# STA 137 Final Project

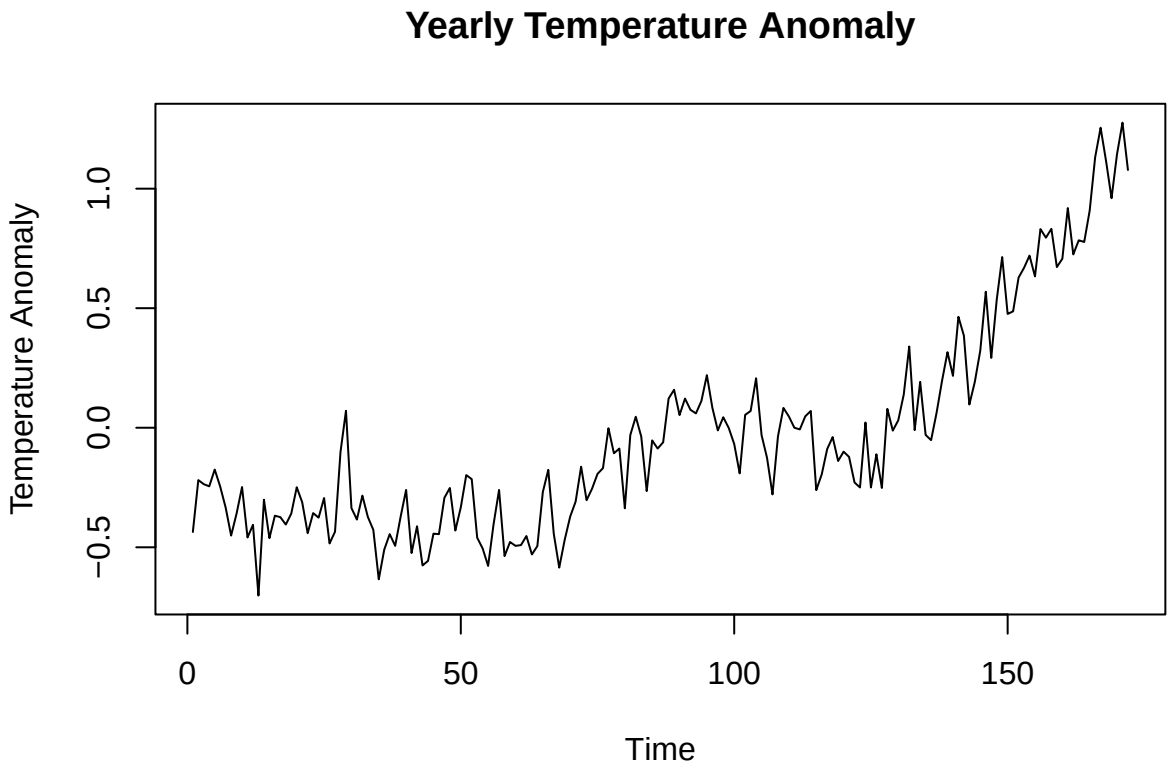
Brian Le 916111559, Aurian Saidi X971615

