DATA MINING 2

Time Series - Stationarity and Forecasting

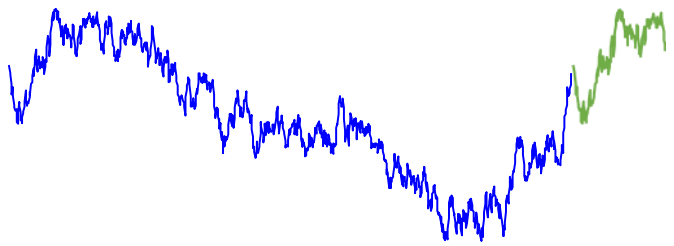
Riccardo Guidotti

a.a. 2019/2020

UNIVERSITÀ DI PISA

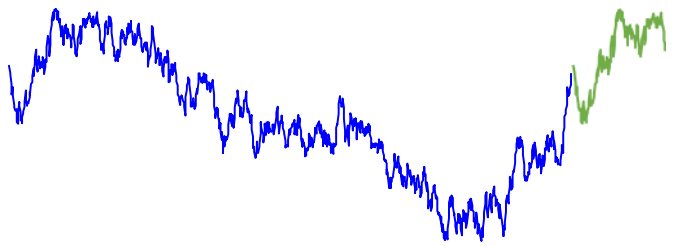
Time Series Forecasting (Prediction)

- Main difference between forecasting and classification: forecasting is about predicting a future state/value, rather than a current one.
- Applications:
  - Temperature, Humidity, CO2 Emissions
  - Epidemics
  - Pricing, Sales Volumes, Stocks
  - Forewarning of Natural Disasters (flooding, hurricane, snowstorm),
  - Electricity Consumption/Demands
- Techniques:
  - Statistical Methods,
  - Machine Learning Classifiers
  - Deep Neural Networks



Forecasting vs Regression

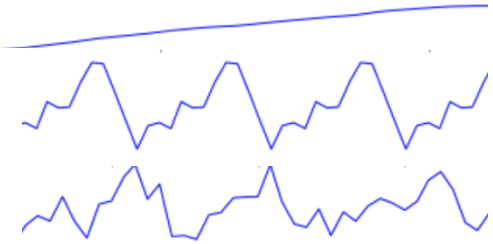
- Forecasting is **time dependent:** the basic assumption of a linear regression model that the observations are independent does not hold.
- Along with an increasing or decreasing **trend**, most TS have some form of **seasonality** trends, i.e. variations specific to a particular time frame.



Time Series Characteristics

Time Series Components

- A given TS consists of three systematic components including level, trend, seasonality, and one non-systematic component called noise.
  - **Level**: The average value in the series.
  - **Trend**: The increasing or decreasing value in the series.
  - **Seasonality**: The repeating short-term cycle in the series.
  - **Noise**: The random variation in the series.



- A **systematic** component have consistency or recurrence and can be described and modeled.
- A **Non-Systematic** component cannot be directly modeled.

