Time Series Analysis and Forecasting -Exercise Set 2

Francesca Sallicati

10 marzo 2018

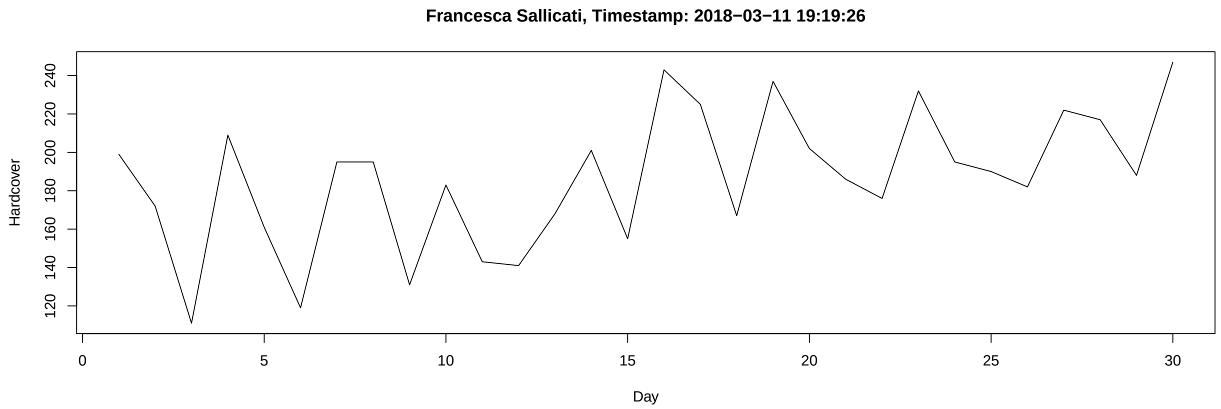
Contents

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.

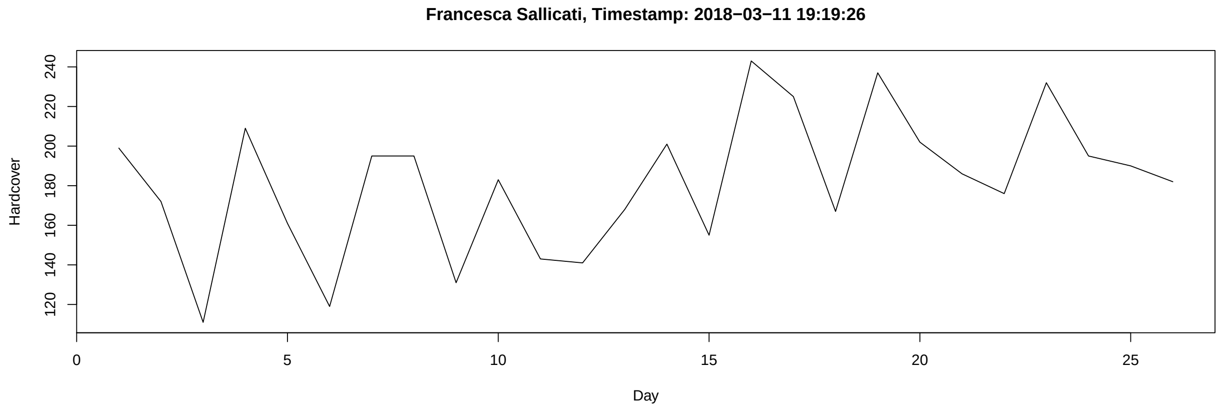
```
data("books")
par(mfrow = c(1, 1))
plot(books[,1], xlab = "Day", ylab = "Hardcover",main=paste("Francesca Sallicati, Timestamp:",Sys.time(
```



