Time Series Analysis and Forecasting - Exercise Set 2

Borja Ruiz

11 marzo 2018

Contents

1 Exercise 1

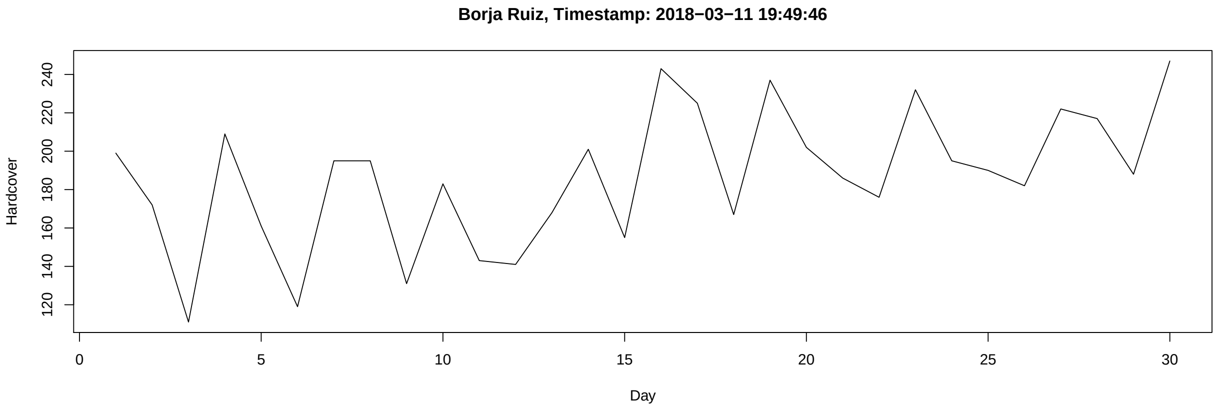
Data set books contains the daily sales of paperback and hardcover books at the same store. The task is to forecast the next four days' sales for hardcover books (data set books).

A) Plot the series and discuss the main features of the data.

```
library(fpp)
```

```
## Loading required package: forecast

## Loading required package: fma

## Loading required package: expsmooth

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: tseries
```

```
library(fma)
data("books")
par(mfrow = c(1, 1))
plot(books[,1], xlab = "Day", ylab = "Hardcover",main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```

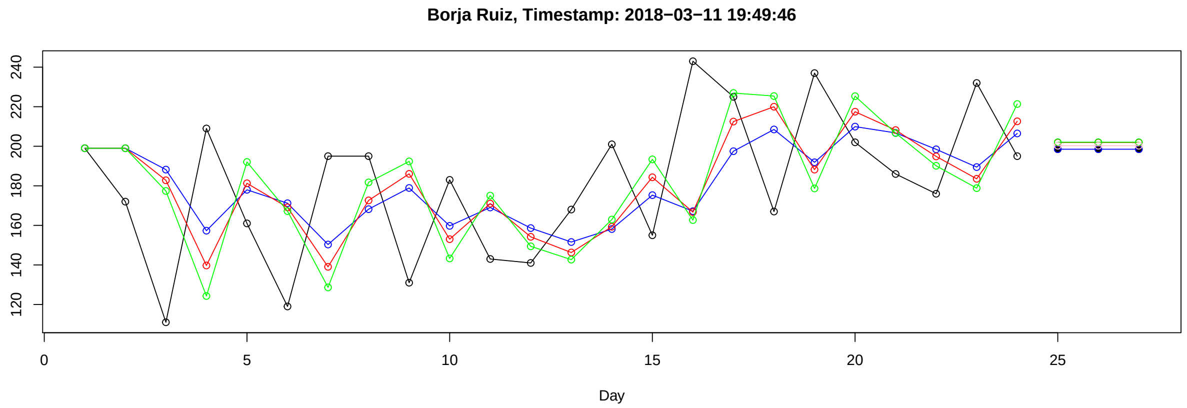


We can appreciate a positive trend a some possible cyclic behaviour in the data.

B) Use simple exponential smoothing with the ses function (setting initial = "simple") and explore different values of α for the paperback series. Record the within-sample SSE for the one-step forecasts. Plot SSE against α and find which value of α works best. What is the effect of α on the forecasts?

```
books1 <- window(books[,1], start = 1, end = 24)
bfit1 <- ses(books1, alpha = 0.2, initial = "simple", h = 3)
bfit2 <- ses(books1, alpha = 0.4, initial = "simple", h = 3)
bfit3 <- ses(books1, alpha=0.6, initial = "simple",h = 3)
bfit4 <- ses(books1, alpha=0.8, initial = "simple",h = 3)
```

```r
plot(bfit1, PI=FALSE, ylab="",
     xlab="Day", main=paste("Borja Ruiz, Timestamp:",Sys.time()), fcol=1, type="o")
lines(fitted(bfit2), col="blue", type="o")
lines(fitted(bfit3), col="red", type="o")
lines(fitted(bfit4), col="green", type="o")
lines(bfit1$mean, col="blue", type="o")
lines(bfit2$mean, col="red", type="o")
lines(bfit3$mean, col="green", type="o")
lines(bfit4$mean, col="pink", type="o")
```



```r
books1 <- window(books[,1], start = 25, end = 30)
a1<-accuracy(bfit1, books1);a1
```

```
##                      ME     RMSE      MAE       MPE      MAPE      MASE
## Training set -0.1021140 36.80078 30.18328 -4.307136 18.125211 0.6914496
## Test set     -0.5098526 17.28949 16.16995 -0.989690  8.043788 0.3704272
##                    ACF1 Theil's U
## Training set -0.1030847        NA
## Test set     -0.2857143 0.6952977
```

