Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(4): 3384–3403.
DOI: 10.3934/mbe.2021169
Received: 18 January 2021
Accepted: 11 April 2021
Published: 15 April 2021

Research article

Social contacts, epidemic spreading and health system. Mathematical modeling and applications to COVID-19 infection

Mattia Zanella[1,*], Chiara Bardelli[2], Mara Azzi[3], Silvia Deandrea[3], Pietro Perotti[3], Santino Silva[3], Ennio Cadum[3], Silvia Figini[4] and Giuseppe Toscani[1,5]

[1] Department of Mathematics, University of Pavia, Via Ferrata, 5, 27100 Pavia, Italy

[2] PhD Program in Computational Mathematics and Decision Sciences, University of Pavia, Italy

[3] Health Protection Agency (ATS), Viale Indipendenza, 3-27100 Pavia, Italy

[4] Department of Political and Social Sciences, University of Pavia, Corso Strada Nuova 65, 27100 Pavia, Italy

[5] Institute for Applied Mathematics and Information Technologies (IMATI), Via Ferrata, 1, 27100 Pavia, Italy

* **Correspondence:** Email: mattia.zanella@unipv.it.

Abstract: Lockdown and social distancing, as well as testing and contact tracing, are the main measures assumed by the governments to control and limit the spread of COVID-19 infection. In reason of that, special attention was recently paid by the scientific community to the mathematical modeling of infection spreading by including in classical models the effects of the distribution of contacts between individuals. Among other approaches, the coupling of the classical SIR model with a statistical study of the distribution of social contacts among the population, led some of the present authors to build a *Social* SIR model, able to accurately follow the effect of the decrease in contacts resulting from the lockdown measures adopted in various European countries in the first phase of the epidemic. The Social SIR has been recently tested and improved through a fruitful collaboration with the Health Protection Agency (ATS) of the province of Pavia (Italy), that made it possible to have at disposal all the relevant data relative to the spreading of COVID-19 infection in the province (half a million of people), starting from February 2020. The statistical analysis of the data was relevant to fit at best the parameters of the mathematical model, and to make short-term predictions of the spreading evolution in order to optimize the response of the local health system.

Keywords: epidemic models; disease control; social contacts; nonlinear incidence rate; healthcare system

1. Introduction

The recent spreading of the COVID-19 epidemic led the central governments to introduce restrictions such as social distancing and lockdown policies. These non-pharmaceutical containment measures were adopted with the aim of reducing the epidemic peak, in order to guarantee public health and the efficiency of the national health system in periods of strong increase in the number of infected. The effects of the restrictions were recently tested by resorting to mathematical models of compartmental epidemiology [1, 2], namely models in which the population is divided into compartments in order to simulate the epidemic evolution. In the Italian case, the actual evolution of the number of registered infected people immediately appeared subject to unavoidable incompleteness of the data during the entire emergency. It has therefore proved necessary to develop forecasting techniques to understand the impact of uncertain quantities during a health emergency. Furthermore, the necessary easing of lockdown measures has raised new questions about how to preserve production capacity while guaranteeing public health.

Such non-pharmaceutical measures, however, necessarily entail significant social and economic costs, which can only be incurred over limited time horizons. Therefore, to optimize containment strategies which save at possible these costs, it results necessary to adopt a new approaches to mathematical modeling combining both clinical information of the disease and a realistic social structure of the population.

The study of mathematical models to understand the spread of an epidemic represents a consolidated scientific field of applied mathematics. A fundamental step is furnished by the pioneering work of Kermark and McKendrick on compartmental epidemiology [3], which has given rise to a widely used model in which the initial population is divided into susceptible (S), who can contract the disease, infected (I), who have already contracted it and can transmit it, and recovered (R), who are immune or cured of the disease. This model is often mentioned as the SIR model, using the acronym of the categories just introduced.

The hypothesis made by Kermack and McKendrick is that of *homogeneous mixing*: it is assumed that every individual has the same probability of contacting any other individual in the population. It is understandable that this hypothesis is unrealistic since the intensity and nature of contact with people from our family, our workplace, school class, or a group of friends are very different, and on the other hand we rarely come into contact with those who live in different social realities.

For these reasons, starting from the original SIR model, more articulated models have been examined, capable of representing the population more realistically, dividing it in more detail. Several models are currently able to consider characteristics related to the mobility or age of the subjects, together with further sociological factors that determine the social interactions that can favor the transmission of the infection [1, 2, 4–6].

Nevertheless, the accuracy of a predictive model, at least in the first phase of the spread of a pandemic, is very difficult to assess. A problem that cannot be ignored is the uncertainty present in the official data provided by different countries in relation to the number of infected people. The heterogeneity of the procedures used to carry out the tests of positivity to the disease, the delays in recording and communicating the results and the high percentage of asymptomatic patients make the construction of predictive scenarios affected by high uncertainty. In general, the actual number of infected and recovered people is typically underestimated, causing delays in the implementation of

public health policies in the face of the spread of epidemic fronts.

A few months after the outbreak of the COVID-19 epidemic, and in the face of containment policies operated by governments, we now have a lot of data available [7,8], which allow us to understand how much and in how the measures of social distancing have succeeded in the containment. Starting from these data, an essential aspect for the development of a predictive mathematical model is to build it in such a way as to accurately take into account the influence of social contacts in the contagion mechanism, and which, tested on the various measures of containment operated by different countries, can provide a correct answer in different situations.

Contacts between individuals occur according to interaction matrices that depend on the age of the individuals in the social activity considered (for example, school, work, family). Extensive studies have been carried out on this aspect. Among these, the recent works [9, 10] provided an updated and complete picture of the statistical distribution of the daily social contacts of the population by age, sex, profession and more in various countries. The social contact matrices are very similar for European and non-European countries. These matrices show a peak of contacts in correspondence to people of school age, and of working age. Furthermore, two contact strips are highlighted between people of different ages, corresponding to the interactions between family members.

Taking into account the homogeneity of the social contact matrices with respect to different countries [10], a statistical model in which the populations of susceptible (S), infected (I) and recovered (R) are characterized by the number of daily social contacts has been proposed [11]. The resulting SIR-type model (in brief S-SIR or Social SIR) is in this case characterized by an *inhomogeneous mixing*. Unlike the classical model, the probability of contagion is considered proportional to the average of the contacts of the individuals in the population, and the increase of the infected is directly proportional to the product between the average of the daily contacts of the populations of infected and susceptible. The probability of contact is now characterized by a function that takes into account memory effects linked to the percentage of infected over a limited time span.

The model is further based on the hypothesis that the time scale that controls social contacts is much faster than the time scale on which the pandemic develops, a very reasonable hypothesis on which, among other things, the government provisions aimed at changing social contacts via lockdown policies. The characteristic parameters of the model are obtained starting from the statistical distribution of contacts observed in [9].

In [11], the output of the new model has been tested resorting on public data made available on the web, which allowed to evaluate accurately the key function characterizing the evolution of the epidemic in terms of the amount of social contacts of infectious people. These experiments showed a good agreement in the evolution of compartments predicted by the model with the data relative to the containment of the epidemic consequent to the lockdown policies in various countries.

These first results encouraged further study. This has been made possible by a joint effort of two groups, one from the University of Pavia, the other from the the Health Protection Agency (ATS) of the province area of Pavia, who analyzed the data relative to the COVID-19 infection starting from the early appearance, in February 2020, to extract from them the correct values of the parameters to be inserted into the S-SIR model developed in [11]. These data are relative to a population of about half a million of people in Northern Italy, a population which is sufficiently large to give stable statistics about diffusion and recovery rates in terms of age and sex.

By means of this joint work, it has been possible to improve the original model developed in [11],

and to use the new version to follow at best the numbers of the compartments in the pandemic. A better adherence of the model to real situations, and a prediction in the short term of the availability of places for people to be hospitalized, has been obtained by resorting to the available detailed data. This allowed a statistical but coherent knowledge both of the percentage of people to be hospitalized in relation to the population of infectious, and of the recovery time of hospitalized people.

