Brownian motion prediction for the logistic Covid-19 model infection

Massimo Scalia [1] and Carlo Cattani [2]
[1] Interuniversity Research Centre for Sustainable Development (CIRPS)
[2] Engineering School, DEIM, University "La Tuscia"

ABSTRACT

One of the simplest mathematical models for population growth, the Verhulst logistic curve, is provided to make some prediction on the curve growth of Covid-19 pandemic spread. However, due to the "rigidity" of the logistic curve, a precise forecast can be done only under some very special and well-determined conditions, such as in the case of the time evolution of the infection in China, or, maybe, for seasonal flues. In general, taking into account the halo of randomness associated to each reported daily data and treating the latter as a value of the Brownian motion, we can replace the logistic behavior with a theoretical time series, fairly approximating the real data series. The day of the overlapping between theoretical and real series is what we were looking for, as we show applying this method to Lombardy and its more affected cities: Bergamo, Brescia and Milan. To shorten the calculations to obtain the theoretical values, a linear approximation is provided, a sort of geometric tangent to the data curve in the crucial day, able to dominate the data of the successive evolution of the infection.

Introduction

Many models have been introduced in scientific literature or are underway around the world to describe and try to interpret covid-19 growth and the burden of infected people and deaths. Some of these models involve such parameters that there is in turn the need for a model to evaluate them, others request an impressive level of detail in the modeling of social and spatial interactions. It's surely important to think about age groups and context of interactions for respiratory infections, but it should be avoided the risk, focusing too much on individual-level social behavior, of adding complexity without obtaining more predictability. Among the models which are aimed to give a "dynamics" to the growth of pandemic, interesting are those of SEIR (**S**usceptible, **E**xposed, **I**nfectious, or **R**ecovered) type, based on an ordinary differential equations system. We too prefer something similar to a dynamic model, because the general objective of all models should be, in our opinion, not only to be applicable to the description of a breakout and its diffusion but also to make some predictions useful for providing quantitative reference parameters to decision makers, in order to manage and possibly control the evolution of infection. Starting from the simplest model that can give predictions, e.g. with a single differential equation instead of a system of them, we limit ourselves to the search for the simplest

dynamics able to foresee some important characteristics such as the maximum number of infected people and from which day, starting from the first observed cases, to give a reliable forecasting model. The simplest model is the logistic model, that can give information through the "growth function" and generates the dynamics, as well as the quadratic map in the May's model. The logistic curve represents the growth of the population of a single species (animals, plants and also viruses) and is the integral of the differential equation of Pierre Verhulst, a mathematician who in 1838 corrected the exponential growth, therefore unlimited, proposed by Thomas Malthus in "An Essay on the Principle of Population" (1798). The latter actually intended to warn the Economy - he had read Adam Smith's "The Wealth of Nations" (1776) - of the problem of the exhaustion of physical resources. A basically unheard topic from the mainstream as up to the present day, but that's another story.

1. The population growth

Let N be the number of individuals in a population and ΔN is the change that occurs after a time Δt, the quantity $1/\Delta t$ ($\Delta N/N$) is the average growth rate relative to the time interval Δt. Although N has integer values, let's assume to be a continuous and derivable function of time t, $N = N(t)$, so that we can define the instantaneous growth rate: $\lim_{\Delta t \to 0} \frac{1}{\Delta t}\left(\frac{\Delta N}{N}\right)$, which, if it is constant value, a, immediately gives rise to the equation $dN/dt = a\,N$. It has as solution $N = N_0\, e^{a\,(t\,-t_0)}$, which is the exponential growth of Malthus, where N_0 is the population at the instant in which the observation begins, t_0. Since the exponential growth of a population have never been observed, the Malthus model has been revised in order to take into account a sort of "social friction" that occurs when the population grows (insufficient space, availability of resources, difficulty of reproduction).

We can also assume that, for each kind of population, there is a feasible maximum P for the number of individuals such that once that maximum is exceeded, the population begins to decrease:

$N > P \longrightarrow a < 0$.

The easiest way to introduce this "limiting" factor is to assume that the growth rate is not constant but depends linearly on the difference between the population at the instant t and the maximum population P,

$a = b\,(P - N)$, con $b > 0$, and then

$$dN/dt = bN\,(P - N) = bPN - bN^2 = rN - qN^2\,,$$

where $r = b\,P$ and $q = r/P$ is the coefficient of the quadratic term that corrects the exponential trend, which would occur in its absence. The correction operates as a term of mortality, which translates the social friction within the population.

Thus, we have the logistic equation of Verhulst, which historically is written:

$$1)\ dN/dt = r\,N\,(1 - N/P).$$

The function N (1- N/P) is called the *growth function* and P, in Ecology, is the *carrying capacity*.

The growth rate depends on the reproductive capacity of the species, the carrying capacity of the environment.

The solution of the logistic equation is the "***logistic function***"

$$N(t) = \frac{PN_0\, e^{rt}}{P+N_0\,(e^{rt}-1)}\ ,$$

where N_0 is the population at the time the observations begin, t_0. The carrying capacity is an ***asymptotic value***, *not actually achieved by the population*. In fact, if one divides by $N_0\, e^{rt}$ the numerator and denominator to the second member, gets the expression for the solution

$$2)\quad N(t) = \frac{P}{1+ h\, e^{-rt}}\ ,$$

where $h = \frac{P - N_0}{N_0}$, and ***therefore*** $\lim_{t\to\infty} N(t) = P$.

Each population tends to balance the environment. The equilibrium solutions, which are obtained by canceling the second member of the logistic equation, are: $N = 0$ and $N = P$. As seen in Fig. 1, where k is our P, each solution tends to the carrying capacity P regardless of the number of individuals N_0 who constitute the population at the initial instant t_0, provided that N_0 is higher than the "critical minimum", below which the population cannot grow.

