Hochschule
für Technik
Stuttgart

Bachelorarbeit

vorgelegt von
Thanh Phan Vu
Bachelorstudiengang Mathematik
Fakultät Vermessung, Informatik und Mathematik
Sommersemester 2023

Anwendung von Zeitreihenmodellen auf spezielle Datensätze mit Hilfe von Python

ErstprüferIn: Prof. Dr. Annegret Weng
ZweitprüferIn: Prof. Dr. Harald Bauer

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Abschlussarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Alle Stellen, die wörtlich oder sinngemäß aus den Quellen entnommen wurden, sind als solche kenntlich gemacht. Weiterhin erkläre ich, dass die Arbeit nicht anderweitig veröffentlicht oder an anderer Stelle als Prüfungsleistung vorgelegt wurde.

Stuttgart, den

_____

Unterschrift

Danksagung

Ich möchte mich ganz herzlich bei Frau Dr. Prof. Annegret Weng bedanken. Ohne ihren Themenvorschlag, ihr Material und ihre intensiven Tipps und kritischen Ideen wäre diese Arbeit nicht möglich gewesen. Ich möchte auch dem DWD-Deutschen Wetterdienst für die Bereitstellung eines großen Temperaturdatensatzes danken. Und ich möchte auch Frau Yvonne Bui für die große Hilfe beim Korrekturleseprozess danken.

Zusammenfassung

Mit der Entwicklung von Technologien, die den Prozess der Datenanalyse und -speicherung erheblich vereinfachen, hat sich in den letzten Jahren die Nachfrage nach Datenanalyse und Datenerfassung deutlich erhöht. Zeitreihen stellen wahrscheinlich die gebräuchlichste Form von Daten dar, da sie eine Sammlung von Beobachtungen über einen bestimmten Zeitraum repräsentieren. In dieser Arbeit führen und untersuchen wir einige wichtige Aspekte der Zeitreihenanalyse und -prognose, indem wir grundlegende Zeitreihenmodelle vorstellen und das SARIMA-Modell zur Modellierung und Prognose der durchschnittlichen monatlichen Lufttemperatur einsetzen. Auf der einen Seite ermöglicht die Anwendung auf die durchschnittliche monatliche Lufttemperatur ein tieferes Verständnis der Zeitreihenanalyse, insbesondere der Box-Jenkins-Methode. Die Daten weisen neben einer geringen Menge stochastischer Komponenten auch eine starke saisonale und eine subtile Trendkomponente auf, die den Prognoseprozess nicht übermäßig komplizieren, aber dennoch viele wichtige Faktoren in der Zeitreihenanalyse und -prognose abdecken. Auf der anderen Seite ermöglicht uns die Verwendung des SARIMA-Modells, die Leistungsfähigkeit und die Grenzen dieses Modells für bestimmte Daten - in diesem Fall die durchschnittliche monatliche Lufttemperatur - zu beurteilen. Anhand dieser Informationen kann dann auch eine Entscheidung getroffen werden, ob die Methode auf Daten mit einem bestimmten statistischen Fehler angewendet werden kann.

Die Arbeit wird in drei Teile gegliedert. Der erste Teil liefert eine Einführung in einige grundlegende und wichtige Zeitreihenmodelle und stellt Definitionen und Methoden zur Untersuchung der statistischen Charakteristiken einer Zeitreihe vor. Der zweite Teil beschreibt komplexere Zeitreihenmodelle und zielt darauf ab, das SARIMA-Modell zu präsentieren, das auf die monatliche Durchschnittstemperatur der Luft angewendet wird. Der letzte Teil wendet die Box-Jenkins-Methodik auf das SARIMA-Modell an, um auf Basis der Daten der monatlichen Durchschnittstemperatur eine Prognose für die Temperatur der nächsten zwei Jahre zu erstellen.

Inhaltsverzeichnis

Kapitel 1

Einführung

Dieses Kapitel wird eine Einführung in einige der grundlegenden und wichtigen Zeitreihenmodelle sowie einige Definitionen und Methoden zur Untersuchung der statistischen Merkmale einer Zeitreihe bieten.
Die folgende Einführung basiert auf den Inhalten im Buch von Peter J. Brockwell und Richard A. Davis. (Siehe [4].)

1.1 Definition und erstes Beispiel einer Zeitreihe

Definition 1.1. *Sei t ein Element der Menge $T = \{t_1, t_2, ..., t_n\}$ mit $t_1 < t_2 < ... < t_n$ und sei x_t eine Variable, die einem bestimmten Zeitpunkt t eine bestimmte Ausprägung zuordnet. Eine Zeitreihe $\{X_t\}$ besteht dann aus einer Folge von n Beobachtungen $x_{t_1}, x_{t_2}, ..., x_{t_n}$, die zu den diskreten Zeitpunkten $t_1, t_2, ..., t_n$ aufgenommen wurden und in zeitlicher Reihenfolge indiziert sind.(vgl. [4], S. 1):*

Bemerkung: Das folgende Konzept soll von Anfang an klar definiert werden. In Definition 1.1 definieren wir eine Zeitreihe $\{X_t\}$ als eine Folge von n Beobachtungen $(x_1, x_2, \cdots, x_n)$. Es ist zu bemerken, dass eine Zeitreihe $\{X_t\}$ auch als Folge von n Zufallsvariablen $(X_1, X_2, \cdots, X_n)$ definiert werden kann.

Hier ist ein Beispiel für den Schlusskurs des Nasdaq-Index vom 10.1.2000 bis zum 08.02.2000. Wir können sehen, dass jedem Tag ein Index zugeordnet wird. Wir haben also eine Zeitreihe, bei der eine Periode einem Tag entspricht und die Datenpunkte die täglichen Kurse sind.

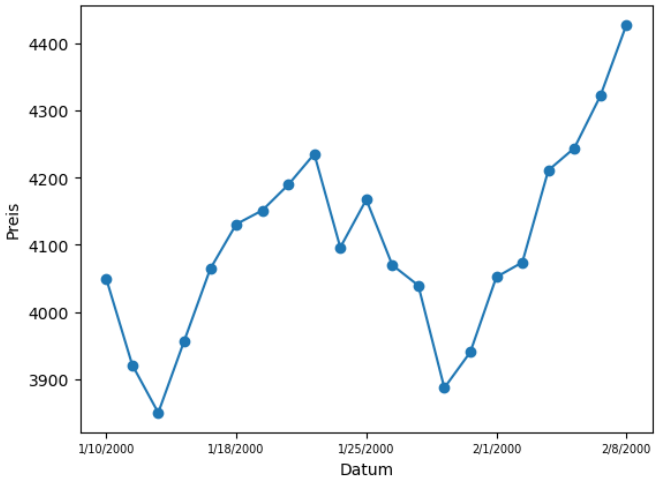

Abbildung 1.1: Nasdaq täglicher Schlusskurs von 10.1.2000- 8.2.2000

Zeitreihen können in verschiedenen Bereichen erhoben werden. In der Wirtschaft haben wir zum Beispiel CPI-Daten und Arbeitslosenzahlen. In der Natur haben wir Wetterdaten, Ozeangezeiten, Daten zu Sonnenflecken und so weiter. Dies sind nur wenige Beispiele aus einer Vielzahl von Zeitreihen, die in unterschiedlichen Feldern anzutreffen sind. Mithilfe geeigneter Methoden und Modelle können wir aus den gesammelten Zeitreihen aussagekräftige Statistiken und Merkmale extrahieren und zukünftige Werte basierend auf früheren Beobachtungswerten vorhersagen.

1.2 Einführung in Zeitreihenmodelle

In diesem Teil geht es um einige einfache, aber sehr wichtige Zeitreihenmodelle, die als grundlegende Bausteine für den Aufbau komplexerer Zeitreihenmodelle dienen. In diesem Abschnitt werden wir uns auf die Zeitreihe, die im Allgemeinen mit einem stochastischen Prozess beschrieben wird. Der Abschnitt wird auch in zwei Kategorien unterteilt sein, einer mit Zeitreihen mit Nullmittelwert und konstanten statistischen Eigenschaften und einer mit Zeitreihen, die sich statistisch verändern.

1.2.1 Zeitreihen mit einem konstanten statistischen Charakter

1.2.1.1 Unabhängig und identisch verteilt (iid Daten)

Definition 1.2. *Eine Folge von Zufallsvariablen $X_1, \ldots, X_n$, die die beiden Bedingungen Unabhängigkeit und identische Verteilung erfüllt, heißt unabhängig und identisch verteilt oder i.i.d und wir definieren die Zufallsvariablen $X \sim IID(\mu, \sigma^2)$. (vgl. [4], Beispiel 1.3.1 auf S. 6):*

Aufgrund der Charakteristik von iid-Daten ist es unmöglich, den Wert von X_{n+h} mit der Kenntnis von $X_1, \cdots, X_n$ vorherzusagen. Das heißt, das Modell ist für Prognostiker ein eher uninteressanter Prozess, spielt aber als Baustein für kompliziertere Zeitreihenmodelle eine wichtige Rolle.

Beispiel 1.1. Zum Beispiel hängt der Aktienkurs heute vom Aktienkurs gestern ab, und die Volatilität kann sich im Laufe der Zeit ändern, was eine Änderung der zugrunde liegenden Verteilung impliziert.

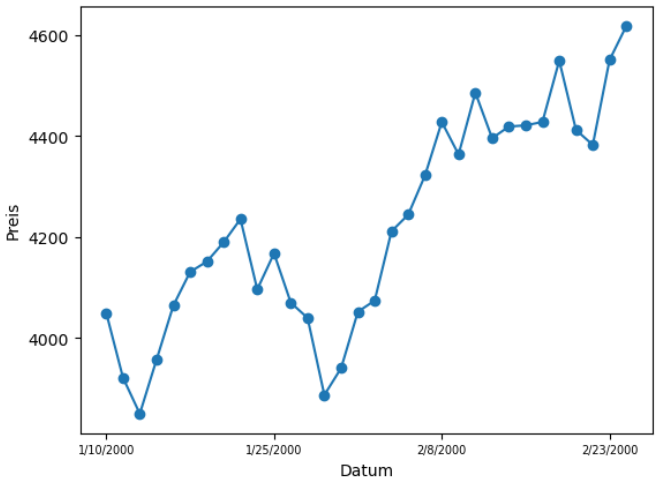

Abbildung 1.2: Nasdaq täglicher Schlusskurs vom 10.1.2000 bis zum 25.2.2000.

In diesen Fällen wollen wir keine Annahme von i.i.d. Daten und haben andere Annahmen, die weder unabhängig noch identisch verteilt sind.

1.2.1.2 White Noise

White Noise ist ein wichtiges Konzept in der Zeitreihenanalyse und -vorhersage. Wie iid Daten ist auch White Noise kein Modell zur Vorhersage zukünftiger Werte, sondern dient vielmehr als wichtiger Baustein

für kompliziertere Zeitreihenmodelle. Wir können White Noise auch als eine "schwächere" Definition von iid sehen, der Grund dafür wird erklärt. Zuerst die Definition:

Definition 1.3. *Eine Zeitreihe ist White Noise, wenn die Variablen unkorreliert sind, einen Mittelwert von Null haben und eine konstante Varianz besitzen. Das bedeutet, dass wir eine Definition von Zufallsvariablen $X \sim WN(0, \sigma^2)$ haben. (vgl. [4], Beispiel 1.4.2 auf S. 14):*

Bemerkung: Die Begriffe "unkorreliert" und "unabhängig" sind zwei gut spezifizierte mathematische Begriffe und sie bedeuten nicht dasselbe:

- **Unabhängige Zufallsvariablen sind immer unkorreliert.**

  Beweis. Seien X und Y zwei unabhängige Zufallsvariablen.
  Die Erwartungswert von zwei Zufallsvariablen ist definiert als: [13]

$$\begin{aligned} E[XY] &= \int\int xyf(x,y)dxdy \\ &= \int\int xyf(x)f(y)dxdy \\ &= \int_x xf(x)dx \cdot \int_y yf(y)dy \\ &= E[X]E[Y] \\ &\Leftrightarrow E[XY] - E[X]E[Y] = 0 \end{aligned}$$

  Die Kovarianz von X und Y berechnet sich wie folgt:: [16]

$$\begin{aligned} cov(X,Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[(XY - XE[Y] - YE[X] + E[X]E[Y])] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

  Das bedeutet, dass $cov(X,Y) = 0$ ist und dass X und Y unkorreliert sind. □

- **Unkorrelierte Zufallsvariaben sind nicht unbedingt unabhängig.**

  Beispiel 1.2. Es sei $\{Zt\}$ ein iid-Noise mit N(0,1)-Verteilung. Definiert wird:

$$X_t = \begin{cases} Z_t & \text{wenn } t \text{ gerade ist} \\ \frac{Z_{t-1}^2 - 1}{\sqrt{2}} & \text{wenn } t \text{ ungerade ist} \end{cases}$$

  Es wird gezeigt, dass X_t ein WN(0,1) ist, aber kein iid-(0,1)-Noise ist.(vgl. [4], Aufgabe 1.8 auf S. 35):

  Lösung:

  * Zunächst wird der Erwartungswert von X_t berechnet:
    Für t gerade:
$$E[X_t] = E[Z_t] = 0.$$
    Für t ungerade:
$$E[X_t] = \frac{1}{\sqrt{2}}(E[Z_{t-1}^2 - 1]) = 0.$$

  * Anschließend suchen wir die Autokovarianzfunktion (ACVF)(Siehe Definition 1.4:
    Für Lag $h = 0$ und t ist gerade:
$$\gamma_X(t,t) = E[Z_t^2] = 1$$

Für Lag h = 0 und t ist ungerade:

$$\begin{aligned}\gamma_X(t,t) &= E[(\frac{Z_{t-1}^2-1}{\sqrt{2}})^2] \\ &= \frac{1}{2}E[Z_{t-1}^4 - 2Z_{t-1}^2 + 1] \text{ (Moment 4 ist Wölbung)} \\ &= \frac{1}{2}(3-2+1) \\ &= 1\end{aligned}$$

Für Lag h = 1 und t ist gerade:

$$\begin{aligned}\gamma_X(t+1,t) &= E[Z_{t+1}Z_t] \\ &= E[\frac{Z_t^2-1}{\sqrt{2}}Z_t] \\ &= \frac{1}{\sqrt{2}}E[Z_t^3 - Zt] \text{ (Moment 3 ist Schiefe)} \\ &= \frac{1}{\sqrt{2}}(0-0) = 0\end{aligned}$$

Für Lag h = 1 und t ist ungerade:

$$\begin{aligned}\gamma_X(t+1,t) &= E[Z_{t+1}Z_t] \\ &= E[Z_{t+1}\frac{Z_{t-1}^2-1}{\sqrt{2}}] \\ &= \frac{1}{\sqrt{2}}E[Z_{t+1}]E[Z_{t-1}^2 - 1] \text{ (Zwei unabhängige Zufallsvariablen)} \\ &= 0\end{aligned}$$

Offensichtlich $\gamma_X(t+h,t) = 0$ für $|h| \geq 2$, Somit:

$$\gamma_X(t+h,t) = \begin{cases} 1 & \text{wenn } h = 0 \\ 0 & \text{wenn } h \neq 0 \end{cases}$$

Daher ist X_t ein WN(0,1). Wenn t ungerade ist, sind X_t und X_{t-1} offentsichtlich voneinander abhängig, daher ist X_t nicht iid(0,1).
Basierend auf dem Unterschied zwischen Korrelation und Unabhängigkeit können wir sagen, dass ein iid(0,1) immer auch ein WN(0,1) ist. Aber ein WN(0,1) ist nicht immer ein iid(0,1).

