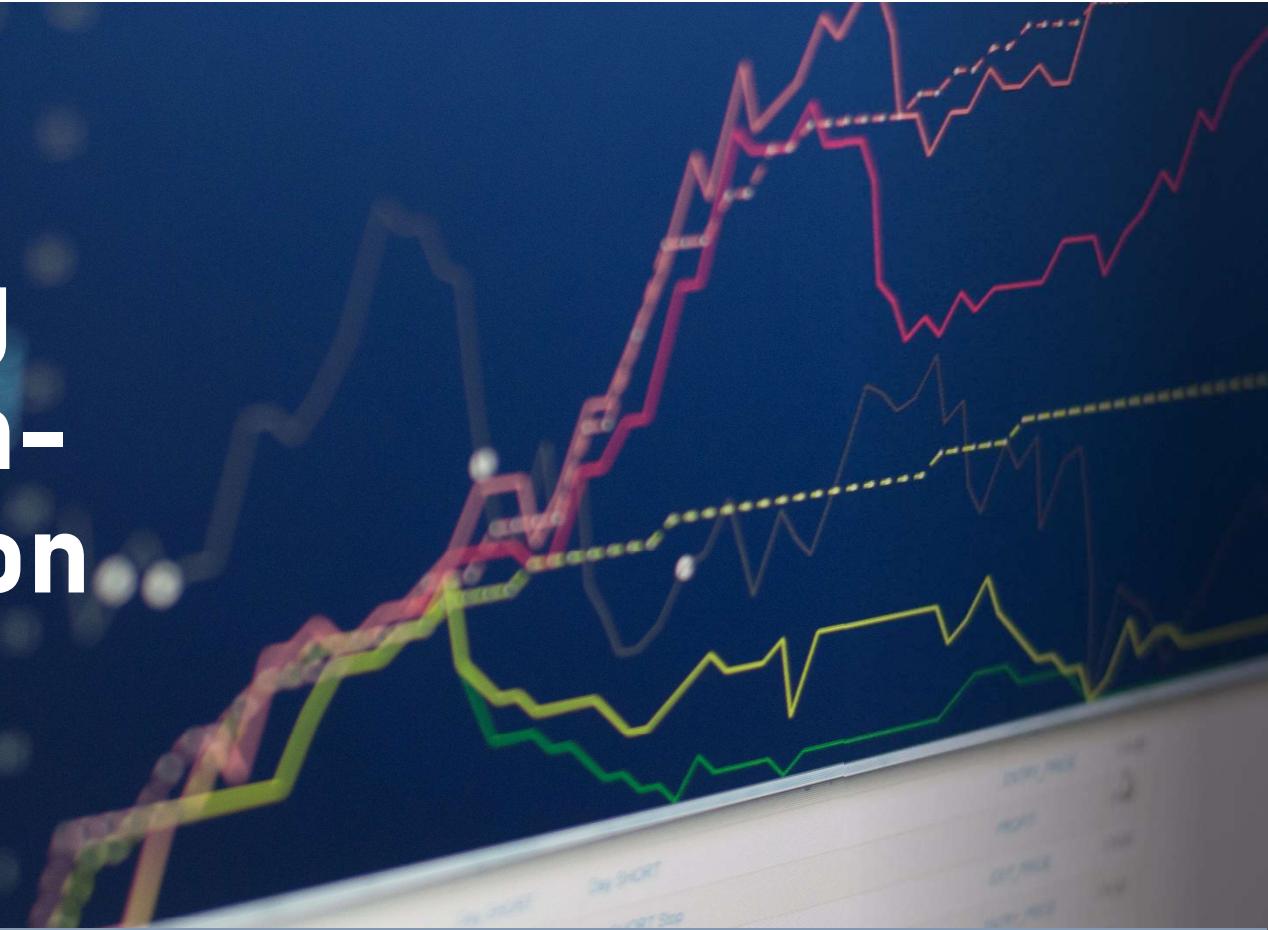


Analyse und Visualisierung von Zeitreihen- Daten in Python



Hmetrica



Analyse und Visualisierung von Zeitreihen- Daten in Python

Tag 1



Erwartungen

1. Ich heiße ...
2. "Fachlich" würde ich sagen bin ich ...
3. Ich arbeite ___ (immer / oft / manchmal / selten / nie) mit Zeitreihen – meist im Kontext von ...
4. Ich arbeite ___ (immer / oft / manchmal / selten / nie) mit Python – meist im Kontext von ...
5. Vom Seminar würde ich gerne ___ mitnehmen.
6. Am meisten interessiert mich dabei ...
7. Wenn ich heute nicht hier wäre, dann würde ich ...





Hmetrica GmbH

www.hmetrica.com |
sina.huber@hmetrica.com

Data Science Beratung

- Statistik, Machine Learning und Künstliche Intelligenz.
- Schwerpunkte: Datenvisualisierung, Machine Learning, Zeitreihenanalyse

Inhalte – Was haben wir vor?

Tag 1

1. Einführung in Zeitreihendaten in Python
2. Zeitreihen und ihre Merkmale visualisieren
3. Zeitreihen vorhersagen (Statistik I): Exponentielle Glättung und Holt-Winters
4. Zeitreihen vorhersagen (Statistik II): ARIMA-Modelle

Tag 2

5. Einblick in andere Zeitreihenmodelle
6. Machine Learning für Zeitreihen: Überblick, Vorbereitung und Vorhersagen
7. Machine Learning für Zeitreihen: Clustering und Klassifikation
8. Deep Learning für Zeitreihen

Modelle für Zeitreihendaten

- **Theta** (Assimakopoulos and Nikolopoulos, 2000)
- **Exponential Smoothing**
 - (SES, Brown, 1956)
 - (ETS, Hyndman, 2008; Holt, 1957; Winters, 1960)
- **Auto-Regressive Integrated Moving Average**
 - (ARIMA, Box and Jenkins, 1990)
 - (TBATS, Livera et al., 2011)
 - (DHR-ARIMA, Hyndman and Athanasopoulos, 2021)
- **Fast Fourier Transform** (Cooley & Tuckey, 1965)
- **FB-Prophet** (Taylor & Letham, 2018)
- **Pooled Regression** (PR, Trapero et al., 2015)
- **Boosted Tree Ensembles**
 - XGBoost (Chen & Guestrin, 2016),
 - Light GBM (Ke et al., 2017),
- **Feed-Forward Neural Networks** (FFNN, Goodfellow et al., 2016)
- **RNN-Based Models: LSTM, GRU** (Hochreiter & Schmidhuber, 1997)
- **DeepAR** (Salinas et al., 2020)
- **N-BEATS** (Oreshkin et al., 2019), **N-HiTS** (Challu et al., 2022)
- **Transformer** (Vaswani et al., 2017)

Statistische vs Machine Learning Modelle

- In der Praxis: Es gibt keinen Königsweg (Godahewa et al., 2021)

Table 13: Mean RMSE results

Dataset	SES	Theta	TBATS	ETS	(DHR)-ARIMA	PR	CatBoost	FFNN	DeepAR	N-BEATS	WaveNet	Transformer
M1 Yearly	193829.49	171458.07	116850.90	167739.02	175343.75	152038.68	237644.50	154309.80	173075.10	192489.80	312821.80	182850.60
M1 Quarterly	2545.73	2282.65	2673.91	2408.47	2538.45	1909.31	2161.01	1871.85	2313.32	2267.27	2271.68	2231.50
M1 Monthly	2725.83	2564.88	2594.48	2263.96	2450.61	2478.88	2461.68	2527.03	2202.19	2183.37	2578.93	3129.84
M3 Yearly	1172.85	1106.05	1386.33	1189.21	1662.17	1181.81	1341.70	1256.21	1157.88	1117.37	1147.62	1084.75
M3 Quarterly	670.56	567.70	653.61	598.73	650.76	605.50	697.96	621.73	606.56	582.83	606.75	819.18
M3 Monthly	893.88	753.99	765.20	755.26	790.76	830.04	874.20	833.15	873.71	796.91	845.30	948.40
M3 Other	309.68	242.13	216.95	224.08	220.77	262.31	349.90	268.99	277.74	248.53	276.97	271.02
M4 Yearly	1154.49	1020.48	1099.95	1052.12	1230.35	1000.18	1065.02	-	-	-	-	-
M4 Quarterly	732.82	673.15	672.74	674.27	709.99	711.93	714.21	735.84	700.32	684.65	696.96	739.06
M4 Monthly	755.45	683.72	743.41	705.70	702.06	720.46	734.79	743.47	740.26	705.21	787.94	902.38
M4 Weekly	412.60	405.17	356.74	408.50	386.30	350.29	420.84	399.10	422.18	330.78	437.26	456.90
M4 Daily	209.75	210.37	208.36	229.97	212.64	213.01	263.13	209.44	343.48	221.69	220.45	233.63
M4 Hourly	1476.81	1483.70	469.87	3830.44	1563.05	312.99	344.62	467.89	1095.10	501.19	468.09	391.22
Tourism Yearly	106665.20	99914.21	105799.40	104701.51	106082.60	89645.61	87489.00	87931.79	78470.68	78241.67	77581.31	80089.25
Tourism Quarterly	17270.57	9254.63	12001.48	10312.34	12564.77	11746.15	12787.97	12182.57	11761.96	1305.95	1154.58	11724.14
Tourism Monthly	7039.35	271.96	366.51	254.96	3132.40	2739.43	3102.76	2584.10	2359.87	2662.21	2691.22	2666.06
CIF 2016	657112.42	894600.19	940000.90	722397.37	526394.02	648800.31	705723.30	1629741.53	353375.00	772924.30	600524.41	4625974.00
Aus. Electricity Demand	766.27	771.51	446.59	1404.02	1234.76	319.98	300.53	330.91	357.00	268.37	286.48	295.22
Dominick	6.48	6.74	8.03	6.59	7.96	9.44	9.15	6.67	9.78	6.81	6.63	-
Bitcoin	5.35×10^{18}	5.35×10^{18}	1.16×10^{18}	1.22×10^{18}	3.96×10^{18}	8.29×10^{17}	2.02×10^{18}	1.57×10^{18}	2.02×10^{18}	1.26×10^{18}	2.55×10^{18}	2.67×10^{18}
Pedestrian Counts	228.14	228.20	261.25	278.26	820.28	61.84	60.78	67.17	65.77	99.33	67.99	70.17
Vehicle Trips	36.53	37.44	25.69	37.61	34.95	31.69	27.28	27.88	26.46	33.56	28.99	32.98
KDD Cup	73.81	73.83	71.21	76.71	82.66	65.71	68.43	80.19	80.39	68.87	76.21	-
Weather	3.85	3.77	2.49	2.49	3.01	9.08	3.09	2.31	2.74	3.09	2.98	2.81
NN5 Daily	8.23	5.28	5.20	5.22	6.05	7.26	5.73	5.79	5.50	6.47	5.75	5.92
NN5 Weekly	18.82	6.65	18.53	18.82	18.55	18.62	18.67	18.29	18.53	17.35	24.16	24.02
Kaggle Daily	590.11	583.32	740.74	650.43	595.43	-	-	-	-	-	-	-
Kaggle Weekly	2970.78	3012.39	2951.87	3369.64	3777.28	4750.26	14040.64	2719.65	2981.91	2820.62	2719.37	3815.38
Solar 10 Minutes	7.23	10.71	7.23	5.55	7.23	8.73	7.21	7.22	6.62	7.23	-	-
Solar Weekly	1331.26	1341.55	1049.01	1264.43	967.87	1168.18	1754.22	1231.54	873.62	1307.78	2569.26	693.84
Electricity Hourly	1026.29	1026.36	743.35	1524.87	1082.44	689.85	582.66	519.06	477.99	510.91	489.91	514.68
Electricity Weekly	77067.87	76935.58	28039.73	70368.97	32594.81	47802.08	37289.74	30594.15	53100.26	35576.83	63916.89	78894.67
Carparts	0.78	0.78	0.84	0.80	0.81	0.73	0.79	0.74	0.74	1.11	0.74	0.74
FRED-MD	3103.00	3898.72	2295.74	2341.72	3312.46	9736.93	2679.38	2631.04	4638.71	2812.97	2779.48	5098.91
Traffic Hourly	0.04	0.04	0.05	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.03	0.02
Traffic Weekly	1.51	1.53	1.53	1.53	1.54	1.50	1.50	1.55	1.51	1.44	1.61	1.94
Rideshare	7.17	8.60	7.35	7.17	4.80	7.18	6.95	7.14	7.15	6.23	3.51	7.17
Hospital	26.55	22.59	21.28	22.02	23.68	23.48	23.45	27.77	22.01	24.18	23.38	40.48
COVID Deaths	403.41	370.14	113.00	102.08	100.46	394.07	607.92	173.14	230.47	186.54	1135.41	479.96
Temperature Rain	10.34	10.36	9.20	10.38	9.22	9.83	8.71	8.89	9.11	11.03	9.07	9.01
Sunspot	4.95	4.95	2.97	4.95	2.96	3.95	2.38	8.43	1.14	14.52	0.66	0.52
Saugeen River Flow	39.79	39.79	42.58	50.39	43.23	47.70	39.32	40.64	45.28	48.91	42.99	49.12
US Births	1369.50	735.51	606.54	607.20	705.51	732.09	618.48	756.77	683.60	697.74	768.81	686.41

