Data Science 2

Zeitreihenanalyse

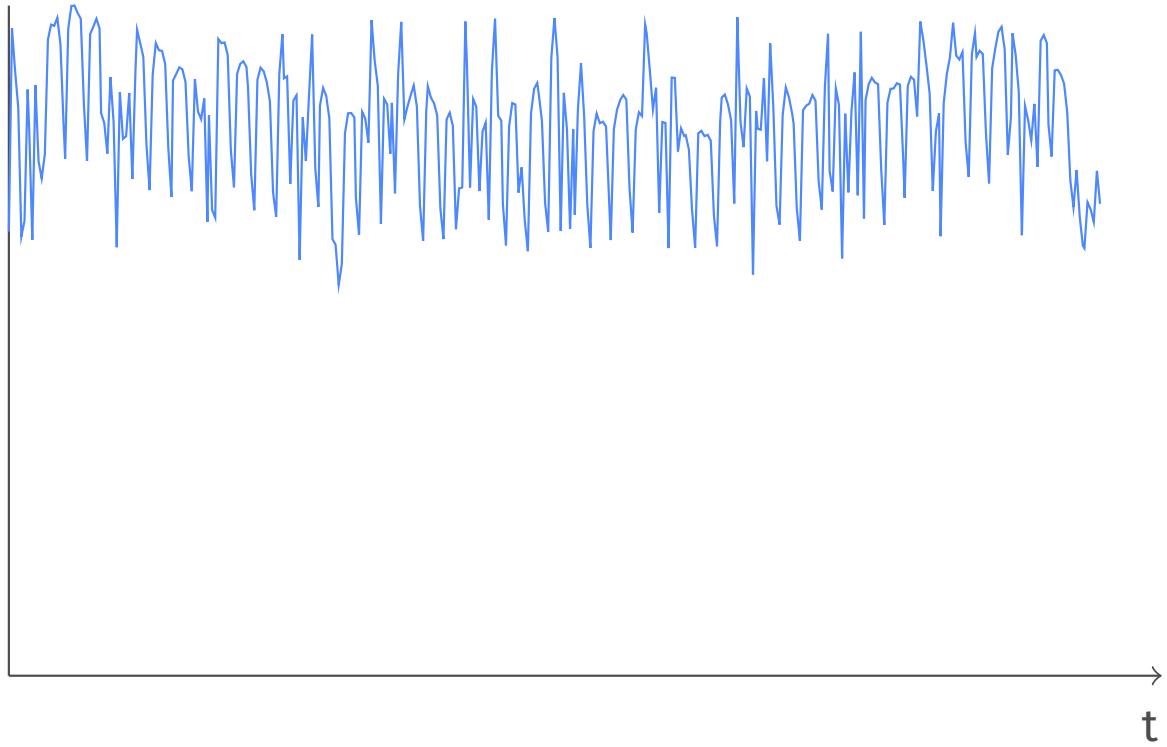
Prof. Dr. Christian Bockermann

Hochschule Bochum

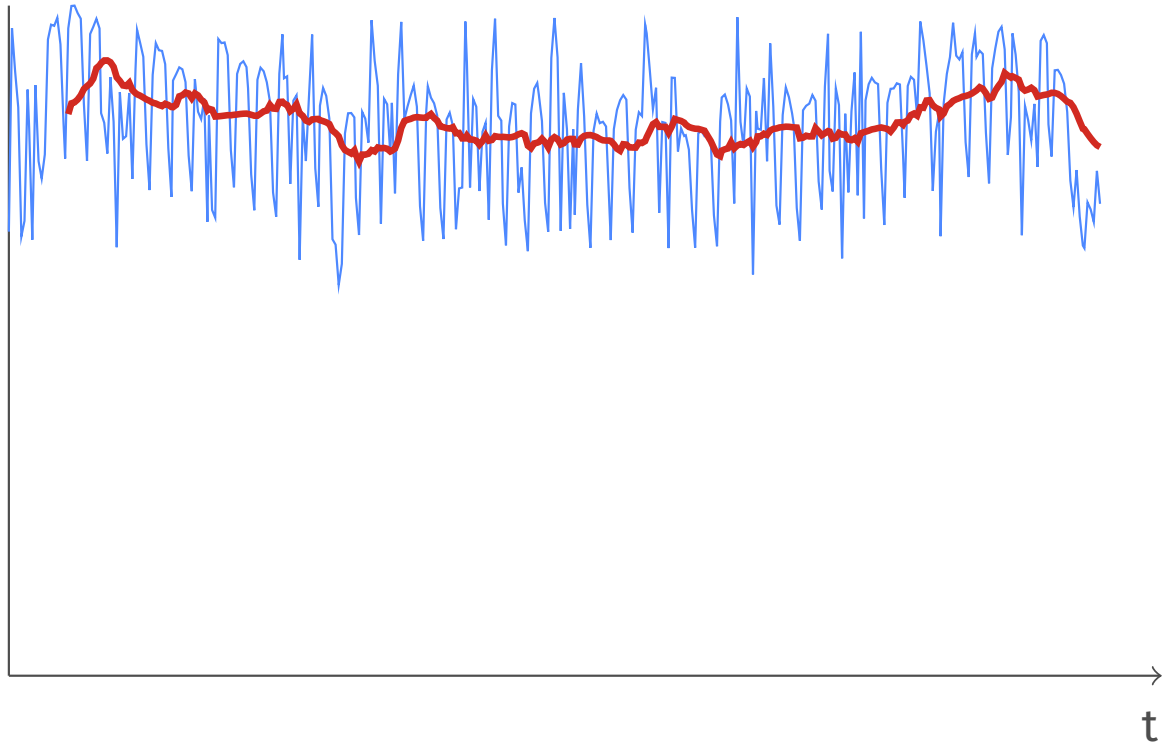
Sommersemester 2024

Inhalt

Beispiel: Stromverbrauch von DE 2019, tägliches Mittel

