Covid-19 in the Netherlands: trends in time dependence of casus data

Han-Kwang Nienhuys (Twitter: @hk_nien)

November 22, 2020

1 Introduction

In the Netherlands, the national health organization RIVM publishes 'casus data' on a daily basis.[1] This is a dataset containing for every (anonymous) individual with a positive test result, among others, a registration date, which can be the date of disease onset (DOO) if known, otherwise the date of positive lab result (DPL) if known, otherwise a date of notification (DON). The DON is the the date the regional health organization (GGD) was notified of the positive lab result. Depending on which lab carried out the test, the DPL and DOO may or may not be registered. For a substantial number of cases, initially the DON is published (usually the day before publication of the casus data), which is then updated a few days later with the actual DPL or DOO date. This may be related to commercial laboratories supplying data on case counts in data formats that need to be imported manually into the national Covid-19 registry. Requests on clarificaton regarding these issues went partially unanswered.[2] In this short report, I will discuss a number of features of this casus data and attempt to correct for these effects.

The code to process the data and the PDF of this report are on Github:
https://github.com/han-kwang/covid19 .

2 Estimating DOO from DPL and DON

I will assume that

$$DOO = DPL - 2 \pm 1 = DON - 3 \pm 1; \tag{1}$$

each casus with a DPL or DON date is mapped to a range of three DPL dates (counting as 1/3 for each of the three). This is reasonable if we assume that those DPL and DON dates are due to commercial test facilities that test individuals quickly after the onset of symptoms, for example, tests arranged by an employer. If those tests are due to, e.g., travellers that need a proof of a negative Covid-19 status, they are likely symptomless and do not have a meaningful DOO.

It is useful to look at cases by DOO because the median incubation time (from infection to disease onset) is about 5 days; the generation interval (for estimating the reproduction number) is taken to be 4 days in the Netherlands. Trends in the cases by DOO can therefore be translated into estimates of the reproduction number.

3 Correction factors for recent DOO data

If we plot the new cases by date of disease onset, for casus data on different publication dates,[3] we typically find data as in Figure 1: the median time for a new case to be registered is around 4 days after the DOO and it takes up to 20 days for the numbers to stabilize. This is problematic if we wish to estimate changes in the reproduction number as soon as possible.

[1] https://data.rivm.nl/covid-19/COVID-19_casus_landelijk.csv
[2] https://twitter.com/ArnoldNiessen/status/1329072350839545858
[3] An archive of casus datafiles is on https://github.com/mzelst/covid-19/tree/master/data-rivm/casus-datasets .

