

# Epidemiemodellierung mit Kompartimentmodellen

am Beispiel der COVID-19-Pandemie in Deutschland

---

MOHAMAD AL FARHAN

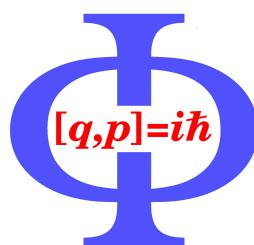
LARS TORBJØRN STUTZER

TORBEN SUNKEL

GEROLD WEBER



GEORG-AUGUST-UNIVERSITÄT  
GÖTTINGEN



*Projektpraktikum Sommersemester 2021*

*Projektbetreuer: PROF. DR. PETER SOLLICH*

*Institut für Theoretische Physik*

*März-Oktober 2021*

*Abgabe am 12.10.2021*

# Inhaltsverzeichnis

<b>1 Einleitung</b>	<b>1</b>
<b>2 Theorie und Methoden</b>	<b>3</b>
2.1 Kompartmentmodell . . . . .	3
2.1.1 SIR-Modell . . . . .	3
2.2 Netzwerkmodell . . . . .	3
2.2.1 Adjazenzmatrix . . . . .	3
2.2.2 Geographische Adjazenzmatrix . . . . .	4
2.2.3 Gewichtete Mobilitäts-Adjazenzmatrix (Pendlermatrix) . . . . .	4
2.2.4 Maß der Vernetzung . . . . .	5
2.3 Charakterisierung der betrachteten Region . . . . .	6
2.4 Erweiterung des Kompartmentmodells . . . . .	7
2.4.1 Der Tod . . . . .	7
2.4.2 Dynamik durch Pendler . . . . .	7
2.5 Die Dynamik an einem Tag . . . . .	8
2.5.1 Der Stufen-Ansatz . . . . .	9
2.5.2 Der konstante Ansatz . . . . .	9
2.6 Datenerhebung . . . . .	10
2.6.1 Verwendete Datensätze . . . . .	10
2.6.2 Zusammenstellung der benötigten Daten . . . . .	10
2.7 Optimierung der Parameter . . . . .	12
2.7.1 Auswahl des zu simulierenden Zeitraums . . . . .	13
2.8 Fehlerpropagation . . . . .	13
2.8.1 Monte-Carlo Verfahren . . . . .	15
2.9 Methoden . . . . .	15
2.9.1 Der klassische Runge-Kutta-Algorithmus . . . . .	15
2.9.2 Parallelisierung . . . . .	15
<b>3 Analyse</b>	<b>16</b>
3.1 Zu der Auswirkung der Parameter . . . . .	16
3.1.1 Variation von $t_0$ . . . . .	16
3.1.2 Variation der $\chi$ -Intervalle . . . . .	16
3.1.3 Das frühe Verhalten der Infektionswelle . . . . .	16
3.1.4 Grenzfall niedriger Anzahl von Infizierten . . . . .	17
3.2 Stabilität des Systems . . . . .	20
<b>4 Ergebnisse</b>	<b>21</b>
4.1 Optimierte Parameter der betrachteten Modelle . . . . .	21
4.2 Simulationen mit optimierten Parametern . . . . .	21
4.3 Bewertung der Qualität des Netzwerkmodells . . . . .	22
4.3.1 Untersuchung auf Korrelation zwischen Vernetzung eines Knotens und Qualität der Simulation . . . . .	22
4.3.2 Betrachtung im Netzwerk . . . . .	24
4.3.3 Vergleich von Simulationen mit verschiedenen breiten Rändern . . . . .	25
4.4 Modellunsicherheit . . . . .	25

<b>5 Zusammenföhrung und Diskussion</b>	<b>26</b>
5.1 Vereinfachungen . . . . .	26
5.1.1 Kompartimente . . . . .	26
5.1.2 Dynamik zwischen Zellen . . . . .	26
5.2 Räumliche Auflösung . . . . .	27
5.2.1 Superspreading . . . . .	28
5.3 Bewertung der Qualität des Modells . . . . .	28
5.4 Bewertung der Optimierung und Unsicherheit der Modelle . . . . .	29
5.5 Mögliche Erweiterungen für das Modell . . . . .	29
<b>Anhang</b>	<b>31</b>
<b>A Genauere Betrachtung der Jacobi-Matrizen</b>	<b>31</b>
<b>B Beweis von Gleichung 25</b>	<b>33</b>
<b>C Matrix der kleinen Parameter</b>	<b>33</b>
<b>D Weitere Plots</b>	<b>34</b>
D.1 Variation von $t_0$ . . . . .	34
D.2 Variation der Intervalle . . . . .	35
D.3 Netzwerkdiagramm des Systems zu verschiedenen Zeiten . . . . .	36
<b>E Weitere optimierte Parameter</b>	<b>37</b>
<b>F Abkürzungen</b>	<b>37</b>
F.1 Interne Identifikationsnummern und Kfz-Kennzeichen der Landkreise . . . . .	37
<b>Quellen</b>	<b>38</b>
Datenverzeichnis . . . . .	38
Literaturverzeichnis . . . . .	38

# 1 Einleitung

Nicht erst seit Beginn der COVID-19-Pandemie Anfang 2020 ist die mathematische Modellierung von zeitlichen und räumlichen Verläufen von Pandemien von Interesse der Wissenschaft. Das einfachste und älteste Modell wurde 1927 von W. O. KERMACK und A. G. MCKENDRICK entwickelt und war bereits in der Lage, die Pestausbrüche von London 1665/1666 und Bombay 1906 sowie die Choleraverbreitung in London 1865, für die Daten vorlagen, gut zu modellieren [1][2]. Dieses Modell teilt die Bevölkerung in sogenannte Kompartimente mit verschiedenen, zur Modellierung relevanten Eigenschaften auf: anfällig (susceptible), infektiös (infectious) und genesen oder verstorben (removed). Die Übergänge zwischen diesen Kompartimenten sind die Essenz der Epidemiemodellierung und werden durch gewöhnliche Differentialgleichungen beschrieben. Das ist das sogenannte SIR-Modell, von dem auch alle Überlegungen in diesem Projekt ausgehen.

Mit Hinblick auf die eklatanten Folgen von Epidemien für das gesellschaftliche Leben, reicht die Forderung, die Ausbreitung von Epidemien zu verstehen, weit über die Wissenschaft hinaus. Zum Zweck der Prävention zukünftiger Epidemien ist dieses Verständnis, dass durch das Studium vergangener Epidemien erwächst, unbedingt notwendig.

Allerdings sind in den vergangenen Jahrzehnten die Ansprüche an Epidemiomodelle wesentlich gewachsen. Einfache Modelle können der Berücksichtigung komplexerer Probleme wie steigender Bevölkerung, einer veränderten Wirtschafts- und Gesellschaftsstruktur, wachsender Urbanisierung und deutlich zunehmender Vernetzung der Menschheit in vielerlei Hinsicht nicht mehr gerecht werden.

Auf der anderen Seite tragen positive Entwicklungen der letzten Jahrzehnte dazu bei, dass sich Epidemiemodellierung mitunter einfacher gestalten kann. Dazu zählen neben einer deutlich umfassenderen Datenlage zum Vergleich von Modellen mit der Realität und der Möglichkeit des computergestützen Rechnens insbesondere Erungenschaften der Netzwerktheorie.

Netzwerke können auf vielen Skalen bei der Epidemiemodellierung zum Einsatz kommen: bei der Begegnung einzelner Menschen oder sozialen Kontakten zwischen Gruppen verschiedener Größe und auf geopolitischer Ebene von einer lokalen Betrachtung bis hin zu einem weltweit agierendem Modell [3, S. 180ff.].

Die Idee, die in diesem Projekt zur Epidemiemodellierung verfolgt wird, ist der Aufbau eines Netzwerks, das die Verbindungen von deutschen Landkreisen in der Umgebung von Göttingen unter Berücksichtigung der zur Arbeit in andere Kreise pendelnden Menschen darstellt. Die Grundlage für die Gestaltung dieses Netzwerks ist der von der Agentur für Arbeit veröffentlichte Pendleratlas [4]. Die zur tatsächlichen Anwendung, Bewertung und Optimierung des Modells wichtigen realen Daten entspringen den umfangreichen vom Robert-Koch-Institut zur Verfügung gestellten Datensätzen im RKI Corona Datenhub [5].

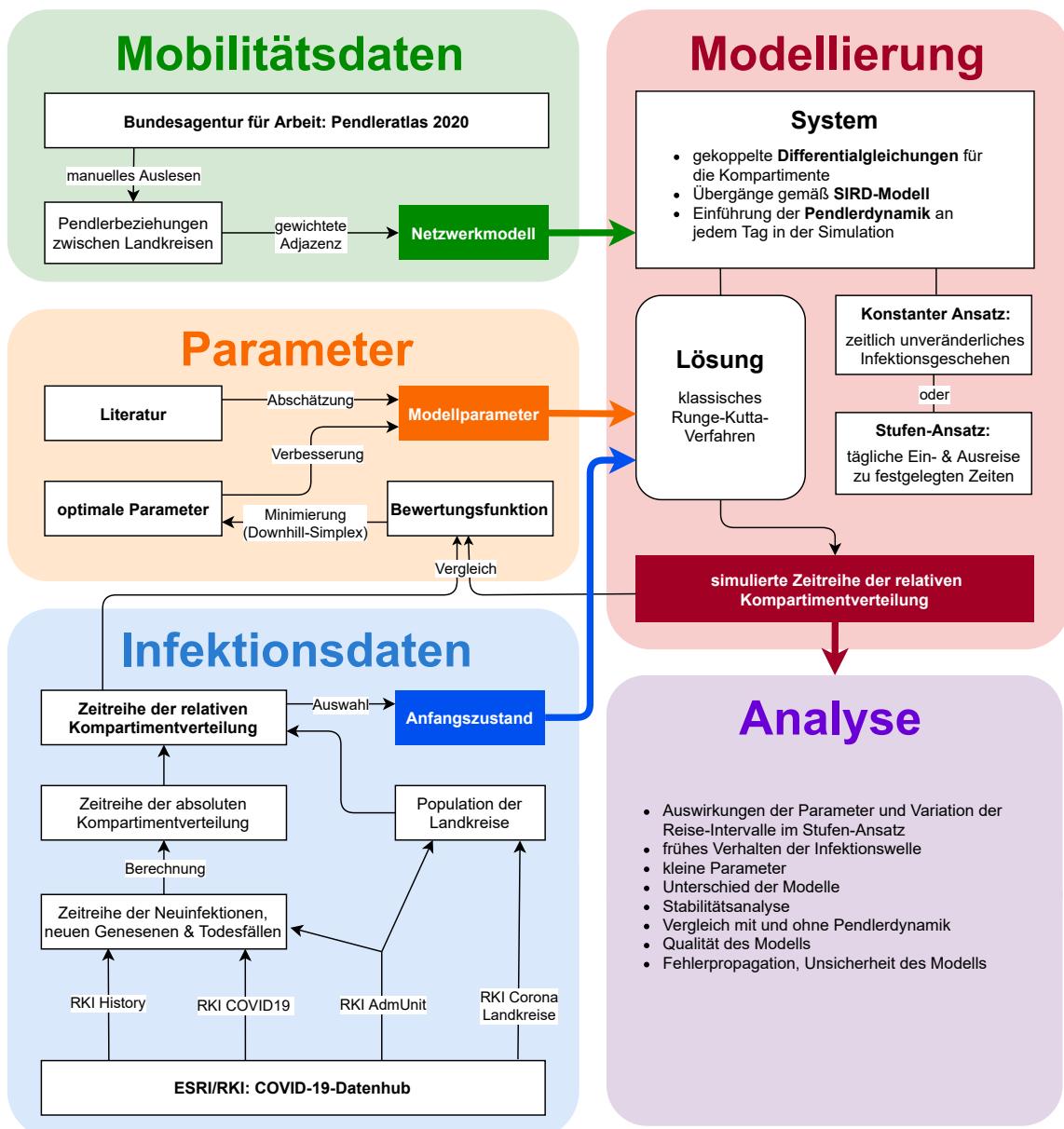
Das Ziel des Projekts ist also die Konstruktion eines Epidemiomodells, das mit einem SIR-Modell mit Pendlerdynamik auf einem Netzwerk aus Landkreisen in der Umgebung von Göttingen arbeitet und nach Optimierung ausgehend von einem gewählten Anfangszustand das tatsächliche Infektionsgeschehen rekonstruieren kann. Dabei liegt der Fokus auf einem Zeitraum von 100 Tagen zwischen Sommer und Herbst des Jahres 2020. Dieser Zeitraum wurde gewählt, da dort die Maßnahmen zur Eindämmung der Pandemie weitgehend unverändert belassen wurden.

Das Projekt charakterisiert sich, ebenso wie das Modell das erarbeitet wird, durch Vernetzung. Die Arbeitsprozesse reichen vom Zusammenstellen und Aufbereiten der notwendigen Daten zum Erstellen eines Modells, mit dem schließlich Simulationen durchgeführt werden können, bis hin zur umfangreichen Untersuchung und Optimierung der Modellierung. Der Ablauf des Projekts ist schematisch in Abbildung 1 dargestellt.

Alle Berechnungen finden computergestützt statt. Programme, Daten und Plots, sowohl die in diesem Dokument als auch weitere, sind in einem Repository auf GitHub<sup>1</sup> einsehbar. Zur Programmierung wurde überwiegend *Python* verwendet, vereinzelt auch C.

---

<sup>1</sup><https://github.com/larstors/Projektpraktikum>



**Abbildung 1:** Ablaufplan der Projektarbeit. Ausgehend von Daten- und Literaturquellen werden Mobilitäts- und Infektionsdaten importiert und Parameter abgeschätzt. Die Mobilitätsdaten bilden die Grundlage für das im Projekt entwickelte Netzwerkmodell mit Pendlerdynamik. Aus den Infektionsdaten wird die Zeitreihe der relativen Kompartimentverteilung berechnet. Einerseits dient diese zum Entnehmen eines Anfangszustands, andererseits zur Optimierung der Modellparameter, die auf dem Vergleich mit der simulierten Zeitreihe der relativen Kompartimentverteilung beruht. Im Netzwerkmodell mit Pendlerdynamik sind zwei Ansätze wählbar. Die Lösung des Systems gekoppelter Differentialgleichungen vollzieht sich durch das klassische Runge-Kutta-Verfahren (RK4). Im weiteren Verlauf werden Eigenschaften und Modifikationen des Modells analysiert.

## 2 Theorie und Methoden

### 2.1 Kompartmentmodell

#### 2.1.1 SIR-Modell

Das SIR-Modell beschreibt den Verlauf der Verbreitung einer Krankheit. Hierfür wird die gesamte Bevölkerung  $N$  in 3 Kompartimente eingeteilt; Anfällige  $S$ , Infizierte  $I$  und Entfernte (d.h. Genesene und Verstorbene)  $R$ . Diese Kompartimente beschreiben die Anteile der Bevölkerung, z.B. die Anzahl der Infizierten ist  $NI$ . Es folgt aus der Aufteilung, dass die Summe der Kompartimente 1 ist. Die zeitliche Änderung der Anfälligen ist sowohl abhängig davon wie viele Anfällige es gibt als auch wie viele Infizierte es gibt. Damit ist [6, S. 727]

$$\frac{dS}{dt} = -\alpha SI, \quad (1)$$

mit  $\alpha$  als konstante Übertragungsrate. Eine anfällige Person kann nur in die Kompartimente der Infizierten übergehen, also ist  $\frac{dI}{dt} \propto \alpha SI$ . Weiter hat die Rate, in der Infizierte genesen oder sterben, eine Genesungsrate  $\beta$ . Daraus erhält man [6, S. 727]:

$$\frac{dI}{dt} = \alpha SI - \beta I. \quad (2)$$

Damit ergibt sich auch die Änderung der Anzahl der Entfernten:

$$\frac{dR}{dt} = \beta I. \quad (3)$$

Zu sehen ist, dass (1), (2) und (3) ein System aus gekoppelten Differentialgleichungen bilden.

### 2.2 Netzwerkmodell

Um die räumliche Verteilung und insbesondere die Vernetzung der Bevölkerung zu erfassen, darzustellen und zwecks Vergleich mit räumlich diskreten, nämlich auf Landkreise<sup>2</sup> bezogenen Daten, zu gruppieren, bieten sich Methoden der Netzwerk- beziehungsweise Graphentheorie an. Zur Abstraktion der Landkreise wird also ein ungerichtetes Netzwerk mit Landkreisen als Knoten verwendet. Die Kanten repräsentieren zunächst einfach die gemeinsamen Landkreisgrenzen, das Netzwerk ist daher im ersten Schritt planar. Die Einführung eines erweiterten Adjazenzbegriffs (Nachbarschaft) eröffnet dann die Möglichkeit der fortschreitenden Vernetzung.

#### 2.2.1 Adjazenzmatrix

Sei  $G = (V, E)$  ein Graph bestehend aus Knoten  $V$  und Verbindungslienien  $E$  zwischen Knoten, so kann die Adjazenzmatrix  $\mathcal{A}$  des Graphen  $G$  geschrieben werden als:

$$[\mathcal{A}_{ij}] = \begin{cases} 1, & (i, j) \in E \\ 0, & \text{sonst.} \end{cases} \quad (4)$$

In diesem Fall bedeutet  $(i, j) \in E$  dass es eine Verbindungslienie zwischen Knoten  $i$  und  $j$  gibt.

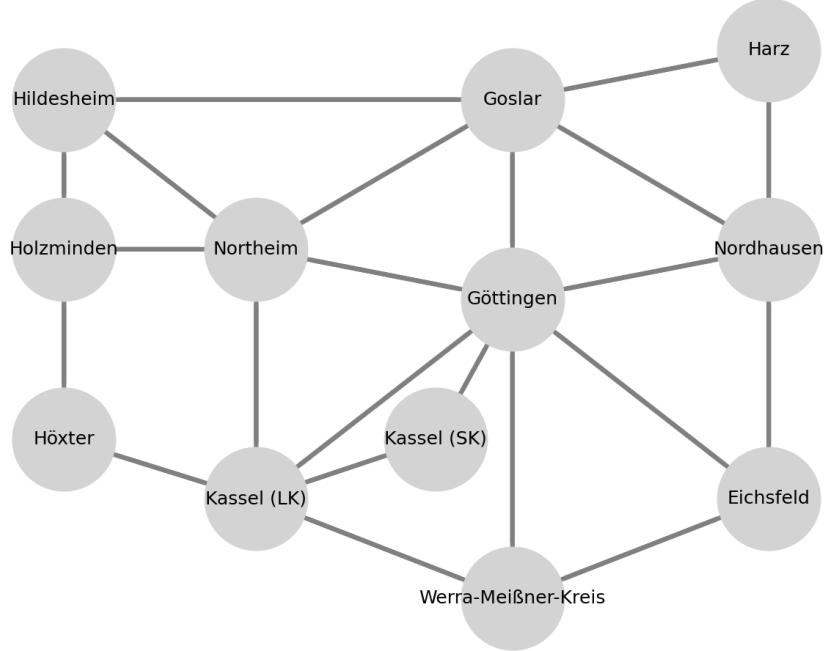
---

<sup>2</sup>Im folgenden werden alle Landkreise, kreisfreie Städte, Regionalverbände und ähnliche administrative Einheiten im Sinne des Kommunalrechts als Landkreise oder Kreise bezeichnet. Demnach gibt es in Deutschland 401 solcher Kreise.

## 2.2.2 Geographische Adjazenzmatrix

Wird nun ein Graph  $G = (V, E)$  betrachtet, in dem jeder Knoten einen Landkreis bezeichnet, so kann die Adjazenzmatrix von  $G$  verwendet werden, um die Geographie der Knoten zu beschreiben. Zunächst wird beispielhaft eine ausgewählte Region von 12 Landkreisen in der Umgebung von Göttingen betrachtet. Nun werden die Landkreise nummeriert und für jeden Knoten die benachbarten Knoten ermittelt, um die Einträge der Adjazenzmatrix  $\mathcal{A}^G$  gemäß Gleichung (4) zu bestimmen. Knoten können mit sich selbst nicht verbunden sein, sodass die Diagonaleinträge von  $\mathcal{A}^G$  null sind.

Unter Verwendung der Python-Bibliothek NetworkX lässt sich die Adjazenzmatrix als Netzwerk darstellen. Dieser Graph ist in Abbildung 2 gezeigt.



**Abbildung 2:** Netzwerkmodell der im ersten Arbeitsschritt betrachteten Region von insgesamt 12 Landkreisen in der Umgebung von Göttingen. Adjazenzen wurden manuell durch Betrachtung einer politischen Landkarte erfasst.

