PSTAT 174 Time Series Final Project

Tanner Berney

December 4, 2021

Table of Contents

1. Abstract. 2

2. Introduction. 2

3. Analysis. 3

4. Conclusion. 12

5. References. 12

6. Appendix. 12

Abstract

Since the Industrial Revolution, we have seen the worlds temperature rise at an alarming rate. For this project, we will be looking at the average temperature of Earth throughout the last 100 years to see if it still rising and see if we can forecast the temperature change for the next few years. Predicting the future data points will give us a better understanding of how much we can expect to see temperatures rise the next few years and how worried we should be with the increase.

I started by first getting a data set that had the average increase and decrease of temperatures compared to the overall average of the past 100 years. I first looked at making the time series data positive and running it through the BoxCox function to see if a transformation was needed. After deciding it was not necessary, I noticed some upward trend and tried to stabilize with a log function. This did not have much affect on the data so I proceeded with the original test data. Using differencing I was able to eliminate this upward trend. After looking at the ACF and PACF of the differenced data, I looked at the following models to examine: ARIMA(0,1,2), ARIMA(1,1,2), ARIMA(2,1,2), ARIMA(3,1,0), ARIMA(3,1,1), ARIMA(3,1,2). For each of the models I looked at the ARIMA fit and found the ARIMA(3,1,0) had the lowest AICC of all the models so I used this one. Once I plotted the forecasted values against the true values, it was clear to see that they are in the confidence intervals giving the assumption that ARIMA(3,1,0) is an appropriate model, however I believe more analysis is needed.

Introduction

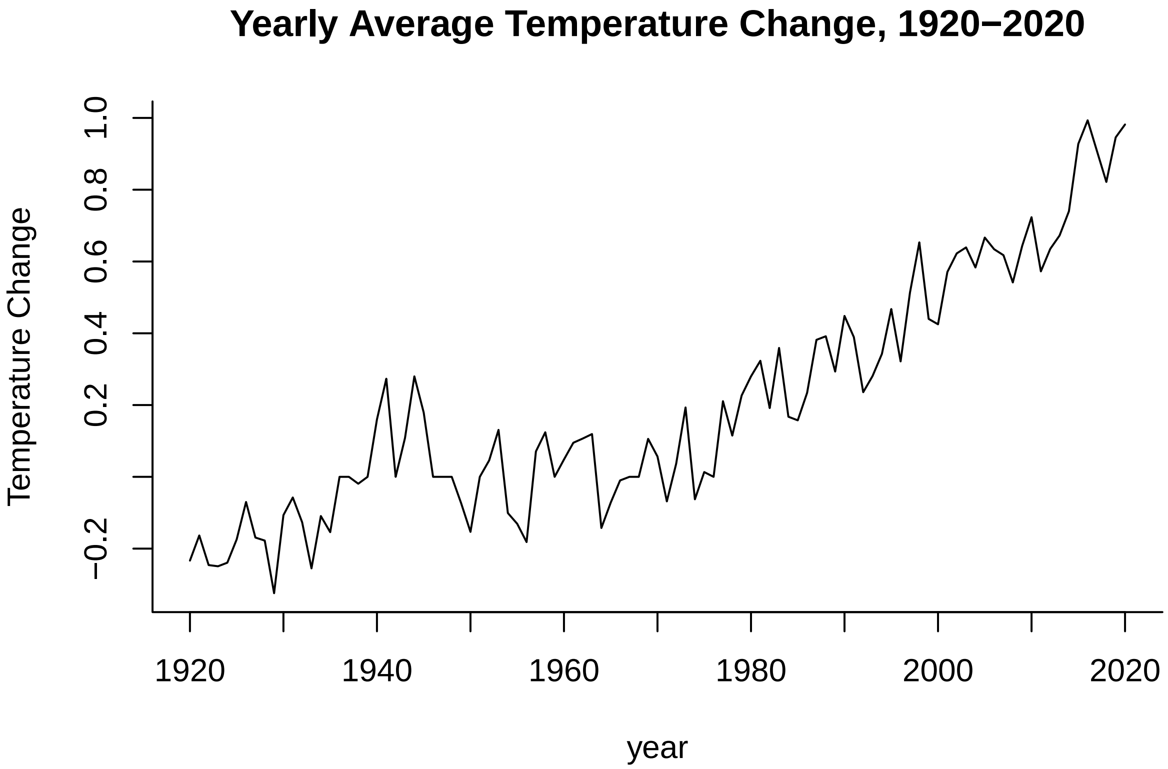
The rising temperatures of Earth are an early sign of global warming. The more information we are able to collect on the topic will better prepare us for the future and identify how big of a problem this is becoming. The data set I used comes from the National Centers for Environmental Information which contains the average temperature of each month from 1920 to 2020. I plan to use this data set to look at how the temperature of Earth has drastically changed over the years and the direction it is heading in for the future.

To start this analysis I first took the data and partitioned the month averages into year averages. This was helpful since the each month varied so much with the different seasons. Having them be yearly averages eliminates this and helps our model have less noise in the calculations. Once this was complete, I shifted values to make them positive in order to do a Box-Cox transformation. After seeing that the Box-Cox method was not needed, I took the log of the data to eliminate some of the upward trend. However, this resulted in a very similar plot, so I continued with the original test data. I applied differencing at lag 1 which helped to eliminate the upward trend in the plots. This led to calculating the ACF and PACF to identify some models to test the differenced data on. I was left with models ARIMA(0,1,2), ARIMA(1,1,2), ARIMA(2,1,2), ARIMA(3,1,0), ARIMA(3,1,1), ARIMA(3,1,2) to do diagnostic checking on. After checking through the confidence intervals of the ARIMA coefficients and comparing AICc's, it was clear that ARIMA(3,1,0) was the most appropriate model. This model seemed to do a fairly good job of forecasting the following data points, as all values were within the confidence intervals. The fitted model however, had many of the forecasted points further than I would have wanted from the true values. Even though this model seems to work, I believe that additional analysis is needed to be more accurate in the prediction for the future. My analysis makes me believe that earths temperature is rising at the same gradual rate, but further analysis is required to get a better estimation of how quickly this rate is changing.

The data used can be found at ncdc.noaa.gov in their time series librbary. All computations and statistical analysis of the time series was done using RStudio software.

