A modeling approach to forecasting data with reporting delay

24º SINAPE | Simpósio Nacional de Probabilidade e Estatística

Izabel Nolau

August, 2022

Instituto de Matemática
Universidade Federal do Rio de Janeiro

Joint work with Dani Gamerman (UFRJ/UFMG) and Leonardo S Bastos (FIOCRUZ)

Motivation

Monitoring the spread of illnesses through a surveillance system is essential!

A surveillance system aims to:

$$\text{observe} \longrightarrow \text{predict} \longrightarrow \begin{array}{c}\text{increase knowledge}\\ \text{and}\\ \text{minimize damage}\end{array}$$

and should present:

- sensitivity
- specificity
- timeliness

Lack of timeliness may be due to:

- laboratory confirmation
- logistical problems
- infrastructure difficulties

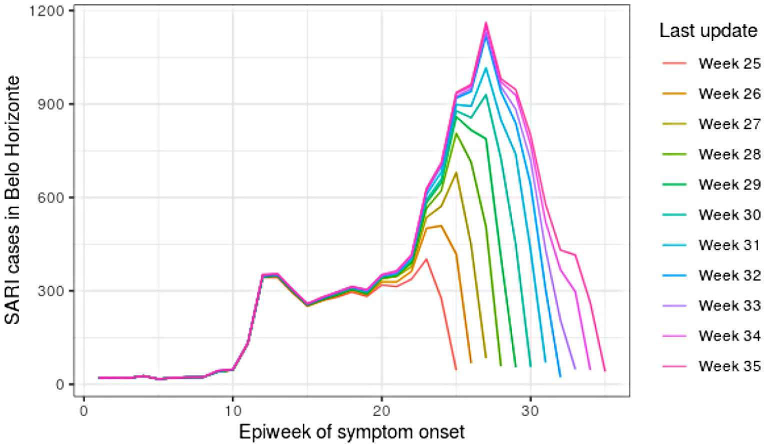
The difference between the reported and the true disease incidence varies according to the reporting delays

This is a problem where the observable data will eventually become available

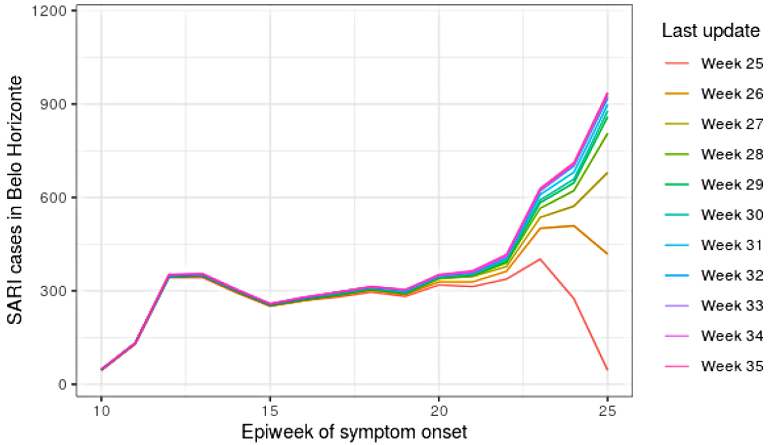
! observed data $\neq$ truth

The reporting delay problem

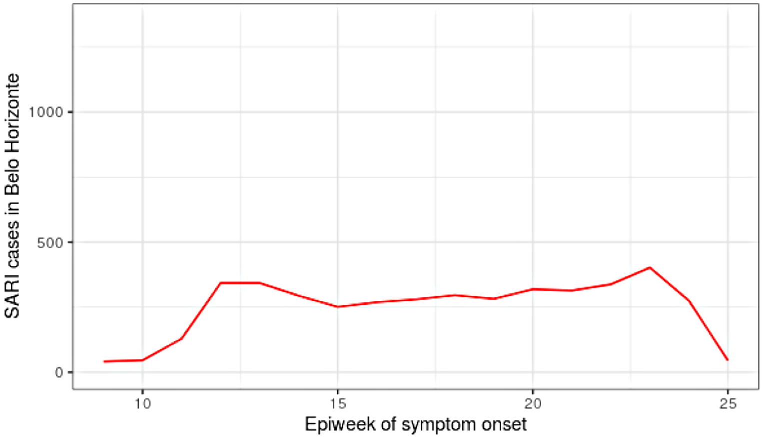
SARI cases in Belo Horizonte, Brazil by reporting date



