ESTIMATION OF COVID-19 UNDER-REPORTING IN BRAZILIAN STATES THROUGH SARI

Balthazar Paixão
CEFET/RJ
balthazar.paixao@eic.cefet-rj.br

Lais Baroni
CEFET/RJ
lais.baroni@eic.cefet-rj.br

Rebecca Salles
CEFET/RJ
rebeccapsalles@acm.org

Luciana Escobar
CEFET/RJ
luciana.vignoli@eic.cefet-rj.br

Carlos de Sousa
FIOCRUZ
carlos.sousa@bio.fiocruz.br

Marcel Pedroso
FIOCRUZ
marcel.pedroso@icict.fiocruz.br

Raphael Saldanha
FIOCRUZ
raphael.saldanha@icict.fiocruz.br

Rafaelli Coutinho
CEFET/RJ
rafaelli.coutinho@cefet-rj.br

Fabio Porto
LNCC
fporto@lncc.br

Eduardo Ogasawara
CEFET/RJ
eogasawara@ieee.org

ABSTRACT

Due to its impact, COVID-19 has been stressing the academy to search for curing, mitigating, or controlling it. However, when it comes to controlling, there are still few studies focused on under-reporting estimates. 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, the objective of this work is to estimate the under-reporting of cases and deaths of COVID-19 in Brazilian states using data from the Infogripe on notification of Severe Acute Respiratory Infection (SARI). The methodology is based on the concepts of inertia and the use of event detection techniques to study the time series of 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 to 2019) 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.

Keywords COVID-19 · under-reporting · time series · SARI · epidemiology

1 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 [4].

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 the one 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 [34].

Decision-making for the adoption of public policies in this pandemic scenario is stressing and, at the same time, 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 [18].

It is estimated that under-reporting is a relevant factor in determining the actual mortality rate and, if not considered, can cause significant misinformation [16]. Therefore, the objective of this work is to estimate the under-reporting of cases and deaths of COVID-19 in Brazilian states. If the possibility of testing the entire population is not viable, data from the Infogripe on notification of Severe Acute Respiratory Infection (SARI) are used.

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 to 2019) 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 over time [10]. 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.

Our paper stands out for estimating the under-reporting of cases and deaths of COVID-19 in Brazilian states. The methodology adopted includes everything from data acquisition and pre-processing to the calculation of under-reporting rates. Event detection methods are used to determine the parameters to be used in the methodology, and the estimate considers the weighted historical record. It adds value to the analysis, allowing a view more faithful to reality.

The results show that under-reporting rates vary significantly between states and that there is no standard for states in the same region in Brazil. It is noticed that the rates of under-reporting of cases are higher in the states of Minas Gerais (MG) and Mato Grosso do Sul (MS), and the highest rate of under-reporting of deaths is in the state of MG. In addition to the under-reporting rates, a brief exploratory analysis is presented, showing some interesting investigations that may help to understand the initial process of the COVID-19 pandemic situation in the country, as well as to analyze epidemic moments in last years.

This article is divided into seven sections in addition to this introduction. In Section 2, the theoretical foundation that supports the adopted methodology is presented, whereas Section 3 presents a summary of the published works regarding the under-reporting of COVID-19. Section 4 discusses the process of under-reporting estimation. Section 5 presents the experimental setup of the scenario in which the methodology was applied. Section 6 presents the most relevant search results. Finally, in Section 7 the main conclusions of the work are pointed out.

2 Background

In this section, we introduce some background for time series (Section 2.1), moving averages (Section 2.2), and event detection (Section 2.3) used in the context of this work.

2.1 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 $<y_1, y_2, \cdots, y_n>$ [9, 31]. A specific observation of a time series is represented as y_i, indexed in time by $i = 1, \ldots, n$, where y_1 represents the first observation and y_n is the most recent observation.

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

The i-th subsequence outdated seasonally in periodicity s of size p in a time series y, represented as $seq_{i,p}^s(y)$, is an ordered sequence of values $<y_{i-(p-1) \cdot s}, y_{i-(p-2) \cdot s}, \ldots, y_i>$, where $|seq_{i,p}^s(y)| = p$ and $p \leq i \leq |y|$. The sequence contains i-th observation and its $p-1$ predecessors outdated seasonally.

2.2 Seasonal Moving Averages

The i-th moving average $\overline{y}_{i,p}$ 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 Equation 1. The i-th 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.

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

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

The i-th seasonal moving average $\overline{y}_{i,p}^{s}$ and the i-th seasonal exponential moving average $\hat{y}_{i,p}^{s}$ of p terms in a time series y are similarly calculated replacing the continuous sequence $seq_{i,p}(y)$ with the seasonal sequence $seq_{i,p}^{s}(y)$, respectively, in Equations 1 and 2.

2.3 Event Detection

Event detection methods include the discovery of anomaly 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 [15]. Change points characterize a transition between different states in a process that generates the time series data [7, 32].

There are several methods to address the detection of anomalies [5, 11] and change points [2]. Among them, there are methods that consider the effects of inertia on time series data. As this work is based on inertial concepts [10], this section presents two methods of this group.

2.3.1 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 Equation 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 Equation 3.

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

2.3.2 Change Points by Change Finder

Change Finder is a technique that detects change points in univariate time series data [32]. 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 Equation 4. The highest scores for s, classified according to Equation 3, indicate the occurrence of 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 Equation 1. The detection of change points is then reduced to the outlier detection problem in $\overline{s}_p$ like the first phase.

$$s_i = (\hat{y}_i - y_i)^2, \; \hat{y}_i = \xi(y)_i \tag{4}$$

3 Related Work

Due to its relevance and novelty, 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.

