New Generation Computing
https://doi.org/10.1007/s00354-021-00125-3

Estimation of COVID-19 Under-Reporting in the Brazilian States Through SARI

Balthazar Paixão[1] · Lais Baroni[1] · Marcel Pedroso[2] · Rebecca Salles[1] · Luciana Escobar[1] · Carlos de Sousa[2] · Raphael de Freitas Saldanha[2] · Jorge Soares[1] · Rafaelli Coutinho[1] · Fabio Porto[3] · Eduardo Ogasawara[1]

Received: 19 December 2020 / Accepted: 4 March 2021

Abstract

Due to its impact, COVID-19 has been stressing the academy to search for curing, mitigating, or controlling it. It is believed that under-reporting is a relevant factor in determining the actual mortality rate and, if not considered, can cause significant misinformation. Therefore, this work aims to estimate the under-reporting of cases and deaths of COVID-19 in Brazilian states using data from the InfoGripe. InfoGripe targets notifications of Severe Acute Respiratory Infection (SARI). The methodology is based on the combination of data analytics (event detection methods) and time series modeling (inertia and novelty concepts) over hospitalized SARI cases. The estimate of real cases of the disease, called novelty, is calculated by comparing the difference in SARI cases in 2020 (after COVID-19) with the total expected cases in recent years (2016–2019). The expected cases are derived from a seasonal exponential moving average. The results show that under-reporting rates vary significantly between states and that there are no general patterns for states in the same region in Brazil. The states of Minas Gerais and Mato Grosso have the highest rates of under-reporting of cases. The rate of under-reporting of deaths is high in the Rio Grande do Sul and the Minas Gerais. This work can be highlighted for the combination of data analytics and time series modeling. Our calculation of under-reporting rates based on SARI is conservative and better characterized by deaths than for cases.

Keywords COVID-19 · Under-reporting · SARI · Time series modeling · Data analytics

✉ Lais Baroni
lais.baroni@eic.cefet-rj.br

[1] Federal Center for Technological Education of Rio de Janeiro, CEFET/RJ, Rio de Janeiro, Brazil

[2] Oswaldo Cruz Foundation, Fiocruz, Rio de Janeiro, Brazil

[3] National Laboratory of Scientific Computing, LNCC, Rio de Janeiro, Brazil

Published online: 14 March 2021

Introduction

In January 2020, the new coronavirus (COVID-19) was considered a Public Health Emergency of International Importance by the World Health Organization (WHO). Later, in March, WHO characterized the disease as a pandemic. Due to its relevance, many efforts are being made to combat COVID-19, either by discovering the characteristics of the virus, methods of prevention, treatment, or directing public policy action [5].

In Brazil, interventional measures such as the creation of field hospitals, surveillance information systems, and actions to reduce the economic impact are being adopted to mitigate the effects caused by COVID-19. Among the main objectives is to slow down the spread of the virus to avoid overloading the health system. In this sense, policies to encourage prevention are adopted, such as, for example, the recommendation or imposition of physical isolation and quarantine [32].

Decision-making for the adoption of public policies in this pandemic scenario is a challenging task. Part of the difficulty comes from the lack of specific information about essential characteristics such as the total number of people infected. There is a lack of availability of tests to confirm the infection by SARS-CoV-2, which ends up being performed only in more severe cases of the disease, with exceptions. Such a scenario makes the capacity of the health system to monitor the evolution of the number of cases uncertain. The discrepancy between the actual amount of infected and diagnosed individuals constitutes under-reporting [21].

It is estimated that under-reporting is a relevant factor in determining the actual mortality rate and, if not considered, can cause significant misinformation [20]. Therefore, this work aims to estimate the under-reporting of cases and deaths of COVID-19 in Brazilian states. Since the possibility of testing the entire population is not viable, data from the InfoGripe is used. InfoGripe targets notifications of Severe Acute Respiratory Infection (SARI).

Our paper stands out for adopting a methodology based on the combination of data analytics (event detection methods) and time series modeling (inertia and novelty). Data analytics is applied to determine the parameters to be used for time series modeling. The estimated parameters consider time series analysis through event detection methods.

The estimate of real cases of the disease, called novelty, is calculated by comparing the difference in SARI cases in 2020 (after COVID-19) with the total expected cases in recent years (2016–2019). The expected cases are derived from a seasonal exponential moving average. The novelty is based on inertial concepts. That is, there is a strength to maintain the values of a time series in a stable state through time [12]. Inertia remains until a rupture occurs. In this case, the rupture is the influence of the COVID-19. Under-reporting, then, is given by the difference between the novelty and the number of reported cases. In the end, under-reporting (cases and deaths) is presented as a rate for each state in Brazil.

For the sake of clarity, it is important to introduce some background for time series, moving averages, and event detection used in the context of this work.

Time Series

A time series is a sequence of observations collected in time. Usually, a time series y can be considered as a stochastic process, i.e., a sequence of n random variables$<y_1, y_2, \dots, y_n>$ [11, 28]. A specific observation of a time series is represented as y_i, indexed in time by $i = 1, \dots, n$, where y_1 represents the first observation, and y_n is the most recent observation.

The ith subsequence of size p in a time series y, represented as $seq_{i,p}(y)$, is a continuous sequence of values $< y_{i-(p-1)}, y_{i-(p-2)}, \cdots, y_i >$, where $|seq_{i,p}(y)| = p$ e $p \leq i \leq |y|$. The sequence contains ith observation and its $p - 1$ predecessors.

The ith subsequence seasonally outdated for time series y, is represented as $seq^s_{i,p}(y)$, is an ordered sequence of values $< y_{i-(p-1)\cdot s}, y_{i-(p-2)\cdot s}, \cdots, y_i >$, where p corresponds to the size of the sequence ($|seq^s_{i,p}(y)| = p$, with $p \leq i \leq |y|$), and s corresponds to the seasonality ($s \ll |y|$). The sequence contains i-th observation and its $p - 1$ predecessors outdated seasonally.

Seasonal Moving Averages

The ith moving average $\bar{y}_{i,p}$ of p terms in a time series y is calculated by the average of t_k observations in the sequence $seq_{i,p}(y)$, as shown in Eq. 1. The ith exponential moving average $\hat{y}_{i,p}$ of p terms in a time series y is calculated by the weighted average of t_k observations in the sequence $seq_{i,p}(y)$ and the weights α_k. The $\hat{y}_{i,p}$ is described in Equation 2, where there is more emphasis on the most recent observations.

$$\bar{y}_{i,p} = \frac{\sum_{k=1}^{p} t_k}{p} \mid t_k \in seq_{i,p}(y),\ p \leq i \leq |y| \tag{1}$$

$$\hat{y}_{i,p} = \frac{\sum_{k=1}^{p} \alpha_k \cdot t_k}{\sum_{k=1}^{p} \alpha_k} \mid t_k \in seq_{i,p}(y), \alpha_k = \left(1 - \frac{2}{p+1}\right)^{p-k},\ p \leq i \leq |y| \tag{2}$$

The i-th seasonal moving average $\bar{y}^s_{i,p}$ and the i-th seasonal exponential moving average $\hat{y}^s_{i,p}$ of p terms in a time series y are similarly calculated replacing the continuous sequence $seq_{i,p}(y)$ with the seasonal sequence $seq^s_{i,p}(y)$ (see “Time series”), respectively, in Eqs. 1 and 2, as shown in the Eqs. 3 and 4.

$$\bar{y}^s_{i,p} = \frac{\sum_{k=1}^{p} t_k}{p} \mid t_k \in seq^s_{i,p}(y),\ p \leq i \leq |y| \tag{3}$$

$$\hat{y}^s_{i,p} = \frac{\sum_{k=1}^{p} \alpha_k \cdot t_k}{\sum_{k=1}^{p} \alpha_k} \mid t_k \in seq^s_{i,p}(y), \alpha_k = \left(1 - \frac{2}{p+1}\right)^{p-k},\ p \leq i \leq |y| \tag{4}$$

Event Detection

Event detection methods include the discovery of anomalies and change points. Anomalies are observations that stand out because they do not appear to have been generated by the same process as the other observations in the time series [19]. Change points characterize a transition between different states in a process that generates the time series data [9, 30].

