Stats 137 Project

Richard Ly (ID: 917270664)

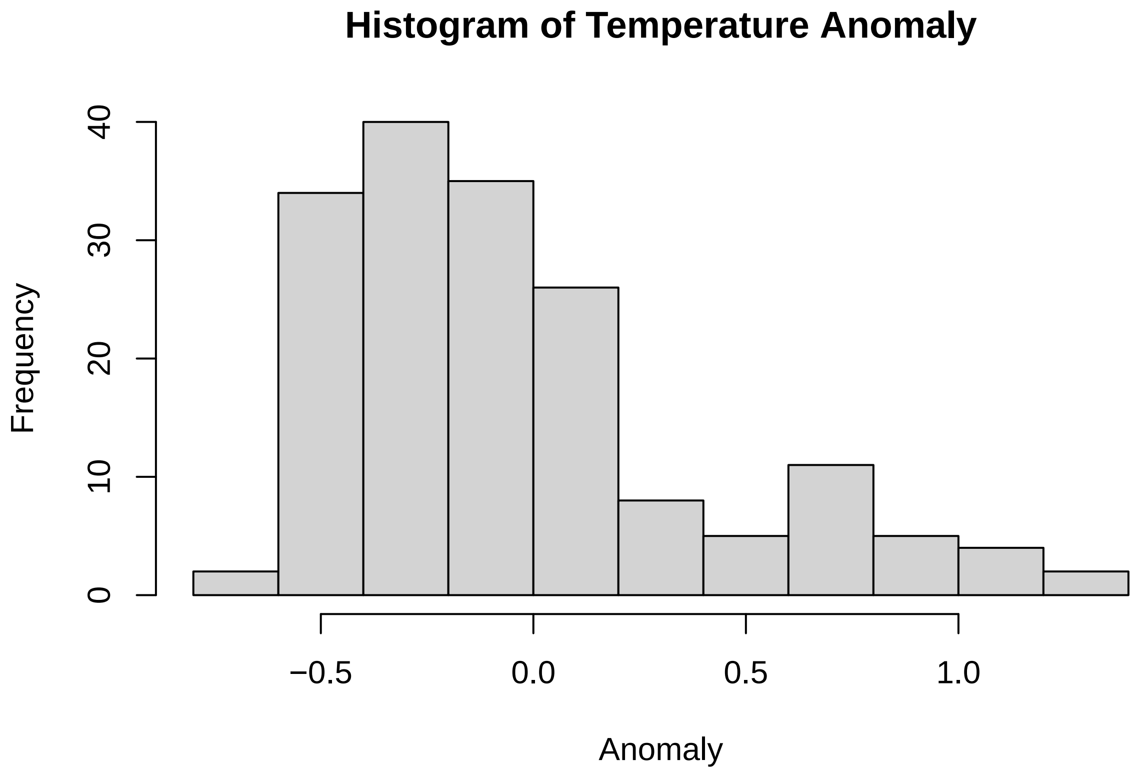
12/4/2022

Introduction (Statement of the problem)

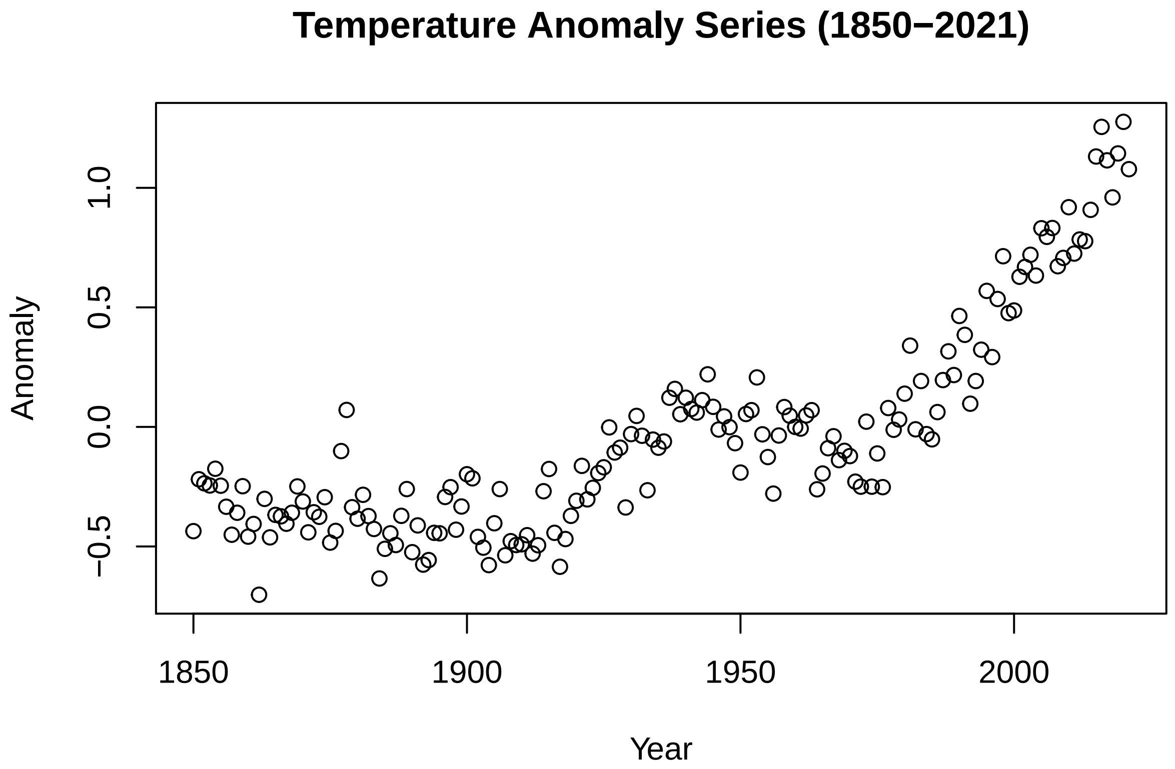
The dataset considered was obtained by the Climate Research Center, University of East Anglia, UK, and concerns the annual temperature anomalies for the northern hemishpere from 1850 to 2021. The data contains an independent variable 'Year' denoting the observed year, and a dependent variable 'Anomaly' denoting the relative temperature anomaly for a given year. Given the nature of the dataset, it may viewed as a time series in which a series of temperature anomalies is measured and plotted against some unit time. Analyzing the behavior of temperature anomalies is important in order to identify whether there exists any trend, increasing or decreasing, for the annual temperature in the northern hemisphere. This may be approached with time series modeling. If such a trend exists, it may be further explored to perhaps identify why the trend in temperature anomalies is occurring, and what may be done to mitigate or nullify the trend, if desired.

