STA 137 Final Project

Hui Qi 916408440

Introduction

The data is an annual temperature anomalies dataset for the northern hemisphere from 1850 to 2021 by the Climate Research Center at the University of East Anglia in the UK. This data is a time series since it is a set of observations, each recorded at a specific time, and this is a discrete set. By selecting the fitted model for this series, this project can draw inferences from this series. By removing the error terms or rough part, this study could provide a long-term trend of the annual temperature anomalies, which could help scientists and governments predict and better prepare for the future phenomenon of temperature anomalies in the northern hemisphere. It is also possible to detect the cyclic behavior of the annual temperature anomalies dataset.

Materials and Methods

There are 172 observations in total, and the mean of this dataset is -0.0345 while its variance is 0.1901463. The primary tool used for building and analyzing the models is the R-Studio package.

This study will start by figuring out whether this dataset is stationary. If it is stationary, then the mean and variance of the dataset do not vary in time. This step is essential because further statistical tests and models rely on it. Achieving this goal requires plotting the dataset and visualizing the data to determine whether it presents some known stationary data property. Then with the Autocorrelation Function (ACF) plot, it will be clearer to figure out its stationarity. However, if it is not stationary, the difference of series will be used to achieve stationarity. After that, by obtaining the ACF and PACF plots of the difference series, it is possible to gain a preliminary identification of a time series model. With this initial identification, an ARIMA model can be fitted to examine the residuals and their properties to detect if this model is appropriate and suitable to the dataset. Next, this paper calculates the model selection criterion AIC to decide the final model.

Moreover, it is essential to do the spectral analysis of the data, which states that any stationary time series can be written as linear combinations of sines and cosines with random coefficients to detect the frequency and the pattern. This paper will also use the final model parameters to forecast temperature anomalies of the last six years and compare them with the actual observational values.

Results:

From the below Figure 1, the data does not fluctuate with a fixed value, which implies that this data is not stationary. This project operates loess with spans=25 to approach the data to understand the nature of variation in the data. The red loess line in the graph is a good fit for the data trend. By deleting the trend from the dataset, this project gets the rough part of the data. It is also clear that the variance is changing through time from Figure 2, as at t=30, there is a significant fluctuation, while at t= 50, the fluctuation is smaller. Therefore, this also shows that the time series is not stationary. From the histogram plot of residues in Figure 3, the residues are left-skewed. Furthermore, from figure 4, loess is fitted to the data since most points are on the 45-degree line in the qq-plot. But the points at the ends of the data do not fall on the line.

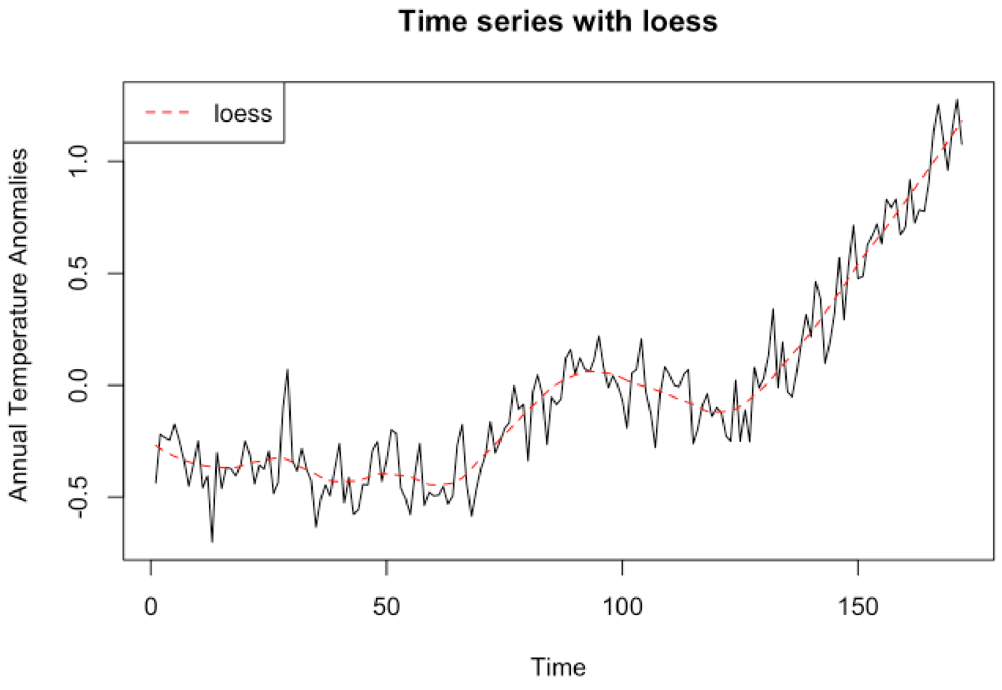


Figure 1

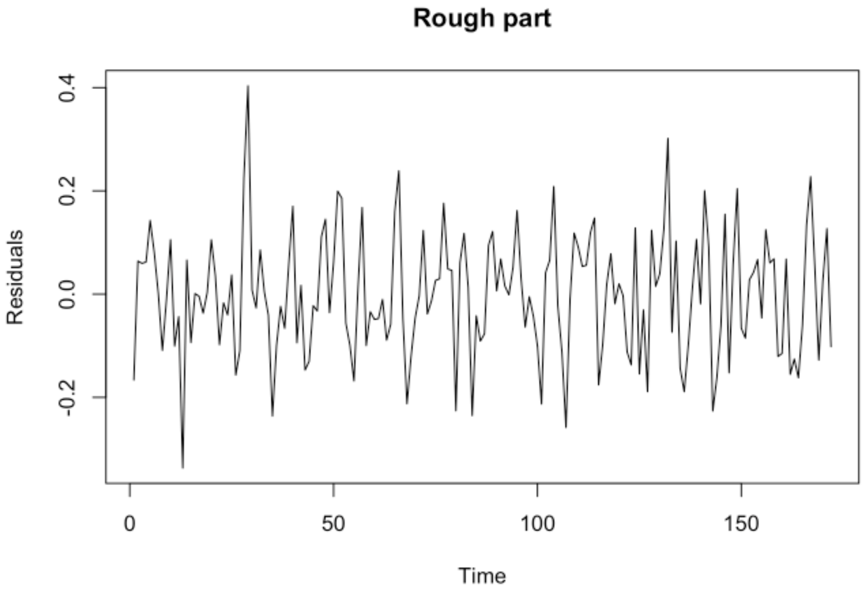


Figure 2

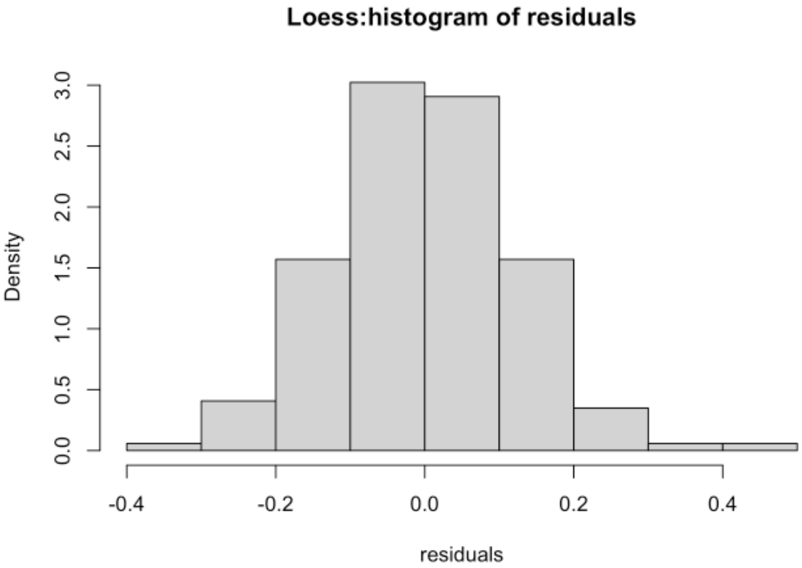


Figure 3

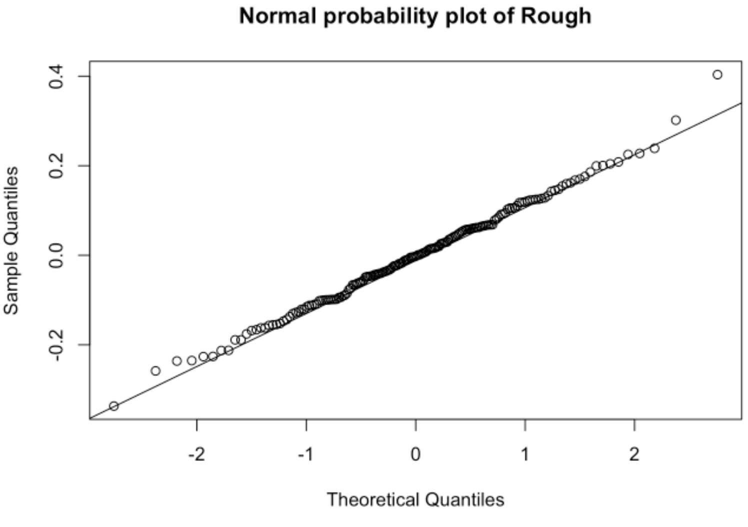


Figure 4

Since the time series is not stationary, this study uses the first difference of the series to achieve stationarity. After getting the first difference series, the plot of the first difference series against time demonstrates to fluctuate around 0, which implies it is stationary from Figure 5.

