

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

Adam Altmejd* Joacim Rocklöv† Jonas Wallin§

June 11, 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 true daily quantities and their uncertainty, estimated using historical reporting delays. The methodology takes into account the observed distribution pattern of the lag. It is derived from the “removal method”, a well-established estimation framework in the field of ecology.

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 (Nature, 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 information is reported. In fact, reporting delays make the most recent days have the least cases accounted for, producing a dangerous illusion of an always improving outlook.

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 reporting delay, thus allowing policymakers to use the latest numbers available without being misled by reporting bias.

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.

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

†Department of Public Health and Clinical Medicine, Umeå University, joacim.rocklov@umu.se

‡Heidelberg Institute of Global Health, University of Heidelberg

§Department of statistics, Lund university, jonas.wallin81@gmail.com

In this paper we describe a statistical methodology for nowcasting the epidemic statistics, such as hospitalizations or deaths, and their degrees of uncertainty, based on the daily reported event frequency and the observed distribution pattern of reporting delays. The prediction model is building on methods developed in ecology, referred to as the “removal method” (Pollock, 1991).

To help motivate why such forecasting is needed, we now turn to the case of Sweden. The model is flexible by design, however, and could easily be applied to other countries as well.

1.1 The current situation of COVID-19 in Sweden

The Swedish Public Health Agency updates the COVID-19 statistics daily¹. During a press conference, they present updates on the number of deaths, admissions to hospitals and intensive care, as well as case counts.

One of the reasons for following these indicators is to enable public health professionals and the public to observe the evolving patterns of the epidemic (Anderson et al., 2020). In relation to policy, it is of specific interest to understand if the growth rates changes, which could indicate the need for a policy response. However, in each daily report only a proportion of the number of recent deaths is yet known, and this bias produces the illusion of a downward trend.

The death counts suffer from the longest reporting delay. In their daily press conference, the Swedish Public Health Agency warns for this by stopping the reported 7-day moving average trend line 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, creating a sense of a downward trend. In fact, this might be the reason why the number of daily deaths have been underestimated repeatedly. At the peak, deaths were initially believed to level out at around 60 per day, but after all cases had been reported more than two weeks later, the actual number was close to 120 (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 exists 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 any previous trend in the data,
- we can generate prediction intervals for the uncertainty about daily true frequencies,

¹The data is published on <https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/bekraftade-fall-i-sverige/>.

- the uncertainty estimates can be carried over to epidemiological models to help create 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 on 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. 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 propose a Bayesian version of the removal model that assumes an overdispersed binomial distribution for the daily observations of deaths in Sweden in COVID-19. We then calculate the posterior distribution, prediction median and 95% prediction intervals of the expected deaths from the reported deaths on each specific day. The method and algorithm is thoroughly described in the Supplementary Information.

To get accurate estimates we apply two institution-specific corrections. First, we only count workdays as constituting reporting delay, as very few deaths are reported during weekends. Second, we apply a constant bias correction to account for the fact that Swedish deaths come from two distinct populations with different trends: deaths in hospitals, and in elderly care.

In Figure 1 we apply the model to the latest statistics from Sweden. The graph shows reported and predicted deaths (with uncertainty intervals) as bars, and a dashed line plots the 7-day (centered) moving average. A version without predictions is used in the Public Health Agency’s daily press briefings. As expected, the model provides estimates of actual deaths considerably above the reported number of deaths. Not how the model predicts additional deaths above the moving average line.

