Submitted: 18 August, 2021 Accepted: 03 November, 2021 Published: 08 May, 2022 DOI:10.22514/sv.2021.253

Open Access
Signa Vitae

ORIGINAL RESEARCH

Data-driven tools for assessing and combating COVID-19 outbreaks in Brazil based on analytics and statistical methods

Raydonal Ospina[1], André Leite[1], Cristiano Ferraz[1], André Magalhães[2], Víctor Leiva[3,*]

[1]Department of Statistics, CASTLab, Universidade Federal de Pernambuco, 51280-000 Recife, Brazil
[2]Department of Economics, Universidade Federal de Pernambuco, 51280-000 Recife, Brazil
[3]School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, 2340000 Valparaíso, Chile

***Correspondence**
victor.leiva@pucv.cl;
victorleivasanchez@gmail.com
(Víctor Leiva)

Abstract

The COVID-19 pandemic is one of the worst public health crises in Brazil and the world that has ever been faced. One of the main challenges that the healthcare systems have when decision-making is that the protocols tested in other epidemics do not guarantee success in controlling the spread of COVID-19, given its complexity. In this context, an effective response to guide the competent authorities in adopting public policies to fight COVID-19 depends on thoughtful analysis and effective data visualization, ideally based on different data sources. In this paper, we discuss and provide tools that can be helpful using data analytics to respond to the COVID-19 outbreak in Recife, Brazil. We use exploratory data analysis and inferential study to determine the trend changes in COVID-19 cases and their effective or instantaneous reproduction numbers. According to the data obtained of confirmed COVID-19 cases disaggregated at a regional level in this zone, we note a heterogeneous spread in most megaregions in Recife, Brazil. When incorporating quarantines decreed, effectiveness is detected in the regions. Our results indicate that the measures have effectively curbed the spread of the disease in Recife, Brazil. However, other factors can cause the effective reproduction number to not be within the expected ranges, which must be further studied.

Keywords

Basic and effective reproduction numbers; Data science; Data visualization; Growth model; SARS-CoV-2; Smart analytics; Time-series models

1. Introduction

The epidemic of COVID-19, caused by the severe acute respiratory syndrome - coronavirus 2 (SARS-CoV-2), has spread rapidly throughout the world. Since February 2020, cases have been reported in more than 26 countries, making it one of the biggest public health crises in the world [1–3]. In Brazil, the first case of the disease was confirmed in the city of São Paulo on 26 February 2020. Since then, the epidemic has spread across the country, forcing Brazilian states to adopt measures of social distancing to contain the virus. Currently, Brazil has accumulated more than 88 thousand deaths by COVID-19, with over 2000 deaths in Recife, the capital of the northeastern Brazilian state of Pernambuco. However, the growth rate of the epidemic still does not show clear signs of diminishing, and the disease continues to expand in the country, albeit at a slower pace [4]. Looking at cumulative curves of deaths attributed to COVID-19 by 01 November 2020, all Brazilian states had progressed to the late growth regime, that is, they were either in the transition to saturation or in the saturation stages (epidemic control), where the cumulative number of cases or deaths tends to a leveling plateau (slowdown in their death curves) [5]. This situation is presented in diverse Latin American countries [3, 6].

In the context of the response to the COVID-19 pandemic, note that it is essential to build strategies based on open and shared knowledge so that the circulation of information is faster. Thus, the iteration of emerging research related to this new disease allows for greater integration of multiple data sources to map and anticipate the spread of COVID-19. Publishing open resources for communication to the public as well as continuing education and widespread dissemination of expertise is also essential. This facilitates the continuity of services and economic activity, which have been strongly affected [7–9], especially when many people are in quarantine or precarious conditions, as imposed by the COVID-19 pandemic.

The complexity and changes of an epidemic or outbreak, such as COVID-19, imply a dynamic response in public health policies, control protocols, and data analysis. Overall, we can identify four stages of the outbreak response. First, the detection stage, which starts with the first case and ends with the first intervention activities (for example, patient isolation, contact tracing, and vaccination), involves surveillance systems [6] and mainly qualitative risk assessment measures. Second, the initial response is the part of the intervention during which

the first simple analysis can take place, essentially centered around the dynamics of infection transmission. Third, this also merges into the intervention stage, where more complex analyses may be needed (for example, vaccination strategies, intermittent lockdown, palliative drugs) [10, 11], which ends once the last reported case recovers or dies. At last, and fourth, post-intervention is the stage where lessons learned may help to improve the protocols and preparedness for the next epidemic.

An essential feature of the epidemic response is the growing focus on exploiting all available data to monitor the stages of this response and provide rapid reaction, enabling evidence-based decision-making [3, 6, 12, 13]. Using data to improve situational awareness is complex because it involves various related functions and techniques, from collecting the data at service points to generating informative situational reports as dashboards, tweets, sensors, etc. [3, 10, 14]. In the case of Recife, Brazil, we can mention the collaborative efforts of:

• Data and Analysis for Decisions and Operations (DADO), whose website is available at dado.recife.br (accessed on 10 August 2021),

• Cooperative Research Network on Modelling the COVID-19 Epidemic and Non-Pharmacological Interventions (MODINTERV), whose website is available at fisica.ufpr.br/redecovid19/index.html and fisica.ufpr.br/modinterv (accessed on 10 August 2021), and

• Computational Agricultural Statistics Laboratory (CAST-Lab), whose website is available at castlab.org (accessed on 10 August 2021).

These organizations provide different indexes, metrics, and data visualizations to inform the population about the stage of the epidemic.

The objective of our investigation is to discuss and provide tools that can be helpful from the perspective of data analytics to respond to the COVID-19 outbreak in Recife, Brazil. The rest of the paper is organized as follows. Section 2 presents the methods used in this research as well as details of the data employed. Section 3 reports the results obtained, while Section 4 concludes and discusses the main findings of the study, as well as challenges and opportunities, limitations, and ideas for future research.

2. Materials and methods

2.1 Materials

For smart analytics, it is needed to employ and optimize various freely accessible data sources to update and select databases [15, 16]. Initial data about the care of COVID-19 suspected patients, or patients who presented moderate symptoms of the disease, are recorded in the Family Health Units. These units consist of a public health network (Brazilian Unique Health System) composed of private clinics, ambulatory centers, and public hospitals located in different neighborhoods of the Brazilian cities, such as Recife. Also, healthcare systems for patients who use supplemental services (Brazilian Supplementary Health) provide data about severe disease symptoms. Another way of tracking and recording data are the call centers created to answer questions about the symptoms and guide the population in their care. Therefore, the Brazilian Ministry of Health implemented the Research Electronic Data Capture (REDCap), e-SUS Epidemiological Surveillance (e-SUS-VE), and Influenza Epidemiological Surveillance Information System (SIVEP-Gripe) platforms to report prospective suspected, probable, and confirmed COVID-19 cases as informed by public and private health services (primary and emergency care). The above-mentioned systems are interrelated on the COVID-19 website (infoms.saude.gov.br/extensions/covid-19_html/covid-19_html.html, accessed on 10 August 2021), which summarizes daily the aggregated counts from both platforms. In the case of Recife, Brazil, governmental and official data sources come from:

• The Center for Strategic Information on Health Surveillance, whose website is available at cievsrecife.wordpress.com (accessed on 10 August 2021).

• The Secretariat of Planning and Management from the Pernambuco State, whose website is available at www.seplag.ce.gov.br (accessed on 10 August 2021).

The Brazilian government and public data sources are available at:

• brasil.io/dataset/covid19/caso (accessed on 10 August 2021).

• dados.gov.br/dataset (accessed on 10 August 2021).

• github.com/wcota/covid19br (accessed on 10 August 2021).

The mobility indexes of the Google Community Mobility Reports (GCMR) are available at www.google.com/covid19/mobility (accessed on 10 August 2021). The GCMR collates data from those accessing Google applications with smartphones or handheld devices, allowing us to record the location history or traceability. In addition, data based on scientific articles reviewed by peers that report primary data as the gold standard for data inclusion provide further information cited in the references of this article.

To find additional details for each case or patient, the data can be augmented with medical reports, registry data, and captured reports (web scraping), primarily through news websites. The primary databases contain epidemiological records used for surveillance, evaluation, and research to address public questions. The administrative records (such as outpatient and hospital data anonymized and privatized) are employed for accounting and controlling the production of the services provided. The medical records contain privatized and anonymized clinical data on patients (in this work, these data were not used because they would have to be approved by the respective ethics committees before they could be used for any kind of implementation). The databases of the Brazilian Ministry of Health have demographic and socio-economic characteristics such as conditional cash transfer and housing programs. Combining these databases is employed to study social determinants of health and evaluate policies for epidemic control. Among the recorded variables to be used, we can cite: (i) daily cases reported; (ii) daily deaths reported; (iii) age; (iv) sex; (v) complete blood analyses (which is a type of blood test) on patients tested and other coronavirus or influenza pathogen diagnoses; (vi) isolation index of confirmed and suspected cases; (vii) number of hospitalizations; (viii) number of hospital beds; (ix) population density for the year 2020 by