2022-12-04

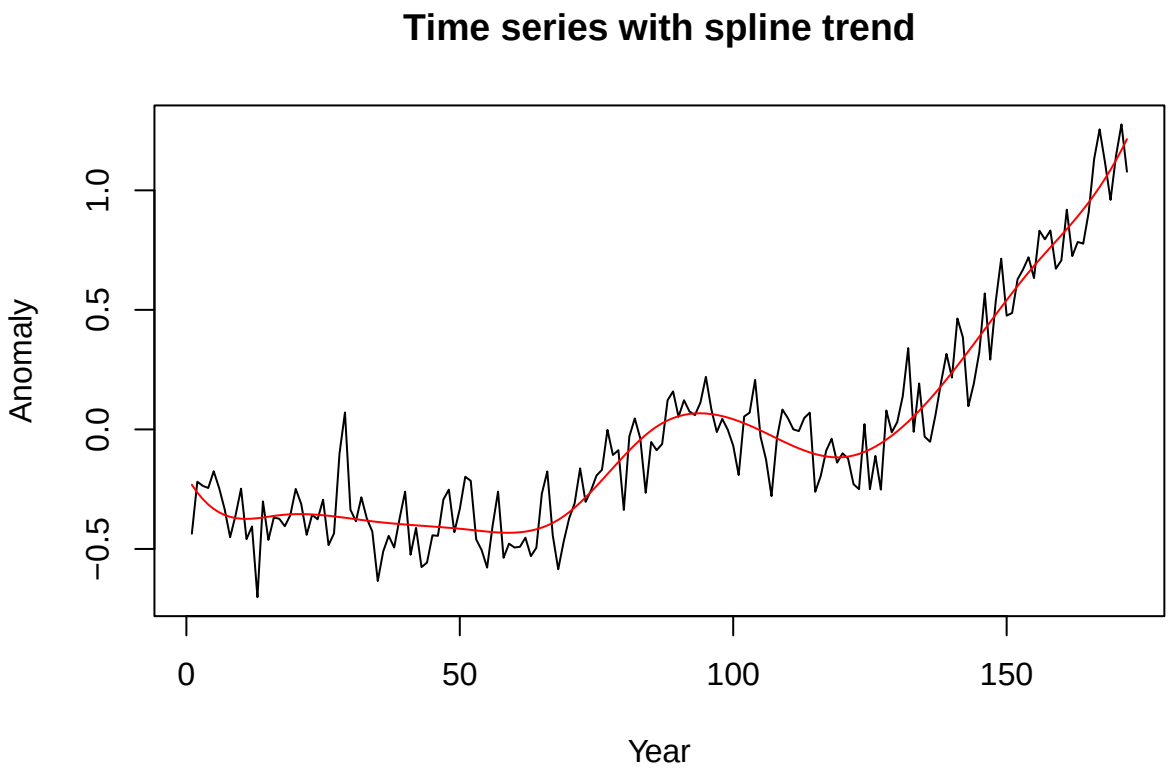
## Introduction and Dataset

For this project, we will be analyzing yearly temperature anomalies between 1850 and 2021 for the northern hemisphere. Our dataset is sourced from the Climate Research Center at the University of East Anglia, UK. Temperature anomalies were collected yearly, for a total of 172 data points. Each data point represents the measured temperature for that year subtracted from a baseline average temperature. We will be using time series statistical techniques to model this data.

## Graphical Analysis

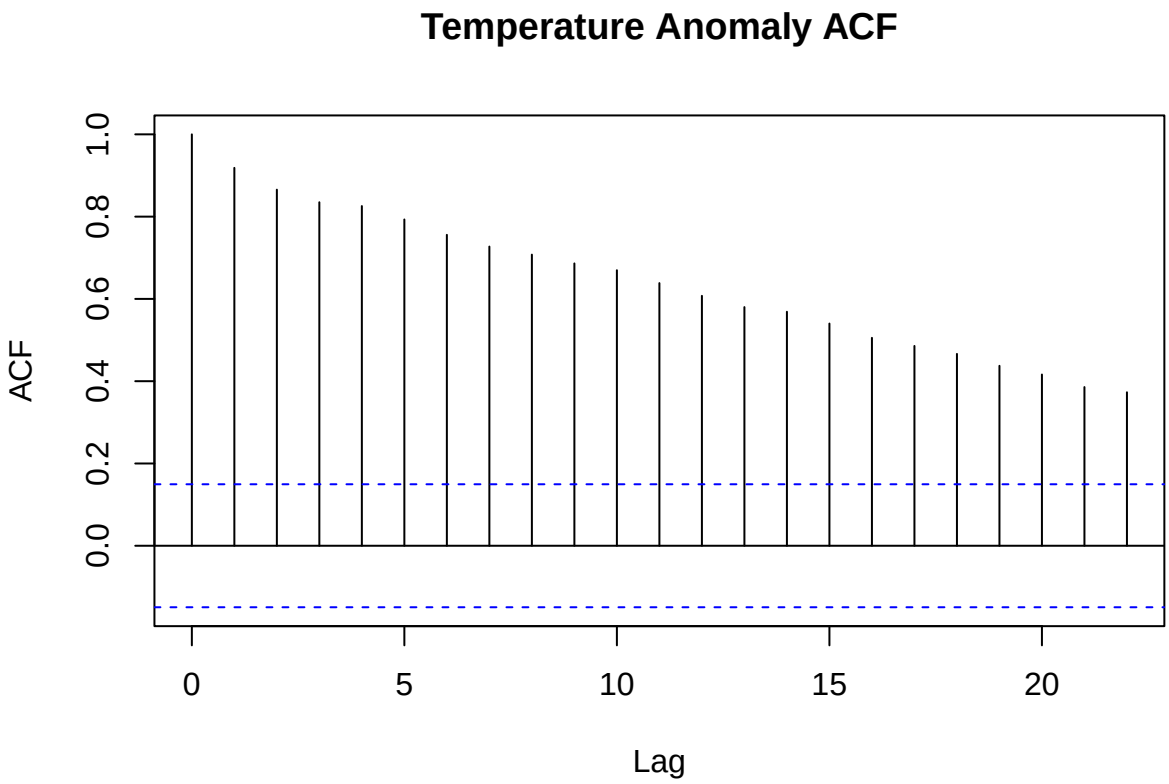



From the plot of the data, we can see that there were periodic changes in temperature anomalies every couple of years. However, the trend appears to be constant until about t = 75, where it starts to increase over time. The data does not appear to fluctuate around a constant mean, which does not satisfy a condition for stationarity.