2. The Social SIR model

One of the main goal of the mathematical models of infectious diseases is the possibility to evaluate control and prevention strategies by comparing their cost with effectiveness, and to support public health decisions [12, 13]. With the sudden spreading of COVID-19 epidemic, and the consequent containment measures adopted to limit its diffusion, special attention was paid by the scientific community to previous researches devoted to estimate the distribution of contacts between individuals (cf. [9, 14–16] and the references therein).

Starting from the analysis of [9], a mathematical framework to connect the distribution of social contacts with the spreading of a disease in a multi agent system was recently developed in [11]. The new model has been obtained by integrating the epidemiological dynamics given by the classical SIR compartmental model [5, 17, 18] with an approach based on kinetic equations, which furnish a detailed mechanism for the formation of social contacts at a rapid scale.

Since this model will be the starting point of the forthcoming analysis, let us briefly recall it with some details. Given a population of individuals, we denote by $f_S(x,t)$, $f_I(x,t)$ and $f_R(x,t)$ the distributions at time $t > 0$ of the number $x \geq 0$ of social contacts of susceptible (S), infected (I), and recovered (R) individuals, respectively. The distribution of contacts of the whole population is then recovered as the sum of the three distributions

$$f(x,t) = f_S(x,t) + f_I(x,t) + f_R(x,t).$$

Assuming for simplicity that the population in the system remains constant, the total distribution of social contacts is assumed to be a probability density for all times $t \geq 0$

$$\int_{\mathbb{R}_+} f(x,t)\,dx = 1.$$

The knowledge of the functions $f_J(x,t)$, $J \in \{S,I,R\}$ allows to compute all relevant observable quantities. The integrals

$$J(t) = \int_{\mathbb{R}^+} f_J(x,t)\,dx, \quad J \in \{S,I,R\} \tag{2.1}$$

denote the fractions at time $t \geq 0$ of the three sub-populations. Analogously we can compute moments, defined, for any given constant $\alpha > 0$ by

$$m_J^\alpha(t) = \int_{\mathbb{R}^+} x^\alpha f_J(x,t)\,dx, \quad J \in \{S,I,R\}.$$

The mean values, corresponding to $\alpha = 1$, are denoted by $m_J(t)$, $J \in \{S,I,R\}$.

The evolution of the densities is built in [11] by assuming that the various classes in the model could act differently in the social process constituting the contact dynamics. The kinetic model then follows combining the epidemic process with the contact dynamics, as modeled in [19]. This gives the system

$$\begin{aligned}
\frac{\partial f_S(x,t)}{\partial t} &= -K(f_S, f_I)(x,t) + \frac{1}{\tau} Q_S(f_S)(x,t) \\
\frac{\partial f_I(x,t)}{\partial t} &= K(f_S, f_I)(x,t) - \gamma f_I(x,t) + \frac{1}{\tau} Q_I(f_I)(x,t) \\
\frac{\partial f_R(x,t)}{\partial t} &= \gamma f_I(x,t) + \frac{1}{\tau} Q_R(f_R)(x,t)
\end{aligned} \tag{2.2}$$

where $\tau \gg 1$ is a relaxation time, $\gamma > 0$ is the recovery rate while the transmission of the infection is governed by the function $K(f_S, f_I)$, the local incidence rate, expressed by

$$K(f_S, f_I)(x,t) = f_S(x,t) \int_{\mathbb{R}^+} \kappa(x,y) f_I(y,t)\, dy. \tag{2.3}$$

According to the literature on the subject [5], in Eq (2.3) the contact function $\kappa(x,y)$ is a nonnegative function growing with respect to the number of contacts y of the population of infected, and such that $\kappa(x,0) = 0$, which expresses the fact that the epidemic can not spreading in absence of infected people. Thus, the evolution of the epidemic depends heavily on the shape of the function $\kappa(\cdot,\cdot)$ used to quantify the rate of possible contagion in terms of the number of social contacts between the classes of susceptible and infected.

Note that the evolution of the mass fractions $J(t)$, $J \in \{S, I, R\}$ then obeys to the classical SIR model by choosing $\kappa(x,y) \equiv \beta > 0$ [5], thus considering the spreading of the disease independent of the intensity of social contacts. In [11] the contact function has been considered of the form $\kappa(x,y) = \beta xy$, leading to an incidence rate of the form

$$K(f_S, f_I)(x,t) = \beta\, x f_S(x,t) m_I(t), \tag{2.4}$$

which implies an incidence of the contagion proportional to the product of the mean values of the populations of infectious and susceptible.

In system (2.2) the operators Q_J, $J \in \{S, I, R\}$ describe the formation of the social contacts of the populations, and are constructed to rapidly push the densities $f_J(x,t)$, $J \in \{S, I, R\}$ towards the equilibrium densities of daily contacts, which are given by suitable Gamma density functions. This choice is consistent with the results of empirical studies based on a precise analysis of a heterogeneous sample of French population, which established that Gamma densities are the most suitable to describe the daily number of contacts in the pre-pandemic scenario [9]. Further, a theoretical basis to the formation of this equilibrium profile has been recently done in the recent paper [11]. A simple choice for the relaxation operators is given by the linear BGK-type model [20]

$$Q_J(f_J)(x,t) = f_J^\infty(x,t) - f_J(x,t)$$

where the Gamma-like equilibrium density is given by

$$f_J^\infty(x,t) = \frac{J(t)\, x^{\nu-1}}{\Gamma(\nu)\, x_J(t)^\nu} \exp\left\{-\frac{\nu x}{x_J(t)}\right\}, \quad J \in \{S, I, R\}$$

where $\nu > 0$ is a parameter measuring the heterogeneity of the population, and

$$x_J(t)J(t) = m_J(t).$$

We remark that the equilibrium density is built to have the same mass and mean value of the function $f_J(x,t)$, $J \in \{S, J, R\}$. Another choice is given by relaxation operators in the form of Fokker-Planck-type operators. We point the interested reader to [11] for a detailed discussion.

Integrating the equations of the system (2.2) with respect to the contact variable x leads to the system for the evolution of the proportions $J(t)$ defined in Eq (2.1), $J \in \{S, I, R\}$

$$\begin{aligned} \frac{\partial S(t)}{\partial t} &= -\beta\, m_S(t) m_I(t), \\ \frac{\partial I(t)}{\partial t} &= \beta\, m_S(t) m_I(t) - \gamma I(t), \\ \frac{\partial R(t)}{\partial t} &= \gamma I(t). \end{aligned} \tag{2.5}$$

Unlike the classical SIR model, system (2.5) is not closed, since the evolution of the mass fractions $J(t)$, $J \in \{S, I, R\}$ depends on the mean values $m_J(t)$.

Hence, fixing the mass of the large time contact distribution of infectious individuals equal to $I(t)$, its momentum is given by

$$m_I(t) = \bar{x}_I I(t) H(I(t)). \tag{2.6}$$

Analogously, if the mass of the large time contact distribution of susceptible individuals is $S(t)$, its momentum is given by

$$m_S(t) = \bar{x}_S S(t) H(I(t)).$$

In Eq (2.6) the constant $\bar{x}_I$ denotes the mean number of social contacts of infectious, while the function H captures the actual intensity of social contacts, in dependence of the amount, at time $t \geq 0$, of infected people. Proceeding as in classical dynamics of fluids, we obtain the closure

$$\begin{aligned} \frac{\partial S(t)}{\partial t} &= -\bar{\beta}\, S(t)\, I(t)\, H^2(I(t)), \\ \frac{\partial I(t)}{\partial t} &= \bar{\beta}\, S(t)\, I(t)\, H^2(I(t)) - \gamma I(t), \\ \frac{\partial R(t)}{\partial t} &= \gamma I(t). \end{aligned} \tag{2.7}$$

In system (2.7) we defined $\bar{\beta} = \beta \bar{x}_I \bar{x}_S$. This identifies the classical transmission parameter of the SIR model, where however now the difference is that this quantities are not postulated but instead derived starting from microscopic considerations.