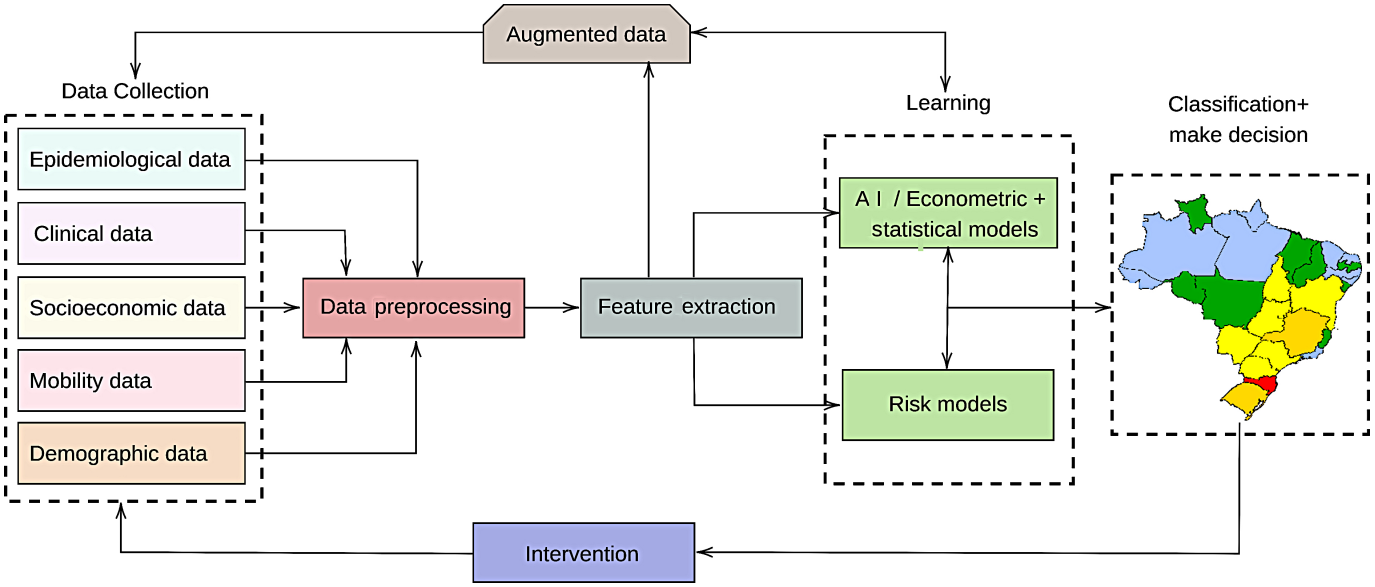

FIGURE 1. Simplified flow diagram on smart analytics for Brazilian COVID-19 data. The process involves feedback of augmented data by using data cleaning, data transformation, and feature extraction by applying learning algorithms that can be employed to prioritize the risks factors according to the quantified judgment of a decision-maker.

age groups and geographical mesoregions; (x) age distribution of confirmed COVID-19 infection rates; (xi) specific rates by age group; and (xii) intensive care unit (ICU) admissions for COVID-19 cases; among others.

The CASTLab, DADO, and MODINTERV platforms integrate several interconnected modules, including tools for data collection, data preprocessing, feature extraction, learning, classifying, making decisions, and data augmentation. These tools and their inter-dependencies are summarized in a simplified workflow sketched in Fig. 1, representing a framework for producing insights using smart analytics to combat outbreaks (SACO). From learning modules, it is possible, for example, to evaluate the epidemic curves via statistical methods, trends models, susceptible, exposed, infectious, recovered (SEIR) epidemic models, and Shewhart cohort chart, among others. Also, rates, averages, and another quantity are provided such that, if correlated with vaccination rates, hospital beds, population density, and under-reported cases, permit us to capture trends, produce epidemiological/clinical profiles, and identify possible risk factors for mortality of Brazilian COVID-19 patients under different aggregated levels, such as population segment, or region.

The classification of epidemic curves by including covariate data, such as vaccination rates, hospital beds, population density, and under-reported cases, implies a certain risk level (high or low). Thus, a decision-maker can choose to implement interventions to reduce the number of fatalities while softening the economic impact of the epidemic. Implementing epidemic curves by using purely statistical and mathematical techniques that evaluate quantities, such as rates and averages, to capture the trends is quite challenging. In this sense, SACO aligns with the emerging data science domain that employs scientific methods, processes, algorithms, and systems to extract helpful knowledge to respond to the COVID-19 epidemic. The SACO intersects health, economy, public policy knowledge, planning, epidemiology, methodological development, and information technologies to collect, select, analyze, visualize, model, simulate, optimize, and report outbreak data of an epidemic in a smart and interconnected way.

The SACO used in Brazil is based on a wide range of approaches including, among others: (i) gathering data, database design and mobile technology [17]; (ii) statistical estimates with the maximum likelihood method [18, 19]; (iii) interactive data visualization [20–22]; (iv) geostatistics [23, 24]; (v) graph theory [25–27]; (vi) Bayesian statistics [28, 29]; (vii) mathematical and computational modeling [30–33]; (viii) genetic analyses [34, 35]; (ix) evidence synthesis methods [36–39]; and (x) statistical quality control charts adapted to monitor COVID-19 deaths and other indexes [40, 41].

After the initial entry of records in the platform, the database should undergo a consistency check and analysis utilizing complementary methodologies such as data anonymization, imputation, and data linkage to identify possible duplicate records and missing values effects. Using the number of confirmed COVID-19 cases and deaths reported up to 10 August 2021, we apply data-driven innovative epidemiology tools. Among them, we can mention the manual verification performed by data curators or the use of statistical and artificial intelligence tools [42], such as probabilistic record linkage, feature hashing, and principal component analysis [43, 44], among others [45, 46]. The epidemiological outbreak of COVID-19 can be investigated with the verified data, including descriptive mapping of occurrences over time and estimation of the main epidemiological parameters employing mathematical models. Specifically, we may gain intelligent insights from epidemic curves that represent the number of new cases or deaths per unit of time, based on the date of symptom onset [47].

We put here only some of the smart tools and statistical methods used by the CASTLab, DADO, and MODINTERV platforms as illustration. We utilize the moving average model to estimate the non-stationarity in the number of confirmed

COVID-19 cases and deaths to identify the moments when there was a significant change in the time-series trend. Note the importance of determining the rhythm of contagion after a policy intervention (such as social distancing, quarantine, closing schools). Also, we use georeferenced standardized data on specific settlements or areas, typically obtained from administrative units such as neighborhoods or census sectors. We utilize choropleth charts as a thematic map in which areas are colored proportionally to a statistical variable that indicates an aggregate summary of a geographic characteristic within each area. These geographical records help to assess the spread of contagion in a macro view and make massive contact tracing between regions easier to monitor. Thus, when implementing successful protocols for containing the epidemic, such as social isolation and containment barriers (for example, vehicles and individuals' flux where are not allowed circulation between different geographical locations), does the handling of the pandemic to be more efficient, that is, the number of deaths and contagions decreases.

The quantification of transmissibility during epidemics is essential to design and adjust public health responses. Epidemics can be measured by the effective or instantaneous reproduction number, that is, the average number of secondary cases caused by an infected individual. Knowledge of the reproduction number is key to understanding the dynamics of any infectious disease, and these should be reevaluated as the epidemic progresses in space and time. The instantaneous reproduction number, which considers the speed with which a disease spreads in a population, is an epidemiological parameter that quantifies the average of contagions that an infected person causes, assuming that disease transmissibility is constant in a window of time [6]. Thus, an effective quarantine restricts the free spread of the disease and maintains this parameter at values less than one to achieve a controlled spread [6]. With the implementation of COVID-19 control measures, the disease transmissibility is likely to change as well. Based on the surveillance data of COVID-19, we calculate the time-varying reproduction number, instantaneous reproduction number, and peak date across basic reproduction number, R_0, namely, after interventions. This is conducted to evaluate its changing dynamics and determine the effectiveness of possible intervention strategies as vaccination policies. Note that R_0 is the expected number of cases generated by one case in a population where all individuals are susceptible to infection.

The epidemic curves are helpful in many ways. They provide a simple visual sketch of demographic dynamics that may be used to assess the growth or decline of an outbreak [48] and evaluate the effect of some measures, such as non-pharmacological interventions. In addition, epidemic curves often form the raw material employed by various modeling techniques for monitoring and forecasting [49–51].

2.2 Statistical methods

Time-series model

We estimate the trend of the number of COVID-19 cases per day in Recife, Brazil, with a time-series model as follows. Let $y(t)$ be the number of confirmed cases at time t. Then, the trend of this series, $m(t)$ say, can be estimated through the moving average on day t given by

$$\hat{\mathrm{m}}(t) = \frac{1}{2d+1} \sum_{i=-d}^{d} y(t+i), \tag{1}$$

with d defined in Eqn. 1 being a constant that regulates the width of the data interval through which the average is calculated, controlling the degree of smoothing. Observe that, as d decreases, the moving average is more sensitive, so it may capture recent trends very well, although it might produce false alerts. Nevertheless, when d is large, these false alerts are avoided, but the identification of current trends is more transient [6]. The choice of d can be made according to the seasonality of the minor scale detected and considering that the confinement measures for the observation period were the weekends.

Epidemic model

When intervention measures (such as lockdown and social displacement) are put in place and a certain proportion of the population gains immunity, interest switches to knowing the time-varying effective reproduction number. Consider an individual who turns infectious on day t. We denote by $R_e(t)$ the expected number of secondary cases that this infectious individual causes. We assume that the times of infecting others and detecting this infectivity (for example, by symptoms or by a test) coincides, that is, on day t, this person also appears in the incidence time-series. The time interval between symptom onset in the primary case and symptom onset in the secondary case is called the serial interval (SI), whose associated statistical model is called the SI distribution. Note that the SI is different from the generation time (GT), which is the period between exposure of the primary and secondary cases. However, since exposure is rarely observable, we must employ the time series of incident symptom onsets as a basis for the statistical inference [52]. Using a renewable equation, we can estimate the time-varying effective reproduction number [53]. Here, we consider a simple growth model and denote by $y(t)$ the expected number of new symptom onsets that we observe on day t.

Let $(g_1, ..., g_M)$ be the probability mass functions of the SI distribution, that is, $P(i) = g_i$, for $i \in \{1, ..., M\}$. The expected number of cases can be described by the homogeneous linear difference equation stated as

$$y(t) = R_e(t-1)g_1 y(t-1) + \cdots + R_e(t-M)g_M y(t-M) = \sum_{i=1}^{M} R_e(t-i)g_i y(t-i), \tag{2}$$

where $t \in \{2, 3, \ldots\}$, and we ignore the terms when $t - M \leq 0$.