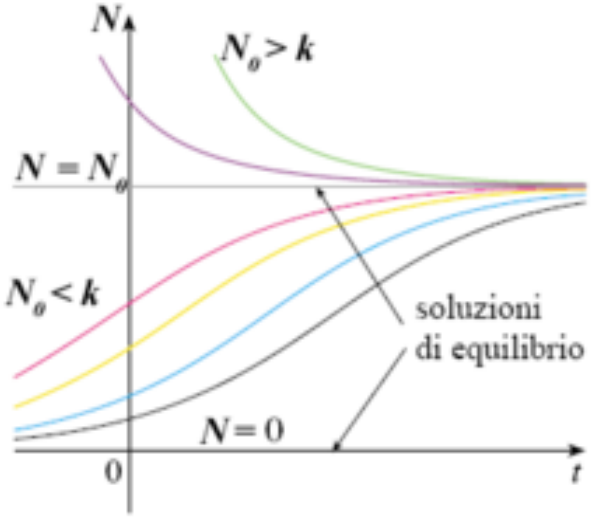

Fig. 1 Graph of the solutions of 1) corresponding to different initial data N_0

If the number of individuals N_0 is less than the carrying capacity P, then the solution is the curve - one of those below $N = N_0 = k$, in Fig.1 - which by its shape is called "sigmoid".

2. Some properties of the logistic function

Let's rename the function $N(t)$ with the more familiar, in the differential calculus, $x(t)$; and $x(t_0)$ is the value of population - e.g., persons infected by covid-19 - at the initial time t_0, that is the time in which observations start. Thus, we can rewrite (2), remembering that $h > 0$ and then we can set $h = e^{\ln(h)}$, as

(2') $\quad x(t) = P/(1+ e^{\ln(h) - rt})$.

In Fig. 2 one finds graphs of a logistic curve and of its growth velocity,

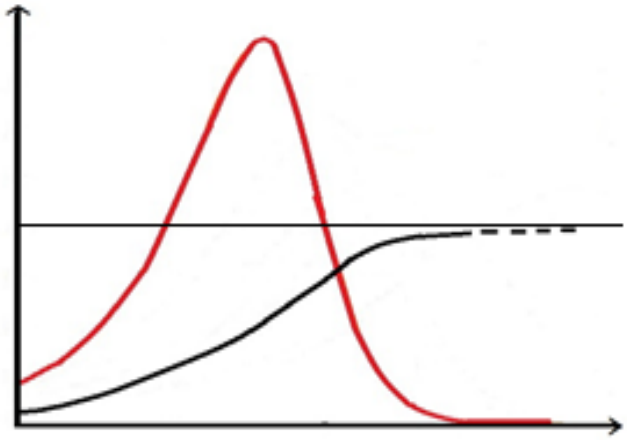

Fig. 2. The black curve is the logistic function for $x_0 < P$ and the red curve is its growth rate (the first derivative of the logistic function). On the abscissa, the time for both curves. In ordinate, the number of individuals for the black curve; the growth rate, i.e. the percentage change, for the red curve.

Two populations of the same species and with the same growth function but with different carrying capacity evolve along two congruent sigmoids, i.e. in some sense "parallel".

The analysis of function 2'), i.e. the calculation of its derivatives, shows us its inflection points and enables us to realize that ***the peak of the red curve*** is obtained just where the ***black one has an inflection***; the corresponding ***population to this point - a day, if one speaks of covid-19 - is the half of the carrying capacity***, that is,

$$x_{max} = P/2,$$

where, by x_{max} we denote the population corresponding to the maximum growth rate or, for the logistic curve, the population at the corresponding inflection point. The latter does not depend on the value of the initial population $x_0 = x(0)$; and all the sigmoids which have the same carrying capacity have the same inflection point. For $x = P/2$, the maximum growth rate υ_{max} is given by

$$\upsilon_{max} = rP/2[1-(1/P)\, P/2] = rP/4.$$

The linearized stability analysis of the equilibrium solution $x^* = P$, the other $x = 0$ is a trivial case, leads to the equation

$$d/dt\,[x(t)] = r\,(1 - 2x/P)\,x,$$

that, in a neighborhood of $x^* = P$, becomes

$$d/dt\,x(t) = -\,rx \quad \Rightarrow \quad x(t) = x_0\, e^{-rt},$$

that is, *the equilibrium solution is stable and "attractive"* (as shown in Fig. 1).

The time $T_r = 1/r$ is called the *characteristic return time* (May, 1976): it is the time needed to restore the previous population level after a disturbance that has altered (reduced) the number of individuals.

Oscillatory solutions (delayed logistic function). The population capable of reproducing itself at time t is that which, generated in an earlier time $t - \delta$, is still presents at t. If this delay is introduced in the growth function, the "quality" of the dynamic changes; in fact, the analysis of linearized stability in a neighborhood of $x^* = P$ shows that there are values of δ for which solution x^* from stable becomes unstable. If we denote with $\underline{\delta}$ the "critical" value of the transition from stability to instability, then it can be shown that this change of stability is a sufficient condition for the old solution, which has become unstable, to be flanked by new, stable ones, in a neighborhood of values $\delta > \underline{\delta}$.

The change in stability has generated new solutions, that is, has produced a bifurcation. When the bifurcation occurs for values greater than the critical value $\underline{\delta}$ it is called supercritical bifurcation; the new solutions have an oscillatory character and their trajectories are superimposed on the sigmoid, deviating from the latter curve precisely at its upper "elbow", the greater the deviation the greater the value of δ.

