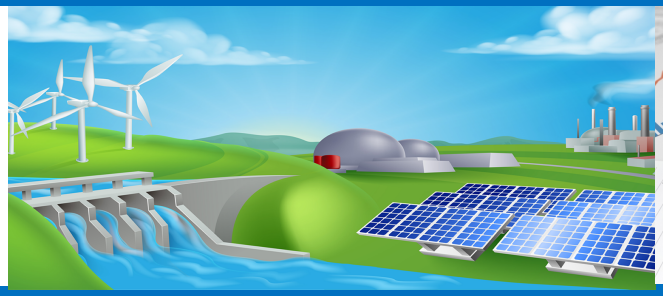



ENV872 - ENVIRONMENTAL DATA ANALYTICS

M7 – Time Series Analysis (TSA)

Luana Lima | John Fay

Spring 2021

Master of Environmental Management Program
Nicholas School of the Environment - Duke University

Learning Goals

- Introduction to Time Series Analysis (TSA)
  - What is TSA?
  - Examples
  - TSA Components (trend, cycle, seasonal, random)
- Autocorrelation Function (ACF)
- Partial Autocorrelation Function (PACF)
- Trend and Seasonal Component
- Stationarity Tests

Introduction to Time Series Analysis

Meaning and definitions

Importance of TSA

Components of TSA

What is a Time Series?

- A set of observations on a variable collected over time
- Discrete and continuous time series
- Example: stock prices, interest rate, retail sales, electric power consumption, etc
- Mathematically representation: a time series is defined by the values $Y_1, Y_2, \ldots$ of a variable Y at times $t_1, t_2, \ldots$

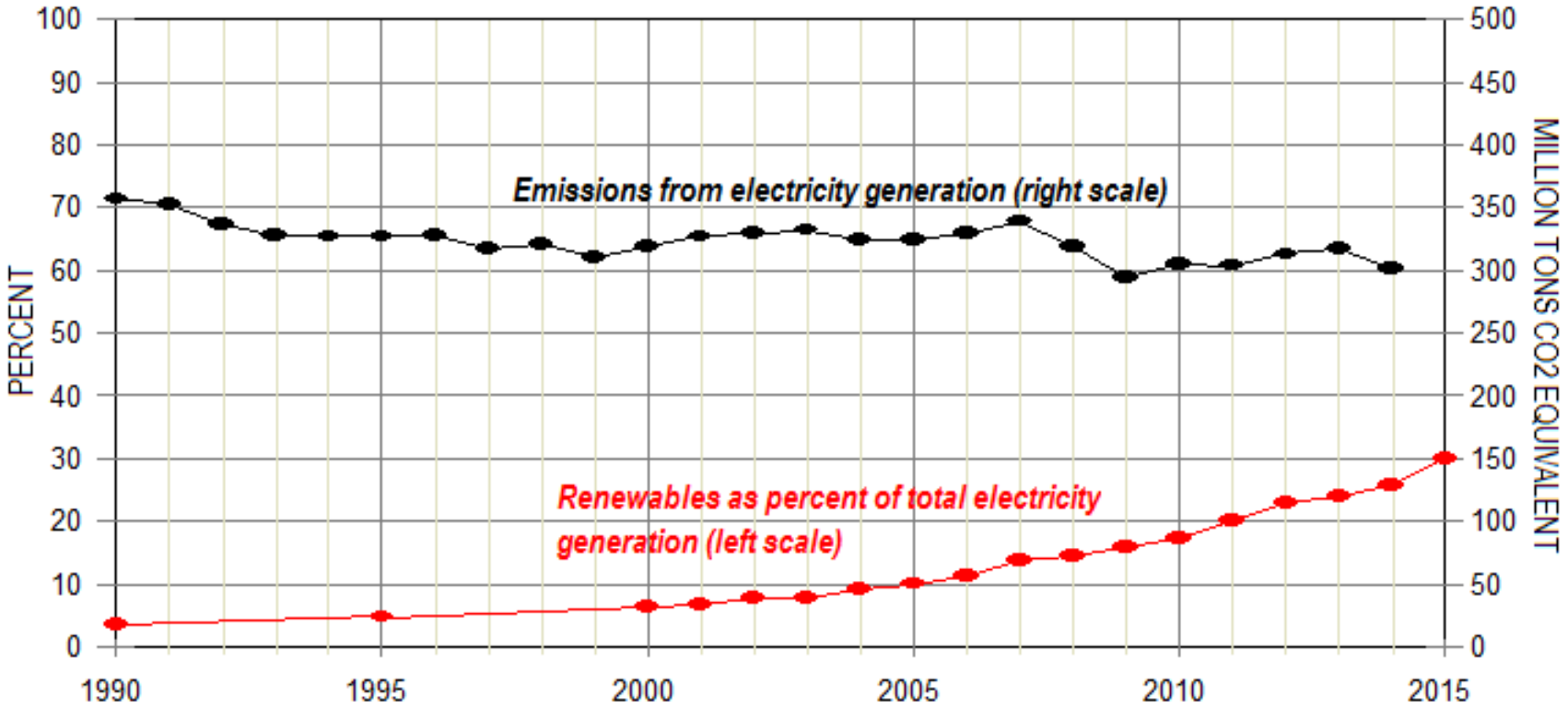
  Thus,

$$Y = F(t)$$

What is Time Series Analysis (TSA)?