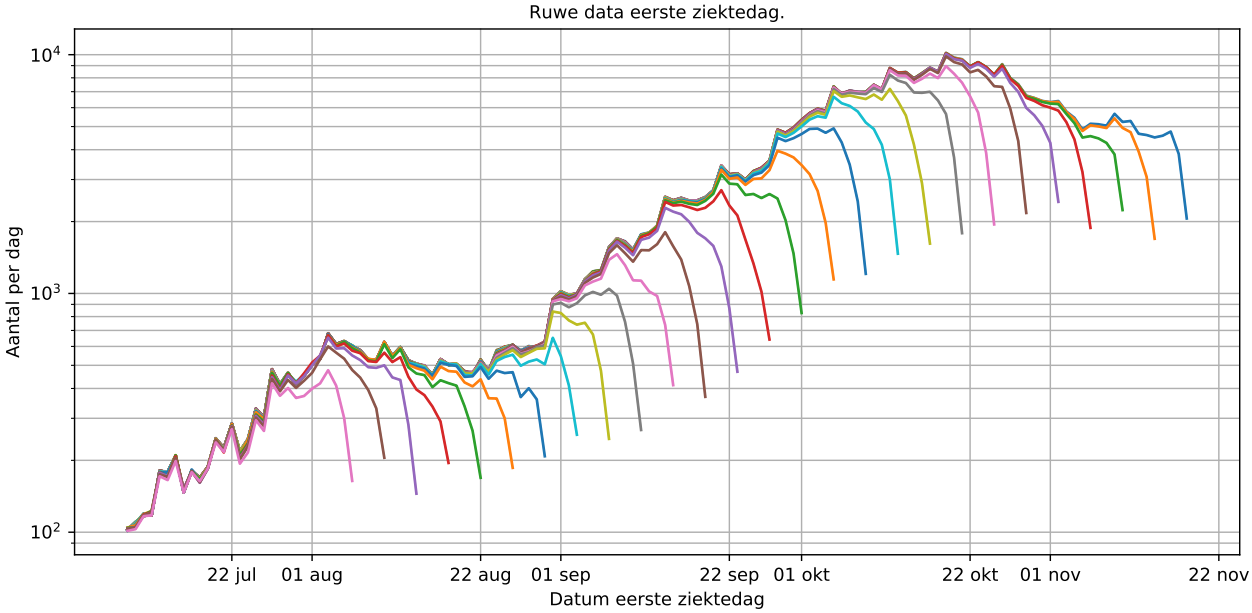

Figure 1: Raw data at different publication dates (publication dates sampled with 4 days interval). Each curve ends at three days prior to the publication date.

We can attempt to correct the recent new cases using a correction factor based on historical data. This was proposed by @bslagter on Twitter.[4] Unfortunately, it is evident from Figure 2 that this correction factor is not constant over time.

The impact of the varying correction factors is illustrated in Figure 3. The correction factors were calibrated for the month of October, In August, there is a systematic overestimate; in early September, a systematic underestimate, and in the beginning of November a systematic overestimate. It is not clear what causes these trends. Possibly, it is related to the ever-changing balance between capacity and demand for tests, which is affected both by scaling up of test facilities and the amount of virus circulaton. The efficiency of data transfer from third-party testing facilities to the regional health organizations may also change. I recommend that the correction factor is based on a longer period so that the error estimate is more realistic.

4 Correction factors based on the day of the week

There are two day-of-the-week (DoW) effects. Firstly, the DoW on which the casus data was published, and secondly, the DoW of disease onset.

The recent-case correction factors have a pronounced DoW effect, as illustrated in Figure 4. If disease onset was on a Sunday, it is typically underreported more than for disease onset on a Monday. Casus data published on Sundays, Mondays, and Tuesdays is typically underreporting more than casus data published on Wednesdays and Thursdays. It appears that the various organizations (GGDs and commercial laboratories) follow a weekly schedule in the processing of new cases.

This publication-DoW correction was applied to the data in Figure 3. The effect of this correction ($\pm 10\%$) is not very noticeable in the data, compared to the huge mismatch ($\pm 50\%$) depending on the month.

The DOO itself also has a pronounced DoW effect, as illustrated in Figure 5. This is quite strange. Although infection event may follow a weekly cycle (due to work, school, and leasure activities), one would not expect the effect to be so pronounced between Sunday (–6%) and Monday (+15%); the distribution in incubation times would not allow such a sharp transition. It is also not likely related to the DOO estimate from DPL and DON: even if DON tends to peak on a certain DoW, the corresponding peak in DOO would be spread out over three days in the present data-processing

[4] https://twitter.com/bslagter/status/1291297749347049474

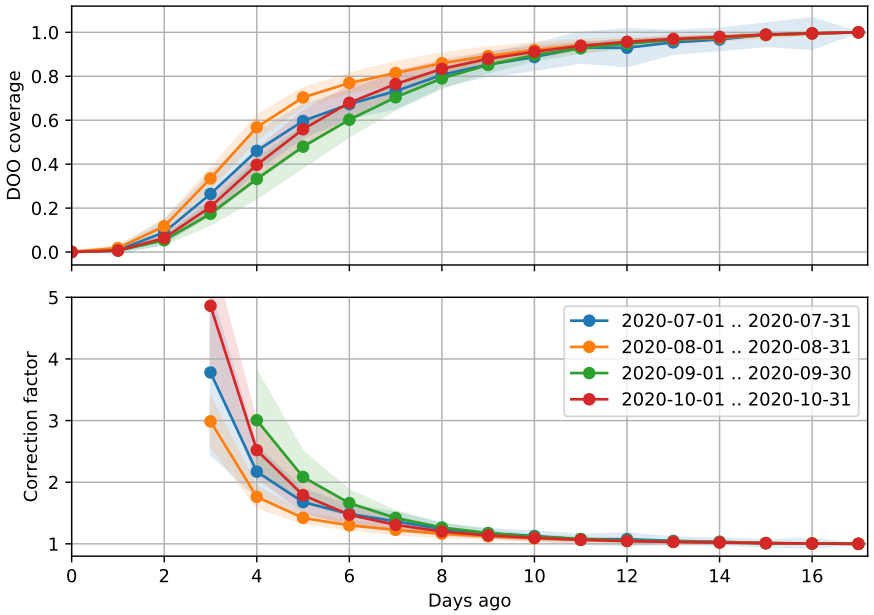

Figure 2: Cumulative coverage (top) and correction factor (bottom) for recent new cases, for different months in 2020. The shaded area indicate one standard deviation.

