DATA-BASED ANALYSIS, MODELLING AND FORECASTING OF THE NOVEL CORONAVIRUS (2019-NCOV) OUTBREAK

A PREPRINT

Cleo Anastassopoulou*
Department of Microbiology, Medical School
University of Athens
Athens, Greece
cleoa@med.uoa.gr

Lucia Russo
Consiglio Nazionale delle Ricerche
Science and Technology for Energy and Sustainable Mobility
Naples, Italy
lucia.russo@cnr.it

Athanasios Tsakris
Department of Microbiology, Medical School
University of Athens
Athens, Greece
atsakris@med.uoa.gr

Constantinos Siettos†
Dipartimento di Matematica e Applizazioni
Università degli Studi di Napoli Federico II
Napoli, Italy
constantinos.siettos@unina.it

February 23, 2020

ABSTRACT

Since the first suspected case of coronavirus diesease-2019 (COVID-19) on December 1st, 2019, in Wuhan, Hubei Province, China, a total of 40,235 confirmed cases and 909 deaths have been reported in China up to February 10, 2020, evoking fear locally and internationally. Here, based on the publicly available epidemiological data (WHO, CDC, ECDC, NHC and DXY) for Hubei from January 11 to February 10, 2020, we provide estimates of the main epidemiological parameters, i.e. the basic reproduction number (R_0) and the infection, recovery and mortality rates, along with their 90% confidence intervals. As the number of infected individuals, especially of those with asymptomatic or mild courses, is suspected to be much higher than the official numbers, which can be considered only as a sample of the actual numbers of infected and recovered cases in the population, we have repeated the calculations under a second scenario that considers twenty times the number of confirmed infected cases and forty times the number of recovered, leaving the number of deaths unchanged. Our computations and analysis were based on a mean field Susceptible-Infected-Recovered-Dead (SIRD) model. Thus, based on the SIRD simulations, the estimated average value of R_0 in both scenarios was found to be 2.5, the one computed considering the period from the 11th of January until the 18th of January, using the official counts of the confirmed cases was found to be 4.6, while the one computed under the second scneario was found to be 3.2. Furthermore, on the estimated parameters from both scenarios, we provide tentative three-week forecasts of the evolution of the outbreak at the epicenter. Our forecasting flashes a note of caution for the presently unfolding outbreak in China. Based on the official counts for confirmed cases, the simulations suggest that the cumulative number of infected will surpass 80,000 (as a lower bound) and could reach 160,000 (with an upper bound of 330,000) by February 29. Regarding the number of deaths, simulations forecast that on the basis of the up to the 10th of February data and estimations of the actual number of infected and recovered in the population, the death toll might exceed 7,000 by February 29. However, our analysis further reveals a significant decline of the mortality rate to which various factors may have contributed, such as the severe control measures taken in Hubei, China (e.g. quarantine and hospitalization of infected individuals), but mainly it is due to the fact that the actual number of infected and recovered cases

*Corresponding author
†Corresponding author

in the population is estimated to be more than twenty times higher than reported, thus resulting in a much lower mortality rate, which according to our computations is of the order of 0.15%.

Keywords Coronavirus · COVID-19 · SARS-CoV-2 · Mathematical Modeling · Data Analysis · Forecasting

1 Introduction

An outbreak of "pneumonia of unknown etiology" in Wuhan, Hubei Province, China in early December 2019 has spiraled into an epidemic that is ravaging China and threatening to reach a pandemic state [1]. The causative agent soon proved to be a new betacoronavirus related to the Middle East Respiratory Syndrome virus (MERS-CoV) and the Severe Acute Respiratory Syndrome virus (SARS-CoV). The novel coronavirus SARS-CoV-2 disease has been named "COVID-19" by the World Health Organization (WHO) and on January 30, the COVID-19 outbreak was declared to constitute a Public Health Emergency of International Concern by the WHO Director-General [2]. Despite the lockdown of Wuhan and the suspension of all public transport, flights and trains on January 23, a total of 40,235 confirmed cases, including 6,484 (16.1%) with severe illness, and 909 deaths (2.2%) had been reported in China by the National Health Commission up to February 10, 2020; meanwhile, 319 cases and one death were reported outside of China, in 24 countries [3]. The origin of COVID-19 has not yet been determined although preliminary investigations are suggestive of a zoonotic, possibly of bat, origin [4, 5]. Similarly to SARS-CoV and MERS-CoV, the novel virus is transmitted from person to person principally by respiratory droplets, causing such symptoms as fever, cough, and shortness of breath after a period believed to range from 2 to 14 days following infection, according to the Centers for Disease Control and Prevention (CDC) [1, 6, 7]. Preliminary data suggest that older males with comorbidities may be at higher risk for severe illness from COVID-19 [6, 8, 9]. However, the precise virologic and epidemiologic characteristics, including transmissibility and mortality, of this third zoonotic human coronavirus are still unknown. Using the serial intervals (SI) of the two other well-known coronavirus diseases, MERS and SARS, as approximations for the true unknown SI, Zhao et al. estimated the mean basic reproduction number (R_0) of SARS-CoV-2 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 [10]. Very similar estimates, 2.2 (95% CI: 1.4-3.9), were obtained for R_0 at the early stages of the epidemic by Imai et al. 2.6 (95% CI: 1.5-3.5) [11], as well as by Li et al., who also reported a doubling in size every 7.4 days [1]. Wu et al. estimated the R0 at 2.68 (95% CI: 2.47–2.86) with a doubling time every 6.4 days (95% CI: 5.8–7.1) and the epidemic growing exponentially in multiple major Chinese cities with a lag time behind the Wuhan outbreak of about 1–2 weeks [12]. Amidst such an important ongoing public health crisis that also has severe economic repercussions, we reverted to mathematical modelling that can shed light to essential epidemiologic parameters that determine the fate of the epidemic [13]. Here, we present the results of the analysis of time series of epidemiological data available in the public domain (WHO, CDC, ECDC, NHC and DXY) from January 11 to February 10, 2020, and attempt a three-week forecast of the spreading dynamics of the emerged coronavirus epidemic in the epicenter in mainland China.

2 Methodology

Our analysis was based on the publicly available data of the new confirmed daily cases reported for the Hubei province from the 11th of January until the 10th of February. Based on the released data, we attempted to estimate the mean values of the main epidemiological parameters, i.e. the basic reproduction number R_0, the infection (α), recovery (β) and mortality rate (γ), along with their 90% confidence intervals. However, as suggested [14], the number of infected, and consequently the number of recovered, people is likely to be much higher. Thus, in a second scenario, we have also derived results by taking twenty times the number of reported cases for the infected and forty times the number for the recovered cases, while keeping constant the number of deaths that is more likely to be closer to the real number. Based on these estimates we also provide tentative forecasts until the 29th of February. Our methodology follows a two-stage approach as described below.

The basic reproduction number is one of the key values that can predict whether the infectious disease will spread into a population or die out. R_0 represents the average number of secondary cases that result from the introduction of a single infectious case in a totally susceptible population during the infectiousness period. Based on the reported data of confirmed cases, we provide estimations of the R_0 from the 16th up to the 20th of January in order to satisfy as much as possible the hypothesis of $S \approx N$ that is a necessary condition for the computation of R_0.

We also provide estimations of β, γ over the entire period using a rolling window of one day from the 11th of January to the 16th of January to provide the very first estimations.

Furthermore, we exploited the SIRD model to provide an estimation of the infection rate α. This is accomplished by starting with one infected person on the 16th of November, which has been suggested as a starting date of the epidemic [6], run the SIR model until the 10th of February and optimize α to fit the reported confirmed cases from the 11th of January to the 10th of February. Below, we describe analytically our approach.

Let us start by denoting with $S(t)$, $I(t)$, $R(t)$, $D(t)$, the number of susceptible, infected, recovered and dead persons respectively at time t in the population of size N. For our analysis we assume that the total number of the population remains constant. Based on the demographic data for the province of Hubei $N = 59m$. Thus, the discrete SIRD model reads:

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

$$I(t) = I(t-1) + \frac{\alpha}{N} S(t-1)I(t-1) - \beta I(t-1) - \gamma I(t-1) \tag{2}$$

$$R(t) = R(t-1) + \beta I(t-1) \tag{3}$$

$$D(t) = D(t-1) + \gamma I(t-1) \tag{4}$$

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

Initially, when the spread of the epidemic starts, all the population is considered to be susceptible, i.e. $S \approx N$. Based on this assumption, by Eq.(2),(3),(4), the basic reproduction number can be estimated by the parameters of the SIRD model as:

$$R_0 = \frac{\alpha}{\beta + \gamma} \tag{5}$$

A problem with the approximation of the epidemiological parameters α, β and γ and thus R_0, from real-world data based on the above expressions is that in general, for large scale epidemics, the actual number of infected $I(t)$ persons in the whole population is unknown. Thus, the problem is characterized by high uncertainty. However, one can attempt to provide some coarse estimations of the epidemiological parameters based on the reported confirmed cases using the approach described next.

2.1 Identification of epidemiological parameters from the reported data of confirmed cases

Lets us denote with $\Delta I(t) = I(t) - I(t-1)$, $\Delta R(t) = R(t) - R(t-1)$, $\Delta D(t) = D(t) - D(t-1)$, the reported new cases of infected, recovered and deaths at time t, with $C\Delta I(t), C\Delta R(t), C\Delta D(t)$ the cumulative numbers of confirmed cases at time t. Thus:

$$\mathbf{C\Delta X}(t) = \sum_{i=1}^{t} \Delta X(t), \tag{6}$$

where, $X = I, R, D$.

Let us also denote by $\mathbf{CX}(t) = [CX(1), CX(2), \cdots, CX(t)]^T$, the $t \times 1$ column vector containing the corresponding cumulative numbers up to time t. On the basis of Eqs.(2), (3), (4), one can provide a coarse estimation of the parameters R_0, β and γ as follows.

