W271 Lab 3: Kalvin Kao

The following functions were created for analyses that were repeated throughout this report. *tsplot*() provides the time plot, histogram, ACF plot, and PACF plot of a time series, with options to also include the Augmented Dickey-Fuller test (null hypothesis: series has a unit root), Ljung-Box test (null hypothesis: series is uncorrelated), and/or Shapiro-Wilk test (null hypothesis: sample distribution is normal) results on the series. *search.sarima.params*() fits a range of seasonal ARIMA models according to the parameter ranges specified in the function call, and summarizes their in-sample fit (AIC and BIC), as well as the results of the Ljung-Box and Shapiro-Wilk tests on their residuals. *get.fcst*() calculates a forecast from a time-series model, including its 95% confidence interval limits.

```
tsplot<-function(series,title,use.adf,use.box,use.shap){par(mfrow=c(2,2));plot.ts(series,main="");
  title(paste("Time Plot of",title));hist(series,main="");title(paste("Histogram of",title))
  acf(series,main="");title(paste("ACF of",title));pacf(series,main="");title(paste("PACF of",title))
  adf.check<-ifelse(use.adf==1,yes=print(paste("ADF Test P-Value for",title,":",
                                               adf.test(series)$p.value)),no=0)
  box.check<-ifelse(use.box==1,yes=print(paste("Ljung-Box Test P-Value for",title,":",
                                               Box.test(series, type="Ljung-Box")$p.value)),no=0)
  norm.check<-ifelse(use.shap==1,yes=print(paste("Shapiro-Wilk Test P-Value for",title,":",
                                                 shapiro.test(series)$p.value)),no=0)}
search.sarima.params<-function(time.series,frequency,max.p,d,max.q,max.P,D,max.Q){results<-data.frame()
  for(p in 0:max.p){;for(q in 0:max.q){;for(P in 0:max.P){;for(Q in 0:max.Q){
          mod<-Arima(time.series,order=c(p,d,q),seasonal=list(order=c(P,D,Q),frequency),method = "ML")
          results<-rbind(results,data.frame(p=p,d=d,q=q,P=P,D=D,Q=Q,AIC=mod$aic,BIC=mod$bic,
                                            box.test=Box.test(mod$residuals, type="Ljung-Box")$p.value,
                                            shapiro.test=shapiro.test(mod$residuals)$p.value))
  };};};};return(head(results[order(results$AIC, results$BIC),],5))}
get.fcst<-function(est.mod,n.steps,ci,st.yr,st.mo,freq){fcst<-predict(est.mod,n.ahead=n.steps,ci=ci)
  pred<-ts(fcst$pred,st=c(st.yr,st.mo),fr=freq);pred.uci<-pred+2*fcst$se;pred.lci<-pred-2*fcst$se;
  return(data.frame(pred=pred,u=pred.uci,l=pred.lci));}
```

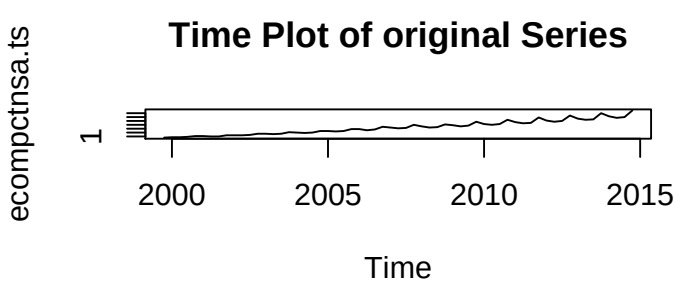
Question 1: Univariate Time Series Analysis and Forecast

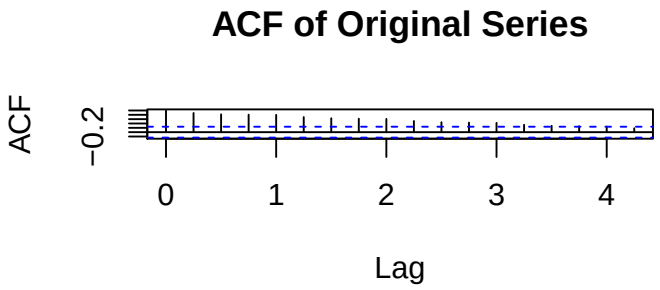
EDTSA: we begin by reading in the data for question 1 and examining its structure.

```
library(tseries); library(vars); library(forecast); #library(GGally); library(car)
ecompctnsa<-read.csv("ECOMPCTNSA.csv",header=TRUE);dim(ecompctnsa);cbind(head(ecompctnsa),tail(ecompctnsa));
## [1] 69  2
##         DATE ECOMPCTNSA       DATE ECOMPCTNSA
## 1 1999-10-01        0.7 2015-07-01        6.8
## 2 2000-01-01        0.8 2015-10-01        8.7
## 3 2000-04-01        0.8 2016-01-01        7.7
## 4 2000-07-01        0.9 2016-04-01        7.5
## 5 2000-10-01        1.1 2016-07-01        7.7
## 6 2001-01-01        1.1 2016-10-01        9.5
```

This dataset contains 69 observations of 2 variables– the second variable, 'ECOMPCTNSA', is the series of interest (quarterly data of E-Commerce Retail Sales as a Percent of Total Sales), and the first variable, 'DATE', provides the date of the observations. The DATE variable shows that this series is quarterly data that begins in Q4 of 1999 and ends in Q4 of 2016. A thorough examination of both variables reveals no missing values and no coding issues or inconsistencies, and we also note that the smallest value of ECOMPCTNSA is 0.7– the data contains no zero values, which is helpful for a later consideration of a log transformation. We therefore continue by respecifying 'ECOMPCTNSA' as a time-series variable– we also create a time series for training that ends in Q4 2014 so that its values in 2015 and 2016 can be used for back-testing (per the assignment instructions).

```
q1.original.ts<-ts(ecompctnsa[,2],start=c(1999,4),frequency=4)
ecompctnsa.ts<-window(q1.original.ts,start=c(1999,4),end=c(2014,4),frequency=4);par(mfrow=c(1,2))
ts.plot(ecompctnsa.ts,main="Time Plot of original Series");acf(ecompctnsa.ts,main="ACF of Original Series")
```

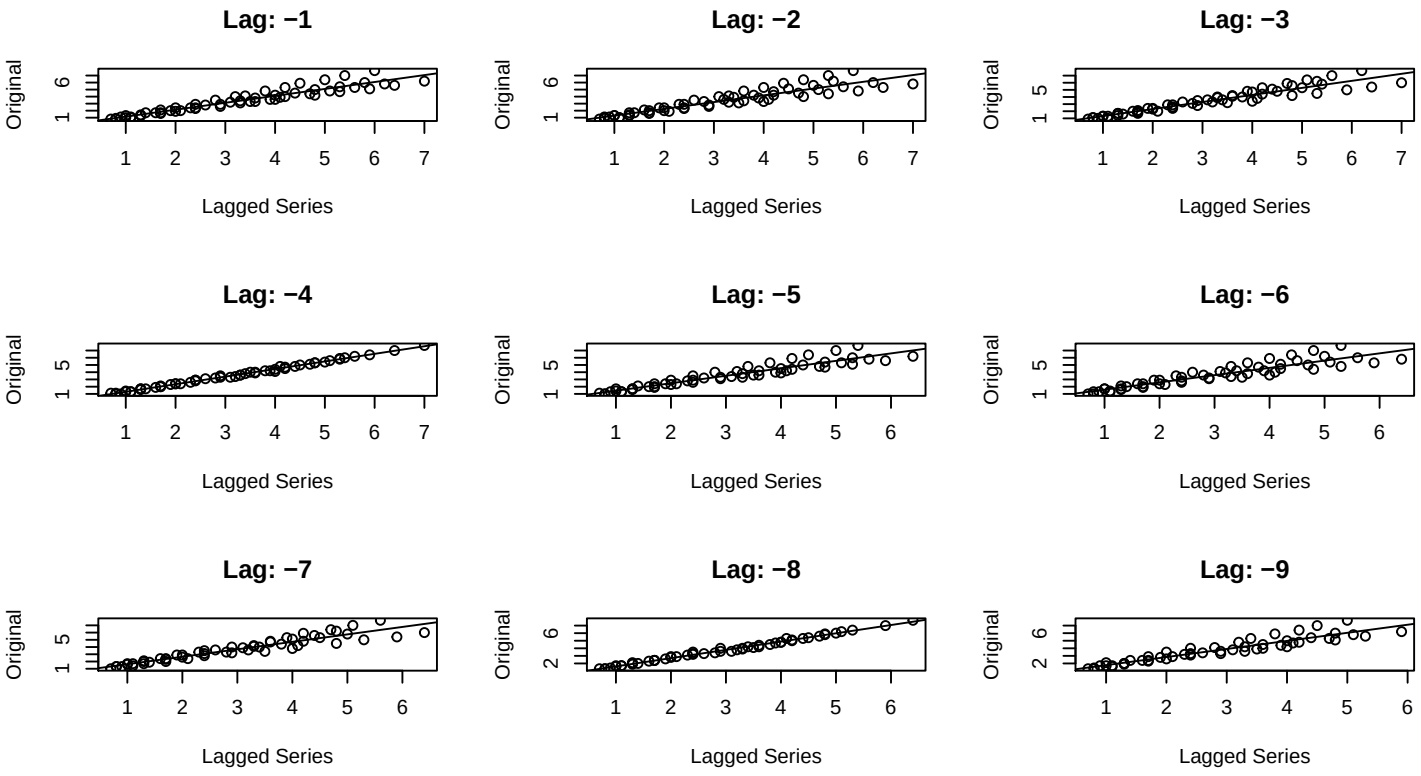





```
print(paste("ADF Test P-Value for Original Series:",adf.test(ecompctnsa.ts)$p.value))
## [1] "ADF Test P-Value for Original Series: 0.99"
```

The plot of the time series clearly shows an increasing trend, seasonality, and increasing volatility over time. The ACF and time series plot together also indicate that the series is not stationary, and the augmented Dickey-Fuller test additionally fails to reject the null hypothesis that this series has a unit root (with the alternative hypothesis that the series is stationary). There is no evidence of extreme values in the series that would require investigation.

```
par(mfrow=c(3,3)); for(lag.level in -1:-9){
  lagged.intersection <- ts.intersect(ecompctnsa.ts, lag(ecompctnsa.ts, lag.level))
  plot(as.vector(lagged.intersection[,2]), as.vector(lagged.intersection[,1]), main="",
       xlab="Lagged Series", ylab="Original"); title(paste("Lag:",lag.level))
  abline(reg=lm(lagged.intersection[,1]~lagged.intersection[,2]))}
```



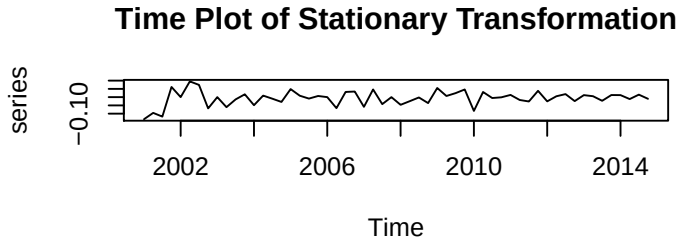
The above plots of the series against its own lags additionally show that the series is highly correlated with its own lags. It also appears as if the correlation is stronger (visually, more clustered around the trend line), at lags 4 and 8, which suggests a seasonal effect in the series.

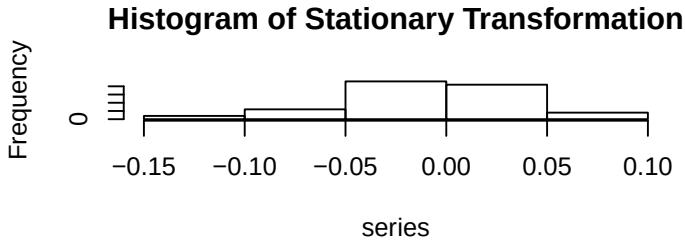
Modeling

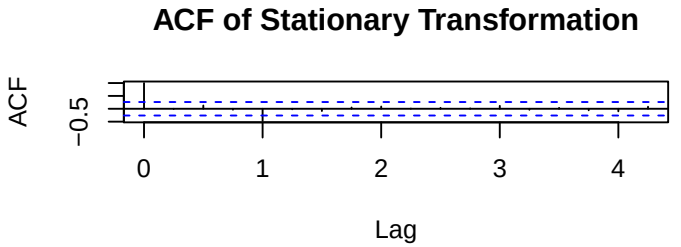
Since I'd like to model the trend in this series and since I do not have insight into the mechanisms that govern the trend, I use differencing to detrend the data, rather than subtracting a trend or curve fit. This also allows the usage of an ARIMA model. Both the original time plot and a time plot of the differenced series (omitted for conciseness) show that its variance is increasing with time. Therefore, prior to differencing, the log transformation of the series is performed in order to stabilize its variance.

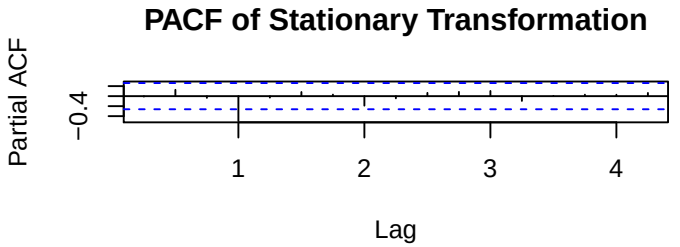
The ACF of the log transformed and first differenced series (again, omitted for conciseness) showed seasonal effects, so a further transformation is performed by taking a seasonal difference. A model which fitted seasonal ARMA components instead of taking a seasonal difference was also investigated and is provided in the Appendix. The following plots are for the series after a log transformation, first difference, and seasonal difference.

```
tsplot(diff(diff(log(ecompctnsa.ts)), lag=4), "Stationary Transformation", 0, 0, 0)
```









The spike in the ACF and PACF at lag=4 suggests a seasonal component, so a parameter search is performed, focusing on the addition of seasonal AR and MA components up to order 2. An MA component in the model is expected, since differencing a linear trend introduces moving average terms into a white noise series.

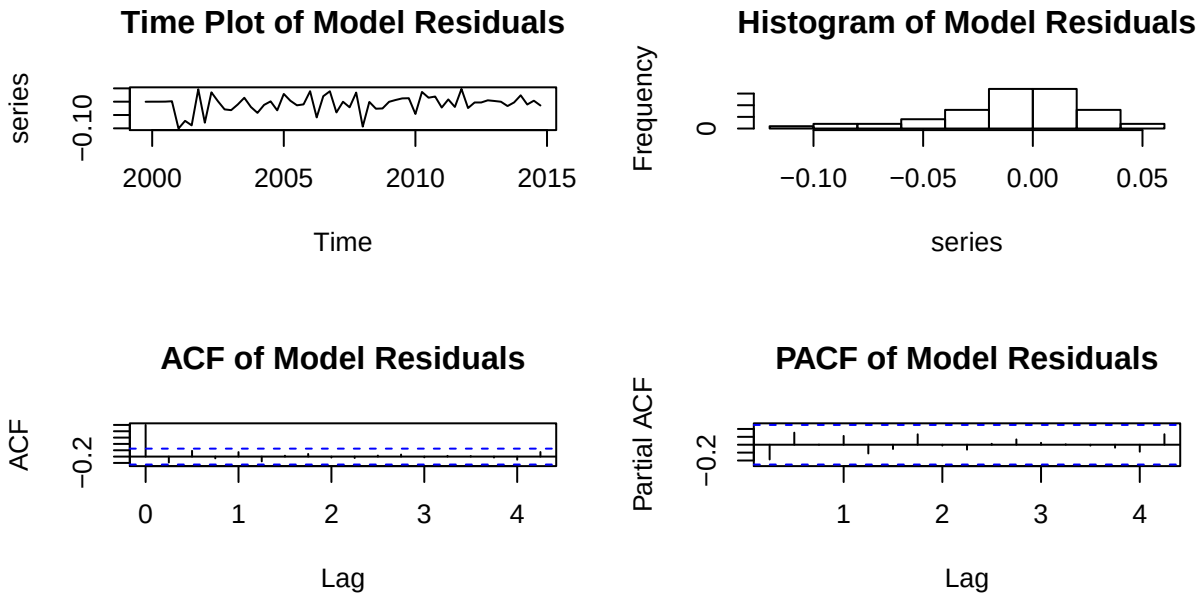
```
search.sarima.params(log(ecompctnsa.ts),4,0,1,0,2,1,2)
##   p d q P D Q       AIC       BIC   box.test shapiro.test
## 3 0 1 0 0 1 2 -207.2204 -201.1444 0.13480565  0.002478424
## 7 0 1 0 2 1 0 -206.4998 -200.4238 0.09472636  0.003334431
## 4 0 1 0 1 1 0 -206.2186 -202.1679 0.12011663  0.002608625
## 5 0 1 0 1 1 1 -205.4989 -199.4228 0.09831768  0.002508739
## 6 0 1 0 1 1 2 -205.4531 -197.3517 0.10025473  0.002033132
```

The summary above shows the top 5 models according to AIC and BIC. All of the Arima(0,1,0)(P,1,Q)_4 models have a low AIC and fail to reject the null hypothesis that the residuals are uncorrelated, when P and Q are each less than 3. The AIC and BIC for the Arima(0,1,0)(0,1,2)_4 model (top model) in particular is amongst the lowest values, and the Ljung-Box test for that model's residuals fails to reject the null hypothesis of no correlation, so this model is selected as a candidate for forecasting. The parameter search also unfortunately shows that none of the top candidates pass the Shapiro-Wilk test for normality (as demonstrated by its significant p-values), which will prevent valid hypothesis tests on its estimates and will introduce some error in the forecast confidence intervals, but the model should still be useful for forecasting. The use two seasonal terms is additionally consistent with the previous EDTSA finding that the series has especially high correlation with its own lags at lags of 4 and 8.

```
q1.mod<-Arima(log(ecompctnsa.ts),order=c(0,1,0),seasonal=list(order=c(0,1,2),4),method="ML");q1.mod
## Series: log(ecompctnsa.ts)
## ARIMA(0,1,0)(0,1,2)[4]
##
## Coefficients:
##          sma1    sma2
##       -0.7714  0.4047
## s.e.   0.1552  0.1343
##
## sigma^2 estimated as 0.001281:  log likelihood=106.61
## AIC=-207.22   AICc=-206.76   BIC=-201.14
```

The estimated Arima(0,1,0)(0,1,2)_4 model has highly statistically significant coefficients, and yields the formula: $x_t = x_{t-1} + x_{t-4} - x_{t-5} + w_t + \Theta_1 w_{t-4} + \Theta_2 w_{t-8}$, where $x_t = log(ecompctnsa.ts)$, Θ_1= -0.77, and Θ_2= 0.4.

```
tsplot(q1.mod$residuals, "Model Residuals", 1, 1, 1)
```