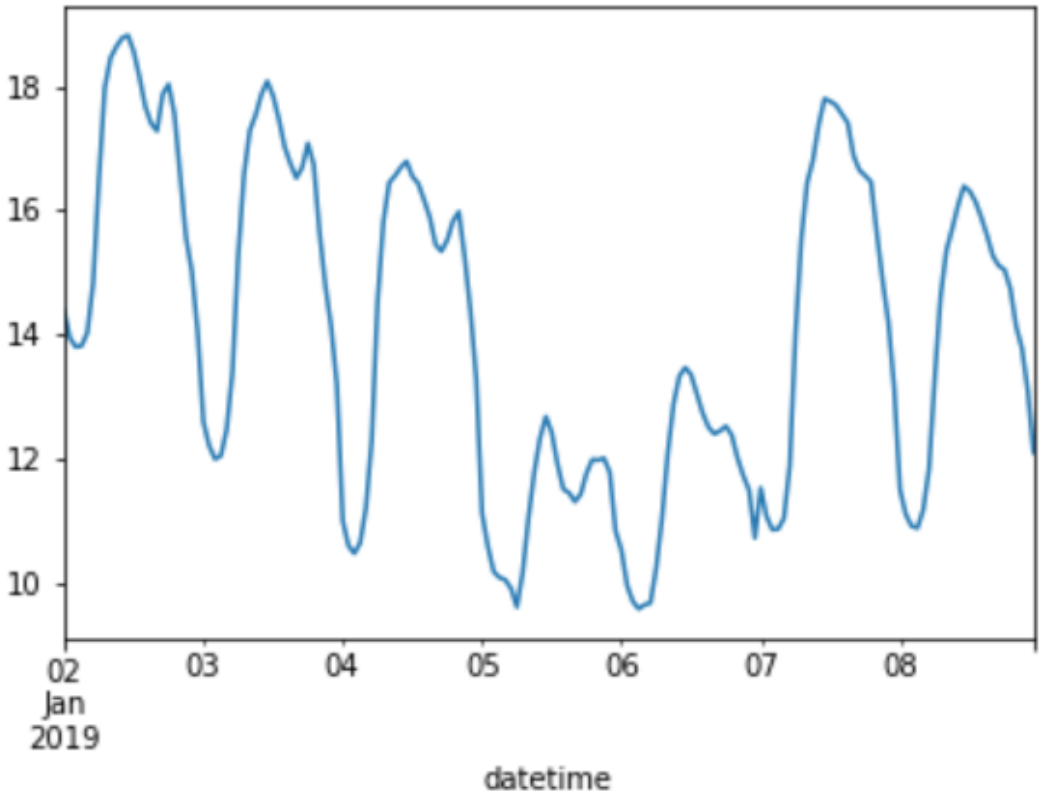


Beispiel: Stromverbrauch von DE 2019, tägliches Mittel

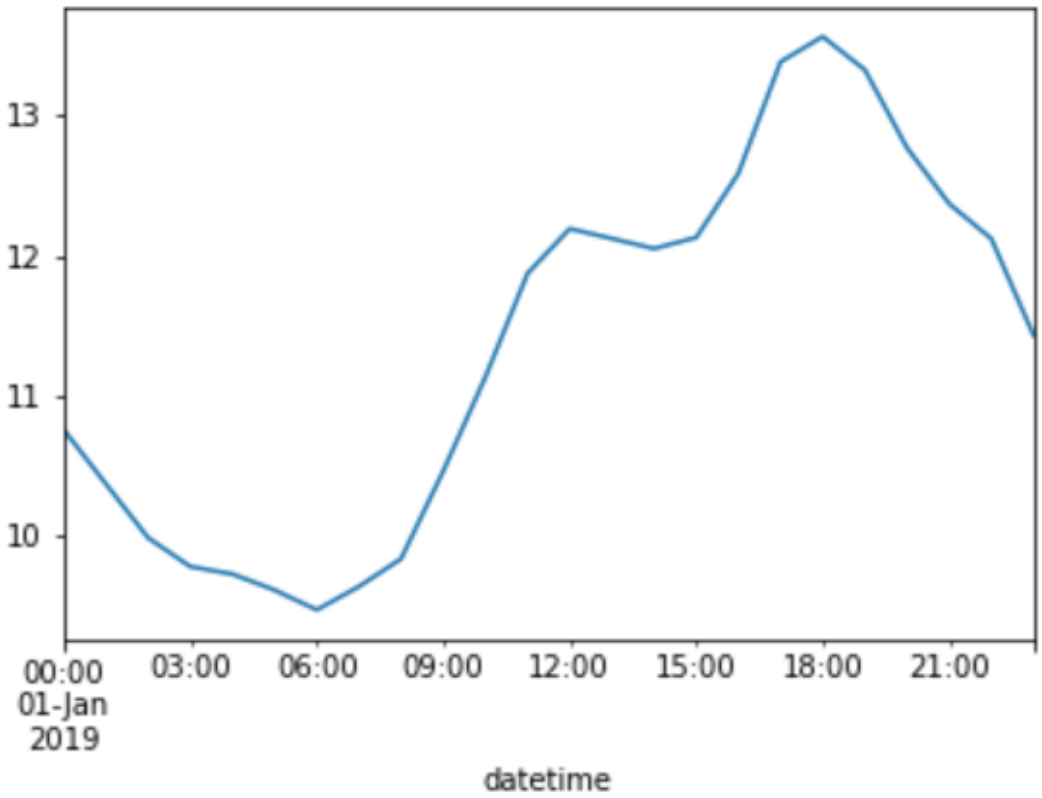


Gleitender 21-Tage Durchschnitt

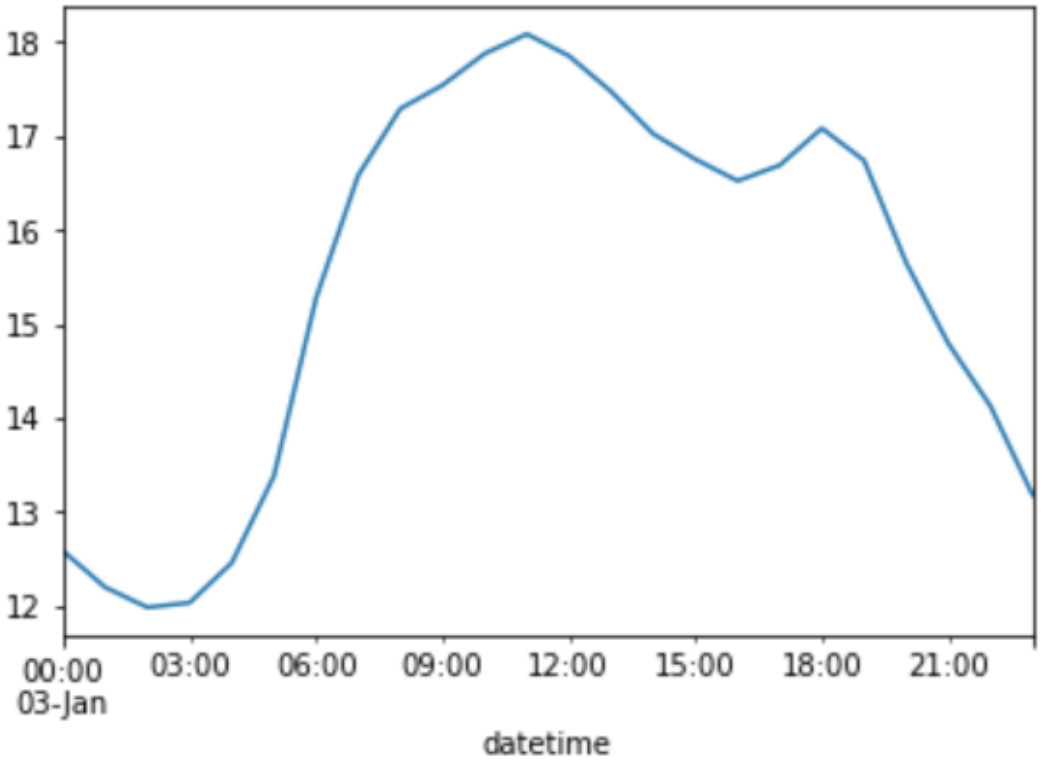
Beispiel: Stromverbrauch von DE, 2019 KW 1



