Estimating the number of COVID-19 victims by combining the general logistic model with the pCN Monte Carlo sampling

June 8, 2020

1 Introduction

A simply and effective way to quantify the impact of an epidemic is given by counting its number of victims. To model its variation in time, we choose a generalized logistic map $X^P(t)$ governed by an unknown parameter $P \in \mathbb{R}^3$. We assume that, when put into the model, the "true" P is capable of coherently fit the past number of victims, as well as predict its future evolution.

Instead of looking for a single precise value, we rely on the Preconditioned Crank-Nicolson Monte Carlo method to establish a multitude of candidates for P, each capable of describing a possible future scenario. We try to establish a "worst", a "best" and an "expected" case in which we are likely to go.

We work with three European Countries (Italy, Germany, France) and our final question is: *under the assumptions above*, how many weeks of data do we need to correctly predict the number of victims for (approximately) at least one future month?

We conclude remarking the importance of measurements errors, showing how they are *probably* lower in the case of Italy and France. Successful predictions until the mid of May by using the first three weeks of April are shown. Surprisingly, they are still remaining valid (today being the 1st of June) despite the change of lockdown measures in all the considered Countries (the virus incubation time being already included in the computations).

2 The generalized logistic model

We choose to model the number of deaths in time with a generalized logistic function. Recall that these maps are all S-shaped curves, intuitively seen as an exponential start, then slowed down until reaching an horizontal equilibrium.

We do not provide any deep epidemiological explanation for such a choice, except that we expect a similar qualitative behavior for the phenomenon under analysis: the number of deaths is supposed to be higher at the beginning,

reaching then a saturation point as long as the health system becomes capable of managing the emergency.

Of course, we are not new in proposing these maps: in the literature, there is a wide range of successful applications in similar fields (REF).

Our general logistic map is given by the following 1-dimension ODE:

$$X'(t) = \frac{q}{v} X(t) \left(1 - \left(\frac{X(t)}{Q}\right)^v\right) \tag{1}$$

with closed form solution:

$$X(t) = \frac{Q}{(1 + A\exp[-qt])^{\frac{1}{v}}} \tag{2}$$

for three *strictly positive* real parameters, $P = \{q, Q, v\}$ and initial time zero condition X_0 (we write $X(t) = X^P(t) = X^P_{X_0}(t)$ to stress this dependence). The letter A abbreviates $A = -1 + \left(\frac{Q}{X_0}\right)^v$.

Due to the limit $\lim_{t\to\infty} X(t) = Q$, we interpret Q as the maximum value reached asymptotically, sometimes referred as the *carrying capacity* in the literature.

The parameter v is related to the symmetry of the curve. Its limit for $v \to 0^+$ produces the *Gompertz* model, while the case $v = 1$ the *simple* logistic map.

The simple logistic model suggests an interpretation for the quantity q, too. In such a case ($v = 1$) the equation would be:

$$X'(t) = qX(t) \left(1 - \left(\frac{X(t)}{Q}\right)\right) \tag{3}$$

For very small time the value $X(t)$ is surely far lower than Q (otherwise the system would be already in equilibrium), therefore we would have $X'(t) \approx qX(t)$, reading q as the rate of this starting exponential growth.

3 The bayesian approach

Let $X^P_{X_0}(t)$ be always the generalized logistic model above, depending on parameters P and with X_0 time zero condition. When using the model in practice, we can only observe a limited amount of points coming from its ODE trajectory, which are furthermore perturbed by a noise due to measurement errors. Let's fix $T + 1$ times $\{t_i\}_{i=0,\dots,T}$. In our experiments, T is generally between 7 and 21 days.

Definition 1. *The* observed vector $\boldsymbol{y} \in \mathbb{R}^{T+1}$ *is the random variable defined componentwise as:*

$$y_i(\omega) = X^P_{X_0}(t_i) + \eta_i(\omega) \tag{4}$$

where $\eta_i \sim \mathcal{N}(0, \sigma_i^2)$. [1]

[1] Sometimes by an abuse of notation we use the symbol η_i to indicate its density function too, so writing $\eta_i(x)$ for $x \in \mathbb{R}$ refers to that.

