Short-term forecasting of total Number of reported COVID-19 cases in South Africa - A Bayesian temporal modeling approch

Belay Birlie Yimer, Ziv Shkedy

Abstract

To be updated.

Author summary

To be updated.

Introduction

In this paper we present (1) South Africa's COVID trajectory to the first 100,000 (22 June 2020) cases and (2) fit a series of non-linear growth models, calibrated to COVID-19 cumulative number of reported case data from 5 March 2020 to 22 June 2020. The models are used to produce short term predictions of the number of reported cases expected for a period of 30 days ahead. These forecasts are generated at the national level

Methods

Data

We downloaded data from Coronavirus COVID-19 (2019-nCoV) Data Repository for South Africa maintained by Data Science for Social Impact research group at the University of Pretoria [ref]. The data repository captures the daily number of new cases, number of tests, number of deaths and recoveries. Our primary outcome of interest was the daily number of newly diagnosed COVID-19 cases and the unit of time used in modelling was a day. We used the daily case reports from March 12, 2020, until February 27, 2021, in our analysis.

Statistical analysis

We considered two widely used temporal models to model the daily number of newly diagnosed COVID-19 cases. We let $Y(t)$ denote the daily number of newly diagnosed COVID-19 cases at time t and $\mu(t)$ represent the expected number of cases at time t. We considered a Negative binomial distribution for $Y(t)$ to account for possible overdispersion. That is, $Y(t) \sim NB(\mu(t), \delta)$, where δ is the overdispersion parameter. We considered two temporal models to capture the trend over time: a random walk of order two ($RW(2)$) and an autoregressive model of order one ($AR(1)$) [1]. We also considerd a $RW(1)$ model, but the model overfits the data (See the supplemntary appendix). Similarly, we considerd an AR model order $p = 2$ but the result is similar to $AR1$ model and we prefer the simpler $AR1$ model.

The $AR(1)$ model [1] is given by,

$$\begin{aligned} Y(t) &\sim NB(\mu(t), \delta) \quad t = 1, \ldots, n, \\ log(\mu(t)) &= \alpha + u_t, \\ u_1 &\sim N(0, \tau_u(1-\rho^2)^{-1}), \\ u_t &= \rho u_{t-1} + \epsilon_t, \quad t = 2, \ldots, n, \\ \epsilon_t &\sim N(0, \tau_\epsilon), \end{aligned}$$

where, α is an intercept, ρ a temporal correlation term (with $|\rho| < 1$) and ϵ_t is a Gaussian error term with zero mean and precision τ_ϵ.

Similarly, the $RW(2)$ model [1] is given by,

$$\begin{aligned} Y(t) &\sim NB(\mu(t), \delta) \quad t = 1, \ldots, n, \\ log(\mu(t)) &= \alpha + u_t, \\ u_t - 2u_{t+1} + u_{t+2} &\sim N(0, \tau_u), \quad t = 2, \ldots, n, \end{aligned}$$

where α is the intercept term as before and τ_u is the precison parameter.

The two models were fitted within the Bayesian framework using *inla* [2]. To complete the specification of both models, we assume the following priors. For the $AR(1)$ model, we denote $\theta_1 = log(\tau_u(1-\rho^2))$ where $\Gamma(10, 100)$ prior is specified for θ_1, and we denote $\theta_2 = log\frac{1+\rho}{1-\rho}$ and assume a $N(0, 0.15)$ prior for θ. Similarly, we represent the precison parameter of $RW(1)$, τ_u, as $\theta = log(\tau_u)$ and assume a $\Gamma(10, 100)$ prior for θ.

To assess the models' accuracy in predicting COVID-19 cases, we present the forecast period's actual observed values and the predicted values. Additionally, the model fits were evaluated by using DIC (Deviance information criteria).

The R-code that we used for our analyses is avaliable at https://github.com/belayb/COVIDincidenceSA/tree/master/COVIDincidenceSA.