The key point in [11] was to quantify in a proper way the function H, describing the action of the variation of the intensity of social contacts on the evolution of the epidemic. For a given constant $N \gg 1$, the function H was assumed to be the decreasing function

$$H(r) = \frac{1}{\sqrt{1 + Nr}}, \quad 0 \leq r \leq 1, \tag{2.8}$$

which describes a possible way in which, in presence of the spread of the disease, the susceptible and the infected population tend to reduce the mean number of daily social contacts $\bar{x}_J$, $J \in \{S, I\}$. This choice produces a SIR model with global incidence rate

$$D(S, I)(t) = \int_{\mathbb{R}^+} K(f_S, f_I)(x, t)\, dx = \bar{\beta} S(t) \frac{I(t)}{1 + N I(t)}, \tag{2.9}$$

that fulfils all the properties required by the non-linear incidence rates considered in [3].

In addition to the form in Eq (2.8), the following function was considered in [11]

$$H(t, r) = \frac{1}{\sqrt{1 + Nr(t) \displaystyle\int_0^t r(s) ds}}, \quad 0 \le r \le 1. \tag{2.10}$$

This function satisfies the same properties of Eq (2.8) in terms of the incidence rate requests detailed in [3] and takes into account memory effects on the population's behavior. In fact, it is reasonable to assume that both the restrictions operated by governments and the individual behavior are based not only in terms of the daily contacts to answer to the actual pandemic situation as expressed by Eq (2.8), but considering the evolution of the epidemic over a certain period of time.

The spreading of the COVID-19 pandemic in presence of lockdown measures has been tested in [11] through system (2.7), where both the form of the function H and the values of the parameters $\bar{\beta}$ and γ have been computed by resorting to public available data. As we shall see, the possibility to have at disposal more detailed data, will lead to an improvement of the Social SIR, and to and its better adherence to epidemic reality. Analogous results have been obtained through an experimental approach in [21] in the region of Wuhan and Shanghai.

3. Dataset description

The Social SIR model has been fitted and tested on the accurate dataset collected by the ATS, representative of a local but extensive territory such as the province of Pavia. The dataset collects all the confirmed COVID-19 cases from February 20, 2020, to November 15, 2020, covering a wide period characterized by different behaviors in the spreading of the disease. For each confirmed case, all the relevant dates related to the personal evolution of the disease are reported: date of the first positive swab test, recovery date, death date, admission and discharge date in case of hospitalization. Age and sex of each COVID-19 patient recorded in the dataset are also available. This information proved essential to make the best estimate of the parameters of the Social SIR model leading to an accurate picture of the spread of COVID-19 infection in the province of Pavia.

In Figure 1 the number of confirmed COVID-19 cases per day is displayed, for a total of 15,768 infectious individuals and 1415 cumulative deaths during the considered period. Figure 1 clearly shows two different crucial periods: the first one represents the initial outbreak of the disease, and the second one, with the restart of the activities.

The detailed information collected in the ATS dataset allows to extrapolate the time to viral clearance (TVC), corresponding to the time from the first positive test to the second negative test before October 12nd or the first negative test after October 12nd, in agreement with the policies of the

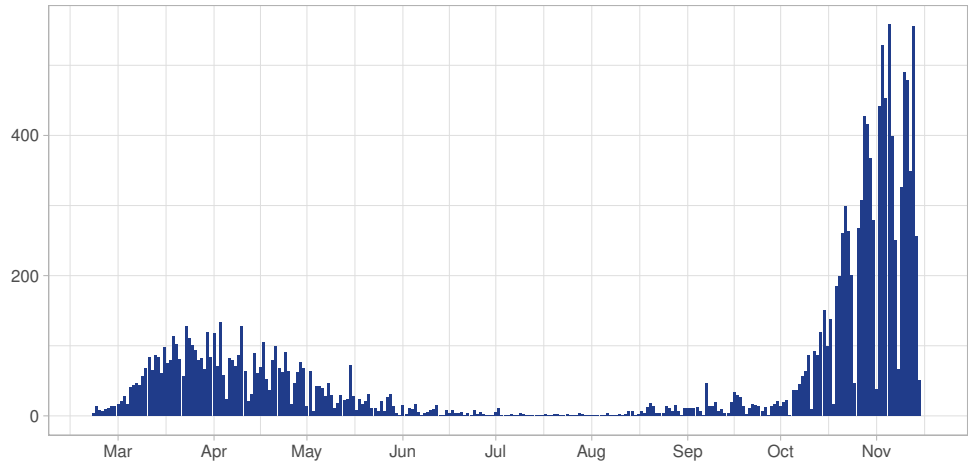


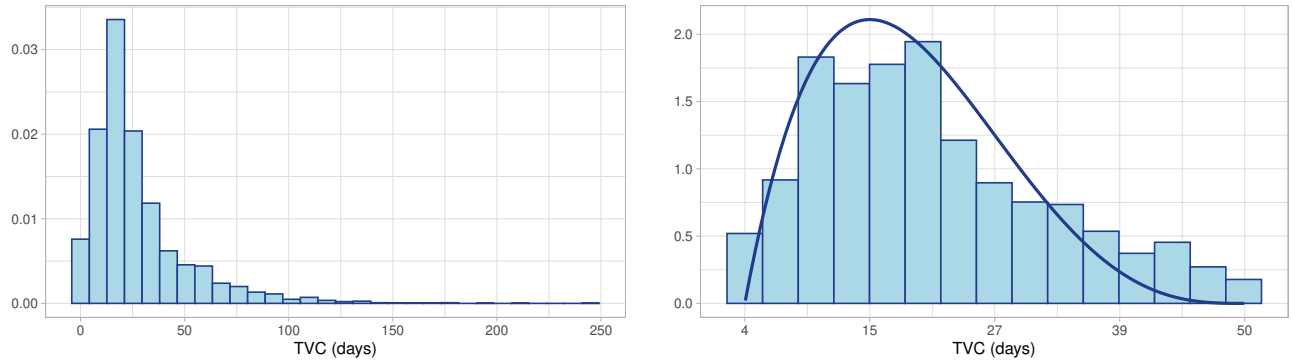
Figure 1. Confirmed COVID-19 cases in the province of Pavia at November 15, 2020.

Ministry of Health. Furthermore, since we considered in the recovered compartment also deceased patients, the computation of the TVC considers in this case the date of the patient's death. Figure 2a shows the global distribution of TVC computed in days with a minimum value of 0 and a maximum value of 250. The values 0 and greater to 100 may be due to anomalies in the registration of swab tests in the period of first emergency. To deeper study the type of distribution of the TVC we focus the analysis on the interval between 4 and 50 days, as reports in Figure 2b. This interval includes the 82% of the recovered patients. The corresponding distribution has been estimated by testing a Beta density function on the empirical distribution of the available data on the bounded interval [4,50]. The method of Maximum Likelihood Estimation has been implemented to estimate the relevant parameters, and to realize that $B(2,4)$ is the Beta distribution capable to best approximate the registered times to viral clearance.

According to several studies, the TVC during the early phase of the epidemic can approximately span from 10 to 24 days (cf. [2,22,23] for further discussions). The present analysis, based on the data of the province of Pavia, is substantially in agreement with the aforementioned works. In Figure 3 we present the boxplots for TVC by month. We may observe how the presence of outliers appears strongly reduced after July 2020, whereas the median TVC remains stable in the considered period.

Since the age component influences significantly the morbidity and mortality of COVID-19 patients [24], we have preliminarily investigated the dependency of the recovery of patients from the age variable $a \in [0, 100]$. Figure 4a shows that heterogeneities in registered time to viral clearance with respect to age exist in terms both of range values and median value. In particular, the median time to viral clearance in days increases with the age of patients, from a minimum value of 15 days if $a \leq 18$ till a value of 25 days if $a \geq 75$.

This difference is presented in Figure 4b where the bootstrap confidence intervals (confidence level equal to 95%) for the median values of the recovery rates are estimated for each value of age reported on x axis. The analysis on median values has been preferred with respect to the mean values since recovery distributions for each group age are strongly positive skewed, as displayed in Figure 4a, and the median value is a more robust statistic with respect to outliers.