In other words, y represents the actual measurements on which we assume the influence of random errors. An important assumption that we make, lies in the belief that the errors follow the Gaussian structure above. In principle, many other possibilities can be chosen. In practice, the choice above is pretty standard and seems to work reasonably well. In our numerical simulations we take into account two possible degree of trust. In the *reliable* case we consider the measurements to carry "up to a 10% of error", in the sense that the "true" value at time i is supposed to be likely in the interval $y_i \pm \frac{y_i}{10}$. In the *unreliable* case, we increase such a level to 100%. Recalling that $\eta_i \sim N(0, \sigma_i^2) \implies \mathbb{P}[-2\sigma_i \leq \eta_i \leq 2\sigma_i] \geq 95\%$, choosing then the standard deviation $\sigma_i \doteq \frac{y_i}{20}$ gives the 10% error tolerance on y_i, while choosing $\sigma_i = \frac{y_i}{2}$ covers the latter case [2].

Once that the model's structure is clarified, the next step is to actually estimate the unknown $P \in \mathbb{R}^3$ *given* the observation of $\mathbf{y}$. If we interpret P as a random variable, it means that the unknown quantity is given by the conditioned probability $\mathbb{P}[P|y]$.

Following the traditional Bayesian approach, we start by choosing a *prior* probability measure $\mathbb{P}[P]$ on $\mathbb{R}^3$, representing our blind guess about P *independently* of the observations $\mathbf{y}$. In practice this is sometimes an hard guess which can strongly influence the results. We will carefully describe our choice in a dedicated section.

If, as just pointed out, we aim at understanding $\mathbb{P}[P|\mathbf{y}]$, then the classical Bayes's law $\mathbb{P}[P|\mathbf{y}] \propto \mathbb{P}[\mathbf{y}|P] \times \mathbb{P}[P]$ allows to find it multiplying the the prior by an hypothetical conditioned law $\mathbb{P}[\mathbf{y}|P]$: this is where the notion of likelihood comes into play.

Definition 2. *For every fixed choice of the parameters P, the likelihood functions for the observation of $\boldsymbol{y}$, given P, is defined to be:*

$$\mathcal{L}(\boldsymbol{y}|P) \doteq \frac{(2\pi)^{-\frac{T}{2}}}{\sigma_0 \cdots \sigma_T} \exp\left(-\frac{1}{2}\sum_{i=0}^{T} \frac{(y_i - X^P_{X_0}(t_i))^2}{\sigma_i^2}\right) \tag{5}$$

By *interpreting* the likelihood as an effective probability conditioning, writing informally "$\mathbb{P}[\mathbf{y}|P] = \mathcal{L}(\mathbf{y}|P)$", not only the formula is explained when looking at the noise distribution $\eta(\mathbf{y} - X^P)$, but the fitting problem has now a clean solution:

Definition 3. *The (Bayesian) answer to the problem "Finding the probability density of the parameters P given the observations $\boldsymbol{y}$" is given by the* posterior *distribution on $\mathbb{R}^3$ defined by:*

$$\mu(dx) \propto \mathcal{L}(\boldsymbol{y}|x)\rho(dx) \tag{6}$$

where $\mathcal{L}$ is the likelihood defined above and ρ the prior probability for P.

[2] The second error is much bigger that the first, but still not huge enough to "destroy" the algorithm as we checked by running multiple simulations on toy models data.

During every use of the Bayesian rule, we constantly omitted the denominator relying always on the proportionality "$\propto$". This is because such a value is always the probability normalization constant, a number that can be completely ignored in practice thanks to the use of suitable Monte Carlo techniques.

