A Simple Model to Reproduce the Covid-19 Propagation

Enrique Castillo[1],[2]

[1] *Spanish Royal Academy of Engineering*
[1]*Spanish Royal Academy of Sciences*

Resumen

A collection of simple models to study the spread of Covid-19 is presented. After normalizing and discussing some properties that the models must have to be valid, it is shown that the accumulated affected (deceased or infected) curves have to be cumulative distribution functions. It is recommended to use models that give the growth of deaths based on their initial number and the time elapsed, demonstrating that these functions cannot be arbitrary, but must satisfy the translation functional equation, which is solved and interpreted appropriately. The solution of this functional equation shows that any interval of the affected curve contains information about all of it, so a sufficiently wide interval of it allows to predict the future, or to recover the past, from it. Application of the particular case of the normal distribution gives incredibly good adjustments for deaths. Finally, the most important conclusions of the work and some considerations on the precautions to consider the end of the quarantine period are given.

Key Words: Covid-19, statistical models for the pandemia propagation, Bayesian methods, functional equations, robust regression models, recommendations and precautions to end the quarantine.

1. Introduction

This brief note presents a very simple model to reproduce the spread of the coronavirus according to the official data provided.[1]. More precisely, a sufficiently regular curve is sought $M = M(t)$, that provides the number of people affected (deceased or infected)[2]. As a relevant result, the curves to be used are identified, which, after being normalized with the final number of deaths, turn out to be distribution functions.

To solve a problem, it is very important to know it well and understand the hypotheses and foundations of the models to be used. In these cases, models such as the SIR or SEIR have been used, in their various versions that include models with and without delays,

[1]The data provided exhibits a very poor quality and manifestly alters the possible models, not only because they do not know exactly what these data measure, but because it is known that the different autonomous communities measure deaths differently and there are some delays that are systematic, such as those that occur on weekends.

[2]From now on we will talk about the deceased up to day t, but understanding that everything also applies to the case of the infected people.

age ranges, time variations in parameters, etc., and many others, which try to reproduce how contagion and transmission of the disease occurs. This is a very valuable physical and medical support for the models. Therefore, the type of curves to be used is quite well studied and discussed, which serves as the basis for our proposal.

In particular, to reprodice the same dynamics of the above models, we are interested in the increasing sigmoidal functions, that is, with an inflection point at $t = t_m$, which corresponds to the maximum of the derivative curve $M'(t)$. This means that initially $M'(t)$ increases and later, it starts decreasing, starting at t_m, which leads to a decrease in infections until they are cancel (either suddenly or asymptotically).

To start, it is important to note that a dimensionless model should always be used. Therefore, in our case, we assume the model

$$M^* = \frac{M}{M_\infty} = M^*(t^*), \tag{1}$$

where M_∞ is the final number of people who died during the whole process, and $t^* = t/t_m$. Note that we used the two most relevant parameters, the number of total deaths at the end of the process, M_∞, and the instant, t_m, at which the maximum of the derived curve (daily) is reached.

Note that we use dimensionless variables, as recommended by Buckingham's Π theorem, and we use asterisks to refer to them.

Since this curve must be increasing and the already normalized range of $M^*(t^*)$ is the interval $[0, 1]$, the function $M^*(t^*)$, of the total number of deceased persons as a function of time, is a distribution function and its derivative, $M^{*\prime}(t^*)$, which gives the number of daily deaths, is the corresponding density function.

Therefore, we have as many models as sigmoidal distribution functions to choose from.

In particular, we will choose the function $M^*(t^*)$ from a parametric family of distributions, which will depend on a number p of parameters, among which it is convenient to choose t_m as one of them, and the previously indicated parameter M_∞. Once this family is chosen, we will look for a family curve, estimating the parameters, or, if we want to do it better, we will look for mixtures of them, as Bayesian models do, for our final proposal [3].

Interesting particular cases are the normal, gamma, logistics, Student t, Weibull, Gumbel and Frechet families, these last three in their maximum and minimum versions, etc. This leads us to a maximum of 4 parameters, in these particular cases. It should be noted that in the normal case, the mean μ is the parameter associated with t_m.