```r
a2<-accuracy(bfit2, books1);a2
```

```
##                      ME     RMSE      MAE       MPE      MAPE      MASE
## Training set  0.3011537 37.98713 32.75275 -3.861192 19.359110 0.7503120
## Test set     -3.8910760 17.71460 17.29703 -2.709850  8.748566 0.3962466
##                    ACF1 Theil's U
## Training set -0.2793339        NA
## Test set     -0.2857143  0.680183
```

```r
a3<-accuracy(bfit3, books1);a3
```

```
##                      ME     RMSE      MAE       MPE      MAPE      MASE
## Training set  0.2116743 40.87900 35.37428 -3.928558 20.744549 0.8103669
## Test set     -4.0481100 17.74976 17.34937 -2.789739  8.781298 0.3974457
##                    ACF1 Theil's U
## Training set -0.3797623        NA
## Test set     -0.2857143 0.6799454
```

```r
a4<-accuracy(bfit4, books1);a4
```

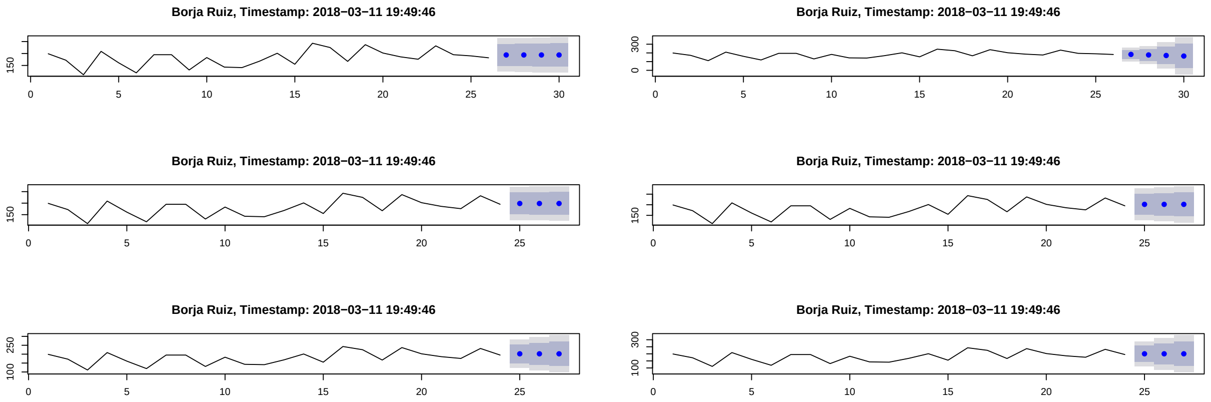
```
##                         ME     RMSE      MAE       MPE     MAPE      MASE
## Training set  0.06630559 44.78986 38.20703 -4.069188 22.20855 0.8752607
## Test set     -2.27306735 17.43082 16.75769 -1.886706  8.41131 0.3838913
##                    ACF1 Theil's U
## Training set -0.4470427        NA
## Test set     -0.2857143 0.6850543
```

$\alpha = 0.8$ seems to be the one to improve the accuracy.

It seems the higher the alpha the smoother the time series seems, therefore high and low peaks disappear progressively, and differente features in the data are appreciated.

C) Now let ses select the optimal value of α. Use this value to generate forecasts for the next four days. Compare your results with (b).

```
books1 <- window(books[,1], start = 1, end = 26)
fit1 <- ses(books1, initial = "simple", h = 4)
fit2 <- holt(books1, initial = "simple", h = 4)
par(mfrow = c(3,2))
plot(fit1, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
plot(fit2, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
plot(bfit1, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
plot(bfit2, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
plot(bfit3, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
plot(bfit4, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



Borja Ruiz, Timestamp: 2018−03−11 19:49:46

Borja Ruiz, Timestamp: 2018−03−11 19:49:46

Borja Ruiz, Timestamp: 2018−03−11 19:49:46

Borja Ruiz, Timestamp: 2018−03−11 19:49:46

Borja Ruiz, Timestamp: 2018−03−11 19:49:46

Borja Ruiz, Timestamp: 2018−03−11 19:49:46

Since we have used Holt's linear trend method we are able to perceive that our predictions now have adquired a slight negative trend. We are also able to make predictions near in time more precisely, though our confidence intervals grow bigger the more h increases.

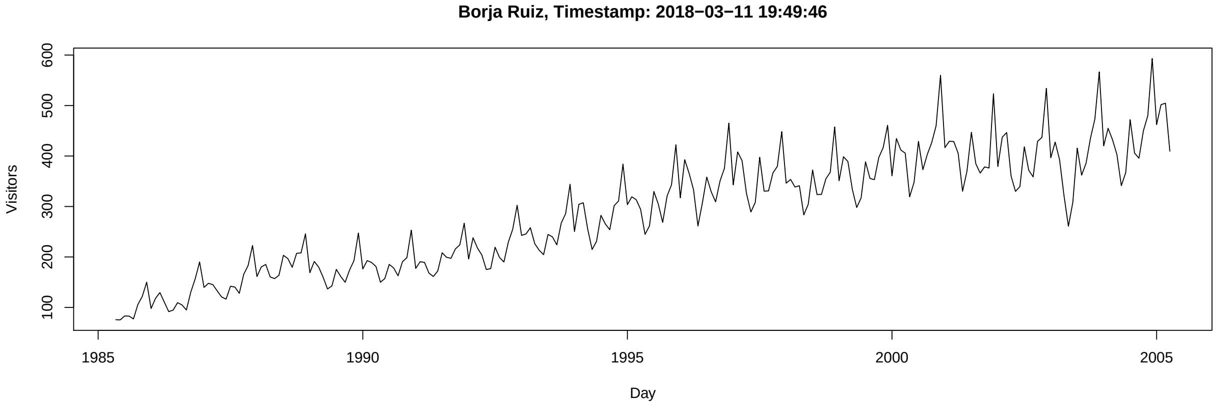
Relative to the ses choosing the value of α, we appreciate that the forecasting produced would seem to have approximately a $\alpha = 0.3$.