Der Graph von X_t mit t im Intervall [0, 100]:

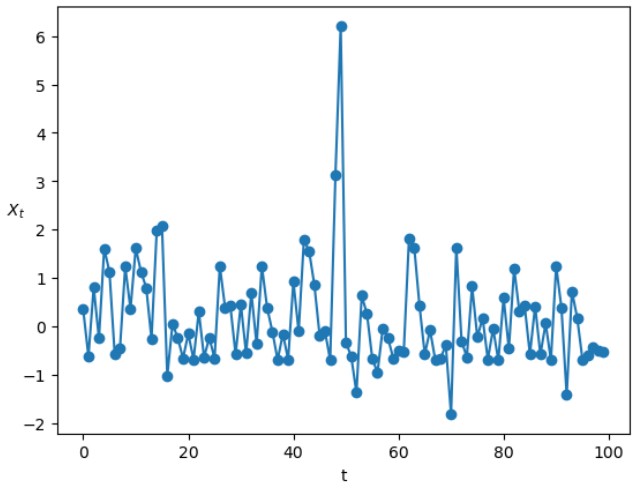

Abbildung 1.3: Plot für obiges Beispiel

Die Autokovarianz für Lag h im Intervall [0, 100] ist im unteren Diagramm dargestellt. Eine Anmerkung: Bei h = 0 entspricht die Autokovarianz der Varianz von X_t

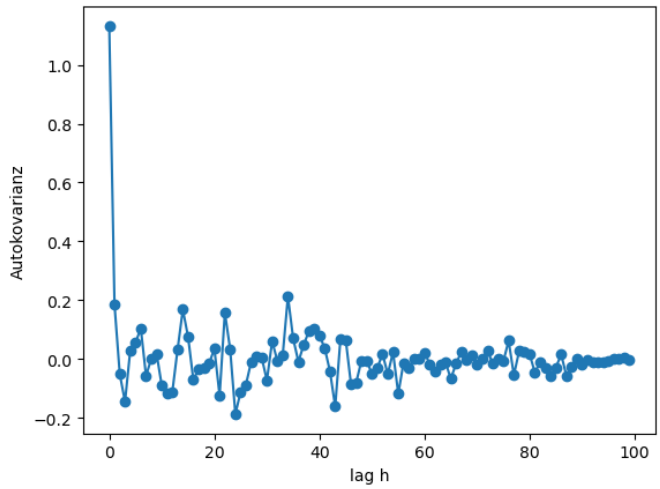

Abbildung 1.4: Autokovarianz für Lag h = [0,100]

1.2.1.3 Random-Walk

Ein Random-Walk ist ein Spezialfall des autoregressiven Modells [12] (Definition in nächste Kapitel) mit der Ordnung $p = 1$. Die Zeitreihe $\{X_t\}$ kann durch einen Random-Walk-Prozess modelliert werden, wenn die Zufallsvariablen X_t wie folgt beschrieben werden können:

$$X_t = \alpha + \phi \cdot X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2). \tag{1.1}$$

Oder die Gleichung kann auch als folgende definieren:

$$X_t = \alpha \cdot t + \sum_{t=1}^{t} \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2). \tag{1.2}$$

- $\alpha \neq 0$ liefert einen Random-Walk mit Drift. Wenn $\alpha = 0$ haben wir einen Random-Walk ohne Drift.
- ϕ ist ein Koeffizient, der angibt, mit welchem Gewicht die Zufallsvariable X_{t-1} zur Beschreibung der Zufallsvariable X_t beiträgt.
- ε_t entspricht einer Beobachtung zum Zeitpunkt t eines WN-Prozesses.

Wir nehmen an, dass $x_0 = 0$. Falls $\alpha \neq 0$ ist, gilt für Gleichung 1.1 der Erwartungswert $E[X_t] = t\alpha$. Falls $\alpha = 0$ ist, beträgt der Erwartungswert $E[X_t] = 0$.

Beispiel 1.3. Wir nehmen an, dass $x_0 = 0$. Es werden 100 Beobachtungen mit dem Random-Walk-Prozess mit $\phi = 0.7$ und $\alpha = 0$ durch Python erstellt:

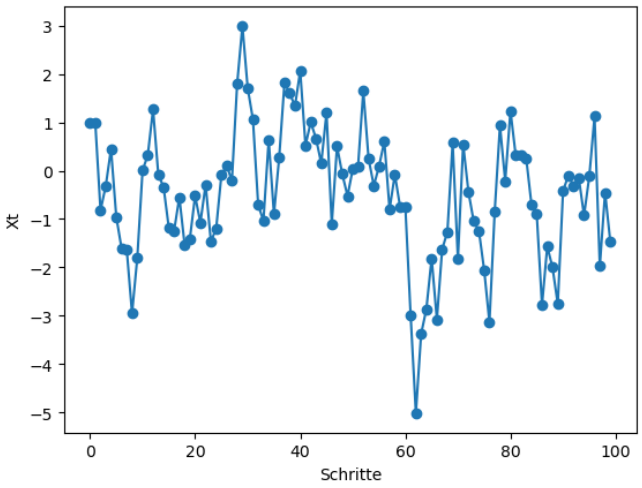

Abbildung 1.5: Ein Random-Walk

Wir können sehen, dass sich das Modell von einer Zufallszahl unterscheidet. Tatsächlich ähneln die erzeugten Daten sehr den Schwankungen von Aktienkursen, die bekanntlich schwer vorherzusagen sind.

1.2.2 Modell mit Trend und Saisonalität

Bisher wurden nur zeitliche Schwankungen diskutiert, die keine klaren Trends oder saisonale Muster aufweisen. Es gibt jedoch auch Zeitreihen, die eine eindeutige Trend- und/oder Saisonalitätskomponente aufweisen. Es ist wichtig, diese beiden Merkmale zu beachten, da sie eine wichtige Rolle beim Aufbau eines Prognosemodells spielen. Ein Beispiel für eine Zeitreihe mit Trend wäre die Lebenserwartung des Menschen, und ein Beispiel für eine Zeitreihe mit Saisonalität wären Temperaturdaten (aufgrund der globalen Erwärmung oder zyklischer Klimaveränderungen könnte dieser Datensatz sogar eine Trendkomponente enthalten).

1.2.2.1 Zeitreihe mit Trendkomponente

Eine Zeitreihe, die eine Trendkomponente aufweist, kann als eine Zeitreihe verstanden werden, bei der die Beobachtungen über einen ausreichend langen Zeitraum hinweg ansteigen oder abnehmen.
Es ist flexibel bei der Auswahl einer Zeitreihe mit Trendkomponente, da der Differenzparameter uns unterschiedliche Trendkomponenten liefert und somit Zeitreihen mit unterschiedlichem statistischem Charakter liefert. Zum Beispiel ist eine Zeitreihe $\{X_t\}$ wie folgt definiert: (vgl. [4], Abschnitt 1.3.2 auf S. 8):

$$X_t = m(t) + \varepsilon_t, \quad \varepsilon_t \sim WN(0, 1). \tag{1.3}$$

- $m(t)$ ist eine sich ändernde Funktion, die als Trendkomponente bekannt ist.
- ε_t ist ein White-Noise-Prozess.

Es werden 10 Beobachtungen von der Zeitreihe, die in der Gleichung 1.3 definiert, durch Python erstellt. Wir setzen die Trendkomponente gleich einer quadratischen Funktion $m(t) = \frac{1}{3}t^2$ und das gibt uns die folgende Zeitreihe:

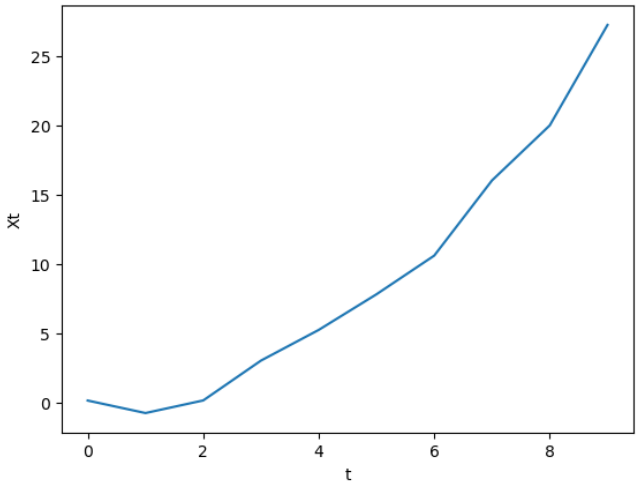

Abbildung 1.6: Zeitreihe mit Trendkomponente $m(t) = \frac{1}{3}t^2$

Die erstellte Zeitreihe in Abbildung 1.6 zeigt einen starken Aufwärtstrend aufgrund der quadratischen Trendkomponente mit einem positiven Koeffizienten und der Tatsache, dass die Trendkomponente von t abhängt, der nur inkrementiert wird.

Wir setzen die Trendkomponente $m(t) = 0.5 \cdot X_{t-1}$ und erhalten die folgende Zeitreihe:

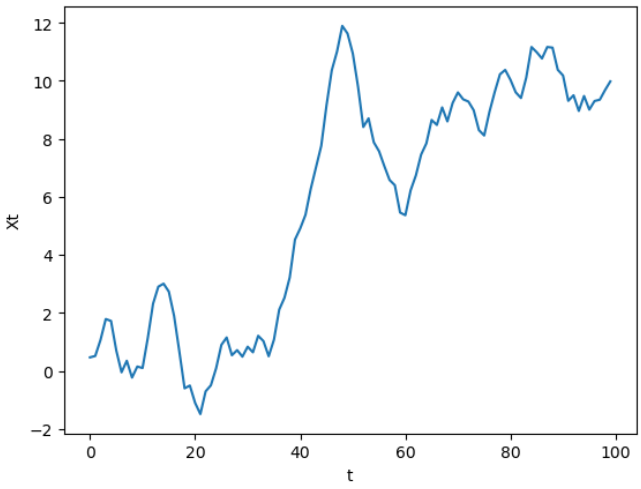

Abbildung 1.7: Zeitreihe mit Trendkomponente $m(t) = 0.5 \cdot X_{t-1}$

Jetzt hängt die Trendkomponente vom letzten Element der Zeitreihe ab und zeigt daher innerhalb der Zeitreihe einen vielfältigeren Trend.

1.2.2.2 Saisonalität

Analog zu Zeitreihen mit Trendkomponente können Zeitreihen auch eine saisonale Komponente haben. Wir definieren eine Zeitreihe mit der gleichen Komponente wie oben, aber hier wird Trendkomponent $m(t)$ nicht mehr da, sondern eine Saisonalitätskomponente $s(t)$: (vgl. [4], Abschnitt 1.3.2 auf S. 11)

$$X_t = s(t) + \varepsilon_t, \quad \varepsilon_t \sim WN(0,1). \tag{1.4}$$

Es werden 200 Beobachtungen von der Zeitreihe 1.4 durch Python erstellt. Wir setzen die Saisonalitätskomponente gleich einer sinus-Funktion $s(t) = A \cdot sin(2\pi t/p)$ mit einer Amplitude von 1 und einer Periode von 30. Das gibt uns die folgende Zeitreihe:

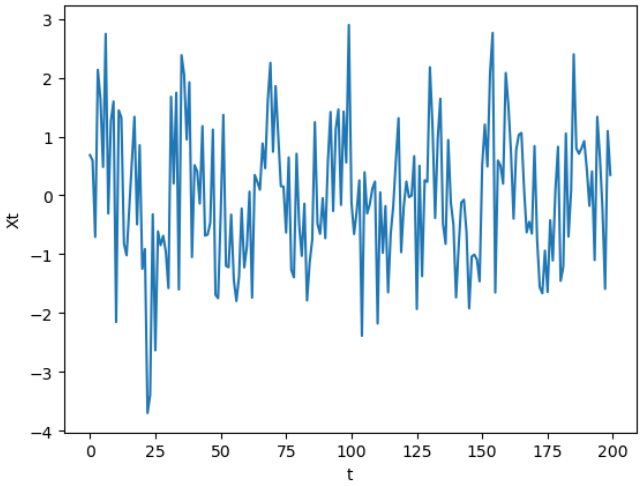

Abbildung 1.8: Zeitreihe mit Saisonalitätskomponente $s(t) = sin(2\pi t/30)$

1.2.2.3 Modell mit Trend und Saisonalität

Zeitreihen können auch Trend- und Saisonalitätskomponenten enthalten. Ein Beispiel dafür sind die monatlichen Temperaturdaten, bei denen es eine Tendenz geben kann, dass aufeinanderfolgende Jahre aufgrund bestimmter Trends und Saisonalitätsmuster wärmer sind als üblich.
Um eine Zeitreihe mit Trend und Saisonalität zu erstellen, können wir einfach die beiden Komponenten zur Zeitreihe hinzufügen, wie folgt:

$$X_t = m(t) + s(t) + \varepsilon_t, \quad \varepsilon_t \sim WN(0,1). \tag{1.5}$$

$s(t)$ wird dieselbe sein wie im obigen Beispiel, $m(t) = \frac{1}{300}t^2$. Wir teilen t durch 100, um die Wirkung des starken Trends zu verringern, den wir im ersten Beispiel des Abschnitts 1.2.2.1 erstellt haben. Das gibt uns folgende Zeitreihe:

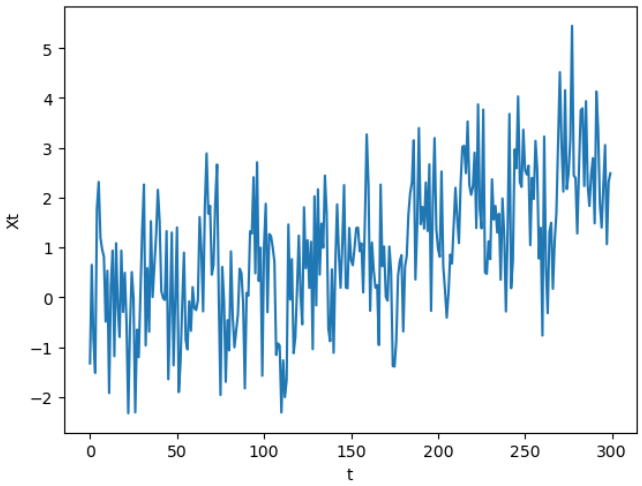

Abbildung 1.9: Zeitreihe mit Trend und Saisonalität

Wir bauen unser Vorhersagemodell auf der Zeitreihe auf, die keine Trend- und Saisonalitätskomponente hat (stationäre Zeitreihe), und fügen dann diese Komponenten wieder hinzu, um Vorhersagen zu treffen. Deshalb ist es wichtig, Trend- und Saisonalitätskomponenten zu identifizieren und sie aus der Zeitreihe zu entfernen. Die Methode zur Entfernung dieser Komponenten wird in einem anderen Abschnitt diskutiert.

1.3 Grundlegende Kennzahlen von Zeitreihen

Das wichtigste Konzept, das die statistischen Merkmale von Zeitreihen beschreibt, wird in diesem Teil diskutiert. Im Prozess der Zeitreihenprognose werden uns diese Konzepte helfen, ein tieferes Verständnis der statistischen Eigenschaften der Zeitreihe zu haben und uns somit bei der Entscheidungsfindung unterstützen, um das Ziel zu erreichen, das beste Modell für die Prognose zu finden.

1.3.1 Autokovarianz

Wenn eine Erhöhung einer Variablen zu einer Erhöhung der anderen Variablen führt, spricht man von einer positiven Kovarianz bei der Variablen und umgekehrt. Kovarianz beschreibt, wie sich zwei Variablen unterscheiden.

Definition 1.4. *Sei $\{X_t\}$ eine Zeitreihe. Die Autokovarianzfunktion von $\{X_t\}$ in Lag h ist: (vgl. [4], Definition 1.4.3 auf S. 13)*

$$\gamma_X(h) = Cov(X_{t+h}, X_t).$$

Die Buchstaben t können als Vergangenheitsvariable und $t+h$ als Gegenwartsvariable verstanden werden. h kann 1,2,... sein, um eine Zeitverschiebung (Lag) darzustellen.