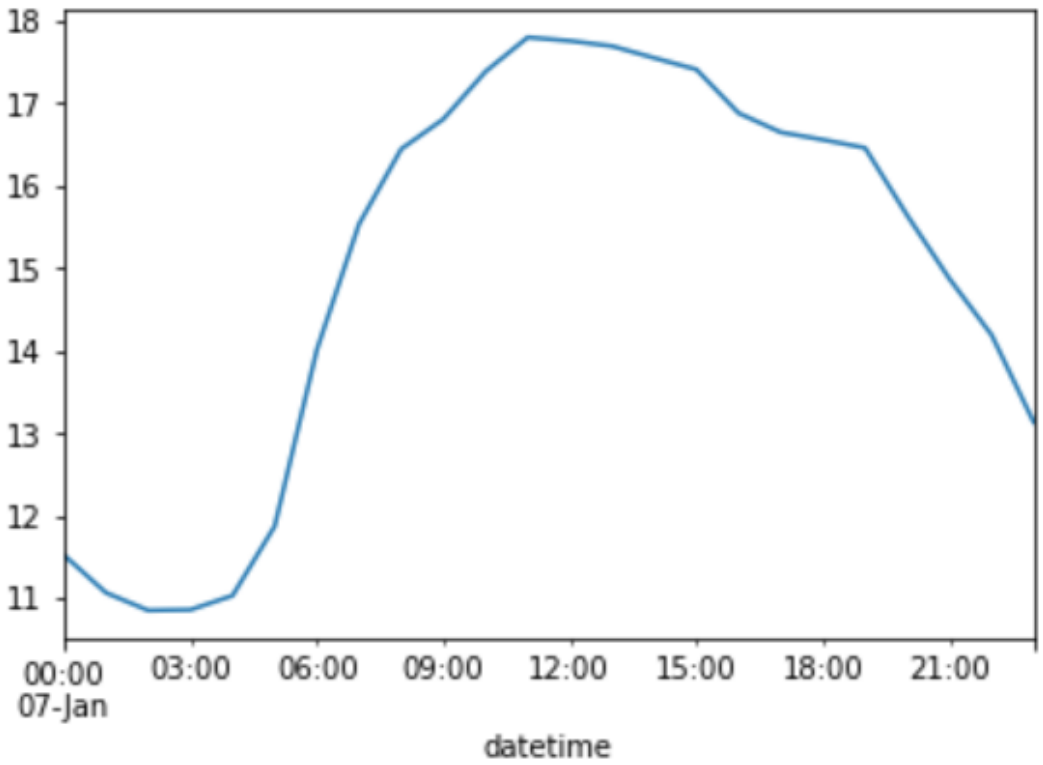
Beispiel: Stromverbrauch von DE, 1.1.2019



Beispiel: Stromverbrauch von DE, 3.1.2019



Beispiel: Stromverbrauch von DE, 7.1.2019



Bisher: “Zeilenweises” Lernen

$X_{Hubraum}$	$X_{Leistung}$	$X_{Zylinder}$	$X_{Gewicht}$	$X_{Beschl.}$	$Y_{l/100km}$
307.0	130	8	3504	12.0	15.7
350.0	165	8	3693	11.5	18.8
318.0	150	8	3436	11.0	15.7
304.0	140	8	3433	12.0	17.6

Bisher: “Zeilenweises” Lernen

$X_{Hubraum}$	$X_{Leistung}$	$X_{Zylinder}$	$X_{Gewicht}$	$X_{Beschl.}$	$Y_{l/100km}$
307.0	130	8	3504	12.0	15.7
350.0	165	8	3693	11.5	18.8
318.0	150	8	3436	11.0	15.7
304.0	140	8	3433	12.0	17.6

Bisher: “Zeilenweises” Lernen

$X_{Hubraum}$	$X_{Leistung}$	$X_{Zylinder}$	$X_{Gewicht}$	$X_{Beschl.}$	$Y_{l/100km}$
307.0	130	8	3504	12.0	15.7
350.0	165	8	3693	11.5	18.8
318.0	150	8	3436	11.0	15.7
304.0	140	8			17.6

Vorhersage

Zeitreihen haben eine andere Lern-Struktur

Datum	Verbrauch
1.1.2019	14.291
⋮	⋮
3.7.2019	13.238
4.7.2019	12.425

Zeitreihen haben eine andere Lern-Struktur

Datum	Verbrauch
1.1.2019	14.291
⋮	⋮
3.7.2019	13.238
4.7.2019	12.425
⋮	⋮
27.4.2021	?

Vorhersage

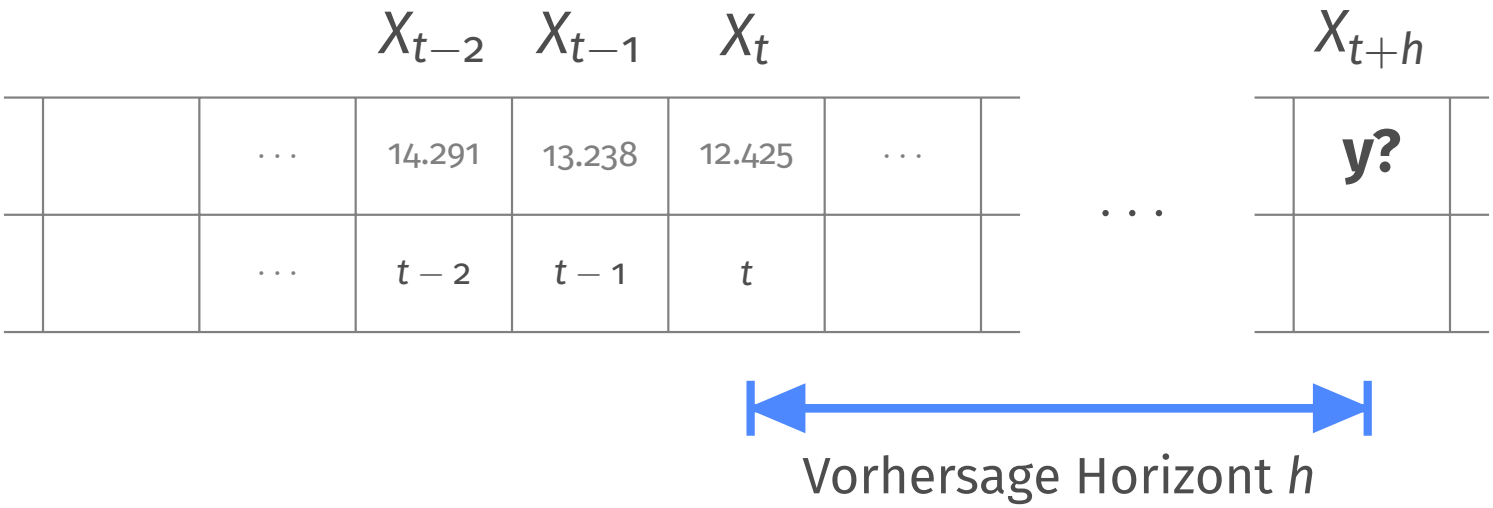
Was haben wir bei der Klassifikation/Regression gemacht?

Typisches Vorgehen:

1. Formalisierung des Problems (Lernaufgabe)
2. Charakterisierung über Funktionen
3. Funktionsklasse für Approximation finden
4. Funktion lernen und “Zukunft” vorhersagen (Training)

Forecast – Vorhersage von Zeitreihenwerten

Abstrakte Darstellung als Zeitstrahl:



Beschreibung der Zeitreihe als Modell m, Vorhersage:

$$m(t+h) = \hat{y}$$

Formalisierung: Zeitreihe

Zeitreihen als (unendliche) Folge von Zufallsvariablen

$$(X_t)_{t \in \mathbb{Z}}$$

Dabei hat jede Zufallsvariable X_t eigene Dichtefunktion $f_t(x)$.

Wie können wir derartige Folgen charakterisieren?

Charakterisierung von Zufallsvariablen

Zufallsvariable X mit Dichtefunktion f lässt sich z.B. durch Erwartungswert $E[X]$ und Varianz $Var(X)$ charakterisieren:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x)\, dx \\ Var(X) &= E[(X - E[X])^2] \end{aligned}$$

Charakterisierung von Zufallsvariablen

Zufallsvariable X mit Dichtefunktion f lässt sich z.B. durch Erwartungswert $E[X]$ und Varianz $Var(X)$ charakterisieren:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x)\, dx \\ Var(X) &= E[(X - E[X])^2] \end{aligned}$$

Übertragung auf Zeitreihen

Zeitreihen sind Folgen von Zufallsvariablen, d.h. es gibt Mittelwerte und Varianzen:

$$\begin{aligned} \mu_t = E(X_t) &= \int_{-\infty}^{\infty} x f_t(x)\, dx \ , \ t \in \mathbb{Z} \\ Var(X_t) &= E[(X_t - E[X_t])^2] \end{aligned}$$

Wie abhängig sind die Zeitpunkte voneinander?