Materials and Methods (Description of the data and the methods used in the analysis)

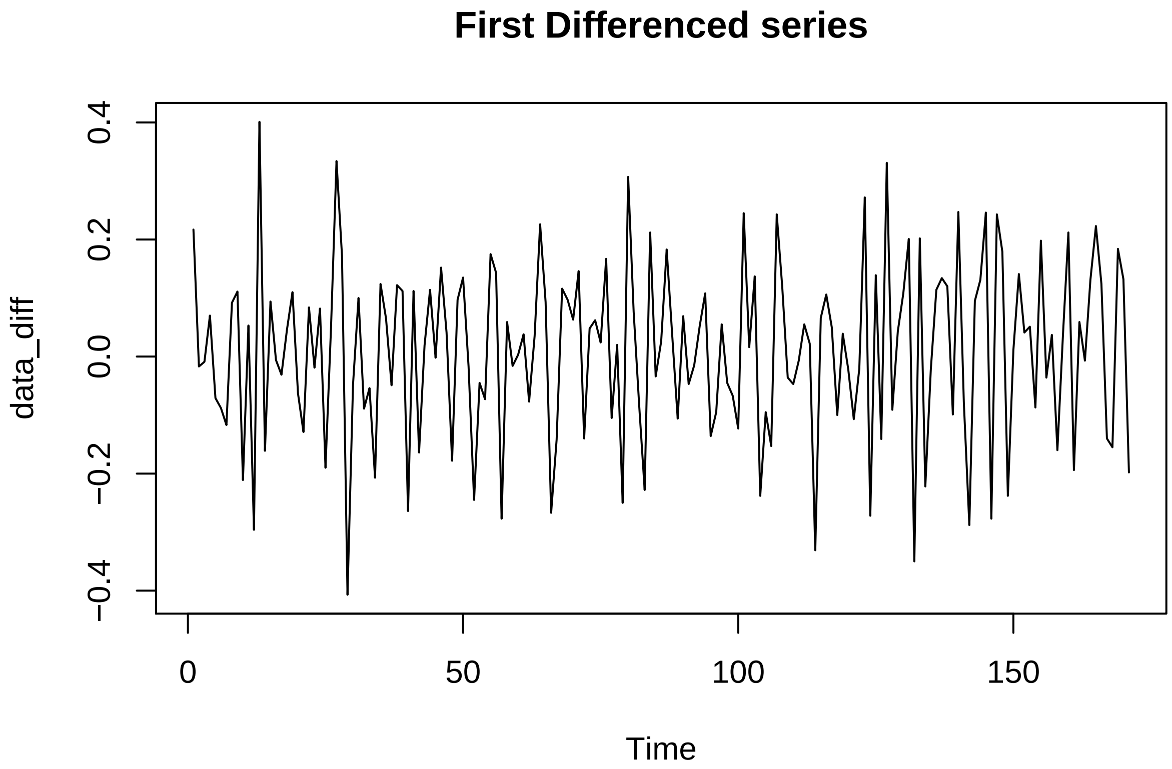
A histogram of the anomalies reveal that the anomalies are right-skewed. A correlation matrix reveals a fairly strong positive correlation between temperature anomaly and year. Additionally, a plot of temperature anomaly against year is given.



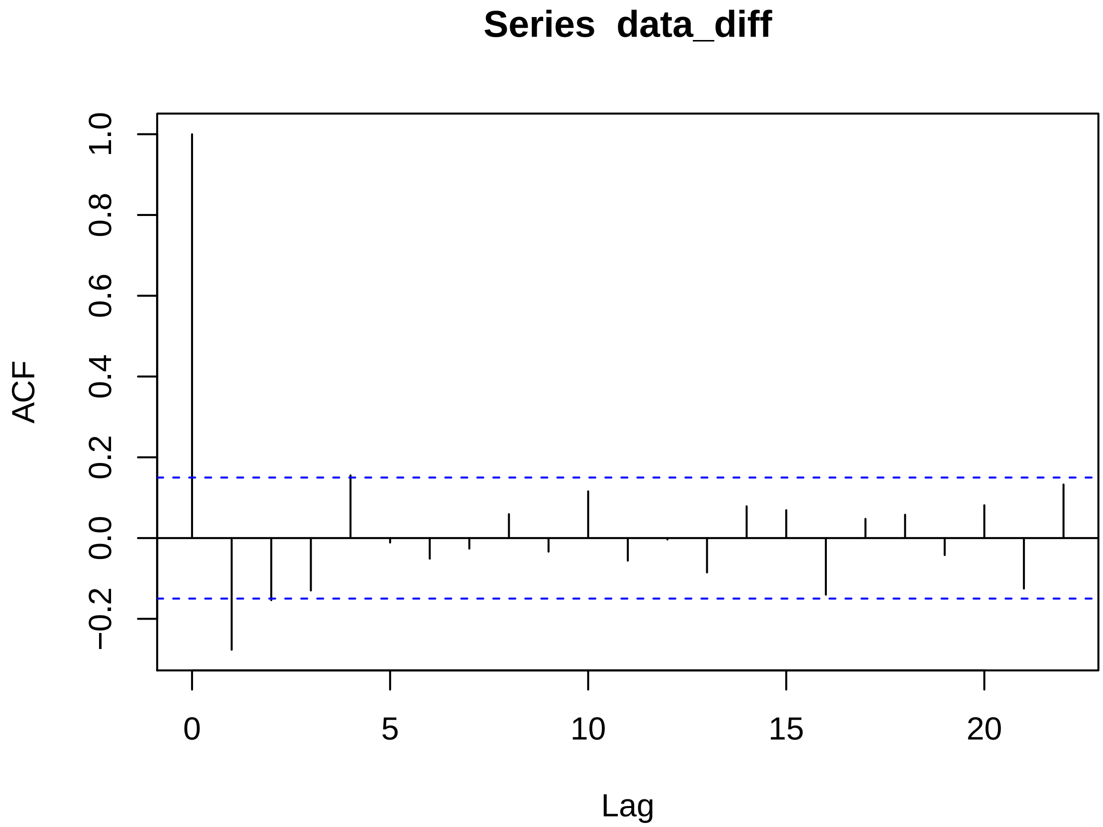
```
##              Year   Anomaly
## Year    1.0000000 0.8216695
## Anomaly 0.8216695 1.0000000
```

