Module MP(C)

Methods for predicting numerical variables

Simon Malinowski

M1 Miage - M1 Data Science, Univ. Rennes

Schedule of this course

1. Part 1 : Time series prediction
   - 4h30 CM
   - 10h30 TD/TP (including a graded work)
2. Part 2 : Prediction using other variables (regression models)
   - 4h30 CM
   - 6h TD/TP
   - 4h30 Project (graded)

     $\rightarrow$ Kaggle competition : ▸ Lien

Time series prediction

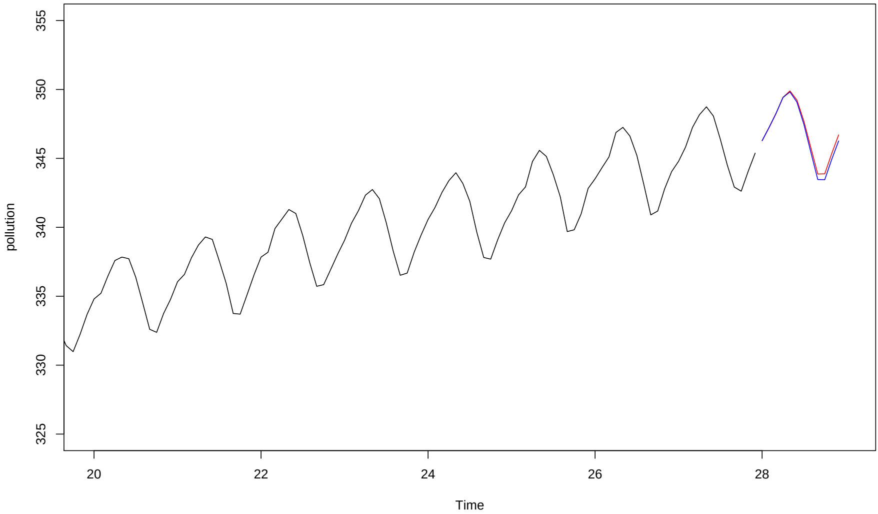
Simon Malinowski

M1 Miage, Univ. Rennes 1

Time series examples

- electricity consumption of users every minute or hour
- number of people in the public transport every day
- average daily temperature in a given city

Our goal: predict the next value(s)



What is a time series?

Definition

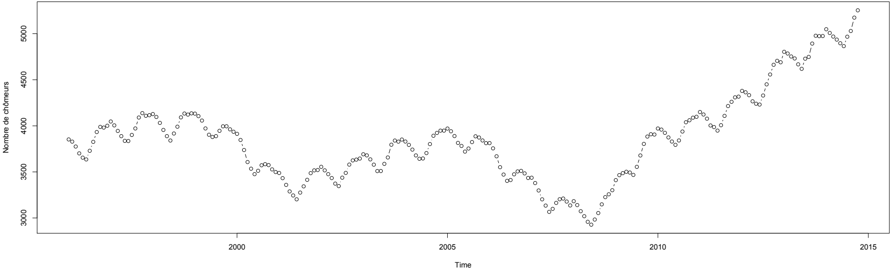
A time series is a **finite series** of **numerical** data points that represents the evolution of a certain quantity over reguar time stamps

Also called sometimes chronolgical series

Denoted by $X(t) = x_1, x_2, \ldots, x_n$, or also $X(t) = x(1), x(2), \ldots, x(n)$

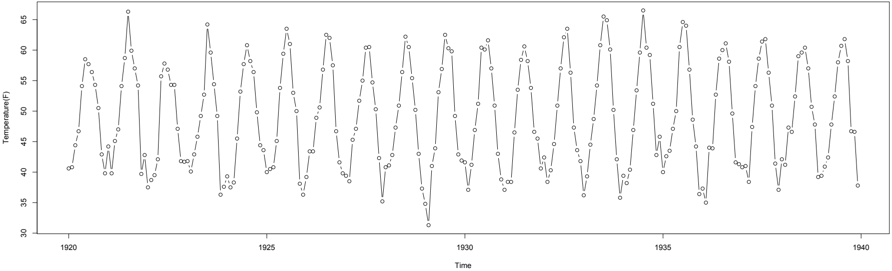
Graphical representations

1) Global representation : the points (t, x_t) are depicted on a plot
Example 1 : number of unemployed people registered to "Pôle Emploi" in France every month from 2001 to 2014



