# Advanced Machine Learning

## 7. Time Series Analytics

### Part Ia - Introduction and Data Preparation

*Prof. Dr. Volker Herbort*

Technische Hochschule Ulm
University of Applied Sciences

# Contents

- Introduction to Time Series
- Data Preparation for Time Series

# Introduction to Time Series

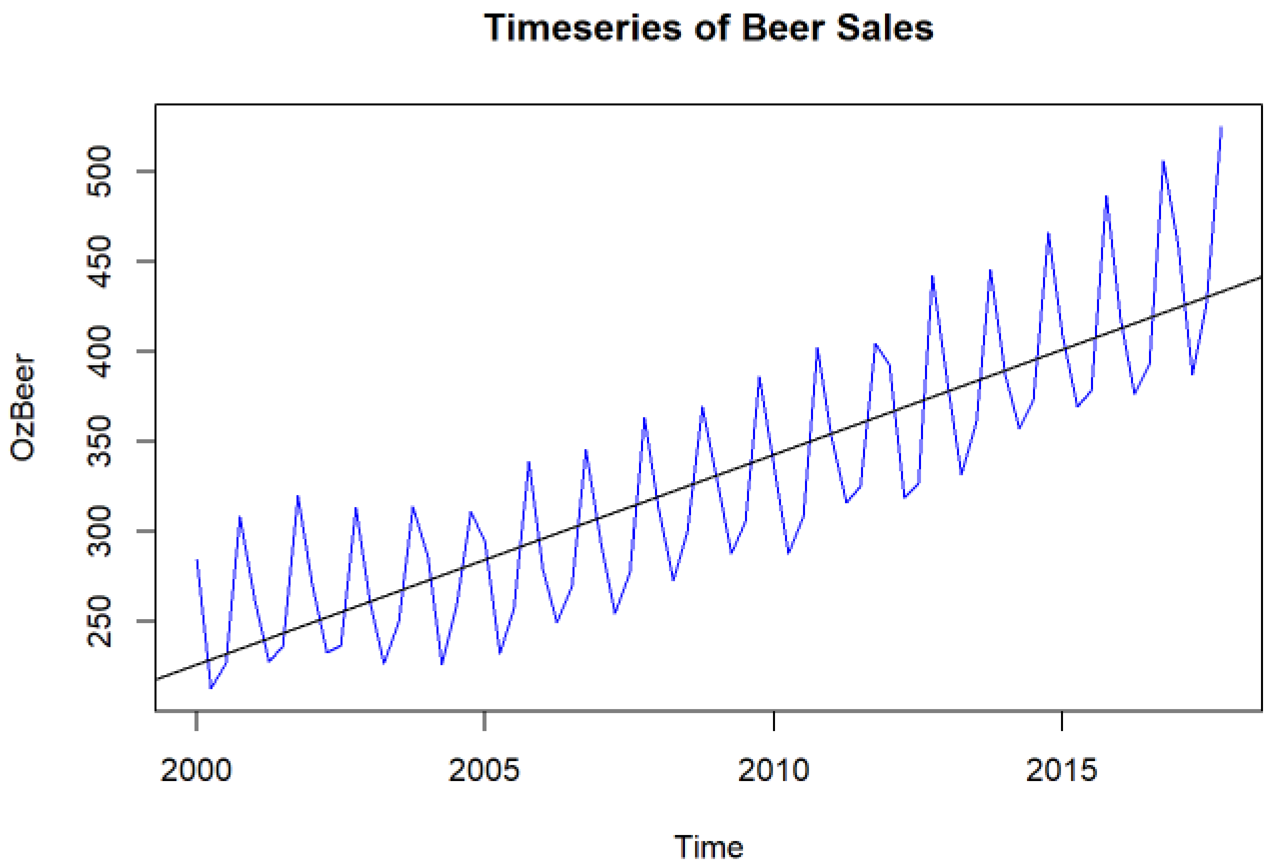
- Definition
  - A time series is a series of data points y indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time *t*. Thus it is a sequence of discrete-time data.[1]
  - A time series can be defined as

    $$Y = \{y_t: t \in T\}$$

    where T can be of any time intervall like seconds, days or years.

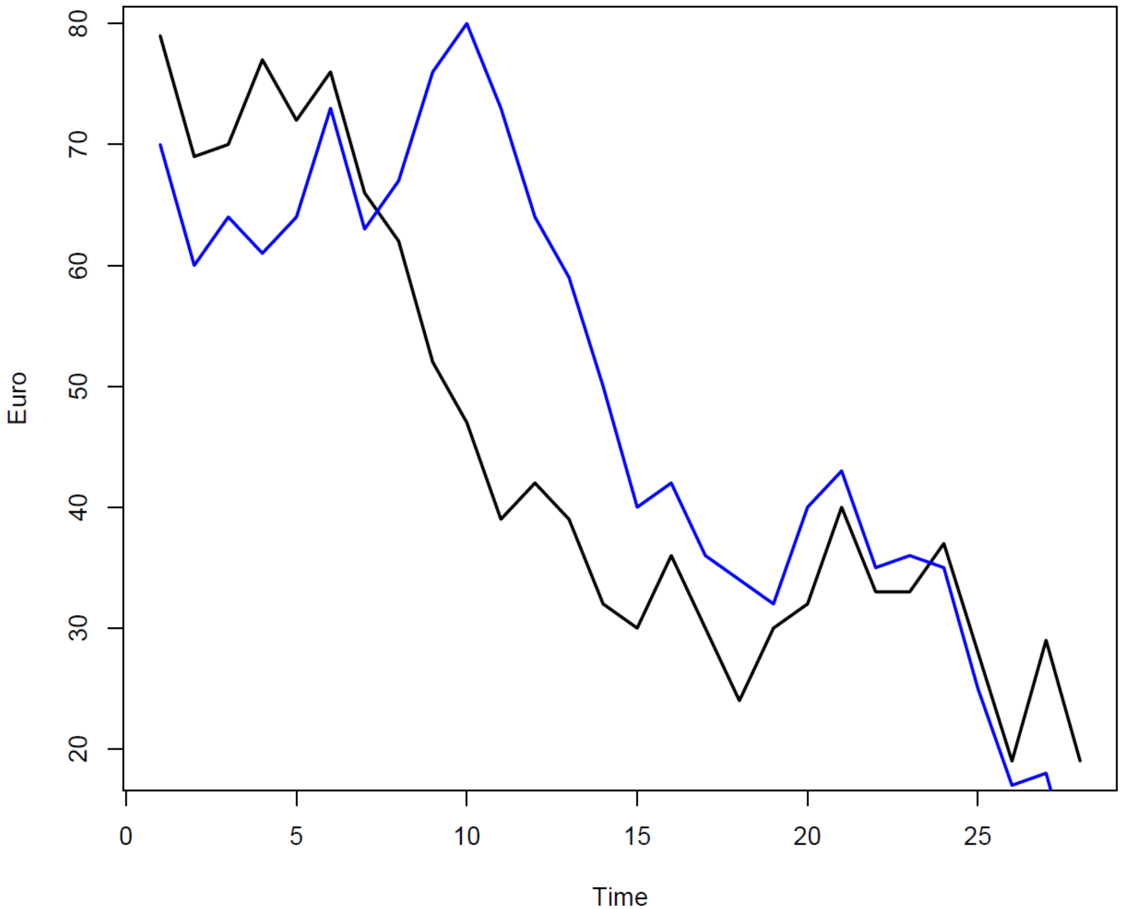
# Time Series Examples

- Beer Sales



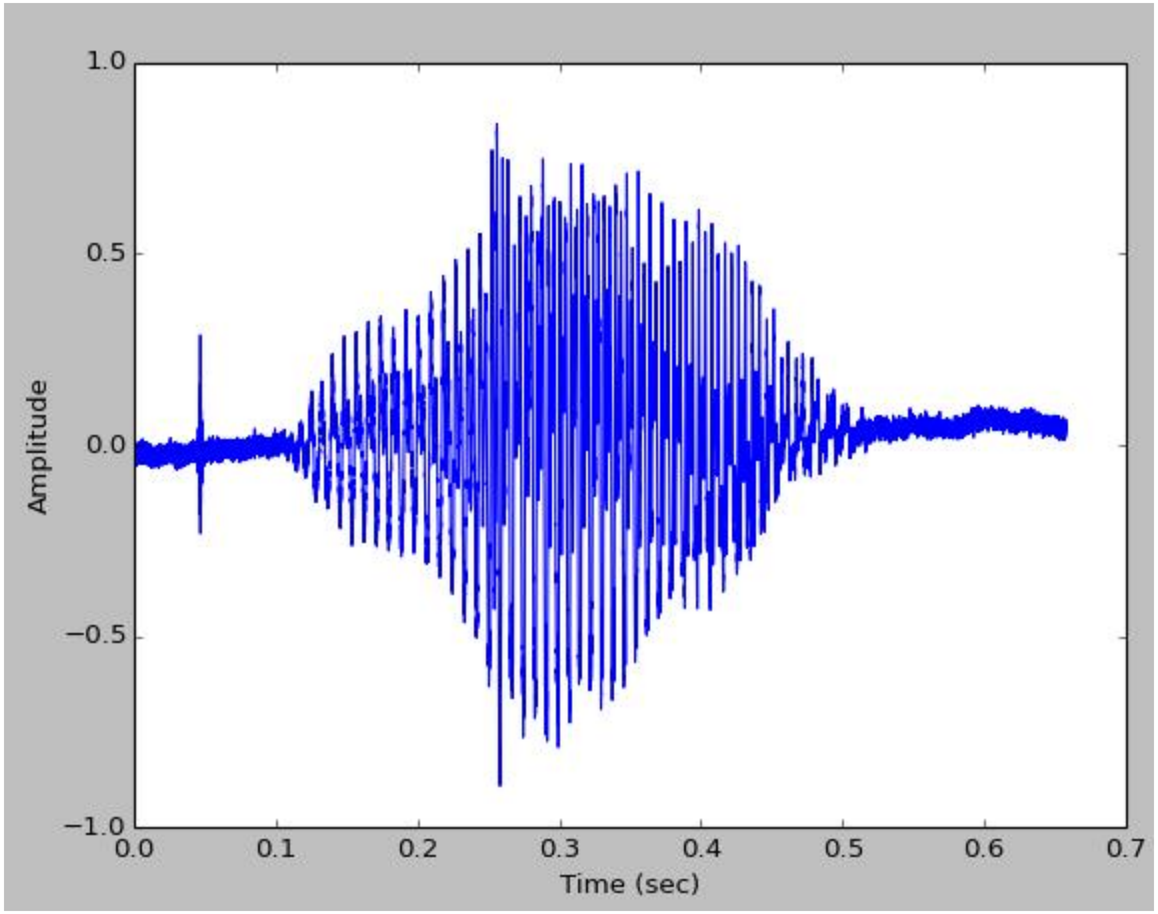
# Time Series Examples ctd.