3.1 The Preconditioned Crank-Nicolson MCMC

In the previous section we revised the Bayesian algorithm as a tool to convert the problem of estimating the parameter $\mathbb{P} \in \mathbb{R}^3$, to the task of sampling from the *posterior* probability μ on $\mathbb{R}^3$. Its statistical properties convey the information that we need: for instance, the mode can be read like "the most probably choice for P", its variance suggests how the uncertainty is spread, and similarly for other quantities. In principle one can analyze the posterior analytically, being explicitly given by formula NUMBER. But usually it is too hard to be done with pen and paper. Our answer it to use numerical tools: first we produce a large amount of samples, then we analyze them statistically.

In order to produce a (single) sample from μ, we use the so-called Preconditioned Crank-Nicolson Markov Chain Monte Carlo (pCN). Briefly speaking, it is a variant of the classic Gaussian Random Walk, but mainly used for much more complicated cases where for instance one can *arbitrarily* refine the ODE trajectory, or when the parameters P belongs to an infinite dimensional Banach space, or simply in the context of PDEs inverse problems. None of them is our case, since we have just $P \in \mathbb{R}^3$ and only a daily limited amount of data.

Therefore there is no solid a priori justification for this specific Monte Carlo strategy, and one is free to choose another sampling method. We plan to test and study their various performances, but not now: since the final results here are reasonably fine, we decided to keep this algorithm and leaving potential room for future improvement.

The reader is invited in consulting REF for a general exhaustive description of the pCN MCMC. In our simpler cases, the algorithm follows.

Given two candidate parameters $u, v \in \mathbb{R}^3$ define the acceptance probability:

$$a(u, v) \doteq \min\{1, \frac{\mathcal{L}(\mathbf{y}|v)}{\mathcal{L}(\mathbf{y}|u)}\} \tag{7}$$

and set the *exploratory* parameter $0 < \beta_{pcn} < 1$ (see step 3). The pCN *requires* the prior distribution $\rho(dx)$ to be a *centered* Gaussian, and used it as proposal step. This might be, in principle, a strong limitation: more comments about the prior will follow in a dedicated section.

To produce a **single sample** from μ, construct a chain $\{x_i\}_{i\in\mathbb{N}}$ as follows:

1. set $x_0 \in \mathbb{R}^n$ **arbitrarily**. Then, for each $k > 0$:
2. sample a point $R \in \mathbb{R}^3$ from the gaussian prior distribution $\rho(dx)$;
3. propose a candidate as $\hat{x}_k = \sqrt{(1-\beta^2)}x_{k-1} + \beta R$;
4. accept it (i.e. set $x_k = \hat{x}_k$) with probability $a(x_{k-1}, \hat{x}_k)$;

5. (accepted or not) repeat from 2;

To avoid overflow problems, it is suggested to work with logarithms in the acceptance formula. Define N to be the integer at which we (always) stop the chain, so that every time that we start from a point x_0, the chain produces a sample x_N. The value for N must be chosen in a way to overtake the known *burning time* issue. We always set it to be $2^{17} > 130000$.

Repeating multiple times the algorithm with *different* starting points x_0, we collect large amounts of samples, $2^{18} > 250000$ in our simulations. The *correlation* between them, a typical issue when when using a *single* traditional Markov Chain, is therefore hopefully well mitigated by the use of multiple chains. Note how every chain can be executed independently, therefore the total sampling procedure is suitable to parallelization (the user must be warned about the use of a proper seed).

Concerning the conservative parameter β, it was usually tuned between 0.01 and 0.001, choose in order to produce chains with an acceptance rate around 25%. Finally, we still have to comment the prior $\rho(dx)$ and the starting points x_0, but this will be done in a section later.

4 Tuning the remaining parameters

4.1 Comments on the assumptions and methodology

We apply the previous theory in order to formulate hypothesis for multiple European Countries concerning the future number of deceased people. We firstly use part of the available data at the beginning of April to predict the behavior until the mid of May (now known), with the intention of verifying them, and then repeating the same methodology for the month of June. Different numbers of weeks are taken as input data in order to compare their prediction' strength. A lot of careful points must be checked.

