Tracing DAY-ZERO and Forecasting the Fade out of the COVID-19 Outbreak in Lombardy, Italy: A Compartmental Modelling and Numerical Optimization Approach

Lucia Russo

Consiglio Nazionale delle Ricerche, Institute of Science and Technology for Energy and Sustainable Mobility, Napoli, Italy

Cleo Anastassopoulou

Department of Microbiology, Medical School, University of Athens, Athens, Greece

Athanasios Tsakris

Department of Microbiology, Medical School, University of Athens, Athens, Greece

Gennaro Nicola Bifulco

Dipartimento di Ingegneria Civile, Edile e Ambientale, Università degli Studi di Napoli Federico II, Napoli, Italy

Emilio Fortunato Campana

Consiglio Nazionale delle Ricerche, Dipartimento di Ingegneria, ICT e Tecnologie per l'Energia e i Trasporti, Roma, Italy

Gerardo Toraldo

Dipartimento di Matematica e Applicazioni "Renato Caccioppoli", Università degli Studi di Napoli Federico II, Napoli, Italy

Constantinos Siettos[1]

Dipartimento di Matematica e Applicazioni "Renato Caccioppoli", Università degli Studi di Napoli Federico II, Napoli, Italy

Abstract

Background. Since the first suspected cluster of cases of coronavirus disease-2019 (COVID-19) on December 1st, 2019, in Wuhan, Hubei Province,

[1]Corresponding author: constantinos.siettos@unina.it

China, a total of 195,892 confirmed infected cases, 80,840 recovered and 7,865 deaths have been reported worldwide up to March 16, 2020. After China, Italy is currently at the forefront of the combat against the epidemic that has now spread to all 22 Italian regions. The disease is sweeping through Lombardy, which remains in lockdown since the 8th of March. As of the same day, the isolation measures taken in Lombardy have been extended to the entire country. On March 11, the WHO declared COVID-19 pandemic. Here, we provide estimates for: (a) the DAY-ZERO of the outbreak in Lombardy, Italy; (b) the actual number of exposed/infected cases in the total population; (c) the basic reproduction number (R_0); (d) the "effective" per-day disease transmission and mortality rates; and, importantly, (e) a forecast for the fade out of the outbreak, on the basis of the released data of confirmed cases for Lombardy from February 21 to March 8, the day of lockdown.

Methods. To deal with the uncertainty in the number of actual exposed/ infected cases in the total population, we address a new compartmental Susceptible/ Exposed/ Infectious/ Recovered/ Dead (SEIRD) model with two compartments of infectious persons: one modelling the total cases in the population and another modelling the confirmed cases. The parameters of the model corresponding to the recovery period, and the time from exposure to the time that an individual starts to be infectious, have been set as reported from clinical studies on COVID-19. For the estimation of the DAY-ZERO of the outbreak in Lombardy, as well as of the "effective" per-day transmission and mortality rates for which no clinical data are available, we have used the SEIRD simulator to fit the data from February 21 to the 8th of March, the lockdown day of Lombardy (and of all Italy). This was accomplished by solving a mixed-integer optimization problem with the aid of genetic algorithms. Based on the computed values, we also provide an estimation of the basic reproduction number R_0. Furthermore, by reducing the estimated transmission rate by 90% on March 8 (to reflect the lockdown of almost all activities), we run the simulator from March 8 to forecast the fade out of the outbreak.

Findings. Based on the proposed methodological procedure, we estimated

that the actual cumulative number of exposed cases in the total population in Lombardy on March 8 was of the order of 15 times the confirmed cumulative number of infected cases. According to this scenario, the DAY-ZERO for the outbreak in Lombardy was the 21st of January 2020. The "effective" per-day disease transmission rate for the period until March 8 was found to be 0.779 (90% CI: 0.777-0.781), while the "effective" per-day mortality rate was found to be 0.0173 (90% CI: 0.0154-0.0192). Based on these values, the basic reproduction rate R_0 was found to be 4.04 (90% CI: 4.03-4.05).

Importantly, by reducing the transmission rate by 90% on March 8 to reflect the suspension of almost all activities in Italy, we run the simulator to forecast the fade out of the epidemic. Simulations show that if the measures continue, the complete fade out of the outbreak in Lombardy is expected to occur by the end of May 2020.

Introduction

The butterfly effect in chaos theory underscores the sensitive dependence on initial conditions, highlighting the importance of even a small change in the state of a nonlinear system. The emergence of a novel coronavirus, SARS-CoV-2, that caused a viral pneumonia outbreak in Wuhan, Hubei province, China in early December 2019 has evolved into the COVID-19 acute respiratory disease pandemic due to its alarming levels of spread and severity, with a total of 195,892 confirmed infected cases, 80,840 recovered and 7,865 deaths in 153 countries as of March 16, 2020 ([1]). The seemingly far from the epicenter, old continent became the second-most impacted region after Asia Pacific to date, mostly as a result of a dramatic divergence of the epidemic trajectory in Italy, where there have been 31,506 total confirmed infected cases, 2,941 recovered and 2,503 deaths as of March 17, 2020 ([1]).

The second largest outbreak outside of mainland China officially started on January 31, 2020, after two Chinese visitors staying at a central hotel in Rome tested positive for SARS-CoV-2; the couple remained in isolation and

was declared recovered on February 26 [2]. A 38-year-old man repatriated back to Italy from Wuhan who was admitted to the hospital in Codogno, Lombardy on February 21 was the first secondary infection case ("patient 1"). "Patient 0" was never identified by tracing the first Italian citizen's movements and contacts. In less than a week, the explosive increase in the number of cases in several bordering regions and autonomous provinces of northern Italy placed enormous strain on the decentralized health system. Following an a dramatic spike in deaths from COVID-19, Italy transformed into a "red zone", and the movement restrictions were expanded to the entire country on the 8th of March. All public gatherings were cancelled and school and university closures were extended through at least the next month.