[3]Note the great difference between choosing a family curve of the selected family and consider any of the possible mixtures among them. This is equivalent to assuming that different persons or groups respond to a curve of the family and that the posterior distribution indicates what weights must be assigned to each of them.

2. Some preliminary considerations

To solve our problem we can choose the function $M^*(t^*)$, which gives the number of death persons as a function of the time t^*, or, better yet, we can search for a family that allows us to proceed in stages, that is, search for a function $M^*(t^*; M_0^*)$ that gives us the number of deceased when starting from an initial number of deceased M_0^* and increasing the time by t^* units.

It should be noted that the function $M(t^*; M_0^*)$ cannot be arbitrary, but must meet an important consistency condition, which establishes the following:

> *The number of deceased that are obtained from any moment in which there are already M_0^* deceased after a time $t_1^* + t_2^*$, should be the same if they are calculated directly, using $M^*(t_1^* + t_2^*; M_0^*)$, than if the deceased are calculated at time t_1^*, and using them as the initial value, those resulting after a time t_2^* are calculated, that is, using the expression $M^*(t_2^*; M^*(t_1^*; M_0^*))$.*

In other words, the function $M^*(t^*; M_0^*)$ must satisfy the functional equation:

$$M^*(t_1^* + t_2^*; M_0^*) = M^*(t_2^*; M^*(t_1^*; M_0^*)), \tag{2}$$

which is the well-known functional equation of the translation, whose solution is [4]:

$$M^*(t^*; M_0^*) = \phi\left(t^* + \phi^{-1}(M_0^*)\right), \tag{3}$$

where $\phi(t^*)$ it is an arbitrary (increasing) invertible function and $\phi^{-1}(M_0^*)$ defines a traslation for each value of M_0^*.

The, the looked for family $M^*(t^*; M_0^*)$ can be generated by the function $\phi(t^*)$ including all possible translations, that is, this family must be stable with respect to translations.

In addition, in the equation (3) we check that, as it must be, we have

$$M^*(0; M_0^*) = \phi\left(\phi^{-1}(M_0^*)\right) = M_0^*, \tag{4}$$

and

$$M^*(0; 0) = \phi\left(\phi^{-1}(0)\right) = 0, \tag{5}$$

It is interesting to see that the degrees of freedom that were initially available to choose $M^*(t^*; M_0^*)$, that is, a function of two arguments, they are severely reduced to choosing a function of a single argument and invertible, if the compatibility condition is to be met, making it clear that, if this function is not chosen this way, the final results will depend on the times selected as initial times to make the propagation.

Since every function of the form (3) fulfills the compatibility condition (2), one wonders if given a function of this family, there will be more than one function ϕ that reproduces it, that is, we can ask ourselves about the uniqueness problem.

[4] This is true under very weak conditions of point continuity of the function.

Theorem 1 (Uniqueness theorem) *If there are two functions ϕ_1 and ϕ_2 that generate the same function, that is, they satisfy the condition:*

$$\phi_1\left(t^* + \phi_1^{-1}(M_0^*)\right) = \phi_2\left(t^* + \phi_2^{-1}(M_0^*)\right), \tag{6}$$

then, we must have

$$\phi_1(t^*) = \phi_2\left(t^* - \mu^*\right), \tag{7}$$

where μ^* is an arbitrary constant.

The equation (6) is a functional equation with two unknown functions, ϕ_1 and ϕ_2, whose solution is (7), which means that the solution is unique, except a translation.

A graphic illustration of this result is presented in figure 1, where it is seen that starting from M_0^*, on the left, we obtain $\phi_1^{-1}(M_0^*)$, using the red curve, we add t^* and then, using the red curve again, we get $\phi\left(t^* + \phi^{-1}(M_0^*)\right)$. It can be seen that doing the same with the blue curve, the same result is obtained, since the blue curve is a horizontal translation of the previous red curve.