Korrelation von Zufallsvariablen X_s, X_t lassen sich über Kovarianz beschreiben:

$$\begin{aligned} \gamma(s,t) &= Kov(X_s, X_t) \\ &= E[(X_s - \mu_s)(X_t - \mu_t)] \end{aligned}$$

Wie abhängig sind die Zeitpunkte voneinander?

Korrelation von Zufallsvariablen X_s, X_t lassen sich über Kovarianz beschreiben:

$$\begin{aligned} \gamma(s,t) &= Kov(X_s, X_t) \\ &= E[(X_s - \mu_s)(X_t - \mu_t)] \end{aligned}$$

Korrelation von X_t zu sich selbst entspricht der Varianz:

$$\begin{aligned} \gamma(t,t) &= Kov(X_t, X_t) \\ &= E[(X_t - \mu_t)(X_t - \mu_t)] \\ &= E[(X_t - E[E_t])^2] \\ &= Var(X_t) \end{aligned}$$

Stationäre Zeitreihen

Eine Zeitreihe $(X_t)_{t\in\mathbb{Z}}$ heißt schwach stationär, wenn

1. $E[X_t] = \mu_t = \mu_s$ für alle $t, s \in \mathbb{Z}$
2. $Kov(X_t, X_{t+h}) = Kov(X_0, X_h)$ für alle $t, h \in \mathbb{Z}$

Stationäre Zeitreihen

Eine Zeitreihe $(X_t)_{t\in\mathbb{Z}}$ heißt schwach stationär, wenn

1. $E[X_t] = \mu_t = \mu_s$ für alle $t, s \in \mathbb{Z}$
2. $Kov(X_t, X_{t+h}) = Kov(X_0, X_h)$ für alle $t, h \in \mathbb{Z}$

Eine schwach stationäre Zeitreihe verhält sich bei konstantem Abstand gleich, d.h. Meßwerte von 12 Uhr und 13 Uhr, verhalten sich genauso wie die von 18 Uhr und 19 Uhr.

Stationäre Zeitreihen

Eine Zeitreihe $(X_t)_{t\in\mathbb{Z}}$ heißt schwach stationär, wenn

1. $E[X_t] = \mu_t = \mu_s$ für alle $t, s \in \mathbb{Z}$
2. $Kov(X_t, X_{t+h}) = Kov(X_0, X_h)$ für alle $t, h \in \mathbb{Z}$

Eine schwach stationäre Zeitreihe verhält sich bei konstantem Abstand gleich, d.h. Meßwerte von 12 Uhr und 13 Uhr, verhalten sich genauso wie die von 18 Uhr und 19 Uhr.

Eine stationäre Zeitreihe hat insbesondere keinen Trend.

Beispiel: Gauß'sches weißes Rauschen

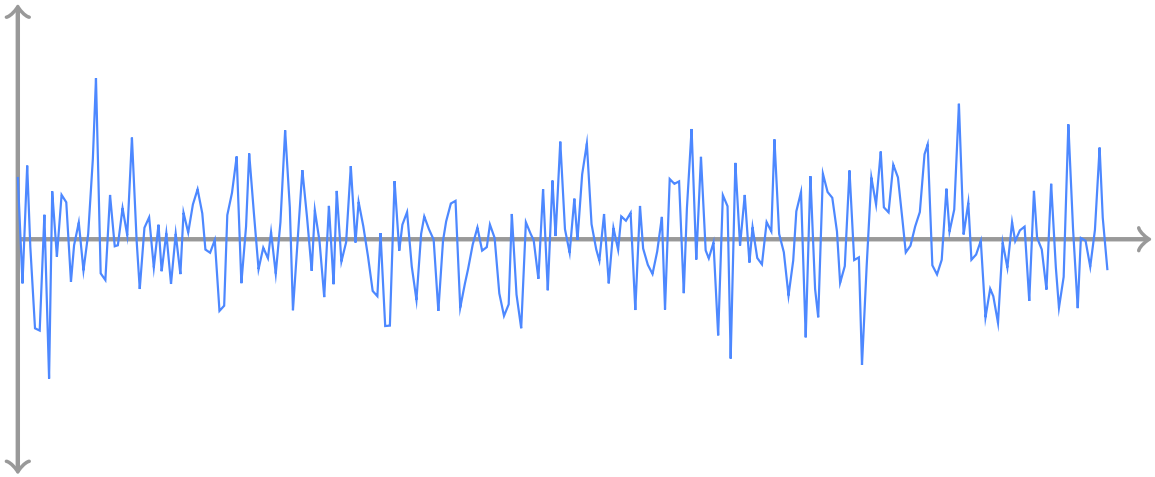
Seien X_t unabhängig normalverteilte Zufallsvariablen, also $X_t \sim N(0, \sigma^2)$, dann ist

$$E[X_t] = 0 \text{ für alle } t \in \mathbb{Z}$$

und

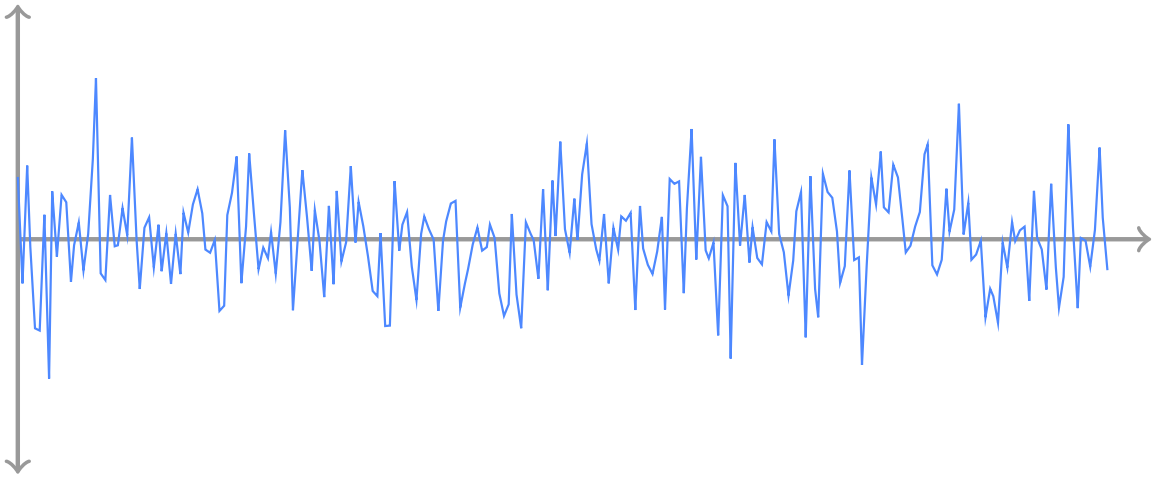
$$Kov(X_i, X_j) = \begin{cases} 0 & \text{für } i \neq j \\ \sigma^2 & \text{für } i = j \end{cases}$$

Beispiel: Gauß'sches weißes Rauschen



Völlig zufälliges Rauschen um 0, d.h. keine Vorhersage möglich.

Beispiel: Gauß'sches weißes Rauschen



Völlig zufälliges Rauschen um 0, d.h. keine Vorhersage möglich.

(Die X_t sind ja unabhängig verteilt!)

Frage: Warum stationäre Zeitreihen?

- gute mathematische Eigenschaften
- einfache Beschreibung/Analyse

In der Wirklichkeit:

- selten stationäre Zeitreihen/Prozesse
- statistische Tests für Stationarität (Dickey-Fuller)
- Transformationen von Zeitreihen in stationäre Zeitreihen

Stationäre Zeitreihen

Für stationäre Zeitreihen X_t können wir

1. den Mittelwert $\mu = \mu_t$ schätzen als

$$\mu = \overline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$$

2. die Kovarianzen schätzen mit

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \overline{X})(X_{t+h} - \overline{X}) \text{ o} \leq h \leq n.$$

Analyse von Zeitreihen

Analyse von Zeitreihen

Was brauchen wir als nächstes?

