Time Series Analysis of Stream Flow

Time Series Analysis Class Project:

David Navar

March 12, 2017

Stats 221

Professor Rick Paik Schoenberg

University of California Los Angeles

INTRODUCTION:

I chose to analyze a monthly streamflow time series dataset of the Liao river in the northeastern region of China. I obtained the data from a colleague in my graduate program, the Civil and Environmental Engineering Department at UCLA. The Chinese government is the agency that collected and stored the data. Obtaining Chinese environmental data is difficult because China does not openly distribute its data. The colleague in my department was able to obtain the data for research and educational purposes. It is a great dataset to run simulations on due to its length and integrity.

Environmental measurements of this time duration and integrity are difficult to find. Streamflow measurements were recorded for a span of 50 years, and consist of a total of 612 data points. The values recorded are the total rainfall accumulation for each month in units of volumetric flow rate, or cubic feet per second (cfs).

The data presented in Figure 1 is the plot of the complete time series dataset analyzed in the project. The data shows a very clear pattern on a yearly basis, as flow rate is markedly higher in the wet season and lower in the dry season. The value of the dry season is consistent from year to year while the height of the wet season varies a lot in a semi-unpredictable pattern. The data appear to be rather stationary as it has an almost constant mean. The value of the wet season has a cyclical nature to it. It seems to increase in small increments over a course of about ten years. This pattern can be seen most clearly in the data points of indexes 260 to 350. This pattern can be seen in the periodogram, once the major yearly cycle is removed.

The autocorrelation function(ACF) and partial autocorrelation function (PACF) graphs are produced to show the correlation between the values of the time series. ACF and PACF graphs are produced in such a way as to omit the display of the 0-lag value, which would distort the presentation of the data. These plots are presented in Figure 3. The ACF shows a strong correlation for values 12 months apart. A negative correlation can be seen for values 6 months apart. This can be seen less so in the PACF however the 12-month positive correlation is still there. The PACF tapers off much more quickly than the ACF.

The data was then analyzed to determine if a trend existed. Upon examination, the data seems to be weakly stationary with a slightly increasing mean. As can be seen in Figure 2, this slope of the mean is hardly detectable to the eye but it can be characterized by the equation y=.1906955x+1184.31. The slope of .1906 can probably be attributed to some natural phenomena like climate change, or shifting atmospheric weather patterns. Once the trend was detected and subsequently removed, graphs were produced to determine the correlation of the data. Figure 4 contains the detrended ACF and PACF graphs. These plots appear be almost identical to the raw data ACF and PACF due to the small trend that was removed.

In Figure 5, differencing the data revealed more about the data. The ACF takes on an interesting shape, rounded on the bottom and sharp top for positive correlation. The ACF also

seems to taper off very slowly. In the differenced PACF, the cutoff point of 11 is obvious. Every PACF value after 11 seems to be statistically insignificant.

REMOVING THE YEARLY CYCLE:

At first, at this point of the analysis, I attempted to fit an ARMA model to the detrended and differenced data. This endeavor was unsuccessful and as a result, I presented incorrect results for my class presentation. Specifically, there is a large amount of autocorrelation in the residuals, and the Q tests were all significant. I was unable to fit a model that would fix this.

In order to successfully fit an ARMA model to my data, I first removed the obvious yearly cycle present in the data. To remove the cycle, I first found the mean value for each of the 12 months over the 50 years. I then subtracted the mean monthly value from each monthly value to produce a new dataset. This yielded a dataset that was closer to white noise as the protruding cycle was not detectable. I then proceeded to detrend the data, as well as take the first difference of the data. I then created ACF and PACF graphs for each of the operations. The trend on the cycle-normalized data can be seen in figure 6, the detrended, and differenced cycle-normalized ACF and PACF can be viewed in Figures 7 and 8.

Analysis of these plots is much more interesting to analyze as the plots show much less cyclical behavior. The ACF in Figure 8 shows strong positive correlations for early lags then negative values for later lags. The PACF seems to have no pattern to its partial correlation. The differenced data virtually eliminated significant auto-correlation. The differenced PACF shows negative correlation for early lags and then quickly fades into insignificance.

FITTING THE ARIMA MODEL:

In order to fit an Autoregressive Integrated Moving Average Model (ARIMA) model to the data, I used the R function: Auto.Arima(). I used the data set which had the cycle removed and I took the first difference of the data. The fit that the Auto.Arima() function recommended is ARIMA (0,1,2) fit with θ_1=-.53 and θ_2=-.388. The residuals analysis is presented in Figure 9. The fit was not satisfactory because the Q-statistic showed significant values although the autocorrelation between values of the residuals seemed to be very low.

To correct for the Q-Statistic failure, the auto-regression(AR) term was also included in the ARIMA fit. I found that the ARIMA (1,1,2) model fit better than the suggested model from Auto.Arima(). The coefficients were found to be ϕ_1=.5012, θ_1=-1.012 and θ_2=0.019. This corrected the Q-test values and also showed that the AR term is significant. The residual analysis produced by the sarima () function in R is shown in figure 10. In this model, the autocorrelation between the residuals remained low, although the Q-Q plot shows a high number of departures from normality. The ARIMA (1,1,2) is a good fit to the data but there is

still room for improvement. The ARIMA (1,1,1) with coefficients ϕ_1=.49 and θ_1=-1.00, is a slightly better fit. This model showed a better fit by the goodness-of-fit score through the Akaike's Information Criterion(AIC), Bayesian Information Criterion(BIC), and the bias corrected AIC, AICc. The fit criteria values are summarized in Figure 12 below.

