Forecasting Anomalies in Global Temperature

Christopher Nguyen | Darcy Wang | Kyle Peters

3/9/2018

Abstract

We based our project on annual global temperature anomalies from 1880 to 2016. We designed this project to answer two principle questions. First, can we come up with a time series that can accurately model the data? Second, can we forecast reliable estimates for future global temperature anomalies for the next ten years, with 2017 being a good value for comparison, given it is known already.

Initiallly, the data has obvious trend and seems seasonal, nonconstant variance. BoxCox method fixed the nonconstant variance, ACF and PACF plots showed this time series only has trend but nonseasonal. Difference one time, then we got stationary time series. ACF, PACF plots and AIC values determined several models to compare. We chose the one passed diagnostic check and also with lowest AIC value. The final model we chose an ARIMA(3,1) model. Through the forecasting plot, we can see predict line is almost horizontal. But the confident intervals shows both up and down. In this case, we cannot conclude that the global temperature is increasing.

Introduction

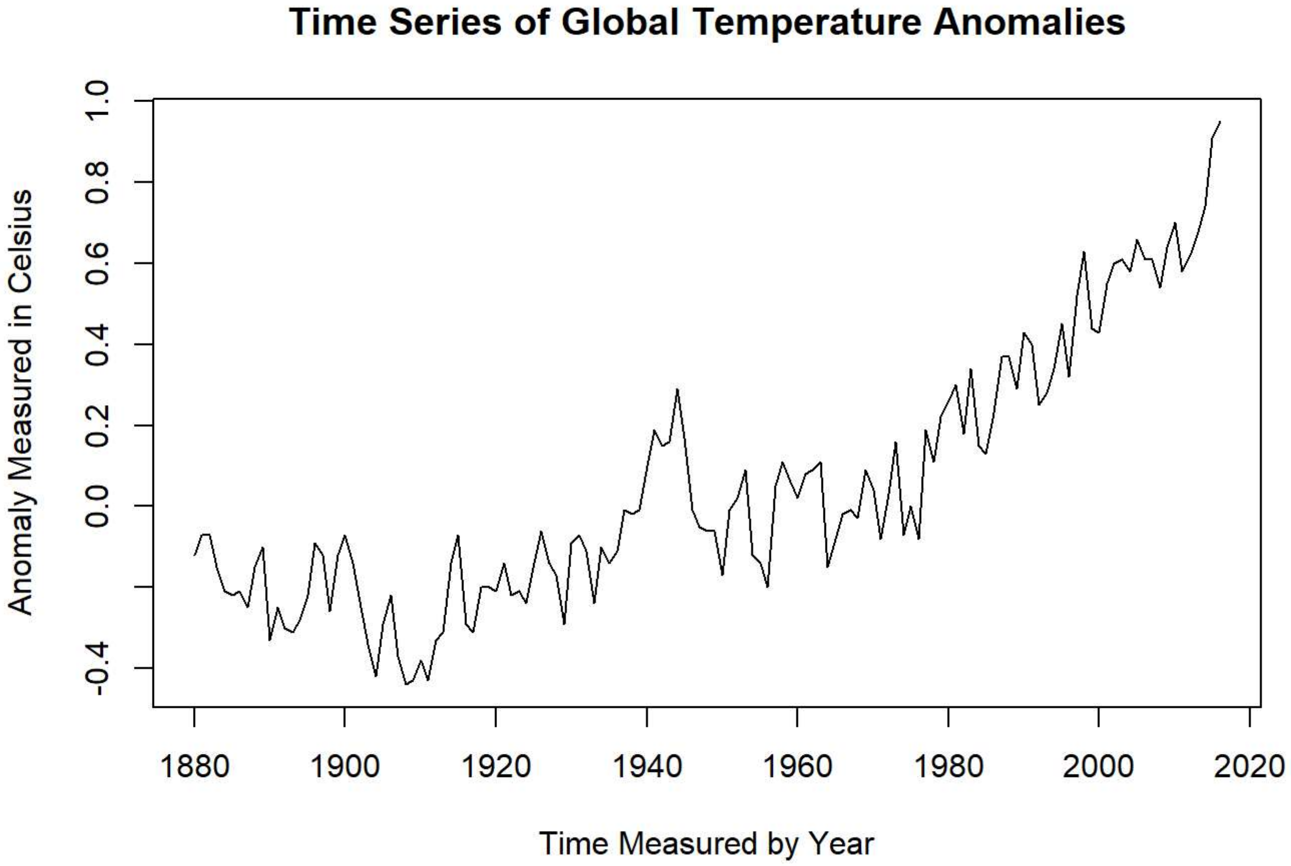
Golbal warming is a well-known issue. We want to use time series analysis to forecast future changes in climate given data on past global temperature anomalies. We believe that this data set is important because climate change, and by extension the future of our planet, has become a controversial and politicised topic. We hope to find insight on such a divisive topic. We believe that it is We plan to use the auto correlation and partial autocorrelation function plots to estimate an ARIMA model that will be able to forecast future data. Our model, ARIMA(1,3) model,\[X_{t} + 0.9X_{t-1} = Z_{t} -0.65Z_{t-1} - 0.47Z_{t-2} - 0.30Z_{t-3} \]. Our data set is also authentic and reliable, coming from a government agency. We obtained our data from the US Department of Commerce, under the NOAA, National Centers for Environmental Information through the NCDC. We used R studio to generate our results.

Packages Used

```
library(readr)
library(qpcR)
library(MASS)
```

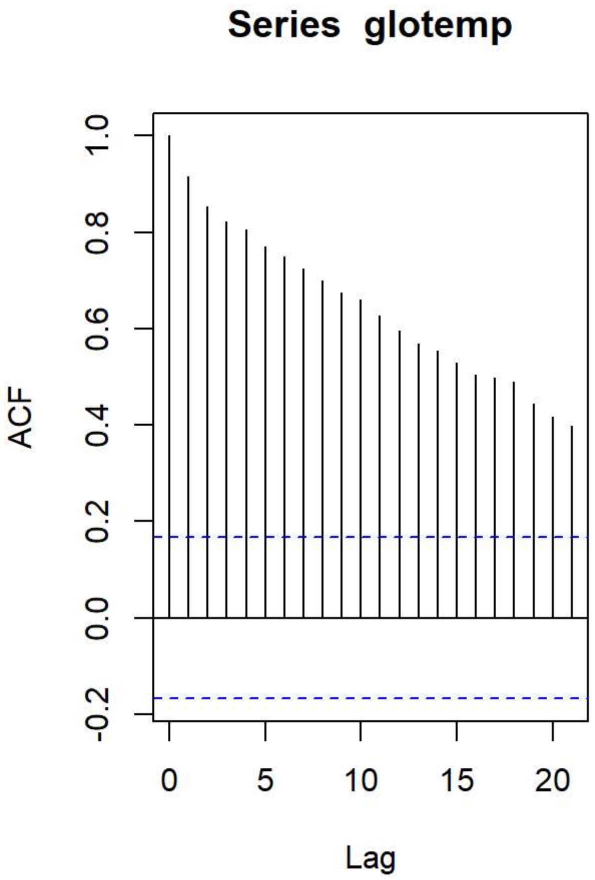
Initial Analysis

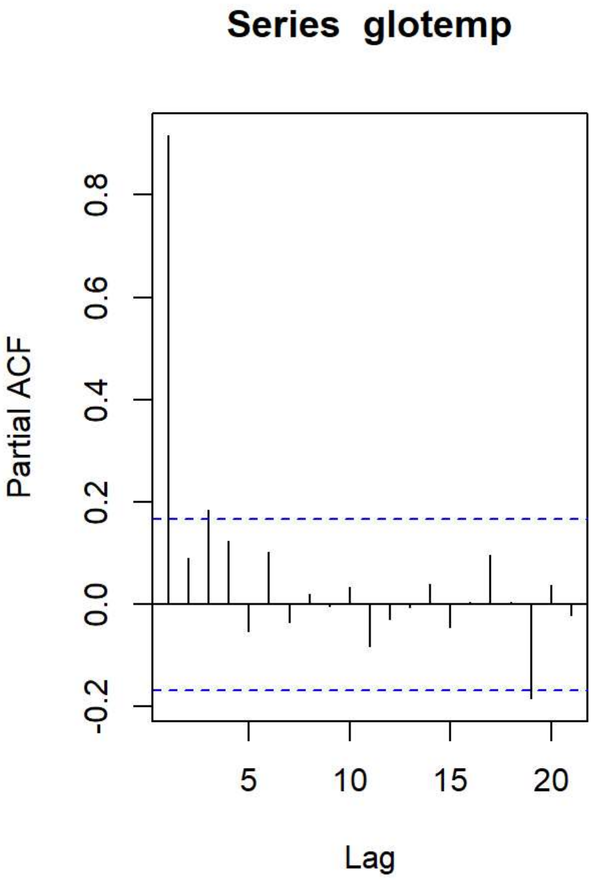
With the necessary packages loaded, we begin by importing our dataset and initializing it as a time series object with specified argument parameters. We then plot the time series and make exploratory observations, checking for indications of trend, seasonality, stationarity, and constant variance.



