SAPIENZA
UNIVERSITÀ DI ROMA

Special Section: COVID-19 Pandemic

Vol. 4, No. 2 (2020)
ISSN: 2532-5876
Open access journal licensed under CC-BY
DOI: 10.13133/2532-5876/16990

A Numerical Method to Estimate the Peak of New Infected and the SARS-CoV-2 Outbreak in Italy

Federico Zullo

[a]DICATAM, University of Brescia, Via Branze 38, Brescia, 25123, Italy

Corresponding author: Fedrico Zullo, Email: federico.zullo@unibs.it

Abstract

We give some numerical observations on the total number of infected by the SARS-CoV-2 in Italy. The analysis is based on a tanh formula involving two parameters. A polynomial correlation between the parameters gives an upper bound for the time of the peak of new infected. A numerical indicator of the temporal variability of the upper bound is introduced. The result and the possibility to extend the analysis to other countries are discussed in the conclusions.

Keywords: COVID-19, SARS-CoV-2, epidemiological models, peak estimation

Citation: Zullo F 2020, "A numerical method to estimate the peak of new infected and the SARS-CoV-2 outbreak in Italy", *Organisms: Journal of Biological Sciences*, vol. 4, no. 2, pp. 111-119. DOI: 10.13133/2532-5876/16990.

Introduction

In most epidemics it is hard to determine the true number of new infected individuals per day. This is the case for the new coronavirus disease, since asymptomatic people or with very mild symptoms may not seek medical assistance and cannot be identified (Baud *et al*. 2020). Realistic data are fundamental to understand the epidemic and to steer the efforts to inhibit the disease in the right direction. Also, the dynamical variables of epidemiological models usually are linked to, or describe directly, the evolution of the true number of infected: the comparison with the empirical data may be problematic if those numbers are not realistic. On this side, researches about the estimation of the real scale of the epidemic or of the proportion of the asymptomatic already appeared in the literature—see e.g. (Li *et al*. 2020) or (Kenji *et al*. 2020).

On the other hand, under very reasonable hypotheses, it is possible to assume that suitable measurable quantities are determined by the relative values of certain characteristics of the population only (in opposition to global absolute values): in this case the knowledge of only a fraction of new infected individuals per day may still be useful to estimate some of the measurable quantities. This property (we will refer to it as "scale invariance") must be reflected in a scale-independent property of the underlying epidemiological model. In this paper we assume that the time of the peak of new infected by the SARS-CoV-2 in Italy has the scale-invariance property. We are aware of the fact that this assumption can be considered at best a rather crude approximation to a very complex system of interactions. In our opinion however, when taken as a working hypothesis, it can provide a well-founded basis, or at least a starting point, to achieve reasonable estimates. This point of view will be justified further in section (1) on the basis of the SIR epidemiological model.

Since the start of the epidemic in China, a certain number of studies appeared in the mathematical community about this subject: the description of the spatial or temporal diffusion of the infected in given regions (Fanelli & Piazza 2020), (Gaeta 2020a; 2020b), (Giuliani *et al.* 2020), the transmission dynamics of the infection (Kucharski *et al.* 2020), the economic and financial consequences of the epidemic (Albulescu 2020), the effect of atmospheric indicators on the spread of the virus (Wang *et al.* 2020), are only a fraction of the topics under investigation in these days. A certain number of epidemiological studies are connected to the SIR model. The SIR model is one of the simplest non-linear deterministic continuous (in time) model of epidemiology: the overall population is divided in three disjoint classes: S, i.e. the number of susceptible individuals, I, the number of infectious individuals and R, the number of recovered individuals. Albeit its non-linearity, the dynamic of the model is fairly uncomplicated and manageable from an analytical point of view and displays very interesting and realistic properties such as the existence of an *epidemic threshold*—see e.g. (Braun 1993) and (Murray 2002).

We must underline that the assumption of the scale invariance is not specific of the SIR model: rather, the SIR model is seen here as an instance among the family of models possessing the scale invariance.

The paper is organized as follows: in section 1 the SIR model is introduced and briefly discussed. In section (2) we analyze the data of the cumulative number of infected in Italy on the base of two simple hypotheses. An upper bound for the time of the peak of new number of infected is obtained. This upper bound is dynamic: when more data are added to the model in the course of the epidemic it may changes in time. In section (3) we will discuss the predictive validity of the model on the basis of a numerical indicator measuring the temporal variability of the upper bound. In the conclusions, we will comment about the results and look for possible extensions.

