# Nowcasting Covid-19 statistics reported with delay: a case-study of Sweden

Adam Altmejd[*] Joacim Rocklöv Jonas Wallin

May 22, 2020

## Abstract

The new corona virus disease - COVID-2019- is rapidly spreading through the world. The availability of unbiased timely statistics of trends in disease events are a key to effective responses. But due to reporting delays, the most recently reported numbers are frequently underestimating of the total number of infections, hospitalizations and deaths creating an illusion of a downward trend. Here we describe a statistical methodology for predicting the actual daily number occurring events a specific day, and its uncertainty, based on the daily reported event frequency. The methodology takes into account the observed distribution pattern of the delay and can be derived from the "removal method", a well-established method in the field of ecology.

Keywords:

[*]SOFI, Stockholm University and Swedish House of Finance, adam@altmejd.se

# 1 Pandemic response demands timely data

The new corona virus pandemic is affecting societies all around the world. As countries are challenged to control and fight back, they are in need of timely, unbiased, data for monitoring trends and making fast and well-informed decisions ("Coronavirus" 2020). Official statistics are usually reported with long delay after thorough verification, but in the midst of a deadly pandemic, real time data is of critical importance for policymakers (Jajosky and Groseclose 2004). The latest data are often not finalized, but change as new statistics are reported. In fact, since these updates are usually in the form of delayed reporting. Since the most recent days then have the least cases, such systematic underreporting gives the dangerous picture of an always improving situation.

Still, these unfinished statistics offer crucial information. If the pandemic is indeed slowing, we should not wait for the data to be finalized before using it. Rather, we argue that actual case counts and deaths should be nowcasted to account for any reporting delay and ensure policymakers are using the most accurate numbers available.

Such predictions provide an additional feature that is perhaps even more important. They explicitly model the uncertainty about these unknown quantities, ensuring that all users of these data have the same view of the current state of the epidemic.

In this paper we describe a statistical methodology for nowcasting the epidemic statistics, such as hospitalizations or deaths, and the degree of uncertainty, based on the daily reported event frequency and the observed distribution pattern of the reporting delay. The prediction model is building on methods developed in ecology, referred to as the "removal methods" or "capture-retain" models (Pollock 1991).

## 1.1 The current situation of COVID-19 in Sweden

Each day at 14:00 the Swedish Public Health Agency holds a press conference where new COVID-19 statistics are presented[1]. Deaths is one of the main indicators to follow for understanding the impact of the pandemic on public health in Sweden, but also the number of new admissions to critical care, to hospitals, and the new confirmed cases are reported. One of the reasons for following these indicators is to enable public health professionals and the public to observe the patterns of the evolving, flattened or suppressed epidemic (Anderson et al. 2020). In relation to policy, it is of further interest to understand if growth rates change, which could indicate a potential response. However, at the daily presentation only a proportion of the number deaths for each of the most recent days is yet known, and this bias causes an artificial, downward, trend in the data.

The death counts suffer from the longest reporting day. In their daily presentation, the Swedish Public Health Agency warns for this by stopping the 7-day moving average trendline 10 days before the latest date. But not only are deaths often reported far further back than 10 days, a bar plot still shows the latest information from the most recent days creating a sense of a downward trend. In fact, this might be the reason why the number of deaths has been underestimated repeatedly. At the peak, deaths were

1. The data is then published on https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/bekraftade-fall-i-sverige/ where we download it every day.

initially believed to level out at around 60 per day, but after all cases had been reported more than two weeks later, the actual level was closer to 100 (Öhman and Gagliano 2020).

# 2 The removal method

We propose to use the removal method, developed in animal management (Pollock 1991), to present an estimate of the actual frequencies at a given day and their uncertainty. The method has a long history dating back at least to the 1930s (Leslie and Davis 1939). However, the first refined mathematical treatment of the method is credited to Moran 1951, more modern derivatives exits today (Matechou et al. 2016). It is a commonly applied method today when analyzing age cohorts in fishery and wildlife management.

The removal method that has three major advantages over simply reporting moving averages:

- it does not relay previous trend in the data,
- we can generate confidence bounds for what is the reasonable range of the uncertainty in the event frequency at a given day,
- the uncertainty in the estimate can be carried over to epidemiological models that uses the estimate as input, and hence give more realistic models.

A classic example where the method proposed to solve this problem has been used is in estimating statistics of trapping a closed population of animals (Pollock 1991). Each day the trapped animals are collected, and kept, and if there is no immigration the number of trapped animals the following days will, on average, decline. This pattern of declining number of trapped animals allows one to draw inference of the underlying population size. Here we replace the animal population with the true number of deaths or cases in a given day. Instead of traps we have the new reports of COVID-19 events. As the number of new reported deaths for a given day declines, we can draw inference on how many actually died that day. In fact, if we assume that the reporting structure is constant over time we can after a while quickly get good estimate of the actual number.

