Computer Methods and Programs in Biomedicine 211 (2021) 106399

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

COVID-19: Estimation of the transmission dynamics in Spain using a stochastic simulator and black-box optimization techniques

Marcos Matabuena[a,*], Pablo Rodríguez-Mier[b], Carlos García-Meixide[c], Victor Leborán[a]

[a] CiTIUS (Centro Singular de Investigación en Tecnoloxías Intelixentes), Universidade de Santiago of Compostela, Santiago de Compostela, Spain
[b] Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, Toulouse 31300, France
[c] Universidade de Santiago de Compostela, Santiago de Compostela, Spain

ARTICLE INFO

Article history:
Received 19 April 2021
Accepted 31 August 2021

Keywords:
Epidemic models
COVID-19
Computing science
Stochastic processes
Evolutionary computations

ABSTRACT

Background and objectives: Epidemiological models of epidemic spread are an essential tool for optimizing decision-making. The current literature is very extensive and covers a wide variety of deterministic and stochastic models. However, with the increase in computing resources, new, more general, and flexible procedures based on simulation models can assess the effectiveness of measures and quantify the current state of the epidemic. This paper illustrates the potential of this approach to build a new dynamic probabilistic model to estimate the prevalence of SARS-CoV-2 infections in different compartments.

Methods: We propose a new probabilistic model in which, for the first time in the epidemic literature, parameter learning is carried out using gradient-free stochastic black-box optimization techniques simulating multiple trajectories of the infection dynamics in a general way, solving an inverse problem that is defined employing the daily information from mortality records.

Results: After the application of the new proposal in Spain in the first and successive waves, the result of the model confirms the accuracy to estimate the seroprevalence and allows us to know the real dynamics of the pandemic a posteriori to assess the impact of epidemiological measures by the Spanish government and to plan more efficiently the subsequent decisions with the prior knowledge obtained.

Conclusions: The model results allow us to estimate the daily patterns of COVID-19 infections in Spain retrospectively and examine the population's exposure to the virus dynamically in contrast to seroprevalence surveys. Furthermore, given the flexibility of our simulation framework, we can model situations —even using non-parametric distributions between the different compartments in the model— that other models in the existing literature cannot. Our general optimization strategy remains valid in these cases, and we can easily create other non-standard simulation epidemic models that incorporate more complex and dynamic structures.

1. Introduction

The spread of SARS-CoV-2 is generating unprecedented health and socio-economical crisis worldwide, being one of the most significant challenges in Europe since World War II. In the light of this emergency, the governments ought to organize an appropriate schedule and optimize political decisions based on scientific evidence to avoid the collapse of the healthcare system, reduce virus-related mortality and minimize the potential effects of an economic recession [39,44,50,51].

Given the vital capacity of the virus to spread and the lack of effectiveness of preventive measures, many countries have been systematically forced to lock down the population temporarily. Although these policies may help control the spread of the virus, they are economically unsustainable over time. In this regard, forecasting the evolution and consequences of the pandemic based on the exposure of the population becomes a critical factor in decision-making [23,40]. However, it is first necessary to assess the current spread of the epidemic to rigorously predict these effects, which is often unknown due to the limited tracking of new infections and active cases.

At the beginning of the 20th century, the first mathematical models to study the dynamics of an epidemic were introduced. Probably the best-known method is the susceptible-infected-

[*] Corresponding author.
E-mail address: marcos.matabuena@usc.es (M. Matabuena).

https://doi.org/10.1016/j.cmpb.2021.106399

recovered model (SIR). SIR model and its variations [33,38] divide the population into compartments, and using differential (deterministic) equations, the number of individuals in each of the compartments over time are estimated. Since then, many new variations of these models that also involved stochastic versions have been introduced in the literature (see for review [3,5,49,70] or other contemporary examples [14,59]).

Despite the enormous progress with these models in recent decades, their direct applications can be limited in several settings. First, most models explain the dynamics of the epidemic at the population level [28,37,41] excluding relevant individual interactions. Second, model-specific assumptions can be restrictive and abstracted from practice. For example, practitioners use Poisson's homogeneous process to handle the mechanism of new infections or parametric distributions that determine time transitions [29,42]. Third, introducing model reformulations in practice can be challenging and time-consuming with the current optimization strategies of the literature-based primary on designed specific procedures with likelihood equations [10]. We believe this is a critical factor limiting the performance of initial experiments and the use of novel and non-standard formulation of epidemic models in a routine and straightforward manner.

As in statistical learning theory, we can say that there is no universal model for all scenarios. Instead, we probably have to design specific models following the existing epidemiological evidence for each situation and introduce the prior knowledge obtained into models.

Simulation techniques are a prominent alternative method to build complex and more realistic epidemiological models at a high computational cost. However, their use is not new, and several agent models have appeared in the literature [75,79], which allow modeling the possible impact of different interventions on the evolution of a pandemic. For instance, we can study the impact of vaccination, social distance, or lockdown policies in the reduction of infections or mortality [32]. More specifically, some of the specific advantages are summarized below:

- We can introduce a wide variety of distributions in the components of the model that can be specified with intractable complex likelihood equations [17] or even non-parametric assumptions.
- Simulation models allow the introduction of personal information of individuals, such as age and other covariates relevant to disease manifestation, without introducing challenges in the model implementation, unlike classical epidemic models.
- Adding some constraints into the model, such as the social interactions between individuals, is not complicated from a model design perspective and only increases computation demands.

A cornerstone in expanding this area of research is the ability to obtain reliable solutions to the underlying optimization problem without resorting to problem-specific optimization strategies. Advances in computational power and the field of Black-Box optimization [66] can be an essential milestone in achieving such a goal and being able to examine different models without consuming much time using general procedures. However, sometimes, this strategy requires high-computing environments. Only by evaluating an objective function can these algorithms learn reasonable solutions performing multiple simulations.

In this paper, we explore this idea. Using a flexible yet straightforward dynamic probabilistic model that we designed based on the biological evidence of the onset of the pandemic, we estimate the seroprevalence in different regions of Spain along different waves. We also reconstruct the dynamics of infections and recoveries in different compartments to answer specific epidemiological questions, such as when the famous infection peaks happened. To do this, we solve an inverse problem with the mortality records to estimate some specific model parameters, such as the daily rate of infections. In this task, we use, for the first time in this area, the CMAES algorithm [30], one of the state-of-the-art Black-Box stochastic optimization methods that have been in our previous tests more competitive than other existing algorithms.

We must note that our primary purpose in mathematical modeling is not to make forecasts about the dynamic evolution of the pandemic. Instead, the aim of our proposal is to perform backcasting: to retrospectively reconstruct the dynamics of infections while estimating seroprevalence in the different compartments of the model. By estimating this information, we can better characterize the concrete mechanisms of virus transmission in the territories analyzed. Thus, for example, we can guide political decisions in a more refined sense by establishing more advanced and personalized epidemiological thresholds to determine lock-down policies, according to each territory's specific socio-economic and healthcare factors and the dynamic evolution of the number of infections drawn by our model in the different compartments.

1.1. Outline

The article structure is as follows: First, we introduce our new mathematical model to estimate the spread of COVID-19 in regions and countries together with the model optimization strategy used. Then, we introduce some historical background on the evolution of the COVID-19 pandemic in Spain. Also, some demographic and economic characteristics of the Spanish population are presented. Next, we evaluate the behavior of the model and we illustrate its usefulness, performing different analyses across several Spanish regions, reporting the day-to-day evolution of susceptible, infected, and recovered patients. Finally, we discuss the results, the model limitations, and the power and value of the new methodology presented in the existing literature.

1.2. Aims of the analysis

In order to show the usefulness and broad potential of our proposal for practitioners, we perform different analyses that allow answering the following epidemiological questions:

1. What was the spread of the virus in the first wave in different regions of Spain like?. For example, when did the peak of infections occur?. How many infected people were there in Spain at the end of the lockdown policies?
2. Using a longer time frame, until March 1, 2021, how were the overall dynamics of SARS-CoV-2 in the Spanish population as a whole?. For example, the healthcare situation was critical in October of 2020, and there were discussions about applying a national lockdown; What could be the real epidemiological situation at that time?
3. Given that, from a theoretical point of view, we can reconstruct the dynamics of infections with our model, how was the actual day-to-day capacity to detect new cases in Spain?

2. Mathematical model and optimization strategy

2.1. Model elements

Suppose that $\mathcal{D} = \{0, 1, \ldots, n\}$ is the set of days under study. Consider the following random processes whose domain is defined on $\mathcal{D}$.

- $\mathcal{S}(t)$: Number of people susceptible to become infected on day t.
- $\mathcal{I}_1(t)$: Number of infected individuals who are incubating the virus on day t.

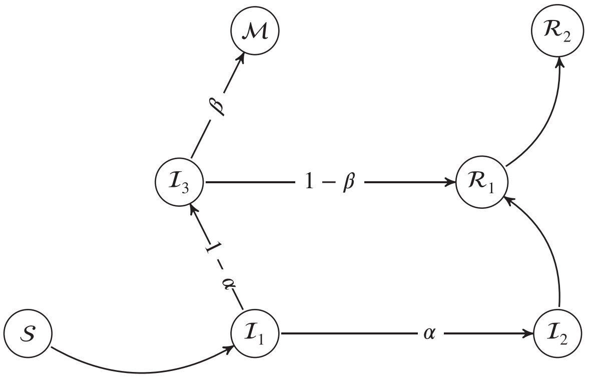

Fig. 1. Diagram of state changes in our model.

- $\mathcal{I}_2(t)$: Number of infected people who have passed the theoretical incubation period and who: (i) don't show symptoms or (ii) symptoms are mild on day t.
- $\mathcal{I}_3(t)$: Number of infected people who have passed the incubation period and do show moderate or severe symptoms on the day t.
- $\mathcal{R}_1(t)$: Number of recovered cases which are still able to infect on the day t.
- $\mathcal{R}_2(t)$. Number of recovered cases that are not able to infect anymore on the day t.
- $\mathcal{M}(t)$: Number of deaths on day t.

Henceforth, we will denote by $\mathcal{I}(t) = \mathcal{I}_1(t) + \mathcal{I}_2(t) + \mathcal{I}_3(t)$ the number of infected people at time $t \in \mathcal{D}$ and $\mathcal{R}(t) = \mathcal{R}_1(t) + \mathcal{R}_2(t)$ the number of recovered people.

The above random processes describe the dynamic of population individuals in separate compartments. We divided the infected and recovered individuals in a broader and specific taxonomy for the particular case of the COVID-19 than the classical epidemiological models [5,38]. There are two main reasons for this. First, the patients tested by healthcare are usually those found in $\mathcal{I}_3$. In this case, there is an essential corpus of prior knowledge about how they evolve, and in case of death, their survival time. Second, there is evidence that there are recovered patients who can still infect others.

2.2. Basic model definition

The causal mechanism of newly infected individuals is introduced below. For each day $t \in \mathcal{D}$, we assume that the new infections $\mathcal{I}_1^{new}(t)$ are generated by the individual interaction of the susceptible people with infected patients and the recovered cases that they can still contaminate (patients that belong of states $\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{I}_3$, $\mathcal{R}_1$).

