Mathematische Modelle zur Beschreibung der COVID-19-Pandemie in Deutschland

von **Ingo Dahn (dahn@dahn-research.eu)**

Was ist ein Jupyter Notebook?

Dies ist ein Jupyter-Notebook, das vor allem für die selbständige, interaktive Arbeit gedacht ist, das aber auch gerne in der Lehre eingesetzt werden kann. Mit diesem Notebook kann der Leser die Grundlagen von drei wichtigen epidemiologischen Modellen verstehen, aber - und das macht Jupyter-Notebooks zu etwas Besonderem - er/sie kann dadurch aktiv mit den Daten zur Ausbreitung von COVID-19 arbeiten und so die Leistungsfähigkeit und die Grenzen der behandelten mathematischen Modelle erfahren.

Anders als in einem Lehrbuch können Sie die durchgeführten Berechnungen ausführen, sie modifizieren und sie auf andere Daten (vielleicht für ein anderes Land?) anwenden. Sie würden manche Berechnungen anders durchführen oder würden gerne zusätzliche Analysen durchführen? Sehr gut! Tun Sie es einfach, ändern Sie die Berechnungen, fügen Sie neue Zellen hinzu - Sie können nichts kaputt machen!

Zusätzlich enthält dieses Notebook einige interaktive Aufgaben, in denen Sie mit Hilfe der Modelle eigene Voraussagen treffen können, deren Gültigkeit dann automatisch bewertet wird (nur für Sie - die Bewertung wird nirgends gespeichert).

Eine gute Online-Umgebung, um mit Jupyter-Notebooks wie diesem zu arbeiten, ist das CoCalc-System (https://cocalc.com). Wenn Sie etwas Zeit haben und auf die interaktiven Aufgaben verzichten können, so können Sie auch das Binder-System (https://mybinder.org/v2/gh/ingodahn/Corona/master?filepath=Deutschland.ipynb) verwenden. Dieses Notebook verwendet übrigens den Kernel SageMath 9.0.

Schließlich können Sie mit diesem Notebook auch im CoCalc-Player (https://dahn-research.eu/corona/Deutschland.html) interaktiv arbeiten.

Praktische Hinweise

In der ausführbaren Ansicht dieses Notebooks in CoCalc bzw. im CoCalc-Player können die Eingabezellen editiert und neu berechnet werden um die Verfahren zu prüfen oder sie auf andere Daten anzuwenden.Viele Zellen haben Schieberegler, mit denen Parameter der Berechnung angepasst werden können. Jede Neuberechnung einer Zelle ändert möglicherweise Daten oder Definitionen, die andere Zellen verwenden

Deshalb müssen die Zellen immer in der gegebenen Reihenfolge ausgeführt werden!

Um eine Zelle auszuführen wählen Sie die Zelle aus und verwenden Sie die Tastenkombination Shift+Return.

Falls Sie Verbesserungsvorschläge, Modifikationen, Erweiterungen oder Anwendungen auf andere Daten haben oder einfach nur dieses Notebook nutzen, so würde ich mich über eine Information dazu freuen.

 Dieses Notebook ist auf GitHub (https://github.com/ingodahn/Corona) verfügbar.

Koblenz im Juli 2020

Dr. Ingo Dahn

Inhalt

Wie gut beschreiben mathematische Modelle reale Prozesse?

In diesem Jupyter-Notebook untersuchen wir die Passung der einfachsten Pandemie-Modelle - des exponentiellen Modells, des SI-Modells und des SIR-Modells - zu den Daten der Pandemie in Deutschland für die Zeit vom 24.2. bis zum 30.6.2020, wie sie vom Robert-Koch-Institut zur Verfügung gestellt wurden.

Zunächst wird ein Überblick über die verwendeten Daten zur Entwicklung der COVID-19-Pandemie gegeben. Es werden taggenaue Daten für COVID-19-Infektionen, -Todesfälle und -Genesungen verwendet.

Der Hauptteil zu den mathematischen Modellen diskutiert zunächst die Frage, wie die Qualität eines (nichtlinearen) mathematischen Modells beurteilt werden kann. Wir definieren dafür die Modell-Toleranz und die Vorhersage-Toleranz, die auf der Berechnung relativer Residuen beruhen.

Danach werden für jedes der drei behandelten Modelle die Überlegungen vorgestellt, die hinter dem jeweiligen Modellansatz stehen und die zur Bestimmung der jeweils dem Modell zugrundeliegenden Differentialgleichungen führen.

Danach versuchen wir, geeignete Werte der Parameter für die einzelnen Modelle zu bestimmen, so dass die Modelle den Verlauf der Pandemie - zumindest in einem gewissen Zeitabschnitt - möglichst gut beschreiben und sinnvolle Vorhersagen ermöglichen.

Für das Verständnis des exponentiellen Modells sind elementare Kenntnisse über Exponentialfunktionen ausreichend, wenn die automatische Bestimmung der Parameter hingenommen wird.

Das logistische Modell beruht auf der Kenntnis der logistischen Funktion, die in diesem Kapitel eingeführt wird. Auch für das logistische Modell kann die Bestimmung geeigneter Parameter automatisch erfolgen. Dabei werden wir sehen, wie das Modell laufend durch Änderungen an den Parametern angepasst werden muss bis es schließlich Anfang Mai 2020 seine Grenzen erreicht. Für das Verständnis des logistischen Modells sind elementare Kenntnisse über Differentialgleichungen hilfreich, jedoch nicht zwingend erforderlich.

Die Anwendung des SIR-Modell auf die Modellierung der COVID-19-Pandemie erfordert schließlich ein gewisses Verständnis für ein System von Differentialgleichungen und dessen näherungsweise numerische Lösung mit dem Runge-Kutta-Verfahren.

Ein abschließendes Fazit fasst die Ergebnisse zusammen und ein Ausblick zeigt Möglichkeiten, dieses Notebook (selbst) weiterzuentwickeln.

Daten

Im Folgenden wird statt mit Kalender-Daten, mit Tagen seit dem 24.2.2020, dem Beginn der täglichen Datenemeldungen des Robert-Koch-Instituts gerechnet. Die folgende Tabelle ermöglicht eine Umrechnung.

Tag Nr.	Datum
0	24.2.2020
20	15.3.2020
40	4.4.2020
60	24.4.2020
80	14.5.2020
100	3.6.2020
120	23.6.2020
140	13.7.2020

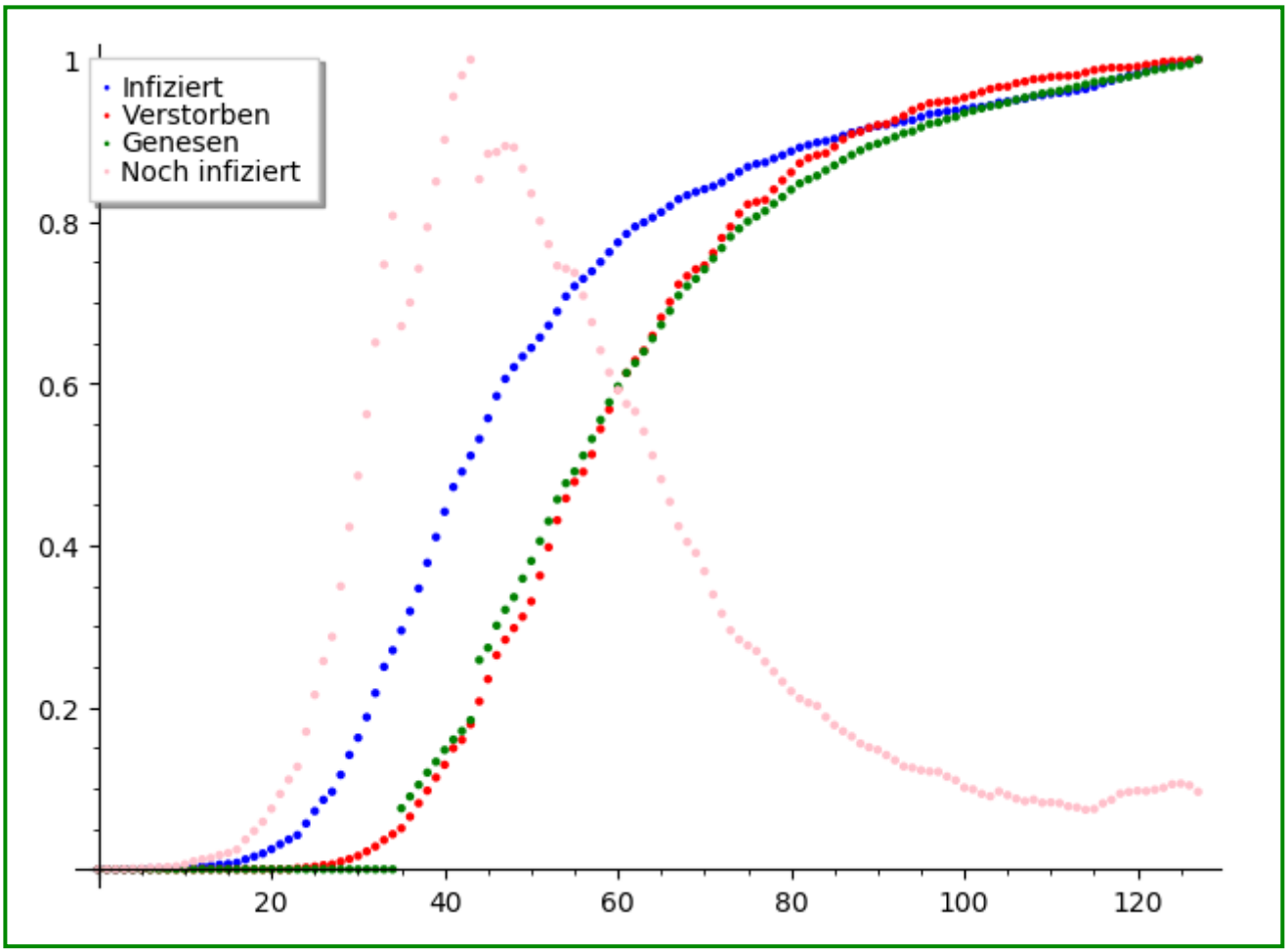
Die Daten sind aktuell bis zum 30.06.2020.

Es stehen taggenaue Meldungen des RKI zu

- gemeldeten Infektionen
- geschätzte Genesungen (auf volle 100 gerundet)
- gemeldete Todesfälle

zur Verfügung. Wir gehen davon aus, dass die absolute Zahl der der Todesfälle realistisch ist, während die Zahlen der realen Infektionen bzw. Genesungen nicht einmal von der Größenordnung her abgeschätzt werden kann, solange keine repräsentativen Querschnittsanalysen vorliegen. Die dazu gemeldeten bzw. geschätzten Zahlen erlauben lediglich das Erkennen von Tendenzen. deshalb stellen wir im Folgenden Diagramm, für eine erste Orientierung, die zur Verfügung stehenden kumulierten Daten auf ihren maximalen Wert normiert dar.

Auch in den folgenden Abschnitten arbeiten wir direkt mit den gemeldeten Daten. Der Leser ist jedoch eingeladen, die Infektionszahlen mit einem Dunkelfaktor seiner Wahl zu multiplizieren (5-15 gelten als realistische Faktoren) und die Modelle neu berechnen zu lassen.



Help (https://sagecell.sagemath.org/help.html) | Powered by SageMath (http://www.sagemath.org)

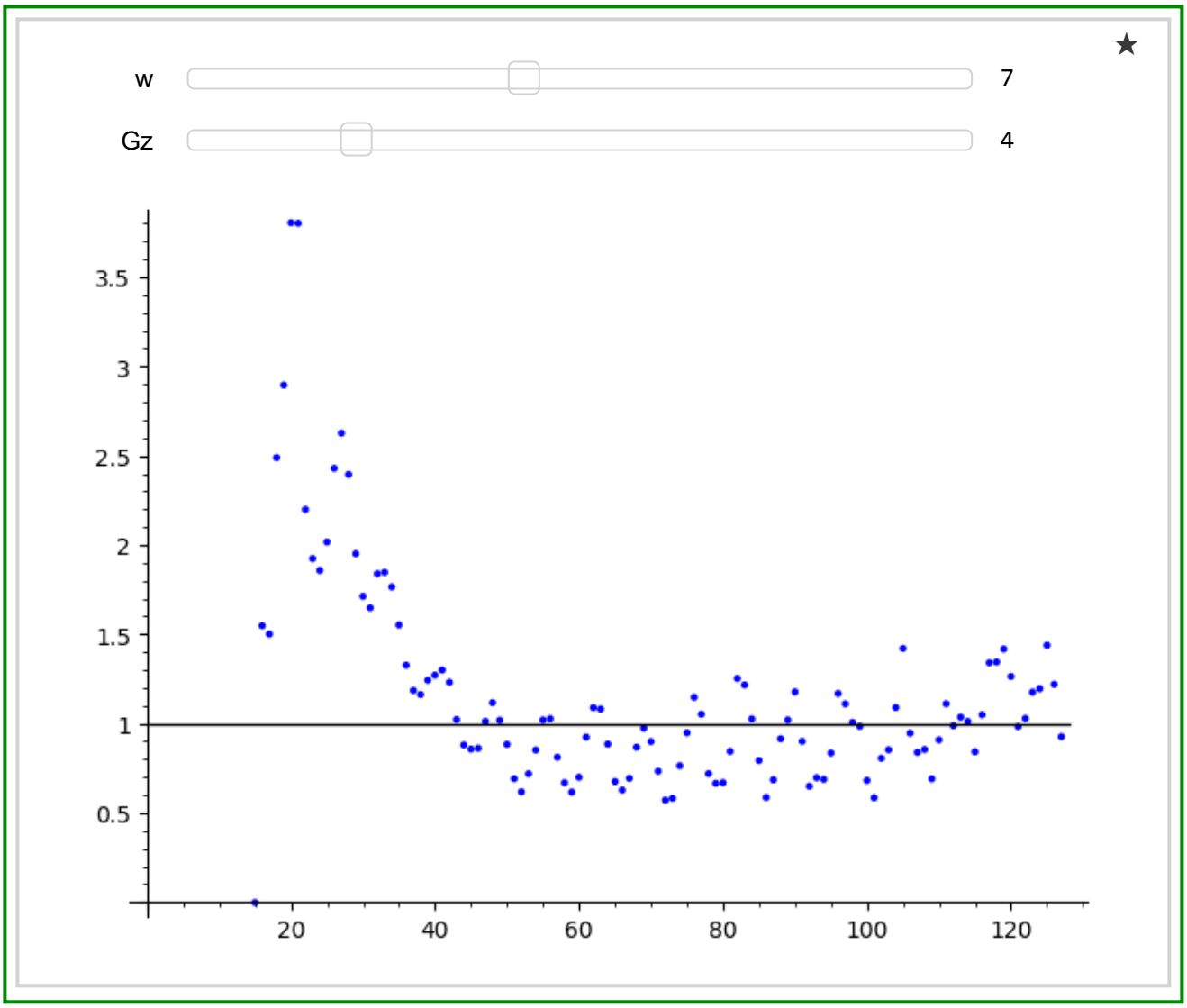
Reproduktionsrate

Hauptziel der Massnahmen zur Eindämmung der Pandemie ist zunächst, eine Überlastung der Intensivpflege in den deutschen Krankenhäusern (ohne zusätzliche Ressourcen) zu verhindern. Dazu müsste die Reproduktionsrate des Virus, d.h. die Zahl der von einem Infizierten angesteckten Personen nach Einschätzung der Deutschen Gesellschaft für Epidemiologie auf 1.1-1.2 gesenkt werden.