## 2.2.3 Gewichtete Mobilitäts-Adjazenzmatrix (Pendlermatrix)

Möchte man weitere Kanten hinzufügen, um die Vernetzung zu erhöhen, so können aus dem Pendleratlas Informationen über Ein- und Auspendler erhalten werden. Diese stellen eine andere Form der Adjazenz dar. Für jeden Landkreis werden dafür die zehn Landkreise mit den meisten Einpendlern betrachtet. Unter Umständen befinden sich solche Landkreise außerhalb der untersuchten Region, entsprechend verbundene Landkreise weisen in diesem begrenzten Modell demnach geringere Vernetzungen auf. Ob dies einen Einfluss auf die Qualität des Modells hat, wird in Abschnitt 4.3 untersucht. Die aufgeführten Pendler machen etwa 10 % der Bevölkerung eines Landkreises aus und sind daher mutmaßlich repräsentativ für gesellschaftliche Verbindungen zwischen Landkreisen. Die betrachtete Region wird als abgeschlossen behandelt. Im Pendleratlas aufgeführte Landkreise, welche Pendelbeziehungen zu Landkreisen in der betrachteten Region aufweisen ohne selbst Teil der Region zu sein, werden infolgedessen nicht berücksichtigt. Die Erhebung der Pendlerdaten ist der wesentliche limitierende Faktor der Projektarbeit. Es existiert vorrangig eine interaktive Aufbereitung der dargestellten Daten, die nicht maschinell eingelesen werden kann. Die zugrunde liegenden Datensätze sind nach Ländern getrennt und nicht zu unserer Zufriedenheit hinreichend angeordnet.<sup>3</sup> Daraus entsteht die Notwendigkeit des manuellen Schreibens der Pendler-

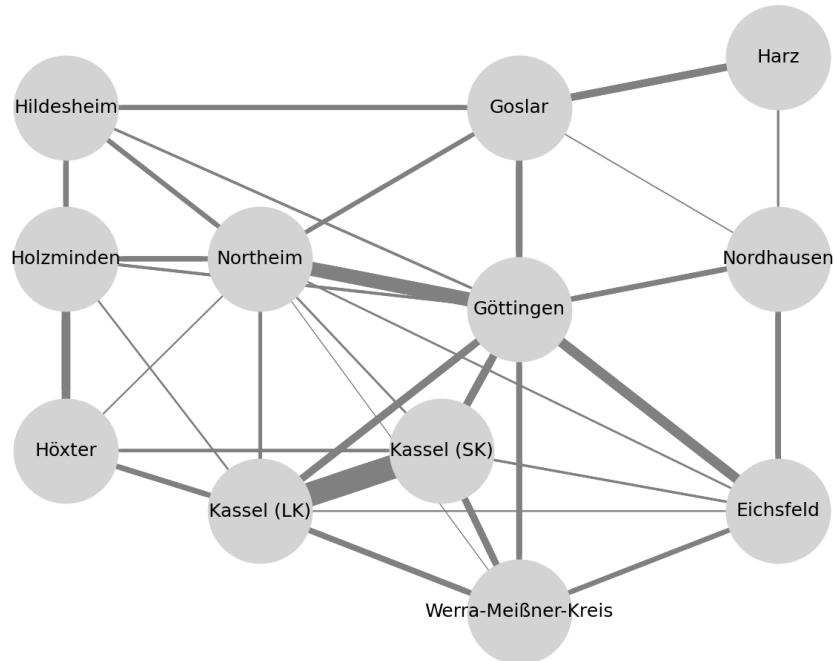
<sup>3</sup>Zweifelsohne wäre das maschinelle Auslesen dieser Datensätze möglich gewesen, allerdings durch das gewählte Format (mehrere Excel-Tabellen mit verschiedenen Blättern) nur aufwendiger als das bereits umfangreiche Aufbereiten der Daten aus dem Corona-Datenhub

matrix. Mit einer Auswahl von 38 Landkreisen<sup>4</sup> wurde ein Kompromiss gefunden, der das manuelle Schreiben der Matrix noch ermöglicht.

Im Eintrag  $P_{ij}$  sind die Anzahl der Pendler von  $j$  nach  $i$  eingetragen. Gegenüber der vorher eingeführten Methode zeichnet sich das in dieser Methode verwendete Netzwerk durch

- Richtung
- Gewichtung
- höhere Vernetzung

aus. Die Matrix ist also nicht symmetrisch und kann von 0 und 1 verschiedene Einträge haben, je nach Ausprägung der Pendelbewegungen. Durch die Auflösung der Bindung an geographische Verhältnisse verliert das Netzwerk allerdings seine Planarität<sup>5</sup>, das heißt: Kanten überschneiden sich unter Umständen. Das Netzwerk ist in Abbildung 3 gezeigt.



**Abbildung 3:** Netzwerkmodell der Region unter Berücksichtigung nicht-geographischer Adjazzenzen in Form von signifikanten Pendlerbewegungen. Im Sinne einer anschaulichen Darstellung ist die Breite der Kanten proportional zur Wurzel des Betrags des Pendlerstroms, also der Summe aus Aus- und Einpendlern, gewählt. Richtungen von Verbindungen werden hier und in folgenden Abbildungen von Netzwerken nicht dargestellt, um insbesondere bei größeren Netzwerken Übersichtlichkeit zu wahren.

## 2.2.4 Maß der Vernetzung

Es existieren verschiedene Größen, die die Ausprägung der Vernetzung eines Netzwerks beschreiben [7]:

- Der Grad eines Knotens gibt an, mit wie vielen anderen Knoten Verbindungen bestehen.
- Die Dichte  $\rho$  gibt an, wie viele der möglichen Verbindungen zwischen Knoten im Netzwerk tatsächlich vorhanden sind. Sie kann also Werte zwischen 0 und 1 annehmen.
- Der globale Clusterkoeffizient  $C$  gibt für die Gesamtheit aller Knoten im Netzwerk an, wie viele Kanten zwischen seinen Nachbarn existieren, gemessen an der Anzahl der möglichen Kanten. Er ist also ein Maß

<sup>4</sup>Dies entspricht 1444 Zellen, davon 266 ungleich null

<sup>5</sup>Bei Netzwerken auf zweidimensionalen Karten ist immer Planarität gegeben, siehe dazu [7], S.123ff.

dafür, wie ausgeprägt die Verbindungen der Nachbarn von Knoten sind und wie nah das Netzwerk einer vollständigen Vernetzung, einer sogenannten Clique, kommt. Er kann ebenfalls Werte zwischen 0 und 1 annehmen.

All diese Größen können mit Methoden der Python-Bibliothek NetworkX, die Netzwerke behandelt, berechnet werden [8]. Eine Einordnung der Verbesserung der Vernetzung der Landkreise in der Region durch Einführung der Pendler-Matrix  $\mathcal{P}$ , die Mobilität berücksichtigt, ist in Tabelle 1 eingetragen.

**Tabelle 1:** Verbesserung der Dichte und des globalen Cluster-Koeffizienten des Netzwerks von 12 Landkreisen durch Einbeziehung nicht-geographischer Adjazzenzen. Gewichtungen sind in dieser Betrachtung nicht berücksichtigt.

Matrix	Netzwerkcharakteristik	Adjazzenzen/Kanten	planar	$\rho$	C
$\mathcal{A}^G$	geographisch	Landkreisgrenzen	ja	0,33	0,52
$\mathcal{P}$	sozio-ökonomisch	signifikante Pendlerbewegungen	nein	0,50	0,72
		<b>Verbesserung</b>	+52 %	+31 %	

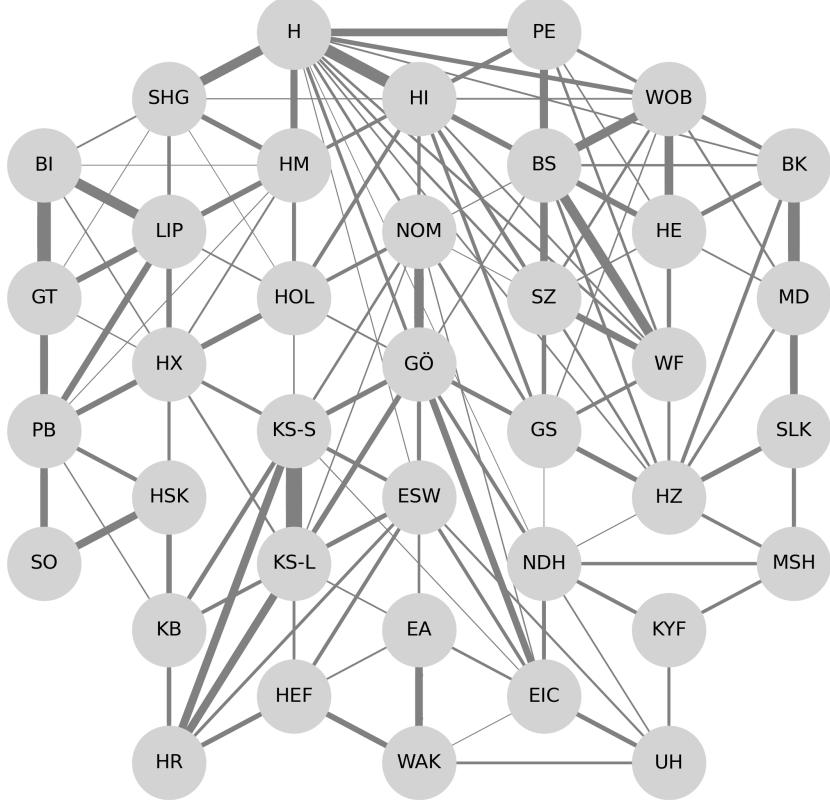
## 2.3 Charakterisierung der betrachteten Region

Während in der Entwicklungsphase die aus lediglich 12 Landkreisen bestehende Region untersucht wird, widmet sich der Hauptteil des Projekts der Simulation in einer größeren Region, die weiterhin den Landkreis Göttingen im Zentrum hat. Damit befindet sie sich auch in der Mitte Deutschlands, sodass Staatsgrenzen und damit andere politische und gesellschaftliche Gegebenheiten eher weit entfernt sind.

Die ausgewählte Region, die in Abbildung 4 auf einer Karte markiert ist, beinhaltet 14 Landkreise in Niedersachsen, 7 in Nordrhein-Westfalen, je 6 in Hessen und Thüringen sowie 5 in Sachsen-Anhalt. Die Populationsgröße beträgt in der Summe 7,6 Millionen und macht damit etwa 9,2 % der Bevölkerung Deutschlands aus. Im Mittel leben in einem Knoten (Landkreis) demnach etwa 200.000 Menschen. Der einwohnerreichste Knoten ist die Region Hannover mit mehr als 1,1 Millionen Einwohnern, der kleinste Knoten ist die Kreisfreie Stadt Eisenach mit etwa 42.000 Einwohnern.



**Abbildung 4:** Lage der zur Untersuchung ausgewählten Region bestehend aus 38 Landkreisen in fünf deutschen Bundesländern



**Abbildung 5:** Netzwerkdarstellung der im Hauptteil des Projekts untersuchten Region. Die Knoten sind die einzelnen Landkreise, markiert mit den jeweiligen Kfz-Kennzeichen entsprechenden Abkürzungen, die Tabelle 4 im Anhang entnommen werden können. Die Kanten stellen Pendlerbeziehungen dar. In dieser Abbildung ist die Breite der Kanten proportional zur Wurzel der Anzahl der zwischen den Landkreisen zur Arbeit pendelnden Personen. Keine Verbindung zwischen zwei Landkreisen bedeutet, dass die Kreise nicht zu den zehn Kreisen mit den meisten Einpendlern in die jeweils anderen gehören. Pendlerbewegungen von und nach Landkreisen außerhalb des Netzwerks sind nicht berücksichtigt.

## 2.4 Erweiterung des Kompartimentmodells

### 2.4.1 Der Tod

Um der Realität etwas näher zu kommen, kann ein weiteres Kompartiment eingeführt werden: die Verstorbenen  $D$ . Infolgedessen bezeichnet das Kompartiment  $R$  im Gegensatz zum SIR-Modell nur noch die Genesenen. Zu beachten ist, dass die Gesamtbevölkerung  $N$  sich nicht ändert, da die Summe über alle Kompartimente konstant bleibt. In diesem Modell werden nur Sterbefälle betrachtet, die durch die Krankheit verursacht werden. Weiter sei  $p$  die Wahrscheinlichkeit, dass eine infizierte Person stirbt statt zu genesen. Damit bleiben  $\frac{dS_i}{dt}$  und  $\frac{dI_i}{dt}$  unverändert, allerdings werden:

$$\begin{aligned} \frac{dR_i}{dt} &= (1-p)\beta I_i, \\ \frac{dD_i}{dt} &= p\beta I_i. \end{aligned} \tag{5}$$

In Deutschland kann die Wahrscheinlichkeit als  $p = 2,64\%$  angenommen werden. [9]

### 2.4.2 Dynamik durch Pendler

Da  $S_i$  und  $I_i$  jeweils die Anteile der Bevölkerung in dem Knoten  $i$  angeben, ist  $N_i S_i$  die Anzahl der Anfälligen des Knotens  $i$ . Analog ist  $N_i I_i$  die Anzahl der Infizierten in Knoten  $i$ . Die Dynamik während der Pendelphase wird wie in Abbildung 6 angenommen. Eine wichtige Annahme ist, dass die gesamte Bevölkerung der Knoten

konstant bleibt. Gleichzeitig sollen die Pendler auch eine Auswirkung auf die anderen Knoten haben. Analog zu den Gleichungen (1), (2) und (3) gilt für den Knoten  $i$  [10] somit:

$$\frac{dS_i}{dt} = -\alpha S_i I_i^{\text{eff}}, \quad (6a)$$

$$\frac{dI_i}{dt} = \alpha S_i I_i^{\text{eff}} - \beta I_i, \quad (6b)$$

$$\frac{dR_i}{dt} = \beta I_i. \quad (6c)$$

Hier ist  $I_i^{\text{eff}}$  der Anteil der Infizierten, die in Knoten  $i$  zu Ansteckungen beitragen, den es nun zu bestimmen gilt. Hierfür sind zwei Bewegungen zu betrachten: die Abwesenheit der Infizierten aus Knoten  $i$  und die Anwesenheit der Infizierten aus Knoten  $j$ . Es ist zu sehen, dass die Anzahl der Infizierten aus Knoten  $i$ , die nicht in andere Knoten pendeln, gegeben ist durch

$$N_i I_i - \sum_j \mathcal{P}_{[i \rightarrow j]} I_i = \left( N_i - \sum_j \mathcal{P}_{[i \rightarrow j]} \right) I_i, \quad (7)$$

wobei  $\mathcal{P}_{[i \rightarrow j]}$  hier die Anzahl der Pendler bezeichnet, die von  $i$  nach  $j$  pendeln<sup>6</sup>. Analog ist die Anzahl der Infizierten, die aus Knoten  $j$  nach  $i$  pendeln, durch

$$\sum_j \mathcal{P}_{[j \rightarrow i]} I_j \quad (8)$$

gegeben. Die Summe der Ausdrücke (7) und (8) ergibt die Anzahl der infizierten Kontaktpersonen während der Pendelphase in Knoten  $i$ . Den entsprechenden Bevölkerungsanteil  $I_i^{\text{pen}}$  erhält man nun, indem man jene Zahl auf die tatsächliche Populationsgröße während der Pendelphase bezieht, das heißt

$$I_i^{\text{pen}} := \frac{N_i^{\text{rest}} I_i + \sum_j \mathcal{P}_{[j \rightarrow i]} I_j}{N_i^{\text{rest}} + \sum_j \mathcal{P}_{[j \rightarrow i]}}, \quad (9)$$

mit der Abkürzung  $N_i^{\text{rest}} = N_i - \sum_j \mathcal{P}_{[i \rightarrow j]}$  für die Nichtpendler aus Ausdruck (7).

Nach der Pendelphase  $dt$  hat sich ein Teil der Nichtpendler  $N_i^{\text{rest}}$  im Heimatknoten  $i$  angesteckt und ein Teil der Pendler in den anderen Knoten  $j \neq i$ , sodass sich insgesamt der Anteil der Suszeptiblen gemäß

$$\frac{dS_i}{dt} = -\alpha \cdot \frac{1}{N_i} \left( N_i^{\text{rest}} S_i I_i^{\text{pen}} + \sum_j \mathcal{P}_{[i \rightarrow j]} S_i I_j^{\text{pen}} \right) = -\alpha S_i \left( \frac{N_i^{\text{rest}}}{N_i} I_i^{\text{pen}} + \sum_j \frac{\mathcal{P}_{[i \rightarrow j]}}{N_i} I_j^{\text{pen}} \right) = -\alpha S_i I_i^{\text{eff}} \quad (10)$$

geändert hat, wobei beim letzten Gleichheitszeichen Gleichung (6a) verwendet wird. Daher beträgt der effektive Anteil der Infizierten im  $i$ -ten Knoten

$$I_i^{\text{eff}} = \frac{N_i^{\text{rest}}}{N_i} I_i^{\text{pen}} + \sum_j \frac{\mathcal{P}_{[i \rightarrow j]}}{N_i} I_j^{\text{pen}}. \quad (11)$$

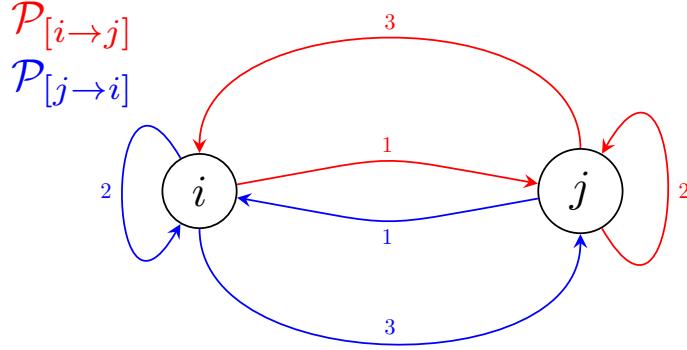
Dies kann nun auch in die Ableitung  $\frac{dI_i}{dt}$  aus Gleichung (6b) eingesetzt werden.

## 2.5 Die Dynamik an einem Tag

Die Differentialgleichungen (6) beschreiben die Dynamik während die Pendler nicht zu Hause sind und die Differentialgleichungen (1), (2) und (3) beschreiben die jeweiligen Knoten während die Pendler zu Hause sind. Es werden nun zwei Modelle vorgestellt, die das dynamische und das statische System zeitlich verbinden.

---

<sup>6</sup>Diese Notation ist anschaulicher als das zuvor verwendete äquivalente Matrixelement  $\mathcal{P}_{ji}$



**Abbildung 6:** Visualisiert ist das Verhalten der Pendler, wie sie im Modell berücksichtigt wird. Bewegung 1 beschreibt das Pendeln vom Heimatknoten in einen anderen Knoten; Bewegung 2 beschreibt das Verbleiben in dem anderen Knoten, welches zu Infektionen beiträgt, und Bewegung 3 ist die Heimreise.

### 2.5.1 Der Stufen-Ansatz

Die charakteristische Funktion des Intervalls  $[a, b]$  ist definiert als:

$$\chi_{[a,b]}(t) = \begin{cases} 1, & t \in [a, b] \\ 0, & \text{sonst.} \end{cases}$$

Ein Tag kann als Zeitintervall  $[0, 1]$  definiert werden. Es werden nun drei Partitionen  $\pi_1$ ,  $\pi_2$  und  $\pi_3$  aus diesem Zeitintervall gebildet, sodass  $\pi_1$  der Zeitraum ist, in dem die Pendler noch nicht ihren Knoten verlassen haben und  $\pi_2$  der Teil des Tages ist, in dem die Pendler in anderen Knoten sind. Der Rest des Tages, an dem die Pendler in den Ausgangsknoten zurückgekehrt sind, ist  $\pi_3$ . Weiter bezeichne  $\chi_1$ ,  $\chi_2$  und  $\chi_3$  die charakteristischen Funktionen der jeweiligen Partitionen, d.h.  $\chi_i(t) \equiv \chi_{\pi_i}(t)$ . Des Weiteren kann eine charakteristische Funktion für den Zeitraum, in dem die Pendler zu Hause sind, definiert werden als  $\chi_{1+3} = \chi_1 + \chi_3$ . Somit ist das dynamische SIRD-Modell:

$$\frac{dS_i}{dt} = -\chi_{(1+3)}\alpha S_i I_i - \chi_2 \alpha S_i I_i^{\text{eff}}, \quad (12a)$$

$$\frac{dI_i}{dt} = \chi_{(1+3)}\alpha S_i I_i + \chi_2 \alpha S_i I_i^{\text{eff}} - \beta I_i, \quad (12b)$$

$$\frac{dR_i}{dt} = (1-p)\beta I_i, \quad (12c)$$

$$\frac{dD_i}{dt} = p\beta I_i. \quad (12d)$$

Für einen Zeitraum, der mehrere Tage umfasst, kann die Aufteilung in Partitionen periodisch verlaufen, d.h. der Zeitraum wird erst in einzelne Tage unterteilt und aus diesen  $\pi_1$ ,  $\pi_2$  und  $\pi_3$  gebildet.

### 2.5.2 Der konstante Ansatz

Statt einen Tag zu teilen kann ein Koeffizient  $t_0$  eingeführt werden, welcher das Infektionsgeschehen während der Pendelphase gewichtet. Der Koeffizient  $t_0$  ist gegeben durch:

$$t_0 = \frac{\text{Zeit, in der Pendler in anderen Knoten sind}}{\text{gesamte Zeit}}.$$

Es ist einfach zu sehen, dass  $0 \leq t_0 \leq 1$ . Weiter beschreibt der Koeffizient  $1 - t_0$  den Anteil des Tages, in dem die Pendler in ihren Heimatknoten sind. Analog zu  $t_0$  ist  $1 - t_0$  auch zwischen 0 und 1. Diese Koeffizienten werden nun als Gewichtung der Infektionsanteile in einem Zeitraum verwendet. Daraus folgt, dass das angepasste SIRD-

Modell die Form

$$\frac{dS_i}{dt} = -(1-t_0)\alpha S_i I_i - t_0 \alpha S_i I_i^{\text{eff}}, \quad (13a)$$

$$\frac{dI_i}{dt} = (1-t_0)\alpha S_i I_i + t_0 \alpha S_i I_i^{\text{eff}} - \beta I_i, \quad (13b)$$

$$\frac{dR_i}{dt} = (1-p)\beta I_i, \quad (13c)$$

$$\frac{dD_i}{dt} = p\beta I_i \quad (13d)$$

erhält. Hier wird angenommen, dass sich das Infektionsverhalten über den Tag nicht stark ändert, daher die Bezeichnung „konstanter Ansatz“.