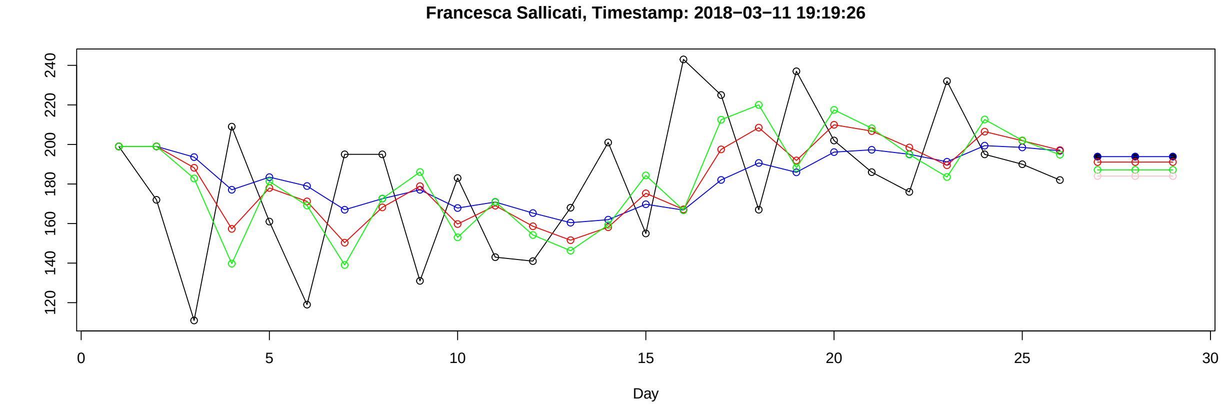
This time series of hardcover books shows a positive trend and a potential cyclic pattern.

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 ?? on the forecasts?

```
books1 <- window(books[,1], start = 1, end = 26)
plot(books1, ylab = "Hardcover", xlab = "Day", main=paste("Francesca Sallicati, Timestamp:",Sys.time())
```



```r
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)
plot(bfit1, PI=FALSE, ylab="",
     xlab="Day",  main=paste("Francesca Sallicati, Timestamp:",Sys.time()), fcol=1, type="o")
lines(fitted(bfit1), col="blue", type="o")
lines(fitted(bfit2), col="red", type="o")
lines(fitted(bfit3), 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")
```



```
##                      ME     RMSE      MAE       MPE      MAPE      MASE
## Training set -0.9910951 35.51534 28.75832 -4.461013 17.216159 0.7069401
## Test set     15.1536943 21.31434 19.05123  6.747337  8.820496 0.4683194
##                     ACF1 Theil's U
## Training set -0.09762906        NA
## Test set     -0.09495549 0.7948302
```

```
##                      ME     RMSE      MAE       MPE      MAPE      MASE
## Training set -0.7614628 36.69153 31.27276 -4.124723 18.430494 0.7687502
## Test set     17.9192126 23.36161 19.97307  8.077731  9.170209 0.4909801
##                     ACF1 Theil's U
## Training set -0.26738666        NA
## Test set     -0.09495549 0.8678334
```

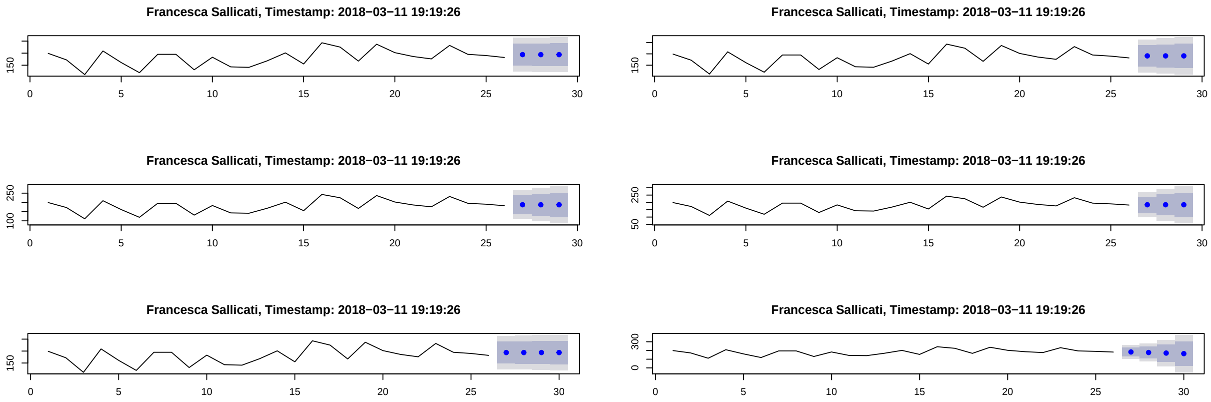
```
##                     ME     RMSE      MAE       MPE      MAPE      MASE
## Training set -0.761045 39.42652 33.60962 -4.141155 19.663608 0.8261951
## Test set     21.872302 26.51536 21.87230  9.979424  9.979424 0.5376672
##                     ACF1 Theil's U
## Training set -0.36793260        NA
## Test set     -0.09495549 0.9933209
```

```
##                      ME     RMSE      MAE       MPE     MAPE      MASE
## Training set -0.7206287 43.12494 36.04986 -4.176612 20.92063 0.8861815
## Test set     24.9890773 29.13967 24.98908 11.478795 11.47880 0.6142841
##                     ACF1 Theil's U
## Training set -0.43761758        NA
## Test set     -0.09495549   1.10483
```

