Project Red Scare

Daniel Truver

5/03/2018

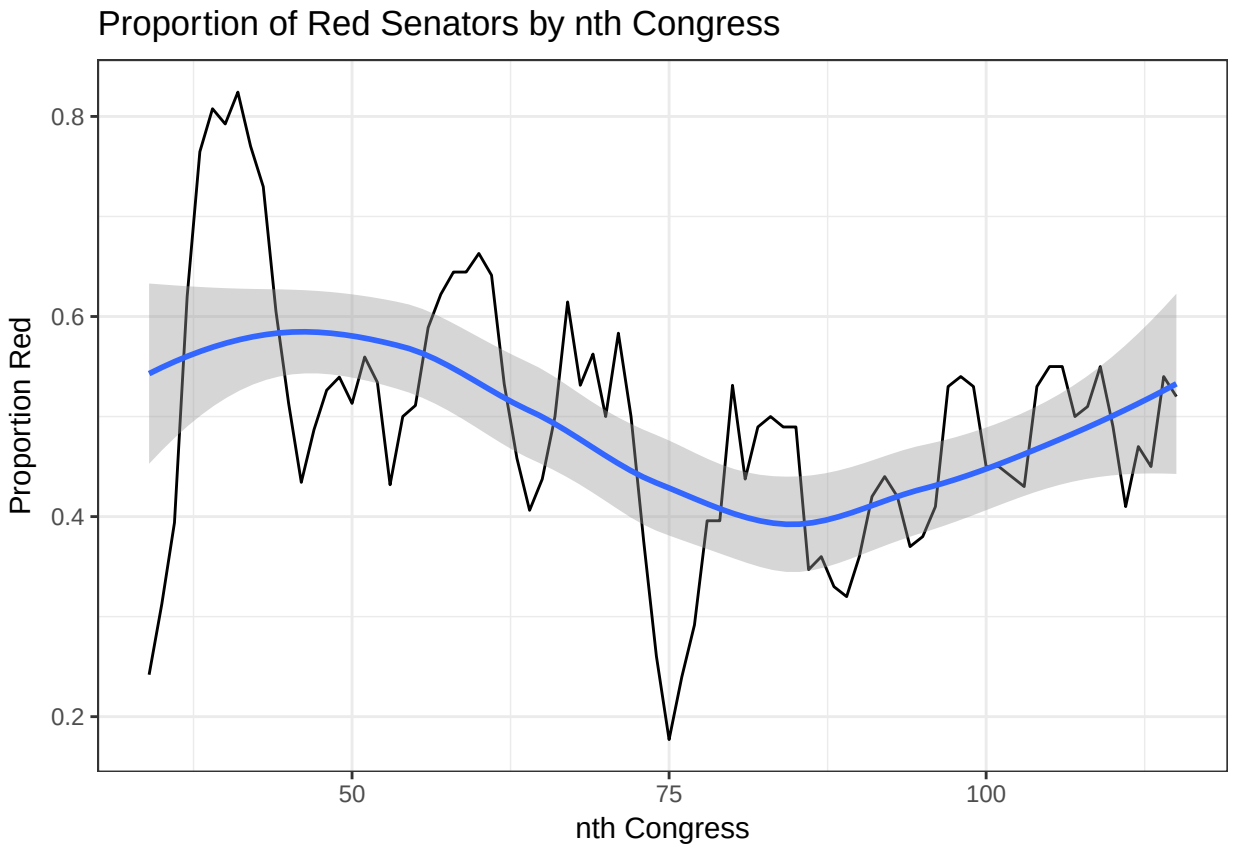
Introduction

After the presidential election of 2016, and once they got through some of the depression, democrats predicted that the 2018 midterm elections would see a blue wave. Of all the times I have heard this claim, I have never seen a statistical model that predicts it. The goal of this project is to construct a time series to model the proportion of seats held by republicans in the House and Senate of the U.S. Congress. We will construct a set of predictors for the mean and acount for additional variation with AR and MA terms. The eventual goal is to predict the proportion of republicans in the 116th Congress.

