Data Understanding and pre-processing Time Series

Anna Monreale and Francesca Naretto
Computer Science Department

Introduction to Data Mining, 2nd Edition
Chapter 1 & Data Exploration (Additional Resources)

Also for time series: know your data

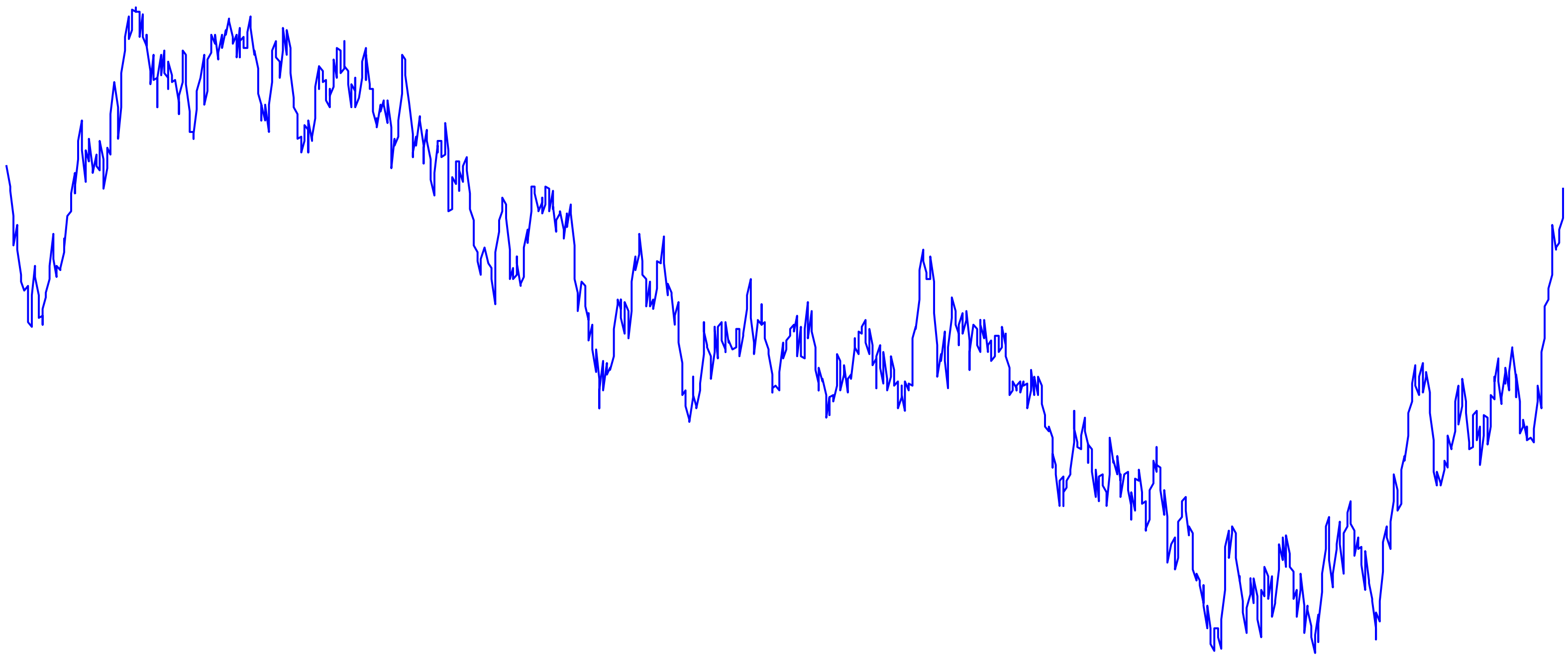
- For preparing data for data mining task it is essential to have an overall understanding of your data
- Gain insight in your data
  - with respect to your project goals
  - and general to understand properties
- Find answers to the questions
  - How is the data quality?
  - What about outliers?

Which is the type of data?

Types of data sets

Time series

A collection of observations that are sequential in time, generally at constant time intervals.



Series

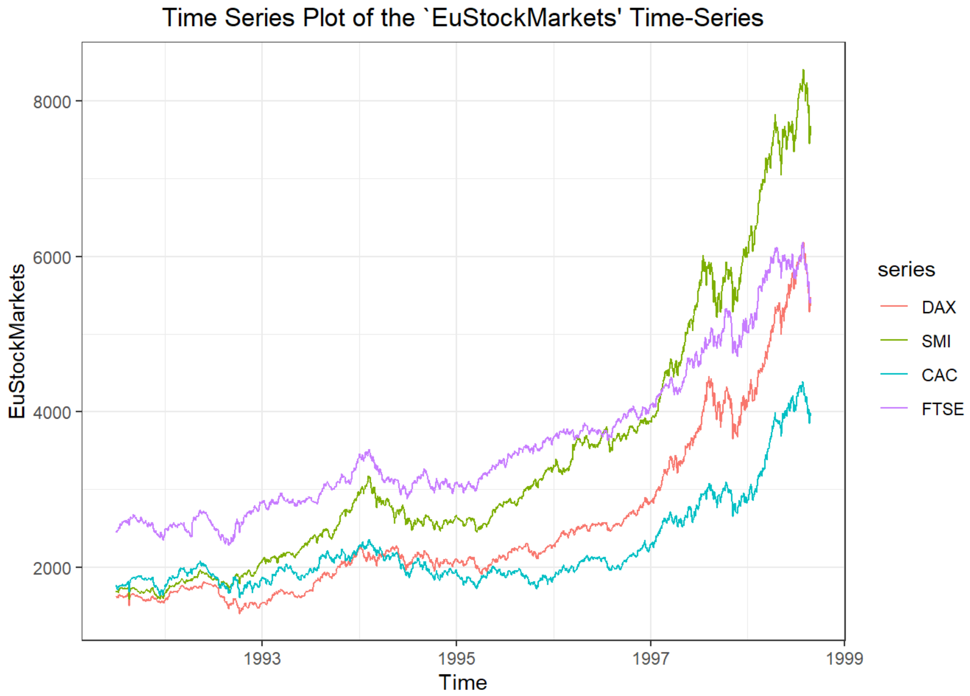
A univariate series x is a sequence of values $[x_1, x_2, \ldots x\_n]$ in a domain X.