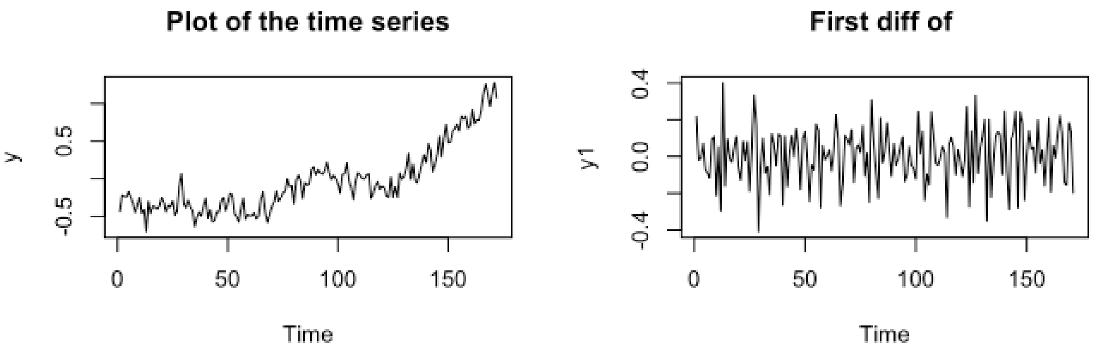


Figure 5

The ACF and PACF of the first difference time series in Figure 6 below could indicate a preliminary identification of the data model. ACF plot helps detecting the moving average models. Since lag 0 and lag 1 are significant while the rest lags are within the confidence interval, it implies MA(1). The PACF plot is useful in detecting autoregressive models. Since lag 1, lag 2, and lag 3 are significant, it implies AR(3). Therefore, the ACF and PACF indicates that ARIMA(3,1,1) is an appropriate model from the annual temperature anomalies.

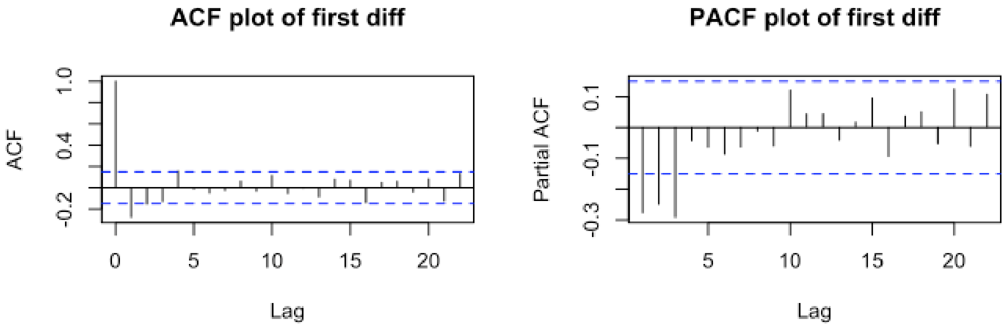


Figure 6

With ARIMA(3,1,1), the R-studio provides the residual plots with this new model. In Figure 7, the fluctuation seems more stable than the previous loess fit one. Moreover, the histogram distributed in a bell shape and more points on the ends of the Normal Probability Plot falls on the 45-degree line from Figure 8.

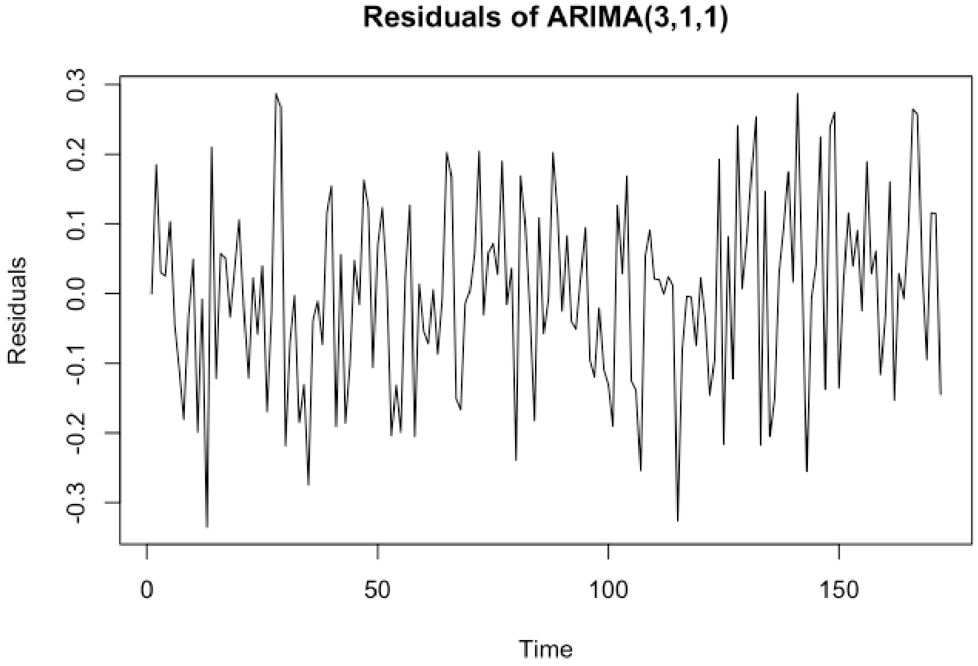


Figure 7

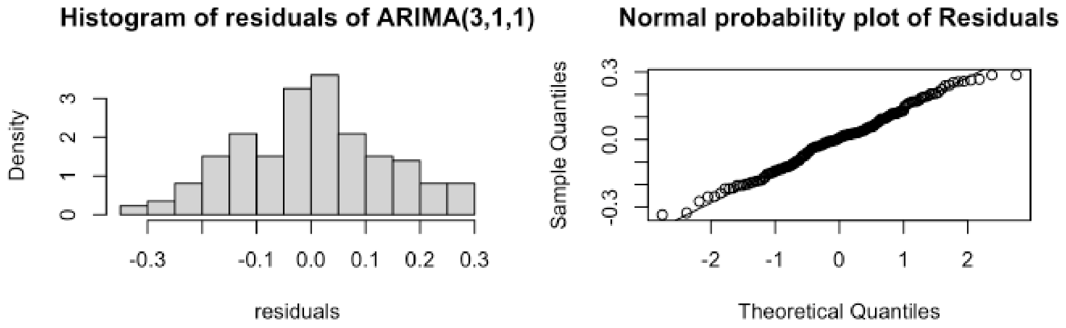
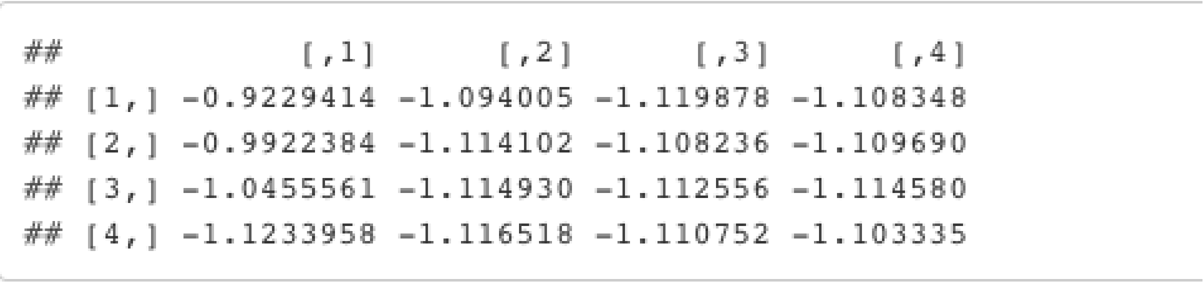


Figure 8

However, to achieve the most appropriate model, it is better to use the model selection criterion. The R-Studio gets the result of the AIC of 16 models with p=0,1 ,2, 3 and q=1, 2, 3 in Figure 9. -1.1233958 in the fourth row and first column is the smallest value in this table, which implies that ARIMA(3,1,0) will be a better fit model for the annual temperature anomalies dataset.



```
##               [,1]      [,2]      [,3]      [,4]
## [1,] -0.9229414 -1.094005 -1.119878 -1.108348
## [2,] -0.9922384 -1.114102 -1.108236 -1.109690
## [3,] -1.0455561 -1.114930 -1.112556 -1.114580
## [4,] -1.1233958 -1.116518 -1.110752 -1.103335
```

Figure 9

The ACF plot of the residuals of ARIMA(3,1,0) in Figure 10 has only significant Lag 0, which shows that this model is a good fit for the dataset.