EDA

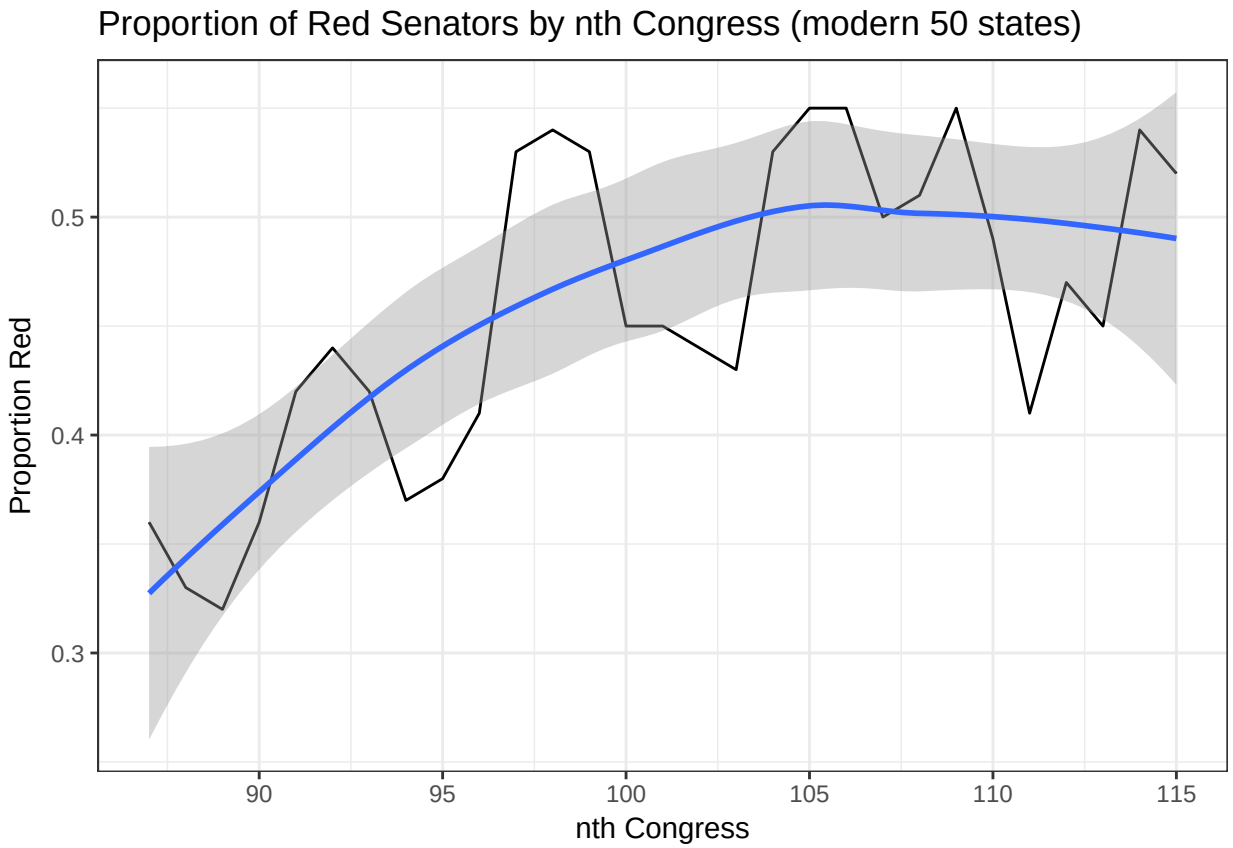
The following data come from the Brooking's institute and were last update in September of 2017. Our first concern should be if a time series model is appropriate. We also want to know if the intended process is stationary and what seasonal trends appear.

```
partisan = read.csv("partisan_congress.csv", stringsAsFactors = FALSE)
partisan$Number.of.representatives[53:54] = c(436,437)
partisan = partisan[1:82,]
for (col in names(partisan) %>% .[!.%in%c("year")] ){
  partisan[,col] = partisan[,col] %>%
    gsub("[^0-9\\.]", "", .) %>%
    as.integer()
}
partisan.full = partisan = partisan %>%
  mutate(red_sen = Republican_sen/Number.of.senators) %>%
  mutate(red_rep = Republican_rep/Number.of.representatives) %>%
  mutate(year_in = as.integer(str_extract(year, "\\d\\d\\d\\d")))
ggplot(data = partisan,
       aes(x = congress, y = red_sen)) +
  geom_line() +
  ggtitle("Proportion of Red Senators by nth Congress") +
  xlab("nth Congress") + ylab("Proportion Red") +
  geom_smooth()
```