In an attempt to assess the dynamics of the outbreak for forecasting purposes as well as to estimate epidemiological parameters that cannot be computed directly based on clinical data, such as the transmission rate of the disease and the basic reproduction number R_0, defined as the expected number of exposed cases generated by one infected case in a population where all individuals are susceptible, many mathematical modelling studies have already appeared since the first confirmed COVID-19 case. The first models mainly focused on the estimation of the basic reproduction number R_0 using dynamic mechanistic mathematical models ([3, 4, 5, 6]), but also simple exponential growth models (see e.g. [7, 8]). Compartmental epidemiological models like SIR, SIRD, SEIR and SEIRD have been proposed to estimate other important epidemiological parameters, such as the transmission rate and for forecasting purposes (see e.g. [6, 9]). Other studies have used metapopulation models, which include data of human mobility between cities and/or regions to forecast the evolution of the outbreak in other regions/countries far from the original epicenter in China [3, 10, 11, 5], including the modelling of the influence of travel restrictions and other control measures in reducing the spread ([12].

Among the perplexing problems that mathematical models face when they are used to estimate epidemiological parameters and to forecast the evolution of the outbreak, two stand out: (a) the uncertainty that characterizes the actual

number of infected cases in the total population, which is mainly due to the large percentage of asymptomatic or mild cases experiencing the disease like the common cold or the flu (see e.g. [13]), and (b) the uncertainty regarding the DAY-ZERO of the outbreak, the knowledge of which is crucial to assess the stage and dynamics of the epidemic, especially during the first growth period.

To cope with the above problems, we herein propose a novel SEIRD with two compartments, one modelling the total infected cases in the population and another modelling the confirmed cases. The proposed modelling approach is applied to Lombardy, the epicenter of the outbreak in Italy, to estimate the scale of under-reporting of the number of actual cases in the total population, the DAY-ZERO of the outbreak and for forecasting purposes. The above tasks were accomplished by the numerical solution of a mixed-integer optimization problem using the publicly available data of cumulative cases for the period February 21-March 8, the day of lockdown of all of Italy.

Methodology

The modelling approach

We address a compartmental SEIRD model that includes two categories of infected cases, namely the confirmed/reported and the unreported (unknown) cases in the total population. Based on observations and studies, our modelling hypothesis is that the confirmed cases of infected are only a (small) subset of the actual number of infected cases in the total population [5, 13, 6]. Regarding the confirmed cases of infected as of February 11, a study conducted by the Chinese CDC which was based on a total of 72,314 cases in China, about 80.9% of the cases were mild and could recover at home, 13.8% severe and 4.7% critical [14].

On the basis of the above findings, in our modelling approach, the unreported cases were considered either asymptomatic or mildly symptomatic cases that recover from the disease relatively soon and without medical care, while the confirmed cases include all the above types, but on average their recovery lasts

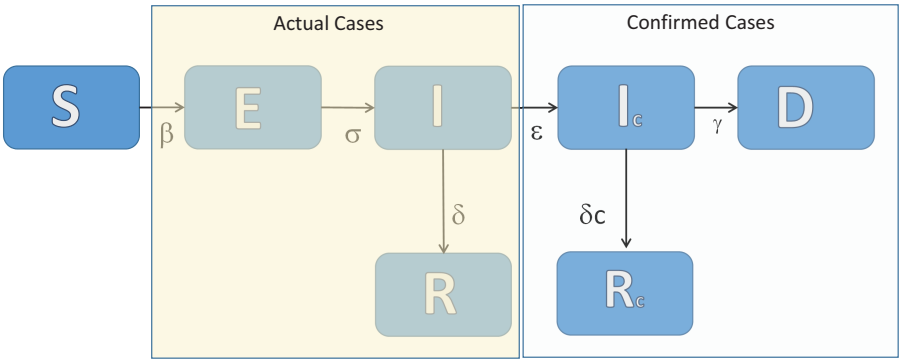

Figure 1: A schematic of the proposed compartmental SEIRD model. The actual number of cases are unknown.

longer than the non-confirmed, they may also be hospitalized and die from the disease.

Based on the above, let us consider a well-mixed population of size N. The state of the system at time t, is described by (see also Figure 1 for a schematic) $S(t)$ representing the number of susceptible persons, $E(t)$ the number of exposed, $I(t)$ the number of unreported infected persons in the total population who are asymptomatic or experience mild symptoms and recover relatively soon without any other complications, $I_c(t)$ the number of confirmed infected cases who may develop more severe symptoms and a part of them dies, $R(t)$ the number of recovered persons in the total population, $R_c(t)$ the number of confirmed recovered cases and $D(t)$ the number of deaths. For our analysis, and for such a short period, we assume that the total number of the population remains constant. Based on demographic data, the total population of Lombardy is $N = 10m$; its surface area is 23,863.09 kmq and the population density is $sim422$ (Inhabitants/Kmq).

The rate at which a susceptible (S) becomes exposed (E) to the virus is proportional to the density of infectious persons I in the total population, excluding the number of dead persons D. Our main assumption here is that upon confirmation, the infected persons I_c go into quarantine, and, thus, they don't

transmit further the disease. The proportionality constant is the "effective" disease transmission rate, say $\beta = \bar{c}p$, where $\bar{c}$ is the average number of contacts per day and p is the probability of infection upon a contact between a susceptible and an infected.

