Explanation of the estimate of the time-varying

Reproduction number R

Robert Koch Institute

May 15, 2020

Summary

The calculation of the 7-day R-value is explained methodically and on the basis of an implementation in the statistical software R. This document is intended for the epidemiological specialist audience.

background

In an der Heiden and Hamouda (2020), the RKI's method for determining the time-varying reproductive number, the so-called R value, was described. The process consists of three steps:

1. Multiple imputation of missing information about the onset of COVID-19 disease
   Cases under a missing-at-random assumption
2nd Correction of the number of new cases for the diagnosis, reporting and
   Delayed transmission using the nowcasting procedure
3rd Calculation of the time-varying number of reproductions assuming a generation time of 4 days

Steps 1 and 2 lead to an estimated epidemic curve, which allows estimates of the trend and extent of the outbreak based on absolute number of cases. Step 3, the calculation of the time-varying R value, corresponds to a trend analysis of this epidemic curve. The R-value is an epidemiological key figure to describe the dynamics of the outbreak.

In this document, the R-value determination (calculations in step
3) are discussed in more detail. The calculation of the so-called **7-day R value** be explained mathematically. This differs from the more sensitive one already reported **R value** through an advanced smoothing that reduces statistical estimation uncertainty. The 7-day R-value is therefore more stable in its temporal dynamics and reacts less sensitively to the current assessment of the epidemic curve by the nowcasting.

Explanation of the R estimate

With the help of those provided by the RKI Excel table The current estimate by imputation and nowcast can be used to calculate the R value of step 3 of the procedure

recalculate and visualize. See also
https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/Nowcast ing.html

The RKI uses R to estimate the time-varying number of reproductions based on the estimated course of the number of new cases •• •• the following formula according to Cori et al. (2013):

$$•• •• = \frac{•• ••}{•• ••,}$$

in which •• •• = ∑ ⋯⋯ = 1 •• •• − •• •• and •• 1, •• 2, ... the discrete probability distribution of the serial interval with carrier 1,2, ... referred to, ie for •• = 1.2, ... applies 0 ≤ •• •• ≤ 1 and the sum over all •• •• is 1. So the formula assumes that the new cases of illness •• •• at the time •• each with a share •• •• of the previously sick people •• •• − •• infected. Technically, it is about •• •• a so-called

instantaneous reproduction number [Cori et al. (2013)], which is defined looking backwards in time.

Assuming a constant generation time and a constant serial interval of 4 days, this results in next the formula

$$•• •• = \frac{•• ••}{•• •• − 4,}$$

because with this assumption the distribution of the serial interval is the same •• •• ≡ •• (•• = 4) is where •• () indicates the indicator function. This means •• •• specifies how many people a person with the onset of illness at the time •• - 4th infected on average. The infected people are then at the time •• observed.

However, the above estimate of R typically behaves relatively uneasily and is normally not used - cf. eg Cori et al. (2013), p. 1506. Instead •• •• just for a time •• can calculate •• •• even over an interval of •• Days are calculated. Cori et al. show that the following Fo sleeves enjoy nt can be:

$$•• ••, •• = \frac{\sum_{⋯⋯ = •• − •• + 1}^{••} •• ••}{\sum_{⋯⋯ = •• − •• + 1}^{••} •• ••},$$

If the serial interval is 4 days, then simplifies si ch this formula too

$$•• ••, •• = \frac{\sum_{⋯⋯ = •• − •• + 1}^{••} •• ••}{\sum_{⋯⋯ = •• − •• + 1}^{••} •• •• − 4}.$$

This formula can also be used as the quotient of two moving averages •• Days of •• •• - Values are described, i.e. as

$$•• ••, •• = \frac{1}{1} \frac{•• \sum_{⋯⋯ = •• − •• + 1}^{••} •• ••}{•• \sum_{⋯⋯ = •• − •• + 1}^{••} •• •• − 4} \equiv \frac{•• ⋯⋯}{•• •• − 4 ••},$$

in which •• ···· = 1 •• ∑ •• •• ···· = •• − •• + 1 the moving average of the number of new cases

••- Designated days. The previous one from the RKI calculated (sensitive) R-value results for
•• = 4, so as

$$•• ••, 4 = \frac{•• ••4}{•• •• − 4\,4 = \sum_{···· = •• − 3}^{···· = •• − 3} •• ••} \cdot \frac{•• ••}{•• •• − 4}$$

The more stable 7-day R-value results For a Smoothing i interval of •• = 7 Days, so as

$$•• ••, 7 = \frac{•• ••7}{•• •• − 4\,7 = \sum_{···· = •• − 6}^{···· = •• − 6} •• ••} \cdot \frac{•• ••}{•• •• − 4}$$