Starting with the estimation of R_0, we note that as the province of Hubei has a population of 59m, one can reasonably assume that for any practical means, at least at the beginning of the outbreak, $S \approx N$. By making this assumption, one can then provide an approximation of the expected value of R_0 using Eq.(5) and Eq.(2), Eq.(3), Eq.(4). In particular, substituting in Eq.(2), the terms $\beta I(t-1)$ and $\gamma I(t-1)$ with $\Delta R(t) = R(t) - R(t-1)$ from Eq.(3), and $\Delta D(t) = D(t) - D(t-1)$ from Eq.(4) and bringing them into the left-hand side of Eq.(2), we get:

$$I(t) - I(t-1) + R(t) - R(t-1) + D(t) - D(t-1) = \frac{\alpha}{N} S(t-1)I(t-1) \tag{7}$$

Adding Eq.(3) and Eq.(4), we get:

$$R(t) - R(t-1) + D(t) - D(t-1) = \beta I(t-1) + \gamma I(t-1) \tag{8}$$

Finally, assuming that for any practical means at the beginning of the spread that $S(t-1) \approx N$ and dividing Eq.(7) by Eq.(8) we get:

$$\frac{I(t) - I(t-1) + R(t) - R(t-1) + D(t) - D(t-1)}{R(t) - R(t-1) + D(t) - D(t-1)} = \frac{\alpha}{\beta + \gamma} = R_0 \tag{9}$$

Note that one can use directly Eq.(9) to compute R_0 with regression, without the need to compute first the other parameters, i.e. β, γ and α. At this point, the regression can be done either by using the differences per se, or by using the corresponding cumulative functions (see that by summing up both sides of Eq.(7) and Eq.(8) over time one can use the cumulative functions instead of the differences for the calculation of R_0 using Eq.(9)). Here, we have chosen the second option as a standard practice to reduce the noise included in the differences.
Thus, based on the above, a coarse estimation of R_0 and its corresponding confidence intervals can be provided by solving a linear regression problem using least-squares problem as:

$$\hat{R_0} = (\mathbf{C\Delta R} + \mathbf{C\Delta D})^{'}(t)(\mathbf{C\Delta R} + \mathbf{C\Delta D})(t))^{-1}(\mathbf{C\Delta R} + \mathbf{C\Delta D})^{'}(t)(\mathbf{C\Delta I} + \mathbf{C\Delta R} + \mathbf{C\Delta D})(t), \tag{10}$$

The prime $'$ is for the transpose operation. The estimation of the mortality and recovery rates was based on the simple formula used by the National Health Commission (NHC) of the People's Republic of China [15] (see also [16]), that is the ratio of the cumulative number of recovered/deaths and that of infected at time t. Thus, a coarse/simplistic estimation of the mortality can be calculated solving a linear regression problem for the corresponding cumulative function by least squares as:

$$\hat{\gamma} = (\mathbf{C\Delta I}^{'}(t)\mathbf{C\Delta I}(t))^{-1}\mathbf{C\Delta I}^{'}(t)(\mathbf{C\Delta D})(t), \tag{11}$$

In a similar manner, the recovery rate can be computed by:

$$\hat{\beta} = (\mathbf{C\Delta I}^{'}(t)\mathbf{C\Delta I}(t))^{-1}\mathbf{C\Delta I}^{'}(t)(\mathbf{C\Delta R})(t), \tag{12}$$

As the reported data are just a sample of the actual number of infected and recovered cases including the asymptomatic and/or mild ones, we have repeated the above calculations by taking twenty times the reported number of infected and forty times the reported number of recovered in the population, while leaving the reported number of deaths the same considering that their cataloguing is close to the actual number of deaths due to COVID-19.

2.2 Estimation of the infection rate from the SIRD model

Having estimated the expected values of the parameters β and γ, an approximation of the infected rate α, that is not biased by the assumption of $S = N$ can be obtained by using the SIRD simulator. In particular, in the SIRD model we set $\hat{\beta}$ and $\hat{\gamma}$, and set as initial conditions one infected person on the 16th of November and run the simulator until the last date for which there are available data (here up to the 10th of February). Then, the value of the infection rate α can be found by "wrapping" around the SIRD simulator an optimization algorithm (such as a nonlinear least-squares solver) to solve the problem:

$$\underset{\alpha}{\operatorname{argmin}}\{\sum_{t=1}^{M}(w_1 f_t(\alpha)^2 + w_2 g_t(\alpha)^2 + w_3 h_t(\alpha)^2)\}, \tag{13}$$

where

$$\begin{aligned} f_t(\alpha) &= C\Delta I^{SIRD}(t) - C\Delta I(t), \\ g_t(\alpha) &= R_t^{SIRD} - C\Delta R(t), \\ h_t(\alpha) &= D_t^{SIRD} - C\Delta D(t) \end{aligned}$$

where, Q_k^{SIRD} ($Q = I, R, D$) are the differences between the cumulative new cases of the infected, recovered and dead people resulting from the SIRD simulator and the reported (or, in the second scenario the five times the reported infected and recovered) real-world cumulative cases at time t; w_1, w_2, w_3 correspond to scalars serving in the general case as weights to the relevant functions. For the solution of the above optimization problem we used the function "lsqnonlin" of matlab [17] using the Levenberg-Marquard algorithm.

3 Results

As discussed, we have derived results using two different scenarios (see in Methodology). For each scenario, we first present the results obtained by solving the least squares problem as described in section 2.1 using a rolling window of an one-day step. The first time window was that from the 11th up to the 16th of January i.e. we used the first six days to provide the very first estimations of the epidemiological parameters. We then proceeded with the calculations by adding one day in the rolling window as described in the methodology until the 10th of February. We also report the corresponding 90% confidence intervals instead of the more standard 95% because of the small size of the data. For each window, we also report, the corresponding coefficients of determination (R^2) representing the proportion of the variance in the dependent variable that is predictable from the independent variables, and the root mean square of error (RMSE). The estimation of R_0 was based on the data until January 20, in order to satisfy as much as possible the hypothesis underlying its calculation by Eq.(9).
Then, we used the SIRD model to provide an estimation of infection rate by "wrapping" around the SIRD simulator the optimization algorithm as described in section 2.2. Finally, we provide tentative forecasts for the evolution of the outbreak based on both scenarios until the end of February.

3.1 Scenario I: Results obtained using the exact numbers of the reported confirmed cases

Figure1 depicts an estimation of R_0 for the period January 16-January 20. Using the first six days from the 11th of January, $\hat{R_0}$ results in $\sim$ 4.80 (90% CI: 3.36-6.67); using the data until January 17, $\hat{R_0}$ results in $\sim$ 4.60 (90% CI: 3.56-5.65); using the data until January 18, $\hat{R_0}$ results in $\sim$ 5.14 (90%CI: 4.25-6.03); using the data until January 19, $\hat{R_0}$ results in $\sim$ 6.09 (90% CI: 5.02-7.16); and using the data until January 20, $\hat{R_0}$ results in $\sim$ 7.09 (90% CI: 5.84-8.35)

Figure2 depicts the estimated values of the recovery (β) and mortality (γ) rates for the period January 16 to February 10. The confidence intervals are also depicted with dashed lines. Note that the large variation in the estimated values of β and γ should be accounted to the small size of the data and data uncertainty. This is also reflected in the corresponding confidence intervals. As more data are taken into account, this variation is significantly reduced. Thus, using all the available data from the 11th of January until the 10th of February, the estimated value of the mortality rate γ is $\sim$ 2.94% (90% CI: 2.89%-3.0%) and that of the recovery rate is $\sim$ 0.05 (90% CI: 0.045-0.055) corresponding to $\sim$ 20 days (90% CI: 18-22). It is interesting to note that as the available data become more, the estimated recovery rate increases significantly from the 31th of January (see Fig.2).

In Figures3,4,5, we show the coefficients of determination (R^2) and the root of mean squared errors ($RMSE$), for $\hat{R_0}$, $\hat{\beta}$ and $\hat{\gamma}$, respectively.

As described in the methodology, we have also used the SIRD simulator to provide an estimation of the infection rate by optimization with w_1=1, w_2=2, w_3=2. Thus, we performed the simulations by setting β=0.05 and γ=0.0294, and as initial conditions one infected, zero recovered and zero deaths on November 16th 2019, and run until the 10th of February. The optimal, with respect to the reported confirmed cases from the 11th of January to the 10th of February, value of the infected rate (α) was $\sim$ 0.201 (90% CI: 0.199-0.203). This corresponds to a mean value of the basic reproduction number $\hat{R_0} \approx 2.53$. Note that this value is different compared to the value that was estimated using solely the reported data.
Finally, using the derived values of the parameters α, β, γ, we have run the SIRD simulator until the end of February. The results of the simulations are given in Figures6,7,8. Solid lines depict the evolution, when using the expected (mean) estimations and dashed lines illustrate the corresponding lower and upper bounds as computed at the limits of the confidence intervals of the estimated parameters.

As Figures 6,7 suggest, the forecast of the outbreak at the end of February, through the SIRD model is characterized by high uncertainty. In particular, simulations result in an expected number of $\sim$ 160,000 infected cases but with a high variation: the lower bound is at $\sim$ 84,000 infected cases while the upper bound is at $\sim$ 330,000 cases. Similarly for the recovered population, simulations result in an expected number of $\sim$ 70,000, while the lower and upper bounds are at $\sim$ 44,000 and $\sim$ 100,000, respectively. Finally, regarding the deaths, simulations result in an average number of $\sim$ 40,000, with lower and upper bounds, $\sim$ 23,000 and $\sim$ 72,000, respectively.
However, as more data are released it appears that the mortality rate is much lower than the predicted with the current data and thus the death toll is expected, hopefully to be significantly less compared with the predictions.
Furthermore, simulations reveal that the actual number of deaths is significantly lower than the lower bound of the simulations. This suggests that the mortality rate is considerably smaller than the estimated one based on the reported official data. Thus, it is expected that the actual numbers of the infected and consequently the recovered cases are considerably larger than the reported ones. Hence, we assessed the dynamics of the epidemics by considering a different scenario that we present in the following subsection.

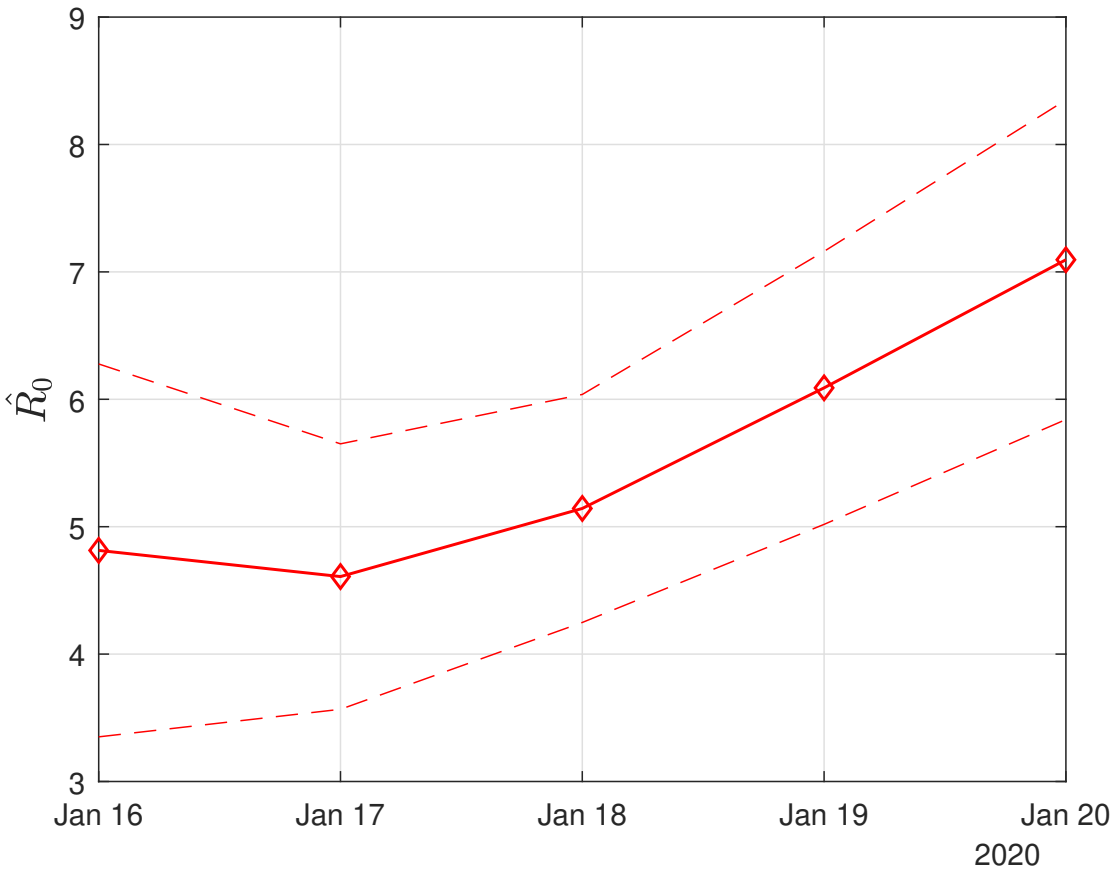

Figure 1: Scenario I. Estimated values of the basic reproduction number (R_0) as computed by least squares using a rolling window with initial date the 11th of January. The solid line corresponds to the mean value and dashed lines to lower and upper 90% confidence intervals.

