NDA: Time Series Analysis - part 1

Lionel Tabourier

LIP6 – CNRS and Sobonne University

first_name.last_name@lip6.fr

January 20th 2021

Bibliography

Formal content:

- Peter Brockwell and Richard Davis
  Introduction to Time Series and Forecasting
- William Thistleton and Tural Sadigov
  MOOC Coursera: *Practical Time Series Analysis*

Informal guide in python:

- `www.machinelearningplus.com/time-series/`

Illustrative datasets:

- `data.world/datasets/time-series`
- `www.kaggle.com/tags/time-series`

Outline

Outline

1. Problem definition
2. Some elementary concepts
3. Some elementary models
4. Decomposing the time series

What is time series analysis?

Definition

Set of observations $\{x_t\}$, recorded at time $t \in T_0$

Think of each x_t as a realization from a distribution

Specificities of the problem

A unique realization of the process

$\Rightarrow$ necessary to make assumptions

- observe time series, identify particularities
- choose a family of models X_t to represent data
- check the goodness of the model

What is time series analysis?

Definition

Set of observations $\{x_t\}$, recorded at time $t \in T_0$

Think of each x_t as a realization from a distribution

Specificities of the problem

A unique realization of the process
$\Rightarrow$ necessary to make assumptions

- observe time series, identify particularities
- choose a family of models X_t to represent data
- check the goodness of the model

Assumptions for this course

Restrictions to a subfamily of problems

- discrete time series (discrete time set)
- fixed time steps (time resolution)
- univariate (one single variable over time) $\neq$ multivariate
  $\rightarrow$ processes have values in $\mathbb{R}$

And only a few approaches (Box-Jenkins)

- *e.g.* no Fourier analysis

Assumptions for this course

Restrictions to a subfamily of problems

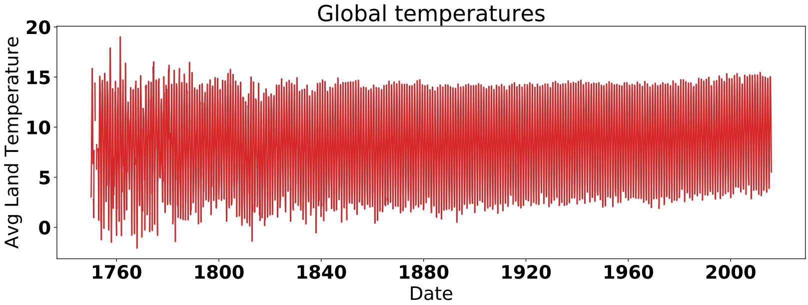
- discrete time series (discrete time set)
- fixed time steps (time resolution)
- univariate (one single variable over time) $\neq$ multivariate
  $\rightarrow$ processes have values in $\mathbb{R}$

And only a few approaches (Box-Jenkins)

- *e.g.* no Fourier analysis

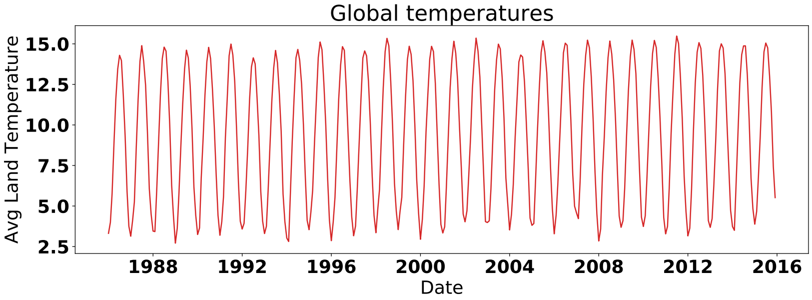
A few examples

Average global land temperature (per month)



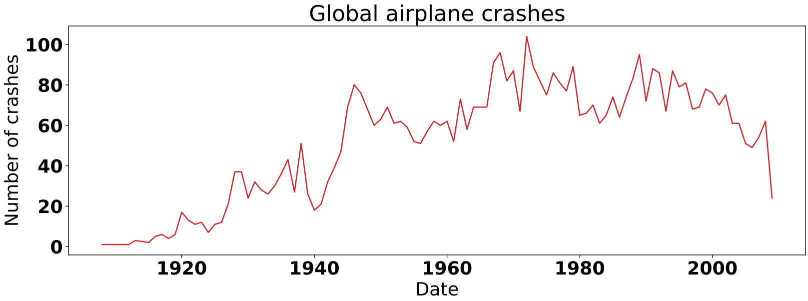
A few examples

Average global land temperature (per month)