Combining Time Series Components

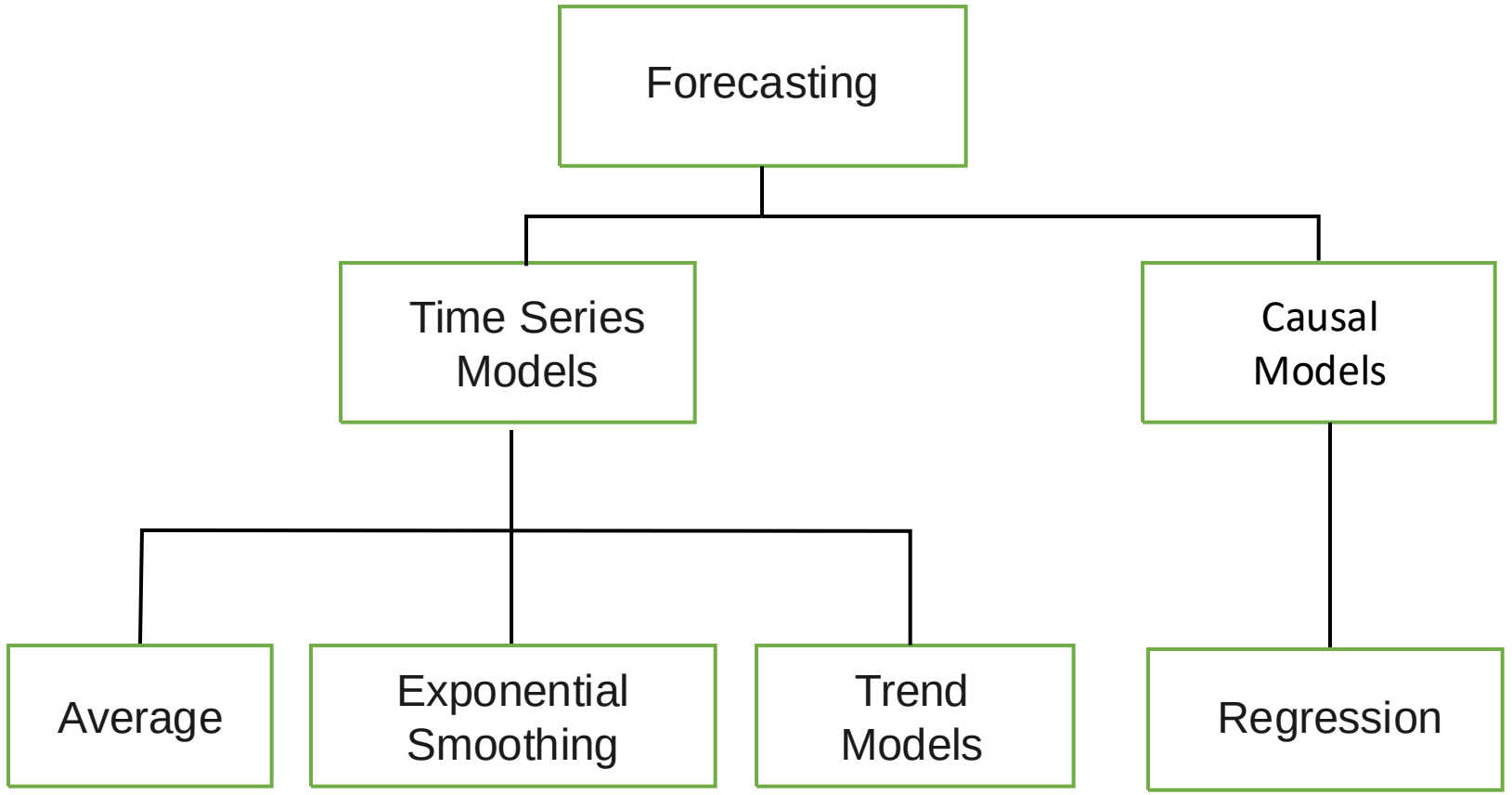
- A TS is an aggregate or combination of these four components.
- All series have a level and noise. The trend and seasonality components are optional.

- **Additive Model**: y(t) = Level + Trend + Seasonality + Noise
  - Changes over time are consistently made by the same amount
  - A linear trend is a straight line.
  - A linear seasonality has the same frequency (width of cycles) and amplitude (height of cycles).

- **Multiplicative Model**: y(t) = Level * Trend * Seasonality * Noise
  - A multiplicative model is nonlinear, such as quadratic or exponential. Changes increase or decrease over time.
  - A nonlinear trend is a curved line.
  - A non-linear seasonality has an increasing/decreasing frequency and/or amplitude over time.

Time Series Forecasting

It’s Difficult to Make Predictions, Especially About the Future



ES and ARIMA models are the two most widely used approaches to time series forecasting, and provide complementary approaches to the problem.

Evaluating Forecast Accuracy

- A forecast “error” is the difference between an observed value and its forecast. An “error” is not a mistake, is the unpredictable part.

