Analyzing Temperature Anomalies Data

Robert Dam and Joshua Choi

2022-11-22

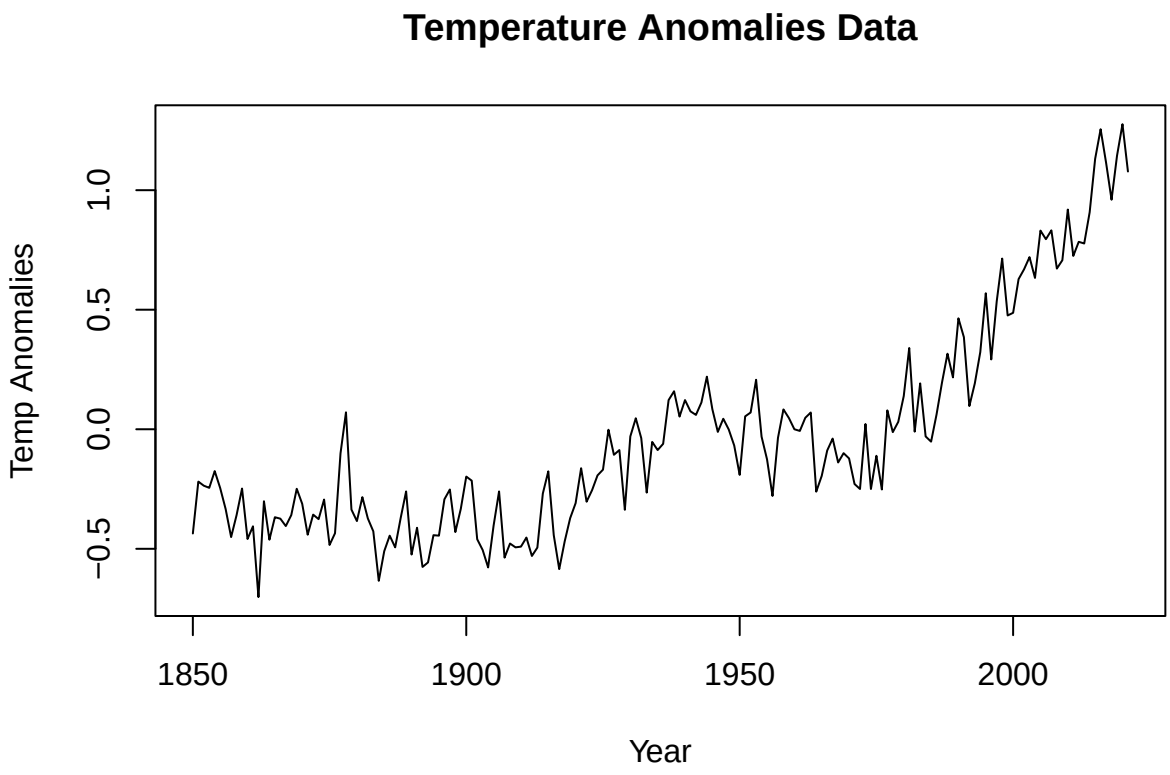
1) Introduction

This dataset is the recorded temperature anomalies throughout 1850 to 2021. Temperature anomaly is the general departure from a reference value or a long term average in a given year. This data set is a time series because the data is record in a consistent sequence taken at successively equally spaced points in time, in this case, the units are years. We expect the variable of interest to have dependency on time. It is important to analyze the temperature anomalies data because if we find a relationship with the anomalies with respect to time it can be helpful to project future anomalies and try to measure the impact it can have in terms of global warming. Short-term forecasts can be useful in detecting extreme weather events in order to sufficiently prepare for them.