We can estimate $R_e(t)$ defined in Eqn. 2 by the method proposed by Wallinga and Teunis (W&T) [54–56]. The probability that case i infected by case j, p_{ij} namely, given its difference in time of symptom onset at $(t_i - t_j)$, may be expressed in terms of the SI distribution, denoted by ω, which is formulated as

$$p_{ij} = \frac{\omega(t_i - t_j)}{\sum_{i \neq k} \omega(t_i - t_j)}. \tag{3}$$

Thus, p_{ij} stated in Eqn. 3 is the probability that case i has been infected by case j, normalized by the probability that case i has been infected by any other case k. The effective reproduction number for case j is the sum of all cases i, weighted by the relative probability that case i has been infected by case j, that is, $R_j = \sum_i p_{ij}$. Therefore, $R_e(t)$ is obtained with the W&T method by averaging the individual reproduction number over all cases detected on the same day j, where $t_j = t$. The analysis can be carried out by utilizing the R0 package of the R software. Another way to evaluate $R_e(t)$ is to use the Robert Koch Institute (RKI) method, where a constant generation time of four days is assumed [57]. The RKI method estimates the time from infection to: (i) first symptoms to be around five days; (ii) being infectious around three days (which results in two days during which an affected person is infectious but might not suspect he/she is sick him/herself); and (iii) infecting other people around four days.

The instantaneous reproduction number $R(t)$ is interpreted as the average number of secondary cases that each symptomatic individual at time t would infect if the conditions remained as they were at time t. Note that $R(t)$ is evaluated by comparing the number of new infections on day t with the infection pressure from the days prior to t, that is, by the expression established as

$$R(t) = \frac{y(t)}{\sum_{s=1}^{t} g_s y(t-s)}. \tag{4}$$

Note that $R(t)$ given in Eqn. 4 is different from the basic reproduction number, which is presented next. Unlike the W&T forward-looking method, the estimate defined in Eqn. 4 is backward in time, whereas the timing of $R(t)$ can make a big difference when comparing it with intervention measures. The estimator of the instantaneous reproduction number is implemented in an R package named EpiEstim. If $R(t) > 1$, it means that a case infects more than one person, resulting in a spread of the virus, whereas $R(t) < 1$ translates into the containment of the virus, as one case infects less than one person.

We use a deterministic SEIR model (one-wave) to describe the spread of the disease because of the uncertainty due to the incomplete identification of the infected population. Note that, at the beginning of the pandemic, there was no clear policy on testing the population for SARS-CoV-2, and there were no test kits that would allow the correct identification of the actual infected cases. As mentioned, the SEIR model classifies the population into four classes: susceptible (S), exposed (E, infected but not yet infectious), infected (and infectious, I), and recovered (R). The dynamics of the transitions between the four different compartments (S, E, I, and R) is described by the non-linear system of ordinary differential equations (ODEs) given by

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta SI}{N}, \\ \frac{dE}{dt} &= \frac{\beta SI}{N} - \sigma E, \\ \frac{dI}{dt} &= \sigma E - (\gamma + \mu) I, \\ \frac{dR}{dt} &= \gamma I, \end{aligned} \tag{5}$$

where the elements defined in Eqn. 5 correspond to $N = S + E + I + R$, σ is the loss of latency rate, γ is the recovery rate, μ is the disease-induced mortality rate, and β is the transmission rate. We assume that those non-pharmacological interventions (for example, the wearing of face masks, social distancing, or self-isolation when sick) can decrease the number of contacts per individual per unit of time during the epidemic. To accommodate such effect, we consider that β change in time t by the dynamic formulated as

$$\frac{d\beta}{dt} = \frac{1}{\tau}\left(\beta_0\beta_1 - \beta(t)\right), t \geq t_0, \tag{6}$$

where t_0 is the starting time of the intervention (initial day) and τ is the average duration of the interventions. Furthermore, β_0 is the initial transmission rate and the product $\beta_0\beta_1$ represents the transmission rate at the end of the epidemic. The explicit solution to the equation given in Eqn. 6, subject to the constraint to $\beta(t_0) = \beta_0$, is stated as

$$\beta(t) = \begin{cases} \beta_0, t < t_0; \\ \beta_0\left(\beta_1 + (1-\beta_1)\, e^{\frac{-(t-t_0)}{\tau}}\right), t \geq t_0; \end{cases} \tag{7}$$

where the elements defined in Eqn. 7 are established in the previous expressions. Once the SEIR model without the time-dependent transmission rate is fitted, the basic reproduction number is estimated [58, 59] as $R_0 = r/(\gamma + \mu)$. Here, r is the exponential rate of increase of cases at the beginning of the epidemic, which, in turn, is estimated by fitting an exponential function employing only the first data. The estimation of parameters is based on a nonlinear curve fitted by minimization of the sum of square errors (by utilizing the least squares method). To solve the ODEs given in Eqn. 5, we use the ode function of an R package named dSolve. This package has implemented stiff and nonstiff integration routines utilizing the ODE-PACK Fortran codes (LSODE, LSODES, LSODA, LSODAR), DVODE, and DASPK2.0. A suite of Runge-Kutta integrators and special-purpose solvers to efficiently integrate 1-, 2- and 3-dimensional partial differential equations are available. The dSolve package includes fixed and adaptive time-step explicit Runge-Kutta solvers and the Euler method. We employ the optim function of the R software that has implemented the Nelder-Mead, quasi-Newton and conjugate-gradient algorithms.

We use curves for the cumulative number of COVID-19 deaths as a function of time. Note that they can show three distinct regions of behavior in the first wave of epidemics. There is an initial phase in which the epidemic curve is typically ex-

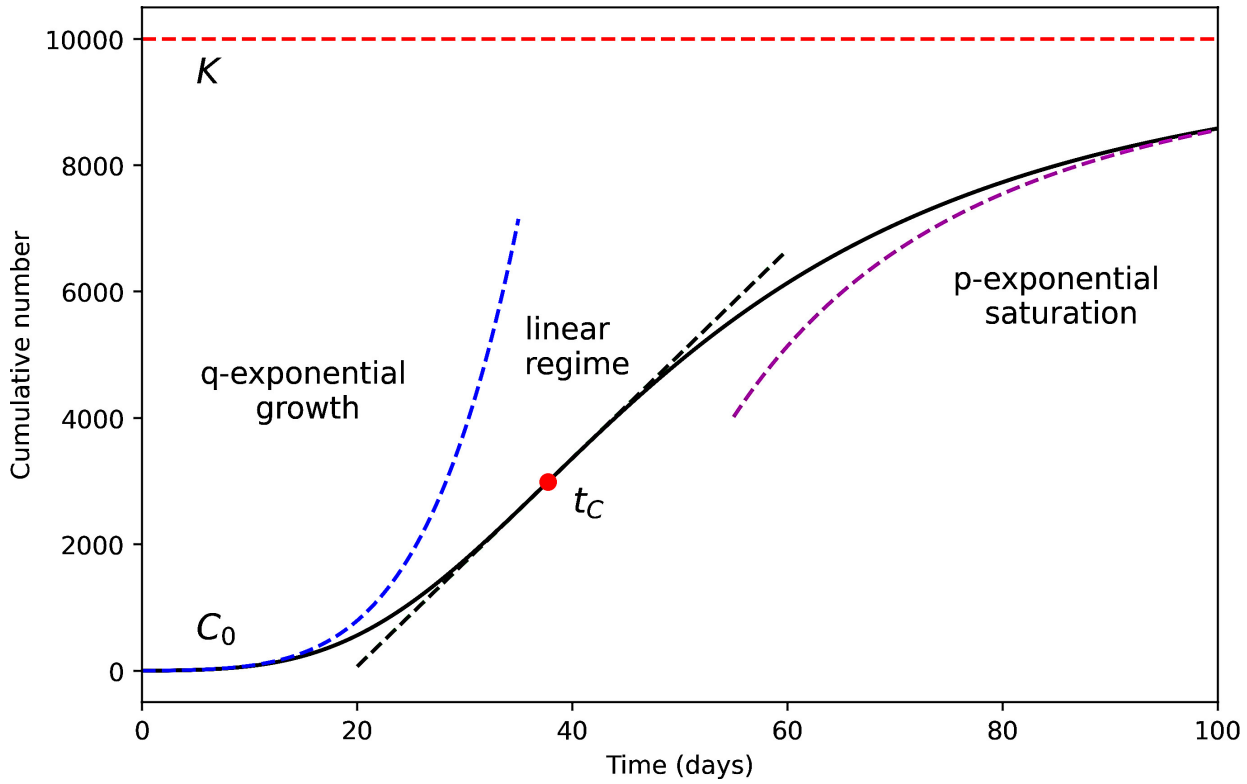

FIGURE 2. Qualitative outline of an epidemic curve (black) for the accumulated number of deaths, with an indication of their different phases in the first wave of the pandemic, with the red circle indicating the change point of the curve.

ponential, based on the number of deaths observed on the first day that a death occurred. This rapid rise is characterized by the growth rate r. Note that contagion may result in a complex network of human contacts or when mitigation measures are adopted at the beginning of the epidemic. However, depending on the complexity of the contagion dynamic, the growth may happen more slowly than an exponential curve characterized by a parameter $q \in (0, 1)$, which interpolates between the linear regime ($q = 0$) and the exponential regime ($q = 1$). This type of epidemic curve presents an inflection point, denoted by t_c and shown in Fig. 2, which corresponds to the instant in which the accumulated curve changes its concavity, that is, the growth rate reaches its maximum value and begins to decrease after that. The final part of the curve, after the inflection point, can be characterized by an α parameter that controls how quickly the epidemic curve moves away from the linear trend and bends towards the plateau. The plateau value, denoted by the parameter K, represents the total number of deaths at the end of the epidemic.