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

- Forecast errors are different from residuals:
  - Residuals are calculated on the training set while forecast errors are calculated on the test set.
  - Residuals are based on one-step forecasts while forecast errors can involve multi-step forecasts.
- We can measure forecast accuracy by summarizing the forecast errors in different ways.

Scale-Dependent Errors

- Cannot be used to make comparisons between TS that involve different units.
- The two most commonly used scale-dependent measures are based on the absolute errors or squared errors:

$$\text{Mean absolute error: MAE} = \text{mean}(|e_t|),$$
$$\text{Root mean squared error: RMSE} = \sqrt{\text{mean}(e_t^2)}.$$

Percentage Errors

- Percentage errors are unit-free, and so are frequently used to compare forecast performances between data sets.
- The percentage error is given by

$$p_t = 100e_t/y_t$$

- The most commonly used measure is:

$$\text{Mean absolute percentage error: MAPE} = \text{mean}(|p_t|)$$

- Total and Median Absolute Percentage Error (TAPE, MedianApe) are also used.

Evaluation Measures from Regression

- **Coefficient of determination** R^2
  - is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)

hat means predicted

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

- **Mean Squared/Absolute Error** MSE/MAE
  - a risk metric corresponding to the expected value of the squared (quadratic)/absolute error or loss

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \qquad \text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

Simple Forecasting Methods

Simple Forecasting Methods

- **Average Method**: the forecasts of all future values are equal to the average (or “mean”) of the historical data.

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

- **Naïve Method**: the forecasts of all future values are equal to the last value of the historical data.

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

- **Drift Method**: increase/decrease last value w.r.t. the amount of change over time (*drift*) as the average change in the historical data.

$$\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1}) = y_T + h\left(\frac{y_T - y_1}{T-1}\right)$$

Exponential Smoothing

Simple Exponential Smoothing (SES)

- Is suitable for data with no clear trend or seasonal pattern.
- SES is in between the average and naive method.
- SES attaches larger weights to more recent observations than to observations from the distant past, while smallest weights are associated with the oldest observations
- Forecasts are calculated using weighted averages, where the weights decrease exponentially as observations come from further in the past.

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots$$

- $0 \leq \alpha \leq 1$ is the smoothing parameter

SES – Formalization in Components

- For SES the only component used is the level.
- Component form representations of SES comprise a forecast equation and a smoothing equation for each of the components in the method.

$$\begin{aligned} \text{Forecast equation} \quad & \hat{y}_{t+h|t} = \ell_t \\ \text{Smoothing equation} \quad & \ell_t = \alpha y_t + (1-\alpha)\ell_{t-1} \end{aligned}$$

- where l_t is the level of the TS at time t

Holt's Linear Trend Method

- Holt extended simple exponential smoothing to allow the forecasting of data with a trend.

$$\begin{aligned} &\text{Forecast equation} && \hat{y}_{t+h|t} = \ell_t + hb_t \\ &\text{Level equation} && \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ &\text{Trend equation} && b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \end{aligned}$$

- where l_t is the level of the TS at time t, b_t estimates the trend of TS, $0 \leq \alpha \leq 1$ is the smoothing parameter for the level and $0 \leq \beta^* \leq 1$ is the smoothing parameter for the trend.

Holt-Winters' Seasonal Method

- Holt (1957) and Winters (1960) extended Holt's method to capture seasonality.
- m denotes the frequency of the seasonality, i.e., the number of seasons in a reference period, while $0 \leq \gamma \leq 1 - \alpha$ is the smoothing parameter for the seasonality.
- The additive method is preferred when the seasonal variations are constant through the TS
- The multiplicative method is preferred when the seasonal variations are changing proportional to the level of the TS.

Holt-Winters' Seasonal Method

- Additive

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$$
$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m},$$

- Multiplicative

$$\hat{y}_{t+h|t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$$
$$\ell_t = \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma\frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}$$

k is the integer part of (h−1)/m, which ensures that the estimates of the seasonal indices come from the final period of the sample.

