Future Prediction of: Insights for Potential Buyers

by Dongmin Wu

September 12, 2024

Abstract

This paper conducts a time series analysis of housing prices in Southern California from 1995 to the present. The goal is to predict future prices and provide insights for potential homebuyers. We have applied Box-Cox transformations for stabilizing the variance and decomposed into trend, seasonal, and residual components with ARMA modeling to the residuals. Finally, a ten-year forecast was generated with predictions from the smooth and rough components, with potential future trends in home prices.

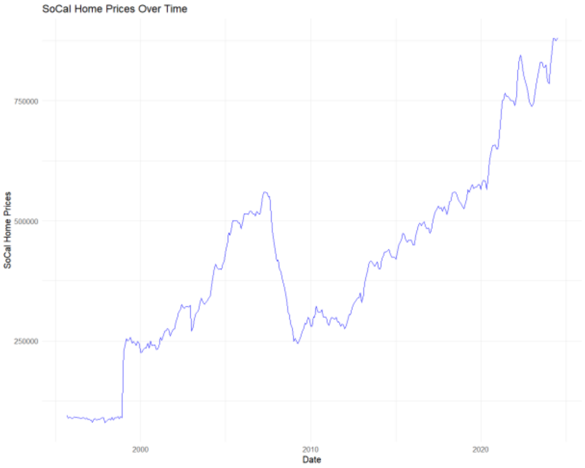
Introduction

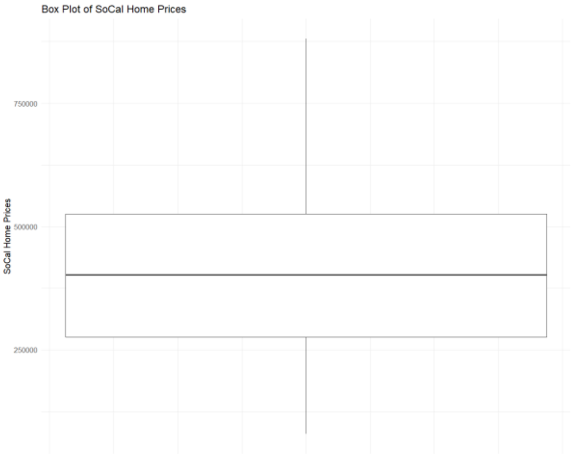
The real estate market in Southern California (SoCal) is a key component of personal financial planning and broader economic trends. Yet it is very unpredictable with home prices constantly fluctuate, which raised concerns for potential homeowners or potential buyer who are planning to make a purchase in the near future. Before making an informed decision, it is essential to forecast future home prices and identify trends that could help determine the best time to buy and understand long term real estate appreciation, i.e. When home prices might be lower? How much value a property might gain over time?

This project aims to forecast home prices over the next 10 years using time series analysis by combining trend, seasonality, and the rough part (residuals) of the data and ARMA modeling, basing on data collected from the California Association of Realtors, specifically the "Median Prices of Existing Detached Homes Historical Data."

The rest of this report breaks down to how the data was analyzed, what method and models were used, and what these forecasts mean for homeowners and homebuyers.

Data Description





The data were used for this report comes from the *Median Prices of Existing Detached Homes Historical Data* provided by the California Association of Realtors (CAR). The dataset contains monthly median home prices across various regions in California, but this report focused specifically on Southern California (SoCal). The recorded data starts from 1990, and the home prices were analyzed up to the latest available data (Jul-2024).