Graphical representation

Global representation : the points (t, x_t) are depicted on a plot
Example 2 : Monthly average temperature in Nottingham
Period 1920-1940

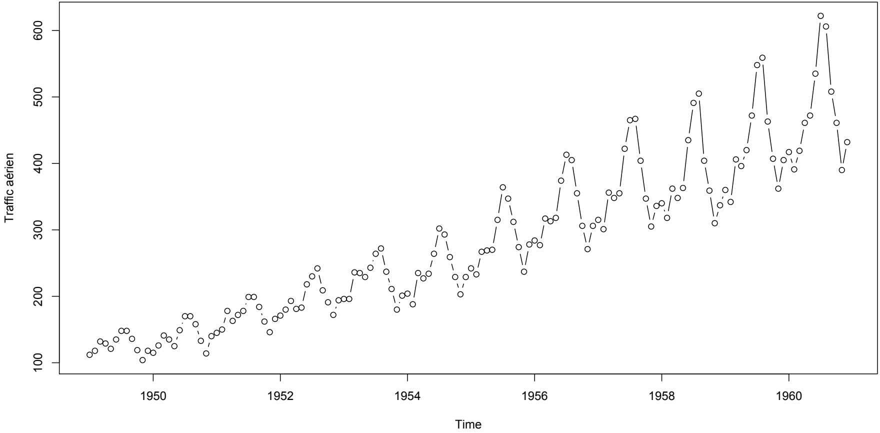


Graphical representations

Global representation : the points (t, x_t) are depicted on a plot

Example 3 : Average monthly air trafic

Period 1949-1961



Analyzing and understanding a time series

Time series often represent a complex phenomenon, difficult to analyze in a straightforward manner

The methods that we will use to predict a time series are based on the decomposition of a TS into elements that are simpler to

- handle with
- control

The elements that are considered are :

1. the trend T_t
2. the seasonal component S_t
3. the residual component E_t

Decomposition models

Additive model

We assume that X_t can be written as

$$X_t = T_t + S_t + E_t$$

Multiplicative model

We assume that X_t can be written as

$$X_t = (T_t * S_t) + E_t$$

$$X_t = T_t * S_t * E_t$$

The trend

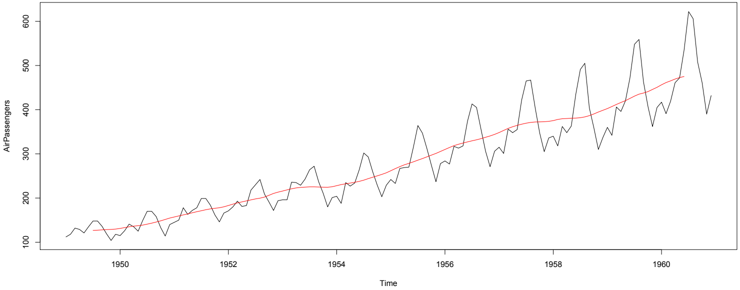
Definition

This component represents the global evolution of the considered phemomenon

Assumptions about this component

- slowly varying
- deterministic
- can be estimated as a (simple) mathematical function

Trend or not trend ?



Kinds of trends

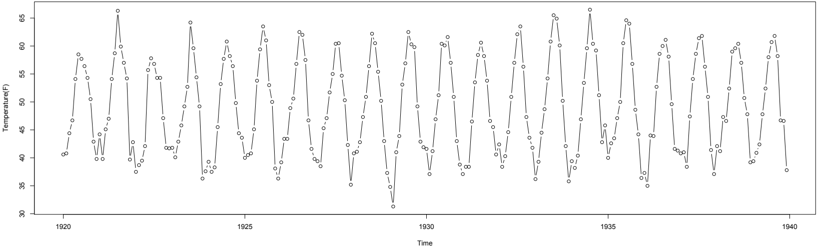
1. Linear: $T_t = a + b \times t$
2. Polynomial : $T_t = a_0 + a_1 \times t + \cdots + a_n \times t^n$
3. Logarithmic : $T_t = a + b \times \log t$
4. Exponential : $T_t = a \times exp(bt)$