We may model the epidemic curves in a deterministic way through a generalized Richards model (GRM) [52, 53]. Here, the GRM is defined by the ODE formulated as

$$\frac{dC}{dt} = r[C(t)]^q \left(1 - \left(\frac{C(t)}{K}\right)^{\alpha}\right), \tag{8}$$

where $C(t)$ is the cumulative number of deaths at t, r is the growth rate in the initial phase, as mentioned, $0 \leq q \leq 1$ is the parameter that enables us to interpolate among linear ($q = 0$), sub-exponential ($q < 1$), and exponential ($q = 1$) growth. The ODE stated in Eqn. 8 must be supplemented with an initial condition given as

$$C(0) = C_0, \tag{9}$$

for a given value of C_0. The exact solution of the expression established in Eqn. 8 for $0 \leq q < 1$, subject to the constraint shown in Eqn. 9, can be written implicitly as

$$t = \frac{1}{r(1-q)} C^{1-q} \;\; {}_2F_1\left(1, \frac{1-q}{\alpha}; 1 + \frac{1-q}{\alpha}; \frac{C^{\alpha}}{K^{\alpha}}\right) - t_i, \tag{10}$$

where ${}_2F_1(a, b; c; x)$ is the hypergeometric Gauss function. The constant t_i given in Eqn. 10 is determined by the initial condition C_0 through the relation defined as

$$t_i = \frac{C_0^{1-q}}{r(1-q)} \;\; {}_2F_1\left(1, \frac{1-q}{\alpha}; 1 + \frac{1-q}{\alpha}; \frac{C_0^{\alpha}}{K^{\alpha}}\right). \tag{11}$$

The inflection point, t_c namely, of the curve $C(t)$ is defined as the time when the second derivative of $C(t)$ with respect to t is zero, that is, $\ddot{C}(t_c) = 0$, where the two points on the top of $C(t)$ indicate the second derivative with respect to the time. Using the expression formulated in Eqn. 11, we find that

$$t_c = \frac{K^{1-q}}{r(1-q)} \left(\frac{q}{q+\alpha}\right)^{(1-q)/\alpha} \;\; {}_2F_1\left(1, \frac{1-q}{\alpha}; 1 + \frac{1-q}{\alpha}; \frac{q}{q+\alpha}\right) - t_i. \tag{12}$$

where the elements defined in Eqn. 12 are established in the previous expressions.

The solution of the GRM given in Eqn. 8 is attained in an implicit form. It does not represent any numerical difficulty since the solution can easily be obtained, for empirical curve

fitting purposes, by simply treating the data in the same implicit form, that is, as a curve of type $C(t)$. In the case where $q = 1$, the GRM defined in Eqn. 8 reduces to the standard Richards model (RM), which has an explicit formula in terms of elementary functions stated as

$$C(t; r, \alpha, K, t_c) = \frac{K}{\left(1 + \alpha e^{-\alpha r (t - t_c)}\right)^{1/\alpha}}. \quad (13)$$

For fitting the formula defined in Eqn. 13 to empirical data, it is convenient to set $C(0) = C_0$, where C_0 is the number of deaths recorded on the first day that a death was reported. Using that $C(0) = K/\left(1 + \alpha e^{-\alpha r t_c}\right)^{1/\alpha}$, we can eliminate t_c in favor of the other parameters, so we work with only three free parameters, namely r, α, and K, to be numerically determined. All statistical fittings may be performed using the Levenberg-Marquardt algorithm to solve the corresponding non-linear least-square optimization problem.

When estimating the model parameters from short time-series, we have overfitting problems due to the redundancy of parameters. This may lead, for example, to the estimation of specific parameters that are outside of biologically reasonable ranges or of other types. For instance, when applied to the number of infected cases in an epidemic, the parameter α should be constrained to the interval $(0, 1)$. We use the GRM instead of the number of deaths, but we assume that the same constraint should be considered. For our purposes, we need to consider the constraints $0 < r < 1$ and $0 < \alpha < 1$ as empirical criteria for validating the GRM, which is unsuitable when the available data do not encompass the inflection point t_c. Nonetheless, as more data are considered, the model is expected to become more accurate. As an empirical criterion, we consider here that the GRM is only acceptable if t_c is less than the time of the last data point. Hence, an intervention strategy (for example, adding covariate data, such as vaccination rates, hospital beds, population density, under-reported cases) may be used to model the data by assuming that its next result is captured by a change in the values of the parameters of the GRM after a given time.

Control charts

Looking at monitoring the number of deaths, adapted Shewhart control charts can be strategically utilized to differentiate variation due to the disease effect and variation due to common causes. Adaptation of these control charts is needed given the exponential behavior of indicators such as COVID-19 death rates. The use of a control chart for monitoring COVID-19 was introduced in [40, 41].

In the context of an epidemic, such as COVID-19, it is possible to analyze the evolution of indicators in a manner analogous to that of a production process. The idea behind the method assumes that observations are made periodically. This is equivalent to having sample sizes $n_t = 1$ analyzed utilizing $Q(t)$ or some statistic indicator as the mean, standard deviation, or percentage. Its choice depends on the characteristic measured in each item and is calculated to summarize the values measured in the sample, $S(t)$ namely, representing the behavior of the process at time t. These statistics are compared to the monitoring limits such as central line (CL) as well as lower (LCL) and upper (UCL) control limits plotted over time. In the case of the indicator of COVID-19 deaths per day, the signal is not constant, and, in general, its growth is exponential. Then, it is possible to model the signal and what must be checked, in terms of stability, is the behavior of the residuals of the fitted model. If the growth is modeled by an exponential function described as

$$Q(t) = y_t = ae^{bt}, \quad (14)$$

where $y(t)$ stated in Eqn. 14 is the number of deaths registered on day $t \in \{1, ..., M\}$, a is the initial value of the growth model, and b is the growth factor. Here, a and b are constants to be estimated. Hence, $log(y(t))$ is a linear function on t formulated as

$$\log(y(t)) = \log(a) + bt. \quad (15)$$

The procedure for adapting the Shewhart chart to our context of COVID-19 is described as follows in Algorithm 1.

Algorithm 1: Adapted Shewhart control chart

Step 1. Fit $log(y(t))$ defined in Eqn. 15 as a function of time t via the linear model stated as

$$\log(y(t)) = \beta_0 + \beta_1 t + \varepsilon_t, \quad (16)$$

where β_0, β_1 are regression coefficients, and ϵ_t is a random error of zero expected value and constant variance.

Step 2. Calculate the residuals of the model stated in Eqn. 16 as

$$\hat{\varepsilon}_t = \log(y(t)) - \hat{\log}(y(t)). \quad (17)$$

Step 3. Obtain the limit of the Shewhart chart drawn up for the residuals defined in Eqn. 17 as

$$\mathrm{L}_{\text{residual}} = \hat{\varepsilon} + 3\hat{\sigma}_{Q\,(\text{individual})}, \quad (18)$$

where $\hat{\sigma}_Q$ is a measure of the variance of the statistic $Q(t)$.

Step 4. Calculate the LCL and UCL of the Shewhart adding and subtract $L_{residual}$ stated in Eqn. 18 to the values fitted by the regression considering

$$\begin{aligned} \mathrm{LCL}_{\text{regression}} &= \hat{\log}(y(t)) - \mathrm{L}_{\text{residual}}; \\ \mathrm{UCL}_{\text{regression}} &= \hat{\log}(y(t)) + \mathrm{L}_{\text{residual}}. \end{aligned} \quad (19)$$

Step 5. The CL and the respective control limits stated in Eqn. 19 of the fitted Shewhart chart are calculated by exponentiating the fitted values and the limit calculated in step 4.

The implementation of Algorithm 1 must be used the logarithm in base 10. In addition, it is possible that, in some situations, null values are observed, that is, $y(t) = 0$, for some $t \in \{1, ..., M\}$. Therefore, for implementing the adapted Shewhart control chart, $(1 + y(t))$ must be employed as the response variable.

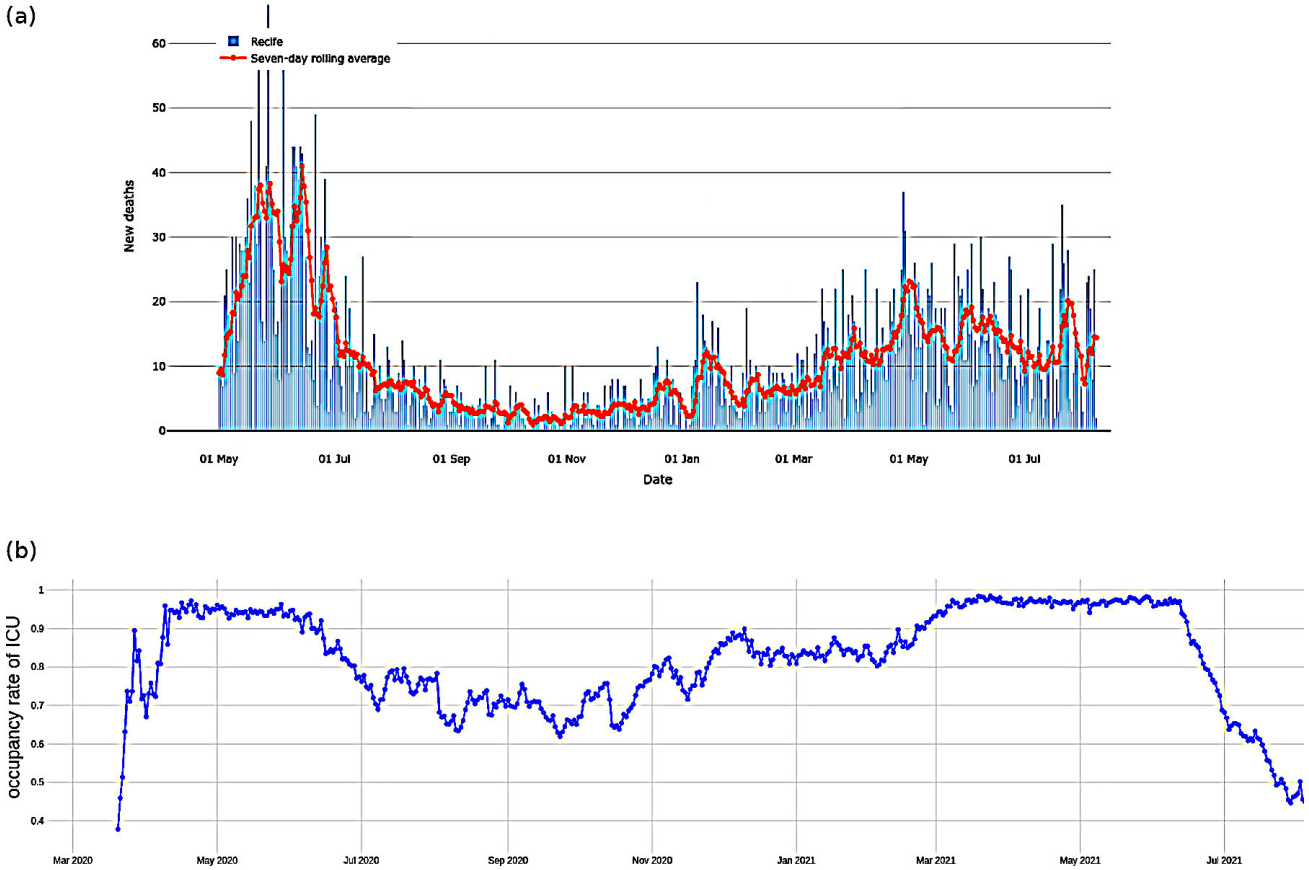

FIGURE 3. Epidemic curves for Recife, Brazil, with (a) temporal evolution of the recorded daily new COVID-19 deaths during epidemic up to 23 July 2021; and (b) temporal evolution of the occupancy rate of ICU beds to evaluate current capacity for clinical hospitalization up to 23 July 2021.