3. Looking for a predictivity

The main problem in a deterministic prediction of population growth is about the possibility to use the logistic function as a simple tool to forecast the behavior in time of the observed phenomenon, like, in our case, the growth of infected people in a country or in some areas (regions, cities) of particular interest (e.g. the situations of high number of cases or of deaths).

Together with its mathematical simplicity, the logistic function, widely used in many sciences of life, couples a substantial rigidity in fitting with the experimental data, mainly due to the linearity of the argument of the exponential: $ln\,(h) - rt$. Anyway, a question reported in many scientific or data collecting sites during the pandemic has been if and when the Covid-19 curves would flatten out. *Is it possible to deliver some forecast useful to take timely decisions for health and administrative management of outbreak*? *And with an anticipation of how many days; or, better, how many days after the "initial time"*?

The behavior of some of the most covid-19 infected countries in the world at April seem to confirm a sigmoid behavior (see, Fig. 3.).

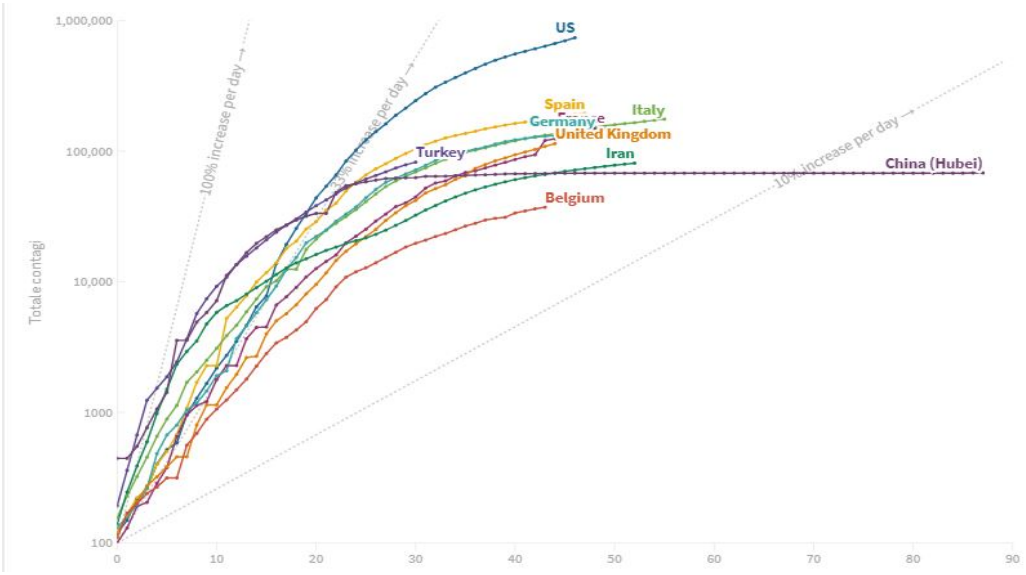

Fig. 3 Each curve starts from hundred cases registered, as WHO requires, and have all begun to flatten after 13th of April 2020. Source: Johns Hopkins, Medicine & Nature https://coronavirus.jhu.edu/data/new-cases

Less "regular" the behavior of new additional cases every day, that is, the rate of growth of the infection (see Fig. 4)

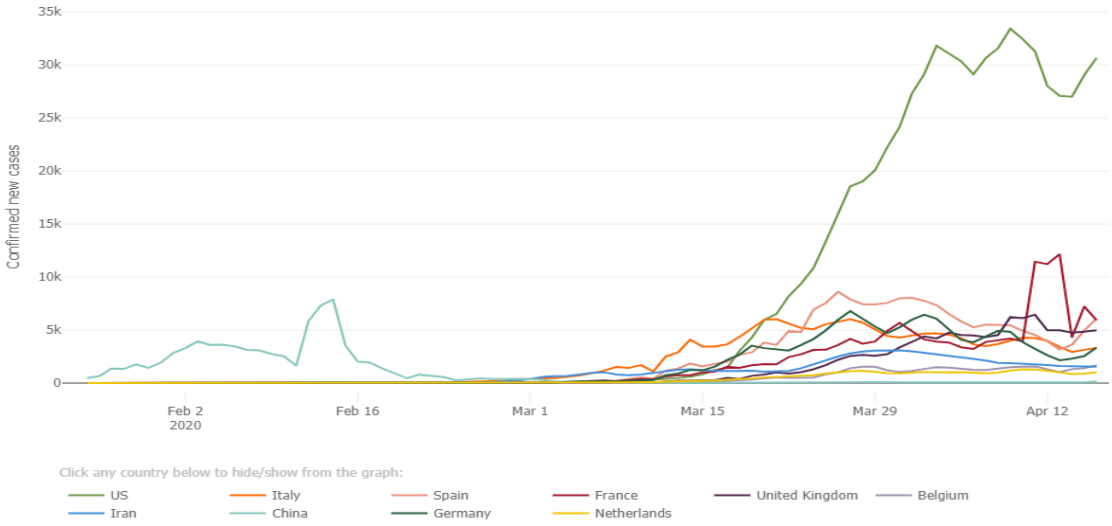

Fig. 4 Confirmed new cases for each day

Both graphs of Fig. 3, 4 represent a description of the situation such as it has evolved; but have substantially failed our attempts to use the logistic function as a tool to give a reliable answer to the above underlined question. For China (Hubei), the peak of velocity, so sharp, has given the day corresponding to the inflection point of the sigmoid - 10th of February, 42708 cases - and therefore yielding an estimate of the carrying capacity $P = 85.416$ cases. A very good estimate because until July 9 the confirmed cases by the WHO were 85.445, at a daily growth rate around 5 per ten thousand, with the curve that has been in its full asymptotic behavior since the last day of May. Moreover, because the isolated peak character could have been recognized just a few days later, the above prediction could have been confirmed 27 days after the first registration of the infection by National Health Authority.

