# Time Series Analysis of Housing Price Data in Alameda County

Sequential observation of the quantity of interest over time is defined as a time series. Observations can be recorded (or measured) on hourly, daily, weekly monthly, quarterly or annually basis. The basic idea behind the time series forecasting is to estimate how the past trend continues into the future. Most of the models used in this chapter can be mathematically written as,

(1)

The time series was only dependent on the past values and no external variable, which may affect the system. The *Error Term* allowed for the random variation and effects not included in the model. The training set for time series models was housing price data of the Alameda County from April 1996 to December 2017. Monthly median housing prices from June 2018 to July 2018 were used as the test set. Performance of each model on the test set was compared using the mean-squared errors (MSE).

## Time Series Data

Figure 1 plots the Alameda County housing price time series. Few features noted from the plot were as follows,

* The time series plot had two regions of positive trend, from 1996-2007 and 2012-2018.
* There was a negative trend from 2007- 2012, indicator of the real-estate financial crisis.
* Just looking at data it was hard to conclude if any seasonality was present in the time series. Seasonality was tested separately as it could affect the model selection (seasonal vs non-seasonal models).

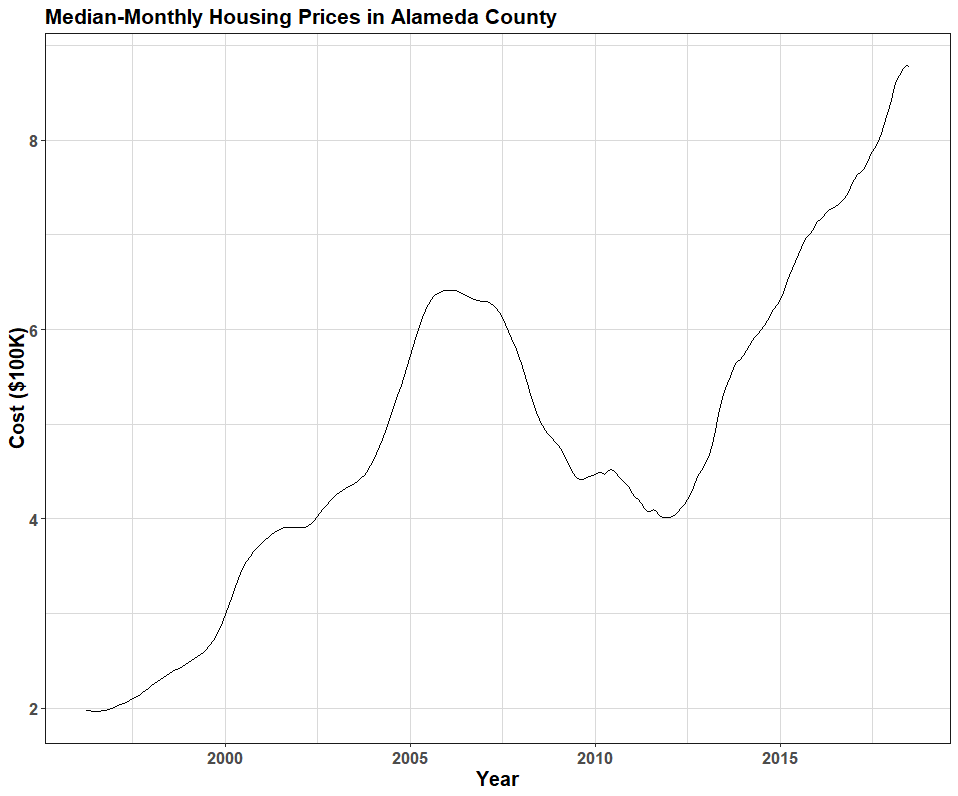


Figure 1 Median-monthly housing price (in $100,000).