We can observe some interesting features in the graph

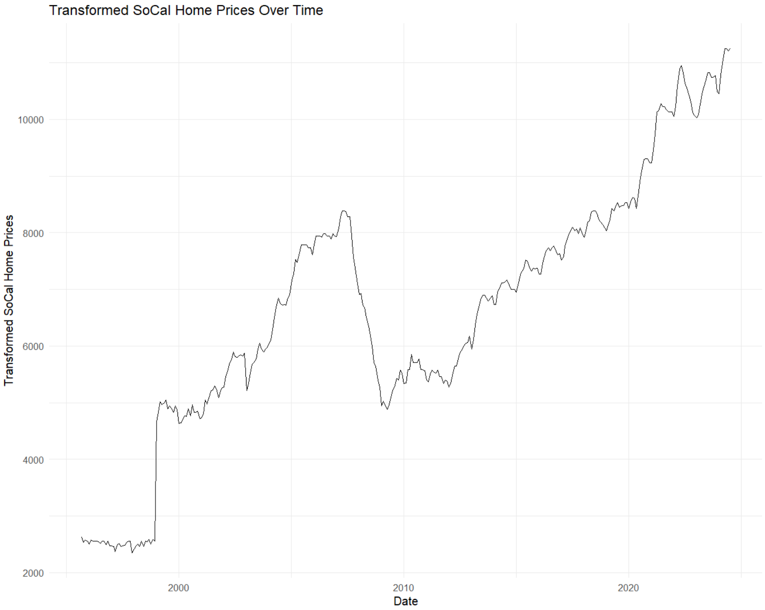
1. Long term trend: Since the 1990s, housing prices have shown an overall upward trend, especially after 2000, with a more significant growth rate. This indicates that the real estate market in Southern California has experienced significant growth over the past few decades.

2. The 2008 financial crisis: As shown in the graph, there was a significant decline around 2008. This is consistent with the real estate collapse during the global financial crisis.

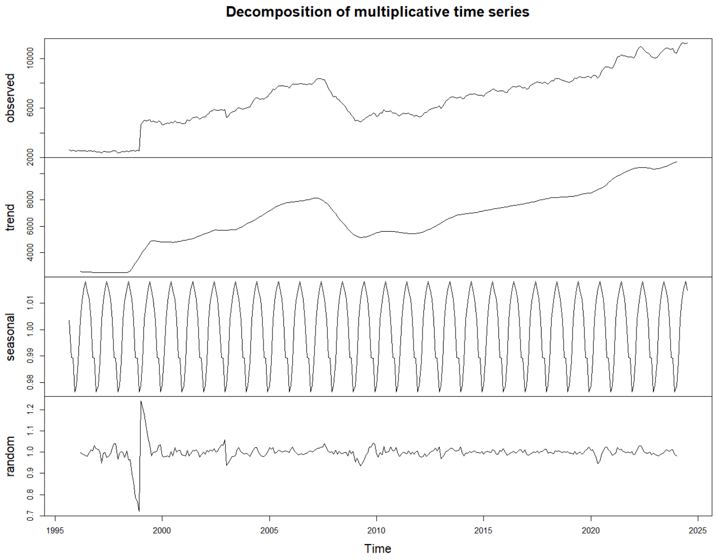
3. Recent increase: Since 2010, housing prices have gradually rebounded, especially after 2020, with growth accelerating. This phenomenon is consistent with the changes in the real estate market during the epidemic.

4. Volatility: Although the overall trend is upward, multiple fluctuations can be seen in the process of rising housing prices, indicating the seasonal behavior of the market.

Although significant fluctuations and changes can be observed in the time series of housing prices, it can be seen from the box plot that there are no obvious outliers in the data. This means that we did not find any individual data points deviating from the overall trend, so there is no need for outlier handling. However, due to significant instability of housing price, the variance of the data shows an increasing trend over time. In order to stabilize the variance and make it more suitable for times series modeling, we used Box Cox transformation and reduced the variance of the data and make it more symmetrical to better capturing trends and seasonal patterns in the time series.