3. Results of smart analytics of the COVID-19 epidemic in Recife, Brazil

The number of deaths and new daily reported cases captured the trend in the data; see Fig. 3(a). We observe that, in the first days of the epidemic, the growth curve is aligned with an average increasing pattern of about 10% per day, which approximately doubles the number of deaths each week. However, during May 2020, the growth vastly exceeded this level, increasing in several ways at a highly alarming pace. This occurs especially about the installed capacity of hospital beds monitored via the occupancy rate of beds, as shown in Fig. 3(b), intended solely for the care of patients with COVID-19 (identified by positive polymerase chain reaction -PCR- or positive serological tests). At the end of May 2020, this growth slowed down and started to grow well below 10% per day.

The geographical distribution of COVID-19 cases and deaths was analyzed via density maps. In Fig. 4(a), we visualize the number of COVID-19 deaths in Recife and mesoregions (spatial aggregate of intermediate level that includes municipalities and/or geographical administrative subdivision) up to 23 July 2021. Note that the highest incidence of deaths is in the Sertão region of Pernambuco and the rural areas of its metropolitan region. This can be explained by the precariousness of the healthcare center network that directly impacts the hospital capacity of Recife (urban area). Furthermore, Fig. 4(b) displays a choropleth map by neighborhoods of Recife, with the number of COVID-19 cases in the last 24 hours up to 23 July 2021. Observe that the highest incidence of new cases is in Boa Viagem zone (near the beachfront), where there is less control of access to the beach.

Fig. 5(a) and (b) show the temporal evolution of the effective reproduction number with intervention dates. We look at the curve over a short period at the beginning of the pandemic, when the first containment measures were implemented. Notice that the intervention measures were effective in reducing the spread of contagion. Also, Fig. 5(c) indicates the peak of the incidence curve after intervention measures (red dashed lines) and variations of the R_0 (calibration of the SEIR model) up to 08 June 2020. Note that delaying the intervention measures increases R_0, that is, the disease increased. As the model did not produce a good fit, then other indicators and models were implemented (this was an important insight into the inability of the SEIR model to follow the dynamics of the epidemic curves). In the case of Recife, it was possible to verify that the epidemic curve via the GRM was in the saturation phase until 30 July 2020, showing that the COVID-19 mitigation measures adopted in Recife had a significant effect on the control of the pandemic; see Fig. 6.

Control charts are strategic to differentiate between common and special causes of variations. As mentioned, they are usually composed of a CL, representing the mean behavior (trend) and LCL and UCL, as introduced in Section

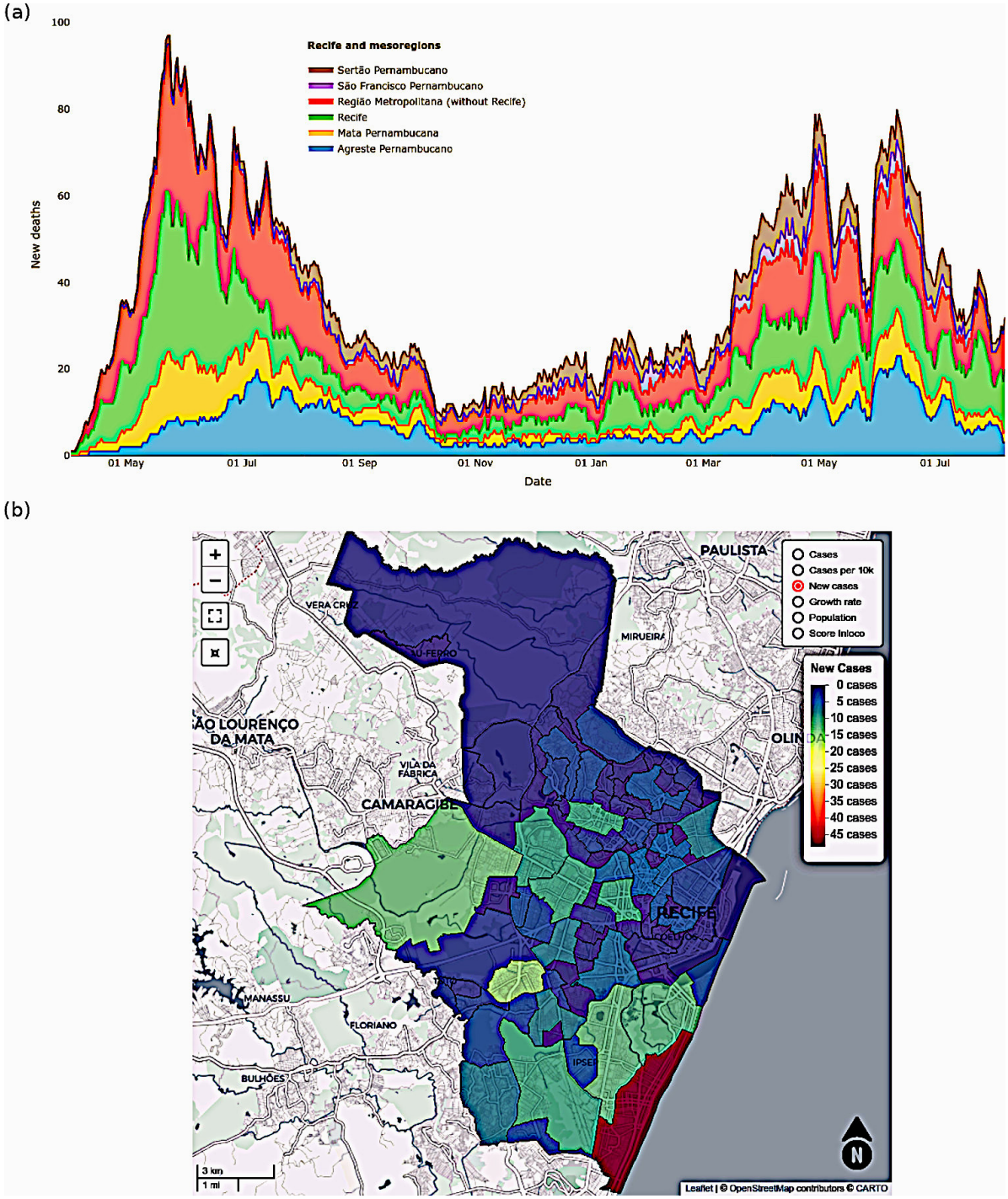

FIGURE 4. Graphical representation of indicators with geographical data, where (a) number of new deaths by mesoregions in Recife, Brazil, up to 23 July 2021; and (b) choropleth map by neighborhoods of Recife, with several cases in the last 24 hours attributed to COVID-19 up to 23 July 2021.

2.2. Observing indicators such as the number of daily deaths, oscillation occurs naturally. It is crucial to have a parameter to distinguish when this behavior simply reproduces common variation due to the phenomenon. A higher (or lower) point does not necessarily mean the phenomenon is changing its behavior, or this behavior is indeed indicating a change in trend. Such reference is provided by the LCL, CL, and UCL of the chart. If the process is stable, oscillation is expected randomly around the CL with almost every point between LCL and UCL. Shewhart control charts have been used successfully to make such type of monitoring for a long time. Rules often known in statistical control processes may be used to monitor the process. This monitoring of daily COVID-19 deaths in Recife employed a seven-point location rule as suggested in [40]. Phases are created each time the rule is implemented, that is, each time, a set of seven consecutive points fall either above or below the CL of the control chart. This means that, within each identified phase, it is reasonable to suppose the phenomenon (daily deaths) behaves consistently with the same parameters (same trend and variation).

