Unit 7

Time Series

ESIGELEC

Instructor: Federico Perea

Glossary

- Time series
- Trend
- Cycle
- Seasonality
- Stationarity
- Irregular component
- Differencing a process
- Integrated process
- Overdifferencing
- *Autoregressive-integrated-moving-average (ARIMA)* model

Introduction

- Time series data are random variable observations collected over a period of time.
- We want to study them in order to predict future values of the random variable of interest.
- Classic methods study time series from its regularities (also called components) which are:
  - Trend
  - Seasonal variation
  - Cyclic variation
  - Irregular variation

  This methodology is too simple. We will instead use ARIMA models.

The main goals of this unit is:

- Defining time series and giving a descriptive analysis, including the classic components.
- Introducing the ARIMA model as a statistics model to analyze time series data

Time series concept. Notation.

Time series data are observations of a random variable collected over a period of time.

Z_t: value of the series at time t.

μ_t, σ_t are the mean and standard deviation of Z_t

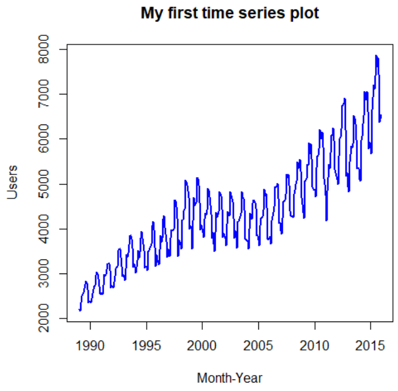
Studying these data will allow you to:

- Build descriptive models about the evolution of the studied variable
- Forecast its future values

See the example of this unit: the users of a train line, measured from January 1989 to December 2015.

Graphic representation

A time series plot is essential for detecting the most important *regularities* in the model.



In R, after Reading the data

#First transform the data USERS into a ts-object

ts_USERS <- ts(USERS, frequency = 12, start = c(1989,1), end = c(2015,12))

#plot the time series

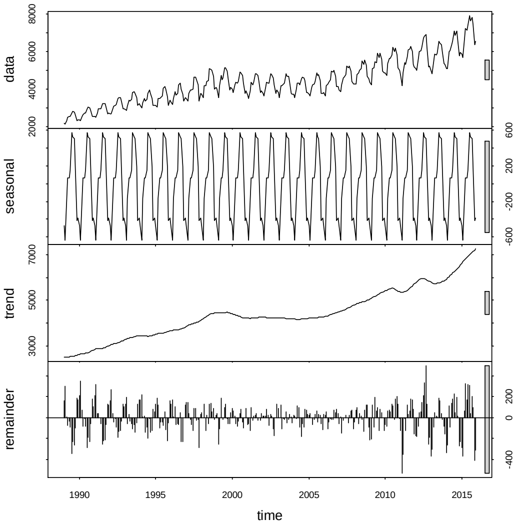
plot(ts_USERS, main = "My first time series plot", type = "l", lwd = 2, col = "blue", xlab = "Month-Year", ylab = "Users")

Classic Components

Every time series consists of four different components:

- Trend component T_t (average behavior in the long term)
- Seasonal component E_t (pattern repeated every certain number of observations: every week, every year, …)
- Cyclical component C_t (pattern repeated every certain large number of observations)
- Irregular (or residual) component R_t (the rest)

It often is difficult to tell the difference between trend and cycle, and therefore they are jointly studied as one unique component.

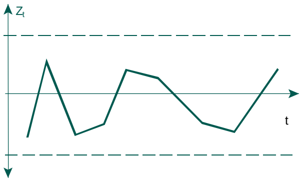


To plot the decomposition of the time series into trend, seasonal component, and remainder (residual)

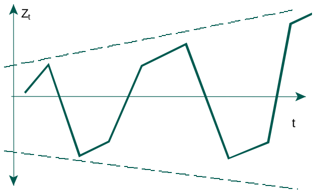
plot(stl(ts_USERS,s.window="periodic"))