1. The SIR Model and the Scale Invariance Property

The SIR model describes the evolution of the individuals in the susceptible *(S)*, infectious *(I)* and recovered *(R)* classes with the following differential equations:

$$\begin{cases} \dfrac{dS}{dt} = -r\dfrac{SI}{N}, \\ \dfrac{dI}{dt} = r\dfrac{SI}{N} - aI, \\ \dfrac{dR}{dt} = aI. \end{cases} \tag{1}$$

The total population $N = S + R + I$ is a conserved quantity from the dynamical point of view, meaning that there are only two independent variables in the set of equations (1). The characteristics of this model are well-known and the interested readers can look for example at the discussions in the classical books of (Braun 1993) and (Murray 2002). Here we will make only few observations, relevant for the next sections.

Some authors do not include the denominator N on the right hand side of (1), since it is a constant and can be absorbed by a re-definition of the parameter r. However, we will keep it: in this way it is indeed evident the scale invariance property of the model: if the initial conditions (S_o, I_o, R_o) are scaled by a common constant factor k, (and so the total population is scaled by a factor k), the solution is scaled by the same factor. Indeed it is enough to observe that, if ($S(t)$, $I(t)$, $R(t)$) are the solutions of equations (1) corresponding to the initial conditions (S_o, I_o, R_o), then ($kS(t)$, $kI(t)$, $kR(t)$) are the solutions corresponding to the initial conditions (kS_o, kI_o, kR_o).

Some temporal properties of this model, like the time corresponding to a maximum in I (the time of the peak of the infected), do not depend on the scaling factor k. This property is very useful, since the actual number of infected or susceptible (and then of recovered) is in general not known. The reasonable assumption that the *same fraction* (with respect to the total) of infected, susceptible and recovered individuals are known, gives the possibility, in this case, to compare the measured data with the properties that are scale-independent.

The solution of the system (1) cannot be given explicitly in terms of known functions. However, if the epidemic is not severe, i.e. the number $R(t)$ can be considered small compared to the overall population, an explicit

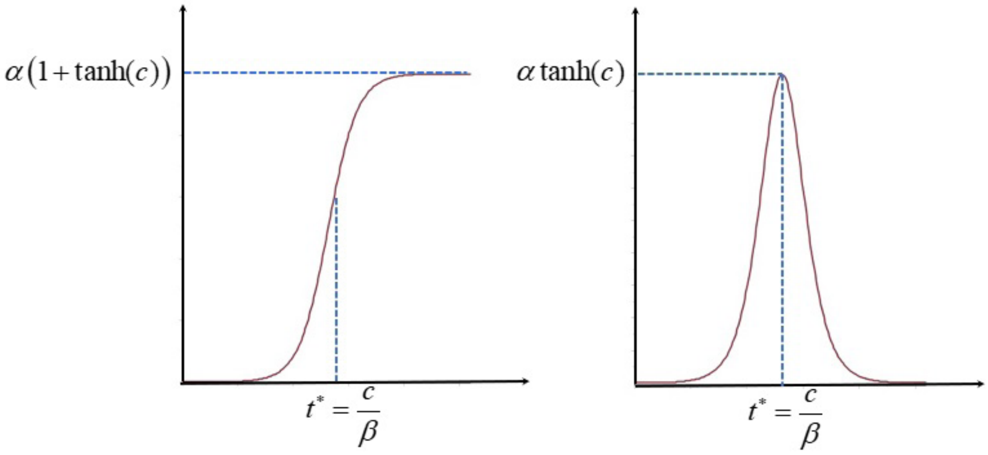


Figure 1: The plot of the function (2) (left) and of its derivative (right) as functions of time for generic values of the parameters (*α, β, c*). The epidemiological interpretation of the parameters is shown.

formula for the number of recovered can be obtained in terms of the hyperbolic tangent function. Here, we will not give the details: the interested reader can look for example in (Kermack & McKendrick 1927) and (Murray 2002). The function reads as