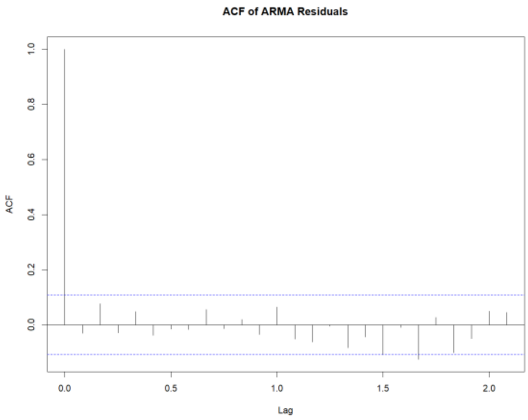


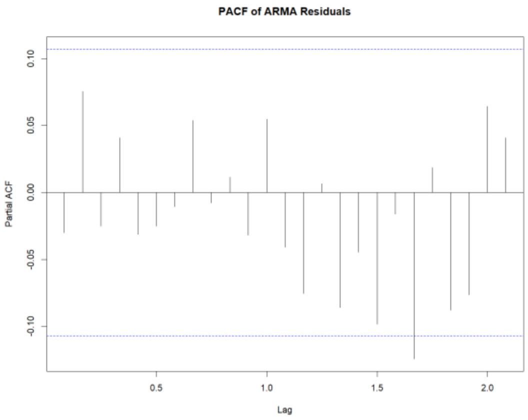
Data Analysis



We first decomposed the time series data into three parts: trend, seasonality, and random. As shown in the figure, we can clearly see the changing trend of each part through decomposition. The trend section presents a long-term upward trend, while the seasonal component shows a regular periodic fluctuation, and the random section contains irregular fluctuations.

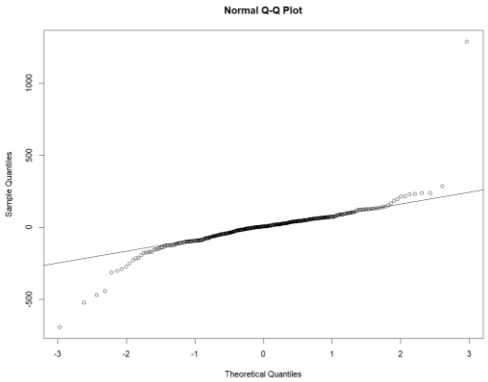
Next, we fitted an ARMA model (Autoregressive Moving Average model) to capture the dependency structure in the data. We conducted several diagnostic tests on the residuals of the ARMA model to confirm that residuals satisfy the assumptions of stationary time series, white noise, and normal distribution.





ACF plot results showed that there was no significant autocorrelation in the residuals, which indicates that the fitted model successfully captured the dependency structure of the sequence.

PACF plot further supports the conclusion that the residuals are white noise.

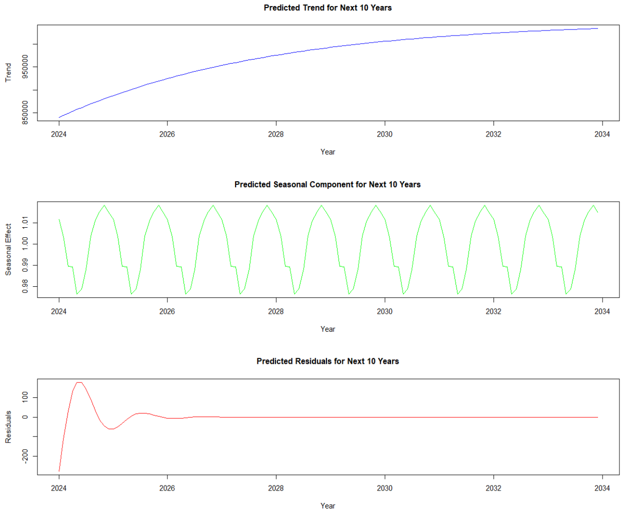


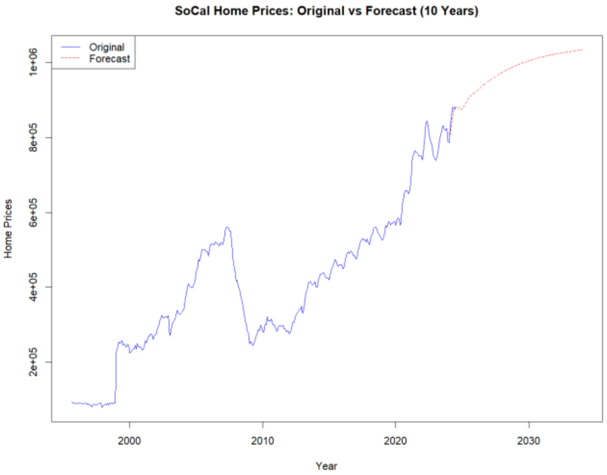
Q-Q plot shows that the majority of the points are arranged along a straight line with a small number of points slightly deviate from the line, which indicate that the residuals are approximately normal distributed.