Im epidemiologischen Bulletin des RKI vom 17.4.2020 ist der folgende Weg zur Schätzung der Reproduktionsrate beschrieben. Sie beruht auf einer Schätzung der sogenannten Generationenzahl, d.i. die Zahl zwischen dem Beginn der Infektion eines Patienten und dem Beginn der Infektion eines von ihm angesteckten Patienten. Die Generationenzahl wird aufgrund des typischen Verlaufs der Infektion mit 4 Tagen angesetzt.

Dann kann die Reproduktionszahl geschätzt werden als Quotient der Anzahl der Neuerkrankungen in 2 aufeinanderfolgenden Perioden von der Länge der Generationenzahl, also $R(n) = \frac{i(n)-i(n-4)}{i(n-4)-i(n-8)}$, wobei $i(n)$ die Zahl der Infizierten am Tag n ist..

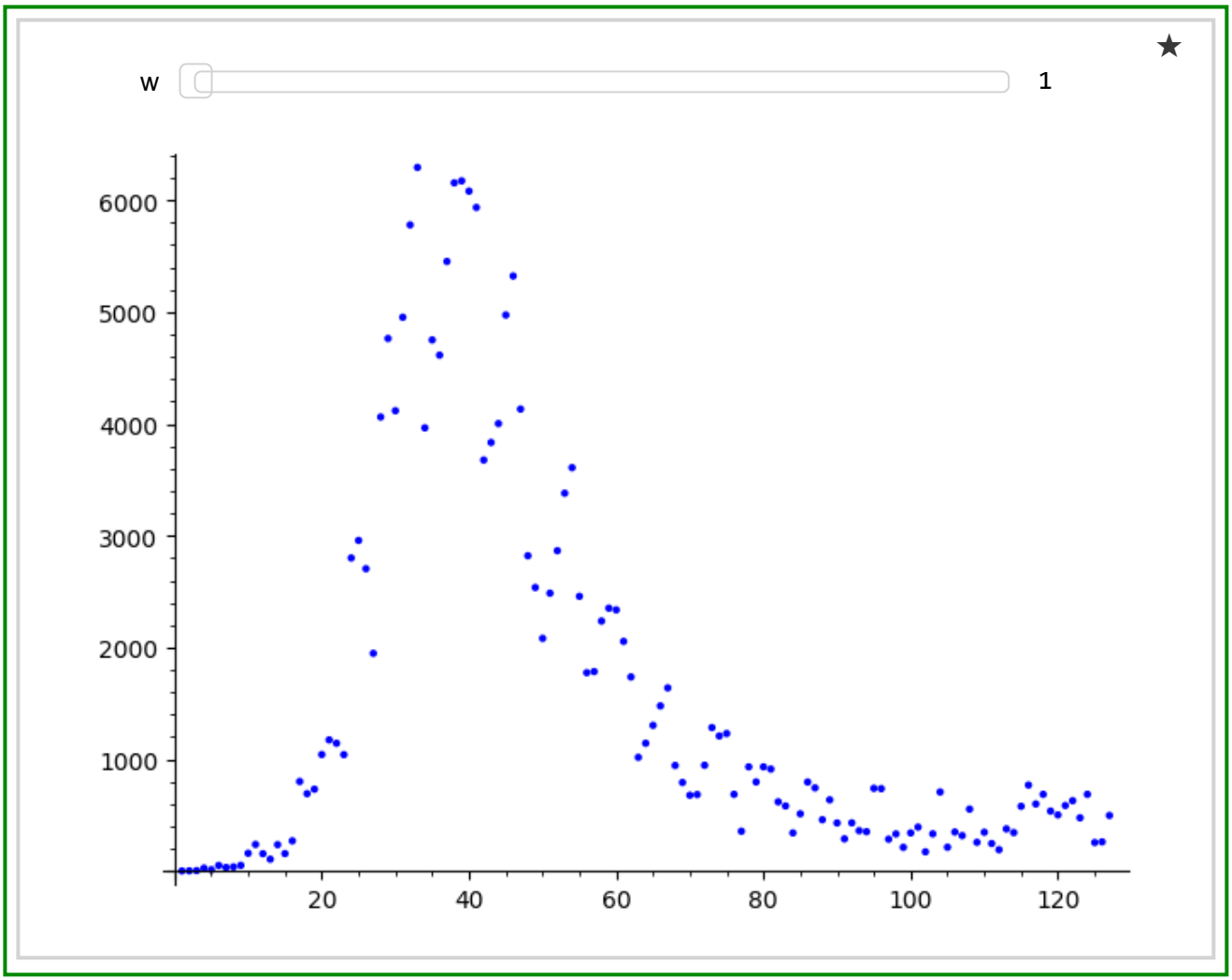
Im folgenden Diagramm kann die für die Berechnung der Reproduktionszahl verwendete Generationenzahl variiert werden. Es wird mit dem Durchschnitt der Infektionszahlen der letzten w Tage gerechnet.



Help (https://sagecell.sagemath.org/help.html) | Powered by SageMath (http://www.sagemath.org)

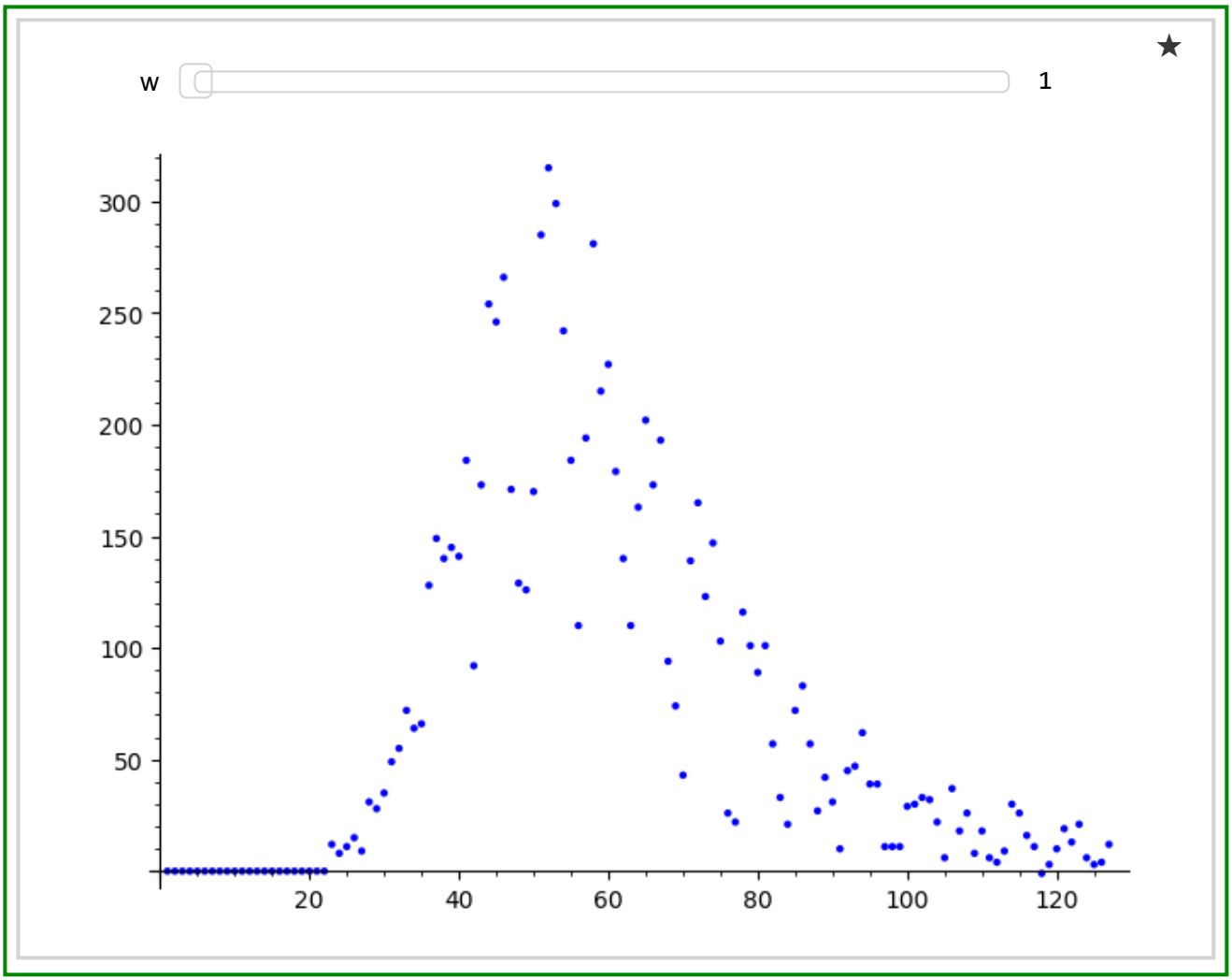
Neuinfektionen

Das folgende Diagramm zeigt für jeden Tag die Zahl der gemeldeten Neuinfektionen pro Tag, gemittelt über die letzten w Tage.



Todesfälle

Das folgende Diagramm zeigt für jeden Tag die Zahl der gemeldeten COVID-19-Todesfälle pro Tag, gemittelt über die letzten w Tage.



Mathematische Modelle

Wir suchen mathematische Funktionen, die den Verlauf der Pandemie möglichst gut beschreiben. Wie wir sehen werden, kann man in der ersten Phase (etwa bis Ende März 2020) dafür Exponentialfunktionen verwenden. In dieser Zeit breitet sich das Virus weitgehend ungehemmt aus. Im April beginnen hemmende Faktoren zu wirken, vermutlich vor allem Kontaktbeschränkungen, die die Ausbreitungsgeschwindigkeit verlangsamen. In dieser Phase liefern logistische Funktionen brauchbare Ergebnisse für die Beschreibung der Entwicklung der Zahl der Infizierten bzw. Verstorbenen.

Wie wir sehen schwanken die Daten der Neuinfektionen und Todesfälle von Tag zu Tag erheblich. Wir werden im Folgenden stets mit kumulierten Daten Arbeiten. Gegenüber diesen Daten sind die täglichen Schwankungen vergleichsweise klein, weshalb wir auf eine Mittelung zum Ausgleich der täglichen Schwankungen verzichten.

Bei der Interpretation der Ergebnisse ist zu berücksichtigen, dass die Daten mit einer erheblichen Ungenauigkeit behaftet sind, da bei Weitem nicht alle Erkrankten Symptome zeigen und deshalb in dem untersuchten Zeitraum gar nicht erfasst wurden. Solange das Testregime dabei bleibt, vorwiegend Personen mit Symptomen zu testen, wie das im März und April der Fall war, können wir annehmen, dass sich die Zahl der Infizierten von der Zahl der gemeldeten um einen unbekannten Dunkelfaktor unterscheidet. Insofern sind in der folgenden Analyse weniger die absoluten Zahlen als deren relative Unterschiede und deren Entwicklung von Interesse.

Wir wollen untersuchen, wie genau sich die reale Entwicklung der Zahl der mit COVID-19 Infizierten in Deutschland mit den verfügbaren mathematischen Modellen vorhersagen ließ.

Aufgabe: Führen Sie eine ähnliche Analyse für die Zahl der Todesfälle mit COVID-19-Bezug durch. Sie können dazu die Code-Zellen in diesem Notebook so modifizieren, dass Sie statt `infections_de` die Liste `deads_de` verwenden. Vergleichen Sie die beiden Analysen!

Versuchen Sie zunächst eine Vorhersage "von Hand":

Es ist umstritten, wie die Passung einer nichtlinearen Regressionsfunktion f zu einer gegebenen Datenreihe `dataP` am Besten definiert wird. Die Datenreihe der kumulierten Infektionszahlen weist Daten mit erheblichen Größenunterschieden auf. Bei größeren absoluten Werten der Daten erscheint auch eine größere absolute Abweichung der Funktionswerte von den Daten als akzeptable. Deshalb messen wir die Qualität der Regression hier mit dem relativen Residuum `rR2(dataP,f)` , das wie folgt definiert wird.

Ist $\vec{vd}$ der Vektor der beobachteten Daten und $\vec{vf}$ der Vektor der entsprechenden Funktionswerte der Funktion f, so definieren wir $rR2(\vec{vd}, f) = \frac{\|\vec{vd}-\vec{vf}\|}{\|\vec{vd}\|}$. Je geringer dieser Wert desto genauer die Approximation.

Das relative Residuum ist zu einem gegebenen Zeitpunkt x für zwei Bereiche von Interesse.

- Für den Bereich der Trainingsdaten von 0 bis x zur Beschreibung der Passgenauigkeit des Modells und
- Für den Vorhersagebereich von $x+1$ bis $x+zh$ zur Einschätzung der Vorhersagequalität.