There are several methods to address the detection of anomalies [6, 13] and change points [1]. Among them, some methods consider the effects of inertia on time series data. As this work is based on inertial concepts [12], two methods of this group are presented.

Anomaly by Adaptive Normalization

Adaptive normalization [23] is used to detect anomalies. This technique uses inertia to address heteroscedastic non-stationary series. Given a time series y, the outlier removal process consists of three stages: (i) inertia calculation, (ii) noise calculation, and (iii) anomaly identification. In the inertia calculation, a moving average for the series $\overline{y}_{i,p}$ with p terms is calculated, as described by Eq. 1. The higher the value of p, the greater the inertia and the lower the adaptation speed. The noise ϵ_i is calculated by the difference between y_i and $\overline{y}_{i,p}$, i.e., $\epsilon_i = y_i - \overline{y}_{i,p}$. Finally, the observations ϵ_i classified as outliers by boxplot correspond to anomalies in Eq. 5.

$$anomaly(y) = \{i\}, \forall i \mid y_i \notin [Q_1(y) - 3 \cdot IQR(y), Q_3(y) + 3 \cdot IQR(y)] \tag{5}$$

Change Points by Change Finder

Change Finder is a technique that detects change points in univariate time series data [30]. Given a time series y, the event detection process consists of two phases. In the first phase, outliers are detected. For this, a learning model ξ is adjusted to the time series y, resulting in $\hat{y}_i = \xi(y)_i$.[1] Next, a score s_i is calculated for each observation in the series related to its deviation from the learned model. This calculation produces a time series s, as presented in Eq. 6. The highest scores for s, classified according to Eq. 5, indicate anomalies.

In the second phase, change points are detected. For this, a new time series $\overline{s}_p$ is produced, composed of moving averages of s with p terms, according to Eq. 1. The detection of change points is then reduced to the outlier detection problem in $\overline{s}_p$ like the first phase.

$$s_i = \left(\hat{y}_i - y_i\right)^2, \; \hat{y}_i = \xi(y)_i \tag{6}$$

[1] in this work, linear regression was used for adjustment.

Related Work

Due to its relevance and recent outbreak, COVID-19 has been attracting much interest in the academy. Therefore, many works on COVID-19 have been published since the beginning of 2020 until today. However, there are still few studies focused on under-reporting estimates. This low number of related publications can be a consequence of the time spent on the execution, review, editing, and publication of papers in scientific journals.

Krantz et al. [17] used harmonic analysis and wavelets to model the under-reporting of COVID-19 in several countries worldwide. They developed susceptibility and infection equations with parameters varied according to the characteristics of each country to build adaptive models. The under-reporting rate was calculated by the difference between the numbers predicted by the model and reported numbers. The result provided the ratio between reported and unreported cases in the format (1 to x) in seven countries. The authors concluded that the results are not entirely accurate due to the lack of some important information that should be included in the model and was not available.

Similarly, to review the numbers of reported COVID-19 cases in several countries, Lachmann et al. [20] also estimated expected cases. For this, the author used demographic data and fixed mortality rates of the countries and the paired comparison with the reference country (South Korea). It presented and discussed estimates of the number of people infected with COVID-19, considering a set of situations that must be true to justify the model.

Ribeiro et al. [25] used regression techniques on hospitalization data in Brazil with a type of acute respiratory syndrome as the cause. They analyzed the time evolution of hospitalizations for each month in the period between 2012 and 2019. They created a mathematical function that replicates the typical behavior of cases of hospitalization for SARI. This function was compared with data from 2020 in the same months to estimate under-reporting. The results showed an under-reporting rate of 7.7:1 for Brazil.

Bastos and Cajueiro [3] modeled and predicted the initial evolution of the COVID-19 pandemic in Brazil using about a month of data provided by the Ministry of Health of Brazil. They sought to model the spread of the virus and evaluate existing countermeasures. For that purpose, they use two variations of the SIR model and we include a parameter that comprises the effects of social distancing measures. They conclude that social distancing policy can fatten the infection pattern of the COVID-19 but that it is only effective if it lasts until mid-June, according to predictions. They also point out the importance of testing the population based on the proportion of asymptomatic individuals.

Silva et al. [29] fitted curves growth models using a Bayesian approach to calculate the total number and daily new cases in the state of Goias, Brazil. Results from the analysis also investigate the possible date of the outbreak peak to the state. The study did not take into consideration possibles changes in government control measures.

Saulo et al. [4] discussed the role of uncertainty in the prediction of the number of infected individuals and deaths. They proposed an adapted susceptible-infected-recovered (SIR) model, which explicitly incorporates the under-reporting and the response of the population to public policies to cast short-term and long-term predictions. As a contribution, it seeks to comprise the role that sub-notification uncertainty plays in the model-based predictions of the COVID-19 contagion, harshly affecting the outlooks for its evolution spread in Brazil.

Our work stands out for estimating the under-reporting of COVID-19 in Brazilian states weekly. The estimate considers the weighted historical record (in which most recent years have more weight than less recent ones) to predict expected SARI cases in 2020. It enriches the analysis allowing an estimate closer to reality. This work can also be highlighted for focusing on time series and using event detection tools in the study. Furthermore, except for the article by Ribeiro et al. [25], as far as we know, the data used in this work to obtain under-reporting rates were not used in any other work with the same or similar purpose.

Methods

In seasonal phenomena, time series are generated by superimposing a seasonal process and random noises. Based on this premise, Eq. 7 models the seasonal component of the time series, where y_i is an observation, $\hat{y}^s_{i-s,p}$ is the seasonal exponential moving average (SEMA) in the previous season, and ϵ_i is the random noise. The obtained seasonal component brings up the inertia concept in time series. It enables the analysis of the intrinsic random noise of the observed phenomenon. At the same time, the influences that determine the behavior of the series are not changed [12].

$$y_i - \hat{y}^s_{i-s,p} - \epsilon_i = 0 \tag{7}$$

In the case of rupture (i.e., a "break" in inertial behavior), we adopt the concept of novelty η. The novelty is the influence introduced in each interval resulting from a rupture in a time series. Once the novelty begins, the modeled SEMA from past data is no longer the only representative process of the new behavior of the time series. In this context, Eq. 7 is expanded to Eq. 8, that expresses novelty η_i and error $\hat{\epsilon}_i$. We have that $\hat{\epsilon}_i$ is approximated by the average error $\overline{\epsilon}$ observed in the pre-novelty period, *i.e.*, $\hat{\epsilon}_i$ is expected to be inside the interval confidence for $\overline{\epsilon}$ ($[\overline{\epsilon}_{min} - \overline{\epsilon}_{max}]$).

$$y_i - \hat{y}^s_{i-s,p} - \eta_i - \hat{\epsilon}_i = 0,\ \hat{\epsilon}_i \approx \overline{\epsilon},\ \hat{\epsilon}_i \in [\overline{\epsilon}_{min} - \overline{\epsilon}_{max}] \tag{8}$$

Until the seasonal component $\hat{y}^s_{i-s,p}$ incorporates the novelty η_i, η_i defines a new phenomenon in the time series. Regarding SARI, we assume that η_i is directly associated with COVID-19, i.e., the new known phenomenon.

From this concept, we first compute the inertial behavior of the time series to estimate under-reporting. Let t be the period in which the rupture y_t occurs. In novelty period (*i.e.*, $t \leq i \leq |y|$), η_i is the subtraction of the observations of the time series y_i by the values of SEMA from the previous period $\hat{y}^s_{i-s,p}$ and the error $\hat{\epsilon}_i$

(approximated by $\overline{\epsilon}$). Equation 8 shows the calculation of the time series with η_i for each i in the novelty period. The novelty η_i estimates the brute number of observations that exceed the expected according to the inertial behavior of the time series and its fundamental error.