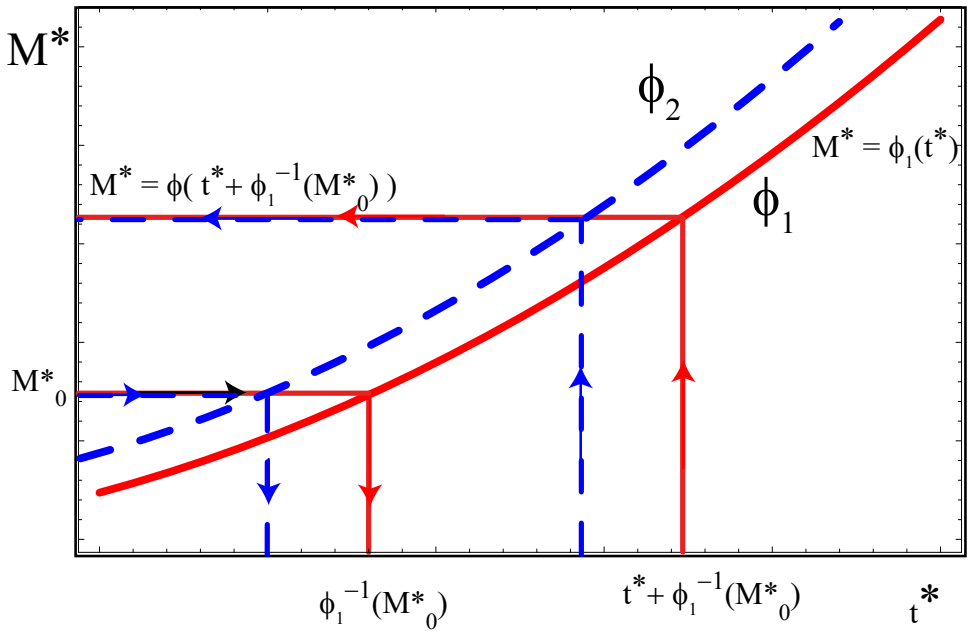

Figura 1: Graphic illustration of the uniqueness theorem.

Another interesting question we can ask ourselves is: what is the $\phi()$ function and how is it chosen in a practical case? The answer to this question is as follows.

From equation (3) we have

$$M_0^*(t^*) = M^*(t^*; 0) = \phi\left(t^* + \phi^{-1}(0)\right), \tag{8}$$

where we have denoted $M_0^*(t^*)$ to $M^*(t^*;0)$, which shows that $M_0^*(t^*) = \phi\,(t^*)$ if $\phi^{-1}(0) = 0$, that is, the function $\phi\,(t^*)$ is one of the functions of the family $M^*(t^*;M_0^*)$.

Consequently, the looked for family $M^*(t^*;M_0^*)$ can be generated from any member of the family, for example $\phi(t^*)$, including all possible translations.

3. The contagion DNA of the curve is contained in any interval of the curve

The relevance of the translation equation (2) goes much further than it initially appears. The expression (3) gives us $M^*(t^*;M_0^*)$, showing that the dependency of M_0^* is defined by the function $\phi^{-1}(M_0^*)$, and that the dependency of t^* is the one given in (3), that is, perfectly known in all compatibility cases.

In fact, the DNA of the time variation of the growth curves of the deaths is contained in any time interval of the same. This means that known $M^*(t^*;M_0^*)$ in any interval of t^*, the complete time variation can already be known. However, since the local information is not sufficiently precise, a sufficiently wide finite time interval must be used, but the idea is still valid, that is, the future behavior of the same can be predicted from the past, and vice versa.

Furthermore, the advantage of using $M^*(t^*;M_0^*)$ instead of $M^*(t^*)$ is that we can choose any initial value M_0^*, since the model is independent of M_0^*. In this way, we can avoid the first data of the spread, since at the origin of the epidemic, the data show great variability. Once the contagion process has started, with higher values of M_0^*, as more people are contagious, the process stabilizes and the estimation is more robust.

If we choose a family of distributions, that is, if $M^*(t^*;M_0^*,\theta^*)$ is used, where θ^* is the vector of its parameters, a time interval can be used, which implies an interval of observations of $M^*(t^*;M_0^*)$, to estimate these and thus predict the future and, if desired, recover the past. Obviously, the closer this interval is to the instant you want to predict and the wider it is, the better estimates you will get.

As indicated, the need for a wide time interval is due to the impossibility of obtaining sufficiently accurate data.