Modelle von Zeitreihen, die wir lernen können (Parameter bestimmen)

Lineare Modelle für Zeitreihen

Seien $a_t, t \in \mathbb{Z}$ unabhängige, identisch verteilte Zufallsvariablen mit $a_t \sim N(0, \sigma^2)$. Eine lineare Zeitreihe $X_1, X_2, \ldots$ ist definiert als

$$\begin{aligned} X_t &= \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots \\ &= \sum_{i=0}^{\infty} \psi_i a_{t-i} \end{aligned}$$

mit konstanten Parametern ψ_i für die gilt

$$\sum_{i=0}^{\infty} \psi_i^2 < \infty.$$

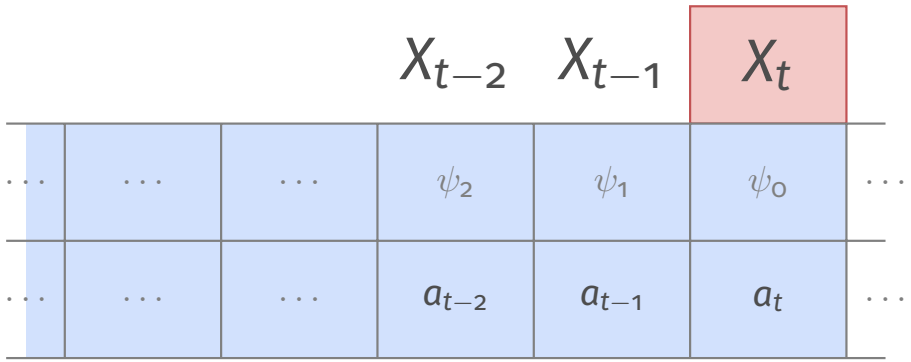
Lineare Modelle für Zeitreihe

			X_{t-2}	X_{t-1}	X_t		
...	...	...	ψ_2	ψ_1	ψ_0	...	*Parameter* ψ_i
...	...	...	a_{t-2}	a_{t-1}	a_t	...	*Grundprozeß*

X_t wird dargestellt als gewichtete Addition von weißem Rauschen:

$$X_t = \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$$

Lineare Modelle für Zeitreihe



X_t wird dargestellt als gewichtete Addition von weißem Rauschen:

$$X_t = \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$$

Lineare Modelle für Zeitreihe

			X_{t-2}	X_{t-1}	X_t		
$\cdots$	$\cdots$	$\cdots$	ψ_2	ψ_1	ψ_0	$\cdots$	*Parameter* ψ_i
$\cdots$	$\cdots$	$\cdots$	a_{t-2}	a_{t-1}	a_t	$\cdots$	*Grundprozeß*

Um ein lineares Modell zu beschreiben brauchen wir

- Verteilungsfunktion der a_t (z.B. Normalverteilung)
- Parameter $\sigma^2, \psi_0, \psi_1, \ldots$ (unendlich viele?)

Linares Modell mit unendlich vielen Parametern

Wie geben wir unendlich viele Parameter an?

Linares Modell mit unendlich vielen Parametern

Wie geben wir unendlich viele Parameter an?

Zwei Lösungsmöglichkeiten:

- Wir setzen alle bis auf ein paar auf 0.
- Wir haben eine Formel für die ψ_i.

Beispiel 1: Weißes Rauschen als lineares Modell

Wir wählen $\psi_0 = 1$ und $\psi_i = 0$ für alle $i > 0$, dann ergibt sich

$$X_t = a_t$$

Beispiel 1: Weißes Rauschen als lineares Modell

Wir wählen $\psi_0 = 1$ und $\psi_i = 0$ für alle $i > 0$, dann ergibt sich

$$X_t = a_t$$

Beispiel 2: MA(1) Modell (*Moving Average*)

Dazu sei $\psi_0 = 1, \psi_1 = 0.5$ und $\psi_i = 0$ für alle $i > 1$, dann ist

$$X_t = a_t + 0.5a_{t-1}$$

Beispiel 3: AR(1) Modell (*Auto Regression*)

Wähle $\psi_i = (0.5)^i$ für alle $i \geq 0$, dann ergibt sich

$$\begin{aligned} X_t &= (0.5)^0 a_t + (0.5)^1 a_{t-1} + (0.5)^2 a_{t-2} + \dots \\ &= a_t + 0.5 a_{t-1} + 0.25 a_{t-2} + 0.125 a_{t-3} + \dots \end{aligned}$$

Beispiel 3: AR(1) Modell (*Auto Regression*)

Wähle $\psi_i = (0.5)^i$ für alle $i \geq 0$, dann ergibt sich

$$\begin{aligned} X_t &= (0.5)^0 a_t + (0.5)^1 a_{t-1} + (0.5)^2 a_{t-2} + \dots \\ &= a_t + 0.5 a_{t-1} + 0.25 a_{t-2} + 0.125 a_{t-3} + \dots \end{aligned}$$

Für die Berechnung müssen wir das natürlich auf eine endliche Anzahl beschränken.

Moving Average - gleitender Durchschnitt

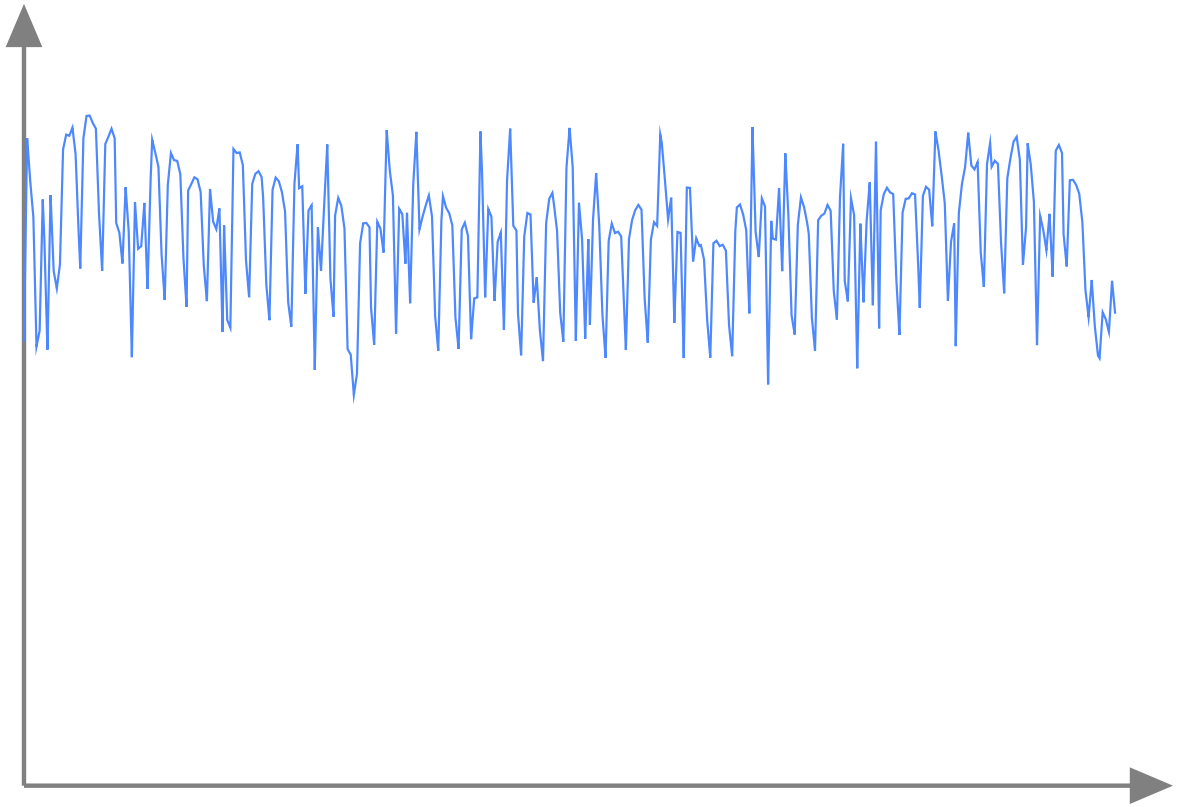
Wähle Parameter $q \in \mathbb{N}$ und setze

$$\psi_i = \begin{cases} \frac{1}{q} & \text{für } i = 1, \ldots, q \\ 0 & \text{für } i > q \end{cases}$$