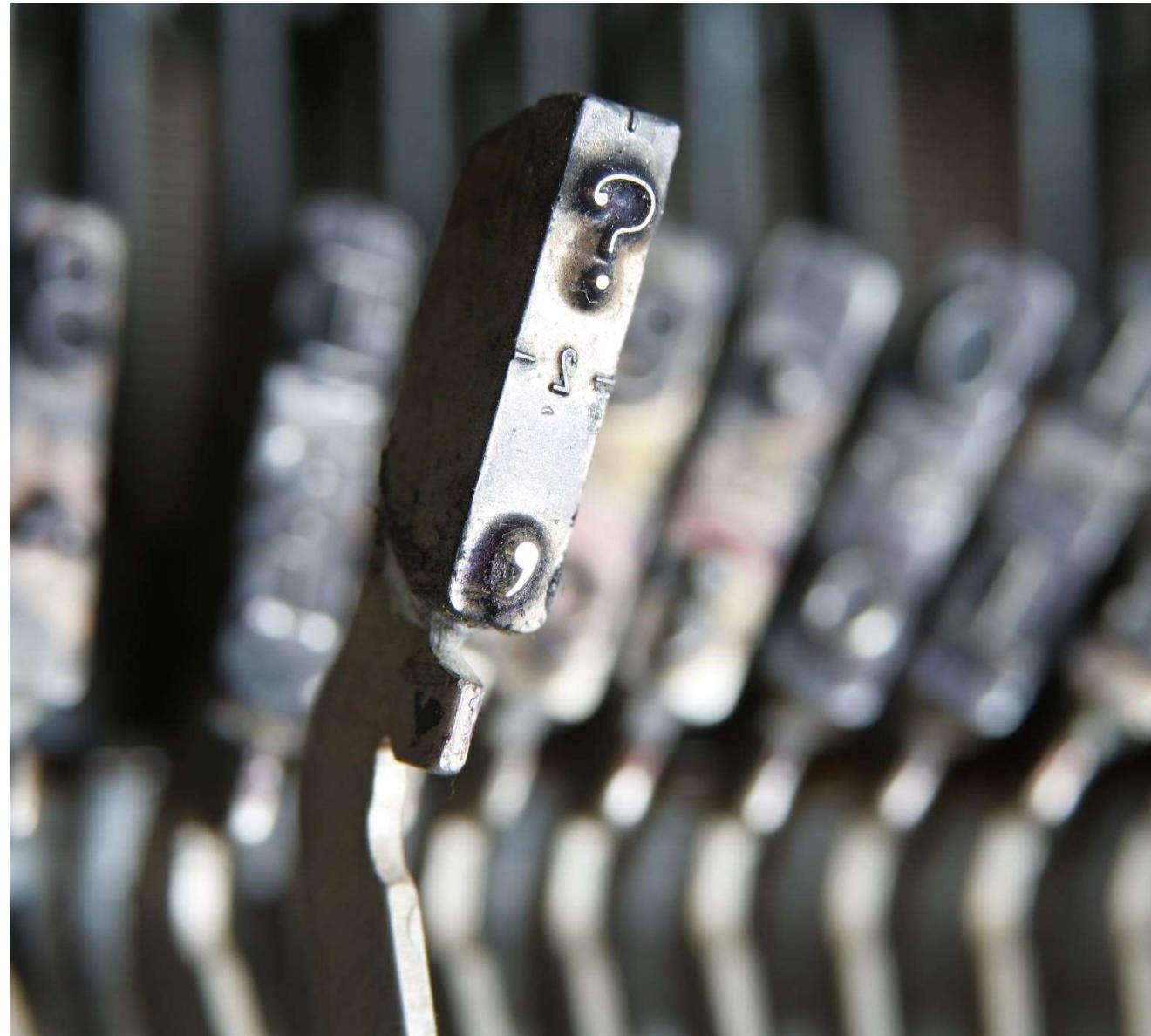
Stats Models

ML Models

Images: Table of model comparisons adapted from Monash Study (Godahewa et al., 2021)

Fragen?

- Mittagessen 12:30 – 13:30 Uhr?



Einführung in Zeitreihendaten in Python

Session 1 (Montag 09:15 – 10:45)



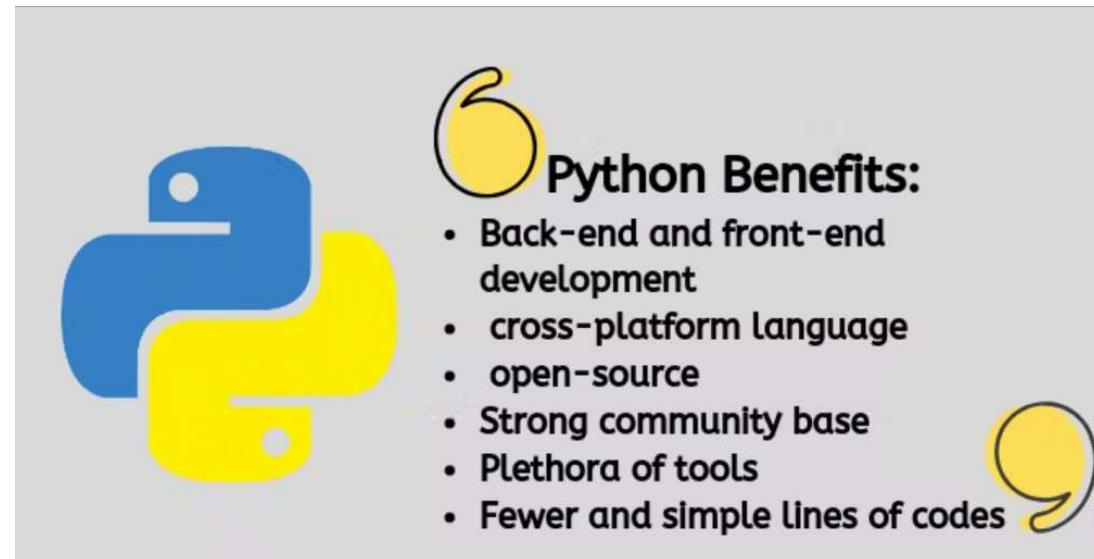
Was haben wir vor?

1. Einführung in Zeitreihendaten in Python
 - 1.1 Einführung in Python: pandas, matplotlib
 - 1.2 Einführung in Zeitreihendaten: Definitionen, einfache Merkmale

Einführung in Python

Python ist eine allgegenwärtige Skriptsprache, benutzt für:

- Webentwicklung,
- App-Entwicklung,
- wissenschaftliche und numerische Bereiche,
- Geschäftsanwendungen,
- GUI-Design,
- Automatisierung,
- künstliche Intelligenz,
- maschinelles Lernen
- und vieles mehr...



<https://rotechnica.com/what-is-python-used-for/>, <https://hackernoon.com/what-is-python-used-for-an-exclusive-answer-1z2xo3xtf>,

Einführung in Zeitreihendaten



Beispiele für Zeitreihenanwendungen



<https://www.analyticssteps.com/blogs/introduction-time-series-analysis-time-series-forecasting-machine-learning-methods-models>

Bsp: Energiewende

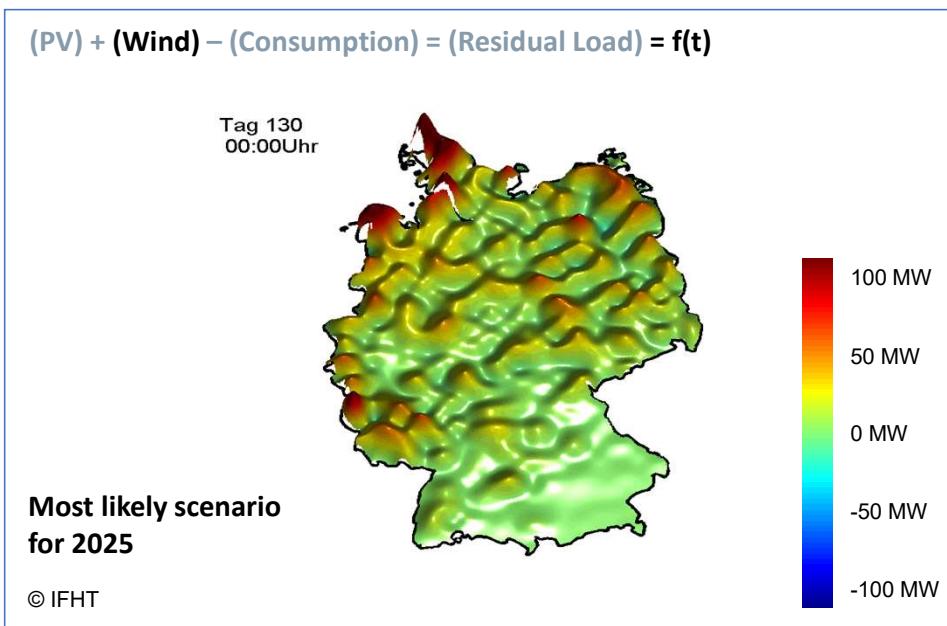


Image source: earth.com; treehugger.com; Siemens, RWTH, Metzger, Huber, et al., 2021;



Was sind Zeitreihen?

