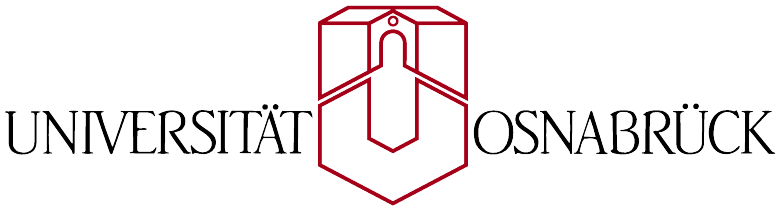
Analyse von Windgeschwindigkeiten mit ARIMA-Modellen

Bachelorarbeit

So Kumneth Sim

Matrikelnummer: 962018

Fachbereich Physik



Oktober 2017

Prüfer:

Dr. Pedro Lind
Prof. Dr. Philipp Maass

Zusammenfassung

Windgeschwindigkeiten weisen ein stochastisches Verhalten auf und können daher nicht mit beliebiger Genauigkeit vorhergesagt werden. Auch, wenn ein Tages- oder Jahresgang, d.h. eine Periodizität vorhanden ist, weichen die Messungen normalerweise von einer periodischen Funktion ab. Von Interesse ist eine Vorhersage für Windkraftwerke, um eine effiziente Nutzung der Windenergie zu ermöglichen. Aber auch meteorologische Vorhersagen beschäftigen sich mit Windgeschwindigkeiten. In dieser Bachelorarbeit werden ARIMA-Prozesse zur Beschreibung von Windgeschwindigkeiten verwendet und beurteilt, inwiefern sich ARIMA-Modelle dafür eignen.

Abstract

Wind speeds show a stochastic behavior and therefore they cannot be forecasted with arbitrary precision. Although there is a diurnal or annual movement, the measurements normally deviate from a periodic function. Wind forecasts are interesting for wind power plants to make an efficient use of the wind energy possible. But also meteorological forecasts deal with wind speeds. In this bachelor thesis ARIMA processes are used to describe wind speeds. These ARIMA models will be judged how far they are suitable for this application.

Inhaltsverzeichnis

1 Einleitung

Viele naturwissenschaftliche, aber auch wirtschaftliche Phänomene liegen in Daten vor, die nicht eindeutig durch mathematische Funktionen oder Folgen beschrieben werden können. Will man die Daten trotzdem numerisch möglichst genau beschreiben, um beispielsweise Ereignisse prognostizieren zu können, eignen sich Zeitreihenanalysen dafür. Neben Prognosen ist ein weiteres Ziel von Zeitreihenanalysen die Erkennung von Veränderungen, beispielsweise beim EEG- oder EKG-Monitoring in der Medizin. Werden in den Zeitreihenanalysen Parameter verwendet, die nicht geeignet sind, besteht die Gefahr, dass falsche Prognosen getroffen werden. Es ist mithilfe von Zeitreihen möglich, sowohl deterministische als auch stochastische Prozesse beschreiben zu können. Auch eine Mischung von deterministischen und stochastischen Prozessen kann durch Zeitreihen dargestellt werden.
In dieser Arbeit wird beschrieben, wie eine vorliegende Zeitreihe zutreffend modelliert werden kann. Zuvor wird darauf eingegangen, was „Zeitreihenanalyse“ bedeutet und welche Modelle zur Beschreibung von Zeitreihen verwendet werden können. Es soll beurteilt werden können, inwiefern ARIMA-Modelle zur Modellierung von Zeitreihen geeignet sind. Ziel dieser Arbeit ist es, die ARIMA-Modelle bei Windgeschwindigkeiten anzuwenden. Dabei sollen Windgeschwindigkeiten vorhergesagt und die Genauigkeit der Vorhersagen für verschiedene ARIMA-Modelle betrachtet werden.

2 ARIMA-Modelle zur Beschreibung stochastischer Prozesse

Bevor *Autoregressive-integrated-moving-average-Modelle* (ARIMA-Modelle) erläutert werden, werden die einzelnen Teilmodelle, die *autoregressiven Modelle* (AR-Modelle) und die *Moving-Average-Modelle* (MA-Modelle) beschrieben. Zum Verständnis wird einführend beleuchtet, was Zeitreihen und Zeitreihenanalyse sind.

2.1 Zeitreihenanalyse

Die Zeitreihenanalyse ist eine Disziplin, die sich mit der statistischen Analyse von Zeitreihen beschäftigt. Sie dient unter anderem dazu, Vorhersagen für die künftige Entwicklung von Zeitreihen machen zu können. Die Aufgabe ist es, zu einer gegebenen Zeitreihe ein geeignetes mathematisches Modell zu finden und dessen Qualität beurteilen zu können. In dieser Arbeit wird das ARIMA-Modell beschrieben, auf das im Folgenden eingegangen wird.

2.2 Zeitreihen

Eine Zeitreihe bezeichnet eine zeitliche Abfolge von Daten. Beispiele sind Wetterbeobachtungen, wirtschaftliche Trends oder fluktuierende Populationen [1, S.4]. Typischerweise sind die Daten stochastischen Charakters oder besitzen zumindest stochastische Eigenschaften. Das Ziel der Zeitreihenanalyse ist die präzise mathematische Beschreibung vorhandener Daten in Form einer Zeitreihe. Um die Daten stochastischen Charakters beschreiben zu können, wird eine Zeitreihe X als Menge von Zufallsvariablen aufgefasst:

$$X = \{x_1, x_2, \dots, x_n\} = \{x_t\} \text{ mit } t \in \mathbb{R} \text{ und } t \in \{t_1, t_2, \dots, t_n\}\,. \tag{2.1}$$

Die Zufallsvariablen x_t sind zeitabhängig, wobei die Zeitpunkte t, zu denen gemessen wird, eine diskrete Menge darstellen. Dabei müssen die zeitlichen Abstände $t_i - t_{i+1}$ zwischen zwei benachbarten Zufallsvariablen nicht zwangsläufig äquidistant sein. In dieser Arbeit wird jedoch eine Äquidistanz zwischen den Zeitpunkten t_i angenommen. Es erfolgt die Darstellung von Zeitreihen mit x_t auf der Ordinate und t auf der Abszisse. Üblicherweise werden die zeitlich benachbarten Punkte mit Linien verbunden, um mögliche Trends und eventuell vorhandene kontinuierliche Muster wie z.B. ein Signal, das von einem Rauschen überlagert ist, erkennen zu können. Die Voraussetzung dafür ist jedoch, dass innerhalb einer Zeiteinheit genügend Messdaten vorhanden sind. Es werden nun einige typische Beispiele für einfache Zeitreihen beschrieben.

Weißes Rauschen. Eine Zeitreihe aus unabhängigen normalverteilten Zufallsvariablen wird in dieser Arbeit *weißes Rauschen* genannt. Das weiße Rauschen wird als w_t notiert und hat, wenn nicht anders erwähnt, den Erwartungswert $\langle w_t \rangle = \mu_w = 0$ und die Varianz σ_w^2. Es liefert die Grundlage für viele der besprochenen stochastischen Modelle. Im Allgemeinen versteht man unter weißem Rauschen jedoch ein Signal mit einem konstanten Leistungsdichtespektrum.

Gleitender Mittelwert (MA). Ein *gleitender Mittelwert* (MA) hebt die groben Oszillationen hervor und entfernt die kleinen Oszillationen teilweise. Ein Beispiel ist

$$v_t = \frac{1}{3}\left(w_{t-1} + w_t + w_{t+1}\right). \tag{2.2}$$

Ein weißes Rauschen w_t lässt sich durch einen gleitenden Mittelwert ersetzen, der zu einer Glättung des weißen Rauschens führt.

Autoregression (AR). Ist eine Zeitreihe deterministischen Charakters oder trägt zumindest eine deterministische Komponente in sich, so eignet sich die Beschreibung durch eine *Autoregression*. Ein Beispiel ist

$$x_t = x_{t-1} - 0.4x_{t-2} + w_t. \tag{2.3}$$

Hierbei hängt der Wert der Zeitreihe zum Zeitpunkt t von den vorigen zwei Werten ab. Die stochastischen Eigenschaften kommen durch w_t zustande.

Random Walk mit Drift. Für die Beschreibung von Zeitreihen mit einem Trend eignen sich *Random Walks*. Ein Random Walk wird beispielsweise durch

$$x_t = \delta + x_{t-1} + w_t \tag{2.4}$$

beschrieben. δ wird *Drift* genannt und parametrisiert die Steigung des Trends. Gl. (2.4) kann auch als

$$x_t = \delta t + \sum_{j=1}^{t} w_j \tag{2.5}$$

geschrieben werden. Für $\delta = 0$ ergibt sich ein einfacher Random Walk, dessen Zufallsbewegung alleine durch w_t determiniert wird. Nichtsdestotrotz steigt die Varianz des Random Walks x_t mit der Zeit an, wie später in Gl. (2.13) gezeigt wird.

2.3 ARIMA-Modelle

Die ARIMA-Modelle dienen zur Modellierung von Zeitreihen verschiedener Ausprägungen. Sie umfassen sowohl rein deterministische Zeitreihen als auch Zeitreihen rein stochastischen Charakters. Ebenso beinhalten sie Modelle, die sowohl deterministische als auch stochastische Eigenschaften besitzen und beschreiben auch Zeitreihen, die einen Trend, entsprechend $\delta \neq 0$ nach Gl. (2.4) und (2.5), aufweisen. Es werden im Folgenden die Teilmodelle des ARIMA-Modells vorgestellt, um sie anschließend zusammenzuführen, sodass das ARIMA-Modell erklärt werden kann. Zuvor soll erwähnt werden, dass sich eine Zeitreihe im Allgemeinen aus einem deterministischen Prozess f und einem stochastischen Prozess g zusammensetzt:

$$\underbrace{\Delta^d[\text{Zeitreihe } X = \underbrace{\underbrace{f\left(x_{t-1},\ldots,x_{t-p}\right) + w_t}_{\text{AR}(p)} + \underbrace{g\left(w_{t-1},\ldots,w_{t-q}\right) + w_t}_{\text{MA}(q)}}_{\text{ARMA}(p,q)}]}_{\text{ARIMA}(p,d,q)}$$

2.3.1 Autoregressive Modelle (AR)

Die autoregressiven Modelle (AR-Modelle) dienen zur Beschreibung von Zeitreihen deterministischen Charakters. Der Wert x_t der Zeitreihe zum Zeitpunkt t soll hierbei aus den vorigen Zeitreihenwerten x_{t-j} errechnet werden können. Ein Beispiel dafür wurde bereits in Gl. (2.3) geliefert. Die allgemeine Definition eines AR-Modells lautet

$$\boxed{x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t = \sum_{j=1}^{p} \phi_j x_{t-j} + w_t.} \tag{2.6}$$

Man spricht bei einem Modell nach Gl. (2.6) von einem *autoregressiven Modell der Ordnung p*, abgekürzt AR(p). Die Konstanten $\phi_j \neq 0$ werden im Folgenden auch AR-Koeffizienten genannt. Es wird im Allgemeinen angenommen, dass w_t einem weißen Rauschen mit dem Erwartungswert $\mu_w = 0$ und der Varianz σ_w^2 entspricht. Der Erwartungswert von Gl. (2.6) ist $\mu_x = 0$. Falls der Erwartungswert $\mu_x \neq 0$ ist, so kann Gl. (2.6) zu

$$x_t - \mu_x = \sum_{j=1}^{p} \phi_j \left(x_{t-j} - \mu_x\right) + w_t \tag{2.7}$$

umgeschrieben werden. Der Grenzfall $w_t = 0$ führt zu einer komplett deterministischen Zeitreihe ohne Zufallskomponente.

2.3.2 Moving-Average-Modelle (MA)

Im Gegensatz zu den AR-Modellen wird bei den Moving-Average-Modellen (MA-Modellen) angenommen, dass die Zeitreihe zum Zeitpunkt t alleine vom weißen Rauschen w_t zum aktuellen Zeitpunkt t und von früheren Rauschtermen w_{t-j} zu den Zeitpunkten $t - j$ abhängig ist. Es folgt daraus die allgemeine Definition

$$\boxed{x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q} = \sum_{k=1}^{q} \theta_k w_{t-k} + w_t.} \tag{2.8}$$

Man spricht bei einem Modell nach Gl. (2.8) von einem *Moving-Average-Modell der Ordnung q*, abgekürzt MA(q). Die Konstanten $\theta_k \neq 0$ werden im Folgenden auch MA-Koeffizienten genannt. Auch hier wird angenommen, dass es sich bei w_t um weißes Rauschen handelt. Damit kann das MA-Modell für Zeitreihen stochastischen Charakters verwendet werden. Da per Definition der Erwartungswert eines weißen Rauschens $\mu_w = 0$ ist, folgt daraus, dass der Erwartungswert von Gl. (2.8) $\mu_x = 0$ ist.

2.3.3 Kombination von AR und MA: ARMA-Modelle

ARMA-Modelle sind eine Kombination aus AR(p)- und MA(q)-Modellen. Sie eignen sich somit für Zeitreihen, die sowohl eine deterministische als auch eine stochastische Komponente aufweisen. Die Kombination der Gleichungen (2.6) und (2.8) liefert

$$\boxed{x_t = \sum_{k=1}^{q} \theta_k w_{t-k} + \sum_{j=1}^{p} \phi_j x_{t-j} + w_t} \tag{2.9}$$

mit $\phi \neq 0$ und $\theta \neq 0$. Man spricht von einem *ARMA(p, q)-Modell der AR-Ordnung p und der MA-Ordnung q*. ARMA(p, q)-Modelle nach Gl. (2.9) haben den Erwartungswert $\mu_x = 0$. Ist jedoch $\mu_x \neq 0$,

kann Gl. (2.9) analog zu Gl. (2.7) modifiziert werden. Dazu wird

$$\alpha = \mu_x \left(1 - \sum_{j=1}^{p} \phi_j\right)$$

definiert, womit Gl. (2.9) als

$$x_t = \alpha + \sum_{k=1}^{q} \theta_k w_{t-k} + \sum_{j=1}^{p} \phi_j x_{t-j} + w_t \tag{2.10}$$

geschrieben werden kann. Die ARMA(p,q)-Modelle umfassen für $p = 0$ die MA(q)-Modelle und für $q = 0$ die AR(p)-Modelle. Sie bilden die Basis für alle ARIMA-Zeitreihenanalysen, da sie sowohl AR- als auch MA-Modelle beinhalten. Bei nichtstationären Zeitreihen wird eine Stationarität erzeugt, um anschließend eine ARMA(p,q)-Analyse durchführen zu können.

2.3.4 Prüfung auf Stationarität

Damit mit der eigentlichen Zeitreihenanalyse begonnen werden kann, muss die vorliegende Zeitreihe stationär sein. Es darf kein Trend wie im Beispiel des Random Walks vorhanden sein. Dafür ist es notwendig, Zeitreihen auf Stationarität zu prüfen. Es wird im Folgenden definiert, was Stationarität bedeutet, um anschließend auf numerische Methoden einzugehen, die eine vorliegende Zeitreihe auf Stationarität überprüfen. Dann wird beschrieben, wie eine nichtstationäre Zeitreihe modifiziert werden kann, um sie stationär zu machen.

Definition der Stationarität einer Zeitreihe. Eine Zeitreihe x_t wird *schwach stationär* genannt, falls

1. der Erwartungswert $\mu_x(t)$ der Zeitreihe konstant ist und nicht von der Zeit t abhängt,
2. die Kovarianzfunktion $\gamma_x(t_1, t_2)$, definiert in Gl. (2.24), von t_1 und t_2 nur bezüglich der Differenz $|t_1 - t_2|$ abhängig ist, und
3. die Varianz endlich ist: $\mathrm{Var}(x_t) = \left\langle (x_t - \mu_t)^2 \right\rangle < \infty, \forall t \in \{t_1, t_2, \ldots, t_n\}$.

Der zweite Punkt besagt, dass sich die Zeitreihe um eine gewisse Zeit verschoben immer gleich stark ähnelt, egal, welchen Zeitpunkt t man betrachtet. Anders gesagt ist eine eine Zeitreihe stationär, wenn ihre stochastischen Eigenschaften nicht zeitlich abhängig sind. Es soll noch erwähnt werden, dass man von *starker Stationarität* spricht, wenn $x_t = x_{t+s}$ für alle s gilt. Dies kommt der zeitlichen Invarianz von Funktionen gleich und tritt beispielsweise im Reellen bei stabilen Fixpunkten von Differentialgleichungen auf. In dieser Arbeit wird im Folgenden die schwache Stationarität lediglich *Stationarität* genannt.

Methoden zur Prüfung auf Stationarität. Ein stochastischer Prozess, der eine sogenannte *Einheitswurzel* aufweist, ist nichtstationär und man spricht von einem *stochastischen Trend*. Eine Einheitswurzel ist vorhanden, wenn 1 eine Nullstelle des charakteristischen Polynoms (2.11) ist. Es wird ein AR(p)-Prozess nach Gl. (2.6) betrachtet, für den $x_0 = 0$ angenommen werde. Wenn $m = 1$ eine einfache oder mehrfache Nullstelle des charakteristischen Polynoms

$$m^p - m^{p-1}\phi_1 - m^{p-2}\phi_2 - \cdots - \phi_p = m^p - \sum_{j=1}^{p} m^{p-j}\phi_j = 0 \tag{2.11}$$

ist, dann hat die Zeitreihe X eine Einheitswurzel.
Als Beispiel für einen nichtstationären Prozess werde ein AR(1)-Prozess

$$x_t = \phi_1 x_{t-1} + w_t = \sum_{k=1}^{t} \phi_1^{t-k} w_k \tag{2.12}$$

mit $x_0 = 0$ angeführt. Für einen Random Walk, welcher nichtstationär ist, gilt $\phi_1 = 1$ und es folgt daraus das charakteristische Polynom $m - \phi_1 = m - 1$ mit der Nullstelle $m = 1$. Die Varianz von Gl. (2.12) mit $\phi_1 = 1$ ist dann

$$\gamma(t,t) = \sigma_x^2(t) = \sum_{k=1}^{t} \sigma_w^2 = t\sigma_w^2, \tag{2.13}$$

womit $\gamma(t,t)$ zeitabhängig ist und daher entsprechend der 2. Eigenschaft stationärer Zeitreihen keine Stationarität vorliegt.
Setzt man in Gl. (2.11) $m = 1$, so erhält man das Kriterium

$$\sum_{j=1}^{p} \phi_j = 1 \tag{2.14}$$

für nichtstationäre Zeitreihen. Das Kriterium (2.14) kann für einen Random Walk nach Gl. (2.4) sofort verifiziert werden. Für stationäre Zeitreihen gilt stattdessen

$$\sum_{j=1}^{p} \phi_j < 1. \tag{2.15}$$

Eine Zeitreihe, für die

$$\sum_{j=1}^{p} \phi_j > 1 \tag{2.16}$$

gilt, heißt *explosiv*. Explosive Prozesse wachsen exponentiell und sind nichtstationär. Damit gilt für nichtstationäre Zeitreihen grundsätzlich

$$\sum_{j=1}^{p} \phi_j \geq 1. \tag{2.17}$$

Algorithmen, die bei einer gegebenen Zeitreihe auf die Existenz oder Abwesenheit einer Einheitswurzel und damit auf (Nicht-)Stationarität prüfen, sind der *Augmented Dickey-Fuller-Test* (ADF), der *Phillips-Perron-Test* (PP) und der *Kwiatkowski-Phillips-Schmidt-Shin-Test* (KPSS). Um als Beispiel den ADF-Test zu beschreiben, werde ein AR(p)-Prozess nach Gl. (2.6) betrachtet. Der ADF-Test versucht, bei einer gegebenen Zeitreihe Nichtstationarität auszuschließen. Dies ist gleichbedeutend mit der Erfüllung von Gl. (2.15). Für den ADF-Test werden daher Werte für die Koeffizienten $\phi_1, \cdots, \phi_p$ geschätzt und überprüft, ob Gl. (2.15) erfüllt wird. Wird Gl. (2.15) erfüllt, gibt es keine Einheitswurzel und der Prozess ist stationär. Um auf Erfüllung von Gl. (2.15) prüfen zu können, muss die Anzahl p der AR-Parameter vorgegeben werden. Als Beispiel für einen ADF-Test [2] werde ein AR(1)-Prozess der Form

$$x_t = \phi x_{t-1} + \delta + w_t$$

betrachtet. Die AR-Ordnung $p = 1$ sei bereits gegeben. Es wird die Nullhypothese

$$H_0 : \phi = 1 \text{ (Random Walk mit Drift)}$$

gegen die Alternativhypothese

$$H_1 : \phi < 1 \text{ (AR(1)-Prozess)}$$

getestet. Setzt man $\varphi := \phi - 1$, kann man den AR(1)-Prozess als

$$\Delta x_t = x_t - x_{t-1} = (\phi - 1)x_{t-1} + \delta + w_t = \varphi x_{t-1} + \delta + w_t$$

schreiben. Nun wird die Nullhypothese

$$H_0 : \varphi = 0 \text{ (Random Walk mit Drift)}$$

gegen die Alternativhypothese

$$H_1 : \varphi < 0 \text{ (AR(1)-Prozess)}$$

getestet. Es wird eine Regression von Δx_t durch x_{t-1} und δ durchgeführt. Mit der Methode der kleinsten Quadrate oder der Maximum-Likelihood-Methode erhält man dann die Schätzwerte $\hat{\delta}, \hat{\varphi}$. Damit wird die Teststatistik

$$T := \frac{\hat{\varphi}}{\sigma_{\hat{\varphi}}}$$

gebildet, die einer spezifischen *Dickey-Fuller-Verteilung* folgt. Dabei ist $\sigma_{\hat{\varphi}}$ die Standardabweichung bezüglich $\hat{\varphi}$, die für eine Zeitreihe mit n Elementen durch

$$\sigma_{\hat{\varphi}} = S \left(\sum_{t=2}^{n} (x_{t-1} - \mu_x)^2 \right)^{-\frac{1}{2}} \quad \text{mit } S^2 = \sum_{t=1}^{n} \left(\Delta x_t + \mu_x \hat{\varphi} - \hat{\varphi} x_{t-1} \right)^2 / (n-3)$$

berechnet wird [2]. Die Nullhypothese wird verworfen, wenn die Teststatistik T kleiner ist als der für die Signifikanz gewählte Schwellenwert. Für eine nichtstationäre Zeitreihe muss T gegen Null gehen, bei stationären Prozessen dahingegen ist $T < 0$. Für AR-Ordnungen $p > 1$ muss in der Berechnung lediglich $\varphi = 1 - \sum_{j=1}^{p} \phi_j$ gesetzt werden.
Um ein Gefühl für den Charakter für verschiedene Fälle von Stationarität und Nichtstationarität von Zeitreihen zu gewinnen, werden Gl. (2.14) und (2.15) anhand von Abb. 2.1 demonstriert. Für $\sum_{i=1}^{p} \phi_i \to 1$ wird ein Prozess im Grenzfall nichtstationär, was in Abb. 2.1 zu erkennen ist.

Erzeugung der Stationarität einer nichtstationären Zeitreihe. Zeigen die Einheitswurzeltests, dass die vorliegende Zeitreihe X nicht stationär ist, so müssen die Zeitreihenwerte x_t wie folgt differenziert werden:

$$\Delta x_t = x_t - x_{t-1}. \tag{2.18}$$

Es sei B der *Backshift-Operator*:

$$B x_t = x_{t-1}.$$

Damit lässt sich Gl. (2.18) auch als

$$(1 - B)\, x_t = x_t - x_{t-1} \tag{2.19}$$

schreiben. Mit dem Backshift-Operator lassen sich mehrfache Differentiationen algebraisch besser beschreiben. Nach der ersten Differentiation wird erneut auf Stationarität in der oben beschriebenen Vorgehensweise geprüft. Liegt immer noch keine Stationarität vor, muss unter Umständen eine mehrfache Differentiation durchgeführt werden. Ist zum Schluss eine Stationarität vorhanden, kann die mehrfach differenzierte stationäre Zeitreihe durch ein ARMA(p, q)-Modell beschrieben werden.

