California House Price Forecasting: ARIMA, ETS, Holt-Winters, NNETAR, Prophet, and Combined Models

Kuanyi Chen

2025-03-11

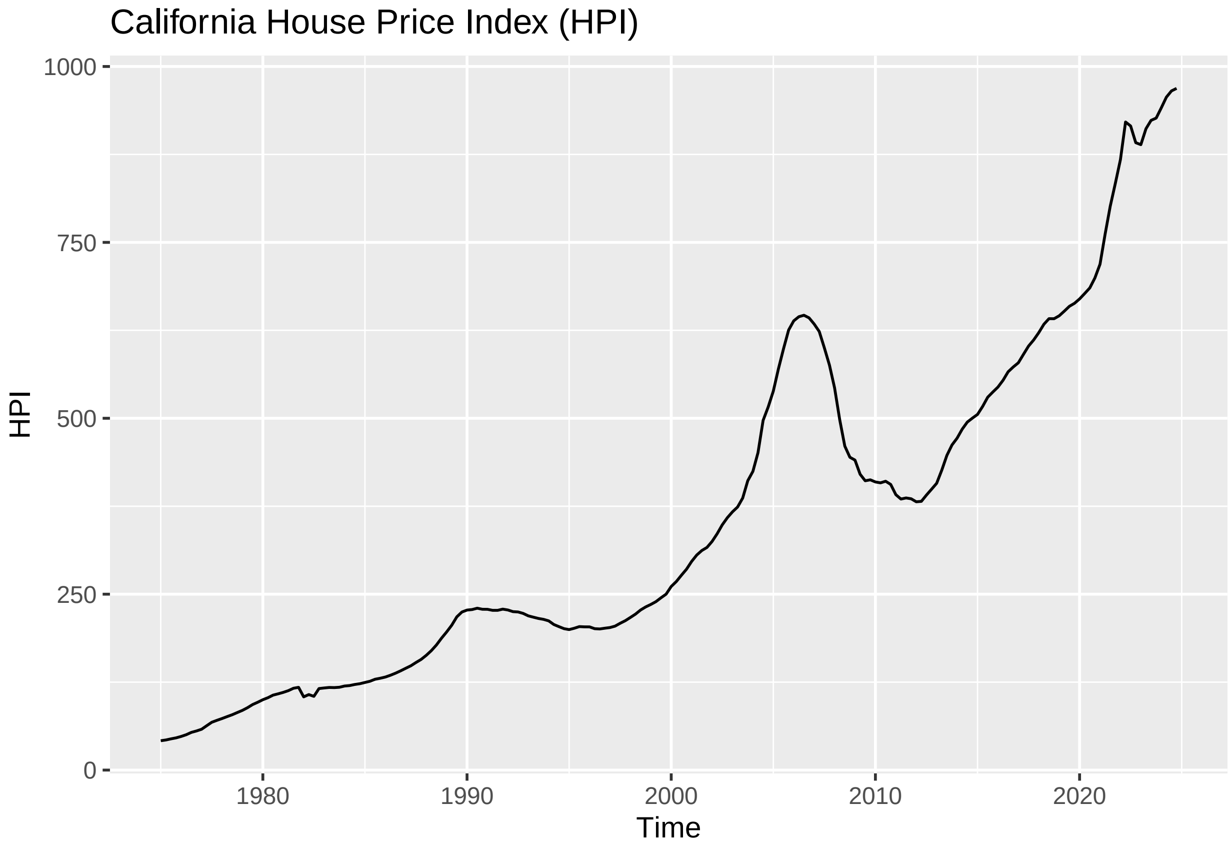
I. Introduction

```
# import and extract data
us_house_price <- read.csv("https://www.fhfa.gov/hpi/download/quarterly_datasets/hpi_at_state.csv")
names(us_house_price) <- c("State", "Year", "Quarter", "Price")
ca_price <- filter(us_house_price, State == "CA")

# create time series
ca_price_ts <- ts(ca_price$Price, start = c(1975, 1), frequency = 4)
```

The time series is the quarterly house price index (HPI) of California single-family houses from 1975 Q1 to 2024 Q2. HPI is a metric that measures changes in the prices of houses, with a certain time set as the base period with index value equal 100. In this data, the base period is 1980 Q1. We will fit different models to the data, including ARIMA, ETS, Holt-Winters, NNETAR, Prophet, and forecasting combination. We will perform different diagnostic tests on the models, including residuals vs. fitted plot, ACF and PACF of residuals, Ljung-Box test of residuals, and CUSUM plot of residuals. We will also identify a preferred model based on the training and testing errors.

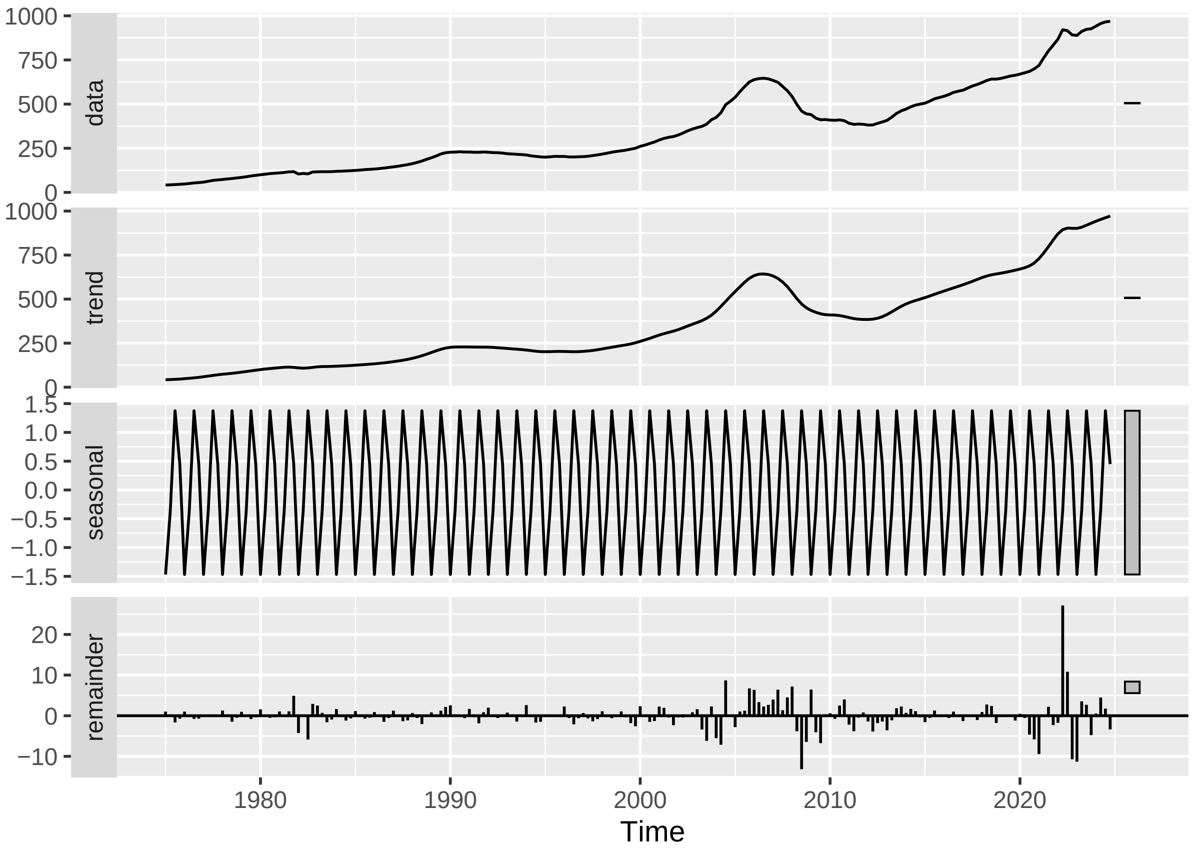
II. Results

```
# plot the time series
autoplot(ca_price_ts) +
  ggtitle("California House Price Index (HPI)") +
  ylab("HPI")
```



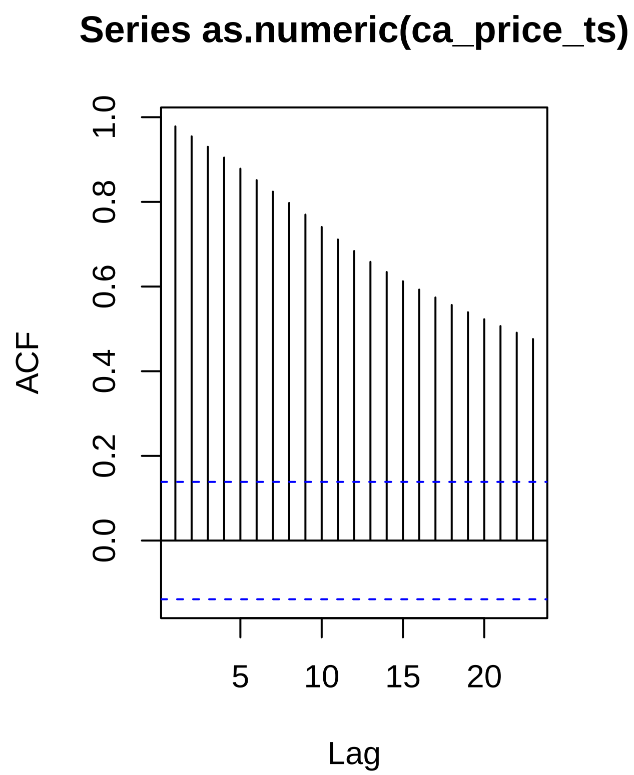
The time series has an overall positive growing trend. There is an irregular cycle pattern, with a sudden spike aroun 2006 and 2007. There seems to be small seasonality, which is not clearly visible in the overall plot.

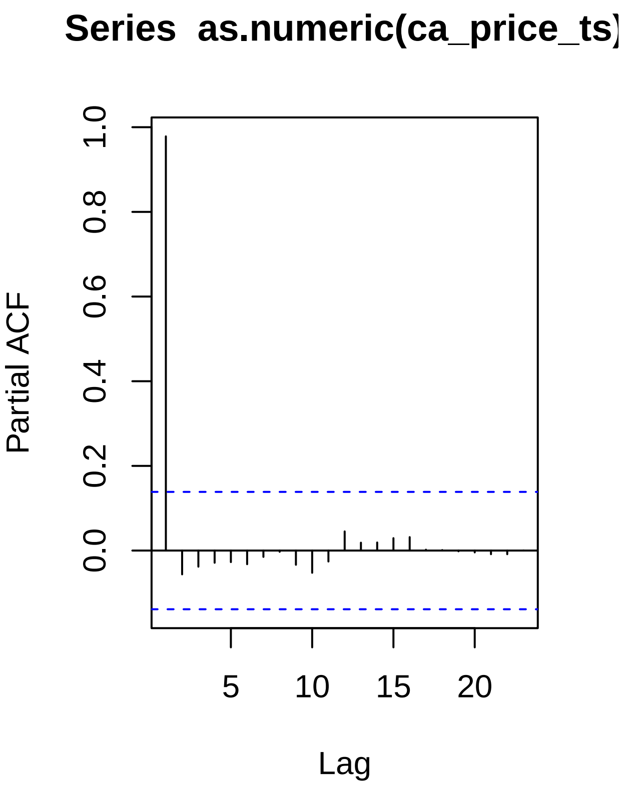
```
# stl decomposition of time series
autoplot(stl(ca_price_ts, s.window = "periodic"))
```



The decomposed graph showcases a positive trend. It also showcases the presence of regular seasonality. There is also presence of some irregular noise.

```r
# plot acf and pacf
par(mfrow = c(1, 2))
acf(as.numeric(ca_price_ts))
pacf(as.numeric(ca_price_ts))
```





The ACF plot indicates strong correlation in the data, and the PACF indicates significance to the first lag to incorporate in an AR model.