Preliminary observations of the plot indicate that there doesn't seem to be a seasonal component, but it could still be possible. It does strongly shows an increasing trend component as time progresses. Furthermore, a strong trend suggests nonstationary behavior. We also notice that variance could potentially be nonconstant.

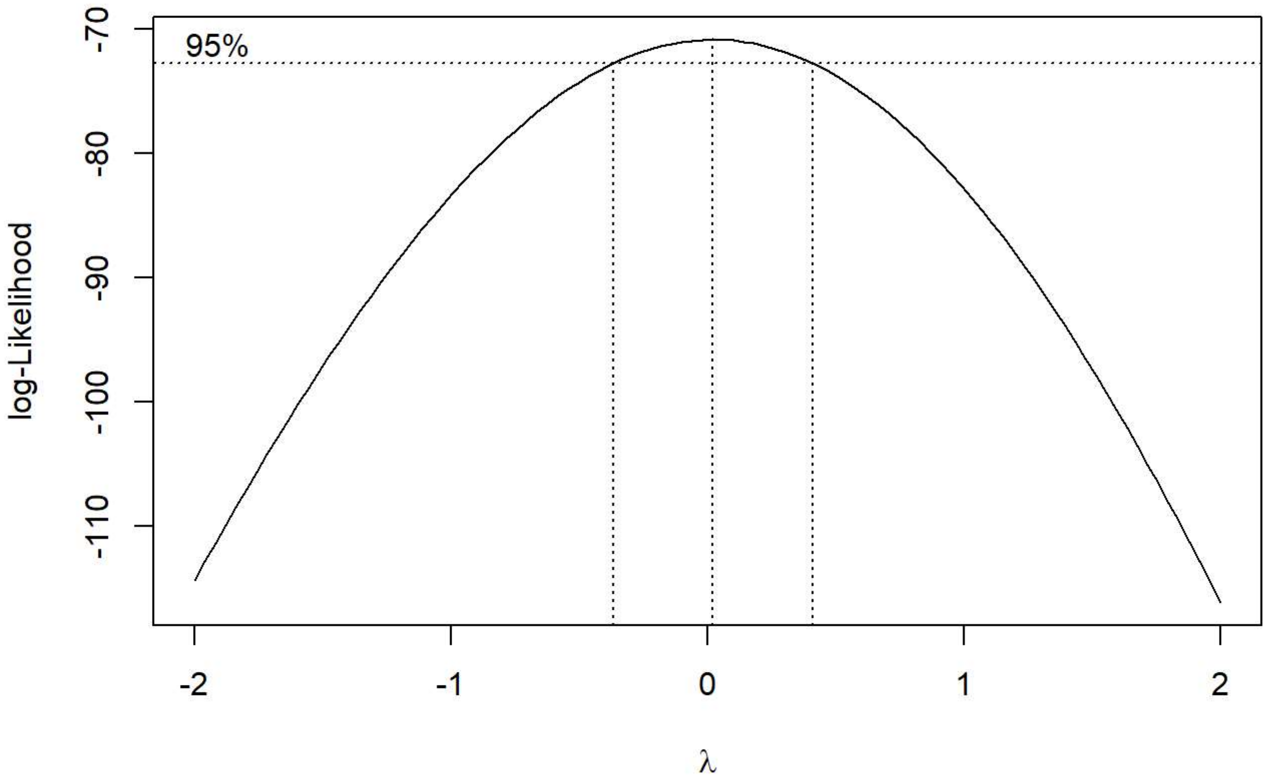
Examining Seasonality and Stationarity via ACF and PACF

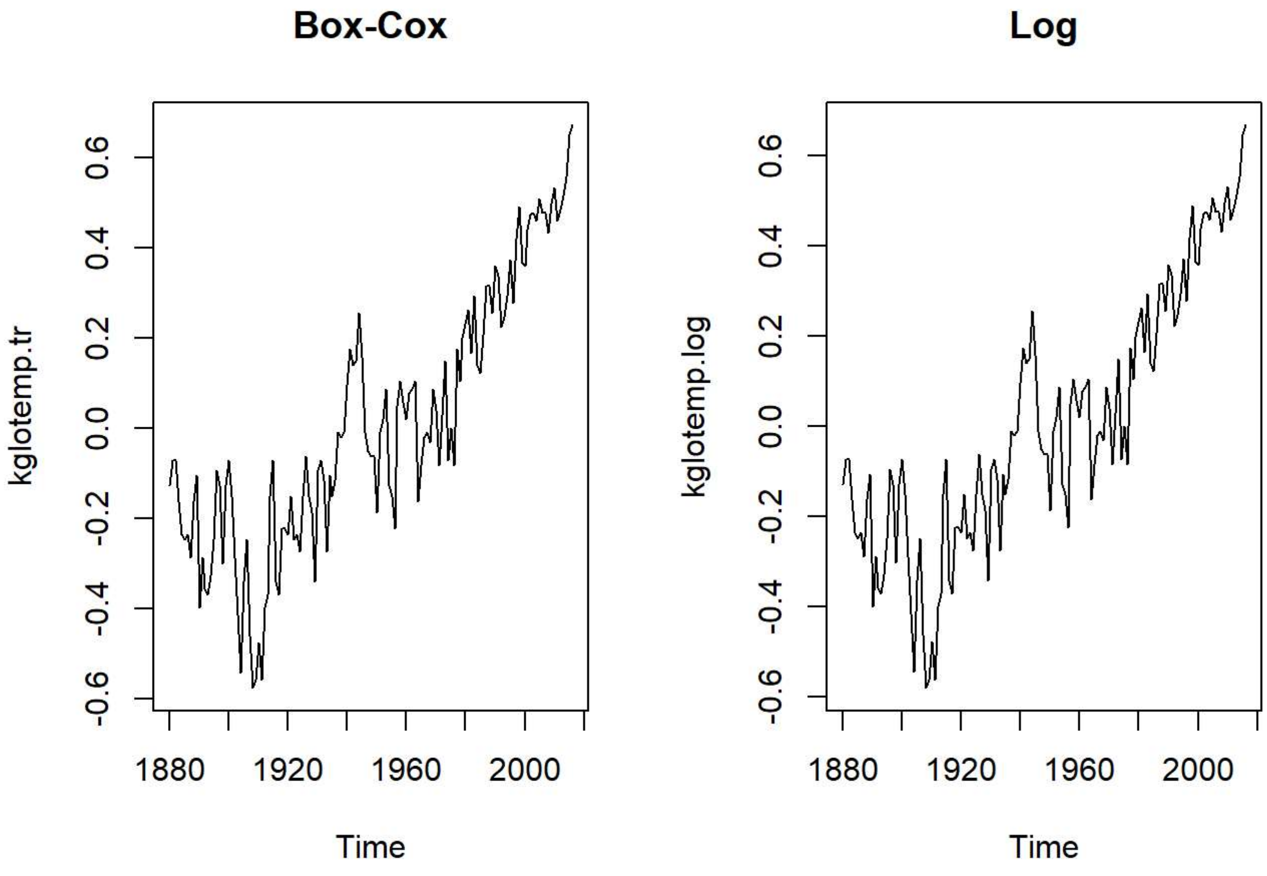




The ACF plot generated above shows large, persistent stabs. The PACF, on the other hand, doesn’t show significant lags, with peculiarities only at lags 2 and 18. This time series is not stationary.