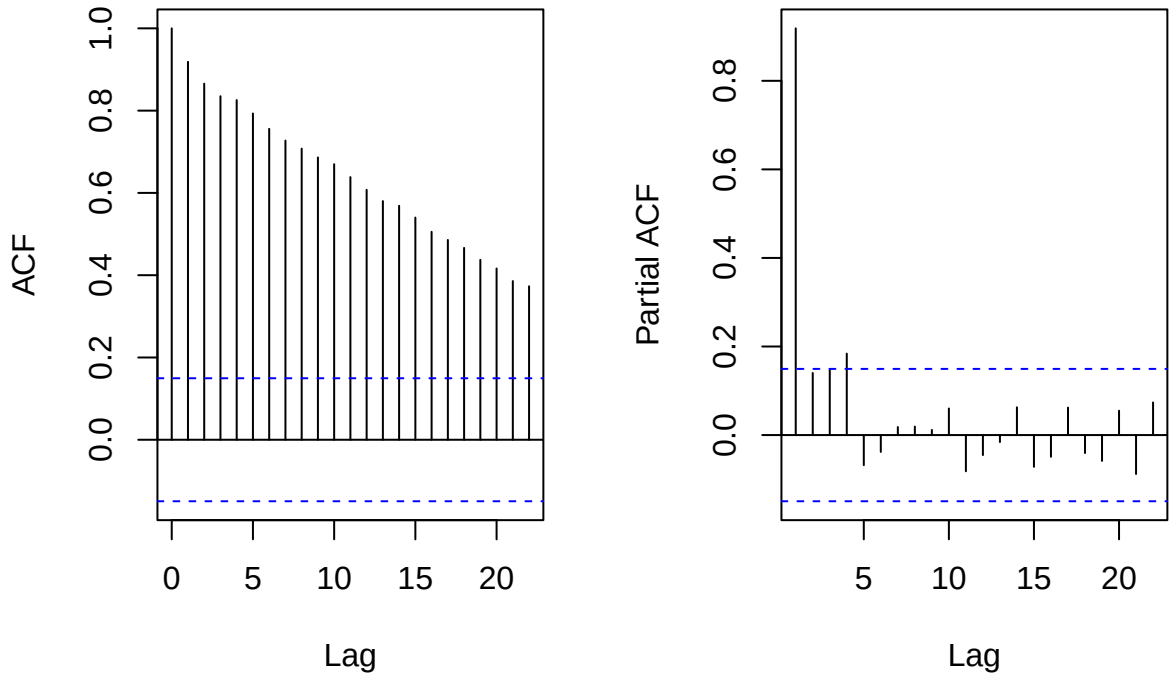
2) Materials and Methods

2.1) Examining Raw Temperature Anomalies Data

We will first examine the temperature anomaly dataset by plotting the data and examining autocorrelation plots in order to get a sense of what the data is like.



When we plot the anomalies with respect the time, our general trend seems to be that the temperature anomalies increase as time goes on. This suggests that global warming and temperature anomalies could be accelerating.



From the autocorrelation plots, we can see that this series is not stationary because graphically in our temperature anomaly data, the data doesn't seem to centered around the "mean" and in the ACF it suggests there is a positive auto correlation for a large time lag further strengthening the evidence that this series is not stationary. Therefore it seems that the model would require differencing of the series in order to proceed the analysis as a stationary series. To analyze this data we will conduct an ARIMA analysis and a trend modeling analysis with splines so that we have two different ways of forecasting this data.

2.2) Methods: ARIMA Model

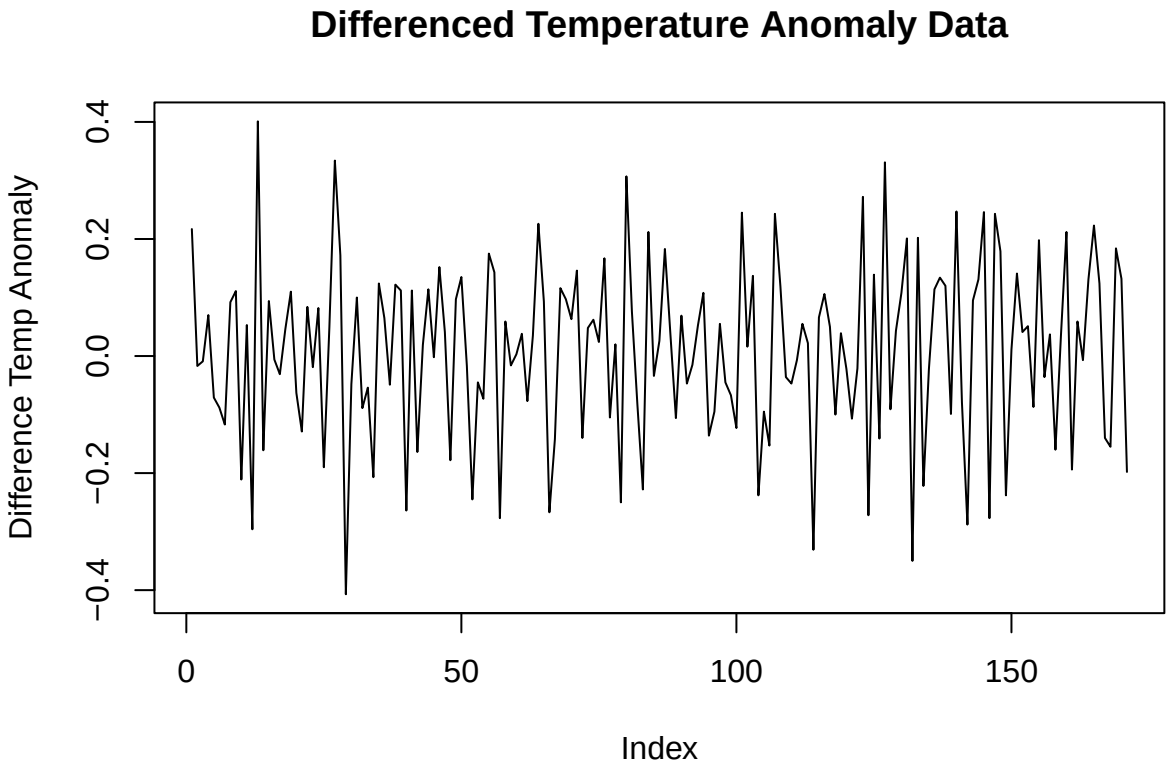
Our first model to consider is an ARIMA model which is a differenced time series model with autoregressive and moving average components. An ARIMA(p,d,q) model has the form

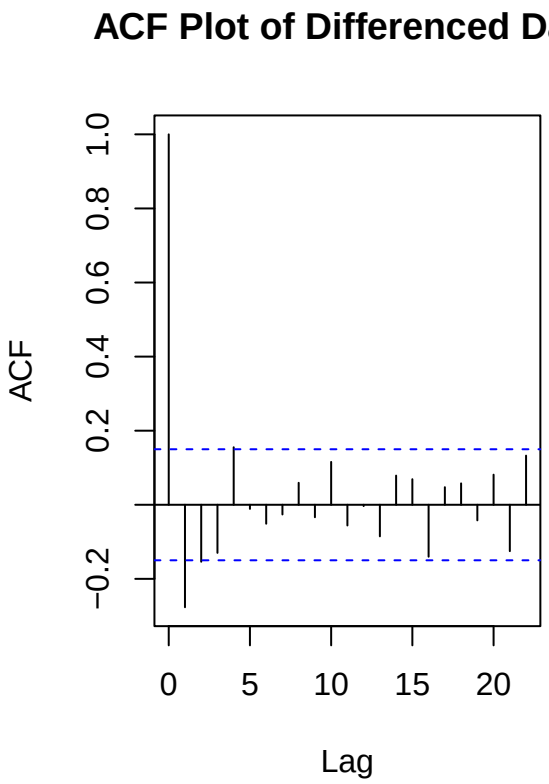
$$\nabla_d Y_t - \mu = \phi_1(\nabla_d Y_{t-1} - \mu) + ... + \phi_p(\nabla_d Y_{t-p} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}$$

