Erläuterung der Schätzung der zeitlich variierenden Reproduktionszahl R

Robert Koch-Institut

15. Mai 2020

Zusammenfassung

Die Berechnung des 7-Tages R-Werts wird methodisch und anhand einer Implementation in der Statistik-Software R erläutert. Dieses Dokument richtet sich an das epidemiologische Fachpublikum.

Hintergrund

In an der Heiden und Hamouda (2020) wurde das Verfahren des RKI zur Bestimmung der zeitlich variierenden Reproduktionszahl, des sogenannten R-Werts, beschrieben. Das Verfahren besteht aus drei Schritten:

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

Die Schritte 1 und 2 führen zu einer geschätzten epidemischen Kurve, welche Einschätzungen zum Trend und Umfang des Ausbruchs anhand von absoluten Fallzahlen erlaubt. Schritt 3, die Berechnung des zeitlich variierenden R-Werts, entspricht einer Trendanalyse dieser epidemischen Kurve. Der R-Wert ist eine epidemiologische Kennzahl, um die Dynamik des Ausbruchsgeschehens zu beschreiben.

In dem vorliegenden Dokument soll auf die R-Wert Bestimmung (Berechnungen in Schritt 3) genauer eingegangen werden. Speziell soll die Berechnung des sogenannten **7-Tages R-Werts** mathematisch erläutert werden. Dieser unterscheidet sich von dem bereits berichteten sensitiveren **R-Wert** durch eine erweiterte Glättung, die die statistische Schätzunsicherheit verringert. Somit ist der 7-Tages R-Wert in seiner zeitlichen Dynamik stabiler und reagiert weniger sensitiv auf die momentane Einschätzung der epidemischen Kurve durch das Nowcasting.

Erläuterung der R-Schätzung

Mit Hilfe der vom RKI bereitgestellten Excel-Tabelle der aktuellen Schätzung durch Imputation und Nowcast lässt sich die R-Wert Berechnung des Schritt 3 des Verfahrens

nachrechnen und visualisieren. Siehe hierzu
https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/Nowcasting.html

Das RKI verwendet zur Schätzung der zeitlich variierenden Reproduktionszahl R aufgrund des geschätzten Verlaufs der Anzahl von Neuerkrankungen E_t die folgende Formel nach Cori et al. (2013):

$$R_t = \frac{E_t}{\Lambda_t},$$

wobei $\Lambda_t = \sum_{s=1}^{t} E_{t-s}\, w_s$ und $w_1, w_2, \ldots$ die diskrete Wahrscheinlichkeits-Verteilung des seriellen Intervalls mit Träger 1,2, … bezeichnet, d.h. für $i = 1,2, \ldots$ gilt $0 \leq w_i \leq 1$ und die Summe über alle w_i ist 1. In der Formel wird also angenommen, dass die neuen Erkrankungsfälle E_t zum Zeitpunkt t sich jeweils bei einem Anteil w_t der früher erkrankten Personen E_{t-s} angesteckt haben. Rein technisch handelt es sich bei R_t um eine sog. *instantaneous reproduction number* [Cori et al. (2013)], welche rückwärts-schauend in der Zeit definiert ist.

Unter der Annahme einer konstanten Generationszeit und eines konstanten seriellen Intervalls von 4 Tagen ergibt sich daraus zunächst die Formel

$$R_t = \frac{E_t}{E_{t-4}},$$

weil bei dieser Annahme die Verteilung des seriellen Intervalls gleich $w_i \equiv I(i = 4)$ ist, wobei $I()$ die Indikatorfunktion angibt. Das heißt R_t gibt an, wieviele Personen eine Person mit Erkrankungsbeginn zum Zeitpunkt $t - 4$ im Durchschnitt ansteckt. Die angesteckten Personen werden dann zum Zeitpunkt t beobachtet.

Die obige Schätzung von R verhält sich allerdings typischerweise relativ unruhig und wird normalerweise nicht verwendet - vgl. z.B. Cori et al. (2013), S. 1506. Statt R_t nur für einen Zeitpunkt t zu berechnen, kann R_t auch über ein Intervall von τ Tagen berechnet werden. Cori et al. zeigen, dass dafür die folgende Formel genutzt werden kann:

$$R_{t,\tau} = \frac{\sum_{s=t-\tau+1}^{t} E_s}{\sum_{s=t-\tau+1}^{t} \Lambda_s},$$

Beträgt das serielle Intervall 4 Tage, dann vereinfacht sich diese Formel zu

$$R_{t,\tau} = \frac{\sum_{s=t-\tau+1}^{t} E_s}{\sum_{s=t-\tau+1}^{t} E_{s-4}}.$$