- Product Prices



# Time Series Examples ctd.

- Speech recognition



# Time Series Characteristics

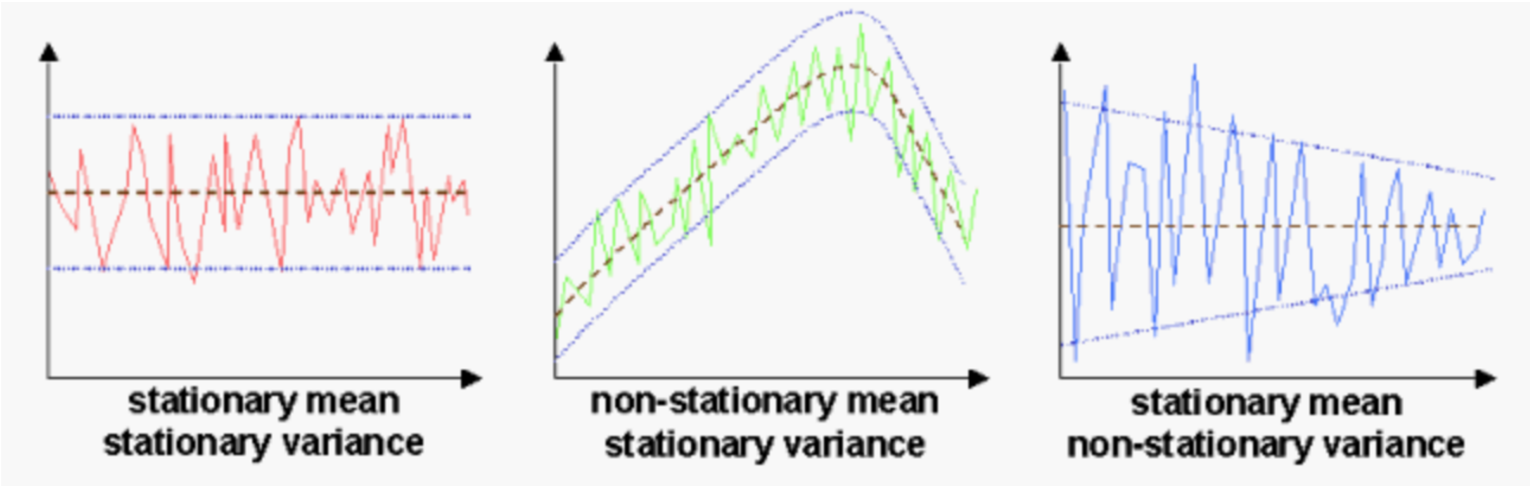
- Special kind of datasets
  - X is always time related

  - Discrete set of values (often discritized from countinous values by recording interval)
    - Univariate: One variable measured over time.
    - Multivariate: Multiple variables measured over time.

  - Can be analysed locally or globally

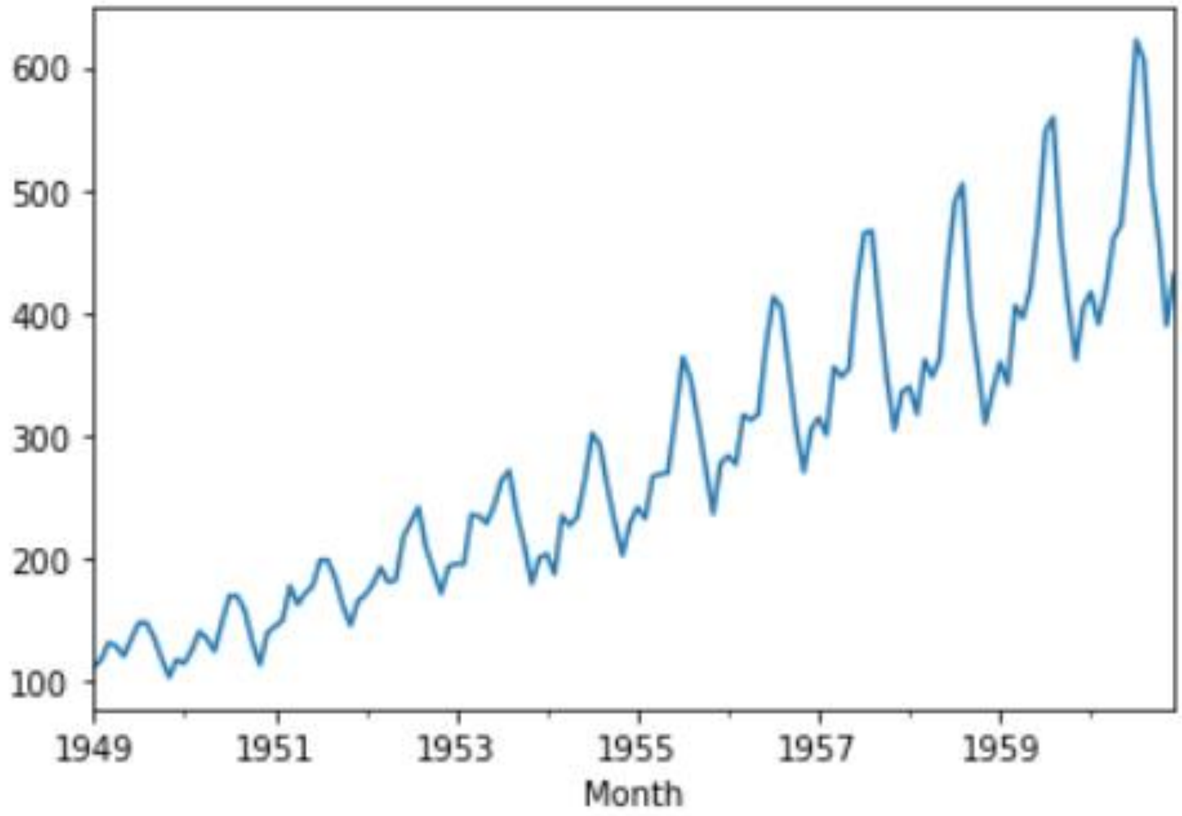
# Stationarity (weak)

- Describes the characteristics of the underlying process. A time series (process) is stationary if
  - the mean (μ) is constant
  - the standard deviation (σ) is constant




  - The auto covariance is independent from time *t*

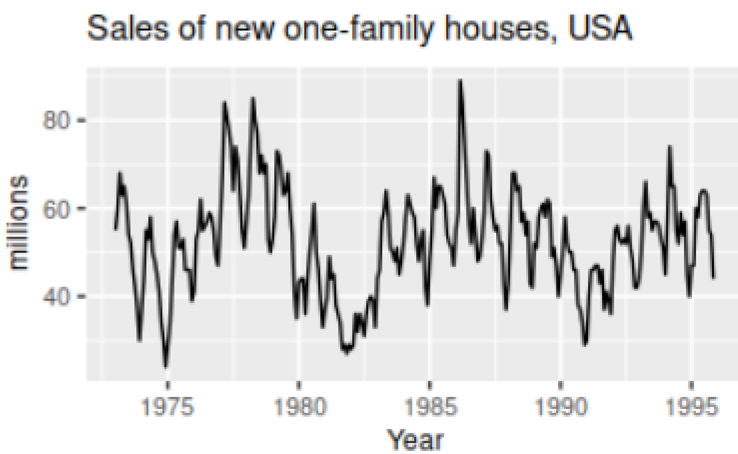
# Why is this time series non- stationary?

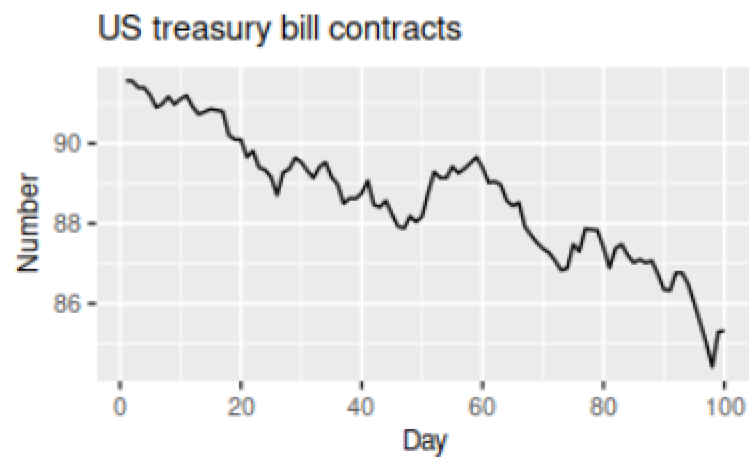


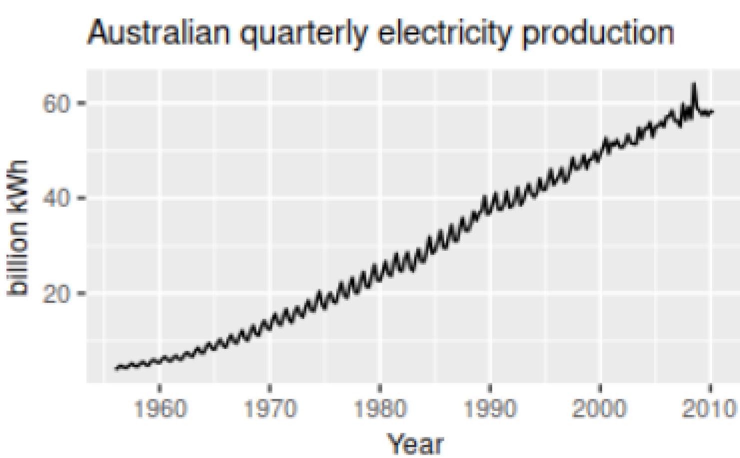
# Time Series Components

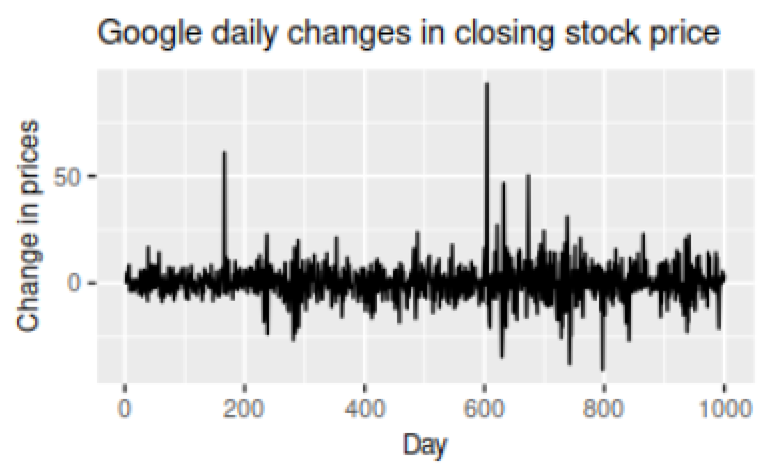
- A time series can be decomposed into:
  - Trend (*T*)
    - Constant increase or decrease
  - Cyclical Component (*C*)
    - Recurring but non-periodic fluctuations
  - Seasonality (*S*)
    - Recurring periodic increase or decrease over time e.g. yearly ups and downs
  - Residuals (*R*)
    - Noise component for random, irregular influences.

