Covid nowcasting

Jonas Wallin
joint work with
Adam Altmejd (SU),
Joacim Rocklöv (UU)

Home
May 21, 2021

LUND
UNIVERSITY

starting point

- This talk is about nowcasting the number of deaths on a given with respect to what is reported so far.

starting point

- This talk is about nowcasting the number of deaths on a given with respect to what is reported so far.
- For a specific day it is only often after several days (weeks) that one has access to how many actually died that day.

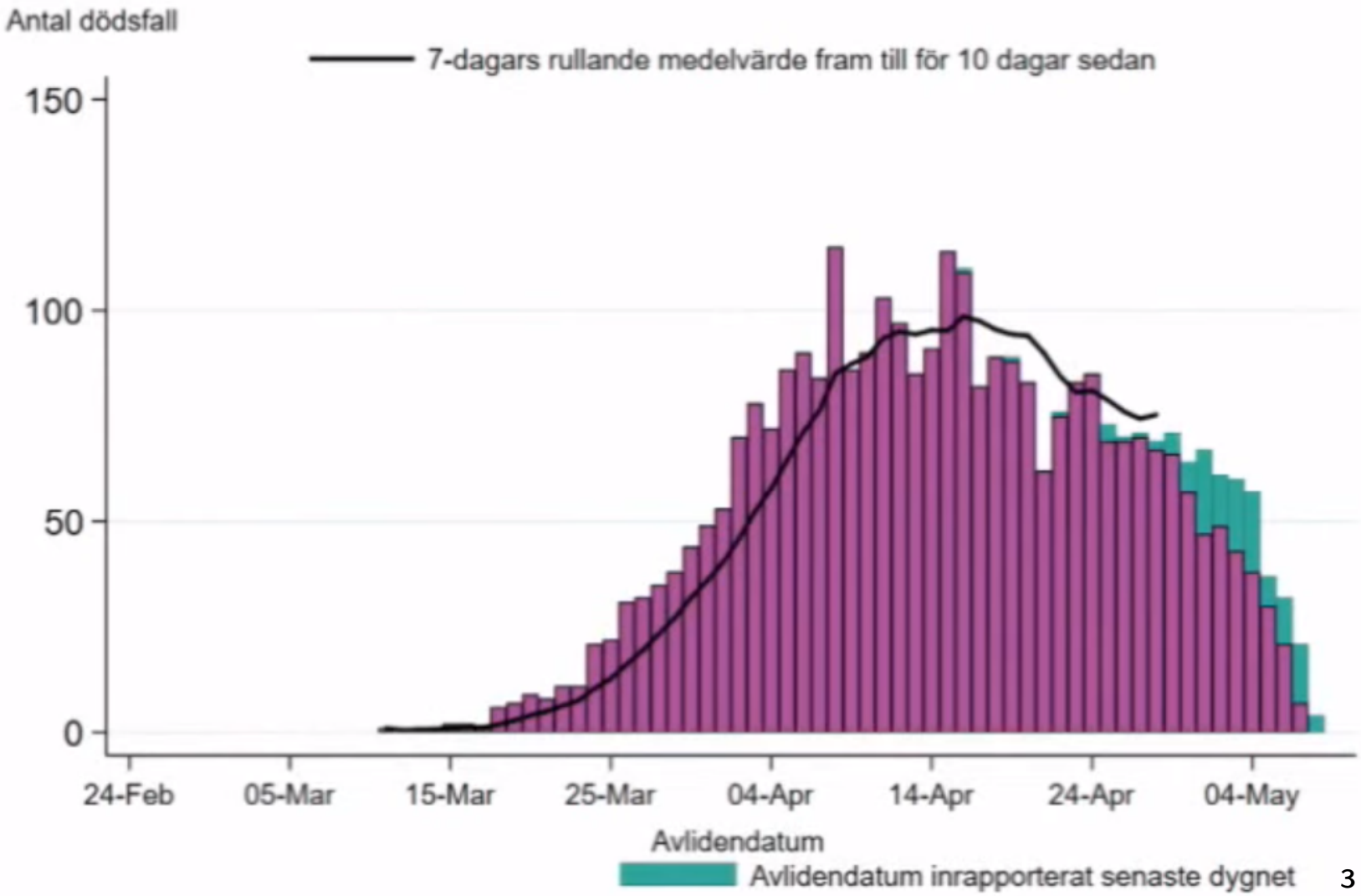
starting point

- This talk is about nowcasting the number of deaths on a given with respect to what is reported so far.
- For a specific day it is only often after several days (weeks) that one has access to how many actually died that day.
- In Sweden and in the UK the data is presented before fully collected.

starting point

- This talk is about nowcasting the number of deaths on a given with respect to what is reported so far.
- For a specific day it is only often after several days (weeks) that one has access to how many actually died that day.
- In Sweden and in the UK the data is presented before fully collected.
- To have early access to the number of death is important for knowing where the epidemic is going and what efforts needs to be put in.

Antal avlidna per dag



Structure

- Our goal is to model how many died on a given day. We denote:

  dea_t – actual number of indivuals that died of covid on day t

- Our data is how many people are reported at a given day. We denote:

  $rep_{t,k}$–How many indivuals reported dead on day t reported at day k

Data

Thus let us examine how many died on 19th of May 2020 (t):

k:	Report date	$(\sum_{j=t}^{k} rep_{t,j})$
t	$05-19$	0
$t+1$	$05-20$	5
$t+2$	$05-21$	12
$t+3$	$05-22$	12
$t+4$	$05-23$	15
$\vdots$	$\vdots$	
$t+349$	$2021-05-05$	55

Capture-Retain

- Long time ago, I worked on estimating populations of natterjack toads, in Archipelago of Bohuslän.

Capture-Retain

- Long time ago, I worked on estimating populations of natterjack toads, in Archipelago of Bohuslän.
- Estimation of population is well studied problem in ecology. Typical methods are type of capture-recapture method.

Capture-Retain

- Long time ago, I worked on estimating populations of natterjack toads, in Archipelago of Bohuslän.
- Estimation of population is well studied problem in ecology. Typical methods are type of capture-recapture method.
- If p is probability of catching an animal and n is the number of catches and N the target population. Then approximately $N \approx \frac{n}{p}$.

Capture-Retain

- Long time ago, I worked on estimating populations of natterjack toads, in Archipelago of Bohuslän.
- Estimation of population is well studied problem in ecology. Typical methods are type of capture-recapture method.
- If p is probability of catching an animal and n is the number of catches and N the target population. Then approximately $N \approx \frac{n}{p}$.
- On the islands there was small isolated populations. This meant that one could use rejection sampling or capture-retain method.

Capture-Retain 2

- One collects as many toads as one can (in a bucket). Then one waits a while and then one do it again. Repeat the procurer T times.
- This gives one a series of observation:

$$toad_1, toad_2, \ldots, toad_T$$

Capture-Retain 2

- One collects as many toads as one can (in a bucket). Then one waits a while and then one do it again. Repeat the procurer T times.
- This gives one a series of observation:

$$toad_1, toad_2, \ldots, toad_T$$

- If one collects the toads from a small populations then one should get fewer toads each time and this actually gives some information of the actual size of the population.

Capture-Retain 2

- One collects as many toads as one can (in a bucket). Then one waits a while and then one do it again. Repeat the procurer T times.
- This gives one a series of observation:

$$toad_1, toad_2, \ldots, toad_T$$

- If one collects the toads from a small populations then one should get fewer toads each time and this actually gives some information of the actual size of the population.
- Suppose the probability of catching a toad is p and the true population is N then:

$$\begin{aligned} toad_1 &\sim Bin(N, p) \\ toad_2 &\sim Bin(N - toad_1, p) \\ toad_3 &\sim Bin(N - toad_1 - toad_2, p) \end{aligned}$$

Now we can estimate both p and N.

Capture-Retain example

- In a Bayesian frame work we can estimate them using posterior distribution. Or treat $\hat{N}$ and $\hat{p}$ as parameters and estimate the ML

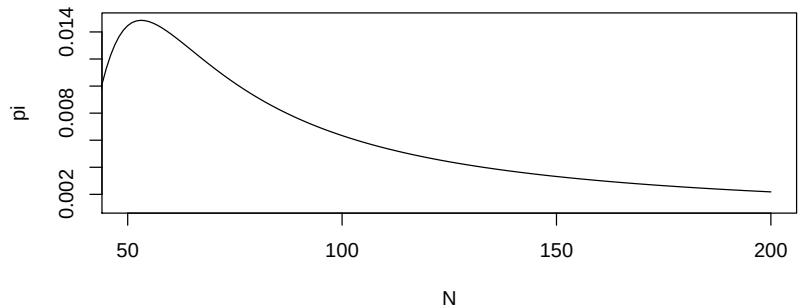
Capture-Retain example

- In a Bayesian frame work we can estimate them using posterior distribution. Or treat $\hat{N}$ and $\hat{p}$ as parameters and estimate the ML
- I realised yesterday that the ML is clearly biased here but it seems easy to correct this, but that for an other day maybe (Does not matter for our data since number of observations will be large).
- Let for example be Bayesian and study an example.

Capture-Retain example

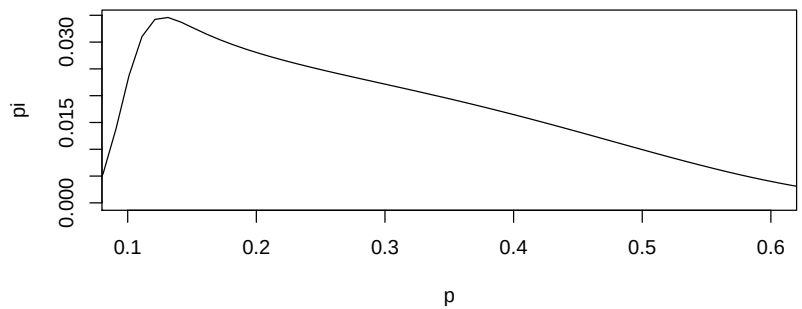
- Observe:

$$toad = [22, 15]$$



- posterior:

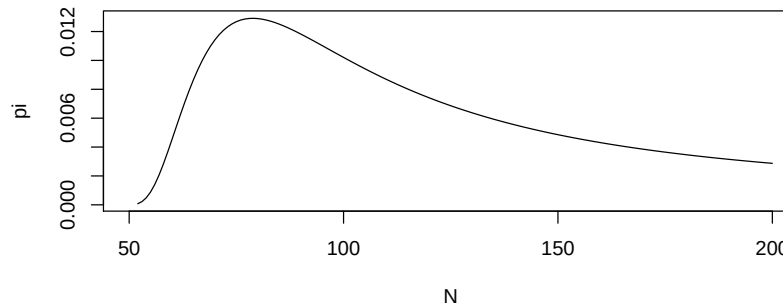
$$\pi(N, p|toad) \propto \binom{N}{22}\binom{N-22}{15} p^{22+15}(1-p)^{2N-22-15}$$



Capture-Retain example

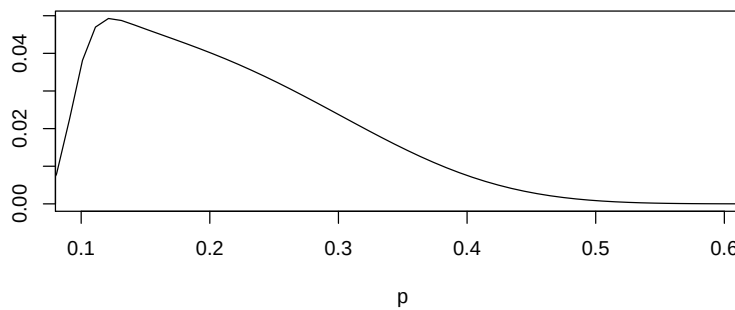
- Observe:

$$toad = [22, 15, 14]$$



- posterior:

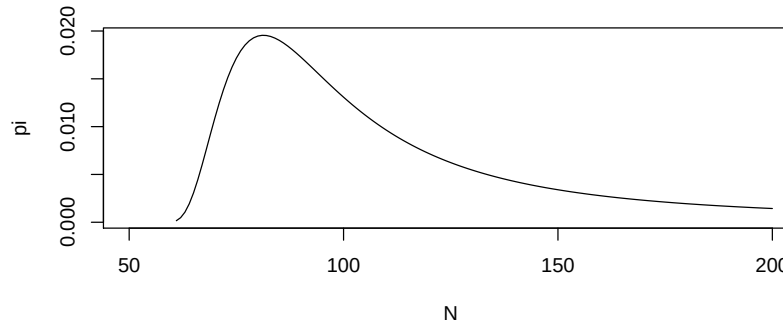
$$\pi(N, p|toad) \propto \left(\prod_{i}^{m} \binom{N - \sum_{1}^{i-1} toad_i}{toad_i} \right) p^{\sum_i toad_i} (1-p)^{mN - \sum_i toad_i}$$



Capture-Retain example

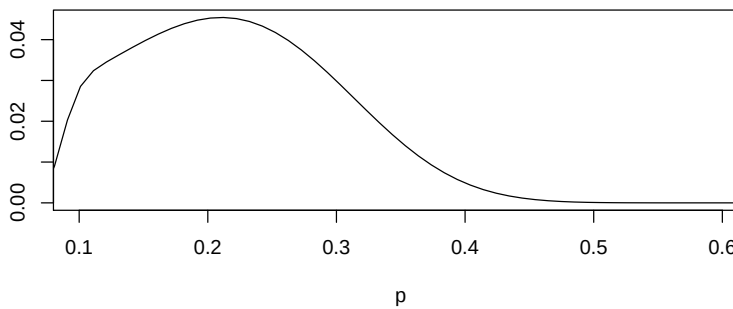
- Observe:

$$toad = [22, 15, 14, 9]$$



- posterior:

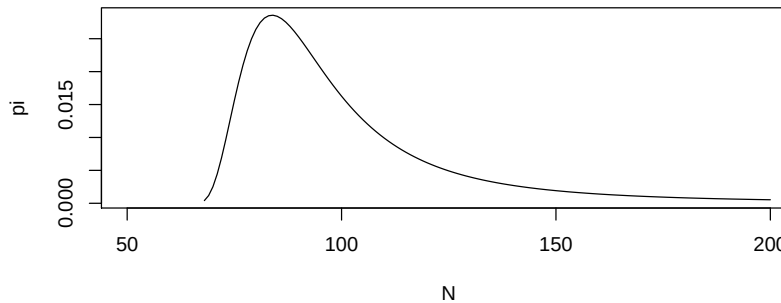
$$\pi(N, p|toad) \propto \left(\prod_{i}^{m} \binom{N - \sum_{1}^{i-1} toad_i}{toad_i} \right) p^{\sum_i toad_i} (1-p)^{mN - \sum_i toad_i}$$



Capture-Retain example

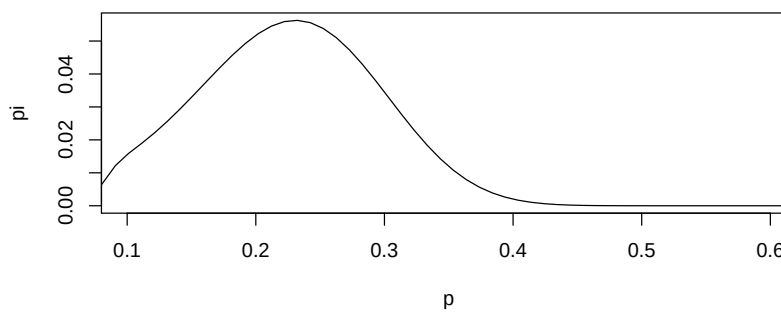
- Observe:

$$toad = [22, 15, 14, 9, 7]$$



- posterior:

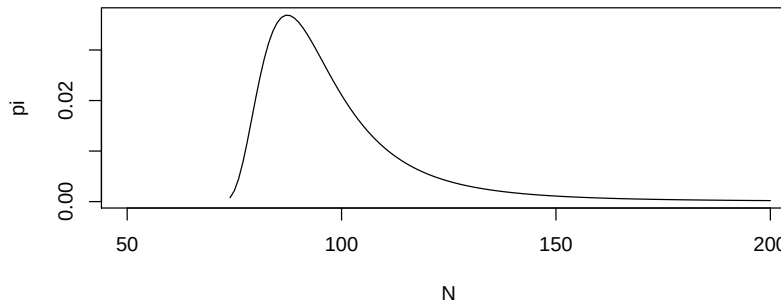
$$\pi(N, p|toad) \propto \left(\prod_i^m \binom{N - \sum_1^{i-1} toad_i}{toad_i} \right) p^{\sum_i toad_i} (1-p)^{mN - \sum_i toad_i}$$



Capture-Retain example

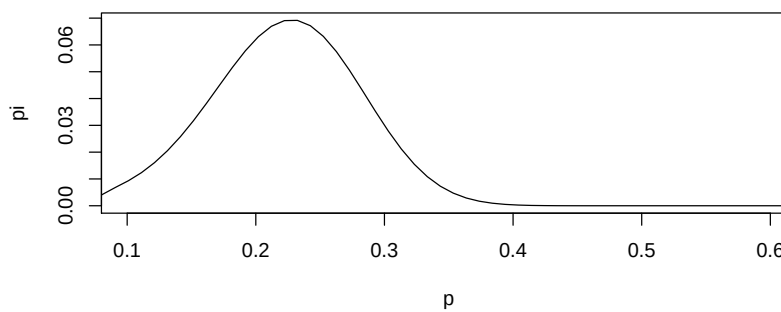
- Observe:

$$toad = [22, 15, 14, 9, 7, 6]$$



- posterior:

$$\pi(N, p|toad) \propto \left(\prod_{i}^{m} \binom{N - \sum_{1}^{i-1} toad_i}{toad_i} \right) p^{\sum_i toad_i} (1-p)^{mN - \sum_i toad_i}$$



Back to the covid-data

- For each given day we can apply the same idea for the reporting

$$rep_{t,.} = [rep_{t,t}, rep_{t,t+1}, ..., rep_{t,T}].$$

and thus estimate dea_t using the model

$$\begin{aligned} rep_{t,t} &\sim Bin(dea_t, p_0) \\ rep_{t,t+1} &\sim Bin(dea_t - rep_{t,t}, p_1) \\ &\vdots \\ rep_{t,t+i} &\sim Bin(dea_t - \sum_{j=0}^{i-1} rep_{t,t+j}, p_i) \end{aligned}$$

Back to the covid-data

- For each given day we can apply the same idea for the reporting

$$rep_{t,.} = [rep_{t,t}, rep_{t,t+1}, ..., rep_{t,T}].$$

and thus estimate dea_t using the model

$$\begin{aligned} rep_{t,t} &\sim Bin(dea_t, p_0) \\ rep_{t,t+1} &\sim Bin(dea_t - rep_{t,t}, p_1) \\ &\vdots \\ rep_{t,t+i} &\sim Bin(dea_t - \sum_{j=0}^{i-1} rep_{t,t+j}, p_i) \end{aligned}$$

Back to the covid-data

- For each given day we can apply the same idea for the reporting

$$rep_{t,.} = [rep_{t,t}, rep_{t,t+1}, ..., rep_{t,T}].$$

and thus estimate dea_t using the model

$$\begin{aligned} rep_{t,t} &\sim Bin(dea_t, p_0) \\ rep_{t,t+1} &\sim Bin(dea_t - rep_{t,t}, p_1) \\ &\vdots \\ rep_{t,t+i} &\sim Bin(dea_t - \sum_{j=0}^{i-1} rep_{t,t+j}, p_i) \end{aligned}$$

- I had missed this for medical data this had been done in previously in [Lawless, J. F. 1994]. (Of course they missed the biological connection dating back to [Moran, P. A. P. 1951])

overdispersion

- Often the data when working with Binomial distribution is overdispersed (higher variability compared to the distribution), here there is no exception.

overdispersion

- Often the data when working with Binomial distribution is overdispersed (higher variability compared to the distribution), here there is no exception.
- Once enough data is collected one can assume that one have the truth. In uk we are very stable after 30 days (much sooner). In this case things reduces to that both $rep_{t,\cdot}$ and dea_t is known and we only need to estimate p.

overdispersion

- Often the data when working with Binomial distribution is overdispersed (higher variability compared to the distribution), here there is no exception.
- Once enough data is collected one can assume that one have the truth. In uk we are very stable after 30 days (much sooner). In this case things reduces to that both $rep_{t,.}$ and dea_t is known and we only need to estimate p.
- Let us examine how many are reported at lag 2 (number of cases reported two days after).

overdispersion

- Often the data when working with Binomial distribution is overdispersed (higher variability compared to the distribution), here there is no exception.
- Once enough data is collected one can assume that one have the truth. In uk we are very stable after 30 days (much sooner). In this case things reduces to that both $rep_{t,\cdot}$ and dea_t is known and we only need to estimate p.
- Let us examine how many are reported at lag 2 (number of cases reported two days after).
- With a bit of simplification we approximately have:

$$\hat{p}_2 = \frac{\sum_t rep_{t,t+2}}{\sum_t dea_t} \approx \frac{\sum_t rep_{t,t+2}}{\sum_t \sum_{i=0}^{30} dea_{t,t+i}}$$

overdispersion

- Given $\hat{p}$ we have approximately

$$rep_{t,t+2} \sim Bin(rep_{t,t+30} \approx dea_t, \hat{p}_2)$$

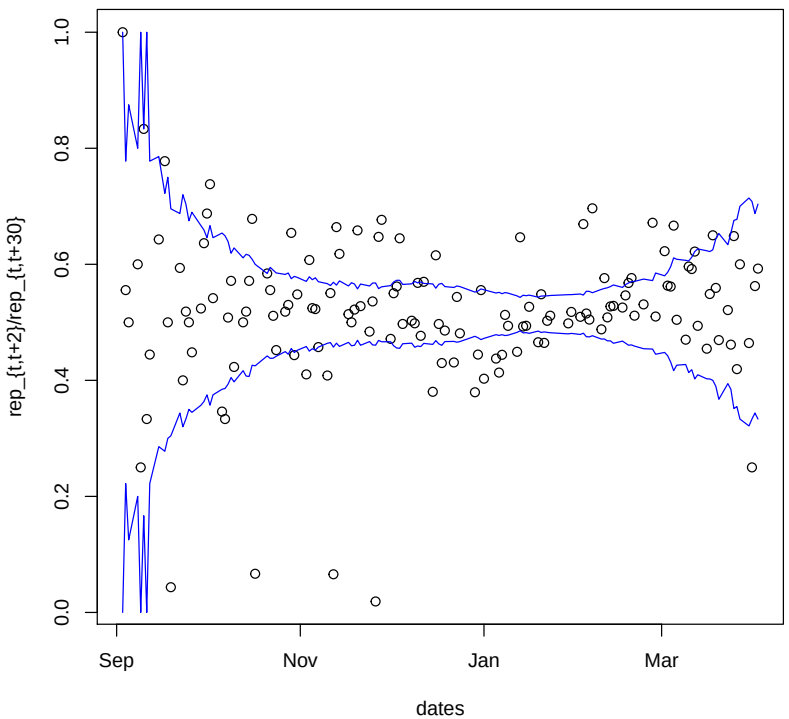
Thus we generate CI for $\frac{rep_{t,t+2}}{dea_t}$ to check the Binomial distribution assumption.

overdispersion

- Given $\hat{p}$ we have approximately

$$rep_{t,t+2} \sim Bin(rep_{t,t+30} \approx dea_t, \hat{p}_2)$$

Thus we generate CI for $\frac{rep_{t,t+2}}{dea_t}$ to check the Binomial distribution assumption.



overdispersion 3

- To address this issue we instead of assuming a Binomial distribution use a Beta-Binomial distribution.

overdispersion 3

- To address this issue we instead of assuming a Binomial distribution use a Beta-Binomial distribution.
- This corresponds to:

$$p \sim Beta(\mu, M),$$

$$rep_{t,t+2} \sim Bin(\sum_{i=0}^{30} dea_{t,t+i} rep_{t,t+i} \approx dea_t, p).$$

One can integrate out p and get

$$rep_{t,t+2} \sim BB(\sum_{i=0}^{30} dea_{t,t+i} rep_{t,t+i} \approx dea_t, \mu, M).$$

here μ is the expected value while M is overdispersion.

