Healthcare Analytics 2 (2022) 100115

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A Bayesian predictive analytics model for improving long range epidemic forecasting during an infection wave

Pedro Henrique da Costa Avelar [a,b,c,d,1,2], Natalia del Coco [a,1], Luis C. Lamb [b], Sophia Tsoka [c], Jonathan Cardoso-Silva [a,c,e,*,1]

[a] Data Science Brigade, Porto Alegre, Rio Grande do Sul, Brazil
[b] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil
[c] Department of Informatics, King's College London, London, United Kingdom
[d] Machine Intellection Department, Institute for Infocomm Research, A*STAR, Singapore
[e] Data Science Institute, London School of Economics and Political Science, London, United Kingdom

ARTICLE INFO

Keywords:
Predictive analytics
Bayesian model
Epidemic forecasting
Infection wave
COVID-19

ABSTRACT

Following the outbreak of the coronavirus epidemic in early 2020, municipalities, regional governments and policymakers worldwide had to plan their Non-Pharmaceutical Interventions (NPIs) amidst a scenario of great uncertainty. At this early stage of an epidemic, where no vaccine or medical treatment is in sight, algorithmic prediction can become a powerful tool to inform local policymaking. However, when we replicated one prominent epidemiological model to inform health authorities in a region in the south of Brazil, we found that this model relied too heavily on manually predetermined covariates and was too reactive to changes in data trends. Our four proposed models access data of both daily reported deaths and infections as well as take into account missing data (e.g., the under-reporting of cases) more explicitly, with two of the proposed versions also attempting to model the delay in test reporting. We simulated weekly forecasting of deaths from the period from 31/05/2020 until 31/01/2021, with first week data being used as a cold-start to the algorithm, after which we use a lighter variant of the model for faster forecasting. Because our models are significantly more proactive in identifying trend changes, this has improved forecasting, especially in long-range predictions and after the peak of an infection wave, as they were quicker to adapt to scenarios after these peaks in reported deaths. Assuming reported cases were under-reported greatly benefited the model in its stability, and modelling retroactively-added data (due to the "hot" nature of the data used) had a negligible impact on performance.

1. Introduction

The World Health Organization (WHO) declared COVID-19 a global pandemic in mid-March 2020, prompting countries to take actions to reduce the spread of the virus in view of the serious respiratory problems that require specialised care in Intensive Care Units (ICU) [1,2]. Little was known about this new strain of coronavirus that threatened to overwhelm health systems, had forced several countries to lockdown and had already been rapidly spreading across Brazil [3].

In many countries, measures such as self-isolation, border closures, testing and social distancing were proposed [4], and it was said that "these protective measures are crucial to managing this disease", although vaccines were considered to be critical [5]. With no drug treatments or vaccines in sight at that time and in the face of a lack of national measures to prevent the spread of the disease [6,7], governors and mayors had to decide independently on the implementation of non-pharmacological measures (NPI) [8]. Amid this scenario and despite the harsh inherent challenges of epidemic modelling, mathematical models offered a timely approach to help understand the regional dynamics of contagion of the disease and to predict how this health crisis could unfold in the weeks and months that followed [9–18]

The MRC Center for Global Infectious Disease Analysis group at Imperial College London introduced a prominent mathematical model in March 2020 [19,20], along with the source code and a technical description of the equations. This model sought, above all, to estimate the impact and effectiveness of NPI measures taken by European countries at that moment. Nonetheless, the model produces other results of interest, such as an estimated number of people infected by SARS-CoV-2 and the variations in the reproduction number (R_t) up until the current date.

* Corresponding author at: Data Science Institute, London School of Economics and Political Science, London, United Kingdom.
E-mail address: J.Cardoso-Silva@lse.ac.uk (J. Cardoso-Silva).
[1] Author no longer holds an affiliation with Data Science Brigade.
[2] Author no longer holds an affiliation with Federal University of Rio Grande do Sul.

https://doi.org/10.1016/j.health.2022.100115
Received 24 June 2022; Received in revised form 17 August 2022; Accepted 26 September 2022

1.1. Contributions

In this paper, we present four variations to the Flaxman et al. model (here called `base` model) that forecast deaths by COVID-19 and overcome limitations we observed after replicating the model on a weekly basis to the seven macro-regions that compose the state of Santa Catarina, a southern state in Brazil. Work on this project started in March 2020, and as we replicated the algorithm every week, we noticed that the original model could not accurately identify the wave pattern so characteristic in an epidemic death curve. The `base` model seemed to assume a linear tendency when forecasting deaths; if deaths had been increasing for the past couple of weeks, they would keep increasing in the following weeks, and vice versa. We conjectured this happened because the `base` model did not consider the reported infections, it only used reported deaths as data input, and therefore it could not anticipate the changes in the infection patterns that had led to the reported deaths.

The covariates used in the model were also inadequate because they limited the effective reproductive number R_t to change only at manually predetermined time points, where an NPI measure came into effect [21]. Other obstacles not directly considered by the `base` were the under-notification of infected cases [22], and the delays in test reporting [23], all common problems at the early stage of the pandemic. From the data available in the official database of Santa Catarina state, we could estimate an average delay of 5 days from RT-PCR test collection until the result was available. Consequently, data from the previous week was guaranteed to be incomplete and uninformative. Despite these limitations, the `base` model was used elsewhere to estimate the impact of NPI measures in two Brazilian states [24].

Our proposed methods aim at overcoming the limitations mentioned above, allowing this mathematical model to be used more effectively for forecasting. Additional equations and algorithmic strategies allow the model to use reported cases to estimate deaths by COVID-19.

2. Literature review

In this section, we comment on related literature to our work. The Covid-19 pandemic brought the research community together, and many predictive analytics methods were tested to help solve different problems in the pandemic. One can find statistical and algorithmic solutions to various related logistic and forecasting problems related to the spread of the disease, such as vaccine allocation [5], metaheuristic feature selection methods for detecting Covid-19 [25], Covid-19 detection with medical images [26], to name a few. In our literature review, however, we focus on three research areas that we consider the most relevant to our study. In Section 2.1 we talk about non-Bayesian methods for Covid-19 forecasting; Section 2.2 focuses on describing Bayesian approaches for forecasting and the main differences between them and our model; finally, in Section 2.3 we describe recent related methods that attempt to perform missing case imputation, as a way to overcome under-reporting and/or delays in test reporting.

2.1. Deterministic forecasting

Traditionally, modelling and forecasting in epidemiology are informed by the use of compartmental models (e.g. SIR, SEIR), where individuals are placed in compartments according to their status (Susceptible, Infectious, Exposed, Recovered, etc.). A recent review, which assessed multiple studies of such type of methods during the COVID-19 pandemic, identified most of these to be "(...) deterministic in nature, by default", with "extensions to stochastic models" being possible [27]. Furthermore, it found that the use of stochastic models was considered to be "more realistic than deterministic models" since deterministic models are valid only where there is a sufficiently large population [28].

But one can also find studies of deterministic approaches that are outside the traditional compartmental literature. For example, the established field of time-series forecasting [29] inspired the models of Parbat and Chakraborty [30] and Sharin et al. [31]. Upon closer look, however, these methods still have room for improvement in their methodology. The cross-validation employed therein was a simple k-fold, whereas standard practice for time series forecasting is to employ some temporally-aware train vs test splitting, such as Walk-forward testing [32,33] or last-block evaluation [34]. In fact, our models use a variation of last-block evaluation, to avoid providing unrealistic predictions and to provide a more grounded model validation [34].

Given the general limitations of deterministic models and the issues we highlighted above, we give more attention in our review of the literature to models that are stochastic in nature or to those in which uncertainty is an integral assumption.

2.2. Bayesian forecasting

We also found compartmental models that use Bayesian inference as part of their solution, the same methodological framework used in our method. In it, prior distributions describe the initial assumptions about the values of random variables in the model. Then, with the foundation of Bayes's Theorem, an optimisation algorithm – typically a variation of a Markov Chain Monte Carlo (MCMC) algorithm – finds solutions to the problem that fit the available data while considering the priors plus any extra custom equations that govern the model [35]. In Roda et al. [11], the authors proposed SIR and SEIR Bayesian compartmental models solved by MCMC, and argued that their simpler SIR model was the most accurate. Whereas this method only uses reported cases as data to fit the model, our models consider previous reported cases and deaths, the patterns of mobility in the geographic region during the period, and it also contains variables to account for missing data and under-reporting. A few other similar models also attempt to model the pandemic without explicitly providing periodic forecasting (e.g., [14,15]).

Many Bayesian models were developed with the intent to measure the impact of public policies on the reproduction rate of virus [10, 18,24], or on the economy [16]. Assessing the effectiveness of public policies is not a goal of our study. Instead, we aim to improve *forecasting* of deaths by COVID-19, even where such policy information is not readily available or unreliable, and therefore we do not compare our proposed models with these in detail. The closest of these models to our case would be [24], which still attempts to measure the impact of NPIs in Brazil using the model made available in Flaxman et al. [20]. Thus, we chose these as baselines and explained them in more detail in Section 2.4

2.3. Missing data imputation

None of the abovementioned work tackle two problems central to COVID forecasting in a low-resource country such as Brazil, namely the under-notification of infected cases [22] and delays in test reporting [23]. These problems, common at the early stage of the pandemic, persisted throughout most of the time in our case study . Although vast, the literature on data imputation is geared mainly towards imputing tabular data (e.g., [36]), which is not directly applicable to our case. Hence, we opted for a data-driven approach that explicitly considers under-notification, delays in test reporting, or both when imputing missing data.

A recent paper by de Nicola et al. [37] used Generalised Additive Models (GAMs) [38] implemented in the `mgcv` R package [39] to explicitly model delays in test-reporting. However, this approach only models delays in test reporting while our model models both delays in test reporting as well as under-reporting. Furthermore, their model is too different to what we propose, as it is not a Bayesian statistical model, nor does it consider other covariates, such as mobility patterns, while our proposed methods model the impact of mobility patterns directly.