Looking for similar work, we searched in the Scopus database in May 2020 with the search string (("covid-19" OR "covid19") AND ("sub-notification" OR "under-reporting" OR "under-reporting")). Only four papers in English were returned by the search. 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. Therefore, we accomplished a search for academic

---

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

works in Google Scholar to complement the research, employing the same words as the search string used in the Scopus database and on the same date.

From the returned works, ten were selected for reading. Most of them discuss the characteristics of COVID-19, such as under-reporting (cases and deaths) and its possible impact on different scenarios [1, 27]. Some works address the specificities of COVID-19 together with other diseases and the under-reporting rate as a factor to be considered [21, 24]. Others make different estimates related to COVID-19 and cite the under-reporting as a limitation or parameter [17, 30]. Three of the returned works are more specific regarding the under-reporting estimate, being more closely related to this work [13, 16, 26].

Krantz et al. [13] used harmonic analysis and wavelets to model the under-reporting of COVID-19 in several countries around the world. 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 were not available.

Similarly, to review the numbers of reported COVID-19 cases in several countries, Lachmann et al. [16] also estimated expected cases. For this, the author used demographic data and fixed mortality rates of the countries as well as 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 certain set of situations that must be true to justify the model.

Ribeiro et al. [26] 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.

Our work stands out for estimating the under-reporting of COVID-19 in Brazilian states weekly. In addition to under-reporting rates being calculated by week and by state, more detail than the cited works, 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.

4 Methodology

In seasonal phenomena, time series are generated by superimposing a seasonal process and random noises. Based on this premise, Equation 5 models the seasonal component of the time series, where y_i is an observation, $\hat{y}_{i-s,p}^s$ is the seasonal exponential moving average (SEMA) in the previous seasonality 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, while the influences that determine the behavior of the series are not changed [10].

$$y_i - \hat{y}_{i-s,p}^s + \epsilon_i = 0 \tag{5}$$

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, Equation 5 is expanded to Equation 6, that expresses novelty η_i and error $\hat{\epsilon}_i$. We have that $\hat{\epsilon}_i$ is approximated by the average error $\bar{\epsilon}$ observed in the pre-novelty period, *i.e.*, $\hat{\epsilon}_i$ is expected to be inside the interval confidence for $\bar{\epsilon}$ ($[\bar{\epsilon}_{min} - \bar{\epsilon}_{max}]$).

$$y_i - \hat{y}_{i-s,p}^s - \eta_i - \hat{\epsilon}_i = 0, \ \hat{\epsilon}_i \approx \bar{\epsilon}, \ \hat{\epsilon}_i \in [\bar{\epsilon}_{min} - \bar{\epsilon}_{max}] \tag{6}$$

Until the seasonal component $\hat{y}_{i-s,p}^s$ 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 $\hat{y}_{i-s,p}^s$ and the error $\hat{\epsilon}_i$ (approximated by $\bar{\epsilon}$). Equation 6 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 7 presents the calculation of the time series with absolute numbers of under-reported observations, where cov_i are observations classified as SARS-CoV-2.

$$sub_i = \eta_i - cov_i, \ t \leq i \leq |y| \tag{7}$$

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 8.

$$cur_i = \sum_{i=t}^{|y|} sub_i, \ t \leq i \leq |y| \tag{8}$$

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 9 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.

$$tx_i = \frac{cur_i}{cov_i} \tag{9}$$

5 Experimental Setup

This section discusses the experimental setup of the scenario in which the methodology was applied. Section 5.1 presents the process of data acquisition and preparation, whereas Section 5.2 describes the methods and parameters applied in the analysis. Section 5.3 presents the implementation details.

5.1 Data Acquisition and Preparation

InfoGripe is the primary dataset 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. These data come from the Influenza Epidemiological Surveillance Information System (SIVEP-Influenza) and present 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 [12]. 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_HM_c) and confirmed deaths (DT_HM_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 [19]. The values are used for purposes of comparison with the results obtained in this work.

---

[2]Data collected on May 28th, 2020

[3]Data collected on May 31th, 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)

5.2 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 [29], symptoms also present in Acute Respiratory Infection (ARI). Fever is another common symptom, even in mild cases of the disease. It is the reason for choosing of 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 [14, 28].

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 being the value that exceeds 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 the 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 to be calculated. For this, three parameters are required: p, i, and s (See Section 2), where 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 [6, 8, 22, 33]. 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. The event detection (targeting both change points and anomalies) in the series DT_SARI_c and DT_SARI_d consistently evidence, in several states, behavior change in two periods: *(i)* between the end of 2015 and the beginning of 2016, and *(ii)* between March and April 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 [3, 20]. The events appear from the 11th epidemiological week of 2020 for most states, *i.e.*, two weeks after the first confirmed case of COVID-19 in Brazil (this occurred in the 9th epidemiological week of 2020).

This result identifies the beginning of the novelty period in the data (t), *i.e.*, the 11th epidemiological week of 2020. Concerning the total number of weeks of the data, it corresponds to week 584 ($t = 584$). So, the model should be executed for the period before this date and extended until the last week of data, which is the week 590 ($|y|$). The parameter i admits values of the COVID-19 influence range (*i.e.*, $584 \leq i \leq 590$).

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 2015/2016 period. Events presented on this Figure 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 four years (2016 to 2019) are considered.