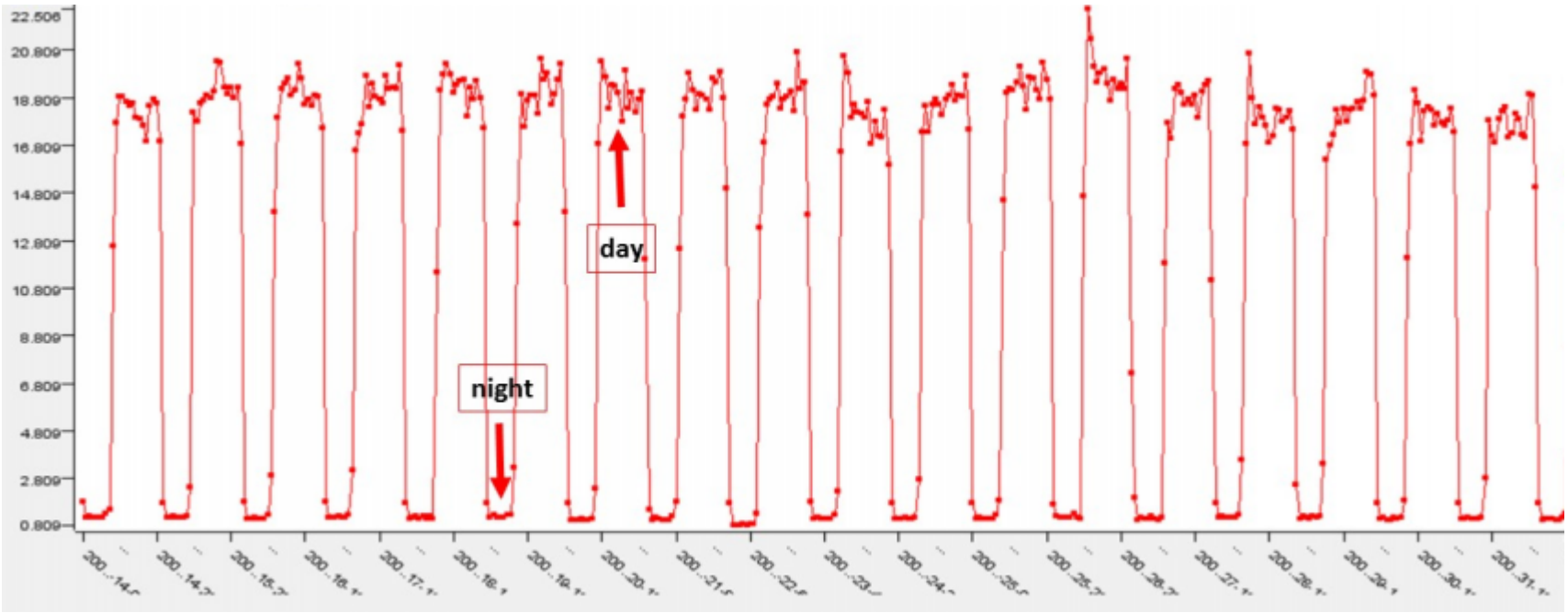
- In TSA, we analyze the past behavior of a variable in order to predict its future behavior
- Causes of variation of Time Series Data
  - Seasons, holidays, etc
  - Natural calamities: earthquake, epidemic, flood, drought, etc
  - Political movements or changes, war, etc

Example of Time Series Data



Percent renewables in Germany's electricity mix versus total greenhouse gas emissions, 1990-2015

Example of Time Series Data



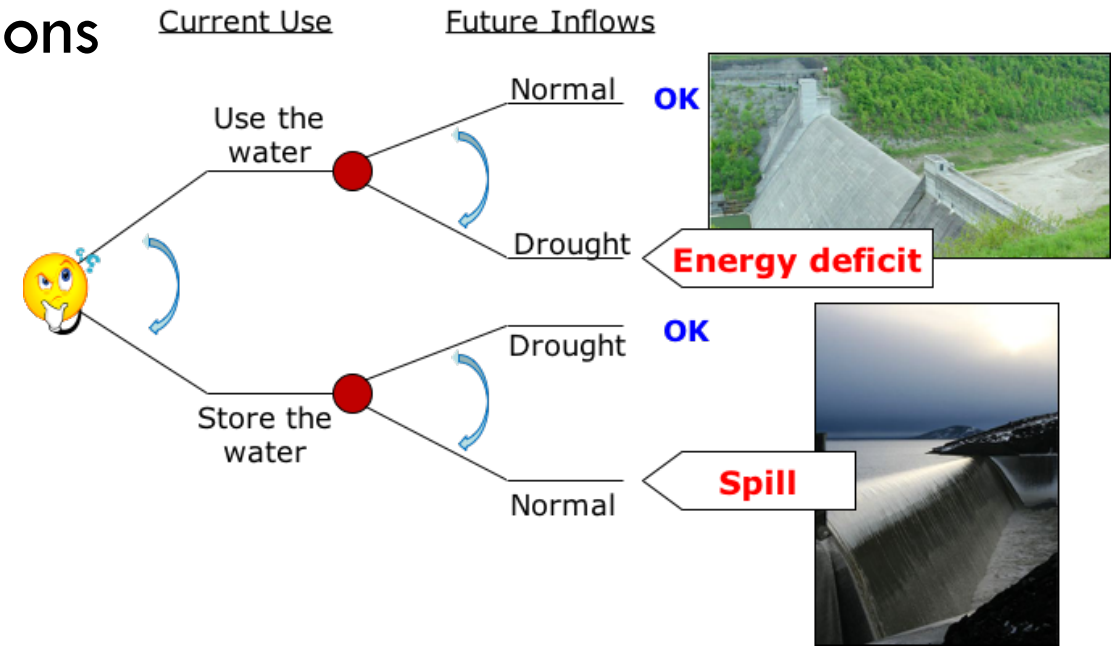
Energy consumption time-series for meter ID 1038. Notice the day/night rhythm.

Source: https://dzone.com/articles/data-chef-etl-battles-energy-consumption-time-seri

Importance of TSA

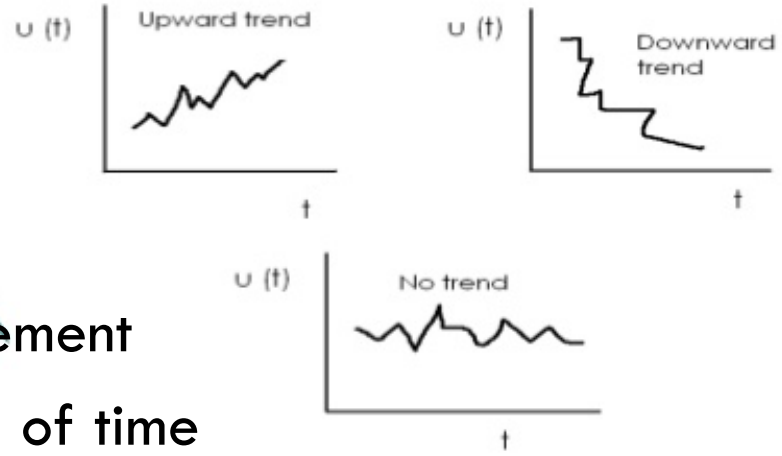
- Very popular tool for business forecasting
- Basis for understanding past behavior