The seasonal component

Definition

The seasonal component represents periodic fluctuations around the mean inside a period (of length p) and that occur almost identically from period to period

Example : Temperature in Nottingham

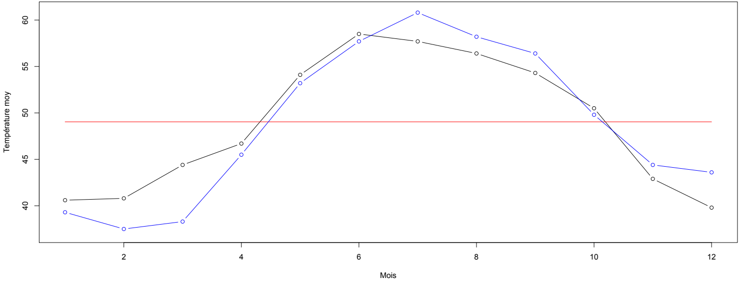


The seasonal component

Definition

The seasonal component represents periodic fluctuations around the mean inside a period (of length p) and that occur almost identically from period to period

Example : Temperature in Nottingham



The seasonal component

This component is completetly defined by ***p* seasonal coefficients** $s_1, \ldots s_p$ that represent the seasonal profile.

We have $S_t = s_1, \ldots, s_p, s_1, \ldots, s_p, \ldots$

Hence, the seasonal component S_t is the repetition of these p coefficients over each period.

The residual component E_t

This component represents irregular and "unpredictable" fluctuations.

It is a random component, considered often as a zero-mean random variable

Determining the composition of a time series

All time series are not composed of both a trend and a seasonal component

The typical kinds of time series are

- completely random : $X_t = E_t$
- with a trend component only : $X_t = T_t + E_t$
- with a seasonal component only : $X_t = S_t + E_t$
- with both a trend and a seasonal component : $X_t = T_t + S_t + E_t$

Next step is a tool to help you determine in which case we are

Autocorrelation function

Let $X_t = x_1, \ldots, x_n$ be a time series, et k an integer $\in [0, n-1]$.
The autocorrelation function of X_t is defined as :

$$\rho(k) = \frac{cov(x_t, x_{t-k})}{var(x_t)} = \frac{\sum_{t=k+1}^{n} (x_t - \overline{x})(x_{t-k} - \overline{x})}{\sum_{t=1}^{n} (x_t - \overline{x})^2}$$

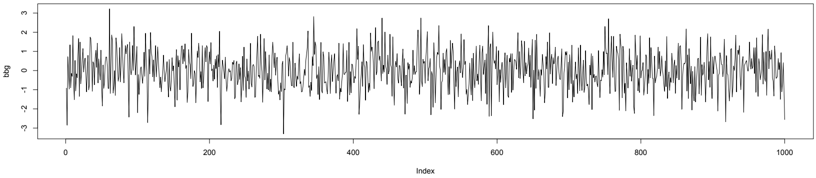
Properties :

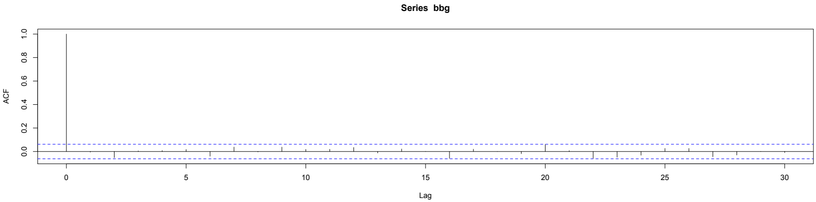
- $\rho(0) = 1$
- $\rho(k) \leq 1, \forall k > 0$