Aus dem direkten Vergleich der Systeme (12) und (13) sind nur die Koeffizienten verschieden. Durch weiteres betrachten der Modelle sind weitere fundamental Unterschiede zu erkennen, z.B. berücksichtigt System (12) das Variieren der Kompartimente über den Tag hinweg, wobei in (13) angenommen wird, dass die Kompartimente sich nicht stark verändern an einem Tag.

## 2.6 Datenerhebung

Die für die Simulation benötigten Anfangswerte für die Verteilung der Kompartimente sowie weitere Daten (z.B. Bevölkerungsanzahl der Landkreise), werden aus den vom Robert-Koch-Institut (RKI) im COVID-19 Datenhub zur Verfügung gestellten Datensätzen gewonnen.

### 2.6.1 Verwendete Datensätze

Dabei kommen die folgenden Datensätze zum Einsatz:

- **RKI AdmUnit:** In den RKI-Datensätzen werden für die Landkreise eindeutige Nummerierungen (sogenannte *AdmUnitIDs*) verwendet. Zum internen Gebrauch im Projekt werden die Landkreise ebenfalls mit einer *internen ID* nummeriert. Es wird also eine bijektive Abbildung zwischen AdmUnitIDs und internen IDs für die betrachtete Region hergestellt.
- **RKI Corona Landkreise:** In diesem Datensatz sind täglich aktuelle Daten enthalten, von denen allerdings kein Gebrauch gemacht wird. Aus RKI Corona Landkreise wird lediglich die Bevölkerungsgröße der Landkreise verwendet, da diese in keinen der anderen verwendeten Datensätzen aufgeführt sind.
- **RKI COVID19:** Dieser umfangreiche Datensatz ist eine Zeitreihe der täglichen Anzahl neuer Infektionen, neuer Sterbefälle sowie neuer Genesenen<sup>7</sup> aufgeschlüsselt nach Landkreis, Geschlecht und Altersgruppe sowie versehen mit der Angabe, ob der entsprechende Tag das Melde- oder Erkrankungsdatum für den Infektionsfall darstellt und ob der Fall daher mehrmals im Datensatz (in verschiedenen Zuordnungen zu Melde- oder Erkrankungsdatum) enthalten ist.
- **RKI History:** Auch dieser Datensatz ist eine Zeitreihe, die die tägliche Anzahl neuer Fälle mit Anmerkung, ob das Datum Melde- oder Erkrankungsdatum ist, sowie den täglichen Stand kumulierter Fälle aufgeschlüsselt nach Landkreis enthält.

### 2.6.2 Zusammenstellung der benötigten Daten

Manche Daten können direkt den im Datenhub zur Verfügung gestellten Datensätzen entnommen werden, während die meisten erst durch einfache Rechnungen, durch vorgegebene oder durch eigene Algorithmen daraus abgeleitet

---

<sup>7</sup>Auf Nachfrage beim Robert-Koch-Institut wurde darauf hingewiesen, dass Daten zu Genesenen auf Schätzungen beruhen und die Genauigkeit auf Kreisebene noch schlechter als auf höheren Ebenen sei. Daher weise man diese Daten nicht aus. Sie sind aber dennoch im Datensatz RKI COVID19 zu finden.

werden. Insbesondere die Verteilung der Bevölkerung auf die Kompartimente S, I, R, D wird vom Robert-Koch-Institut weder erfasst noch direkt geschätzt. Beim Zusammenstellen der Daten werden alle einzelnen Informationen immer den internen IDs der Landkreise zugeordnet. Die folgenden Daten werden für interne Berechnungen, Anfangsbedingungen und Vergleiche benötigt:

- **Populationsgröße:** Diese Daten existieren als fertige Datenreihe in RKI Corona Landkreise.
- **Zeitreihe der Neuinfektionen:** Diese Zeitreihe setzt sich gemäß der Dokumentation des Datensatzes aus der Summe der neu gemeldeten und der neu erkrankten Fälle pro Tag in der Zeitreihe RKI History zusammen.
- **Zeitreihe der Sieben-Tage-Inzidenz:** Der Sieben-Tage-Inzidenzwert pro 100000 Einwohner  $\Omega_{\text{RKI}}(k, t)$  im Landkreis  $k$  wird berechnet als

$$\Omega_{\text{RKI}}(k, t) = \frac{\sum_{d=0}^6 \text{Neuinfektionen in } k \text{ vor } d \text{ Tagen}}{\text{Populationsgröße } N(k)/100000}. \quad (14)$$

- **Zeitreihe der neuen Todesfälle:** Dazu wird RKI COVID19 verwendet. Für jeden Tag und für jeden Landkreis wird die Anzahl neuer Todesfälle aus allen Alters- und Geschlechtsgruppen summiert, falls in der Datenreihe die Indikation „Fall ist in der Publikation für den aktuellen Tag und in der für den Vortag jeweils ein Todesfall“ oder „Fall ist in der aktuellen Publikation ein Todesfall, nicht jedoch in der Publikation des Vortages“ vorliegt.<sup>8</sup>
- **Zeitreihe der neuen Genesenen:** Das Vorgehen ähnelt der Zusammenstellung der Zeitreihe der neuen Todesfälle. Es wird ebenfalls RKI COVID19 verwendet. Für jeden Tag und für jeden Landkreis wird die Anzahl neuer Genesener aus allen Alters- und Geschlechtsgruppen summiert, falls in der Datenreihe die Indikation „Fall ist in der Publikation für den aktuellen Tag und in der für den Vortag jeweils Genesen“ oder „Fall ist in der aktuellen Publikation Genesen, nicht jedoch in der Publikation des Vortages“ vorliegt.<sup>9</sup> Stichprobenartige Vergleiche mit mehreren Mitteilungen von Behörden von Landkreisen zeigen jedoch Abweichungen von den so ermittelten Genesenzahlen, die unter einer Verschiebung von 14 Tagen in die Zukunft weitgehend verschwinden.<sup>10</sup>
- **Zeitreihen der kumulierten Infektionen/ Todesfälle/ Genesenen:** Diese Zeitreihen können einfach durch Summation über die entsprechenden Fälle bis zum betreffenden Zeitpunkt erhalten werden. Bemerkungen: Die Anzahl der kumulierten Todesfälle entspricht trivialerweise der Anzahl der aktuell im Kompartiment der Verstorbenen befindlichen Personen, da Verstorbene nicht in ein weiteres Kompartiment übergehen. Analog entspricht die Anzahl der kumulierten Genesenen der Anzahl der aktuell im Kompartiment der Genesenen befindlichen Personen. Dieser Umstand ist allerdings auf das gewählte Modell zurückzuführen, in dem aus dem Kompartiment der Genesenen kein Übergang in andere Kompartimente möglich ist. Dementsprechend sind für die Verstorbenen und für die Genesenen die Zeitreihen der entsprechenden Fallzahlen identisch mit den Zeitreihen der Anzahl der Individuen in den entsprechenden Kompartimenten.
- **Zeitreihe der Anzahl an aktuell Infizierten:** Diese Zeitreihe lässt sich zu jedem Zeitpunkt berechnen aus der Differenz zwischen den kumulierten Infektionen und der Summe aus Entfernten, das heißt Genesenen und Verstorbenen.

---

<sup>8</sup>Dieses Vorgehen ist nicht intuitiv, folgt aber der Dokumentation des Datensatzes und wurde auf Nachfrage von technischen Experten des Environmental System Research Institute (ESRI) Corona Dashboard Teams Germany and Switzerland bestätigt.

<sup>9</sup>Auch dieses Vorgehen ist verifiziert.

<sup>10</sup>Die Ursache für diese notwendige Anpassung ist nicht vollständig geklärt. Allerdings geht das Robert-Koch-Institut bei der Berechnung der Anzahl der Genesenen davon aus, dass Infizierte, die nicht hospitalisiert werden, nach 14 Tagen genesen. Mit einer entsprechenden Verschiebung stimmen die Genesenzahlen mit auf Kreisebenen veröffentlichten Zahlen besser überein. Die Genesenzahlen ausschließlich aus den Angaben von Kreisverwaltungen etc. zu beziehen, kommt bei der Anzahl der betrachteten Landkreise nicht mehr in Betracht. Eine eigene Berechnung der Genesenzahlen ist nicht möglich, da der vom Robert-Koch-Institut angegebene Algorithmus Daten zur Hospitalisierungsquote voraussetzt, die im COVID-19 Datenhub nicht zur Verfügung stehen.

- **Zeitreihe der Anzahl an aktuell Suszeptiblen:** Diese Zeitreihe lässt sich zu jedem Zeitpunkt berechnen aus der Differenz zwischen der Populationsgröße und der Summe aus der Anzahl der Individuen in allen anderen Kompartimenten.
- **Zeitreihe der relativen Kompartimentverteilung:** Diese Zeitreihe gibt zum jeden Zeitpunkt und für jeden Landkreis an, welcher Anteil der Bevölkerung sich in welchem der vier betrachteten Kompartimente befindet. Sie wird einfach berechnet aus dem Quotienten der oben berechneten Anzahl der Personen in einem Kompartiment und der Populationsgröße. Sie erfüllt also  $S_i(t), I_i(t), R_i(t), D_i(t) \in [0, 1]$   
 $\wedge S_i(t) + I_i(t) + R_i(t) + D_i(t) = 1 \forall i \in K$  (Menge der betrachteten Landkreise), Zeiten  $t$  und soll direkte Vergleichbarkeit mit den Ergebnissen der Simulation gewährleisten.

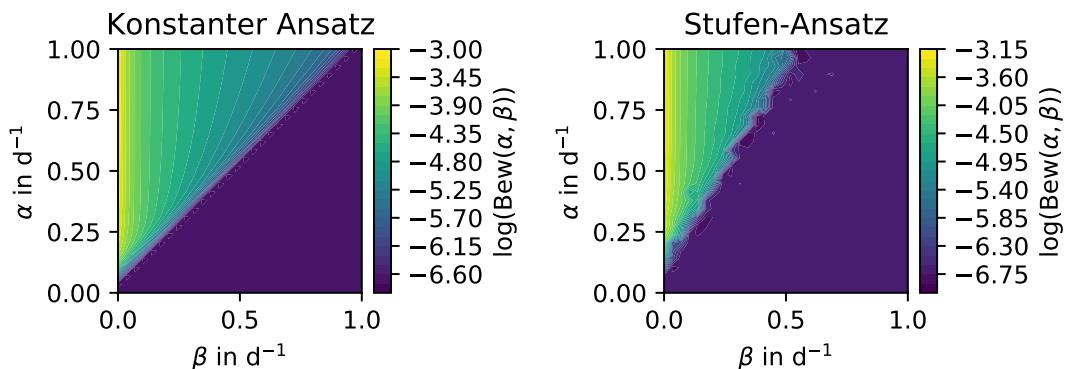
Aufgrund der im Datensatz der Genesenen vorgenommenen Zeitverschiebung sind alle davon abgeleiteten Daten, insbesondere die Kompartimentverteilung für die letzten zwei Wochen des Zeitraums, für den die Daten importiert wurden, nicht brauchbar. Da aber alle historischen Daten im Mai 2021 importiert wurden und die im Projekt betrachteten Zeiträume nicht über November 2020 hinaus reichen, ist diese Problematik bedeutungslos. Die Zeitreihe der relativen Kompartimentverteilung ist am relevantesten für das Projekt, da sie an einem gewählten Zeitpunkt die Anfangswerte liefert und weil ihr Verlauf als Vergleichswert verwendet wird. Wenn im folgenden auf die RKI-Daten verwiesen wird, ist die von den vom Robert-Koch-Institut zur Verfügung gestellten Daten abgeleitete Zeitreihe der relativen Kompartimentverteilung gemeint.

## 2.7 Optimierung der Parameter

Unter der Annahme, dass es Modellparameter  $\boldsymbol{\theta}_{\min}$  gibt, unter Verwendung derer das Modell den Verlauf der Pandemie im betrachteten Zeitraum zufriedenstellend abbildet, definieren wir in Gleichung (15) eine Bewertungsfunktion  $\text{Bew}(\boldsymbol{\theta})$ , welche die Abweichung zwischen Modellergebnissen und Messdaten des tatsächlichen Pandemieverlaufs widerspiegeln soll. Eine passende Auswahl hierfür ist die mit der Bevölkerungsanzahl gewichtete mittlere Abweichung der Infizierten zwischen simuliertem und tatsächlichem Modell:

$$\text{Bew}(\boldsymbol{\theta}) = \frac{1}{|K||D|} \sum_{k \in K} \sum_{d \in D} |I_{\text{RKI}}(d, k) - I_{\text{sim.}}(d, k, \boldsymbol{\theta})| N_k. \quad (15)$$

Dabei ist  $D$  die Menge der simulierten Tage und  $K$  die Menge der betrachteten Landkreise. Der Faktor  $\frac{1}{|K||D|}$  dient zur Normierung der Funktion, sodass deren Werte für verschiedene Regionen direkt miteinander verglichen werden können. Die Funktion  $\text{Bew}$  ist in Abhängigkeit der Parameter  $\alpha$  und  $\beta$  in Abbildung 7 dargestellt.



**Abbildung 7:** Logarithmische Darstellung des Verlaufs der Bewertungsfunktion (Gleichung 15) in Abhängigkeit der Parameter  $\alpha$  und  $\beta$  für das Modell des konstanten Ansatzes (links) und des Stufen-Ansatzes (rechts). Berechnet wurde dies für 12 Landkreise über den Zeitbereich in 2.7.1.

Beachte, dass in diesem Fall die Messdaten  $\langle I_{\text{RKI}}(d) \rangle$  einen mithilfe der Einwohnerzahlen gewichteten Mittelwert über die miteinbezogenen Landkreise darstellen.

Mit der Anforderung an  $\theta_{\min}$ , dass die Funktion Bew bei dieser Stelle ein globales Minimum hat, wird mit Hilfe eines Optimierungsverfahrens der Wert von  $\theta_{\min}$  ermittelt. Dies geschieht mit Hilfe des Downhill-Simplex-Algorithmus, welches in Abhängigkeit eines Startwertes ein Minimum der Bewertungsfunktion findet [11].

So werden für viele Startwerte aus der Menge  $\epsilon = \{(\alpha, \beta, p) \mid \alpha \in [0; 0,1], \beta \in [0; 0,2], p = 0,0264\}$  einige Parameter optimiert, um die Menge der optimalen Parameter (Opt) zu bilden. Der Wert des globalen Minimums  $\theta_{\min}$  schätzen wir durch den Vektor  $\theta \in \text{Opt}$ , bei dem das tiefste Minimum vorliegt:

$$\theta_{\min} := \min (\{\text{Bew}(\theta) \mid \theta \in \text{Opt}\}) \quad (16)$$

Aus der Standardabweichung der Komponenten von  $\theta \in \text{Opt}$  wird der Fehler der Parameter entnommen:

$$\sigma_{\theta_i} = \sqrt{\frac{\sum_{\theta^s \in \text{Opt}} (\theta_i^s - \langle \theta_i \rangle)^2}{(n-1)}} \quad \text{mit} \quad \langle \theta_i \rangle = \sum_{\theta^s \in \text{Opt}} \frac{\theta_i^s}{n} \quad (17)$$

Dabei ist  $n$  die Anzahl der Elemente in der Menge Opt,  $\theta_i^s$  die i-te Komponente des Parametervektors  $\theta^s$ , wobei dieser der s-te Element in Opt ist.

Anzumerken ist, das der Parameter  $p$ , welcher die Wahrscheinlichkeit darstellt, dass ein Infizierter stirbt, auf die Bewertungsfunktion keinen Einfluss hat. Die Entscheidung, jeweils das Kompartiment der Verstorbenen in der Bewertungsfunktion nicht zu berücksichtigen, stammt daher, dass sich unsere Modelle für den Tod nicht grundsätzlich von der Literatur unterscheiden. Die Pendlerdynamik beeinflusst nur  $I$ . Daher wäre eine Optimierung dieses Parameters eine Wiederholung der Arbeit in [9]. Sein Wert wird daher aus dieser Quelle entnommen.

### 2.7.1 Auswahl des zu simulierenden Zeitraums

Für die Bestimmung der Parameter wird vorausgesetzt, dass sie zeitlich und räumlich konstant sind. Hierfür ist es wichtig, sicherzustellen, dass die Vergleichsdaten des simulierten Zeitraums möglichst keine Variation der Parameter aufweisen. Dies kann aber z.B. der Fall sein, wenn verschiedene Maßnahmen implementiert werden, oder, wenn mehrere Varianten des Virus vorhanden sind. Um möglichst konstante Bedingungen zu gewährleisten, wird ein Zeitraum betrachtet, bei dem gilt:

- Kein Lockdown wird währenddessen durchgesetzt.
- Das Testangebot und die Möglichkeiten, sich zu testen, ändern sich nicht drastisch.
- So weit wie möglich ist nur eine einzige Variante (der sogenannte Wildtyp) prävalent.
- Die Impfkampagne hat noch nicht begonnen.

Ein passender Zeitbereich für diese Voraussetzungen ist das Intervall von 100 Tagen beginnend mit dem 24. Juli 2020. Das Intervall endet am 01. November 2020, genau einen Tag vor dem Beginn des sogenannten „Lockdown Light“, dem ersten wesentlichen Verstärken von Maßnahmen im Herbst des Jahres 2020. [12]

## 2.8 Fehlerpropagation

Die numerische Lösung des Systems lässt sich auf die Form:

$$\mathbf{x}_{j+1} = \mathbf{x}_j + h_j \mathbf{F}(\theta, \mathbf{x}_j) \quad (18)$$

vereinfachen, wobei  $\mathbf{F}(\theta, \mathbf{x}_j)$  modellspezifisch,  $h_j$  Zeitschrittgröße und  $\mathbf{x}_j$  der Zustand des Systems zum Zeitpunkt  $t_j$  ist:

$$\mathbf{x}_j = [S_1, \dots, S_k, I_1, \dots, I_k, R_1, \dots, R_k, D_1, \dots, D_k]^T \quad (t = t_j). \quad (19)$$

Der Anfangswert  $\mathbf{x}_0$  unterliegt einer Unsicherheit und kann nicht exakt angegeben werden. Dies gilt ebenfalls für die Parameter, weshalb die Parameter durch die Zufallsvariable:

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}, \quad (20)$$

mit konstantem Wert  $\boldsymbol{\theta}_0$  und einer normalverteilten Abweichung  $\Delta\boldsymbol{\theta} \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ , angegeben werden. Der Wert von  $\boldsymbol{\theta}_0$  sowie die Kovarianzmatrix  $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$  werden aus den optimalen Parametern aus 2.7 entnommen, wobei  $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \text{diag}(\sigma_{\boldsymbol{\theta}_i})$  aus der Unsicherheit der optimalen Parametern berechnet wird. Wir definieren  $\mathbf{u}_j$  als die Abweichung der numerischen Lösung am Zeitpunkt  $t_j$  vom korrekten Wert:

$$\mathbf{x}_{j, \text{corr.}} = \mathbf{x}_j + \mathbf{u}_j. \quad (21)$$

Analog zu den Parametern wird angenommen, dass  $\mathbf{u} \sim N(0, \boldsymbol{\Sigma}_{\mathbf{X}})$  als eine normalverteilte Zufallsgröße mit dem Mittelwert  $\mu = 0$  und der Kovarianz  $\boldsymbol{\Sigma}_{\mathbf{X}} = \text{diag}(\sigma_{S_1}, \dots, \sigma_{S_k}, \sigma_{I_1}, \dots, \sigma_{I_k}, \sigma_{R_1}, \dots, \sigma_{R_k}, \sigma_{D_1}, \dots, \sigma_{D_k})$  angegeben werden kann.