To estimate the brute number of under-reported time series, we use the number of observations classified as SARS-CoV-2 (Severe Acute Respiratory Infection Coronavirus 2) in the novelty period. Equation 9 presents the calculation of the time series with absolute numbers of under-reported observations, where cov_i are observations classified as SARS-CoV-2.

$$sub_i = \eta_i - cov_i,\ t \le i \le |y| \tag{9}$$

As we assume that the modeled novelty in time series η_i represents COVID-19 cases, the time series sub_i defines the number of under-reported observations per week. Then, the estimates sub_i are added together to form the accumulated number of under-reported observations in the period, represented as cur_i in Equation 10.

$$cur_i = \sum_{i=t}^{|y|} sub_i,\ t \le i \le |y| \tag{10}$$

The under-reporting rate is estimated by dividing the accumulated number of under-reported time series cur_i by the accumulated number of total time series cov_i for the period. Equation 11 describes the under-reporting rate, denoted as tx_i, where $tx_{|y|}$ is the final rate. In this work, this calculation provides the estimated under-reporting rates for cases and deaths of COVID-19 for each Brazilian state individually. Thus, these rates allow for a comparable interpretation between the states.

$$tx_i = \frac{cur_i}{cov_i} \tag{11}$$

Experimental Setup

This section discusses the experimental setup of the scenario in which the methodology was applied. The next section presents the process of data acquisition and preparation, whereas the following section describes the methods and parameters applied in the analysis. The next section presents the implementation details.

Data Acquisition and Preparation

InfoGripe is the principal data source used for the analysis and development of the work.[2] It is an initiative of the Oswaldo Cruz Foundation (Fiocruz) with the Getulio Vargas Foundation (FGV) and the Brazilian Health Surveillance System of the Ministry of Health. It records weekly SARI reported cases since January 2009. The

[2] Data collected on July 27th, 2020.

Table 1 Attributes of processed datasets *DT_SARI_c* and *DT_SARI_d*

Attribute	Description
Year	The epidemiological year of first symptoms
Week	The epidemiological week of first symptoms
State	The state name
Total	The total number of recorded cases (*DT_SARI_c*) / deaths (*DT_SARI_d*)
SARS-CoV-2	The total number of cases with positive results for COVID-19 (*DT_SARI_c*) / deaths by COVID-19 (*DT_SARI_d*)

data comes from the Influenza Epidemiological Surveillance Information System (SIVEP-Gripe). It presents the cases following the criteria: (fever) AND (cough OR sore throat) AND (dyspnoea OR oxygen saturation < 95% OR respiratory difficulty) AND (hospitalization OR death), symptoms equivalent to SARI international records [16]. For the sake of simplicity, we are calling the dataset *DT_SARI*.

To keep only the relevant data, we apply the following filter: $type$ = "State" $\wedge$ $gender$ = "Total" $\wedge$ $scale$ = "Cases". The resulting dataset shows the number of cases or deaths per epidemiological week of a given year for each state. Besides, it specifies the number of observations that correspond to Influenza A, Influenza B, SARS-CoV-2, Respiratory Syncytial Virus (RSV), Parainfluenza 1, Parainfluenza 2, Parainfluenza 3, and Adenovirus.

It is then performed the differentiation of the case observations that evolved to death. For this, we apply a second filter that resulted in two datasets, one with cases (*DT_SARI_c*) and another with deaths (*DT_SARI_d*). Finally, five attributes of interest are selected: Year, Week, State, Total, and SARS-CoV-2. Table 1 describes these attributes.

In addition to these data, we use the number of confirmed cases (*DT_MH_c*) and confirmed deaths (*DT_MH_d*) from COVID-19 by state, provided by the Ministry of Health.[3] These numbers are updated daily on the COVID-19 Portal, the official communication channel on the epidemiological situation of COVID-19 in Brazil [14]. The values are used for purposes of comparison with the results obtained in this work.

Method and Parameter Selection

The method and parameter selection are a determining factor for the quality of the results obtained in the research. This section aims at justifying the applied methodology, which includes the choice of the used dataset, and the methods and parameters adopted in the data analysis.

Datasets. The most severe cases of COVID-19 manifest respiratory symptoms, such as difficulty in breathing or shortness of breath, and chest pain or pressure [27]. These symptoms are also present in Acute Respiratory Infection (ARI). Fever is

[3] Data collected on July 09th, 2020.

Table 2 Change point (CP) dates that occurred in 2020

State	Cases	Deaths	State	Cases	Deaths
Acre	Mar. 28	Feb. 08	Paraíba	Mar. 14	Mar. 14
Alagoas	Mar. 14	Mar. 21	Pernambuco	Mar. 07	Mar. 07
Amazonas	Mar. 14	Mar. 14	Piauí	Feb. 29	Mar. 07
Amapá	Mar. 14	Mar. 07	Paraná	–	Mar. 14
Bahia	Mar. 07	Mar. 07	Rio de Janeiro	Mar. 14	Mar. 07
Ceará	Mar. 07	Mar. 07	Rio G. do Norte	Mar. 21	Mar. 14
Dirito Federal	Mar. 07	Mar. 07	Rondônia	Mar. 28	Mar. 28
Espírito Santo	Mar. 14	Mar. 14	Roraima	Mar. 14	Mar. 14
Goiás	Mar. 14	Mar. 14	Rio G. do Sul	Mar. 21	Mar. 21
Maranhão	Feb. 01	Feb. 08	Santa Catarina	Mar. 28	Mar. 14
Minas Gerais	Mar. 14	Mar. 14	Sergipe	Mar. 14	Mar. 07
Mato G. do Sul	Mar. 14	Mar. 14	São Paulo	Mar. 07	Mar. 07
Mato Grosso	Mar. 07	Mar. 14	Tocantins	Mar. 07	Feb. 08
Pará	Mar. 14	Feb. 29			

another common symptom, even in mild cases of the disease. It is the reason for choosing SARI data (*DT_SARI*) instead of ARI data (*DT_ARI*). *DT_SARI* is a subset of *DT_ARI*. They differ only in the manifestation of fever. Therefore, we consider that the probable cases of COVID-19 with severe symptoms also present fever, making *DT_SARI* the most suitable dataset to estimate the under-reporting of the disease [18, 26].

SEMA for Inertial Model. It is necessary to identify the SARI observations that correspond to the COVID-19 to compute the under-reporting of COVID-19 in Brazil. For this, data from years predating COVID-19 should be observed to model the expected inertial behavior if there was no pandemic. Thus, it is possible to estimate the COVID-19 case number as the value exceeding the expected for the same period in the year.

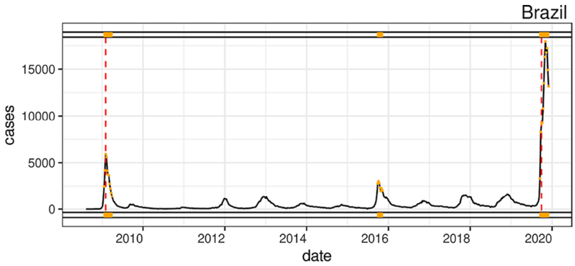
SEMA provides an appropriate method to create the inertial function since it is a trend indicator that assigns more weight to the most recent data considering a seasonal pattern. It is efficient to estimate an inertial behavior of a time series if the series has not undergone any significant behavior change in the period.

First, we define the time series for which SEMA is calculated. For this, three parameters are required: p, i, and s (see "Introduction" section). The i represents the time index of the reference time series, p is the number of predecessors, and s is the seasonality to be considered. Note that p and s are defined based on the locality of i.

The s is chosen based on the seasonal variation of respiratory viral diseases. The annual epidemics of the common cold and the flu affect the human population of temperate regions in the winter season [7, 10, 22, 31]. Therefore, s is defined as 52, since 52 corresponds to the number of weeks in the year. In this way, we guarantee the analysis of comparable observation sequences in the SARI series.

The parameters p and i are based on the response of the event detection algorithms in each state. The event detection (targeting both change points and

Fig. 1 Anomalies (yellow) and change points (red) detected in SARI cases of Brazil



anomalies) in the series *DT_SARI_c* and *DT_SARI_d* evidence consistently, in several states, behavior change in two periods: (i) between the end of 2015 and the beginning of 2016, and (ii) between February and March 2020. Table 2 shows the dates of events detected in 2020 for each state.

The events detected in 2020 are a consequence of COVID-19 in Brazil. These events coincide with the first record of the disease in the country, considering the time for the disease spread and the manifestation of symptoms [2, 15]. The events appear for most of the states from March 07 and March 14. They correspond, respectively, to the 11th and 12th epidemiological week, two or three weeks after the first confirmed case of COVID-19 in Brazil.