The day •• Reported R-Value refers to the nowcasting up to the point in time
•• = •• - 4 and thus in the sensitive variant on new cases in the period
•• - 7, ⋯, •• - 4 and in the more stable variant on new cases in the period
•• - 10, ⋯, •• - 4. Both variants of the R-value relate to intervals and are assigned to a single day for illustration purposes only. If one also includes the incubation period of 4 to 6 days, this describes the day ••

reported number of reproductions •• •• in the sensitive variant, the new infections in the period
•• - 13, ⋯, •• - 8 and in the more stable variant the new infections in the period
•• - 16, ⋯, •• - 8. In comparison, the latter interval goes back longer and can be more easily compared to the interval •• - 14, ⋯, •• - 9 than with the interval •• - 13, ⋯, •• - 8 to compare. In order to better compare the R-value and the 7-day R-value, the 7-day R-value is dated back one day. See also Figure 2. As an example: In the RKI management report on May 15, 2020, the specified sensitive R value refers to the infection process in the period from May 2, 2020 to May 7, 2020. The stable R value relates to the period April 29, 2020 to May 7, 2020. The 95% prediction intervals for these two R values for a specific day ••

result from the application of the above formula for the R-values of the 200 realizations of the Nowcasting. They cannot be easily generated from the prediction intervals for the new diseases specified in the Excel table. It is important that both R-values are a statistical estimate, which is why the prediction intervals contain important information for the safety of the estimate.

Implementation in R

```
#  Download the latest nowcast from the RKI website
data_file <- str_c ( " Nowcasting_numbers- " , Sys.Date (), ". xlsx " )
if ( ! file.exists ( data_file)) {
    file_url <-
"https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/No wcasting_Zahlen.xlsx? __ blob =
publicationFile"
  download.file ( url = file_url, destfile = data_file, mode = "wb" )}
```

```
#  Read Excel file
data <- xlsx :: read.xlsx ( file = data_file, sheetName = "Nowcast_R" , encoding = "UTF-8" ) data <- data [, 1 : 13 ]


#  Renaming the column names to shorter variable names
names ( data) <- c ( " Date" , "New Err" , "lb_NeuErkr" , "ub_NeuErkr" ,
"NeuErkr_ma4" , "lb_NeuErkr_ma4" , "ub_NeuErkr_ma4" , "R" , "lb_R" , "ub_R" ,
"R_7days" , "lb_R_7days" , "ub_R_7days" )

#  R-value calculation with a serial interval of 4 days
R_value <- rep ( N / A , nrow ( data))
for ( t in 8th : nrow ( data)) {
    R_value [t] <- sum ( data $ NewDr -0 : 3rd ]) / sum ( data $ NewDr -4 : 7 ])} data <- data %>% dplyr :: mutates ( R_value

= round ( R_value, digits = 2nd ))

#Compare with the R values in the Excel spreadsheet
data %>% select ( Date, R, R_value) %>% tail ()

##                  date            R R_value
##   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
```

Differences in the third decimal place of the recalculated R value result from the slightly different use of rounding to whole numbers.

```
#  plot
ggplot ( data = data, aes ( x = Date)) +
    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 = "Reproduction number R" ) +
    scale_x_date ( date_breaks = "2 days" , labels =
scales :: date_format ( "% dm." )) +
    scale_y_continuous ( labels = function ( x) format ( x, big.mark = "." ,
decimal.mark = "," , scientific = FALSE )) +
    theme ( axis.text.x = element_text ( angle = 90 , vjust = 0 ))
```

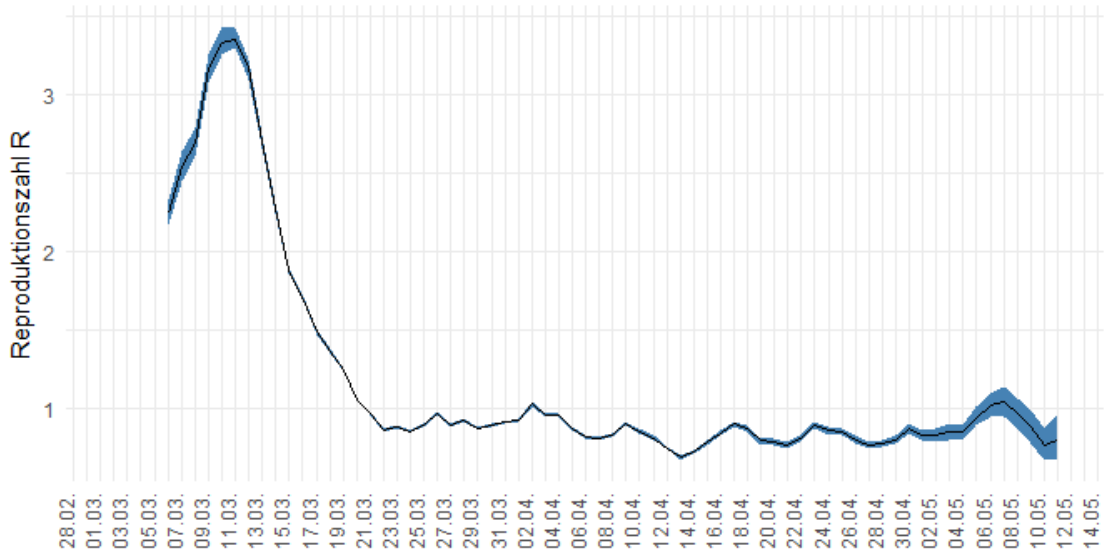

Figure 1: Estimated number of reproductions during the COVID-19 epidemic in Germany.

