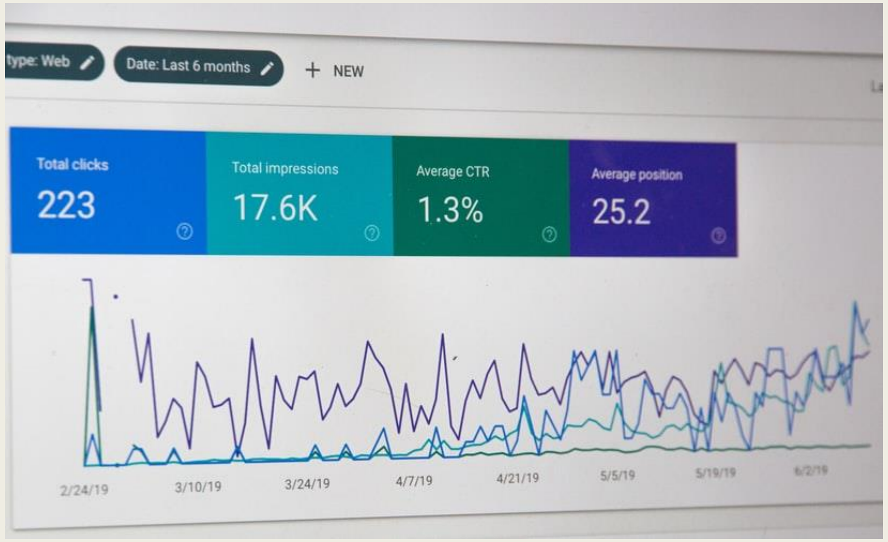
TIME SERIES ANALYSIS

Photo by Riho Kroll on Unsplash

Time series

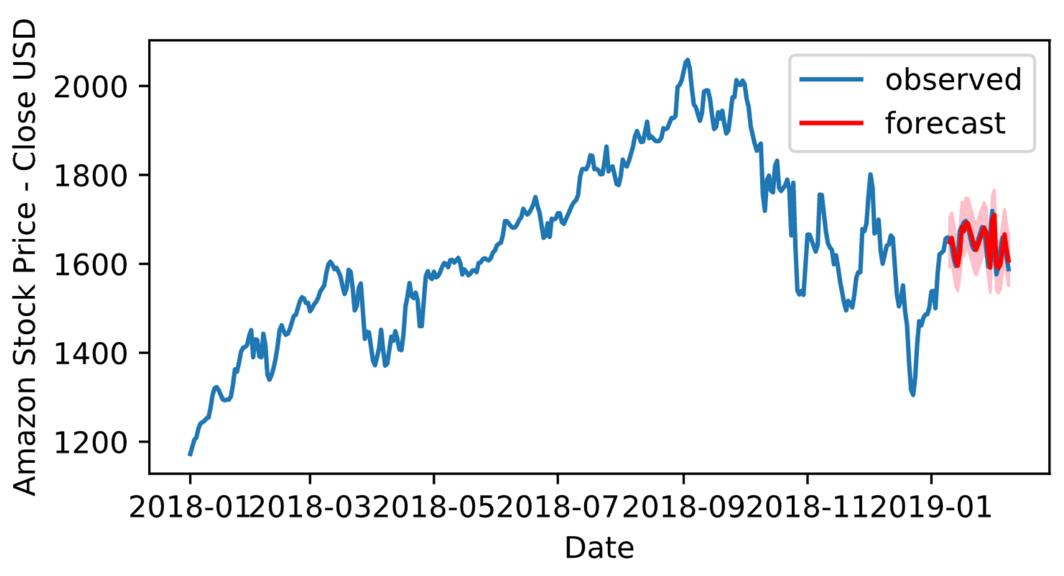
- A time series is a series of data points indexed (or listed or graphed) in time order (usually at successive equally spaced points in time):

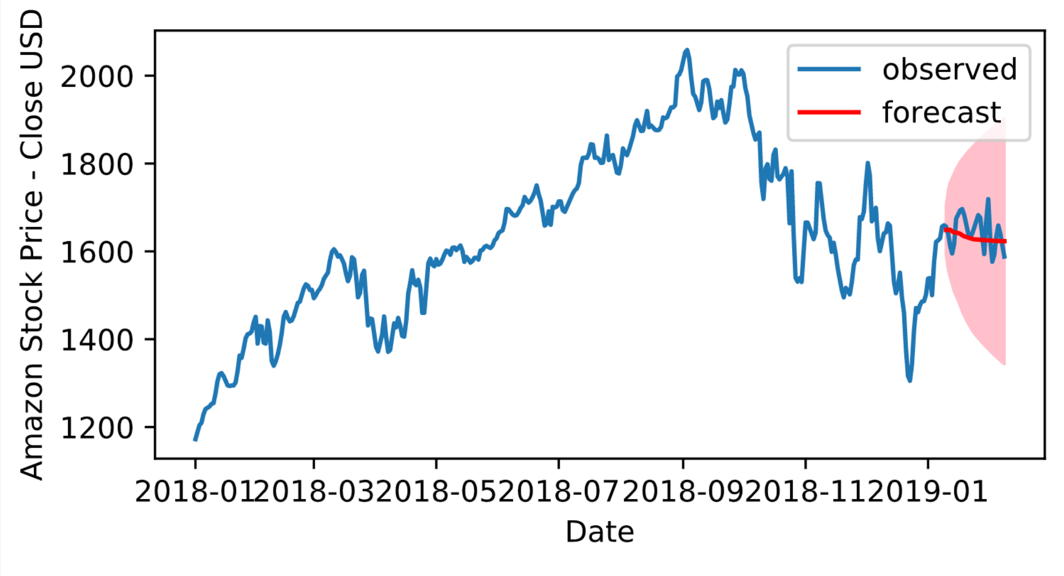
$$y_1, y_2, \dots, y_t, \dots, \qquad y_t \in R$$



Goal

- Identify and evaluate the model of the process under study,
- Predict future values for one step ahead and dynamic



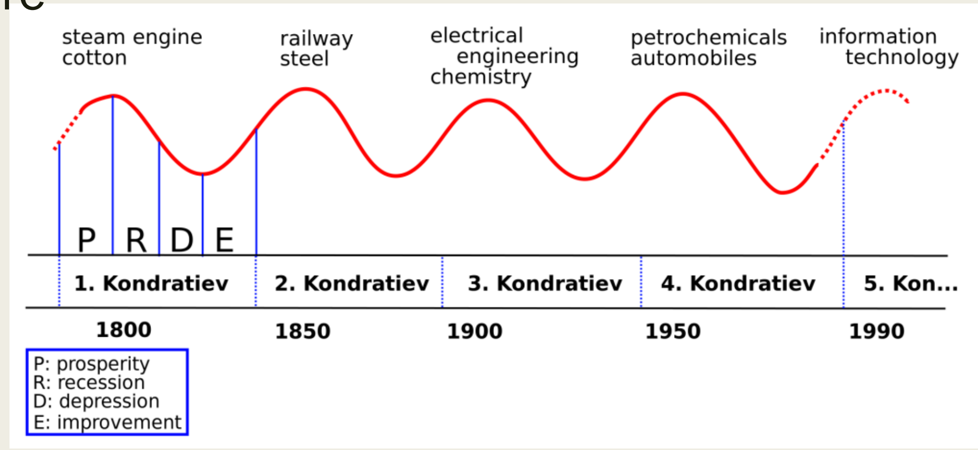


Time series decomposition

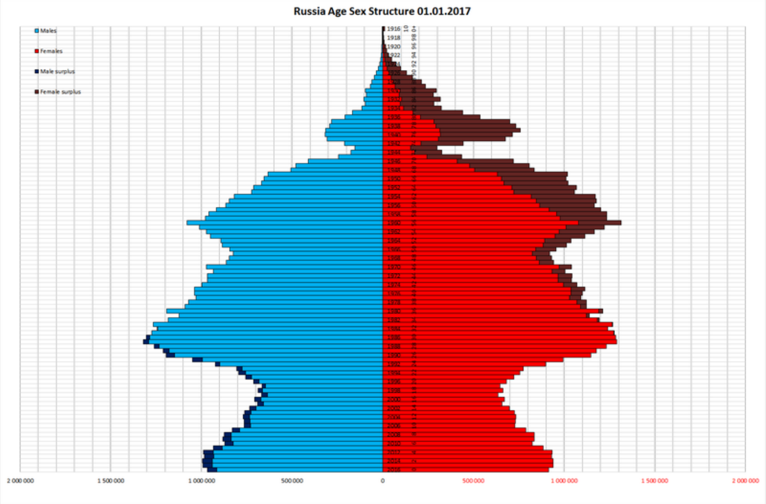
- Usually the following components of the time series are considered:

$$y_t = u_t + v_t(+c_t) + \varepsilon_t,$$

- where u_t - a trend that describes the impact of long-term factors and long-term trends (like economic and population growth),
- v_t - seasonal component that reflects the cyclicity of economic processes over a short period of time (changes in sales volumes at different times of the year),
- c_t - a cyclical component that reflects the repeatability of economic processes over long periods (Kondratiev waves of economic activity or demographic “pits”),
- ε_t - a random component that reflects the influence of random factors that cannot be accounted for or registered