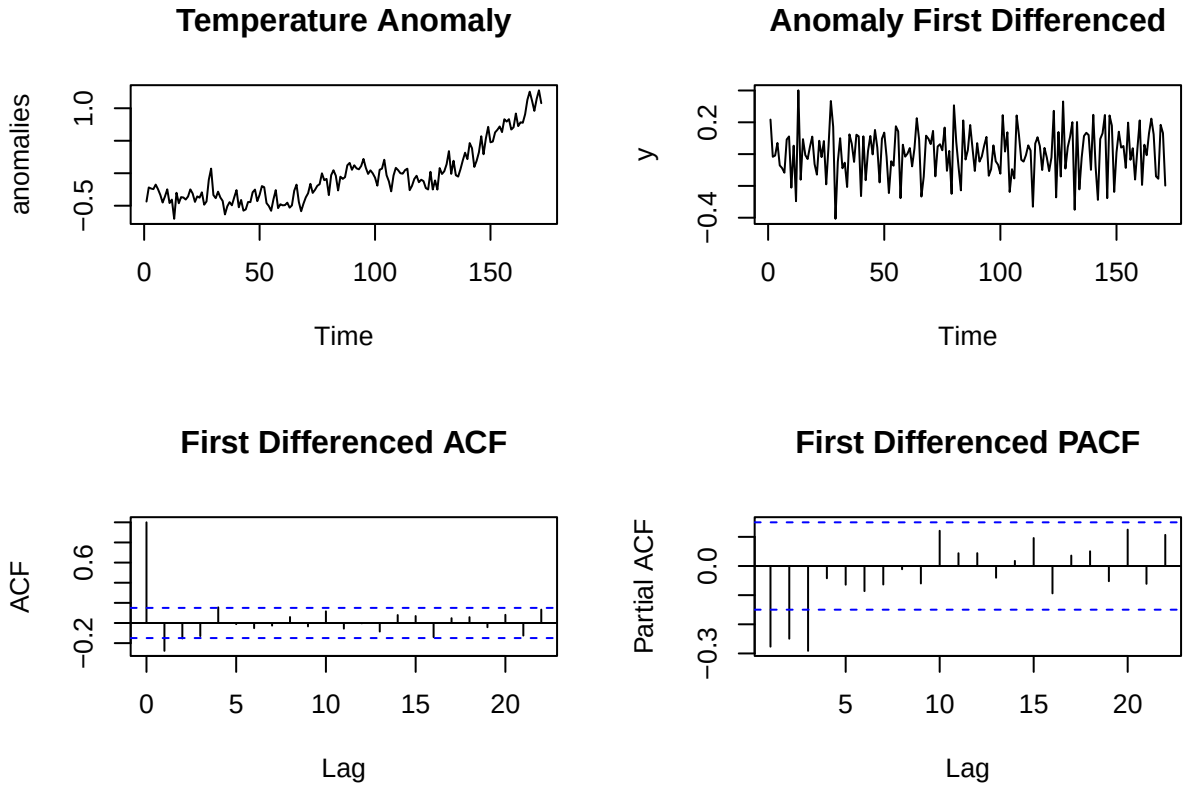

To obtain the trend, we fit a spline trend curve. By fitting a curve to the data, we can get a better idea of how the data behaves. Initially, the temperature anomaly is stationary, with a mean of about -0.3 and constant variation around that mean. Over time, we see that the mean increases, to about 0.1 at t = 100 (year 1950). We also see a large linear spike starting at t = 120 (year 1970). The variation around the curve appears constant.

Because there isn't a constant mean for the data, we will be applying differencing to remedy this. That is, we will consider the difference between observations of temperature anomaly and its previous observation.

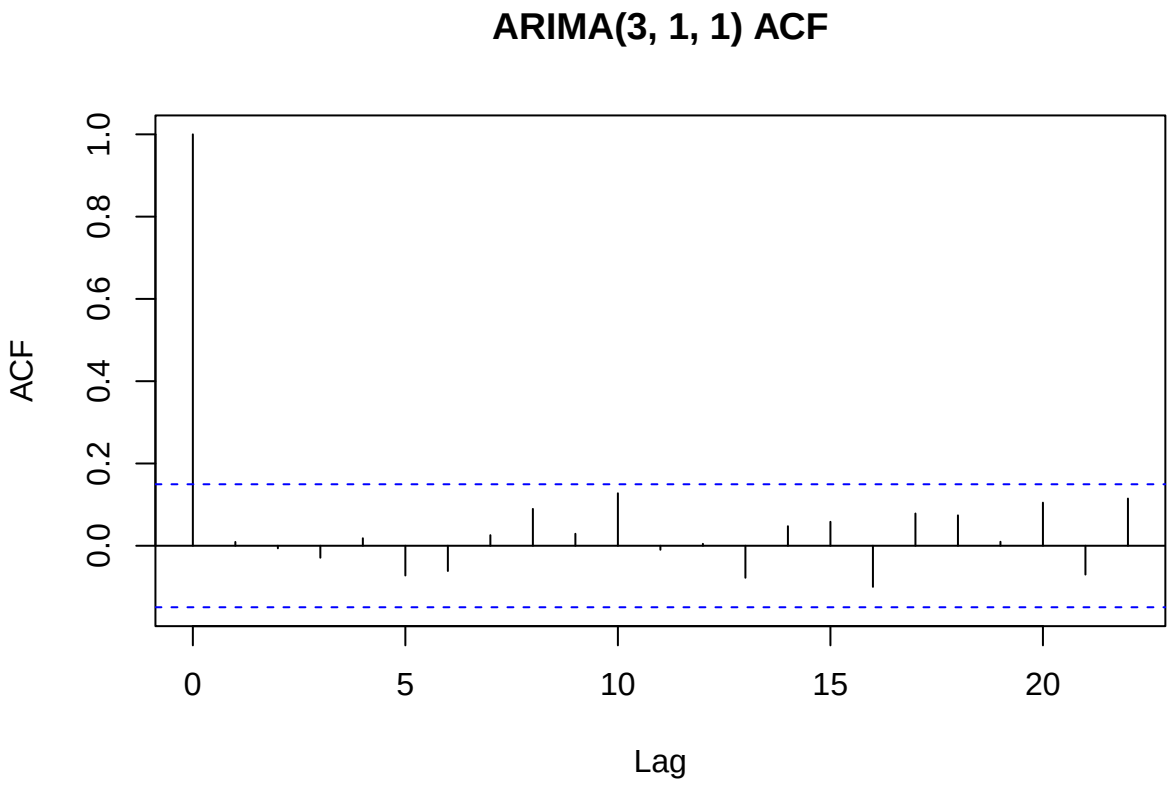



A non-stationary time series' ACF will decrease slowly, like we see here, so we will need to correct this for further analysis.