Results

Figure 1 presents the daily number of reported COVID-19 cases from 12 March 2020 to 27 February 2021. Similar to elsewhere in the world, South Africa pass through a two-wave pandemic. The pandemic's first peak was on 07 July 2020, where up to 13944 new COVID-19 cases reported, followed by a second peak in January 2021, where more than 21,000 daily cases reported. Figure 2 presents the cumulative number of new reported COVID-19 cases and tests performed. To date, 8,838,937 tests have been conducted, and a total of 1,500,677 cases reported.

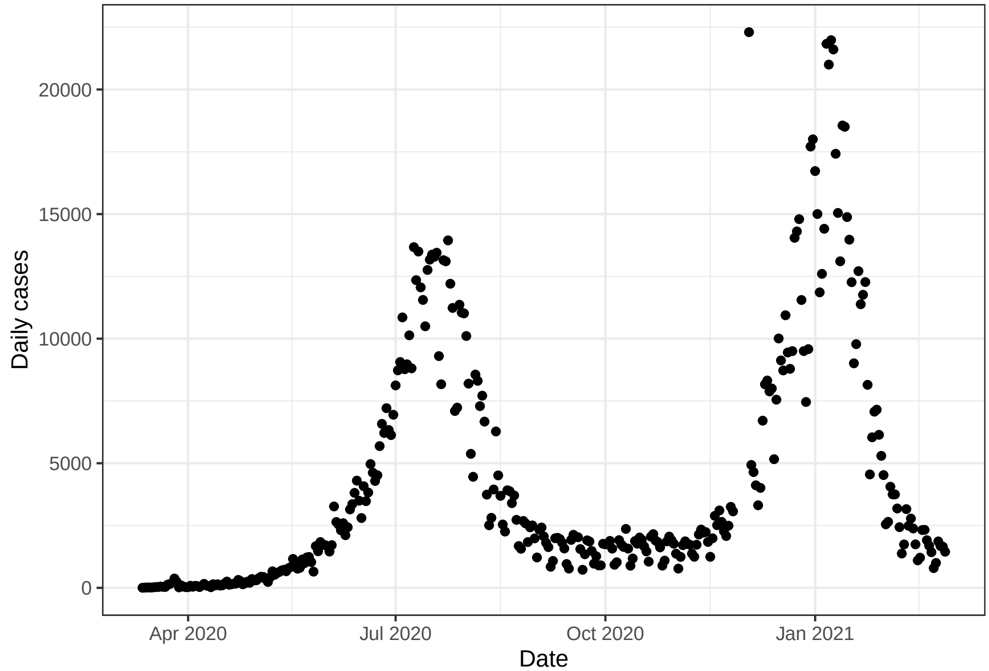

Fig 1. Daily number of COVID-19 cases in South Africa from 12/03/2020-27/02/2021.