$$R(t) = \alpha \tanh(\beta t - c) + \alpha \tanh(c), \qquad (2)$$

where we used the initial value $R(0) = 0$. The important point for the rest of the paper is not the exact solution of the system of equations (1), neither the behavior of the solution. Rather, the possibility to represent an epidemiological curve with a simple and manageable formula like (2) will be crucial in this study. The parameters (*α, β, c*) possess an explicit representation in terms of the parameters *a* and *r* of the SIR model (1) and of the initial conditions (S_o, I_o, R_o). The formulas are quite cumbersome and the interested reader can look for example in (Murray 2002). Our aim here is not to analyze the data to fit the solutions of the system (1), but to make use of the explicit formula (2) in a way that will be clear in the next pages. In passing, we would like to underline that, if on the one hand the SIR model gives a mathematical basis to formula (2), on the other hand it would be possible to consider (2) as a postulate and to assess the soundness of this postulate from the truth value of the conclusions obtained. From this point of view, formula (2) can be considered as one of the examples of the so-called s-shaped epidemiological curve (with a peaked derivative, the function sech^2) that universally describes an infection disease (Braun 1993). As can be seen from the second formula in the set of equations (1), once the value of *R* is given, the value of the number of infected can be obtained by derivation, i.e. $aI(t) = \alpha\beta\text{sech}(\beta t - c)^2$. When considering the cumulative number of infected, $R + I$, the contribution of sech^2 is negligible on the tails, whereas it is more pronounced in correspondence of the maximum of sech^2, but it is however small if the value of the parameter β is less than one. In this case, the value of $R + I$ is well approximated by a tanh formula like (2), with a certain different value of *c*. For the sake of clearness, we report in Figure (1) a plot of the function (2) (left) and a plot of its derivative (right): as can be seen from the figures, an epidemiological interpretation of the parameters (*α, β, c*) can be the following: $t^* = c/\beta$ is the time of the peak of new number of infected, $\alpha(1 + \tanh(c))$ is the cumulative final number of people infected, whereas $\alpha \tanh(c)$ is the maximum of the new number of infected.

As a final remark we want to make two observations about the scale invariance and the usefulness of formula (2). The first observation is the following: the scale invariance property assumes that the same fraction of the true number of infected, susceptible and recovered is measured. However it is tacitly assumed that this fraction does not vary in time. If the epidemic persists in time, there is the possibility that the value of such fraction changes significantly. For this reason the number of data to be analyzed must span a limited interval of time. In the following section we will take the data of the outbreak in Italy from the 6th of March (15 days after the 21st of February, when the outbreak started, in order to have enough statistical data) to the 2nd of April, for a total of 28 days. The second observation is about the usefulness of formula (2). We are aware that this formula is a very crude approximation of the real curve, but, for limited intervals of time (like the interval that we are going to consider), it is able to incorporate, in a simple way, the main characteristics of the epidemic. For large time, the epidemic curve may be asymmetric and surely its differential will develop with different velocities. For the above reasons, in the next section we are going to use the above formula only for a limited amount of data. As we will see, this is enough to obtain some relevant information about the time of the peak of new infected.

2. Analysis of Data with a tanh Model

The discussion made at the end of the previous section, despite to be very basic, has the advantage to be manageable and to incorporate the main properties of the SIR model. It is not by chance that the first application of the SIR model (the Bombay plague of 1905) by Kermack and McKendrick (Kermack & McKendrick 1927) used precisely the tanh formula above.

In the following we will base our analysis on two hypothesis:

1. We assume that the cumulative number of infected is described by a tanh model, when the data analyzed span a limited interval of time (as explained in section (2)). Although this assumption is coherent with the founding of the SIR model, it does not depend on the particular dynamical model considered.
2. We assume that, whatever it is the underlying model describing the evolution of the number of infected, this model is scale invariant, in the sense specified in the previous section.

The second hypothesis is fundamental since we are going to look at scale-independent quantities: even in the case the measured number of infected and recovered individuals are different from the actual values, it is possible to estimate these quantities.

The cumulative total number of infected that will be considered in the next lines are those of the entire Italian territory. There are at least two reasons that suggested to not taking regional or local data: the first one is that the epidemic started to spread across three different regions (Lombardy, Veneto and Emilia-Romagna), and there could not be a correspondence between the locality where a certain fraction of inhabitants reside and the region where this fraction was infected. This is also true at a national level, but the fraction is assumed to be smaller. The second reason is that a non-negligible number of workers and students moved, just before the lockdown, from the northern regions to their regions of origin in the center and south of Italy.

The possibility that a non-negligible flow of infected people passed from the north to other regions should be taken into consideration. By taking the entire national set of data, we overpass the above issues.

The data can be taken for example from WHO (World Health Organization, 2020) or from Worldometer (Worldometer, 2020). The cumulative total number of infected will be indicated by F_n, with $F_1 = 21$ corresponding to the number of infected on 21st of February 2020. The subscript n stays for the number of days from the starting of epidemic. These data will be opposed to the continuous formula

$$f(t, \alpha, \beta, c) = \alpha \tanh(\beta t - c) + \alpha \tanh(c). \quad (3)$$

The value of β will be taken to be constrained by the equation

$$\alpha \tanh(\beta - c) + \alpha \tanh(c) = F_1. \quad (4)$$

The function f (3) then depends on two parameters, α and c. When necessary, to stress the dependence on these parameters, we will denote the function with $f_{\alpha,c}(t)$. The cumulative final number of infected expected from formula (3) is given by $f_\infty = \alpha(1 + \tanh(c))$. It is possible to estimate the parameters α and c by minimizing the difference between the actual and predicted number of cases, i.e. minimizing

$$S_n = \sum_{i=2}^{n} (F_n - f(n))^2 \quad (5)$$

In order to have a reasonable minimum number of data, we start the analysis by taking $n \geq 15$. The values of the parameters minimizing the sum S_n are reported in table (1). Please note that some values are slightly different from those of the preprint version of this paper since some data on the cumulative number of infected F_n have been corrected according to (Worldometer, 2020). These values have been obtained by equating the derivative of the expression (5) with respect to α and the derivative with respect to c both to zero. The solutions, for each given n, have been found numerically. Due to the strong non-linearity of equation (3) it is not easy to find estimates for the confidence intervals.