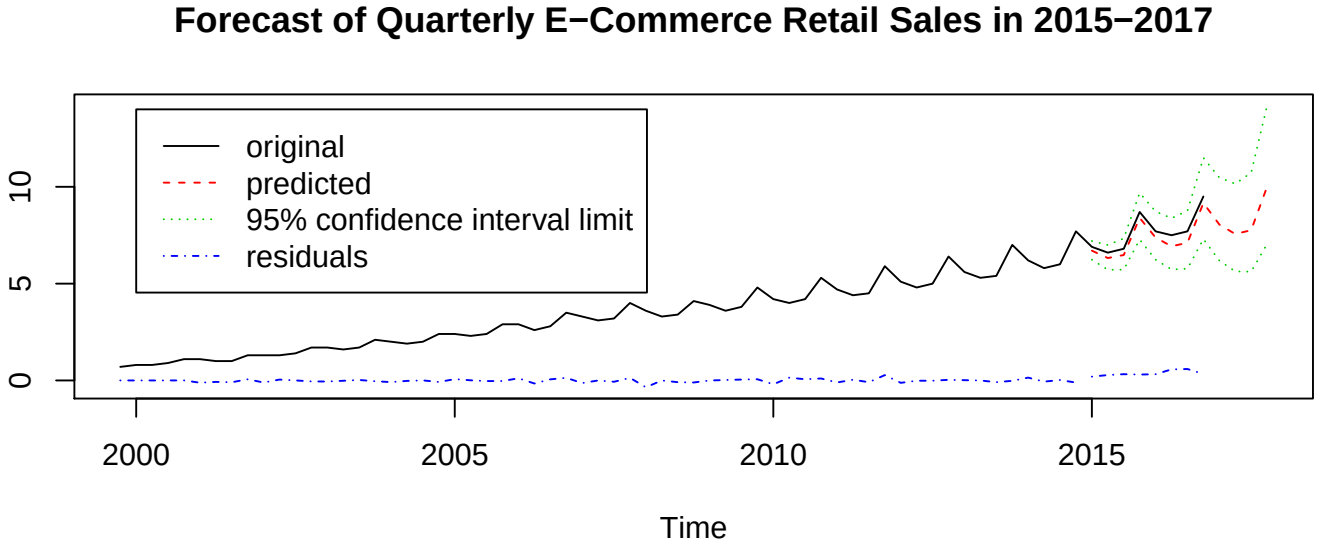



```
## [1] "ADF Test P-Value for Model Residuals : 0.0122645526708789"
## [1] "Ljung-Box Test P-Value for Model Residuals : 0.134805654302829"
## [1] "Shapiro-Wilk Test P-Value for Model Residuals : 0.0024784238726727"
data.frame(roots=abs(polyroot(c(1, q1.mod$coef[1], q1.mod$coef[2]))),
           Theta1.CI=q1.mod$coef[1] + c(-2,2)*sqrt(q1.mod$var.coef)[1],
           Theta2.CI=q1.mod$coef[2] + c(-2,2)*sqrt(q1.mod$var.coef)[4])
##      roots  Theta1.CI Theta2.CI
## 1 1.571891 -1.0818019 0.1360217
## 2 1.571891 -0.4609929 0.6734187
```

The above dataframe shows the 95% confidence intervals of the model parameters, which do not include zero, as well as the roots of the seasonal MA characteristic equation– both lie outside the unit circle, meaning that the model's seasonal MA process is invertible (which in turn means that it can be expressed as a function of previous values), and can therefore be used for forecasting. An MA process is also always stationary. The overall model residuals have several low values in the beginning and middle of the series that are slightly concerning, however, their ACF and PACF do resemble a white noise series. The Augmented Dickey-Fuller and Box-Ljung tests on the residuals additionally reject the null hypothesis of a unit root and fail to reject the null hypothesis of no correlation, respectively. The distribution of the residuals appears approximately normal– its negative skew is likely a consequence of the log transformation of the series, and an empirical adjustment is therefore made prior to forecasting. Although the Shapiro-Wilk test on the residuals strongly rejects the null hypothesis of a normal distribution, the other evidence indicates that the residuals are uncorrelated and could still be a realization of a white noise process– the model is next used for forecasting with the understanding that the non-normality of the residuals may introduce some error.

Forecasting

```
q1.pred<-get.fcst(q1.mod,12,0.95,2015,1,4)
ts.plot(cbind(q1.original.ts,exp(q1.pred$pred),exp(q1.pred$u),exp(q1.pred$l),
              q1.original.ts-exp(ts.union(q1.mod$fitted, q1.pred$pred))),lty=c(1,2,3,3,4,4),
        col=c(1,2,3,3,4,4),main="Forecast of Quarterly E-Commerce Retail Sales in 2015-2017")
legend(x=2000,y=14,legend=c("original","predicted","95% confidence interval limit","residuals"),
       lty=c(1,2,3,4),col=c(1,2,3,4))
```



The above plot shows the estimated model's forecasts for 2015 - 2017 (red), after fitting to the series (black). The model residuals are shown in blue, with points after 2014 showing only a small error in back-testing (2015/2016 RMSE: 0.39)– the actual observations in 2015 and 2016 are within the 95% confidence bounds of the forecast (green), so this model is able to forecast well. The forecast according to the model shows the trend and seasonality continuing to increase, however, its 95% confidence interval also shows that a decrease in the trend cannot be ruled out.

Many other modeling approaches were investigated, and the notable alternatives are included in the Appendix. One alternative fits a seasonal ARMA on the log transformed, first differenced series (i.e. ARIMA(0,1,0)(1,0,2)[4])– this approach resulted in slightly better in-sample fit, but also less parsimony and evidence of correlation in the model residuals. Another modeling approach fits a seasonal ARMA on the seasonally differenced series (i.e. ARIMA(2,0,0)(2,1,0)[4])– such an approach satisfied all of the ARIMA model assumptions (including normality of its residuals), however, its in-sample fit and back-testing results were significantly worse and the resulting models were thus unfavorable for forecasting.

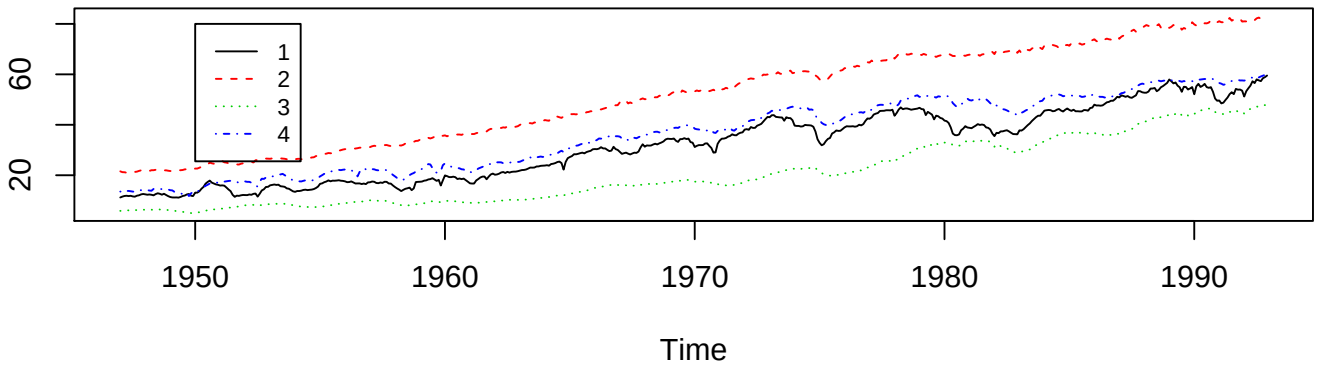
Question 2: Multivariate Time Series Analysis and Forecast

EDTSA

```
varData<-read.table("data_2018Spring_MTS_v2.txt",header=TRUE); str(varData); #summary(varData)
## 'data.frame':    564 obs. of  6 variables:
##  $ year   : int  1947 1947 1947 1947 1947 1947 1947 1947 1947 1947 ...
##  $ mon    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ series1: num  11.2 11.5 11.8 11.9 11.7 ...
##  $ series2: num  21.6 21.2 21.2 21 20.9 ...
##  $ series3: num  5.87 5.94 5.97 6.05 6.08 ...
##  $ series4: num  13.5 13.6 14.1 13.7 13.8 ...
```

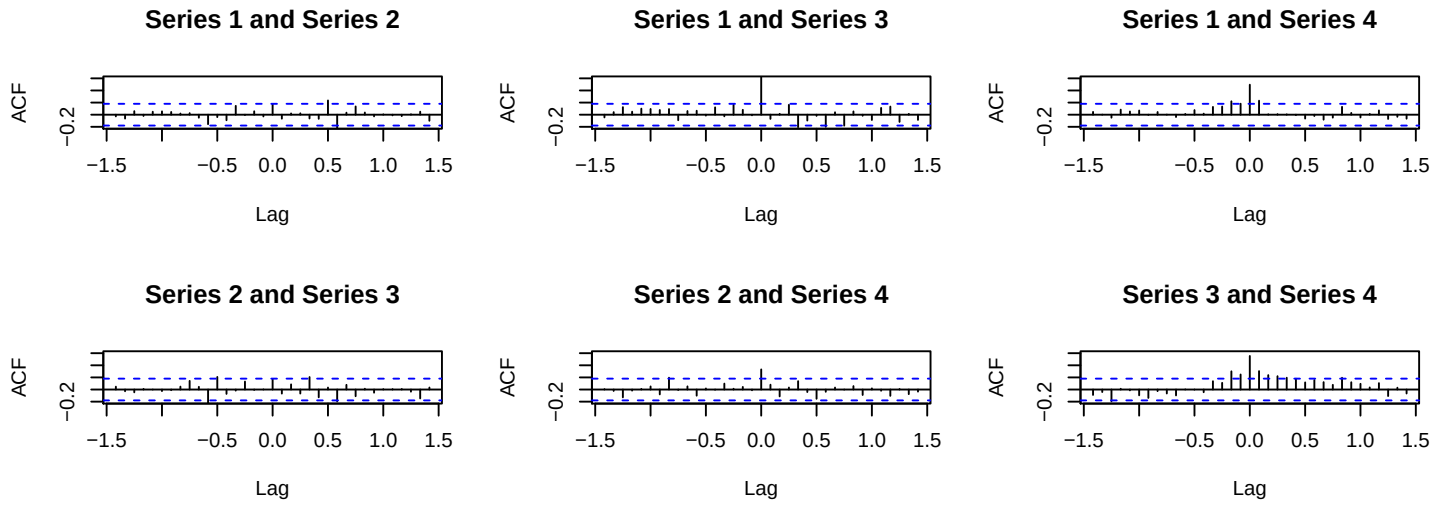
The dataset contains 6 variables and 792 observations. Variables year and mon indicate the time period of each observation– no skipped or missing values were found. The code that was used for this integrity check is omitted for conciseness and is instead available in the Appendix. The remaining 4 variables are 4 monthly time series that begin in January 1947 and end in December 2012– their original data format is that of a continuous numerical variable so they are re-specified as time-series variables, and the data following 1992 is excluded per the assignment instructions. The following time plot of those series also shows that all values are above zero (which is important for a later consideration of a log transform), and do not show potential outliers.

```
q2Data<-ts(varData[varData$year<1993,3:6],start=c(1947,1),frequency=12)
q2Test<-ts(varData[varData$year>1992,3:6],st=c(1993,1),fr=12)
ts.plot(q2Data,main="Time Plot of Original 4 Series",lty=c(1:4),col=c(1:4))
legend(x=1950,y=80,legend=c(1:4),lty=c(1:4),col=c(1:4), cex=0.75)
```



All 4 series are highly persistent and are generally increasing, with series 1, 3, and 4 appearing to follow the same trends– series 3 and 4 in particular appear to be very closely related. These series are very long, with changes in their trends occurring multiple times. The goal is to produce a forecast, so only data beginning in January 1983 are selected for modeling, since this more recent data seems to be more representative of the increasing trend towards the end of the original data. Other models which include the full dynamics of the entire original series were also investigated and their results are included in the Appendix. The ACF plots for the shortened series (beginning in January 1983) show again that the series are highly persistent (they are omitted here for conciseness and are instead available in the Appendix). It is known that two series with trends can easily have a high correlation (also known as spurious correlation), so a first difference is taken on each series before examining their cross-correlations.

```
series1<-window(q2Data[,1],start=c(1983,1),frequency=12)
series2<-window(q2Data[,2],start=c(1983,1),frequency=12)
series3<-window(q2Data[,3],start=c(1983,1),frequency=12)
series4<-window(q2Data[,4],start=c(1983,1),frequency=12)
allseries<-cbind(series1,series2,series3,series4)
par(mfrow=c(2,3));for(a in 1:4){print(paste("ADF Test P-Value for Series",a,"(after differencing):",
  adf.test(diff(allseries[,a]))$p.value));for(b in 1:4){if(a!=b & b>a){
      ccf(diff(allseries[,a]),diff(allseries[,b]),ylim=c(-0.2,0.6),
          main=paste("Series",a,"and Series",b));};};}
## [1] "ADF Test P-Value for Series 1 (after differencing): 0.01"
## [1] "ADF Test P-Value for Series 2 (after differencing): 0.01"
## [1] "ADF Test P-Value for Series 3 (after differencing): 0.0338580860460096"
```



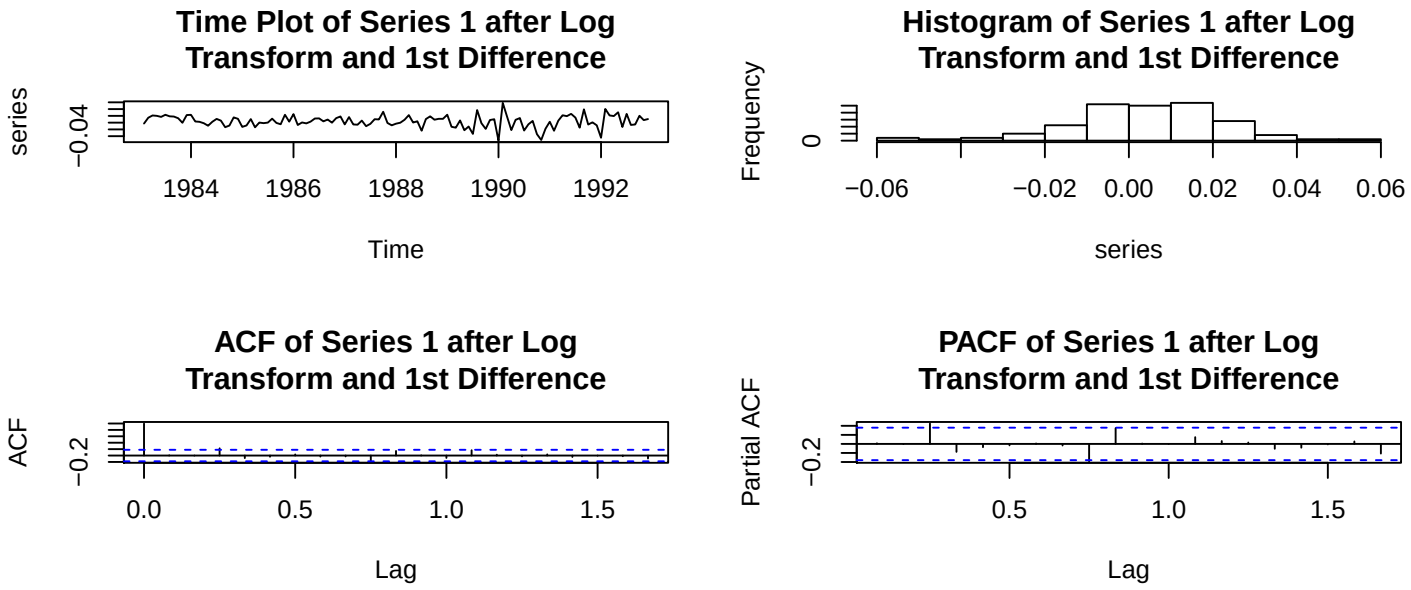

```
## [1] "ADF Test P-Value for Series 4 (after differencing): 0.0537714870857964"
```

After first differencing, the Augmented Dickey-Fuller test (calculated above) is able to reject the null hypothesis of a unit root for each of the 4 series, and the cross-correlation between each differenced series can thus be checked for relationships that are

more likely to be real (or which can at least help explain simultaneous trends). The CCF plots above are for each combination of differenced series after January 1983– there appears to be a small correlation between series 4 differences and past differences of series 1, but it is clear however that series 4 differences are correlated with both lagged *and* future differences of series 3. This also means that series 3 differences are correlated with lagged values of series 4 differences– this relationship between series 3 and 4 is important because it allows the usage of a vector autoregressive (VAR) model in which each series in the model is regressed on lagged values of all other series in the model. Therefore, the approach for simultaneously forecasting these 4 series is to use univariate ARIMA models for series 1 and 2, and a VAR model for series 3 and 4.