Analysis

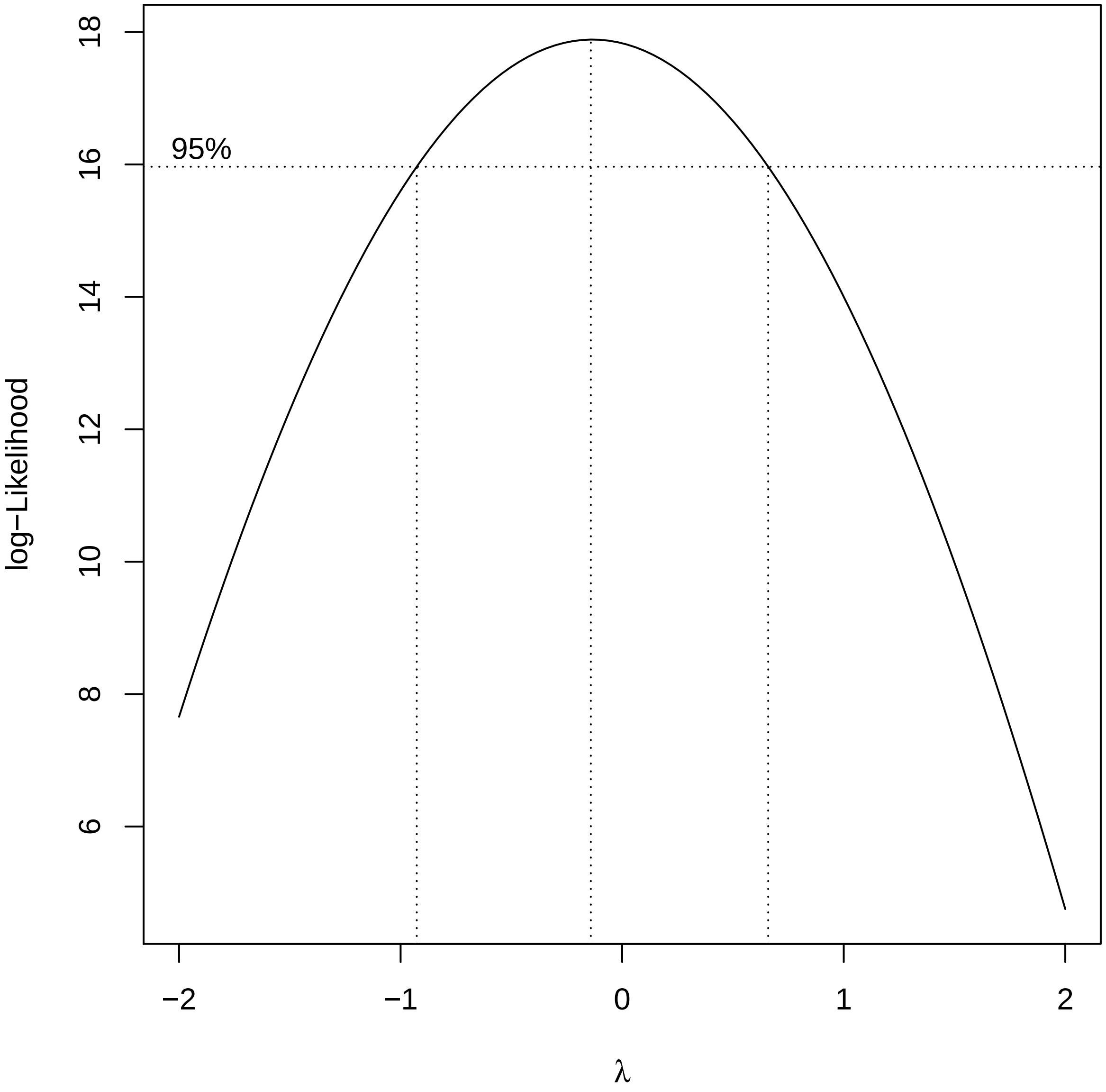
We will start the analysis by looking at the plotted time series.



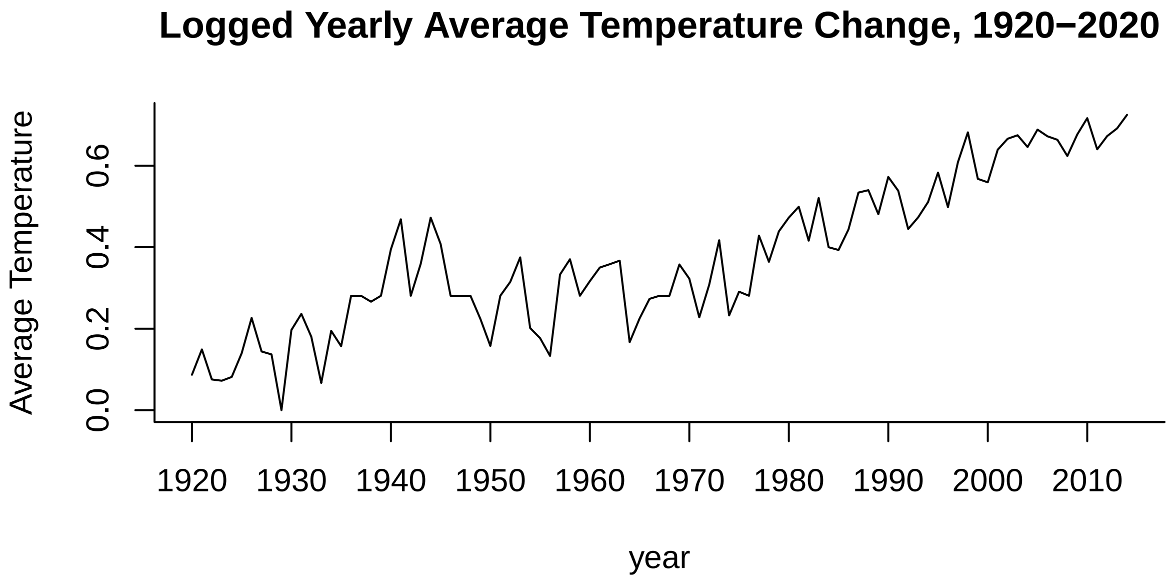
Looking at the plot we can see right away that the current data shows some signs of being non-stationary. From the 1920's to about the 1970's the average seems to remain about the same, but after that it looks to fluctuate in the upward direction. The variance of the plot does not seem to have any big issues, except for it being very small at certain points. From this plot, we can see that there is an upward trend that needs to be dealt with and I will also look at the normality of the data points to see if transformations are required.