Formally, we assume that if an individual can contaminate, it does so according to a random variable $X \sim Poisson(\mathcal{R}_i(t))$, being $\mathcal{R}_i(t)$ the average number of new infections that can cause each person in the day t. It is natural to assume that the function $\mathcal{R}_i(t)$ follows a decreasing trend in the first months of the epidemic, basically due to two reasons: (i) quarantine policies have been systematically introduced along with different countries and regions. (ii) the number of susceptible people decreases over time, while the number of infected people can increase. These facts indicate that in our particular setting, it is more complicated to interact with non-infected people.

Once a new infected person arrives to the model (see Fig. 1), we assume that the transitions between the different graph states are modeled by a probability law that verifies the following conditions: (i) the transition probabilities are independent of the absolute instant when such transition takes place (ii) the probabilities depend only on the current state of the patient regardless of the previous path in the graph. In particular, given the $States = \{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3, \mathcal{R}_1, \mathcal{R}_2, \mathcal{M}\}$, and $\alpha, \beta \in [0,1]$, we have: $\mathbf{P}(\mathcal{I}_2|\mathcal{I}_1) = \alpha$, $\mathbf{P}(\mathcal{I}_3|\mathcal{I}_1) = 1-\alpha$, $\mathbf{P}(\mathcal{M}|\mathcal{I}_3) = \beta$, $\mathbf{P}(\mathcal{R}_1|I_3) = 1-\beta$, $\mathbf{P}(\mathcal{I}_3|\mathcal{I}_2) = 1$, $\mathbf{P}(\mathcal{R}_2|\mathcal{I}_1) = 1$; all other transitions take a value equal to zero in probability. More schematically, the P, probability transition matrix, between events is shown in the Eq. (1).

Table 1
Random variables of the time of each transition.

Transition	Random variable	Used references
$\mathcal{I}_1 \rightarrow \mathcal{I}_2$	$Gamma(5.807, 0.948)$	Abdel-Salam and Mollazehi [1], Lauer et al. [43]
$\mathcal{I}_1 \rightarrow \mathcal{I}_3$	$Gamma(5.807, 0.948)$	Abdel-Salam and Mollazehi [1], Lauer et al. [43]
$\mathcal{I}_2 \rightarrow \mathcal{R}_1$	$Uniform(5, 10)$	
$\mathcal{I}_3 \rightarrow \mathcal{R}_1$	$Uniform(9, 14)$	Abdel-Salam and Mollazehi [1]
$\mathcal{I}_3 \rightarrow \mathcal{M}$	$Gamma(6.67, 2.55)$	Abdel-Salam and Mollazehi [1], Novel et al. [57], Salje et al. [68], Verity et al. [81]
$\mathcal{R}_1 \rightarrow \mathcal{R}_2$	$Uniform(7, 14)$	Bi et al. [7], Ehmann et al. [22]

Table 2
Probability of each transition.

Coefficient	Value	Used references
α	0.8	Day [19], Mizumoto et al. [53], Nishiura et al. [56], Tabata et al. [77]
β	0.06	Dudel et al. [20], Fauci et al. [24], Mahase [48], Rajgor et al. [65], Verity et al. [80], Wu et al. [83]

$$P = \left[\begin{array}{ccccccc} & I1 & I2 & I3 & R1 & M & R2 \\ I1 & 0 & \alpha & 1-\alpha & 0 & 0 & 0 \\ I2 & 0 & 0 & 0 & 1 & 0 & 0 \\ I3 & 0 & 0 & 0 & 1-\beta & \beta & 0 \\ R1 & 0 & 0 & 0 & 0 & 0 & 1 \\ M & 0 & 0 & 0 & 0 & 1 & 0 \\ R2 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]. \qquad (1)$$

Additionally, Table 1 shows the random variables that model the time between transitions together with the references used in our elections for real examples of COVID-19 in Spain.

Table 2 shows the values used to model the transition probabilities. The Supplementary Material provides specific details about how the mentioned parameters and functions were selected.

Finally, as we defined the model above, we have a continuous-time probabilistic model. However, the surrogate variables to fit and compare the model results are recorded daily in real-world situations. Consequently, in our implementation we perform the simulation between transitions and new infections on a daily basis and we truncate the corresponding continuous time in days.

2.3. Stochastic model implementation

Our model does not have a closed-form solution. Therefore, in a real-world setting, it is necessary to use statistical simulation methods to approximate specific population characteristics of the stochastic process as quantile functions. Also, we must fit some parameters of the model to characterize the behavior of the study population. For this purpose, we use a sample of the deceased patients $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_s\}$ along the set of days $\mathcal{O} = \{1, \ldots, s\}$.

Next, we suppose that our model (M) is dependent on a vector of parameters $\theta = (\theta_1, \theta_2) \in \mathbf{R}^{p_1} \times \mathbf{R}^{p_2}$ (with $p_1 + p_2 = p$), where θ_1 is a vector of dimension p_1, defined in beforehand, and θ_2 must be estimated from the sample. Furthermore, let us assume that the initial state of the system is characterized by $\mathcal{S} = (\mathcal{S}(0), \mathcal{I}_1(0), \mathcal{I}_2(0), \mathcal{I}_3(0), \mathcal{R}_1(0), \mathcal{R}_2(0), \mathcal{M}(0)) \in \mathbf{N}^7$ and $\mathcal{T} = (\mathcal{T}_1(0), \mathcal{T}_2(0), \mathcal{T}_3(0), \mathcal{T}_4(0), \mathcal{T}_5(0), \mathcal{T}_6(0)) \in \mathbf{N}^m \times \ldots \times \mathbf{N}^m$. $\mathcal{S}$ has the number of elements for each compartment of the model on day 0. $\mathcal{T}$ also contains the amount of remaining days to complete the transition they are in for each individual in the initial state, being m a natural number that represents the maximum number of registered days.

To simplify the notation, for each day $t \in \mathcal{D}$, we denote the average dead trajectory by the function $Mean(\theta_1, \theta_2, \mathcal{S}, \mathcal{T})(t)$.

The next step is to estimate $\hat{\theta}_2$. To do this, we propose to solve the following optimization problem:

$$\hat{\theta}_2 = \arg \min_{\theta_2 \in \mathcal{S} \subset \mathbf{R}^{p_2}} \sum_{i=1}^{s} \omega^i (\mathcal{M}_i - Mean(\theta_1, \theta_2, \mathcal{S}, \mathcal{T})(i))^2, \qquad (2)$$

where $\omega = (\omega^1, \ldots, \omega^s)$ is a weighted vector that can help to improve model estimation. Examples of these weights may be:

$\omega^i = \mathcal{M}_i / \sum_{i=1}^{s} \mathcal{M}_i$ or $\omega^i = (1/\mathcal{M}_i)/(\sum_{i=1}^{s} 1/\mathcal{M}_i)$ $(i = 1, \ldots, s)$.

At this point, it is relevant to note that the above optimization problem (2), as formulated, includes the possibility of introducing constraints in the parameters' space. The previous fact is essential because we often have prior knowledge of the range of parameters, or we can even invoke the biological interpretation of parameters to answer this question. Thus, by introducing this knowledge, we can spread up the speed of optimization Black-Block techniques significantly.

In Eq. (2), we have used the real mean trajectory. However, in practice, this is unknown, and we must approximate it using simulation. Next, we run B different simulations, and we denote by $\overline{M}(\theta_1, \theta_2, \mathcal{S}, \mathcal{T}) = \frac{1}{B} \sum_{i=1}^{B} M^i(\theta_1, \theta_2, \mathcal{S}, \mathcal{T})$, the estimated mean trajectory. $M^i(\theta, \theta_2, \mathcal{S}, \mathcal{T})$ $(i = 1, \ldots, B)$ denotes the result of simulation number i.

So the optimization problem to be solved is:

$$\hat{\theta}_2 = \arg \min_{\theta_2 \in \mathcal{S} \subset \mathbf{R}^{p_2}} \sum_{i=1}^{s} \omega^i (M_i - \overline{M}(\theta_1, \theta_2, \mathcal{S}, \mathcal{T})(i))^2. \qquad (3)$$

Schematically, the overall optimization process is described below.

1. Define an initial θ_2^0 and run B times $M(\theta_1, \theta_2, \mathcal{S}, \mathcal{T})$. We denote by $M^1(\theta_1, \theta_2^0, \mathcal{S}, \mathcal{T}), \ldots, M^B(\theta_1, \theta_2^0, \mathcal{S}, \mathcal{T})$, different results are obtained.
2. Estimate the mean trajectory $\overline{M}(\theta_1, \theta_2^0, \mathcal{S}, \mathcal{T}) = \frac{1}{B} \sum_{i=1}^{B} M^i(\theta_1, \theta_2^0, \mathcal{S}, \mathcal{T})$
3. Estimate the mean square error $R\hat{S}S^0 = \sum_{i=1}^{s} \omega^i (\overline{M}(\theta_1, \theta_2^0, \mathcal{S}, \mathcal{T})(i) - \mathcal{M}_i)^2$.
4. To construct a succession of vectors $\{\theta_2^j\}_{j=1}^{R+1}$ so that $R\hat{S}S^0 > R\hat{S}S^1 > R\hat{S}S^2 > \ldots > R\hat{S}S^{R+1}$. For example with a stochastic optimization solver.
5. Stop after $R+1$ iterations and return θ_2^{R+1} as the optimal parameter of the problem.

In our particular setting, θ_2 contains the parameter of the $\mathcal{R}_i(t)$ function that are defined in Section 2.2. In our preliminary experiments, we assumed that their functional form is equal to $\mathcal{R}_i(t) = \min\{C, ae^{-(bt+ct^2+dt^3+et^4+ft^5)}\}$ where $a \in [0, 3]$, $b \in [-1, 1]$, $c \in [0, 1]$, $d \in [0, 1]$, $e \in [0, 1]$, $f \in [0, 1]$ with $\theta_2 = (a, b, c, d, e, f) \in [0, 3] \times [-1, 1] \times [0, 1] \times [0, 1] \times [0, 1] \times [0, 1]$ and C is a positive constant fixed 0.005. However, our final election after several experiments and sensitivity analysis with different family of functions that include the previously exponential or inverse sigmoid/logit functions is, $\mathcal{R}_i(t) = d + \frac{a}{1+b^{-(t-c)}}$, where $a \in [0, 5]$, $b \in [0, 1]$, $c \in [-30, 30]$, $d \in [0, 0.1]$.

2.4. Structural limitations of the model in COVID-19 pandemic

The behavior of our model is primarily determined by the parameters α, β, and the function $\mathcal{R}_i(t)$, while small variations in the distributions function of the transitions times should not have a significant impact on the seroprevalence estimations. $\mathcal{R}_i(t)$ is estimated from observed mortality records. However, the parameters α, β are fixed, with statics values over time, and perhaps, in practice, their value should vary in successive waves. In addition, α, β, determine the infection fatality rate (*IFR*), the quotient between fatalities, and the number of infections. In particular, in our model *IFR*, is given, $IFR\% = ((1-\alpha) * \beta) * 100$.

Some studies have shown that in the current pandemic, *IFR* is the gold standard epidemiological indicator for monitoring the severity with which the virus has affected different countries [58,60]. However, dynamic estimation of *IFR* is challenging to perform since a precise approximation of the real number of infected people is generally only possible in a single time point, thanks to seroprevalence studies.