Autocorrelation function

For a completely random time series ($X_t = E_t$),

$$|\rho(l)| < 1.96 * \sqrt{1/n}, \forall l > 0$$

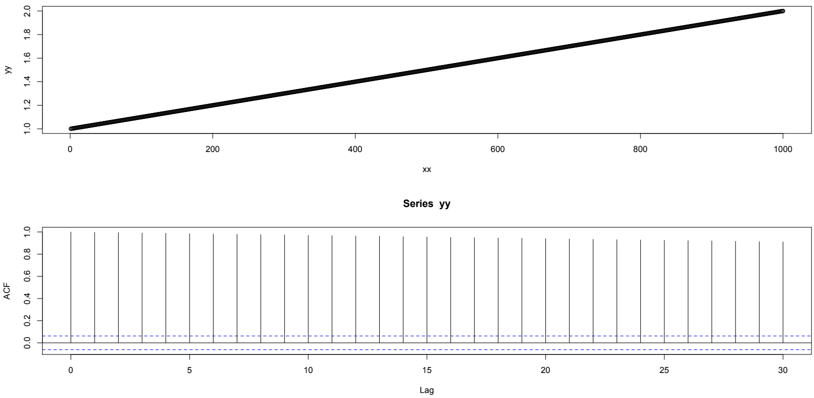




Autocorrelation function

For a time series with a trend ($X_t = T_t + E_t$), we have:

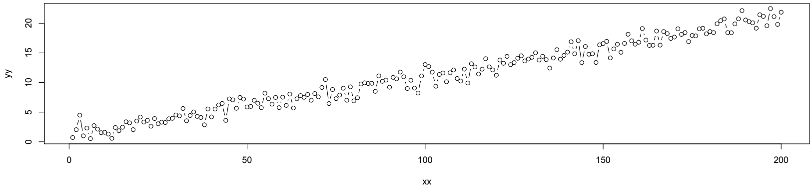
$$\rho(l) \to 1, \forall l > 0$$

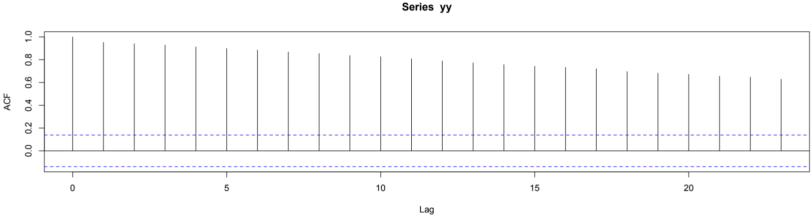


Autocorrelation function

For a time series with a trend ($X_t = T_t + E_t$), we have:

$$\rho(l) \to 1, \forall l > 0$$

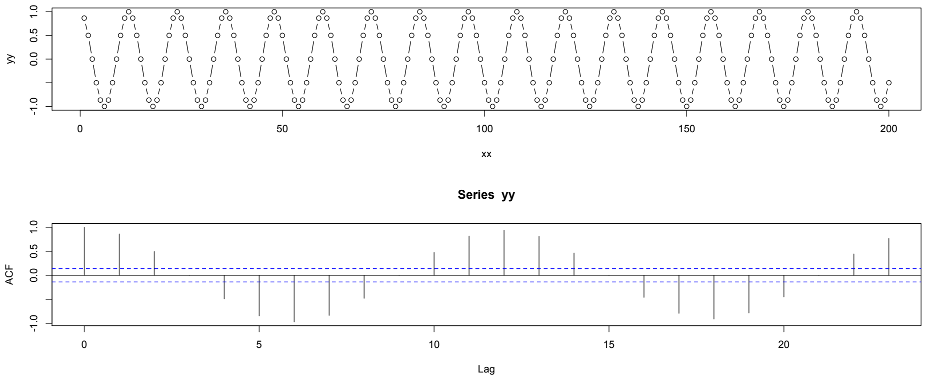




Autocorrelation function

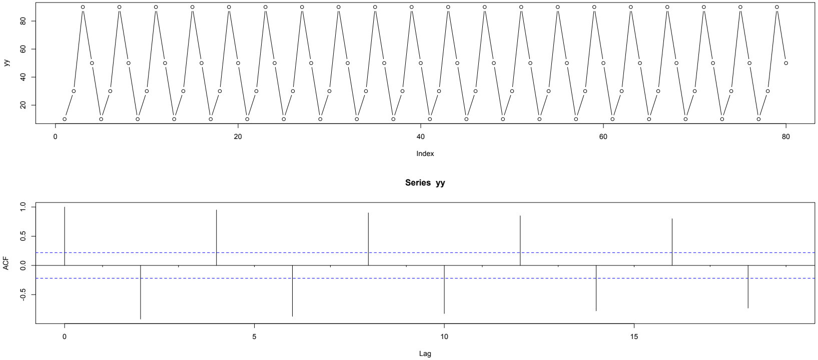
For a time series of the kind $x_t = cos(\frac{2\pi}{T}t)$ (with a seasonal component), we have :

$$\rho(l) \rightarrow cos(\frac{2\pi}{T}t), \forall l > 0$$



Autocorrelation function

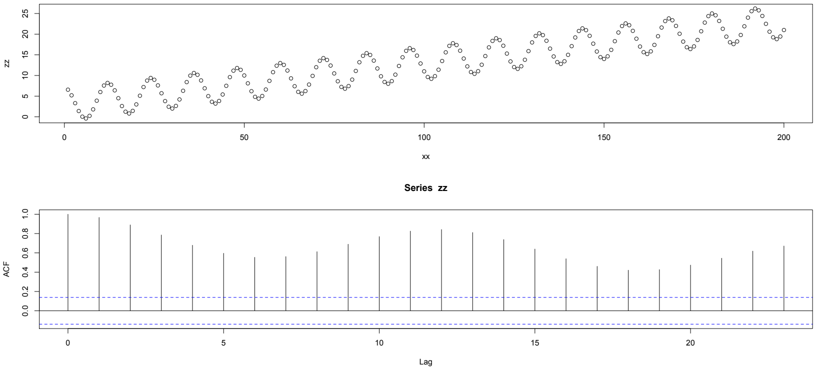
Example for a time series with a period of length 4



Autocorrelation function

And for a series with both a trend and a seasonal component:

$$x_t = a + b \times t + cos(\frac{2\pi}{T} t)$$



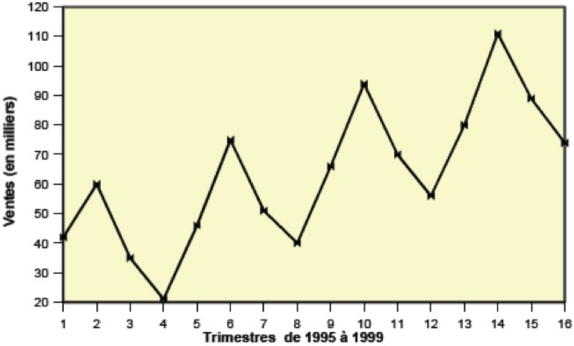
Detection of trend and/or seasonality

Summary:

1. Intuition about the phenomenon
2. Graphical representations of the time series
3. Analysis of the autocorrelation function
   - slow decrease of the coefficients $\rightarrow$ trend
   - periodicity of the coefficients $\rightarrow$ seasonality
   - combination of both : trend + seasonality

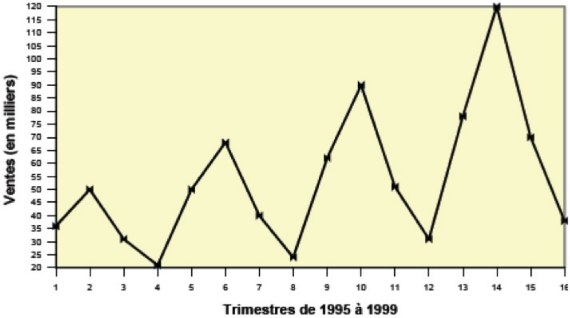
Seasonal component: additive or multiplicative ?

- For an additive model, the time series stays inside a constant "strip" around the trend



Seasonal component: additive or multiplicative ?

- For a multiplicative model, the ”strip” arounf the mean is not constant over time, i.e. the seasonal components depend on the trend



Example

Prediction of a series without trend nor seasonality

It is the simplest case, rarely used in practice. But it allows to understand and apprehend some important things that will be used later.
Hypothesis: the time series takes its values around a stable level a , it has no trend an no seasonal component.

$$X_t = a + e_t,$$

with e_t a residual component.