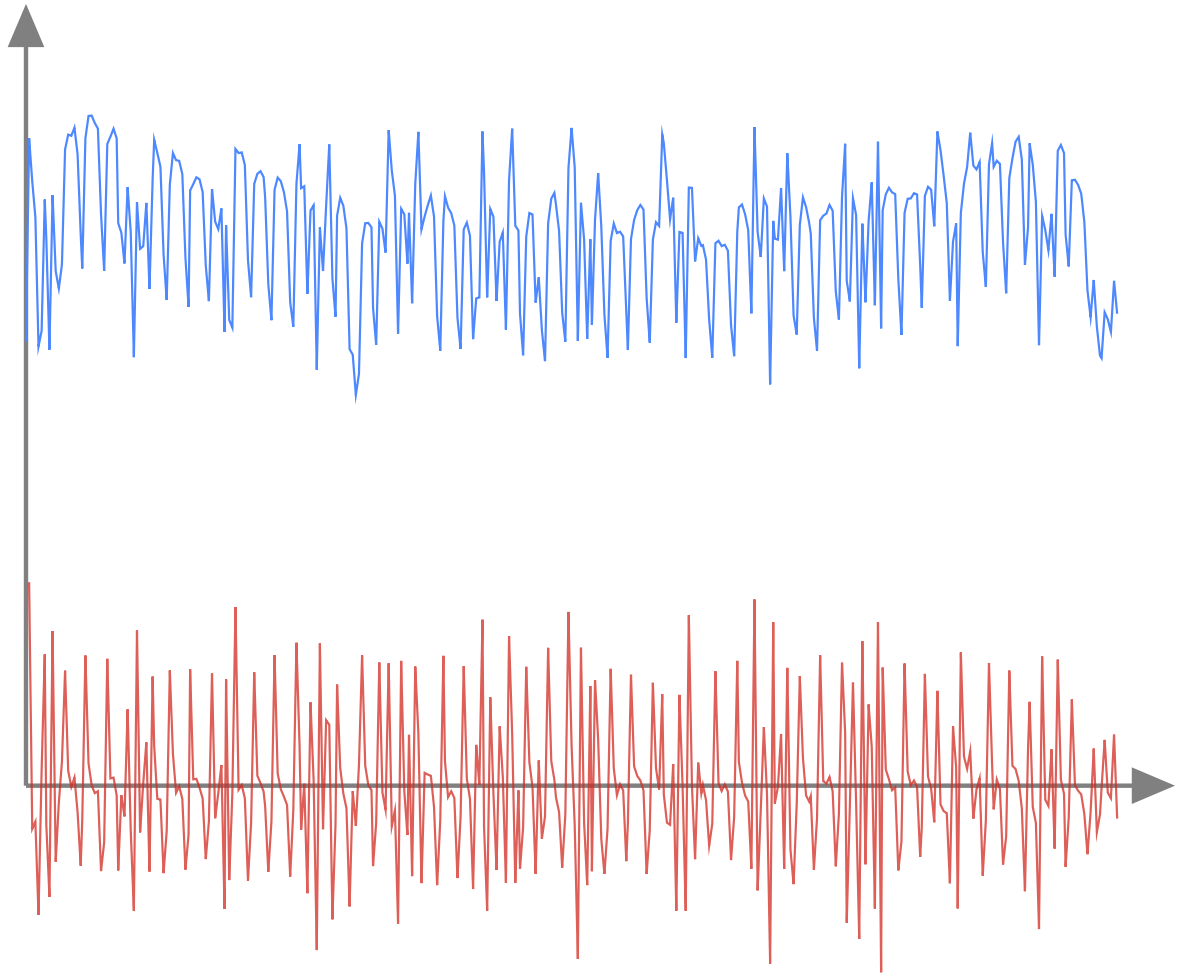
dann ist

$$X_t = a_t + \frac{1}{q} \sum_{i=1}^{q} a_{t-1}$$

Wie kommen wir zu stationären Zeitreihen?



Wie kommen wir zu stationären Zeitreihen?



Bisher: Einfache Modelle für Zeitreihen

Aber uns fehlt ja noch

- die Modellierung von Trends
- die Modellierung von periodischen Saisonalitäten
- ... andere Anwendungsspezifische Eigenschaften

Bisher: Einfache Modelle für Zeitreihen

Aber uns fehlt ja noch

- die Modellierung von Trends
- die Modellierung von periodischen Saisonalitäten
- ... andere Anwendungsspezifische Eigenschaften

Trends haben wir ja in linearer Regression schonmal modelliert...

Das Additive Komponenten-Modell

Annahme, dass die Zeitreihe aus k Komponenten zusammengesetzt werden kann:

$$X_t = K_{1,t} + K_{2,t} + \ldots + K_{k,t} + e_t$$

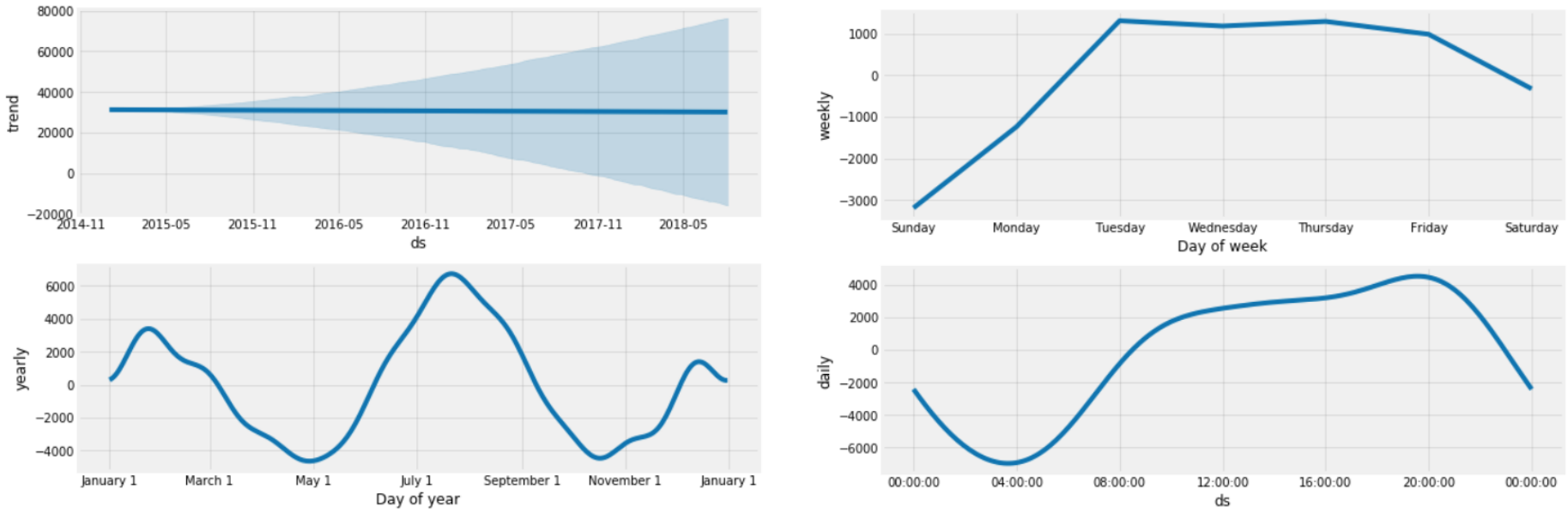
dabei ist $K_{i,t}$ die i-te Komponente, e_t der "Rest".