3.2 Scenario II. Results obtained based by taking twenty times the number of infected and forty times the number of recovered people with respect to the confirmed cases

For our illustrations, we assumed that the number of infected is twenty times the number of the confirmed infected and forty times the number of the confirmed recovered people. Figure9 depicts an estimation of R_0 for the period January 16-January 20. Using the first six days from the 11th of January to 16th of January, $\hat{R}_0$ results in 3.25 (90% CI: 2.37-4.14); using the data until January 17, $\hat{R}_0$ results in 3.13 (90% CI: 2.50-3.76); using the data until January 18, $\hat{R}_0$ results in 3.42 (90% CI: 2.91-3.93); using the data until January 19, $\hat{R}_0$ results in 3.96 (90% CI: 3.36-4.57) and using the data until January 20, $\hat{R}_0$ results in 4.58 (90% CI: 3.81-5.35).
It is interesting to note that the above estimation of R_0 is close enough to the one reported in other studies (see in the Introduction for a review).

Figure10 depicts the estimated values of the recovery (β) and mortality (γ) rates for the period January 16 to February 10. The confidence intervals are also depicted with dashed lines. Note that the large variation in the estimated values of β and γ should be accounted to the small size of the data and data uncertainty. This is also reflected in the corresponding confidence intervals. As more data are taken into account, this variation is significantly reduced. Thus,using all the (scaled) data from the 11th of January until the 10th of February, the estimated value of the mortality rate γ now drops to $\sim$ 0.147% (90% CI: 0.144%-0.149%) while that of the recovery rate is $\sim$ 0.1 (90% CI: 0.091-0.11) corresponding to

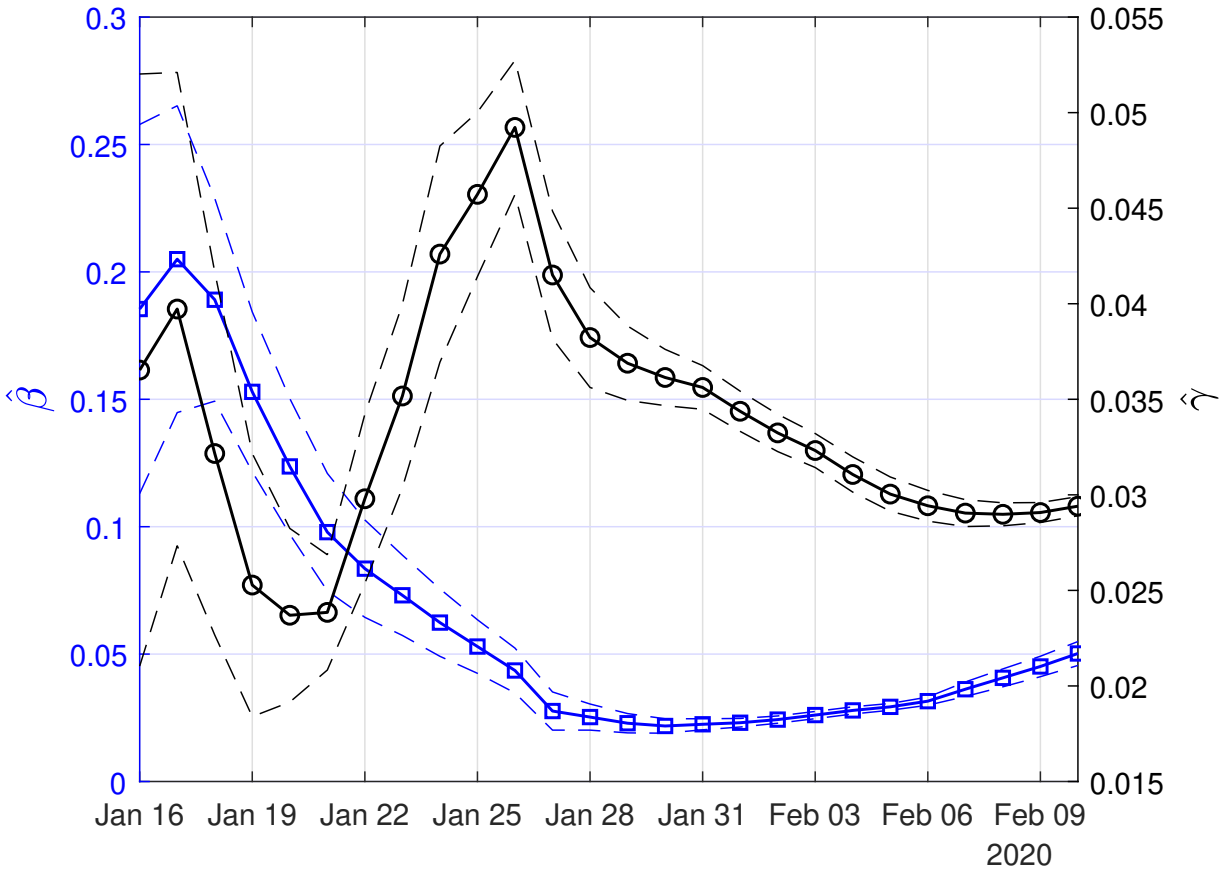

Figure 2: Scenario I. Estimated values of the recovery ($\hat{\beta}$) and mortality ($\hat{\gamma}$) rate as computed by least squares using a rolling window (see section 2.1). Solid lines correspond to the mean values and dashed lines to lower and upper 90% confidence intervals

$\sim$ 10 days (90% CI: 9-11 days). It is interesting also to note, that as the available data become more, the estimated recovery rate increases slightly (see Fig.10), while the mortality rate seems to be stabilized at a rate of $\sim$ 0.147%.

In Figures 11,12,13, we show the coefficients of determination (R^2) and the root of mean squared errors ($RMSE$), for $\hat{R}_0$, $\hat{\beta}$ and $\hat{\gamma}$, respectively.

Again, we used the SIRD simulator to provide estimation of the infection rate by optimization setting $w_1 = 1$, $w_2 = 400$, $w_3 = 1$ to balance the residuals of deaths with the scaled numbers of the infected and recovered cases. Thus, to find the optimal infection run, we used the SIRD simulations with $\beta = 0.1$, and $\gamma = 0.00147$ and as initial conditions one infected, zero recovered, zero deaths on November 16th 2019, and run until the 10th of February. The optimal, with respect to the reported confirmed cases from the 11th of January to the 10th of February value of the infected rate (α) was $\sim$ 0.25 (90% CI: 0.248-0.252). This corresponds to a mean value of the basic reproduction number $\hat{R}_0 \approx 2.46$. Finally, using the derived values of the parameters α, β, γ, we have run the SIRD simulator until the end of February. The simulation results are given in Figures14,15,16. Solid lines depict the evolution, when using the expected (mean) estimations and dashed lines illustrate the corresponding lower and upper bounds as computed at the limits of the confidence intervals of the estimated parameters.

Again as Figures 15,16 suggest, the forecast of the outbreak at the end of February, through the SIRD model is characterized by high uncertainty. In particular, in Scenario II, by February 29, simulations result in an expected actual number of 1.9m infected cases in the whole population with a lower bound at 660,000 and an upper bound at $\sim$4.6m cases. Similarly, for the recovered population, simulations result in an expected actual number of 1.3m, while the lower and upper bounds are at 540,000 and 2.9m, respectively. Finally, regarding the deaths, simulations under this scenario

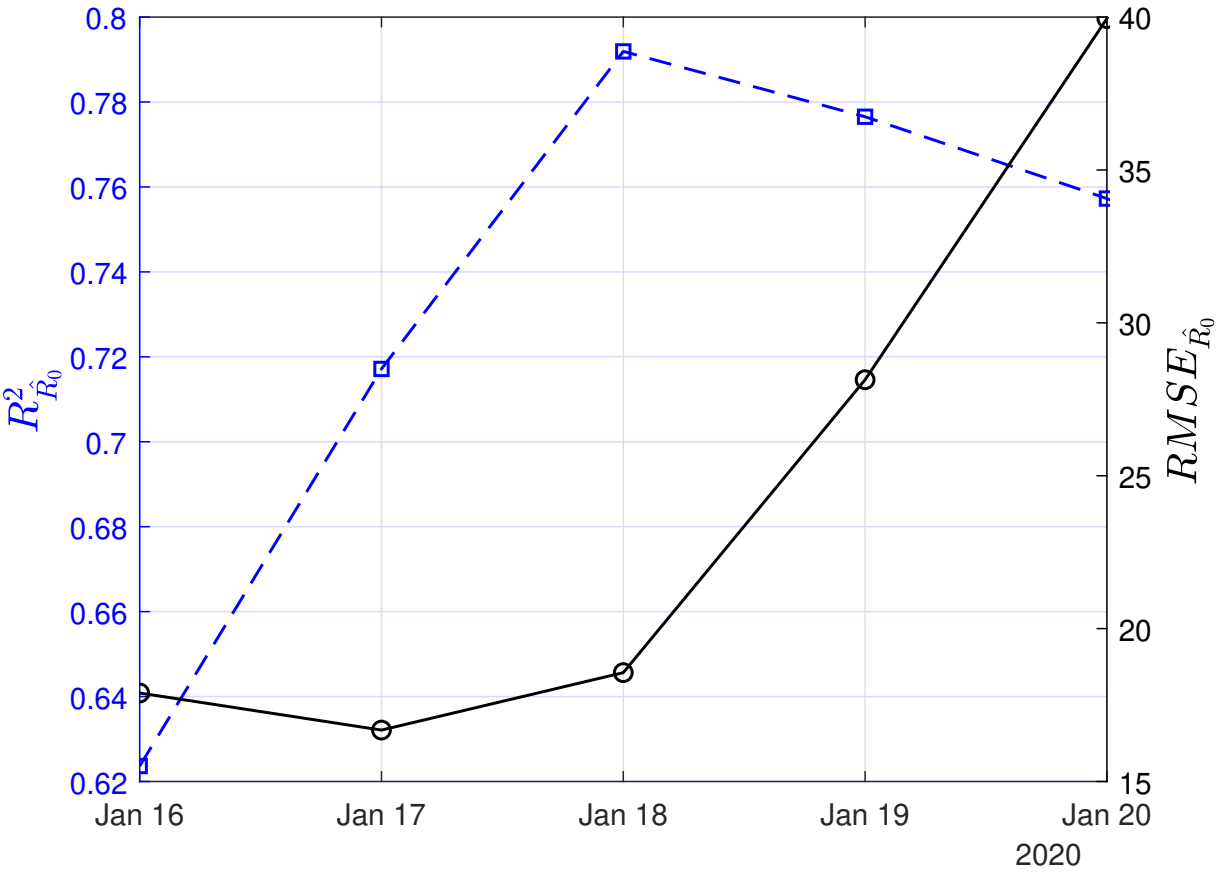

Figure 3: Scenario I. Coefficient of determination (R^2) and root mean square error ($RMSE$) resulting from the solution of the linear regression problem with least-squares for the basic reproduction number (R_0)

result in an average number of 19,000, with lower and upper bounds, 7,000 and 47,000.
Table 1 summarizes the above results for both scenarios.
We note, that the results derived under the Scenario II seem to better reflect the actual situation as the reported number of deaths is within the average and lower limits of the SIRD simulations. In particular, the reported number of deaths on the 22th February was 2,346, while the lower bound of the forecast is 2,900. This indicates an even lower mortality rate than that of 0.147%, thus an even larger actual number of infected (and recovered) cases in the whole population. At this point we should make clear and underline that the above forecasts should not be considered by any means as actual/ real- world predictions for the evolution of the epidemic as they are based on yet incomplete knowledge of the characteristics of the epidemic, the high uncertainty regarding the actual number of already infected and recovered people and the effect of the measures already taken to combat the epidemic.