Thus, our discrete mean field compartmental SEIRD model reads:

$$S(t) = S(t-1) - \frac{\beta}{N - D(t)} S(t-1)I(t-1) \tag{1}$$

$$E(t) = E(t-1) + \frac{\beta}{N - D(t)} S(t-1)I(t-1) - \sigma E(t-1) \tag{2}$$

$$I(t) = I(t-1) + \sigma E(t-1) - \delta I(t-1) - \epsilon I(t-1) \tag{3}$$

$$I_c(t) = I_c(t-1) + \epsilon I(t-1) - \delta_c I_c(t-1) - \gamma I_c(t-1) \tag{4}$$

$$R(t) = R(t-1) + \delta I(t-1) \tag{5}$$

$$R_c(t) = R_c(t-1) + \delta_c I(t-1) \tag{6}$$

$$D(t) = D(t-1) + \gamma I_c(t-1) \tag{7}$$

The above system is defined in discrete time points $t = 1, 2, \ldots$, with the corresponding initial condition at the very start of the outbreak (DAY-ZERO): $S(0) = N - 1$, $I(0) = 1$, $E(0) = 0$ $I_c(0) = 0$, $R(0) = 0$, $R_c(0) = 0$, $D(0) = 0$.

The parameters of the model are:

- $\beta(d^{-1})$ is the "effective" transmission rate of the disease,
- $\sigma(d^{-1})$ is the average per-day "effective" rate at which an exposed person becomes infective,
- $\delta(d^{-1})$ is the average per-day "effective" recovery rate within the group of unreported (asymptomatic/mild) cases in the total population,
- $\delta_c(d^{-1})$ is the average per-day "effective" recovery rate within the subset of confirmed infected cases
- $\gamma(d^{-1})$ is the average per-day "effective" mortality rate within the subset of confirmed infected cases,

- $\epsilon(d^{-1})$ is the per-day rate of the all cases of infected in the total population that get confirmed. This proportionality rate quantifies the uncertainty in the actual number of unreported cases in the total population.

Here, we should note the following: As new cases of recovered and dead at each time t appear with a time delay (which is generally unknown but an estimate can be obtained by clinical studies) with respect to the corresponding infected cases, the above per-day rates are not the actual ones; thus, they are denoted as "effective/apparent" rates.

The values of the epidemiological parameters σ, δ, δ_c that were fixed in the proposed model were chosen based on clinical studies.

In particular, in many studies that use SEIRD models, the parameter σ is set equal to the inverse of the mean incubation period (time from exposure to the development of symptoms) of a virus. *However, the incubation period does not generally coincide with the time from exposure to the time that someone starts to be infectious.* Regarding COVID-19, it has been suggested that an exposed person can be infectious well before the development of symptoms [15]. With respect to the incubation period for SARS-CoV-2, a study in China [16] suggests that it may range from 2–14 days, with a median of 5.2 days. Another study in China, using data from 1,099 patients with laboratory-confirmed 2019-nCoV ARD from 552 hospitals in 31 provinces/provincial municipalities suggested that the median incubation period is 4 days (interquartile range, 2 to 7). In our model, as explained above, $\frac{1}{\sigma}$ represents the period from exposure to the onset of the contagious period. Thus, based on the above clinical studies, for our simulations, we have set $\frac{1}{\sigma} = 3$.

Regarding the recovery period, the WHO-China Joint Mission in a study that is based on 55,924 laboratory-confirmed cases has reported a median time of 2 weeks from onset to clinical recovery for mild cases, and 3-6 weeks for severe or critical cases [17]. Based on the above and on the fact that within the subset of confirmed cases the mild cases are the 81% [14], we have set the recovery period for the confirmed cases' compartment to be $\delta_c = 1/21$ in order to balance

the recovery period with the corresponding characterization of the cases (mild, severe/critical). The average recovery period of the unreported/non-confirmed part of the infected population, which in our assumptions experiences the disease like the flu or a common cold, is set equal to one week [18], i.e. we have set $\delta = 1/7$.

Note, that per-day mortality rate in the model does not coincide with the case fatality (mortality) ratio (CFR) which is usually reported and computed as the ratio between the reported deaths and the cumulative number of infected. Furthermore, the transmission rate cannot be obtained by clinical studies, but only by mathematical models.

Finally, regarding DAY-ZERO in Lombardy, what has been reported is just the date on which the first infected person was confirmed to be positive for SARS-CoV-2. That day was February 21, 2020, which is the starting date of public data release of confirmed cases.

Estimation of the DAY-ZERO of the outbreak, the scale of data uncertainty, the disease transmission and mortality rates

In order to provide a coarse estimation of the scale of under-reporting of the number of actual cases in the total population, we have considered five indicative values of the corresponding variable ϵ (0.01, 0.05, 0.1, 0.15, 0.2). Thus, for each one of the above values of ϵ, the DAY-ZERO of the outbreak, the per-day "effective" transmission rate β and the "effective" per-day mortality rate γ, were computed by the numerical solution of a mixed-integer optimization problem with the aid of genetic algorithms to fit the reported data of confirmed cumulative cases from February 21 to March 8, the day of the lockdown of Lombardy.

Here, for our computations, we have used the genetic algorithm "ga" provided by the Global Optimization Toolbox of Matlab [4] to minimize the following objective function:

$$f(t_0,\alpha,\gamma) = \underset{t_0,\alpha,\gamma}{\operatorname{argmin}}\{ \sum_{t=t_0}^{February29} (w_1 f_t((t_0,\alpha,\gamma|\beta,\beta_c,\delta,\delta_c,\epsilon)^2 + w_2 g_t((t_0,\alpha,\gamma|\beta,\beta_c,\delta,\delta_c,\epsilon)^2 + w_3 h_t((t_0,\alpha,\gamma|\beta,\beta_c,\delta,\delta_c,\epsilon)^2)\}, \quad (8)$$

where

$$f_t(t_0,\alpha,\gamma|\beta,\beta_c,\delta,\delta_c,\epsilon) = \sum \Delta I^{SEIRD}(t) - \sum \Delta I(t),$$
$$g_t(t_0,\alpha,\gamma|\beta,\beta_c,\delta,\delta_c,\epsilon) = \sum \Delta R^{SEIRD}(t) - \sum \Delta R(t),$$
$$h_t(t_0,\alpha,\gamma|\beta,\beta_c,\delta,\delta_c,\epsilon) = \sum \Delta D^{SEIRD}(t) - \sum \Delta D(t) \quad (9)$$

where, $\sum \Delta X^{SEIRD}(t)$, $(X = I, R, D)$ are the cumulative cases resulting from the SEIRD simulator at time t; w_1, w_2, w_3 correspond to scalars serving in the general case as weights to the relevant functions.