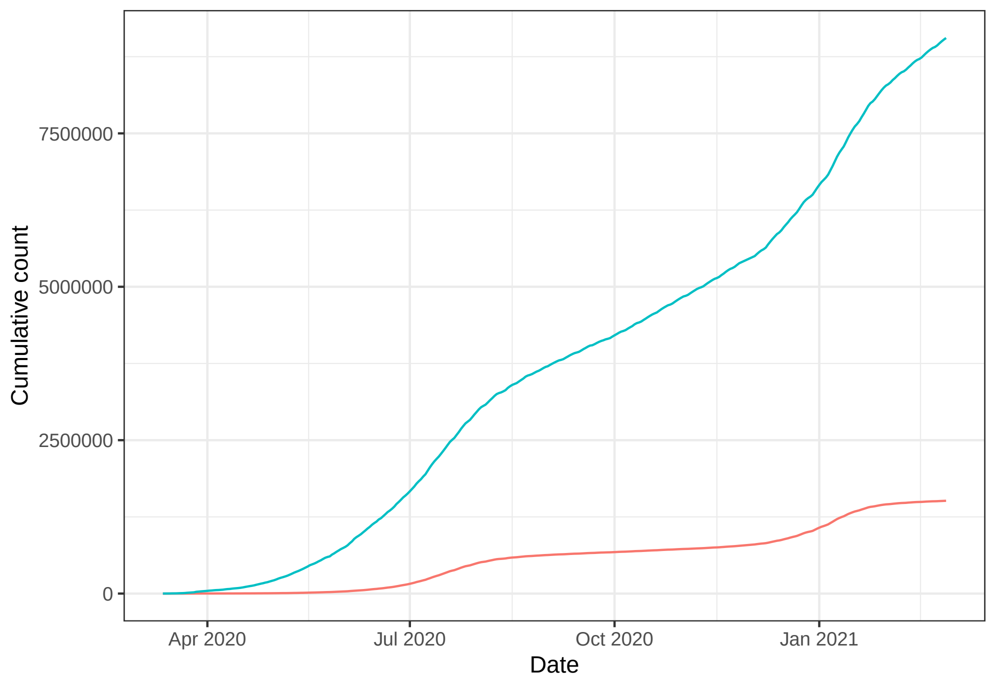

Fig 2. The cummulative number of COVID-19 cases and Cummulative number of tests in South Africa from 12/03/2020-27/02/2021. Red-line denote the number of cases and blue-line denotes the number of tests.