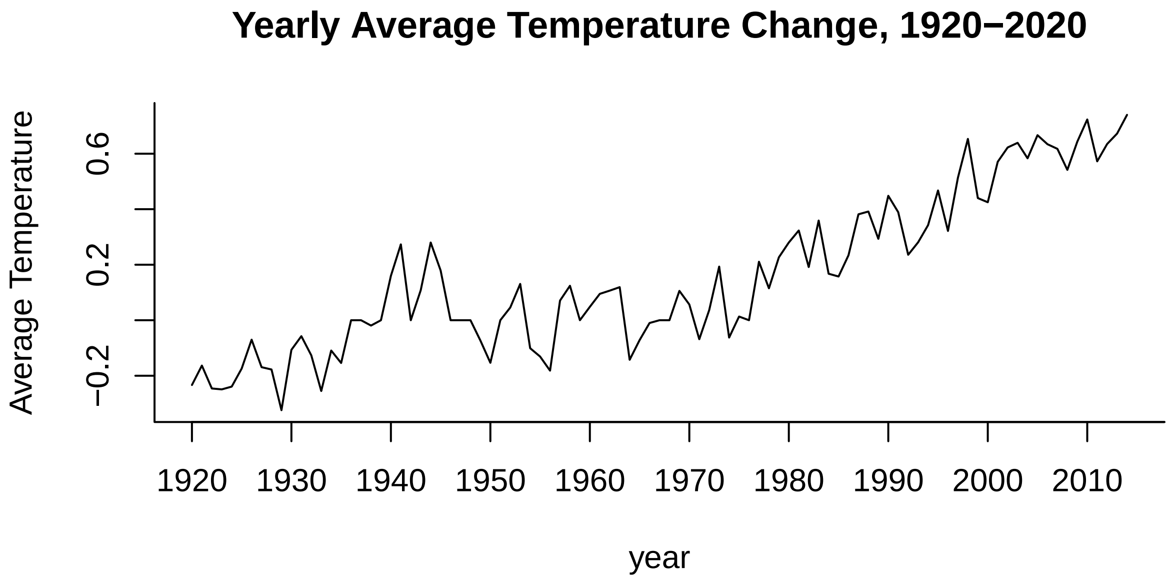
Transformations

I first looked at doing a Box-Cox transformation on the data to see if it would make it look more normal. To do this, I first changed the data point values to be all positive since it is required to run a Box-Cox transformation. For these values I added the absolute value of the smallest observation to each one to center them around 0, and then added a constant of 1 to them.



Looking at the Box-Cox plot, we can see that zero lies within the confidence interval. This lead me to believe that a Box-Cox transformation was not needed and proceeded to try a log transformation on the data instead.

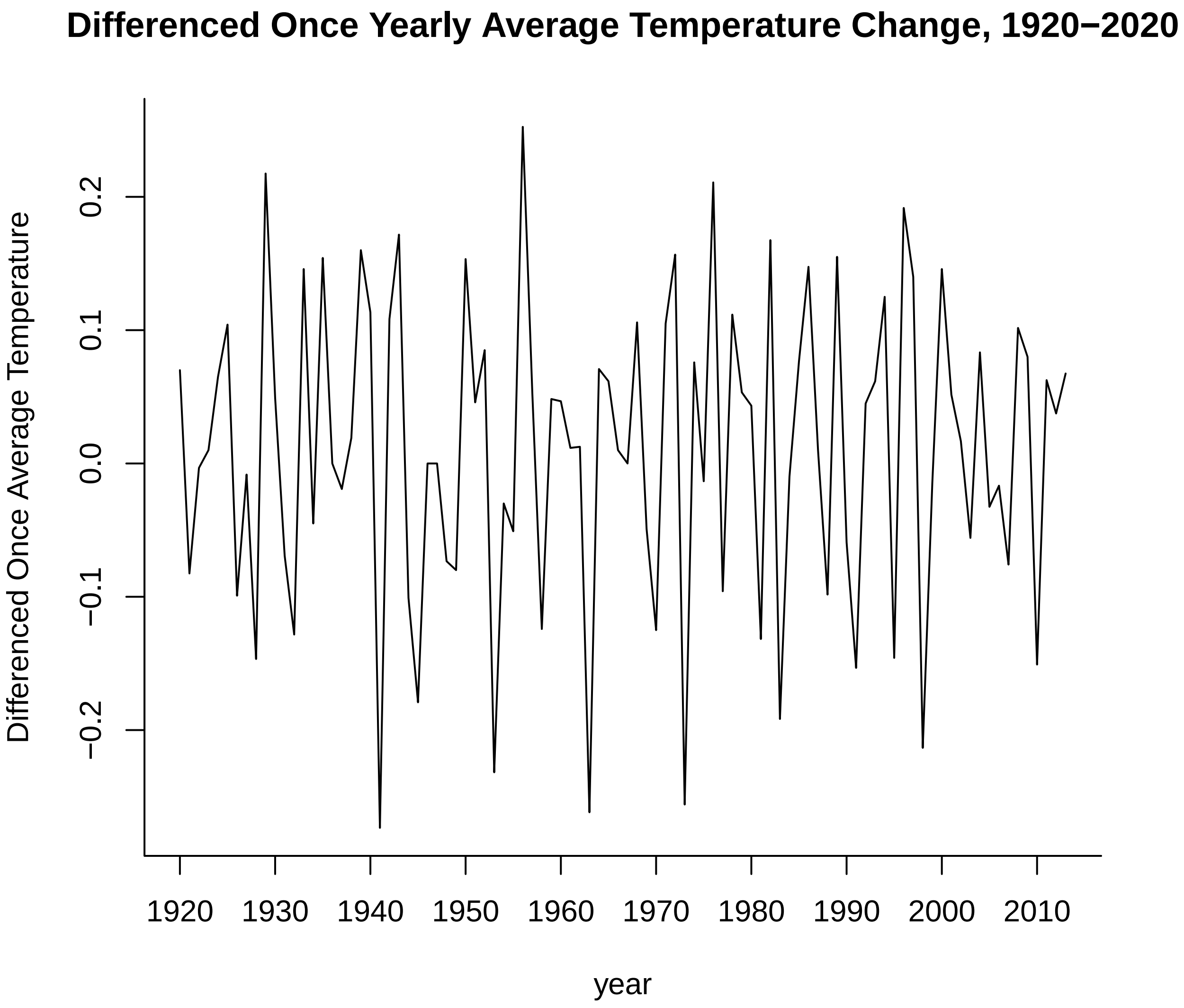




I was hoping that when I logged the data it would help to make the upward trend more linear as well as make the variance be more stable. However, the plot of the logged data did not differ that much from the original data. With the complications of interpretation that come with logging the data, I believed it was better to stick with the original data rather than try and continue with the logged data.

