Universität Potsdam
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Mathematik
Arbeitsgruppe Stochastik

Masterarbeit

Ein multityper Verzweigungsprozess als Modell zur Untersuchung der Ausbreitung von Covid-19

vorgelegt von: Andrea Hübner
Email: andrea.erna.huebner@gmail.com
Erste Gutachterin: Prof. Dr. Sylvie Roelly
Zweite Gutachterin: Frau Dr. Franziska Göbel

Potsdam, den 21.03.2021

Online veröffentlicht auf dem
Publikationsserver der Universität Potsdam:
https://doi.org/10.25932/publishup-50922
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-509225

Inhaltsverzeichnis

1 Einleitung

Im Zuge der Covid-19 Pandemie werden zwei Werte täglich diskutiert: Die zuletzt gemeldete Zahl der neu Infizierten und die sogenannte Reproduktionsrate. Sie gibt wieder, wie viele weitere Menschen ein an Corona erkranktes Individuum im Durchschnitt ansteckt. Für die Schätzung dieses Wertes gibt es viele Möglichkeiten - auch das Robert Koch-Institut gibt in seinem täglichen Situationsbericht stets zwei R-Werte an: Einen 4-Tage-R-Wert und einen weniger schwankenden 7-Tage-R-Wert (siehe [11]). Diese Arbeit soll eine weitere Möglichkeit vorstellen, einige Aspekte der Pandemie zu modellieren und den R-Wert zu schätzen.

In der ersten Hälfte der Arbeit werden die mathematischen Grundlagen vorgestellt, die man für die Modellierung benötigt. Hierbei wird davon ausgegangen, dass der Leser bereits ein Basisverständnis von stochastischen Prozessen hat. Im Abschnitt *Grundlagen* werden Verzweigungsprozesse mit einigen Beispielen eingeführt und die Ergebnisse aus diesem Themengebiet, die für diese Arbeit wichtig sind, präsentiert. Dabei gehen wir zuerst auf einfache Verzweigungsprozesse ein und erweitern diese dann auf Verzweigungsprozesse mit mehreren Typen. Um die Notation zu erleichtern, beschränken wir uns auf zwei Typen, das Prinzip lässt sich aber auf eine beliebige Anzahl von Typen erweitern.

Vor allem soll die Wichtigkeit des Parameters λ herausgestellt werden. Dieser Wert kann als durchschnittliche Zahl von Nachfahren eines Individuums interpretiert werden und bestimmt die Dynamik des Prozesses über einen längeren Zeitraum. In der Anwendung auf die Pandemie hat der Parameter λ die gleiche Rolle wie die Reproduktionsrate R.

In der zweiten Hälfte dieser Arbeit stellen wir eine Anwendung der Theorie über Multitype Verzweigungsprozesse vor. Prof. Yanev und seine Mitarbeiter in ihrer Veröffentlichung *Branching stochastic processes as models of Covid-19 epidemic development* modellieren die Ausbreitung des Corona Virus' über einen Verzweigungsprozess mit zwei Typen. Wir werden dieses Modell diskutieren und Schätzer daraus ableiten: Ziel ist es, die Reproduktionsrate zu ermitteln. Außerdem analysieren wir die Möglichkeiten, die Dunkelziffer (die Zahl nicht gemeldeter Krankheitsfälle) zu schätzen. Wir wenden die Schätzer auf die Zahlen von Deutschland an und werten diese schließlich aus.

2 Grundlagen

2.1 Monotype Verzweigungsprozesse

Zunächst betrachten wir Verzweigungsprozesse in einfachster Form. Die Definition orientiert sich am Skript *Grundlegende Eigenschaften von Bienaymé - Galton - Watson - Verzweigungsprozessen in diskreter Zeit*, siehe [13].

Es wird die Zahl X_n der Individuen in einer Bevölkerung zur Zeit n betrachtet. Zu jedem Zeitschritt produziert jedes Individuum Nachkommen nach einem Reproduktionsgesetz $\mathbf{r} := (r_i)_{i\in\mathbb{N}}$ und stirbt danach selbst. Mit $Y_{n,k}$ beschreiben wir die Nachkommen des k-ten Individuums der n-ten Generation. Diese Zufallsvariablen sind unabhängig und identisch verteilt in der Form:

$$\mathbb{P}(Y_{n,k} = i) = r_i \qquad \forall i \in \mathbb{N}$$

Die Anzahl neuer Individuen in der nächsten Generation berechnet sich also wie folgt:

$$\begin{cases} X_{n+1} = \sum_{k=1}^{X_n} Y_{n,k} & \text{, falls } X_n \geq 1 \\ X_{n+1} = 0 & \text{, falls } X_n = 0 \end{cases}$$

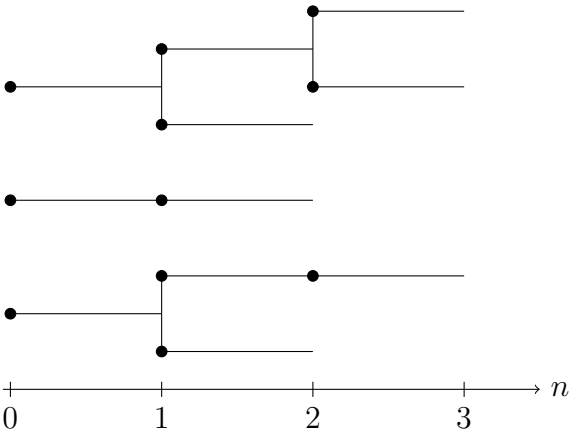
Es handelt sich also um eine Summe von Zufallsvariablen mit der Besonderheit, dass die Anzahl der Summanden ebenfalls zufällig ist.

Für $s \in [0,1]$ definieren wir die erzeugende Funktion $f(s)$ von $Y_{n,k}$, welche äquivalent ist zur erzeugenden Funktion von X_1, für den Fall, dass $X_0 = 1$.

$$f(s) = \mathbb{E}[s^{Y_{n,k}}] = \mathbb{E}[s^{X_1|(X_0=1)}] = \sum_{i=0}^{\infty} \mathbb{P}(Y_{n,k} = i)s^i = \sum_{i=0}^{\infty} r_i s^i$$

f ist stetig und monoton wachsend.

Beispiel 1. *Ein mögliches Reproduktionsgesetz könnte* $\mathbf{r} = \frac{1}{3}(1,1,1)$ *sein. Das würde bedeuten: jedes Individuum kann mit gleicher Wahrscheinlichkeit kein, ein oder zwei Nachfahren bekommen. So ein Prozess mit Anfangsbevölkerung* $X_0 = 3$ *könnte zum Beispiel so aussehen:*



Die erzeugende Funktion von $Y_{n,k}$ *wäre hier:*

$$f(s) = \sum_{i=0}^{\infty} r_i s^i = \frac{1}{3} + \frac{1}{3}s + \frac{1}{3}s^2$$

Diese Prozesse erfüllen die **Verzweigungseigenschaft**, auch additive Eigenschaft genannt: Das bedeutet, dass ein Prozess, der mit i Individuen beginnt, äquivalent ist zu der Summe von i Prozessen, die mit einem Individuum beginnen. Also für alle $i, k \in \mathbb{N}$ gilt:

$$\mathbb{P}((X_{n+k})_{n\geq 0} \in \cdot | X_k = i) = \mathbb{P}((\sum_{j=1}^{i} X_n^{(j)})_{n\geq 0} \in \cdot),$$
$$\text{wobei } \mathbb{P}(X_n^{(j)} \in \cdot) = \mathbb{P}(X_n \in \cdot | X_0 = 1)$$

Die Verzweigungseigenschaft liefert uns eine interessante Eigenschaft der erzeugenden Funktion:

Theorem 1. *Bezeichnen wir mit f_n die erzeugende Funktion von X_n, startend in $X_0 = 1$, so gilt:*

$$f_n = \underbrace{f \circ f \circ ... \circ f}_{n \text{ mal}} \quad \text{und} \quad \mathbb{E}[s^{X_n} | X_0 = i] = (f_n(s))^i$$

Der zweite Fakt ergibt sich direkt aus der Verzweigungseigenschaft und daraus, dass die erzeugende Funktion der Summe von Zufallsvariablen sich aus dem Produkt der einzelnen erzeugenden Funktionen ergibt.

Sowohl die Verzweigungseigenschaft, als auch dieses Theorem finden sich in [13], in Kapitel 1 von [3], [6] und [15] . Wir geben den ausführlichen Beweis aus [13] wieder, welcher induktiv erfolgt:

Beweis.

$$\begin{aligned} f_{n+1}(s) &= \mathbb{E}[s^{X_{n+1}} | X_0 = 1] = \sum_{i\geq 0} \mathbb{E}[\mathbb{1}_{X_n=i} s^{X_{n+1}} | X_0 = 1] \\ &= \sum_{i\geq 0} \mathbb{E}[\mathbb{1}_{X_n=i} s^{\sum_{k=1}^{i} Y_{n,k}} | X_0 = 1] = \sum_{i\geq 0} \mathbb{E}[\mathbb{1}_{X_n=i} | X_0 = 1](\mathbb{E}[s^{X_1}] | X_0 = 1])^i \\ &= \sum_{i\geq 0} \mathbb{P}(X_n = i | X_0 = 1)(f(s))^i = f_n \circ f(s) \end{aligned}$$

□

2.2 Multitype Verzweigungsprozesse

Beim vorher beschriebenen Verzweigungsprozess waren alle Individuen der Bevölkerung gleichartig. Bei multitypen Verzweigungsprozessen kann die Bevölkerung in d verschiedene Typen eingeteilt werden. Da wir nun oft mit Vektoren und ihren Einträgen hantieren, wollen wir Vektoren hervorheben:

$$\mathbf{X}_n = (X_n^{(1)}, X_n^{(2)}), \quad \mathbf{s} = (s_1, s_2) \text{ für } \mathbf{s} \in \mathbb{R}^2, \quad \mathbf{f} = (f^{(1)}, f^{(2)}), \text{ usw.}$$

Dies erlaubt uns gelegentlich auch eine verkürzte Schreibweise.

2.2.1 Einführung des Modells und der Notation

Die hier vorgestellte Konstruktion ist eine vereinfachte Version von der in Kapitel 5 des Buches *Branching Processes* ([3]). Die Gesamtzahl der Bevölkerung ist eine Zusammenstellung der Individuen der d verschiedenen Typen: $\mathbf{X}_n = (X_n^{(1)}, X_n^{(2)}, ..., X_n^{(d)})$.

Hierbei ist $X_n^{(l)}$ die Zahl der Individuen vom Typ l in der n-ten Generation.

Jeder Typ kann eigene Reproduktionsgesetze haben. Sei $Y_{n,k,j}^{(l)}$ die Zahl der Kinder vom Typ l, die das k-te Individuum vom Typ j in der n-ten Generation hat. Diese Zufallsvariablen sind unabhängig voneinander. Das Wahrscheinlichkeitsmaß p_j auf $\mathbb{N}^d$ ist das Reproduktionsgesetz für den j-ten Typ.

$$\mathbb{P}(\mathbf{Y}_{n,k,j} = (i_1, i_2, ..., i_d)) = p_j(i_1, i_2, ..., i_d) \text{ , für } (i_1, i_2, ..., i_d) \in \mathbb{N}_0^d$$

Im Folgenden wollen wir uns auf **zwei Typen** beschränken. Dann berechnet sich die Zahl neuer Individuen vom Typ 1 wie folgt:

$$X_{n+1}^{(1)} = \sum_{k=1}^{X_n^{(1)}} Y_{n,k,1}^{(1)} + \sum_{k=1}^{X_n^{(2)}} Y_{n,k,2}^{(1)}$$

Und die Zahl der neuen Individuen vom zweiten Typ entsprechend:

$$X_{n+1}^{(2)} = \sum_{k=1}^{X_n^{(1)}} Y_{n,k,1}^{(2)} + \sum_{k=1}^{X_n^{(2)}} Y_{n,k,2}^{(2)}$$

Die Reproduktionsgesetze sind zwei Maße auf $\mathbb{N}_0^2$:

$$\begin{aligned} &\mathbb{P}(\mathbf{Y}_{n,k,1} = (i_1, i_2)) = p_1(i_1, i_2), \\ &\mathbb{P}(\mathbf{Y}_{n,k,2} = (i_1, i_2)) = p_2(i_1, i_2), \text{ für } i_1, i_2 \in \mathbb{N}_0 \end{aligned}$$

Für feste l und j sind $Y_{n,k,j}^{(l)}$ für verschiedene n und k identisch verteilt. Daher werden in späteren Betrachtungen n und k nicht spezifiziert, um die Notation nicht zu überladen.

Die **Verzweigungseigenschaft** gilt auch im multitypen Fall:

$$\begin{aligned} \mathbb{P}((\mathbf{X}_{n+k})_{n\geq 0} \in \cdot | \mathbf{X}_k = (i_1, i_2)) &= \mathbb{P}((\sum_{j_1=1}^{i_1} (\hat{X}_n^{(1),j_1}, 0) + \sum_{j_2=1}^{i_2} (0, \tilde{X}_n^{(2),j_2}))_{n\geq 0} \in \cdot), \\ \text{wobei } \mathbb{P}(\hat{X}_n^{(1),j_1} \in \cdot) &= \mathbb{P}(X_n^{(1)} \in \cdot | \mathbf{X}_0 = (1,0)) \\ \text{und } \mathbb{P}(\tilde{X}_n^{(2),j_2} \in \cdot) &= \mathbb{P}(X_n^{(2)} \in \cdot | \mathbf{X}_0 = (0,1)) \end{aligned}$$

Des Weiteren führen wir Relationen auf $\mathbb{R}^2$ ein:

Definition 1. - *Für* $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^2$ *ist* $\boldsymbol{s} \geq \boldsymbol{t} \Leftrightarrow s_1 \geq t_1$ *und* $s_2 \geq t_2$.

- *Für* $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^2$ *ist* $\boldsymbol{s} < \boldsymbol{t} \Leftrightarrow \boldsymbol{s} \leq \boldsymbol{t}$ *und* $s_1 < t_1$ *oder* $s_2 < t_2$.

Definition 2. *Für* $0 \leq \boldsymbol{s} \leq 1$ *definieren wir die erzeugende Funktion* $\boldsymbol{f}(\boldsymbol{s}) = (f^{(1)}(s_1, s_2), f^{(2)}(s_1, s_2))$ *von* $\boldsymbol{X}_1 = (X_1^{(1)}, X_1^{(2)})$*:*

$$f^{(1)}(s_1, s_2) = \mathbb{E}[s_1^{Y_{n,k,1}^{(1)}} s_2^{Y_{n,k,1}^{(2)}}] = \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} p_1(i_1, i_2) s_1^{i_1} s_2^{i_2}$$

$$f^{(2)}(s_1, s_2) = \mathbb{E}[s_1^{Y_{n,k,2}^{(1)}} s_2^{Y_{n,k,2}^{(2)}}] = \sum_{i_1=0}^{\infty} \sum_{i_2=0}^{\infty} p_2(i_1, i_2) s_1^{i_1} s_2^{i_2}$$

$\boldsymbol{f}$ *ist die Erzeugendenfunktion von* $\boldsymbol{X}_1$ *und* $\boldsymbol{f}_n$ *ist die Erzeugendenfunktion für* $\boldsymbol{X}_n$.

$f^{(i)}$ ist monoton wachsend und stetig. Mit ähnlichen Argumenten wie in Theorem 1 (siehe zum Beispiel [7] Seite 411) kann auch gezeigt werden, dass:

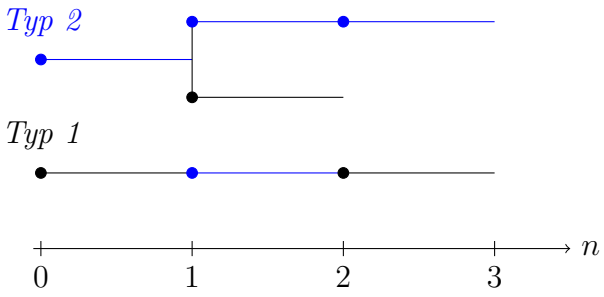
$$\mathbf{f}_n = \underbrace{\mathbf{f} \circ \mathbf{f} \circ \ldots \circ \mathbf{f}}_{n \text{ mal}} \tag{1}$$

Beispiel 2. *Ein bityper Verzweigungsprozess könnte zum Beispiel die Reproduktionsgesetze haben:*

$$p_1(0,0) = p_1(0,1) = p_1(1,0) = \frac{1}{3}$$

$$p_2(0,0) = 0, \quad p_2(0,1) = p_2(1,0) = p_2(1,1) = \frac{1}{3}$$

Nur Individuen vom ersten Typ können aussterben, während der zweite Typ immer mindestens einen Nachfahren hat. Eine Realisation dieses Prozesses mit Anfangsbedingung $X_0 = (1,1)$ *könnte so aussehen:*



Die erzeugenden Funktionen wären hier:

$$f^{(1)}(s_1, s_2) = \frac{1}{3}(1 + s_1 + s_2)$$

$$f^{(2)}(s_1, s_2) = \frac{1}{3}(s_1 + s_2 + s_1 s_2)$$

Beispiel 3. *Ist die Anzahl möglicher Nachfahren für jeden Typ endlich, so ergibt sich für die erzeugende Funktion stets ein Polynom. Sei* m *die maximale Anzahl an Nachkommen, die ein Individuum von irgendeinem Typ haben kann. Und seien* p_1 *und* p_2 *Wahrscheinlichkeitsmaße auf* $\{0, 1, ..., m\}^2$*. Dann ist:*

$$f^{(1)}(s_1, s_2) = \sum_{i_1=0}^{m} \sum_{i_2=0}^{m} p_1(i_1, i_2) s_1^{i_1} s_2^{i_2}$$

$$f^{(2)}(s_1, s_2) = \sum_{i_1=0}^{m} \sum_{i_2=0}^{m} p_2(i_1, i_2) s_1^{i_1} s_2^{i_2}$$

2.2.2 Asymptotisches Verhalten in Abhängigkeit zum Parameter λ

Das asymptotische Verhalten solcher Prozesse lässt sich nicht intuitiv ablesen. Um die Dynamik besser betrachten zu können, definieren wir die Durchschnittsmatrix, auch Mutationsmatrix genannt:

Definition 3. *Für einen bitypen Verzweigungsprozess sei* $M = (m_{i,j})_{i,j=1,2} \in \mathbb{R}^{2\times 2}$ *die Durchschnittsmatrix mit:*

$$m_{i,j} = \mathbb{E}[Y^{(j)}_{n,k,i}] = \sum_{i_1,i_2} p_i(i_1,i_2) i_j$$

$m_{i,j}$ ist also die durchschnittliche Zahl der Nachkommen vom Typ j, die ein Individuum vom Typ i hat. Wir gehen im Folgenden stets davon aus, dass alle Einträge von M endlich sind, also für alle $i,j \in \{1,2\}$ gilt: $\mathbb{E}[|Y^{(j)}_{n,k,i}|] = \mathbb{E}[Y^{(j)}_{n,k,i}] < \infty$.

