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.

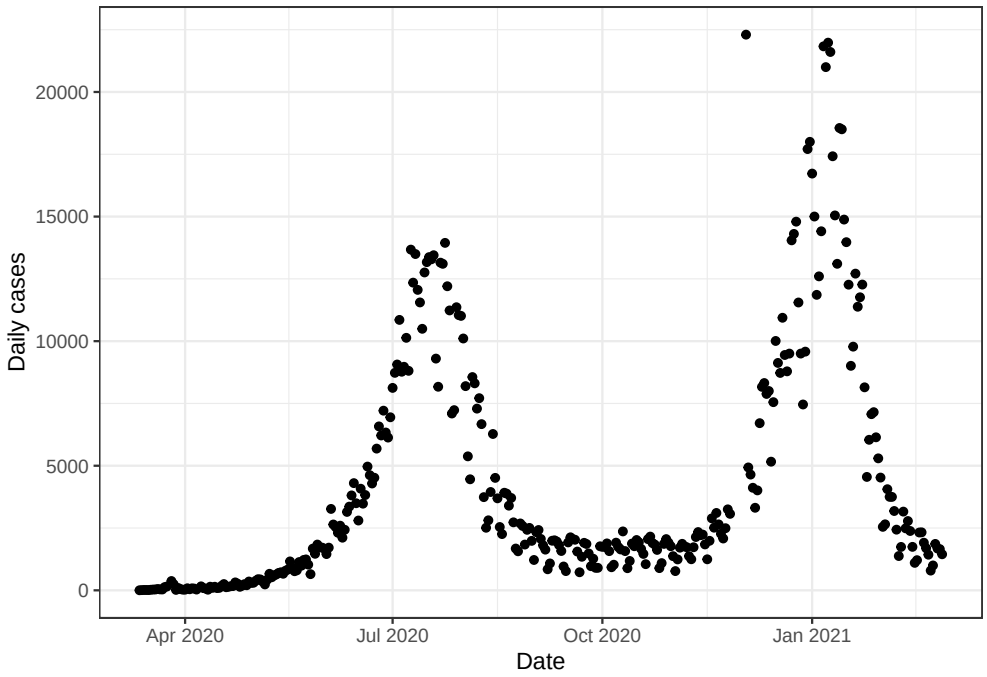

Fig 1. Daily number of COVID-19 cases in South Africa from 12/03/2020-27/02/2021.

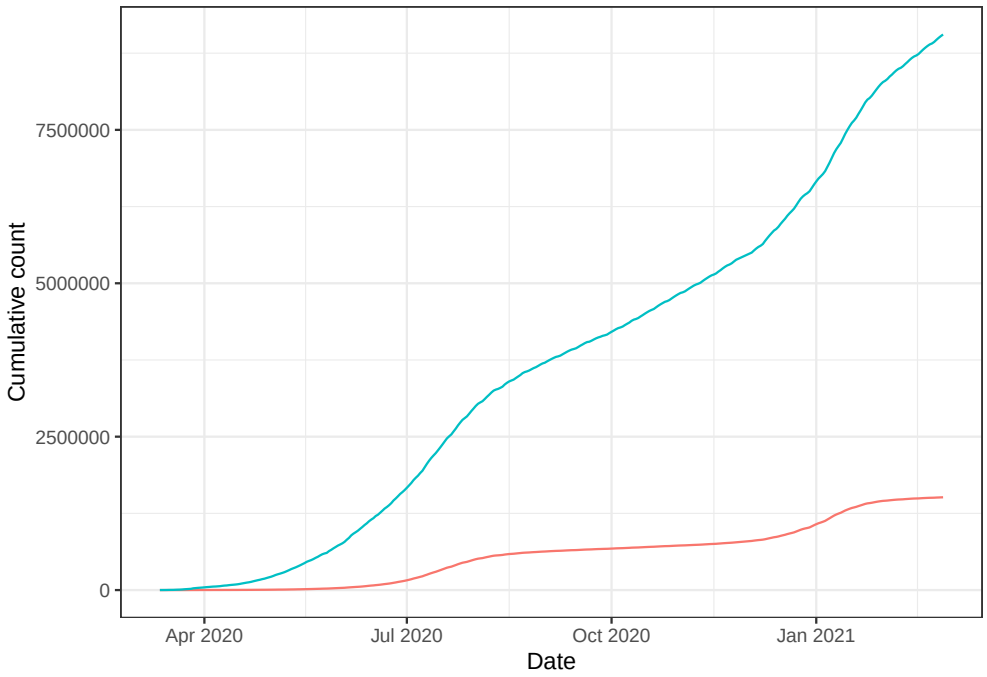

Fig 2. The cummulative number of COVID-19 cases and Cummulative number of tests in South Africa from 12/03/2020-27/02/2021. Red-line denote the number of cases and blue-line denotes the number of tests.