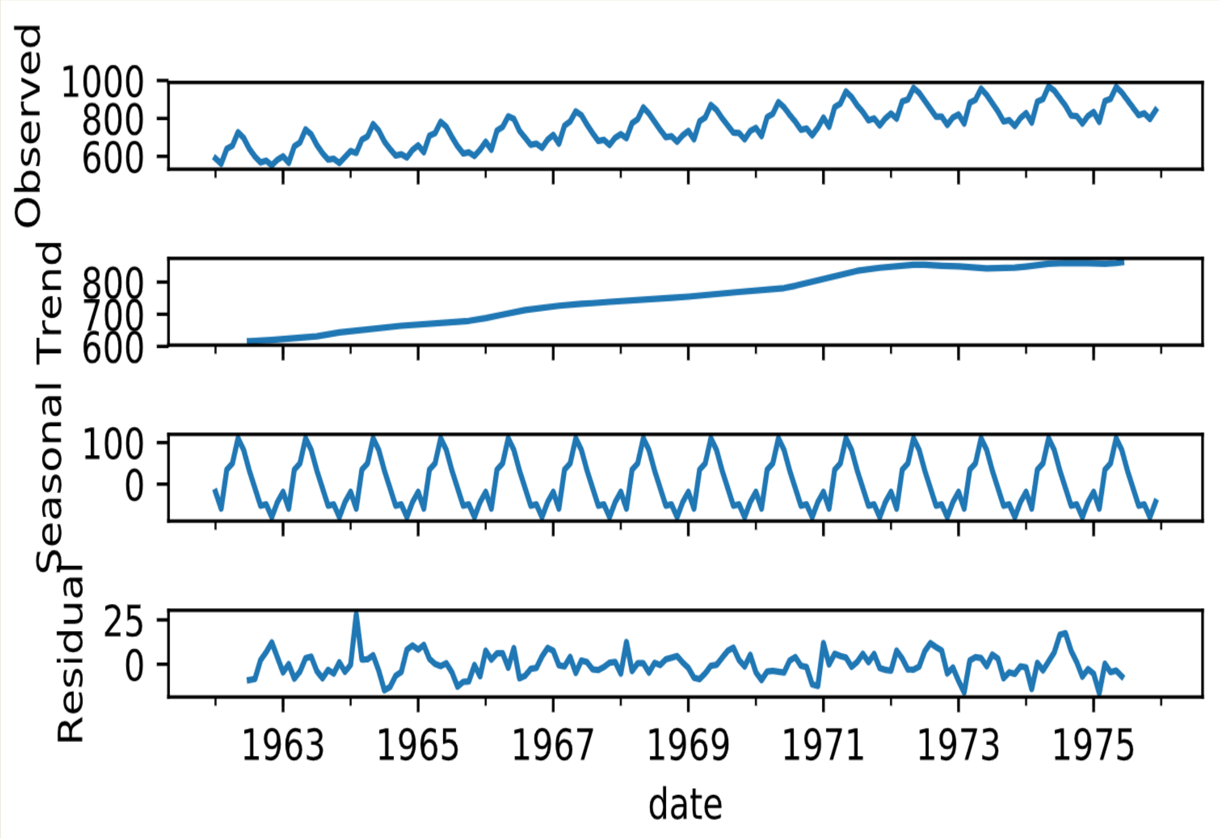
Made by Rursus, source: wikipedia

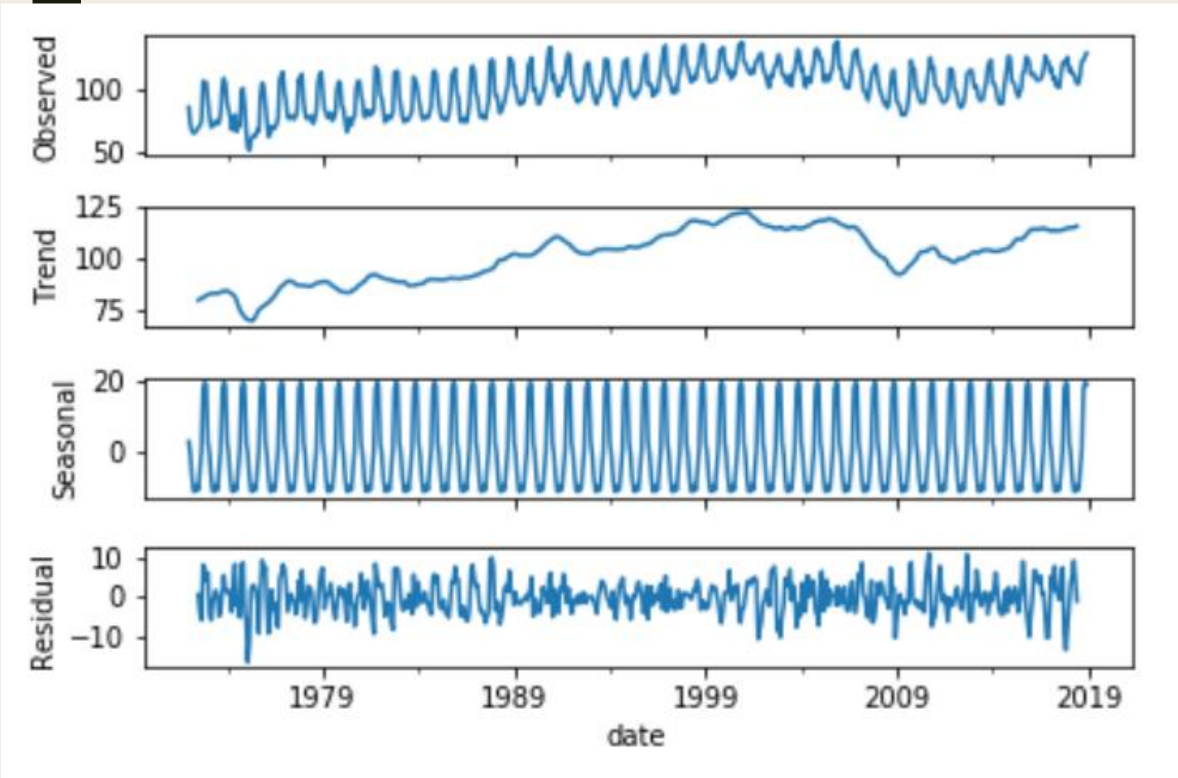


Made by Rickky1409, source: wikipedia

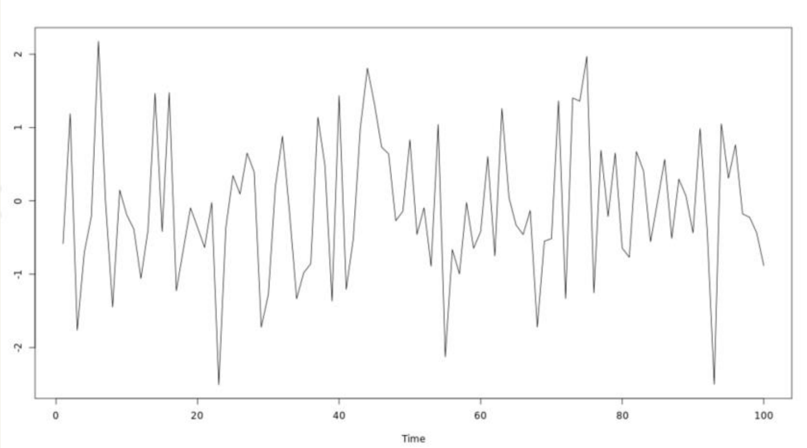
Time series decomposition

- $y_t = u_t + v_t(+c_t) + \varepsilon_t$,
- u_t - trend,
- v_t - seasonal component,
- ε_t - random component.



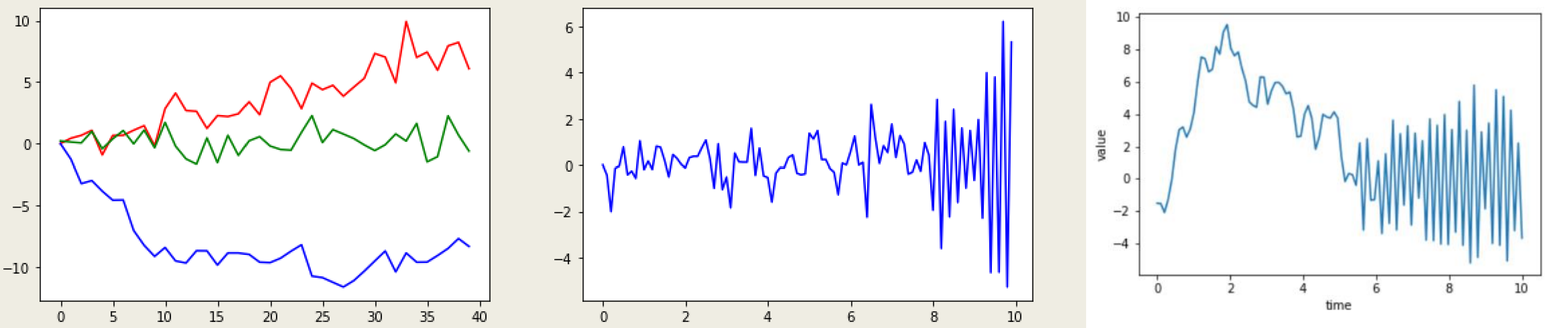


Stationary time series



Example of the stationary time series

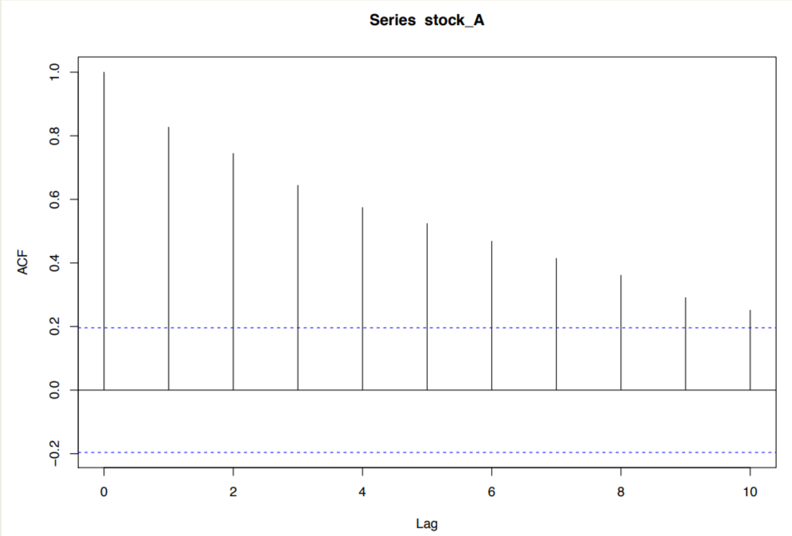
- Stationary time series are those whose probabilistic properties do not change over time :
  - *mean, variance, and autocovariance/autocorrelation are independent of time :* $M(y_t) = const, D(y_t) = const, cov(y_t, y_s) = K(t-s)$ *(covariance/correlation depends only on the time difference* $t-s$*)*

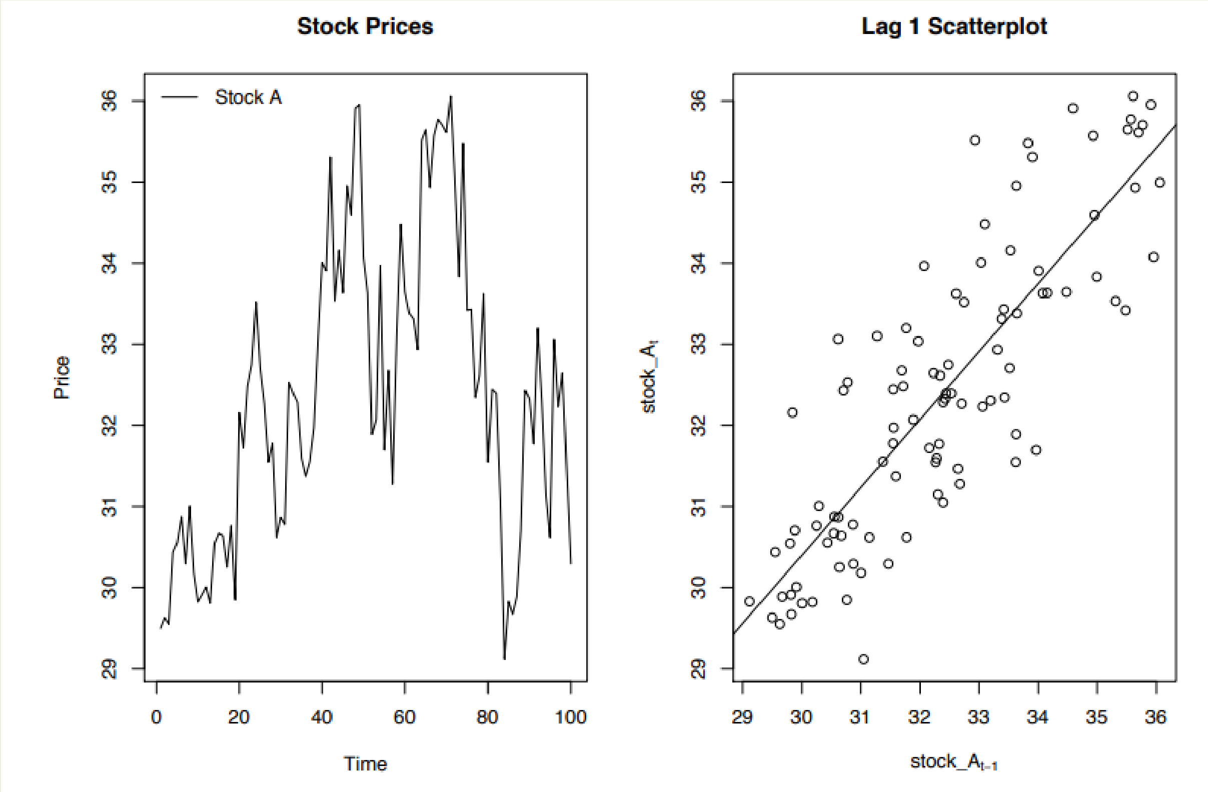


Examples of non-stationary time series

Autocorrelation function(ACF)