More on Exponential Smoothing

- ES methods are not restricted to those we have presented.

Trend	Seasonal		
	N	**A**	**M**
N	$\hat{y}_{t+h\|t} = \ell_t$ $\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = \ell_t s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t/\ell_{t-1}) + (1-\gamma)s_{t-m}$
A	$\hat{y}_{t+h\|t} = \ell_t + hb_t$ $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t/(\ell_{t-1} + b_{t-1})) + (1-\gamma)s_{t-m}$
A_d	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t$ $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = (\ell_t + \phi_h b_t)s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t/(\ell_{t-1} + \phi b_{t-1})) + (1-\gamma)s_{t-m}$

ARIMA Models

Auto-Regressive Integrated Moving Averages

- The ARIMA forecasting for a stationary time series is a linear equation (like a linear regression).
- While *ES* are based on a *description of the trend and seasonality*, *ARIMA* models aim to describe the *autocorrelations* in the data.
- Before we introduce ARIMA models, we recall the concept of stationarity and the technique of differencing TS.

Stationarity (again)

- A stationary TS is one whose properties do not depend on the time at which the series is observed.
- TS with trends, or with seasonality, are not stationary: the trend and seasonality affect the value of the TS at different times.
- A white noise series is stationary: it does not matter when you observe it, it looks much the same at any point in time.

Differencing (again)

- Differencing: compute the differences between consecutive observations.
- It is a possible transformation to make a non-stationary TS stationary.
- Indeed, it can help stabilize the mean of a TS by removing changes in the level, and thus eliminating (or reducing) trend and seasonality.
- In addition, transformations such as logarithms can help to stabilize the variance of a time series.

Autoregressive Models

- In multiple *regression* model, we predict the variable of interest using a linear combination of predictors.
- In an autoregression model, we forecast the variable of interest using a linear combination of past values of the variable.
- The term *autoregression* indicates that it is a regression of the variable against itself.
- An autoregressive model of order *p* can be written as

white noise

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

- This is as an **AR(p) model** of order *p* (p = lag in the past)

Autoregressive Models

- We normally restrict AR models to stationary data, in which case some constraints on the values of the parameters are required.
- For AR(1): $-1 \leq \phi_1 \leq 1$
- For AR(2): $-1 \leq \phi_2 \leq 1,\ \phi_1 + \phi_2 < 1,\ \phi_2 - \phi_1 < 1$
- When *p>2* the restriction are much more complicated.

Moving Average Models

- Rather than using past values of the forecast variable in a regression, a MA model uses past forecast errors in a regression-like model.

white noise

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

- This is as a **MA(q) model** of order *q* (q = lag in the past).
- MA models should not be confused with the moving average smoothing.
- It is possible to write any stationary AR(p) as MA(∞)

Moving Average Models

- It is possible to write any stationary AR(p) as MA(∞)
- The reverse result holds if we impose some constraints on the MA parameters.
- Then the MA model is called **invertible.**
- The invertibility constraints for other models are similar to the stationarity constraints.
- For MA(1): $-1 \leq \theta_1 \leq 1$
- For MA(2): $-1 \leq \theta_2 \leq 1, \theta_1+\theta_2> -1, \theta_1-\theta_2< 1$
- When p>2 the restriction are much more complicated.

ARIMA Models (Non-Seasonal)