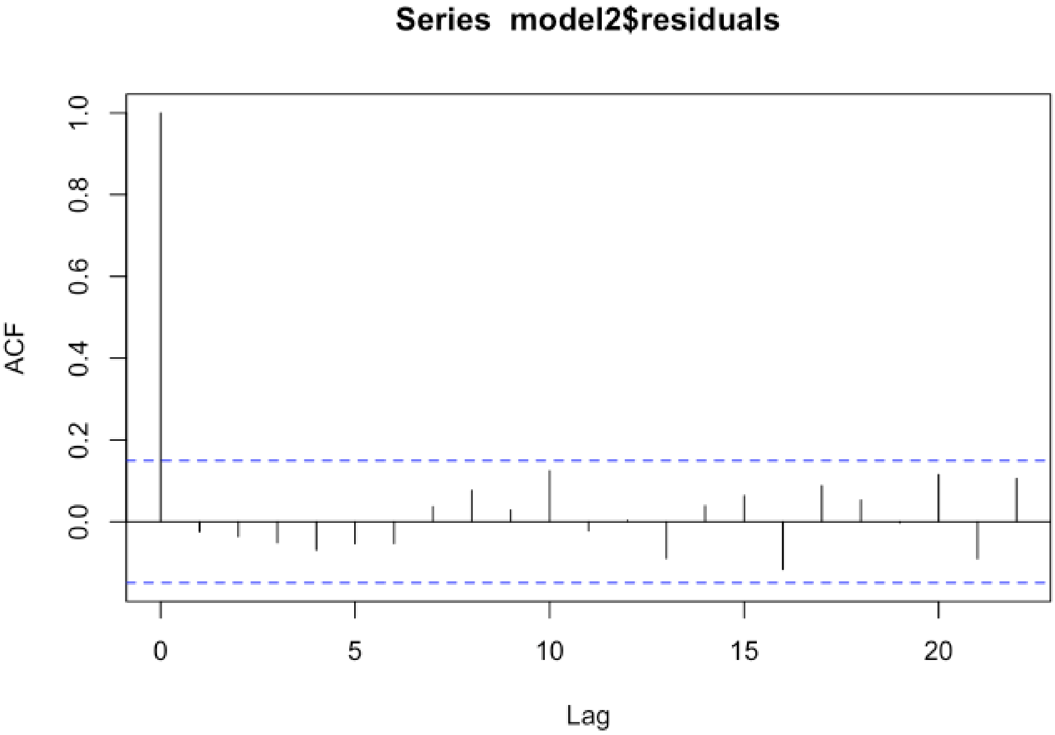


Figure 10

Moreover, the histogram and the Normal Probability plots of residuals in Figure 11 also indicate that this model is suitable for analyzing the data because the Histogram plot is approximately symmetric. Also, almost all the points fall on the 45-degree line.

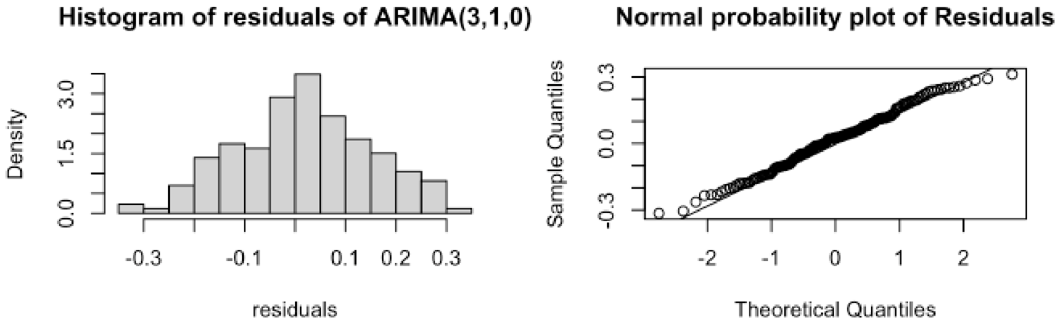
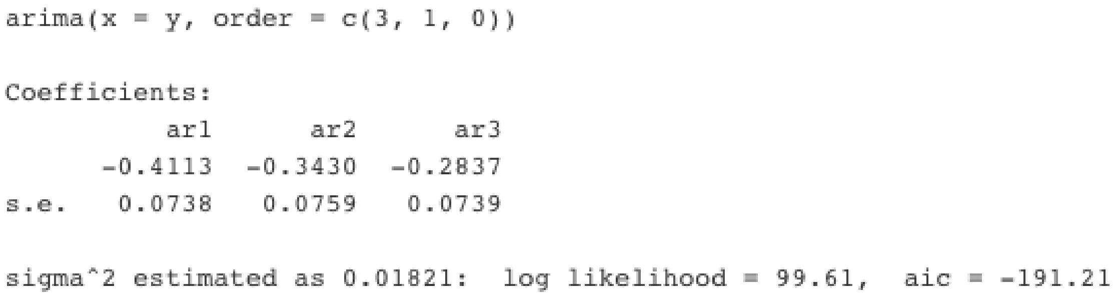


Figure 11

The ARIMA(3,1,0) model has parameters estimated ar1 = -0.4113, ar2 = -0.3430, and ar3 = -0.2837 as shown in Figure 12. Their standard errors are 0.0738, 0.0759, and 0.0739 respectively. Notice that ar1 means phi1; ar2 means phi2; ar3 means phi3.



```
arima(x = y, order = c(3, 1, 0))

Coefficients:
          ar1      ar2      ar3
      -0.4113  -0.3430  -0.2837
s.e.   0.0738   0.0759   0.0739

sigma^2 estimated as 0.01821:  log likelihood = 99.61,  aic = -191.21
```

Figure 12

In order to obtain the values of the criterion function for obtaining the optimal number of neighbors for spectral density estimate for modifies Daniell’s method, Figure 13 demonstrates that the point at c = 12 has the lowest value un the criterion function. Therefore, $spans = 2 \times 12 + 1 = 25$.

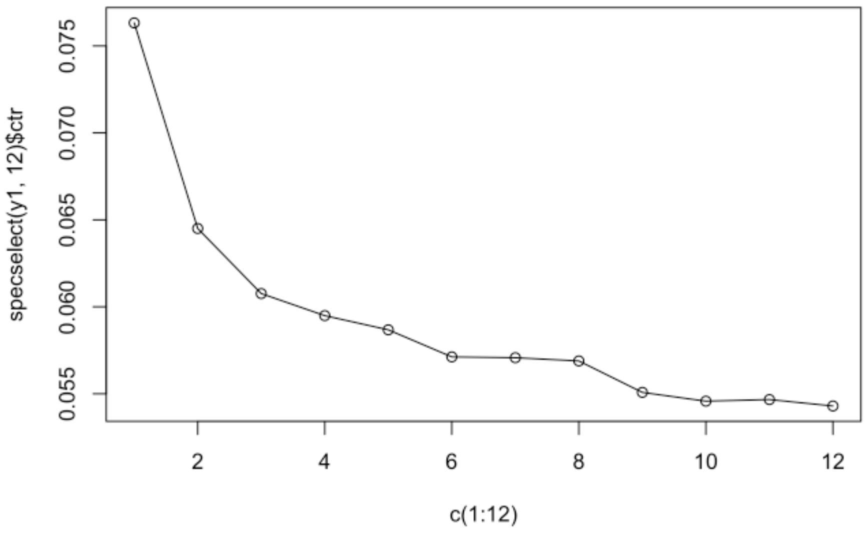


Figure 13

Figure 14 shows the plot of the spectral density function and the smooth periodogram with a chosen number of smoothing spans=2×12+1=25. The black and blue lines have similar trends and shapes. They are similar for frequencies between 0.0 and 0.17. they are also similar for frequencies between 0.3 and 0.45. but they differ for frequencies between 0.17 and 0.3. the sample size is n = 172 hear. If the sample size were much larger than 127, the periodogram line would be smoother over more neighbors. Moreover, the differences between the smoothed periodogram and the densities from fitted models may not be so significant.