But the "confirmed new cases" graphs reported in Fig. 4 for the other countries are enough to show that China was an exception, probably due to the special strict rules adopted by the authorities; anyway, the best fit of the data with the logistic curve, as in the case of China, could turn useful when, studying one of those epidemies which periodically hit many areas of our planet, like flus do, we can promptly recognize an "isolated" peak in the epidemic growth rate. Otherwise and in general, is a forecast possible, and with how many days of advantage?

4. Starting from data

Lombardy, the most populous and rich region of Italy, has been fiercely hit by covid-19. At the 17th of May: 84,844 infected, 37,6% of infections of all Italy; 15,519 deaths (48,6%) and a mortality ratio of 18,3%, the highest in the world, compared with regions of a similar demographic weight (for example, at the same day: NYCity: 192,990 case; 17,031 deaths and 8,8% mortality ratio).

The problem of applying the logistic model to the data isn't to know P and the rate r (see, (2')), because also having a little more than twenty daily data one can try to guess the next behavior and obtain the two mentioned values by successive attempts, in a sort of recursive process, but the laboriousness and the questionability of this method. In a nutshell, the attempts done have convinced us that reliable results could be obtained with a less rigid function than the logistic one, (2'), such as the "delayed" logistic function (see sect. 1) – an additional unknown parameter - or a parametrized family of logistic functions.

That is, mathematically more complicated methods. Because we have elected Occam's razor as our polar star, is there an easier method that, starting from the daily data available from the first thirty, or less, days, enables us to a reliable prediction on the successive evolution?

Here, the table 1 of daily data and, immediately after, the **graphs corresponding to those data**

Table 1: Number of Covid-19 infected

day	Lombardy	Milan	Bergamo	Brescia
1st March	984	46	209	40
2	1254	58	243	60
3	1520	93	372	86
4	1820	145	423	127
5	2251	197	537	155
6	2612	267	623	182
7	3420	361	761	413
8	4189	406	997	501
9	5469	506	1245	739
10	5791	592	1472	790
11	7280	925	1816	1351
12	8725	1146	2136	1598
13	9820	1307	2368	1784
14	11685	1551	2864	2122
15	13272	1750	3416	2473
16	14649	1983	3760	2918
17	16220	2326	3993	3300
18	17713	2644	4305	3784
19	19984	3278	4465	4247
20	22264	3804	5154	4648
21	25515	4672	5869	5028
22	27206	5096	6216	5317
23	28761	5326	6471	5905
24	30703	5701	6728	6298
25	32346	6074	7072	6597
26	34885	6922	7458	6931
27	37298	7469	8060	7305
28	39415	7783	8349	7678
29	41007	8329	8527	8013
30	42161	8676	8664	8213
31	43208	8911	8803	8367
1st April	44773	9522	9039	8598
2	46065	10004	9171	8757
3	47250	10391	9315	9014
4	49118	10819	9588	9180
5	50455	11230	9712	9340
6	51354	11508	9790	9467
7	52025	11787	9868	9594
8	53414	12039	9931	9909
9	54731	12393	10041	10139
10	56048	12748	10151	10369
11	57555	13214	10230	10618
12	59062	13680	10309	10868
13	60314	14161	10391	11058
14	61326	14418	10431	11122
15	62155	14675	10472	11187
16	63094	14952	10518	11355
17	64135	15227	10590	11567
18	65381	15546	10629	11758
19	66236	15825	10689	11946
20	66971	16112	10738	12004
21	67931	16601	10793	12091
22	69092	17000	10848	12178
23	70165	17277	10946	12308
24	71256	17689	11002	12475
25	71969	17908	11047	12540
26	72889	18371	11113	12564
27	73479	18559	11150	12599
28	74348	18837	11196	12691
29	75134	19121	11291	12806
30	75732	19337	11313	12861
1st May	76469	19701	11360	12929
2	77002	19950	11394	12999
3	77528	20068	11453	13028
4	78105	20254	11538	13122
5	78605	20398	11550	13168
6	79369	20711	11587	13267
7	80089	20893	11622	13391
8	80723	21094	11671	13480
9	81225	21272	11717	13506
10	81507	21367	11741	13550
11	81871	21490	11791	13620
12	82904	21626	12294	13748
13	83298	21731	12318	13842
14	83820	21900	12347	13948
15	84119	21966	12371	14008
16	84518	22041	12397	14091
17	84844	22151	12443	14147

In Fig. 5 the graphs are obtained dividing the value of each daily data by the number of the inhabitants of the corresponding city.