4 Discussion and Concluding Remarks

In the digital and globalized world of today, new data and information on the novel coronavirus and the evolution of the outbreak become available at an unprecedented pace. Still, crucial questions remain unanswered and accurate answers for predicting the dynamics of the outbreak simply cannot be obtained at this stage. We emphatically underline the uncertainty of available official data, particularly pertaining to the true baseline number of infected (cases), that may lead to ambiguous results and inaccurate forecasts by orders of magnitude, as also pointed out by other investigators [1, 14, 18]. Of note, COVID-19, which is thought to be principally transmitted from person to person by respiratory droplets and fomites without excluding the possibility of the fecal-oral route [19] had been spreading for at least over a month and a half before the imposed lockdown and quarantine of Wuhan on January 23, having thus infected unknown

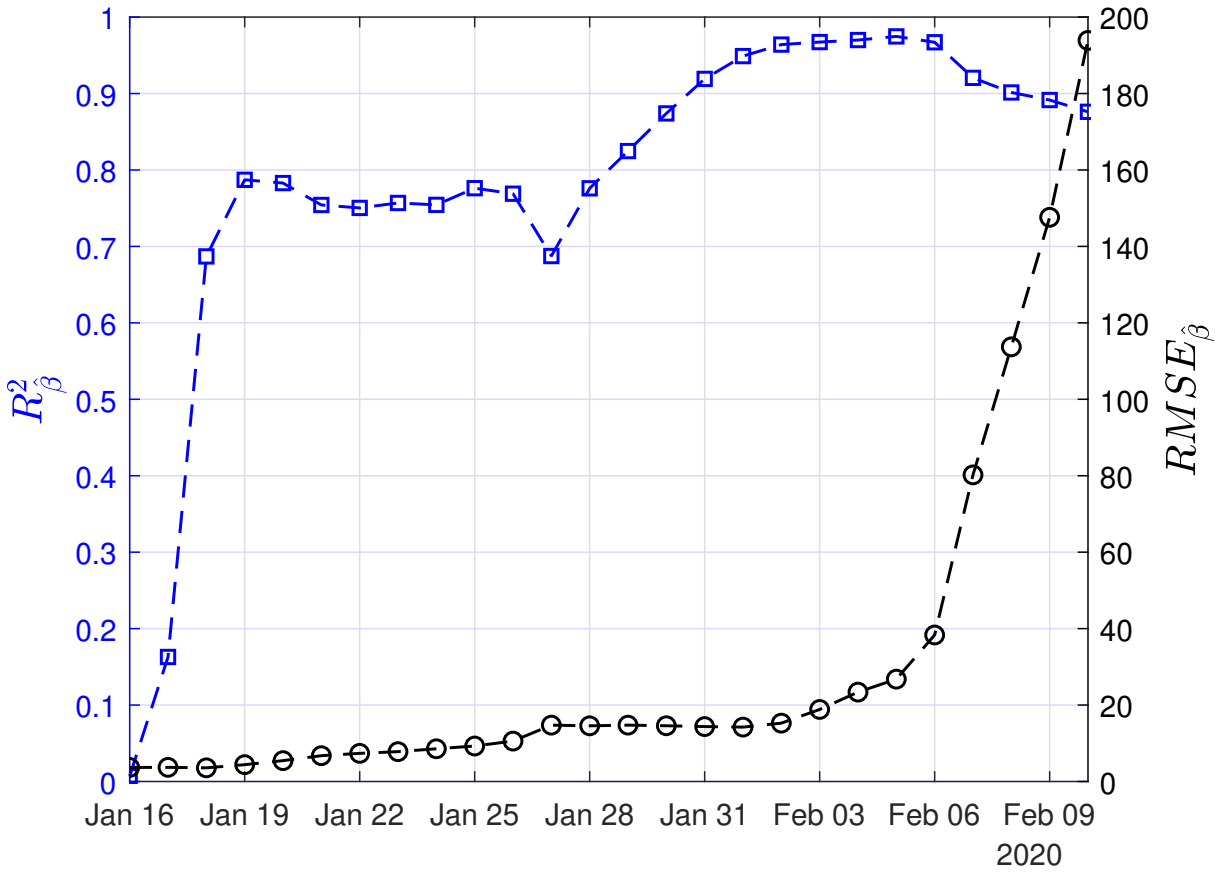

Figure 4: Scenario I. Coefficient of determination (R^2) and root mean square error ($RMSE$) resulting from the solution of the linear regression problem with least-squares for the recovery rate (β)

numbers of people. The number of asymptomatic and mild cases with subclinical manifestations that probably did not present to hospitals for treatment may be substantial; these cases, which possibly represent the bulk of the COVID-19 infections, remain unrecognized, especially during the influenza season [18]. This highly likely gross under-detection and underreporting of mild or asymptomatic cases inevitably throws severe disease courses calculations and death rates out of context, distorting epidemiologic reality. Another important factor that should be taken into consideration pertains to the diagnostic criteria used to determine infection status and confirm cases. A positive PCR test was required to be considered a confirmed case by China's Novel Coronavirus Pneumonia Diagnosis and Treatment program in the early phase of the outbreak [20]. However, the sensitivity of nucleic acid testing for this novel viral pathogen may only be 30-50%, thereby often resulting in false negatives, particularly early in the course of illness. To complicate matters further, the guidance changed in the recently-released fourth edition of the program on February 6 to allow for diagnosis based on clinical presentation, but only in Hubei province [20]. The swiftly growing epidemic seems to be overwhelming even for the highly efficient Chinese logistics that did manage to build two new hospitals in record time to treat infected patients. Supportive care with extracorporeal membrane oxygenation (ECMO) in intensive care units (ICUs) is critical for severe respiratory disease. Large-scale capacities for such level of medical care in Hubei province, or elsewhere in the world for that matter, amidst this public health emergency may prove particularly challenging. We hope that the results of our analysis contribute to the elucidation of critical aspects of this outbreak so as to contain the novel coronavirus as soon as possible and mitigate its effects regionally, in mainland China, and internationally.

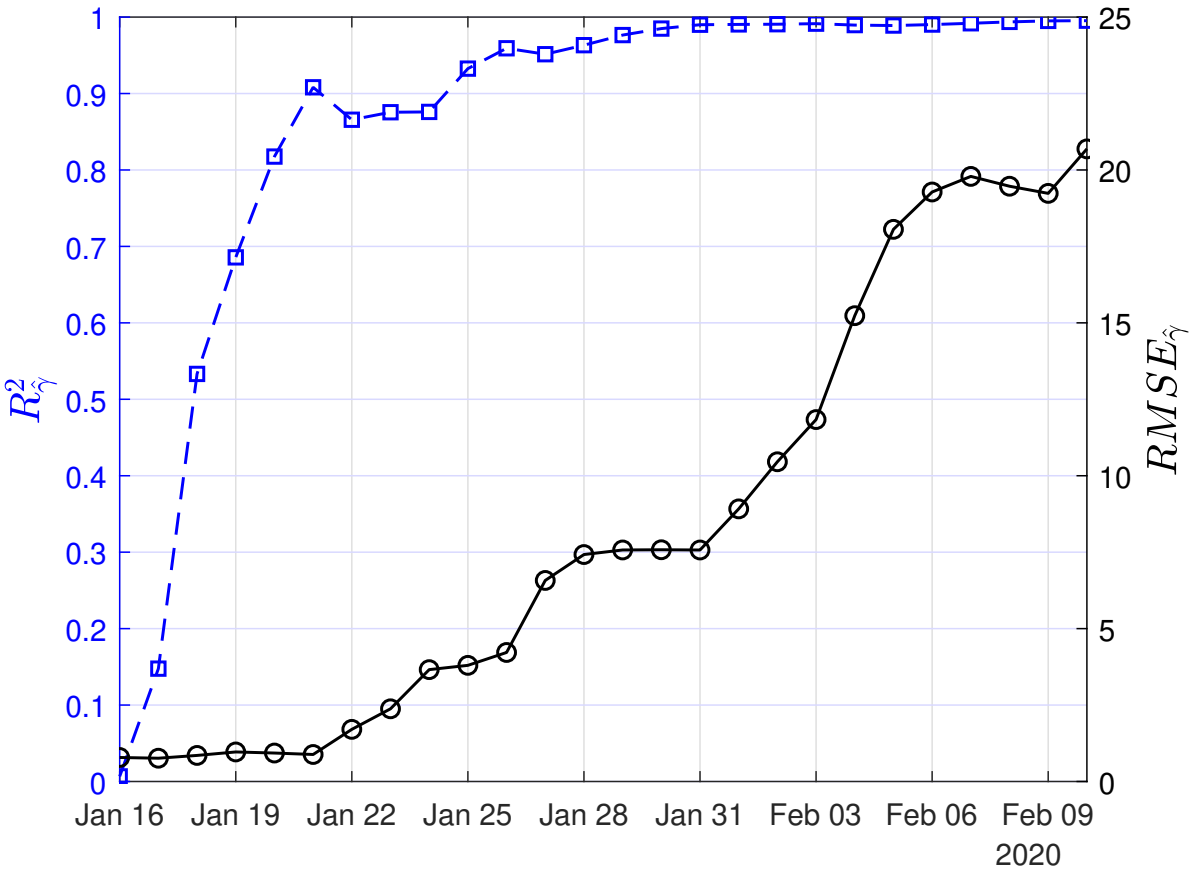

Figure 5: Scenario I. Coefficient of determination (R^2) and root mean square error (RMSE) resulting from the solution of the linear regression problem with least-squares for the mortality rate (γ)

References

[1] Q Li, X Guan, P Wu, and et al. *Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia*, jan 2020.

[2] World Health Organization. *WHO Statement Regarding Cluster of Pneumonia Cases in Wuhan, China*, 2020.

[3] World Health Organization. *Novel coronavirus(2019-nCoV). Situation report 21. Geneva, Switzerland: World Health Organization; 2020*, 2020.

[4] Roujian Lu, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, Yuhai Bi, Xuejun Ma, Faxian Zhan, Liang Wang, Tao Hu, Hong Zhou, Zhenhong Hu, Weimin Zhou, Li Zhao, Jing Chen, Yao Meng, Ji Wang, Yang Lin, Jianying Yuan, Zhihao Xie, Jinmin Ma, William J Liu, Dayan Wang, Wenbo Xu, Edward C Holmes, George F Gao, Guizhen Wu, Weijun Chen, Weifeng Shi, and Wenjie Tan. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, jan 2020.