  Deshalb definieren wir noch $rR3(\vec{vd}, f, x_0, x_1) = rR2(\vec{vd}|[x_0, x_1], f|[x_0, x_1])$, d.h. wir schränken die Argumente von $rR2$ auf den Bereich $[x_0, x_1]$ ein, insbesondere für $x_0 = 0, x_1 = x$ und $x_0 = x+1, x_1 = x+zh$.

Damit wird die Modelltoleranz $MT(\vec{vd}, f, x) = rR3(\vec{vd}, f, 0, x)$ und die Vorhersagetoleranz $VT(\vec{vd}, f, x, zh) = rR3(\vec{vd}, f, x+1, x+zh)$. Je kleiner diese Werte sind, desto besser beschreibt das Modell die Trainigsdaten bzw. desto besser ist die Vorhersage.

Es sei zugegeben, dass in der Wahl dieser Definitionen eine gewisse Willkür steckt. **Der Leser wird deshalb ausdrücklich ermutigt, die Definitionen in der folgenden Zelle durch eigene zu ersetzen!**

Residuen sind definiert

Exponentielles Modell

In der ersten Phase der Epidemie, etwa bis zum 31.3.2020 (Tag 36), beschreibt ein einfaches Modell des exponentiellen Wachstums die Entwicklung der Zahl der vom Robert-Koch-Institut gemeldeten Infektionsfälle recht gut. Die realen Fallzahlen schwanken in geringem Maße um den Verlauf einer approximierenden Exponentialfunktion. Wie wir sehen werden, wird die Abweichung Ende März größer als die vorherige zufällige Schwankungsbreite - das Modell eines exponentiellen Verlaufs stößt an seine Grenzen.

Das exponentielle Modell geht von der einzigen Annahme aus, dass die Zahl der Neuinfektionen in einem kleinen Zeitraum Δx proportional ist zur Zahl der zu Beginn Infizierten $I(x)$ und zur Länge des Zeitraums:

$$I(x + \Delta x) = c \cdot \Delta x \cdot I(x).$$

c ist dabei proportional zur Wahrscheinlichkeit, dass sich eine Person infiziert. Für $\Delta x \to 0$ ergibt sich daraus die Differentialgleichung $I' = c \cdot I(x)$ mit der Lösung $I(x) = ae^{cx}$, wobei $a = I(0)$ der Anfangswert ist. c bestimmt, wie steil die Exponentialfunktion ansteigt.

Das exponentielle Modell beschreibt ein ungehemmtes und unbegrenztes Wachstum und vernachlässigt Genesungen - wie wir sehen werden in der Anfangsphase der Pandemie durchaus zu Recht.

Für die Berechnung der passenden Parameter a, c stellt uns SageMath die funktion `find_fit` zur Verfügung. Diese Funktion versucht mittels Regressionsverfahren eine Funktion zu finden, die möglichst gut zu den gegebenen Werten passt.

Zahl der Infektionen

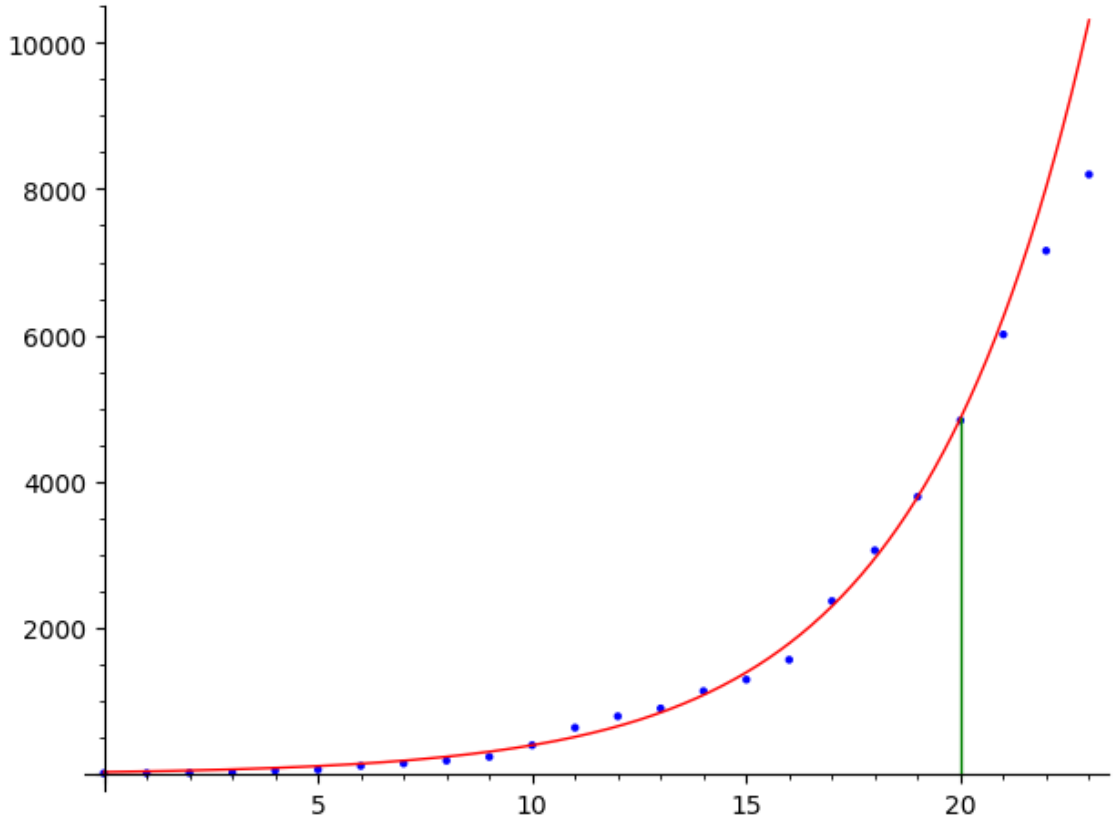
Das folgende Diagramm zeigt die kumulierte Zahl der Infektionen in Deutschland nach Angaben des Robert-Koch-Instituts vom 24.2.2020 (Tag 0) bis zu dem mit dem Schieberegler bestimmten Tag (grüne Linie). Die rote Kurve stellt eine Exponentialfunktion dar, die den Verlauf der Infektionszahlen bis zu diesem Tag möglichst gut approximiert und für den gewählten Zeithorizont `zh` darüber hinaus. Für diesen Tag werden Modell-Tolearanz und Vorhersage-Toleranz angezeigt.

Tag 20

$$32.81299410243268\, e^{(0.24998790149213163\, x)}$$

$\texttt{Modell-Toleranz:}0.0474288341988228$

$\texttt{Vorhersage-Toleranz:}0.184519373369907$

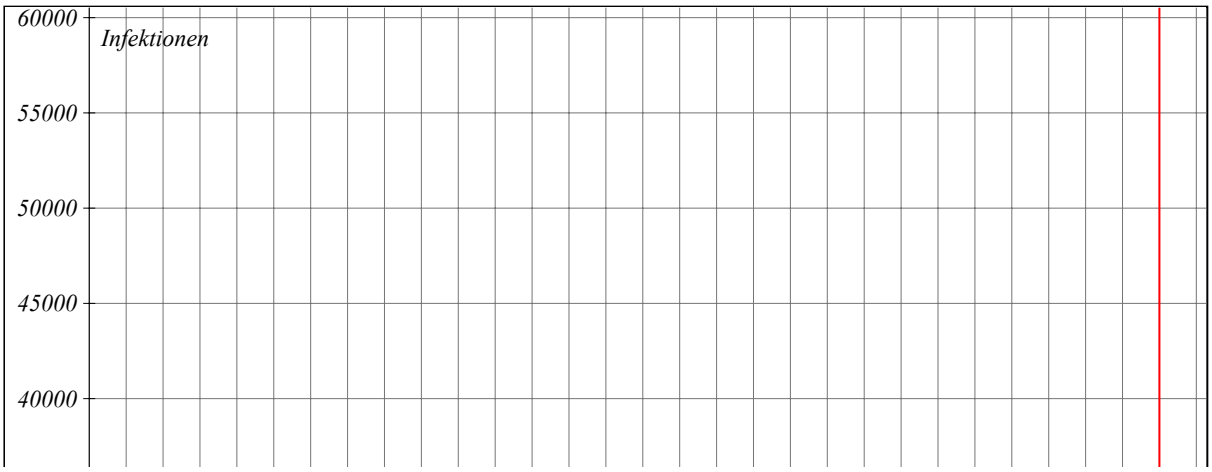


Bei einer rein exponentiellen Entwicklung ae^{cx} sollen sich die Logarithmen der Werte entsprechend einer linearen Funktion $cx + \ln(a)$ entwickeln. eine solche Funktion können wir mit Hilfe der linearen Regression näherungsweise bestimmen.

Sehen wir uns deshalb die Logarithmen dieser Werte an. Dazu zeichnen wir eine passende Regressionsgerade und einen Korridor um diese Gerade, der die zufälligen Schwankungen begrenzt.

Das folgende Diagramm zeigt die Anzahl der im Labor bestätigten Infektionen mit dem Coronavirus in Deutschland nach Angaben de Robert-Koch-Instituts vom 24.2.2020 (Tag 0) bis zum 23.3.2020.

Zeichnen Sie eine Kurve, die die Zahl der Infektionen bis zum 23.3.2020 (Tag 28) möglichst gut annähert:



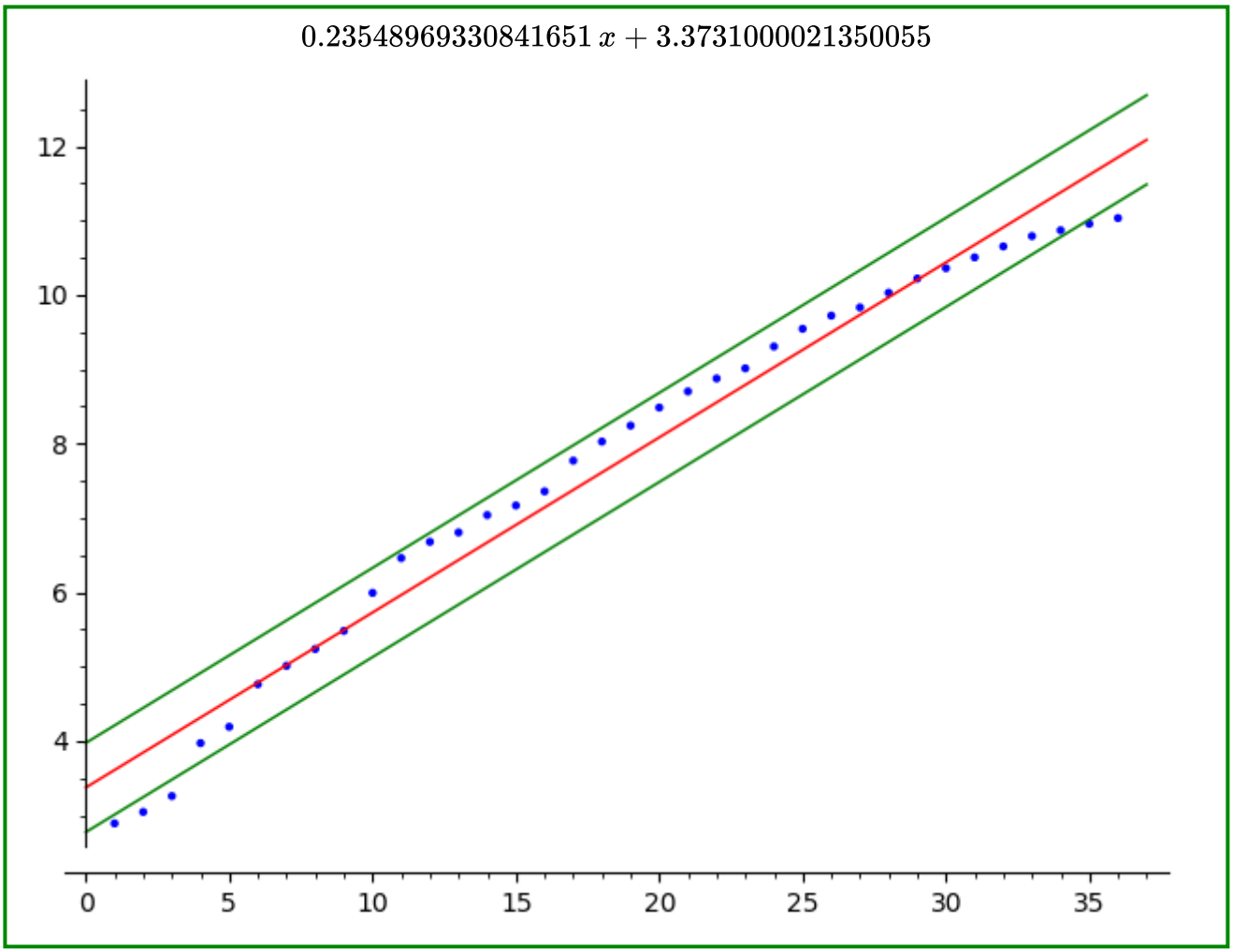
Qualität der Modelle

Die zu erwartende Genauigkeit einer Vorhersage hängt offenbar vom Zeithorizont zh der Voraussage ab - davon, über wie viele Tage im Voraus eine Aussage gemacht werden soll. Den Zeithorizont, mit dem im Folgenden gerechnet wird, können Sie mit diesem Regler festlegen.

zh (Tage) 3

Der Zeithorizont für die Vorhersagen beträgt 3 Tage.