A series is defined by:

- **Type**: discrete, e.g., nucleotide bases, or continuous, e.g., stock values in a financial market
- **Sampling rate**: How often values are sampled, e.g., daily
- **Amplitude**: Values sampled, e.g., value of the stock on a particular day

Series

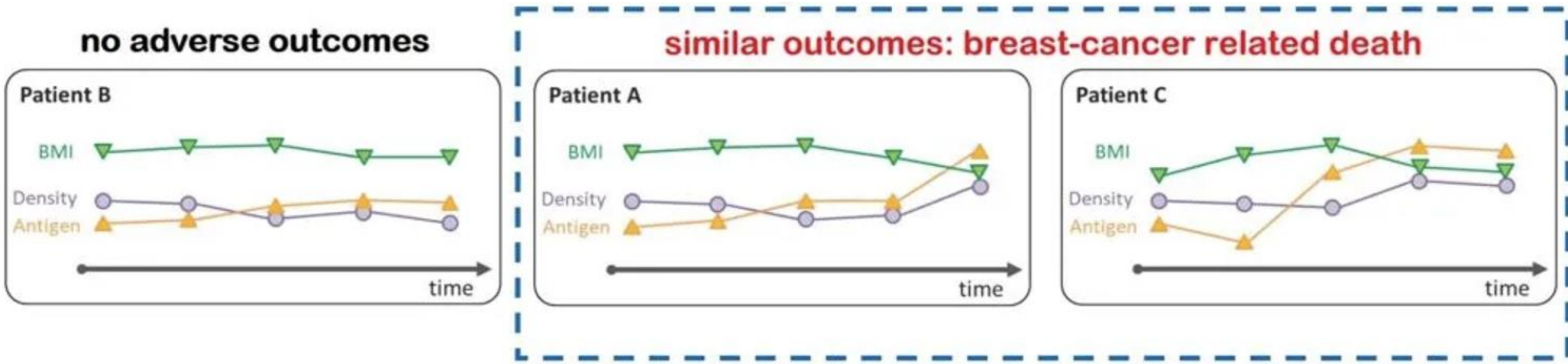
A multivariate time series x is a sequence that generalizes to multiple variables. Each instance is comprised of multiple time series, each representing a different feature.

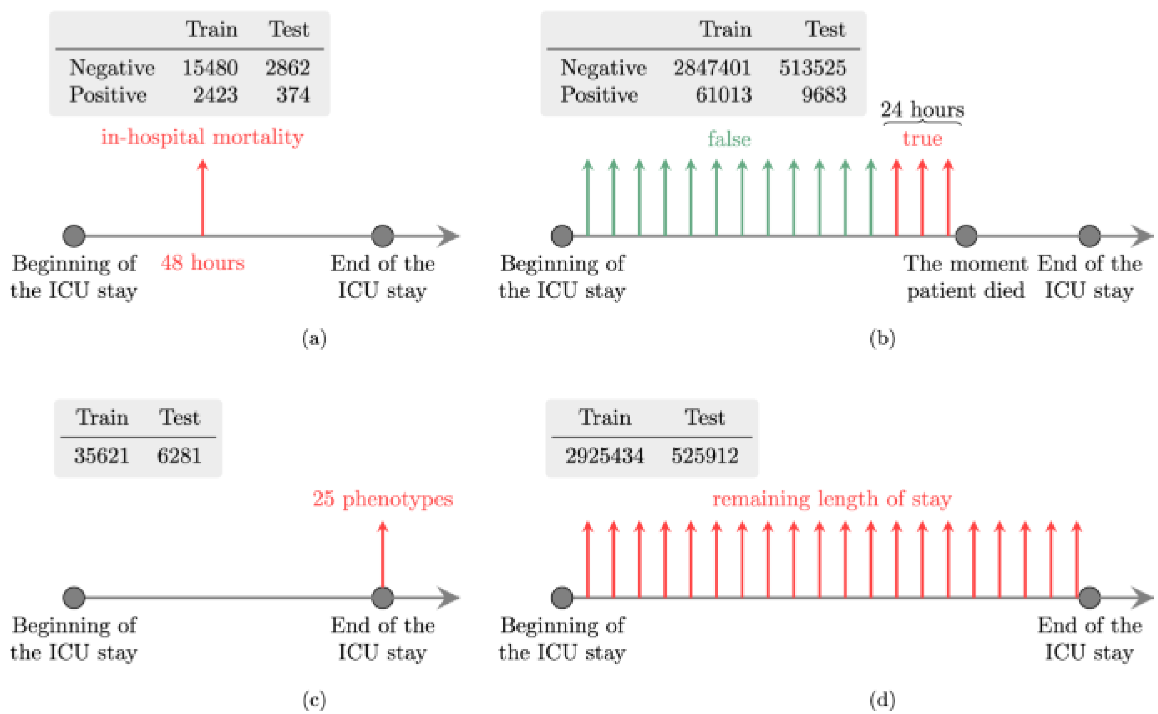


TS are ubiquitos

- You can measure many things … and things change over time.
  - Blood pressure
  - Donald Trump’s popularity rating
  - The annual rainfall in Pisa
  - The value of your stocks
- In addition other data type can be considered as time series
  - Text data: words count
  - Images: edges displacement
  - Videos: object positioning

TS are ubiquitos





Time series characteristics

- Large amount of data.
- Similarity is not easy to estimate.
- Different data formats.
- Different sampling rates.
- Noise, missing values, etc.