Differencing the Data

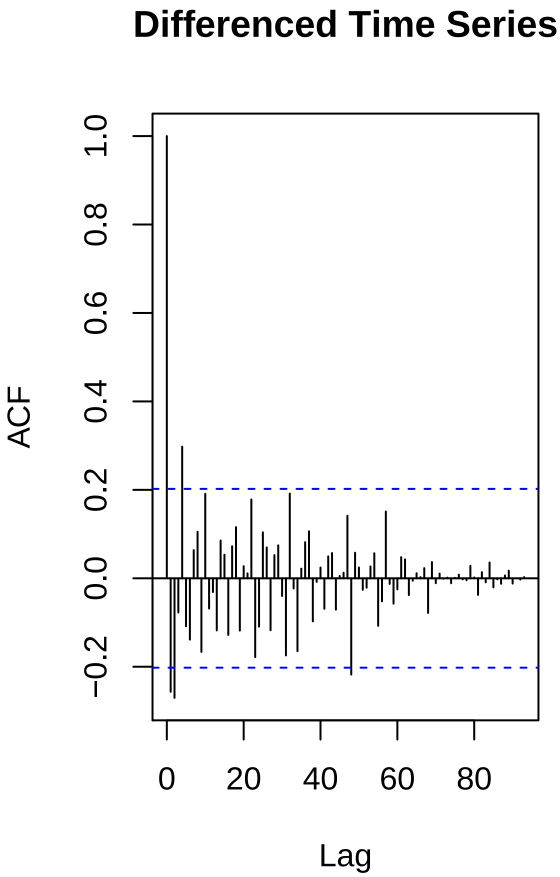
Since the Box-Cox and log transformations were not necessary, I decided to try to difference the data in order to remove the upward trend we see.

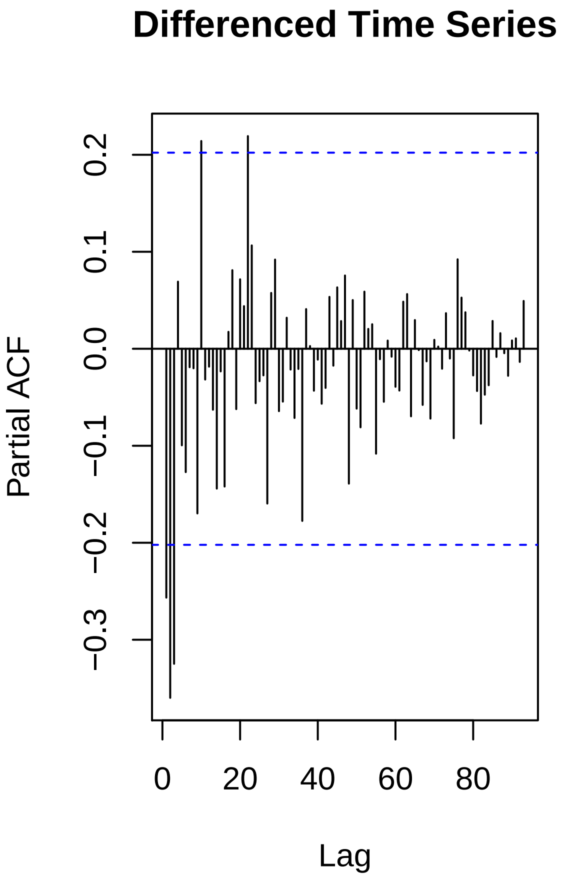


The plot above shows the data that has been differenced at lag 1. This difference has pulled the plot down around zero allowing it to follow a much more stable mean making it stationary. The variance of the data does not seem to have any large spikes or intrusions to cause concern for us. To check for any over differencing I took the variance from the data before and after the difference to compare them. I found that before differencing the data had a variance of 0.08297254 and after it was 0.01355168. This huge decrease in variance proves that differencing at lag 1 was a reasonable move to make in order to eliminate trend and have a more stable variance.

ACF and PACF

We will now look at the ACF and PACF of the data differenced at lag 1 in order to determine a model that best fits the data.





Looking at the resulting ACF and PACF, I stated to think about some models that could work with the data. Some of the ideas I started with was an AR(3) process or MA(2) because of the lag spikes from the PACF and ACF at 3 and 2 respectively. This also led me to the model of ARMA(3,2) to test and see if it worked better. I decided it was best to try some models that had more or less terms to get a bigger picture of which model worked best. This led me to the following ARIMA models(Not ARMA since we used a difference at lag 1): ARIMA(0,1,2), ARIMA(1,1,2), ARIMA(2,1,2), ARIMA(3,1,0), ARIMA(3,1,1), ARIMA(3,1,2). After identifying these models I ran ARIMA models and looked at the terms along with their 95% confidence intervals.