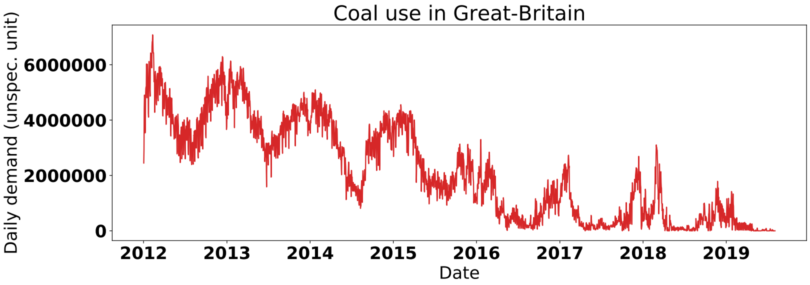
A few examples

Number of civil airplane crashes (per year)

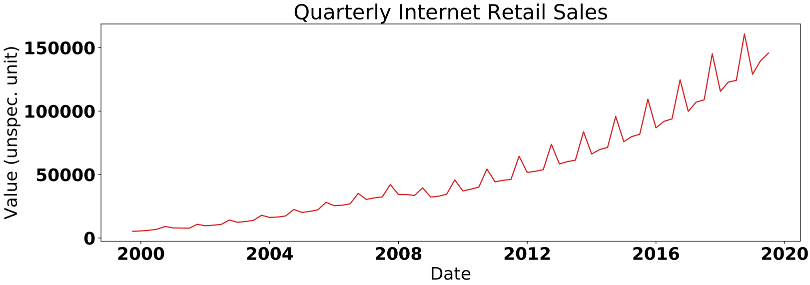


A few examples

Daily demand of power obtained with coal in GB



A few examples



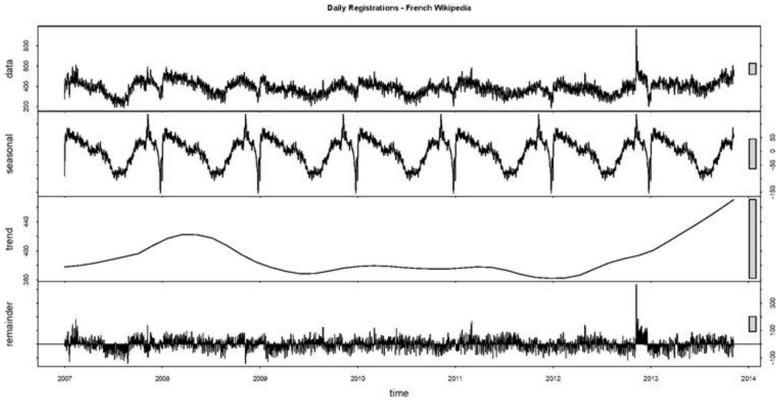
Goals of time series analysis

- Have a simplified description of the data
  $\rightarrow$ improve our understanding (*ex: climate data*)
- Test an assumption
  ex: is there a significant measurable global warming?
- Filter: separate signal from noise
  ex: known physical signal broadcast $\rightarrow$ *filter noise*
- Predict future values
  ex: predict the future demand for a product
- Simulate a process in a complex model
  ex: expectation for the GDP to predict economic activity

How to analyze a time series? (1)

Analyze from Greek *análusis* $\sim$ unravel $\Rightarrow$ decompose

Decompose the time series into parts, for example:

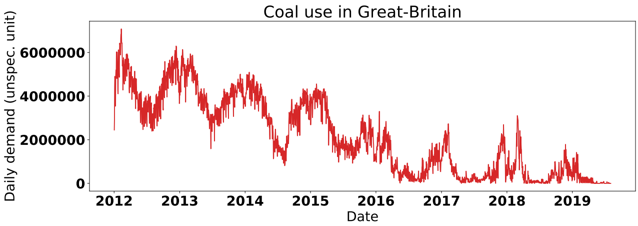


How to analyze a time series? (1)

First step

Plot the time series to:

- identify the existence of a trend *(tendance)*
- uncover seasonal variations *(variations saisonnières)*
- detect changes of behavior
- spot outliers *(valeurs aberrantes)*

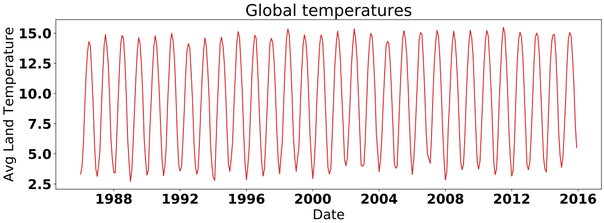


How to analyze a time series? (1)

First step

Plot the time series to:

- identify the existence of a trend *(tendance)*
- uncover seasonal variations *(variations saisonnières)*
- detect changes of behavior
- spot outliers *(valeurs aberrantes)*



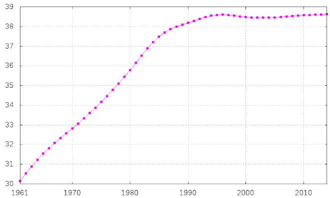
How to analyze a time series? (1)

First step

Plot the time series to:

- identify the existence of a trend *(tendance)*
- uncover seasonal variations *(variations saisonnières)*
- detect changes of behavior
- spot outliers *(valeurs aberrantes)*

Poland population

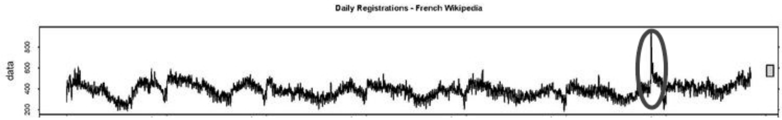


How to analyze a time series? (1)

First step

Plot the time series to:

- identify the existence of a trend *(tendance)*
- uncover seasonal variations *(variations saisonnières)*
- detect changes of behavior
- spot outliers *(valeurs aberrantes)*



How to analyze a time series? (1)

First step

Plot the time series to:

- identify the existence of a trend *(tendance)*
- uncover seasonal variations *(variations saisonnières)*
- detect changes of behavior
- spot outliers *(valeurs aberrantes)*

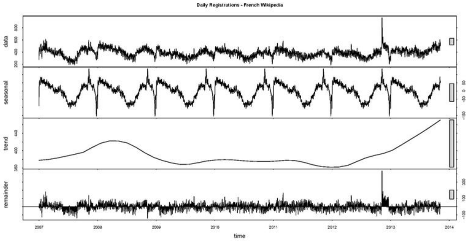
$\rightarrow$ subjective components in this analysis

The classical decomposition

Classical decomposition of the time series

$$X_t = s_t + m_t + r_t$$

- seasonality s_t
- trend m_t
- remainder r_t



Outline

Mean and covariance of a time series

Two fundamental definitions

Let $\{X_t\}$ a time series with $\mathbb{E}[X_t^2] < \infty$ (finite variance)
rk: here we consider X_t as a model

- **mean function** of X_t, defined for all t:

$$\mu_X(t) = \mathbb{E}[X_t]$$

- **covariance function** of X_t, defined for all r, s:

$$\gamma_X(r, s) = Cov(X_r, X_s) = \mathbb{E}[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

Stationarity

Informal definition

A process is **stationary** *(stationnaire)* if its statistical properties are similar when shifted in time

Remarks:

- stationarity is a property of a model (not of data)
- stationary processes are simpler to investigate
  $\Rightarrow$ usual to transform a TS to obtain a stationary process

Stationarity

Informal definition

A process is **stationary** *(stationnaire)* if its statistical properties are similar when shifted in time

Formal definition

A process is said to be **weakly stationary** if

- the mean function $\mu_X(t)$ is independent of $t \Rightarrow \mu_X$
- $\gamma_X(t+h, t)$ is independent of t for any h (including $h = 0$)
  h is called the *lag* *(décalage)*

$$\gamma_X(t+h, t) = \mathbb{E}[(X_{t+h} - \mu_X)(X_t - \mu_X)] = \gamma_X(h)$$

Stationarity

Informal definition

A process is **stationary** *(stationnaire)* if its statistical properties are similar when shifted in time

Formal definition

A process is said to be **strictly (or strongly) stationary** if

- $\forall n$ and $\forall h$

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_{1+h} = x_1, \ldots, X_{n+h} = x_n)$$

Unless specified otherwise, we talk about weak stationarity in the following

Stationarity

Informal definition

A process is **stationary** *(stationnaire)* if its statistical properties are similar when shifted in time

Formal definition

A process is said to be **strictly (or strongly) stationary** if

- $\forall n$ and $\forall h$

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_{1+h} = x_1, \ldots, X_{n+h} = x_n)$$

Unless specified otherwise, we talk about weak stationarity in the following

Autocorrelation function

Notice that for a stationary time series: $\gamma_X(t+h,t) = \gamma_X(h)$
$\Rightarrow$ the covariance function γ_X has one variable (the lag)

Definition

For a stationary time series:

- the **autocovariance function** at lag h is:

$$\gamma_X(h) = Cov(X_{t+h}, X_t)$$

- the **autocorrelation function** (ACF) at lag h is:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}$$