[5] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, Hui-Dong Chen, Jing Chen, Yun Luo, Hua Guo, Ren-Di Jiang, Mei-Qin Liu, Ying Chen, Xu-Rui Shen, Xi Wang, Xiao-Shuang Zheng, Kai Zhao, Quan-Jiao Chen, Fei Deng, Lin-Lin Liu, Bing Yan, Fa-Xian Zhan, Yan-Yi Wang, Geng-Fu Xiao, and Zheng-Li Shi. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, feb 2020.

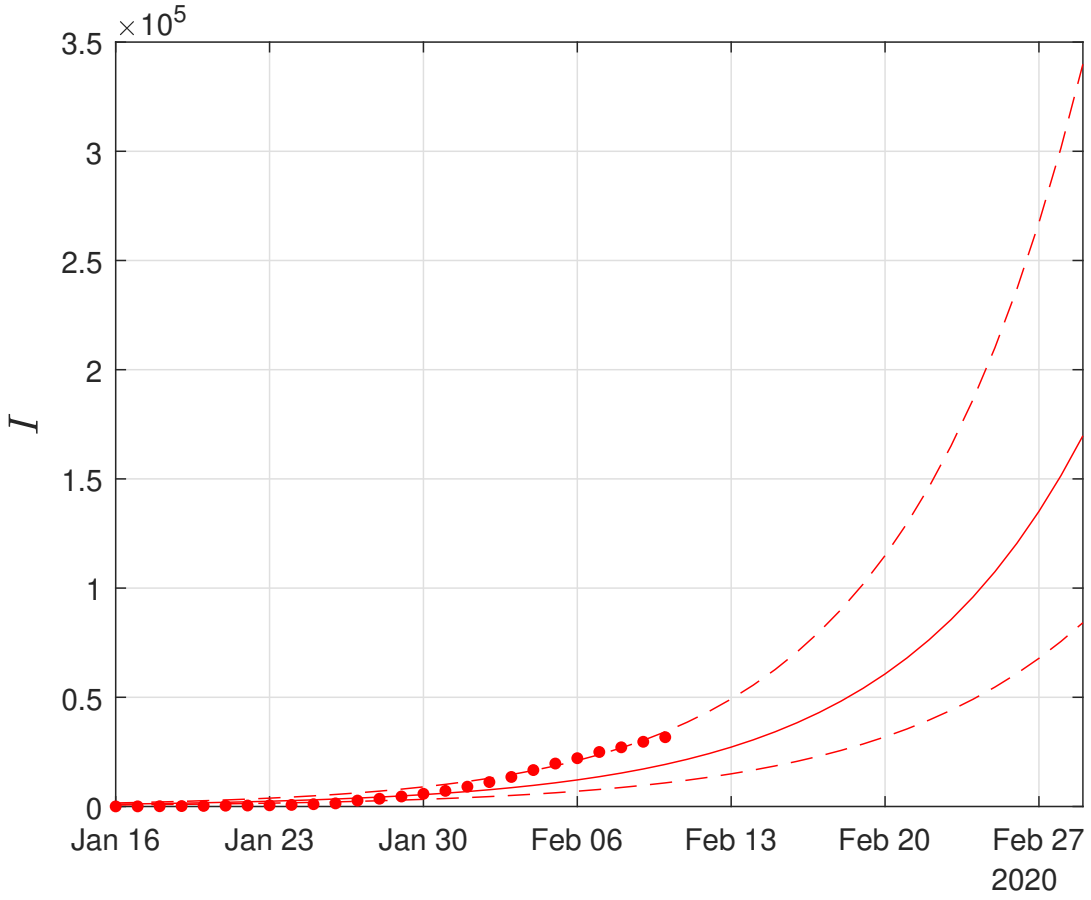

Figure 6: Scenario I. Simulations until the 29th of February of the cumulative number of infected as obtained using the SIRD model. Dots correspond to the number of confirmed cases from the 16th of January to the 10th of February. The initial date of the simulations was the 16th of November with one infected, zero recovered and zero deaths. Solid lines correspond to the dynamics obtained using the estimated expected values of the epidemiological parameters $\alpha = 0.201$, $\beta = 0.05$, $\gamma = 0.0294$;dashed lines correspond to the lower and upper bounds derived by performing simulations on the limits of the confidence intervals of the parameters.

[6] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, Jiaan Xia, Ting Yu, Xinxin Zhang, and Li Zhang. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The Lancet*, jan 2020.

[7] A Patel, DB Jernigan, and 2019 nCoV CDC Response Team. Initial public health response and interim clinical guidance for the 2019 novel coronavirus outbreak - united states, december 31, 2019-february 4, 2020. *MMWR Morb Mortal Wkly Rep*, feb 2020.

[8] C Hunag, Y Wang, X Li, and et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *Lancet*, jan 2020.

[9] Dawei Wang, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu, Jing Zhang, Binbin Wang, Hui Xiang, Zhenshun Cheng, Yong Xiong, Yan Zhao, Yirong Li, Xinghuan Wang, and Zhiyong Peng. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *JAMA*, feb 2020.

[10] Shi Zhao, Qianyin Lin, Jinjun Ran, Salihu S Musa, Guangpu Yang, Weiming Wang, Yijun Lou, Daozhou Gao, Lin Yang, Daihai He, and Maggie H Wang. Preliminary estimation of the basic reproduction 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 Dis*, jan 2020.

[11] N Imai, A Cori, I Dorigatti, and et al. Report 3: Transmissibility of 2019-ncov. *Int J Infect Dis*, dec 2019.

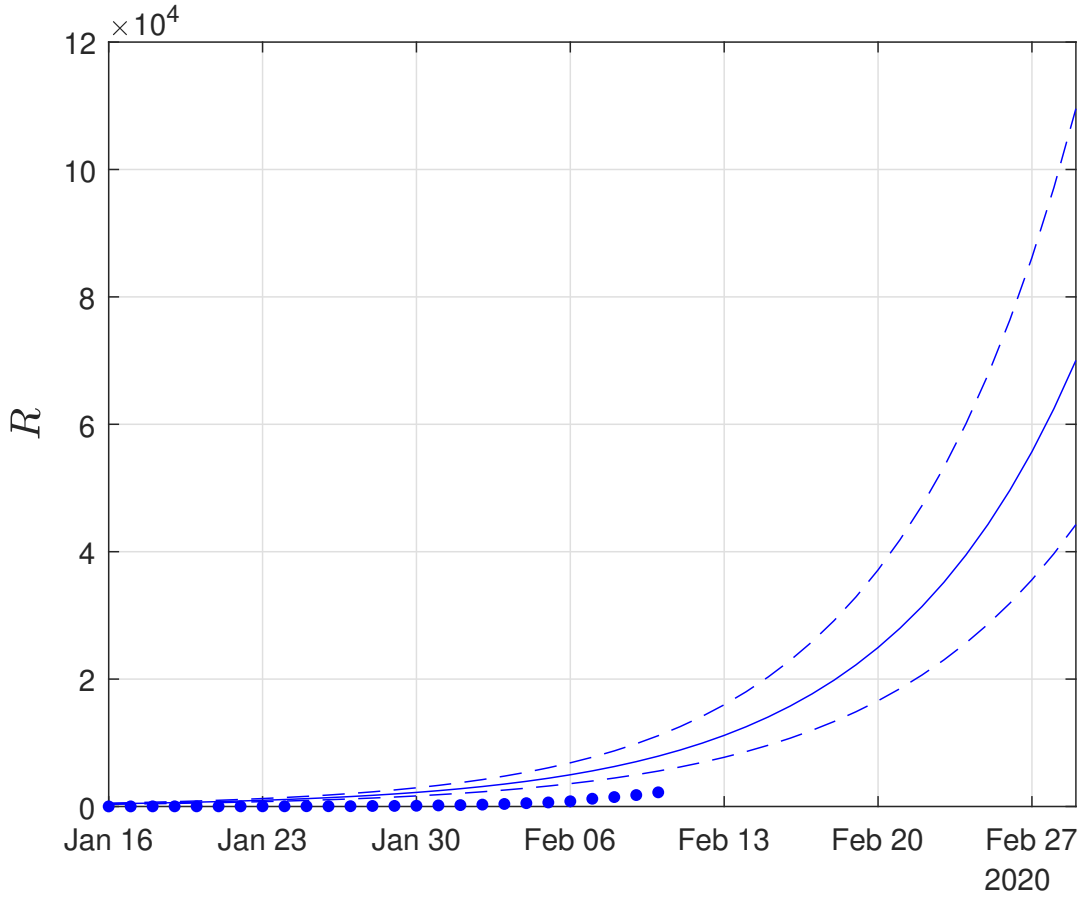

Figure 7: Scenario I. Simulations until the 29th of February of the cumulative number of recovered as obtained using the SIRD model. Dots correspond to the number of confirmed cases from the 16th of January to the 10th of February. The initial date of the simulations was the 16th of November with one infected, zero recovered and zero deaths. Solid lines correspond to the dynamics obtained using the estimated expected values of the epidemiological parameters $\alpha = 0.201$, $\beta = 0.05$, $\gamma = 0.0294$;dashed lines correspond to the lower and upper bounds derived by performing simulations on the limits of the confidence intervals of the parameters.

[12] Joseph T Wu, Kathy Leung, and Gabriel M Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in wuhan, china: a modelling study. *The Lancet*, jan 2020.

[13] Constantinos I. Siettos and Lucia Russo. Mathematical modeling of infectious disease dynamics. *Virulence*, 4(4):295–306, may 2013.

[14] Peng Wu, Xinxin Hao, Eric H Y Lau, Jessica Y Wong, Kathy S M Leung, Joseph T Wu, Benjamin J Cowling, and Gabriel M Leung. Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in wuhan, china, as at 22 january 2020. *Eurosurveillance*, 25(3), jan 2020.

[15] NHC. *NHS Press Conference, Feb. 4 2020 - National Health Commission (NHC) of the People's Republic of China*, feb 2020.

[16] A. C. Ghani, C. A. Donnelly, D. R. Cox, J. T. Griffin, C. Fraser, T. H. Lam, L. M. Ho, W. S. Chan, R. M. Anderson, A. J. Hedley, and G. M. Leung. Methods for Estimating the Case Fatality Ratio for a Novel, Emerging Infectious Disease. *American Journal of Epidemiology*, 162(5):479–486, 09 2005.

[17] The Mathworks, Inc., Natick, Massachusetts. *MATLAB R2018b*, 2018.

