Data-Science 1

forecasting

Karel de Grote Hogeschool

26/04/2023

Inhoud

- voorbeeld
- op basis van het verleden
- betrouwbaarheid van voorspelling
- op basis van een model

Voorbeeld

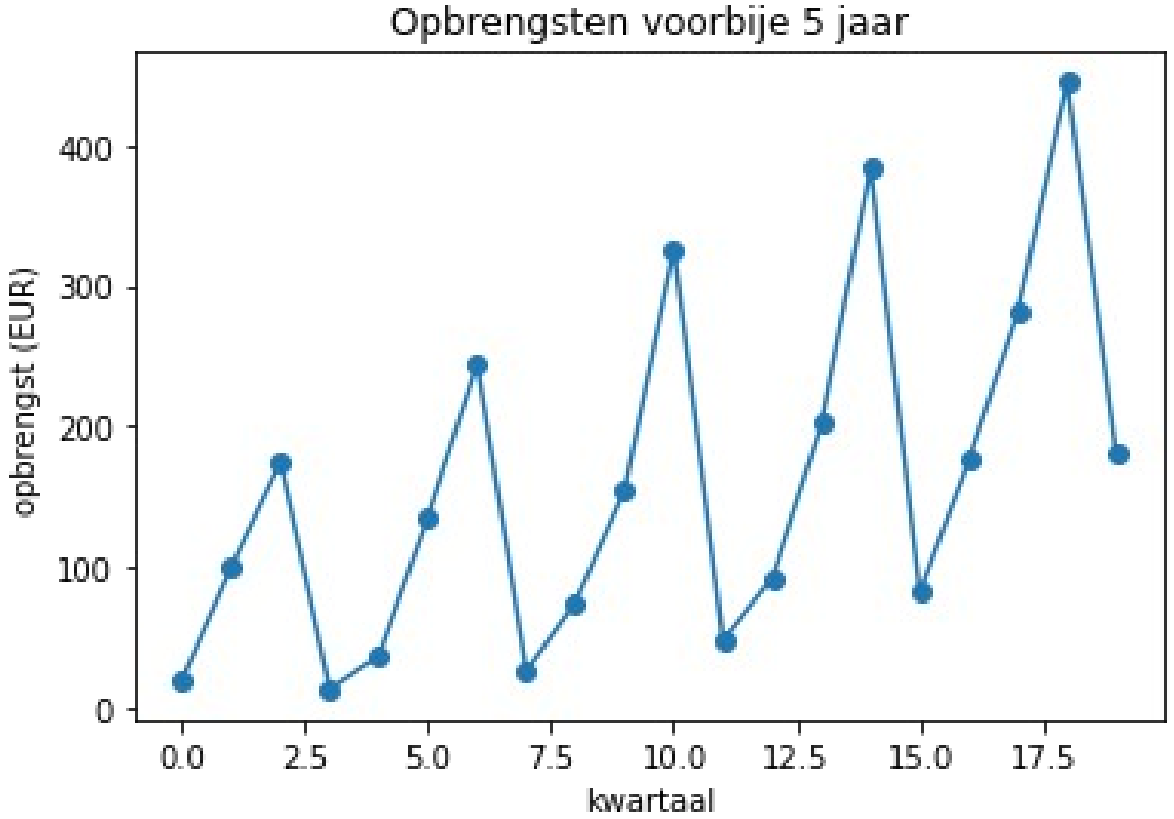
Voorbeeld voor deze les

- gegeven: opbrengsten van een bedrijfje van de laatste 5 jaar (per kwartaal):

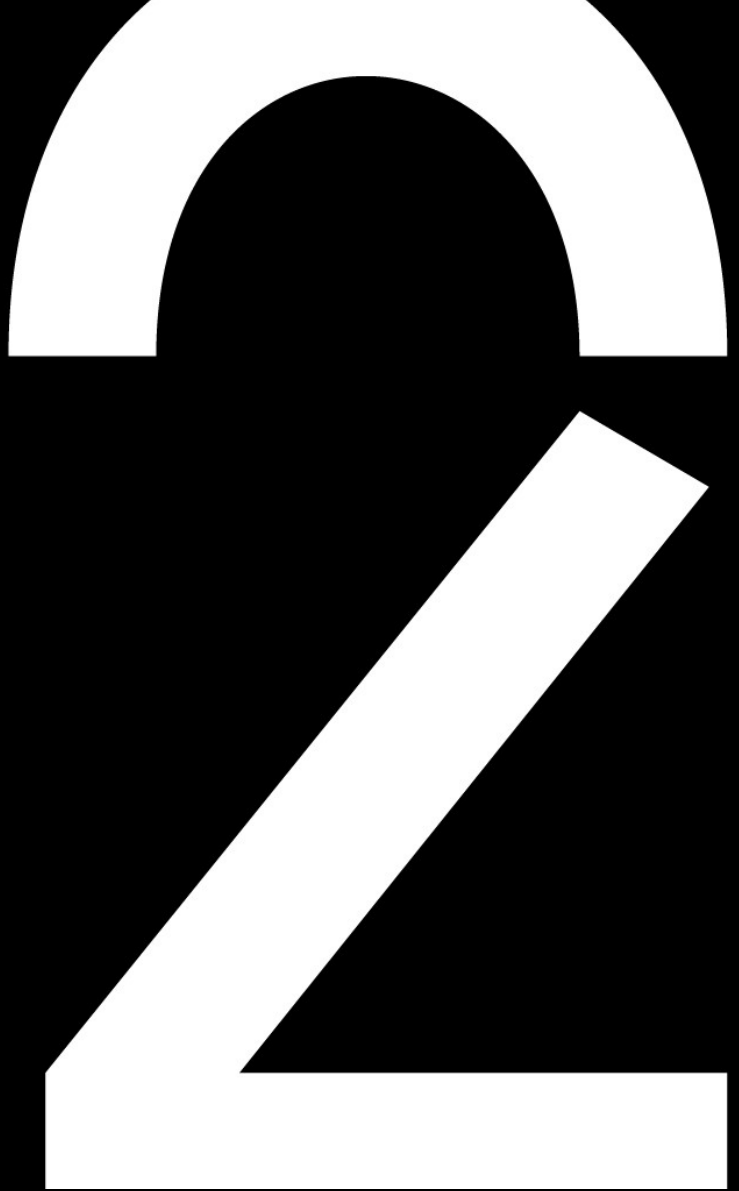
  opbrengsten = [20, 100, 175, 13, 37, 136, 245, 26, 75, 155, 326, 48, 92, 202, 384, 82, 176, 282, 445, 181]

- gevraagd: opbrengsten voor volgend jaar (volgende 4 waarden)

Scatterplot



Op basis van het verleden



Inhoud

- naïeve forecasting
- gemiddelde
- voortschrijdend gemiddelde
- lineaire combinatie