In order to get the 90% confidence intervals for β and γ (as these are not provided by the genetic algorithm), we fixed the DAY-ZERO for the simulations and run the Levenberg-Marquard around the optimal solution as implemented by the "lsqnonlin" function of matlab [19].

Thus, for each one of the five values of *epsilon*, we have repeated the above numerical optimization procedure fifty times and we kept the best fitting outcome.

At this point we should note that the above optimization problem may in principle have more than one near-optimal solutions, which may be attributed to the fact that the tuning of both DAY-ZERO and the transmission rate may in essence result in nearby values of the objective function. As a consequence, starting from different sets of optimal values for DAY-ZERO and the transmission rate (the closer DAY-ZERO is to the day of the initial report of the data, the larger the transmission rate will be and vice versa), simulations may result in numbers of cases that differ significantly.

To validate the model as resulting for the estimated level of scaling, we used it to forecast the confirmed reported cases from March 9 to March 16, 2020.

Since the complete lockdown of almost all activities was decided as of March 8, we have taken an 90% reduction in the corresponding "effective" transmission rate to reflect the drop in the per-day average contacts per person.

Finally, we also attempt a forecasting of the fade out of the outbreak.

For the optimization procedure, we set as initial guesses (the intervals within which the optimal estimates were sought are also given in parentheses): for the DAY-ZERO (t_0) the 16th of January (1st January-15th of February), β=0.5 (0.1-0.95), $\gamma = 0.01$ (0.001-0.05).

Estimation of the basic reproduction number R_0 from the SEIRD model

Initially, when the spread of the epidemic starts, all the population is considered to be susceptible, i.e. $S \approx N$. On the basis of this assumption, we computed the basic reproduction number based on the estimates of the epidemiological parameters computed using the data from the 21st of February to the 8th of March with the aid of the SEIRD model given by Eq.(1)-(7) as follows.

Note that there are three infected compartments, namely E, I, I_c that determine the outbreak. Thus, considering the corresponding equations given by Eq.(2),(3),(4), and that at the very first days of the epidemic $S \approx N$ and $D \approx 0$, the Jacobian of the system as evaluated at the disease-free state reads:

$$\boldsymbol{J} = \frac{\partial(E(t), I(t), I_c(t))}{\partial(E(t-1), I(t-1), I_c(t-1))} = \begin{bmatrix} 1-\sigma & \beta & 0 \\ \sigma & 1-(\delta+\epsilon) & 0 \\ 0 & -\epsilon & 1-(\delta_c+\gamma) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} -\sigma & \beta & 0 \\ \sigma & -(\delta+\epsilon) & 0 \\ 0 & \epsilon & -(\delta_c+\gamma) \end{bmatrix} \tag{10}$$

The eigenvalues (that is the roots of the characteristic polynomial of the Jacobian matrix) dictate if the disease-free equilibrium is stable or not, that is

if an emerging infectious disease can spread in the population. In particular, the disease-free state is stable, meaning that an infectious disease will not result in an outbreak, if and only if all the norms of the eigenvalues of the Jacobian $\boldsymbol{J}$ of the discrete time system are bounded by one. Jury's stability criterion [20] (the analogue of Routh-Hurwitz criterion for discrete-time systems) can be used to determine the stability of the linearized discrete time system by analysis of the coefficients of its characteristic polynomial. The characteristic polynomial of the Jacobian matrix reads:

$$F(z) = a_3 z^3 + a_2 z^2 + a_1 z + a_0 \tag{11}$$

where

$$\begin{aligned}
a_3 &= 1 \\
a_2 &= \delta + \delta_c + \epsilon + \gamma + \sigma - 3 \\
a_1 &= \delta\delta_c - 2\delta_c - 2\epsilon - 2\gamma - 2\sigma - 2\delta + \delta_c\epsilon + \delta\gamma + \epsilon\gamma - \beta\sigma + \delta\sigma + \delta_c\sigma + \epsilon\sigma + \gamma\sigma + 3 \\
a_0 &= \delta + \delta_c + \epsilon + \gamma + \sigma - \delta\delta_c - \delta_c\epsilon - \delta\gamma - \epsilon\gamma + \beta\sigma - \delta\sigma - \delta_c\sigma - \epsilon\sigma - \gamma\sigma - \beta\delta_c\sigma + \\
&\quad \delta\delta_c\sigma - \beta\gamma\sigma + \delta_c\epsilon\sigma + \delta\gamma\sigma + \epsilon\gamma\sigma - 1
\end{aligned} \tag{12}$$

The necessary conditions for stability read:

$$F(1) > 0 \tag{13}$$

$$(-1)^3 F(-1) > 0 \tag{14}$$

The sufficient conditions for stability are given by the following two inequalities:

$$|a_0| < a_3 \tag{15}$$

$$|b_0| > |b_2|, \tag{16}$$

where,

$$b_0 = \begin{vmatrix} a_0 & a_3 \\ a_3 & a_0 \end{vmatrix}, b_2 = \begin{vmatrix} a_0 & a_1 \\ a_3 & a_2 \end{vmatrix} \tag{17}$$

It can be shown that the second necessary condition (14) and the first sufficient condition (15) are always satisfied for the range of values of the epidemiological parameters considered here.

The first inequality (13) results in the necessary condition:

$$\frac{\beta}{\delta + \epsilon} < 1 \tag{18}$$

It can be also shown that for the range of the parameters considered here, the second sufficient condition (16) is satisfied if the necessary condition (18) is satisfied. Thus, the necessary condition (18) is also a sufficient condition for stability. Hence, the disease-free state is stable, if and only if, condition (18) is satisfied.