2 Exercise 2

Use the monthly Australian short-term overseas visitors data, May 1985-April 2005. (Data set: visitors)

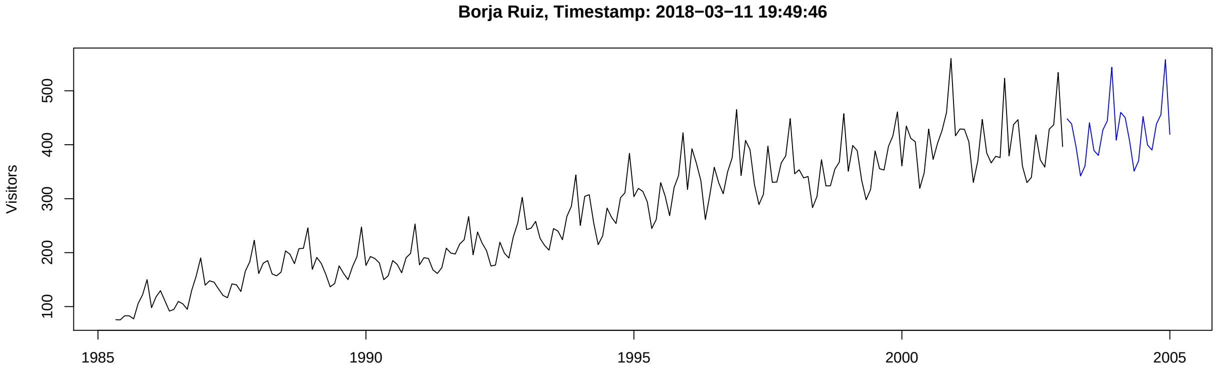
A) Make a time plot of your data and describe the main features of the series.

```
data("visitors")
par(mfrow = c(1, 1))
plot(visitors, xlab = "Day", ylab = "Visitors",main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



B) Forecast the next two years using Holt-Winters' multiplicative method.

```
visitors1<-window(visitors,end=2003)
fit6 <- hw(visitors1, h=24, seasonal = "multiplicative")
par(mfrow = c(1,1))
plot(fit6, ylab="Visitors", main=paste("Borja Ruiz, Timestamp:",Sys.time()), flwd=1, PI=FALSE)
```



C) Why is multiplicative seasonality necessary here?

Multiplicative method is necessary since we observe a growth in the seasonal variation in the data along time, and this is best addressed with multiplicative methods.

D) Experiment with making the trend exponential and/or damped.

```
fit7 <- holt(visitors1,h=24)
fit8 <- holt(visitors1,h=24, exponential = TRUE)
fit9 <- holt(visitors1,h=24, damped = TRUE)
fit10 <- holt(visitors1,h=24, exponential = TRUE, damped = TRUE)

par(mfrow=c(2,2))
accuracy(fit8, window(visitors, start = 2003))
```

```
##                        ME     RMSE      MAE         MPE     MAPE     MASE
## Training set   2.066836 41.04792 31.22731  -0.2278252 11.59589 1.202569
## Test set     -35.586793 75.72756 58.35689 -11.6835362 15.73868 2.247334
##                   ACF1 Theil's U
## Training set 0.2208060        NA
## Test set     0.4460329  1.278936
```

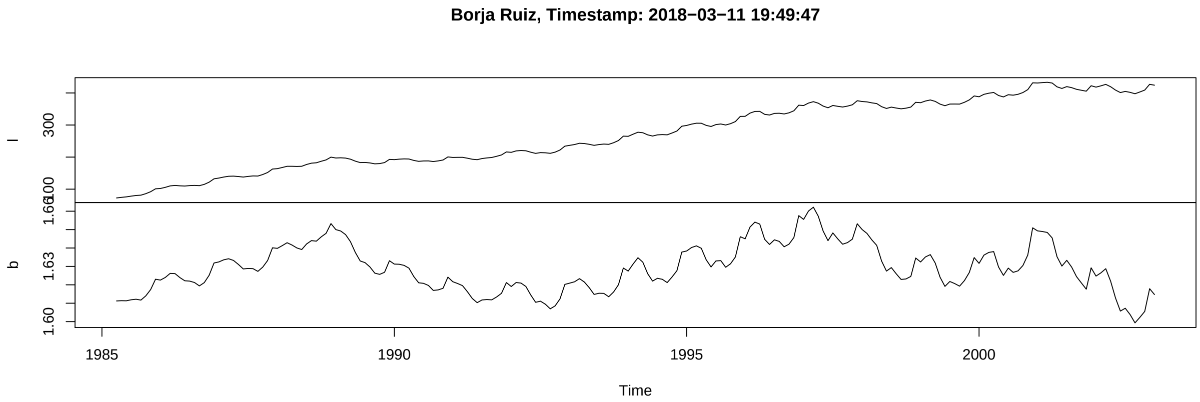
```r
accuracy(fit9, window(visitors, start = 2003))
```