It is possible to identify the beginning period (t) of the novelty for a determined state.[4] The online method consists of seeking a change point in 2020, running it weekly since the first week of 2020 until it detects a change point in the year. When the change point is detected, the method stops and considers that week as the beginning of the period. So, for each state, the parameter i admits values after t and extended until the last week of data ($|y|$), which corresponds the week 26 of 2020 (i.e., June 27, 2020).

Figure 1 shows the events detected in the SARI cases curve in Brazil. In addition to 2009 (H1N1) and 2020 (COVID-19), events are observed in the 2016 period. Events presented in Fig. 1 correspond to abnormal behavior. They can affect the previous inertial behavior of the series. For this reason, the value attributed to p is 4, meaning that the previous 4 years (2016–2019) are considered.

The model errors (random noise) for this period for both the cases and deaths in each state are, respectively, described in Tables 3 and 4. Since ϵ_i follows a non-normal distribution, the interval confidence for $\overline{\epsilon}$ is computed by bootstrap with 1000 repetitions. These values are important to determine the novelty calculation, reducing the chance of an increase generated by a random event.

Implementation

The adopted methodology was implemented in R [24]. The code description and Jupyter notebook also developed in R complements this work.[5] In it, it is possible

[4] According to the corresponding epidemiological week identified by change points. They are presented in Table 2. For the state of Paraná, the date detected for deaths was used instead.

[5] Available at https://eic.cefet-rj.br/~dal/covid-19-under-report/.

Table 3 Errors of the models (cases)

State	$\overline{\epsilon}$	$[\overline{\epsilon}_{min}, \overline{\epsilon}_{max}]$
Acre	1.727	[1.177, 2.305]
Paraíba	2.198	[1.740, 2.855]
Alagoas	1.482	[0.972, 2.028]
Pernambuco	11.537	[9.336, 13.903]
Amazonas	9.770	[6.689, 14.797]
Piauí	26.50	[1.735, 3.796]
Amapá	0.299	[0.163, 0.457]
Paraná	24.465	[19.121, 31.052]
Bahia	10.211	[7.582, 13.304]
Rio de Janeiro	9.788	[6.700, 13.761]
Ceará	6.967	[4.397, 10.813]
Rio Grande do Norte	1.230	[0.705, 1.833]
Distrito Federal	13.036	[11.223, 14.998]
Rondônia	0.502	[0.141, 0.959]
Espírito Santo	4.021	[2.853, 5.709]
Roraima	−0.012	[−0.119, 0.117]
Goiós	6.349	[3.413, 10.248]
Rio Grande do Sul	7.516	[2.343, 14.642]
Maranhão	9.80	[6.35, 14.58]
Santa Catarina	4.396	[1.655, 7.998]
Minas Gerais	6.320	[1.580, 12.928]
Sergipe	1.851	[1.370, 2.341]
Mato Grosso do Sul	9.276	[6.377, 12.874]
São Paulo	49.934	[22.327, 91.296]
Mato Grosso	1.515	[0.843, 23.07]
Tocantins	1.172	[0.889, 1.454]
Pará	6.403	[4.842, 8.280]

to check the entire process on the calculation of the under-reporting rates and all numerical and graphical results. The graphics with the cases and deaths series from the *DT_SARI* and the marking of the detected events are presented in this notebook for all states. Also, the site contains graphics with the evolution of under-reported records over the weeks after COVID-19 for each state. There it is possible to see whether under-reported records increase, decrease or remain constant over time.

The Harbinger[6] framework was used for detecting events in time series (adaptive normalization and change finder). It receives the time series and parameters and returns the detected events. The parameters used are those defined in "method and parameter selection" section.

[6] Available at https://eic.cefet-rj.br/~dal/harbinger/.

Table 4 Errors of the models (deaths)

State	$\overline{\epsilon}$	$[\overline{\epsilon}_{min}, \overline{\epsilon}_{max}]$
Acre	0.480	[0.284, 0.683]
Paraíba	0.585	[0.402, 0.816]
Alagoas	0.293	[0.146, 0.452]
Pernambuco	0.325	[0.128, 0.552]
Amazonas	0.670	[0.391, 1.075]
Piauí	0.185	[0.024, 0.417]
Amapá	0.047	[0.007, o.102]
Paraná	3.015	[2.086, 4.005]
Bahia	0.847	[0.571, 1.142]
Rio de Janeiro	1.066	[0.531, 1.660]
Ceará	0.670	[0.381, 1.107]
Rio Grande do Norte	0.409	[0.238, 0.634]
Distrito Federal	0.422	[0.271, 0.618]
Rndônia	0.056	[−0.025, 0.155]
Espírito Santo	0.381	[0.150, 0.661]
Roraima	0.009	[−0.020, 0.053]
Goiós	0.940	[0.496, 1.454]
Rio Grande do Sul	0.902	[0.175, 1.870]
Maranhão	0.093	[0.029, 0.186]
Santa Catarina	0.632	[0.247, 1.054]
Minas Gerais	0.993	[0.147, 2.085]
Sergipe	0.119	[0.047, 0.210]
Mato Grosso do Sul	0.976	[0.451, 1.592]
São Paulo	3.941	[1.178, 8.057]
Mato Grosso	0.246	[0.076, 0.457]
Tocantins	0.302	[0.197, 0.432]
Pará	0.449	[0.225, 0.694]

For each state, two time series were submitted to the process described in "Methods" section, both from the InfoGripe dataset on hospitalizations for SARI (DT_SARI). The first is the weekly series with information on the number of registered SARI cases in the state. The second is the weekly series with information on the number of SARI deaths in the state.

Under-reporting rates were calculated for states where it was found that there were, in fact, novelty and under-reported notification. For this, two independent tests were carried out using the Wilcoxon test. The average error observed in the pre-novelty period ($\overline{\epsilon}$) was compared with the novelty (η_i) to check if there was a novelty. To check if there was an under-reported notification, the number of novelty calculated (η_i) was compared with the number classified as SARS-CoV-2 at InfoGripe data (cov_i) in a paired test. Then, in both cases, only when there is a relevant difference at a significance level of 0.05, the under-reporting rates were calculated.

Results

This work focuses on estimating under-reporting rates for cases and deaths of COVID-19. In “Data analytics” section, exploratory analysis is conducted. It contains discussions based on event detection (change points and anomaly) over the SARI time series. These findings bring valuable information to help understand the disease scenario in the most affected states. Besides, they helped to evaluate the choice of the method and the confidence of the estimates. Then, “Under-reporting rates” section briefly discusses the characteristics of the under-reporting rates calculated. Finally, “Evolution of the under-reporting rates” section presents the evolution of under-reporting in the period considered in this work.

Data Analytics

The detection of change points and anomalies in the time series of SARI hospitalization in Brazil was an important aspect to understand the beginning process of the pandemic situation of COVID-19 in the country. It also enabled the analyses of epidemic moments over the last years. In Figs. 2 and 3, it is possible to observe the behavior of data and specificity of the most affected Brazilian state.[7]

Amazonas state is the epidemic center in the North region, and its capital, Manaus, was the first capital from Brazil to suffer from a wave of deaths. The state presented in 2019 an increase in the number of hospitalizations. This increase is also observed in other states from 2016 until 2019. The Amazonas time series shows some anomalies, but just one change point for both the number of cases and deaths. The change point in the number of cases and deaths is marked in the 11th epidemiological week of 2020. The state reaches its peak of hospitalizations and deaths at the 17th epidemiological week and now presents a decrease in the curve.

In the Northeast region, it is possible to highlight the cases and deaths at Ceará, Pernambuco, and Bahia. Both Ceará and Pernambuco displayed the highest numbers in the region. All three states present both of the change points in the 10th week. Pernambuco and Ceará, respectively, reached their peaks of hospitalizations in the 18th (more than 1000 cases) and 19th week (more than 1800 cases). The peak for deaths for both of these states is located in the 18th week. In Bahia and Pernambuco, the number of cases and deaths show, between 2016 and 2019, a similar increase and decrease in shaping a curve between March and July.