(**a**) Global distribution of times to viral clearance.

(**b**) Distribution of times to viral clearance excluding outliers and the fitted Beta density function .

Figure 2. Time to viral clearance (in days).

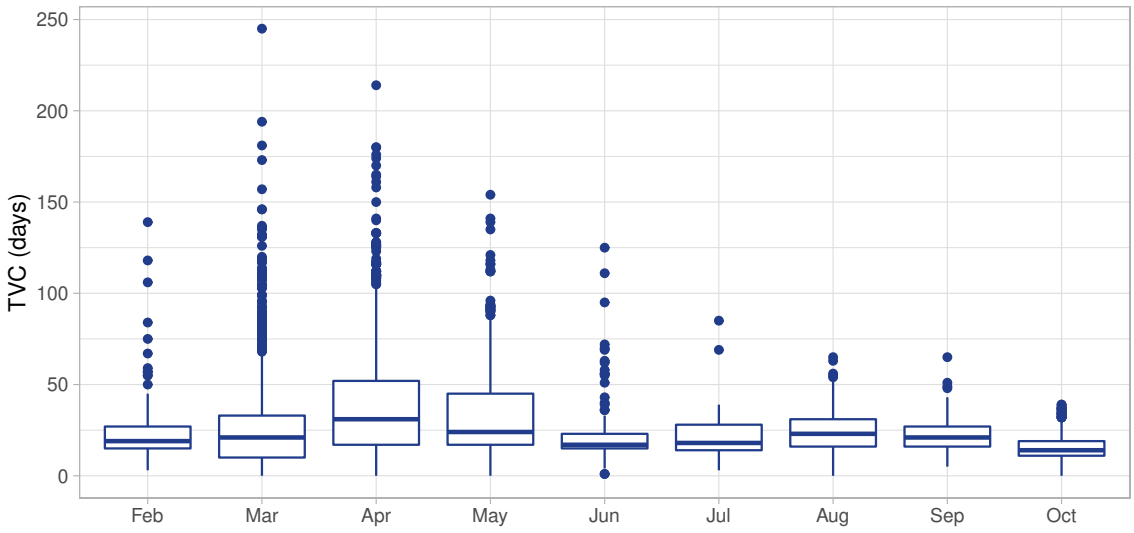
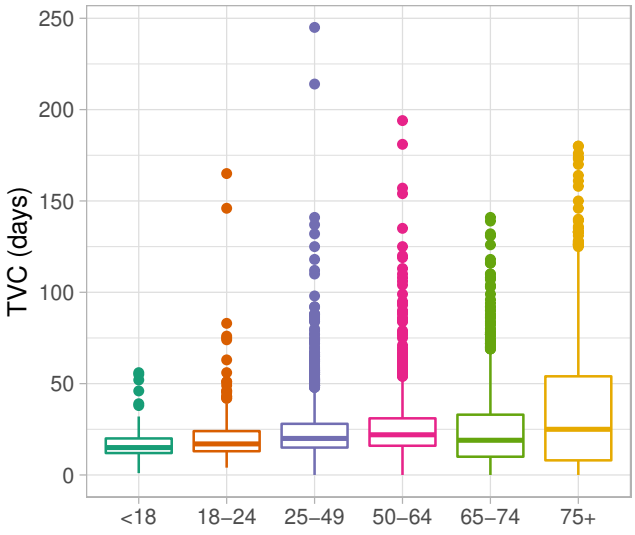


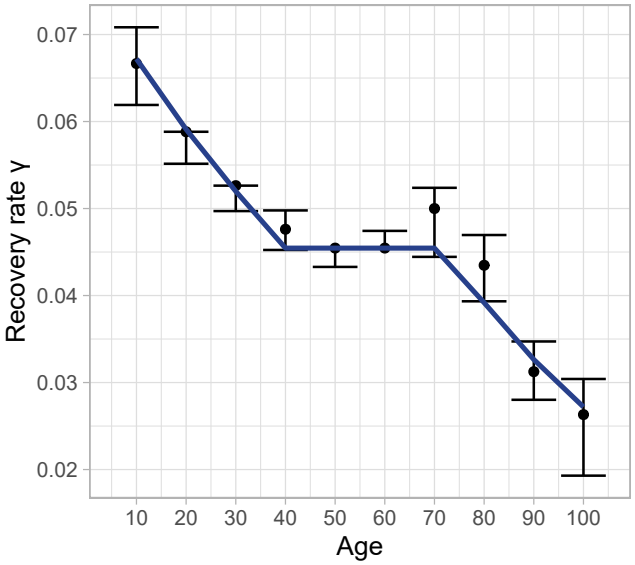
Figure 3. Distribution of times to viral clearance (in days) for each month registered in the Province of Pavia. The month represents the first swab test date.

Table 1. Estimated parameters of the age dependent recovery rate (3.1).

Age interval	p_i	q_i
$i = 1$	0.0764	0.0128
$i = 2$	0.0455	0.000
$i = 3$	0.169	0.0183



(**a**) Distribution of recovery in days for each age group.



(**b**) Confidence intervals for the median of the recovery rates and interpolation with a piecewise function.

Figure 4. Dependencies between recovery rates and age.

Resorting to these results the constant recovery rate γ of the original continuous model (2.7) can be substantially improved. An age-dependent recovery rates $\gamma(a)$ in the form of the following piecewise function

$$\gamma(a) = \begin{cases} p_1 e^{-q_1 a} & 0 \le a \le 40 \\ p_2 e^{-q_2 a} & 40 < a \le 70 \\ p_3 e^{-q_3 a} & a > 70. \end{cases} \tag{3.1}$$

is depicted in Figure 4b. The parameters p_1, p_2, p_3 and q_1, q_2, q_3 can be obtained by solving a least square optimization problem in the three introduced subcategories. The results of the optimization problem are reported in Table 1.

Recovery distribution has been also analyzed comparing different months of the considered period in order to understand if significant changes have occurred during the evolution of the epidemic. Figure 3 displays a particular trend in the first months of the outbreak (March, April, May and June) characterized by a right-skewed distribution and a wide range of times to viral clearance. From July to September the distributions become more and more stable around the median value of 25 days.

This significant difference, between the first months of the epidemic and the following ones, can be explained considering that collections of data of first period have been affected by lot of imprecisions

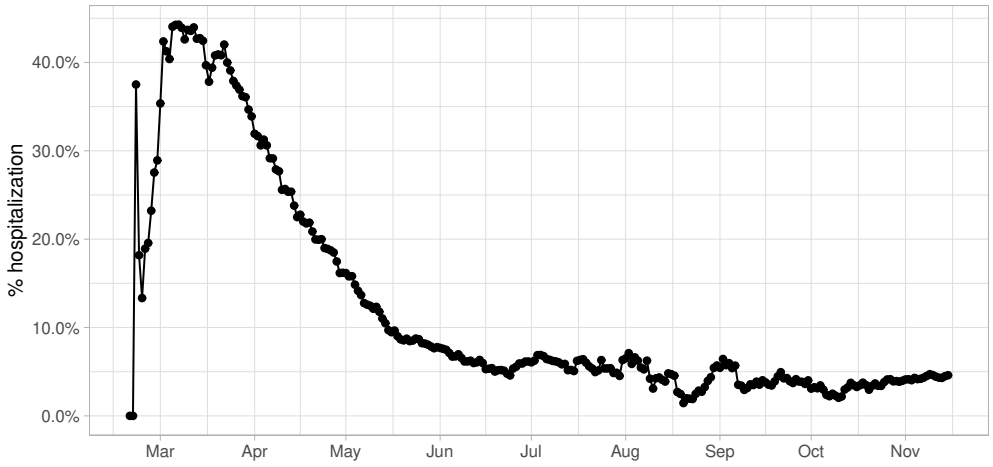


Figure 5. Percentage of hospitalized COVID-19 patients per each day.

due to the overwhelming activities which have involved the ATS of Pavia because of this new emergency.

