Master 2 Actuariat Parcours Data science pour l'actuariat

TIME SERIES

Cheikh MbackÃľ BEYE

Table des matières

Environment

```
knitr::opts_chunk$set(echo = TRUE,
                      fig.width=12,
                      comment = NA,
                      message = FALSE,
                      warning=FALSE,
                      background="#ccffcc"
                      )
```

```
library(ggplot2)
library(astsa)
library(xts)
```

INTRODUCTION

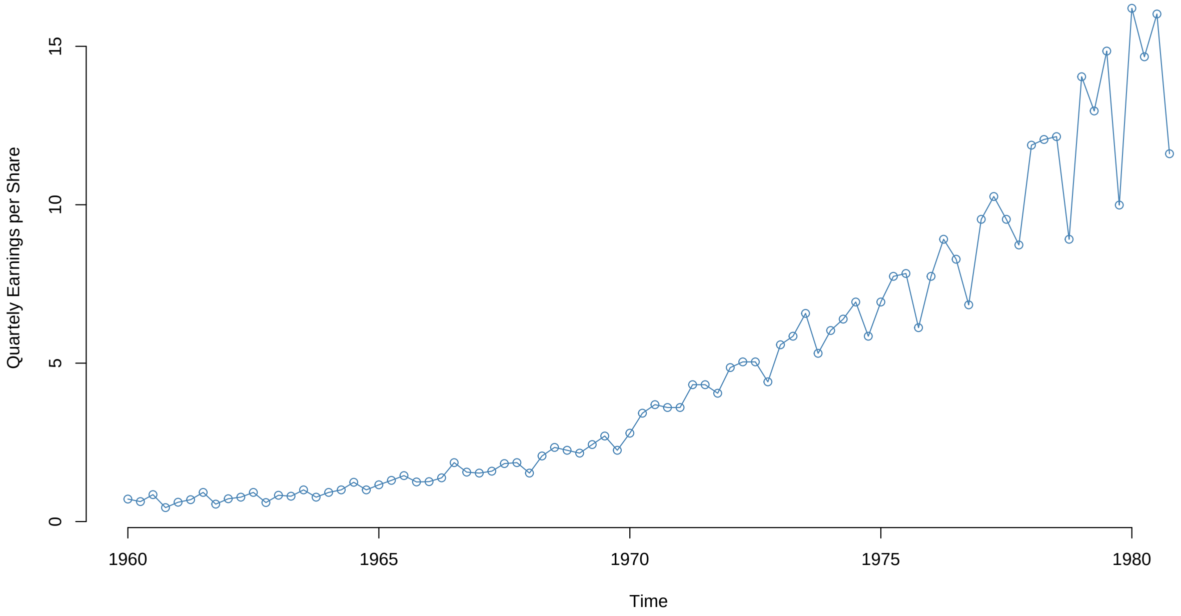
Data obtained from observations collected sequentially over time are extremely common. In bussines we observe weekly interest rates, daily closing stock prices, monthly prices indices, yearly sales figures, and so forth. In meterology, we observe daily temperatures, annual precipitation and drought indices and hourly wind speeds. In agriculture, we record annual figures for crop and livestock prooduction, soil erosion, and export sales. In the biological science , we observe the electrical activity of the heart at millisecond intervals. In ecology, we record the abundance of animal species. The list of areas in which **times series** are studies is virtually endless. The purpose of time series analysis is generaly twofold : to undestand or model the stochastic mechanism that gives rise to an observed series and to predict or forecast the future values of a series based on the history of that series and possibily other related factors. We will introduces a variety of examples of time series from diverse areas of application. A somewhat unique feature of time series and their models is that we usually cannot assume that the observations are independently from a common population. Studying models that incoperate dependence is the key concept in time series analysis.

TIME SERIES EXAMPLES

Johnson and Johnson Quarterly Earnings Per Share

```
jj

          Qtr1      Qtr2      Qtr3      Qtr4
1960  0.710000  0.630000  0.850000  0.440000
1961  0.610000  0.690000  0.920000  0.550000
1962  0.720000  0.770000  0.920000  0.600000
1963  0.830000  0.800000  1.000000  0.770000
1964  0.920000  1.000000  1.240000  1.000000
1965  1.160000  1.300000  1.450000  1.250000
1966  1.260000  1.380000  1.860000  1.560000
1967  1.530000  1.590000  1.830000  1.860000
1968  1.530000  2.070000  2.340000  2.250000
1969  2.160000  2.430000  2.700000  2.250000
1970  2.790000  3.420000  3.690000  3.600000
1971  3.600000  4.320000  4.320000  4.050000
```

```
1972  4.860000  5.040000  5.040000  4.410000
1973  5.580000  5.850000  6.570000  5.310000
1974  6.030000  6.390000  6.930000  5.850000
1975  6.930000  7.740000  7.830000  6.120000
1976  7.740000  8.910000  8.280000  6.840000
1977  9.540000 10.260000  9.540000  8.729999
1978 11.880000 12.060000 12.150000  8.910000
1979 14.040000 12.960000 14.850000  9.990000
1980 16.200000 14.670000 16.020000 11.610000

tseries<-data(jj)
plot(jj,
     type='o',
     ylab='Quartely Earnings per Share',
     frame=FALSE,
     col='steel blue'
     )
```