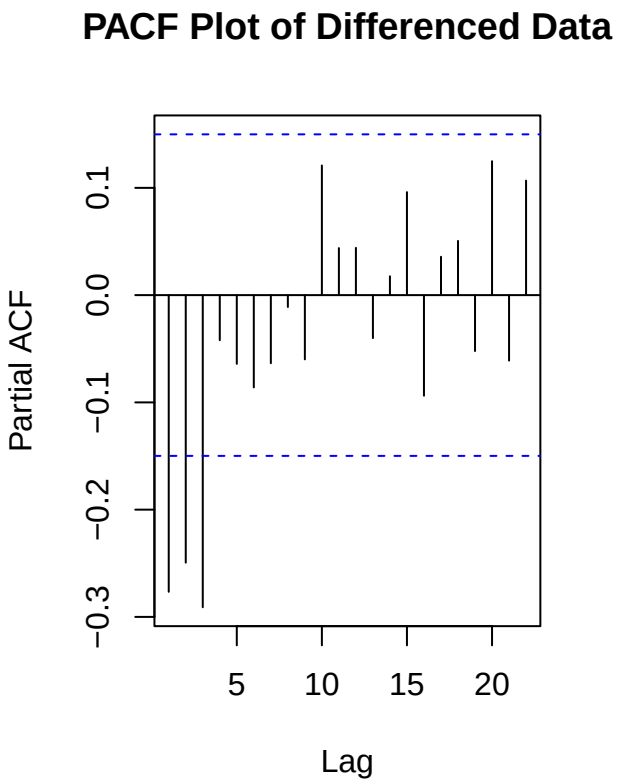
Where $\nabla_d Y_t$ is the d'th difference of the data at time t, ϕ_i are autoregressive coefficients, μ is the mean of the series, θ_i are the moving average coefficients, and $\{\varepsilon\}$ are iid with $Var(\varepsilon_t) = \sigma^2$.

To see if we can use an ARIMA model let's first examine if the first differenced data is stationary in order to see if it is sufficient for ARIMA. The first differenced data is defined as

$$\nabla Y_t = Y_t - Y_{t-1}$$


Differenced Temperature Anomaly Data
Difference Temp Anomaly
−0.4
−0.2
0.0
0.2
0.4
0
50
100
150
Index