Modeling

```
tsplot(diff(log(series1)), "Series 1 after Log\nTransform and 1st Difference", 0, 0, 0)
```



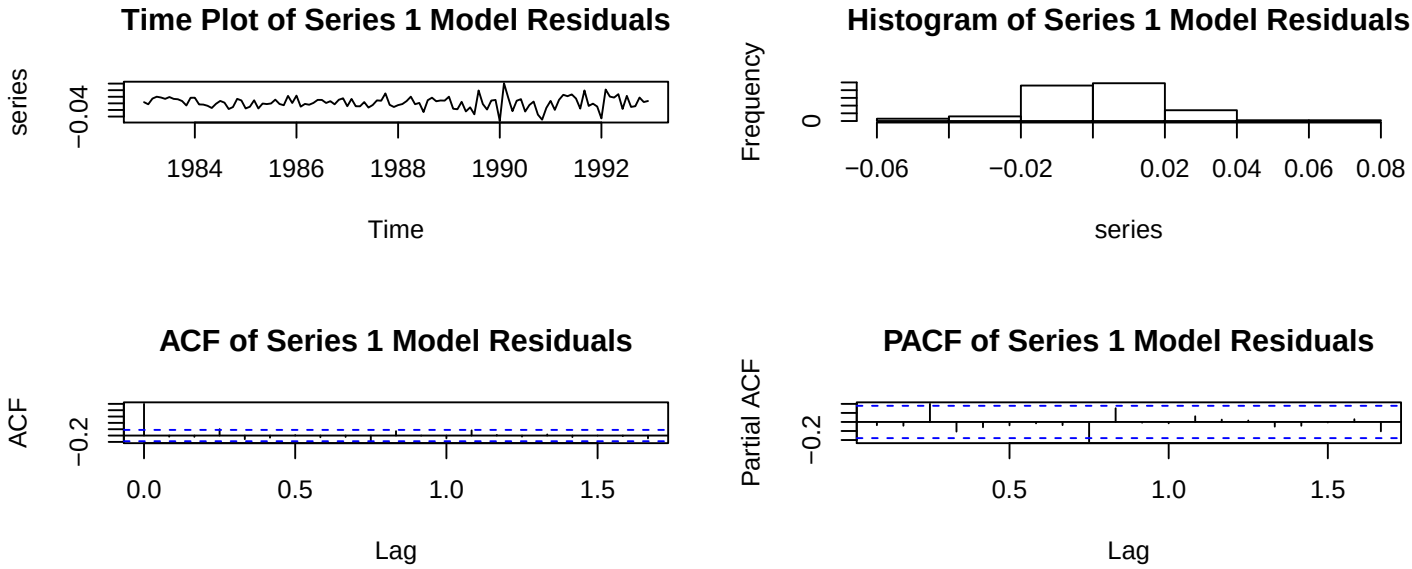
Prior to modeling Series 1, a first difference is taken to remove the trend. The plot of the differenced series (included in the Appendix), as well as subsequent model residuals, showed evidence of increasing variance, so a log transform is also taken in order to stabilize the variance. The above time plot, histogram, ACF plot, and PACF plot are for series 1 after a log-transformation and first difference. Other than a small spike in both the ACF and PACF at a lag of 3, the plots show no clear patterns, and a parameter search is therefore focused low orders of both seasonal and non-seasonal ARMA terms.

```
search.sarima.params(log(series1),12,2,1,2,1,0,1)
##    p d q P D Q       AIC       BIC  box.test shapiro.test
## 1  0 1 0 0 0 0 -615.2433 -612.4642 0.9345407   0.01946923
## 3  0 1 0 1 0 0 -613.6172 -608.0590 0.8517899   0.01601600
## 17 1 1 1 0 0 0 -613.6007 -605.2633 0.6919270   0.05551553
## 2  0 1 0 0 0 1 -613.5871 -608.0288 0.8624988   0.01567172
## 13 1 1 0 0 0 0 -613.5536 -607.9953 0.6317203   0.01802902
```

An ARIMA(0,1,0)(0,0,0)_12 model (a random walk for the log transformation of series 1) is the best candidate in terms of both AIC and BIC. The Ljung-Box test on the residuals of all top models fails to reject the null hypothesis of no correlation, but all of the residuals also show evidence of non-normality (via the Shapiro-Wilk test). However, the ARIMA(1,1,1)(0,0,0)[12] model has competitive AIC and BIC, and also fails to reject the null hypothesis of normality in its residuals (at the 5% level), so this model is selected as a candidate for forecasting. While hypothesis testing with its estimates would be questionable due to the low p-value of the Shapiro-Wilk test on its residuals, the model should still be useful for forecasting.

```
q2.mod.1 <- Arima(log(series1),order=c(1,1,1),seasonal=list(order=c(0,0,0),12),method="ML");q2.mod.1
## Series: log(series1)
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1      ma1
##       0.9462  -0.8923
## s.e.  0.4201   0.5619
##
## sigma^2 estimated as 0.0003261:  log likelihood=309.8
```

```
## AIC=-613.6   AICc=-613.39   BIC=-605.26
tsplot(q2.mod.1$residuals, "Series 1 Model Residuals", 1, 1, 1)
```

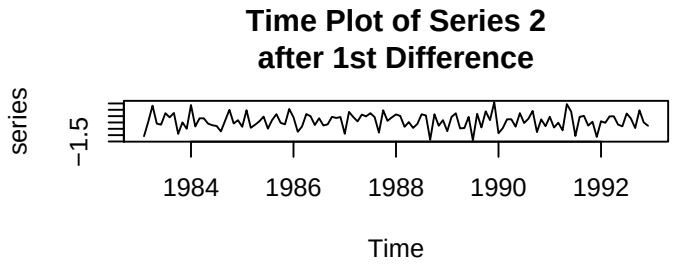


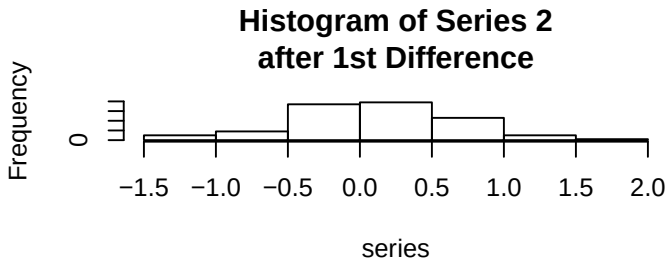

```
## [1] "ADF Test P-Value for Series 1 Model Residuals : 0.01"
## [1] "Ljung-Box Test P-Value for Series 1 Model Residuals : 0.691926990762174"
## [1] "Shapiro-Wilk Test P-Value for Series 1 Model Residuals : 0.055515525832056"
data.frame(phi1.CI=q2.mod.1$coef[1] + c(-2,2)*sqrt(q2.mod.1$var.coef)[1],
           theta1.CI=q2.mod.1$coef[2] + c(-2,2)*sqrt(q2.mod.1$var.coef)[4])
##     phi1.CI  theta1.CI
## 1 0.1060669 -2.0162111
## 2 1.7863118  0.2315617
```

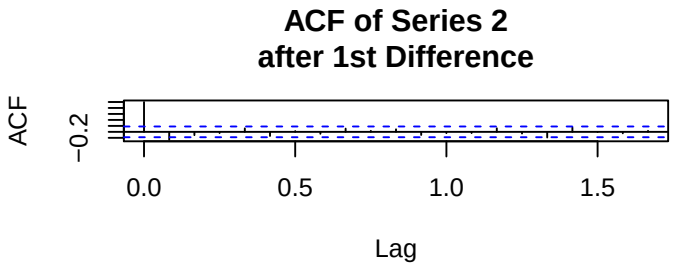
The series 1 model summary above shows that the AR term is statistically significant– its 95% confidence interval, shown in the dataframe above, does not include zero. The MA term, however, has only borderline significance. Since the coefficient on the AR term is 0.95, the characteristic equation for the AR process is (1-0.95B), which has a root > 1 and is therefore stationary. The coefficient on the MA component is -0.89, so its characteristic equation is (1-0.89B), which has a root > 1 and is therefore invertible. Note again that MA processes are always stationary. The time plot, histogram, ACF plot, and PACF plot of the residuals resemble a realization of a white noise process, and as mentioned previously, the Ljung-Box and Shapiro-Wilk tests on the residuals fail to reject the null hypotheses of no correlation and normality (at the 5% level), respectively. The Augmented Dickey-Fuller test on the residuals also rejects the null hypothesis of a unit root. Since the residuals resemble a white noise process, and since the stationarity and invertibility conditions are satisfied, this model will be used for forecasting.

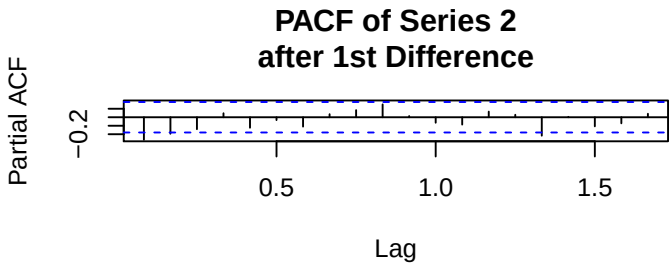
Prior to modeling Series 2, a first difference is taken to remove the trend. The following time plot, histogram, ACF plot, and PACF plot are for series 2 after a first difference.

```
tsplot(diff(series2), "Series 2\nafter 1st Difference", 1, 1, 1)
```









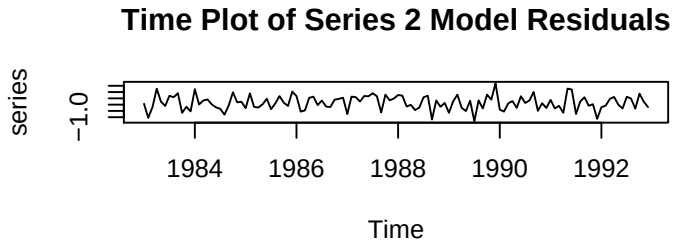

```
## [1] "ADF Test P-Value for Series 2\nafter 1st Difference : 0.01"
## [1] "Ljung-Box Test P-Value for Series 2\nafter 1st Difference : 0.00333933569748068"
## [1] "Shapiro-Wilk Test P-Value for Series 2\nafter 1st Difference : 0.780386636031753"
```

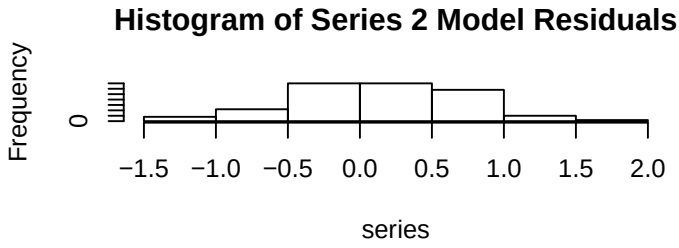
Both the ACF and PACF graphs show negative correlations up to lag 3, so the following parameter search focuses on the addition of non-seasonal AR and MA components up to order 3. The spike at lag 16 is noted, but no action is taken against it since 5% of lags are expected to be significant at the 95% confidence level.

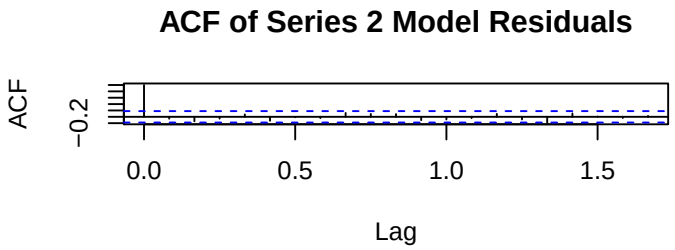
```
search.sarima.params(series2,12,3,1,3,0,0,0)
##   p d q P D Q      AIC      BIC  box.test shapiro.test
## 8 1 1 3 0 0 0 212.1932 226.0888 0.9766180    0.9789040
## 2 0 1 1 0 0 0 214.7784 220.3367 0.5112958    0.8658214
## 5 1 1 0 0 0 0 216.1249 221.6831 0.3455567    0.8292628
## 9 2 1 0 0 0 0 216.2715 224.6089 0.3783339    0.9233949
## 3 0 1 2 0 0 0 216.5759 224.9132 0.3754386    0.9169607
```

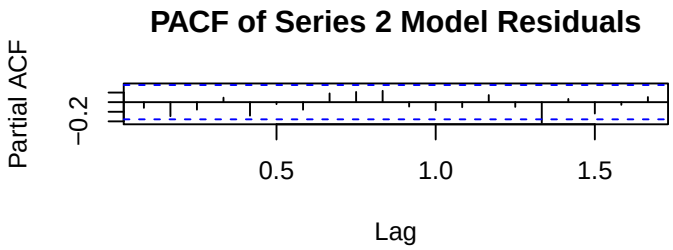
This parameter search yields multiple SARIMA models with comparably low AIC and BIC, and no evidence of correlation or non-normality in their residuals, according to the Ljung-Box and Shapiro-Wilk tests. The ARIMA(0,1,1)(0,0,0)[12] model is selected because it is more parsimonious than the other top models while having a comparable in-sample fit– in fact, its BIC is the lowest in the model search.

```
q2.mod.2<-Arima(series2,order=c(0,1,1),seasonal=list(order=c(0,0,0),12),method="ML");q2.mod.2
## Series: series2
## ARIMA(0,1,1)
##
## Coefficients:
##           ma1
##       -0.2658
## s.e.   0.0903
##
## sigma^2 estimated as 0.3469:  log likelihood=-105.39
## AIC=214.78   AICc=214.88   BIC=220.34
tsplot(q2.mod.2$residuals, "Series 2 Model Residuals", 1, 1, 1)
```










```
## [1] "ADF Test P-Value for Series 2 Model Residuals : 0.01"
## [1] "Ljung-Box Test P-Value for Series 2 Model Residuals : 0.51129576210539"
## [1] "Shapiro-Wilk Test P-Value for Series 2 Model Residuals : 0.865821359953802"
print(paste("95% Confidence Interval for the MA Parameter: (",q2.mod.2$coef[1]-
              (2*sqrt(q2.mod.2$var.coef)[1]),",",q2.mod.2$coef[1]+(2*sqrt(q2.mod.2$var.coef)[1]),")"))
## [1] "95% Confidence Interval for the MA Parameter: ( -0.446362500441944 , -0.0851659873674173 )"
```

The series 2 model summary shows that the MA term coefficient is significant– its 95% confidence interval, also calculated above, does not include zero. Since the coefficient on the MA term is -0.27, the characteristic equation for the MA process is (1-0.27B), which has a root > 1 and the MA component is therefore invertible. The time plot, histogram, ACF plot, and PACF plot of the residuals resemble a realization of a white noise process, and as mentioned previously, the Ljung-Box and Shapiro-Wilk tests on the model residuals fail to reject the null hypotheses of no correlation and normality, respectively. The Augmented Dickey-Fuller test on the residuals additionally rejects the null hypothesis of a unit root. Since the residuals resemble a white noise process, and since the stationarity and invertibility conditions are satisfied, this model will be used for forecasting.

Next, series 3 and 4 are modeled with a VAR model. The EDTSA indicated growing variance in series 3 (see Appendix), so a log transform on that series is first performed in order to stabilize the variance. The VARselect function is used to identify an optimal VAR model order. Both a constant and a trend are included in the search because it was noted earlier that all series have both a trend and a non-zero mean.

```
VARselect(cbind(log(series3),series4), lag.max=4, type="both")$selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      3      3      1      3
```

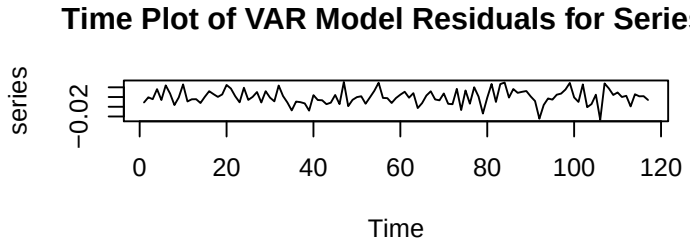
The information criteria in the VARselect function indicate that a VAR(1) or a VAR(3) would be appropriate (these values are omitted here for conciseness). Since there is an interest in forecasting, a more parsimonious VAR(1) model was first fitted– this model showed good performance, however, a Portmanteau test for this model showed marginal evidence of autocorrelation in the residuals. Therefore, a VAR(3) model is selected to model series 3 and 4, which in addition to having good performance also fails to reject the null hypothesis of no autocorrelation in its residuals (via the Portmanteau test).

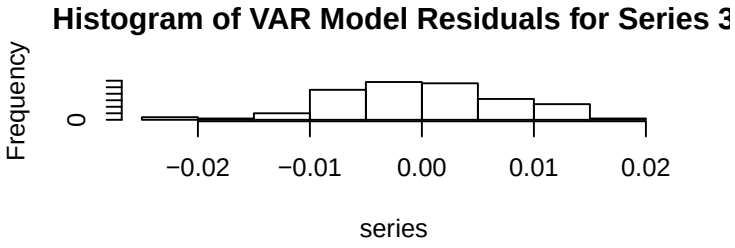
```
var.fit<-VAR(cbind(log(series3),series4),p=3,type="both");summary(var.fit)$varresult$log.series3.$coefficients
##                      Estimate   Std. Error    t value     Pr(>|t|)
## log.series3..l1  0.6240720411 0.1121914972  5.5625609 1.917095e-07
## series4.l1       0.0098907139 0.0031863721  3.1040674 2.432134e-03
## log.series3..l2  0.2807721992 0.1237256892  2.2693121 2.521753e-02
## series4.l2      -0.0054519384 0.0043615449 -1.2500017 2.139761e-01
## log.series3..l3 -0.1438910254 0.1054032010 -1.3651485 1.750174e-01
## series4.l3       0.0015502849 0.0032817249  0.4723994 6.375862e-01
## const            0.5456815912 0.1293627581  4.2182279 5.108078e-05
## trend            0.0002181664 0.0000833413  2.6177470 1.010917e-02
summary(var.fit)$varresult$series4$coefficients;round(roots(var.fit),2)
##                     Estimate  Std. Error    t value     Pr(>|t|)
```

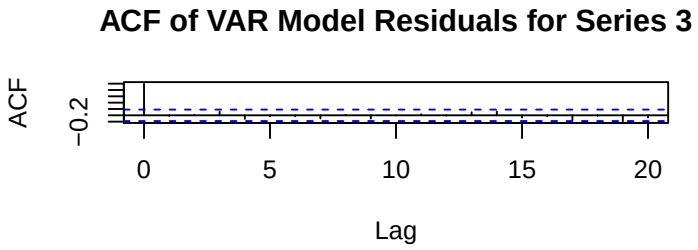
```
## log.series3..l1   5.58471808 3.789179865  1.4738593 1.434017e-01
## series4.l1        1.02349604 0.107617220  9.5105230 5.599265e-16
## log.series3..l2   2.64408563 4.178738157  0.6327474 5.282238e-01
## series4.l2       -0.13133609 0.147307760 -0.8915762 3.745841e-01
## log.series3..l3 -12.62952083 3.559910480 -3.5477074 5.744346e-04
## series4.l3        0.14274571 0.110837687  1.2878807 2.005150e-01
## const            13.67413036 4.369125740  3.1297177 2.245594e-03
## trend             0.01136759 0.002814787  4.0385257 1.003428e-04
## [1] 0.96 0.96 0.28 0.28 0.21 0.06
```

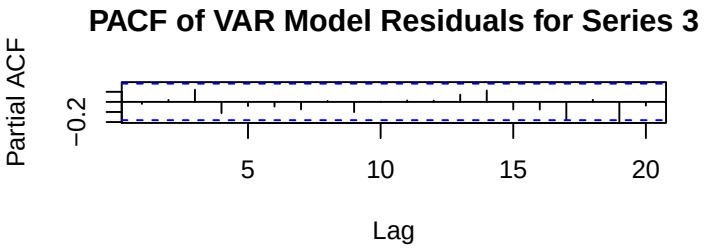
The roots() function above calculates the eigenvalues of the model's companion matrix– since these are less than 1 in this case, this model is considered to be stable. The model estimates are also displayed above. For the log transform of series 3, its first 2 lags, as well as the first lag of series 4, have significant p-values. For series 4, its own first lag, as well as the 3rd lag of the log transform of series 3, have significant p-values. This is an interesting result and it would be interesting to further study whether series 3 has a causal effect on series 4. The constant and trend terms for both series also have significant p-values.

```
var.fit.residuals<-resid(var.fit);par(mfrow=c(2,1));for(a in 1:2){
  tsplot(var.fit.residuals[,a],paste("VAR Model Residuals for Series",a+2),1,0,0)}
```