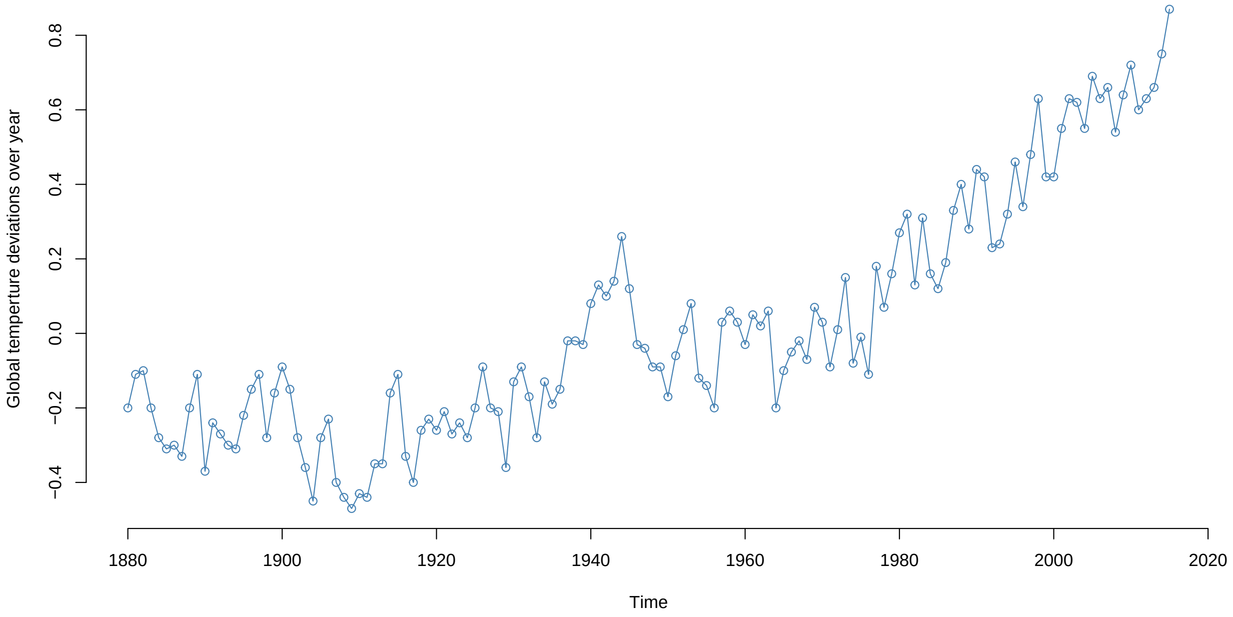


This figure shows quaterly earnings per shape for the US company Johnson and Johnson, furnished by Professor Paul Griffin. There are 84 quaters (84/4=21 years) measured from the first quarter of 1960 to the last quarter of 1980. Modelling such a series begins by observing the primary patterns in the time history ?. In this case, note the gradualy increasing underlying trend anf the rather regular variation superimposedon on the trend that seems to repeat over quarters .

Global mean land-ocean temperature deviations to 2015

```
globtemp

Time Series:
Start = 1880
```

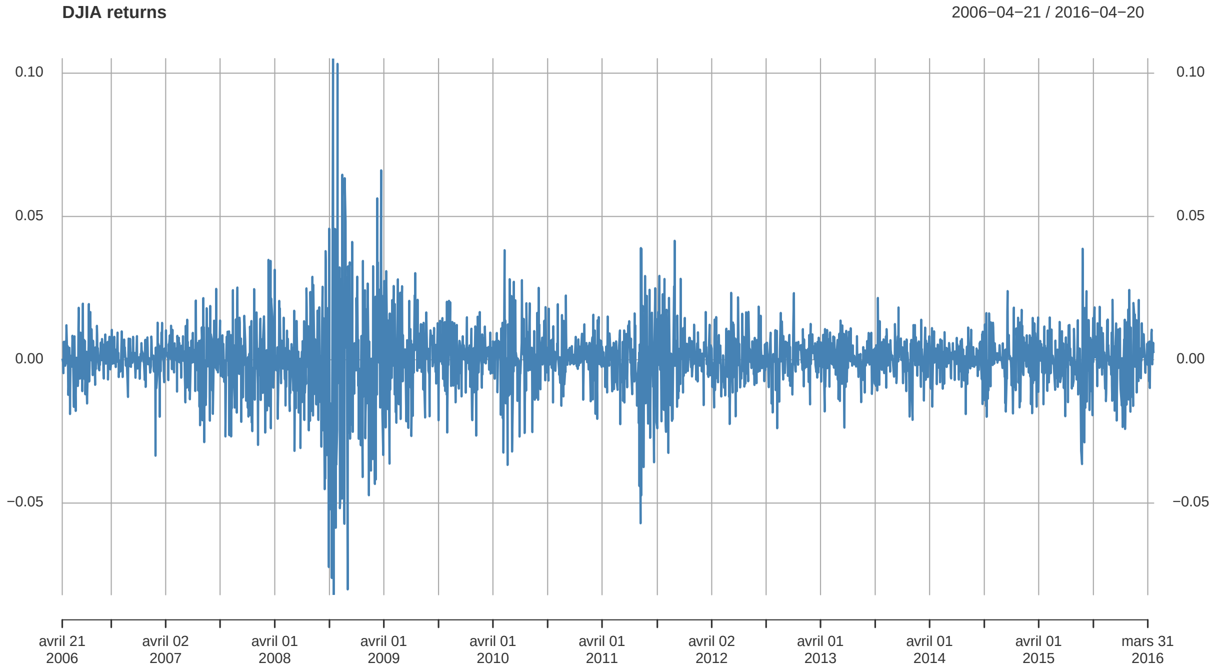
```
End = 2015
Frequency = 1
  [1] -0.20 -0.11 -0.10 -0.20 -0.28 -0.31 -0.30 -0.33 -0.20 -0.11 -0.37 -0.24
 [13] -0.27 -0.30 -0.31 -0.22 -0.15 -0.11 -0.28 -0.16 -0.09 -0.15 -0.28 -0.36
 [25] -0.45 -0.28 -0.23 -0.40 -0.44 -0.47 -0.43 -0.44 -0.35 -0.35 -0.16 -0.11
 [37] -0.33 -0.40 -0.26 -0.23 -0.26 -0.21 -0.27 -0.24 -0.28 -0.20 -0.09 -0.20
 [49] -0.21 -0.36 -0.13 -0.09 -0.17 -0.28 -0.13 -0.19 -0.15 -0.02 -0.02 -0.03
 [61]  0.08  0.13  0.10  0.14  0.26  0.12 -0.03 -0.04 -0.09 -0.09 -0.17 -0.06
 [73]  0.01  0.08 -0.12 -0.14 -0.20  0.03  0.06  0.03 -0.03  0.05  0.02  0.06
 [85] -0.20 -0.10 -0.05 -0.02 -0.07  0.07  0.03 -0.09  0.01  0.15 -0.08 -0.01
 [97] -0.11  0.18  0.07  0.16  0.27  0.32  0.13  0.31  0.16  0.12  0.19  0.33
[109]  0.40  0.28  0.44  0.42  0.23  0.24  0.32  0.46  0.34  0.48  0.63  0.42
[121]  0.42  0.55  0.63  0.62  0.55  0.69  0.63  0.66  0.54  0.64  0.72  0.60
[133]  0.63  0.66  0.75  0.87

tseries<-data(jj)
plot(globtemp,
     type='o',
     ylab='Global temperture deviations over year',
     frame=FALSE,
     col='steel blue'
     )
```