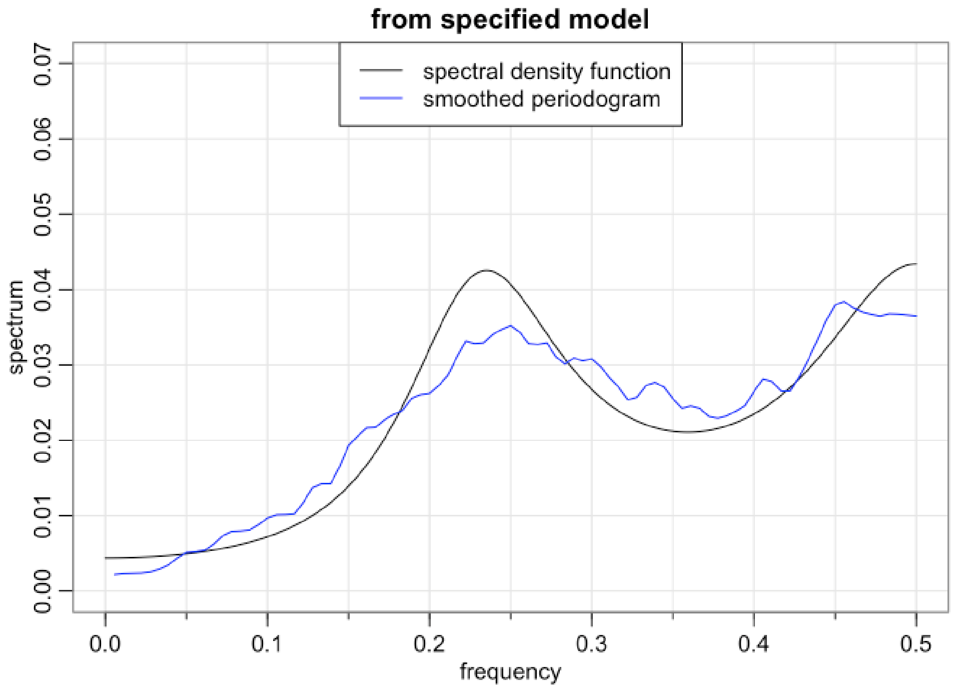


Figure 14

This project operates all the data except for the last six years and the ARIMA(3,1,0) model to forecast the temperature anomalies from 2016 to 2021. As shown in Figure 15, the blue line is the fitted line, and the black line is the actual data's true line. The blue bounded region is the 95% confidence interval of the true line. This model is a good prediction since the fitted line falls in the blue region.

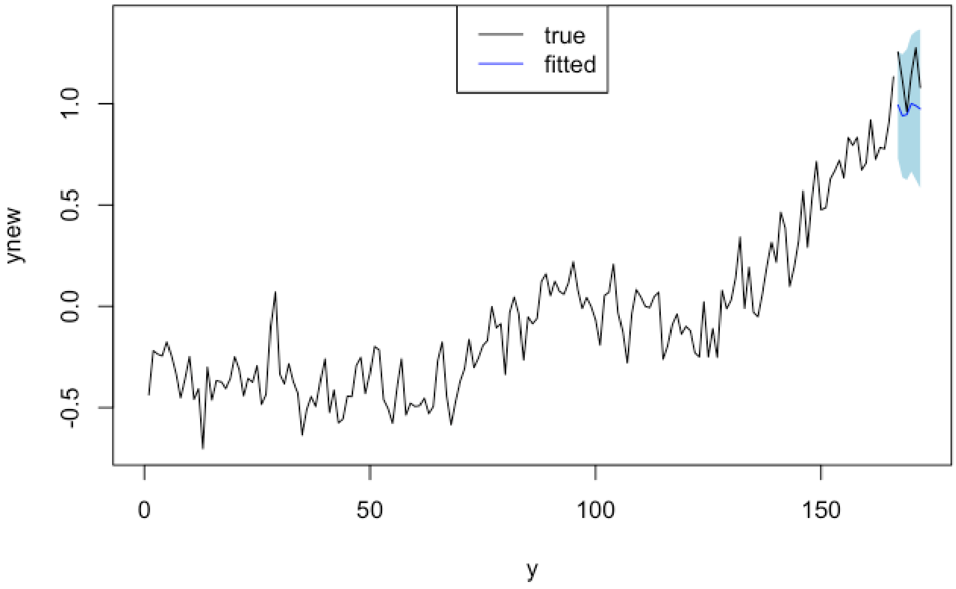


Figure 15

Conclusion

This project approaches a dataset by selecting and diagnosing three models to find the final one that is most appropriate to the data. It demonstrates the process of visualizing, analyzing, and making predictions of a dataset with the help of R-Studio. The plots help detect the pattern, stationarity, and normality within the whole study. Although the forecasted line falls in the confidence interval region from the final model, the fitted line from ARIMA(3,1,0) still significantly differs from the actual value. The fitted model and the prediction can be improved by increasing the sample size. Though the data of annual temperature anomalies are a natural event, it will have times that the variance suddenly grows in a year. But it is still necessary for Researchers to collect and analyze the dataset to help the citizens better prepare the possible weather changes.

Project

Hui Qi

3/13/2022

R Markdown

```
library(readxl)
TempNH <- read_excel("TempNH_1850_2021.xlsx", col_names = c('x','y'))
x = TempNH[,1]
y = TempNH[,2]
y = unlist(y)
```

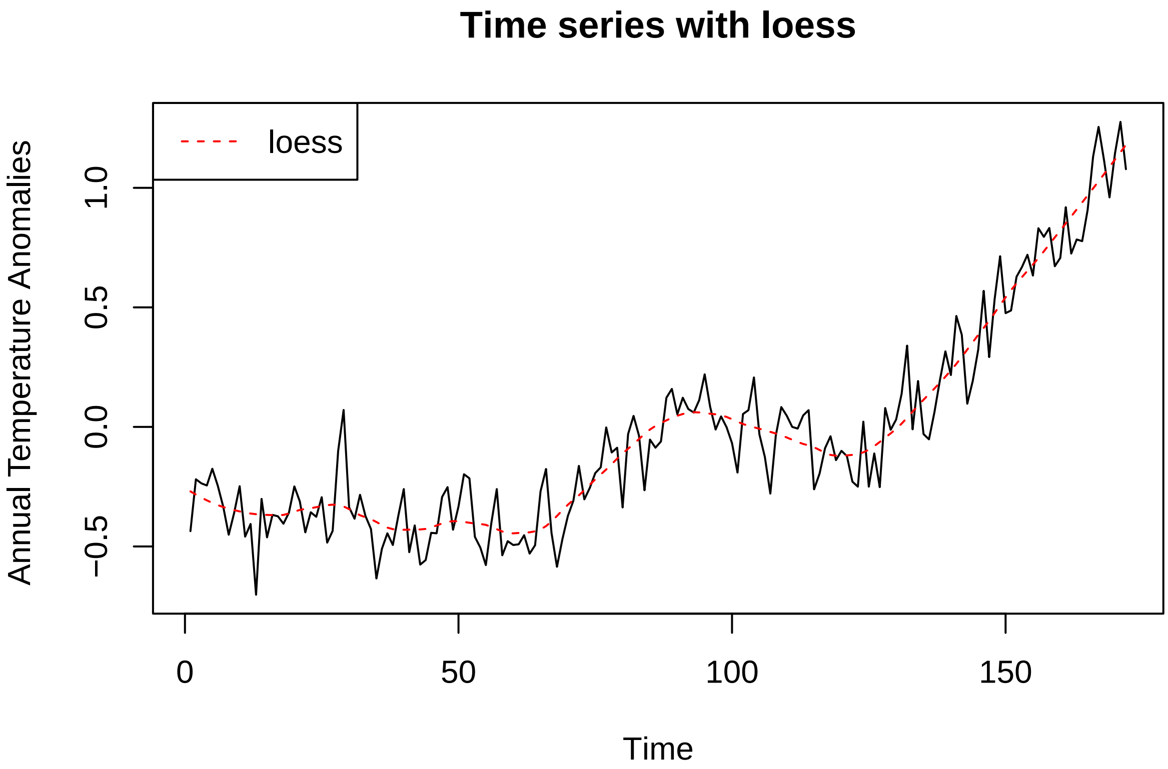
```
summary(y)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.70200 -0.36125 -0.12400 -0.03456  0.10075  1.27600
```