2.3.5 Berücksichtigung von nichtstationären Zeitreihen: ARIMA-Modelle

Zeitreihenmodelle, die auch den Fall mitberücksichtigen, dass eine vorliegende Zeitreihe nichtstationär sein kann und damit einen Trend aufweisen kann, sind die ARIMA-Modelle. Bevor die Definition einer ARIMA-Zeitreihe erklärt wird, wird die Problemstellung anhand eines Beispiels erläutert.
Betrachtet werde ein Random Walk mit Drift wie in Gl. (2.5) beschrieben. Der Erwartungswert des Random Walks mit Drift berechnet sich zu

$$\mu_x = \langle x_t \rangle = \left\langle \delta t + \sum_{j=1}^{t} w_j \right\rangle = \langle \delta t \rangle + \sum_{j=1}^{t} \underbrace{\langle w_j \rangle}_{=0} = \delta t. \tag{2.20}$$

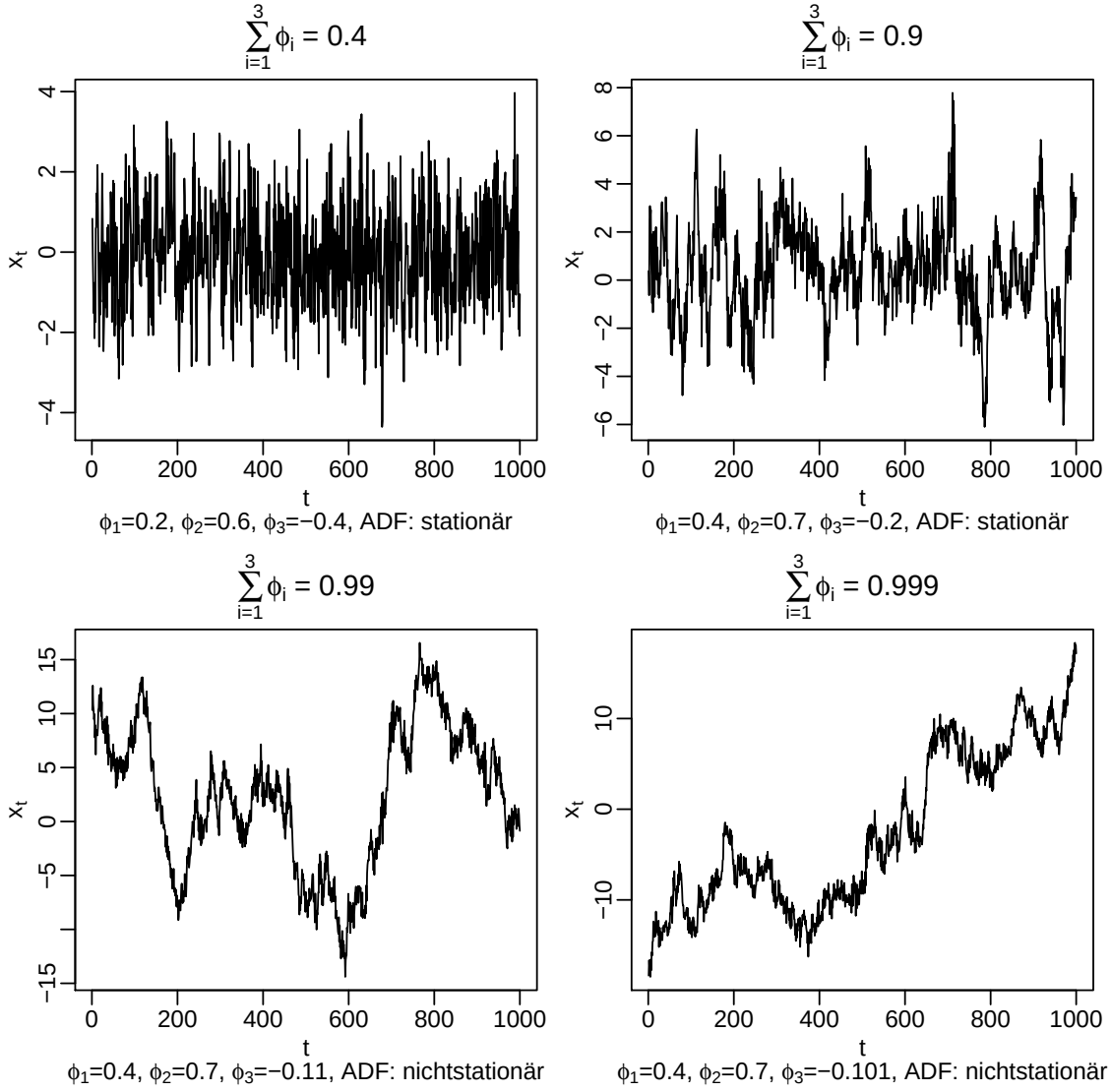


Abbildung 2.1: Verschiedene AR(3)-Prozesse zur Demonstration der Nichtstationarität.

Damit ist gezeigt, dass ein Random Walk mit Drift nicht stationär ist: Entsprechend der 1. Eigenschaft einer stationären Zeitreihe darf ihr Erwartungswert nicht zeitabhängig sein, was jedoch der Fall ist, wie Gl. (2.20) aussagt. Auch die Varianz eines Random Walks wächst trotz eines nicht vorhandenen Drifts linear mit der Zeit, wie Gl. (2.13) zeigt. Damit wird die 3. Eigenschaft einer stationären Zeitreihe für $t \to \infty$ verletzt.
Wird Gl. (2.5) differenziert, wie es in Gl. (2.18) beschrieben ist, erhält man

$$\Delta x_t = x_t - x_{t-1} = \left(\delta t + \sum_{j=1}^{t} w_j\right) - \left(\delta(t-1) + \sum_{j=1}^{t-1} w_j\right) = \delta + w_t. \tag{2.21}$$

Der Drift δ ist der Erwartungswert der differenzierten Zeitreihe Δx_t, was verständlich ist, wenn man bedenkt, dass es sich bei Δx_t um die Steigung eines Random Walks mit Drift handelt. Das weiße Rauschen w_t ist laut Definition unkorreliert und dessen Varianz σ_w^2 bleibt zeitlich konstant. Da der Drift δ konstant ist, folgt daraus, dass sowohl der Erwartungswert der differenzierten Zeitreihe Δx_t als auch deren Varianz für alle Zeiten konstant sind. Somit lässt sich sagen, dass Gl. (2.21) eine stationäre

Zeitreihe darstellt, was durch die Differentiation von Gl. (2.5) erreicht wurde. Ein weiteres Beispiel wird in Abschnitt 3.3 geliefert.
Der Random Walk ist ein vereinfachendes Beispiel für eine Zeitreihe, die einen Trend beinhaltet. Für andere Zeitreihen, die einen Trend aufweisen, muss man auf analoge Art vorgehen, um eine Stationarität zu induzieren, sodass eine ARMA(p, q)-Analyse durchgeführt werden kann. Gegebenenfalls muss die zu untersuchende Zeitreihe wie in Gl. (2.18) beschrieben d-fach differenziert werden. Es werde die d-fach differenzierte Zeitreihe als

$$\Delta^d x_t =: (1 - B)^d \, x_t \tag{2.22}$$

mit dem Backshift-Operator B geschrieben. Ein stochastischer Prozess x_t ist eine ARIMA(p, d, q)-Zeitreihe, wenn $(1 - B)^d \, x_t$ eine ARMA(p, q)-Zeitreihe ist:

$$\boxed{(1 - B)^d \, x_t = \sum_{k=1}^{q} \theta_k w_{t-k} + \sum_{j=1}^{p} \phi_j \, (1 - B)^d \, x_{t-j} + w_t.} \tag{2.23}$$

Mit den ARIMA(p, d, q)-Modellen werden demnach auch nichtstationäre Zeitreihen berücksichtigt. Letztendlich kann jede ARIMA(p, d, q)-Zeitreihe durch d-fache Differentiation wie in Gl. (2.18) beschrieben in eine ARMA(p, q)-Zeitreihe überführt werden.

2.4 Identifizierung der Abhängigkeiten zwischen den Werten einer Zeitreihe

Um eine Zeitreihenanalyse durchführen zu können, ist es von wesentlicher Bedeutung, die Abhängigkeiten zwischen den Werten einer Zeitreihe quantifizieren zu können. Es werden im Folgenden Instrumente vorgestellt, die dafür notwendig sind.

2.4.1 Autokovarianz

Die *Autokovarianz* einer Zeitreihe mit den Werten $x_t = x(t)$ und dem Erwartungswert μ_x ist als

$$\gamma_x \, (t_1, t_2) = \langle (x(t_1) - \mu_x(t_1)) \, (x(t_2) - \mu_x(t_2)) \rangle =: \mathrm{Cov} \, (x(t_1), x(t_2)) \tag{2.24}$$

definiert. Sie beschreibt die Korrelation einer Zeitreihe mit sich selbst zu einem früheren Zeitpunkt. Bei stationären Zeitreihen gibt sie an, wie stark die Ähnlichkeit der Zeitreihe $x(t)$ zu einem Zeitpunkt t mit der um den Lag $\tau := |t_1 - t_2|$ verschobenen Zeitreihe $x(t + \tau)$ ist. Besteht zwischen den Gliedern der Zeitreihe eine Beziehung, die nicht rein zufällig ist, weicht die Autokovarianz signifikant von Null ab und die Werte der Zeitreihe sind *autokorreliert*. Für $\tau = 0$ entspricht die Autokovarianz $\gamma_x \, (t, t)$ der Varianz σ_x^2, was durch die Definition in (2.24) klar wird.
Für einen stationären Prozess sind weder die Varianz noch der Erwartungswert zeitabhängig. Wie bereits in der 2. Eigenschaft von stationären Zeitreihen gefordert, ist die Autokovarianz dann nur vom Lag τ zwischen t_1 und t_2 abhängig:

$$\gamma_x \, (\tau) = \langle (x_t - \mu_x) \, (x_{t+\tau} - \mu_x) \rangle = \mathrm{Cov} \, (x_t, x_{t+\tau}) \, . \tag{2.25}$$

Für stationäre Zeitreihen gilt die Eigenschaft

$$\gamma(\tau) = \gamma(-\tau). \tag{2.26}$$

2.4.2 Autokorrelationsfunktion (ACF)

Die *Autokorrelationsfunktion* (ACF) einer Zeitreihe entspricht der normierten Autokovarianz. Es sei $\sigma_x(t_i)$ die Standardabweichung von $x(t_i)$ zum Zeitpunkt t_i. Die ACF ist als

$$\rho_x(t_1, t_2) = \frac{\gamma_x(t_1, t_2)}{\sigma_x(t_1)\,\sigma_x(t_2)} \tag{2.27}$$

definiert. Durch die Normierung hat die ACF die Eigenschaft

$$-1 \leq \rho_x(t_1, t_2) \leq 1.$$

Da für einen stationären stochastischen Prozess die Varianz $\gamma_x(t,t) = \sigma_x^2$ zeitunabhängig ist, ist auch die Standardabweichung $\sigma_x = \sqrt{\gamma_x(t,t)}$ nicht zeitlich abhängig. Das Produkt im Nenner der Definition (2.27) entspricht dann lediglich der zeitunabhängigen Varianz σ_x^2. Mit der Schreibweise $\tau = |t_1 - t_2|$ ist die ACF für eine stationäre Zeitreihe

$$\rho_x(t_1, t_2) = \rho_x(\tau) = \frac{\gamma_x(\tau)}{\sigma_x^2} = \frac{\gamma_x(\tau)}{\gamma_x(0)}, \tag{2.28}$$

da $\sigma_x^2 = \gamma_x(\tau = 0)$ gilt. Aus Gl. (2.28) folgt, dass immer $\rho(\tau = 0) = 1$ ist; eine nicht verschobene Zeitreihe entspricht schließlich exakt sich selbst. Für stationäre Zeitreihen gilt zusätzlich die Eigenschaft

$$\rho(\tau) = \rho(-\tau). \tag{2.29}$$

2.4.3 Partielle Autokorrelationsfunktion (PACF)

Die *partielle Autokorrelationsfunktion* (PACF) ist für stationäre Zeitreihen x_t als

$$\varphi_{kk} = \mathrm{Corr}\left(x_t, x_{t-k} | x_{t-1}, \cdots, x_{t-k+1}\right) \tag{2.30}$$

definiert. Dabei bezeichnet $\mathrm{Corr}(\cdot,\cdot|\cdot)$ die *bedingte Korrelation*. Diese entspricht der Autokorrelation, wenn die Werte für $x_{t-1}, \cdots, x_{t-k+1}$ bekannt sind. Die bedingte Korrelation wird aus den bedingten Varianzen und Kovarianzen berechnet:

$$\mathrm{Corr}\left(x_t, x_{t-k} | x_{t-1}, \cdots, x_{t-k+1}\right) =$$

$$\frac{\mathrm{Cov}\left(x_t, x_{t-k} | x_{t-1}, \cdots, x_{t-k+1}\right)}{\sqrt{\mathrm{Var}\left(x_t | x_{t-1}, \cdots, x_{t-k}\right)\mathrm{Var}\left(x_{t-k} | x_{t-1}, \cdots, x_{t-k+1}\right)}}.$$

Die bedingte (Ko-)Varianz kann wie folgt berechnet werden:

$$\text{bedingte Varianz: } \mathrm{Var}(\mathcal{A}|\mathcal{C}) = \left\langle \mathcal{A}^2 | \mathcal{C} \right\rangle - \langle \mathcal{A}|\mathcal{C}\rangle^2$$

$$\text{bedingte Kovarianz: } \mathrm{Cov}(\mathcal{A}, \mathcal{B}|\mathcal{C}) = \langle \mathcal{A}\mathcal{B}|\mathcal{C}\rangle - \langle \mathcal{A}|\mathcal{C}\rangle \langle \mathcal{B}|\mathcal{C}\rangle .$$

Bedingte statistische Größen entsprechen den normalen Größen, wenn bereits andere Werte bekannt sind. Beispielsweise beschreibt der bedingte Erwartungswert $\langle \mathcal{A}|\mathcal{C}\rangle$ den Wert, den man für das Ereignis $\mathcal{A}$ im Mittel erwartet, wenn das Ereignis $\mathcal{C}$ bereits eingetreten ist. Der bedingte Erwartungswert von $\mathcal{A}$, wenn $\mathcal{C}$ bereits eingetreten ist, kann durch

$$\langle \mathcal{A}|\mathcal{C}\rangle = \frac{\langle \mathcal{A}\rangle}{P(\mathcal{C})}$$

berechnet werden, wobei $P(\mathcal{C})$ die Wahrscheinlichkeit für das Ereignis $\mathcal{C}$ bezeichnet.
Die PACF beschreibt den linearen Zusammenhang zwischen den Werten x_t und x_{t-k}, ohne den Einfluss der dazwischen liegenden Werte $x_{t-1}, \cdots, x_{t-k+1}$ zu berücksichtigen.

2.5 Algorithmische Schätzung der ARIMA-Koeffizienten

Es liege eine ARMA(p,q)-Zeitreihe vor, für die die Parameter p und q bereits bekannt seien. Ziel ist es, die Koeffizienten $\phi_1, \cdots, \phi_p$ und $\theta_1, \ldots, \theta_q$ und zudem σ_w^2 abzuschätzen. Es werden im Folgenden Algorithmen vorgestellt, die jeweils zur Berechnung der AR- und der MA-Koeffizienten dienen.

Schätzung der AR-Koeffizienten und von σ_w^2

Es liege ein AR(p)-Zeitreihe nach Gl. (2.6) vor. Es sollen die AR-Koeffizienten ϕ_i geschätzt werden. Die Autokovarianz einer AR(p)-Zeitreihe mit $\mu_x = 0$ ist

$$\begin{aligned}\gamma(\tau) &= \mathrm{Cov}\left(x_t, x_{t+\tau}\right) \\ &= \left\langle x_t \left(\sum_{j=1}^{p} \phi_j x_{t+\tau-j} + w_{t+\tau} \right) \right\rangle \\ &= \sum_{j=1}^{p} \phi_j \gamma\left(\tau - j\right) + \mathrm{Cov}\left(w_{t+\tau}, x_t\right), \ \tau \geq 1. \end{aligned} \tag{2.31}$$

Man beachte, dass die Eigenschaft (2.26), $\gamma(\tau) = \gamma(-\tau)$, gilt. Dafür muss eine Stationarität vorliegen, was für eine ARMA(p,q)-Zeitreihe der Fall ist. Weiterhin ist

$$\mathrm{Cov}\left(w_{t+\tau}, x_t\right) = \begin{cases} \sigma_w^2 & \text{für } \tau = 0 \\ 0 & \text{sonst.} \end{cases} \tag{2.32}$$

Aus den Gleichungen (2.31) und (2.32) folgen die *Yule-Walker-Gleichungen*:

$$\gamma(\tau) = \phi_1 \gamma\left(\tau - 1\right) + \cdots + \phi_p \gamma\left(\tau - p\right) = \sum_{j=1}^{p} \phi_j \gamma\left(\tau - j\right), \ \tau \geq 1 \tag{2.33}$$

$$\sigma_w^2 = \gamma(0) - \phi_1 \gamma(1) - \cdots - \phi_p \gamma\left(p\right) = \gamma(0) - \sum_{j=1}^{p} \phi_j \gamma\left(j\right). \tag{2.34}$$

Die Yule-Walker-Gleichungen (2.33) und (2.34) können in Matrixnotation formuliert werden:

$$\underbrace{\begin{bmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(-p+1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(-p+2) \\ \vdots & \vdots & \vdots & \ddots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{bmatrix}}_{\boldsymbol{\Gamma}_p} \underbrace{\begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix}}_{\boldsymbol{\phi}} = \underbrace{\begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix}}_{\boldsymbol{\gamma}_p} \tag{2.35}$$

und

$$\sigma_w^2 = \gamma(0) - \boldsymbol{\phi}^T \boldsymbol{\gamma}_p = \gamma(0) - \underbrace{\left(\phi_1, \phi_2, \ldots, \phi_p\right)}_{\boldsymbol{\phi}^T} \underbrace{\begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix}}_{\boldsymbol{\gamma}_p}. \tag{2.36}$$

Dabei sind $\boldsymbol{\Gamma}_p = \left\{\gamma\left(k-j\right)\right\}_{k,j=1,\ldots,p}$ eine $(p \times p)$-Matrix, $\boldsymbol{\phi} = \left(\phi_1, \ldots, \phi_p\right)^T$ und $\boldsymbol{\gamma}_p = \left(\gamma(1), \ldots, \gamma(p)\right)^T$ p-Vektoren.

Um die Koeffizienten ϕ_i abschätzen zu können, werden die Gleichungen (2.35) und (2.36) mithilfe der Inversen der Matrix umgestellt. Dies liefert

$$\begin{aligned}\hat{\boldsymbol{\phi}} &= \boldsymbol{\Gamma}_p^{-1}\boldsymbol{\gamma}_p \\ \hat{\sigma}_w^2 &= \gamma(0) - \boldsymbol{\gamma}_p^T\boldsymbol{\Gamma}_p^{-1}\boldsymbol{\gamma}_p = \gamma(0) - \boldsymbol{\gamma}_p^T\hat{\boldsymbol{\phi}},\end{aligned} \tag{2.37}$$

die *Yule-Walker-Schätzer*. Es ist üblich, anstelle der Autokovarianz mit der ACF zu rechnen. Dafür müssen lediglich in Gl. (2.37) alle Autokovarianzen $\gamma(\tau)$ durch die dazugehörigen ACF $\rho(\tau)$ ersetzt und für die Varianz σ_w^2 die Normierung rückgängig gemacht werden. Es bezeichne $\boldsymbol{R}_p = \{\rho(k-j)\}_{k,j=1,\dots,p}$ eine $(p \times p)$-Matrix und $\boldsymbol{\rho}_p = (\rho(1), \dots, \rho(p))^T$ einen p-Vektor. Dann lauten die Yule-Walker-Schätzer

$$\begin{aligned}\hat{\boldsymbol{\phi}} &= \boldsymbol{R}_p^{-1}\boldsymbol{\rho}_p \\ \hat{\sigma}_w^2 &= \gamma(0)\left(1 - \boldsymbol{\rho}_p^T\boldsymbol{R}_p^{-1}\boldsymbol{\rho}_p\right) = \gamma(0)\left(1 - \boldsymbol{\rho}_p^T\hat{\boldsymbol{\phi}}\right).\end{aligned} \tag{2.38}$$

Anschließend soll ein Zahlenbeispiel angeführt werden. Es werde das AR(2)-Modell

$$x_t = 1.5x_{t-1} - 0.75x_{t-2} + w_t$$

mit $\sigma_w = 1$ betrachtet. Die Autokovarianzen und -korrelationen $\gamma(0) = 8.434$, $\rho(1) = 0.834$ und $\rho(2) = 0.476$ seien bereits nach Gl. (2.25) und Gl. (2.28) berechnet worden. Dies liefert für die AR-Koeffizienten

$$\begin{aligned}\hat{\boldsymbol{\phi}} = \begin{pmatrix}\hat{\phi}_1 \\ \hat{\phi}_2\end{pmatrix} &= \begin{bmatrix}\rho(0) & \rho(-1) \\ \rho(1) & \rho(0)\end{bmatrix}^{-1}\begin{pmatrix}\rho(1) \\ \rho(2)\end{pmatrix} = \begin{bmatrix}1 & 0.834 \\ 0.834 & 1\end{bmatrix}^{-1}\begin{pmatrix}0.834 \\ 0.476\end{pmatrix} \\ &= \begin{pmatrix}1.439 \\ -0.725\end{pmatrix},\end{aligned}$$

was den wirklichen Werten ϕ_1, ϕ_2 nahe kommt. Für die geschätzte Varianz $\hat{\sigma}_w^2$ ergibt sich

$$\begin{aligned}\hat{\sigma}_w^2 &= \gamma(0)\left(1 - \begin{pmatrix}\rho(1), & \rho(2)\end{pmatrix}\begin{pmatrix}\hat{\phi}_1 \\ \hat{\phi}_2\end{pmatrix}\right) \\ &= 8.434\left(1 - \begin{pmatrix}0.834, & 0.476\end{pmatrix}\begin{pmatrix}1.439 \\ -0.725\end{pmatrix}\right) = 1.215.\end{aligned}$$

Die Schätzung $\hat{\sigma}_w = \sqrt{\hat{\sigma}_w^2} = 1.102$ liegt damit in der Nähe von $\sigma_w = 1$.

Schätzung der MA-Koeffizienten

Es sei eine ARMA(p, q)-Zeitreihe gegeben. Die AR-Koeffizienten seien bereits geschätzt worden. Nun sollen die MA-Koeffizienten mit dem *Innovations-Algorithmus* [2] geschätzt werden. Da die Herleitung sehr umfangreich ist, wird hier auf sie verzichtet. Der Algorithmus berechnet die Koeffizienten $\theta_{q1}, \theta_{q2}, \cdots, \theta_{qq}$ für verschiedene Ordnungen q. Danach wird mithilfe von Informationskriterien validiert, welche Ordnung q am besten passt. Die Informationskriterien sind im nächsten Abschnitt beschrieben. Es sei $\{x_t\}$ eine Zeitreihe mit dem Mittel $\mu_x = 0$ und den Erwartungswerten $\left\langle |x_t|^2 \right\rangle < \infty$ für alle t. Weiterhin wird

$$\langle x_i x_j \rangle =: \kappa(i, j)$$

definiert. Die Koeffizienten können dann rekursiv durch

$$\begin{aligned} v_0 &= \kappa(1,1) \\ \theta_{q,q-k} &= v_k^{-1}\left(\kappa(q+1,k+1) - \sum_{j=0}^{q-1}\theta_{k,k-j}\theta_{q,q-j}v_j\right) \\ v_q &= \kappa(q+1,q+1) - \sum_{j=0}^{q-1}\theta_{q,q-j}^2 v_j \end{aligned}$$

berechnet werden. Zuerst wird v_0 berechnet, dann folgen (θ_{11}, v_1); $(\theta_{22}, \theta_{21}, v_2)$; ... bis zur gewünschten Ordnung q.