Here $\alpha = 0.8$ works better and it seems like that greater the values of α the smoother gets the times series,

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)
par(mfrow = c(3,2))
plot(bfit1,main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
plot(bfit2,main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
plot(bfit3,main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
plot(bfit4,main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
fit2<-holt(books1, initial = "simple", h = 4)
plot(fit1,main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
plot(fit2,main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
```



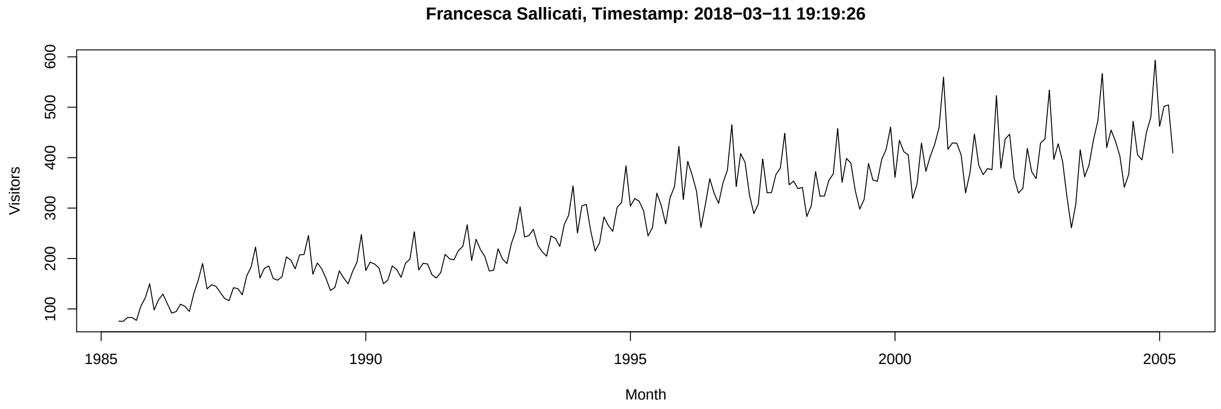
We can see that ses (plot (3,1)) select an α between 0.2 and 0.4, however it is not able to capture the trend and the prediction intervals are quite large. While using the Holt's method (plot (3,2)) the predictions are more accurate with narrower intervals and shows a slightly negative trend, therefore is more similar to the result obtained in the previous point.

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.

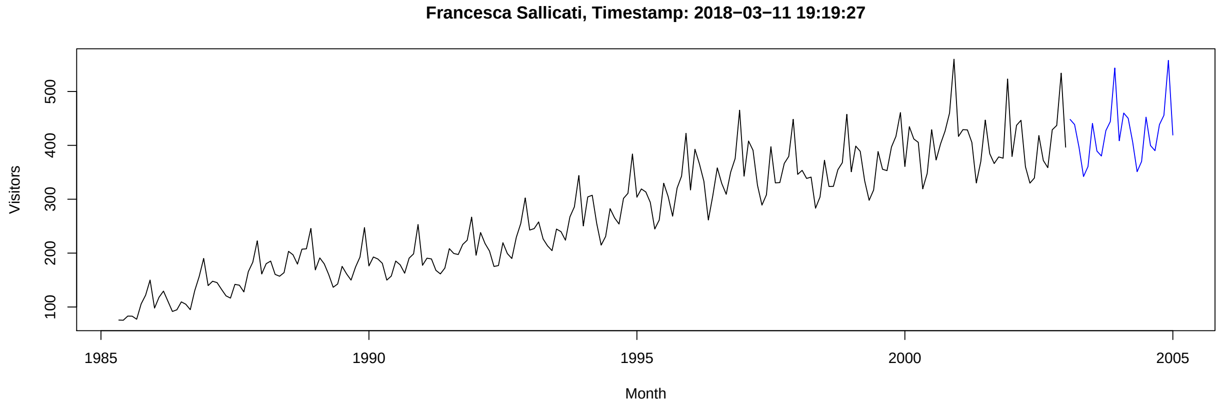
```
data("visitors")
par(mfrow = c(1, 1))
plot(visitors, xlab = "Month", ylab = "Visitors",main=paste("Francesca Sallicati, Timestamp:",Sys.time(
```



The time series of monthly overseas visitors in Australia clearly shows seasonality with a positive trend.

(*b*) Forecast the next two years using Holt-Winters' multiplicative method.