Stabilizing the Variance through Transformation

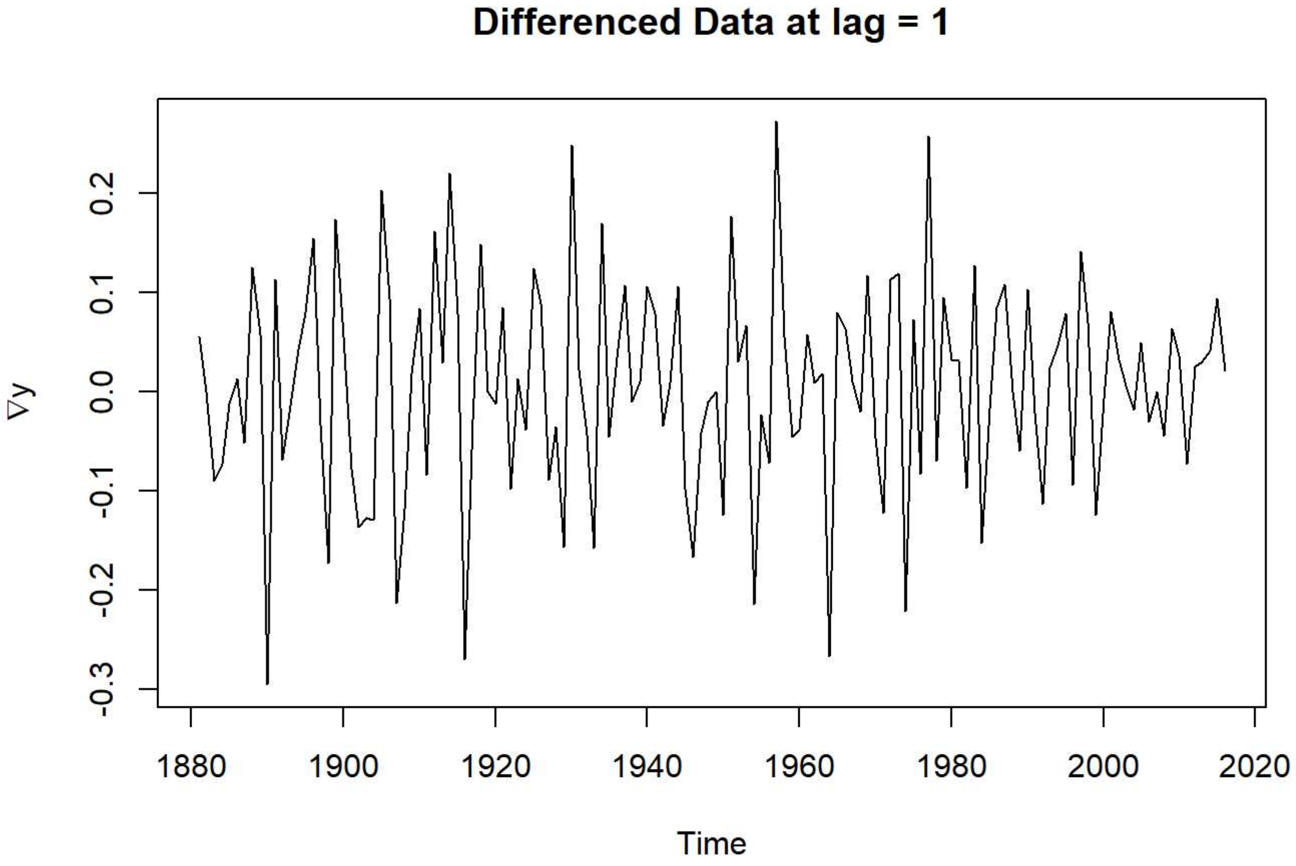

95%
log-Likelihood
-70
-80
-90
-100
-110
-2
-1
0
1
2
λ

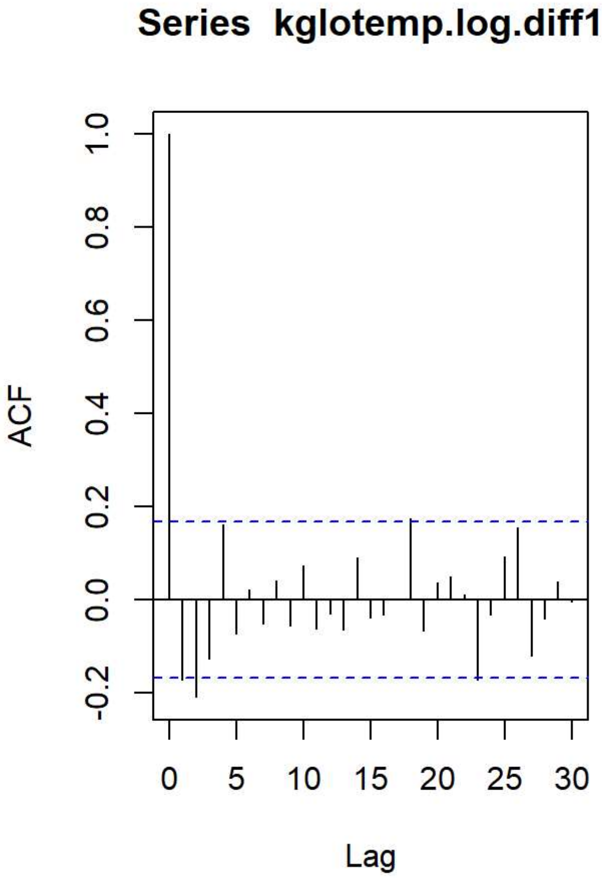

Box-Cox
Log
kglotemp.tr
kglotemp.log
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
1880
1920
1960
2000
Time
Time

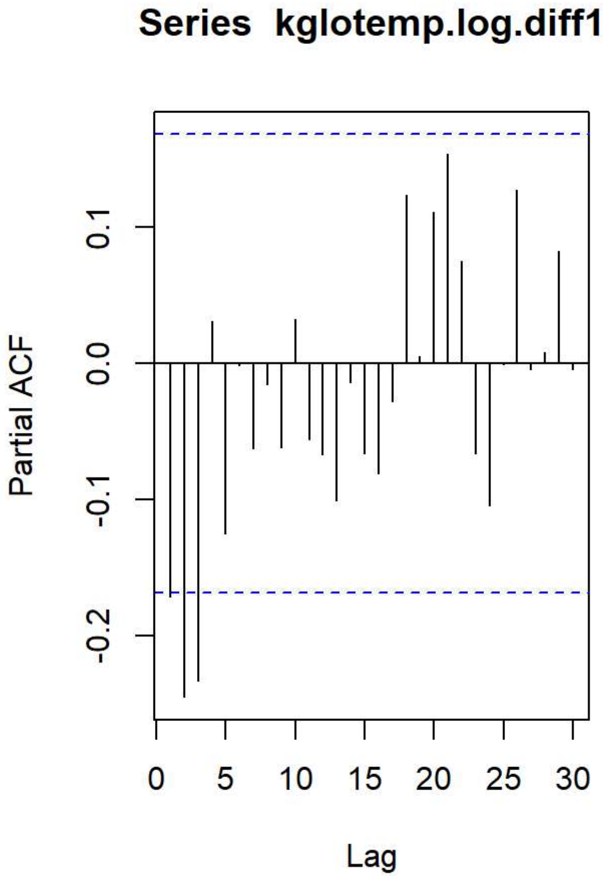
```
## [1] 0.08233008
```

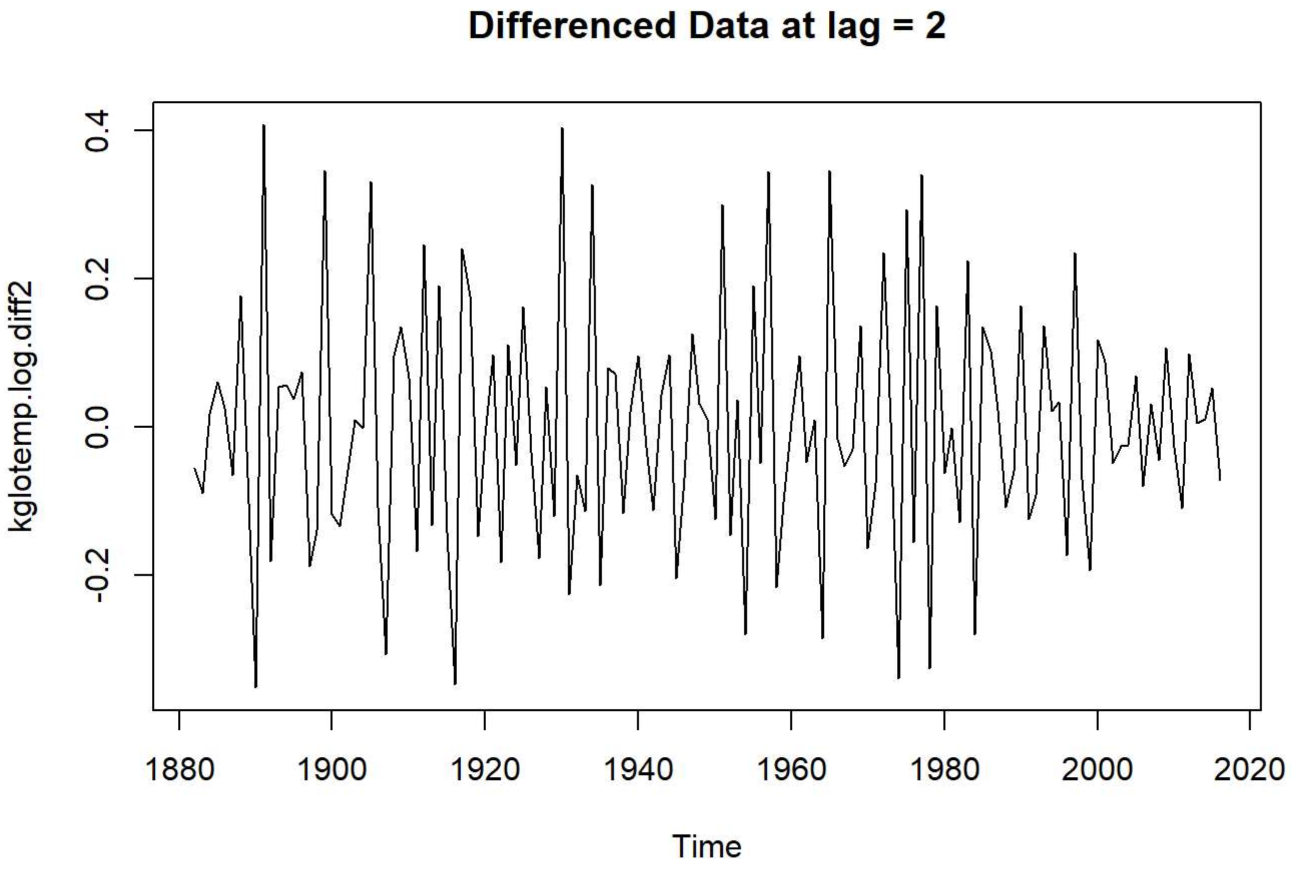
The original data contains negative values so we added 1 to make all data has positive values. Now BoxCox is plausible. The optimal value is 0.02, very close to 0, also 0 in 95% confident intervals, and time series plots after Box-Cox and log transformations looks almost the same, therefore, we move forward with the log transformation because reverting the log transformation for forecasting after will prove easier. With variance constant, we move forward with differencing to remove the trend.