# Which components can you spot?



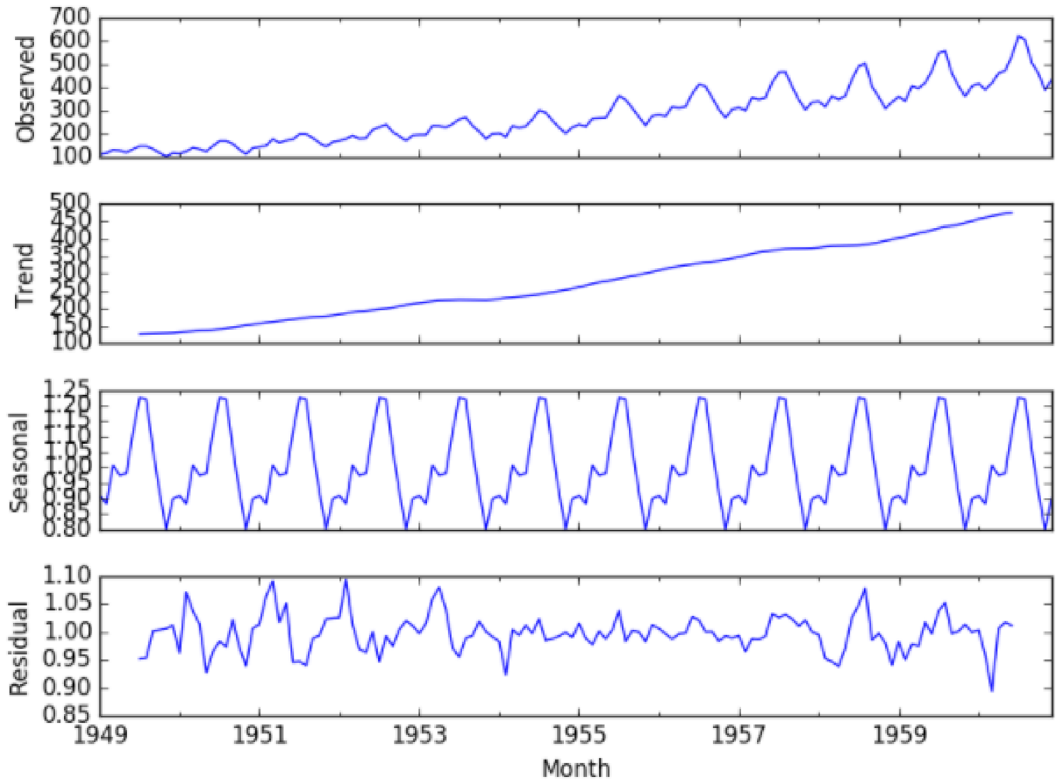




Source: Forecasting: Principles and Practice,Rob J Hyndman and George Athanasopoulos,Monash University, Australia

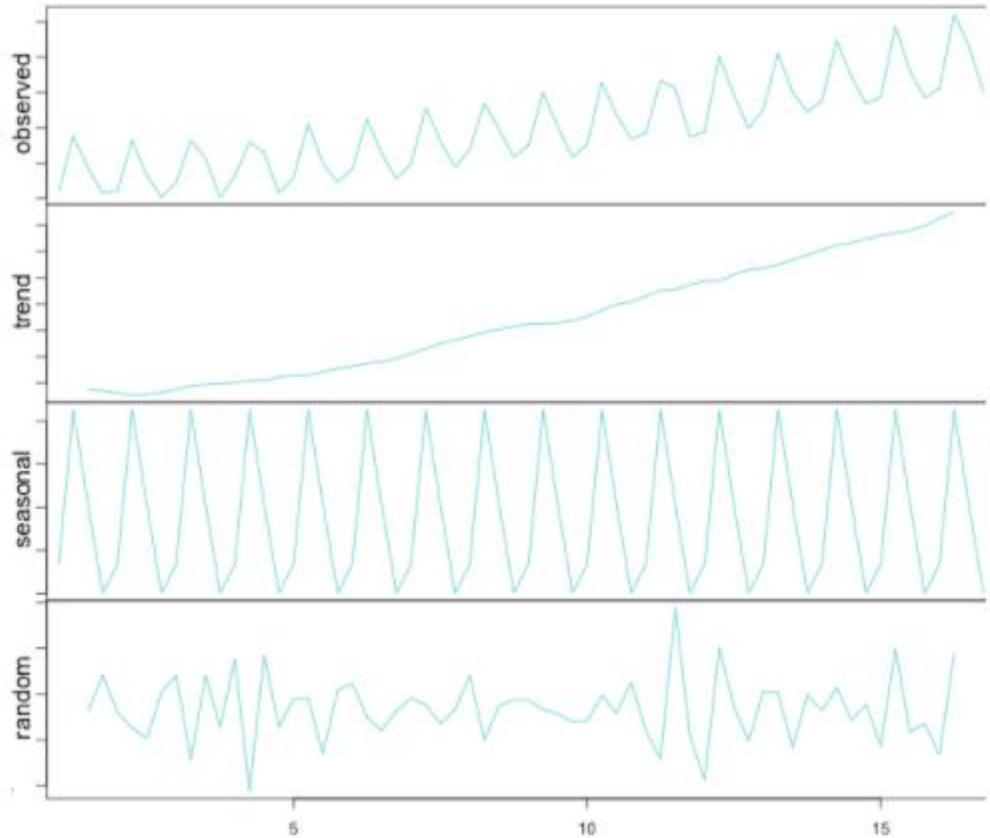
# Time Series Components

- Decomposed time series (multiplicative)
  - $Y_t = t_t * s_t * r_t * c_t$



# Time Series Components

- Decomposed time series (additive)
  - $Y_t = t_t + s_t + r_t + c_t$



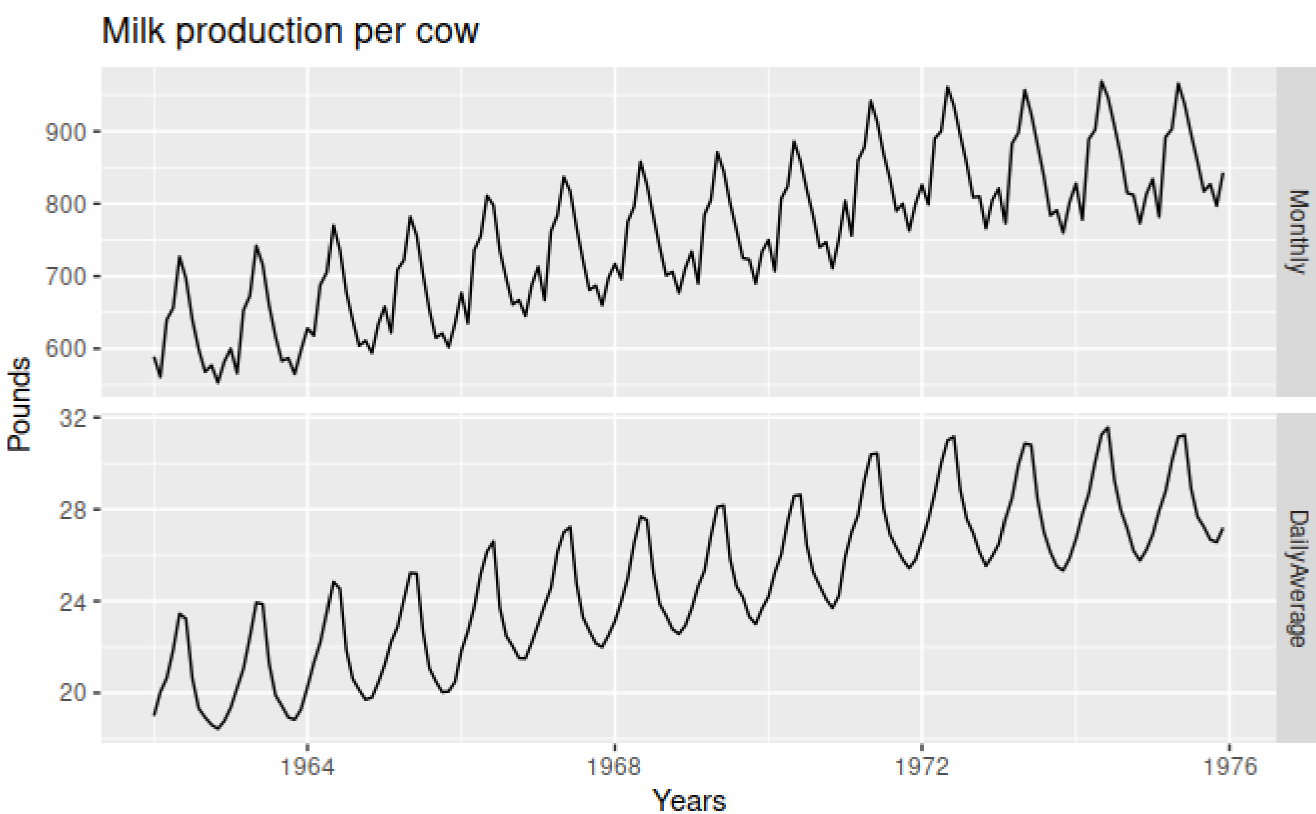
# Contents

# Data Preparation for Time Series

- Idea
  - Transform the series data to a simpler form
  - Easier to predict => simpler models may be used
  - Common Preparations
    - Calendar adjustments
      - Affects the granularity of the data
      - Depends on the type of prediction
    - Population adjustments
      - Normalizes data
    - Inflation adjustments
      - Only applicable for money related data