The explanation for all this is that the chosen model, that is, the fact that they are distribution functions after normalization, together with the specific families of chosen distributions already include the dynamics of the pandemic system. The same occurs with medical doctors, who, after examining the patient and diagnosing the disease, will be able to predict in advance the stages that the patient will follow and the closer the prediction time is to the predicted event, the more precise it will be.

All this highlights the depth and beauty of functional equations and mathematical models, which are capable of reproducing many aspects of reality that seem implausible [5].

[5]This behavior we have already observed in other models, especially fatigue models , in which we have been working for many years. The DNA of fatigue behavior of a material, that is, the progress of the

4. Proposed model

Having clarified the above ideas, we propose, as a simple example, the normal family of distributions for our model, which turns out to be:

$$M^*(t^*) = \Phi\left(\frac{t^* - \mu^*}{\sigma^*}\right), \tag{9}$$

where $\Phi()$ is the cumulative distribution function of the standard normal distribution $N(0,1)$, that is, our parameters are $M_\infty, \mu^* = \mu/t_0$ and $\sigma^* = \sigma/t_m$, where the Greek letters μ and σ represent the means and standard deviations, respectively

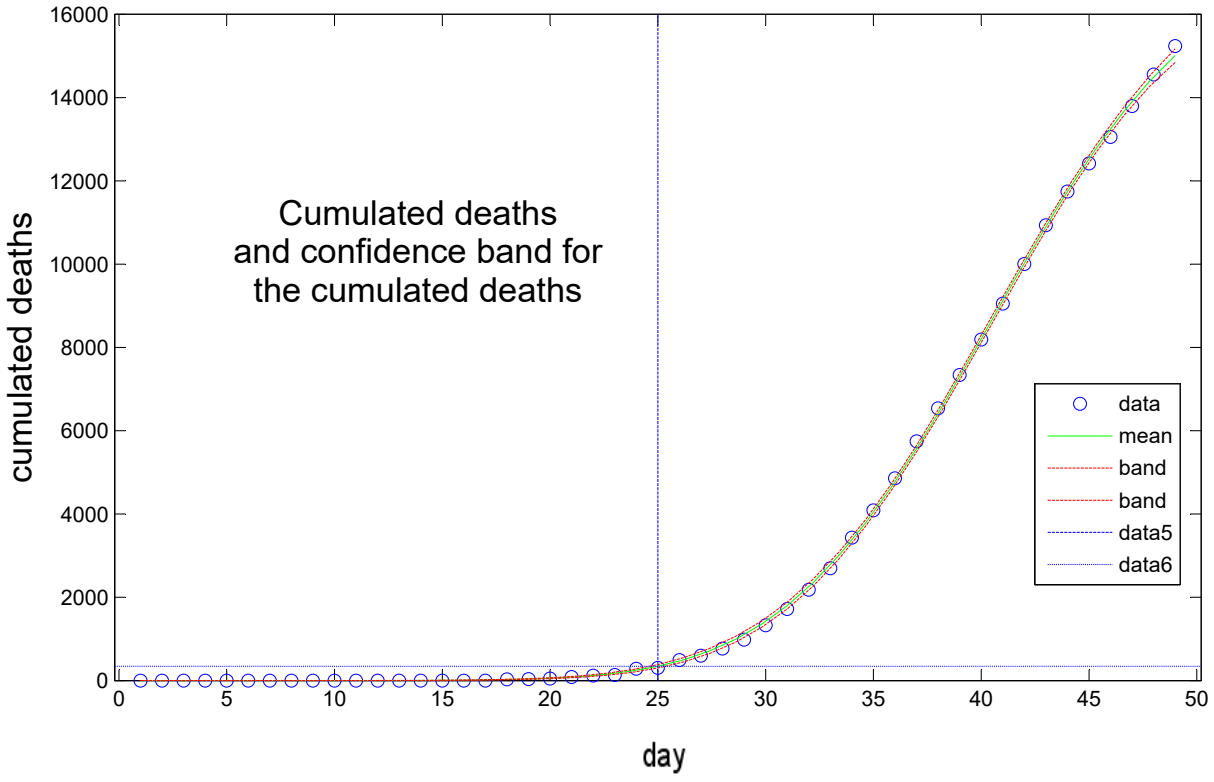

Figura 2: Accumulated deaths and confidence band 0.95 for the mean of the curve.