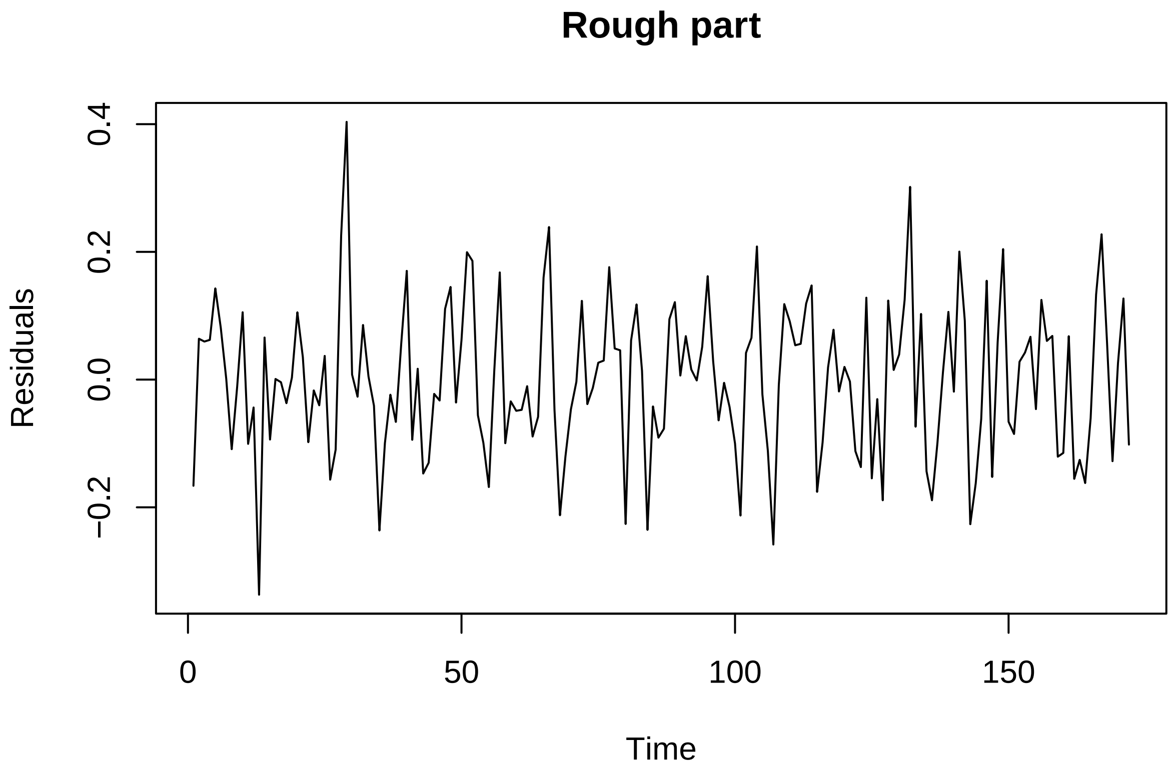
```
var(y)
```

```
## [1] 0.1901463
```

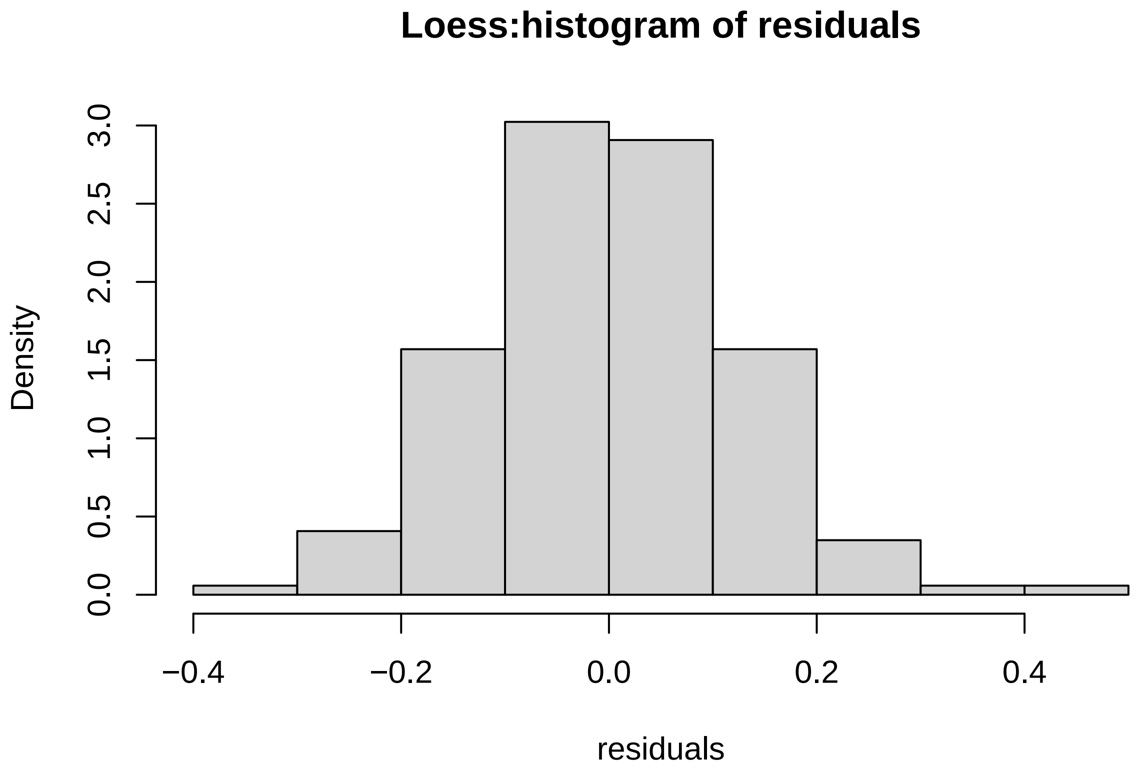
```
tm = 1:172
loesstrnd=loess(y~tm, span  = 0.25)

plot(tm, y, type="l", lty=1, xlab="Time", ylab="Annual Temperature Anomalies ", main="Time series with 
points(tm, loesstrnd$fitted, type="l", lty=2, col= "red")
legend("topleft", "loess", lty = 2, col = "red")
```



```
rough = loesstrnd$residuals
plot(rough, type='l',xlab = 'Time', ylab='Residuals', main = 'Rough part')
```



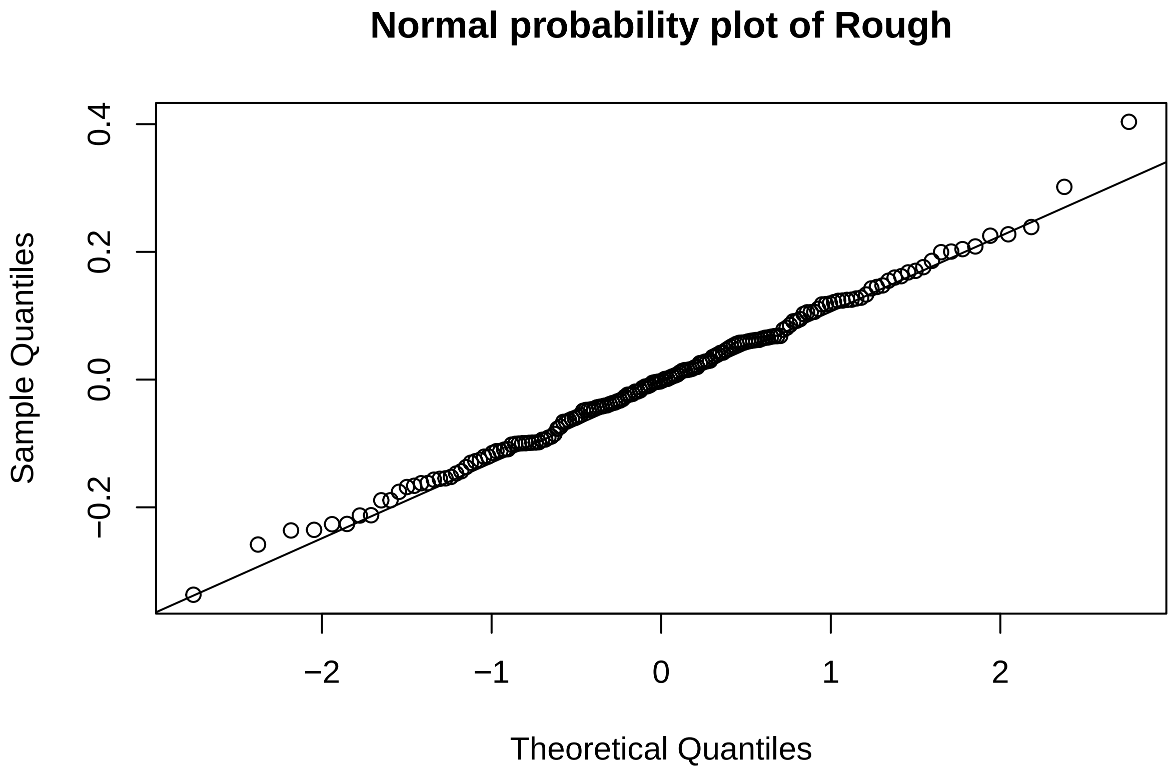
```
hist(loesstrnd$residuals, freq = F, xlab="residuals", main="Loess:histogram of residuals")
```