A plot of $df_{\alpha 42, c42}/dt$ and of $F_n+1 - F_n$ is reported in Figure (2). A fundamental observation is that the function S_n actually has a basin of depressed values, showed in detail in Figure (3) for a given value of n. This basin of

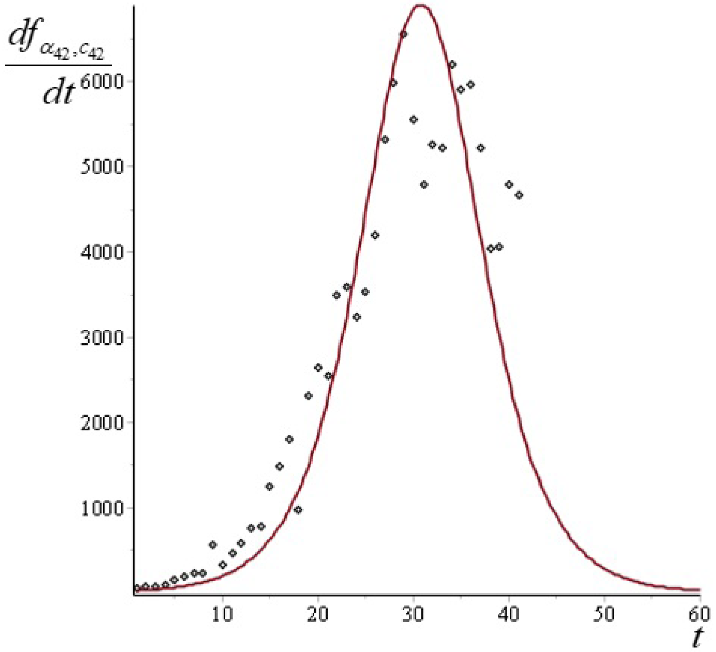


Figure 2: The number of new infected and the continuous curve given by $df_{\alpha 42,c42}/dt$

n	α_n	c_n	n	α_n	c_n
15	3514.0	2.506	29	30995.0	3.403
16	4703.0	2.618	30	35166.7	3.455
17	6482.9	2.749	31	38541.2	3.492
18	8757.1	2.876	32	40737.4	3.515
19	8748.9	2.862	33	42602.6	3.533
20	9908.2	2.926	34	44297.6	3.548
21	12158.0	3.011	35	46317.8	3.566
22	14102.9	3.074	36	48397.7	3.582
23	16946.5	3.151	37	50527.0	3.599
24	19752.4	3.216	38	52472.3	3.613
25	21453.2	3.251	39	54030.3	3.624
26	22862.5	3.278	40	55380.4	3.633
27	24679.2	3.309	41	56736.1	3.642
28	27439.8	3.353	42	58161.2	3.650

Table 1: the estimated values of α_n and c_n

minimum seems to indicate that there is a given function $\alpha(c)$ giving a family of tanh curves with reduced values of S_n. The curve $\alpha(c)$ is quite stable by varying n (see section (4)) and suggests to look at the values of α_n as functions of c_n. In Figure (4) we report the plot of the values of α_n and c_n given in table (1) as a function of n, whereas in Figure (5) the plot of the values (c_n, α_n). The values of α_n vs c_n, as explained above, describe the basin of depressed values for S_n as a function of $\alpha(c)$. We make a cubic t, with linear coefficients, in order to get a rough description of the curve $\alpha(c)$:

$$\alpha = \sum_{k=0}^{3} a_k c^k \qquad (6)$$

Clearly, by considering a number N of values of α_n and c_n to fit a_k, $k=0,...,3$, we will obtain a set of values $\{a_{k,N}\}$. By fitting all the data available (i.e. by taking $N = 28$), we get the following values for the coefficients a_k:

$$a_0 = -999707,\ a_1 = 1077192,\ a_2 = -389358,\ a_3 = 47555. \qquad (7)$$

It is possible to get more terms in the sum (6), but the cubic term is sufficient to get a formula accurate enough to what we are going to say.

The plot of the fit is given in Figure (6), together with the values of the residuals,

$$\alpha_n - \sum_{k=0}^{3} a_k c_n^k,$$

where the values ak are those given in equation (7).