Setzt man dies in Gleichung 21 ein, kombiniert mit Gleichung 18 erhält man:

$$\mathbf{x}_{j+1, \text{corr.}} = \mathbf{x}_j + \mathbf{u}_j + h_j \mathbf{F}(\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}, \mathbf{x}_j + \mathbf{u}_j). \quad (22)$$

Wird eine normalverteilte Zufallsgröße  $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  mit der Beziehung  $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{b}$  transformiert, erhält man eine normalverteilte Zufallsgröße  $\mathbf{y}$  mit dem Mittelwert  $\boldsymbol{\mu}' = \mathbf{M}\boldsymbol{\mu} + \mathbf{b}$  und der Kovarianzmatrix [13].:

$$\boldsymbol{\Sigma}' = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^T \quad (23)$$

Um dies auf  $\mathbf{u}_0$  anzuwenden, muss das System  $\mathbf{F}(\boldsymbol{\theta}, \mathbf{x})$  erst mit Hilfe der Taylorentwicklung linearisiert werden. Zu beachten ist, dass nur bis zur ersten Ordnung erweitert wird, sodass eine Linearität erhalten wird. Mit der Linearisierung des Systems mit der Taylorentwicklung<sup>11</sup> von  $\mathbf{F}$  um die Stelle  $(\boldsymbol{\theta}_0, \mathbf{x}_j)$  [14, S. 154] erhält man eine induktive Definition der Abweichung der Simulation:

$$\mathbf{u}_{j+1} = [\mathbf{id} + h_j \mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta}, \mathbf{x}_j)] \mathbf{u}_j + h_j \mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{x}_j) \Delta\boldsymbol{\theta} \quad (24)$$

deren explizite Form<sup>12</sup> sich wie folgt ergibt:

$$\mathbf{u}_n = \underbrace{\prod_{i=0}^{n-1} [\mathbf{id} + h_i \mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta}, \mathbf{x}_i)] \mathbf{u}_0}_{\mathbf{A}} + \underbrace{\sum_{i=0}^{n-1} \prod_{j=i+1}^{n-1} [\mathbf{id} + h_j \mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta}, \mathbf{x}_j)] h_i \mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{x}_i) \Delta\boldsymbol{\theta}}_{\mathbf{B}}. \quad (25)$$

Das System kann näherungsweise durch die lineare Gleichung

$$\mathbf{u}_n = (\mathbf{A} \quad \mathbf{B}) \begin{pmatrix} \mathbf{u}_0 \\ \Delta\boldsymbol{\theta} \end{pmatrix} \quad (26)$$

aufgefasst werden. Damit kann mit Hilfe von Gleichung 23 die Kovarianzmatrix der Simulation zu beliebigen Zeitpunkten ausgerechnet werden. Um daraus den Fehler der Simulation zu entnehmen gilt (nach Definition):

$$\sigma_{x_i}^2 = e_{x_i}^T \boldsymbol{\Sigma} e_{x_i} \quad (27)$$

<sup>11</sup>Zur Definition von  $\mathbf{J}_{\mathbf{x}}, \mathbf{J}_{\boldsymbol{\theta}}$  sowie deren Herleitung siehe Anhang A

<sup>12</sup>Siehe Anhang B für den Beweis

### 2.8.1 Monte-Carlo Verfahren

Voraussetzung für das vorherige Verfahren ist die Differenzierbarkeit des Systems, da dies bei dessen Linearisierung notwendig ist. Beim Stufen-Ansatz ist die Differenzierbarkeit aufgrund der charakteristischen Funktionen (vgl. Gleichung 13) nicht gewährleistet. Daher nehmen wir einen anderen Ansatz für die Fehlerfortpflanzung. Ausgehend von der Verteilung der Fehler der Anfangswerte  $\mathbf{u}_0 \sim N(0, \Sigma)$  sowie der Fehler der Parameter  $\Delta\theta \sim N(0, \Sigma_\theta)$  wählen wir mehrfach zufällig verschiedene Parameter und Startwerte für die Simulation mit der entsprechenden Verteilung. Den Fehler  $\sigma_{\mathbf{x}_i^s}$  der  $s$ -ten Komponente der Simulation zum  $i$ -ten Zeitpunkt erhalten wir dann durch Berechnen der Standardabweichung der gleichen Komponenten zu den gleichen Zeitpunkten über alle resultierenden Simulationen.

## 2.9 Methoden

### 2.9.1 Der klassische Runge-Kutta-Algorithmus

Das hier betrachtete Problem ist ein System aus gewöhnlichen Differentialgleichungen. Um dieses System numerisch zu lösen bietet sich unter anderem der klassische Runge-Kutta-Algorithmus (ab hier RK4 genannt) an. Der Algorithmus folgt der folgenden Vorschrift [15, S. 1028]:

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, t), \quad (28a)$$

$$\mathbf{k}_1 = \mathbf{F}(\mathbf{x}_i, t_i), \quad (28b)$$

$$\mathbf{k}_2 = \mathbf{F}\left(\mathbf{x}_i + \frac{h}{2}\mathbf{k}_1, t_i + \frac{h}{2}\right), \quad (28c)$$

$$\mathbf{k}_3 = \mathbf{F}\left(\mathbf{x}_i + \frac{h}{2}\mathbf{k}_2, t_i + \frac{h}{2}\right), \quad (28d)$$

$$\mathbf{k}_4 = \mathbf{F}(\mathbf{x}_i + h\mathbf{k}_3, t_i + h), \quad (28e)$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \frac{h}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4). \quad (28f)$$

Der RK4-Algorithmus hat eine Abweichung der Ordnung  $\mathcal{O}(h^5)$ . Die numerische Implementation in diesem Projekt erfolgt über die SciPy-Bibliothek, um genauer zu sein mit der Funktion `scipy.integrate.solve_ivp`, welche einen RK4 verwendet. [16].

### 2.9.2 Parallelisierung

Den Code zu parallelisieren, kann die Laufzeit eines Codes drastisch verringern. Bei einem normalen Prozess, der nicht parallelisiert ist, werden die Arbeitsaufgaben hintereinander ausgeführt. Die Problematik an diesem Verlauf ist, dass die nächste Arbeitsaufgabe erst anfängt, wenn die vorherige Aufgabe beendet ist. Bei einem parallelisierten Prozess werden die Aufgaben nicht hintereinander ausgeführt, sondern parallel. Es gibt jedoch einige Beschränkungen für die Parallelisierung:

- **Unabhängigkeit:** Wenn die Arbeitsaufgaben nicht voneinander unabhängig sind, kann keine isolierte Parallelisierung verwendet werden. Die Prozessorkerne kommunizieren im Normalfall nicht untereinander.
- **Prozessorkerne:** Die Anzahl der Prozessorkerne, auf die sich die Aufgaben verteilen können, beeinflusst stark die Laufzeit des ganzen Prozesses.

Eine schematische Darstellung einer Parallelisierung eines Prozesses ist in Abbildung 8 dargestellt.

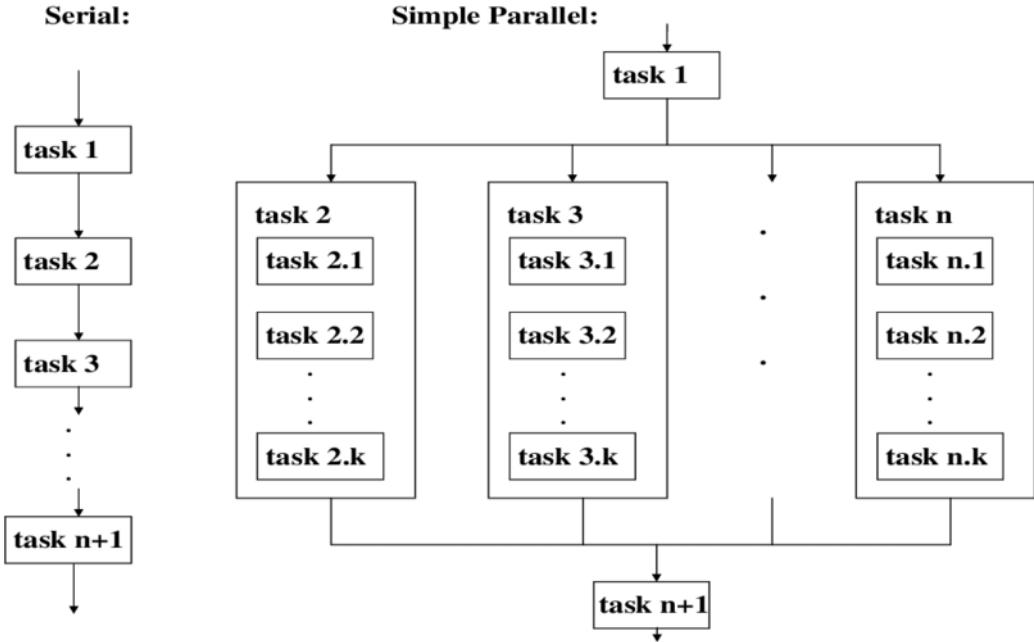


Abbildung 8: Visuelle Darstellung des Verlaufs eines parallelisierten Prozesses [17].

### 3 Analyse

#### 3.1 Zu der Auswirkung der Parameter

Im Folgenden werden die Parameter genauer betrachtet, insbesondere wie das System durch Änderungen der Parameter sich beeinflussen lassen.

##### 3.1.1 Variation von $t_0$

Zunächst ist festzuhalten, dass in Gleichung (13), die Systemgleichung des konstanten Ansatzes, der Ausdruck für  $\frac{dS_i}{dt}$  aus einem gewichtetem Mittelwert besteht, wobei der Koeffizient  $t_0$  die Gewichtung angibt. Anders ausgedrückt, die Änderung der Anfälligen und der Infizierten ist das gewichtete Mittel aus dem Anteil wenn Pendler zu Hause und ausgependelt sind. In Abbildung 9 sind die Kurven für verschiedene  $t_0$  aufgetragen. Die Kurven sind also jeweils durch unterschiedliche gewichtete Mittelwerte entstanden. In dieser Abbildung ist zu sehen, dass es für den Landkreis Göttingen keinen großen Unterschied macht, den Parameter  $t_0$  zu variieren. Im Landkreis Eichsfeld jedoch, welcher in Abbildung 21 dargestellt ist, sind deutliche Unterschiede zu erkennen.

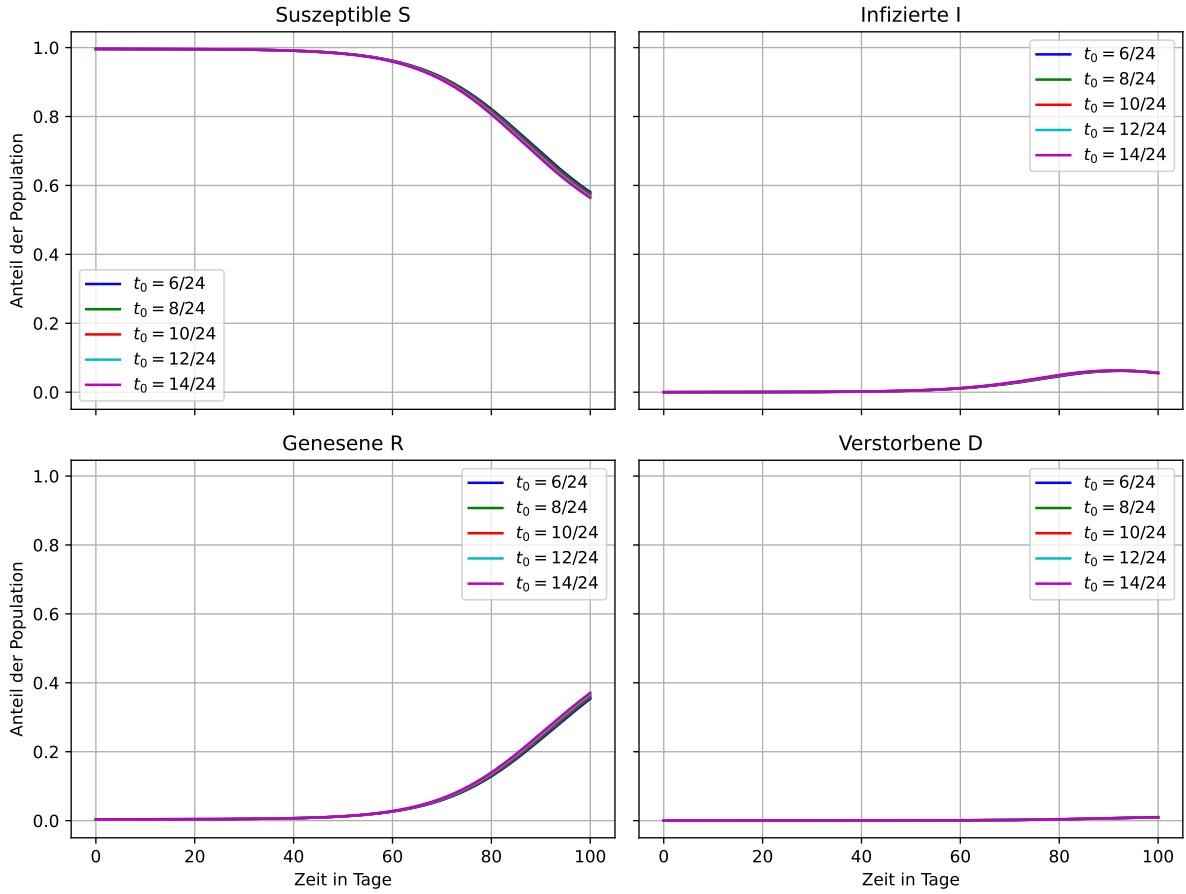
##### 3.1.2 Variation der $\chi$ -Intervalle

In Abbildung 10 sind die numerischen Lösungen der Differentialgleichungen mit dem Stufen-Ansatz für verschiedene Anfangswerte der Partitionen aufgetragen. Wobei die allgemeine Form der Kurven sehr gleich sind, können eindeutig Unterschiede erkannt werden. Interessanterweise wächst der Anteil der Infizierten am schnellsten für die mittlere Aufteilung, bei der  $\pi_2 = [0,4; 0,8]$  ist.

##### 3.1.3 Das frühe Verhalten der Infektionswelle

Es ist bekannt, dass die normale SIR-Gleichung für kleine Zeiten ein exponentielles Anwachsen der Infizierten vorhersagt. Um zu sehen, wie die dynamischen Modelle sich für kleine Zeiten entwickeln, sind in Abbildung 11 die Infizierten der beiden Ansätze, mit einer logarithmischen Skala, gegen die Zeit aufgetragen. Um die Systeme

Zeitentwicklung in Knoten 5 für verschiedene  $t_0$  im konstanten Ansatz;  $\alpha = 0.3$ ,  $\beta = 1/5$ ,  $p = 0.0264$



**Abbildung 9:** Plot der numerischen Lösungen der Differentialgleichungen im konstanten Ansatz für verschiedene  $t_0$  im Landkreis Göttingen (Knoten 5).

besser zu verstehen, werden diese für vier verschiedene  $\alpha$  aufgetragen. Für den konstanten Ansatz ist eindeutig zu erkennen, dass sich  $I$  exponentiell verhält, für den Stufen-Ansatz jedoch nicht.

### 3.1.4 Grenzfall niedriger Anzahl von Infizierten

Wenn die Infektionen sehr niedrig sind, ist zu erwarten, dass die Modelle ähnliche Verläufe haben. In Abbildung 13 ist eine solche Situation dargestellt. Unter diesen Voraussetzungen kann also  $S_i = 1$  für alle  $i$  angenommen werden. Somit wird

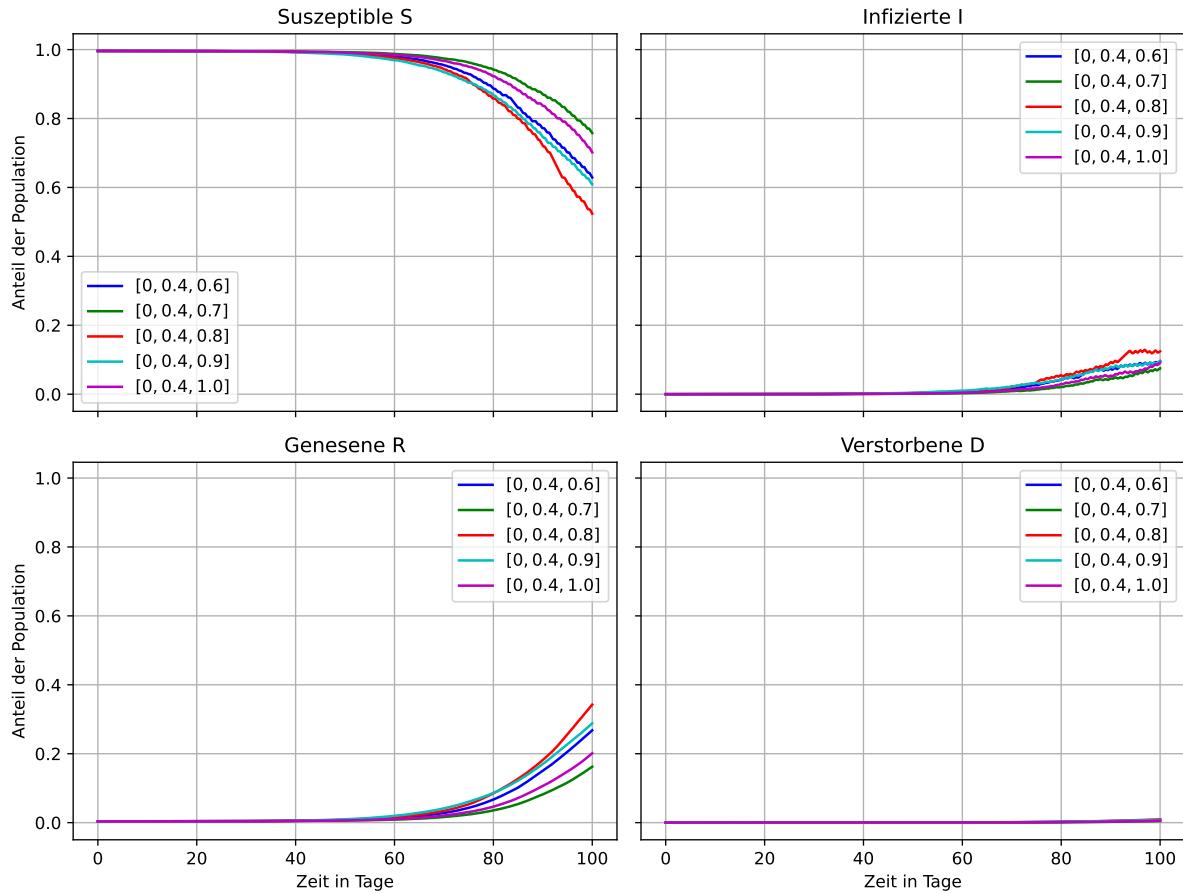
$$\frac{dI_i}{dt} = \gamma_1 \alpha I_i + \gamma_2 \alpha \left[ \frac{N_i^{\text{rest}}}{N_i} \frac{N_i^{\text{rest}} I_i + \sum_j \mathcal{P}_{[j \rightarrow i]} I_j}{N_i^{\text{rest}} + \sum_j \mathcal{P}_{[j \rightarrow i]}} + \sum_j \frac{\mathcal{P}_{[i \rightarrow j]}}{N_i} \frac{N_j^{\text{rest}} I_j + \sum_k \mathcal{P}_{[k \rightarrow j]} I_k}{N_j^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow j]}} \right] - \beta I_i. \quad (29)$$

Hierbei sind  $\gamma_1$  und  $\gamma_2$  eine Verallgemeinerung der Gewichtung, welche durch  $t_0$  oder charakteristische Funktionen gegeben sind. Dies ist eine lineare Gleichung, die geschrieben werden kann als

$$\frac{d\mathbf{I}}{dt} = \mathbf{M}\mathbf{I}, \quad (30)$$

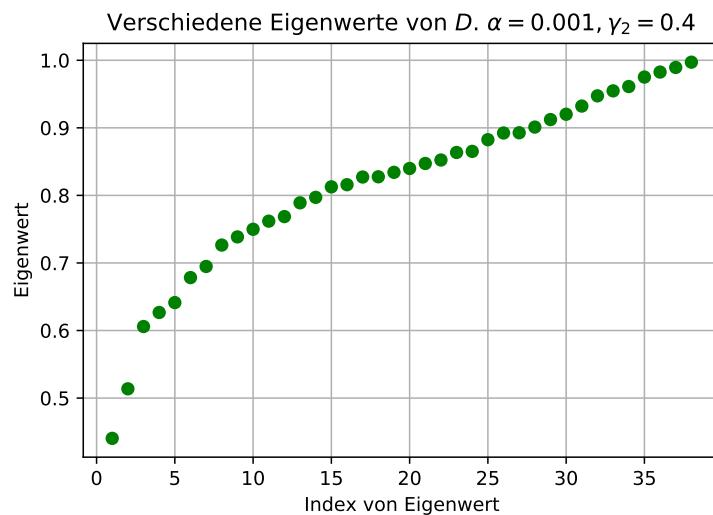
wobei  $\mathbf{I}$  ein Vektor aller Infizierten ist. Die Matrix  $\mathbf{M} = (\gamma_1 \alpha - \beta) \mathbf{id} + \gamma_2 \alpha \mathbf{D}$  ist durch Gleichung (29) gegeben. Die genaue Form ist in Sektion C im Anhang dargestellt. Gleichung (30) wird durch

Zeitentwicklung in Knoten 5 für verschiedene  $\pi_1$ ,  $\pi_2$  &  $\pi_3$  im Stufen-Ansatz;  $\alpha = 0.4$ ,  $\beta = 1/4$ ,  $p = 0.0264$



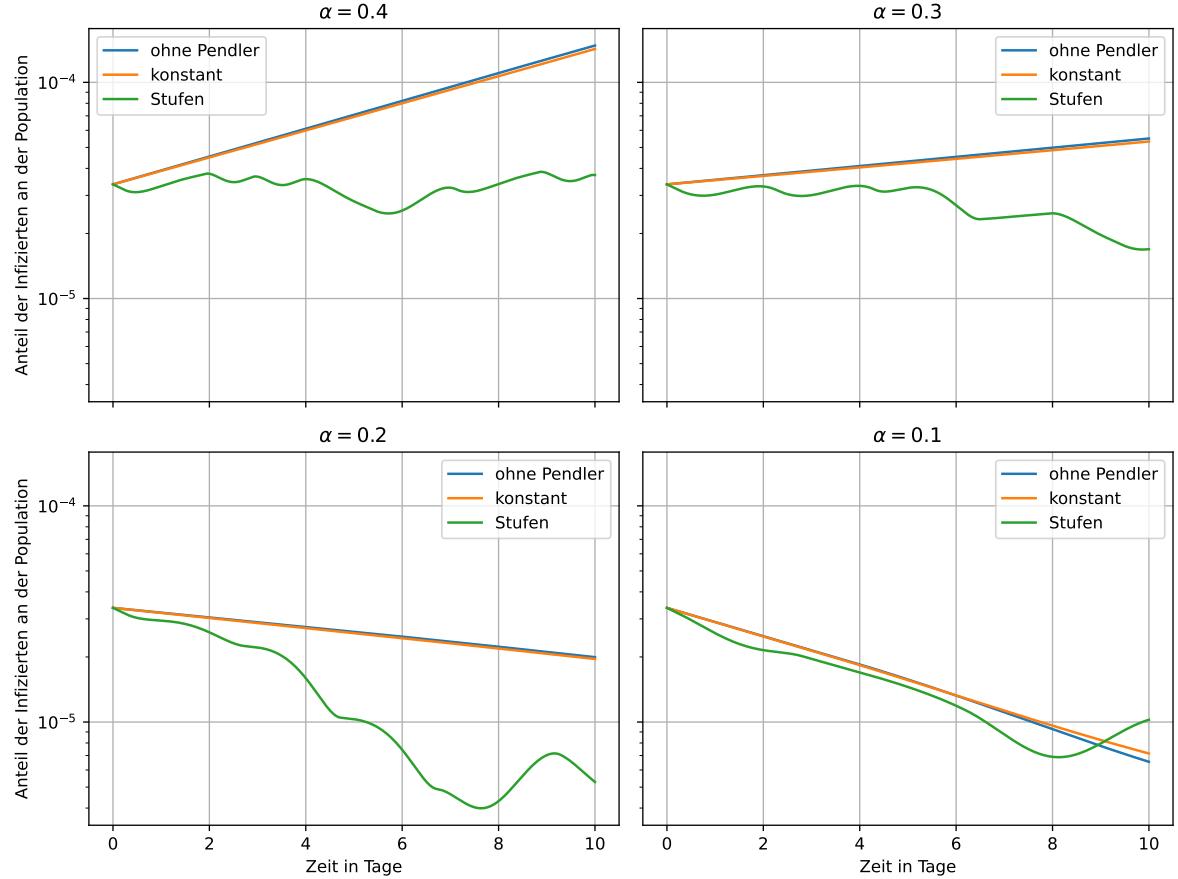
**Abbildung 10:** Plot der numerischen Lösungen der Differentialgleichungen für verschiedene Grenzen der charakteristischen Funktionen  $\chi_{[a,b]}$  im Landkreis Göttingen (Knoten 5).

$$\mathbf{I}(t) = \exp\{\mathbf{M}t\}\mathbf{I}_0 \quad (31)$$



**Abbildung 12:** Die Eigenwerte der Matrix  $\mathbf{D}$  mit 38 Landkreisen.

Anfängliches Verhalten der Infektion in Knoten 5,  $\beta = 1/4$ ,  $p = 0.0264$ ,  $t_0 = 0.4$ ,  $\pi_1 = [0, 0.4]$ ,  $\pi_2 = [0.4, 0.8]$ ,  $\pi_3 = [0.8, 1]$



**Abbildung 11:** Aufgetragen sind die Verläufe von  $I$  nach Simulation der ersten zehn Tage für verschiedene  $\alpha$  für den einfachen Ansatz (ohne Pendlerdynamik) sowie die beiden Ansätze mit Pendlern

gelöst, wobei  $\mathbf{I}_0$  der Vektor mit den Infizierten bei  $t = 0$  ist. Die Eigenwerte der Matrix  $\mathbf{D}$  zeigen, wie stark die Pendlerdynamik das System beeinflusst. Der erste Teil von  $\mathbf{M}$  ist ein Vielfaches der Identität und hat somit nur gleiche Eigenwerte. Die Matrix  $\gamma_2 \alpha \mathbf{D}$ , die die Pendlerdynamik beschreibt, hat jedoch verschiedene Eigenwerte. Diese Eigenwerte sind in Abbildung 12 aufgetragen.