If we estimate the parameters using this regression model, using matlab and its "nlinfit" function for robust regression, and the "nlintool" tool, we obtain the models in figures 2 and 3 and the one in figure 4, if we adjust the distribution function and the density function, respectively.

cracks are marked by these before we test them, and, in fact, it is not necessary to arrive at their break to predict when it will occur. And not only this, but we can guess what the previous process of cracking was like, even when it was not observable.

The fit is amazing for the accumulated curve and quite good for the density function [6]

The different width of the bands is due to the fact that one of them refers to the prediction of the mean and the other to that of isolated data for each day, the prediction of the first being much easier (more accurate) than that of the second [7].

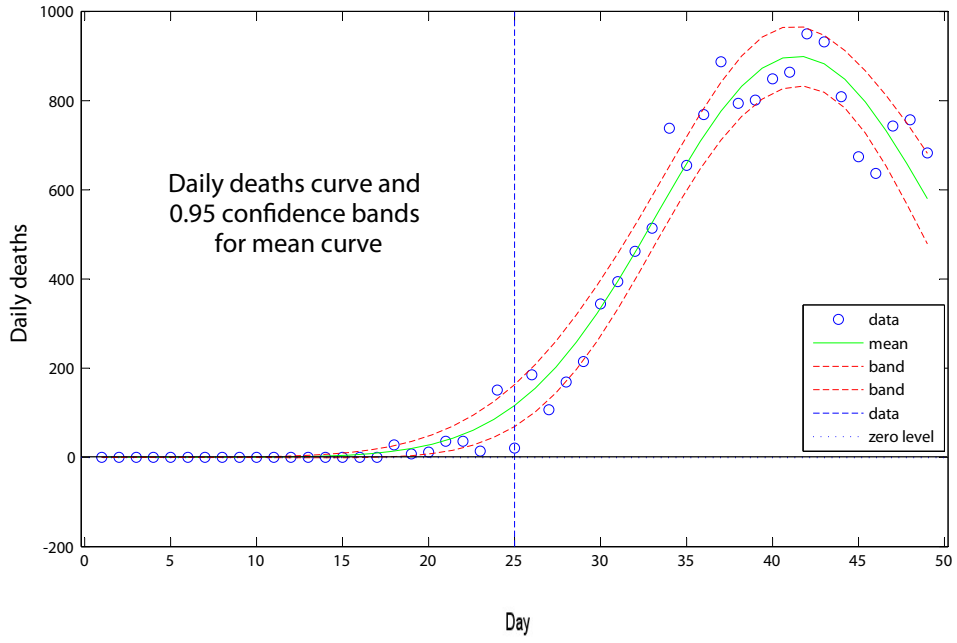

Figura 3: Daily deaths and confidence band 0.95 for the mean of the curve.

Finally, it should be noted that this normal model implies a symmetry in the propagation curves on both sides of the maximum of the daily deaths curve, which might or might not correspond to reality. Therefore, the other indicated models or even different ones can be used, when applicable.

In closing, it would be very interesting to use these models considered as Bayesians with MCMC sampling, such as those used by WinBUGS or OpenBUGS. Unfortunately, I was unable to fit the model that I expected, possibly due to deficiencies in the implementation of these methods in free software or poor data quality, which affect the result.

Finally, indicate that Bayesian methods, when considering the random parameters, include as candidate curves all the convex linear combinations (mixtures) of those of the

[6] Roberto Mínguez and Javier Girón tell me that the logistic distribution also gives very good results, and Javier Samper that this normal model fits both deceased very well like infected.

[7] Daily data is clearly altered, especially on weekends, and could be corrected at a preprocessing stage.