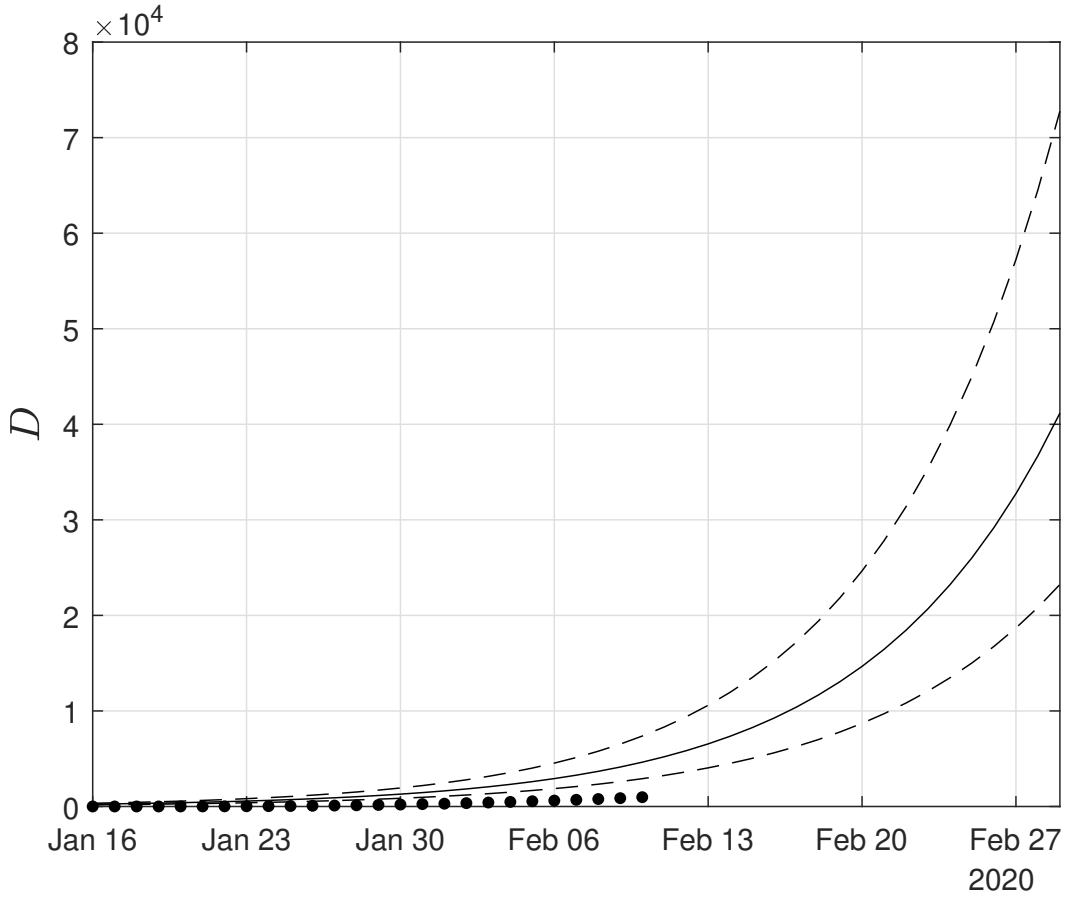

Figure 8: Scenario I. Simulations until the 29th of February of the cumulative number of deaths as obtained using the SIRD model. Dots correspond to the number of confirmed cases from 16th of January to the 10th of February. The initial date of the simulations was the 16th of November with one infected, zero recovered and zero deaths. Solid lines correspond to the dynamics obtained using the estimated expected values of the epidemiological parameters $\alpha = 0.201$, $\beta = 0.05$, $\gamma = 0.0294$;dashed lines correspond to the lower and upper bounds derived by performing simulations on the limits of the confidence intervals of the parameters.

[18] Manuel Battegay, Richard Kuehl, Sarah Tschudin-Sutter, Hans H. Hirsch, Andreas F. Widmer, and Richard A. Neher. 2019-novel coronavirus (2019-nCoV): estimating the case fatality rate – a word of caution. *Swiss Medical Weekly*, feb 2020.

[19] J Gale. *Coronavirus May Transmit Along Fecal-Oral Route, Xinhua Reports*, feb 2020.

[20] Johns Hopkins Center for Health Security. *Daily updates on the emerging novel coronavirus from the Johns Hopkins Center for Health Security. February 9, 2020*, feb 2020.

.

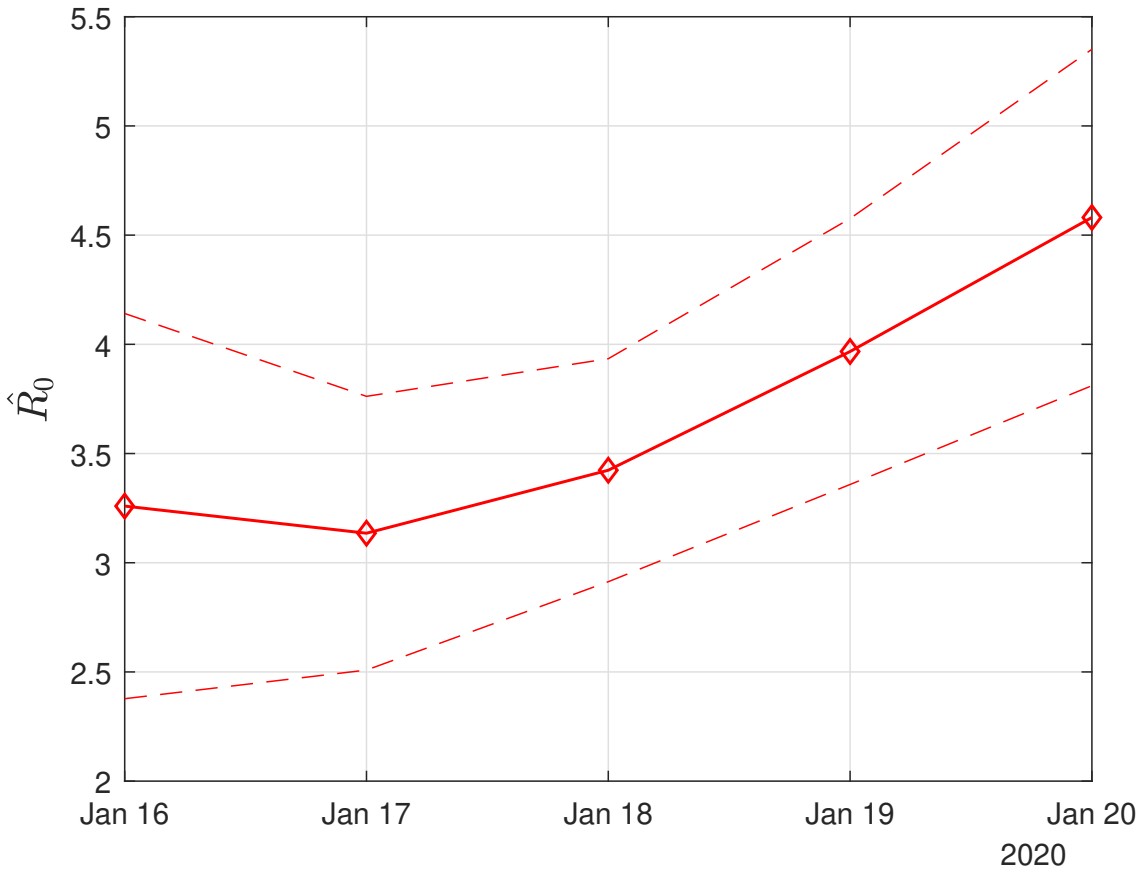

Figure 9: Scenario II. Estimated values of the basic reproduction number (R_0) as computed by least squares using a rolling window with initial date the 11th of January. The solid line corresponds to the mean value and dashed lines to lower and upper 90% confidence intervals.

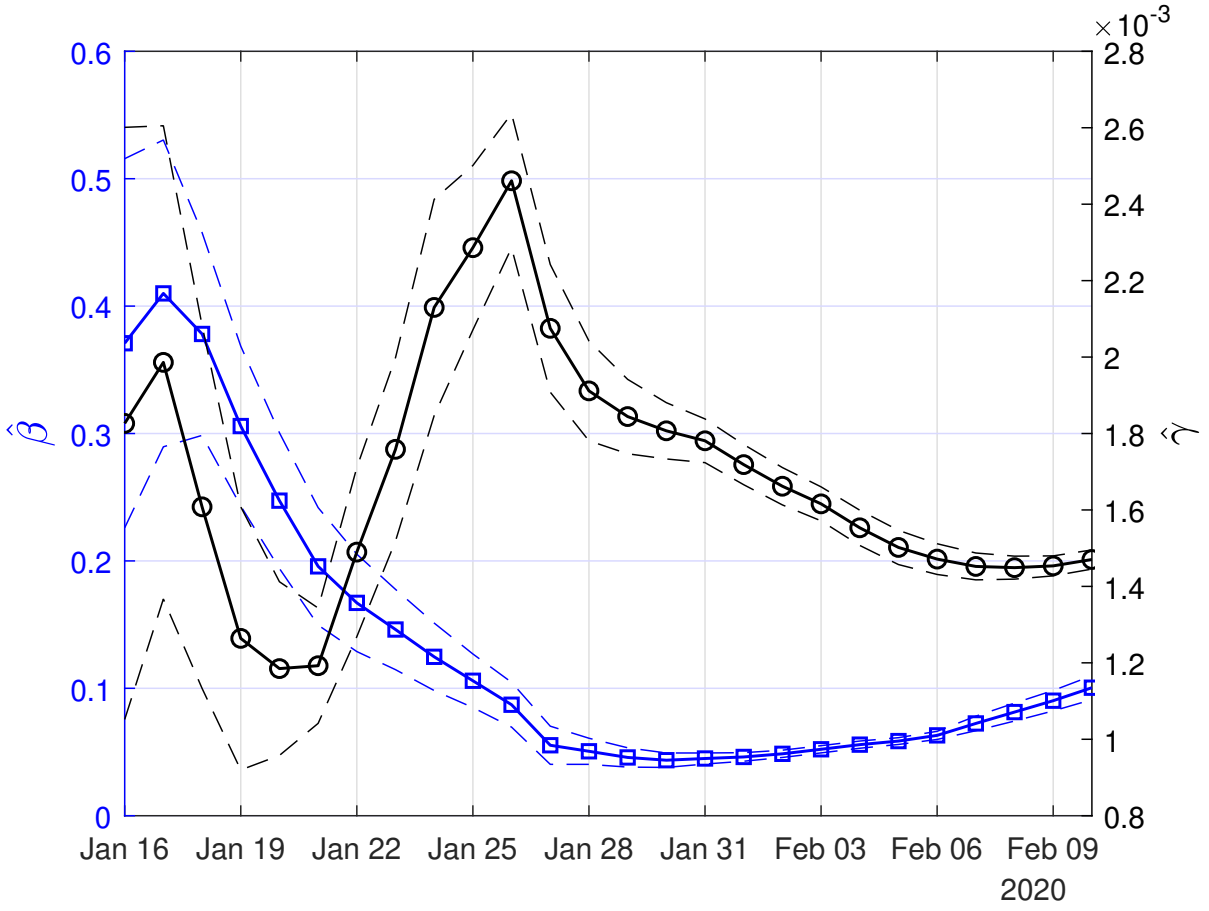

Figure 10: Scenario II. Estimated values of the recovery ($\hat{\beta}$) rate and mortality ($\hat{\gamma}$) rate, as computed by least squares using a rolling window (see section 2.1). Solid lines correspond to the mean values and dashed lines to lower and upper 90% confidence intervals