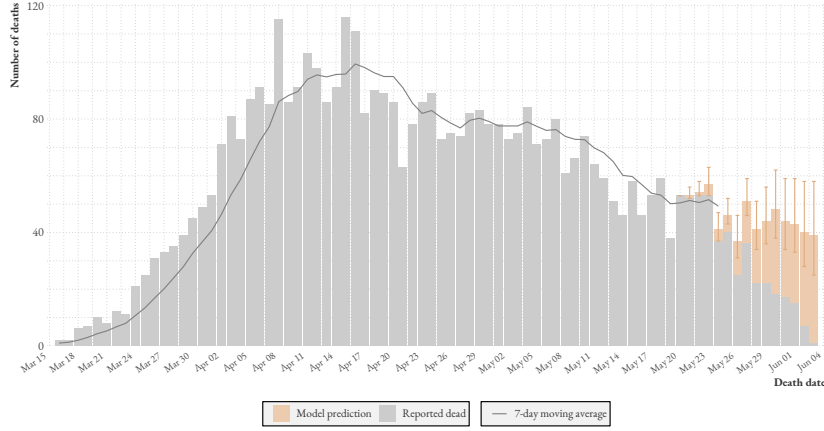


Figure 1: Swedish Covid-19 deaths as of 2020-06-04 and model predictions

3.1 Model Performance

To judge whether or not the model is accurate we need to compare it to a benchmark. The moving average of reported deaths is not useful, since it is biased for deaths that occurred within the last week. Instead, we create a benchmark prediction by a Normal distribution where the mean and standard deviation is taken from the historical lags from the last two weeks to the reported numbers².

Figure 2 depicts four randomly chosen dates where the model is compared to the benchmark. The model and the benchmark are tasked with predicting the total number of individuals who have died at a given date and have been reported within 14 days of that date. As time progresses, more deaths are reported and the dashed grey line approaches the horizontal line. Meanwhile model uncertainty decreases.

Figure 3 shows model performance compared to the benchmark for three difference performance metrics. All three graphs are based on predictions of reported deaths within 14 days, and show how performance increases as more data has been reported. Each data point is the average of all dates where predictions can be evaluated. SCRPS is a measure of accuracy that rewards precision, it is a proper scoring rule like the continuous probability rank score or the Brier score (see definition in Appendix) (Bolin and Wallin, 2019). The central plot shows the width of the prediction intervals, and the rightmost one the proportion of PIs that cover the true value.

Benchmark and model point estimates are similarly close to the truth. The model produces tighter prediction intervals. For 8-5 days of reporting lag (see Figure 3), the intervals are too tight. This is likely because the Public Health Agency queries the Swedish death registry for Covid-19 deaths only once or twice a week. Since we do not know the process, it has not been explicitly modeled.

²For a death date two days ago we add the mean of deaths reported after 3 days, 4 days, etc. We use the sum of standard deviations to generate the prediction intervals, assuming that lags are independent across days. The exact calculation is described in the appendix.