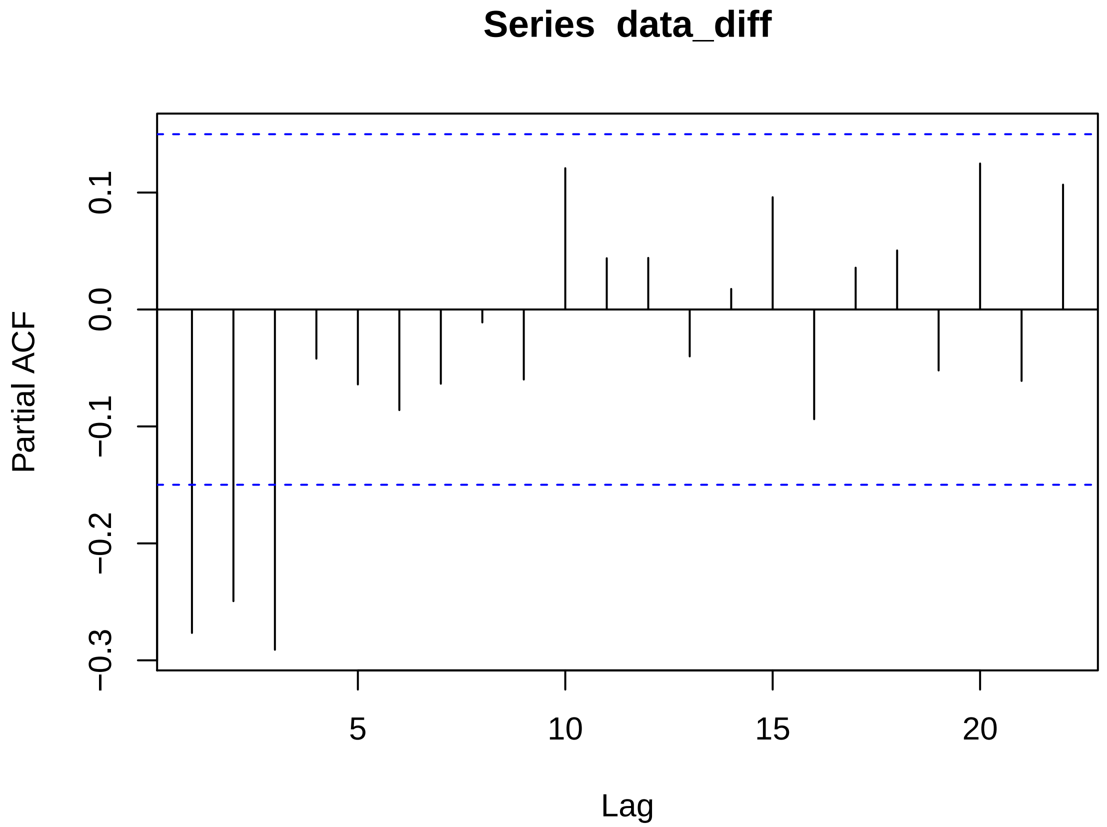


From the time series plot of temperature anomaly against year, a trend of increasing temperature anomaly can be observed beyond around the year 1910. Because such a trend exists, the series is not stationary. To obtain a stationary sequence to model and analyze, it is desirable to difference the temperature anomaly series.



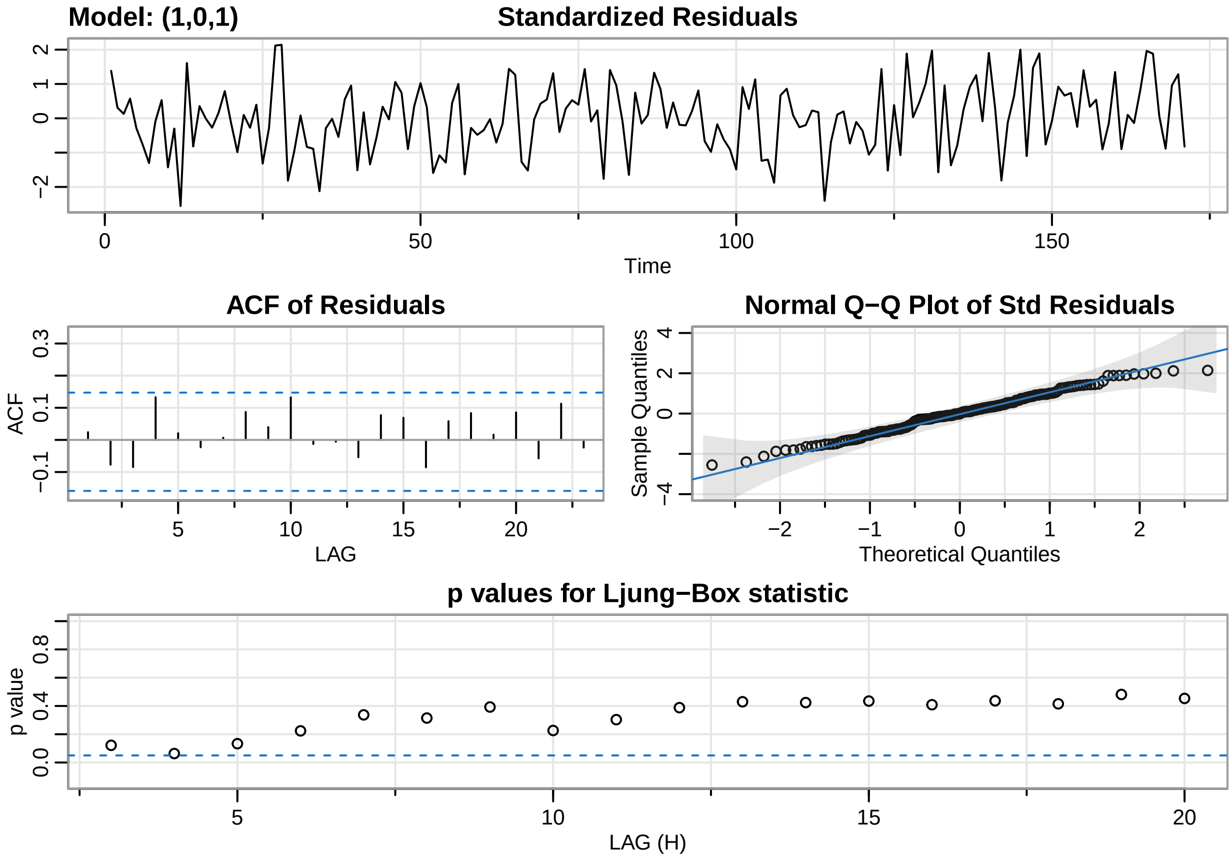
We obtain and plot the first differenced series for temperature anomaly above. We observe that the series seems to fluctuate about a mean zero, and seems to maintain a constant variance, throughout the observed time. Thus, the first differenced series yields a stationary sequence for which we may conduct time series modeling, and furthermore analyze.


Series data_diff
ACF
Lag
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
20



The ACF plot for the differenced series indicates that only autocorrelations of lag 1 are present, although lag 2 and lag 4 may be considered as being present as well. The PACF plot for the differenced series indicates that only partial autocorrelations for lags 1-3 are present.

