[redacted]

IT3011 Theory and Practices in Statistical Modeling

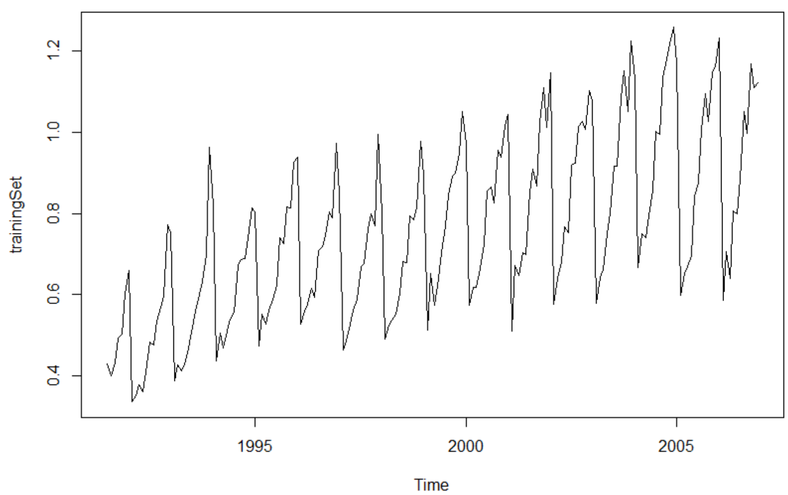
Take home Assignment IV – Time Series Analysis

1) Divide dataset into two sections as "training set" and "testing set". Use the "training set" to answer the following sections. Use the data up to December 2006 to represent the training set and the remaining set as the testing set.

```
install.packages("fpp")
library(fpp)
data(h02)
trainingSet = ts(trainingSet,frequency = 12,start=c(1991,7),end =c(2006,12))
trainingSet
testingSet = ts(testingSet,frequency = 12,start=c(2007,1),end =c(2008,6))
testingSet
```

2) Plot the time series. Interpret the behavior of the time series

```
plot(trainingSet)
```



- TrainingSet graph shows long term mean level change in between 1991 to 2006. (This is a upward trend). So we can say this graph has trend component. And, within one year period pattern of the graph has repeated over years. So pattern of the less than one year period has repeated over the time therefore we can say this graph has seasonal component too. We do not have enough evidence to say this graph has cyclical component. Because this graph doesn't show any drastic increases and decreases .and also because of we can't see any irregular pattern we can't say this has irregular component too.

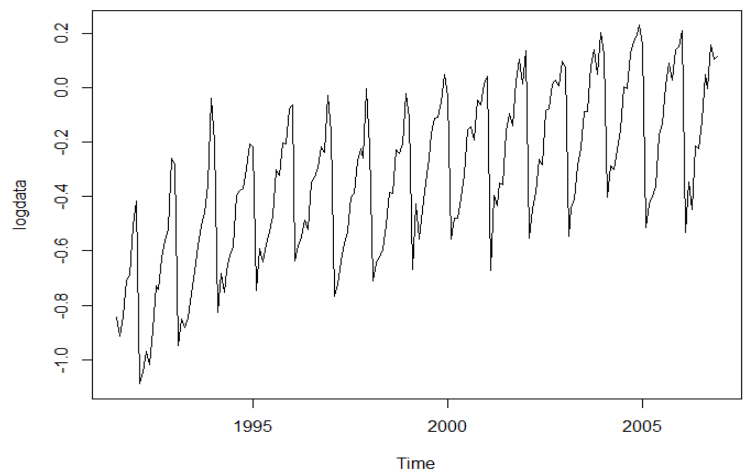
[redacted]

3) **Do the data need transforming in order to stabilize the variance? Give reasons for your answer. Apply a suitable transformation on the data if necessary**.

Yes.h02 graph shows long term variance change. So, variance is not constant in this graph. To stabilize the variance, need to perform log transformation.

logdata <- log(trainingSet)

plot(logdata)



4) **Test the stationary of data using kpss test? State the hypothesis you tested and your conclusion about stationary**

H1: Data is not stationary vs H0: Data is stationary

P value = 0.01

At 5% significance level p value is less than alpha (0.05). so reject HO. we can conclude that at 5% significance level data is not stationary

kpss.test(logdata)

[redacted]

5) i. If data are stationary, plot the ACF and time series plot.

ii. If data are not stationary, obtain the stationary series by applying differencing. Prove that the differenced series is stationary.