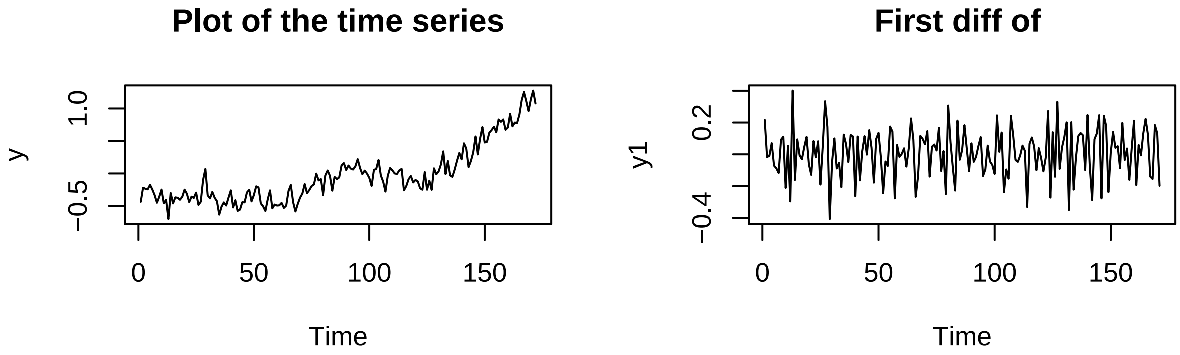
```
sse <- sum((loesstrnd$residuals)^2)
ssto <- sum((y-mean(y))^2)
R2 <- 1-sse/ssto
R2
```

```
## [1] 0.9264963
```

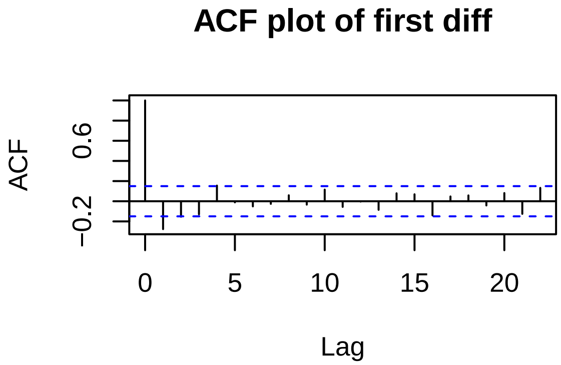
```
qqnorm(rough, main = "Normal probability plot of Rough")
qqline(rough)
```

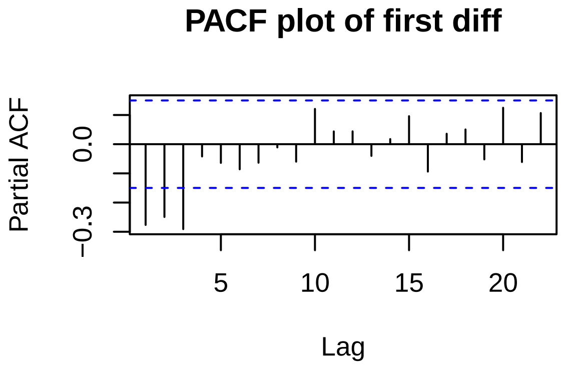


```
par(mfrow =c(2,2))
plot.ts(y, main = 'Plot of the time series')

y1 = diff(y,1)
plot.ts(y1, main = 'First diff of ')

par(mfrow =c(1,1))
```

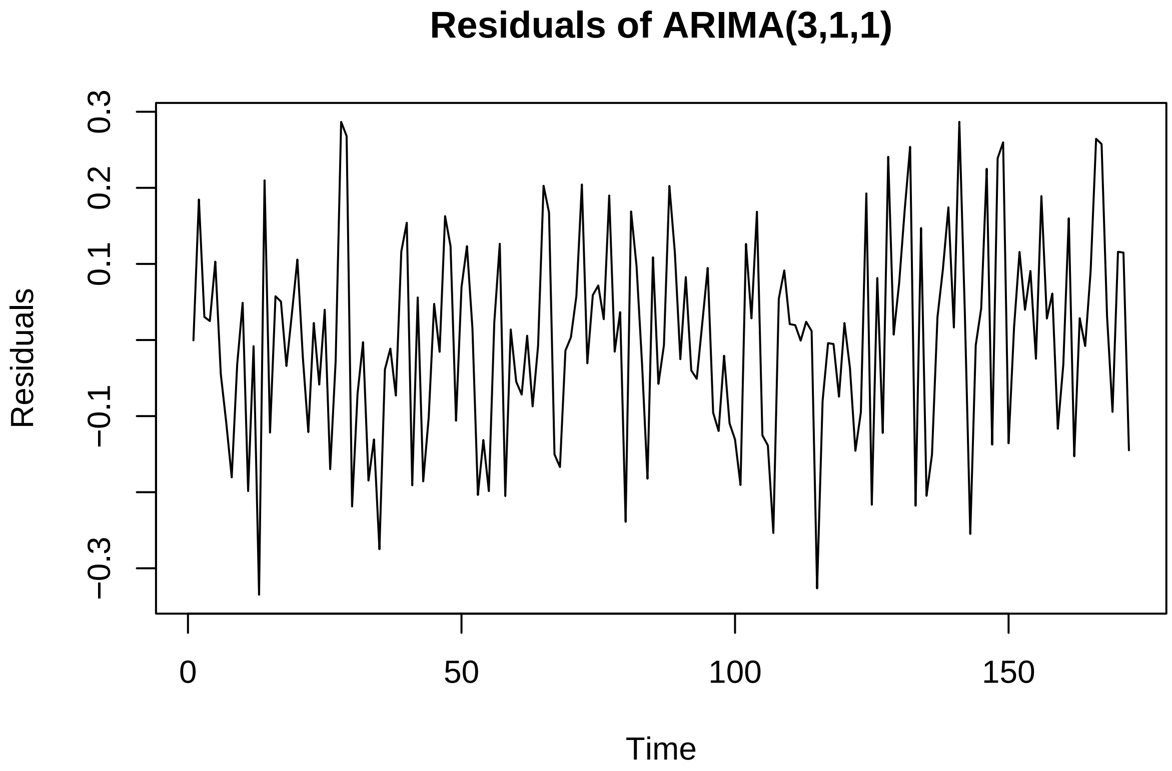


```
par(mfrow =c(2,2))
acf(y1, main = 'ACF plot of first diff')
pacf(y1, main = 'PACF plot of first diff')

par(mfrow =c(1,1))
```

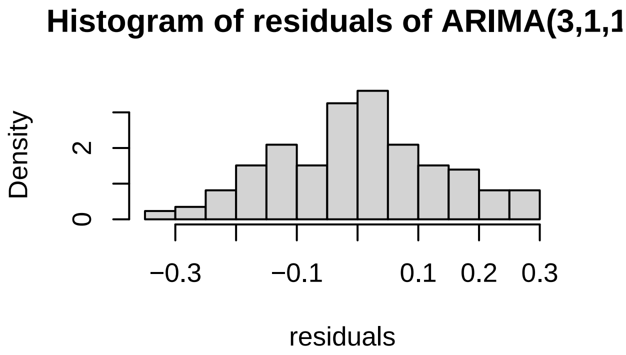


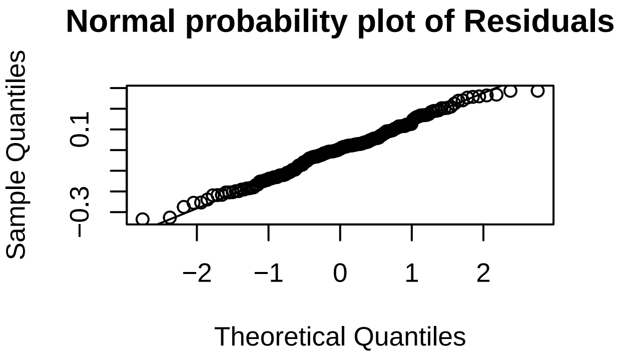


```
library(astsa)
model1 = sarima(y, p=3,d=1,q=1,details=FALSE)

model1res= model1$fit$residuals

plot(model1res, type='l',xlab = 'Time', ylab='Residuals', main = 'Residuals of ARIMA(3,1,1)')
```



```
par(mfrow=c(2,2))
hist(model1res, freq = F, xlab="residuals", main="Histogram of residuals of ARIMA(3,1,1)")

qqnorm(model1res, main = "Normal probability plot of Residuals")
qqline(model1res)

par(mfrow=c(1,1))
```