Distrito Federal, located in the Central-West region of Brazil, was then considered one of the main focuses of COVID-19 contagion beside Rio de Janeiro and São Paulo. Previously, the peak of the number of cases in Distrito Federal was August of 2009, during the H1N1 epidemic. The pandemic superseded this high number in 2020. Besides, when analyzing the number of deaths caused by H1N1, it was not as expressive as the number of deaths registered by COVID-19.

[7] The graphics for all states are available at https://eic.cefet-rj.br/~dal/covid-19-under-report/.

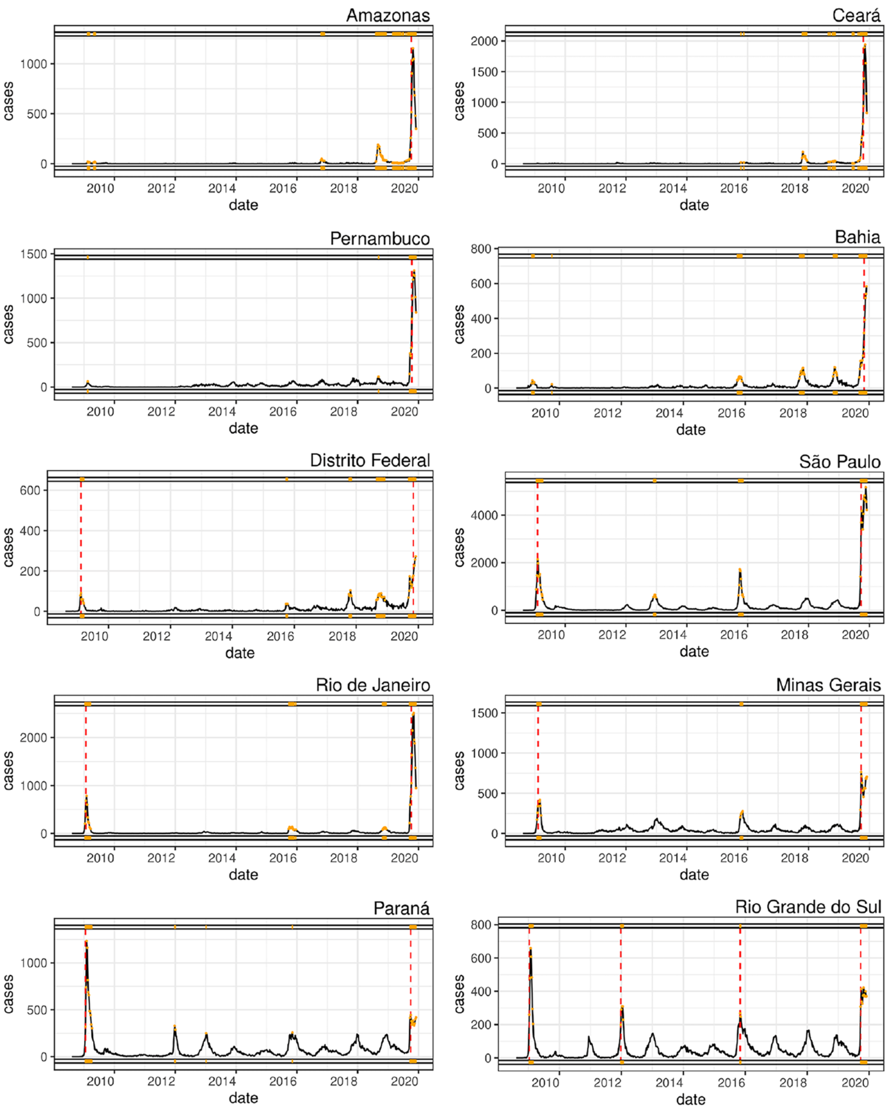


Fig. 2 Event detection in time series of cases

The Southeast is the most populous region and the most infected area in the country. São Paulo was the first state to register a case (February) and death (March) by COVID-19. It is still the epicenter of the disease in Brazil. The state has the mark of the change point for cases and deaths in the 10th week.

Rio de Janeiro, also in the Southeast region, was impacted by SARS-CoV-2. It is possible to observe in cases two change points. The first one is 2009 and the second in 2020. However, the number of observed change points for the number of deaths occurred only once, in 2020, showing the seriousness of this pandemic.

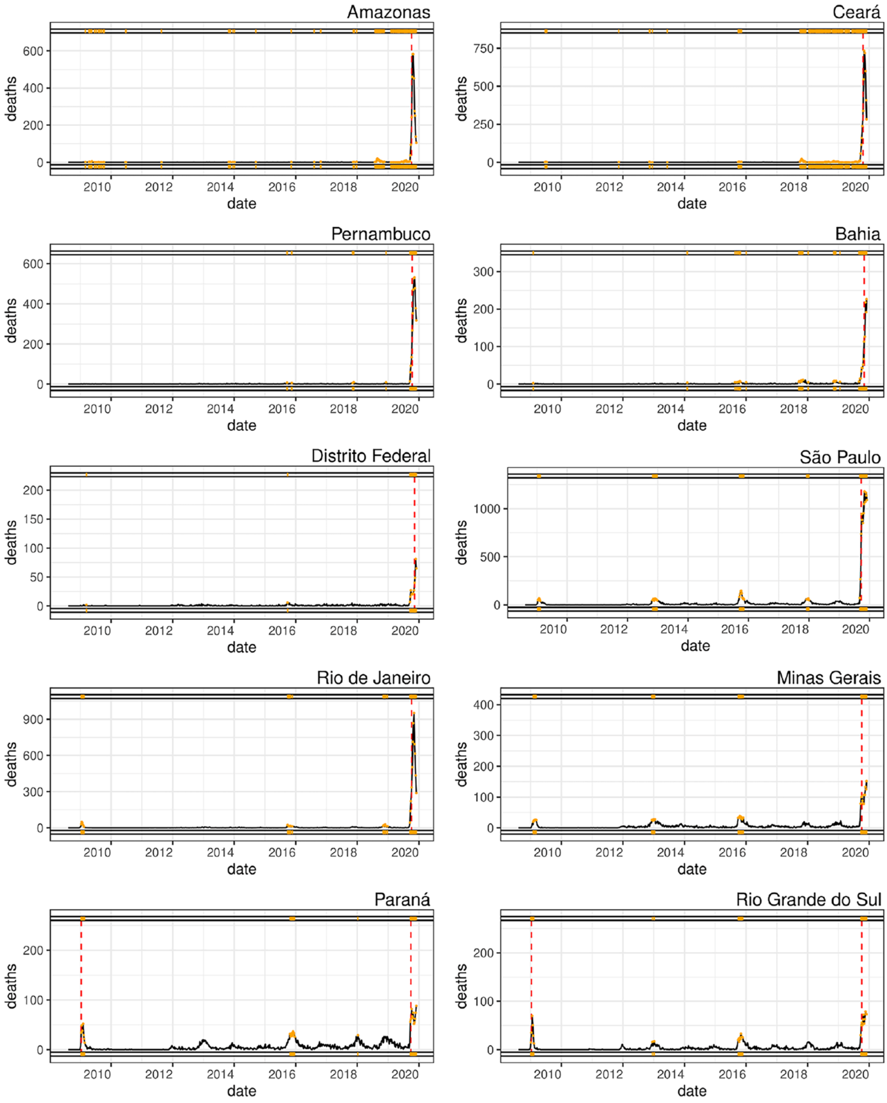


Fig. 3 Event detection in time series of deaths

Another state in the Southeast is Minas Gerais. It registered outliers in 2015 and more stable behavior between 2017 and 2019 for the numbers of cases and deaths. In 2020 the change point was detected in the 11th epidemiological for both cases and deaths.