Diese Formel kann äquivalent auch als Quotient zweier gleitendender Mittel über τ Tage der E_s-Werte beschrieben werden, also als

$$R_{t,\tau} = \frac{\frac{1}{\tau}\sum_{s=t-\tau+1}^{t} E_s}{\frac{1}{\tau}\sum_{s=t-\tau+1}^{t} E_{s-4}} \equiv \frac{\overline{E}_t^{\tau}}{\overline{E}_{t-4}^{\tau}},$$

wobei $\overline{E}_t^{\tau} = \frac{1}{\tau}\sum_{s=t-\tau+1}^{t} E_s$ den gleitende Mittelwert der Anzahl von Neuerkrankungen über τ-Tage bezeichnet. Der bisherige vom RKI berechnete (sensitive) R-Wert ergibt sich für $\tau = 4$, also als

$$R_{t,4} = \frac{\overline{E}_t^4}{\overline{E}_{t-4}^4} = \frac{\sum_{s=t-3}^{t} E_s}{\sum_{s=t-3}^{t} E_{s-4}}.$$

Der stabilere 7-Tages R-Wert ergibt sich für ein Glättungsintervall von $\tau = 7$ Tagen, also als

$$R_{t,7} = \frac{\overline{E}_t^7}{\overline{E}_{t-4}^7} = \frac{\sum_{s=t-6}^{t} E_s}{\sum_{s=t-6}^{t} E_{s-4}}.$$

Der an Tag u berichtete R-Wert bezieht sich auf das Nowcasting bis zum Zeitpunkt $t = u - 4$ und damit in der sensitiven Variante auf Neuerkrankungen im Zeitraum $u - 7, \cdots, u - 4$ und in der stabileren Variante auf Neuerkrankungen im Zeitraum $u - 10, \cdots, u - 4$. Beide Variantes des R-Wertes beziehen sich also auf Intervalle und werden nur zu Darstellungszwecken einem einzelnen Tag zugeordnet.

Bezieht man noch die Inkubationszeit von 4 bis 6 Tagen mit ein, so beschreibt die am Tag u berichtete Reproduktionszahl R_t in der sensitiven Variante die Neuinfektionen im Zeitraum $u - 13, \cdots, u - 8$ und in der stabileren Variante die Neuinfektionen im Zeitraum $u - 16, \cdots, u - 8$. Dieses letztere Intervall reicht im Vergleich länger zurück und lässt sich eher mit dem Intervall $u - 14, \cdots, u - 9$ als mit dem Intervall $u - 13, \cdots, u - 8$ vergleichen. Um den R-Wert und den 7-Tage R-Wert besser vergleichen zu können, wird daher der 7-Tage R-Wert um einen Tag zurück datiert. Siehe dazu auch Abbildung 2.

Als Beispiel: Im RKI-Lagebericht am 15. Mai 2020 bezieht sich der angegebene sensitive R-Wert auf das Infektionsgeschehen im Zeitraum vom 02. Mai 2020 bis 07. Mai 2020. Der stabile R-Wert auf den Zeitraum 29. April 2020 bis 07. Mai 2020.

Die 95% Prädiktionsintervalle für diese beiden R-Werte für einen spezifischen Tag t ergeben sich durch Anwendung der obigen Formel für die R-Werte der 200 Realisationen des Nowcastings. Sie können nicht auf einfache Weise aus den in der Excel-Tabelle angegeben Prädiktionsintervallen für die Neuerkrankungen erzeugt werden. Wichtig ist, dass es sich bei beiden R-Werten um eine statistische Schätzung handelt, weshalb die Prädiktionsintervalle wichtige Informationen zur Sicherheit der Schätzung enthalten.

Implementation in R

```
# Lade neuesten Nowcast von der RKI Webseite
daten_file <- str_c("Nowcasting_Zahlen-",Sys.Date(),".xlsx")
if (!file.exists(daten_file)) {
  file_url <-
"https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/Nowcasting_Zahlen.xlsx?__blob=publicationFile"
 download.file(url=file_url,destfile= daten_file, mode="wb")
}
```

```r
# Lese Excel-File
data <- xlsx::read.xlsx(file = daten_file, sheetName = "Nowcast_R", encoding
= "UTF-8")
data <- data[,1:13]
# Umbennung der Spalten Namen zu kürzeren Variabelnamen
names(data) <- c("Datum", "NeuErkr", "lb_NeuErkr", "ub_NeuErkr",
"NeuErkr_ma4", "lb_NeuErkr_ma4", "ub_NeuErkr_ma4", "R", "lb_R", "ub_R",
"R_7Tage", "lb_R_7Tage", "ub_R_7Tage")

# R-Wert Berechnung bei einem seriellen Intervall von 4 Tagen
R_Wert <- rep(NA, nrow(data))
for (t in 8:nrow(data)) {
  R_Wert[t] <- sum(data$NeuErkr[t-0:3]) / sum(data$NeuErkr[t-4:7])
}
data <- data %>% dplyr::mutate(R_Wert = round(R_Wert, digits = 2))

#Vergleiche mit den R-Werten in der Excel-Tabelle
data %>% select(Datum, R, R_Wert) %>% tail()
```

```
##         Datum    R R_Wert
## 66 2020-05-06 1.02   1.02
## 67 2020-05-07 1.04   1.04
## 68 2020-05-08 0.97   0.97
## 69 2020-05-09 0.88   0.88
## 70 2020-05-10 0.77   0.77
## 71 2020-05-11 0.80   0.80
```