overdispersion 3

- To address this issue we instead of assuming a Binomial distribution use a Beta-Binomial distribution.
- This corresponds to:

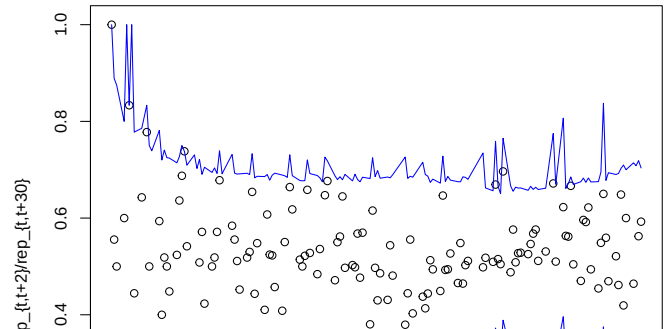
$$p \sim Beta(\mu, M),$$

$$rep_{t,t+2} \sim Bin(\sum_{i=0}^{30} dea_{t,t+i} rep_{t,t+i} \approx dea_t, p).$$

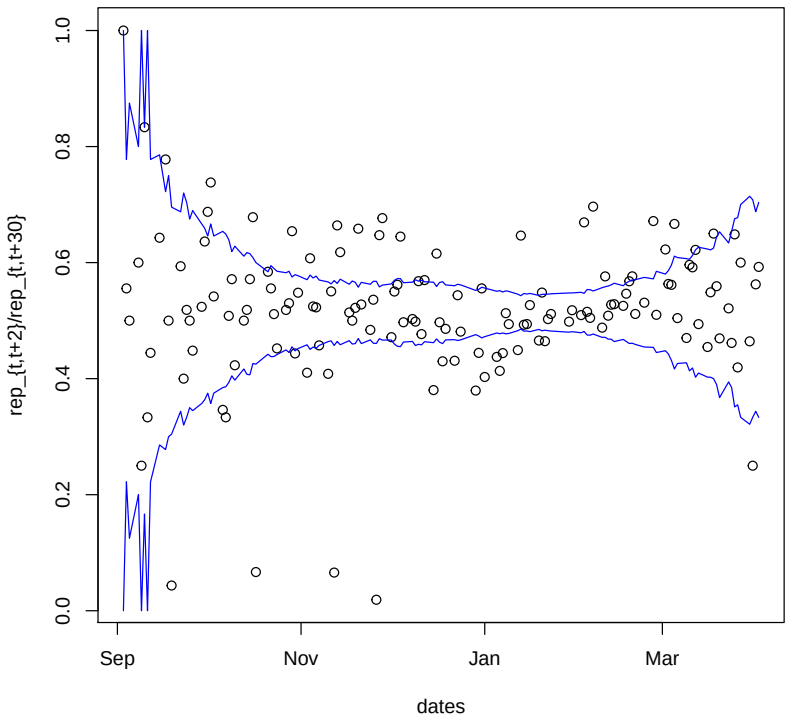
One can integrate out p and get

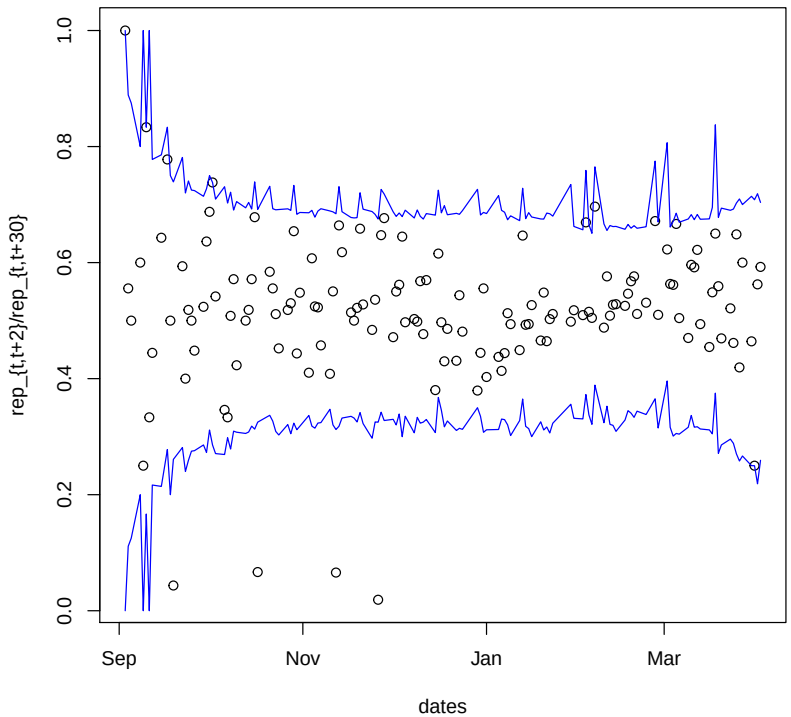
$$rep_{t,t+2} \sim BB(\sum_{i=0}^{30} dea_{t,t+i} rep_{t,t+i} \approx dea_t, \mu, M).$$

here μ is the expected value while M is overdispersion.



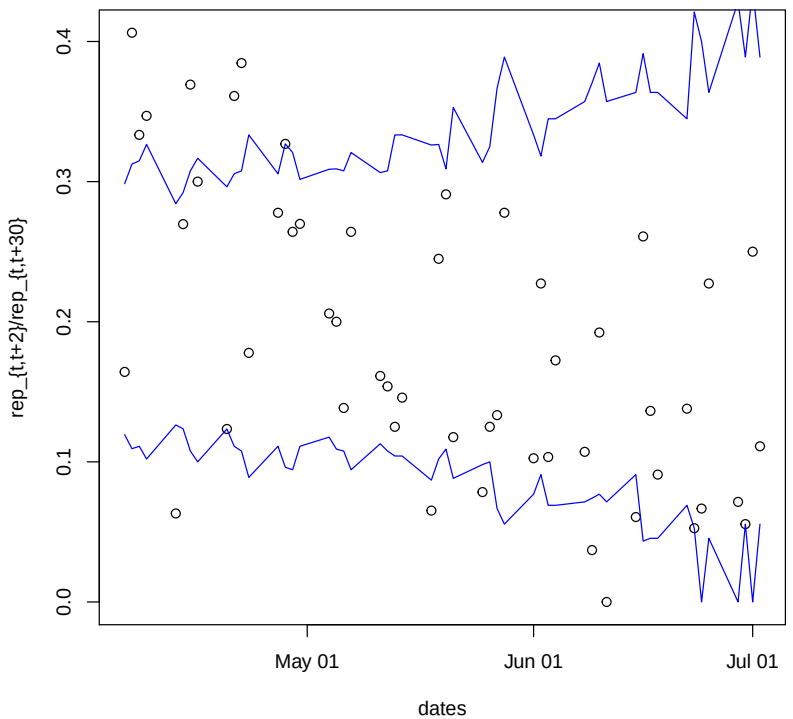
overdispersion 3





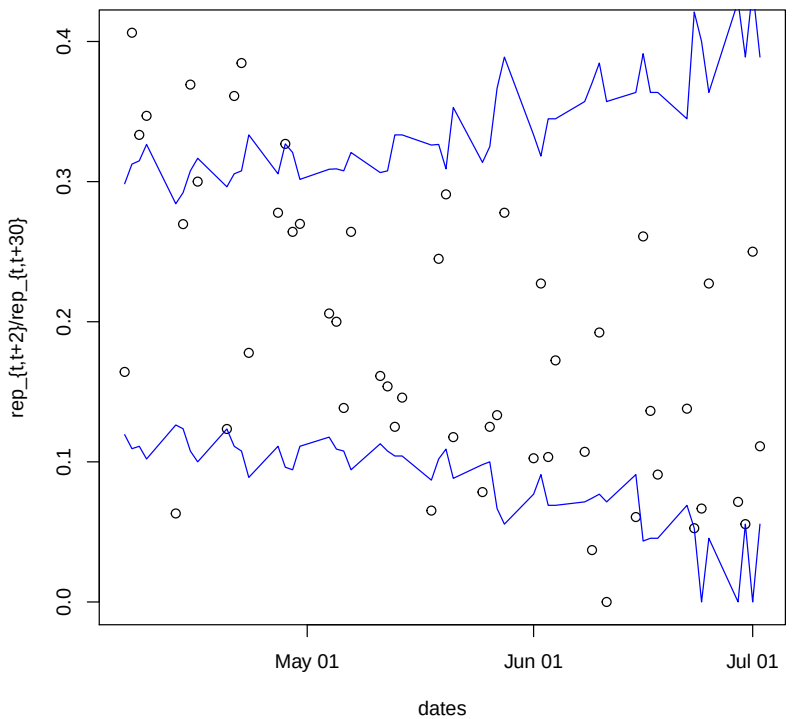
Sweden

- For Sweden things are a bit more complicated due to clear time trend in the data (at the beginning of the pandemic)



Sweden

- For Sweden things are a bit more complicated due to clear time trend in the data (at the beginning of the pandemic)
- We will ignore this for now and focus on the second wave.



result

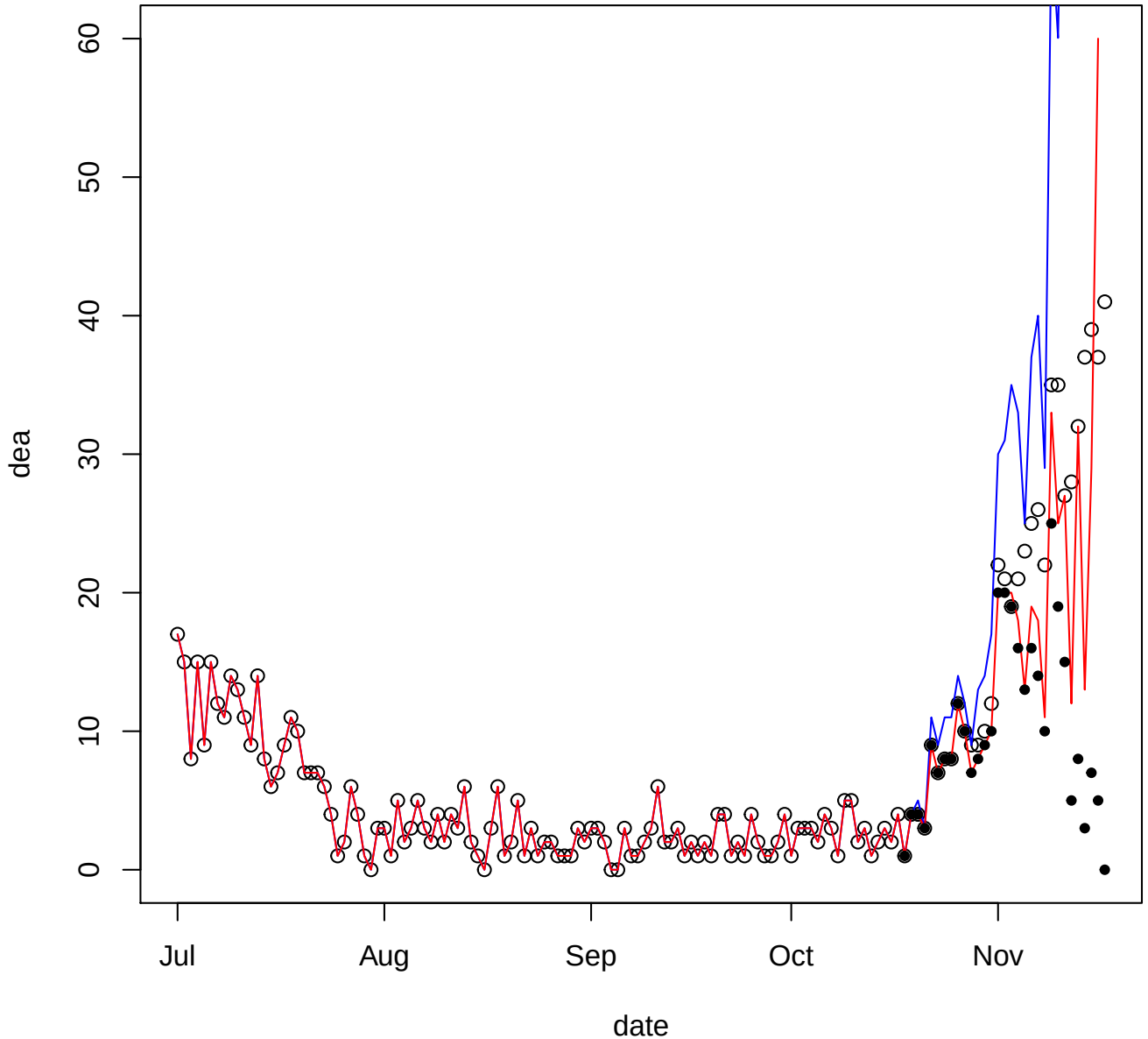
- So our model at day t is now

$$rep_{t,} \sim BB(dea_t, \theta_{BB})$$

  where θ_{BB} is set of parameters controlling for different lags, report days etc.
- Let us examine the result for Sweden at 17th of November 2020. Where we fitted θ_{BB} with ML removing the thirty latest dates and then sample from the posterior distribution i,e.

