which median mortgage rate and unemployment rate are treated as exogenous variables and are combined with prior history and noise to predict median house price. However, the SARIMA model family has such high cardinality of possible parameter combinations that an exhaustive grid search with one step cross validation is too expensive to be feasible. In this case, we used stepwise search and found parameters which minimized AIC. Using the seasonal order as determined by AIC, we then

performed grid search with rolling forward cross validation to find the trend order which would minimize the RMSE. This secondary search showed SARIMAX (0, 0, 1) (0, 1, 2, 12) to have the lowest cross validation RMSE.

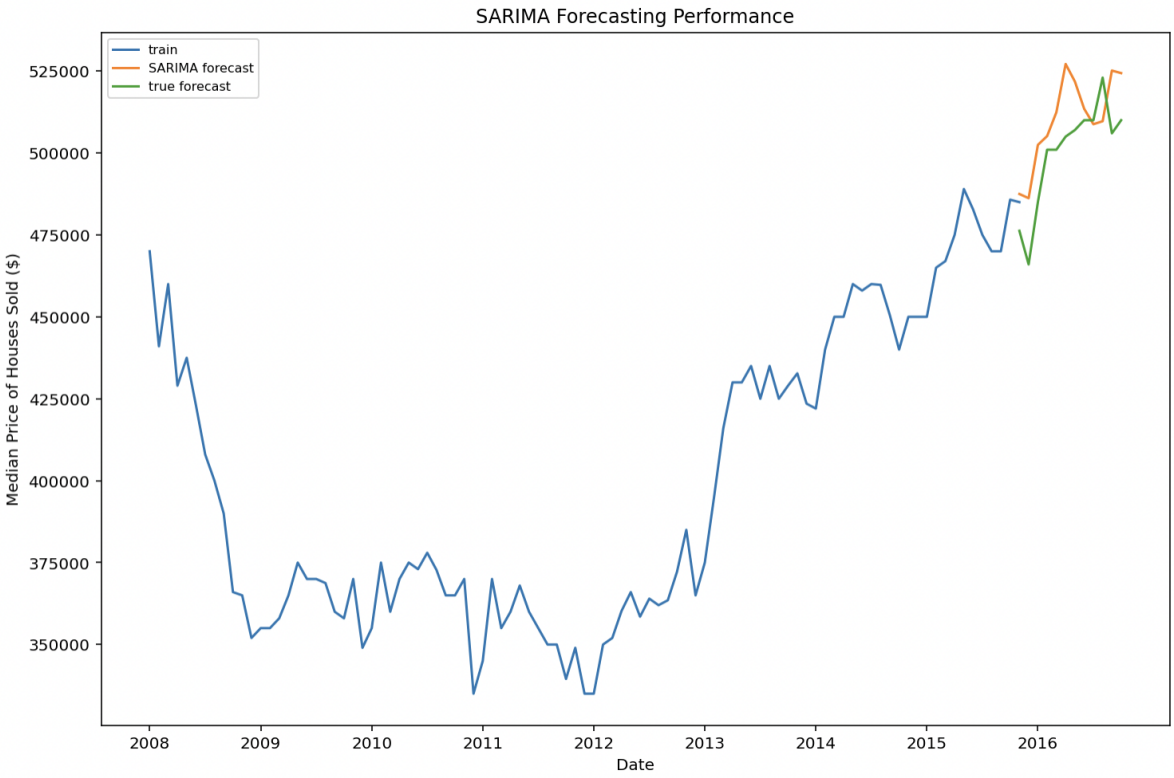
IV. Model Selection

At this point, we have found four candidate models - one from each model family. We used the validation set that we set aside at the beginning of analysis (observations from January 2015 - December 2015) to choose a final model. We used each candidate model to forecast the validation set and chose the model with the smallest RMSE.

Model	RMSE on validation set
SARIMA (3, 1, 5) (0, 1, 0, 12)	$ 10,384.30
ETS (no trend, additive seasonality)	$ 25,724.05
VAR (15)	$ 21,019.66
SARIMAX (0, 0, 0) (0, 1, 2, 12)	$ 18,159.42

V. Model Evaluation

At this point, we have our final model: SARIMA (3, 1, 5) (0, 1, 0, 12). We used the final test set, consisting of observations from January to December of 2016, to calculate a final RMSE score for our model. Our RMSE on the test set was: **$ 14,322.89.** This means our model's predictions of median monthly price of houses sold in California for the 2016 year were $14, 322.89 away from the actual values on Zillow.



VI. Appendix

Model	**Description**
SARIMA (p, d, q) (P, D, Q, m)	seasonal auto regressive integrated moving average
ETS	exponential smoothing techniques
VAR (p)	Multivariate vector autoregressive
SARIMAX (p, d, q) (P, D, Q, m)	Multivariate seasonal auto regressive integrated moving average

Key Term	**Definition**
trend	Overall change in response over time
seasonality	Repeated change in response over specific periods
stationary	A time series without trend or seasonality - random noise

Parameter	Description
p	Order of previous observations to consider in trend
d	Order of trend differencing needed to achieve stationarity
q	Order of previous noises to consider in trend
P	Order of previous observations to consider in seasonality
D	Order of seasonal differencing needed to achieve stationarity
Q	Order of previous noises to consider in trend
m	Seasonal lag