# Adjustments in Time Series

- Calendar adjustments
  - E.g. monthly data is affected by days in month



# Adjustments in Time Series

- Population adjustments
  - Total data may be affected by the population size
    - Use relative data whenever suitable
    - e.g. average, sum or count per capita
- Inflation adjustments
  - Data related to money should usually be adjusted
    - Choose a reference year and adjust the series
    - E.g. Consumer Price Index (CPI)

# Transformations

- Data shows in- or decreasing variation with level
  - Log transformation
    - $w_t = \log(y_t)$
  - Power transformations
    - e.g. Square or Square Root
    - $w_t = y_t^p$

# Transformations

- More sophisticated…
  - Box-Cox transformation

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^\lambda - 1)/\lambda & \text{otherwise.} \end{cases}$$

    - log is always ln (natural logarithm)
    - λ has to be configured according to visual shape
      - if λ=1 then $w_t = y_t - 1$ => no change in shape
      - else shape will change
  - Revert transformation

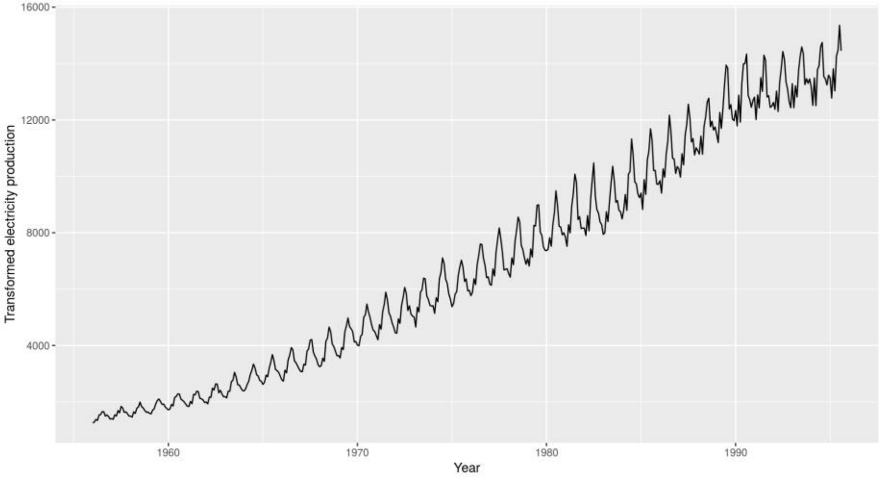
$$y_t = \begin{cases} \exp(w_t) & \text{if } \lambda = 0; \\ (\lambda w_t + 1)^{1/\lambda} & \text{otherwise.} \end{cases}$$

# Choosing λ (lambda)

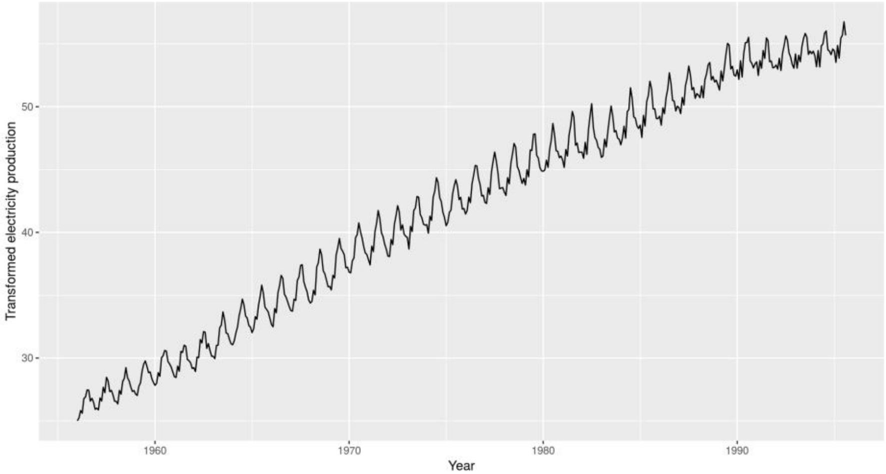
- λ has to be configured according to visual shape
  - λ = -1. is a reciprocal transform.
  - λ = -0.5 is a reciprocal square root transform.
  - λ = 0.0 is a log transform.
  - λ = 0.5 is a square root transform.
  - λ = 1.0 is no transform.

# Box-Cox: Electricity production

- λ=1




- λ=0.3



# Rolling/Sliding Window

- Combination of values of n time stamps per time step
  - E.g. n = 5

| *i* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* |
|---|---|---|---|---|---|---|---|---|---|---|
| *t0* | 112 | 118 | 132 | 129 | 121 | 135 | 148 | 148 | 136 | 119 |
| *t1* | 112 | 118 | 132 | 129 | 121 | 135 | 148 | 148 | 136 | 119 |
| *t2* | 112 | 118 | 132 | 129 | 121 | 135 | 148 | 148 | 136 | 119 |
| *t..* | … | | | | | | | | | |

  - Compute aggregates like sum, min, max, average.

# Moving Average (MA)

- Simple Moving Average (SMA): Compute the average of the **preceeding** window to t.
  - $m_{MA}^{(n)} = \frac{1}{n} \sum_{j=0}^{n-1} y_{t-j}$

- Weighted MA (WMA)
  - $m_{MA}^{(n)} = \frac{1}{n} \sum_{j=0}^{n-1} \alpha_{t-j} y_{t-j}$

# Moving Average Smoothing

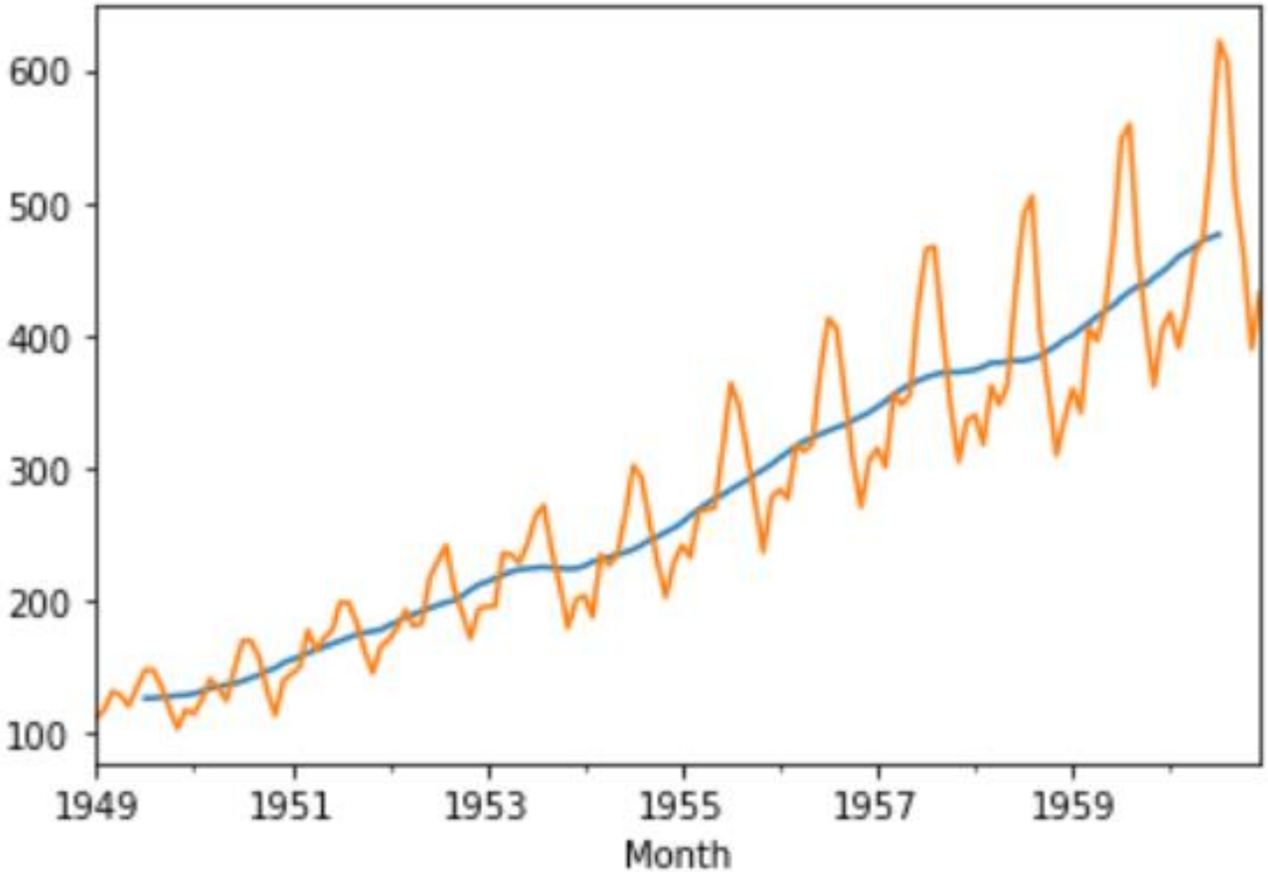
- Centered Moving Average of order $n$

$$\hat{T}_t = \frac{1}{n}\sum_{j=-k}^{k} y_{t+j},\ n = 2k + 1$$

  - Computes the average of preceeding and following values to t
  - E.g. $m_{MA}^{(3)} = \frac{1}{3}(y_{t-1} + y_t + y_{t+1})$