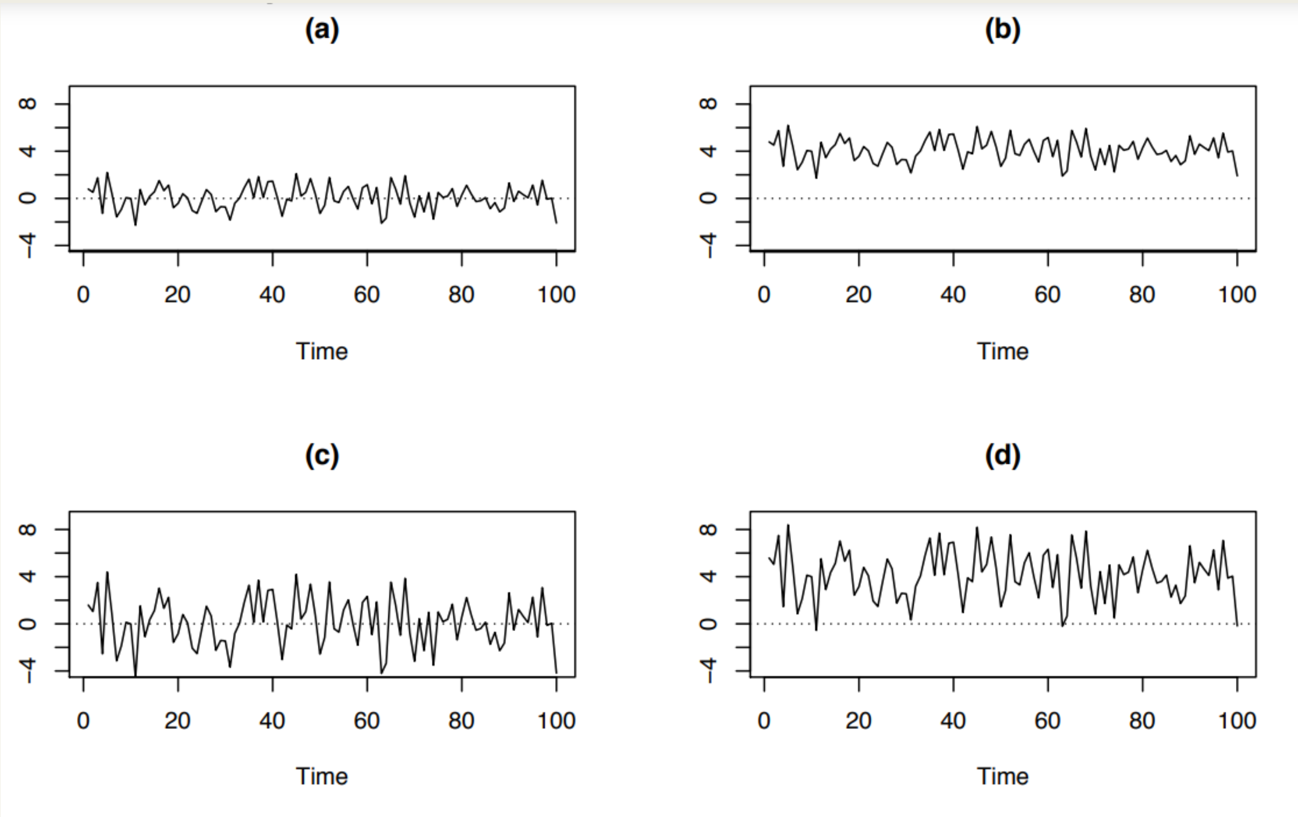
- Autocorrelation with a lag of 1: compare the values of the time series at the current moment y_t with values shifted one step (lag) back
- We calculate the correlation = we estimate the degree of linear dependence of today's values on yesterday's values
- Similarly, you can calculate autocorrelation with a lag n





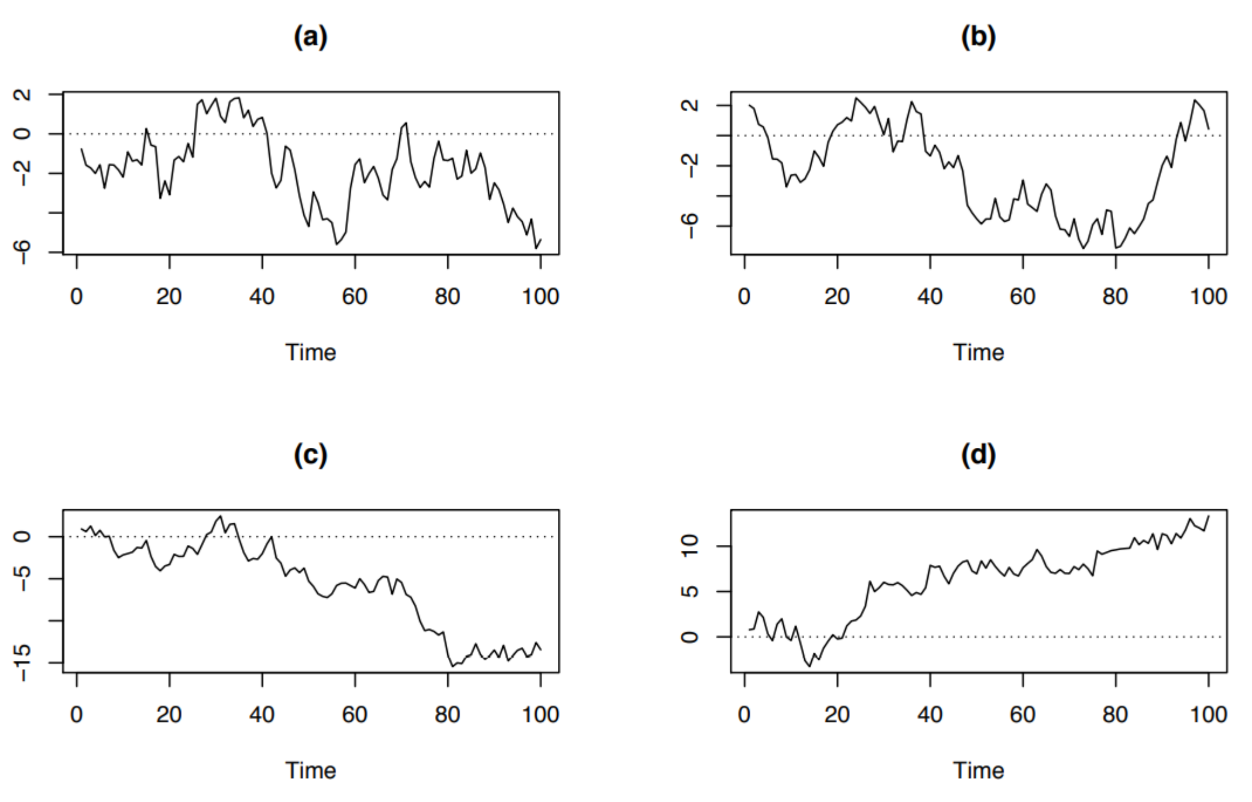
White noise

- The simplest example of a stationary series is white noise: $y_t = \varepsilon_t$ (mean and variance are constant, covariance is 0)



Random walk

- This is the simplest example of a non-stationary time series : $y_t = y_{t-1} + \varepsilon_t$
  - *no average value or variance,*
  - *high correlation,*
  - ε_t *is a white noise with mean 0*

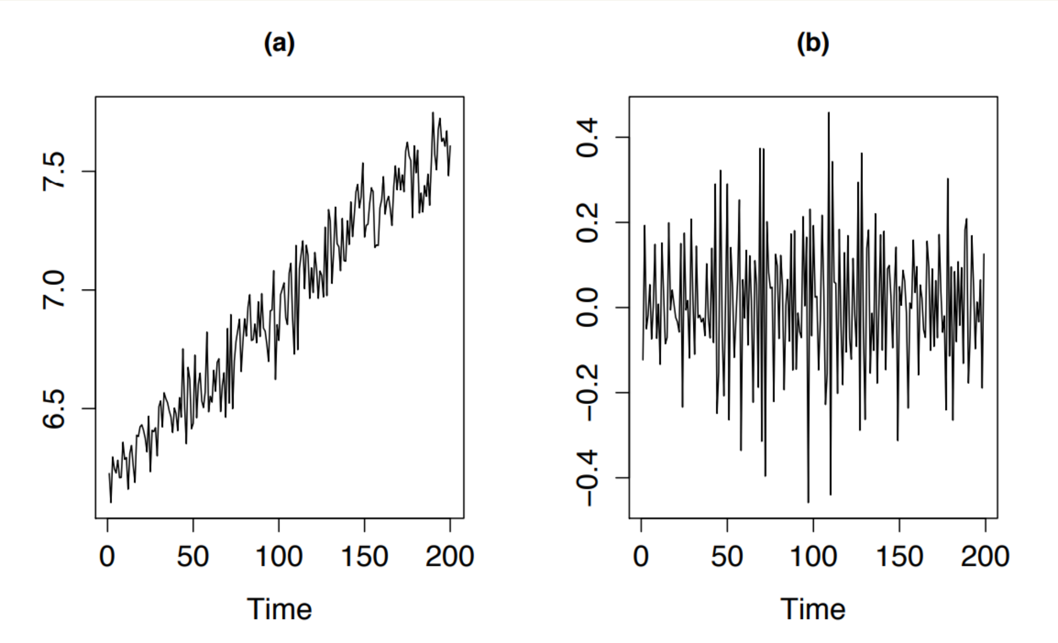


How can we check the stationarity?