The models had very similar AIC, AICc and BIC values, and the best fitting model is the ARIMA, (1,1,1) model, which had the lowest values for the parameters all around. The residual analysis of the ARIMA (1,1,1) model is shown in figure 11. There is a low amount of significant auto-correlation between the residuals. The Q-Q plot shows much deviation from the normality and a high number of outliers appear off of the normal line. The Q-Statistic shows that all the values are insignificant and therefore it is a good fit.

EVALUATING THE PERIODOGRAM:

A periodogram was first produced with the original data (Figure 13), which is the data before the removal of the yearly cyclical trend. This periodogram shows the strong yearly and half-yearly cycles that dwarf any other periodic trend in the data. The raw periodogram shows a very high spectral density for the yearly cycle. The smoothed periodogram shows a better variance for the approximation of spectral density, but still dwarfs the cycles below the main yearly and half-yearly peaks. The spectral density estimates are displayed in Figure 14 below. It should be noted that the value of spectral density for the raw periodogram is very high and shows itself in the plot through very large and consistent yearly peaks and troughs.

When the dominant cycles are removed from the data, the more subtle cycles can be observed. The cycle-normalized periodogram is presented in Figure 15. Here, we can see many steep peaks at the regions of lower frequency, and the peaks seem to lessen in intensity as the frequency increases. The most interesting peaks occur around the range of 10-13 years. These peaks, I suspect, are due to sun-spot cycles which typically last 11-12 years. The frequency with the highest peak is .0065 which corresponds to a period of 12.8 years. The second highest peak corresponds to a period of 2 years. There are many peaks very close together in the raw periodogram and therefore estimating the spectral density is difficult. Figure 16 shows the values of the estimation of spectral density. It should be noted that the spectral density of the cycle-normalized data set is smaller by about an order of magnitude. Figure 16 also shows that the smoother periodogram produced a reduced variance in the estimation of spectral density.

Since the cycle-normalized periodogram has many steep peaks in very proximity, when smoothed by a Danielle kernel smoother, the peak position changes frequencies corresponding to the combination of the peaks. I therefore changed the frequency at which I took the spectral density estimate so that my calculations corresponded to the two maximum peaks produced by the smoothing. The variance of the spectral estimation was also reduced due to Danielle kernel smoothing.

The raw time series corresponds sharply to the yearly cycle and its half-yearly harmonic cycle. If the yearly trend is not removed, it dominates the trend of the time series and therefore the smoothness of the data. Once the yearly cycle is removed, the periodogram of the normalized data unveils the intricacies of the more subdued cycles. The normalized periodogram shows that the lower frequencies contain a large amount of variance in the periodogram. These lower frequencies play a large role in the smoothness of the data although it is complementary to the dominant yearly cycle.

CONCLUSIONS:

If I were to continue the analysis on this dataset, I would investigate the cycles in the periodogram which were observable once the yearly cycle was removed. The lower frequencies of the periodogram are full of sharp peaks which could have interesting drivers behind them. I would be interested in parsing these patterns and removing them. The remainder could be studied to create a very precise tool for forecasting or at least quantifying expected patterns in streamflow. Even without the delving into the precision of the cycles, I think it would be a useful to create forecasting models for this data set to predict streamflow in the future.

I further would like to understand the harmonics in the data, especially in these hidden more subdued cycles and understand what is causing the harmonics. Along those same lines, the ARIMA model can also be improved my implementing seasonal parameters to the fit. Environmental data is certainly subject to seasonal variation and I think it would be useful to understand how seasonality is represented in the data.

FIGURES:

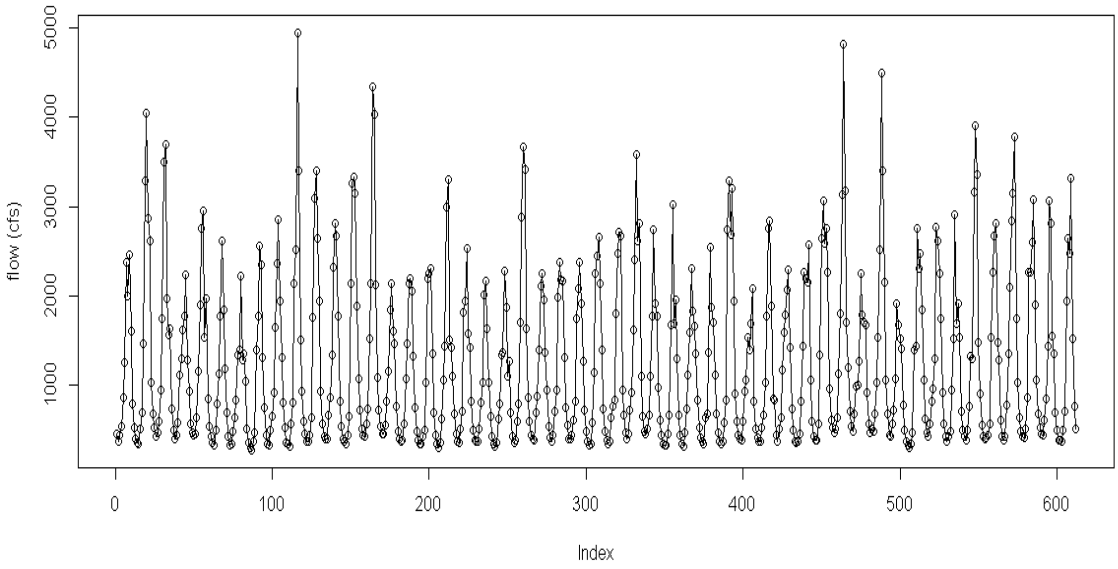

Figure 1 Time Series Plot of Monthly Streamflow

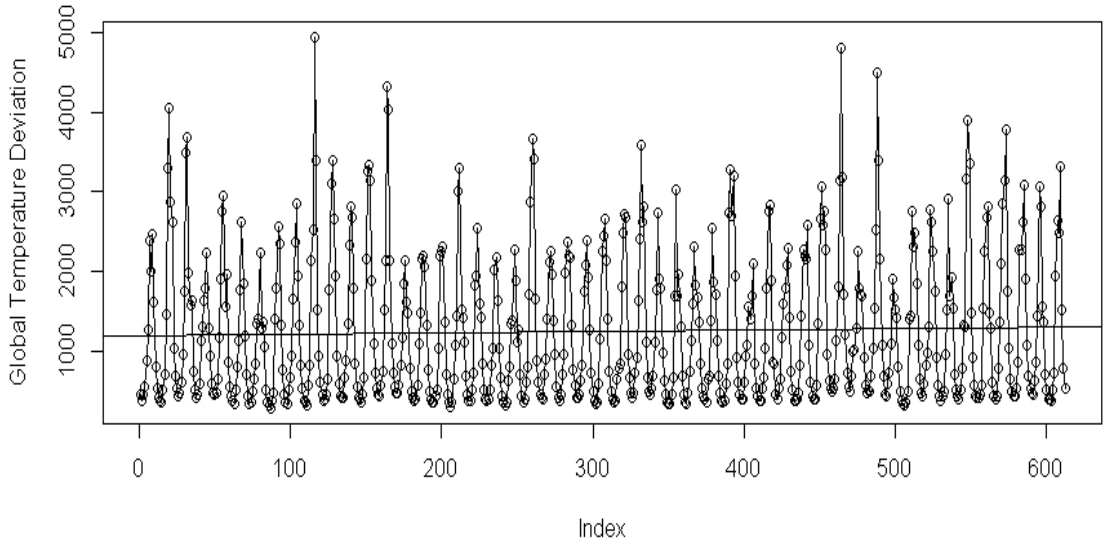

Figure 2 Time Series Plot of Monthly Streamflow with Trendline

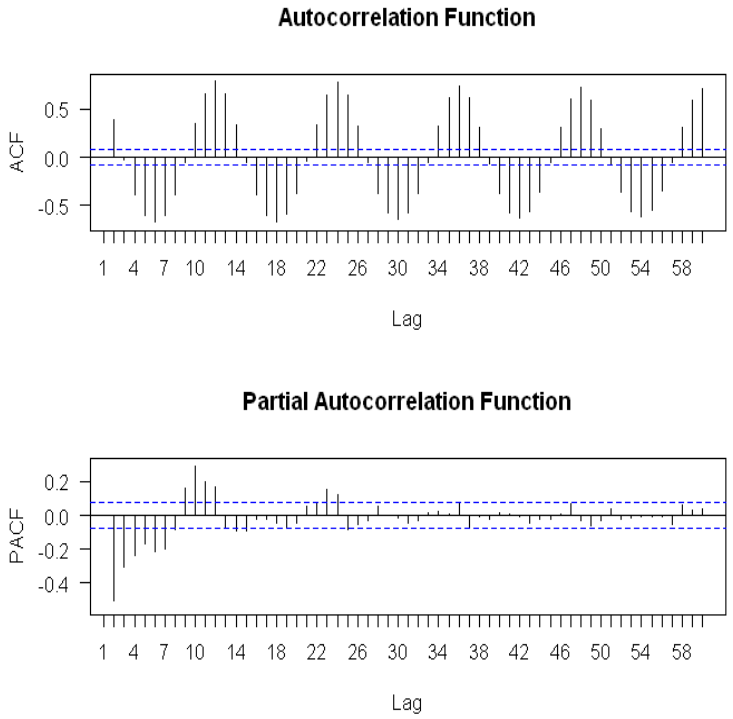

Figure 3 ACF and PAF of Raw Data