Wenn das System homogen ist, d.h.  $I_i = I_j$  für alle  $i, j$ , so kann Gleichung (29) geschrieben werden als

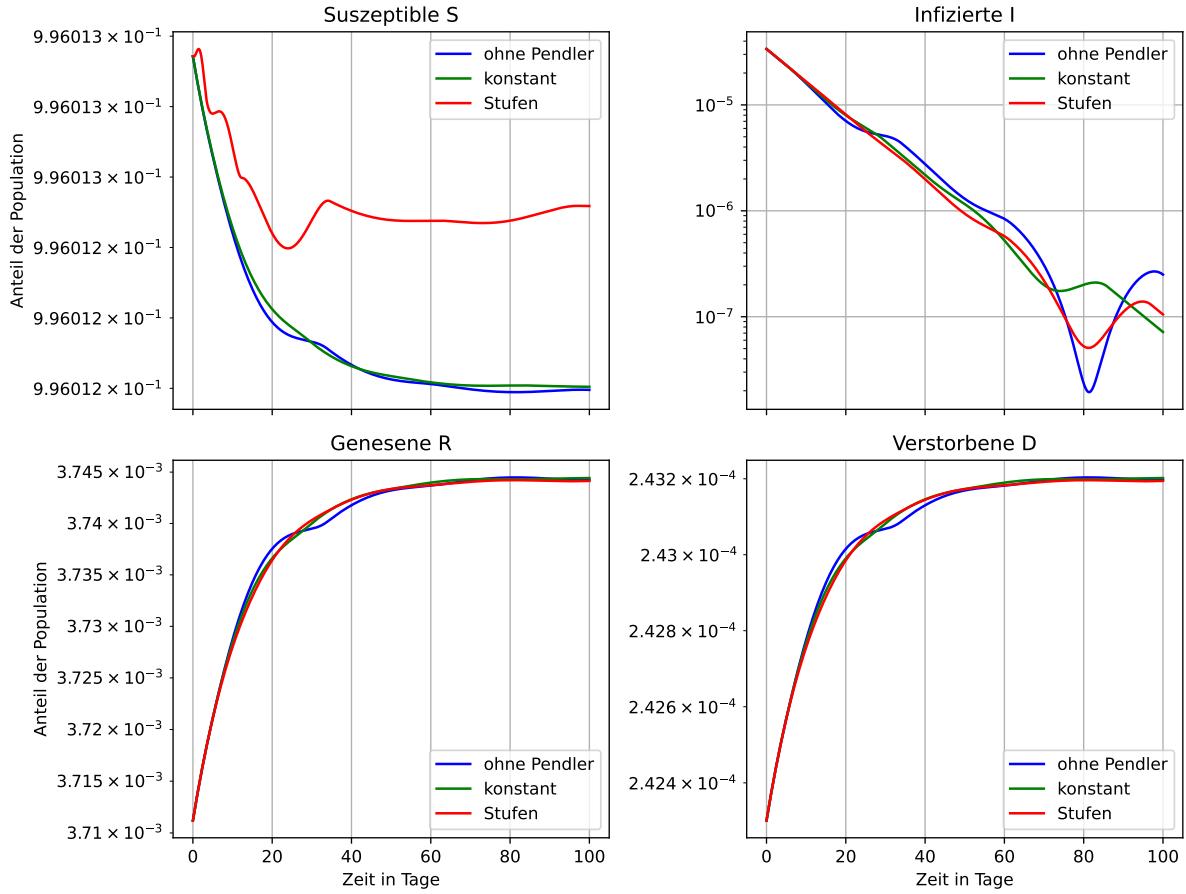
$$\frac{dI_i}{dt} = \gamma_1 \alpha I_i + I_i \left( \gamma_2 \alpha \left[ \frac{N_i^{\text{rest}}}{N_i} \frac{N_i^{\text{rest}} + \sum_j \mathcal{P}_{[j \rightarrow i]}}{N_i^{\text{rest}} + \sum_j \mathcal{P}_{[j \rightarrow i]}} + \sum_j \frac{\mathcal{P}_{[i \rightarrow j]}}{N_i} \frac{N_j^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow j]}}{N_j^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow j]}} \right] \right) - \beta I_i.$$

und mit  $N_i = N_i^{\text{rest}} + \sum_j \mathcal{P}_{[i \rightarrow j]}$  reduziert sich diese Gleichung auf

$$\frac{dI_i}{dt} = \gamma_1 \alpha I_i + \gamma_2 \alpha I_i - \beta I_i = \alpha I_i - \beta I_i. \quad (32)$$

Für ein homogenes System mit niedrigen Infizierten ist also der Eigenwert der Pendlermatrix  $\mathbf{D}$  1, welches auch in der Abbildung 12 zu sehen ist.

Vergleich der Modelle in Knoten 5 für  $\alpha = 0.001$ ,  $\beta = 1/14$ ,  $t_0 = 0.4$ ,  $\pi_1 = [0, 0.4]$ ,  $\pi_2 = [0.4, 0.8]$ ,  $\pi_3 = [0.8, 1]$



**Abbildung 13:** Vergleich der drei DGL-Ansätze für ein sehr kleines  $\alpha$  im Landkreis Göttingen (Knoten 5). Um eine niedrige Infektion zu gewährleisten wird  $\alpha$  sehr klein gewählt.

### 3.2 Stabilität des Systems

Nun wird untersucht, ob das System einen stabilen Zustand haben kann. Die Bedingung hierfür ist, dass

$$\frac{dI_i}{dt} = 0. \quad (33)$$

Es gibt den trivialen Fall, bei dem  $\mathbf{I} = \mathbf{0}$ . Für den nicht trivialen Fall folgt nach einigem Umstellen der Gleichung die Bedingung

$$\frac{\beta}{\alpha} = \left[ \gamma_1 + \gamma_2 \cdot \left( \frac{N_i^{\text{rest}}}{N_i} \frac{N_i^{\text{rest}} + \sum_j \mathcal{P}_{[j \rightarrow i]} I_{ij}^{\text{rel}}}{N_i^{\text{rest}} + \sum_j \mathcal{P}_{[j \rightarrow i]}} + \sum_j \frac{\mathcal{P}_{[i \rightarrow j]} I_{ij}^{\text{rel}} N_j^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow j]} I_{ik}^{\text{rel}}}{N_j^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow j]}} \right) \right] \cdot S_i, \quad (34)$$

mit  $I_{ij}^{\text{rel}} = \frac{I_j}{I_i}$ . Unter der Annahme, dass die Koeffizienten  $\alpha$  und  $\beta$  konstant sind, kann das ganze System nicht stabil sein. Hierfür müsste auch die rechte Seite zeitlich unabhängig sein. Alles in der Klammer ist zeitlich unabhängig, da  $\frac{dI_i}{dt} = 0$  vorausgesetzt wird, allerdings ist  $S_i$  nicht zeitunabhängig.

## 4 Ergebnisse

### 4.1 Optimierte Parameter der betrachteten Modelle

Die Modell-Parameter  $\alpha$  und  $\beta$  für den Stufen-Ansatz sowie den konstanten Ansatz werden mit Hilfe des in Abschnitt 2.7 erklärten Verfahrens optimiert. Dabei werden für jedes Modell 12 und 38 Landkreise simuliert<sup>13</sup>. Die Bewertungsfunktion  $Bew$  berücksichtigt hierbei alle simulierten Landkreise. Für die Pendlerdynamik des konstanten Ansatzes wird  $t_0$  auf 0,4 festgesetzt, da 40% eine plausible Abschätzung für den Anteil des Tages ist, in dem die Pendler ausgependelt sind. Analog wird diese Pendlerdynamik durch die  $\chi$ -Intervalle des Stufen-Ansatzes  $\pi_1 = [0; 0,4]$ ,  $\pi_2 = [0,4; 0,8]$  und  $\pi_3 = [0,8; 1]$  widergespiegelt. In allen Fällen werden 100 verschiedene Anfangswerte der Parameterpaare aus der Menge  $\epsilon$  optimiert.

Die optimalen Parameter sind in Tabelle 2 aufgetragen. Zum Vergleich sind ebenfalls in Tabelle 2 einige Literaturwerte dargestellt.

**Tabelle 2:** Übersicht über die optimalen Parameter  $\alpha$  und  $\beta$  für die untersuchten Pendlermodelle. Optimierte wurden hierfür Simulationen des jeweiligen Modells mit  $N$  Landkreisen.

Quelle	Zeitbereich	Modell	$N$	$\alpha$ in $d^{-1}$	$\beta$ in $d^{-1}$
Optimierung	24.07.-01.11.2020	Stufen-Ansatz	12	$1,0 \pm 2,8$	$1,0 \pm 1,0$
Optimierung	—”—	Konstanter Ansatz	12	$0,40 \pm 0,11$	$0,37 \pm 0,11$
Optimierung	—”—	Stufen-Ansatz	38	$0,45 \pm 0,20$	$0,26 \pm 0,17$
Optimierung	—”—	Konstanter Ansatz	38	$0,33 \pm 0,08$	$0,30 \pm 0,08$
Literatur [18]	10.12.2019 - 09.04.2020	SIR	—	0.83	0,74
Literatur [19]	24.02.-21.03.2020	SIR (stochastisch)	—	0.38	$0,135 - 0,45$
Literatur [20]	22.01.- 22.03.2020	SIR	—	0.28	0,1
Literatur [21]	15.02. - 25.05.2020	SIR	—	0.4	1/14

Die optimalen Werte der Parameter liegen für die Pendlermodelle mit Ausnahme eines Ausreißers (Stufen-Ansatz für 12 Landkreise) im  $1\sigma$ -Intervall voneinander entfernt<sup>14</sup>. Bemerkenswerterweise weisen die Literaturwerte der Parameter unter sich große Unterschiede auf. Diese lassen sich auf die verschiedenen Modelle, Zeiträume sowie Annahmen der verwendeten Quellen zurückführen:

- In [18] wird der frühe Pandemie-Verlauf bei der Berechnung der Parameter berücksichtigt. Ein Teil dieser Periode kennzeichnet sich durch Mangel an Maßnahmen und eine nicht genau geschätzte Dunkelziffer. Dies kann ein Grund für den relativ hohen Wert von  $\alpha$  sein.
- In [19] und [20] werden bei der Modellierung Maßnahmen, wie Testen, Lockdowns sowie Quarantäne berücksichtigt. In [20] ist  $\beta$  per Annahme auf 1/10 festgesetzt.

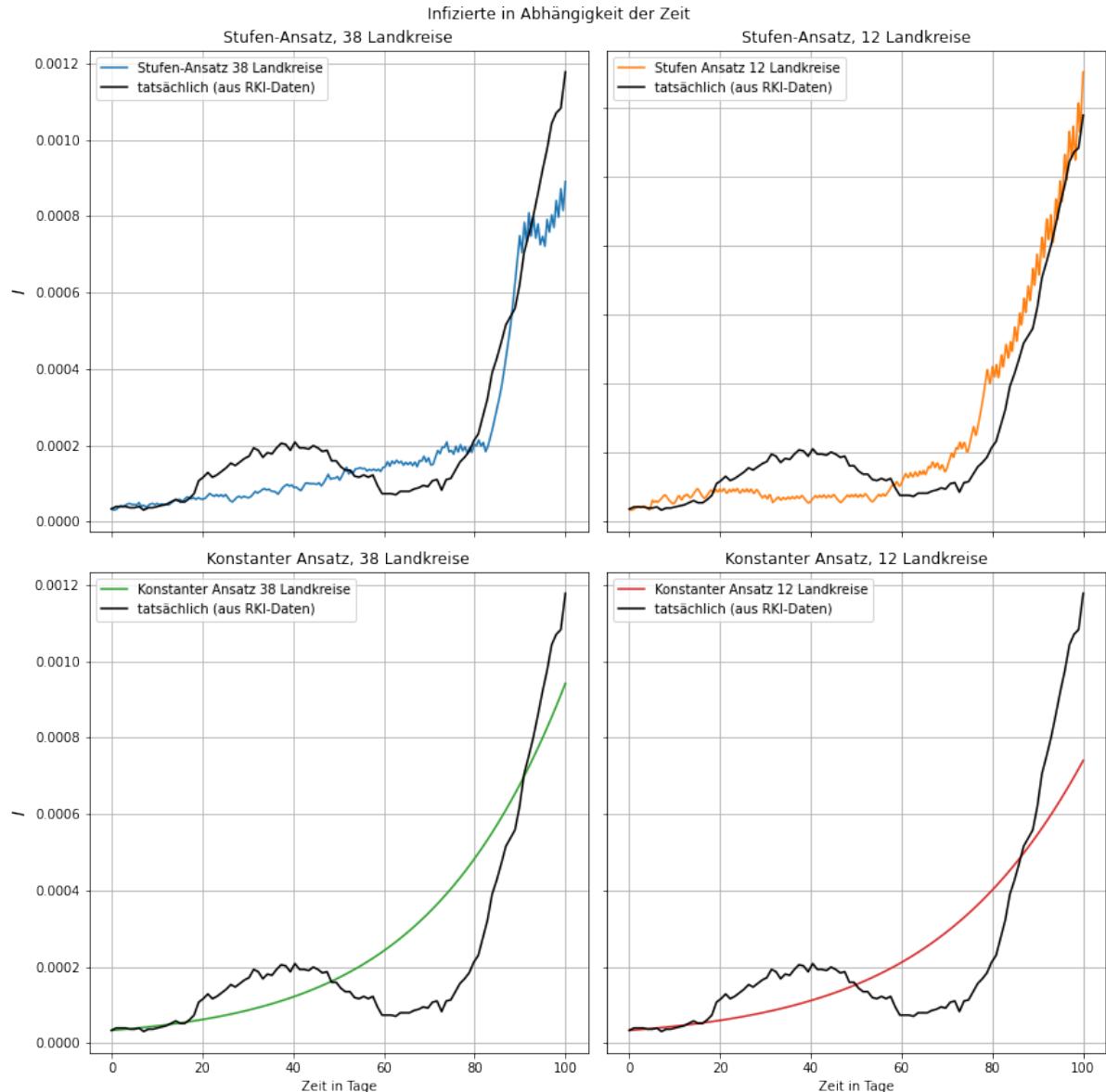
Mit Ausnahme der Parameterwerte aus [18] sind die Literaturwerte im  $1\sigma$ -Intervall der optimalen Parametern. Jedoch ist erkennbar, dass die optimierten Parameter eine zu große Streuung aufweisen. Dies wird in 5.4 genauer diskutiert.

### 4.2 Simulationen mit optimierten Parametern

Unter Berücksichtigung der im vorherigen Abschnitt ermittelten Parameter können nun Simulationen durchgeführt werden, von denen eine deutliche Realitätsnähe zu erwarten ist. Dargestellt wird also in Abbildung 14 der simulierte Anteil der aktuell Infizierten im Landkreis Göttingen in dem Zeitraum in 2.7.1. Zum Vergleich werden die verschiedenen Modelle mit den optimalen Parametern aus 4.1 zusammen mit den RKI-Vergleichsdaten

<sup>13</sup>Jeweils die ersten  $N$  Landkreise aus Tabelle 4 (siehe dafür Anhang.)

<sup>14</sup>Dies gilt sogar für alle, wenn die exakten numerischen Werte betrachtet werden. Hier ist das aufgrund des wissenschaftlichen Rundens nicht ersichtlich.



**Abbildung 14:** Vergleich von simulierten Epidemieverläufen mit optimierten Parametern und den tatsächlichen Verläufen im Landkreis Göttingen für den Zeitbereich in 2.7.1.

aufgetragen. Erkennbar ist, dass die Simulation mit dem Stufen-Ansatz eine bessere Übereinstimmung mit dem tatsächlichen Verlauf aufweist.

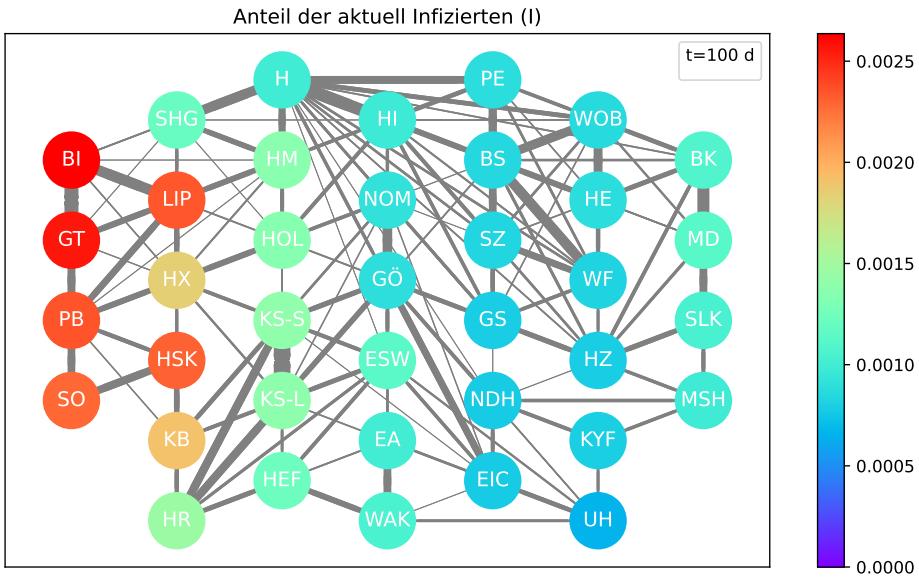
### 4.3 Bewertung der Qualität des Netzwerkmodells

Im Modell wird die Region als abgeschlossen betrachtet, es existieren ausschließlich Pendlerbewegungen zwischen Landkreisen in der betrachteten Region. Um zu untersuchen, ob dadurch das Infektionsgeschehen in Landkreisen, die tatsächlich zusätzliche Pendlerverbindungen mit externen Landkreisen aufweisen, möglicherweise schlechter modelliert<sup>15</sup> ist, werden drei Verfahren zur Bewertung in Betracht gezogen:

#### 4.3.1 Untersuchung auf Korrelation zwischen Vernetzung eines Knotens und Qualität der Simulation

Zur quantitativen Auswertung und Bewertung des Modells muss sowohl ein Maß für die Qualität einer Simulation als auch für die Relevanz des Modells gefunden werden.

<sup>15</sup>das heißt: die Abweichung des Infektionsgeschehens in diesem Landkreis ist größer als im Mittel aller Landkreise in der Region



**Abbildung 15:** Der Zustand des Netzwerks nach Ende der Simulation. Eine Zeitreihe dieser Abbildung ist mit Abbildung 23 im Anhang D.3 enthalten. Darüber hinaus existiert zur Darstellung des zeitlichen Verlaufs ein Video im Repository (Media/network\_timeline.mp4).