$0.23548969330841651\,x + 3.3731000021350055$



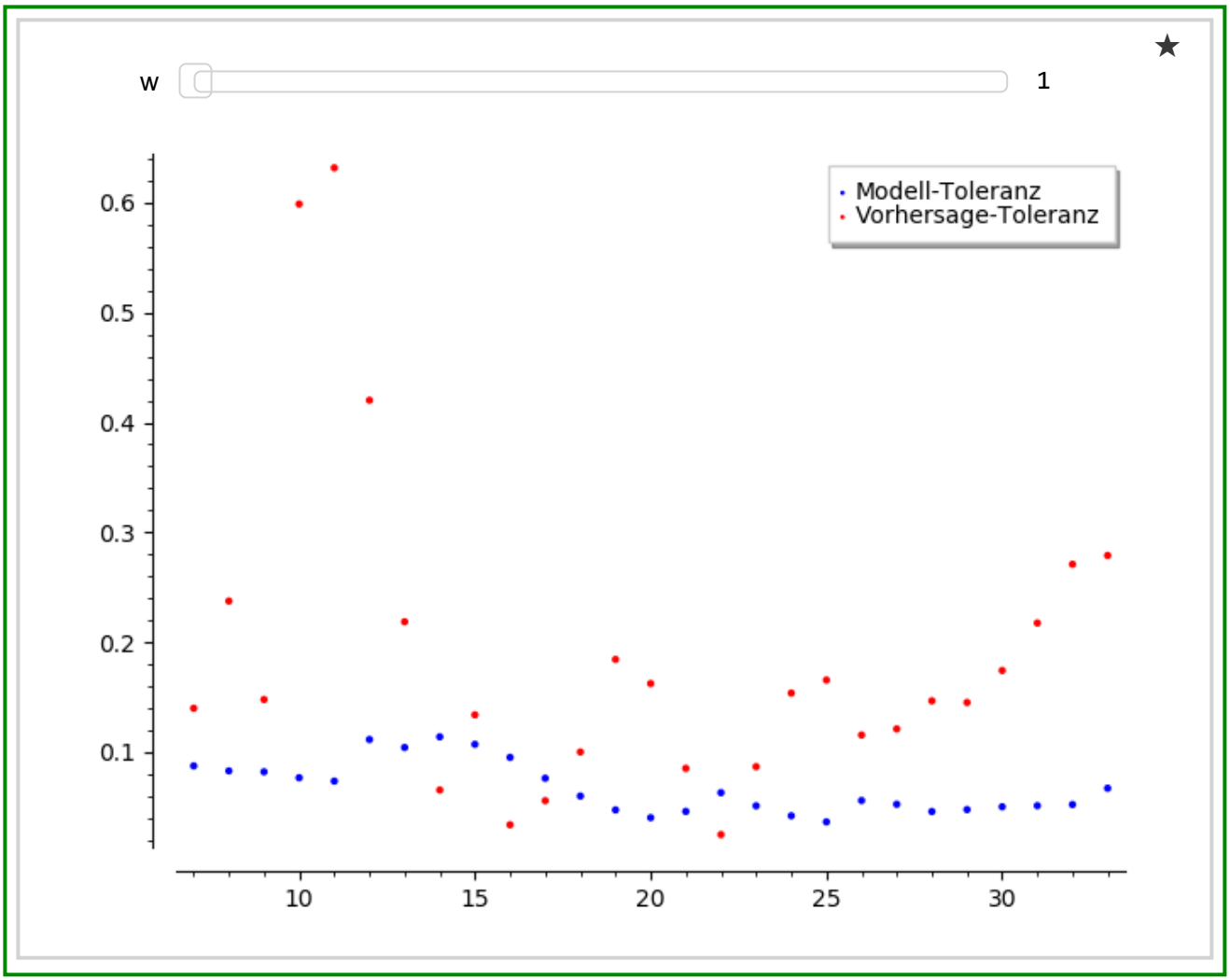
Help (https://sagecell.sagemath.org/help.html) | Powered by SageMath (http://www.sagemath.org)

Wie man sieht, verlassen die Logarithmen der kumulierten Infektionszahlen um den Tag 36 (31.3.2020) den Korridor der zufälligen Schwankungen. Dieser Trend setzte sich in den folgenden Tagen fort wie man sehen kann, wenn man die Berechnungen mit $dataExpNr > 36$ wiederholt.

Sehen wir uns die Entwicklung der Modell-Toleranz und der Vorhersage-Toleranz während der exponentiellen Phase an. Wir sehen, dass eine gute Passung des Modells zu den bisher beobachteten Daten kein Garant für eine gute Vorhersage ist. Dies gilt insbesondere in der frühen Phase, als noch relativ wenige Daten vorlagen, als auch gegen Ende der exponentiellen Phase, als das Modell beginnt, seine Passgenauigkeit zu verlieren.

Es könnte sein, dass eine kontinuierliche Verschlechterung der Vorhersagequalität ein Frühindikator dafür ist, dass das Modell an seine Grenzen stößt, auch wenn die bisherige Passung des Modells zu den Daten noch recht gut ist.

Mit dem Schieberegler im folgenden Diagramm können Sie die Werte dadurch glätten, dass jeweils mit dem arithmetischen Mittel der kumulierten Infektionszahlen der letzten w Tage gerechnet wird ($w = 0 \ldots zh$).



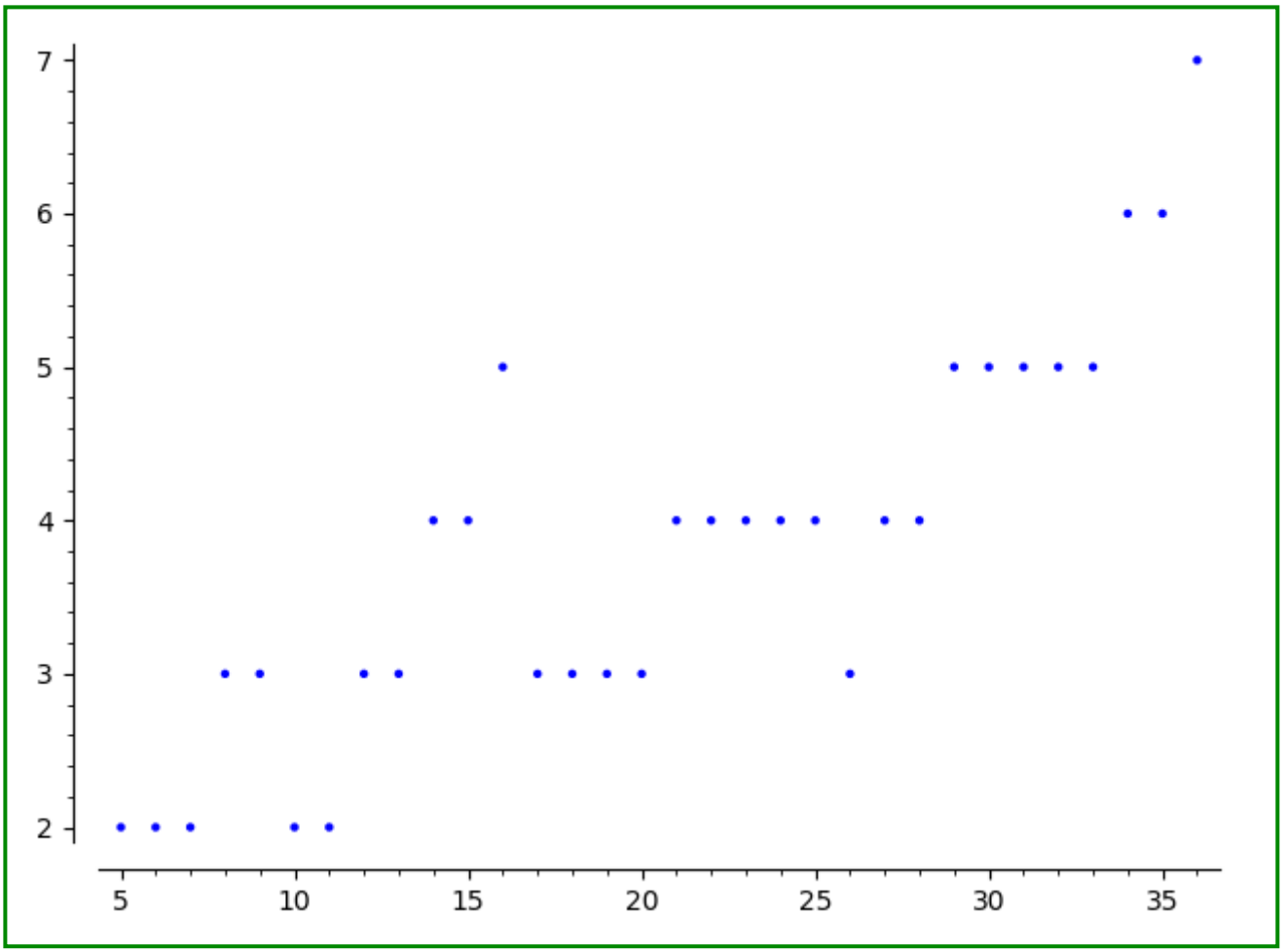
Help (https://sagecell.sagemath.org/help.html) | Powered by SageMath (http://www.sagemath.org)

Verdopplungsrate

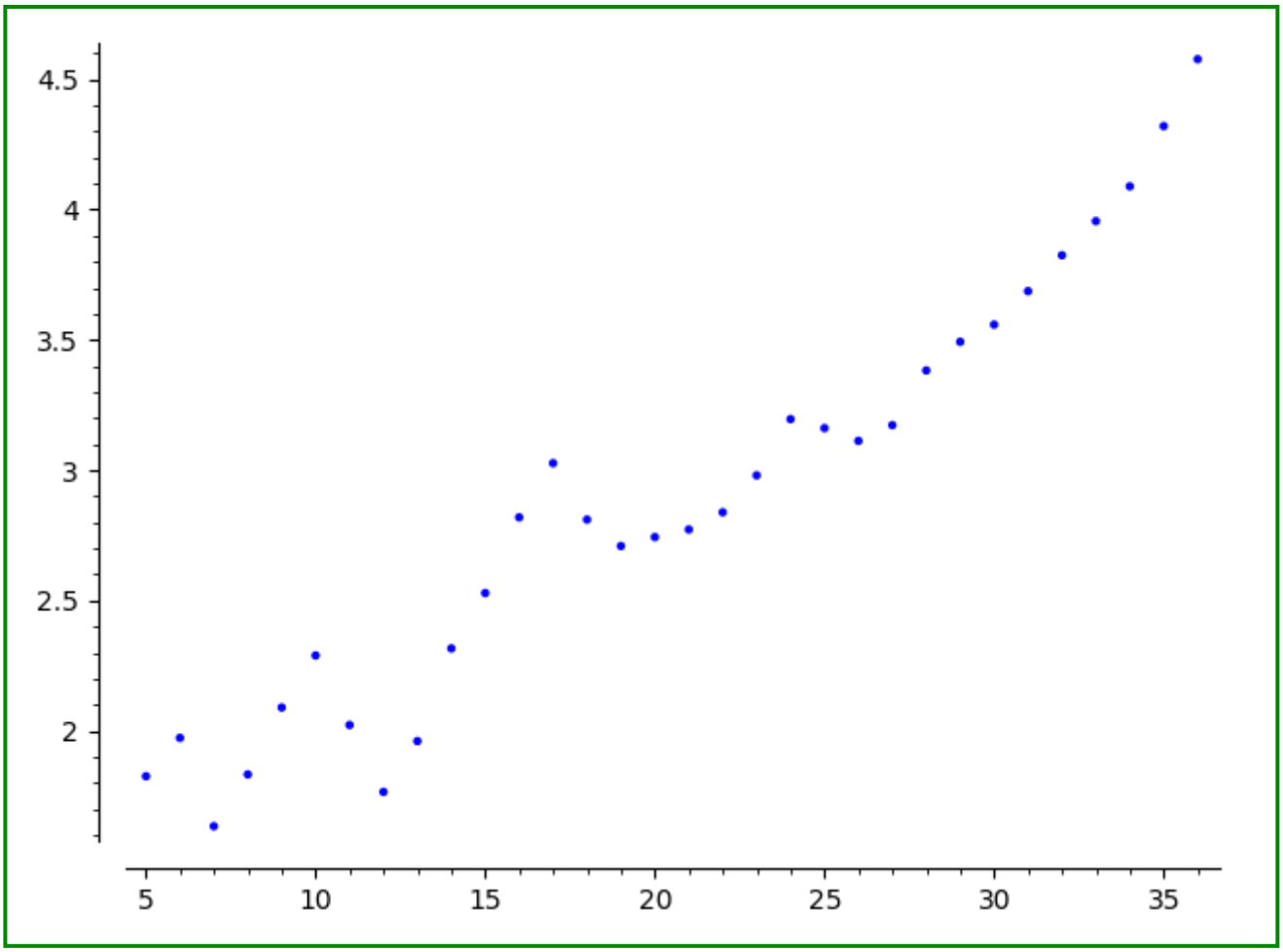
Am 28.3.2020 erklärte Kanzleramtschef Braun (https://www.tagesspiegel.de/politik/kanzleramtschef-erteilt-rascher-lockerung-eine-absage-bis-20-april-bleiben-alle-coronavirus-massnahmen-bestehen/25690036.html):

> Wenn wir es schaffen, die Infektionsgeschwindigkeit so zu verlangsamen, dass wir zehn, zwölf oder noch mehr Tage haben bis zu einer Verdopplung, dann wissen wir, dass wir auf dem richtigen Weg sind.

Wir definieren die Verdopplungsrate zu einem Zeitpunkt x als die kleinste Zahl d, so dass die kumulierte Zahl der Infizierten zum Zeitpunkt $x - d$ höchstens halb so groß ist, wie die kumulierte Zahl der Infizierten zum Zeitpunkt x. Betrachten wir die Verdopplungsrate in der exponentiellen Phase im März 2020.



Wie in dieser Aufgabe berechnen wir die Verdopplungsrate des zu einem Zeitpunkt besten exponentiellen Modells im Laufe des März 2020:



Help (https://sagecell.sagemath.org/help.html) | Powered by SageMath (http://www.sagemath.org)

Wir bemerken, dass die Entwicklung der Verdopplungsrate der exponentiellen Modelle ein deutlicheres Bild von der Entwicklung der Pandemie vermittelt als die tag-genaue Bestimmung der Verdopplungsrate in den realen Daten, ohne Berücksichtigung der Modelle.

Da die Berechnungen nur taggenau erfolgen können Verzögerungen in der Meldepraxis durchaus zu Sprüngen von ± 1 Tag führen. Wie wir sehen, ist die Verdopplungsrate im März 2020 weit vom Zielwert von 10-12 entfernt, auch wenn ein Anstieg festzustellen ist.

