DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Time Series

Francesco Marcelloni
Artificial Intelligence Group
IT2PAO Coordinator
Department of Information Engineering
Largo Lucio Lazzarino 1
56123 PISA

E-mail: francesco.marcelloni@unipi.it

1

University of Pisa

Francesco Marcelloni

Material

The slides are extracted from the book:
Marco Peixeiro: Time series forecasting in Python
Manning Publications Co.
2022

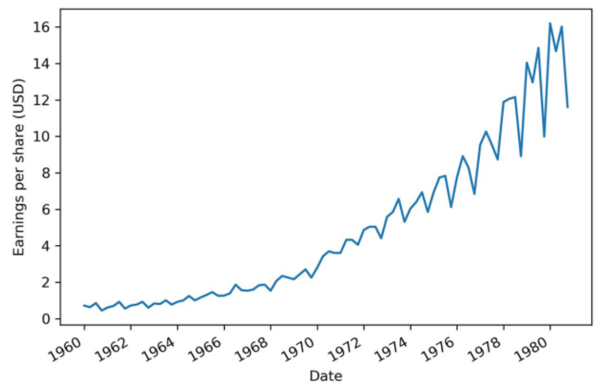
2

University of Pisa

Francesco Marcelloni

Time series

- Time series
  - **Set of data points ordered in time**
  - The data is equally spaced in time: for instance, it was recorded at every hour, minute, month or quarter.



Quarterly earnings of Johnson & Johnson in USD from 1960 to 1980 showing a positive trend and a cyclical behavior

3

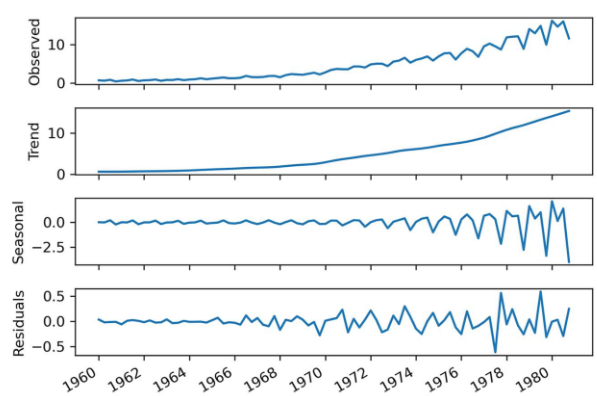
University of Pisa

Francesco Marcelloni

Time series decomposition

Decomposition: Statistical task that separates a time series into its different components trend, seasonality and residuals

- **Trend**: slow-moving changes in a time series
- **Seasonality**: seasonal pattern in the series
- **Residuals**: behavior that cannot be explained by the trend and seasonality



Decomposition of quarterly earnings of Johnson & Johnson in USD from 1960 to 1980 showing a positive trend and a cyclical behavior

4

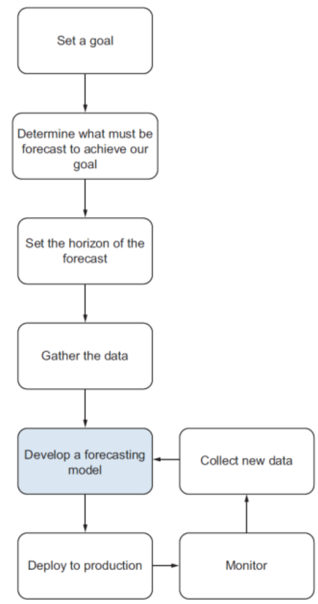
University of Pisa

Francesco Marcelloni

Time series forecasting

- Forecasting is predicting the future using historical data and knowledge of future events that might affect our forecasts

- Why different from other regression tasks?
  - Time series have an order
  - It is possible to forecast time series without the use of features



Life Cycle

5

University of Pisa

Francesco Marcelloni

Defining a baseline model

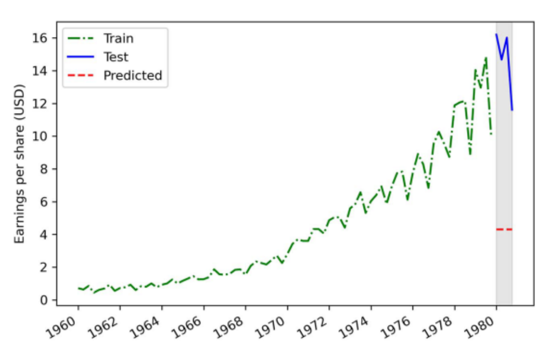
- **Goal: to predict the quarterly earnings per share (EPS) of Johnson & Johnson**
- **Baseline model**: a trivial solution to the forecasting problem under consideration
  - **compute the mean of the values over a certain period and assume that future values will be equal to that mean**

Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \times 100$$

A_i is the actual value at point i in time, and F_i is the forecast value at point i in time; n is simply the number of forecasts

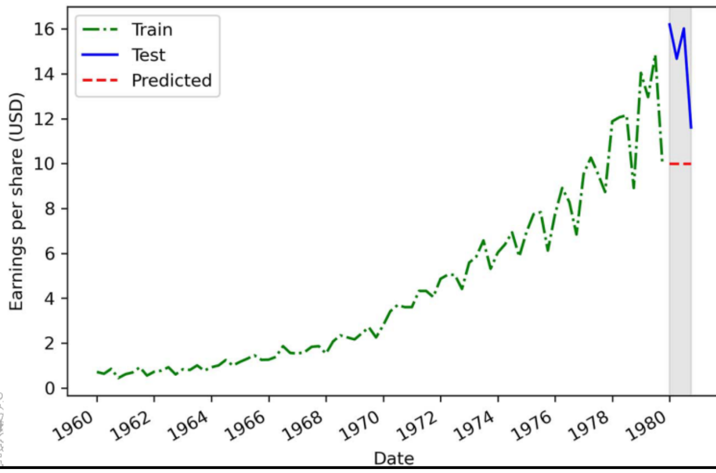
MAPE=70%



6

Defining a baseline model

- **Predict using the last known value**



MAPE=30.45%

EPS displays a cyclical behavior, where it is high during the first three quarters and then falls at the last quarter. Using the last known value does not take the seasonality into account

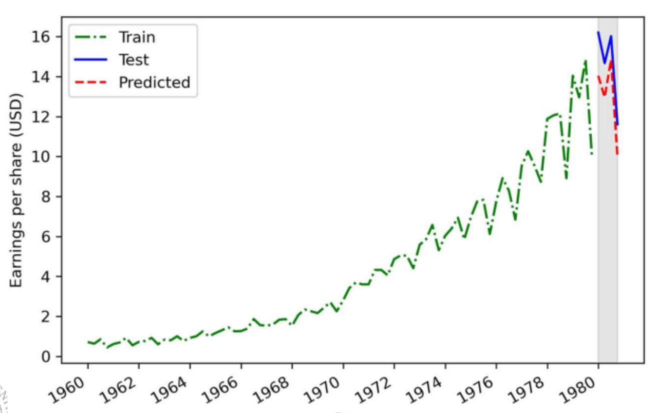
7

University of Pisa

Francesco Marcelloni

Defining a baseline model

- **Predict using the naïve seasonal forecast:** takes the last observed cycle and repeats it into the future.



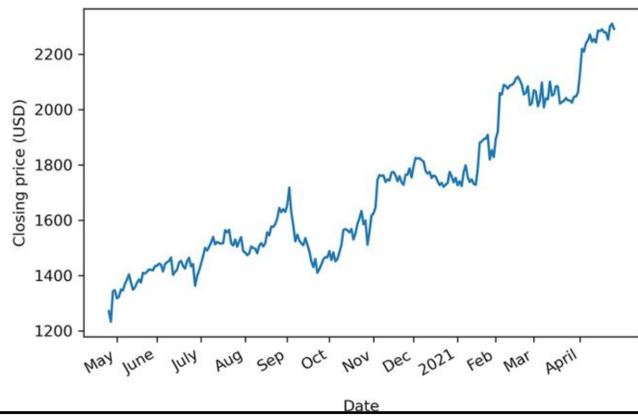
MAPE=11.56%

Seasonality has a significant impact on future values, since repeating the last season into the future yields fairly accurate forecasts.

8

Random walk process

- A **random walk** is a process in which there is an equal chance of going up or down by a random number.
- Let us assume that we would like to buy shares of a company. We would like to buy if the **closing prices of the stock is expected to go up** in the future



Long term trend is increasing but there exist abrupt changes

The daily closing price can be modeled using the random walk model

9

Random walk process

- A **random walk** is a process in which there is an equal chance of going up or down by a random number. Mathematically, we can express a random walk as:

$$y_t = C + \alpha_1 y_{t-1} + \epsilon_t$$

the current value y_t is a function of the value at the previous timestep y_{t-1}, a constant C, and a random number ϵ_t, also termed **white noise**.