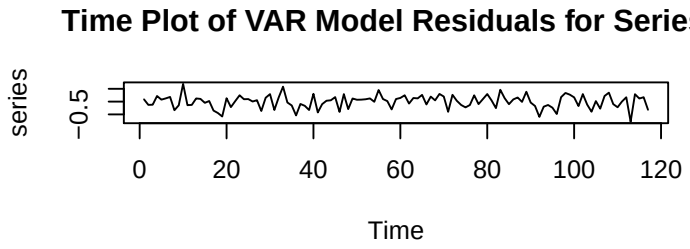


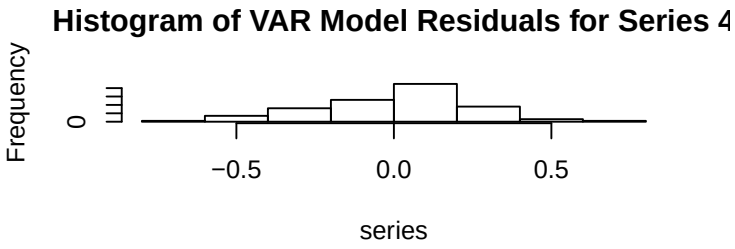


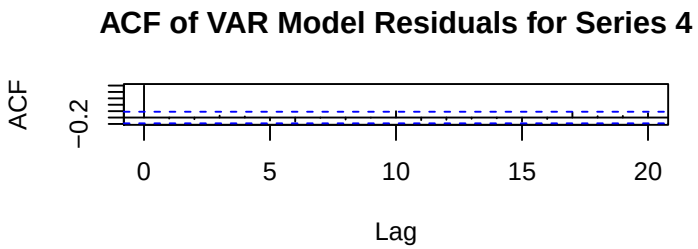


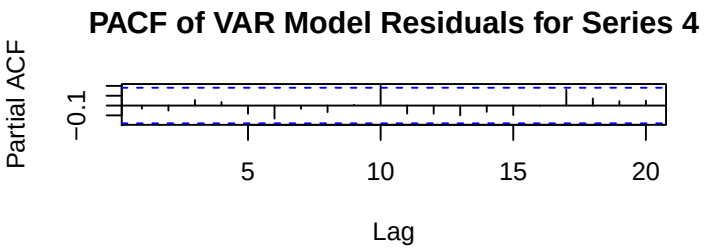



```
## [1] "ADF Test P-Value for VAR Model Residuals for Series 3 : 0.01"
```







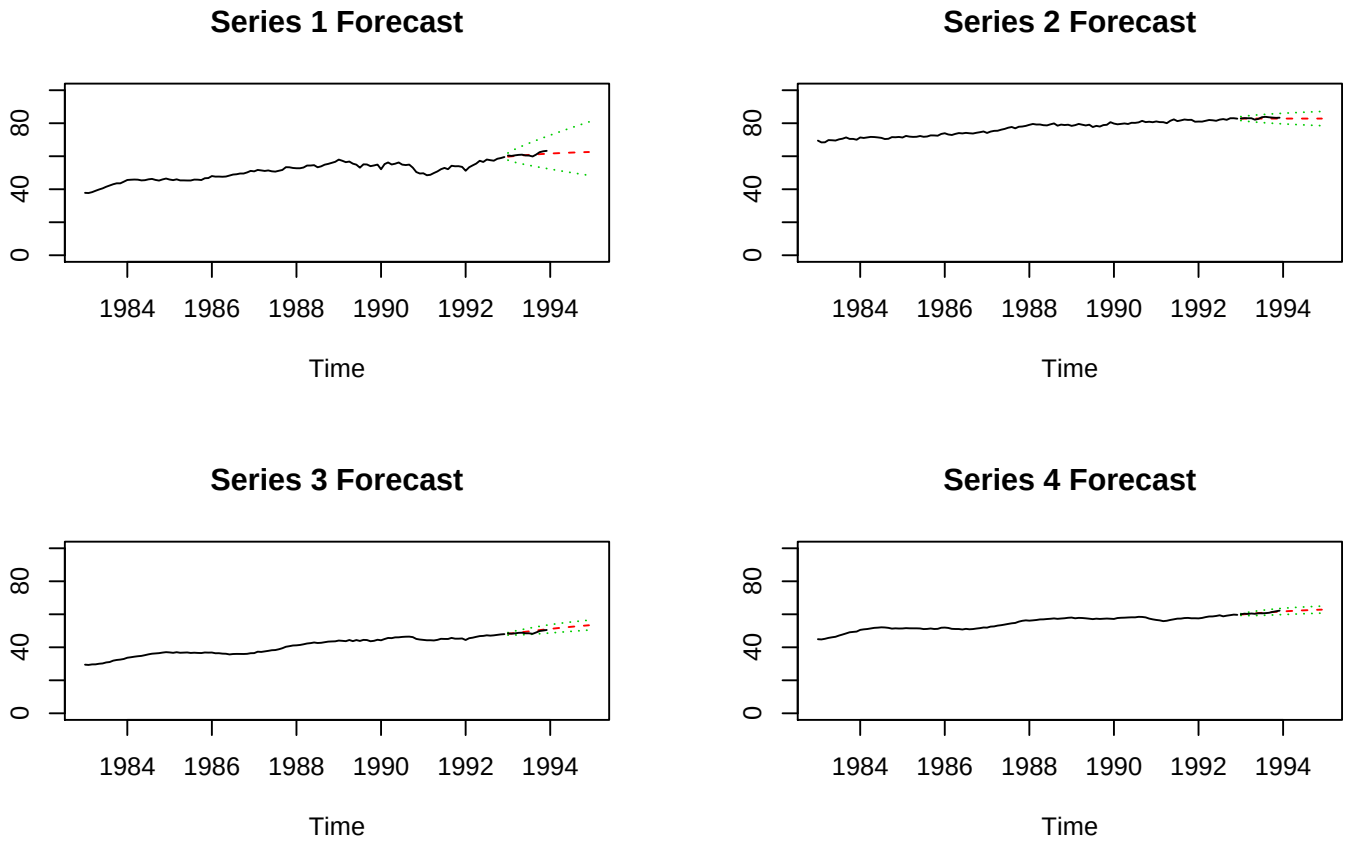


```
## [1] "ADF Test P-Value for VAR Model Residuals for Series 4 : 0.01"
print(paste("Portmanteau Test P-Value:",serial.test(var.fit,lags.pt=12,type="PT.adjusted")$serial$p.value))
## [1] "Portmanteau Test P-Value: 0.759502237544689"
print(paste("Multivariate Jarque-Bera Test P-Value:",
            normality.test(var.fit,multivariate.only=TRUE)$jb.mul$JB$p.value))
## [1] "Multivariate Jarque-Bera Test P-Value: 0.301460653000829"
```

The above time plots, histograms, ACF plots, and PACF plots are for the residuals of the VAR(3) model. Both sets of plots indicate that the VAR(3) model residuals could be a realization of a white noise process. Their Augmented Dickey-Fuller test results indicate that the residuals are stationary, and the Portmanteau and Jarque-Bera tests on the VAR model fail to reject the null hypotheses that the residuals have no autocorrelation and that the residuals are normally distributed, respectively. Since the model is stable and since its residuals are a realization of a white noise process, it is next used for forecasting.

Forecasting

```
fit.pr<-predict(var.fit,n.ahead=24,ci=0.95);q2.pred.1<-get.fcst(q2.mod.1,24,0.95,1993,1,12)
q2.pred.2<-get.fcst(q2.mod.2,24,0.95,1993,1,12);q2.pred.3<-ts(fit.pr$fcst$log.series3.,st=c(1993,1),fr=12)
q2.pred.4<-ts(fit.pr$fcst$series4,st=c(1993,1),fr=12);par(mfrow=c(2,2))
ts.plot(cbind(series1,exp(q2.pred.1$pred),exp(q2.pred.1$l),exp(q2.pred.1$u),q2Test[,1]),
        lty=c(1,2,3,3,1),col=c(1,2,3,3,1),main="Series 1 Forecast",ylim=c(0,100))
ts.plot(cbind(series2,q2.pred.2$pred,q2.pred.2$l,q2.pred.2$u,q2Test[,2]),
        lty=c(1,2,3,3,1),col=c(1,2,3,3,1),main="Series 2 Forecast",ylim=c(0,100))
ts.plot(cbind(series3,exp(q2.pred.3[,1]),exp(q2.pred.3[,2]),exp(q2.pred.3[,3]),q2Test[,3]),
        lty=c(1,2,3,3,1),col=c(1,2,3,3,1),main="Series 3 Forecast",ylim=c(0,100))
ts.plot(cbind(series4,q2.pred.4[,1],q2.pred.4[,2],q2.pred.4[,3],q2Test[,4]),
        lty=c(1,2,3,3,1),col=c(1,2,3,3,1),main="Series 4 Forecast",ylim=c(0,100))
```



The above plot shows the simultaneous forecasts for 1993 and 1994 (red), after fitting to the series (black). The model residuals are omitted here due to the small size of the plots– the 1993 RMSEs from back-testing are: 0.9, 0.57, 0.8, and 0.29, for series 1, 2, 3, and 4, respectively. The higher RMSE for series 1 and 3 are due to a small drop in the actual observations in August of 1993. 95% confidence bounds on the forecasts are shown in green– the actual observations in 1993 fall within these bounds for all series, so this model is able to forecast well. Series 2 has a flat forecast, while the other series are forecasted to continue increasing. Note that the 95% confidence intervals also show that for both series 1 and 2, a decrease in the trend cannot be ruled out.

Appendix

Q1 Additional EDTSA

Since the data for question 1 is so small, all of it can be displayed to check its integrity:

```
ecompctnsa$DATE
```

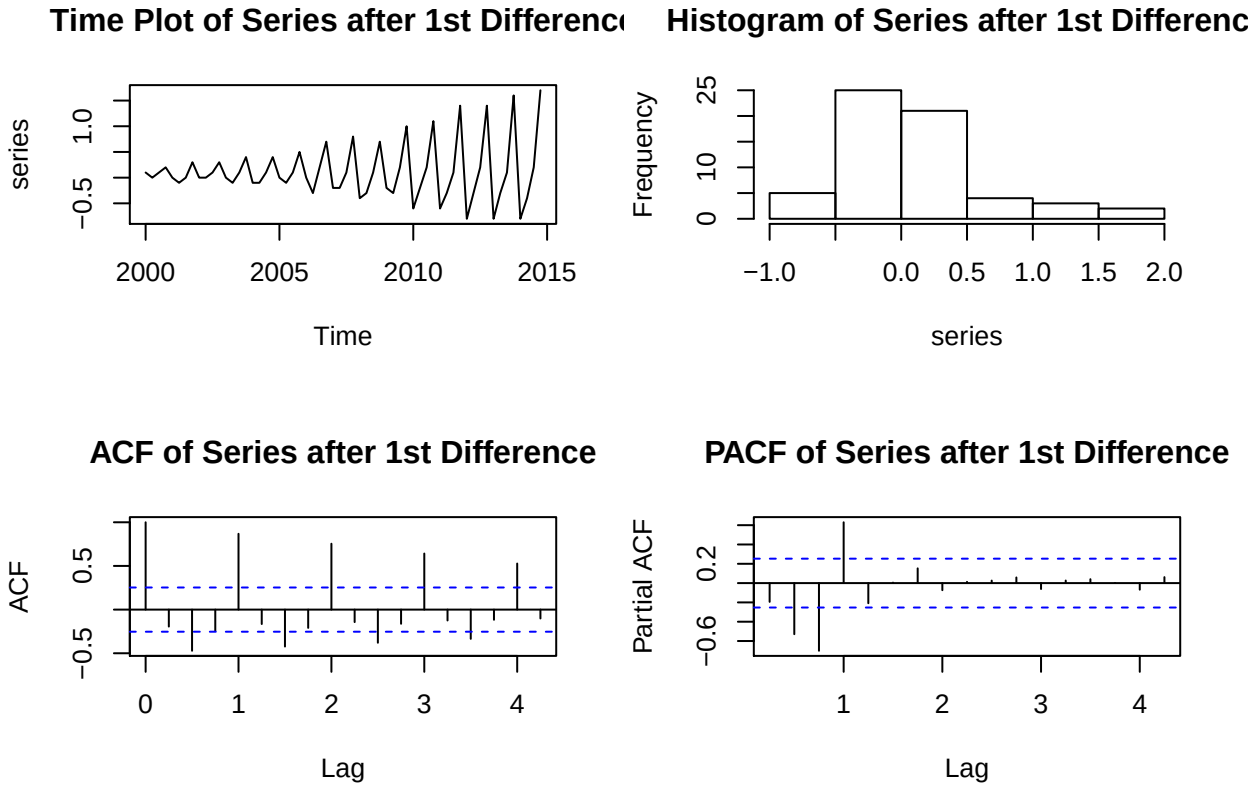
```
##  [1] 1999-10-01 2000-01-01 2000-04-01 2000-07-01 2000-10-01 2001-01-01
##  [7] 2001-04-01 2001-07-01 2001-10-01 2002-01-01 2002-04-01 2002-07-01
## [13] 2002-10-01 2003-01-01 2003-04-01 2003-07-01 2003-10-01 2004-01-01
## [19] 2004-04-01 2004-07-01 2004-10-01 2005-01-01 2005-04-01 2005-07-01
## [25] 2005-10-01 2006-01-01 2006-04-01 2006-07-01 2006-10-01 2007-01-01
## [31] 2007-04-01 2007-07-01 2007-10-01 2008-01-01 2008-04-01 2008-07-01
## [37] 2008-10-01 2009-01-01 2009-04-01 2009-07-01 2009-10-01 2010-01-01
## [43] 2010-04-01 2010-07-01 2010-10-01 2011-01-01 2011-04-01 2011-07-01
## [49] 2011-10-01 2012-01-01 2012-04-01 2012-07-01 2012-10-01 2013-01-01
## [55] 2013-04-01 2013-07-01 2013-10-01 2014-01-01 2014-04-01 2014-07-01
## [61] 2014-10-01 2015-01-01 2015-04-01 2015-07-01 2015-10-01 2016-01-01
## [67] 2016-04-01 2016-07-01 2016-10-01
## 69 Levels: 1999-10-01 2000-01-01 2000-04-01 2000-07-01 ... 2016-10-01
```

```
ecompctnsa$ECOMPCTNSA
```

```
##  [1] 0.7 0.8 0.8 0.9 1.1 1.1 1.0 1.0 1.3 1.3 1.3 1.4 1.7 1.7 1.6 1.7 2.1
## [18] 2.0 1.9 2.0 2.4 2.4 2.3 2.4 2.9 2.9 2.6 2.8 3.5 3.3 3.1 3.2 4.0 3.6
## [35] 3.3 3.4 4.1 3.9 3.6 3.8 4.8 4.2 4.0 4.2 5.3 4.7 4.4 4.5 5.9 5.1 4.8
## [52] 5.0 6.4 5.6 5.3 5.4 7.0 6.2 5.8 6.0 7.7 6.9 6.6 6.8 8.7 7.7 7.5 7.7
## [69] 9.5
```

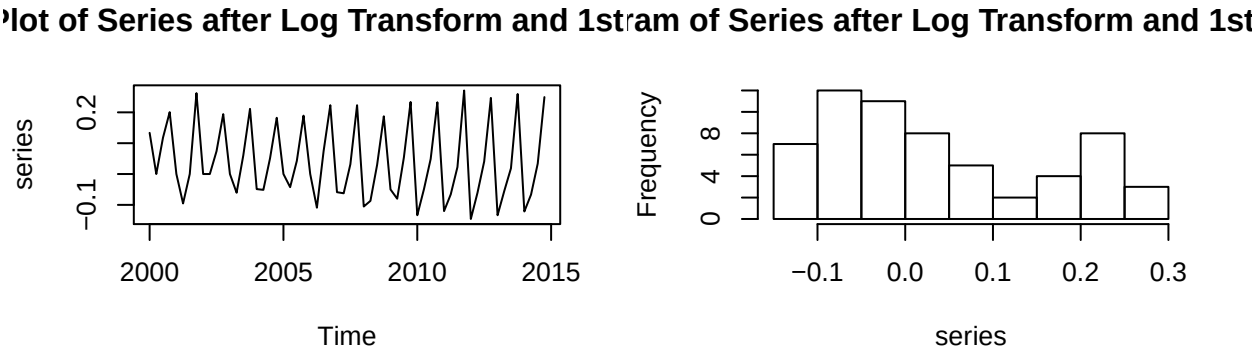
The following plots show the increasing variance of the series that led to a log transform.