Time series understanding

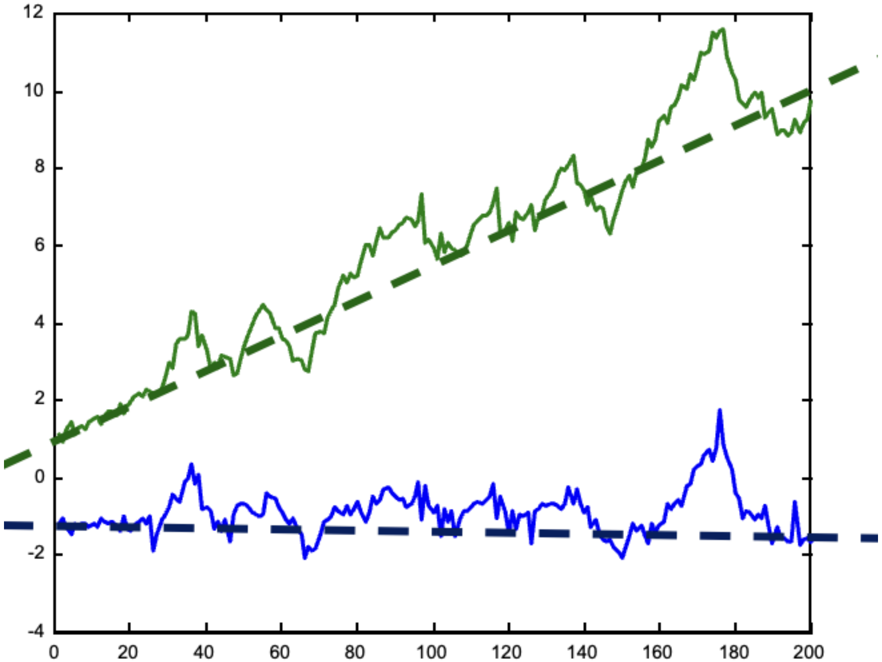
1. Look for trends
2. Check for seasonality, cyclicity, irregularities
3. Look for noise

Time series statistics

- **Mean**: the expected value of the time series
- **Variance**: variance of the time series
- **Trends**: the slope of a linear model that models the time series behavior
- **Interquartile ranges**: check the distributions
- **Skewness**: is the distribution symmetric?
- **Kurtosis**: what is the probability mass on the tails?

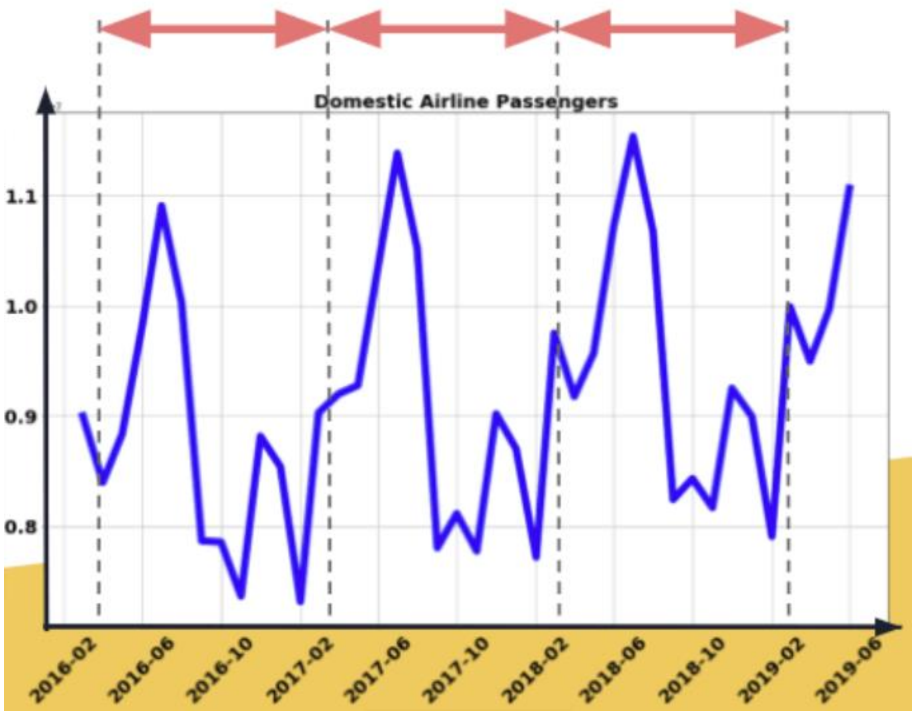
Trend

It is a long-term movement of the time series. It is non repeating. Technically, it is a slope $delta$ of a linear model, modelling the time series *x*.



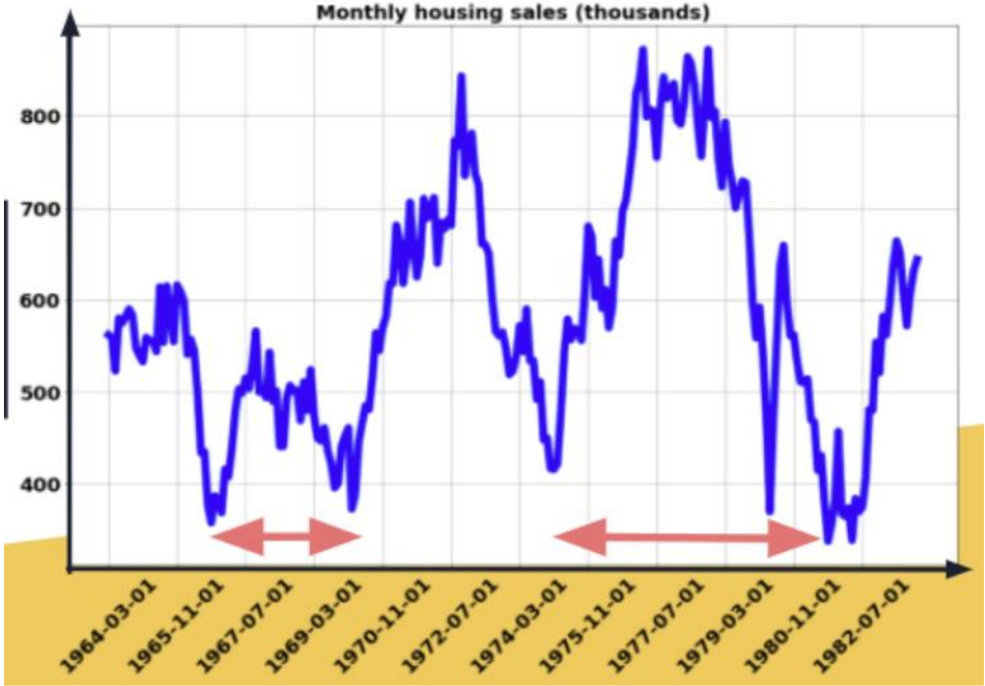
Seasonality

It is a regular periodic occurrence within a time interval, usually smaller than a year.