Unterschiede in der dritten Nachkommastelle des nachgerechneten R-Wertes entstehen durch etwas unterschiedliche Verwendung von Rundungen auf ganze Zahlen.

```r
# Plot
ggplot(data=data, aes(x=Datum)) +
  geom_ribbon(aes(ymin = lb_R, ymax = ub_R), stat="identity",
fill="steelblue")+
  geom_line(aes(y = R), stat="identity", fill="steelblue")+
  theme_minimal() +
   labs(title = "",
       x = "",
       y = "Reproduktionszahl R")  +
  scale_x_date(date_breaks = "2 days", labels =
scales::date_format("%d.%m.")) +
  scale_y_continuous(labels = function(x) format(x, big.mark = ".",
decimal.mark = ",", scientific = FALSE)) +
  theme(axis.text.x  = element_text(angle=90, vjust=0))
```

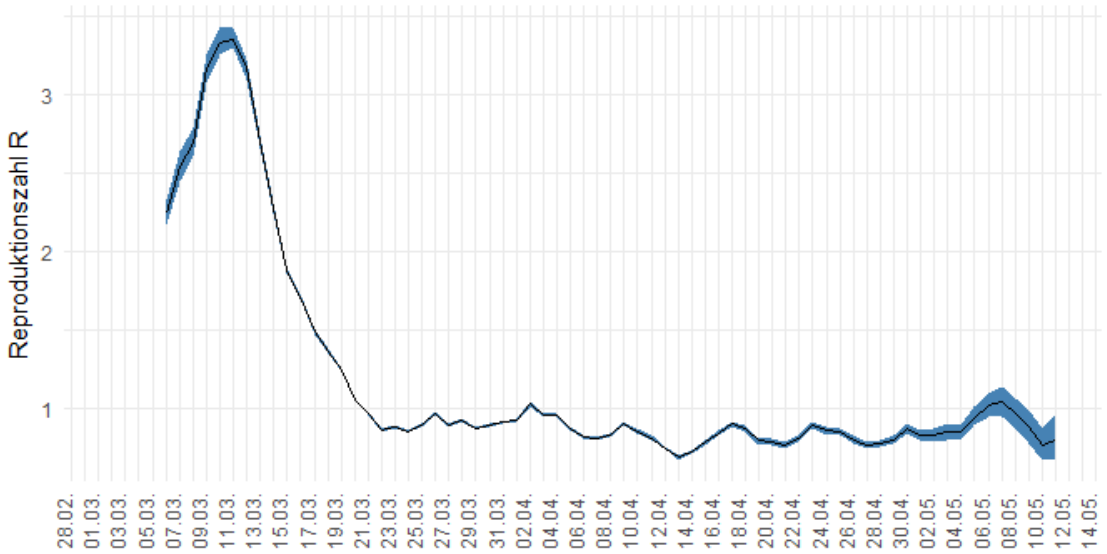

Figure 1: Geschätzte Reproduktionszahl im Verlauf der COVID-19 Epidemie in Deutschland.

Der 7-Tage R-Wert lässt sich auf ähnliche Art wie der bereits bekannte R-Wert bestimmen:

```
#Berechnung des 7-Tage R-Werts
R7_Wert <- rep(NA, nrow(data))
for (t in 11:nrow(data)) {
  R7_Wert[t-1] <- sum(data$NeuErkr[t-0:6]) / sum(data$NeuErkr[t-4:10])
}
data <- data %>% dplyr::mutate(R7_Wert = round(R7_Wert, digits = 2))
#Vergleiche mit den R-Werten in der Excel-Tabelle
data %>% select(Datum, R_7Tage, R7_Wert) %>% tail()

##         Datum R_7Tage R7_Wert
## 66 2020-05-06    0.92    0.92
## 67 2020-05-07    0.94    0.94
## 68 2020-05-08    0.93    0.93
## 69 2020-05-09    0.89    0.89
## 70 2020-05-10    0.90    0.90
## 71 2020-05-11      NA      NA
```