Definition 1.5. *Der Schätzer für die Autokovarianzfunktion mit Lag h für eine Stichprobegröße n ist wie folgt definiert: (vgl. [4], Definition 1.4.4 auf S. 16):*

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

Bemerkung: Durch die Anwendung des Schätzers der Autokovarianzfunktion verlieren wir h Elemente in der Stichprobe, wodurch die Schätzung nicht mehr zuverlässig ist. Eine Empfehlung ist, dass die Stichprobengröße n mindestens 50 betragen sollte und $h \leq \frac{n}{4}$ sein sollte. (vgl. [4], Bemerkung 1 auf S. 52)

Beispiel 1.4. Als Beispiel können wir auch IID-Zeitreihen $X_t \sim IID(0, \sigma^2)$ verwenden. Nun suchen wir die Autokovarianzfunktion $\gamma_X(t+h, t)$.

- Für $h = 0$: $\gamma_X(t,t) = E[(X_t - \mu)(X_t - \mu)] = E[(X_t - \mu)^2]$, welches ist Varianz und gleich σ^2
- Für $h \neq 0 : \gamma_X(t+h,t) = E[(X_t - \mu)(X_{t+h} - \mu)] = E[(X_t)(X_{t+h})] = 0$, (Dabei haben wir verwendet, dass $\mu = 0$ und X_t unabhangig für alle t sind

Wir sehen hier einen Autokovarianzwert, der unabhängig von der Änderung von t und h ist.

Beispiel 1.5. Im letzten Beispiel hatten wir eine Zeitreihe mit einer Autokovarianz, die unabhängig von t und h ist. Nun betrachten wir eine Zeitreihe mit einer Autokovarianz, die von t und h abhängt. Zuerst definieren wir einen Random-Walk-Prozess ohne Drift mit $\phi = 1$ und die Beobachtung $x_0 = 0$:

$$X_t = X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0,1). \tag{1.6}$$

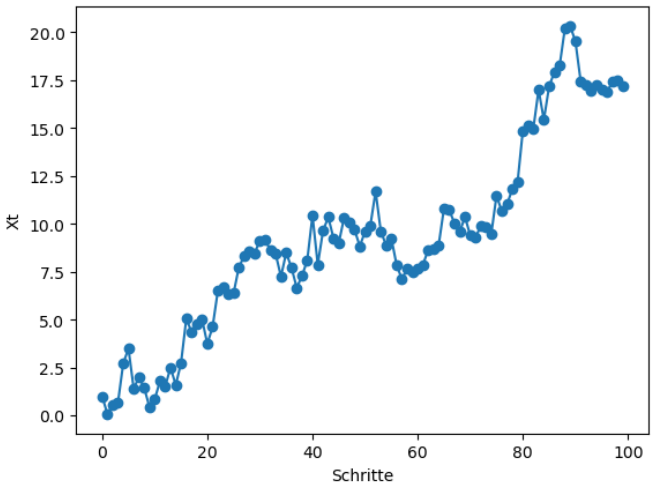

Abbildung 1.10: Random-Walk mit Drift = 1

Die Gleichung 1.6 hat den Erwartungswert von 0 und die Varianz von $t\sigma^2$. Für $h > 0$ gilt:

$$\gamma_X(t+h,t) = Cov(X_{t+h}, X_t) = Cov(X_t + X_{t+1} + \cdots + X_{t+h}, X_t) = Cov(X_t, X_t) = t\sigma^2.$$

Wir können sehen, dass die Kovarianz jetzt von t abhängt.

1.3.2 Autokorrelationfunktion

Die Charakteristik von Autokorrelation ist im Grunde genommen dieselbe wie bei Autokovarianz. Der Unterschied in der Definition besteht darin, dass die Autokorrelation sich selbst durch die Kovarianz mit Verzögerung 0, die die Varianz ist, teilen, um einen Wert im Bereich von -1 bis 1 zu erhalten. Das kann ein Faktor sein, der bei der Analyse von Prozessen hilft, da es intuitiver und verständlicher ist.

Definition 1.6. *Sei $\{X_t\}$ eine Zeitreihe. Die Autokorrelationfunktion von $\{X_t\}$ mit der Verzögerung h ist: (vgl. [4], Definition 1.4.3 auf S. 13)*

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = Cor(X_{t+h}, X_t).$$

Wir sehen, dass für $h = 0$ die Kovarianz $\gamma_X(0)$ die Varianz ist. Zum Beispiel für $h = 1$:

$$\rho_X(1) = \frac{\gamma_X(1)}{\sigma^2}.$$

Definition 1.7. *Der Schätzer für die Autokorrelationfunktion mit Lag h für eine Stichprobe ist wie folgt definiert: (vgl. [4], Definition 1.4.4 auf S. 16):*

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n$$

Beispiel 1.6. Hierfür betrachten wir die White-Noise-Zeitreihen $X_t \sim WN(0, \sigma^2)$. Für alle h ist die erwartete Autokorrelation also 0, was auf eine Nullkorrelation zwischen allen Daten in White-Noise-Zeitreihen hinweist.

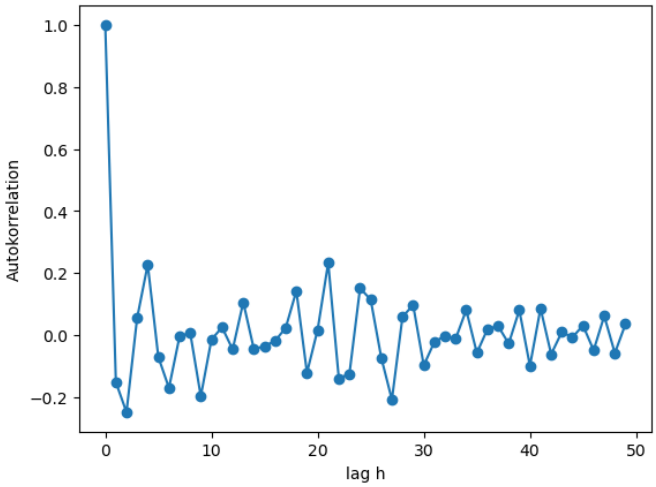

Abbildung 1.11: Autokorrelationsdiagramm mit einer Verzögerung (lag) von 0 bis 50 einer White-Noise-Zeitreihe.

1.3.3 Partial Autokorrelationfunktion

Der folgende Abschnitt basiert auf den Inhalten in [5] und in [18]

Die Partielle Autokorrelationsfunktion (PACF) misst wie die Autokorrelationsfunktion (ACF) die Korrelation zwischen Variablen mit Verzögerungen von k. Es gibt jedoch einen Unterschied zwischen den beiden Korrelationsfunktionen. Während die Autokorrelationsfunktion die Korrelation zwischen Variablen und ihren Verzögerungen k misst und die Wirkung anderer Verzögerungen $x_{t-k+1}, \cdots, x_t$ nicht berücksichtigt, berücksichtigt die Partielle Autokorrelationsfunktion die Wirkung anderer Verzögerungen $x_{t-k+1}, \cdots, x_t$ auf die Korrelation zwischen Variablen und ihren Verzögerungen k.
Im Folgenden wird ein Beispiel gezeigt, wie die partielle Autokorrelationsfunktion berechnet wird:

Wir möchten die PACF mit einer Verzögerung von 3 für eine Zeitreihe X_t berechnen. Da PACF auch die Auswirkungen anderer Verzögerungen von $t-k$ bis t berücksichtigt, regressieren wir die Variable X_t auf X_{t-1}, X_{t-2} und X_{t-3}. Die Auswirkung des Koeffizienten von X_{t-1} und X_{t-2} wird berücksichtigt, aber nur der Koeffizient von X_{t-3} wird als PACF verwendet:

$$X_t = \alpha + \beta_{t-1}X_{t-1} + \beta_{t-2}X_{t-2} + \underbrace{\beta_{t-3}}_{\text{lag 3}}X_{t-3}$$

Die Autokorrelationsfunktion (ACF) wird oft als Indikator für die Wahl der Ordnung des Moving-Averages-Modell verwendet, während die partielle Autokorrelationsfunktion (PACF) oft für das autoregressive Modell verwendet wird. Im nächsten Kapitel werden wir dies näher erläutern.

1.3.4 Stationarität und Korrelation

Dies ist ein wichtiger Begriff in der Zeitreihenprognose. In Abschnitt 1.2.2.3 wird bereits erwähnt, dass eine Zeitreihe frei von Trend- und Saisonkomponenten sein sollte, bevor darauf ein Modell angewendet wird. Eine Zeitreihe, die frei von Trend- und Saisonkomponenten ist, kann eine konstante statistische Eigenschaft aufweisen. Wir verwenden den Begriff 'stationär', um eine Zeitreihe zu beschreiben, die über die Zeit hinweg eine konstante statistische Eigenschaft aufweist.

Definition 1.8. *Eine Zeitreihe* $\{X_t, t = 0, \pm 1, ...\}$ *wird als stationär bezeichnet, wenn ihre statistischen Eigenschaften ähnlich sind wie die der "zeitverschobenen Reihe"* $\{X_{t+h}, t = 0, \pm 1, ...\}$ *für jede ganze Zahl h. (vgl. [4], Definition 1.4.2 auf S. 13)*

1.3.4.1 (Schwach) stationär

Wir unterscheiden auch zwischen den Definitionen von stationären Zeitreihen. Während für (streng) stationäre Zeitreihen die Wahrscheinlichkeitsdichtefunktion timeshift-invariant ist, gelten für schwach stationäre Zeitreihen nur der Mittelwert, die Varianz und die Kovarianz als timeshift-invariant.

Definition 1.9. *Eine Zeitreihe* X_t *ist (Schwach) stationär, wenn folgende Bedingungen erfüllt sind: (vgl. [4], Definition 1.4.2 auf S. 13)*

1. *Der Erwartungswert* $\mu_X(t)$ *ist unabhängig von t und damit konstant.*
2. *Die Autokovarianzfunktion* $\gamma_X(t + h, t)$ *ist unabhängig von t für alle h und damit konstant.*

Der zweite Punkt bedeutet, dass für eine Verzögerung (lag) h gleich 0 eine konstante Varianz über die Zeitreihe hinweg besteht und für eine Verzögerung (lag) ungleich 0 die Zeitreihe eine konstante Veränderung für alle t aufweist. Mit anderen Worten, für eine gegebene Verzögerung (lag) h hat die Zeitreihe dieselbe Veränderung, unabhängig von dem Zeitpunkt t, den wir beobachten.
In dieser Arbeit werden wir den Begriff stationär für die Definition von schwach stationär und stark stationär verwenden, da die Differenz keinen Einfluss auf das Ergebnis hat, ist auch der Nachweis für eine schwache Stationarität einfacher.
Ein Beispiel für eine nicht-stationäre Zeitreihe ist das Beispiel 1.5. Die Autokovarianz ist berechnet und das Ergebnis ist $t\sigma^2$. Dadurch ist $\gamma_X(t + h, h)$ nicht konstant.

1.3.4.2 Prüfung auf Stationarität

Dieser Abschnitt basiert auf [10], [17] und Abschnitt 6.3 von [4].

Eine Methode, die verwendet wird, um statistisch zu testen, ob eine Zeitreihe stationär ist oder nicht, ist der Dickey-Fuller-Test und seine erweiterte Version, der Augmented-Dickey-Fuller-Test. Der Dickey-Fuller-Test testet die Nullhypothese, dass in einem AR(1)-Prozess (Siehe AR(p)-Prozess Definition 2.1) eine Einheitswurzel vorhanden ist. Für diese Nullhypothese ist die Zeitreihe nicht stationär, während die alternative Hypothese besagt, dass die Zeitreihe stationär ist. Die erweiterte Version, der Augmented Dickey-Fuller-Test, erweitert den Test auf einen AR(p)-Prozess.

Zuerst wird das Konzept der Einheitswurzel in einem AR(p)-Prozess erklärt. Ein AR(p)-Prozess für eine Zeitreihe $\{X_t\}$ mit $x_0 = 0$ ist wie folgt definiert:

$$x_t = \phi_1 \cdot x_{t-1} + \phi_2 \cdot x_{t-2} + ... + \phi_p \cdot x_{t-p} + \varepsilon_t \tag{1.7}$$

Unter Verwendung des Lag-Operators (Siehe 2.7) kann die Gleichung 1.7 wie folgt geschrieben werden:

$$(1 - \phi_1 \cdot B - \phi_2 \cdot B^2 - ... - \phi_p \cdot B^p) \cdot x_t = \varepsilon_t \tag{1.8}$$

Wir ersetzen den Lag-Operator B durch eine Variable z, setzen das resultierende Polynom gleich 0 und erhalten die charakteristische Gleichung:

$$1 - \phi_1 \cdot z - \phi_2 \cdot z^2 - ... - \phi_p \cdot z^p = 0 \tag{1.9}$$

Die charakteristischen Wurzeln sind die Werte von z, wenn wir die Gleichung 1.9 lösen. Es gibt p-Lösungen von z. Wenn der Wert von z gleich 1 oder -1 ist, spricht man von einer Einheitswurzel.

Definition 1.10. *Ein AR(p)-Prozess ist nicht stationär, wenn seine charakteristische Gleichung mindestens eine Einheitswurzel aufweist oder wenn es Wurzeln gibt, deren Werte zwischen -1 und 1 liegen.*

Beispiel 1.7. Ein Beispiel hierfür ist ein AR(1)-Prozess. Die charakteristische Gleichung 1.9 für diesen Fall würde wie folgt geschrieben werden:

$$1 - \phi_1 \cdot z = 0 \Leftrightarrow z = \frac{1}{\phi_1}$$

Nach Definition 1.10 ist ein AR(1)-Prozess nicht stationär, wenn der Koeffizient ϕ_1 in einem der beiden Intervalle liegt: $(-\infty, -1]$ und $[1, \infty)$
In Beispiel 1.3 haben wir einen AR(1)-Prozess mit $\phi_1 = 0.7$, der als stationär gilt. In Beispiel 1.5 haben wir einen AR(1)-Prozess mit $\phi_1 = 1$, der nachgewiesenermaßen nicht stationär ist.

Jetzt kehren wir zum Dickey-Fuller-Test zurück. Zuerst nehmen wir die erste Differenz des AR(1)-Prozesses: $\Delta x_t = x_t - x_{t-1} = \delta \cdot x_{t-1} + \varepsilon_t$, wobei $\delta := \phi_1 - 1$. Es gibt drei Versionen des Dickey-Fuller-Tests:

1. Test auf eine Einheitswurzel: $\Delta x_t = \delta \cdot x_{t-1} + \varepsilon_t$.
2. Test auf eine Einheitswurzel mit Konstante: $\Delta x_t = \alpha_0 + \delta \cdot x_{t-1} + \varepsilon_t$.
3. Test auf eine Einheitswurzel mit Konstante und deterministischem Trend: $\Delta x_t = \alpha_0 + \alpha_1 \cdot t + \delta \cdot x_{t-1} + \varepsilon_t$.

Das Hypothesenpaar lautet:

- H_0: $\delta = 0$, also $\phi_1 = 1$
- H_1: $-2 < \delta < 0$, also $-1 < \phi_1 < 1$

Für jede Version wird δ mithilfe der Methode der kleinsten Quadrate (OLS) geschätzt. Der geschätzte δ-Wert wird als $\hat{\delta}$ bezeichnet. Die geschätzte Standardabweichung von $\hat{\delta}$ ist:

$$\hat{\sigma}_{\hat{\delta}} = S(\sum_{t=2}^{n}(x_{t-1} - \overline{x})^2)^{\frac{-1}{2}} \quad \text{mit} \quad S^2 = \sum_{t=1}^{n}(\Delta x_t + \overline{x} \cdot \hat{\delta} - \hat{\delta} \cdot x_{t-1})^2/(n-3)$$

Die Teststatistik wird wie folgt definiert::

$$T := \frac{\hat{\delta}}{\hat{\sigma}_{\hat{\delta}}}$$

Dann kann der p-Wert basierend auf der Dickey-Fuller-Tabelle berechnet werden (siehe Tabelle 8.5.2 in [4]).