Notatie

- aantal waarden = n
- gemeten waarden = x_i (i = 0 tot n-1)
- voorspelde waarden = f_i (i = n tot ...)
- voorbeeld: n = 3:

$$x_0, x_1, x_2, f_3, f_4, \ldots$$

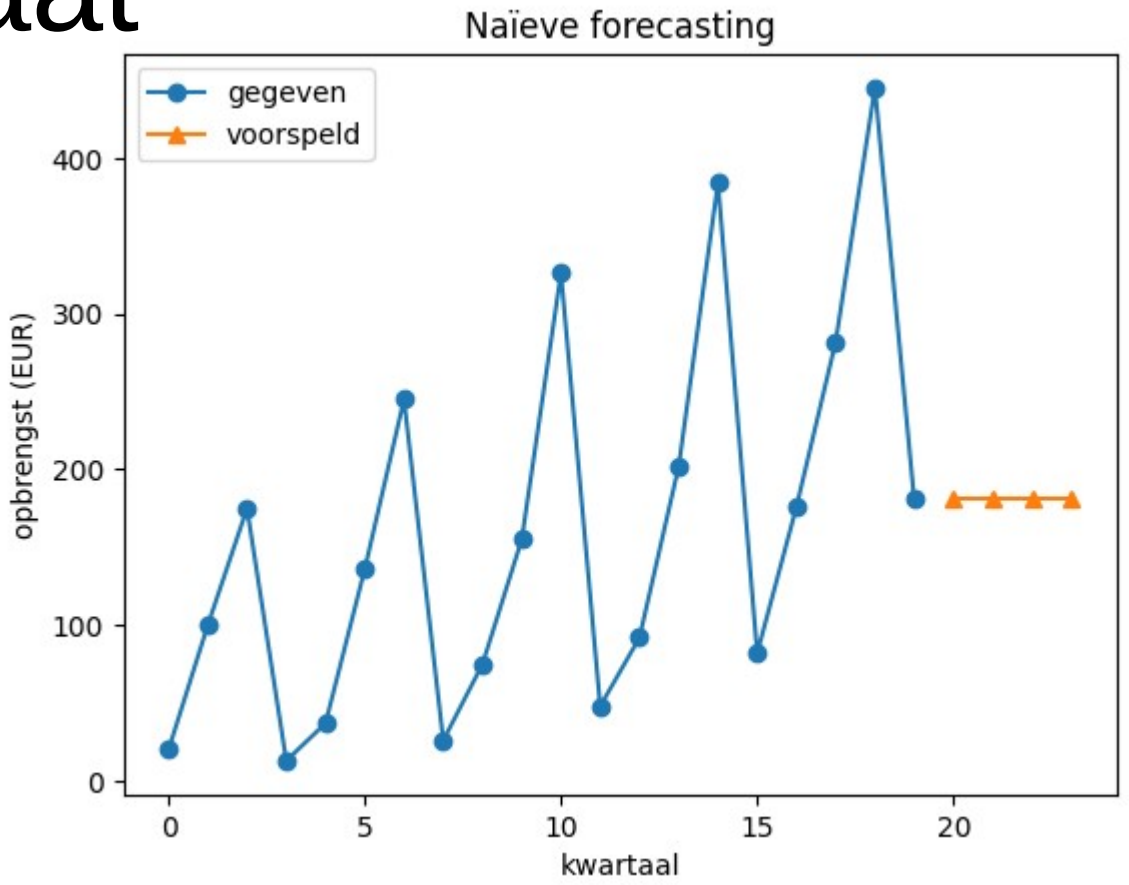
Naïeve forecasting

- volgende waarde = laatste waarde

$$f_n = x_{n-1}$$

- zie python

Resultaat


Naïeve forecasting
gegeven
voorspeld
opbrengst (EUR)
0
100
200
300
400
0
5
10
15
20
kwartaal

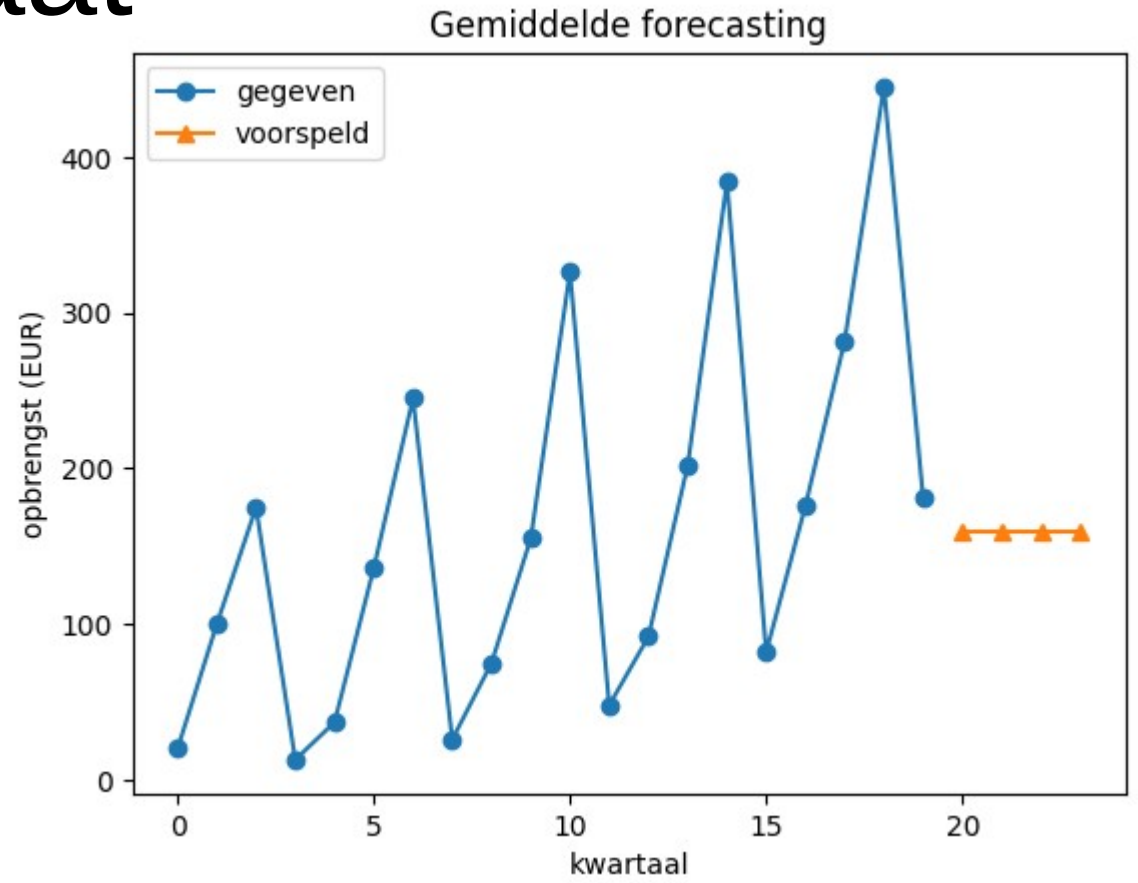
Gemiddelde forecasting

- volgende waarde = gemiddelde van alle voorgaande waarden

$$f_n = \frac{1}{n} \sum_{i=0}^{n-1} x_i$$

- zie python

Resultaat


Gemiddelde forecasting
gegeven
voorspeld
opbrengst (EUR)
0
100
200
300
400
0
5
10
15
20
kwartaal