Split Training and Testing Data

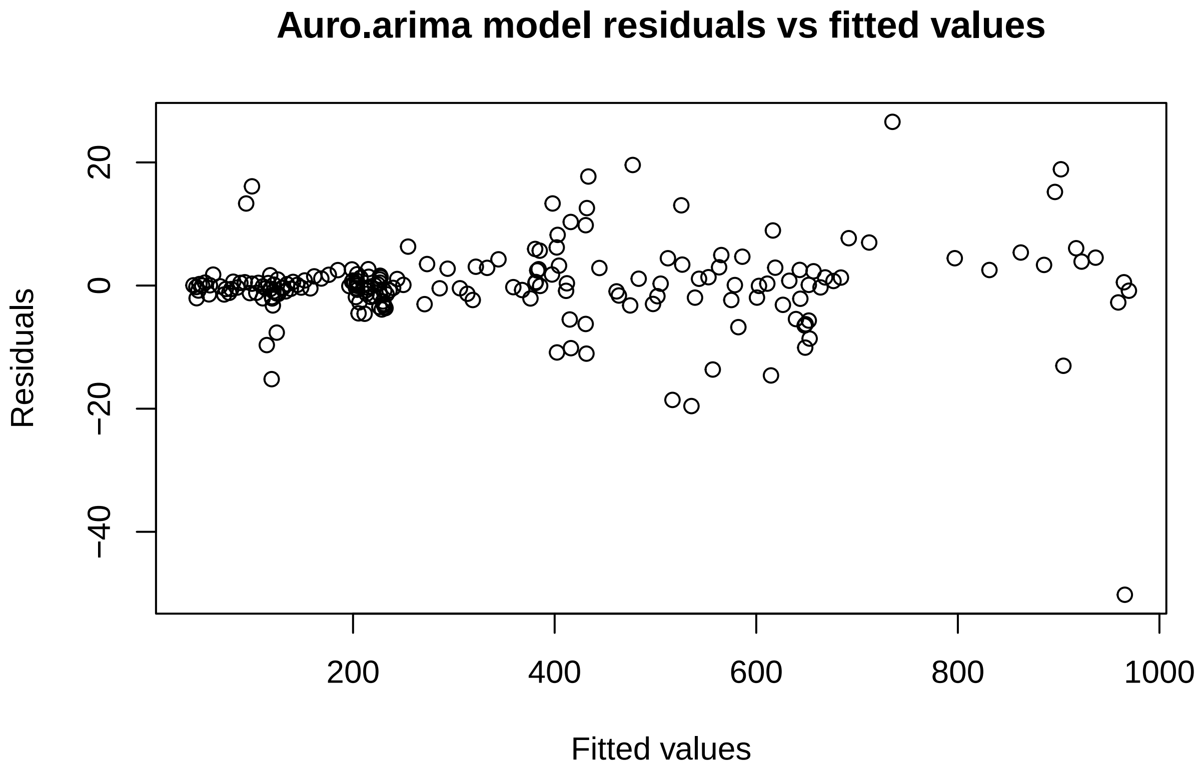
We split the data into 70% training and 30% testing, which is the standard train test split proportion.

```
# 70% training data, 30% testing data
ca_price_train <- subset(ca_price_ts, end = length(ca_price_ts) * 0.7)

ca_price_test <- subset(ca_price_ts, start = length(ca_price_ts) * 0.7 + 1)
```

1. ARIMA

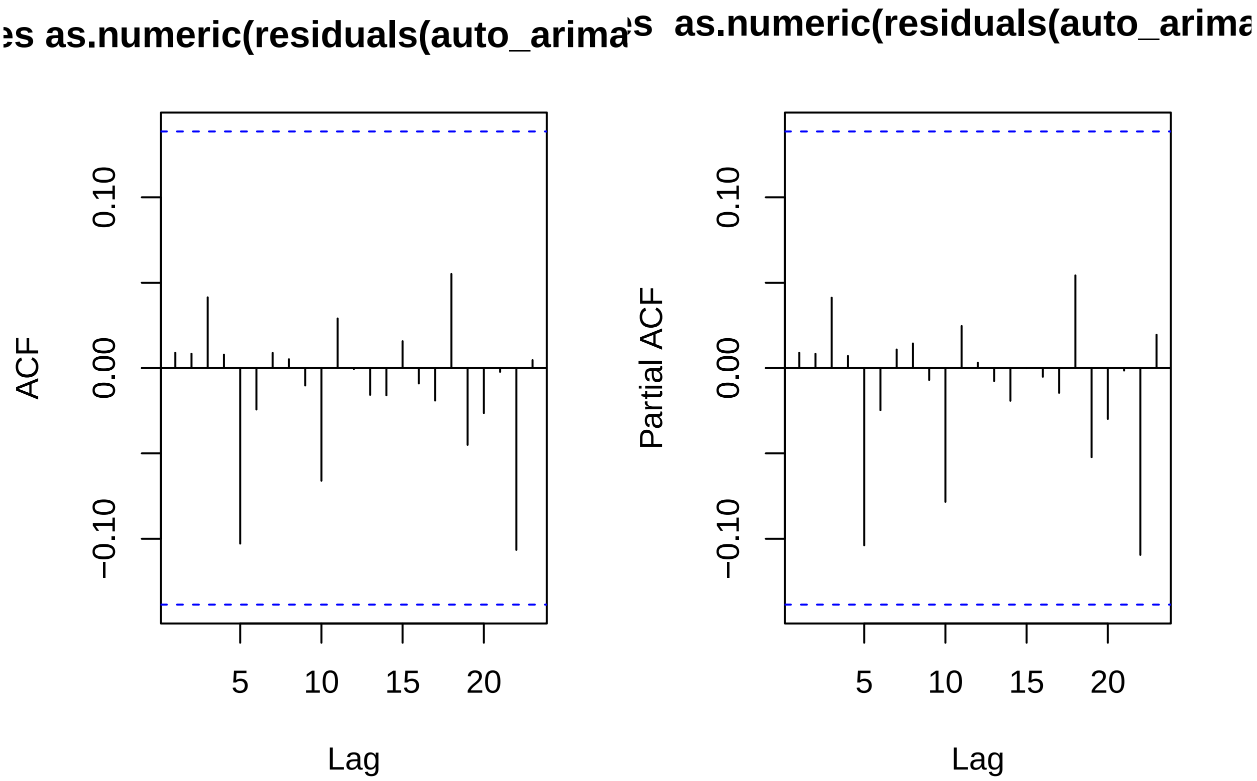
```
# auto.arima
auto_arima_model <- auto.arima(ca_price_ts)
```

```
# plot residuals vs fitted values
plot(fitted(auto_arima_model), residuals(auto_arima_model),
     main = "Auro.arima model residuals vs fitted values",
     xlab = "Fitted values",
     ylab = "Residuals")
```



The residuals are generally evenly scattered around 0. They are more clustered in the beginning and become more spread out as the fitted values increase.

```r
# plot ACF and PACF of residuals
par(mfrow = c(1, 2))
acf(as.numeric(residuals(auto_arima_model)))
pacf(as.numeric(residuals(auto_arima_model)))
```



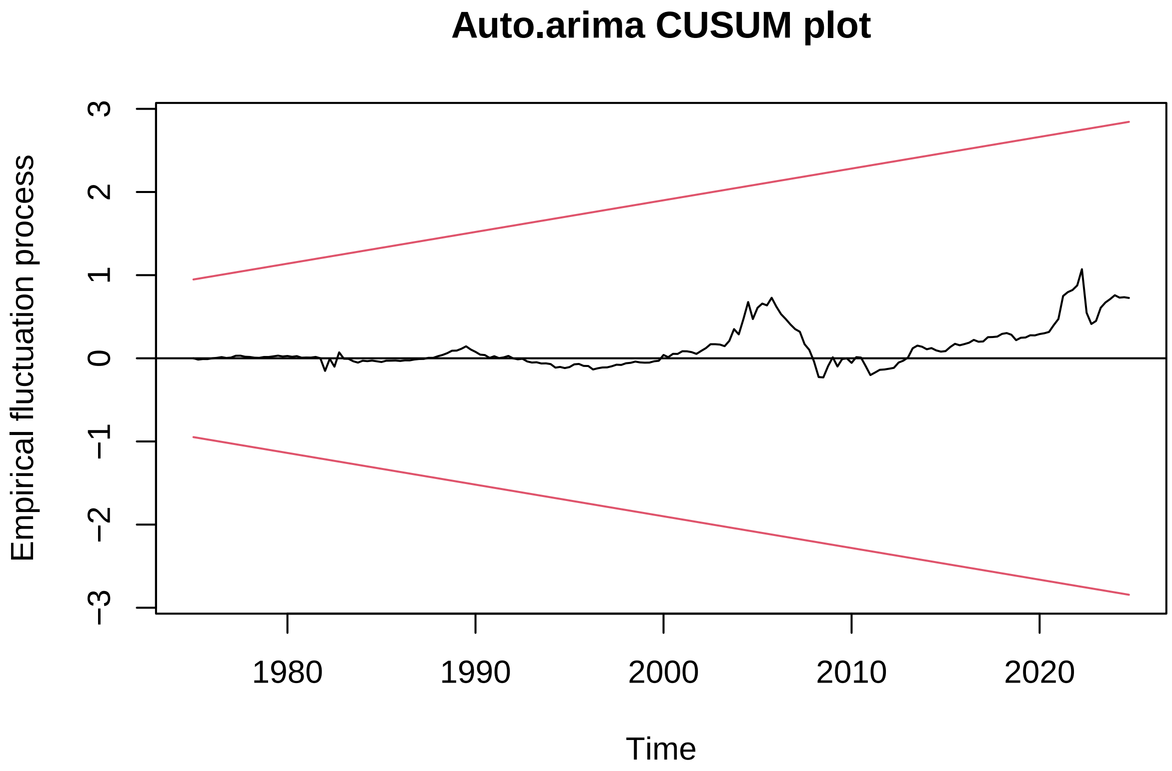
The residuals are insignificant at all the lags, indicating no pattern or autocorrelation. The residuals are white noise.

```r
# ljung-box test on residuals
Box.test(residuals(auto_arima_model))
```

```
##
##  Box-Pierce test
##
## data:  residuals(auto_arima_model)
## X-squared = 0.016155, df = 1, p-value = 0.8989
```

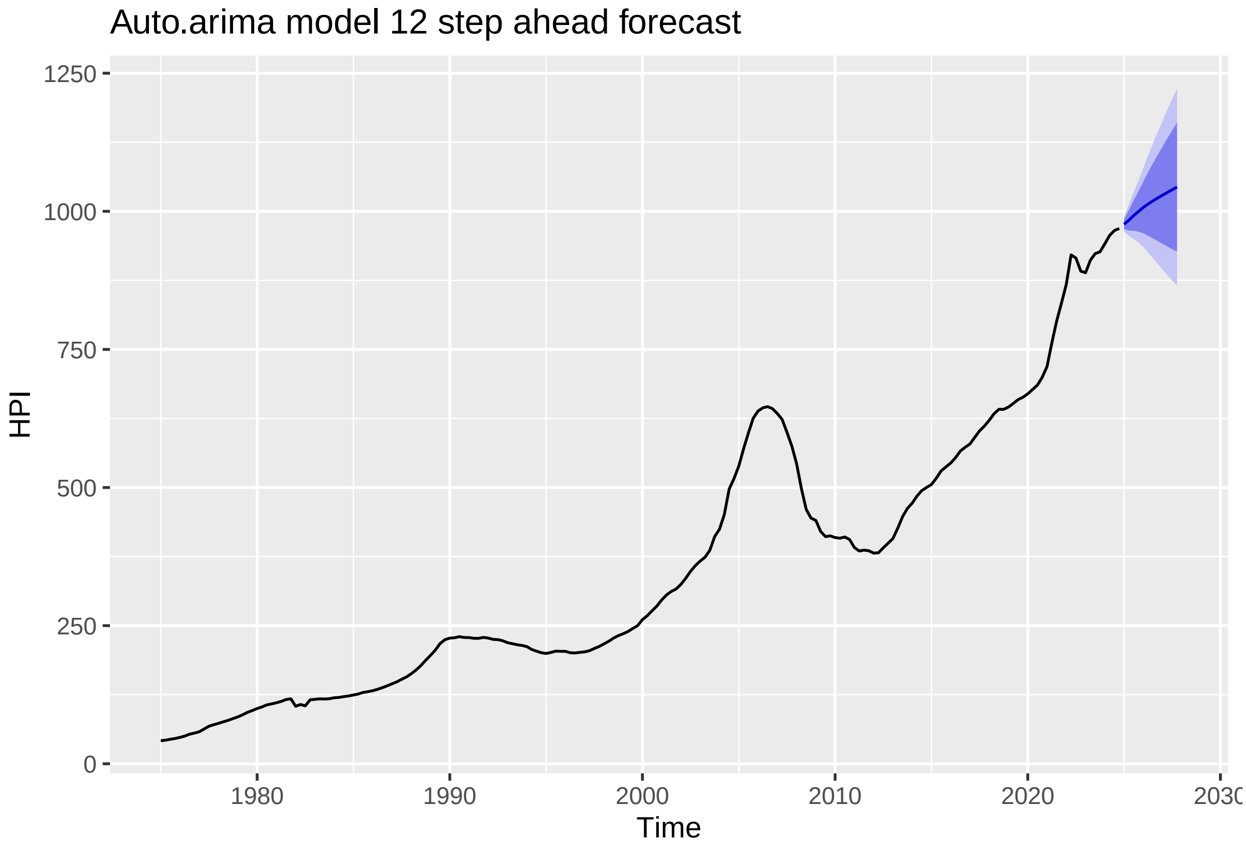
The p-value of the Ljung-box test is not significant at the 95% confidence level, so we fail to reject the null hypothesis and conclude that there is no autocorrelation in the residuals.

```r
# plot the CUSUM
plot(efp(residuals(auto_arima_model) ~ 1, type = "Rec-CUSUM"),
main = "Auto.arima CUSUM plot")
```



According to the CUSUM plot, the cumulative sum of residuals starts out with little fluctuation around 0, then the fluctuation starts to increase past 2000.