Formally, a **random walk is a series whose first difference is stationary and uncorrelated: the process moves completely at random.**

$$y_0 = 0$$

$$y_1 = y_0 + \epsilon_1 = 0 + \epsilon_1 = \epsilon_1$$

$$y_1 = \epsilon_1$$

$$y_2 = y_1 + \epsilon_2 = \epsilon_1 + \epsilon_2$$

10

Stationarity

- A **stationary time series** is one whose statistical properties do not change over time. In other words, it has a constant mean, variance, and autocorrelation, and these properties are independent of time.
- Many forecasting models assume stationarity. Thus, **we need to find ways to transform the time series to make it stationary**.
- **The simplest transformation is differencing**, which calculates the change from one timestep to another, thus stabilizing the mean.

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

- It is possible to difference a time series many times to obtain a stationary series

11

Stationarity

- Differencing allows us to obtain a constant mean through time
- **To obtain a constant variance, we can use logarithms, which help stabilize the variance**
- **Note that** when we model a time series which has been transformed, we have to **untransform it** to return the results of the model to the original units of measurements.
  - For instance, if you apply a log transformation to your data, make sure you raise your forecast values to the power of 10 in order to bring the values back to their original magnitude

12

The augmented Dickey-Fuller test

- We recall that the time series can be expressed as

$$y_t = C + \alpha_1 y_{t-1} + \epsilon_t$$

- **α_1 is the root of the time series**. This time series will be stationary only if the root lies within the unit circle. Therefore, its value must be between –1 and 1. Otherwise the series is non-stationary.

Let us consider the following processes:

$y_t = 0.5 y_{t-1} + \epsilon_t$ Stationary

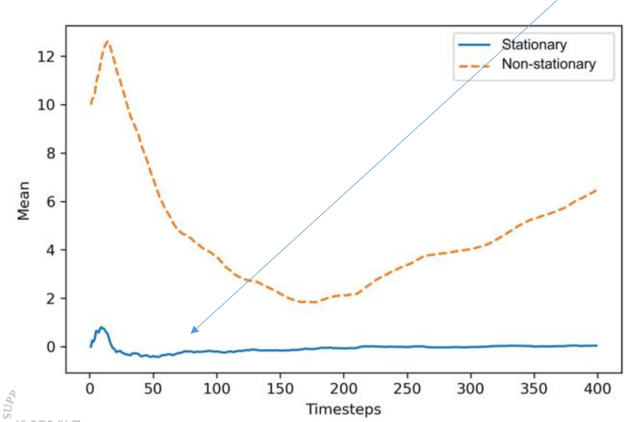
$y_t = y_{t-1} + \epsilon_t$ Non-Stationary

13

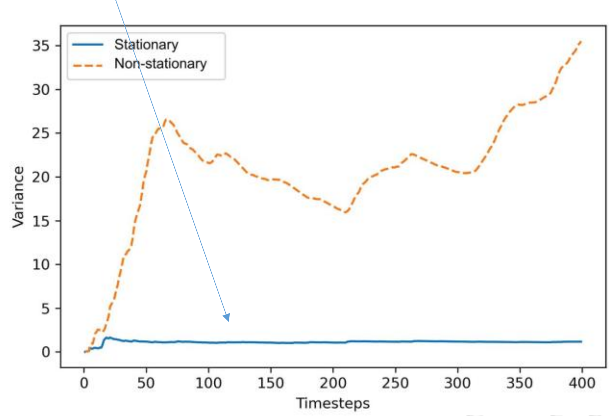
The augmented Dickey-Fuller test

The mean and variance of the stationary process becomes constant after the first few timesteps

Mean



Variance



14

The augmented Dickey-Fuller test

- The **augmented Dickey-Fuller (ADF)** test helps us determine if a time series is stationary by **testing for the presence of a unit root**. If a unit root is present, the time series is not stationary.
- **Null Hypothesis**: a unit root is present, meaning that the time series is not stationary
- If the test returns a p-value less than a certain significance level, typically 0.05 or 0.01, then we can reject the null hypothesis, meaning that there are no unit roots, and so the series is stationary.
- Once we have a stationary series, we must determine whether there is **autocorrelation or not**. Remember that a random walk is a series whose first difference is stationary and uncorrelated.

15

Autocorrelation

- The **autocorrelation function (ACF) measures the linear relationship between lagged values of a time series** (the lag is simply the number of timesteps separating two values). In other words, it measures the correlation of the time series with itself.
- For example, the autocorrelation coefficients r_1, r_2, r_3, .. are computed between, respectively, y_t and y_{t-1}, y_t and y_{t-2}, y_t and y_{t-3}, …

$$r_k = \frac{\sum\limits_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum\limits_{t=1}^{T} (y_t - \bar{y})^2}$$

- In the **presence of a trend**, a plot of the ACF will show that **the coefficients are high for short lags**, and they will decrease linearly as the lag increases.
- If the data is **seasonal**, the **ACF plot will also display cyclical patterns**.

16

University of Pisa

Francesco Marcelloni

Autocorrelation

- Let us suppose to adopt the following random walk and to generate a series with T samples.

$$y_t = y_{t-1} + \epsilon_t$$

- Once generated the series, we apply the ADF statistic and the ACF function.
  - ADF statistics is equal to -0.97 with p-value = 0.77: we cannot reject the null hypothesis.

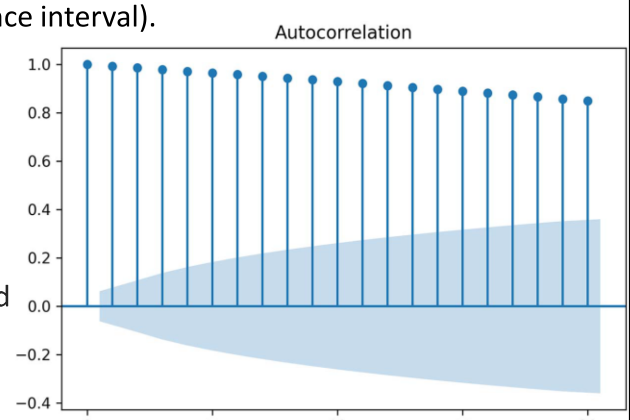
17

University of Pisa

Francesco Marcelloni

Autocorrelation

- **the autocorrelation coefficients slowly decrease as the lag increases**, which is a clear indicator that our random walk is not a stationary process (shaded area represents a confidence interval).



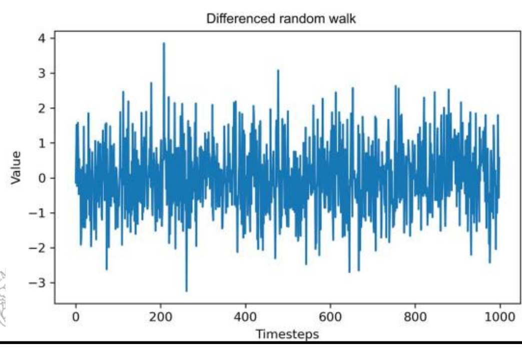
The shaded area represents 95% confidence interval: If a point is within the shaded area, then it is not significantly different from 0. Otherwise, the autocorrelation coefficient is significant. Generally, the evaluation of the shaded area is based on the Bartlett's standard errors

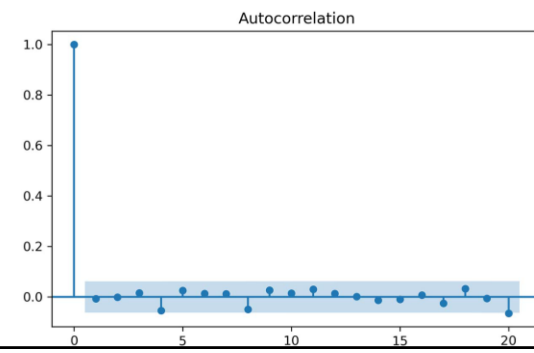
$$(1 + 2\sum_{j=1}^{k-1} r_j^2)^{1/2} n^{-1/2}$$

18

Autocorrelation

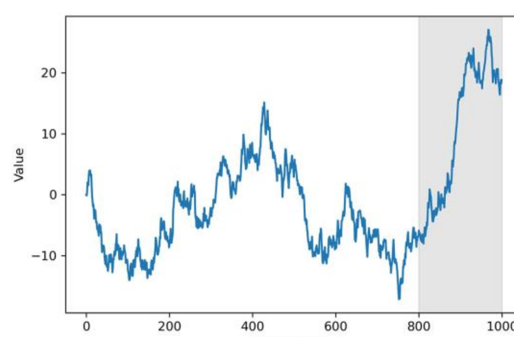
- We need to apply a **transformation**
  - Since the series displays changes in trend without seasonal patterns, we apply a first-order differencing and recompute ADF statistics and autocorrelation
    - ADF statistic is –31.79 with a p-value of 0. Null hypothesis is rejected.
    - Stationary process completely random (white noise)





Forecasting a random walk on a long horizon

- In case of **random walk**, since the values change randomly, **no statistical learning model can be applied**
  - We can only use **naïve forecasting methods or baselines** (historical mean, last known value and drift method)
  - We split in training (white area) and test (grey area)