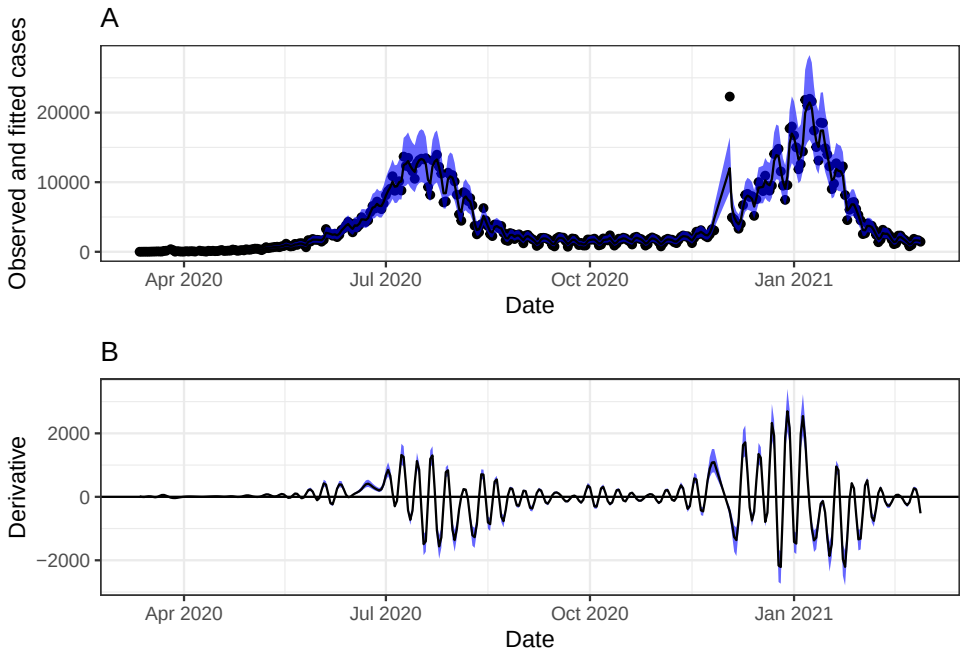

Fig 3. Fitted and observed data AR(1) model

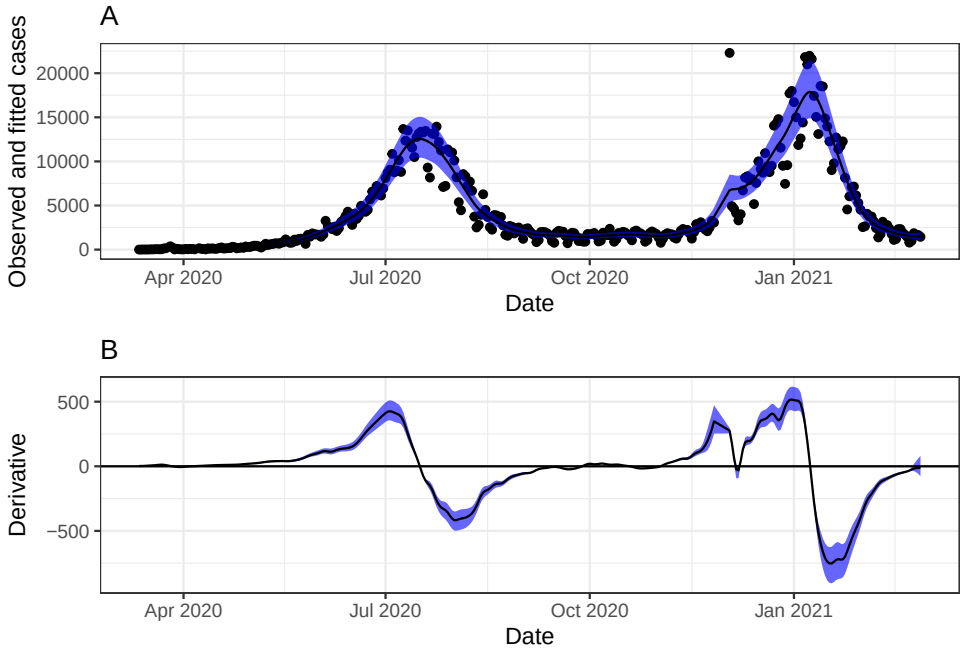

Fig 4. Fitted and observed data RW(2) model

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.