Several studies have investigated which factors influence the value of the *IFR*. The primary sources of variations are age and sex [76]. If the distribution of new infections is uniform along time regarding these variables, we can assume constants and statics values for α and β in different time-spans.

Testing that assumption is necessary to determine if we must vary the coefficients α, and β, over different waves. However, since we do not know about the true infections in successive waves, it is not trivial to validate this hypothesis empirically, and some statistical estimations are needed.

In Spain, heath institutions performed an ambitious and unique longitudinal epidemiological study to know the patterns of virus expansion in several time points that allow us to draw estimation about *IFR*. In particular, we have information on infections at the beginning of June and the end of November 2020. Using this information, we can estimate the differences in *IFR* between the first and successive waves. We must note that the Spanish situation in the first wave was critical, with many problems in the elderly population, particularly in nursing homes, and and so changes in the *IFR* along time are expected.

2.5. Model implementation to handle multi-waves

Analogous to an intervention analysis in the context of time series, we need to update the daily infection function to be able to model well the reality in at least the following two situations: at the end of each lockdown, and a new return to normal, or shortly before an explosive growth in the number of new cases or deaths occurs, and no lockdown policies are applied. In these situations, there are abrupt changes in daily infection trends, and therefore the functional form of the function $\mathcal{R}_i(t)$ needs to be modified.

Consider $t_0 = 0 < t_1 < t_2 < \ldots < t_m$, m temporal points, defined with expert knowledge, and, in which, we hypothesize that the trend of the daily infection function is modifiable between different $\{t_s\}_{s=0}^{m}$, for example between two waves. Then, we propose to define the function $\mathcal{R}_i(t)$, as a piecewise function, dependent on the local functions $\mathcal{R}_i^{t_1}(t), \ldots, \mathcal{R}_i^{t_m}(t)$, that is,

$$\mathcal{R}_i(t) = \begin{cases} \mathcal{R}_i^{t_1}(t) & t \in [0, t_1) \\ \mathcal{R}_i^{t_2}(t) & t \in [t_1, t_2) \\ \vdots & \\ \mathcal{R}_i^{t_m}(t) & t \in [t_{m-1}, t_m). \end{cases}$$

In our fits in successive waves, we assume that the functional form of each $\mathcal{R}_i^{t_s}(t)$ $(s = 1, \ldots, m)$ is identical, and equal to prior function $\mathcal{R}_i^{t_s}(t) = d_s + \frac{a_s}{1+b_s^{-(t-c_s)}}$ (for all $t \in [t_{s-1}, t_s)$, and any $s \in \{1, \ldots, m\}$), where, the sub-index s, denotes the dependence of parameters, to time-period $[t_{s-1}, t_s)$. Then, the number of initial free-model parameters is multiplied by the number of periods, m, considered.

Regarding parameters α and β, our implementation allows varying these parameters between $\{t_s\}_{s=0}^{m}$. Let be α_s and β_s, the mentioned parameter for the interval $[t_{s-1}, t_s)$. In this paper $\alpha = \alpha_1 = \ldots = \alpha_m$. However, in the global Spanish analysis along different waves, β will be vary dynamically.

2.6. Conformal simulation bands to quantify model uncertainty

Quantify model uncertainty is a critical point in order to interpretate the results obtained together with their natural limits. In epidemic modeling, according to [21], we can decompose the model uncertainty in three different sources of error:

- *Data uncertainty*: Uncertainty in specified model parameters, estimated externally from data, or in data to which models are fitted.
- *Stochastic uncertainty*: Uncertainty derived from the method of simulation.
- *Structural uncertainty*: Uncertainty in the optimal model structure, or derived from the use of more than one model structure for a given question.

In our settings, we use the dynamic evolution of mortality as a source of information. Then, directly reporting the "Stochastic Uncertainty" drawn for the stochastic simulator as a measure of uncertainty is unrealistic, since we can restrict the number of possible scenarios that happen in practice, according to mortality residuals. More formally, our source of information is a correlated time series that determines the possible reality that occurred, and for a given configuration of the vector parameter θ, we should be removed a significant fraction of unrealistic simulation scenarios.

"Data uncertainty" is not trivial to incorporate in this type of model, and the non-parametric bootstrap solution proposed in D'Agostino McGowan et al. [21] have important limitations, since it does not consider the dynamic and correlation structure of daily reports. More general, bootstrap strategies that can handle the specific dependence structure of daily reports can be adapted as our setting, such as block or wild bootstrap, but it is something out of the scope of this work.

Testing "Structural uncertainty" is also challenging. There may be no universal model in an epidemic modeling context, and perhaps the best model for each situation is a combination of broad dictionary of simpler models that change dynamically as the pandemic evolves.

In this paper, we propose to use a solution similar to the one proposed in Shen et al. [73], which consists of selecting a small fraction of simulations, according to error criteria, e.g., mean squared error, that shape the possible real infection trajectories. With the remaining trajectories, we apply specific conformal inference techniques that, to the best of our knowledge, have not been applied previously to the context of stochastic simulation models. Conformal inference methods are a general methodology to quantify model uncertainty [72], with well-established theoretical foundations [45], and were used in an extensive list of machine learning and statistics problems (see for example a contemporany application [46]).

Below, we introduce the specific mathematical detail of the used conformal simulation bands that can handle heterocedastic noise. First, suppose that $\theta = (\theta_1, \hat{\theta}_2)$ is the optimal parameter configuration, where $\hat{\theta}_2$ was estimated according to methodology proposed in the Section 2.3. Then,

1. Perform $B = 10,000$ simulation of the model $M(\theta_1, \hat{\theta}_2, \mathcal{S}, \mathcal{T})$ and evaluate the mean square error metric (RSS). $\widehat{RSS}^s$ and $M^s(\theta_1, \hat{\theta}_2, \mathcal{S}, \mathcal{T})$ $(s = 1, \ldots, B)$, denote the results for iteration s of mean square error and the estimation of mortaly records in the simulation model respectively.
2. Let $Sel = \{i \in \{1, \ldots, B\} : \widehat{RSS}^i \leq \widehat{RSS}_{(1000)}\}$, the set of index of simulation with the lesser or equal 1000 value of *RSS* estimations. $\widehat{RSS}_{(1000)}$ denote the element 1000, considering the order sample of $\{\widehat{RSS}^s\}_{s=1}^{B}$.
3. Using the subsample of death simulation trajectories $\{M^i(\theta_1, \hat{\theta}_2, \mathcal{S}, \mathcal{T})\}_{i \in Sel}$, estimate pointwise the standart deviation $\hat{\sigma}(t)$, $\forall t \in \mathcal{O}$.
4. Define the conformal score $Score_i = \max_{t \in \mathcal{O}} \frac{|M^i(\theta_1, \hat{\theta}_2, \mathcal{S}, \mathcal{T})(t) - \mathcal{M}_t|}{\hat{\sigma}(t)}$ if $\hat{\sigma}(t) > 0$, $\forall i \in Sel$. Otherwise $Score_i$ is equal to 0.
5. Calculate the quantile $q_\alpha = \arg\min_{t \in \mathbb{R}^+} \{\frac{\sum_{s=1}^{1000} 1\{Score_s \leq t\}}{1000} \geq \alpha\}$, with $\alpha = 0.95$, to guarantee distributional intervals that cover a confidence level of 90%.
6. Define for each $t \in \mathcal{O}$, the confidence interval prediction as $[\overline{M}(\theta_1, \hat{\theta}_2, \mathcal{S}, \mathcal{T})(t) - \hat{\sigma}(t) q_\alpha, \overline{M}(\theta_1, \hat{\theta}_2, \mathcal{S}, \mathcal{T})(t) + \hat{\sigma}(t) q_\alpha]$, where $\overline{M}(\theta_1, \hat{\theta}_2, \mathcal{S}, \mathcal{T})$ denote the simulation mean that correspond with our given mortality estimations.
7. Finally, to build confidence bands for the rest of the stochastic process that makes up our epidemic model, we must select simulation trajectories that lead to the mortality outcome falling within the mortality band calculated in step 6).

A key element of conformal inference is the selection of conformal scores. In this paper, the conformal score has been estimed using the geometry of the supreme norm $|| \cdot ||_\infty$. $|| \cdot ||_\infty$ is often used in the analysis of stochastic processes in the field of functional data analysis to estimate confidence bands (see for example [27]).

2.7. Our probabilistic model proposal in the literature

The literature on epidemic modeling is very broad and includes both mechanistic models built from causal epidemiological knowledge and more statistical approaches that exploit information from historical data with purely predictive models [9,41]. Nowadays, it is not easy to establish a boundary between both approaches since, on many occasions, both methodologies are used from the same point of view or even jointly.

Our model definition is not very complicated from a mathematical point of view. However, it introduces new challenges from the computational and modeling point of view: the simulation of the trajectory of each individual along the population, the introduction of probability distributions that go beyond the exponential law as traditional models do [5], and the use of latent infections models through a non-homogeneous Poisson process, extending in this sense the Markovian property [5]. In [25], the authors define a model similar to ours and propose a resolution framework with Bayesian estimation methods. However, we assume that specific parameters are known from the epidemiological scientific evidence. Our approach using mortality records [61] allow us to obtain reliable seroprevalence estimates, but the difficulty of the estimation increases. We also do not introduce a Bayesian approach but a frequentist approach with its advantages and disadvantages as the need in the Bayesian paradigm of selecting prior functions. Finally, with the philosophy of our simulation model, we can consider complex extensions without making too many changes in the implementation.

2.8. Model optimization with a CMA-ES black-box solver

In our setting, we must find optimal parameters taking into account randomness in approximating the mean. To do this, we should resort to stochastic optimization algorithms.

Many stochastic algorithms are available in the literature. Still, based on the excellent preliminary results, we have decided to use a state-of-the-art evolutionary algorithm: the CMA-ES [30]. CMA-ES is an evolutionary-based derivative-free optimization technique that can optimize a wide variety of functions, including noisy functions, like the one we use in our method. One survey of Black-Box optimization strategies found that CMA-ES outranked 31 other optimization algorithms, performing exceptionally well on "difficult" functions or larger dimensional search spaces [31]. From a theoretical point of view, CMA-ES can be seen as a particular case of the Expectation-Maximization algorithm (EM) [8].

We provide specific algorithm steps in Supplementary Material.

2.9. Inverse problem behavior

The model fit with the mortality records invokes new challenges in model identification so that the inverse problem is well-defined. We fixed some model parameters according to existing scientific evidence to address this issue to make the problem more regularized. A potential alternative to fitting more parameters with the data is to transform the objective function into a multiobjective optimization problem, considering the daily cases or other ICU indicators as additional sources of information. However, in the early stages of the pandemic, and even nowadays, there were essential doubts about the real capacity of detection of new infections.