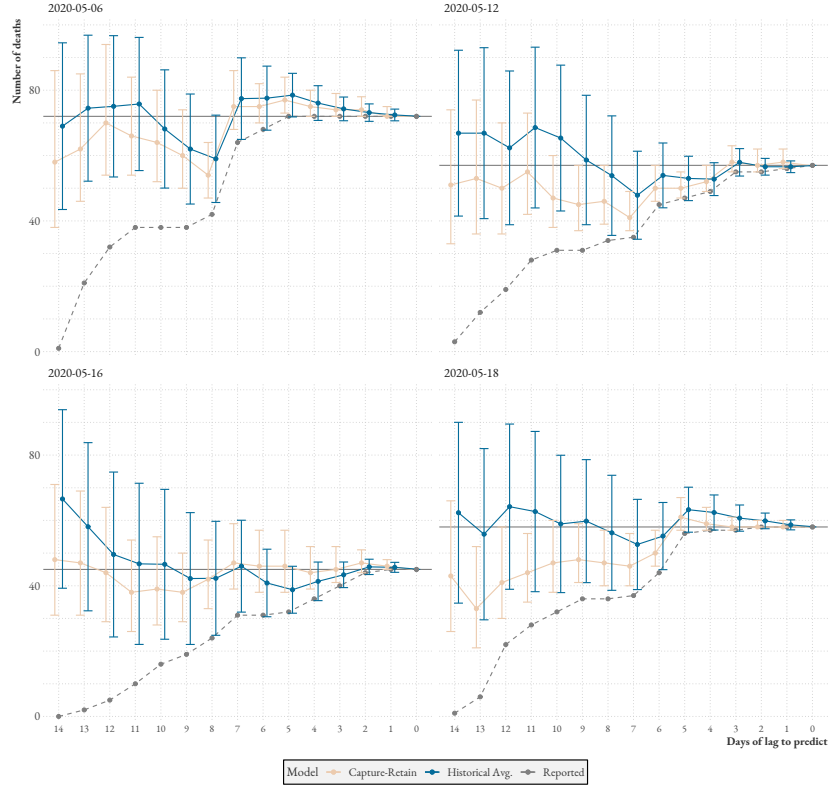


Figure 2: Model accuracy over time for four dates, compared to benchmark

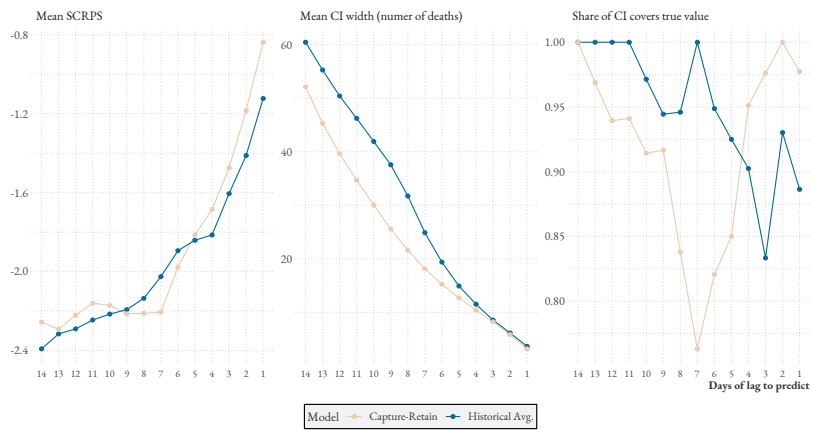


Figure 3: Model accuracy over time for four dates, compared to a benchmark

4 Implications and limitations

The model proposed here can 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 overcomes 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 users 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 than 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.

Moreover, deaths are reported from two distinct populations that seem to follow different trends. At the time of writing, the daily deaths in elderly care, reported with a longer delay, seem to be decreasing slower than hospital deaths. But statistics offer only aggregate numbers, prohibiting us from modeling two distinct processes. However, we have noted a clear decline in proportions of deaths reported the two first working days. For example the number of deaths occurring at the second of April $\approx 30\%$ of deaths were reported within the first two working days whereas for the eighteenth of May only $\approx 10\%$ were reported during the two first working days. We address this by assuming that the deaths reported during the two first working days comes from a different population than the remainder of days.

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, 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. When the covariates to the reporting delay pattern are known, the model can incorporate

them and provide more accurate predictions.

5 Conclusion

In this paper, we provide a method to accurately nowcast daily Covid-19 statistics that are reported with delay. By systematically modelling the delay, policy makers can avoid dangerous illusory downward trends. Our model also gives precise uncertainty intervals, making sure users of these statistics are aware of the fast-paced changes that are possible during this pandemic.

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.
- 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.
- Bolin, D. and J. Wallin (2019). Scale dependence: Why the average crps often is inappropriate for ranking probabilistic forecasts.
- 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.
- 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.
- 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.
- Møller, J., A. R. Syversveen, and R. P. Waagepetersen (1998). Log Gaussian Cox Processes. *Scandinavian Journal of Statistics* 25(3), 451–482.
- Moran, P. A. P. (1951). A Mathematical Theory of Animal Trapping. *Biometrika* 38(3/4), 307–311.
- Nature (2020). Coronavirus: Three things all governments and their science advisers must do now. *Nature* 579(7799), 319–320.
- Öhman, D. and A. Gagliano (2020). Antalet virusdöda har underskattats. *Sveriges Radio*.
- 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.

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

A Appendix

B Model

B.1 Notation

Before presenting the model we describe some notation used through out the appendix. For a $m \times n$ matrix r we use the following broadcasting notation $\mathbf{r}_{k,j:l} = [r_{k,j}, r_{k,j+1}, \dots, r_{k,l}]$. Further $x|y \sim \pi(\cdot)$ implies that the random variable x if we conditioning on y follows distribution $\pi(\cdot)$. 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.
α	$K \times 1$	Latent prior parameter for \mathbf{p}
β	$K \times 1$	Latent prior parameter for \mathbf{p}
α^H	2×1	parameter for the probability, \mathbf{p} for holiday adjustment.
β^H	2×1	parameter for the probability, \mathbf{p} for holiday adjustment.
μ	$T \times 1$	μ_i is the intensity of the expected number of deaths at day i .
σ^2	1×1	Variation of the random walk prior for the log intensity.
ϕ	1×1	overdispersion parameter for negative binomial distribution.
p_0	1×1	probability of reporting for a low reporting event.
p_i	1×1	probability of a low reporting event.