```
visitors1<-window(visitors,end=2003)
fit5 <- hw(visitors1, h=24, seasonal="multiplicative")
plot(fit5, ylab="Visitors",xlab="Month", main=paste("Francesca Sallicati, Timestamp:",Sys.time()),
     flwd=1, PI=FALSE)
```



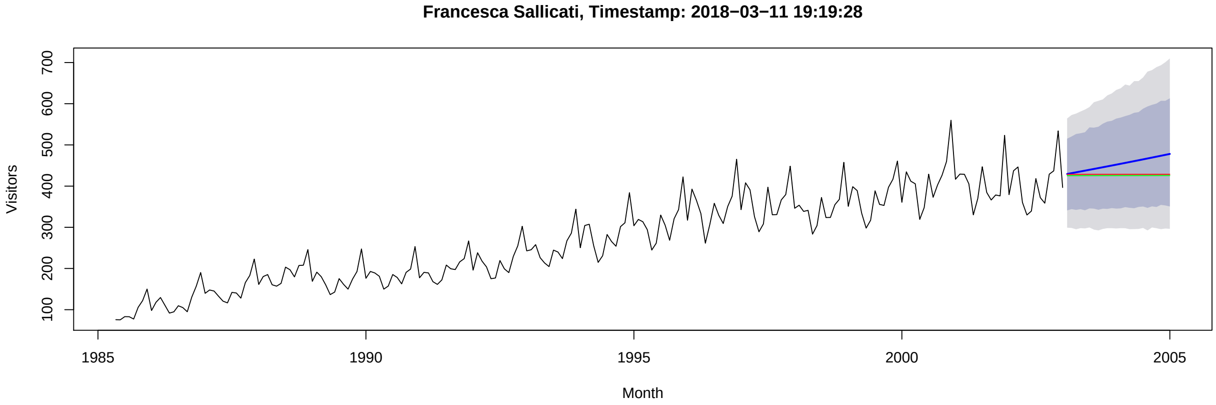
(*c*) Why is multiplicative seasonality necessary here?

The Holt-Winter's method works better with multiplicative seasonality beacuse the variance of the time series is increasng and it is able to capture the higher variance in the final part.

(*d*) Experiment with making the trend exponential and/or damped.

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

par(mfrow=c(1,1))
plot(fit8,ylab="Visitors",xlab="Month",main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
lines(fit9$mean, col="red")
lines(fit10$mean, col="green")
```



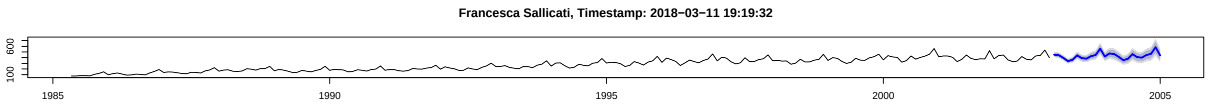
We can learn that making the trend exponential work better than damped.

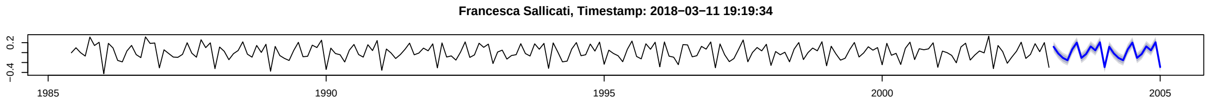
(*e*) Now fit each of the following models to the same data: 1. an ETS model 2. an additive ETS model applied to a Box-Cox transformed series 3. an STL decomposition applied to the Box-Cox transformed data followed by an ETS model applied to the seasonally adjusted (transformed) data.
Plot all the forecasts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