The figure shows the global temperature series record. The data are the global mean land-ocean temperature index from 1880 to 2015. We note an apparent upward trend in the series during the latter part of the twentieth century that has been used as an argument for the gloabal warming hypothesis. Note also the leveling off at about 1935 and then another rather sharp upward trend at about 1970.

Dow Jones Industrial Average

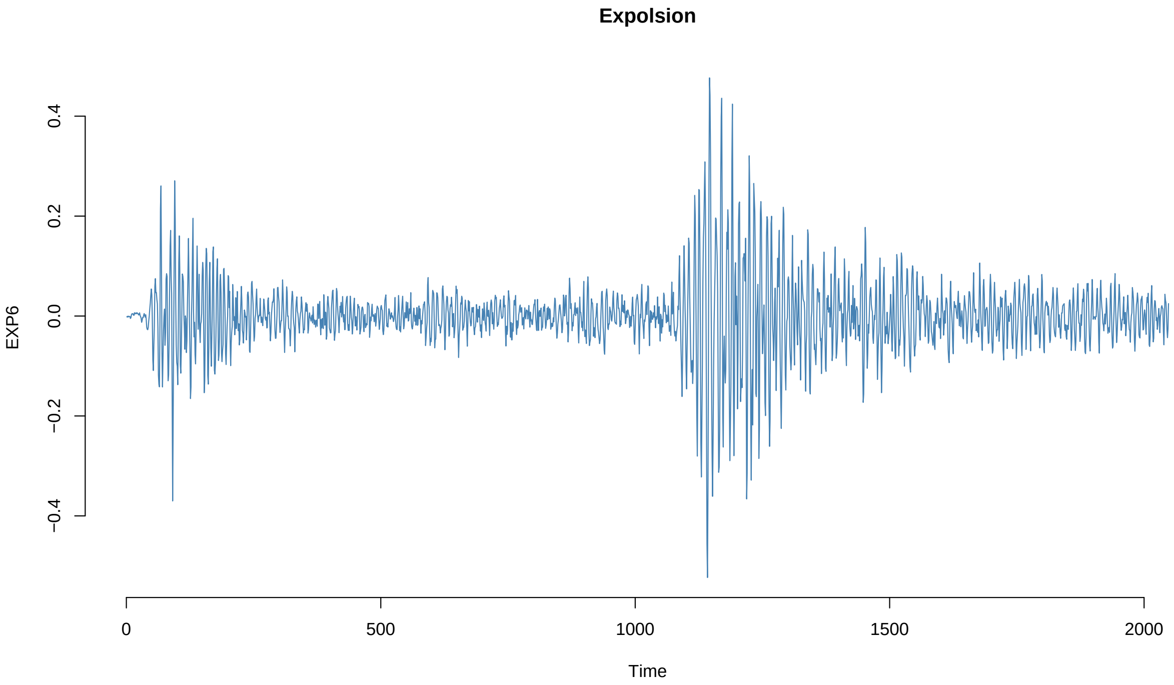
```
djia[1:10]

                 Open     High      Low    Close    Volume
2006-04-20 11278.53 11384.11 11275.05 11342.89 336420000
2006-04-21 11343.45 11405.88 11316.79 11347.45 325090000
2006-04-24 11346.81 11359.70 11305.83 11336.32 232000000
2006-04-25 11336.56 11355.37 11260.84 11283.25 289230000
2006-04-26 11283.25 11379.87 11282.77 11354.49 270270000
2006-04-27 11349.53 11416.93 11275.30 11382.51 361740000
2006-04-28 11358.33 11417.66 11347.21 11367.14 738440000
2006-05-01 11367.78 11428.37 11329.44 11343.29 365970000
2006-05-02 11345.21 11427.65 11345.13 11416.45 335420000
2006-05-03 11414.69 11424.93 11362.42 11400.28 380540000

djiaReturns<-diff(log(djia$Close))[-1]
plot(djiaReturns,
     main='DJIA returns',
     frame=FALSE,
     col='steel blue'
     )
```