2.6 Informationskriterien: AIC und BIC

Die sogenannten *Informationskriterien* sind Kriterien zur Auswahl eines statistischen Modells. Sie werden in dieser Arbeit als Maße für die Qualität eines Fits zu einer Zeitreihe verwendet. Die Idee dabei ist es, die *Maximum-Likelihood-Methode* zu verwenden, während gleichzeitig die Anzahl der für die Modellbildung nötigen Parameter möglichst gering sein soll. Ein gutes Modell darf nicht zu komplex sein und muss dabei die statistischen Eigenschaften einer vorliegenden Datenmenge trotzdem zutreffend beschreiben können.

Maximum-Likelihood-Methode

Bei der *Maximum-Likelihood-Methode* wird angenommen, dass es sich bei der vorliegenden Datenmenge um eine Realisierung X eines Zufallsexperiments handelt. Um nun diese Realisierung beschreiben zu können, sind Schätzparameter ϑ notwendig. Die Realisierung kann mit einer Wahrscheinlichkeitsdichtefunktion beschrieben werden:

$$\varrho : \Omega \to [0;1], \; X \mapsto \varrho(X|\vartheta), \tag{2.39}$$

wobei Ω den Raum aller Realisierungen bezeichne. Zu einer beobachteten Realisierung wird

$$L : \Theta \to [0;1], \; \vartheta \mapsto \varrho(X|\vartheta), \tag{2.40}$$

als *Likelihood-Funktion* definiert. Θ ist dabei der Raum aller möglichen Parameterwerte. Für eine bestimmte Parametermenge ϑ entspricht die Likelihood-Funktion der Wahrscheinlichkeit, die Realisierung X zu erhalten. Bei der Maximum-Likelihood-Methode wird die Schätzung mithilfe von $\vartheta \in \Theta$ derart durchgeführt, dass das Auftreten der beobachteten Realisierung, der vorliegenden Datenmenge, am wahrscheinlichsten wird. Als *Maximum-Likelihood-Schätzer* wird diejenige Parametermenge $\hat{\vartheta}$ bezeichnet, die die Wahrscheinlichkeit, die beobachtete Realisierung zu erhalten, maximiert.
Es werde die Menge $\{x_1, x_2, \cdots, x_n\}$ als eine Realisierung mit der Wahrscheinlichkeitsverteilung $\varrho(x)$ betrachtet. Weiterhin sei die Likelihood-Funktion als eine Klasse von Wahrscheinlichkeitsdichtefunktionen $L(x|\vartheta)$, durch die $\varrho(x)$ beschrieben werden kann, gegeben. Die durchschnittliche logarithmierte Likehood-Funktion [3] ist dann

$$\frac{1}{n}\sum_{i=1}^{n}\ln L(x_i|\vartheta). \tag{2.41}$$

Für $n \to \infty$ konvergiert Gl. (2.41) gegen

$$S(\varrho, \ln L) = \int \varrho(x)\ln L(x|\vartheta)\mathrm{d}x. \tag{2.42}$$

Die Differenz

$$S(\varrho, \varrho) - S(\varrho, \ln L) = \int \varrho^2(x)\mathrm{d}x - \int \varrho(x) \ln L(x|\vartheta)\mathrm{d}x \tag{2.43}$$

heißt *Kullback-Leibler-Divergenz* und muss für gute Fits gegen Null gehen. Dies wird erreicht, wenn $S(\varrho, \ln L)$ maximiert wird. Die Maximierung von $S(\varrho, \ln L)$ als Funktion von ϑ liefert dann die Maximum-Likelihood-Schätzer $\hat{\vartheta}$. Es bezeichne k im Folgenden die Anzahl der Maximum-Likelihood-Schätzer, d.h. in dieser Arbeit die Anzahl der Modellparameter $\mu_w, \sigma_w, \phi_i, \theta_i$.

Akaikes Informationskriterium (AIC)

Akaikes Informationskriterium (AIC) ist als

$$\mathrm{AIC}\left(\hat{\vartheta}\right) = -2 \ln L\left(\hat{\vartheta}\right) + 2k \tag{2.44}$$

definiert. Es ist dasjenige Modell zu wählen, bei dem das AIC minimal ist. Der erste Term behandelt die logarithmierte Likelihood-Funktion $\ln L\left(\hat{\vartheta}\right)$, die es zu maximieren gilt. Der zweite Term zeigt, dass wenn die Anzahl k der geschätzten Parameter größer wird, die AIC steigt. Demnach wird eine hohe Anzahl von geschätzten Parametern „bestraft“, weswegen der zweite Term auch *Strafterm* genannt wird. Der Strafterm muss eine bezüglich k streng monoton wachsende Funktion sein, damit kleine k bevorzugt werden. Für das AIC wird $2k$ gewählt.

Bayessches Informationskriterium (BIC)

Bayessches Informationskriterium (BIC) berechnet sich durch

$$\mathrm{BIC}\left(\hat{\vartheta}\right) = -2 \ln L\left(\hat{\vartheta}\right) + k \ln n, \tag{2.45}$$

wobei n die Anzahl der in der Realisierung enthaltenen Zufallsvariablen bezeichne. Der Term, der die Likelihood-Maximierung behandelt, bleibt verglichen mit dem AIC (2.44) unverändert. Der Nachteil des AIC gegenüber zum BIC liegt darin, dass die Größe des Strafterms unabhängig von der Größe n des Datensatzes ist. Beim BIC wird daher der Strafterm derart definiert, dass er logarithmisch mit n anwächst. Ab einer Größe $n = 8$ des Datensatzes bestraft das BIC zusätzliche Parameter k stärker als das AIC. Dies folgt aus der Betrachtung

$$\begin{aligned} \mathrm{BIC} &> \mathrm{AIC} \\ k \ln n &> 2k \\ n &> e^2 \approx 7.4. \end{aligned}$$

Das BIC reagiert damit empfindlicher auf eine hohe Anzahl k von geschätzten Parametern als das AIC.

2.7 Bestimmung der Ordnungsparameter p, d, q

Es wird nun beleuchtet, wie die Ordnungsparameter einer ARIMA(p, d, q)-Zeitreihe, unter anderem mithilfe der Software R [4] , bestimmt werden können.
Bevor die Ordnungen p, q der AR- und MA-Terme bestimmt werden können, muss die vorliegende Zeitreihe stationär sein. Zur Überprüfung der Stationarität eignen sich die in Abschnitt 2.3.4 erwähnten Einheitswurzeltests (PP, ADF, KPSS). Ist die vorliegende Zeitreihe noch nicht stationär, muss sie wie in Gl. (2.18) beschrieben differenziert werden. Da für die Einheitswurzeltests eine AR-Ordnung p angegeben werden muss, beginnt die Analyse mit der Schätzung von p.
Dann wird erneut auf Stationarität geprüft. Liegt diese immer noch nicht vor, muss entsprechend

	ACF	PACF
AR(p)	nimmt langsam ab	schneidet nach dem Lag p scharf ab
MA(q)	schneidet nach dem Lag q scharf ab	nimmt langsam ab
ARMA(p, q)	beides möglich	beides möglich

Tabelle 2.1: Verhalten der ACF und der PACF für ARMA-Modelle.

Gl. (2.18) weiter differenziert werden, bis die d-fach differenzierte Zeitreihe stationär ist. Die Anzahl der benötigten Differentiationen zum Erreichen der Stationarität entspricht dem Parameter d. Im `forecast`-Package von R ist die Funktion `ndiffs()` implementiert, die bei einer gegebenen Zeitreihe die Anzahl d der Differentiationen liefert, die notwendig sind, um eine Stationarität zu erreichen. Dabei kann die Art des Einheitswurzeltests ausgewählt werden. Der Parameter d sollte unabhängig von der Art des Einheitswurzeltests sein.
Ist nun die Zeitreihe stationär, können die Ordnungen p, q der AR- und MA-Terme bestimmt werden:

Wenn die PACF scharf abschneidet, wohingegen die ACF langsam abflacht, sind AR-Terme vorhanden. Der Lag τ, bei dem die PACF gegen 0 geht, entspricht der Anzahl der vorhandenen AR-Terme und damit der Ordnung p.

Schneidet dahingegen die ACF bei einem Lag τ scharf ab und geht gegen 0, so sind MA-Terme vorhanden und derjenige Lag τ, der gegen 0 geht, entspricht der Anzahl der vorhandenen MA-Terme und damit der Ordnung q.

Für alle anderen Fälle kann ein ARMA(p, q)-Prozess mit $p \geq 1$ und $q \geq 1$ vermutet werden. Beim ARMA(p, q)-Prozess können die ACF und die PACF sowohl langsam abnehmen als auch schnell gegen Null gehen; es sind je nach Größenordnung der Koeffizienten jedoch auch andere ACF- oder PACF-Charakteristiken zu verzeichnen.

Die ACF und die PACF lassen sich zwar zur Abschätzung der Parameter p, q, heranziehen, lassen jedoch keine eindeutige Bestimmung zu. Daher muss eine Zeitreihenanalyse für verschiedene vermutete Paare p, q durchgeführt werden, womit die Koeffizienten ϕ_i, θ_i berechnet werden können. Anhand des AIC und des BIC kann beurteilt werden, ob die Parameter p, q gut gewählt worden sind: Es ist das Modell zu wählen, bei dem das AIC und das BIC minimal sind. Daher sollten mehrere Paare p, q ausprobiert werden, um die Güte des Modells zu validieren und abzuschätzen, ob die Parameter p, q richtig sind. In Tab. 2.1 ist zusammenfassend dargestellt, wie die Wahl der Parameter p, q anhand der Betrachtung der ACF und der PACF abläuft.

2.8 Zusammenfassung zur Identifikation eines ARIMA-Modells

In diesem Abschnitt wird eine allgemeine Herangehensweise bei Zeitreihenanalysen beschrieben. Das Ziel ist es, bei einem gegebenen Datensatz das richtige Modell zu finden. Dabei wird heuristisch vorangegangen. Die gesamte Herangehensweise für Zeitreihenanalysen ist zudem im Flussdiagramm in Abb. 2.2 abgebildet.

1. Plot der Zeitreihe und ggf. Elimination fehlerhafter Daten.
2. Falls nötig, wird die Varianz der Datenpunkte durch Transformationen stabilisiert.
3. Plot der ACF und der PACF und Wahl möglicher Parameter p, q wie in Abschnitt 2.7 besprochen.
4. Modellierung mit den in Schritt 3 gewählten Parametern. Betrachtung der Informationskriterien (AIC/BIC) zur Suche nach besseren Fits.
5. Prüfung auf Stationarität unter Hinzunahme von Einheitswurzeltests. Differentiation, bis eine Stationarität erreicht ist. Der Parameter d wird hier gefunden.
6. Überprüfung der Residuen, die nicht korreliert sein sollten. Dazu werden die ACF und die PACF der Residuen betrachtet.
   Überprüfung, ob sich die Residuen wie ein weißes Rauschen verhalten. Dazu können Histogramme erstellt werden.
   Sind die Residuen korreliert und nicht normalverteilt, zurück zu Schritt 3. Ansonsten weiter zu Schritt 7.
7. Nutzung des Modells: Vorhersagen und Modellierung.

Alternativ kann nach Schritt 2 auch die Funktion `auto.arima()` verwendet werden. Dann wird direkt Schritt 6 durchgeführt.
Die *Residuen* r_x in Schritt 6 entsprechen dem numerisch berechneten weißen Rauschen w_t bei den ARIMA-Modellen und sind vom *Fehler*

$$\varepsilon_t := x_t - \hat{x}_t \tag{2.46}$$

zu unterscheiden, wobei $\hat{x}_t$ die Fitzeitreihe bezeichne. Bei einer Simulation können die Residuen r_x zur Validierung der Zeitreihenanalyse mit dem weißen Rauschen w_t verglichen werden. Die Differenz $r_x - w_t$ muss dann gegen Null gehen. Im Idealfall sollten die Residuen r_x und das weiße Rauschen w_t identisch sein, d.h. $r_x - w_t = 0$ für alle t. Bei empirischen Daten muss für die Residuen geprüft werden, ob deren statistischen Eigenschaften mit denen des weißen Rauschens, d.h. keine Autokorrelation und eine Normalverteilung, übereinstimmen.

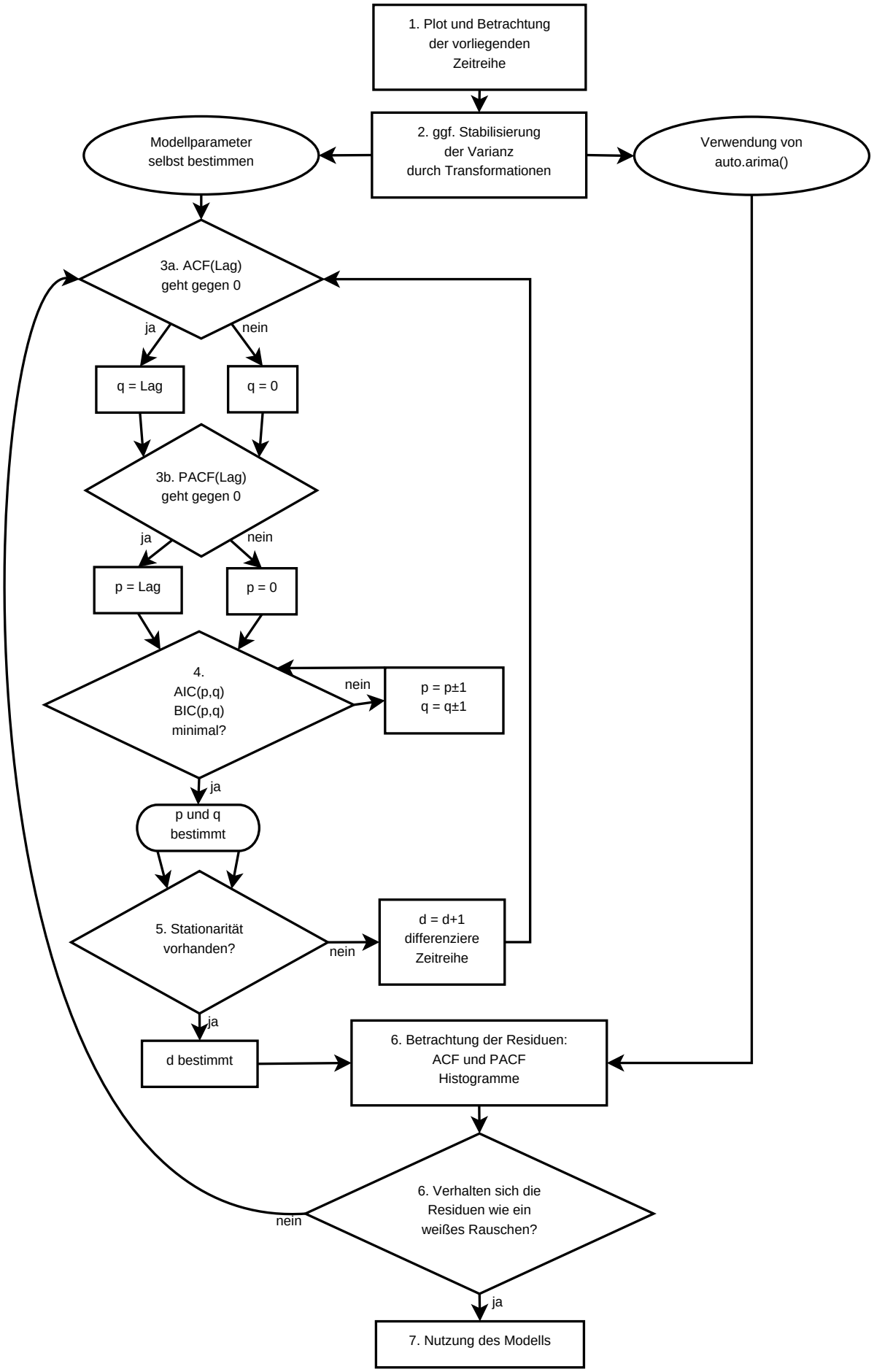


Abbildung 2.2: Herangehensweise bei Zeitreihenanalysen.

3 Beispiele zu ARIMA-Prozessen

Es werden Zeitreihen mit selbst gewählten Parametern und Koeffizienten erzeugt, um diese anschließend mithilfe von ARIMA(p, d, q)-Modellen analysieren zu können. Die für die Zeitreihenanalysen nötigen Plots sind jeweils in einer Grafik zusammengefasst. Bei allen Beispielen wird wie in Abb. 2.2 beschrieben vorgegangen.

3.1 ARIMA(1,0,1)-Prozess

Erzeugung und Visualisierung

Es soll ein ARIMA(1,0,1)-Modell in der Form

$$x_t = \phi x_{t-1} + w_t + \theta w_{t-1} \tag{3.1}$$

implementiert werden. Die Koeffizienten θ, ϕ und die Anfangsbedingungen x_0, w_0 werden selbst gewählt. Mit R sollen anschließend unter Angabe sowohl korrekter als auch falscher p, q die Koeffizienten θ, ϕ gefunden werden und mit den selbst gewählten Werten verglichen werden, um die Qualität der Zeitreihenanalyse beurteilen zu können. Die Daten werden mit einem selbst geschriebenen Programm in C erzeugt und in `*.dat`-Dateien abgespeichert. In diesen sind die Werte für x_t und w_t in ihrer Reihenfolge abgespeichert. Der Hauptalgorithmus im C-Programm ist im Folgenden dargestellt:

```
for(unsigned int t = 1; t <= N; t++)
{
    w = rand_normal(0.0, 1.0);
    x = phi*x0 + w + theta*w0;
    fprintf(datei_x, "%f    \n", x);
    fprintf(datei_w, "%f    \n", w);
    w0 = w;
    x0 = x;
}
```

Es werden $x_0 = 1$, $w_0 = 0.3$ und $\phi = 0.9$, $\theta = 0.05$ gewählt. Mit `N` $= 20000$ werden jeweils 20000 Werte erstellt; `rand_normal()` liefert gaußverteilte Zufallszahlen mit der Varianz $\sigma_w^2 = 1$ und dem Erwartungswert $\mu_w = 0$. Die erstellten Daten werden in R importiert und als Zeitreihe eingelesen. Die Zeitreihe $\{x_t\}$ und das weiße Rauschen w_t werden mit den jeweiligen Histogrammen in Abb. 3.1(a,b) gezeigt. Die Histogramme werden mit der Funktion `hist()` in R erstellt. Es lässt sich die Kovarianzmatrix für x_t und w_t berechnen. Die mit den erzeugten Daten berechnete Kovarianzmatrix lautet

$$\Sigma\left(x_t, w_t\right) = \begin{bmatrix} 5.351 & 0.926 \\ 0.926 & 0.996 \end{bmatrix} \tag{3.2}$$

Für die Berechnung der ACF ist in R die Funktion `acf()` implementiert, für die PACF wird `pacf()` verwendet. Beide sind in Abb. 3.1(c,d) zu sehen. Es ist bei der ACF zu erkennen, dass aufeinander folgende Werte x_t, $x_{t+\tau}$ mit kleinem $\tau \geq 1$ am stärksten miteinander korrelieren. Bei der PACF hingegen ist die lineare Abhängigkeit nur für einen Lag von $\tau = 1$ signifikant groß.

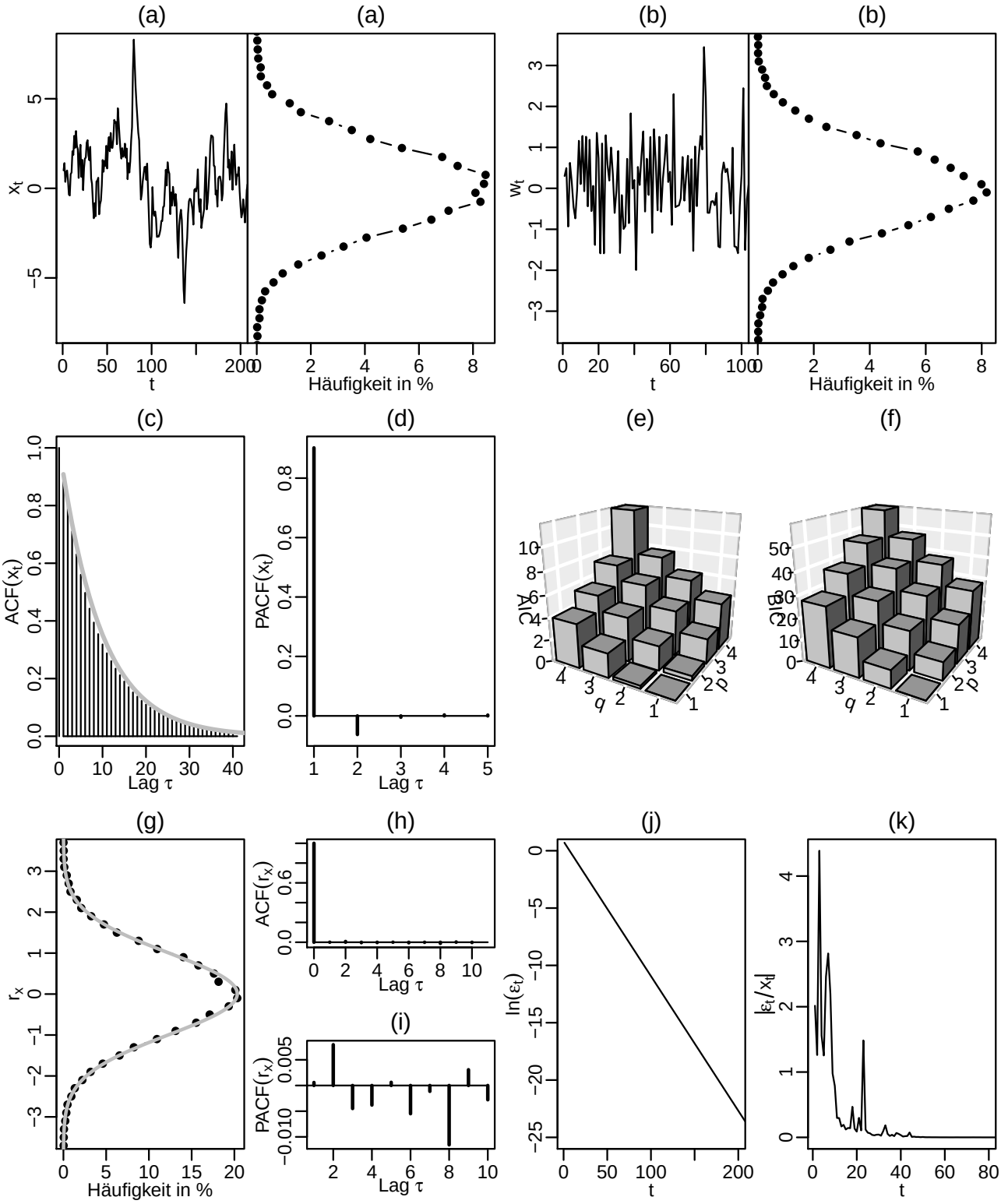


Abbildung 3.1: Zeitreihenanalyse für einen ARIMA(1,0,1)-Prozess nach Gl. (3.1). (a) Zeitreihe $\{x_t\}$ mit Histogramm: $\phi = 0.9, \theta = 0.05$. (b) Weißes Rauschen w_t mit Histogramm: $\mu_w = 4.732 \cdot 10^{-3}, \sigma_w = 0.998$. (c-d) ACF/PACF der Zeitreihe $\{x_t\}$. (e-f) AIC/BIC für verschiedene p, q und $d = 0$. Der kleinste Wert ist jeweils gleich Null. (g) Histogramm der Residuen r_x: $\mu_r = 5.182 \cdot 10^{-3}, \sigma_r = 0.998$. (h-i) ACF/PACF der Residuen r_x. (j) Logarithmierte Fehler $\ln \varepsilon_t = \ln (x_t - \hat{x}_t)$. (k) Beträge $|\varepsilon_t / x_t|$ der relativen Fehler.