$$dea_t|rep_{t,}, \theta_{BB} \sim BB(rep_{t,}; dea_t, \theta_{BB})$$

result



result

- It is easy to see that the deaths are correlated which the model does not take into account.

result

- It is easy to see that the deaths are correlated which the model does not take into account.
- The underlying factor is the number of infected (also by age group matters).

result

- It is easy to see that the deaths are correlated which the model does not take into account.
- The underlying factor is the number of infected (also by age group matters).
- We can separately (from the reporting) try to model the epidemic.

$$dea_t | dea_{t-1:1} \sim ?$$
$$rep_{t,} \sim BB(dea_t, \theta_{BB})$$

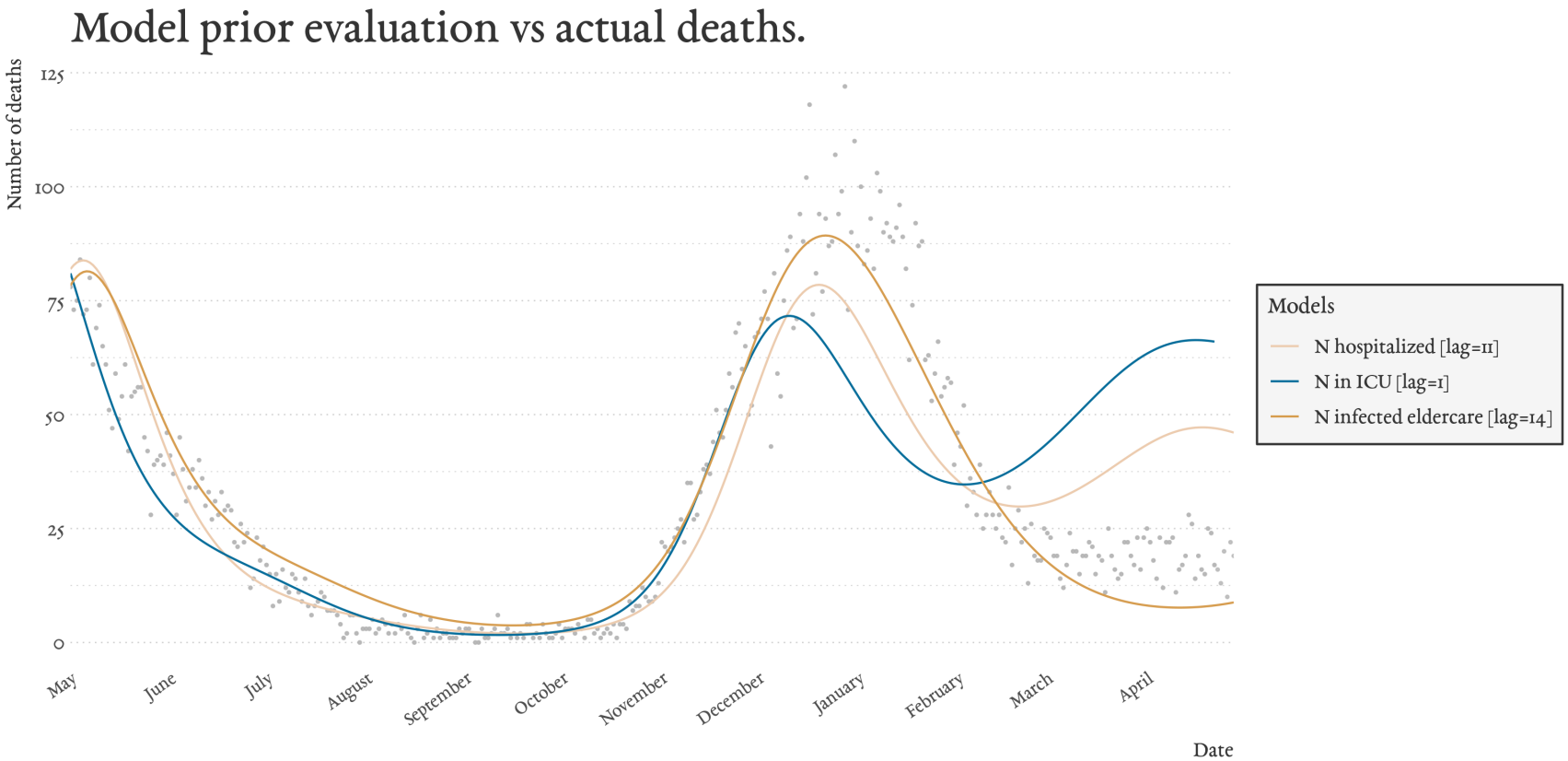
Adding other information

- We also have other data that correlates with the deaths: ICU, hospitalized with Covid, and infected in eldercare.
- Let M_t denote the one of these time series (or a smoothed version) we can use

$$dea_t \sim Po(exp\,(\beta_0 + \beta_1 M_{t-c}\,))$$

  here we need to estimate β_0, β_1, c.
- We in-fact included $M_t = ICU_t$ in our model.... at the completely wrong time.
- Issue they only correlate with the dea_t and this correlation can change over time

Adding other information



Gaussian processes ideal

- We instead would like to incorporate an underlying pandemic trend.

Gaussian processes ideal

- We instead would like to incorporate an underlying pandemic trend.
- Ideally we would like to model as follows

$$
\begin{aligned}
X(t) &\sim GP(t; \theta_{GP}), \\
\lambda_t &= \exp(X(t)), \\
dea_t &\sim Poisson(\lambda_t), \\
rep_{t,} &\sim BB(dea_t, \theta_{BB}).
\end{aligned}
$$

Gaussian processes ideal

- We instead would like to incorporate an underlying pandemic trend.
- Ideally we would like to model as follows

$$
\begin{aligned}
X(t) &\sim GP(t; \theta_{GP}), \\
\lambda_t &= \exp(X(t)), \\
dea_t &\sim Poisson(\lambda_t), \\
rep_{t,} &\sim BB(dea_t, \theta_{BB}).
\end{aligned}
$$

- Then sampling from the posterior distribution $dea_t | rep_{t,}$ one gets predictions.

Gaussian processes ideal

- We instead would like to incorporate an underlying pandemic trend.
- Ideally we would like to model as follows

$$
\begin{aligned}
X(t) &\sim GP(t; \theta_{GP}), \\
\lambda_t &= \exp(X(t)), \\
dea_t &\sim Poisson(\lambda_t), \\
rep_{t,} &\sim BB(dea_t, \theta_{BB}).
\end{aligned}
$$

- Then sampling from the posterior distribution $dea_t|rep_{t,}$ one gets predictions.
- However, this requires a bit of coding and also mentoring of the MCMC chains which we (I) am to lazy to do.

Gaussian processes ideal

- We instead would like to incorporate an underlying pandemic trend.
- Ideally we would like to model as follows

$$
\begin{aligned}
X(t) &\sim GP(t; \theta_{GP}), \\
\lambda_t &= \exp(X(t)), \\
dea_t &\sim Poisson(\lambda_t), \\
rep_{t,} &\sim BB(dea_t, \theta_{BB}).
\end{aligned}
$$

- Then sampling from the posterior distribution $dea_t | rep_{t,}$ one gets predictions.
- However, this requires a bit of coding and also mentoring of the MCMC chains which we (I) am to lazy to do.
- Instead we look to approximate steps in the chain to get a faster algorithm.

Gaussian processes approximation

- Since the data seems to be Poisson (or negative binomial) distributed we use a square root transformation

$$dea_t^{sqrt} = \sqrt{dea_t}$$

this since it gives the right relation between variance and mean.

Gaussian processes approximation

- Since the data seems to be Poisson (or negative binomial) distributed we use a square root transformation

$$dea_t^{sqrt} = \sqrt{dea_t}$$

this since it gives the right relation between variance and mean.

- Then the latent part becomes

$$X(t) \sim GP(t; \theta)$$
$$dea_t^{sqrt} \sim \mathcal{N}\left(X(t), \sigma^2\right)$$

Gaussian processes approximation

- Since the data seems to be Poisson (or negative binomial) distributed we use a square root transformation

$$dea_t^{sqrt} = \sqrt{dea_t}$$

this since it gives the right relation between variance and mean.

- Then the latent part becomes

$$X(t) \sim GP(t; \theta)$$
$$dea_t^{sqrt} \sim \mathcal{N}\left(X(t), \sigma^2\right)$$

- Yet the likelihood component

$$rep_{t,} \sim BB(dea_t, \theta_{BB})$$

is also still intractable and this we will approximate in the next step.

Gaussian processes approximation

- Here it is easy to sample from $dea_t|rep_{t,}$ from the likelihood only model (denoted $\mathcal{F}_{simple}$):

$$rep_{t,} \sim BB(dea_t, \theta_{BB}).$$

- From these samples we approximate the posterior distribution with the following

$$f(\sqrt{dea_t}|rep_{t,}) \approx \mathcal{N}\left(\sqrt{dea_t}; \mu_t, \sigma^2\right)$$

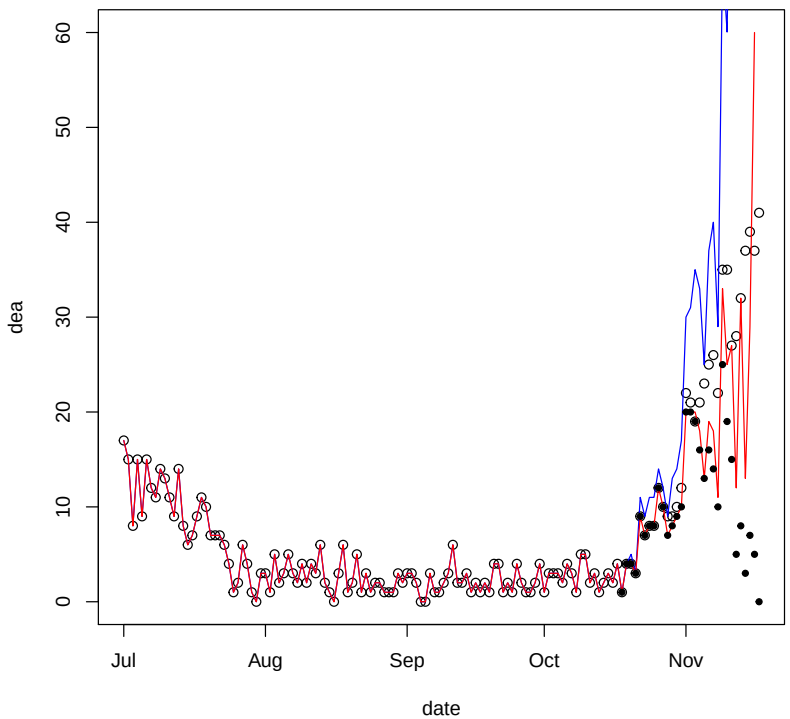
Where

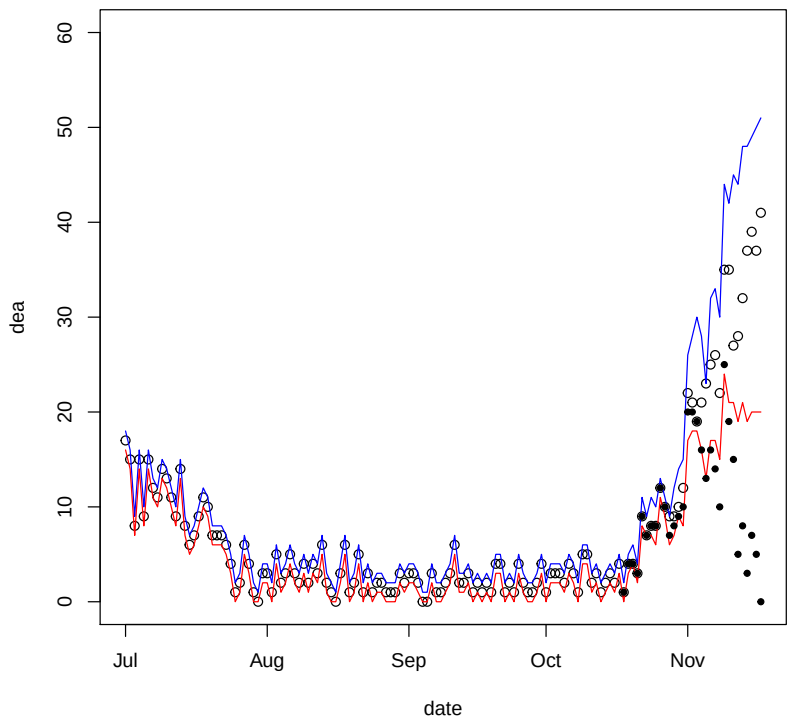
$$\mu_t = \mathbb{E}^{MC}\left[\sqrt{dea}_t|rep_{t,}, \mathcal{F}_{simple}\right]$$

and

$$\sigma_t^2 = \mathbb{V}^{MC}\left[\sqrt{dea}_t|rep_{t,}, \mathcal{F}_{simple}\right]$$