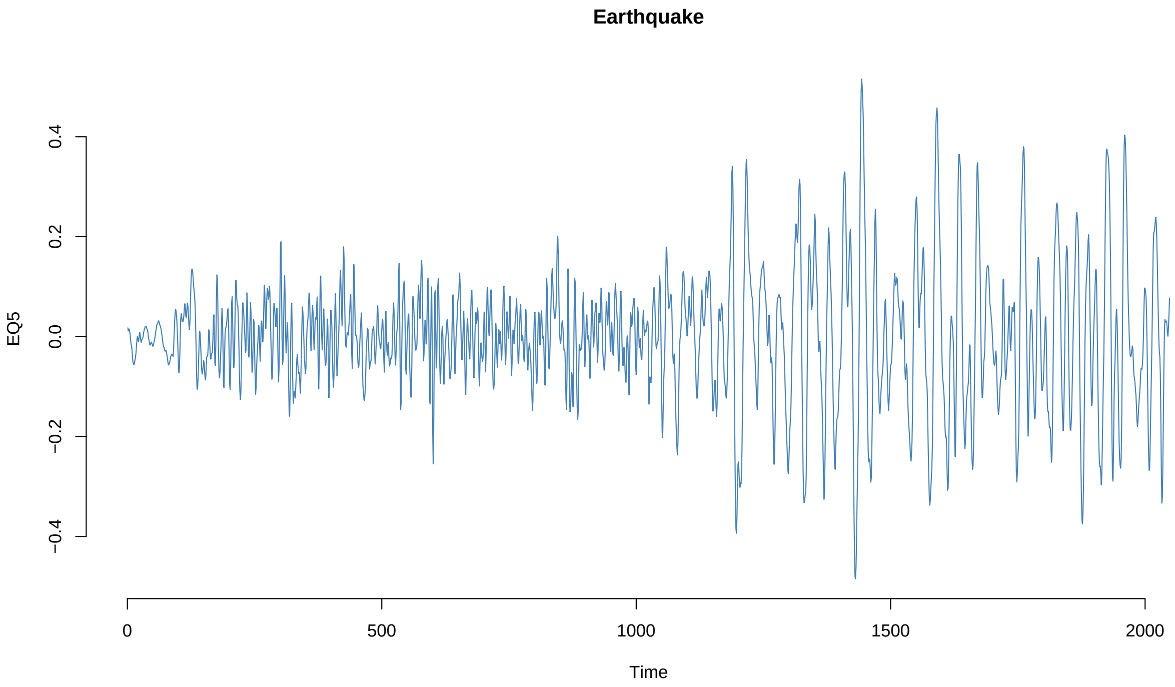
This an example of financial time series data. It shows the daily returns of the Dow Jones Industrial Average from april 20, 2006 to april 20, 2016. This a typical of return data. The mean of the series appears to be stable with an average return approximately zero. A problem in the nalaysis of these type of financial data is to forcast the volatility of future returns. It's easy to spot the financial crisis of 2008.

Seismic Trace of Explosion and Earthquake

```
plot(EXP6,
     main='Expolsion',
     frame=FALSE,
     col='steel blue'
     )
```



```
plot(EQ5,
     main='Earthquake',
     frame=FALSE,
     col='steel blue'
)
```

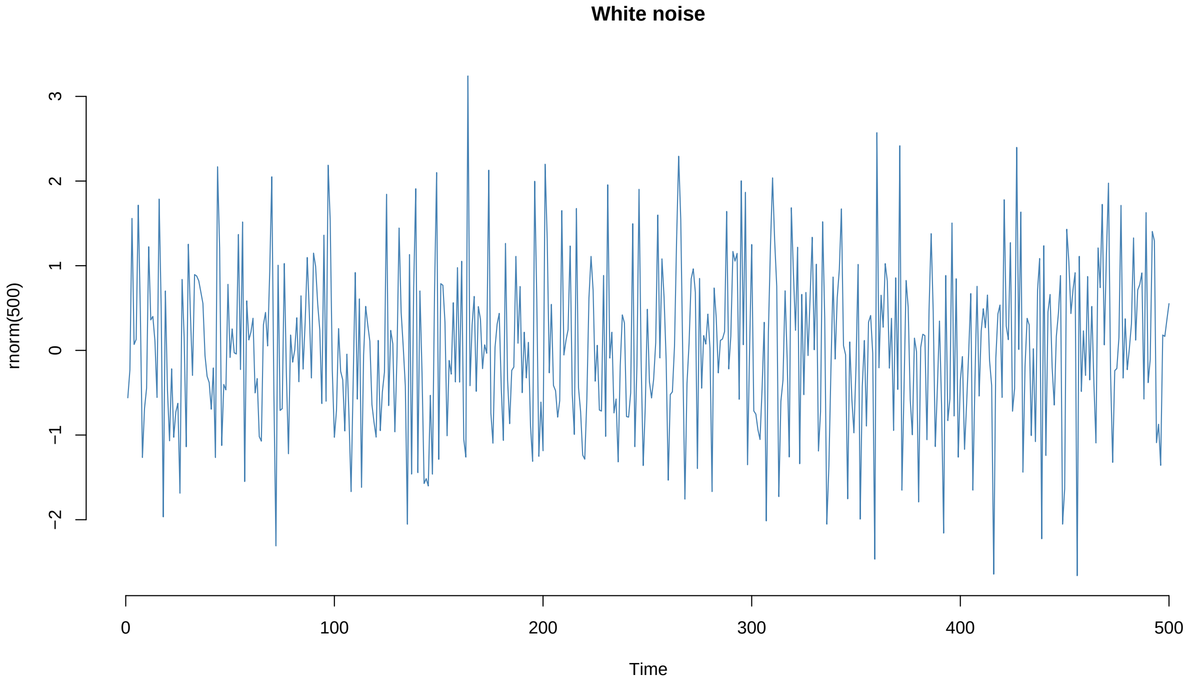


These two last examples represent two phases denoted by $P(t = 1, ...; 1024)$ and $S(t = 1025, ..., 2048)$ at a seismic recording station. The recording instruments in Scandinavia are observing earthquake and mining explosion. The general problem of intrest is in distinguishing or discriminating between waveforms generated by earthquakes and those generated by explosion . We can also focus oÃğn the amplitude ratios between the two phases,which tend to be smaller for earhtquakes than for explosions.

WHITE NOISE

A simple kind of generated series might be a collection of uncorrelated random variables, $(\epsilon_t)_{t\in\mathbb{Z}}$ with mean 0 and finite variance σ^2. The time series generated from uncorrelated variables is used as a model for noise in engenerring applications where it is called *white noise*. The designation white originates from the analogy with white light and indicates that all possible periodic oscilllations are present with equal strength.
We will sometimes require the noise to be independent and identically distributed (iid). A particulary useful white noise series is Gaussian white noise where $\epsilon_t \approx \mathcal{N}(0, \sigma^2)$

```
set.seed(123)
plot.ts(rnorm(500),
        frame=FALSE,
        main='White noise',
        col='steel blue'
        )
```



We note mixture of many diffirent kinds of oscillations. But the white noise alone cannot explained all time series behavior. If it was the case, classical statistical methods would suffice. To model a time series for forcasting or predicting purpose, we should take account of serial correlation between observations.