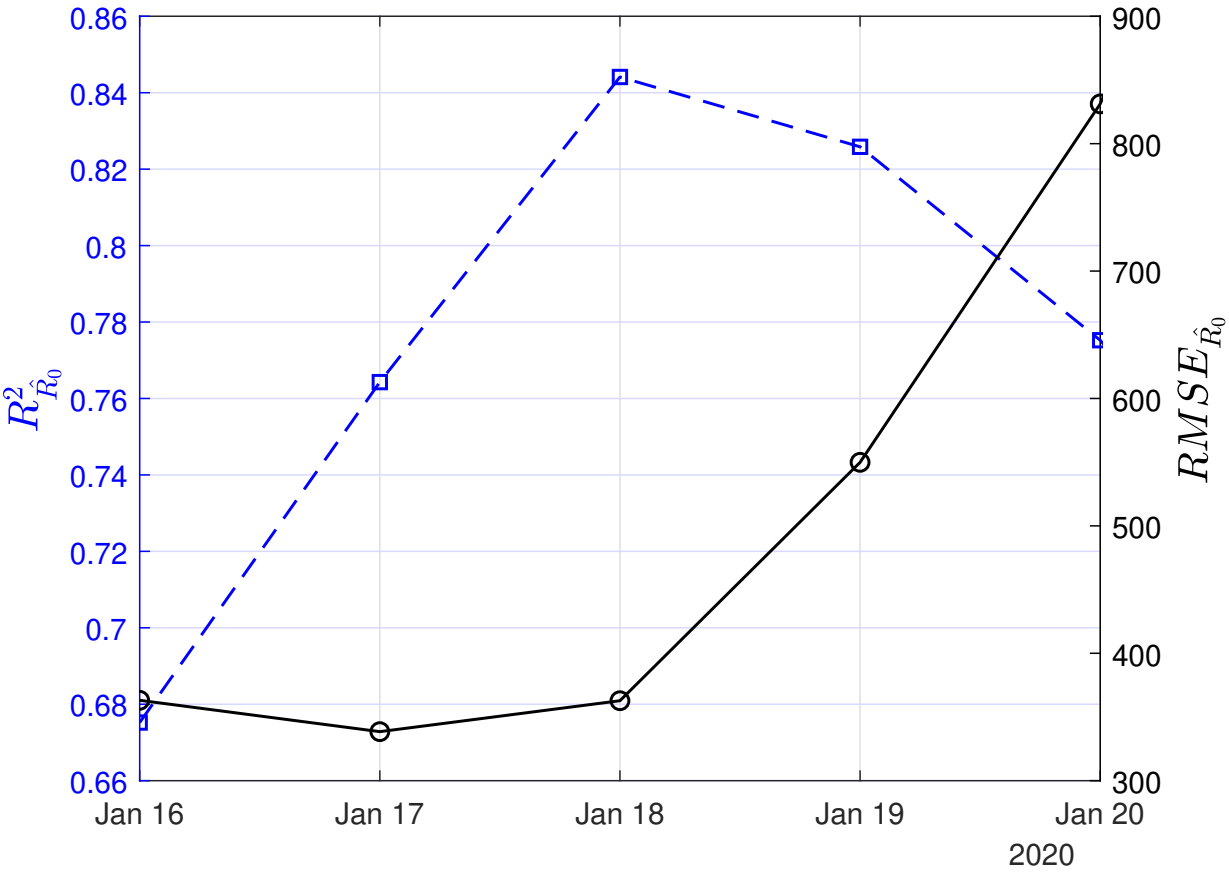

Figure 11: Scenario II. Coefficient of determination (R^2) and root mean square error ($RMSE$) resulting from the solution of the linear regression problem with least-squares for the basic reproduction number (R_0).

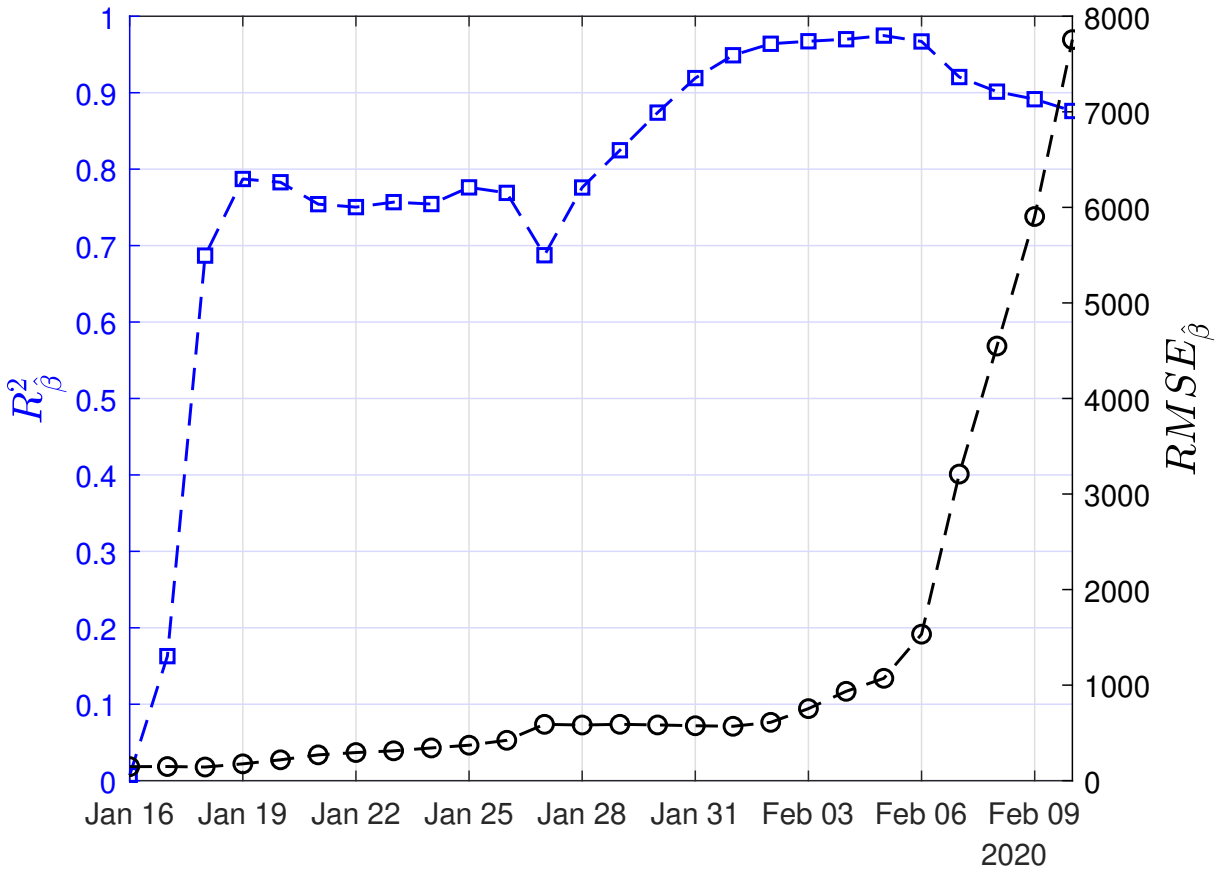

Figure 12: Scenario II. Coefficient of determination (R^2) and root mean square error ($RMSE$) resulting from the solution of the linear regression problem with least-squares for the recovery rate (β).

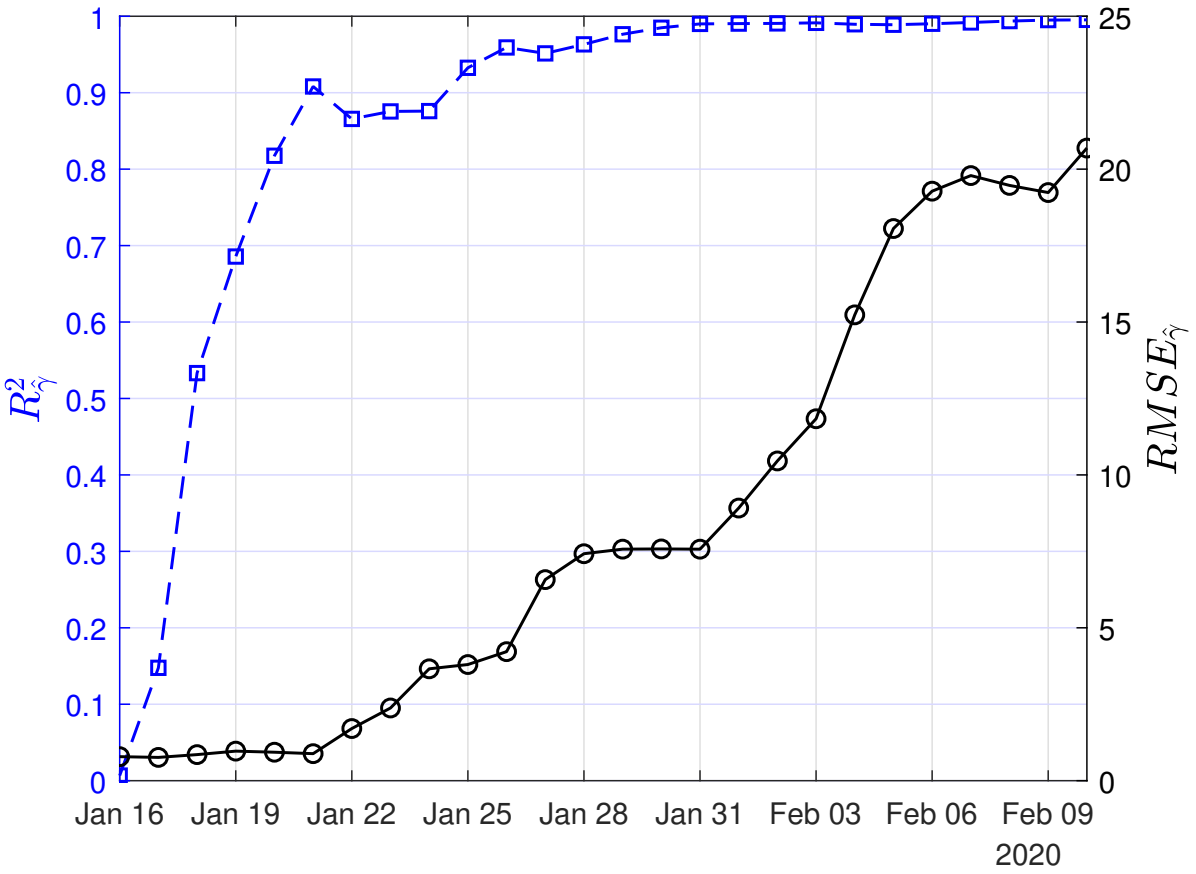

Figure 13: Scenario II. Coefficient of determination (R^2) and root mean square error ($RMSE$) resulting from the solution of the linear regression problem with least-squares for the mortality rate (γ).