Forecasting a random walk on a long horizon

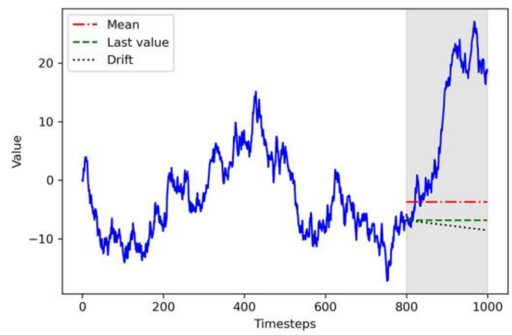
- **Drift Method**: the last value is equivalent to calculating the **slope between the first and last value of the training set** and simply extrapolating this straight line into the future.

slope

$$forecast = \frac{y_f - y_i}{\#timesteps - 1} \cdot timestep$$

where y_f and y_i are the last and first values in the training set

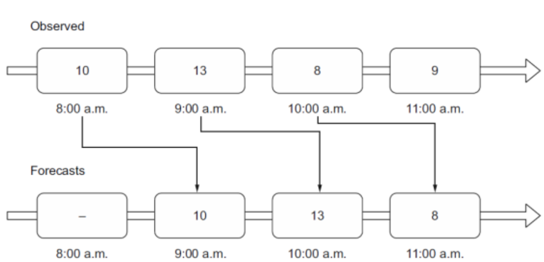
training (white area) test (grey area)

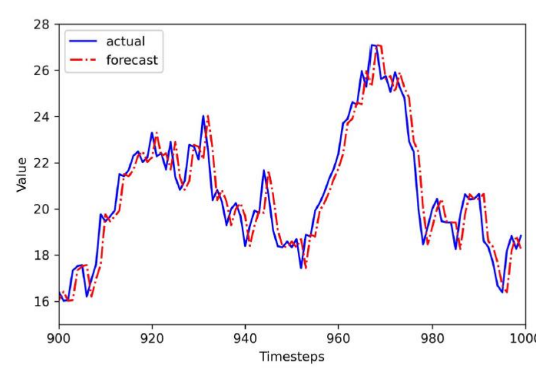


Forecasting a random walk on a long horizon does not make sense.

Forecasting the next step of a random walk

- Predict the last known value
- **Very limited approach**

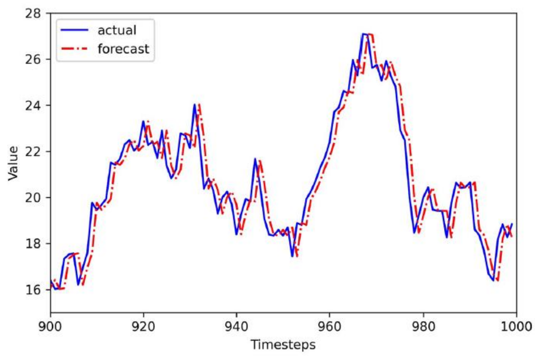


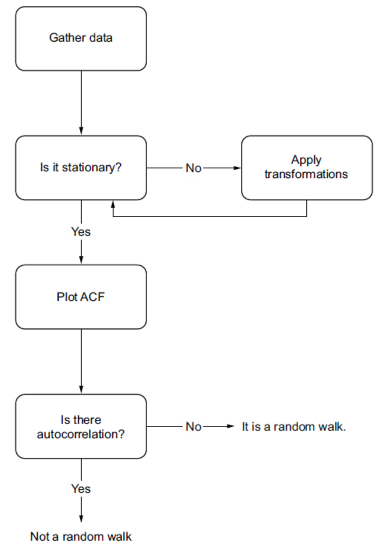


University of Pisa — Francesco Marcelloni

Forecasting the next step of a random walk

- Predict the last known value
- **Very limited approach**





23

University of Pisa — Francesco Marcelloni

Moving Average Model

Moving average model (MA): the current value depends linearly on the **mean of the series**, the **current error term** and **past error terms**

- **MA(q)**, where q is the order (past error terms)

MA(q) = $y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}$

Mean of the series

Past error terms

Current error term. Normally distributed white noise with mean 0 and constant variance (for instance 1)

Coefficients which determine the impact

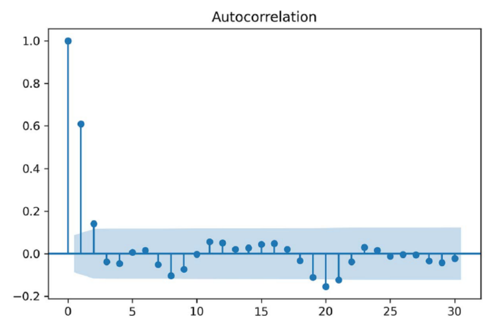
24

Moving Average Model

MA(q) = $y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}$

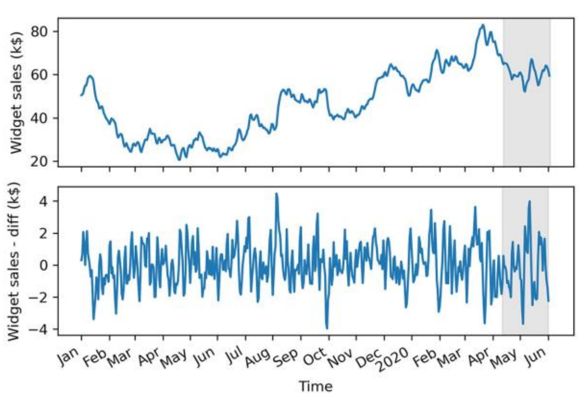
How can you choose q?

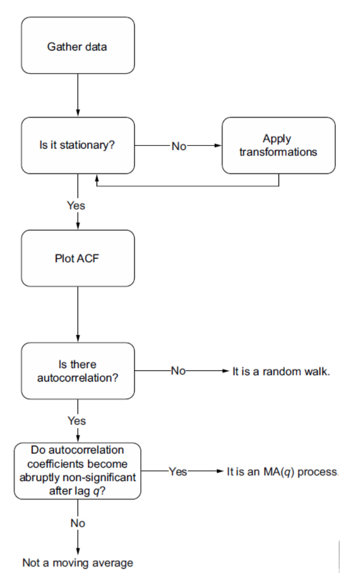
- order *q* of the moving average model determines the **number of past error** terms that affect the present value
- The larger *q* is, the more past error terms affect the present value.
- **Observe the autocorrelation plot**: the plot presents **significant correlation** coefficients up until lag *q*, after which all coefficients will be non-significant. If that is the case, then we have a moving average process of order q.



Moving Average Model

Differenced series



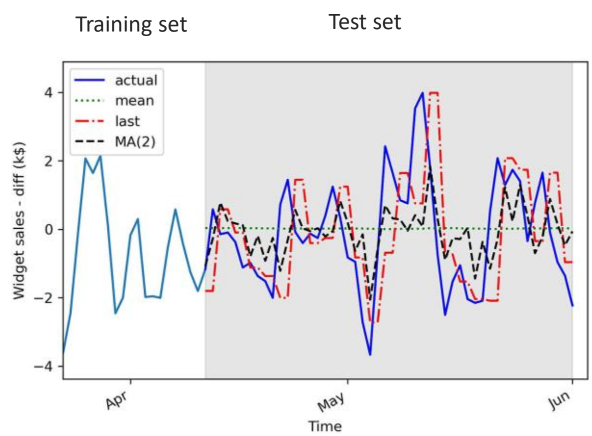


90% of the data for training set and 10% for the test set, meaning that we must forecast 50 timesteps into the future

Rolling Forecasts

When using an MA(q) model, **forecasting beyond q** steps into the future will **simply return the mean**, because there are no error terms to estimate beyond q steps. We can use **rolling forecasts to predict up to *q* steps at a time** in order avoid predicting only the mean of the series.

We need to develop a function that will predict *q* timesteps or less at a time, until all predictions in the test sets are made



We train on the first 449 timesteps and predict timesteps 450 and 451. Then, on the second pass, we will train on the first 451 timesteps, and predict timesteps 452 and 453…..

27

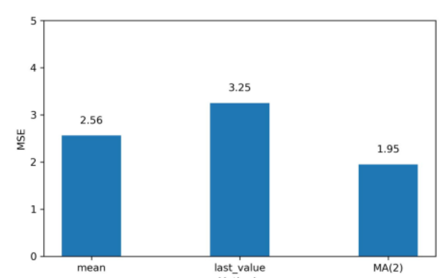
University of Pisa Francesco Marcelloni

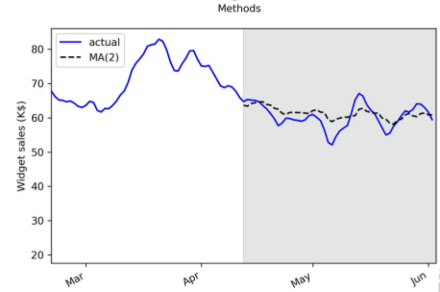
Rolling Forecasts

You do not need forecasts *q* steps ahead: you can forecast *q*-1, *q*-2, etc.