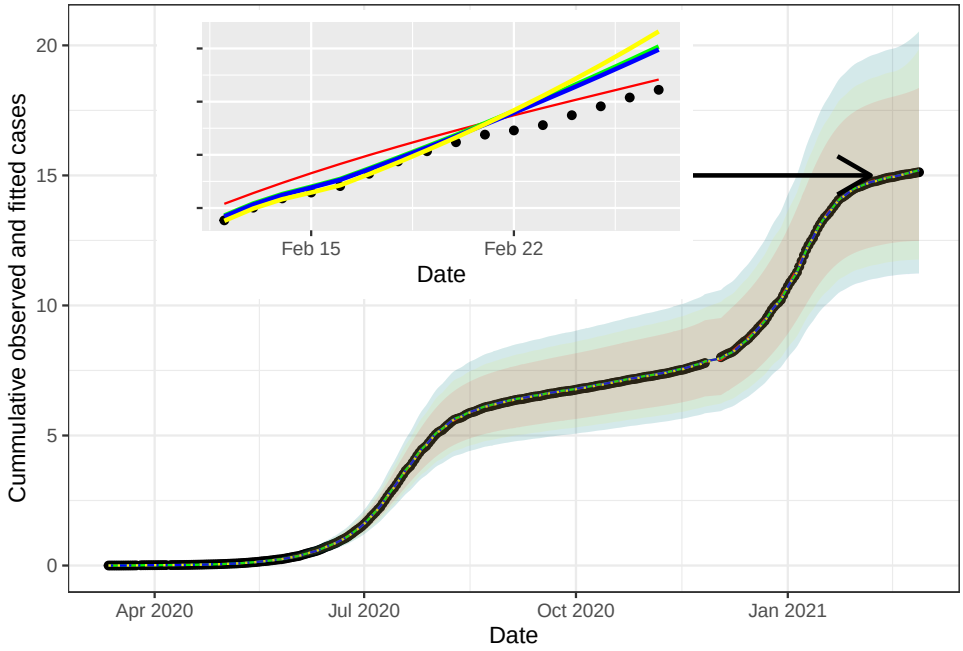

Fig 5. Predicted cummulative COVID-19 cases in South Africa under the RW(1) and AR(1) model. Estimation period 12/03/2020-20/02/2021. The black dots are the observed cummulative cases. The red dashed lines are for RW(2) and the blue line for AR(1).The shaded bands are the prediction intervals.

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
2021-02-18	1498766	1501550	(1243364.14-1808126.44)	-2784.084
2021-02-19	1500677	1503091	(1244303.4-1810572.89)	-2414.449
2021-02-20	1502367	1504581	(1245125.49-1813126.02)	-2213.984
2021-02-21	1503796	1506026	(1245838.81-1815814.63)	-2229.618
2021-02-22	1504588	1507432	(1246452.98-1818667.83)	-2844.351
2021-02-23	1505586	1508808	(1246978.01-1821716.68)	-3222.321
2021-02-24	1507448	1510161	(1247424.13-1824995.35)	-2712.894
2021-02-25	1509124	1511498	(1247800.94-1828542.17)	-2373.747
2021-02-26	1510778	1512827	(1248117.46-1832400.55)	-2048.995
2021-02-27	1512225	1514157	(1248381.83-1836620.02)	-1932.342

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
2021-02-18	1498766	1499490	(1118111.85-1994222.05)	-723.8101
2021-02-19	1500677	1501589	(1119005.83-1998492.97)	-911.8181
2021-02-20	1502367	1503742	(1119771.48-2003403.42)	-1374.5498
2021-02-21	1503796	1505948	(1120441.58-2008928.37)	-2152.4467
2021-02-22	1504588	1508210	(1121036.44-2015054.48)	-3622.1451
2021-02-23	1505586	1510527	(1121570.32-2021775.18)	-4941.3574
2021-02-24	1507448	1512901	(1122053.58-2029087.04)	-5452.8340
2021-02-25	1509124	1515331	(1122494.04-2036989.06)	-6207.3481
2021-02-26	1510778	1517820	(1122897.79-2045481.78)	-7041.6895
2021-02-27	1512225	1520367	(1123269.68-2054566.81)	-8141.6604