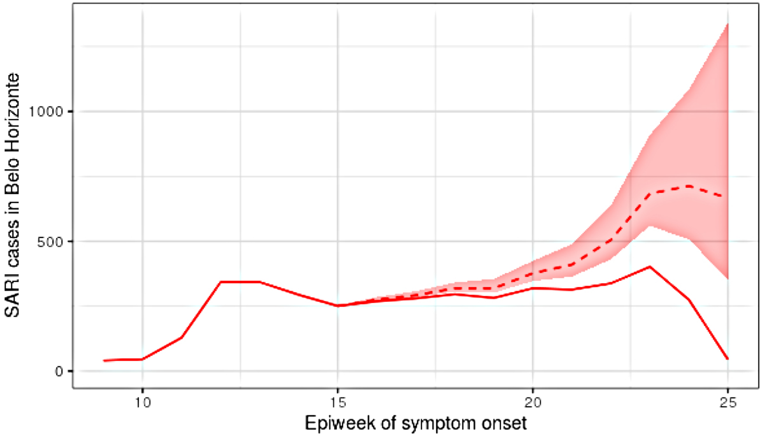
SARI cases in Belo Horizonte by reporting date up to week 25



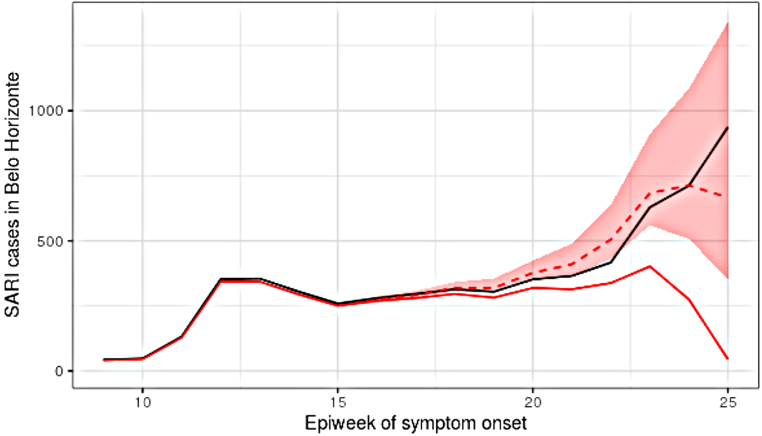
Suppose today is 20/6/2020 (week 25)



We aim to predict the occur-but-not-yet-reported cases



and in the future, compare with what actually happened



Now + Forecasting = Nowcasting!

There are different approaches in the literature:

- Regression-like approach using proxies (Ginsberg et al., 2009)
- Model historic delay to correct present data (Brookmeyer and Damiano, 1989)
- Bayesian hierarchical modeling approach (Bastos et al., 2019)

In these models, there is no specific component relating to the disease dynamics that collaborate toward forecasting

We aim to propose a Bayesian model that enables nowcasting and forecasting!

- T: current time (today)
- t: time index, varying in $\{1, 2, \ldots, T + H\}$
- D: maximum relevant delay
- d: delay index, varying in $\{k, k + 1, \ldots, D\}$
- $n_{t,d}$: number of events occurred at time t recorded after d units of time
- $N_t = \sum_{d=k}^{D} n_{t,d}$: total number of events occurred at time t

! when there is concomitant available information, set $k = 0$

$t \backslash d$	0	1	$\cdots$	$D-1$	D	N
1	$n_{1,0}$	$n_{1,1}$	$\cdots$	$n_{1,D-1}$	$n_{1,D}$	N_1
2	$n_{2,0}$	$n_{2,1}$	$\cdots$	$n_{2,D-1}$	$n_{2,D}$	N_2
3	$n_{3,0}$	$n_{3,1}$	$\cdots$	$n_{3,D-1}$	$n_{3,D}$	N_2
$\vdots$			$\cdots$			$\vdots$
$T-D$	$n_{T-D,0}$	$n_{T-D,1}$	$\cdots$	$n_{T-D,D-1}$	$n_{T-D,D}$	N_{T-D}
$T-D+1$	$n_{T-D+1,0}$	$n_{T-D+1,1}$	$\cdots$	$n_{T-D+1,D-1}$	$n_{T-D+1,D}$	N_{T-D+1}
$\vdots$			$\cdots$			$\vdots$
$T-1$	$n_{T-1,0}$	$n_{T-1,1}$	$\cdots$	$n_{T-1,D-1}$	$n_{T-1,D}$	N_{T-1}
T	$n_{T,0}$	$n_{T,1}$	$\cdots$	$n_{T,D-1}$	$n_{T,D}$	N_T
$T+1$	$n_{T+1,0}$	$n_{T+1,1}$	$\cdots$	$n_{T+1,D-1}$	$n_{T+1,D}$	N_{T+1}
$T+2$	$n_{T+2,0}$	$n_{T+2,1}$	$\cdots$	$n_{T+2,D-1}$	$n_{T+2,D}$	N_{T+2}
$\vdots$			$\cdots$			$\vdots$
$T+H$	$n_{T+H,0}$	$n_{T+H,1}$	$\cdots$	$n_{T+H,D-1}$	$n_{T+H,D}$	N_{T+H}

Tabela 1: Data structure in a reporting delay problem.