Suppose for example that on day one, 4 individuals are reported dead for that day. On the second day, 10 deaths are recorded for day two. Then, with no further information, it is reasonable to assume that more people died on day two. If the proportion reported on the first day is 3%, the actual number of deaths would be 133 for day one and 333 for day two.

If additionally, 60 deaths are reported during the second day to have happened during day one, and on the third day, only 40 are reported for day two, we now have conflicting information. From the first-day reports it seemed like more people had died during day two, but the second day-reports gave the opposite indication. The model we propose systematically deals with such data, and handles many other sources of systematic variation in reporting delay. In fact, the Swedish reporting lag follows a calendar pattern. The number of events reported during weekends is much smaller. To account for this,

we allow the estimated proportions of daily reported cases to follow a probability distribution taking into consideration what type of day it is.

# 3 Applying the model to COVID-19 in Sweden

We prorpose a Bayesian version of the removal model that assumes anoverdispersed binomial distribution for the the daily observations of deaths in Sweden in COVID-19. We then calculate the posterior distribution and prediction median and 95% uncertainty intervals of the expected deaths from the reported deaths on each specific day. The method and algorithm is thoroughly described in the Supplementary Information.

In Figure 1 we illustrate the similarity and difference between the 7-day moving average and the new Bayesian prediction model with 95% prediction intervals to the reported number of deaths from COVID-19 in Sweden. The model provides estimates of actual deaths considerably above the reported number of deaths and the uncertainty of the estimate.

# 4 Implications and limitations

The model proposed here has much better ability to estimate the trends in surveillance data with reporting delays, such as the daily COVID-19 reports in Sweden. To generate accurate estimates of the actual event frequencies based on these reports is highly relevant and can have large implications for interpretations of the trends and evolution of disease outbreaks. In Sweden, delays are considerable and exhibit a weekday and holiday pattern that need to be accounted for to draw conclusions from the data. The method and algorithm proposed overcome major shortcomings in the daily interpretation and practice analyzing and controlling the novel Corona virus pandemic. It also provides valuable measures of uncertainty around these estimates, showing readers how large the range of possible outcomes can be.

Whenever case statistics are collected from multiple sources and attributed to its actual event date in the middle of a public health emergency, similar reporting delays to the ones in Sweden will necessarily occur. The method described thus has implications and value beyond Sweden, for any situation where nowcasts of disease event frequencies are of relevance to public health.

Nevertheless, the method also has its limitations. As presented, the model assumes that all deaths are reported in the same manner, given there exists many regions in Sweden this is unlikely to be the case. For example, it is easy to see that the Swedish region Västra Götland follows a different reporting structure compared to Stockholm. Building a model for each region separately would most likely give better results and make the assumptions more reasonable. Unfortunately we do not currently have access to the high resolution data required to do so. Another limitation is that the model assumes that the number of new reported deaths for a given day cannot be negative, which is not actually true, due to miscount or misclassification of days. The number of such cases is very small, however, and its removal should not make much difference. The central assumption of the model is that the proportions deaths reported each day is fixed (up to the known covariates). If actual reporting standards change over time and its not explicitly modelled by a covariate, the model will not be able to account for this. But reporting likely becomes faster as the crisis infrastructure improves. One can imagine that after a while the reporting improves, or is changed, if this is not accounted for by a covariate in the model, it will report incorrect

numbers. Of course, there might be unknown variables that we have failed to incorporate, but at the least the model is an improvement from the estimates using moving averages.

# 5 Conclusion

# References


Anderson, R. M., H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth. 2020. "How Will Country-Based Mitigation Measures Influence the Course of the COVID-19 Epidemic?" *The Lancet* 395 (10228): 931–934. https://doi.org/10.1016/S0140-6736(20)30567-5.

Atchadé, Y. F. 2006. "An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift." *Methodology and Computing in Applied Probability* 8 (2): 235–254. https://doi.org/10.1007/s11009-006-8550-0.

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng. 2011. *Handbook of Markov Chain Monte Carlo.* CRC Press.

Jajosky, R. A., and S. L. Groseclose. 2004. "Evaluation of Reporting Timeliness of Public Health Surveillance Systems for Infectious Diseases." *BMC Public Health* 4 (1): 29. https://doi.org/10.1186/1471-2458-4-29.

Leslie, P. H., and D. H. S. Davis. 1939. "An Attempt to Determine the Absolute Number of Rats on a Given Area." *Journal of Animal Ecology* 8 (1): 94–113. https://doi.org/10.2307/1255.

