Katherine Olson

998992204

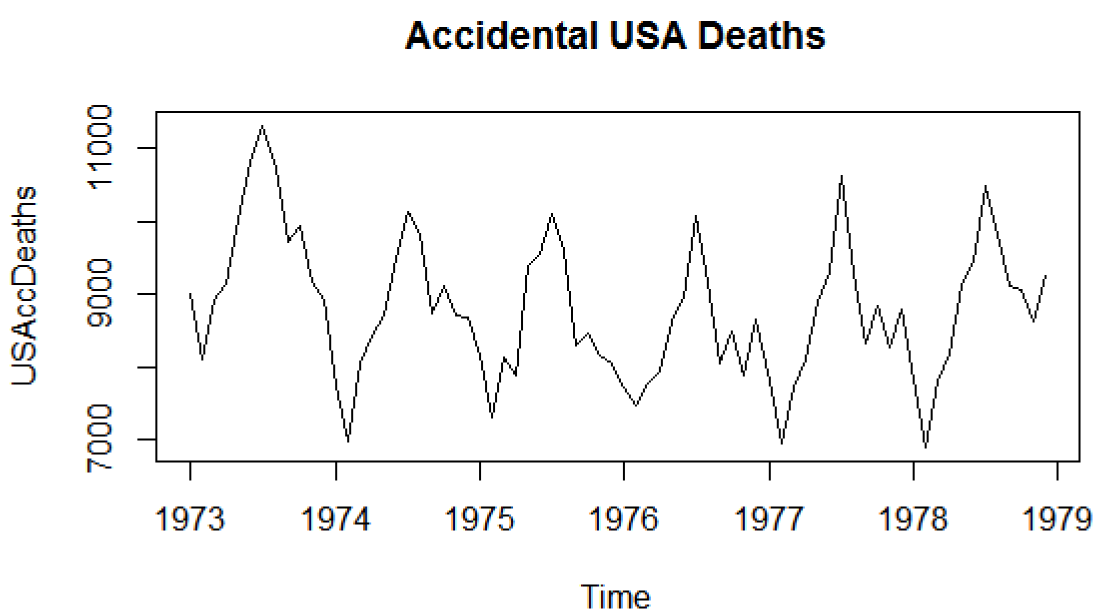
STA 137 Final Project

Introduction

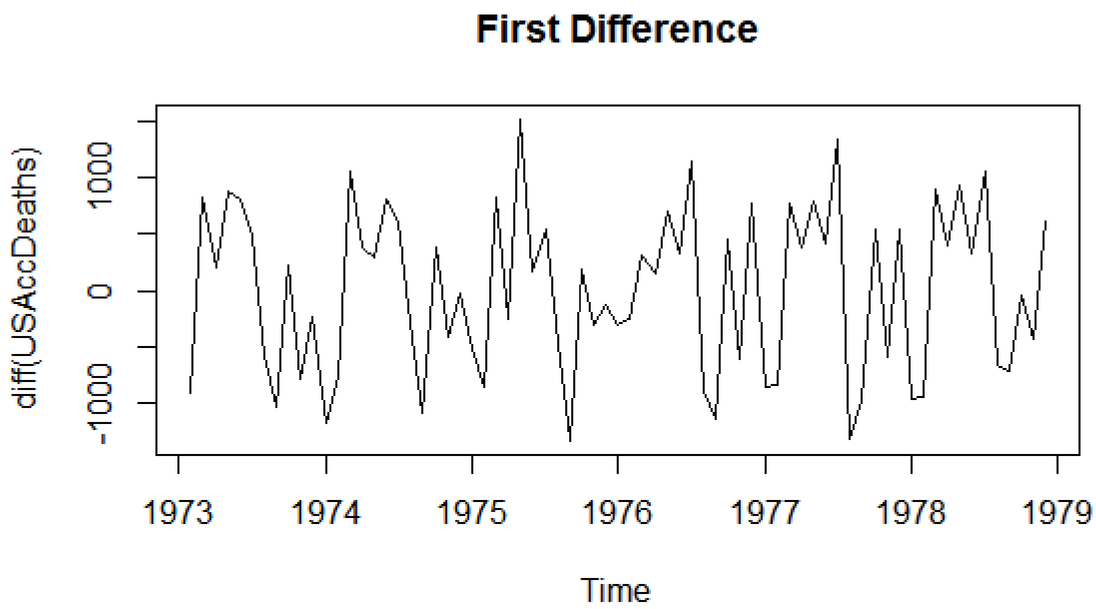
I am working with a time series dataset form the dataset R package called USAccDeaths. It contains the number of accidental deaths per month in the United States from 1973 to 1978. The dataset contains 72 observations. This is a time series dataset because it is taken over a period of time, in this case, per month. This dataset is important to analyze because we would like to lower the number of accidental deaths. So if the data shows some sort of pattern that can help solve why the accidental deaths are happening, that would be very helpful. It will be especially useful to see if the dataset has a seasonal component. If the accidental deaths are higher in some months, there is interest in figuring out why.