Nun hat jede Exponentialfunktion e^{kx} eine konstante Verdopplungsrate von $\frac{\ln 2}{k}$. Eine Änderung der Verdopplungsrate zeigt also an, dass der Parameter k des exponentiellen Modells angepasst werden muss - bzw. dass das exponentielle Modell an die Grenzen seiner Leistungsfähigkeit stößt.

In der folgenden Aufgabe geht es um die Berechnung der Verdopplungsrate - nicht aufgrund der realen Daten sondern aufgrund des Ende März bestmöglichen exponentiellen Modells. Vergleichen Sie die Ergebnisse mit obigem Diagramm.

Wenn Sie auf "Auswerten" klicken, so wird eine Exponentialfunktion $a \cdot e^{bx}$ berechnet, die die Entwicklung der Zahl der bestätigten Corona-Infektionen in Deutschland (24.2.2020 - 31.3.2020) (nach Informationen des Robert-Koch-Instituts) annähert.

```
1  data=[(0,16),(1,18),(2,21),(3,26),(4,53),(5,66
2  var('a,b')
3  f(x) = a*e^(x*b);
4  q=find_fit(data, f, solution_dict = True)
5  show(f(a=q[a],b=q[b]))
6  list_plot(data)+plot(f(a=q[a],b=q[b]), 0, 36,
```

Auswerten

Die Zahl der Infizierten verdoppelt sich alle _____ Tage. Übernehmen

Tip

Antwort prüfen Neue Frage

Logistisches Modell

Ab Anfang April 2020, kann das Wachstum der Zahl der gemeldeten Infektionen immer schlechter durch Exponentialfunktionen beschrieben werden. Das Virus stößt - aus welchen Gründen auch immer - auf Faktoren, die seine Ausbreitung behindern (z.B. Senkung der Reproduktionsrate durch Einschränkung sozialer Kontakte, hoher Grad der Durchseuchung der Bevölkerung).

Solche begrenzenden Faktoren werden im logistischen Modell (SI-Modell (https://de.wikipedia.org/wiki/SI-Modell)) berücksichtigt. Dieses Modell nimmt an, dass es eine obere Grenze N für die Zahl der Infizierten gibt. Dabei kann N durchaus kleiner als die reale Gesamtbevölkerung sein, denn manche Teile der Bevölkerung kommen z.B. durch Kontaktbeschränkungen, Vorsichtsmaßnahmen oder geographische Beschränkungen nie in Kontakt mit dem Virus. Die Gruppe der für die Virusausbreitung relevanten N Personen wird zu jedem Zeitpunkt x unterteilt in die $I(x)$ infizierten und die $S(x)$ noch infizierbaren (Susceptibles). Damit ist für alle x $S(x) + I(x) = N$ - $S(x)$ ergibt sich für festes x aus $I(x)$ durch $S = N - I$.

Wir bemerken, dass $\frac{S(x)}{N}$ die Wahrscheinlichkeit dafür ist, dass eine zum Zeitpunkt x zufällig ausgewählte Person infizierbar ist. Das logistische Modell nimmt an, dass der Zuwachs an Infektionen in einem kleinen Zeitraum proportional zur Länge des Zeitraums, zur Zahl der zu Beginn Infizierten und zu dieser Wahrscheinlichkeit ist:

$$I(x + \Delta x) - I(x) = c \cdot \Delta x \cdot \frac{S(x)}{N} \cdot I(x) = c \cdot \Delta x \cdot \frac{N - I(x)}{N} \cdot I(x)$$

Daraus ergibt sich für $\Delta x \to 0$ die logistische Differentialgleichung

$$I' = c \cdot \frac{N - I}{N} \cdot I.$$

Die allgemeine Lösung der logistischen Differentialgleichung (http://statistik.wu-wien.ac.at/~leydold/MOK/HTML/node183.html) wird durch eine logistische Funktion

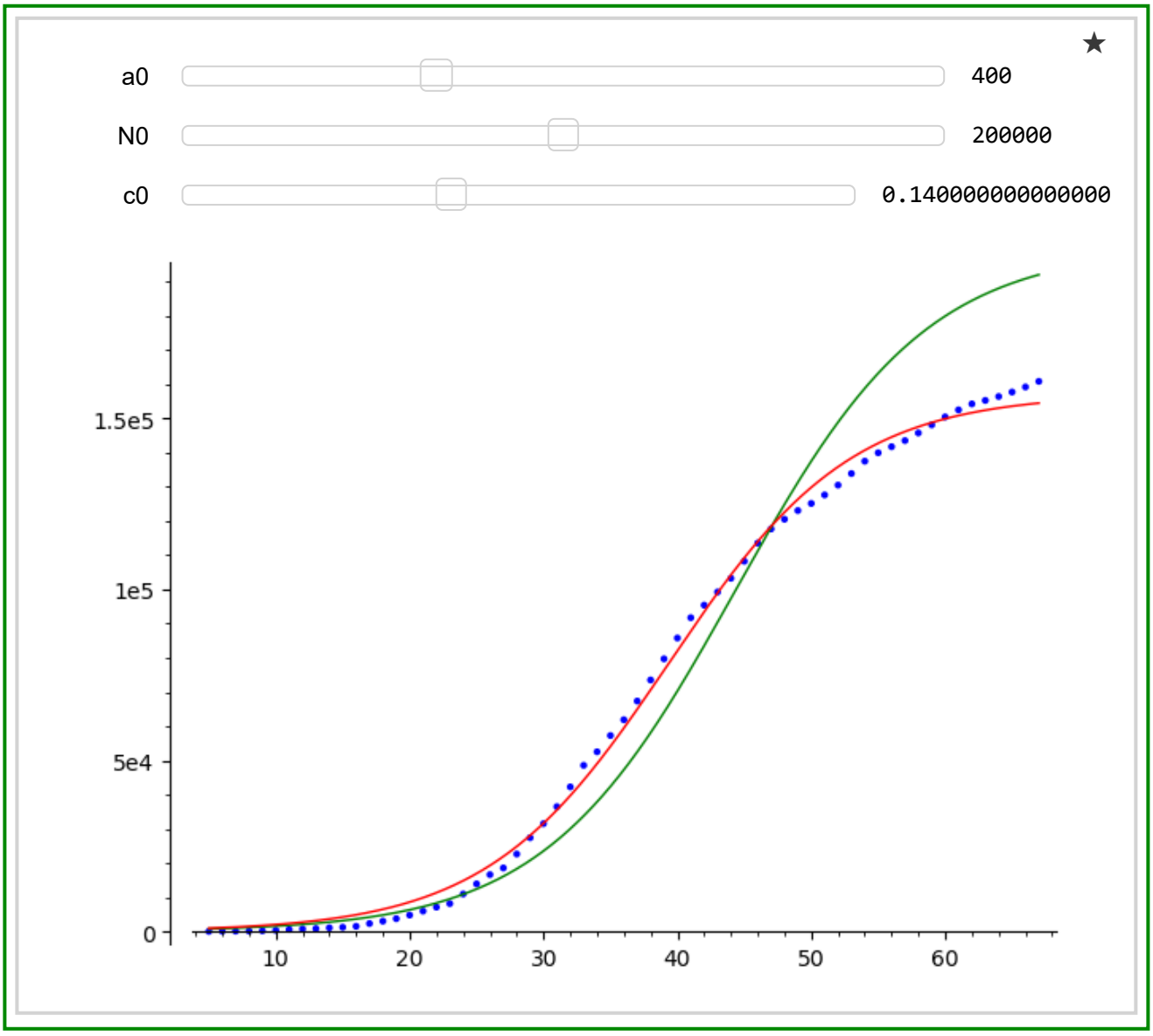
$$I(x) = a \cdot \frac{N}{a + (N - a) \cdot e^{-cx}}$$

beschrieben. Dabei ist a der Anfangswert bei $x = 0$, N ist eine angenommene Sättigungsgrenze, der sich die Zahl der Infizierten asymptotisch annähert (Kapazitätsgrenze) und c ist ein Parameter der die Geschwindigkeit dieser Annäherung beschreibt.

Das logistische Modell betrachtet alle Infizierten als infektiös, berücksichtigt also Todesfälle und Immunität nicht. Dies erfolgt im verfeinerten SIR-Modell.

Versuchen wir nun automatisch eine logistische Funktion für die Entwicklung der COVID-19-Pandemie in Deutschland zu bestimmen, welche die Zahl der Infizierten bis Ende März (Tag 67) möglichst gut beschreibt. Als Startwerte für die Regression nehmen wir $a = 400$ und $k = 0.14$, was den Koeffizienten des exponentiellen Modells Ende März entspricht und $S = 2000000$, was von der Größenordnung her plausibel erscheint.

Mit den Schiebereglern können Sie die Startwerte ändern. Die Startfunktion wird grün dargestellt, das Ergebnis der logistischen Regression in rot.



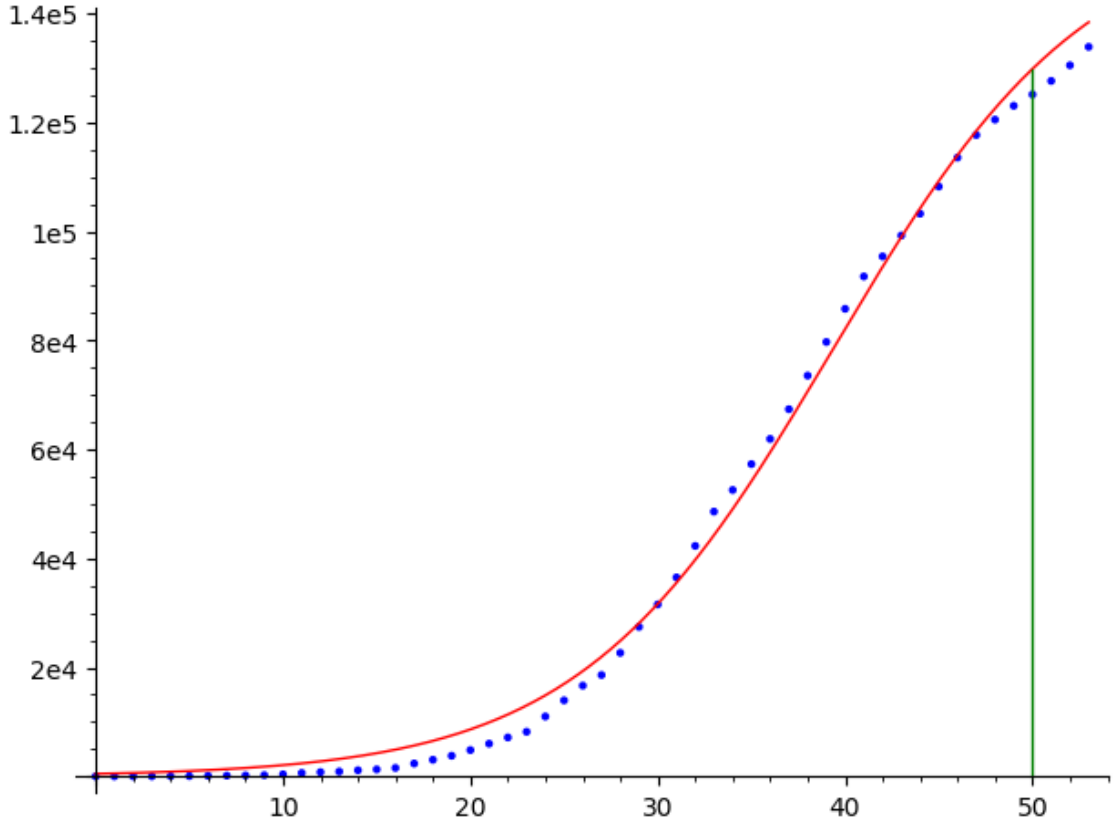
Help (https://sagecell.sagemath.org/help.html) | Powered by SageMath (http://www.sagemath.org)

Wie für das exponentielle Modell können nun für jeden Tag (grün) Modell- und Vorhersage-Toleranz des optimalen logistischen Modells (rot) berechnen.

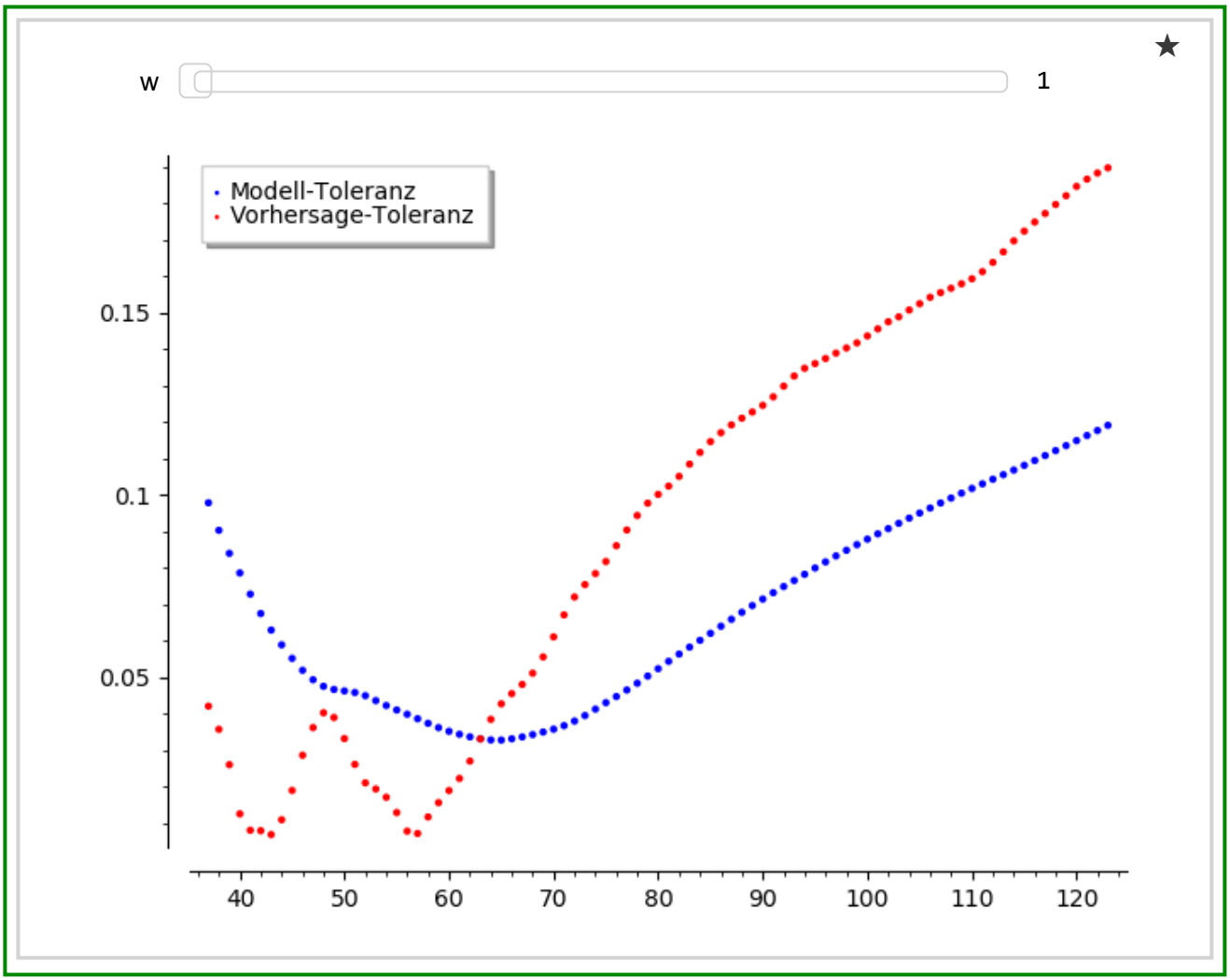
Frage: Wie verhalten sich Toleranzen des logistischen Modells und des exponentiellen Modells (s.o.) in der exponentiellen Phase (also vor Tag 36) zueinander?