- Can forecast future activities/planning for future operations



Components of TSA

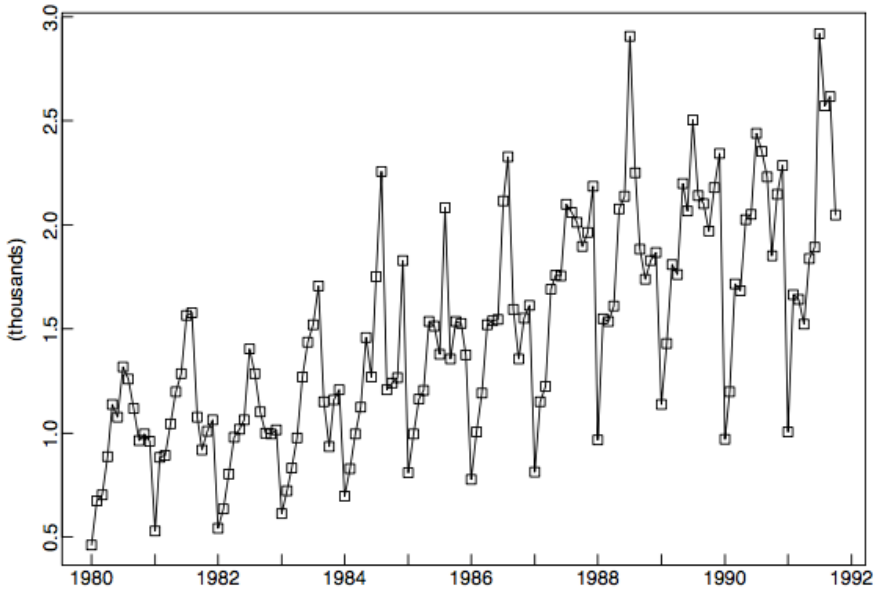
- Time frame: short, medium and long-term
  - How far can we predict?
- Trend
  - General tendency to grow or decline over a long period
  - Easiest to detect
  - Maybe linear or non-linear
- Cycle
  - An up and down repetitive movement
  - Repeat itself over a long period of time
  - Example: business cycle (prosperity, decline, depressions, recovery)



Components of TSA (cont'd)

- Seasonal Variation
  - An up and down repetitive movement occurring periodically (short duration)
  - Factor that cause seasonal variations: climate and weather condition or custom traditions and habits

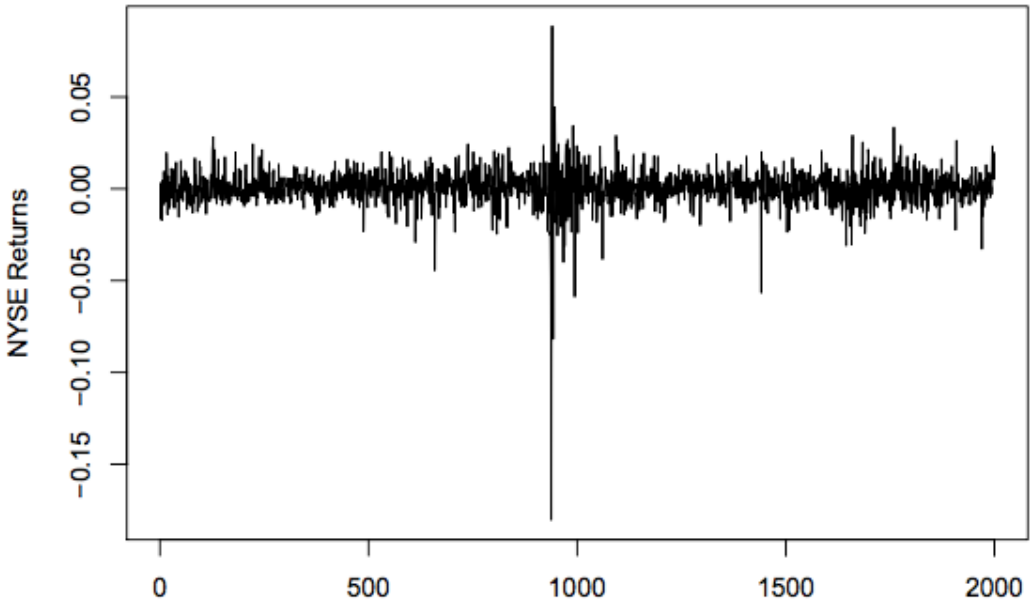
Australian Red Wine Sales



Source: Brockwell and Davis, Introduction to Time Series and Forecasting

Components of TSA (cont'd)

- Random Variations
  - Erratic movements that are not predictable because they don't follow a pattern
  - Example: strike, fire, war, flood, earthquake, etc..



Source: Brockwell and Davis, Introduction to Time Series and Forecasting

TSA Terms

- Stationary Data - a time series variable exhibiting no significant upward or downward trend over time
- Nonstationary Data - a time series variable exhibiting a significant upward or downward trend over time
- Seasonal Data - a time series variable exhibiting a repeating patterns at regular intervals over time

Autocorrelation Function

Meaning of Autocorrelation Function

- Recap: What is correlation?

*From stats: covariance and correlation measure **joint variability** of two variables.*

Meaning of Autocorrelation Function

- Recap: What is correlation?

  Is a measure of linear dependence between two variables

- In TSA: What is autocorrelation?

  Is a measure of dependence between two adjacent values of the same variables

- The prefix *auto* is to convey the notion of self-correlation, that is, correlation between variables from the same time series

How to compute autocorrelation?

- In the context of a single variable, Y_t is the original series and Y_s is a lagged version of the series

Y_t	Y_s
Y_1	Y_2
Y_2	Y_3
Y_3	Y_4
Y_4	Y_5
⋮	⋮
Y_{N-3}	Y_{N-2}
Y_{N-2}	Y_{N-1}
Y_{N-1}	Y_N
Y_N	

Compute lag 1 autocorrelation

$$\rho_{t,s} = Corr(Y_t, Y_s)$$

How to compute autocorrelation?

- In the context of a single variable, Y_t is the original series and Y_s is a lagged version of the series

Y_t	Y_s
Y_1	Y_3
Y_2	Y_4
Y_3	Y_5
Y_4	Y_6
$\vdots$	$\vdots$
Y_{N-3}	Y_{N-1}
Y_{N-2}	Y_N
Y_{N-1}	
Y_N	