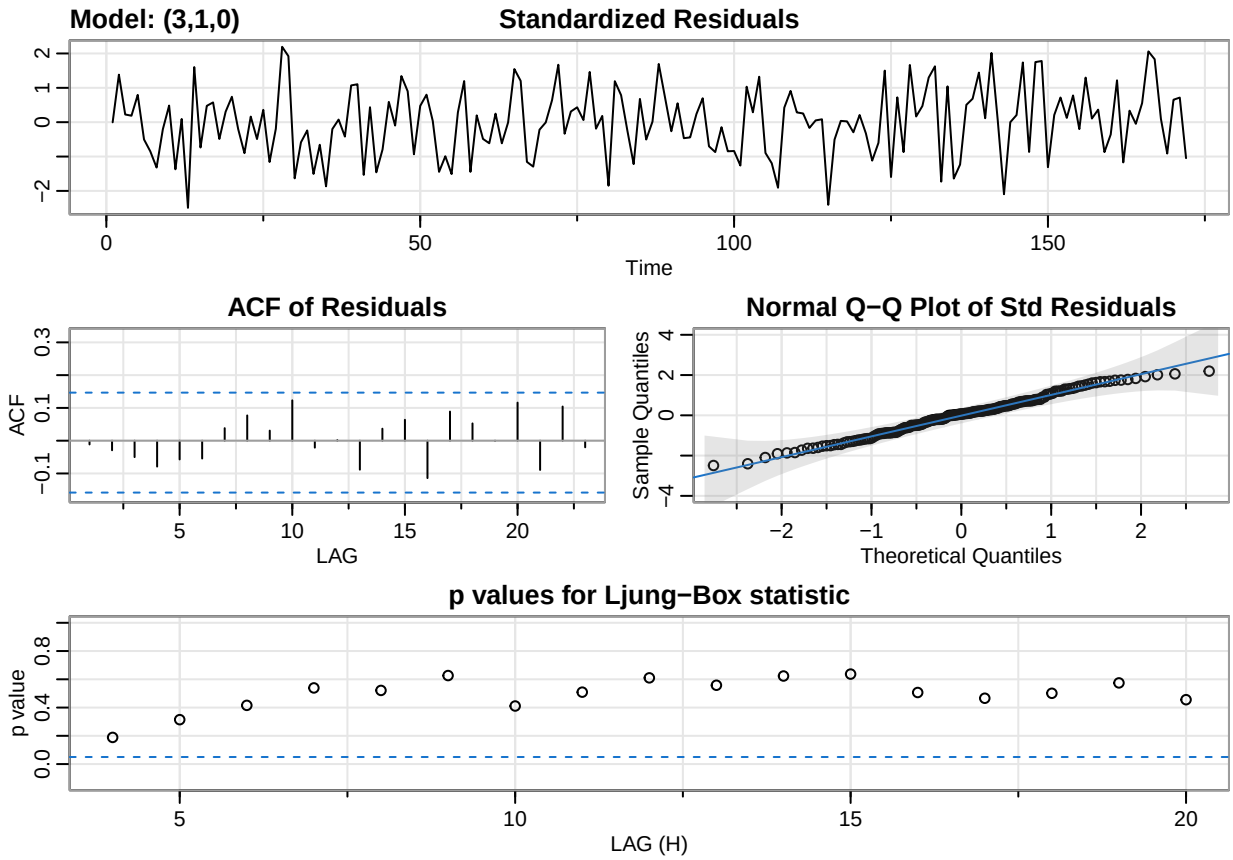




From plotting the differenced data, we can see that it is much more stationary than the original data. It seems to fluctuate about 0 and the variance across time seems mostly stable. We can examine autocorrelations and partial autocorrelations to see that there is no long term leftover autocorrelations so there likely is no trend in this data. All this together suggests that the differenced data can be seen as stationary.

We can make a preliminary diagnosis of which ARIMA model would be suitable by looking at the ACF and PACF plots further. Looking the PACF plot, the appropriate time series model may be AR(3) because there are 3 spikes in the PACF plot. Looking at the ACF plot it is difficult to draw a definitive conclusion because there is a tail off after time lag 1, but not in an obvious way.

So now, let's examine the diagnostics from fitting an ARIMA(3,1,0) model on our data.



From the residuals plots, we can see that the residual diagnostics indicate a good fit. The residuals look stationary centered around 0 and from the Normal QQ plot they very closely follow the normal quantiles. The ACF plot and the Ljung-Box test p-values also indicate that the residuals can be seen as having 0 autocorrelations. So most leftover information is accounted for by the ARIMA(3,1,0) model.

Finally, for this model we can verify which ARIMA model would be most suitable by computing the AIC model selection criterion for all combinations of AR and MA orders from 0 to 3 each.

Table 1: AIC values for different p,q for differenced data

	q=0	q=1	q=2	q=3
p=0	-0.923	-1.094	-1.120	-1.108
p=1	-0.992	-1.114	-1.108	-1.109
p=2	-1.046	-1.115	-1.112	-1.115
p=3	-1.123	-1.117	-1.111	-1.103

Using the AIC criterion we look for the smallest AIC value, which we see the most fit for the graph. From the above table, ARIMA(3,1,0) is best with the smallest value of -1.123, which we already fit from the preliminary examination.

2.3) Methods: Trend Modeling with Splines

Another approach to modeling time series data with trend is to model the trend and the residuals to produce forecasts. In this project, we will work with cubic regression splines and then an ARMA model on the residuals from the trend estimation. Splines are a method of fitting piecewise cubic polynomials to

regions of the data joined smoothly at knots. We will be using a function that finds the suitable number of knots and their locations. After fitting the trend, we can model the residuals from this trend with an ARMA model.

Results and Forecasting

Now, let's examine our final ARIMA(3,1,0) model and perform forecasts with this model.

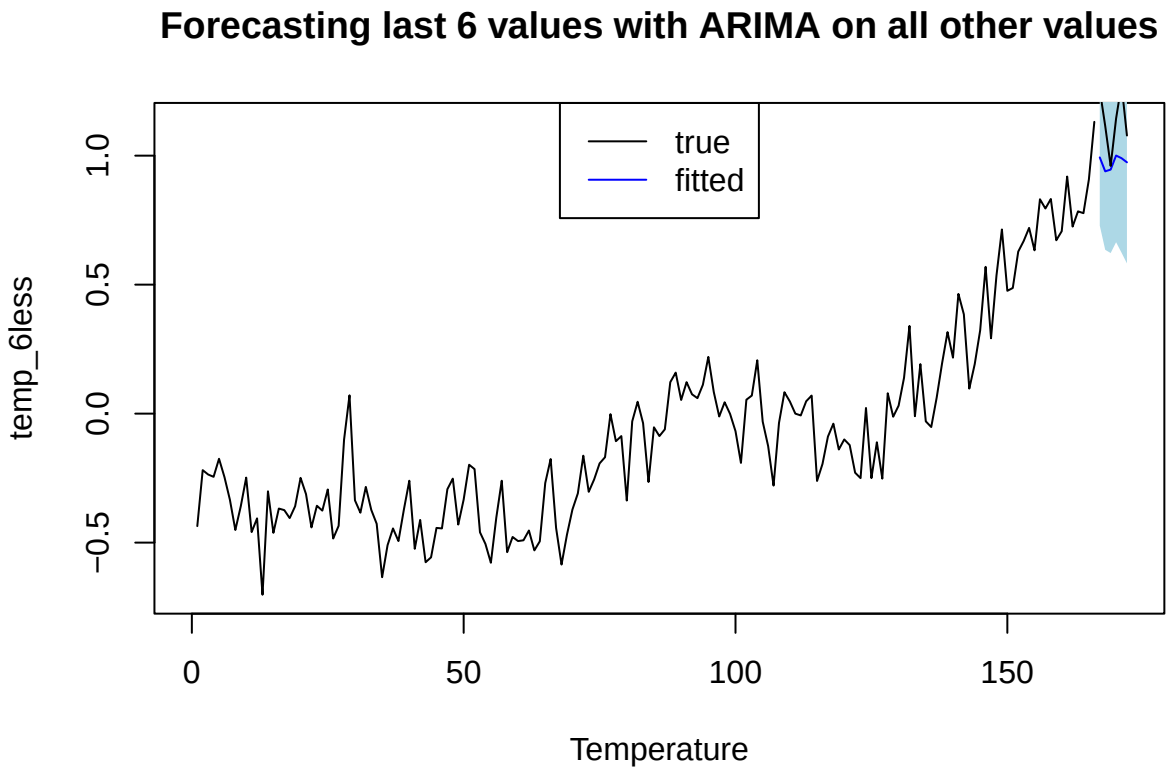
ARIMA(3,1,0) coefficients table

	ar1	ar2	ar3
	-0.4113	-0.343	-0.2837
s.e.	0.0738	0.07587	0.07386

From this summary table, we can see that all of the coefficients are signficant at level 0.05 since each have a z-value over 1.96. We can also see that our final model writes our data as approximately