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
2021-02-18	1498766	1498569	(1171371.01-1905725.92)	197.3859
2021-02-19	1500677	1500866	(1172282.33-1910642.87)	-188.6823
2021-02-20	1502367	1503262	(1173049.22-1916495.57)	-894.6407
2021-02-21	1503796	1505761	(1173711.18-1923283.63)	-1964.7423
2021-02-22	1504588	1508368	(1174292.27-1931023.86)	-3779.5408
2021-02-23	1505586	1511087	(1174808.38-1939743.02)	-5500.8398
2021-02-24	1507448	1513924	(1175270.86-1949474.01)	-6475.6849
2021-02-25	1509124	1516883	(1175688.37-1960253.97)	-7759.3666
2021-02-26	1510778	1519971	(1176067.58-1972123.43)	-9193.4282
2021-02-27	1512225	1523194	(1176413.76-1985125.8)	-10968.6766

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
2021-02-18	1498766	1499297	(1117410.1-1994516.93)	-530.9910
2021-02-19	1500677	1501378	(1118308.28-1998709.43)	-701.0323
2021-02-20	1502367	1503507	(1119078.99-2003512.61)	-1140.4836
2021-02-21	1503796	1505686	(1119754.71-2008896.75)	-1889.5478
2021-02-22	1504588	1507912	(1120355.57-2014844.05)	-3324.2951
2021-02-23	1505586	1510188	(1120895.7-2021343.34)	-4601.9776
2021-02-24	1507448	1512513	(1121385.41-2028386.98)	-5064.8673
2021-02-25	1509124	1514887	(1121832.47-2035969.53)	-5763.2504
2021-02-26	1510778	1517311	(1122242.95-2044087.07)	-6533.4176
2021-02-27	1512225	1519786	(1122621.69-2052736.67)	-7560.6602

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

Table 7. Information Creteria for AR1 and RW2 models

	DIC	WAIC
AR1	5278.871	5286.902
RW1	5447.397	5460.804

Internal validation

59

Table 8. Accuracy metrics of forecasting for AR1, AR2, RW1, and RW2 models for 1-10 days forcasting.

	MAE RW1	MAE RW2	MAE AR1	MAE AR2
One day	566.3424	3403.057	1100.229	975.1169
Two day	1395.5437	3048.163	1503.359	1412.4527
Three day	2310.3943	2756.451	2273.296	2200.8456
Four day	3224.4373	2669.102	3057.519	2971.4183
Five day	4077.2533	2722.422	3756.599	3653.5679
Six day	4678.4217	2760.629	4178.428	4055.0046
Seven day	5122.3159	2691.440	4415.616	4268.1508
Eight day	5898.4226	2928.807	5036.350	4844.4758
Nine day	6041.7974	2896.331	5119.785	4921.4926
Ten day	6219.2886	2884.664	5229.782	5024.2168

Table 9. Accuracy metrics of forecasting for AR1, AR2, RW1, and RW2 models for 1-10 days forcasting.

	MAPE RW1	MAPE RW2	MAPE AR1	MAPE AR2
One day	0.0003802	0.0022844	0.0007385	0.0006546
Two day	0.0009357	0.0020427	0.0010073	0.0009465
Three day	0.0015471	0.0018446	0.0015218	0.0014734
Four day	0.0021567	0.0017841	0.0020448	0.0019873
Five day	0.0027239	0.0018177	0.0025094	0.0024406
Six day	0.0031217	0.0018410	0.0027877	0.0027055
Seven day	0.0034132	0.0017926	0.0029418	0.0028437
Eight day	0.0039274	0.0019499	0.0033531	0.0032256
Nine day	0.0040218	0.0019282	0.0034079	0.0032762
Ten day	0.0041386	0.0019204	0.0034802	0.0033437

Table 10. Accuracy metrics of forecasting for AR1, AR2, RW1, and RW2 models for 1-10 days forcasting.