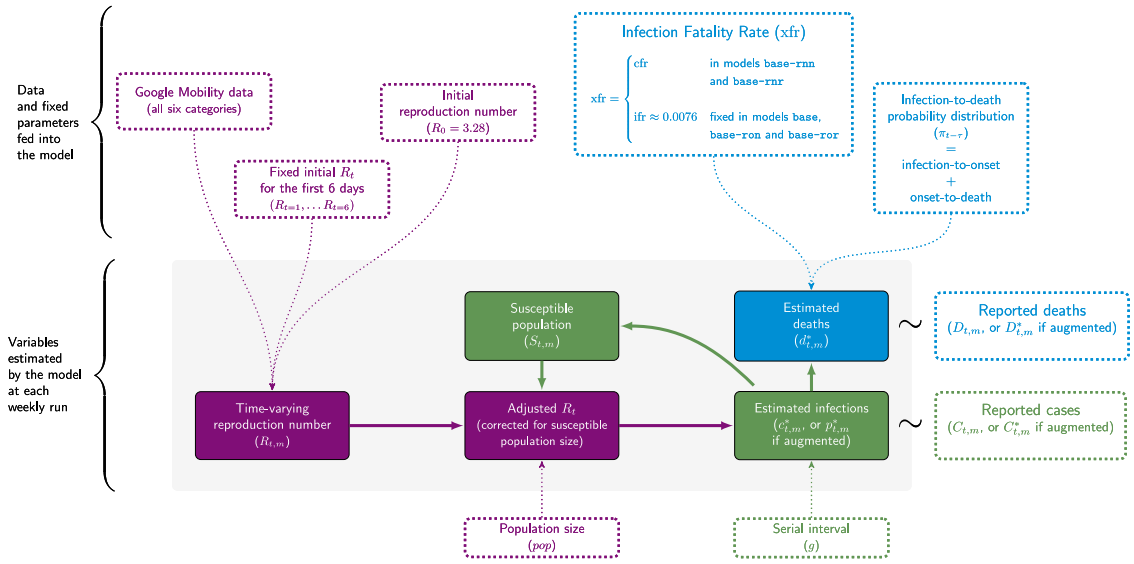


Fig. 1. A high-level depiction of the models proposed and reported in this study, including the baseline model (`base`) and all our four proposed models (`base-ron`, `base-rnn`, `base-ror` and `base-rnr`). Data and fixed parameters passed to the models are represented by the boxes with dashed-line borders, whereas the coloured boxes represent the variables inferred by the models in the middle of the diagram. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.4. Baseline

Due to the abovementioned reasons, we chose Flaxman et al.'s previously-proposed [20] model as our main baseline, referred to us as `base`. This baseline model is almost identical to the MCMC-based model made available by Flaxman et al. [20], except for the covariates used. Instead of manually curated non-pharmaceutical intervention measures, we used Google Mobility data as covariates to the model since our primary goal is to forecast new infections and not measure the effect of NPIs on the reproduction rate of the pandemic.

This strategy also counteracts an implicit confirmation bias of the model [21] and have been used by the original group in a subsequent work [40,41]. In contrast to European countries targeted by the original `base` model, the State of Santa Catarina did not impose strict state-wide measures consistently during this period. Local governments (cities and regional associations of municipalities) were responsible for independently deciding on their social distancing measures [8], which rendered changes in legislation impractical.

3. Proposed models

In this section, we describe the equations and modifications of our proposed models, `base-ron`, `base-rnn`, `base-ror` and `base-rnr`. The reader is referred to Fig. 1 for a high level depiction of the fixed parameters, data and variables used in this section. Each weekly run, all methods receive the same information: the reported cases and deaths, fatality rate, infection-to-death probability distribution, fixed parameters (population size and serial interval), and the Google Mobility covariates. What distinguishes each model is how they treat this data. Whereas the `base` model implicitly treats reported cases as "ground-truth", our proposed variations all "augment" (i.e., impute) the reported data to account for delays in test reporting or under-notification of infections by adding variables to the Bayesian inference model. Table 1 offers for a more technical comparison between `base` model and our proposals.

3.1. Proposed models

Since our Bayesian models rely on the same distributions as the `base` model, we will use the same symbols to mean the same distributions whenever possible. We refer the reader to the original paper for a full description of the symbols [20]. In mathematical terms, we observe daily deaths $D_{t,m}$ and daily cases $C_{t,m}$ for days $t \in \{1, \ldots, n\}$ and geographical regions $m \in \{1, \ldots, M\}$. These values might be augmented (i.e., imputed) in some models (e.g. `base-ror` and `base-rnr`) to a $D^*_{t,m}$ and $C^*_{t,m}$, which are explained later in this section.

Deaths by COVID-19 are statistically modelled as:

$$D^*_{t,m} \sim \text{NegativeBinomial}\left(d_{t,m}, d_{t,m} + \frac{d^2_{t,m}}{\phi}\right), \tag{1}$$

where $d_{t,m}$ represents the number of modelled cases, following:

$$d_{t,m} = \text{xfr}^*_m \sum_{\tau=0}^{t-1} c_{\tau,m} \pi_{t-\tau}, \tag{2}$$

where xfr is either the region-specific case-fatality rate cfr (in models `base-rnn` and `base-rnr`) or the infection-fatality rate $\text{ifr} \approx 0.0076$ estimated for Brazil in [40,41] (in models `base`, `base-ron`, and `base-ror`). All models have an uncertainty factor added in the same way as in the original model.

Reported infections, on the other hand, are modelled as in the original model:

$$c_{t,m} = S_{t,m} R_{t,m} \sum_{\tau=0}^{t-1} c_{\tau,m} g_{t-\tau}, \tag{3}$$

where number of reported cases ($c_{t,m}$) depends on the number of susceptible individuals S, reproduction rate R, and the generation distribution g, a fixed polynomial that is also usually referred to as "serial interval" [42].

A key component in all variations of the proposed algorithm is that we do not use $c_{t,m}$ directly, but instead, we "augment" (i.e., impute) it. Thus, rather than relying solely on the death-based data to model the pandemic – arguably the most reliable source – we also use the **reported infections** data as a way to make the model less reactive. In some model variations, we assume that the number of reported infections $C_{t,m}$ is under-reported and we model this by adding a normally distributed **overestimate** variable $\text{overestimate}_m \sim \mathcal{N}(11.5, 2.0)$ to model under-reporting explicitly, following estimates that the number of COVID-19 cases in Brazil was about 11 times higher than what was officially being reported [22].

Table 1
The models differ mainly in which objectives they optimise and how one can interpret the model. For our proposed models, we use infection data (with and without an estimate of under-reporting) to try to make the model less reactive, since data depending on deaths may only reflect the situation from a few weeks back. $d_{t,m}$ represents the number of predicted deaths and $D^*_{t,m}$ is the number of deaths used as an input to the model, in the same fashion $c_{t,m}$ and $C^*_{t,m}$ are the number of predicted cases and the number used as an input to the model.

Model name	Description	Optimisation objectives		Model interpretation		Model inputs
		$d_{t,m} \propto k_d * D^*_{t,m}$	$c_{t,m} \propto k_c * C^*_{t,m}$	$d_{t,m} \propto XFR * c_{t,m}$	$c_{t,m}$	$C^*_{t,m}, D^*_{t,m}$
`base`	Baseline model	$k_d = 1$	Not Used	$XFR = IFR$	Real Cases	As reported
`base-ron`	Includes reported cases and overestimate infections	$k_d \approx 1$[a]	$k_c \sim \mathcal{N}(11.5, 2.0)$[a,b]	$XFR = IFR$	Real Cases[c]	Augmented[d]
`base-rnn`	Includes reported cases but does not attempt to overestimate infections	$k_d \approx 1$[a]	$k_c \approx 1$[a]	$XFR = CFR$	Reported Cases	Augmented[d]
`base-ror`	Includes reported cases, model retroactive data and overestimate infections	$k_d \approx 1$[a]	$k_c \sim \mathcal{N}(11.5, 2.0)$[a,b]	$XFR = IFR$	Real Cases[c]	As reported
`base-rnr`	Includes reported cases, model retroactive data but does not attempt to overestimate infections	$k_c \approx 1$[a]	$k_d \approx 1$[a]	$XFR = CFR$	Reported Cases	As reported

[a]These values may not be exact, since the model has to take into consideration the number of both cases and deaths.
[b]$\mathcal{N}(11.5, 2.0)$ follows from estimates that the number of COVID-19 cases in Brazil was about 11 times higher than officially reported [22].
[c]These cases are interpreted as the number of real cases as long as all the assumptions of the model hold true, which most likely they do not.
[d]Cases $C_{t,m}$ and deaths $D_{t,m}$ from the last week of a data snapshot are augmented (i.e., imputed) according to how historically these values had been retroactively changed, by having $C^*_{t,m} \approx C_{t,m} k_{c,t,m}$ and $D^*_{t,m} \approx D_{t,m} k_{d,t,m}$.

Hence, in addition to Eq. (1), we also rely on Eq. (4) below to calibrate our model:

$$C^*_{t,m} \sim \text{NegativeBinomial}\left(p_{t,m}, p_{t,m} + \frac{p^2_{t,m}}{\Phi}\right), \tag{4}$$

where $C^*_{t,m}$ is the new (augmented) number of reported infections and $p_{t,m} = \frac{c_{t,m}}{\text{overestimate}_m}$ estimates the actual number of people infected at time t considering an overestimate. That is, the augmented case number is where our model performs imputation due to delayed case notification, while the overestimate parameters is where our method imputes missing data due to under-reporting.