Compute lag 2 autocorrelation

$$\rho_{t,s} = Corr(Y_t, Y_s)$$

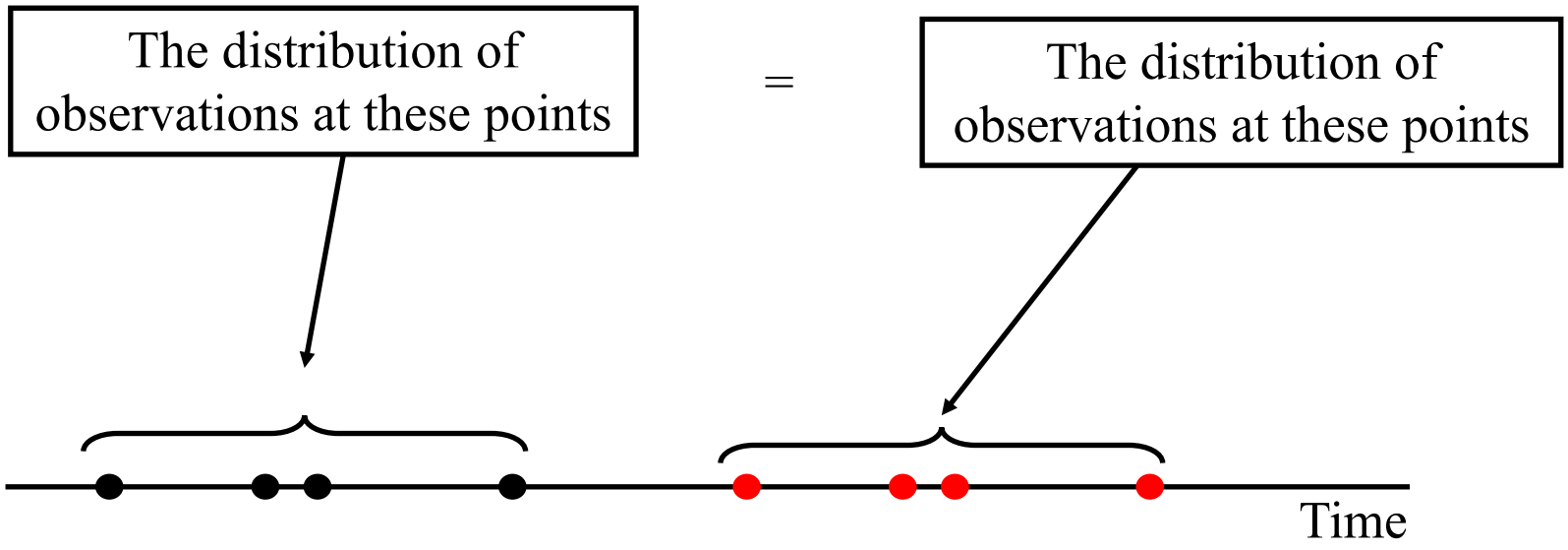
Main Conclusion

Autocovariance and autocorrelation function give information about the dependence structure of a time series

Stationary Process

Stationary Process

- The basic idea of stationarity is that the probability laws that govern the behavior of the process do not change over time



Consequences of Stationarity

- Distribution of Y_t is the same of Y_{t-k} for all t and k
- Then,
  - $E(Y_t) = E(Y_{t-k})$ for all t and k so the **mean function is constant** for all time
  - $Var(Y_t) = Var(Y_{t-k})$ for all t and k so the **variance is also constant** over time

- And what happens with the autocovariance function?

**adapted from Ch. 2 of Cryer and Shan*

Consequences of Stationarity (cont'd)

- If the process is stationary, then

$$\gamma_{t,s} = Cov(Y_t, Y_s) = Cov(Y_{t-k}, Y_{s-k})$$

For $k = s \quad \rightarrow \quad Cov(Y_t, Y_s) = Cov(Y_{t-s}, Y_0)$

For $k = t \quad \rightarrow \quad Cov(Y_t, Y_s) = Cov(Y_0, Y_{s-t})$

Thus, $\gamma_{t,s} = Cov(Y_0, Y_{|t-s|}) = \gamma_{0,|t-s|}$

- In other words, the covariance between Y_t and Y_s depends only on the time difference $|t - s|$ and not on the actual times *t* and *s*

**adapted from Ch. 2 of Cryer and Shan*

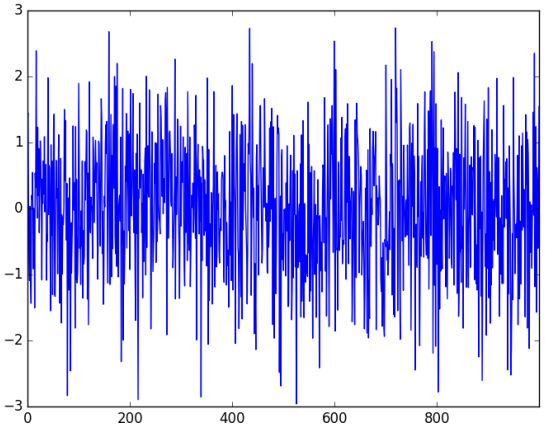
White Noise Series

- Example of a stationary process: **white noise** series
- The white noise series is a sequence of independent, identically distributed (i.i.d.) random variables $\{e_t\}$
- $\{e_t\}$ is a stationary process, then

$$\mu_t = \mathrm{E}(\mathrm{e}_t)$$

$$\gamma_k = \begin{cases} Var(e_t) & for\ k = 0 \\ 0 & for\ k \neq 0 \end{cases}$$

$$\rho_k = \begin{cases} 1 & for\ k = 0 \\ 0 & for\ k \neq 0 \end{cases}$$



- In time series modeling we usually assume that the white noise process has mean zero and $Var(e_t) = \sigma_e^2$

**adapted from Ch. 2 of Cryer and Shan*

Partial Autocorrelation Function

Partial Autocorrelation Function

Recap: The ACF of a stationary process Y_t at lag h

$$\rho_{t,t-h} = Corr(Y_t, Y_{t-h})$$

measures the linear dependency among the process variables Y_t and Y_{t-h}.

But the dependency structure among the **intermediate variables**

$$Y_t, \boxed{Y_{t-1}, Y_{t-2}, \cdots Y_{t-h+2}, Y_{t-h+1}}, Y_{t-h}$$

also plays an important role on the value of the ACF.

Partial Autocorrelation Function (cont'd)

Imagine if you could **remove** the influence of all these intermediate variables…

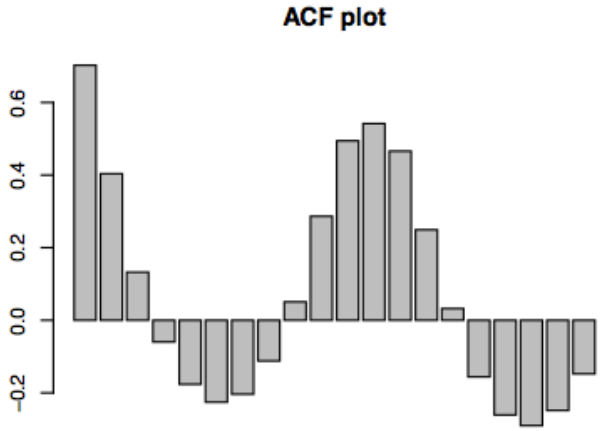
You would have only the directly correlation between Y_t and Y_{t-h}

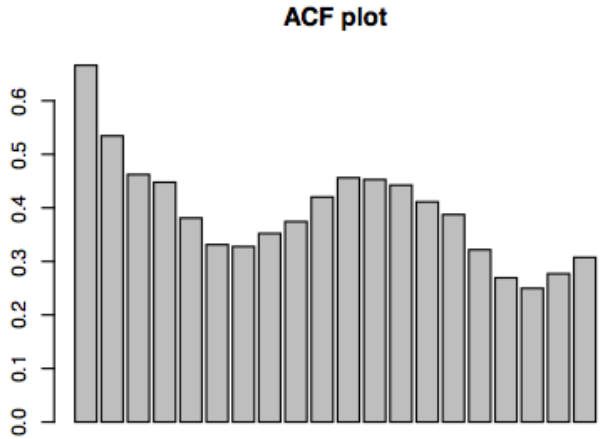
That's the so called **partial autocorrelation function (PACF)**