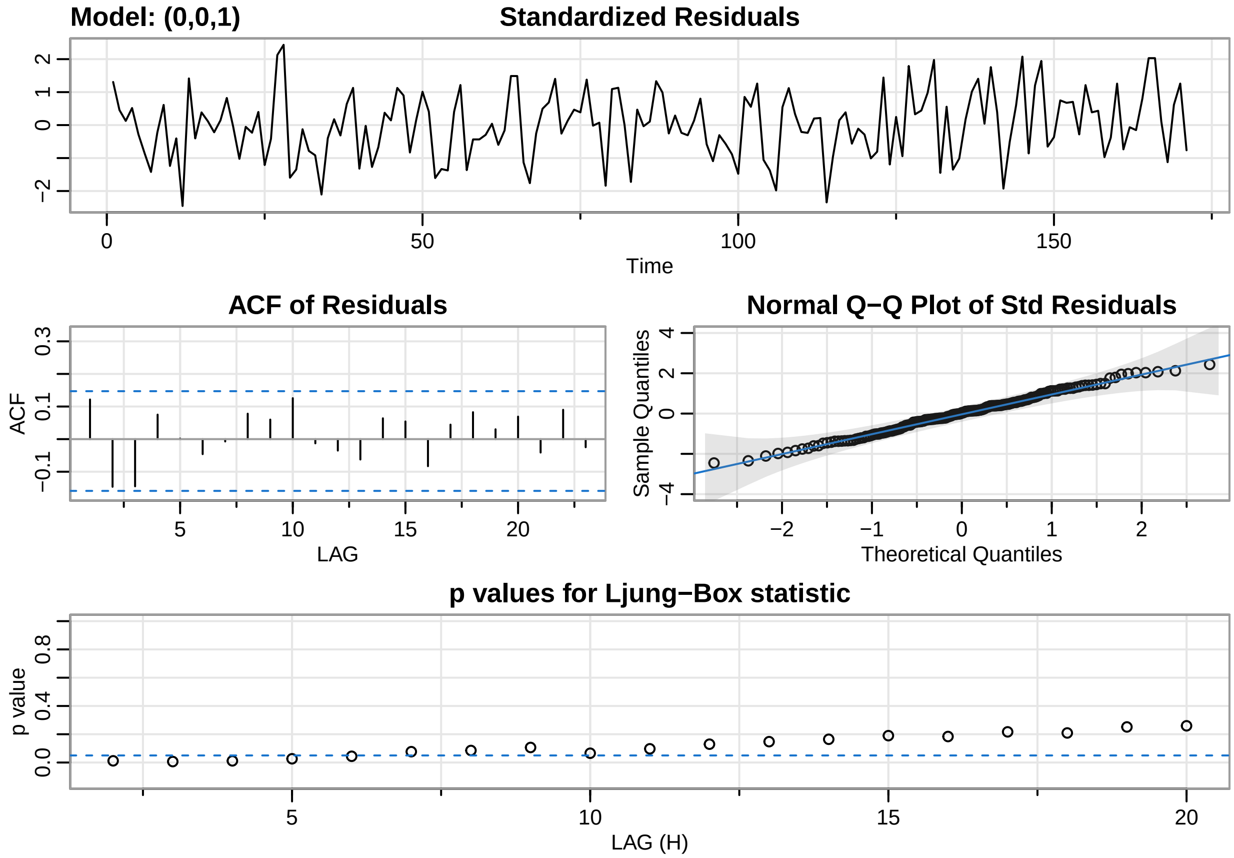
With the above observations, a preliminary model for the first differenced temperature anomaly series can be given by ARIMA(1,0,1).



We use the preliminary model of ARIMA(1,0,1) to fit the first differenced temperature anomaly series. The residuals of the preliminary model are plotted above. Additionally, the ACF of residuals and the Normal Q-Q plot of the residuals are plotted above, and indicate that the residuals are i.i.d. The p-values obtained for various lag are also shown for the Ljung-Box test. The p-values are fairly small, and so the claim of independent residuals is reasonable. With these results, the residuals can be claimed to resemble white noise, indicating that the preliminary model of ARIMA(1,0,1) is reasonable for the first differenced series.

```
##           [,1]      [,2]      [,3]      [,4]
## [1,] -157.8230 -187.0748 -191.4991 -189.5276
## [2,] -169.6728 -190.5115 -189.5084 -189.7570
## [3,] -178.7901 -190.6531 -190.2470 -190.5932
## [4,] -192.1007 -190.9245 -189.9385 -188.6703
```

To identify the final model, the AIC criterion is used. Models of the form ARIMA(p,0,q) are considered, where p = 0,. . . ,3 and q = 0,. . . ,3. The AIC values for the 16 models are given above in a table with the columns corresponding to the value of p, and the rows corresponding to the value of q. The model with the lowest AIC value is selected as the final model. Thus, the final model for the first differenced temperature anomaly series is given by ARIMA(0,0,3).



As with the preliminary model, diagnostics are performed on the final model's residuals. The plot of the residuals against time seem to resemble white noise. In support of this, the ACF values of the residuals are within the blue confidence lines, suggesting independence. The Normal Q-Q plot of the residuals suggest that the residuals are normally distributed. Finally, the p-values for the Ljung-Box test are significantly low. These results allow for a reasonable claim towards the final model's residuals being i.i.d.

The final model for the first differenced temperature anomaly series is given by ARIMA(0,0,3). Equivalently, the final model may be described as a MA(3) model:

$$X_t - \mu = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \theta_3\epsilon_{t-3}$$

$$\implies X_t = \mu + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \theta_3\epsilon_{t-3} + \epsilon_t$$

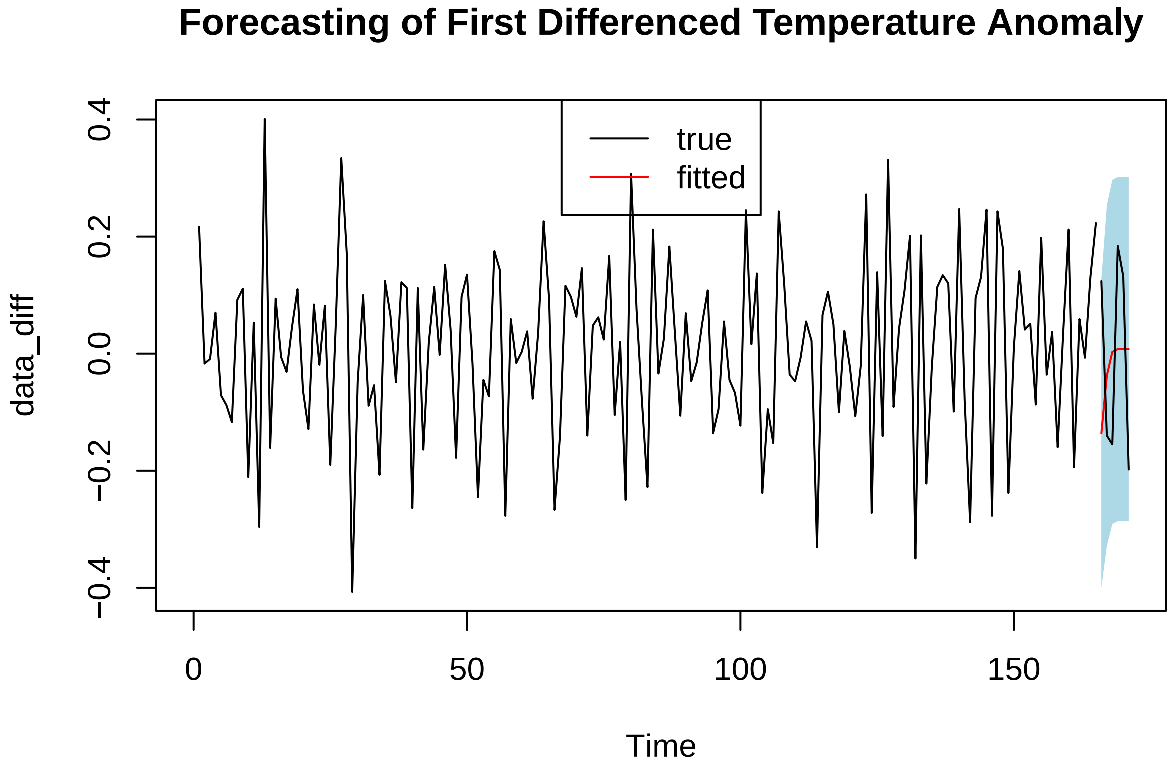
```
##          ma1          ma2          ma3    intercept
## -0.452065198 -0.173042918 -0.013508577  0.008524323
```

The estimated parameters of the final model are given above. This yields the final model as:

$$X_t = 0.00852 - 0.452\epsilon_{t-1} - 0.173\epsilon_{t-2} - 0.0135\epsilon_{t-3} + \epsilon_t$$