Problem : how to estimate a in order to predict the future values of X_t ?

Prediction of a series without trend nor seasonality

Let $X_t = x_1, \ldots, x_n$ a series witout trend nor seasonality. We aim at predicting the next value : $\hat{x}_{n+1}$ (and why not the next ones).

1. Naïve : the last value x_n is used as the prediction at time $n+1$

$$\hat{x}_{n+1} = x_n$$

   + next value is likely to be close to the previous one
   − only the previous value is used to predict (the past information is lost)

On our example:

Prediction of a series without trend nor seasonality

2. prediction is the mean of all past values :

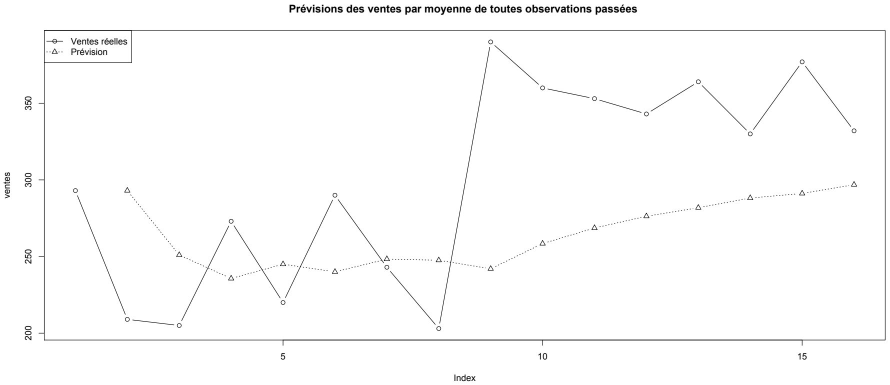
$$\hat{x}_{n+1} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

+ all past values are used for prediction

−

On our example:

Prediction using the mean of past values



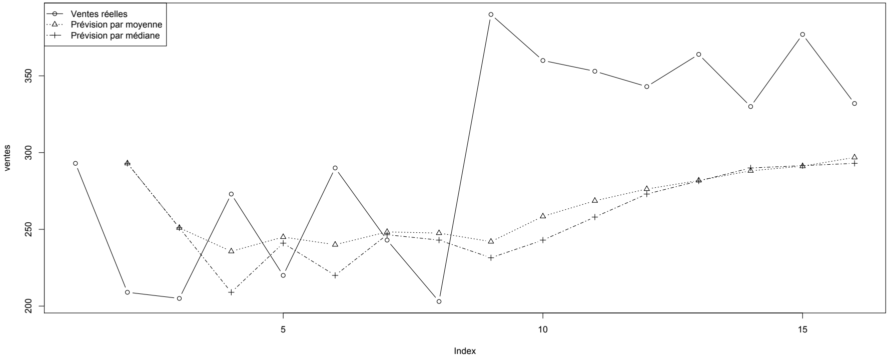
Prediction of a series without trend nor seasonality

3. prediction is the median of all past values

+ all available data is taken into account
+ extreme values have less influence
− the close past is as important than the far past

On our example:

Prediction using the median of past values



Prediction of a series without trend nor seasonality

4. Weighted sum of all past values

$$\hat{x}_{n+1} = \frac{\sum_{i=1}^{n} \omega_i x_i}{\sum_{i=1}^{n} \omega_i}$$

+ we can decide to give more influence to some past values
+ all available data is used
− choice of ω_i is tedious

On our example:

Prediction of a series without trend nor seasonality

5. Simple exponential smoothing

+ more influence to recent values
+ all available data is used
+ flexible and easy to apply