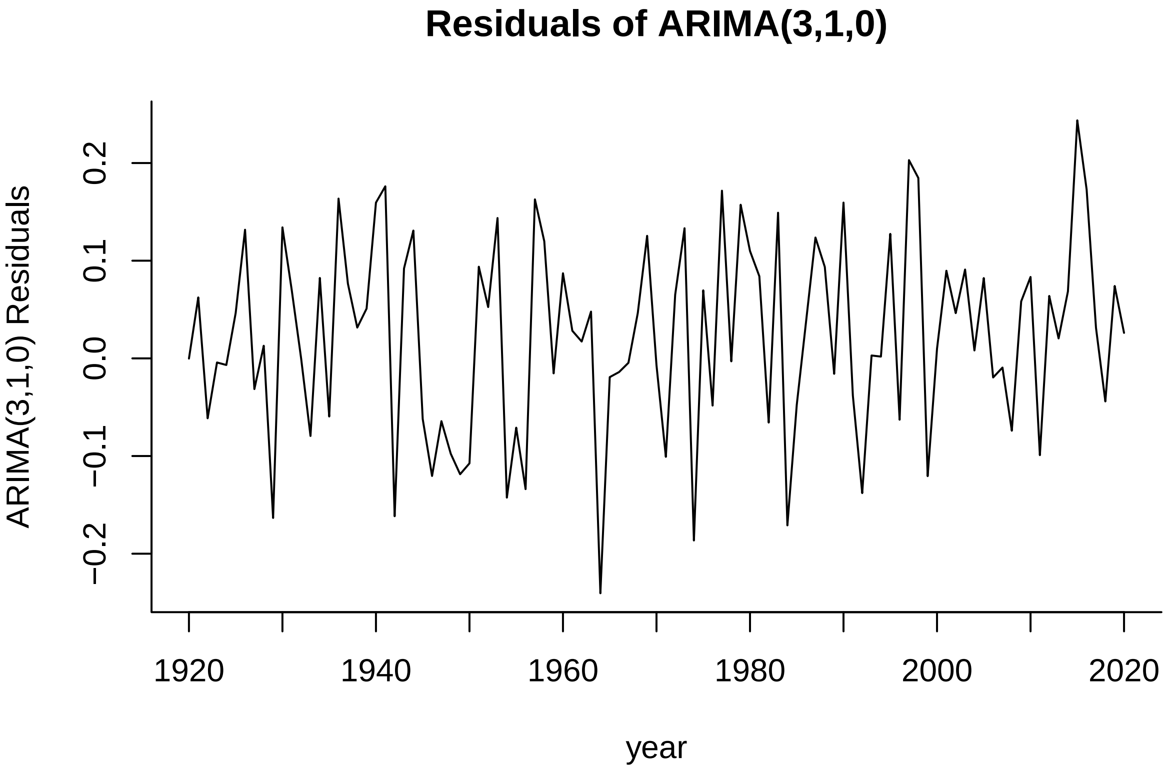
	MA(q)	Lower CI	Upper CI	AR(p)	Lower CI	Upper CI
ARIMA(0,1,2)	-0.1937	-0.3615	-0.026	NA	NA	NA
ARIMA(1,1,2)	-0.228	-0.48	0.0241	-0.0913	-0.6389	0.4564
ARIMA(2,1,2)	0.2419	-0.2017	0.6854	-0.4483	-0.825	-0.0716
ARIMA(3,1,0)	NA	NA	NA	-0.2529	-0.4414	-0.0643
ARIMA(3,1,1)	0.405	-0.1751	0.9852	-0.3796	-0.5897	-0.1695
ARIMA(3,1,2)	0.2292	-0.5562	1.0146	-0.4015	-0.6277	-0.1754

The table above shows the estimated coefficients for the ARIMA models as well as 95% confidence intervals of the coefficients. We can look at these confidence intervals of the values to help us decide on which models we can eliminate. Looking at ARIMA(1,1,2) we see that the MA term confidence interval contains zero. This presents a problem as we can not have the defining term in our model be equal to zero. The same can be said for ARIMA(2,1,2), ARIMA(3,1,1), and ARIMA(3,1,2) with the MA terms confidence intervals all containing zero. This allows us to eliminate these models from selection as well, leaving us with ARIMA(0,1,2) and ARIMA(3,1,0) models.

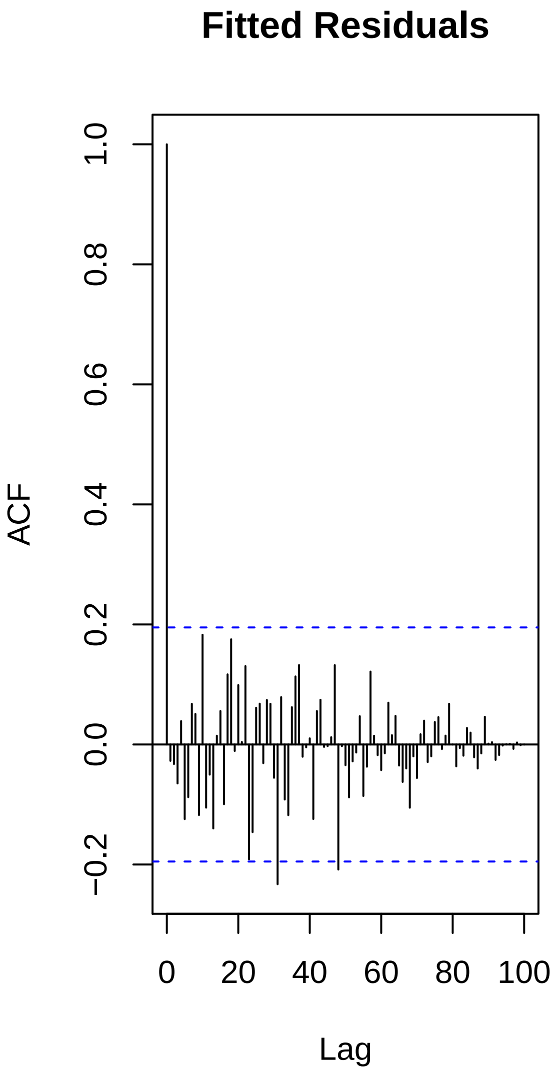
From these models, we will look at the AICC to determine which model has the best fit. The AICC of ARIMA(0,1,2) is -159.81 and for ARIMA(3,1,0) it is -163.68. Taking the model with the lowest AICC will leave us with ARIMA(3,1,0) as being the best model candidate for this data. In addition, the AICC of ARIMA(3,1,0) was the lowest of all the models I looked at, making me confident in the decision to use it.