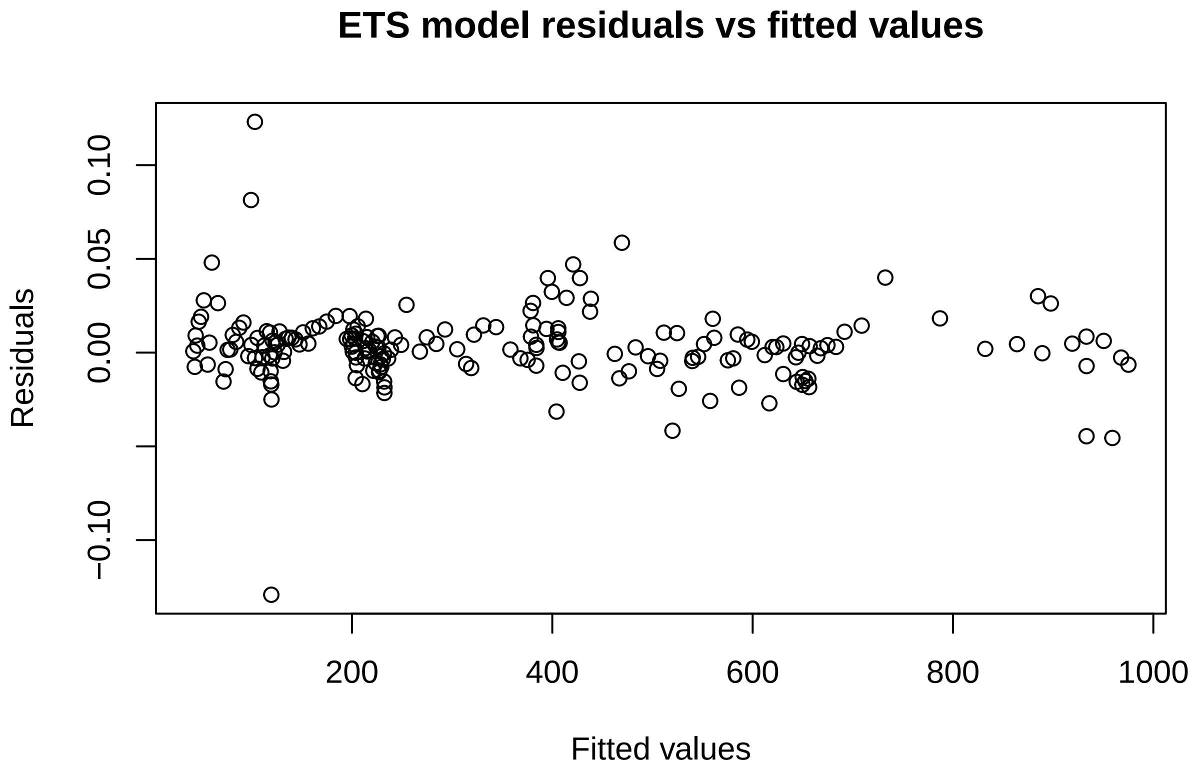
```
# plot the 12 quarter ahead forecast
autoplot(forecast(auto_arima_model, h = 12)) +
  ggtitle("Auto.arima model 12 step ahead forecast") +
  ylab("HPI")
```



2. ETS

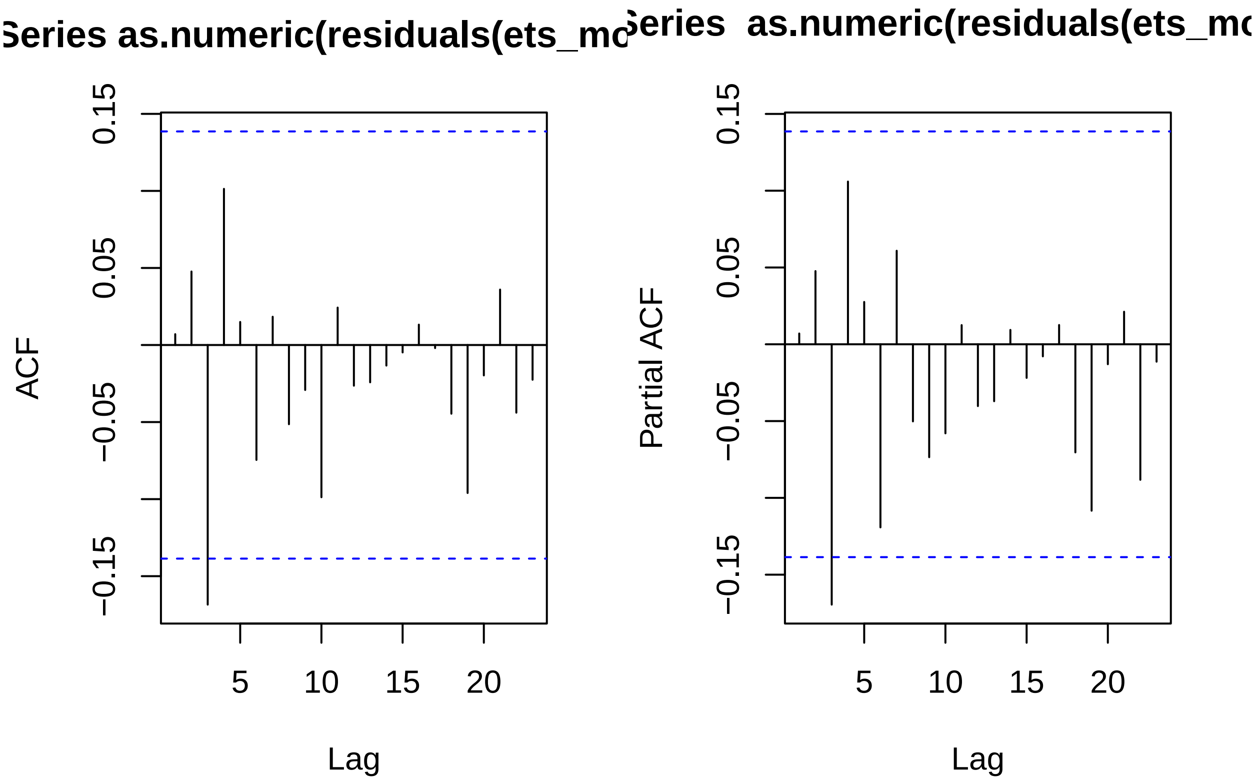
```r
# ets
ets_model <- ets(ca_price_ts)
```

```r
# plot residuals vs fitted values
plot(fitted(ets_model), residuals(ets_model),
     main = "ETS model residuals vs fitted values",
     xlab = "Fitted values",
     ylab = "Residuals")
```



The residuals are generally evenly scattered around 0. They are more clustered in the beginning and become more spread out as the fitted values increase.

```
# plot ACF and PACF of residuals
par(mfrow = c(1, 2))
acf(as.numeric(residuals(ets_model)))
pacf(as.numeric(residuals(ets_model)))
```



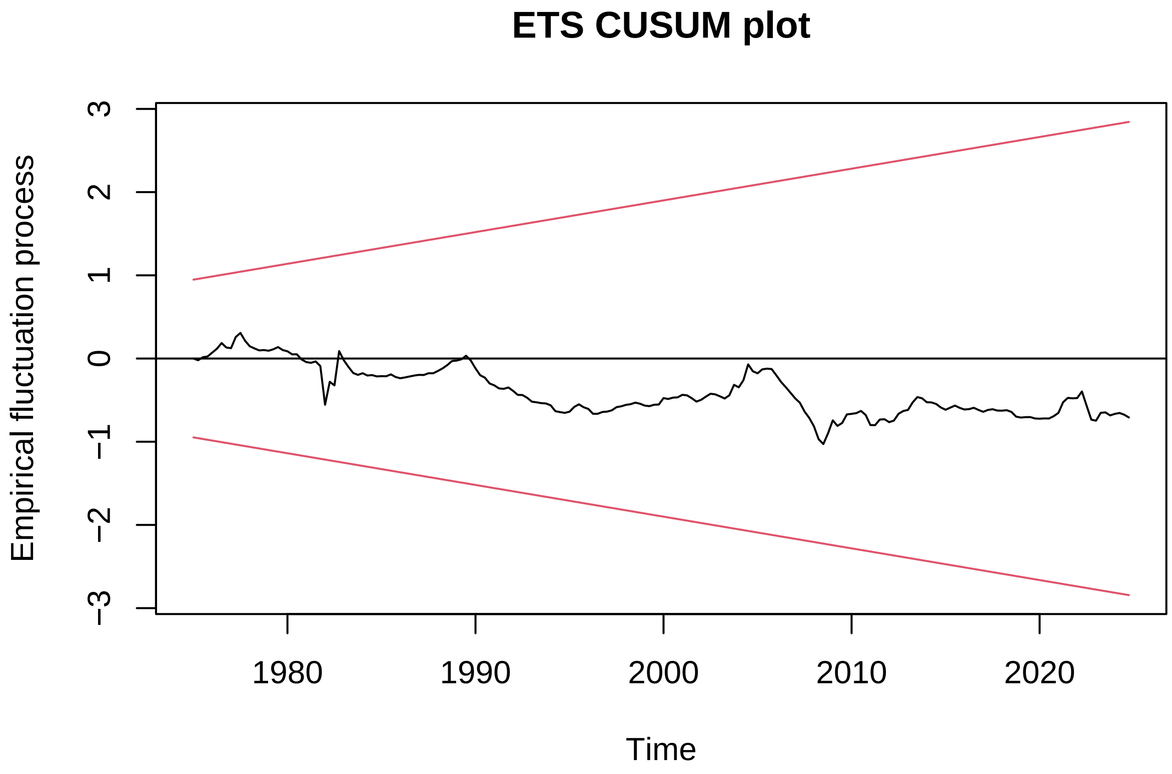
Most of the residuals are insignificant at the lags, except lag 3. So there is almost no pattern or autocorrelation, closely resembling white noise.

```
# ljung-box test on residuals
Box.test(residuals(ets_model))
```

```
##
##  Box-Pierce test
##
## data:  residuals(ets_model)
## X-squared = 0.0099016, df = 1, p-value = 0.9207
```

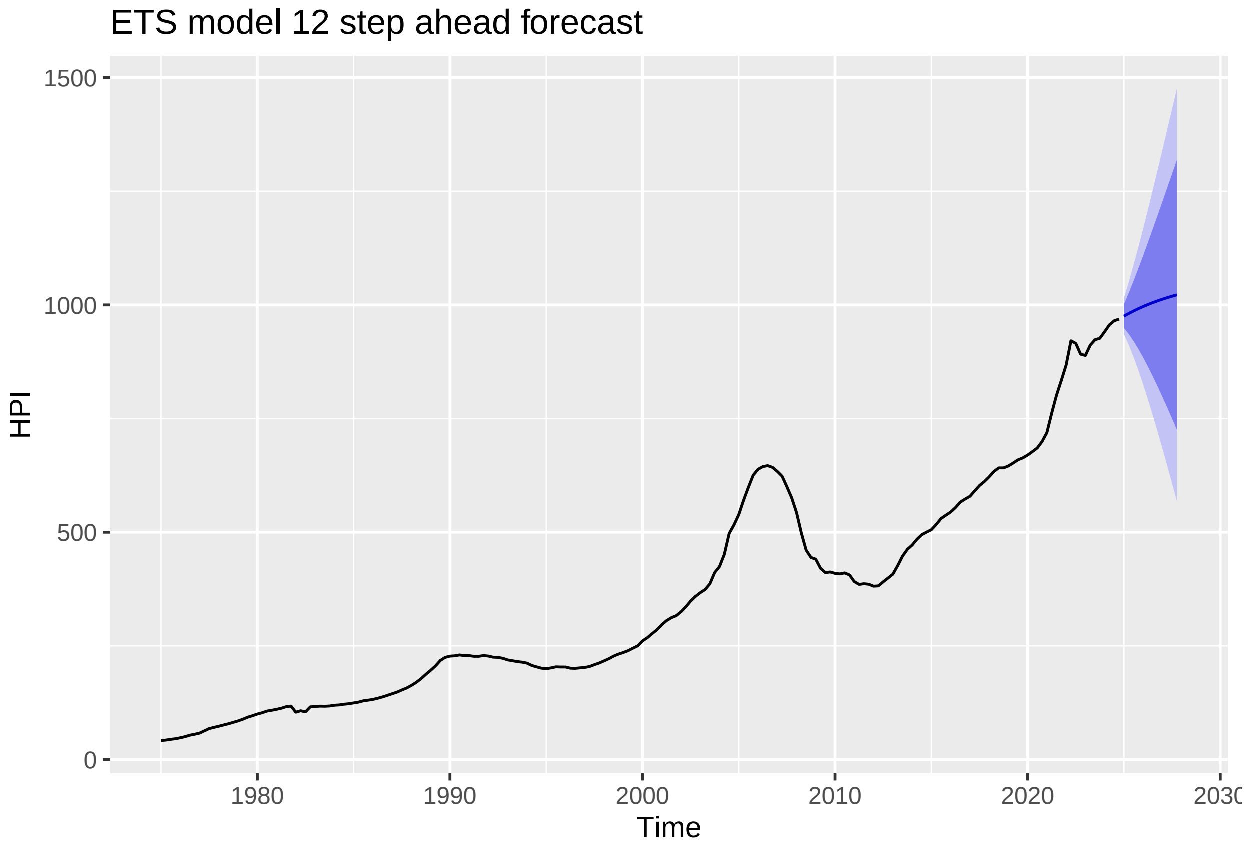
The p-value of the Ljung-box test is not significant at the 95% confidence level, so we fail to reject the null hypothesis and conclude that there is no autocorrelation in the residuals.

```
# plot the CUSUM
plot(efp(residuals(ets_model) ~ 1, type = "Rec-CUSUM"),
main = "ETS CUSUM plot")
```



According to the CUSUM plot, the cumulative sum of residuals starts out with little fluctuation around 0, but then soon deviates towards the negative direction.