(p,d,q)	ϕ_1	ϕ_2	θ_1	θ_2	AIC	BIC	σ_w^2	$\Delta\sigma^2$
(1,0,0)	0.902	-	-	-	56753.38	56777.09	0.9994	0.0034
(0,0,1)	-	-	0.751	-	73494.43	73518.14	2.308	1.312
(1,0,1)	0.889	-	0.068	-	**56678.49**	**56710.11**	0.9956	0.0004
(2,0,1)	1.024	-0.122	-0.067	-	56678.91	56718.42	0.9956	0.0004
(1,0,2)	0.887	-	0.071	0.011	56678.75	56718.27	0.9955	0.0005
(2,0,2)	0.720	0.148	0.238	0.022	56680.72	56728.14	0.9956	0.0004

Tabelle 3.1: AIC, BIC, Koeffizienten ϕ_i, θ_i und Varianz σ_w^2 verschiedener Paare p, q für eine gegebene ARIMA(1,0,1)-Zeitreihe. Alle Werte werden von R geliefert. Es ist weiterhin $\Delta\sigma^2 := |\sigma_{\text{cov}}^2 - \sigma_w^2|$ mit $\sigma_{\text{cov}}^2 = 0.996$, der in der Kovarianzmatrix (3.2) dargestellten Varianz für w_t.

Bestimmung der Ordnungsparameter p, d, q

Im R-Package `forecast` ist eine Funktion `auto.arima()` implementiert, die algorithmisch die am besten geeigneten ARIMA-Ordnungsparameter p, d, q sucht und anschließend die Koeffizienten ϕ_i und θ_i anhand der gefundenen p, d, q berechnet. Um zu verstehen, wie `auto.arima()` funktioniert, wird zuvor das Verfahren manuell durchgeführt und dann mit dem Ergebnis der Analyse mittels `auto.arima()` verglichen. Die Parameter für das ARIMA(1,0,1)-Modell sind zwar aufgrund der eigenen Implementation bereits bekannt. Nun wird jedoch angenommen, dass eine Zeitreihe ohne die Kenntnis ihrer Parameter vorliegt. Es kann mithilfe der ACF und der PACF abgeschätzt werden, welche Parameter p, q sinnvoll für die ARIMA-Analyse sind. Bevor die Parameter p, q gewählt werden, muss die vorliegende Zeitreihe stationär sein. Zur Überprüfung der Stationarität werden Einheitswurzeltests (KPSS, ADF, PP) verwendet. In der in R implementierten Funktion `ndiffs()` kann ein Einheitswurzeltest ausgewählt werden und auf die Zeitreihe angewendet werden. `ndiffs()` liefert dann die Anzahl der Differentiationen, die notwendig sind, um eine Stationarität zu erreichen. Für Gl. (3.1) wird $d = 0$ geliefert; demnach ist die vorliegende Zeitreihe bereits stationär und es kann mit der Schätzung der Parameter p, q begonnen werden. Die ACF in Abb. 3.1(c) sinkt langsam ab, was darauf schließen lässt, dass AR-Terme vorhanden sind. Die PACF in Abb. 3.1(d) zeigt deutlich, dass AR-Terme vorhanden sind. Aufgrund des Peaks für einen Lag von $\tau = 1$ und der danach gegen Null gehenden PACF ist zu vermuten, dass mindestens ein AR-Term vorhanden ist, d.h. $p \in \{0, 1, 2\}$. Anhand der ACF kann keine Aussage über die Ordnung q der MA-Terme gemacht werden. Da jedoch für ein ARMA(p, q)-Modell sowohl die ACF als auch die PACF langsam abnehmen können, muss davon ausgegangen werden, dass auch MA-Terme, d.h. $q \in \{0, 1, 2\}$, vorhanden sein können. Die präferierten Modelle sind somit ARIMA$(p, 0, q)$-Modelle mit $p \in \{0, 1, 2\}$ und $q \in \{0, 1, 2\}$.

Schätzung der Koeffizienten ϕ, θ

Es wird jeweils für verschiedene Paare p, q eine ARIMA-Analyse mithilfe der in R implementierten Funktion `Arima()` durchgeführt und das AIC/BIC in Tab. 3.1 dargestellt. Da es sich bei ARIMA$(p, 0, q)$-Modellen um stationäre Zeitreihen handelt, wird $d = 0$ gesetzt. Das AIC und das BIC müssen für die Parameter p, d, q minimal sein, wenn letztere passend gewählt worden sind. In Tab. 3.1 und in den Abb. 3.1(e,f)[1] und 3.2 ist zu erkennen, dass sowohl das AIC/BIC als auch $\Delta\sigma^2$ für die Parameter $(p, d, q) = (1, 0, 1)$ minimal sind. Dies entspricht den Erwartungen, da die für Gl. (3.1) verwendeten Parameter $p = 1$ und $q = 1$ sind. Wird $p = 0$ gesetzt und damit der AR-Term in Gl. (3.1) vernachlässigt, folgt daraus eine große Abweichung der von R geschätzten Varianz σ^2 und das AIC/BIC ist viel höher als bei denjenigen Berechnungen, die einen AR-Term, d.h. $p \geq 1$, beinhalten. Es wird vermutet, dass aufgrund $\phi > \theta$ dem AR-Term in Gl. (3.1) eine größere Gewichtung zuzuordnen ist,

[1] Es werden einige Werte des AIC und BIC nicht dargestellt, da das Minimum sonst nicht zu erkennen wäre.

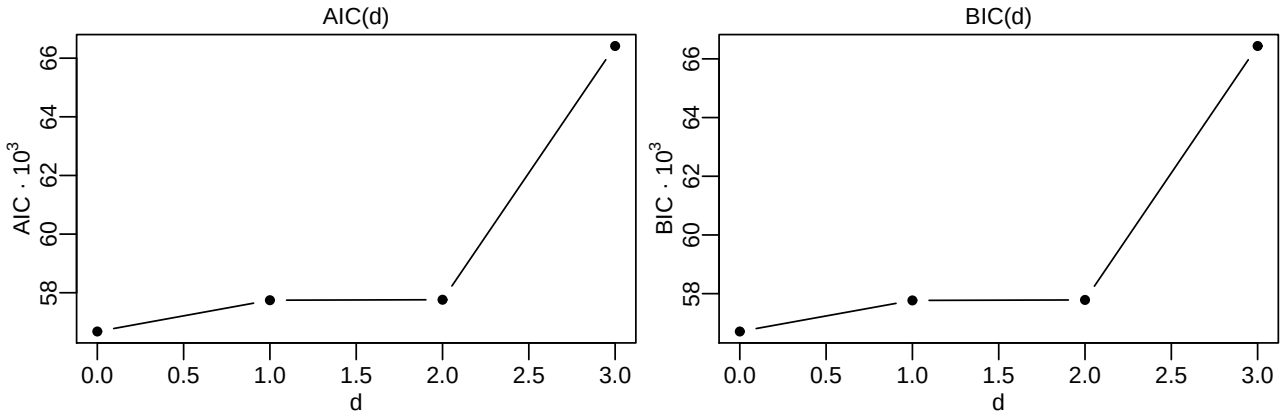


Abbildung 3.2: AIC und BIC einer ARIMA-Analyse der Zeitreihe nach Gl. 3.1 für verschiedene d. Es sind $p = 1$ und $q = 1$ fest.

sodass eine Vernachlässigung des AR-Terms zu fehlerhaften Ergebnissen führt.
Die nach Gl. (2.38) berechneten Koeffizienten $\hat{\phi}, \hat{\theta}$ betragen

$$\boxed{\begin{aligned}\hat{\phi} &= 0.889 \pm 0.004 \\ \hat{\theta} &= 0.068 \pm 0.008\end{aligned}} \tag{3.3}$$

für $(p, d, q) = (1, 0, 1)$. Die angegebenen Fehler werden von R geliefert und reichen nicht dafür aus, die berechneten Koeffizienten $\hat{\phi}, \hat{\theta}$ innerhalb der Fehlergrenzen auf die im Programmcode zu Gl. (3.1) gewählten Koeffizienten $\phi = 0.9$ und $\theta = 0.05$ zu bringen. Es ist vermutlich die Anzahl der Werte $\texttt{N} = 20000$ nicht groß genug, um eine statistisch genauere Analyse durchzuführen. Dennoch ist zu erkennen, dass es sich um die richtigen Koeffizienten handeln muss, da $\hat{\phi}$ nahe an ϕ liegt. Die Abweichungen der geschätzten Werte von den gewählten Werten betragen

$$\boxed{\Delta\phi := \left|\frac{\phi - \hat{\phi}}{\phi}\right| \hat{=} 1.2\%} \tag{3.4}$$

$$\boxed{\Delta\theta := \left|\frac{\theta - \hat{\theta}}{\theta}\right| \hat{=} 36\%.} \tag{3.5}$$

Die Abweichung $\Delta\theta$ ist damit sehr groß. Generiert man stattdessen $\texttt{N} = 4 \cdot 10^5$ Daten mit den selben Koeffizienten und Anfangsbedingungen, erhält man durch die Zeitreihenanalyse

$$\hat{\phi} = 0.899 \pm 0.001$$
$$\hat{\theta} = 0.051 \pm 0.002$$

mit den relativen Fehlern $\Delta\phi = 0.6‰$ und $\Delta\theta = 2.7\%$. Diese Abweichungen sind kleiner als in Gl. (3.5), womit gezeigt ist, dass die Genauigkeit der Analyse mit der Anzahl der Daten sinkt. Die Qualität der Fits mithilfe von R soll nun im Folgenden untersucht werden.

Validierung des geschätzten Modells

Aufgrund des kleinsten AIC und BIC für die Parameter $(p, d, q) = (1, 0, 1)$ wird nun davon ausgegangen, dass die Parameter korrekt gewählt worden sind. Zur Beurteilung der Qualität der nach Gl. (2.38) berechneten Koeffizienten $\hat{\phi}, \hat{\theta}$ werden die Residuen r_x mit dem weißen Rauschen w_t verglichen. Ein Vergleich von Abb. 3.1(g) mit Abb. 3.1(b) zeigt, dass die Residuen eine Häufigkeitsverteilung wie

das weißen Rauschen aufweisen. Die Standardabweichungen $\sigma_w = \sigma_r$ des weißen Rauschens und der Residuen stimmen überein. Die ACF und die PACF in Abb. 3.1(h,i) legen nahe, dass die Residuen unkorreliert sind, was für einen passenden Fit gewünscht ist. Es wird zusätzlich eine Zeitreihe in der Form

$$\hat{x}_t = \hat{\phi}\hat{x}_{t-1} + \hat{w}_t + \hat{\theta}\hat{w}_{t-1} \tag{3.6}$$

mit den gleichen Initialwerten x_0, w_0, wie bereits im Programmcode zu Gl. (3.1) verwendet, erzeugt. Dabei werden für das weiße Rauschen $\hat{w}_t$ die durch die ARIMA-Analyse gewonnenen Residuen r_x verwendet. Dann werden die Fehler ε_t nach Gl. (2.46) zwischen den Zeitreihen (3.1) und (3.6) gebildet. Die logarithmierten Fehler $\ln \varepsilon_t$ und die Beträge $|\varepsilon_t/x_t|$ der relativen Fehler sind in Abb. 3.1(j,k) zu sehen. In Abb. 3.1(j) ist ein exponentieller Abfall der absoluten Fehler zu verzeichnen. Für große t sind kleine Fluktuationen zu beobachten. Die Beträge der relativen Fehler fluktuieren erst und gehen für größere t gegen Null. Große relative Fehler können dadurch zustande kommen, dass die entsprechenden Werte der Zeitreihe nah bei Null liegen. Es soll kurz gezeigt werden, warum die absoluten Fehler ε_t exponentiell abfallen. Dazu werden Gl. (3.1) und Gl. (3.6) in Gl. (2.46) eingesetzt:

$$\varepsilon_t = x_t - \hat{x}_t = (\phi x_{t-1} + w_t + \theta w_{t-1}) - \left(\hat{\phi}\hat{x}_{t-1} + \hat{w}_t + \hat{\theta}\hat{w}_{t-1}\right). \tag{3.7}$$

Bei einer passenden Zeitreihenanalyse entsprechen die Schätzungen der Koeffizienten den wirklichen Werten, sodass in Gl. (3.7) $\phi = \hat{\phi}$ und $\theta = \hat{\theta}$ gesetzt werden können:

$$\varepsilon_t = \phi \underbrace{(x_{t-1} - \hat{x}_{t-1})}_{\varepsilon_{t-1}} + \theta\,(w_{t-1} - \hat{w}_{t-1}) + (w_t - \hat{w}_t)\,.$$

Mit der Schreibweise $\Delta w_t := w_t - \hat{w}_t$ folgt daraus, dass

$$\varepsilon_t = \phi\varepsilon_{t-1} + \theta\Delta w_{t-1} + \Delta w_t \tag{3.8}$$

selbst eine autoregressive Zeitreihe bildet. Bei guten Fits ist davon auszugehen, dass die Residuen $r_x = \hat{w}_t$ dem weißen Rauschen w_t entsprechen und daher Δw_t gegen Null geht bzw. für große t zu kleinen Fluktuationen führt. Setzt man in Gl. (3.8) $\Delta w_t = 0$, folgt daraus die Zeitreihe

$$\varepsilon_t = \phi\varepsilon_{t-1}, \tag{3.9}$$

die aufgrund $\phi < 1$ einen exponentiellen Abfall beschreibt. Es wird zur Überprüfung der Ergebnisse noch eine Zeitreihenanalyse mit der Funktion `auto.arima()` in R durchgeführt. Für die nach Gl. (3.1) erstellte Zeitreihe lautet die Ausgabe

```
ARIMA(1,0,1) with zero mean
Coefficients:
         ar1     ma1
      0.8892  0.0681
s.e.  0.0036  0.0077
sigma^2 estimated as 0.9956:  log likelihood=-28335.52
AIC=56678.49   AICc=56678.49   BIC=56710.11
```

Die Parameter $(p, d, q) = (1, 0, 1)$ sind von `auto.arima()` gefunden worden und entsprechen denjenigen, die in Tab. 3.1 das kleinste AIC und BIC liefern. Die berechneten Koeffizienten gleichen ebenfalls denen in Tab. 3.1. Damit kann abschließend bemerkt werden, dass die Parameter für die Zeitreihenanalyse von x_t nach Gl. (3.1) passend gewählt worden sind. Es muss jedoch erwähnt werden, dass bis auf $(p, d, q) = (0, 0, 1)$ und $(p, d, q) = (1, 0, 0)$ auch die anderen in Tab. 3.1 genannten Modelle passend sind, da deren AIC/BIC nicht stark vom minimalen AIC/BIC zum ARIMA(1,0,1)-Modell abweicht.

3.2 ARIMA(2,0,2)-Prozess

Es wird nun eine ARIMA(2,0,2)-Zeitreihe in der Form

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} \tag{3.10}$$

erzeugt. Dabei wird analog zu Abschnitt 3.1 vorgegangen. Jedoch wird die Zeitreihe nicht mit einem C-Programm, sondern in R erzeugt. Weiterhin soll die Genauigkeit der geschätzten Koeffizienten verbessert werden, indem ein größerer Datensatz erzeugt wird. Es soll die Zeitreihe (3.10) mit `n` $= 4 \cdot 10^5$ Werten erstellt werden. Dabei werden die Parameter $\phi_1 = 0.7, \phi_2 = -0.3$ und $\theta_1 = 0.4, \theta_2 = 0.2$ gewählt. Die Varianz des weißen Rauschens soll $\sigma_w^2 = 0.25$ betragen. Es werden x_t und w_t mit Histogrammen in den Abb. 3.4(a,b) gezeigt. Die Kovarianzmatrix zu x_t und w_t von Gl. (3.10) lautet

$$\Sigma\left(x_t, w_t\right) = \begin{bmatrix} 0.6778 & 0.2507 \\ 0.2507 & 0.2501 \end{bmatrix}. \tag{3.11}$$

Es werden nun stichpunktartig die Ergebnisse der Zeitreihenanalyse zusammengefasst:

- Die Funktion `ndiffs()` liefert $d = 0$, daher ist x_t stationär.
- Mit der ACF und der PACF (Abb. 3.4(c,d)) können $p \leq 3$ und $q \leq 3$ geschätzt werden.
- Die Ergebnisse der ARIMA-Analyse für verschiedene Parameter p, q sind in Tab. 3.2 dargestellt.
- Das AIC und BIC in Abb. 3.4(e,f)[2] und Abb. 3.3 zeigen je ein Minimum für $(p, d, q) = (2, 0, 2)$.
- Die nach Gl. (2.38) berechneten Koeffizienten $\hat{\phi}_i, \hat{\theta}_i$ betragen für $(p, d, q) = (2, 0, 2)$

$$\boxed{\begin{aligned} \hat{\phi}_1 &= 0.709 \pm 0.006 \\ \hat{\phi}_2 &= -0.304 \pm 0.004 \\ \hat{\theta}_1 &= 0.391 \pm 0.006 \\ \hat{\theta}_2 &= 0.196 \pm 0.003. \end{aligned}} \tag{3.12}$$

  Die von R gelieferten Fehlergrenzen reichen nur für $\hat{\phi}_2$ aus, um die eigentlich gewählten Werte für die Koeffizienten zu erreichen.
- Die Abweichungen der geschätzten Werte von den gewählten Werten betragen

$$\boxed{\begin{aligned} \Delta\phi_1 = 1.3\%, \Delta\phi_2 = 1.3\%, \\ \Delta\theta_1 = 2.3\%, \Delta\theta_2 = 2.0\%, \end{aligned}} \tag{3.13}$$

  wobei die Berechnung analog zu (3.5) erfolgt. Die Abweichungen sind verglichen mit denen von Gl. (3.5) um einiges kleiner. Es wurde für den ARIMA(2,0,2)-Prozess ein 20-fach größerer Datensatz gewählt als in Abschnitt 3.1, sodass die Schätzung der Koeffizienten genauer erfolgt.
- Der Vergleich von Abb. 3.4(g) mit Abb. 3.4(b) zeigt, dass die Häufigkeitsverteilung von w_t mit der der Residuen r_x übereinstimmt. Es ist $\sigma_w = \sigma_r = 0.5$.
- Die ACF und die PACF bezüglich der Residuen (Abb. 3.4(h,i)) zeigen, dass keine Korrelation vorhanden ist. Dies erfüllt die gewünschte Eigenschaft unabhängiger Zufallszahlen.
- Die logarithmierten Fehler $\ln \varepsilon_t$ und die Beträge $|\varepsilon_t / x_t|$ der relativen Fehler sind in Abb. 3.4(j,k) dargestellt. Die absoluten Fehler gehen für große t gegen Null. Bei den relativen Fehlern ist ein Ausreißer für $t = 4$ zu verzeichnen; alle anderen Werte sind jedoch vernachlässigbar klein. Dass

[2] Es werden einige Werte des AIC und BIC nicht dargestellt, da das Minimum sonst nicht zu erkennen wäre.

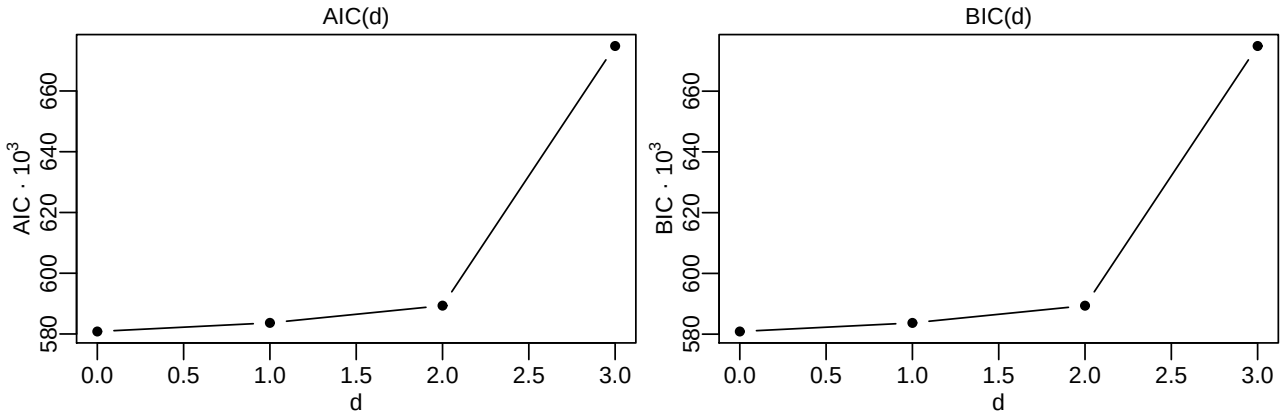


Abbildung 3.3: AIC und BIC einer ARIMA-Analyse der Zeitreihe nach Gl. 3.10 für verschiedene d. Es sind $p = 2$ und $q = 2$ fest.

(p,d,q)	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	θ_3	AIC	BIC	σ_w^2	$\Delta\sigma^2$
(1,0,0)	0.716	-	-	-	-	-	692327.5	692360.2	0.3305	0.0804
(0,0,1)	-	-	-	0.732	-	-	698170.1	698202.8	0.3354	0.0853
(1,0,1)	0.545	-	-	0.492	-	-	608235.6	608279.2	0.2679	0.0178
(2,0,1)	0.962	-0.412	-	0.132	-	-	583637.9	583692.4	0.2519	0.0018
(1,0,2)	0.300	-	-	0.794	0.350	-	586032.5	586087	0.2534	0.0033
(2,0,2)	0.708	-0.303	-	0.392	0.196	-	**580811.3**	**580876.7**	0.2501	0
(3,0,2)	0.718	-0.312	0.004	0.382	0.195	-	580813.7	580890	0.2501	0
(2,0,3)	0.710	-0.303	-	0.390	0.195	-0.001	580813.2	580889.5	0.2501	0
(3,0,3)	0.362	-0.055	-0.107	0.738	0.329	0.066	580814.9	580902.1	0.2501	0
(2,1,2)	0.962	-0.413	-	-0.868	-0.132	-	583648.5	583703	0.2519	0.0018
(2,2,2)	1.088	-0.502	-	-1.981	0.981	-	589321.6	589376.1	0.2555	0.0054

Tabelle 3.2: AIC, BIC, Koeffizienten ϕ_i, θ_i und Varianz σ_w^2 verschiedener Paare p, q für eine gegebene ARIMA(2,0,2)-Zeitreihe. Alle Werte werden von R geliefert. Es ist weiterhin $\Delta\sigma^2 := |\sigma_{\text{cov}}^2 - \sigma_w^2|$ mit $\sigma_{\text{cov}}^2 = 0.2501$, der in der Kovarianzmatrix (3.11) dargestellten Varianz für w_t.

die absoluten Fehler exponentiell sinken, kann analog zu den Gleichungen (3.7)-(3.9) erklärt werden. Hier kann der Fehler als

$$\varepsilon_t = \sum_{j=1}^{2} \phi_j \varepsilon_{t-j} + \sum_{k=0}^{2} \theta_k \Delta w_{t-k}$$

mit $\theta_0 = 1$ geschrieben werden, wobei die Terme $\sum_{k=0}^{2} \theta_k \Delta w_{t-k}$ zu einem fluktuierenden Fehler für große t führen. Aus $\phi_1 + \phi_2 < 1$ folgt ein exponentieller Abfall.

Die Zeitreihenanalyse `auto.arima()` liefert die Ausgabe

```
ARIMA(2,0,2) with zero mean
Coefficients:
         ar1      ar2     ma1     ma2
      0.7098  -0.3037  0.3904  0.1955
s.e.  0.0057   0.0035  0.0057  0.0034
sigma^2 estimated as 0.2501:  log likelihood=-290400.4
AIC=580811.3   AICc=580811.3   BIC=580876.7
```

Es entsprechen die von `auto.arima()` gelieferten Koeffizienten denen in (3.12). Damit kann davon ausgegangen werden, dass p und q richtig gewählt worden sind.