Our models have $K = 7$ covariates. Six of them are the Google Mobility indicators (described in Section 4.1), and the remaining covariate is the percentage of the population of a region Susceptible to infection $S_{t,m}$. The reproduction rate $R_{t,m}$ is assumed to vary with the covariates:

$$R_{t,m} = R_{0,m} \exp^{-\sum_{k=1}^{K} I_{k,t,m}(\alpha_{k,m}+\alpha^*_k) - S_{t,m}(\alpha_{pop,m}+\alpha^*_{pop})}, \tag{5}$$

where the percentage of the population that is susceptible to the disease $S_{t,m}$ (the seventh covariate) was modelled with a similar impact measure α_{pop} as the mobility data. On both the `base` and our proposed models we use an extra-region impact measure α_k as well as a per-region impact measure $\alpha_{k,m}$. To consistently simulate the baseline algorithm, our simulations with `base` model did not include $S_{t,m}$ as a covariate, and we also used the same way of weighting these covariates, with both a state-wide α_k and a per-region $\alpha_{k,m}$.

Some of our models (`base-ror` and `base-rnr`) also try to take into account the delay between PCR test collection and test result notification. At any given week, we expect the number of people getting infected to be higher than reported. When we look back at the tally of infections for the same week a few weeks later, we will notice that infections can generally be between 2.5x to 7.5x higher than their initial reported values, and from 60% up to more than 85% of cases are reported with 5 days of delay. See Figures S1 and S2 for a comparison of the delay in the number of reported cases and reported deaths, respectively. Deaths follow a similar but less volatile pattern, rarely passing values 2x higher than their initial reports and possibly having 50% of the data reported after this period. While this confirms that deaths are the most reliable source of information, this also shows an issue in using such "hot" data for modelling: it can often be incomplete and lead to a false decrease in the number of cases and deaths.

In the models mentioned above, we compensate for this delay by augmenting the number of cases and deaths of the past week by a percentage, $\Delta\text{retroactive}_{c,t,m}$ and $\Delta\text{retroactive}_{d,t,m}$, respectively. These values vary per region and were calculated every week based on historical data up until that point. A sample of their distributions can be seen in Figures S3 and S4. This gives us the augmented values:

$$C^*_{t,m} = C_{t,m}(1 + \Delta\text{retroactive}_{c,t,m}) \tag{6}$$

and

$$D^*_{t,m} = D_{t,m}(1 + \Delta\text{retroactive}_{d,t,m}) \tag{7}$$

Another minor modification was that the original model assumed the onset-to-death distribution to follow the distribution $\text{Gamma}(17.8, 0.45)$. We kept modelling this as a gamma function but we estimated the average and deviation from the data at each snapshot. For reference, this value was close to $\text{Gamma}(20.67, 0.76)$ on the latest simulated weeks and the distribution of onset-to-death in Santa Catarina can be seen in Figure S7.

3.2. Test workflow and validation

Our main goal is to assess which combination of equations would most accurately forecast the curve of deaths by COVID-19 for the seven demographic macro-regions within Santa Catarina. We ran the models as close as possible to what happened in real life. For every week, we only used the data available at that point in time, akin to the last-block evaluation methods normally used for time series forecasting [34]. Forecasts of the regions were then aggregated to compose the overall prediction for the entire state.

The models produce an average prediction which we compare to the "ground-truth" number of deaths using Root Mean Squared Error (RMSE) and the Mean Average Error (MAE) metrics. To validate our results, we selected the data snapshot from 07/03/2021 to represent the "ground-truth" – 1 month after the last run simulation – to account for the notification delays discussed in the previous section. We also calculated RMSE and MAE of the upper and lower confidence intervals and took their average. The resulting metrics, RMSE_conf and MAE_conf, provide a measure of how distant the borders of the confidence interval are from the truth values.

Another important aspect of our testing procedure is how we set up the priors for the weekly simulations (a summary can be seen in Algorithm 1). At any given week, except the first one, posteriors inferred from the previous week were used as starting points for the current models. This practice of updating the priors with previous estimations is known as Sequential Bayesian Updating, or simply Bayesian

Algorithm 1 Test Workflow

```
procedure TEST
    last−date ← NULL
    for current−date in dates do
        if last−date is NULL then
            run−and−save−model(current−date, initial−hyperparameters)
        else
            run−and−save−model(current−date, hyperparameters, load−last−model(last−date))
            last−date ← current−date
        end if
    end for
end procedure
```

Updating [43], and has been used for similar purposes in related literature [44], as well as in other academic fields [45,46].

This strategy allowed the inference optimisation to converge much faster in our experiments, so we were able to run fewer iterations of the algorithm than if we had to build the model from the ground up every week. The number of iterations (see Table S1) was chosen after a preliminary test phase where we analysed the trade-off between reliability and execution time. While this sequential nature of the experiments means that we only report a single run of the model for each week, we still believe the sample size produced is more than enough to allow us to identify which models are better.

4. Data and results

4.1. Data

We were granted access to the state government big data platform Plataforma BoaVista [47], from where we obtained anonymised data on every confirmed case of COVID-19 in SC along with the date of onset of first symptoms, date of PCR test collection, date of death and municipality of residence. We collected data every week, starting from 31/05/2020 – when daily snapshots of data became available in the official database system of the state – until 31/01/2021. We also downloaded mobility data from Google Mobility community reports [48]. This data describes how people's mobility has changed during the pandemic and were available per day and in six categories: Grocery & pharmacy, Parks, Transit stations, Retail & Recreation, Residential, and Workplaces. In practice, when simulating the weekly runs, Google Mobility data from the past week onward was unavailable; therefore, we assumed that these covariates would remain constant from one week before the snapshot date.

The state of Santa Catarina has over 7 million inhabitants organised in 297 municipalities organised in 6 distinct geographical macro-regions distributed across 95 square kilometres of land area. Estimated population in each of the seven regions of Santa Catarina were obtained from the Brazilian Institute of Geography and Statistics, IBGE [49]. The state government imposed suspensions of many economic activities after the first deaths were confirmed in SC in March 2020 but ended up relaxing social distancing measures, eventually leading to a decree in June 2020 after which municipal governments would be responsible for most decisions regarding NPI measures. By 24 March 2021, when we completed this study, over 764,000 cases and over 9800 deaths by COVID-19 had been confirmed, hospitals were fully occupied, and local news reported that at least 397 people were on the waiting list for ICU beds [50].

4.2. Results

Results of our simulations of the `base` model and the four variations of our proposed model, (`base-ron`, `base-rnn`, `base-ror`, `base-rnr`) is presented here. All of our models add the **newly reported infections** to the equations but they differ in how the number of infections is augmented (i.e., imputed) and whether an estimated percentage of cases and deaths are added retroactively in the data before running the model to account for **delay in test reporting**.

Table 2
This table shows the test error values (RMSE and $avg(\pm std)$ MAE) for the predicted value and their confidence interval counterparts in a 7 and 30-day forecasting window.

Model	RMSE7		RMSE30	
	pred	conf	pred	conf
base	2.55	2.46	5.79	3.57
base-ron	2.29	2.55	2.92	3.06
base-rnn	2.48	2.55	3.14	3.10
base-ror	**2.28**	**2.20**	3.05	**2.82**
base-rnr	2.31	2.43	**3.02**	3.01
Model	MAE7		MAE30	
	pred	conf	pred	conf
base	2.19 (±1.05)	2.09 (±1.05)	4.77 (±2.71)	2.96 (±1.60)
base-ron	**1.96** (±0.95)	2.21 (±1.01)	**2.40** (±1.32)	2.51 (±1.38)
base-rnn	2.13 (±0.98)	2.18 (±1.06)	2.56 (±1.45)	2.54 (±1.41)
base-ror	**1.97** (±0.92)	**1.86** (±0.96)	2.52 (±1.36)	**2.27** (±1.32)
base-rnr	**1.97** (±0.96)	2.07 (±1.01)	2.48 (±1.37)	2.45 (±1.38)

On average, all models had a similar prediction accuracy on the first 7–10 days of forecast (P-value > 0.05, One-way ANOVA) but our proposed models outperformed the `base` model in the medium term (P-value $< 10^{-14}$, One-way ANOVA). This is illustrated in Fig. 2, where we show the average residual errors for all weekly forecasts aggregated to the entire state of SC. Notice how the margin of prediction errors made by `base` model grew wider over time while our models maintained a more stable error throughout the forecasting period. Table 2 also provide a numerical comparison of these errors for a window of 7 and 30 days, respectively. In terms of both RMSE and MAE for a 30-day forecasting window, all of our models were considered to be significantly different from the `base` model (P-value $< 10^{-5}$, independent T-test), while for a 7-day window all models except `base-rnn` were individually considered significantly different (P-value < 0.01, independent T-test). On Table S2, one could also inspect the predictive performance of the models for each of the seven individual geographic regions that compose the state of Santa Catarina.