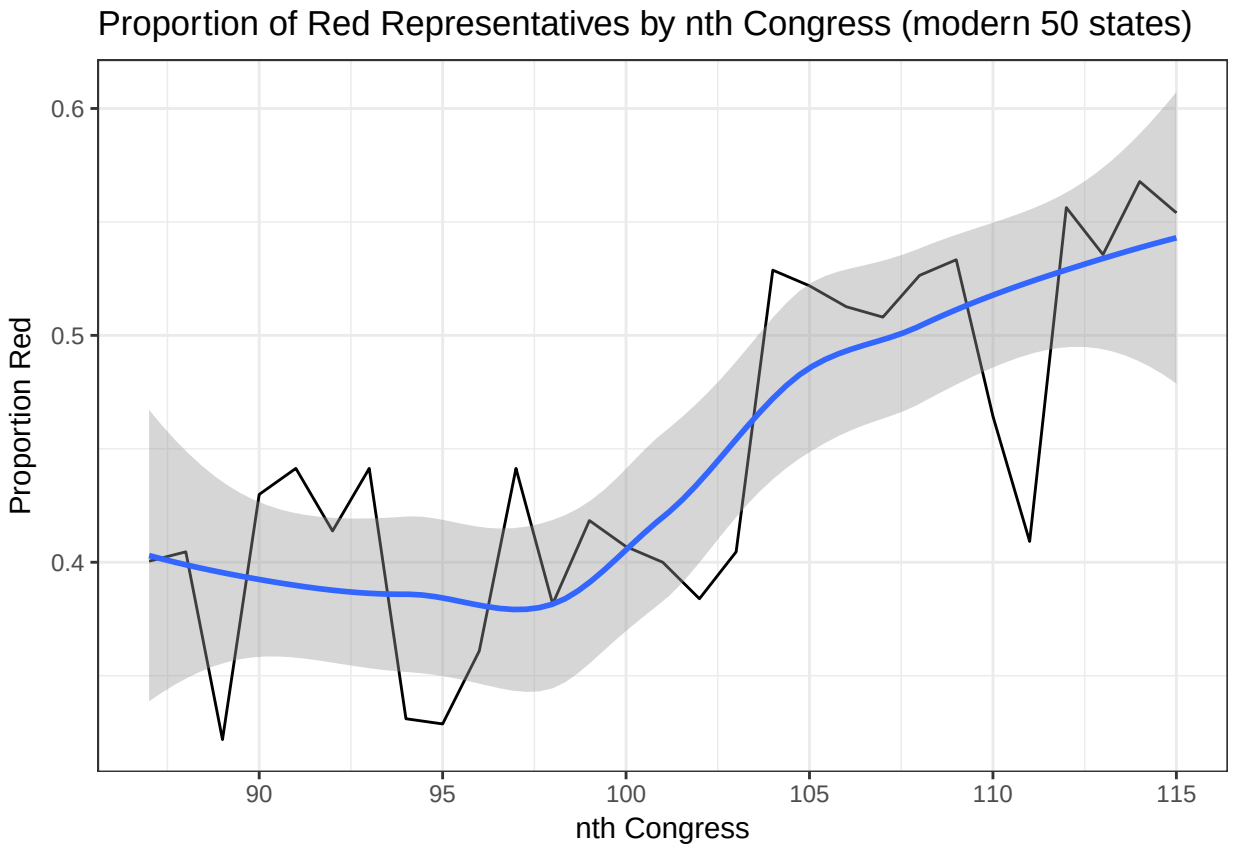
The above figure shows the proportion of Senate seats occupied by republicans. There does seem to be dependence in time. This figure is all years for which we have data, but moving forward, we will only incorporate years for which all 50 modern states have representatives in the Senate. The adjusted figure is below.

```
partisan = partisan %>% filter(Number.of.senators == 100)
ggplot(data = partisan,
       aes(x = congress, y = red_sen)) +
  geom_line() +
  ggtitle("Proportion of Red Senators by nth Congress (modern 50 states)") +
  xlab("nth Congress") + ylab("Proportion Red") +
  geom_smooth()
```