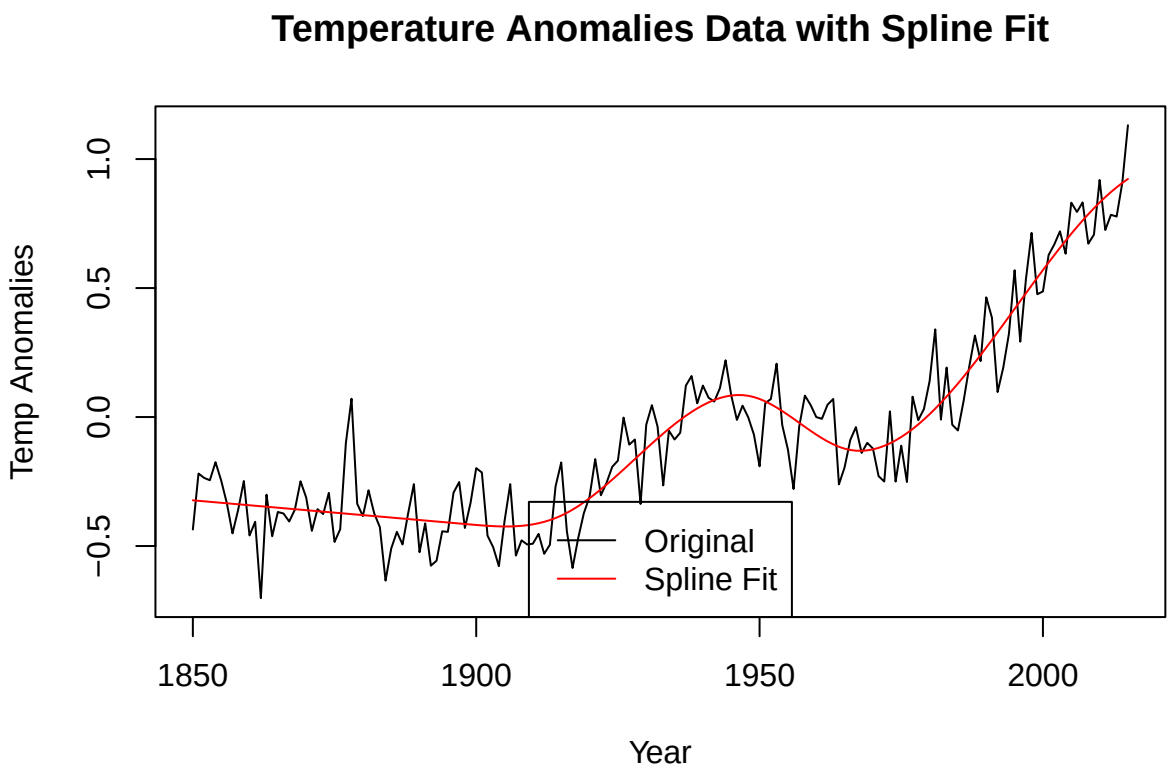
$$\nabla Y_t = -0.4113\nabla Y_{t-1} - 0.343\nabla Y_{t-2} - 0.2837\nabla Y_{t-3}$$

Now, we would like to use this model in predictions. To do this, we will fit the same model on all the data without the last 6 data points then see the performance of the forecasts compared to the last 6 values.