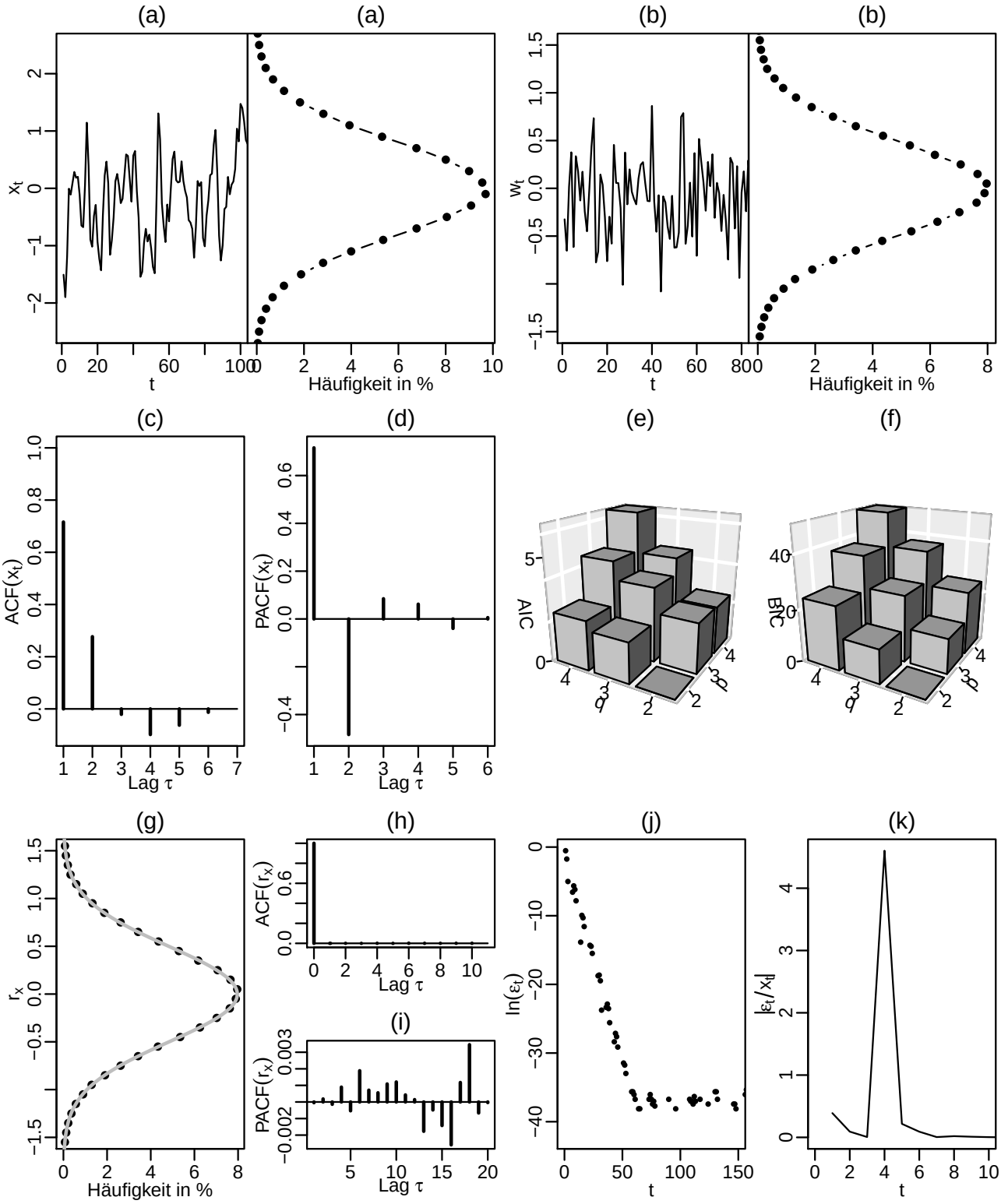


Abbildung 3.4: Zeitreihenanalyse für einen ARIMA(2,0,2)-Prozess nach Gl. (3.10). (a) Zeitreihe $\{x_t\}$ mit Histogramm: $\phi_1 = 0.7, \phi_2 = -0.3, \theta_1 = 0.4, \theta_2 = 0.2$. (b) Weißes Rauschen w_t mit Histogramm: $\mu_w = -9.375 \cdot 10^{-4}, \sigma_w = 0.5$. (c-d) ACF/PACF der Zeitreihe $\{x_t\}$. (e-f) AIC/BIC für verschiedene p, q und $d = 0$. Der kleinste Wert ist jeweils gleich Null. (g) Histogramm der Residuen r_x: $\mu_r = -9.378 \cdot 10^{-4}, \sigma_r = 0.5$. (h-i) ACF/PACF der Residuen r_x. (j) Logarithmierte Fehler $\ln \varepsilon_t = \ln (x_t - \hat{x}_t)$. (k) Beträge $|\varepsilon_t / x_t|$ der relativen Fehler.

3.3 ARIMA(0,1,0)-Prozess (Random Walk)

Ein ARIMA(0,1,0)-Prozess ist gleichbedeutend mit einem Random Walk. Dies soll genauer untersucht werden. Es wird ein Random Walk der Form

$$x_t = x_{t-1} + \underbrace{w_t + \delta}_{=:W_t} \tag{3.14}$$

simuliert. Dabei ist $W_t = w_t + \delta$ ein weißes Rauschen mit dem Erwartungswert $\mu_W = \delta$. Es wird ein C-Programm für die Erstellung der Daten verwendet. Für das weiße Rauschen W_t werden der Erwartungswert $\mu_W = \delta = 0.01$ und die Standardabweichung $\sigma_W = 1$ gewählt. Die Zeitreihe beinhaltet `N` $= 10^5$ Werte. Die Zeitreihe (3.14) und das dazugehörige weiße Rauschen sind mit Histogrammen in Abb. 3.5(a,b) dargestellt. Es ist deutlich ein Trend für die Zeitreihe in Abb. 3.5(a) zu erkennen. Die Stärke des Drifts ist von $\mu_W = \delta$ abhängig. Auch deutlich zu erkennen ist, dass die Häufigkeitsverteilung für x_t in Abb. 3.5(a) nicht normalverteilt ist. Aufgrund des Trends kann festgestellt werden, dass x_t nicht stationär ist. Der Erwartungswert ist zeitabhängig und wurde in Gl. (2.20) berechnet. Eine Zeitreihe der Form (3.14) hat eine Einheitswurzel, das charakteristische Polynom nach Gl. (2.11) lautet $m-1$ und die Varianz ist wie in Gl. (2.13) gezeigt zeitabhängig. Daraus folgt, dass der ARIMA(0,1,0)-Prozess in Form eines Random Walks mit Drift nicht stationär ist. Die Differentiation von Gl. (3.14) liefert

$$\Delta x_t = W_t = w_t + \delta, \tag{3.15}$$

was bereits in Gl. (2.21) berechnet wurde. Die Differenzen Δx_t bilden selbst ein weißes Rauschen mit dem Erwartungswert δ. Mit R werden durch die Funktion `ndiffs()` Einheitswurzeltests durchgeführt. Es wird für den Random Walk (3.14) der Wert $d = 1$ geliefert. Demnach ist der Random Walk eine Zeitreihe mit dem Parameter $d = 1$.
Um Δx_t zu erhalten, wird in R

```
rwdiff <- diff(rw, differences=1)
```

ausgeführt. Es ist nun zu prüfen, ob Δx_t stationär ist. In R wird durch `ndiffs(rwdiff)` der Wert $d = 0$ geliefert. Damit ist Δx_t stationär. Weiterhin kann Gl. (3.15) verifiziert werden. Dazu werden Δx_t und W_t in das selbe Diagramm eingetragen. Es ist in Abb. 3.5(c) zu erkennen, dass Δx_t und W_t identisch sind. Damit kann Gl. (3.15) bestätigt werden. Anhand der ACF und der PACF wird noch überprüft, ob die InkrementeΔx_t nicht korreliert sind. In den Abb. 3.5(d,e) ist keine Korrelation zu erkennen und es kann daher davon ausgegangen werden, dass es sich bei Δx_t um das weißes Rauschen w_t mit dem Erwartungswert $\mu_w = \delta$ handelt. Die Anwendung der Funktion `auto.arima()` auf x_t liefert die Ausgabe

```
ARIMA(0,1,0) with drift
Coefficients:
       drift
      0.0101
s.e.  0.0032
sigma^2 estimated as 0.9961:  log likelihood=-141698.9
AIC=283401.7   AICc=283401.7   BIC=283420.7
```

Abschließend kann damit gesagt werden, dass ein ARIMA(0,1,0)-Prozess einem Random Walk entspricht. Beide haben gleiche statistische Eigenschaften. Der Erwartungswert ist abhängig vom Drift δ und ist zeitabhängig; die Varianz wächst mit der Zeit entsprechend Gl. (2.13) an. Ein ARIMA(0,1,0)-Prozess kann jedoch einen Drift beinhalten oder auch keinen Drift; die Angabe des Parameters $d = 1$ erlaubt keine Aussage über das Vorhandensein eines Drifts. Der Drift entspricht dem Erwartungswert der differenzierten Zeitreihe Δx_t und führt zu einem deterministischen Trend, abhängig von δ.

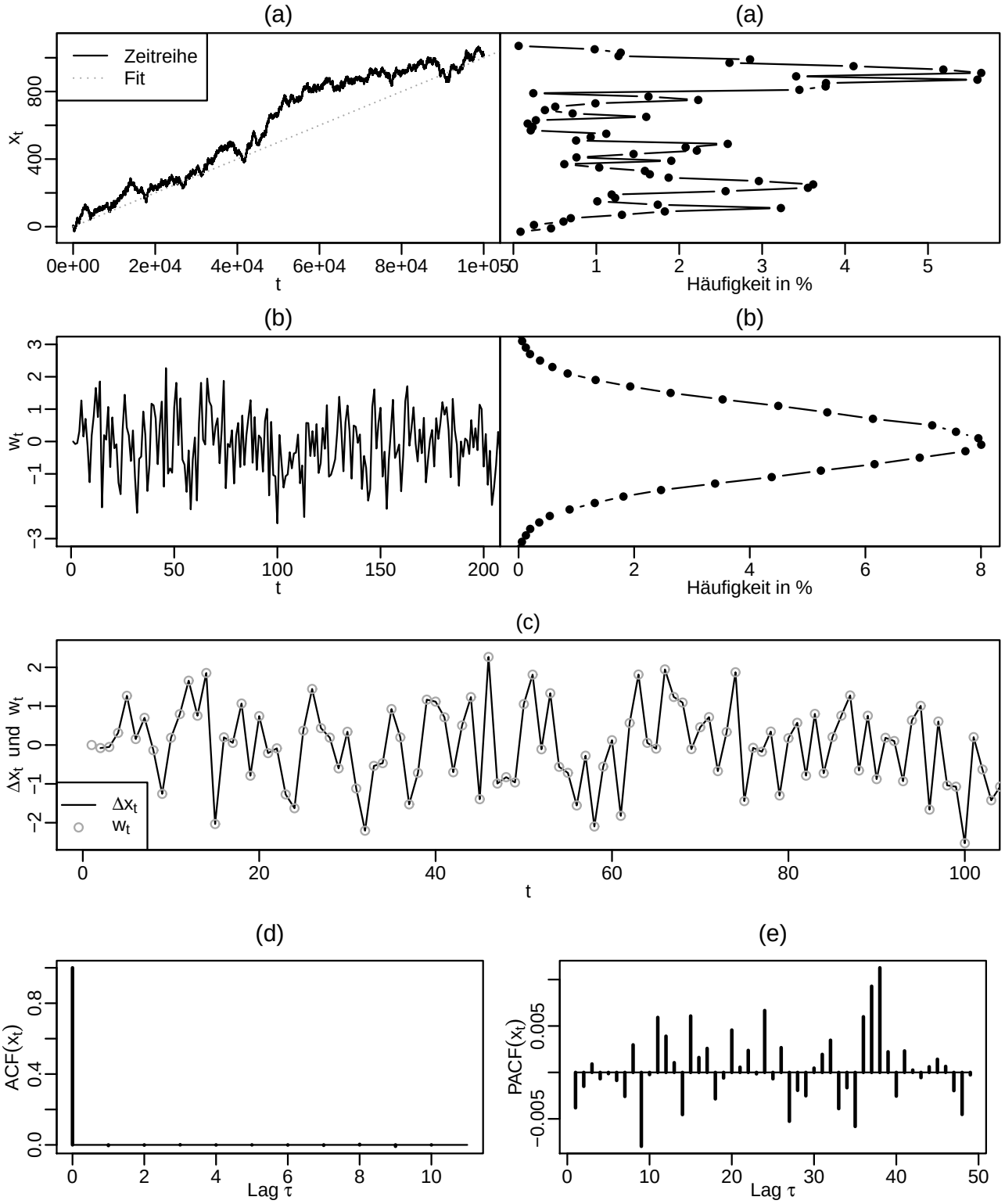


Abbildung 3.5: Zeitreihenanalyse für einen ARIMA(0,1,0)-Prozess nach Gl. (3.14) (Random Walk). (a) Zeitreihe $\{x_t\}$ mit Histogramm: $\delta = 0.01, \sigma_W = 1$. (b) Weißes Rauschen W_t mit Histogramm: $\mu_W = 1.013 \cdot 10^{-2}, \sigma_w = 0.998$. (c) Vergleich der Differenzen $\Delta x_t = x_t - x_{t-1}$ mit dem weißen Rauschen w_t. (d-e) ACF/PACF der Differenzen Δx_t bzw. des weißen Rauschens w_t.

3.4 ARIMA(2,1,2)-Prozess

Eine ARIMA(2,1,2)-Zeitreihe in der Form

$$\Delta x_t = \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} \tag{3.16}$$

soll erstellt werden. Die Zeitreihe soll mit R erzeugt werden. Der ARIMA(2,1,2)-Prozess (3.16) soll $\mathtt{n} = 4 \cdot 10^5$ Werte beinhalten. Dabei werden die Parameter $\phi_1 = 0.5, \phi_2 = -0.1$ und $\theta_1 = 0.3, \theta_2 = -0.2$ gewählt. Die Varianz des weißen Rauschens soll $\sigma_w^2 = 0.01$ betragen. Die Zeitreihe $\{x_t\}$ und das dazugehörige weiße Rauschen sind mit den jeweiligen Histogrammen in Abb. 3.6(a,b) zu sehen.

- Die Trends von x_t für gewisse Zeitintervalle in Abb. 3.6(a) weisen auf eine Nichtstationarität hin. Einheitswurzeltests mithilfe von `ndiffs()` liefern $d_x = 1$.
- Durch eine Differentiation erhält man den Prozess in der Form von Gl. (3.16), welcher in Abb. 3.6(c) zu sehen ist. Einheitswurzeltests auf Δx_t liefern $d_{\Delta x} = 0$, womit der Parameter $d_x = 1$ feststeht. Die gewohnte Zeitreihenanalyse erfolgt nun mit Δx_t.
- Mit der ACF und der PACF in Abb. 3.6(d,e) können die Parameter

  $$\begin{aligned} &q \in \{1, 2\} \text{ und} \\ &p \in \{2, 3, 4, 5\} \end{aligned} \tag{3.17}$$

  geschätzt werden.
- Eine Darstellung des AIC und des BIC für die nach Gl. (3.17) möglichen Parameterpaare p, q befindet sich in Abb. 3.6(e,f)[3]. Ein Minimum ist jeweils für $(p, d, q) = (2, 1, 2)$ zu verzeichnen.
- Die nach Gl. (2.38) berechneten Koeffizienten $\hat{\phi}_i, \hat{\theta}_i$ betragen für $(p, d, q) = (2, 1, 2)$

  $$\boxed{\begin{aligned} \hat{\phi}_1 &= 0.499 \pm 0.033 \\ \hat{\phi}_2 &= -0.101 \pm 0.005 \\ \hat{\theta}_1 &= 0.299 \pm 0.033 \\ \hat{\theta}_2 &= -0.199 \pm 0.022. \end{aligned}} \tag{3.18}$$

  Die Fehlergrenzen reichen aus, um auf die ursprünglich gewählten Werte zu kommen.
- Die Abweichungen der geschätzten Werte von den gewählten Werten betragen

  $$\boxed{\begin{aligned} &\Delta\phi_1 = 0.6‰, \Delta\phi_2 = 1.2\%, \\ &\Delta\theta_1 = 0.6‰, \Delta\theta_2 = 5.7‰, \end{aligned}} \tag{3.19}$$

  wobei die Berechnung analog zu (3.5) erfolgt. Die Abweichungen (3.19) sind noch kleiner als in den vorigen Abschnitten.
- Ein Vergleich der Verteilung der Residuen in Abb. 3.6(h) mit der des weißen Rauschens in Abb. 3.6(b) zeigt, dass beide die selbe Häufigkeitsverteilung mit $\sigma_w = \sigma_r = 0.1$ aufweisen.
- Die ACF und die PACF der Residuen in Abb. 3.6(i,j) weisen keine Korrelation auf, wie es für die Residuen gewünscht ist.
- Es wird weiterhin ein ARIMA(2,1,2)-Prozess $\hat{x}_t$ mit den Parametern (3.18) simuliert. Dabei werden die berechneten Residuen r_x von x_t als weißes Rauschen für $\hat{x}_t$ verwendet.
- Die Fehler nach Gl. (2.46) werden berechnet. Die Häufigkeitsverteilung der absoluten Fehler ist in Abb. 3.6(k), die der relativen in Abb. 3.6(l) zu sehen. Beide sind normalverteilt, zeigen jedoch einen Erwartungswert $\mu_\varepsilon > 0$. Damit beinhaltet der Fit durchschnittlich kleinere Werte als die originale Zeitreihe.

[3] Es sind nicht alle Werte dargestellt, da das Minimum sonst nicht mehr zu erkennen wäre.

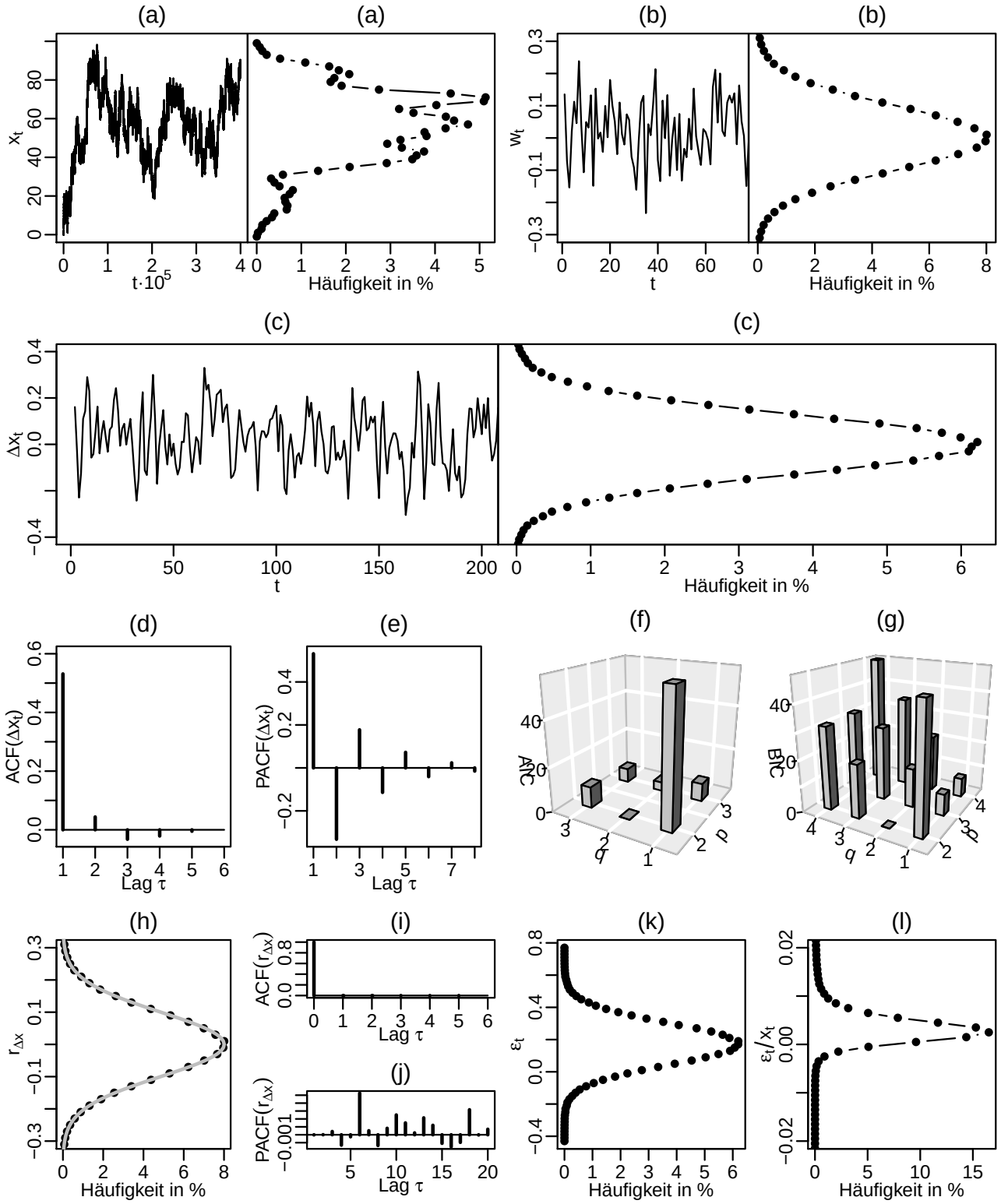


Abbildung 3.6: Zeitreihenanalyse für einen ARIMA(2,1,2)-Prozess nach Gl. (3.16). (a) Zeitreihe $\{x_t\}$ mit Histogramm: $\phi_1 = 0.5, \phi_2 = -0.1, \theta_1 = 0.3, \theta_2 = -0.2$. (b) Weißes Rauschen w_t mit Histogramm: $\mu_w = 1.115 \cdot 10^{-4}, \sigma_w = 0.1$. (c) Differenzen Δx_t mit Histogramm. (d-e) ACF/PACF der Differenzen Δx_t. (f-g) AIC/BIC für verschiedene p, q und $d = 1$. Der kleinste Wert ist jeweils gleich Null. (h) Histogramm der Residuen r_x: $\mu_r = 1.115 \cdot 10^{-4}, \sigma_r = 0.1$. (i-j) ACF/PACF der Residuen r_x. (k) Häufigkeitsverteilung der absoluten Fehler $\varepsilon_t = x_t - \hat{x}_t$: $\mu_\varepsilon > 0, \sigma_\varepsilon = 0.128$. (l) Häufigkeitsverteilung der relativen Fehler ε_t / x_t: $\mu_{\varepsilon/x} > 0, \sigma_{\varepsilon/x} = 4.621 \cdot 10^{-2}$.

Modell: (p, d, q)	$(1, 0, 1)$	$(2, 0, 2)$	$(2, 1, 2)$
Anzahl n der Daten	$2 \cdot 10^4$	$4 \cdot 10^5$	$4 \cdot 10^5$
$\phi_1, \hat{\phi}_1$	$0.9, 0.889$	$0.7, 0.709$	$0.5, 0.499$
$\phi_2, \hat{\phi}_2$	-	$-0.3, -0.304$	$-0.1, -0.101$
$\theta_1, \hat{\theta}_1$	$0.05, 0.068$	$0.4, 0.391$	$0.3, 0.299$
$\theta_2, \hat{\theta}_2$	-	$0.2, 0.196$	$-0.2, -0.199$
$\Delta\phi_1$	1.2%	1.3%	0.6‰
$\Delta\phi_2$	-	1.3%	1.2‰
$\Delta\theta_1$	36%	2.3%	0.6‰
$\Delta\theta_2$	-	2.0%	5.7‰
$\Delta\sigma = \lvert\sigma_w - \sigma_r\rvert / \sigma_w$	0.3‰	2.2 ppm	1.6 ppm
$\Delta\mu = \lvert\mu_w - \mu_r\rvert / \lvert\mu_w\rvert$	9.5%	0.3‰	0.2‰
AIC	56678.49	580811.3	−706825.7
BIC	56710.11	580876.7	−706771.2

Tabelle 3.3: Ergebnisse für die Beispiele zu ARIMA-Modellen. Die relativen Fehler $\Delta\phi_i, \Delta\theta_i$ werden wie in Gl. (3.5) berechnet.

Die automatische Zeitreihenanalyse `auto.arima()` für die ursprüngliche Zeitreihe ohne Differentiation liefert die Ausgabe

```
ARIMA(2,1,2)
Coefficients:
         ar1      ar2     ma1      ma2
      0.4987  -0.1005  0.2987  -0.1992
s.e.  0.0334   0.0048  0.0332   0.0223
sigma^2 estimated as 0.01:  log likelihood=353417.8
AIC=-706825.7   AICc=-706825.7   BIC=-706771.2
```

Die Analyse mit `auto.arima()` für die differenzierte Zeitreihe Δx_t liefert die selben Werte, lediglich ist $d = 0$. Es entsprechen die von `auto.arima(x212)` gelieferten Koeffizienten denen in (3.18). Damit müssen sind die Parameter p, d, q richtig gewählt worden sein.