Um die Qualität zu bewerten, wird die Kontrollgröße

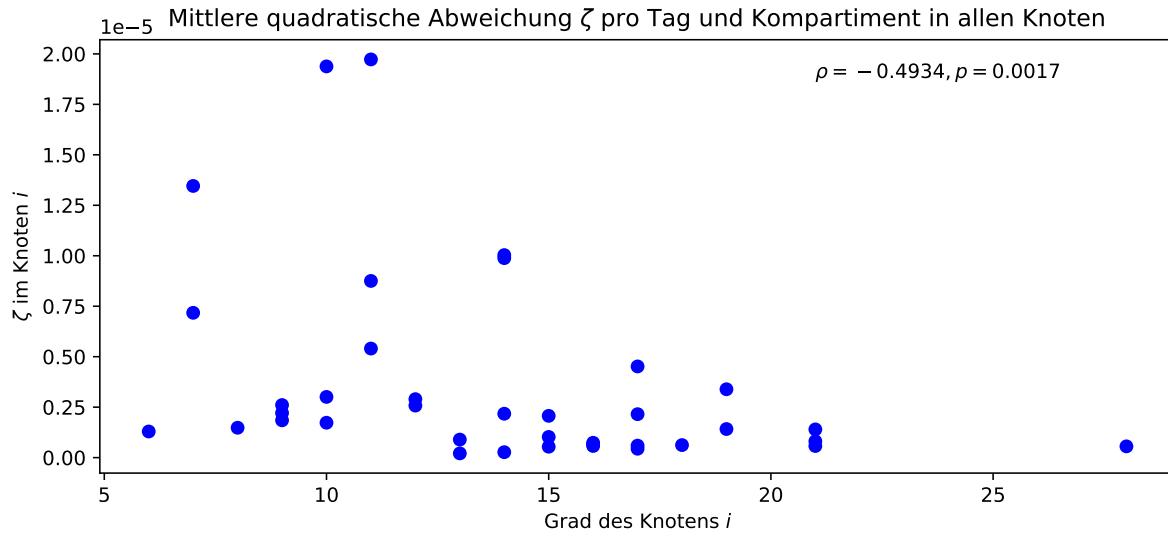
$$\zeta_i = \frac{\sum_{C=S,I,R,D} \sum_{t=0}^{100} (\Delta C_i(t))^2}{100 \text{ Tage} \cdot 4 \text{ Kompartimente}} \quad (35)$$

eingeführt. Sie gibt die quadratische Abweichung zwischen realem und simuliertem Verlauf an, gemittelt über alle Tage (also Zeiten  $t$ ) und Kompartimente  $C$ .

Der Grad eines Knotens ist ein Maß dafür, wie stark der Knoten in das Netzwerk integriert ist. Daraus können allerdings keine Rückschlüsse auf die geographische Lage des Knotens gezogen werden. Wir definieren den Grad als Summe der Anzahl der Ein- und Auspendler-Beziehungen. Es ist davon auszugehen: Je höher der Grad eines Knotens ist, desto stärker ist er eingebunden in die Pendlerdynamik auf dem Netzwerk. Wenn also eine negative Korrelation zwischen Grad und bewertender Kontrollgröße  $\zeta$  besteht, ist davon auszugehen, dass das Modell erfolgreich arbeitet, in dem Sinn, dass die Berücksichtigung der Pendlerdynamik zu besseren Simulationsergebnissen führt. Abbildung 16 zeigt ein Streudiagramm, in dem jeder Punkt einen Landkreis darstellt. Auf der Ordinatenachse ist sein Grad angegeben und auf der Abszissenachse die Kontrollgröße  $\zeta$ . Tatsächlich ist optisch erkennbar, dass die mittlere quadratische Abweichung  $\zeta$  für Landkreise mit geringem Knoten größer ist. Um eine quantitative Aussage zu treffen, wird der Spearman-Korrelationskoeffizient<sup>16</sup> der Verteilung mit `scipy.stats` berechnet. Er ist robuster gegenüber Ausreißern als der sonst übliche Pearson-Korrelationskoeffizient. Daher ist seine Verwendung, insbesondere vor dem Hintergrund von Superspreading (diskutiert in Abschnitt 5.2.1), sinnvoll. Außerdem erkennt nicht nur lineare sondern allgemeine monotone Korrelationen [22, S.143ff.], was ebenfalls sinnvoll ist, da die Art einer möglichen Korrelation zwischen Grad und mittlerer Abweichung a priori nicht bekannt ist. Im Fall des in Abbildung 16 gezeigten Streudiagramms beträgt die Spearman-Korrelation zwischen den Variablen Grad und mittlerer Abweichung  $\rho = -0,4934$ . Demnach existiert eine negative Korrelation mit einer gewissen Streuung. Darüber hinaus kann mit `scipy.stats` die Signifikanz  $p$  dieser Korrelation berechnet werden. Sie gibt den  $\alpha$ -Fehler eines Hypothesentests mit der Nullhypothese „Grad und mittlere Abweichung sind unkorreliert“ an [22, S.419] und beträgt in diesem Fall  $p = 0,0017$ . Die Annahme, dass zwischen Grad des Knotens  $i$  und Qualität der Simulation gemessen an  $\zeta_i$  kein Zusammenhang besteht, ist also lediglich mit einer Wahrscheinlichkeit von 0,17%

<sup>16</sup>Korrelationskoeffizienten nehmen für eine perfekte positive Korrelation den Wert +1, für eine perfekt negative Korrelation den Wert -1 und für keine Korrelation den Wert 0 an. Werte dazwischen geben die Streuung an.

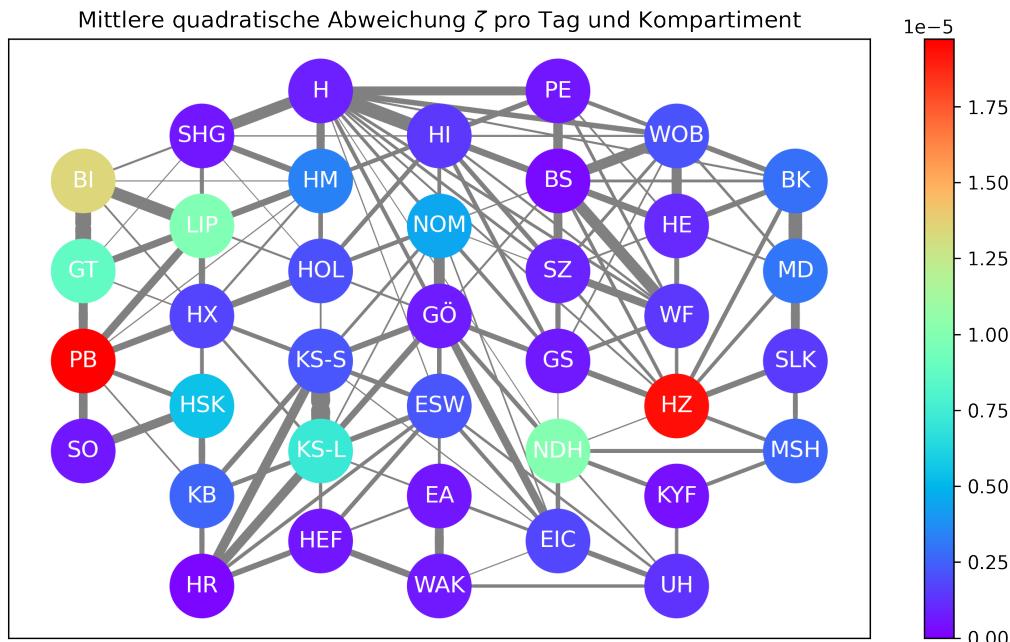
wahr. Die Schlussfolgerungen aus diesen Beobachtungen werden in Abschnitt 5.3 der Diskussion erläutert.



**Abbildung 16:** Mittlere quadratische Abweichung zwischen modelliertem und realem Verlauf pro Tag und Kompartiment für jeden Knoten aufgetragen gegen den Grad des Knotens. Die Simulation wurde mit dem Stufen-Ansatz für die tägliche Pendlerdynamik durchgeführt.

#### 4.3.2 Betrachtung im Netzwerk

Abbildung 17 zeigt die Größe  $\zeta$  farblich eingetragen in ein Netzwerkdigramm. Augenscheinlich ist die Qualität in Regionen mit hoher Vernetzung (im Plot: viele und dicke Verbindungslien, zum Beispiel rund um HI, BS, SZ, u.s.w.) sehr gut. Knoten, die im Netzwerk nur wenige (zum Beispiel BI, GT, PB) oder schwache (z.B. HZ) Pendlerverbindungen aufweisen, sind eher schlechter simuliert. Es existieren Ausnahmen, außerdem lässt dieser optische Vergleich keine verallgemeinernden Schlussfolgerungen zu.



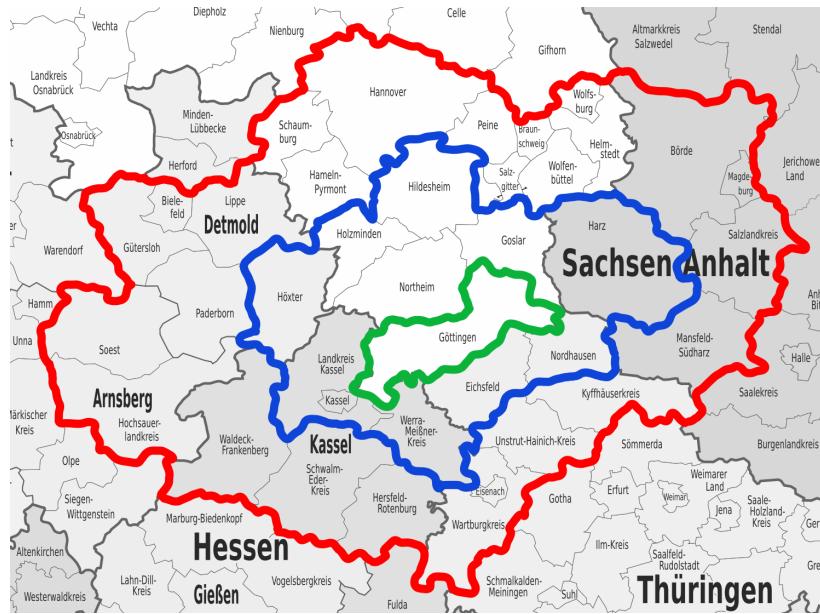
**Abbildung 17:** Mittlere Abweichung  $\zeta$  in jedem Knoten des Netzwerks.

### 4.3.3 Vergleich von Simulationen mit verschiedenen breiten Rändern

Dieser Ansatz sieht vor, den simulierten Anteil an Infizierten verglichen mit dem tatsächlichen Anteil an Infizierten nur für den Landkreis Göttingen (der sich zentral in der betrachteten Region befindet) zu ermitteln. Dazu werden zwei Simulationen durchgeführt, die sich durch eine unterschiedliche Anzahl an betrachteten Landkreisen unterscheiden:

- Simulation der Region aus 12 Landkreisen. Es wird also ein Rand aus 11 Landkreisen um Göttingen gebildet.
- Simulation der Region aus 38 Landkreisen. Es wird also ein Rand aus weiteren 26 Landkreisen um die bisherige Region gebildet.

Die so entstehenden Schalen rund um den Landkreis Göttingen sind in Abbildung 18 visualisiert. Danach werden



**Abbildung 18:** Der Landkreis Göttingen (grün) zentral eingebettet in eine größere Region aus 12 Landkreisen (blau) und eine noch größere Region aus 38 Landkreisen (rot).

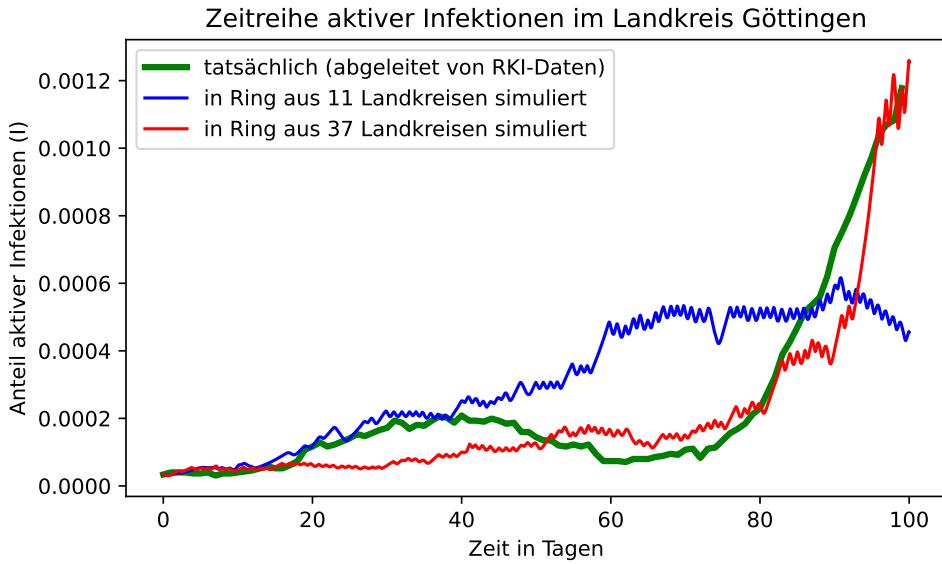
für jede Simulation nur die Ergebnisse für den Landkreis Göttingen (mit für den Landkreis Göttingen optimierten Parametern) untersucht und geprüft, ob die Berücksichtigung umliegender Landkreise zu einer verbesserten Beschreibung des tatsächlichen Verlaufs beiträgt. Das Ergebnis ist in Abbildung 19 gezeigt.

Erkennbar ist, dass der Verlauf bei einer Modellierung in einem Rand von mehr (37) Landkreisen näher an die Realität kommt als bei einer Modellierung in einem Rand von weniger (11) Landkreisen. Zu beachten ist, dass die Optimierungen nicht vollständig vergleichbar sind. Um diesen Umstand zu erklären, wird der Wert der Bewertungsfunktion, die in Gleichung (15) vorgestellt wurde<sup>17</sup>, für die Parameter zu den in Abbildung 19 dargestellten Modellverläufen berechnet. Während sie für den Ring aus 11 Landkreisen  $Bew \approx 8,6 \cdot 10^{-8}$  beträgt, nimmt sie für den größeren Ring einen geringeren Wert von  $Bew \approx 3,4 \cdot 10^{-8}$  an. Es wurden also tatsächlich für den größeren Ring bessere Parameter gefunden, als für den kleineren Ring. Die möglichen Gründe und Schlussfolgerungen sind Teil der Diskussion in Abschnitt 5.3.

## 4.4 Modellunsicherheit

Für die Unsicherheit des Anfangszustands  $\mathbf{x}_0$  wird 1% dessen Werts angenommen. Da jedoch manche Kompartimente Startwerte von  $X_i = 0$  haben und dies eine unrealistische Unsicherheit von  $\sigma = 0$  bedeutet, wird die

<sup>17</sup>Die Menge der betrachteten Landkreise für die Bewertungsfunktion ist hier nur die Menge mit dem Landkreis Göttingen  $\cap$  Menge der simulierten Landkreisen (bestehend aus entweder 12 oder 38 Regionen), das heißt: der Menge mit dem Landkreis Göttingen.



**Abbildung 19:** Resultate der Simulation im Landkreis Göttingen mit verschiedenen breiten Ringen darum. Dargestellt ist der zeitliche Verlauf des Anteils der aktuell Infizierten im Landkreis Göttingen. Dabei wurde für beide Ringe jeweils die Optimierung durchgeführt, die ausschließlich den Verlauf im Landkreis Göttingen vergleicht.

Unsicherheit auf mindestens 1 Person festgelegt. Also gilt, wenn  $N$  die Anzahl der Bevölkerung des Kompartiments  $X_i$  ist die Standardabweichung des Startwerts gegeben durch:

$$\sigma_{X_i^s} = \max(1/N, U_0^s/100) \quad (36)$$

Mit dieser Abschätzung sowie der Unsicherheit der Parameter aus 4.1 wird nun die Unsicherheit der gesamten Simulation berechnet. Für den Stufen-Ansatz wird diese mit Hilfe des Monte-Carlo-Verfahrens aus 2.8.1 ermittelt. Hierfür ist eine Stichprobengröße von 100 Simulationen zu verwenden. Für den konstanten Ansatz wird das Verfahren aus 2.8 eingesetzt. Dabei ist aufgrund des enormen Rechenaufwands die Zeit der Simulation auf 100 Schritte diskretisiert. Nun wird exemplarisch in Abbildung 20 die relative Unsicherheit des Kompartimentes der Infizierten im Landkreis Göttingen in Abhängigkeit der Zeit dargestellt. Diese lässt sich mit dem Quotienten  $\frac{\sigma_I}{I}$  berechnen, wobei  $I$  der Anteil der Infizierten im Landkreis und  $\sigma_I$  dessen Unsicherheit ist.

## 5 Zusammenfassung und Diskussion

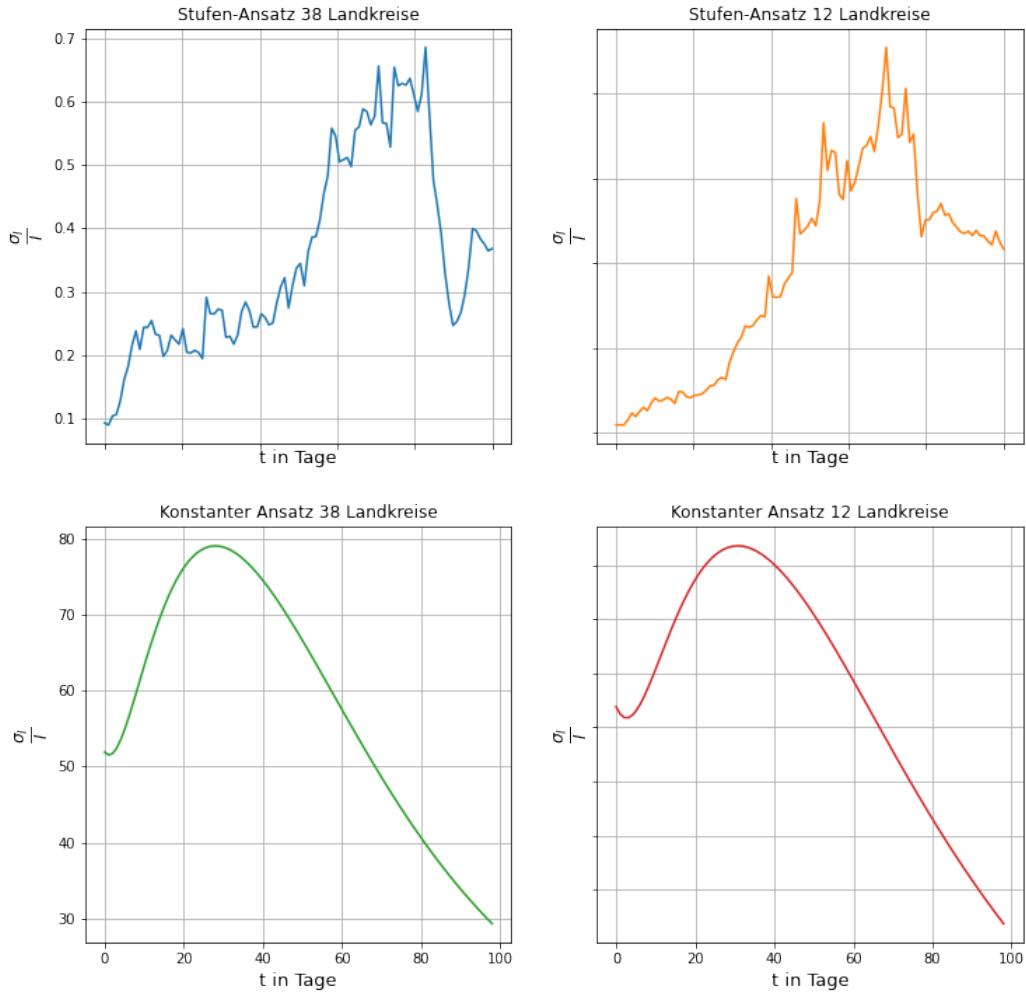
### 5.1 Vereinfachungen

#### 5.1.1 Kompartimente

Das SIR-Modell basiert auf der Einteilung der Populationen in Kompartimente. Da die betrachteten Populationen relativ groß sind, ist es offensichtlich, dass mit steigender Anzahl der Kompartimente die Genauigkeit des Modells der Realität näher kommt. Um eine zufriedenstellende Vergleichbarkeit zu gewährleisten, wurden die Daten des Robert-Koch-Institutes verwendet; diese erlauben jedoch nur die Aufteilung in wenige Kompartimente. Deswegen wurden hier nur 4 Kompartimente pro Landkreis betrachtet.

#### 5.1.2 Dynamik zwischen Zellen

In Sektion 2.2 wurden schon die Vorteile einer Pendlervernetzung gegenüber einer geografischen Vernetzung gezeigt. In Wahrheit werden allerdings weitere Vernetzungen das System beeinflussen: Reisende auf Durchfahrt, Sportveranstaltungen, Besuche, etc. Problematisch wird nur die Datensituation der zusätzlichen Dynamiken. Für



**Abbildung 20:** Relative Unsicherheit  $\sigma_I$  des Kompartiments der Infizierten I im Landkreis Göttingen für den Zeitbereich 2.7.1.

die Anzahl der Pendler gibt es den Pendleratlas [4], die anderen Vernetzungen haben jedoch keine solche Datenbanken. Aus diesem Grund müssten für eine erweiterte Dynamik Abschätzungen zur Größe gemacht werden.