MOVING AVERAGE

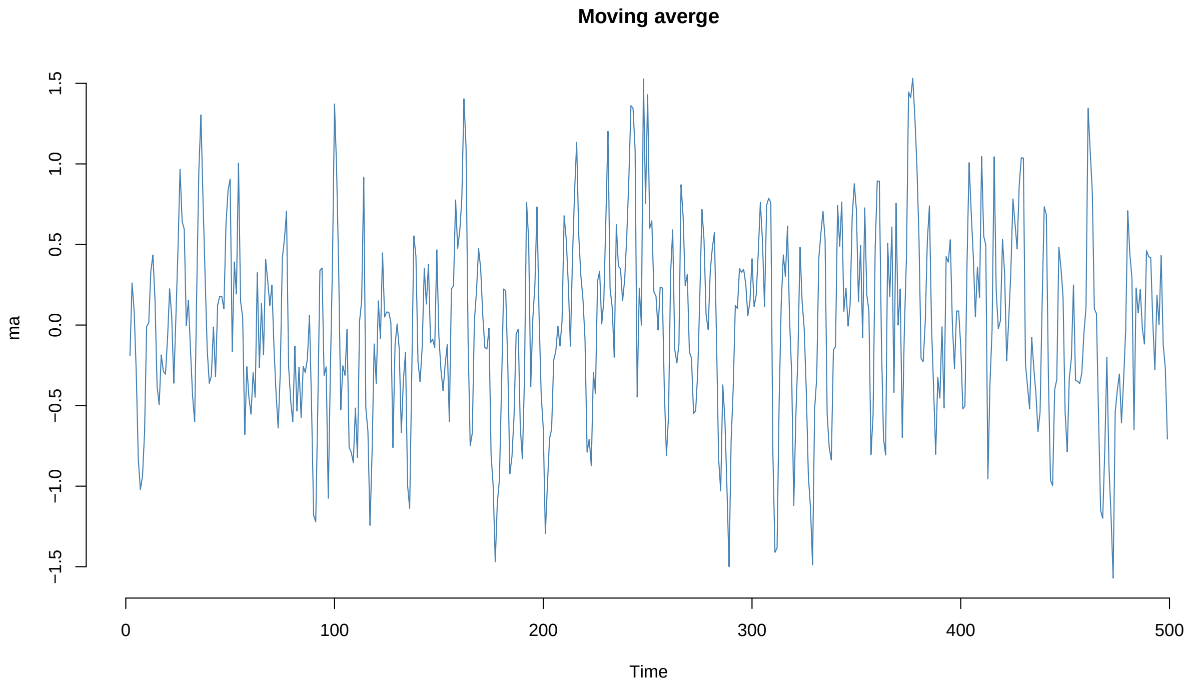
We might replace the white noise series ϵ_t by a moving average that smooths the series. For example, consider replacing ϵ_t by an average of its current value and its immediate neighbors in the past and future. That is, let

$$X_t = \frac{1}{3}(\epsilon_{t-1} + \epsilon_t + \epsilon_{t-2})$$

To get the moving averge we can use the command *filter* in *stats* package.
filter(*x*, *filter*, *method*= c("convolution", "recursive"),*sides*= 2, *circular* = FALSE)
Arguments :

1. x a univariate or multivariate time series.
2. filter a vector of filter coefficients in reverse time order (as for AR or MA coefficients).
3. method Either *convolution* or *recursive* (and can be abbreviated). If "convolution" a moving average is used : if "recursive" an autoregression is used.
4. sides for convolution filters only. If sides = 1 the filter coefficients are for past values only ; if sides = 2 they are centred around lag 0. In this case the length of the filter should be odd, but if it is even, more of the filter is forward in time than backward.
5. circular for convolution filters only. If TRUE, wrap the filter around the ends of the series, otherwise assume external values are missing (NA).

```
ma<-stats::filter(rnorm(500),
                  sides = 2,
                  method = 'convolution',
                  filter=rep(1/3,3),
                )
plot.ts(ma,
     main = 'Moving averge',
     frame.plot = FALSE,
     col='steel blue'
     )
```



Oscillations are more apparent and some of the faster oscillations are taken out.

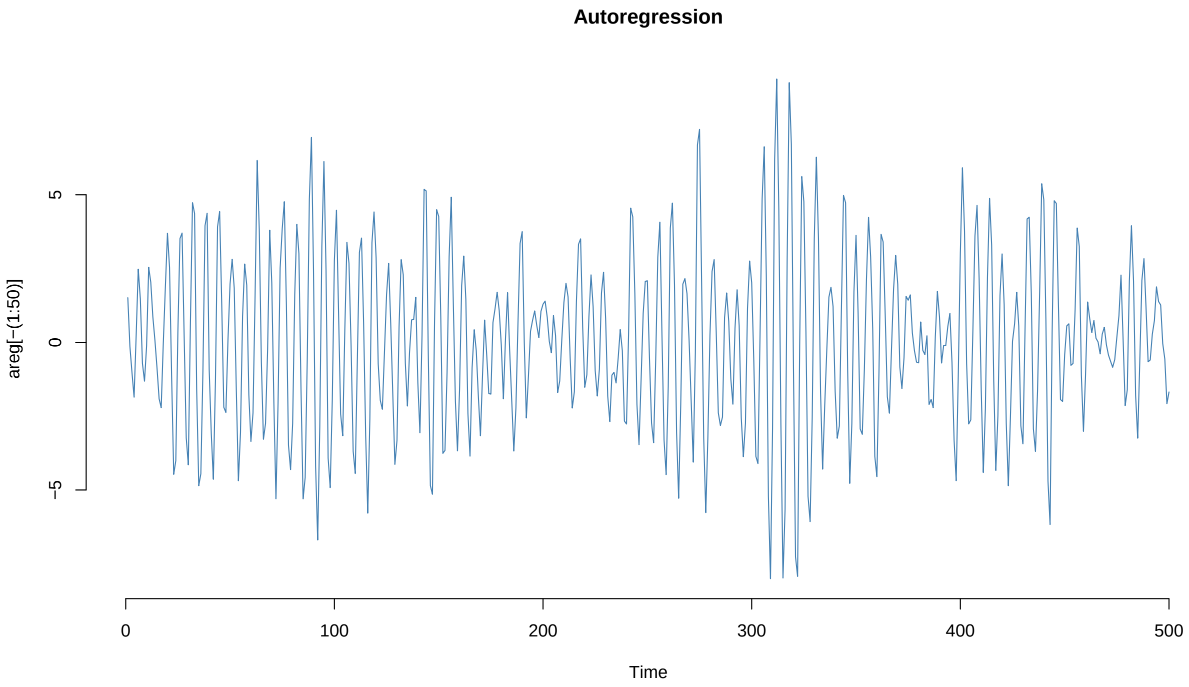
AUTOREGRESSIONS

$$X_t = X_{t-1} - 9X_{t-2} + \epsilon_t$$

This equation represnts a regression or prediction of the current value X_t of a times series as function of the past two values of the series. A problem with startup values exists because the equation depends on the initial conditions.