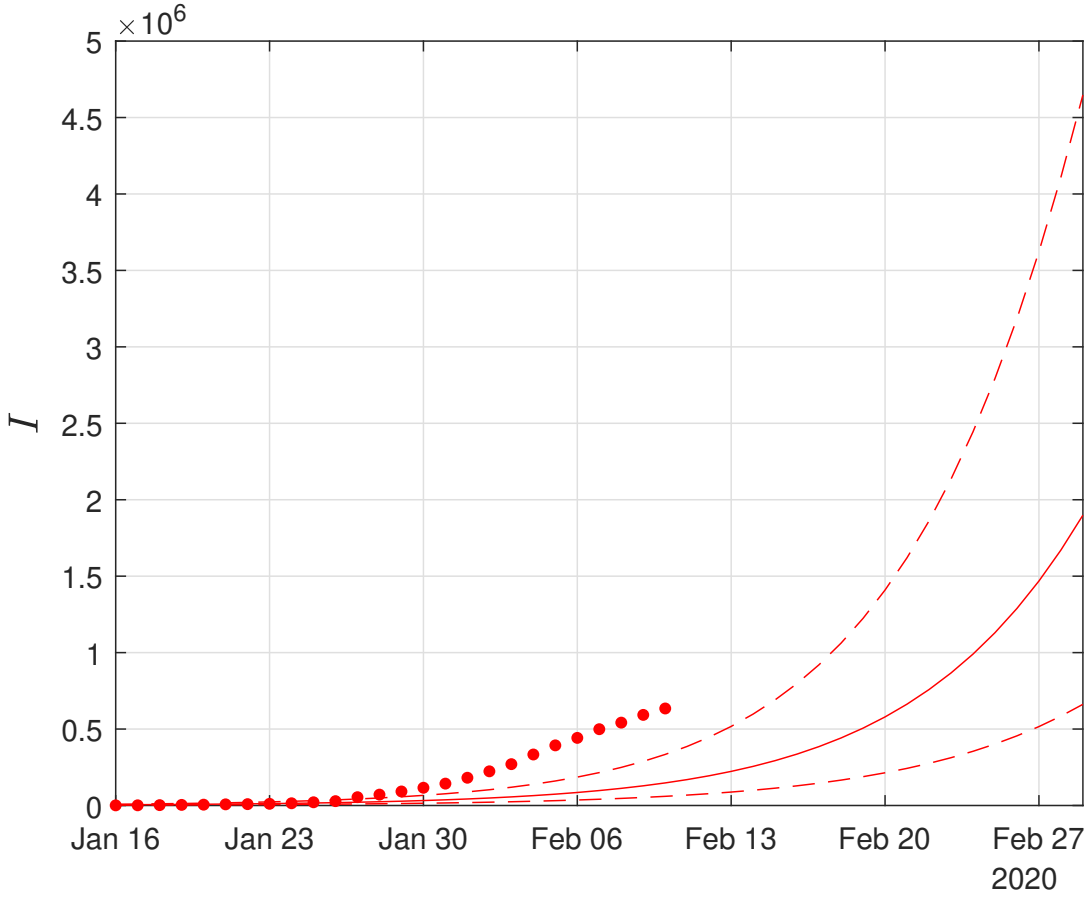

Figure 14: Scenario II. Simulations until the 29th of February of the cumulative number of infected as obtained using the SIRD model. Dots correspond to the number of confirmed cases from 16th of Jan to the 10th of February. The initial date of the simulations was the 16th of November with one infected, zero recovered and zero deaths. Solid lines correspond to the dynamics obtained using the estimated expected values of the epidemiological parameters $\alpha = 0.25$, $\beta = 0.1$, $\gamma = 0.00147$; dashed lines correspond to the lower and upper bounds derived by performing simulations on the limits of the confidence intervals of the parameters.

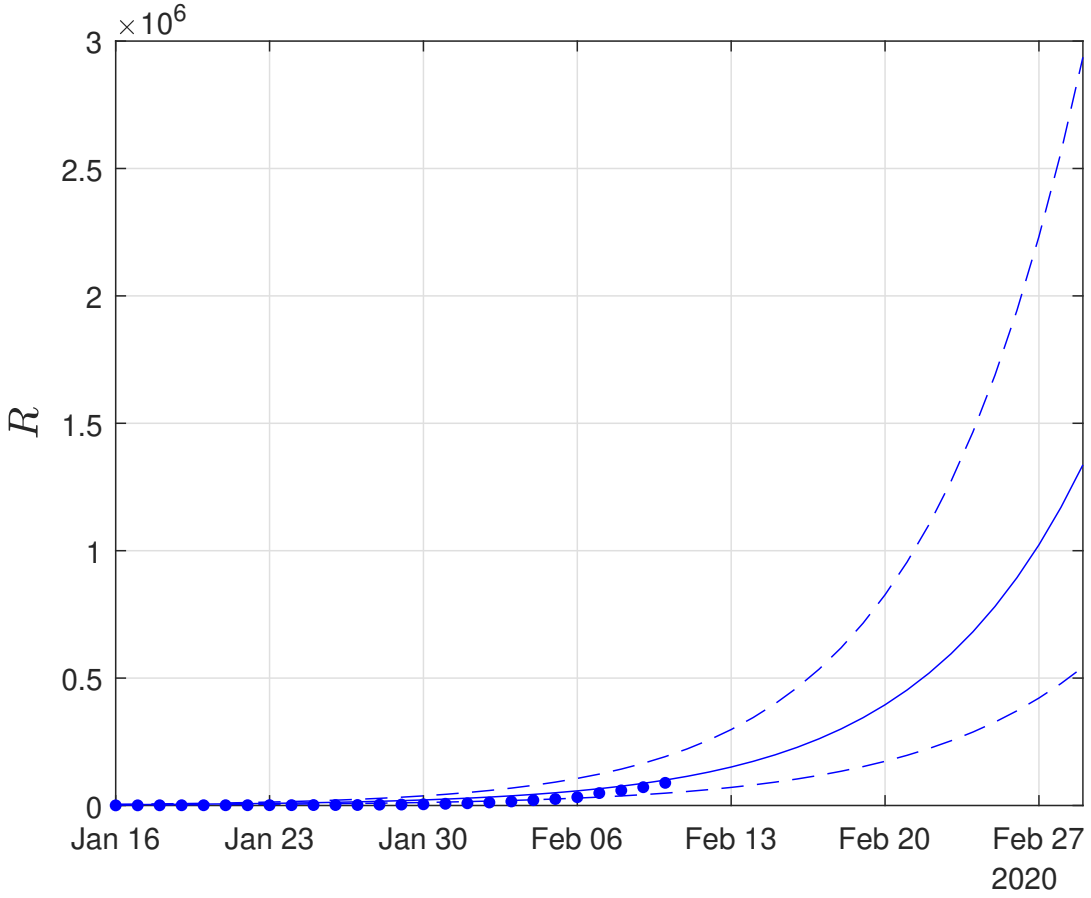

Figure 15: Scenario II. Simulations until the 29th of February of the cumulative number of recovered as obtained using the SIRD model. Dots correspond to the number of confirmed cases from 16th of January to the 10th of February. The initial date of the simulations was the 16th of November, with one infected, zero recovered and zero deaths. Solid lines correspond to the dynamics obtained using the estimated expected values of the epidemiological parameters $\alpha = 0.25$, $\beta = 0.1$, $\gamma = 0.00147$; dashed lines correspond to the lower and upper bounds derived by performing simulations on the limits of the confidence intervals of the parameters.

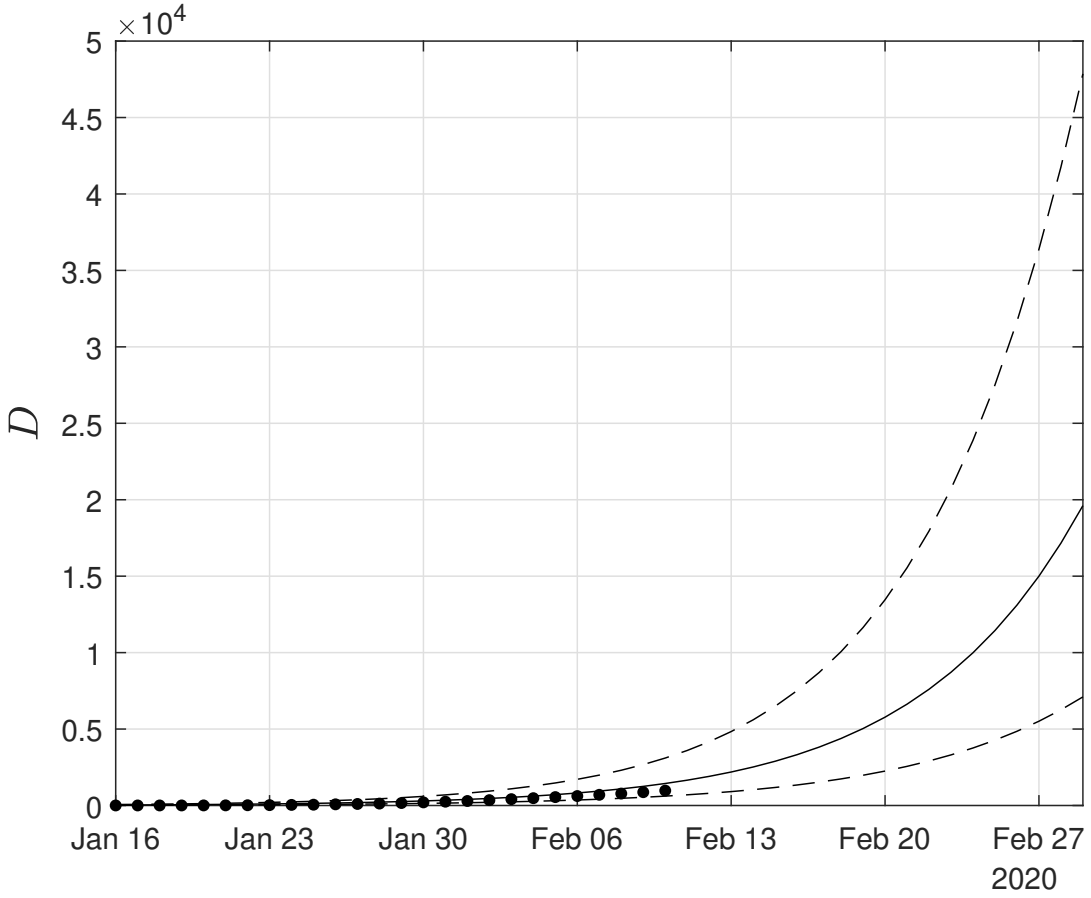

Figure 16: Scenario II. Simulations until the 29th of February of the cumulative number of deaths as obtained using the SIRD model. Dots correspond to the number of confirmed cases from the 16th of November to the 10th of February. The initial date of the simulations was the 16th of November with zero infected, zero recovered and zero deaths. Solid lines correspond to the dynamics obtained using the estimated expected values of the epidemiological parameters $\alpha = 0.25$, $\beta = 0.1$, $\gamma = 0.00147$; dashed lines correspond to the lower and upper bounds derived by performing simulations on the limits of the confidence intervals of the parameters.

	symbol	parameter	computed values	90% CI
Scenario I Exact numbers for confirmed cases	R_0	Basic reproduction number		
		11-16 Jan	4.80	3.36-6.67
	Based on linear regression	11-17 Jan	4.60	3.56-5.65
		11-18 Jan	5.14	4.25-6.03
	Based on the SIRD simulator	Nov 16-Feb 10	2.53	-
	α	infection rate	0.201	0.199-0.203
	β	recovery rate	0.050	0.045-0.055
		recovery time	20 days	18-22 days
	γ	mortality rate	2.94%	2.89%-3.0%
Forecast to Feb 29				
		infected	160,000	84,000-330,000
		recovered	70,000	44,000 - 100,000
		deaths	40,000	23,000 - 72,000
Scenario II x20 Infected, x40 recovered of confirmed cases	R_0	Basic reproduction number		
		11-16 Jan	3.25	2.37-4.14
	Based on linear regression	11-17 Jan	3.13	2.50-3.76
		11-18 Jan	3.42	2.91-3.93
	Based on the SIRD simulator	Nov 16-Feb 10	2.46	-
	α	infection rate	0.25	0.248-0.252
	β	recovery rate	0.10	0.091-0.11
		recovery time	10 days	9-11 days
	γ	mortality rate	0.147%	0.144%-0.149%
Forecast to Feb 29				
		infected	1.9m	660,000-4.6m
		recovered	1.3m	540,000 - 2.9m
		deaths	19,000	7,000 - 47,000

Table 1: Model parameters, their computed values and forecasts for the Hubei province under two scenarios: (I) using the exact values of confirmed cases or (II) using estimations for infected and recovered (twenty and forty times the number of confirmed cases, respectively).