Die erweiterte Version wird als Augmented Dickey-Fuller-Test bezeichnet und hat ebenfalls drei Versionen wie der Dickey-Fuller-Test. Sie erweitert den Test von AR(1) auf AR(p). Das δ zur Überprüfung in der Hypothese wird nun als $\delta = \sum_{i=1}^{p} \phi_i - 1$ definiert. Die Nullhypothese lautet ebenfalls $\delta = 0$. Die Methode zur Berechnung des p-Werts ist analog zur Version des Dickey-Fuller-Tests.
Wir können den Augmented Dickey-Fuller-Test einfach mithilfe der Funktion `adfuller()` aus der Bibliothek `statsmodels` anwenden. Die Version des Tests kann über den Parameter `regression` festgelegt werden, wobei die Anzahl der Lags für die Regression standardmäßig durch das AIC-Kriterium (Siehe Abschnitt 2.2.5) gewählt wird. Das AIC wählt den Lag mit dem niedrigsten AIC-Wert aus, um die Korrelationen in den Residuen zu minimieren.

1.3.4.3 Prüfung auf Korrelation

Eine stationäre Zeitreihe, wie bereits erwähnt, hat einen konstanten Mittelwert, eine konstante Varianz und eine konstante Autokovarianz. Die Korrelation zwischen den Beobachtungen kann jedoch weiterhin bestehen. Diese Informationen sind hilfreich, um uns bei der Modellierung zu unterstützen und diese Struktur einzufangen. Nachdem ein ausgewähltes Modell angepasst wurde, sollten die Residuen idealerweise einem

White-Noise-Prozess entsprechen, was bedeutet, dass keine Korrelation zwischen den Beobachtungen vorhanden sein sollte. Dies deutet darauf hin, dass das Modell gut mit den Daten funktioniert hat und nur noch unberechenbares Rauschen übrig bleibt. Um statistisch zu überprüfen, ob eine Zeitreihe Autokorrelation zwischen den Beobachtungen aufweist, können wir den Ljung-Box-Test verwenden. [9]
Der Ljung-Box-Test hat folgendes Hypothesenpaar:

- H_0: Die Daten sind unabhängig verteilt.
- H_1: Die Daten sind nicht unabhängig verteilt; sie zeigen eine Serienkorrelation.

Die Teststatistik ist wie folgt definiert:

$$Q = n \cdot (n+2) \cdot \sum_{k=1}^{h} \frac{\hat{\rho}^2(h)}{n-k}$$

Hierbei steht n für die Anzahl der Beobachtungen und $\rho(h)$ für die Autokorrelationskoeffizienten bei der Verzögerung h. Unter H_0 folgt die Teststatistik Q einer Chi-Quadrat-Verteilung mit h Freiheitsgraden.
Die Verwendung des Ljung-Box-Tests kann mithilfe der Funktion `acorr_ljungbox` in der Bibliothek `statsmodels` erfolgen. Wir setzen die zu testenden Verzögerungen hier über den Parameter lags. Es wird empfohlen [8], für nicht saisonale Zeitreihen eine Verzögerung von $h = 10$ zu verwenden und für saisonale Zeitreihen eine Verzögerung von $2m$ (m ist die Periode).

Kapitel 2

Modellierung von Zeitreihendaten

Dieses Kapitel wird das Modell beschreiben, das mit der Box-Jenkins-Methodik auf die in Abschnitt 2.1 vorgestellten Daten angewendet wird.

2.1 Daten

In dieser Arbeit werden Daten verwendet, die vom Climate Data Center (CDC) des Deutschen Wetterdienstes (DWD) stammen. Der DWD verfügt über ein Netzwerk von Wetterstationen in ganz Deutschland. Die Daten umfassen die mittleren monatlichen Lufttemperaturen aller Bundesländer von 1881-2022 in Deutschland. Die Daten für einzelne Bundesländer stammen aus den durchschnittlichen Werten aller Messstationen innerhalb der Region. (Für weitere Informationen können Sie sich an Axel.Kuschnerow@dwd.de wenden.)

Wir werden uns ausschließlich auf die monatliche Lufttemperatur von Baden-Württemberg beschränken, was zu einer Zeitreihe mit 1704 Beobachtungen führt.

2.2 Modell

Das autoregressive (AR)-Modell, das gleitende Durchschnitts (MA)-Modell und die kombinierte Version ARMA wurden von den Statistikern G.U. Yule (1927 [20]) und E. Slutsky (1927 [14]) vorgeschlagen. Sie beobachteten, dass eine zufällige Reihe, die durch diese beiden Prozesse beschrieben wird, eine Zeitreihe erzeugen kann, die zyklische Eigenschaften aufweist und als Beschreibung für andere Zeitreihen in der realen Welt betrachtet wird. Aufbauend auf dieser Arbeit lieferten Box und Jenkins (1970 [3]) Vorschläge für die Analyse, Prognose und Kontrolle von Zeitreihen basierend auf ARMA-Modellen.
Für die in Abschnitt 2.1 vorgestellten Daten wird das SARIMA-Modell angewendet. Das SARIMA-Modell wird auf Daten angewendet, die ein periodisches Merkmal aufweisen. Um den Begriff des SARIMA-Modells zu erläutern, werden im Folgenden die Begriffe erläutert, die benötigt werden, um das größere Konzept des SARIMA-Modells zu erklären.

2.2.1 Autoregressives Modell (AR-Modell)

Eine Zeitreihe $\{X_t\}$ kann durch ein autoregressives Modell modelliert werden, falls die Abhängigkeit der Zeitreihe von einer Zufallsvariable zum Zeitpunkt t und von anderen Zufallsvariablen $X_{t-1}, X_{t-2}, \cdots, X_{t-p}$, wobei p der Parameter des autoregressiven Modells AR(p) ist, beschrieben werden kann. Wir haben bereits ein Beispiel für ein AR(1)-Modell, welches ein Random-Walk ist. Die Notation AR(p) steht für ein autoregressives Modell der Ordnung p. Wir definieren: (vgl. [7], Abschnitt 4.1 auf S. 102)

Definition 2.1. *Sei $\{X_t\}$ eine Zeitreihe. $\{X_t\}$ ist ein AR(p)-Prozess für alle t, wenn für die Zufallsvariable*

X_t gilt:

$$X_t = \sum_{i=1}^{p} \phi_i \cdot X_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2). \tag{2.1}$$

Wir nehmen an, dass die Gleichung 2.1 einen Erwartungswert von 0 hat. Dann ist X_t ein AR(p)-Prozess mit einem Erwartungswert von μ, wenn die Zeitreihe $\{X_t - \mu\}$ die Gleichung 2.1 erfüllt.

Ein kritischer Schritt bei der Anwendung des AR-Modells ist die Auswahl des Parameters p. Die Verwendung von PACF zur Auswahl der Ordnung p für AR(p) ist geeignet, da PACF den Einfluss der Zwischenverzögerungen berücksichtigt.

2.2.2 Moving-Average-Modell (MA-Modell)

Während im AR-Modell die aktuelle Zufallsvariable X_t auf der Beziehung zu ihren vorherigen Werten bis zum Lag k und einem aktuellen Fehlerterm Z_t basiert, der als zufälliger Prozess betrachtet wird, basiert die aktuelle Zufallsvariable X_t im Moving-Average- (MA-) Modell auf einem Wert μ, um den sich der Prozess bewegt, einem aktuellen Fehlerterm und den vorherigen Fehlern bis zum Lag k. Die Notation MA(q) steht für ein Moving-Averages-Modell der Ordnung q. Wir definieren: (vgl. [7], Abschnitt 4.2 auf S. 104)

Definition 2.2. *Sei $\{X_t\}$ eine Zeitreihe. $\{X_t\}$ ist ein MA(q)-Prozess für alle t wenn für die Zufallsvariable X_t gilt:*

$$X_t = \mu + Z_t + \sum_{i=1}^{q} \theta_i \cdot Z_{t-i}, \quad Z_t \sim WN(0, \sigma^2). \tag{2.2}$$

Die Gleichung 2.2 hat den Erwartungswert von μ.
Da die statistische Charakteristik dieser Fehler zufällig und unkorreliert ist, ist die Verwendung des ACF zur Auswahl der Ordnung p für MA(p) geeignet, da das ACF nur die Beziehung zwischen dem aktuellen Term und dessen Verzögerung berücksichtigt, ohne den Effekt der Zwischenverzögerungen zu kontrollieren.

2.2.3 ARMA Modell/ARIMA Modell

ARMA-Modell: Das ARMA-Modell ist eine Kombination aus AR(p)- und MA(q)-Modellen, mit der eine Zeitreihe flexibler modelliert werden kann. Die Notation für das ARMA-Modell lautet ARMA(p,q), wobei p der Parameter des AR-Modells und q der Parameter des MA-Modells entspricht. Eine wichtige Voraussetzung für die Anwendung des ARMA(p,q)-Modells auf eine Zeitreihe ist, dass die Zeitreihe zuerst stationär sein muss. Wir definieren: (vgl. [4], Definition 3.1.1 auf S. 74)

Definition 2.3. *Sei $\{X_t\}$ eine stationäre Zeitreihe. $\{X_t\}$ ist ein ARMA(p,q)-Prozess für alle t, wenn es gilt für die Zufallsvariable X_t:*

$$X_t = \sum_{i=1}^{p} \phi_i \cdot X_{t-i} + \sum_{j=1}^{q} \theta_i \cdot Z_{t-j} + Z_t, \quad Z_t \sim WN(0, \sigma^2) \tag{2.3}$$

Wir nehmen an, dass die Gleichung 2.3 einen Erwartungswert von 0 hat. Dann ist X_t ein ARMA(p,q)-Prozess mit einem Erwartungswert von μ, wenn die Zeitreihe $\{X_t - \mu\}$ die Gleichung 2.3 erfüllt.

ARIMA-Modell: Für die Modellierung einer Zeitreihe mit einem ARMA(p,q)-Prozess muss die Zeitreihe zuerst stationär sein. ARIMA(p,d,q) wird vorgeschlagen, um diese Voraussetzung zu erfüllen. Die Notation ARIMA steht für Autoregressive Integrated Moving Average, und der Parameter d entspricht der Anzahl der erforderlichen Differenzen, um die Zeitreihe in einen stationären Zustand zu versetzen.

Um den Prozess knapper definieren zu können, definieren wir zunächst den Backshift-Operator. Die Gleichung 2.3 kann wie folgt durch den Backshift-Operator definiert werden: (vgl. [4], Abschnitt 3.1 auf S. 74):

$$\Phi(B)X_t = \Theta(B)Z_t. \tag{2.4}$$

Wobei $\Phi(.)$ und $\Theta(.)$ die Polynomgrade p und q darstellen:

$$\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p. \tag{2.5}$$

und

$$\Theta(z) = 1 - \theta_1 z - \cdots - \theta_p z^p. \tag{2.6}$$

B entspricht dem Lag-Operator:

$$B^j X_t = X_{t-j}. \tag{2.7}$$

und

$$B^j Z_t = Z_{t-j}.$$

Definition 2.4. *Sei $\{X_t\}$ eine stationäre Zeitreihe, $\{X_t\}$ ist ein ARIMA(p,d,q)-Prozess wenn $(1-B)^d X_t$ ein ARMA(p,q)-Prozess ist. (vgl. [4], Definition 6.1 auf S. 158).*

2.2.4 SARIMA Modell

Das SARIMA-Modell ist eine Version des ARIMA-Modells, das einen saisonalen Term enthält. Das Modell wird mit der Notation SARIMA(p, d, q)(P, D, Q)m angegeben. Wir verwenden den Lag-Operator, um das SARIMA-Modell wie folgt zu definieren:

Definition 2.5. *Sei $\{X_t\}$ eine stationäre Zeitreihe, $\{X_t\}$ ist ein SARIMA(p,d,q)(P,D,Q)m-Prozess, wenn $Y_t = (1-B)^d(1-B^s)^D X_t$ ein ARMA-Prozess ist, der wie folgt definiert: (vgl. [4], Definition 6.5.1 auf S. 177).*

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t, \quad Z_t \sim WN(0, \sigma^2).$$

Beispiel 2.1. Wir betrachten das Beispiel SARIMA(1,0,0)(0,1,1)4. Unter Verwendung des Lag-Operators kann das Modell wie folgt geschrieben werden:

$$(1 - \phi_1 \cdot B)(1 - B^4)X_t = (1 + \Theta_1 \cdot B^4)Z_t$$

Im Vergleich zum ARIMA-Modell ist SARIMA in Bezug auf die Anzahl der Parameter komplexer. Diese Erweiterung ist hilfreich, wenn das Ziel darin besteht, eine Zeitreihe zu modellieren, die saisonale Komponenten aufweist.

2.2.5 AIC-Kriterium

Ein Modell mit mehr Parametern passt besser zu den Trainingsdaten. Es kann jedoch auch zu einem Überanpassungsproblem führen und eine schlechte Prognose für Daten liefern, die es nicht gesehen hat. Irgendwann lohnt es sich nicht mehr, weitere Parameter hinzuzufügen, da die Vorteile eines besseren Anpassens geringer werden. Es gibt viele Kriterien, die nicht nur berücksichtigen, wie gut das Modell zu den Daten passt, sondern auch einen Strafterm für die Anzahl der Parameter enthalten. In dieser Arbeit werden wir das AIC-Kriterium [1] verwenden. Sei $y = x_1, x_2, x_3, ...$ eine Menge, die die Realisierungen von $\{X_t\}$ darstellt. $L(\hat{\theta}, y)$ ist die Likelihood-Funktion des Modells mit dem Maximum-Likelihood-Schätzer $\hat{\theta}$ für den

Parametervektor θ. k ist die Anzahl der Parameter im Vektor θ. Der Statistiker H. Akaike definiert das Kriterium wie folgt:

$$AIC = -2L(\hat{\theta}, y) + 2 \cdot p \tag{2.8}$$

Die Gleichung 2.8 leitet sich aus dem Schätzer der Kullback-Leibler-Divergenz ab:

$$E[K(\widehat{f_{\theta_k}}, f_0)] = -L(\hat{\theta}, y) + p + \int \log f_0(y) f_0(y) dy. \tag{2.9}$$

Die Kullback-Leibler-Divergenz wird wie folgt definiert:

$$K(f_\theta, f_0) = \int [\log(f_0(y)) - \log(f_\theta(y))] \cdot f_0(y) dy. \tag{2.10}$$

Die Kullback-Leibler-Divergenz misst, wie gut das Modell f_θ (das Modell, das θ als Parameter verwendet) zum Modell f_0 (das das wahre Modell darstellt, das die Daten y generiert) passt. p ist die Anzahl der Parameter in θ. Wenn $\hat{\theta}$ der Maximum-Likelihood-Schätzer (MLE) von θ ist, kann $K(f_\theta, f_0)$ verwendet werden, um zu sehen, wie gut das Modell θ zu den Daten passt. Da f_0 nicht sicher bekannt ist, wird θ_k verwendet (θ_k sind die unbekannten Parameter), um die Gleichung 2.10 zu minimieren (Siehe 7. Anhang [19]).
Da das Ziel darin besteht, das Kriterium zur Vergleichbarkeit zwischen Modellen zu verwenden, kann der Term $\int \log f_0(y) f_0(y) dy$ in Gleichung 2.9 entfernt werden, da er nur das reale Modell enthält und für alle zu vergleichenden Modelle gleich ist. Dies führt zur Gleichung 2.8. Ein niedrigeres Ergebnis der Gleichung 2.9 würde bedeuten, dass das Modell f_θ besser zum realen Modell passt. Dies deutet darauf hin, dass ein niedrigerer AIC uns ein besseres Modell liefert.