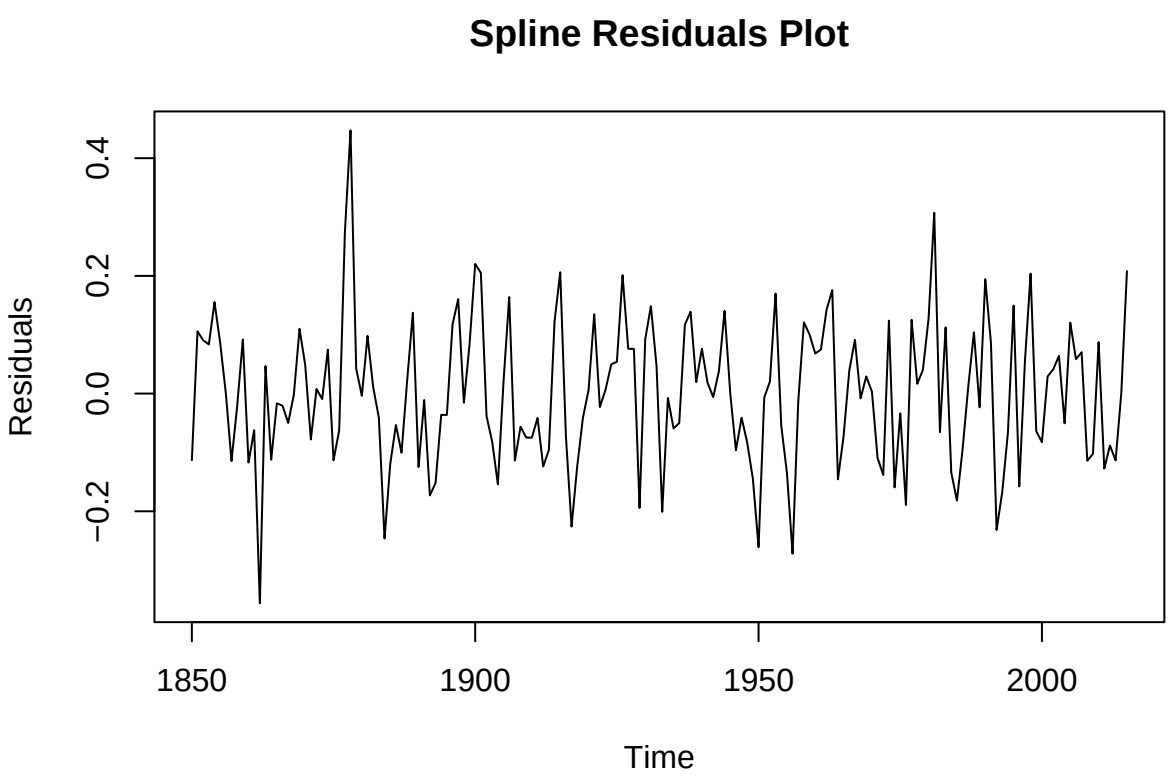


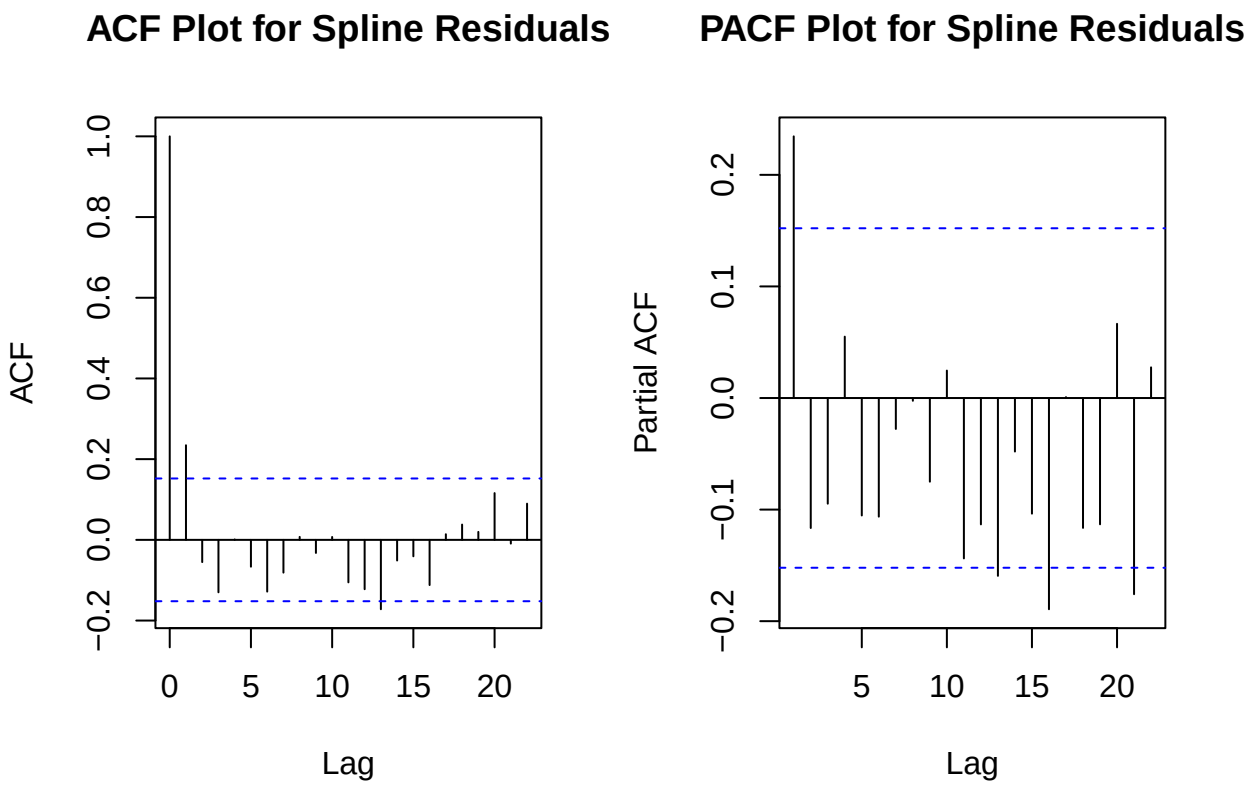
We can see that from this forecasting, thetrue values are within the prediction bands of the ARIMA(3,1,0) method which suggests that we have an effective forecast. The prediction bars are a bit large, but there is always inherent uncertainty with forecasting.

Now, we can also examine the results of the spline fit.



From looking at the spline fit with the original data, we can see that the fit does seem to capture the overall trend so it seems suitable for this task. Now, we look at modelling the residuals from this fit.