You can use SARIMAX function from the statsmodels library.

If we estimate the first-order difference, we need add the initial value y0 of the test set to the first differenced value to determine the first predicted value



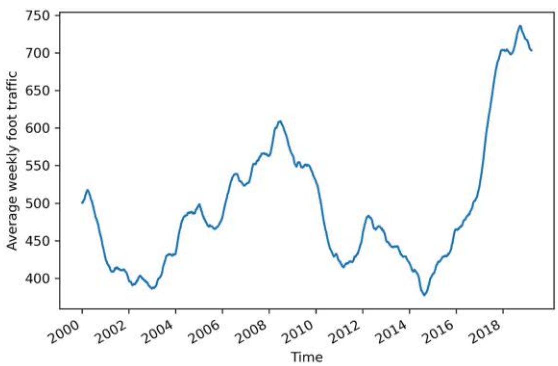


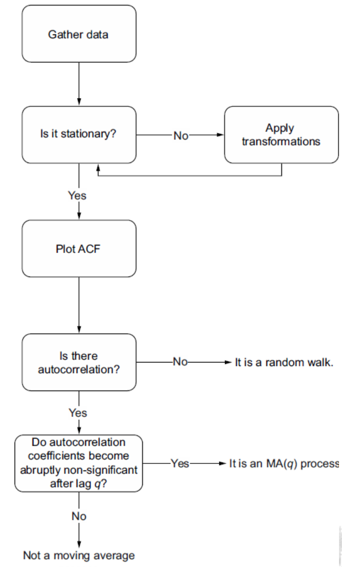
28

University of Pisa Francesco Marcelloni

Autoregressive process

We aim to forecast the average weekly foot traffic in a retail store so that the store manager can better manage the staff's schedule.





29

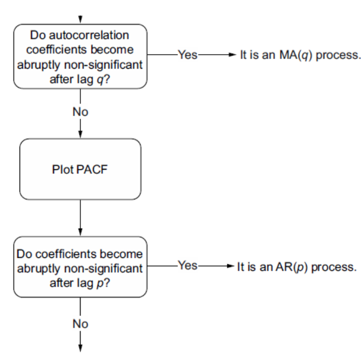
University of Pisa Francesco Marcelloni

Autoregressive process

An **autoregressive process** is a **regression** of a variable **against itself**. In a time series, this means that the present value is linearly dependent on its past values. The autoregressive process is denoted as AR(*p*), where *p* is the order. The general expression of an AR(*p*) model is:

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

The random walk is a special case of an autoregressive process (*p*=1)

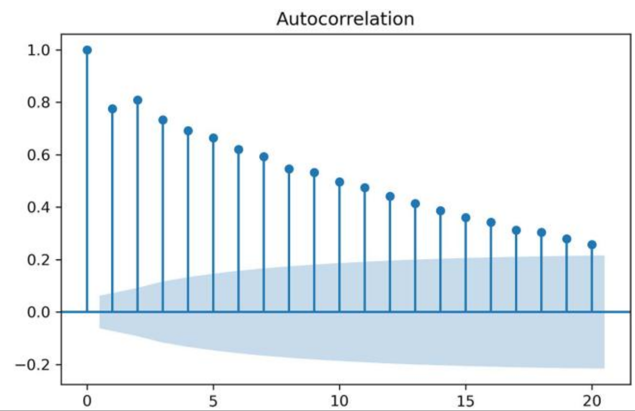


30

Autoregressive process

- In the example, ADF = -1.18 with a p-value = 0.68. The null Hypothesis cannot be rejected.
- Differenced series: ADF = -5.27 with a p-value = $6.36 \cdot 10^{-6}$. The null hypothesis is rejected and therefore stationary.
- We compute the autocorrelation

There is no lag at which the coefficients abruptly become nonsignificant.
We do not have a moving average process and that we are likely studying an autoregressive process.



Autoregressive process

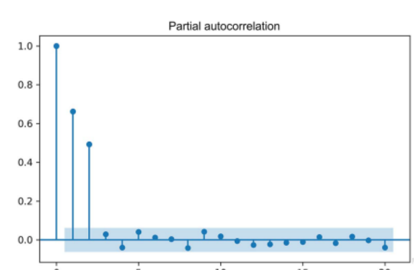
$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots \phi_p y_{t-p} + \epsilon_t$$

- How can we determine the order *p*?
- **Partial Autocorrelation Function**: measures the **correlation between lagged values in a time series when we remove the influence of correlated lagged values in between**. We can plot the partial autocorrelation function to determine the order of a stationary AR(p) process. The coefficients will be non-significant after lag p.

$$y_t = 0.33 y_{t-1} + 0.50 y_{t-2}$$

We wish to measure how y_t relates to y_{t-2}.

When we measure the autocorrelation between y_t and y_{t-2} we are not taking into account that y_{t-1} has an influence on both y_t and y_{t-2}. This means that we are not measuring the real impact of y_{t-2} on y_t.

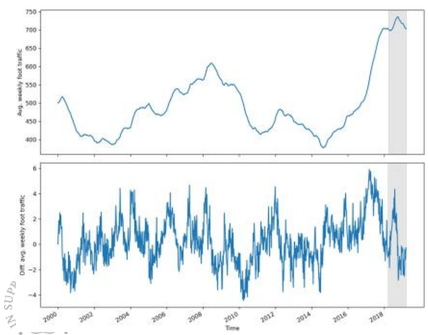


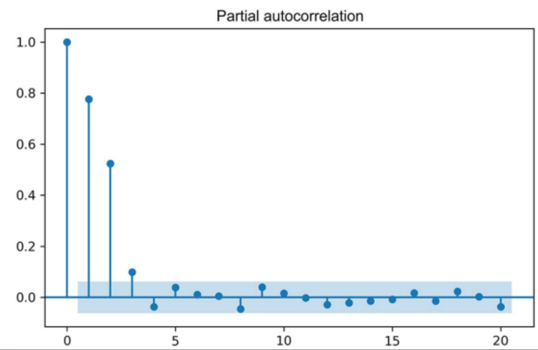
Autoregressive process

- We adopt 947 weeks in the training set and 52 weeks in the test set
- We compute PACF for the differenced series

We can observe that **there are no significant coefficients** after lag 3.

Therefore, the **differenced average weekly foot traffic is an autoregressive process** of order 3, which can also be denoted as AR(3).





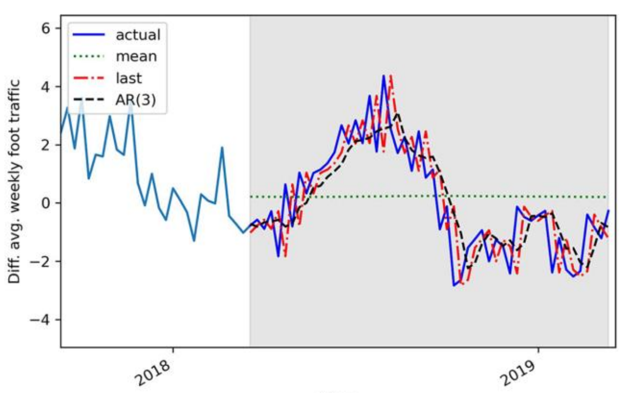
33

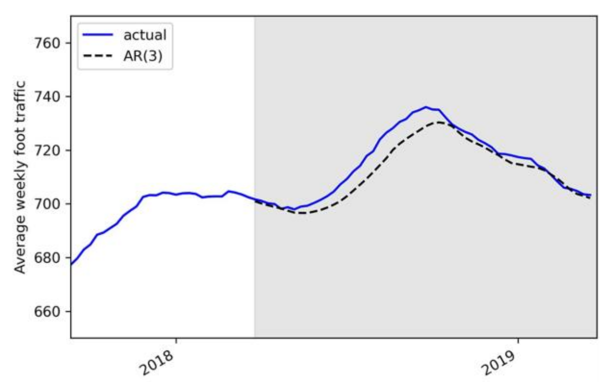
University of Pisa

Francesco Marcelloni

Rolling Forecasts

- We adopt the rolling forecast with window 1



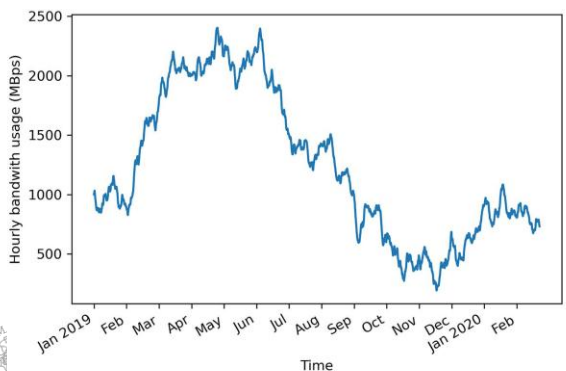


34

University of Pisa | Francesco Marcelloni

Autoregressive moving average

- We aim to **predict bandwidth usage for a large data center**.
- Bandwidth is defined as the maximum rate of data that can be transferred. Its base unit is bits per second (bps)



First, we **check whether the series is stationary**. It is not.
Second, we **transform by using differences and verify whether the differenced series is stationary**. It is.
Third, we plot the ACF functions and find that there are significant autocorrelation coefficients after lag 0 and these coefficients slowly decay.
Fourth, we plot the PACF function and will verify that they have a sinusoidal pattern.
Thus, it **is not a purely moving average process** and it **is not a purely autoregressive process**

35

University of Pisa | Francesco Marcelloni

Autoregressive moving average

Autoregressive moving average process

- **combination of the autoregressive process and the moving average process**.
- It is denoted as **ARMA(p,q)**, where p is the order of the autoregressive process, and q is the order of the moving average process. The general equation of the ARMA(p,q) model is

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

- An ARMA(0,q) process is equivalent to an MA(q) process, since the order p = 0 cancels the AR(p) portion.
- An ARMA(p,0) process is equivalent to an AR(p) process, since the order q = 0 cancels the MA(q) portion.

