Modellierung der Covid-19 Fälle

Zusammenfassung

Ziel der Modellierung ist die Prognose der Covid-19 Infektionen für die kommenden Wochen sowie die Berechnung der dadurch benötigten Betten und Intensivbetten in den Krankenhäusern. Ausschlaggebend für die Entwicklung der Fälle ist die Reproduktionsrate, welche angibt, wie viele weitere Personen eine Person im Schnitt ansteckt. Diese wird zunächst aus den Daten geschätzt und anschließend mit Hilfe eines Zeitreihenmodells für jeden Landkreis vorhergesagt. Unsicherheiten werden durch ein Prognoseintervall berücksichtigt. Außerdem werden verschiedenen Szenarien bezüglich der politischen Maßnahmen betrachtet. Um räumliche Abhängigkeiten zu berücksichtigen und die Instabilität bei manchen Landkreisen zu beseitigen, folgt eine Glättung der Raten mit Hilfe eines generalisierten additiven Regressionsmodells. Die sich ergebenden Raten werden verwendet, um die Infektionszahlen für die kommenden Wochen zu bestimmen. Dazu wird das epidemiologische SEIR-Modell herangezogen.

1 Schätzung der Reproduktionsrate

Im ersten Schritt wird die Reproduktionsrate aus den gemeldeten Infektionszahlen des RKI geschätzt. Die Rate gibt an, wie viele Personen eine Person im Schnitt ansteckt. Zunächst muss dafür bestimmt werden, wie viele Personen sich wann infizieren. Dazu werden folgende Annahmen getroffen:

- Es wird eine Dunkelziffer angenommen. Die Daten zur Schätzung der Dunkelziffer stammen aus Seropositivitätsdaten von "The Economist". Es wird dabei angenommen, dass die Dunkelziffer aufgrund einer steigenden Anzahl von durchgeführten Covid-Tests über die Zeit abnimmt.
- Der Meldeverzug beträgt 7 Tage. Es vergehen also 7 Tage von der Infektion bis zur Meldung durch das RKI.

Nachdem die tatsächlichen Infektionszahlen dadurch berechnet wurden, kann die Reproduktionsrate an einem Tag wie folgt bestimmt werden:

$$\mathrm{R}_t = \frac{\sum_{i=t}^{t+(s-1)} NewInf_i}{\sum_{i=t-s}^{t-1} NewInf_i}$$

wobei folgendes gilt:

- t: Index für einen Tag
- R_t: Reproduktionsrate am Tag t
- $NewInf_t$: Neue Infektionen am Tag t

- s: Hyperparameter für die "Fenstergröße"

Die Formel entspricht der Schätzfunktion des RKI. Der darin enthaltene Parameter s wird mit 7 festgelegt, um eine gewisse Stabilität zu erzeugen und um die Rate um die wöchentliche Saisonalität zu bereinigen (Am Wochenende werden weniger Fälle gemeldet als unter der Woche). Wenn max.date der letzte Tag ist, an dem die gemeldeten Infektionszahlen verfügbar sind, kann die Reproduktionsrate dadurch bis zum Tag max.date.rate = max.date - 7 (Meldeverzug) - 7 (Fenstergröße s) geschätzt werden. Ab dem Tag danach sind Forecasts notwendig.
Anschließend wird diese Reproduktionsrate um den Anteil der aktuellen "Susceptibles" (also die Personen, die sich aktuell infizieren können) bereinigt. Die Anzahl der aktuellen Susceptibles wird mit S_t bezeichnet und kann mit Hilfe folgender weiterer Annahmen berechnet werden:

- Dauer der Infektion: 10 Tage (als Mittelwert)
- Dauer der Immunität nach überstandener Infektion (6 Monate)
- Dunkelziffer wie oben beschrieben

Berechnung:

$\mathrm{R}_t^C = R_t/(S_t/Gesamtpopulation)$

Diese korrigierte Reproduktionsrate hängt im Gegensatz zur ursprünglichen Reproduktionsrate nicht mehr davon ab, wie hoch der Anteil der Susceptibles aktuell ist und lässt sich somit auch über verschiedene Zeitpunkte, Landkreise etc. vergleichen. Sie lässt sich als die Anzahl der Personen, die eine Person im Mittel anstecken würde, wenn die gesamte Population "Susceptible" wäre, interpretieren.
Die Schätzung der (korrigierten) Reproduktionsrate erfolgt separat für jeden Landkreis, aber auch für Gesamtdeutschland. Da die Rate bei kleinen Infektionszahlen instabil ist, wird sie mit dem Wert 3 gedeckelt.