$t \backslash d$	k	$k+1$	$\cdots$	$D-1$	D	N
1	$n_{1,k}$	$n_{1,k+1}$	$\cdots$	$n_{1,D-1}$	$n_{1,D}$	N_1
2	$n_{2,k}$	$n_{2,k+1}$	$\cdots$	$n_{2,D-1}$	$n_{2,D}$	N_2
3	$n_{3,k}$	$n_{3,k+1}$	$\cdots$	$n_{3,D-1}$	$n_{3,D}$	N_2
$\vdots$			$\cdots$			$\vdots$
$T-D+k$	$n_{T-D+k,k}$	$n_{T-D+k,k+1}$	$\cdots$	$n_{T-D+k,D-1}$	$n_{T-D+k,D}$	N_{T-D+k}
$T-D+k+1$	$n_{T-D+k+1,k}$	$n_{T-D+k+1,k+1}$	$\cdots$	$n_{T-D+k+1,D-1}$	$n_{T-D+k+1,D}$	$N_{T-D+k+1}$
$\vdots$			$\cdots$			$\vdots$
$T-1$	$n_{T-1,k}$	$n_{T-1,k+1}$	$\cdots$	$n_{T-1,D-1}$	$n_{T-1,D}$	N_{T-1}
T	$n_{T,k}$	$n_{T,k+1}$	$\cdots$	$n_{T,D-1}$	$n_{T,D}$	N_T
$T+1$	$n_{T+1,k}$	$n_{T+1,k+1}$	$\cdots$	$n_{T+1,D-1}$	$n_{T+1,D}$	N_{T+1}
$T+2$	$n_{T+2,k}$	$n_{T+2,k+1}$	$\cdots$	$n_{T+2,D-1}$	$n_{T+2,D}$	N_{T+2}
$\vdots$			$\cdots$			$\vdots$
$T+H$	$n_{T+H,k}$	$n_{T+H,k+1}$	$\cdots$	$n_{T+H,D-1}$	$n_{T+H,D}$	N_{T+H}

Tabela 2: Data structure in a reporting delay problem, with first delay k.

We assume the following structure for N_t

$$\begin{aligned} N_t &\sim NegBin(\theta_t, \phi) \\ \theta_t &= \frac{a_\theta c_\theta f_\theta \exp(-c_\theta t)}{[b_\theta + \exp(-c_\theta t)]^{f_\theta + 1}} \end{aligned}$$

for $t = 1, \ldots, T + H$, such that

$$E[N_t] = \theta_t \quad \text{and} \quad Var[N_t] = \theta_t \left(1 + \frac{\theta_t}{\phi}\right)$$

! when $\phi \to \infty$, the Negative Binomial reduces to the Poisson

! for $t > T - D + k$, N_t is a function of unobserved quantities

We assume the following structure for $n_{t,d}$

$$\begin{aligned} n_{t,d} &\sim NegBin(\lambda_{t,d}, \sigma) \\ \log(\lambda_{t,d}) &= \alpha_t + \beta_d \end{aligned}$$

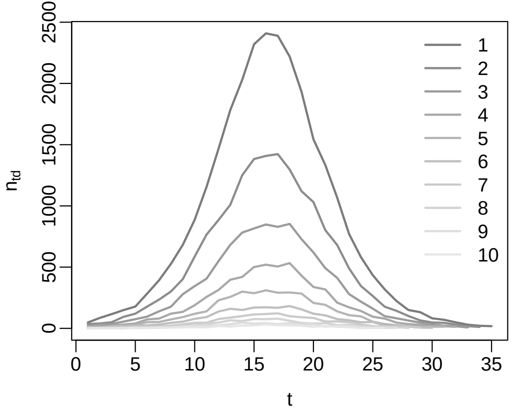
for $t = 1, \ldots, T+H$, $d = k+1, \ldots, D$, where

$$\exp(\alpha_t) = \frac{a_\alpha c_\alpha f_\alpha \exp(-c_\alpha t)}{[b_\alpha + \exp(-c_\alpha t)]^{f_\alpha + 1}} \quad \text{and} \quad \beta_d = \gamma d$$