Stationary time series

- A time series Z_t is stationary if its mean and variance are constant over time and the value of the covariance between any two periods depends only on the distance (or gap or lag) between them, and not on the actual time at which the covariance is computed $(\mu_t = \mu\,, \sigma_t^2 = \sigma^2, Cov(Z_t, Z_{t-k}) = \gamma_k \ \forall\ t,)$.
- Otherwise the series is *nonstationary*.
- Note: series with trend and/or seasonal behavior are not stationary.
- Is the series *USERS* stationary? No, as there is a trend and a seasonal behavior



STATIONARY



NONSTATIONARY

Models for integrated processes

- It is possible to formulate ARIMA models for stationary series, or for non-stationary processes that, when being differenced, they become stationary
- Such processes are called the *autoregressive-integrated-moving average* processes, ARIMA.

- The most general ARIMA model is
- $ARIMA(p,d,q)x(P,D,Q)_S$
  $\left(1 - sar1 \cdot B^S - sar2 \cdot B^{2S} - \cdots - sarP \cdot B^{SP}\right)(1 - ar1 \cdot B - \cdots - arp \cdot B^p)$
  $(1 - B^s)^D (1 - B)^d \ \bar{Z}_t =$
  $(1 - ma1 \cdot B - \cdots - maq \cdot B^q)\left(1 - sma1 \cdot B^S - \cdots - smaQ \cdot B^{SQ}\right)\alpha_t$
- Autorregresive regular order: p (order of AR)
- Number of regular differences to remove trend: d
- Moving average regular order: q (order MA)
- Autorregressive seasonal order: P (order of SAR)
- Number of seasonal differences to remove seasonality: D
- Moving average seasonal order: Q (order of SMA)
- Period of seasonality: S

The values of d and D are typically 0,1,2.

Values for *p,q,P,Q* need to be found out.

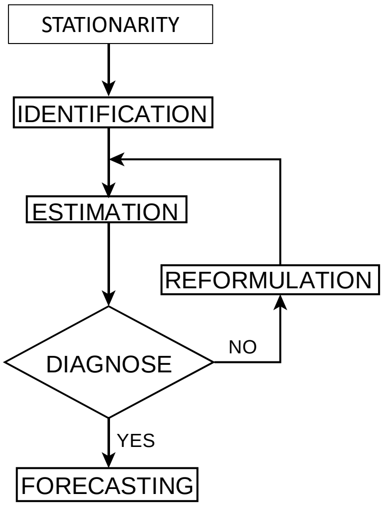
The values of the parametes ar, ma, sar, sma need to be estimated.

Box-Jenkins Methodology

- One key aspect is to determine the orders of the ARIMA model *(p,d,q,P,D,Q,S).*
- The Box-Jenkins (B-J) methodology comes in handy in answering this question.
- In this short introduction, we will just jump to the auto.arima function of R, which gives you a “good” ARIMA model automatically (go to slide Find a model and do predictions)
- If you have some time, have a look at the Box-Jenkings methodology, in following slides.

Box-Jenkins methodology

- In order to solve the problem of how to identify the parameters of the ARIMA model that describes a time series, we have the Box-Jenkins methodology



STATIONARITY

IDENTIFICATION

ESTIMATION

REFORMULATION

DIAGNOSE

NO

YES

FORECASTING

1. **Stationarity:** apply the necessary transformations so the process becomes stationary (find out *d, D,* and *S*).

2. **Identification**: find out the appropriate values of *p,q,P,Q*.

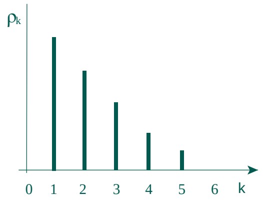
3. **Estimation**: Estimate the parameters of the autoregressive and moving average terms included in the model.

4. **Diagnostic checking**: Is the estimated model appropriate for our data? One simple way is to see if the residuals estimated from this model are white noise. If they are, we proceed to step 5. If they are not, start over. (Not to be seen in this introductory unit)

5. **Forecasting**: one of the main goals. Forecasts obtained by this method are often more reliable than those obtained from the traditional approach, particularly for short-term forecasts.