B.2 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 recorded 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

	Reported date				
Death 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.

$p_{i,j}$, i.e.

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

Typically in removal sampling one would set the probability of reporting uniform, i.e. $p_{i,j} := p$. However for this data this is clearly not realistic given weekly patterns in reporting –very little reporting during the weekends. Instead we assume that we have k different probabilities. Further, to account for overdispersion, we assume that each probability rather being a fixed scalar is a random variable with a Beta distribution. The Beta distribution has two parameters α and β . This resulting the following distribution for the probabilities

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

Here, if $j \in H$ then day j is a holidays or 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. Finally we add an extra mixture component that allows for very low reporting.

B.3 Priors

For the α and β parameters we use an (improper) uniform prior. For the deaths, 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), but instead of Poisson distribution we use a negative binomial to handle possible over dispersion. The latent Gaussian processes has a intrinsic random walk distribution (Rue and Held, 2005) i.e.

$$\begin{aligned} \log(\mu_i) - \log(\mu_{i-1}) &\sim N(0, \sigma^2), \\ d_i|\lambda_i &\sim \text{NegBin}(\mu_i, \phi). \end{aligned}$$

This model is created to create a temporal smoothing between the reported deaths. For the hyperparameter σ^2 we impose a inverse Gamma distribution, this prior is suitable here because it guarantees that the process is not constant ($\sigma^2 = 0$) which we know is not the case.

B.4 Full model

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

$$\begin{aligned}
\sigma^2 &\sim \Gamma(1, 0.01) \\
\phi &\sim \Gamma(1, 0.01) \\
\alpha_k &\sim U[0, \infty] \\
\beta_k &\sim U[0, \infty] \\
\alpha_k^H &\sim U[0, \infty] \\
\beta_k^H &\sim U[0, \infty] \\
\log(\mu_i) - \log(\mu_i - 1) &\sim N(0, \sigma^2) \\
d_i | \lambda_i &\sim \text{NegBin}(\mu_i, \phi) \\
p_{i,j} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}^H, \boldsymbol{\beta}^H &\sim \text{Beta}(\alpha_j^H \alpha_{\min(j-i,k)}, \beta_j^H \beta_{\min(j-i,k)}) \\
r_{i,j} | d_i, \mathbf{r}_{1:1:j}, p &\sim \pi \text{Bin}(d_i - \sum_{k=1}^{j-1} r_{i,k}, p_0) + (1 - \pi) \text{Bin}(d_i - \sum_{k=1}^{j-1} r_{i,k}, p_{i,j}),
\end{aligned}$$

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

C Inference

As the main goal to generate inference of the number of death \mathbf{d} is through the posterior distribution of number of deaths \mathbf{d} given the observations \mathbf{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 $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}^H, \boldsymbol{\beta}^H | \mathbf{d}, \mathbf{r}$ using the fact that one can integrate out p in the model, and then $\mathbf{d} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}^H, \boldsymbol{\beta}^H, \mathbf{r}, \boldsymbol{\lambda}$ follows a Beta-Binomial distribution. Here to we use an adaptive MALA (Atchadé, 2006) to sample from these parameters.
- To sample $\mathbf{d} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}^H, \boldsymbol{\beta}^H, \mathbf{r}, \boldsymbol{\lambda}$, that each death, d_i is conditionally independent, and we just use a Metropolis Hastings random walk to sample each one.
- To sample $\boldsymbol{\lambda} | \mathbf{d}, \sigma^2$ we again use an adaptive MALA.
- Finally We sample $\sigma^2 | \mathbf{d}$, and p_0, π directly since this distribution is explicit, and ϕ using a MH-RW.

D Model Benchmark

In this section, we present additional comparison of the model to the benchmark. We first describe the benchmark model in detail.