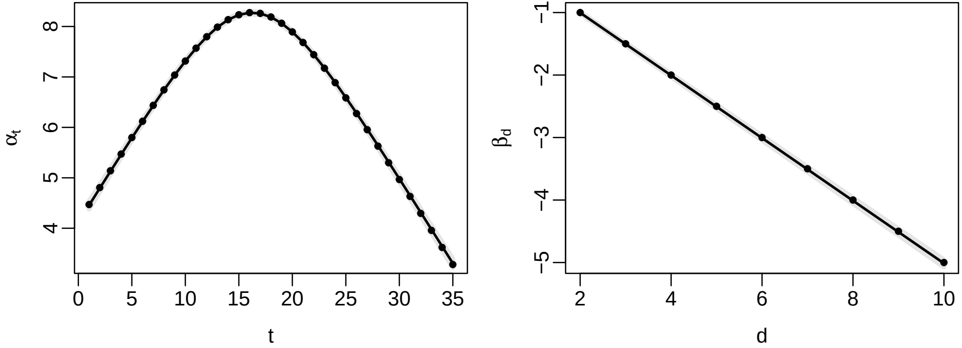
! we do not specify a distribution for $n_{t,k}$ since $n_{t,d} = N_t - \sum_{d=k+1}^{D} n_{t,d}$

! θ_t must be greater than $\sum_{d=k+1}^{D} \lambda_{t,d}$

To assess the effectiveness of the proposed model, artificial data was generated with $\phi \to \infty$ and $\sigma \to \infty$



The results accurately recovered the real parameters' values



t \ d	2	3	4	5	6	7	8	9	10	N_t
27									5 2 (0;6)	567 589 (542;642)
28								4 3 (0;7)	3 2 (0;5)	406 425 (383;468)
29							5 4 (0;8)	4 2 (0;5)	2 1 (0;4)	311 306 (273;343)
30						3 4 (1;9)	6 2 (0;6)	1 1 (0;4)	1 1 (0;3)	209 219 (190;249)
31					8 5 (1;10)	2 3 (0;7)	5 2 (0;5)	2 1 (0;3)	0 0 (0;3)	185 157 (130;182)
32				9 6 (2;11)	3 3 (0;8)	1 2 (0;6)	1 1 (0;4)	0 1 (0;3)	1 0 (0;2)	124 112 (92;133)
33			9 7 (3;13)	8 4 (1;9)	2 2 (0;6)	6 1 (0;4)	1 1 (0;3)	1 0 (0;2)	0 0 (0;2)	83 79 (62;98)
34		8 8 (3;15)	4 5 (2;10)	4 3 (0;7)	2 2 (0;5)	1 1 (0;3)	2 0 (0;3)	0 0 (0;2)	0 0 (0;2)	54 58 (44;73)
35	8 10 (4;16)	3 6 (2;11)	5 4 (0;8)	2 2 (0;5)	2 1 (0;4)	1 1 (0;3)	0 0 (0;2)	0 0 (0;2)	0 0 (0;2)	39 40 (29;55)

Tabela 3: True values (upper number), posterior median (lower number), and respective 95% credibility interval (in parenthesis) for the non-observed counts.

We are proposing a promising model to nowcast and forecast

The model can be used in real-time decision-making as well as in making short-term and long-term predictions

The example shows the model's ability to recover the parameters accurately and nowcast the unobserved values

Future work: extend this model to accommodate more waves

Referências

Bastos, L. S., Economou, T., Gomes, M. F., Villela, D. A., Coelho, F. C., Cruz, O. G., Stoner, O., Bailey, T., and Codeço, C. T. "A modelling approach for correcting reporting delays in disease surveillance data." *Statistics in medicine*, 38(22):4363–4377 (2019).

Brookmeyer, R. and Damiano, A. "Statistical methods for short-term projections of AIDS incidence." *Statistics in Medicine*, 8(1):23–34 (1989).

Gamerman, D. and Lopes, H. F. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC press (2006).

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. "Detecting influenza epidemics using search engine query data." *Nature*, 457(7232):1012–1014 (2009).

Stoner, O., Economou, T., and Drummond Marques da Silva, G. "A hierarchical framework for correcting under-reporting in count data." *Journal of the American Statistical Association*, 114(528):1481–1492 (2019).

Thank You!