Note that in this necessary and sufficient condition (18), the first term in the parenthesis, i.e. $\frac{1}{(\delta+\epsilon)}$ is the average infection time of the compartment I Thus, the above expression reflects the basic reproduction number R_0 which is qualitatively defined by $R_0 = \beta \frac{1}{infection\ time}$. Hence, our model results in the following expression for the basic reproduction number:

$$R_0 = \frac{\beta}{\delta + \epsilon} \tag{19}$$

Note that for $\epsilon = 0$, the above expression simplifies to R_0 for the simple SIR model.

Results

As discussed in the Methodology, we used five different values of ϵ (0.01, 0.05, 0.1, 0.15, 0.2) to assess the actual number of cases in the total population. Thus, for our computations, we run 50 times the numerical optimization procedure and for further analysis we kept the value of ϵ that gave the smaller fitting error

over all runs. In particular, by using the genetic algorithm as described in the Methodology, the best fitting to the reported data was obtained with ϵ=0.05 (best fitting residual norm of the objective function: 301,231). For this scale of under-reporting, we present the results for DAY-ZERO, the "effective" per-day transmission and mortality rates as computed with the reported data from February 21 to March 8. The solution of the mixed-integer optimization problem for the DAY-ZERO and the "effective" per-day transmission and mortality rates of the compartmental SEIRD model resulted in the following values (with w_1=1, w_2=4, w_3=16 to balance for the different scales of the number of infected vs. the number of recovered and dead).

For the computed "effective" per-day transmission and mortality rates, we also report the corresponding 90% confidence intervals instead of the more standard 95% CI because of the small size of the data. Under this scenario, the DAY-ZERO for the outbreak in Lombardy was found to be the 21st of January. The "effective" per-day transmission rate was found to be $\beta = 0.779$ (90% CI: 0.777-0.781) and the "effective" per-day mortality rate for the confirmed cases was found to be $\gamma = 0.0173$ (90% CI: 0.0154-0.0192). Based on the derived value of the "effective" per-day disease transmission rate, the basic reproduction number was found to be R_0= 4.04 (90% CI: 4.03-4.05).

Thus, based on the computed expression for R_0, the "effective" per-day transmission rate should have been below ∼0.19, implying that the average contacts per person should have dropped by at least ∼75% for the outbreak to fade out at the initial stage.

Using these estimated values for the epidemiological parameters, we ran the simulator from DAY-ZERO (21st of January) to March 8. On March 8, simulations resulted in the following numbers for the cumulative cases: $\sum \Delta E(t) = 66,016$, $\sum \Delta I(t) = 37,691$, $\sum \Delta I_c(t) = 4,232$, $R(t) = 12,123$, $R_c(t) = 623$, $D(t) = 220$. The reported cumulative numbers for the day of lockdown were $\sum \Delta I_c(t) = 4,189$, $R_c = 550$, $\sum D(t) = 267$. Figures (2),(3),(4) depict the simulation results based on the optimal estimates, starting from the 21st of January to the 8th of March. As shown, the predictions of the model are quite close to

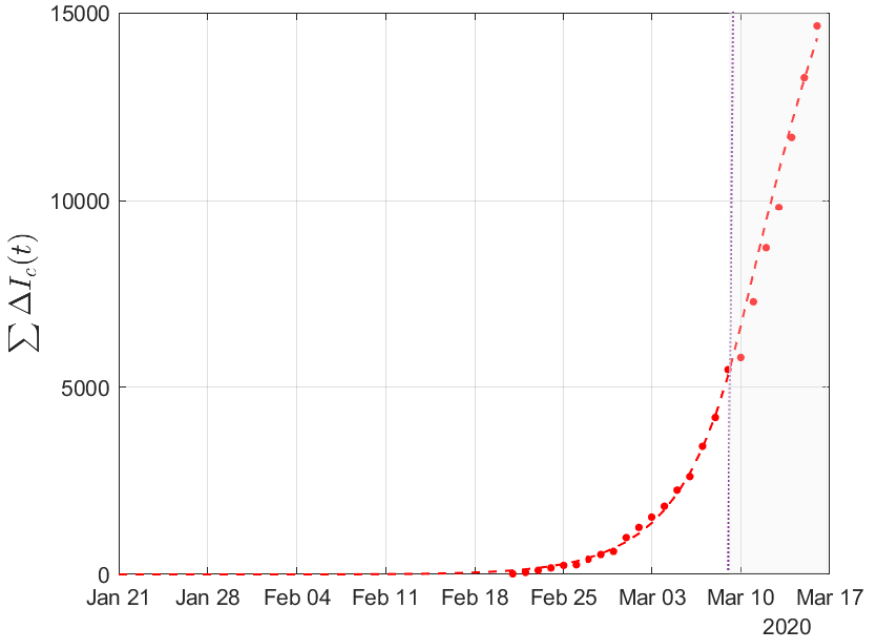

Figure 2: Cumulative number of (confirmed) infected cases resulting from simulations from DAY-ZERO (January 21) until the 16th of March. The DAY-ZERO, $\beta = 0.779$, $\gamma = 0.0173$ were computed by solving the mixed-integer optimization problem for the period DAY-ZERO to March 8. The validation of the model was performed using the reported data of confirmed cases from March 9 to March 16 (shaded area) by taking a 90% reduction in the β due to the lockdown of March 8. Dots correspond to the reported data of confirmed cases.

the reported number of confirmed cases for that period.

Thus, according to the above results, on the 8th of March, the actual cumulative number of infected cases in the total population (taking into account the exposed cases to the virus) was of the order of 15 times more the confirmed cumulative number of infected cases.