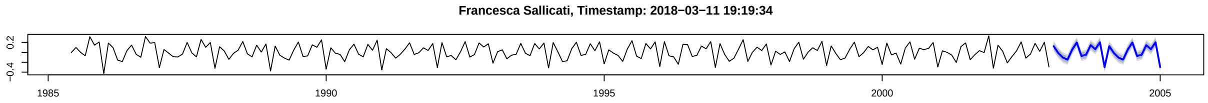
```
par(mfrow=c(3,1))
vdata <- window(visitors, end = 2003)
fit <- ets(vdata)
plot(forecast(fit), main=paste("Francesca Sallicati, Timestamp:",Sys.time()))

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

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





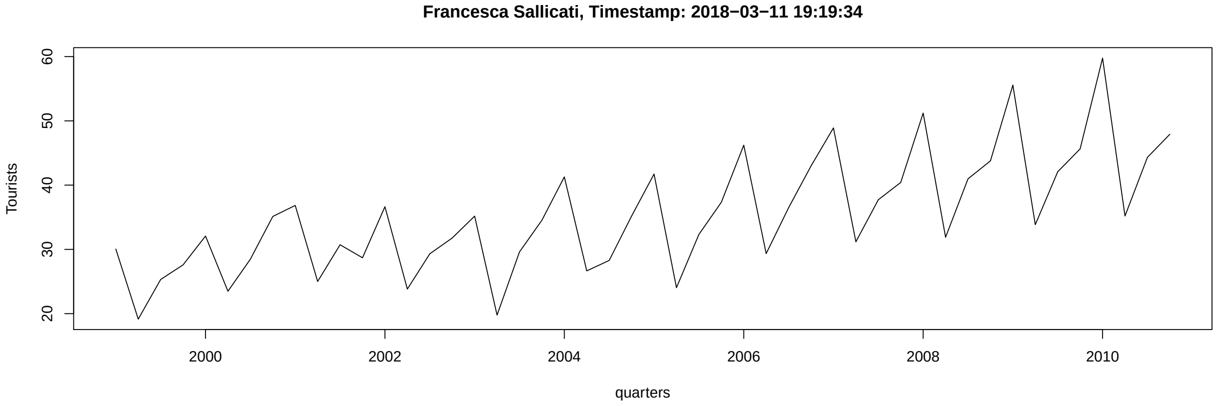


3 Exercise 3:

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

(*a*) Describe the time plot.

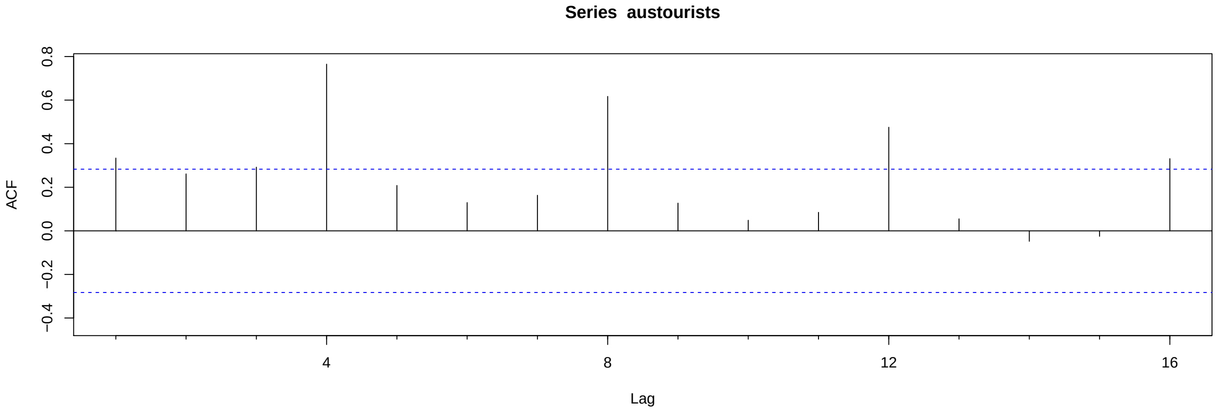
```
data("austourists")
par(mfrow = c(1, 1))
plot(austourists, xlab = "quarters", ylab = "Tourists",main=paste("Francesca Sallicati, Timestamp:",Sys
```



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

(*b*) What can you learn from the ACF graph?

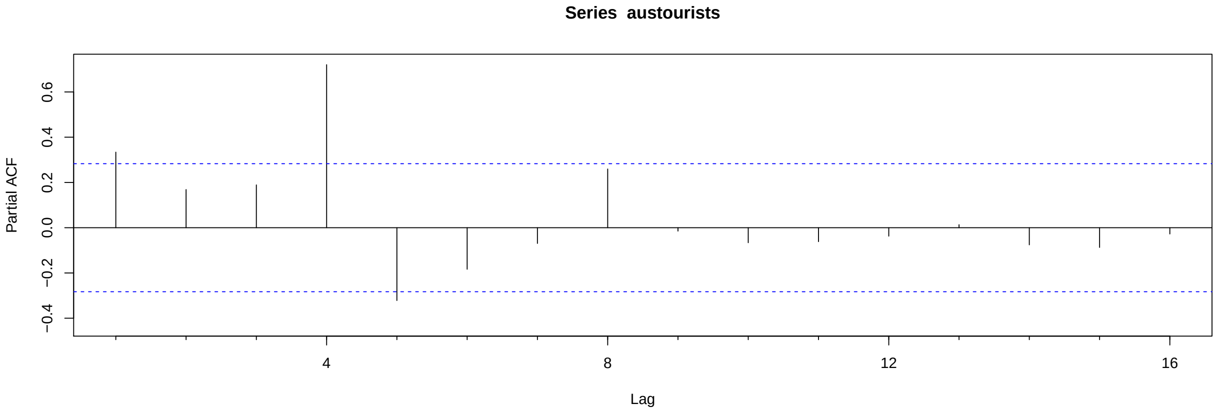
```
Acf(austourists)
```