Tag 50

a=483.4611401420046 N=157086.77914178697 c=0.14679874435677764
Modell-Toleranz:0.0467073035254511
Vorhersage-Toleranz:0.0390077318915248



Das folgende Diagramm stellt die Entwicklung der Modell- und Vorhersage-Toleranz für das SI-Modell nach dem Ende der exponentiellen Phase bis Ende Juni 2020 dar. Mit dem Regler $w = 0 \ldots zh$ kann bestimmt werden, ob mit den Original-Infektionszahlen ($w = 1$) oder mit dem gleitenden Durchschnitt der letzten w Tage gerechnet wird.



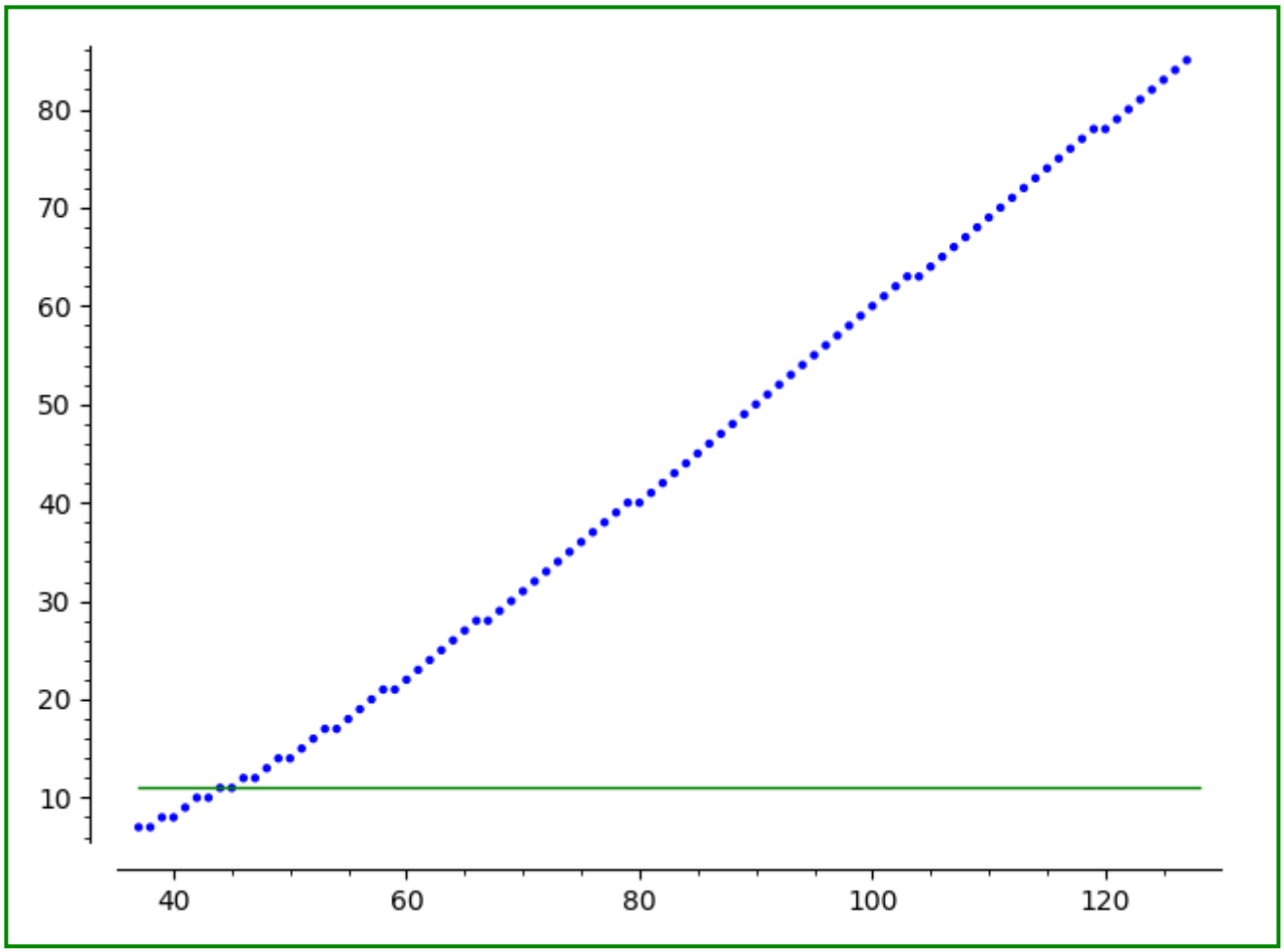
Help (https://sagecell.sagemath.org/help.html) | Powered by SageMath (http://www.sagemath.org)

Wie wir sehen, erreicht das logistische Modell um Tag 64 (28.4.2020) mit Modell- und Vorhersage-Toleranzen von etwa 3.3% seine maximale Leistungsfähigkeit bei einem 3-Tages-Horizont für die Vorhersagen.

Bis zu diesem Zeitpunkt ist die Vorhersage-Toleranz sogar besser, als die Modell-Toleranz. Wir können dies dadurch erklären, dass sich das Verhalten der Pandemie von Tag zu Tag besser durch ein logistisches Modell beschreiben lässt.

Verdopplungsraten

Sehen wir uns zunächst die realen Verdopplungszahlen an, d.h. wir bestimmen für die Zeit nach der exponentiellen Phase zu jedem Tag x wie viele Tage vorher die kumulierten Infektionszahlen halb so hoch waren. Die grüne Linie zeigt die angestrebte Verdopplungsrate von 11.



Die Verdopplungsrate von 11 Tagen wird am Tag 44 (8.4.2020) erreicht. Wir beobachten eine kontinuierlichere - sogar fast lineare -Entwicklung als in der exponentiellen Phase. Dies wird noch deutlicher, wenn man den Beginn des dargestellten Zeitraums `dataExpMr_de` etwa durch 5 ersetzt.

Wie für Exponentialfunktionen können wir auch für logistische Funktionen $L(x) = a\frac{N}{a+(N-a)e^{-cx}}$ eine Verdopplungsrate berechnen. Gesucht wird dafür zu gegebenem x ein Wert d, so dass $L(x-d) = \frac{1}{2}L(x)$. Dieser Wert d hängt natürlich von a, S, c ab.

SageMath liefert uns eine Lösung für diese Gleichung und damit für die Verdopplungsrate des jeweiligen logistischen Modells:

$$d(a,N,c,x) = \frac{cx + \log\left(\frac{\left(ae^{(cx)}+2\,N-2\,a\right)e^{(-cx)}}{N-a}\right)}{c}$$

Damit könnte man versuchen, die Verdopplungsrate für die Zukunft vorherzusagen. Es ist jedoch zu erwarten, dass im Falle der COVID-19-Pandemie eine einfache lineare Regression der realen Verdopplungsraten ein besseres Ergebnis liefert, als es logistische Modelle mit ihrer schließlich wachsenden Modell- und Vorhersage-Toleranz liefern könnten. Wir verzichten deshalb darauf, dies weiter zu untersuchen.

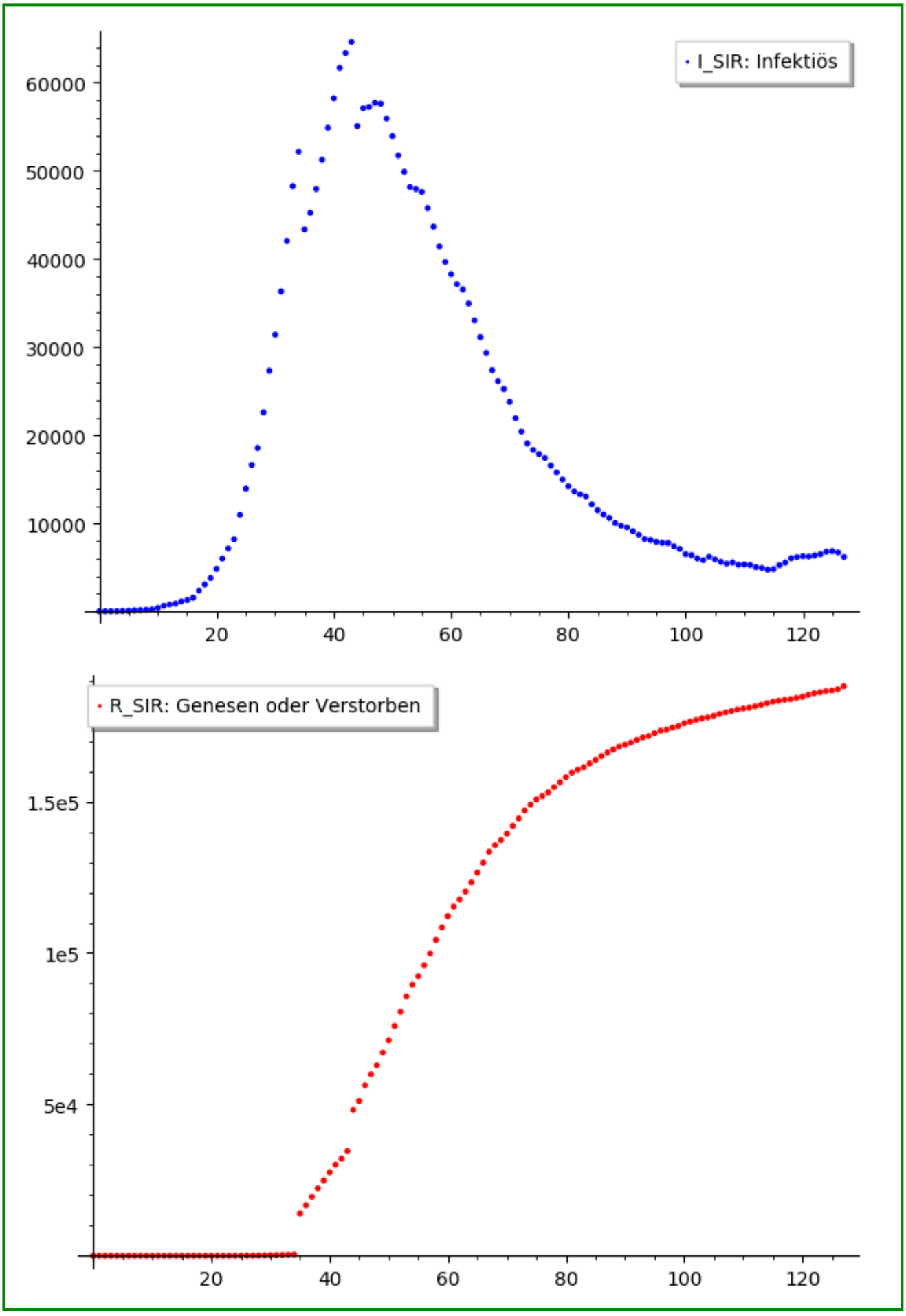
SIR-Modell

Das SIR-Modell unterscheidet sich vom exponentiellen und logistischen (SI-)Modell dadurch, dass es Genesene und Verstorbene berücksichtigt und davon ausgeht, dass diese nicht mehr infektiös sind. Wir fassen Genesene und verstorbene in der Gruppe R (recovered) zusammen. Die Größe dieser Gruppe zum Zeitpunkt x bezeichnen wir mit $R(x)$.