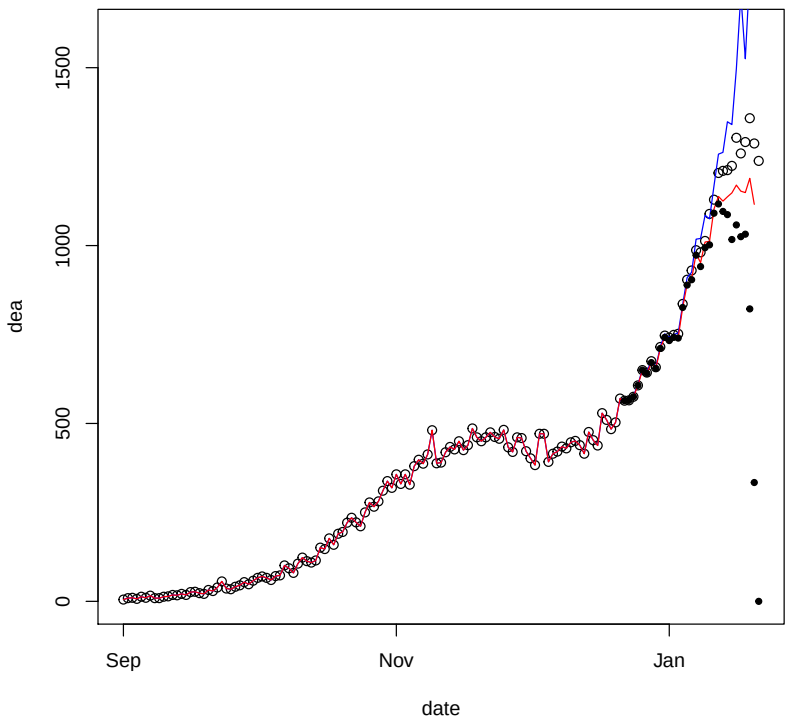
result

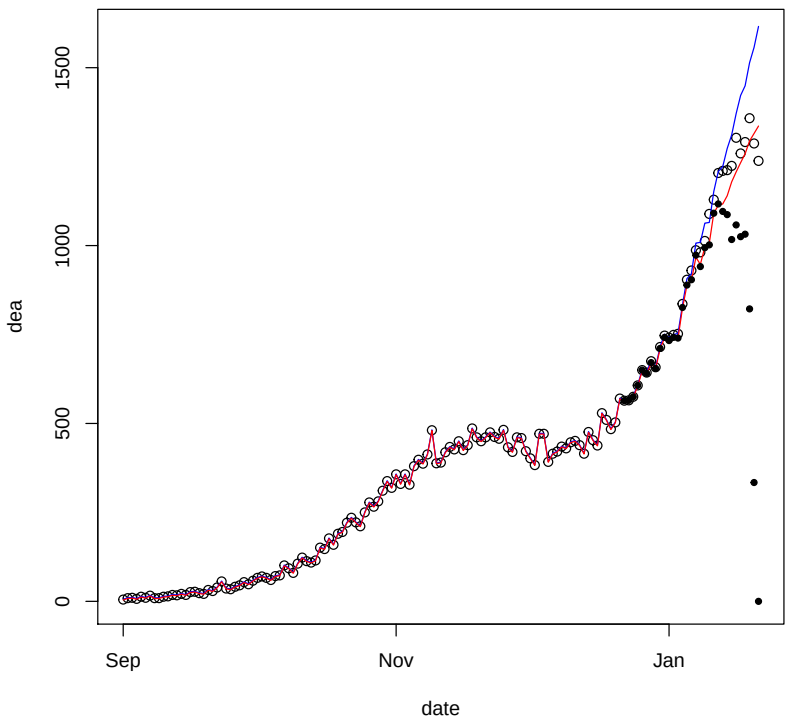

dea
0
10
20
30
40
50
60
Jul
Aug
Sep
Oct
Nov
date


dea
0
10
20
30
40
50
60
Jul
Aug
Sep
Oct
Nov
date

Gaussian processes issues

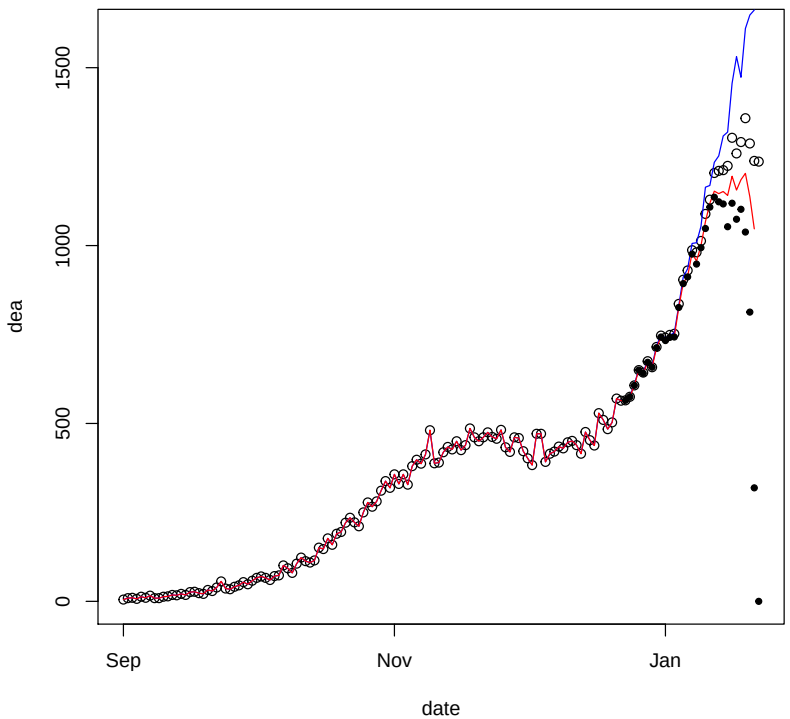
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

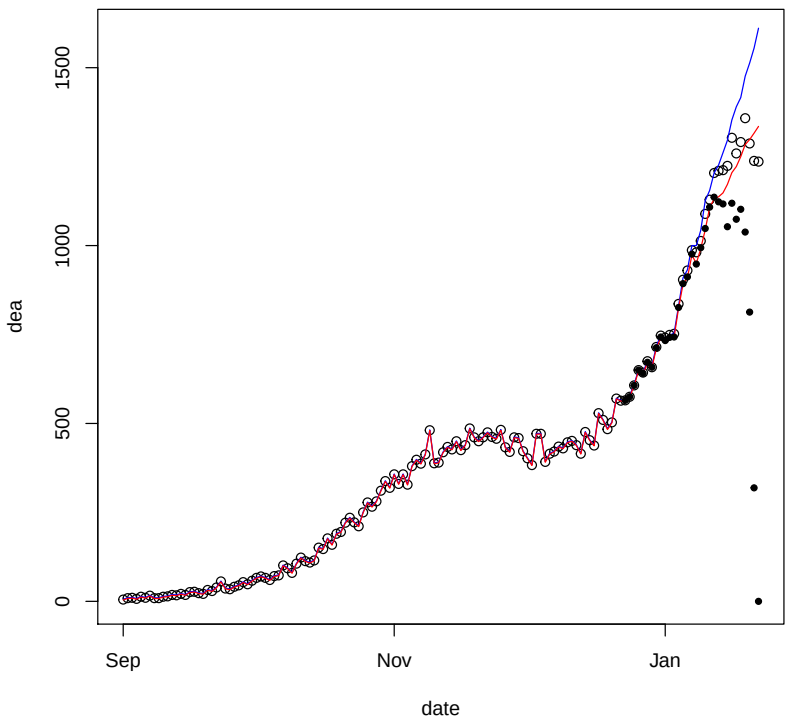




Gaussian processes issues

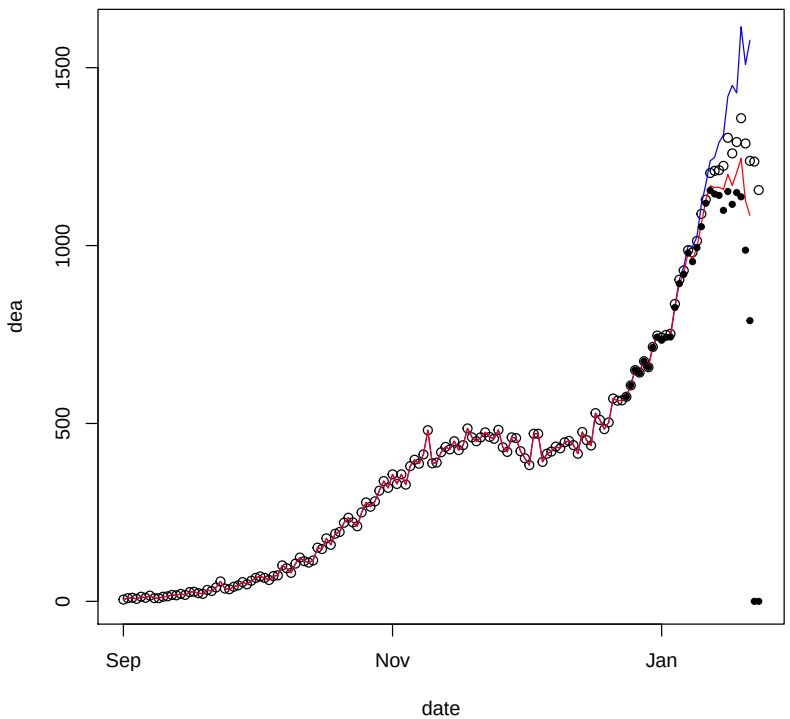
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