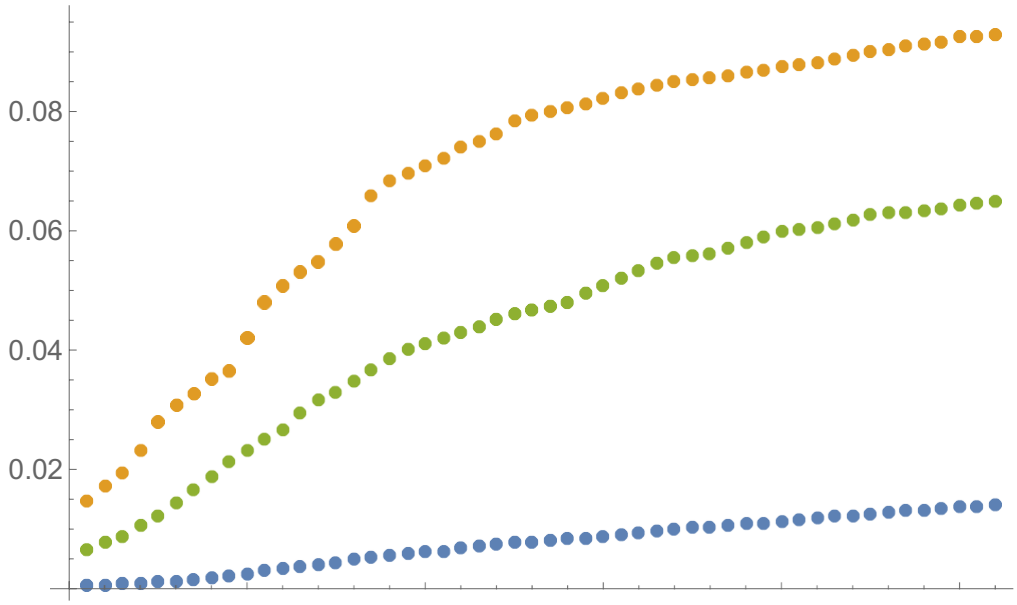

Fig. 5 In the graphs every dot represents the total number of infected people divided by the number of inhabitants, x_n/N, where N = 1390434 (Milan, blue), N = 122383 (Bergamo, yellow), N = 199415 (Brescia, green).

In order to make a comparison of data with respect to the different population of infected these values can be normalized substituting the value x_n of *the number of infected at* the n-th day by the normalized number, $\boldsymbol{x}_n$:

$$\boldsymbol{x}_n = (x_n - x_{min})/(x_{max} - x_{min}), \text{ where } x_{min} = min(x_n) \text{ and } x_{max} = max(x_n), \ (n=1,\ldots,N).$$

The normalized data are shown in Fig. 6

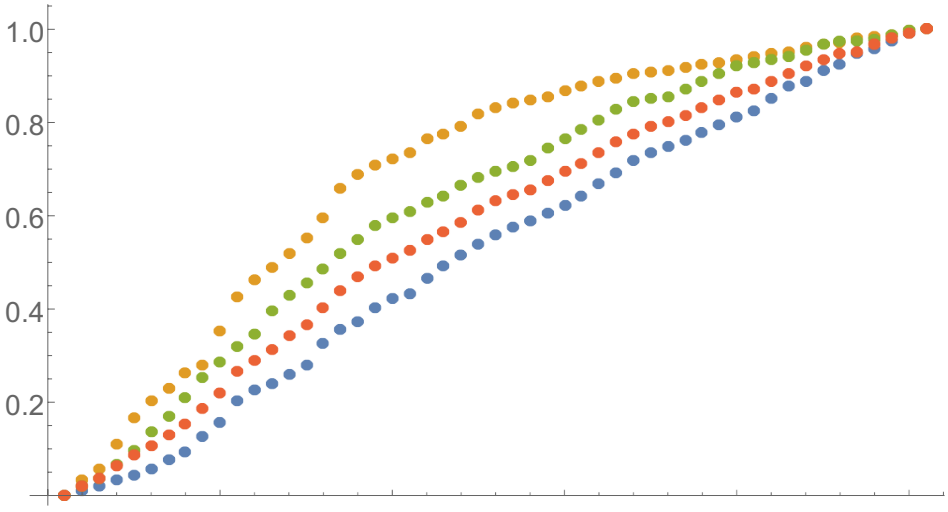

Fig. 6 Curves obtained normalizing data (see above): blue (Milan); yellow (Bergamo); green (Brescia); red (Lombardy)

5. Predictability, at a first glance, based on Brownian motion

Although many of the curves in the previous figures recall a logistic sigmoid, we still believe that the logistic model might be feasible only when there are some special conditions, as we have seen that none of the velocity growth graph exhibits a "Chinese" peak. However, some features of the logistic growth can be retained such as the total number of infected people N (limit value of the logistic sigmoid) and the first day where a trustable prediction can be performed (inversion point of the logistic). In this section we provide a forecasting model able to determine:

1) the value towards which tends the total number of the infected people, P;

2) the first day from which the forecast 1) applies.

Let x_n be the total number of cases in the n-th day, then we define the *mean* of the numbers x_k, when $k = 1, 2, .. , n$, as usual,

$$\mu_n = (\frac{1}{\mathrm{n}} \sum_1^n x_k),$$

and let's denote the standard deviation of the set $\{x_2, .., x_n\}$, by

$$\sigma_n = \sqrt{\frac{1}{n} \sum_k^n (x_k - \mu_n)} \ , \ n \geq 2.$$

Each daily data x_k is affected, for many reasons easy to understand, by an inevitable halo of randomness; thus, we'll treat the infected time series as the values of a *Brownian motion*.

The botanist Robert Brown observed minute particles, ejected by the pollen grains suspended in water he was studying under a microscope, executing a jittery motion: a Brownian motion. By repeating the experiment with particles of inorganic matter he was able to rule out that the motion was life-related (1827).

Albert Einstein provided a mathematical solution of the problem, mainly as a way to indirectly confirm the existence of atoms and molecules
Einstein recognized that Brownian particles as a collective set, because Mechanics is not able to follow the motion of each particle, obey to a diffusion equation whose solution gives the density of the Brownian particles, $\rho(x,t)$, at any position in space x and any instant in time t.

This solution is a gaussian type curve, well known to be characterized by its mean $\mu = 0$ and the square deviation σ that, in this model, *linearly* depends on the diffusion coefficient of the collective motion and *on time*. In fact, when time grows the curve enlarges itself, in such a way that at successive times $t_1 < t_2 < t_3$ ... the density of Brownian particles becomes flatter and flatter.