## Modeling




We see that the first differenced series does not have an increasing trend like our original series. The fluctuations appear to be constant around the trend , so we will use this differenced series to continue our analysis. We can see that the ACF plot is insignificant after lag 1 and the PACF plot cuts off at lag 3 due to the large difference between lag 3 and lag 4. Based on these plots, our preliminary model for analysis is ARIMA(3, 1, 1).

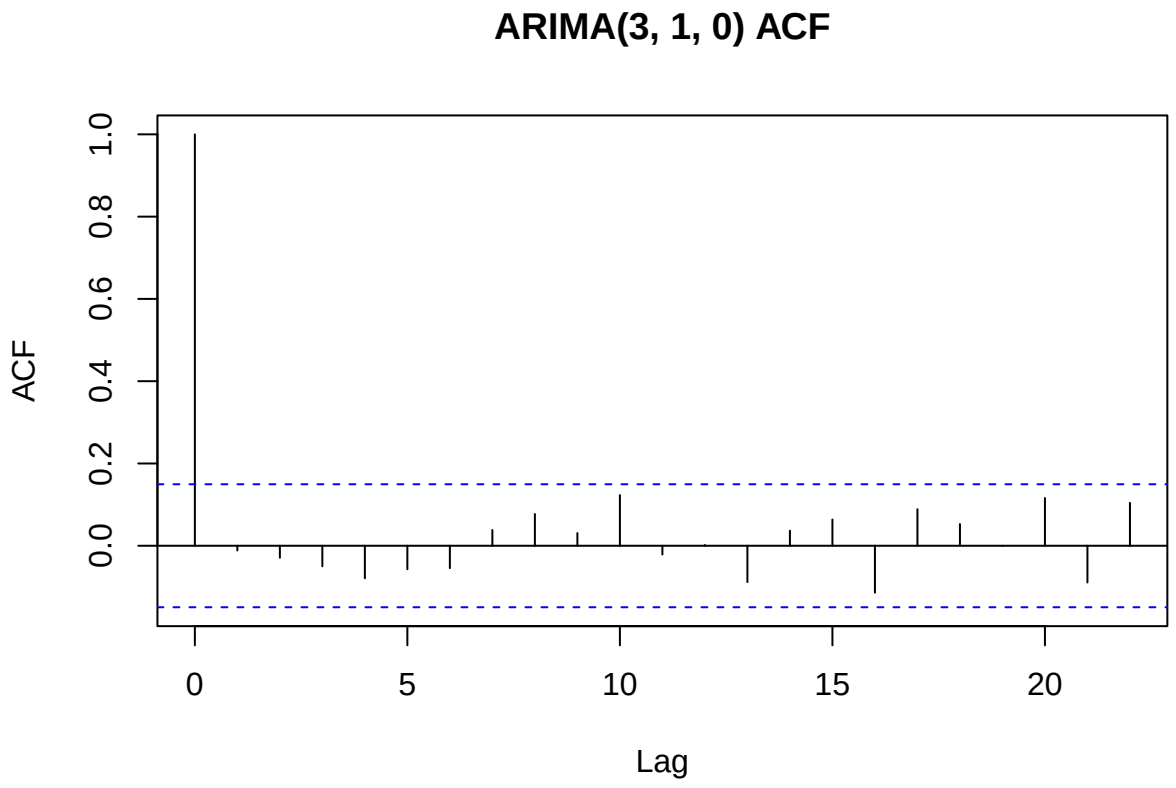



We see that all of the lags lie within the confidence bands, so the residuals resemble white noise. The ARIMA(3, 1, 1) model is reasonable. However, to confirm the best model, we will consider ARIMA(p, 1, q) models, where $p = 0, ..., 3$ and $q = 0, ..., 3$, selected using the AIC criterion.

Table 1: AIC Table

| | | | |
|---|---|---|---|
| -0.9228030 | -1.093587 | -1.119037 | -1.106939 |
| -0.9918207 | -1.113262 | -1.106827 | -1.107563 |
| -1.0447157 | -1.113521 | -1.110429 | -1.111585 |
| -1.1219867 | -1.114391 | -1.107756 | -1.099317 |

Based on the AIC values, an ARIMA(3, 1, 0) is the best fit since it has the lowest AIC.




The residuals resemble white noise, so the ARIMA(3, 1, 0) model is reasonable.