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

UF	CP Cases	CP Deaths	UF	CP Cases	CP Deaths
AC	-	-	PB	14/03/2020	14/03/2020
AL	04/04/2020	04/04/2020	PE	21/03/2020	28/03/2020
AM	28/03/2020	04/04/2020	PI	14/03/2020	14/03/2020
AP	21/03/2020	28/03/2020	PR	-	14/03/2020
BA	14/03/2020	21/03/2020	RJ	21/03/2020	28/03/2020
CE	28/03/2020	28/03/2020	RN	21/03/2020	14/03/2020
DF	14/03/2020	14/03/2020	RO	28/03/2020	28/03/2020
ES	14/03/2020	21/03/2020	RR	14/03/2020	14/03/2020
GO	14/03/2020	14/03/2020	RS	21/03/2020	21/03/2020
MA	22/02/2020	29/02/2020	SC	28/03/2020	14/03/2020
MG	14/03/2020	14/03/2020	SE	14/03/2020	14/03/2020
MS	14/03/2020	14/03/2020	SP	14/03/2020	14/03/2020
MT	14/03/2020	21/03/2020	TO	14/03/2020	18/04/2020
PA	04/04/2020	04/04/2020			

Table 3 summarizes the used parameters. The model errors (random noise) for this period for both the cases and deaths in each state are, respectively, described in Tables 4 and 5. Since ϵ_i follows a non-normal distribution, the interval confidence for $\bar{\epsilon}$ is computed by bootstrap with 1000 repetitions. Under-reporting rates were calculated for states where it was found that there were, in fact, novelty. Therefore, average error observed in the pre-novelty period ($\bar{\epsilon}$) was compared with the novelty (η_i) and assessed whether there is a relevant difference at a significance level of 0.05 using the Wilcoxon test.
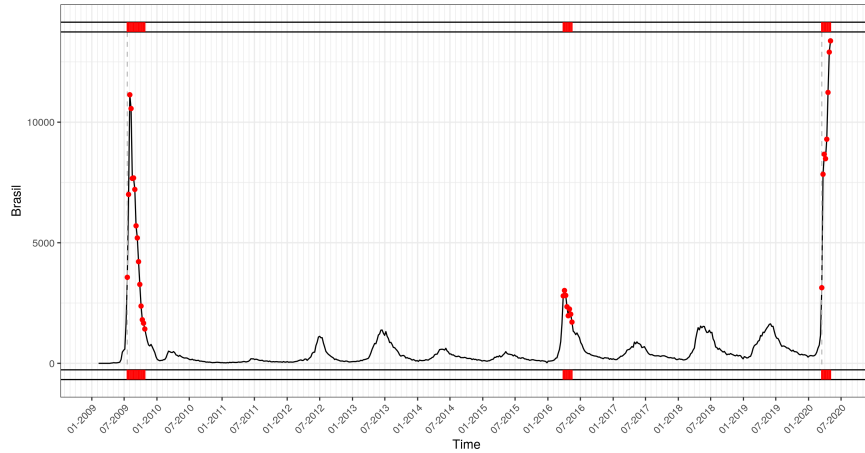

Figure 1: Events detected in the SARI cases curve in Brazil. The red dots mark anomalies (Adaptive Normalization), and the gray dotted lines mark the change points (Change Finder).

Table 3: Parameters

Parameter	Value
i	$t \leq i \leq |y|$
t	584
p	4
s	52

5.3 Implementation

The adopted methodology was implemented in R [25]. The code description and Jupyter notebook also developed in R complements this work[4]. In it, it is possible to check the entire process on the calculation of the under-reporting rates

---

[4]available at `https://eic.cefet-rj.br/~dal/covid-19-under-report/`

Table 4: Errors of the models (cases)

UF	$\bar{\epsilon}$	$[\bar{\epsilon}_{min}, \bar{\epsilon}_{max}]$	UF	$\bar{\epsilon}$	$[\bar{\epsilon}_{min}, \bar{\epsilon}_{max}]$
AC	1.727	[1.166, 2.344]	PB	2.198	[1.700, 2.821]
AL	1.482	[0.959, 2.092]	PE	11.537	[9.311, 13.81]
AM	9.770	[6.82, 14.343]	PI	2.651	[1.758, 3.979]
AP	0.299	[0.181, 0.478]	PR	24.465	[18.79, 31.21]
BA	10.211	[7.478, 13.31]	RJ	9.788	[6.514, 14.28]
CE	6.967	[4.372, 11.14]	RN	1.230	[0.705, 1.841]
DF	13.036	[11.19, 15.11]	RO	0.502	[0.162, 0.970]
ES	4.021	[2.789, 5.562]	RR	-0.012	[-0.12, 0.119]
GO	6.349	[3.787, 10.31]	RS	7.516	[1.965, 14.86]
MA	0.980	[0.617, 1.535]	SC	4.396	[1.316, 8.088]
MG	6.320	[1.449, 12.34]	SE	1.851	[1.391, 2.382]
MS	9.276	[6.668, 13.15]	SP	49.934	[21.59, 85.15]
MT	1.516	[0.855, 2.333]	TO	1.172	[0.909, 1.484]
PA	6.403	[5.012, 8.195]			

Table 5: Errors of the models (deaths)

UF	$\bar{\epsilon}$	$[\bar{\epsilon}_{min}, \bar{\epsilon}_{max}]$	UF	$\bar{\epsilon}$	$[\bar{\epsilon}_{min}, \bar{\epsilon}_{max}]$
AC	0.480	[0.298, 0.688]	PB	0.586	[0.383, 0.815]
AL	0.293	[0.151, 0.481]	PE	0.325	[0.120, 0.555]
AM	0.670	[0.399, 1.094]	PI	0.185	[0.015, 0.376]
AP	0.047	[0.007, 0.100]	PR	3.015	[2.129, 4.137]
BA	0.847	[0.566, 1.182]	RJ	1.066	[0.563, 1.662]
CE	0.670	[0.378, 1.082]	RN	0.409	[0.236, 0.613]
DF	0.423	[0.266, 0.603]	RO	0.056	[-0.028, 0.165]
ES	0.381	[0.161, 0.655]	RR	0.009	[-0.017, 0.050]
GO	0.940	[0.462, 1.466]	RS	0.902	[0.089, 1.717]
MA	0.093	[0.028, 0.169]	SC	0.632	[0.283, 1.075]
MG	0.993	[0.088, 2.061]	SE	0.119	[0.049, 0.196]
MS	0.976	[0.460, 1.658]	SP	3.941	[1.110, 8.098]
MT	0.246	[0.046, 0.443]	TO	0.302	[0.198, 0.418]
PA	0.449	[0.226, 0.719]			

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.