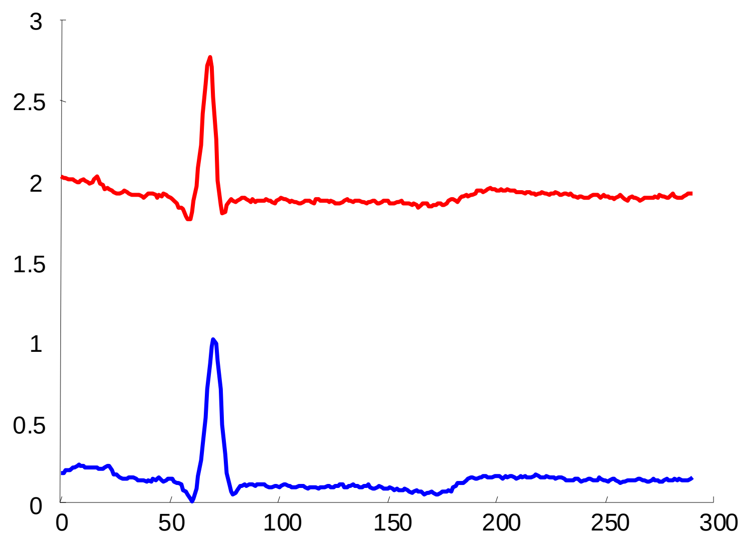
Cycle

It is a repeated fluctuation long in duration but not as much as a trend.



Time series analysis

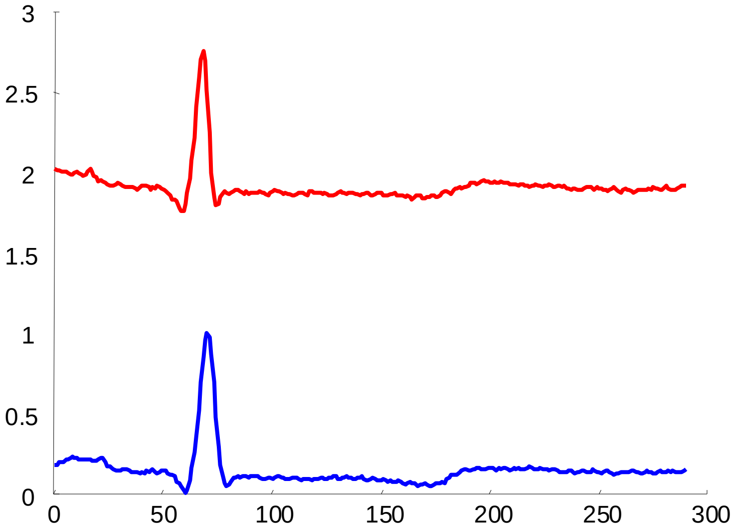
To analyze and compare different time series, we first need to pre-process them such that they have all the same format.

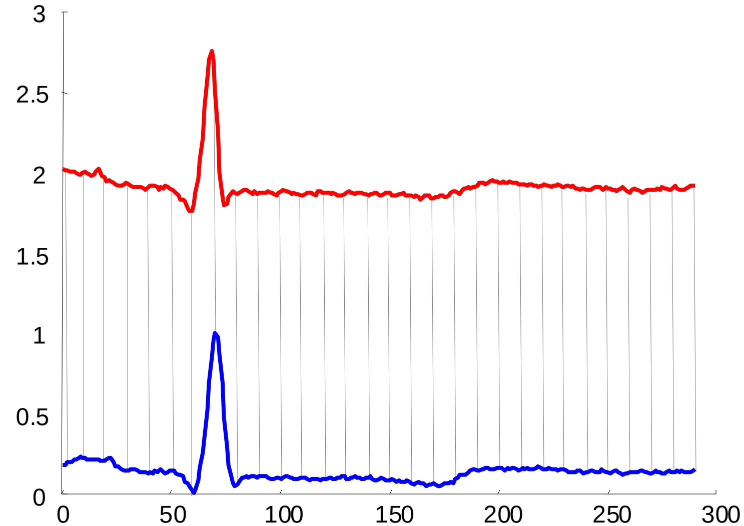


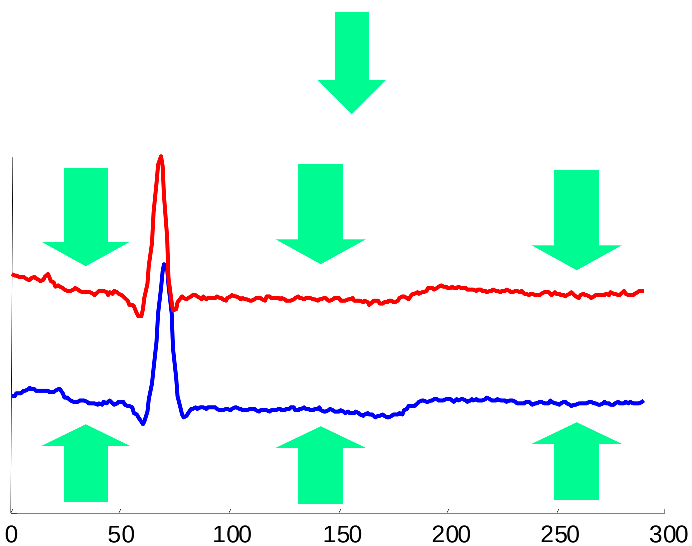
Time series analysis

- Often we need to employ Euclidean distance to analyze/compare time series. Euclidean distance is very sensitive to “distortions” in the data.
- These distortions are dangerous and should be removed.
- Most common distortions:
  - Offset Translation
  - Amplitude Scaling
  - Linear Trend
  - Noise
- They can be removed by using the appropriate transformations.