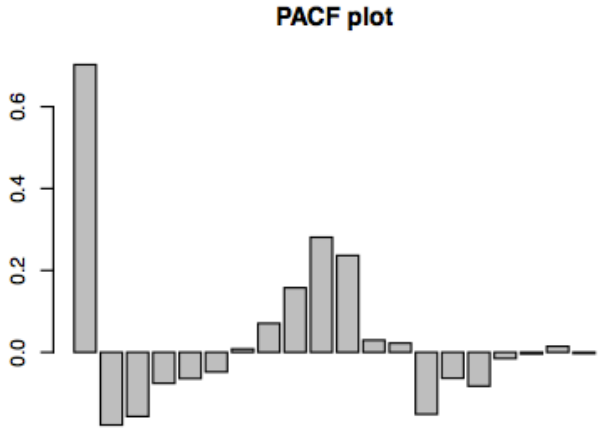
Partial Autocorrelation Function (cont'd)

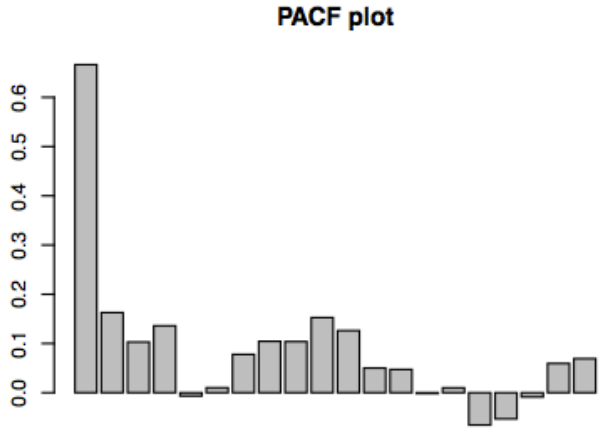
- The PACF is a little more difficult to compute
- We will talk about that later when we discuss the Yule Walker equations
- In summary:
  - The ACF and PACF measure the temporal dependency of a stochastic process
  - You will always build the ACF and PACF before fitting a model to a stochastic process
  - The ACF and PACF give us information about the **auto-regressive component** of the series

Examples of ACF and PACF plots







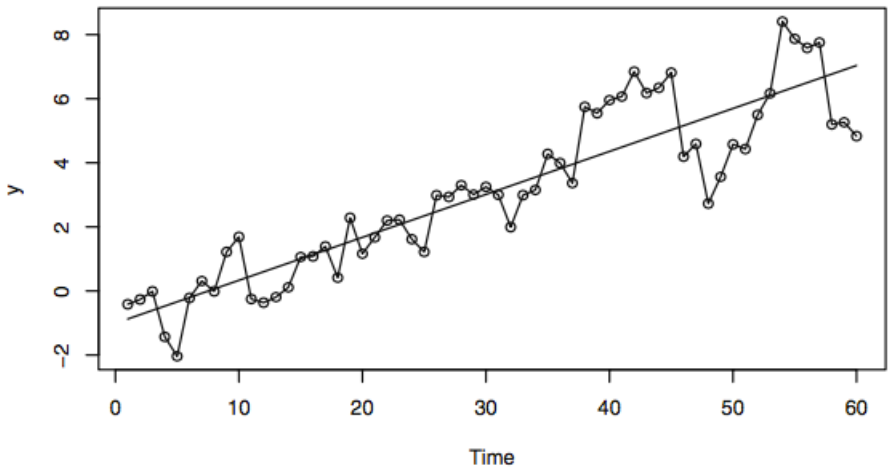


Trend Component

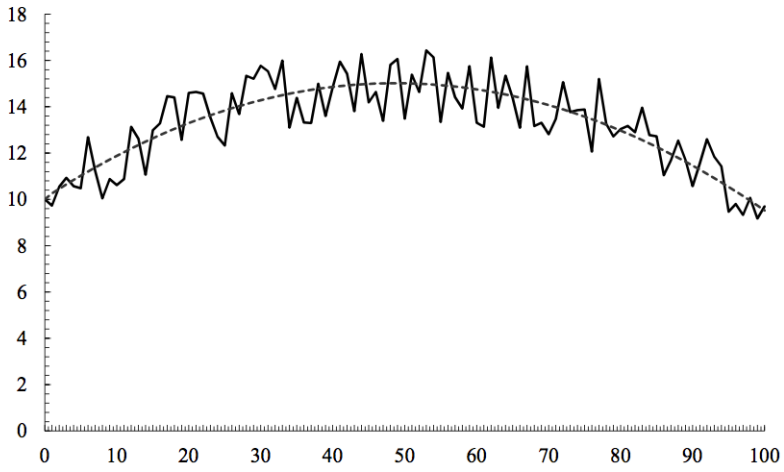
Trend Component

- Long-term tendency
  - Increase (upward movement) or
  - Decrease (downward movement)
- Trend can be linear or non-linear