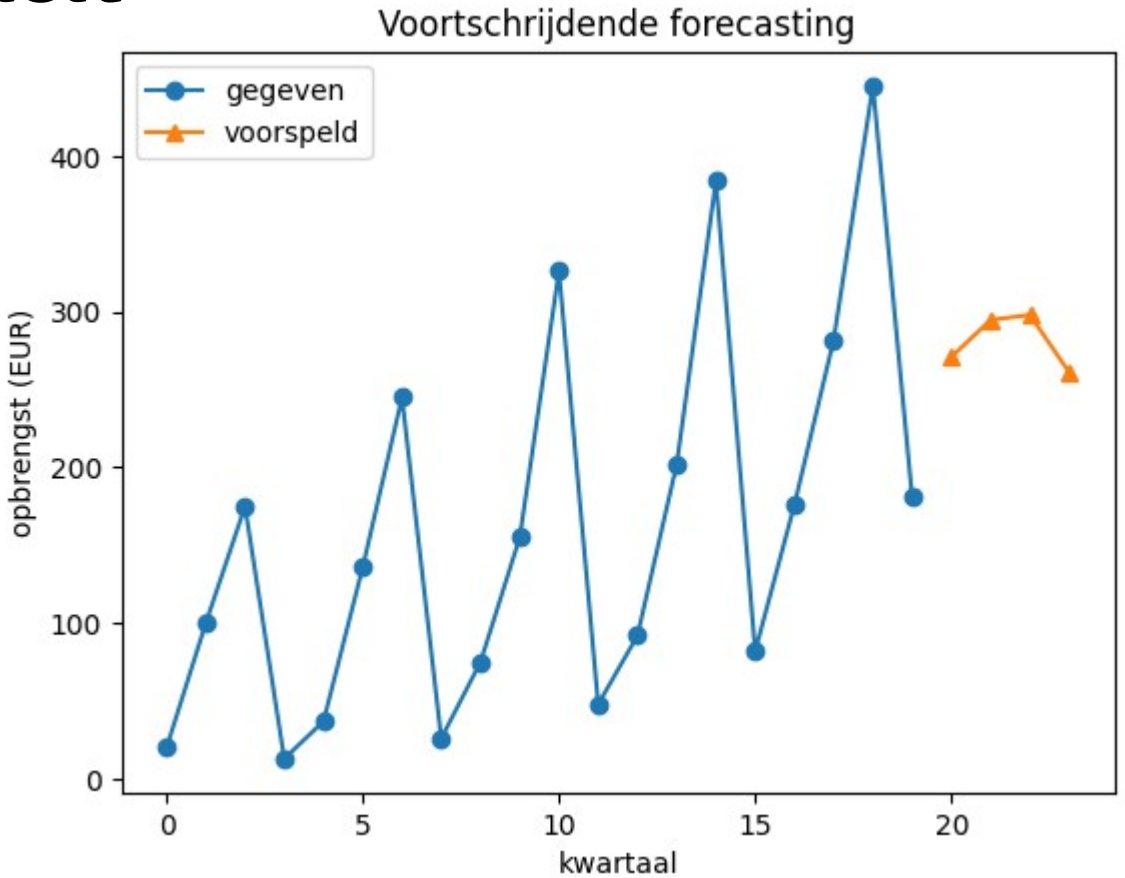
Voortschrijdend gemiddelde

- volgende waarde = gemiddelde van de laatste m waarden

$$f_n = \frac{1}{m} \sum_{i=n-m}^{n-1} x_i$$

- opm: als m=1, dan is dit naïeve forecasting, als m=n, dan is dit het gemiddelde van alle voorgaande waarden
- zie python

Resultaat


Voortschrijdende forecasting
gegeven
voorspeld
opbrengst (EUR)
kwartaal
0
100
200
300
400
0
5
10
15
20

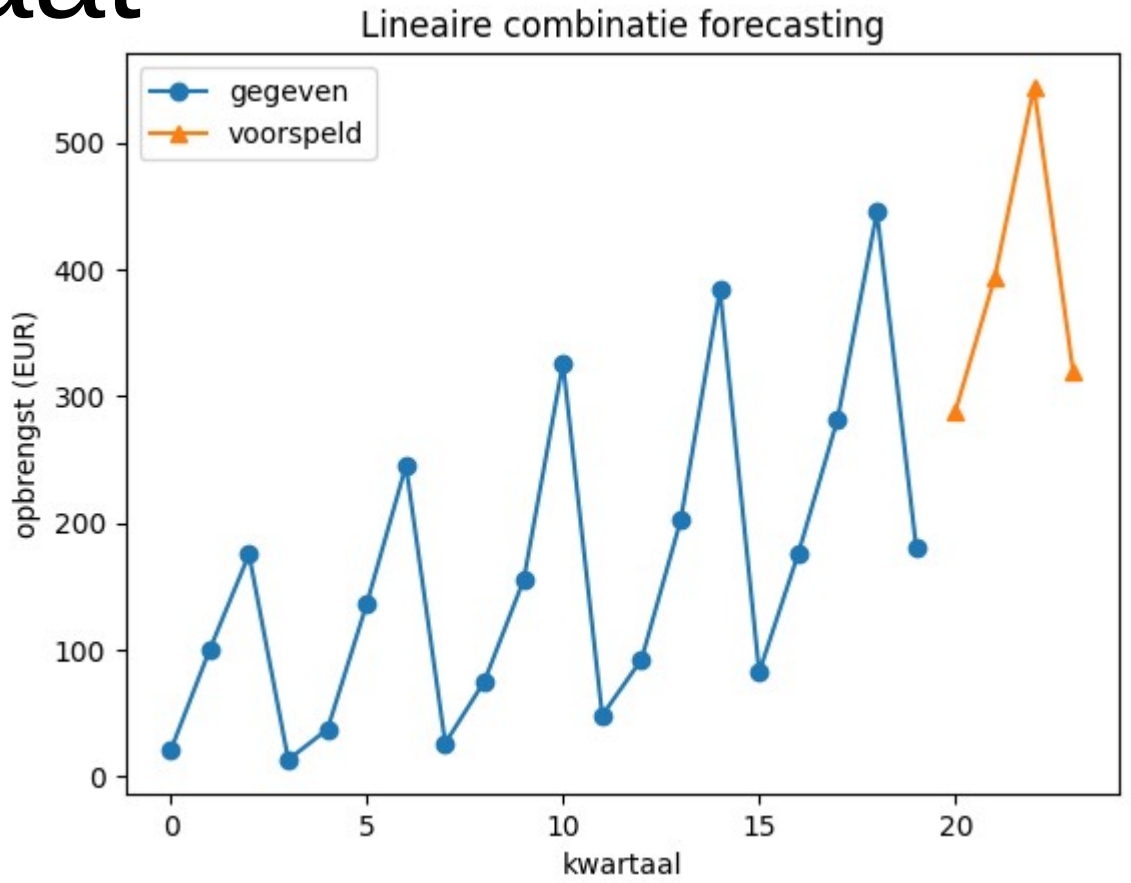
Lineaire combinatie

- volgende waarde = gewogen gemiddelde van m voorgaande waarden

$$f_n = \sum_{i=n-m}^{n-1} a_{i-n+m} \cdot x_i$$

- opm.: als a_i = 1/m, dan is dit het voortschrijdend gemiddelde
- men noemt dit een "lineaire combinatie" van de m voorgaande waarden
- je kan de waarden van a_i berekenen aan de hand van de laatste 2m waarden (zie python)