Offset translation to remove distortions



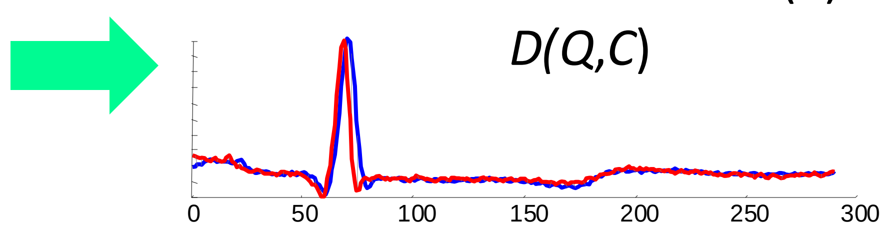




$Q = Q$ - mean(Q)

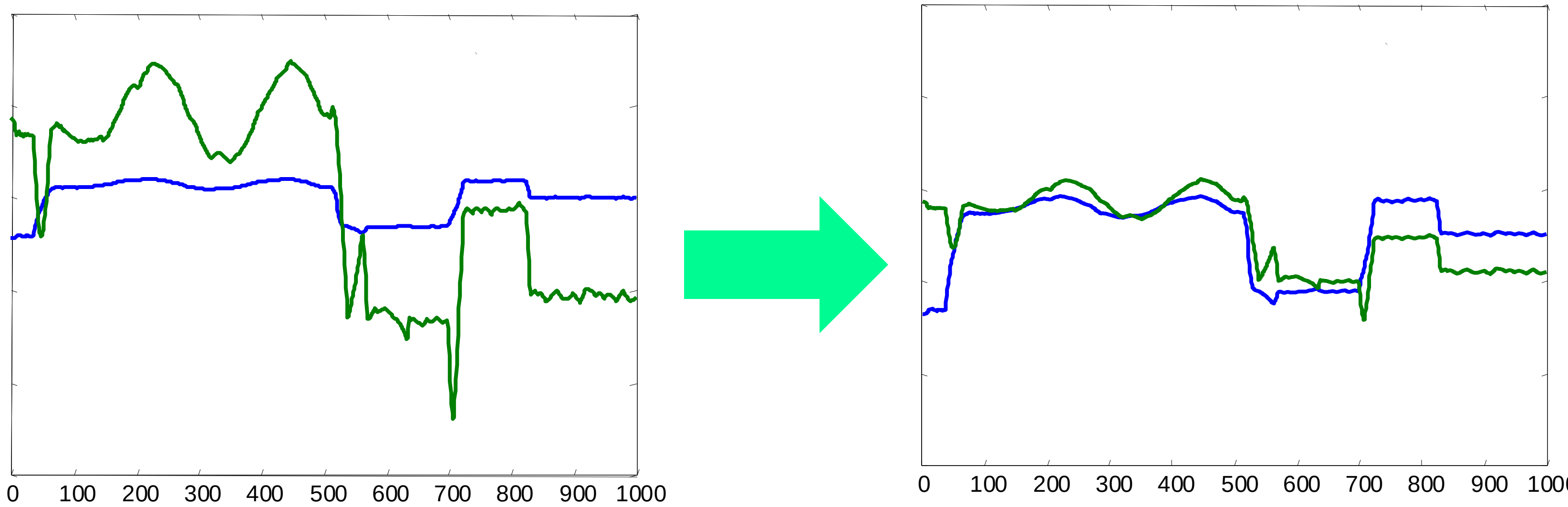
$C = C$ - mean(C)

$D(Q,C)$



Amplitude scaling

Objective: compare inherent patterns in different TS independently of their magnitudes.
Normalize the amplitude: divide by the standard deviation of the TS.



$$Q = (Q - \text{mean}(Q)) / \text{std}(Q)$$

$$C = (C - \text{mean}(C)) / \text{std}(C)$$

$D(Q,C)$

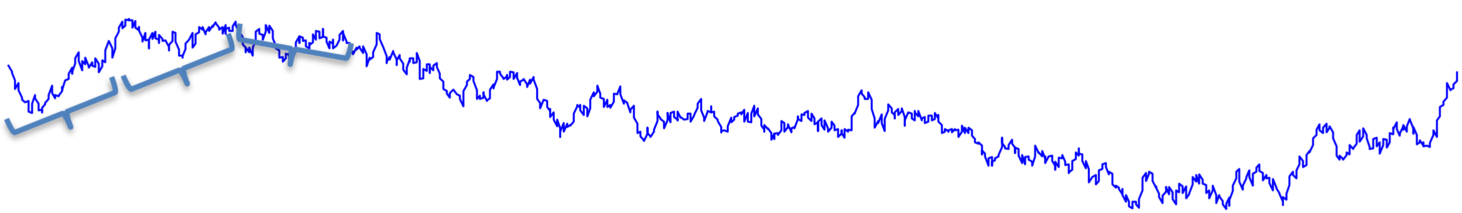
TS: rolling statistics

The time series may be huge. Analyzing it in its entirety may be difficult.

A solution is to employ the ‘rolling’ method, in which the TS is analyzed extracting a series of consecutive subsequences of fixed length. Each sub-series gives a different view on TS and its called window.

Given a window, each locality can now be described.

Examples are: rolling mean, rolling std etc.