Ex: Upward Linear Trend



Ex: Quadratic Trend

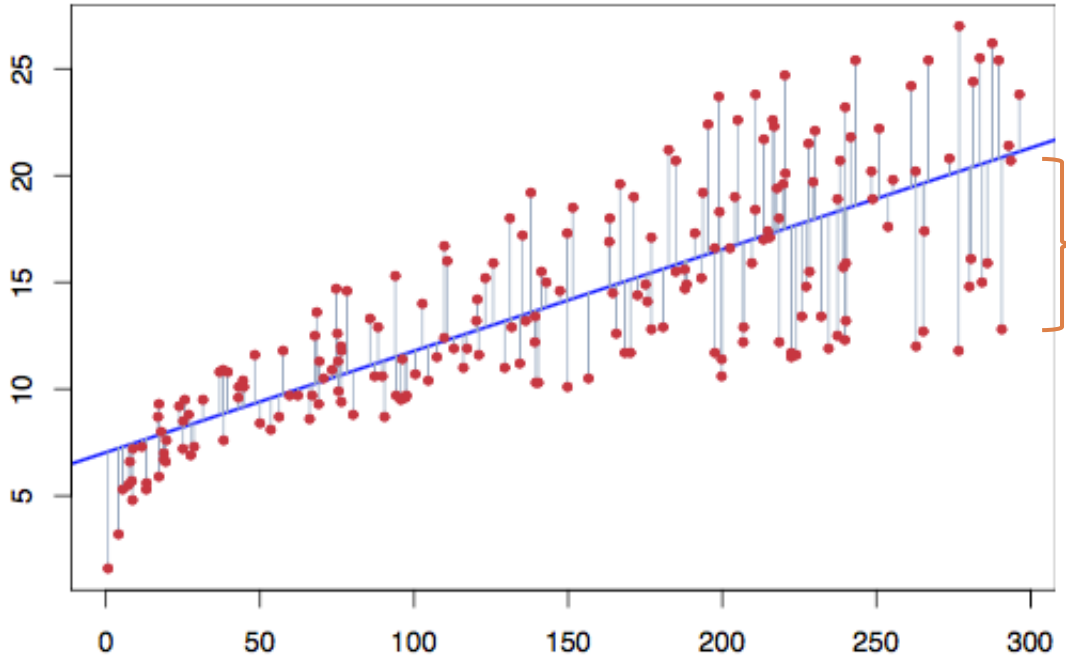


Linear Trend Component

- For a linear trend we can write

$$Y_i = \beta_0 + \beta_1 t_i + \varepsilon_i$$

- **Slope** (β_1) and the **intercept** (β_0) are the unknown parameters, and ε_i is the **error term**



$\widehat{Y}_i = \beta_0 + \beta_1 t_i$

The error term or residual is the distance from point Y_i to the estimate $\widehat{Y}_i$

$$\varepsilon_i = Y_i - \widehat{Y}_i$$

Linear Trend Estimation and Removal

1. Model the trend: find β_0 and β_1

2. For each observation t remove trend

$$Y_{detrend_t} = Y_t - (\beta_0 + \beta_1 t)$$

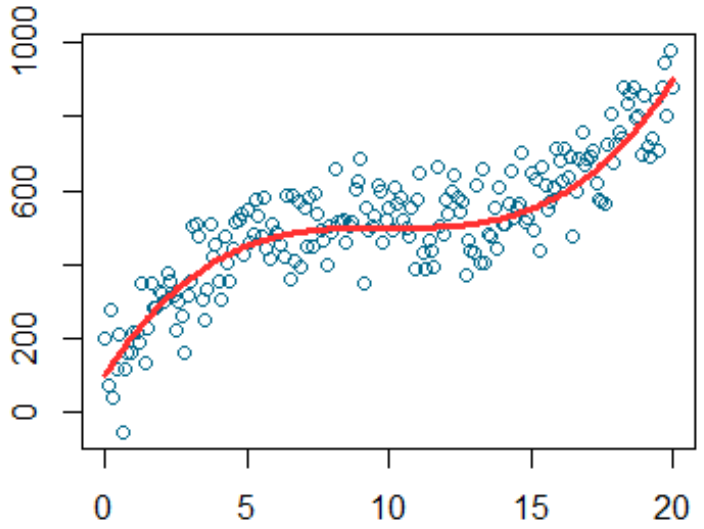
Non-linear Trend

Polynomial trend

- Example: quadratic trend

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 T_i^2 + \varepsilon_i$$

- Or any other order

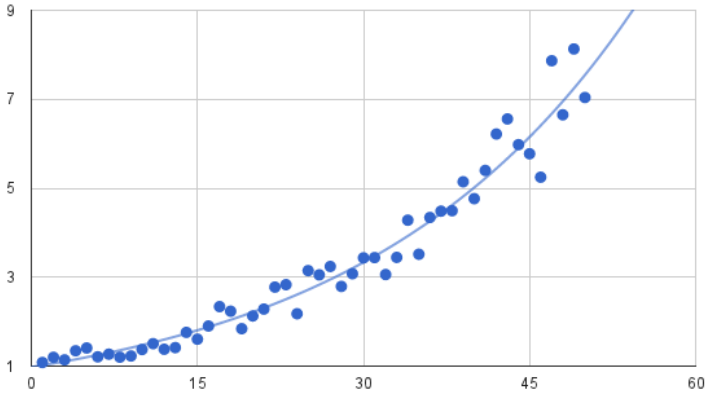


Exponential trend

$$Y_i = (e^{\beta_0 + \beta_1 T_i})\varepsilon_i$$

- Can be transformed into linear trend

$$\ln Y_i = \beta_0 + \beta_1 T_i + \ln \varepsilon_i$$

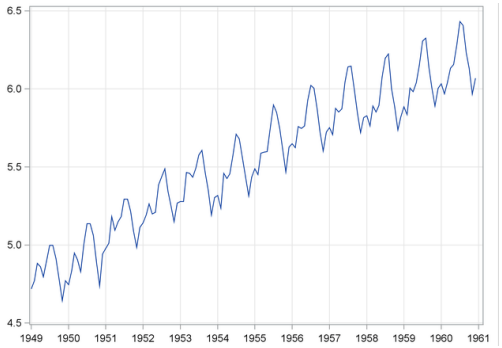


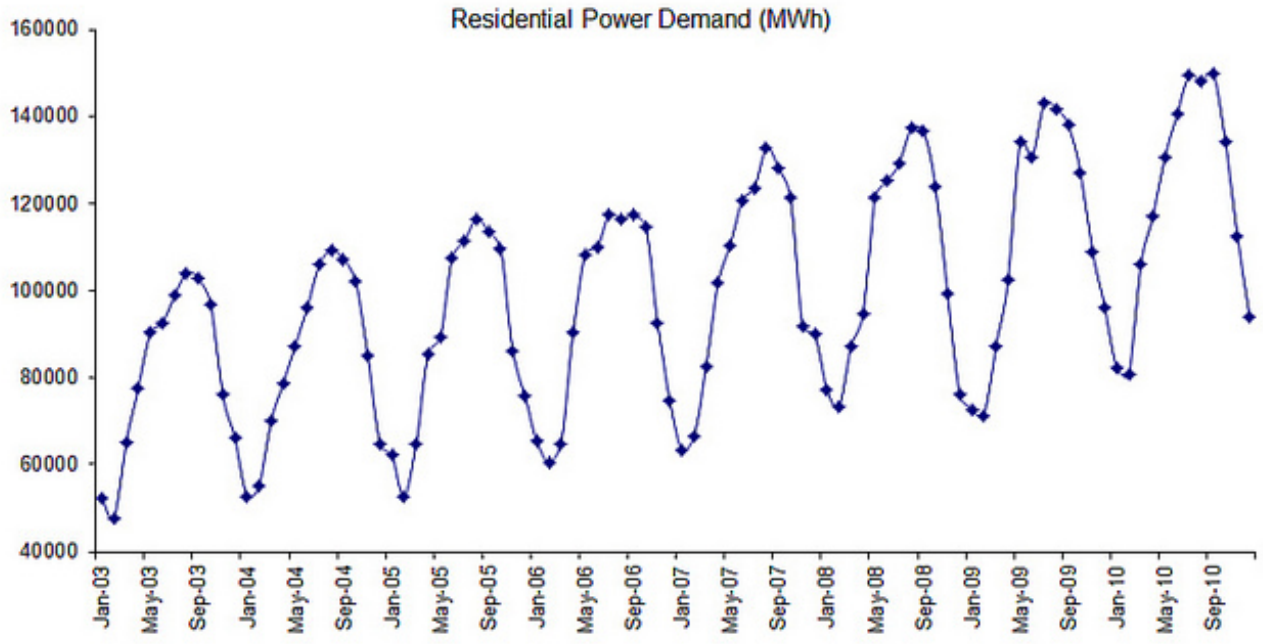
Most of the time we assume a linear trend to simplify the analysis