3.5 Zusammenfassung der Ergebnisse für die Beispiele zu ARIMA-Modellen

Abschließend werden die Ergebnisse für die Beispiele zu den ARIMA-Modellen in Tab. 3.3 zusammenfassend dargestellt. Das Beispiel zum ARIMA(0,1,0)-Modell (Random Walk) wird nicht aufgeführt, da es keine ARIMA-Modellparameter bis auf $d = 1$ beinhaltet. Es ist anhand von Tab. 3.3 zu bemerken, dass die Abweichungen $\Delta\sigma$ und $\Delta\mu$ mit der Anzahl n der Daten sinken. Eine größere Anzahl von Daten führt zu genaueren Fits.

4 Anwendung der ARIMA-Modelle auf Windgeschwindigkeiten

Die ARIMA-Modelle werden zur Analyse und Vorhersage von Windgeschwindigkeiten verwendet. Dabei sollen die Windgeschwindigkeiten, die bei einer Frequenz von 1 Hz gemessen wurden, transformiert und standardisiert werden [5]. Die Messungen erstrecken sich über einen Zeitraum von 20 Monaten (Sep. 2015 bis Apr. 2017). Alle hier verwendeten Windgeschwindigkeiten wurden mithilfe eines Schalenanemometers bei einer Höhe von 100 m gemessen. Die Daten stammen von der Forschungsplattform FINO1 [6], welche sich etwa 45 km nördlich von Borkum befindet. Die Vorgehensweise wird im Folgenden zusammen mit den Ergebnissen für die vorliegenden Daten erläutert.

4.1 Transformation und Standardisierung der Windgeschwindigkeiten

Für jeden der 20 Monate wird eine Transformation und Standardisierung der Windgeschwindigkeiten separat durchgeführt. Von Interesse sind die stündlichen Mittelwerte, weswegen zuerst die arithmetischen Mittel über je eine Stunde gebildet werden. Es bezeichnen $v(t)$ die stündlichen Windgeschwindigkeiten, welche in Abb. 4.1 für jeden Monat gezeigt werden. Es wird dann eine Transformation [7] der Windgeschwindigkeiten, die der Weibull-Verteilung

$$f_{g,b}(v) = \frac{g}{b}\left(\frac{v}{b}\right)^{g-1} \exp\left(-(v/b)^g\right) \tag{4.1}$$

folgen, durchgeführt:

$$u(t) = \sqrt{v(t)}. \tag{4.2}$$

Die transformierten Windgeschwindigkeiten $u(t)$ nach Gl. (4.2) können nun durch eine Gaußverteilung genähert werden. In dieser Arbeit ist das weiße Rauschen w_t, das den berechneten Residuen entspricht, stets gaußverteilt. Eine Linearkombination gaußverteilter Zufallszahlen w_t folgt wieder einer Gaußverteilung, von der bei den in Kap. 2 besprochenen Modellen ausgegangen wird. Daher ist es wünschenswert, dass die transformierten Daten über einen langen Zeitraum einer Gaußverteilung folgen. Es werden in Abb. 4.2 die Häufigkeitsverteilungen von $v(t)$ und $u(t)$ gezeigt. Dann werden die stündlichen Erwartungswerte

$$\mu(t) = \frac{1}{T}\sum_{d=0}^{T-1} u\left(t + 24d\right), \quad t = 1, 2, \ldots, 24 \tag{4.3}$$

und Standardabweichungen

$$\sigma(t) = \sqrt{\frac{1}{T}\sum_{d=0}^{T-1} \left(u\left(t + 24d\right) - \mu(t)\right)^2}, \quad t = 1, 2, \ldots, 24 \tag{4.4}$$

für jeden Monat gebildet. Dabei bezeichnet T die Anzahl der Tage des entsprechenden Monats. Es wird eine 24-stündige Periodizität angenommen. Um eine Tendenz über 20 Monate erkennen zu können,

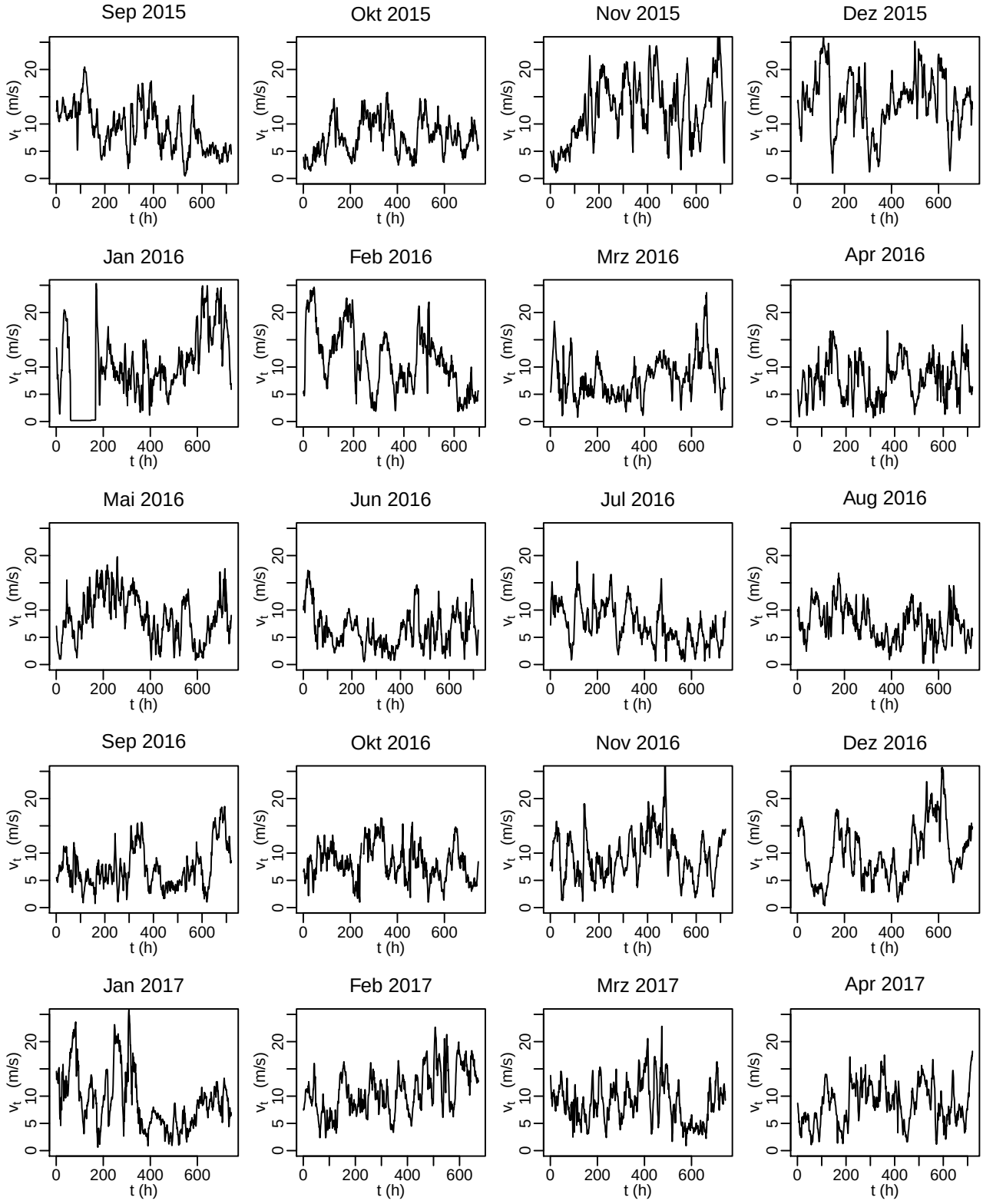


Abbildung 4.1: Stündliche Windgeschwindigkeiten $v(t)$ von 20 Monaten (Sep. 2015 bis Apr. 2017).

wird jeweils für jeden Monat der monatliche Mittelwert

$$\bar{\mu} = \frac{1}{24} \sum_{t=1}^{24} \mu(t) \tag{4.5}$$

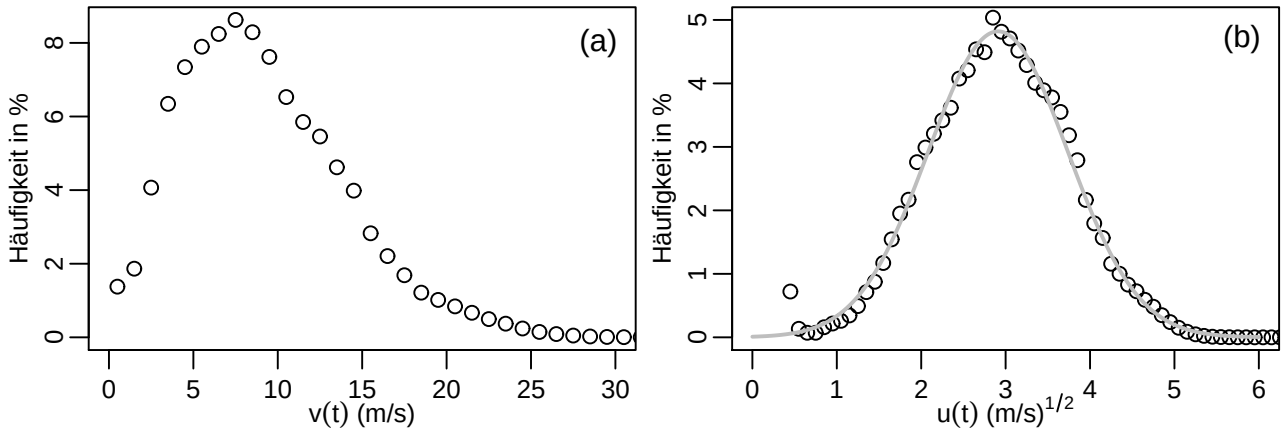


Abbildung 4.2: Häufigkeitsverteilungen über 20 Monate: (a) Windgeschwindigkeiten $v(t)$. (b) Nach Gl. (4.2) transformierte Windgeschwindigkeiten $u(t)$ mit $\mu_u = 2.91\ \ (\mathrm{m/s})^{1/2}$ und $\sigma_u = 0.83\ \ (\mathrm{m/s})^{1/2}$.

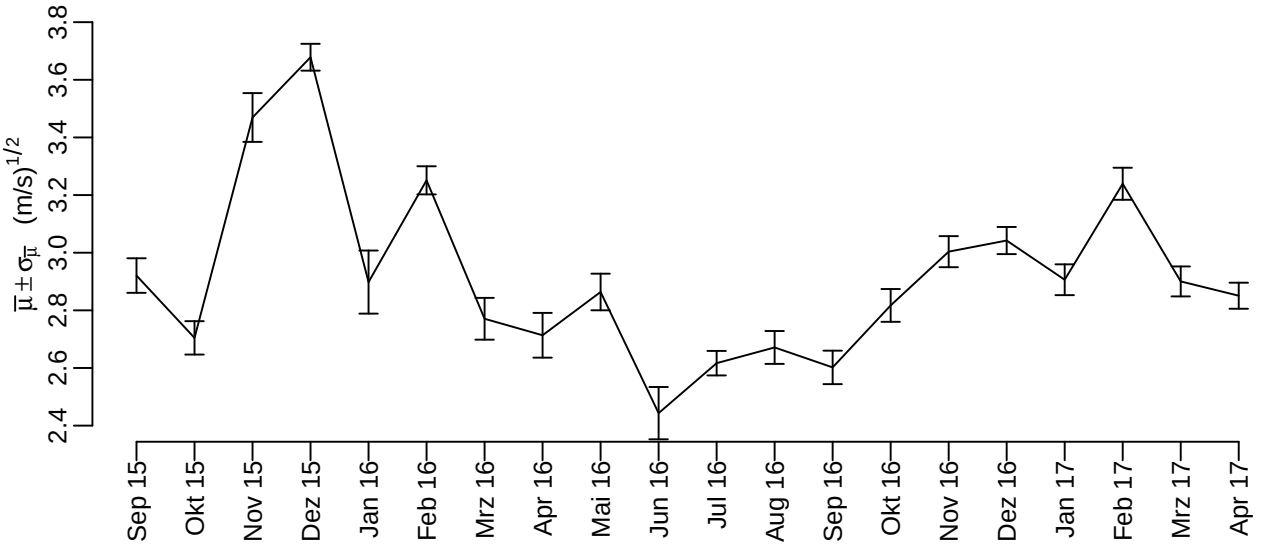


Abbildung 4.3: Monatliche Mittelwerte $\bar{\mu}$ mit den dazugehörigen Standardabweichungen $\sigma_{\bar{\mu}}$ der transformierten Windgeschwindigkeiten $u(t)$.

und die dazugehörige Standardabweichung

$$\sigma_{\bar{\mu}} = \sqrt{\frac{1}{24}\sum_{t=1}^{24}\left(\mu(t) - \bar{\mu}\right)^2} \tag{4.6}$$

berechnet. In Abb. 4.3 werden die monatlichen Mittelwerte $\bar{\mu}$ mit den Standardabweichungen $\sigma_{\bar{\mu}}$ abgebildet. Es sind in den Frühlingsmonaten tendenziell geringere Windgeschwindigkeiten zu verzeichnen. Insgesamt ist eine Variabilität bezüglich der monatlichen Mittelwerte $\bar{\mu}$ zu erkennen. Die stündlichen Erwartungswerte $\mu(t)$ schwanken in einem Monat um einen konstanten Wert, was in Abb. 4.3 daran erkannt werden kann, dass die Standardabweichungen $\sigma_{\bar{\mu}}$ klein im Vergleich zu den monatlichen Mittelwerten $\bar{\mu}$ sind. Mit μ und σ erhält man durch

$$u'(t) = \frac{u(t) - \mu(t)}{\sigma(t)} \tag{4.7}$$

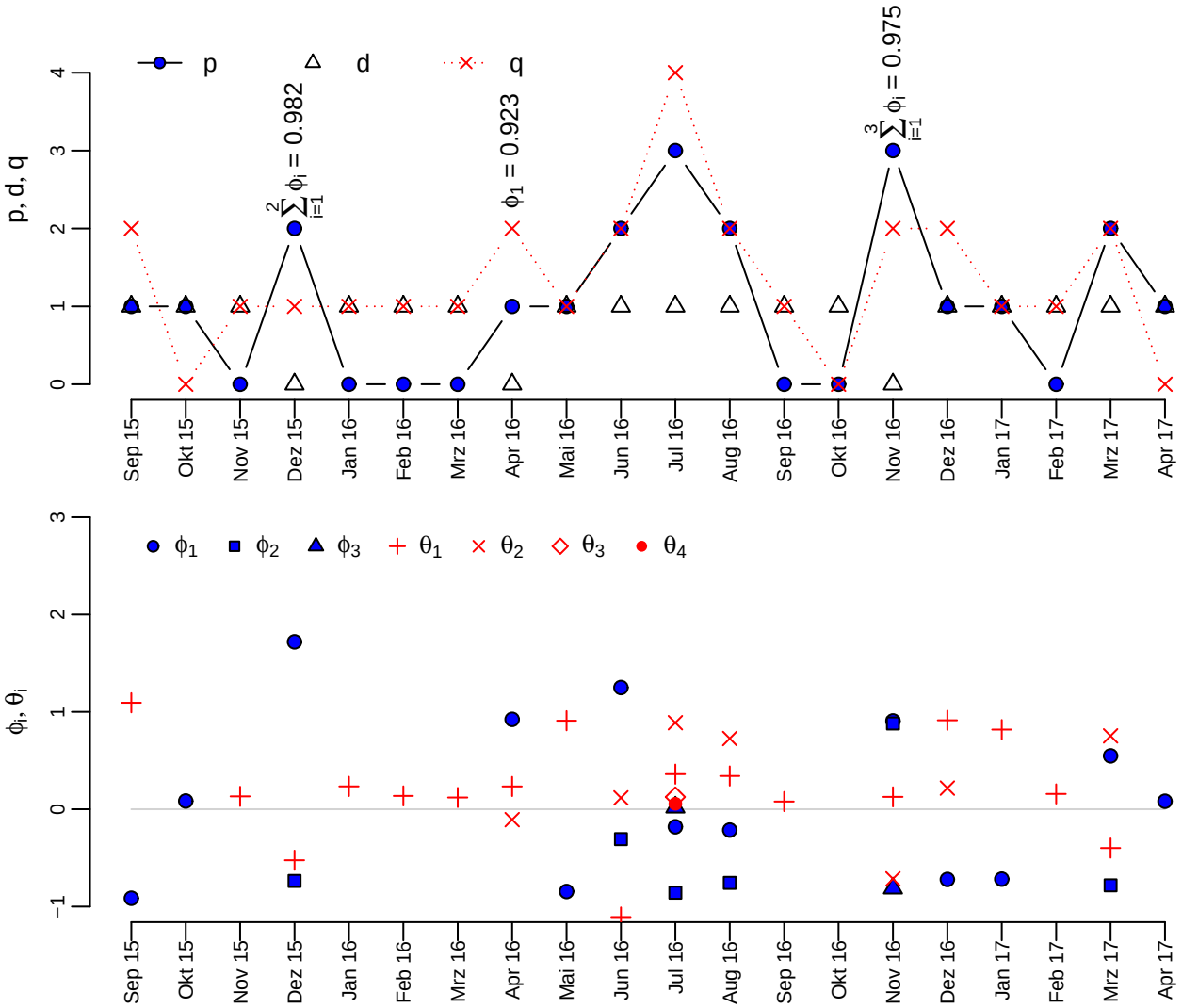


Abbildung 4.4: ARIMA-Parameter p, d, q und Koeffizienten ϕ_i, θ_i mit dem für den jeweiligen Monat geringsten BIC.

die standardisierten Windgeschwindigkeiten $u'(t)$ für je einen Monat. Die standardisierten Windgeschwindigkeiten $u'(t)$ zeigen dann über den jeweiligen Monat ein arithmetisches Mittel $\mu' \approx 0$ und eine Standardabweichung $\sigma' \approx 1$. Mit den transformierten und standardisierten Windgeschwindigkeiten $u'(t)$ wird dann die Zeitreihenanalyse durchgeführt.

4.2 Zeitreihenanalyse für jeden Monat

Für jeden Monat wird eine Zeitreihenanalyse nach dem in Abb. 2.2 besprochenen Verfahren durchgeführt. Es wird in Abb. 4.4 gezeigt, welche Parameter p, d, q für die jeweiligen Monate das minimale BIC aufweisen. Weiterhin sind dort die Koeffizienten ϕ_i, θ_i dargestellt. Es gibt kein ARIMA-Modell, welches für alle Monate das geringste BIC aufweist. Bis auf drei Monate ist jedoch $d = 1$ als Parameter

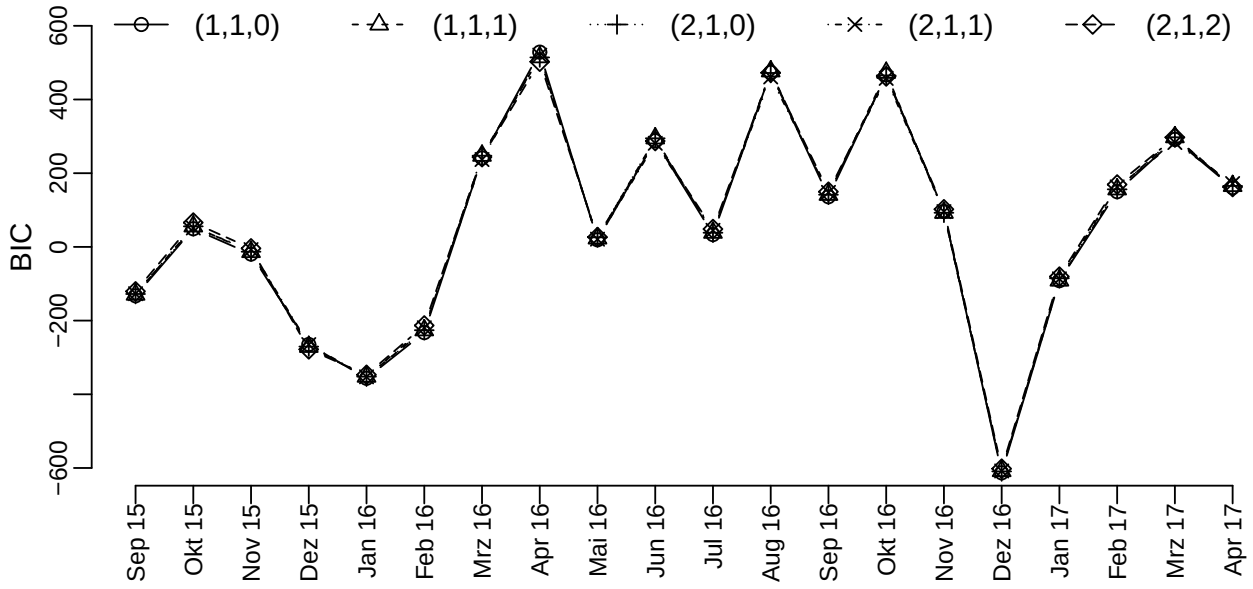


Abbildung 4.5: BIC für alle Monate für die in Gl. (4.8) genannten Parameterpaare.

zu verzeichnen. Diejenigen drei Monate, für die $d = 0$ gefunden wird, zeigen jedoch die Tendenz

$$\sum_{i=1}^{p} \phi_i \to 1,$$

was dem Grenzfall eines nichtstationären Prozesses nach Gl. (2.14) entspricht. Daher wird im nächsten Abschnitt für alle Monate $d = 1$ gewählt.