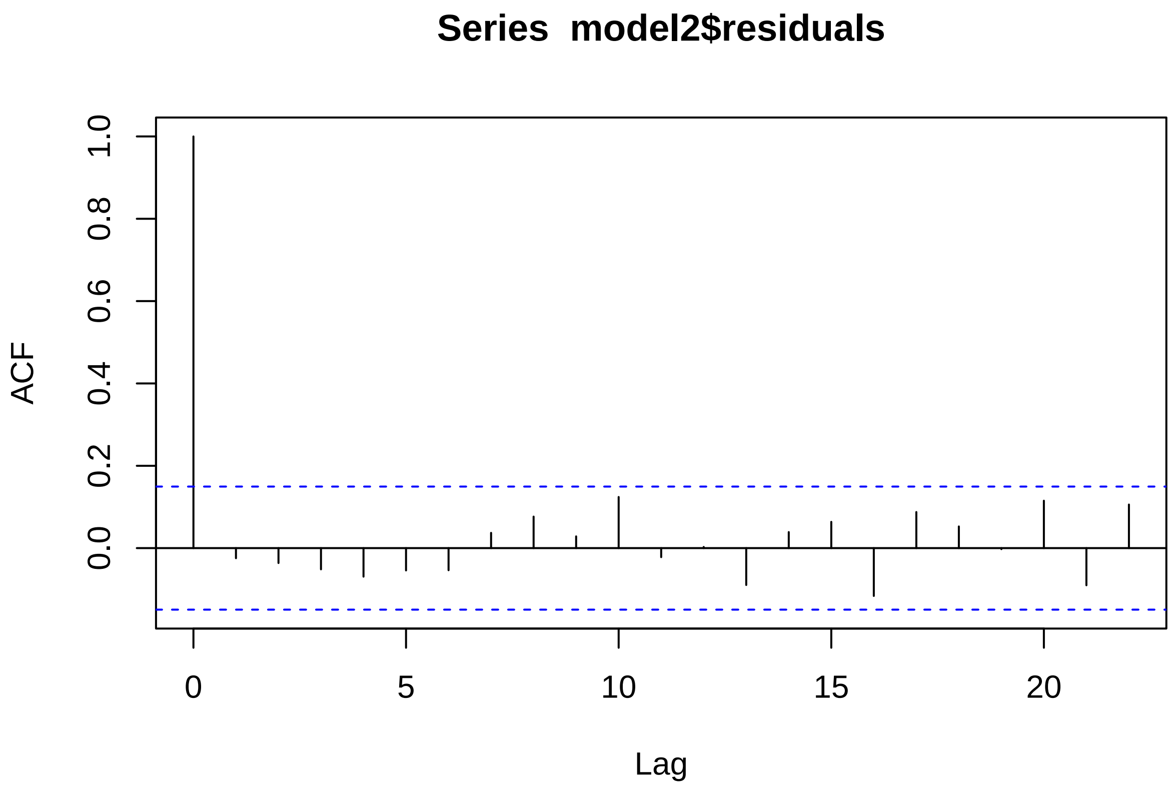
```
AIC = matrix(0,4,4)
for (i in 1:4){
  for (j in 1:4){
    AIC[i,j] <- sarima(y, p = i-1, d=1, q = j-1, details=FALSE)$AIC
  }
}
```

```
## Warning in sqrt(diag(fitit$var.coef)): NaNs produced

## Warning in sqrt(diag(fitit$var.coef)): NaNs produced
```

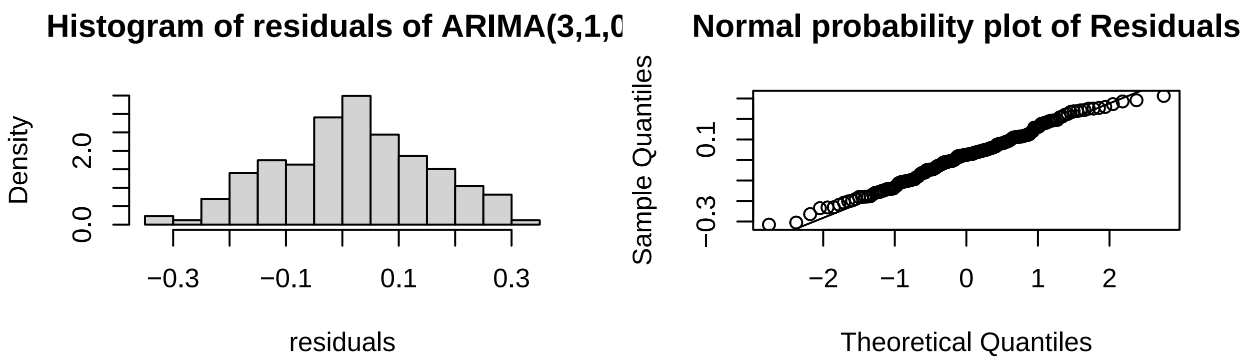
```
AIC
```

```
##            [,1]      [,2]      [,3]      [,4]
## [1,] -0.9229414 -1.094005 -1.119878 -1.108348
## [2,] -0.9922384 -1.114102 -1.108236 -1.109690
## [3,] -1.0455561 -1.114930 -1.112556 -1.114580
## [4,] -1.1233958 -1.116518 -1.110752 -1.103335
```

```
# -1.1233958 is the smallest ARIMA(3,1,0)

model2<-arima(y,order=c(3,1,0))
acf(model2$residuals)
```



```
par(mfrow=c(2,2))
hist(model2$residuals, freq = F, xlab="residuals", main="Histogram of residuals of ARIMA(3,1,0)")

qqnorm(model2$residuals, main = "Normal probability plot of Residuals")
qqline(model1res)

par(mfrow=c(1,1))
```



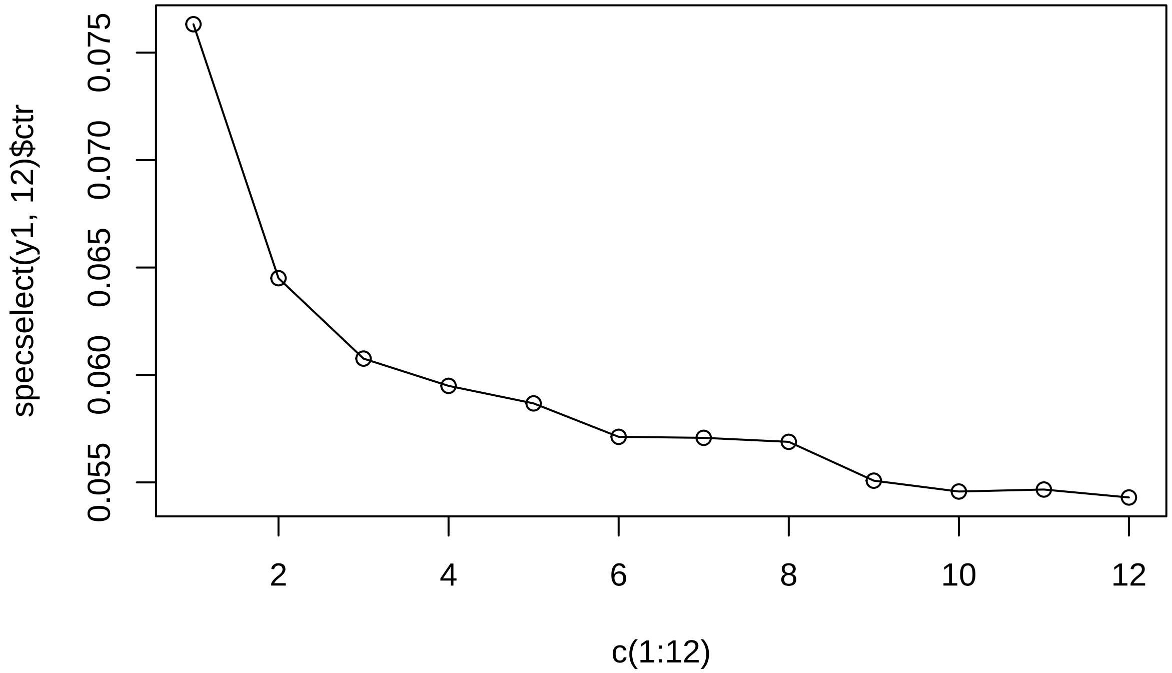
```
model2
```

```
##
## Call:
## arima(x = y, order = c(3, 1, 0))
##
## Coefficients:
##           ar1      ar2      ar3
##       -0.4113  -0.3430  -0.2837
## s.e.   0.0738   0.0759   0.0739
##
## sigma^2 estimated as 0.01821:  log likelihood = 99.61,  aic = -191.21
```

```
specselect=function(y,kmax){
# Obtains the values of the criterion function for
# obtaining the optimal number of neighbors for
# spectral density estimate for modified Daniell's method.
# input: y, observed series; kmax=max number of neighbors to
# be considered
# output: ctr - the criterion function
# output: kopt - the value of k at which the criterion function # is minimized
ii=spec.pgram(y,log="no",plot=FALSE)
ii=ii$spec
cc=norm(as.matrix(ii),type="F")^2
ctr=rep(1,kmax) ###criterion function
for(k in 1:kmax) {
ss=2*k+1; kk=1/(2*k)
ff=spec.pgram(y,spans=ss,log="no",plot=FALSE)
fspec=ff$spec
ctr[k]=norm(as.matrix(ii-fspec),type="F")^2+kk*cc
}

kopt=which.min(ctr)
result=list(ctr=ctr,kopt=kopt)
return(result)
}

specselect(y1,12)
```

```
## $ctr
##  [1] 0.07632212 0.06450387 0.06076200 0.05949365 0.05867776 0.05712129
##  [7] 0.05707398 0.05688819 0.05508067 0.05457552 0.05466916 0.05429911
##
## $kopt
## [1] 12
```

```
plot(c(1:12),specselect(y1,12)$ctr,type="o")
```