The 2009 H1N1 crisis also impacted the states in the south region. According to the time series, it is noticeable that Paraná and the Rio Grande do Sul were affected in the number of cases. On the other hand, if we compare the number of deaths, we can observe and analyze the lethality between these two epidemic moments. Paraná

Table 5 Under-reporting rates of cases of COVID-19 for the states of Brazil

State	Cum. novelty (*DT_SARI_c*)	Cum. cases (*DT_SARI_c*)	Cases rate	Disclosed cum. cases (*DT_MH_c*)
Acre	356	297	0.198 ± 0.027	12913
Alagoas	2856	1520	0.879 ± 0.006	33521
Amazonas	7453	5080	0.467 ± 0.016	69022
Amapá	488	337	0.450 ± 0.008	27901
Bahia	4416	2936	0.504 ± 0.018	65244
Ceará	13,028	7804	0.669 ± 0.008	106628
Distrito Federal	2569	2094	0.227 ± 0.016	42766
Espírito Santo	1039	924	0.124 ± 0.029	41652
Goiás	2298	1306	0.760 ± 0.048	21620
Maranhão	3144	1597	0.969 ± 0.007	78115
Minas Gerais	10,076	3584	1.811 ± 0.029	40966
Mato Grosso do Sul	852	530	–•	7307
Mato Grosso	1945	884	1.200 ± 0.015	13805
Pará	10,924	7449	0.467 ± 0.004	99313
Paraíba	2213	1272	0.740 ± 0.009	44242
Pernambuco	8987	5418	0.659 ± 0.008	57089
Piauí	2558	1535	0.666 ± 0.013	18665
Paraná	4000	2238	0.787 ± 0.047	19819
Rio de Janeiro	18,786	11483	0.636 ± 0.006	108803
Rio Grande do Norte	1873	1361	0.376 ± 0.006	24253
Rondônia	631	523	0.207 ± 0.012	19273
Roraima	401	260	0.541 ± 0.008	13078
Rio Grande do Sul	4896	2515	0.947 ± 0.043	25000
Santa Catarina	1767	1101	0.605 ± 0.046	23808
Sergipe	810	558	0.451 ± 0.014	23319
São Paulo	57,546	37,025	0.554 ± 0.019	265581
Tocantins	630	389	0.619 ± 0.013	9966

• The difference between computed novelty and reported values as SARS-CoV-2 was not statistically significant

is an example of that analysis, where the maximum point of cases in 2009 surpasses 5,000 records. Meanwhile, the top of 2020 cases (until the current moment) is less than 1000. Nonetheless, when observing the number of deaths, the highest numbers occur in 2020, during the COVID-19 pandemic.

Under-Reporting Rates

The under-reporting rates were computed according to the proposed methodology. Tables 5 and 6 show the values of the under-reporting rates of cases and deaths for the 27 states of Brazil (columns cases rate and deaths rate, respectively). The rates

Table 6 Under-reporting rates of deaths by COVID-19 for the states of Brazil

State	Cum. novelty (*DT_SARI_d*)	Cum. deaths (*DT_SARI_d*))	Death rate	Disclosed cum. deaths (*DT_MH_d*)
Acre	135	160	–•	351
Alagoas	1082	797	0.357 ± 0.003	993
Amazonas	3288	2169	0.516 ± 0.003	2772
Amapá	242	154	0.574 ± 0.006	406
Bahia	1523	1133	0.345 ± 0.005	1697
Ceará	4437	3543	0.252 ± 0.002	5981
Distrito Federal	595	465	0.280 ± 0.007	537
Espírito Santo	689	643	0.072 ± 0.007	1507
Goiás	585	454	0.288 ± 0.018	429
Maranhão	1480	1080	0.371 ± 0.002	1943
Minas Gerais	1582	853	0.855 ± 0.021	882
Mato Grosso do Sul	104	89	–•	68
Mato Grosso	238	198	0.202 ± 0.017	527
Pará	4176	3263	0.280 ± 0.002	4834
Paraíba	789	629	0.255 ± 0.006	896
Pernambuco	3520	2773	0.269 ± 0.002	4708
Piauí	457	352	0.297 ± 0.011	592
Paraná	761	471	0.616 ± 0.034	575
Rio de Janeiro	5573	4170	0.337 ± 0.003	9789
Rio Grande do Norte	646	546	0.184 ± 0.007	909
Rondônia	206	187	0.102 ± 0.007	476
Roraima	307	195	0.574 ± 0.003	281
Rio Grande do Sul	984	496	0.983 ± 0.029	554
Santa Catarina	334	250	0.337 ± 0.027	304
Sergipe	222	194	0.143 ± 0.008	605
São Paulo	13,253	9458	0.401 ± 0.007	14263
Tocantins	160	136	0.177 ± 0.020	191

• The difference between computed novelty and reported values as SARS-CoV-2 was not statistically significant

shown are calculated for the period between the week detected by the event detection methods (see Table 2) and the epidemiological week 26 (which corresponds to the date 27/06/2020). Thus, the periods considered vary for cases and deaths and between states.

The second column of both tables (cum. novelty) presents the novelty values (η_i) computed according to the methodology. In the third column (cum. cases *DT_SARI_c* and cum. deaths *DT_SARI_d*) are the number of cases/deaths classified as SARS-CoV-2 in InfoGripe data. In the fifth column (disclosed cum. cases *DT_MH_c* and disclosed cum. deaths *DT_MH_d*) are the number of cases/deaths reported by the Ministry of Health, for comparison purposes. The information published by the Ministry of Health is all confirmed cases/deaths of COVID-19. They are presented regardless of whether there was hospitalization for SARI or not, so they capture a broader number of reported records.

The under-reporting rates presented in this paper can be applied to compute the under-reported cases or deaths of COVID-19 in each state. It is calculated by multiplying the under-reporting rates with the number of confirmed cases or deaths of COVID-19. The result can be added to reported cases/deaths to estimate the expected number of cases or deaths of COVID-19 in the state.

The under-reporting rates of cases vary between 0.124 and 1.811, while the under-reporting rates of deaths vary between 0.072 and 0.983. Among the states for which it was possible to calculate the two rates, most had a higher under-reporting rate of cases than under-reporting rate of deaths. Only the states of Rio Grande do Sul, Roraima, Distrito Federal, Amazonas, and Amapá behaved differently.

There is no dominant pattern between states in each region of Brazil. It suggests that under-reporting is a characteristic of each state. The regional similarity is not a relevant factor. The states of Minas Gerais and Mato Grosso have the highest rates of under-reporting of cases. The rate of under-reporting of deaths is high in the Rio Grande do Sul and the Minas Gerais.

The Distrito Federal, São Paulo, and Rio de Janeiro are identified as the focus of the contagion of COVID-19 in Brazil. Nevertheless, these states are not among the ones with the highest rates of under-reporting. It may be because they might be better structured and less susceptible to reporting failures. This same observation is not valid for the states Mato Grosso and Minas Gerais. They are respectively from the mid-west and Southeast regions. They have the highest rates of under-reporting of cases across Brazil.

The proposed model did not capture the under-reporting of cases in the Mato Grosso do Sul. Similar behavior occurred for under-reporting deaths in the states of Acre and Mato Grosso do Sul. These are the cases in which under-reporting cannot be observed ($^{\bullet}$).

Regarding the margin of error considered for the case rates, the states of the south region are highlighted. A factor that may have been determinant for this result is their historical temperature. As they have low temperatures, they generally, a higher number of SARI records. Thus, the novelty modeled in this work takes longer to be noticed, as it needs to reach even higher values to provide statistically significant changes.

Evolution of the Under-Reporting Rates

To create a better characterize the behavior of underrates-report, we analyze them week by week. It is important to have in mind that the COVID-19 tests were not available in most states at the beginning of the pandemic (11th week). Therefore, aiming for a better comparison, we present the analysis from the 12th week for all states.