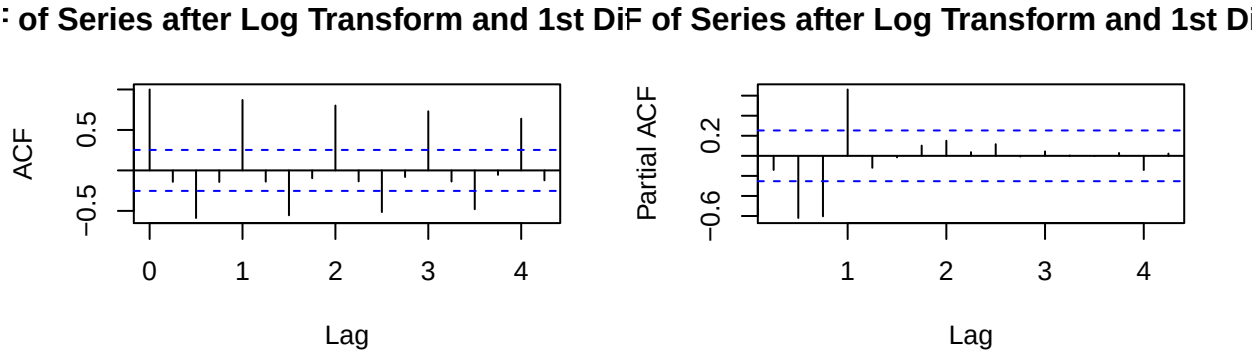
```
tsplot(diff(ecompctnsa.ts), "Series after 1st Difference", 0, 0, 0)
```



The following plots show the strong seasonality of the series.

```
tsplot(diff(log(ecompctnsa.ts)), "Series after Log Transform and 1st Difference", 0, 0, 0)
```

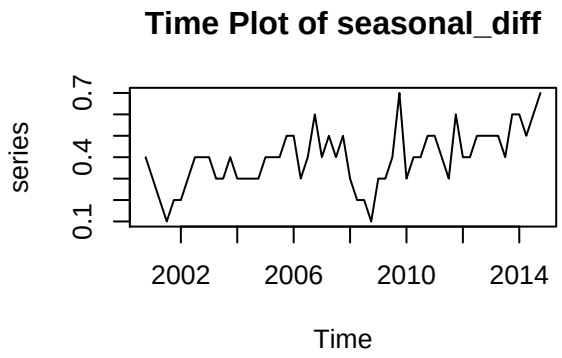


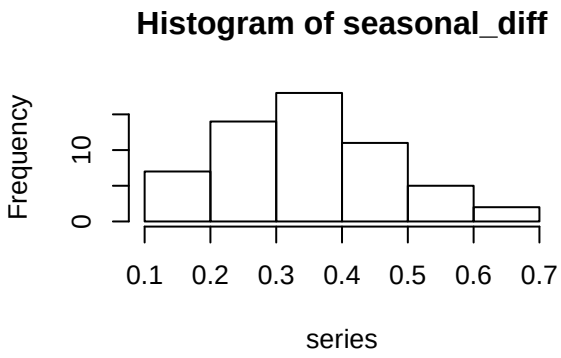


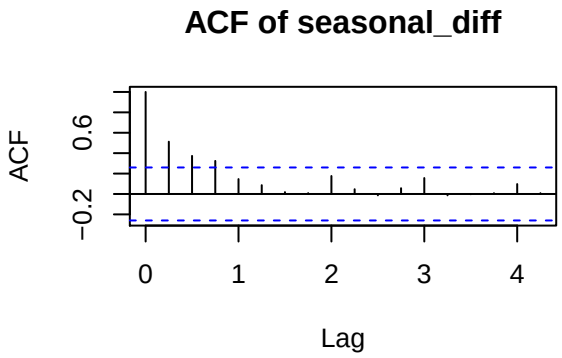
Q1 Alternate Modeling Approaches

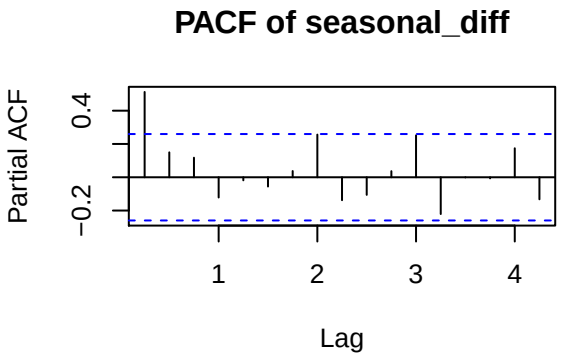
Alternative 1: seasonal difference and AR/SAR model

```
tsplot(diff(ecompctnsa.ts, lag=4), title="seasonal_diff", 1, 1, 0)
```










```
## [1] "ADF Test P-Value for seasonal_diff : 0.108401678485301"
## [1] "Ljung-Box Test P-Value for seasonal_diff : 7.14007753663815e-05"
```

After the seasonal difference, the series shows high persistence, and some lingering seasonal effects. The gradual decline in the ACF and spike at lag 1 in the PACF suggest an AR(1) model.

```
search.sarima.params(ecompctnsa.ts, 4, 2, 0, 4, 2, 1, 4)
```

```
##     p d q P D Q       AIC       BIC  box.test shapiro.test
## 158 2 0 0 1 1 2 -91.97008 -79.71177 0.6155019    0.1901356
## 98  1 0 1 1 1 2 -91.81242 -79.55411 0.6218790    0.1587412
## 159 2 0 0 1 1 3 -90.45109 -76.14974 0.6239279    0.3365087
## 163 2 0 0 2 1 2 -90.41170 -76.11034 0.6265054    0.3534835
## 162 2 0 0 2 1 1 -90.24375 -77.98544 0.6190233    0.5547386
```

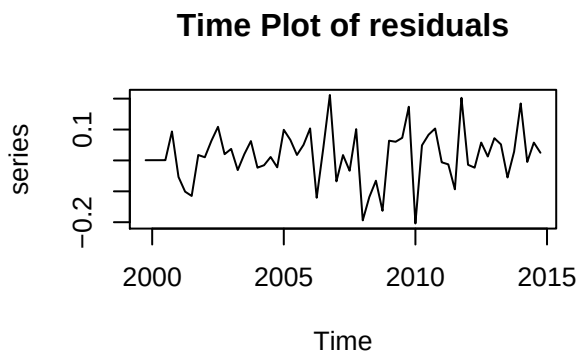
```
#search.sarima.params(log(ecompctnsa.ts), 4, 2, 0, 4, 2, 1, 4)
```

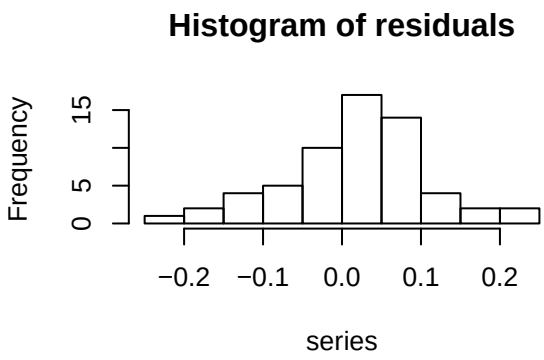
All of the above models seem to be equally good candidates.

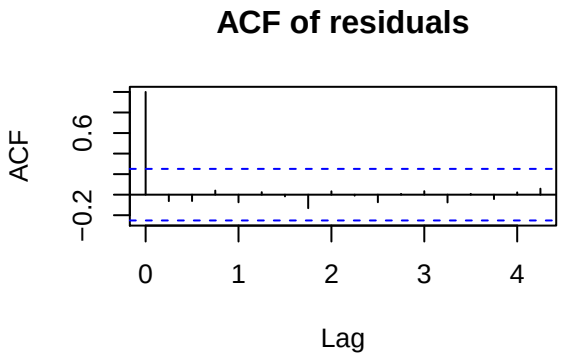
```
q1.mod <- Arima(ecompctnsa.ts, order=c(2,0,0), seasonal=list(order=c(1,1,2),4), method="ML")
q1.mod;
```

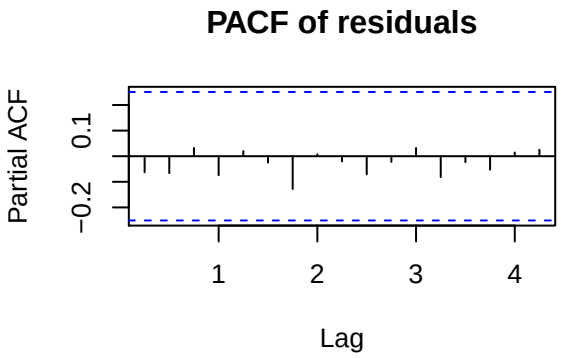
```
## Series: ecompctnsa.ts
## ARIMA(2,0,0)(1,1,2)[4]
##
## Coefficients:
##          ar1     ar2    sar1     sma1    sma2
##       0.5620  0.3368  0.9715  -1.3549  0.5614
## s.e.  0.1253  0.1262  0.0317   0.1413  0.1398
##
## sigma^2 estimated as 0.008778:  log likelihood=51.99
## AIC=-91.97   AICc=-90.29   BIC=-79.71
```

```
tsplot(q1.mod$residuals, "residuals", 1, 1, 1)
```



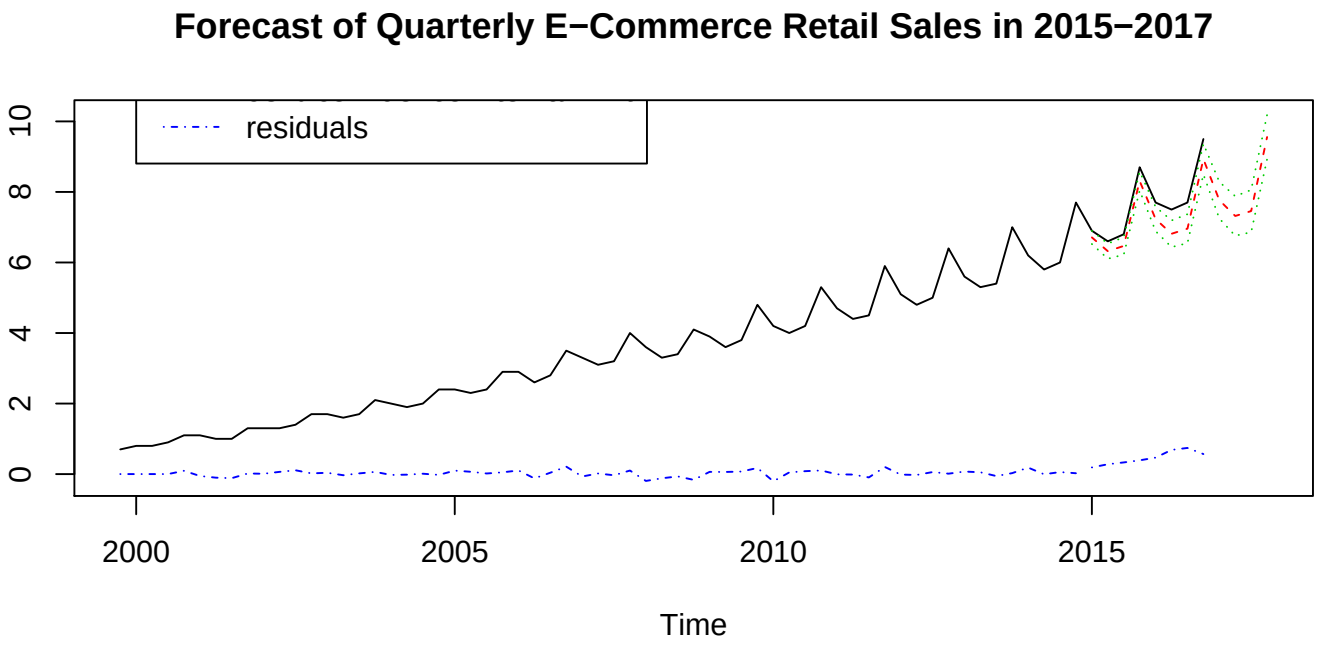







```
## [1] "ADF Test P-Value for residuals : 0.013365148851757"
## [1] "Ljung-Box Test P-Value for residuals : 0.61550185271121"
## [1] "Shapiro-Wilk Test P-Value for residuals : 0.190135617261871"
```

```
q1.pred<-get.fcst(q1.mod,12,0.95,2015,1,4)
ts.plot(cbind(q1.original.ts,(q1.pred$pred),(q1.pred$u),(q1.pred$l),
              q1.original.ts-(ts.union(q1.mod$fitted, q1.pred$pred))),lty=c(1,2,3,3,4,4),
        col=c(1,2,3,3,4,4),main="Forecast of Quarterly E-Commerce Retail Sales in 2015-2017")
legend(x=2000,y=14,legend=c("original","predicted","95% confidence interval limit","residuals"),
       lty=c(1,2,3,4),col=c(1,2,3,4))
```



The model diagnostics show that this model passes all ARIMA assumptions, however both its relatively weak in-sample fit and back-testing results show that this model is less favorable for forecasting than the one selected.

Alternative 2: seasonal ARMA on the 1st differenced log transform

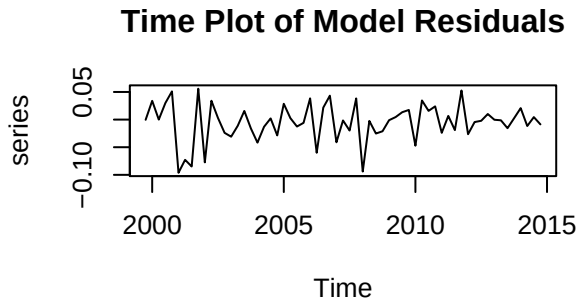
```
#search.sarima.params(log(ecompctnsa.ts),4,4,1,4,4,0,4)
search.sarima.params(log(ecompctnsa.ts),4,3,1,3,3,0,3)
```

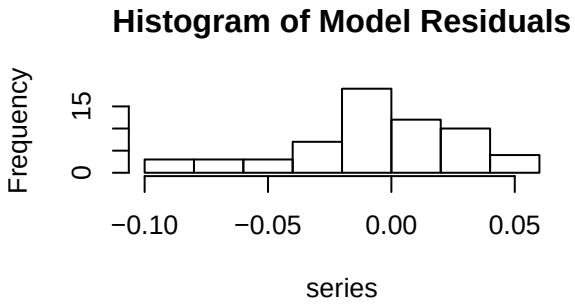
```
##    p d q P D Q       AIC       BIC   box.test shapiro.test
## 11 0 1 0 2 0 2 -210.7033 -200.2316 0.10277352  0.004093671
## 7  0 1 0 1 0 2 -210.1162 -201.7388 0.08569823  0.016011974
## 13 0 1 0 3 0 0 -209.4148 -201.0374 0.06861598  0.024823073
## 9  0 1 0 2 0 0 -209.2038 -202.9207 0.08008544  0.014253641
## 10 0 1 0 2 0 1 -208.4299 -200.0525 0.06971993  0.017784873
```

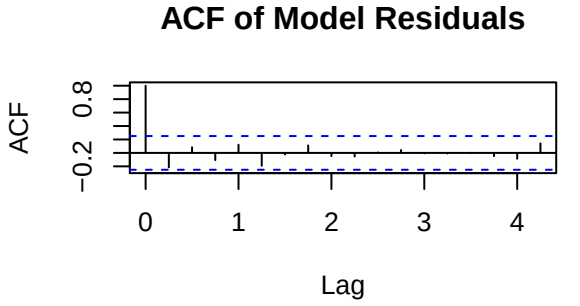
```
mod.3<-Arima(log(ecompctnsa.ts),order=c(0,1,0),seasonal=list(order=c(1,0,2),4),method="ML");mod.3
```

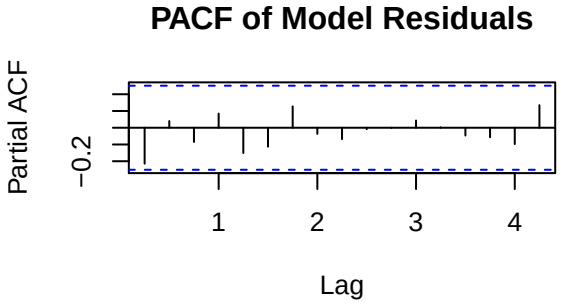
```
## Series: log(ecompctnsa.ts)
## ARIMA(0,1,0)(1,0,2)[4]
##
## Coefficients:
##         sar1     sma1    sma2
##       0.9864  -0.7612  0.4024
## s.e.  0.0111   0.1576  0.1347
##
## sigma^2 estimated as 0.001294:  log likelihood=109.06
## AIC=-210.12   AICc=-209.39   BIC=-201.74
```

```
tsplot(mod.3$residuals, "Model Residuals", 1, 1, 1)
```










```
## [1] "ADF Test P-Value for Model Residuals : 0.01"
## [1] "Ljung-Box Test P-Value for Model Residuals : 0.085698229474127"
## [1] "Shapiro-Wilk Test P-Value for Model Residuals : 0.016011973817517"
```

```
abs(polyroot(c(1, mod.3$coef[1])))
```

```
## [1] 1.013794
```

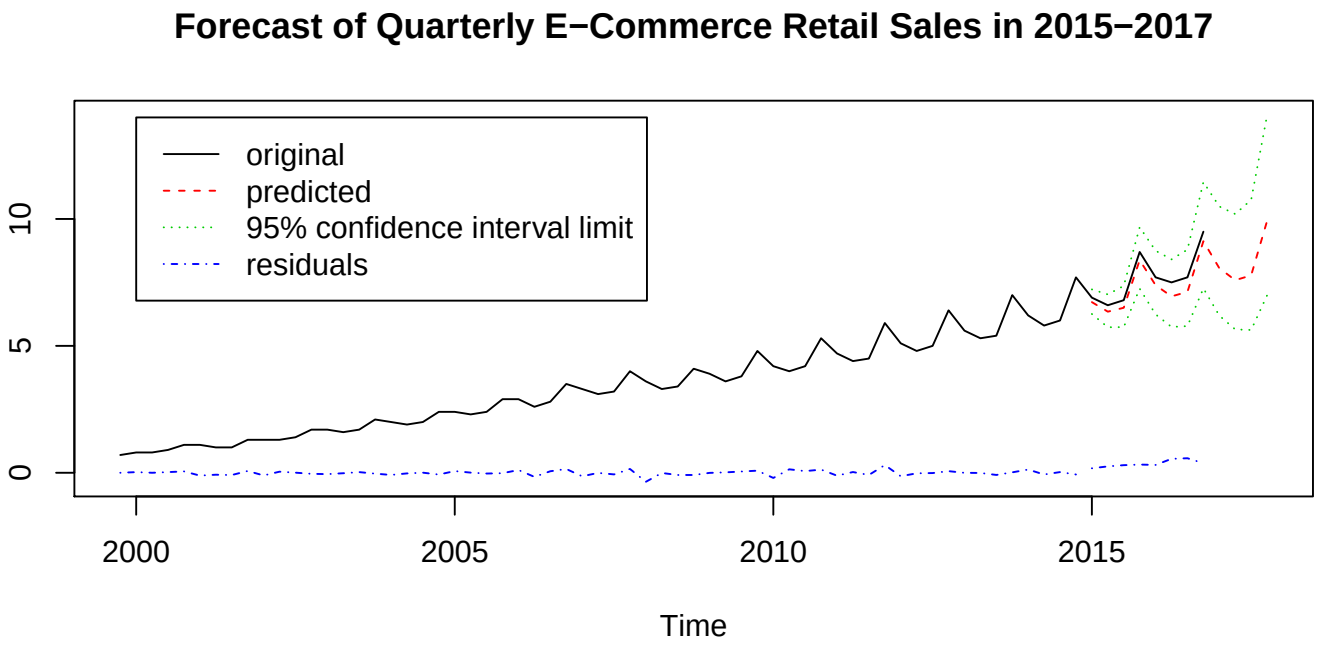
```
data.frame(roots=abs(polyroot(c(1, mod.3$coef[2], mod.3$coef[3]))),
           Phi1.CI=mod.3$coef[1] + c(-2,2)*sqrt(mod.3$var.coef)[1],
           Theta1.CI=mod.3$coef[2] + c(-2,2)*sqrt(mod.3$var.coef)[5],
           Theta2.CI=mod.3$coef[3] + c(-2,2)*sqrt(mod.3$var.coef)[9])
```

```
##      roots   Phi1.CI  Theta1.CI Theta2.CI
## 1 1.576405 0.9642309 -1.0763984 0.1329796
## 2 1.576405 1.0085568 -0.4459382 0.6718323
```

```
shapiro.test(mod.3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod.3$residuals
## W = 0.95096, p-value = 0.01601
```

```
mod.3.prediction <- predict(mod.3, n.ahead=12)
mod.3.prediction.preds <- ts(mod.3.prediction$pred, st=c(2015,1), fr=4)
mod.3.prediction.uci<-mod.3.prediction.preds+2*mod.3.prediction$se
mod.3.prediction.lci<-mod.3.prediction.preds-2*mod.3.prediction$se
ts.plot(cbind(q1.original.ts, exp(mod.3.prediction.preds), exp(mod.3.prediction.uci), exp(mod.3.prediction.lci)
              q1.original.ts-exp(ts.union(mod.3$fitted, mod.3.prediction.preds))),lty=c(1,2,3,3,4,4),
        col=c(1,2,3,3,4,4),main="Forecast of Quarterly E-Commerce Retail Sales in 2015-2017")
legend(x=2000,y=14,legend=c("original","predicted","95% confidence interval limit","residuals"),
       lty=c(1,2,3,4),col=c(1,2,3,4))
```