- Take a look at the picture!
- Augmented Dickey–Fuller test:
- The null hypothesis: the time series is non-stationary and a unit root is present in a time series
- The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity.
- Test statistics (DF-statistics) similar to the usual t-statistics for checking the significance of linear regression coefficients $y_t = ay_{t-1} + \varepsilon_t$ and H_0: $a = 1$, H_1: $|a| < 1$

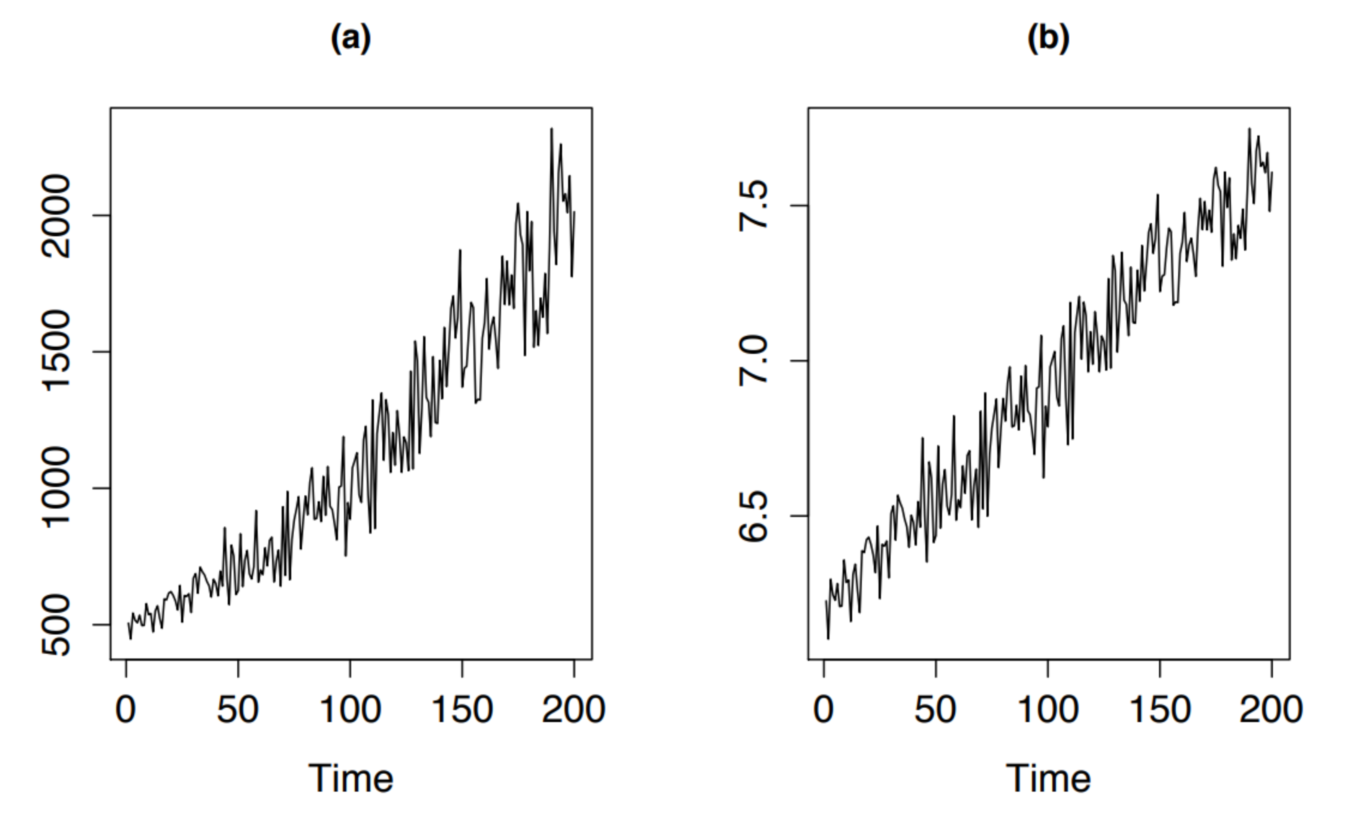
Transformation to stationary series

- If trend is linear, then usually we use differences: $\Delta y_t = y_t - y_{t-1}$,
- Transformed series of differences Δy_t should be checked for stationarity by Dickey–Fuller test
- If not, we can take differences of differences, and again differences of differences, …



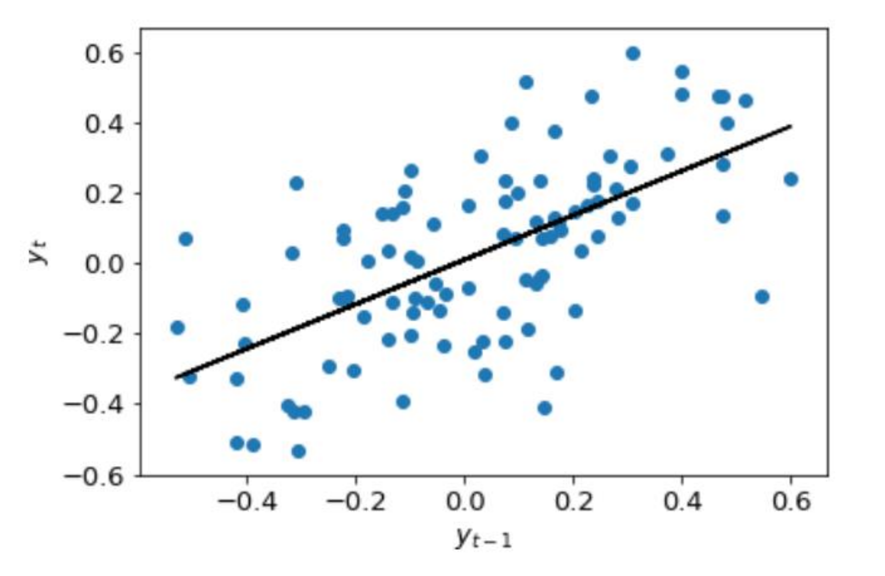
Transformation to stationary series

- Other types of transformation also can be used: $z_t = \ln(y_t)$, $z_t = \sqrt{y_t}$, …



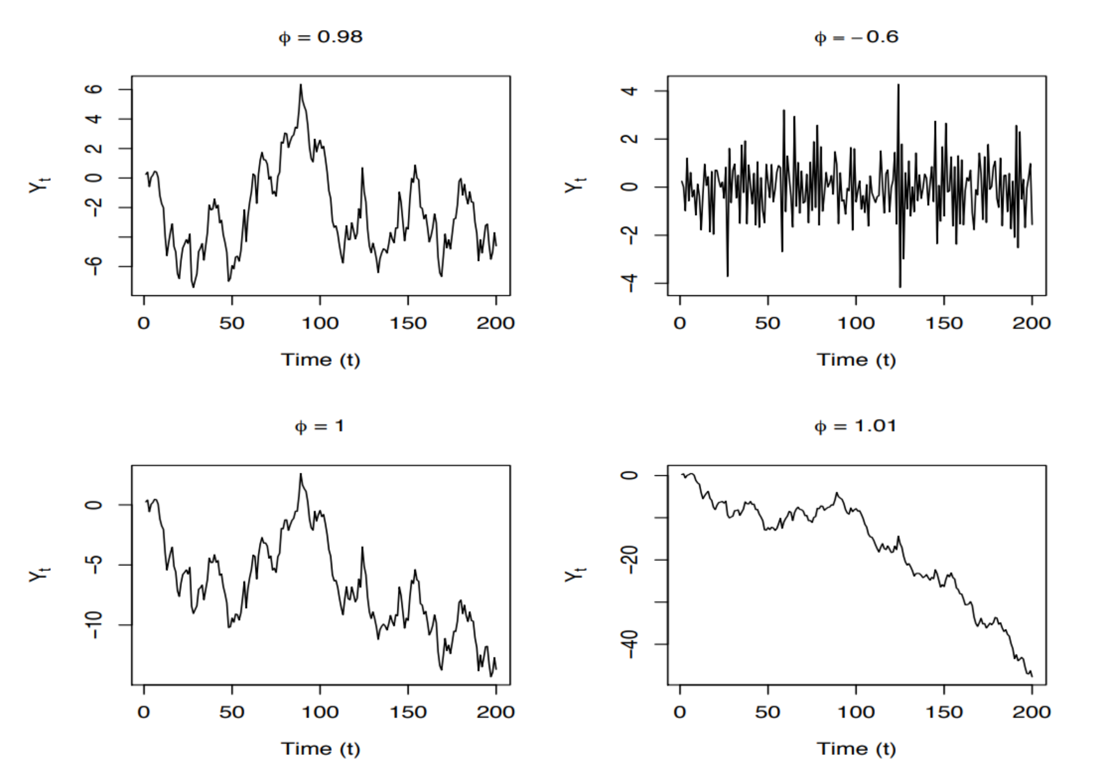
Autoregressive model (AR)

- Simplest autoregressive model AR(1): $y_t = a_1 y_{t-1} + \varepsilon_t$, where ε_t is a white noise
- Order of the AR model = number of lags
- AR(2): $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t$
- AR(p): $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \varepsilon_t$

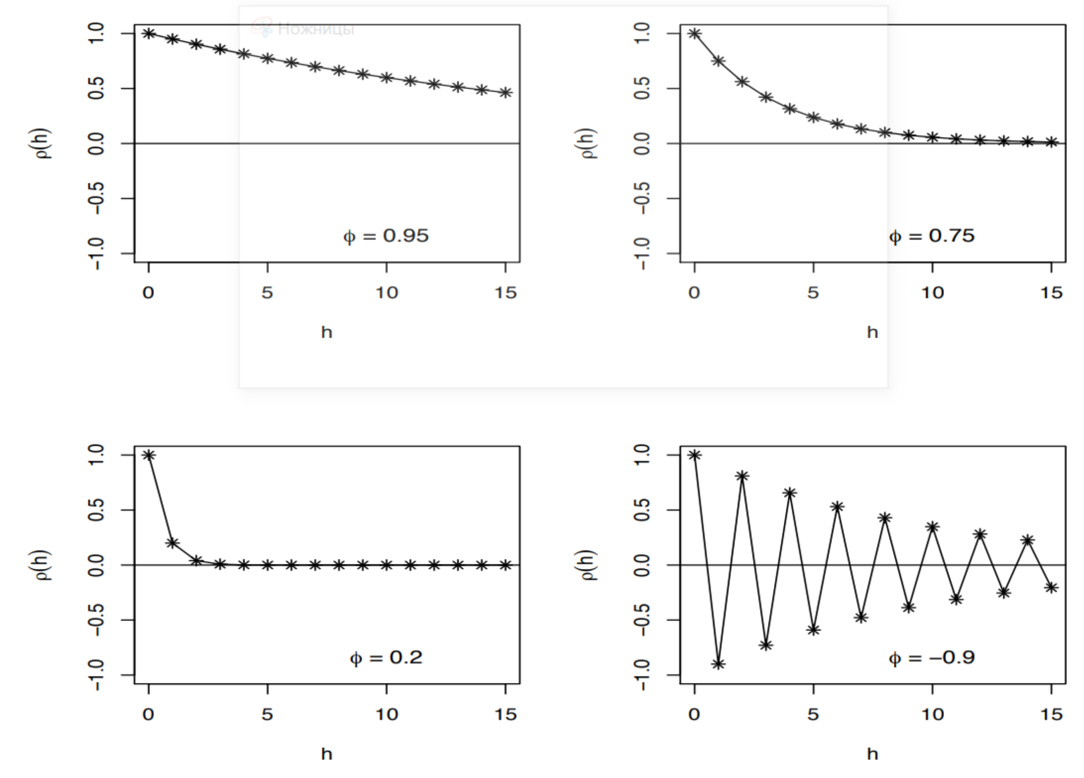


Autoregressive model (AR)

- Simplest autoregressive model AR(1): $y_t = a_1 y_{t-1} + \varepsilon_t$, where ε_t is a white noise