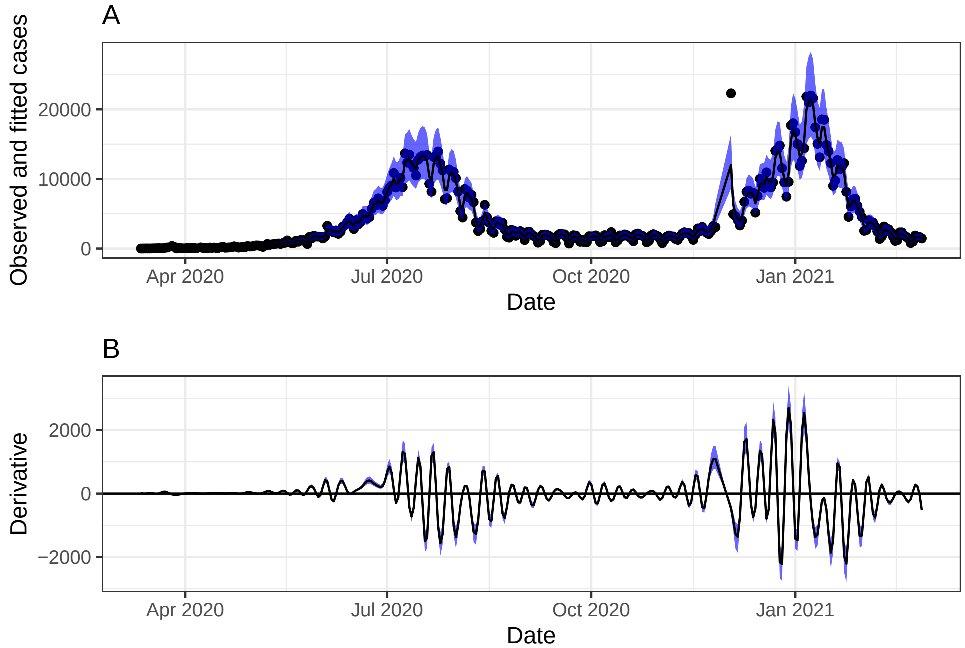

Fig 3. Fitted and observed data AR(1) model

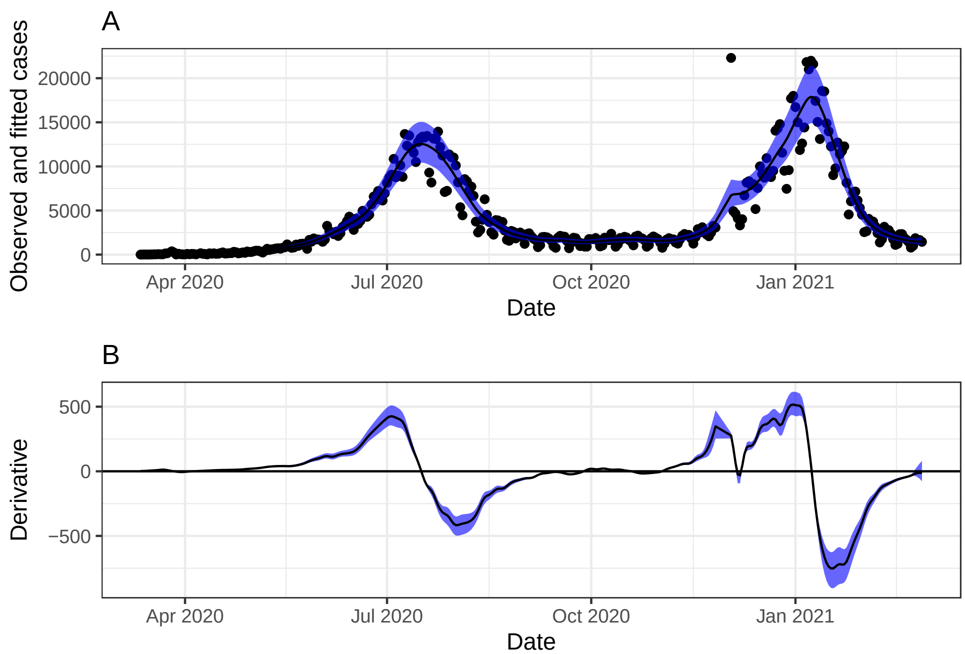

Fig 4. Fitted and observed data RW(2) model

Short-term prediction of the total number of reported COVID-19 cases

We fit the two models described in the previous section to the daily reported new COVID-19 cases. The parameter estimates for the two models are presented in Supplementary Table 1. As depicted in Figure 3, two models fitted to the data appear to fit the observed data (within the estimation period) well with a narrow confidence interval obtained for the $RW(2)$ model. The two models provides similar predictions over the 7-day ahead period.