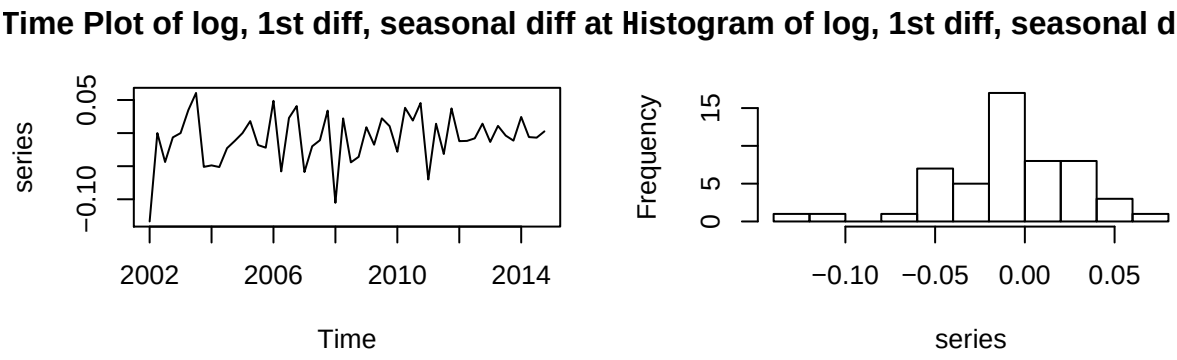


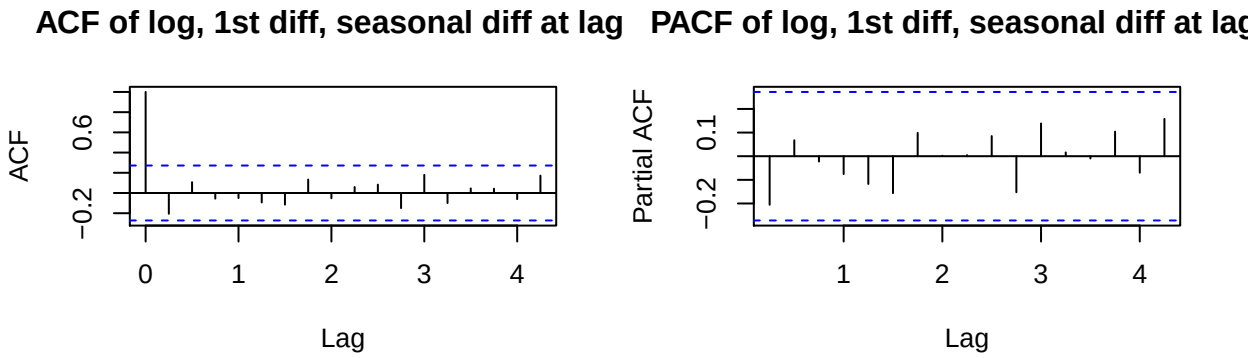
This model had good in-sample fit, but was less parsimonious and had correlated residuals.

Interestingly, after a log transformation, first difference, and seasonal difference at a lag of 8, the series becomes white noise:

```
tsplot(diff(diff(log(ecompctnsa.ts)), lag=8), "log, 1st diff, seasonal diff at lag 8", 1, 1, 1)
```

```
## Warning in adf.test(series): p-value smaller than printed p-value
```






```
## [1] "ADF Test P-Value for log, 1st diff, seasonal diff at lag 8 : 0.01"
## [1] "Ljung-Box Test P-Value for log, 1st diff, seasonal diff at lag 8 : 0.127086498135534"
## [1] "Shapiro-Wilk Test P-Value for log, 1st diff, seasonal diff at lag 8 : 0.0440116117145459"
```

Q2 Additional EDTSA

The following function was used to check the integrity of the time variables (year and mon), and to produce summary statistics for the remaining variables.

```
year.month.check <- rep(0,length(varData$year))
iter <- 1
for (a in 1947:1993) {
  for (b in 1:12) {
    year.check <- ifelse(test=varData$year[iter]==a, yes=0, no=1)
    month.check <- ifelse(test=varData$mon[iter]==b, yes=0, no=1)
    year.month.check[iter] <- year.check + month.check
    iter <- iter + 1
  }
}
summary(year.month.check); summary(varData[,3:6])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

```
##     series1         series2         series3          series4
##  Min.   :11.13   Min.   :20.86   Min.   : 5.007   Min.   :11.73
##  1st Qu.:17.57   1st Qu.:33.23   1st Qu.: 9.317   1st Qu.:22.04
##  Median :33.69   Median :53.63   Median :17.389   Median :38.31
##  Mean   :32.61   Mean   :52.29   Mean   :21.313   Mean   :36.90
##  3rd Qu.:44.78   3rd Qu.:68.57   3rd Qu.:32.539   3rd Qu.:50.48
##  Max.   :63.27   Max.   :83.87   Max.   :50.493   Max.   :62.22
```

```
print(paste("Series 1 Minimum Value:",min(varData$series1)))
```

```
## [1] "Series 1 Minimum Value: 11.1307"
```

```
print(paste("Series 2 Minimum Value:",min(varData$series2)))
```

```
## [1] "Series 2 Minimum Value: 20.8575"
```

```
print(paste("Series 3 Minimum Value:",min(varData$series3)))
```

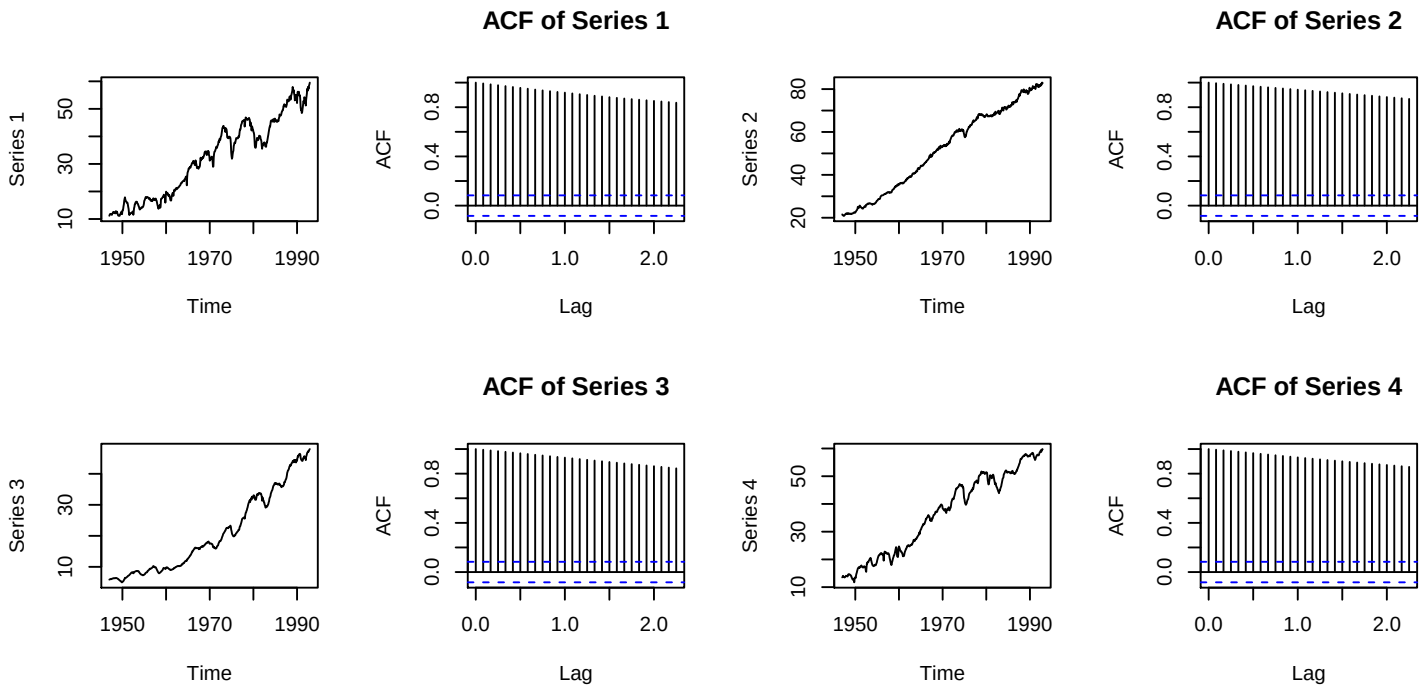
```
## [1] "Series 3 Minimum Value: 5.0072"
```

```
print(paste("Series 4 Minimum Value:",min(varData$series4)))
```

```
## [1] "Series 4 Minimum Value: 11.7309"
```

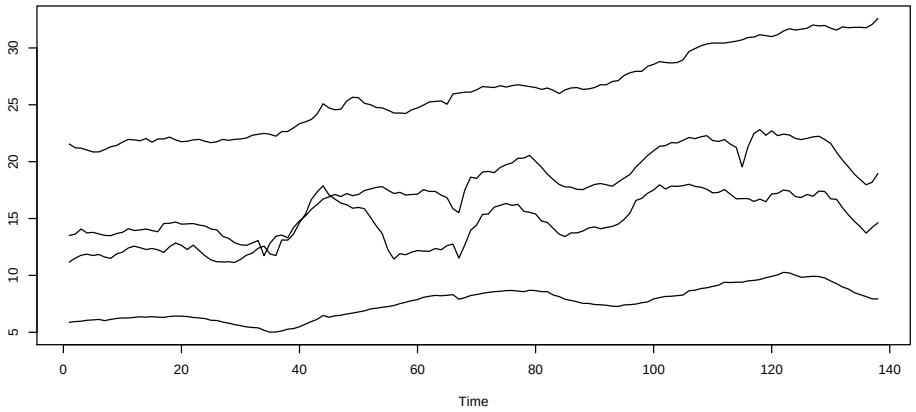
The following time plots and ACF plots show that the original series each have a strong increasing trend and are each highly persistent.

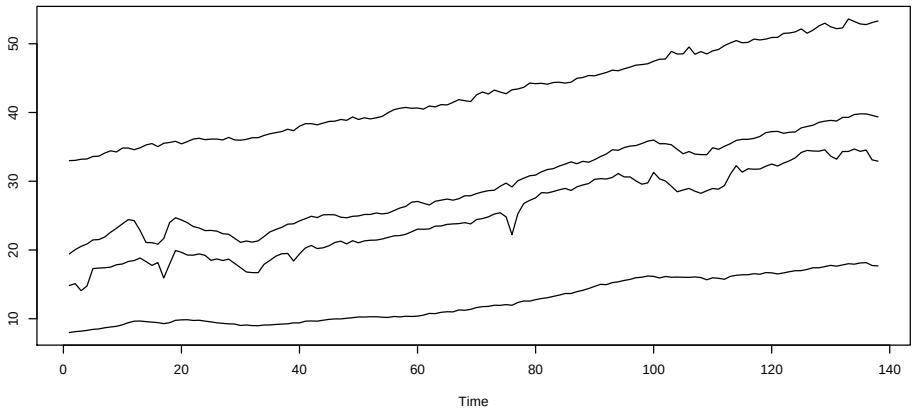
```
q2Data<-varData[varData$year<1993,]
series1<-ts(q2Data$series1,start=c(1947,1),frequency=12)
series2 <- ts(q2Data$series2, start=c(1947,1), frequency=12)
series3 <- ts(q2Data$series3, start=c(1947,1), frequency=12)
series4 <- ts(q2Data$series4, start=c(1947,1), frequency=12)
allseries <- cbind(series1, series2, series3, series4)
par(mfrow=c(2,4));for(num in 1:4){
plot.ts(allseries[,num],ylab=paste("Series",num));acf(allseries[,num],main=paste("ACF of Series",num))}
```

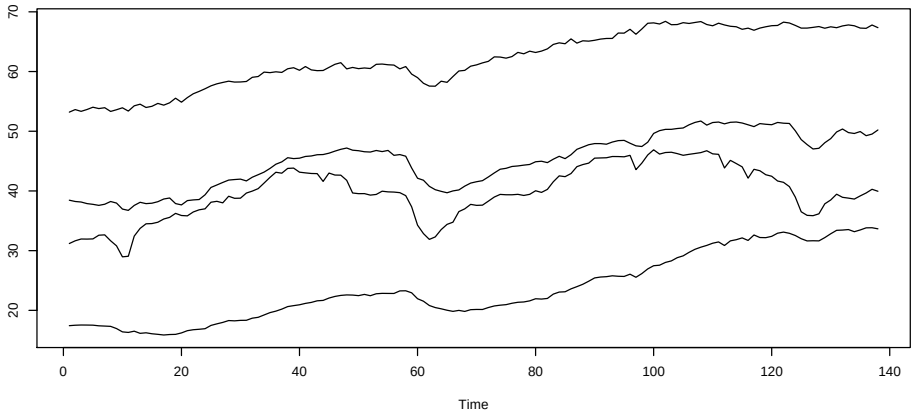


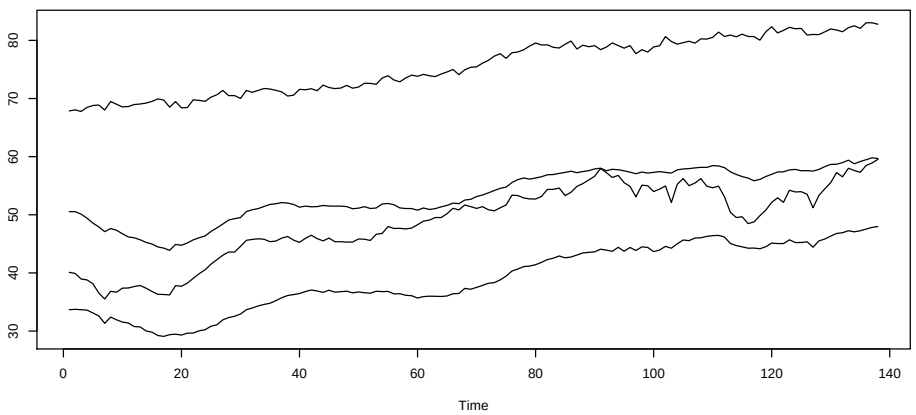
The following plots display the original series in 4 segments in order to more closely examine their trends and select the data for training.

```
par(mfrow=c(4,1));ts.plot(allseries[1:138,]);ts.plot(allseries[139:276,])
ts.plot(allseries[277:414,]);ts.plot(allseries[415:552,])
```


5
10
15
20
25
30
0
20
40
60
80
100
120
140
Time


10
20
30
40
50
0
20
40
60
80
100
120
140
Time

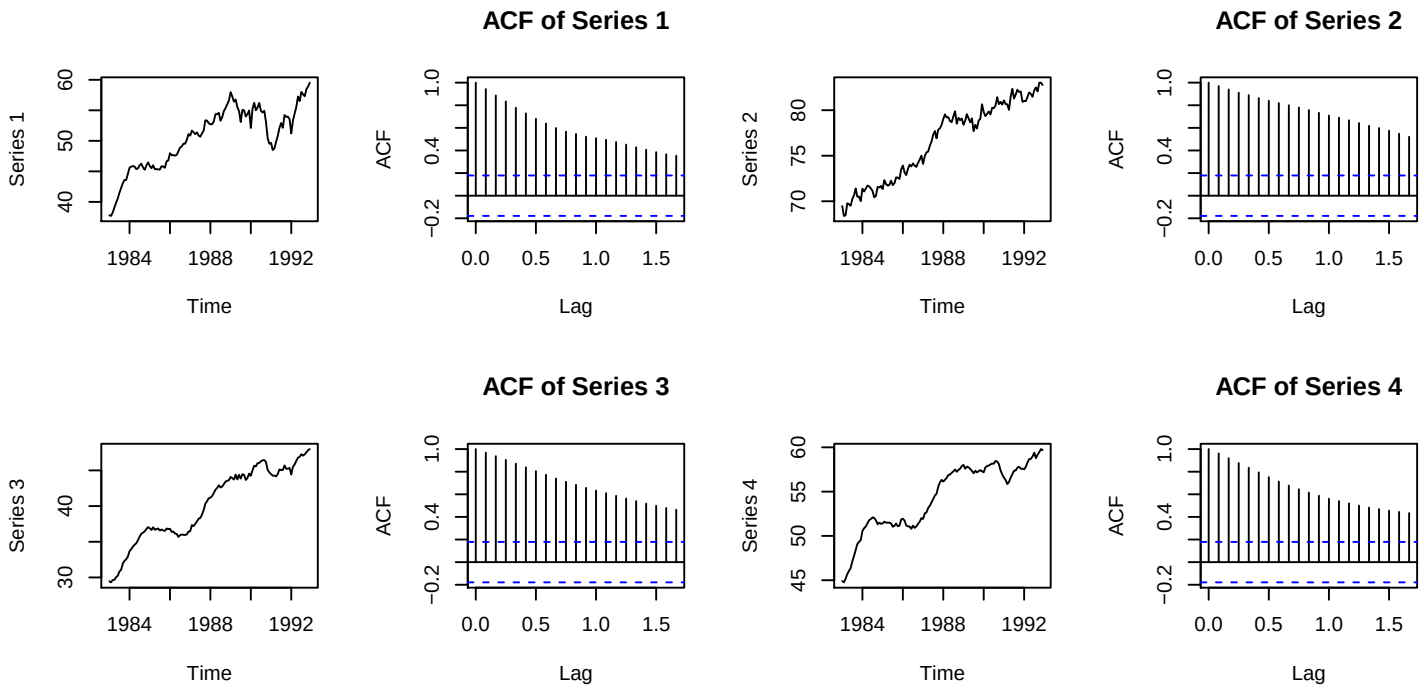

20
30
40
50
60
70
0
20
40
60
80
100
120
140
Time


30
40
50
60
70
80
0
20
40
60
80
100
120
140
Time