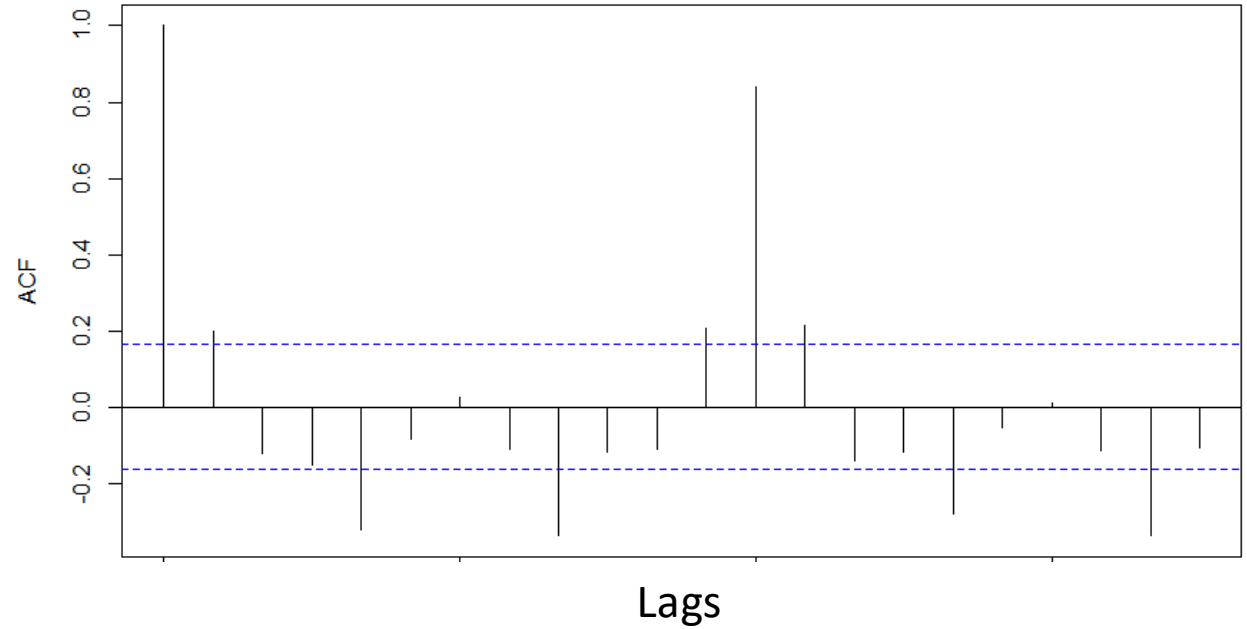
- If we combine differencing with an AR model and a MA model, we obtain a non-seasonal ARIMA model. ARIMA is an acronym for AutoRegressive Integrated Moving Average ("integration" is the reverse of differencing).

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

- where y'_t is the differenced series.
- We call this model **ARIMA(p,d,q) model**, where *p* is the order of the autoregressive part, *d* is the degree of first differencing involved, *q* is the order of the moving average part