If one passes from the scheme in which the time is a continuous variable, like in the Einstein's reasoning, to a discrete one, more proper to describe the *daily* number (discrete time) of infected, it

can be shown that each new step – the n-th theoretical value to be compared with the n-th data - in the time interval dt is characterized by the sum of two terms: a linear trend depending on the data average μ and on time interval dt and a term depending on the standard deviation σ and the square root of time interval.

This means that two consecutive values of the series have a difference given by

$$x_n^* - x_{n-1}^* = x_{n-1}^*(\mu_{n-1}dt + \sigma_{n-1}\sqrt{dt})$$

For the time interval, dt, we set

$$dt = 0.00273973,$$

where $0.00273973 = 1/365$, that is, the time interval of one day in a year; then, the data of Table1 can be approximated by the following sequence:

$$x^*_n = x_1, \qquad n = 1,$$

(3)

$$x^*_n = x^*_{n\text{-}1} + x^*_{n\text{-}1} \,(0.00273973\ \mu_{n\text{-}1} + \sigma_{n\text{-}1}\ \sqrt{0.00273973}\ ), \quad n \geq 2,$$

In Fig. 7 are represented the graph of the values given by approximation (3) together with the sequence of data drawn from the third column of Table 1 (Bergamo); analogue graphs can be obtained for the other columns (Lombardy, Milan and Brescia).

Coming back to (3), each term x^*_n depends, for $n \geq 2$, on the knowledge both of the mean value $\mu_{n\text{-}1}$ and the standard deviation $\sigma_{n\text{-}1}$ of all the data of the day preceding the n-th ones.

As n increases, the sequence (3) will tend to a limit value $x^*\wp$; in that limit it will be:

(3') $\qquad x^*_n \approx x^*_{n\text{-}1} = x^*\wp \qquad$ and $\qquad 0.00273973\ \mu_{n\text{-}1} \approx 0;$

correspondingly the standard deviation too tends to zero $\lim_{x^*_n \to x^*_l} \sigma_n = 0.$

There is no need of a rigorous proof of the limit asserted in (3') because it can be caught at a glance.

In the worst case, the sequence (3) could oscillate tending to $x^*\wp$; however, our interest is that the difference between two successive terms x^*_n and $x^*_{n\text{-}1}$ – whether the oscillations exist or not – reduces itself under a fixed limit.

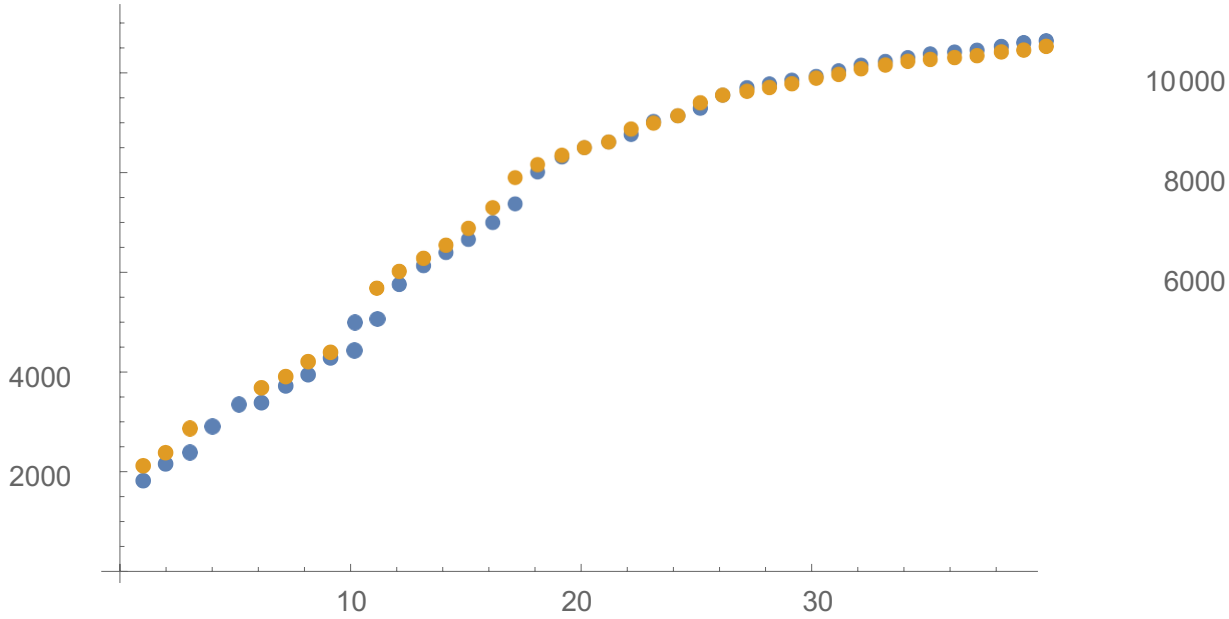

Fig. 7 Blue dots are the values estimated with (3), yellow dots (Bergamo) are the data from the third column of Table 1

Looking at Fig. 7 it's evident that the two sequences almost coincide starting from about the 20th day, showing that approximation (3) is fair enough, and mainly, that a day exists – $n^* = 20$, for Bergamo – such that, starting from it, the growth of cases could be reliably dominate by a suitable straight line, and the same will happen for the graphs relative to the other columns of Table 1.