After performing kpss test we got to know that data is not stationary

ii. First have to remove seasonal component from the graph using below R code

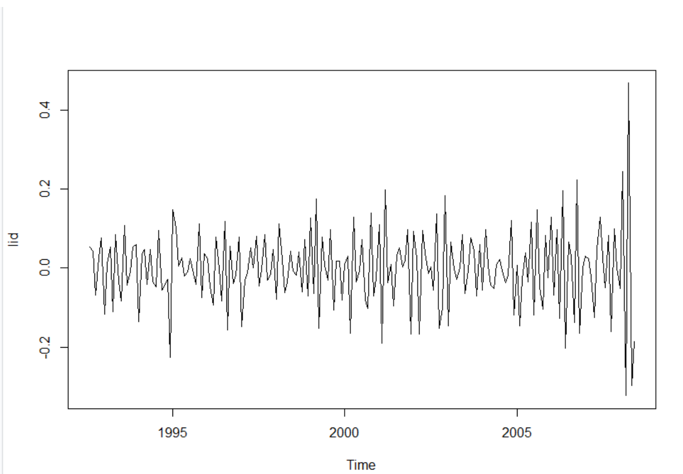
sdiff <- diff(logdata,12)

plot(sdiff)

Then using lag 1 differencing we can stabilize the mean of the data set using below R code.

lid <- diff(sdiff,1)

plot(lid)



(Graph after performing differencing)

Then perform kpss test to verify this graph is stationary or not.

H1: Data is not stationary vs H0: Data is stationary

P value = 0.1

At 5% significance level p value is greater than alpha (0.05). so, we do not reject HO. we can conclude that at 5% significance level data is stationary.

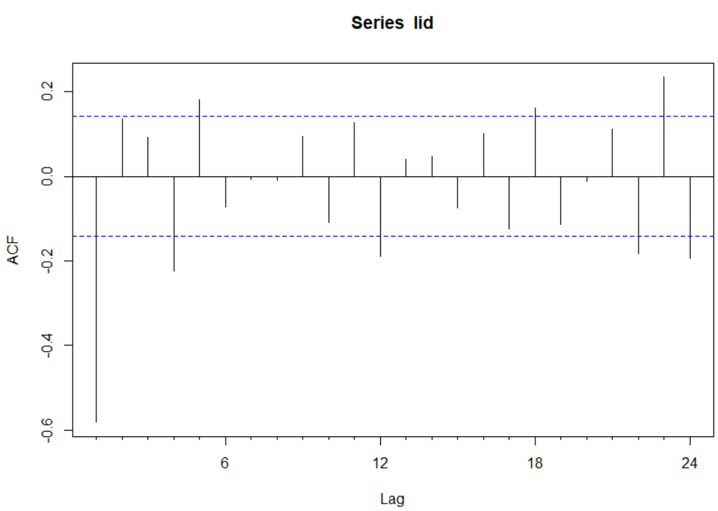
kpss.test(lid)

Now we can plot the ACF and time series using below R codes

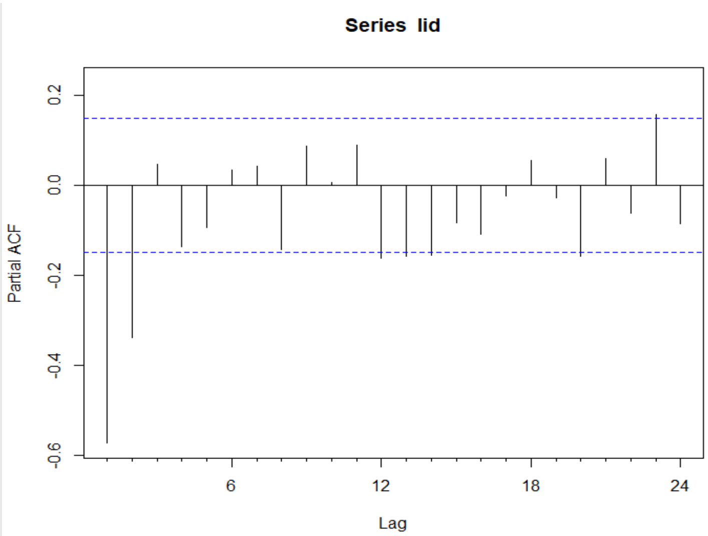
Acf(lid)

Pacf(lid)