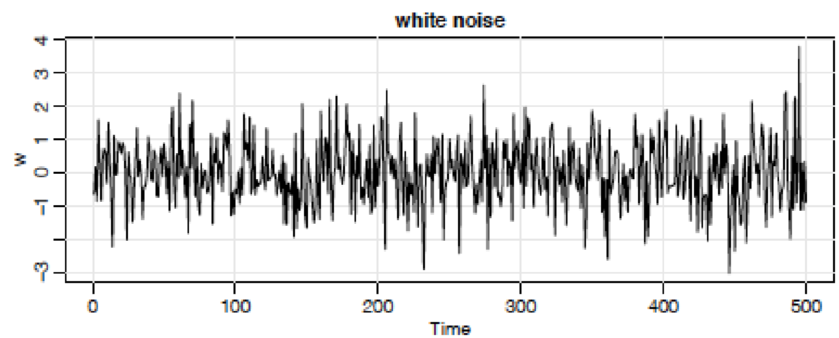
Moving average for noise removal

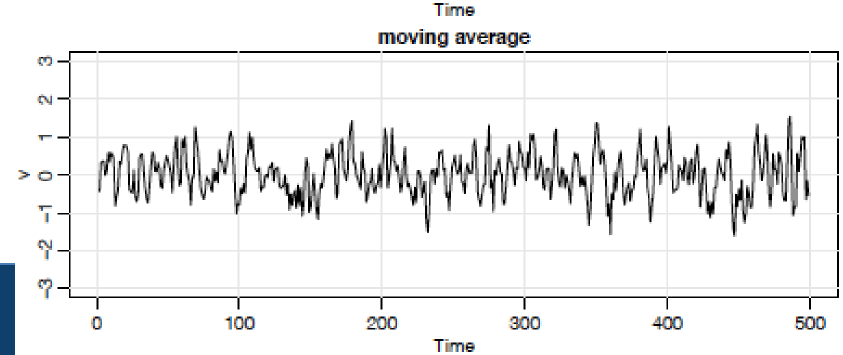
- Noise can be removed by a **moving average** (MA) that smooths the TS.
- Given a window of length *w* and a TS *t*, the MA is applied as follows
- $t_i = \frac{1}{w} \sum_{j=i-w/2}^{w/2} t_j$ for $i = 1, \ldots, n$

- For example, if w=3 we have
- $t_i = \frac{1}{3} \left(t_{i-1} + t_i + t_{i+1}\right)$

w=3

time	value	ma
t1	20	-
t2	24	22.0
t3	22	24.0
t4	26	24.3
t5	25	-

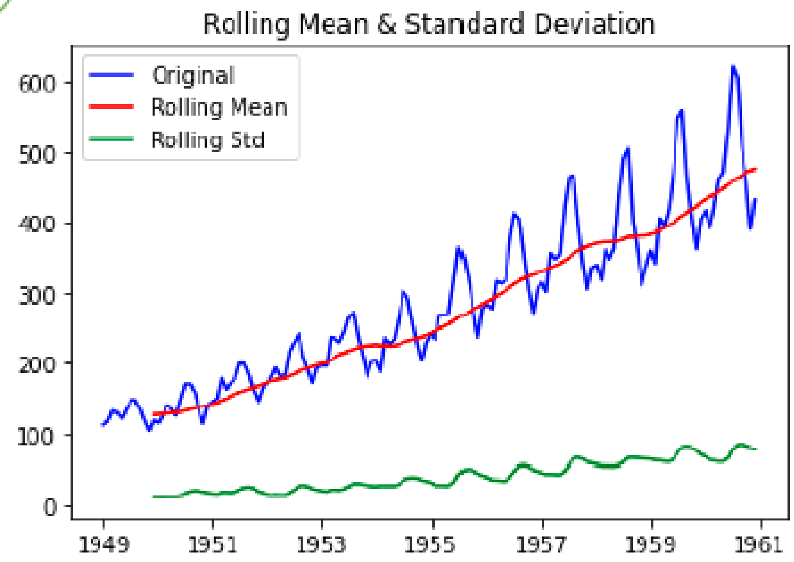




TS: rolling statistics

1

Rolling Statistics (Visual)



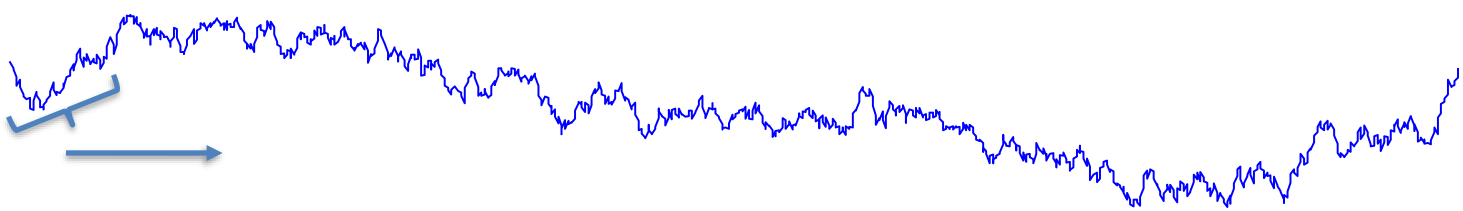
Plot the moving average or moving variance to check if it varies with time.

Notice the mean and variance **increase** constantly

TS: sliding statistics

Given a window, we can slide it through the entire TS and compute some kind of metric on the entire TS.

Similar to the convolution, we apply to the TS a mask sliding over the entire TS.



TS: sliding statistics – auto-covariance

How much does a component of a TS correlate with the previous and future components?

How to: compute the covariance between two components of the TS using the formula:

$$\rho_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

where:

- ρ_k: Autocorrelation at lag k
- Y_t: Value of the series at time t
- $\bar{Y}$: Mean of the series
- n: Number of observations

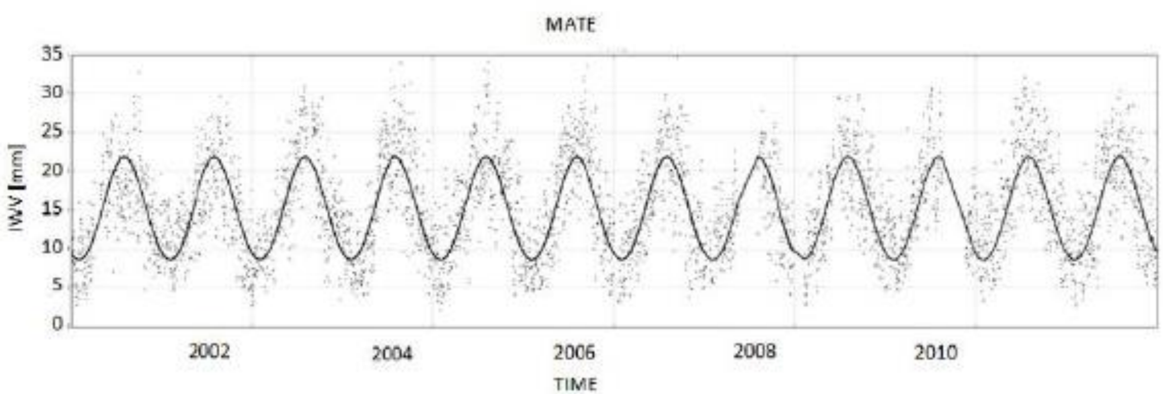
High auto-covariance may indicate seasonality

TS: why all of these statistics?