Beispiel: Trend und Saisonalität

$$X_t = T_t + S_t + e_t$$

mit Trendkomponente T_t und Saisonkomponente S_t.

Komponente eines Komponentenmodells



Komponenten aus Prophet Modell

`www.kaggle.com/robikscube/time-series-forecasting-with-prophet`

Weiterführende Literatur, Quellen

- Skript *Zeitreihenanalyse Teil 2*
  (Prof. Dr. W. Zucchini, Uni Göttingen)
- Hinweis auf kommenden Gastvortrag

Zeitreihen mit Pandas

Pandas Series als Zeitreihe

Pandas Series ist ein Datentyp für Zeitreihen, bisher:

- Series als Folge von Werten
- Ganzzahliger Index für Wert an Position *i*

Für den Umgang mit Zeitreihen wäre folgendes hilfreich:

- Zugriff über Zeitpunkte/Zeitintervalle
- Aggregation/Umwandlung in andere Granularitäten (z.B. Tag $\rightarrow$ Monat)

DatetimeIndex für Pandas Series Daten

Pandas benutzt `datetime46` als Zeitstempel:

- Hochauflösende Zeitstempel (Nanosekunden)
- Parser für Zeitstempel vorhanden (`to_datetime`)

Beispiel:

```
df = ... # Daten lesen

# Zeitreihe aus Tabelle extrahieren
zr = df['Werte']

# Zeitstempel erzeugen und index setzen:
zr.index = pd.to_datetime(df['Datum'])
```

DatetimeIndex manuell erzeugen

Die `date_range` Funktion erzeugt Folgen von Zeiten:

```
data = ... # Sequenz von Messwerten

# Tageweiser Index:
idx = pd.date_range('2021-01-01', '2021-01-31',
                    freq='D')

# Series mit index erzeugen:
series = pd.Series(data, index=idx)
```

`date_range` unterstützt Granularitäten (`freq`)

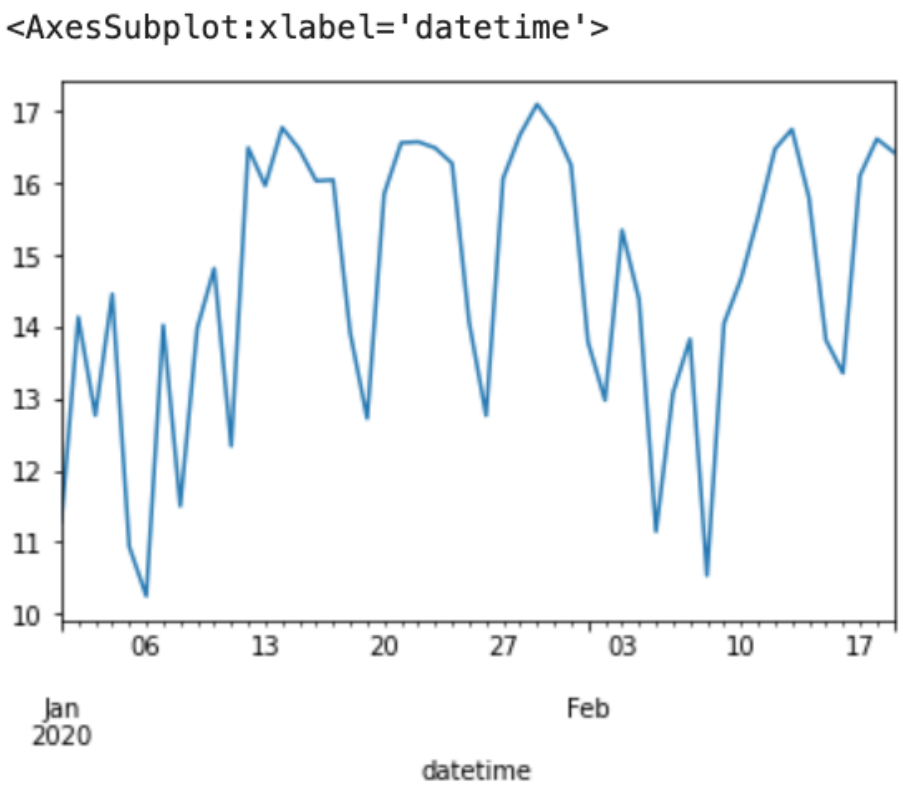
```
# Index mit 16 Stunden ab Startzeitpunkt:
idx = pd.date_range(start='2021-01-01 00:00',
                    periods=24, freq='H')
```

D	Tag
B	Werktag (*business day*)
H	Stunde
M, MS	Monatsende, Monatsbeginn (*month start*)
W-MON, W-TUE,..	Wöchentlich Montag, wöchentlich Dienstag,..
…	…
A-JAN, A-FEB,..	Jährlich Januar, Februar,..

Plotten von Pandas Series

Pandas Series bringen bereits eine `plot()` Funktion mit:

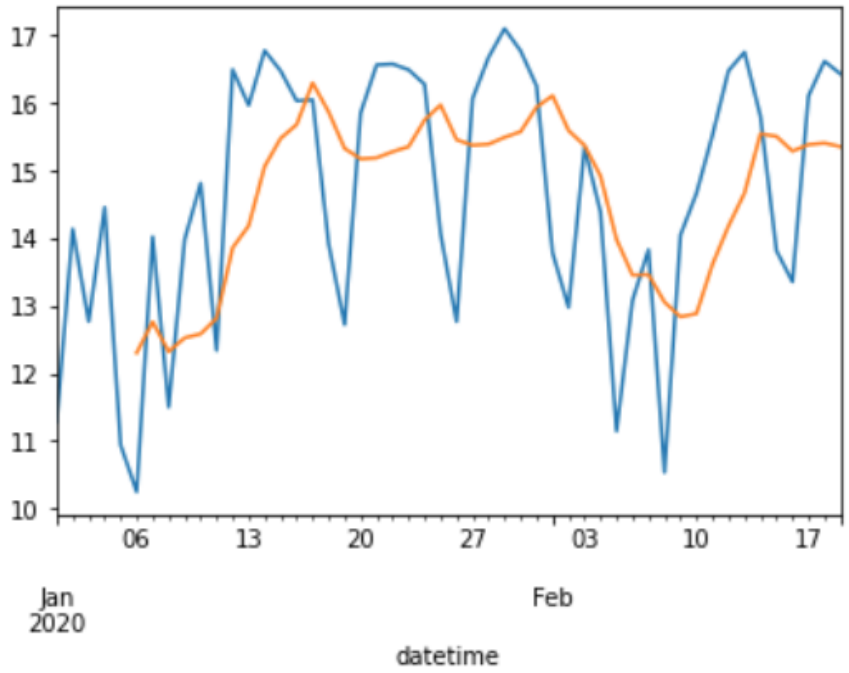
```
series.plot()
```



Hinweis:
Wenn Sie mehrere .plot() Aufrufe in einer Code-Zelle ausführen, erhalten Sie einen Plot mit mehreren Series-Verläufen:

```
s1.plot()
s1.rolling(6).mean().plot()
```

```
<AxesSubplot:xlabel='datetime'>
```



Resampling von Zeitreihen

Resampling Funktion ermöglicht die Anpassung der Frequenz:

```
# Laden der Zeitreihe
series = ...

# Monatlichen Durchschnitt als neue Zeitreihe
monthly = series.resample('M').mean()
```