To validate the model with respect to the reported data of confirmed cases from March 9 to March 16, we have set as $\beta = 0.078$, (i.e. taking a 90% reduction in the value of β found with the optimization procedure) and as initial conditions the values resulting from the simulation on March 8. Based on the above, the model resulted in the following numbers for the cumulative cases for March 16:

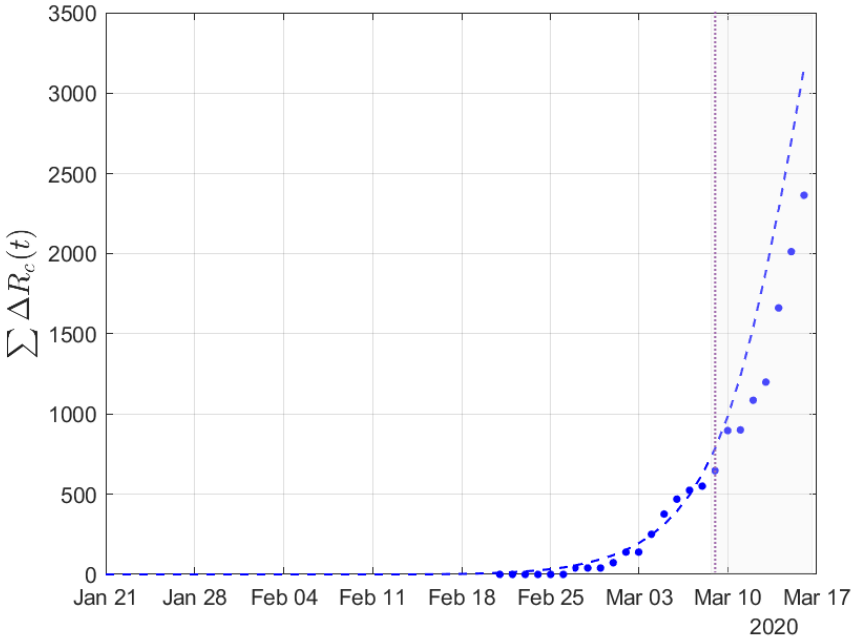

Figure 3: Cumulative number of (confirmed) recovered cases resulting from simulations from DAY-ZERO (January 21) until the 16th of March. The DAY-ZERO, $\beta = 0.779$, $\gamma = 0.0173$ were computed by solving the mixed-integer optimization problem for the period DAY-ZERO to March 8. The validation of the model was performed using the reported data of confirmed cases from March 9 to March 16 (shaded area) by taking a 90% reduction in the β due to the lockdown of March 8. Dots correspond to the reported data of confirmed cases.

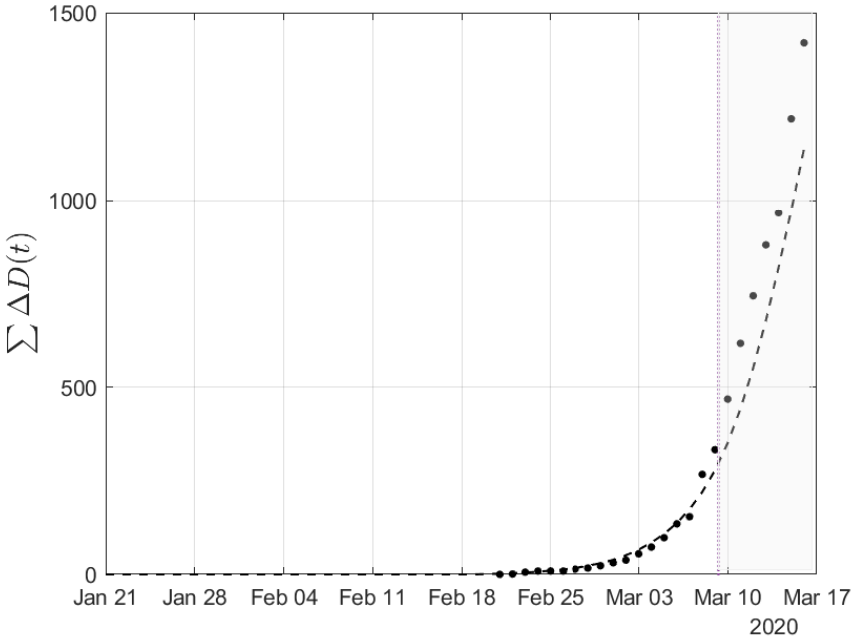

Figure 4: Cumulative number of (confirmed) deaths resulting from simulations from DAY-ZERO (January 21) until the 16th of March. The DAY-ZERO, $\beta = 0.779$, $\gamma = 0.0173$ were computed by solving the mixed-integer optimization problem for the period DAY-ZERO to March 8. The validation of the model was performed using the reported data of confirmed cases from March 9 to March 16 (shaded area) by taking a 90% reduction in the β due to the lockdown of March 8. Dots correspond to the reported data of confirmed cases.