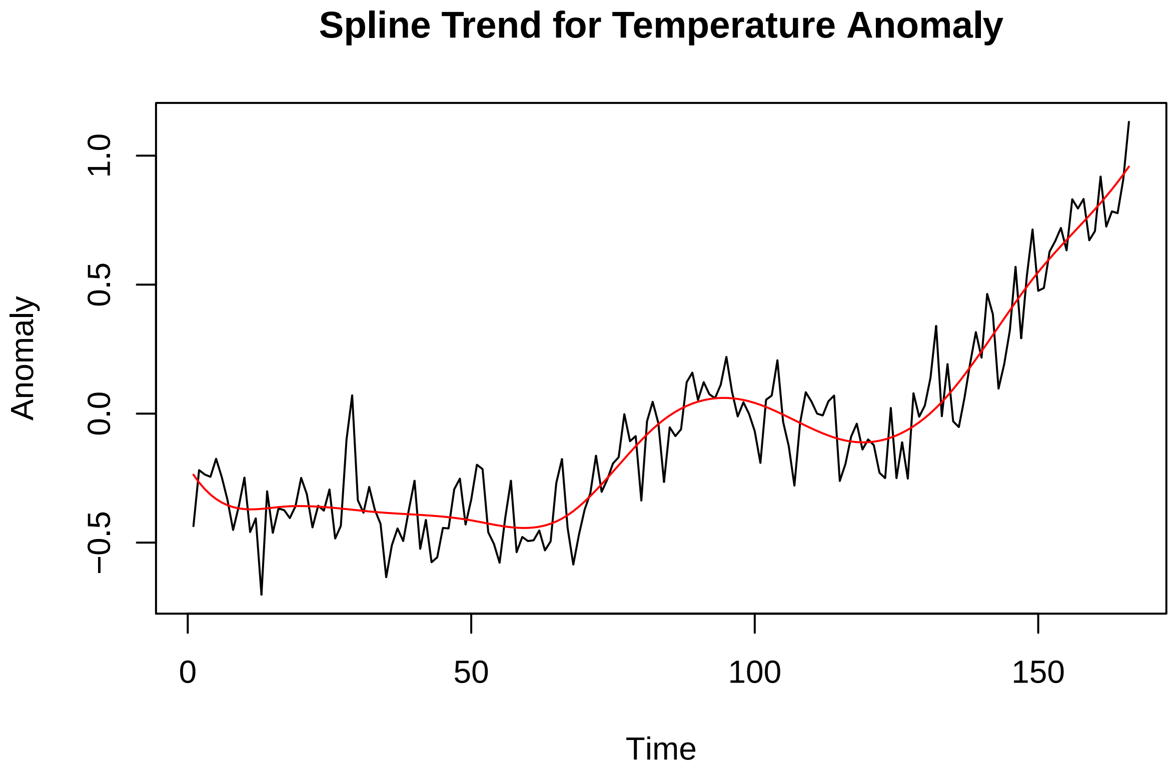
```
##         ma1         ma2         ma3   intercept
## 0.083937930 0.071580479 0.080239724 0.003779355
```

The standard errors for the estimated model parameters are given above.

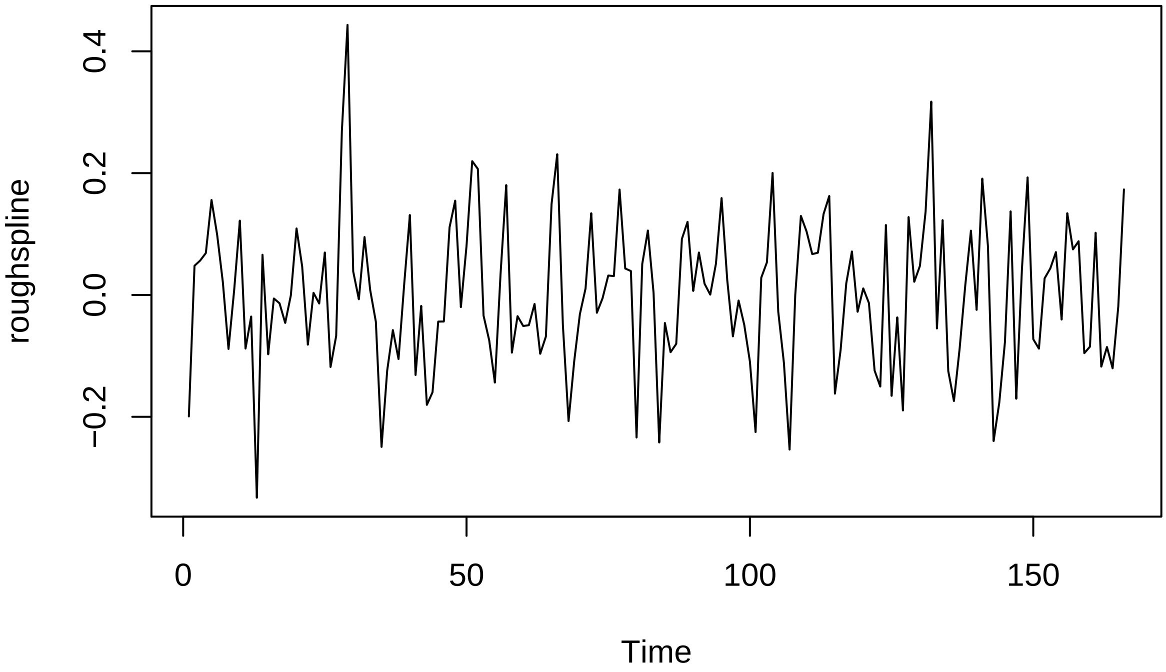


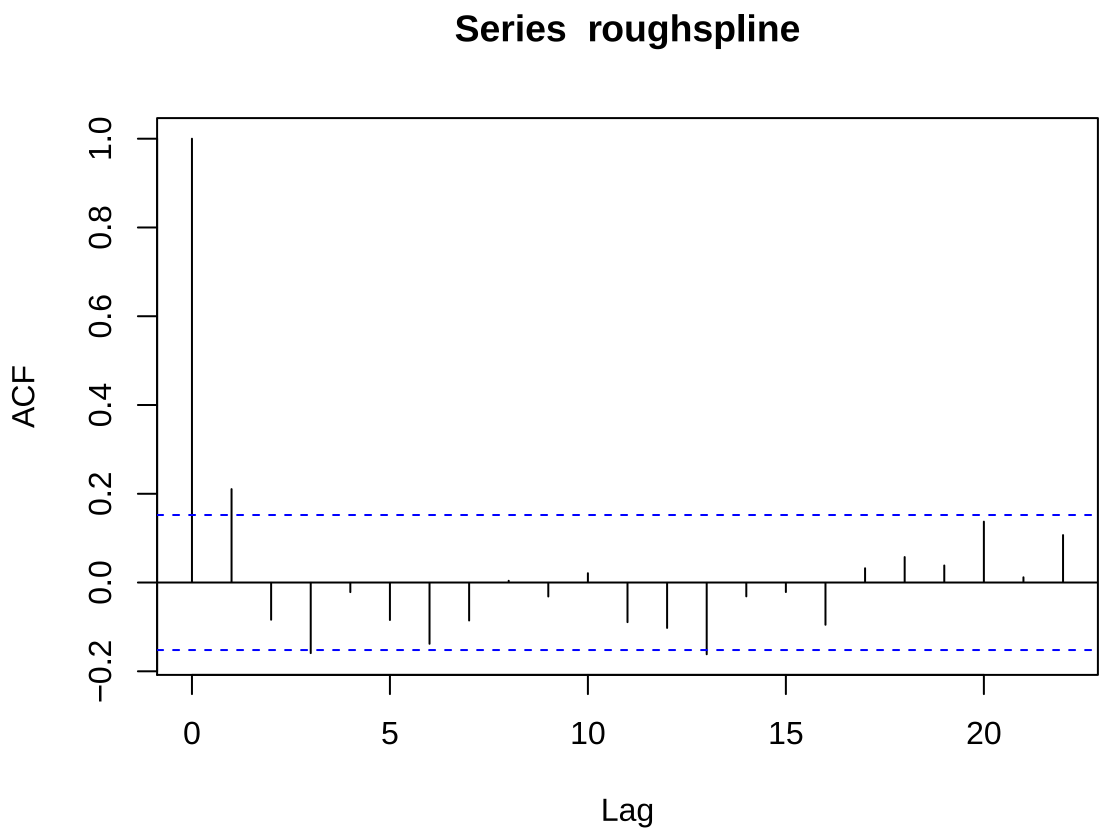
Above we plot the observed first differenced temperature anomaly series. In addition, the last 6 years of the first differenced series is forecasted using all years prior to the last 6 years. The observed values and the forecasted values are shown on the plot. A 95% confidence band for the 6 year forecast is plotted as well. We note that the final 6 years observed fall within the calculated 95% confidence band, and so the final model does reasonably well in forecasting.