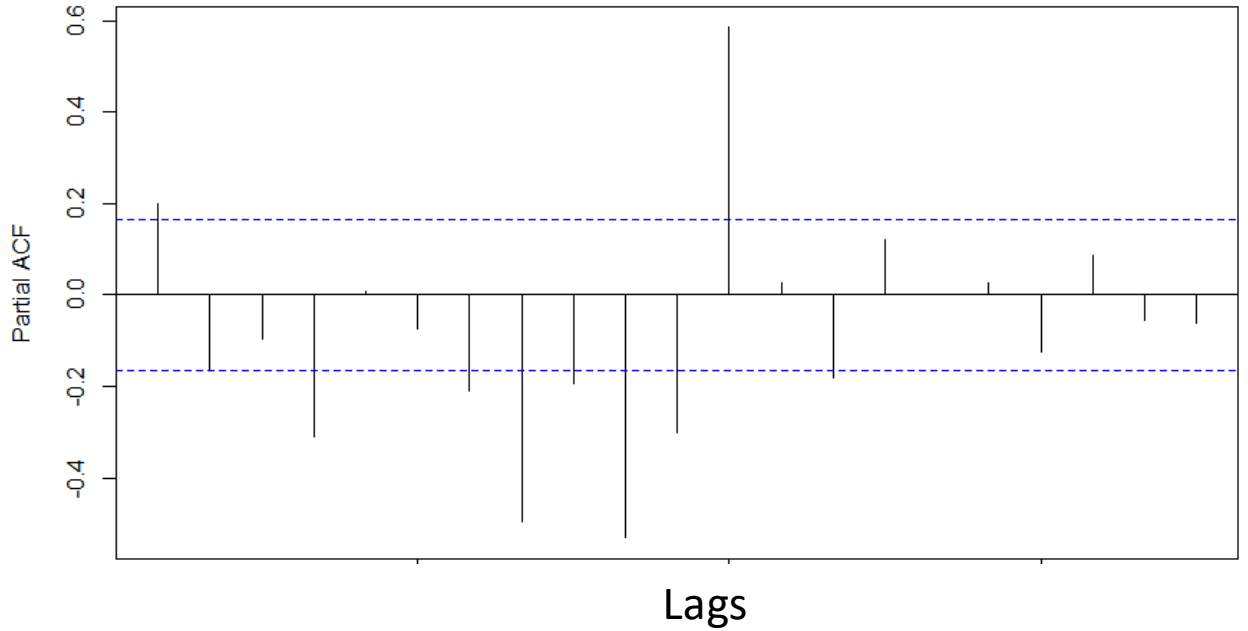
ACF plot

- The ACF plot shows the total correlation between different lag functions by calculating the correlation for TS with observations with previous time steps, called lags.
- Thus we calculate the ACF for x_t with x_{t+1} x_{t+2}, *etc.*