The gap between the baseline model and the proposed method over time is most noticeable at particular points in time, as indicated by RMSE plots for 7-day and 30-day forecasting windows of the models in Fig. 3. The `base` model showed the largest short-term error in the middle of August 2020 and at the end of November 2020. On the 30-day window, the baseline algorithm is clearly making the worst predictions, particularly during August 2020 and the beginning of 2021 (Fig. 3(b)). The dates where we observe higher errors on `base` models correspond to predictions made on dates during or immediately after the peaks in the daily number of deaths, as highlighted in Fig. 4. Diagnostic graphs produced by the model for these dates confirm that `base` was unable to reflect major changes in the trend of death data. The model predicted that the number of infections was growing even though data regarding new reported infections already displayed a downward trend (Figures S5a–S5c). One could contrast the diagnostic

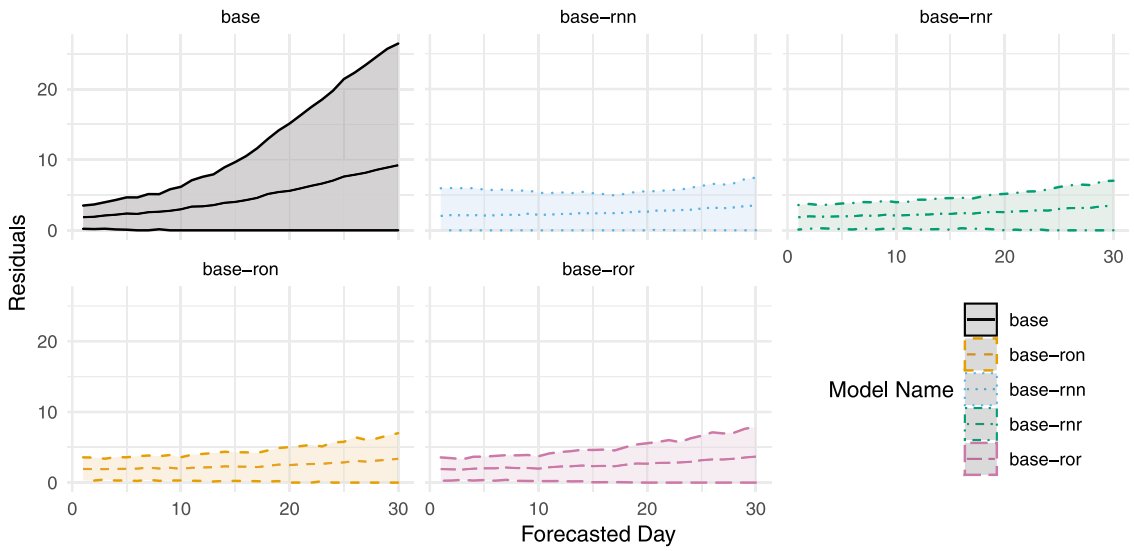


Fig. 2. Average absolute prediction error of baseline and proposed models, aggregated to the entire state of SC over all (35) weeks encompassed in the study. The lines and areas show the average and standard deviation of absolute prediction error produced by the forecast of the models for a time window of 30 days.

plots above to the ones obtained by `base-ron` model for the same dates in Figures S6a–S6c, where this misdirection in predicting death trend was not present in our proposed models.

Another way to visualise these results is by comparing the graph of cumulative deaths with predictions made by `base` and one of the best performing models `base-ron` (Fig. 5). Scenarios 01 and 03 show models' 95% confidence interval around the average prediction indicated as Scenario 02. It is clear that our model has improved the predictions, providing a narrower confidence interval which was closer to the real value.

5. Discussion

Our methods outperform the baseline in nearly all runs, with the best algorithms being the ones with the "overestimate" variable: `base-ron` and `base-ror`. These methods yielded the most stable predictions over both forecasting periods examined. Interestingly, the posterior distribution of the "overestimate" parameter in these models resulted in values much smaller compared to the priors we set ($\text{overestimate}_m \sim \mathcal{N}(11.5, 2.0)$) – see Figure S8. For example, the mean value of overestimate_m for the macro-region "Foz do Rio Itajai" was close to 7.5, and lower than 2.5 for the "Grande Oeste" macro-region. The fact that the MCMC inference algorithm automatically converged to smaller values consistently across macro-regions suggests that, although present and significant, the sub-notification in the state of Santa Catarina was not as high as we had assumed. On the other hand, `base-rnn` and `base-rnr` exhibited larger errors in predictions for certain periods in time, for example the middle of July 2020, later August 2020 or at the end of December 2020 (Fig. 3(a)).

One interesting finding of is that augmenting the number of infections by a percentage of estimated retroactive data did not seem to contribute much to the predictive accuracy, as this feature was present both in one of the best-performing models (`base-ror`) and in one of the less predictive ones, `base-rnr`. We refer the reader to Section 3 and in Table 1 for other assumptions embedded in each model variation.

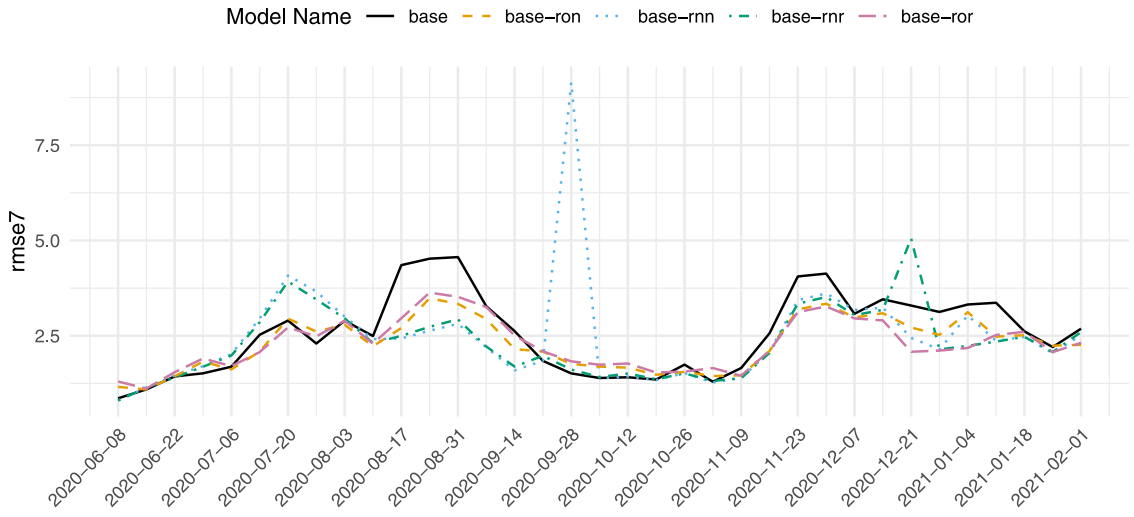
There is evidence in the literature that COVID-19 models are inefficient for long-range forecasting [51]. However, these results show that one can use "hot" data (i.e., the most recent data that is constantly being updated and might still not be complete) to update a model every week and achieve higher accuracy. Although we concede that such "hot" data come with some issues, such as unreliability and delays in updating, when forecasting epidemic outbreaks, one has to address real-time uncertainties and changes as closely as possible. A model which does not take into account the fast pace production of data is bound to underperform in a real-time setting.

6. Conclusions

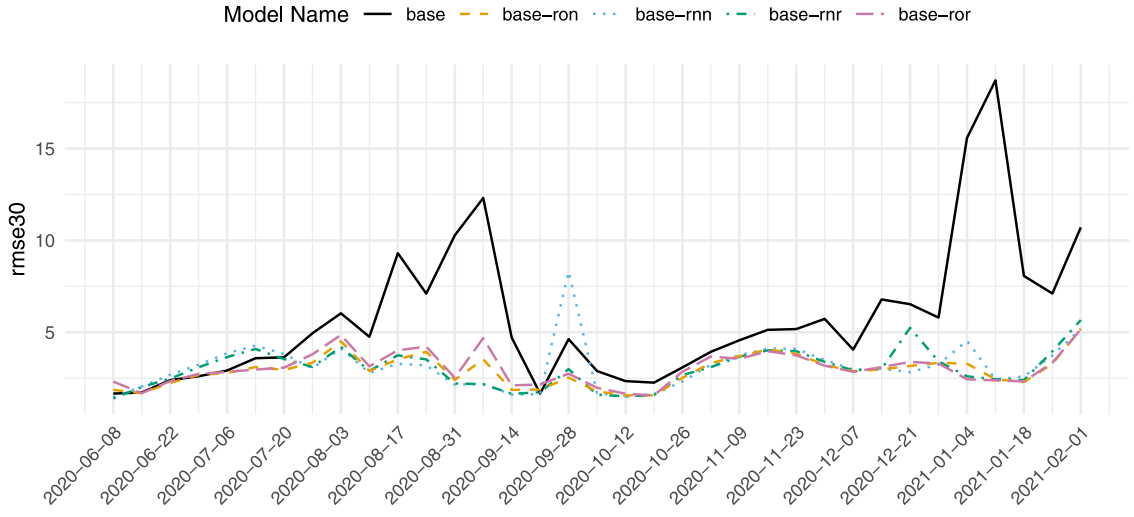
In this paper, we propose Bayesian inference models to overcome some limitations of the original algorithm in Flaxman et al. [20], mostly by introducing original equations that allow the model to access data regarding daily reported infections and letting to account for under-reporting in a more explicit way. We show that these changes increase the predictive accuracy of forecasting, not only in the near future (a week) but even in the medium term (thirty days). We have also tested some variations in the algorithm to account for the delay within test collection and notification but these did not prove as useful in predicting the death curve in the state of Santa Catarina.

These alternative models, however, are not without their failings, of course. While predictions have improved and some assumptions of the model could be confirmed by inspection of the data – for example, the onset-to-death indeed seems to follow a gamma distribution with parameters very close to what the original model assumed – there are just too many assumptions in the original model that have not been thoroughly validated [52] (for example, the value of initial reproduction number, the R_0 parameter). Another issue is that, from the point of view of the optimisation algorithm, estimated R_t values and estimated number of people infected daily are interchangeable. That is, the algorithm could reach two opposing configurations that are equally valid and optimal: one where the reproduction rate is low, but there is a large pool of infected people in the population, and another separate solution in which the number of infected people is small but R_t is larger.

Also, even though adding infection data into the model has given it more adaptability, the same data could make the model more fragile in the future. If the dynamics of infections change (e.g. because of new, more transmissible strains of the virus), the model might be biased to readjust its fitting of the historical data to compensate. In theory, this could be counteracted by more reliable epidemiological data from other sources (i.e., tracking the prevalence of new virus strains, information about age, or information on entry and exit into ICU) or by including even more granular mobility data, at the expense of the citizens' privacy, all of which are generally more expensive or infeasible to obtain. One could also consider adding assumptions to the model, such as the cultural orientation of the population [53] or the peculiarities of the test strategy in place in the modelled region [54].