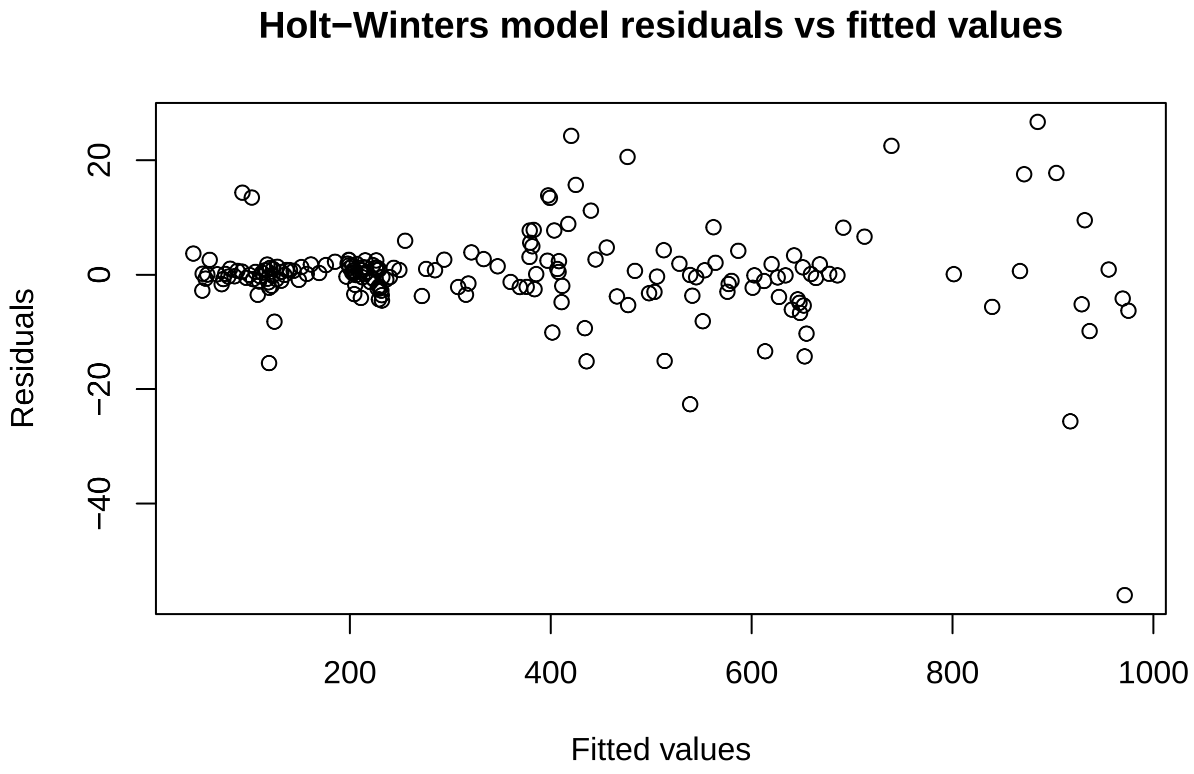
```
# plot the 12 quarter ahead forecast
autoplot(forecast(ets_model, h = 12)) +
  ggtitle("ETS model 12 step ahead forecast") +
  ylab("HPI")
```



3. Holt-Winters

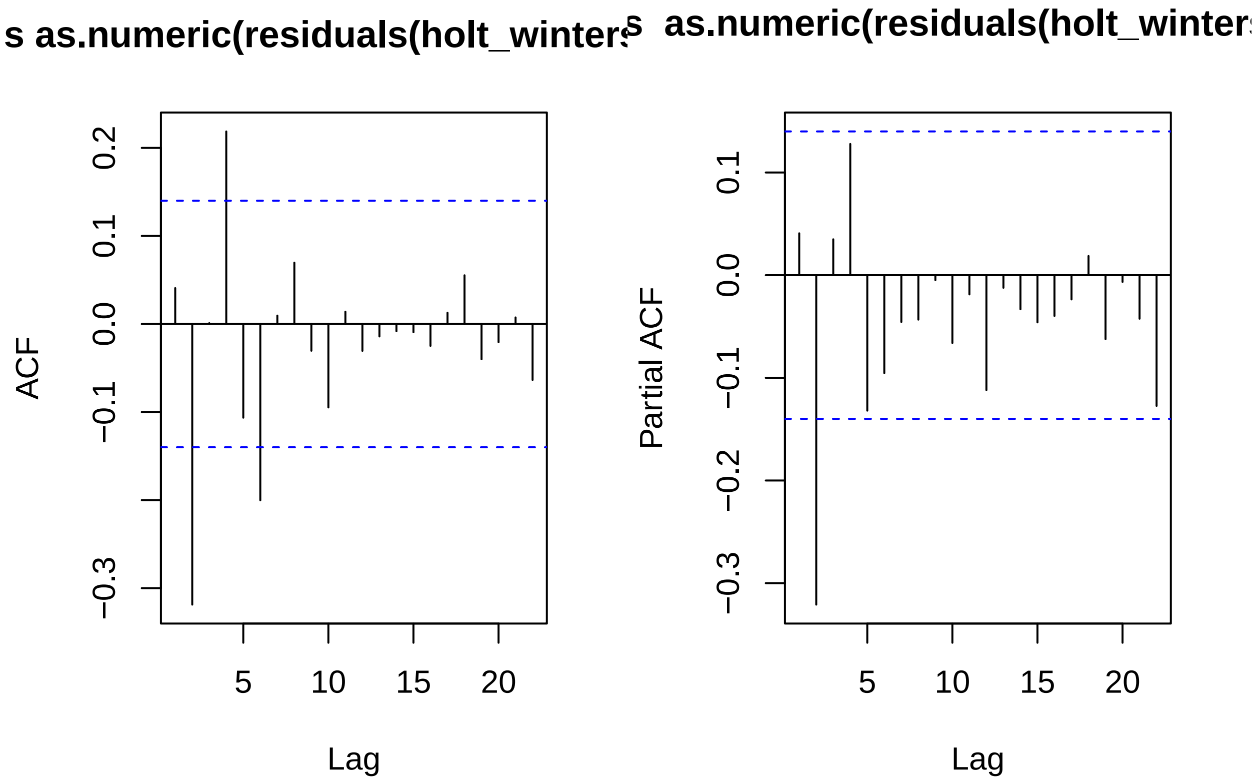
```r
# holt-winters method
holt_winters_model <- HoltWinters(ca_price_ts)
```

```r
# plot residuals vs fitted values
plot(fitted(holt_winters_model)[,1], residuals(holt_winters_model),
     main = "Holt-Winters model residuals vs fitted values",
     xlab = "Fitted values",
     ylab = "Residuals")
```



The residuals are generally evenly scattered around 0. They are more clustered in the beginning and become more spread out as the fitted values increase.

```
# plot ACF and PACF of residuals
par(mfrow = c(1, 2))
acf(as.numeric(residuals(holt_winters_model)))
pacf(as.numeric(residuals(holt_winters_model)))
```



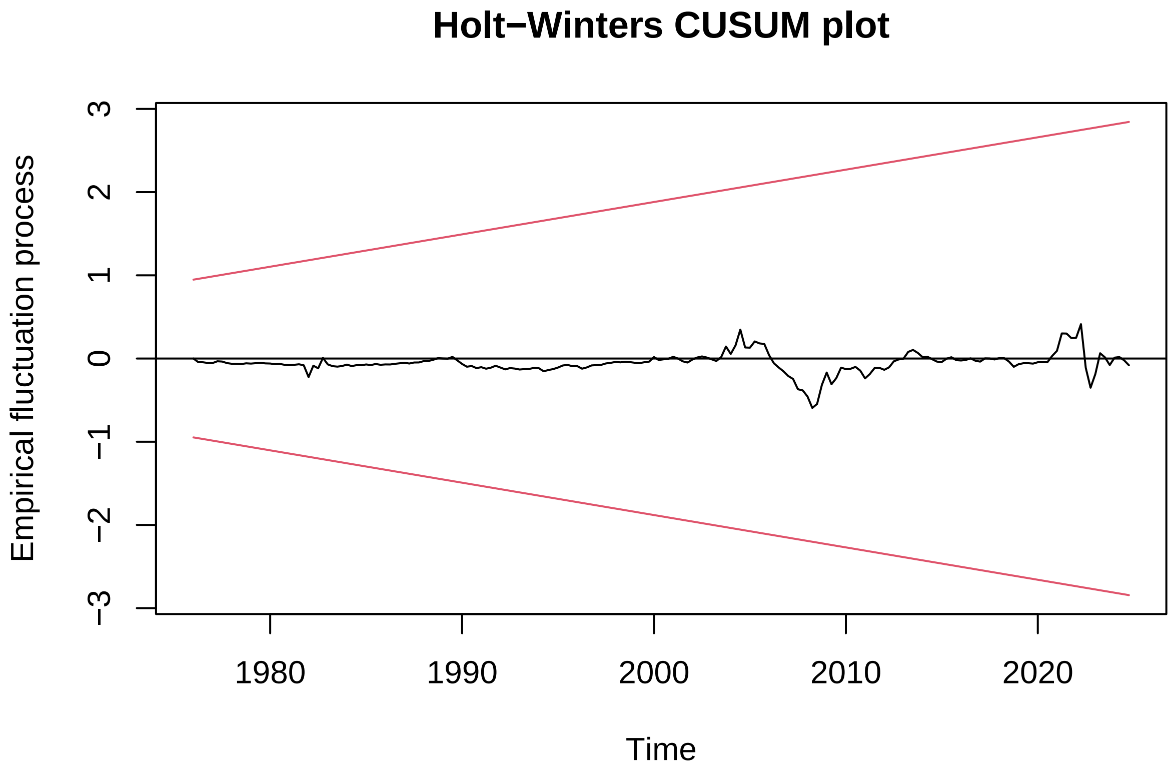
Most of the residuals are insignificant at the lags, except lag 2, 4, and 6. So there is almost no pattern or autocorrelation, closely resembling white noise.

```
# ljung-box test on residuals
Box.test(residuals(holt_winters_model))
```

```
##
##  Box-Pierce test
##
## data:  residuals(holt_winters_model)
## X-squared = 0.32636, df = 1, p-value = 0.5678
```

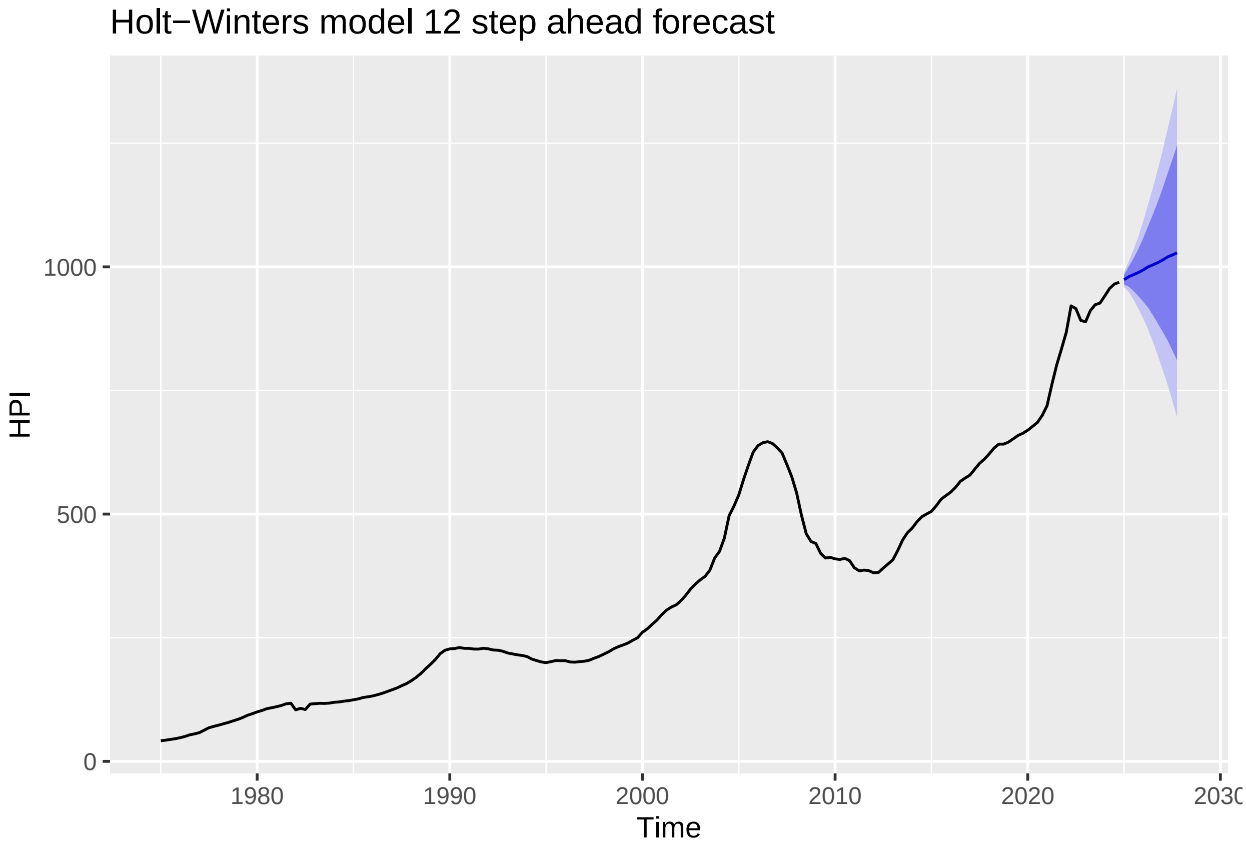
The p-value of the Ljung-box test is not significant at the 95% confidence level, so we fail to reject the null hypothesis and conclude that there is no autocorrelation in the residuals.

```
# plot the CUSUM
plot(efp(residuals(holt_winters_model) ~ 1, type = "Rec-CUSUM"),
main = "Holt-Winters CUSUM plot")
```



According to the CUSUM plot, the cumulative sum of residuals starts out with little fluctuation around 0, then has some minor fluctuations in the 2000s.