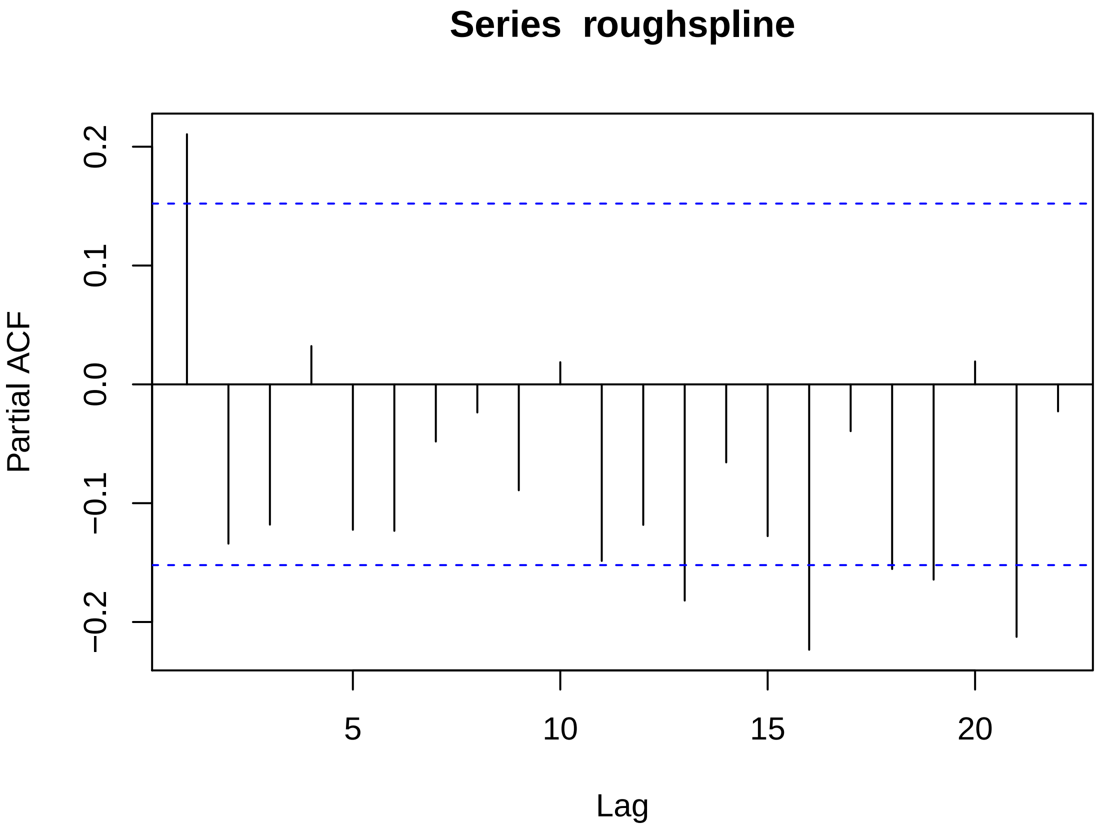
To compare forecasting with a different method, we estimate the trend by spline using all but the last 6 years of observation. Below the spline trend estimate is plotted over the observed series.



Below we plot the rough of the spline trend estimate, as well as plot the ACF and PACF values for the rough.


roughspline
−0.2
0.0
0.2
0.4
0
50
100
150
Time


Series roughspline
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
20
Lag



```
##            [,1]      [,2]      [,3]      [,4]
## [1,] -228.5292 -235.6150 -233.6283 -238.1587
## [2,] -234.2372 -233.6198 -233.7616 -254.8977
## [3,] -235.4474 -256.2489 -234.9504 -253.3076
## [4,] -235.8595 -234.6004 -254.0339 -251.3119
```

The rough for the spline trend estimate can be modeled by ARMA(p,q). The model with the lowest AIC value is selected. The AIC values for ARMA(p,q), p = 0,..,3, q = 0, . . . ,3 are given above, with columns corresponding to p and rows corresponding to q. This gives that the rough may be modeled as ARMA(1,2).

- Forecast the temperature anomalies for the years 2016-2021 by guessing the trend for these years (you should use the function ApproxExtrap in the Hmisc package) and forecasting the rough. Plot the observed and the forecasts from these two methods against time on the same graph.

```
## Loading required package: lattice
```

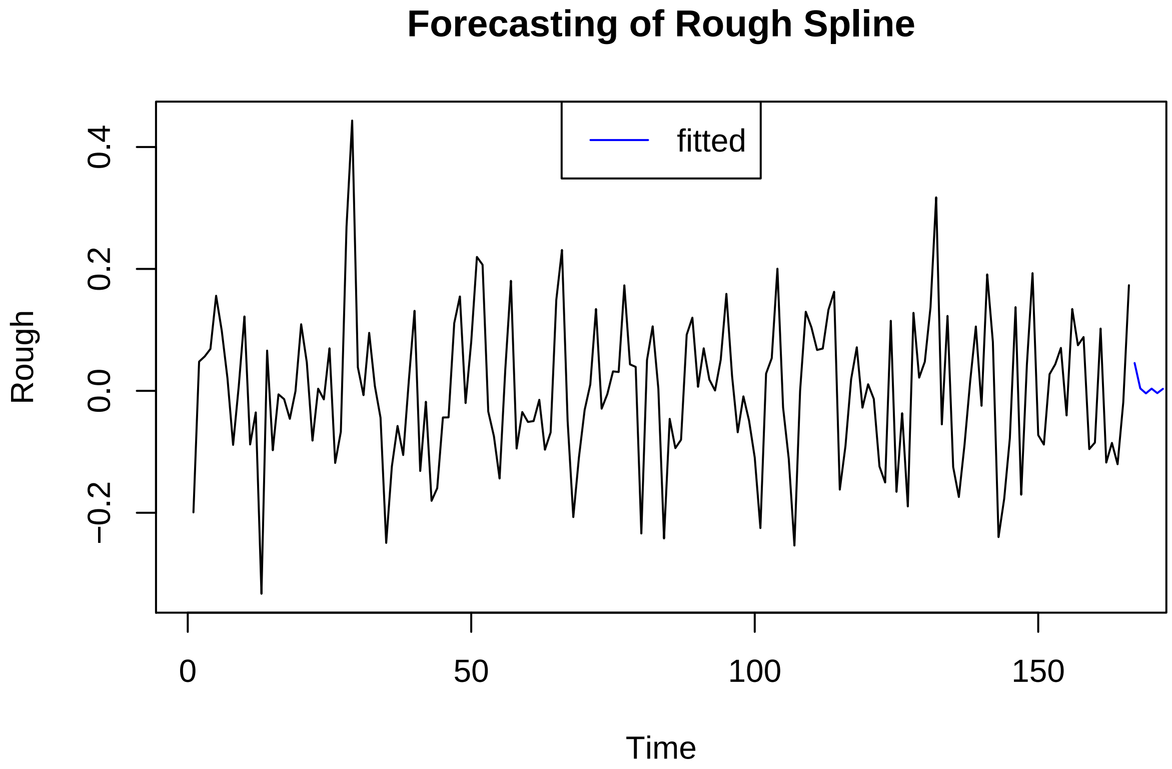
```
## Loading required package: survival
```

```
## Loading required package: Formula
```