(a) Forecast errors over a time window of 7 days after the day of prediction.



(b) Error on the forecasts over a time window of 30 days after the day of prediction.

Fig. 3. Comparison of Root Mean Square Error (RMSE) of predictions made by the models at each weekly snapshot.

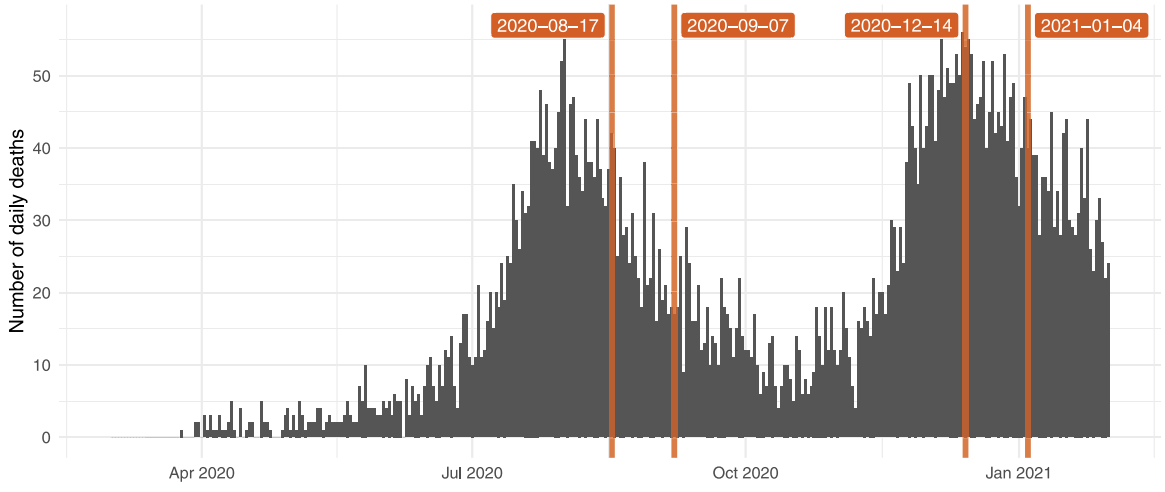


Fig. 4. The curve of daily deaths by COVID-19 in the state of Santa Catarina. Highlighted are the dates in which prediction made by `base` model were worse.

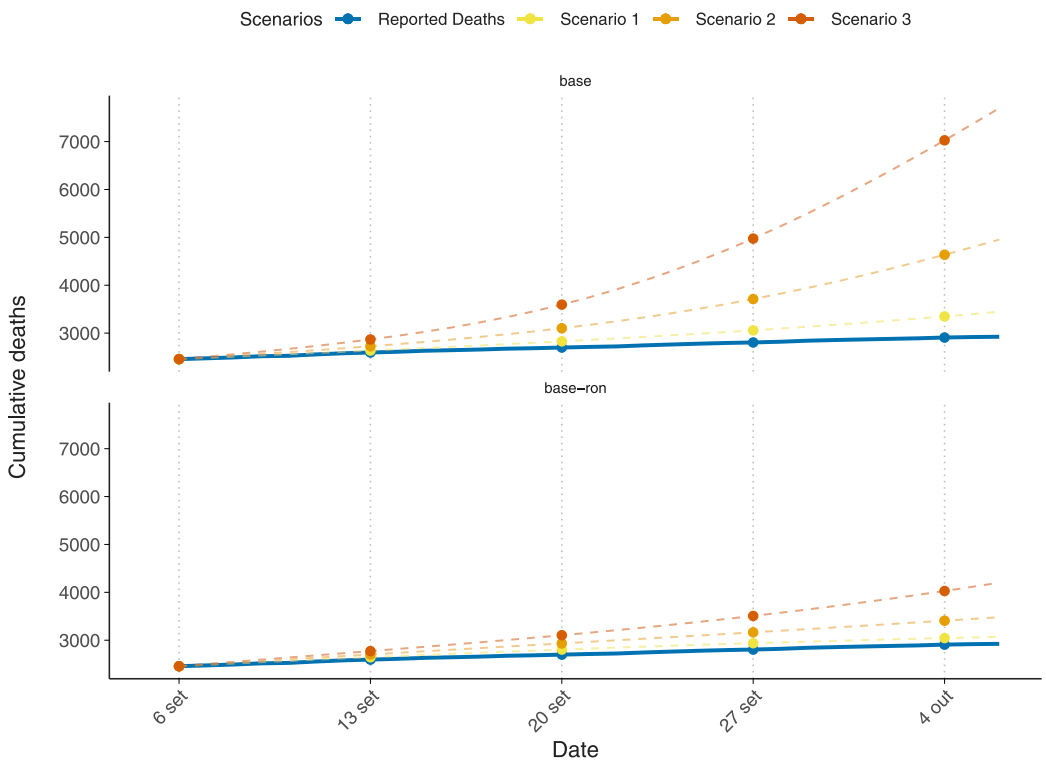


Fig. 5. The forecast scenarios produced by `base` vs `base-ron` in 07/09 and the four following weeks. Scenarios 01 and 03 are the models' 95% confidence interval around the average prediction indicated as Scenario 02. Our model clearly improved the predictions, providing a narrower confidence interval much closer to the real value.

New or data of higher resolution can alleviate some challenges in epidemic modelling but, importantly, new observations can help us revise assumptions of existing models in search of a more accurate description of real-world cases [51,55]. Our proposed model is one step in that direction of scientific inquiry. We show that a model can become more accurate by adding one more data source and a new assumption about under-reporting the tests, and therefore more useful for forecasting and decision making. As more immediate plans for future works, we also want to extend our models to account for the geographical spread of disease using concepts from network analysis in a principled way [56]. We plan to adapt algorithms our group has previously developed to analyse other types of networks in tasks involving regression, clustering and temporal data [57–59]. We also intend to review the other assumptions built into the original model and continue to investigate how much the changes we have introduced are sustained in the face of new epidemic waves, new variants of the virus and new political measures that affect the dynamics of contagion.

CRediT authorship contribution statement

Pedro Henrique da Costa Avelar: Conceived the experiments, Programmed and conducted the experiments, Analysed the results, Reviewed the manuscript. **Natalia del Coco:** Conducted some experiments, Discussed the results, Reviewed the manuscript. **Luis C. Lamb:** Discussed the results, Reviewed the manuscript. **Sophia Tsoka:** Discussed the results, Reviewed the manuscript. **Jonathan Cardoso-Silva:** Conceived the experiments, Analysed the results, Reviewed the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research presented in this paper was conceived in the context of the Intersectorial Data Intelligence Center for COVID-19 (NI-IDC), a group of volunteers that ran from March until June 2020 from academia, private companies and public local state departments dedicated to discussing data analytic strategies with potential to inform decision-makers in the southern Brazilian state of Santa Catarina (SC) [60,61]. Special thanks to Dr. Ana Luiza Curi Hallal for the valuable exchanges and discussions about epidemic modelling in the early days of this group. We thank CIASC (SC) for providing us the credentials to download data from Platform BoaVista. Thanks, Bang Wong, for providing a colour scheme inclusive for colour-blind people [62]. Data Science Brigade acknowledges funding from ICASA (SC). P.H.C.A. and L.C.L. acknowledge that this study was financed in part by CNPq (Brazilian Research Council) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Finance Code 001. P.H.C.A. acknowledges that during his stay at KCL and A*STAR he is partly funded by King's College London and by the A*STAR Research Attachment Programme (ARAP).

Code availability

The source code to replicate this study and to generates the figures in this paper are available at: https://github.com/jonjoncardoso/paper-covid19-modelling-2022-healthcare-analytics. The source code of the models, written in R and Stan, can be found on the following Github repository: https://github.com/jonjoncardoso/modelo-epidemiologico-sc.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.health.2022.100115.

References

[1] Domenico Cucinotta, Maurizio Vanelli, WHO declares COVID-19 a pandemic, Acta Bio Med. Atenei Parmensis 91 (1) (2020) 157–160, http://dx.doi.org/10.23750/abm.v91i1.9397, ISSN 25316745, 03924203.

[2] Tedros Adhanom Ghebreyesus, WHO Director-General's opening remarks at the media briefing on COVID-19 - 16 March 2020, Technical report, World Health Organization (WHO), 2020, pp. 26–28, URL https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---16-march-2020 Publication Title: World Health Organization.

[3] Talha Burki, COVID-19 in latin america, Lancet Infect. Dis. (ISSN: 14733099) 20 (5) (2020) 547–548, http://dx.doi.org/10.1016/S1473-3099(20)30303-0, URL https://linkinghub.elsevier.com/retrieve/pii/S1473309920303030.

[4] L. Coudeville, G.B. Gomez, O. Jollivet, R.C. Harris, E. Thommes, S. Druelles, A. Chit, S.S. Chaves, C. Mahé, Exploring uncertainty and risk in the accelerated response to a COVID-19 vaccine: Perspective from the pharmaceutical industry, Vaccine (ISSN: 0264410X) 38 (48) (2020) 7588–7595, http://dx.doi.org/10.1016/j.vaccine.2020.10.034, URL https://linkinghub.elsevier.com/retrieve/pii/S0264410X20313281.

[5] Madjid Tavana, Kannan Govindan, Arash Khalili Nasr, Mohammad Saeed Heidary, Hassan Mina, A mathematical programming approach for equitable COVID-19 vaccine distribution in developing countries, Ann. Oper. Res. (2021) http://dx.doi.org/10.1007/s10479-021-04130-z, URL https://link.springer.com/10.1007/s10479-021-04130-z ISSN 0254-5330, 1572-9338.

[6] The Lancet, COVID-19 in Brazil: "So what?", Lancet (ISSN: 01406736) 395 (10235) (2020) 1461, http://dx.doi.org/10.1016/S0140-6736(20)31095-3, URL https://linkinghub.elsevier.com/retrieve/pii/S0140673620310953.