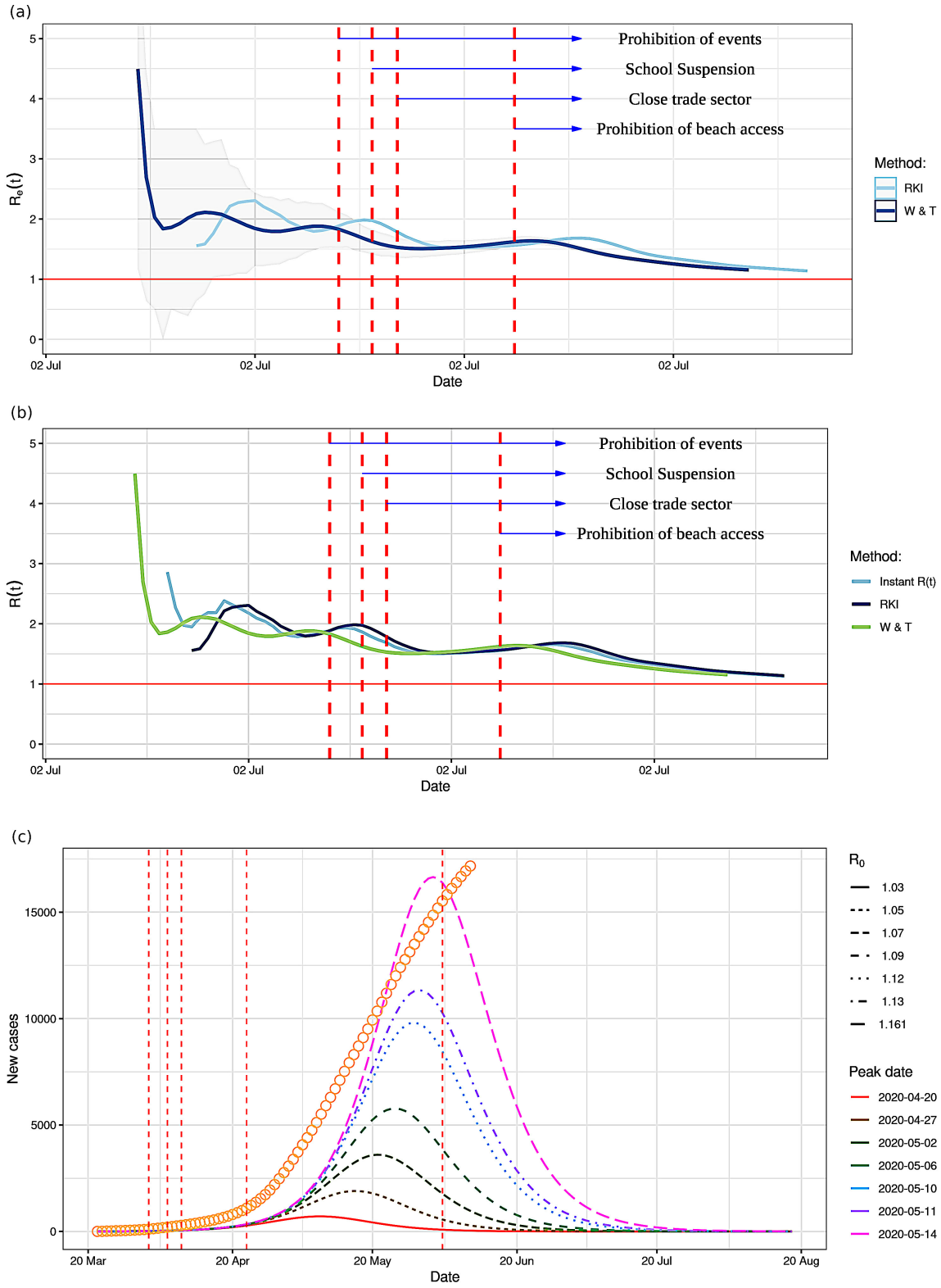

FIGURE 5. Epidemic curves for Recife, Brazil, with (a) temporal evolution of the effective reproduction rate and dates of intervention up to 30 July 2020; (b) temporal evolution of the effective reproduction number and dates of intervention up to 30 July 2020; and (c) incidence curve at peak date dependently of basic reproduction number after interventions measures (calibration of the SEIR model) up to 08 June 2020.

Fig. 7 illustrates an adapted Shewhart control chart for daily deaths on each day in Recife up to 10 August 2021. Each point (circle) represents a given number of deaths. Most of the points are in black. Red points are used to mark the beginning of each phase. Blue and green points represent Saturday and Sunday days, respectively. Although point colors were kept in the graph, they are enhancements, not main components of the method, considering they play a secondary role in the monitoring. The lines between points mainly show the growth rate or trends at regular time intervals (days in our case).

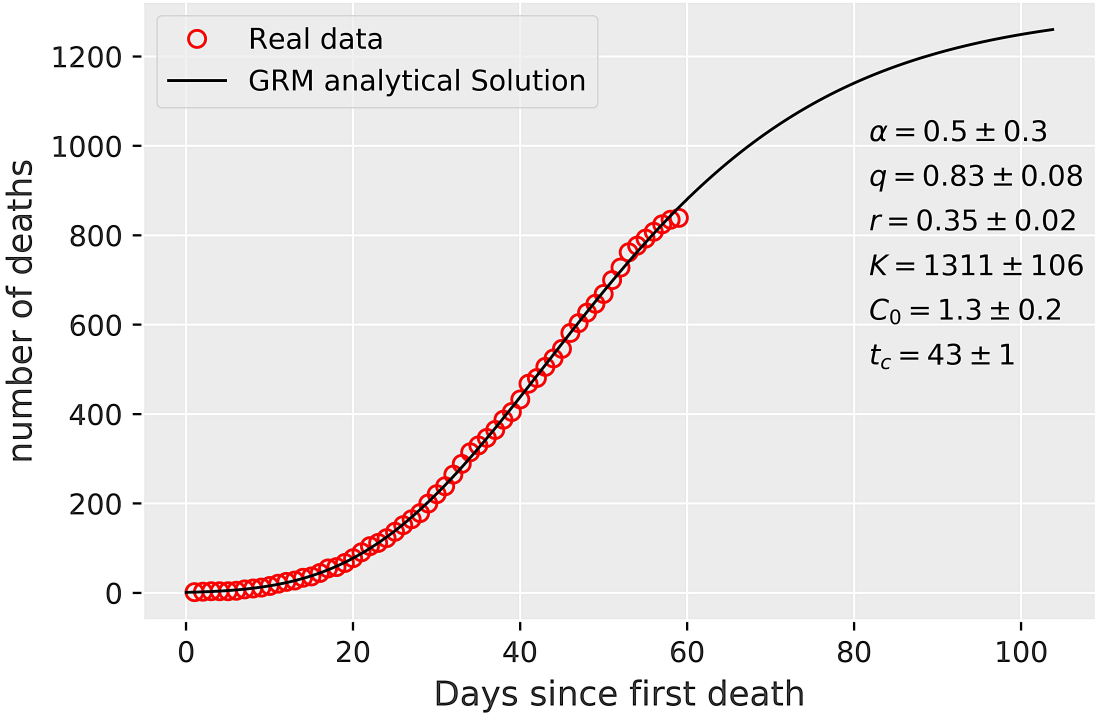

FIGURE 6. Qualitative outline of an epidemic curve (black) for the accumulated number of COVID-19 deaths, with an indication of their different phases in the first wave of the pandemic, with the red circle indicating the change point of the curve. The GRM fits the cumulative number of deaths by COVID-19 in Recife up to 30 July 2020, with the red circles representing the data and the black curve being the fitted model.

The chart shows points representing the observed number of deaths registered each day, along with the CL, indicating the trend and control limits, LCL and UCL. Phases were identified to monitor the situation and are shown in different colors. Looking at the last phase (in yellow), note that, during the period of 24 June 2021 to 21 August 2021, the number of daily deaths was still under a descendant trend, with parameters being consistent with the anterior phase (phase 10, between 24 April 2021 and 23 June 2021).

In addition to monitoring the number of COVID-19 deaths by adapted Shewhart charts, openly available data should be continuously selected and employed in modeling and simulation. Nevertheless, suppose the epidemic continues to grow. In that situation, public health agencies may tend to omit cases, only reporting the estimated number of COVID-19 confirmed or suspected cases, as it has occurred in previous large outbreaks, such as the H1N1 influenza pandemic [55].

The data available primarily on COVID-19 are insufficient to investigate the impact of assumptions on future projections related to the planning of public health policies. As detailed data become less and less available as the epidemic grows, we can have an augmented database structure that contains calibrated parameters of models utilized to train machine learning algorithms [42] and produce only reports of the total of new cases by location. Then, data/features augmentation, survey data via probabilistic sampling design, and deep learning can provide more insights into the outbreak.

4. Discussion, challenges and opportunities, conclusions, limitations, and future research

The city of Recife in Brazil is taking crucial steps to creating a culture of building scenarios for designing and evolving strategies for cities based on helpful tools of machine learning, artificial intelligence, and other modern data analysis [42]. In Recife, such technologies have been implemented very fast. For example, Shewhart control charts, machine learning models, and automatic classification are innovations of the existing protocols for outbreak control. The classification of epidemic curves by including covariate data, such as vaccination rates, hospital beds, population density, and underreported cases, implies a certain risk level (high or low). Then, implementing such curves by using purely statistical and mathematical techniques that evaluate quantities, such as rates and averages, to capture the trends is quite challenging.

In our analysis of Recife, Brazil, based on mesoregion data, it is possible to generate clusters and explore the evolution of the infection over time. We assessed the impacts on local productive arrangements and, via data analysis, that isolation barriers could have been implemented more effectively. Using the data from the neighborhoods, it was possible to understand which neighborhoods in Recife were more likely to be infected by some contaminated cases coming from other cities. Considering only the neighborhoods of Recife, it is

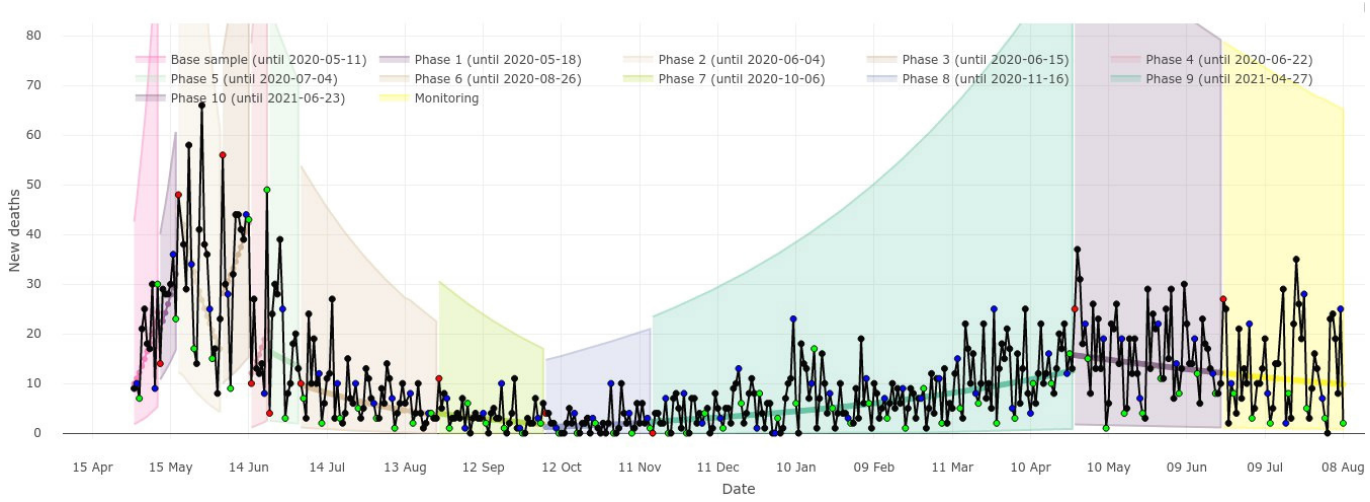

FIGURE 7. Adapted Shewhart chart for daily deaths in Recife, Brazil, with phases up to 10 August 2021. Red points mark the beginning of each phase. Blue and green points represent Saturdays and Sundays, respectively.

possible to understand associations between socioeconomic indicators, epidemic data, and risk measures for neighborhood prioritization. The results obtained via SACO make it possible to provide helpful information supporting disease monitoring and progressive economic recovery in the state and city studied as well as in other regions of Brazil and other countries.