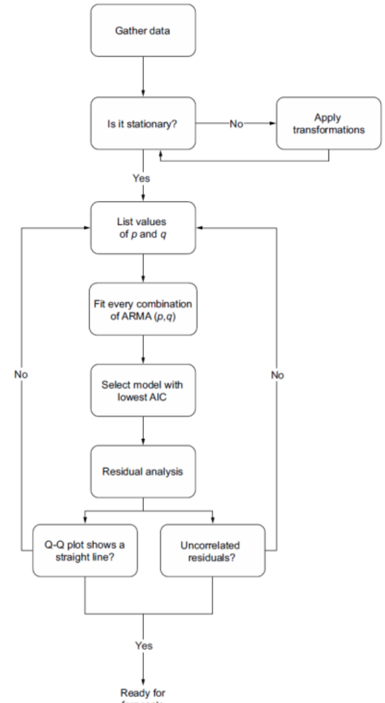
36

University of Pisa

Autoregressive moving av

How can we determine p and q?

- We need a new procedure!
- With a list of possible values for p and q, we can fit every unique combination of ARMA(p,q) and evaluate it by using the **Akaike Information criterion (AIC)**
- The model with the lowest AIC is selected
- Then, we analyze the model's residuals, which is the difference between the actual and predicted values of the model. If the residuals look like white noise, then they are uncorrelated and independently distributed and the model is ready for forecasting; otherwise, a different set of values for p and q has to be evaluated.



37

University of Pisa

Autoregressive moving av

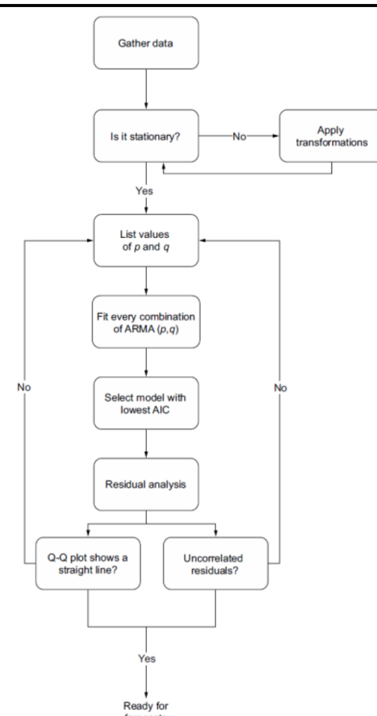
Akaike Information criterion (AIC)

- The AIC is a function of the number of parameters k in a model and the maximum value of the likelihood function :

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

- The **lower the value** of the AIC, the **better the model**.
- AIC allows us to keep a balance between the complexity of a model and its goodness of fit to the data.

- In our case, k=p + q
- The likelihood function measures the goodness of fit of a model. In other words, it measures "How likely is it that my observed data is coming from an ARMA(1,1) model?"



38

Autoregressive moving average

Example

- Let us assume that p and q vary from 0 and 3. Thus, we have 16 possible combinations
- We iterate for each possible combination (p,q), compute the corresponding ARMA(p,q) model and evaluate the AIC with the available data
- We select the model with the lowest AIC.
- Consider the simulated model ARMA(1,1) = $y_t = 0.33y_{t-1} + 0.9\epsilon_{t-1} + \epsilon_t$
- Let us suppose that we estimate the coefficients perfectly

$$\widehat{y_t} = 0.33y_{t-1} + 0.9\epsilon_{t-1}$$

- Then

$$\text{residuals} = 0.33y_{t-1} + 0.9\epsilon_{t-1} + \epsilon_t - (0.33y_{t-1} + 0.9)$$

$$\text{residuals} = \epsilon_t$$

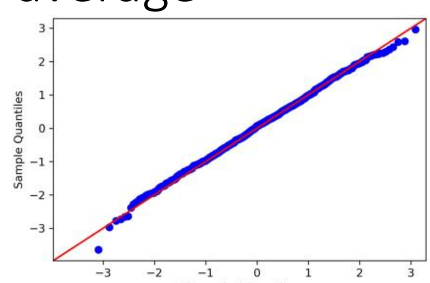
39

Autoregressive moving average

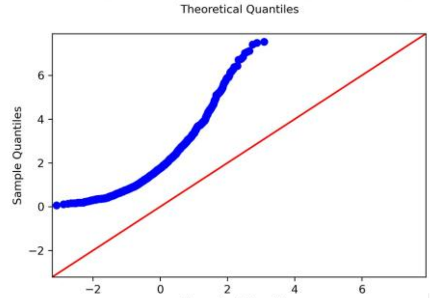
How can we analyse the residuals?

- **Qualitative analysis: Q-Q plot**
  - The Q-Q plot is constructed by plotting the quantiles of the residuals on the y-axis against the quantiles of a theoretical distribution, in this case the **normal distribution**, on the x-axis.
  - If the model is a good fit, the residuals are similar to white noise

Normal distribution



Non-Normal distribution



40

Autoregressive moving average

How can we analyse the residuals?

- **Quantitative analysis: Ljung-Box test**
  - The Ljung-Box test is a statistical test that determines whether the **autocorrelation of a group of data is significantly different from 0**.
  - Null hypothesis: the data is independently distributed, meaning that there is no autocorrelation.
  - If the **p-value is larger than 0.05**, we cannot reject the null hypothesis, meaning that the residuals are independently distributed. Therefore, there is no autocorrelation, the residuals are similar to white noise, and the model can be used for forecasting.
  - If the **p-value is less than 0.05**, we reject the null hypothesis, meaning that our residuals are not independently distributed and are correlated. The model cannot be used for forecasting.