# Smoothing using SMA

- e.g. m = 12



# Advanced Machine Learning

## 7. Time Series Analytics

### Part Ib - Basic Models for Forecasting

*Prof. Dr. Volker Herbort*

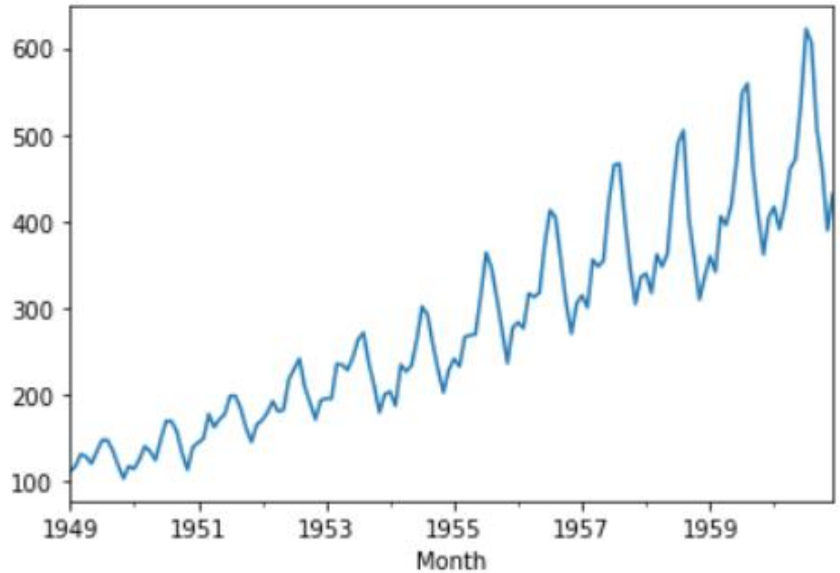
Technische Hochschule Ulm
University of Applied Sciences

# Basic Models for Time Series

- Time Series Analytics
  - Prediction/Forecasting
    - e.g. infections, share price or sales figures
  - Classification
    - e.g. speech recognition, event detection or malfunction
  - Clustering
    - e.g. energy consumption behavior

# Predicting Time Series Values

- How can you predict how many people use a plane in January 1962?




- Historic values are determined by value of *t*.
- What are available features for future values???

# Lag

- A lag describes the size of the time shift of a timeseries. The k$^{th}$ lag is the time period that happened "k" time points before time t.
- For example:
  - $lag1(y_2) = y_1 \; or \; lag4(y_9) = y_5$

| Index | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Time Series | 112 | 118 | 132 | 129 | 121 |
| lag1 | | 112 | 118 | 132 | 129 |
| lag2 | | | 112 | 118 | 132 |
| lag3 | | | | 112 | 118 |
| … | | | | | 112 |

# Simple Methods

- Sometimes good prediction can be done by using…
  - Average method

    $$\hat{y}_{T+h|T} = \bar{y} = (y_1 + \cdots + y_T)/T$$

  - Naive method
    - Prediction is equal to preceeding value

    $$\hat{y}_{T+h|T} = y_T$$

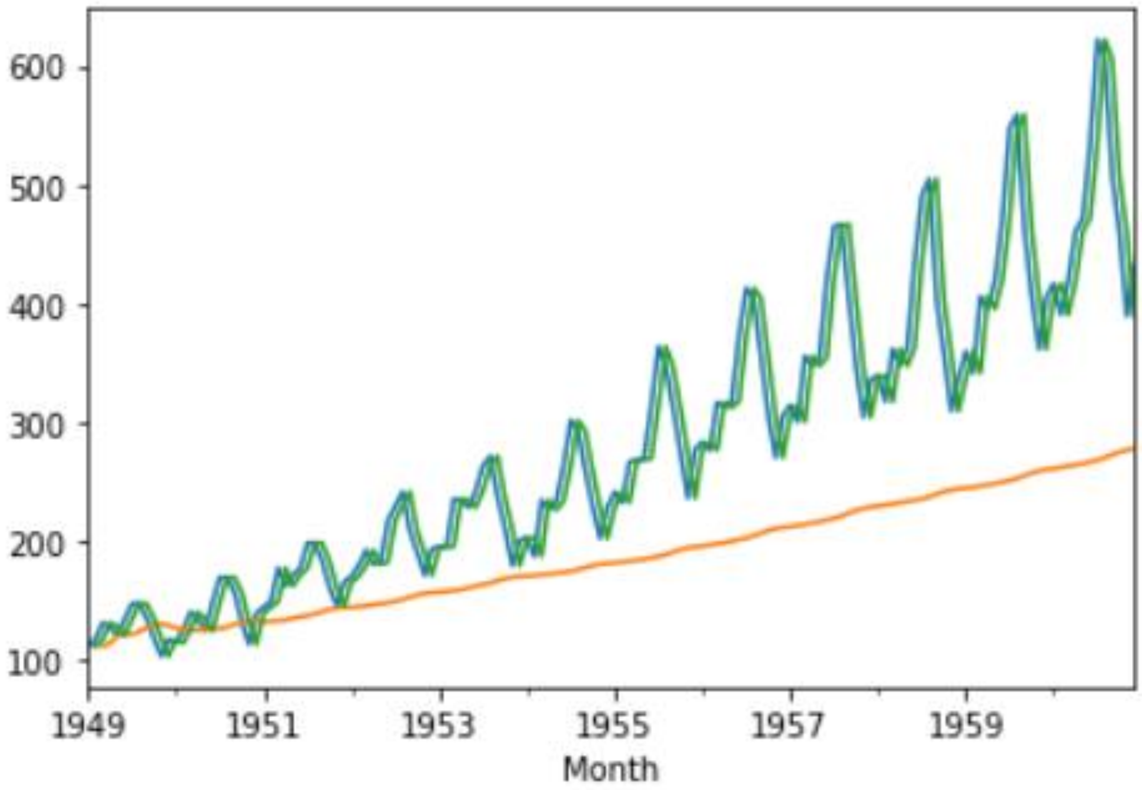
# Digression: Random Walk Model

- A.k.a. Naive estimator, next value is last value + error

$$y_t = y_{t-l} + \varepsilon_t$$

- Assumptions:
  - future movements are unpredictable
  - are equally likely to be up or down.

# Results for Airline passengers



# Simple Methods ctd.

- More sophisticated naive methods
  - Seasonal Naive method
    - Prediction is equal to preceeding value in season

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$$
$$(m = season, k = complete\ years\ (floor))$$

  - Drift method
    - Includes modelling of a change over time (trend)

$$\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1})$$

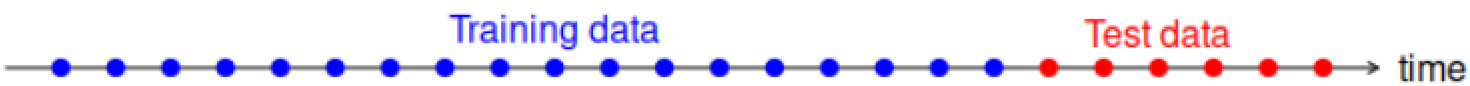
# Digression: Fitted Values

- Results of a one-step forcast model
  - Based on all values $y_1 \dots y_{t-1}$ in a time series
  - Denoted as $\hat{y}_{t|t-1} = \hat{y}_t$

- Estimate is based on all the data at hand
  - No real forecast values
  - Usually includes some deviation from the actual values -> Residuals

# Evaluate Model Quality

- Actual forecast quality without overfitting
  - Hold-Out Method
    - Test Data $\{y_{T+1}, y_{T+2}, \dots\}$ of at least the size of forecast interval
    - Training Data $\{y_1, y_2, \dots, y_T\}$ used for fitting the model




    - Forecast error

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T}$$

      - Based on Test Data
      - Can be computed for multi-step forecast
      - Outcome ist much more representative for new data!

# Evaluate Forecast Accurracy

- Ways to summarize Forecast Error
  - Scale dependent $e_t$
    - Mean absolute error: MAE=$\text{mean}(|e_t|)$,
    - Root mean squared error: RSME = $\sqrt{\text{mean}(e_t^2)}$
  - Percentage
    - $p_t = 100e_t/y_t$
    - Mean absolute percentage error: MAPE = $\text{mean}(|p_t|)$

# Advanced Machine Learning

## 7. Time Series Analytics

### Part Ic - Basic Models for Forecasting ctd.

*Prof. Dr. Volker Herbort*

Technische Hochschule Ulm
University of Applied Sciences

# Auto Regression (AR)

- Estimate a value $y_t$ based on it‘s predecessor $y_{t-1}$ (lag 1) and the use of a linear function
  - $y_t = AR(1) = a + \beta y_{t-1} + \varepsilon_t$
  - $a = level\ (e.g.\ mean)$
  - $\beta = weight$
  - $\varepsilon_t = white\ noise => ideal\ residuals$

# Degression: White Noise

- A time series is white noise if the variables **are independent** and identically distributed with a **mean of zero**. => stationary!
- This means that all variables have the **same variance** ($\sigma^2$) and each value has a **zero correlation** with all other values in the series.
- Relevance
  - **Predictability**: White noise is random.

    => You cannot reasonably model it and make predictions.
  - **Model Diagnostics**: Series of errors from a time series forecast model should ideally be white noise.

# Auto Regression (AR)

- General Definition AR(p):
  - Use the latest p lags for the model and use a linear function to predict the next value $y_t$.
  - $y_t = AR(p) = a + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t$
- Parameter estimation for AR Models
  - Mean Squared Errors

# MA Model

- MA can also be used as a model
  - $y_t = c + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}$
    - q = order (number of previous values)
    - ε = error
    - c = constant (often mean)
  - Similar to AR models but combination of previous prediction errors (no real regression of values).