Examples



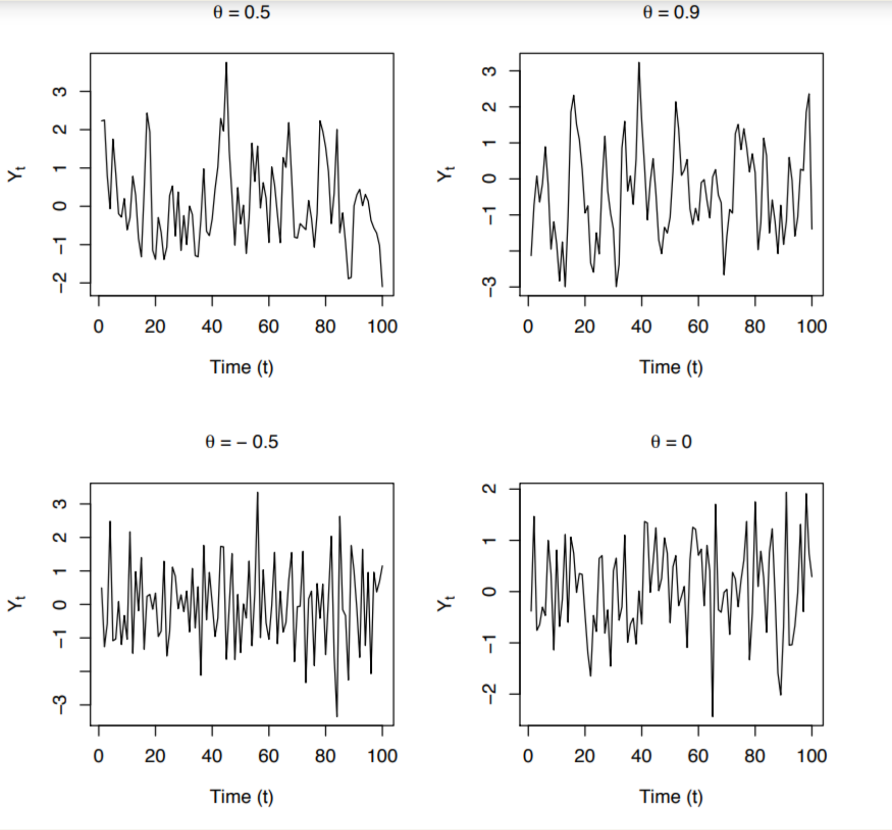
Examples of ACF for AR(1)

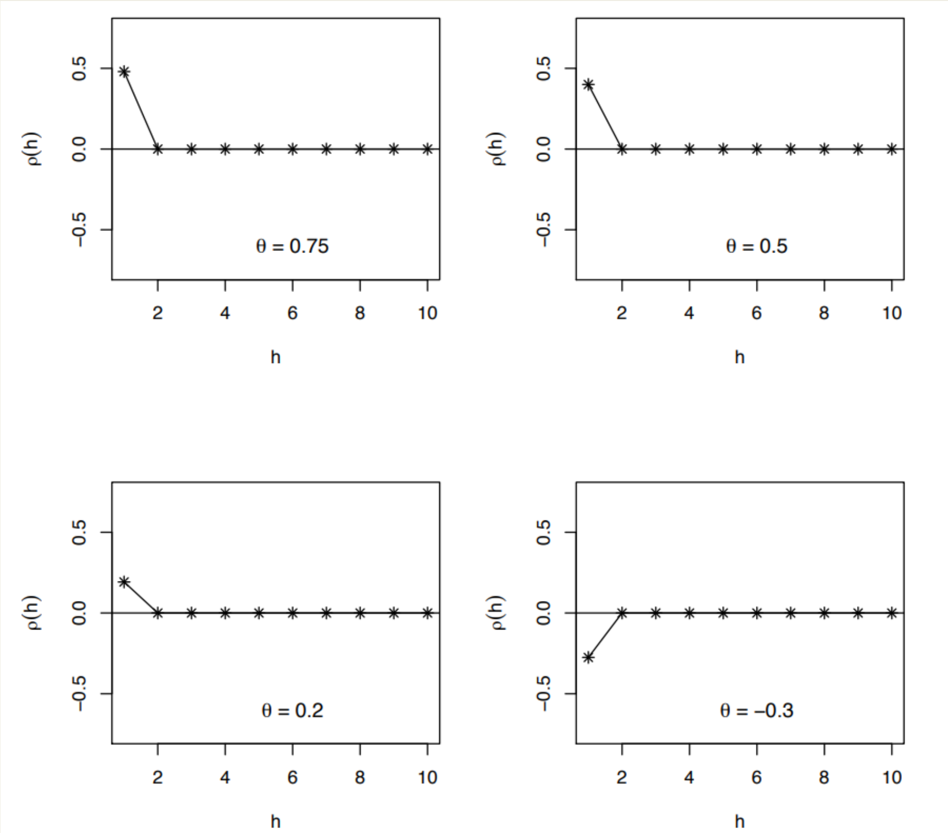
Moving Average Model (MA)

- Simplest model of moving average MA(1): $y_t = m_1 \varepsilon_{t-1} + \varepsilon_t$, where ε_t is a white noise
- Order of the MA model = number of lags
- MA(2): $y_t = m_1 \varepsilon_{t-1} + m_2 \varepsilon_{t-2} + \varepsilon_t$
- MA(q): $y_t = m_1 \varepsilon_{t-1} + m_2 \varepsilon_{t-2} + \cdots + m_q \varepsilon_{t-q} + \varepsilon_t$

Moving Average Model (MA)

- Simplest model of moving average MA(1): $y_t = m_1\varepsilon_{t-1} + \varepsilon_t$, where ε_t is a white noise





ARMA model

- ARMA (1,1): $y_t = a_1 y_{t-1} + m_1 \varepsilon_{t-1} + \varepsilon_t$
- ARMA (p,q): $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + m_1 \varepsilon_{t-1} + m_2 \varepsilon_{t-2} + \cdots + m_q \varepsilon_{t-q} + \varepsilon_t$

ARMAX model

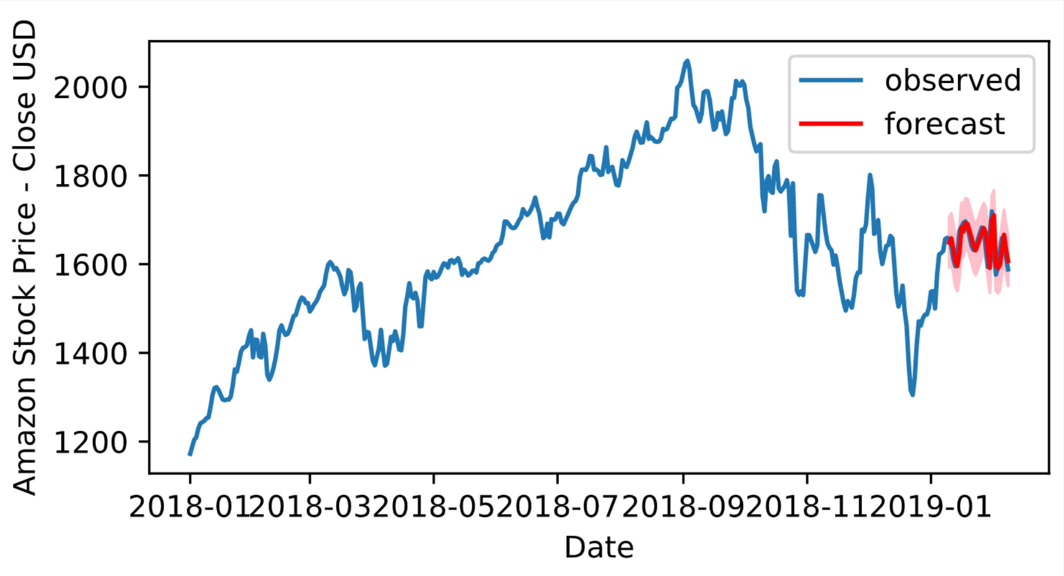
- We can use also exogenous variables like in usual regression
- Autoregressive–moving-average model with exogenous inputs model (ARMAX model)
- Examples: my productivity today might depend on my previous productivity and also on how may hours I slept today
- ARMAX (1,1): $y_t = b_1 x_t + a_1 y_{t-1} + m_1 \varepsilon_{t-1} + \varepsilon_t$

ARIMA model (Autoregressive Intergrated Moving Average)

- ARIMA (p,d,q) = we turn a non-stationary series into a stationary one by differentiating it d times in a row and apply the model ARMA (p,q)
- ARIMA (p,0,q) = ARMA (p,q)
- ARIMA (0,1,0): $y_t - y_{t-1} = \varepsilon_t$ is a white noise

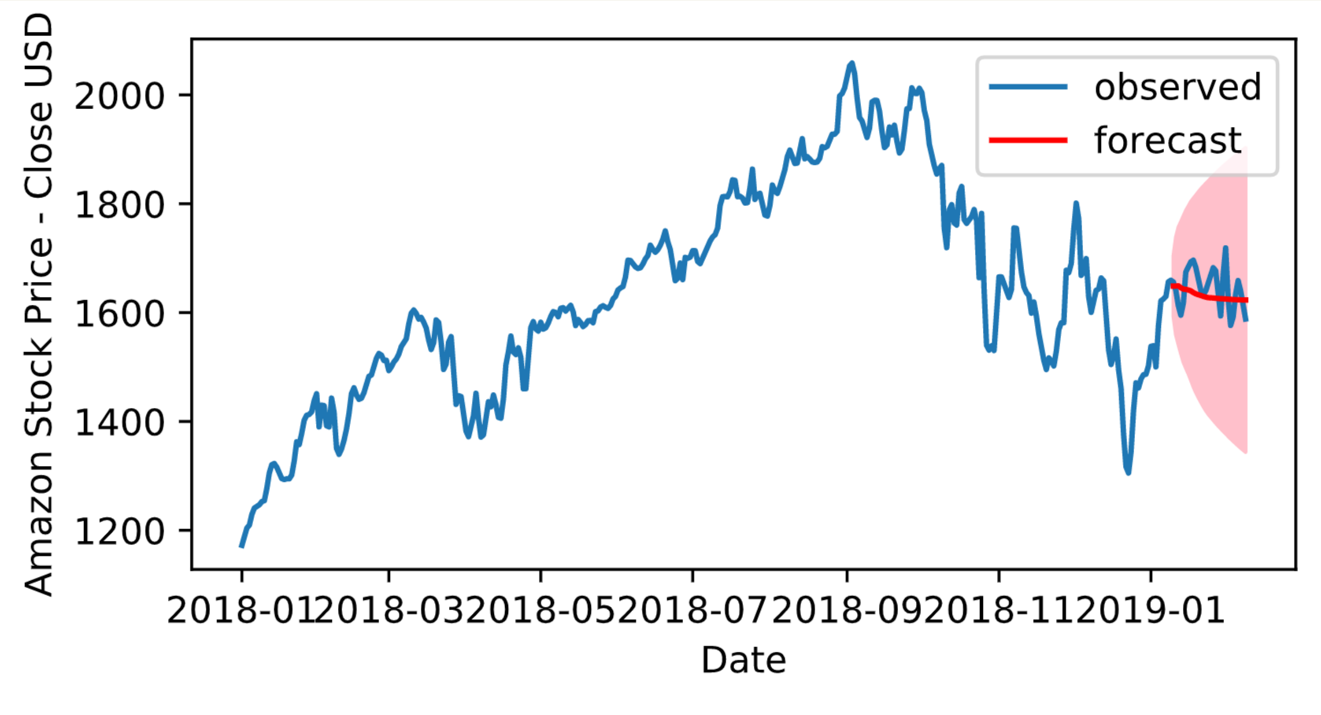
Forecasting

- One-step-ahead forecast: define the model, evaluate the next value and the standard deviation will describe the possible uncertainty