2.10. Tuning parameter

We performed multiple experiments with CMA-ES to check how the optimization solver behaves. At the same time, through statistical simulation, we estimate the variance of the empirical mean by varying fatalities in different settings. After those initial experiments, we decided to estimate the mean at 300 repetitions, that is, $B = 300$. In addition, we have allowed CMA-ES to run 3000 iterations, starting the optimization algorithm from different random points. To obtain the results of this paper, CMA-ES, was able to find the optimal solution with the function $\mathcal{R}_i(t)$ selected in less than two hours. Finally, the loss function used is Mean Square Error (MSE) with $w^i = 1$ $(i = 1, \ldots, s)$ (see Section 2.3).

2.11. Software details and resources

Our proposal has been implemented in several programming languages- C++, Python and R- although the results that are shown in this article have been obtained with Python. We optimized the parameters using library `pycma` [2], and `numpy` has been used for mathematical operations.

In the different performed statistical analyses, we have used R. Plots have been made both in *R* with `ggplot2` library and in Python with `matplotlib`.

Finally, the training data used to fit the models in the first wave can be downloaded at [67], and [18], that represent the daily Spanish statistics of COVID-19 fatalities. In the most extensive analysis of the overall Spanish population, we use the excess of mortality as a source of information. The raw data to estimated excess of mortality can be obtained in the public web interface related to MoMo-daily Spanish mortality surveillance system (https://momo.isciii.es/public/momo/dashboard/momo_dashboard.html), coordinated by National Institute of Health Carlos III (ISCIII).

We release the code used in this paper for the benefit of the scientific community at (https://github.com/covid19-modeling).

3. COVID-19 in Spain

Spain was one of the first countries worldwide to experience the effects of COVID-19, after China and Italy. However, the consequences were more dramatic despite the delayed outbreak start with respect to these countries. To give a better context to the evolution of Coronavirus in Spain, in the first wave, and compare it with other countries, we introduce some historical background:

- January 31st. The first positive result was confirmed on Spanish territory in La Gomera. At that time, there were around 10,000 confirmed cases worldwide.
- February 12th. The Mobile World Congress, one of the most remarkable technological congresses in the world, to be held in Barcelona, was cancelled.
- March 8th. Multitudinous marches were celebrated in Spain. Also, sports competitions and other events were held as usual.
- March 13th. Madrid reported 500 new cases of Cov-19 in one day (64 deaths total). Wuhan had gone into lockdown with 400 new cases per day (17 deaths total).
- March 14th. With the increase in the outbreak of infections, the government declared a quarantine throughout the country.
- March 21st. Due to an overloaded health system, the first patients started to arrive at new makeshift hospitals.
- April 3rd. Spain accounts for a total of 117,710 confirmed cases, surpassing Italy for the first time.
- April 6th. Spain becomes the country in the world with more deaths per million inhabitants.
- April 9th. The FMI forecasts that 170 countries are going to fall into recession this year in the worst crisis since the Great Depression.
- April 18th. The Spanish government changes protocols for the daily statistics of COVID-19.

Fig. 2, shows the accumulated number of cases and deceases respectively in the previous periods in Spain, Italy, China, United Kingdom, and the United States according to the data supplied by the different governments.

Following the statistics of the Population Reference Bureau, Spain is the 20th country with the world's oldest population [12]. The country demographic structure, poverty rates, and epidemiological profiles are essential to compare mortality between countries. In the Coronavirus disease, relative and absolute case-fatality risk (CFR) [26] increases dramatically with age and with comorbidity, as evidenced by the current literature. Relative risk can increase by more than 900% in patients over 60 [85].

Subsequently, we perform a descriptive analysis in the regions of Spain that we analyze in this paper: Galicia, País Vasco, Castilla y León, Madrid, Cataluña. Table 3 contains the essential demographic and socioeconomic characteristics of these regions. We can see that Castilla y León is the region with the highest proportion of elderly people. At the same time, Castilla y León has the most delocalized population centres, and the País Vasco is the region with the lowest poverty rate. The national poverty rate is higher than the other analyzed regions because we do not include the most poverty regions.

Spain is a multicultural country where there are significant economic, geographical, social, and demographic differences throughout the regions. All these peculiarities make Spain an interesting country to extrapolate the effects of COVID-19 spread to other regions and countries.

Finally, in Fig. 3, we show the evolution of infections and fatalities among the regions under consideration in the first wave. As we can see, Madrid is the most affected region, while Galicia is the least affected, despite its older population. However, it is essential to note that the outbreak began later, and the containment was carried out earlier than in Madrid.

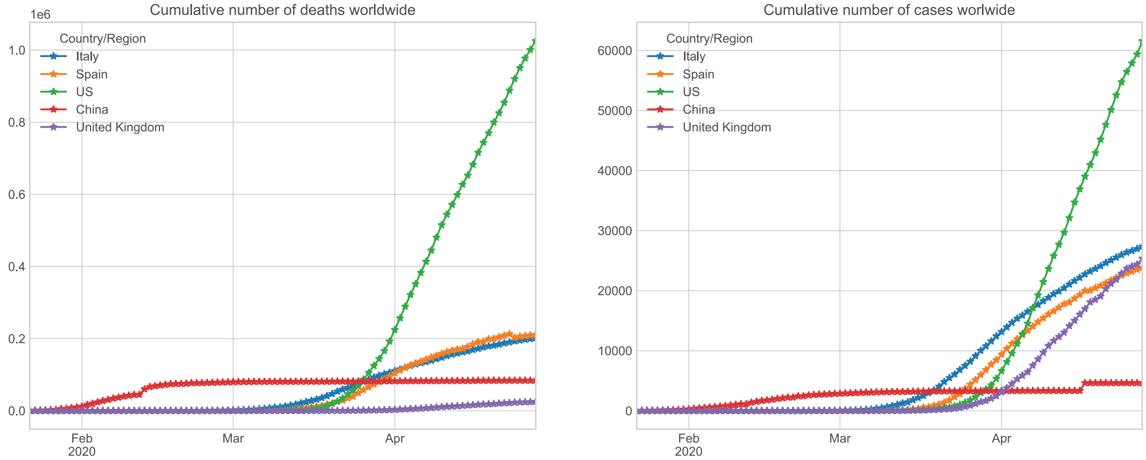

Fig. 2. Spread and number of deaths of Coronavirus in Spain, Italy, China, and the United States. Number of accumulated infected patients (left) and the number of accumulated deaths (right) [13].

Table 3
Demographic and socioeconomics characteristics of the Spanish population throughout some regions: Galicia, País Vasco, Castilla y León, Cataluña, Madrid [34].

	Galicia	País Vasco	Castilla León	Cataluña	Madrid	Spain
Population	2,698,763	2,181,916	2,553,301	7,609,497	6,685,470	47,100,396
At-risk-of-poverty rate	18.8	8.6	16.1	13.6	16.1	21.5
Population density	91.28	305.19	25.47	239.01	830.02	93.08
Percentage of population by age group						
0–9	7.50	8.93	7.51	9.78	9.92	9.28
9–18	7.56	8.78	7.79	9.74	9.44	9.37
18–30	10.39	10.66	10.67	12.72	12.80	12.42
30–45	21.18	20.28	19.47	22.24	23.24	22.07
45–60	22.77	23.28	23.60	21.77	22.23	22.61
60–80	22.60	21.41	22.30	18.24	17.33	18.70
from 81 on	8.01	6.67	8.65	5.50	5.03	5.56

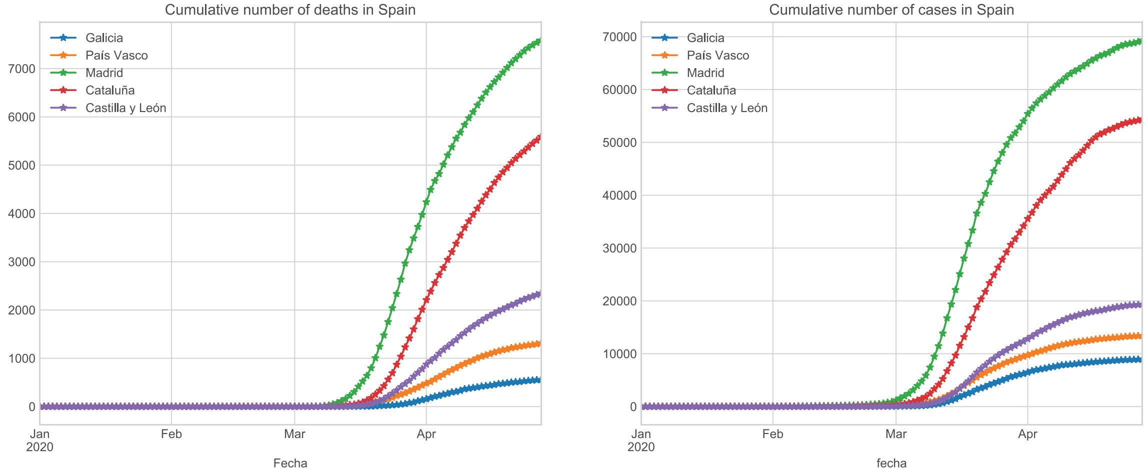

Fig. 3. Evolution of accumulated infected (left) and death patients (right) in Galicia, País Vasco, Castilla y León, Madrid, Cataluña.

4. Results

4.1. First wave analysis

In order to explore the limits of the model in a more challenging scenario, we start the analysis with the first wave. In this period, most Spanish seroprevalence surveys were performed, and therefore we have a reliable estimation of the number of infections accross different Spanish regions in a single time point, allowing us to evaluate our model performance. In addition, the information is of poor quality, and epidemiological evidence is scarce; thus, making estimations in this scenario is more complicated. More specifically, we restrict the model analysis to April 26st in Galicia, País Vasco, Castilla y León, Madrid and Cataluña. As there is considerable uncertainty about mortality records, we assuming that these two scenarios hold:

1. We assume that the number of real deaths due to Coronavirus is reflected in official records.

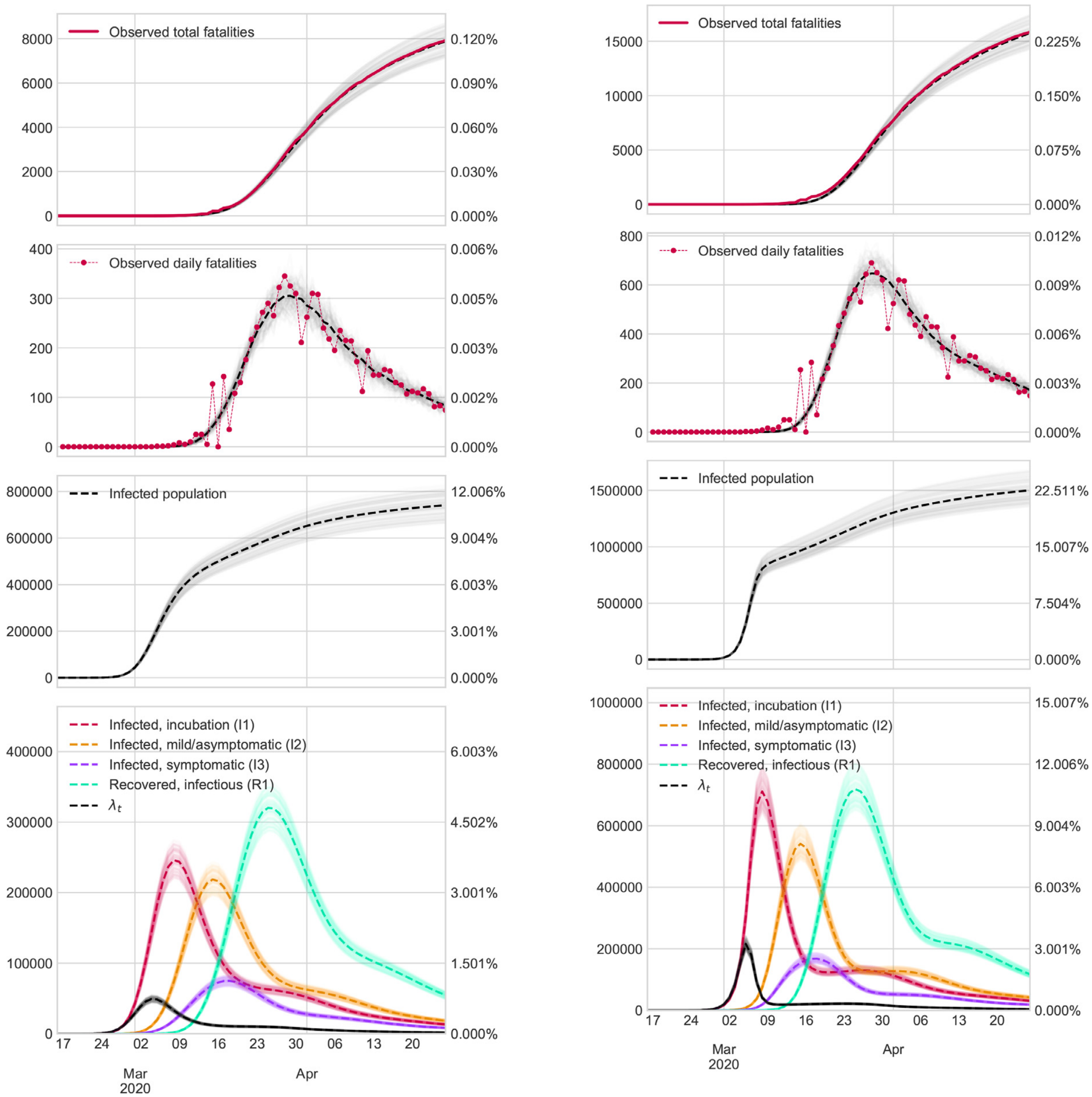

Fig. 4. Results in Madrid.

2. We suppose a more pessimistic scenario. We assume that many people have died of Coronavirus, but they have not been included in the records because a diagnostic test was not performed. In particular, we shall suppose that the number of deaths is twice as high as those indicated in the official records each day.

To display results in an easy-to-view format, we graphically represent the evolution of some states defined at the beginning of Section 2.1 in the two cases considered. In addition, to gain further insights into the results, we show (i) the number of people who may be contaminated or have already transmitted the virus as a percentage of the population size; and (ii) the rate of new infections each day (denoted in the Figures as λ_t).

Finally, we introduce confidence bands of our estimations using methodology described in Section 2.6.

Here, we only show the Figures that contain results in Galicia and Madrid. The rest of the Figures are available in Supplementary Material.

4.2. Multi-wave analysis

To show more recent and informative results on Coronavirus dynamics in Spain, we adjusted the model for the total Spanish population until 1 March 2021. To avoid choosing between the two scenarios above, we use excess mortality as a source of information to feed our model. α has been selected with the same criteria as the first wave. However, β is fixed with a value equal to 0.085-in the first period, while the rest with 0.0425. These values were established to guarantee an *IFR* of 1.7% in the first wave and 0.85% in successive periods. Specific details about *IFR* estimations are relegated to the Supplementary Material. Daily infection function $\mathcal{R}_i(t)$ was fitted as a piecewise function (see Section 2.5 for details). In particular, the cut-off points selected for the jumps are as specified below (Figs. 4–6).

1. 1 March to 30 May.
2. 1 June to 5 July.
3. 6 July to 15 August.

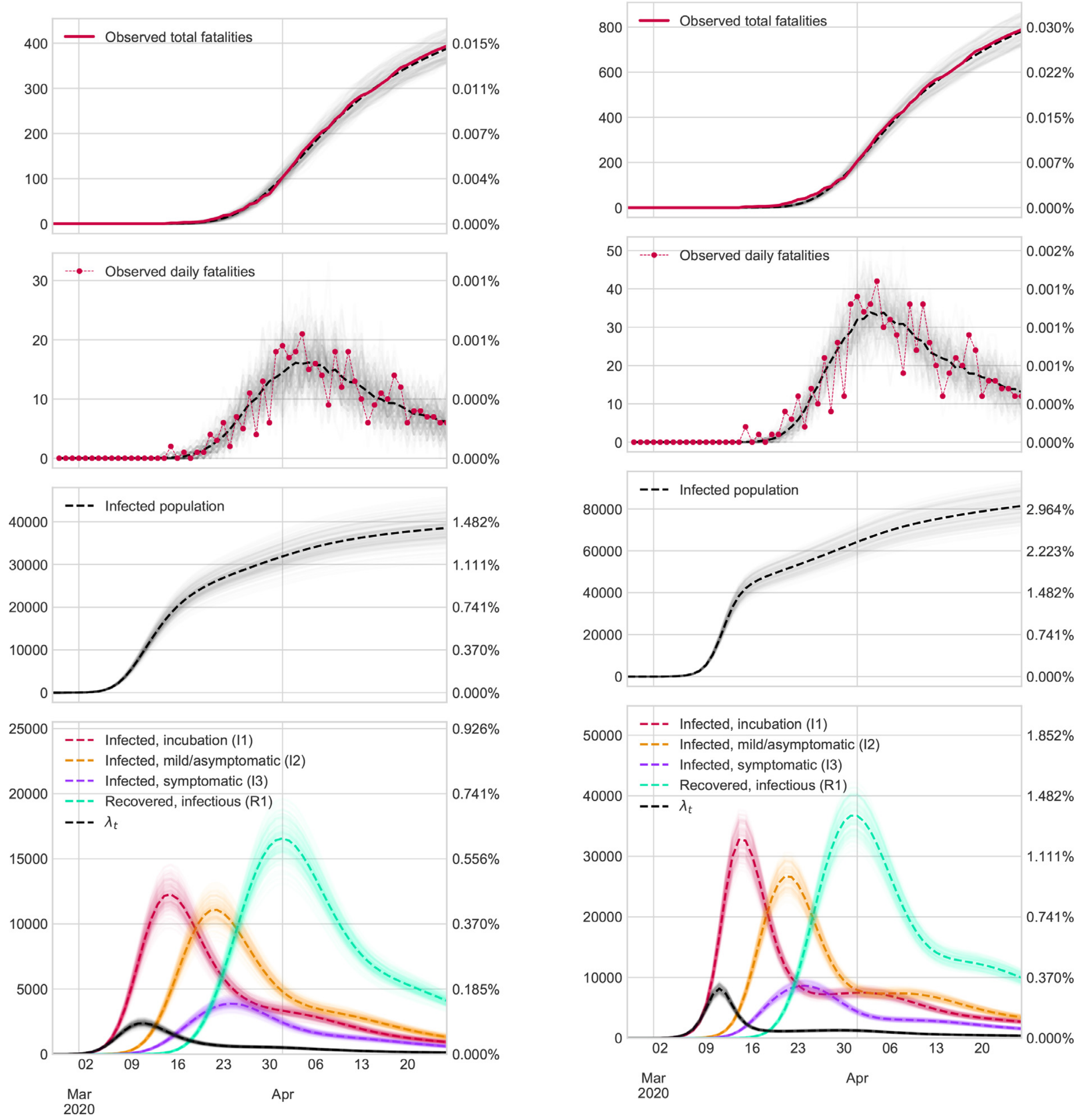

Fig. 5. Results in Galicia.

4. 16 August to 30 September.
5. 1 October to 24 December
6. 25 December to 1 March.

It is important to note that these periods correspond to critical events in pandemic evolution, such as a change in lockdown politics, holidays, or other events that led to abrupt changes in the dynamics of infections.

4.3. Analysis of results

The most relevant results in the first wave (1st March 2020 to 26th April 2020) are outlined below:

- Madrid was the most affected region by COVID-19. If we consider an extreme setting (e.g., the number of deaths is double that reported by the Government), 22.5% of the population could have been infected or recovered from the virus. On April 26st, there may have been almost 1,2 million of patients recovered.
- Galicia was the region that suffered the mildest effects. The percentage of infected people was less than 2.9%.
- Castilla y León, Pais Vasco and Cataluña could have suffered the effects of COVID-19 with a proportional magnitude. In those regions, the percentage of infections could have been between 6 and 12% of the population.
- The peak of new infections probably occurred at the start of quarantine, while the peak of people who can contaminate took place between March 17 and 24.
- The number of new infections have been dramatically reduced after the introduction of containment measures.
- The most accurate scenario is the pessimistic scenario. The analysis of the excess mortality reported in the Supplementary

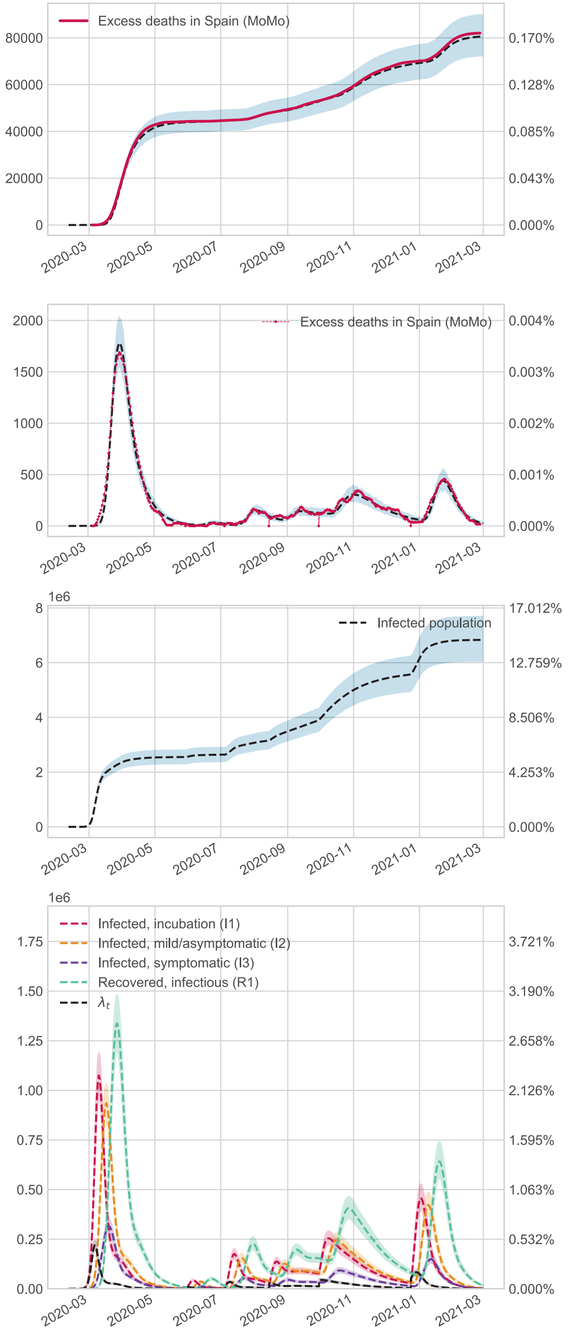

Fig. 6. Results in Spain.