Um das Problem zu erklären, ist es einfacher ein Beispiel heranzuziehen. Hierfür eignen sich Reisende:

1. Die Reisenden, die von Zelle B in Zelle A reisen, können auch weitere Zellen besuchen. Auch auf dem Weg in Zelle A können die Reisenden in anderen Zellen halten.
2. Reisen sind abhängig von den Ferien, damit nicht zeitlich konstant. Da man in den Ferien mehr Reisende erwartet, also eine größere Dynamik, muss dies in der Abschätzung berücksichtigt werden. Zudem nehmen die Ferien nicht einen Zeitraum von 100 Tagen ein, also müssen mindestens zwei Abschätzungen gemacht werden:
3. Die Ferien sind in Deutschland nicht zeitgleich, welches an den Schulferien [23] gesehen werden kann.

## 5.2 Räumliche Auflösung

Die hier verwendete örtliche Auflösung, d.h. diskret verteilte Landkreise mit internen Bevölkerungen, die nur mit Pendlerbewegungen die Infektion in anderen Landkreisen beeinflussen können, kommen auch in der wissenschaftlichen Literatur vor. Ein solches Beispiel ist *Stochastic Simulations of a Spatial SIR Model* von Camacho, J. et al.,

siehe [10]. In dieser Publikation wird das SIR-Modell auf einem  $10 \times 10$  Gitter betrachtet, wobei nur die benachbarten Gitterpunkte sich gegenseitig beeinflussen können. Camacho erwähnt auch ein räumliches Modell, welches auf einer stetigen Verteilung aufbaut. Ein solches Modell erlaubt es, den räumlichen Austausch mit u.a. Diffusion zu simulieren. Da die Daten des RKI nur diskret sind, und es keine Datenbank mit besserer Auflösung im Ort gibt, kann keine stetige Verteilung mit sinnvollen Vergleichsdaten modelliert werden.

### 5.2.1 Superspreading

Unter Superspreading versteht man die Ausbreitung von Infektionen durch besondere, unvorhersehbare Ereignisse, bei denen sehr wenige Infizierte sehr viele Suszeptible anstecken (sogenannte Superverbreitungsergebnisse). Im Fall der COVID-19-Pandemie ist davon auszugehen, dass viele Infektionen auf solche Ereignisse zurückzuführen sind [24]. Auf der im Projekt betrachteten Systemebene von Landkreisen ist die Modellierung solcher Ereignisse nicht möglich, dazu müssten soziale Systeme auf niedrigeren Skalen mit deutlich präziseren Daten zu Bewegungen und Verhalten betrachtet werden. In dem in diesem Projekt entwickelten Modell ist Superspreading allenfalls implizit in den Systemparametern (insbesondere nach der Optimierung) enthalten, die das Auftreten solcher Ereignisse bereits im Mittel annehmen. In Einzelfällen, nämlich bei Superverbreitungsergebnissen mit außergewöhnlich vielen Infizierten, kann das Netzwerkmodell auf der Ebene von Landkreisen keine Vorhersagen treffen.

## 5.3 Bewertung der Qualität des Modells

Der in Abschnitt 4.3.1 ermittelte Spearman-Korrelationskoeffizient  $\rho = -0,4934$  und das zugehörige Signifikanzniveau  $p = 0,0017$  legen nahe, dass die mittlere quadratische Abweichung der simulierten von den tatsächlichen Verläufen in einem Knoten mit dem Grad des Knotens negativ korreliert. Die Vermutung, dass Knoten mit hohem Grad eine niedrigere Abweichung aufweisen, weil sie mehr Einwohner besitzen könnten und daher stärker bei der Optimierung berücksichtigt werden, kann ausgeschlossen werden; da Einwohnerzahl und Grad eines Knotens im ausgewählten Netzwerk nicht korrelieren ( $\rho = 0,0446$ ,  $p = 0,7903$ ). Die Einwohnerzahl ist also keine Drittvariable, die für die Korrelation verantwortlich ist. Allerdings gibt es drei Gründe, warum diese Korrelationsprüfung das Modell nicht ohne weiteres positiv bewerten kann:

- Das Signifikanzniveau  $p$  allein sollte nicht dazu verwendet werden, zu folgern, dass ein ganzes Modell gut funktioniert [25].<sup>18</sup>
- Die Existenz möglicher Drittvariablen kann nicht ausgeschlossen werden.
- Von Korrelation kann nicht auf Kausalität geschlossen werden.
- Die Unsicherheit des Modells ist zu groß.

Dass die Parameteroptimierung, wie anhand der mit ihnen durchgeföhrten Simulationen in Abschnitt 4.3.3 dargestellt, für den Landkreis Göttingen in einem Ring aus 37 Landkreisen einen niedrigeren Wert der Bewertungsfunktion liefert, und daher besser ist als die Optimierung für Göttingen in einem Ring aus 11 Landkreisen, könnte zwei Gründe haben:

- Die Optimierung hat sich für den kleineren Ring nicht auf das globale Minimum der Bewertungsfunktion konzentriert und daher auch nicht die optimalen Parameter gefunden. Das ist nicht unwahrscheinlich, da eine eher geringe Dichte an Startwerten gewählt wurde, um die Laufzeit der Optimierung akzeptabel zu halten. Gleichermaßen gilt theoretisch auch für den größeren Ring, dort wurden allerdings offenbar bessere Parameter gefunden.

---

<sup>18</sup>Der Transparenz halber sei hier außerdem erwähnt, dass noch weitere Korrelationsprüfungen mit anderen Definitionen von  $\zeta$  durchgeführt wurden, von denen nicht alle eine signifikante Korrelation aufzeigen. Weil die gewählte Definition von  $\zeta$  jedoch die gängigste ist, und ihre Festlegung daher vertretbar ist, wurde hier kein „p-Hacking“ betrieben.

- Die Optimierung hat tatsächlich das globale Minimum gefunden und das Netzwerk mit einem Ring aus 11 Landkreisen lässt keine bessere Optimierung zu. Ein breiter Rand würde also tatsächlich bessere Ergebnisse liefern. In Anbetracht der ersten Möglichkeit ist das eher unwahrscheinlich, würde aber für den Erfolg der Berücksichtigung der Pendlerdynamik im Kontext eines Netzwerkmodells sprechen.

## 5.4 Bewertung der Optimierung und Unsicherheit der Modelle

Die in Tabelle 2 aufgelisteten optimierten Parameter, insbesondere diejenigen des Stufen-Ansatzes mit 12 Landkreisen, weisen eine große Streuung auf.

Dies kann auf verschiedene Faktoren zurückgeführt werden. Einerseits kann ein zu flaches globales Minimum der Bewertungsfunktion diese Streuung verursachen. Anhand Abb. 7 lässt sich jedoch erkennen, dass ein globales Minimum keine sehr flache Umgebung haben kann. Weiter ist durch genaues Betrachten erkennbar, dass die Bewertungsfunktion viele lokale Minima und kein ausgezeichnetes globales Minimum hat. Möglicherweise ist die Numerik des Optimierungsverfahrens nicht robust genug, um die Minima der Funktion zu finden<sup>19</sup>. Darüber hinaus ist aufgrund der zeitlichen Zerlegung des simulierten Zeitintervalls ein Rauschen in der Bewertungsfunktion zu erwarten. Die Simulation wird für viele Zeitschritte erstellt, jedoch nimmt die Bewertungsfunktion nur 100 Zeitpunkte davon, um diese dann mit den RKI-Daten zu vergleichen. Der Vergleich hängt daher davon ab, ob der Zeitpunkt vor Beginn des Tages oder nach Beginn des Tages gerundet wird.

Um die Streuung der Parameter zu verringern, eignet sich der Ansatz, eine größere Menge an Startwerten zu optimieren. Dies erwies sich jedoch aufgrund des numerischen Aufwands und der damit verbundenen Rechenzeit schwer umsetzbar. Dazu kommt in Betracht, eine andere Bewertungsfunktion, welche klare Minima aufweist, zu verwenden. Vermutlich gehen wegen der Summation über die Zeit und die verschiedenen Landkreise viele Informationen verloren.

Nun wird die relative Unsicherheit  $\sigma_I/I$  in Abbildung 20 betrachtet. Zu erkennen ist, dass diese für beide Simulationen mit dem konstanten Ansatz sehr groß ausfallen. Hier sind die kleinen Werte von  $I$  sowie die große Unsicherheit der Parameter keine zufriedenstellende Erklärung des beobachteten Verlaufs, da dies ebenfalls für den Stufen-Ansatz gilt. Der Verlauf der relativen Unsicherheit des Stufen-Ansatzes ist deutlich kleiner trotz größerer Abweichung der Parameter. Wahrscheinlich ist das verwendete Verfahren zur Berechnung der Unsicherheit des konstanten Ansatzes 2.8 die Ursache der großen Abweichung. Die Diskretisierung der Zeit auf 100 Schritten für Gleichung 25 ist vermutlich eine zu grobe Approximation. Mit zunehmend feinerer Aufteilung des Zeitintervalls ist zu erwarten, dass der Fehler gegen den tatsächlichen Fehler konvergiert. Jedoch ist, wie bereits erwähnt, der Rechenaufwand hier der begrenzende Faktor. Gleichung 25 benötigt Operationen mit schätzungsweise  $\mathcal{O}((4 \cdot k)^{3t})$  mit  $k$  Anzahl der Landkreise und  $t$  Anzahl der Zeitschritte.

## 5.5 Mögliche Erweiterungen für das Modell

Zukünftige Arbeiten an den Modellen, um sie vollständiger zu machen, sind in Vielfalt möglich. Erstens kann die Anzahl der Kompartimente erweitert werden, zum Beispiel für Geimpfte oder nichtansteckende Infizierte. Seit Ende 2020 werden Menschen in Deutschland geimpft [26][27]. Vor allem kann mit weiteren Unterteilungen des Geimpften-Kompartiments zwischen den Impfstoffen differenzieren. Um Impfungen mit zu berücksichtigen, gibt es allerdings auch alternative Ansätze ohne die Einführung weiterer Kompartimente, z.B. mit einem direkten Übergang zwischen  $S$  und  $R$  [28].

Auch die Beschränkung von Kontakten zwischen Menschen, im Sinne von Lockdown oder anderen Maßnahmen, kann simuliert werden. Diese Methoden wurden immer wieder von Regierungen etabliert um die COVID-19 Pandemie zu kontrollieren. Hierfür können mehrere Ansätze gemacht werden, unter anderem Parameter, die sich zu bestimmten Zeiten oder bei Überschreiten von Infektionen/Infektionsraten ändern.

---

<sup>19</sup>Es gibt keinen Beweis dafür, dass das verwendete Optimierungsverfahren immer gegen ein Minimum konvergiert [11].

Darüber hinaus liegt eine offensichtliche Erweiterung in der Erhöhung der betrachteten Anzahl an Landkreisen. In diesem Projekt wurden maximal 38 Landkreise betrachtet, Gründe dafür sind der sehr hohe Aufwand zum Erstellen großer Pendlermatrizen sowie der Speicherbedarf, insbesondere von Infektionsdaten aus dem Datenhub und darüber hinaus die Laufzeit, allen voran die der Optimierung. Mit weiteren Ressourcen an Hardware, Software und Zeit sind diese Erweiterungen denkbar.

# Anhang

## A Genaue Betrachtung der Jacobi-Matrizen

Hier wird die Jacobi-Matrix aus Sektion 2.8 genauer betrachtet und ein analytischer Ausdruck hergeleitet. Hierfür teilen wir die Jacobi-Matrix auf in eine Blockform:

$$\mathbf{J}_x = \begin{pmatrix} \mathbf{J}_{SS} & \mathbf{J}_{SI} & \mathbf{J}_{SR} & \mathbf{J}_{SD} \\ \mathbf{J}_{IS} & \mathbf{J}_{II} & \mathbf{J}_{IR} & \mathbf{J}_{ID} \\ \mathbf{J}_{RS} & \mathbf{J}_{RI} & \mathbf{J}_{RR} & \mathbf{J}_{RD} \\ \mathbf{J}_{DS} & \mathbf{J}_{DI} & \mathbf{J}_{DR} & \mathbf{J}_{DD} \end{pmatrix}. \quad (37)$$

und

$$\mathbf{J}_\theta = \begin{pmatrix} \mathbf{J}_{S\alpha} & \mathbf{J}_{S\beta} & \mathbf{J}_{Sp} \\ \mathbf{J}_{I\alpha} & \mathbf{J}_{I\beta} & \mathbf{J}_{Ip} \\ \mathbf{J}_{R\alpha} & \mathbf{J}_{R\beta} & \mathbf{J}_{Rp} \\ \mathbf{J}_{D\alpha} & \mathbf{J}_{D\beta} & \mathbf{J}_{Dp} \end{pmatrix}. \quad (38)$$

Der erste Index eines Blocks beschreibt, welches Kompartiment und der zweite Index nach welchem Kompartiment beziehungsweise Parameter abgeleitet wird. Zum Beispiel ist  $J_{SR}^{ij} = \frac{\partial}{\partial R_j} \frac{dS_i}{dt}$  und  $J_{S\alpha}^i = \frac{\partial}{\partial \alpha} \frac{dS_i}{dt}$ . Damit ist  $\mathbf{J}_x$  eine quadratische Matrix der Form  $(m, m)$  und  $\mathbf{J}_\theta$  eine  $(m, 3)$ -Matrix, wobei  $m$  die Anzahl der Landkreise ist. Weiter wird jetzt nicht zwischen (12) und (13) differenziert, sondern die Koeffizienten mit der allgemeinen Funktion  $\Xi_{1/2}$  vertauscht. Diese kann dann je nach Modell für die jeweiligen Koeffizienten wieder ausgetauscht werden. Die Differentialgleichungen sind also:

$$\begin{aligned} \frac{dS_i}{dt} &= -\Xi_1 \alpha S_i I_i - \Xi_2 \alpha S_i I_i^{\text{eff}}, \\ \frac{dI_i}{dt} &= \Xi_1 \alpha S_i I_i + \Xi_2 \alpha S_i I_i^{\text{eff}} - \beta I_i, \\ \frac{dR_i}{dt} &= (1-p)\beta I_i, \\ \frac{dD_i}{dt} &= p\beta I_i \end{aligned} \quad (39)$$

$\mathbf{J}_{SS}$

Aus (39) ist zu sehen, dass nur  $S_i$  in  $\frac{dS_i}{dt}$  vorkommt, also ist:

$$J_{SS}^{ij} = \frac{\delta_{ij}}{S_j} \frac{dS_i}{dt}. \quad (40)$$

$\mathbf{J}_{SS}$  ist also eine diagonale Matrix. Die Diagonalelemente können auch geschrieben werden als

$$J_{SS}^{ii} = -\Xi_1 \alpha I_i - \Xi_2 \alpha I_i^{\text{eff}}$$

$\mathbf{J}_{SI}$

Die Matrixeinträge des Blocks  $\mathbf{J}_{SI}$  sind:

$$J_{SI}^{ij} = -\alpha \delta_{ij} \left[ \Xi_1 + \Xi_2 \frac{N_i^{\text{rest}}}{N_i} \frac{N_i^{\text{rest}}}{N_i^{\text{rest}} + \sum_j \mathcal{P}_{[j \rightarrow i]}} \right] S_i \\ - \alpha \Xi_2 \left[ \frac{N_i^{\text{rest}}}{N_i} \frac{\mathcal{P}_{[j \rightarrow i]}}{N_i^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow i]}} + \frac{N_j^{\text{rest}}}{N_i} \frac{\mathcal{P}_{[i \rightarrow j]}}{N_i^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow i]}} + \sum_l \frac{\mathcal{P}_{[i \rightarrow l]}}{N_i} \frac{\mathcal{P}_{[j \rightarrow l]}}{N_l^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow l]}} \right] S_i. \quad (41)$$

$\mathbf{J}_{SR}, \quad \mathbf{J}_{SD}, \quad \mathbf{J}_{IR}, \quad \mathbf{J}_{ID} \quad \mathbf{J}_{RS}, \quad \mathbf{J}_{DS}, \quad \mathbf{J}_{RR}, \quad \mathbf{J}_{DR}, \quad \mathbf{J}_{RD} \quad \& \quad \mathbf{J}_{DD}$

Es ist einfach zu sehen, dass diese Blöcke immer Nullmatrizen sind.

$\mathbf{J}_{IS}$

Wegen der Symmetrie der Differentialgleichung ist

$$\mathbf{J}_{IS} = -\mathbf{J}_{SS} \quad (42)$$

$\mathbf{J}_{II}$

Wieder ist die Symmetrie zu verwenden, nur dass hier nun ein Zusatzterm hinzukommt:

$$\mathbf{J}_{II} = -\mathbf{J}_{SI} - \beta \mathbf{id} \quad (43)$$

$\mathbf{J}_{RI}$

$$J_{RI}^{il} = \delta_{il}(1-p)\beta \quad (44)$$

$\mathbf{J}_{DI}$

$$J_{DI}^{il} = \delta_{il}p\beta \quad (45)$$

$\mathbf{J}_{S\beta}, \quad \mathbf{J}_{Sp}, \quad \mathbf{J}_{Ip}, \quad \mathbf{J}_{R\alpha}, \quad \mathbf{J}_{D\alpha}$

Hier ist ebenfalls ersichtlich, dass die Einträge Null sind.

$\mathbf{J}_{S\alpha}$

$$\mathbf{J}_{S\alpha}^i = -\Xi_1 S_i I_i - \Xi_2 S_i I_i^{\text{eff}} \quad (46)$$

$\mathbf{J}_{I\alpha}$

$$\mathbf{J}_{I\alpha}^i = \Xi_1 S_i I_i + \Xi_2 S_i I_i^{\text{eff}} \quad (47)$$

$\mathbf{J}_{I\beta}$

$$\mathbf{J}_{I\beta}^i = -I_i \quad (48)$$

$$\mathbf{J}_{R\beta}$$

$$\mathbf{J}_{R\beta}^i = (1-p)I_i \quad (49)$$

$$\mathbf{J}_{Rp}$$

$$\mathbf{J}_{R\beta}^i = -p\beta I_i \quad (50)$$

$$\mathbf{J}_{D\beta}$$

$$\mathbf{J}_{R\beta}^i = pI_i \quad (51)$$

$$\mathbf{J}_{Dp}$$

$$\mathbf{J}_{R\beta}^i = \beta I_i \quad (52)$$

## B Beweis von Gleichung 25

Betrachte die induktive Definition  $\mathbf{u}_{j+1} = \mathbf{A}_j \mathbf{u}_j + \mathbf{B}_j \mathbf{v}$ . Wir zeigen mit einem Induktionsbeweis, dass  $\mathbf{u}_n$  durch

$$\mathbf{u}_n = \prod_{i=0}^{n-1} \mathbf{A}_i \mathbf{u}_0 + \sum_{i=0}^{n-1} \prod_{j=i+1}^{n-1} \mathbf{A}_j \mathbf{B}_i \mathbf{v}. \quad (53)$$

gegeben ist. Die Gleichung gilt offenbar für  $n = 0$ . Mit der Annahme, dass die Relation für  $\mathbf{u}_n$  gilt, betrachte  $\mathbf{u}_{n+1}$ :

$$\begin{aligned} \mathbf{u}_{n+1} &= \mathbf{A}_n \mathbf{u}_n + \mathbf{B}_n \mathbf{v} \\ &= \mathbf{A}_n \left( \prod_{i=0}^{n-1} \mathbf{A}_i \mathbf{u}_0 + \sum_{i=0}^{n-1} \prod_{j=i+1}^{n-1} \mathbf{A}_j \mathbf{B}_i \mathbf{v} \right) + \mathbf{B}_n \mathbf{v} \\ &= \prod_{i=0}^n \mathbf{A}_i \mathbf{u}_0 + \left( \mathbf{A}_n \sum_{i=0}^{n-1} \prod_{j=i+1}^{n-1} \mathbf{A}_j \mathbf{B}_i \mathbf{v} + \mathbf{B}_n \right) \mathbf{v} \\ &= \prod_{i=0}^n \mathbf{A}_i \mathbf{u}_0 + \sum_{i=0}^n \prod_{j=i+1}^n \mathbf{A}_j \mathbf{B}_i \mathbf{v}. \end{aligned} \quad (54)$$

Damit gilt diese Beziehung für alle positiven Ganzzahlen  $n$ .