The function $filter$ uses zero for the initial values. To fix that we can run $filter$ for more than needed and remove the initial valaues.

```
areg<-stats::filter(rnorm(550),
                  method = 'recursive',
                  filter=c(1,-.9)
                )
plot.ts(areg[-(1:50)],
     main = 'Autoregression',
     frame.plot = FALSE,
     col='steel blue'
     )
```



RANDOM WALK WITH DRIFT

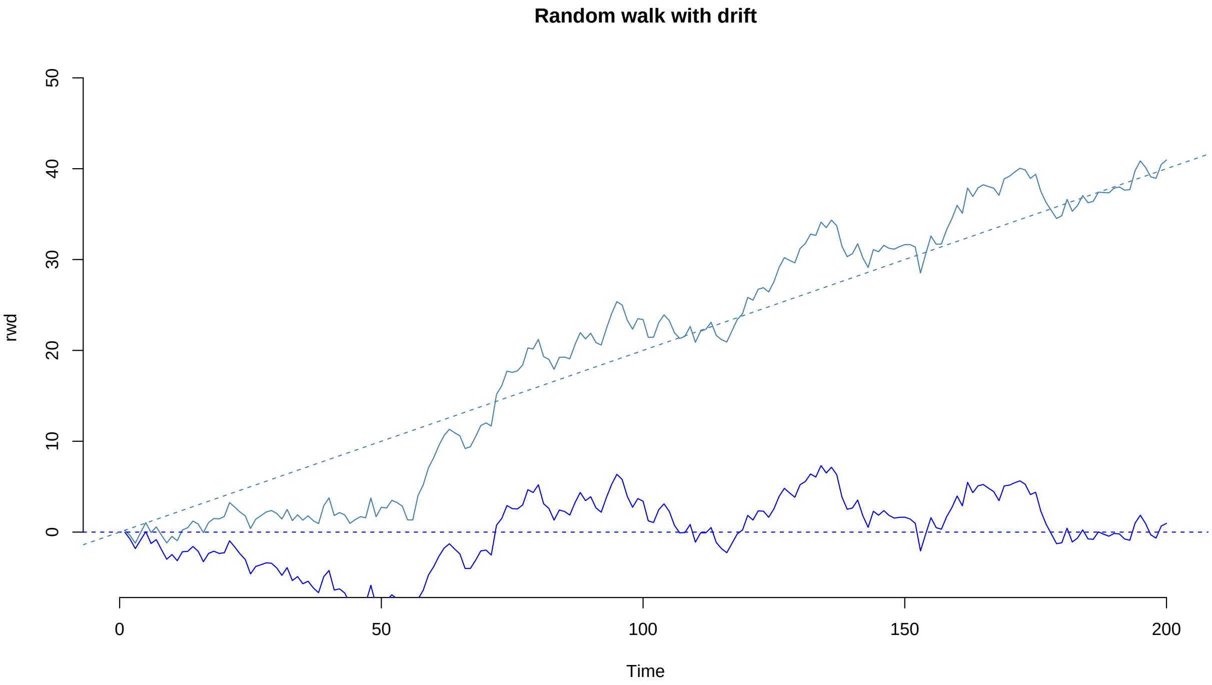
A model for anlysing trend such as seen in the global temperature data is the random walk with drift model given by

$$X_t = \delta + X_{t-1} + \epsilon_t; \quad X_0 = 0$$

δ is called drift and when $\delta = 0$ is called simply a random walk.

$$X_t = \delta t + \sum_{i=1}^{t} \epsilon_i$$

```
wn<-rnorm(200)
rwd<-cumsum(wn+0.2)
rw<-cumsum(wn)
plot.ts(rwd,
        main='Random walk with drift',
        col='steel blue',
        frame=FALSE,
        ylim=c(-5,50)
        )
lines(rw,
      col='blue'
      )
abline(h=0,
       lty=2,
       col='blue'
       )
abline(a=0,
       b=0.2,
       lty=2,
       col=' steel blue'
       )
```