Betrachten wir den Zusammenhang mit der erzeugenden Funktion:

Da $s_1, s_2 \leq 1$ und da sich die Koeffizienten p_j zu 1 aufsummieren und positiv sind, konvergiert die Reihe:

$$f^{(i)}(s_1,s_2) = \mathbb{E}[s_1^{Y^{(1)}_{n,k,i}} s_2^{Y^{(2)}_{n,k,i}}] = \sum_{i_1=0}^{\infty}\sum_{i_2=0}^{\infty} p_i(i_1,i_2) s_1^{i_1} s_2^{i_2} \leq \sum_{i_1=0}^{\infty}\sum_{i_2=0}^{\infty} p_i(i_1,i_2) = 1$$

Somit ist die Potenzreihe $\mathbf{f}$ für $0 < \mathbf{s} < 1$ unendlich oft differenzierbar.

Wir bezeichnen den Grenzwert der Ableitung von links: $\lim_{(s_1,s_2)\uparrow(1,1)} \frac{\partial f^{(i)}}{\partial s_j}(s_1,s_2)$ abkürzend mit $\frac{\partial f^{(i)}}{\partial s_j}(1,1)$. Der Grenzwert existiert, da $f^{(i)}$ monoton wachsend und stetig ist. Es gilt: $\frac{\partial f^{(i)}}{\partial s_j}(1,1) = \infty$ genau dann, wenn $\mathbb{E}[Y^{(j)}_{n,k,i}] = \infty$.

$$\begin{aligned}
\frac{\partial f^{(i)}}{\partial s_1} &= \sum_{i_1=1}^{\infty}\sum_{i_2=0}^{\infty} p_i(i_1,i_2) i_1 s_1^{i_1-1} s_2^{i_2} \\
\frac{\partial f^{(i)}}{\partial s_1}(1,1) &= \sum_{i_1=1}^{\infty}\sum_{i_2=0}^{\infty} p_i(i_1,i_2) i_1 \\
\Rightarrow m_{i,1} &= \frac{\partial f^{(i)}}{\partial s_1}(1,1) \text{ und analog: } m_{i,2} = \frac{\partial f^{(i)}}{\partial s_2}(1,1) \\
\Rightarrow m_{i,j} &= \frac{\partial f^{(i)}}{\partial s_j}(1,1)
\end{aligned}$$

In unserem **Beispiel 2** ergäbe sich:

$$\begin{aligned}
m_{1,1} &= \mathbb{E}[Y^{(1)}_{n,k,1}] = \sum_{i_1=0}^{1}\sum_{i_2=0}^{1} p_1(i_1,i_2) i_1 = p_1(1,0) = \frac{1}{3} \\
m_{1,2} &= \frac{\partial f^{(1)}}{\partial s_2}(1,1) = \sum_{i_1=0}^{1}\sum_{i_2=1}^{1} p_1(i_1,i_2) i_2 = p_1(0,1) = \frac{1}{3} \\
m_{2,1} &= \mathbb{E}[Y^{(1)}_{n,k,2}] = \sum_{i_1=0}^{1}\sum_{i_2=0}^{1} p_2(i_1,i_2) i_1 = p_2(1,0) + p_2(1,1) = \frac{2}{3} \\
m_{2,2} &= \frac{\partial f^{(2)}}{\partial s_2}(1,1) = \sum_{i_1=0}^{1}\sum_{i_2=1}^{1} p_1(i_1,i_2) i_2 = p_2(0,1) + p_2(1,1) = \frac{2}{3} \\
M &= \frac{1}{3}\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}
\end{aligned}$$

Jeder Prozess hat genau eine solche Matrix, aber mit der Matrix allein kann man den Prozess nicht rekonstruieren, da bei ihr Informationen verloren gehen. Folgende Reproduktionswahrscheinlichkeiten würden beispielsweise die gleichen Durchschnittswerte liefern:

$$p_1(0,0) = \frac{1}{2}, \quad p_1(1,0) = p_1(0,1) = p_1(1,1) = 1/6$$
$$p_2(0,0) = p_2(2,0) = p_2(0,2) = \frac{1}{3}$$

Definition 4. *Eine Matrix M heißt* strikt positiv*, falls es ein $n \in \mathbb{N}$ gibt, sodass in M^n jeder Eintrag positiv ist. Das heißt $m_{i,j}^{(n)} > 0$ für alle $i, j \in \{1, 2\}$, wenn $m_{i,j}^{(n)}$ den (i, j)-ten Eintrag in M^n bezeichnet. Ein Verzweigungsprozess, dessen Durchschnittsmatrix diese Eigenschaft hat, heißt* positiv regulär*.*

Bei positiv regulären Prozessen kann bei jeder Anfangsbedingung jeder Zustand in endlicher Zeit mit positiver Wahrscheinlichkeit erreicht werden. Um den Verlauf solcher Prozesse zu beobachten, hilft uns das Theorem aus [3] aus Kapitel 2.

Theorem 2. *Jede strikt positive Matrix M mit nichtnegativen Einträgen hat einen maximalen Eigenwert $\lambda \in \mathbb{R}^+$. Das heißt: für alle Eigenwerte $\lambda' \in \mathbb{C}, \lambda' \neq \lambda$ gilt: $|\lambda'| < \lambda$. Dieser Eigenwert ist einfach, hat also die algebraische Vielfachheit eins. Seien $\boldsymbol{u}$ der dazugehörige Rechtseigenvektor und $\boldsymbol{v}$ der assoziierte Linkseigenvektor. Seien $\boldsymbol{u}$ und $\boldsymbol{v}$ normiert, sodass $\boldsymbol{u} \cdot \boldsymbol{v} = 1$ und $\boldsymbol{u} \cdot (1, 1) = 1$. Dann ist die Matrix C mit $C_{i,j} = u_i v_j$ strikt positiv und es gilt:*

$$M^n = \lambda^n C + R^n$$

wobei $CR = RC = 0$ und für alle $i, j = 1, 2$ gilt: $(R^n)_{i,j} = \mathcal{O}(\lambda_0^n)$ für ein $\lambda_0 < \lambda$ bzw. $(R^n)_{i,j} = o(\lambda^n)$.

Für zwei Vektoren $\mathbf{u}$ und $\mathbf{v}$ bezeichnet $\mathbf{u} \cdot \mathbf{v}$ das Skalarprodukt.

Die Dynamik des Prozesses für große n kann also durch den maximalen Eigenwert λ von der Durchschnittsmatrix M kontrolliert werden. Daher kann λ als Wert für die durchschnittliche Zahl der Nachfahren eines Individuums interpretiert werden. Bei $\lambda < 1$ nennt man den Prozess subkritisch, bei $\lambda = 1$ kritisch und bei $\lambda > 1$ superkritisch.

Dass der maximale Eigenwert selbst bei zweidimensionalen Matrizen schon nicht intuitiv zu raten ist, soll folgendes Beispiel illustrieren:

Beispiel 4. *Betrachten wir einen Verzweigungsprozess mit folgender Durchschnittsmatrix:*

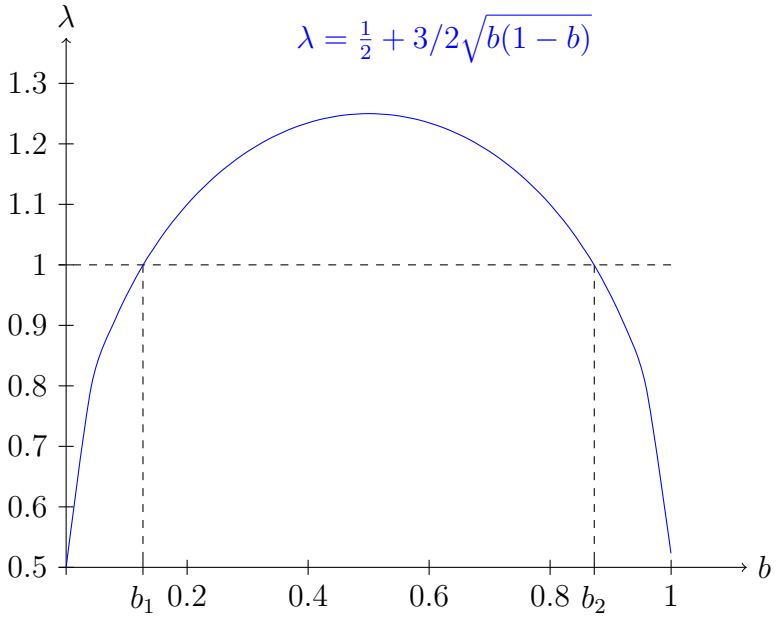
$$M = \begin{pmatrix} \frac{1}{2} & \frac{9}{4}b \\ 1 - b & \frac{1}{2} \end{pmatrix}$$

Mit $b \in [0, 1]$. Dann ist der maximale Eigenwert die Lösung von $\det(M - \lambda \mathbb{1}) = 0$:

$$\lambda = \frac{1}{2} + 3/2\sqrt{b(1-b)}.$$

Der Prozess wird kritisch bei $\lambda = 1$. Das passiert, wenn $b(1 - b) = 1/9$. Somit sind die kritischen Werte für b:

$$b_1 = \frac{1}{2} - \frac{\sqrt{5}}{6} \approx 0.127$$
$$b_2 = \frac{1}{2} + \frac{\sqrt{5}}{6} \approx 0.873.$$



Der Prozess ist also subkritisch ($\lambda < 1$) für $b \in [0, b_1) \cup (b_2, 1]$,
kritisch für $\lambda = 1$, also $b = b_1$ oder $b = b_2$
und superkritisch ($\lambda > 1$) dazwischen, also für $b \in (b_1, b_2)$.

2.2.3 Aussterbewahrscheinlichkeit

Wir wollen triviale Prozesse ausschließen.

Definition 5. *Ein Prozess X_n heißt* singulär *oder* einfach, *wenn es eine Matrix A mit nichtnegativen Einträgen gibt, sodass für die erzeugende Funktion gilt: $\boldsymbol{f}(\boldsymbol{s}) = A\boldsymbol{s}$.*

Ein Prozess ist singulär, wenn jedes Individuum immer genau einen Nachfahren hat und nur noch zufällig ist, von welchem Typ der Nachfahre ist. Die Gesamtzahl der Population bleibt dann konstant und der Prozess ist für unsere Zwecke uninteressant.

Beispiel 5. *Mit den Reproduktionsgesetzen:*

$$p_1(1,0) = \frac{1}{2}, \quad p_1(0,1) = \frac{1}{2}$$
$$p_2(1,0) = 1$$

ist der Prozess singulär. Betrachte die erzeugende Funktion:

$$f_1(s_1, s_2) = \frac{1}{2}(s_1 + s_2)$$
$$f_2(s_1, s_2) = s_1$$

Dann finden wir eine nichtnegative Matrix A, für die gilt: $\boldsymbol{f}(\boldsymbol{s}) = A\boldsymbol{s}$.

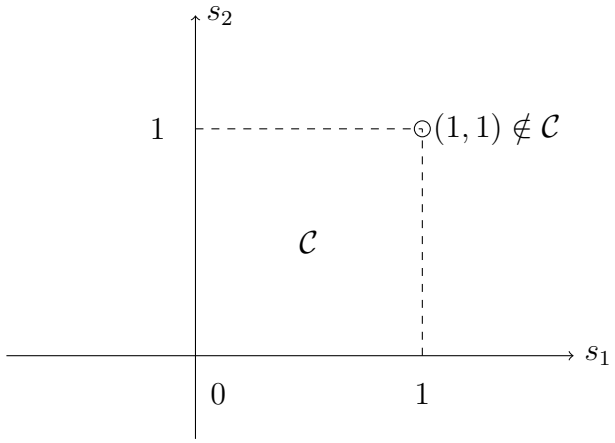
$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$$
$$A\boldsymbol{s} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(s_1 + s_2) \\ s_1 \end{pmatrix} = \boldsymbol{f}(\boldsymbol{s}).$$

Definition 6. *Mit $\boldsymbol{a} = (a_1, a_2)$ bezeichnen wir den Vektor mit Aussterbewahrscheinlichkeiten: $a_i \in [0,1]$ ist die* Aussterbewahrscheinlichkeit *eines Prozesses, der mit genau einem Individuum des Types i beginnt.*

$$a_1 = \mathbb{P}(\boldsymbol{X}_n = (0,0) \text{ für ein } n \in \mathbb{N} | \boldsymbol{X}_0 = (1,0)).$$
$$a_2 = \mathbb{P}(\boldsymbol{X}_n = (0,0) \text{ für ein } n \in \mathbb{N} | \boldsymbol{X}_0 = (0,1)).$$

Des Weiteren führen wir das positive Einheitsquadrat auf $\mathbb{R}^2$ ein:

Definition 7. $\mathcal{C} = \{(s_1, s_2) \in \mathbb{R}^2 | (0,0) \leq \boldsymbol{s} < (1,1)\}$



Mit folgendem Theorem aus [3] Abschnitt 3 stellen wir die Aussterbewahrscheinlichkeit in Abhängigkeit zum größten Eigenwert der Durchschnittsmatrix dar:

Theorem 3. *Sei $\boldsymbol{X}_n$ positiv regulär und nicht singulär. Sei λ der maximale Eigenwert der Durchschnittsmatrix M. Dann:*

i) Falls $\lambda \leq 1$, so ist $\boldsymbol{a} = (1,1)$. Falls $\lambda > 1$, so ist $\boldsymbol{a} < (1,1)$.

ii) $\lim_{n\to\infty} \boldsymbol{f}_n(\boldsymbol{s}) = \boldsymbol{a}$ für alle $\boldsymbol{s} \in \mathcal{C}$.

iii) $\boldsymbol{a}$ ist die einzige Lösung von $\boldsymbol{f}(\boldsymbol{s}) = \boldsymbol{s}$ für $\boldsymbol{s} \in \mathcal{C}$.

Bei kritischen und subkritischen Prozessen stirbt die Population also fast sicher aus. Außerdem liefert das Theorem eine Möglichkeit, die Aussterbewahrscheinlichkeiten für superkritische Prozesse zu berechnen.

Wenden wir dieses Theorem auf ein Beispiel an. Interessant ist nur der superkritische Fall. Also berechnen wir λ in Abhängigkeit vom Parameter b und setzen $\lambda > 1$ voraus. Wir müssen uns mögliche Reproduktionsgesetze ausdenken, welche die entsprechenden Durchschnitte produzieren:

Beispiel 6.

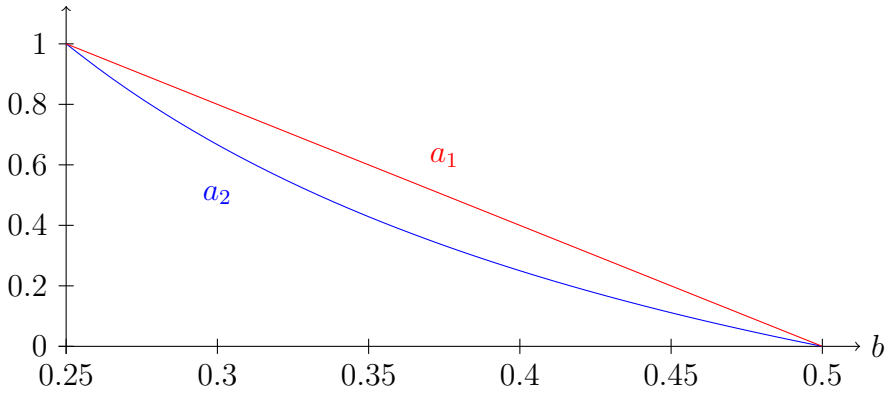
$$
\begin{aligned}
p_1(1,0) = \frac{1}{2}, \quad p_1(0,1) = b, \quad p_1(0,0) = \frac{1}{2} - b \qquad (0 \le b \le \frac{1}{2}) \\
p_2(1,0) = \frac{1}{2}, \quad p_2(1,1) = \frac{1}{2} \\
f^{(1)}(s_1,s_2) = \frac{1}{2} - b + \frac{1}{2}s_1 + bs_2 \\
f^{(2)}(s_1,s_2) = \frac{1}{2}s_1 + \frac{1}{2}s_1 s_2 \\
M = \begin{pmatrix} \frac{1}{2} & b \\ 1 & \frac{1}{2} \end{pmatrix}, \\
0 = (\frac{1}{2} - \lambda)^2 - b \\
\Rightarrow \lambda = \frac{1}{2} + \sqrt{b} \\
\Rightarrow b > \frac{1}{4}, \text{ damit } \lambda > 1.
\end{aligned}
$$

Da $f^{(2)}$ ein Polynom vom Grad größer eins ist, ist der Prozess nicht singulär und M hat offenbar nur positive Einträge, da $b > 0$. Also ist der Prozess auch positiv regulär. Finden wir nun den Fixpunkt $\boldsymbol{a}$ von $\boldsymbol{f}(\boldsymbol{s})$ in $\mathcal{C}$.

$$
\begin{aligned}
&\begin{cases} f^{(1)}(s_1,s_2) &= \frac{1}{2} - b + \frac{1}{2}s_1 + bs_2 = s_1 \\ f^{(2)}(s_1,s_2) &= \frac{1}{2}s_1 + \frac{1}{2}s_1 s_2 = s_2 \end{cases} \\
\Leftrightarrow &\begin{cases} s_1 &= 1 - 2b + 2bs_2 \\ s_2 &= \frac{1}{4b} \pm \frac{1}{4b}\sqrt{1 - 8b + 16b^2} \end{cases}
\end{aligned}
$$

Da a_2 eine Wahrscheinlichkeit sein soll, muss $s_2 \in [0,1]$ sein und da für $b \in (\frac{1}{4}, \frac{1}{2}]$ gilt: $\frac{1}{4b} \in [\frac{1}{2}, 1)$, ist klar, dass wir $s_2 = \frac{1}{4b} - \frac{1}{4b}\sqrt{1 - 8b + 16b^2}$ wählen müssen. Insgesamt gilt für die Aussterbewahrscheinlichkeit also:

$$
(a_1, a_2) = \left(1 - 2b + 2b\Big(\frac{1}{4b} - \frac{1}{4b}\sqrt{1 - 8b + 16b^2}\Big), \frac{1}{4b} - \frac{1}{4b}\sqrt{1 - 8b + 16b^2}\right)
$$



Für $b < \frac{1}{4}$ wissen wir nach Theorem 3, dass die Aussterbewahrscheinlichkeit eins ist. Das stimmt mit unserem Randwert überein. Für $b = \frac{1}{2}$ ist $p_1(0,0) = \frac{1}{2} - b = 0$, somit ist es unmöglich, dass ein Individuum keinen Nachfahren hat und es ist logisch, dass dann die Aussterbewahrscheinlichkeit Null ist.