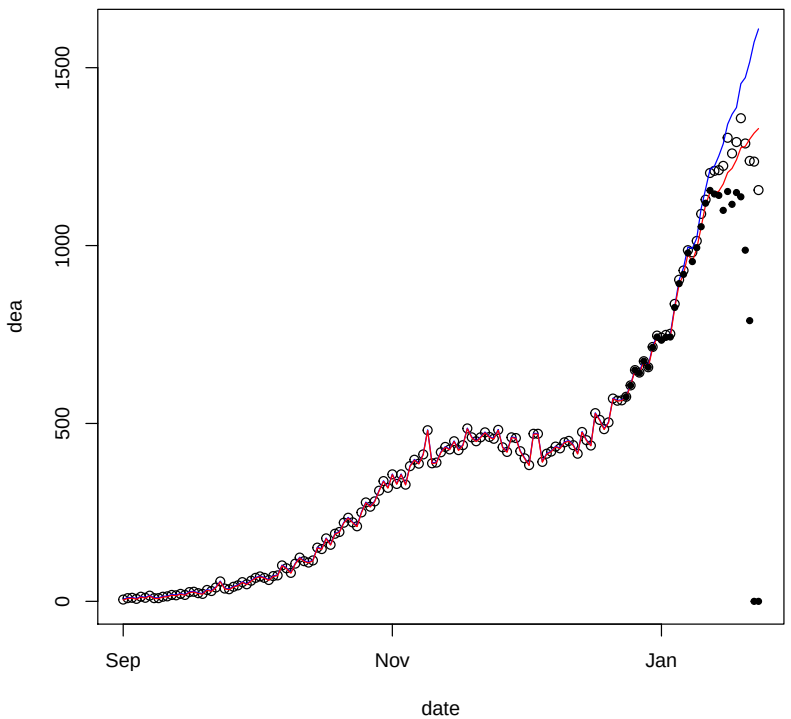




Gaussian processes issues

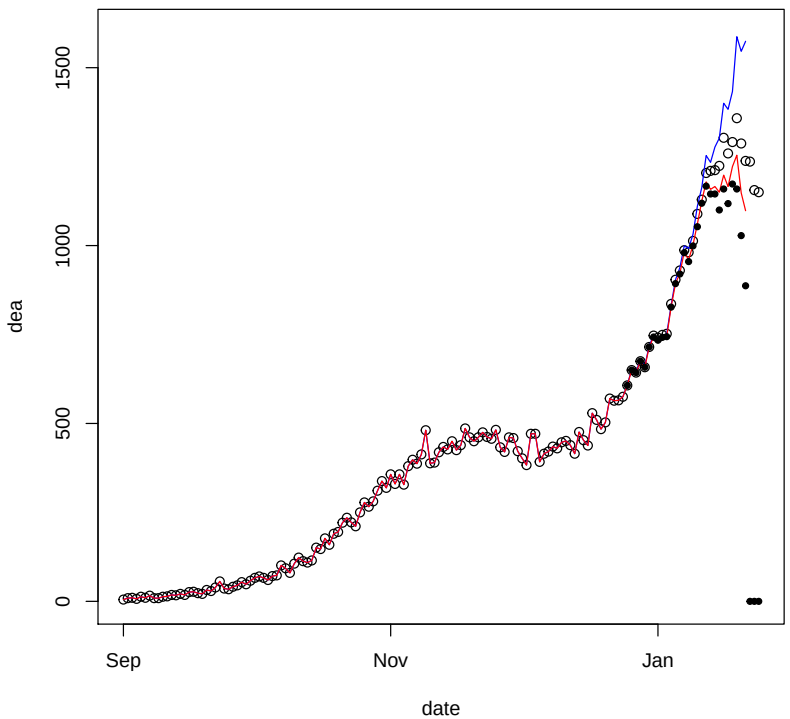
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

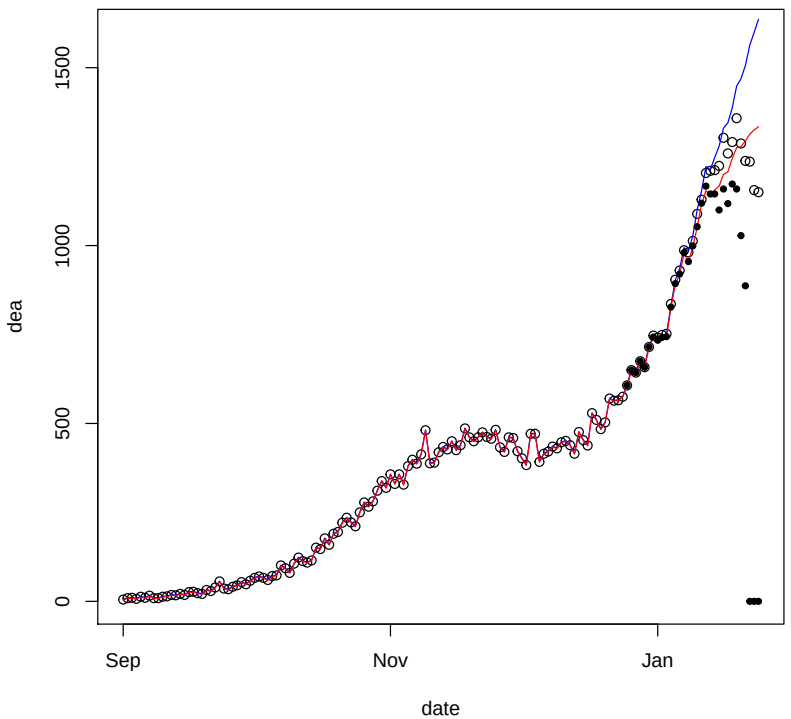




Gaussian processes issues

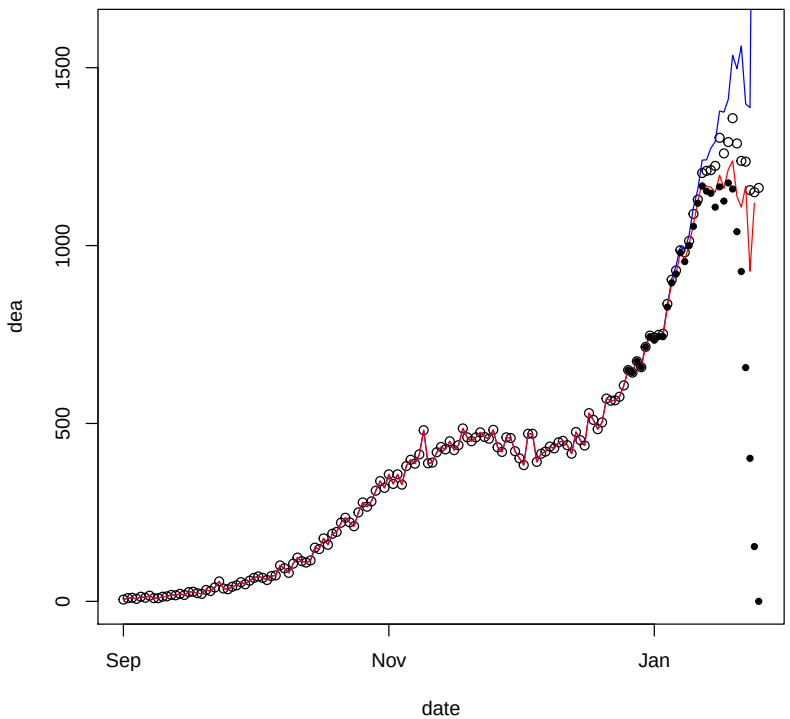
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

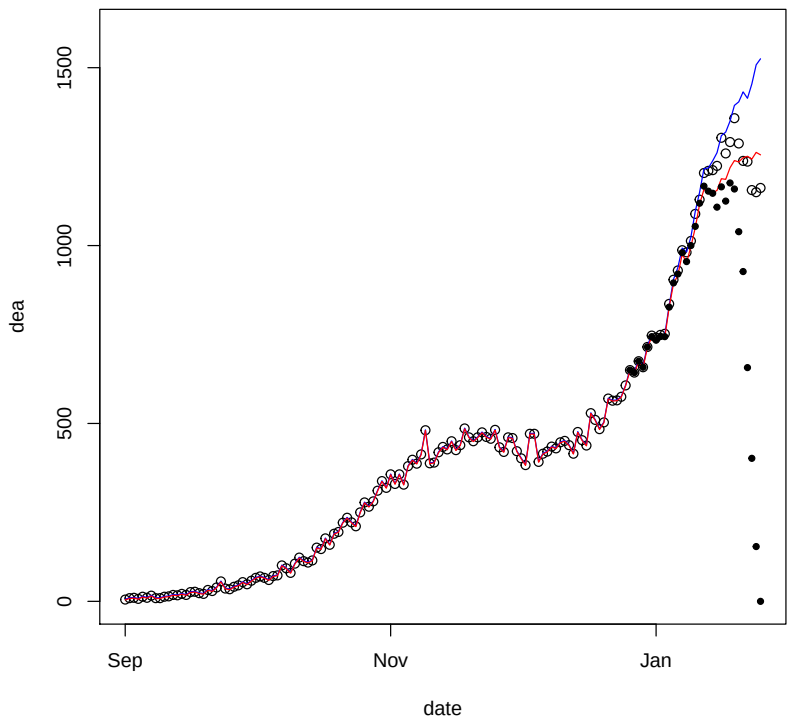




Gaussian processes issues

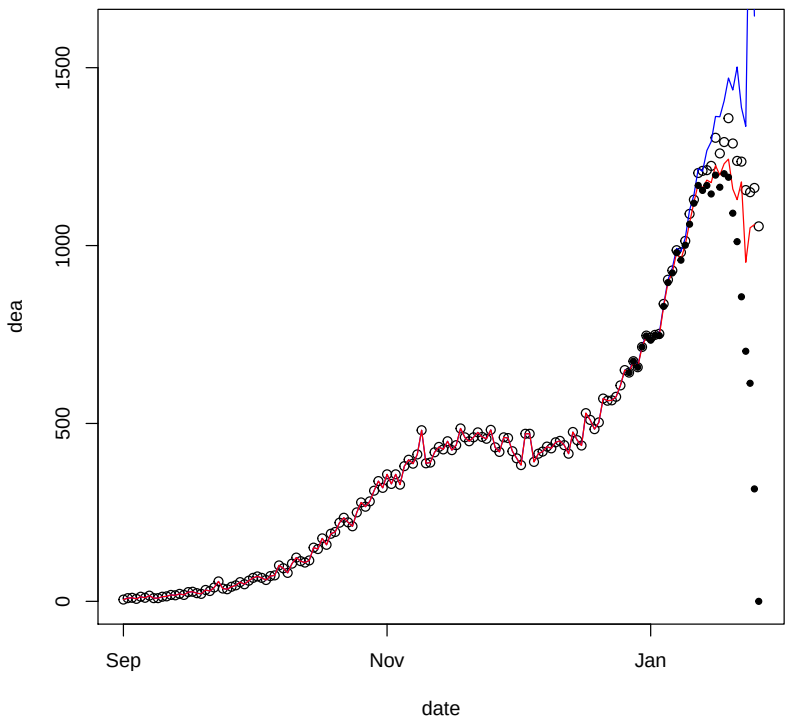
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

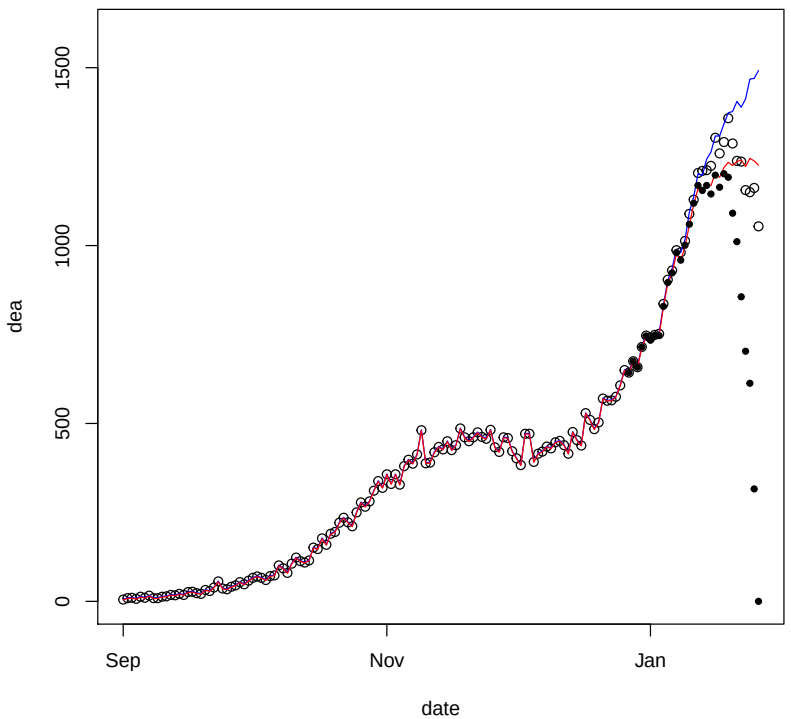




Gaussian processes issues

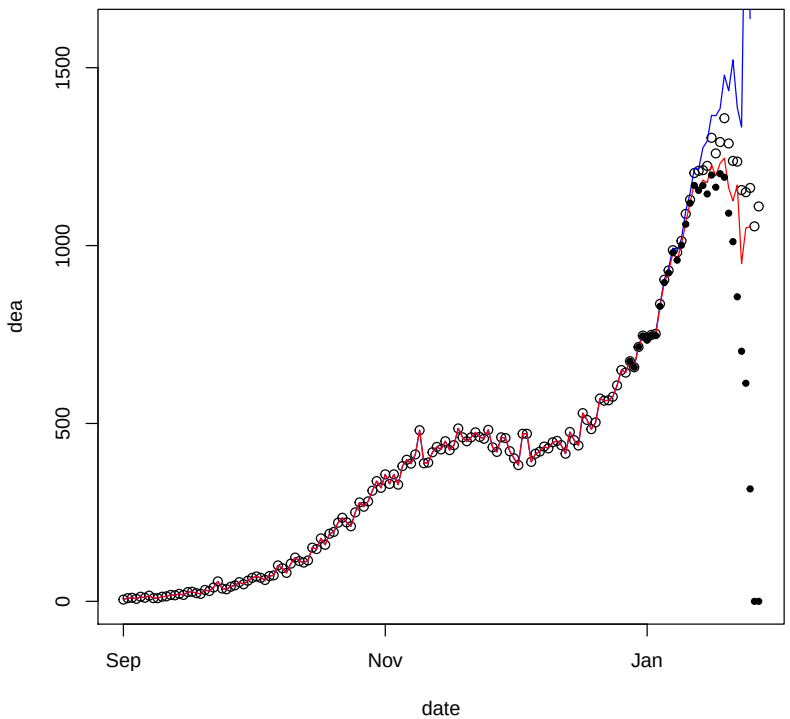
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