Finally, we conducted Box Ljung test and the results showed a p-value of 0.5805, which cannot reject the hypothesis that the residuals are white noise.

In summary, the fitting and diagnostic results of the ARMA model shows that the residual sequence meets the requirements of white noise and normal distribution, which established the precondition for predictions.

Then we conducted the last stage of data analysis, prediction. We first use the Box Cox inverse transform to convert the data back to its original scale. This ensures that the prediction is meaningful in the context of housing prices. Then we forecasted the trend, seasonal, and residual components for the next 10 years and successfully separated the smooth components (trends and seasonality) and rough components (residuals) of the data, then we plot the final forecast of the home prices in the next 10 years. The result is reflected in the plots and data below.





Mon\Yr	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033
1	808152.4	880791.5	924343.2	953674.3	975758.1	992777.2	1005890.0	1015972.9	1023716.7	1029659.3
2	831931.4	885666.5	926812.9	955662.9	977342.0	994005.1	1006834.6	1016698.3	1024273.2	1030086.0
3	853189.5	891228.4	929449.7	957634.0	978890.8	995205.0	1007757.7	1017407.0	1024816.8	1030502.6
4	869271.0	896860.7	932193.9	959593.5	980409.8	996380.3	1008662.5	1018102.0	1025350.2	1030911.8
5	879179.1	902077.1	934965.8	961535.3	981897.8	997528.7	1009545.9	1018780.3	1025870.4	1031310.5
6	883378.9	906590.9	937696.8	963455.9	983359.0	998654.3	1010412.0	1019445.7	1026381.	1031702.3
7	883321.7	910310.2	940331.1	965345.9	984792.7	999756.8	1011260.1	1020097.5	1026881.6	1032086.4
8	880896.2	913309.1	942837.5	967197.4	986198.4	1000836.7	1012090.7	1020735.8	1027372.0	1032462.9
9	877932.0	915767.3	945206.3	969003.3	987573.8	1001892.6	1012902.3	1021359.3	1027850.6	1032830.2
10	875858.6	917911.3	947449.5	970761.3	988919.4	1002925.5	1013695.9	1021968.8	1028318.4	1033189.1
11	875536.2	919956.2	949590.6	972471.6	990235.1	1003935.9	1014472.2	1022564.8	1028775.8	1033540.0
12	877245.3	922067.2	951656.6	974135.8	991520.6	1004923.7	1015230.8	1023147.1	1029222.5	1033882.4

Based on the final forecast, housing prices are expected to steadily rise over the next 10 years, with seasonal fluctuations recurring as anticipated.

From 2024 to 2033, housing prices show a stable growth trend, gradually rising from $800000 to $1.03 million. The predicted increase is relatively stable, without any significant fluctuations or sudden drops. Despite the overall upward trend in housing prices, there are some seasonal fluctuations for example from June 2024 to November 2024, there was a slight decline in housing prices and followed by a recovery in prices at December 2024. By the end of 2033, the predicted housing price will exceed 1.03 million dollars, which is an increase of approximately 230,000 dollars compared to 800,000 dollars at the beginning of 2024.

Discussion

Based on our prediction results, we can answer some practical questions, such as “When is the best time to buy a house?” and “How much is the expected appreciation after buying a house?” Which are important questions for people who want to invest or purchase property.

Based on our result, buying a house as early as possible is clearly a good choice, since house prices are expected to significantly increase in the coming years. Specifically, if you buy a house in early 2024, you may save about $70000 compared to buying a house in early 2025.