Finally, we consider the number of hospitalizations in terms of admission and discharge dates in the hospitals of Pavia. This allows to estimate the daily percentage of hospitalized patients over all the daily confirmed cases. The percentage trend is represented in Figure 5. This value shows a peak during the first outbreak of the epidemic, between March and April, achieving 50% and becomes more and more constant starting from the beginning of June, with an average value equal to 4.7%. It is reasonable to consider this last one as a more accurate estimate of the real hospitalization rate, since data collected from March to May are altered by the low number of swab tests performed during this period.

4. Calibration of S-SIR model and forcastings

In this section we perform the estimation of epidemiological parameters of the introduced S-SIR model taking into account the ATS dataset. It is worth to mention that the estimation of such parameters for compartmental models is generally a difficult task and is strongly related on the assumptions of the adopted model. A comparison between several available models and methods has been recently presented in [25]. Furthermore, available epidemiological data represent typically a lower bound of the real diffusion of the epidemic since according to recent results of the serological campaigns promoted in Italy around 80% of infected are asymptomatic [26]. Another source of uncertainty is given by the limited testing capacity which is particularly evident at the beginning of the epidemic. Therefore data are partial and heterogeneous with respect to their assimilation making the fitting problem challenging. We mention in this direction the detailed discussions in [27–30].

Several approaches have been developed to infer the time-course of infections of COVID-19 in Italy from available aggregated data through specific compartmental analysis of the population. In particular, a common trait in existing literature relies in considering the compartment of hospitalized or infected acutely symptomatic whose evolution depends on specific choice of parameters [1, 2, 31].

Nevertheless, we have observed in Figure 5 how the percentage of hospitalizations in the Province of Pavia remain quite stable from June 2020 up to November 15 around the 5% of the current cases. Interestingly enough, from February to April 2020 the registered hospitalizations increased a lot in the reference area with a peak of more than 40% of the current cases. Hence, the hospitalizations decreased constantly from April to June 2020. This fact may be explained by the limited testing capacity in the early phases of the epidemic. In agreement with [26] the real number of infected has been largely underestimated.

We will predict the number of hospitalized patients using the percentage of hospitalizations estimated from the last available month before the time interval considered for the prediction. As already pointed out, we will show that this percentage appears stable from July to October.

4.1. Extrapolation of the contact function

In the following, we present a calibration approach which is based on a strategy with two main optimization horizons related to the pre-lockdown and lockdown time spans. We consider first the time interval $t \in [t_0, t_\ell]$, being t_ℓ the day in which lockdown started in Italy and t_0 the day in which reported cases hit 30 units. The lower bound has been imposed to reduce the limited testing capacity in the early phase of the infection.

We estimate the recovery rate $\gamma > 0$ from data by computing the median value of the empirical recovery rates distribution, $\gamma = 0.0476$ corresponding approximatively to a median time to viral clearance of 21 days. Hence, to determine β, we solve a least square problem based on the minimization of a cost functional $\mathcal{J}$ which takes into account the sum of two relative $L^2([t_1, t_2])$ norms over the time horizon $[t_1, t_2]$. In particular, we consider a cost functional taking into account the difference between the reported number of infected and the reported total cases, i.e., $\hat{I}(t)$ and $\hat{I}(t) + \hat{R}(t)$, and the evolution of $I(t)$ and $I(t) + R(t)$ given by the S-SIR model (2.7) with $H \equiv 1$ and $t \in [t_1, t_2]$. In details, we solve the constrained optimization problem

$$\min_{\beta \in [0,1]} \mathcal{J}(\hat{I}, \hat{R}, I, R), \qquad t \in [t_0, t_\ell], \tag{4.1}$$

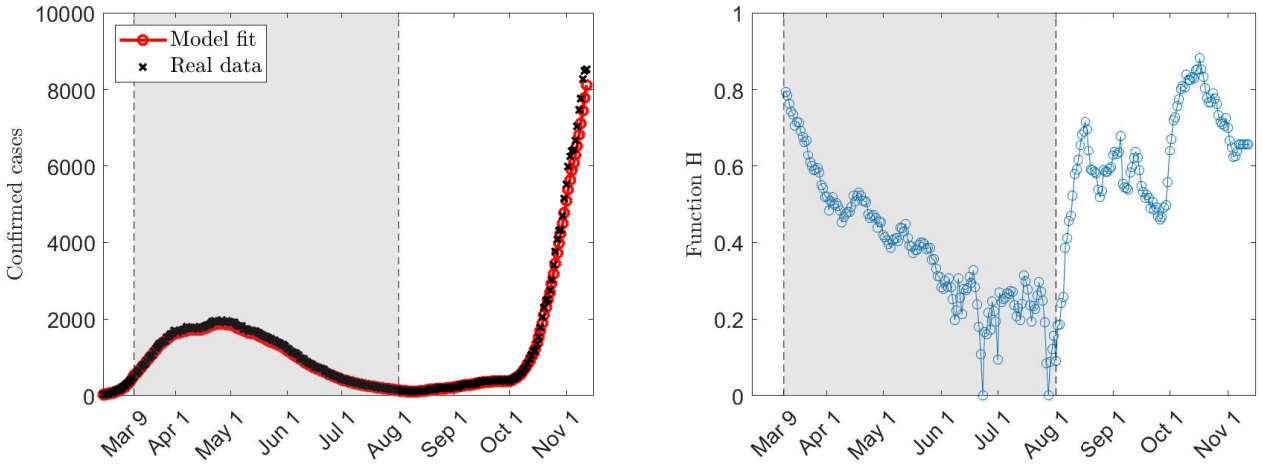
subject to the dynamics in Eq (2.7). In Eq (4.1) the cost functional $\mathcal{J}$ is a convex combination between two relative L^2 norms

$$\mathcal{J}(\hat{I}, \hat{R}, I, R) = \mu \frac{\|\hat{I}(t) - I(t)\|_{L^2([t_1,t_2])}}{\|\hat{I}(t)\|_{L^2([t_1,t_2])}} + (1-\mu) \frac{\|\hat{I}(t) + \hat{R}(t) - I(t) - R(t)\|_{L^2([t_1,t_2])}}{\|\hat{I}(t) + \hat{R}(t)\|_{L^2([t_1,t_2])}},$$

with $\mu = 10^{-2}$. We solved the introduced optimization problem confronting different integration methods and we obtained $\beta = 0.218$.

Once the epidemiological parameters have been estimated in the pre-lockdown time interval we can proceed with the estimation of the shape of the contact function H in the time span $[t_\ell, T]$. In order to reduce the fluctuations due to possible delays in the registration procedure, we solve the corresponding least square problems for a sequence of time steps t_i in $[t_i + \kappa_r, t_i + \kappa_\ell]$ where $\kappa_r, \kappa_r \geq 1$ are integers chosen to cover a window of seven days. We considered $\kappa_r = 2$ and $\kappa_\ell = 5$ for regularization along one week of available data. To this end, we considered a second optimization problem

$$\min_{H \in [0,1]} \mathcal{J}(\hat{I}, \hat{R}, I, R), \qquad t \in [t_i + \kappa_r, t_i + \kappa_\ell], \tag{4.2}$$



(**a**) Real curve of confirmed cases and fitting of the Social SIR model.

(**b**) Punctual estimate of contact function H.

Figure 6. Fitting of the Social SIR model on the ATS dataset.

where the parameters β, γ have been determined as a result of the optimization problem in Eq (4.1). These optimization problems have been solved testing different optimization methods in combination with RK4 integration method of the system of ODEs.

The available data start on February 20, 2020, when moderate social restrictions were already enforced by the national government, and the lockdown started on March 9, 2020. In Figure 6 we present the evolution of the current confirmed cases together with the result of the described fitting procedure. The punctual evolution of the contact function $H(t)$, $t \in [t_\ell, T]$, is presented in the left column. We highlighted a shaded gray zone to indicate the first wave of the infection for which we can clearly observe that the contact function can be coherently approximated through a decreasing function.

On the other hand, the beginning of the second wave of infections can be detected in the first days of August since the function H returned rapidly to its early pandemic values remaining quite stable till November 15th.