2.2.4 Grenzwertbetrachtung im subkritischen Fall

Das folgende Theorem aus Abschnitt 4 in [3] nutzt die in Theorem 2 definierten Rechts- und Links-Eigenvektoren $\mathbf{u}$ und $\mathbf{v}$.

Theorem 4. *Für einen Verzweigungsprozess mit $\lambda < 1$ existiert eine Funktion $\gamma : \mathbb{R}^2 \to \mathbb{R}$, sodass gilt:*

$$\frac{\boldsymbol{v} \cdot [(1,1) - \boldsymbol{f}_n(\boldsymbol{s})]}{\lambda^n} \downarrow \gamma(\boldsymbol{s}) \geq 0, \text{ für } n \to \infty \text{ und } \boldsymbol{s} \in \mathcal{C}. \tag{2}$$

Hierbei ist $\gamma(\cdot)$ monoton fallend und > 0, genau dann, wenn $\mathbb{E}\Big[\|\boldsymbol{X}_1\| \log \|\boldsymbol{X}_1\|\Big] < \infty$.

$$\lim_{n\to\infty} \frac{(1,1) - \boldsymbol{f}_n(\boldsymbol{s})}{\lambda^n} = \gamma(\boldsymbol{s})\boldsymbol{u}, \tag{3}$$

und:

$$\lim_{n\to\infty} \mathbb{P}(\boldsymbol{X}_n \neq (0,0) | \boldsymbol{X}_0 = \boldsymbol{k}) = \gamma(0,0)(\boldsymbol{k} \cdot \boldsymbol{u}). \tag{4}$$

Im subkritischen Fall wissen wir bereits, dass der Prozess fast sicher aussterben wird. Dieses Theorem liefert eine exponentielle Konvergenzrate der erzeugenden Funktion mit $\mathbf{f}_n(\mathbf{s}) \sim (1,1) + \lambda^n \gamma(\mathbf{s})\mathbf{u}$. Da die Verteilung von $\mathbf{X}_n$ durch $\mathbf{f}_n$ eindeutig bestimmt ist, erlaubt uns dies folgende Interpretation: $\mathbf{X}_n$ konvergiert in Verteilung exponentiell gegen eine entartete Variable mit der erzeugenden Funktion $(1,1)$. Eine Variable mit dieser erzeugenden Funktion ist fast sicher Null, denn sie ist konstant und somit ist jeder Moment der Variable Null.

2.2.5 Grenzwertbetrachtung im kritischen Fall

Für unsere weitere Betrachtung müssen wir zunächst den Vektor aus Matrizen $\mathbf{Q} = (Q_1, Q_2)$ einführen, wobei Q_l die zweiten Momente des Typs l beinhaltet:

Definition 8. $\boldsymbol{Q} = (Q_1, Q_2)$*, hierbei sind $Q_l \in \mathbb{R}^{2,2}$ Matrizen mit Einträgen:*

$$\begin{aligned}(Q_l)_{i,j} =& q_l(i,j) = \mathbb{E}[Y^{(i)}_{n,k,l} Y^{(j)}_{n,k,l} - \delta_{i,j} Y^{(i)}_{n,k,l}] \\ \overset{*}{=}& \frac{\partial^2 f^{(l)}}{\partial s_i \partial s_j}(1,1).\end{aligned}$$

Außerdem definieren wir die quadratische Form:

$$Q_l(\boldsymbol{s}) = (s_1, s_2) Q_l \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \frac{1}{2} \sum_{i,j=1}^{2} s_i q_l(i,j) s_j$$

Somit ist $\boldsymbol{Q}(\boldsymbol{s}) = (Q_1(\boldsymbol{s}), Q_2(\boldsymbol{s}))$ eine Funktion von $\mathbb{R}^2$ nach $\mathbb{R}^2$.

Rechnen wir die mit $*$ markierte Gleichung nach:

$$q_l(i,j) = \mathbb{E}[Y^{(i)}_{n,k,l} Y^{(j)}_{n,k,l} - \delta_{i,j} Y^{(i)}_{n,k,l}] = \sum_{k_1,k_2=0}^{\infty} k_i k_j p_l(k_1,k_2) - \delta_{i,j} k_i p_l(k_1,k_2).$$

Um $f^{(l)}$ abzuleiten, müssen wir die Fälle unterscheiden: Für $i \neq j$ können wir aufgrund der Symmetrie $i = 1, j = 2$ voraussetzen und:

$$\frac{\partial^2 f^{(l)}}{\partial s_1 \partial s_2} = \frac{\partial^2}{\partial s_1 \partial s_2} \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} p_l(k_1, k_2) s_1^{k_1} s_2^{k_2} = \sum_{k_1,k_2=1}^{\infty} k_1 k_2 p_l(k_1, k_2) s_1^{k_1-1} s_2^{k_2-1}.$$

$$\frac{\partial^2 f^{(l)}}{\partial s_1 \partial s_2}(1,1) = \sum_{k_1,k_2=0}^{\infty} k_1 k_2 p_l(k_1, k_2).$$

Hier ist die Gleichheit sofort zu sehen. Nehmen wir $i = j$ und $i = 1$ an. Der Fall $i = 2$ läuft dann analog:

$$\frac{\partial^2 f^{(l)}}{\partial s_1^2} = \frac{\partial^2}{\partial s_1^2} \sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} p_l(k_1, k_2) s_1^{k_1} s_2^{k_2} = \sum_{k_1=2}^{\infty} \sum_{k_2=0}^{\infty} k_1(k_1-1) p_l(k_1, k_2) s_1^{k_1-2} s_2^{k_2}.$$

$$\frac{\partial^2 f^{(l)}}{\partial s_1^2}(1,1) = \sum_{k_1=2}^{\infty} \sum_{k_2=0}^{\infty} k_1(k_1-1) p_l(k_1, k_2)$$

Durch Ausklammern erkennen wir auch hier die Gleichheit.

Beispiel 7. *Nehmen wir die Reproduktionsgesetze aus Beispiel 6 und berechnen $\boldsymbol{Q}$:*

$$\begin{aligned}
p_1(1,0) = \frac{1}{2}, \quad p_1(0,1) = b, \quad p_1(0,0) = \frac{1}{2} - b \qquad (0 \leq b \leq \frac{1}{2}) \\
p_2(1,0) = \frac{1}{2}, p_2(1,1) = \frac{1}{2} \\
f^{(1)}(s_1, s_2) = \frac{1}{2} - b + \frac{1}{2} s_1 + b s_2 \\
f^{(2)}(s_1, s_2) = \frac{1}{2} s_1 + \frac{1}{2} s_1 s_2 \\
M = \begin{pmatrix} \frac{1}{2} & b \\ 1 & \frac{1}{2} \end{pmatrix} \\
\lambda = \frac{1}{2} + \sqrt{b}
\end{aligned}$$

Da wir hier den subkritischen Fall betrachten, setzen wir $b < \frac{1}{4}$ voraus. Die Matrizen Q_1 und Q_2 sind nichts weiter als die Hessematrizen für $f^{(1)}$ respektive $f^{(2)}$.

$$H_{f_1} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = Q_1, H_{f_2} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} = Q_2.$$

Somit ist $Q_1(\boldsymbol{s}) = 0$ und $Q_2(s_1, s_2) = \frac{1}{2} s_1 s_2$. Insgesamt also $\boldsymbol{Q}(\boldsymbol{s}) = (0, \frac{1}{2} s_1 s_2)$.

Folgendes Theorem aus [3] Abschnitt 5 nutzt die zuvor definierten Eigenvektoren und das gerade eingeführte $\mathbf{Q}$, um das asymptotische Verhalten des Prozesses zu beschreiben:

Theorem 5. *Sei ein multipler Verzweigungsprozess mit $\lambda = 1$ und $\mathbb{E}[\|\boldsymbol{X}_1\|^2] < \infty$ gegeben. Seien $\boldsymbol{u}$ und $\boldsymbol{v}$ die Rechts-und Links-Eigenvektoren der Durchschnittsmatrix M. Für $\boldsymbol{w} \in \mathbb{R}^2$ mit $\boldsymbol{w} \cdot \boldsymbol{v} > 0$ konvergiert $\frac{1}{n}(\boldsymbol{X}_n \cdot \boldsymbol{w})$ bedingt auf das Nichtaussterben $\boldsymbol{X}_n \neq 0$ in n zu einer Zufallsvariable mit Verteilungsfunktion:*

$$f(x) = \begin{cases} \frac{1}{\rho} e^{-\frac{x}{\rho}}, \text{ für } x \geq 0 \\ 0 \text{ für } x < 0 \end{cases} \tag{5}$$

$$\text{wobei } \rho = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\boldsymbol{v} \cdot \boldsymbol{Q}(\boldsymbol{u})} \tag{6}$$

Auch im kritischen Fall stirbt die Bevölkerung fast sicher aus, jedoch liefert das Theorem interessante Ergebnisse, wenn man den Prozess darauf bedingt, nicht auszusterben. Hier konvergiert $\frac{1}{n}(\mathbf{X}_n \cdot \mathbf{w})$ in Verteilung gegen eine Zufallsvariable Z, wobei Z die Verteilungsfunktion f hat. Z ist exponentialverteilt mit Parameter $\frac{1}{\rho}$ und Erwartungswert ρ.

2.2.6 Exponentielles Wachstum im superkritischen Fall

Im superkritischen Fall wissen wir bisher nur, dass der Prozess nicht fast sicher ausstirbt. In den meisten Fällen bleibt die Wahrscheinlichkeit, dass der Prozess ausstirbt auch für große λ nicht Null. Wir können aber zumindest eine Grenzwertbetrachtung der Variable $\mathbf{X}_n$ durchführen und feststellen, dass sie mit exponentieller Geschwindigkeit explodiert. Genauer: die Variable $\frac{\mathbf{X}_n}{\lambda^n}$ konvergiert gegen eine Zufallsvariable, die unter gewissen Bedingungen nicht entartet ist. Das Theorem aus [3] Abschnitt 6 beschreibt den Sachverhalt genau:

Theorem 6. *Sei $\boldsymbol{X}_n$ ein superkritischer nicht singulärer und positiv regulärer Verzweigungsprozess mit Durchschnittsmatrix M. Sei $\lambda > 1$ der größte Eigenwert von M und $\boldsymbol{v}$ der Linkseigenvektor. Dann existiert eine nicht negative Zufallsvariable W, sodass:*

$$\lim_{n\to\infty} \frac{\boldsymbol{X}_n}{\lambda^n} = \boldsymbol{v}W \text{ a.s. .} \tag{7}$$

Weiterhin gilt:

$$\mathbb{P}(W > 0) > 0 \tag{8}$$

genau dann, wenn:

$$\mathbb{E}[Y^{(l)}_{1,1,j} \log Y^{(l)}_{1,1,j}] < \infty \text{ für alle } j, l \in \{1, 2\} \tag{9}$$

Wenden wir dieses Theorem auf den Prozess aus Beispiel 6 an. Hierbei ist $b \in (\frac{1}{4}, \frac{1}{2}]$, sodass der Prozess superkritisch ist

Beispiel 8.

$$\begin{aligned} p_1(1,0) = \frac{1}{2}, \quad p_1(0,1) = b, \quad p_1(0,0) = \frac{1}{2} - b \\ p_2(1,0) = \frac{1}{2}, \quad p_2(1,1) = \frac{1}{2} \\ \lambda = \frac{1}{2} + \sqrt{b} \end{aligned}$$

Schreibe verkürzend $Y^{(l)}_j$ für $Y^{(l)}_{1,1,j}$ und prüfe (9):

$$\begin{aligned} \mathbb{E}[Y^{(j)}_l \log Y^{(j)}_l] &= \sum_{k_1,k_2=0}^{\infty} k_j \log(k_l) p_j(k_1, k_2). \\ \mathbb{E}[Y^{(1)}_1 \log Y^{(1)}_1] &= \frac{1}{2} \cdot 1 \log(1) + b \cdot „0 \log(0)“ = 0 \\ \mathbb{E}[Y^{(2)}_1 \log Y^{(2)}_1] &= b \cdot 1 \log(1) = 0 \\ \mathbb{E}[Y^{(1)}_2 \log Y^{(1)}_2] &= \frac{1}{2} \cdot 1 \log(1) + \frac{1}{2} \cdot 1 \log(1) = 0 \\ \mathbb{E}[Y^{(2)}_2 \log Y^{(2)}_2] &= \frac{1}{2} \cdot 1 \log(1) = 0. \end{aligned}$$

$\Rightarrow$(9) *ist erfüllt und somit gilt Theorem 6 mit* $\mathbb{P}(W > 0) > 0$.

Somit divergiert $\boldsymbol{X}_n$ in exponentieller Geschwindigkeit: $\boldsymbol{X}_n \sim \lambda^n \boldsymbol{v} W$, wobei W mit positiver Wahrscheinlichkeit größer als 0 ist.

Nur die Teilmenge des Ereignisraums Ω ist interessant, auf der $\{W > 0\}$ gilt. Für diese Teilmenge liefert das Theorem eine exakte Explosionsrate λ^n. Im anderen Fall kann die Bevölkerung $\mathbf{X}_n$ durchaus auch explodieren, nur nicht so schnell wie λ^n.

Beispiel 9. *Folgendes ist ein Beispiel für einen Verzweigungsprozess, der* (9) *nicht erfüllt. Damit es möglichst unkompliziert bleibt, konstruieren wir einen Multitypen Verzweigungsprozess, der sich wie ein Monotyper Verzweigungsprozess lesen lässt. Das schaffen wir, indem wir festlegen, dass Individuen vom Typ 1 nur Kinder vom Typ 1 zeugen können:*

$$\begin{aligned} &p_2(i,j) = 0, \text{ für alle } i,j \in \mathbb{N}_0 \\ &p_1(0,0) = p, \quad p_1(1,0) = q, \quad p_1(k,0) = \frac{1}{k^2 \log(k)^2}, \text{ für } k \geq 2 \\ &p + q + \sum_{k=2}^{\infty} p_1(k,0) = 1 \end{aligned}$$

Dass die Reihe $\sum_{k=2}^{\infty} p_1(k,0)$ konvergiert, folgt mithilfe des Vergleichskriteriums daraus, dass $\mathbb{E}[|Y_1^{(1)}|] < \infty$, was wir gleich zeigen werden. p und q müssen dann so gewählt werden, dass p_1 ein Wahrscheinlichkeitsmaß ist. Da wir für jeden unserer Verzweigungsprozesse voraussetzen, dass die durchschnittliche Anzahl von Kindern $m_{i,j}$ für jeden Typ existieren, zeigen wir $\mathbb{E}[|Y_1^{(1)}|] < \infty$.

Dafür nutzen wir das Cauchy-Verdichtunskriterium (siehe [8]), welches besagt:

Für eine monoton fallende Folge nichtnegativer Zahlen (a_k) gilt:

$$\sum_k a_k < \infty \Leftrightarrow \sum_k 2^k a_{2^k} < \infty \tag{10}$$

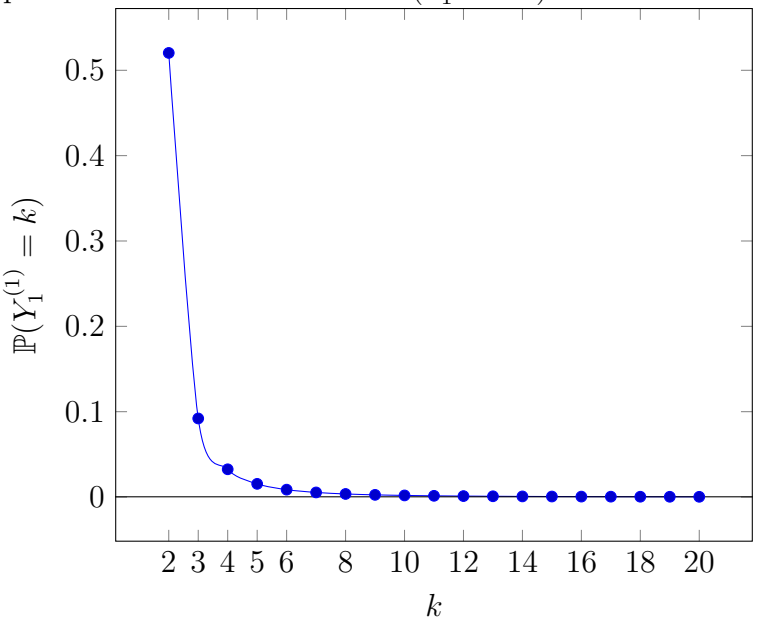
Wir zeigen $\mathbb{E}[|Y_1^{(1)}|] < \infty$, aber $\mathbb{E}[|Y_1^{(1)} \log(Y_1^{(1)})|] = \infty$:

$$\begin{aligned} \mathbb{E}[|Y_1^{(1)}|] = q + \sum_{k=2}^{\infty} \frac{1}{k \log(k)^2} &< \infty \\ \Leftrightarrow \sum_{k=2}^{\infty} \frac{2^k}{2^k \log(2^k)^2} = \sum_{k=2}^{\infty} \frac{1}{k^{(2)} \log(2)^2} &< \infty \ \ w.A. \\ \mathbb{E}[|Y_1^{(1)} \log(Y_1^{(1)})|] = \sum_{k=2}^{\infty} \frac{1}{k \log(k)} &= \infty \\ \Leftrightarrow \sum_{k=2}^{\infty} \frac{1}{\log(2^k)} = \sum_{k=2}^{\infty} \frac{1}{k \log(2)} &= \infty \ \ w.A. \end{aligned}$$

Somit ist (9) *für $j = l = 1$ nicht erfüllt, was bedeutet, dass es keine Teilmenge vom Ereignisraum Ω mit Maß größer Null gibt, auf der unser gerade konstruierter Verzweigungsprozess so schnell wie λ^n explodiert. Numerisch angenähert entspricht $\lambda = \mathbb{E}[|Y_1^{(1)}|] \approx 2 + q$.*

Die Äquivalenz von (8) und (9) erscheint unintuitiv: um (9) nicht zu erfüllen, müssen, wie in Beispiel 9, viele Kinder mit positiver Wahrscheinlichkeit gezeugt werden können. Intuitiv würde das auch bedeuten, dass der Prozess dann auch schneller explodiert. Das

Gegenteil ist jedoch der Fall. Das liegt daran, dass zumindest immer $\mathbb{E}[|Y_l^{(j)}|] < \infty$ gefordert wird. Dadurch müssen die Wahrscheinlichkeiten für viele Kinder rapide abfallen. In Beispiel 9 ist dies auch der Fall: $\mathbb{P}(Y_1^{(1)} < 2) \approx 0.3$ und:



Ein ausführlicher Beweis für die monotype Version dieses Theorems findet sich in [1] auf den Seiten 18 bis 26. Dort wird auch ein weiteres Argument vorgebracht, welches die Zusammenhänge erklärt: Betrachte man zwei Prozesse mit gleichen Parameter λ und fordere man für den zweiten Prozess, dass die Wahrscheinlichkeiten für viele Kinder größer sind als beim ersten. Dann ist klar, dass auch die Wahrscheinlichkeiten für ein oder keine Kinder größer sein müssen, als beim ersten Prozess, damit der gleiche Durchschnitt λ zustande kommt.