Das SIR-Modell versucht, Funktionen zu berechnen, die

- die Zahl der Infektiösen I_{SIR}
- die Zahl der Infizierbaren S_{SIR} und
- die Zahl der Genesenen bzw. Verstorbenen R_{SIR}

beschreiben. Es gilt also für die im exponentiellen und logistischen Modell approximierte kumulative Zahl der Infizierten $I = I_{SIR} + R_{SIR}$ bzw. $I_{SIR} = I - R_{SIR}$. Die Zahl der Verstorbenen und Genesenen haben wir zu Beginn in die Variablen `deadsAll_de` bzw. `recoveredAll_de` abgespeichert und können damit nun auch R_{SIR}, I_{SIR} berechnen:

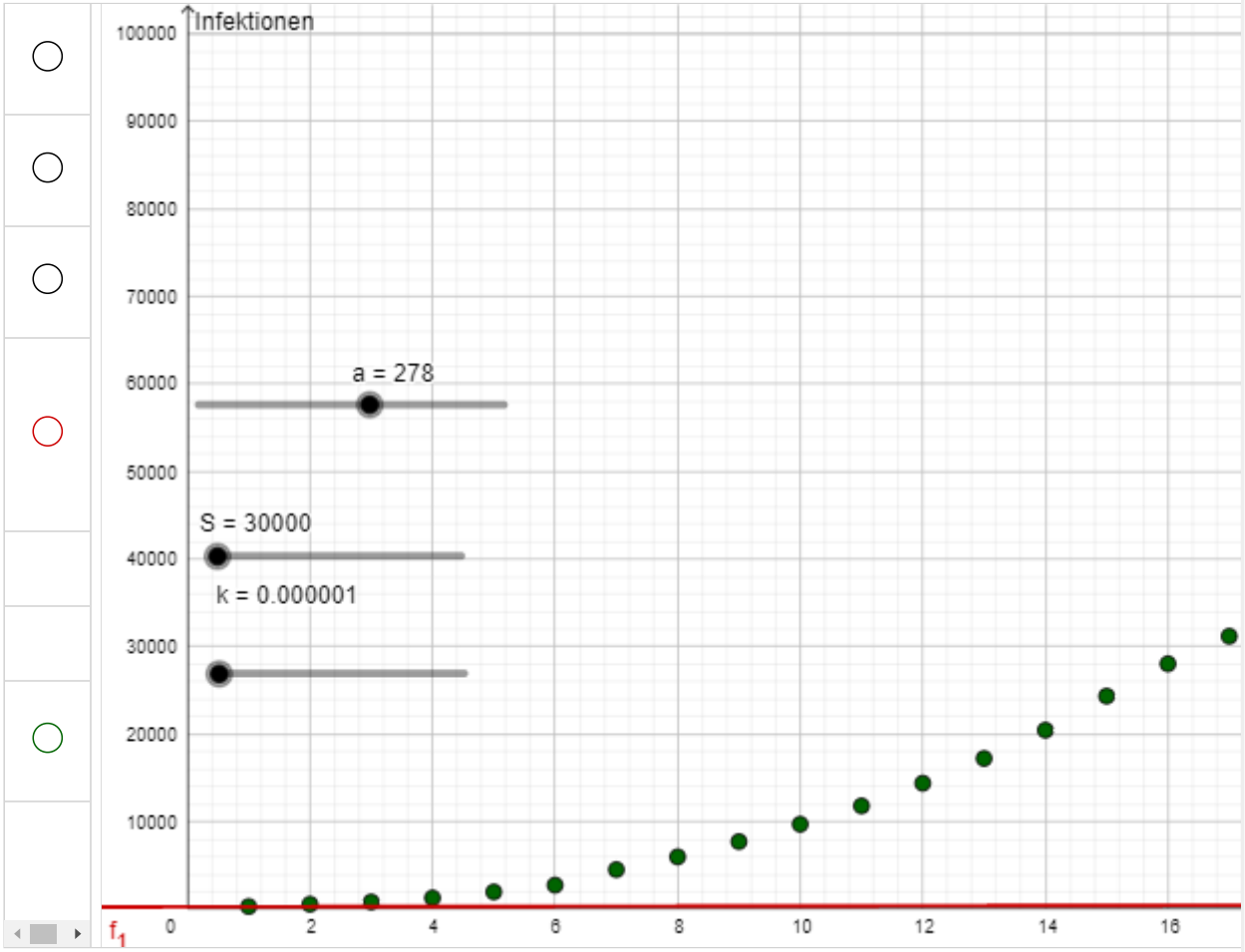


Zahl der Infektionen

Versuchen Sie in der folgenden Aufgabe mit Hilfe der Schieberegler eine logistische Funktion zu finden, die möglichst genau zu den Daten passt, die den Verlauf der Epidemie in der VR China im Januar 2020 darstellen.

In der Literatur werden unterschiedliche Bezeichnungen verwendet; so werden unsere Konstanten N, c in der folgenden Aufgabe mit $S, k \cdot S$ bezeichnet.

Das folgende Diagramm zeigt die Anzahl der im Labor bestätigten Infektionen mit dem Coronavirus in China nach WHO-Daten; Tag 0 ist der 21.1.2020 (der Beginn der täglichen Reports der WHO):



Durch Ziehen der Regler setzen Sie die Parameter S und k für die logistische Funktion $a \cdot \frac{S}{a + (S - a) \cdot e^{-kS \cdot x}}$, die dann graphisch dargestellt wird; $a = 278$ ist der Wert am Tag 0.

Zeichnen Sie so eine Kurve, die die Zahl der Infektionen möglichst gut annähert.

Die Berechnung der Zahl der Infizierbaren $S_{SIR} = N - I_{SIR} - R_{SIR}$ ist leider nicht unmittelbar möglich, da sie vom Parameter N des Modells abhängt, der zunächst nicht bekannt ist.

Funktion $S_{SIR}(N)$ definiert

Das SIR-Modell nimmt an, dass die Zahl der in einem kurzen Zeitraum Genesenen oder Verstorbenen proportional zur vergangenen Zeit und zur Zahl der Infizierten ist:

$$R_{SIR}(x + \Delta x) - R_{SIR}(x) = w \cdot \Delta x \cdot I_{SIR}(x)$$

Daraus ergibt sich die Differentialgleichung

$$R'_{SIR} = w \cdot I_{SIR}$$

Wir bemerken, dass bei dieser Annahme nicht berücksichtigt wird, dass Genesung bzw. Versterben erst mehrere Tage (10-14 Tage) nach (Registrierung der) Infektion eintreten, so dass eher von einer Proportionalität mit $I_{SIR}(x - 10)$ auszugehen wäre. Dies wird im SIR-Modell aber nicht berücksichtigt.

Die Veränderung der Anzahl der Infizierten ergibt sich aus der Zahl der Neuinfektionen, die wie im SI-Modell berechnet wird, vermindert um die Zahl der in diesem Zeitraum Genesenen bzw. Verstorbenen:

$$I_{SIR}(x + \Delta x) - I_{SIR}(x) = \Delta x \cdot c \cdot \frac{S_{SIR}(x)}{N} \cdot I_{SIR}(x) - (R_{SIR}(x + \Delta x) - R_{SIR}(x))$$

woraus sich für $\Delta x \to 0$ die Differentialgleichung

$$I'_{SIR} = c \cdot \frac{S_{SIR}}{N} \cdot I_{SIR} - R'_{SIR} = c \cdot \frac{S_{SIR}}{N} \cdot I_{SIR} - w \cdot I_{SIR}$$

ergibt.

Aus diesen beiden Differentialgleichungen und dem Fakt, dass die Anzahl der Infizierten, Genesenen bzw. Verstorbenen und Infizierbaren als konstant gleich N angenommen wird ergibt sich schließlich die dritte Differentialgleichung

$$S'_{SIR} = -c\frac{S_{SIR}}{N} I_{SIR}$$

Wir bemerken, dass das SI-Modell der Spezialfall des SIR-Modells mit der Genesungsrate $w = 0$ ist.

Wir speichern die rechten Seiten des SIR-Differentialgleichungssystems

$S'_{SIR} = -c\frac{S_{SIR}}{N} I_{SIR}$

$I'_{SIR} = c \cdot \frac{S_{SIR}}{N} \cdot I_{SIR} - w \cdot I_{SIR}$

$R'_{SIR} = w \cdot I_{SIR}$

in der Funktion `DGL_right` ab, die von den Parametern N, c, w abhängt:

Differentialgleichungssystem definiert

Leider lässt sich diese Differentialgleichungssystem nicht in geschlossener Form lösen, so dass wir mit numerischen Näherungslösungen vorlieb nehmen müssen. SageMath stellt uns dafür die Funktion `desolve_system_rk4` zur Verfügung, die Differentialgleichungssysteme numerisch mit dem Runge-Kutta-Verfahren (https://de.wikipedia.org/wiki/Klassisches_Runge-Kutta-Verfahren) löst.

Dabei werden die Anfangswerte zum Beginn einer zu untersuchenden Periode vorgegeben und das Runge-Kutta-Verfahren berechnet schrittweise den Verlauf der Lösungen des Differentialgleichungssystems während dieser Periode. Um Datenreihen für die Lösung des SRI-Differentialgleichungssystems ab einem Tag x_0 zu berechnen benötigen wir also

- Werte für die Parameter N, c, w des Differentialgleichungssystems und
- Anfangswerte $i_0 = I_{SRI}(x_0), r_0 = R_{SRI}(x_0), s_0 = S_{SRI}(x_0)$.

Außerdem benötigen wir x_0 als Start der Periode, das Ende der Periode und die Schrittweite für die x-Werte, für die die Lösung berechnet werden soll.

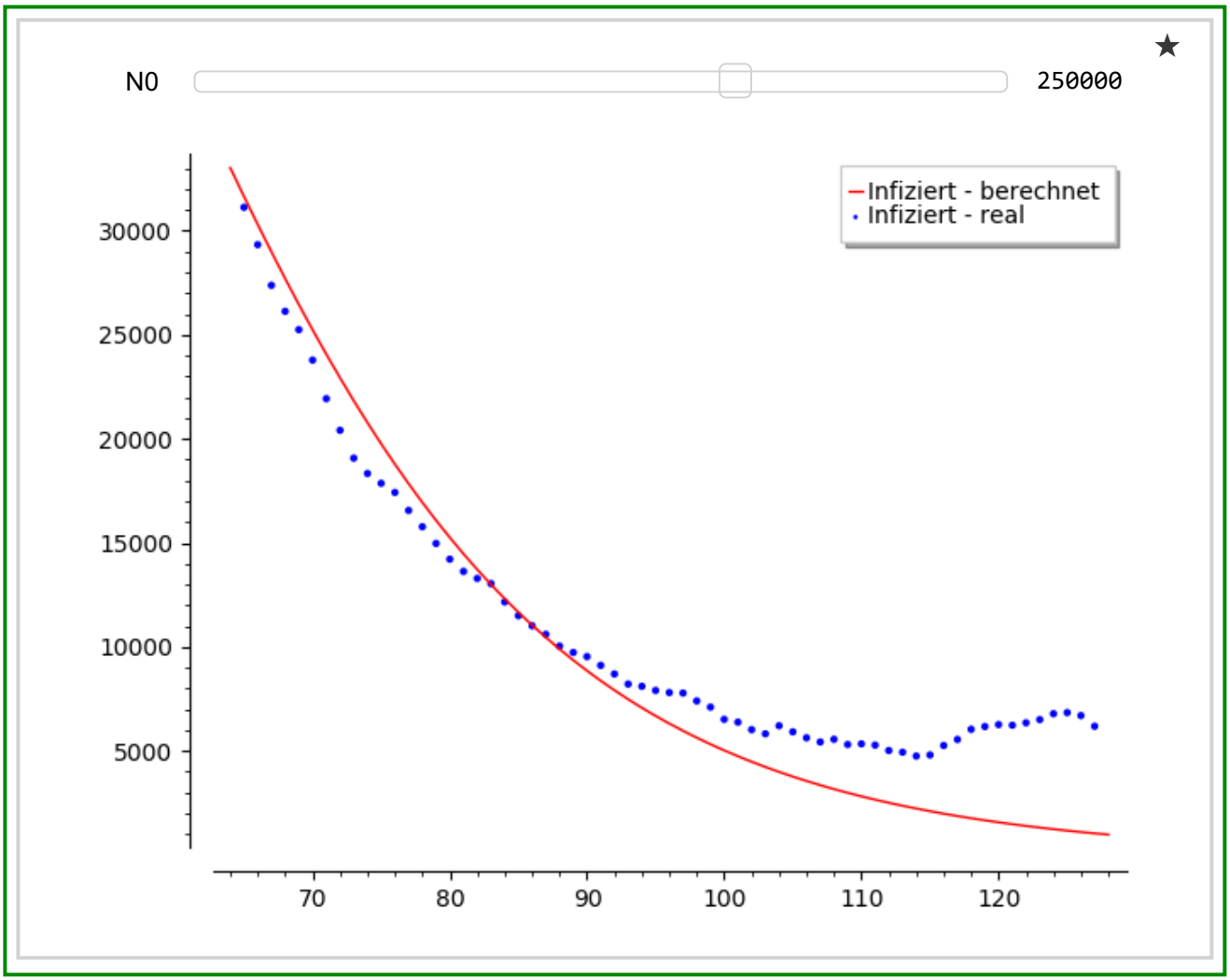
Lösung des DGL-Systems definiert

Berechnen wir als Erstes einmal eine Lösung für den Tag $x_0 = 64$ (28.4.2020), den Tag der besten Passung des logistischen Modells. Die logistische Regression lieferte für diesen Tag $N_0 = 157086.77914178697, c_0 = 0.14679874435677764$. Für w_0 wählen wir - angeregt durch die dritte Differentialgleichung unseres System $w_0 = \frac{R_{SIR}(x_0+1)-R_{SIR}(x_0)}{I_{SIR}(x_0)}$. Als Anfangswerte wählen wir $I_0 = I_{SIR}(x_0), R_0 = R_{SIR}(x0), S_0 = N_0 - R_0 - I_0$.

```
w0= 0.0969597868217054
I0= 33024.0000000000  R0= 123313.000000000  S0= 749.779141786974
```


Was dabei auffällt: Der berechnete Wert S_0 von etwa 750 noch zu infizierenden ist - verglichen mit den Werten für I_0, R_0 - sehr klein. Dies könnte darauf hindeuten, dass N_0 zu klein gewählt ist.

Aufgabe: Versuchen Sie, mit dem Schieberegler einen Wert für N_0 zu finden, der zu einer möglichst guten Approximation der realen Zahl der Infizierten führt.