Material justifies why this scenario is more reliable for seroprevalence estimations.

In the overall Spanish analysis, the main findings are specified below:

- In the first wave, the number of infections reaches 5%, while at the end of the year, the total reaches around 10%. These results are in agreement with the national seroprevalence study [60].
- After July, several lockdowns were carried out locally in different areas during specific periods. Our results show that this measure only partially controls the spread of the virus but does not sufficiently reduce the number of cases to reach a low-risk threshold of new infections. For example, active infections decreased after the global and national measures were taken from October 26th until after Christmas, but the epidemiological risk was still high during the period.
- In the summer, the spread of the virus starts again, probably due to imported cases of tourists and high mobility patterns.
- The peak of new infections was less severe in successive waves than at the pandemic's beginning.
- After Christmas, the epidemiological situation was critical. However, thanks to the new measures and the large-scale vaccination strategy, the risk of an explosion of infections was contained.

4.4. Comparison of predictions vs. seroprevalence studies during the first wave

The Spanish government launched an ambitious longitudinal epidemiological study with over 60,000 participants to assess the level of exposure of the population to the virus in at least two temporal points. The preliminary results of this study by regions can be freely downloaded at [52], and the publisher's scientific results are available in the following article [64]. As can be seen, assuming a pessimistic scenario, our model estimates are close to the percentage of the serological survey, although there are some variations in some cases, such as in Madrid. The limitations of the epidemiological study may explain part of these discrepancies. In addition, we must consider the following factors: (i) the methodology followed in the design of the survey and in their statistical analysis; (ii) our model may need a more careful tuning of the parameters according to the study population, or should introduce more reliable sources of information. The local government of Galicia also conducted its study with more than 40,000 participants, and the results are similar [71] to those that appear in a national survey despite the different nature of the study design. Finally, the city council of Torrejón de Ardóz (Madrid), one of the most affected places in Spain, carried out its epidemiological study with more than 100,000 participants, estimating that 22 percent of the population was infected [4].

4.5. Evaluation of the testing capacity

A distinctive feature of our model is that we can retrospectively reconstruct day-to-day infection dynamics with a low cost and non-invasive procedure, unlike other epidemic monitoring strategies. For example, serological surveys measure antibodies at a fixed temporal point, with the inherent limitation that we cannot answer many interesting questions about the dynamics of virus spread. Given the inherent advantage of our approach, an interesting question is to know the real number of infections detected in screening campaigns before starting large-scale vaccination. This is an essential question, as asymptomatic and pre-asymptomatic patients may be the primary transmitters of new infections. Then, this index can assess the quality and limitation of screening protocols and, more specifically, the ability of countries to manage the pandemic.

Fig. 7 shows the percentage of active cases that can be detected by public health agencies on a daily basis. At the beginning of the pandemic, we can observe that the detection capacity is minimal, and their capacity increases in May, with the end of the lockdowns and the reduction in the number of infected people. It is important to note that in the Summer, control is weak, perhaps because it coincides with the holiday period, and national mobility patterns increase, which would explain the increase of epidemiological risk in the coming months. In autumn, control is more reliable and stable. Finally, near Christmas, maximum effectiveness is obtained, partially motivated by upcoming family events, which drive a prior increase in protective measures, such as better testing capacity or stricter control of non-pharmacological measures such as social distance. In any case, in general, the Spanish government's ability to detect daily active infections was less than 10% during the period examined.

5. Discussion

The Coronavirus pandemic is causing an unprecedented crisis in health, economic, and social terms. To help design better policies, estimating the number of people infected by the virus is crucial. We explore the possibility of using statistical simulation models to estimate the spread of an epidemic and Black Block Optimization techniques as a general procedure to obtain model parameters.

The estimations obtained over several regions of Spain have practical relevance. The results reveal that the number of infections was much higher than reported by the Spanish government. Two factors may explain these discrepancies: (i) The tests were carried out in a limited way among the general population, and (ii) no random sampling [84] was done among the different regions throughout the territory, shading the absolute magnitude of the pandemic. It is also worth noting that its impact was uneven across regions. Madrid was the most affected region, while Galicia was the least affected. However, according to our model, in Cataluña, Castilla y León, and the País Vasco, the effects were surprisingly similar, even though the outbreak of infections theoretically started at different time points. In addition, the geographical dispersion of Castilla y León is more significant (see Table 3). Perhaps the older population of Castilla y León (see Table 3) has a higher case fatality rate than the other communities. Therefore, our model is overestimating the number of infected people in that region. As previously mentioned, the demographic structure of the regions, together with the epidemiological profiles and other socioeconomic characteristics, is fundamental in the analysis of results.

A critical concern in epidemic modeling is feeding the models with reliable data, which is challenging due to governments' limited tracking capacities worldwide. In this paper, the mortality records are used as a source of information to increase the reliability of estimations. However, although the information provided by this variable is more accurate than other epidemic monitoring variables as daily infections rates, our approach presents certain disadvantages. In general, fewer deaths are registered than the number of those that happened. Both Spanish and international press echoed this problem in the current Coronavirus crisis. More than twice as many deaths than those officially announced can happen in certain regions, at least at the beginning of the pandemic. A paper released in the early stages of the pandemic reported that the actual number of deaths might be even three times higher in Italy [11].

In order to alleviate this practical limitation, we consider two scenarios. An optimistic scenario in which the official number of COVID-19 caused deaths is correct, and a more extreme and pessimistic scenario, in which we assume that fatality cases are twice as much as those recorded by the government. In addition, in the overall Spanish analysis until 1 March, we use the excess of mortality as a source of information. For this purpose, to calibrate the model better, we retrospectively estimate the IFR parameter and update the β coefficients dynamically.

From a practical point of view, our approach can be essential to assess the current state of the epidemic. In particular, we can know: (i) The degree of immunity of each population; (ii) The number of people who might be infected today; (iii) The total number of recovered patients. In addition, our proposal work can be helpful in a retrospective sense to rebuild the past dynamic of infections and help plan better future decisions based on prior evaluation of the impact of epidemic policies performed in the past. As a relevant example of this modeling strategy, we estimate the capacity detection of active infections on a daily basis, which show that in the summer, the control was poor, which contributed to the substantial expansion of the virus until Christmas.

In other epidemics as influenza, data-driven approaches may be state-of-the-art because there is much evidence about the dynamic of the epidemic in many situations and the response of different agents as public health systems [9]. In this case, we point out that mechanistic models should be preferred at the beginning and in the middle of the pandemic [41]. When more information about the current epidemic and the performance of measures is gathered over time, more purely predictive models can be used. With all this information, medical institutes and governments will be able to guide the elaboration of more personalized policies for normally restoration and establish decisions thresholds to perform critical interventions as local lockdowns according to specific characteristics of each territory.

As a relevant example of our model, we can describe how the peak happened. From an epidemiological point of view, we can identify two different peaks. According to the results reported by our model in the first wave, the first one is the absolute maximum of new infections between 9 and 16 March. The second peak is related to the maximum number of potential contaminators predicted to have happened on March 17 and 24.

In general, assessing the performance of epidemic models is very hard. Practitioners do not know real epidemiological situations and monitor epidemic solutions with surrogate variables that only provide partial knowledge of epidemic dynamics. One of the most critical problems in this pandemic was that many research groups fitted their models with an official daily report of active cases, which led to inconsistent and unrealistic estimations about the number of infections and the population at epidemic risk.

In our particular case, we evaluate the performance of the model during the first wave —the most challenging scenario— against the broad number of seroprevalence studies performed in Spain. In the mentioned analysis, we restrict the expert information introduced in the models to the evidence of the first wave to evaluate model performance in an uncertain setting. In general, our results agree with different epidemiological studies in Spain and several parts of the world (https://covid19-modeling.github.io). In case of discrepancies, these can be explained as a consequence of the lack of appropriate parameter tuning. Also, there are essential discrepancies between the results of national survey studies versus more local ones such as the one carried out in Torrejón de Ardoz [4]. This might be explained by the fact that the effect of the pandemic was worse in Torrejón de Ardoz than in the majority of the population of Madrid. Another critical factor to highlight is the limitations in the statistical survey methodology employed in the national seroprevalence study. Concretely, there are more refined ways to carry on statistical inference modeling than the standard application of the Inverse Probability weight estimator. In this sense, we suggest using the recent methodology described in Ma and Wang [47] to provide more reliable results, in particular in the age groups related to the elderly population. Also, as multiple comparisons are not performed in the confidence interval estimation, coverage uncertainty is somewhat optimistic.

Despite the potential limitations, both in the parameters of our model and in the methodology of the survey, in most territories examined and assuming a pessimistic scenario, the results confirm the accuracy of our approach in the real world. In Supplementary Material, specific details are provided to explain why the pessimistic scenario was more reliable in the first wave estimates when less information on the pandemic was available.

In the more extensive analysis up to 1st March 2021, we use the excess mortality as a specific source of information. In this case, we update the daily infection function dynamically between successive waves, and β was calibrated using our IFR estimation,

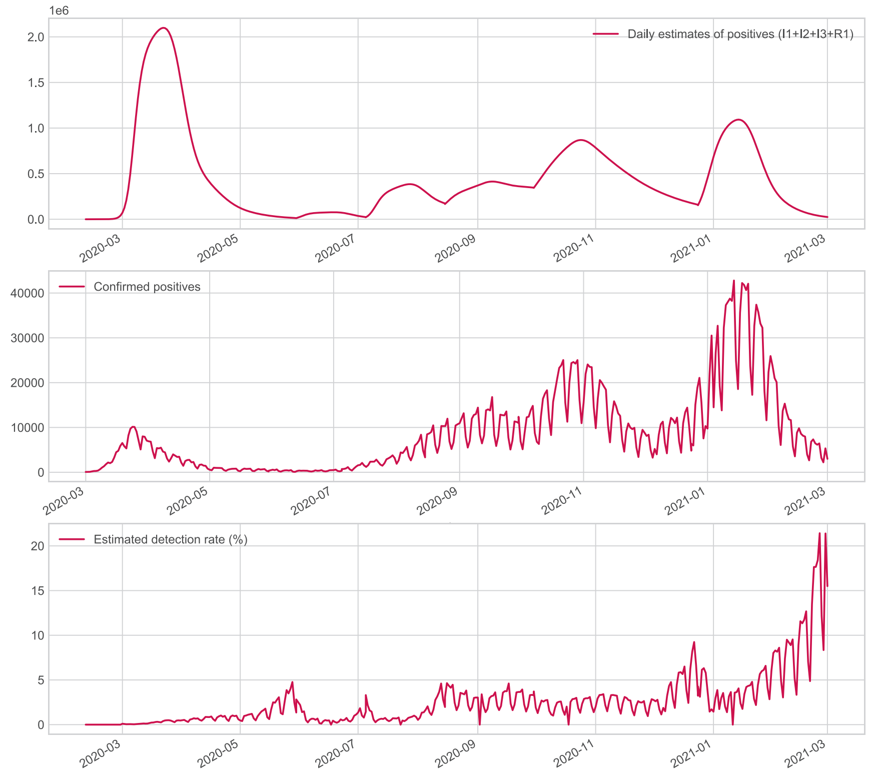

Fig. 7. Dynamic evolution of the number of active infections in the Spanish population (top Figure). Daily cases reported by the Spanish government (middle Figure). The day-to-day percentage of active infections in the Spanish population by health agencies (bottom Figure).

using the results provided by seroprevalence surveys and excess mortality records. In this sense, our results show the limited monitoring capacity of Spanish agencies and the need for more efficient large-scale testing to control the epidemic.