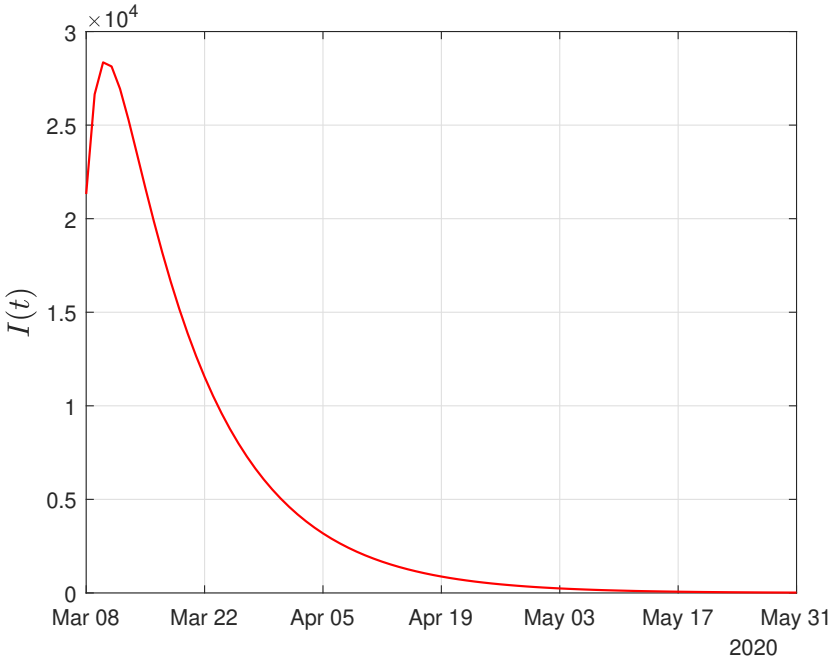

Figure 5: Estimated number of infected cases in the total population at each day resulting from simulations from March 8 to May 31 (the day of the lockdown of all Italy) using $\beta = 0.779 - 90\%0.779$, $\gamma = 0.0173$ as computed by solving the mixed-integer optimization problem for the period DAY-ZERO (21st January) to March 8.

$\sum \Delta E(t) = 81,619$, $\sum \Delta I(t) = 75,149$, $\sum \Delta I_c(t) = 14,317$, $R(t) = 40,946$, $R_c(t) = 3,158$, $D(t) = 1,139$ (the reported confirmed cumulative numbers on March 16 were $\sum \Delta I_c(t) = 14,649$, $R_c = 2,368$, $D(t) = 1,420$). Thus, the model predicted fairly well the period from March 9 to March 16.

As discussed in the Methodology, we also attempted to forecast the evolution of the outbreak based on our analysis. To do so, we have considered a 90% reduction in the effective transmission rates starting on March 8, the day of lockdown not only of Lombardy, but of all Italy. The result of our forecast is depicted in Figure 5. As predicted by simulations, if the strict isolation measures continue to hold, the outbreak in Lombardy will fade out by the end of May, 2020.

Discussion

The crucial questions about an outbreak is how, when (DAY-ZERO), why it started and when it will end. Answers to these important questions would add critical knowledge in our arsenal to combat the pandemic. The tracing of DAY-ZERO, in particular, is of outmost importance. It is well known that minor perturbations in the initial conditions of a complex system, such as the ones of an outbreak, may result in major changes in the observed dynamics. No doubt, a high level of uncertainty for DAY-ZERO, as well as the uncertainty in the actual numbers of exposed people in the total population, raise several barriers in our ability to correctly assess the state and dynamics of the outbreak, and to forecast its evolution and its end. Such pieces of information would lower the barriers and help public health authorities respond fast and efficiently to the emergency.

This study aimed exactly at shedding more light into this problem, taking advantage of state-of-the-art tools of mathematical modelling and numerical analysis/optimization tools. To achieve this goal, we addressed a new compartmental SEIRD with two infectious compartments in order to bridge the gap between the number of reported cases and the actual number of cases in the total population. By following the proposed methodological framework, we found that the DAY-ZERO in Lombardy was the 21th of January, a date that precedes by one month the fate of the first confirmed case in Lombardy. Furthermore, our analysis revealed that the actual cumulative number of infected cases in the total population is around 15 times more the cumulative number of confirmed infected cases. Importantly, based on our simulations, we predict that the fade-out of the outbreak in Lombardy will be by the end of May, if the strict isolation measures continue to hold.

To this end, we would like to make a final comment with respect to the basic reproduction number R_0, the significance and meaning of which are very often misinterpreted and misused, thereby leading to erroneous conclusions. Here, we found an $R_0 \sim 4$, which is similar to the values reported by many studies in

China. For example, Zhao et al. estimated R_0 to range between 2.24 (95% CI: 1.96-2.55) and 3.58 (95% CI: 2.89-4.39) in the early phase of the outbreak [7]. Similar estimates, were obtained for R_0 by Imai et al. 2.6 (95% CI: 1.5-3.5) [4], Li et al. [21], Wu et al. 2.68 (95% CI: 2.47–2.86), as well as by Anastassopoulou et al. recently 3.1 (90% CI: 2.5-3.7) [6].

However, we would like to stress that R_0 is NOT a biological constant for a disease as it is affected not only by the pathogen, but also by many other factors, such as environmental conditions, the demographics as well as, importantly, by the social behavior of the population (see for example the discussion in [22]). Thus, a value for R_0 that is found in a part of world (and even in a region of the same country) cannot be generalized as a global biological constant for other parts of world (or even for other regions of the same country). Obviously, the environmental factors and social behavior of the population in Lombardy are different from the ones, for example, prevailing in Hubei.

We hope that the results of our analysis help to mitigate some of the severe consequences of the currently uncontrolled pandemic.

Funding

We did not receive any specific funding for this study.

Data Availability

The data used in this paper are given in the Supporting information.

References

[1] W. H. Organization, Coronavirus disease 2019 (COVID-19). Situation report 51 (2020).
URL https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_10

[2] F. Carinci, Covid-19: preparedness, decentralisation, and the hunt for patient zero, BMJ 368. arXiv:https://www.bmj.com/content/368/bmj.m799.full.pdf, doi:10.1136/bmj.m799.
URL https://www.bmj.com/content/368/bmj.m799

[3] J. T. Wu, K. Leung, G. M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in wuhan, china: a modelling study, The Lancetdoi:10.1016/s0140-6736(20)30260-9.
URL https://doi.org/10.1016%2Fs0140-6736%2820%2930260-9

[4] N. Imai, A. Cori, I. Dorigatti, et al., Report 3: Transmissibility of 2019-ncov, Int J Infect Disdoi:10.1016/j.ijid.2020.01.050.
URL https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-2019-nCoV-transmissibility.pdf

[5] D. Li, J. Lv, G. Botwin, J. Braun, W. Cao, L. Li, D. P. McGovern, Estimating the scale of covid-19 epidemic in the united states: Simulations based on air traffic directly from wuhan, china, medRxivarXiv:https://www.medrxiv.org/content/early/2020/03/08/2020.03.06.20031880.full.pdf, doi:10.1101/2020.03.06.20031880.
URL https://www.medrxiv.org/content/early/2020/03/08/2020.03.06.20031880