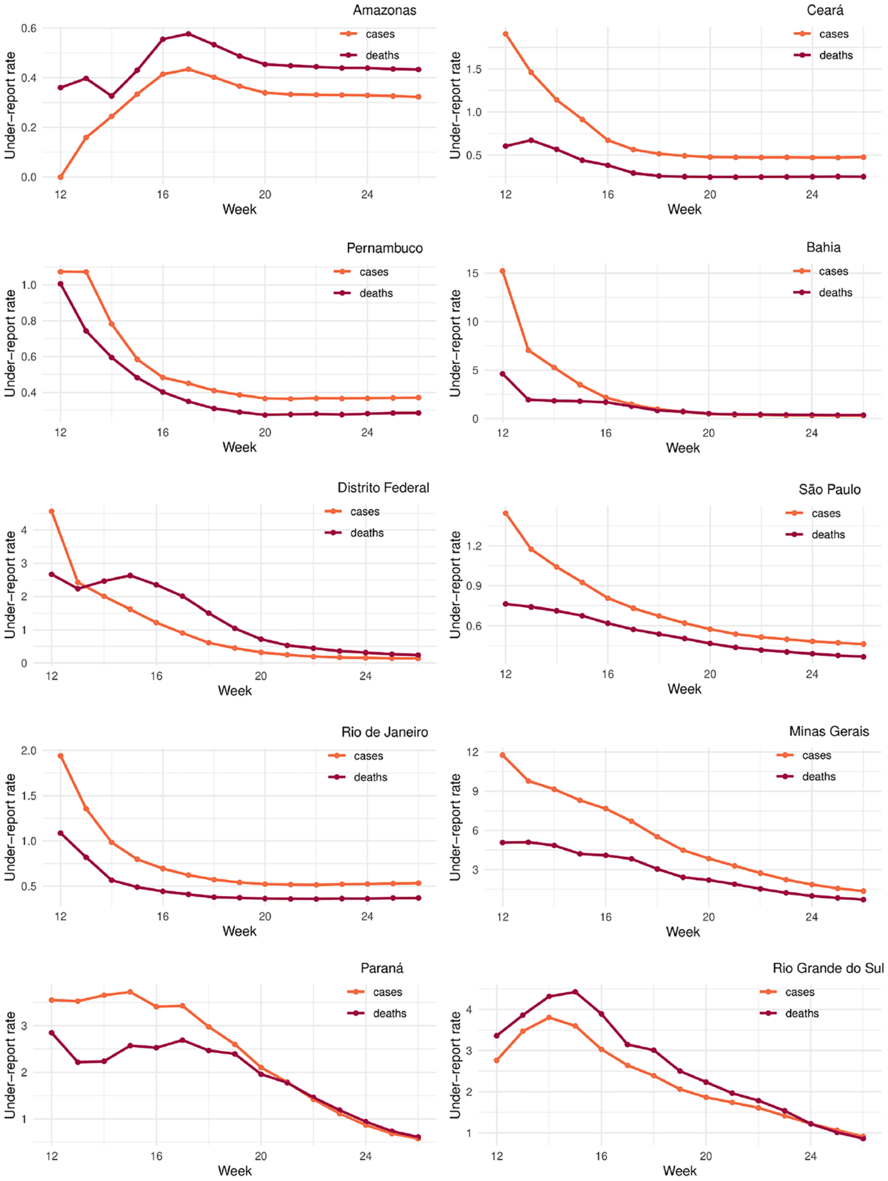


Fig. 4 Under-report rates

The lack of tests for the population results in an increased rate of under-report in the beginning. Over time, tests are expected to occur more, and the rates start to decrease. This explanation can be observed in the weekly rates graphs (Fig. 4).

As it can be observed, under-report rates tend to stabilize throughout time. This convergence enables more confidence in computed under-report rates. Besides, it shows that even when more tests for COVID-19 are available, there is still a high under-reporting rate for some states like Minas Gerais and the Rio Grande do Sul.

Discussion

The three sections of the results complement each other. Data analytics (with results presented in “Data analytics”) is used to set the parameters to be applied in the modeling of time series and determinant to calculate under-reporting rates. The subsequent analysis (with results presented in “Evolution of the under-reporting rates” section) shows the trend towards stability for the behavior of the calculated under-reporting rates. When rates are stable, the long-term estimation is more reliable, as there is no significant change in rate values over time.

Limitations should be noted. One limitation is inherent to the dataset used. In times of epidemic, health services tend to be more sensitive and report more occurrences. Thus, the increase in the number of SARI cases in 2020 is partially justified by the over-notification of health units. This super notification, however, is mitigated when only hospitalized cases are observed.

Another limitation is due to random noise ϵ_i. The states with higher ϵ_i are slower to characterize the novelty η_i. Again, the computed under-reporting rates presented in this paper are conservative. They can be improved by predicting ϵ_i using autoregressive models.

Since the under-reporting is inferred from SARI data, estimates are limited to cases of COVID-19, who were hospitalized from the specific symptoms: fever, cough or sore throat, dyspnoea, or oxygen saturation below 95% and difficulty to breathe. It corresponds to a portion of the cases of COVID-19, as many individuals have milder symptoms or are even asymptomatic. Thus, we can consider the computed under-reporting rates as conservative since it only considers symptomatic and hospitalized disease cases.

For this same reason, we believe that the results are better characterized for under-reporting of deaths than cases. It is reasonable since people who died are much more likely to have been hospitalized and, therefore, present in SARI data. It is quite clear when looking at Tables 5 and 6. The cases reported by the Ministry of Health mostly account for more cases than those determined by novelty. Conversely, the number of deaths found by novelty is sometimes even higher than the ones presented by the Ministry of Health.

An important observation that must be highlighted is the occurrence of under-reporting with the impact of COVID-19 on the Health System. From the moment that health surveillance fails to identify cases—due to under-reporting at times—it becomes more difficult to control its dissemination. With that, the dynamics and the complexity of the disease changes, and the Health System is overloaded. A consequence of that is to preclude people from getting the proper treatment not just for COVID-19 but also for other diseases, leading to an increase of deaths without medical assistance and ill-defined causes compared to last years [8].

Conclusions

This paper estimates the rates of under-reporting of cases and deaths in the states of Brazil. The methodology studies the time series of hospitalized SARI cases as a proxy variable for COVID-19. The paper contributes by combining data analytics (event detection methods) and time series modeling (inertia and novelty concepts). Data analytics ensures transparency and consistency in the choice of the adopted parameters. In contrast, novelty and inertia enable an understandable approach to estimate under-report.

COVID-19 causes a rupture in the SARI series inertial behavior, changing the statistical properties of the time series. Event detection techniques identify this rupture. Assuming that the change that occurred is due to COVID-19, the computed novelty then corresponds to estimates of the values of cases and deaths from the disease. From this, under-reporting rates were computed for both cases and deaths.

The rates of under-reporting of cases were estimated for all states except for Mato Grosso do Sul. The values vary between 0.124 (Espírito Santo) and 1.811 (Minas Gerais), thus reaching almost two under-reported cases for each notified case. The novelty observed by our SARI analysis in the states is lower, in their majority, compared to the cases reported by the Ministry of Health. It is expected since many diagnosed cases of COVID-19 are asymptomatic.

Under-reporting rates for deaths were estimated for 25 of the 27 states in Brazil. For the states of Acre and Mato Grosso do Sul, the under-report was not verified, and, therefore, death rates were not calculated for these states. Rates vary between 0.072 (Espírito Santo) and 0.983 (the Rio Grande do Sul), thus indicating that there may be more than twice as many deaths as reported. The novelties for death cases using SARI analysis in the states are commonly higher than those notified by the Ministry of Health. It helps to corroborate the justification that the death rates are better estimated since SARI covers most of the individuals who die.

No pattern of behavior was observed for the events detected or for the evolution and values of under-reporting rates between states in the same Brazilian region. Therefore, it is observed that the states behave in different and independent ways concerning the occurrence/notification of COVID-19. The analysis for each state allows heads of state to make strategic decisions about avoiding the spread of the disease in each geographic area.

The methodology developed in this paper can be adapted to support the under-report rate for other diseases as long as it exists a proxy variable that presents an inertial behavior. Besides, the methodology can also support the detection of outbreaks, as it uses both the combination of event detection and inertia concepts.

Acknowledgements The authors thank CNPq, CAPES (finance code 001), and FAPERJ for partially funding this research.

Author Contributions All authors contributed equally to the study. EO conceptualized the study design. MP and RFS acquired the data. BP and LB conducted data analysis and interpretation. RS, LE, CS, RC, FP, and JS revised it critically for intellectual content. All authors have the approval of the final version.

Funding BP, LB, FP were supported by CNPq. RS was supported by CAPES (finance code 001). MP was supported by FAPERJ. EO was supported by both CNPq and FAPERJ. The content is solely the responsibility of the authors. It does not necessarily represent the official views of the funding agencies. The funding agencies had no role in the study design, data collection, and analyses, decision to publish, or preparation of the manuscript.