4.1.1 Why choosing an autonomous ODE

In the first section we justified the choice of a logistic map by looking at its S-shaped qualitative behavior, but we didn't comment the importance of keeping a curve described by an autonomous ODE.

Since the data in all Countries start around the mid of February, one might be tempted to use the *entire* collection (until, e.g. the 10th of April) relying on the intuition "more data, more accuracy".

As a first intuitive remark, recall that the general logistic model is connected to the simple logistic map ($v = 1$). For the latter one can show that, when used to model the number of infected, the map can be reformulated as a SIS compartmental model where q is also connected to the average number of people's daily social contacts. But since lockdown measures have been adopted, this quantity strongly changed in time. As a consequence of that, although

we have no 100% rigorous justification, we prefer **not** to trust the entire available dataset, but rather limiting the observations on time period with *constant* (stronger or weaker) lockdown measures. When following this logic, three cases can happen:

1 the "true" process is actually a general logistic one, but it already started *before* we began collecting data;

2 same as above, but the ODE *still* has to start, implying that part of the initial data that we read belongs to another model/trajectory;

3 the "true" model is not a general logistic one at all.

Of course, we cannot do nothing against 3, but the final results suggest that a general logistic model might be a "not-too-bad" choice. Conversely, by adding some days of delay we can increase the possibility of not being in 2. Finally, dealing with the problem number 1, it is necessary to ensure that the skipped days (e.g. including the one skipped in 2) can actually be "forgotten". But when choosing an autonomous ODE, this is precisely what claimed by the semigroup property of its associated flow. Briefly speaking, no matter if we start observing a value, say V_n, at day n, or V_{n+1} at day $n+1$: the trajectory produced from day $n+1$ (with initial conditions V_{n+1}) is precisely the same as the one produced by beginning at day n (with initial condition V_n). The two trajectories *are ruled by the same values of* P, therefore they infer the same parameter P (i.e. one can safely start the observations from the day $n+1$). Of course, one should not abuse this fact: the more are the observed data belonging the *same* trajectory, the better the estimation algorithms generally work. In other words, if one is sure that the ODE started at day n, one should take it as starting day and not $n+1$, in order to improve the estimator's performance.

4.2 Choosing the prior distribution

The goal of this section is to explain in which way we choose the prior probability measure $\rho(dx)$ for $\mathcal{P} = \{q, Q, v\}$. Since we used the pCN MCMC, it is required to be a centered Gaussian. It might represent a strong limitation, especially if we consider the fact that all the parameters in the general logistic equation are supposed to be positive. In order to circumvent this problem, we tried with a very straightforward approach (and it worked): we started the chain always with positive values, and when negative were proposed, we simply discarded them.

Consequently we are technically not considering the full posterior distribution, but just its part with positive values. It might open some theoretical issues that we have not yet carefully explored, but for this specific work our emphasis went on producing quick and effective results, postponing for the future a classic complete mathematical analysis.

Remember that the role of ρ is to represent the expected range of the searched values. For simplicity we define it with an easy covariance form:

$$\begin{pmatrix} \sigma_q^2 & 0 & 0 \\ 0 & \sigma_Q^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{pmatrix} \tag{8}$$

splitting so the prior into three independent one dimensional Gaussians, one for every parameters. Recall that if $X \sim N(0, \sigma^2)$, the classical quantile property claims $\mathbb{P}[-2\sigma \leq X \leq 2\sigma] \geq 0.95$. In the first section we recall how we prefer having $v \in [0, 1]$, so by using the formula above we suggest $\sigma_v = 0.5$.

We remind that q is connected to the simple logistics's starting exponential growth, while Q relates to the maximum number of possible cases. Our idea is simple: as a first move, we do a rough exponential interpolation on the given dataset. Since we always obtained very small exponential powers, we use this knowledge to set $\sigma_q = 0.05$ (although we are using a general logistic map, and not just the simple logistic model). Furthermore, the exponential law gives a rough prediction about the number of the deaths around, we fixed, the $20th$ of May (i.e. one Month after the datasets); call it N_e. We set $\sigma_Q = \frac{N_e}{2}$, in order to represent the idea "a good candidate for the true number of deceased people is likely below the prediction done with the exponential case, being that the most aggressive accepted model".