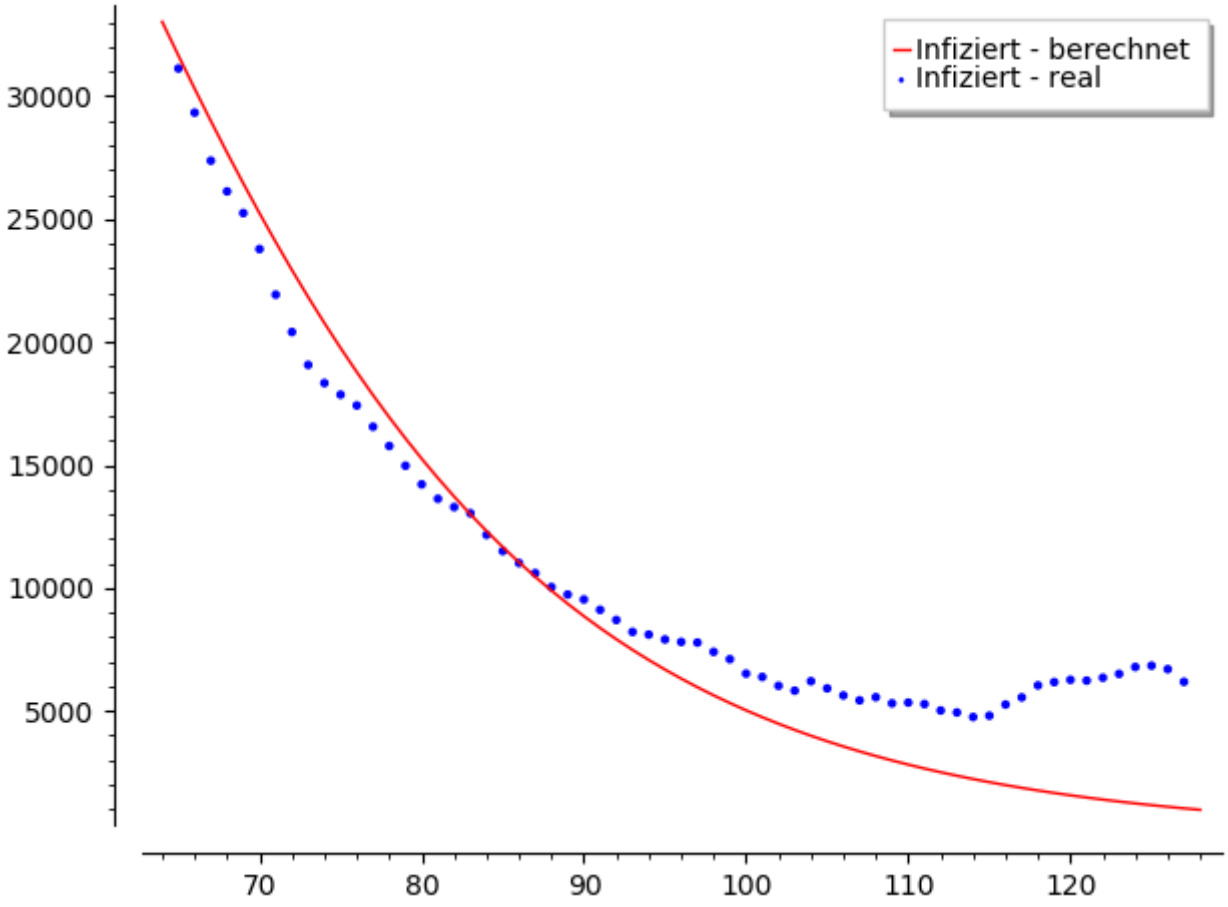
Da das Runge-Kutta-Verfahreneine Datenreihe, keine Funktion, liefert (auch wenn wir dies hier durch eine Kurve darstellen), müssen wir unser Toleranzmaß neu definieren, wobei Funktionswerte durch die Daten ersetzt werden:

Residuen sind für Datenreihen definiert

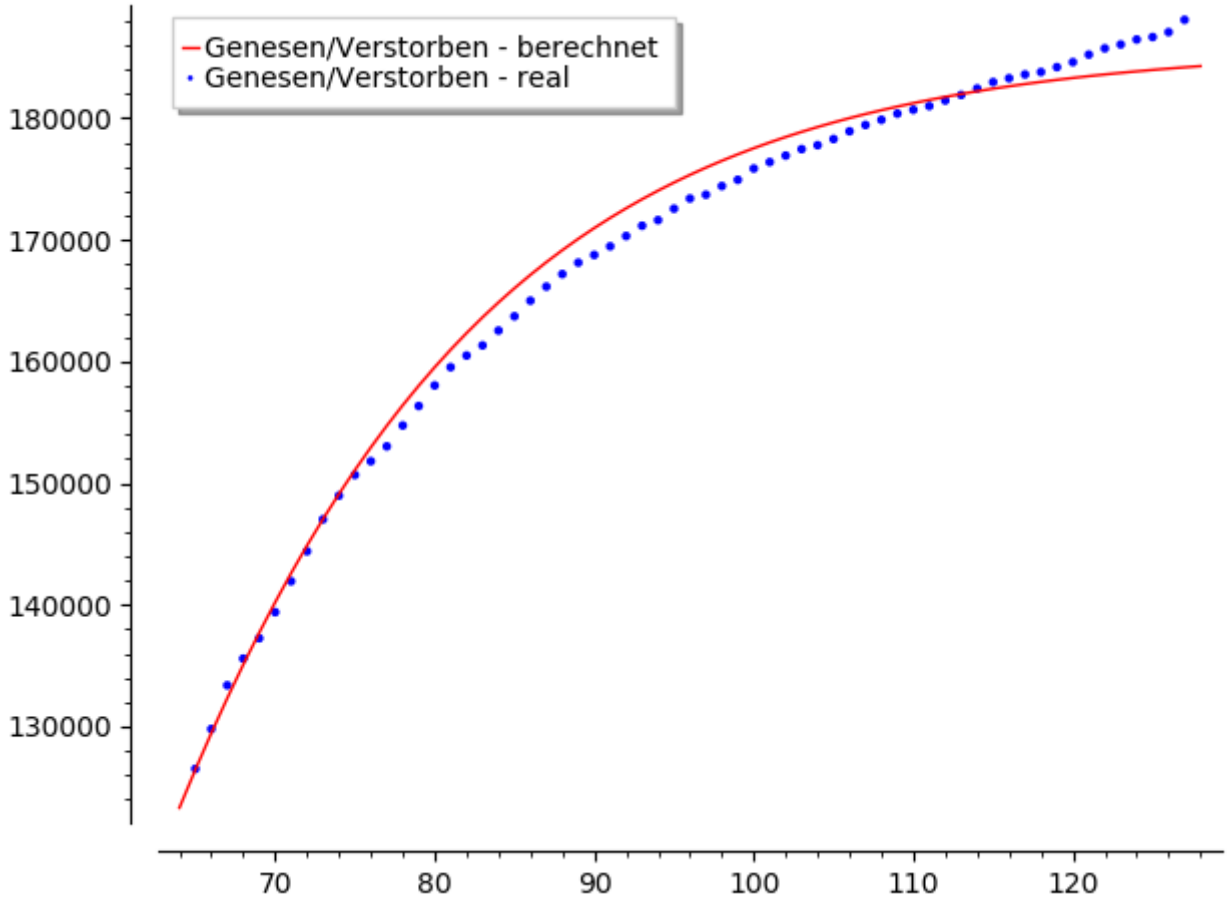
Da das Runge-Kutta-Verfahren historische Werte vor den Anfangswerten nicht approximiert, macht die Berechnung der Modell-Toleranz keinen Sinn. Wir berechnen stattdessen die Vorhersage-Toleranz für die Zeit von Tag 64 bis Tag 104 für die drei Datenreihen, die mit Hilfe des Runge-Kutta-Verfahrens gefunden werden.

Wir verwenden dafür den von Ihnen gefundenen Wert N_0 =250000. Passen Sie diesen Wert ggf. mit dem obigen Schieberegler an und berechnen Sie die folgende Zelle neu.

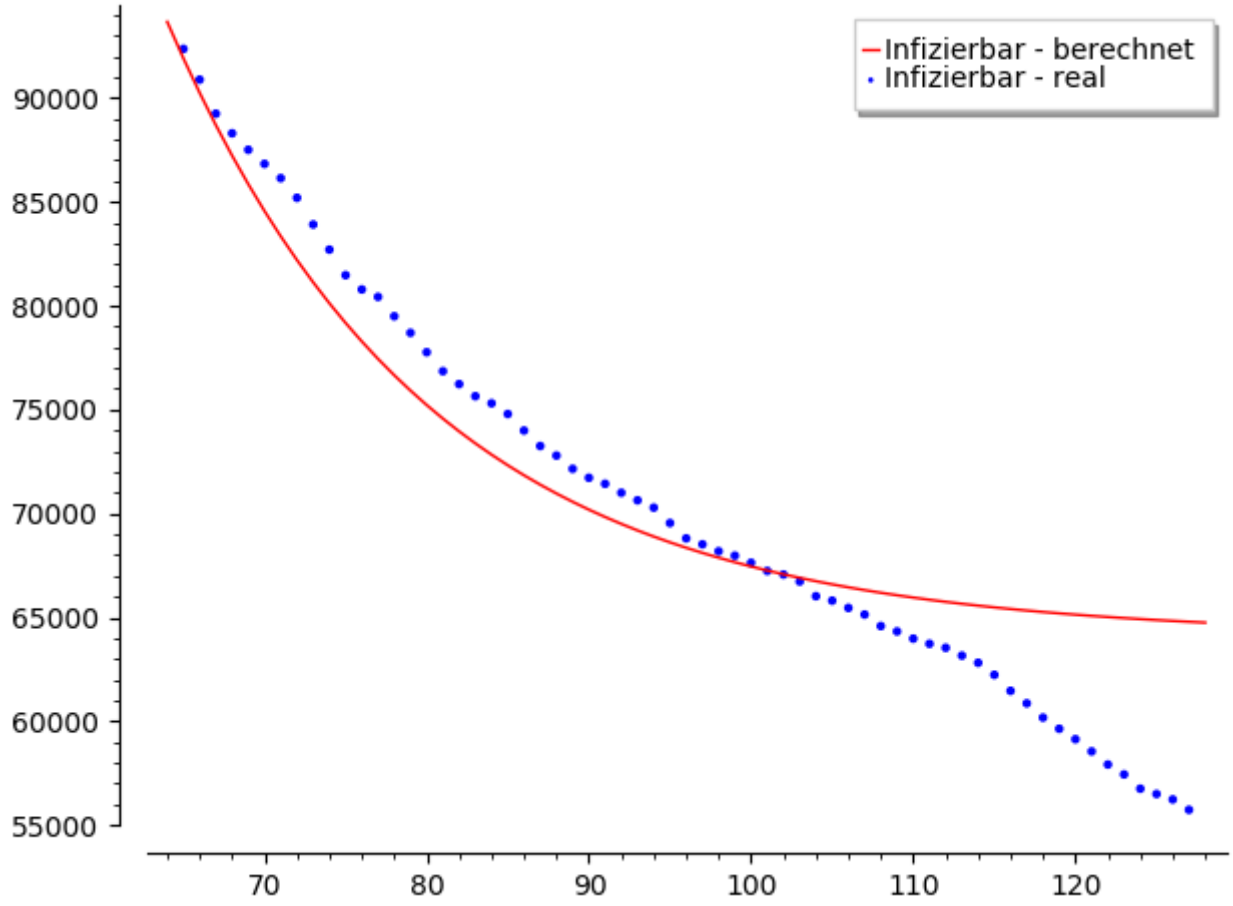
Vorhersage-Toleranz für Infizierte (Tag 64-104):0.084624



Vorhersage-Toleranz für Genesene/Verstorbene (Tag 64-104):0.010418



Vorhersage-Toleranz für Infizierbare (Tag 64-104):0.023824



Maximale Voehersage-Toleranz: 0.084624

Wenn Sie mit unterschiedlichen Werten für N_0, bzw. allgemein mit unterschiedlichen Parametern N, c, w für das SRI-Differentialgleichungssystem experimentieren so werden Sie bemerken, dass sich bei manchen Änderungen der Parameter die Vorhersage-Toleranz für eine der Datenreihen verbessert, während sie sich für eine andere verschlechtert. Somit ist es vom Ziel der Analyse abhängig, wie man ein Maß für die Qualität eines konkreten SIR-Modells definieren wird.

Fazit

Wir haben gesehen, dass die COVID-19-Pandemie sich in Deutschland 2020 in mehreren Phasen entwickelt hat, in denen Sie sich durch unterschiedliche mathematische Modelle mehr oder weniger gut beschreiben lässt.

- In der ersten Phase, etwa bis Ende März erfolgt die Ausbreitung des Virus im Wesentlichen ungehemmt, so dass sie sich am Besten mit einem exponentiellen Modell beschreiben lässt.
- In der zweiten Phase, die etwa bis Anfang Mai 2020 geht, wird die Ausbreitung des Virus gehemmt - das SI-Modell erweist sich als optimal zur Beschreibung der Situation.
- Die dritte Phase umfasst den Zeitraum von Anfang Mai bis Mitte Juni 2020. Die Lockdown-Maßnahmen wirken, Infektionsketten werden unterbrochen, immer mehr Personen scheiden nach überstandener Krankheit aus dem Infektionsgeschehen aus. Jetzt kann das SIR-Modell die Entwicklung am Besten beschreiben.
- Schließlich werden in Phase 4 Lockdown-Maßnahmen gelockert, das Infektionsgeschehen ist örtlich sehr unterschiedlich. Damit verlieren Deutschland-weite Modelle an Bedeutung; andere Modelle, insbesondere zur Modellierung der räumlichen Ausbreitung des Virus, gewinnen an Bedeutung.

Wir haben auch gesehen, wie Modelle altern. Um die Vorhersagequalität eines Modells zu erhalten ist es erforderlich, seine Parameter kontinuierlich anzupassen - solange wie dies möglich ist und kein kein besseres Modell zur Verfügung steht.

Ausblick

Dieses Notebook stellt notwendigerweise eine subjektive Auswahl der Möglichkeiten und Ansätze zur Modellierung einer Pandemie dar. So gibt es auch viele ebenso berechtigte Möglichkeiten zur Analyse der Pandemie-Daten, die sich durch Erweiterung und/oder Modifikation dieses Notebooks implementieren lassen. Hier einige Anregungen.

- Analyse der Todesfallzahlen, deren Daten in der Variablen `deadsAll_de` zur Verfügung stehen
- Regression der Parameter des SIR-Modells zur Modell-Optimierung
- Analyse der Entwicklung der für den jeweiligen Tag optimierten Modell-Parameter

So möge dieses Notebook vor allem Anregung zu eigener aktiver Beschäftigung mit mathematischen Modellen sein.