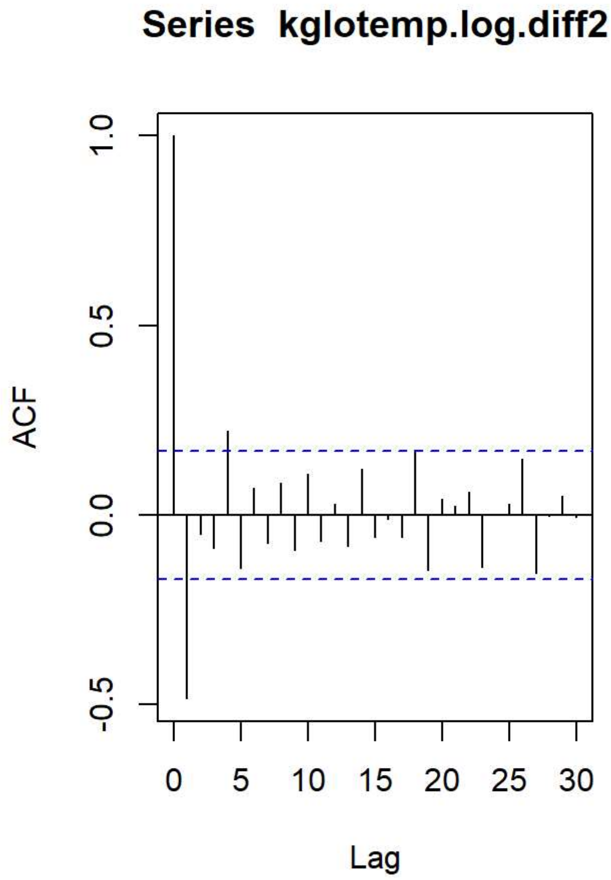
Trend Removal by Differencing

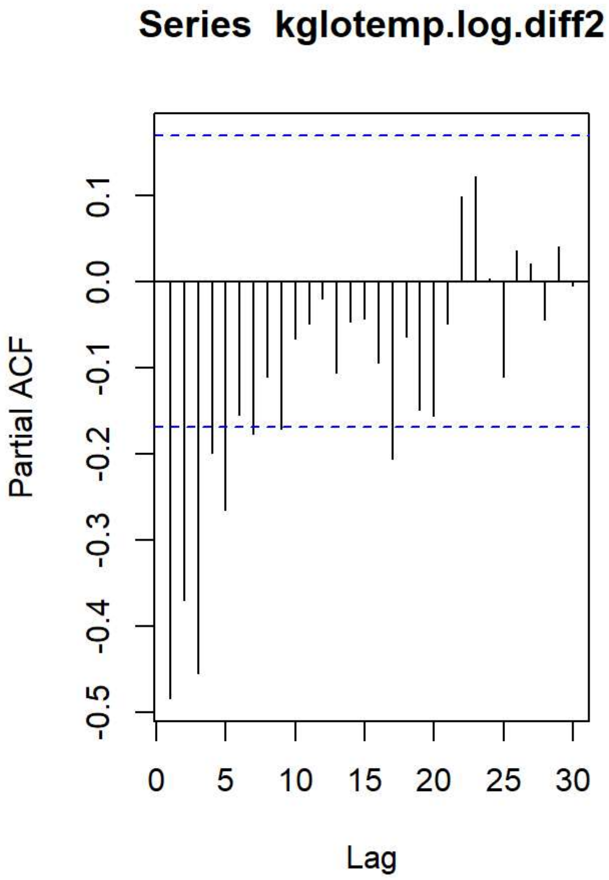

Differenced Data at lag = 1
∇y
0.2
0.1
0.0
-0.1
-0.2
-0.3
1880
1900
1920
1940
1960
1980
2000
2020
Time

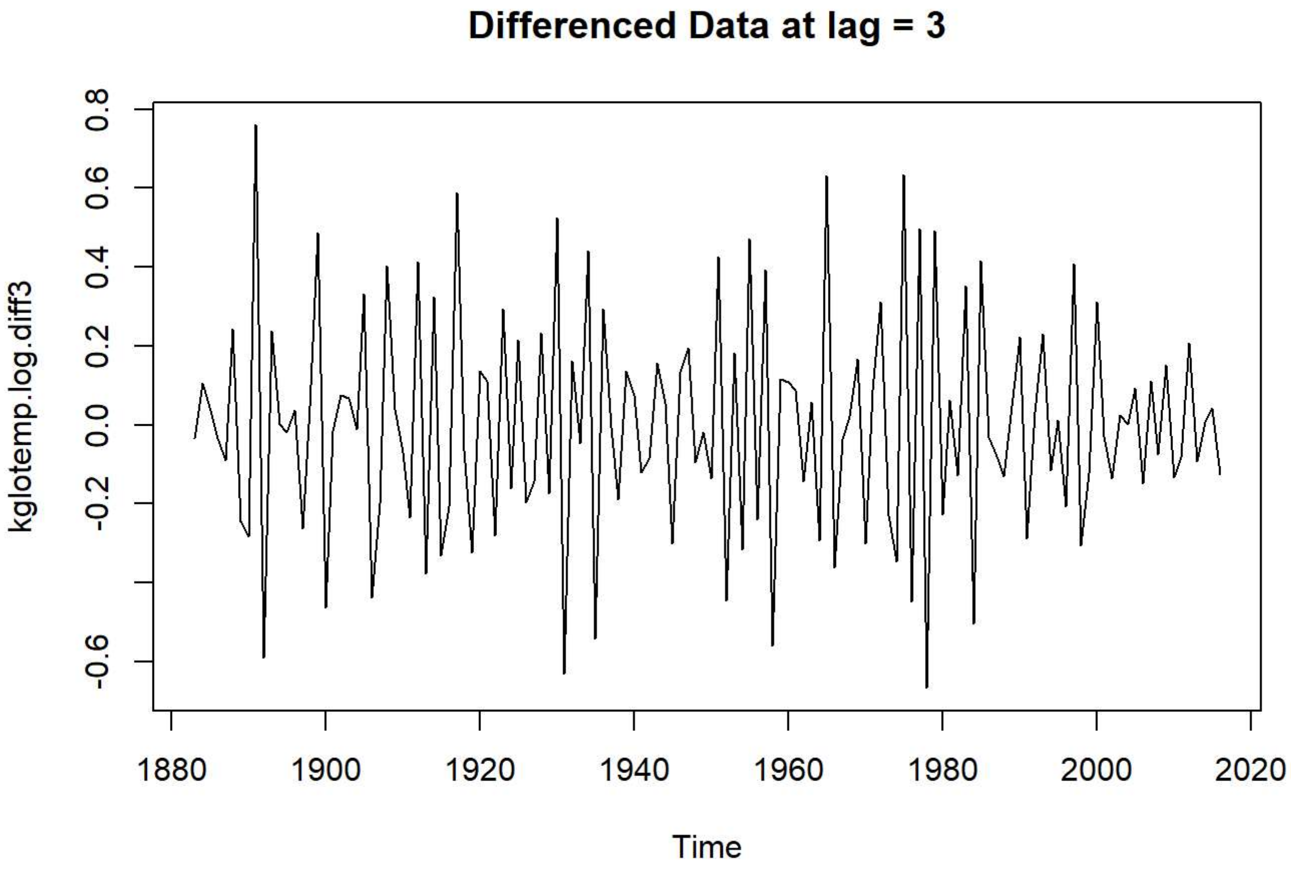

Series kglotemp.log.diff1
ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20
25
30
Lag