## C Matrix der kleinen Parameter

Aus Gleichung (29) und (30) ist:

$$(\mathbf{DI})_i = \left( \frac{N_i^{\text{rest}}}{N_i} \frac{N_i^{\text{rest}} I_i + \sum_j \mathcal{P}_{[j \rightarrow i]} I_j}{N_i^{\text{rest}} + \sum_j \mathcal{P}_{[j \rightarrow i]}} + \sum_j \frac{\mathcal{P}_{[i \rightarrow j]}}{N_i} \frac{N_j^{\text{rest}} I_j + \sum_k \mathcal{P}_{[k \rightarrow j]} I_k}{N_j^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow j]}} \right). \quad (55)$$

Hieraus können die Diagonalelemente abgelesen werden:

$$\mathbf{D}_{ii} = \frac{N_i^{\text{rest}}}{N_i} \frac{N_i^{\text{rest}}}{N_i^{\text{rest}} + \sum_j \mathcal{P}_{[j \rightarrow i]}}. \quad (56)$$

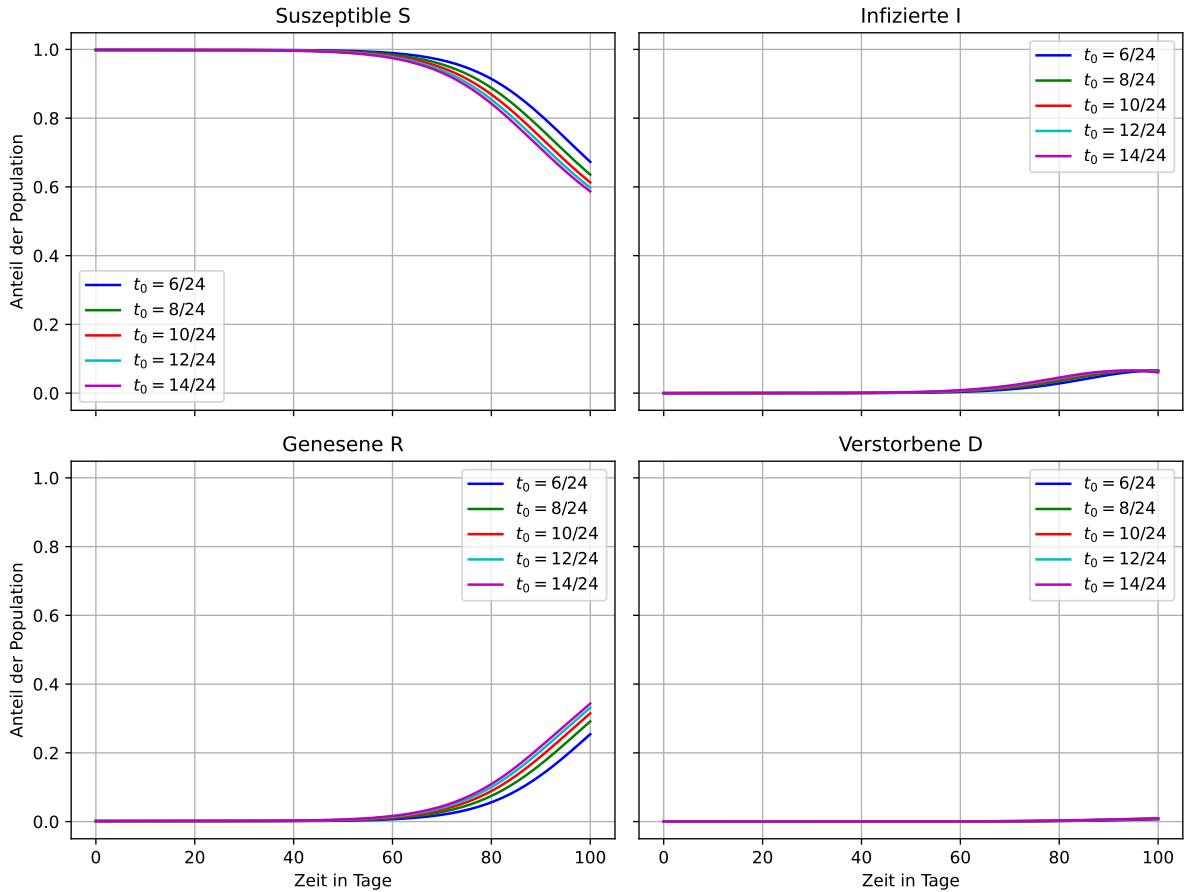
Alle restlichen Elemente sind gegeben durch

$$\mathbf{D}_{ij} = \left[ \frac{N_i^{\text{rest}}}{N_i} \frac{\mathcal{P}_{[j \rightarrow i]}}{N_i^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow i]}} + \frac{N_j^{\text{rest}}}{N_i} \frac{\mathcal{P}_{[i \rightarrow j]}}{N_i^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow i]}} + \sum_l \frac{\mathcal{P}_{[i \rightarrow l]}}{N_i} \frac{\mathcal{P}_{[j \rightarrow l]}}{N_l^{\text{rest}} + \sum_k \mathcal{P}_{[k \rightarrow l]}} \right]. \quad (57)$$

## D Weitere Plots

### D.1 Variation von $t_0$

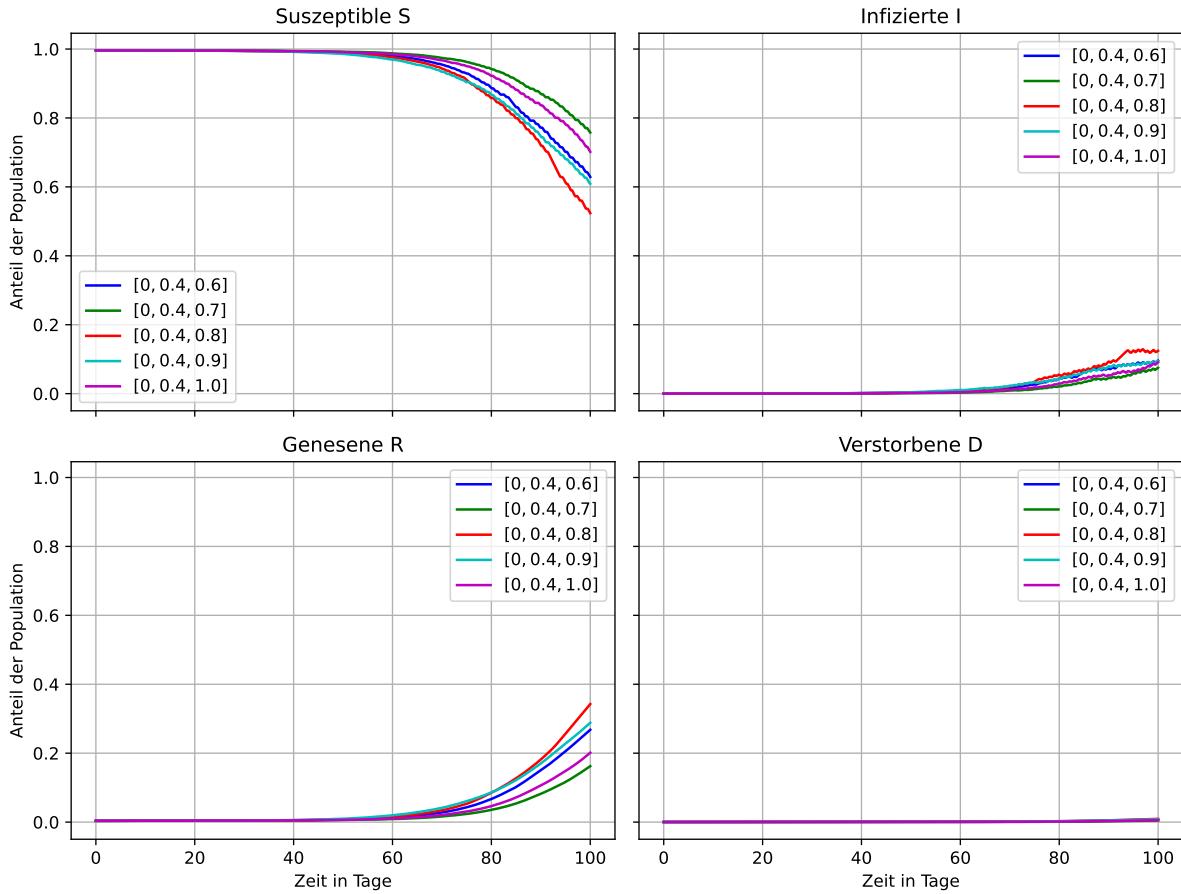
Zeitentwicklung in Knoten 10 für verschiedene  $t_0$  im konstanten Ansatz;  $\alpha = 0.3$ ,  $\beta = 1/5$ ,  $p = 0.0264$



**Abbildung 21:** Plot der numerischen Lösungen der Differentialgleichungen des konstanten Ansatzes für verschiedene  $t_0$  im Landkreis Eichsfeld (Knoten 10).

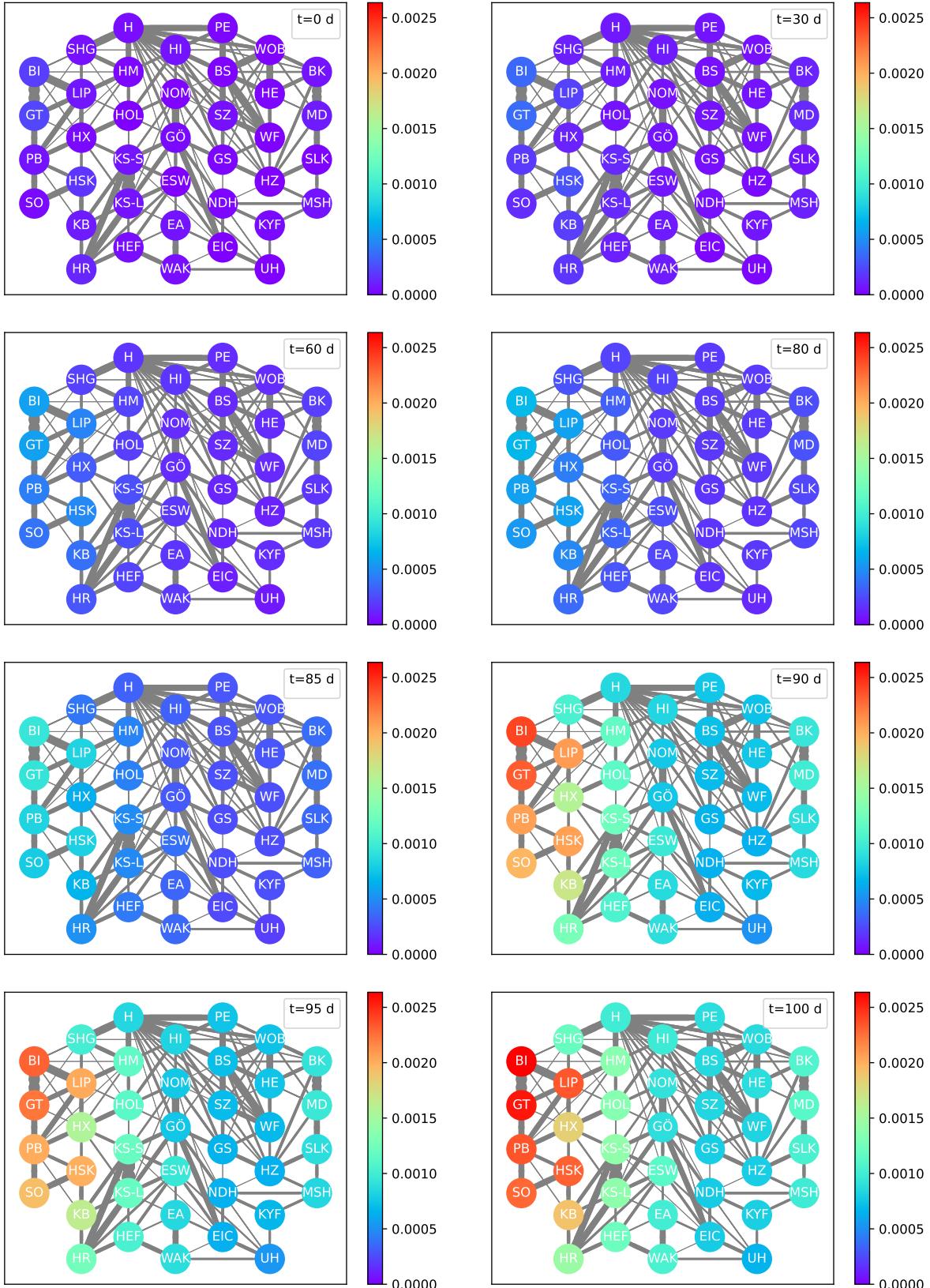
## D.2 Variation der Intervalle

Zeitentwicklung in Knoten 5 für verschiedene  $\pi_1$ ,  $\pi_2$  &  $\pi_3$  im Stufen-Ansatz;  $\alpha = 0.4$ ,  $\beta = 1/4$ ,  $p = 0.0264$



**Abbildung 22:** Plot der numerischen Lösungen der Differentialgleichungen für verschiedene Grenzen der charakteristischen Funktionen  $\chi_{[a,b]}$  im Landkreis Schaumburg (Knoten 13).

### D.3 Netzwerkdigramm des Systems zu verschiedenen Zeiten



**Abbildung 23:** Anteil der aktuell Infizierten in jedem Knoten zu ausgewählten Zeitpunkten dargestellt im Netzwerkdigramm. Bei der zugrundeliegenden Simulation wurde der Stufen-Ansatz gewählt und entsprechend optimiert. Erkennbar ist das Ausbreiten einer Infektionswelle auf dem Netzwerk ausgehend vom Westen der Region. Zu beachten ist, dass die Zeitpunkte nicht äquidistant gewählt wurden, um die schnellere Ausbreitung gegen Ende des betrachteten Zeitraums besser darzustellen.

## E Weitere optimierte Parameter

**Tabelle 3:** Übersicht über die optimalen Parameter  $\alpha$  und  $\beta$  für die untersuchten Pendlermodelle. Optimierte wurden hierfür Simulationen des jeweiligen Modells mit  $N$  Landkreisen, allerdings wurde dabei einzig das Verhalten im Knoten Göttingen verglichen. Das ist der wesentliche Unterschied zu den Parametern, die in Tabelle 2 aufgeführt sind, denn diese sind optimal für die Gesamtheit des Netzwerks.

Modell	N	$\alpha$ in $d^{-1}$	$\beta$ in $d^{-1}$
Stufen-Ansatz	12	$0,3 \pm 2,8$	$0,2 \pm 2,8$
Konstanter Ansatz	12	$0,33 \pm 0,08$	$0,30 \pm 0,08$
Stufen Ansatz	38	$0,49 \pm 0,17$	$0,26 \pm 0,17$
Konstanter Ansatz	38	$0,32 \pm 0,07$	$0,28 \pm 0,07$

## F Abkürzungen

### F.1 Interne Identifikationsnummern und Kfz-Kennzeichen der Landkreise

**Tabelle 4:** Interne Identifikationsnummern und Abkürzungen für Namen der im Projekt verwendeten Landkreise. Die Abkürzungen sind die primären Kfz-Kennzeichen des Landkreises, mit Ausnahme von Kassel, dort wurde für Stadtkreis und Landkreis zur Unterscheidung der Zusatz -S beziehungsweise -L gewählt.

ID	Abkürzung	Name des Landkreises	ID	Abkürzung	Name des Landkreises
0	HI	Hildesheim	19	SO	Soest
1	HOL	Holzminden	20	HSK	Hochsauerlandkreis
2	GS	Goslar	21	KB	Waldeck-Frankenberg
3	HX	Höxter	22	HR	Schwalm-Eder-Kreis
4	NOM	Northeim	23	HEF	Hersfeld-Rotenburg
5	GÖ	Göttingen	24	WAK	Wartburgkreis
6	HZ	Harz	25	EA	Eisenach
7	KS-L	Kassel (Land)	26	UH	Unstrut-Hainich-Kreis
8	KS-S	Kassel (Stadt)	27	KYF	Kyffhäuserkreis
9	ESW	Werra-Meißner-Kreis	28	MSH	Mansfeld-Südharz
10	EIC	Eichsfeld	29	SLK	Salzlandkreis
11	NDH	Nordhausen	30	MD	Magdeburg
12	H	Region Hannover	31	BK	Börde
13	SHG	Schaumburg	32	HE	Helmstedt
14	HM	Hameln-Pyrmont	33	WOB	Wolfsburg
15	LIP	Lippe	34	WF	Wolfenbüttel
16	BI	Bielefeld	35	BS	Braunschweig
17	GT	Gütersloh	36	SZ	Salzgitter
18	PB	Paderborn	37	PE	Peine

# Quellen

## Datenverzeichnis

- [4] Pendleratlas. Statistik der Bundesagentur. (Datenstand Juni 2020). <https://statistik.arbeitsagentur.de/DE/Navigation/Statistiken/Interaktive-Angebote/Pendleratlas/Pendleratlas-Nav.html>. Aufgerufen am 19.06.2021.
- [5] RKI AdmUnit, RKI COVID19, RKI History und RKI Corona Landkreise im COVID-19-Datenhub. <https://npgeo-corona-npgeo-de.hub.arcgis.com/search?collection=Dataset&q=RKI>. Aufgerufen am 19.06.2021, Datensätze zu verschiedenen Zeitpunkten abgerufen.
- [9] COVID-19: Fallzahlen in Deutschland und weltweit. [https://www.rki.de/DE/Content/InfAZ/N/Neuartiges\\_Coronavirus/Fallzahlen.html](https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Fallzahlen.html). Aufgerufen am 02.05.2021.
- [18] Ihor Nesteruk. „Simulations and predictions of COVID-19 pandemic with the use of SIR model“. In: (2020).
- [19] Ashutosh Simha, R. Venkatesha Prasad und Sujay Narayana. A simple Stochastic SIR model for COVID 19 Infection Dynamics for Karnataka: Learning from Europe. 2020. arXiv: 2003.11920 [q-bio.PE].
- [20] Alexis Akira Toda. Susceptible-Infected-Recovered (SIR) Dynamics of COVID-19 and Economic Impact. 2020. arXiv: 2003.11221 [q-bio.PE].
- [21] Priyanka und Vicky Verma. Study of lockdown/testing mitigation strategies on stochastic SIR model and its comparison with South Korea, Germany and New York data. 2020. arXiv: 2006.14373 [physics.soc-ph].

## Literaturverzeichnis

- [1] W. O. Kermack und A. G. McKendrick. „A Contribution to the Mathematical Theory of Epidemics“. In: *Proc. Roy. Soc. Lond. A* 115, 700-721 (1927).
- [2] Eric W. Weisstein. „Kermack-McKendrick Model“. In: *From MathWorld—A Wolfram Web Resource*. <https://mathworld.wolfram.com/Kermack-McKendrickModel.html> () .
- [3] Alain Barrat, Marc Barthélemy und Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge: Cambridge University Press, 2008. DOI: 10.1017/CBO9780511791383.
- [6] Hans Petter Langtangen. *A Primer on Scientific Programming with Python*. Fifth edition. Berlin Heidelberg: Springer-Verlag, 2016.
- [7] Mark Newman. *Networks*. 2nd ed. Oxford: Oxford University Press, 2018.
- [8] Aric A. Hagberg, Daniel A. Schult und Pieter J. Swart. „Exploring Network Structure, Dynamics, and Function using NetworkX“. In: *Proceedings of the 7th Python in Science Conference*. Hrsg. von Gaël Varoquaux, Travis Vaught und Jarrod Millman. Pasadena, CA USA, 2008, S. 11–15.
- [10] Judit Camacho, Fernando Carreón, Derik Castillo-Guajardo u. a. „Stochastic Simulations of a Saptial SIR Model“. In: *Biometrics Unit Technical Reports; Number BU-1366-M* (1996).
- [11] `scipy.optimize.fmin`. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.fmin.html>. Aufgerufen am 11.10.2021.
- [12] *Bund-Länder-Beschluss: Neue Maßnahmen zur Eindämmung der Pandemie*. <https://www.bundesregierung.de/breg-de/suche/bund-laender-beschluss-1804936>. Aufgerufen am 10.10.2021.
- [13] Taboga, Marco (2010): "Linear combinations of normal random variables"; in: *Lectures on probability and statistics*. <https://www.statlect.com/probability-distributions/normal-distribution-linear-combinations>. Aufgerufen am 11.10.2021.
- [14] Tom Lindstrøm und Klara Hveberg. *Flervariabel Analyse med Lineær Algebra*. 2. utgave. Oslo: Gyldendal Norsk Forlag AS, 2016.

- [15] K.F. Riley, M.P. Hobson und S.J. Bence. *Mathematical Methods for Physics and Engineering*. Cambridge: Cambridge University Press, 2006.
- [16] *scipy.integrate.solve\_ivp*. [https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.solve\\_ivp.html](https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.solve_ivp.html).
- [17] Guangbao Guo. „Financial Data Analysis Based on Parallel Statistical Computing“. Diss. Apr. 2012. DOI: 10.13140/RG.2.1.4846.1527.
- [22] *Statistik : Der Weg zur Datenanalyse*. ger. Fünfte, verbesserte Auflage. Springer-Lehrbuch | Business and Economics. Online-Ressource, v.: digital. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2004. ISBN: 9783540350378 | 978-3-540-35037-8. DOI: <https://doi.org/10.1007/3-540-35037-3>. URL: <https://doi.org/10.1007/3-540-35037-3>.
- [23] *Sommerferien 2021, 2022*. [https://www.schulferien.org/Schulferien\\_nach\\_Ferien/Sommerferien/sommerferien.html](https://www.schulferien.org/Schulferien_nach_Ferien/Sommerferien/sommerferien.html). Aufgerufen am 09.08.2021.
- [24] Felix Wong und James J. Collins. „Evidence that coronavirus superspreading is fat-tailed“. In: *Proceedings of the National Academy of Sciences* 117.47 (2020), S. 29416–29418. ISSN: 0027-8424. DOI: 10.1073/pnas.2018490117. eprint: <https://www.pnas.org/content/117/47/29416.full.pdf>. URL: <https://www.pnas.org/content/117/47/29416>.
- [25] Ronald L. Wasserstein und Nicole A. Lazar. „The ASA Statement on p-Values: Context, Process, and Purpose“. In: *The American Statistician* 70.2 (2016), S. 129–133. DOI: 10.1080/00031305.2016.1154108. eprint: <https://doi.org/10.1080/00031305.2016.1154108>. URL: <https://doi.org/10.1080/00031305.2016.1154108>.
- [26] *Impfdashboard.de*. <https://impfdashboard.de/>. Aufgerufen am 09.10.2021.
- [27] *Aktuelle Informationen zur COVID-19-Impfung*. <https://www.bundesgesundheitsministerium.de/coronavirus/faq-covid-19-impfung.html>. Aufgerufen am 09.10.2021.
- [28] Randy L. Caga-anan, Michelle N. Raza, Grace Shelda G. Labrador u. a. „Effect of Vaccination to COVID-19 Disease Progression“. In: 2021.