# Auto Correlation Function (ACF)

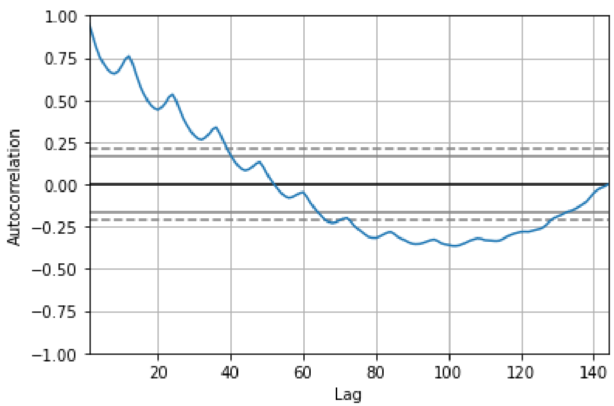
- Use Case:
  - Find out how dependent values $y_t$ are from their predecessor $y_{t-1}$ (or *Y* from its lagged series)
  - Auto Correlation Function (ACF) γ for $y_1$ and $y_2$
    - $\gamma = \frac{\tau(y_1,y_2)}{\sigma_{y1}\sigma_{y2}}$
    - $\tau(y_1, y_2) = Cov(y_1, y_2) = (y_1 - \mu_1)(y_2 - \mu_2)$
    - $\mu = mean$
    - $\sigma = standard\ deviation$

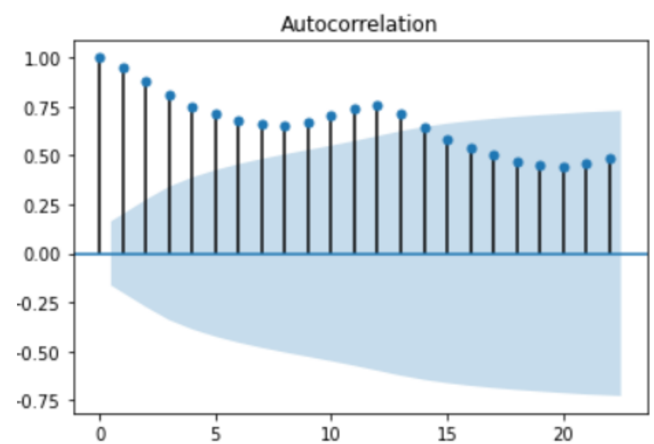
# ACF Ctd.

- Assuming stationarity it is a bit simpler
  - Mean and standard deviation are constant
    - $\mu_{y1} = \mu_{y2} = \mu_Y$
    - $\sigma_{y1} = \sigma_{y2} = \sigma_Y$
  - Therfore
    - $\mu_Y = \frac{1}{n}\sum_{i=1}^{n} y_i$
    - $\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\mu_Y)^2}{n}}$
    - $\gamma = \frac{\sum_{i=1}^{n-1}(y_i-\mu_y)(y_{i+1}-\mu_y)}{\sum_{i=1}^{n}(y_i-\mu_Y)^2}$

# Correlogram Airline Passengers

- Correlation between $y_t$ and its lag series $y_{t-l}$




- Can you spot the components?

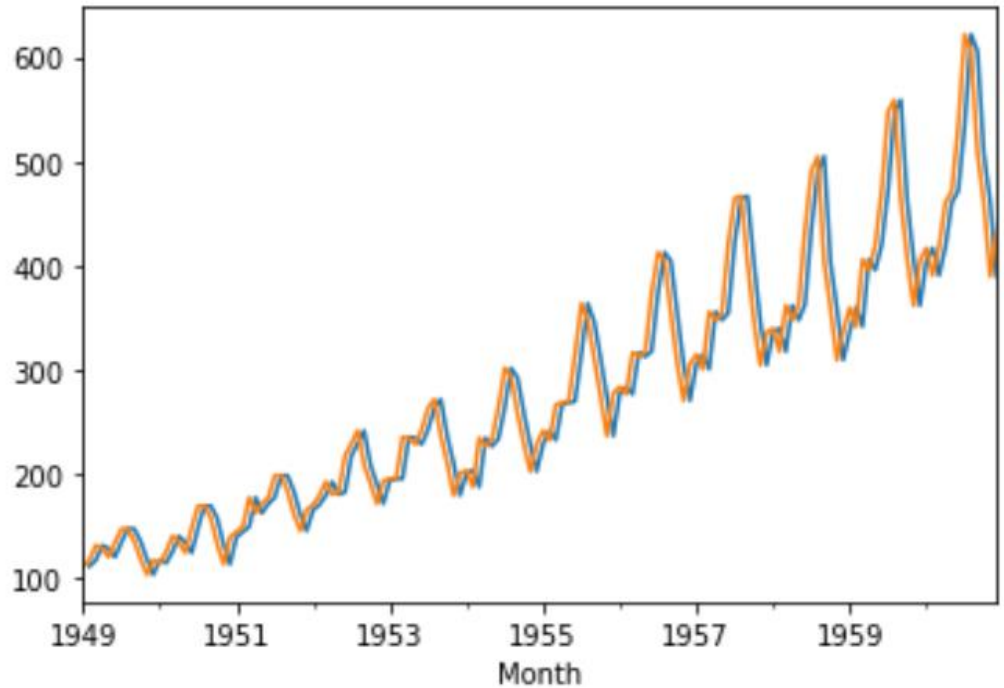
# Stationarity and ACF

- ACF plot is useful for identifying non-stationary time series.
  - stationary time series: ACF drops to zero relatively quickly
  - non-stationary: ACF decreases slowly

- Also, for non-stationary data, the value of $r_1$ is often large and positive.

# Auto Correlation Example

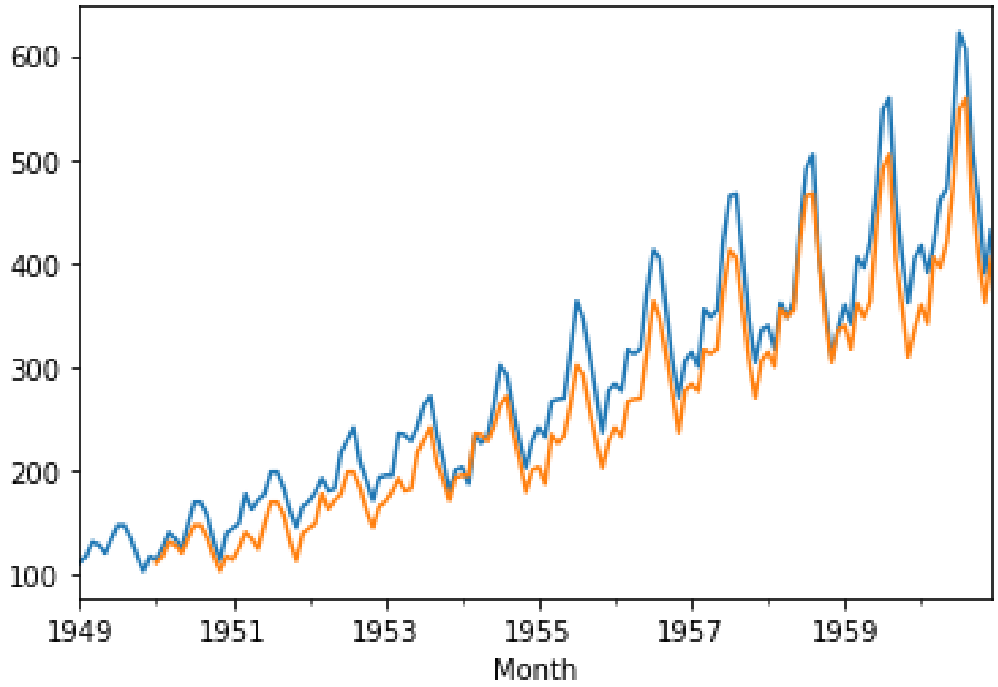
- Correlation between $y_t$ and $y_{t-1}$
  - r = 0,96…




- Is there even a higher correlation possible? What about $y_{t-12}$?

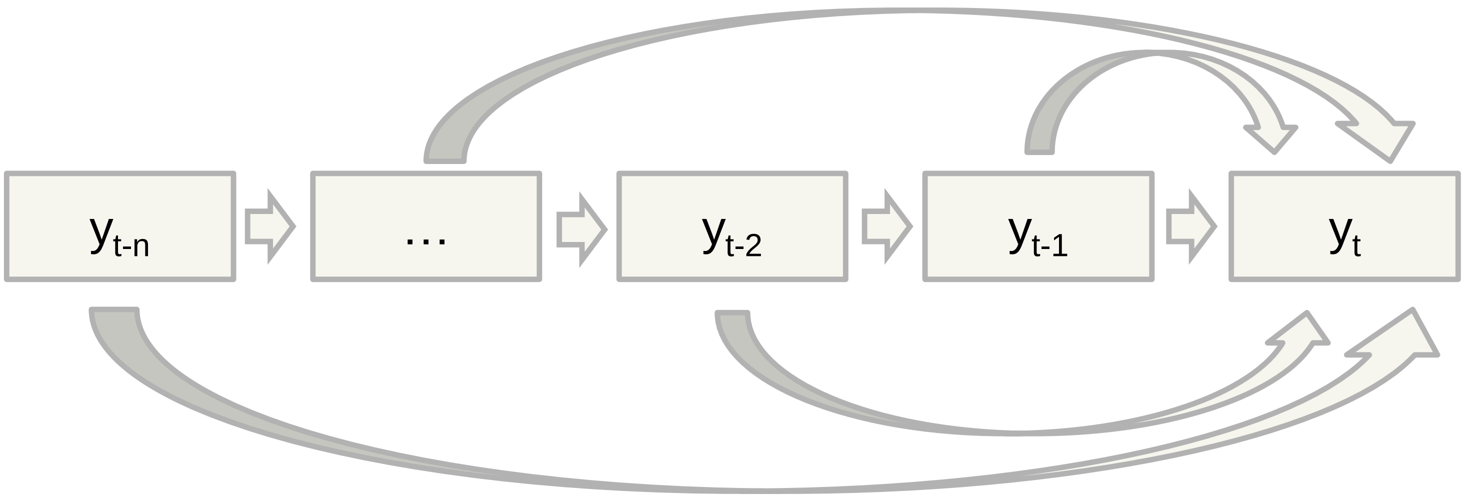
# Auto Correlation Example

- Correlation between $y_t$ and $y_{t-12}$
  - r = 0,99…



# Partial Autocorrelation Function (PACF)

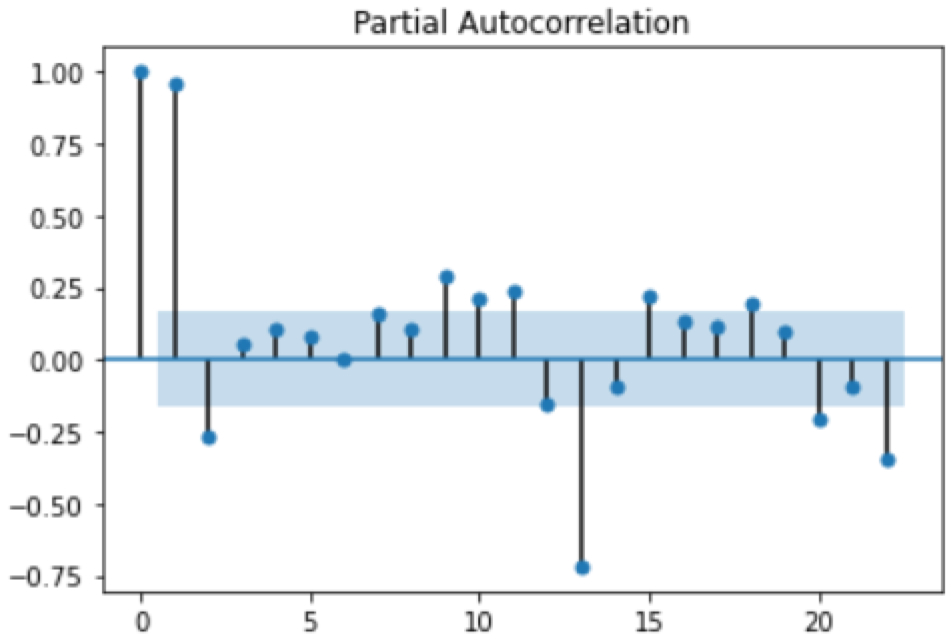
- Influences on current value




- How does e.g. $y_{t-2}$ directly affect $y_t$?