Wenn beide Prozesse $\mathbb{E}[|Y_l^{(j)}|] < \infty$ erfüllen und der zweite (9) nicht erfüllt, so muss beim zweiten Prozess auch die Wahrscheinlichkeit für wenig Kinder oder das Aussterben höher sein. Somit ist die Wahrscheinlichkeit, dass der Prozess so schnell wächst wie λ^n für $n \to \infty$ Null.

3 Ein multityper Verzweigungsprozess als Modell zur Untersuchung der Ausbreitung von Covid-19

Nachdem wir in der ersten Hälfte der Arbeit die theoretischen Grundlagen geschaffen haben, wenden wir diese nun auf das Beispiel der Corona Pandemie an. Hier soll anhand der Fallzahlen die Reproduktionsrate der Krankheit geschätzt werden. Wir erklären zuerst das Modell, stellen noch einmal die Bedeutung der Reproduktionsrate heraus und stellen Schätzer für diese auf. Außerdem schätzen wir andere interessante Parameter. Schließlich wenden wir die Schätzer auf die Daten von Deutschland an und diskutieren die Ergebnisse.

3.1 Einführung des Modells

In seinem Artikel [17] stellt Yanev ein Modell für die Verbreitung des Coronavirus' vor. Mit dessen Hilfe möchte er die Reproduktionsrate und das Verhältnis zwischen behördlich gemeldeten Fällen und unbekannten Krankheitsfällen schätzen.

Ein Verzweigungsprozess soll die Zahl der infizierten Menschen $\mathbf{X}_n$ in einem Land modellieren. Hierbei unterscheiden wir zwei Typen: $\mathbf{X}_n = (X_n^{(1)}, X_n^{(2)})$. Individuen vom Typ 2 sind Menschen, die am Tag n als krank diagnostiziert und gemeldet wurden. Die Zahl dieser Menschen lässt sich erfassen. Genauer: $X_n^{(2)}$ ist die Zahl der am Tag n gemeldeten *neuen* Covid-19 Fälle in einem Land - nicht zu verwechseln mit der monoton wachsenden Gesamtzahl der Fälle in einem Land.

Menschen vom Typ 1 sind nicht diagnostizierte Infizierte. Also Menschen, die die Krankheit in sich tragen (damit also auch andere anstecken können), von denen aber nicht bekannt ist, dass sie das Virus haben. Sie können einen asymptomatischen Krankheitsverlauf haben, oder ihr Test wurde noch ausgewertet, etc. Ihre Zahl kann nicht gemessen, sondern nur vermutet werden. $X_n^{(1)}$ bezeichnet also die Zahl der Menschen, die am Tag n das Virus in sich tragen, ohne dass sie diagnostiziert wurden.

Sei $Y_{n,k,j}^{(l)}$ die Zahl der „Nachfahren" vom Typ l, die das k-te Individuum vom Typ j in der n-ten Generation hat. „Nachfahren" sind hier Personen, die von diesem Individuum angesteckt wurden. Bleibt die Person selbst krank, so wird sie im nächsten Schritt als ihr eigener „Nachfahre" gewertet. Es errechnet sich die Menge der kranken Individuen am nächsten Tag durch:

$$X_{n+1}^{(1)} = \sum_{k=1}^{X_n^{(1)}} Y_{n,k,1}^{(1)} + \sum_{k=1}^{X_n^{(2)}} Y_{n,k,2}^{(1)}$$
$$X_{n+1}^{(2)} = \sum_{k=1}^{X_n^{(1)}} Y_{n,k,1}^{(2)} + \sum_{k=1}^{X_n^{(2)}} Y_{n,k,2}^{(2)}$$

Für feste l und j sind $Y_{n,k,j}^{(l)}$ identisch unabhängig voneinander verteilt. Daher lassen wir, wenn wir über ihre Verteilung sprechen, die unrelevanten Indizes n und k aus.

Sprechen wir über die Reproduktionsgesetze: Wir gehen davon aus, dass Individuen vom Typ 2 aus dem Prozess der Ausbreitung des Virus' entfernt werden, weil sie sich in Quarantäne begeben oder sterben. Daher ist $\mathbf{Y}_2 = (0, 0)$.

$$X_{n+1}^{(1)} = \sum_{k=1}^{X_n^{(1)}} Y_{n,k,1}^{(1)}$$
$$X_{n+1}^{(2)} = \sum_{k=1}^{X_n^{(1)}} Y_{n,k,1}^{(2)}$$

Interessant sind also nur die Reproduktionsgesetze der Individuen vom Typ 1:

Ein infizierter Mensch von Typ 1 kann an jedem Tag 0 bis $B-1$ andere Menschen infizieren. Diese sind zunächst vom Typ 1, da sie selbst noch nicht wissen, dass sie infiziert wurden. Dann ist $Y_1^{(1)} \in \{1, ..., B\}$, da das Individuum selbst auch noch krank bleiben kann. Wir legen hierbei $B \in \mathbb{N}$ nicht fest, wissen aber, dass B durch die Zahl der Menschen, die in dem betrachteten Land leben, begrenzt ist.

Außerdem kann der infizierte Mensch mit der Wahrscheinlichkeit p_0 selbst gesund werden ($\mathbf{Y}_1 = (0,0)$), oder erfahren, dass er krank ist, wodurch er zu Typ 2 wird ($\mathbf{Y}_1 = (0,1)$). Die Wahrscheinlichkeit dafür bezeichnen wir mit q. Somit ergibt sich für die Reproduktionsgesetze:

$$\mathbb{P}(\mathbf{Y}_1 = (i_1, i_2)) = p_1(i_1, i_2) \text{ für } i_1, i_2 \in \mathbb{N}_0$$
$$p_1(0,0) = p_0; \quad p_1(j,0) = p_j, \text{ für } j = 1, ..., B; \quad p_1(0,1) = q = 1 - \sum_{j=0}^{B} p_j$$

Anmerkung 1. *Wir stellen fest:* $\mathbb{P}(Y_1^{(2)} = 0) = 1 - q$ *und* $\mathbb{P}(Y_1^{(2)} = 1) = q$. *Ob ein nicht diagnostiziertes infiziertes Individuum im nächsten Schritt gemeldet wird, ist Bernoulliverteilt zum Parameter* q. *Somit ist die Zahl der gesamt als krank gemeldeten Menschen am* $n+1$*-ten Tag Binominial-verteilt zu den Parametern* $X_n^{(1)}$ *und* q*:*

$$\mathbb{P}(X_{n+1}^{(2)} = i | X_n^{(1)} = l) = \binom{l}{i} q^i (1-q)^{l-i}, \quad i = 0, 1, ..., l; \quad l = 0, 1, 2, ...$$

q kann also als Quote zwischen den am Tag n gemeldeten Fallzahlen und der Zahl der nicht diagnostizierten Infizierten am Vortag gesehen werden.

Der Prozess startet am Tag 0 mit $X_0^{(1)} > 0$ nicht diagnostizierten Erkrankten und ohne Individuen vom Typ 2. In unseren Betrachtungen werden wir von $X_0^{(1)} = m_0 \in \mathbb{N}$ ausgehen.

3.2 Die Reproduktionsrate R

Öffentlich wurde schon viel darüber diskutiert, welche Bedeutung die Reproduktionsrate R hat und dass es wichtig ist, dass sie kleiner als Eins ist. In [9] ist ein Ausschnitt einer Rede von Angela Merkel verlinkt, in dem sie über die Implikationen des Wertes spricht. In diesem Abschnitt liefern wir die Implikationen des Wertes von R für den gerade konstruierten Verzweigungsprozess. Wir stellen fest, dass dieser Wert dem im Abschnitt *Grundlagen* hervorgehobenen Parameter λ entspricht und wiederholen die Konsequenzen, die wir in diesem Abschnitt herausgestellt haben für λ beziehungsweise R größer, kleiner und gleich eins.

Definition 9. *Sei* R *die durchschnittliche Anzahl von Typ 1 Infizierten, die von einem Typ 1 Individuum an einem Tag erzeugt werden:*

$$R = \mathbb{E}[Y_1^{(1)}] = \sum_{j=1}^{B} j p_j$$

Stellen wir die Durchschnittsmatrix auf:

$$M = \begin{pmatrix} R & q \\ 0 & 0 \end{pmatrix} \tag{11}$$

Betrachten wir die offensichtlichen Eigenwerte 0 und R, so stellen wir fest, dass der größte Eigenwert der Durchschnittsmatrix $\lambda = R$ entspricht. Dies erlaubt uns die Einteilung des Prozesses in subkritisch für $R < 1$, bei $R = 1$ kritisch und bei $R > 1$ superkritisch. Die Implikationen der verschiedenen Klassifikationen wurden im Grundlagenbereich diskutiert:

- Für $R \leq 1$ stirbt der Prozess fast sicher aus
- Für $R < 1$ strebt der Prozess in exponentieller Geschwindigkeit in Verteilung gegen Null.
- Im superkritischen Fall explodiert die Zahl der Infizierten in Geschwindigkeit R^n auf einem nicht trivialen Teil des Wahrscheinlichkeitsraums, falls: $\mathbb{E}[|Y_1^{(1)} \log(Y_1^{(1)})|] < \infty$ (9).

Bedingung (9) ist erfüllt, da $|Y_1^{(1)}|$ nach oben durch B beschränkt ist.

Außerdem können wir die Bedeutung von R für den durchschnittlichen Verlauf des Prozesses sehr einfach herausstellen:

$$M^n = \begin{pmatrix} R^n & q \cdot R^{n-1} \\ 0 & 0 \end{pmatrix}$$

Definieren wir mit $M^{(i)}(n)$ die erwartete Anzahl an Typ i Infizierten am n-ten Tag. Es gilt:

$$\mathbb{E}[(X_n^{(1)}, X_n^{(2)})] = (M^{(1)}(n), M^{(2)}(n)) = (M^{(1)}(0), M^{(2)}(0))M^n$$

Da wir von $\mathbf{X}_0 = (m_0, 0)$ ausgehen, ergibt sich:

$$\mathbb{E}[(X_n^{(1)}, X_n^{(2)})] = (M^{(1)}(n), M^{(2)}(n)) = (m_0, 0) \begin{pmatrix} R^n & q \cdot R^{n-1} \\ 0 & 0 \end{pmatrix} = (m_0 \cdot R^n, m_0 \cdot q \cdot R^{n-1})$$

Mit $M^{(1)}(n) = m_0 \cdot R^n$ können wir einfach nachzuvollziehende Aussagen über den durchschnittlichen Verlauf des Prozesses treffen: Für $R < 1$ fällt der Durchschnitt der Infizierten exponentiell auf Null, während er für $R > 1$ in exponentieller Geschwindigkeit explodiert. Im kritischen Fall bleibt der Erwartungswert konstant.

3.3 Ermittlung der Schätzer für R und weiterer interessanter Größen

In diesem Abschnitt werden wir verschiedene Möglichkeiten diskutieren, die Reproduktionsrate zu schätzen. Außerdem überlegen wir, wie wir mithilfe von R die Zahl der nicht gemeldeten Fälle ermitteln können.

3.3.1 Harris Schätzer für R

Nach unseren Berechnungen im vorherigen Abschnitt ist:

$$R = \frac{m_0 \cdot R^{n+1}}{m_0 \cdot R^n} = \frac{M^{(2)}(n+1)}{M^{(2)}(n)} = \frac{\mathbb{E}[X_{n+1}^{(2)}]}{\mathbb{E}[X_n^{(2)}]}$$

Harris entwickelte in [6] für monotype Verzweigungsprozesse einen Schätzer für den Parameter λ: Sind in einem monotypen Verzweigungsprozess die Werte $X_1, .., X_{n+1}$ bekannt, so ist der Maximum-Likelyhood-Schätzer (MLE) für λ:

$$\lambda_n = \frac{\sum_{i=2}^{n+1} X_i}{\sum_{i=1}^{n} X_i}$$

Dieser Schätzer ergibt intuitiv sehr viel Sinn: Als Schätzer für den Erwartungswert für $X_{n+1}^{(2)}$ wird das Mittel über die Werte $X_2^{(2)}, ..., X_{n+1}^{(2)}$ genommen und als Schätzer für den Erwartungswert für $X_n^{(2)}$ das Mittel über $X_1^{(2)}$ bis $X_n^{(2)}$.

Für unser Modell ergibt sich als Schätzer:

$$\tilde{R}_n = \frac{\sum_{i=2}^{n+1} X_i^{(2)}}{\sum_{i=1}^{n} X_i^{(2)}} \text{ für } n = 1, 2, ..., N \tag{12}$$

Wir schätzen also R für jeden Tag, wobei jeweils die Tage 1 bis $n+1$ einbezogen werden. Hierbei soll N (die Zahl der betrachteten Tage) groß genug sein und es ist zu beachten, dass die Schätzer für kleine n unpräzise sind.

Harris stellte außerdem fest, dass dieser Schätzer im superkritischen Fall und bedingt auf das Nicht-Aussterben konsistent ist. Also: unabhängig vom tatsächlichen Wahrscheinlichkeitsmaß des Modells ist der geschätzte Wert fast sicher sehr nah am zu schätzenden Wert, solange man eine beliebig große Stichprobe zulässt.

Diese Eigenschaft ist für unseren Fall leider nicht sehr nützlich. Die Reproduktionsrate ist nicht konstant, sondern ändert sich im Verlauf der Zeit: zum Beispiel, weil im betrachteten Land Quarantänemaßnahmen erlassen bzw. aufgehoben werden. Auch die Urlaubssaison oder große Events können die Reproduktionsrate verändern. Welche Maßnahmen R wie genau beeinflussen, soll jedoch nicht Teil dieser Arbeit sein.

Da der Harris Schätzer ein Maximum-Likelyhood-Schätzer ist, ist $\tilde{R}_n$ für große n näherungsweise normalverteilt um den wahren Wert R. Die obere und untere Grenze des 95%-Konfidenzintervalls um $\tilde{R}_n$ ist somit:

$$\tilde{R}_n^{\pm} = \tilde{R}_n \pm z_{1-\frac{0.05}{2}} \frac{\sigma}{\sqrt{n}} \tag{13}$$

Hierbei ist $z_{1-\frac{0.05}{2}} = z_{0.975} = 1,96$ das 0.975-Quantil der Standardnormalverteilung und σ die Wurzel der Varianz des Schätzers. Da wir die Wahrscheinlichkeitsverteilung der Daten nicht kennen, muss auch σ geschätzt werden mit:

$$\tilde{\sigma}_n^2 = s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}(\tilde{R}_i - \bar{\tilde{R}}_n)^2$$

Hierbei ist $\bar{\tilde{R}}_n$ der statistische Mittelwert des Schätzers:

$$\bar{\tilde{R}}_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{R}_i$$

Einen tieferen Einblick in das Erstellen von Konfidenzintervallen für unbekannte Wahrscheinlichkeitsverteilungen findet man in [4].

Da R zeitabhängig ist, können $\tilde{R}_n$ mit großen n nicht immer näher an einen tatsächlichen konstanten R Wert kommen. Stattdessen wollen wir dynamischere Schätzer erstellen. Daher betrachten wir außerdem den Harris Type Schätzer, der nur die letzten L Daten benutzt:

$$\tilde{R}_{n,L} = \frac{\sum_{i=n-L+1}^{n+1} X_i^{(2)}}{\sum_{i=n-L}^{n} X_i^{(2)}} \text{ für } n = L, L+1, ..., N \tag{14}$$

Diesen werden wir für $L = 7, 14$ und 28 verwenden.
In Deutschland wurden zwischen dem 13.02.2019 und 25.02.2019 keine neuen Fälle gemeldet. Daher müssen wir für den Schätzer $\tilde{R}_{n,L=7}$ Sonderfälle einfügen: In den Fällen, in denen der Schätzer erfordert, durch Null zu teilen, ergänzen wir die Werte mit $\tilde{R}_{n,L} = 1$. Außerdem können wir auch $\tilde{R}_{n,L} = 0$ nicht zulassen. An dieser Stelle ersetzen wir den Schätzer ebenfalls mit 1.

Da diese Schätzer aus dem Harris Schätzer hervorgehen, leiten wir auch die Grenzen der Konfidenzintervalle analog ab:

$$\tilde{R}_{n,L}^{\pm} = \tilde{R}_{n,L} \pm z_{0.975}\frac{s_{n,L}}{\sqrt{L}} \tag{15}$$

mit:

$$s_{n,L}^2 = \frac{1}{L-1}\sum_{i=n-L+1}^{n}(\tilde{R}_{i,L} - \bar{\tilde{R}}_{n,L})^2$$

und:

$$\bar{\tilde{R}}_{n,L} = \frac{1}{L}\sum_{i=n-L+1}^{n}\tilde{R}_{i,L}.$$

3.3.2 Wie das RKI die Reproduktionsrate berechnet

Das Robert Koch-Institut beruft sich in seiner Methodik auf den von Matthias an der Heiden und Osamah Hamouda am 22.04.2020 veröffentlichten Artikel. (Siehe: [2]).

Die Methodik wird in [10] folgendermaßen zusammengefasst:

> „Das Verfahren besteht aus drei Schritten:
>
> 1. Multiple Imputation fehlender Information zum Erkrankungsbeginn von COVID-19- Fällen unter einer Missing-at-Random Annahme
> 2. Korrektur der Anzahl von Neuerkrankungen für den Diagnose-, Melde- und Übermittlungsverzug mittels des Nowcasting-Verfahren
> 3. Berechnung der zeitlich variierenden Reproduktionszahl unter der Annahme einer Generationszeit von 4 Tagen“

Gehen wir genauer auf Schritt 3 ein und betrachten, wie die Reproduktionsrate vom RKI berechnet wird. Den Schätzer des RKI für die Reproduktionsrate am Tag n nennen wir $R_{RKI,n}$. Aus den ersten beiden Schritten ergeben sich E_n, die geschätzten Zahlen der Neuerkrankten am Tag n. Es wird davon ausgegangen, dass ein krankes Individuum am Tag $n-4$ im Durchschnitt $R_{RKI,n}$ Mitmenschen infiziert, welche alle genau am Tag n als krank gemeldet werden. Diese Annahme beruht auf dem sogenannten *seriellen Intervall*, welches in [12] anschaulich erklärt wird: Die Inkubationszeit gibt den Zeitintervall zwischen Ansteckung und Beginn der Erkrankung an. Das serielle Intervall beschreibt hingegen die Zeit zwischen der Erkrankung eines Individuums und der Erkrankung eines von diesem Individuum angesteckten Falles. Im Median schätzt das Robert Koch-Institut diesen Wert auf 4 Tage.