Materials, Methods, and Results

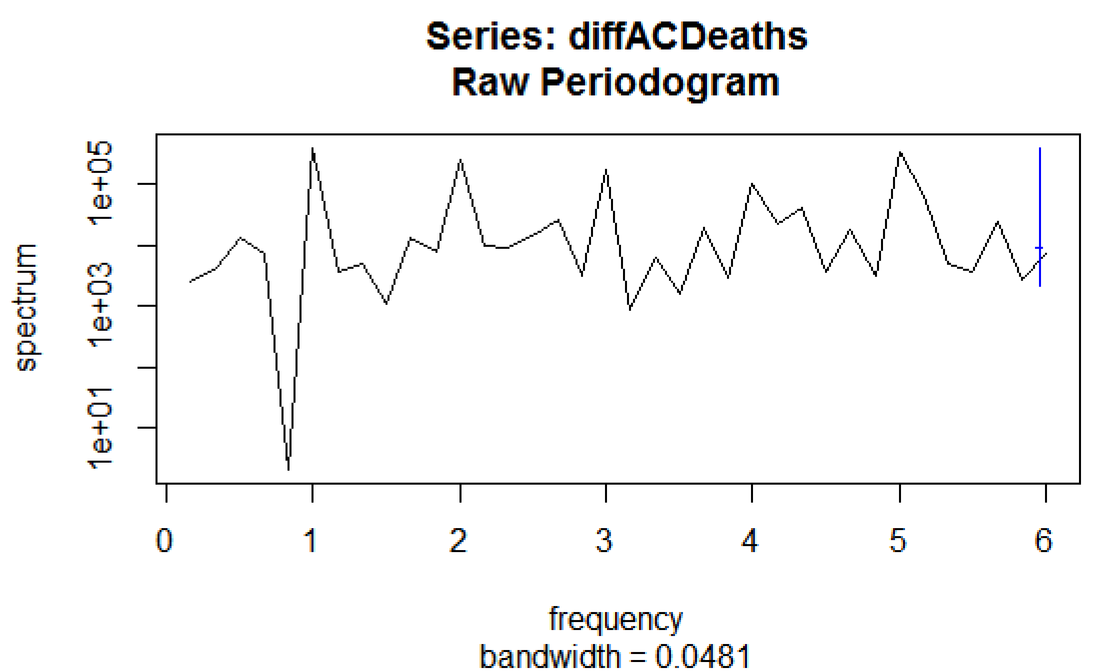
I started by plotting the series:



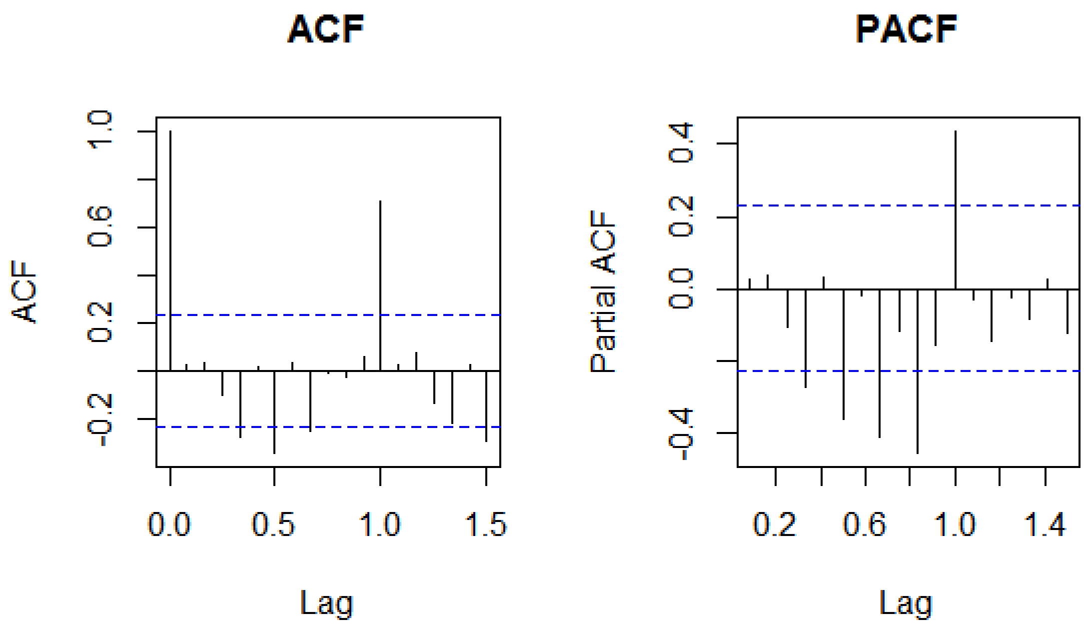
Looking at the series, there is definitely a seasonal component, which we will come back to later. The series looks mostly stationary, but at some points it looks like the mean function is a bit higher than other points. It looks higher at the edges and lower in the middle. Do to deal with this small dependency on time, I added a first difference. This plot looks more stationary. The variance appears to be about constant throughout, so I do not think that any variance smoothing transformations are necessary.



Next I looked at the periodogram to examine the periodicity of the data. There is a large drop right before frequency = 1. There are small peaks at each increment of one. But otherwise, it looks smooth.



Then I looked at the ACF and PACF:



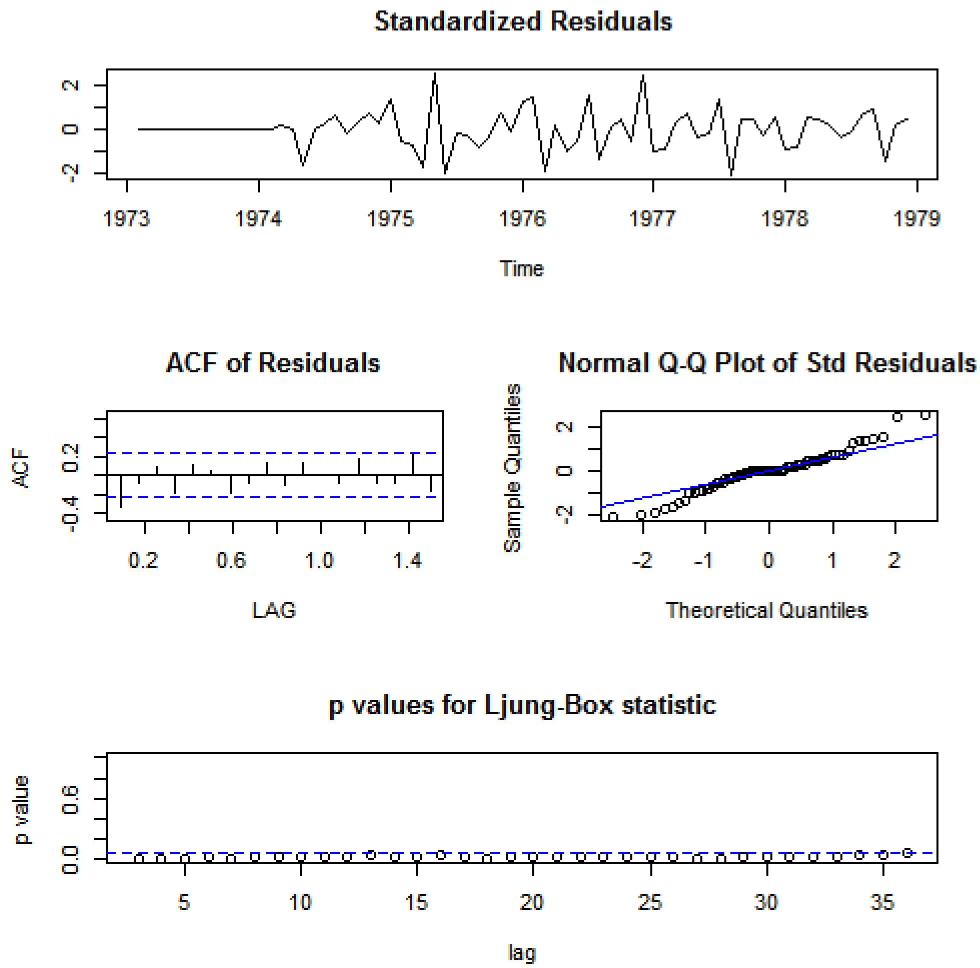
Starting the AR and MA parts of fitting a model, the ACF has significance at lag 1 with about 0.7 and appears to cut after that. The PACF appears to tail off. So the ACF and PACF suggest that the model is MA(1). Then, since we already took a first order difference we have ARIMA(0, 1, 1). Now we need to look at the seasonal component. The ACF cuts off after 1 lag. So Q = 1. We know s = 12 from looking at the original plot and given the fact that we are working with monthly data. The PACF only goes for so many lags, but we can tell that there is a cut off after about we lags around lag = 1.0. So both parts point to a seasonal component of $(0, 1, 1)_{12}$. As a whole, the PACF and ACF suggest the model ARIMA(0, 1, 1)X$(0, 1, 1)_{12}$.