Resultaat


Lineaire combinatie forecasting
gegeven
voorspeld
opbrengst (EUR)
kwartaal
0
100
200
300
400
500
0
5
10
15
20

Betrouwbaarheid

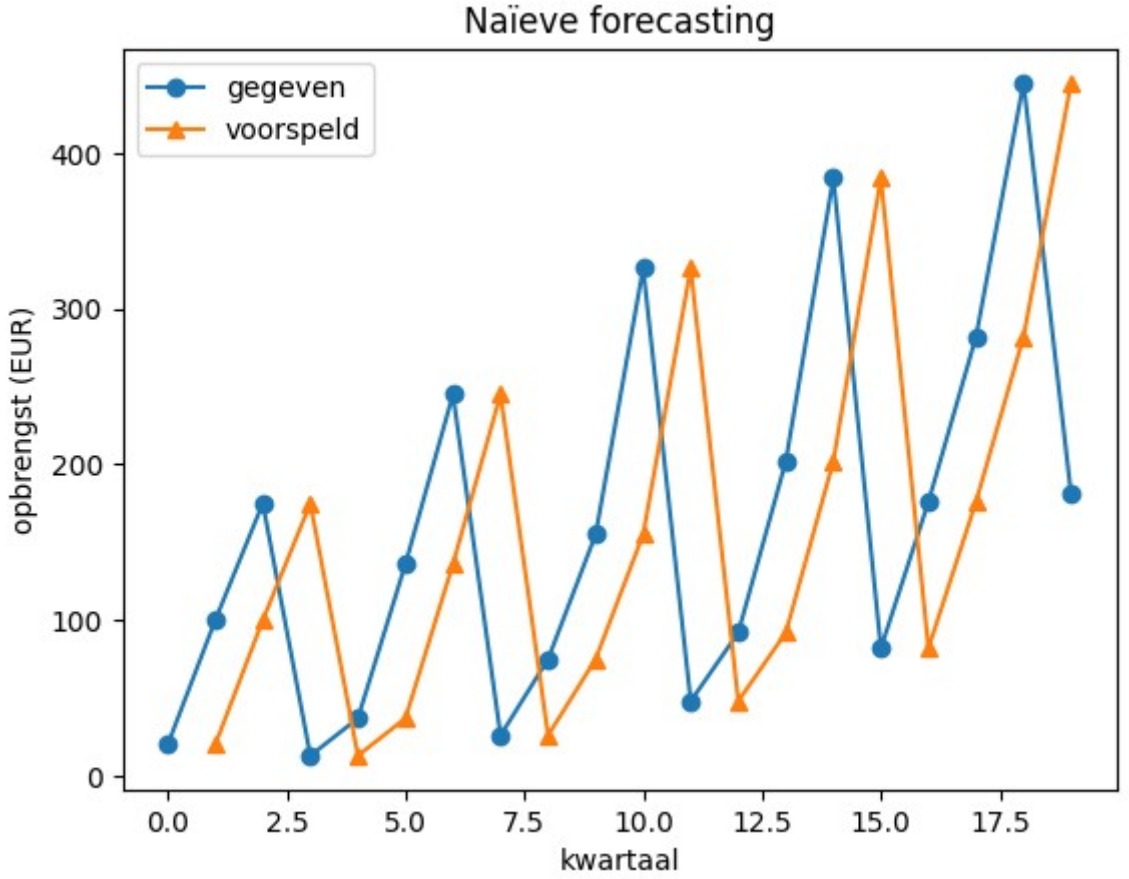
Betrouwbaarheid

- hoe betrouwbaar is een voorspelling?
- moeilijk te bepalen: je kent de toekomst niet
- oplossing?

Betrouwbaarheid

- gebruik de voorspellingsmethode in het verleden
- bepaal dus de f_i voor i=1 tot n
- bepaal dan de fout $e_i = x_i - f_i$
- je kan nu de gemiddelde fout berekenen

Betrouwbaarheid



Betrouwbaarheid

- gebruik de voorspellingsmethode in het verleden
- bepaal dus de f_i voor i=1 tot n
- bepaal dan de fout $e_i = x_i - f_i$
- je kan nu de gemiddelde fout berekenen

Drie maten

- Mean Absolute Error: $MAE = \frac{1}{n}\sum |e_i|$
- Root Mean Squared Error: $RMSE = \sqrt{\frac{1}{n}\sum e_i^2}$
- Mean Absolute Percentage Error: $MAPE = \frac{1}{n}\sum \left|\frac{e_i}{x_i}\right|$

Resultaten

voorspeller	MAE	RMSE	MAPE
naïef	137,4211	158,6978	2,070257
gemiddelde	103,0048	130,8062	1,03622
voortschrijdend gemiddelde	92,90625	113,3213	0,77707
lineaire combinatie	28,26923	32,7929	0,1742834

In Python

- zie code

4

Op basis van een model

Inhoud

- trend
- seasonal decomposition
- (seasonal trend forecasting)

Trend forecasting

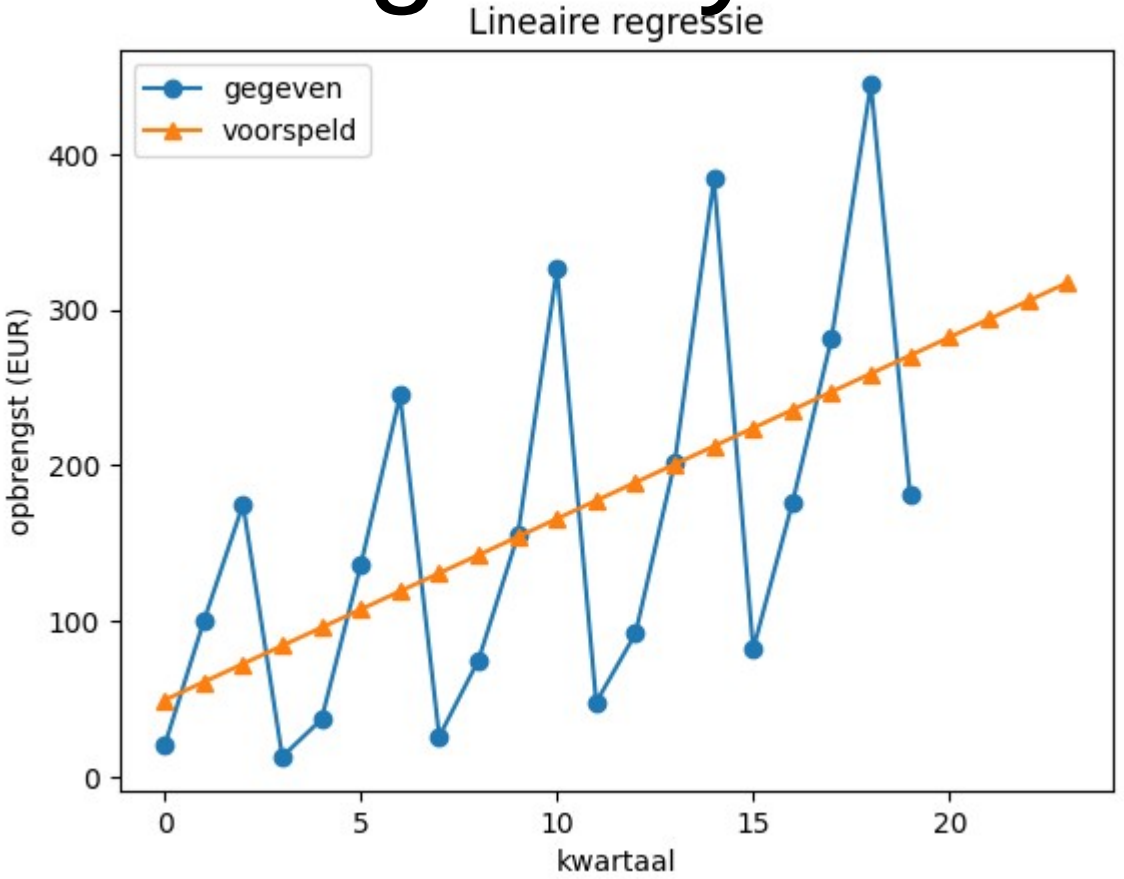
- trend forecasting = (lineaire) regressie
- dus: zoek lijn door de punten en gebruik die als model
- in ons voorbeeld:

$$f_i = 49{,}3286 + 11{,}6496 \cdot i$$

- het resultaat is dus een functie (en niet een volgende waarde)

Trend forecasting in Python

- zie code



Betrouwbaarheid

- verschil: we gebruiken nu het model om het verleden te voorspellen
- resultaten (zie python):
  - MAE = 85
  - RMSE = 100,6167 (dit is ook de s_e!)
  - MAPE = 1,149211

Seasonal forecasting

- data bestaat dikwijls uit:
  - (lineaire) trend
  - weerkerend patroon ("seizoen")
  - ruis
- we kunnen de data terug ontleden en deze 3 factoren identificeren = seasonal decomposition

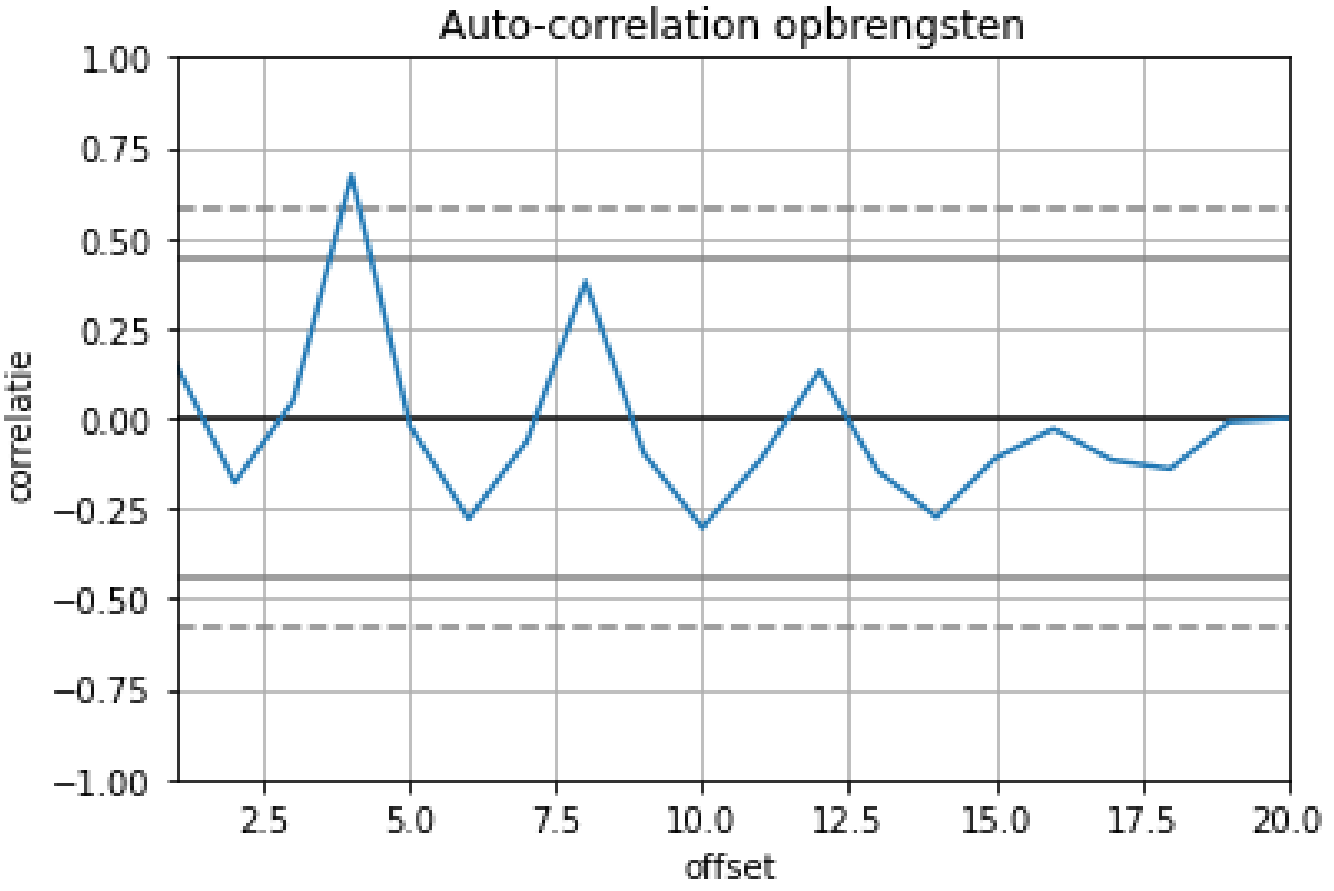
Seasonal decomposition

- dus ("additief")
  - data = trend + seizoen + ruis
  - $x_i = T(i) + S(i) + R(i)$
- kan ook "multiplicatief"
  - $x_i = T(i) \cdot S(i) \cdot R(i)$

Bepalen vd seizoensgrootte

- om het seizoen te kunnen identificeren, moeten we weten hoe groot dit is (m)
- je kan dit visueel meestal zien
- ook mogelijk met "auto correlation function"