Matechou, E., R. S. McCrea, B. J. T. Morgan, D. J. Nash, and R. A. Griffiths. 2016. "Open Models for Removal Data." *Annals of Applied Statistics* 10 (3): 1572–1589. https://doi.org/10.1214/16-AOAS949.

Møller, J., A. R. Syversveen, and R. P. Waagepetersen. 1998. "Log Gaussian Cox Processes." *Scandinavian Journal of Statistics* 25 (3): 451–482. https://doi.org/10.1111/1467-9469.00115.

Moran, P. A. P. 1951. "A Mathematical Theory of Animal Trapping." *Biometrika* 38 (3/4): 307–311. https://doi.org/10.2307/2332576.

"Coronavirus: Three Things All Governments and Their Science Advisers Must Do Now." 2020. *Nature* 579, no. 7799 (7799): 319–320. https://doi.org/10.1038/d41586-020-00772-4.

Öhman, D., and A. Gagliano. 2020. "Antalet virusdöda har underskattats." *Sveriges Radio: Nyheter (Ekot).* Accessed May 22, 2020. https://sverigesradio.se/sida/artikel.aspx?programid=83&artikel=7459277.

Pollock, K. H. 1991. "Review Papers: Modeling Capture, Recapture, and Removal Statistics for Estimation of Demographic Parameters for Fish and Wildlife Populations: Past, Present, and Future." *Journal of the American Statistical Association* 86 (413): 225–238. https://doi.org/10.1080/01621459.1991.10475022.

Rue, H., and L. Held. 2005. *Gaussian Markov Random Fields: Theory and Applications.* CRC Press.

| Death date \ Reported date | | | | | |
|---|---|---|---|---|---|
| | $r_{11}$ | $r_{12}$ | $\cdots$ | $\cdots$ | $r_{1T}$ |
| | | $r_{22}$ | $\cdots$ | $\cdots$ | $r_{2T}$ |
| | | | $r_{33}$ | $\cdots$ | $r_{3T}$ |
| | | | | $\ddots$ | $\vdots$ |
| | | | | | $r_{TT}$ |

Table 1. The table describes the observations data.

# Appendix

## Model

### Notation

Before presenting the model we describe some notation used through out the article for a $m \times n$ matrix $r$ we use the following broadcasting notation $\mathbf{r}_{k,j:l} = [r_{k,j}, r_{k,j+1}, \ldots, r_{k,l}]$. Further $x|y \sim \pi(.)$ implies that the random variable $x$ if we conditioning on $y$ follows distribution $\pi(.)$. The relevant variables in the model are the following:

| Variable name | Dimension | Description |
|---|---|---|
| $\mathbf{d}$ | $T \times 1$ | $d_i$ is the number of deaths that occurred day $i$. |
| $\mathbf{r}$ | $T \times T$ | $r_{ij}$ is number of death recorded for day $i$ at day $j$. Note that $r_{ij}$ for $i < j$ is not defined. |
| $\mathbf{p}$ | $T \times T$ | $p_{ij}$ is the probability of that a death for day $i$ not yet recorded is recorded at day $j$. Note that $p_{ij}$ for $i < j$ is not defined. |
| $\alpha$ | $K \times 1$ | Latent prior parameter for $\mathbf{p}$ |
| $\beta$ | $K \times 1$ | Latent prior parameter for $\mathbf{p}$ |
| $\alpha^H$ | $2 \times 1$ | parameter for the probability, $\mathbf{p}$ for holiday adjustment. |
| $\beta^H$ | $2 \times 1$ | parameter for the probability, $\mathbf{p}$ for holiday adjustment. |
| $\lambda$ | $T \times 1$ | $\lambda_i$ is the intensity of the expected number of deaths at day $i$. |
| $\sigma^2$ | $1 \times 1$ | Variation of the random walk prior for the log intensity. |

### likelihood

The most complex part of our model is the likelihood, i.e. the density of the observations given the parameters. Here the data consist the daily report of recored deaths for the past days. This can conveniently be represented upper triangular matrix, $\mathbf{r}$, where $r_{i,j}$ represents number of new reported deaths for day $i$ reported at day $j$. This matrix is displayed on the left in Table 1.