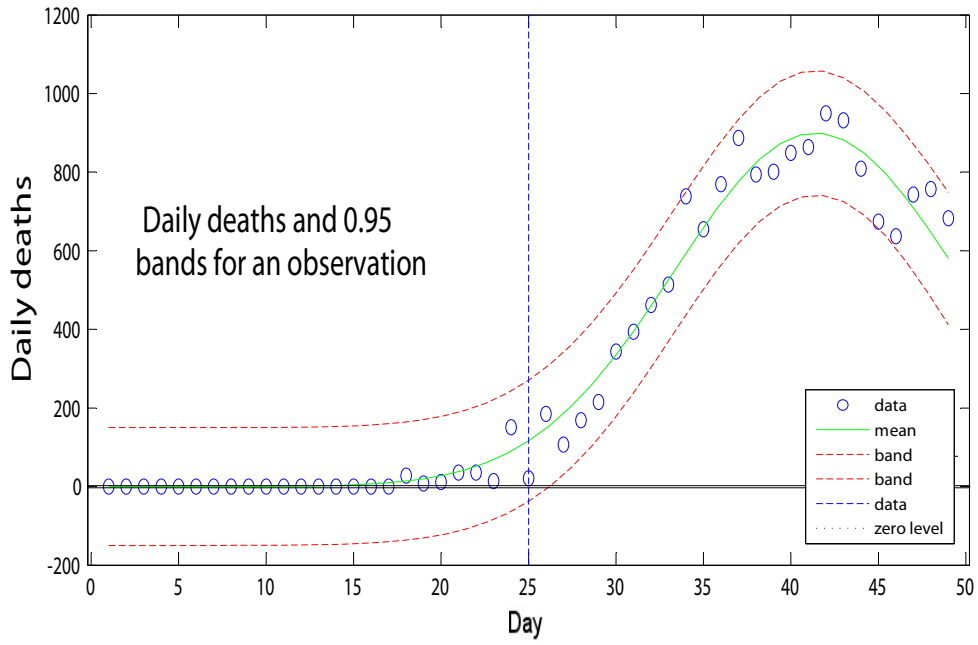

Figura 4: Daily deaths and confidence band 0.95 for an observation.

chosen family. Furthermore, if the sample is large, the results are practically independent of the chosen "a priori" distributions.

5. Conclusions

The most important conclusions that can be derived from the above are the following:

1. It is important to always use dimensionless models to solve problems, which is still valid in this case of spread of the epidemic.

2. Given the random nature of the problem, although there are many valuable deterministic methods to study it and derive its dynamic, they are clearly insufficient and it is necessary to use stochastic or random methods.

3. In the case at hand, increasing sigmoidal models with an inflection point should be used, which marks the expected and desired maximum of the curve.

4. We must be careful to choose models that allow updating the accumulated number of deaths at any moment, for which an arbitrary function cannot be chosen. The functional equation of the translation clarifies for us which are valid and which are not.

5. The ϕ functions, which generate these valid functions and must be invertible, have been given and explained what they are and represent.

6. To any increasing function that covers the total range, we must add a location parameter (translation) to obtain a valid family for this indicated update, although more parameters can and should be added.

7. With the indicated dimensioning, the resulting curves are cumulative distribution functions, so depending on the trends, we can know which ones could be optimal in each case.

8. The time variation of the curve is contained in any interval of the curve and can be calculated from a sufficiently wide finite interval of the curve. This allows predicting the future and recovering the past, even if it is not observable.

9. Bayesian methods consider not only family curves, but convex linear combinations of them (mixtures), which implies totally general models.

10. Given the lack of quality and the deficiencies of the available data, all of the above is only an aid in the search for solutions to get out of the confined state, making no sense to improve the methods used with such poor data and changing criteria.

11. Indicate that the maximum of the derived curve corresponds to the first phase of the process, there being a second phase, which may be longer than the previous one. Therefore, I consider it very risky to start leaving the confinement as soon as possible. An error in this sense would have catastrophic consequences, since to the damage of having to repeat the previous process, the psychological consequences of it would be added.

12. Finally, since the lethality of the virus is almost nil in healthy young people, they should be the first to emerge, but only in the case of not being in contact with people of high age and/or risk. Governments should not forget that to decide who leaves the confinement, it is not enough to analyze the risks of transportation and workplace of the workers who leave, but it is necessary to consider, mainly, the risk and age of the people they live with. Special consideration should be given to nursing homes, to which, in almost all cases, the virus has been brought from outside them.