We hope that the recent proposals, such as the use of telecommunications information [63], or more efficient pooled testing strategies [16,54], will allow new ways to control and monitor pandemics in a more reliable way [15,55].

From a methodological perspective, this work aims to explore the idea that complex statistical simulation models, together with Black Box optimization techniques, are a powerful way to define new epidemic models with good empirical performance. Black-box optimization is a promising approach to optimize complex models for which it is not easy to find optimal parameters using traditional methods, such as a likelihood or gradient-based methods, due to the complex and stochastic nature of the model. The use of black-box optimization further increases the flexibility of the approach as it can be adapted to new non-standard complex epidemic formulations from a general perspective. For example, an exciting modification of our model can include non-parametric distributions to model the distribution times between model transitions. Some papers claim that this is a good general approach for estimating other parameters, for example, the virus incubation times [29].

We are unable to make a complete comparative study due to a large number of literature articles on epidemics. However, this paper aims to propose general and flexible modeling strategies to create epidemic models in a simple way, the limit of which depends on computational capabilities.

The epidemic models introduced here can improve in several directions. First, we may obtain more reliable inferences by fitting the model at simultaneous locations, for example, through multilevel simulation models. In this sense, a promising optimization strategy is to employ bayesian optimization. Second, to handle more realistic infection mechanisms, we can use other distributions such as Conway–Maxwell–Poisson [74] to handle over-under dispersion situations. Another exciting direction is introducing social interactions or genetic information about how the virus changes [78,82]. However, nowadays, data related to this are not public in most countries or are very limited or difficult to obtain. Another significant contribution could be the use of several sources of information simultaneously to obtain more accurate estimations. In a recent paper [35], the authors used the results of seroprevalence surveys of the U.S. population to explore this idea. Finally, using machine learning approaches to accelerate parameter fitting is a promising area that may be necessary for more complex simulation models [6]. In preliminary tests with our approach, speed gains are considerable.

As for the new library for simulation that we developed, we fill some gaps in the literature. There are few computational approaches to reconstruct infection dynamics from mortality reports [36,62], and most of the existing models are not dynamic. In this direction, we provide a well-built and reliable python library that allows estimation of seroprevalence, with interpretable and easily comparable parameters across countries. However, obtaining parameters of some prior contributions in different countries can be very challenging [62].

Traditional epidemic models are solved using well-known optimization criteria. Their properties are much better understood, but this epidemic showed the weaknesses of these classical models and the need to model and manage epidemics in an efficient way [69]. We think a general framework such as ours is an exciting and flexible direction towards adapting and designing new models that consider the particular characteristics of each territory without wasting time.

6. Conclusions

This paper has illustrated the potential of combining simulation models with a Black-Box optimization techniques to obtain new epidemic models. Despite our model's simplicity, we have shown that the estimated results are close to those produced by different epidemiological studies. Furthermore, the proposed method can be extended in a simple way to handle more complex situations as graph structures or performing estimations in other epidemics using specific biological expert knowledge, while the same Black-Box optimization strategy remains valid.

Declaration of Competing Interest

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

Acknowledgments

We want to thank the reviewers for their thoughtful comments that undoubtedly helped improve our initial manuscript's quality. The authors also thank Oscar Hernán Madrid Padilla for their helpful suggestions. Finally, we are also grateful to Felipe Muñoz López for his help in the computer implementations at the early stages of this research.

This work has received financial support from the Spanish Ministry of Science, Innovation, and Universities under Grant RTI2018-099646-B-I00, the Consellería de Educación, Universidade e Formación Profesional and the European Regional Development Fund under Grant ED431G-2019/04.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cmpb.2021.106399

References

[1] A. Abdel-Salam, M. Mollazehi, Modeling survival time to recovery from COVID-19: acase study on singapore (2020).
[2] Y. Akimoto, yoshihikoueno, D. Brockhoff, M. Chan, ARF1, CMA-ES/pycma: r3.0.3, 2020, 10.5281/zenodo.3764210
[3] L.J. Allen, Some discrete-time SI, SIR, and SIS epidemic models, Math. Biosci. 124 (1) (1994) 83–105.
[4] Ayuntamiento, Seroprevalence report torrejón de Ardoz, 2020, (https://www.ayto-torrejon.es/noticia/nota-de-prensa/el-estudio-de-seroprevalencia-de-torrejon-de-ardoz-revela-una-prevalencia-de).
[5] F. Ball, C. Larédo, D. Sirl, V.C. Tran, Stochastic Epidemic Models with Inference, 2255, Springer Nature, 2019.
[6] Y. Bengio, A. Lodi, A. Prouvost, Machine learning for combinatorial optimization: A methodological tour d'Horizon, Eur. J. Oper. Res. 290 (0) (2021) 405–421, doi:10.1016/j.ejor.2020.07.063.
[7] Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S.A. Truelove, T. Zhang, et al., Epidemiology and transmission of COVID-19 in Shenzhen China: analysis of 391 cases and 1286 of their close contacts, MedRxiv (2020).
[8] D.H. Brookes, A. Busia, C. Fannjiang, K. Murphy, J. Listgarten, A view of estimation of distribution algorithms through the lens of expectation-maximization, arXiv preprint arXiv:1905.10474(2019).
[9] L.C. Brooks, D.C. Farrow, S. Hyun, R.J. Tibshirani, R. Rosenfeld, Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions, PLOS Comput. Biol. 14 (6) (2018) 1–29, doi:10.1371/journal.pcbi.1006134.
[10] F. Bu, A.E. Aiello, J. Xu, A. Volfovsky, Likelihood-based inference for partially observed epidemics on dynamic networks, J. Am. Stat. Assoc. (2020) 1–17.
[11] P. Buonanno, S. Galletta, M. Puca, News from the COVID-19 epicenter, Available at SSRN 3567093 (2020).
[12] P.R. Bureau, 2020, (https://www.prb.org/countries-with-the-oldest-populations/).
[13] J.H.C.R. Center, 2020, (https://coronavirus.jhu.edu/map.html).
[14] B. Choi, G.A. Rempala, Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling, Biostatistics 13 (1) (2012) 153–165.
[15] S.J. Clark, A.N. Turner, Monitoring epidemics: lessons from measuring population prevalence of the coronavirus, Proc. Natl. Acad. Sci. 118 (9) (2021), doi:10.1073/pnas.2026412118.
[16] S. Comess, H. Wang, S. Holmes, C. Donnat, Statistical modeling for practical pooled testing during the COVID-19 pandemic, arXiv preprint arXiv:2107.05619(2021).
[17] K. Cranmer, J. Brehmer, G. Louppe, The frontier of simulation-based inference, Proc. Natl. Acad. Sci. 117 (48) (2020) 30055–30062.
[18] W. Datadista, 2020, (https://github.com/datadista/datasets/tree/master/COVID2019).
[19] M. Day, COVID-19: four fifths of cases are asymptomatic, China figures indicate, BMJ 369 (2020), doi:10.1136/bmj.m1375.
[20] C. Dudel, T. Riffe, E. Acosta, A.A. van Raalte, M. Myrskyla, Monitoring trends and differences in COVID-19 case fatality rates using decomposition methods: contributions of age structure and age-specific fatality, medRxiv (2020).
[21] L. D'Agostino McGowan, K.H. Grantz, E. Murray, Quantifying uncertainty in mechanistic models of infectious disease, Am. J. Epidemiol. 190 (7) (2021) 1377–1385, doi:10.1093/aje/kwab013.
[22] Ehmann, et al., Virological assessment of hospitalized cases of coronavirus disease 2019 (2019).
[23] M. Enserink, K. Kupferschmidt, Mathematics of life and death: How disease models shape national shutdowns and other pandemic policies, Sci. Mag. 10 (2020).
[24] A.S. Fauci, H.C. Lane, R.R. Redfield, COVID-19 - navigating the uncharted, N. Engl. J. Med. 382 (13) (2020) 1268–1269, doi:10.1056/NEJMe2002387.
[25] J. Fintzi, X. Cui, J. Wakefield, V.N. Minin, Efficient data augmentation for fitting stochastic epidemic models to prevalence data, J. Comput. Graph. Stat. 26 (4) (2017) 918–929.
[26] A. Ghani, C. Donnelly, D. Cox, J. Griffin, C. Fraser, T. Lam, L. Ho, W. Chan, R. Anderson, A. Hedley, et al., Methods for estimating the case fatality ratio for a novel, emerging infectious disease, Am. J. Epidemiol. 162 (5) (2005) 479–486.
[27] J. Goldsmith, S. Greven, C. Crainiceanu, Corrected confidence bands for functional data using principal components, Biometrics 69 (1) (2013) 41–51.
[28] J. Gomez-Gardees, L. Lotero, S.N. Taraskin, F.J. Perez-Reche, Explosive contagion in networks, Sci. Rep. 6 (1) (2016) 19767, doi:10.1038/srep19767.
[29] P. Groeneboom, Estimation of the incubation time distribution for COVID-19, Stat. Neerl. 75 (2021) 161–179, doi:10.1111/stan.12231.
[30] N. Hansen, The CMA evolution strategy: a tutorial, arXiv preprint arXiv:1604.00772(2016).
[31] N. Hansen, A. Auger, R. Ros, S. Finck, P. Povsik, Comparing results of 31 algorithms from the black-box optimization benchmarking BBOB-2009, in: Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation, GECCO 10, Association for Computing Machinery, New York, NY, USA, 2010, pp. 1689–1696, doi:10.1145/1830761.1830790.
[32] N. Hoertel, M. Blachier, C. Blanco, M. Olfson, M. Massetti, M.S. Rico, F. Limosin, H. Leleu, A stochastic agent-based model of the SARS-CoV-2epidemic in France, Nat. Med. 26 (9) (2020) 1417–1421.
[33] A. Huppert, G. Katriel, Mathematical modelling and prediction in infectious disease epidemiology, Clin. Microbiol. Infect. 19 (11) (2013) 999–1005, doi:10.1111/1469-0691.12308.
[34] (INE), T.N.S.I., 2018, (https://www.ine.es/jaxiT3/Datos.htm?t=9963).
[35] N.J. Irons, A.E. Raftery, Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys, Proc. Natl. Acad. Sci. 118 (31) (2021), doi:10.1073/pnas.2103272118.
[36] J. Johndrow, P. Ball, M. Gargiulo, K. Lum, Estimating the number of SARS-CoV-2 infections and the impact of mitigation policies in the united states, Harv. Data Sci. Rev. (2020), doi:10.1162/99608f92.7679a1ed. https://hdsr.mitpress.mit.edu/pub/9421kmzi
[37] M.J. Keeling, K.T.D. Eames, Networks and epidemic models, J. R. Soc., Interface 2 (4) (2005) 295–307, doi:10.1098/rsif.2005.0051.
[38] M.J. Keeling, P. Rohani, Modeling Infectious Diseases in Humans and Animals, Princeton University Press, 2011.
[39] I. Kickbusch, G. Leung, Response to the emerging novel coronavirus outbreak, BMJ 368 (2020), doi:10.1136/bmj.m406.
[40] S.M. Kissler, C. Tedijanto, E. Goldstein, Y.H. Grad, M. Lipsitch, Projecting the transmission dynamics of SARS-CoV-2through the postpandemic period, Science 368 (6493) (2020) 860–868, doi:10.1126/science.abb5793.
[41] J.S. Koopman, J.W. Lynch, Individual causal models and population system models in epidemiology, Am. J. Public Health 89 (8) (1999) 1170–1174, doi:10.2105/ajph.89.8.1170.
[42] T. Kypraios, P.D. O'Neill, et al., Bayesian nonparametrics for stochastic epidemic models, Stat. Sci. 33 (1) (2018) 44–56.
[43] S.A. Lauer, K.H. Grantz, Q. Bi, F.K. Jones, Q. Zheng, H.R. Meredith, A.S. Azman, N.G. Reich, J. Lessler, The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application, Ann. Intern. Med. (2020) 577–582.
[44] S.P. Layne, J.M. Hyman, D.M. Morens, J.K. Taubenberger, New coronavirus outbreak: framing questions for pandemic prevention, Sci. Transl. Med. 12 (534) (2020), doi:10.1126/scitranslmed.abb1469.
[45] J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, J. Am. Stat. Assoc. 113 (523) (2018) 1094–1111.
[46] L. Lei, E.J. Candès, Conformal inference of counterfactuals and individual treatment effects, arXiv preprint arXiv:2006.06138(2020).
[47] X. Ma, J. Wang, Robust inference using inverse probability weighting, J. Am. Stat. Assoc. 0 (0) (2019) 1–10, doi:10.1080/01621459.2019.1660173.
[48] E. Mahase, COVID-19: death rate is 0.66% and increases with age, study estimates, BMJ 369 (2020) m1327, doi:10.1136/bmj.m1327.
[49] S. Mandal, R.R. Sarkar, S. Sinha, Mathematical models of malaria-a review, Malar. J. 10 (1) (2011) 202.