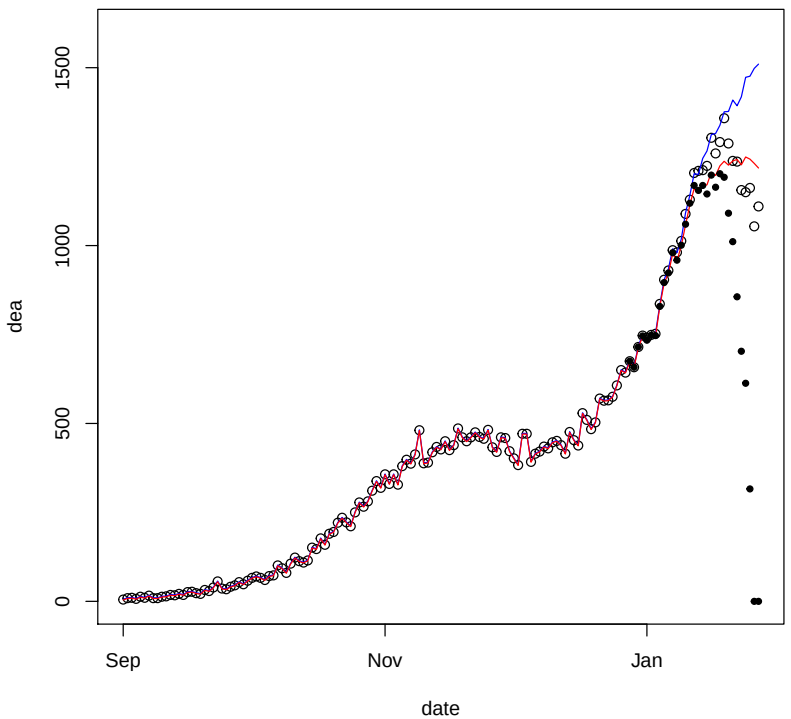




Gaussian processes issues

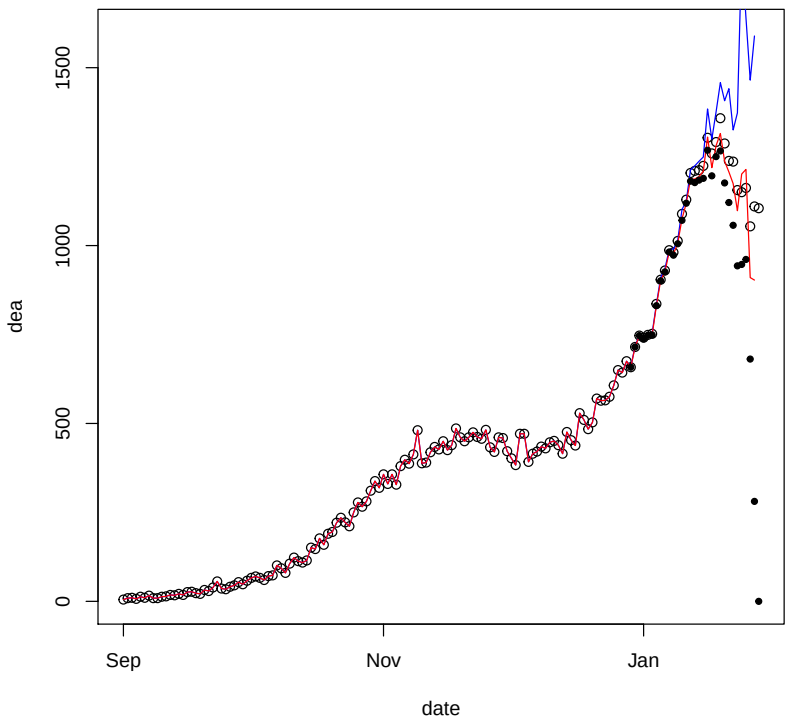
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

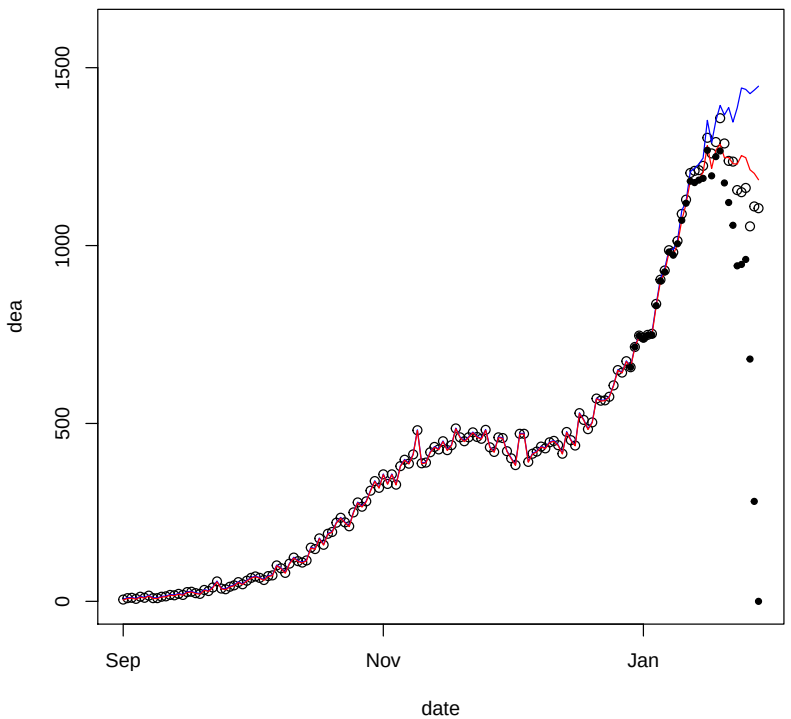




Gaussian processes issues

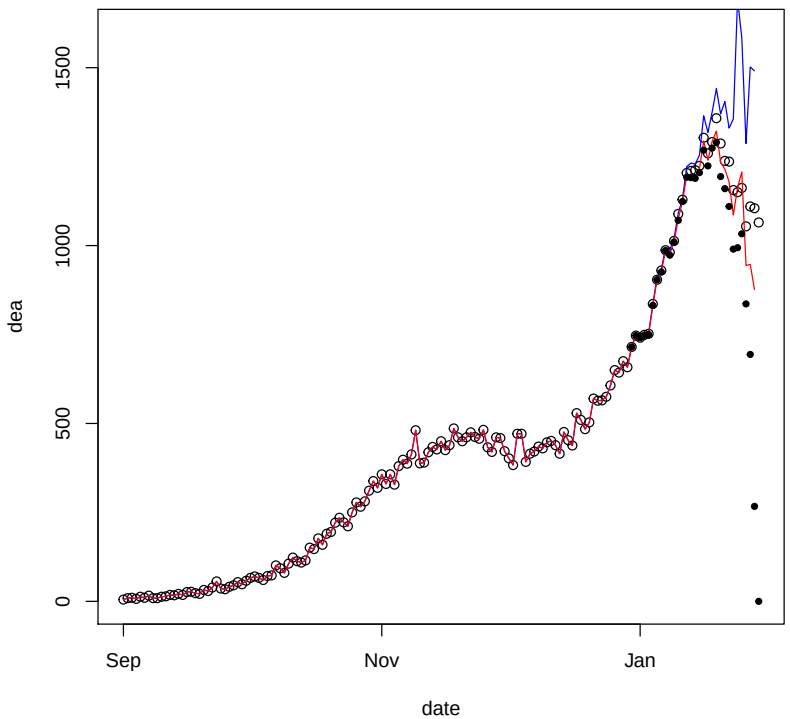
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

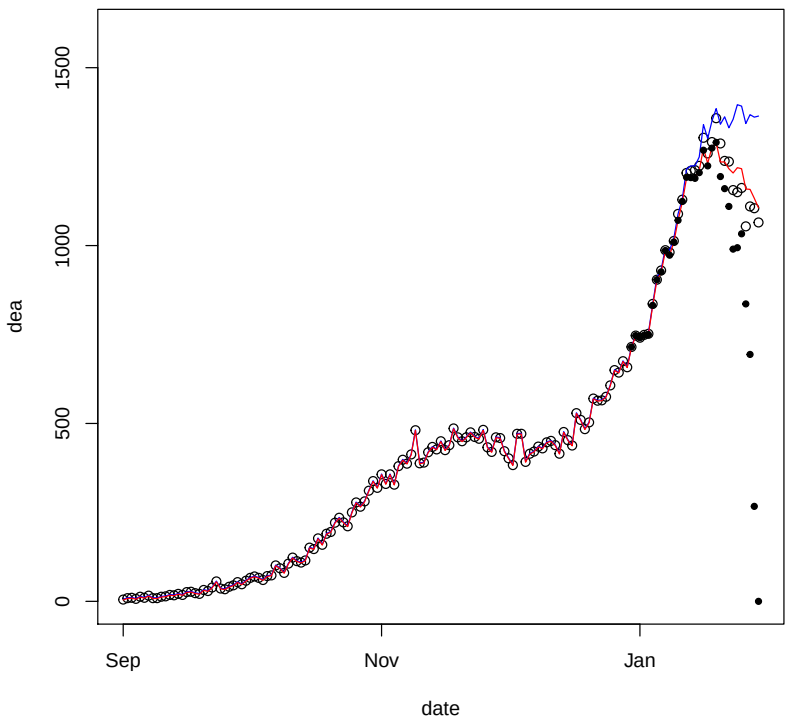




Gaussian processes issues

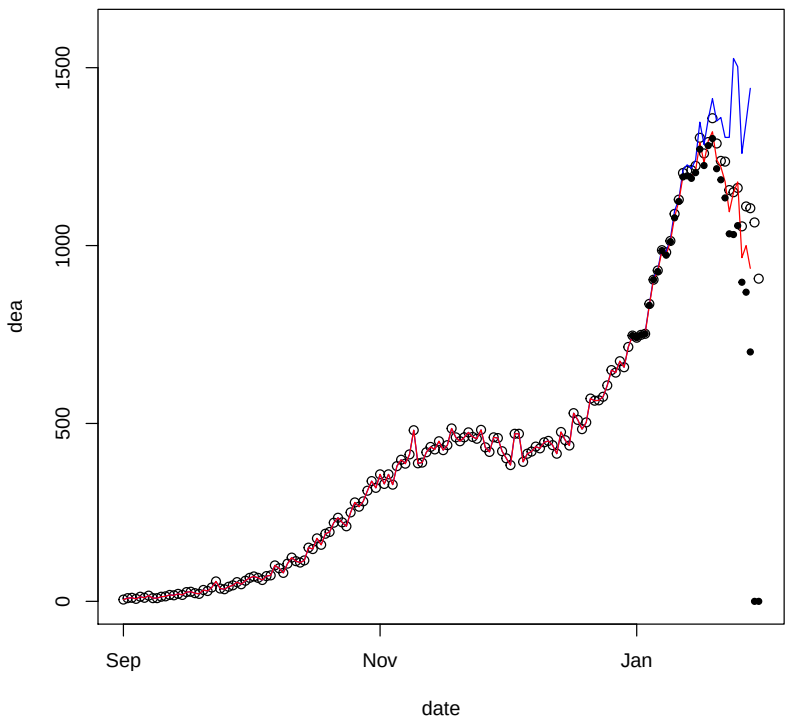
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