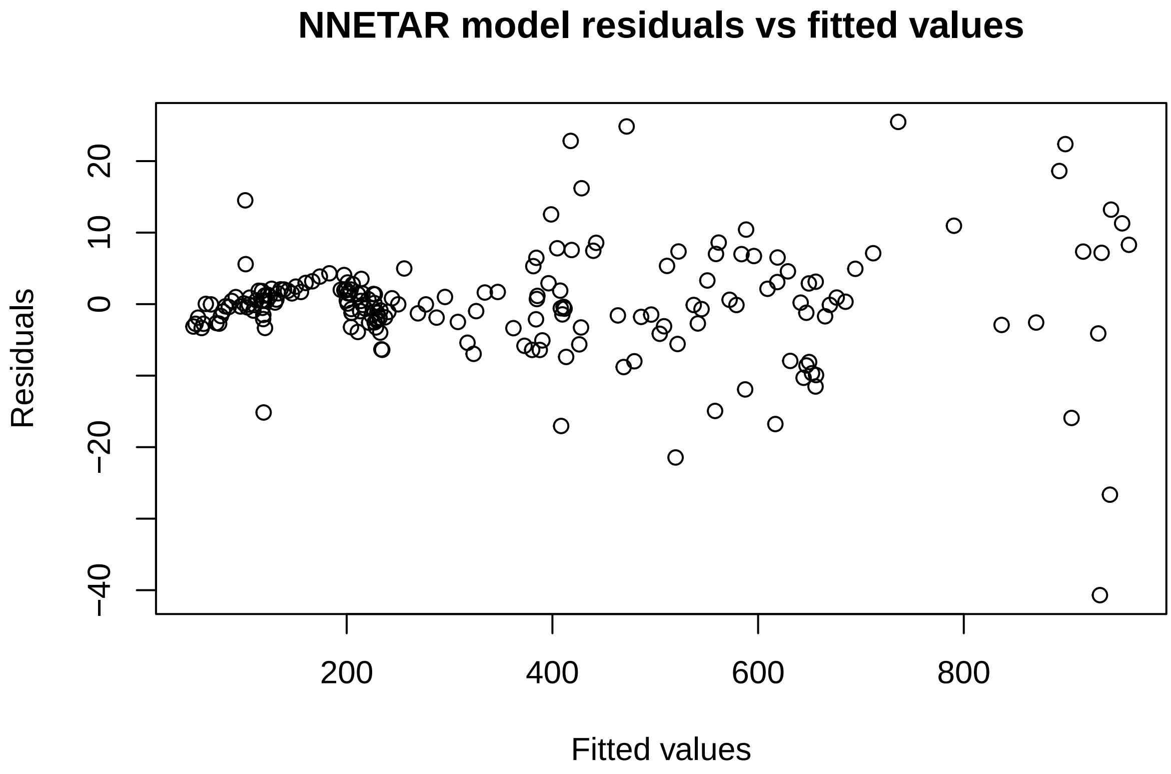
```r
# plot the 12 quarter ahead forecast
autoplot(forecast(holt_winters_model, h = 12)) +
  ggtitle("Holt-Winters model 12 step ahead forecast") +
  ylab("HPI")
```



4. NNETAR

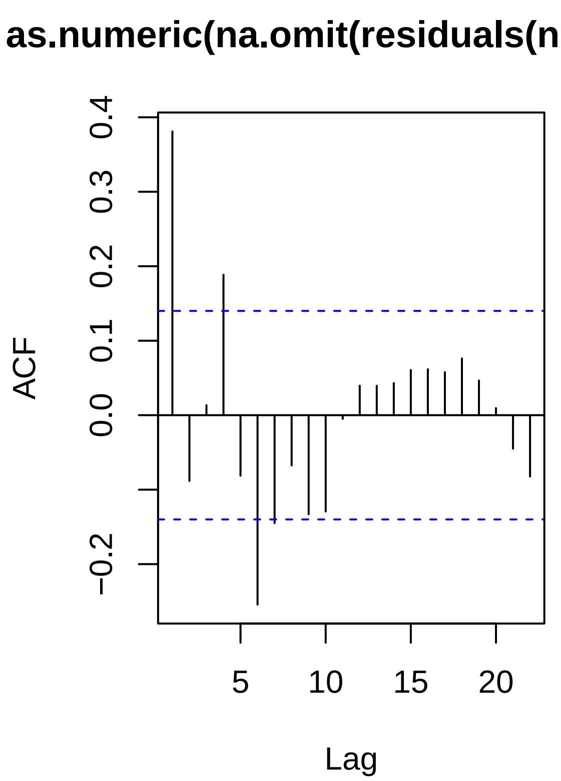
```r
# NNETAR method
nnetar_model <- nnetar(ca_price_ts)
```

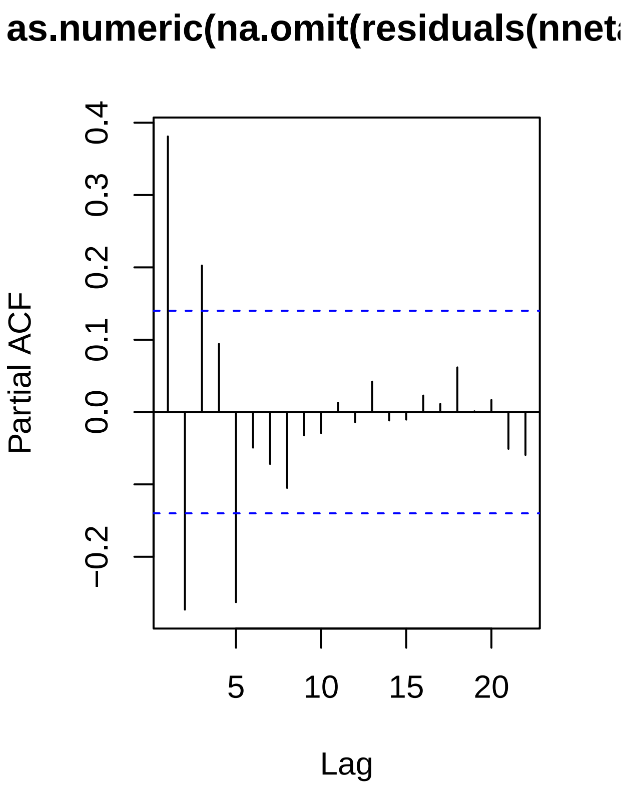
```r
# plot residuals vs fitted values
plot(fitted(nnetar_model), residuals(nnetar_model),
     main = "NNETAR model residuals vs fitted values",
     xlab = "Fitted values",
     ylab = "Residuals")
```



The residuals are generally evenly scattered around 0. They are more clustered in the beginning and become more spread out as the fitted values increase.

```
# plot ACF and PACF of residuals
par(mfrow = c(1, 2))
acf(as.numeric(na.omit(residuals(nnetar_model))))
pacf(as.numeric(na.omit(residuals(nnetar_model))))
```





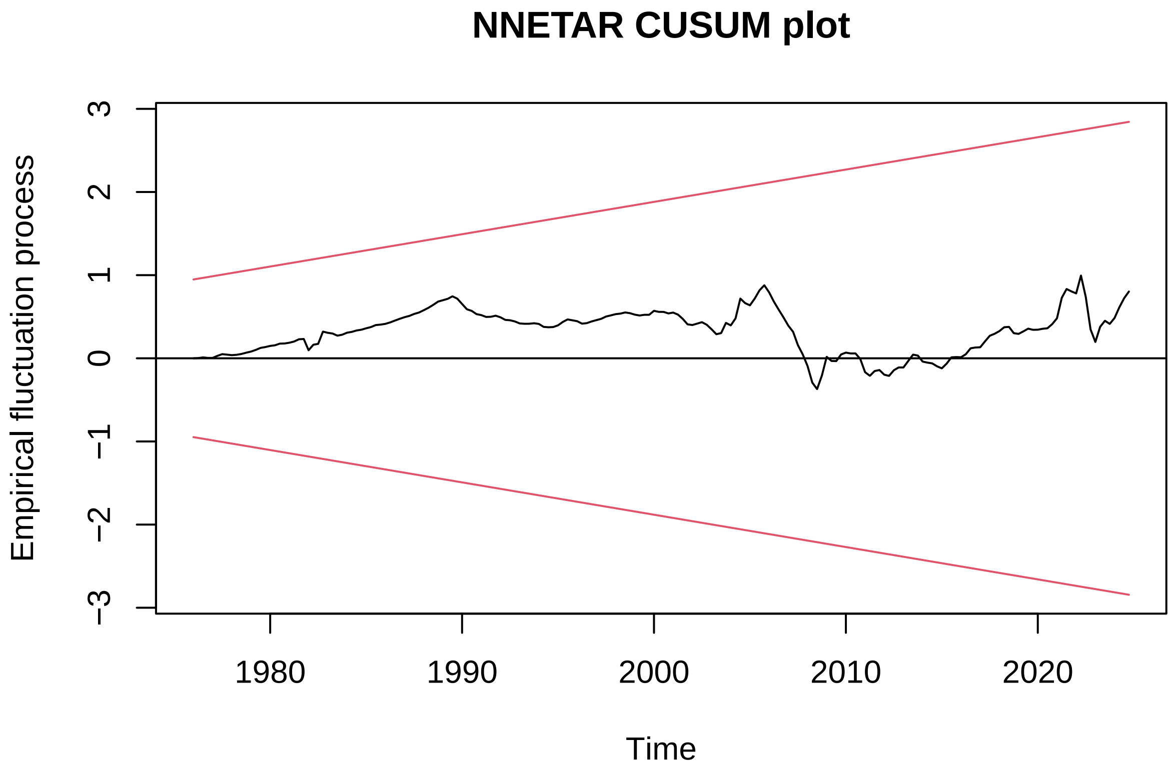
There seems to be some pattern and autocorrelation in the residuals, as they are significant at multiple lags. So the residuals do not resemble white noise.

```
# ljung-box test on residuals
Box.test(na.omit(residuals(nnetar_model)))
```

```
##
##  Box-Pierce test
##
## data:  na.omit(residuals(nnetar_model))
## X-squared = 28.449, df = 1, p-value = 9.618e-08
```