The path to follow to determine such a straight line could be to impose that the modulus of the difference between a data of the sequence, x_n, and the corresponding value given by formula (3), x^*_n, is less than a fixed number ε

$$|x_n - x^*_n| < \varepsilon \; ; \tag{4}$$

then, let be n^* the first day for which (4) is satisfied, from that day we can write the sequence

$$x^{**}_n = x_{n^*} + (0.00273973\,\mu_n + \sigma_n \sqrt{0.00273973}\,)\,(n - n^*), \quad n > n^*, \tag{5}$$

that represents the straight line passing through the point

$$(n^*, x_{n^*})$$

and allows to perform an estimate (for excess) of the cases for all n after n^*. In the equation (4) the value ε can be estimated by a least square method, but it is enough to take as n^* the first day where the standard deviation is less than 0.01 (see Fig. 8).

By this way we can provide a straight line of prediction for the three cases we are studying, that is, for Bergamo (Fig. 9), Brescia (Fig. 10) and Milan (Fig. 11).

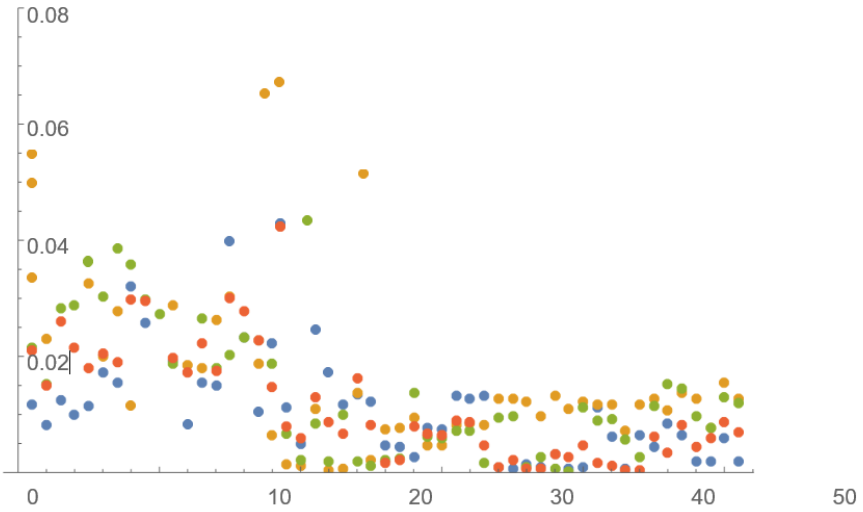

Fig. 8: Standard deviation (percentage) of Milan (blue), Bergamo (yellow), Brescia (green), Lombardy (red).

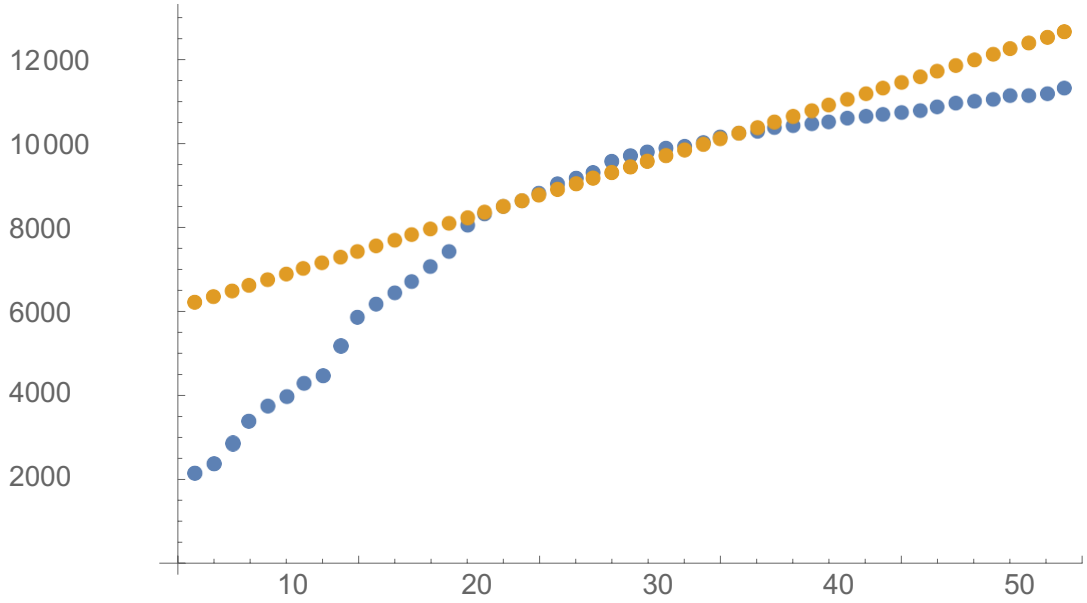

Fig. 9: Straight line of prediction for Bergamo: starting from $n^* = 20$, prediction points from (5) in orange; data series in blue