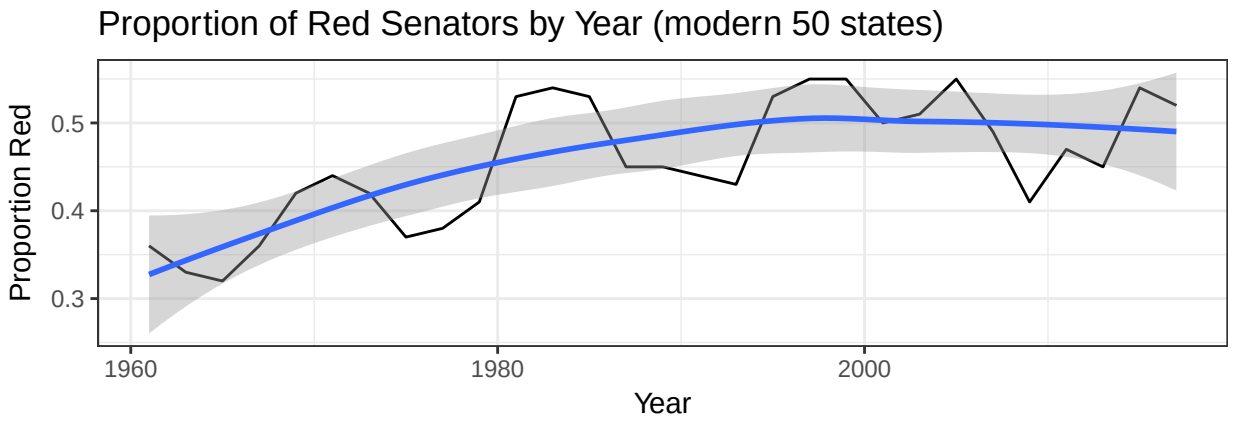
We lose a large fraction of our data by making this change, but we are most interested in modeling the modern trends in Congress anyway, and our target predictions involve all 50 states. We see a different overal trend in these data than we saw when including all previous years. There does still appear to be autocorrelation. Let's take a look at the House.

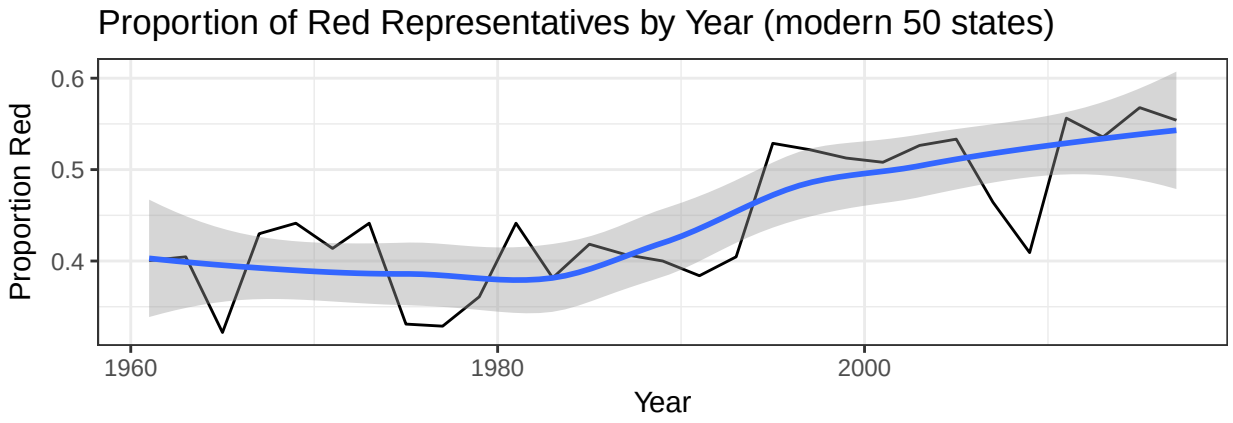
```
ggplot(data = partisan,
       aes(x = congress, y = red_rep)) +
  geom_line() +
  ggtitle("Proportion of Red Representatives by nth Congress (modern 50 states)") +
  xlab("nth Congress") + ylab("Proportion Red") +
  geom_smooth()
```



The trend is incresing as with the senate, but the pattern is not the same. The trend by year for both bodies is below.