[7] Pedro Baqui, Ioana Bica, Valerio Marra, Ari Ercole, Mihaela van der Schaar, Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study, The Lancet Glob. Health (ISSN: 2214109X) 8 (8) (2020) e1018–e1026, http://dx.doi.org/10.1016/S2214-109X(20)30285-0, URL https://linkinghub.elsevier.com/retrieve/pii/S2214109X20302850.

[8] Leandro Pereira Garcia, Jefferson Traebert, Alexandra Crispim Boing, Grazielli Faria Zimmer Santos, Lucas Alexandre Pedebôs, Eleonora d'Orsi, Paulo Inacio Prado, Maria Amelia de Sousa Mascena Veras, Giuliano Boava, Antonio Fernando Boing, O potencial de propagação da COVID-19 e a tomada de decisão governamental: uma análise retrospectiva em Florianópolis, Brasil, Revis. Brasileira Epidemiologia 23 (2020) e200091, http://dx.doi.org/10.1590/1980-549720200091, URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-790X2020000100208&tlng=pt ISSN 1980-5497, 1415-790X.

[9] Nicholas P. Jewell, Joseph A. Lewnard, Britta L. Jewell, Predictive mathematical models of the COVID-19 pandemic: Underlying principles and value of projections, JAMA (ISSN: 0098-7484) 323 (19) (2020) 1893, http://dx.doi.org/10.1001/jama.2020.6585, URL https://jamanetwork.com/journals/jama/fullarticle/2764824.

[10] Ernesto Estrada, COVID-19 and SARS-CoV-2. Modeling the present, looking at the future, Phys. Rep. (ISSN: 03701573) 869 (2020) 1–51, http://dx.doi.org/10.1016/j.physrep.2020.07.005, URL https://linkinghub.elsevier.com/retrieve/pii/S0370157320302544.

[11] Weston C. Roda, Marie B. Varughese, Donglin Han, Michael Y. Li, Why is it difficult to accurately predict the COVID-19 epidemic? Infect. Dis. Model. (ISSN: 24680427) 5 (2020) 271–281, http://dx.doi.org/10.1016/j.idm.2020.03.001, URL https://linkinghub.elsevier.com/retrieve/pii/S2468042720300075.

[12] Andrea L. Bertozzi, Elisa Franco, George Mohler, Martin B. Short, Daniel Sledge, The challenges of modeling and forecasting the spread of COVID-19, Proc. Natl. Acad. Sci. 117 (29) (2020) 16732–16738, http://dx.doi.org/10.1073/pnas.2006520117, URL https://pnas.org/doi/full/10.1073/pnas.2006520117 ISSN 0027-8424, 1091-6490.

[13] Steven Sanche, Yen Ting Lin, Chonggang Xu, Ethan Romero-Severson, Nick Hengartner, Ruian Ke, High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2, Emerg. Infect. Diseases 26 (7) (2020) 1470–1477, http://dx.doi.org/10.3201/eid2607.200282, URL http://wwwnc.cdc.gov/eid/article/26/7/20-0282_article.htm ISSN 1080-6040, 1080-6059.

[14] Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, Nicholas Davies, Amy Gimma, Kevin van Zandvoort, Hamish Gibbs, Joel Hellewell, Christopher I Jarvis, Sam Clifford, Billy J Quilty, Nikos I Bosse, Sam Abbott, Petra Klepac, Stefan Flasche, Early dynamics of transmission and control of COVID-19: a mathematical modelling study, Lancet Infect. Dis. (ISSN: 14733099) 20 (5) (2020) 553–558, http://dx.doi.org/10.1016/S1473-3099(20)30144-4, URL https://linkinghub.elsevier.com/retrieve/pii/S1473309920301444.

[15] Timothy W Russell, Joseph T Wu, Sam Clifford, W John Edmunds, Adam J Kucharski, Mark Jit, Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study, Lancet Public Health (ISSN: 24682667) 6 (1) (2021) e12–e20, http://dx.doi.org/10.1016/S2468-2667(20)30263-2, URL https://linkinghub.elsevier.com/retrieve/pii/S2468266720302632.

[16] Uri Goldsztejn, David Schwartzman, Arye Nehorai, Public policy and economic dynamics of COVID-19 spread: A mathematical modeling study, in: Laurent Pujo-Menjouet (Ed.), PLOS ONE (ISSN: 1932-6203) 15 (12) (2020) e0244174, http://dx.doi.org/10.1371/journal.pone.0244174, URL https://dx.plos.org/10.1371/journal.pone.0244174.

[17] You Li, Harry Campbell, Durga Kulkarni, Alice Harpur, Madhurima Nundy, Xin Wang, Harish Nair, The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries, Lancet Infect. Dis. (ISSN: 14733099) 21 (2) (2021) 193–202, http://dx.doi.org/10.1016/S1473-3099(20)30785-4, URL https://linkinghub.elsevier.com/retrieve/pii/S1473309920307854.

[18] Jan M. Brauner, Sören Mindermann, Mrinank Sharma, David Johnston, John Salvatier, Tomáš Gaveniak, Anna B. Stephenson, Gavin Leech, George Altman, Vladimir Mikulik, Alexander John Norman, Joshua Teperowski Monrad, Tamay Besiroglu, Hong Ge, Meghan A. Hartwick, Yee Whye Teh, Leonid Chindelevitch, Yarin Gal, Jan Kulveit, Inferring the effectiveness of government interventions against COVID-19, Science 371 (6531) (2021) eabd9338, http://dx.doi.org/10.1126/science.abd9338, URL https://www.science.org/doi/10.1126/science.abd9338 ISSN 0036-8075, 1095-9203.

[19] S Flaxman, S Mishra, A Gandy, H Unwin, H Coupland, T Mellan, H Zhu, T Berah, J Eaton, P Perez Guzman, N Schmit, L Cilloni, K Ainslie, M Baguelin, I Blake, A Boonyasiri, O Boyd, L Cattarino, C Ciavarella, L Cooper, Z Cucunuba Perez, G Cuomo-Dannenburg, A Dighe, A Djaafara, I Dorigatti, S Van Elsland, R Fitzjohn, H Fu, K Gaythorpe, L Geidelberg, N Grassly, W Green, T Hallett, A Hamlet, W Hinsley, B Jeffrey, D Jorgensen, E Knock, D Laydon, G Nedjati Gilani, P Nouvellet, K Parag, I Siveroni, H Thompson, R Verity, E Volz, C Walters, H Wang, Y Wang, O Watson, P Winskill, X Xi, C Whittaker, P Walker, A Ghani, C Donnelly, S Riley, L Okell, M Vollmer, N Ferguson, S Bhatt, Report 13: estimating the Number of Infections and the Impact of Non-Pharmaceutical Interventions on COVID-19 in 11 European Countries, Technical report, Imperial College London, 2020, http://dx.doi.org/10.25561/77731, URL http://spiral.imperial.ac.uk/handle/10044/1/77731.

[20] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette, Unwin T., Thomas A. Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W. Eaton, Mélodie Monod, Imperial College COVID-19 Response Team, Pablo N. Perez-Guzman, Nora Schmit, Lucia Cilloni, Kylie E.C. Ainslie, Marc Baguelin, Adhiratha Boonyasiri, Olivia Boyd, Lorenzo Cattarino, Laura V. Cooper, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Ilaria Dorigatti, Sabine L. van Elsland, Richard G. FitzJohn, Katy A.M. Gaythorpe, Lily Geidelberg, Nicholas C. Grassly, William D. Green, Timothy Hallett, Arran Hamlet, Wes Hinsley, Ben Jeffrey, Edward Knock, Daniel J. Laydon, Gemma Nedjati-Gilani, Pierre Nouvellet, Kris V. Parag, Igor Siveroni, Hayley A. Thompson, Robert Verity, Erik Volz, Caroline E. Walters, Haowei Wang, Yuanrong Wang, Oliver J. Watson, Peter Winskill, Xiaoyue Xi, Patrick G.T. Walker, Azra C. Ghani, Christl A. Donnelly, Steven Riley, Michaela A.C. Vollmer, Neil M. Ferguson, Lucy C. Okell, Samir Bhatt, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe, Nature 584 (7820) (2020) 257–261, http://dx.doi.org/10.1038/s41586-020-2405-7, URL http://www.nature.com/articles/s41586-020-2405-7 ISSN 0028-0836, 1476-4687.

[21] Christof Kuhbandner, Stefan Homburg, Commentary: Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe, Front. Med. (ISSN: 2296-858X) 7 (2020) 580361, http://dx.doi.org/10.3389/fmed.2020.580361, URL https://www.frontiersin.org/articles/10.3389/fmed.2020.580361/full.

[22] Marcelo Freitas do Prado, Bianca Brandão de Paula Antunes, Leonardo dos Santos Lourenço Bastos, Igor Tona Peres, Amanda de Araújo Batista da Silva, Leila Figueiredo Dantas, Fernanda Araújo Baião, Paula Maçaira, Silvio Hamacher, Fernando Augusto Bozza, Analysis of COVID-19 under-reporting in Brazil, Revis. Brasileira Terapia Intensiva (ISSN: 0103-507X) 32 (2) (2020) http://dx.doi.org/10.5935/0103-507X.20200030, URL http://rbti.org.br/artigo/detalhes/0103507X-32-2-7.

[23] Katelyn M. Gostic, Lauren McGough, Edward B. Baskerville, Sam Abbott, Keya Joshi, Christine Tedijanto, Rebecca Kahn, Rene Niehus, James A. Hay, Pablo M. De Salazar, Joel Hellewell, Sophie Meakin, James D. Munday, Nikos I. Bosse, Katharine Sherrat, Robin N. Thompson, Laura F. White, Jana S. Huisman, Jérémie Scire, Sebastian Bonhoeffer, Tanja Stadler, Jacco Wallinga, Sebastian Funk, Marc Lipsitch, Sarah Cobey, Practical considerations for measuring the effective reproductive number, Rt, in: Virginia E. Pitzer (Ed.), PLoS Comput. Biol. (ISSN: 1553-7358) 16 (12) (2020) e1008409, http://dx.doi.org/10.1371/journal.pcbi.1008409, URL https://dx.plos.org/10.1371/journal.pcbi.1008409.