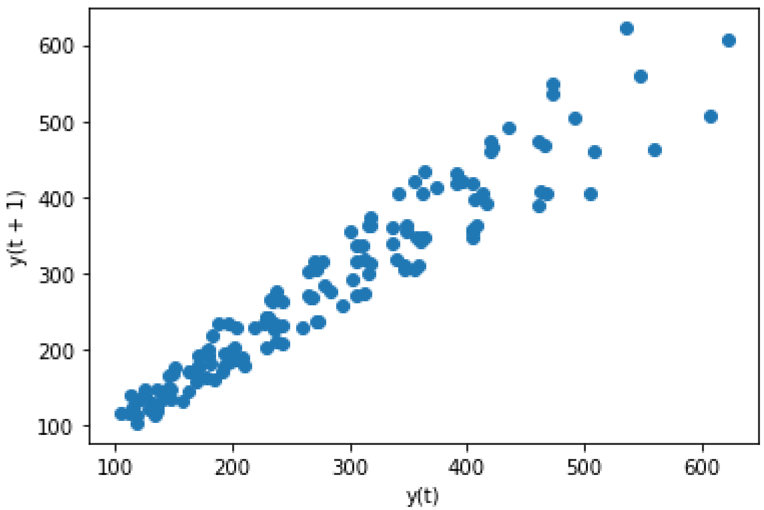
# PACF Plot

- Basically determined by the Coefficients of AR Model

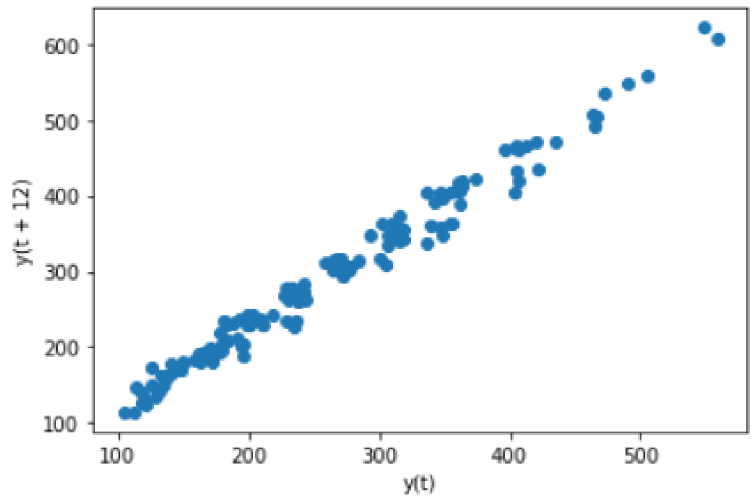


# Lag Plot

- Scatterplot for $y_t$ vs $y_{t+l}$

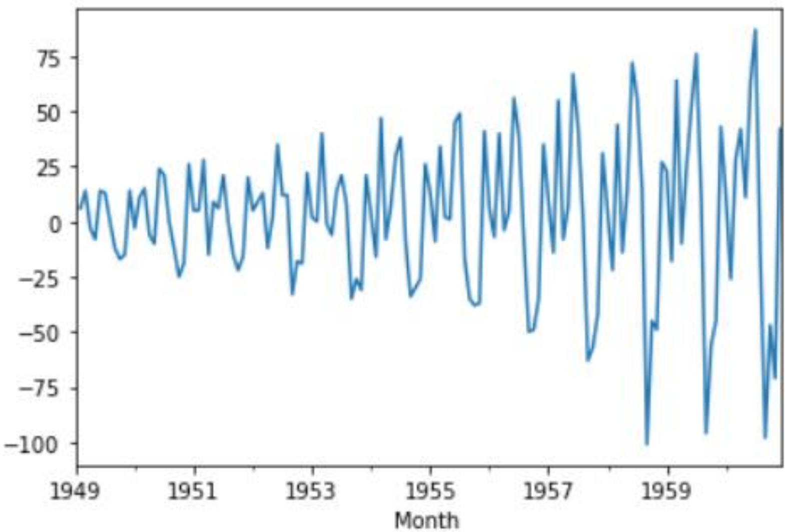


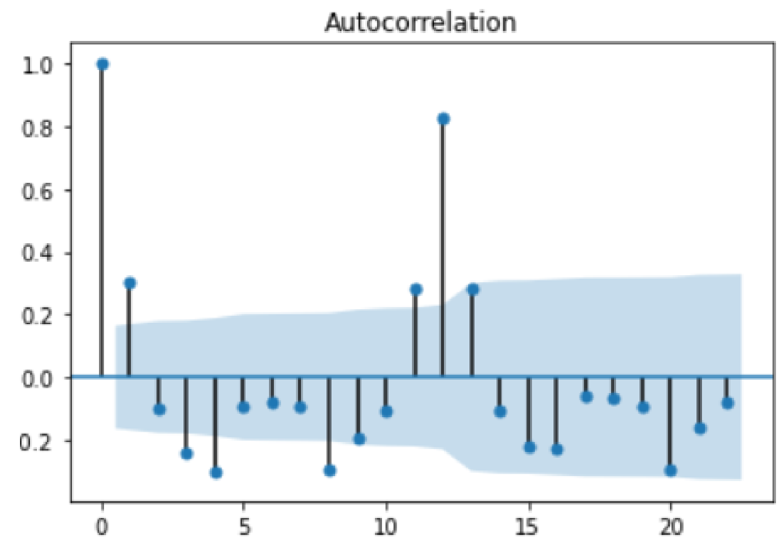

l=1




l=12

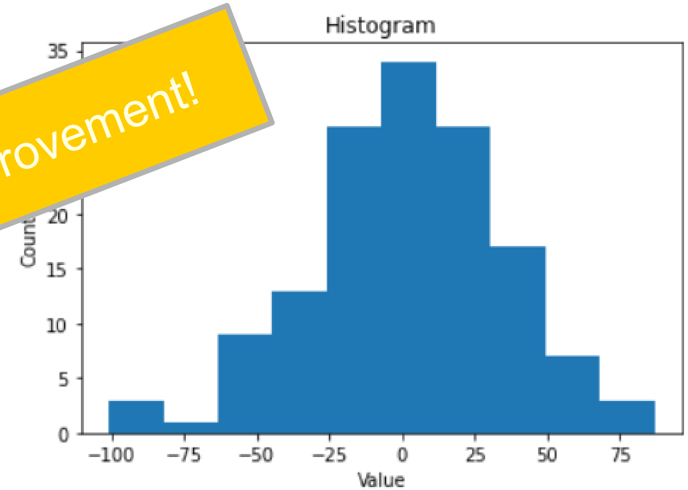
# Quality of Naive Passenger Model



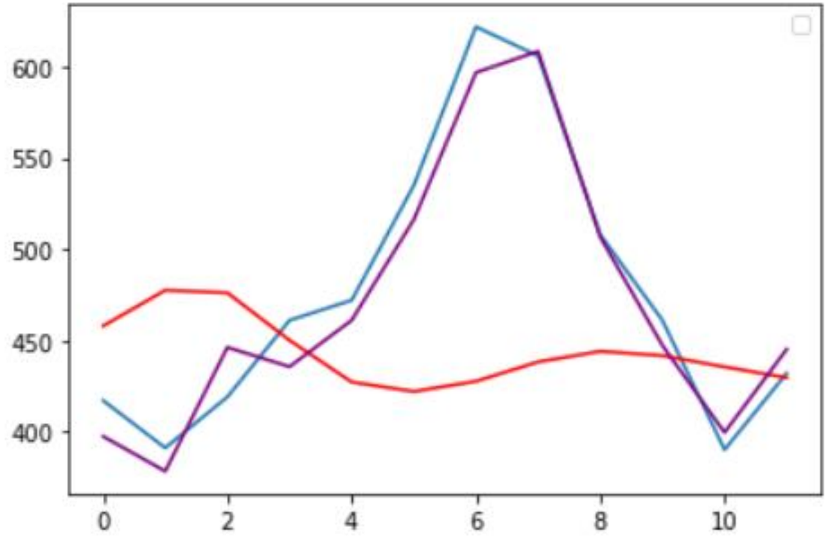

Mean: 2.237762237762238

There is obviously room for improvement!