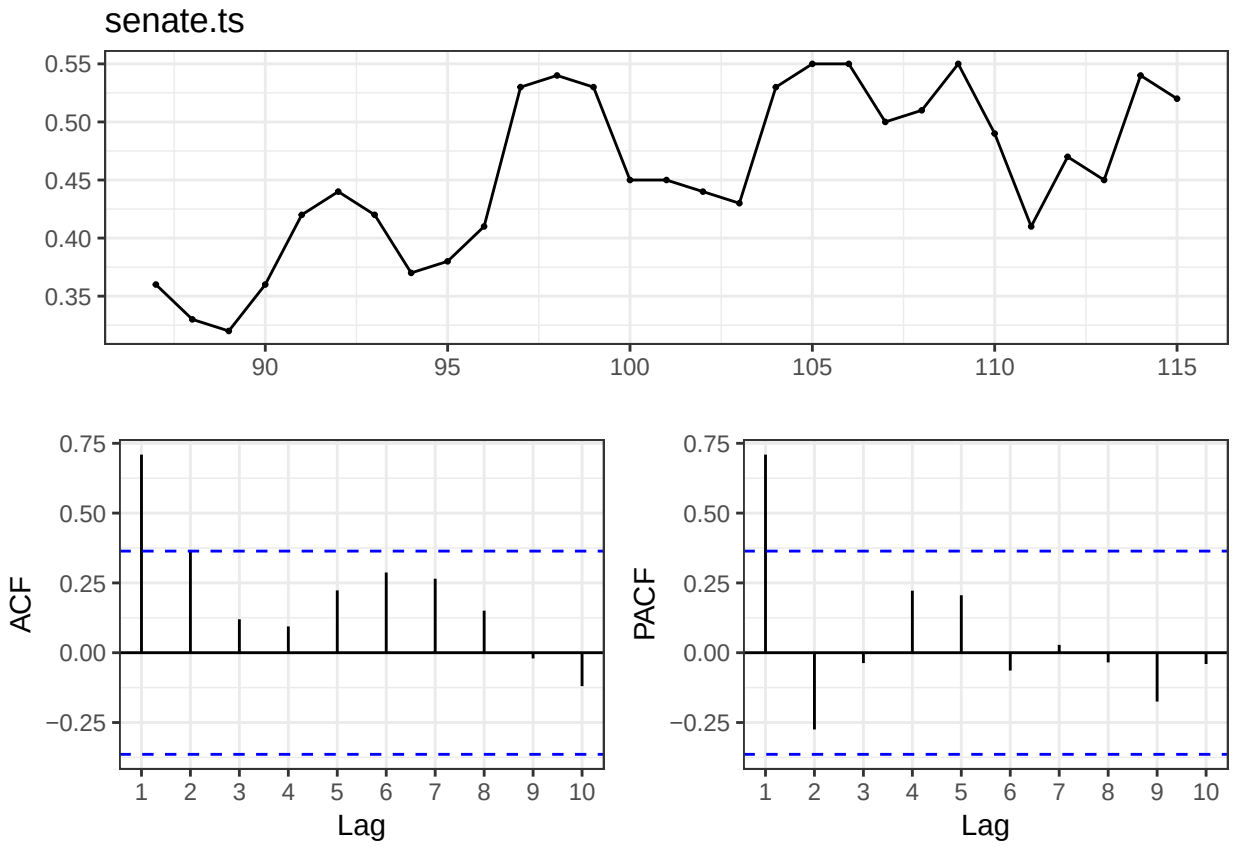
```
library(gridExtra)
g_rep = ggplot(data = partisan,
       aes(x = year_in, y = red_rep)) +
  geom_line() +
  ggtitle("Proportion of Red Representatives by Year (modern 50 states)") +
  xlab("Year") + ylab("Proportion Red") +
  geom_smooth()
g_sen = ggplot(data = partisan,
       aes(x = year_in, y = red_sen)) +
  geom_line() +
  ggtitle("Proportion of Red Senators by Year (modern 50 states)") +
  xlab("Year") + ylab("Proportion Red") +
  geom_smooth()
grid.arrange(g_sen, g_rep, nrow = 2)
```