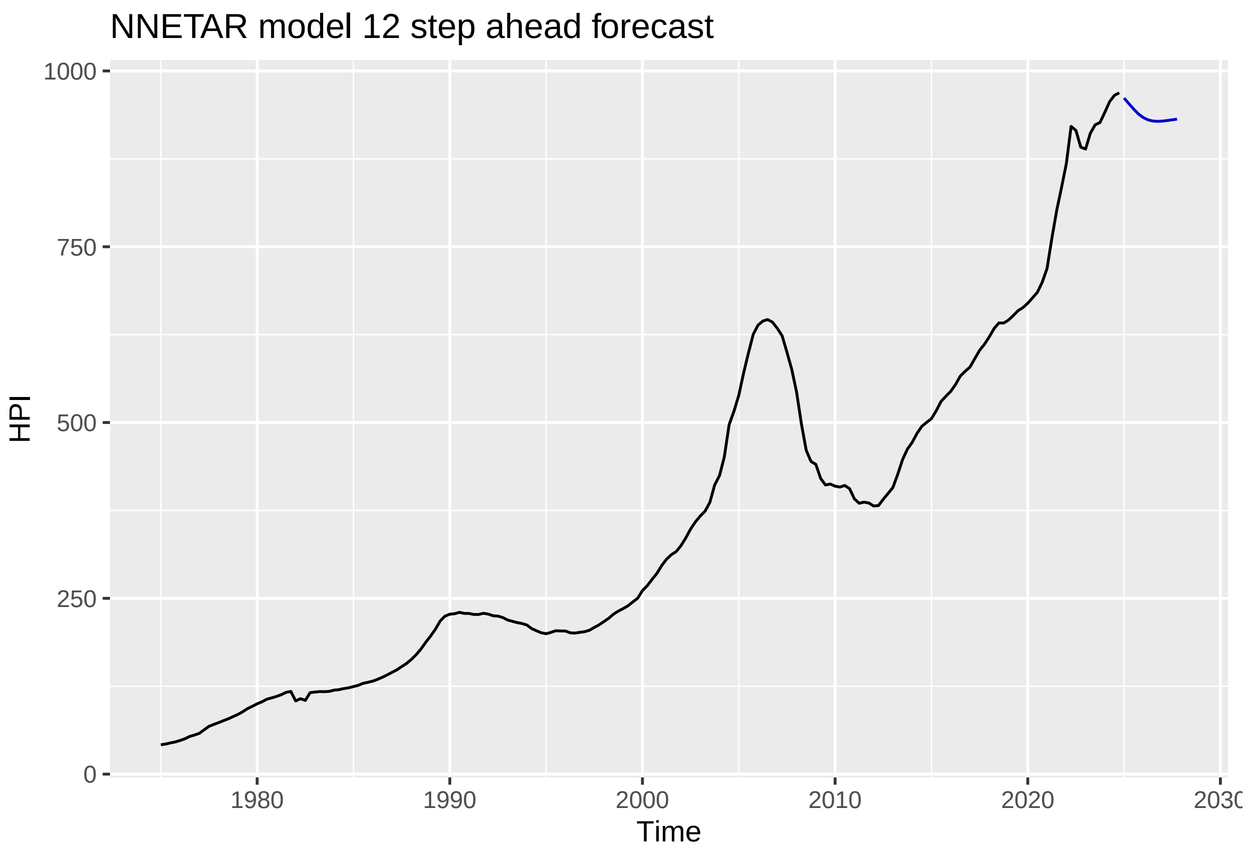
The p-value of the Ljung-box test is significant at the 95% confidence level, so we reject the null hypothesis and conclude that there is autocorrelation in the residuals, indicating a lack of fit of the model.

```
# plot the CUSUM
plot(efp(na.omit(residuals(nnetar_model)) ~ 1, type = "Rec-CUSUM"),
main = "NNETAR CUSUM plot")
```



According to the CUSUM plot, the cumulative sum of residuals starts to deviate from 0 early on, but does not deviate in a particular positive or negative direction, instead stays around 0.

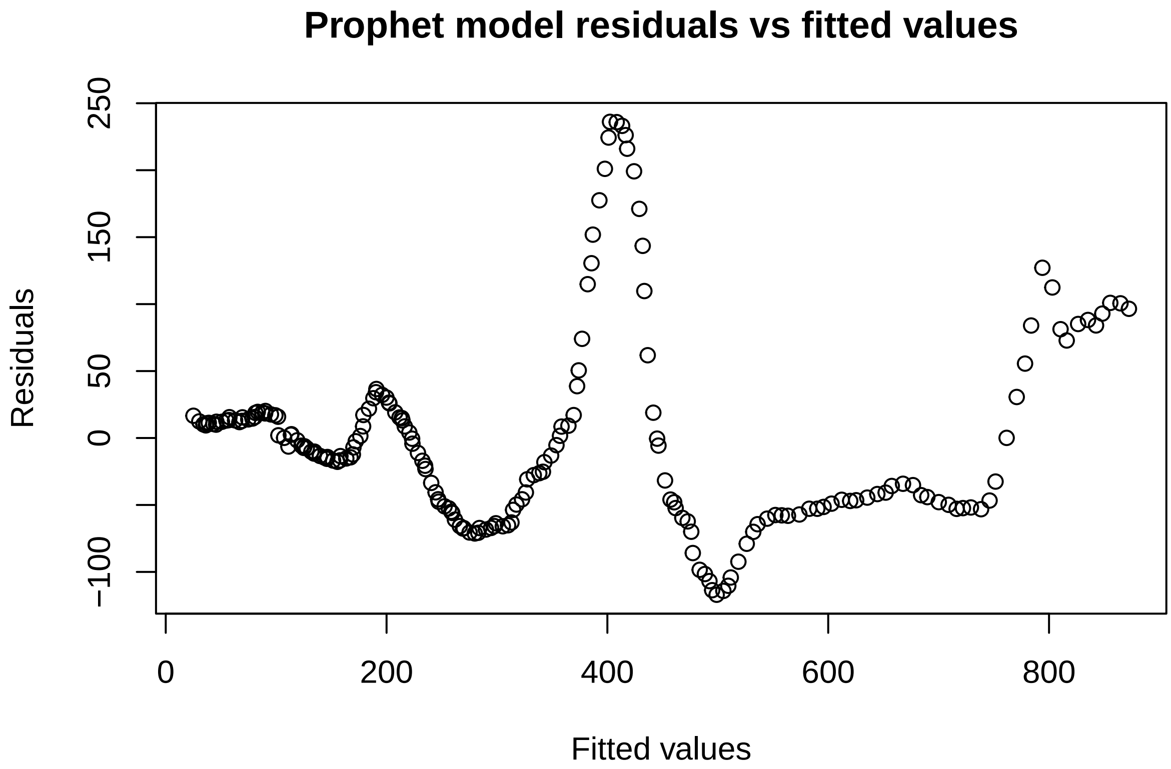
```r
# plot the 12 quarter ahead forecast
autoplot(forecast(nnetar_model, h = 12)) +
  ggtitle("NNETAR model 12 step ahead forecast") +
  ylab("HPI")
```



5. Prophet

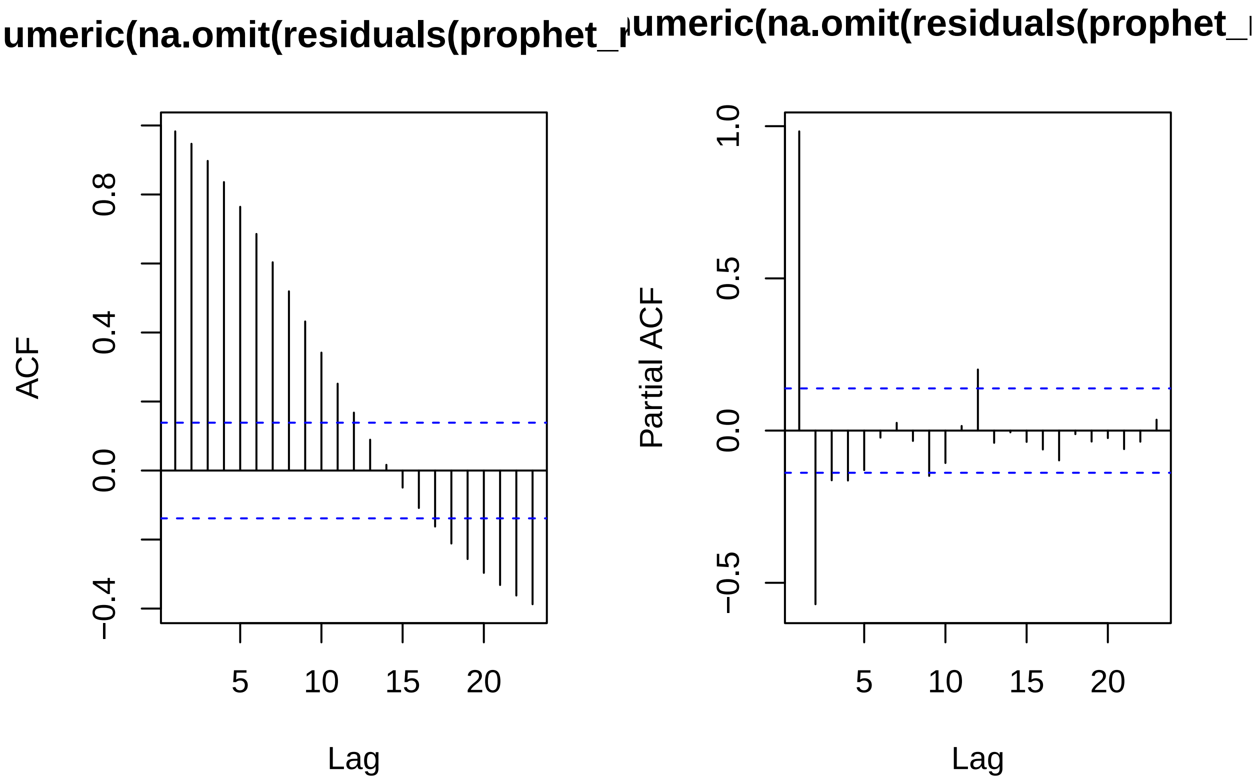
```r
# prophet method
ca_prices_tsibble <- tsibble(
  Quarter = yearquarter(seq(as.Date("1975-01-01"),
                            as.Date("2024-10-01"), by = "quarter")),
  Price = ca_price_ts)

prophet_model <- ca_prices_tsibble %>%
  model(prophet = prophet(Price ~ season(period = 4, order = 4,
                                        type = "additive")))
```

```r
# plot residuals vs fitted values
plot(fitted(prophet_model)[,3][[1]], residuals(prophet_model)[,3][[1]],
     main = "Prophet model residuals vs fitted values",
     xlab = "Fitted values",
     ylab = "Residuals")
```



There is a strong pattern and autocorrelation in the residuals, and the residuals are not evenly scattered around 0.

```r
# plot ACF and PACF of residuals
par(mfrow = c(1, 2))
acf(as.numeric(na.omit(residuals(prophet_model)[,3][[1]])))
pacf(as.numeric(na.omit(residuals(prophet_model)[,3][[1]])))
```



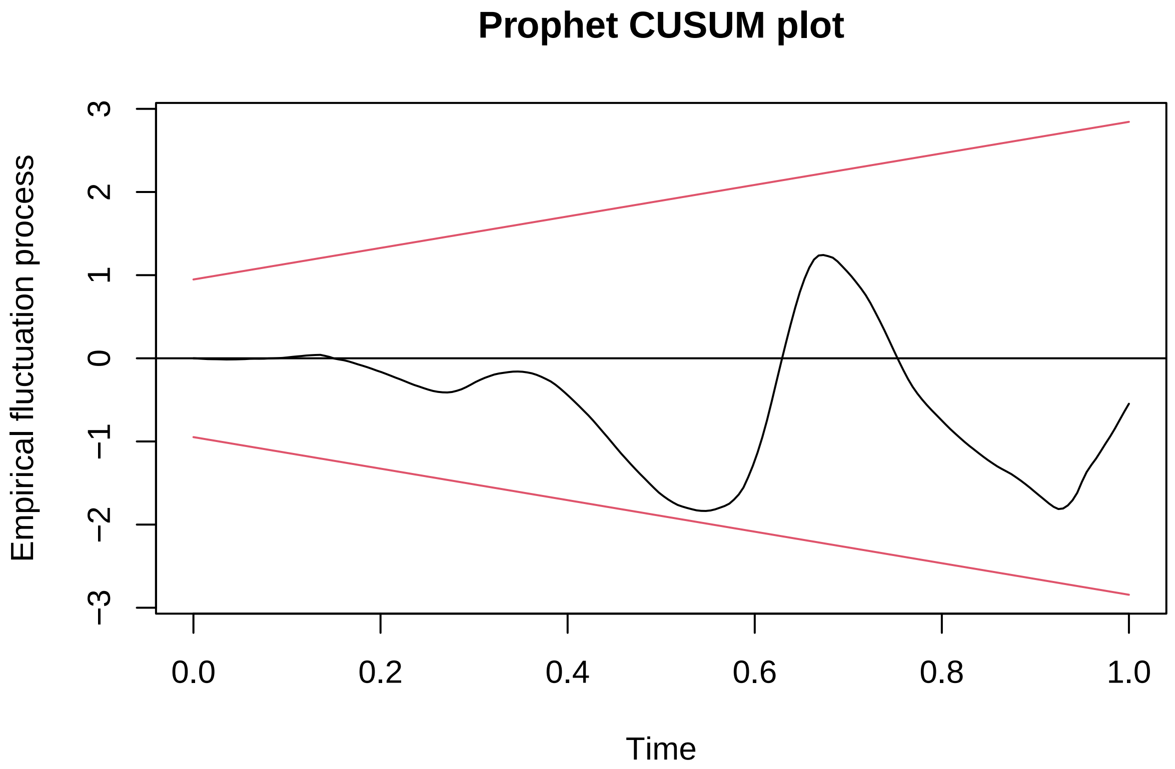
The ACF and PACF plot indicate strong autocorrelation in the residuals.

```r
# ljung-box test on residuals
Box.test(residuals(prophet_model)[,3][[1]])
```

```
##
##  Box-Pierce test
##
## data:  residuals(prophet_model)[, 3][[1]]
## X-squared = 193.24, df = 1, p-value < 2.2e-16
```