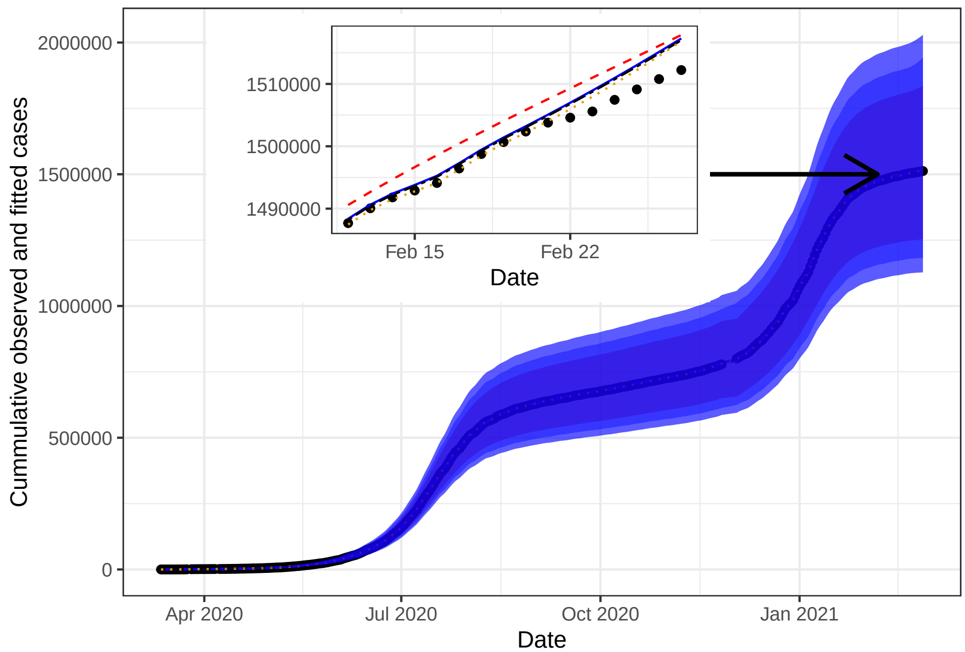

Fig 5. Predicted cummulative COVID-19 cases in South Africa under the RW(1) and AR(1) model. Estimation period 12/03/2020-20/02/2021. The black dots are the observed cummulative cases. The red dashed lines are for RW(2) and the blue line for AR(1).The shaded bands are the prediction intervals.

Table 1. Short-term predictions of total number of reported cases at the national level under the rw2 model. Estimation period 12/03/2020-20/02/2021

Date	Total	Prediction	Prediction Interval	Total - Prediction
2021-02-21	1503796	1507571	(1248627.23-1814986.69)	-3774.807
2021-02-22	1504588	1509293	(1249674.68-1817727.08)	-4705.011
2021-02-23	1505586	1511002	(1250614.28-1820670.52)	-5415.872
2021-02-24	1507448	1512704	(1251449.87-1823859.4)	-5255.732
2021-02-25	1509124	1514406	(1252187.22-1827340.63)	-5281.620
2021-02-26	1510778	1516115	(1252833.39-1831167.8)	-5337.390
2021-02-27	1512225	1517842	(1253396.46-1835402.22)	-5616.872

Table 2. Short-term predictions of total number of reported cases at the national level under the AR1 model. Estimation period 12/03/2020-20/02/2021

Date	Total	Prediction	Prediction Interval	Total - Prediction
2021-02-21	1503796	1505071	(1124486.84-1998035.3)	-1274.995
2021-02-22	1504588	1506982	(1125286.9-2001911.8)	-2393.764
2021-02-23	1505586	1508942	(1125972.18-2006363.99)	-3356.174
2021-02-24	1507448	1510953	(1126572.19-2011372.04)	-3504.598
2021-02-25	1509124	1513014	(1127105.33-2016924.89)	-3889.609
2021-02-26	1510778	1515126	(1127584.21-2023016.99)	-4347.854
2021-02-27	1512225	1517290	(1128017.98-2029646.04)	-5065.020

Table 3. Short-term predictions of total number of reported cases at the national level under the RW1 model. Estimation period 12/03/2020-20/02/2021

Date	Total	Prediction	Prediction Interval	Total - Prediction
2021-02-21	1503796	1504037	(1179288.4-1906832.48)	-241.2192
2021-02-22	1504588	1505956	(1180039.06-1910889.23)	-1367.7046
2021-02-23	1505586	1507956	(1180669.91-1915712.14)	-2370.0667
2021-02-24	1507448	1510042	(1181213.73-1921302.31)	-2593.7729
2021-02-25	1509124	1512217	(1181690.53-1927674.02)	-3092.5479
2021-02-26	1510778	1514484	(1182113.69-1934850.96)	-3706.3254
2021-02-27	1512225	1516849	(1182492.77-1942862.03)	-4624.2376

Table 4. Short-term predictions of total number of reported cases at the national level under the AR2 model. Estimation period 12/03/2020-20/02/2021

Date	Total	Prediction	Prediction Interval	Total - Prediction
2021-02-21	1503796	1504897	(1123980.48-1998091.99)	-1101.411
2021-02-22	1504588	1506802	(1124788.11-2001923.68)	-2213.521
2021-02-23	1505586	1508752	(1125480.87-2006309.88)	-3165.515
2021-02-24	1507448	1510747	(1126088.3-2011227)	-3299.344
2021-02-25	1509124	1512789	(1126628.81-2016661.08)	-3665.250
2021-02-26	1510778	1514878	(1127115.05-2022603.28)	-4099.507
2021-02-27	1512225	1517012	(1127556.18-2029047.89)	-4787.420