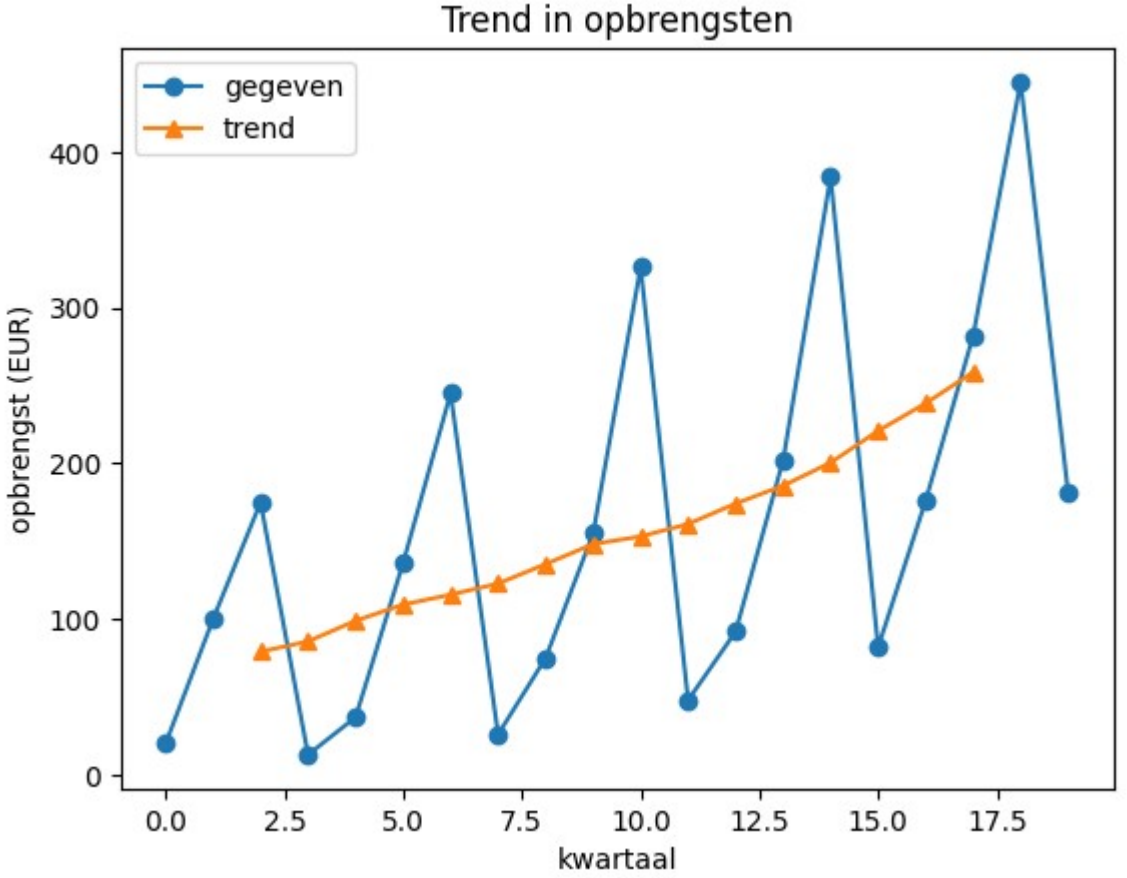
ACF


Auto-correlation opbrengsten
correlatie
offset
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

Bepalen van de trend

- trend = data - seizoen - ruis
- seizoen en ruis hebben een hoge frequentie
- filter de lage frequenties uit de data
  - dit doe je door het voortschrijdend gemiddelde te nemen van m waarden

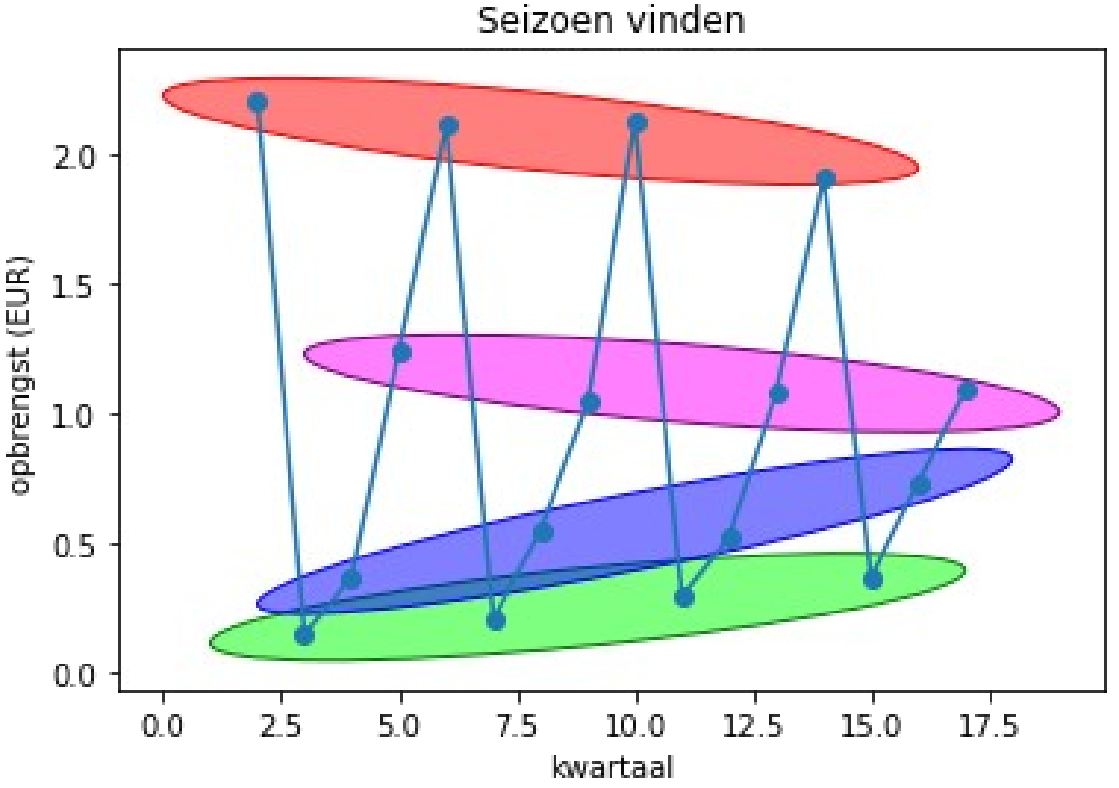
Bepalen van de trend



Seizoen bepalen

- trek de trend af van de data
  - resultaat = enkel seizoen en ruis
  - data - trend = seizoen + ruis
- seizoen is wederkerend patroon van steeds m punten