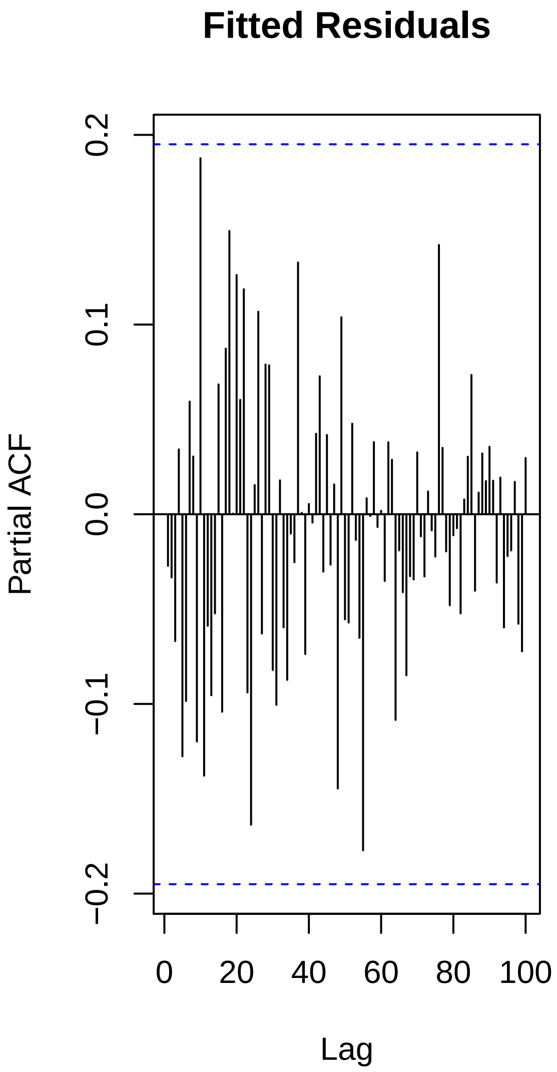
Model Testing

Once I decided on the model, I plotted the residuals to analyze as well as the ACF and PACF.

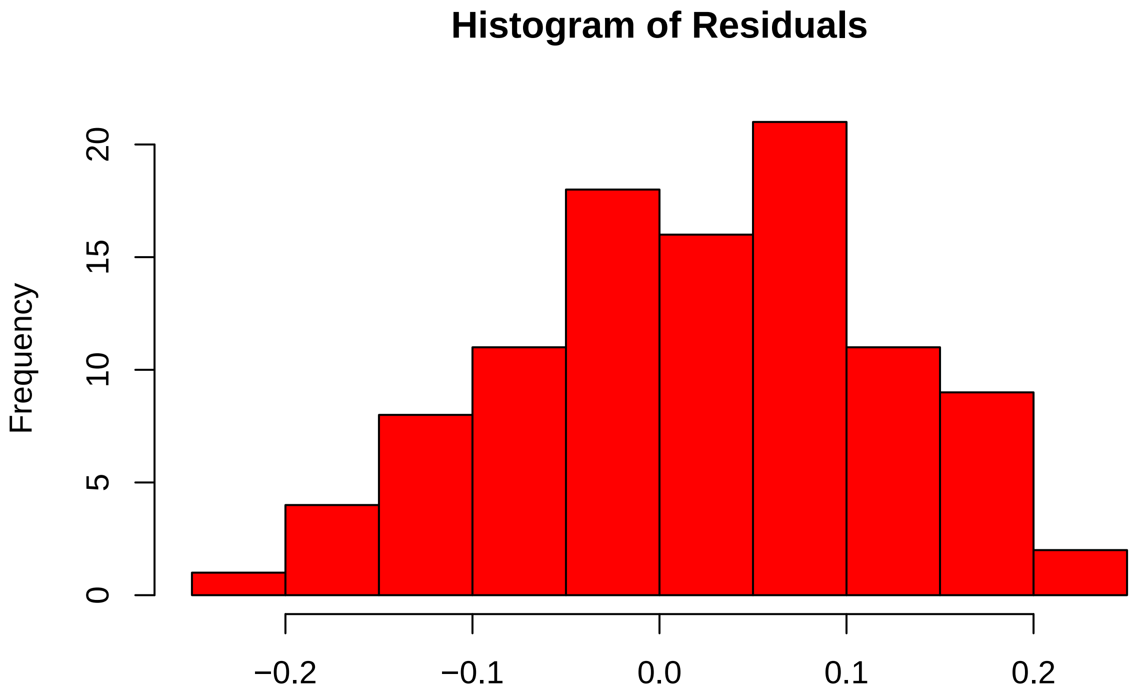


First looking at the plot of the residuals, we can see that it does not seem to display any trend or change in variance. This is a good indication of normality but lets also look at the ACF and PACF.

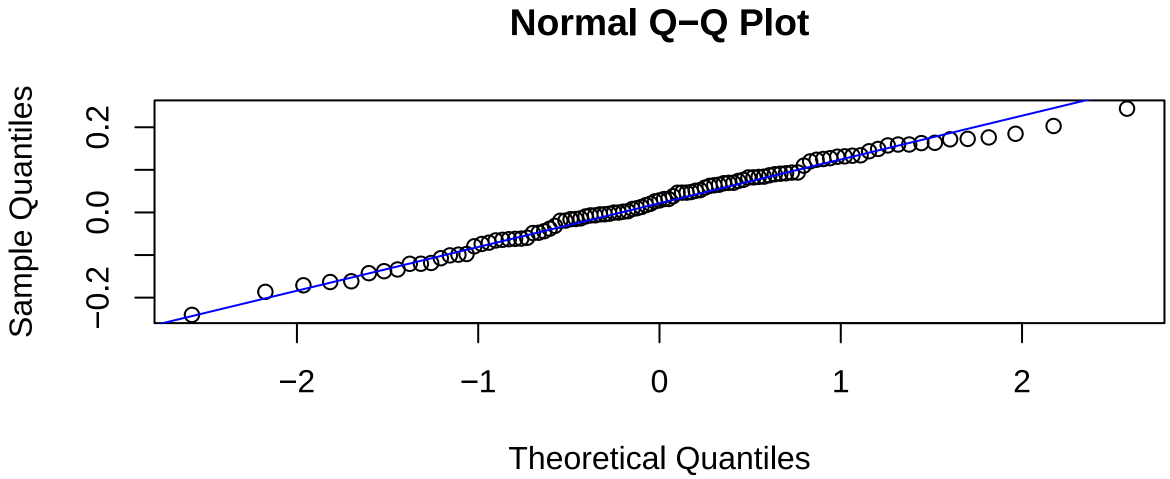




The ACF of the residuals shows that most fall within the confidence interval, except for lags 31 and 47. However, when looking at the PACF we can see that the values all fall within the confidence interval.This leads me to believe there may be an issue due to the outside noise. However, I am confident the model fits and since the residuals seem to show some behavior of normality I continued with ARIMA(3,1,0).



The histogram seems to be normally distributed with the right side having a higher number of values which suggest there could be values within the data that are not normally distributed.



Taking a look at the Q-Q Plot, This further leads me to believe that there are indeed some points within the distribution that are not normally distributed. They follow close enough to the line for me to classify them as resembling a normal distribution. I run Shapiro-Wilk and Box-Ljung tests to get a better sense of the data.