Table 5. Parameter estimates AR1 model

	mean	sd	0.025quant	0.975quant
(Intercept)	6.109	2.441	0.145	10.569
Size	28.955	8.389	16.805	49.397
Precision for time	0.158	0.112	0.023	0.438
Rho for time	0.995	0.004	0.985	0.999

Table 6. Parameter estimates RW2 model

	mean	sd	0.025quant	0.975quant
(Intercept)	7.603	0.017	7.570	7.637
Size	10.422	0.911	8.734	12.319
Precision for time	0.036	0.014	0.016	0.070

Internal validation

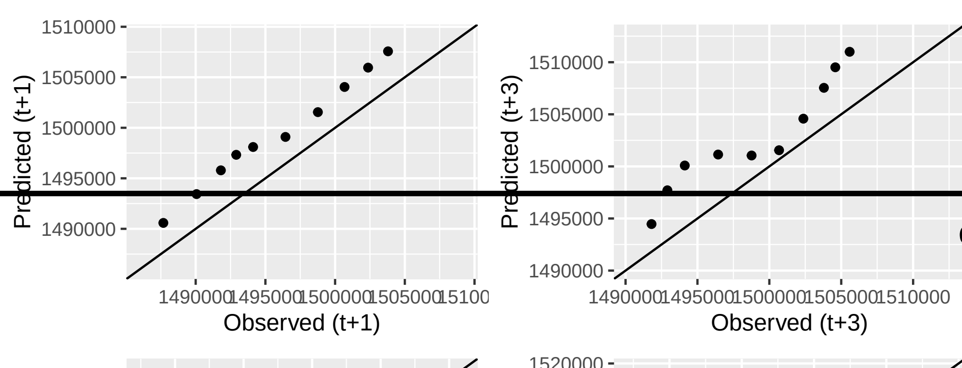

Table 7. Information Creteria for AR1 and RW2 models

	DIC	WAIC
AR1	5278.871	5286.902
RW1	5447.398	5460.805

Table 8. Accuracy metrics of forecasting for AR1, AR2, RW1, and RW2 models for 1-7 days forcasting.

	MAE RW1	MAE RW2	MAE AR1	MAE AR2
One day	449.8629	3482.437	1201.786	1056.635
Two day	1309.7940	3575.328	1871.106	1716.237
Three day	2352.1872	3758.634	2777.020	2636.543
Four day	3328.7112	3890.365	3620.173	3460.322
Five day	4227.9129	3951.368	4348.963	4164.685
Six day	4982.9150	3909.469	4893.645	4679.546
Seven day	5822.3071	3959.652	5479.801	5230.147

Table 9. Accuracy metrics of forecasting for AR1, AR2, RW1, and RW2 models for 1-7 days forcasting.

	MAPE RW1	MAPE RW2	MAPE AR1	MAPE AR2
One day	0.0003009	0.0023281	0.0008036	0.0007065
Two day	0.0008749	0.0023875	0.0012498	0.0011463
Three day	0.0015696	0.0025075	0.0018531	0.0017593
Four day	0.0022192	0.0025927	0.0024135	0.0023069
Five day	0.0028156	0.0026303	0.0028962	0.0027734
Six day	0.0033145	0.0025989	0.0032550	0.0031125
Seven day	0.0038684	0.0026292	0.0036407	0.0034748