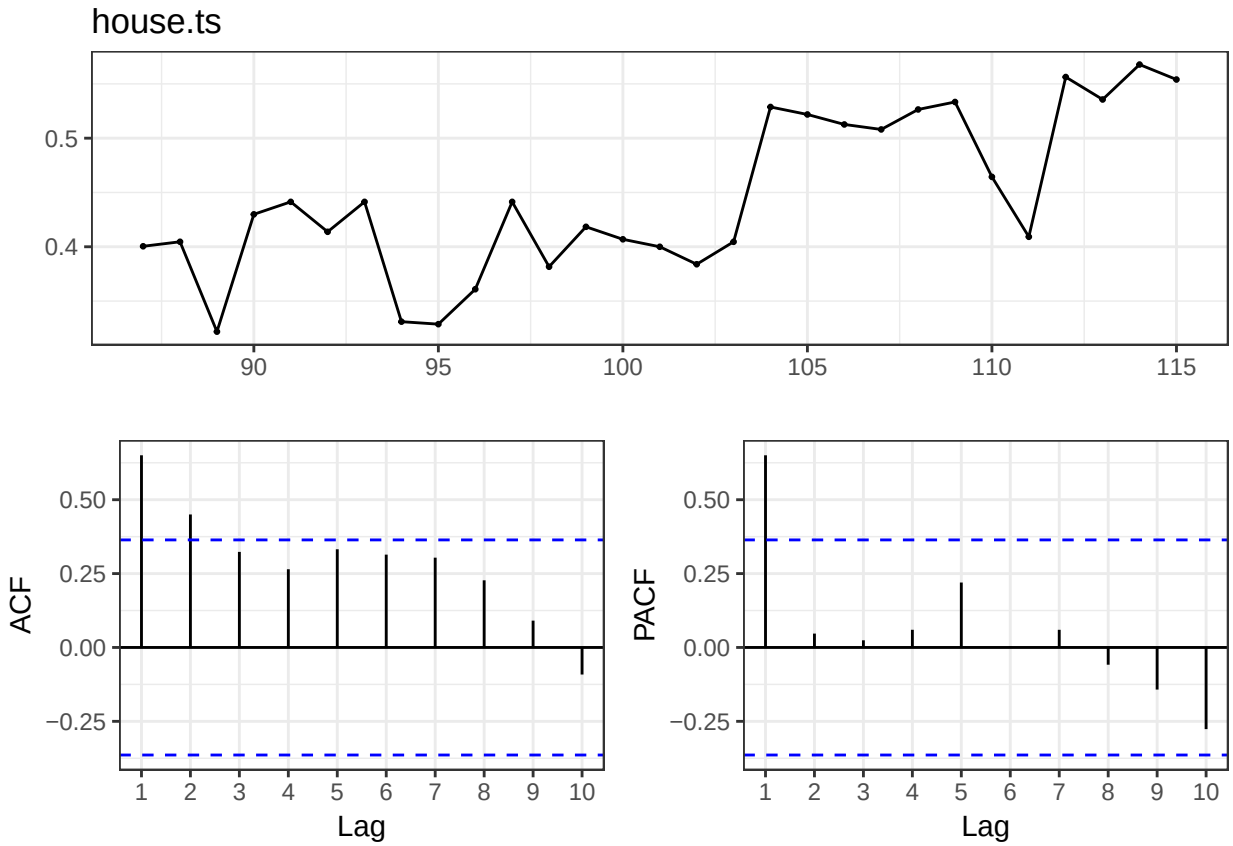


Technical EDA

```
library(forecast)
senate.ts = ts(data = partisan %>% select(red_sen), start = 87)
ggtsdisplay(senate.ts)
```




```
house.ts = ts(data = partisan %>% select(red_rep), start = 87)
ggtsdisplay(house.ts)
```