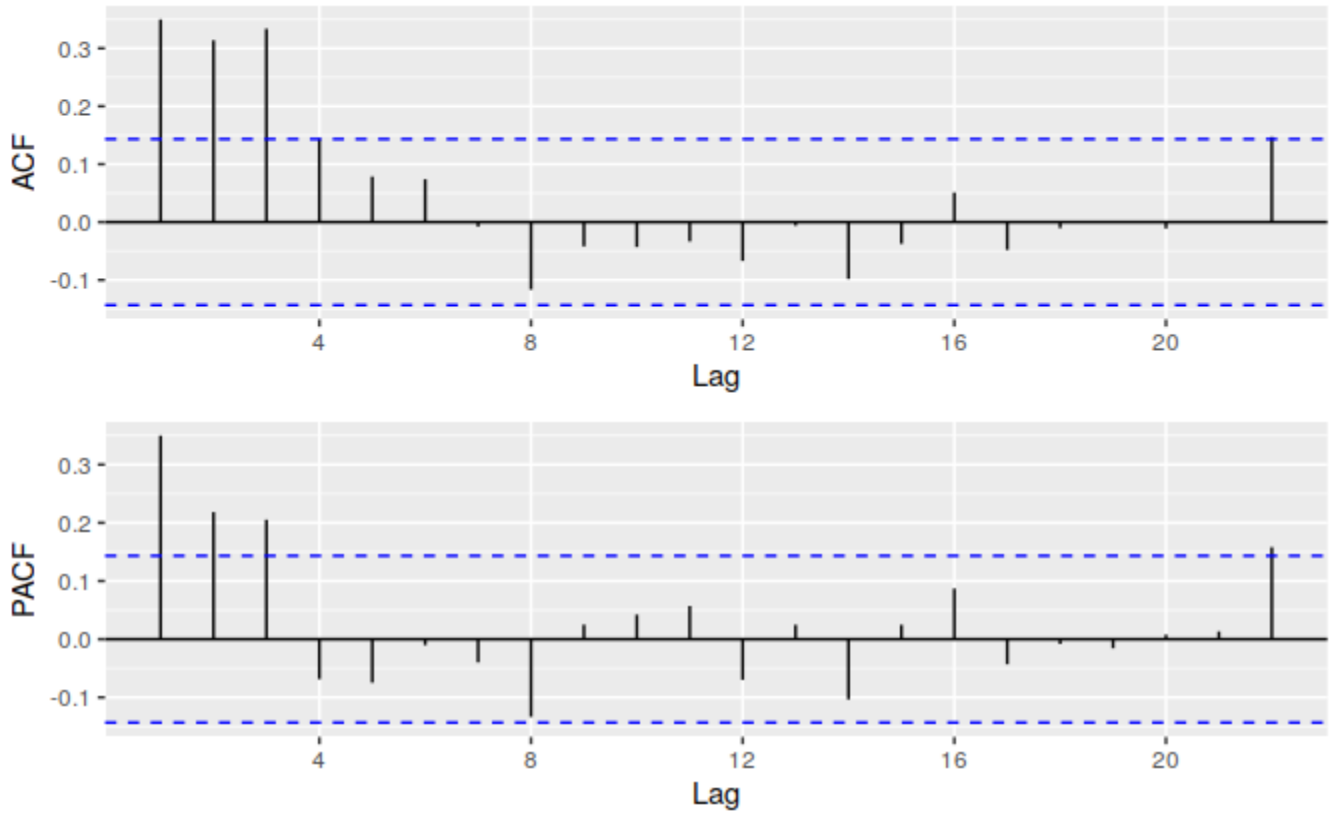
PACF plot

- A partial autocorrelation is a summary of the relationship between an observation in a TS with observations at prior time steps with the relationships of intervening observations *removed*.
- The partial autocorrelation at lag *k* is the correlation that results after removing the effect of any correlations due to the terms at shorter lags.



ACF and PACF plots - Example

- There are three spikes in the ACF, followed by an almost significant spike at lag 4. In the PACF, there are three significant spikes, and then no significant spikes.
- The pattern in the first three spikes is what we would expect from an ARIMA(3,0,0), as the PACF tends to decrease.
- So in this case, the ACF and PACF lead us to think an ARIMA(3,0,0) model might be appropriate.



ARIMA – Parameters Estimation

- Once the model order has been identified (i.e., the values of *p,d,q*), we need to estimate the parameters $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_p$.
- *Maximum Likelihood Estimation* (MLE) can be used to find the values for these parameters.
- For ARIMA models, MLE is similar to the *least squares* estimates that would be obtained by minimizing