For the execution of the event detection methods, Adaptive Normalization and Change Finder, the Harbinger[5] framework was used for detecting events in time series. It receives the time series and parameters and returns the detected events. Thus, it was not necessary to implement these two techniques, just to invoke them from Harbinger. The parameters used are those defined in Section 5.2.

For each state, two time series were submitted to the process described in Section 4, 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, and the second is the weekly series with information on the number of SARI deaths. Under-reporting rates were calculated for states where it was found that there were, in fact, under-reported notification. Therefore, the number of novelty calculated (η_i) was compared with the number classified as SARS-CoV-2 at Infogripe data (cov_i) and assessed whether there is a relevant difference at a significance level of 0.05 using the Wilcoxon test.

6 Results

This work focuses on estimating under-reporting rates for cases and deaths of COVID-19. In Section 6.1 an exploratory analysis is conducted. It contains discussions that are based on the results of event detection (change points and anomaly) over the SARI time series. These findings bring valuable information to help understand the disease scenario

---

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

in the most affected states. Besides, they helped to evaluate the choice of the method and the confidence of the estimates. Then, the actual under-reporting rates are presented in Section 6.2.

6.1 Exploratory Data Analysis

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 Figures 2 and 3, it is possible to observe the behavior of data and specificity of the most affected Brazilian state[6].

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 (Figure 2a) and deaths (Figure 3a). The change point in the number of deaths and cases is marked, respectively, in the last week of March 2020 and one week later, which corresponds to the thirteenth and fourteenth epidemiological weeks.

In the Northeast region, it is possible to highlight the cases and deaths that occurred at Ceará (Figures 2b and 3b), Pernambuco (Figures 2c and 3c), and Bahia (Figures 2d and 3d). Both Ceará and Pernambuco displayed the highest numbers in the region. The Ceará state shows the same behavior as Amazonas, presenting the change points in the thirteenth and fourteenth weeks. Meanwhile, in Pernambuco, both deaths and cases occurred one week early. 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. The peak of the number of cases (Figure 2e) in Distrito Federal is in August of 2009, during the H1N1 epidemic. However, the number of deaths (Figure 3e) caused by H1N1 was not as expressive as the numbers registered by COVID-19.

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 and death by COVID-19. They, respectively, occurred in February and March. It is still the epicenter of the disease in Brazil. The state has the mark of the change point for cases (Figure 2f) and deaths (Figure 3f) at the eleventh epidemiological week. It quickly reached the highest registered numbers, more than 4000 cases and 800 deaths in a week.

Rio de Janeiro, also a southeast region, was impacted by SARS-CoV-2. It is possible to observe in the cases (Figure 2g) 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 (Figure 3g) occurred only once, in 2020, showing the seriousness of this pandemic.

Another southern state is Minas Gerais. It registered outliers in 2015 and more stable behavior between 2017 and 2019 for the numbers of cases (Figure 2h) and deaths (Figure 3h). In 2020 the method used detected the change point in the same epidemiological week not only for cases but also for the number of deaths.

The southern states were also impacted by the 2009 H1N1 crisis. According to the time series it is noticeable that Paraná and the Rio Grande do Sul were affected in the number of cases (respectively Figures 2i and 2j). On the other hand, if we compare the number of deaths, we can observe and analyze the lethality between these two epidemic moments. Paraná is an example of that analysis, where the maximum point of cases in 2009 surpasses 5,000. Meanwhile, the top of 2020 cases (until the current moment) is less than 1,000. Nonetheless, when observing the number of deaths (Figure 3i), the highest numbers occurs in 2020.

6.2 Under-Reporting Rates

The under-reporting rates were computed according to the proposed methodology. Tables 6 and 7 show the values of the under-reporting rates of cases and deaths for the 27 states of Brazil. In the second column ($cum.novelty$) are the novelty values (η_i) calculated in 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 ($cum.cases\ DT\_HM\_c$ and $cum.deaths\ DT\_HM\_d$) are the number of cases/deaths reported by the Ministry of Health, for comparison purposes. The information published by the Ministry of Health are all confirmed cases/deaths of COVID-19, regardless of whether there was hospitalization for SARI or not, so they capture a broader number of reported records.