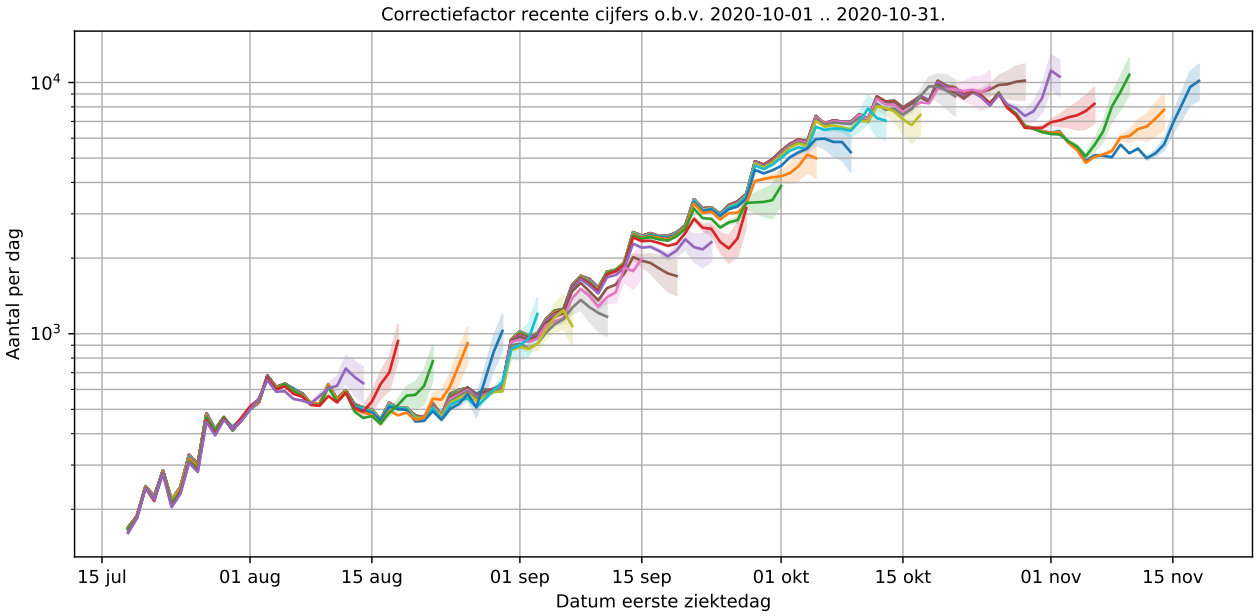

Figure 3: New cases after correction of recent new cases, for different publication dates. Correction factors calibrated for the month of October. Shaded areas indicate one estimated standard error.

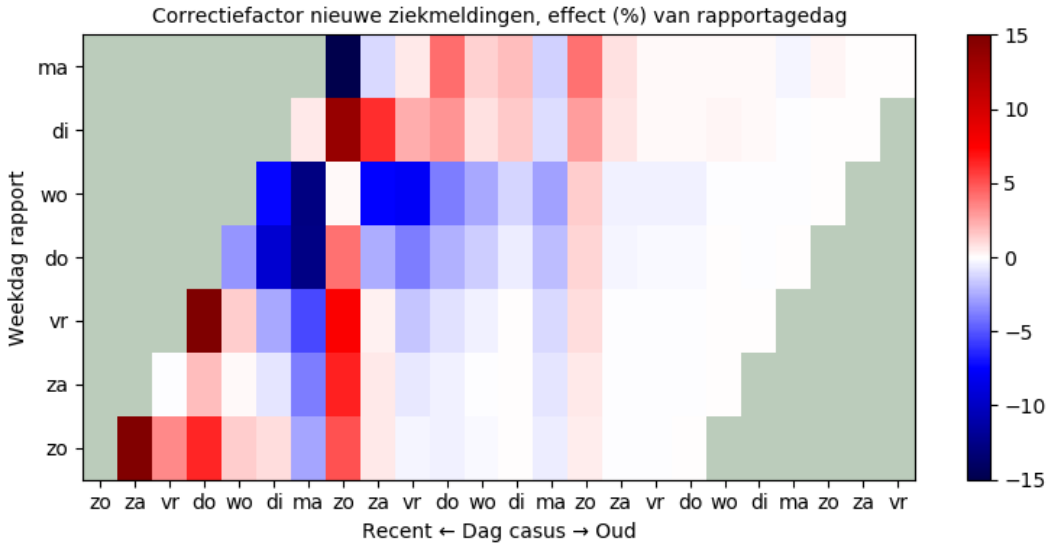

Figure 4: Publication-DoW effect on recent-case correction factors, based on 3 months of data preceding this document.

workflow. The effect is likely to have psychological causes; apparently people are more likely to recognize symptoms on the first weekday of the week than in the weekend. The DOO-DoW effect seems to be decreasing over the period 1 Oct to 5 Nov.