[50] M. Matabuena, O.H.M. Padilla, F.-J. Gonzalez-Barcala, Statistical and mathematical modeling in the coronavirus epidemic: some considerations to minimize biases in the results, Arch. Bronconeumol. (2020).

[51] M. McKee, D. Stuckler, If the world fails to protect the economy, COVID-19will damage health not just now but also in the future, Nat. Med. (2020), doi:10.1038/s41591-020-0863-y.

[52] MISAN, Estudio nacional de sero-epidemiología de la infección por SARS-CoV-2 en España, 2020, (https://www.mscbs.gob.es/ciudadanos/ene-covid/home.htm).

[53] K. Mizumoto, K. Kagaya, A. Zarebski, G. Chowell, Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the diamond princess cruise ship, Yokohama, Japan, 2020, Eurosurveillance 25 (10) (2020), doi:10.2807/1560-7917.ES.2020.25.10.2000180.

[54] L. Mutesa, P. Ndishimye, Y. Butera, J. Souopgui, A. Uwineza, R. Rutayisire, E.L. Ndoricimpaye, E. Musoni, N. Rujeni, T. Nyatanyi, et al., A pooled testing strategy for identifying SARS-CoV-2at low prevalence, Nature 589 (7841) (2021) 276–280.

[55] M.I. Nelson, Tracking the UK SARS-CoV-2 outbreak, Science 371 (6530) (2021) 680–681, doi:10.1126/science.abg2297.

[56] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A.R. Akhmetzhanov, N.M. Linton, Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19), medRxiv (2020). 10.1101/2020.02.03.20020248

[57] C.P.E.R.E. Novel, et al., The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China, Zhonghua Liu Xing Bing Xue Za Zhi 41 (2) (2020) 145.

[58] M. O'Driscoll, G. Ribeiro Dos Santos, L. Wang, D.A.T. Cummings, A.S. Azman, J. Paireau, A. Fontanet, S. Cauchemez, H. Salje, Age-specific mortality and immunity patterns of SARS-CoV-2, Nature 590 (7844) (2021) 140–145, doi:10.1038/s41586-020-2918-0.

[59] D. Osthus, K.S. Hickmann, P.C. Caragea, D. Higdon, S.Y. Del Valle, Forecasting seasonal influenza with a state-space SIR model, Ann. Appl. Stat. 11 (1) (2017) 202.

[60] R. Pastor-Barriuso, B. Pérez-Gómez, M.A. Hernán, M. Pérez-Olmeda, R. Yotti, J. Oteo-Iglesias, J.L. Sanmartín, I. León-Gómez, A. Fernández-García, P. Fernández-Navarro, I. Cruz, M. Martín, C. Delgado-Sanz, N. Fernández de Larrea, J. León Paniagua, J.F. Muñoz-Montalvo, F. Blanco, A. Larrauri, M. Pollán, Infection fatality risk for SARS-CoV-2 in community dwelling population of spain: nationwide seroepidemiological study, BMJ 371 (2020), doi:10.1136/bmj.m4509.

[61] T.A. Perkins, S.M. Cavany, S.M. Moore, R.J. Oidtman, A. Lerch, M. Poterek, Estimating unobserved SARS-CoV-2 infections in the united states, Proc. Natl. Acad. Sci. 117 (36) (2020) 22597–22602.

[62] T.A. Perkins, S.M. Cavany, S.M. Moore, R.J. Oidtman, A. Lerch, M. Poterek, Estimating unobserved SARS-CoV-2 infections in the united states, Proc. Natl. Acad. Sci. 117 (36) (2020) 22597–22602, doi:10.1073/pnas.2005476117.

[63] J. Persson, J.F. Parie, S. Feuerriegel, Monitoring the COVID-19 epidemic with nationwide telecommunication data, Proc. Natl. Acad. Sci. 118 (26) (2021), doi:10.1073/pnas.2100664118.

[64] M. Pollán, B. Pérez-Gómez, R. Pastor-Barriuso, J. Oteo, M.A. Hernán, M. Pérez-Olmeda, J.L. Sanmartín, A. Fernández-García, I. Cruz, N.F. de Larrea, et al., Prevalence of SARS-CoV-2 in spain (ENE-COVID): a nationwide, population-based seroepidemiological study, Lancet (2020) 535–544.

[65] D.D. Rajgor, M.H. Lee, S. Archuleta, N. Bagdasarian, S.C. Quek, The many estimates of the COVID-19 case fatality rate, Lancet Infect. Dis. (2020) 776–777.

[66] L.M. Rios, N.V. Sahinidis, Derivative-free optimization: a review of algorithms and comparison of software implementations, J. Global Optim. 56 (3) (2013) 1247–1293.

[67] C. Rubén Fernández Casal, 2020, (https://github.com/rubenfcasal/COVID-19).

[68] H. Salje, C.T. Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hoze, J. Richet, C.-L. Dubost, et al., Estimating the burden of SARS-CoV-2 in France 368(6493) 860-868 (2020).

[69] A. Saltelli, G. Bammer, I. Bruno, E. Charters, M. Di Fiore, E. Didier, W.N. Espeland, J. Kay, S.L. Piano, D. Mayo, et al., Five ways to ensure that models serve society: a manifesto, 2020,

[70] J. Satsuma, R. Willox, A. Ramani, B. Grammaticos, A. Carstea, Extending the SIR epidemic model, Phys. A 336 (3–4) (2004) 369–375.

[71] SERGAS, Estudo de seroprevalencia fronte a covid, 2020, (https://saladecomunicacion.sergas.gal/Paginas/DetalleNova.aspx?idioma=es&idNova=10430&vista=clasica).

[72] G. Shafer, V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. 9 (3) (2008) 371–421.

[73] M. Shen, J. Zu, C.K. Fairley, J.A. Pagán, B. Ferket, B. Liu, S.Y. Stella, E. Chambers, G. Li, Y. Guo, et al., Effects of New York's executive order on face mask use on COVID-19 infections and mortality: a modeling study, J. Urban Health 98 (2) (2021) 197–204.

[74] G. Shmueli, T.P. Minka, J.B. Kadane, S. Borle, P. Boatwright, A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution, J. R. Stat. Soc. 54 (1) (2005) 127–142, doi:10.1111/j.1467-9876.2005.00474.x.

[75] A. Shoukat, S.M. Moghadas, Agent-based modelling: an overview with application to disease dynamics, arXiv preprint arXiv:2007.04192(2020).

[76] C. Staerk, T. Wistuba, A. Mayr, Estimating effective infection fatality rates during the course of the COVID-19 pandemic in Germany, BMC Public Health 21 (1) (2021) 1073, doi:10.1186/s12889-021-11127-7.

[77] S. Tabata, K. Imai, S. Kawano, M. Ikeda, T. Kodama, K. Miyoshi, H. Obinata, S. Mimura, T. Kodera, M. Kitagaki, M. Sato, S. Suzuki, T. Ito, Y. Uwabe, K. Tamura, Non-severe vs. severe symptomatic COVID-19: 104 cases from the outbreak on the cruise ship "diamond princess" in Japan, medRxiv (2020). 10.1101/2020.03.18.20038125

[78] M. Tang, Fitting Stochastic Epidemic Models to Multiple Data Types, 2019 Ph.D. thesis.

[79] S. Venkatramanan, B. Lewis, J. Chen, D. Higdon, A. Vullikanti, M. Marathe, Using data-driven agent-based models for forecasting emerging infectious diseases, Epidemics 22 (2018) 43–49.

[80] R. Verity, L.C. Okell, et al., Estimates of the severity of coronavirus disease 2019: a model-based analysis, Lancet Infect. Dis. (2020), doi:10.1016/S1473-3099(20)30243-7.

[81] R. Verity, et al., Estimates of the severity of coronavirus disease 2019: a model-based analysis, Lancet Infect. Dis. (2020), doi:10.1016/S1473-3099(20)30243-7.

[82] C. Viboud, A. Vespignani, The future of influenza forecasts, Proc. Natl. Acad. Sci. 116 (8) (2019) 2802–2804, doi:10.1073/pnas.1822167116.

[83] J.T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P.M. de Salazar, B.J. Cowling, M. Lipsitch, G.M. Leung, Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China, Nat. Med. (2020), doi:10.1038/s41591-020-0822-7.

[84] C.T. Yiannoutsos, P.K. Halverson, N. Menachemi, Bayesian estimation of SARS-CoV-2 prevalence in Indiana by random testing, Proc. Natl. Acad. Sci. 118 (5) (2021), doi:10.1073/pnas.2013906118.

[85] X. Zhao, B. Zhang, P. Li, C. Ma, J. Gu, P. Hou, Z. Guo, H. Wu, Y. Bai, Incidence, clinical characteristics and prognostic factor of patients with COVID-19: a systematic review and meta-analysis, medRxiv (2020). 10.1101/2020.03.17.20037572