---

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

(a) Amazonas cases

(b) Ceará cases

(c) Pernambuco cases

(d) Bahia cases

(e) Distrito Federal cases

(f) São Paulo cases

(g) Rio de Janeiro cases
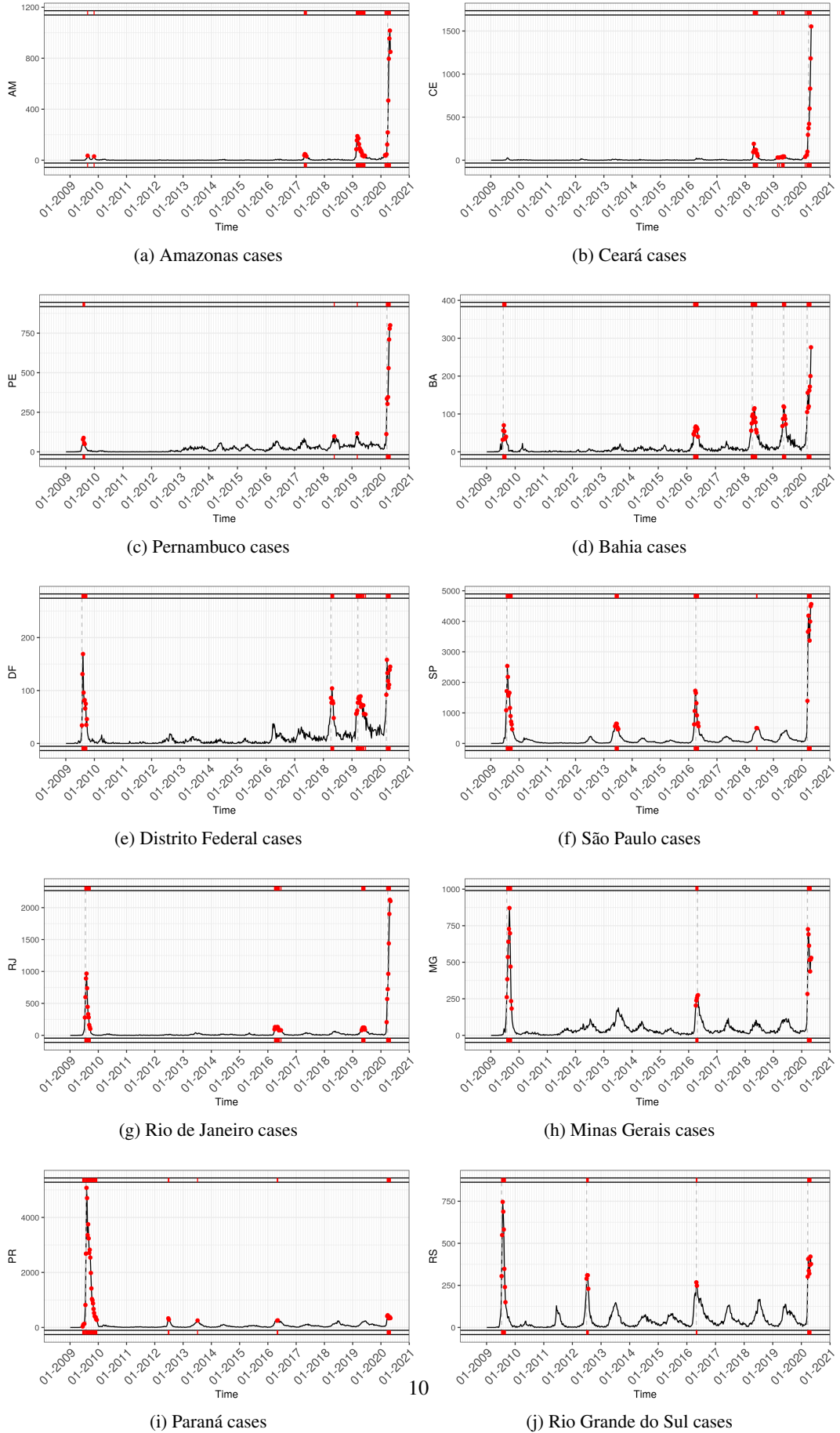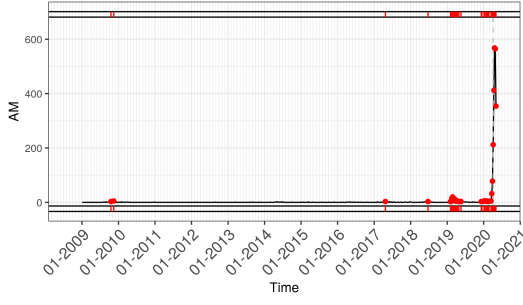
(h) Minas Gerais cases
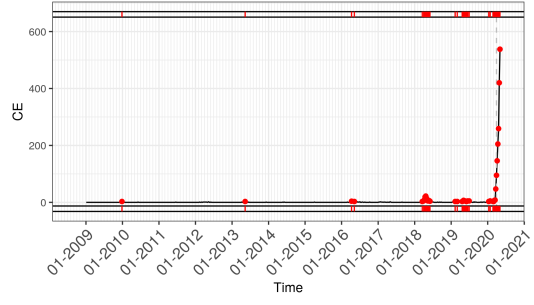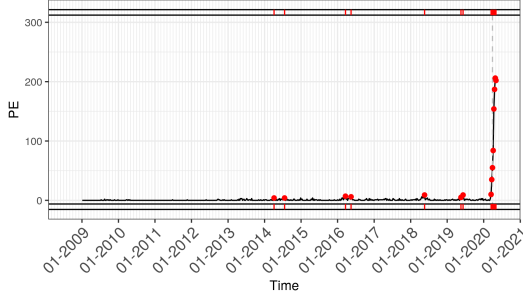
(i) Paraná cases