```
##                       ME     RMSE      MAE       MPE     MAPE     MASE
## Training set  2.877313 41.14654 31.47813 -1.301350 11.94075 1.212228
## Test set     -7.691850 73.28643 54.90297 -5.128036 14.17644 2.114323
##                   ACF1 Theil's U
## Training set 0.1775351        NA
## Test set     0.5316251  1.191356
```

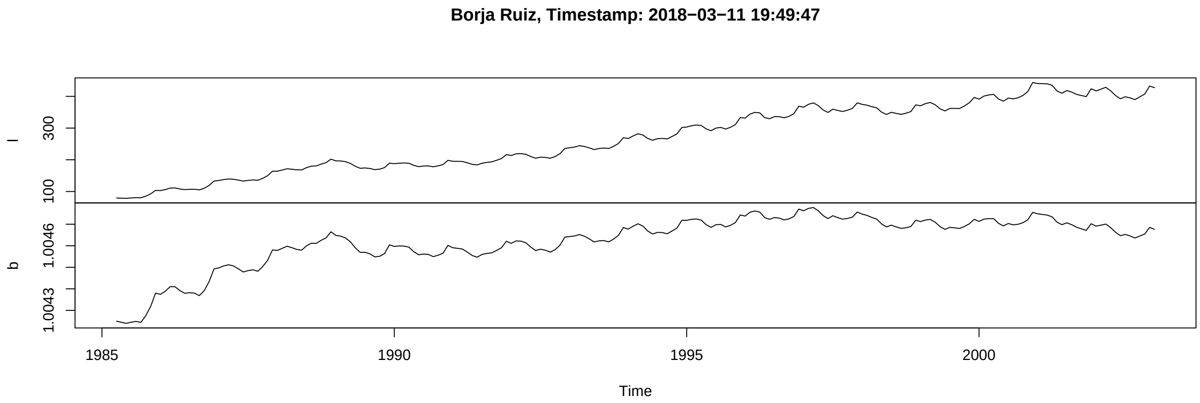
```r
accuracy(fit10, window(visitors, start = 2003))
```

```
##                        ME     RMSE      MAE       MPE     MAPE     MASE
## Training set   4.817694 41.23793 31.33221  0.100957 11.62430 1.206609
## Test set     -10.801875 73.72285 55.27896 -5.898647 14.36921 2.128802
##                   ACF1 Theil's U
## Training set 0.1423545        NA
## Test set     0.5321676  1.207447
```

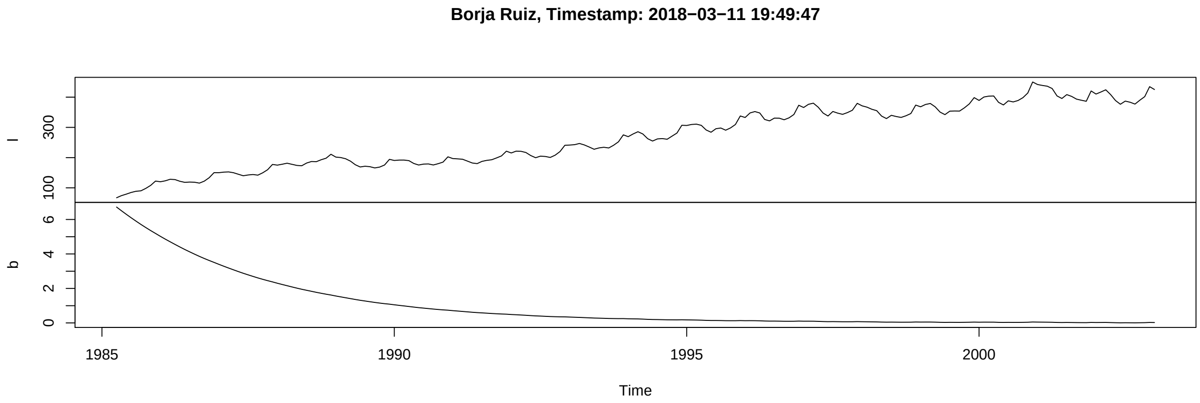
```r
par(mfrow=c(2,2))
plot(fit7$model$state, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