[24] Darlan S. Candido, Ingra M. Claro, Jaqueline G. de Jesus, William M. Souza, Filipe R.R. Moreira, Simon Dellicour, Thomas A. Mellan, Louis du Plessis, Rafael H.M. Pereira, Flavia C.S. Sales, Erika R. Manuli, Julien Thézé, Luiz Almeida, Mariane T. Menezes, Carolina M. Voloch, Marcilio J. Fumagalli, Thaís M. Coletti, Camila A.M. da Silva, Mariana S. Ramundo, Mariene R. Amorim, Henrique H. Hoeltgebaum, Swapnil Mishra, Mandev S. Gill, Luiz M. Carvalho, Lewis F. Buss, Carlos A. Prete, Jordan Ashworth, Helder I. Nakaya, Pedro S. Peixoto, Oliver J. Brady, Samuel M. Nicholls, Amilcar Tanuri, Átila D. Rossi, Carlos K.V. Braga, Alexandra L. Gerber, Ana Paula de C. Guimarães, Nelson Gaburo, Cecila Salete Alencar, Alessandro C.S. Ferreira, Cristiano X. Lima,

José Eduardo Levi, Celso Granato, Giulia M. Ferreira, Ronaldo S. Francisco, Fabiana Granja, Marcia T. Garcia, Maria Luiza Moretti, Mauricio W. Perroud, Terezinha M.P.P. Castiñeiras, Carolina S. Lazari, Sarah C. Hill, Andreza Aruska de Souza Santos, Camila L. Simeoni, Julia Forato, Andrei C. Sposito, Angelica Z. Schreiber, Magnun N.N. Santos, Camila Zolini de Sá, Renan P. Souza, Luciana C. Resende-Moreira, Mauro M. Teixeira, Josy Hubner, Patricia A.F. Leme, Rennan G. Moreira, Maurício L. Nogueira, Genomics Brazil-UK Centre for Arbovirus Discovery, Epidemiology (CADDE) Genomic Network, Neil M. Ferguson, Silvia F. Costa, José Luiz Proenca-Modena, Ana Tereza R. Vasconcelos, Samir Bhatt, Philippe Lemey, Chieh-Hsi Wu, Andrew Rambaut, Nick J. Loman, Renato S. Aguiar, Oliver G. Pybus, Ester C. Sabino, Nuno Rodrigues Faria, Evolution and epidemic spread of SARS-CoV-2 in Brazil, Science 369 (6508) (2020) 1255–1260, http://dx.doi.org/10.1126/science.abd2161, URL https://www.science.org/doi/10.1126/science.abd2161 ISSN 0036-8075, 1095-9203.

[25] Mohammad H. Nadimi-Shahraki, Hoda Zamani, Seyedali Mirjalili, Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study, Comput. Biol. Med. (ISSN: 00104825) 148 (2022) 105858, http://dx.doi.org/10.1016/j.compbiomed.2022.105858, URL https://linkinghub.elsevier.com/retrieve/pii/S0010482522006126.

[26] Toufique A. Soomro, Lihong Zheng, Ahmed J. Afifi, Ahmed Ali, Ming Yin, Junbin Gao, Artificial intelligence (AI) for medical imaging to combat coronavirus disease (COVID-19): a detailed review with direction for future research, Artif. Intell. Rev. 55 (2) (2022) 1409–1439, http://dx.doi.org/10.1007/s10462-021-09985-z, URL https://link.springer.com/10.1007/s10462-021-09985-z ISSN 0269-2821, 1573-7462.

[27] Jose M. Martin-Moreno, Antoni Alegre-Martinez, Victor Martin-Gorgojo, Jose Luis Alfonso-Sanchez, Ferran Torres, Vicente Pallares-Carratala, Predictive models for forecasting public health scenarios: Practical experiences applied during the first wave of the COVID-19 pandemic, Int. J. Environ. Res. Public Health (ISSN: 1660-4601) 19 (9) (2022) 5546, http://dx.doi.org/10.3390/ijerph19095546, URL https://www.mdpi.com/1660-4601/19/9/5546.

[28] M.S. Bartlett, Measles periodicity and community size, J. R. Statist. Soc. Ser. A (Gen.) (ISSN: 00359238) 120 (1) (1957) 48, http://dx.doi.org/10.2307/2342553, URL https://www.jstor.org/stable/2342553?origin=crossref.

[29] K.-R. Müller, A.J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, V. Vapnik, in: Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Wulfram Gerstner, Alain Germond, Martin Hasler, Jean-Daniel Nicoud (Eds.), Artificial Neural Networks — ICANN'97, Vol. 1327, in: Lecture Notes in Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg, 1997, pp. 999–1004, http://dx.doi.org/10.1007/BFb0020283, URL http://link.springer.com/10.1007/BFb0020283 ISBN 978-3-540-63631-1 978-3-540-69620-9.

[30] Debanjan Parbat, Monisha Chakraborty, A python based support vector regression model for prediction of COVID19 cases in India, Chaos Solitons Fractals (ISSN: 09600779) 138 (2020) 109942, http://dx.doi.org/10.1016/j.chaos.2020.109942, URL https://linkinghub.elsevier.com/retrieve/pii/S0960077920303416.

[31] Siti Nurhidayah Sharin, Mohamad Khairil Radzali, Muhamad Shirwan Abdullah Sani, A network analysis and support vector regression approaches for visualising and predicting the COVID-19 outbreak in Malaysia, Healthc. Anal. (ISSN: 27724425) 2 (2022) 100080, http://dx.doi.org/10.1016/j.health.2022.100080, URL https://linkinghub.elsevier.com/retrieve/pii/S2772442522000338.

[32] Iebeling Kaastra, Milton Boyd, Designing a neural network for forecasting financial and economic time series, Neurocomputing (ISSN: 09252312) 10 (3) (1996) 215–236, http://dx.doi.org/10.1016/0925-2312(95)00039-9, URL https://linkinghub.elsevier.com/retrieve/pii/0925231295000399.

[33] L.J. Cao, F.E.H. Tay, Support vector machine with adaptive parameters in financial time series forecasting, IEEE Trans. Neural Netw. (ISSN: 1045-9227) 14 (6) (2003) 1506–1518, http://dx.doi.org/10.1109/TNN.2003.820556, URL http://ieeexplore.ieee.org/document/1257413/.

[34] Christoph Bergmeir, José M. Benítez, On the use of cross-validation for time series predictor evaluation, Inform. Sci. (ISSN: 00200255) 191 (2012) 192–213, http://dx.doi.org/10.1016/j.ins.2011.12.028, URL https://linkinghub.elsevier.com/retrieve/pii/S0020025511006773.

[35] Richard McElreath, Statistical Rethinking: A Bayesian Course with Examples in R and Stan, second ed., Chapman and Hall/CRC, ISBN: 9780429029608, 2020, http://dx.doi.org/10.1201/9780429029608, URL https://www.taylorfrancis.com/books/9780429642319.

[36] Mohammad H. Nadimi-Shahraki, Saeed Mohammadi, Hoda Zamani, Mostafa Gandomi, Amir H. Gandomi, A hybrid imputation method for multi-pattern missing data: A case study on type II diabetes diagnosis, Electronics (ISSN: 2079-9292) 10 (24) (2021) 3167, http://dx.doi.org/10.3390/electronics10243167, URL https://www.mdpi.com/2079-9292/10/24/3167.

[37] Giacomo De Nicola, Marc Schneble, Göran Kauermann, Ursula Berger, Regional now- and forecasting for data reported with delay: toward surveillance of COVID-19 infections, AStA Adv. Stat. Anal. (2022) http://dx.doi.org/10.1007/s10182-021-00433-5, URL https://link.springer.com/10.1007/s10182-021-00433-5 ISSN 1863-8171, 1863-818X.

[38] Trevor Hastie, Robert Tibshirani, Generalized Additive Models, Chapman & Hall/CRC, Boca Raton, Fla, ISBN: 978-0-412-34390-2, 1999.

[39] Simon N. Wood, Generalized Additive Models: an Introduction with R, 0 ed., Chapman and Hall/CRC, ISBN: 978-0-429-09315-9, 2006, http://dx.doi.org/10.1201/9781420010404, URL https://www.taylorfrancis.com/books/9781420010404.

[40] T Mellan, H Hoeltgebaum, S Mishra, C Whittaker, R Schnekenberg, A Gandy, H Unwin, M Vollmer, H Coupland, I Hawryluk, N Rodrigues Faria, J Vesga, H Zhu, M Hutchinson, O Ratmann, M Monod, K Ainslie, M Baguelin, S Bhatia, A Boonyasiri, N Brazeau, G Charles, L Cooper, Z Cucunuba Perez, G Cuomo-Dannenburg, A Dighe, A Djaafara, J Eaton, S Van Elsland, R Fitzjohn, K Fraser, K Gaythorpe, W Green, S Hayes, N Imai, B Jeffrey, E Knock, D Laydon, J Lees, T Mangal, A Mousa, G Nedjati Gilani, P Nouvellet, D Olivera Mesa, K Parag, M Pickles, H Thompson, R Verity, C Walters, H Wang, Y Wang, O Watson, L Whittles, X Xi, L Okell, I Dorigatti, P Walker, A Ghani, S Riley, N Ferguson, C Donnelly, S Flaxman, S Bhatt, Report 21: Estimating Covid-19 Cases and Reproduction Number In Brazil, Technical report, Imperial College London, 2020, http://dx.doi.org/10.25561/78872, URL http://spiral.imperial.ac.uk/handle/10044/1/78872.