In der folgenden Grafik werden der R-Wert sowie der 7-Tages R-Wert abgebildet.

```
# Plot
ggplot(data=data, aes(x=Datum, y = R, color="R")) +
  geom_ribbon(aes(ymin = lb_R, ymax = ub_R, color=NULL),  fill="steelblue") +
  geom_ribbon(aes(ymin = lb_R_7Tage, ymax = ub_R_7Tage, color=NULL),
fill="orange") +
  geom_line(aes(y = R, color="R")) +
  geom_line(aes(y = R_7Tage, color="R_7Tage"), size = 1) +
    theme_minimal() +
   labs(title = "",
```

```
       x = "",
       y = "Reproduktionszahl R")  +
  scale_x_date(date_breaks = "2 days", labels =
scales::date_format("%d.%m.")) +
  scale_y_continuous(labels = function(x) format(x, big.mark = ".",
decimal.mark = ",", scientific = FALSE)) +
scale_color_manual(name="Methode:",     values=c("darkblue","orangered")) +
  guides(color=guide_legend(override.aes=list(fill=NA))) +
  theme(axis.text.x  = element_text(angle=90, vjust=0)) +
  theme(legend.position="bottom")
```

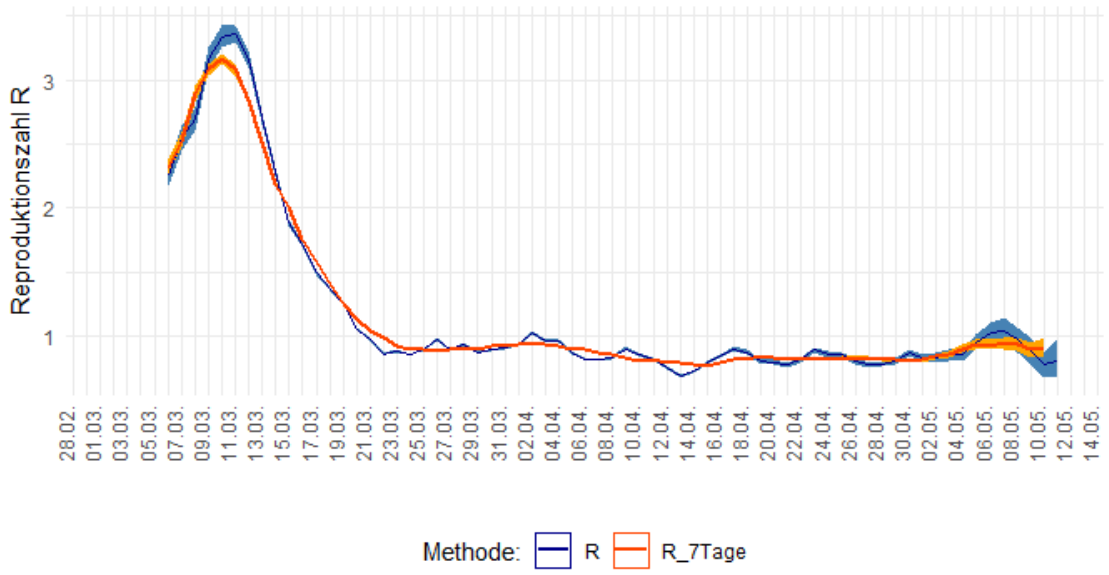

Figure 2: Geschätzte Reproduktionszahl im Verlauf der COVID-19 Epidemie in Deutschland, Vergleich der sensitiven und stabilen Variante.

Diskussion

Sowohl der R-Wert als auch der 7-Tages R-Wert bauen auf der gleichen statistischen Vorgehensweise zur Bestimmung der epidemischen Kurve auf. Das 7-Tages-R stellt dabei eine etwas stärker geglättete Version des R-Werts dar, die Wochentagseffekte in der Schätzung der Anzahl von Neuerkrankungen ausgleicht. Er entspricht einem gewichteten Durchschnitt aus 4 benachbarten R-Werten.

Die gewählte methodische Vorgehensweise zur Berechnung der R-Werte erlaubt es weitere Entwicklungen, wie z.B. die Berücksichtigung einer Verteilung des seriellen Intervalls und die Schätzunsicherheit einer solchen Verteilung, im methodischen Rahmen von Cori et al. (2013) und dem zugehörigen R-Paket EpiEstim zu behandeln.

Literatur

- an der Heiden, M, Hamouda, O, “Schätzung der aktuellen Entwicklung der SARS-CoV-2-Epidemie in Deutschland - Nowcasting”, Epid Bull 2020;17:10–16, https://doi.org/10.25646/6692.4
- Cori A, Ferguson NM, Fraser C, and Cauchemez S, “A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics”, American journal of epidemiology 178(9), 1505-1512, https://doi.org/10.1093/aje/kwt133