The simple autocorrelation function (ACF)

- This function is calculated in stationary processes from the autocorrelation coefficients $\rho_k = \gamma_k / \gamma_0$
- The simple autocorrelation function is the graphical representation of these coefficients as a function of the lag *k.*



- The ACF helps in detecting the existence of trend and seasonality
- Slow or linear decrease in first 6-8 coefficients implies trend.
- Picks in seasonal coefficients (12, 24, 36,… if the series has period 12) imply the existence of a seasonal component.

Partial autocorrelation function (PACF)

- Sometimes the relation between two variables separated by a lag of *k* can be direct or indirect.
- The effect of Z_{t-2} over Z_t is indirect (through Z_{t-1})

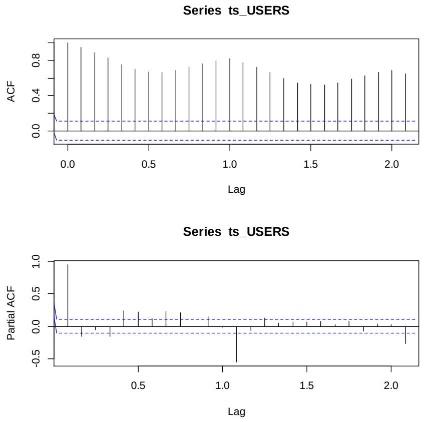
$$\begin{tikzcd} Z_{t-2} \arrow[r] & Z_{t-1} \arrow[r] & Z_t \end{tikzcd}$$

- The effect of Z_{t-2} over Z_t is direct and indirect

$$\begin{tikzcd} Z_{t-3} \arrow[r] \arrow[rr, bend left] & Z_{t-2} \arrow[r] \arrow[rr, bend right] & Z_{t-1} \arrow[r] & Z_t \end{tikzcd}$$

- The partial autocorrelation coefficient measures the direct relationship between two variables.
- The simple autocorrelation coefficient measures the total relation (both direct and indirect) between two variables.

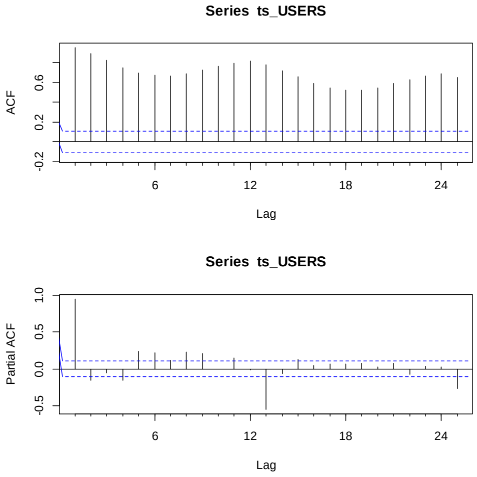
ACF and PACF in R



par(mfrow=c(2,1))
acf(ts_USERS)
pacf(ts_USERS)
par(mfrow=c(1,1))

These functions will be useful to design ARIMA models

ACF and PACF with package “forecast”



library(forecast)
par(mfrow=c(2,1))
Acf(ts_USERS)
Pacf(ts_USERS)
par(mfrow=c(1,1))

In ACF we see a slow linear decrease in the first coefficients → Trend
We also observe picks in coefficient number 12 and 24, → Seasonality

Integrated models: remove the trend

- Most of economics models are not stationary because their mean changes with time.
- It is often the case that such processes become stationary when we take differences (when we *difference* it).
- We call this process W_t **integrated of order 1**

$$W_t = Z_t - Z_{t-1}$$

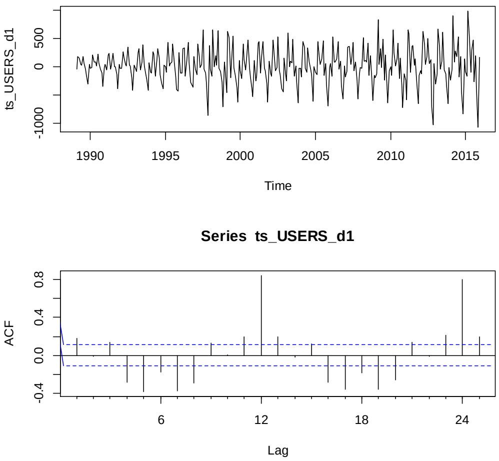
Integrated time series

- A model that has to be differenced to become stationary is called an integrated stationary process.
- If a time series has to be differenced once, the process is integrated of order 1
- If it has to be differenced twice (take the first difference of the first differences) to make it stationary, we call such a time series integrated of order 2.
- And so on.