# How about an AR Model?

- Original time series (blue) vs. one step prediction using
  - AR(6)(red) and
  - AR(1) (purple) with lag=12



# Advanced Machine Learning

## 7. Time Series Analytics

### Part Id - Basic Models for Forecasting ctd.

*Prof. Dr. Volker Herbort*

Technische Hochschule Ulm
University of Applied Sciences

# Autoregressive Moving Average (ARMA)

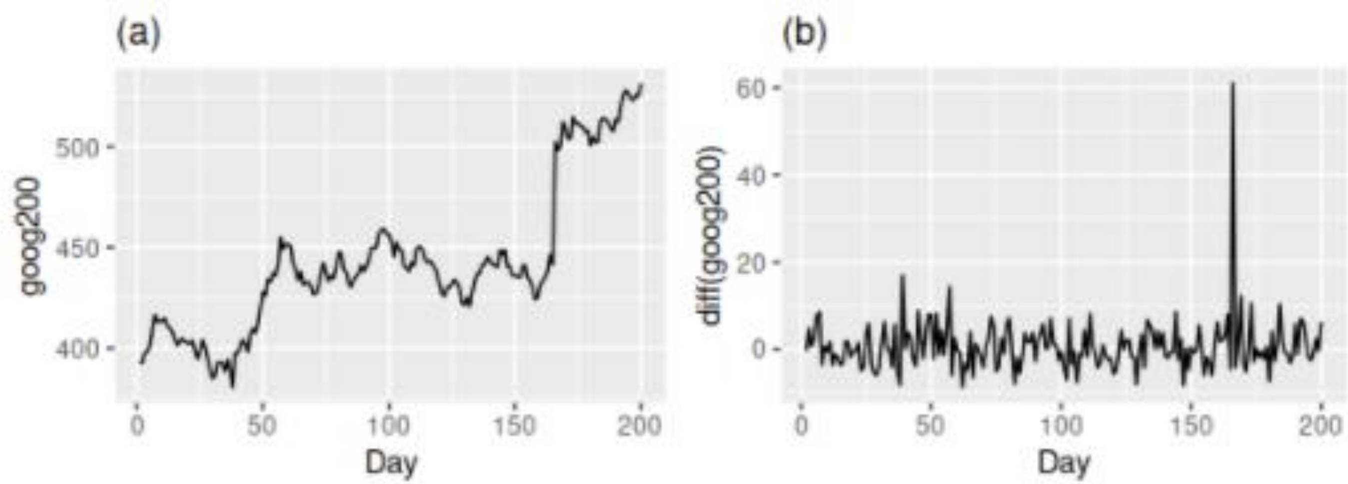
- Combination of both techniques
  - AR: regressing of the series with its own lagged version
  - MA: modeling the value as a linear combination of previous error terms

$$y_t = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}$$

  - A.k.a. **ARMA**(**p**,**q**) Models
- Disadvantage=> Only for stationary time series

# Autoregressive Integrated Moving Average (ARIMA)

- Used for non-stationary time series with trend
  - Idea: integrate the detrending in the model




Google stock price Daily change in Google stock price

  - By differencing it becomes stationary

Source: Forecasting: Principles and Practice,Rob J Hyndman and George Athanasopoulos,Monash University, Australia

# Differencing

- Compute the change between series and a lagged series

$$y'_t = y_t - y_{t-l}$$

  - 1st Order

$$y'_t = y_t - y_{t-1}$$

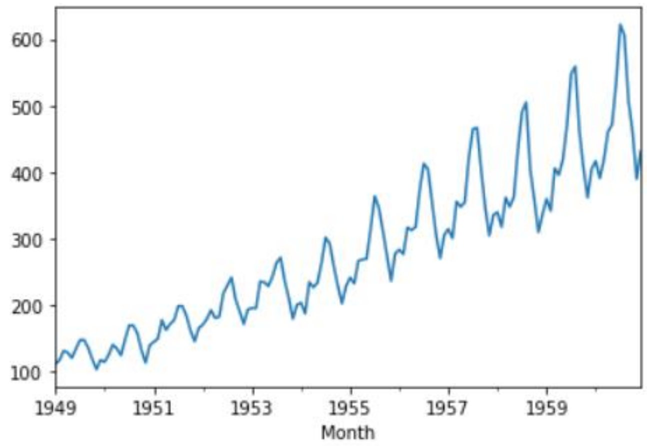
  - 2nd Order

$$y''_t = y'_t - y'_{t-1}$$

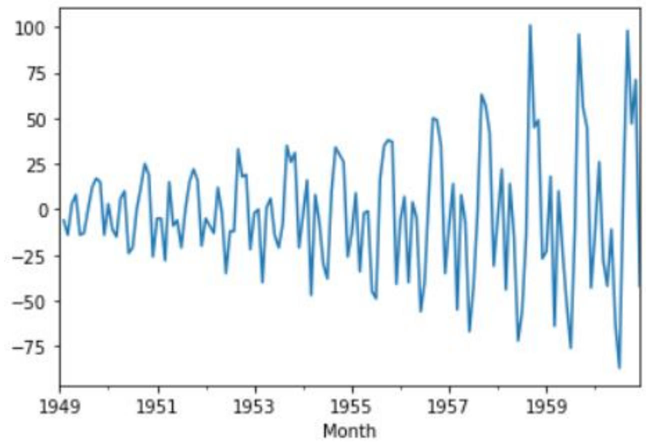
  - Can be used for removal of trend and/or seasonality

# Differencing for Trend Removal

- Using 1st order with lag=1 differencing on airplane passengers




$y_t$

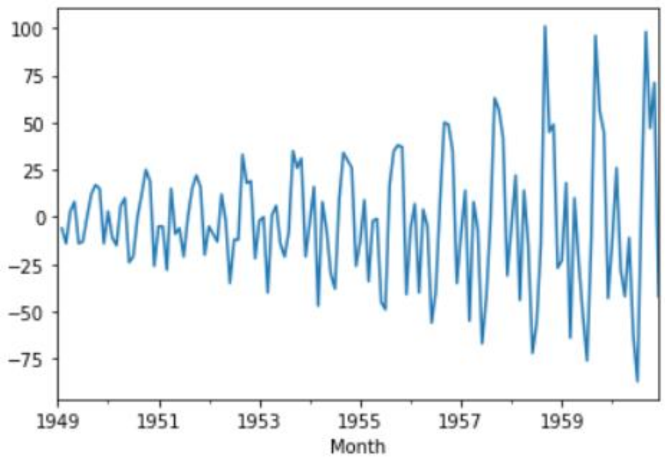



$y_t^T$

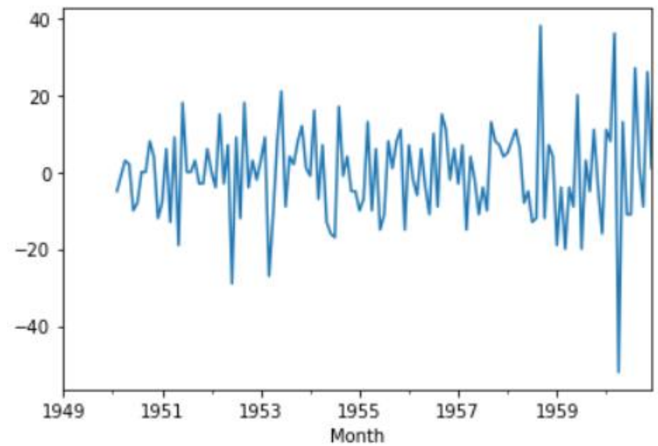
- Observation:
  - In $y_t^T = y_t'$ trend is gone, but saisonality still remains

# Differencing for Season Removal

- Using 1st order with lag=12 differencing on $y_t^T$




$y_t^T$




$y_t^S$

- Observation:
  - In $y_t^S = y_t''$ seasonality is gone only white noise remains

=> stationary time series

# ARIMA

- Non-seasonal ARMA on differenced series

$$y'_t = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i y'_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}$$

Where $y'_t$ can be differenced more than once…

- This is called an ***ARIMA(p,d,q) model*** with
  - p = order of AR part
  - d = degree of first differencing
  - q = order of MA part

# ARIMA ctd.

- Above models can be described using ARIMA

| Model | ARIMA(p,d,q) |
|---|---|
| White noise | (0,0,0) |
| Random walk | (0,1,0) |
| Autoregression | (p,0,0) |
| Moving Average | (0,0,q) |

# Estimation / Model Order

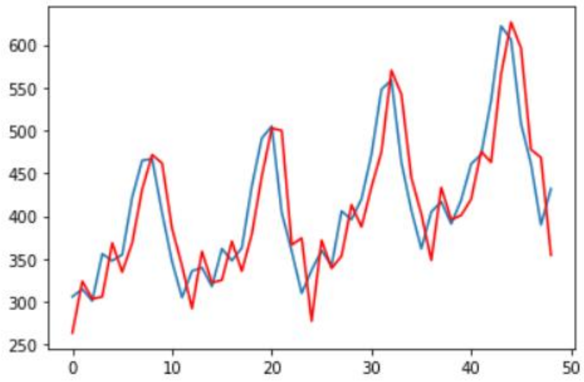
- Maximum Likelihood Estimation (MLE)
  - Used to find the paramters $\alpha_i$ and $\beta_i$
  - Values are computed based on order of model (p, d and q)
  - Idea: How likely is it to obtain the values in series t using this model order and parameters?

$$\sum_{t=1}^{T} \varepsilon_t^2$$

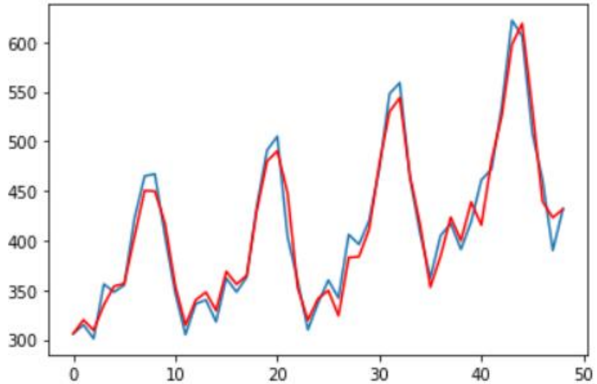
=> Basically a mean squared estimator for regression.

# ARIMA Airline Passengers

- ARIMA(1,1,1)
  - Mean Squared Error: 1988,75




- ARIMA(12,1,12)
  - Mean Squared Error: 257,18



# Seasonal ARIMA (SARIMA)

- ARIMA Model with **seasonal components**
  - $SARIMA(p,d,q)(\boldsymbol{P,D,Q})_m$ model with
    - m = number of observations per year
    - P = order of AR part
    - D = degree of first differencing
    - Q = order of MA part
  - Seasonal terms are simply multiplied by the non-seasonal terms.

# Seasonal ARIMA (SARIMA)

- Seasonal components can be seen in lags of PACF/ACF

  - $ARIMA(0,0,0)(0,0,1)_{12}$
    - a spike at lag 12 in the ACF but no other significant spikes;
    - exponential decay in the seasonal lags of the PACF (i.e., at lags 12, 24, 36, …).

  - $ARIMA(0,0,0)(1,0,0)_{12}$
    - exponential decay in the seasonal lags of the ACF;
    - a single significant spike at lag 12 in the PACF.

# References


- [1] https://en.wikipedia.org/wiki/Time_series