Next, I used AIC to pick between a few models. The PACF and ACF had a clear direction for the model, but I wanted to look at some other similar models to be sure.

Model	**AIC**
ARIMA(0, 1, 1)X$(0, 1, 1)_{12}$	12.72764
ARIMA(0, 2, 1)X$(0, 2, 1)_{12}$	13.74445
ARIMA(0, 1, 1)X$(0, 2, 1)_{12}$	13.26676
ARIMA(0, 2, 1)X$(0, 1, 1)_{12}$	13.76663

So the final best model is ARIMA(0, 1, 1)X$(0, 1, 1)_{12,}$ with the smallest AIC.

Fitting that model we have:



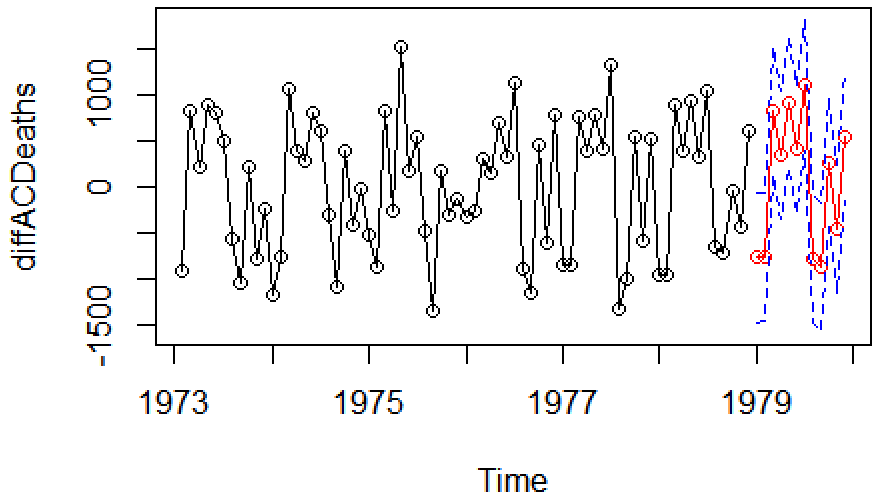
The residuals look mostly like white nose. The ends are a little off, but nothing bad.

