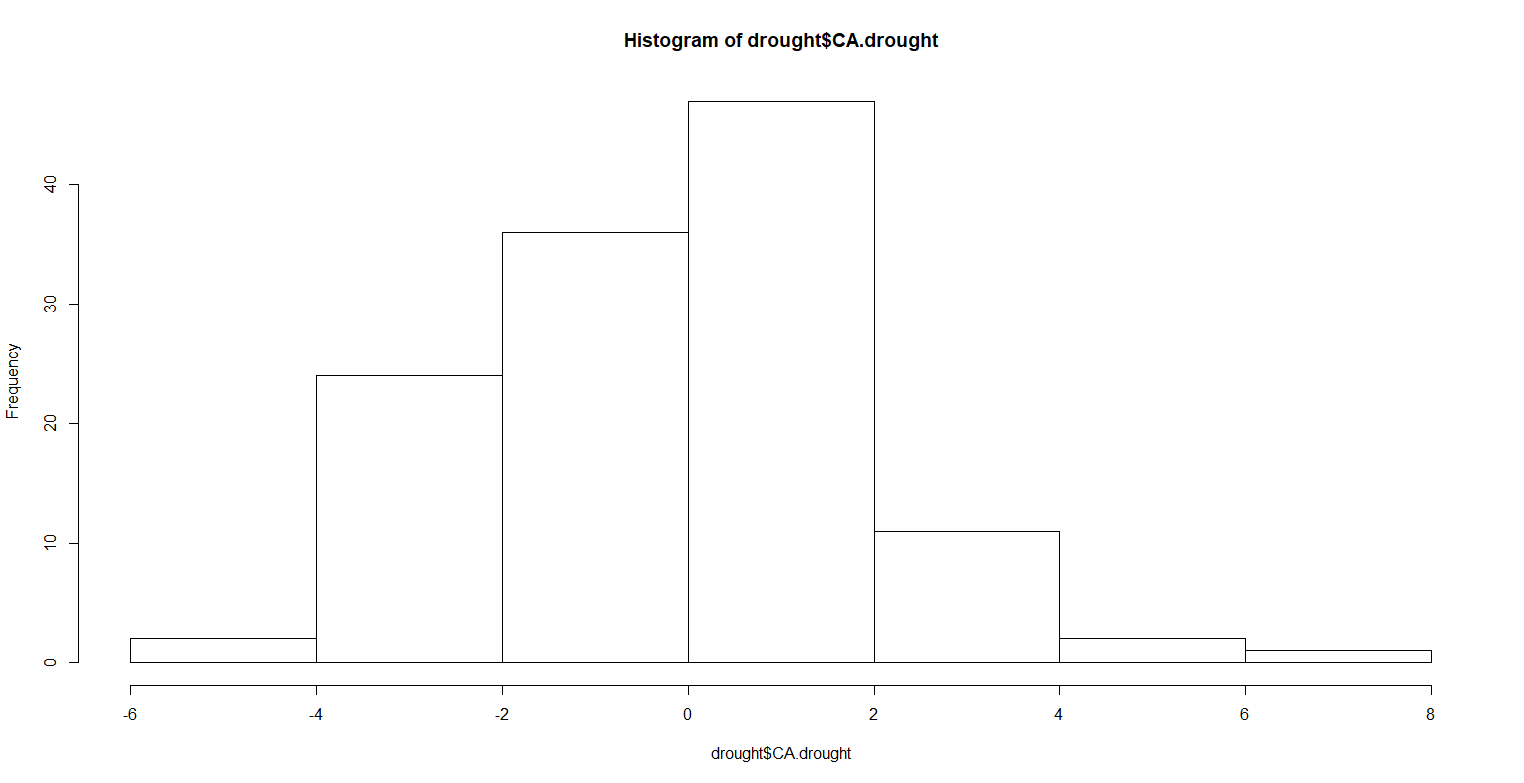
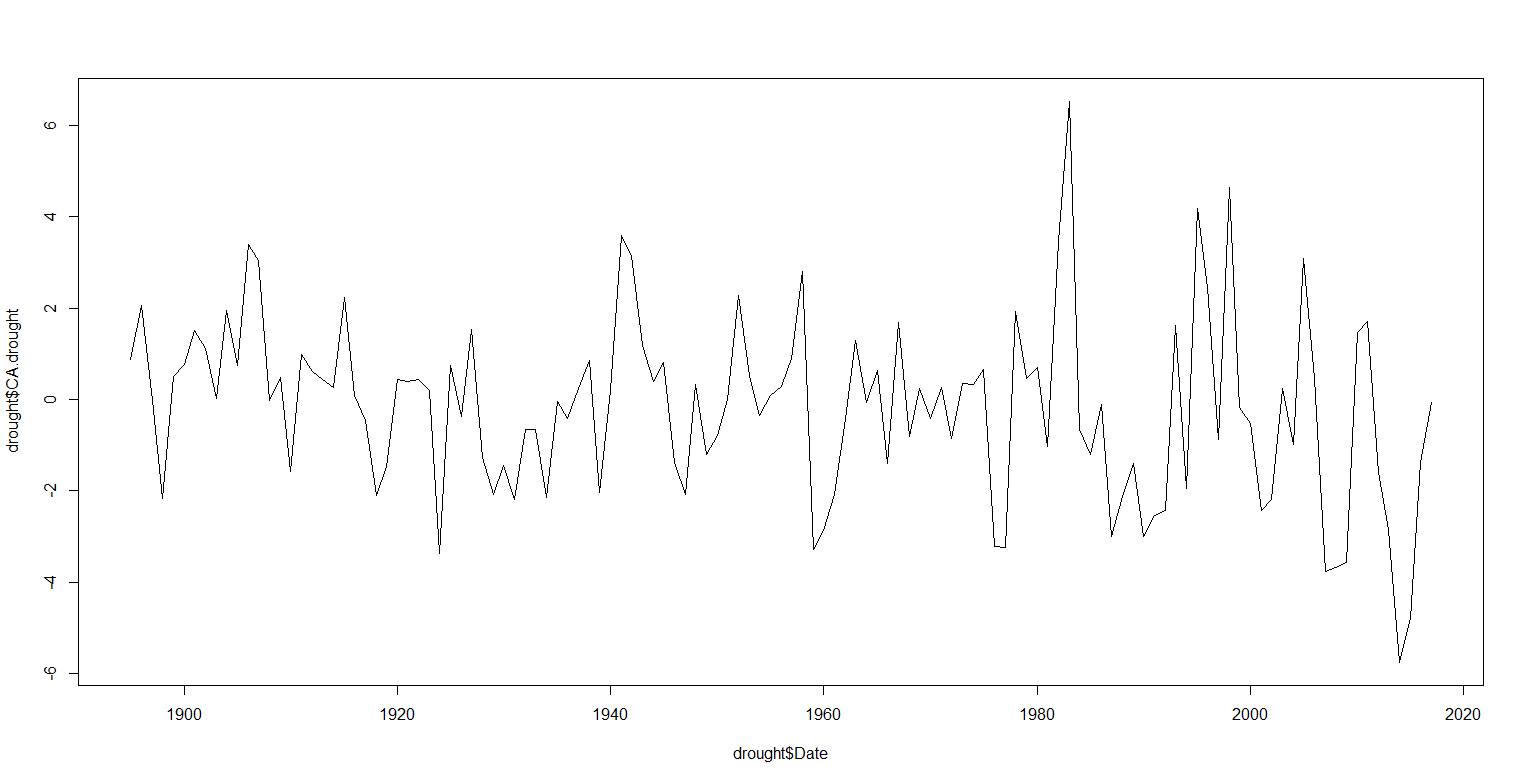
**Analysis of California Drought Severity Index from 1895-2017**