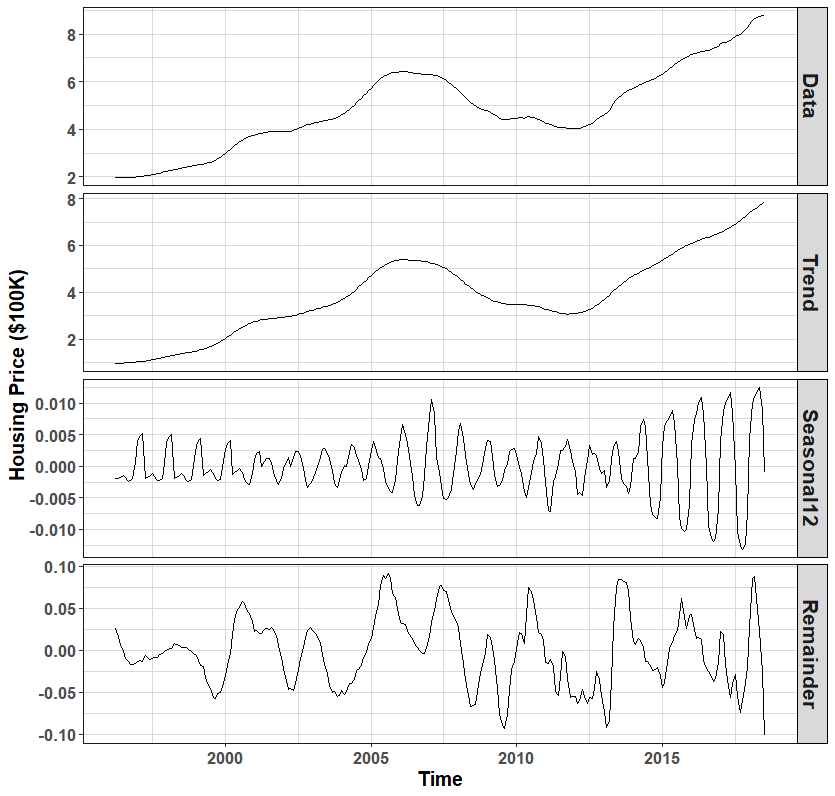


Figure 2 Decomposition of the additive time series.

# Forecasting Models

ARIMA(*p,d,q*) model of the differenced series can be written as,

(2)

Rearranging the above equation and using back difference operator equation 4.10 can be reformulated as,

(3)

## ARIMA(0,2,3)

Table 1 shows the output of the auto.arima model for the housing prices in the Alameda County. ARIMA(0,2,3) was selected by R based as the best fit model on the training data. Using the coefficient estimates in Table 4.4, the housing price time series of Alameda County was,

,

The error terms had extremely low variance and zero mean, thus error impact on present housing prices was almost negligible. The model was approximated as,

,

ARIMA(0,2,3) model suggested that the present value of housing price was dependent only on previous months’ housing price and change in the housing price. Table 1 lists the prediction summary and the test MSE. Test MSE of ARIMA(0,2,3) was 0.72%. Figure 3 shows the residual plots for ARIMA(0,2,3) model. Both ACF and PACF plots suggested that the residuals were correlated as their magnitudes exceeded the significant bound at multiple lags. Thus the model was not a reliable forecasting model for the Alameda County housing prices. For the 36 lags shown, a model was acceptable as long as no more than 2 consecutive lag values exceeded the bound.

Table 1 ARIMA(0,2,3) Model (R output).

|  |
| --- |
| Model Summary |
| ARIMA(0,2,3)  Coefficients:  ma1 ma2 ma3  0.9007 -0.1267 -0.5625  s.e. 0.0584 0.0617 0.0654  sigma^2 estimated as 9.991e-05: log likelihood=825.5  AIC=-1642.99 AICc=-1642.83 BIC=-1628.76  Training set error measures:  ME RMSE MAE MPE MAPE  Training set 0.0003006419 0.00989911 0.006937339 0.01359208 0.1424329  MASE ACF1  0.1577909 -0.1653788 |
| Forecast Summary |
| Time TestData Forecast Residuals Residual^2  2018-01-01 8.423 8.385408 0.03759224 1.413177e-03  2018-02-01 8.569 8.480777 0.08822271 7.783246e-03  2018-03-01 8.654 8.575230 0.07876959 6.204648e-03  2018-04-01 8.698 8.669684 0.02831646 8.018221e-04  2018-05-01 8.755 8.764137 -0.00913666 8.347856e-05  2018-06-01 8.786 8.858590 -0.07258978 5.269277e-03  2018-07-01 8.783 8.953043 -0.17004291 2.891459e-02 |
| Test MSE for ARIMA (0,2,3) = 0.72% |