41

Autoregressive moving average

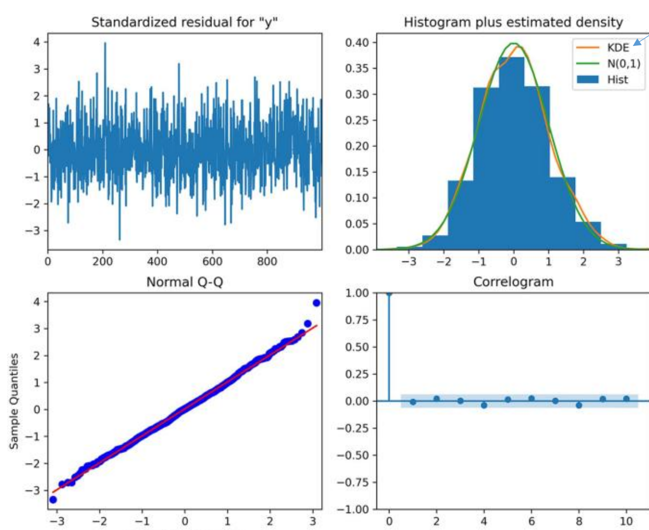
estimated density of residuals

Residuals

Histogram of the residuals

Q-Q plot

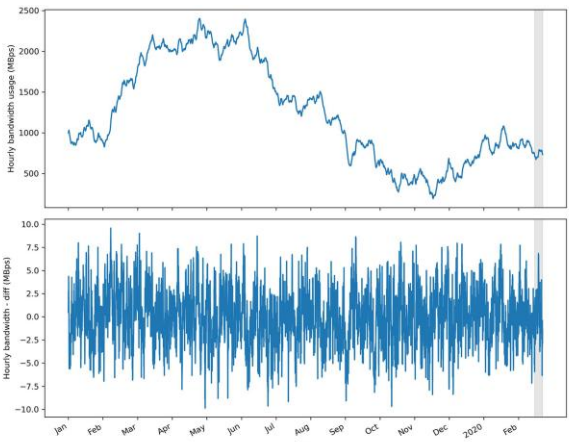
ACF plot of the residuals

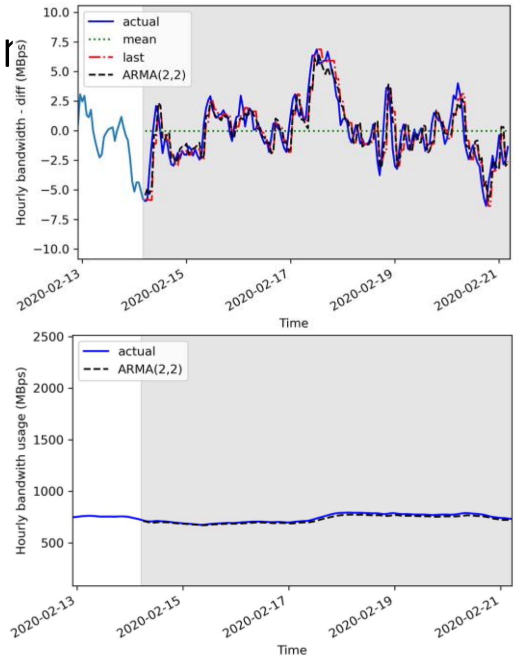


42

Autoregressive movi

Example





43

Non-stationary time series

We observed that the ARMA model is suitable only for stationary time series.
For non-stationary time series we need to add a component, the **integration order**.

- **Autoregressive Integrated Moving Average (ARIMA)**
  - is the combination of the AR(p) and MA(q) processes, but in terms of the differenced series.
  - is denoted as ARIMA(p,d,q), where **p** is the **order of the AR(p) process**, **d** is the order of integration, and **q** is the order of the MA(q) process.
  - Integration is the reverse of differencing, and the order of integration d is equal to the number of times the series has been differenced to be rendered stationary.

44

Non-stationary time series

- **Autoregressive Integrated Moving Average (ARIMA)**
  - The general equation of the ARIMA(p,d,q) process is

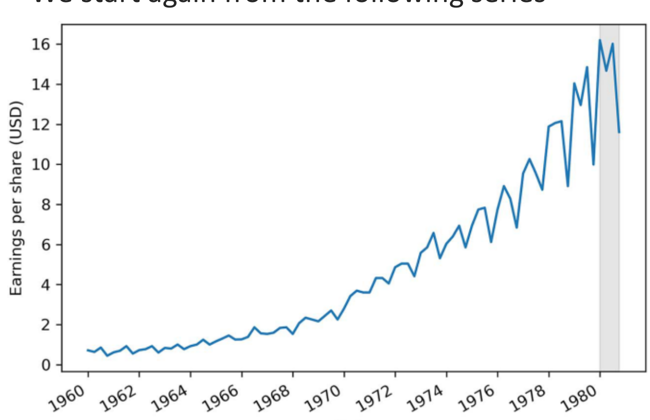
$$y'_t = C + \varphi_1 y'_{t-1} + \cdots \varphi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

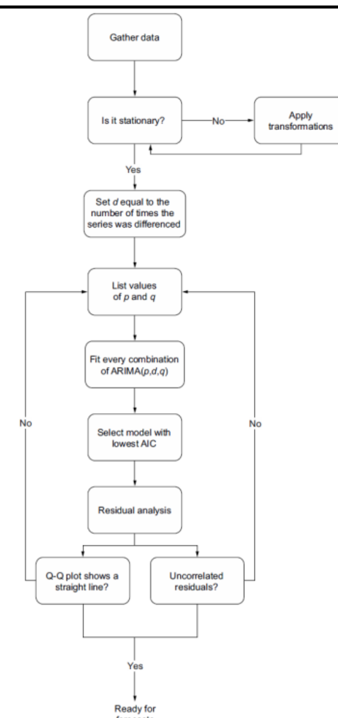
  - Note that y'_t represents the differenced series, and it may have been differenced more than once.
  - **Order *d* is the order of integration**, where integration is the reverse of differencing.
  - We have to find the order of integration, which corresponds to the minimum number of times a series must be differenced to become stationary

45

Non-stationary time serie

- **Autoregressive Integrated Moving Average (ARIMA)**
  - We need to change a little the previous procedure
  - We start again from the following series



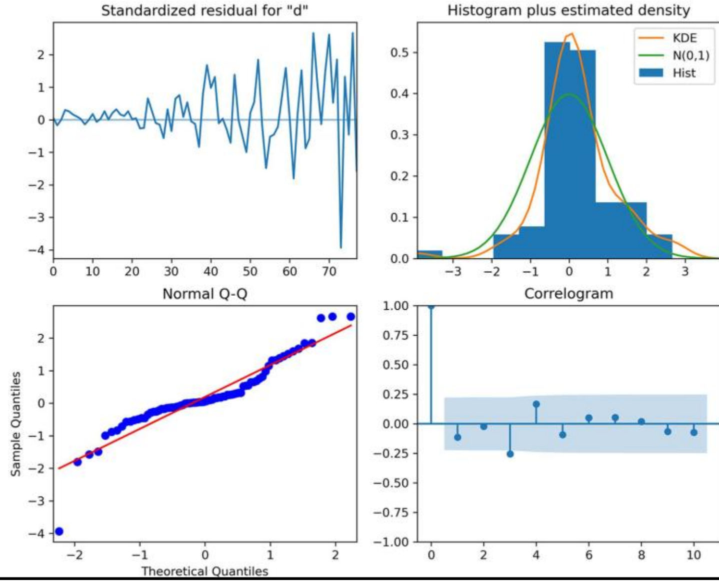


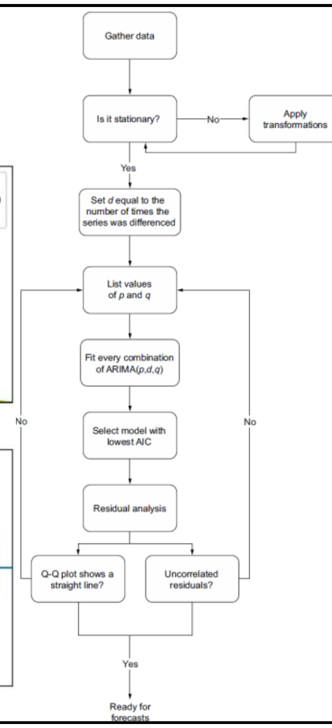
46

University of Pisa

Non-stationary time serie

After the application of the process, we determine that the best model is ARIMA(3,2,3). Indeed, the Ljung-Box test on the first 10 lags of the model's residuals returns a list of p-values that are all larger than 0.05.





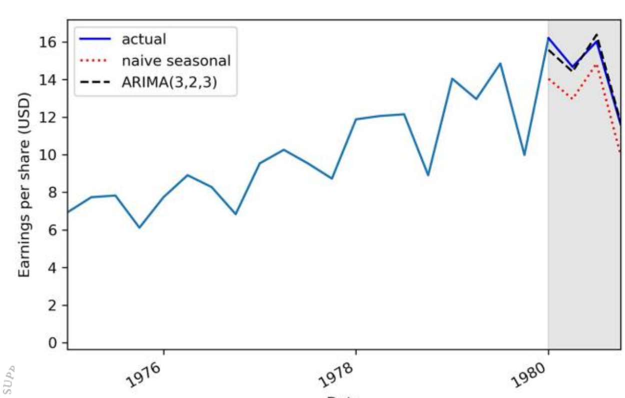
47

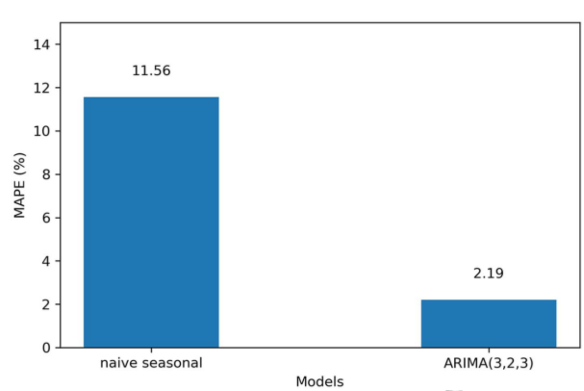
University of Pisa

Francesco Marcelloni

Non-stationary time series

Naive seasonal method: the first quarter of 1979 is used to forecast for the EPS of the first quarter of 1980; the EPS of the second quarter of 1979 is used as a forecast for the EPS of the second quarter of 1980, and so on.