Table 2: ARIMA(3, 1, 0) Parameter Estimates

| | Estimate | SE | t.value | p.value |
|---|---|---|---|---|
| ar1 | -0.4234 | 0.0735 | -5.7590 | 0.0000 |
| ar2 | -0.3557 | 0.0756 | -4.7063 | 0.0000 |
| ar3 | -0.2947 | 0.0735 | -4.0092 | 0.0001 |
| constant | 0.0085 | 0.0050 | 1.7199 | 0.0873 |

Using the first order differenced series $Y_t$, the final model is:

$Y_t = \mu + \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \phi_3(Y_{t-3} - \mu) + \epsilon_t$

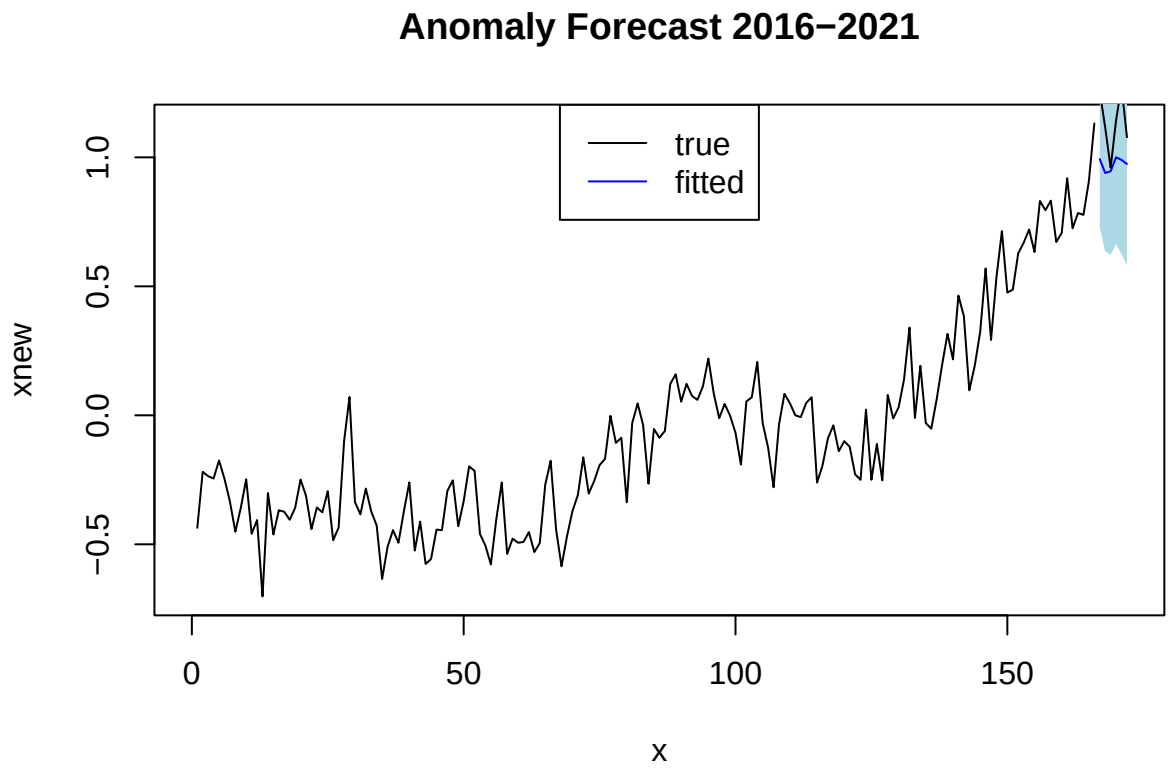
which is an third order autoregressive sequence.

The parameter estimates with their standard errors are:

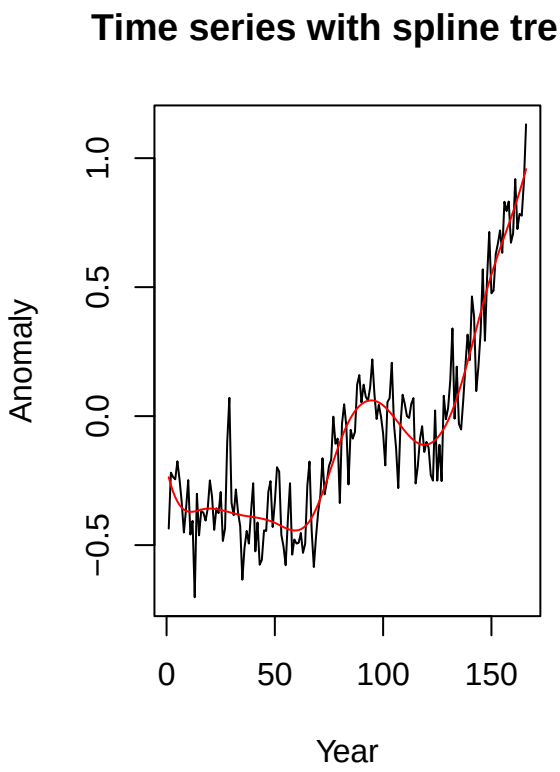
- $\mu = 0.0085, SE = 0.0050$
- $\phi_1 = -0.4234, SE = 0.0735$
- $\phi_2 = -0.3557, SE = 0.0756$
- $\phi_3 = -0.2947, SE = 0.0735$