Daraus ergibt sich zunächst die Formel:

$$R_{RKI,n} = \frac{E_n}{E_{n-4}}$$

Diese Reproduktionsrate wird als zu instabil bewertet und daher nicht weiter verwendet. Statt dessen wird ein gleitendes Mittel über vier oder sieben Tage betrachtet:

$$R_{RKI,n,4} = \frac{\bar{E}^4_n}{\bar{E}^4_{n-4}} = \frac{\sum_{i=n-3}^{n} E_i}{\sum_{i=n-3}^{n} E_{i-4}} \tag{16}$$

$$R_{RKI,n,7} = \frac{\bar{E}^7_n}{\bar{E}^7_{n-4}} = \frac{\sum_{i=n-6}^{n} E_i}{\sum_{i=n-6}^{n} E_{i-4}} \tag{17}$$

Obwohl das RKI in seinen Artikeln stets beide Werte veröffentlicht, wird von den meisten Nachrichtenagenturen primär der 4-Tage-R-Wert diskutiert.

Die Reproduktionsrate bezieht sich stets auf ein Intervall, der Tage in der Vergangenheit liegt. $R_{RKI,n,4}$ bezieht sich auf die Tage $n-7$ bis $n-4$. Wenn man dazu die Inkubationszeit von 4 bis 6 Tagen in Betracht zieht, so beschreibt die Reproduktionszahl $R_{RKI,n,4}$ am Tag n das Verhalten der Pandemie in den Tagen $n-13$ bis $n-8$ und die stabilere Variante $R_{RKI,n,7}$ bezieht sich auf die Tage $n-16$ bis $n-8$.

Die Reproduktionsrate des RKI ist also mit starker Verzögerung zu betrachten. Im Gegensatz dazu versucht unsere Methode, die Reproduktionszahl für den direkten Vortag zu messen. Dank der vorhergehenden Glättung der Zahlen für die Neuinfektionen ist die geschätzte Reproduktionsrate weniger anfällig für Schwankungen. Diese Glättung nehmen

wir nicht vor, dafür verwendet unser Schätzer mehr Datensätze, was ihn stabiler macht - teilweise auch zu stabil. Denn der Harris Schätzer, der alle bisherigen Daten verwendet, geht nicht darauf ein, dass die Reproduktionsrate der Krankheit nicht gleich bleibt, sondern sich den Gegebenheiten entsprechend stetig und ständig verändert. Daher betrachten wir auch Versionen des Schätzers, die nur die letzten L Tage berücksichtigen.

3.3.3 Schätzung der Dunkelziffer

Wie zuvor erklärt, ergibt sich unter der Annahme einer konstanten Reproduktionsrate und $\mathbf{X}_0 = (m_0, 0)$ für die durchschnittliche Zahl nicht diagnostizierter Infizierter:

$$M^{(1)}(n) = m_0 R^n \tag{18}$$

Unter der Annahme $m_0 = 1$ ergibt sich der aus dem Harris Schätzer für R resultierende Schätzer für $M^{(1)}$:

$$\tilde{M}^{(1)}(n) = \tilde{R}_N^n \tag{19}$$

wobei N die Zahl der Tage ist, deren Daten in die Schätzung eingehen. Dieser Schätzer beruht auf der Annahme, dass das Virus sich frei exponentiell gemäß einer konstanten Reproduktionsrate ausbreiten kann. Da dies weit von der Realität entfernt ist, wollen wir diesen Schätzer nicht weiter verwenden.

Sinnvoller ist ein Schätzer für $M^{(1)}$, der die Zeitinhomogenität des Prozesses berücksichtigt. Dafür nutzen wir die zuvor erstellen Schätzer $\tilde{R}_{n,L}$, die nur die letzten L Tage berücksichtigen und in (13) definiert wurden.

Für die zeitabhängigen Schätzer ergibt sich:

$$\tilde{M}_L^{(1)}(n) = \tilde{R}_{L,L}^{L-1} \prod_{i=L}^{n} \tilde{R}_{i,L} \text{ für } n > L, \quad L = 7, 14, 28 \tag{20}$$

Außerdem wollen wir mithilfe dieser Schätzer das Verhältnis $\alpha(n)$ von diagnostizierten Kranken zu der Gesamtzahl Infizierter ermitteln. Wir stellen also die Schätzer auf:

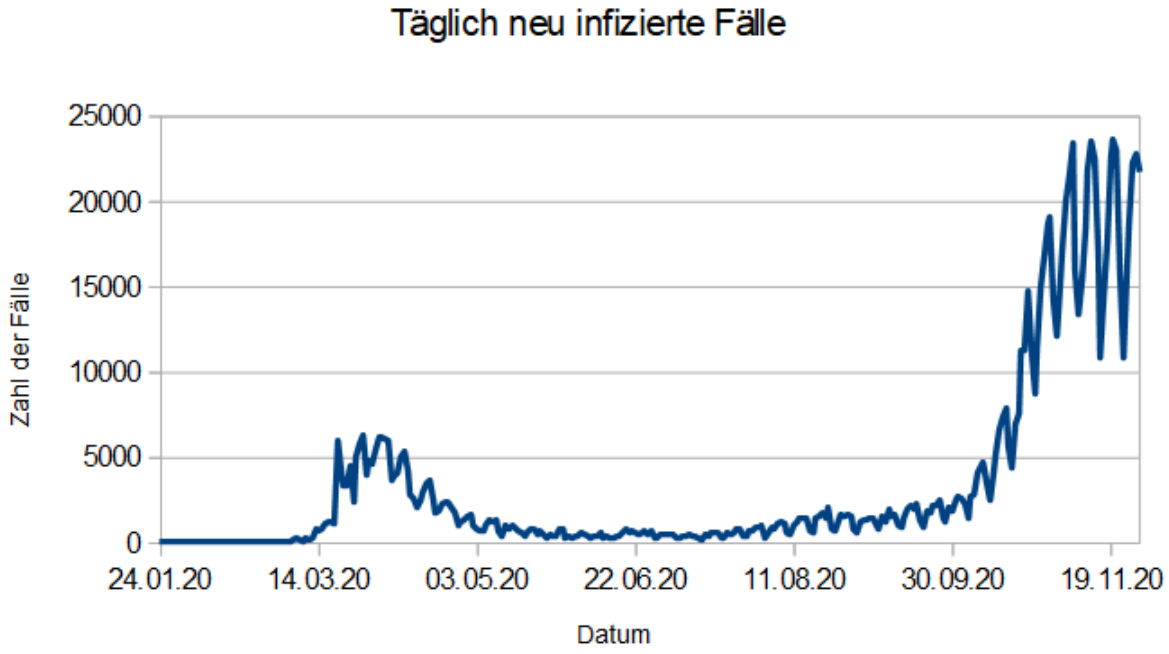
$$\tilde{\alpha}_L(n) = \frac{X_n^{(2)}}{X_n^{(2)} + \tilde{M}_L^{(1)}(n)} \text{ für } n > L, \quad L = 7, 14, 28 \tag{21}$$

3.4 Schätzungen und Evaluation der Schätzer am Beispiel der Daten aus Deutschland

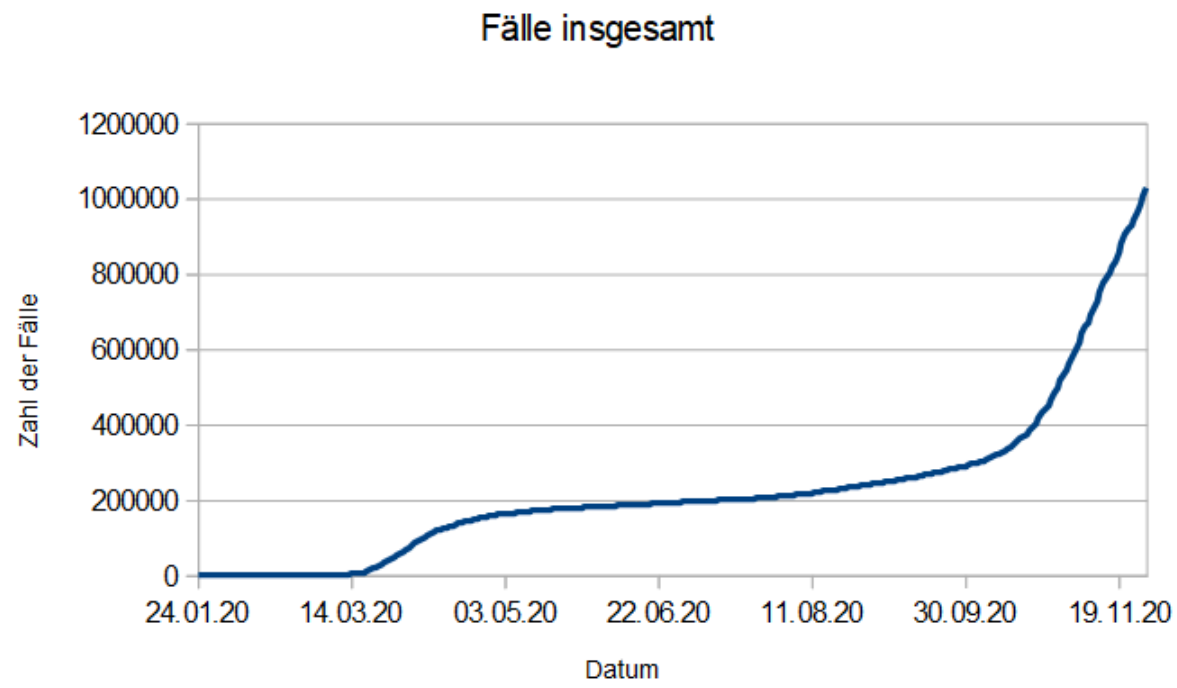
Mit den zuvor erklärten Formeln erstellen wir Graphen, die unsere Schätzer in einer Anwendung auf die Fallzahlen in Deutschland veranschaulichen sollen.

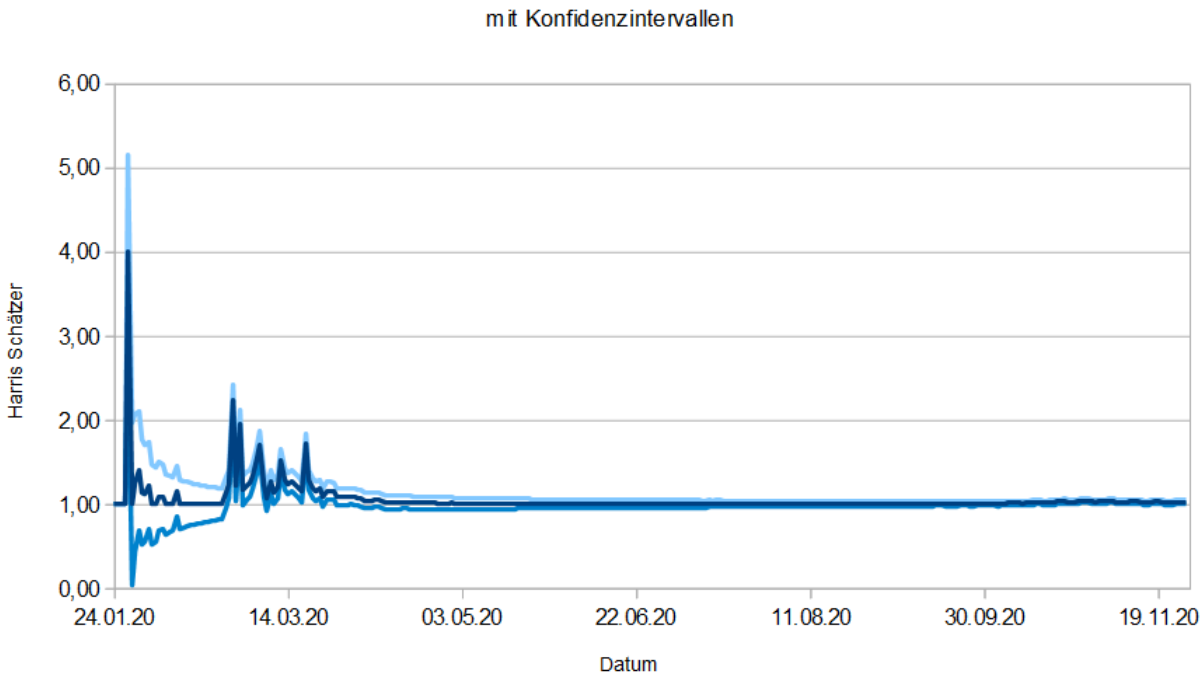
Der erste gemeldete Corona-Fall in Deutschland war am 28.01.2020. Es ist unmöglich, zu wissen, wie viele nicht diagnostizierte erkrankte Individuen zu der Zeit in Deutschland waren. Wir müssen daher eine Annahme machen und berufen uns auf das serielle Intervall des RKI. Wir gehen davon aus, dass genau vier Tage zuvor eine unbekannte Zahl nicht diagnostizierter kranker Individuen im Land war. Diese Zahl legen wir in einer weiteren Annahme auf *ein* Individuum fest. Somit ist Tag Null der 24.01.20 mit $m_0 = 1$.

Die Daten über tägliche Neuinfektionen werden vom European Centre of Disease Prevention and Control bereitgestellt. Auf ihrer Website [5] findet man die Daten für alle Länder bis zum 14.12.2020. Nach diesem Datum wurde zu einer wöchentlichen Berichterstattung gewechselt. Für unsere Betrachtung gehen wir auf die Daten zwischen 24.01.2020 und 28.11.2020 ein.



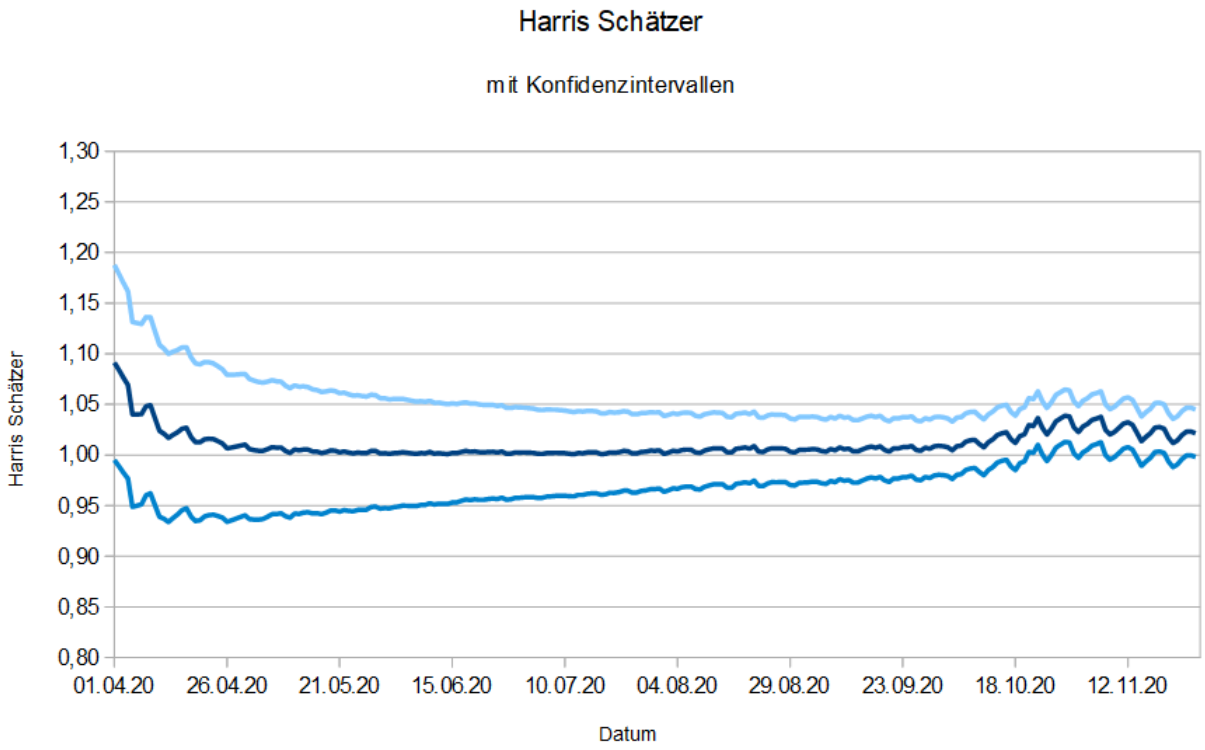
Hier kann man sehr eindeutig zwei Infektionswellen ausmachen: eine erste zwischen dem 14.03.2020 und dem 03.05.2020 und eine zweite, beginnend im September 2020. Auch kann man deutlich die starken Schwankungen innerhalb jeder Woche sehen. An den Wochenenden werden weniger Zahlen erfasst, da die Labore und Gesundheitsämter nicht voll besetzt sind. Die Fälle werden verspätet unter der Woche gemeldet, sodass selbst bei klaren Aufwärtstrends wie im September 2020 vom einen Tag auf den nächsten die Fallzahl sinken kann. Dies ließ Raum für Fehlinterpretationen und ist einer der Gründe, wieso auf 7-Tage-Summen gewechselt wurde.