- Umgangssprachlich: Daten, die sich auf aufeinanderfolgende Zeitpunkte (oder Zeiträume) beziehen
- Statistik betrachtet: Zeitreihe als “Realisation” eines “stochastischen Prozesses”
- Maschinelles Lernen: Zeitreihen als Datensequenzen mit Mustern über die Zeit hinweg

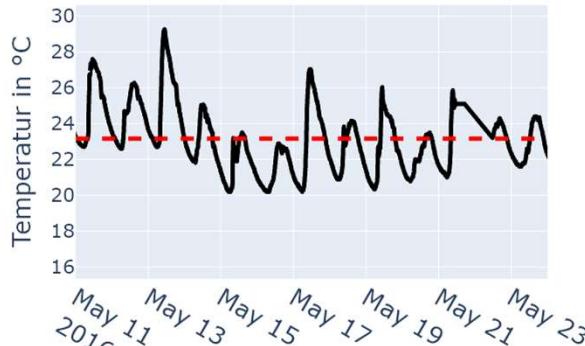
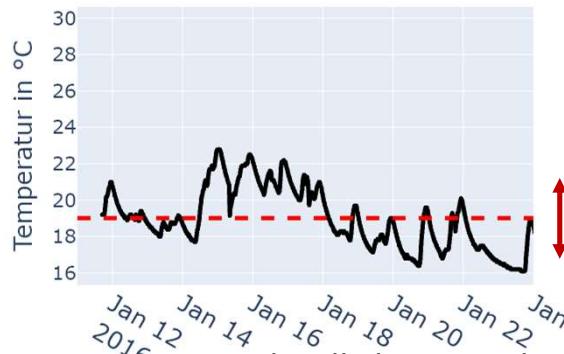
Stochastischer Prozess $\{X_t\}_{t \in T}$ ist eine Folge von Zufallsvariablen (Funktionen $X_t: \Omega \rightarrow \mathbb{R}$), die einen Index t aus einer Indexmenge T (meist Zeit gemessen als \mathbb{N}_0 oder \mathbb{R}_+) haben

Zeitreihen beschreiben

Beispiel

Sie beschreiben die Temperatur in einem Gebäude im Januar und Mai

Wie unterscheiden sich diese beiden Zeitreihen?



Sie haben einen unterschiedlichen Durchschnitt – und streuen auch unterschiedlich darum herum

Stochastischer Prozess X_1, X_2, X_3, \dots

Erwartungswert $\mu_t = \mathbb{E}[X_t]$

Varianz $\text{Var}(X_t) = \mathbb{E}[(X_t - \mu_t)^2] = \gamma_t(0)$

Autokovarianz

$\text{Cov}(X_{t+h}, X_t) = \mathbb{E}[(X_{t+h} - \mu_{t+h})(X_t - \mu_t)] = \gamma_t(h)$

Daten $x_1, x_2, x_3, \dots, x_t$

Mittelwert $\bar{x}_t = \hat{\mu}_t = \frac{1}{t} \sum_{i=1}^t x_i$

Stichprobenvarianz $\hat{\gamma}_t(0) = \frac{1}{t} \sum_{i=1}^t (x_i - \bar{x})^2$

Stichprobautokovarianz $\hat{\gamma}_t(h) =$

$\frac{1}{t} \sum_{i=1}^{t-|h|} (x_{i+|h|} - \bar{x})(x_i - \bar{x})$

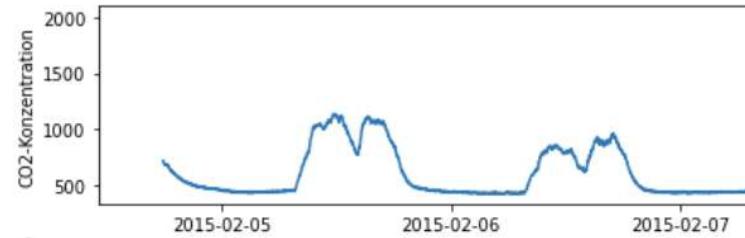
Hängt von der Zeit ab!

Warum sind Zeitreihen besonders?

- Bei Zeitreihen:
- Beobachtung zu Zeitpunkt t : *könnte etwas zu tun haben* mit Beobachtungen zu anderem Zeitpunkt $t + h$
- Deshalb: Kovarianz der Zeitreihe mit sich selbst (zeitverzögert) wichtig!

Beispiel

- (A) Ich messe heute Mittag die CO₂-Konzentration in 72 Büros in München (**keine Zeitreihe**)
(B) Ich messe 72 Stunden lang die CO₂-Konzentration in meinem München (**Zeitreihe**)



Stochastischer Prozess $X_1, X_2, X_3, \dots, X_n$

Kovarianz

$$\text{Cov}(X_{t+h}, X_t) = \mathbb{E}[(X_{t+h} - \mu_{t+h})(X_t - \mu_t)] = \gamma_t(h)$$

Daten $x_1, x_2, x_3, \dots, x_n$

Stichprobenkovarianz

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x})$$

- Die **Autokorrelationsfunktion ACF** (engl. Auto Correlation Function) ist definiert als

$$\rho_t(h) = \frac{\gamma_t(h)}{\gamma_t(0)}$$

Zeitreihen modellieren

Mathematisch, können wir schreiben

- Ein stochastischer Prozess Z_t heißt **Weißes Rauschen**, wenn $\mu_t = E[Z_t] = 0$, $\text{Var}(Z_t) = \sigma^2$ und $\text{Cov}(Z_t, Z_{t+k}) = 0$ für $k \neq 0$
- Ein stochastischer Prozess X_t heißt **Random Walk**, wenn

$$X_t = X_{t-1} + Z_t$$

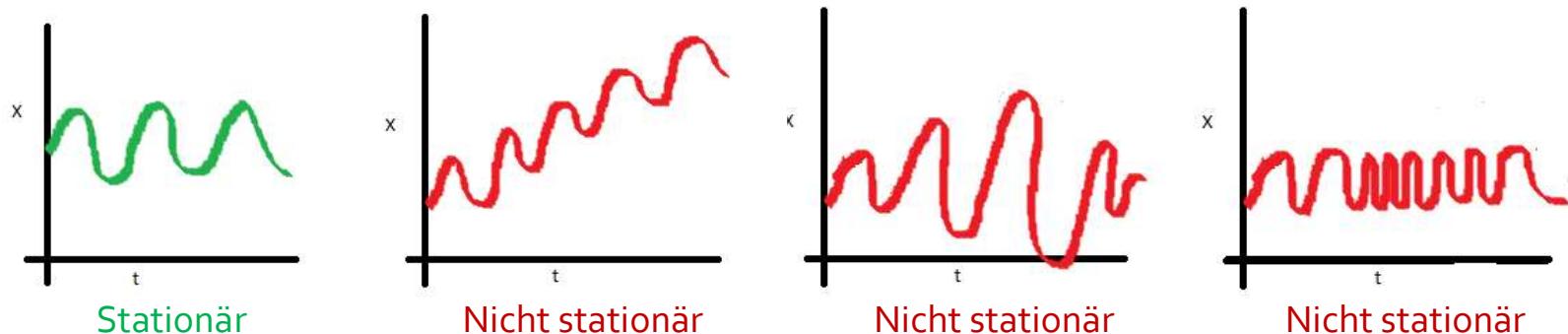
- wobei Z_t weißes Rauschen ist.

Zeitreihen modellieren

Eine Zeitreihe heißt **stationär**, wenn

$$\mu_t = \mu \quad \text{und} \quad \gamma_t(h) = \gamma(h)$$

- Mittelwertfunktion und die Autokovarianzfunktion (und somit auch die Varianzfunktion) nicht abhängig von t sind.



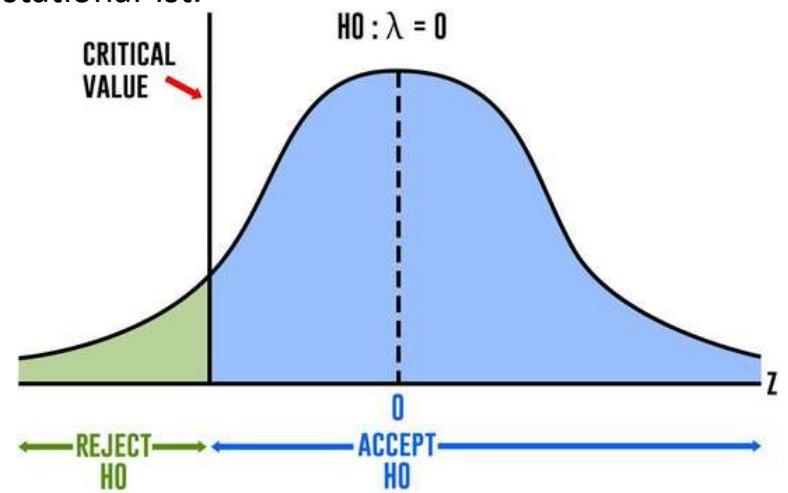
Zeitreihen modellieren

Was für Elemente machen einen Prozess nicht-stationär?