Equivalent on real data

Concepts well defined on models, but what about real data?
Let $\{x_1, \ldots, x_n\}$ be a series of observations

Sample mean

- the **sample mean** estimator is

$$\overline{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

Equivalent on real data

Concepts well defined on models, but what about real data?
Let $\{x_1, \ldots, x_n\}$ be a series of observations

Sample autocovariance function

- the **sample autocovariance function** estimator is

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \overline{x}).(x_t - \overline{x}) \ , \ -n < h < n$$

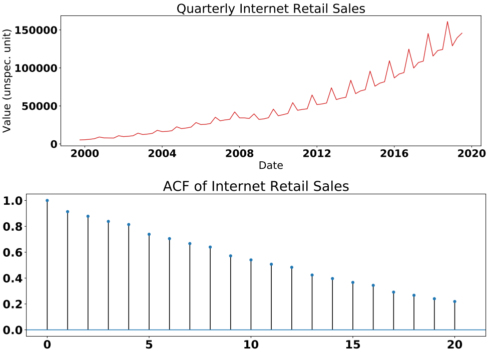
remark: notice the denominator (because of mathematical properties)

- the **sample autocorrelation function** estimator is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \ , \ -n < h < n. \text{ Note that } \hat{\rho}(h) \in [-1; 1]$$

Equivalent on real data

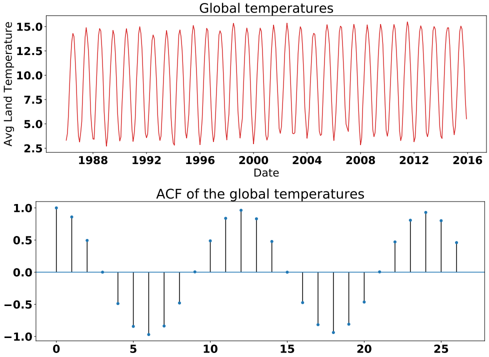
Data with strong trend:



slow decay of correlations with h

Equivalent on real data

Data with strong seasonality:



periodicity on the ACF (here monthly measures $\Rightarrow$ period = 12)

Outline

What is a time series model?

Definition

Time series model: specification of the joint distributions of a sequence of random variables X_t of which the observed data is supposed to be the realization

Remarks:

- suppose to know $\forall t$ the distribs $P(x_1, \dots, x_t, \dots, x_n)$
  $\Rightarrow$ in most case too many parameters to handle
- in practice, we focus on first and second order moments:
  - expected values $\mathbb{E}[X_t]$
  - and expected products $\mathbb{E}[X_{t+h}X_t]$, $h = 1, 2, \dots$

What is a time series model?

Definition

Time series model: specification of the joint distributions of a sequence of random variables X_t of which the observed data is supposed to be the realization

Remarks:

- suppose to know $\forall t$ the distribs $P(x_1, \dots, x_t, \dots, x_n)$
  $\Rightarrow$ in most case too many parameters to handle
- in practice, we focus on first and second order moments:
  - expected values $\mathbb{E}[X_t]$
  - and expected products $\mathbb{E}[X_{t+h}X_t]$, $h = 1, 2, \dots$

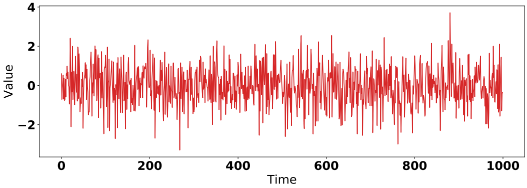
Independent Identically Distributed noise model

IID noise

- independant:
  $P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$
- identically distributed: $P(X_t = x) = P(X_{t'} = x)$

IID noise is obviously stationary

ex: repeated coin flipping with heads=1, tails=-1 should be IID noise



Independent Identically Distributed noise model

IID noise

- independant:
  $P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$
- identically distributed: $P(X_t = x) = P(X_{t'} = x)$

IID noise is obviously stationary

ex: repeated coin flipping with heads=1, tails=-1 should be IID noise

White noise *(bruit blanc)*

Special case IID noise with

- 0 mean: $E[X_t] = 0$
- autocovariance function:
  $\gamma_X(h) = \sigma^2$ if $h = 0$ and $\gamma_X(h) = 0$ if $h \neq 0$

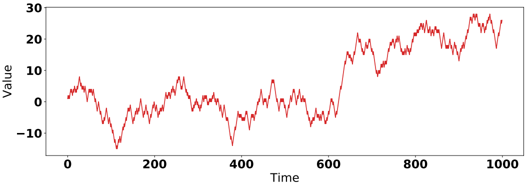
Random Walk model

How to build a random walk? *(marche aléatoire)*

Suppose $\{X_t\}$ is IID noise, then $\{S_t\}$ defined as:

$$S_t = X_1 + \ldots + X_t$$

is a random walk



Random Walk model

How to build a random walk? *(marche aléatoire)*

Suppose $\{X_t\}$ is IID noise, then $\{S_t\}$ defined as:

$$S_t = X_1 + \ldots + X_t$$

is a random walk

Remarks:

- is a random walk stationary?
- it's a summation of an IID process
- and conversely $X_t = S_t - S_{t-1}$, or $S_t = S_{t-1} + X_t$

Random Walk model

How to build a random walk? *(marche aléatoire)*

Suppose $\{X_t\}$ is IID noise, then $\{S_t\}$ defined as:

$$S_t = X_1 + \ldots + X_t$$

is a random walk

Remarks:

- is a random walk stationary? No
- it's a summation of an IID process
- and conversely $X_t = S_t - S_{t-1}$, or $S_t = S_{t-1} + X_t$

Outline

How to analyze a time series? (2)

Second step:

- (if necessary) transform data
- remove the trend and seasonal components to get stationary residuals *(résidus)*

Residual time series obtained (remainder) should be stationary, but not necessarily IID noise...

Back to the classical decomposition

Classical decomposition of the time series

$$X_t = s_t + m_t + r_t$$

- seasonality s_t
- trend m_t
- remainder r_t

What is the difference between seasonality and trend?

- seasonality is periodic $s_{t+d} = s_t$
- seasonality is null on average $\sum_{j=1}^{d} s_j = 0$

Back to the classical decomposition

Classical decomposition of the time series

$$X_t = s_t + m_t + r_t$$

- seasonality s_t
- trend m_t
- remainder r_t

What is the difference between seasonality and trend?

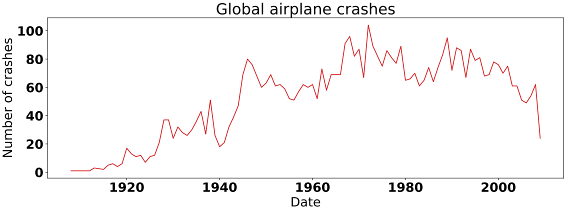
- seasonality is periodic $\quad s_{t+d} = s_t$
- seasonality is null on average $\quad \sum_{j=1}^{d} s_j = 0$

Isolate the trend component

Model and regression

→ cf. course *Regression*

eg. 2nd order polynomial model with least squares regression



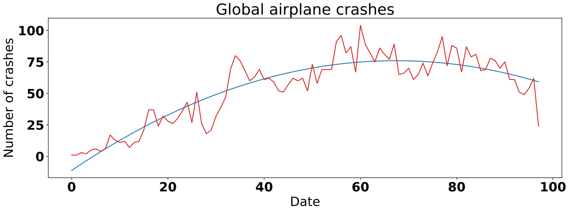
Isolate the trend component

Model and regression

$\rightarrow$ cf. course *Regression*

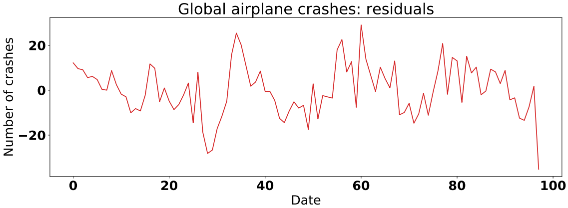
eg. 2nd order polynomial model with least squares regression

$$\text{Minimize } \sum_{t=1}^{n} (x_t - m_t)^2, \text{ with } m_t = a_0 + a_1 t + a_2 t^2$$



Isolate the trend component

Then we plot the residuals $\{x_t - m_t\}$



Questions to ask oneself:

- Is there a perceptible trend? Is it smooth? Do we see stretches (séquences) of values of the same sign?
- Does it look stationary? Does it look like noise?

Isolate seasonal component

Model and regression

Which model?

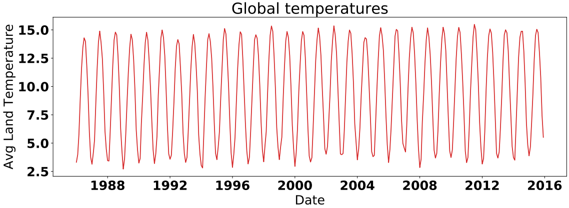
Isolate seasonal component

Model and regression

Which model? Harmonic regression

$$s_t = (a_0 +) \sum_{j=1}^{k} a_j cos\left(\frac{2\pi t}{T_j}\right) + b_j sin\left(\frac{2\pi t}{T_j}\right)$$