Kapitel 3

Anwendung des SARIMA-Modells auf monatliche Temperaturdaten

Im folgenden Abschnitt wird eine Vorhersage basierend auf den genannten Daten mithilfe des SARIMA-Modells abgeleitet. Gemäß Box-Jenkins (1970) kann eine Vorhersage in vier Schritten aus einem Modell abgeleitet werden: [2]

- (1) Modellidentifikation: Der Standardansatz besteht darin, mithilfe der ACF und PACF nach einem Modell zu suchen, das zur Selbstkorrelation innerhalb der Daten passt (beispielsweise aufgrund von Trend- oder Saisonalitätskomponenten).
- (2) Schätzung der Modellparameter: Nachdem das Modell identifiziert wurde, werden die Modellparameter geschätzt.
- (3) Diagnoseprüfung: Hier wird die Analyse der Residual-Zeitreihe durchgeführt.
- (4) Anwendung des Modells zur Prognose: Schließlich wird das validierte Modell verwendet, um Vorhersagen für zukünftige Zeitpunkte zu machen.

3.1 Datenvisualisierung und -vorbereitung

Die Daten werden zuerst von der Website `https://opendata.dwd.de/climate_environment/CDC/regional_averages_DE/monthly/air_temperature_mean/` mithilfe von Python heruntergeladen. Anschließend wird ein Dataframe mit nur einer Spalte erstellt, die die mittlere Lufttemperatur von Baden-Württemberg enthält. Aufgrund der Länge der Beobachtungsreihe werden die Daten in zwei Teile aufgeteilt und visualisiert.

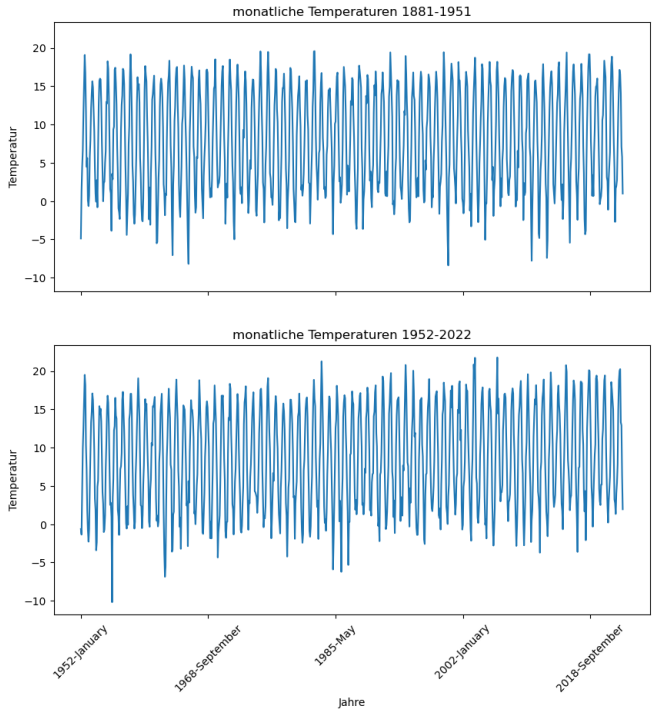

Abbildung 3.1: Die mittleren monatlichen Lufttemperaturen von Baden-Württemberg von 1881 bis 2022.

Hier ist eine klare Periodizität erkennbar, die der jährlichen Veränderung der Jahreszeiten in einem 12-Monats-Zyklus entspricht.

3.2 Datenverschiebung und Logarithmierung

Der Logarithmus wird auf die Daten angewendet, um die Zeitreihe stabiler zu machen, so dass ein Modell leichter auf der Zeitreihe aufgebaut werden kann. Da die Daten jedoch negative Werte aufweisen, wird zunächst ein Schritt der Datenverschiebung durchgeführt.
Da der niedrigste Wert in der Abbildung 3.1 einen Wert von -10,19 aufweist, werden alle Werte in der Zeitreihe um 15 addiert. Dann werden die Daten logarithmiert. Der Prozess wird mit Python durchgeführt und liefert uns die folgende Abbildung:

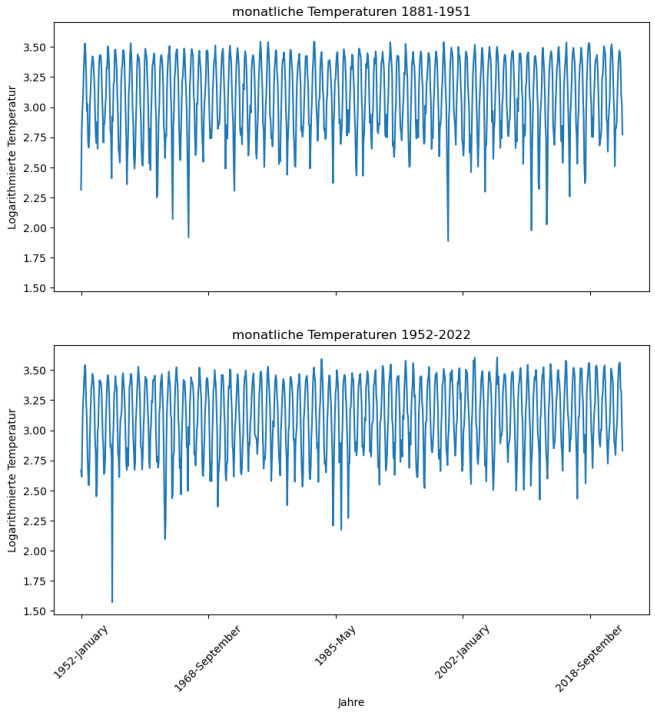

Abbildung 3.2: verschobene und logarithmierte Daten

Ab jetzt werden weitere Analysen auf den verschobenen und logarithmierten Daten durchgeführt.

3.3 Trend- und Periodizitätskomponenten entfernen

Wie bereits in Abschnitt 1.2.2.3 erwähnt wurde, müssen gegebenenfalls Trend- und Periodizitätskomponenten aus der Zeitreihe entfernt werden (stationär werden), um sie in weiteren Modellen modellieren zu können. Dieser Schritt ist wichtig, da er uns eine stationäre Zeitreihe liefert, auf der wir weitere Modelle anwenden können. Es gibt verschiedene Ansätze für diesen Prozess, und für das SARIMA-Modell werden Differenzierungsmethoden angewendet.
Die Funktion `statsmodels.tsa.stattools.adfuller()` wird jetzt angewendet, um die Daten aus Abschnitt 3.2 auf das Vorhandensein von Einheitswurzel zu überprüfen. Hier wird der Parameter "regressionïn der adfuller()-Funktion auf 'n' gesetzt, was eine Regression ohne Intercept bedeutet. Das liefert uns folgende p-Werte und Teststatistik:

```
Test statistic = 0.2388
p-value = 0.7579
```

In dem Fall, dass das Signifikanzniveau 0.05 beträgt, ergibt das Ergebnis eine sehr hohe Wahrscheinlichkeit dafür, dass die Zeitreihe nicht stationär ist. Durch Visualisierung kann die Trend- und Periodizitätskomponente noch sichtbar gemacht werden.

3.3.1 Die Komponenten visualisieren

Eine Zeitreihe kann durch drei Komponenten definiert werden, wie in Gleichung 1.5 beschrieben. Mithilfe der Funktion `season_decompose()` aus der Bibliothek `statsmodels` in Python können die Daten aus Abbildung 3.2 in die drei Komponenten aufgeteilt werden:

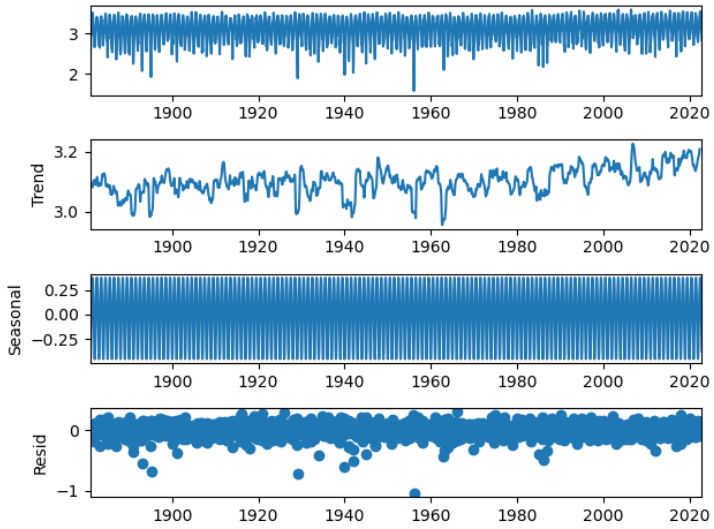

Abbildung 3.3: Die Zeitreihe in drei Teile aufgeteilt.

Laut der Dokumentation von Statsmodel [15] handelt es sich um einen naiven Decomposition-Ansatz, der uns jedoch einen guten Überblick über die Daten liefert. Um die Trendkomponente zu berechnen, wird entlang der Zeitreihe eine gleitende Durchschnittsbildung mit einem Zeitfenster, das der Periodizität der Zeitreihe entspricht (in diesem Fall 12), durchgeführt (siehe Abschnitt 1.5.1, Method 1a, S. 21 [4]). Diese Methode glättet die Zeitreihe und hilft, den zugrunde liegenden Trend zu erkennen. Nachdem die Trendkomponente berechnet wurde, wird sie von der Zeitreihe abgezogen, um die Periodizitätskomponente zu berechnen. Die Periodizitätskomponente ergibt sich hier aus den durchschnittlichen Werten jeder Periode (in diesem Fall 12) der trendbereinigten Zeitreihe. Der Residual wird erhalten, indem man die Trend- und Periodizitätskomponente von der Zeitreihe abzieht. Dieser Ansatz ist hilfreich, um einen Überblick über eine Zeitreihe zu erhalten, aber für eine detaillierte Modellierung ist er aufgrund seiner Einfachheit nicht geeignet. Die Saisonalität ist hier gut mit einer Periodizität von 12 erkennbar, und der Trend bewegt sich im Bereich von 1881 bis 1962 innerhalb einer Range. Ab 1962 ist ein Aufwärtstrend erkennbar. Somit ist die Zeitreihe aus visueller Perspektive sicherlich nicht stationär.

3.3.2 Differenzierungsmethode

Um die Trendkomponente einer Zeitreihe zu entfernen, kann man dies auch durch Erstdifferenzierung erreichen. Wir nehmen die Erstdifferenzierung für Gleichung 1.2 als Beispiel (ein Random-Walk mit Drift):

$$\Delta X_t = X_t - X_{t-1} = (\alpha \cdot t + \sum_{t=1}^{t} \varepsilon_t) - (\alpha \cdot (t-1) + \sum_{t=1}^{t-1} \varepsilon_t) = \alpha + \varepsilon_t$$

Wir können sehen, dass $E[\Delta X] = \alpha$ und $Var[\Delta X] = \alpha + \sigma^2$ für alle t ist. Daraus folgt, dass uns die Erstdifferenzierung eine stationäre Zeitreihe liefert.
Falls eine Zeitreihe eine Periodizitätskomponente besitzt, kann sie auch durch Differenzierung entfernt werden. Für die monatlichen Temperaturendaten kann eine 12-Differenzierungsmethode genutzt werden, um die Saisonalitätskomponente zu entfernen. Es ist auch möglich, dass mehrere Differenzierungen erforderlich sind oder eine Kombination aus Trend- und Periodizitätsdifferenzierung. Ein allgemeiner Prozess wird im nächsten Kapitel erklärt. Jetzt werden die monatlichen Temperaturen durch verschiedene Differenzierungen untersucht, indem die differenzierten Daten aus dem Abschnitt 3.2 sowie ihre ACF und PACF geplottet werden. Dazu wird auch ein ADF-Test durchgeführt, um zu testen, ob die differenzierte Zeitreihe stationär ist. Zu bemerken ist hier, dass das berechnete Konfidenzintervall von ACF und PACF normalerweise im Bereich von -0.2 bis 0.2 liegt. Aufgrund der großen Menge an Daten wird das Konfidenzintervall jedoch kleiner. Die Bestimmung des Konfidenzintervalls wird automatisch von der Funktion `plot_pacf` und `plot_acf` berechnet.

- **Erstdifferenzierung**:

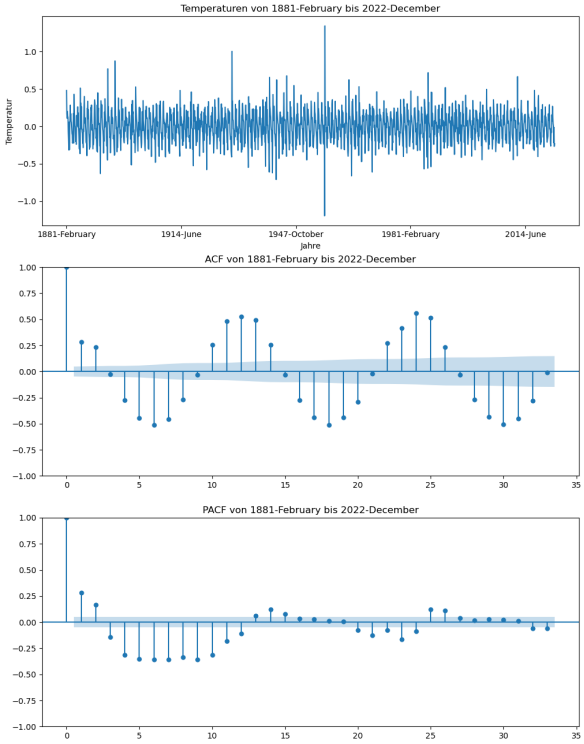

Abbildung 3.4: Erstdifferenzdaten und ihre ACF und PACF.

Die ADF-Test liefert den p-Wert und die Teststatistik:

```
Test statistic = -14.7368
p-value = 0.0000
```

Durch die Erstdifferenzierung wurde die Trendkomponente bereits eliminiert. Es besteht jedoch immer noch eine starke Periodizitätskomponente, wie aus dem ACF-Plot ersichtlich. Die Zeitreihe nach der Erstdifferenzierung stellt bereits eine stationäre Zeitreihe dar. Es ist jedoch nicht optimal, weitere Analysen auf den erstedifferenzierten Daten durchzuführen, da noch zu viel Struktur der Originalzeitreihe in den differenzierten Zeitreihen vorhanden ist.

- **Periodendifferenz**:

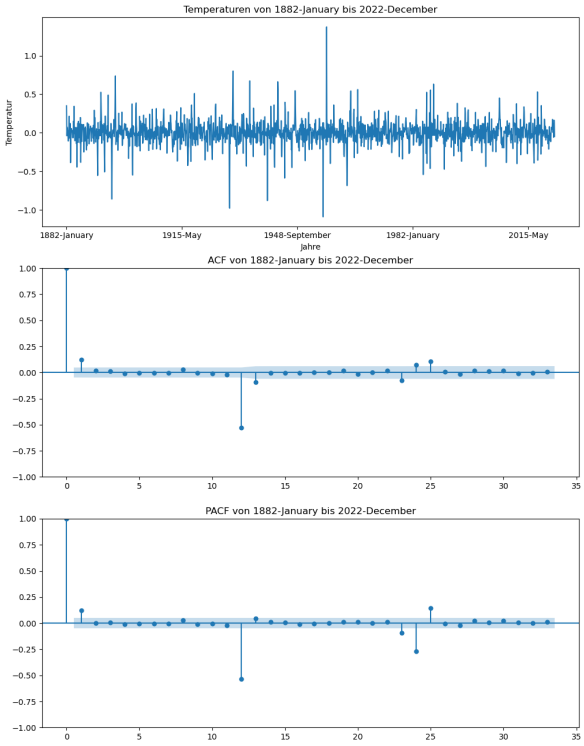

Abbildung 3.5: Periodisch differenzierte Daten und ihre ACF und PACF.