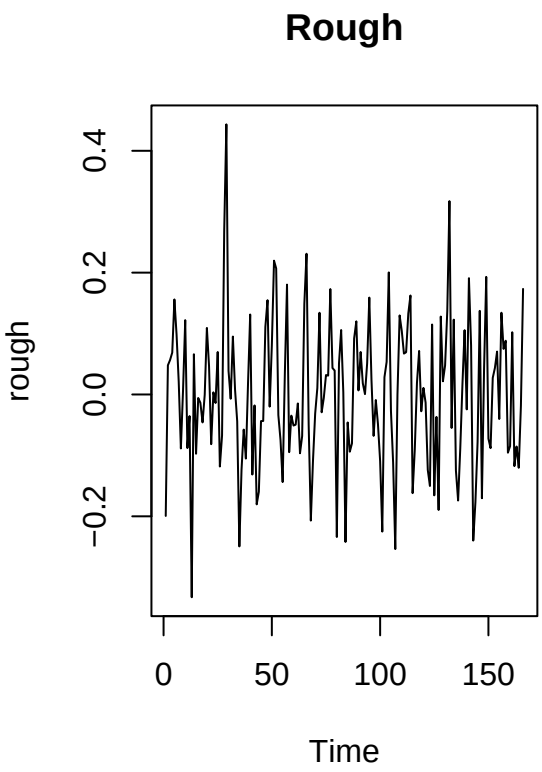
### Forecasting

To forecast the last six years of the data, we will fit an ARIMA(3, 1, 0) model to the the data, excluding the last six years.

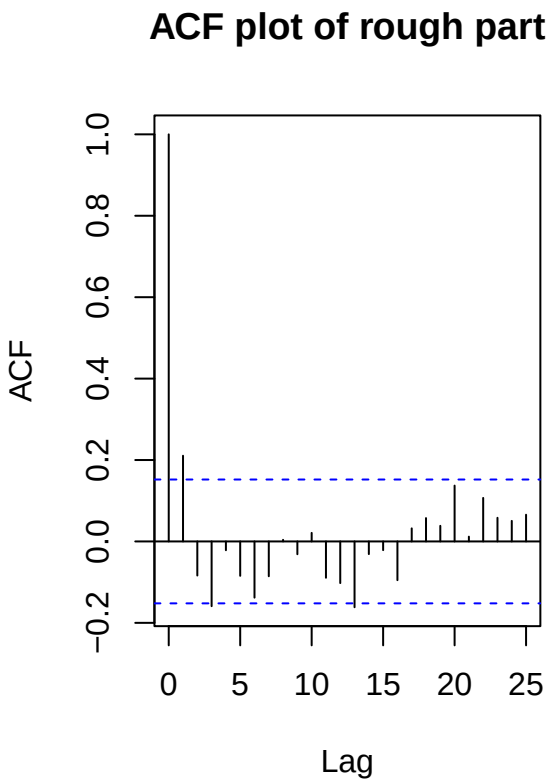


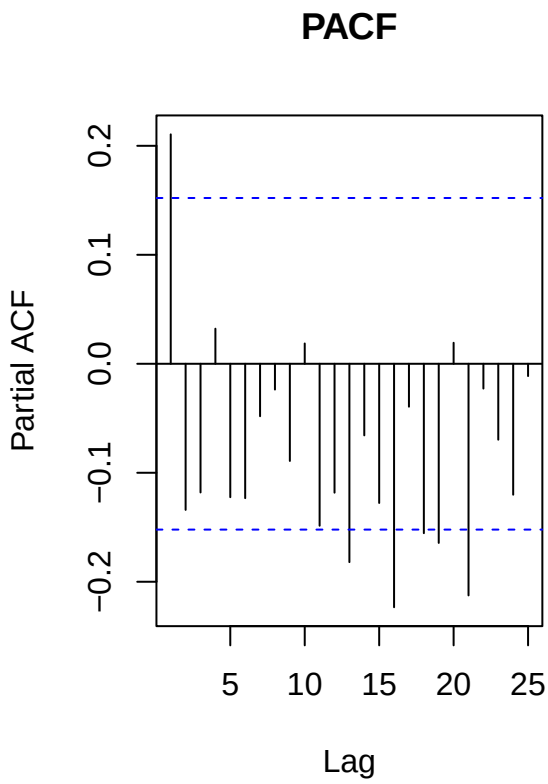

The plot shows the observed data and the data fitted from the ARIMA(3, 1, 0) model. The fitted line does not match exactly to the observed data, but the observed data is within the 95% confidence bars, so the observed data for the last six years is not unreasonable for the ARIMA(3, 1, 0) model.




The rough appears to have a mean $\mu = 0$ with consistent fluctuations around that mean, so we may consider this white noise.




The ACF is insignificant after lag 1 and the PACF cuts off at lag 1 due to the difference between lag 1 and lag 2. At a glance, this appears to fit an ARMA(1,1) model. However, we will consider ARIMA(p, 0, q) models (equivalently ARMA(p,q)), where $p = 0, ..., 3$ and $q = 0, ..., 3$, selected using the AIC criterion, to model the rough.