From this ACF plot we can learn the seasonal pattern in the data with peaks which tends to 4 quarters apart, while there are no negative throughs. This might suggest to use an AR(4) in the ARIMA model for the 4 significative lags.

(c)What can you learn from the PACF graph?

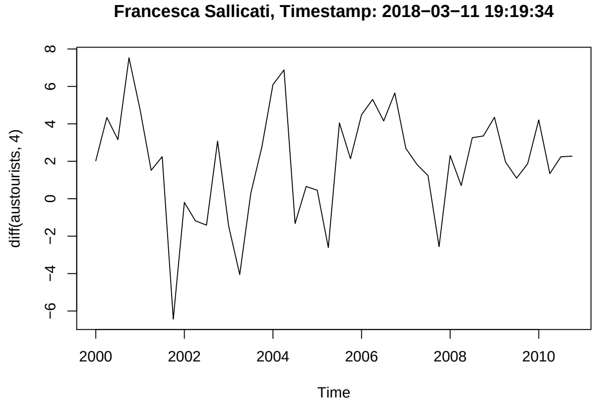
```r
Pacf(austourists)
```

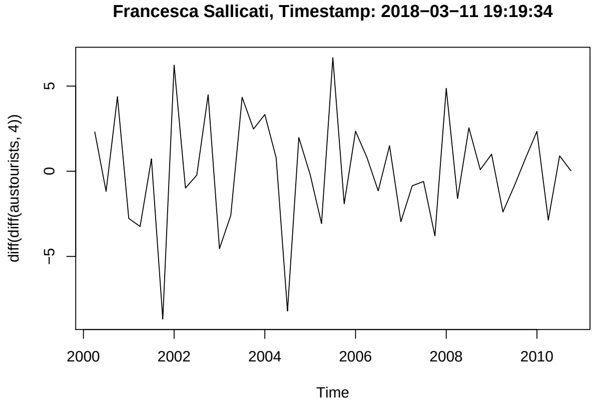


From this plot of partial autocorrelation we can see that the only significant peak is at lag(4), this means that the number of period per season actually $m = 4$ and that the autoregressive part of the model can be only AR(1) referred to lag4 only.

(d) Produce plots of the seasonally differenced data $(1 - B^4)Y_t$. What model do these graphs suggest?

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





Seasonally differenced data suggest that it may be worth to use $(.,1,.)_4$ or $(.,2,.)_4$ in the seasonal part of the ARIMA model in order to make the series stationary.

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