4.2. Forecasts of hospitalizations

In this Section we present the prediction results furnished by the improved S-SIR model for expected hospitalizations due to COVID-19 acute symptoms. In particular, we will take into account the information on the times to viral clearance presented in Section 3 where we have observed that the experimental data are well approximated with a Beta distribution $p(z)$ of the form

$$\frac{1}{\gamma(z)} = h_0 z + h_1, \qquad z \sim B(2,4), \tag{4.3}$$

and $h_0 = 46$, $h_1 = 4$. In the following we will indicate the support of z with $\Gamma = [4, 50]$.

The S-SIR model with stochastic recovery rate can be obtained in the same theoretical setting

described in Section 2 and reads

$$\begin{aligned}
\frac{\partial S(z,t)}{\partial t} &= -\bar{\beta}\, S(z,t)\, I(z,t)\, H^2(z, I(z,t)),\\
\frac{\partial I(z,t)}{\partial t} &= \bar{\beta}\, S(z,t)\, I(z,t)\, H^2(z, I(z,t)) - \gamma(z) I(z,t),\\
\frac{\partial R(z,t)}{\partial t} &= \gamma(z) I(z,t).
\end{aligned} \tag{4.4}$$

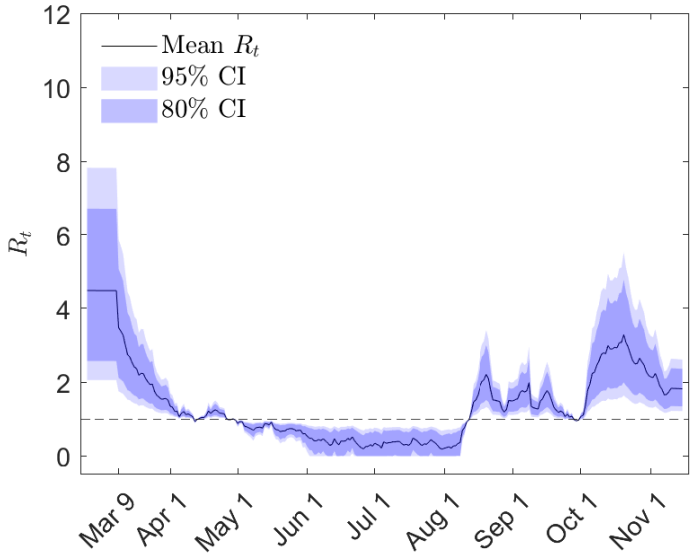


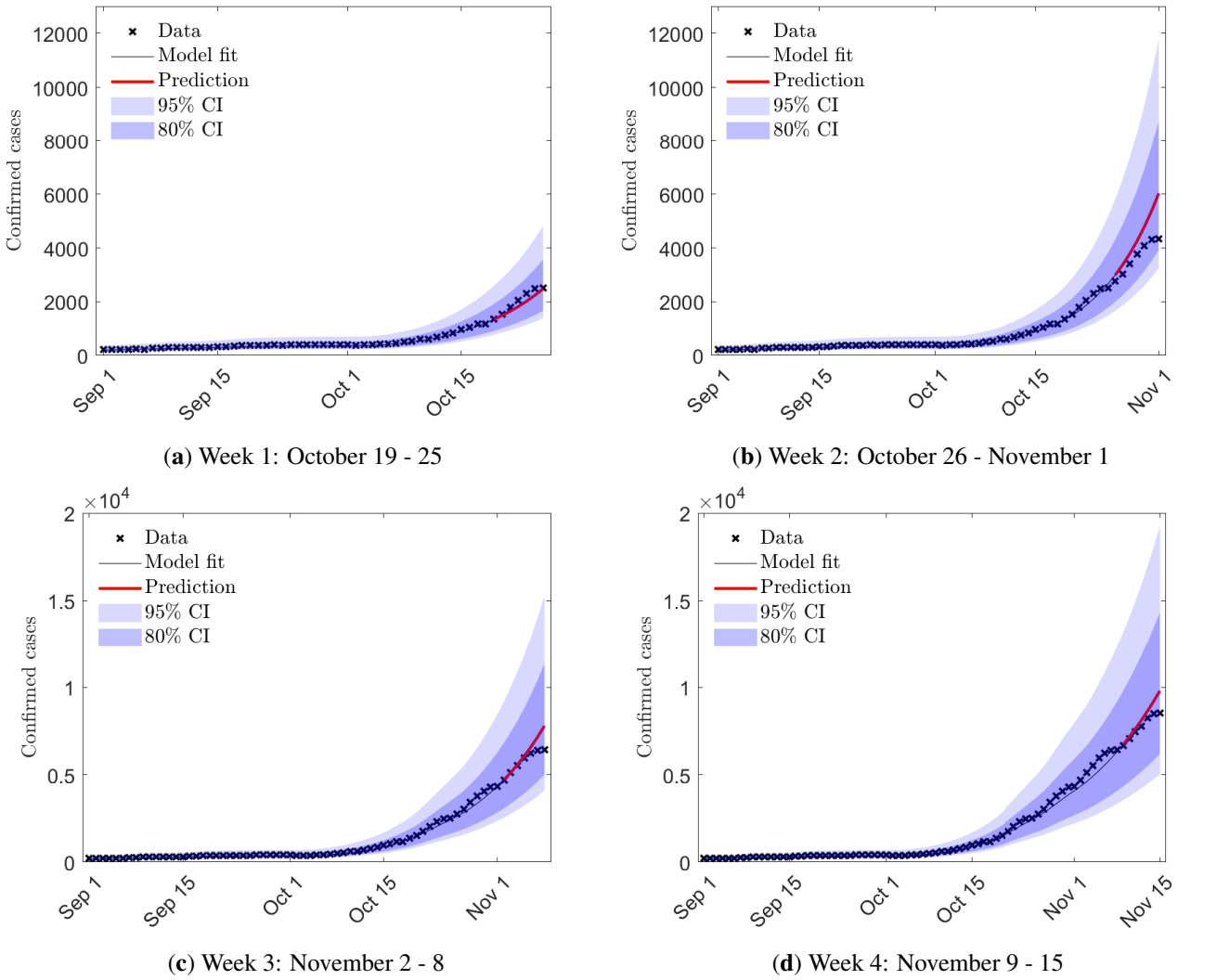
Figure 7. Evolution of $\mathbb{E}[R_t]$ of the S-SIR model (4.4) with $\gamma(z)$ defined in Eq (4.3).

In order to handle efficiently the introduced stochasticity in the dynamics we will adopt a stochastic collocation approach based on stochastic Galerkin methods (see [32] for an introduction and [27, 30] for their application to epidemiological dynamics). These methods allow to handle efficiently the stochasticities of a system when information on the uncertainties distribution are available thanks to the fast convergence properties holding under suitable regularity assumptions.

We construct the sample $\{\gamma_0, \dots, \gamma_M\}$ from $\gamma(z)$. The samples are obtained in a collocation approach using Gauss-Jacobi polynomials with $M = 10$. For each fixed value γ_j, $j = 0, \dots, M$, we solve the optimization problem (4.1) subject to the dynamics in Eq (2.7) to obtain the value $\beta_j > 0$. Hence, we estimate the contact function $H_j(t)$, $j = 0, \dots, M$, in time by solving the subsequent optimization problem (4.2) in the time interval $[t_i + k_\ell, t_i + k_r]$, with $\kappa_r = 2$ and $\kappa_\ell = 5$, and where $\gamma = \gamma_j, \beta = \beta_j$ are fixed.

In Figure 7 we report the expected value of R_t together with the 95% and 80% confidence intervals with respect to the stochastic parameter z. The results show that, in presence of the marked reduction of social contacts enforced by the lockdown measures of the government, the reproduction number R_t has been drastically reduced and its expected value fell below one in the first week of April. Then, the observed R_t remains stably below unity from May to the first week of August.