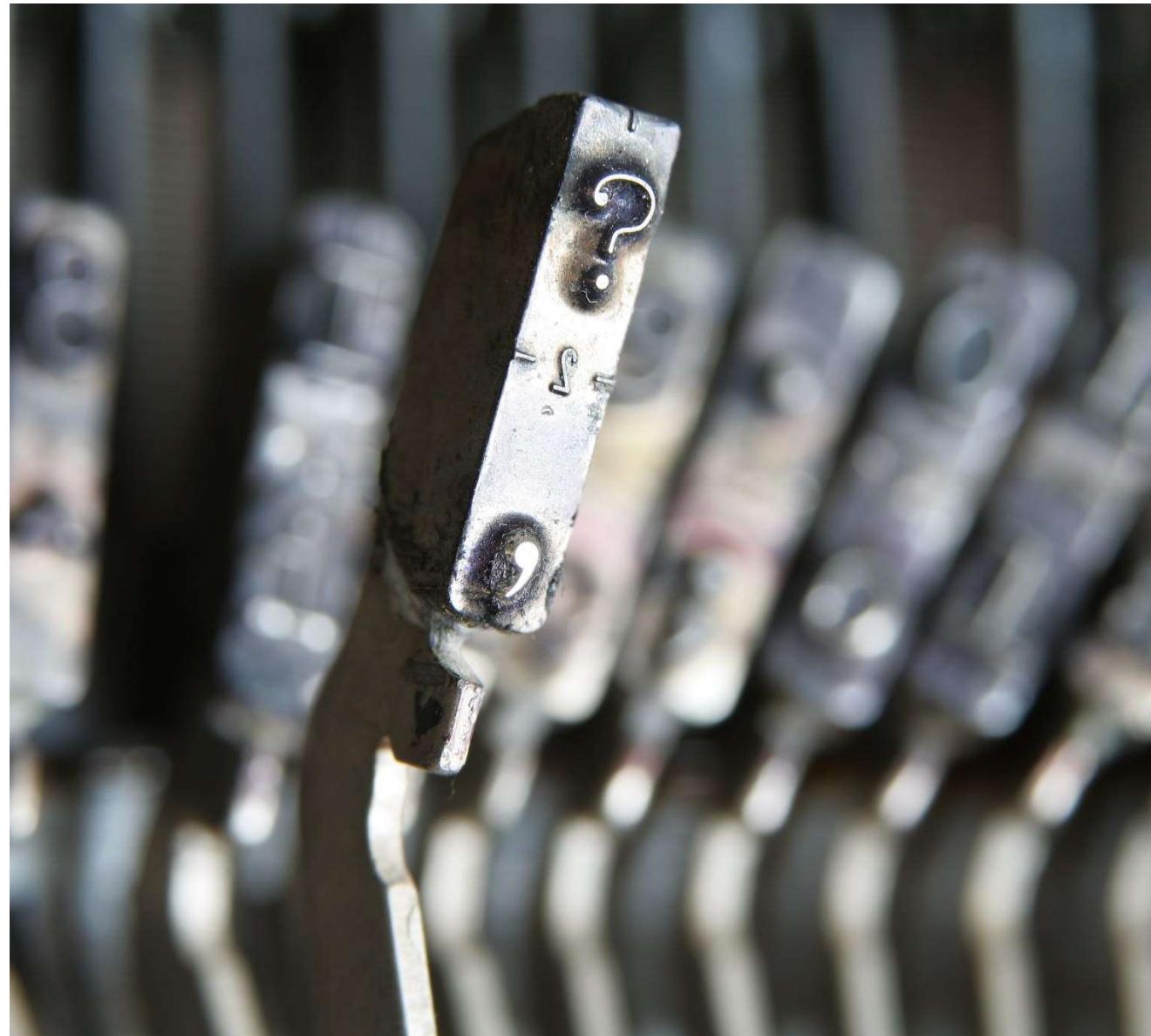
1. Trend
2. Saisonalität (hier kommt es theoretisch drauf an – praktisch aber sehr wichtig)
3. Heteroskedastizität (wenn die Varianz über die Zeit variiert)
4. Existenz einer “Einheitswurzel” → s App

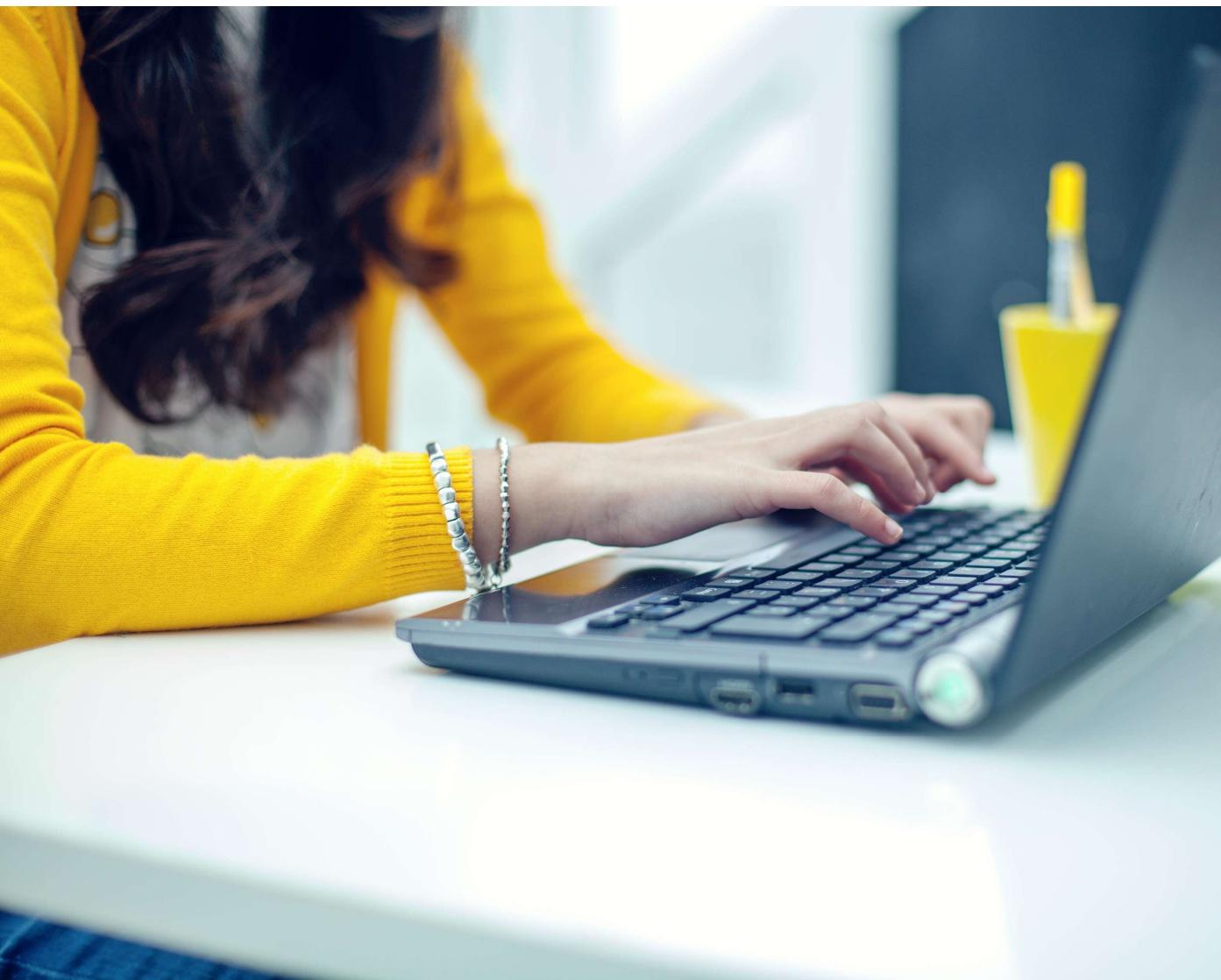
Stationarität testen

- Der **Augmented Dickey-Fuller-Test** ist ein statistischer Test, mit dem festgestellt werden kann, ob eine Zeitreihe stationär oder nichtstationär ist.
- Beliebter Test in der Zeitreihenanalyse.
- Die Nullhypothese des Adfuller-Tests lautet, dass die Zeitreihe nicht stationär ist.
- Die Alternativhypothese ist, dass die Zeitreihe stationär ist.



Fragen?





**Ab in die App
und ins
Jupyter
Notebook**

Zeitreihen und ihre Merkmale visualisieren

Session 2 (Montag 11:00 – 12:30)



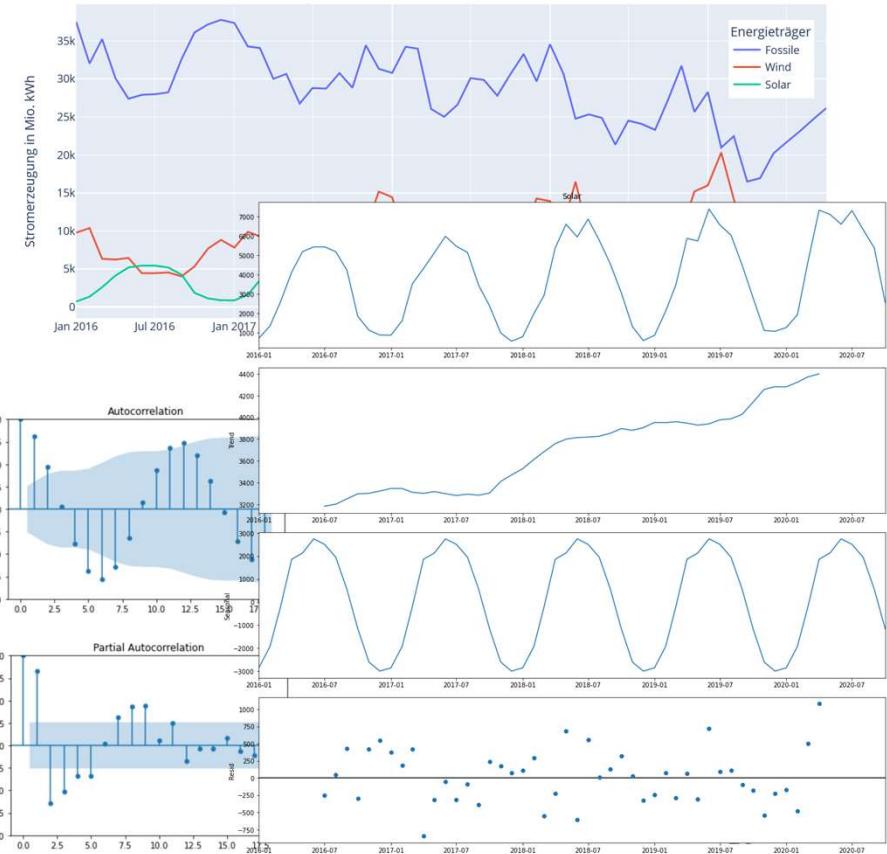
Zeitreihen und ihre Merkmale visualisieren

- Das **Plotten von Zeitreihendaten** ist eine häufige Aufgabe bei der Datenanalyse und -visualisierung.
- Die Visualisierung von Zeitreihendaten kann helfen, **Trends, Muster und Anomalien zu erkennen**, was ein besseres Verständnis und eine bessere Analyse der zugrunde liegenden Informationen ermöglicht.
- Wichtige Überlegungen:
 - **Auswahl des richtigen Visualisierungstools:** z. B. Matplotlib, Seaborn und Plotly in Python.
 - **Vorverarbeitung von Daten:** z.B. Umgang mit fehlenden Werten, die Aggregation von Daten zu bestimmten Zeiträumen (z. B. täglich, monatlich) oder die Anwendung von Glättungsverfahren zur Hervorhebung von Trends bei gleichzeitiger Reduzierung des Rauschens beinhalten