Seizoen bepalen



Ruis vinden

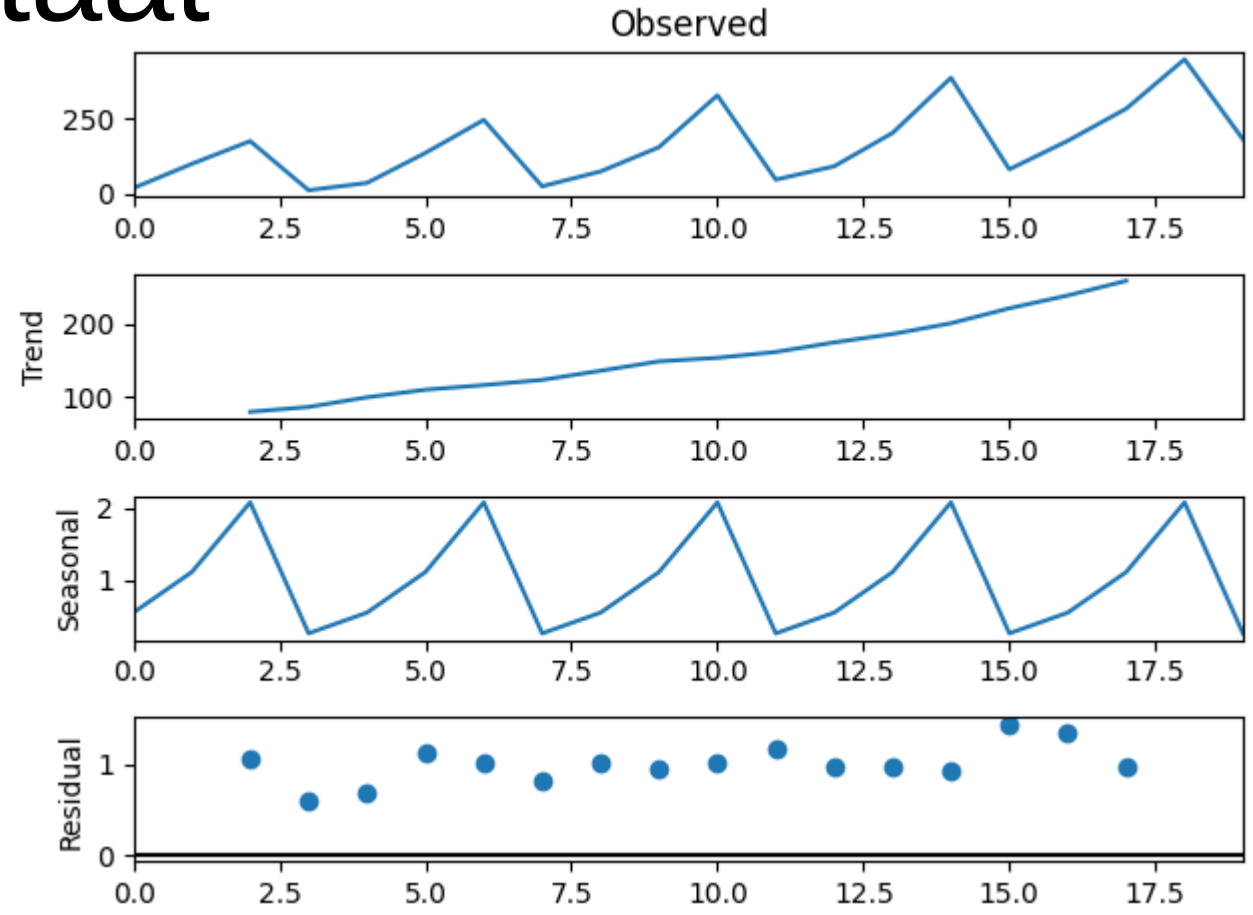
- ruis = data - trend - seizoen
- ruis kan je niet voorspellen in de toekomst
- de standaardafwijking van de ruis is wel een indicatie voor de kwaliteit van het model (dit is de RMSE als het model additief is!)
- we gebruiken daarom als model:

  forecast = trend + seizoen

Voorbeeld in Python

- zie code
  - in dit geval "multiplicatief"
  - data / trend = seizoen * ruis
  - ruis = data / trend / seizoen

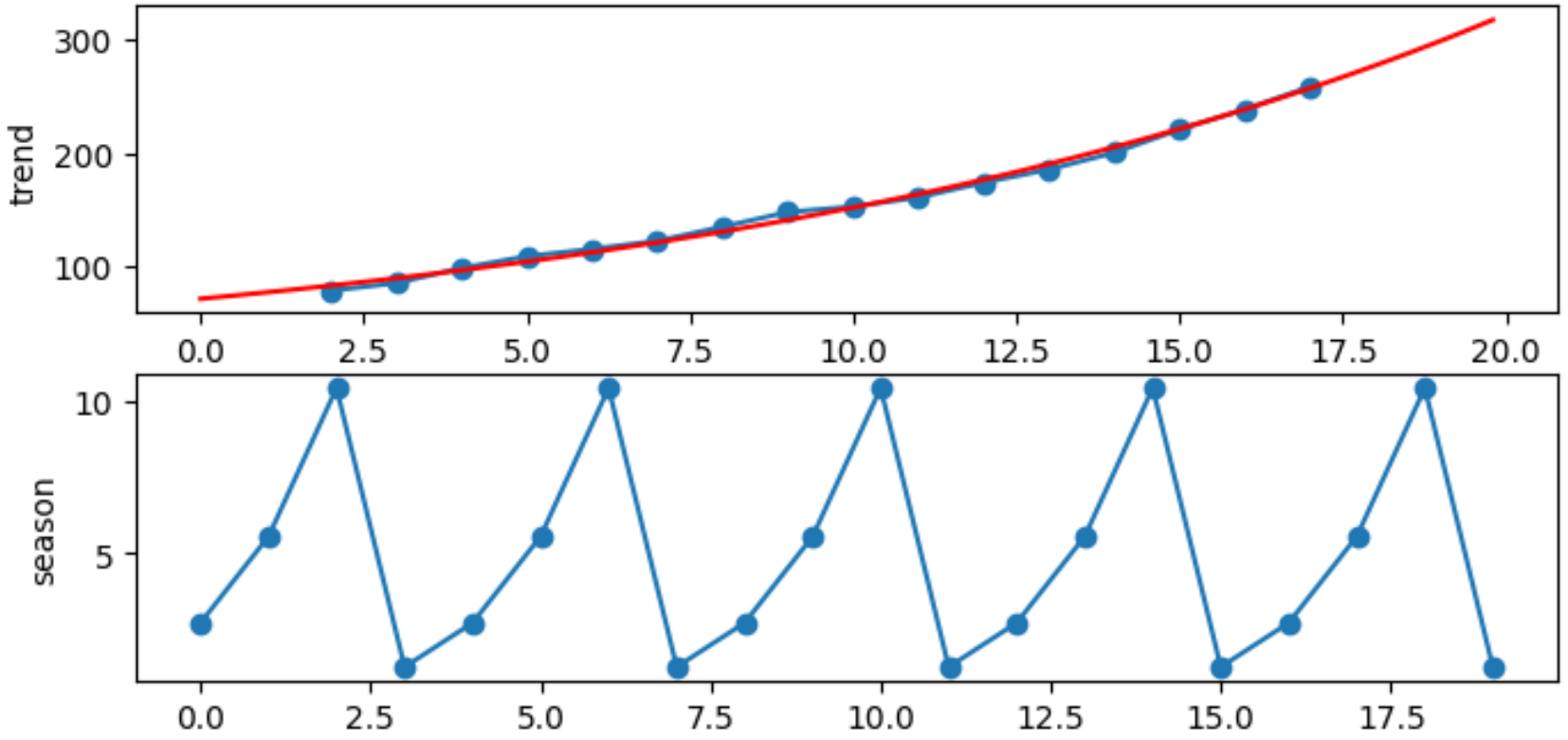
Resultaat


Observed
250
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Trend
200
100
Seasonal
2
1
Residual
1
0