Table 3: AIC Table for Rough

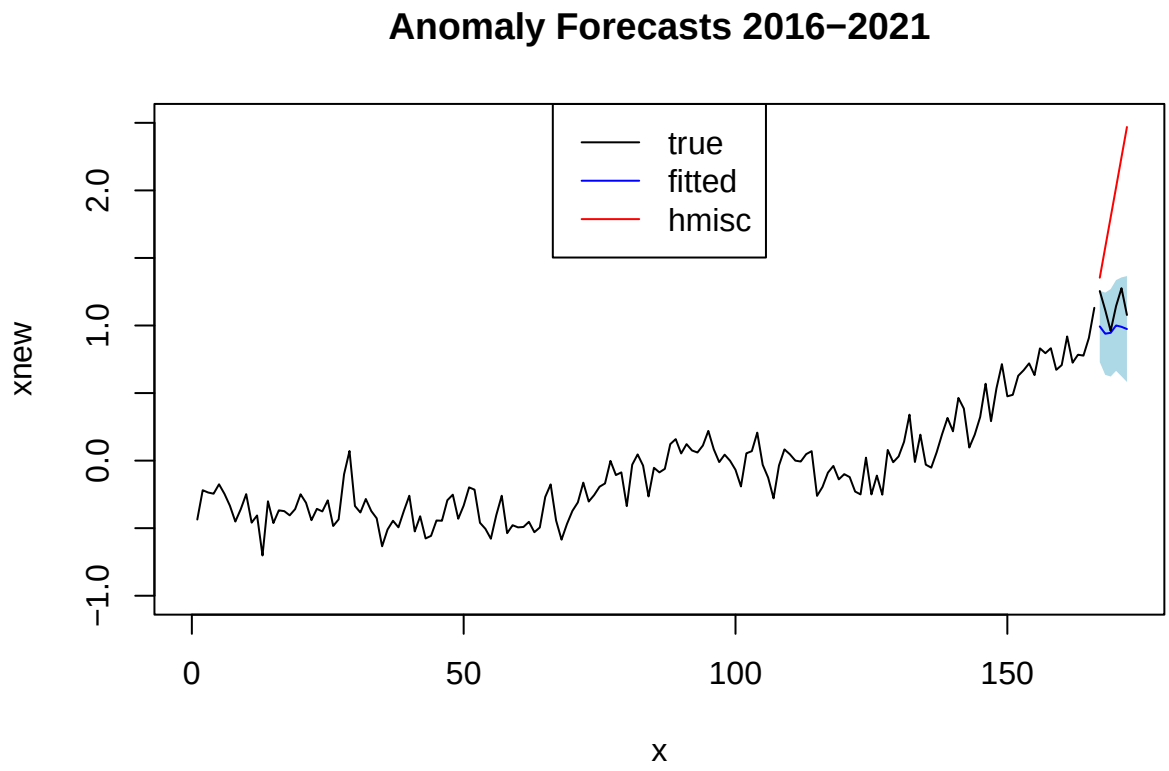
| | | | |
|---|---|---|---|
| -1.376535 | -1.418924 | -1.406507 | -1.433194 |
| -1.410624 | -1.406456 | -1.406706 | -1.533269 |
| -1.417466 | -1.542171 | -1.413105 | -1.522767 |
| -1.419344 | -1.410996 | -1.527142 | -1.509657 |

According to the AIC values, the best model for the rough is ARIMA(2,0,1), or equivalently, ARMA(2,1).

We will use a different forecasting method using the function ApproxExtrap in the Hmisc package to linearly extrapolate the last six years of anomalies.

The temperature anomaly forecasts for the last six years using this method are:

- 1.354
- 1.577
- 1.800
- 2.023
- 2.246
- 2.469




It appears that this method predicted the increasing linear trend we observe in the original data. It is quite different from the forecast obtained using the ARIMA(3, 1, 0) model, an forecasts an even greater increase in temperature anomaly for 2016-2021.

## Conclusion

In this analysis, we analyzed yearly temperature anomalies between 1850 and 2021 for the northern hemisphere. We observed that the original data was not stationary, so we used the first ordered differenced series for the analysis. This was necessary to ensure the series we were using was stationary, as analyzing non-stationary data would affect the accuracy of our final model. Through analyzing the ACF and PACF plots we were able to visually identify models by seeing where the ACF is insignificant and where the PACF cuts off. However, we considered ARIMA(p, 1, q) models where $p = 0, ..., 3$ and $q = 0, ..., 3$ to fit the final model, which was ARIMA(3, 1, 0).

We then fitted the same model to the data without the last six observations, to obtain forecasts for 2016-2021. We saw that the forecasts obtained through the ARIMA(3, 1, 0) did not match the forecasts obtained using the Hmisc package. The latter appeared to consider the increasing linear trend in the second half of the data, and predicted an even greater increase in temperature anomalies.

Some caveats with this analysis is that the temperature anomaly data was not stationary. Due to climate change, temperature anomalies have increased significantly over the years and we had to use the differenced time series for the analysis. We may be able to better predict the increase in temperature anomalies though other methods.