Shapiro-Wilk and Box-Ljung Tests

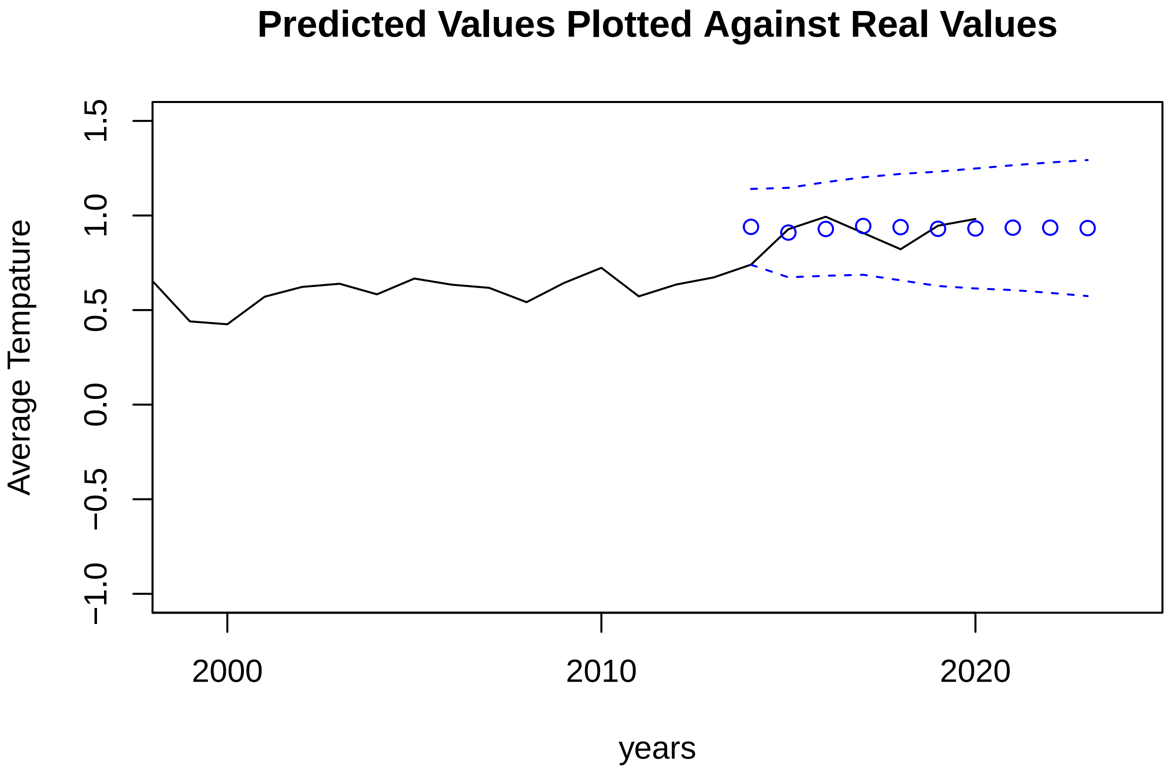
```
##
##  Shapiro-Wilk normality test
##
## data:  fit310.res
## W = 0.98813, p-value = 0.5103
```

The Shapiro-Wilk test will help us to test for normality among the residuals. Since it tests for normality, a p-value less than 0.05 means we reject the null hypothesis that these residuals are normally distributed. Luckily for us, we got a value of 0.5103 which means the data is normally distributed.

```
##
##  Box-Ljung test
##
## data:  fit310.res
## X-squared = 10.784, df = 7, p-value = 0.1483
```

The Box-Ljung test will allow us to see if the samples are independent from one another. The p-value we get here is 0.1483 which means we once again fail to reject the null hypothesis. The residuals seem to be independent and normally distributed so we can move on to forecasting.

Forecasting



Taking a look of our forecasted values on the actual data, we can see that the predicted values follow the true observations somewhat accurately. The first forecasted point is too high and the following seems to be much more linear compared to the actual points. It seems at the end the forecasted values correct themselves, but I am unsure at how precise they actually will be. Overall, all of the observations fall within the confidence interval which suggest that the model we have chosen does fit the data.

Conclusion

Upon completion of this project I found that the ARIMA(3,1,0) model was the most appropriate with the model being: $X_t = -0.3746X_{t-1} - 0.4027X_{t-2} - 0.2529X_{t-3} + Z_t$

The goal of this project was to identify and forecast the next few years temperature changes using data from the past 100 years. To do this I first made the data stationary and then looked at the ACF and PACF to help identify some models to fit the data. After doing diagnostic checking and multiple tests I settled on using the ARIMA(3,1,0). Although this model fit the data the best compared to the other models, I am not confident in the predictions it made. The forecasted points only followed the real data points slightly and had big differences when the data made a sharp upward or downward move. All the data points fell within the confidence intervals which leads me to believe this model is not the worst, however I wish the forecasted values resembled the true values more. Overall, this model does a satisfactory job of forecasting data points, but I believe there may be a model I did not consider that could forecast even more effectively.