A comparison between the curve $\alpha(c)$ and the basin of minima for S_n has been plotted in Figure (7): the red curve is the function (6) with the black dots giving the actual values of (c_n, α_n) in table (1).

The function $\alpha(c)$ denotes a trend in the data that may be useful. If in the next days the values of the infection continue to rise, it is reasonable to expect that the values of α and c will be constrained closely by the same curve. Clearly, the model used here is rough, but it can give at least an idea about the future trend of the data. We are tacitly assuming that there will be no other cluster of infection around Italy in the next days: the point will be discussed later.

Now we consider the function f in (3) as a function of t and c alone, since the value of a is constrained by the curve (6). The plot of the derivative of this function (with respect to t) gives the time of the peak of infections. The plot is reported in Figure (8): we notice that the maximum of the derivative of the cumulative number of infected increases with n up to c ~ 4.3 and then *decreases* by increasing c. This gives an upper bound for the peak of new number of infected (the point where the

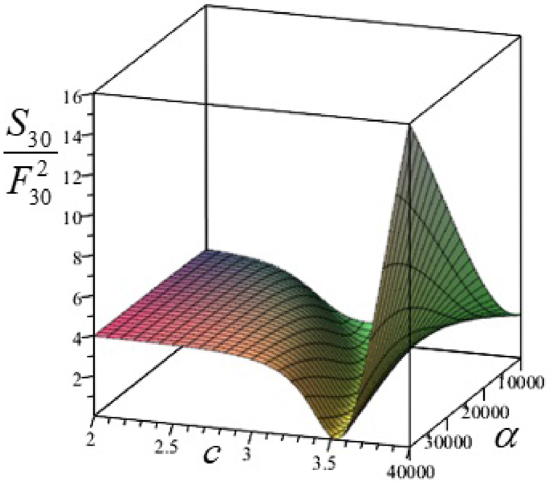
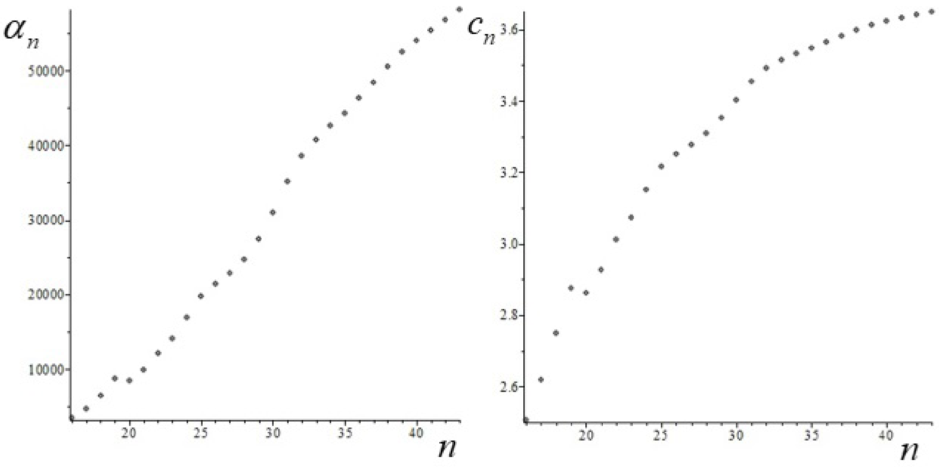


Figure 3 (left): The basin of depressed values of S_{30}: the values have been re-scaled to F_{30}^2 for easy of plotting.

Figure 4 (down): The values of α_n and c_n vs n as given in table (1)



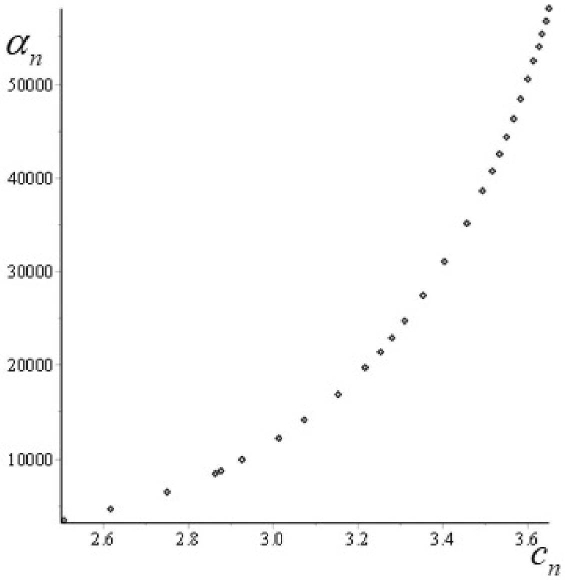


Figure 5: The values of α_n vs the values of c_n as given in table (1)

second derivative of $f(t)$ (3) is zero), given by 36 days after the data corresponding to F_1 (21st of February).