where T_j are the expected periods of the process



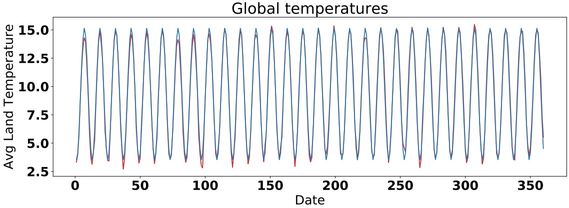
Isolate seasonal component

Model and regression

Which model? Harmonic regression

$$s_t = (a_0 +) \sum_{j=1}^{k} a_j cos\left(\frac{2\pi t}{T_j}\right) + b_j sin\left(\frac{2\pi t}{T_j}\right)$$

where T_j are the expected periods of the process



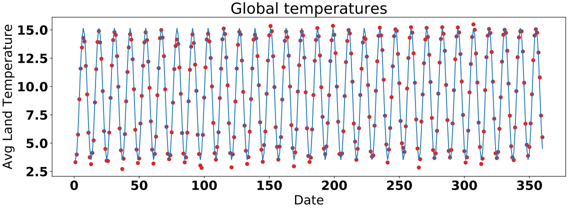
Isolate seasonal component

Model and regression

Which model? Harmonic regression

$$s_t = (a_0 +) \sum_{j=1}^{k} a_j cos\left(\frac{2\pi t}{T_j}\right) + b_j sin\left(\frac{2\pi t}{T_j}\right)$$

where T_j are the expected periods of the process



About pre-processing

Second step:

- (if necessary) transform data
- remove the trend and seasonal components to get stationary residuals

When is it necessary to transform data?

Some cases

- if outliers $\rightarrow$ if justified, discard them
  ex: external stimulus, mistake in data acquisition, ...
- if obvious different regimes
  $\rightarrow$ break data into homogeneous segments
- if noise or seasonality increase with trend level
  $\rightarrow$ logarithmic transformation of the data

About pre-processing

Second step:

- (if necessary) transform data
- remove the trend and seasonal components to get stationary residuals

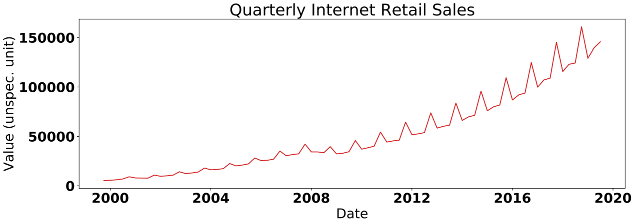
When is it necessary to transform data?

Some cases

- if outliers $\rightarrow$ if justified, discard them
  ex: external stimulus, mistake in data acquisition, ...
- if obvious different regimes
  $\rightarrow$ break data into homogeneous segments
- if noise or seasonality increase with trend level
  $\rightarrow$ logarithmic transformation of the data

Logarithmic transformation

If fluctuations (seasonality, noise) grow with magnitude…



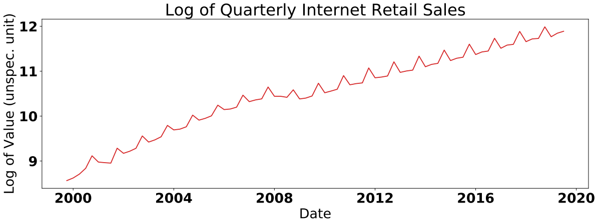
…after logarithmic transform

→ c.f. course *Regression (heteroscedasticity)*

Conduct similar analysis on the transformed time series and reverse the transformations in the end to model the original data

Logarithmic transformation

If fluctuations (seasonality, noise) grow with magnitude…



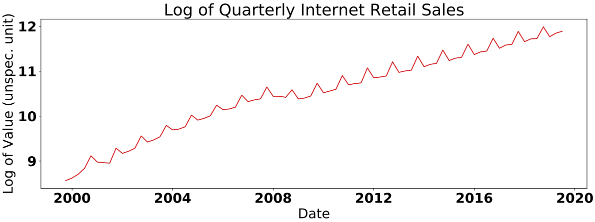
…after logarithmic transform

$\rightarrow$ c.f. course *Regression (heteroscedasticity)*

Conduct similar analysis on the transformed time series and reverse the transformations in the end to model the original data