2 Forecast der Reproduktionsrate

Als Nächstes erfolgt der Forecast der Reproduktionsrate ab dem Tag max.date.rate + 1. Dafür wird ein Zeitreihenmodell (ARMA) herangezogen, da Abhängigkeiten der Rate bei aufeinanderfolgenden Zeitpunkten angenommen werden können.
In diesem wird auch ein "Shutdown-Effekt", welcher die politischen Maßnahmen zur Eindämmung des Corona-Virus repräsentiert, berücksicktigt. Die entsprechende Variable ShD ist zu Beginn der Covid-19 Pandemie 1 und sinkt im Zeitraum 02.03.2020 bis 23.03.2020 (Zeitraum, in dem die ersten Shutdown Maßnahmen ergriffen wurden) linear bis auf 0. Ab dem 20.04.2020 steigt sie zunächst wieder schrittweise linear an. Dies hängt von den Maßnahmen der Politik ab und kann sich auch von Bundesland zu Bundesland unterscheiden. Ab September sinkt

die Variable aufgrund strengerer Maßnahmen wieder. Es ergibt sich dadurch ein ARMAX Modell.

Die Modellierung gliedert sich in zwei Schritte:

1. Modelltraining: Das ARMAX Modell wird auf Deutschland-Ebene trainiert. Für die Rate wird eine Lognormalverteilung angenommen. Dadurch beschränkt sich der Wertebereich auf positive Werte.
   $Log(R_t^C) = c + \sum_i^p a_i Log(R_{t-i}^C) + \sum_j^q b_j \epsilon_{t-j} + \beta * ShD_t + \epsilon_t$
   p und q werden mit Hilfe eines Auto-ARMA datengestützt automatisch bestimmt.

2. Anwendung des Modells für jeden Landkreis: Das Modell wird nur auf Deutschland-Ebene trainiert, aber auf jedem Landkreis separat ausgewertet. Unsicherheiten werden durch ein 90% Prognoseintervall berücksichtigt. Zudem werden neben dem Basisszenarie (Fortsetzung der bestehenden Maßnahmen) je nach aktueller Infektionslage und politischer Situation weitere Szenarien modelliert. Diese können folgende enthalten:

   - Zusätzliche Lockerungen ab einem bestimmten Datum
   - Vollständige Aufhebung aller Maßnahmen ab einem bestimmten Datum
   - Verschärfung der Maßnahmen ab einem bestimmten Datum
   - Erneuter Lockdown ab einem bestimmten Datum
   - Maßnahmen wie in einem bestimmten Bundesland werden in Gesamtdeutschland ergriffen

3 Glättung der Reproduktionsrate

Es liegen nun Reproduktionsraten über die Zeit für jeden Landkreis vor. Diese werden mit Hilfe eines generalisierten linearen additiven Regressionsmodells (GAM) geglättet. Dies hat folgende Zwecke:

- Räumliche Glättung: Im Modell werden Latitude und Longitude der Landkreise (Zentroide) als Regressoren verwendet. Dies führt zu einem glatten räumlichen Verlauf. Dies ist in der Hinsicht plausibel, dass Reproduktionsraten benachbarter Landkreise sich tendenziell ähnlicher sind, als Raten entfernter Landkreise.

- Bereinigung von Landkreisen mit geringen Infektionszahlen: Die Schätzungen für Landkreise mit wenigen Daten und/oder kleinen Infektionszahlen ist sehr instabil. Auch hier hat die Glättung einen dementsprechend positiven Effekt.

- Zusätzliche zeitliche Glättung

Zusätzlich zu den Variablen Zeit und Ort (Lat, Lon) werden die Bevölkerungsdichte BevD und das Bundesland BL als Regressoren berücksichtigt. Die Modellformel lautet:

$$Log(R_t^C) = f_1(Lat, Lon) + f_{2,BL}(t) + beta_{BL} + beta_{BevD} * BevD + \epsilon_t$$

Das Modell wird im Gegensatz zu Schritt 2 über alle Landkreise hinweg trainiert und ausgewertet. Es wird dazu lediglich das Basis-Szenario herangezogen. Die geglätteten Raten der weiteren Szenarien und die Raten der Prognoseintervallgrenzen erhält man proportional zur nicht geglätteten Variante.

4 Modellierung der Infektionszahlen

Um aus den Reproduktionsraten die Infektionszahlen zu bestimmen, wird das epidemiologische SEIR-Modell verwendet. Dieses umfasst vier Stages:

- S - Susceptible: Personen, die sich infizieren können
- E - Exposed: Infizierte, aber noch nicht infektiöse Fälle
- I - Infectious: Infektiöse Fälle
- R - Removed: Personen, die sich aktuell nicht infizieren können (aufgrund von Immunität oder Tod)

Dieses Modell wird um zusätzliche Stages erweitert bzw. in weitere Stages unterteilt:

- Infektiöse Fälle:
  - $I_{Presymp}$: Infektiös bevor man möglicherweise Symptome entwickelt
  - I_{Asymp}: Infektiös ohne Symptome
  - $I_{SympPreHosp}$: Infektiös mit Symptomen (bevor man möglicherweise schwer erkrankt und eine Hospitalisierung notwendig wird)
  - I_{Symp}: Infektiös mit Symptomen (keine schwere Erkrankung/Hospitalisierung)
  - $I_{SympHospPreICU}$: Infektiös und in stationärer Behandlung (bevor möglicherweise eine intensivmedizinische Behandlung erfolgt)
  - $I_{SympHosp}$: Infektiös und in stationärer Behandlung (keine intensivmedizinische Behandlung nötig)
  - $I_{SympHospIcu}$: Infektiös und in intensivmedizinischer Behandlung
- Removed Fälle:
  - $R_{Infection}$: immun dank überstandener Infektion
  - R_{Dead}: tot

Die Personen werden im Modell durch die Stages "geschoben". Die Parameter des Modells umfassen dabei die Verweildauern in den Stages und die Übergangswahrscheinlichkeiten zwischen den Stages.

- Verweildauern:
  - E: 3 Tage
  - $I_{Presymp}$: 2 Tage
  - I_{Asymp}: 7 Tage
  - $I_{SympPreHosp}$: 1 Tag
  - I_{Symp}: 10 Tage
  - $I_{SympHospPreICU}$: 1 Tag
  - $I_{SympHosp}$: 7 Tage
  - $I_{SympHospIcu}$: 10 Tage
  - $R_{Infection}$: 180 Tage
  - S und R_{Dead} (absorbierender Zustand) haben keine feste Dauer
- Übergangswahrscheinlichkeiten:
  - Übergangswahrscheinlichkeiten bezüglich Schwere der Erkrankung: Schätzung aus historischen Daten (abhängig von der Altersgruppe)
  - Übergangswahrscheinlichkeit bezüglich Infektion (S - E): abhängig von der Reproduktionsrate und der Anzahl der aktuell infektiösen Fälle

Folgende Abbildung zeigt die Stages inklusive der möglichen Übergänge:

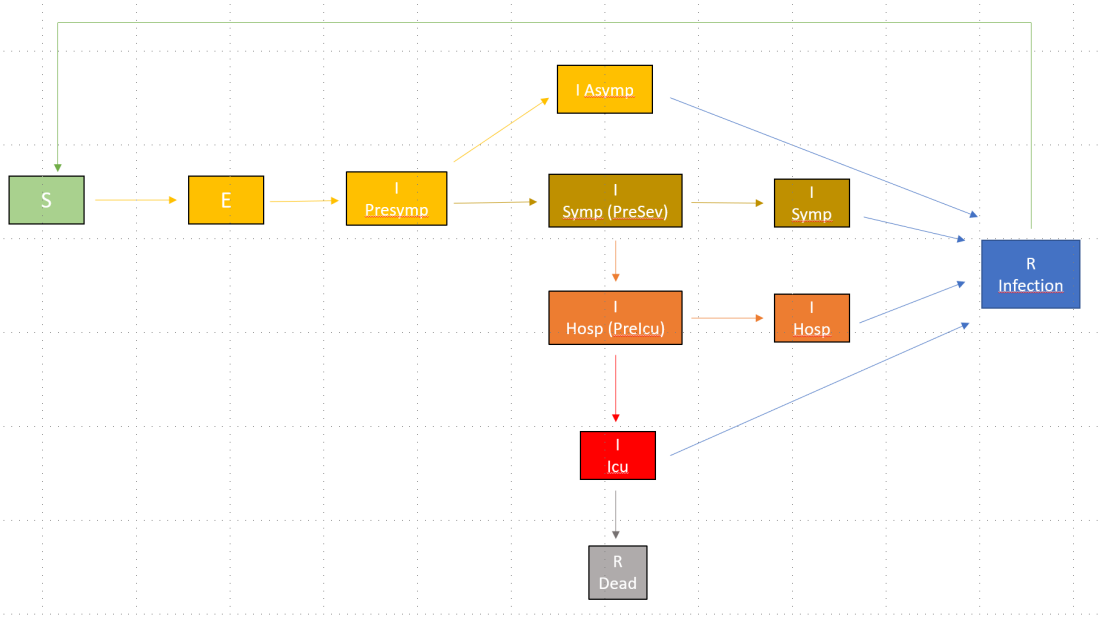

Figure 1: Erweitertes SEIR Modell

Auf Basis der Parameter, der historischen RKI Zahlen und der Dunkelziffer lassen sich die Personenzahlen in jedem Stage an jedem Tag bestimmen (Bis zum Tag max.date - 8): Mit Hilfe der vorhergesagten Reproduktionsraten (Aus den Schritten 1 - 3) lassen sich auch entsprechend Werte für die Zukunft kalkulieren.

Dieses erweiterte SEIR-Modell wird

- für jeden Landkreis
- für jedes Szenario
- für den Erwartungswert und die Prognoseintervallgrenzen der Reproduktionswerte

berechnet.