Zeitreihen und ihre Merkmale visualisieren

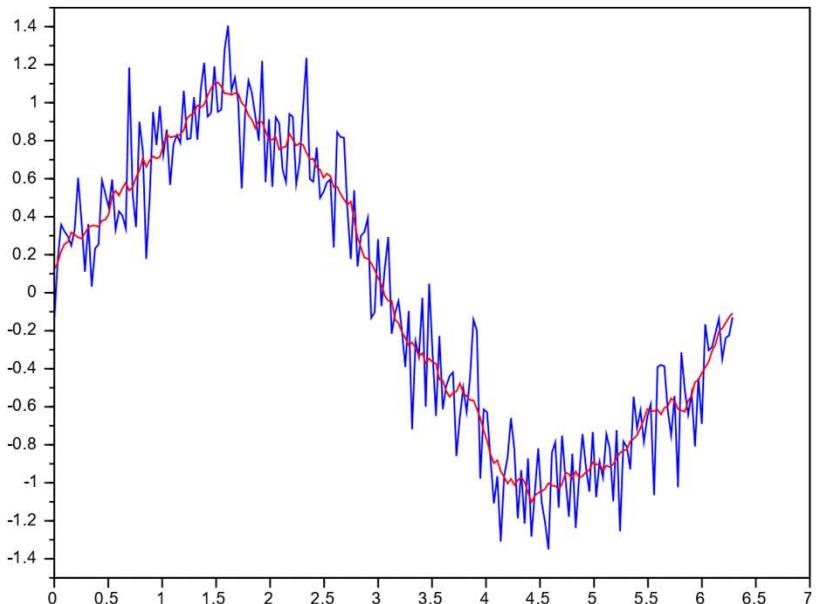
- **Liniendiagramme:** So lassen sich Trends und Muster im Zeitverlauf beobachten.
- **Gleitende Mittel:** einfache Glättung
- **Plots der Autokorrelationsfunktion (ACF) und der partiellen Autokorrelationsfunktion (PACF):** geben Aufschluss über die Korrelation zwischen Beobachtungen bei verschiedenen Verzögerungen (Lags).
- **Saisonale Muster:** oft saisonale Muster, z. B. regelmäßige Schwankungen (jährlich, täglich, ...). Um diese Muster zu visualisieren: saisonale Zerlegung von Zeitreihen

Stromerzeugung in Deutschland nach Energieträger



Gleitende Mittel

- Gleitende Mittel = grundlegende Technik in der Zeitreihenanalyse
- Dient dazu, kurzfristige Schwankungen zu glätten und langfristige Trends hervorzuheben
- Häufig zur „Glättung“ von Zeitreihendaten eingesetzt, um die zugrunde liegenden Trends besser zu erkennen
- Achtung: Wahl des „Fensters“ entscheidend.
- Achtung: Basieren auf Vergangenheitsdaten: „Hinken hinterher“

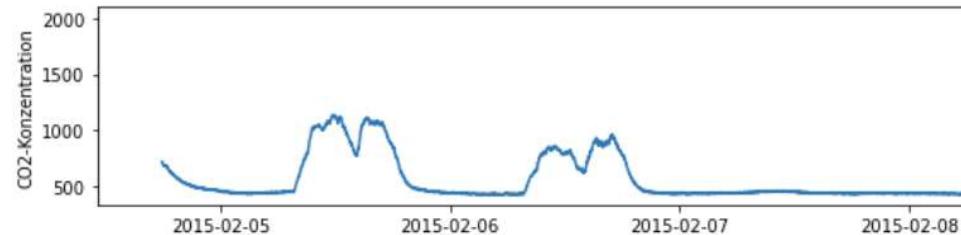


Autokorrelation

- Bei Zeitreihen:
- Beobachtung zu Zeitpunkt t : könnte etwas zu tun haben mit Beobachtungen zu anderem Zeitpunkt $t+h$
- Deshalb: Kovarianz der Zeitreihe mit sich selbst (zeitverzögert) wichtig!

Beispiel

- (A) Ich messe heute Mittag die CO₂-Konzentration in 72 Büros in München (**keine Zeitreihe**)
(B) Ich messe 72 Stunden lang die CO₂-Konzentration in meinem Büro in München (**Zeitreihe**)



Stochastischer Prozess $X_1, X_2, X_3, \dots, X_n$

Kovarianz

$$\text{Cov}(X_{t+h}, X_t) = \mathbb{E}[(X_{t+h} - \mu_{t+h})(X_t - \mu_t)] = \gamma_t(h)$$

Daten $x_1, x_2, x_3, \dots, x_n$

Stichprobenkovarianz

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x})$$

Die **Autokorrelationsfunktion ACF** (engl. Auto Correlation Function) ist definiert als

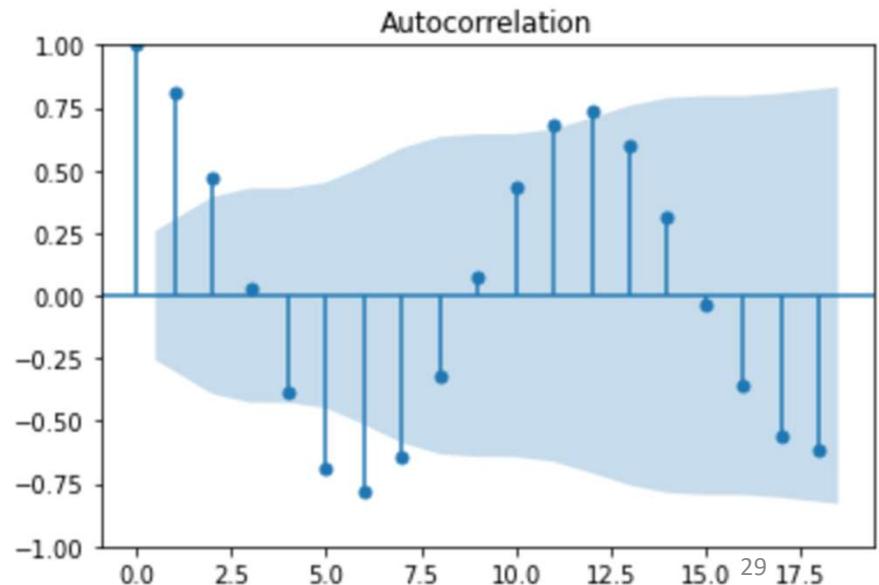
$$\rho_t(h) = \frac{\gamma_t(h)}{\gamma_t(0)}$$

Autokorrelationsfunktion (ACF) Plot

Der Lag-Operator (L) wirkt auf ein Element einer Zeitreihe, um das vorhergehende Element zu erzeugen.

$$LX_t = X_{t-1}$$

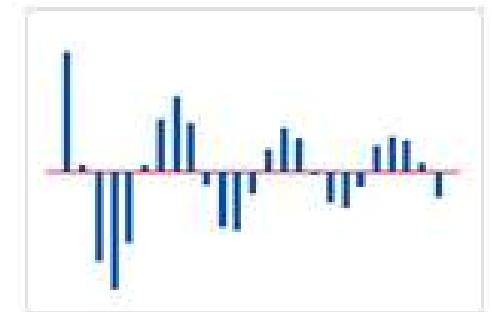
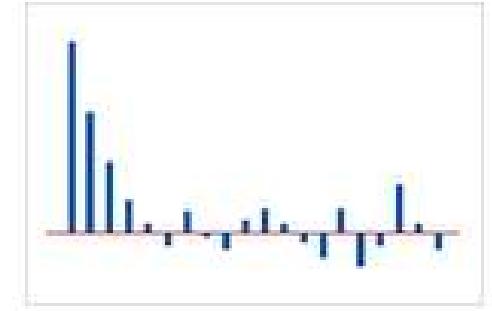
- Die ACF misst die **Korrelation zwischen einer Zeitreihe und ihren verzögerten Werten** (Lags).
- Hilft, die Beziehung zwischen einer Beobachtung und ihren historischen Werten bei verschiedenen Lags zu identifizieren.
- Die ACF-Darstellung zeigt die **Korrelationskoeffizienten auf der y-Achse** an, während die **x-Achse die Lags** darstellt.



Autokorrelationsfunktion (ACF) Plot

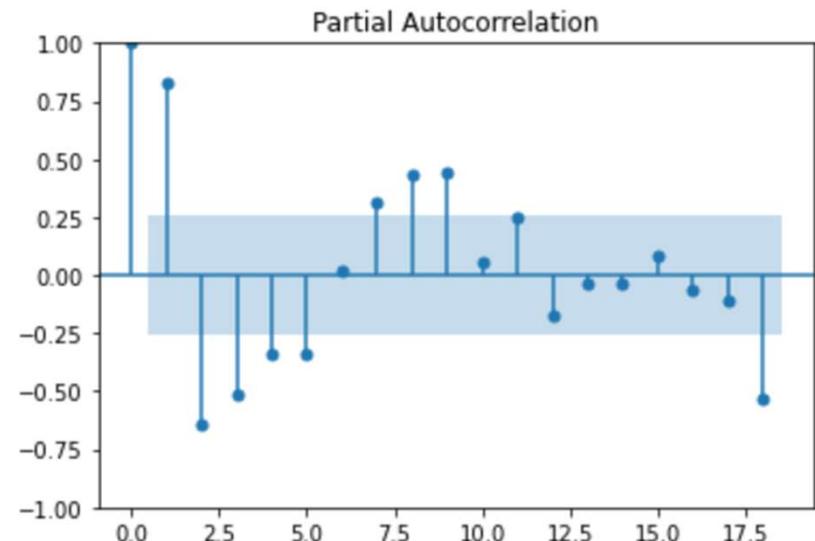
Im ACF Plot:

- Der Korrelationskoeffizient **bei Lag 0 ist immer 1**, da er die Korrelation zwischen einer Beobachtung und sich selbst darstellt.
- ACF-Werte **nahe bei 1** deuten auf eine starke positive Korrelation hin, während Werte **nahe bei -1** eine starke negative Korrelation anzeigen.
- ACF-Werte **nahe 0** deuten darauf hin, dass keine signifikante Korrelation zwischen den Beobachtungen mit diesem Lag besteht.



Partielle Autokorrelationsfunktion (PACF) Plot

- Der PACF misst die **Korrelation zwischen einer Zeitreihe und ihren verzögerten Werten**, nachdem die Auswirkungen von Zwischenlags entfernt wurden.
- Mit anderen Worten, er misst die direkte Beziehung zwischen einer Beobachtung und ihren historischen Werten bei bestimmten Verzögerungen, wobei der Einfluss früherer Lags berücksichtigt wird.
- Werte jenseits des **Konfidenzintervalls** (in der Grafik oft schattiert) gelten als statistisch signifikant
- PACF: **Anzahl der signifikanten Lags, bei denen die Korrelation nicht durch kürzere Lags erklärt wird.**