- The process is integrated of order 2 when it is obtained when differencing W_t.

$$Y_t = W_t - W_{t-1} = Z_t - 2Z_{t-1} + Z_{t-2}$$

- A process is integrated of order *h* if it becomes stationary when differencing it *h* times.
- The differences of a stationary process is a stationary process, therefore *overdifferencing* does not affect the process.



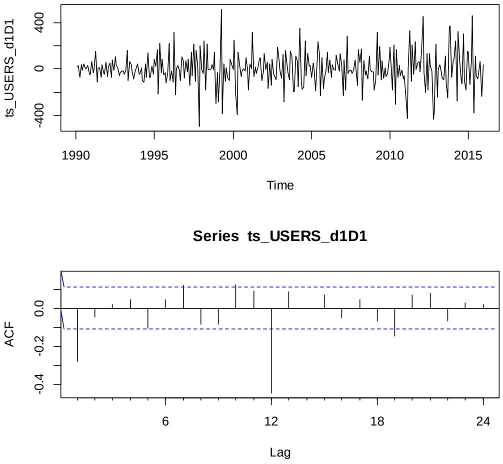
#Take regular differences to remove the trend
ts_USERS_d1 <-diff(ts_USERS,differences=1)

par(mfrow=c(2,1))
plot(ts_USERS_d1)
Acf(ts_USERS_d1)
par(mfrow=c(1,1))

#After this difference it seems that the trend is removed, but still there is seasonality, since we observe a similar pattern repeated in the series plot, and picks in the ACF, coefficients 12 and 24. This implies a seasonal behavior of period 12.

Removing Seasonality

- Seasonality can be removed by *differencing seasonally.*
- The current value of the variable minus the value of the variable in the previous season. For instance, assuming a repetitive behavior each year
  - If the variable is monthly measured: $Y_t = Z_t - Z_{t-12}$
  - If the variable is quarterly measured: $Y_t = Z_t - Z_{t-4}$
- Seasonality can be observed from
  - The time series plot
  - The simple autocorrelation function (picks every 12 observations)



```
#Let us add a seasonal difference of order 1
ts_USERS_d1D1 <- diff(ts_USERS_d1,lag = 12,
differences=1)

par(mfrow=c(2,1))
plot(ts_USERS_d1D1)
Acf(ts_USERS_d1D1)
par(mfrow=c(1,1))

#It now seems that the series is stationary
```

Variance

- When the variance of a process is not constant, this can be solved by:
  - Taking logarithms of observations
  - Taking square roots of observations
  - Differencing
- Sometimes we may need more than one transformation to obtain stationarity. In such cases the order in which the transformations are applied matters.
- The simple autocorrelation function does not help to detect problems with non-constant variance.
- But the representation of the time series does help in deciding whether or not the variance is constant.
- For our example, we will continue assuming that the variance is constant.

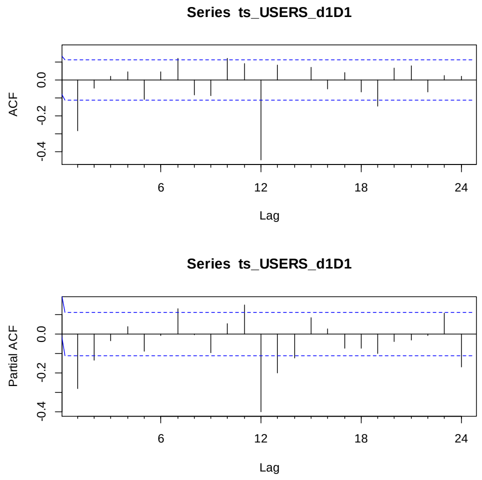
Finding p,q,P,Q

1. Regular part
   q = number of significant ACF coefficients.
   p = number of significant PACF coefficients.
   In both cases, check only the first 6-8 coefficients.
2. Seasonal part
   Q = number of significant seasonal ACF coefficients.
   P = number of significant seasonal PACF coefficients.
   Seasonal coefficients are:
      12, 24, 36,… if S=12
      4,8,12,… if S =4
      …
   In both cases, check only the first 2-3 coefficients.