Anne Polyakov

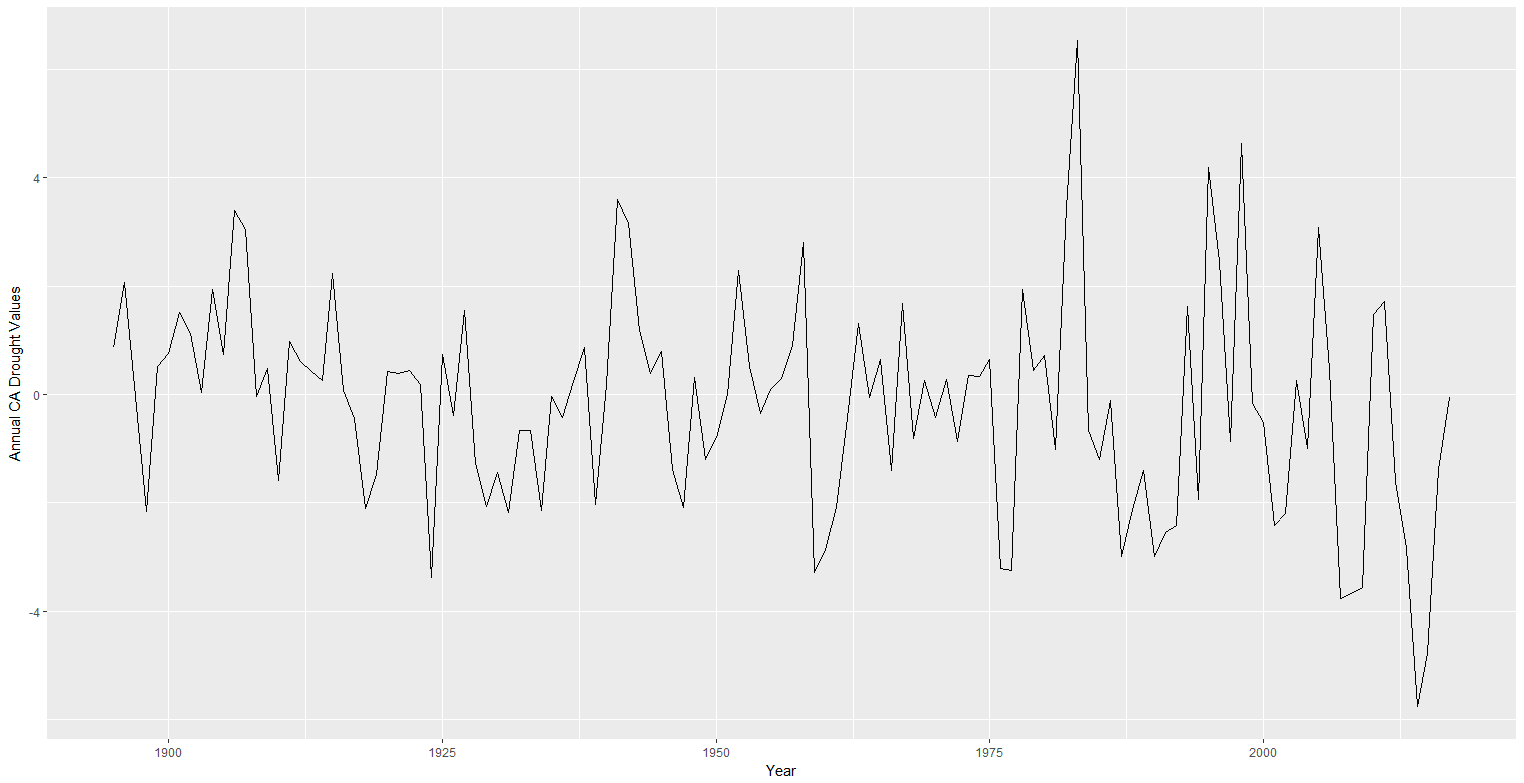
Initially, we look at the distribution of our data:



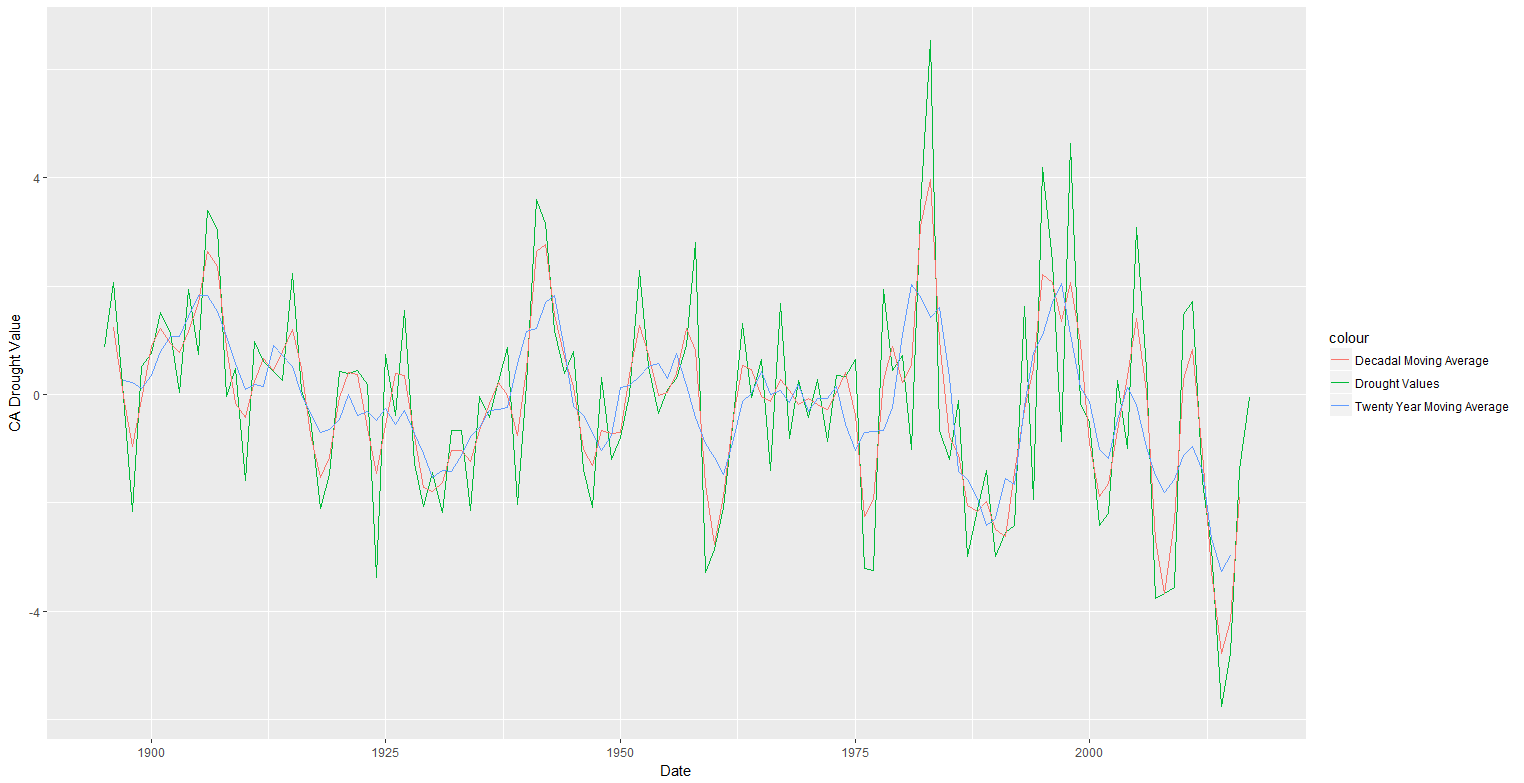
And then we plot the time series of our data, and examine for any outliers, volatility or irregularities. It looks like decline happening towards the end of the time series, and some volatility.



We used ts() to identify and replace outliers using series smoothing and decomposition. The cleaned version is not any different (probably because there were no outliers).