```
## 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
```

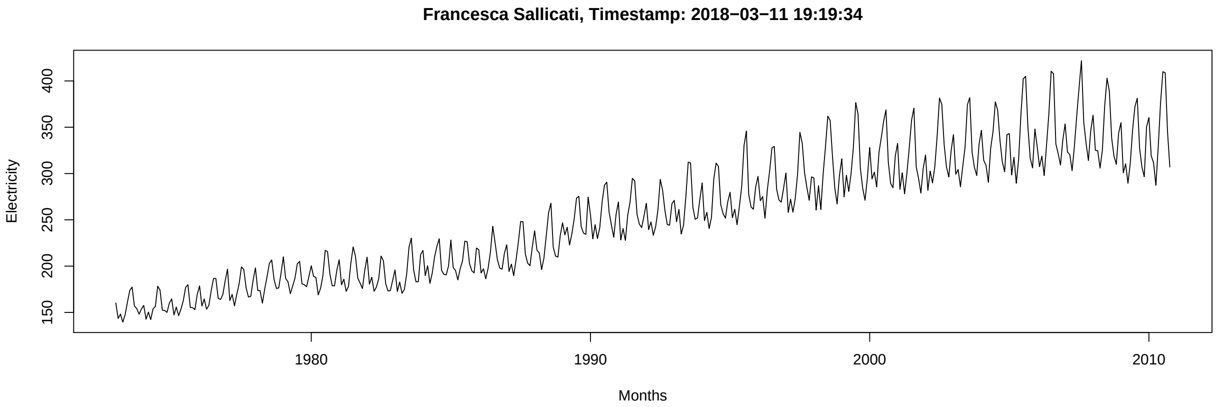
The auto.arima select a model with $D = 1$ in the seasonal part and an AR(1) in both the non-seasonal and seasonal part of the model, this quite agrees with what learned from the ACF and PACF graphs.

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.

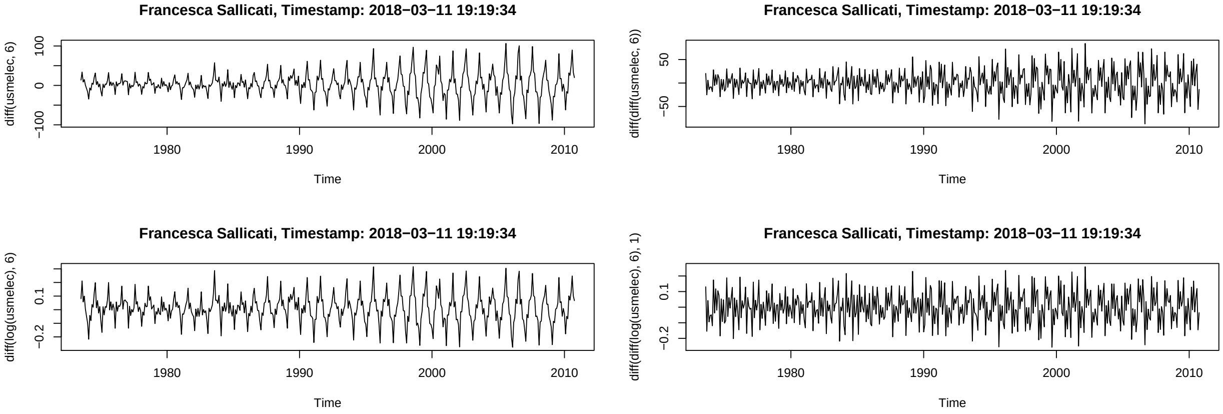
```
data("usmelec")
par(mfrow = c(1, 1))
plot(usmelec, xlab = "Months", ylab = "Electricity",main=paste("Francesca Sallicati, Timestamp:",Sys.ti
```



This seasonal time series shows a positive trend and an increase in the variance as well, this suggest that it could be useful to apply a transformation to the data.

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

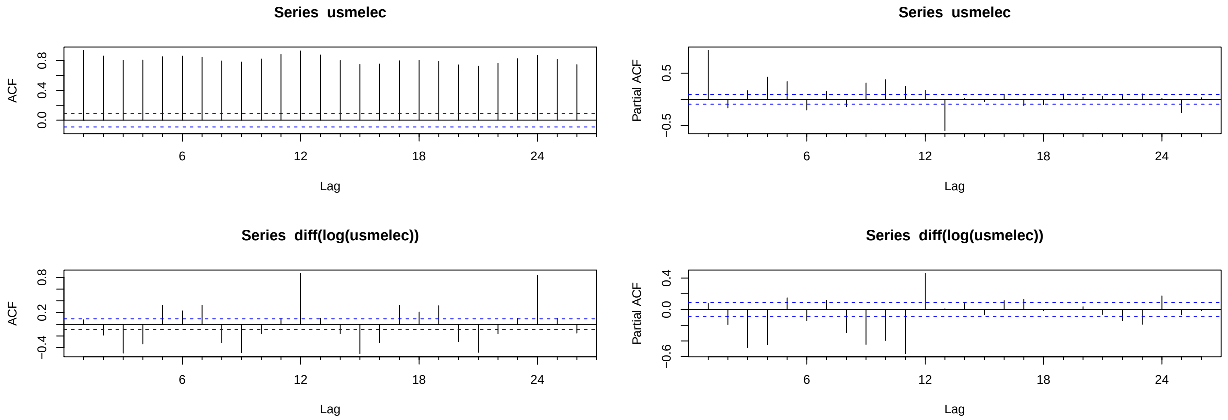
```
par(mfrow=c(2,2))
plot(diff(usmelec,6),main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
plot(diff(diff(usmelec,6)),main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
plot(diff(log(usmelec),6),main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
plot(diff(diff(log(usmelec),6),1),main=paste("Francesca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