Removing the DoW effect on DOO is useful for identifying changes in the trend and estimating the reproduction number.

5 Recommendation for R estimates

Because changes in the correction factors G^{-1} can only be confirmed about two weeks after the fact, it is best to assume a large error margin on the corrected data. See Figure 6: DOO dates become reliable within ±5% not earlier than 9 days after publication of the casus data. Note that a ±5% per day change in case numbers corresponds to a reproduction number $R_t = 1.0 \pm 0.2$, which is far too inaccurate to draw conclusions upon. IN order to estimate R_t with ±0.05 accuracy, we need the case counts to be accurate to ±1.2%.

A better way may be not to attempt to get accurate correction factors. Instead, keep the correction-factor list fixed and compare data from different publication dates that had the same age, so that the error in the correction factor is always the same. One may still correct for DoW effects. In the formulas (see appendix), the case age is called j. A demonstration is shown in Figure 7. If we assume that the data for $j = 17$ is the truth (only known 17 days after disease onset), then we see that the trend in the data is fairly accurate for $j = 7$ (slope and change of slope), and even better for $j \geq 9$. Unfortunately, it appears that the DOO-DoW effect is also changing and age-dependent. An age-dependency can be incorporated, but this is yet to be done.

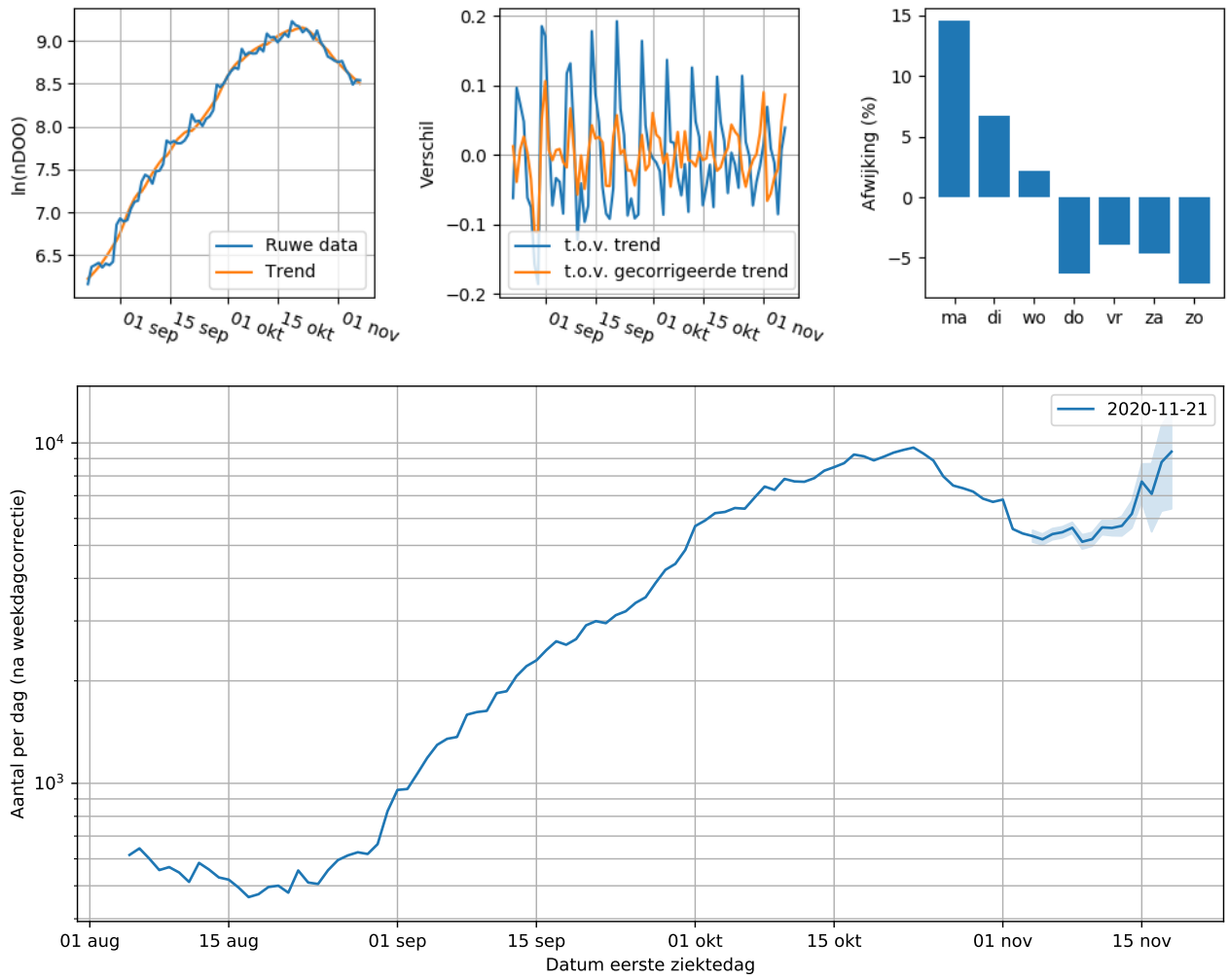

Figure 5: Disease-onset DoW effect on recent-case correction factors, based on 3 months of data preceding this document. Top row, left: new cases (logarithmic; raw and smoothened); center: difference raw-smooth and residual after subtracting DoW effect; right: DoW effect. Bottom: daily case numbers after correction for DOO-DoW effects (much smoother than the highest-values data in Figure 1.

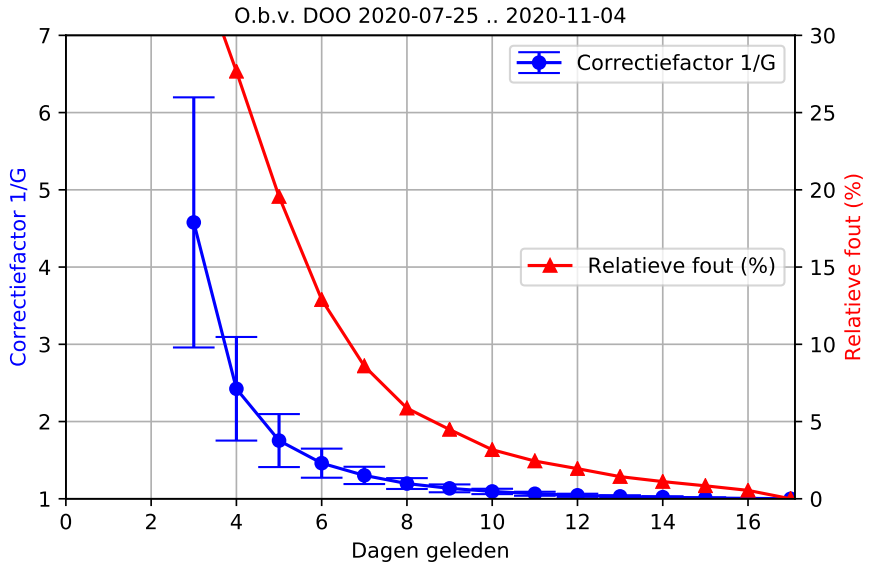

Figure 6: Recommended correction factors G^{-1} for recent casus data.