3. Combine the models chosen for the regular part with the models chosen for the seasonal part.

Regular	Seasonal	ARIMA(p,d,q)x(P,D,Q)
q	Q	(0,d,q)x(0,D,Q)
q	P	(0,d,q)x(P,D,0)
p	Q	(p,d,0)x(0,D,Q)
p	P	(p,d,0)x(P,D,0)

- EXAMPLE: ACF and PACF of the stationary series (obtained after one regular difference and one seasonal difference).



#Let us obtain the ACF and PACF of the differenced series

par(mfrow=c(2,1))
Acf(ts_USERS_d1D1)
Pacf(ts_USERS_d1D1)
par(mfrow=c(1,1))

We have: p=2, q=1, P=2, Q = 1.

- Which ARIMA(p,d,q)x(P,D,Q)$_s$?
- In our example we have four possible combinations

Regular	Seasonal	ARIMA(p,d,q)x(P,D,Q)
q=1	Q=1	(0,1,1)x(0,1,1)
q=1	P=2	(0,1,1)x(2,1,0)
p=2	Q=1	(2,1,0)x(0,1,1)
p=2	P=2	(2,1,1)x(2,1,0)

- Which one should you choose? The one with best significance of parameters, and lowest error variance, best residuals,….
- This is too much for this introductory unit, and deserves more time.
- Instead, we will use the R-function *auto.arima*, within the package “forecast”

model_best <- auto.arima(ts_USERS)

- In this example you get model_1

Find a model and do predictions

With *auto.arima*, within the package “forecast”, you get the “best” possible ARIMA model.

model_best <- auto.arima(ts_USERS)

- The ARIMA model allows you to estimate the values of the variable studied.
- Find the forecasts for the next 9 months using model_best

predictions <- predict(model_best,n.ahead=9)

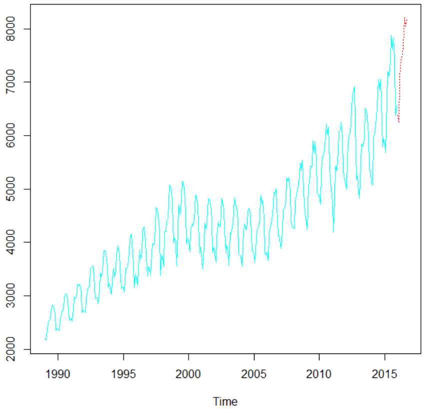
predictions

The result is:

6379.561, 6236.035, 7088.315, 7458.473, 7489.022, 7708.362, 8213.007, 8044.732, 8183.267

Plot the series with the predictions

ts.plot(ts_USERS, predictions$pred, lty = c(1,3), col=c(5,2))



How to compute and plot the confidence intervals of the predictions

- #First build the confidence intervals

upper = predictions$pred + 1.96*predictions$se

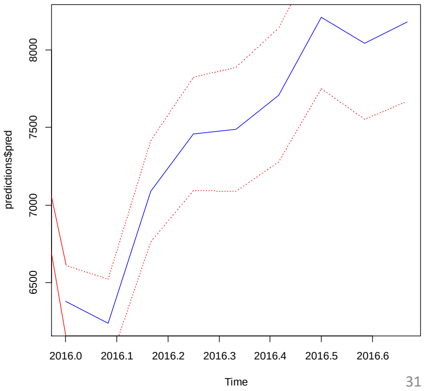
lower = predictions$pred - 1.96*predictions$se

- Then plot both the predictions and the CI

plot(predictions$pred,col="blue")

lines(upper,col="red",lty=3)

lines(lower,col="red",lty=3)