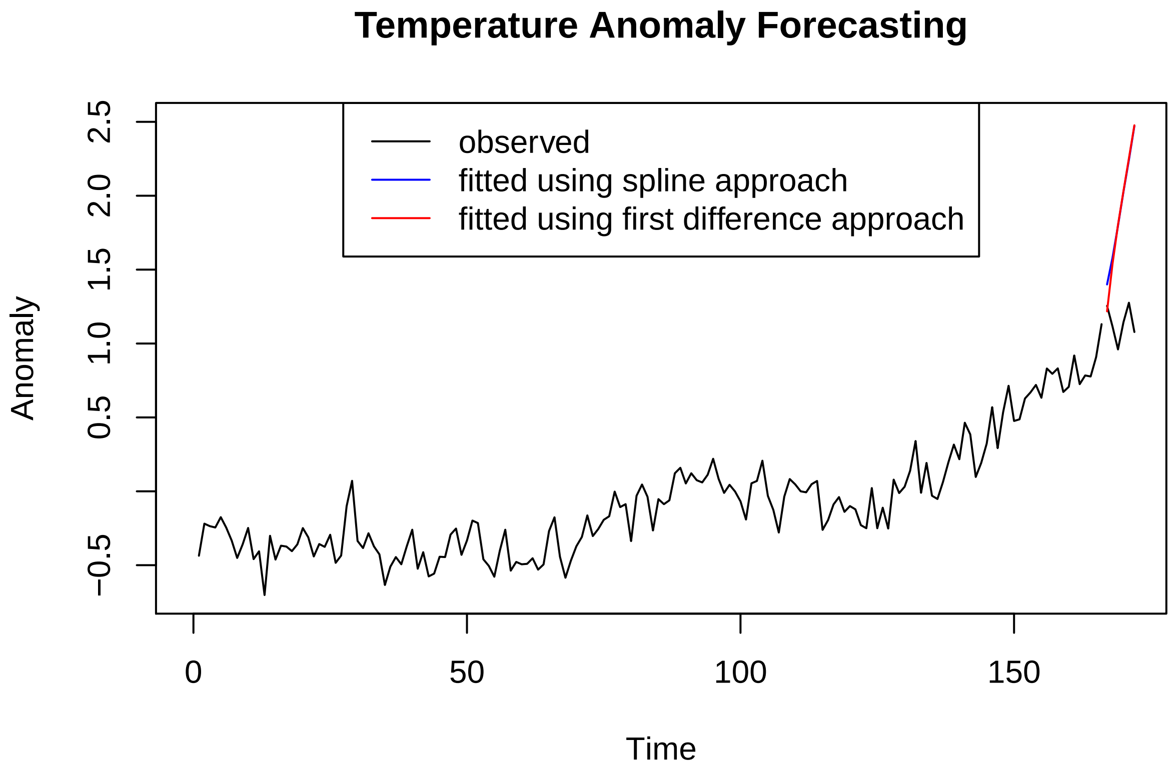
```
## Loading required package: ggplot2
```