	Chisq RW1	Chisq RW2	Chisq AR1	Chisq AR2
One day	3.519941	80.43338	10.14578	7.642623
Two day	16.537758	71.66416	22.21648	19.346128
Three day	41.855697	70.77741	46.44967	42.301887
Four day	79.209191	71.26839	79.45425	72.947524
Five day	127.990433	72.73753	119.51466	109.013108
Six day	185.907637	72.10287	162.88745	146.770096
Seven day	254.232830	71.57873	210.31240	187.416759
Eight day	293.941945	75.30636	235.73996	209.342520
Nine day	309.855289	74.35368	242.98123	215.549116
Ten day	333.234490	74.04449	253.91151	225.029683

Appendix

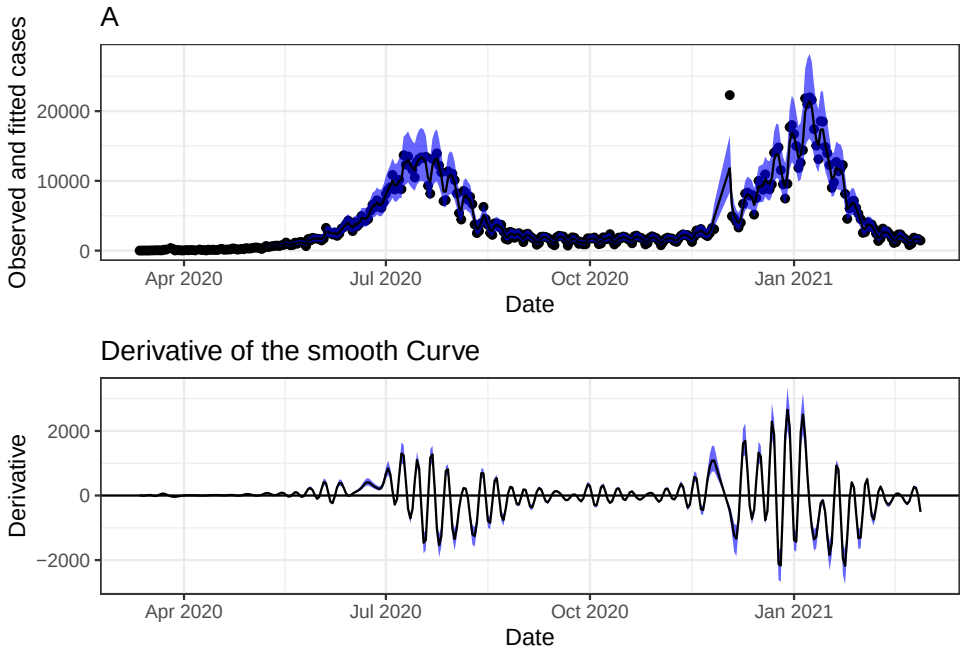

Fig S1. Fitted and observed data AR(2) model

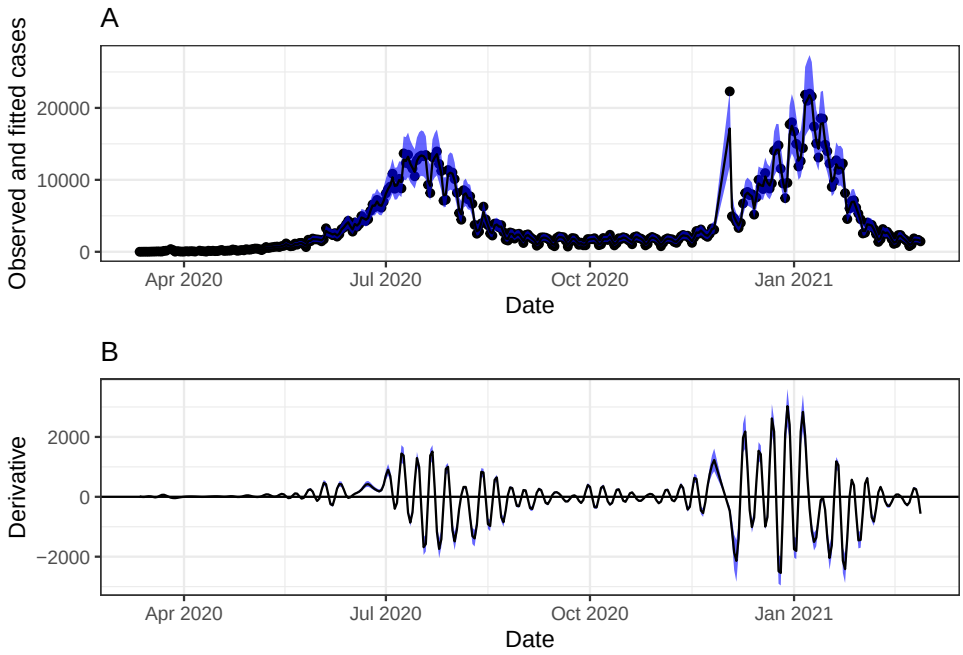

Fig S2. Fitted and observed data RW(1) model

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.