Seasonal Component

Seasonal Component

- Short-term regular wave-like patterns
- Observed within 1 year
- Often monthly or quarterly





Seasonal Trend Estimation

- How do we estimate seasonal trend?
- Assume the observed series can be represented as

$$Y_t = \mu_t + X_t$$

where $E[X_t] = 0$

- For monthly seasonal data assume 12 parameters such as

$$\mu_t = \begin{cases} \beta_1 & for\ t = 1,13,25,\cdots \\ \beta_2 & for\ t = 2,14,26,\cdots \\ \vdots & \\ \beta_{12} & for\ t = 12,24,36,\cdots \end{cases}$$

Seasonal Means Model

- The number of seasons may be less than 12.

Seasonal Trend Removal

1. Model the seasonal trend

$$Y_t = \sum_{s=1}^{12} \mu_t D_{t,s} \text{ where } \mu_t = \begin{cases} \beta_1 & for\ t = 1,13,25,\cdots \\ \beta_2 & for\ t = 2,14,26,\cdots \\ & \vdots \\ \beta_{12} & for\ t = 12,24,36,\cdots \end{cases}$$

1. For each observation t remove seasonal trend

$$Y_{deseason_t} = Y_t - (\sum_{s=1}^{12} \beta_s D_{t,s})$$

Stochastic versus deterministic trend

Series with Deterministic Trend

- Deterministic linear trend process

$$Y_i = \beta_0 + \beta_1 t_i + \varepsilon_i$$

- Or more generally, for a polynomial trend

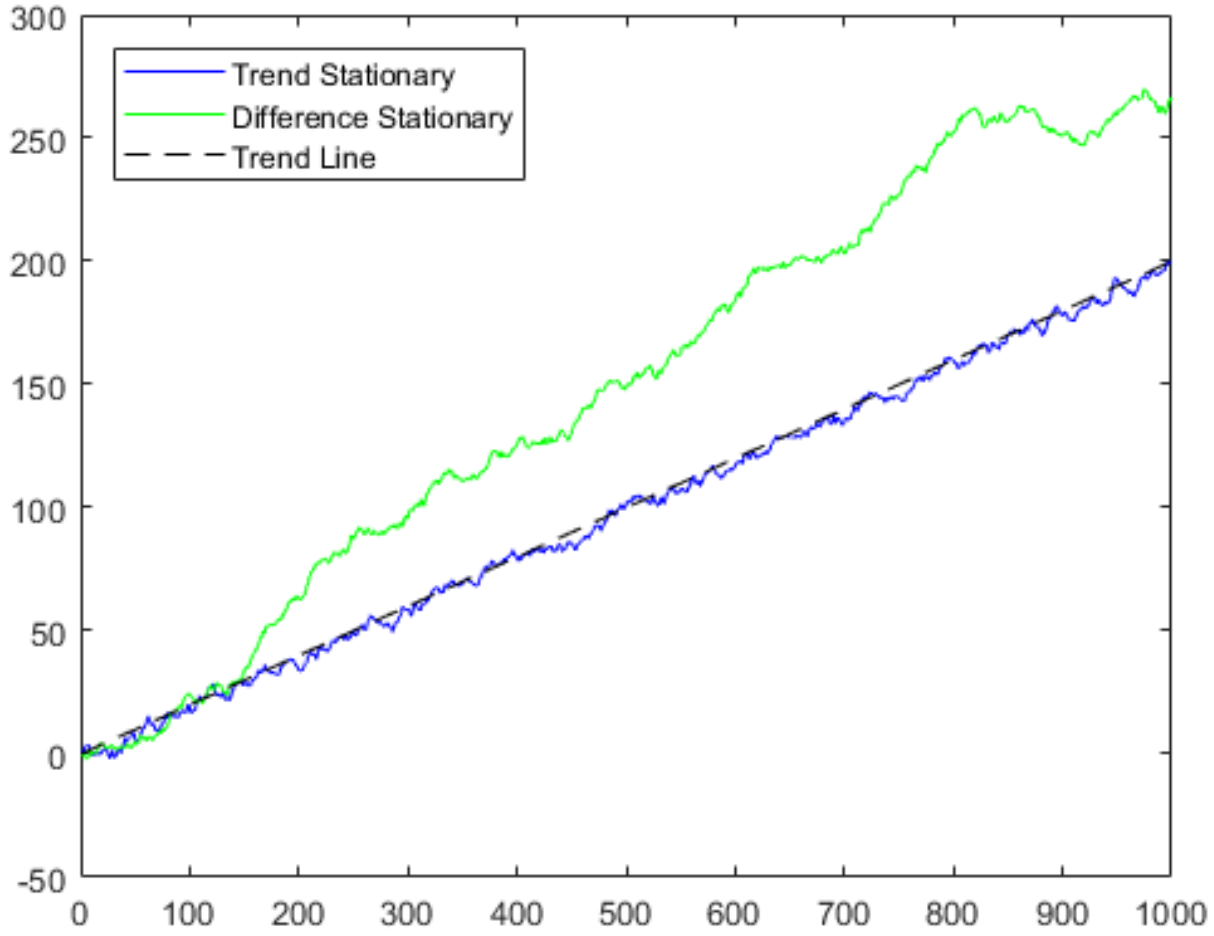
$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 T_i^2 + \cdots + \beta_n T_i^n + \varepsilon_i$$

- Detrending is accomplished by running a regression and obtaining the series of residuals. The residuals will give you the detrended series
- That's what we call **trend-stationarity**

Series with Stochastic Trend

- But some series have what we call **difference-stationarity**
- Although trend-stationary and difference-stationary series are both "trending" over time, the stationarity is achieved by a distinct procedure
- In the case of difference-stationarity, stationarity is achieved by differencing the series
- Sometimes we need to difference the series more than once

Trend-stationarity vs difference-stationarity



Stationarity Tests

Stationarity Assessment

- **Mann-Kendall Test– monotonic trend**
- **Spearman's Rank Correlation Test – monotonic trend**