Expressing the final model in terms of the estimated parameters:

MA(1)	sMA(1)
-1.0	-0.5548

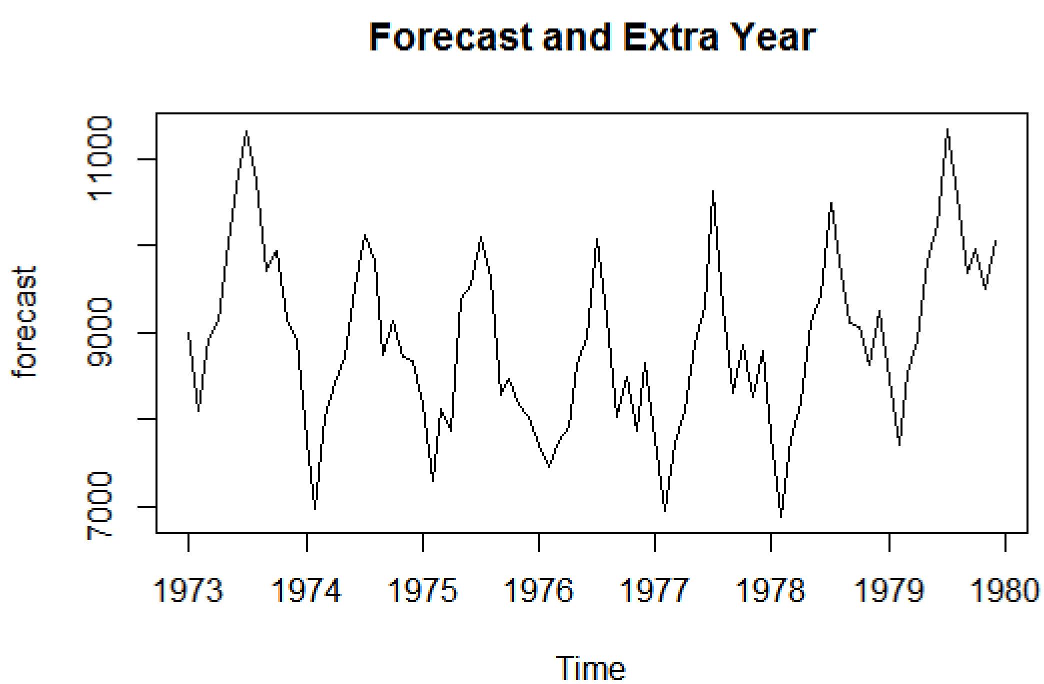
Sigma^2 = 117160

Next, I forecasted future values for the next 12 months:

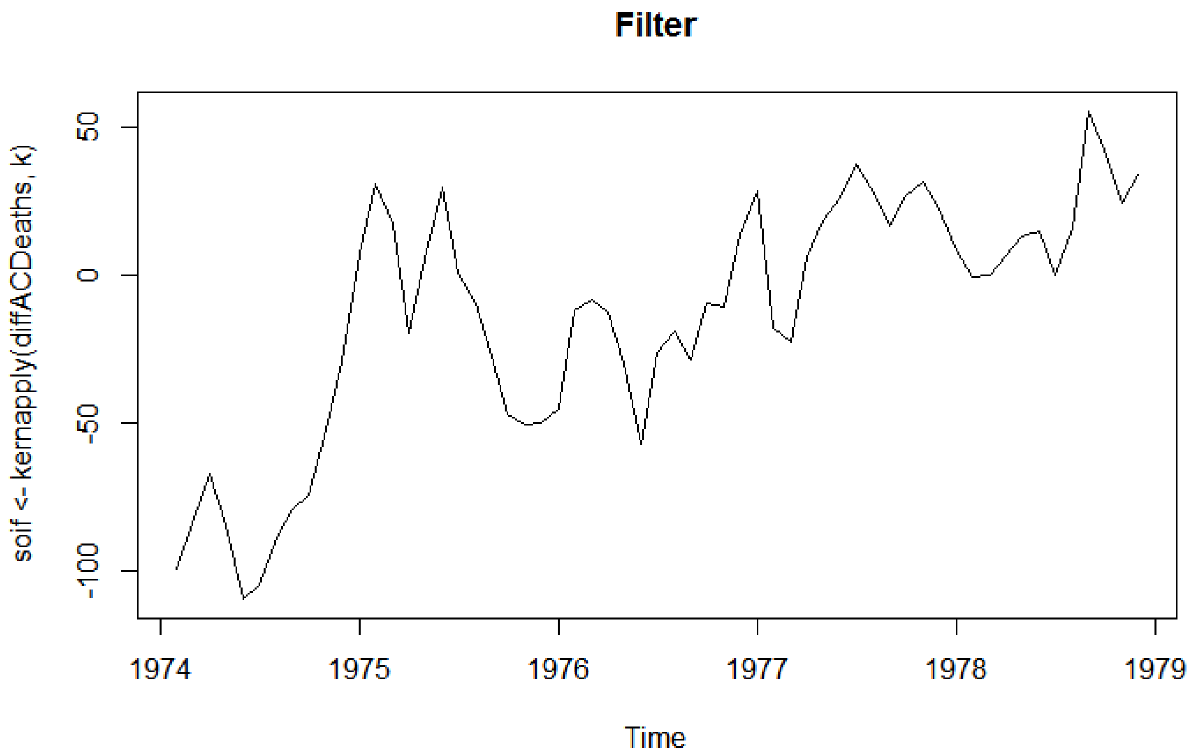


Un-differencing the data we get the values: 8471.388, 7716.100, 8547.165, 8897.285, 9817.704, 10237.469, 11332.815, 10560.700, 9686.768, 9955.184, 9504.237, and 10043.178.

Then adding them to the original plot:



I applied a modified Daniell kernel filter with m = 6:



Conclusion and Discussion

I used a time series of accidental deaths per month in the US from 1973 to 1978 for exploratory data analysis. Right away there was a clear seasonal pattern at s = 12. To make the data stationary, I applied a first difference. I then looked at the periodogram. Looking at the ACF and PACF, the suggested model was ARIMA(0, 1, 1)X(0, 1, $1)_{12}$, the residuals looked good, and that model was the final model selected using AIC to compare with a few other models. I expressed the final model in terms of the estimated parameters. I then applied a linear filter. Approaching this dataset, I was very interested in finding a seasonal pattern and I found a yearly one. This seasonal pattern should be useful in providing insight into why accidental deaths are higher in certain times during the year. It appears from the original plot that accidental deaths are higher in the summer. Finding the cause for this would be very useful, and hopefully help lower the accidental deaths overall.

Code Appendix

```
library(datasets)
library(astsa)
length(USAccDeaths)
ts.plot(USAccDeaths, main = "Accidental USA Deaths")
ts.plot(diff(USAccDeaths), main = "First Difference")
diffACDeaths = diff(USAccDeaths)
spec.pgram(diffACDeaths)
par(mfrow = c(1, 2))
acf(diffACDeaths, main = "ACF")
pacf(diffACDeaths, main = "PACF")
sarima(diffACDeaths, 0, 1, 1, 0, 1, 1, 12)
sarima(diffACDeaths, 0, 2, 1, 0, 1, 2, 12)
sarima(diffACDeaths, 0, 1, 1, 0, 2, 1, 12)
sarima(diffACDeaths, 0, 2, 1, 0, 1, 1, 12)
fit = sarima(diffACDeaths, 0, 1, 1, 0, 1, 1, 12)
par(mfrow = c(1, 1))
forecast = sarima.for(diffACDeaths, 12, 0, 1, 1, 0, 1, 1, 12)
preds = forecast$pred[1:12]
addPred = c(diffACDeaths, ts(preds, start = 1979))
addPred2 = ts(addPred, start = c(1973, 2), end = c(1979, 12), frequency = 12)
forecast = diffinv(addPred2, xi = 9007)
plot.ts(forecast, main = "Forecast and Extra Year")
forecast[73:84]
k = kernel("modified.daniell", 6) # filter weights
plot(soif <- kernapply(diffACDeaths, k), main = "Filter") # plot 12 month filter
```