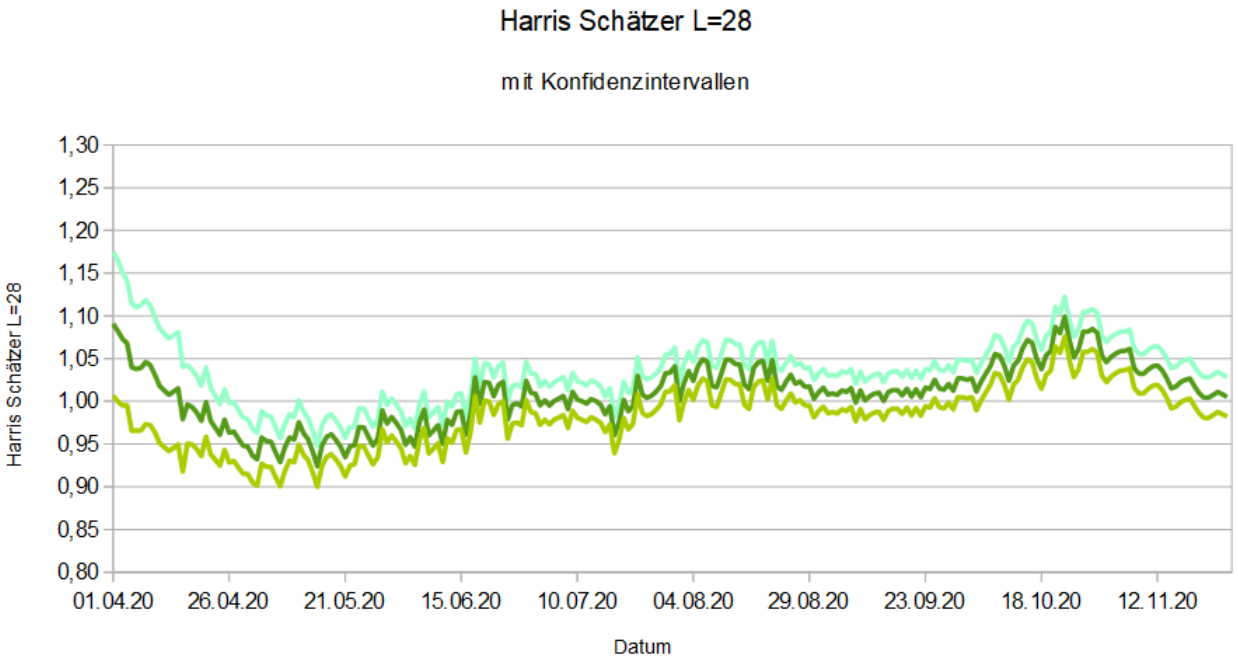
Hier wird dargestellt: der in Formel (12) erläuterte Schätzer für den R-Wert: $\tilde{R}_n$, mit dem in Formel (13) erläuterten 95%-Konfidenzintervall. Für die ersten Monate schwanken unsere Schätzer sehr stark.

Betrachten wir den Graphen zwischen April und November 2020:

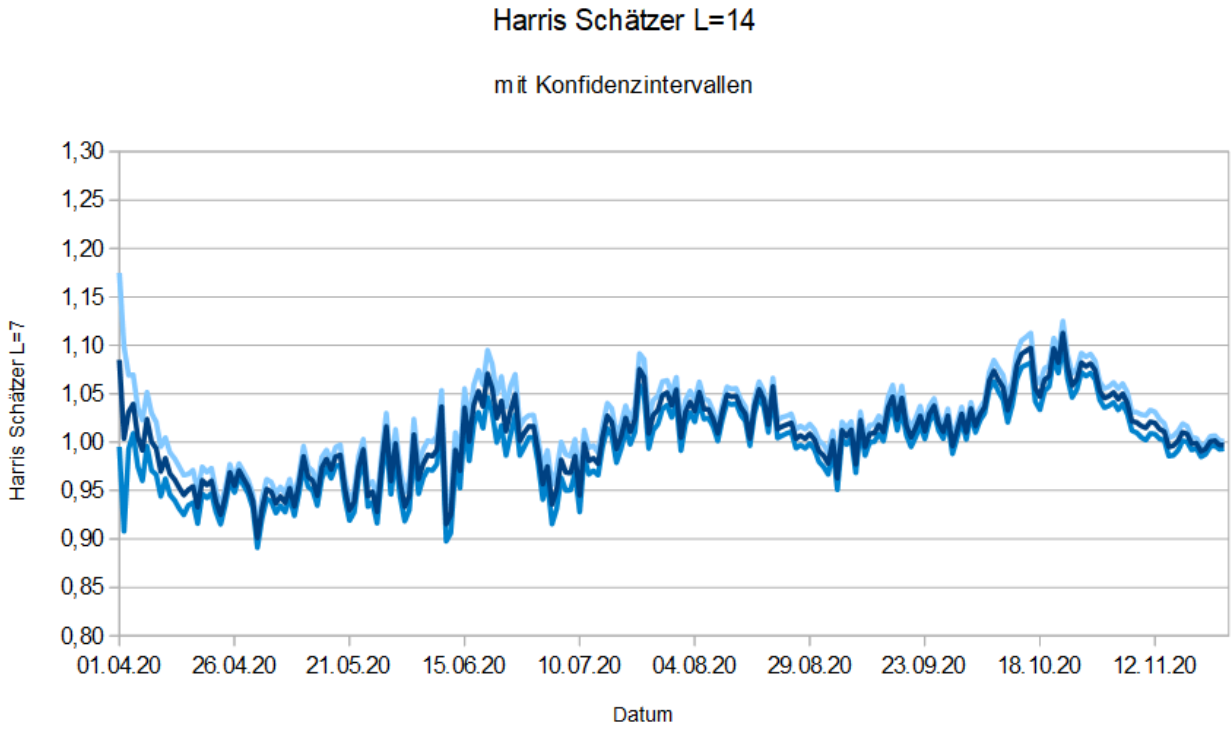


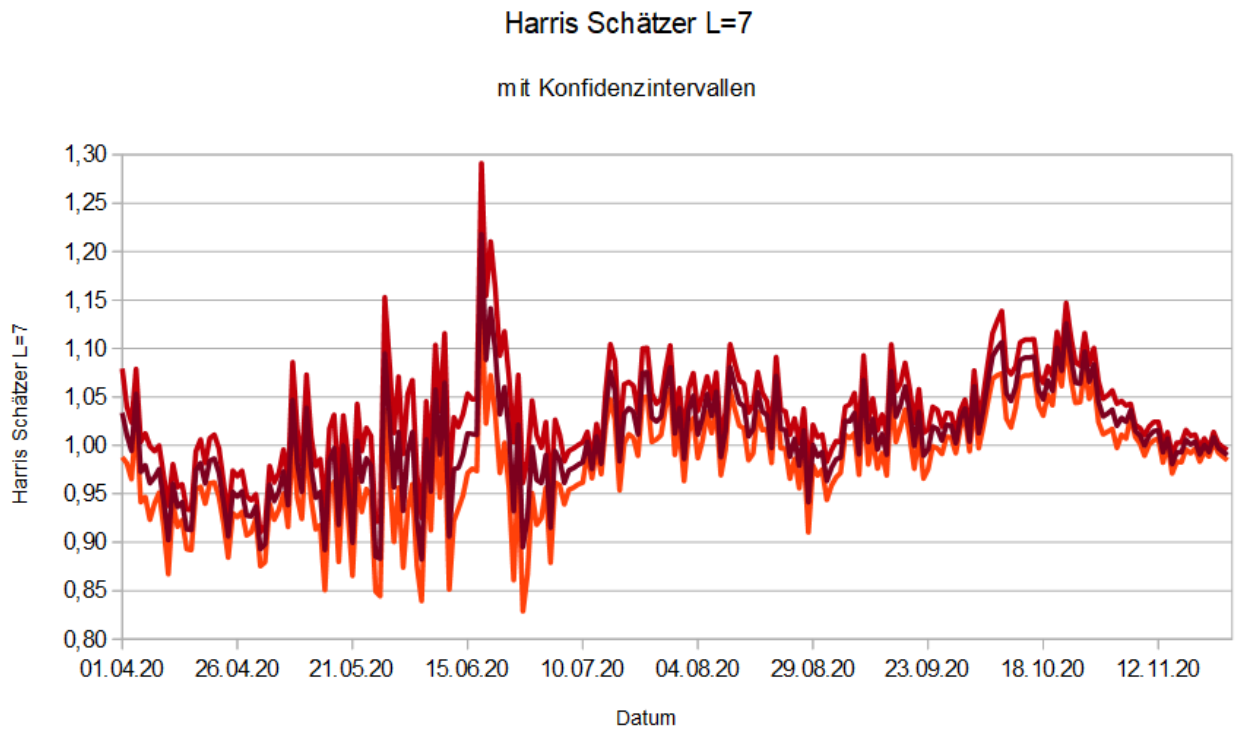
Hier kann man sehen, wieso die Annahme der Zeitunabhängigkeit dem Schätzer schadet: Der R-Wert wird zu konstant und der Schätzer geht zu wenig auf neue Daten ein: Obwohl die zweite Welle stärker ist, als die erste, liefert der Harris Schätzer für diese Zeit fast keine echte Aussage. Der Punktschätzer ist knapp über 1, das Konfidenzintervall schließt jedoch Werte über und unter 1 ein. Lediglich im Oktober und November gibt es kurze Perioden, an denen der R-Wert laut Konfidenzintervall über 1 ist.

Die Variationen des Harris Schätzers, die nur auf die letzten L Daten eingehen, werden mit $\tilde{R}_{n,L}$ bezeichnet und in Formel (14) erklärt. Die dazugehörigen 95%-Konfidenzintervalle wurden gemäß Formel (15) berechnet.



Für $L = 28$ werden nur die letzten 28 Tage betrachtet und der Punktschätzer schwankt stärker. Hier sieht man den Vorteil gegenüber dem Harris Schätzer, der alle Daten nutzt: Nach diesem Schätzer ist das 95%-Konfidenzintervall des R-Werts am Ende der ersten Welle eindeutig unter 1 und im Oktober/ November 2020 eindeutig über 1.



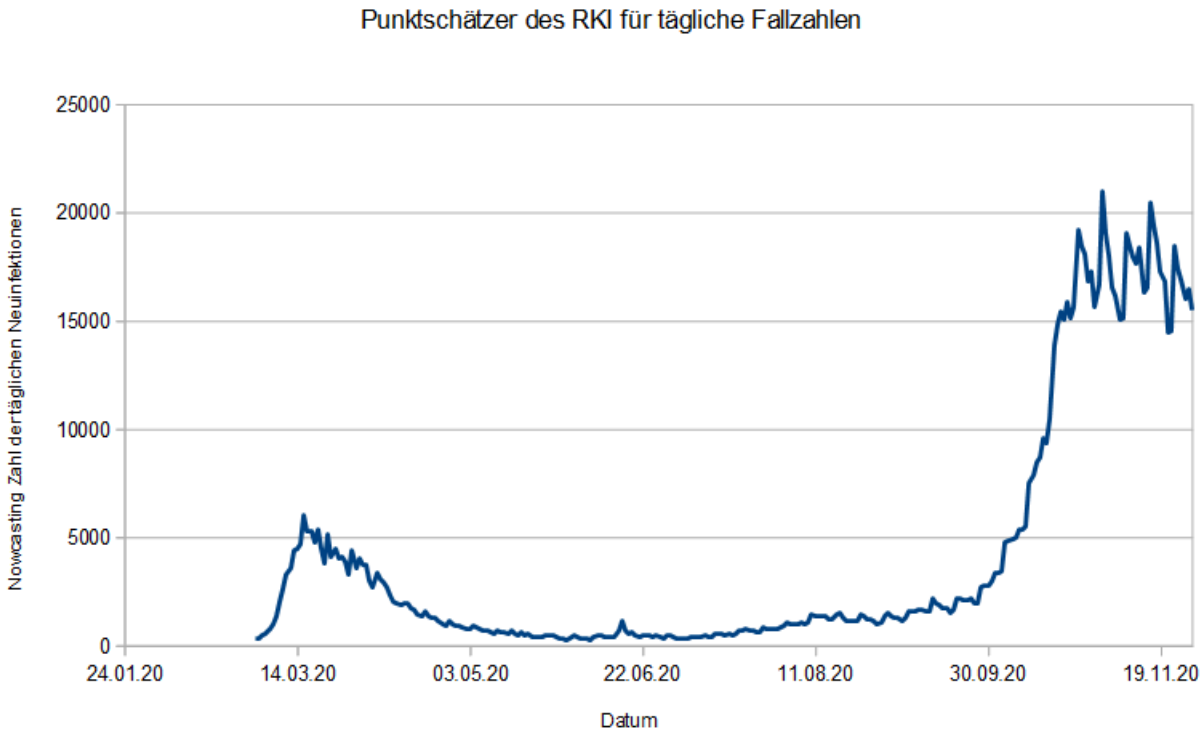


In den letzten beiden Graphen gibt es noch weitere eindeutige Bereiche, bei denen R unter bzw. über 1 liegt. Das am besten zu lesende Resultat liefert hier der Harris Schätzer mit $L = 14$.

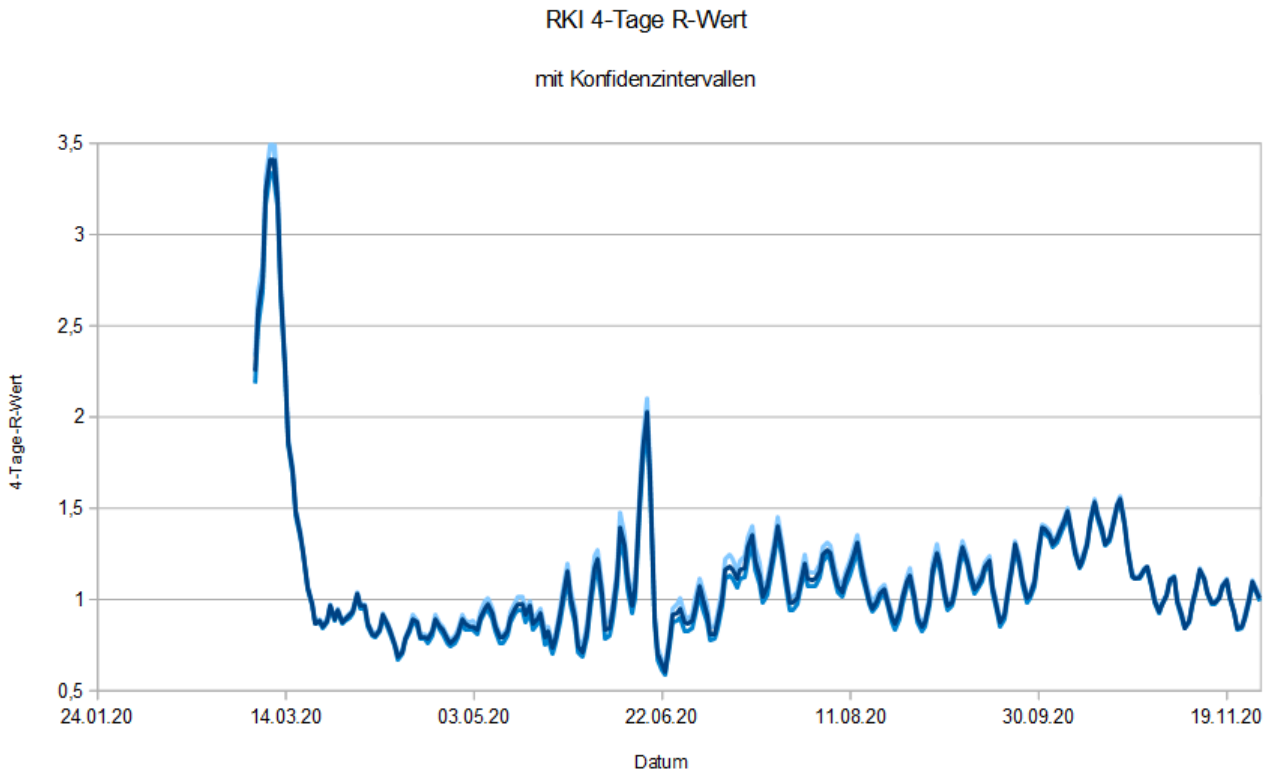
Mitte April bis Mai ist $\tilde{R}_{n,L} < 1$, was man erwarten würde, da in diesem Zeitraum die erste Welle endete. Im Juni gibt es eine kurze Periode, in der der Schätzer mit 95% über 1 liegt, obwohl die Fallzahlen zu dem Zeitpunkt kaum ansteigen. Eindeutig zu erkennen ist der Beginn der zweiten Welle im November. Auch auf das Abflachen der Welle am Ende des Novembers 2020 reagieren die beiden Schätzer wie man erwarten würde. Mit den Schätzern für $L = 14$ und $L = 7$ lassen sich also echte Aussagen treffen.

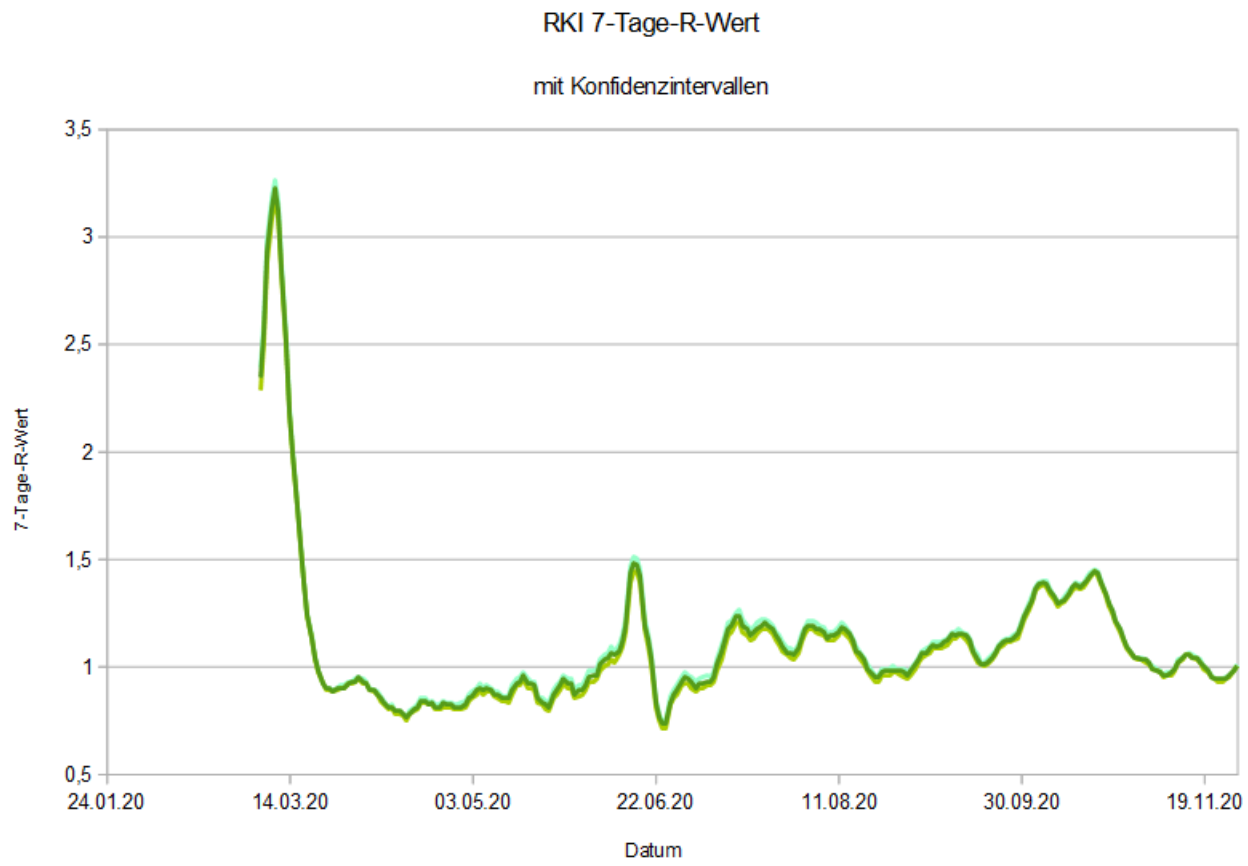
Nimmt man zu viele Daten, so geht der Harris Schätzer zu wenig auf neue Daten ein. Betrachtet man zu wenig Tage, so schwankt der Schätzer zu sehr. $\tilde{R}_{n,14}$, der Harris Schätzer, der auf die letzten 14 Tage eingeht, scheint hier einen guten Punkt zwischen den beiden Extremen zu treffen.