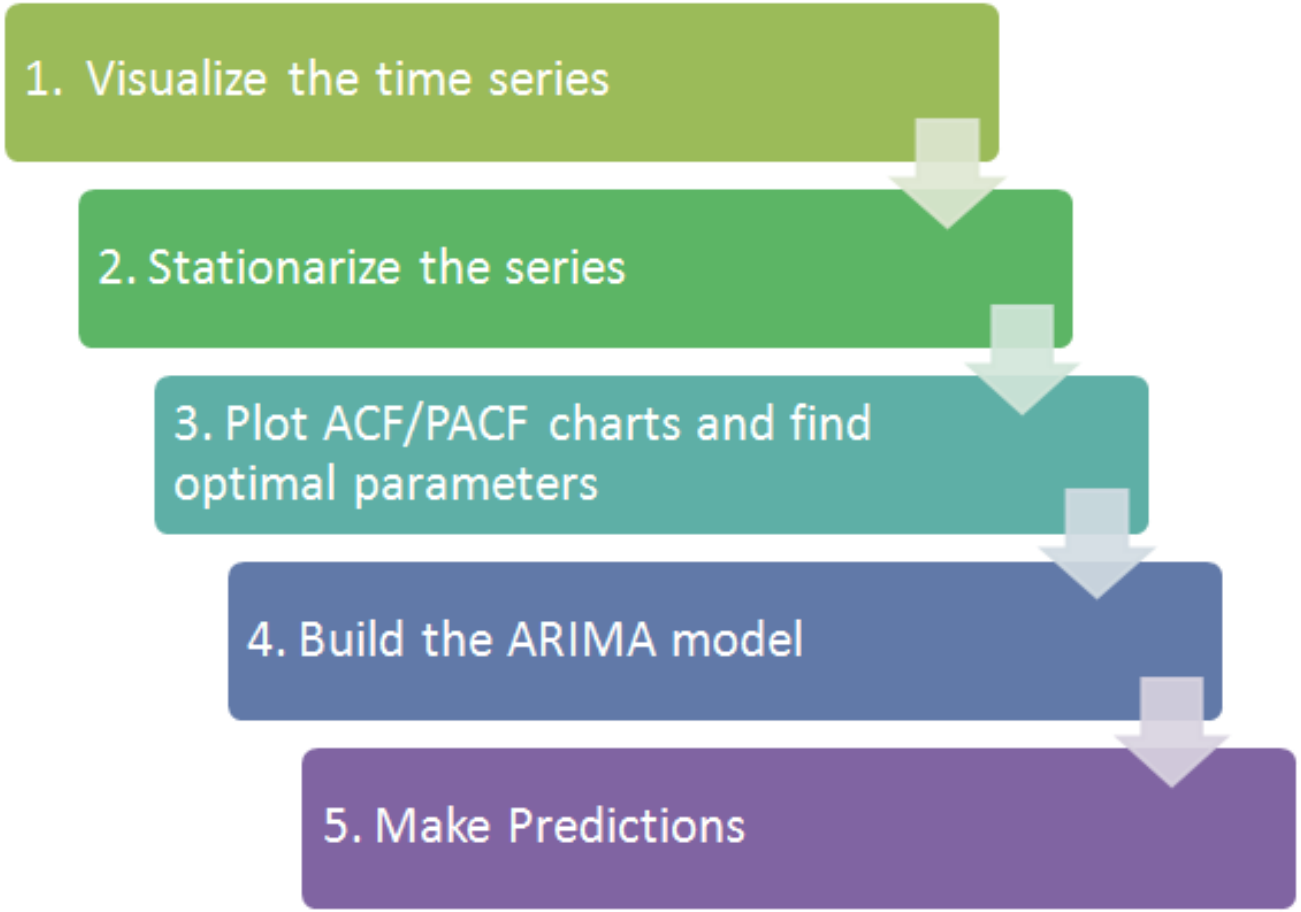
$$\sum_{t=1}^{T} \varepsilon_t^2$$

- Once the parameters are estimated they are placed in the equation and used to make the prediction of $y_{t+1}, y_{t+2}, \ldots, y_{t+n}$

Determining the order of an ARIMA model

- Akaike's Information Criterion (AIC) $\mathrm{AIC} = -2\log(L) + 2(p+q+k+1)$
- Bayesian Information Criterion (BIC) $\mathrm{BIC} = \mathrm{AIC} + [\log(T) - 2](p+q+k+1)$
- k=1 if c=0, k=0 otherwise
- Good models are obtained by minimizing the AIC, or BIC
- We highlight that AIC, or BIC are not good guides to selecting the appropriate *d*, but only for selecting *p* and *q*.
- This is because the differencing changes the data on which the likelihood is computed, making the AIC values between models with different orders of differencing not comparable.

Modelling Procedure



Advanced Forecasting Methods

Advanced Forecasting Methods

- Machine Learning models in form of (auto-)regressors can be used for time series forecasting.
- Decision Tree Regressors
- (Deep) Neural Networks Regressors
  - Convolutional Neural Networks
  - Recurrent Neural Networks
- Ensemble Regressors
  - Bagging
  - Bootstrapping
  - Random Forest Regressors

References

- Forecasting: Principles and Practic. Rob J Hyndman and George Athanasaopoulus. (https://otexts.com/fpp2/)
- Time Series Analysis and Its Applications. Robert H. Shumway and David S. Stoffer. 4th edition.(https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf)
- Mining Time Series Data. Chotirat Ann Ratanamahatana et al. 2010. (https://www.researchgate.net/publication/227001229_Mining_Time_Series_Data)
- Dynamic Programming Algorithm Optimization for Spoken Word Recognition. Hiroaki Sakode et al. 1978.
- Experiencing SAX: a Novel Symbolic Representation of Time Series. Jessica Line et al. 2009
- Compression-based data mining of sequential data. Eamonn Keogh et al. 2007.