4.3 Wahl eines ARIMA-Modells für alle Monate

Es wird versucht, ein passendes ARIMA(p, d, q)-Modell für alle Monate zu finden. Die Parameterpaare, die dafür in Frage kommen, sind

$$(p, d, q) \in \{(1, 1, 0), (1, 1, 1), (2, 1, 0), (2, 1, 1), (2, 1, 2)\}\,. \tag{4.8}$$

In Abb. 4.5 wird das BIC für die in Gl. (4.8) erwähnten Modelle für alle Monate gezeigt. Es ist in Abb. 4.5 auffällig, dass sich das BIC für verschiedene ARIMA-Modelle in einem Monat kaum unterscheidet. Dies ist ein Hinweis darauf, dass mehrere ARIMA-Modelle für je einen Monat passend sein könnten. Dass sich das BIC je nach Monat ändert, liegt an unterschiedlichen Maxima der Likelihood-Funktion. Es sei $\mathrm{BIC}_{\min}$ dasjenige BIC, das für einen Monat minimal ist. Weiterhin bezeichne $\mathrm{BIC}_{(p,d,q)}$ das BIC zu den entsprechend Gl. (4.8) gewählten Parameterpaaren (p, d, q). Bei der Wahl eines ARIMA(p, d, q)-Modells für alle Monate ist es von Interesse, ein ARIMA(p, d, q)-Modell zu wählen, bei dem die durchschnittliche Abweichung von $\mathrm{BIC}_{(p,d,q)}$ zu $\mathrm{BIC}_{\min}$ klein ist. Aus diesem Grund wird die relative quadratische Abweichung

$$\Delta\mathrm{BIC}^2_{(p,d,q)} := \frac{\left(\mathrm{BIC}_{\min} - \mathrm{BIC}_{(p,d,q)}\right)^2}{\mathrm{BIC}^2_{\min}} \tag{4.9}$$

für jeden Monat berechnet, wobei die Parameterpaare (4.8) verwendet werden. Es sind diejenigen ARIMA(p, d, q)-Modelle zu bevorzugen, bei denen der Mittelwert $\overline{\Delta\mathrm{BIC}^2_{(p,d,q)}}$ der relativen quadratischen Abweichungen über alle Monate klein ist. Die Ergebnisse sind in Abb. 4.6 zu sehen. Der größte

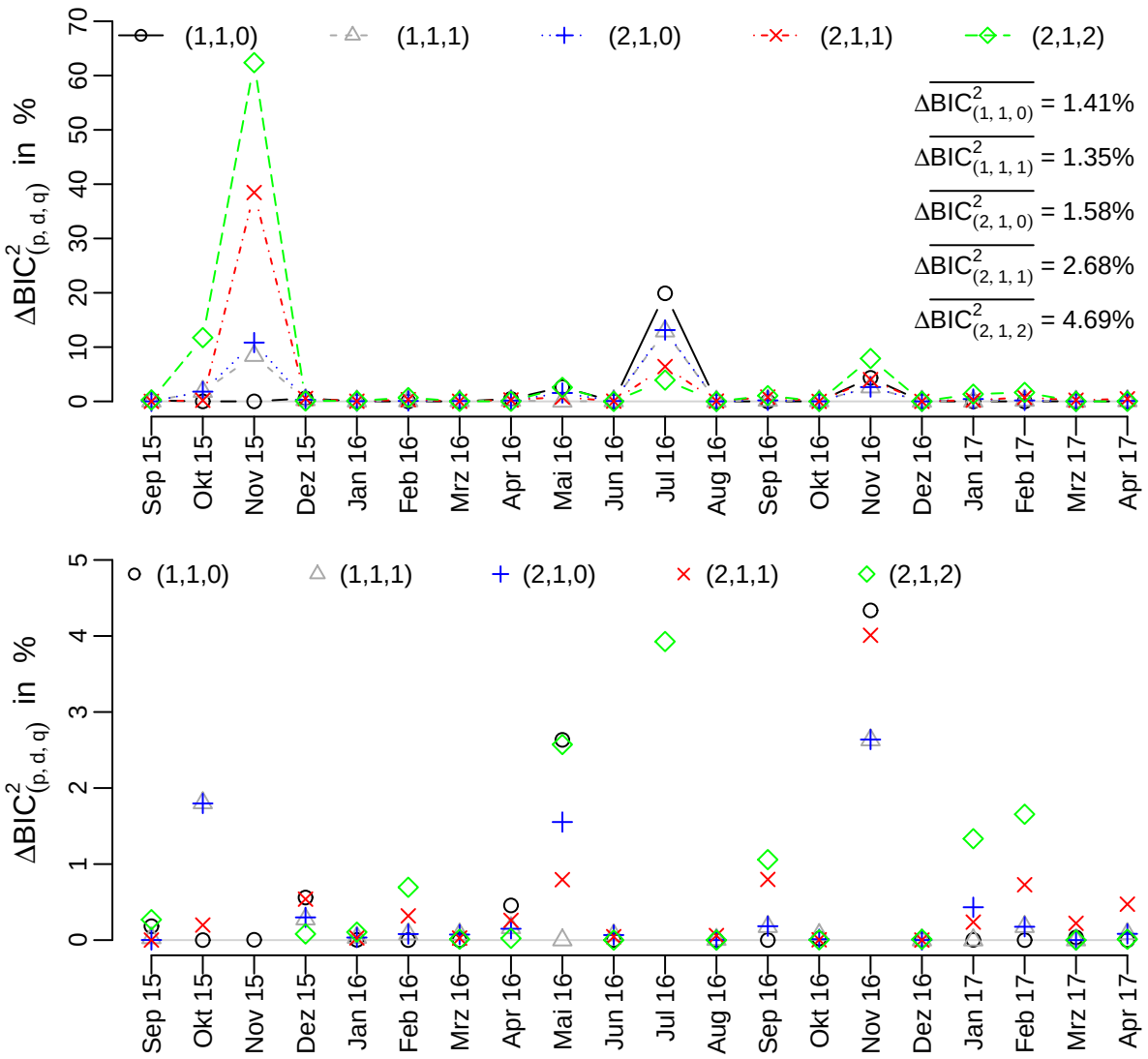


Abbildung 4.6: Relative Abweichungen nach Gl. (4.9) und deren Mittelwerte über alle Monate. Im unteren Diagramm sind nur Abweichungen bis zu 5% eingetragen, um Unterschiede sichtbar zu machen.

Teil der relativen quadratischen Abweichungen $\Delta\mathrm{BIC}^2_{(p,d,q)}$ liegt unter 2%. Dies ist aus Abb. 4.7 ersichtlich. Als mögliche Parameterpaare werden diejenigen (p, d, q) gewählt, für die die mittlere relative quadratische Abweichung $\overline{\Delta\mathrm{BIC}^2_{(p,d,q)}}$ kleiner als 2% ist. Dazu gehören

$$(p, d, q) \in \{(1, 1, 0), (1, 1, 1), (2, 1, 0)\}\,. \tag{4.10}$$

Im Folgenden werden die ARIMA-Koeffizienten ϕ_i, θ_i und die Fehler der Voraussagen für die in Gl. (4.10) genannten Fälle behandelt. Dabei handelt es sich noch nicht um die eigentlichen Voraussagen, da noch mit transformierten Daten gearbeitet wird.
Es wird für alle Monate eine ARIMA(1,1,0)-Analyse durchgeführt. Das Modell lautet

$$\begin{aligned} \Delta u'_t = u'_t - u'_{t-1} &= \phi \Delta u'_{t-1} + w_t \quad \text{bzw.} \\ u'_t &= (1+\phi)\, u'_{t-1} - \phi u'_{t-2} + w_t. \end{aligned} \tag{4.11}$$

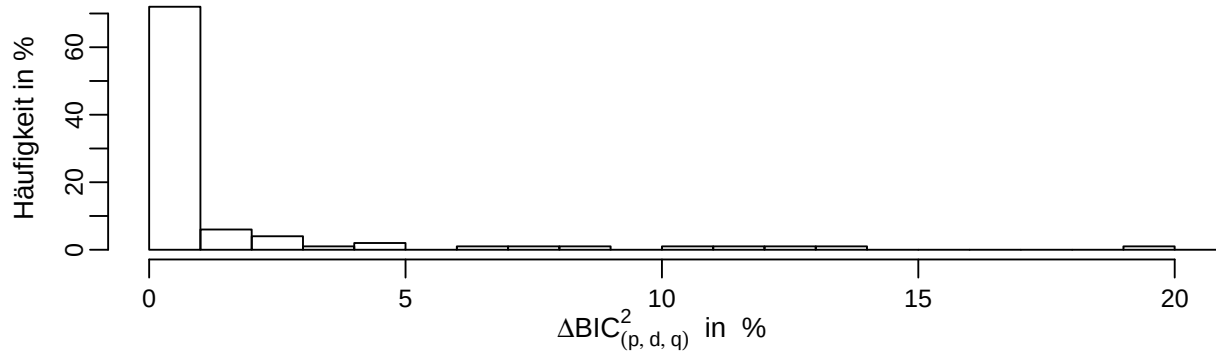


Abbildung 4.7: Häufigkeitsverteilung der relativen quadratischen Abweichungen nach Gl. (4.9).

Die monatlichen Werte für die Schätzung $\hat{\phi}$ sind in Abb. 4.8(a) dargestellt. Für alle Monate ist $\hat{\phi} > 0$. Die monatlichen Mittelwerte μ_ε der Fehler ε_t nach Gl. (2.46) sind auch zu sehen. Sie berechnen sich zu

$$\mu_\varepsilon = \frac{1}{24T-2}\sum_{t=3}^{24T}\varepsilon_t = \frac{1}{24T-2}\sum_{t=3}^{24T}u'_t - \underbrace{\left(\left(1+\hat{\phi}\right)u'_{t-1} - \hat{\phi}u'_{t-2}\right)}_{\hat{u}'_t}, \tag{4.12}$$

wobei T die Anzahl der Tage im jeweiligen Monat bezeichne. Die Fehler haben hier den Mittelwert $\mu_\varepsilon \approx 0\,(\mathrm{m/s})^{1/2}$. Die Standardabweichungen sind durch

$$\sigma_\varepsilon = \sqrt{\frac{1}{24T-2}\sum_{t=3}^{24T}\left(\varepsilon_t - \mu_\varepsilon\right)^2} \tag{4.13}$$

gegeben und weisen hier eine Schwankung auf.
Das ARIMA(1,1,1)-Modell lautet

$$\begin{aligned}\Delta u'_t = u'_t - u'_{t-1} &= \phi\Delta u'_{t-1} + \theta w_{t-1} + w_t \quad \text{bzw.}\\ u'_t &= (1+\phi)\,u'_{t-1} - \phi u'_{t-2} + \theta w_{t-1} + w_t.\end{aligned} \tag{4.14}$$

Die monatlichen Werte für $\hat{\phi}$ und $\hat{\theta}$ sind in Abb. 4.8(b) dargestellt. Es fällt auf, dass der Koeffizient $\hat{\phi}$ in einem Monat groß ist, wenn $\hat{\theta}$ klein ist und umgekehrt. Die monatlichen Mittelwerte der Fehler sind durch

$$\mu_\varepsilon = \frac{1}{24T-2}\sum_{t=3}^{24T}\varepsilon_t = \frac{1}{24T-2}\sum_{t=3}^{24T}u'_t - \underbrace{\left(\left(1+\hat{\phi}\right)u'_{t-1} - \hat{\phi}u'_{t-2} + \theta\,\overbrace{\hat{w}_{t-1}}^{r_x(t-1)}\right)}_{\hat{u}'_t} \tag{4.15}$$

gegeben. Für das weiße Rauschen w_t werden die Residuen $r_x(t)$ verwendet. Die Standardabweichungen werden analog zu Gl. (4.13) berechnet. Wie beim ARIMA(1,1,0)-Modell haben die Fehler den Mittelwert $\mu_\varepsilon \approx 0\,(\mathrm{m/s})^{1/2}$. Beim Vergleich der Standardabweichungen σ_ε in Abb. 4.8(b) mit denen in Abb. 4.8(a) (ARIMA(1,1,0)) fällt auf, dass sie sich ähnlich verhalten.
Das ARIMA(2,1,0)-Modell kann durch

$$\begin{aligned}\Delta u'_t = u'_t - u'_{t-1} &= \phi_1\Delta u'_{t-1} + \phi_2\Delta u'_{t-2} + w_t \quad \text{bzw.}\\ u'_t &= (1+\phi_1)\,u'_{t-1} + (\phi_2-\phi_1)\,u'_{t-2} - \phi_2 u'_{t-3} + w_t\end{aligned} \tag{4.16}$$

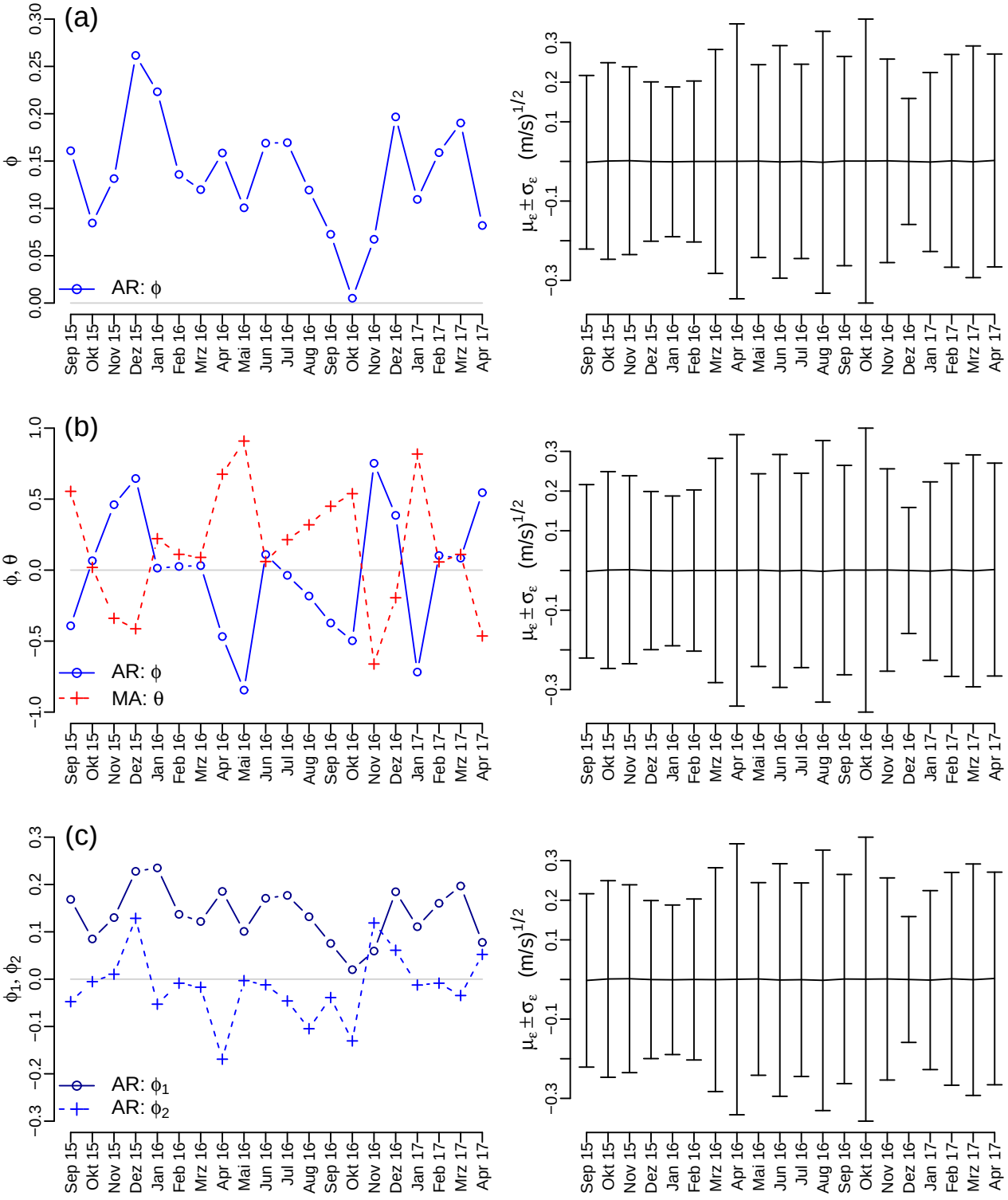


Abbildung 4.8: ARIMA-Analyse für Windgeschwindigkeiten: Parameter ϕ_i, θ_i und mittlere Fehler μ_ε der Fits. (a) $(p, d, q) = (1, 1, 0)$. (b) $(p, d, q) = (1, 1, 1)$. (c) $(p, d, q) = (2, 1, 0)$.

beschrieben werden. Die monatlichen Werte für $\hat{\phi}_1$ und $\hat{\phi}_2$ sind in Abb. 4.8(c) dargestellt. Beim Vergleich des Koeffizienten $\hat{\phi}_1$ mit dem Koeffizienten $\hat{\phi}$ in Abb. 4.8(a) (ARIMA(1,1,0)) fällt auf, dass sich die monatlichen Verläufe stark ähneln. Es handelt sich um die gleichen AR-Terme, wobei beim ARIMA(2,1,0)-Modell noch ein zusätzlicher AR-Term vorhanden ist. Die Fehler erhält man mit

$$\mu_\varepsilon = \frac{1}{24T-3}\sum_{t=4}^{24T}\varepsilon_t = \frac{1}{24T-3}\sum_{t=4}^{24T}u'_t - \underbrace{\left(\left(1+\hat{\phi}_1\right)u'_{t-1} + \left(\hat{\phi}_2-\hat{\phi}_1\right)u'_{t-2} - \hat{\phi}_2 u'_{t-3}\right)}_{\hat{u}'_t}. \tag{4.17}$$

Die Standardabweichungen σ_ε in Abb. 4.8(c) verhalten sich ähnlich verglichen mit denen in Abb. 4.8(a) (ARIMA(1,1,0)) und Abb. 4.8(b) (ARIMA(1,1,1)). Sie werden durch

$$\sigma_\varepsilon = \sqrt{\frac{1}{24T-3}\sum_{t=4}^{24T}\left(\varepsilon_t-\mu_\varepsilon\right)^2} \tag{4.18}$$

berechnet.
Es kann bezüglich der Fehler kein großer Unterschied zwischen den Modellen erkannt werden, weswegen es noch nicht möglich ist, eines der Modelle für die Voraussagen auszuwählen. Alle Modelle weisen einen Fehler von $\mu_\varepsilon \approx 0$ auf, was zeigt, dass kein Offset vorhanden ist. Setzt man in Gl. (4.13) und Gl. (4.18) $\mu_\varepsilon = 0$, erkennt man, dass die Standardabweichung im Wesentlichen den mittleren Betrag der Fehler beschreibt. Es ist in einigen Monaten, wie z.B. April 2016, zu erkennen, dass die Standardabweichung größer ist als in den anderen Monaten. Dies bedeutet, dass die Fehler der Vorhersagen aller gewählten Modelle in diesem Monat größer sind als in anderen Monaten. Einige Monate können schlechter angepasst werden als andere. Es wird im Folgenden die Qualität der Vorhersagen zu den gemessenen Windgeschwindigkeiten geprüft, um eine Aussage darüber machen zu können, ob die drei Modelle geeignet für eine Vorhersage sind. Zudem soll herausgefunden werden, ob sich eines der drei Modelle besonders für Voraussagen eignet.

4.4 Vorhersagen für Windgeschwindigkeiten

Für die in Gl. (4.10) genannten Modelle soll eine Vorhersage gemacht werden. Dabei erfolgt die Vorhersage für je eine Stunde voraus, wobei die Windgeschwindigkeiten der vorigen Stunden als Messdaten vorliegen. Es konnte von Grigonytė et al. gezeigt werden, dass die Ungenauigkeit von Voraussagen steigt, je weiter in die Zukunft vorausgesagt wird [8]. Daher erfolgt hier lediglich eine Voraussage für eine Stunde voraus. Die Vorhersagen für je eine Stunde voraus sind durch die Terme $\hat{u}'_t$ in Gl. (4.12), (4.15) und (4.17) gegeben. Zur Kontrolle, ob sich die Modelle wirklich für Vorhersagen eignen, wird auch eine Vorhersage für ein ARIMA(0,1,0)-Modell der Form

$$\hat{u}'_t = u'_{t-1} \tag{4.19}$$

durchgeführt. Die Vorhersagen $\hat{u}'_t$ beziehen sich jedoch auf die transformierten Daten, weswegen eine Rücktransformation erfolgen muss. Durch Umstellen von Gl. (4.2) und Gl. (4.7) erhält man die Rücktransformationen

$$\hat{v}(t) = \left[\hat{u}'(t)\cdot\sigma(t \bmod 24) + \mu(t \bmod 24)\right]^2, \tag{4.20}$$

die die eigentlichen Vorhersagen $\hat{v}(t)$ darstellen. Es werden in Abb. 4.9 für die vier Modelle jeweils die gemessenen Windgeschwindigkeiten $v(t)$ und deren Vorhersagen $\hat{v}(t)$ in einem Zeitraum von 3 Tagen gezeigt. Alle vier Beispiele zeigen, dass die großen Fluktuationen der Windgeschwindigkeiten durch die Voraussagen erhalten werden können. Dabei ist zu erkennen, dass sich die Vorhersagen im Wesentlichen wie die originale Zeitreihe, jedoch um einen Lag $\tau = 1$ verschoben, verhalten. Um die

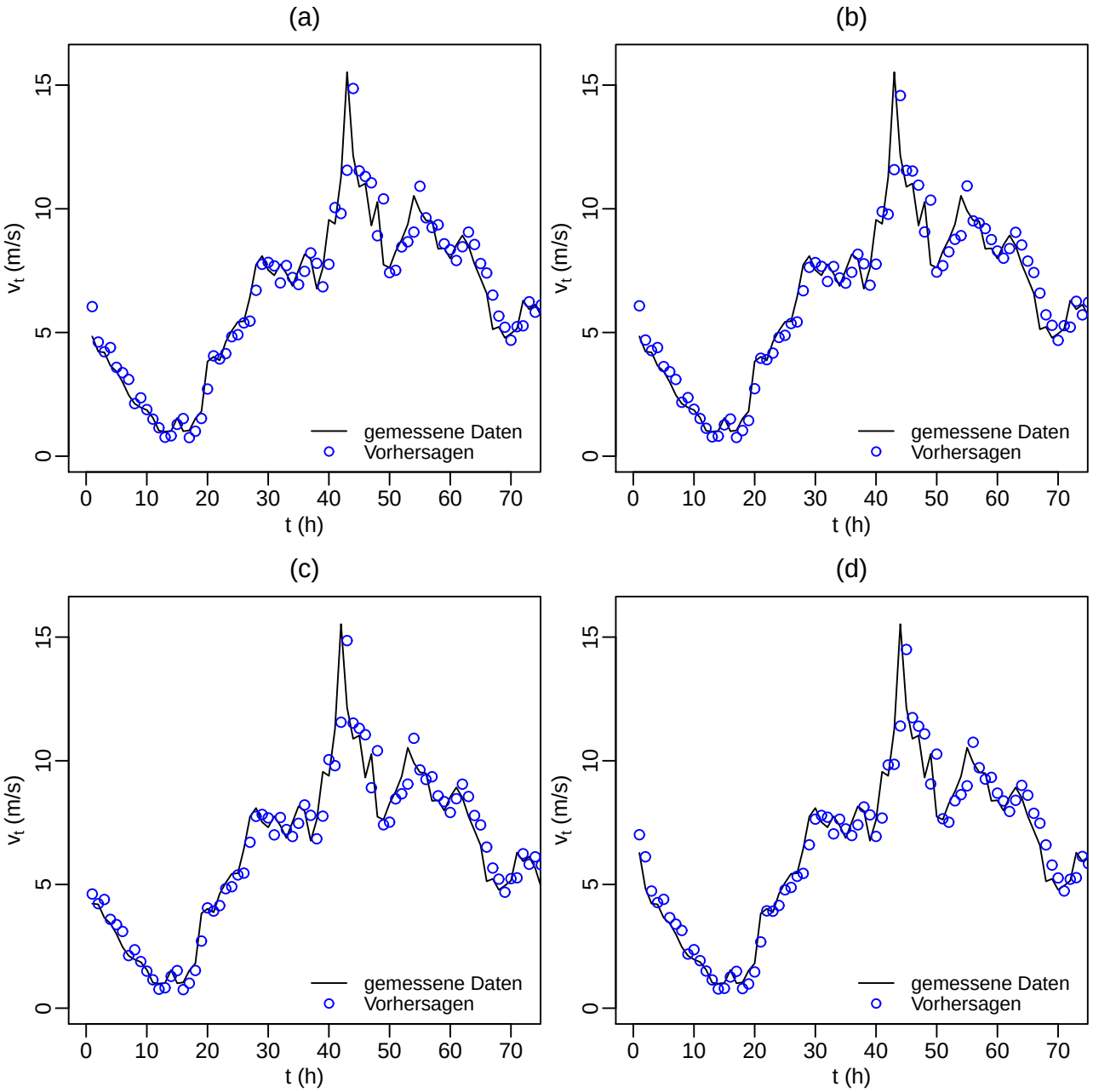


Abbildung 4.9: Vorhersagen für Windgeschwindigkeiten für je eine Stunde voraus. Gezeigt werden die ersten drei Tage vom Mai 2016. (a) $(p, d, q) = (1, 1, 0)$. (b) $(p, d, q) = (1, 1, 1)$. (c) $(p, d, q) = (2, 1, 0)$. (d) $(p, d, q) = (0, 1, 0)$.