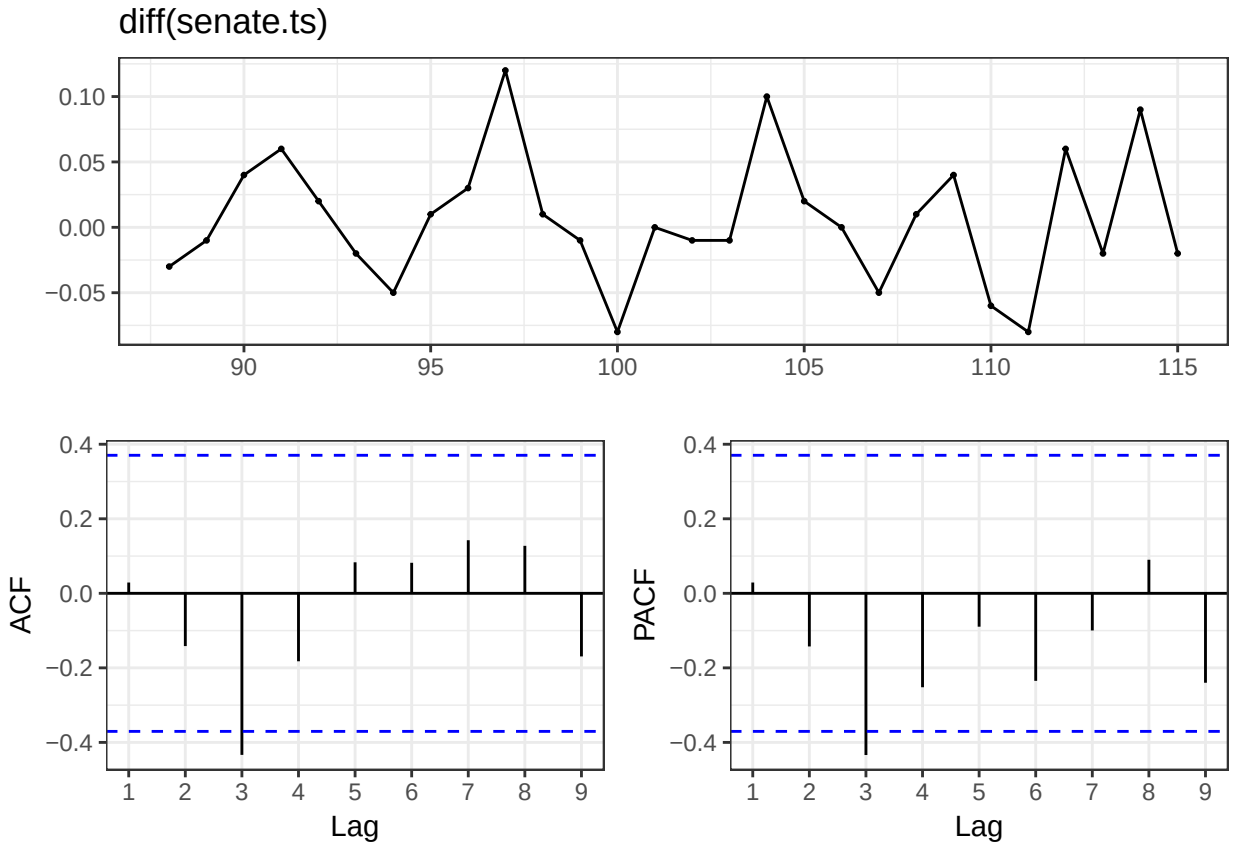


There is autocorrelation in both of these series, but it does not look the same. There could also be some stationarity issues. We're going to ignore R's suggested significance levels and just look at the ACF and PACF spikes; we have to remember that we do not have an abundance of data. The Senate's ACF has a strange wave pattern to it, but it seems to die out at lag 2. Similary, the Senate's PACF spikes are lag 1 and lag 2.

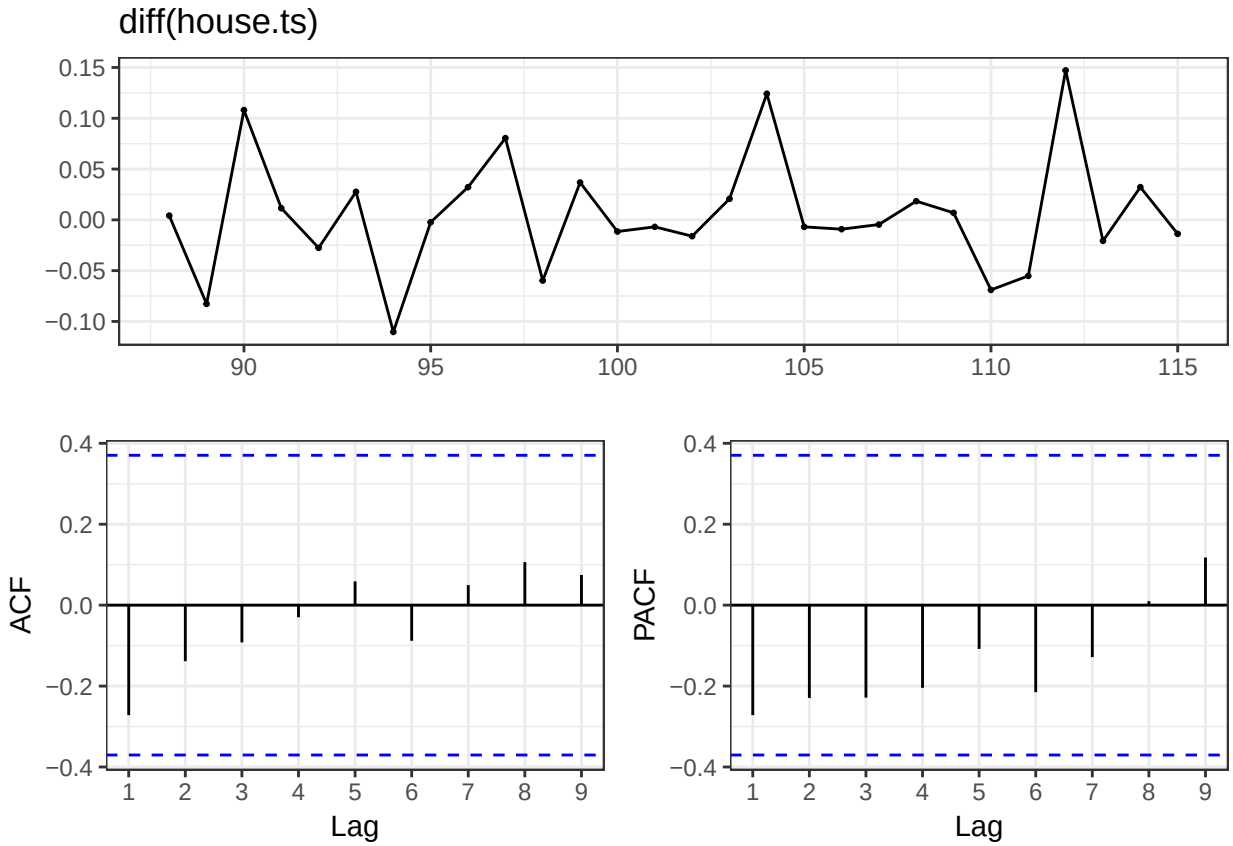
The House's ACF takes a lot longer to die off to the same level as the Senate's, and it still has the wave pattern. The House also shows PACF spike at 5 and 10, possibly a seasonal component?