Finally, it remains to tune the starting point $x_0 \in \mathbb{R}^3$ for every Markov Chain. Let D be the least read data from the dataset, i.e. the current number of deaths. Not surprisingly, we initialize x_0 uniformly randomly in $[0, 1] \times [D, N_e] \times [0, 1]$, but we highlight how such a choice does not play an active role: we also tried with various intervals on comparable ranges, and the results were the same (as it must be).

5 Numerical results

We took into account Italy, France and Germany with the idea of answering to the following question. Let's suppose to be in a stable non-exponential situation. How many weeks of data do we need to perform a long-lasting prediction for the number of deaths?

For each Country, we looked for the day when the lockdown measures started, and then waited around 15 days (i.e. the virus incubation time, in order to counterbalance potential influence of the past). From such a day we perform the predictions using initially 1, then 2 and finally 3 weeks of data, trying to estimate the number of victims for at least others 3-4 weeks. Since the starting time is always around the beginning of April, it means that we try to predict until the mid of May, which is furthermore the day when the Countries changed their policies.

Summing up, the idea is: if we manage to succeed in predicting what happened in May by using the data in April, we can repeat the same strategy concerning the month of June.

For the sake of completeness, the data are anyway pictured until the 1st of June instead of stopping at the mid of May, to point out the effects of changing the lockdown regimes.

Recall that the fitting procedure is done by using a Bayesian algorithm, producing so a probability measure for the parameters. We split its density into clusters, producing multi dimensional histograms, and considered only the 99% of its mass, eliminating so the very unlikely extreme value. We paid attention to the "worst" case scenario, i.e. the one with the highest number of deaths, the "best" one, and the expectation value.

The whole reasoning is repeated two times, using two different error tolerance for the given dataset (roughly speaking, 10% and 100% as explained in details in the section SEC).

We started observing the German data from the $8th$ of April, the French from the $3rd$ and the Italian from the $1st$.

Summing up, we have three predictions for each Country (using 1,2,3 weeks of data), three Countries (Italy, France, Germany), and two error possibilities (10%, 100%), producing therefore a total of 18 plots.