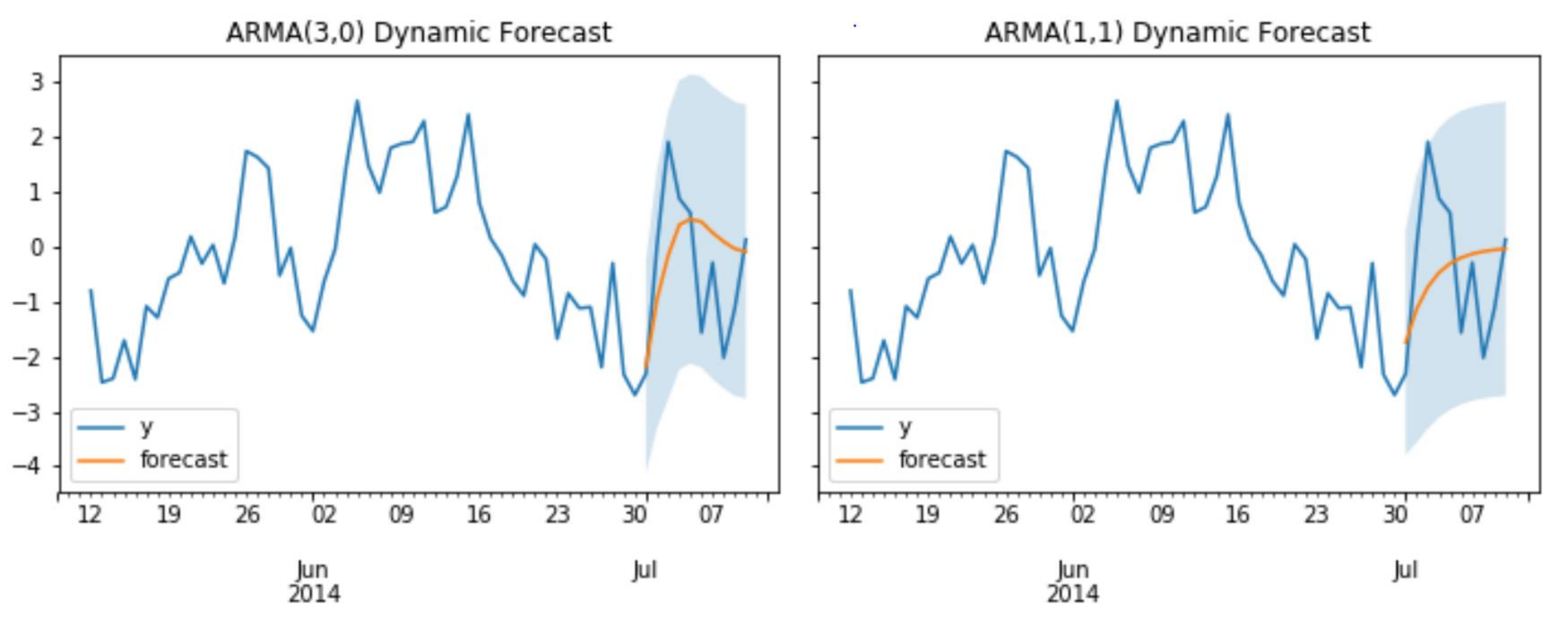


Forecasting

- Dynamic forecast– calculate the one-step-ahead forecast and use it as the final point to define the next one-step-ahead forecast

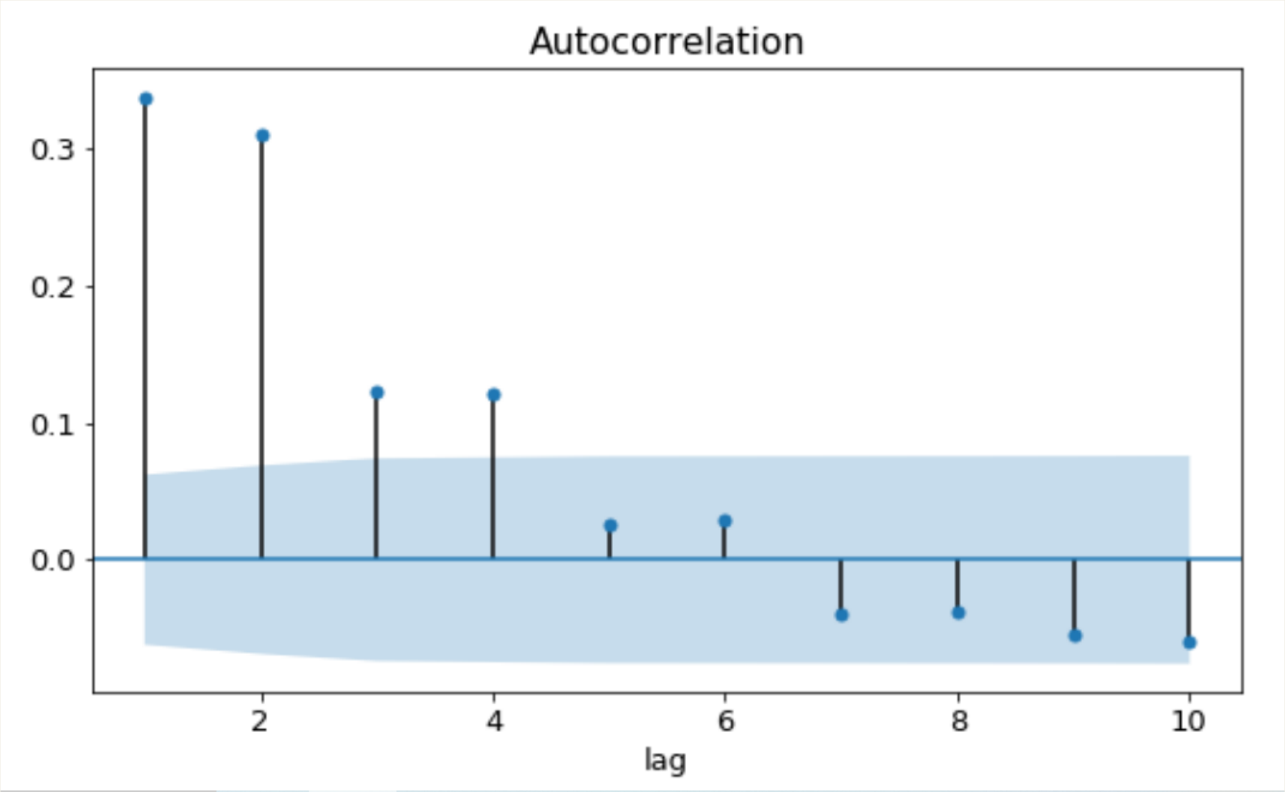


Good forecast = right model



How can we choose type of model?

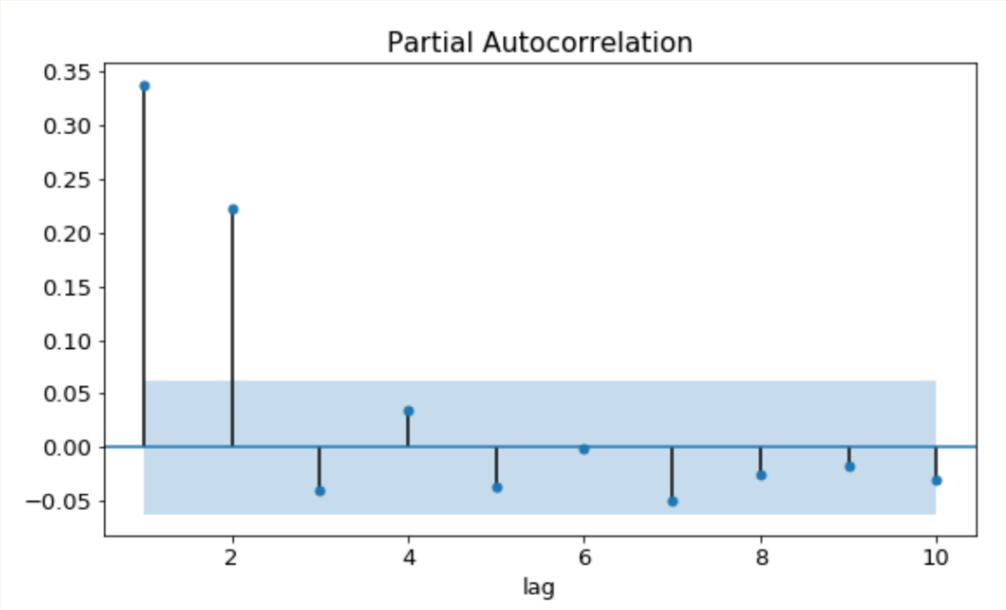
- The ACF and PACF function can give a hint!



How can we choose type of model?

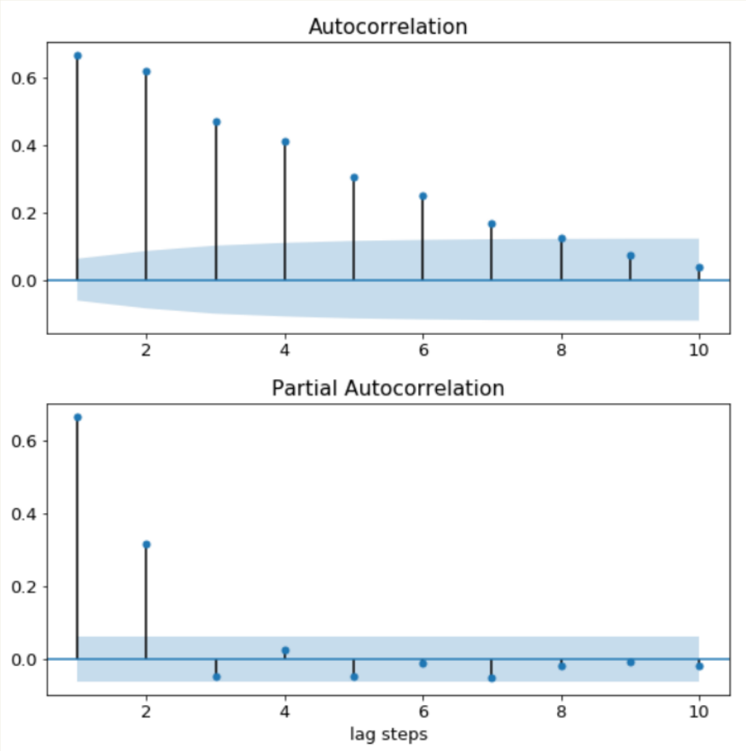
- The ACF and PACF function can give a hint!
- PACF(n) is the correlation between the values of the time series and the same values shifted n lags back, after the influence of all intermediate lags was excluded:

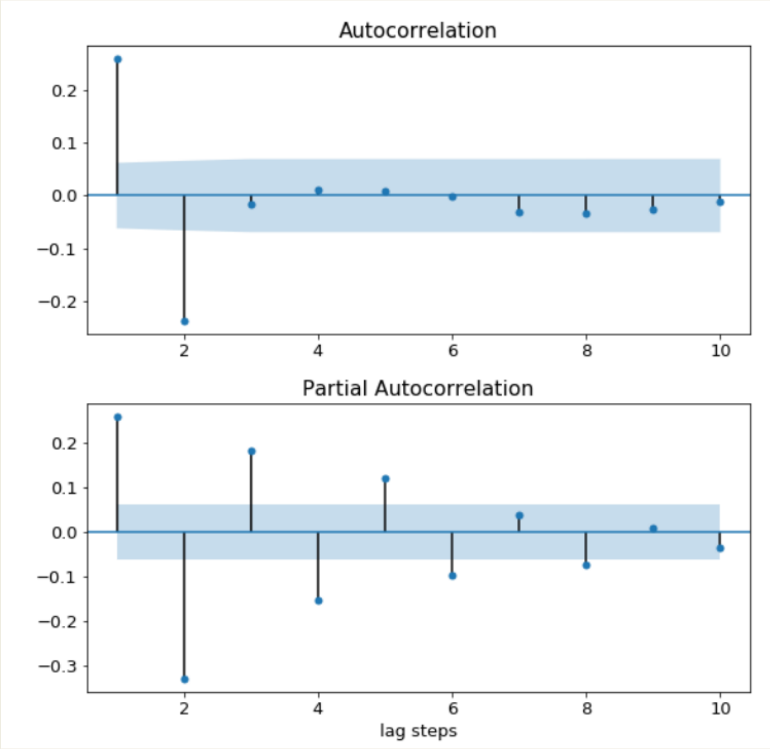
$$PACF(n) = \beta_n \text{ in regression } y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_n y_{t-n} + \varepsilon_t$$

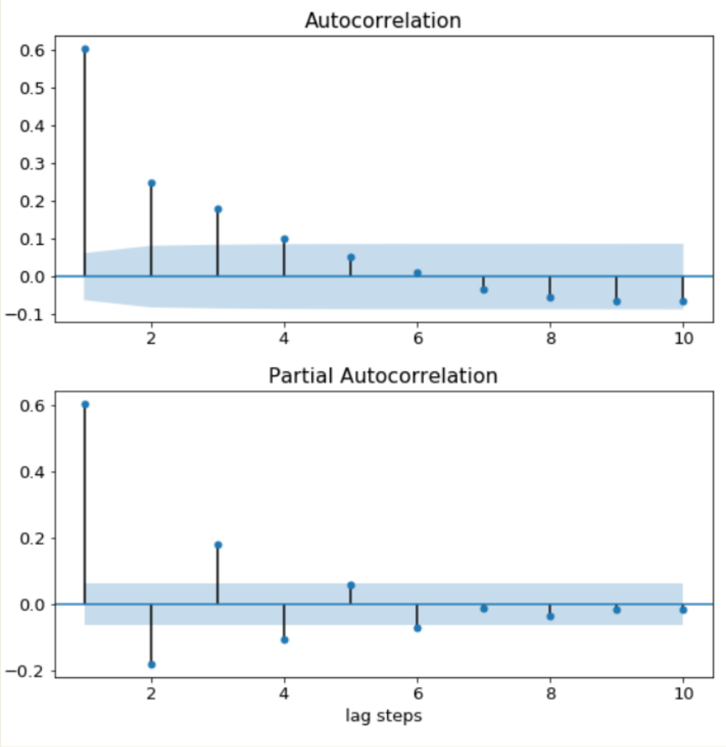


How can we choose the model?

AR(p)	MA(q)	ARMA(p,q)
ACF exponentially decrease	ACF =0 after lag q	ACF exponentially decrease
PACF =0 after lag p	PACF exponentially decrease	PACF exponentially decrease





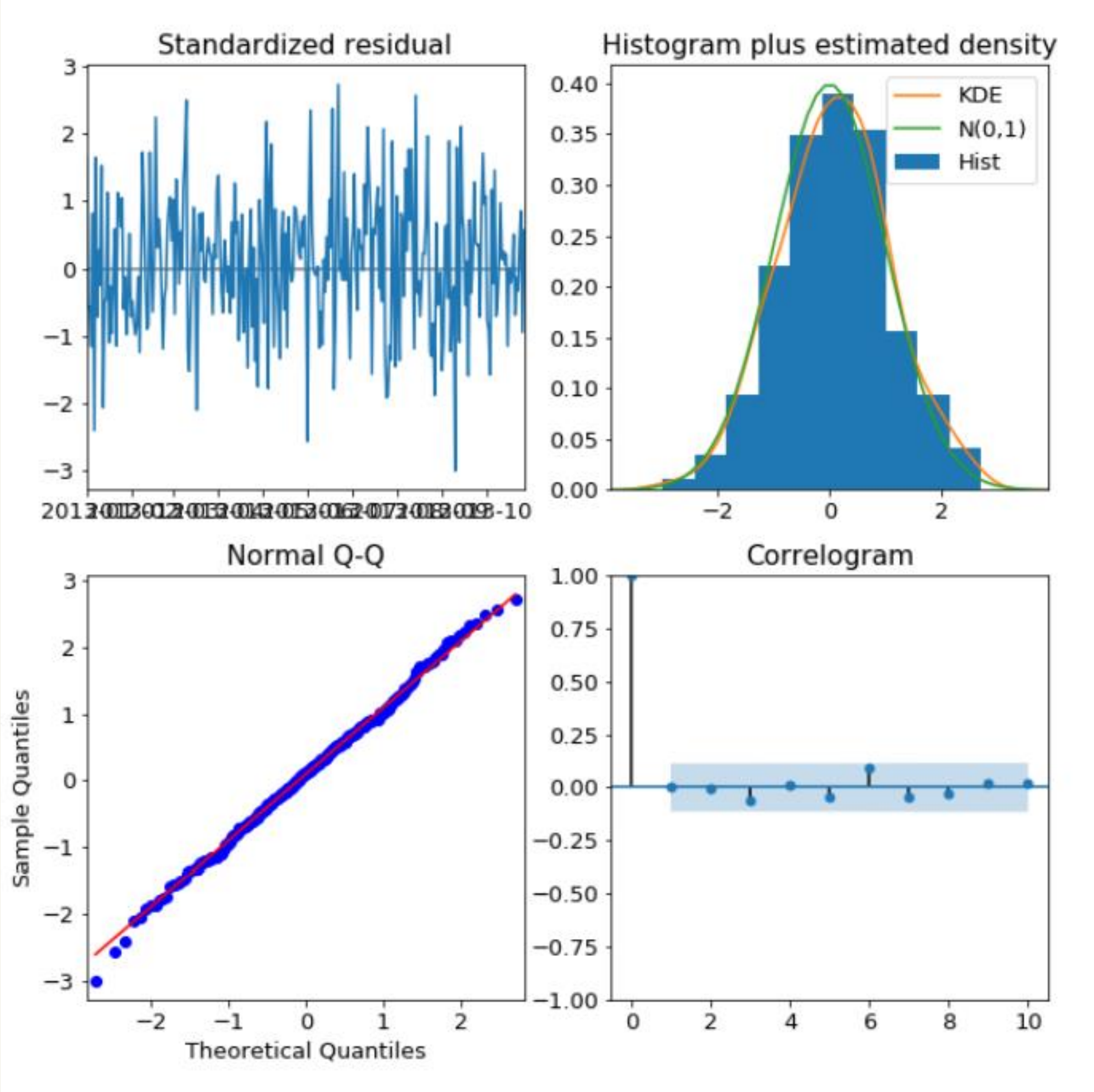


How can we choose type of model?

- Make a series of models with different orders p and q, fit models
- Use the Akaike (AIC) and Bayesian (BIC) information criteria and choose the model with the lowest value

Model diagnostics

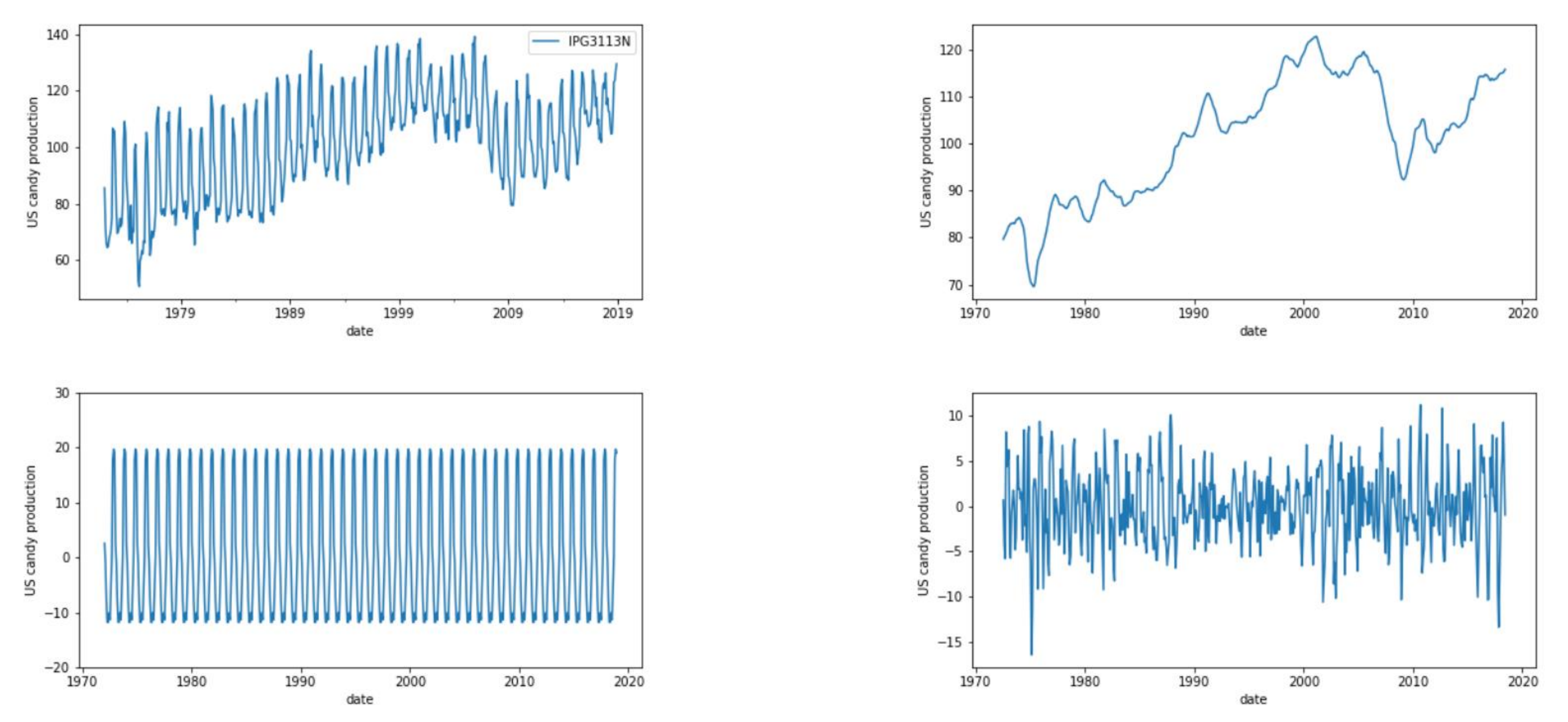
- Look at the residuals and check whether do they look like a white noise
- Compare the empirical distribution function for the residuals and the theoretical one for $N(0,1)$
- Check how much the quantiles of the empirical and theoretical distributions match at Q-Q plot
- Look at the correlogram (the ACF of residuals): all its values for positive lags must be equal to 0