If you purchase a property at a price of approximately $808,000 in early 2024, according to forecast data, the house price will reach around $1.01 million by 2030, with an appreciation of approximately $220000 and an average annual appreciation rate of about 4%. This expected increase in value is ideal for buyers.

Seasonal fluctuations may affect homebuyers' decisions in the short term i.e. home prices tend to decrease in the second half of the year. But this short-term price fluctuation is not significant in long-term investments.

However, there are some shortcomings that need to be noted in my analysis process.

Firstly, although I conducted a normality test and the results showed residuals were close to a normal distribution, it did not fully conform to normality. Fortunately, residuals are approximately normal distributed and it met the requirements of most statistical inferences.

Secondly, the predicted results indicate that without social emergencies such as the 2008 financial crisis or COVID-19 pandemic, housing prices in Southern California will continue to increase. However, in reality housing prices are influenced by many complex factors, including economics, politics, natural disasters, and so on. Thus, my predictions are very limited under such uncertainty. If significant economic or social events occur, the actual trend of housing prices may differ significantly from the predicted values. Due to time and technological limitations, my predictive model cannot take into account more complex factors.

Conclusion

In this analysis and report, we predicted the housing price in Southern California over the next 10 years through time series analysis. The results show that housing prices will steadily rise, expected to increase from 808000 in 2024 to around 1.01 million in 2030, with an appreciation of approximately 220000. Buying a house as early as possible can help avoid high housing prices in the future and obtain better appreciation returns.

However, the predictions still have limitations due to its inability to fully capture external influences. Overall, this analysis provides data support for home purchases, but actual decisions still might need to consider broader economic and policy factors.

References:

https://www.car.org/marketdata/data/housingdata

https://lassho.edu.vn/incredible-compilation-of-4k-house-images-over-999-exquisite-selections/