Quelle: Eigene Darstellung

ACF vs PACF Plot

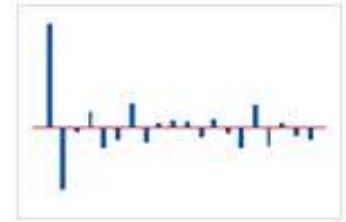
- Den Unterschied zwischen ACF und PACF lässt sich am besten durch die entsprechenden Regressionsgleichungen verstehen
- Beispiel: Autokorrelationsfunktion (ACF) zu Lag 6

$$y = a_6 x_{t-6}$$

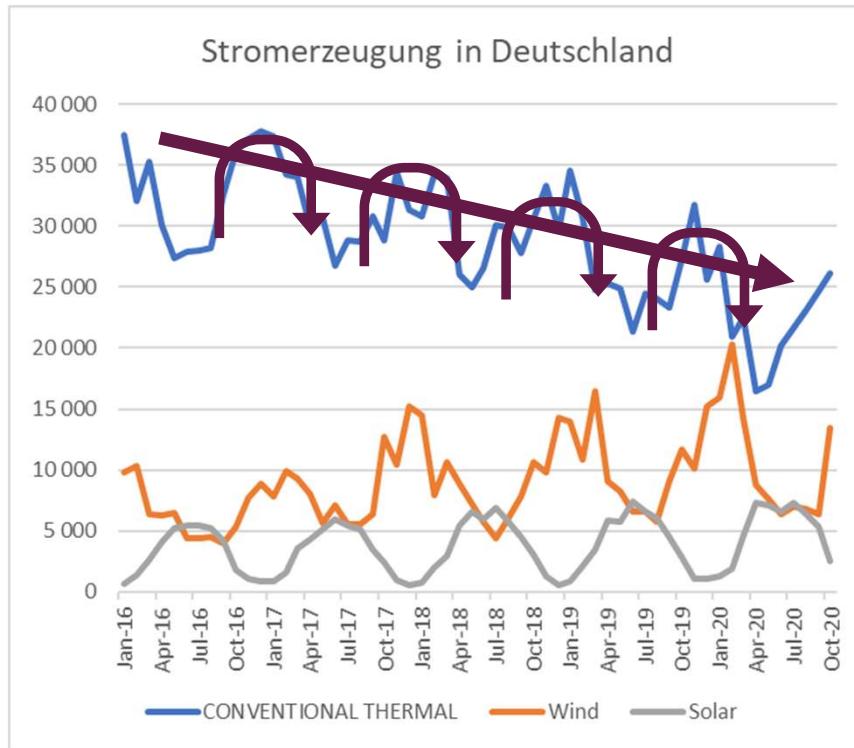
- Bestimme Gewicht a_6
- Partielle Autokorrelationsfunktion (PACF) zu Lag 6

$$y = a_1 x_{t-1} + a_2 x_{t-2} + a_3 x_{t-3} + a_4 x_{t-4} + a_5 x_{t-5} + a_6 x_{t-6}$$

- Bestimme Gewicht a_6



Komponenten von Zeitreihen



Beispiel

Sie beschreiben die Entwicklung der Stromproduktion aus fossilen Energieträgern in Deutschland ab 2016

Komponenten von Zeitreihen

- Die saisonale Zerlegung ist eine Technik, die in der Zeitreihenanalyse verwendet wird, um eine Zeitreihe in ihre **zugrunde liegenden Komponenten** zu zerlegen: **Trend, Saison und Residuum**.
- Diese Zerlegung hilft dabei, die einzelnen Beiträge dieser Komponenten zu verstehen und Muster und Verhaltensweisen innerhalb der Daten zu erkennen.
- Konzept auch wichtig für additive Modelle (z.B. Prophet)
- Zwei Ansätze:
 - Additive Zerlegung: $x_t = T_t + S_t + R_t$
 - Multiplikative Zerlegung: $x_t = T_t \times S_t \times R_t$
wobei T_t die Trend(-Zyklus)-Komponente, S_t die saisonale Komponente, R_t die Restkomponente

Komponenten von Zeitreihen

- **Trend:** Die Trendkomponente stellt das langfristige, anhaltende Verhalten oder die Richtung der Zeitreihe dar. Sie erfasst das allgemeine steigende, fallende oder stationäre Muster im Zeitverlauf und ignoriert die kurzfristigen Schwankungen und saisonalen Effekte.
- **Saisonal:** Die saisonale Komponente erfasst die regelmäßigen, sich wiederholenden Muster, die innerhalb einer Zeitreihe auftreten. Sie stellt die systematischen Schwankungen dar, die in festen Zeitintervallen auftreten, z. B. in täglichen, monatlichen oder jährlichen Zyklen. Die saisonale Komponente ist häufig durch ein festes Muster von Höchst- und Tiefstwerten gekennzeichnet, was auf die Saisonalität der Daten hinweist.
- **Residual (oder Fehler):** Die Residualkomponente steht für die unregelmäßigen und zufälligen Schwankungen, die nach Abzug der Trend- und Saisonkomponenten verbleiben. Sie umfasst die unvorhersehbaren oder unerklärlichen Schwankungen in den Daten, die nicht durch den Trend und die saisonalen Muster erklärt werden können.

Quiz

- Was ist eine Zeitreihe?
- Warum ist Autokorrelation wichtig für Zeitreihen?
- Wann ist eine Zeitreihe stationär?
- Nenne mindestens drei Komponenten, in die sich eine Zeitreihe zerlegen lässt.
- Sternchenfrage: Was sieht man in ACF und PACF-Plots?

Quiz Antworten

- Was ist eine Zeitreihe?
 - Z.B. Eine Zeitreihe ist eine Folge von Datenpunkten, die in zeitlicher Reihenfolge erfasst wurden.
- Warum ist Autokorrelation wichtig für Zeitreihen?
 - Autokorrelation ist wichtig für Zeitreihen, da sie misst, inwiefern ein Wert in der Reihe mit vorherigen Werten zusammenhängt. Dies ist z.B. entscheidend für Vorhersagemodelle.
- Wann ist eine Zeitreihe stationär?
 - Eine Zeitreihe ist stationär, wenn ihre statistischen Eigenschaften wie Mittelwert, Varianz und Autokorrelation über die Zeit konstant bleiben.
- Nenne mindestens drei Komponenten, in die sich eine Zeitreihe zerlegen lässt.
 - Eine Zeitreihe lässt sich in mindestens drei Komponenten zerlegen: Trendkomponente, saisonale Komponente und zufällige (oder residuale) Komponente.

Quiz Antworten

1.1 Grundlagen der Zeitreihenanalyse

- Sternchenfrage: Was sieht man in ACF und PACF-Plots?
 - ACF steht für Autokorrelationsfunktion und PACF für partielle Autokorrelationsfunktion. Beide sind grafische Darstellungen, die verwendet werden, um die Stärke und den Typ der Beziehung zwischen verschiedenen Zeitpunkten in einer Zeitreihe zu bestimmen.
 - ACF-Plot: Zeigt die Korrelation zwischen der Zeitreihe und ihrer selbst, verschoben um verschiedene Zeiteinheiten (sog. Lags). Der ACF-Plot gibt einen Überblick über die gesamte Korrelation zwischen verschiedenen Lags.
 - PACF-Plot: Zeigt die Korrelation, die nicht durch vorherige Lags erklärt wird. Im Wesentlichen misst der PACF-Plot die direkte Beziehung zwischen einem Beobachtungspunkt und einem anderen Punkt bei einer bestimmten Verzögerung, nachdem die Beziehungen zu den Punkten dazwischen bereinigt wurden.

CODING

Stromerzeugung in Deutschland nach Energieträger

- Visualisieren
- Autokorrelation
- Gleitende Mittel
- Komponentenzerlegung

Zeitreihen vorhersagen (Statistik I): Exponentielle Glättung und Holt-Winters

Session 3 (Montag 13:30 – 15:00)

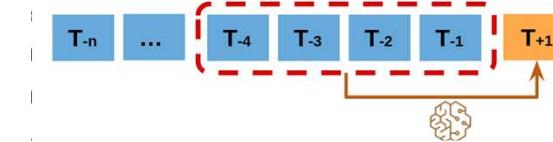
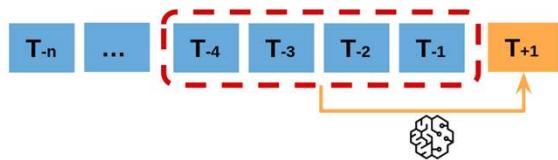


Zeitreihen vorhersagen (Statistik I): Exponentielle Glättung und Holt-Winters

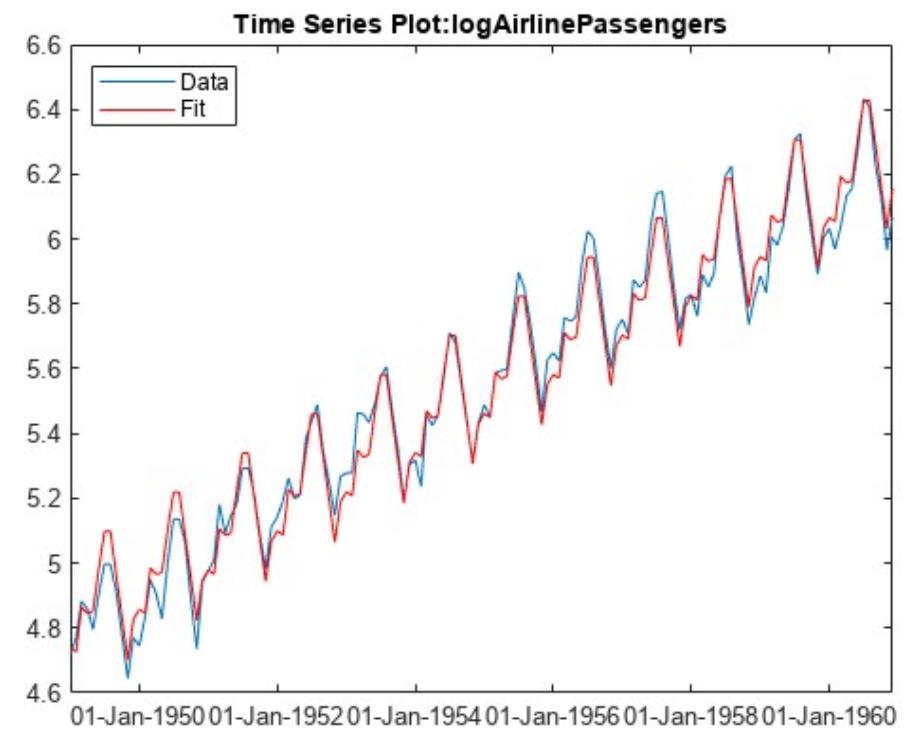
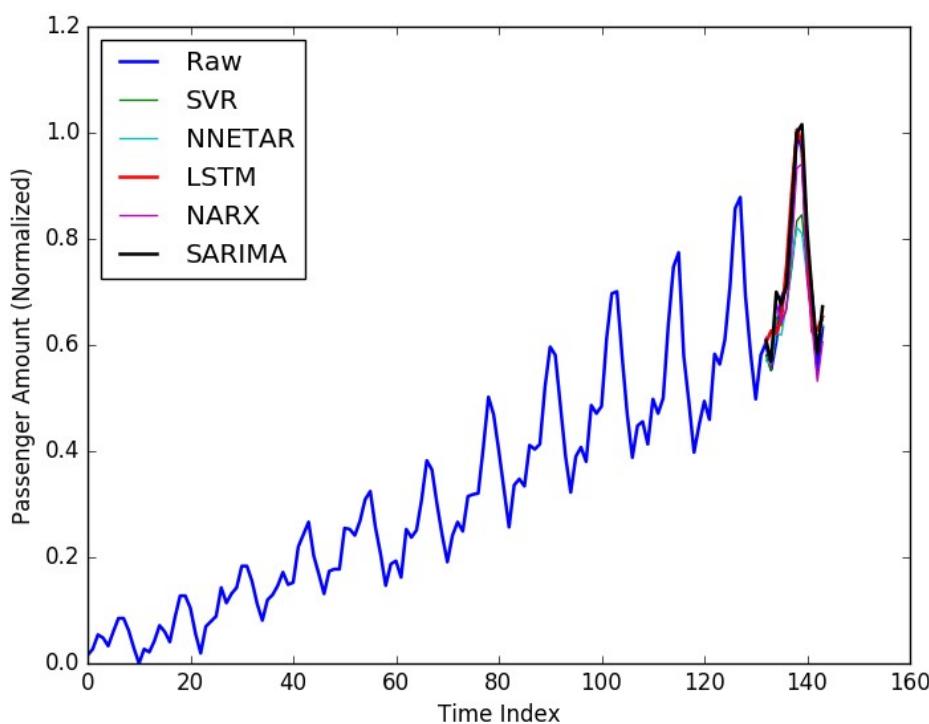
- Naive Vorhersagen
- einfaches exponentielles Glätten (SES)
- Holt-Verfahren / Holt-Winters-Verfahren

Vorhersagen in Zeitreihen

- Ein-Schritt- vs Mehr-Schritt-Vorhersagen



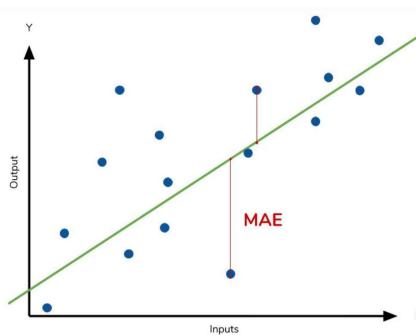
Vorhersagen bewerten



Vorhersagen bewerten

Zeitreihen Vorhersagen

- Um die **Genauigkeit und Güte einer Zeitreihenvorhersage zu messen**, werden üblicherweise verschiedene Bewertungsmaßstäbe verwendet. Die Wahl der Bewertungsmetrik hängt von den spezifischen Merkmalen der Daten und dem Ziel der Analyse ab.
- Der mittlere absolute Fehler oder der mittlere quadratische Fehler sind gängige Optionen.