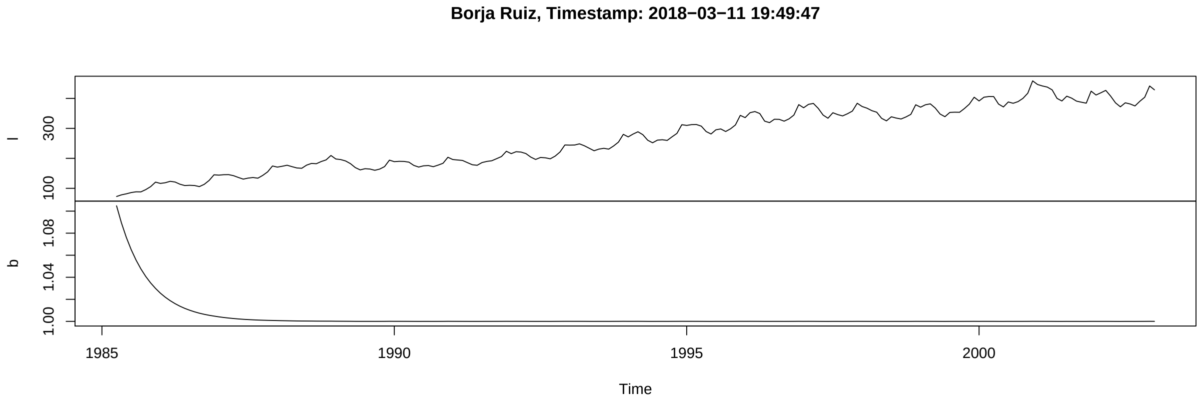
```r
plot(fit8$model$state, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



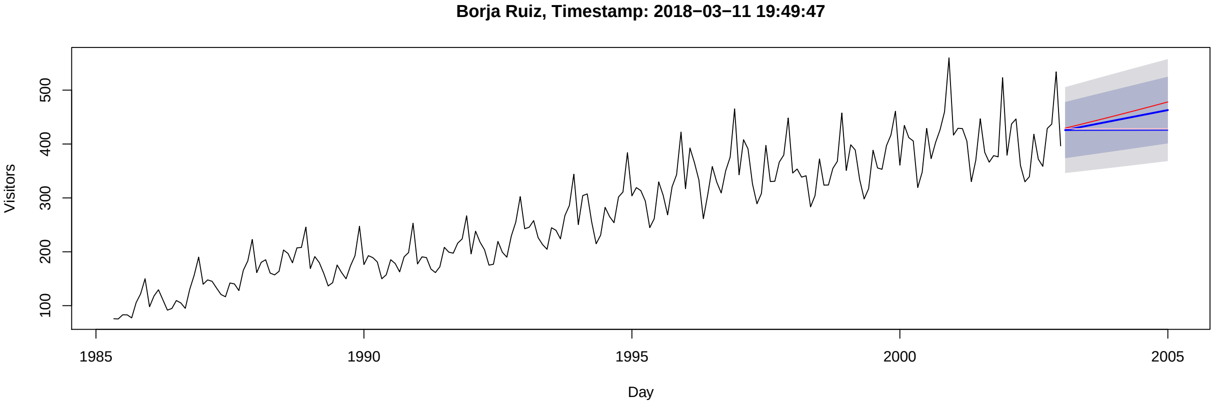
```r
plot(fit9$model$state, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



```r
plot(fit10$model$state, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



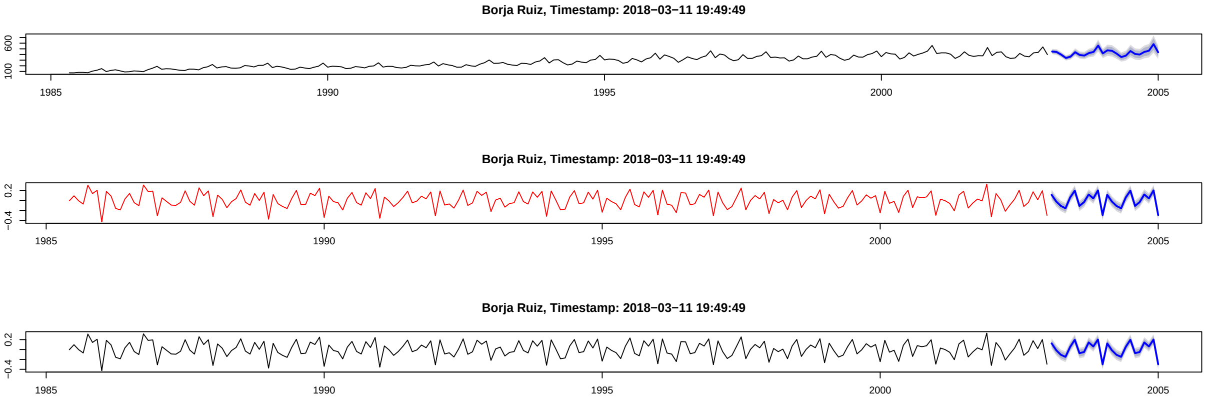
```r
plot(fit7, xlab = "Day", ylab = "Visitors",main=paste("Borja Ruiz, Timestamp:",Sys.time()))
lines(fit8$mean, col = "red")
lines(fit9$mean, col = "blue")
lines(fit10$mean, col = "pink")
```


Borja Ruiz, Timestamp: 2018−03−11 19:49:47

We are able to appreciate how the slope of the trends gets modified.

E) Now fit each of the following models to the same data: 1. an ETS model 2. an additive ETS model applied to a Box-Cox transformed series 3. an STL decomposition applied to the Box-Cox transformed data followed by an ETS model applied to the seasonally adjusted (transformed) data.
Plot all the forecasts together.

```
par(mfrow=c(3,1))

vdata <- window(visitors, end = 2003)
fit1 <- ets(vdata)
plot(forecast(fit1), main=paste("Borja Ruiz, Timestamp:",Sys.time()))


vdatadiff <- window(diff(log(visitors)), end = 2003)
fit2<- ets(vdatadiff)
plot(forecast(fit2), col="red", main=paste("Borja Ruiz, Timestamp:",Sys.time()))

fit3 <- stl(vdatadiff, t.window = 50, s.window="periodic", robust=TRUE)
plot(forecast(fit3), main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```


Borja Ruiz, Timestamp: 2018−03−11 19:49:49

Borja Ruiz, Timestamp: 2018−03−11 19:49:49

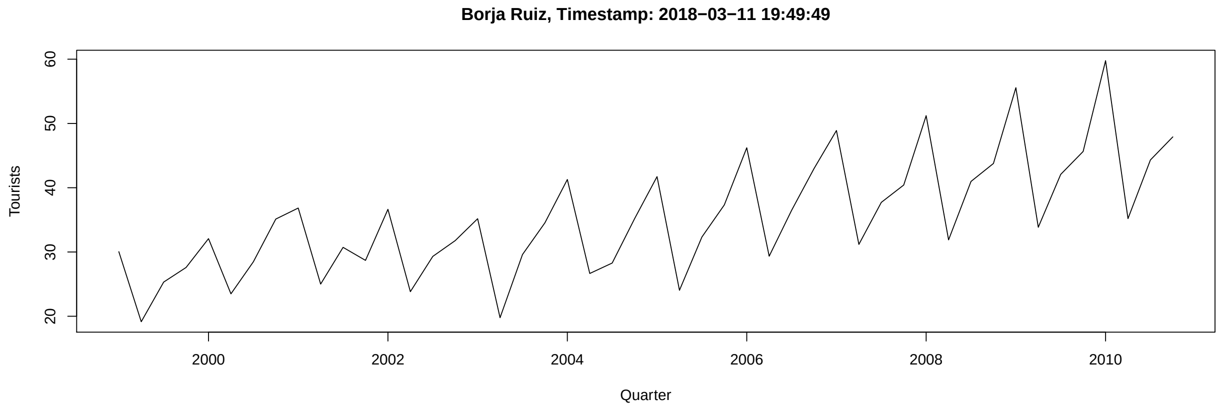
Borja Ruiz, Timestamp: 2018−03−11 19:49:49

3 Exercise 3

Consider the quarterly number of international tourists to Australia for the period 1999-2010. (Data set austourists.)

A) Describe the time plot.