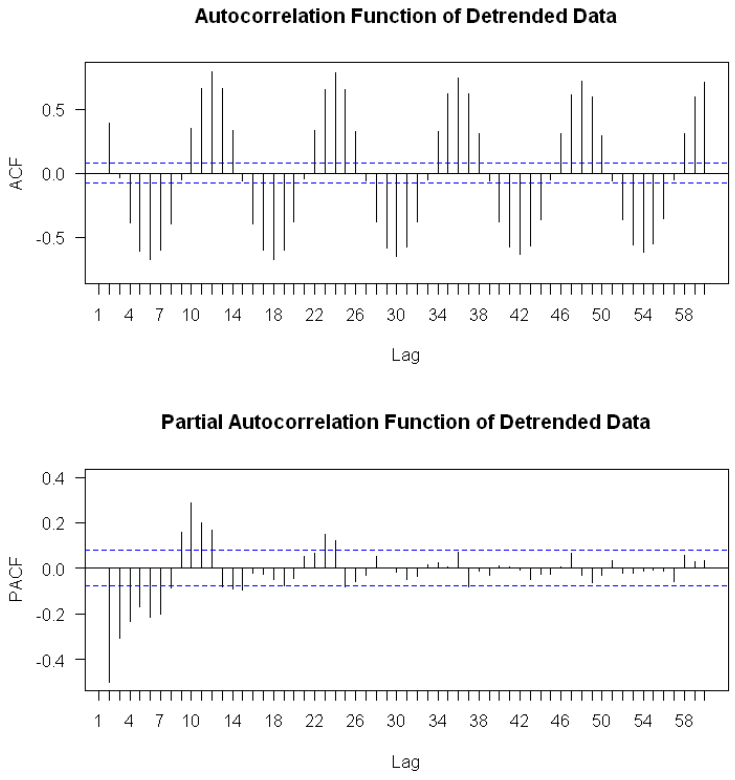

Figure 4 ACF and PACF of Detrended Data

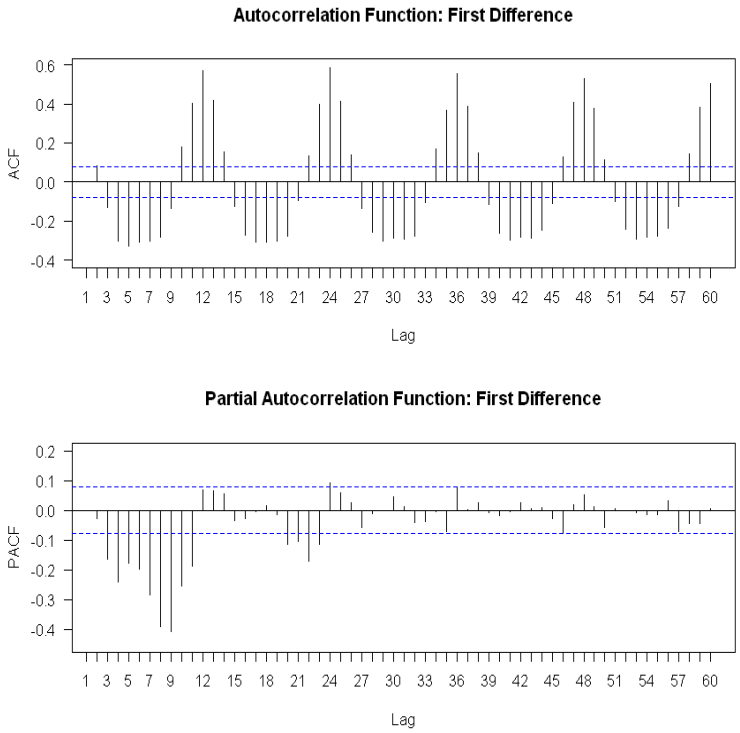

Figure 5 ACF and PACF of the First Difference

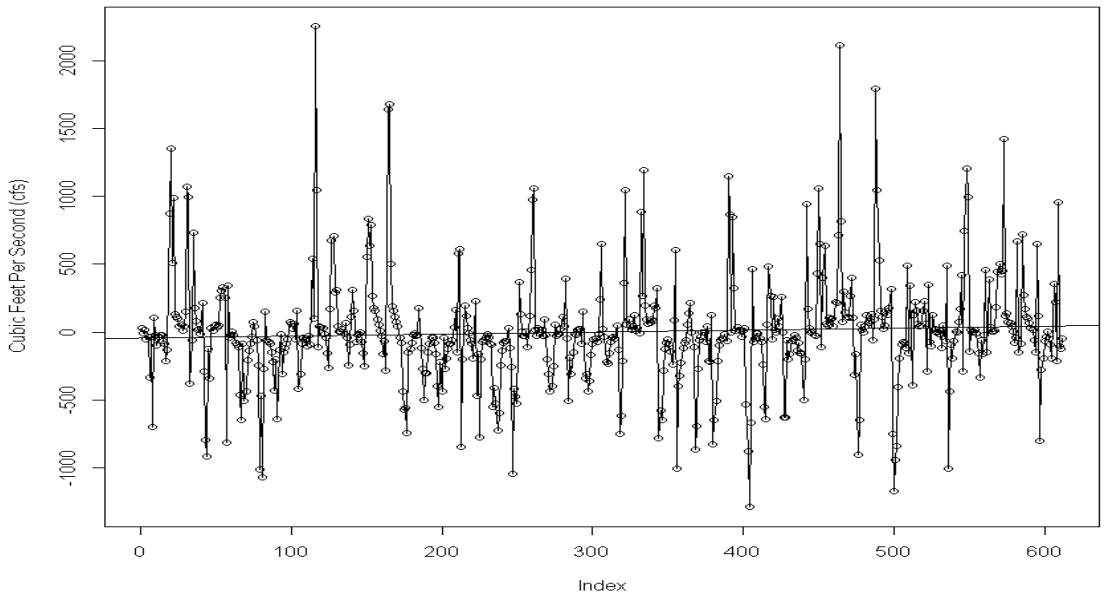

Figure 6 Plot of Data with the Yearly Cycle Removed

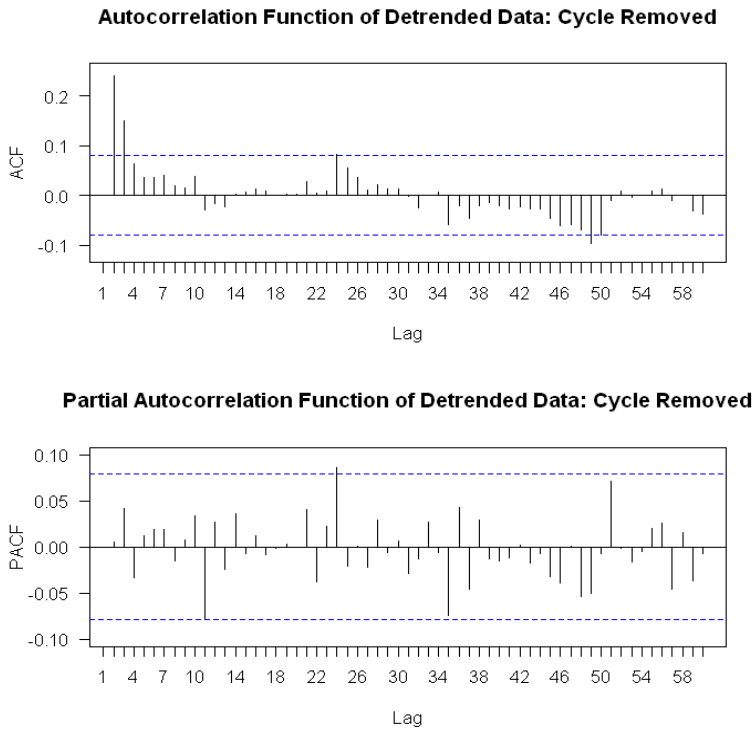