The 7-day R-value can be determined in a similar way to the already known R-value:

```
#Calculation of the 7-day R value
R7_value <- rep ( N / A , nrow ( data))
for ( t in 11 : nrow ( data)) {
    R7_value [t -1 ] <- sum ( data $ NewDr -0 : 6 ]) / sum ( data $ NewDr -4 : 10th ])} data <- data %>% dplyr :: mutates ( R7_value

= round ( R7_value, digits = 2nd ))
#Compare with the R values in the Excel spreadsheet
data %>% select ( Date, R_7days, R7_value) %>% tail ()

##                 Date R_7 days R7_value
##  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                 N / A          N / A
```

The following graphic shows the R-value and the 7-day R-value.

```
#  plot
ggplot ( data = data, aes ( x = Date, y = R, color = "R" )) +
    geom_ribbon ( aes ( ymin = lb_R, ymax = ub_R, color = ZERO ), fill = "steelblue" ) +
    geom_ribbon ( aes ( ymin = lb_R_7days, ymax = ub_R_7days, color = ZERO ),
fill = "orange" ) +
    geom_line ( aes ( y = R, color = "R" )) +
    geom_line ( aes ( y = R_7days, color = "R_7days" ), size = 1 ) +
       theme_minimal () +
     labs ( title = "" ,
```

```
            x = "" ,
            y = "Reproduction number 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 = "Method:" ,                    values = c ( " darkblue " , "orangered" )) +
    guides ( color = guide_legend ( override.aes = list ( fill = N / A ))) +
    theme ( axis.text.x = element_text ( angle = 90 , vjust = 0 )) +
    theme ( legend.position = "bottom" )
```

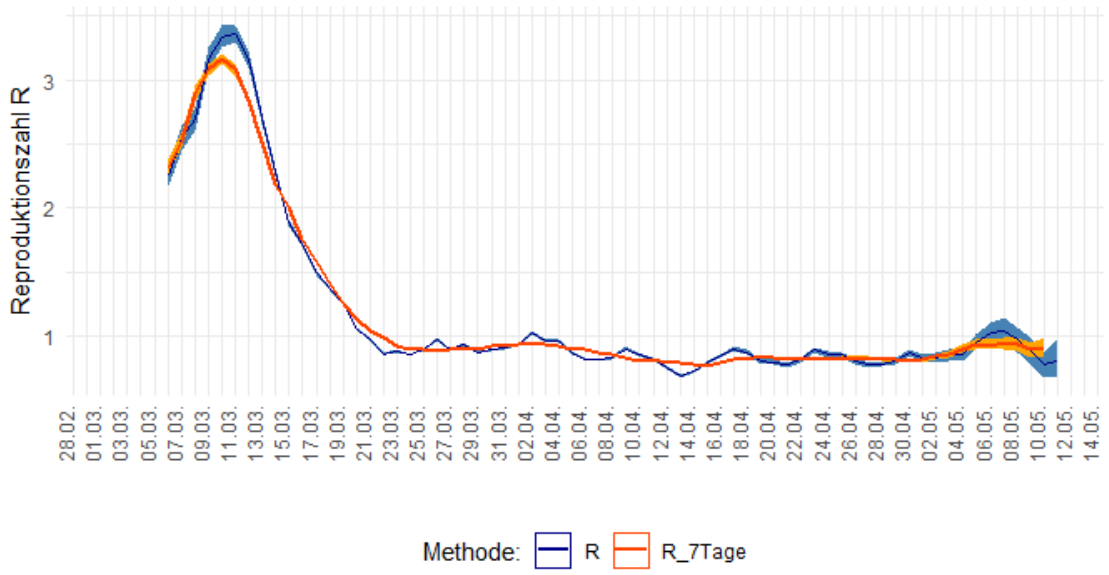

Figure 2: Estimated number of reproductions in the course of the COVID-19 epidemic in Germany, comparison of the sensitive and stable variant.

discussion

Both the R-value and the 7-day R-value are based on the same statistical procedure for determining the epidemic curve. The 7-day R represents a slightly more smoothed version of the R value, which compensates for weekday effects in the estimate of the number of new cases. It corresponds to a weighted average of 4 neighboring R values.

The methodological approach chosen to calculate the R values allows further developments, such as taking into account a distribution of the serial interval and the estimation uncertainty of such a distribution, within the methodological framework of Cori et al. (2013) and the associated R package EpiEstim to treat.

literature

- an der Heiden, M, Hamouda, O, “Estimating the Current Development of the SARS-CoV-2 Epidemic in Germany - 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