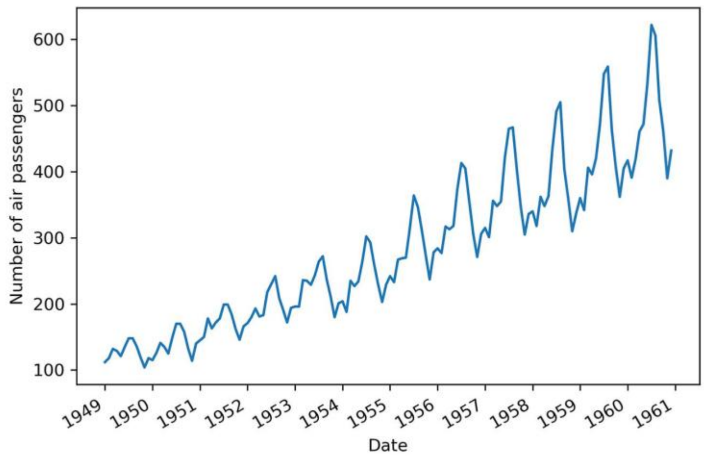




48

Non-stationary time series with seasonality

- We take into consideration the monthly total number of air passengers for an airline.



49

Non-stationary time series with seasonality

- **Seasonal autoregressive integrated moving average (SARIMA) model**
  - **SARIMA model adds seasonal parameters to the ARIMA(p,d,q) model**.
  - It is denoted as **SARIMA(p,d,q)(P,D,Q)$_m$**, where P is the order of the seasonal AR(P) process, D is the seasonal order of integration, Q is the order of the seasonal MA(Q) process, and m is the frequency, or the number of observations per seasonal cycle.
  - Note that a SARIMA(p,d,q)(0,0,0)$_m$ model is equivalent to an ARIMA(p,d,q) model.

50

Non-stationary time series with seasonality

- **Appropriate frequency m (some example)**

Data collection	Frequency *m*
Annual	1
Quarterly	4
Monthly	12
Weekly	52

Data collection	Frequency *m*				
	Minute	Hour	Day	Week	Year
Daily				7	365
Hourly			24	168	8766
Every minute		60	1440	10080	525960
Every second	60	3600	86400	604800	31557600

51

University of Pisa
Francesco Marcelloni

Non-stationary time series with seasonality

- **How can we choose P and Q?**
  - Let us consider an example with *m*=12.
  - If we are considering P=2, this means that we are including two past values of the series at a lag that is a multiple of *m*. Therefore, we include the values at y_{t-12} and y_{t-24}.
  - Similarly, if D = 1, this means that a seasonal difference makes the series stationary. With *m*=12, this corresponds to $y'_t = y_t - y_{t-12}$
  - If Q = 2, we include past error terms at lags that are a multiple of *m*. Therefore, we include the errors ϵ_{t-12} and ϵ_{t-24}.

52

University of Pisa

Francesco Marcelloni

Non-stationary time series with seasonality

- **How can we identify seasonal patterns?**
  - **Visually** by plotting the series
  - Through **time series decomposition** (STL function from the statsmodels library), that is, the statistical task that separates the time series into its three main components: a trend component, a seasonal component, and the residuals.
    - The **trend component** represents the long-term change in the time series. This component is responsible for time series that increase or decrease over time.
    - The **seasonal component** is the periodic pattern in the time series. It represents repeated fluctuations that occur over a fixed period of time.
    - The **residuals**, or the noise, express any irregularity that cannot be explained by the trend or the seasonal component.

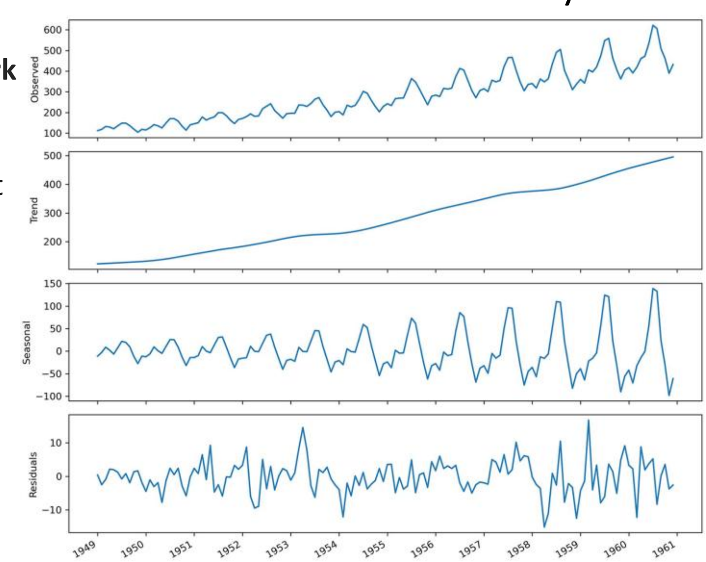
53

University of Pisa

Francesco Marcelloni

Non-stationary time series with seasonality

- **Time series decomposition at work**
  - We are using m = 12
  - We can observe that the seasonal component is present



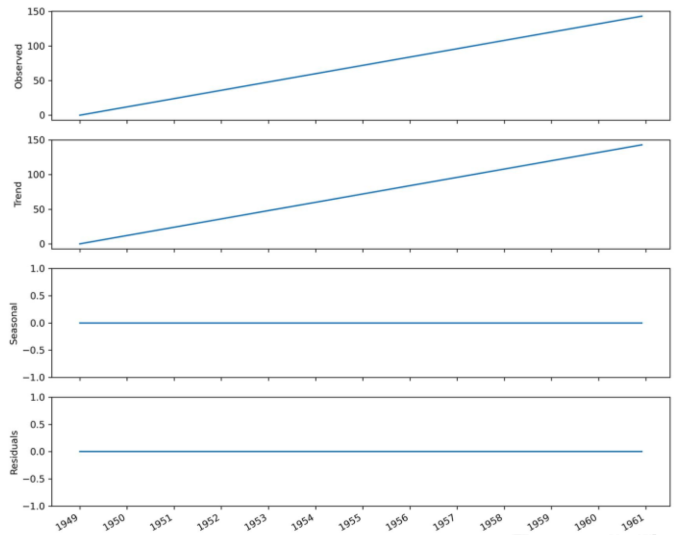
54

University of Pisa

Francesco Marcelloni

Non-stationary time series with seasonality

- **Time series decomposition at work**
  - Example of a simulated linear series (no seasonal pattern)
  - We are using m = 12
  - We can observe that the seasonal component corresponds to a flat horizontal line at 0. Actually, no seasonal pattern!

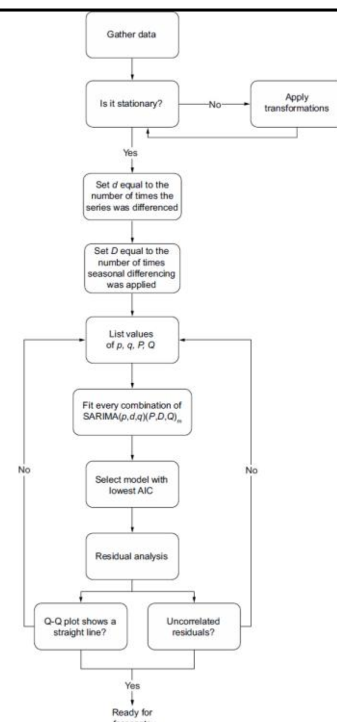


55

University of Pisa

Non-stationary time series with se

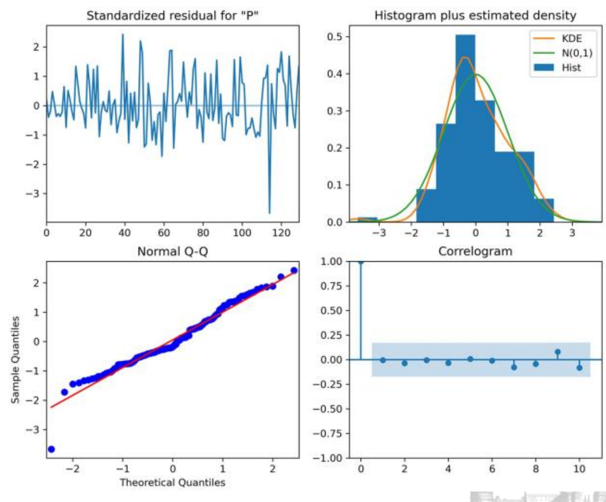
- **New procedure**
  - We have to check for all the possible parameters
  - Two experiments
    - SARIMA(p,d,q)(0,0,0)$_m$ which corresponds to ARIMA(p,d,q)
    - SARIMA(p,d,q)(P,D,Q)$_m$



56

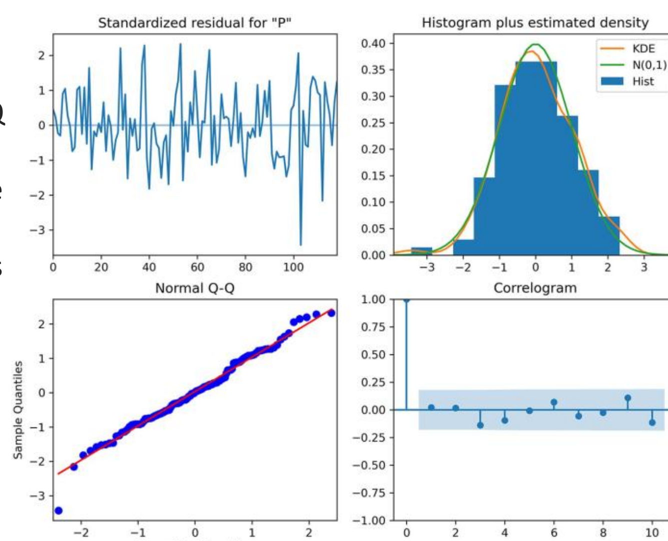
Non-stationary time series with seasonality

- SARIMA(p,d,q)(0,0,0)$_m$
- Ljung-Box test: The returned p-values are all greater than 0.05 except for the first two values. Thus, we can conclude that the residuals are uncorrelated starting at lag 3.
- The test is suggesting that the model is not perfect although it is not preforming so bad.