Before the application of any statistical model to TS, we need to analyze and pre-process them so that we can have stationary data.

Stationary: consistent means, variance and covariance over time. No trends, seasonality and so on. No predictable patterns.

Obtaining stationary TS may simplify tasks as classification or forecasting.

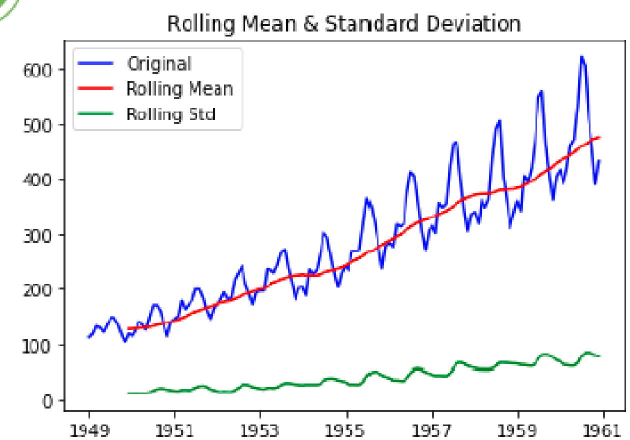


Stationary series

TS: how to know if they are stationary?

1. Apply rolling statistics and see if there are high covariance (means presence of trends or seasonality).
2. In the plot, we can see that the std is constant but mean is going up as trend. Hence it is not stationary

1 **Rolling Statistics (Visual)**



Plot the moving average or moving variance to check if it varies with time.

Notice the mean and variance **increase** constantly

TS: how to know if they are stationary?

We can conduct statistical tests, such as the Augmented Dickey-Fuller test: The test statistic looks for a unit root: if there is, it means that the TS is not stationary.

Dickey Fuller test (Statistical)

2

```
Test Statistic                    0.815369
p-value                           0.991880
#Lags Used                       13.000000
Number of Observations Used     130.000000
Critical Value (1%)              -3.481682
Critical Value (5%)              -2.884042
Critical Value (10%)             -2.578770
dtype: float64
```

Null Hypothesis = TS is non-stationary

If 'Test Statistic' < 'Critical Value',
Reject the null hypothesis

TS: how to make them stationary?

We can conduct statistical tests, such as the Augmented Dickey-Fuller test: The test statistic looks for a unit root: if there is, it means that the TS is not stationary.

Dickey Fuller test (Statistical)

2

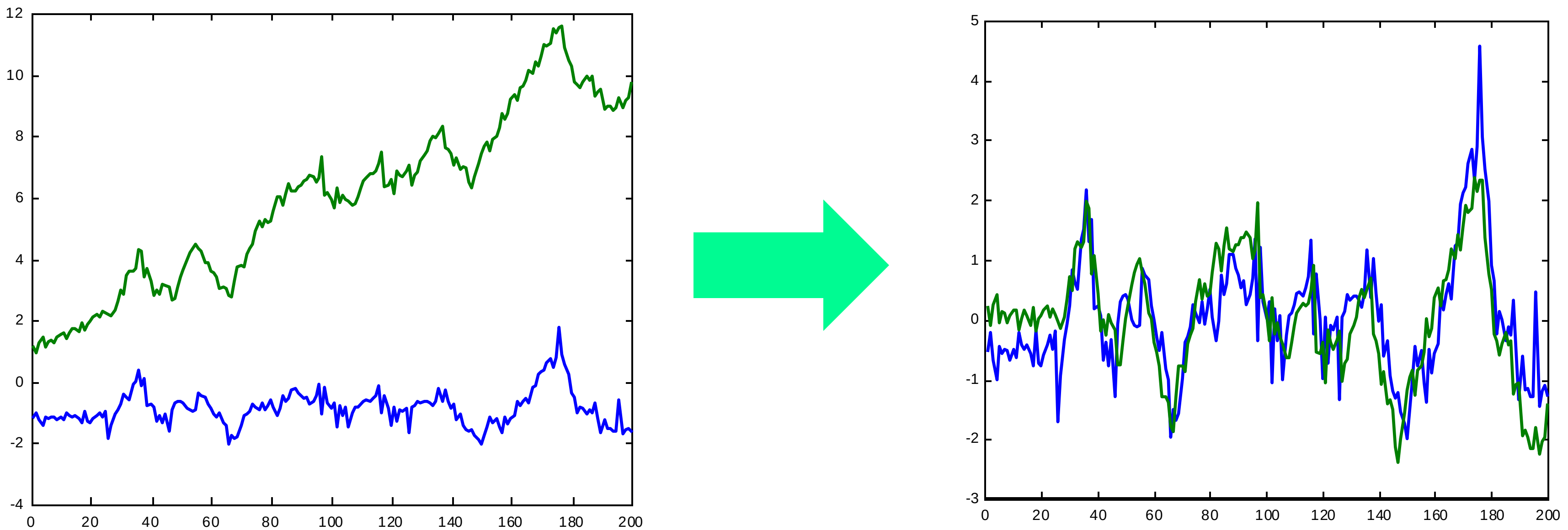
```
Test Statistic                   0.815369
p-value                          0.991880
#Lags Used                      13.000000
Number of Observations Used    130.000000
Critical Value (1%)             -3.481682
Critical Value (5%)             -2.884042
Critical Value (10%)            -2.578770
dtype: float64
```

Null Hypothesis = TS is non-stationary

If 'Test Statistic' < 'Critical Value',
Reject the null hypothesis

Linear trend

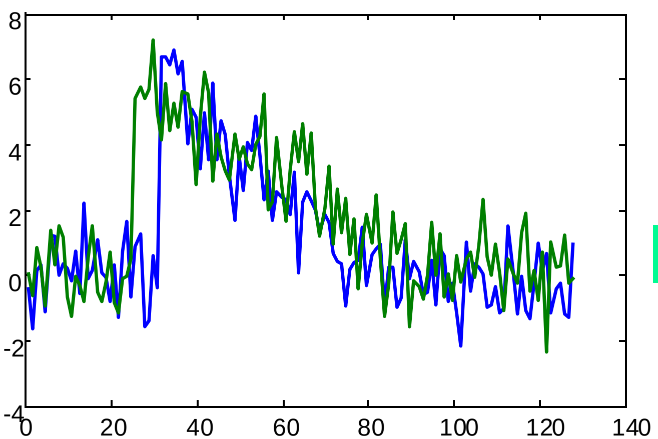
Removing linear trend: fit the best fitting straight line to the time series, then subtract that line from the time series.