3. Temporal variability of the upper bound

The basin of minimum of S_n, for each fixed n, is described by a function $g_n = \alpha(c)$: this function gives a family of tanh curves with reduced values of S_n. The curve $\alpha(c)$ has been described in the previous section by fitting the values of α_n and c_n in table (1) with a cubic formula. We obtained just one curve by making use of all the data available, i.e. 28 couples (c_n, α_n). It is possible to ask how the curves $g_n(c)$ depend on the number of the data available: if the curves g_n have a temporal stability, then they can be used to make a reliable estimation of the time of the peak of new infected. To address this question we fit the data (c_i, α_i), with the index i from 15 to N, by varying the number of data taken (i.e. by varying N). The fit is again cubic, like in (6). In this way we follow the temporal variation of the curve $g_N(c)$. In plot (9) we report the curves $g_N(c)$ for N in the interval [30, 42]: they correspond to the last thirteen curves (in order of time). It is possible to see that indeed the convexity of the curves is slowly increasing, so that the differences are more pronounced for higher values of c. To have a measure of the variability of these curves we introduce a parameter giving the relative increase of $g_n(c)$ for c fixed and equal to the last available value (i.e. $c = c_{42}$ in table (1))

$$P_N = \left.\frac{\sum_{k=0}^{3} a_{N,k}c^k}{\sum_{k=0}^{3} a_{30,k}c^k}\right|_{c=c_{42}} \qquad (8)$$

The values of P_N, $N = 30...42$, are also reported in Figure (10): the maximum value is obtained for $N = 42$ corresponding to $P_{42} = 1.058$.

4. Discussion

Optimal responses to public health emergencies must be tailored to the regional context and must take into account different aspects of the epidemic, like its severity and the type of risks, and different aspects of the local capacity to manage and mitigate the spreading. From this side the quantity and quality of healthcare infrastructures and the coordination across public and private sectors appear to be crucial for a successful implementation of the control strategy. On the other side, mathematical modeling tools can surely be decisive to properly assess the intensity, evolution and

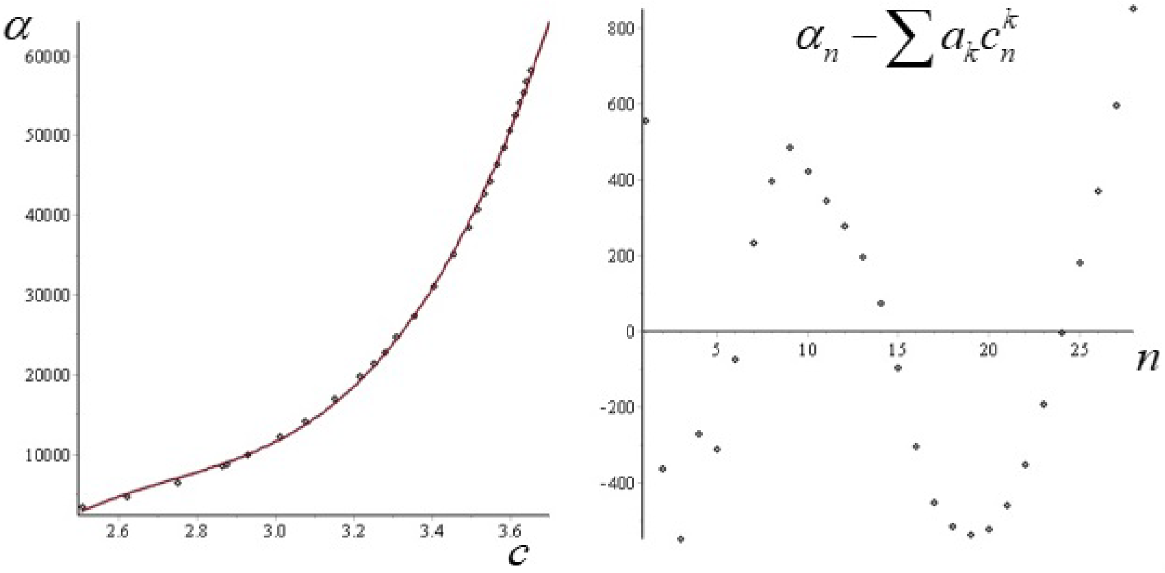
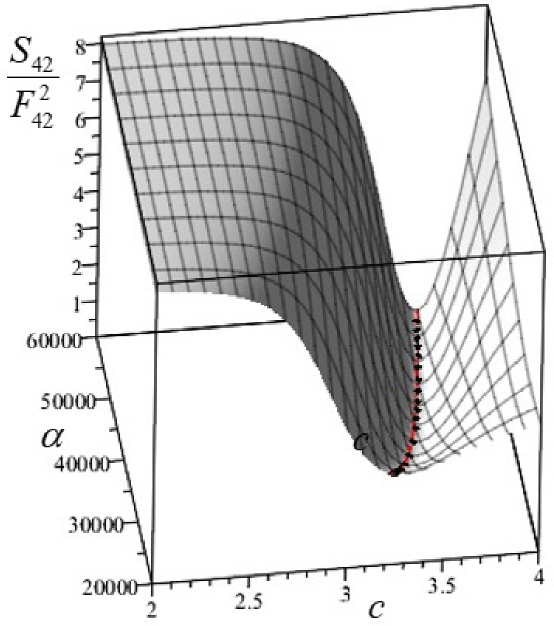