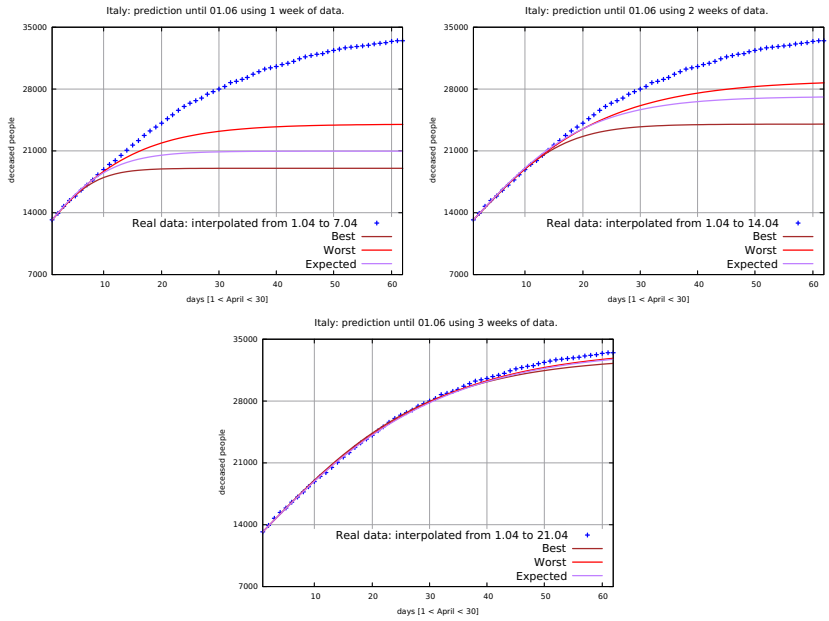

Figure 1: Italy - plot with the 10% of "error". Using 1 or 2 weeks of data seems not enough, but with 3 weeks the results are far better. The real data start slightly exceeding the worst-case-scenario from day 45, i.e. the $15th$ of May. It can probably be read as a sign of weakening the lockdown measures. Notably, the plot is surprisingly "valid" until today, the 1st of June, far beyond our initial expectations.

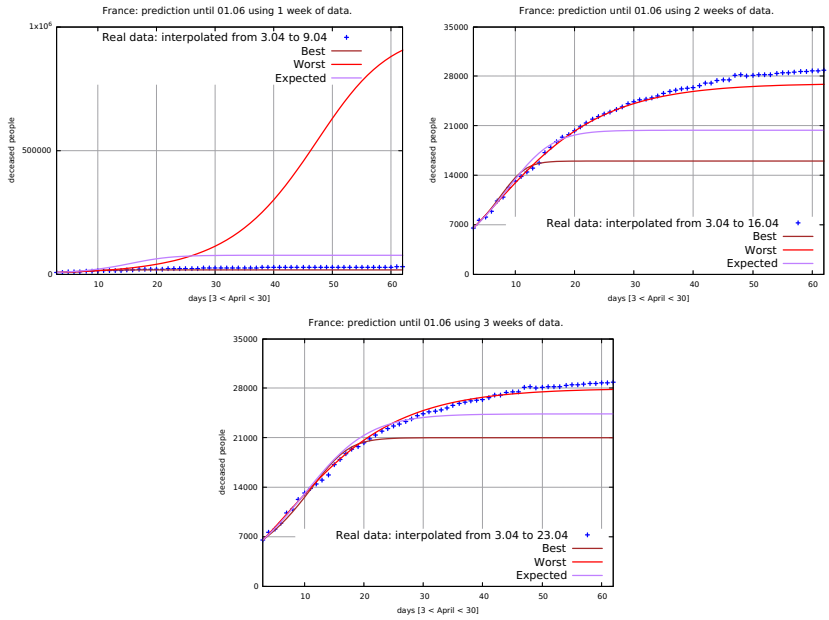

Figure 2: France - plot with 10% of "error". After one week of data, the predictions are terrible, but they improve progressively when using 2 and 3 weeks. Notably, as in the Italian case, the real data slightly exceed the worst-case prediction around day 45, i.e. the 15th of May, and are reasonably "valid" until the 1st of June.

6 Conclusions

We used a Monte Carlo technique combined with a general logistic equation to predict the number of deaths from Coronavirus in Italy, Germany and France. In particular, the algorithm required a parameter connected to the measurement error. When tuned to 10%, 3 weeks of data allowed "reasonably" fine predictions for Italy and France, but not for Germany. By increasing it to 100%, not only the results for Germany are working quite well, but there is a general improvements for Italy and France, too. It can of course suggest that Germany carries higher measurements errors, but we remark how the model is very experimental and based on multiple assumptions. They can just hold "better" for France and Italy rather than for Germany, with their effects influenced by the error parameters.

Using 1 or 2 weeks of data didn't seem to be a good strategy at all, producing both strong under- and over- estimations.

Therefore we speculate that: by using the algorithm explained in this review, with 3 weeks of data and a 100% "error", it is possible to successfully predict the number of victims in Italy, Germany and France for the duration of minimum 1 month.

Our next step will be to gathering of data from the $20th$ of May to the $9th$ of June, computing then new estimations until the mid of July.