$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

MAE = mean absolute error
 y_i = prediction
 x_i = true value
 n = total number of data points

Loss	Formula	Calculation
Mean-Squared error (MSE)	$\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$	$\frac{1}{2} ((17.112 - 16.5)^2 + (7.05 - 4.3)^2) = 7.937$
Mean absolute error (MAE)	$\frac{1}{N} \sum_{i=1}^N y_i - \hat{y}_i $	$\frac{1}{2} (17.112 - 16.5 + 7.05 - 4.3) = 3.362$

Vorhersagen treffen

Naive Vorhersage

$$\hat{x}_{t+h} = x_t$$

Wie können wir die verbessern?

Beispiel

Wie viel Solarstrom wird gerade auf dem Dach meines Hauses produziert?

Wenn ich Ihnen sagen würde, dass es vor einer Stunde 1,5 kW war?

Vorhersagen treffen

Naive Vorhersage II (Durchschnittsmethode)

$$\hat{x}_{t+h} = \frac{1}{t} \sum_{i=1}^t x_i$$

Wie können wir die verbessern?

Beispiel

Was ist der Preis von 1 MWh Strom an der Strombörsen?

Was wäre Ihre Schätzung, wenn ich Ihnen sage, dass er durchschnittlich bei 42.6 Euro/MWh liegt?

Exponentielle Glättung

- 1. Idee: Neuere Beobachtungen stärker zu gewichten als Beobachtungen aus der fernen Vergangenheit
- Dies ist genau das Konzept hinter der einfachen exponentiellen Glättung.
- Prognosen werden mit gewichteten Mittelwerten berechnet, wobei die Gewichte exponentiell abnehmen, wenn die Beobachtungen weiter in der Vergangenheit liegen:

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

Einfache Exponentielle Glättung (SES)

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

- Glättungsparameter alpha zwischen null und eins
- Die einfachste der exponentiell glättenden Methoden wird natürlich als einfache exponentielle Glättung („simple exponential smoothing“ SES) bezeichnet.
- Diese Methode eignet sich für die Prognose von Daten **ohne klaren Trend oder saisonales Muster**

- Holt Verfahren beinhaltet eine Trendkomponente
- Holt-Winter Verfahren: eine Trend- und eine Saisonkomponente

Holt's Methode (Double ES)

- Alternative Schreibweise SES

Forecast equation $\hat{y}_{t+h|t} = \ell_t$

Smoothing equation $\ell_t = \alpha y_t + (1 - \alpha)\ell_{t-1},$

- Holt Verfahren beinhaltet eine Trendkomponente

Forecast equation $\hat{y}_{t+h|t} = \ell_t + hb_t$

Level equation $\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$

Trend equation $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1},$

Manchmal wird die Trendkomponente auch **multiplikativ** ins Modell aufgenommen

Holt-Winter's Methode

- Holt Verfahren beinhaltet eine Trendkomponente

Forecast equation

$$\hat{y}_{t+h|t} = \ell_t + hb_t$$

Level equation

$$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$$

Trend equation

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1},$$

- Holt-Winter Verfahren: eine Trend- und eine Saisonkomponente

Man **muss** nicht
schlau aus
diesen
Gleichungen
werden

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$$

$$\ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + b_{t-1})$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1}$$

$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m},$$

CODING

Vorhersage Stromlast im Gebäude

- Vorhersagen generell
- Naive Vorhersage
- Exponentielle Glättung



Zeitreihen vorhersagen (Statistik II): ARIMA-Modelle

Session 4 (Dienstag 15:15 – 17:00)



Zeitreihen und ihre Merkmale visualisieren

- Autoregressive- (AR) und Moving Average (MA)
- Mit ARMA und ARIMA-Modellen vorhersagen
- Mit SARIMA-Modellen Saisonalität berücksichtigen

Das AR(p)-Modell “autoregressive”

Naïve Vorhersage

$$\hat{x}_{t+h} = x_t$$

Wie können wir die verbessern?

Ein **autoregressives Modell der Ordnung p** (AR(p)) kann geschrieben werden als

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \epsilon_t$$

wobei ϵ_t weißes Rauschen ist.

Bsp: AR(2)-Schätzer: $\hat{x}_{t+1} = c + \phi_1 x_t + \phi_2 x_{t-1}$

mit (zuvor) geschätzten Parametern c, ϕ_1, ϕ_2 und Daten x_t, x_{t-1}

Stationäre Daten → Beschränkungen für Parameter (Werden in der Praxis von Software übernommen)

Beispiel

Was ist der Preis von 1 MWh Strom (Intraday Preis) an der Strombörse in einer Stunde?

Was wäre Ihre Schätzung, wenn ich Ihnen sage, dass er gerade bei 42.6 Euro/MWh liegt?

Das MA(q)-Modell “moving average”

Naive Vorhersage II (Durchschnittsmethode)

$$\hat{x}_{t+h} = \frac{1}{t} \sum_{i=1}^t x_i$$

Wie können wir die verbessern?

Ein **gleitendes Durchschnittsmodell der Ordnung q** (MA(q)) kann geschrieben werden als

$$x_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t$$

wobei ϵ_t weißes Rauschen ist.

Beispiel

Wie viel Strom wird heute, am 25.2.2021, in Bayern aus Windkraft erzeugt?

Wenn ich Ihnen jetzt sage, dass es im Durchschnitt täglich 12,6 Gwh sind?

Wenn ich zusätzlich sage, dass meine Schätzung gestern +0,6 Gwh daneben lag?

MA-Prozesse sind besser darin, Abhängigkeiten in Beobachtungen zu modellieren, die kurz auftreten und sich dann wieder auflösen, während AR-Prozesse zeitliche Abhängigkeiten modellieren, die von Dauer über die Zeitreihe auftreten.

Das ARMA (p,q)-Modell

Ein **ARMA(p,q)-Modell** (AutoRegressive Moving Average) wird geschrieben als

$$x_t = c + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

wobei ϵ_t weißes Rauschen ist.

- Erweiterung: **ARIMA(p,d,q)-Modell** (AutoRegressive *Integrated* Moving Average)
- für nicht-stationäre Zeitreihen: ist erste Differenz $x_t - x_{t-1}$ stationär? d-mal wiederholen
- Erweiterung: **SARIMA(p,d,q)(P,D,Q)_m-Modell**
- Saisonales ARIMA-Modell
- Saisonalität m (z.B. 12 bei monatlichen Daten, 4 bei Quartalsdaten, ...)
- (m = Saisonalität, p = Ordnung des autoregressiven Teils, d = Grad der benötigten ersten Differenzen, q = Ordnung des gleitenden Durchschnittsteils – Großbuchstaben für den saisonalen Teil)

Das ARIMA (p,d,q)-Modell

Differenzen- und Backshiftoperator

$$\Delta_h x_t = x_t - x_{t-h},$$

$$Bx_t = x_{t-1}$$

$$\text{also: } \Delta_1 x_t = x_t - x_{t-1} = (1 - B)x_t$$

Beispiel

Ich messe die Raumtemperatur in meinem Schlafzimmer jede Stunde. Es ist neun Uhr und ich will die Temperatur für zehn Uhr vorhersagen. Welche Informationen hätten Sie gern von mir?

Ein **ARIMA(p,d,q)-Modell** (AutoRegressive Integrated Moving Average) wird geschrieben als

$$x'_t = c + \phi_1 x'_{t-1} + \cdots + \phi_p x'_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

wobei ϵ_t weißes Rauschen und $x'_t = (1 - B)^d x_t$ ist.

p = Ordnung des autoregressiven Teils, d = Grad der benötigten ersten Differenzen,
q = Ordnung des gleitenden Durchschnittsteils.

Das SARIMA(p,d,q)(P,D,Q)_m Modell

Erweiterung: Saisonales ARIMA-Modell

SARIMA(p,d,q)(P,D,Q)_m

Beispiel

Ich messe die Raumtemperatur in meinem Schlafzimmer jede Stunde. Es ist neun Uhr und ich will die Temperatur für zehn Uhr vorhersagen. Welche Informationen hätten Sie gern von mir?