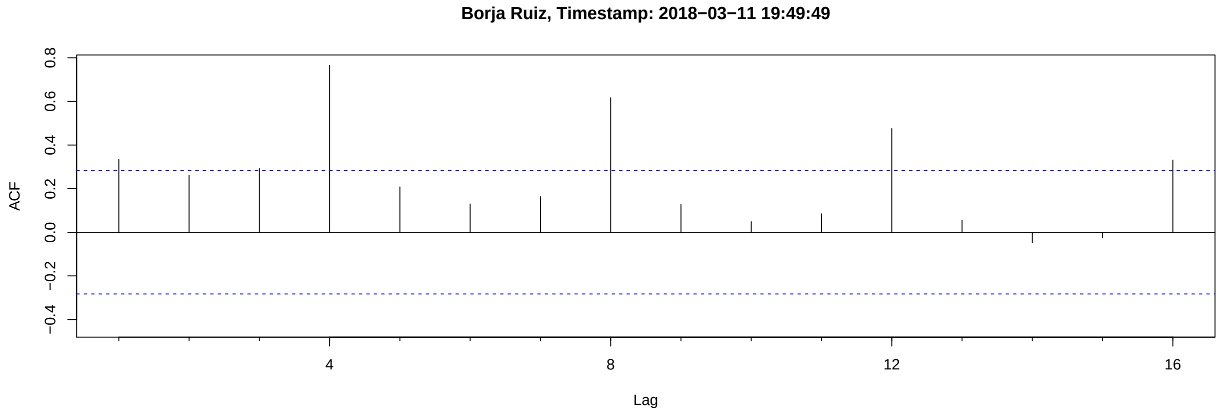
```r
data("austourists")
par(mfrow = c(1, 1))
plot(austourists, xlab = "Quarter", ylab = "Tourists", main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



The time series shows a seasonality pattern within a positive trend.

B) What can you learn from the ACF graph?

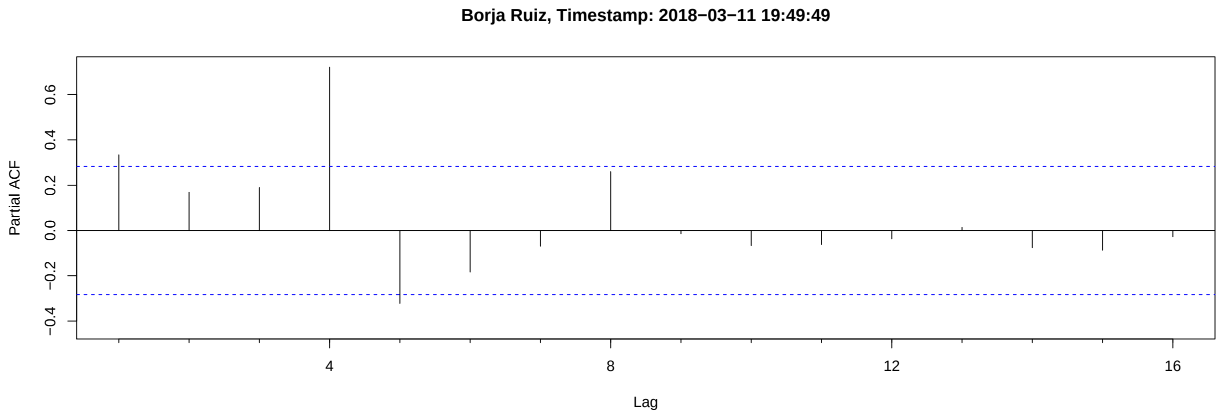
```r
Acf(austourists, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



Since we can observe high autocorrelations in seasonality and in the trend we assume that the dataset is not stationary. We can see that we can choose 4 possible values of m for the AR model, 4, 8, 12 and 16.

C) What can you learn from the PACF graph?

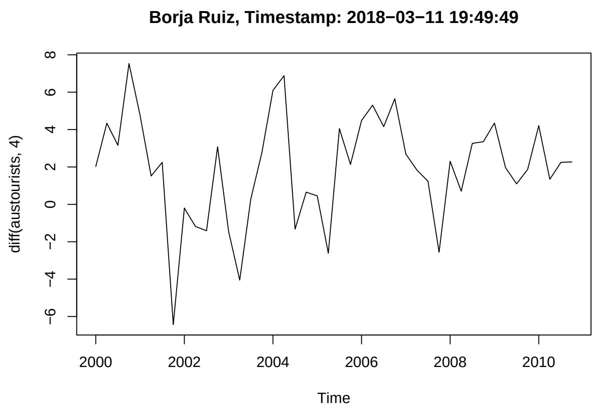
```r
Pacf(austourists, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



After eliminating the partial correlations we can infer that the optimum lag value is $m = 4$ in a AR(1) model since we have $p = 1$.

D) Produce plots of the seasonally differenced data $(1 - B^4)Y_t$.

```r
par(mfrow = c(1,2))
plot(diff(austourists, 4), main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



What model do these graphs suggest?

We confirm that $m = 4$ is a good value since the graph seems a lot more stationary and $D = 1$ could be a possible value.

E) Does auto.arima give the same model that you chose? If not, which model do you think is better?

```r
auto.arima(austourists)
```

```
## Series: austourists
## ARIMA(1,0,0)(1,1,0)[4] with drift
##
## Coefficients:
##          ar1     sar1   drift
##       0.4493  -0.5012  0.4665
## s.e.  0.1368   0.1293  0.1055
##
## sigma^2 estimated as 5.606:  log likelihood=-99.47
## AIC=206.95   AICc=207.97   BIC=214.09
```