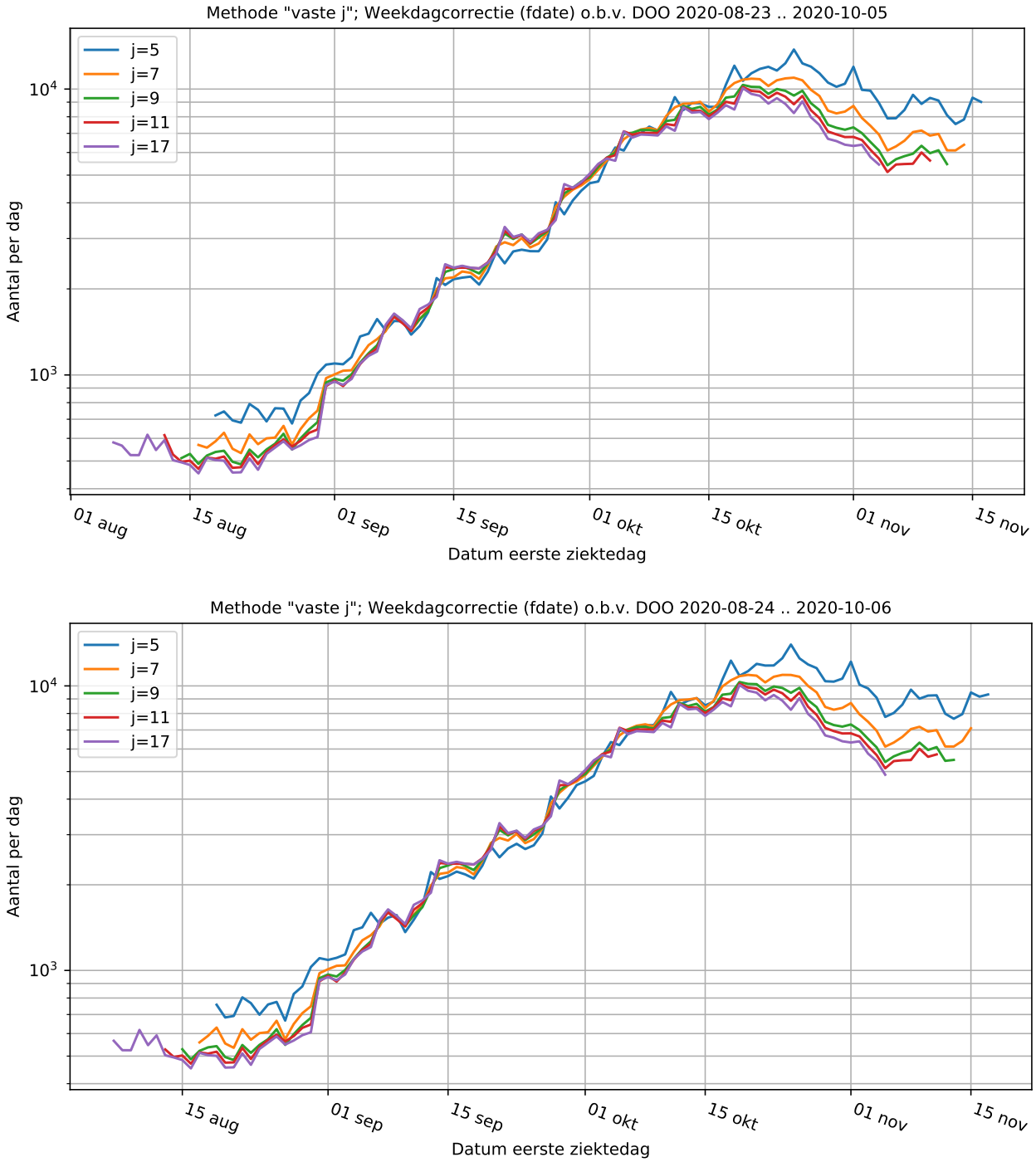

Figure 7: Estimated new cases from the fixed-j method, for various values of j. Top panel: with DOO-DoW and publication-DoW correction; bottom panel: with only publication-DoW correction.

Appendix A Correction factors for recent data

The correction-factor approach is based on the assumption that there is a fixed discrete probability distribution $h[k]$ that a patient with a given DOO date first appears in the daily casus data k days later. If the reported number of cases $r[i, j]$ on reporting date i days ago and DOO $i + j$ days ago and the true number of new cases $i + j$ days ago is given by $f[i + j]$, then

$$r[i, j] = f[i + j] \cdot G[j], \tag{2}$$

where

$$G[j] = \sum_{k=0}^{j} h[k]. \tag{3}$$

We can estimate $f[i + j]$ by assuming that $r[i, j]$ (the daily case data) is correct after $m - 1$ days (I take $m = 18$), i.e., $f[i + j] = r[+j - m + 1, m - 1]$ for $j = m$. In the Python code, all arrays use zero-based indexing. Hence,

$$G[j] = \left\langle \frac{r[i, j]}{r[i + j - m + 1, m - 1]} \right\rangle, \tag{4}$$

where the averaging is over i. The correction factor is then the inverse, $1/G[j]$.

In the fixed-j method, we estimate the number of new cases by DOO as

$$f_j[i] = r[i - j, j] \cdot G^{-1}[j], \tag{5}$$

where j is constant.

Appendix B Disease-onset DoW effect

The DoW effects were separated from the general trend by applying a Savitsky-Golay filter ($n = 15$, order 2) to the logarithm of the daily case numbers. This filter has a good performance in eliminating weekly cycles and other noise, while preserving the trend even when the trend changes from growing to shrinking, such as around 20 October.