Logarithmic transformation

If fluctuations (seasonality, noise) grow with magnitude…



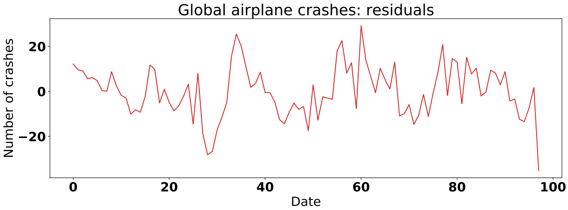
…after logarithmic transform

$\rightarrow$ c.f. course *Regression (heteroscedasticity)*

Conduct similar analysis on the transformed time series and reverse the transformations in the end to model the original data

Stationarity of the residuals

Trend and seasonal components modeled → residuals

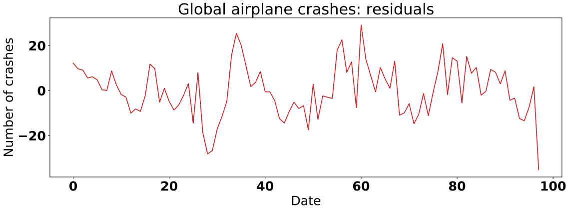


Visual evaluation of stationarity

- Is there a perceptible trend? Is it smooth? Do we see stretches (séquences) of values of the same sign?
- Does it look stationary? Does it look like noise?

Stationarity of the residuals

Trend and seasonal components modeled → residuals



Quantitative evaluation of stationarity?

Testing stationarity of a time series

A general method in Time Series Analysis:

- assume a model
- compute consequences of this model
- test if observations are compatible

$\rightarrow$ c.f. course *Hypothesis testing*

We illustrate this logic with the Dickey-Fuller test

Testing stationarity of a time series

A general method in Time Series Analysis:

- assume a model
- compute consequences of this model
- test if observations are compatible

$\rightarrow$ c.f. course *Hypothesis testing*

We illustrate this logic with the Dickey-Fuller test

Illustration: RW and AR(1) processes

Auto-regressive processes

Consider the random walk model (not stationary):

$$X_t = X_{t-1} + W_t \text{ where } W_t \text{ is the “error”, some white noise}$$

A specific kind of auto-regressive model: AR(1)

$$X_t = \phi X_{t-1} + W_t$$

$\Rightarrow$ Random Walk is AR(1) with $\phi = 1$

Illustration: RW and AR(1) processes

Auto-regressive processes

Consider the random walk model (not stationary):

$$X_t = X_{t-1} + W_t \text{ where } W_t \text{ is the “error”, some white noise}$$

A specific kind of auto-regressive model: AR(1)

$$X_t = \phi X_{t-1} + W_t$$

$\Rightarrow$ Random Walk is AR(1) with $\phi = 1$

Illustration: RW and AR(1) processes

AR(1) processes

$$X_t = \phi X_{t-1} + W_t$$

W_t is a white noise (mean 0, variance σ^2)

Stationary? (necessary conditions)

- $\mathbb{E}[X_t] = 0$
- $\gamma_X(h) = \phi^{|h|}\gamma_X(0) = \phi^{|h|}\frac{\sigma^2}{1-\phi^2} \rightarrow$ problem if $\phi = 1$

Illustration: RW and AR(1) processes

AR(1) processes

$$X_t = \phi X_{t-1} + W_t$$

W_t is a white noise (mean 0, variance σ^2)

Stationary? (necessary conditions)

- $\mathbb{E}[X_t] = 0$
- $\gamma_X(h) = \phi^{|h|}\gamma_X(0) = \phi^{|h|}\frac{\sigma^2}{1-\phi^2} \rightarrow$ problem if $\phi = 1$

Testing for unit roots

Notion of characteristic equation

$$X_t - \phi X_{t-1} - W_t = 0$$
$$\longrightarrow 1 - \phi x = 0$$

here, 1 is a root of this equation if $\phi = 1$

More generally, if we have a process modeled by the equation:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} + \ldots - \phi_p X_{t-p} - W_t = 0$$
$$\longrightarrow 1 - \phi_1 x - \phi_2 x^2 + \ldots - \phi_p x^p = 0$$

if 1 is a root, the process is not stationary (admitted)

Stationarity tests look for unit roots

Testing for unit roots

Notion of characteristic equation

$$X_t - \phi X_{t-1} - W_t = 0$$

$$\longrightarrow 1 - \phi x = 0$$

here, 1 is a root of this equation if $\phi = 1$

More generally, if we have a process modeled by the equation:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} + \ldots - \phi_p X_{t-p} - W_t = 0$$

$$\longrightarrow 1 - \phi_1 x - \phi_2 x^2 + \ldots - \phi_p x^p = 0$$

if 1 is a root, the process is not stationary (admitted)

Stationarity tests look for unit roots

Testing for unit roots

Notion of characteristic equation

$$X_t - \phi X_{t-1} - W_t = 0$$
$$\longrightarrow 1 - \phi x = 0$$

here, 1 is a root of this equation if $\phi = 1$

More generally, if we have a process modeled by the equation:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} + \ldots - \phi_p X_{t-p} - W_t = 0$$
$$\longrightarrow 1 - \phi_1 x - \phi_2 x^2 + \ldots - \phi_p x^p = 0$$

if 1 is a root, the process is not stationary (admitted)

Stationarity tests look for unit roots

Dickey-Fuller test

Is the model $X_t = X_{t-1} + W_t$ consistent with data?