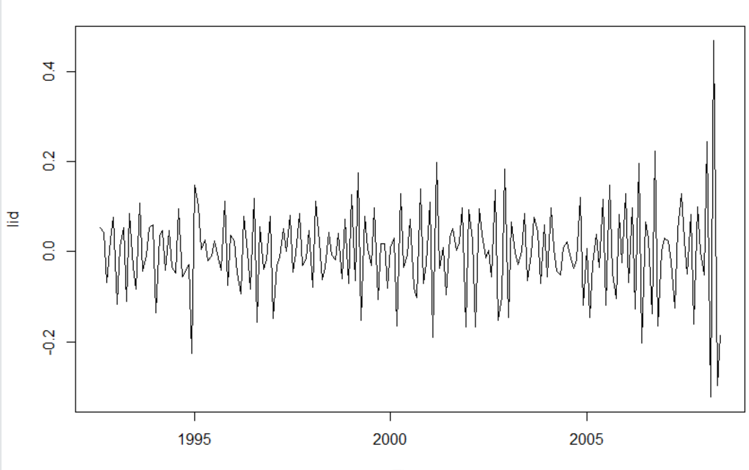
[redacted]



Series lid
ACF
0.2
0.0
-0.2
-0.4
-0.6
6
12
18
24
Lag



Series lid
Partial ACF
0.2
0.0
-0.2
-0.4
-0.6
6
12
18
24
Lag



lid
0.4
0.2
0.0
-0.2
1995
2000
2005
Time

[redacted]

6) **Identify a set of ARIMA models that might be useful in describing the time series. What is the best model? Give reasons for your selection.**

ARIMA (2,1,1)

ARIMA (0,1,1)

ARIMA (2,1,0)

```
fit1=Arima(trainingSet,order=c(2,1,1))
> summary(fit1)
Series: trainingSet
ARIMA(2,1,1)

Coefficients:
         ar1      ar2      ma1
      0.6858  -0.1198  -0.9456
s.e.  0.0747   0.0740   0.0195

sigma^2 estimated as 0.01912:  log likelihood=104.42
AIC=-200.83   AICc=-200.61   BIC=-187.95                Minimum value

Training set error measures:
                    ME      RMSE        MAE        MPE     MAPE      MASE
   ACF1
Training set 0.02108669 0.1367969 0.09441665 -0.1691392 13.5521 1.576804 -
0.04378112
>
>
>
> fit2=Arima(trainingSet,order=c(0,1,1))
> summary(fit2)
Series: trainingSet
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.1110
s.e.   0.0798

sigma^2 estimated as 0.02238:  log likelihood=89.44
AIC=-174.88   AICc=-174.81   BIC=-168.44

Training set error measures:
                     ME      RMSE        MAE       MPE     MAPE      MASE
    ACF1
Training set 0.004173037 0.1488049 0.09103613 -2.048868 13.81167
1.520348 0.007096728
>
>
>
> fit3=Arima(trainingSet,order=c(2,1,0))
> summary(fit3)
Series: trainingSet
ARIMA(2,1,0)

Coefficients:
          ar1      ar2
      -0.1009  -0.0755
s.e.   0.0731   0.0730
```

[redacted]

```
sigma^2 estimated as 0.02241:  log likelihood=89.83
AIC=-173.65   AICc=-173.52   BIC=-163.99

Training set error measures:
                      ME      RMSE        MAE       MPE     MAPE     MASE
     ACF1
Training set 0.004383123 0.1484893 0.09256621 -2.080551 13.97973 1.545901 -
0.003763627


>
```

Minimum BIC and AIC value is belongs to fit1 (ARIMA(2,1,1)). Therefore We can conclude that best model for this training set is ARIMA(2,1,1)

7) **Perform diagnostic testing on the residuals. Do the residuals resemble white noise? Give reasons for your answer.**

NO. when we perform Ljung Box Test to the fit 1 residuals p value = 2.2e-16

H0: Data are Independently distributed vs H1: Data are not independently distributed

At 5% significance level p value is less than 0.05. reject H0. Therefore, we can conclude that at 5% significance level data are not independently distributed. So, residuals do not resemble white noise

```
Box.test(residuals(fit1),lag=24,type="Ljung")

coeftest(fit1)

shapiro.test(residuals(fit1))
```

Box-Ljung test

data: residuals(fit1)

X-squared = 296.45, df = 24, p-value < 2.2e-16

output

[redacted]

8) Derive forecasts for the testing set. Plot the forecasts.

forecast(testingSet)

plot(forecast(testingSet))