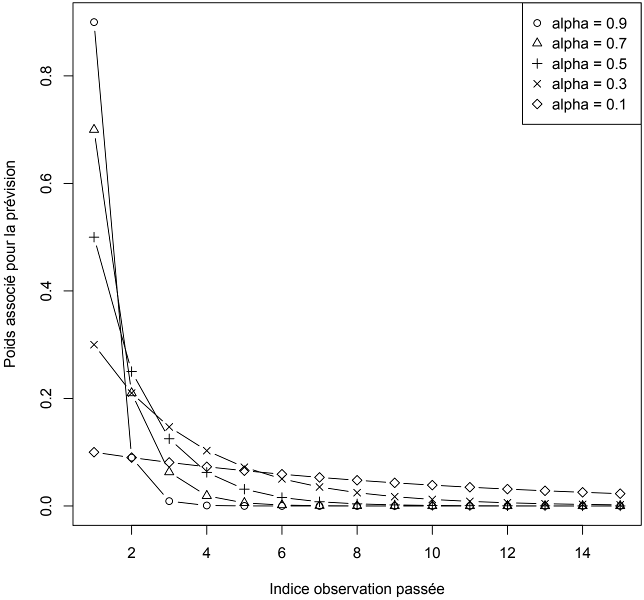
Principle:

Simple exponential smoothing : analysis

$$
\begin{aligned}
\hat{x}_{n+1} &= \alpha\, x_n + (1-\alpha)\, \hat{x}_n \\
&= \alpha\, x_n + (1-\alpha)\,(\, \alpha\, x_{n-1} + (1-\alpha)\, \hat{x}_{n-1}\,) \\
&= \alpha\, x_n + \alpha(1-\alpha)\, x_{n-1} + (1-\alpha)^2(\alpha x_{n-2} + (1-\alpha)\hat{x}_{n-2}) \\
&= \dots \\
&= \alpha\, x_n + \alpha(1-\alpha)\, x_{n-1} + \alpha(1-\alpha)^2\, x_{n-2} + \cdots + \\
&\alpha(1-\alpha)^k\, x_{n-k} + \alpha(1-\alpha)^{n-1}\, x_1 + (1-\alpha)^n\, \hat{x}_1
\end{aligned}
$$

Simple Exponential smoothing : example

Simple exponential smoothing : analysis



Influence of the parameter α

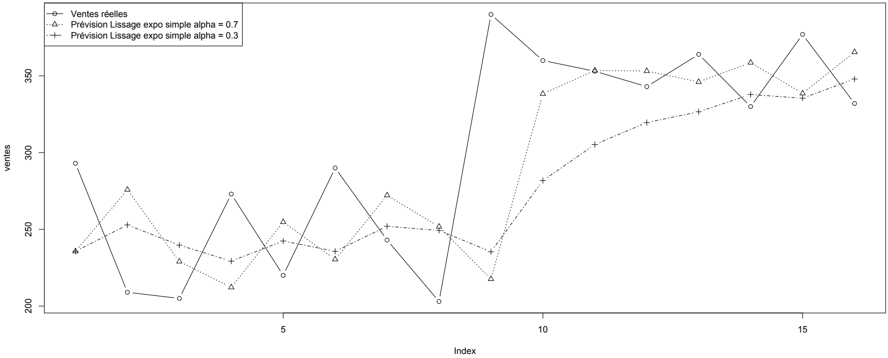
- if α is big (close to 1), we give a lot of influence to recent values. At the limit ($\alpha = 1$), I predict the last observed value

  $\rightarrow$ predictions are fluctuating; quick reaction to changes in the data

- if α is small (close to 0), more influence is given to past values (tends toward the mean of all past values)

  $\rightarrow$ predictions are stable but slow reactions to changes in the data

Simple Exponential Smoothing



Choice of α and $\hat{x}_0$

Auto-regressive models (AR)

Principle :

$$\hat{x}_n = \beta_0 + \sum_{k=1}^{p} \beta_k x_{n-k}$$

- p is the order of the model : AR(p)
- the coefficients β_j are estimated based on the original series

Auto-regressive models : example

How to choose the order p ?

For a given p, when a model is fitted (coefficients are estimated), we can compute the BIC score (Bayesian Information Criterion).

It is a tradeoff between accuracy of the intermediate predictions and the complexity of the model (number of coefficients)

The best value of p for a given time series can be chosen according to this score (the lower BIC, the better)

Summary