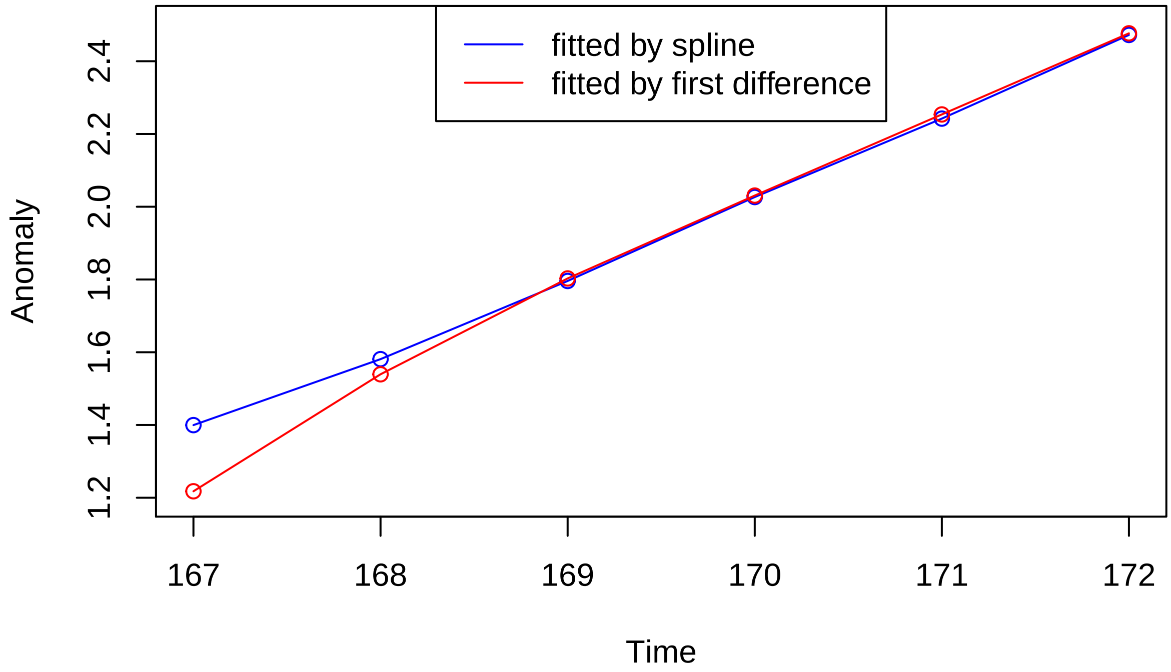
```
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

For the years 2016-2021, the trend of the temperature anomaly series is guessed using the function approxExtrap() in the Hmisc package. The rough is forecasted. The forecasted values for temperature anomalies for the years 2016-2021 is then the sum of the trend and the forecasted rough. The forecasted temperature anomalies are plotted below for the spline approach.



To compare the forecasting done with the first difference approach, both approach's forecasts are plotted together with the true observed values. The full plot and a magnified plot are given below.


Temperature Anomaly Forecasting
observed
fitted using spline approach
fitted using first difference approach
Anomaly
−0.5
0.5
1.0
1.5
2.0
2.5
0
50
100
150
Time



Results

Regarding the first difference approach to forecasting, the forecasts were found to be reasonable. The preliminary model seemed reasonable, but a final model selected by AIC criterion was instead used. The true, observed anomaly values were within the 95% confidence band of the forecast. The final model presented can therefore be accepted as a reasonable time series model for the first differenced series.

The spline approach to forecasting required estimated the trend of the data. Through this, the rough of the spline estimate was obtained, and a model was fit to the rough with respect to the AIC criterion. The estimated trend and the rough were together used to forecast. The forecasted last 6 years of temperature anomalies was found to be very similar to those forecasted values found in the first difference approach.

Conclusion and Discussion

The original dataset concerning the annual temperature anomalies of the northern hemisphere (1850-2021) contained a trend, and could not be considered a stationary series. To remedy this, the first differenced series was considered instead, as it was found to be a stationary series. Using the ACF and PACF plots for the first differenced series, a preliminary model was obtained. While the preliminary time series model seemed reasonable, a final model was selected in accordance to the AIC criterion. The final model is given by an ARIMA(0,0,3) model, or equivalently a MA(3) model. The model and its parameters are:

$$X_t = 0.00852 - 0.452\epsilon_{t-1} - 0.173\epsilon_{t-2} - 0.0135\epsilon_{t-3} + \epsilon_t$$

This final model was used to forecast the 6 most recent observations, using all observations excluding these 6 most recent years. The forecast was found to be reasonable. In addition to this forecast found through the first differenced series, an alternate forecast was found by estimating the trend of the original temperature anomalies through splines. The rough of the spline trend estimate was used together with the estimated trend to similarly forecast the 6 most recent years, using all observations excluding these 6 most recent years.

The forecasts for the first difference approach and the spline trend approach were found to be very similar. The first difference forecasts were of slightly lower value for the first two forecasts, but were nearly identical for subsequent forecasts. From this we may draw a conclusion that for the dataset used, both approaches considered are reasonable and perform similarly well.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
library(readxl)
data <- read_excel("C:/Users/rly75/OneDrive/Desktop/137/TempNH_1850_2021.xlsx")
hist(data$Anomaly,main = 'Histogram of Temperature Anomaly',xlab='Anomaly')
cor(data)
plot.ts(data$Year,data$Anomaly, xlab='Year', ylab='Anomaly', main='Temperature Anomaly Series (1850-202
data_diff = diff(data$Anomaly,1)
plot.ts(data_diff, main = "First Differenced series")
par(mar=c(4,4,4,4))
acf(data_diff)
par(mar=c(4,4,4,4))
pacf(data_diff)
library(astsa)
model_prelim = arima(data_diff,order=c(1,0,1))

# model_prelim$coef
# model_prelim$var.coef #diagonal yields variance
# sqrt(diag(model_prelim$var.coef)) #gives standard error
# model_prelim$sigma2
#ts.plot(model_prelim$residuals)
sarima(data_diff,p=1,d=0,q=1)
aic_table = matrix(rep(0,16),nrow=4,ncol=4)
for (i in 0:3){
  for (j in 0:3){
    aic_table[i+1,j+1]=arima(data_diff,order=c(i,0,j))$aic
  }
}

aic_table
sarima(data_diff,p=0,d=0,q=1)
model_final = arima(data_diff,order=c(0,0,3))

model_final$coef
#model_final$var.coef #diagonal yields variance
sqrt(diag(model_final$var.coef)) #gives standard error
#model_final$sigma2
n = length(data_diff)
xnew <- data_diff[1:(n-6)]
xlast <- data_diff[(n-5):n]
#fit
model1 <- arima(xnew,order = c(0,0,3))
#prediction
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
plot.ts(xnew, xlim = c(0,n), xlab = "Time",ylab="data_diff",main="Forecasting of First Differenced Tempe
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$pred,col='red')
lines(x=m+(1:h), y=xlast,col='black')
legend("top", legend = c("true","fitted"), lty=c(1, 1), col = c("black","red"))
trend_spline=function(y, lam){
  n=length(y)
  p=length(lam)
  rsq=rep(0, p)
  y=sapply(y,as.numeric)
  tm=seq(1/n, 1, by=1/n)
  xx=cbind(tm, tm^2, tm^3)
  knot=seq(.1, .9, by=.1)
  m=length(knot)
  for (j in 1:m){
    u=pmax(tm-knot[j], 0); u=u^3
    xx=cbind(xx,u)
  }
  for (i in 1:p){
  if (lam[i]==0){
    ytran=log(y)
    } else {
        ytran=(y^lam[i]-1)/lam[i]
      }
    ft=lm(ytran~xx)
    res=ft$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)
}
end = length(data$Anomaly)-6
y = data$Anomaly[1:end]
tm = 1:length(y)
splinetrnd=trend_spline(y, 1) ##note: lam = 1 means no transformation
plot(tm, y, type="l", lty=1, xlab="Time", ylab="Anomaly", main="Spline Trend for Temperature Anomaly")
points(tm, splinetrnd$fitted, type="l", lty=1, col = "red")
roughspline = splinetrnd$residual
plot.ts(roughspline)
par(mar=c(4,4,4,4))
acf(roughspline)
pacf(roughspline)
aic_table_spline = matrix(rep(0,16),nrow=4,ncol=4)
for (i in 0:3){
  for (j in 0:3){
    aic_table_spline[i+1,j+1]=arima(roughspline,order=c(i,0,j))$aic
  }
}

aic_table_spline
library(Hmisc)
m = length(data$Anomaly) - 6
fcast_trend = approxExtrap(1:m,data$Anomaly[1:m],xout= seq(length(data$Anomaly)-5,length(data$Anomaly),
ynew = roughspline
h <- 6
m = length(roughspline)
model2 = arima(ynew,order = c(1,0,2))
fcast_roughspline <- predict(model2, n.ahead=h)
plot.ts(roughspline, xlab = "Time",ylab="Rough",main="Forecasting of Rough Spline")
lines(x=m+(1:h), y=fcast_roughspline$pred,col='blue')
legend("top", legend = c("fitted"), lty=c(1), col = c("blue"))
fcast_spline = fcast_trend + fcast_roughspline$pred
fcast_diff = fcast_trend + fcast$pred
plot(tm, y, type="l", xlim=c(0,n),ylim =c(-0.7,2.5) ,lty=1, xlab="Time", ylab="Anomaly", main="Temperatu
lines(x=m+(1:h), y=data$Anomaly[167:172],col='black')
lines(x=m+(1:h), y=fcast_spline,col='blue')
lines(x=m+(1:h), y=fcast_diff,col='red')
legend("top", legend = c("observed","fitted using spline approach", "fitted using first difference appro

plot(x=m+(1:h), y=fcast_spline,ylim=c(1.2,2.5),type = 'o', col='blue', ylab ="Anomaly",xlab = "Time")
lines(x=m+(1:h), y=fcast_diff,type='o',col='red')
legend("top", legend = c("fitted by spline", "fitted by first difference"), lty=c(1,1), col = c("blue",
```