In Figure 8 we present the dynamics of the confirmed cases with respect to the available data in four successive weeks from October 15, 2020 to November 15, 2020. In particular, we performed the



(**a**) Week 1: October 19 - 25

(**b**) Week 2: October 26 - November 1

(**c**) Week 3: November 2 - 8

(**d**) Week 4: November 9 - 15

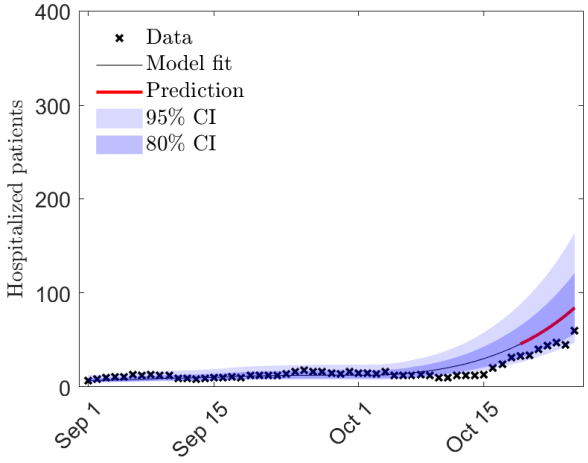
Figure 8. Prediction of the confirmed cases over four weeks from October 15 to November 15 in the province of Pavia. The results are based on S-SIR model with recovery rate of the form of Eq (4.3) .

fitting procedure of the model up to the week of prediction. We fixed for the whole prediction horizon $H_j(t) = H_j(T)$, $t \in [T, T+7]$. We highlight in red the expected value of the predicted cases, i.e., $\mathbb{E}[I(z,t)] = \int_\Gamma I(z,t)p(z)dz$. Together with the expected trends we plot the confidence intervals (CI) that are obtained by imposing

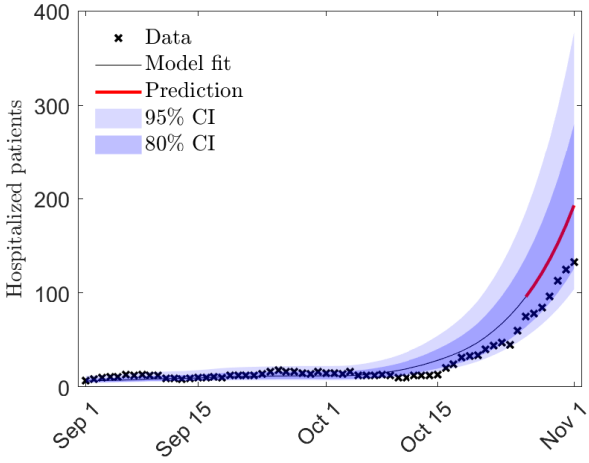
$$\text{Prob}\,[I_\ell \le I(z,t) \le I_r] = 1 - c,$$

and $c = 0.05$ for the 95% CI and $c = 0.2$ for the 80% CI. For any given week of interest, the previous seven days have been neglected in the fitting procedure, and then plotted together with predictions to establish the performance of the modelling setting. We can observe how through S-SIR model with stochastic recovery rate we are capable to catch the epidemic dynamics with a good level of accuracy.

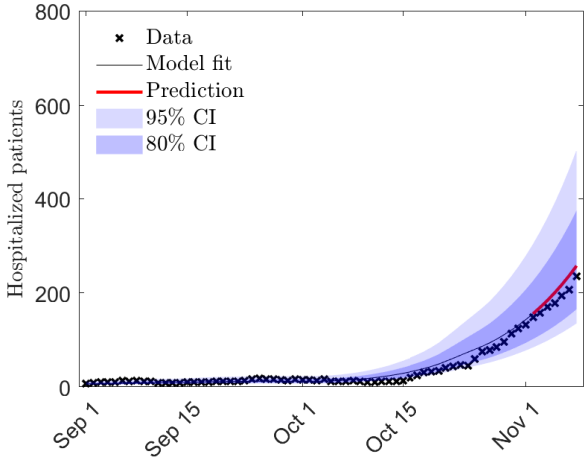
Once a sufficiently precise evolution of the infections has been obtained, we can infer the expected hospitalizations thanks to the preliminary statistical results introduced in Section 3. In details, for



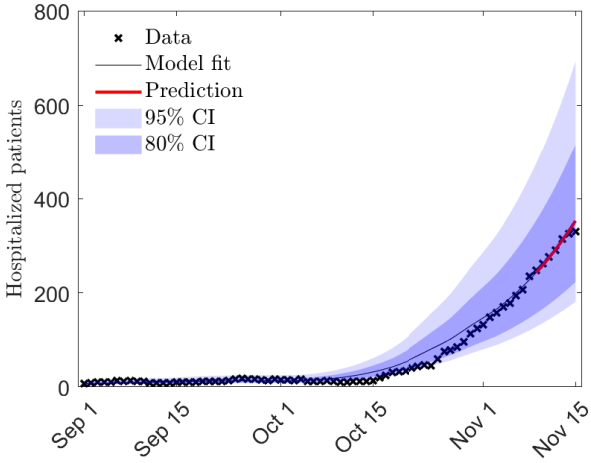
(**a**) Week 1: October 19 - 25, HPP = 3.4%



(**b**) Week 2: October 26 - November 1, HPP = 3.2%



(**c**) Week 3: November 2 - 8, HPP =3.3%



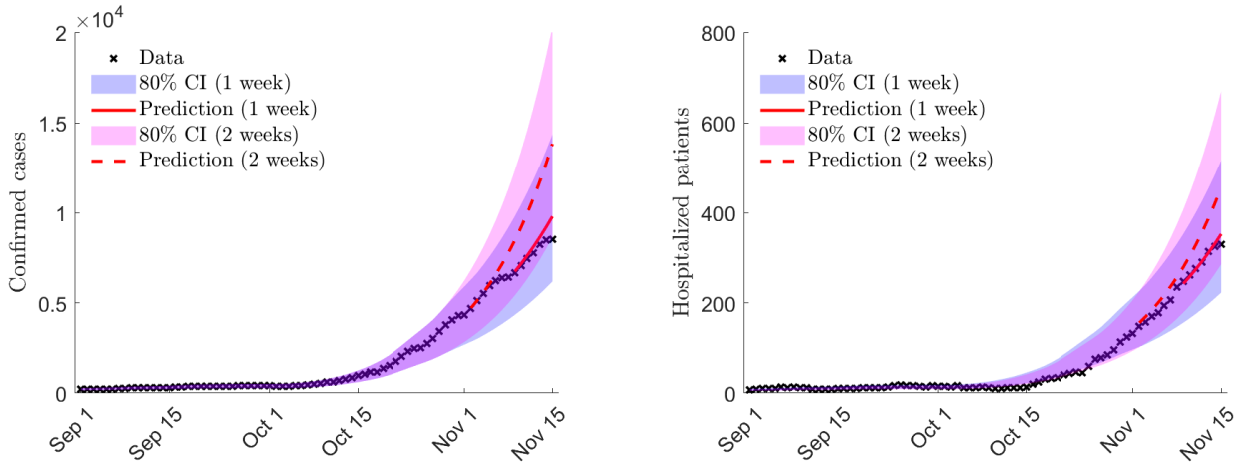
(**d**) Week 4: November 9 - 15, HPP = 3.6%

Figure 9. Seven-days prediction of number of hospitalized individuals based on Social SIR model.

each prediction week, we computed the percentage of hospitalized patients (HPP) taking into account the previous 30 days of observations. Hence, the prediction is obtained by multiplying the predicted cases with HPP. In Figure 9 we present the obtained predictions of hospitalized patients from the S-SIR model with stochastic recovery rate (4.3). As before, we neglected the previous seven days of observations for each week of interest to estimate the parameters of the model. The performance of the model are compatible with the expected hospitalized patients $\mathbb{E}[I(z,t)]$.

We considered a prediction horizon of one week, since a complex phenomenon like the COVID-19 diffusion in Italy embed a huge number of hidden uncertainties and variable factors. We briefly discussed some of them in the previous paragraphs. In Figure 10 we present a comparison between a 1 week prediction horizon, i.e., the S-SIR model is calibrated taking into account data until November 8th 2020, and a 2 weeks prediction horizon, where the calibration is based only on data until November 1st 2020. For the left figure we also used the HPP relative to the week November 2-8, i.e., HPP = 3.3%.