Betrachten wir die Daten des RKI, die ab dem 02.03.2020 zur Verfügung stehen:



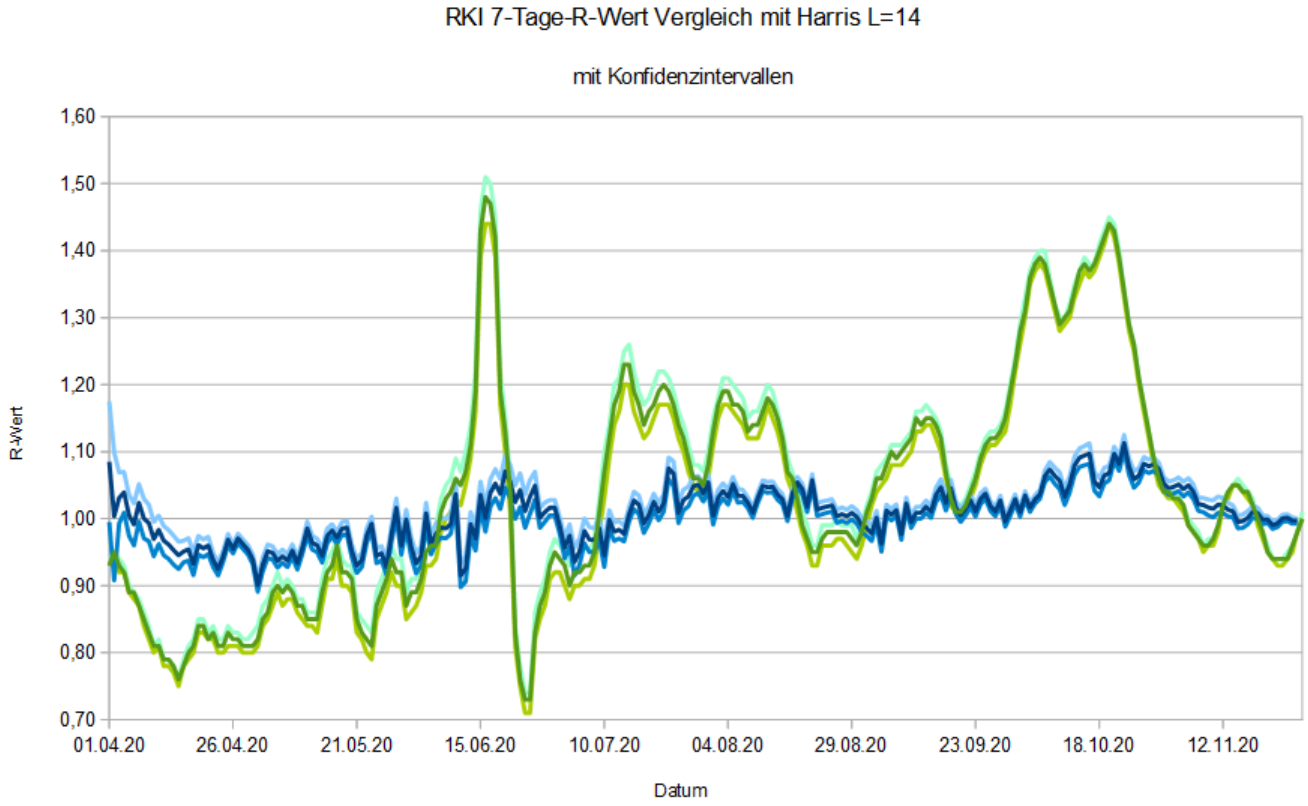
Der Verlauf der Fallzahlen, die das RKI für seine Schätzer nutzt, sieht dem Graphen der Daten vom European Centre of Disease Prevention and Control sehr ähnlich. Nur die Schwankungen innerhalb einer Woche sind weniger prägnant. Die Daten für diesen Graphen sind in [10] zu finden. Die *Tabelle mit Nowcasting-Zahlen zur R-Schätzung* dort wird täglich aktualisiert und liefert: die korrigierten Fallzahlen, mit denen das RKI rechnet und die Schätzer für den 4- beziehungsweise 7-Tage-R-Wert. Alle Werte werden mit ihren 95%-Konfidenzintervallen angegeben.





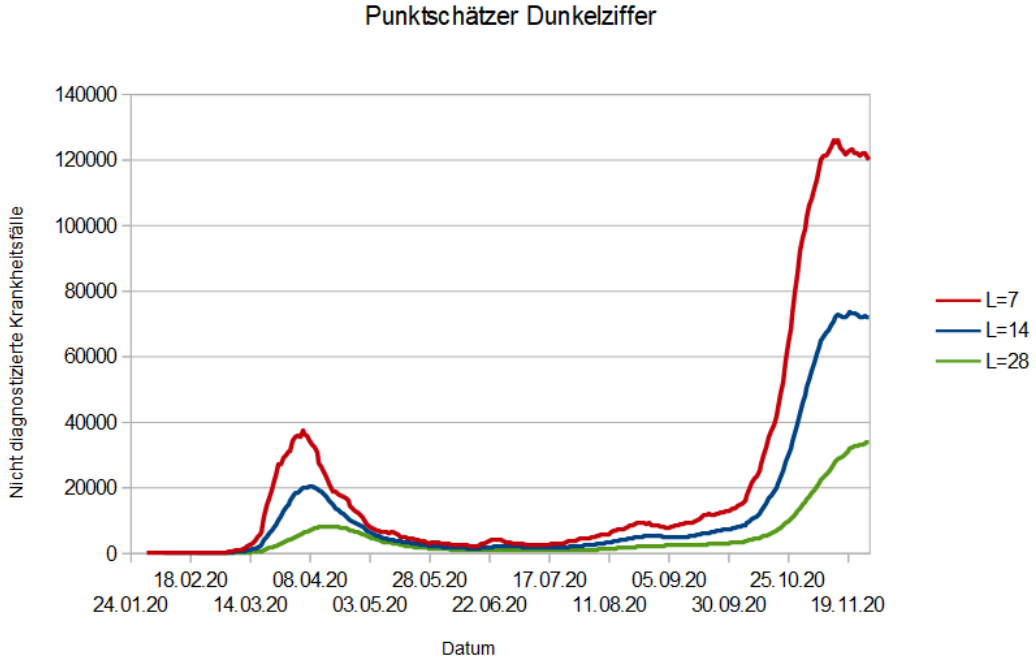
Die Formeln zu $R_{RKI,n,4}$ und $R_{RKI,n,7}$ wurden in (16) und (17) beschrieben. Im Vergleich sieht man, dass der 7-Tage-R-Wert weniger schwankt, aber trotzdem eindeutige Trends darstellt.

Im folgenden Graphen wird der RKI Schätzer für den 7-Tage-R-Wert (grün) dem Harris Schätzer $\tilde{R}_{n,14}$ (blau) mit jeweiligen Konfidenzintervallen gegenübergestellt.



Die Graphen haben auf den ersten Blick vielleicht wenig gemeinsam. Betrachtet man sie jedoch unter der Fragestellung, wann die Schätzer einen R-Wert über bzw. unter dem kritischen Wert Eins liefern, so ähneln sich $\tilde{R}_{n,14}$ und $R_{RKI,n,7}$ durchaus. Der Harris Schätzer liefert die gleichen Änderungen allerdings um etwa eine Woche verzögert. Das sieht man zum Beispiel in der zweiten Welle: Hier zeigt der RKI-Schätzer schon am 23.09.2020 einen Wert eindeutig über 1, während das Konfidenzintervall des Harris Schätzer erst zum Anfang Oktober komplett über 1 ist.

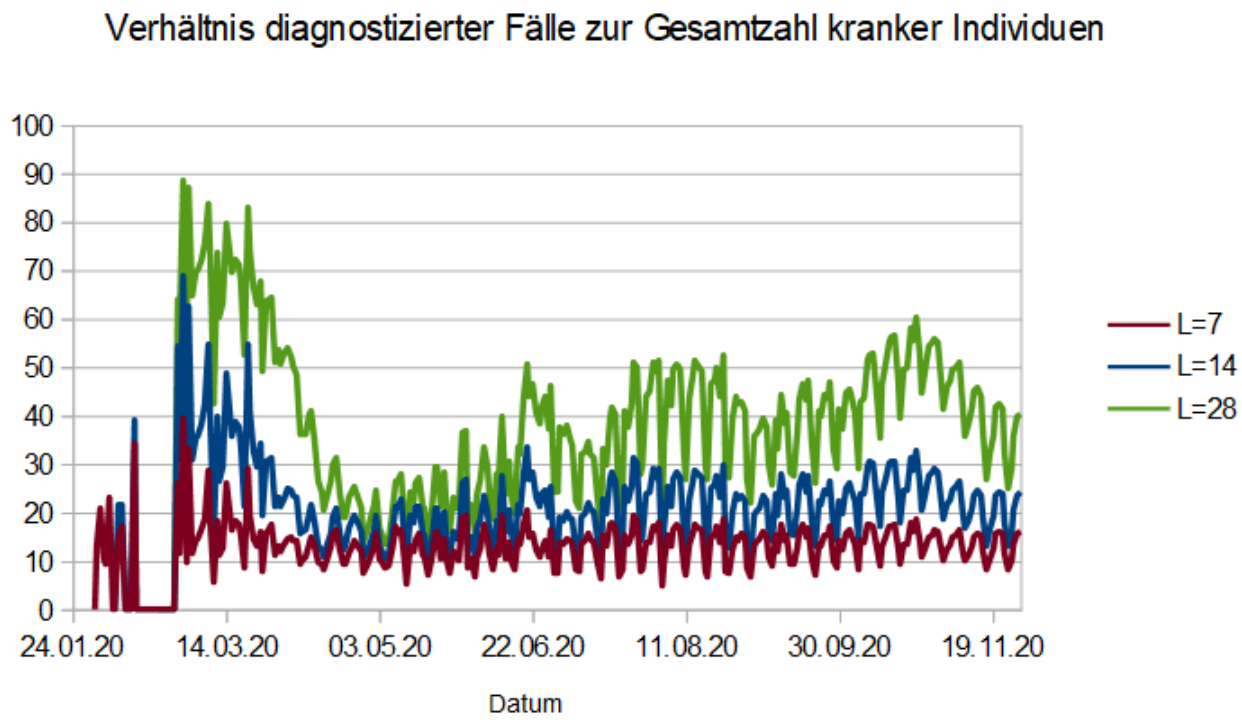
Die Schätzer für den Verlauf der Dunkelziffer $\tilde{M}_L^{(1)}(n)$ sind wie in (20) beschrieben auf Grundlage der $\tilde{R}_{n,L}$ berechnet worden. Sie sehen sich für unterschiedliche L sehr ähnlich, unterscheiden sich aber massiv in ihrer Skala:



Für $L = 7$ wird die Dunkelziffer auf bis zu 120.000 geschätzt - eine äußerst pessimistische Aussicht. Für $L = 14$ erreicht die Dunkelziffer nur die Hälfte dieses Wertes. Und für $L = 28$ ist der höchste Wert bei 33.000, also in ähnlicher Größenordnung wie die täglich gemeldeten Fälle.

Diese großen Unterschiede entstehen durch die stark schwankenden Daten zum Beginn der Pandemie. Außerdem ist unser Schätzer für die Dunkelziffer sehr anfällig für Fehler: Verschätzen wir uns beispielsweise bei einem Anfangswert um die Hälfte, so setzt sich dieser Fehler linear fort. Wir sind am Anfang von $m_0 = 1$ ausgegangen. Ist in Wirklichkeit $m_0 = 2$, so müssten alle folgenden Werte doppelt so groß sein.

Leider sind diese Schätzer wenig aussagekräftig. Um das 95% Konfidenzintervall für $\tilde{M}_L^{(1)}(n)$ zu berechnen, müssten wir in (20) die obere bzw. untere Grenze der Konfidenzintervalle der jeweiligen Harris Schätzer einsetzen. Dann ergeben sich für die Grenzen der Intervalle jedoch 0 und quasi unendlich.

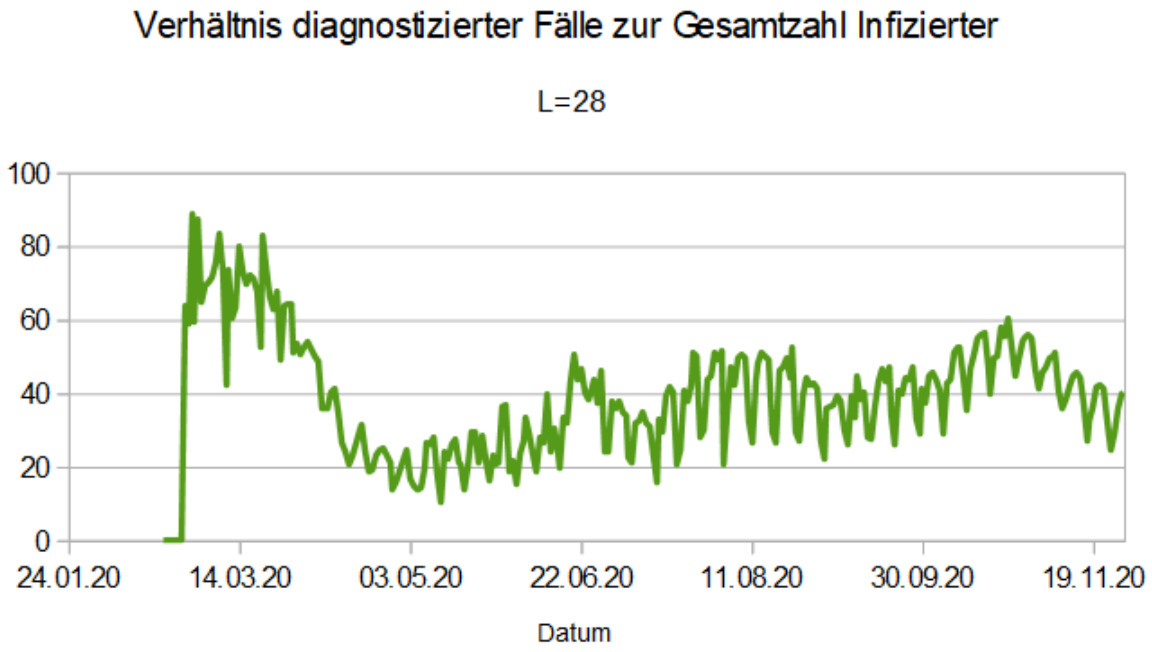


Diese Verhältnisse $\tilde{\alpha}_L(n)$ wurden wie in (21) berechnet.

- Für $L = 7$ liefert die Schätzung, dass nur etwa 12% der Infizierten auch diagnostiziert sind.
- Für $L = 14$ liefert die Schätzung, dass etwa 25% der Infizierten diagnostiziert sind.
- Für $L = 28$ liefert die Schätzung, dass etwa 40% der Infizierten diagnostiziert wurden.

Im Artikel [14] wurden Ergebnisse für Schätzungen dieses Verhältnisses im März bis Juni 2020 für verschiedene Länder weltweit präsentiert. Sie lagen zwischen 71% (in Chile) und 5% (in Frankreich). Das kann neben dem tatsächlichen Infektionsgeschehen auch an unterschiedlicher Testpolitik liegen. Unsere Ergebnisse befinden sich in diesem Bereich, schwanken aber für verschiedene L so stark, dass wir keine genaue Aussage für Deutschland treffen können.