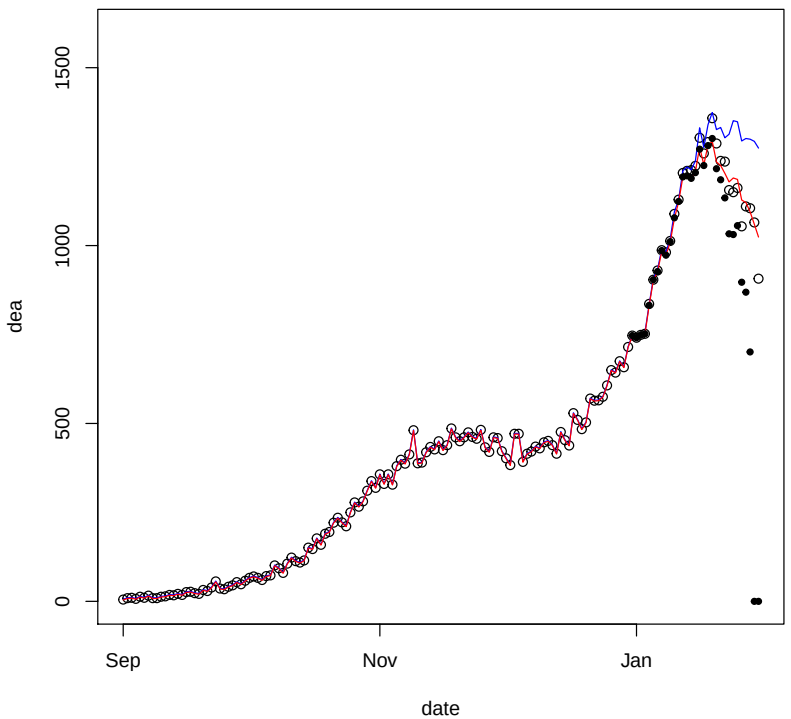




Gaussian processes issues

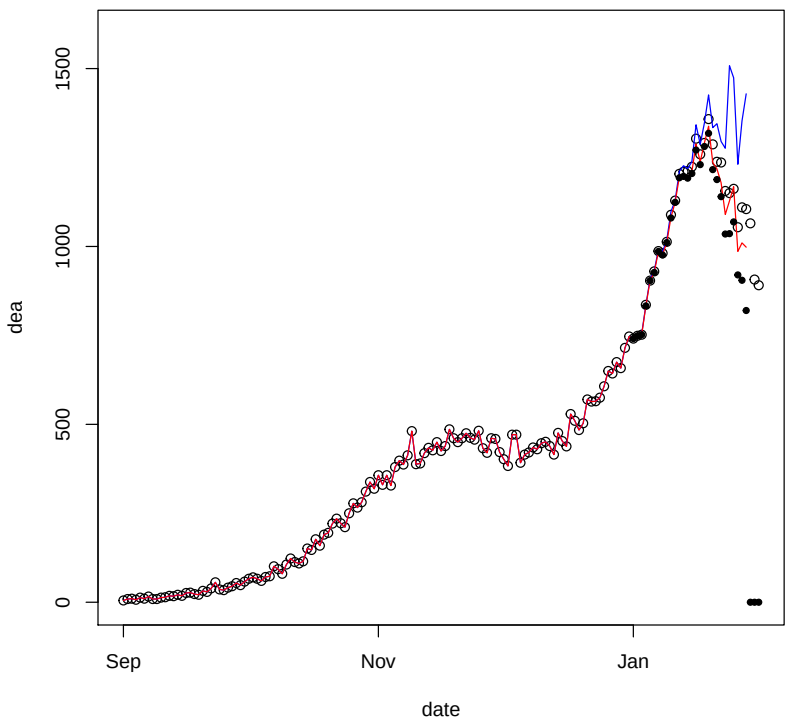
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.

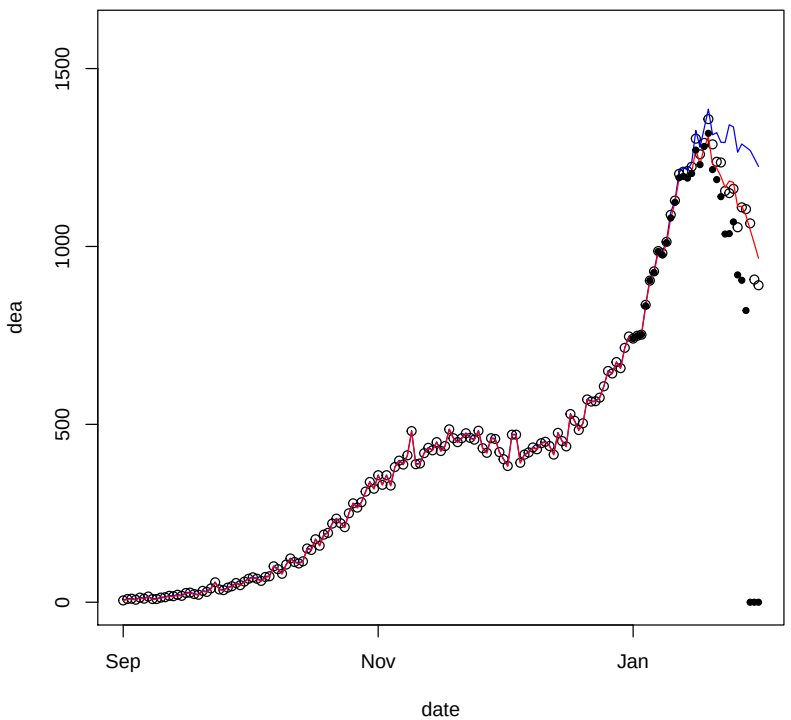




Gaussian processes issues

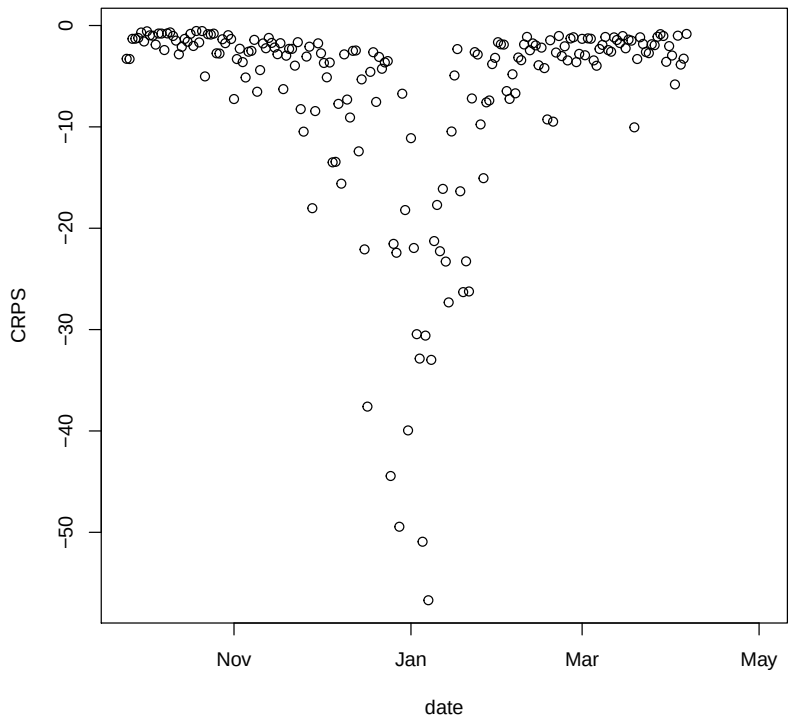
- The GP assumes a certain (we fit from data) a certain smoothness of the latent process.



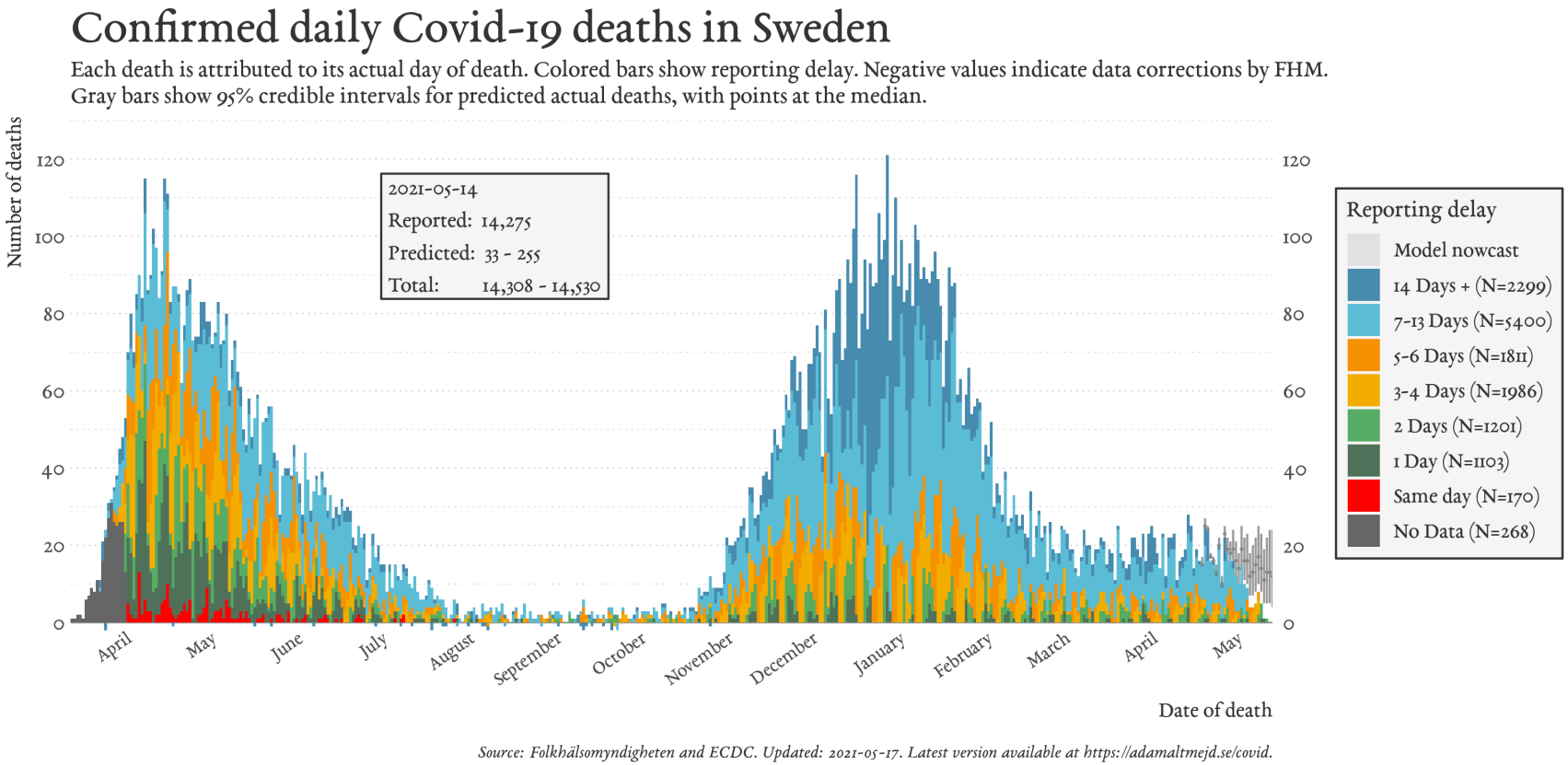


further issues

- We are using a metric for prediction fit known as continuous probability rank score (CRPS).



Sweden model final



Uk data final