Die ADF-Test liefert den p-Wert und die Teststatistik:

```
Test statistic = -11.5748
p-value = 0.0000
```

Durch die Periodendifferenzierung wird die Zeitreihe stationär. Im Vergleich zur Erstdifferenz wurden viele Strukturen der Originalzeitreihe bereits entfernt. In diesem Fall ist es sinnvoll, ein Modell darauf anzuwenden, da noch einige Strukturen vorhanden sind. Wir sehen ein Spike im ersten Lag über dem Signifikanzniveau sowohl im ACF als auch im PACF, aber die folgenden Lags sind nicht signifikant. Wir können zunächst einen Parameter $q = 1$ für das MA(q)-Modell im ARIMA-Teil auswählen. Es ist auch zu beachten, dass jede Periode von 12 eine signifikante Lag in beiden ACF und PACF aufweist, was auf Parameter für den AR- und MA-Teil im saisonalen Modell hinweist. Es ist jedoch schwierig zu bestimmen, ob es sich um einen AR- oder MA-Parameter handelt. Wir werden noch verschiedene Modelle ausprobieren und das AIC-, BIC-Kriterium vergleichen, um herauszufinden, welches Modell am besten geeignet ist.

- **Erstdifferenz der saisonalen Differenz**

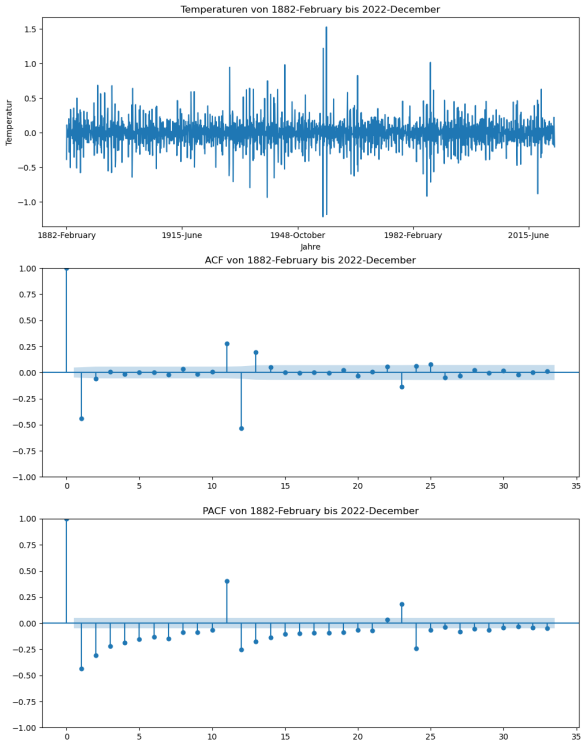

Abbildung 3.6: Erste Differenz der saisonalen Differenz Daten und ihre ACF und PACF.

Die ADF-Test liefert den p-Wert und die Teststatistik:

```
Test statistic = -14.9729
p-value = 0.0000
```

Die erste Differenz der saisonalen Differenz liefert ebenfalls ein sinnvolles Residuum, das stationär ist und noch modellierbare Strukturen aufweist. Der Spike in der ersten Lag in der ACF deutet darauf hin, dass möglicherweise ein Parameter q = 1 für den MA(q)-Teil erforderlich ist. Die allmählich abnehmende Korrelation in der PACF ist in diesem Fall schwer zu beurteilen, daher können wir verschiedene p-Parameter ausprobieren. Der Spike um die Periodenlag 12 zeigt, dass eine saisonale Struktur erfasst werden sollte, und daher werden wir auch verschiedene Parameter P und Q ausprobieren. Im Folgenden werden verschiedene SARIMA-Modelle untersucht, die auf den Daten nach der Periodendifferenz und der Erstdifferenz der saisonalen Differenz basieren.

3.4 Modellidentifikation

Verschiedene Methoden können verwendet werden, um ein geeignetes Modell auszuwählen, und jede Methode kann unterschiedliche 'beste Modelle' liefern. Zunächst werden Grenzen für die Parameter anhand der ACF und PACF festgelegt. Anschließend werden in dieser Arbeit die beiden Methoden Brute-Force (Auswahl nach AIC-Kriterium) und Backward-Elimination angewendet, um das beste Modell zu ermitteln.
Für die Periodendifferenz setzen wir folgende Grenzen für die Parameter:
`SARIMA(2,0,2)(3,1,3)12`.
Und für die Erstdifferenzierung der Saisonalitätsdifferenzierung setzen wir folgende Grenzen:
`SARIMA(3,1,3)(3,1,3)12`.

3.4.1 Trainings- und Testdaten

Von nun an werden wir die Zeitreihe in einen Trainingsdatensatz und einen Testdatensatz aufteilen. Die Modelle werden anhand des Trainingsdatensatzes ausgewählt und anschließend auf dem Testdatensatz getestet, um zu sehen, welches Modell geeignet ist. Der Trainingsdatensatz wird die ersten 1363 Beobachtungen und der Testdatensatz wird die letzten 341 Beobachtungen enthalten, was ein Verhältnis von 80:20 anzeigt.

3.4.2 Backward-Elimination

Die Anwendung von Backward-Elimination ist eine Form der schrittweisen Regression, die häufig in der multiplen Regression verwendet wird, um Koeffizienten auszuwählen. Das Ziel ist es, nicht signifikante Koeffizienten zu eliminieren. Dies geschieht durch einen t-Test mit der Nullhypothese H_0: Koeffizient = 0 für jeden Koeffizienten. Dieser Ansatz dient dazu, das Modell zu vereinfachen und Overfitting zu vermeiden. Bei der Berechnung des p-Werts verwenden wir die Teststatistik: $t = \frac{(\beta_i - 0)}{\sigma_{\beta_i}}$. Dabei ist der Freiheitsgrad gleich der Stichprobengröße minus der Anzahl der Koeffizienten. Für das Signifikanzniveau wählen wir üblicherweise 0.05. Obwohl die Anwendung von Backward-Elimination in der Box-Jenkins-Methodik nicht üblich ist, können wir das SARIMA-Modell dennoch als eine Art von multipler Regression betrachten und die Backward-Elimination anwenden. (Weitere Informationen zu dieser Anwendung [11].)

Für `SARIMA(2,0,2)(3,1,3)12` erhalten wir:

$$X_t - X_{t-12} = W_t \tag{3.1}$$

$$\begin{aligned} W_t = {} & w + \phi_1 W_{t-1} + \phi_2 W_{t-2} + \phi_3 W_{t-12} + \phi_4 W_{t-24} \\ & + \phi_5 W_{t-36} + \theta_1 \varepsilon_{w_{t-1}} + \theta_2 \varepsilon_{w_{t-2}} \\ & + \theta_3 \varepsilon_{w_{t-12}} + \theta_4 \varepsilon_{w_{t-24}} + \theta_5 \varepsilon_{w_{t-36}} + \varepsilon_{w_t}. \end{aligned} \tag{3.2}$$

Mit Hilfe der Bibliothek SARIMAX von statsmodels werden die Koeffizienten geschätzt und die p-Werte berechnet. Dabei werden unsignifikante Koeffizienten eliminiert. Dieser Prozess wird fortgesetzt, bis alle Koeffizienten signifikant sind. Im Folgenden haben wir eine Tabelle, die den Prozess darstellt. Die fettgedruckte Linie kennzeichnet den zu eliminierenden Parameter.

Tabelle 3.1: Backward-Elimination für SARIMA(2,0,2)(3,1,3)12

1) SARIMA(2,0,2)(3,1,3)12				2) SARIMA(1,0,2)(3,1,3)12			
Parameter	Schätzung	StdErr	p-Wert	Parameter	Schätzung	StdErr	p-Wert
AR(1)	0.3503	19.426	0.98561	**AR(1)**	**-0.142220**	**1.010852**	**0.88811**
AR(2)	**-0.041890**	**4.26**	**0.99217**	MA(1)	0.323325	1.009808	0.74883
MA(1)	-0.183211	19.427	0.99248	MA(2)	0.061447	0.176090	0.72712
MA(2)	0.020507	1.041312	0.98429	SAR(1)	-1.288512	0.137238	0.00000
SAR(1)	-1.207482	0.171994	0.00000	SAR(2)	-0.450098	0.146599	0.00214
SAR(2)	-0.444189	0.180167	0.01368	SAR(3)	0.079464	0.020466	0.00010
SAR(3)	0.067207	0.022993	0.00347	SMA(1)	0.283417	0.145203	0.05095
SMA(1)	0.212768	0.174620	0.22305	SMA(2)	-0.756182	0.099771	0.00000
SMA(2)	-0.684611	0.152914	0.00001	SMA(3)	-0.517050	0.140424	0.00023
SMA(3)	-0.512155	0.171094	0.00276				

3) SARIMA(0,0,2)(3,1,3)12				4) SARIMA(0,0,1)(3,1,3)12			
Parameter	Schätzung	StdErr	p-Wert	Parameter	Schätzung	StdErr	p-Wert
MA(1)	0.160392	0.018734	0.00000	MA(1)	0.157363	0.018279	0.00000
MA(2)	**0.040258**	**0.027904**	**0.14910**	SAR(1)	-1,780351	0.047233	0.00000
SAR(1)	-1.556658	0.100026	0.00000	SAR(2)	-0.878240	0.058190	0.00000
SAR(2)	-0.658956	0.109696	0.00000	SAR(3)	0.028735	0.020578	0.16259
SAR(3)	0.069623	0.019721	0.00041	SMA(1)	0.785023	2.446928	0.74835
SMA(1)	0.573085	0.297064	0.05371	**SMA(2)**	**-0.884180**	**4.369805**	**0.83965**
SMA(2)	-0.839754	0.425905	0.04864	SMA(3)	0.900639	0.207475	0.68328
SMA(3)	-0.731594	0.224678	0.00113				

5) SARIMA(0,0,1)(3,1,[1,0,1])12				6) SARIMA(0,0,1)(2,1,[1,0,1])12			
Parameter	Schätzung	StdErr	p-Wert	Parameter	Schätzung	StdErr	p-Wert
MA(1)	0.163704	0.019353	0.00000	MA(1)	0.154046	0.019191	0.00000
SAR(1)	-0.458134	0.165589	0.00566	SAR(1)	-0.340948	0.166994	0.04118
SAR(2)	-0.374907	0.167921	0.02557	**SAR(2)**	**-0.263680**	**0.168024**	**0.11658**
SAR(3)	**0.011024**	**0.024906**	**0.65804**	SMA(1)	-0.647455	1.171523	0.00016
SMA(1)	-0.528414	0.170160	0.00190	SMA(3)	-0.340523	0.159345	0.03260
SMA(3)	-0.456354	0.162769	0.00505				

7) SARIMA(0,0,1)(1,1,[1,0,1])12			
Parameter	Schätzung	StdErr	p-Wert
MA(1)	0.155712	0.019096	0.00000
SAR(1)	-0.073862	0.031927	0.02085
SMA(1)	-0.918293	0.199935	0.0000
SMA(3)	-0.080801	0.023603	0.00064

Das beste Modell, das mit der Backward-Elimination für SARIMA(2,0,2)(3,1,3) ausgewählt wurde, ist SARIMA(0,0,1)(1,1,[1,0,1])12. Die AIC-Kriterium dafür ist: -2089.66009

Für `SARIMA(2,1,2)(3,1,3)12` erhalten wir:

$$X_t - X_{t-12} = W_t \tag{3.3}$$

$$W_t - W_{t-1} = Y_t \tag{3.4}$$

$$\begin{aligned} Y_t = {} & y + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-12} + \phi_4 Y_{t-24} \\ & + \phi_5 Y_{t-36} + \theta_1 \varepsilon_{y_{t-1}} + \theta_2 \varepsilon_{y_{t-2}} + \theta_3 \varepsilon_{y_{t-12}} + \theta_4 \varepsilon_{y_{t-24}} \\ & + \theta_5 \varepsilon_{y_{t-36}} + \varepsilon_{y_t}. \end{aligned} \tag{3.5}$$

Tabelle 3.2: Backward-Elimination für SARIMA(2,1,2)(3,1,3)12

1) SARIMA(2,1,2)(3,1,3)12				2) SARIMA(2,0,[0,1])(3,1,3)12			
Parameter	Schätzung	StdErr	p-Wert	Parameter	Schätzung	StdErr	p-Wert
AR(1)	-0.786690	0.119312	0.00000	AR(1)	-0.142220	1.010852	0.00000
AR(2)	0.156436	0.035311	0.00000	AR(2)	0.323325	1.009808	0.00000
MA(1)	**-0.040638**	**0.118017**	**0.73059**	MA(2)	0.061447	0.176090	0.00000
MA(2)	-0.957674	0.119167	0.00000	SAR(1)	-1.288512	0.137238	0.00000
SAR(1)	-1.277025	0.083898	0.00000	SAR(2)	-0.450098	0.146599	0.00000
SAR(2)	-0.830759	0.087470	0.00000	SAR(3)	0.079464	0.020466	0.08170
SAR(3)	0.026936	0.024210	0.26587	SMA(1)	0.283417	0.145203	0.86700
SMA(1)	0.286173	0.267524	0.28475	**SMA(2)**	**-0.756182**	**0.099771**	**0.91889**
SMA(2)	-0.403089	0.318627	0.00016	SMA(3)	-0.517050	0.140424	0.83533
SMA(3)	-0.880962	0.233121	0.00276				
3) SARIMA(2,0,[0,1])(3,1,[1,0,1])12				4) SARIMA(2,0,[0,1])(3,1,[0,0,1])12			
Parameter	Schätzung	StdErr	p-Wert	Parameter	Schätzung	StdErr	p-Wert
AR(1)	0.160392	0.018734	0.00000	AR(1)	0.157363	0.018279	0.00000
AR(2)	0.040258	0.027904	0.00000	AR(2)	-1,780351	0.047233	0.00000
MA(2)	-1.556658	0.100026	0.00000	MA(2)	-0.878240	0.058190	0.00000
SAR(1)	-0.658956	0.109696	0.00000	SAR(1)	0.028735	0.020578	0.00000
SAR(2)	0.069623	0.019721	0.00000	SAR(2)	0.785023	2.446928	0.00000
SAR(3)	0.573085	0.297064	0.72990	**SAR(3)**	**-0.884180**	**4.369805**	**0.95064**
SMA(1)	**-0.839754**	**0.425905**	**0.92201**	SMA(3)	0.900639	0.207475	0.00514
SMA(3)	-0.731594	0.224678	0.00113				

5) SARIMA(2,0,[0,1])(2,1,[0,0,1])12			
Parameter	Schätzung	StdErr	p-Wert
AR(1)	-0.828501	0.018655	0.00000
AR(2)	0.160705	0.018802	0.00000
MA(2)	-0.992828	0.012020	0.00000
SAR(1)	-0.994623	0.029225	0.00000
SAR(2)	-0.996170	0.53104	0.00000
SMA(3)	-0.994651	0.125828	0.00000

Das beste Modell für `SARIMA(2,1,2)(3,1,3)` ist `SARIMA(2,1,[0, 1])(2,1,[0,0,1])`. Die AIC-Kriterium dafür ist: -2073.2228

3.4.3 Brute-Force

Mit der Brute-Force-Methode werden Grenzen für mögliche Parameter für das SARIMA-Modell basierend auf der ACF und PACF festgelegt. Anschließend wird für jede Kombination von Parametern das AIC-Kriterium berechnet, und das Modell mit dem geringsten AIC-Wert wird ausgewählt. In diesem Abschnitt wird die

Brute-Force-Methode sowohl für die Periodendifferenz als auch für die Erstdifferenzierung der Periodendifferenz angewendet.

Hier wird das Python-Skript in dem Buch von Jochen Hirschle (siehe [7] Abschnitt 4.6.2) angewendet.

Für Periodendifferenz: `SARIMA(2,0,2)(3,1,3)12` erhalten wir das beste Model:
`SARIMA(1,0,0)([0,1],1,[1,0])12` mit AIC-Kriterium = -2092.4219. Für Erstdifferenz von Periodendifferenz: `SARIMA(2,1,2)(3,1,3)12` erhalten wir das beste Model:
`SARIMA(1,1,1)([0,1],1,1)12` mit AIC-Kriterium = -2080.2926. .