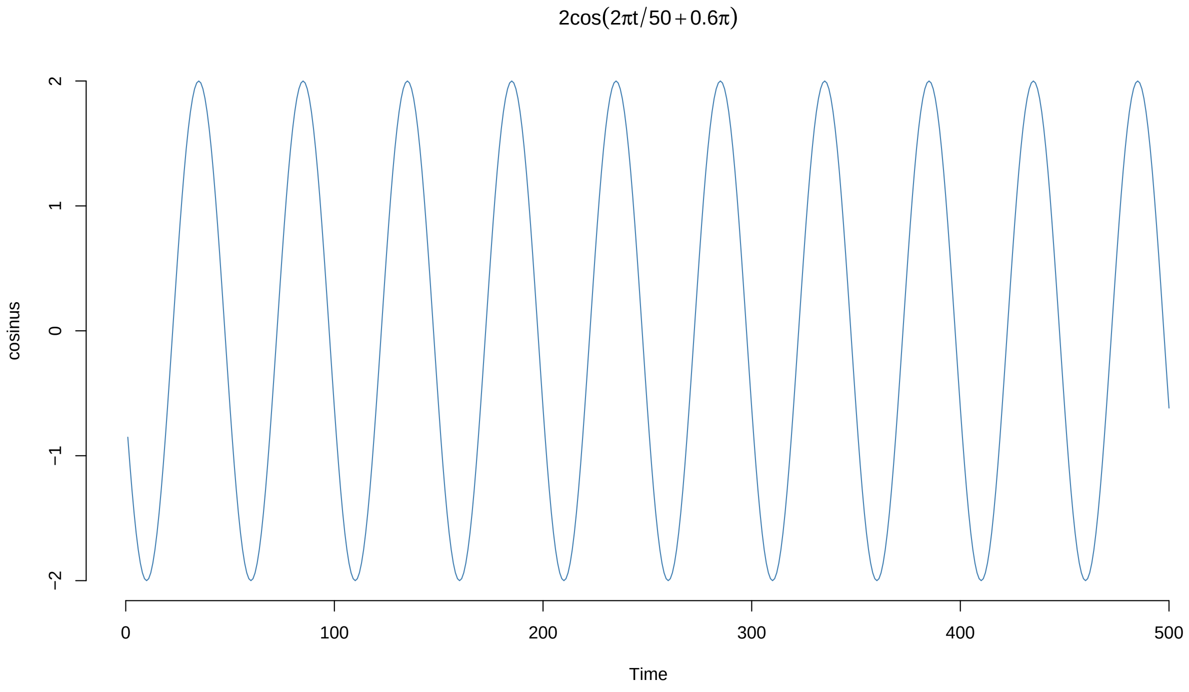
SIGNAL IN NOISE

Many realistic models for generating time series assume an undeerlying signal with some consitent periodic variation , contaminated by adding a random noise.

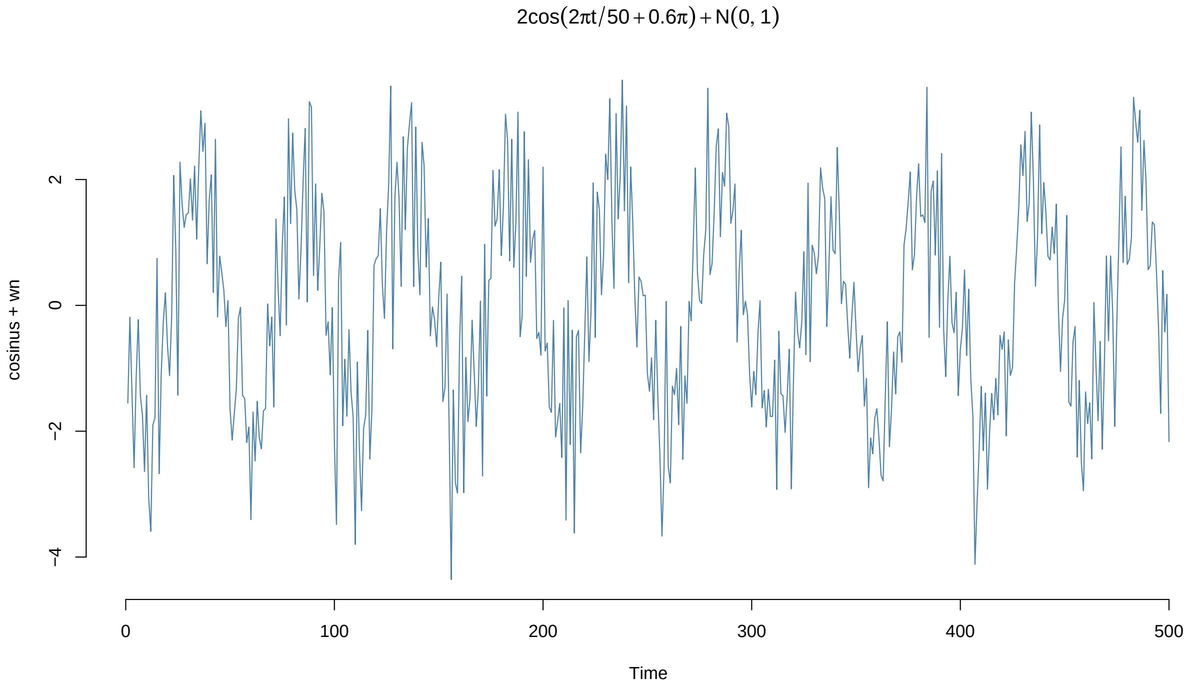
$$X_t = 2\cos\left(2\pi\frac{t+15}{50}\right) + \epsilon_t$$

1. A=2 is the amplitude
2. $\omega = \frac{1}{50}$ is the frequency of oscillations (one cycle every 50 time points)
3. $\phi = 2\pi\frac{15}{50} = 0.6\pi$ is the phase shift