Availability of data and materials The datasets analyzed during the current study and additional documentation is freely and openly available. It corresponds to weekly aggregated of anonymized records of patients contained in the SIVEP-Gripe. The Ministry of Health of Brazil is committed to respecting the ethical precepts and guaranteeing the privacy and reliability of the data. The continuously updated SARI data was obtained from the GitLab repository of Infogripe at https://gitlab.procc.fiocruz.br/mave/repo/-/blob/master/Dados/InfoGripe/dados_semanais_faixa_etaria_sexo_virus.csv. In this paper, we used a copy of Infogripe made on July 27th, 2020. It can be accessed on the GitHub repository at https://github.com/balthapaixao/Covid19_BR_underreport/tree/master/Aux_arqs.

The dataset used in this paper has not been reported in any other submission by us or anyone else.

The authors are committed to keeping the under-reporting rates updated. It means that the under-reporting rates will be recalculated periodically, provided that new data referring to SARI are made available by InfoGripe. The new under-reporting rates to be included will undergo the same methodological process described in this paper.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethics approval and consent to participate DATASUS provided the datasets used in this study. They were produced by aggregating and anonymizing all personal information of SARI registers contained in the SIVEP-Gripe. The Ministry of Health of Brazil is committed to respecting the ethical precepts and guaranteeing the privacy and reliability of the data.

References

1. Aminikhanghahi, S., Cook, D.: A survey of methods for time series change point detection. Knowl. Inf. Syst. **51**(2), 339–367 (2017)
2. Bastos, L., Niquini, R., Lana, R., Villela, D., Cruz, O., Coelho, F., Codeço, C., Gomes, M.: COVID-19 and hospitalizations for SARI in Brazil: a comparison up to the 12th epidemiological week of 2020. Cadernos de Saude Publica **36**(4) (2020)
3. Bastos, S.B., Cajueiro, D.O.: Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil (2020). arXiv:2003.14288
4. Bastos, S.B., Morato, M.M., Normey-Rico, D.O.: The Covid-19 (sars-cov-2) uncertainty tripod in Brazil: assessments on model-based predictions with large under-reporting (2020). arXiv:2006.15268
5. Callaway, E., Cyranoski, D., Mallapaty, S., Stoye, E., Tollefson, J.: The coronavirus pandemic in five powerful charts. Nature **579**(7800), 482–483 (2020)
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 1–58 (2009)
7. Chew, F., Doraisingham, S., Ling, A., Kumarasinghe, G., Lee, B.: Seasonal trends of viral respiratory tract infections in the tropics. Epidemiol. Infect. **121**(1), 121–128 (1998)
8. Covid-19, M.: Óbitos em excesso, dentro e fora de hospitais, mostram quadro de desassistência á saúde no município do rio de janeiro. Tech. rep., FIOCRUZ (2020). https://bigdata-covid19.icict.fiocruz.br/nota_tecnica_14.pdf
9. Ding, J., Xiang, Y., Shen, L., Tarokh, V.: Multiple change point analysis: fast implementation and strong consistency. IEEE Trans. Signal Process. **65**(17), 4495–4510 (2017)
10. Dowell, S., Shang Ho, M.: Seasonality of infectious diseases and severe acute respiratory syndrome—what we don't know can hurt us. Lancet Infect. Dis. **4**(11), 704–708 (2004)

11. Esling, P., Agon, C.: Time-series data mining. ACM Comput. Surv. **45**(1), 1–34 (2012)
12. Gujarati, D.: Basic Econometrics, 4th edn. McGraw-Hill/Irwin, Boston (2002)
13. Gupta, M., Gao, J., Aggarwal, C., Han, J.: Outlier detection for temporal data: a survey. IEEE Trans. Knowl. Data Eng. **26**(9), 2250–2267 (2014)
14. Ministry of Health, H.S.S.: Covid-19 epidemiological surveillance guide. Tech. rep. (2020). https://covid.saude.gov.br/
15. Ministry of Health, H.S.S.: Special epidemiological bulletin 14: Coronavirus Disease 2019. Tech. rep. (2020). https://portalarquivos.saude.gov.br/
16. InfoGripe.: Weekly bulletin—Week 18 of 2020. Tech. rep. (2020). https://covid-19.procc.fiocruz.br/
17. Krantz, S.G., Rao, A.S.R.S.: Level of underreporting including under diagnosis before the first peak of COVID-19 in various countries: preliminary retrospective results based on wavelets and deterministic modeling. Infect. Control Hosp. Epidemiol. **41**(7), 857–859 (2020)
18. Ksiazek, T., Erdman, D., Goldsmith, C., Zaki, S., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J., Lim, W., Rollin, P., Dowell, S., Ling, A.E., Humphrey, C., Shieh, W.J., Guarner, J., Paddock, C., Roca, P., Fields, B., DeRisi, J., Yang, J.Y., Cox, N., Hughes, J., LeDuc, J., Bellini, W., Anderson, L.: A novel coronavirus associated with severe acute respiratory syndrome. N. Engl. J. Med. **348**(20), 1953–1966 (2003)
19. Kuchar, J., Ashenfelter, A., Kliegr, T.: Outlier (anomaly) detection modelling in PMML. In: CEUR Workshop Proceedings, vol. 1875 (2017)
20. Lachmann, A., Jagodnik, K.M., Giorgi, F.M., Ray, F.: Correcting under-reported COVID-19 case numbers: estimating the true scale of the pandemic. medRxiv p. 2020.03.14.20036178 (2020)
21. Marson, F., Ortega, M.: COVID-19 in Brazil. Pulmonology **26**(4), 241–244 (2020)
22. Moriyama, M., Hugentobler, W.J., Iwasaki, A.: Seasonality of respiratory viral infections. Annu. Rev. Virol. **7**(1), 83–101 (2020)
23. Ogasawara, E., Martinez, L., De Oliveira, D., Zimbrão, G., Pappa, G., Mattoso, M.: Adaptive normalization: a novel data normalization approach for non-stationary time series. In: Proceedings of the International Joint Conference on Neural Networks (2010)
24. R Core Team.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2014)
25. Ribeiro, L.C., Bernardes, A.T.: Estimate of underreporting of COVID-19 in Brazil by Acute Respiratory Syndrome hospitalization reports. Tech. rep. (2020). https://econpapers.repec.org/paper/cdptecnot/tn010.htm
26. Rota, P., Oberste, M., Monroe, S., Nix, W., Campagnoli, R., Icenogle, J., Peñaranda, S., Bankamp, B., Maher, K., Chen, M.H., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J., Chen, Q., Wang, D., Erdman, D., Peret, T., Burns, C., Ksiazek, T., Rollin, P., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Fouchier, R., Günther, S., Osterhaus, A., Drosten, C., Pallansch, M., Anderson, L., Bellini, W.: Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science **300**(5624), 1394–1399 (2003)
27. Rothan, H., Byrareddy, S.: The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J. Autoimmun. **109**, 1–4 (2020)
28. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and Its Applications: With R Examples, 4th edn. Springer, New York (2017)
29. Silva, R.R., Velasco, W.D., da Silva Marques, W., Tibirica, C.A.G.: A Bayesian analysis of the total number of cases of the Covid 19 when only a few data is available. A case study in the state of Goias, Brazil. medRxiv (2020)
30. Takeuchi, J.I., Yamanishi, K.: A unifying framework for detecting outliers and change points from time series. IEEE Trans. Knowl. Data Eng. **18**(4), 482–492 (2006)
31. Tchidjou, H., Vescio, F., Boros, S., Guemkam, G., Minka, E., Lobe, M., Cappelli, G., Colizzi, V., Tietche, F., Rezza, G.: Seasonal pattern of hospitalization from acute respiratory infections in Yaoundé. Cameroon. J. Trop. Pediatr. **56**(5), 317–320 (2010)
32. Zheng, Z., Peng, F., Xu, B., Zhao, J., Liu, H., Peng, J., Li, Q., Jiang, C., Zhou, Y., Liu, S., Ye, C., Zhang, P., Xing, Y., Guo, H., Tang, W.: Risk factors of critical & mortal COVID-19 cases: a systematic literature review and meta-analysis. J. Infect. **81**(2), 16–25 (2020)