Der Koeffizient und der p-Wert für jeden Koeffizienten sind:

```
       coef  std err  t value p value
ar.L1   0.166164 0.018501 8.981276 0.00000
ar.S.L24 0.065807 0.016770 3.924132 0.00009
ma.S.L12 -0.999843 0.867901 -1.152024 0.24931
sigma2  0.011901 0.010300 1.155435 0.24791
```

3.5 Modelldiagnose

Durch den Schritt der Modellidentifikation in Abschnitt 3.3 werden die folgenden SARIMA-Modelle diagnostiziert:

- `SARIMA(0,0,1)(1,1,[1,0,1])12` (Backward-Elimination, AIC = -2089.66)
- `SARIMA(2,1,[0, 1])(2,1,[0,0,1])12` (Backward-Elimination, AIC = -2073.2228)
- `SARIMA(1,0,0)([0,1],1,1)12` (Brute-Force, AIC = -2092.4219)
- `SARIMA(1,1,1)([0,1],1,1)12` (Brute-Force, AIC = -2080.2926)

Ein gutes SARIMA-Modell sollte ein Residuum erzeugen, das einem White-Noise-Prozess folgt. Dies bedeutet, dass alle möglichen Strukturen in den Daten bereits vom Modell erfasst wurden und das, was übrig bleibt, eine Zufälligkeit ist, die keinen Vorhersagewert hat. Bei der Überprüfung des Modells wird das Residuum auf White-Noise getestet, was bedeutet, dass das Residuum einen konstanten Mittelwert haben sollte und keine Korrelation zwischen den Beobachtungen aufweisen darf. Eine konstante Varianz und eine normalverteilte Verteilung des Residuums sind hilfreich, aber nicht zwingend erforderlich, um das Vorhersageintervall zu berechnen (siehe Abschnitt 3.3, [6]). Hierbei wird der Ljung-Box-Test verwendet, um die Autokorrelation zu testen, der ADF-Test zur Überprüfung der Stationarität und der Residuen vs vorhergesagte Werte Graph zur Überprüfung der konstanten Varianz (Heteroskedastizität). Außerdem werden die Residuen mit ihrem Histogramm und einem Normal-Probability-Plot dargestellt. Für den Ljung-Box-Test wird empfohlen, dass die Lag-Einstellung das Zweifache des saisonalen Faktors beträgt [8].

3.5.1 `SARIMA(0,0,1)(1,1,[1,0,1])12`-Modelldiagnose

SARIMA(0,0,1)(1,1,[1,0,1])12		
Ljung-Box-Test	Ljung-Box-Teststatistik	15.986772
	Ljung-Box-p-Wert	0.888553
ADF-Test	Teststatistik	-24.9715
	p-Wert	0.0000

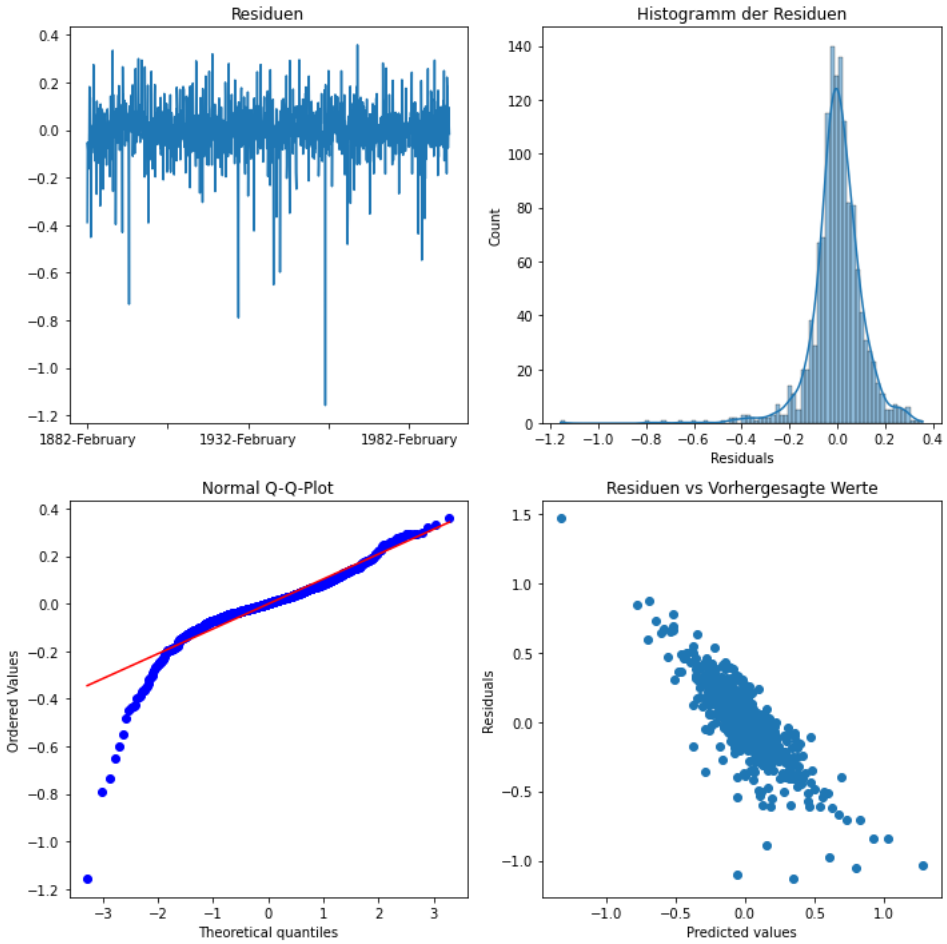

Abbildung 3.7: Modelldiagnose

3.5.2 SARIMA(2,1,[0, 1])(2,1,[0,0,1])12-Modelldiagnose

SARIMA(2,1,[0,1])(2,1,[0,0,1])12		
Ljung-Box-Test	Ljung-Box-Teststatistik	23.083788
	Ljung-Box-p-Wert	0.514856
ADF-Test	Teststatistik	-36.9101
	p-Wert	0.0000

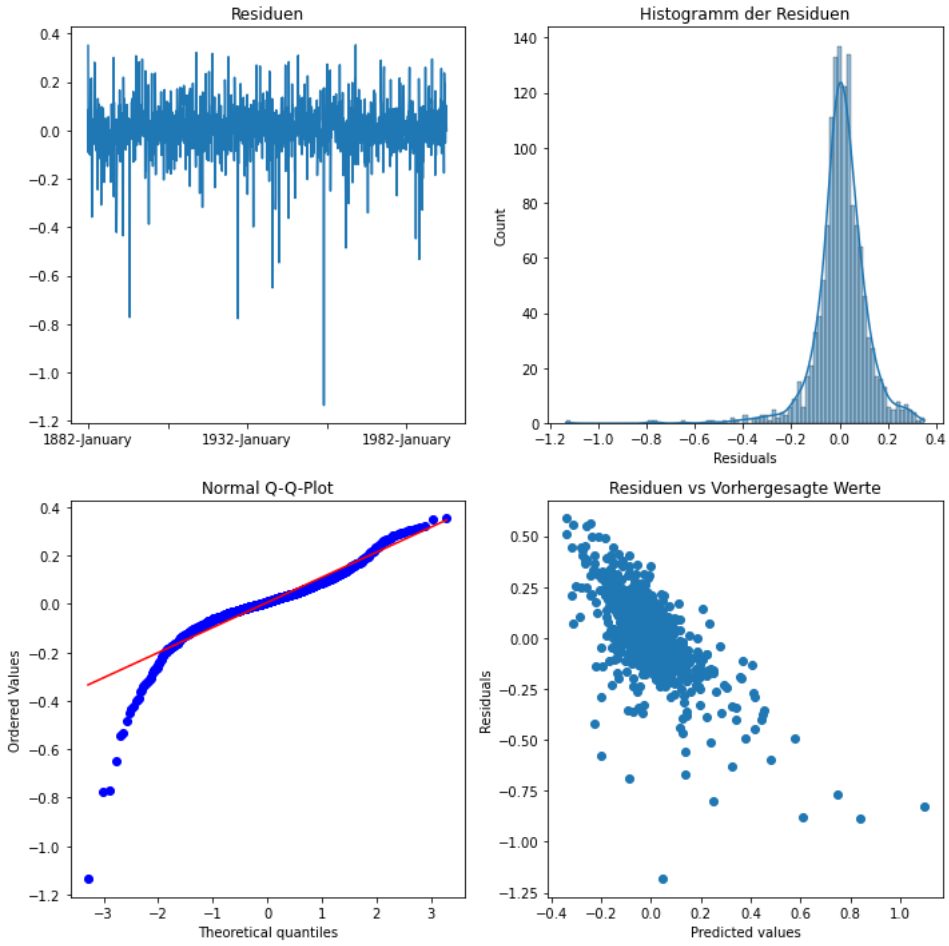

Abbildung 3.8: Modelldiagnose

3.5.3 SARIMA(1,0,0)([0,1],1,1)12-Modelldiagnose

SARIMA(1,0,0)([0,1],1,1)12		
Ljung-Box-Test	Ljung-Box-Teststatistik	15.39517
	Ljung-Box-p-Wert	0.908662
ADF-Test	Teststatistik	-36.8324
	p-Wert	0.0000

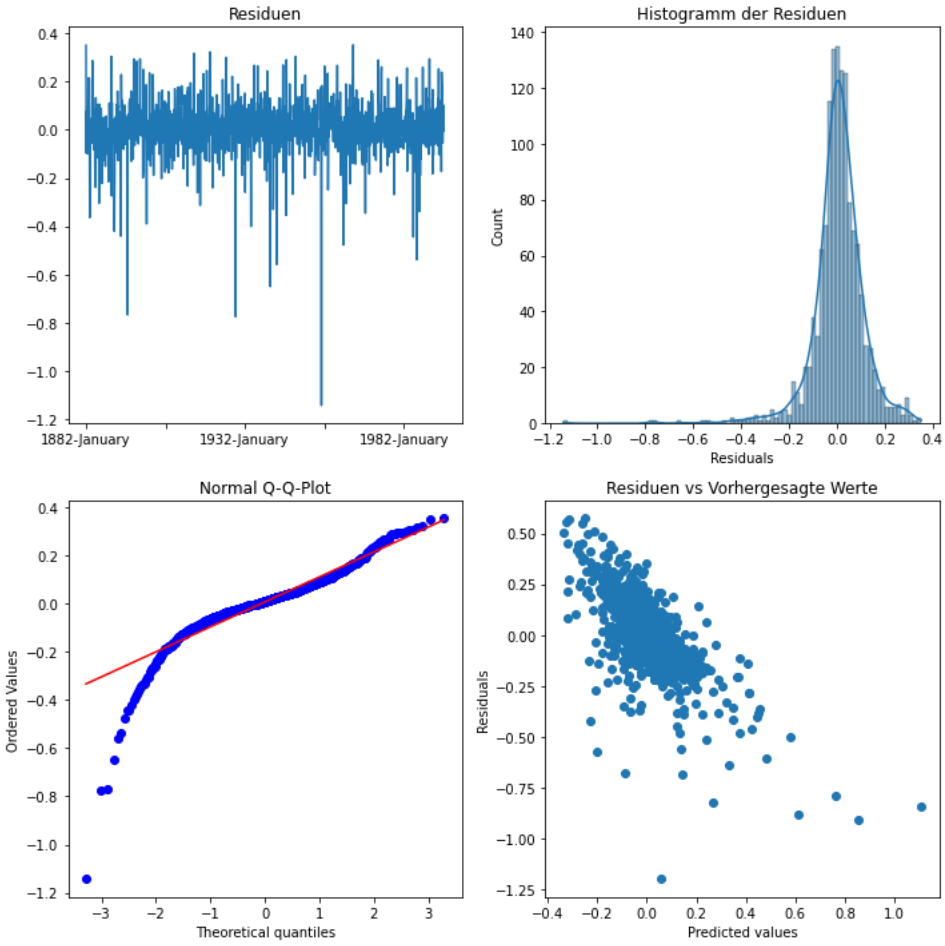

Abbildung 3.9: Modelldiagnose

3.5.4 SARIMA(1,1,1)([0,1],1,1)12-Modell

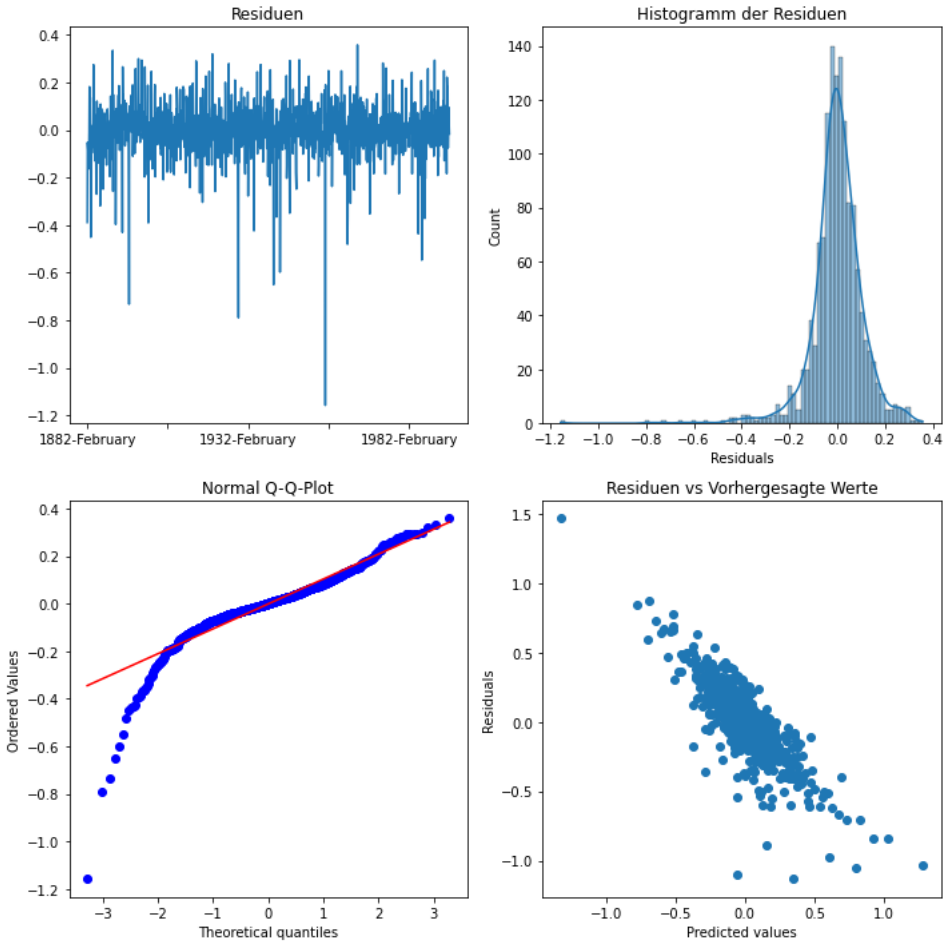

Abbildung 3.10: Modelldiagnose

Die Ergebnisse zeigen ein akzeptables diagnostisches Ergebnis, um zum nächsten Schritt für alle gewählten Modelle überzugehen. Die Residuen scheinen keiner normalen Verteilung zu folgen, und es besteht immer noch eine Korrelation zwischen den Residuen und den vorhergesagten Werten. Dies zeigt, dass das Modell gut für normale Werte funktioniert, aber bei den extremen Werten der wahren Werte schlecht abschneidet.