We may employ machine learning algorithms that automatically classify epidemic curves after training on augmented data to solve this problem. These new tools will permit an automatic epidemic analysis and offer fresh opportunities to operationalize previously unexplored and rapidly growing data sources as big data [46, 60], as well as to synthesize data helping to understand epidemiological and risk patterns in the city. Also, this may help to develop quantitative evidence and decision-making in public policy for the health area. In this sense, the approaches analyzed in [61–63] may help to implement improved methodologies in the city for disaster control. Potential uses of machine learning and artificial intelligence will enable us to improve the diagnostic accuracy and tools for COVID-19 [42] (or any new disease or calamity), so obtaining more reliable prognosis, targeted treatments, and increasing the operational efficiency of health systems [64–68].

Potential future implementations of machine learning for SACO in Brazil include disruptive technologies, as image-based deep learning, which show clinical promise for fighting the pandemic. For example, deep learning-based algorithms improve the accuracy of pathology diagnosis compared to experienced physicians [42, 69, 70]. Also, natural language processing may be used as a tool to extract information from structured and unstructured text data embedded in electronic health records [71–75] to improve comprehension of the pandemic and to enhance decision-making systems.

We are beginning to understand the richness of opportunities offered by these smart tools. There is a growing concern in the academic community and in the public sector that results based on automatic analysis are not perceived in the same way as other interventions (for example, pharmacological) [74, 76]. These tools do not have clear guidelines for their development and rarely come under the same degree of scrutiny [77]. Several high-quality publications have demonstrated a lack of transparency, replicability, ethics, and effectiveness in reporting and evaluating predictive models based on machine learning and artificial intelligence.

This growing body of evidence suggests that while many recommendations in best practices for designing, conducting, analyzing, reporting, assessing, and implementing clinical tools can be borrowed from the traditional literature on economics, health systems, medical statistics, and public policies, they are not sufficient to guide the use of machine learning and artificial intelligence in research [78–83]. The production of such guidelines in Brazil is a significant undertaking given the growing battery of machine learning and artificial intelligence algorithms that have been developed, as well as the multifaceted nature of performance assessment of clinical, political, and social impact.

AUTHOR CONTRIBUTIONS

RO, AL, CF designed the study, collected the data, and created the models and codes. RO, AL, CF, AM performed the data analysis. RO and VL drafted and wrote the final version of the paper. RO and VL revised the article critically for important intellectual content. All authors approved the final version to be published.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

ACKNOWLEDGMENT

The authors thank MODINTERV, DADO, Porto Digital, and Recife City Hall, PE/Brazil. Also, the authors thank the three reviewers for their constructive comments on an earlier version of this paper.

FUNDING

This research was partially supported by the National Council for Scientific and Technological Development (CNPq) through the grant 305305/2019-0 (RO), and Comissão de Aperfeiçoamento de Pessoal do Nível Superior (CAPES), from the Brazilian government; and by FONDECYT, grant number 1200525 (V. Leiva), from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science, Technology, Knowledge, and Innovation.

CONFLICT OF INTEREST

The authors declare no conflict of interest. Víctor Leiva is serving as one of the Editorial Board members of this journal. We declare that Víctor Leiva had no involvement in the peer review of this article and has no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to ASA.

DATA AVAILABILITY

The data used to support the findings of this study are available upon request.

REFERENCES

[1] Johns Hopkins, C. S. S. E. Coronavirus COVID-19 global cases by the center for systems science and engineering (CSSE) at Johns Hopkins University (JHU). 2020. Available at: https://coronavirus.jhu.edu/map.html (Accessed: 5 August 2021).

[2] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020; 395: 497–506.

[3] Martin-Barreiro C, Ramirez-Figueroa JA, Cabezas X, Leiva V, Galindo-Villardón MP. Disjoint and functional principal component analysis for infected cases and deaths due to COVID-19 in South American countries with sensor-related data. Sensors. 2021; 21: 4094.

[4] Brum AA, Duarte-Filho GC, Vasconcelos GL. Application Modinterv COVID-19. 2020. Available at: https://fisica.ufpr.br/modinterv (Accessed: 5 August 2021).

[5] Duarte-Filho GC, Brum AA, Ospina R, Almeida FA, Macêdo AM, Vasconcelos GL. Recife e Belém são atualmente as únicas capitais que já estão na fase de saturação da Covid-19 no Brasil. Scielo Preprints. (in press)

[6] Jerez-Lillo N, Álvarez BL, Gutiérrez JM, Figueroa-Zúñiga JI, Leiva V. A statistical analysis for the epidemiological surveillance of COVID-19 in Chile. Signa Vitae. 2021. (in press)

[7] Liu Y, Mao C, Leiva V, Liu S, Silva Neto WA. Asymmetric autoregressive models: Statistical aspects and a financial application under COVID-19 pandemic. Journal of Applied Statistics. 2021. (in press)

[8] Chahuán-Jiménez K, Rubilar R, de la Fuente-Mella H, Leiva V. Breakpoint analysis for the COVID-19 pandemic and its effect on the stock markets. Entropy. 2021; 23: 100.

[9] de la Fuente-Mella H, Rubilar R, Chahuán-Jiménez K, Leiva V. Modeling COVID-19 cases statistically and evaluating their effect on the economy of the countries. Mathematics. 2021; 9: 1558.

[10] Cabezas X, García S, Martin-Barreiro C, Delgado E, Leiva V. A two-stage location problem with order solved using a Lagrangian algorithm and stochastic programming for a potential use in COVID-19 vaccination based on sensor-related data. Sensors. 2021; 21: 5352.

[11] Rojas F, Leiva V, Huerta M, Martin-Barreiro C. Lot-size models with uncertain demand considering its skewness/kurtosis and stochastic programming applied to hospital pharmacy with sensor-related COVID-19 data. Sensors. 2021; 21: 5198.

[12] Cauchemez S, Fraser C, Van Kerkhove MD, Donnelly CA, Riley S, Rambaut A, *et al*. Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. The Lancet Infectious Diseases. 2014; 14: 50–56.

[13] Cori A, Donnelly CA, Dorigatti I, Ferguson NM, Fraser C, Garske T, *et al*. Key data for outbreak evaluation: building on the Ebola experience. Philosophical Transactions of the Royal Society B. 2017; 372: 20160371.

[14] Yozwiak NL, Schaffner SF, Sabeti PC. Data sharing: Make outbreak research open access. Nature. 2015; 518: 477–479.

[15] Ienca M, Vayena E. On the responsible use of digital data to tackle the COVID-19 pandemic. Nature Medicine. 2020; 26: 463–464.

[16] Moorthy V, Henao Restrepo AM, Preziosi MP, Swaminathan S. Data sharing for novel coronavirus (COVID-19). Bulletin of the World Health Organization. 2020; 98: 150.

[17] Drew DA, Nguyen LH, Steves CJ, Menni C, Freydin M, Varsavsky T, *et al*. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. Science. 2020; 368: 1362–1367.

[18] Zhuang Z, Zhao S, Lin Q, Cao P, Lou Y, Yang L, *et al*. Preliminary estimation of the novel coronavirus disease (COVID-19) cases in Iran: a modelling analysis based on overseas cases and air travel data. International Journal of Infectious Diseases. 2020; 94: 29–31.

[19] Plohl N, Musil B. Modeling compliance with COVID-19 prevention guidelines: the critical role of trust in science. Psychology, Health and Medicine. 2020; 26: 1–12.

[20] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases. 2020; 20: 533–534.

[21] Dey SK, Rahman MM, Siddiqi UR, Howlader A. Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. Journal of Medical Virology. 2020; 92: 632–638.

[22] Hamzah FB, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, *et al*. CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. Bull World Health Organ. 2020; 1.

[23] Cheshmehzangi A. How Cities Cope in Outbreak Events? The City in Need (pp. 17–39). Springer: Singapore. 2020.

[24] Fronterre C, Read JM, Rowlingson B, Bridgen J, Alderton S, Diggle PJ, *et al*. COVID-19 in England: spatial patterns and regional outbreaks. medRxiv. 2020. (in press)

[25] Nagraj VP, Randhawa N, Campbell F, Crellen T, Sudre B, Jombart T. epicontacts: Handling, visualisation and analysis of epidemiological contacts. F1000Research. 2018; 7: 566.

[26] Liu C, Wu X, Niu R, Wu X, Fan R. A new SAIR model on complex networks for analysing the 2019 novel coronavirus (COVID-19). Nonlinear Dynamics. 2020; 101: 1777–1787.

[27] Block P, Hoffman M, Raabe IJ, Dowd JB, Rahal C, Kashyap R, *et al*. Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. Nature Human Behaviour. 2020; 4: 588–596.

[28] Parag KV, Donnelly CA. Using information theory to optimise epidemic models for real-time prediction and estimation. PLOS Computational Biology. 2020; 16: e1007990.

[29] Chowell G, Luo R, Sun K, Roosa K, Tariq A, Viboud C. Real-time forecasting of epidemic trajectories using computational dynamic ensembles. Epidemics. 2020; 30: 100379.

[30] Liang K. Mathematical model of infection kinetics and its analysis for COVID-19, SARS and MERS. Infection, Genetics and Evolution. 2020; 82: 104306.