(j) Rio Grande do Sul cases

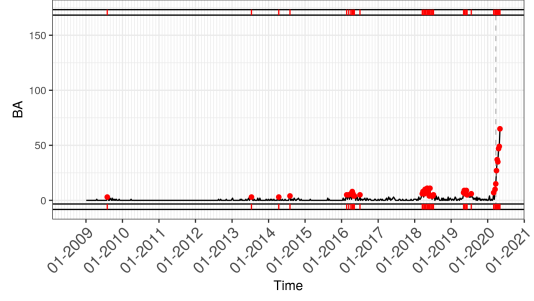Figure 2: Event detection in time series of cases
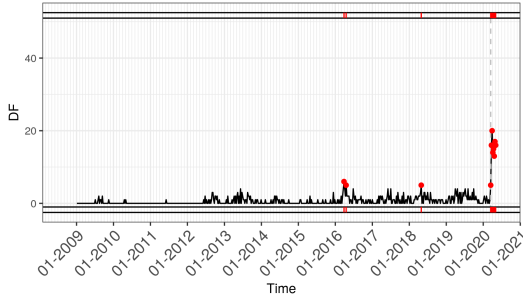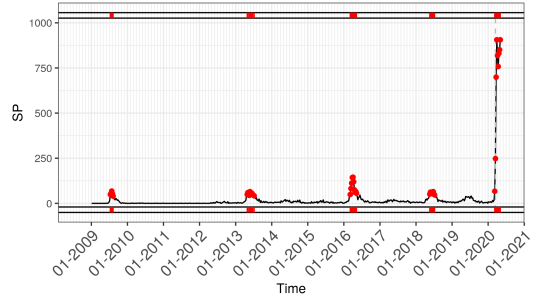
(a) Amazonas deaths

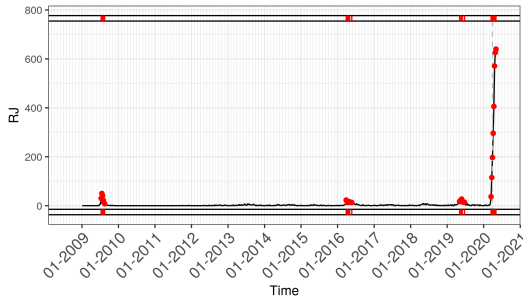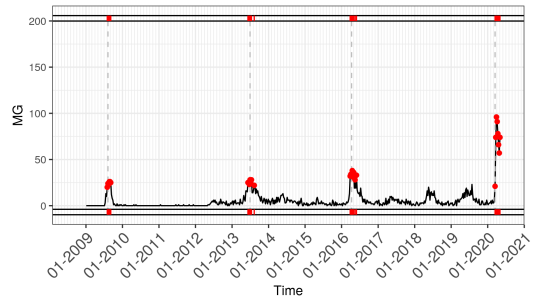(b) Ceará deaths

(c) Pernambuco deaths
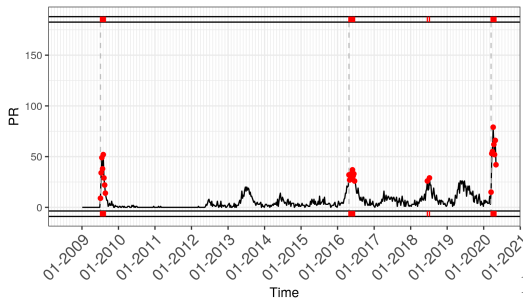
(d) Bahia deaths

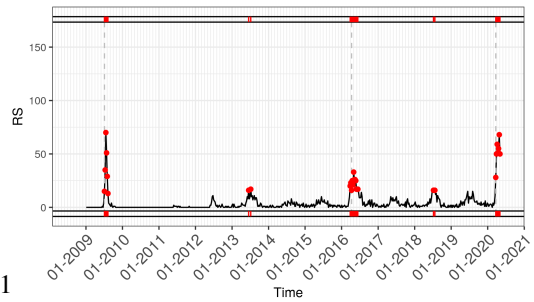(e) Distrito Federal deaths

(f) São Paulo deaths

(g) Rio de Janeiro deaths

(h) Minas Gerais deaths

(i) Paraná deaths

(j) Rio Grande do Sul deaths

11

Figure 3: Event detection in time series of deaths

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 computed 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.

Table 6: Under-reporting rates of cases of COVID-19 for the states of Brazil

UF	cum. novelty (DT_SARI_c)	cum. cases (DT_SARI_c)	cases rate	cum. cases (DT_HM_c)
AC	0	13	- °	553
AL	308	152	1.026 ± 0.026	1372
AM	3824	2165	0.766 ± 0.018	6062
AP	83	39	1.128 ± 0.026	1187
BA	832	350	1.377 ± 0.071	3267
CE	4704	2085	1.256 ± 0.015	8231
DF	401	251	0.598 ± 0.064	1566
ES	243	152	0.599 ± 0.086	2948
GO	363	162	1.241 ± 0.191	825
MA	650	132	3.924 ± 0.030	3805
MG	3553	484	6.341 ± 0.024	2023
MS	420	53	6.925 ± 0.110	266
MT	360	85	3.235 ± 0.071	331
PA	1390	909	0.529 ± 0.017	3460
PB	619	168	2.685 ± 0.030	1034
PE	3158	976	2.236 ± 0.018	8145
PI	602	186	2.237 ± 0.048	665
PR	1779	389	3.573 ± 0.136	1492
RJ	8069	3679	1.193 ± 0.009	10546
RN	386	207	0.865 ± 0.024	1366
RO	27	15	- °	653
RR	71	45	0.578 ± 0.022	668
RS	2175	615	2.537 ± 0.093	1619
SC	972	303	2.208 ± 0.096	2346
SE	92	62	0.484 ± 0.065	601
SP	25938	13057	0.987 ± 0.025	31174
TO	141	38	2.711 ± 0.053	191

° The difference between computed novelty and random noise was not statistically significant.