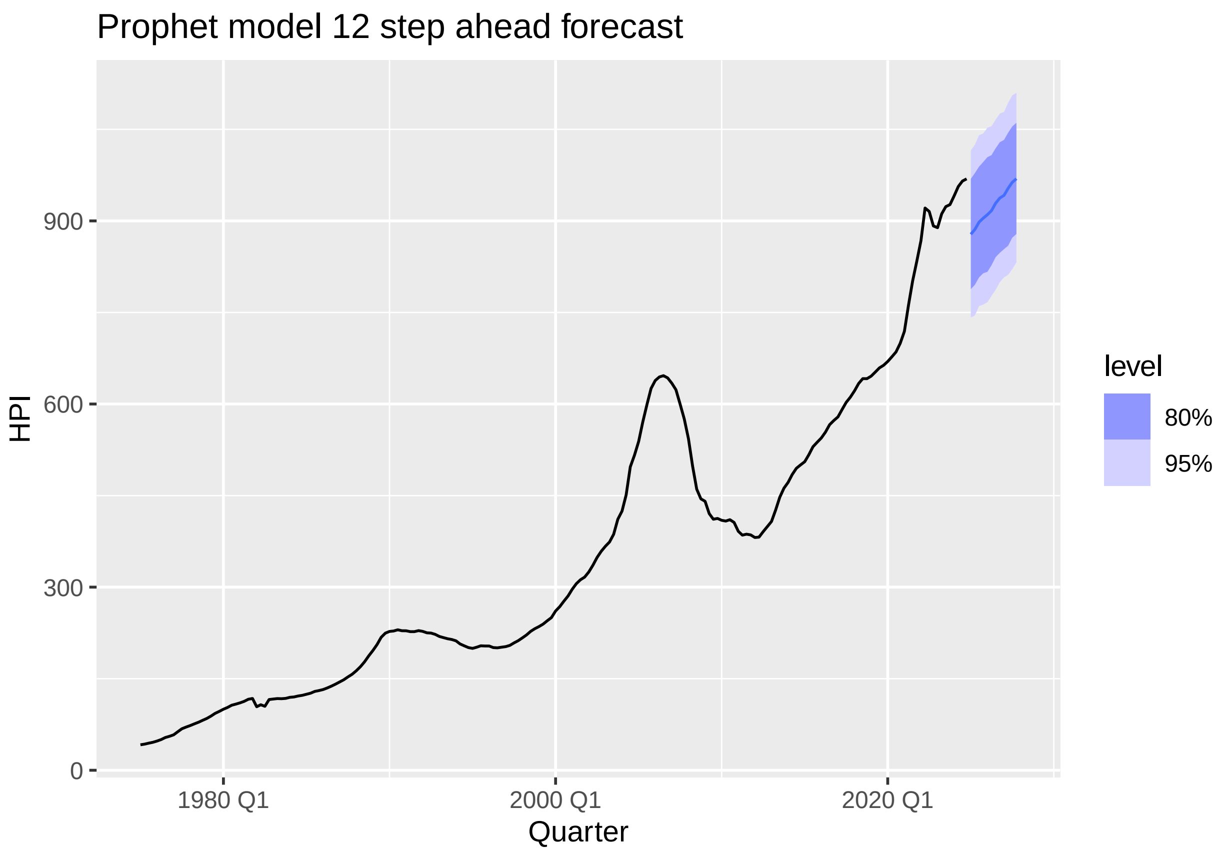
The p-value of the Ljung-box test is significant at the 95% confidence level, so we reject the null hypothesis and conclude that there is autocorrelation in the residuals, indicating a lack of fit of the model.

```r
# plot the CUSUM
plot(efp(residuals(prophet_model)[,3][[1]] ~ 1, type = "Rec-CUSUM"),
main = "Prophet CUSUM plot")
```



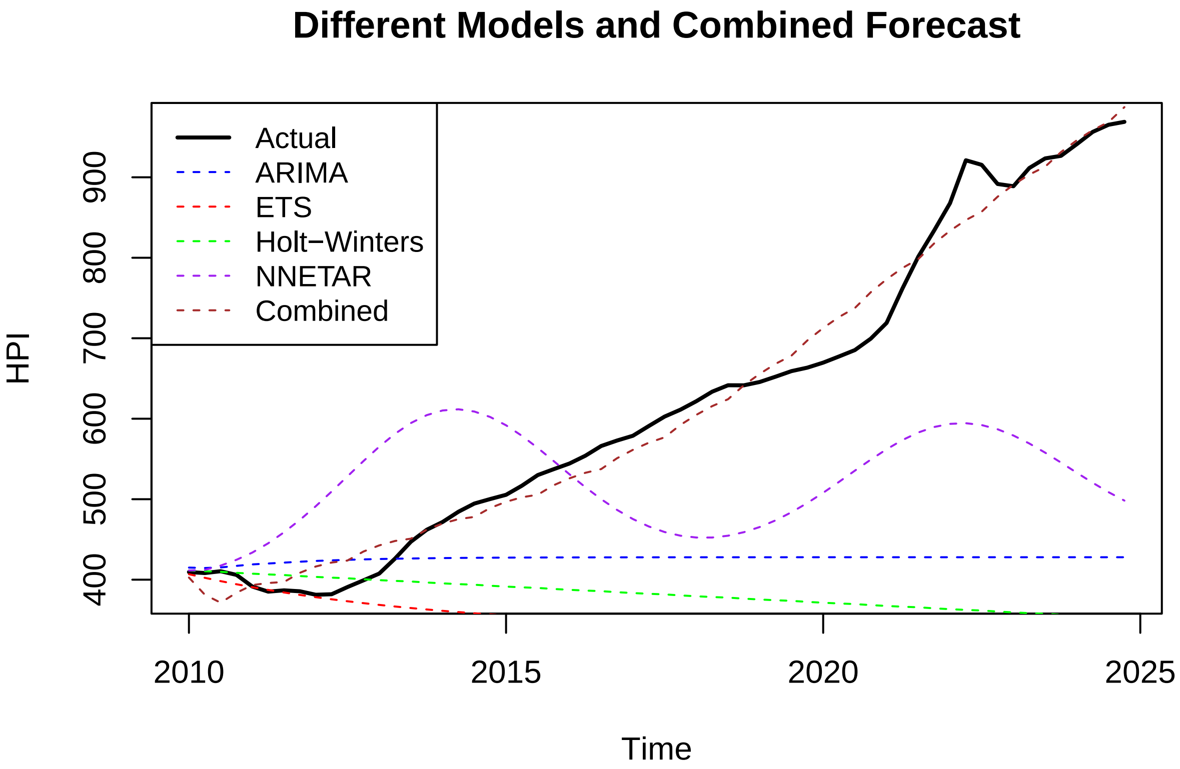
According to the CUSUM plot, the cumulative sum of the residuals have large deviations and fluctuations from 0, indicating a lack of fit of the model.

```
# plot the 12 quarter ahead forecast
forecast(prophet_model, h = 12) %>%
  autoplot(ca_prices_tsibble) +
  ggtitle("Prophet model 12 step ahead forecast") +
  ylab("HPI")
```



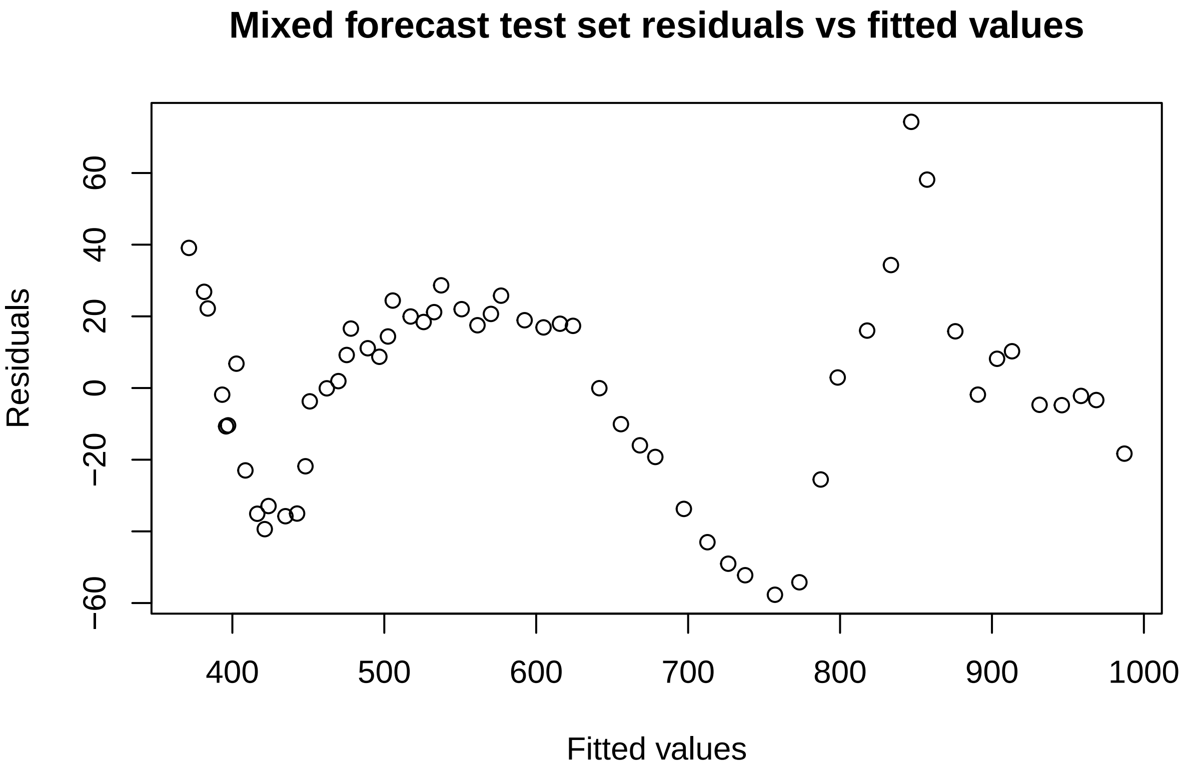
6. Forecast Combination

```
auto_arima_model_train <- auto.arima(ca_price_train)
ets_model_train <- ets(ca_price_train)
holt_winters_model_train <- HoltWinters(ca_price_train)
nnetar_model_train <- nnetar(ca_price_train)
ca_prices_train_tsibble <- tsibble(Quarter = yearquarter(seq(as.Date("1975-01-01"),
                                  as.Date("2009-10-01"), by = "quarter")),
        Price = ca_price_train)
prophet_model_train <- ca_prices_train_tsibble %>%
  model(prophet = prophet(Price ~ season(period = 4, order = 4,
                                    type = "additive")))


fore_auto_arima <- forecast(auto_arima_model_train, h = length(ca_price_test))$mean
fore_ets <- forecast(ets_model_train, h = length(ca_price_test))$mean
fore_holt_winters <- forecast(holt_winters_model_train, h = length(ca_price_test))$mean
fore_nnetar <- forecast(nnetar_model_train, h = length(ca_price_test))$mean
fore_prophet <- forecast(prophet_model_train, h = length(ca_price_test))[,4][[1]]

combine_df <- data.frame(
  Actual = ca_price_test,
  ARIMA = fore_auto_arima,
  ETS = fore_ets,
  HoltWinters = fore_holt_winters,
  NNETAR = fore_nnetar,
  Prophet = fore_prophet
  )

combined_model <- lm(Actual ~ 0 + ARIMA + ETS + HoltWinters + NNETAR,
                     data = combine_df)

combined_weights <- coef(combined_model)

combined_fore_weighted <-
  combined_weights["ARIMA"] * fore_auto_arima +
  combined_weights["ETS"] * fore_ets +
  combined_weights["HoltWinters"] * fore_holt_winters +
  combined_weights["NNETAR"] * fore_nnetar

plot(ca_price_test, type = "l", col = "black", lwd = 2,
     main = "Different Models and Combined Forecast",
     ylab = "HPI")
lines(fore_auto_arima, col = "blue", lty = 2)
lines(fore_ets, col = "red", lty = 2)
lines(fore_holt_winters, col = "green", lty = 2)
lines(fore_nnetar, col = "purple", lty = 2)
lines(combined_fore_weighted, col = "brown", lty = 2)
legend("topleft", legend = c("Actual", "ARIMA", "ETS", "Holt-Winters", "NNETAR", "Combined"),
       col = c("black", "blue", "red", "green", "purple", "brown"), lty = c(1,2,2,2,2,2),
       lwd = c(2,1,1,1,1,1), cex = 0.9)
```