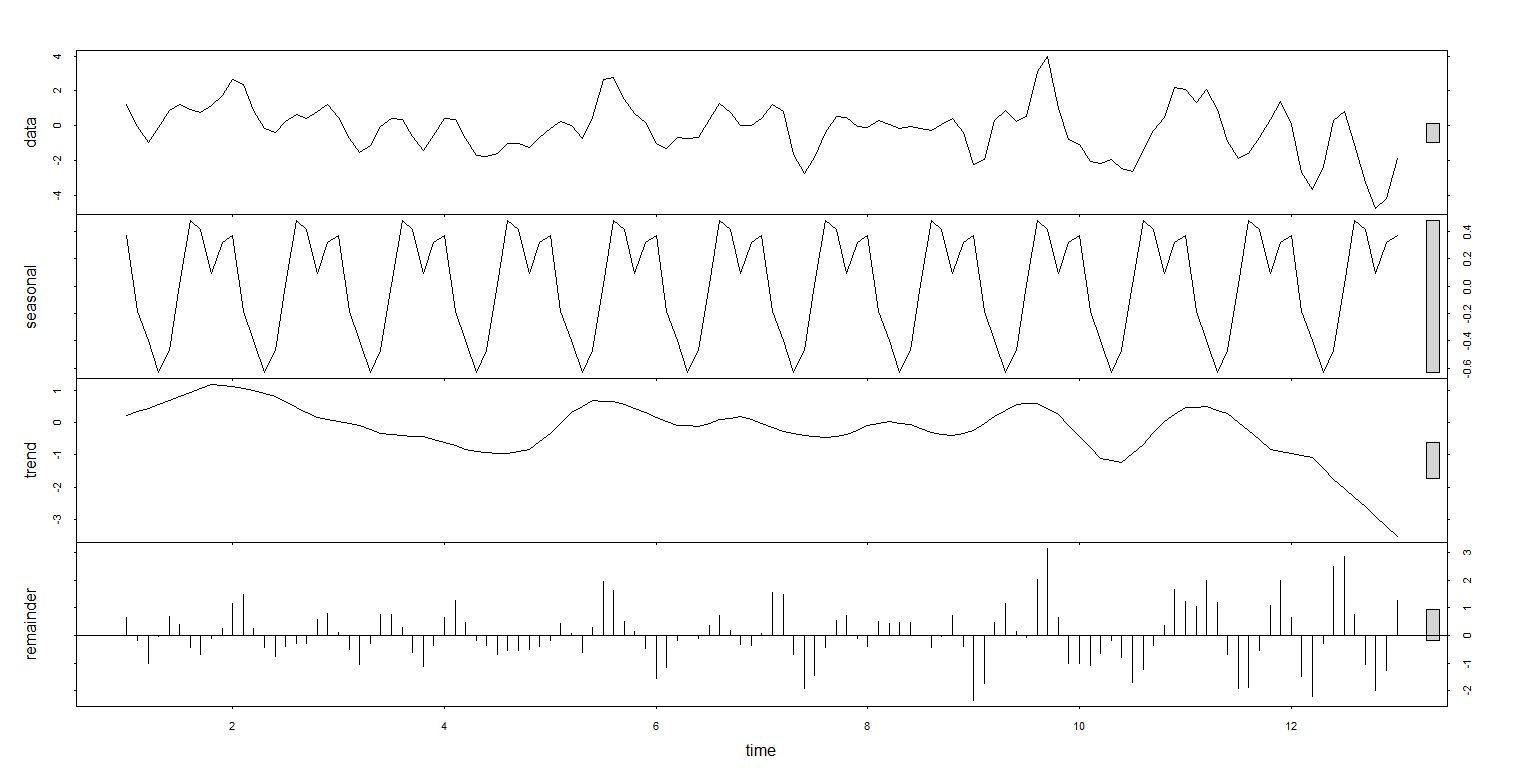


However, it looks like our time series is still pretty volatile, so we would like to smooth the data into a more stable predictable series. We will use moving average as a data smoothing technique (not the same thing as the moving average in the framework of the ARIMA model). For the window of the moving average, we chose a 2-year moving average (should we have chosen differently?) and a 5 year period. This smooths the series into something more stable and therefore predictable.



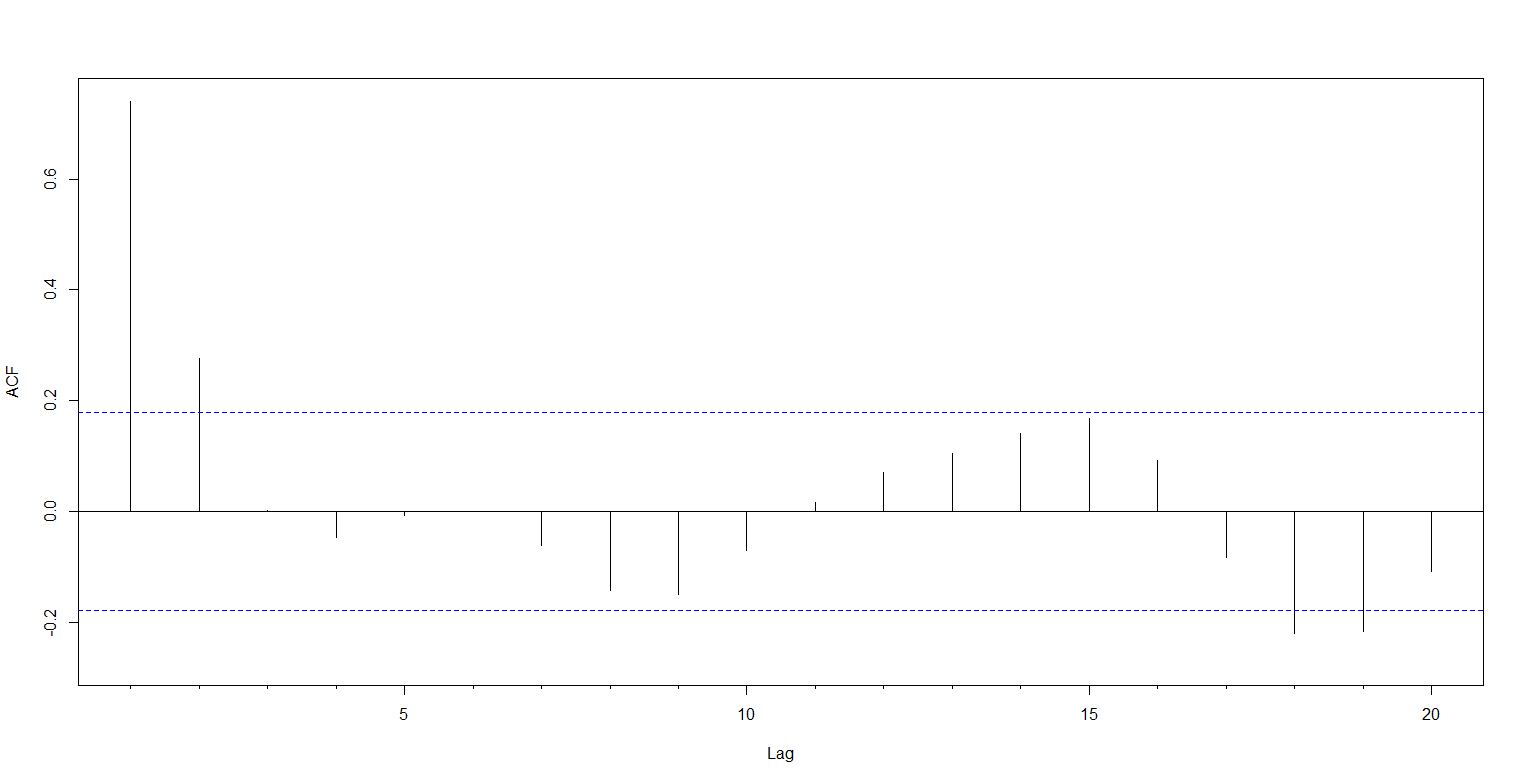
Now we decompose our data into seasonal, trend and cycle components. Part of the series that can't be attributed to any of these components is the residual or error. First, decide if we will use an additive or multiplicative model for modeling these components. Since the seasonality component seems to change with the trend of the series (more frequent seasonality with decreasing trend), multiplicative model would probably be best.