The under-reporting rates of cases vary between 0.484 and 6.925, while the under-reporting rates of deaths vary between 0.143 and 3.617. Among the states for which it was possible to calculate the two rates, the majority had higher under-reporting rate of cases than under-reporting rate of deaths. Only the states RS, DF and SE behave differently. DF is highlighted because it has a death rate almost 3 times higher than that of cases.

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 MG and MS have the highest rates of under-reporting of cases. The rate of under-reporting of deaths is high in the MG and the RS.

The DF, SP and RJ are identified as the focus of the contagion of COVID-19 in Brazil. Nevertheless, both DF and SP are not 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 MS and MG in the same regions (mid-west and southeast regions, respectively), which have the highest rates of under-reporting of cases across Brazil.

The proposed model did not capture under-reporting of cases in the AC and RO or deaths in the states of AC, RO, MS, MT, TO, GO, RR, AP, and ES. These are the cases in which either a novelty cannot be detected (°) or under-reporting cannot be observed (•). MS stands out since, despite having a high-rate of under-reporting of cases (second highest among states), the under-reporting of deaths was not observed.

Regarding the margin of error considered for the case rates, the states of the mid-west and south regions are highlighted. A factor that may have been determinant for this result is their historical temperature. As they have low temperatures,

Table 7: Under-reporting rates of deaths by COVID-19 for the states of Brazil

UF	cum. novelty (DT_SARI_d)	cum. deaths (DT_SARI_d)	death rate	cum. deaths (DT_HM_d)
AC	0	13	- °	21
AL	49	34	0.441 ± 0.029	58
AM	2023	1147	0.764 ± 0.003	501
AP	26	21	- •	40
BA	200	110	0.818 ± 0.027	123
CE	1429	983	0.454 ± 0.004	614
DF	90	33	1.727 ± 0.061	31
ES	91	84	- •	102
GO	61	36	- •	30
MA	60	31	0.935 ± 0.032	224
MG	434	94	3.617 ± 0.096	88
MS	26	8	- °	9
MT	34	19	- •	12
PA	473	414	0.143 ± 0.005	273
PB	133	86	0.547 ± 0.023	74
PE	653	369	0.770 ± 0.005	628
PI	86	34	1.529 ± 0.059	26
PR	287	83	2.458 ± 0.096	90
RJ	2236	1577	0.418 ± 0.003	951
RN	83	68	0.221 ± 0.029	59
RO	3	4	- °	23
RR	17	16	- •	9
RS	303	77	2.935 ± 0.104	62
SC	104	48	1.167 ± 0.062	52
SE	22	13	0.692 ± 0.077	14
SP	5131	3207	0.600 ± 0.010	2586
TO	13	16	- °	4

° The difference between computed novelty and random noise was not statistically significant.
• The difference between computed novelty and reported values was not statistically significant.

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.

7 Final Remarks

This study aimed to estimate the rates of under-reporting of cases and deaths in the states of Brazil. The methodology is based on the concepts of inertia and the use of event detection techniques to study the time series of hospitalized SARI cases. All methods and parameters used in the methodology are justified, based on the modeling or available data.

We introduced the concept of novelty about SARI analysis to observe the under-reporting of COVID-19. Consequently, COVID-19 causes a rupture in the SARI series inertial behavior, changing the statistical properties of the time series. This break is identified by event detection techniques. If the change 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.

Since the under-reporting is inferred from SARI data, estimates are limited to cases of COVID-19 that manifested specific symptoms (fever, cough or sore throat, dyspnea or oxygen saturation below 95% and difficulty to breathe) and were hospitalized. 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 of under-reporting rates as very conservative since it only considers symptomatic and hospitalized cases of the disease.

For this same reason, we believe that the results are better characterized for deaths than for cases, since people who died are much more likely to have been hospitalized and, therefore, present in SARI data. This is quite clear when looking at the Tables 6 and 7. While in the table of cases (Table 6) the data from the Ministry of Health mostly account

for many more cases than those determined in the novelty, in the Table of deaths (Table 7) the number of deaths found of the novelty are higher.

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 in which ϵ_i were higher 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.

Acknowledgments

References

[1] A. Abbara, D. Rayes, O. Fahham, O. Alhiraki, M. Khalil, A. Alomar, and A. Tarakji. Coronavirus 2019 and health systems affected by protracted conflict: The case of Syria. *International Journal of Infectious Diseases*, 96:192–195, 2020.

[2] S. Aminikhanghahi and D. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

[3] L. Bastos, R. Niquini, R. Lana, D. Villela, O. Cruz, F. Coelho, C. Codeço, and M. Gomes. 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.

[4] E. Callaway, D. Cyranoski, S. Mallapaty, E. Stoye, and J. Tollefson. The coronavirus pandemic in five powerful charts. *Nature*, 579(7800):482–483, 2020.

[5] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009.

[6] F. Chew, S. Doraisingham, A. Ling, G. Kumarasinghe, and B. Lee. Seasonal trends of viral respiratory tract infections in the tropics. *Epidemiology and Infection*, 121(1):121–128, 1998.

[7] J. Ding, Y. Xiang, L. Shen, and V. Tarokh. Multiple Change Point Analysis: Fast Implementation and Strong Consistency. *IEEE Transactions on Signal Processing*, 65(17):4495–4510, 2017.

[8] S. Dowell and M. Shang Ho. Seasonality of infectious diseases and severe acute respiratory syndrome - What we don't know can hurt us. *Lancet Infectious Diseases*, 4(11):704–708, 2004.

[9] P. Esling and C. Agon. Time-series data mining. *ACM Computing Surveys*, 45(1), 2012.

[10] D. Gujarati. *Basic Econometrics*. McGraw-Hill/Irwin, Boston; Montreal, 4 edition, Mar. 2002.