Figure 7 ACF and PACF of Detrended Data with Yearly Cycle Removed

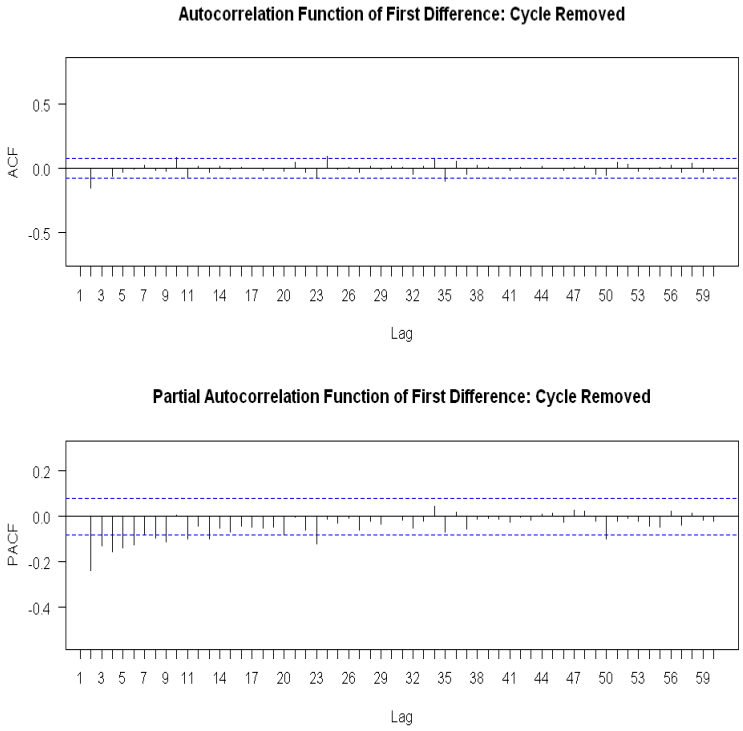

Figure 8 ACF and PACF of First Difference of Data with Yearly Cycle Removed

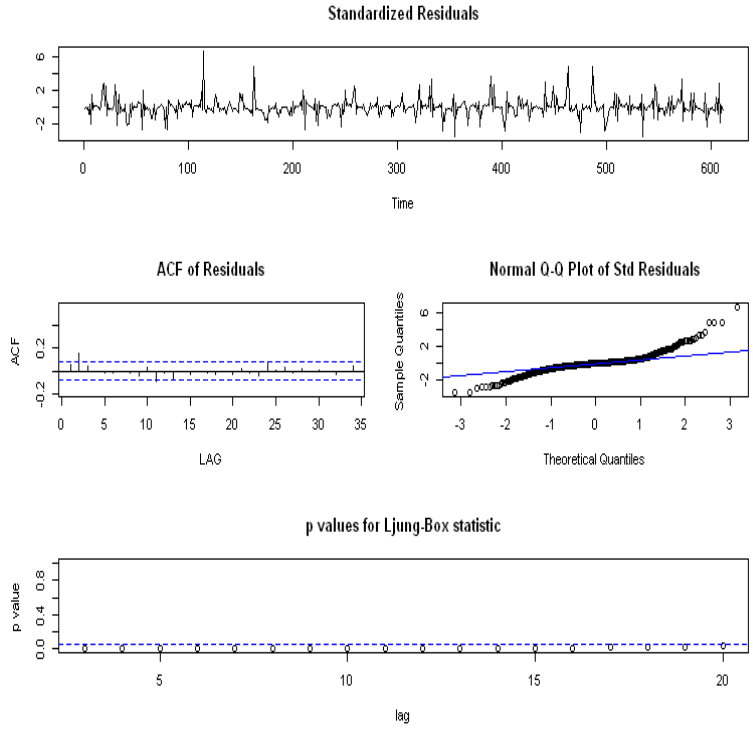

Figure 9 Residual Analysis for ARIMA (0,1,2)