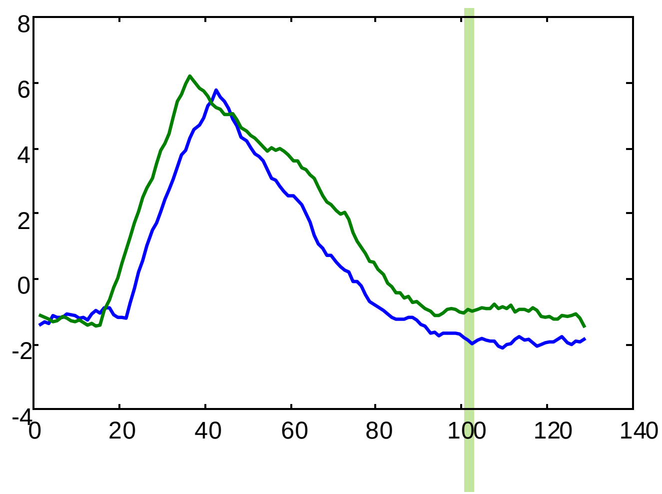


Removed linear trend,
offset translation,
amplitude scaling

Noise removal

The intuition behind removing noise is to average each datapoints value with its neighbors.





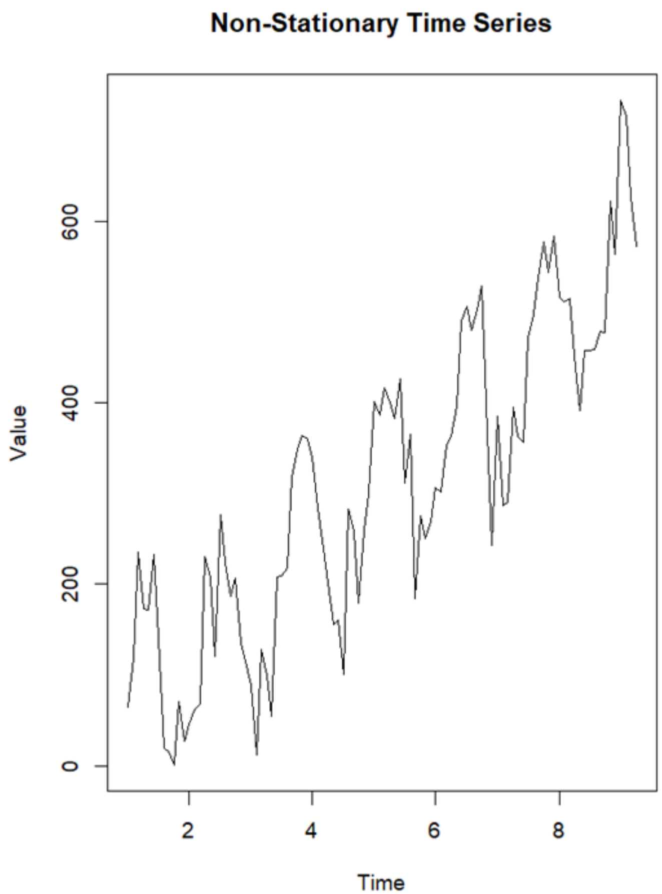
$Q = \text{smooth}(Q)$

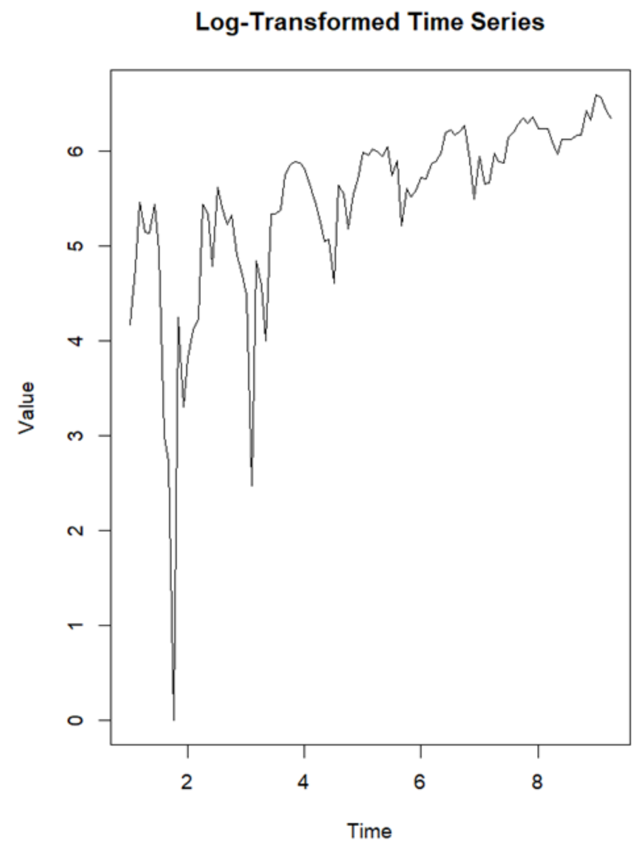
$C = \text{smooth}(C)$

$D(Q,C)$

Log transformation

You can apply the natural logarithm or the base 10 logarithm for stabilizing the variance, for linearizing trends, improve normal distribution.





Log transformation: pros/cons

1. Data must be positive
2. A little bit more difficult to interpret since the space of the data is changed, hence the info and pattern may be more difficult to comprehend
3. Masking not always easy (how to handle zero?)