- Beispiel SARIMA(1,1,1)(1,1,1)₄

$$(1 - \phi_1 B) (1 - \Phi_1 B^4) (1 - B) (1 - B^4) y_t = (1 + \theta_1 B) (1 + \Theta_1 B^4) e_t.$$

(Non-seasonal)
AR(1)

(Seasonal)
AR(1)

(Non-seasonal)
difference

(Seasonal)
difference

(Non-seasonal)
MA(1)

(Seasonal)
MA(1)

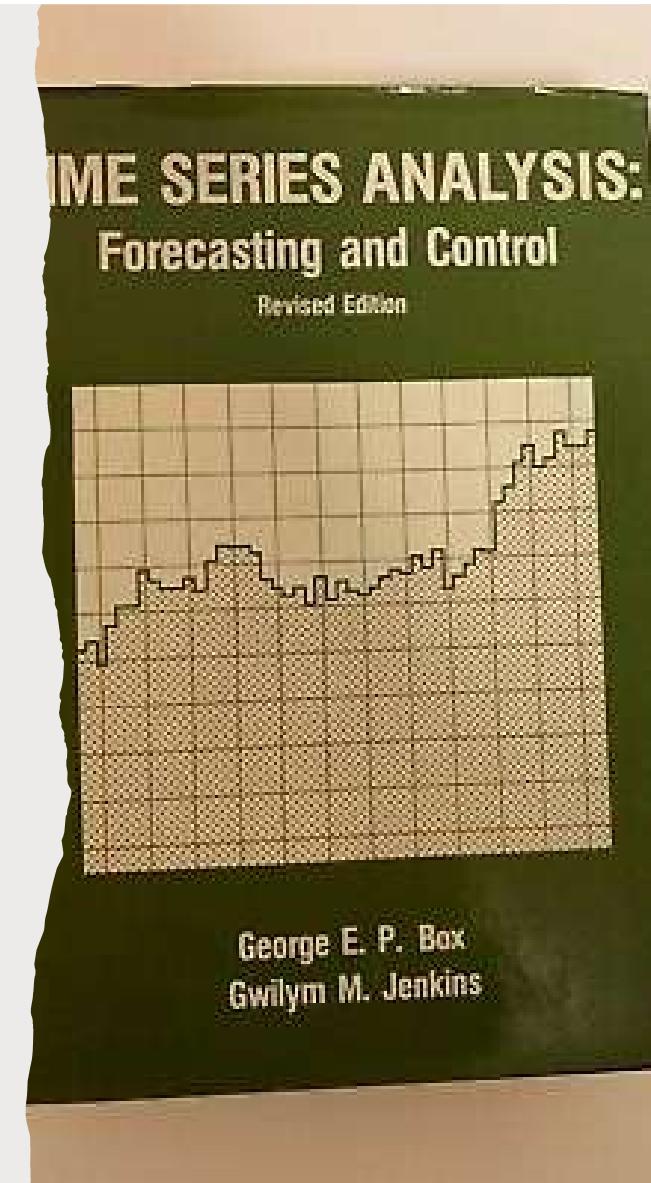


Box-Jenkins-Methode: Mit SARIMA Vorhersagen

"Alle Modelle sind falsch, aber
einige sind nützlich,"

**George Edward Pelham
Box (1919 - 2013)**
britischer Statistiker

Image source: Wikipedia, xyz



Box-Jenkins-Methode: Mit SARIMA Vorhersagen

1. Modellidentifikation

- Stationär? Sonst d Differenzen bilden
- Passende p und q wählen (ACF, PACF; AIC, AICc, BIC)

2. Schätzung

- Software schätzt $c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \Phi_1, \dots, \Phi_p, \Theta_1, \dots, \Theta_q$

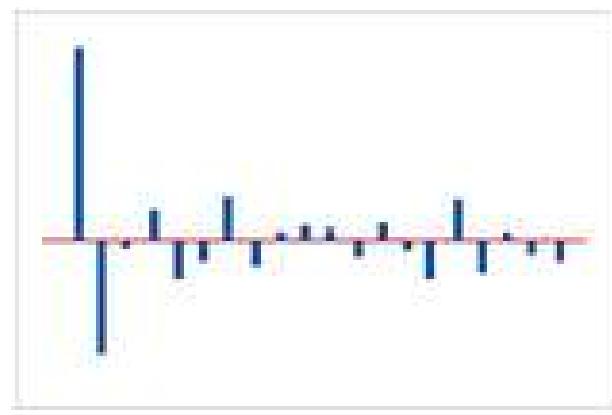
3. Validierung

- z.B. prüfen, ob die Residuen, also die geschätzten ϵ_t unkorreliert sind und sich wie weißes Rauschen verhalten

4. Anwendung: Vorhersage

- Einschritt-Prognose: Differenzengleichung des geschätzten ARMA-Modells eine Periode in die Zukunft schieben und den bedingten Erwartungswert berechnen.
- Mehrschritt-Prognosen: dies rekursiv wiederholen

ACF und PACF nutzen



ACF

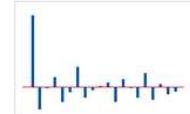
- Die Autokorrelationsfunktion ist ein Maß für die Korrelation zwischen Beobachtungen einer Zeitreihe, die durch k Zeiteinheiten (y_t und y_{t-k}) getrennt sind.

Muster	Bedeutung	Beispiel
Großer Ausschlag bei Lag 1, der nach ein paar Lags abnimmt.	Ein autoregressiver Term in den Daten. Verwenden Sie die partielle Autokorrelationsfunktion, um die Ordnung des autoregressiven Terms zu bestimmen.	
Großer Ausschlag bei Lag 1, gefolgt von einer abnehmenden Welle, die zwischen positiven und negativen Korrelationen wechselt.	Ein Autoregressionsterm höherer Ordnung in den Daten. Verwenden Sie die partielle Autokorrelationsfunktion, um die Ordnung des autoregressiven Terms zu bestimmen.	
Signifikante Korrelationen beim ersten oder zweiten Lag, gefolgt von Korrelationen, die nicht signifikant sind.	Ein gleitender Durchschnittsterm in den Daten. Die Anzahl der signifikanten Korrelationen gibt die Ordnung des gleitenden Durchschnittsterms an.	

<https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/autocorrelation/interpret-the-results/autocorrelation-function-acf/>

PACF

- Die partielle Autokorrelationsfunktion ist ein Maß für die Korrelation zwischen Beobachtungen einer Zeitreihe, die durch k Zeiteinheiten (y_t und y_{t-k}) voneinander getrennt sind, nach Bereinigung um alle anderen Terme mit kürzeren Lags ($y_{t-1}, y_{t-2}, \dots, y_{t-k-1}$).

Muster	Bedeutung	Beispiel
Großer Ausschlag bei Lag 1, der nach ein paar Lags abnimmt.	Ein gleitender Durchschnittsterm in den Daten. Verwenden Sie die Autokorrelationsfunktion, um die Ordnung des gleitenden Durchschnittsterms zu bestimmen.	
Großer Ausschlag bei Lag 1, gefolgt von einer abnehmenden Welle, die zwischen positiven und negativen Korrelationen wechselt.	Ein gleitender Durchschnittsterm höherer Ordnung in den Daten. Verwenden Sie die Autokorrelationsfunktion, um die Ordnung des gleitenden Durchschnittsterms zu bestimmen.	
Signifikante Korrelationen beim ersten oder zweiten Lag, gefolgt von Korrelationen, die nicht signifikant sind.	Ein autoregressiver Term in den Daten. Die Anzahl der signifikanten Korrelationen deutet auf die Ordnung des autoregressiven Terms hin.	

[to/autocorrelation/interpret-the-results/autocorrelation-function-pacf](#)

CODING

Vorhersage Stromlast im Gebäude

- Vorhersagen generell
- AR, MA, ARMA, ARIMA,
- SARIMA
- Vergleich (Errormaße)



Analyse und Visualisierung von Zeitreihen-Daten in Python

Tag 2



Inhalte – Was haben wir vor?

Tag 1

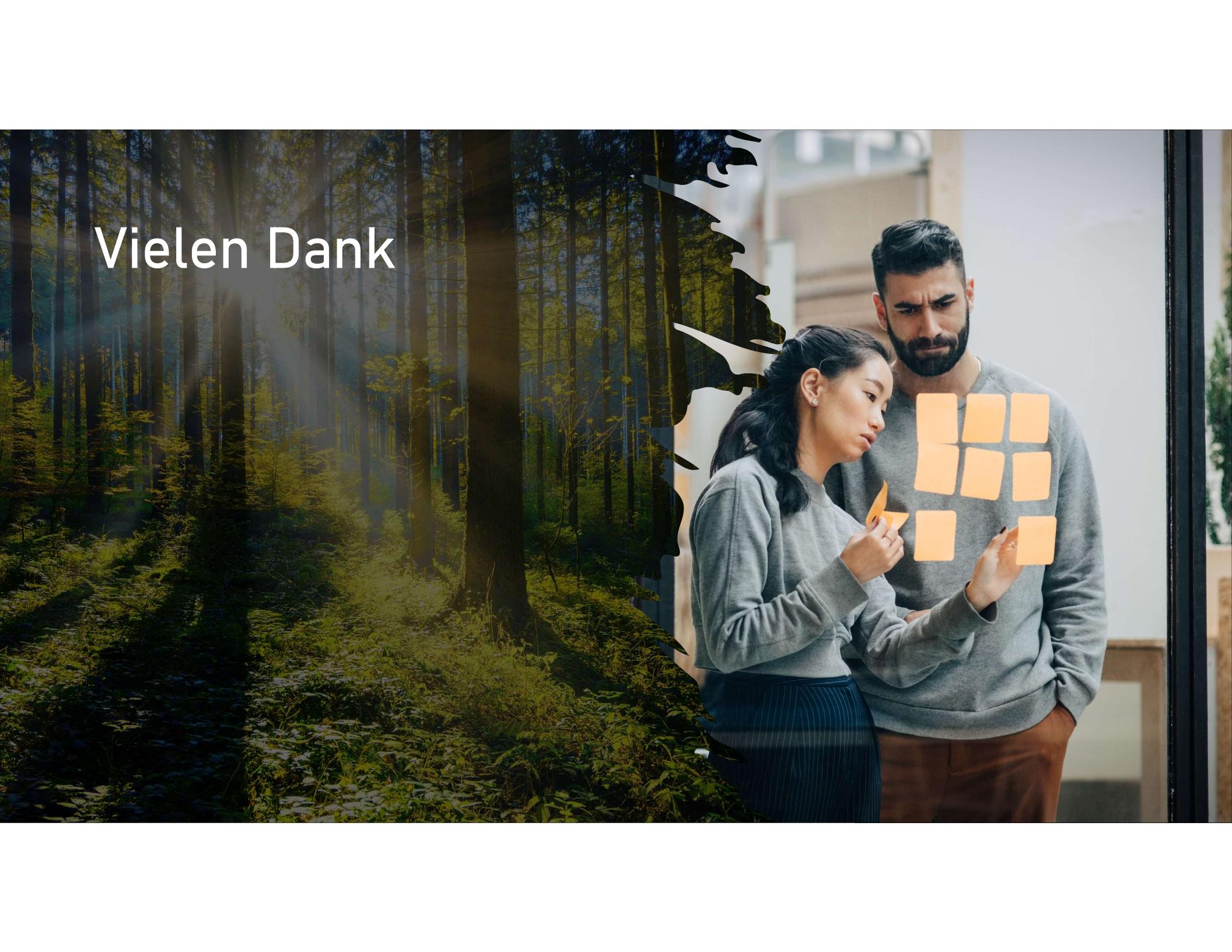
1. Einführung in Zeitreihendaten in Python
2. Zeitreihen und ihre Merkmale visualisieren
3. Zeitreihen vorhersagen (Statistik I): Exponentielle Glättung und Holt-Winters
4. Zeitreihen vorhersagen (Statistik II): ARIMA-Modelle

Tag 2

5. Einblick in andere Zeitreihenmodelle
6. Machine Learning für Zeitreihen: Überblick, Vorbereitung und Vorhersagen
7. Machine Learning für Zeitreihen: Clustering und Klassifikation
8. Deep Learning für Zeitreihen

Feedback



A composite image. On the left, a dense forest with tall trees and sunlight filtering through the canopy. On the right, two people, a man and a woman, are standing in an office setting. They are looking at a whiteboard that has several orange sticky notes pinned to it. The man has a beard and is wearing a grey sweatshirt. The woman is wearing a grey sweatshirt and blue pants. They appear to be discussing something.

Vielen Dank