Appendix:

```
library(readxl)
library(ggplot2)
library(lubridate)
library(zoo)
library(dplyr)
library(forecast)
library(tseries)
data<-
read_excel("E:/Study/STA137/Project/MedianPricesofExistingDetachedHomesHistoricalData.xlsx",skip = 7)
#no data before line 69
data <- data %>%
  slice(69:n()) %>%
  select(`Mon-Yr`, SoCal) #'Mon-Yr''SoCal' col

head(data)

##############################################################1

# make $### to number
data$SoCal <- as.numeric(data$SoCal)

# origin plot
ggplot(data, aes(x = `Mon-Yr`, y = SoCal)) +
  geom_line(color = "blue") +
  labs(title = "SoCal Home Prices Over Time",
       x = "Date",
       y = "SoCal Home Prices") +
  theme_minimal()

# boxplot check for outliers
ggplot(data, aes(x = 1, y = SoCal)) +
  geom_boxplot() +
  labs(title = "Box Plot of SoCal Home Prices", y = "SoCal Home Prices") +
  theme_minimal() +
  theme(axis.title.x = element_blank(), axis.text.x = element_blank(), axis.ticks.x = element_blank())

#clean NA
socalClean <- na.omit(data)


# boxcox transformation
```

```
lamda <- BoxCox.lambda(data$SoCal, method = "loglik")
dataTransform <- BoxCox(data$SoCal, lamda)
print(lamda)

#plot transformed data
ggplot(data, aes(x = `Mon-Yr`, y = dataTransform)) +
  geom_line() +
  labs(title = "Transformed SoCal Home Prices Over Time", x = "Date", y = "Transformed SoCal Home
Prices") +
  theme_minimal()



# Convert transformed data to time series
socal_ts <- ts(dataTransform, start = c(1995, 9), frequency = 12)

# Decompose time series
dataDecomp <- decompose(socal_ts, type = "multiplicative")
plot(dataDecomp)

# Extract trend and seasonal components
trend <- dataDecomp$trend
season <- dataDecomp$seasonal


# Calculate residuals and remove NA values
residuals <- socal_ts - trend - season
residualsClean <- na.omit(residuals)

# ACF and PACF for residuals
acf(residualsClean, main = "ACF of Residuals")
pacf(residualsClean, main = "PACF of Residuals")

# Ljung-Box test for whiteness
Box.test(residualsClean, type = "Ljung-Box")

# QQ plot and Shapiro-Wilk test for normality
qqnorm(residualsClean)
qqline(residualsClean)
shapiTest <- shapiro.test(residualsClean)
print(shapiTest)

#####################################################analyze 4
# Fit ARMA model
```

```
armaModel <- auto.arima(residualsClean, max.p = 5, max.q = 5, stationary = TRUE, seasonal = FALSE)
summary(armaModel)

# Check ACF and PACF
armaResiduals <- residuals(armaModel)

acf(armaResiduals, main = "ACF of ARMA Residuals")
pacf(armaResiduals, main = "PACF of ARMA Residuals")

# Ljung-Box test for checking white noise
Box.test(armaResiduals, type = "Ljung-Box")

# Shapiro-Wilk test for normality
shapiTest2 <- shapiro.test(armaResiduals)
print(shapiTest2)

# QQ plot for normality
qqnorm(armaResiduals)
qqline(armaResiduals)

###################################################################analyze 5

#predict
trendClean <- na.omit(trend)

# Predict the future trend 10 years
trendForecast <- forecast(trendClean, h = 120)

#Predict the seasonal component
seasonForecast <- rep(tail(season, 12), 10)#12 month repeat 10 times

# predict the rough component
armaModel <- auto.arima(residualsClean)   #fit arma for cleaned residuals
armaForecast <- forecast(armaModel, h = 120)   # 10 year forcast

# Combine smooth and rough forecasts
smoothForecast <- trendForecast$mean + seasonForecast
finalForecast <- smoothForecast + armaForecast$mean



#original scale
actFinalForecast <- InvBoxCox(finalForecast, lamda)

actTrendForecast <- InvBoxCox(trendForecast$mean, lamda)
```

```
actual_socal_ts <- InvBoxCox(socal_ts, lamda)

#Plot forecasted with original
plot(actFinalForecast, main = "Forecast of SoCal Home Prices (Original Scale)",
    xlab = "Date", ylab = "SoCal Home Prices", col = "blue", type = "l")
summary(season)


# original
lines(actual_socal_ts, col = "black")

legend("topright", legend = c("Original", "Forecast"), col = c("black", "blue"), lty = 1)

ts.plot(actual_socal_ts, actTrendForecast, col = c("blue", "red"), lty = 1:2)
legend("topright", legend = c("Original", "Forecast"), col = c("blue", "red"), lty = 1:2)


length(seasonForecast)

timeSeq <- seq(2024, 2024 + (120 - 1) / 12, by = 1/12)

length(timeSeq)


par(mfrow = c(3, 1))


plot(timeSeq, actTrendForecast , type = "l",
    main = "Predicted Trend for Next 10 Years",
    xlab = "Year", ylab = "Trend", col = "blue")

plot(timeSeq, seasonForecast, type = "l",
    main = "Predicted Seasonal Component for Next 10 Years",
    xlab = "Year", ylab = "Seasonal Effect", col = "green")

plot(timeSeq, armaForecast$mean, type = "l",
    main = "Predicted Residuals for Next 10 Years",
    xlab = "Year", ylab = "Residuals", col = "red")

par(mfrow = c(1, 1))

ts.plot(actual_socal_ts, actFinalForecast, col = c("blue", "red"), lty = 1:2,
      main = "SoCal Home Prices: Original vs Forecast (10 Years)",
    xlab = "Year", ylab = "Home Prices")
legend("topleft", legend = c("Original", "Forecast"), col = c("blue", "red"), lty = 1:2)


# original scale
actFinalForecast <- InvBoxCox(finalForecast, lamda)
#print actual value
forecast_years <- seq(2024, 2024 + (length(actFinalForecast) - 1) / 12, by = 1/12)
forecast_df <- data.frame(Year = forecast_years, Predicted_Home_Prices = actFinalForecast)
print(forecast_df)
```