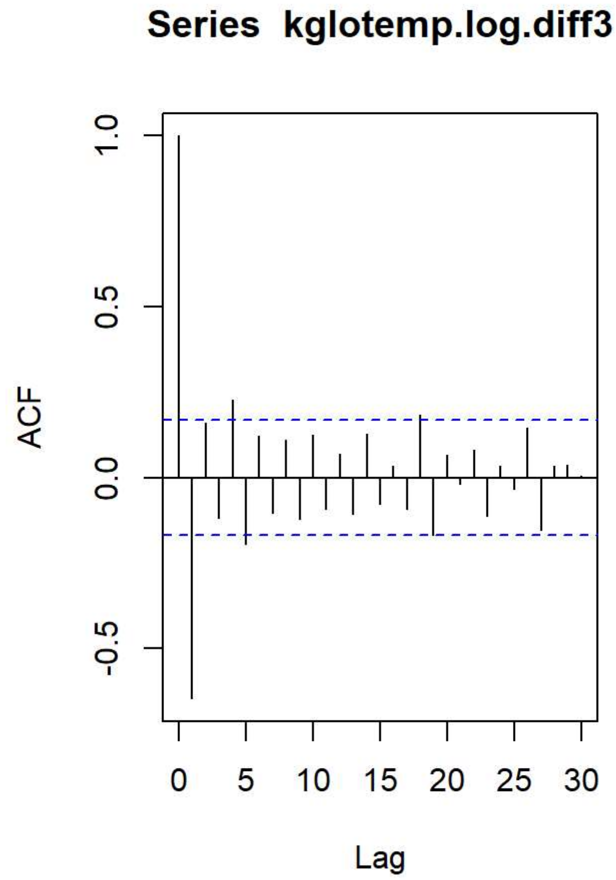

Series kglotemp.log.diff1
Partial ACF
0.1
0.0
-0.1
-0.2
0
5
10
15
20
25
30
Lag

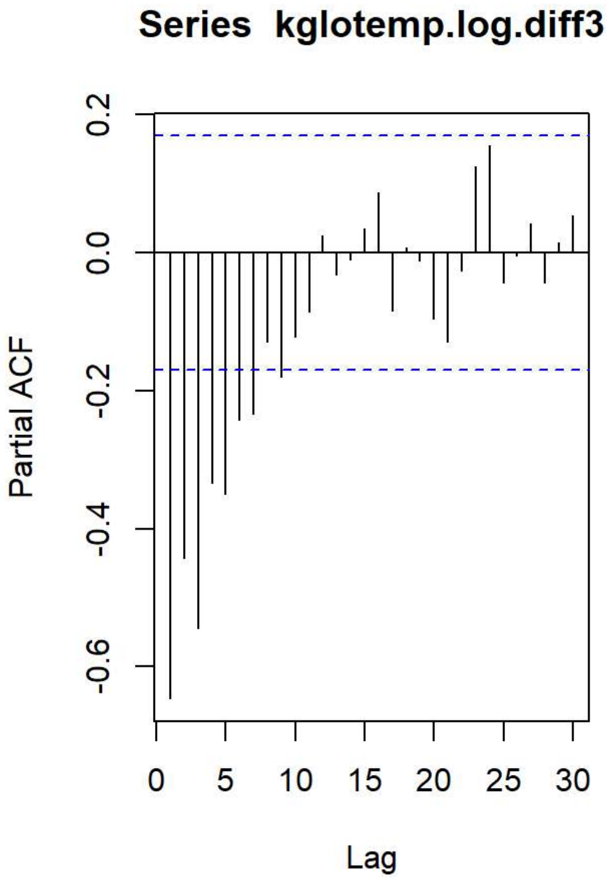

Differenced Data at lag = 2
kglotemp.log.diff2
0.4
0.2
0.0
-0.2
1880
1900
1920
1940
1960
1980
2000
2020
Time


Series kglotemp.log.diff2


Series kglotemp.log.diff2


Differenced Data at lag = 3





```
##   var.kglotemp.log. var.kglotemp.log.diff1. var.kglotemp.log.diff2.
## 1        0.08233008              0.01133661              0.02673375
##   var.kglotemp.log.diff3.
## 1              0.07988196
```

For convenience, we constructed a table with the all the variances for analysis. A lag 1 difference significantly reduced our variance from 0.08233008 to 0.01133661. A second difference slightly increased the variance to 0.02673375.because we also have significant sample ACF at lags 2, 4, 18, 23, so the time series could have a seasonal. From the interval between these number, the season could be 2, 4, 6. Then we tried 2, 4, 6 lag differnce, but all of them give the PACF more stabs, so this time series doesn't have a seasonality. Hence, we stay with lag 1 difference. We could consider ACF as tails off and PACF as cut off at 3, shows behavior of AR3.

Determining p and q for Model Selection

```
## 
## Call:
## ar(x = kglotemp.log.diff1, method = "yule-walker")
## 
## Coefficients:
##       1        2        3  
## -0.2703  -0.2946  -0.2333  
## 
## Order selected 3  sigma^2 estimated as  0.01
```

```
##
## Call:
## ar(x = kglotemp.log.diff1, method = "mle")
##
## Coefficients:
##       1        2        3  
## -0.2700  -0.2943  -0.2325  
##
## Order selected 3  sigma^2 estimated as  0.009691
```

Yule-Walker method and MLE method both suggest AR3 model. It’s the same as we see from ACF and PACF.

Picking the Best ARIMA Model

However, we also can consider they are both tails off. ACF and PACF of ARIMA(p,d,q) models should both be tails off at the p or q (determined by which one is bigger). In this case, this time series could be ARIMA model contains parameters respectively less than 3. That’s the reason why we use AIC to pick up ARIMA model.

```
# Calculate AICc for ARMA models with p and q running from 0 to 3
aiccs <- matrix(NA, nr = 4, nc = 4)
dimnames(aiccs) = list(p=0:3, q=0:3)
for(p in 0:3) {
  for(q in 0:3) {
    aiccs[p+1,q+1] = AICc(arima(kglotemp.log.diff1, order = c(p,0,q), method="ML"))
    }
}
aiccs
```

```
##    q
## p           0         1         2         3
##   0 -220.2643 -227.0033 -234.1256 -234.0302
##   1 -222.2285 -234.1942 -233.1589 -234.9711
##   2 -228.5012 -233.3171 -231.1640 -232.9706
##   3 -233.9972 -232.6166 -232.7377 -230.8898
```

```
(aiccs==min(aiccs))
```

```
##    q
## p       0     1     2     3
##   0 FALSE FALSE FALSE FALSE
##   1 FALSE FALSE FALSE  TRUE
##   2 FALSE FALSE FALSE FALSE
##   3 FALSE FALSE FALSE FALSE
```

Potential models to be considered include the smallest AIC value model, ARIMA(1,3), and the second smallest, ARIMA(1,1).

```
##
## Call:
## arima(x = kglotemp.log, order = c(3, 1, 0), method = c("ML"))
##
## Coefficients:
##           ar1      ar2      ar3
##       -0.2627  -0.2877  -0.2255
## s.e.   0.0834   0.0826   0.0831
##
## sigma^2 estimated as 0.009788:  log likelihood = 121.48,  aic = -234.96
```

```
##
## Call:
## arima(x = kglotemp.log, order = c(1, 1, 1), method = c("ML"))
##
## Coefficients:
##          ar1      ma1
##       0.4587  -0.7707
## s.e.  0.1400   0.0975
##
## sigma^2 estimated as 0.01004:  log likelihood = 119.75,  aic = -233.5
```

```
##
## Call:
## arima(x = kglotemp.log, order = c(1, 1, 3), method = c("ML"))
##
## Coefficients:
##           ar1     ma1      ma2      ma3
##       -0.9006  0.6478  -0.4700  -0.3014
## s.e.   0.2247  0.2224   0.1238   0.0813
##
## sigma^2 estimated as 0.009623:  log likelihood = 122.58,  aic = -235.16
```

In summary, the models to be compared:

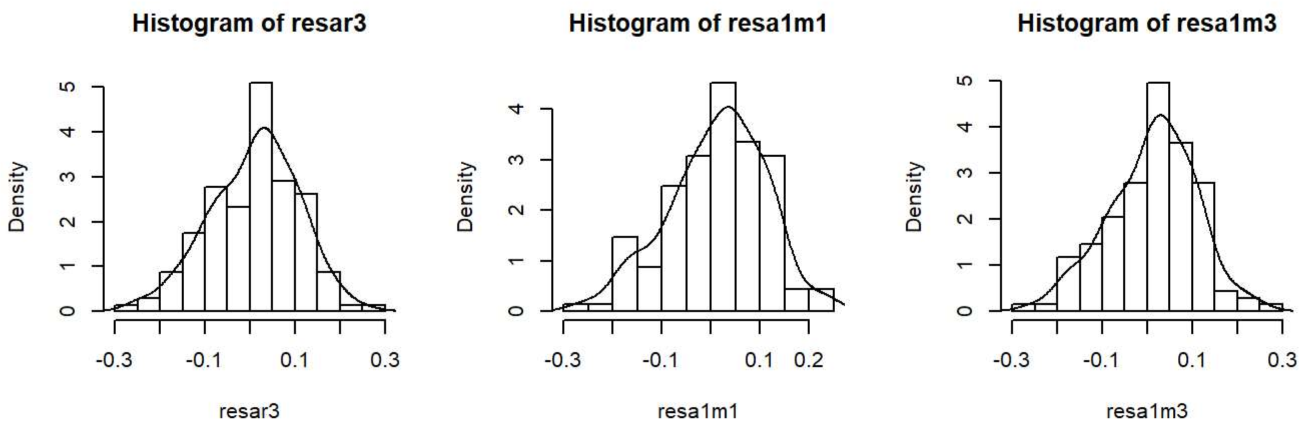
AR(3):\[X_{t} + 0.27 X_{t-1 }+ 0.29X_{t-2} + 0.23 X_{t-3} = Zt \] ARIMA(1,3):\[X_{t} + 0.9X_{t-1} = Z_{t} -0.65Z_{t-1} - 0.47Z_{t-2} - 0.30Z_{t-3} \] ARIMA(1,1):\[X_{t} - 0.46 X_{t-1} = Z_{t} - 0.78Z_{t-1} \]