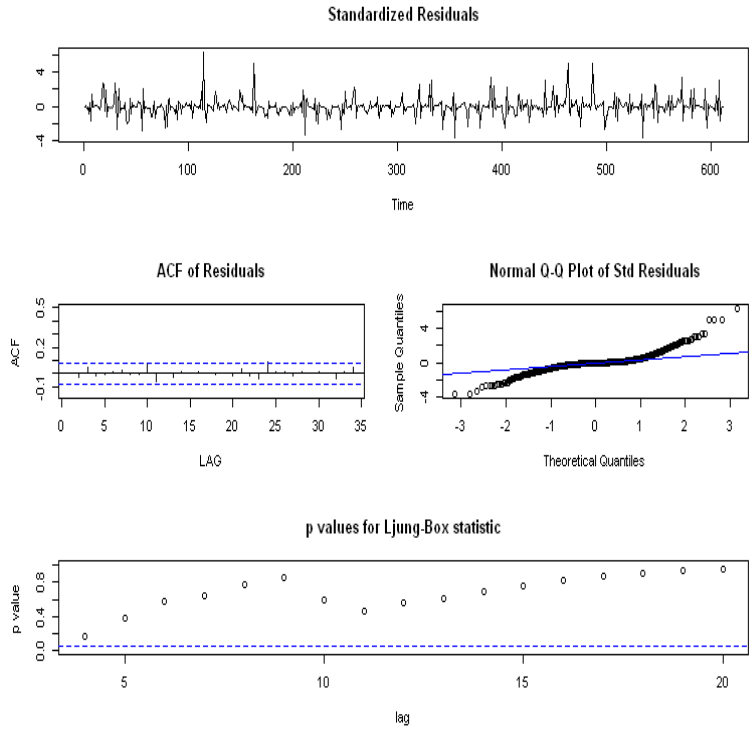

Figure 10 Residual Analysis of ARIMA (1,1,2)

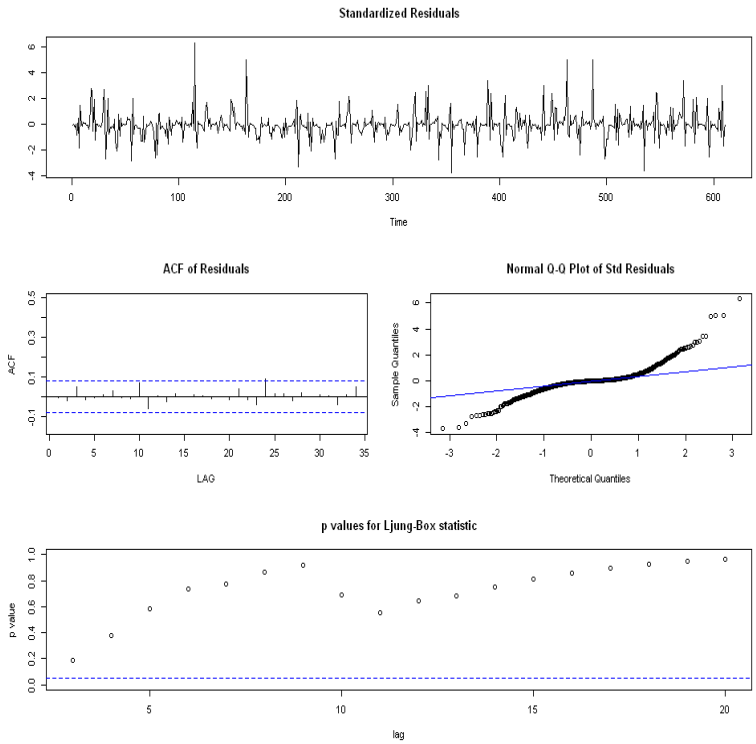

Figure 11 Residual Analysis of ARIMA (1,1,1)

ARMA	AIC	AICc	BIC	aic
(0,1,2)	12.872	12.875	11.886	8989.910
(1,1,2)	12.719	12.723	11.749	8902.120
(1,1,1)	12.716	12.720	11.738	8900.150