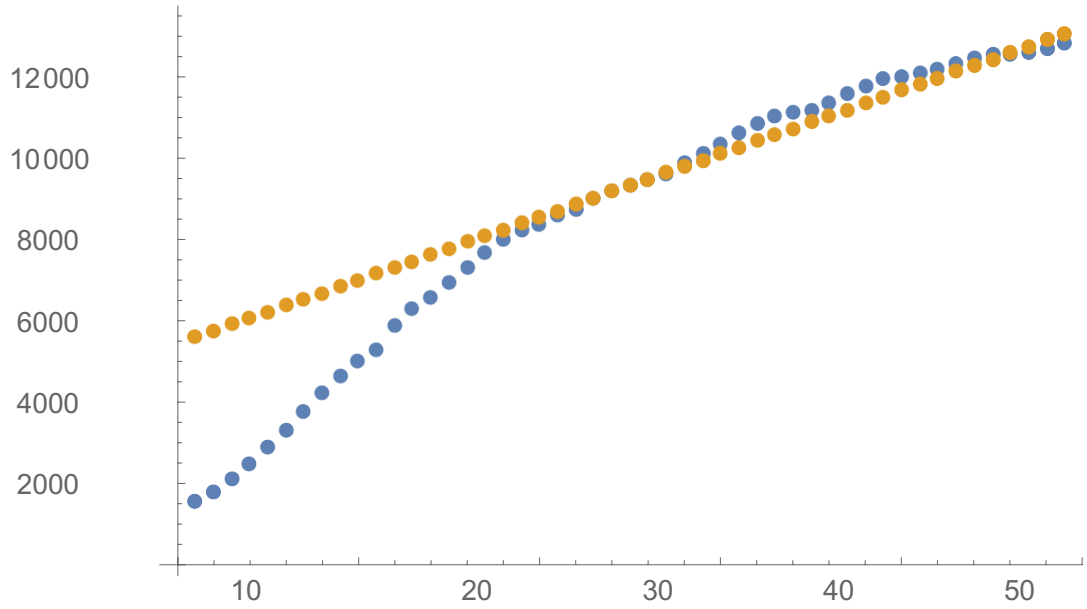

Fig. 10: Straight line of prediction for Brescia: starting from $n^* = 26$, prediction points from (5) in orange; data series in blue

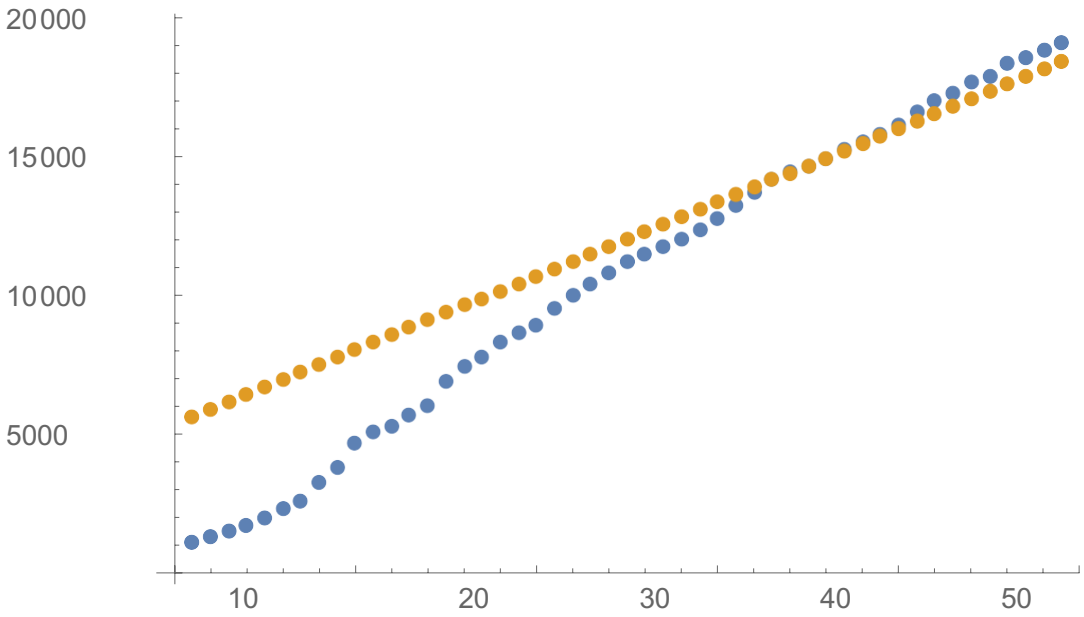

Fig. 11: Straight line of prediction for Milan: starting from n^* = 38, prediction points from (5) in orange; data series in blue

Results and discussion

The logistic model is characterized by many advantages but also well-known limits. However, preferring simplicity, we did not use more flexible but also mathematically more complicated models and probably capable of a more precise description of the covid-19 evolution and better predictive power.

The theoretical *n*-series (5) constructed from the *n*-data, taking in account the substantially random character of the data, fit well with data series, but oblige to perform some calculations not difficult but cumbersome when *n* grows. A calculus program could overcome this aspect.

If one hasn't got such a program, applying the "straight line prediction" yields interesting results. In the provided examples, on the 50th day, i.e. 9 May because for the initial day has been assumed 11th of March, only in the case of Bergamo the percentage difference between forecast value and data is about 9%; in the other two cases - Brescia and Milan - the percentage difference is under 2% (see Fig.10 and 11). On the whole, a not scarce approximation, especially when compared with many "official" dashboards that record data and represent them in figure such as Fig. 4, in which the "straight lines" are shown with a slope of 10%, 33% and 100% per day and a consequent great deviation from the real values.

Another significant feature of the latter approach, in the performed examples, is that day from which one can provide the prediction is, in the worst case (Milan), n* = 38, i.e., 46 days after the achievement of the first 100 cases, as WHO requires. A questionably timely prediction,

the latter, even though after that day, 22 April, the new cases added up to 9 July represent more than 40% plus. In the cases of Brescia and Bergamo, the prediction seems acceptable with respect to the time too, because $n^* = 26$ for Brescia is 33 days after the starting day (4 March); and for Bergamo $n^* = 20$ implies 30 days after its starting point (1 March).

Finally, our affection for the logistic curve from which we started relies on its being the more suggestive mathematical representation, with its asymptotic behavior, about how long the growth of infection lasts. An important feature, especially since many people speak of a "second wave" of Sars-CoV-2, ignoring that many viruses continue to live for generations in our environment or in our body, having weakened their viral load and lethality.