Qualität der Voraussagen besser beurteilen zu können, werden das MAE (*mean absolute error*) und das RMSE (*root mean square error*) definiert:

$$\mathrm{MAE} = \frac{1}{24T - z} \sum_{t=z+1}^{24T} |v_t - \hat{v}_t| \tag{4.21}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{24T - z} \sum_{t=z+1}^{24T} (v_t - \hat{v}_t)^2}. \tag{4.22}$$

Dabei bezeichnen T die Anzahl der Tage im jeweiligen Monat und z die Anzahl der Windgeschwindigkeiten der vorigen Stunden, die nötig sind, um eine Voraussage $\hat{u}'_t$ entsprechend Gl. (4.12), (4.15) und (4.17) tätigen zu können. Das MAE und das RMSE quantifizieren die durchschnittliche Größenordnung der Fehler von Vorhersagen. Das RMSE in Gl. (4.22) wird bei gleichen Fehlern $v_t - \hat{v}_t$ aufgrund der Wurzel größer als das MAE in Gl. (4.21). Es lassen sich zum MAE und zum RMSE Standardabweichungen definieren, die durch

$$\sigma_{\mathrm{MAE}} = \sqrt{\frac{1}{24T - z} \sum_{t=z+1}^{24T} \left(|v_t - \hat{v}_t| - \mathrm{MAE}\right)^2} \tag{4.23}$$

$$\sigma_{\mathrm{RMSE}} = \sqrt{\frac{1}{24T - z} \sum_{t=z+1}^{24T} \left(|v_t - \hat{v}_t| - \mathrm{RMSE}\right)^2} \tag{4.24}$$

gegeben sind. Für das ARIMA(0,1,0)-Modell und die in Gl. (4.10) genannten Modelle werden das MAE und das RMSE und deren Standardabweichungen jeweils für jeden Monat berechnet. Die Ergebnisse sind in Abb. 4.10 zu sehen. Es fällt auf, dass in einigen Monaten die Standardabweichungen eher groß bzw. klein für alle vier Modelle sind. Weiterhin zeigen das MAE und RMSE für die Modelle einen ähnlichen Verlauf. Dies deutet wie zuvor erwähnt darauf hin, dass es Monate gibt, die durch alle Modelle schlechter angepasst werden können als in anderen Monaten. Damit die Fehler der Voraussagen durch ein Modell quantifiziert werden können, werden die Mittelwerte des MAE und des RMSE über alle zwanzig Monate gebildet:

$$\overline{\mathrm{MAE}} = \frac{1}{20} \sum_{m=1}^{20} \mathrm{MAE}_m \tag{4.25}$$

$$\overline{\mathrm{RMSE}} = \frac{1}{20} \sum_{m=1}^{20} \mathrm{RMSE}_m. \tag{4.26}$$

Dazu können deren Standardabweichungen $\sigma_{\overline{\mathrm{MAE,RMSE}}}$ berechnet werden:

$$\sigma_{\overline{\mathrm{MAE}}} = \sqrt{\frac{1}{20} \sum_{m=1}^{20} \left(\mathrm{MAE}_m - \overline{\mathrm{MAE}}\right)^2} \tag{4.27}$$

$$\sigma_{\overline{\mathrm{RMSE}}} = \sqrt{\frac{1}{20} \sum_{m=1}^{20} \left(\mathrm{RMSE}_m - \overline{\mathrm{RMSE}}\right)^2}. \tag{4.28}$$

Die Größen $\overline{\mathrm{MAE}}$ und $\overline{\mathrm{RMSE}}$ sollten für dasjenige Modell, das am besten für Voraussagen geeignet ist, minimal sein. Für die vier Modelle werden $\overline{\mathrm{MAE}}$ und $\overline{\mathrm{RMSE}}$ nach Gl. (4.25) und Gl. (4.26) in Abb. 4.10 angegeben. Die Mittelwerte $\overline{\mathrm{MAE}}$ und $\overline{\mathrm{RMSE}}$ sind für das ARIMA(1,1,1)-Modell am geringsten. Danach folgen das ARIMA(2,1,0)- und dann das ARIMA(1,1,0)-Modell. Die größten Fehler weist das Kontrollmodell, das ARIMA(0,1,0)-Modell, auf. Das durchschnittliche MAE ist hier um etwa 1 cm/s größer als bei den anderen Modellen. Wird Abb. 4.6 betrachtet, kann man erkennen, dass sich das ARIMA(1,1,1)-Modell mit der durchschnittlich kleinsten relativen quadratischen Abweichung $\overline{\Delta\mathrm{BIC}^2_{(1,1,1)}}$ des BIC am besten für Voraussagen eignet. Gemessen an der Abweichung $\overline{\Delta\mathrm{BIC}^2_{(1,1,0)}}$ folgt danach das ARIMA(1,1,0)-Modell , welches sich jedoch, anders als die Abweichung $\overline{\Delta\mathrm{BIC}^2_{(1,1,0)}}$ erwarten lässt, schlechter für Vorhersagen eignet als das ARIMA(2,1,0)-Modell. Es ist insgesamt zu erwähnen, dass sich alle Modelle etwa gleich gut für Vorhersagen eignen, da der Unterschied in der Genauigkeit bei 1 mm/s liegt, was einen nur sehr kleinen Teil der mittleren Windgeschwindigkeit $\langle v_t \rangle = 9.1$ m/s ausmacht. Die Untersuchung hat gezeigt, dass die Beschränkung auf eines der

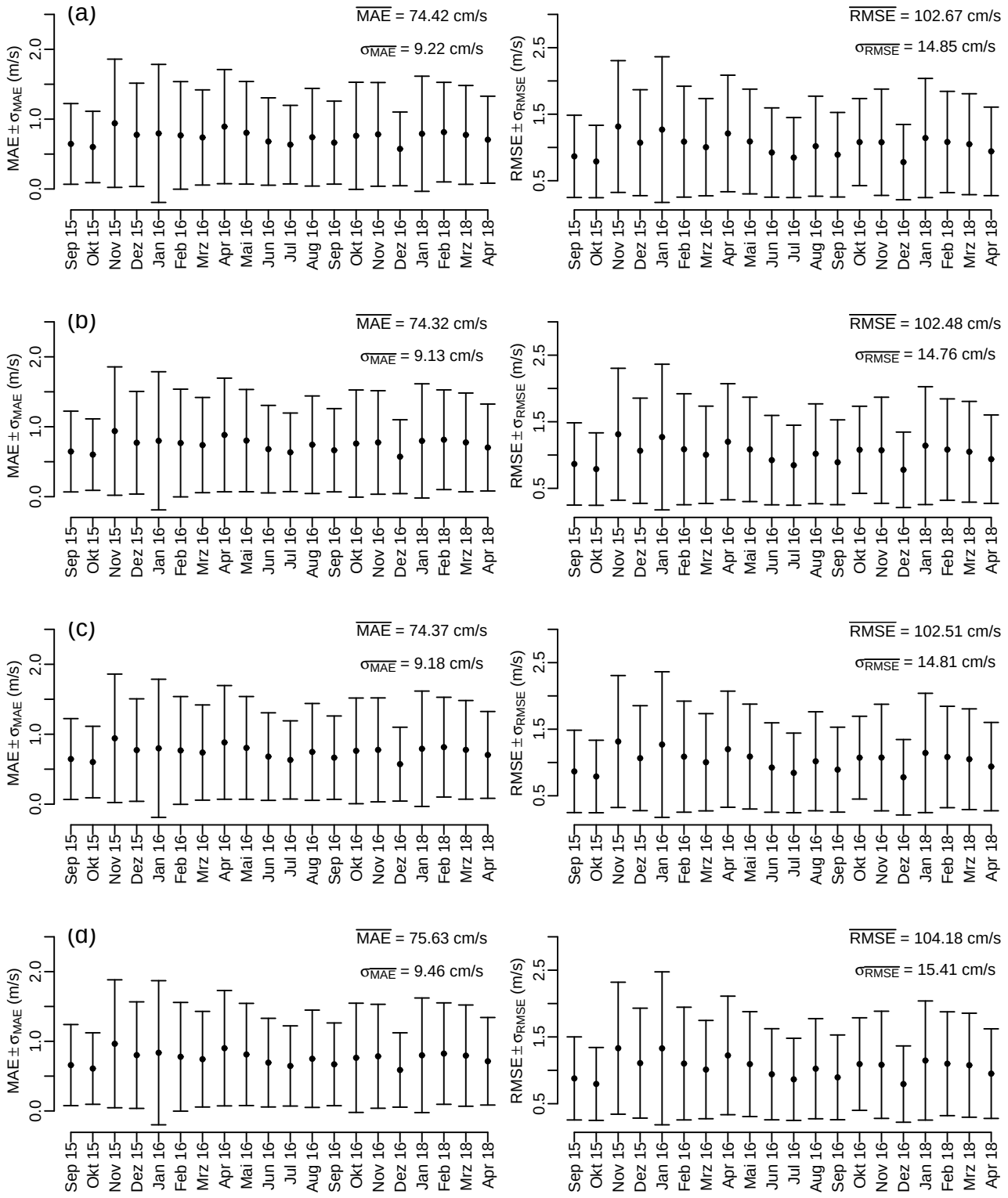


Abbildung 4.10: MAE (Gl. (4.21)) und RMSE (Gl. (4.22)) für Voraussagen aller 20 Monate. (a) $(p, d, q) = (1, 1, 0)$. (b) $(p, d, q) = (1, 1, 1)$. (c) $(p, d, q) = (2, 1, 0)$. (d) $(p, d, q) = (0, 1, 0)$.

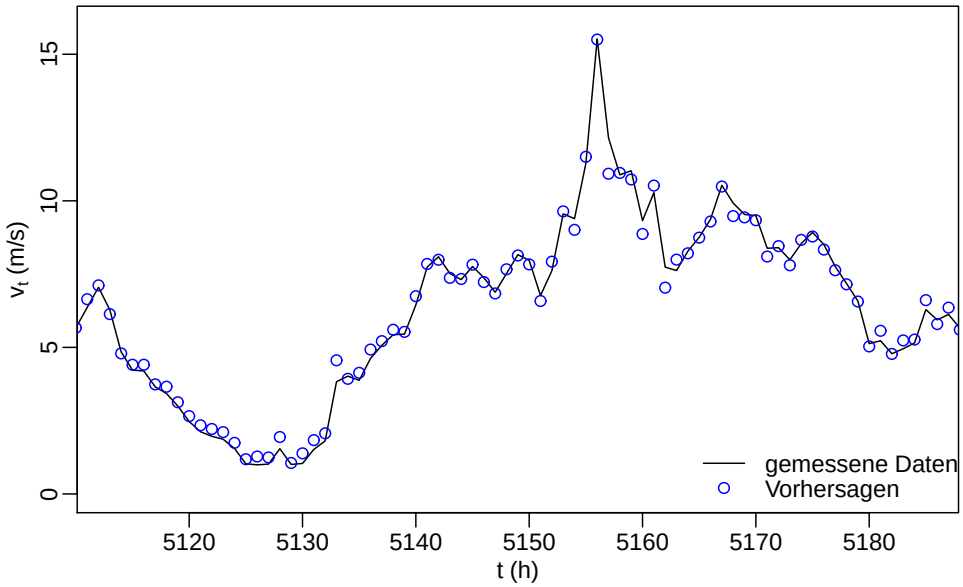


Abbildung 4.11: Vorhersagen von Windgeschwindigkeiten mithilfe aller möglichen ARIMA-Modelle für je eine Stunde voraus. Gezeigt werden die ersten drei Tage vom Mai 2016.

ARIMA-Modelle zu ähnlichen Ergebnissen bei der Voraussage führt, unabhängig davon, welches Modell verwendet wird. Im Folgenden wird daher eine eigene Methode zur Verbesserung der Genauigkeit von Vorhersagen entwickelt.
Es werden die gemessenen stündlichen Windgeschwindigkeiten $v(t)$ entsprechend Gl. (4.2) transformiert, wobei s stündliche Werte vorliegen. Anders als bei der bisherigen Vorgehensweise sollen nun alle ARIMA(p, d, q)-Modelle zugelassen werden. Die ARIMA-Analyse erfolgt jeweils für $H = 720$ h (Stunden) mit den gemessenen Werten und wird für jede Stunde t einzeln durchgeführt. Die Voraussage wird für die Stunde $t + H$ gemacht und im nächsten Schritt wird t um eine Stunde inkrementiert, bis die letzte Voraussage für die Stunde s getätigt werden kann. Die Vorgehensweise ist im Folgenden schematisch dargestellt:

$$\begin{array}{llll}
\text{ARIMA-Analyse} & \text{für 1 h bis 720 h} & \longrightarrow & \text{Vorhersage für 721 h} \\
\text{ARIMA-Analyse} & \text{für 2 h bis 721 h} & \longrightarrow & \text{Vorhersage für 722 h} \\
\vdots & \vdots & \vdots & \vdots \\
\text{ARIMA-Analyse} & \text{für } (s-720) \text{ h bis } (s-1) \text{ h} & \longrightarrow & \text{Vorhersage für } s.
\end{array}$$

Nach der gesamten Analyse liegen $s - H$ Vorhersagen vor. Diese müssen rücktransformiert werden, sodass man die geschätzten Windgeschwindigkeiten v_t erhält. Die Vorhersagen sind mit den gemessenen Windgeschwindigkeiten in Abb. 4.11 für einen Zeitraum von 3 Tagen dargestellt. Wie zuvor werden nun das MAE und das RMSE gebildet:

$$\text{MAE} = \frac{1}{s-H} \sum_{t=H+1}^{s} |v_t - \hat{v}_t| \tag{4.29}$$

$$\text{RMSE} = \sqrt{\frac{1}{s-H} \sum_{t=H+1}^{s} (v_t - \hat{v}_t)^2}. \tag{4.30}$$

	$(1,1,0)$	$(1,1,1)$	$(2,1,0)$	$(0,1,0)$	Zulassung aller Modelle
MAE in cm/s	74.42	74.32	74.37	75.63	15.27
RMSE in cm/s	102.67	102.48	102.51	104.18	24.12
MAE/ $\langle v_t \rangle$	8.129%	8.118%	8.123%	8.261%	1.668%

Tabelle 4.1: MAE, RMSE und relative Genauigkeit für verschiedene Vorgehensweisen von Vorhersagen von Windgeschwindigkeiten. Die mittlere Windgeschwindigkeit beträgt $\langle v_t \rangle = 9.1$ m/s.

Zusätzlich werden die Standardabweichungen von Gl. (4.29) und (4.30) analog zu Gl. (4.23) und (4.24) berechnet. Man erhält

$$\begin{aligned} \text{MAE} &= 15.27 \text{ cm/s} \\ \sigma_{\text{MAE}} &= 18.67 \text{ cm/s} \\ \text{RMSE} &= 24.12 \text{ cm/s} \\ \sigma_{\text{RMSE}} &= 20.66 \text{ cm/s}, \end{aligned}$$

womit die Genauigkeit im Vergleich zur vorigen Vorgehensweise, bei der lediglich je ein Modell zugelassen wird, deutlich erhöht wird.
Abschließend werden die Ergebnisse zur Genauigkeit verschiedener Vorhersagen in Tab. 4.1 zusammengefasst. Es ist zu bemerken, dass die Zulassung aller ARIMA-Modelle zur höchsten Genauigkeit führt. Die Verwendung lediglich eines Modells über alle Monate liefert in allen Fällen ähnliche Genauigkeiten.

5 Zusammenfassung

In Kap. 2 wurden die ARIMA-Modelle und die Methoden, mit denen die passenden Parameter gefunden werden können, beschrieben. Es ist aus Kap. 3 ersichtlich, dass die Genauigkeit der Analyse mit wachsender Datenmenge steigt. Weiterhin wurde die Vorgehensweise bei der ARIMA-Zeitanalyse in Kap. 3 praktisch erlernt, um in Kap. 4 Windgeschwindigkeiten analysieren zu können. Zuerst wurden die gemessenen Windgeschwindigkeiten im Abschnitt 4.1 stündlich gemittelt und anschließend nach der Arbeit von Brown et al. transformiert und standardisiert [5], damit der Datensatz für eine ARIMA-Analyse verwendet werden kann. Aus Abschnitt 4.2 kann erkannt werden, dass die stündlichen Windgeschwindigkeiten grundsätzlich eine Nichtstationarität aufweisen. Es wird in Abschnitt 4.3 versucht, je ein Modell für alle Monate zu verwenden, wobei die geeigneten Modelle anhand des BIC ausgewählt werden. Dabei ist zu bemerken, dass die Unterschiede des BIC eines Modells zum BIC eines anderen Modells klein sein können, was ein erster Hinweis darauf ist, dass mehrere Modelle ähnlich gut geeignet sind. Mit den in Abschnitt 4.3 gewählten Modellen werden in Abschnitt 4.4 Vorhersagen erstellt. Es wird hier gezeigt, dass die Genauigkeit der verschiedenen Modelle nur sehr leicht variiert. Daher wird abschließend eine eigene Methode für Vorhersagen zu Windgeschwindigkeiten entwickelt. Es kann bei der Analyse der Windgeschwindigkeiten die Genauigkeit der Vorhersagen erhöht werden, indem bei der Zeitreihenanalyse nicht nur ein ARIMA(p, d, q)-Modell zugelassen wird. Die Genauigkeit der Vorhersagen beträgt dann ca. 1.7% der durchschnittlichen Windgeschwindigkeit. Die üblichen ARIMA-Modelle sind Linearkombinationen des weißen Rauschens w_t, weswegen sie sich zur Beschreibung gaußverteilter Prozesse eignen. Jedoch sind die Windgeschwindigkeiten, wie aus Abb. 4.2 ersichtlich ist, weibullverteilt. Es wird die Genauigkeit der Vorhersagen vermutlich durch die Transformation nach Gl. (4.2) erhöht. In der Arbeit von Grigonytė et al., in der letztendlich ein höheres MAE als in dieser Bachelorarbeit erhalten wird, wird die ARIMA-Analyse durchgeführt, ohne dass die Windgeschwindigkeiten transformiert werden [8]. Es stellt sich die Frage, wie sich die Genauigkeit der Vorhersagen innerhalb der ARIMA-Modellfamilie verbessern lässt. Beispielsweise könnte die Transformation nach Gl. (4.2) verbessert werden. Weiterhin bleibt offen, ob und wie sich die Modellparameter verändern, wenn die zeitlichen Intervalle, über die die Windgeschwindigkeiten gemittelt werden, nicht eine Stunde betragen. Es wurde in dieser Bachelorarbeit stets über einen Zeitraum von einem Monat analysiert, weswegen es auch von Interesse ist, die Länge dieses Zeitraums derart zu variieren, dass die Vorhersagen maximal genau werden. Ist die beste Genauigkeit erreicht, stellt sich weiterhin die Frage, ob andere Modelle, wie z.B. SARIMA-Modelle, bei der Periodizitäten mitberücksichtigt werden können, eine noch höhere Genauigkeit erzielen können.
Andere Arbeiten, die sich mit der Vorhersage von Windgeschwindigkeiten beschäftigen, nutzen oft andere mit der ARIMA-Modellfamilie verwandte Modelle. Beispielsweise können, wenn Messungen von unterschiedlichen Orten vorliegen, die mehrdimensionalen VARIMA-Modelle, bei der unterschiedliche Zeitreihen miteinander gekoppelt werden, verwendet werden [9, 10]. Weiterhin kann auch die vorliegende Zeitreihe in hoch- und niedrigfrequente Anteile zerlegt werden und erst dann eine Analyse durchgeführt werden [11]. In der Arbeit von Zhang et al. dahingegen wird die originale Zeitreihe in periodische und nichtlineare Anteile zerlegt, wobei der periodische Anteil durch ein SARIMA-Modell und der nichtlineare durch ein ANFIS-Modell, welches auf neuronalen Netzwerken basiert, beschrieben werden [12]. Da neben den linearen Zeitreihenmodellen auch nichtlineare Modelle wie z.B. GARCH- oder NARX-Modelle existieren, kann versucht werden, letztere unter Hinzunahme weiterer physikalischer Parameter wie z.B. der Temperatur auf Windgeschwindigkeiten anzuwenden [13, 14], um eine

vorige Transformation der Daten umgehen zu können. Letztendlich kann versucht werden, auf die klassische Zeitreihenanalyse zu verzichten, wobei stattdessen Modelle neuronaler Netzwerke verwendet werden. Ein Vergleich sehr unterschiedlicher Modelle für die Analyse von Windgeschwindigkeiten ist in der Arbeit von Ma und Liu zu finden [15]. Viele der anderen Arbeiten erreichen ein höheres MAE, d.h. eine schlechtere Genauigkeit, als in dieser Bachelorarbeit. Es wäre wünschenswert, zu wissen, wovon die Genauigkeit der Vorhersagen abhängt. Vermutlich eignen sich ARIMA-Modelle besonders für gaußverteilte Prozesse, da Zeitreihen im Bild der ARIMA-Modelle Linearkombinationen des weißen Rauschens darstellen und daher selbst gaußverteilt sein müssen.
Abgesehen von der Analyse von Windgeschwindigkeiten ist zu erwähnen, dass Random Walks durch ARIMA(0,1,0)-Modelle, Ornstein-Uhlenbeck-Prozesse durch ARIMA(1,0,0)-Modelle beschrieben werden können. Es ist wünschenswert, zu wissen, ob und wie weitere stochastische Differentialgleichungen als ARIMA-Prozesse geschrieben werden können. Umgekehrt kann auch versucht werden, einen vorliegenden ARIMA-Prozess in eine stochastische Differentialgleichung zu überführen. Von Interesse ist die Verbindung zwischen der Langevin-Gleichung und der ARIMA-Modellfamilie, da Windgeschwindigkeiten auch durch stochastische Differentialgleichungen beschrieben werden können.

Literaturverzeichnis

[1] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications With R Examples*. Springer, Pittsburgh, PA, 2nd edition, 2006.

[2] P. J. Brockway and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer, Fort Collins and New York, 3rd edition, 2016.

[3] A. Hirotugu. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[4] The R Project for Statistical Computing. `https://cran.r-project.org`. Aufgerufen: 11.09.2017.

[5] Brown, Katz, and Murphy. Time Series Models to Simulate and Forecast Wind Speed and Wind Power. *Journal of Applied Meteorology and Climatology*, 23:1184–1195, 1984.

[6] FINO1-Projekt und Daten. `http://www.bsh.de`, 2016. Das FINO-Projekt wird durch die deutsche Regierung über das Bundesministerium für Wirtschaft und Energie und dem Forschungszentrum Jülich unterstützt.

[7] D. Hinkley. On quick choice of power transformation. *Applied Statistics*, 26:67–69, 1977.

[8] E. Grigonytė and E. Butkevičiutė. Short-term wind speed forecasting using ARIMA model. *Energetika*, 62:45–55, 2016.

[9] T. Filik. Improved Spatio-Temporal Linear Models for Very Short-Term Wind Speed Forecasting. *Energies*, 9(168), 2016.

[10] Yunus, Chen, and Thiringer. Modelling spatially and temporally correlated wind speed time series over a large geographical area using VARMA. *IET Renewable Power Generation*, 11:132–142, 2016.

[11] Yunus, Chen, and Thiringer. ARIMA-Based Frequency-Decomposed Modeling of Wind Speed Time Series. *IEEE Transactions on Power Systems*, 31(4):2546–2556, 2016.

[12] Zhang, Wei, Tan, Wang, and Tian. A Hybrid Method for Short-Term Wind Speed Forecasting. *Sustainability*, 9(596), 2017.

[13] G. Reikard. Using Temperature and State Transitions to Forecast Wind Speed. *Wind Energy*, 11:431–443, 2008.

[14] Cadenas, Rivera, Campos-Amezcua, and Heard. Wind Speed Prediction Using a Univariate ARIMA Model and a Multivariate NARX Model. *Energies*, 9(109):716–723, 2016.

[15] X. Ma and D. Liu. Comparative Study of Hybrid Models Based on a Series of Optimization Algorithms and Their Application in Energy System Forecasting. *Energies*, 9(640), 2016.

Danksagung

Zuerst danke ich Herrn Prof. Dr. Philipp Maaß für die Aufnahme in die Arbeitsgruppe *Statistische Physik*. Außerdem möchte ich Herrn Dr. Pedro Lind für die zahlreichen Denkanstöße während der Bachelorarbeit danken. Ein besonderer Dank geht an die Forschungsplattform FINO1 für die Daten zu Windgeschwindigkeiten. Das FINO-Projekt wurde finanziell vom Bundesministerium für Wirtschaft und Energie und dem Forschungszentrum Jülich unterstützt. Nicht zuletzt möchte ich den restlichen Mitgliedern der Arbeitsgruppe für die herzliche Aufnahme und ein angenehmes Zusammensein während und neben der Arbeit danken.

Eidesstattliche Versicherung

Ich versichere, dass ich die eingereichte Bachelorarbeit selbstständig und ohne unerlaubte Hilfe verfasst habe. Anderer als der von mir angegebenen Hilfsmittel und Schriften habe ich mich nicht bedient. Alle wörtlich oder sinngemäß den Schriften anderer Autoren entnommenen Stellen habe ich kenntlich gemacht.

Ort, Datum

Unterschrift