Figure 12 Goodness of Fit Measurements for ARIMA models

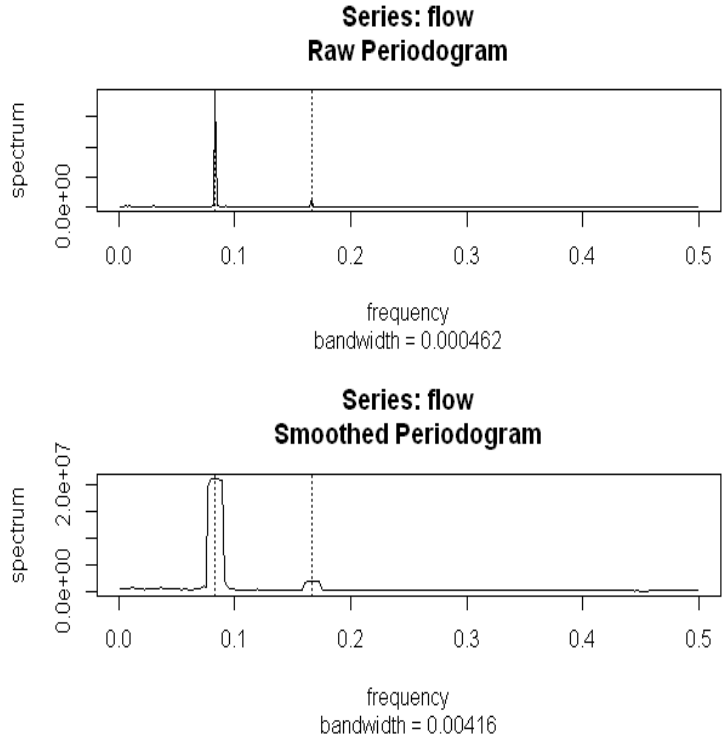

Figure 13 Periodogram of the Raw Data

Type	Frequency	Period	Power	Lower	Upper
Raw Periodogram	(1/12)	1 year	187121954	50725961	7390922404
	(1/6)	6 months	14256467	3864715	563100351
Smoothed Periodogram	(1/12)	1 year	27246286	14099694	73187193
	(1/6)	6 moths	2240606	1159492	6018568

Figure 14 Spectral Density Estimation for Raw Periodogram

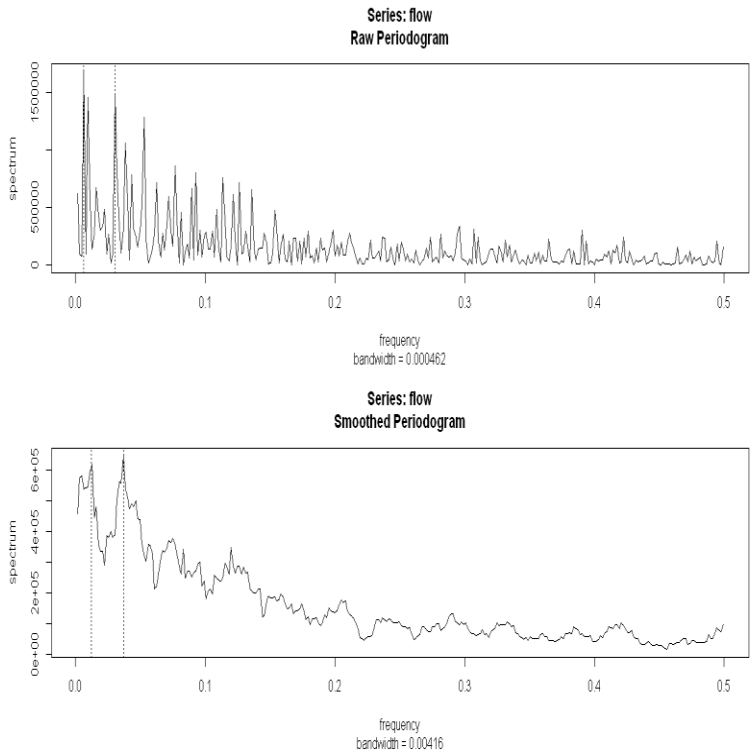

Figure 15 Periodogram of Data with Yearly Cycle Removed

Type	Frequency	Period	Power	Lower	Upper
Raw Periodogram	0.00650	12.8years	1670158	452755	65967737
	0.03040	2 years	1389831	376762.5	54895411
Smoothed Periodogram	0.01240	6.72 years	597436.4	330283.9	1393240
	0.03700	2.25 years	646914.4	1393240	1508625

Figure 16 Spectral Density Estimation of Data with Yearly Cycle Removed

APPENDIX: Contains sample code for creating each of the graphs in the report. Details of the graphs may change for each graph but the general methodology is presented here.

Plotting and Fitting a Regression:

```
library(astsa)
setwd("C:/Users/David/Desktop/R Projects/")
load("tsa3.rda")
flow1<-read.csv("statsprojectdata1cycleremoved.csv")
flow<-flow1[,2]
number<-flow1[,1]

plot(flow, type="o", ylab="flow (cfs)")         #main plot


summary(fit <- lm(flow~number) )                #regress gtemp on time
plot(flow, type="o", ylab="Cubic Feet Per Second (cfs)")
abline(fit)                                     #add regression line to the plot

mod <- lm(flow ~ number); coef(mod)             #There is a very slight increasign trend
mean(diff(gtemp)) (drift)
sd(diff(gtemp))/sqrt(length(diff(gtemp)))       # (SE)-standard error
fit = lm(flow~number , na.action=NULL)          # regress gtemp on time
```

Generate ACF and PACF Trend Removed with cycle:

```
library(astsa)
setwd("C:/Users/David/Desktop/R Projects/")
load("tsa3.rda")
flow1<-read.csv("statsprojectdata1.csv")
flow<-flow1[,2]
number<-flow1[,1]

#fit = lm(flow~number , na.action=NULL)            #regress flow on number(month number)
j<-(resid(fit))                                   #this is the detrended graph

# Create an "acf" object called z and Create an "pacf" object called w
z<- acf(j,60)
w<- pacf(j,60)

par(mfrow=c(2,1))                                 #create space for two graphs
#Plot the autocorrelation function without lag 0
plot(z[2:60],
     type="h",
     main="Autocorrelation Function: Trend Removed",
     xlab="Lag",
     ylab="ACF",
     ylim=c(-0.1,0.25), # this sets the y scale to -0.35 to 0.35
     las=1,
     xaxt="n")
abline(h=0)
# Add labels to the x-axis
x <- c(1:60)
y <- c(1:60)
axis(1, at=x, labels=y)

#Plot the partialautocorrelation function without lag 0

plot(w[2:60],
     type="h",
     main="Partial Autocorrelation Function: Trend Removed",
     xlab="Lag",
     ylab="PACF",
     ylim=c(-0.1,0.1), # this sets the y scale
     las=1,
     xaxt="n")
abline(h=0)
# Add labels to the x-axis
x <- c(1:60)
y <- c(1:60)
axis(1, at=x, labels=y)
```