We may observe how we lose some accuracy in a longer prediction horizon even if we are qualitatively able to catch the observed trends.



(**a**) Number of infectious individuals.

(**b**) Number of hospitalized patients.

Figure 10. Comparison between fourteen-days and seven-days predictions (November 2–15) based on Social SIR model.

5. Conclusions

The recent spreading of the COVID-19 epidemic led the scientific community to improve the existing mathematical models of infections spreading by adding to them the possible effects of social contacts about individuals. In this direction, the social S-SIR model introduced in [11] aimed to quantify the effects of the lockdown policies operated by governments to limit the spreading of COVID-19 epidemic in terms of a computable reduction of the contacts. Months after the early phase of the epidemic, the huge amount of data at disposal of the authorities can be fruitfully used to calibrate the parameters of the epidemic mathematical models to furnish a better knowledge of the evolution of infections. In the present paper, by resorting to the ATS dataset relative to the province of Pavia, we were able to correctly evaluate the key parameters of the S-SIR model. In particular, the knowledge of data relative to the recovery rate of infectious allowed to substitute in the S-SIR model the constant recovery rate parameter with a Beta distributed random variable, thus providing a better adherence to real data. Further, a critical inspection of the dataset outlined the importance of considering a novel mathematical model in which, in addition to the social contacts, one has to take into account in a proper way the age variable, mainly responsible of the expected number of hospitalized, and consequently of the possible critical situations in the health system. Nevertheless, in absence of this new type of refined mathematical model, the merging of the evolution of spreading obtained by means of the S-SIR model with a reasoned use of the dataset, led to an accurate output on a one-week time horizon, with a uniform goodness in a period of four weeks, ranging from the middle of October, 2020 to the middle of November, 2020. Based on the prediction performance of this modeling approach, this joint research represents a first step towards a easy-to-handle fruitful numerical approach to the mechanism of control of a local health system during the period of epidemic spreading.

Acknowledgments

This work has been written within the activities of the Agreement 60/2020, Protocol number 154,768 of December 21, 2020, between the Department of Political and Social Sciences of the University of Pavia and the Health Protection Agency (ATS) of the province area of Pavia, with object "Mathematical modeling and statistics for the forecast of the COVID-19 epidemic in the territory of the Province of Pavia". M. Z. and G. T. acknowledge support from the GNFM group of INdAM (National Institute of High Mathematics), from MIUR project "Optimal mass transportation, geometrical and functional inequalities with applications", and by the Italian Ministry of Education, University and Research (MIUR): Dipartimenti di Eccellenza Program (2018–2022)–Dept. of Mathematics "F. Casorati", University of Pavia.

Conflict of interest

The authors declare no conflict of interest.

References

1. G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, et al., Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy, *Nat. Med.*, **26** (2020), 855–860.
2. M. Gatto, E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi, et al., Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures, *PNAS*, **117** (2020), 10484–10491.
3. W. O. Kermack, A. G. McKendrick, Contributions to the mathematical theory of epidemics, *Proc. Roy. Soc. London Ser. A*, **115** (1927), 700–721.
4. G. Dimarco, L. Pareschi, G. Toscani, M. Zanella, Wealth distribution under the spread of infectious diseases, *Phys. Rev. E*, **102** (2020), 022303.
5. H. W. Hethcote, The mathematics of infectious diseases, *SIAM Rev.*, **42** (2000), 599–653.
6. M. Iannelli, F. A. Milner, A. Pugliese, Analytical and numerical results for the age-structured SIS epidemic model with mixed inter-intracohort transmission, *SIAM J. Math. Anal.*, **23** (1992), 662–688.
7. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, *Lancet Infect. Dis.*, **20** (2020).
8. S. Flaxman, S. Mishra, A. Gandy, H. Juliette, T. Unwin, T. A. Mellan, et al., Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries, *Nature*, **584** (2020), 257–261.
9. G. Beraud, S. Kazmercziak, P. Beutels, D. Levy-Bruhl, X. Lenne, N. Mielcarek, et al., The French connection: The first large population-based contact survey in France relevant for the spread of infectious diseases, *PLoS ONE*, **10** (2015), e0133203.
10. K. Prem, A. R. Cook, M. Jit, Projecting social contact matrices in 152 countries using contact surveys and demographic data, *PLoS ONE*, **13** (2017), e1005697.

11. G. Dimarco, B. Perthame, G. Toscani, M. Zanella, Kinetic models for epidemic dynamics with social heterogeneity, preprint, arXiv:2009.01140v1.
12. N. M. Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, D. S. Burke, Strategies for mitigating an influenza pandemic, *Nature*, **442** (2006), 448–452.
13. S. Riley, C. Fraser, C. A. Donnelly, A. C. Ghani, L. J. Abu-Raddad, A. J. Hedley, et al., Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions, *Science*, **300** (2003),1961–1966.
14. J. Dolbeault, G. Turinici, Heterogeneous social interactions and the COVID-19 lockdown outcome in a multi-group SEIR model, *Math. Model. Nat. Pheno.*, **15** (2020), 36.
15. L. Fumanelli, M. Ajelli, P. Manfredi, A. Vespignani, S. Merler, Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread, *PLoS Comput. Biol.*, **8** (2012), e1002673.
16. J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, et al., Social contacts and mixing patterns relevant to the spread of infectious diseases, *PLoS Med.*, **5** (2008), e74.
17. F. Brauer, C. Castillo-Chavez, Z. Feng, *Mathematical Models in Epidemiology*, Springer, New York, 2019.
18. O. Diekmann, J. A. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*, John Wiley & Sons, Chichester, UK, 2000.
19. G. Dimarco, G.Toscani, Kinetic modeling of alcohol consumption, *J. Stat. Phys.*, **177** (2019), 1022–1042.
20. P. L. Bhatnagar, E. P. Gross, M. Krook, A model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems, *Phys. Rev.*, **94** (1954), 511–525.
21. J. Zhang, M. Litvinova, Y. Liang, Y. Wang, S. Zhao, Q. Wu, et al., Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China, *Science*, **368** (2020), 1481–1486.
22. J. Chen, H. Lu, G. Melino, S. Boccia, M. Piacentini, W. Ricciardi, et al., COVID-19 infection: the China and Italy perspectives, *Cell Death Dis.*, **11** (2020), 438.
23. E. Lavezzo, E. Franchin, C. Ciavarella, G. Cuomo-Dannenburg, L. Barzon, C. Del Vecchio, et al., Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo', *Nature*, **584** (2020), 425–429.
24. S. J. Kang, S. I. Jung, Age-related morbidity and mortality among patients with COVID-19, *Infect. Chemother.*, **52** (2020), 154.
25. Y. Liu, A. A. Gayle, A. Wilder-Smith, J. Rocklöv, The reproductive number of COVID-19 is higher compared to SARS coronavirus, *J. Travel Med.*, **27** (2020), 1–4.
26. Istituto Nazionale di Statistica, Primi risultati dell'indagine di sieroprevalenza sul SARS-CoV-2. Available from: https://www.istat.it/it/files//2020/08/ReportPrimiRisultatiIndagineSiero.pdf.
27. G. Albi, L. Pareschi, M. Zanella, Control with uncertain data of socially structured compartmental epidemic models, preprint, arxiv:2004.13067.
28. A. Capaldi, S. Behrend, B. Berman, J. Smith, J. Wrigth, A. L. Lloyd, Parameter estimation and uncertainty quantification for an epidemic model, *Math. Biosci. Eng.*, **9** (2012), 553–576.

29. G. Chowell, Fitting dynamics models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecast, *Infect. Dis. Model.*, **2** (2017), 379–398.
30. M. G. Roberts, Epidemic models with uncertainty in the reproduction, *J. Math. Biol.*, **66** (2013), 1463–1474.
31. A. Pugliese, S. Sottile, Inferring the COVID-19 infection curve in Italy, preprint, arXiv:2004.09404.
32. D. Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press, 2010.

AIMS Press