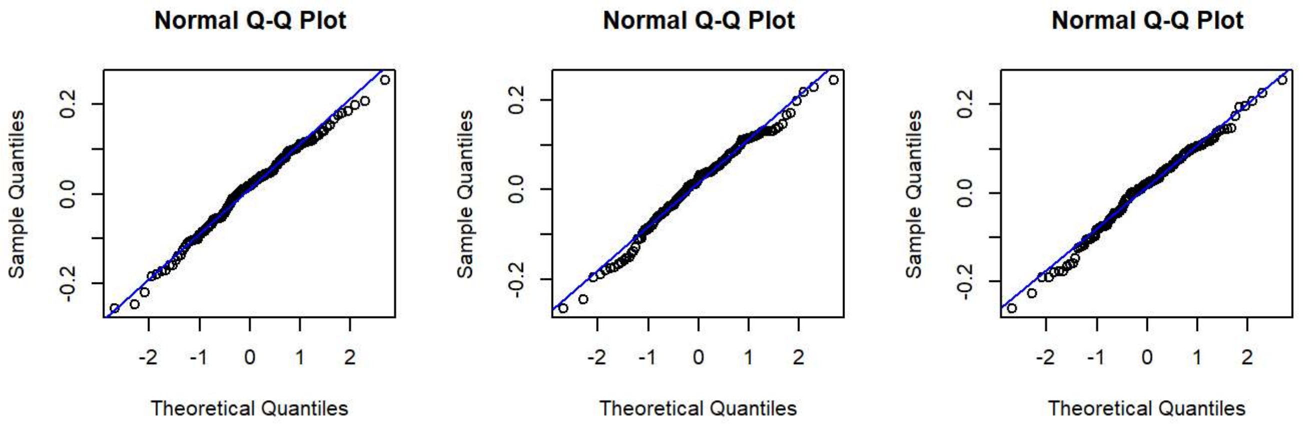
Diagnostic Check: Residual Test for Normality

```
##
##  Shapiro-Wilk normality test
##
## data:  resar3
## W = 0.99101, p-value = 0.5294
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resa1m1
## W = 0.98561, p-value = 0.1615
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resa1m3
## W = 0.98958, p-value = 0.3973
```





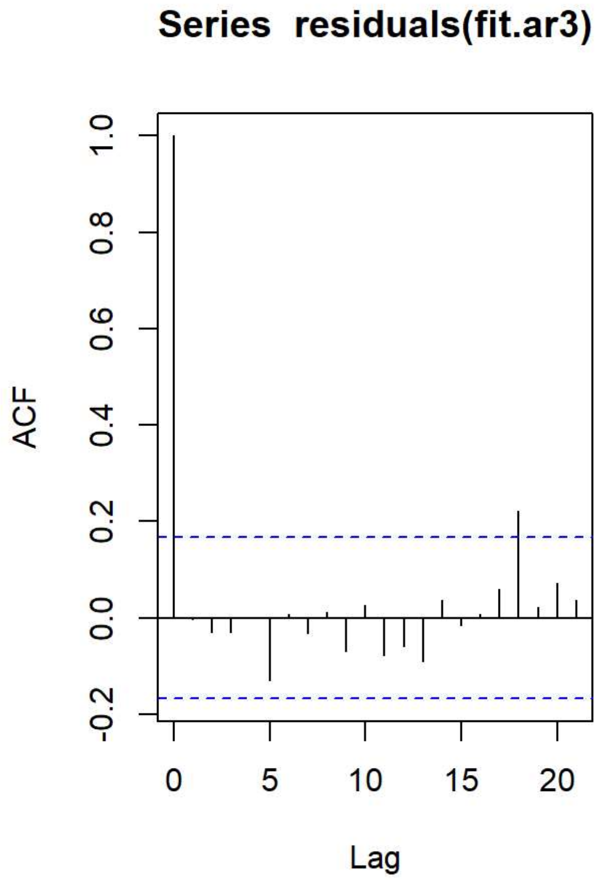
The histogram of ARIMA(1,1) is skewed, but they all passed the Shapiro-Wilk test.

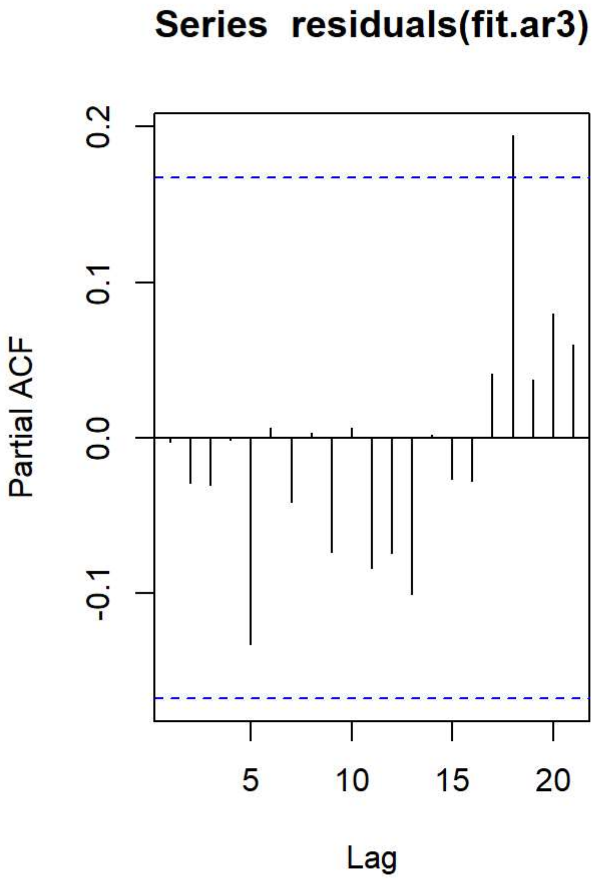
Diagnostic Check: Correlation Indepedent

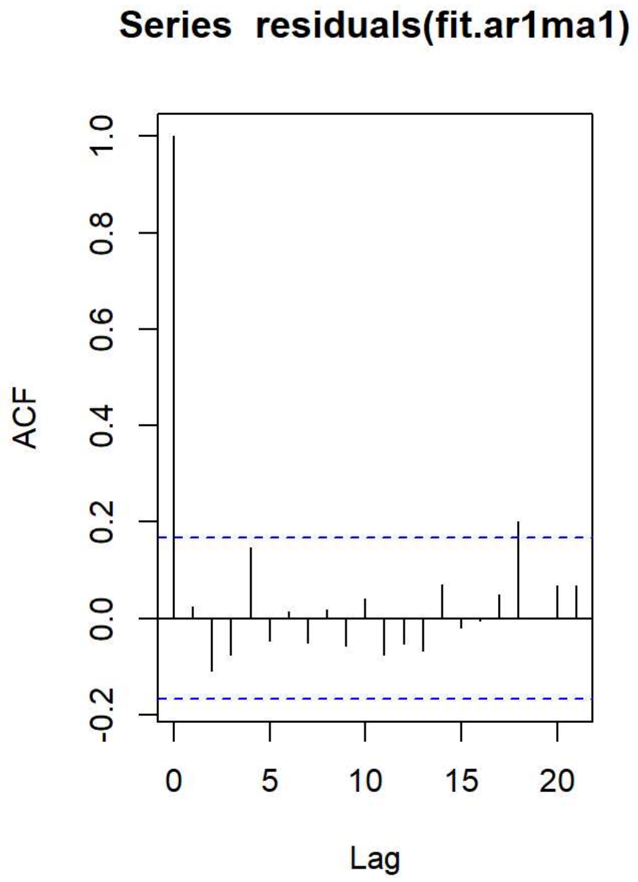
```
##
##  Box-Ljung test
##
## data:  residuals(fit.ar3)
## X-squared = 5.1354, df = 9, p-value = 0.8223
```

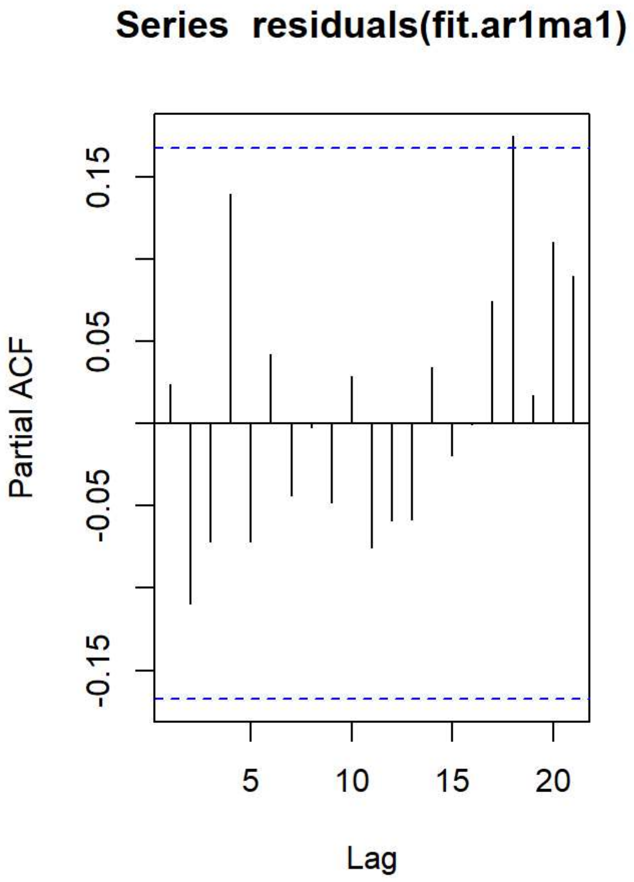
```
## 
##  Box-Ljung test
## 
## data:  residuals(fit.ar1ma1)
## X-squared = 8.3917, df = 10, p-value = 0.5906
```