Appendix

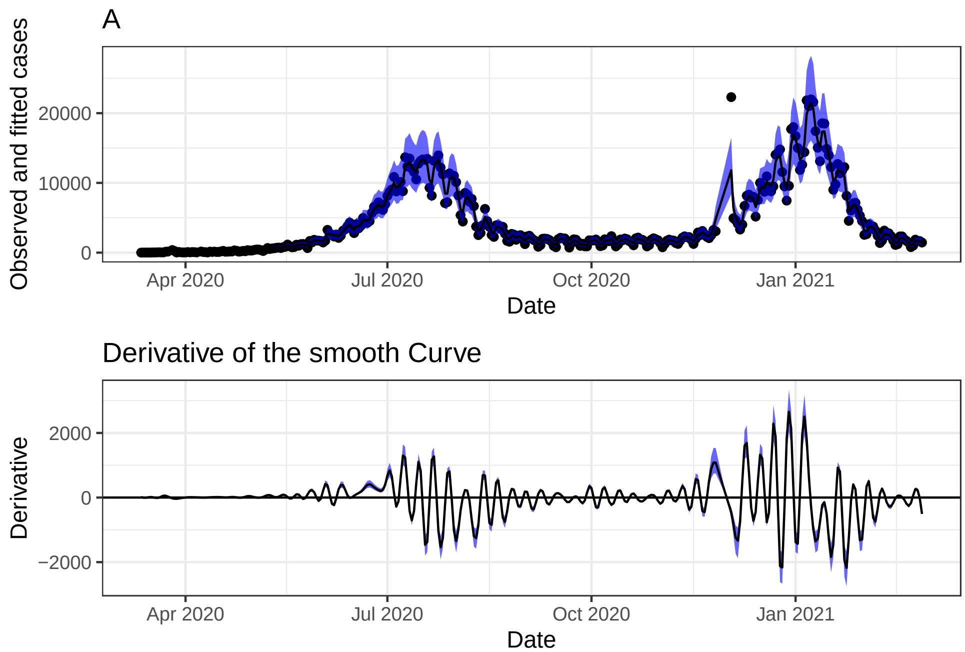

Fig S1. Fitted and observed data AR(2) model

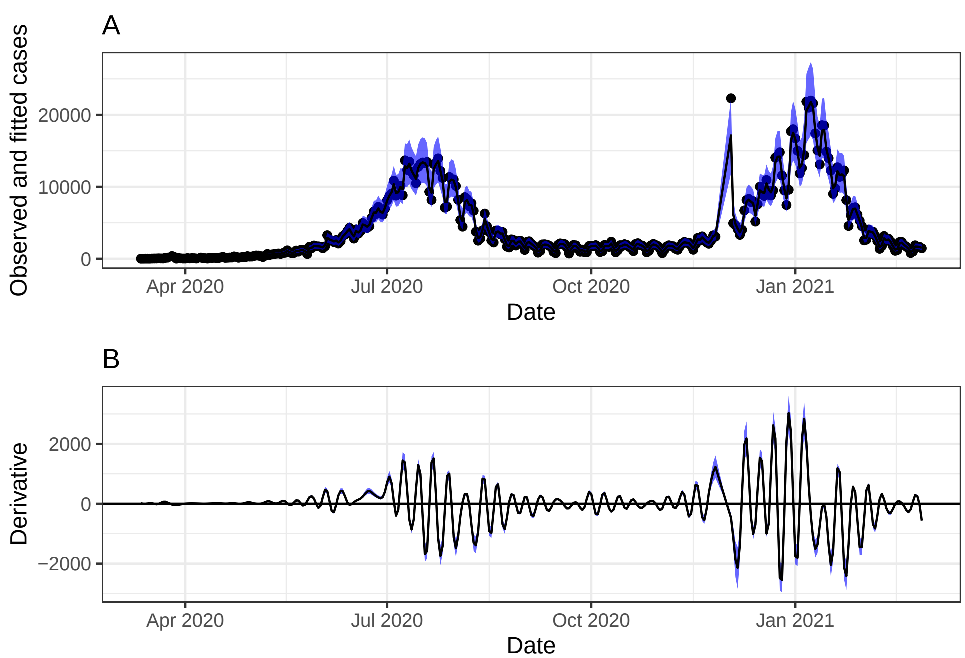

Fig S2. Fitted and observed data RW(1) model

References

1. Gómez-Rubio V. Bayesian inference with inla. CRC Press; 2020.

2. Martins TG, Simpson D, Lindgren F, Rue H. Bayesian computing with inla: New features. Computational Statistics & Data Analysis. Elsevier; 2013;67: 68–83.