```
par(mfrow=c(1,1));ts.plot(window(allseries, st=c(1983,1)))
```

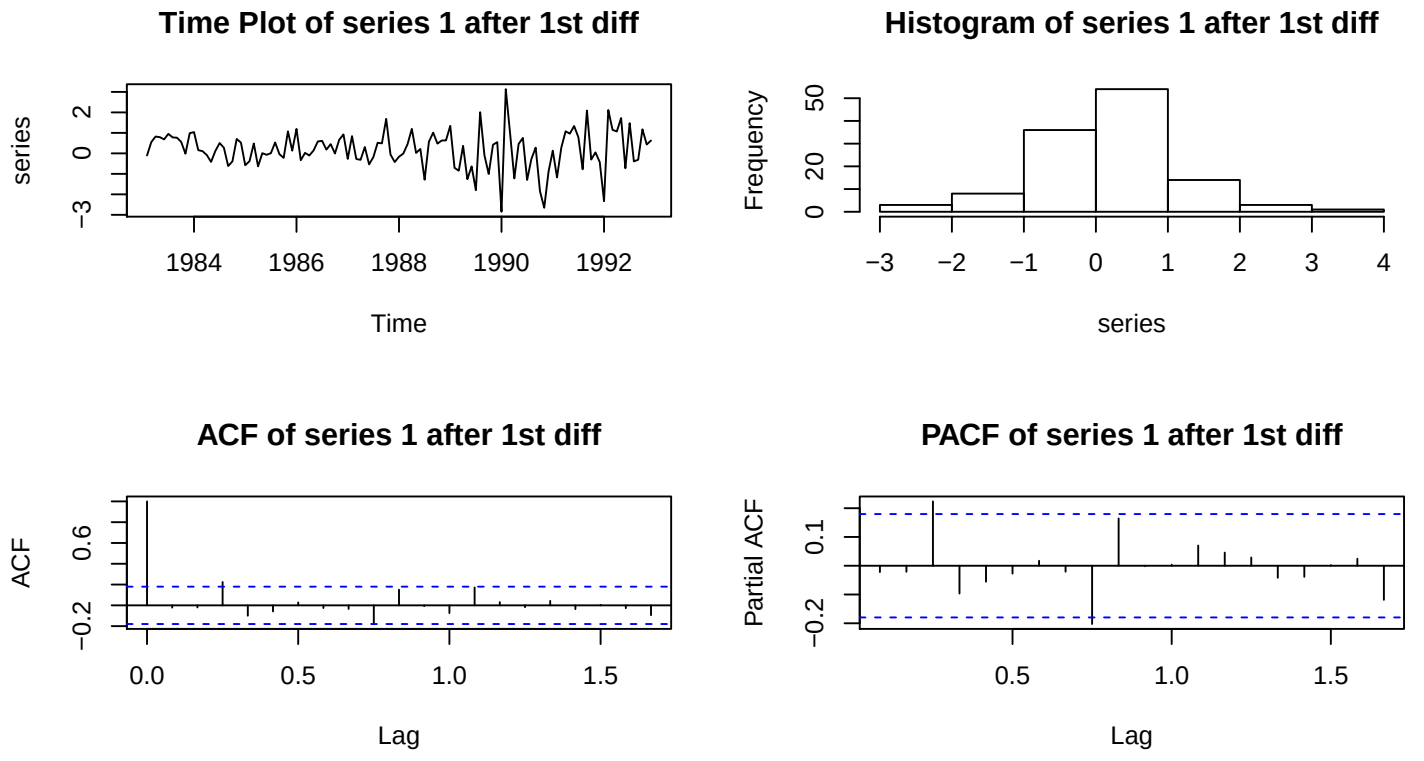
The following plots show that in the shortened training data, each series is still highly persistent, with a strong increasing trend. The time plots again show the close relationship between series 3 and series 4, especially in this time frame selected for training the model.

```
q2Data<-varData[varData$year<1993,]
series1<-window(ts(q2Data$series1,start=c(1947,1),frequency=12), start=c(1983,1),frequency=12)
series2 <- window(ts(q2Data$series2, start=c(1947,1), frequency=12), start=c(1983,1), frequency=12)
series3 <- window(ts(q2Data$series3, start=c(1947,1), frequency=12), start=c(1983,1), frequency=12)
series4 <- window(ts(q2Data$series4, start=c(1947,1), frequency=12), start=c(1983,1), frequency=12)
allseries <- cbind(series1, series2, series3, series4)
par(mfrow=c(2,4));for(num in 1:4){
plot.ts(allseries[,num],ylab=paste("Series",num));acf(allseries[,num],main=paste("ACF of Series",num))}
```



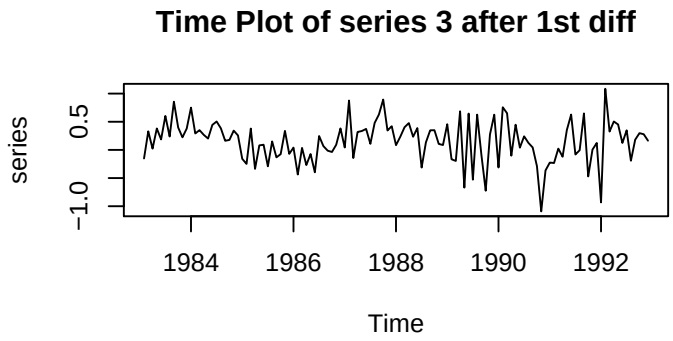
The following plots show the increasing variance in Series 1 that called for a log transformation to stabilize the variance.

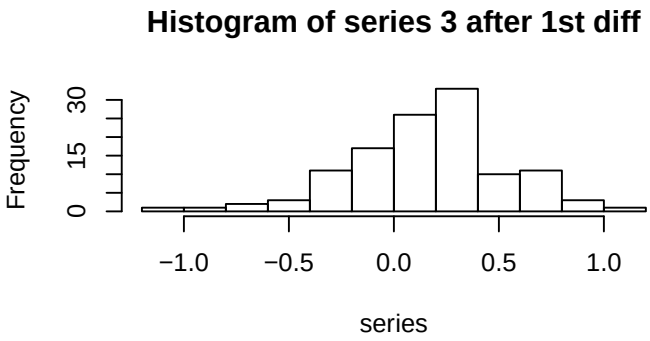
```
tsplot(diff(series1), "series 1 after 1st diff", 0, 0, 0)
```

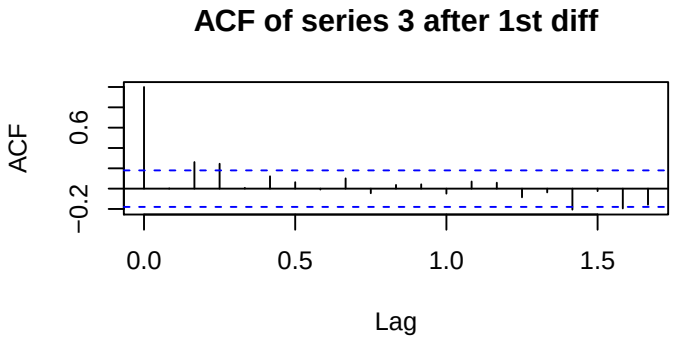


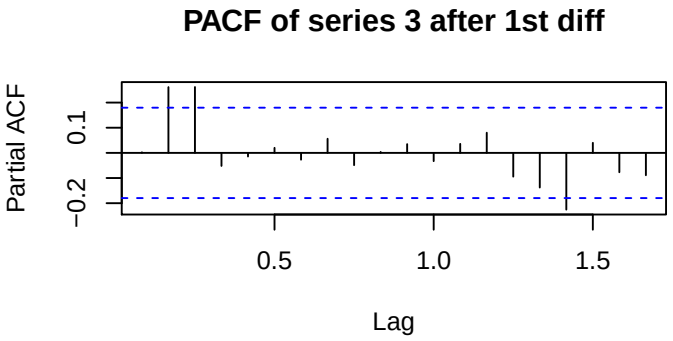
The following plots show the increasing variance in Series 3 that called for a log transformation to stabilize its variance.

```
tsplot(diff(series3), "series 3 after 1st diff", 0, 0, 0)
```







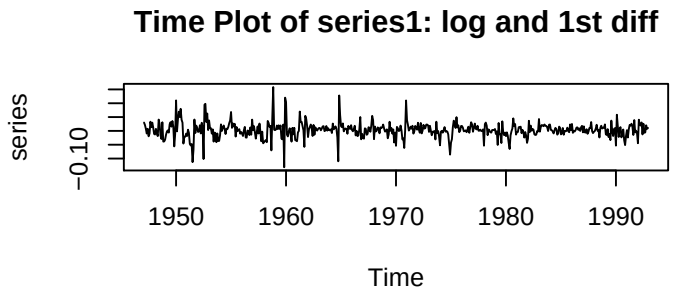


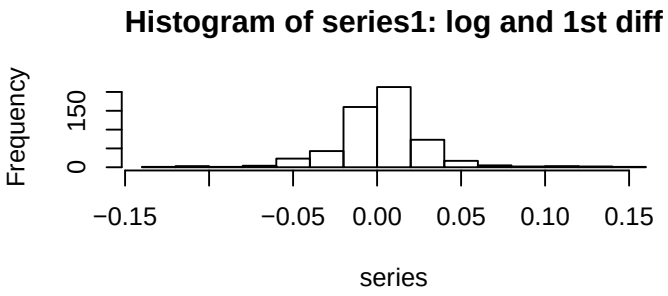
Q2 Alternate Modeling Approaches

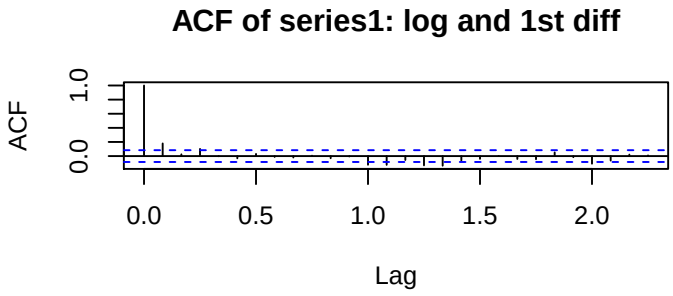
Approach 1: Arima for Series 1 and 2; VAR for Series 3 and 4 (using full data)

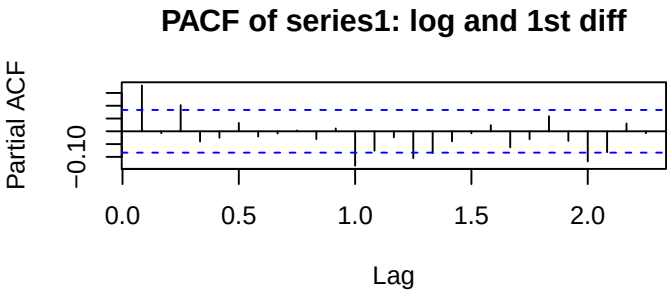
```
series1<-ts(q2Data$series1,start=c(1947,1),frequency=12)
series2 <- ts(q2Data$series2, start=c(1947,1), frequency=12)
series3 <- ts(q2Data$series3, start=c(1947,1), frequency=12)
series4 <- ts(q2Data$series4, start=c(1947,1), frequency=12)
allseries <- cbind(series1, series2, series3, series4)
```

```
tsplot(diff(log(series1)), "series1: log and 1st diff", 1, 1, 0)
```










```
## [1] "ADF Test P-Value for series1: log and 1st diff : 0.01"
## [1] "Ljung-Box Test P-Value for series1: log and 1st diff : 2.48212092824884e-05"
```

Prior to modeling Series 1, a first difference is taken to remove the trend. The plot of the differenced series, as well as subsequent model residuals, showed evidence of increasing variance, so a log transform is also taken in order to stabilize the variance. The above time plot, histogram, ACF plot, and PACF plot are for series 1 after a log-transformed and first difference. The spike in the PACF at lag 1 indicates an AR term, and spikes in the ACF and especially the PACF at lags 12 and 24 suggest the need for seasonal components as well. A parameter search is therefore focused on non-seasonal AR and MA components up to order 1, and seasonal AR and MA components up to order 2.

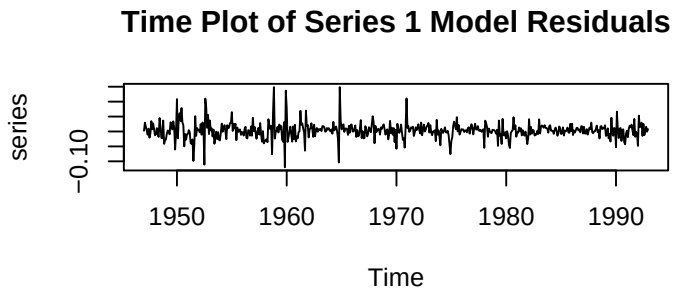
```
search.sarima.params(log(series1),12,1,1,1,2,0,2)
```

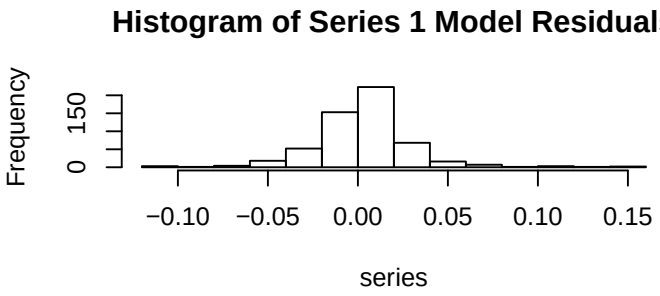
An ARIMA(1,1,0)(0,0,2)_12 model is the best candidate in terms of both AIC and BIC, and a Ljung-Box test on its residuals fails to reject the null hypothesis of no correlation.

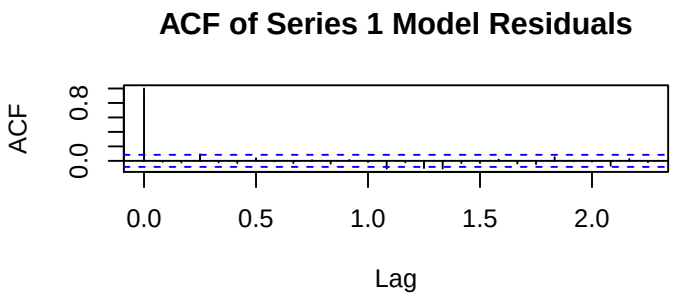
```
q2.mod.1 <- Arima(log(series1),order=c(1,1,0),seasonal=list(order=c(0,0,2),12),method="ML");q2.mod.1
```

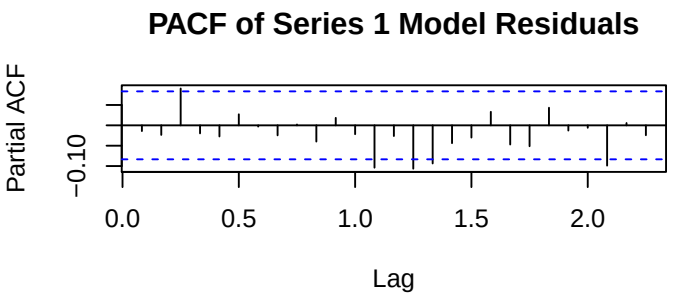
```
## Series: log(series1)
## ARIMA(1,1,0)(0,0,2)[12]
##
## Coefficients:
##          ar1     sma1     sma2
##       0.1786  -0.1053  -0.1001
## s.e.  0.0420   0.0434   0.0452
##
## sigma^2 estimated as 0.0007564:  log likelihood=1199.44
## AIC=-2390.89   AICc=-2390.81   BIC=-2373.64
```

```
tsplot(q2.mod.1$residuals, "Series 1 Model Residuals", 1, 1, 1)
```










```
## [1] "ADF Test P-Value for Series 1 Model Residuals : 0.01"
## [1] "Ljung-Box Test P-Value for Series 1 Model Residuals : 0.738931762715525"
## [1] "Shapiro-Wilk Test P-Value for Series 1 Model Residuals : 2.3739397987512e-18"
```

```
data.frame(sma.roots=abs(polyroot(c(1,q2.mod.1$coef[2],q2.mod.1$coef[3]))),
           phi1.CI=q2.mod.1$coef[1] + c(-2,2)*sqrt(q2.mod.1$var.coef)[1],
           Theta1.CI=q2.mod.1$coef[2] + c(-2,2)*sqrt(q2.mod.1$var.coef)[5],
           Theta2.CI=q2.mod.1$coef[3] + c(-2,2)*sqrt(q2.mod.1$var.coef)[9])
```

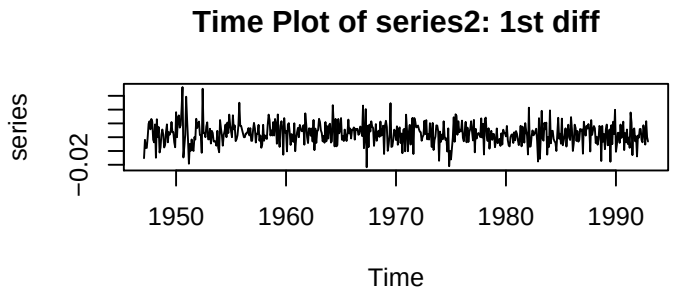
```
##   sma.roots    phi1.CI   Theta1.CI    Theta2.CI
## 1  2.677924 0.09452925 -0.19198016 -0.190504236
## 2  3.729055 0.26271819 -0.01853808 -0.009773655
```

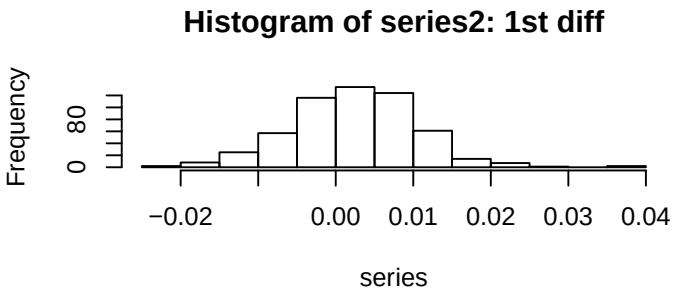
The series 1 model summary shows that each of the coefficients is statistically significant– their 95% confidence intervals, shown in the dataframe above, do not include zero. Since the coefficient on the AR term is 0.18, the characteristic equation for the AR process is (1-0.18B), which has a root > 1 and is therefore stationary. The roots of the characteristic equation for the seasonal MA component are also shown in the dataframe above– since they are also greater than 1, the seasonal MA process is invertible. The time plot, histogram, ACF plot, and PACF plot of the residuals resemble a realization of a white noise process, and as mentioned previously, the Box-Ljung test fails to reject the null hypothesis of no correlation in the residuals. Since the residuals resemble a white noise process, and since the stationarity and invertibility conditions are satisfied, this model can be used for forecasting.

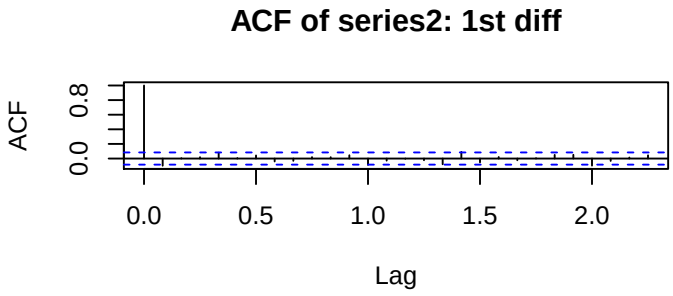
Similar transformations are made to series 2 prior to modeling:

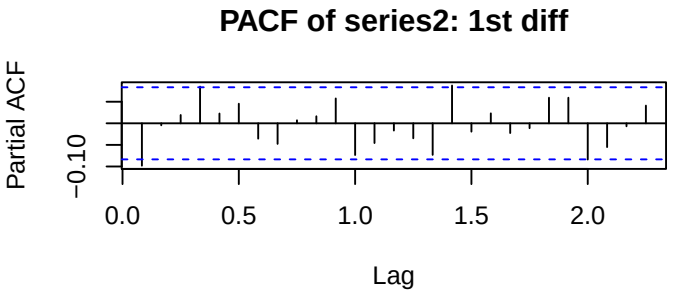
```
tsplot(diff(log(series2)), "series2: 1st diff", 1, 1, 0)
```

```
## Warning in adf.test(series): p-value smaller than printed p-value
```










```
## [1] "ADF Test P-Value for series2: 1st diff : 0.01"
## [1] "Ljung-Box Test P-Value for series2: 1st diff : 0.0207934509408901"
```

Prior to modeling Series 2, a first difference is taken to remove the trend. The plot of the differenced series, as well as subsequent model residuals, showed evidence of increasing variance, so a log transform is also taken in order to stabilize the variance. The above time plot, histogram, ACF plot, and PACF plot are for series 2 after a log-transformation and first difference. There are no clear patterns in the ACF and PACF graphs– only small spikes at a lag of 1 in both graphs, so the following parameter search focuses on the addition of non-seasonal AR and MA components up to order 2. It is notable that the small spikes at lags 12 and 24 are almost significant, suggesting that seasonal components may be helpful. A model that included seasonal components was tested– the AIC and BIC were improved, however, this improvement was only minor and we therefore omit seasonal components for greater model parsimony, which is favorable for forecasting.