Box–Jenkins method

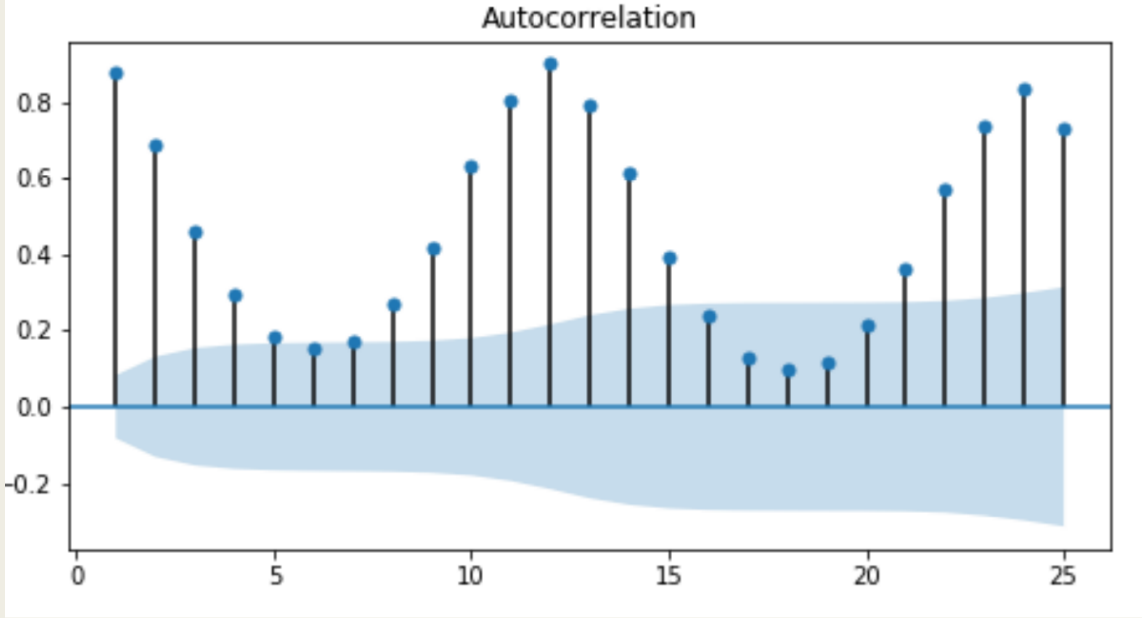
- Determine whether the time series is stationary and make differencing to achieve stationarity if needed
- Identify the order (i.e. the p and q) of the autoregressive and moving average terms
- Find the estimates of the model
- Check the goodness-of-fit and find the most appropriate model
- Make the model diagnostics
- Iterate steps 2-5 if needed
- Make a prediction

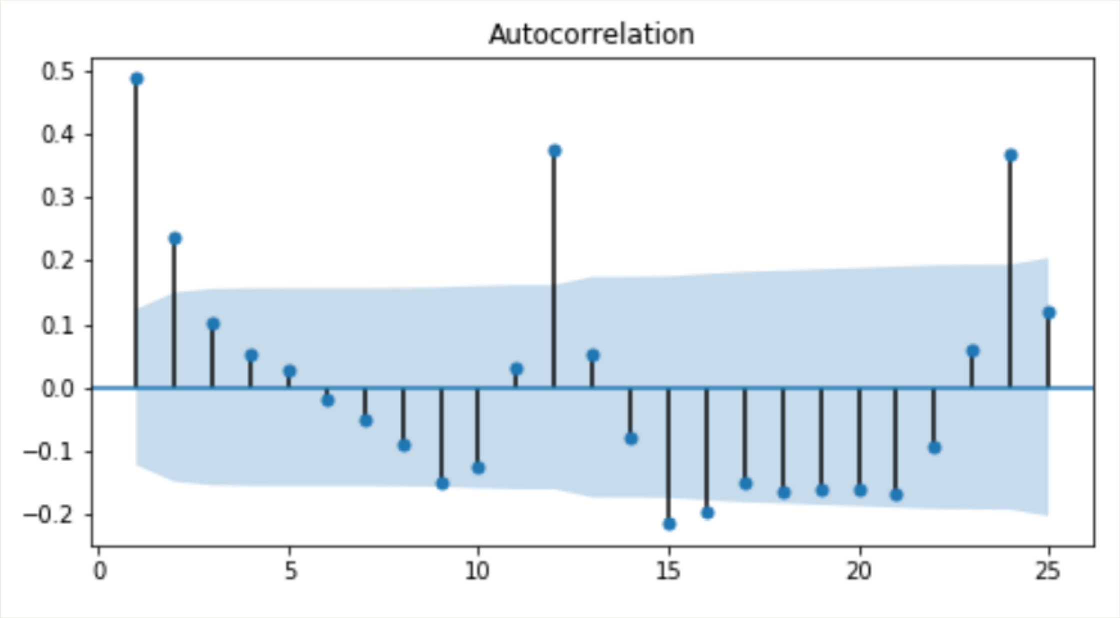
Seasonality



Seasonality

- Also can be found in the ACF
- The lag with the largest value of ACF (except the 1 lag) will define the length of the cycle



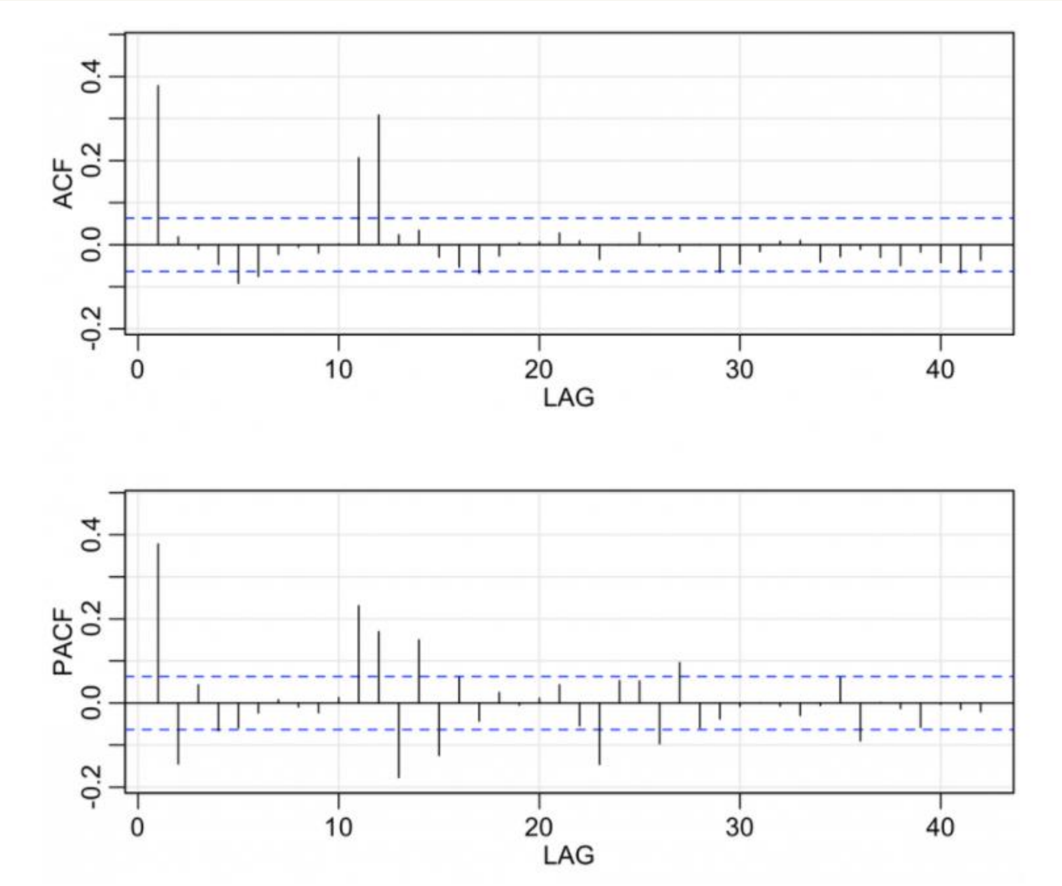


SARIMA model (Seasonal ARIMA)

- SARIMA(p,d,q)(P,D,Q)$_S$
- (p,d,q) are the orders for non-seasonal part of the time series , a (P,D,Q)$_S$ are the orders for seasonal part of the length of cycle S
- For example, ARIMA(2,0,1) is $y_t = a_1 y_{t-1} + a_2 y_{t-2} + m_1 \varepsilon_{t-1} + \varepsilon_t$ and SARIMA(0,0,0)(2,0,1)$_{12}$ is $y_t = a_{12} y_{t-12} + a_{24} y_{t-24} + m_{12} \varepsilon_{t-12} + \varepsilon_t$
- SARIMA(0,0,1)(1,0,0)$_7$ is $y_t = a_7 y_{t-7} + m_1 \varepsilon_{t-1} + \varepsilon_t$
- Pure seasonal model SARIMA(0,0,0)(1,0,0)$_{12}$ $y_t = a_{12} y_{t-12} + \varepsilon_t$ (the temperature)

Examples

- SARIMA(0,0,1)(0,0,1)$_{12}$ is $y_t = m_{13}\varepsilon_{t-13} + m_{12}\varepsilon_{t-12} + m_1\varepsilon_{t-1} + \varepsilon_t$



Identifying a Seasonal Model

- **Step 1: Do a time series plot**
- Define the parameter S from time series plot from the length of cycle patterns
- If it's hard to find the pattern see the ACF and PACF
- **Step 2: Do any necessary differencing**
- Define the parameter D from how many times you have to use seasonal differencing: df_season = df.diff(S).dropna()
- If there is linear trend and no obvious seasonality, then take a first difference. If there is a curved trend, consider a transformation of the data before differencing.
- If there is both trend and seasonality, apply a seasonal difference to the data and then re-evaluate the trend. If a trend remains, then take first differences.

Identifying a Seasonal Model

- **Step 3: Suggest the reasonable models**
- First insect the seasonal components 1S, 2S, 3S, 4S,..
- For example, ACF cuts off at lag 1S and PACF tails off at 1S, 2S,…, then suggest seasonal MA(1) model
- Then focus on nonseasonal lags 1,2,3,…,
- Spikes in the ACF (at low lags) indicate non-seasonal MA terms. Spikes in the PACF (at low lags) indicate possible non-seasonal AR terms.
- For example, if both ACF and PACF both tail off, then we might suggest ARMA(1,1) model.

Identifying a Seasonal Model

- **Step 4: Estimate the models from the previous step**
- Use SARIMAX
- **Step 5: Examine the residuals and find the best model**
- Compare AIC or BIC values to choose the best model
- You can do steps 2-5 iteratively if needed

Courses

- Time Series Analysis in R, author David S. Matteson, https://learn.datacamp.com/courses/time-series-analysis-in-r
- ARIMA Models in Python, author James Fulton, https://learn.datacamp.com/courses/arima-models-in-python

THANK YOU!

Photo by Riho Kroll on Unsplash