- **Dickey-Fuller (ADF) Test – unit root**
- Phillips-Perron (PP) Test – unit root
- Kitawoski-Phillips-Schmidt-Shin (KPSS) – unit root

- And others…

Mann-Kendall Test

- Commonly employed to detect deterministic trends in series of environmental data, climate data or hydrological data
- **Cannot be applied to seasonal data**
- Hypothesis Test

$$\begin{cases} H_0: & Y_t \quad is \quad i.i.d.(stationary) \\ H_1: & Y_t \quad follow\ a\ trend \end{cases}$$

Mann-Kendall Test

- Mann-Kendall statistic is

$$S = \sum_{k=1}^{N-1} \sum_{j=k+1}^{N} sgn(Y_j - Y_k)$$

where

$$sgn(Y_j - Y_k) = \begin{cases} 1 & if \quad Y_j - Y_k > 0 \\ 0 & if \quad Y_j - Y_k = 0 \\ -1 & if \quad Y_j - Y_k < 0 \end{cases}$$

- The test will check the magnitude of S and its significance based on the number of observations
- In other words, the bigger the number of observations the higher S will need to be

Mann-Kendall test in R

- The Mann-Kendall test in R is done with the command MannKendall() from package "Kendall"

Description

This is a test for monotonic trend in a time series z[t] based on the Kendall rank correlation of z[t] and t.

Usage

```
MannKendall(x)
```

Arguments

x	a vector of data, often a time series

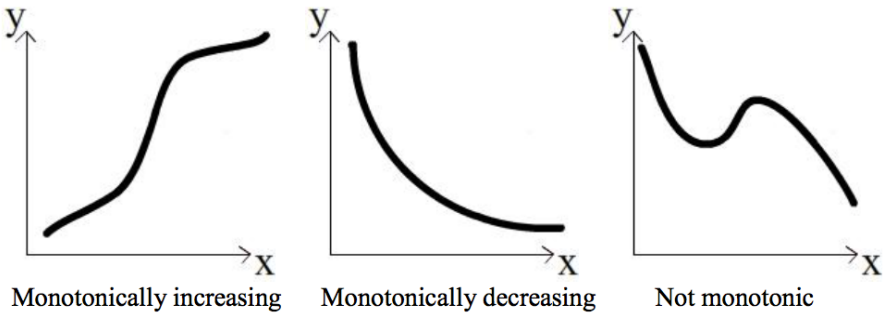
Details

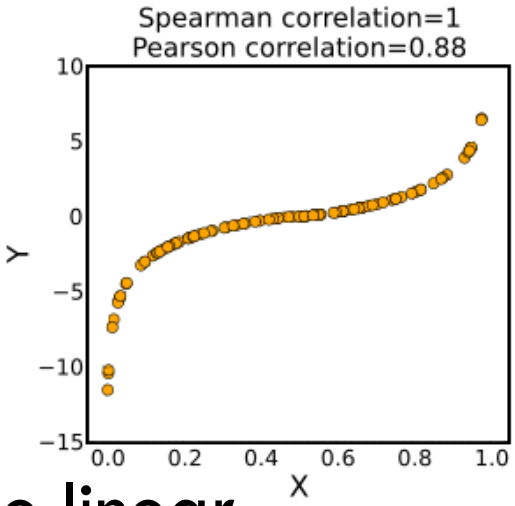
The test was suggested by Mann (1945) and has been extensively used with environmental time series (Hipel and McLeod, 2005). For autocorrelated time series, the block bootstrap may be used to obtain an improved signficance test.

- For seasonal data you can use SeasonalMannKendall() from the same package

Spearman's Rank Correlation Coefficient

- Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship





- Unlike Pearson's correlation, the relationship does not need to be linear
- In other words, if one variable increases so do does the other, it does not matter the proportion of the increase

Spearman's Rank Correlation Coefficient

- To verify a monotonic trend in your data, compute the spearman correlation between your data and series T

Y_t	T
Y_1	1
Y_2	2
Y_3	3
$\vdots$	$\vdots$
Y_{N-2}	$N-2$
Y_{N-1}	$N-1$
Y_N	N

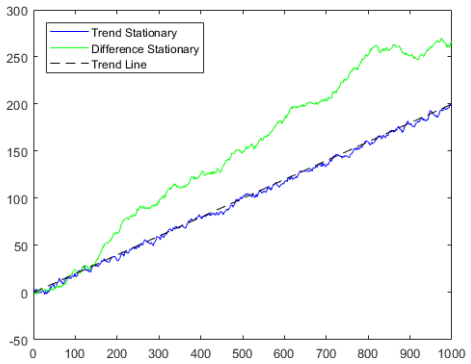
- If the correlation is close to 0, then there is no trend
- The function to compute spearman correlation is cor() or the cor.test() from package "stats". The latter provides the significance of the coefficient

Dick-Fuller Test

- The first work on testing for a unit root in time series was done by Dickey and Fuller
- Consider the model

$$Y_t = a + \phi Y_{t-1} + \epsilon_t$$

White noise series



- The objective is to test

$$\begin{cases} H_0: & \phi = 1 \ \ (i.e. \ \ contain\ a\ unit\ root) \\ H_1: & \phi < 1 \ \ (i.e. \ \ is\ stationary) \end{cases}$$

- More general case can include more lags, the so called Augmented Dickey-Fuller (ADF) test

Dick-Fuller Test in R

- The ADF test in R is done with the command adf.test() from package "tseries"

Description

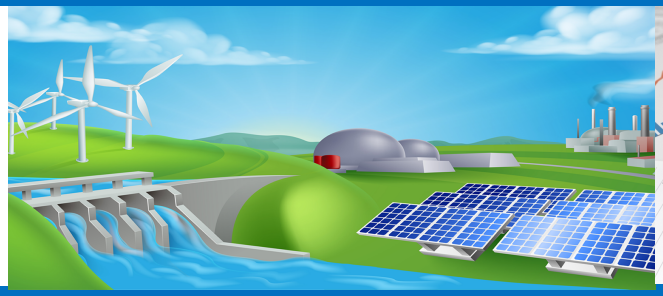
Computes the Augmented Dickey-Fuller test for the null that x has a unit root.

Usage

```
adf.test(x, alternative = c("stationary", "explosive"),
         k = trunc((length(x)-1)^(1/3)))
```

Arguments

`x`	a numeric vector or time series.
`alternative`	indicates the alternative hypothesis and must be one of `"stationary"` (default) or `"explosive"`. You can specify just the initial letter.
`k`	the lag order to calculate the test statistic.




THANK YOU !

luana.marangon.lima@duke.edu

Spring 2021

Master of Environmental Management Program
Nicholas School of the Environment - Duke University