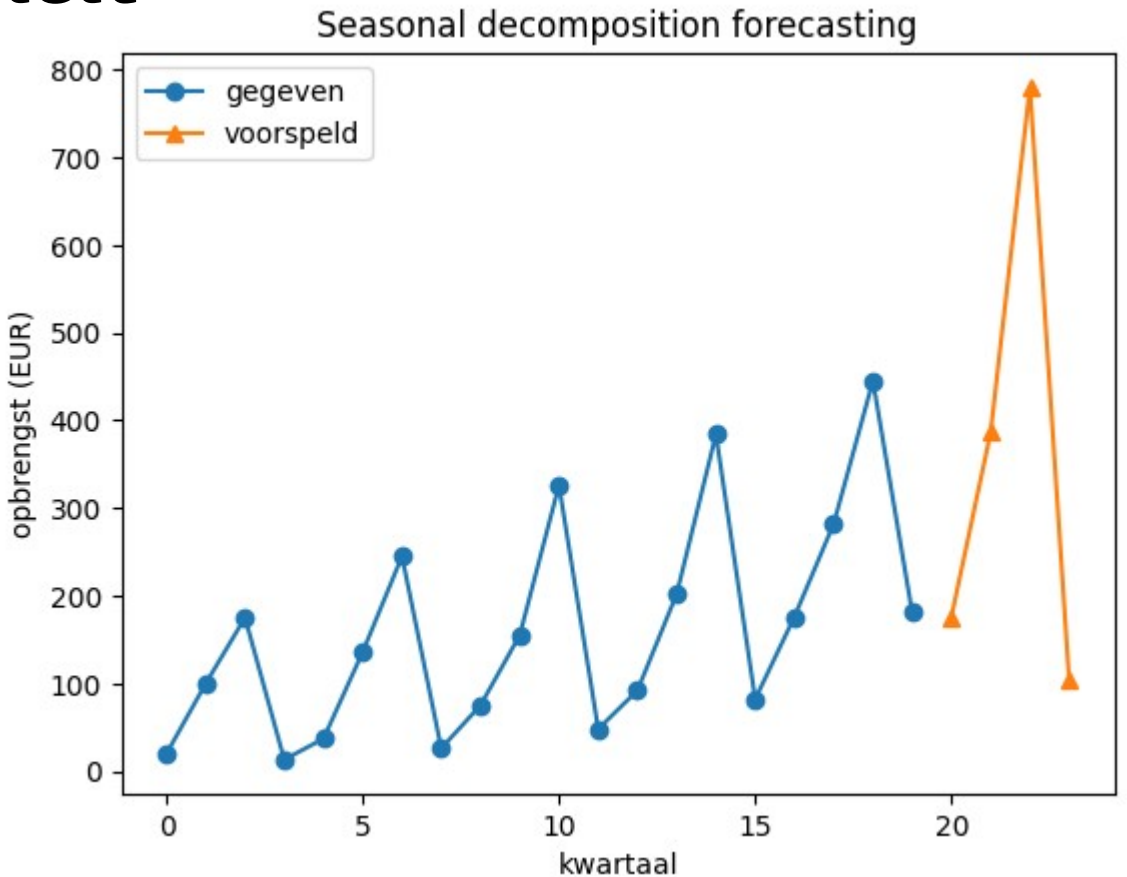
Voorspellingen maken

- zoek een formule voor de trend (bv met regressie)
- zet het patroon gewoon verder
- twee mogelijkheden
  - additief: voorspelde waarde = trend + patroon
  - multiplicatief: voorspelde waarde = trend * patroon

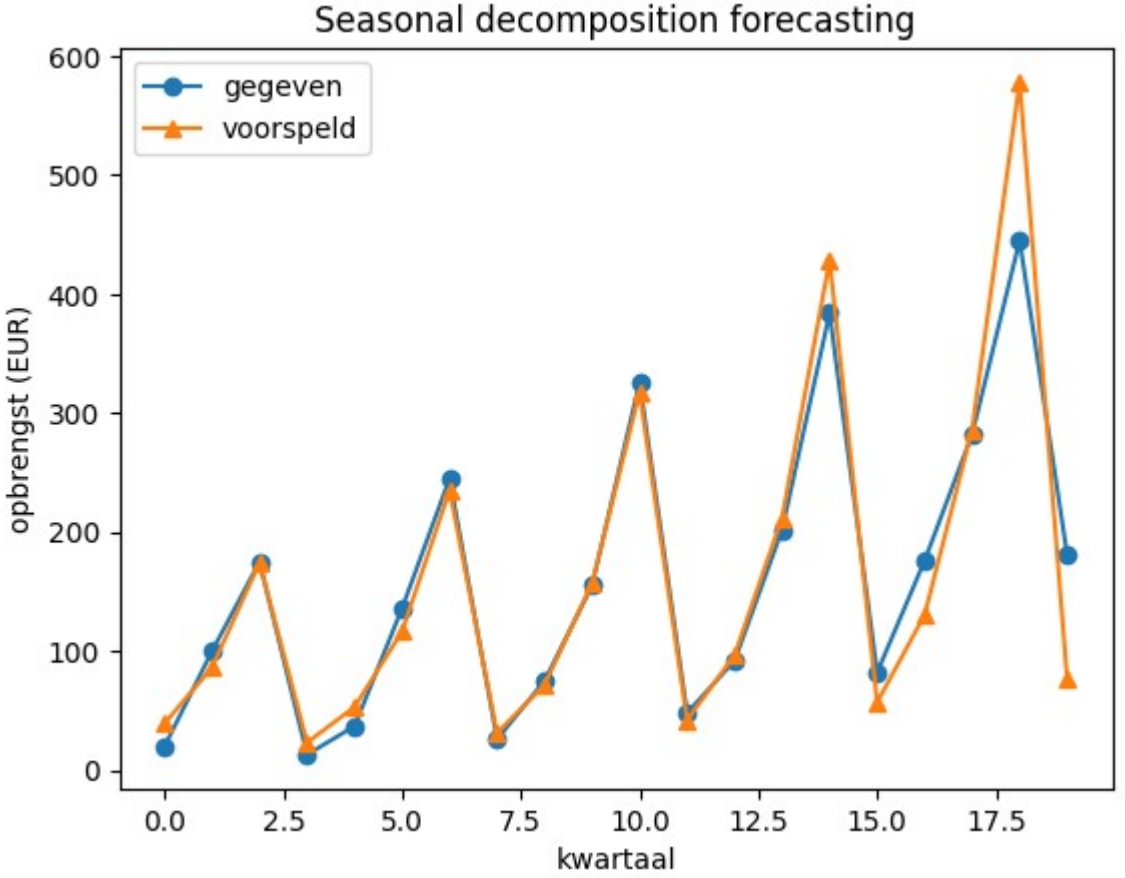
Voorspellingen maken



Resultaat


Seasonal decomposition forecasting
gegeven
voorspeld
opbrengst (EUR)
kwartaal
0
100
200
300
400
500
600
700
800
0
5
10
15
20

Betrouwbaarheid



Betrouwbaarheid

- resultaat (zie python)
  - MAE = 24,29
  - RMSE = 41,81
  - MAPE = 0,229

Oefeningen

Oefeningen

- forecasting
  - populariteit app
  - pretpark
- call center