Spline Residuals Plot
Residuals
0.4
0.2
0.0
−0.2
1850
1900
1950
2000
Time



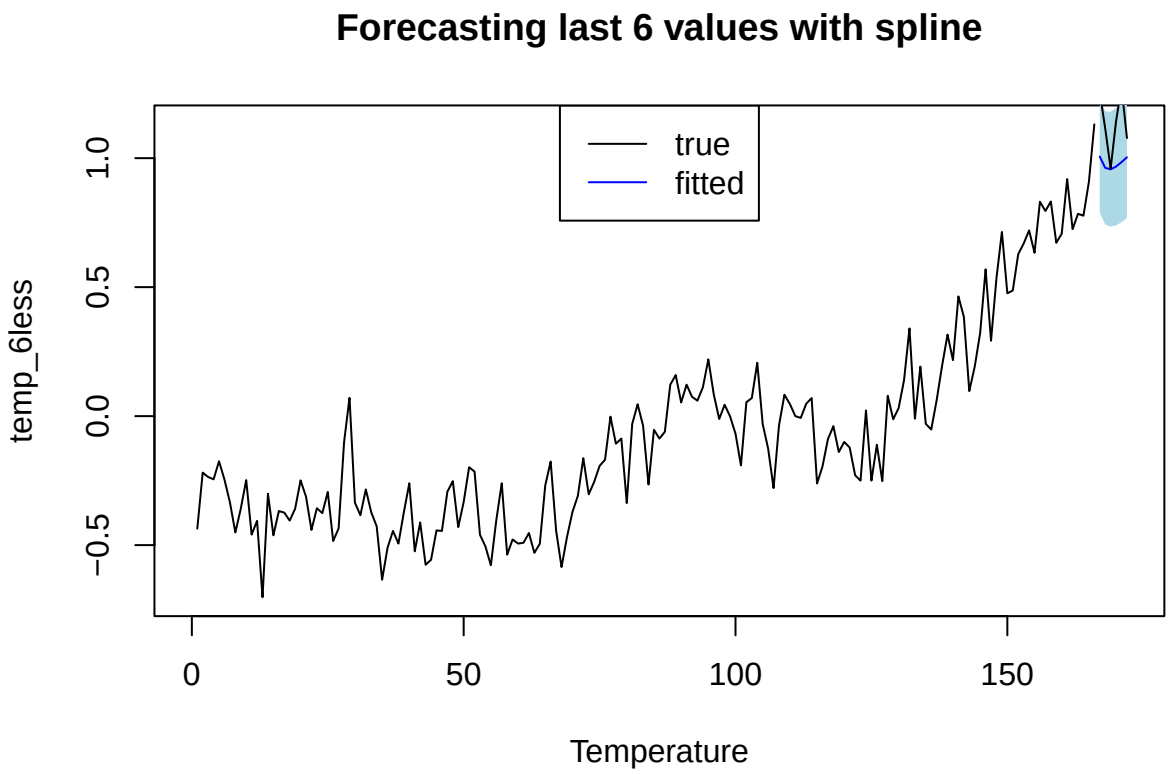
We can see that the residuals do look stationary centered around 0 with stable variance. The PACF plot does not suggest a strong AR or MA model so it could need a mixed ARIMA type of model. Again, we examine the most suitable model by computing the AIC model selection criterion for values of AR and MA order from 0 to 3 each.

Table 3: AIC values for different p,q for spline residuals

	q=0	q=1	q=2	q=3
p=0	-224.1407	-232.8736	-230.9194	-233.5559
p=1	-231.6972	-230.8931	-231.1319	-245.4820
p=2	-232.1223	-246.7585	-231.7642	-244.2188
p=3	-231.7092	-230.9565	-244.3313	-242.3293

Looking at the AIC criterion the AIC value is the lowest with p=2, and q=1, which gives an ARMA(2,1) with value -246.7585. So to produce predictions we will use an ARMA(2,1) model. So in this case, we also seem to have an effective fit by modeling the trend with splines as well.

Now, let's also perform some forecasting with the trend splines model by seeing the performance on the last 6 data points. We will perform a linear extrapolation of the trend and model the rough with our ARMA(2,1) estimation.



Again, the true values are mostly in the prediction intervals, but there might be a bit of it that is not in the interval. The prediction here also performs pretty well, but it seems a bit more off than the ARIMA model. This shows that both methods are suitable to this task, but which one performs better depends on the task.

Conclusion

In our project it was important to identity the nature of data and recognize that this is a time series, and how our background knowledge of this time series explains the nature of variation of the data. Once we understand the trend of the data, we used the first difference of the series in order to use ARMA in the model. Determining if this data is a stationary series is crucial thus after we took the first difference we conducted various graphical and hypothetical tests to confirm that it was stationary. We also figure we would be able to use a different method of forecasting data through estimating trend through spline. A good indicator of whether our models are a good predictor of forecasting future data is to use the data that excludes the last few points of the data, then we can compare the "predicted" data against the actual last few points of data to determine retroactively if our model is a good indicator of data points that did not happen/recorded yet. In our case for we compared the last 6 years and it appears that both of the models (ARMA, spline) are reasonable predictors as compared to the actual data for the last 6 years. Though something that remains to be desired, is we could not factor the seasonality into the model because there is a short term natural variability such as LA Niña, that couldn't be confirmed with our tools we used in this project. Additionally, another shortcoming of this analysis is that the prediction intervals for the spline estimates may be narrow since it only takes into account the variability in modelling the rough, but not the extrapolation of the trend. This means that the trend estimates may be a bit over-optimistic. However this project served its purpose on the grounds of trying to estimate what our next few years would be if the trend of global warming shall continue, through forecasting we can raise awareness in this issue that requires global cooperation and statistical analysis to solve the problem.