[41] Thomas A Mellan, Henrique H Hoeltgebaum, Swapnil Mishra, Charlie Whittaker, Ricardo P Schnekenberg, Axel Gandy, H Juliette T Unwin, Michaela A C Vollmer, Helen Coupland, Iwona Hawryluk, Nuno Rodrigues Faria, Juan Vesga, Harrison Zhu, Michael Hutchinson, Oliver Ratmann, Mélodie Monod, Kylie E C Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Nicholas Brazeau, Giovanni Charles, Zulma Cucunuba, Gina Cuomo-Dannenburg, Amy Dighe, Jeff Eaton, Sabine L van Elsland, Katy A M Gaythorpe, Will Green, Edward Knock, Daniel Laydon, John A Lees, Andria Mousa, Gemma Nedjati-Gilani, Pierre Nouvellet, Kris V Parag, Hayley A Thompson, Robert Verity, Caroline E Walters, Haowei Wang, Yuanrong Wang, Oliver J Watson, Lilith Whittles, Xiaoyue Xi, Ilaria Dorigatti, Patrick Walker, Azra C Ghani, Steven Riley, Neil M Ferguson, Christl A Donnelly, Seth Flaxman, Samir Bhatt, Subnational Analysis of the COVID-19 Epidemic in Brazil, Technical report,, Epidemiology, 2020, http://dx.doi.org/10.1101/2020.05.09.20096701, URL http://medrxiv.org/lookup/doi/10.1101/2020.05.09.20096701.

[42] Muluneh Alene, Leltework Yismaw, Moges Agazhe Assemie, Daniel Bekele Ketema, Wodaje Gietaneh, Tilahun Yemanu Birhan, Serial interval and incubation period of COVID-19: a systematic review and meta-analysis, BMC Infect. Dis. (ISSN: 1471-2334) 21 (1) (2021) 257, http://dx.doi.org/10.1186/s12879-021-05950-x, URL https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-021-05950-x.

[43] Jacques Balayla, Bayesian updating and sequential testing: overcoming inferential limitations of screening tests, BMC Med. Inf. Decis. Making (ISSN: 1472-6947) 22 (1) (2022) 6, http://dx.doi.org/10.1186/s12911-021-01738-w, URL https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01738-w.

[44] Maria L. Daza-Torres, Marcos A. Capistrán, Antonio Capella, J. Andrés Christen, Bayesian sequential data assimilation for COVID-19 forecasting, Epidemics (ISSN: 17554365) 39 (2022) 100564, http://dx.doi.org/10.1016/j.epidem.2022.100564, URL https://linkinghub.elsevier.com/retrieve/pii/S1755436522000196.

[45] F.H. Petzschner, S. Glasauer, Iterative Bayesian estimation as an explanation for range and regression effects: A study on human path integration, J. Neurosci. 31 (47) (2011) 17220–17229, http://dx.doi.org/10.1523/JNEUROSCI.2028-11.2011, URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2028-11.2011 ISSN 0270-6474, 1529-2401.

[46] Bin Yao, Richard T.R. Qiu, Daisy X.F. Fan, Anyu Liu, Dimitrios Buhalis, Standing out from the crowd – an exploration of signal attributes of Airbnb listings, Int. J. Contemp. Hosp. Manag. (ISSN: 0959-6119, 0959-6119) 31 (12) (2019) 4520–4542, http://dx.doi.org/10.1108/IJCHM-02-2019-0106, URL https://www.emerald.com/insight/content/doi/10.1108/IJCHM-02-2019-0106/full/html.

[47] CIASC-SC, Plataforma BoaVista, 2021, URL https://www.sc.gov.br/boavista/.

[48] Ahmet Aktay, Shailesh Bavadekar, Gwen Cossoul, John Davis, Damien Desfontaines, Alex Fabrikant, Evgeniy Gabrilovich, Krishna Gadepalli, Bryant Gipson, Miguel Guevara, Chaitanya Kamath, Mansi Kansal, Ali Lange, Chinmoy Mandayam, Andrew Oplinger, Christopher Pluntke, Thomas Roessler, Arran Schlosberg, Tomer Shekel, Swapnil Vispute, Mia Vu, Gregory Wellenius, Brian Williams, Royce J. Wilson, Google COVID-19 community mobility reports: Anonymization process description (version 1.1), 2020, URL http://arxiv.org/abs/2004.04145 arXiv:2004.04145 [cs].

[49] IBGE, Estimativas populacionais dos municípios em 2019, Technical report, Instituto Brasileiro de Geografia e Estatística, 2019, URL https://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2019/estimativa_dou_2019.pdf.

[50] Valéria Martins, SC tem 9,5 mil mortes por Covid-19 e 397 pacientes na fila por UTI (PT-BR), 2021, G1 Santa Catarina, URL https://g1.globo.com/sc/santa-catarina/noticia/2021/03/21/sc-tem-95-mil-mortes-por-covid-19-e-397-pessoas-na-fila-por-uti.ghtml. Place: Florianópolis, SC Publication Title: G1 Santa Catarina.

[51] Shiva Moein, Niloofar Nickaeen, Amir Roointan, Niloofar Borhani, Zarifeh Heidary, Shaghayegh Haghjooy Javanmard, Jafar Ghaisari, Yousof Gheisari, Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan, Sci. Rep. (ISSN: 2045-2322) 11 (1) (2021) 4725, http://dx.doi.org/10.1038/s41598-021-84055-6, URL http://www.nature.com/articles/s41598-021-84055-6.

[52] Kristian Soltesz, Fredrik Gustafsson, Toomas Timpka, Joakim Jaldén, Carl Jidling, Albin Heimerson, Thomas B. Schön, Armin Spreco, Joakim Ekberg, Örjan Dahlström, Fredrik Bagge Carlson, Anna Jöud, Bo Bernhardsson, The effect of interventions on COVID-19, Nature 588 (7839) (2020) E26–E28, http://dx.doi.org/10.1038/s41586-020-3025-y, URL http://www.nature.com/articles/s41586-020-3025-y ISSN 0028-0836, 1476-4687.

[53] Gamaliel A. Palomo-Briones, Mario Siller, Arnaud Grignard, An agent-based model of the dual causality between individual and collective behaviors in an epidemic, Comput. Biol. Med. (ISSN: 00104825) 141 (2022) 104995, http://dx.doi.org/10.1016/j.compbiomed.2021.104995, URL https://linkinghub.elsevier.com/retrieve/pii/S0010482521007897.

[54] Miguel Guzmán-Merino, Christian Durán, Maria-Cristina Marinescu, Concepción Delgado-Sanz, Diana Gomez-Barroso, Jesus Carretero, David E. Singh, Assessing population-sampling strategies for reducing the COVID-19 incidence, Comput. Biol. Med. (ISSN: 00104825) 139 (2021) 104938, http://dx.doi.org/10.1016/j.compbiomed.2021.104938, URL https://linkinghub.elsevier.com/retrieve/pii/S0010482521007320.

[55] Jordana Cepelewicz, The hard lessons of modeling the coronavirus pandemic, Quanta Mag. (2021) 1–28, URL https://www.quantamagazine.org/the-hard-lessons-of-modeling-the-coronavirus-pandemic-20210128/.

[56] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, Alessandro Vespignani, Epidemic processes in complex networks, Rev. Modern Phys. 87 (3) (2015) 925–979, http://dx.doi.org/10.1103/RevModPhys.87.925, URL https://link.aps.org/doi/10.1103/RevModPhys.87.925 ISSN 0034-6861, 1539-0756.

[57] Jonathan C. Silva, Laura Bennett, Lazaros G. Papageorgiou, Sophia Tsoka, A mathematical programming approach for sequential clustering of dynamic networks, Eur. Phys. J. B 89 (2) (2016) 39, http://dx.doi.org/10.1140/epjb/e2015-60656-5, URL http://link.springer.com/10.1140/epjb/e2015-60656-5 ISSN 1434-6028, 1434-6036.

[58] Lingjian Yang, Jonathan Silva, Lazaros Papageorgiou, Sophia Tsoka, Community structure detection for directed networks through modularity optimisation, Algorithms (ISSN: 1999-4893) 9 (4) (2016) 73, http://dx.doi.org/10.3390/a9040073, URL http://www.mdpi.com/1999-4893/9/4/73.

[59] Jonathan Cardoso-Silva, Lazaros G. Papageorgiou, Sophia Tsoka, Network-based piecewise linear regression for QSAR modelling, J. Comput. Aided Mol. Des. 33 (9) (2019) 831–844, http://dx.doi.org/10.1007/s10822-019-00228-6, URL http://link.springer.com/10.1007/s10822-019-00228-6 ISSN 0920-654X, 1573-4951.

[60] Coordenadoria de Comunicação Social do MPSC, Municípios recebem Sala de Situação Digital (SSD) para apoiar o combate à covid-19, Ministério Público de Santa Catarina, Florianópolis, SC, 2020, URL https://www.mpsc.mp.br/noticias/municipios-recebem-sala-de-situacao-digital-ssd-para-apoiar-o-combate-a-covid-19.

[61] Ângelo Medeiros, Ferramenta de apoio no combate à Covid-19 é lançada com a contribuição do Judiciário, Assessoria de Imprensa/Núcleo de ComunicaÇÃo Institucional, Tribunal de JustiÇa de Santa Catarina, Florianópolis, SC, Brazil, 2020, URL https://www.tjsc.jus.br/web/imprensa/-/ferramenta-de-apoio-no-combate-a-covid-19-e-lancada-com-a-contribuicao-do-judiciario.

[62] Bang Wong, Points of view: Color blindness, Nature Methods 8 (6) (2011) 441, http://dx.doi.org/10.1038/nmeth.1618, URL http://www.nature.com/articles/nmeth.1618 ISSN 1548-7091, 1548-7105.