Code for first difference ACF and PACF:

```
library(astsa)
setwd("C:/Users/David/Desktop/R Projects/")
load("tsa3.rda")
flow1<-read.csv("statsprojectdata1.csv")
flow<-flow1[,2]
number<-flow1[,1]

j<-(diff(flow))                                    #this is the detrended graph

# Create an "acf" object called z and Create an "pacf" object called w
z<- acf(j,60)
w<- pacf(j,60)

par(mfrow=c(2,1))                                  #create space for two graphs
#Plot the autocorrelation function without lag 0
plot(z[2:60],
     type="h",
     main="Autocorrelation Function: First Difference",
     xlab="Lag",
     ylab="ACF",
     ylim=c(-0.4,0.59), # this sets the y scale to -0.35 to 0.35
     las=1,
     xaxt="n")
abline(h=0)
# Add labels to the x-axis
x <- c(1:60)
y <- c(1:60)
axis(1, at=x, labels=y)

#Plot the partialautocorrelation function without lag 0

plot(w[2:60],
     type="h",
     main="Partial Autocorrelation Function: First Difference",
     xlab="Lag",
     ylab="PACF",
     ylim=c(-0.45,0.2), # this sets the y scale
     las=1,
     xaxt="n")
abline(h=0)
# Add labels to the x-axis
x <- c(1:60)
y <- c(1:60)
axis(1, at=x, labels=y)
```

Code for ACF and PACF: detrended, with cycle removed

```
library(astsa)
setwd("C:/Users/David/Desktop/R Projects/")
load("tsa3.rda")
flow1<-read.csv("statsprojectdata1cycleremoved.csv")
flow<-flow1[,2]
number<-flow1[,1]

fit = lm(flow~number , na.action=NULL)          # regress gtemp on time
#flow2 = diff(flow)
z<-(resid(fit))     #this is the detrended graph

# Create an "acf" object called z and Create an "pacf" object called w
z<- acf(z,60)
w<- pacf(z,60)
par(mfrow=c(2,1))

#Plot the autocorrelation function without lag 0
plot(z[2:60],
     type="h",
     main="Autocorrelation Function of Detrended Data: Cycle Removed",
     xlab="Lag",
     ylab="ACF",
     ylim=c(-0.12,0.25), # this sets the y scale to -0.35 to 0.35
     las=1,
     xaxt="n")
abline(h=0)
# Add labels to the x-axis
x <- c(1:60)
y <- c(1:60)
axis(1, at=x, labels=y)

#Plot the partialautocorrelation function without lag 0

plot(w[2:60],
     type="h",
     main="Partial Autocorrelation Function of Detrended Data: Cycle Removed",
     xlab="Lag",
     ylab="PACF",
     ylim=c(-0.1,0.1), # this sets the y scale
     las=1,
     xaxt="n")
abline(h=0)
# Add labels to the x-axis
x <- c(1:60)
y <- c(1:60)
axis(1, at=x, labels=y)
```

Code used for fitting an ARMA model:

```
library(astsa)
library(forecast)
setwd("C:/Users/David/Desktop/R Projects/")
load("tsa3.rda")
flow1<-read.csv("statsprojectdata1cycleremoved.csv")
flow<-flow1[,2]
number<-flow1[,1]

#fit = lm(flow~number , na.action=NULL)
#z<-(resid(fit))     #this is the detrended data

flow2 = diff(flow)  # First difference of data

auto.arima(flow2, stepwise=FALSE, approximation=FALSE)

sarima(flow2, 0, 0, 1)  # ARMA
sarima(flow2, 1, 0, 2)  # ARMA
sarima(flow2, 1, 0, 1)  # ARMA
```

Sample Code for producing a Periodogram and Spectral Density Estimation

```
library(astsa)
setwd("C:/Users/David/Desktop/R Projects/")
load("tsa3.rda")
flow1<-read.csv("statsprojectdata1.csv")
flow<-flow1[,2]
number<-flow1[,1]

par(mfrow=c(2,1))
# to plot the raw perdiogram
rp <- spec.pgram(flow,taper=0,log="no")
abline(v=c(.083333,.1666666), lty="dotted")

nextn(612)#=625

rp$spec[52.0833312] #this is at one year
rp$spec[104.166662] #this is at 6 months

# conf intervals for raw
U = qchisq(.025,2)
L = qchisq(.975,2)
2*rp$spec[52.0833312]/L
2*rp$spec[52.0833312]/U
2*rp$spec[104.166662]/L
2*rp$spec[104.166662]/U


#smoothed periodogram
k = kernel("daniell",4)
sp= spec.pgram(flow, k, log="no")
abline(v=c(.083333,.1666666), lty="dotted")



sp$spec[52.0833312] #this is at one year
sp$spec[104.166662] #this is at 6 months

# conf intervals for smooth
df=sp$df
U = qchisq(.025,df)
L = qchisq(.975,df)
df*sp$spec[52.0833312]/L
df*sp$spec[52.0833312]/U
df*sp$spec[104.166662]/L
df*sp$spec[104.166662]/U
```