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(astsa)
library(forecast)
library(Hmisc)
temp <- read.csv("TempNH_1850_2021.csv", header=T)

plot(temp$Year,temp$Anomaly,type="l",main="Temperature Anomalies Data",ylab="Temp Anomalies",xlab="Year
par(mfrow=c(1,2))
acf(temp$Anomaly, main="ACF Plot of Temperature Anomaly Data")
pacf(temp$Anomaly, main="PACF Plot of Temperature Anomaly Data")
diff_dat <- diff(temp$Anomaly)
plot(diff_dat, main="Differenced Temperature Anomaly Data", ylab="Difference Temp Anomaly", type="l")

par(mfrow=c(1,2))
acf(diff_dat, main="ACF Plot of Differenced Data")
pacf(diff_dat, main="PACF Plot of Differenced Data")
arima_fit <- capture.output(sarima(temp$Anomaly,p=3,q=0,d=1))
res <- arima(temp$Anomaly,order=c(3,1,0))$residuals
Box.test(res, type="Ljung-Box")
#Model selection for for 0<=p<=3, 0<=q<=3 ARIMA(p,1,q)
aic_table = matrix(nrow=4,ncol=4)

for (i in 0:3){
  for (j in 0:3){
    aic_table[i+1,j+1] = sarima(temp$Anomaly,p=i,q=j,d=1,details=F)$AIC
  }
}

aic_table
trend_spline=function(y, lam) {
  # Fits cubic spline estimate of trend
  # If lam contains a single number, then the corresponding
  # Box-Cox transformation is made, and a spline model is fitted
  # If lam is a vector, then the best transformation is obtained from   the
  # candidates in 'lam', and then spline is
  # fitted for the best transformation after
  # deleting knots using backward stepwise regression
  #Output,
  # 1. transformed y: ytran (if lam is a vector, this corresponds to   the
  # best transformation)
  #  2. trend: the fitted spline estimate
  # 3. residual: the remainder, ie, ytran-trend
  # 4. rsq, R^2 values for different transformations
  # 5. lamopt: the best chosen transformation from 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$resid; sse=sum(res^2)
    ssto=(n-1)*var(ytran);
    rsq[i]=1-sse/ssto
  }
  ii=which.max(rsq); lamopt=lam[ii]
  if (lamopt==0) {
    ytran=log(y)
  } else {
    ytran=y^lamopt
  }
  newdat=data.frame(cbind(ytran,xx))
  ft=lm(ytran~.,data=newdat);
  best_ft=step(ft, trace=0)
  fit=best_ft$fitted; res=best_ft$resid
  result=list(ytrans=ytran, fitted=fit, residual=res, rsq=rsq, lamopt=lamopt)
  return(result)
}
#summary(arima(temp$Anomaly,order=c(3,1,0)))
#pander::pander(arima(temp$Anomaly,order=c(3,1,0)))
n <- nrow(temp)

#Divide data into n-6 and last 6
temp_6less <- temp$Anomaly[1:(n-6)]
temp_last6 <- temp$Anomaly[(n-5):n]

#Fit on all but last 5 values
arima_6less <- arima(temp_6less,order=c(3,1,0))

#Get predictions
h <- 6
m <- n-h
forecast <- predict(arima_6less,n.ahead=6)
upper <- forecast$pred + 1.96*forecast$se
lower <- forecast$pred - 1.96*forecast$se

#Plot forecasts
plot.ts(temp_6less, xlim = c(0,n), xlab = "Temperature", main="Forecasting last 6 values with ARIMA on a
polygon(x=c(m+1:h,m+h:1), y=c(upper,rev(lower)), col='lightblue', border=NA)
lines(x=m+(1:h), y=forecast$pred,col='blue')
lines(x=m+(1:h), y=temp_last6,col='black')
legend("top", legend = c("true","fitted"), lty=c(1, 1), col = c("black","blue"))
```

```r
spline_fit <- trend_spline(temp$Anomaly[1:(n-6)], 1)
plot(temp$Year[1:(n-6)],temp$Anomaly[1:(n-6)],type="l",main="Temperature Anomalies Data with Spline Fit
lines(temp$Year[1:(n-6)],spline_fit$fitted, col="red")
legend("bottom", legend=c("Original","Spline Fit"), lty=c(1,1), col=c("black","red"))
plot(temp$Year[1:(n-6)],spline_fit$residual,type="l",main="Spline Residuals Plot",xlab="Time",ylab="Resi
par(mfrow=c(1,2))
acf(spline_fit$residual,main="ACF Plot for Spline Residuals")
pacf(spline_fit$residual,main="PACF Plot for Spline Residuals")
#Model selection for for 0<=p<=3, 0<=q<=3 ARIMA(p,1,q)
aic_table_spline = matrix(nrow=4,ncol=4)

for (i in 0:3){
  for (j in 0:3){
    aic_table_spline[i+1,j+1] = arima(spline_fit$residual,order=c(i,0,j))$aic
  }
}

aic_table_spline
rough_arma <- arima(spline_fit$residual,order=c(2,0,1))
trend_extrap <- approxExtrap(temp$Year[1:(n-6)],spline_fit$fitted,xout=temp$Year[(n-5):n])
forecast_trend <- trend_extrap$y
forecast_rough <- predict(rough_arma,n.ahead=6)
upper <- forecast_rough$pred + 1.96*forecast_rough$se
lower <- forecast_rough$pred - 1.96*forecast_rough$se

#Plot forecasts
plot.ts(temp_6less, xlim = c(0,n), xlab = "Temperature", main="Forecasting last 6 values with spline")
polygon(x=c(m+1:h,m+h:1), y=c(upper+forecast_trend,rev(lower+forecast_trend)), col='lightblue', border=
lines(x=m+(1:h), y=forecast_rough$pred+forecast_trend,col='blue')
lines(x=m+(1:h), y=temp_last6,col='black')
legend("top", legend = c("true","fitted"), lty=c(1, 1), col = c("black","blue"))
```