We assume that given the true number of deaths at day $i$, $d_i$, that each reported day $j$ the remaining death $d_i - \sum_{k=1}^{j-1} r_{i,k}$ each recored with probability $p_{ij}$, i.e.

$$r_{i,j}|D_i, r_{1,1:j}.p \sim Bin(d_i - \sum_{k=1}^{j-1} r_{i,k}, p_{i,j}).$$

Typically in removal sampling one let $p_{i,j} := p$ however for this data this is clearly not realistic– given delaying in the reporting. Instead we assume that we have $k$ different probabilities. Further to account for overdispertion we assume that each probability instead follows a Beta distribution. The Beta distibution has two parameters $\alpha$ and $\beta$. This resulting the following distribution for the probabilities

$$p_{i,j}|\alpha, \beta, \alpha^H, \beta^H \sim Beta(\alpha_j^H \alpha_{min(j-i,k)}, \beta_j^H \beta_{min(j-i,k)}).$$

Here, we let $H$ denote holidays and weekends and the parameters above are

$$\alpha_j^H = \begin{cases} \alpha_1^H \alpha_2^H & \text{if } \{j \in H\} \cup \{j-1 \in H\}, \\ \alpha_1^H & \text{if } \{j \in H\} \cup \{j-1 \in H^c\}, \\ \alpha_2^H & \text{if } \{j \in H^c\} \cup \{j-1 \in H\}, \\ 1 & \text{else,} \end{cases}$$

and

$$\beta_j^H = \begin{cases} \beta_1^H \beta_2^H & \text{if } \{j \in H\} \cup \{j-1 \in H\}, \\ \beta_1^H & \text{if } \{j \in H\} \cup \{j-1 \in H^c\}, \\ \beta_2^H & \text{if } \{j \in H^c\} \cup \{j-1 \in H\}, \\ 1 & \text{else.} \end{cases}$$

These extra parameters are created to account for the under-reporting that occurs during weekend and holidays.

## Priors

For the $\alpha$ and $\beta$ parameters we use an (improper) uniform prior. For the deaths, $\mathbf{d}$, one could imagine several different prior ideally some sort of epidemiological model. However, here we just assume a log-Gaussian Cox processes Møller et al. 1998 where the Gaussian processes has a intrinsic random walk distribution Rue and Held 2005 i.e.

$$\log(\lambda_i) - \log(\lambda_{i-1}) \sim N(0, \sigma^2),$$
$$d_i|\lambda_i \sim Po(\lambda_i).$$

This model is created to create a temporal smoothing between the reported deaths. For the hyperparameter $\sigma^2$ we impose a inverse Gamma distribution.

### Full model

Putting the likelihood and priors together we get the following hierarchical Bayesian model

$$
\begin{aligned}
\sigma^2 &\sim \Gamma^{-1}(0.01, 0.01) \\
\alpha_k &\sim U[0, \infty] \\
\beta_k &\sim U[0, \infty] \\
\alpha_k^H &\sim U[0, \infty] \\
\beta_k^H &\sim U[0, \infty] \\
\log(\lambda_i) - \log(\lambda_{i-1}) &\sim N(0, \sigma^2) \\
d_i|\lambda_i &\sim Po(\lambda_i) \\
p_{i,j}|\alpha, \beta, \alpha^H, \beta^H &\sim Beta(\alpha_j^H \alpha_{min(j-i,k)}, \beta_j^H \beta_{min(j-i,k)}) \\
r_{i,j}|d_i, \mathbf{r}_{1,1:j}, p &\sim Bin(d_i - \sum_{k=1}^{j-1} r_{i,k}, p_{i,j}),
\end{aligned}
$$

where where and $j \leq i$ and $i = 1, \ldots, T$.

## Inference

As the main goal to generate inference of the number of death **d** is through the posterior distribution of number of deaths **d** given the observations **r**. In order to generate samples from this distribution we use a Markov Chain Monte Carlo method Brooks et al. 2011. In more detail we use a blocked Gibbs sampler, which generates samples in the following sequence:

- We sample $\alpha, \beta, \alpha^H, \beta^H|\mathbf{d}, \mathbf{r}$ using the fact that one can integrate out $p$ in the model, and then $\mathbf{d}|\alpha, \beta, \alpha^H, \beta^H, \mathbf{r}, \lambda$ follows a Beta-Binomial distribution. Here to we use an adaptive MALA Atchadé 2006 to sample from these parameters.
- To sample $\mathbf{d}|\alpha, \beta, \alpha^H, \beta^H, \mathbf{r}, \lambda$, that each death, $d_i$ is conditionally independent, and we just use a Metropolis Hastings random walk to sample each one.
- To sample $\lambda|\mathbf{d}, \sigma^2$ we again use an adaptive MALA.
- Finally we sample $\sigma^2|\mathbf{d}$ directly since this distribution is explicit.

## Model Benchmark

In this section, we present additional comparison of the model to a simple constant model. The benchmark model is the sum of average reporting lag for the preceding 14 days.