Refrences

https://www.ncdc.noaa.gov

Appendix

```
# Download the data set
data <- read.csv("https://www.ncdc.noaa.gov/cag/global/time-series/globe/land_ocean/all/1/1920-2020/data
```

```
# Create data set that only contains temperature values
tempatures <- ts(as.numeric(data$Global.Land.and.Ocean.Temperature.Anomalies[5:1216]))
# Create loop to get average temperature of each year
yearly <- 1:(length(tempatures)/12)
for (i in yearly){
  count <- (i-1)*12 + 1
  if (sum(tempatures[count:(count+11)] == 0)){
    yearly[i] <- 0
  }
  else{
  yearly[i] <- sum(tempatures[count:(count+11)])/12
  }
}
ts.yearly <- ts(yearly)
```

```r
# partition data into test set
len <- as.numeric(length(ts.yearly))
test.yearly <- ts.yearly[-c(len:(len-5))]
```

```r
#plotting the time series
par(mfrow= c(1,1))
plot.ts(ts.yearly,xaxt="n",xlab="year",ylab="Temperature Change",main="Yearly Average Temperature Chang
axis(1,at=seq.int(1,101,by=10),labels=seq(1920,2020,by=10))
```

```r
#Box-Cox Transformation
library(MASS)
postive.yearly <- test.yearly + abs(min(test.yearly)) + 1
BCtrans <- boxcox(postive.yearly~as.numeric(1:length(postive.yearly)))
```

```r
lambda <- BCtrans$x[which(BCtrans$y==max(BCtrans$y))]
```

```r
#Log data
log.yearly <- log(postive.yearly)
#plot 2
par(mfrow=c(2,1))
plot.ts(log.yearly,xaxt="n",xlab="year",ylab="Average Temperature",main="Logged Yearly Average Temperatu
axis(1,at=seq.int(1,101,by=10),labels=seq(1920,2020,by=10))
plot.ts(test.yearly,xaxt="n",xlab="year",ylab="Average Temperature",main="Yearly Average Temperature Cha
axis(1,at=seq.int(1,101,by=10),labels=seq(1920,2020,by=10))
```

```r
par(mfrow=c(1,1))
test.yearly.diff <- diff(test.yearly,lag=1)
plot.ts(test.yearly.diff,xaxt="n",xlab="year",ylab="Differenced Once Average Temperature",main="Differen
axis(1,at=seq.int(1,101,by=10),labels=seq(1920,2020,by=10))
```

```r
#variance check
#var(test.yearly.diff) < var(test.yearly)
```

```r
# Plot ACF and PACF
par(mfrow=c(1,2))
acf(test.yearly.diff,lag.max=length(test.yearly.diff),main="Differenced Time Series")
pacf(test.yearly.diff,lag.max=length(test.yearly.diff),main="Differenced Time Series")
```

```r
#ARIMA models
#arima(0,1,2)
fit012 <- arima(ts.yearly,order=c(0,1,2),method="ML")
c(fit012$coef[2] -1.96*.0913, fit012$coef[2] +1.96*.0913)
#arima(1,1,2)
fit112 <- arima(ts.yearly,order=c(1,1,2),method="ML")
c(fit112$coef[1] -1.96*.2985, fit112$coef[1] +1.96*.2985) #ar 1
c(fit112$coef[3] -1.96*.1463, fit112$coef[3] +1.96*.1463)
#arima(2,1,2)
fit212 <- arima(ts.yearly,order=c(2,1,2),method="ML")
c(fit212$coef[2] -1.96*.1988, fit212$coef[2] +1.96*.1988) #ar2
c(fit212$coef[4] -1.96*.2321, fit212$coef[4] +1.96*.2321)
#arima(3,1,0)
```

```
fit310 <- arima(ts.yearly,order=c(3,1,0),method="ML")
c(fit310$coef[3] -1.96*.1, fit310$coef[3] +1.96*.1)
#arima(3,1,1)
fit311 <- arima(ts.yearly,order=c(3,1,1),method="ML")
c(fit311$coef[3] -1.96*.1181, fit311$coef[3] +1.96*.1181) #ar 3
c(fit311$coef[4] -1.96*.3248, fit311$coef[4] +1.96*.3248)
#arima(3,1,2)
fit312 <- arima(ts.yearly,order=c(3,1,2),method="ML")
c(fit312$coef[3] -1.96*.1234, fit312$coef[3] +1.96*.1234)
c(fit312$coef[5] -1.96*.3603, fit312$coef[5] +1.96*.3603)
```

```
#table for p and q terms
arima.table <- matrix(ncol=6,nrow=6)
colnames(arima.table) <-c("MA(q)","Lower CI","Upper CI","AR(p)","Lower CI","Upper CI")
rownames(arima.table) <- c("ARIMA(0,1,2)","ARIMA(1,1,2)","ARIMA(2,1,2)","ARIMA(3,1,0)","ARIMA(3,1,1)","A
arima.table[1,1:6] <- c(-.1937,round(fit012$coef[2] -1.96*.0856,4),round(fit012$coef[2] +1.96*.0856,4),
arima.table[2,1:6] <- c(-.2280,round(fit112$coef[3] -1.96*.1286,4),round(fit112$coef[3] +1.96*.1286,4),-
arima.table[3,1:6] <- c(.2419,round(fit212$coef[4] -1.96*.2263,4),round(fit212$coef[4] +1.96*.2263,4),-
arima.table[4,1:6] <- c("NA","NA","NA",-.2529,round(fit310$coef[3] -1.96*.0962,4),round(fit310$coef[3] +
arima.table[5,1:6] <- c(.405,round(fit311$coef[4] -1.96*.296,4),round(fit311$coef[4] +1.96*.296,4),-.379
arima.table[6,1:6] <- c(.2292,round(fit312$coef[5] -1.96*.4007,4),round(fit312$coef[5] +1.96*.4007,4),-
last <- as.table(arima.table)
```

```
# Used to make table look nice
library(knitr)
kable(last)
```

```
# Plot residuals
fit310.res <- residuals(fit310)
plot.ts(fit310.res,xaxt="n",xlab="year",ylab="ARIMA(3,1,0) Residuals",main="Residuals of ARIMA(3,1,0)",
axis(1,at=seq.int(1,101,by=10),labels=seq(1920,2020,by=10))
```

```
# Plot ACF and PACF of residuals
par(mfrow=c(1,2))
acf(fit310.res,lag.max = length(fit310.res),main="Fitted Residuals")
pacf(fit310.res,lag.max = length(fit310.res),main="Fitted Residuals")
```

```
# Plot histogram
hist(fit310.res,main="Histogram of Residuals",col="red",xlab="")
```

```
# Plot qqnorm
qqnorm(fit310.res)
qqline(fit310.res,col="blue")
```

```
# Compute the shapiro-wilk test
shapiro.test(fit310.res)
```

```
# compute the Ljung test
Box.test(fit310.res,lag=11,type="Ljung",fitdf=4)
```

```
# Forecast predicted values onto original data
fit310.prediction <- predict(fit310,n.ahead = 10)
i <- as.numeric(length(test.yearly))
# Plot the forcasted against the original
plot(ts.yearly,main="Predicted Values Plotted Against Real Values",ylab="Average Tempature",xlab="years
axis(1,at=seq.int(1,101,by=10),labels=seq(1920,2020,by=10))
points(x = i:(i+9),fit310.prediction$pred,col="blue")
lines(i:(i+9),fit310.prediction$pred+1.96*fit310.prediction$se,lty=2,col="blue")
lines(i:(i+9),fit310.prediction$pred-1.96*fit310.prediction$se,lty=2,col="blue")
```