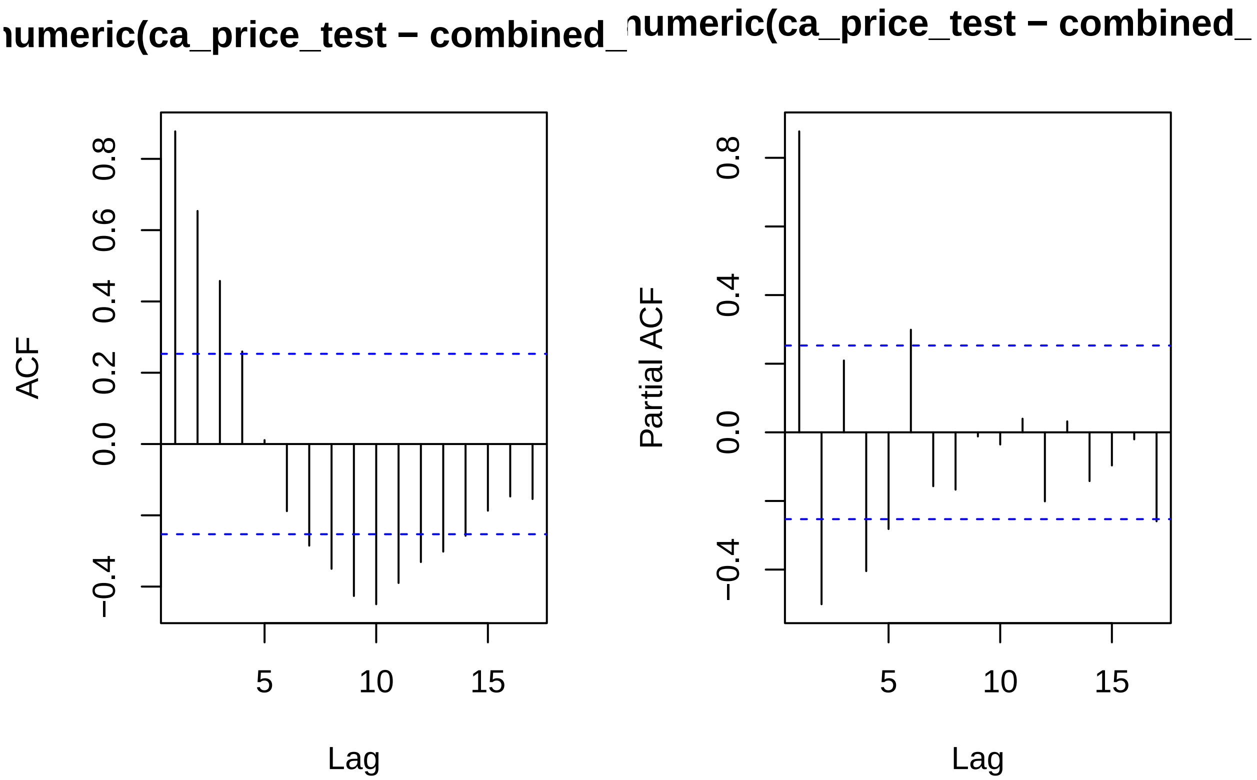
According to the plot of the fit of the individual models and combined forecasts, the combined forecasts (dotted brown line) fits the actual data (solid black line) the best. This indicates the combined forecast is preferred over the individual models. We will further verify this by looking at the error statistics of the models.

```r
# plot residuals vs fitted values
plot(as.numeric(combined_fore_weighted), as.numeric(ca_price_test - combined_fore_weighted),
     main = "Mixed forecast test set residuals vs fitted values",
     xlab = "Fitted values",
     ylab = "Residuals")
```



There is a strong pattern and autocorrelation in the residuals, and the residuals are not evenly scattered around 0.

```
par(mfrow = c(1, 2))
acf(as.numeric(ca_price_test - combined_fore_weighted))
pacf(as.numeric(ca_price_test - combined_fore_weighted))
```



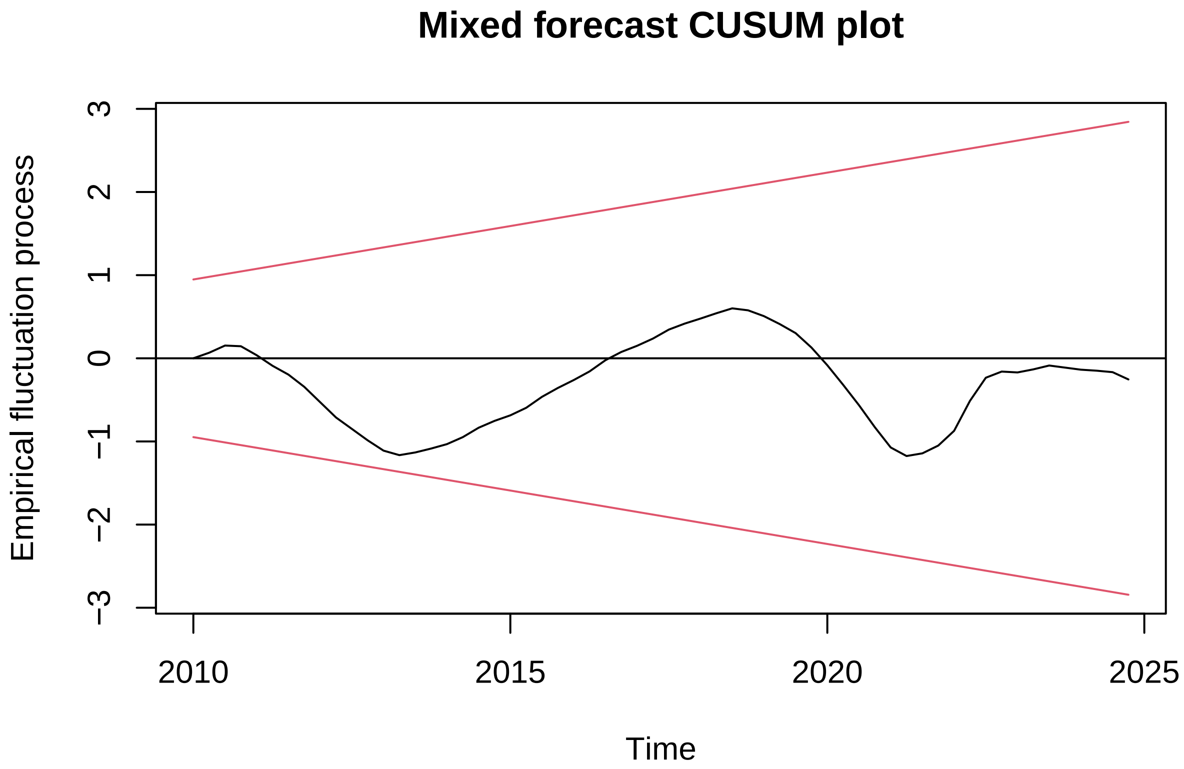
The ACF and PACF plot indicate strong autocorrelation in the test set residuals.

```
# ljung-box test on residuals
Box.test(ca_price_test - combined_fore_weighted)
```

```
##
##  Box-Pierce test
##
## data:  ca_price_test - combined_fore_weighted
## X-squared = 46.166, df = 1, p-value = 1.086e-11
```

The p-value of the Ljung-box test is significant at the 95% confidence level, so we reject the null hypothesis and conclude that there is autocorrelation in the residuals, indicating a lack of fit of the model.

```
# plot the CUSUM
plot(efp((ca_price_test - combined_fore_weighted) ~ 1, type = "Rec-CUSUM"),
main = "Mixed forecast CUSUM plot")
```



According to the CUSUM plot, the cumulative sum of the residuals have large deviations and fluctuations from 0 (though not as large as the residuals in the Prophet model), indicating a lack of fit of the model.

Training and Testing Error

```
# auto.arima model
accuracy(forecast(auto_arima_model_train, h = length(ca_price_test)), ca_price_test)
```

```
##                       ME       RMSE        MAE       MPE      MAPE      MASE
## Training set   0.3998171   5.523836   3.229838  0.296704  1.364039 0.1231328
## Test set     194.2959084 268.971665 205.078748 24.986943 27.753636 7.8183272
##                     ACF1 Theil's U
## Training set -0.08176642        NA
## Test set      0.95794352  14.42269
```

```
# ETS model
accuracy(forecast(ets_model_train, h = length(ca_price_test)), ca_price_test)
```

```
##                       ME       RMSE        MAE        MPE      MAPE       MASE
## Training set   0.2380864   6.012881   3.450141  0.2709519  1.328373  0.1315316
## Test set     265.7531737 334.540849 265.823940 36.5979244 36.616297 10.1341488
##                   ACF1 Theil's U
## Training set 0.3699647        NA
## Test set     0.9560344   18.6259
```

```r
# Holt-Winters model
accuracy(forecast(holt_winters_model_train, h = length(ca_price_test)), ca_price_test)
```

```
##                        ME       RMSE        MAE         MPE      MAPE      MASE
## Training set  -0.02829005   5.795343   3.328961  0.06617084  1.376396 0.1269118
## Test set     238.41149159 314.380203 242.984442 31.75976549 32.942332 9.2634264
##                    ACF1 Theil’s U
## Training set 0.01614141        NA
## Test set     0.95707387  17.11623
```

```r
# NNETAR model
accuracy(forecast(nnetar_model_train, h = length(ca_price_test)), ca_price_test)
```

```
##                       ME       RMSE      MAE        MPE      MAPE      MASE
## Training set -0.02286136   4.482035   2.8184 -0.1162329  1.294594 0.1074474
## Test set     98.08367431 204.507795 167.9112  8.7955345 24.983777 6.4013696
##                     ACF1 Theil’s U
## Training set -0.05032877        NA
## Test set      0.95631577  12.09702
```

```r
# Prophet model
ca_prices_test_tsibble <- tsibble(Quarter = yearquarter(seq(as.Date("2010-01-01"),
                                  as.Date("2024-10-01"), by = "quarter")),
        Price = ca_price_test)
accuracy(forecast(prophet_model_train, h = length(ca_price_test)),
         ca_prices_test_tsibble)
```

```
## # A tibble: 1 x 10
##   .model  .type    ME  RMSE   MAE   MPE  MAPE  MASE RMSSE  ACF1
##   <chr>   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 prophet Test  -146.  167.  149. -29.4  29.7   NaN   NaN 0.956
```

```r
# Combined model
accuracy(combined_fore_weighted, ca_price_test)
```

```
##                  ME     RMSE      MAE        MPE     MAPE      ACF1 Theil’s U
## Test set 0.01230304 27.03598 21.53807 -0.2144641 3.769051 0.8771749  2.049542
```

The errors in the test set of the combined forecast model are the lowest out of all the models and are significantly lower than the other model errors. This shows the combined forecat model is the preferred model based on lowest error. We focus on testing error because it shows the model performs well on unsees data and will have better forecasting performance, providing better real-world application of the data.

III. Conclusions and Future Work

Conclusions This study analyzed the quarterly House Price Index (HPI) of California from 1975 Q1 to 2024 Q2, employing various time series forecasting models, including ARIMA, ETS, Holt-Winters, NNETAR, Prophet, and a combined forecasting approach. The goal was to identify the most suitable model for predicting future house prices by evaluating model performance using diagnostic tests and error metrics.

The results indicate that the combined forecast model outperformed individual models in terms of minimizing prediction errors. The ARIMA model displayed strong residual properties, indicating white noise behavior, but its forecasting ability was constrained by its reliance on past linear patterns. The ETS and Holt-Winters models captured trend and seasonality effectively but had higher error rates. The NNETAR and Prophet models exhibited significant autocorrelation in residuals, suggesting inadequate fit for this dataset. The combined forecast model leveraged the strengths of multiple models, leading to the lowest testing errors and superior generalization capability.

Overall, the findings demonstrate that a hybrid approach of integrating multiple forecasting models can improve prediction accuracy and robustness, as it incorporates the strengths of each model. This insight is particularly valuable for stakeholders in the housing market, policymakers, and financial analysts who rely on accurate forecasts for decision-making.

Future Work While the combined forecast model showed promising results, there is space for future research and improvement:

1. **Incorporation of External Variables:** Future studies could enhance the forecasting model by integrating macroeconomic indicators such as interest rates, inflation, employment rates, and mortgage rates to improve prediction accuracy.
2. **Refinement of Forecast Combination Methods:** The current combined forecasting approach assigns weights based on regression coefficients. Alternative techniques, such as directly taking the mean of the different model forecasts or weighted averaging based on model performance metrics, could be explored to further enhance predictive power.
3. **Comparative Analysis with Other States:** A cross-state analysis comparing California's housing market trends with other high-cost states like New York or Texas could provide valuable insights into regional housing dynamics and price determinants.
4. **Impact of Policy Changes:** Future research could examine how government policies, such as tax incentives, housing subsidies, and zoning regulations, impact California's housing market. Incorporating these factors into the forecasting model could provide insights into policy effectiveness and housing affordability.

By addressing these aspects, future research can contribute to a more comprehensive understanding of housing price dynamics and further improve predictive capabilities for economic decision-making.

IV. References

https://www.fhfa.gov/data/hpi/datasets?tab=quarterly-data

https://www.fhfa.gov/hpi/download/quarterly_datasets/hpi_at_state.csv