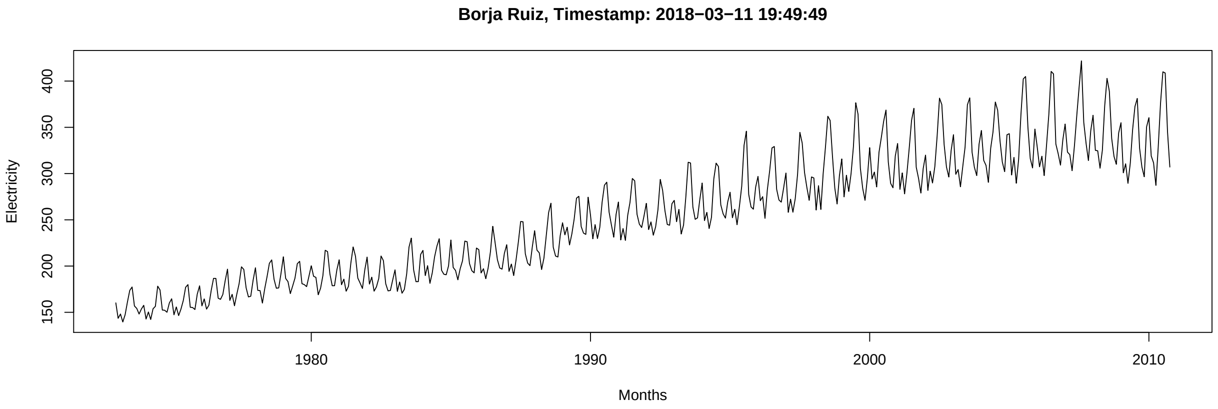
Auto arima chose the same model as me.

4 Exercise 4

Consider the total net generation of electricity (in billion kilowatt hours) by the U.S. electric industry (monthly for the period 1985-1996). (Data set usmelec.) In general there are two peaks per year: in mid-summer and mid-winter.

A) Examine the 12-month moving average of this series to see what kind of trend is involved.

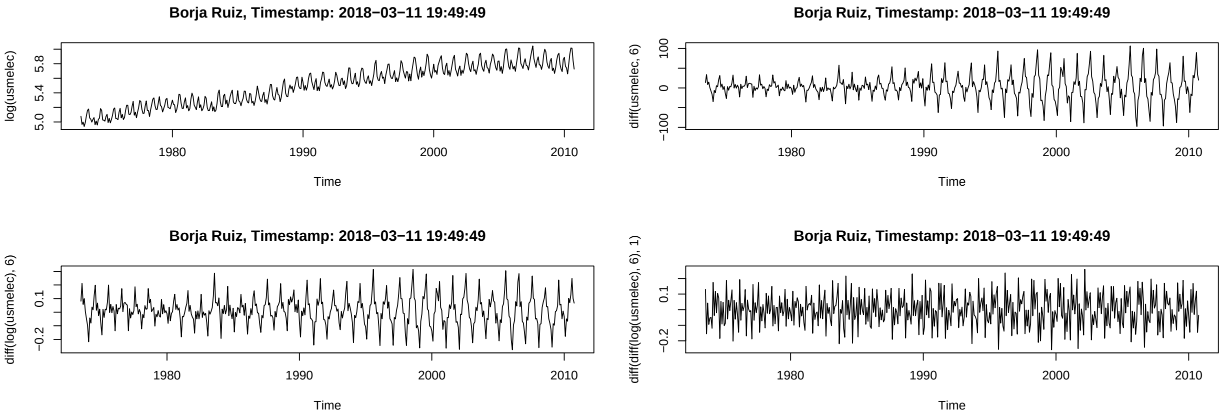
```r
data("usmelec")
par(mfrow = c(1, 1))
plot(usmelec, xlab = "Months", ylab = "Electricity",main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



We appreciate a positive trend and a seasonal variation.

B) Do the data need transforming? If not, find an appropriate differencing which yields stationary data.

```r
par(mfrow=c(2,2))
plot(log(usmelec), main=paste("Borja Ruiz, Timestamp:",Sys.time()))
plot(diff(usmelec,6), main=paste("Borja Ruiz, Timestamp:",Sys.time()))
plot(diff(log(usmelec),6), main=paste("Borja Ruiz, Timestamp:",Sys.time()))
plot(diff(diff(log(usmelec),6),1), main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```

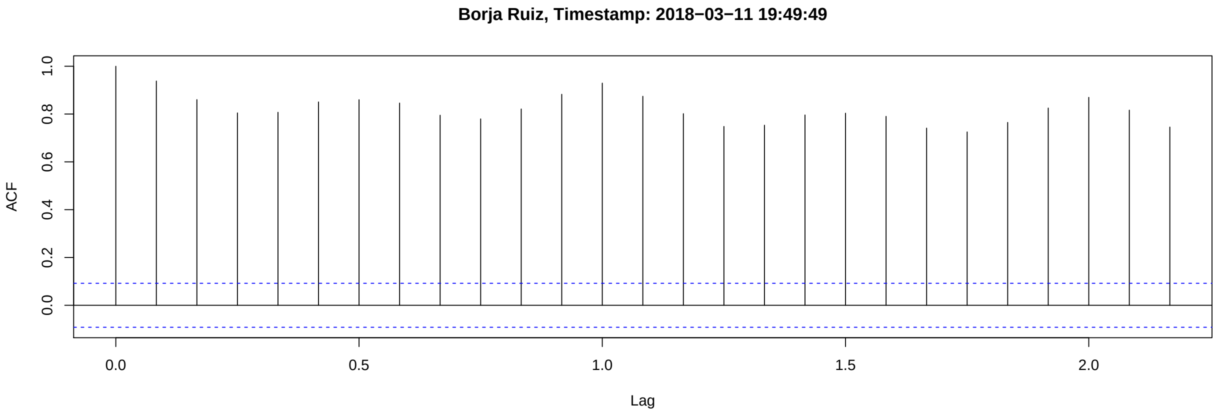


To transform the data we first demonstrate that using the log function minimizes the seasonal variance. Furthermore applying the difference between intervals will eliminate the seasonality. 6 is a good value for the difference stimation since we appreciate peaks every 6 months in the original data.