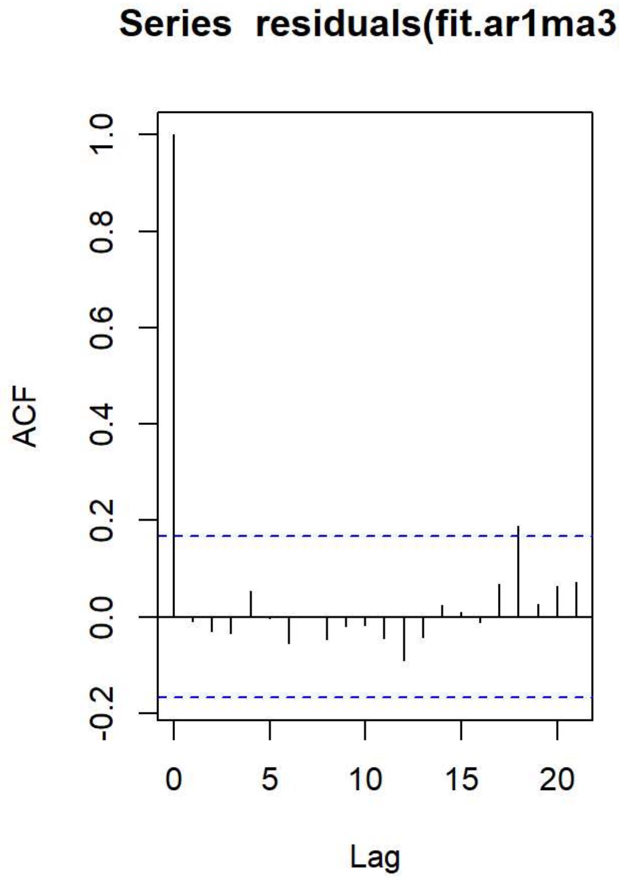
```
## 
##  Box-Ljung test
## 
## data:  residuals(fit.ar1ma3)
## X-squared = 3.1319, df = 8, p-value = 0.9258
```

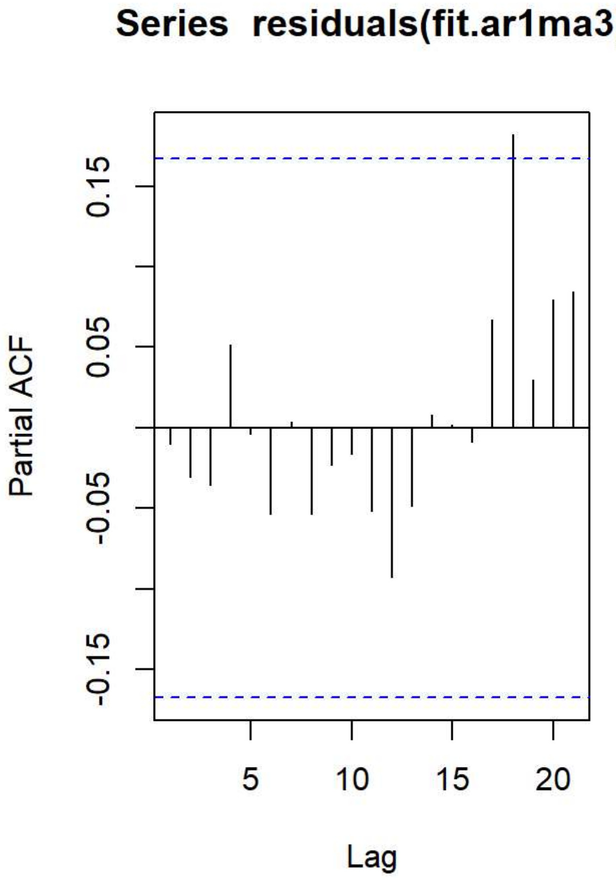

Series residuals(fit.ar3)
ACF
Lag
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20


Series residuals(fit.ar3)
Partial ACF
Lag
0.2
0.1
0.0
-0.1
5
10
15
20


Series residuals(fit.ar1ma1)
ACF
Lag
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20


Series residuals(fit.ar1ma1)
Partial ACF
Lag
0.15
0.05
-0.05
-0.15
5
10
15
20





We used Ljung-Box test to see if correlation is indepedent. They all pasted the teat. They all have few stabs at the same place, but we consider they all fit well enough, because an unexplained increase of temperature is in the data set. The extremity is between 1940 and 1945, the years World War 2 occurred. The models' abnormal stabs are due to the war's strain on the climate.

Diagnostic Check: Constant Variance

After log transformation, variances are constant.

We choose ARIMA(1,3) as final model. Because it has the lowest AIC values, and most possibly passed all diagnostic checks.

Forecast on Original Data

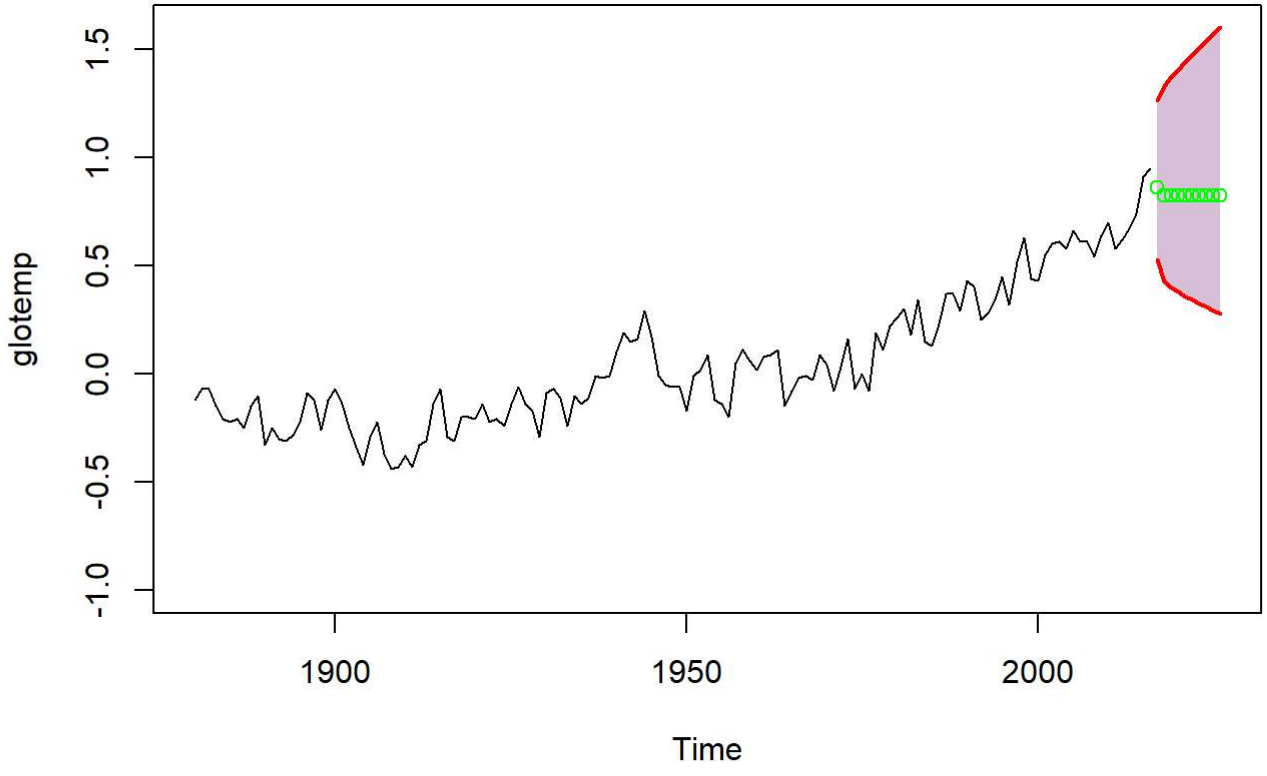
```
pred.tr <- predict(fit.ar1ma3, n.ahead=10)
U.tr= pred.tr$pred+ 2*pred.tr$se
D.tr= pred.tr$pred- 2*pred.tr$se
pred.orig = exp(pred.tr$pred) - 1
pre2017= print(pred.orig)[1]
```

```
## Time Series:
## Start = 2017
## End = 2026
## Frequency = 1
##  [1] 0.8597316 0.8213958 0.8209489 0.8213514 0.8209889 0.8213154 0.8210213
##  [8] 0.8212862 0.8210476 0.8212625
```

```
print(pre2017)
```

```
## [1] 0.8597316
```

```
U.orig = exp(U.tr) - 1
D.orig = exp(D.tr) - 1
ts.plot(glotemp, xlim=c(1880, 2026),ylim=c(-1,max(U.orig)))
polygon(c(2017:2026,rev(2017:2026)),c(D.orig,rev(U.orig)),col="thistle"
,border=NA)
points(2017:2026,pred.orig, col = "green", pch = 1)
lines(U.orig, col = "red", lwd=2)
lines(D.orig, col = "red", lwd=2)
```



To get the prediction of global temperature, we need to forecast on original data. We transformed the data to fit models, now we need to transform back. We added 1 then log transformed it. We got prediction on transformed data first, then we applied exponential transform on it and minus 1.