3.6 Vorhersage mit dem ausgewählten Modell

Da die Daten differenziert, logarithmiert und transponiert wurden, müssen die folgenden Schritte durchgeführt werden, um den vorhergesagten Wert zu erhalten: Zuerst wird die Exponentialfunktion verwendet, um den Log-Effekt zu entfernen. Dann wird 15 von den Daten abgezogen. Anschließend können wir mit Hilfe der Funktion 3.1 für die Periodendifferenz und der Funktionen 3.3 und 3.4 für die Erstdifferenz der Periodendifferenz den ursprünglichen Wert X_t finden. Um den Prozess zu vereinfachen, wird hier die Funktion `SARIMAX` aus der `statmodels`-Bibliothek verwendet, um den vorhergesagten Wert einer logarithmierten Datenreihe basierend auf einem gegebenen SARIMA-Modell zu generieren. Nach der Vorhersage sind nur noch die Verwendung der Exponentialfunktion und die Subtraktion von 15 erforderlich.
Zur Bewertung der Vorhersageleistung auf den Testdatensatz und Trainingsdatensatz werden zwei Metriken verwendet: RMSE (Root Mean Square Error) und MAE (Mean Absolute Error). Anschließend wird die Vorhersage des besten Modells in einem bestimmten Intervall visualisiert. Da die Trendkomponente in

der ursprünglichen Zeitreihe nicht so deutlich erkennbar ist, werden wir auch die aus der Vorhersage resultierende Trendkomponente und die ursprüngliche Zeitreihe mithilfe der Funktion `seasonal_decompose` analysieren, um zu sehen, ob das Modell die Trendkomponente in der ursprünglichen Zeitreihe erfassen kann.

	SARIMA(0,0,1)(1,1,[1,0,1])12	SARIMA(1,0,0)([0,1],1,1)12	SARIMA(1,1,1)([0,1],1,1)12-Modell	SARIMA(2,1,[0,1])(2,1,[0,0,1])12
RMSE	3.76370	3.75813	4.29740	4.31025
MAE	1.42230	1.421595	1.45083	1.45468

Tabelle 3.3: RMSE und MAE für Trainingsdatensatz

	SARIMA(0,0,1)(1,1,[1,0,1])12	SARIMA(1,0,0)([0,1],1,1)12	SARIMA(1,1,1)([0,1],1,1)12	SARIMA(2,1,[0,1])(2,1,[0,0,1])12
RMSE	3.37096	3.36472	3.11144	3.05933
MAE	1.46710	1.46472	1.41009	1.39520

Tabelle 3.4: RMSE und MAE für Testdatensatz

Die Ergebnisse des Trainingsdatensatzes zeigen, dass das Modell mit dem niedrigsten AIC-Kriterium ausgewählt wurde. Die Ergebnisse des Testdatensatzes zeigen jedoch, dass das Modell, das auf der ersten Differenz der saisonalen Differenz basiert, besser abschneidet. Wir werden das Modell SARIMA(2,1,[0,1])(2,1,[0,0,1])12 für weitere Vorhersagen auswählen.

3.6.1 Vorhersage

Jetzt werden der Unterschiedsteil der vorhergesagten Zeitreihe sowie die Trendkomponente der vorhergesagten Zeitreihe mit der ursprünglichen verglichen und dargestellt. Zuerst wird das Intervall von 2000 bis 2010 visualisiert.

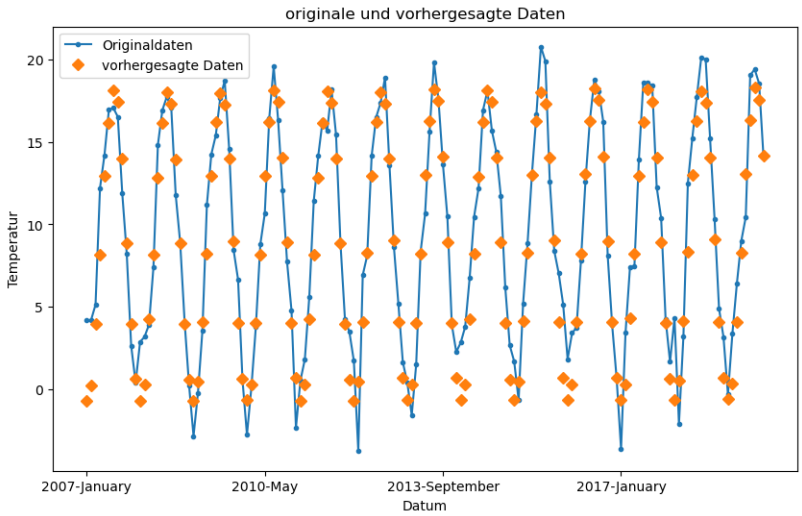

Abbildung 3.11: Testdatensatzvorhersagen

Anschließend wird eine Vorhersage für die nächsten 2 Jahre gemacht.

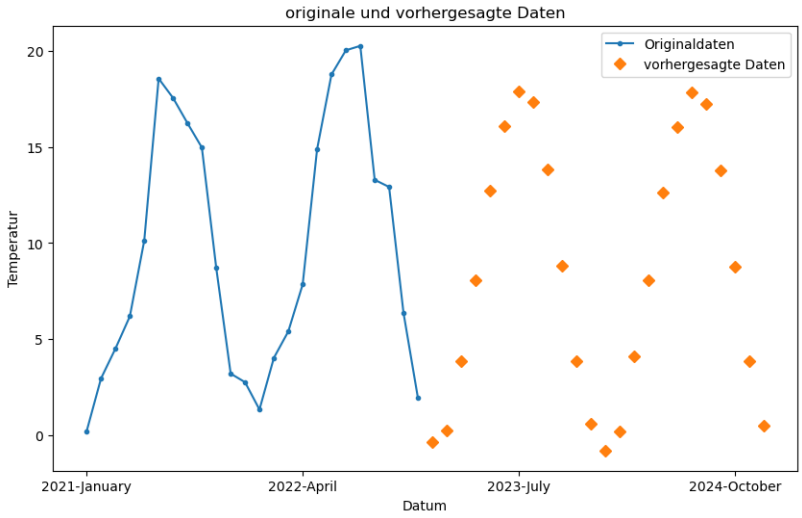

Abbildung 3.12: Vorhersagen Januar-2023 bis Demzember-2024

Die Trendkomponente wurde vom Modell nicht besonders gut erfasst. Der Grund dafür könnte sein, dass die Trainingsdaten nur Werte bis Dezember 1994 enthalten und der Trend erst nach 1960 begonnen hat. Obwohl der Trend in der vorhergesagten Zeitreihe in Bezug auf die Varianz stabiler ist, kann dies darauf zurückzuführen sein, dass in der ursprünglichen Zeitreihe mehr extreme Werte vorhanden sind (siehe Modell-Diagnose).

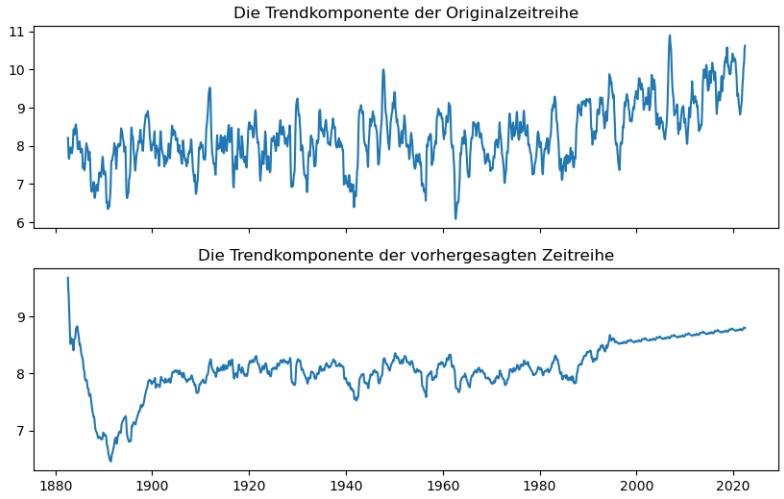

Abbildung 3.13: Die Trendkomponente der Originalzeitreihe und der vorhergesagten Zeitreihe.

Kapitel 4

Zusammenfassung

In Kapitel 1 und 2 wird das Konzept der Zeitreihenanalyse und spezieller das SARIMA-Modell vorgestellt. In Kapitel 3 wird mit Hilfe der Box-Jenkins-Methodik ein SARIMA-Modell ausgewählt, um die nächsten 24 Monate der durchschnittlichen Monatstemperatur von Baden-Württemberg vorherzusagen. Das ausgewählte Modell hat eine Leistung auf dem Testdatensatz mit einem RMSE von 3.05933 und einem MAE von 1.39520. Es gibt auch folgende Punkte und Probleme, die gründlicher betrachtet werden könnten:

- Erstens wurde in dieser Arbeit der erweiterte Dickey-Fuller-Test (augumented-Dicky-Fuller-Test) Version 1 verwendet, um die Stationarität der Zeitreihe vor und nach der Differenzierung zu testen. Der Test ist im Grunde ein Regressionsproblem und die differenzierte Version des Tests führt zu unterschiedlichen Koeffizienten und damit zu unterschiedlichen Teststatistiken, die die Zuverlässigkeit des Tests beeinflussen könnten. Bei komplexeren Fällen sollten mehr Analysen an den abhängigen und unabhängigen Variablen durchgeführt werden, um sicherer bei der Wahl einer Testversion zu sein, oder falls die Entscheidung nicht getroffen werden kann, kann stattdessen ein anderer Stationaritätstest angewendet werden.
- In dieser Arbeit wurden drei Arten von Differenzenbildung in Zeitreihen durchgeführt, um Stationarität zu erreichen. Abschnitt 1.4.5 zeigt, dass die Residuen der Differenzmodelle immer noch keiner Normalverteilung folgen und eine lange rechte Schleppe aufweisen. Das Diagramm der Residuen und prognostizierten Werte zeigt auch, dass die Residuen tendenziell größer sind, wenn der tatsächliche Wert im extremen Intervall liegt, was zeigt, dass das Modell bei extremen Werten schlecht abschneidet. Der große RMSE-Wert (3.05933) belegt diesen Punkt ebenfalls. Weitere Differenzierungen könnten durchgeführt werden, um die Residuen und die Modellleistung zu untersuchen, auch das Problem der Überdifferenzierung könnte überprüft werden, möglicherweise durch die Überprüfung auf eine Einheitswurzel im MA-Modell (siehe Abschnitt 6.3.2, Seite 171, [4]).
- Eine ähnliche Arbeit von Tara Ahmed Chawsheen und Mark Broom [11] verwendet ebenfalls das SARIMA-Modell zur Prognose der durchschnittlichen monatlichen Lufttemperatur mit einem Trainingsdatensatz von 287 Beobachtungen und erzielt einen RMSE von 1.696, der deutlich kleiner ist als der RMSE dieser Arbeit. Da wir sehr große Datensätze haben, könnten auch kleinere Beobachtungen in verschiedenen Intervallen in Betracht gezogen und ihre Leistung überprüft werden. Diese Implementierung könnte auch bestätigen, ob die Trendkomponente besser erfasst werden kann, wenn der Trainingsdatensatz eine konstante Trendkomponente enthält.
- Die schlechte Leistung bei extremen Werten könnte auch mit anderen Modellen überprüft werden, zum Beispiel einem multivariaten Modell, das auch andere Faktoren berücksichtigt, die diese extremen Werte vorhersagen könnten, die im SARIMA-Modell als zufälliger Faktor betrachtet werden könnten. Die Modelle, die sich dynamisch ändern können, könnten hier ebenfalls in Betracht gezogen werden, um mit dem Problem umzugehen, dass die Trendkomponente nicht im Trainingsdatensatz konsistent war.

Durch diese Arbeit wurden einige wichtige Aspekte der Zeitreihenanalyse und Prognose eingeführt und untersucht, indem grundlegende Zeitreihenmodelle vorgestellt und das SARIMA-Modell zur Modellierung und Vorhersage der durchschnittlichen monatlichen Lufttemperatur angewendet wurden. Zeitreihenmodellierung und Prognose sind Aufgaben, die, wenn sie gut gemacht werden sollen, ein tiefes Verständnis der zu prognostizierenden Daten und ein tiefes Verständnis der Methoden, die verwendet werden sollen, oder idealerweise vieler anderer Methoden erfordern, um die bestmögliche auszuwählen (George E. P. Box: 'All models are wrong, but some are useful'). Obwohl das von uns erzielte Ergebnis nicht wirklich optimal war, hat es einige kritische Probleme aufgezeigt, die bei der Zeitreihenanalyse und Prognose beachtet werden müssen.

Literatur

[1] Hirotugu Akaike. “Information Theory and an Extension of the Maximum Likelihood Principle”. In: 1973.

[2] John Boland. *Box-Jenkins Models*. URL: https://lo.unisa.edu.au/pluginfile.php/1156111/mod_resource/content/1/Box-Jenkins.pdf. (letzte Zugriff 12.05.2023).

[3] George EP Box u. a. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[4] Peter J Brockwell und Richard A Davis. *Introduction to time series and forecasting*. Springer Wesley, 2016.

[5] Sachin Date. *Understanding Partial Auto-correlation And The PACF*. URL: https://timeseriesreasoning.com/contents/partial-auto-correlation/. (letzte Zugriff 12.04.2023).

[6] Rob J Hyndman und George Athanasopoulos. *Forecasting: Principles and Practice*. Otexts, 2018.

[7] Jochen Hirschle. *Machine Learning für Zeitreihen*. Hanser, 2021.

[8] Rob J Hyndman. *Thoughts on the Ljung-Box test*. URL: https://robjhyndman.com/hyndsight/ljung-box-test/. (letzte Zugriff 16.05.2023).

[9] G. M. LJUNG und G. E. P. BOX. “On a measure of lack of fit in time series models”. In: *Biometrika* 65.2 (Aug. 1978), S. 297–303. ISSN: 0006-3444. DOI: 10.1093/biomet/65.2.297. eprint: https://academic.oup.com/biomet/article-pdf/65/2/297/649058/65-2-297.pdf. URL: https://doi.org/10.1093/biomet/65.2.297.

[10] Lonnie Magee und Winter. “Unit Roots , Cointegration , VARs and VECMs”. In: 2013.

[11] Tara Ahmed Chawsheen und Mark Broom. “Seasonal time-series modeling and forecasting of monthly mean temperature for decision making in the Kurdistan Region of Iraq”. In: *Journal of Statistical Theory and Practice* 11.4 (2017), S. 604–633.

[12] Robert Nau. *Random walk model*. URL: https://people.duke.edu/~rnau/411rand.htm. (letzte Zugriff 12.04.2023).

[13] Andrew Rothman. *Uncorrelated vs Independent Random Variables— Definitions, Proofs, Examples*. URL: https://towardsdatascience.com/uncorrelated-vs-independent-random-variables-definitions-proofs-examples-26422589a5d6. (letzte Zugriff 01.02.2023).

[14] E Slutsky. “The summation of random causes as the source of cyclic processes (Russian with English summary)”. In: *Prob. Ec. Cond., Inst. Ec. Conj. Moscow* (1927).

[15] *statsmodels.tsa.seasonal.seasonal$_d$ecompose*. URL: https://www.statsmodels.org/dev/generated/statsmodels.tsa.seasonal.seasonal_decompose.html. (letzte Zugriff 29.04.2023).

[16] Wikipedia. *Covariance*. URL: https://en.wikipedia.org/wiki/Covariance. (letzte Zugriff 01.02.2023).

[17] Wikipedia. *Dickey-Fuller-Test*. URL: https://de.wikipedia.org/wiki/Dickey-Fuller-Test. (letzte Zugriff 22.05.2023).

[18] Wikipedia. *Partielle Autokorrelationsfunktion*. URL: https://de.wikipedia.org/wiki/Partielle_Autokorrelationsfunktion. (letzte Zugriff 12.04.2023).

[19] Simon Wood. *Generalized additive models: an introduction with R*. Taylor Francis, 2017.

[20] George Udny Yule. "VII. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 226.636-646 (1927), S. 267–298.