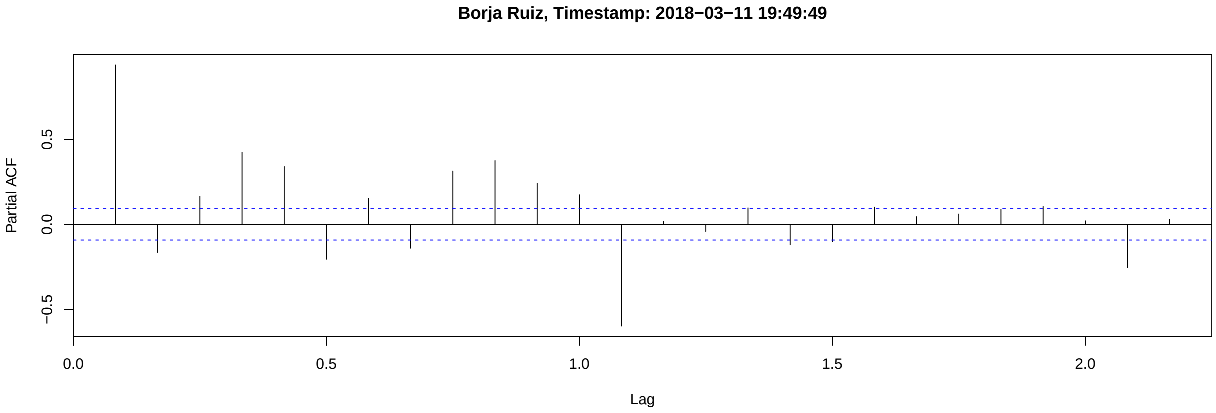
We finally applied both transformations to obtain a stationary structure from our original data. We can also apply another difference transformation to obtain white noise.

C) Identify a couple of ARIMA models that might be useful in describing the time series. Which of your models is the best according to their AICc values?

```
acf(usmelec, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



```
pacf(usmelec, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```



```
Arima(usmelec, order = c(2,2,0), seasonal = c(2,2,0))
```

```
## Series: usmelec
## ARIMA(2,2,0)(2,2,0)[12]
##
## Coefficients:
##           ar1      ar2     sar1     sar2
##       -0.8179  -0.5338  -0.9360  -0.4213
## s.e.   0.0408   0.0411   0.0441   0.0447
##
## sigma^2 estimated as 204.9:  log likelihood=-1751.94
## AIC=3513.87   AICc=3514.02   BIC=3534.17
```

```r
Arima(usmelec, order = c(1,1,0), seasonal = c(1,1,0))
```

```
## Series: usmelec
## ARIMA(1,1,0)(1,1,0)[12]
##
## Coefficients:
##           ar1     sar1
##       -0.2828  -0.4631
## s.e.   0.0458   0.0425
##
## sigma^2 estimated as 81.77:  log likelihood=-1597.31
## AIC=3200.62   AICc=3200.68   BIC=3212.89
```

```r
Arima(usmelec, order = c(2,1,0), seasonal = c(2,1,0))
```

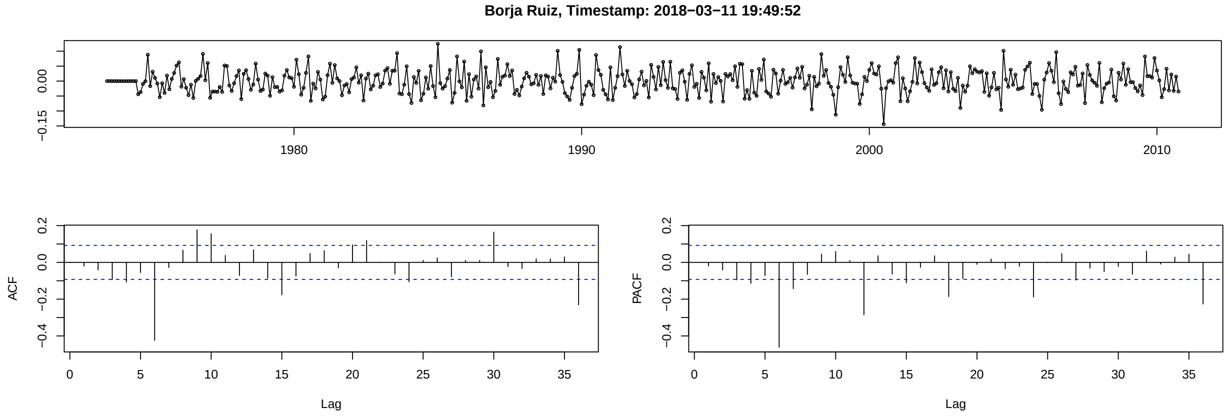
```
## Series: usmelec
## ARIMA(2,1,0)(2,1,0)[12]
##
## Coefficients:
##           ar1      ar2     sar1     sar2
##       -0.3700  -0.2740  -0.5778  -0.2878
## s.e.   0.0461   0.0465   0.0471   0.0469
##
## sigma^2 estimated as 68.26:  log likelihood=-1557.5
## AIC=3125   AICc=3125.14   BIC=3145.45
```

Since we can observe in the ACF and PACF graphs two possible values for m, we decide to play with values of $p = 1, 2$. Finally our best arima model will be A(2,1,0)(2,1,0). We now try this value with our transformed data:

```r
fitarima <- Arima(diff(log(usmelec),6), order = c(2,1,0), seasonal = c(2,1,0))
```

D) Estimate the parameters of your best model and do diagnostic testing on the residuals. Do the residuals resemble white noise? If not, try to find another ARIMA model which fits better.

```r
res <- residuals(fitarima)
tsdisplay(res, main=paste("Borja Ruiz, Timestamp:",Sys.time()))
```


Borja Ruiz, Timestamp: 2018−03−11 19:49:52
ACF
Lag
PACF
Lag