```
search.sarima.params(log(series2),12,2,1,2,0,0,0)
```

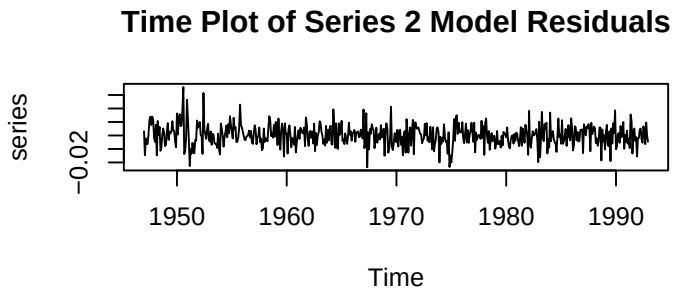
```
##   p d q P D Q       AIC       BIC   box.test shapiro.test
## 6 1 1 2 0 0 0 -3754.202 -3736.955 0.99065830 0.0013404449
## 8 2 1 1 0 0 0 -3754.092 -3736.845 0.97883946 0.0012823762
## 9 2 1 2 0 0 0 -3752.198 -3730.639 0.98930260 0.0013281631
## 7 2 1 0 0 0 0 -3707.678 -3694.742 0.03551637 0.0008342876
## 1 0 1 0 0 0 0 -3707.013 -3702.702 0.02027545 0.0009382657
```

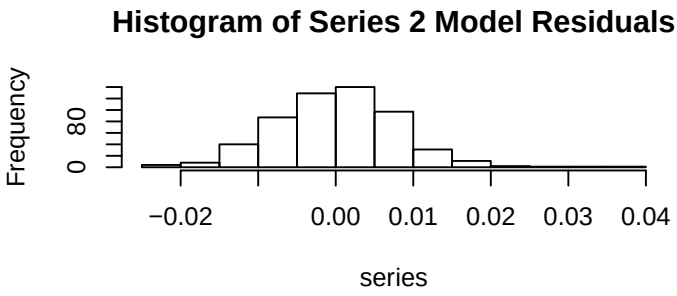
In this parameter search, an ARIMA(1,1,2)(0,0,0)_12 model is the best candidate in terms of both AIC and BIC, and a Ljung-Box test on its residuals fails to reject the null hypothesis of no correlation.

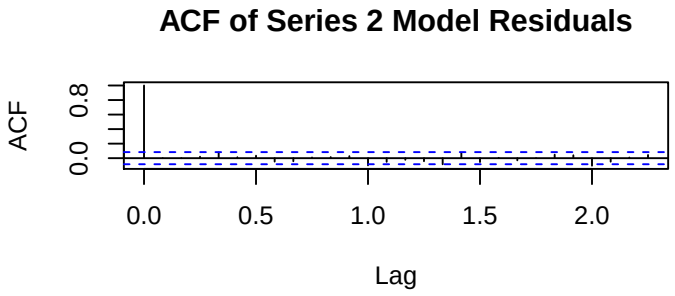
```
q2.mod.2<-Arima(log(series2),order=c(1,1,2),seasonal=list(order=c(0,0,0),12),method="ML");q2.mod.2
```

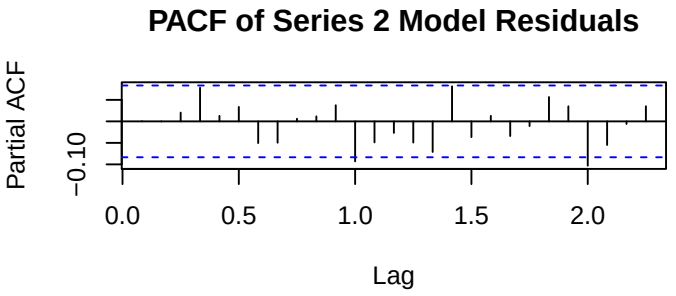
```
## Series: log(series2)
## ARIMA(1,1,2)
##
## Coefficients:
##          ar1      ma1     ma2
##       0.9994  -1.0992  0.1092
## s.e.  0.0010   0.0425  0.0423
##
## sigma^2 estimated as 6.358e-05:  log likelihood=1881.1
## AIC=-3754.2   AICc=-3754.13   BIC=-3736.95
```

```
tsplot(q2.mod.2$residuals, "Series 2 Model Residuals", 1, 1, 1)
```










```
## [1] "ADF Test P-Value for Series 2 Model Residuals : 0.01"
## [1] "Ljung-Box Test P-Value for Series 2 Model Residuals : 0.990658302014778"
## [1] "Shapiro-Wilk Test P-Value for Series 2 Model Residuals : 0.00134044490550966"
```

```
data.frame(sma.roots=abs(polyroot(c(1,q2.mod.2$coef[2],q2.mod.2$coef[3]))),
           phi1.CI=q2.mod.2$coef[1] + c(-2,2)*sqrt(q2.mod.2$var.coef)[1],
           theta1.CI=q2.mod.2$coef[2] + c(-2,2)*sqrt(q2.mod.2$var.coef)[5],
           theta2.CI=q2.mod.2$coef[3] + c(-2,2)*sqrt(q2.mod.2$var.coef)[9])
```

```
##   sma.roots   phi1.CI theta1.CI  theta2.CI
## 1  1.011436 0.9973925 -1.184145 0.02471444
## 2  9.052095 1.0014481 -1.014185 0.19373083
```

The series 2 model summary shows that each of the coefficients is statistically significant– their 95% confidence intervals, shown in the dataframe above, do not include zero. Since the coefficient on the AR term is 1, the characteristic equation for the AR process is (1-1B), which has a root > 1 and is therefore stationary. The roots of the characteristic equation for the MA component are also shown in the dataframe above– since they are also greater than 1, the MA process is invertible. The time plot, histogram, ACF plot, and PACF plot of the residuals resemble a realization of a white noise process, and as mentioned previously, the Box-Ljung test fails to reject the null hypothesis of no correlation in the residuals. Since the residuals resemble a white noise process, and since the stationarity and invertibility conditions are satisfied, this model can be used for forecasting.

Next, series 3 and 4 are modeled with a VAR model. The EDTSA indicated growing variance in series 3, so a log transform on that series is performed in order to stabilize the variance. The VARselect function is used to identify an optimal VAR modle order. Both a constant and a trend are included in the search because it was noted earlier that all series have both a trend and a non-zero mean.

```
VARselect(cbind(log(series3),series4), lag.max=8, type="both")
```

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      7      4      4      7
##
## $criteria
##                  1             2             3             4
## AIC(n) -1.054732e+01 -1.076261e+01 -1.083293e+01 -1.086512e+01
## HQ(n)  -1.052260e+01 -1.072553e+01 -1.078350e+01 -1.080333e+01
## SC(n)  -1.048410e+01 -1.066778e+01 -1.070649e+01 -1.070707e+01
## FPE(n)  2.626390e-05  2.117679e-05  1.973871e-05  1.911360e-05
##                  5             6             7             8
## AIC(n) -1.086404e+01 -1.087107e+01 -1.087594e+01 -1.086436e+01
```

```
## HQ(n)  -1.078988e+01 -1.078456e+01 -1.077707e+01 -1.075313e+01
## SC(n)  -1.067438e+01 -1.064980e+01 -1.062306e+01 -1.057987e+01
## FPE(n)  1.913444e-05  1.900051e-05  1.890841e-05  1.912886e-05
```

```
#VARselect(cbind(series1,series3,series4), lag.max=8, type="both")
```

The information criteria in the VARselect function indicate that a VAR(4) or a VAR(7) would be appropriate. Since there is an interest in forecasting, a more parsimonious VAR(4) model was first fitted– this model showed good performance, however, a Portmanteau test for this model rejected the null hypothesis of no autocorrelation in the residuals. Therefore, a VAR(7) model is selectedd to model series 3 and 4, which in addition to good performance also fails to reject the null hypothesis of no autocorrelation in its residuals (via the Portmanteau test).

```
var.fit<-VAR(cbind(log(series3),series4),p=7,type="both")
summary(var.fit)$varresult$log.series3.$coefficients
```

```
##                      Estimate   Std. Error    t value     Pr(>|t|)
## log.series3..l1  1.020272e+00 4.745378e-02 21.5003256 3.686988e-74
## series4.l1       8.292512e-03 1.306516e-03  6.3470433 4.723567e-10
## log.series3..l2  1.433789e-01 6.759406e-02  2.1211767 3.437139e-02
## series4.l2      -8.166040e-03 2.071042e-03 -3.9429629 9.130775e-05
## log.series3..l3 -6.635932e-02 6.784963e-02 -0.9780351 3.285040e-01
## series4.l3       2.728693e-03 2.119473e-03  1.2874396 1.985041e-01
## log.series3..l4 -4.328794e-02 6.764162e-02 -0.6399602 5.224755e-01
## series4.l4      -3.269053e-03 2.107564e-03 -1.5511052 1.214745e-01
## log.series3..l5 -6.816227e-02 6.786162e-02 -1.0044304 3.156304e-01
## series4.l5       5.299200e-04 2.109107e-03  0.2512533 8.017159e-01
## log.series3..l6  1.126513e-01 6.781135e-02  1.6612452 9.725671e-02
## series4.l6      -1.504305e-03 2.088185e-03 -0.7203887 4.716039e-01
## log.series3..l7 -1.324708e-01 4.547258e-02 -2.9132024 3.728656e-03
## series4.l7       2.011019e-03 1.341348e-03  1.4992520 1.344048e-01
## const            5.114057e-02 1.149404e-02  4.4493122 1.050868e-05
## trend            8.319448e-05 2.400343e-05  3.4659411 5.713180e-04
```

```
summary(var.fit)$varresult$series4$coefficients;round(roots(var.fit),2)
```

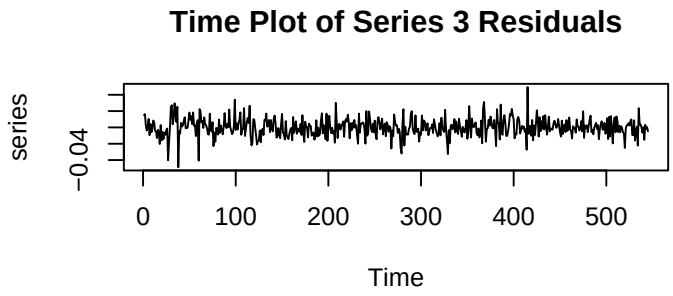
```
##                      Estimate   Std. Error    t value      Pr(>|t|)
## log.series3..l1  0.523526225 1.7416332643  0.3005950  7.638415e-01
## series4.l1       1.339800511 0.0479513212 27.9408466 3.270946e-106
## log.series3..l2  2.509423941 2.4808153564  1.0115319  3.122243e-01
## series4.l2      -0.404114322 0.0760107008 -5.3165451  1.562466e-07
## log.series3..l3 -4.331742422 2.4901952125 -1.7395192  8.252512e-02
## series4.l3       0.097531191 0.0777882078  1.2538043  2.104669e-01
## log.series3..l4  4.712426526 2.4825608260  1.8982119  5.821219e-02
## series4.l4      -0.073817875 0.0773511231 -0.9543220  3.403566e-01
## log.series3..l5 -4.581531998 2.4906351826 -1.8395034  6.640134e-02
## series4.l5      -0.115363205 0.0774077650 -1.4903312  1.367332e-01
## log.series3..l6  0.975659601 2.4887904294  0.3920216  6.952001e-01
## series4.l6       0.173111469 0.0766399079  2.2587641  2.430444e-02
## log.series3..l7 -0.603383242 1.6689199135 -0.3615412  7.178393e-01
## series4.l7      -0.025306667 0.0492297283 -0.5140525  6.074299e-01
## const            1.467174019 0.4218505518  3.4779474  5.468793e-04
## trend            0.004001784 0.0008809662  4.5424945  6.892183e-06
```

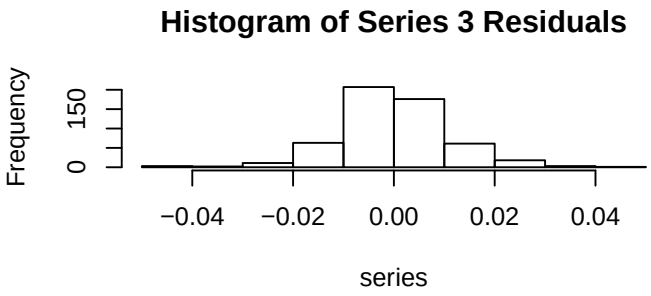
```
##  [1] 0.97 0.94 0.94 0.72 0.72 0.70 0.70 0.68 0.68 0.67 0.67 0.65 0.65 0.24
```

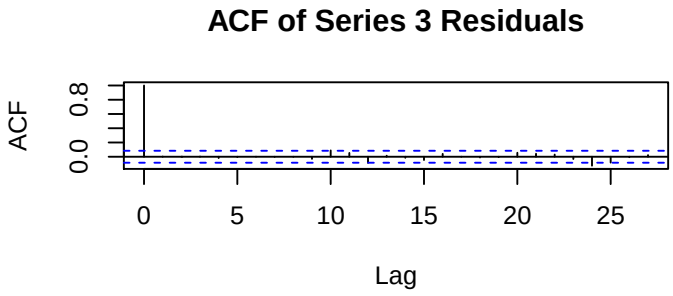
The model summary shows that the roots of the characteristic equation are all less than unity, indicating that the transformed model is stationary.

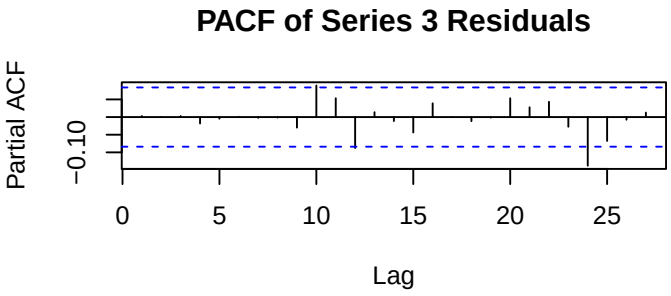
```
var.fit.residuals<-resid(var.fit);par(mfrow=c(2,1));for(a in 1:2){
  tsplot(var.fit.residuals[,a],paste("Series",a+2,"Residuals"),1,1,1)}
```

```
## Warning in adf.test(series): p-value smaller than printed p-value
```



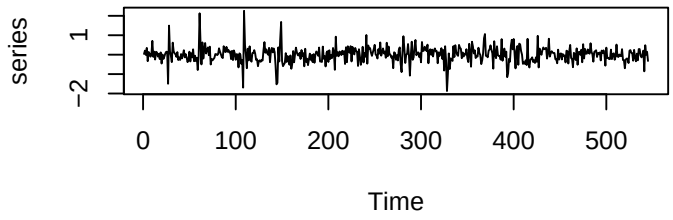


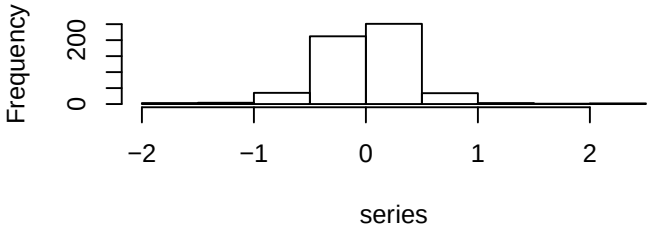


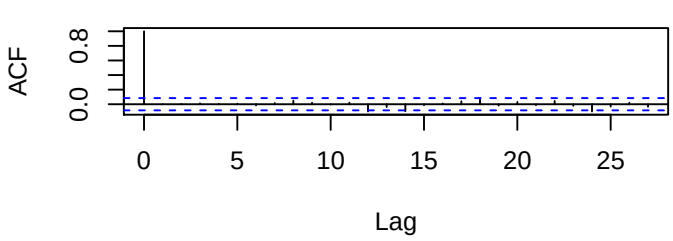


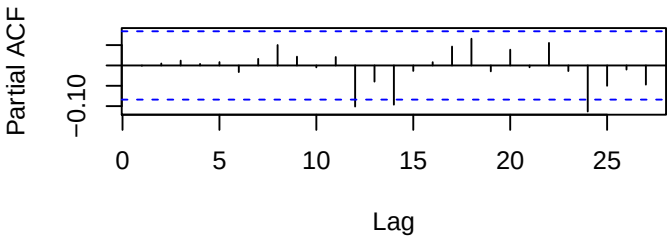

```
## [1] "ADF Test P-Value for Series 3 Residuals : 0.01"
## [1] "Ljung-Box Test P-Value for Series 3 Residuals : 0.943590709014302"
## [1] "Shapiro-Wilk Test P-Value for Series 3 Residuals : 7.10968224463641e-06"
```

```
## Warning in adf.test(series): p-value smaller than printed p-value
```










```
## [1] "ADF Test P-Value for Series 4 Residuals : 0.01"
## [1] "Ljung-Box Test P-Value for Series 4 Residuals : 0.983061928432045"
## [1] "Shapiro-Wilk Test P-Value for Series 4 Residuals : 3.22808156553588e-14"
```

```
serial.test(var.fit, lags.pt=12, type="PT.adjusted")
```

```
##
##  Portmanteau Test (adjusted)
##
```

```
## data:  Residuals of VAR object var.fit
## Chi-squared = 25.755, df = 20, p-value = 0.1741
```