[6] C. Anastassopoulou, L. Russo, A. Tsakris, C. Siettos, Data-based analysis, modelling and forecasting of the covid-19 outbreak, medRxivarXiv:https://www.medrxiv.org/content/early/2020/03/12/2020.02.11.20022186.full.pdf, doi:10.1101/2020.02.11.20022186.
URL https://www.medrxiv.org/content/early/2020/03/12/2020.02.11.20022186

[7] S. Zhao, Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He, M. H. Wang, Preliminary estimation of the basic repro-

duction number of novel coronavirus (2019-nCoV) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak, Int J Infect Disdoi:10.1101/2020.01.23.916395.
URL https://doi.org/10.1101%2F2020.01.23.916395

[8] A. Remuzzi, G. Remuzzi, COVID-19 and italy: what next?, The Lancetdoi:10.1016/s0140-6736(20)30627-9.
URL https://doi.org/10.1016%2Fs0140-6736%2820%2930627-9

[9] W.-K. Ming, J. Huang, C. J. P. Zhang, Breaking down of the healthcare system: Mathematical modelling for controlling the novel coronavirus (2019-nCoV) outbreak in wuhan, chinadoi:10.1101/2020.01.27.922443.
URL https://doi.org/10.1101%2F2020.01.27.922443

[10] H.-Y. Yuan, M. P. Hossain, M. M. Tsegaye, X. Zhu, P. Jia, T.-H. Wen, D. Pfeiffer, Estimating the risk on outbreak spreading of 2019-ncov in china using transportation data, medRxivarXiv:https://www.medrxiv.org/content/early/2020/02/04/2020.02.01.20019984.full.pdf, doi:10.1101/2020.02.01.20019984.
URL https://www.medrxiv.org/content/early/2020/02/04/2020.02.01.20019984

[11] P. M. De Salazar, R. Niehus, A. Taylor, C. O. Buckee, M. Lipsitch, Using predicted imports of 2019-ncov cases to determine locations that may not be identifying all imported cases, medRxivarXiv:https://www.medrxiv.org/content/early/2020/02/11/2020.02.04.20020495.full.pdf, doi:10.1101/2020.02.04.20020495.
URL https://www.medrxiv.org/content/early/2020/02/11/2020.02.04.20020495

[12] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. P. y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, A. Vespignani, The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak, Science

(2020) eaba9757doi:10.1126/science.aba9757.
URL https://doi.org/10.1126%2Fscience.aba9757

[13] C. William Feuer, Current US coronavirus cases are "just the tip of the iceberg," former USAID director says. (2020).
URL https://www.cnbc.com/2020/03/05/us-coronavirus-cases-just-the-tip-of-the-iceberg-ex-usaid-director.html

[14] T. N. C. P. E. R. E. Team, The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) in china, 2020, China CDC Weekly 2 (2020) 113.
URL http://weekly.chinacdc.cn//article/id/e53946e2-c6c4-41e9-9a9b-fea8db1a8f51

[15] C. for Disease Control, Prevention, How COVID-19 Spreads (2020).
URL https://www.cdc.gov/coronavirus/2019-ncov/about/transmission.html

[16] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, Z. Feng, Early transmission dynamics in wuhan, china, of novel coronavirus infected pneumonia, New England Journal of Medicine 0 (0) (0) null. arXiv:https://doi.org/10.1056/NEJMoa2001316, doi:10.1056/NEJMoa2001316.
URL https://doi.org/10.1056/NEJMoa2001316

[17] W. H. Organization, Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) (2020).
URL https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf

[18] t. . F. Fernando Duarte, BBC. [link].
URL https://www.cdc.gov/flu/symptoms/symptoms.htm

[19] The Mathworks, Inc., Natick, Massachusetts, MATLAB R2018b (2018).

[20] E. I. Jury, L. Stark, V. V. Krishnan, Inners and stability of dynamic systems, IEEE Transactions on Systems, Man, and Cybernetics SMC-6 (10) (1976) 724–725. doi:10.1109/tsmc.1976.4309436.
URL https://doi.org/10.1109%2Ftsmc.1976.4309436

[21] Q. Li, X. Guan, P. Wu, et al., Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus Infected Pneumonia (2020). doi:10.1056/NEJMoa2001316.
URL https://doi.org/10.1088%2F0951-7715%2F16%2F2%2F308

[22] P. L. Delamater, E. J. Street, T. F. Leslie, Y. T. Yang, K. H. Jacobsen, Complexity of the basic reproduction number (r0), Emerging Infectious Diseases 25 (1) (2019) 1–4. doi:10.3201/eid2501.171901.
URL https://doi.org/10.3201%2Feid2501.171901

Supporting information

All the relevant data used in this paper are publicy available and accessible at `https://lab.gedidigital.it/gedi-visual/2020/coronavirus-i-contagi-in-italia/`. In Table S1 are given the reported cumulative numbers from Febrary 21 to March 16. The data from February 21 to March 8 have been used for the calibration of the model parameters and the data from March 9 to March 16 have been used for the validation of the model.

S1 Table..

Table 1: Reported cumulative numbers of cases for the Hubei region, China for the period January 11-February 10

Date	Infected	Deaths	Recovered
Feb 21	15	0	0
22	54	0	1
23	1101	0	6
24	172	0	9
25	240	0	9
26	258	0	9
27	403	40	14
28	531	40	17
29	615	40	23
March 01	984	73	31
02	1254	139	38
03	1529	139	55
04	1820	250	73
05	2251	376	98
06	2612	469	135
07	3420	524	154
08	4189	550	267
09	5469	646	333
10	5791	896	468
11	7280	900	617
12	8725	1085	744
13	9820	1198	880
14	11685	1660	966
15	13272	2011	1218
16	14649	2368	1420