Figure 6: The plot of the fit (6) (left) and the values of the residuals $\alpha_n - \sum_{k=0}^{3} a_k c_n^k$



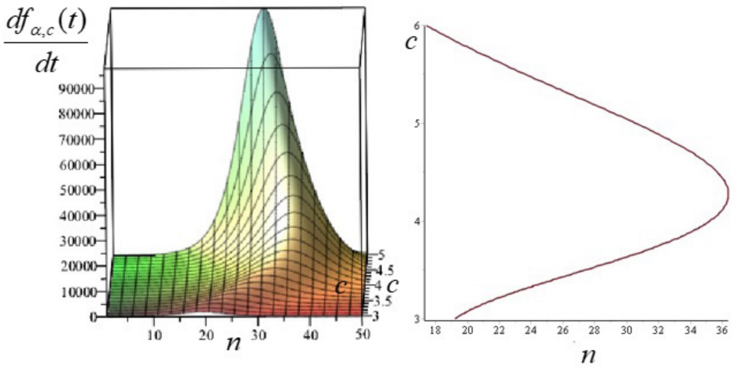
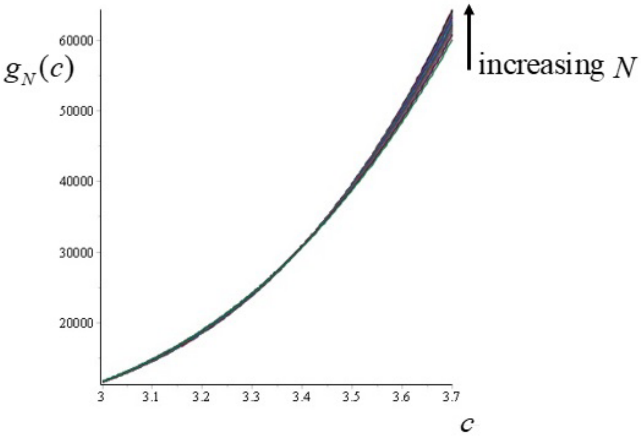


Figure 8 (top): The values of the derivative of the function f (t) vs n and c (left) and the corresponding values of maxima as a function of n and c (the points where the second derivative of f (t) vanishes) (right).

Figure 7 (left): The basin of minima together with the curve (6) (in red) and the actual values (c_n, α_n) (black dots).



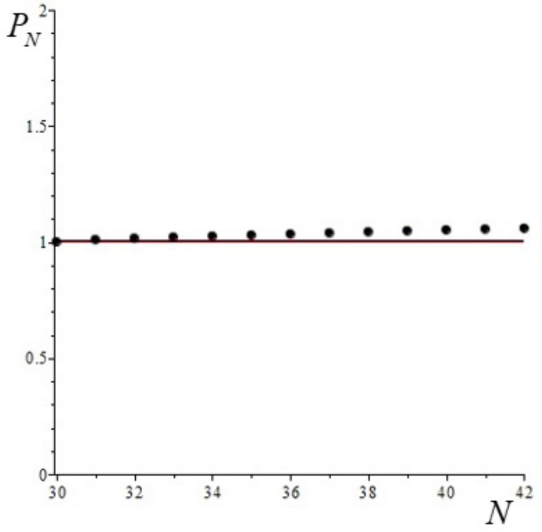


Figure 9: The functions gN (c), for N = 30...42: the curves overlap in a small region of the plane (α, c).

Figure 10: The values of the parameters P_N , N = 30...42, measuring the temporal variability of the curves g_N (c).

duration of the emergency. A reasonable estimate of the peak of new infected in epidemic outbreaks may be useful not only to rightly evaluate the dynamic of the infection, but also to give an (a posteriori) assessment of the effectiveness of the containment strategies: the perseverance of the infections may denote some degeneration of the control strategy or some change in the social situation. Clearly, the peak of new infected is just one of the multiple indicators that should be taken into account: deterministic or stochastic models, r_o and r_t indexes, risk parameters and spatio-temporal simulations are all fundamental to provide a quantification of the evolution of the infectious disease and develop a rapid decision making process.