Resampling funktioniert auch mit eigenen Funktionen:

```
def fn(x):
    return sum(x) / len(x)

daily = series.resample('D').apply(fn)
```

Shift Operation von Series

Zeitreihen lassen sich mit `shift` um Stellen verschieben:

```
series = ...

verschoben_um_1 = series.shift(1)
```

Shift Operation von Series

Zeitreihen lassen sich mit `shift` um Stellen verschieben:

```
series = ...

verschoben_um_1 = series.shift(1)
```

Beispiel: Änderungsrate berechnen

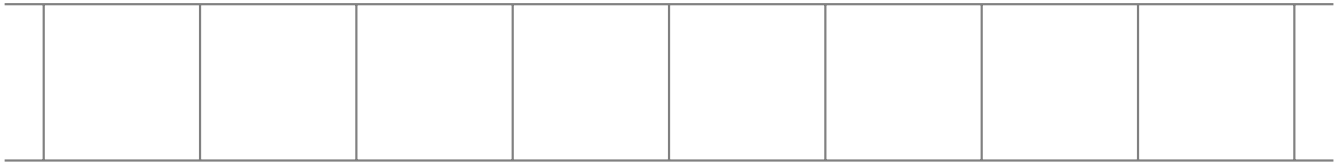
```
ts = ... # Series einlesen, aus DF selektiere,...

# Prozentuale Aenderung berechnen:
rate = ts / ts.shift(1) - 1
```

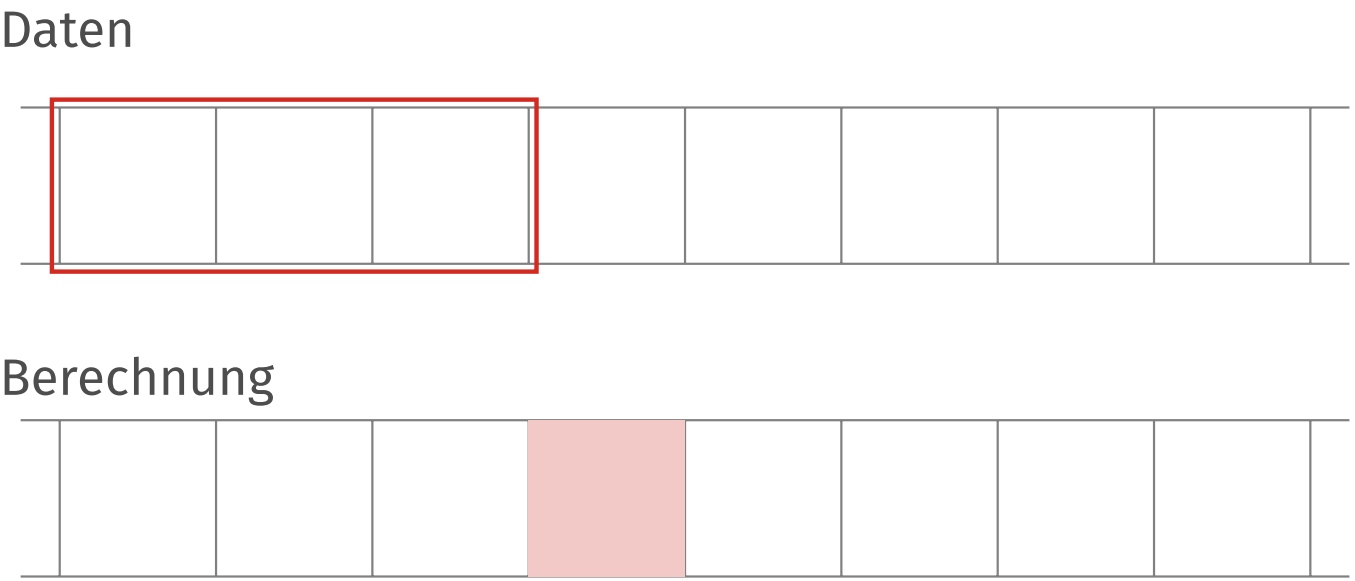
Sliding Windows

Daten

Berechnung

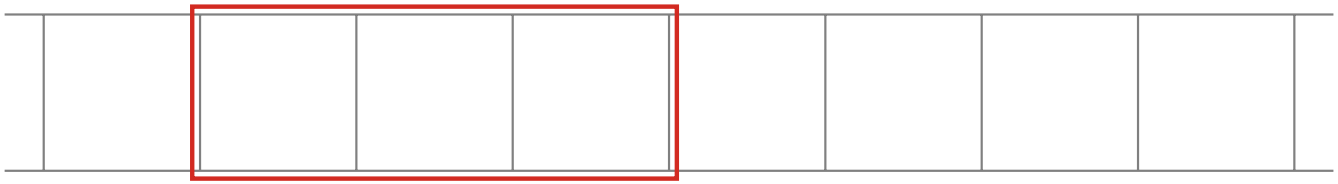


Sliding Windows

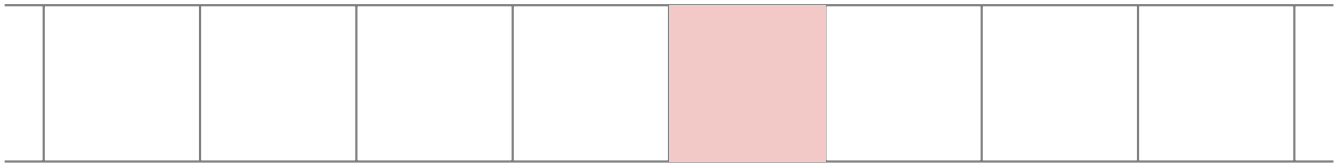


Sliding Windows

Daten

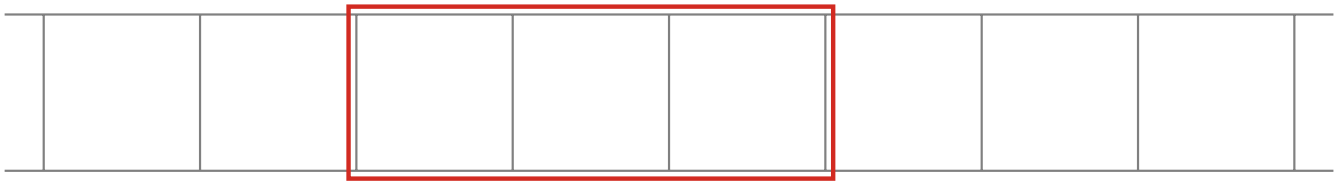


Berechnung



Sliding Windows

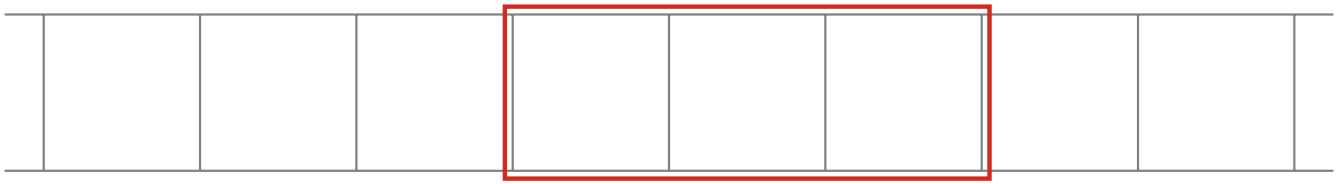
Daten



Berechnung

Sliding Windows

Daten



Berechnung

Sliding Windows

Daten

Berechnung

Sliding (rolling) Windows mit Pandas

Pandas Series hat `rolling` Funktion, die sich aggregieren läßt:

```
# Series lesen
daten = ...

# Mittelwert auf letzten 4 Werten berechnen
#
berechnet = daten.rolling(window=4).mean()
```