Prediction of 2017 is 0.8597. We can obtain the 2017 average temperature of real world from the data resourse, which is 0.85. Our prediction was very precise.

Conclusion

At the beginning of our project, we set out with fulfilling two straight forward goals. We wanted to find a time series model that accureately reflects the data and forecast estimates for the next estimates for the next ten years. Absolutely, we accomplished them. Indeed, it was difficult work transforming and differencing the data, but with all the diagnostics checks verified, we have our model, ARIMA(1,1,3): \[X_{t} + 0.9X_{t-1} = Z_{t} -0.65Z_{t-1} - 0.47Z_{t-2} - 0.30Z_{t-3} \]. We forecasted an estimated value that was excpetionally close to the actual real life value. According to our forecasting, we can’t say global temperature is increasing. We woud like to acknowledge and thank Professor Bapat for teaching us this powerful tool, as well as our TAs, Zhipu Zhou and Patricia Ning for the guidance they provided throughout the course.

Reference

R Studio Statistical Software

NOAA National Centers for Environmental information, Climate at a Glance: Global Time Series, published February 2018, retrieved on March 11, 2018 from http://www.ncdc.noaa.gov/cag/ (http://www.ncdc.noaa.gov/cag/)

https://www.ncdc.noaa.gov/cag/global/time-series/globe/land_ocean/ytd/12/1880-2016 (https://www.ncdc.noaa.gov/cag/global/time-series/globe/land_ocean/ytd/12/1880-2016)

Appendix

```
#Load libraries
library(readr)
library(qpcR)
library(MASS)

#import and plot time series
X1880 <- read_csv("climatechange.csv")
glotemp.data = as.numeric(X1880$Value)
glotemp <- ts(data = glotemp.data, start = c(1880), frequency = 1)
plot.ts(glotemp, main = "Time Series of Global Temperature Anomalies",xlab = "Time Measured by Y
ear", ylab = "Anomaly Measured in Celsius")



#plot acf and pacf
op <- par(mfrow=c(1,2))
acf(glotemp)
pacf(glotemp)
par(op)

#transform data
ktemp.data <- glotemp.data + 1
kglotemp<- ts(data = ktemp.data, start = c(1880), frequency = 1)
bcTransform <- boxcox(kglotemp ~ as.numeric(1:length(kglotemp)))
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
print(lambda) ##almost 0
kglotemp.tr <- (1/lambda)*(kglotemp^lambda - 1) # Box-cox
kglotemp.log <- log(kglotemp) # log, since log also in CI
op <- par(mfrow = c(1,2))
ts.plot(kglotemp.tr,main = "Box-Cox")
ts.plot(kglotemp.log,main = "Log")
var(kglotemp.log)

#difference data
kglotemp.log <- log(kglotemp)
kglotemp.log.diff1 <- diff(kglotemp.log,1)
ts.plot(kglotemp.log.diff1,main = "Differenced Data at lag = 1",ylab=expression(paste(nabla,y)))
op <- par(mfrow=c(1,2))
acf(kglotemp.log.diff1, lag.max = 30)
pacf(kglotemp.log.diff1, lag.max = 30)
par(op)
##There is a significant sample ACF at lag 2, 4, 18, 23.
##PACF looks tail off.
var(kglotemp.log.diff1)

kglotemp.log.diff2 <- diff(kglotemp.log.diff1,1)
ts.plot(kglotemp.log.diff2,main = "Differenced Data at lag = 2")
op <- par(mfrow=c(1,2))
acf(kglotemp.log.diff2, lag.max = 30)
pacf(kglotemp.log.diff2, lag.max = 30)
par(op)
#var(kglotemp.log.diff2)
##Time Series plot looks not  Variance goes up. Pacf has more stabs.
```

```
kglotemp.log.diff3 <- diff(kglotemp.log.diff2,1)
ts.plot(kglotemp.log.diff3,main = "Differenced Data at lag = 3")
op <- par(mfrow=c(1,2))
acf(kglotemp.log.diff3, lag.max = 30)
pacf(kglotemp.log.diff3, lag.max = 30)
par(op)
#var(kglotemp.log.diff3)
## Variance goes up. Pacf becomes worse.
##Stay with difference 1.
##ACF tails off, PACF cuts off after lag 3, suggest AR model.
var.change<-data.frame(var(kglotemp.log),var(kglotemp.log.diff1),var(kglotemp.log.diff2),var(kgl
otemp.log.diff3))
print(var.change)


#calculate ar coeff.
fit.ar3 <- ar(kglotemp.log.diff1, method="yule-walker")
print(fit.ar3)


# Calculate AICc for ARMA models with p and q running from 0 to 3
aiccs <- matrix(NA, nr = 4, nc = 4)
dimnames(aiccs) = list(p=0:3, q=0:3)
for(p in 0:3) {
  for(q in 0:3) {
    aiccs[p+1,q+1] = AICc(arima(kglotemp.log.diff1, order = c(p,0,q), method="ML"))
    }
}
aiccs
(aiccs==min(aiccs))

#check for nor
fit.ar3<- arima(kglotemp.log, order = c(3,1,0), method = c("ML"))
print(fit.ar3)
resar3 = residuals(fit.ar3)
fit.ar1ma3 <- arima(kglotemp.log, order = c(1, 1, 3), method = c("ML"))
print(fit.ar1ma3)
fit.ar1ma1 <- arima(kglotemp.log, order = c(1, 1, 1), method = c("ML"))
print(fit.ar1ma1)## One coefficient is not significant.


#normality check
resar3 = residuals(fit.ar3)
resa1m1 = residuals(fit.ar1ma1)
resa1m3 = residuals(fit.ar1ma3)
shapiro.test(resar3)
shapiro.test(resa1m1)
shapiro.test(resa1m3)

op <- par(mfrow = c(2,3))
hist(resar3, prob=TRUE)
lines(density(resar3))
hist(resa1m1, prob=TRUE)
```

```
lines(density(resa1m1))  # add a density estimate with defaults
hist(resa1m3, prob=TRUE)
lines(density(resa1m3))
##Histograms both look Okay
qqnorm(resar3)
qqline(resar3,col ="blue")
qqnorm(resa1m1)
qqline(resa1m1,col ="blue")##Yhis one looks worst.
qqnorm(resa1m3)
qqline(resa1m3,col ="blue")


#correlation independent check
Box.test(residuals(fit.ar3), lag = 12, type="Ljung", fitdf = 3)
Box.test(residuals(fit.ar1ma1), lag = 12, type="Ljung", fitdf = 2)##Didn't past the Ljung check.
Box.test(residuals(fit.ar1ma3), lag = 12, type="Ljung", fitdf = 4)
op <- par(mfrow=c(1,2))
acf(residuals(fit.ar3))
pacf(residuals(fit.ar3))
par(op)
##AR(3) has some stabs.
op <- par(mfrow=c(1,2))
acf(residuals(fit.ar1ma1))
pacf(residuals(fit.ar1ma1))
par(op)
##ARIMA(1,1) has some stabs.
op <- par(mfrow=c(1,2))
acf(residuals(fit.ar1ma3))
pacf(residuals(fit.ar1ma3))
par(op)
##ARIMA(1,3) pasted acf and pacf test, all in confidence intervals.


# constant variance check
#According from the Time Series plot after 1 difference, variances are constant.

#forecasting
pred.tr <- predict(fit.ar1ma3, n.ahead=10)
U.tr= pred.tr$pred+ 2*pred.tr$se
D.tr= pred.tr$pred- 2*pred.tr$se
ts.plot(kglotemp.log, xlim=c(1880, 2026), ylim=c(-1,max(U.tr)))
polygon(c(2017:2026,rev(2017:2026)),c(D.tr,rev(U.tr)),col="thistle"
,border=NA)
points(2017:2026,pred.tr$pred, col = "green", pch = 24)
lines(U.tr, col = "red", lwd=2)
lines(D.tr, col = "red", lwd=2)

pred.orig = exp(pred.tr$pred) - 1
U.orig = exp(U.tr) - 1
D.orig = exp(D.tr) - 1
ts.plot(glotemp, xlim=c(1880, 2026),ylim=c(-1,max(U.orig)))
polygon(c(2017:2026,rev(2017:2026)),c(D.orig,rev(U.orig)),col="thistle"
,border=NA)
points(2017:2026,pred.orig, col = "green", pch = 1)
```

```
lines(U.orig, col = "red", lwd=2)
lines(D.orig, col = "red", lwd=2)
```