[11] M. Gupta, J. Gao, C. Aggarwal, and J. Han. Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267, 2014.

[12] InfoGripe. Weekly bulletin - Week 18 of 2020. Technical report, https://covid-19.procc.fiocruz.br/, May 2020.

[13] S. G. Krantz and A. S. R. S. Rao. Level of underreporting including underdiagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling. *Infection Control & Hospital Epidemiology*, 41(7):857–859, July 2020.

[14] T. Ksiazek, D. Erdman, C. Goldsmith, S. Zaki, T. Peret, S. Emery, S. Tong, C. Urbani, J. Comer, W. Lim, P. Rollin, S. Dowell, A.-E. Ling, C. Humphrey, W.-J. Shieh, J. Guarner, C. Paddock, P. Roca, B. Fields, J. DeRisi, J.-Y. Yang, N. Cox, J. Hughes, J. LeDuc, W. Bellini, and L. Anderson. A novel coronavirus associated with severe acute respiratory syndrome. *New England Journal of Medicine*, 348(20):1953–1966, 2003.

[15] J. Kuchar, A. Ashenfelter, and T. Kliegr. Outlier (anomaly) detection modelling in PMML. In *CEUR Workshop Proceedings*, volume 1875, 2017.

[16] A. Lachmann, K. M. Jagodnik, F. M. Giorgi, and F. Ray. Correcting under-reported COVID-19 case numbers: estimating the true scale of the pandemic. *medRxiv*, page 2020.03.14.20036178, Apr. 2020.

[17] H. Lau, V. Khosrawipour, P. Kocbach, A. Mikolajczyk, H. Ichii, J. Schubert, J. Bania, and T. Khosrawipour. Internationally lost COVID-19 cases. *Journal of Microbiology, Immunology and Infection*, 2020.

[18] F. Marson and M. Ortega. COVID-19 in Brazil. *Pulmonology*, 2020.

[19] H. S. S. Ministry of Health. Covid-19 epidemiological surveillance guide. Technical report, https://covid.saude.gov.br/, June 2020.

[20] H. S. S. Ministry of Health. Special epidemiological bulletin 14: Coronavirus Disease 2019. Technical report, https://portalarquivos.saude.gov.br/, Apr. 2020.

[21] P. Mohindra, C. Buckey, S. Chen, T. Sio, and Y. Rong. Radiation therapy considerations during the COVID-19 Pandemic: Literature review and expert opinions. *Journal of Applied Clinical Medical Physics*, 2020.

[22] M. Moriyama, W. J. Hugentobler, and A. Iwasaki. Seasonality of Respiratory Viral Infections. *Annual Review of Virology*, Mar. 2020.

[23] E. Ogasawara, L. Martinez, D. De Oliveira, G. Zimbrão, G. Pappa, and M. Mattoso. Adaptive Normalization: A novel data normalization approach for non-stationary time series. In *Proceedings of the International Joint Conference on Neural Networks*, 2010.

[24] J. Ong, B. Young, and S. Ong. COVID-19 in gastroenterology: A clinical perspective. *Gut*, 2020.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[26] L. C. Ribeiro, A. T. Bernardes, and others. Estimate of underreporting of COVID-19 in Brazil by Acute Respiratory Syndrome hospitalization reports. Technical report, https://econpapers.repec.org/paper/cdptecnot/tn010.htm, 2020.

[27] V. Ricoca Peixoto, C. Nunes, and A. Abrantes. Epidemic Surveillance of Covid-19: Considering Uncertainty and Under-Ascertainment. *Portuguese Journal of Public Health*, 2020.

[28] P. Rota, M. Oberste, S. Monroe, W. Nix, R. Campagnoli, J. Icenogle, S. Peñaranda, B. Bankamp, K. Maher, M.-H. Chen, S. Tong, A. Tamin, L. Lowe, M. Frace, J. DeRisi, Q. Chen, D. Wang, D. Erdman, T. Peret, C. Burns, T. Ksiazek, P. Rollin, A. Sanchez, S. Liffick, B. Holloway, J. Limor, K. McCaustland, M. Olsen-Rasmussen, R. Fouchier, S. Günther, A. Osterhaus, C. Drosten, M. Pallansch, L. Anderson, and W. Bellini. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*, 300(5624):1394–1399, 2003.

[29] H. Rothan and S. Byrareddy. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *Journal of Autoimmunity*, 109, 2020.

[30] T. W. Russell, J. Hellewell, S. Abbott, C. Jarvis, K. van Zandvoort, C. n. w. group, S. Flasche, A. Kucharski, and others. Using a delay-adjusted case fatality ratio to estimate under-reporting. Technical report, https://cmmid.github.io/topics/covid19/global_cfr_estimates.html, 2020.

[31] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer, New York, NY, 4 edition, Apr. 2017.

[32] J.-I. Takeuchi and K. Yamanishi. A unifying framework for detecting outliers and change points from time series. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):482–492, 2006.

[33] H. Tchidjou, F. Vescio, S. Boros, G. Guemkam, E. Minka, M. Lobe, G. Cappelli, V. Colizzi, F. Tietche, and G. Rezza. Seasonal pattern of hospitalization from acute respiratory infections in Yaoundé, Cameroon. *Journal of Tropical Pediatrics*, 56(5):317–320, 2010.

[34] Z. Zheng, F. Peng, B. Xu, J. Zhao, H. Liu, J. Peng, Q. Li, C. Jiang, Y. Zhou, S. Liu, C. Ye, P. Zhang, Y. Xing, H. Guo, and W. Tang. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *Journal of Infection*, 2020.