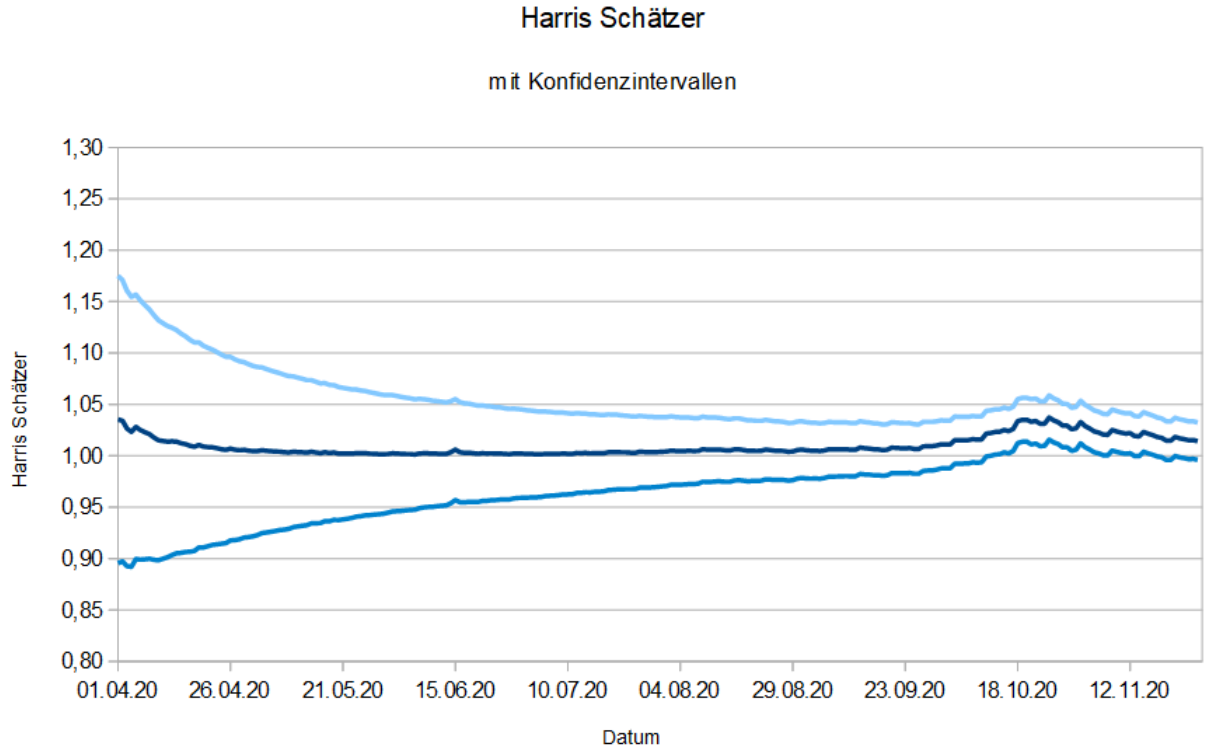
Am Verlauf dieser Graphen sollte uns etwas stören, was anhand des Graphen für $L = 28$ verdeutlicht werden soll:



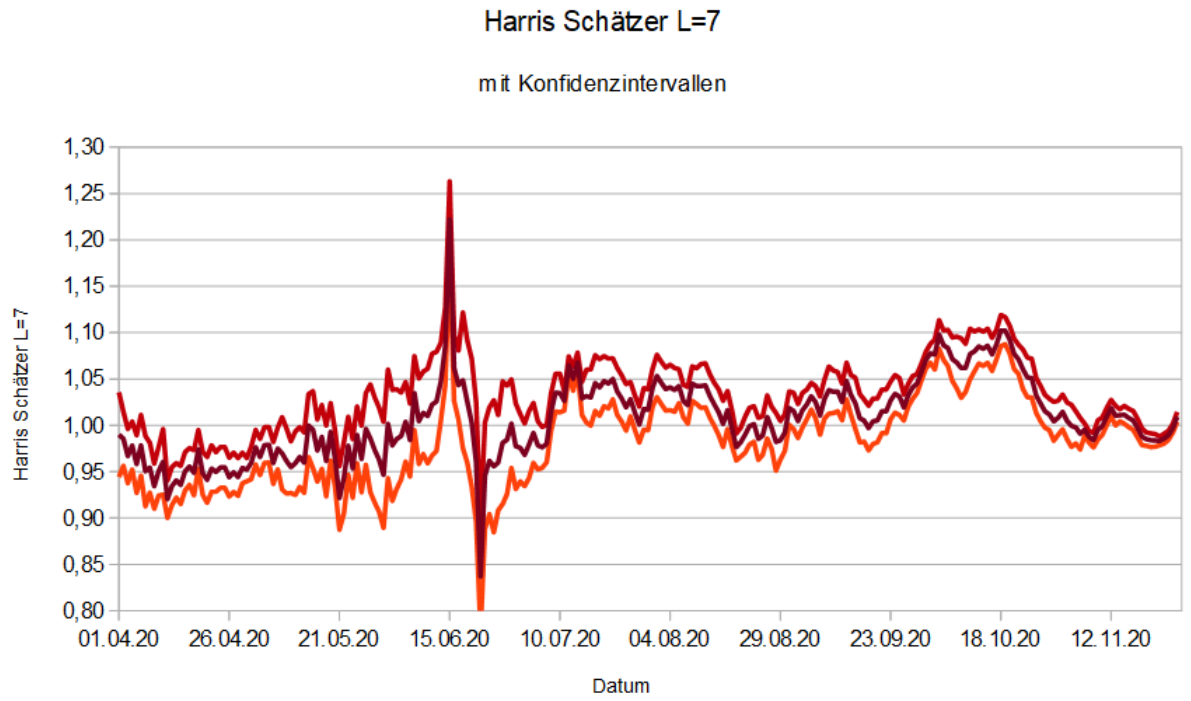
Hier sollte uns erstaunen, wie sich das geschätzte Verhältnis diagnostizierter Fälle zur Gesamtzahl der Fälle verhält: Es wäre logisch, wenn zu Zeiten weniger gemeldeter Fälle verhältnismäßig wenige nicht diagnostizierte Menschen in der Bevölkerung sind. Währenddessen sollte bei hohen Fallzahlen davon ausgegangen werden, dass im Verhältnis mehr kranke Menschen nicht diagnostiziert werden. Dafür gibt es viele Gründe: Zum Beispiel können Teststellen überlastet sein, sodass nur Leute mit Symptomen getestet werden, oder Menschen lassen sich nicht testen, weil sie aufgrund der hohen Fallzahlen Angst haben, sich bei den überfüllten Teststellen anzustecken. Der Graph verhält sich aber gegenteilig: Während der ersten Welle wird der Prozentsatz diagnostizierter Infizierter auf etwa 80 geschätzt. In der ruhigen Phase zwischen den Wellen sinkt dieser Wert auf bis zu 20%. Dann während der zweiten Welle steigt der Wert wieder an.

Ein Argument für diesen Verlauf des Graphen ist folgendes: zu Zeiten hoher Fallzahlen könnte die Motivation innerhalb der Bevölkerung, sich testen zu lassen, besonders hoch sein. Grund dafür könnte die breite Medienaufmerksamkeit sein und das Gefühl, solidarisch eine Krise abwenden zu müssen. Wenn sich besonders viele Menschen testen lassen, bleiben wahrscheinlich weniger nicht identifizierte Fälle übrig. Zu Zeiten weniger Testfälle hingegen haben die Menschen möglicherweise wieder ein Gefühl von Normalität und weniger das Bedürfnis, sich auf Corona testen zu lassen. In diesem Szenario wäre eine verhältnismäßig höhere Dunkelziffer die Folge.

Eine weitere Möglichkeit ist es, nicht die tatsächlich gemeldeten täglichen Fälle für unsere Schätzer zu verwenden, sondern vorverarbeitete Daten, wie es das RKI tut:

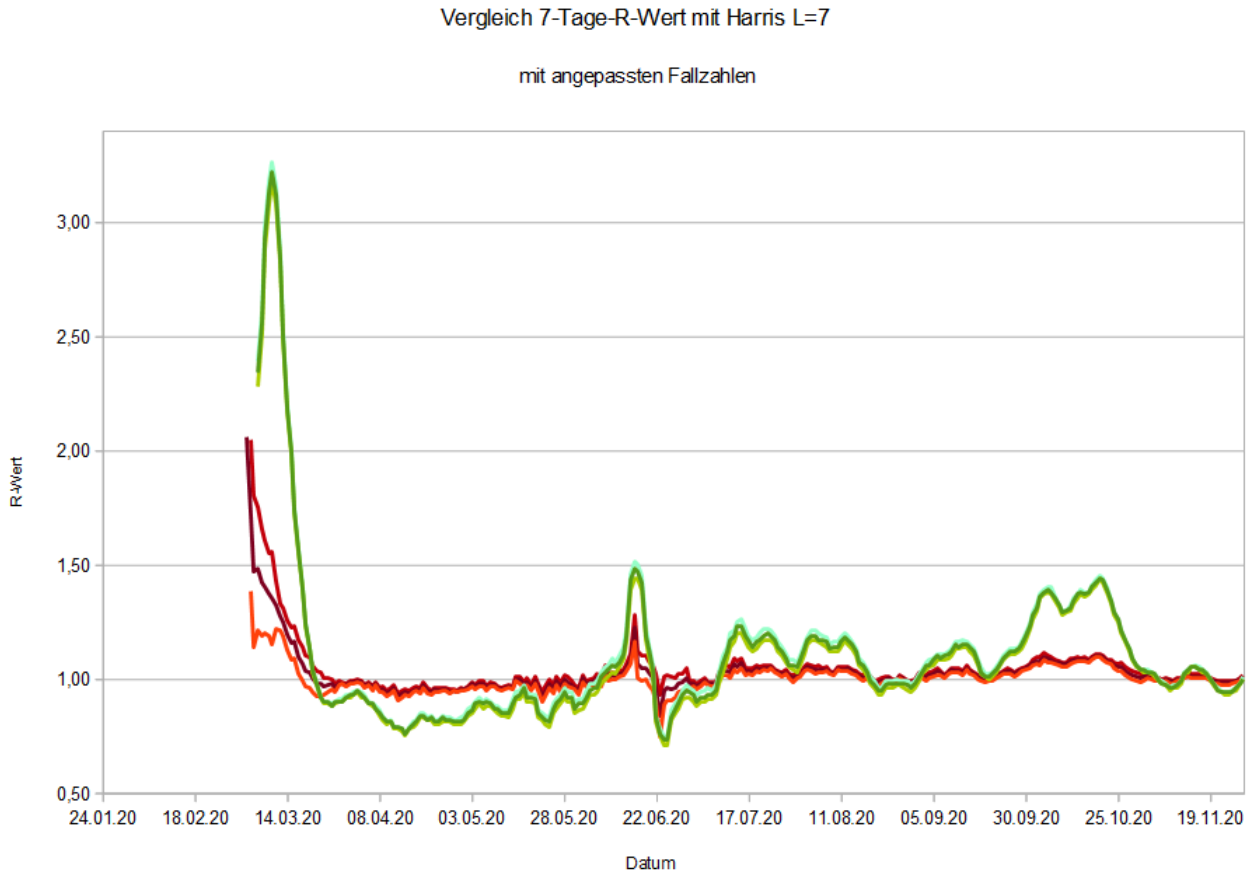


Wir können sehen, dass diese Herangehensweise den Verlauf des Graphen zwar glättet, er aber prinzipiell gleich bleibt in dem Sinne, in welchen Bereichen er über oder unter 1 ist.



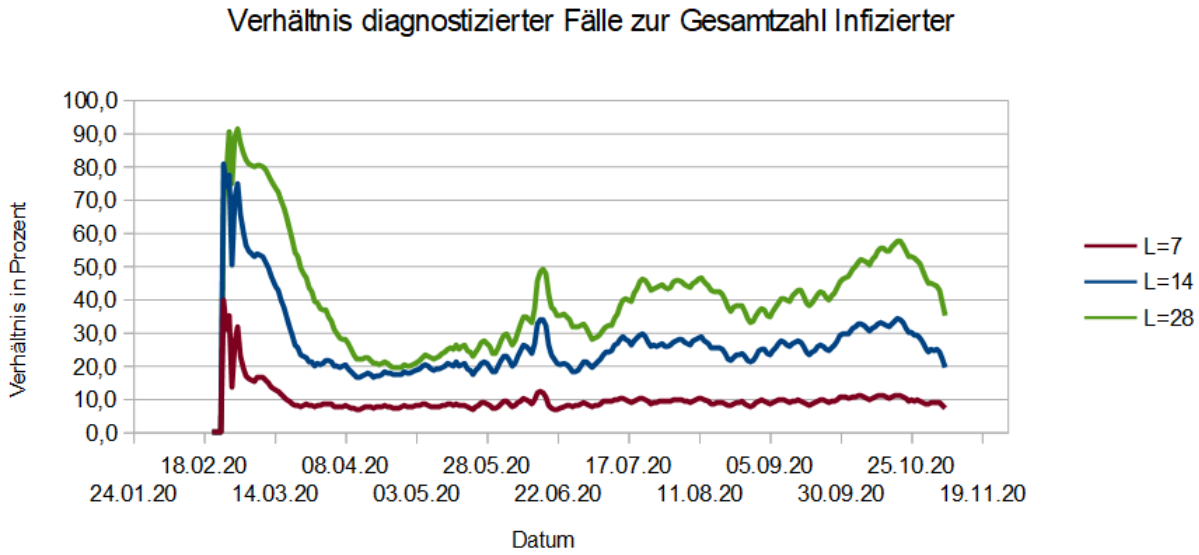
Dieser Graph ist um einiges lesbarer geworden.

Hier vergleichen wir den Graphen für $\tilde{R}_{n,7}$ (rot) mit dem 7-Tage-R-Wert des RKI ($R_{RKI,n,7}$, grün):



Im Gegensatz zum letzten Vergleich liegt hier zwischen den beiden Schätzern keine sichtbare Verzögerung vor. Der Schätzer des RKI schlägt jedoch um einiges stärker aus.

Der Graph, der den Prozentsatz diagnostizierter Personen im Verhältnis zur Gesamtzahl Infizierter anzeigt, wird durch die Glättung der Fallzahlen auch viel lesbarer:



4 Ausblick

In seinem Paper [17] geht Yanev auch auf die Möglichkeit ein, den Verzweigungsprozess mit einem Term zu ergänzen, der Immigration von Infizierten in den Prozess einbindet. Dafür werden die identisch unabhängig verteilten Zufallsvariablen $\{I_n\}$ eingeführt. I_n gibt die Zahl von nicht diagnostizierten infizierten Individuen an, die am Tag n ins Land immigriert sind. Der große Unterschied zum vorherigen Modell ist, dass der Prozess hier nicht ausstirbt, selbst wenn im Land zu einem Zeitpunkt keine Krankheitsfälle vorhanden sind. Die Zahl der Infizierten im nächsten Schritt ergibt sich dann aus:

$$X_{n+1}^{(1)} = \sum_{k=1}^{X_n^{(1)}} Y_{n,k,1}^{(1)} + I_n$$

$$X_{n+1}^{(2)} = \sum_{k=1}^{X_n^{(1)}} Y_{n,k,1}^{(2)}$$

Dieses Modell liefert tatsächlich die gleiche Formel für den Harris Schätzer des R-Werts, wie das Modell ohne Immigration. Nur die Dunkelziffer wird hier anders berechnet.

Da diese Krankheit ein globales Phänomen ist, gibt es auch global unzählige wissenschaftliche Artikel darüber, die unmöglich alle hier aufzulisten sind. Wer weiterführende Lektüre sucht, kann beispielsweise
https://www.biomedcentral.com/collections/GlobalOutbreaksandResponses
aufsuchen, wo unter anderem in [14] das Verhältnis von diagnostizierten Kranken zu nicht diagnostizierten Infizierten geschätzt wurde.

5 Schlusswort

In dieser Arbeit haben wir Verzweigungsprozesse mit mehreren Typen eingeführt und einige wichtige Ergebnisse aus diesem Themenbereich zusammengetragen. Vor allem haben wir eine Möglichkeit eingeführt, trotz komplizierterer Übergangsmatrizen herauszufinden, wie sich der dazugehörige Prozess auf lange Zeit verhält: Abhängig vom größten Eigenwert λ. Die Bedeutung dieses Parameters konnten wir dann in unserer Anwendung auf das Corona-Virus auf die Reproduktionsrate R übertragen. Das Modell, das wir in dieser Arbeit vorgestellt haben liefert simple Schätzer, die wir mithilfe von Graphen auswerteten: Der Harris Schätzer wäre besonders sinnvoll bei einer Epidemie, deren Bedingungen sich mit der Zeit nicht ändern. Das trifft auf die Covid-19 Pandemie nicht zu, da das Virus sich offenbar nicht nur wie die Influenza zu kalten Jahreszeiten schneller ausbreitet, sondern seine Ausbreitung durch Lockdowns zeitweise massiv eingedämmt wurde. Die Harris Schätzer, die sich auf die letzten L Tage beschränkten, lieferten aussagekräftige Ergebnisse. Auch die daraus resultierende geschätzte Zahl der nicht diagnostizierten infizierten Individuen war interessant, schwankte aber in L so sehr, dass wir uns in der Hinsicht auf kein Ergebnis festlegen konnten.

In der massiven Zahl an wissenschaftlichen Artikeln, die zur Pandemie herausgebracht wurden, gibt es viele detailliertere Modelle, die mehr auf all das Wissen eingehen, das mittlerweile über das Virus existiert. Zum Beispiel fließen in unser Modell wenige Ergebnisse von Studien ein, die die Inkubationszeit und Verzögerung zwischen Test und Meldung untersuchen. Trotzdem ist die von Yanev entwickelte Herangehensweise interessant: Sie veranschaulicht ein aktuelles Anwendungsbeispiel für multitype Verzweigungsprozesse und erstaunt uns darin, dass sie trotz eines komplizierten Modells sehr simple Schätzer liefert.

Literatur

[1] Alsmeyer, G.: Galton-Watson Prozesse. Skript, Institut für Mathematische Stochastik, Arbeitsgruppe Gerold Alsmeyer (Erneuerungstheorie, Verzweigungsprozesse), Universität Münster, Kapitel 1
https://www.uni-muenster.de/Stochastik/Arbeitsgruppen/Alsmeyer/

[2] an der Heiden M, Hamouda O: Schätzung der aktuellen Entwicklung der SARS-CoV-2-Epidemie in Deutschland ? Nowcasting 2020
https://edoc.rki.de/handle/176904/6650.4

[3] Athreya, K.B., Ney, P.E.: Branching Processes. Springer Verlag Berlin Heidelberg New York 1972, Kapitel 5

[4] Dalitz, C.: Konstruktionsmethoden für Konfidenzintervalle. Technischer Bericht Nr. 2017-01, Fachbereich Elektrotechnik und Informatik, Hochschule Niederrhein 2017
https://www.hs-niederrhein.de/fileadmin/dateien/FB03/Technische_Berichte/fb03-tb-2017-01-de.pdf

[5] European Centre of Disease Prevention and Control Website:
https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

[6] Harris, T.E.: Branching Processes. The Annals of Mathematical Statistics , Dec., 1948, Vol. 19, No. 4, Institute of Mathematical Statistics, S. 474-494
http://www.jstor.com/stable/2236017

[7] Joffe, A., Spitzer, F.: On multitype branching processes with $\rho \leq 1$. Journal of mathematical analysis and applications 19, 1967, S. 409-430

[8] Königsberger, K.: Analysis 1. Springer Verlage Berlin u.a. 2004, Seite 78

[9] Merkel zur Bedeutung von Infektionszahlen:
https://www.youtube.com/watch?v=Ki5vx3jEvJI#t=35m20

[10] RKI offizielle Website:
https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus /Projekte_RKI/Nowcasting.html

[11] RKI Aktuelle Situationsberichte: https://www.rki.de/DE/Content/InfAZ/N/ Neuartiges_ Coronavirus/ Situationsberichte/Gesamt.html

[12] RKI Steckbrief Coronavirus: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Steckbrief.html
Abschnitt 5: Inkubationszeit und serielles Intervall

[13] Roelly, S.: Grundlegende Eigenschaften von Bienaymé-Galton-Watson-Verzweigungsprozessen in diskreter Zeit. Skript, Institut für Mathematik, Universität Potsdam 2010

[14] Russell, T.W., Golding, N., Hellewell, J. et al. Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. BMC Med 18, 332 (2020).
https://doi.org/10.1186/s12916-020-01790-9

[15] Vatutin, V.A.: Branching Processes and their Applications. Skript, Department of Discrete Mathematics, Steklov Mathematical Institute, Moskau 2005

[16] Yanev, N., Stoimenova V. und Atanasov D.: Branching stochastic processes as models of Covid-19 epidemic development. Version 1. Quantitative Biology, Cornell University 29.04.2020
https://arxiv.org/abs/2004.14838v1

[17] Yanev, N., Stoimenova V. und Atanasov D.: Branching stochastic processes as models of Covid-19 epidemic development. Version 2. Quantitative Biology, Cornell University 04.05.2020
https://arxiv.org/abs/2004.14838

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die von mir angegebenen Hilfsmittel genutzt habe.

Potsdam, den 21.03.2021