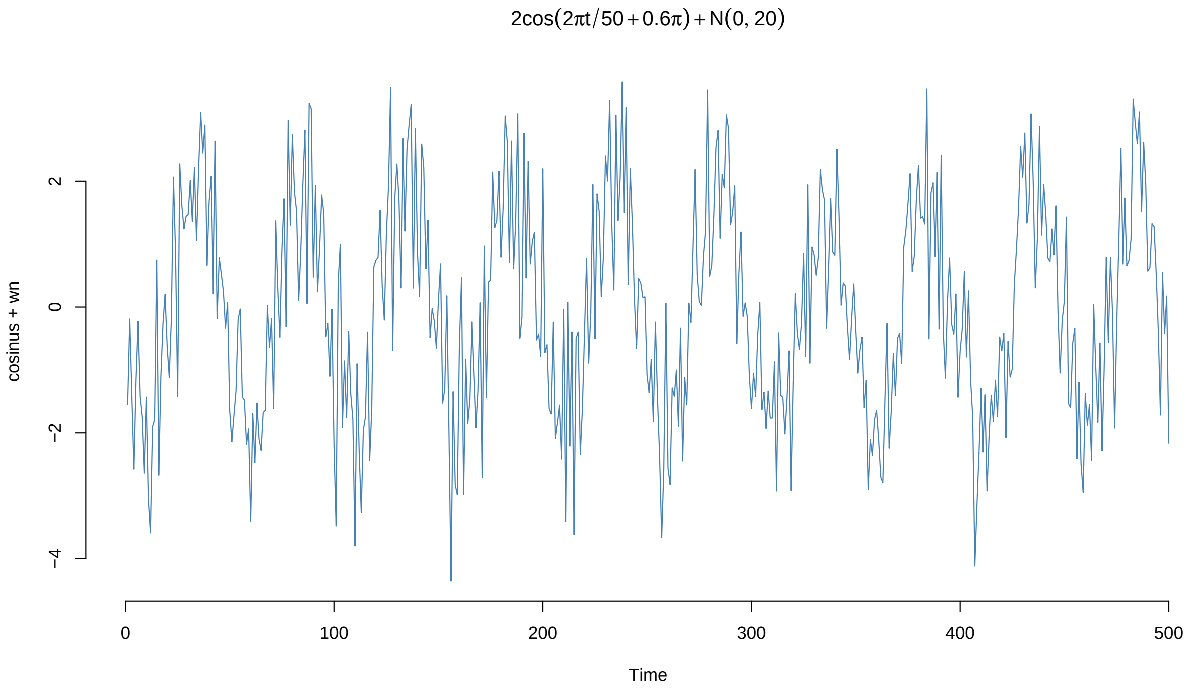
```
wn<-rnorm(500)
cosinus<-2*cos(2*pi*1:500/50+.6*pi)
plot.ts(cosinus,
        col='steel blue',
        frame=FALSE,
        main=expression(2*cos(2*pi*t/50+.6*pi))
        )
```



```
plot.ts(cosinus+wn,
        col='steel blue',
        frame=FALSE,
        main=expression(2*cos(2*pi*t/50+.6*pi)+N(0,1))
        )
```



```r
plot.ts(cosinus+wn,
        col='steel blue',
        frame=FALSE,
        main=expression(2*cos(2*pi*t/50+.6*pi)+N(0,20))
        )
```



CLASSICAL REGRESSION IN TIMES SERIES CONTEXT

We begin our discusssion of linear regression in the time series context by assiming some ouput or dependaent time series say , X_t for $t = 1, ..., t = n$ is being influenced by a collection of possible inputs or independent series say, $Z_{t_1}, ..., Z_{t_q}$, where we first regard the inputs as fixed and known. This assumption is for applying conventional linear regression. We express this relationship through the model

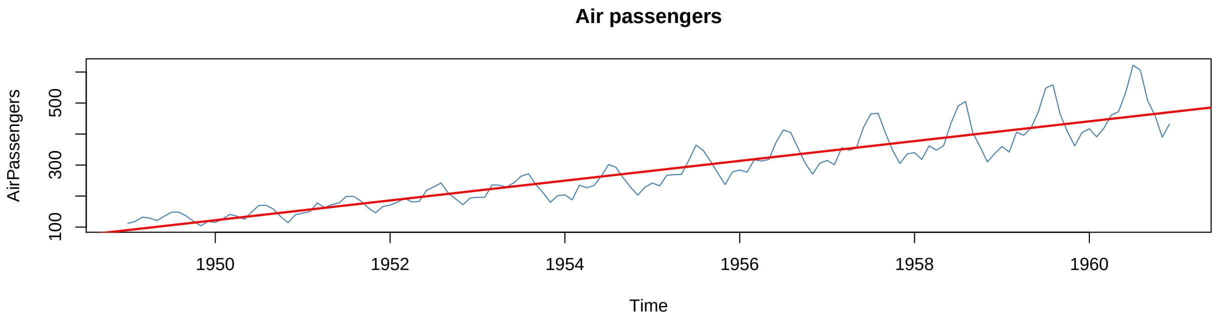
$$X_t = \beta_0 + \beta_1 Z_{t_1} + ... + \beta_q Z_{t_q} + \epsilon_t$$

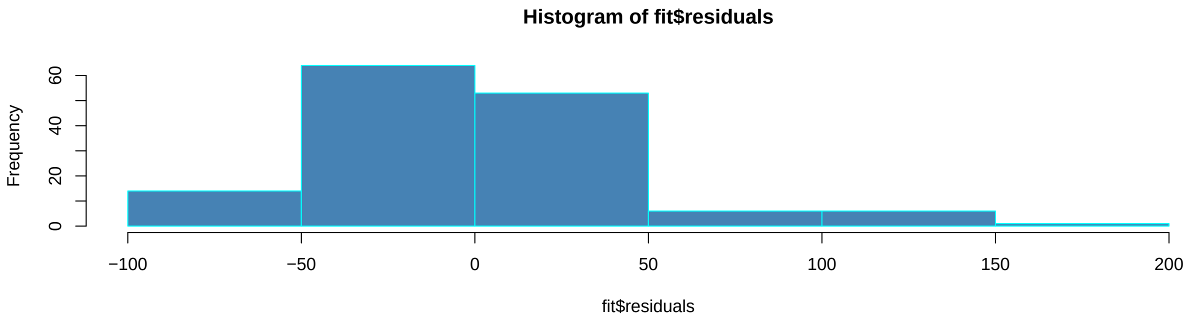
Where the paremeters $\beta_0, \beta_1, ..., \beta_q$ are unknown and $(_t)$ is random error or noise process consisting of indepndent and identically distrinuted normal variables.

Example

```r
fit<-lm(AirPassengers~time(AirPassengers),na.action = NULL)
summary(fit)


Call:
lm(formula = AirPassengers ~ time(AirPassengers), na.action = NULL)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-93.858 -30.727  -5.757  24.489 164.999 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -62055.907   2166.077  -28.65   <2e-16 ***
time(AirPassengers)     31.886      1.108   28.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.06 on 142 degrees of freedom
Multiple R-squared:  0.8536,Adjusted R-squared:  0.8526 
F-statistic: 828.2 on 1 and 142 DF,  p-value: < 2.2e-16

par(mfrow=c(2,1))
plot(AirPassengers,col='steel blue',main='Air passengers')
abline(fit,col='red',lwd=2)
hist(fit$residuals,col = 'steel blue', border = 'cyan')
```


Air passengers
AirPassengers
Time


Histogram of fit$residuals
Frequency
fit$residuals