The log-differenciated series seems the best option to make the series stationary. A seasonality of 6 months has been chosen due to mid-winter and mid-summer peaks of the data. In particular the logarithm addresses the problem of the different variance while the one time differenciating makes the times series stationary. Taking two times the difference in logarithm further eliminates seasonality producing a random walk.

(*c*) Identify a couple of ARIMA models that might be useful in describing the ti me series. Which of your models is the best according to their AICc values?

```
par(mfrow=c(2,2))
Acf(usmelec)
Pacf(usmelec)
Acf(diff(log(usmelec)))
Pacf(diff(log(usmelec)))
```



From the first two plot of the original data we can see autocorrelation in the time series, which is no longer present in the differenced data taken in log.

```
a1<-Arima(log(usmelec), order = c(1, 1, 0), seasonal = c(1, 1, 0))
a2<-Arima(log(usmelec), order = c(0, 1, 1), seasonal = c(0, 1, 1))
a3<-Arima(log(usmelec), order = c(1, 2, 0), seasonal = c(1, 2, 0))
a4<-Arima(log(usmelec), order = c(1, 2, 0), seasonal = c(1, 2, 0))
a5<-Arima(log(usmelec), order = c(1, 2, 1), seasonal = c(1, 2, 1)) #worse
a6<-Arima(log(usmelec), order = c(2, 2, 0), seasonal = c(2, 2, 0)) #worse
a7<-Arima(log(usmelec), order = c(1, 2, 1), seasonal = c(2, 2, 1)) #worse
a8<-Arima(log(usmelec), order = c(1, 2, 1), seasonal = c(1, 2, 2)) #worse
```

After having tried $D = 1$,$D = 2$ in the seasonal part gives a lower value of the AIC; moreover by introducing a MA() or increasing the AR() the AIC grows.

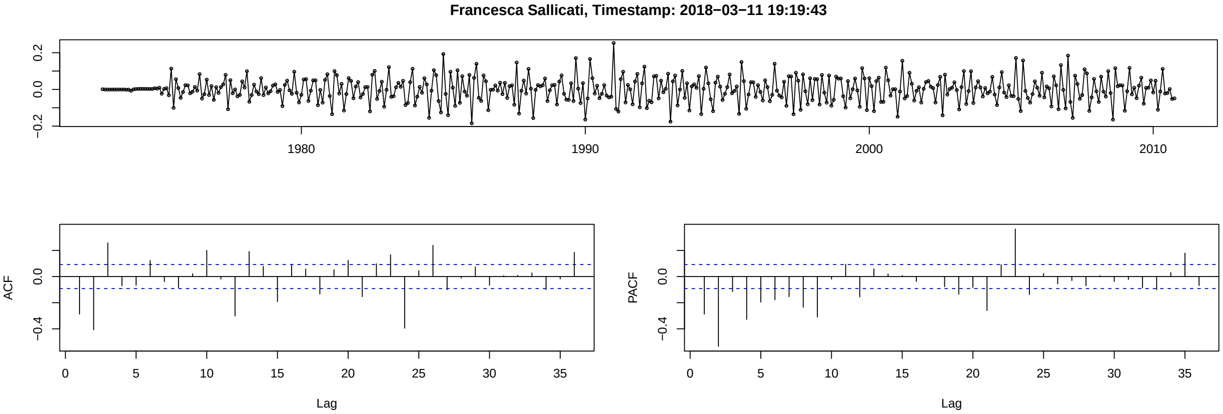
Therefore the best model among this is:

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

```
## Series: log(usmelec)
## ARIMA(1,2,0)(1,2,0)[12]
##
## Coefficients:
##           ar1      sar1
##       -0.5268   -0.6207
## s.e.   0.0410    0.0371
##
## sigma^2 estimated as 0.004883:  log likelihood=528.09
## AIC=-1050.17   AICc=-1050.12   BIC=-1037.99
```

(*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.

```
res <- residuals(a4)
tsdisplay(res,main=paste("Francesca Sallicati, Timestamp:",Sys.time()))
```



From the residual plot we can see that they quite follow a white noise, apart from the first part.