First, we will calculate seasonal component of the data. STL() calculates the seasonal component using smoothing and adjusts the original series by subtracting seasonality. What should we specify our periodicity to be? I chose decadal periodicity, but am really not sure. This is the result with a moving average smoothing of 2 years and a frequency of 10 years:

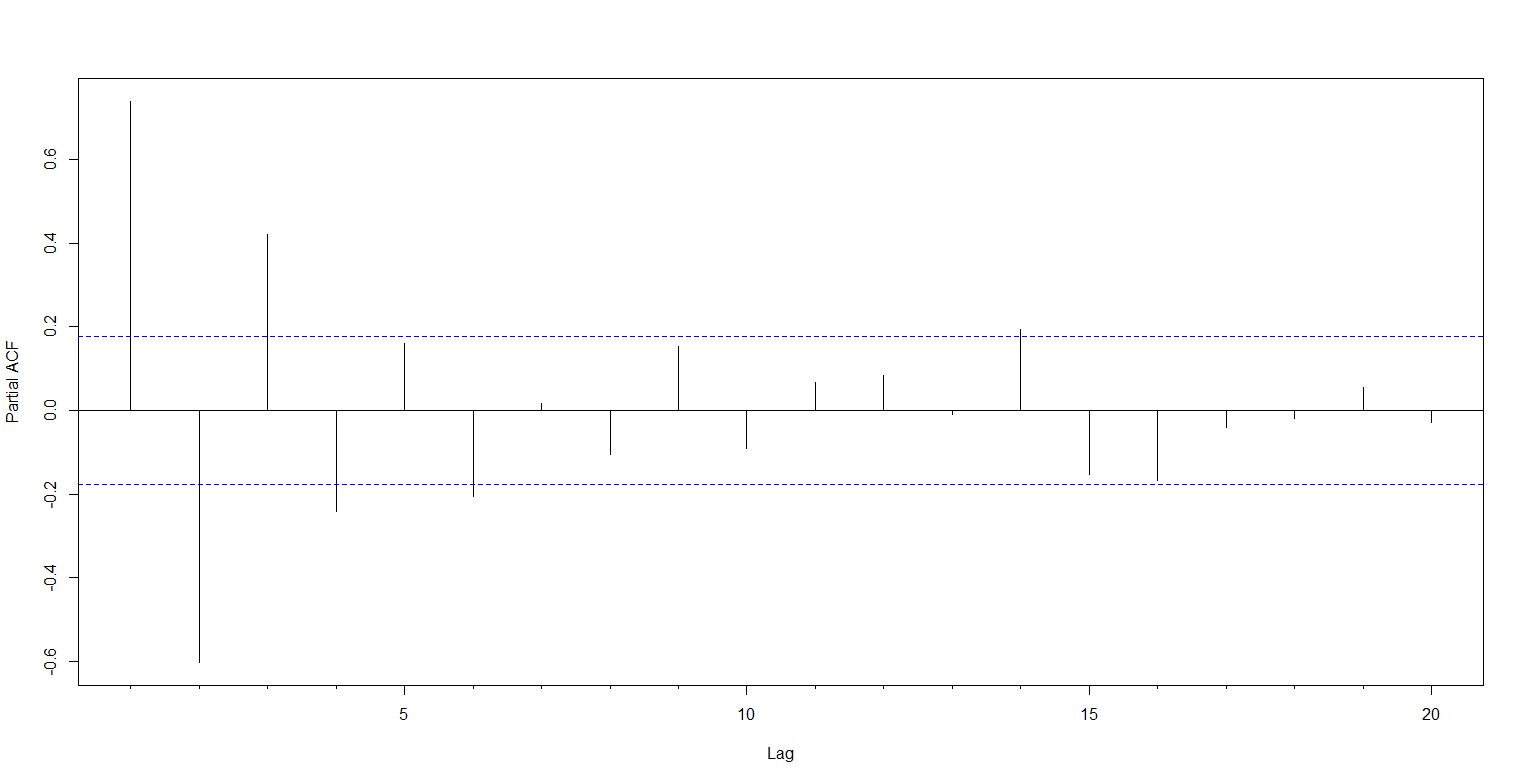


We now run the augmented Dickey-Fuller test to determine whether the time series is stationary. The null hypothesis is that the time series is non-stationary, and the p-value we found was 0.0283, thus our time series is stationary.