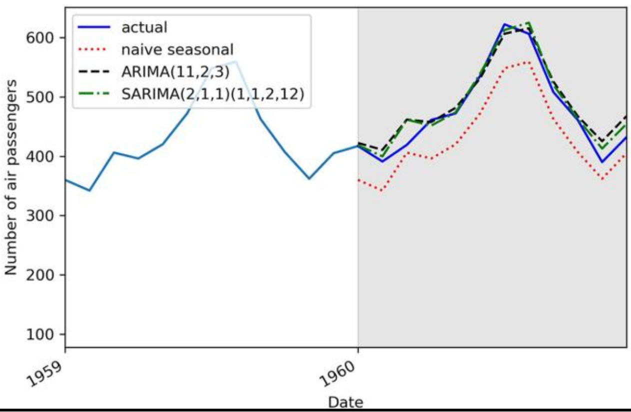
Non-stationary time series with seasonality

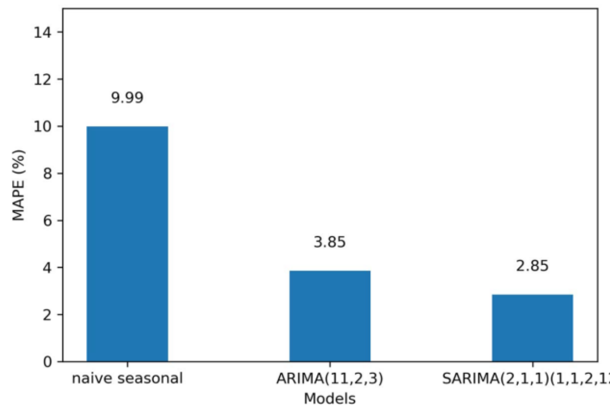
- SARIMA(p,d,q)(P,D,Q)$_m$
- We use d and D equal to 1
- We try values in [0,1,2,3] for p, q, P, Q
- We determine that SARIMA(2,1,1)(1,1,2)$_{12}$ model has the lowest AIC
- Ljung-Box test: The returned p-values are all greater than 0.05



Non-stationary time series with seasonality

- Comparing the different approaches





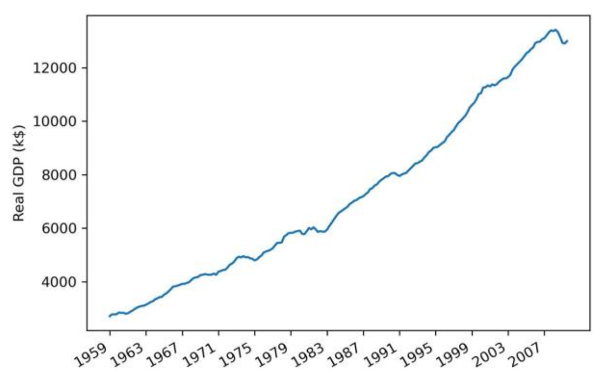
59

University of Pisa
Francesco Marcelloni

Non-stationary time series with seasonality and external variables

- Often, **external variables are also predictive of time series**
- For instance, Gross Domestic Product (GDP) which is the total market value of all the finished goods and services produced within a country, is defined as the **sum of consumption C**, **government spending G**, **investments I**, and **net exports NX**. Each element is probably affected by external variables

$$\mathrm{GDP} = C + G + I + NX$$



60

Non-stationary time series with seasonality and external variables

- **SARIMAX model**, where X denotes exogenous variables, adds simply **linear combination of exogenous variables to the SARIMA model**. This allows modelling the impact of external variables on the future value of a time series.

- We can loosely define the SARIMAX model as follows:

$$y_t = SARIMA(p,d,q)(P,D,Q)_m + \sum_{i=1}^{n} \beta_i X_t^i$$

- The SARIMAX model is the most general model for forecasting time series. You can see that if you have no seasonal patterns, it becomes an ARIMAX model. With no exogenous variables, it is a SARIMA model. With no seasonality or exogenous variables, it becomes an ARIMA model.

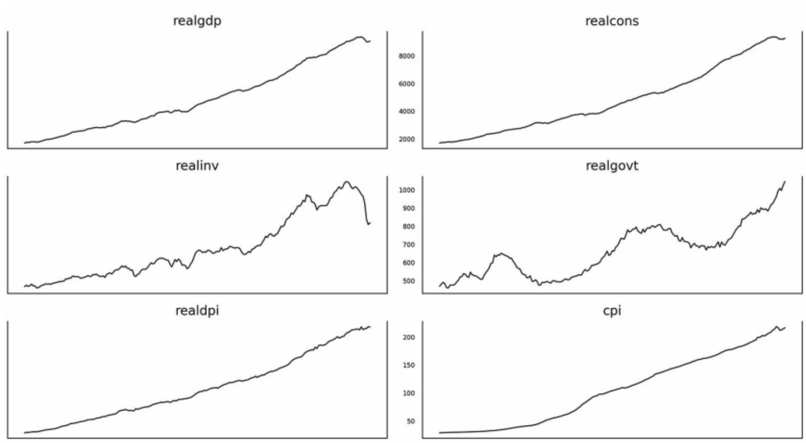
Non-stationary time series with seasonality and external variables

- **In the example, we take into consideration the following variables**

Variable	Description
realgdp	Real gross domestic product (the target variable or endogenous variable) ← Target
realcons	Real personal consumption expenditure
realinv	Real gross private domestic investment
realgovt	Real federal consumption expenditure and investment
realdpi	Real private disposable income
cpi	Consumer price index for the end of the quarter

Non-stationary time series with seasonality and external variables

- **In the example, we take into consideration the following variables**



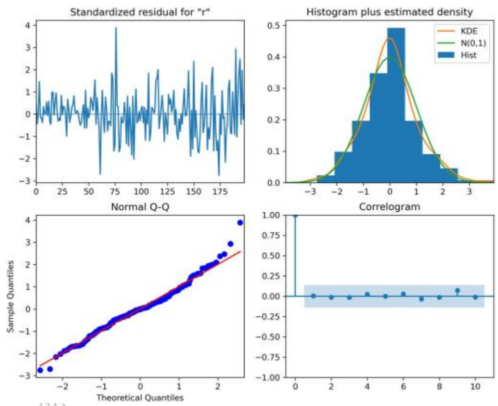
Non-stationary time series with seasonality and external variables

- **Warning: what if you wish to predict two timesteps into the future?** While this is possible with a SARIMA model, the SARIMAX model requires us to forecast the exogenous variables too.
- The only way to avoid that situation is **to predict only one timestep into the future and wait to observe the exogenous variable before predicting the target for another timestep into the future**.
- On the other hand, if your exogenous variable is easy to predict, meaning that it follows a known function that can be accurately predicted, there is no harm in forecasting the exogenous variable and using these forecasts to predict the target.

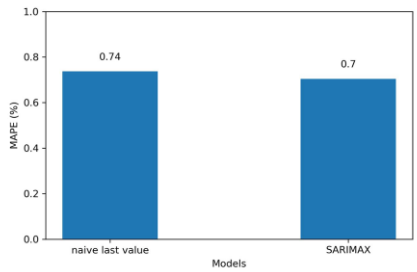
University of Pisa

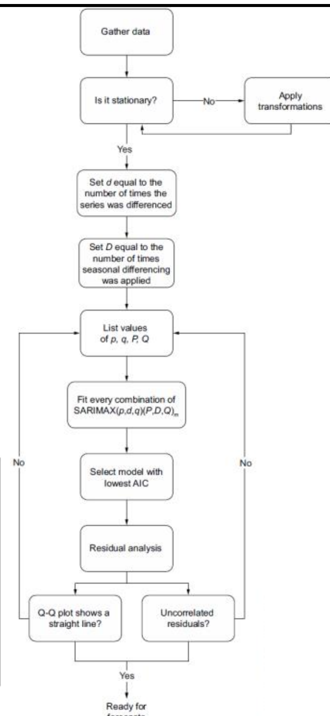
Non-stationary time series with s and external variables

- **General procedure**
  - No real changes with respect to SARIMA procedure, except for the use of SARIMAX as fitting model



Optimal Model:
ARIMAX(3,1,3)
No seasonability





65