The benchmark model simply takes the sum of average historical reporting lags for the preceding 14 days. As before r_{ij} is the number of deaths that happened on day i and were recorded on day j . To predict the number of people that died on a given day, we first calculate lag averages:

$$\hat{r}_{i,i+L} = \frac{\sum_{k=i-14}^{i+L} r_{k-L,k}}{14}, \quad (1)$$

where $\hat{r}_{i,i+L}$ is the average number of deaths reported with a lag of L days, based on the 14 reports closest preceding day i . If we are looking at data released 2020-04-28 and call this day 0, the latest death date that we have 10-day ($L = 10$) reporting lag observation for is $r_{-10,0}$. The average for $Lag(0, 10)$ is therefore taken over the 14 days between $r_{-24,-14}$ and $r_{-10,0}$ (2020-04-04 and 2020-04-18). For this reason, some of the earlier predictions will not have data from 14 days. The average is then taken over all available reports.

In the comparisons we aim at predicting the total number of deaths that will have been reported within 14 days of the death date. To do so, we sum over the average lag that has yet to be reported. If we are predicting the number of people that have yet to be reported dead for day -3, we already know the true values for $r_{-3,-3}$, $r_{-3,-2}$, $r_{-3,-1}$, and $r_{-3,0}$ so we only need to predict $r_{-3,1} \dots r_{-3,10}$. The prediction is then

$$Benchmark(i, j) = \sum_{l=i}^j r_{i,l} + \sum_{l=j}^{14} \hat{r}_{i,l}. \quad (2)$$

As confidence interval we simply use a Normal assumption with standard deviations of the reporting lags, assuming independence, i.e. this is just the square root of the sum of $Var(\hat{r})$.

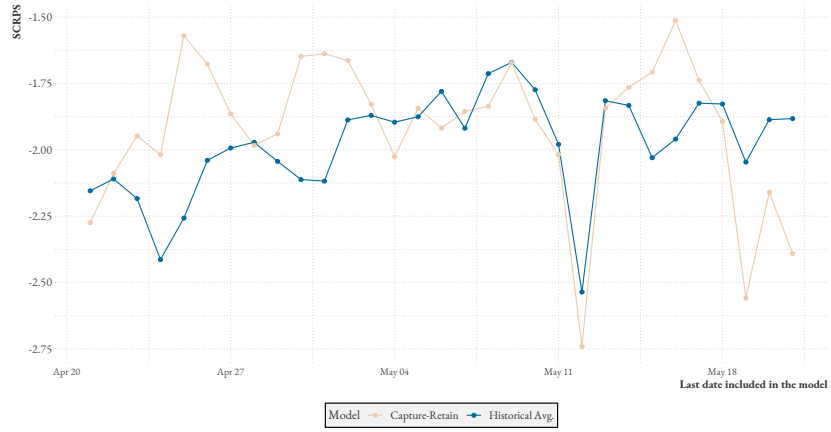


Figure 4: Average SCRPS as the pandemic progresses.

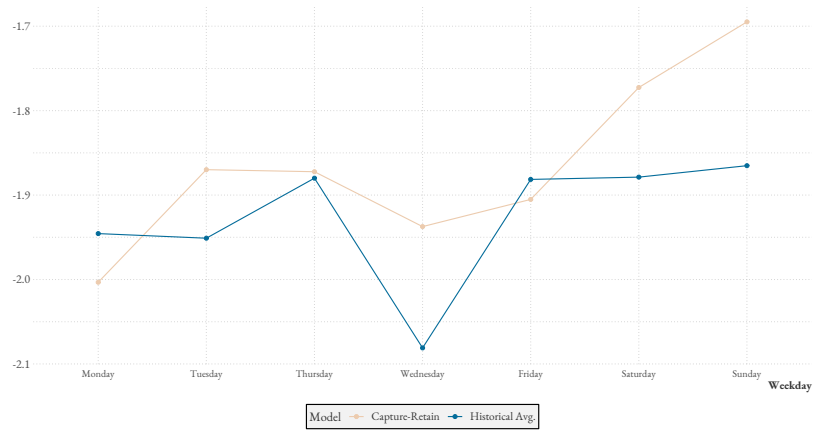


Figure 5: Average SCRPS per weekday.