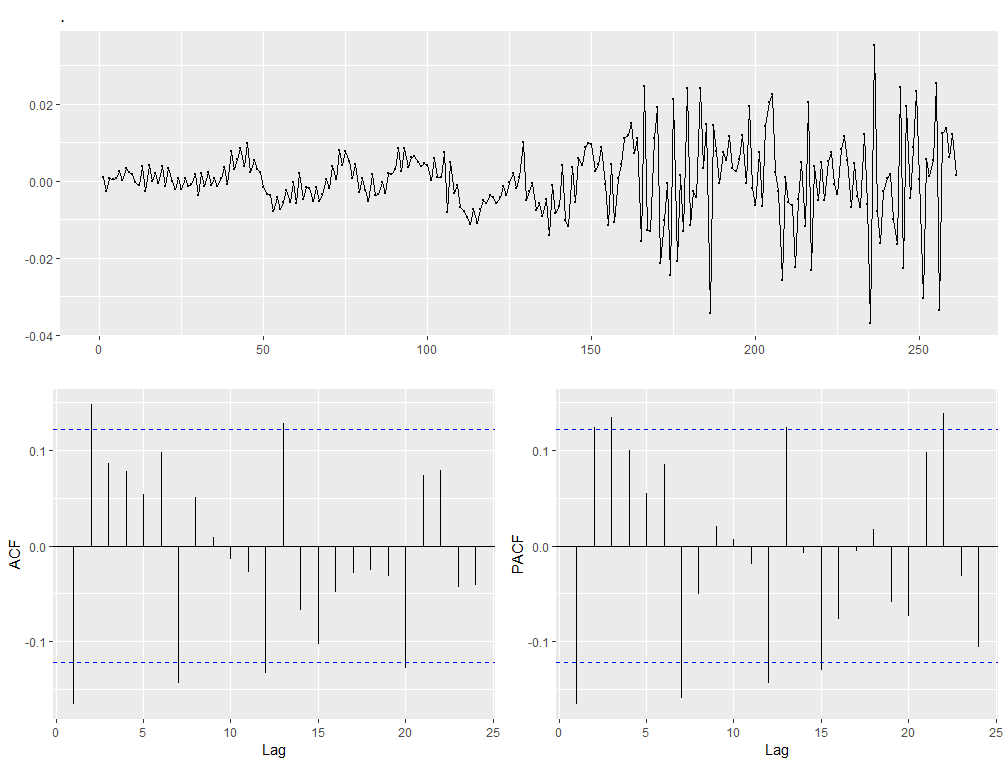


Figure 3 Residual data for the ARIMA(0,2,3) model.

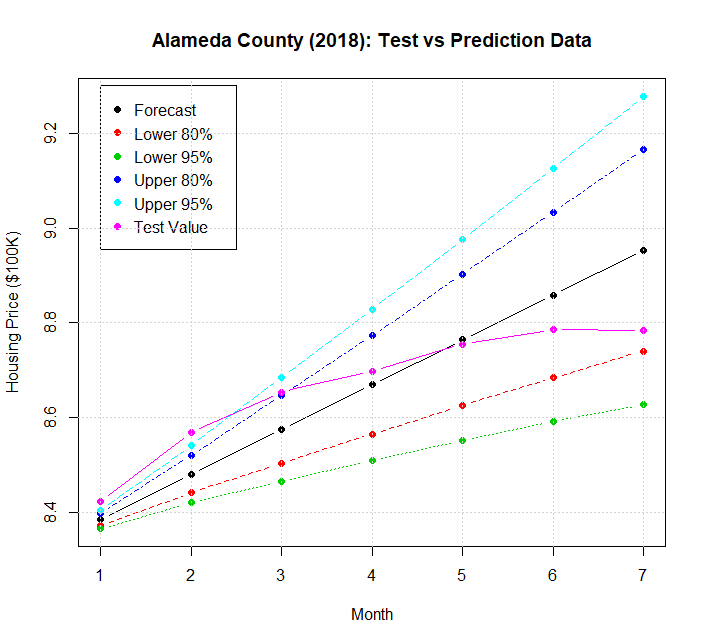


Figure 4 Test data and forecast of ARIMA(0,2,3).

Based on the above plots the time series analysis results from the standard R package are not reliable. Other options are:

1. To consider train-and-error method to find a suitable model fit or, (2) to perfume interval analysis around the housing price crisis time zone.