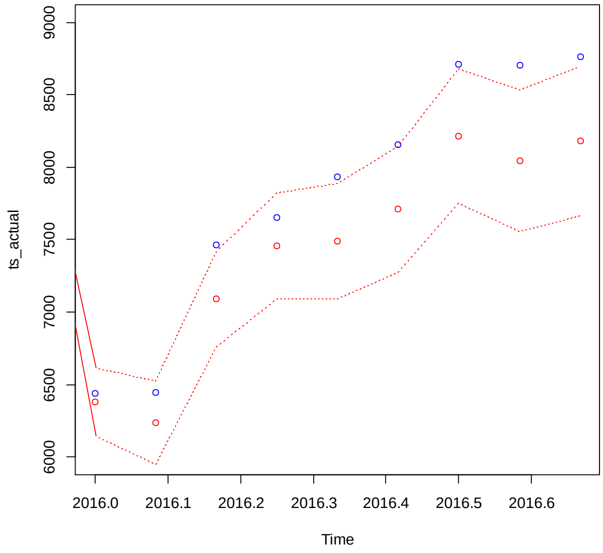


Are the predictions good?

- The actual values of the variable USERS for the first 9 months of 2016 were
- 6441.1; 6446.6; 7466.3; 7652.31; 7935.89; 8154.57; 8712.82; 8706.77; 8762.54
- Comparing these values with your predictions before, is the model proposed forecasting correctly?
- You want (most of) the actual values of the variable to be within the confidence intervals of the predictions. More or less half of the actual values should be over the predicted value, the other half below.

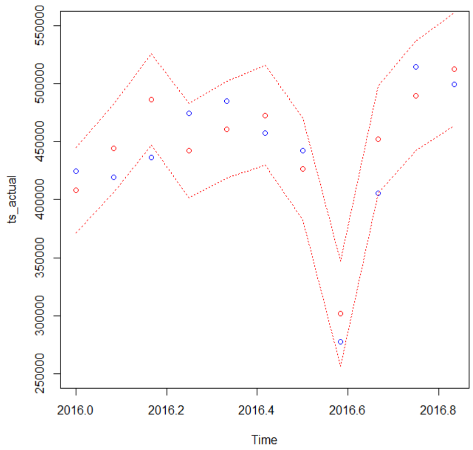
Actual	Pred	Lower	Upper	In CI?	Over/Under
6441	6379	6416	6612		
6446	6236	5951	6520		
7466	7088	6760	7415		
7652	7458	7092	7824		
7935	7489	7088	7889		
8154	7708	7276	8140		
8712	8213	7751	8674		
8706	8044	7555	8534		
8762	8123	7667	8699		

```
actual = c(6441.1, 6446.6, 7466.3, 7652.31, 7935.89, 8154.57, 8712.82, 8706.77, 8762.54)
ts_actual <- ts(actual, frequency = 12, start = c(2016,1), end = c(2016,9))
#Now plot the actual values together with the forecasts and their C.I.
#plot the actual values in the forecasted periods as red points
plot(ts_actual,col="red", type="p",ylim=c(6000, 9000))
#add to the plot the confidence intervals of the forecasts as dotted blue lines
lines(upper,col="blue",lty=3)
lines(lower,col="blue",lty=3)
#add to the plot the forecasted values as blue points
points(predictions$pred,col="blue")
```



It seems that 5 out of 9 actual values are out of the confidence intervals, and one is on the edge. It also seems that our predictions (red points) are always below the actual values (blue points). This is called Underestimation. So, the model does not seem to be predicting correctly, and it should be fixed. How? We don't have time in this introductory unit

Example of good preditions



This is an example of predictions (red points) that seem to be good, because

1) We see that most of actual values (blue points) are within the confidence intervals of the predictions (red lines).
2) Some of the actual values are above the predictions, some are below (more or less 50% - 50%).

End

- This was only a short introductory unit to time series
- The Box-Jenkins methodology is more complex tan what we have seen here
- After you propose an ARIMA model, you should always check that the residuals are a White noise process. Otherwise, the model is not good and should be reformulated.
- Time series analysis is quite extensive!