**Code Appendix**

```
knitr::opts_chunk$set(warning = FALSE, echo = FALSE)
library(readxl)
data <- as.data.frame(read_excel("TempNH_1850_2021.xlsx"))
anomalies = data[,2]
ts.plot(anomalies, main = "Yearly Temperature Anomaly", ylab = "Temperature Anomaly")
trend_spline=function(y, lam){
  n=length(y)
  p=length(lam)
  rsq=rep(0, p)
  y=sapply(y,as.numeric)
  tm=seq(1/n, 1, by=1/n)
  xx=cbind(tm, tm^2, tm^3)
  knot=seq(.1, .9, by=.1)
  m=length(knot)
  for (j in 1:m){
    u=pmax(tm-knot[j], 0); u=u^3
    xx=cbind(xx,u)
  }
  for (i in 1:p){
    if (lam[i]==0){
      ytran=log(y)
    } else {
      ytran=(y^lam[i]-1)/lam[i]
    }
    ft=lm(ytran~xx)
    res=ft$resid; sse=sum(res^2)
    ssto=(n-1)*var(ytran)
    rsq[i]=1-sse/ssto
  }
  ii=which.max(rsq); lamopt=lam[ii]
  if (lamopt==0) {
    ytran=log(y)
  } else {
  ytran=y^lamopt
  }
  ft=lm(ytran~xx);
  best_ft=step(ft, trace=0)
  fit=best_ft$fitted; res=best_ft$resid
  result=list(ytrans=ytran, fitted=fit, residual=res, rsq=rsq, lamopt=lamopt)
  return(result)
}
tm = 1:nrow(data)
splinetrnd=trend_spline(anomalies, 1)
plot(tm, anomalies, type="l", lty=1, xlab="Year", ylab="Anomaly", main="Time series with spline trend")
points(tm, splinetrnd$fitted, type="l", lty=1, col = "red")
x<-unlist(anomalies)
acf(x, main = "Temperature Anomaly ACF")
par(mfrow = c(2,2))
plot.ts(anomalies, main = "Temperature Anomaly")

#differenced series
```

```r
y<-diff(x,1)
plot.ts(y, main = "Anomaly First Differenced")
acf(y, main = "First Differenced ACF")
pacf(y, main = "First Differenced PACF")
par(mfrow = c(1,1))
library(astsa)
#preliminary model, ARIMA(3,1,1)
pre_mod = sarima(x,p=3,d=1,q=1,details=FALSE)
acf(pre_mod$fit$residuals, main = "ARIMA(3, 1, 1) ACF")
AICc<-matrix(0,4,4)
for (i in 1:4){
  for (j in 1:4) {
    AICc[i,j]<-sarima(x,p=i-1,d=1,q=j-1,details=FALSE)$AICc
  }
}
knitr::kable(AICc, "pipe", caption = "AIC Table")
final_mod = sarima(x,p=3,d=1,q=0,details=FALSE)
acf(final_mod$fit$residuals, main = "ARIMA(3, 1, 0) ACF")
knitr::kable(final_mod$ttable, "pipe", caption = "ARIMA(3, 1, 0) Parameter Estimates")
n <- length(anomalies)
xnew <- anomalies[1:(n-6)]
xlast <- x[(n-5):n]

model1 <- arima(xnew,order = c(3,1,0))

h <- 6
m <- n - h

fcast <- predict(model1, n.ahead=h)

upper <- fcast$pred+1.96*fcast$se
lower <- fcast$pred-1.96*fcast$se

plot.ts(xnew, xlim = c(0,n), xlab = "x", main = "Anomaly Forecast 2016-2021")
polygon(x=c(m+1:h,m+h:1), y=c(upper,rev(lower)), col='lightblue', border=NA)
lines(x=m+(1:h), y=fcast$pred,col='blue')
lines(x=m+(1:h), y=xlast,col='black')
legend("top", legend = c("true","fitted"), lty=c(1, 1), col = c("black","blue"))
y <- data$Anomaly[1:166]
tm <- 1:166
tmout <- 167:172

splinetrnd=trend_spline(y, 1)
par(mfrow=c(1,2))
plot(tm, y, type="l", lty=1, xlab="Year", ylab="Anomaly", main="Time series with spline trend")
points(tm, splinetrnd$fitted, type="l", lty=1, col = "red")
rough = splinetrnd$residual
ts.plot(rough, main = "Rough")
x1 = splinetrnd$residual

par(mfrow=c(1,2))
acf(x1, lag.max = 25, main = "ACF plot of rough part")
pacf(x1,main="PACF", lag.max = 25)
```

```
aic<-matrix(0,4,4)
for (i in 1:4){
  for (j in 1:4) {
    aic[i,j]<-sarima(splinetrnd$residual,p=i-1,d=0,q=j-1, details = FALSE)$AICc
  }
}
knitr::kable(aic, "pipe", caption = "AIC Table for Rough")
library(Hmisc, quietly = T)
fcast_hmisc = approxExtrap(tm, y, tmout)
plot.ts(xnew, xlim = c(0,n), ylim = c(-1, 2.5), xlab = "x", main = "Anomaly Forecasts 2016-2021")
polygon(x=c(m+1:h,m+h:1), y=c(upper,rev(lower)), col='lightblue', border=NA)
lines(x=m+(1:h), y=fcast$pred,col='blue')
lines(x=m+(1:h), y=xlast,col='black')
lines(x=m+(1:h), y=fcast_hmisc$y,col='red')
legend("top", legend = c("true","fitted","hmisc"), lty=c(1, 1), col = c("black","blue","red"))
```