(and actually some other models that we ignore in this course)

We suppose that we follow this model and we test for $\phi = 1$

- H_0: $\phi = 1$ and process is not stationary (unit root)
- H_1: $|\phi| < 1$ and process is stationary (no unit root)

Estimate (ϕ -1) from data

compute $\frac{\hat{\phi}-1}{\hat{\sigma}/\sqrt{N}}$ and compare to Dickey-Fuller distribution

Dickey-Fuller test

Is the model $X_t = X_{t-1} + W_t$ consistent with data?

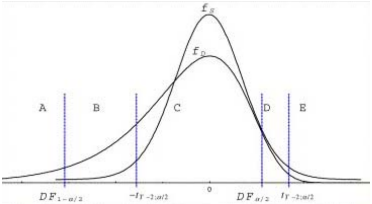
(and actually some other models that we ignore in this course)

We suppose that we follow this model and we test for $\phi = 1$

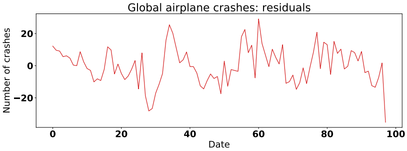
- H_0: $\phi = 1$ and process is not stationary (unit root)
- H_1: $|\phi| < 1$ and process is stationary (no unit root)

Estimate (ϕ -1) from data

compute $\frac{\hat{\phi}-1}{\hat{\sigma}/\sqrt{N}}$ and compare to Dickey-Fuller distribution



Testing in practice (airplane crashes residuals)

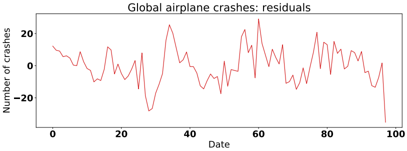


$$\frac{\hat{\phi}-1}{\sigma/\sqrt{N}} = -2.39 \text{ from data}$$

critical value from table (for $p = 0.05$): $\frac{\hat{\phi}-1}{\sigma/\sqrt{N}} < -1.944$

critical value from table (for $p = 0.025$): $\frac{\hat{\phi}-1}{\sigma/\sqrt{N}} < -2.234$

Testing in practice (airplane crashes residuals)



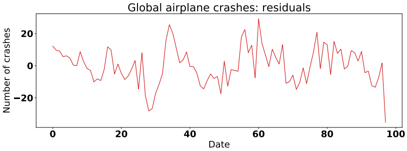
$$\frac{\hat{\phi}-1}{\sigma/\sqrt{N}} = -2.39 \text{ from data}$$

critical value from table (for $p = 0.05$): $\frac{\hat{\phi}-1}{\sigma/\sqrt{N}} < -1.944$

critical value from table (for $p = 0.025$): $\frac{\hat{\phi}-1}{\sigma/\sqrt{N}} < -2.234$

$\Rightarrow$ can reject the random walk model, possibly stationary

Testing in practice (airplane crashes residuals)



$$\frac{\hat{\phi}-1}{\sigma/\sqrt{N}} = -2.39 \text{ from data}$$

critical value from table (for $p = 0.05$): $\frac{\hat{\phi}-1}{\sigma/\sqrt{N}} < -1.944$

critical value from table (for $p = 0.025$): $\frac{\hat{\phi}-1}{\sigma/\sqrt{N}} < -2.234$

$\Rightarrow$ can reject the random walk model, possibly stationary

- other tests available, with other models
  Augmented Dickey-Fuller, Phillips-Perron, . . .
- in general implemented in statistic software

Toward ARMA models

A very standard process model:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + W_t$$

autoregressive with a memory of length $p \rightarrow$ AR(p)

Toward ARMA models

A very standard process model:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + W_t$$

autoregressive with a memory of length $p \rightarrow$ AR(p)

$\longrightarrow$ family of ARMA models (next course)

Studying time series in python

Among several options, `pandas` library

A few useful functions:

- Load data as dataframe:
  `read_csv` from `pandas` library
- Fitting:
  `curve_fit` in `scipy.optimize` library
- Autocorrelation function:
  `plot_ACF` in `statsmodels` library