[31] Van den Broeck W, Gioannini C, Gonçalves B, Quaggiotto M, Colizza V, Vespignani A. The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. BMC Infectious Diseases. 2011; 11: 37.

[32] Vasconcelos GL, Macêdo AMS, Ospina R, Almeida FAG, Duarte-Filho GC, Brum AA, *et al*. Modelling fatality curves of COVID-19 and the effectiveness of intervention strategies. PeerJ. 2020; 8: e9421.

[33] Moore S, Rogers T. Predicting the Speed of Epidemics Spreading in Networks. Physical Review Letters. 2020; 124: 068301.

[34] Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling. 2017; 2: 379–398.

[35] Heymann DL, Shindo N. COVID-19: what is next for public health? The Lancet. 2020; 395: 542–545.

[36] Sun J, He WT, Wang L, Lai A, Ji X, Zhai X, *et al*. COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. Trends in Molecular Medicine. 2020; 26: 483–495.
[37] Birrell PJ, De Angelis D, Presanis AM. Evidence Synthesis for Stochastic Epidemic Models. Statistical Science. 2018; 33: 34–43.
[38] Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, *et al*. Early dynamics of transmission and control of COVID-19: A mathematical modelling study. The Lancet Infectious Diseases. 2020; 20: 553–558.
[39] Alahmadi A, Belet S, Black A, Cromer D, Flegg JA, House T, *et al*. Influencing public health policy with data-informed mathematical models of infectious diseases: Recent developments and new challenges. Epidemics. 2020; 32: 100393.
[40] Ferraz C, Petenate AJ, Wanderley AL, Ospina R, Torres J, Peruzzi-Moreira A. COVID-19: Monitoring by Shewhart charts. Revista Brasileira de Estatística. 2020; 78: 23–41. (In Portuguese)
[41] Perla RJ, Provost SM, Parry GJ, Little K, Provost LP. Understanding variation in COVID-19 reported deaths with a novel Shewhart charts application. International Journal for Quality in Health Care. 2021; 33: mzaa069.
[42] Bustos N, Tello M, Droppelmann G, Garcia N, Feijoo F, Leiva V. Machine learning techniques as an efficient alternative diagnostic tool for COVID-19 cases. Signa Vitae. 2022; 18: 23-33.
[43] Ramirez-Figueroa JA, Martin-Barreiro C, Nieto-Librero AB, Leiva V, Galindo-Villardón MP. A new principal component analysis by particle swarm optimization with an environmental application for data science. Stochastic Environmental Research and Risk Assessment. 2021; 35: 1969–1984.
[44] Pita R, Pinto C, Sena S, Fiaccone R, Amorim L, Reis S, *et al*. On the Accuracy and Scalability of Probabilistic Data Linkage over the Brazilian 114 Million Cohort. IEEE Journal of Biomedical and Health Informatics. 2018; 22: 346–353.
[45] Duboue P. The Art of Feature Engineering: Essentials for Machine Learning. Cambridge University Press: Cambridge. 2020.
[46] Khashan EA, Eldesouky AI, Fadel M, Elghamrawy SM. A big data-based framework for executing complex query over COVID-19 datasets (COVID-QF). arXiv. 2020. (in press)
[47] Jombart T, Kamvar ZN, FitzJohn R, Cai J, Bhatia S, Schumacher J, *et al*. Incidence: Compute, Handle, Plot and Model Incidence of Dated Events. R package version. 2020; 1.
[48] Matrajt L, Leung T. Evaluating the effectiveness of social distancing interventions to delay or flatten the epidemic curve of Coronavirus disease. Emerging Infectious Diseases. 2020; 26: 1740–1748.
[49] Vasconcelos GL, Macêdo AM, Duarte-Filho GC, Araújo AA, Ospina R, Almeida FA. Complexity signatures in the COVID-19 epidemic: Power law behaviour in the saturation regime of fatality curves. medRxiv. 2020. (in press)
[50] Wu K, Darcet D, Wang Q, Sornette D. Generalized logistic growth modeling of the covid-19 outbreak in 29 provinces in China and in the rest of the world. medRxiv. 2020. (in press)
[51] Velasco H, Laniado H, Toro M, Catano-Lopez A, Leiva V, Lio Y. Modeling the risk of infectious diseases transmitted by Aedes aegypti using survival and aging statistical analysis with a case study in Colombia. Mathematics. 2021; 9: 1488.
[52] Svensson A. A note on generation times in epidemic models. Mathematical Biosciences. 2007; 208: 300–311.
[53] Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. American Journal of Epidemiology. 2004; 160: 509–516.
[54] Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings of the Royal Society B. 2006; 274: 599–604.
[55] Ali ST, Kadi AS, Ferguson NM. Transmission dynamics of the 2009 influenza a (H1N1) pandemic in India: the impact of holiday-related school closure. Epidemics. 2013; 5: 157–163.
[56] Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS ONE. 2007; 2: e758.
[57] an der Heiden M, Hamouda O. Schätzung der aktuel-len entwicklung der Sars-Cov-2-epidemie in Deutschland-Nowcasting. Epidemiologisches Bulletin. 2020; 17: 10–15.
[58] Fierro R, Leiva V, Balakrishnan N. Statistical Inference on a Stochastic Epidemic Model. Communications in Statistics. 2015; 44: 2297–2314.
[59] Diekmann O, Heesterbeek JA, Metz JA. On the definition and the computation of the basic reproduction ratio Ro in models for infectious diseases in heterogeneous populations. Journal of Mathematical Biology. 1990; 28: 365–382.
[60] Bürger R, Chowell G, Lara-Díıaz LY. Comparative analysis of phenomenological growth models applied to epidemic outbreaks. Mathematical Biosciences and Engineering. 2019; 16: 4250–4273.
[61] Khan U, Mehta R, Arif MA, Lakhani OJ. Pandemics of the past: A Narrative Review. The Journal of the Pakistan Medical Association. 2020; 70: S34–S37.
[62] Aykroyd RG, Leiva V, Ruggeri F. Recent developments of control charts, identification of big data sources and future trends of current research. Technological Forecasting and Social Change. 2019; 144: 221–232.
[63] Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: a systematic review. Journal of Biomedical Informatics. 2020; 108: 103500.
[64] Lanza F, Seidita V, Chella A. Agents and robots for collaborating and supporting physicians in healthcare scenarios. Journal of Biomedical Informatics. 2020; 108: 103483.
[65] Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, *et al*. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. British Medical Journal. 2020; 368: l6927.
[66] Badie-Modiri A, Karsai M, Kivelä M. Efficient limited-time reachability estimation in temporal networks. Physical Review E. 2020; 101: 052303.
[67] Topirceanu A, Udrescu M, Marculescu R. Centralized and decentralized isolation strategies and their impact on the COVID-19 pandemic dynamics. arXiv. 2020. (in press)
[68] Alamo T, Reina DG, Millán P. Data-driven methods to monitor, model, forecast and control COVID-19 pandemic: Leveraging data science, epidemiology, and control theory. arXiv. 2020. (in press)
[69] Nikolaou P, Dimitriou L. Identification of critical airports for controlling global infectious disease outbreaks: Stress-tests focusing in Europe. Journal of Air Transport Management. 2020; 85: 101819.
[70] Ruiz-Estrada MA, Koutronas E. The application of the 2019-nCoV global economic impact simulator (the 2019-nCoV-GEI-Simulator) in China. Social Science Research Network. 2020. (in press)
[71] Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, *et al*. A deep learning algorithm using CT images to screen for Coronavirus disease (COVID-19). European Radiology. 2021; 31: 6096–6104.
[72] Pereira RM, Bertolini D, Teixeira LO, Silla CN Jr, Costa YMG. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. Computer Methods and Programs in Biomedicine. 2020; 194: 105532.
[73] Son H, Hyun C, Phan D, Hwang HJ. Data analytic approach for bankruptcy prediction. Expert Systems with Applications. 2019; 138: 112816.
[74] Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. JMIR Public Health and Surveillance. 2020; 6: e18828.
[75] Jahanbin K, Rahmanian V. Using twitter and web news mining to predict COVID-19 outbreak. Asian Pacific Journal of Tropical Medicine. 2020; 13: 378–380.
[76] Bukowski M, Farkas R, Beyan O, Moll L, Hahn H, Kiessling F, *et al*. Implementation of eHealth and AI integrated diagnostics with multidisciplinary digitized data: are we ready from an international perspective? European Radiology. 2020; 30: 5510–5524.
[77] Foraker R, Mann DL, Payne PRO. Are Synthetic Data Derivatives the Future of Translational Medicine? JACC: Basic to Translational Science. 2018; 3: 716–718.
[78] Hollingsworth TD, Medley GF. Learning from multi-model comparisons: Collaboration leads to insights, but limitations remain. Epidemics. 2017; 18: 1–3.
[79] Király FJ, Mateen B, Sonabend R. NIPS-not even wrong? A systematic review of empirically complete demonstrations of algorithmic effectiveness in the machine learning and artificial intelligence literature. arXiv. 2018. (in press)

[80] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation. 2015; 131: 211–219.

[81] Al-Shahi Salman R, Beller E, Kagan J, Hemminki E, Phillips RS, Savulescu J, *et al*. Increasing value and reducing waste in biomedical research regulation and management. The Lancet. 2014; 383: 176–185.

[82] Nor AKM, Pedapati SR, Muhammad M, Leiva V. Overview of explainable artificial intelligence for prognostic and health management of industrial assets based on preferred reporting items for systematic reviews and meta-analyses. Sensors. 2021; 21: 8020.

[83] Cortes C, Jackel LD, Chiang WP. Limits on learning machine accuracy imposed by data quality. In Tesauro G, Touretzky D, Leen T (eds.) Advances in Neural Information Processing Systems (pp. 239–246). MIT Press: MA, US. 1995.

How to cite this article: Raydonal Ospina, André Leite, Cristiano Ferraz, André Magalhães, Víctor Leiva. Data-driven tools for assessing and combating COVID-19 outbreaks in Brazil based on analytics and statistical methods. Signa Vitae. 2022; 18(3): 18-32. doi:10.22514/sv.2021.253.