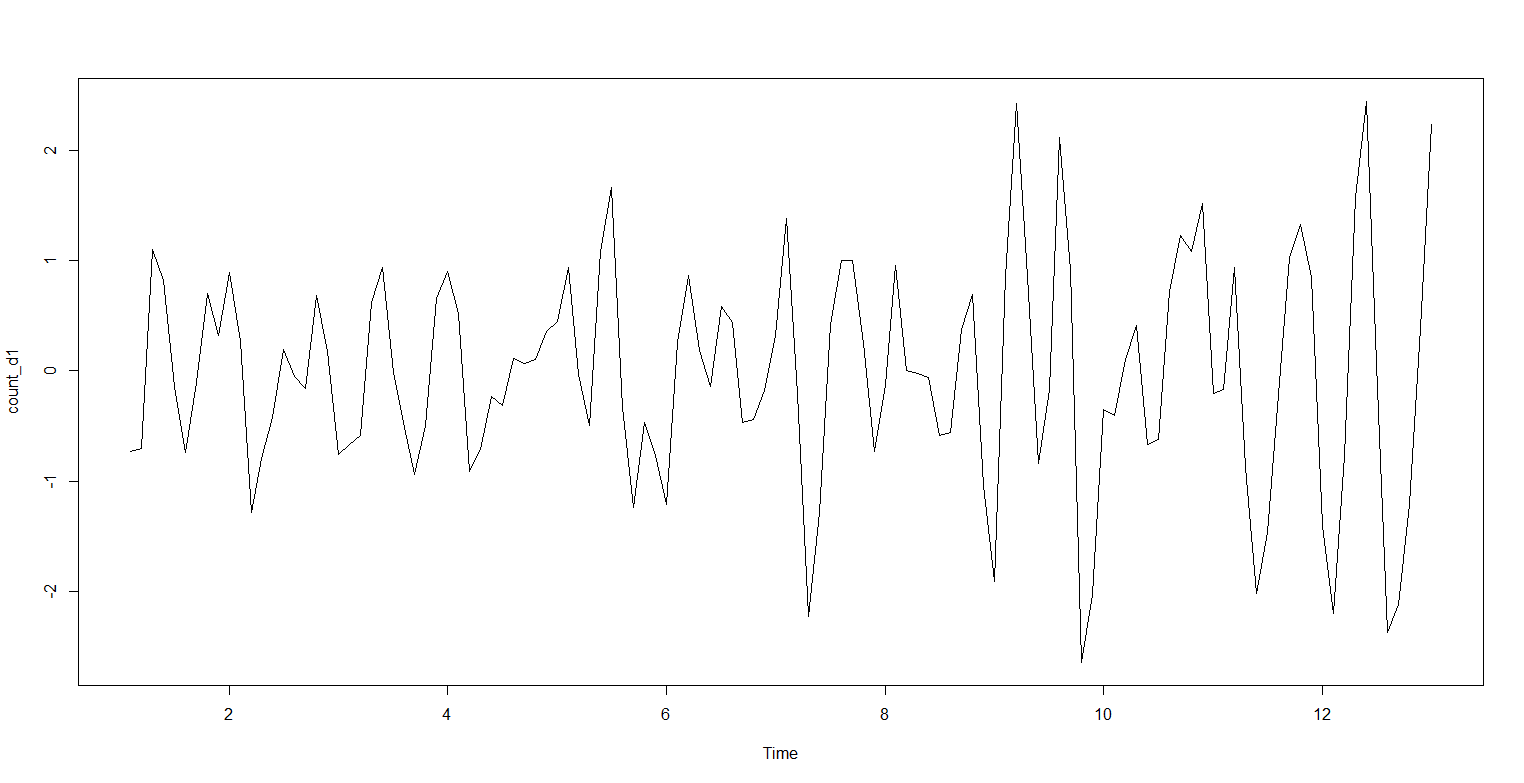
Now we take a look at the autocorrelation plots. The 95% significance boundaries are the blue dotted lines. It looks like there is significant autocorrelation with lag 1, 2, and 18 and 19.



The partial autocorrelation plot display correlation between a variable and its lags that is not explained by previous lags. The PACF shows autocorrelation at lag 1-4.



Start with order of d=1 and re-evaluate whether further differencing is needed. Plot the difference series, and see if there is a strong visible trend. If not, then differencing of order 1 terms is sufficient and should be included in the model.



Now we can run the auto.arima function which returns the best ARIMA model according to AIC, AICc and BIC:

Series: deseasonal\_CA.drought

ARIMA(3,1,2)

Coefficients:

ar1 ar2 ar3 ma1 ma2

1.0353 -0.6498 0.2094 -0.2078 -0.7380

s.e. 0.1068 0.1345 0.1062 0.0663 0.0584

sigma^2 estimated as 0.3608: log likelihood=-108.34

AIC=228.68 AICc=229.43 BIC=245.41

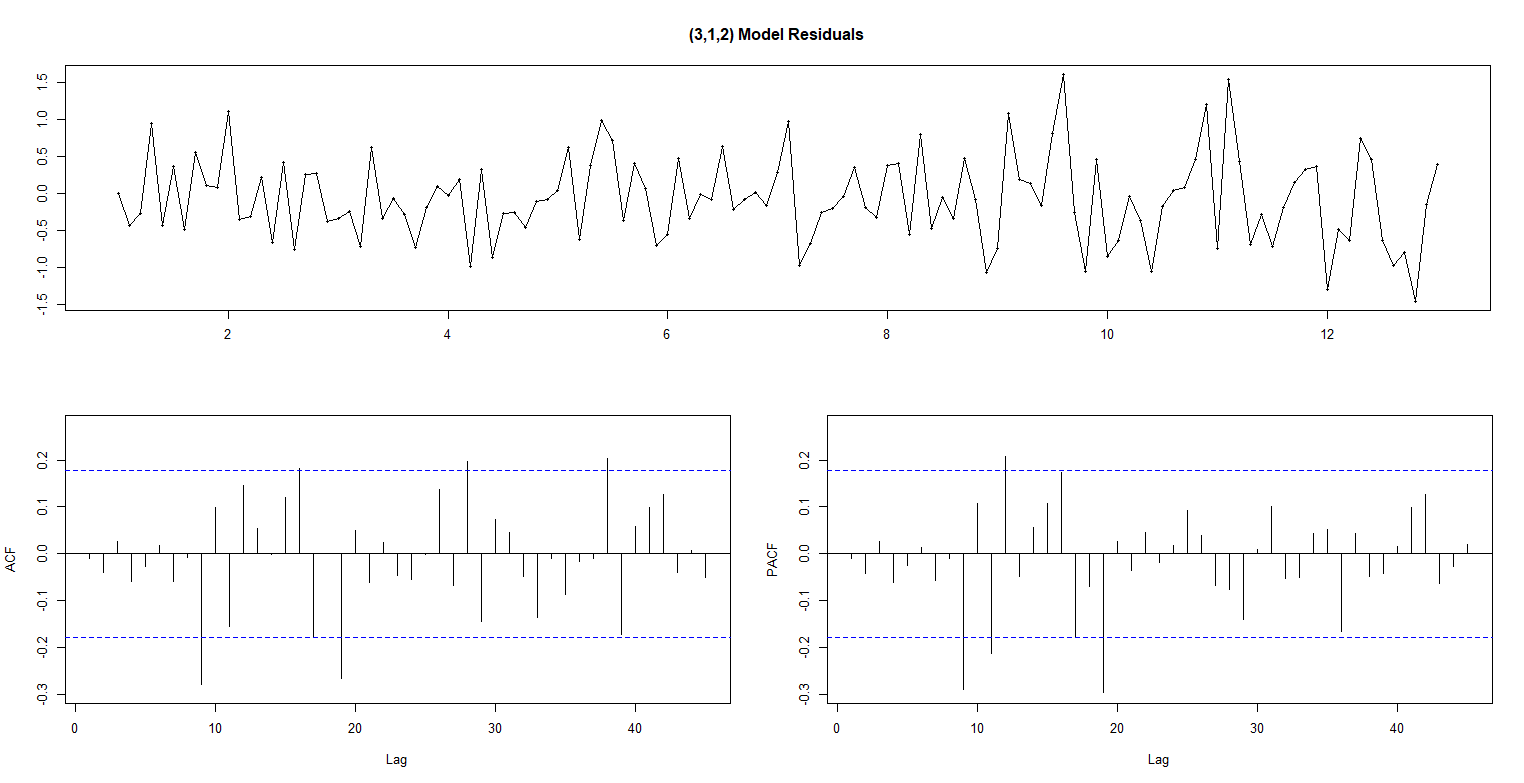
It looks like the best fitting

P=3 = autoregressive term of third lag

D=1 = model incorporates differencing of degree 1

Q=2 = moving average model of order 1

Can we trust this model though? We examine ACF and PACF plots for model residuals



It looks like there is a clear pattern in ACF/PACF with a lag 9 and 19. We can repeat the fitting process allowing for the MA(9) and MA(19) component and examine the diagnostics plots again.

arima(x = deseasonal\_CA.drought, order = c(3, 1, 9))

Coefficients:

ar1 ar2 ar3 ma1 ma2 ma3 ma4 ma5 ma6 ma7

0.7976 -0.3609 -0.2474 0.0572 -0.9544 0.3054 0.5368 -0.0501 -0.2299 -0.2010

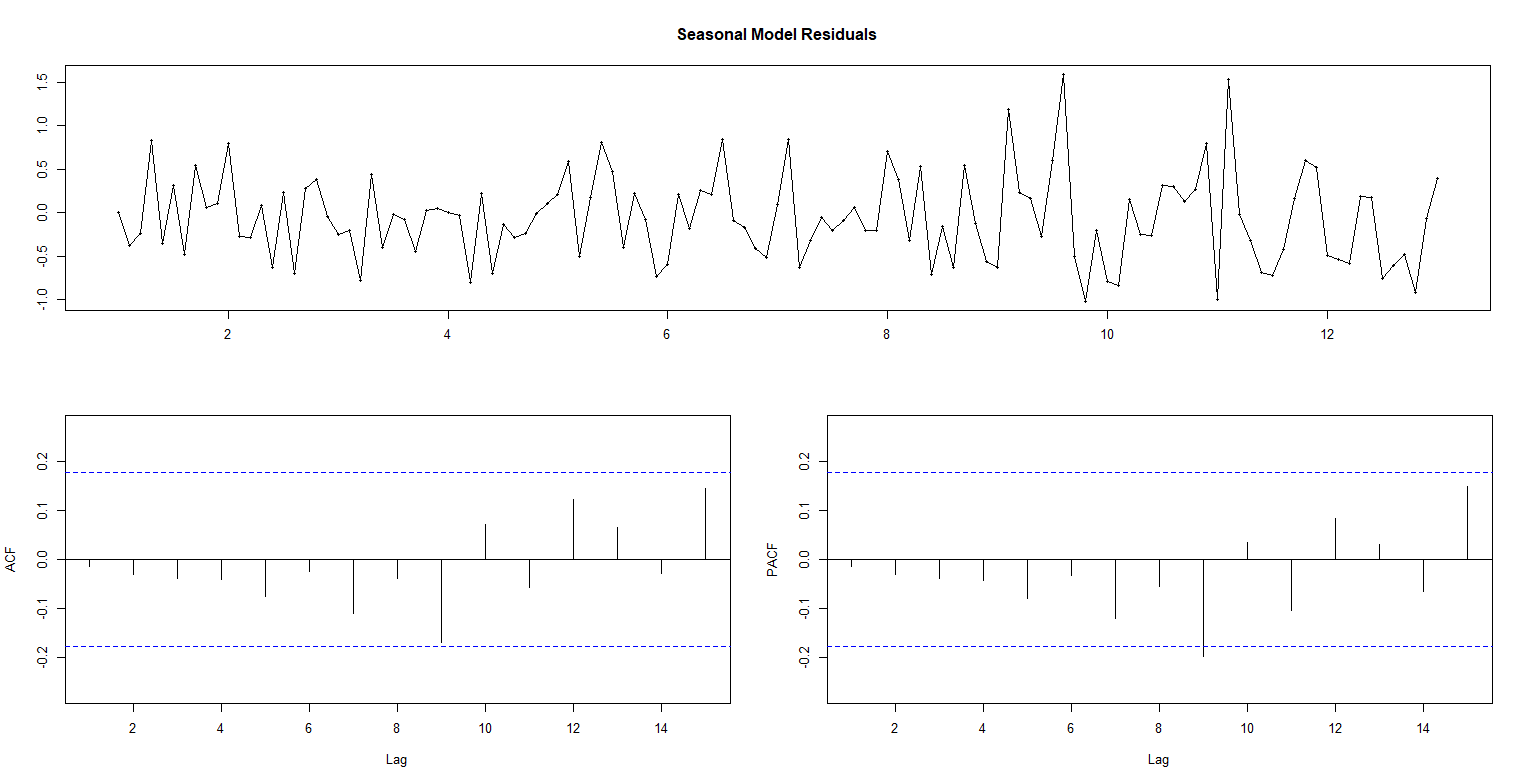
s.e. 0.3128 0.4303 0.2746 0.3080 0.2551 0.3268 0.3177 0.2022 0.1999 0.1916