Conclusions

The above analysis, despite using a rough function for the total number of infected, is able to give an upper bound for the time of the peak of new infected (27th of March) thanks to the observation that the values of αn are, in a certain sense, not independent on the values of cn and are well described by a polynomial interpolation with linear coefficient. The hypothesis about the scale invariance of the underlying model (that, we repeat, not necessarily is represented by the SIR model) and the low temporal variability of the upper bound are fundamental for the accuracy of the result. Another underlying assumption is that the restrictive measures will be kept and observed in the next days and there will be no other clusters in the south of Italy (in the SIR model language, the values of S_0 are below the epidemic threshold, see e.g. (Murray, 2002)). In the unfortunate case that there will be other clusters (the preprint version of the paper appeared before March 27, 2020 on arXiv, see: https://arxiv.org/abs/2003.11363v1), it is possible to think of a substitution of the tanh curve by a combination of such functions: if there are two clusters of comparable magnitude, then we will have

$$f(t) = \alpha_1 \tanh(\beta_1 t - c_1) + \alpha_1 \tanh(c_1) + \alpha_2 \tanh(\beta_2 t - c_2) + \alpha_2 \tanh(c_2) \qquad (9)$$

A statistical analysis of the data given in table (1) will surely help to improve the results here given and will be provided in a next paper, where other sets of data, from different countries, will be analyzed.

Acknowledgments

The author would like to acknowledge the support of the University of Brescia, GNFM-INdAM and INFN.

References

Albulescu C 2020, “Coronavirus and oil price crash”, *Working Papers, HAL*, https://ideas.repec.org/p/hal/wpaper/hal-02507184.html.

Baud D *et al.* 2020, “Real estimates of mortality following covid-19 infection”, *The Lancet Infectious Diseases*, vol. 20, no. 7, p. 773. DOI: https://doi.org/10.1016/S1473-3099(20)30195-X.

Braun M 1993, *Differential Equations and Their Applications: An Introduction to Applied Mathematics*. Berlin: Springer.

Fanelli D, & Piazza F 2020, “Analysis and forecast of COVID-19 spreading in China, Italy and France”, *Chaos, Solitons and Fractals*, vol. 134, 109761. DOI: https://doi.org/10.1016/j.chaos.2020.109761.

Gaeta G 2020a, “Data analysis for the COVID-19 early dynamics in northern Italy”, *eprint Arxiv*, https://arxiv.org/abs/2003.02062.

Gaeta G 2020b “Data analysis for the COVID-19 early dynamics in northern Italy. The effect of first restrictive measures”, *eprint Arxiv,* https://arxiv.org/ abs/2003.03775.

Giuliani, D, Dickson M M, Espa G, & Santi F 2020, “Modelling and predicting the spatio-temporal spread of COVID-19 in Italy”. *BMC Infectious Diseases*, vol. 20, art. no. 700, DOI: https://doi.org/10.1186/s12879-020-05415-7.

Kenji M, Katsushi K, Alexander Z, & Gerardo C 2020, “Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship”, *Euro Surveillance*, vol. 25, no. 10, DOI: 10.2807/1560-7917.ES.2020.25.10.2000180.

Kermack W, & McKendrick A 1927, “A contribution to the mathematical theory of epidemics”, *Proceedings of the Royal Society A*, vol 115, no. 772, pp. 700–721, DOI: http://doi. org/10.1098/rspa.1927.0118.

Kucharski A *et al.* 2020 “Early dynamics of transmission and control of COVID-19: A mathematical modelling study”, *Lancet Infectious Diseases*, vol. 20, no. 5, pp. 553–558, DOI: 10.1016/S1473-3099(20)30144-4.

Li D *et al.* 2020, “Estimating the scale of COVID-19 epidemic in the United States: Simulations based on air traffic directly from Wuhan, China”, *medRxiv* [preprint], DOI: 10.1101/2020.03.06.20031880.

Murray J 2002, *Mathematical Biology*, vol. 1, Berlin: Springer.

Wang J *et al.* 2020, “Impact of temperature and relative humidity on the transmission of COVID-19: A modeling study in China and the United States” *BMJ Open, Forthcoming*, Available at SSRN: https://ssrn.com/abstract=3551767 or http://dx.doi.org/10.2139/ssrn.3551767

World Health Organization 2020, *Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update*, [online] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

Worldometer 2020, *Coronavirus, 2020*, [online], https://www. worldometers.info/coronavirus/country/italy/.