We attempt differencing.

```
ggtsdisplay(diff(senate.ts))
```

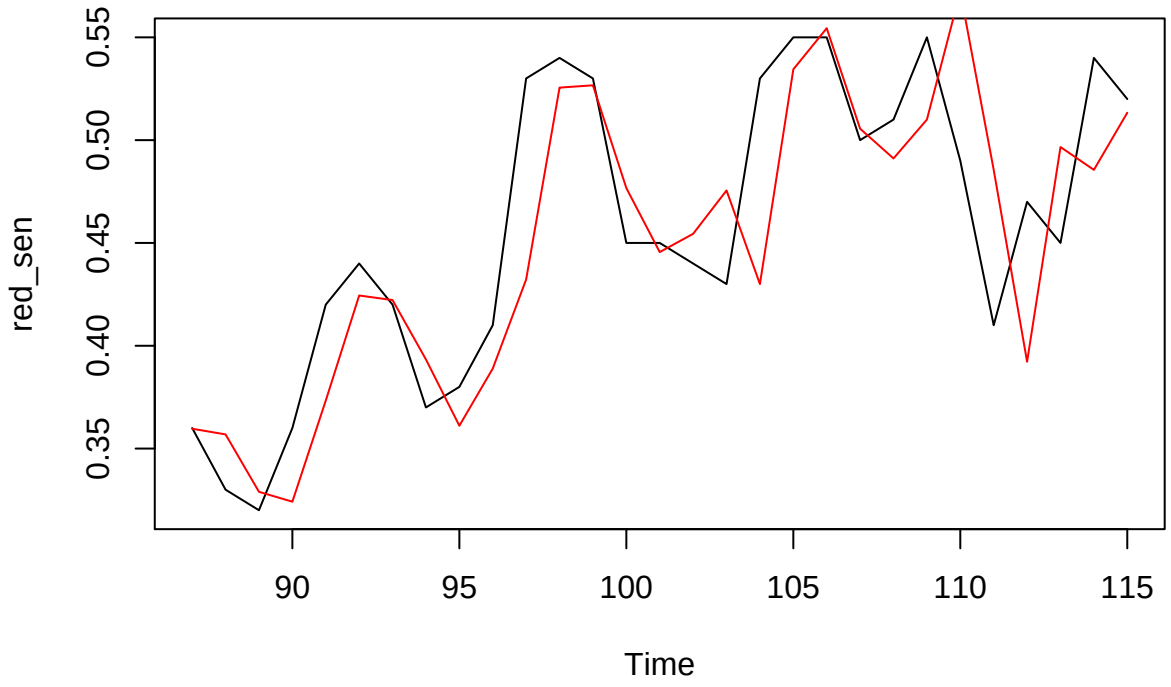



```
ggtsdisplay(diff(house.ts))
```



Differencing the Senate seems to reveal as seasonal component every 3 congresses. If we recall our high school government classes, this makes sense; Senate seats only come up for re-election every six years (3 session of Congress).

```
basicSenate = Arima(senate.ts, order = c(0,1,0), seasonal = list(order = c(1,0,0), period = 3))
{
  plot(senate.ts)
  # points(time(fitted(basicSenate))-deltat(basicSenate), basicSenate$fitted,
  #        col = "red", type = "l")
  points( basicSenate$fitted, col = "red", type = "l")
}
```




```
basicHouse = Arima(house.ts, order = c(1,0,0))
{
  plot(house.ts)
  points(basicSenate$fitted, col = "red", type = "l")
}
```