ma8 ma9

-0.1053 -0.2525

s.e. 0.1631 0.1996

sigma^2 estimated as 0.2599: log likelihood = -94.8, aic = 215.61



This model fits better (no more significant autocorrelation).

However, we can also simply look at the California drought data using stl() and see that it is non-periodic.

> stl(drought$CA.drought)

Error in stl(drought$CA.drought) :

series is not periodic or has less than two periods

Then, we can skip all the steps above and simply run an ARIMA model to get a ARIMA(0,0,1) model.

> fit <- auto.arima(drought$CA.drought,xreg=drought$Date)

> fit

Series: drought$CA.drought

Regression with ARIMA(0,0,1) errors

Coefficients:

ma1 xreg

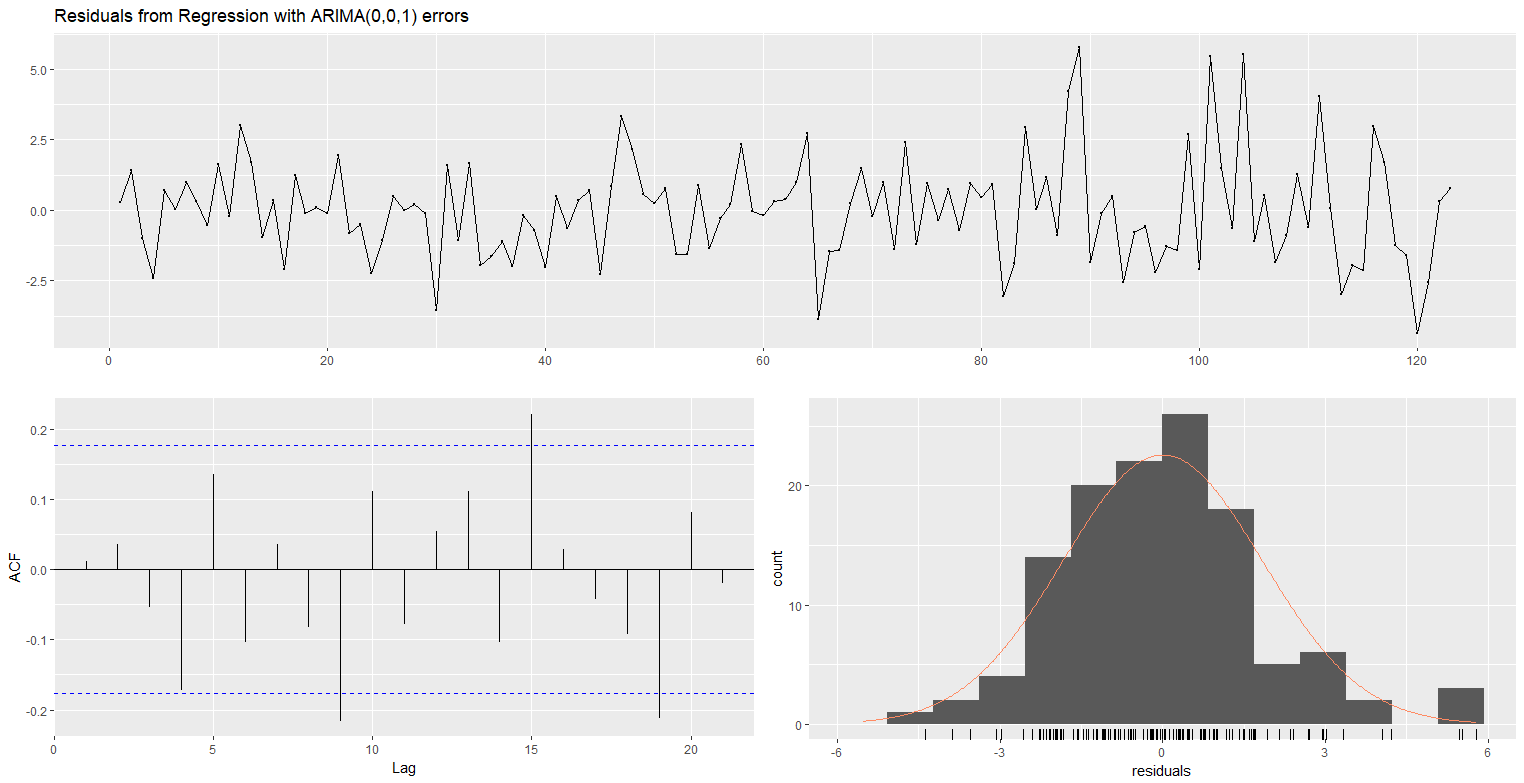
0.3033 -0.0124

s.e. 0.0853 0.0061

sigma^2 estimated as 3.436: log likelihood=-249.48

AIC=504.96 AICc=505.16 BIC=513.39

Then we can look to see how well the model fits by running checkresiduals()



It looks like the model fits well!