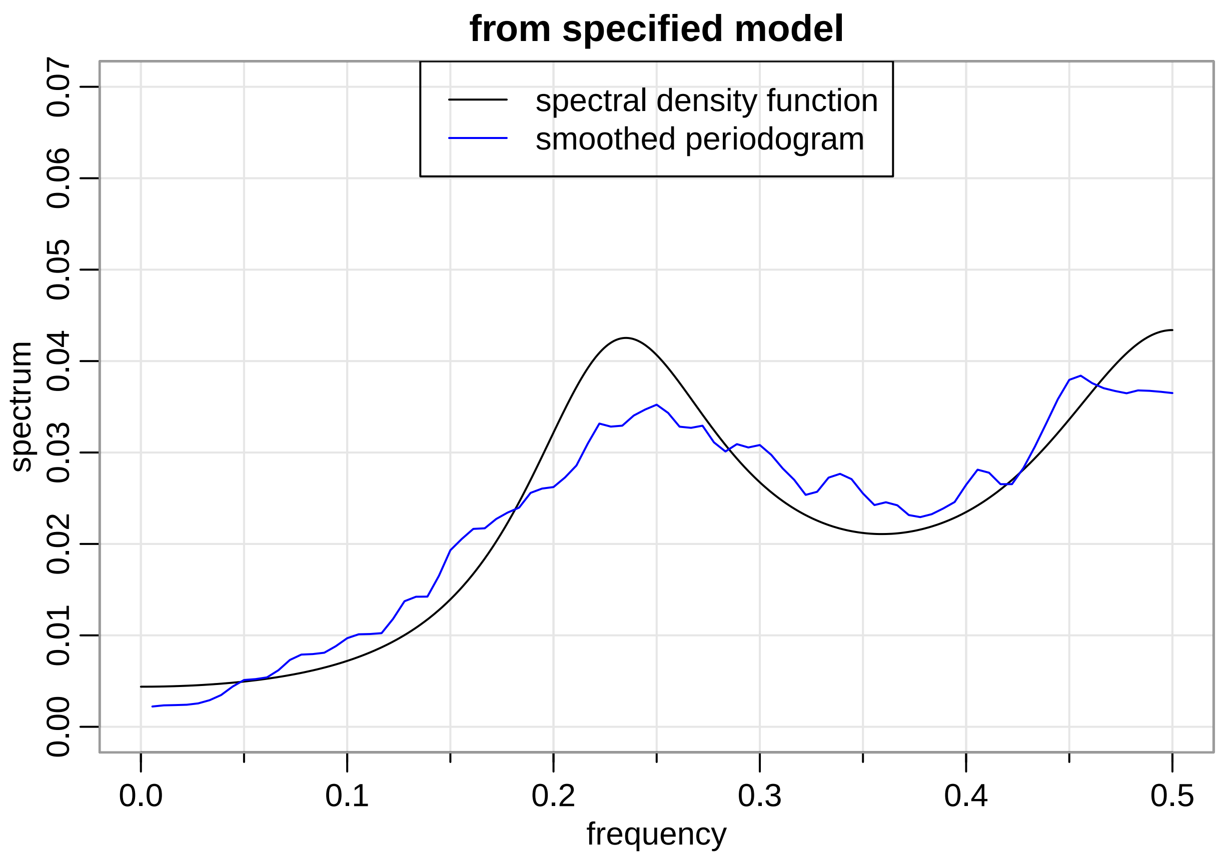
```
koptimal<-specselect(y1,12)$kopt ##the one which minimizes the criterion function
spans<-koptimal*2+1 ##optimal span
spans
```

```
## [1] 25
```

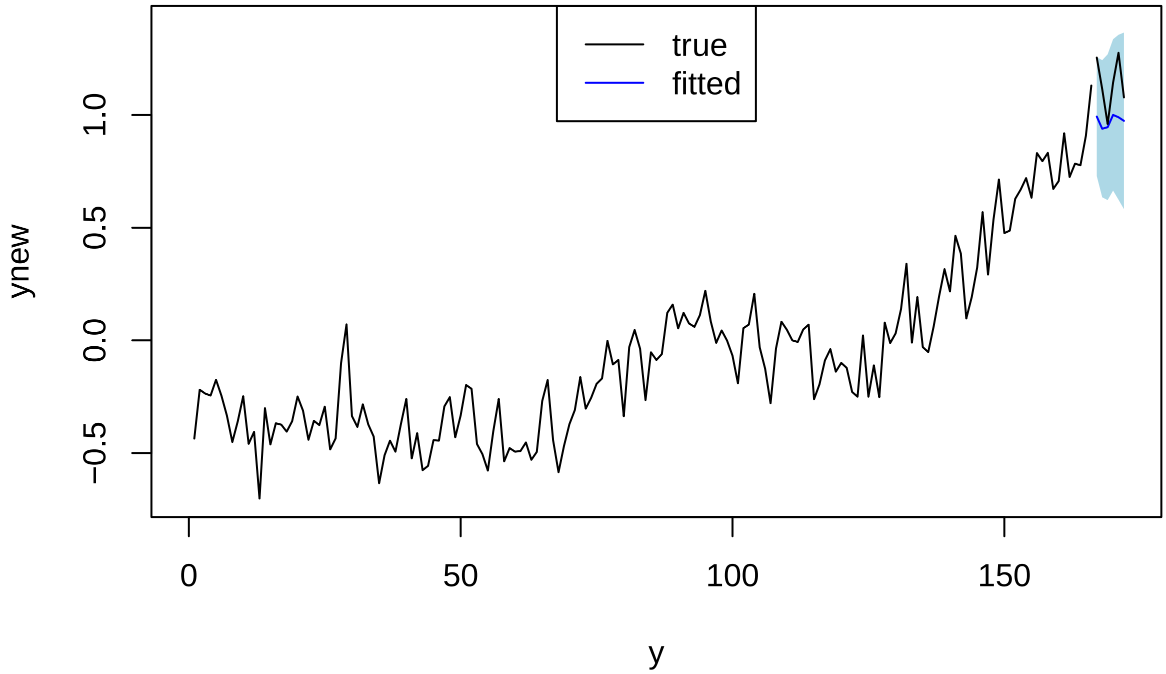
```
model2
```

```
##
## Call:
## arima(x = y, order = c(3, 1, 0))
##
## Coefficients:
##           ar1      ar2      ar3
##       -0.4113  -0.3430  -0.2837
## s.e.   0.0738   0.0759   0.0739
##
## sigma^2 estimated as 0.01821:  log likelihood = 99.61,  aic = -191.21
```

```
spec_smooth <- spec.pgram(y1, spans=25, log="no", plot=FALSE)
freq <- spec_smooth$freq
spec <- spec_smooth$spec

library(astsa)
arma.spec(ar = model2$coef[1:3], var.noise = model2$sigma2, ylim = c(0,0.07))
lines(freq,spec, col = "blue")
legend("top", legend = c("spectral density function", "smoothed periodogram"), lty = c(1,1), col = c("bl
```



```
n = length(y)
ynew<- y[1:(n-6)]
ylast<- y[(n-5):n]

model4<-arima(ynew, order = c(3,1,0))
h <- 6
m <- n-h
fcast <- predict(model4, n.ahead=h)
upper <- fcast$pred+1.96*fcast$se
lower <- fcast$pred-1.96*fcast$se
#plot
plot.ts(ynew, xlim = c(0,n), xlab = "y", ylim=c(-0.7,1.4))
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=ylast, col="black")
legend("top", legend = c("true","fitted"), lty=c(1, 1), col = c("black","blue"))
```


true
fitted
ynew
−0.5
0.0
0.5
1.0
0
50
100
150
y