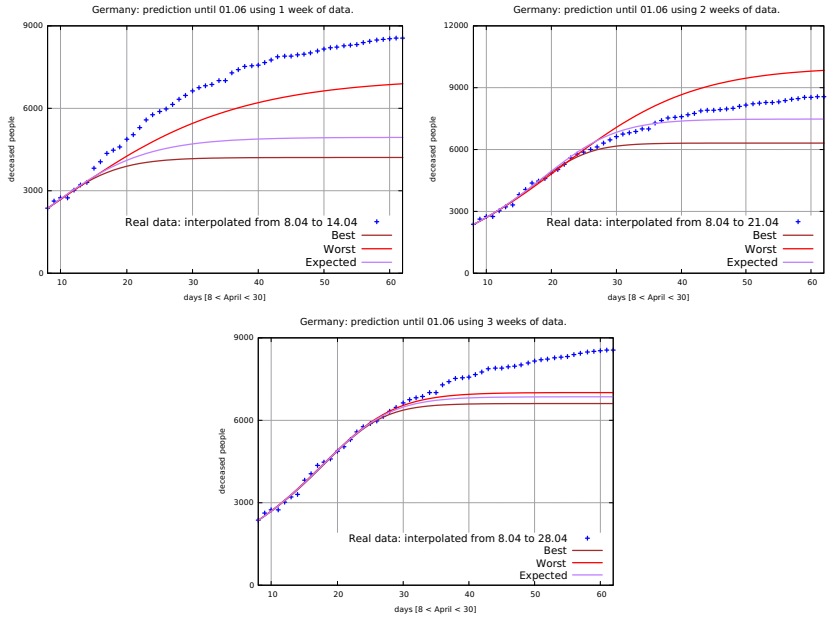

Figure 3: Germany - plot with 10% of "error". Interestingly, the worst results are plotted with 1 *and* 3 weeks of data, being acceptable when using only 2 weeks. This model does not seem to work well with Germany.

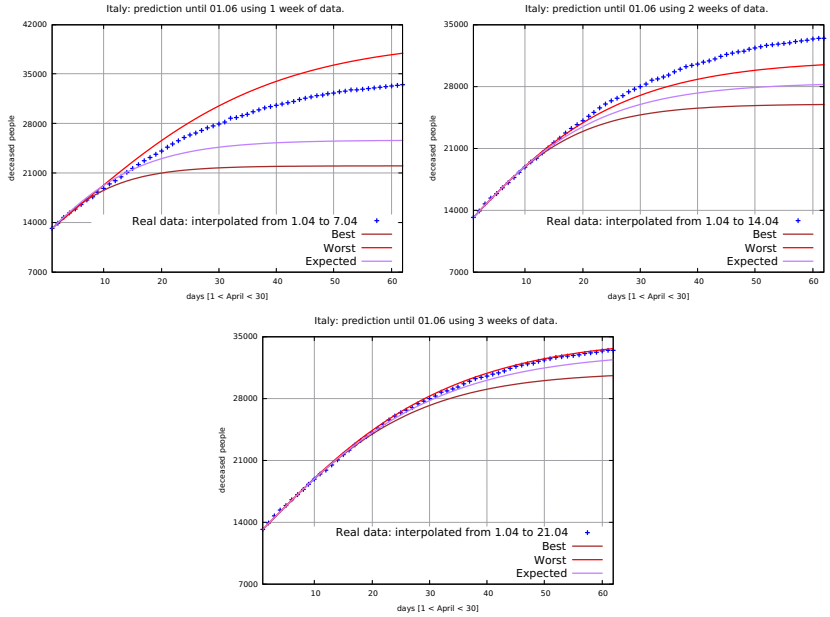

Figure 4: Italy - plot with a tolerance "error" of 100% (to be interpreted as pointed out in section SEC). Notably, the results after 1 and 2 weeks are improved, although the latter seem worse than the former. But the great news come when using 3 weeks of data: not only the plot is positively similar to the previous case, but the real does not seem to exceed the worst-case-scenario. Of course, one should not exaggerate and take into account possible approximation errors due to the very subtle figure. It is extremely surprisingly to see such a clear pattern lasting until the 1of June, although drastic lockdown changes happened meanwhile.

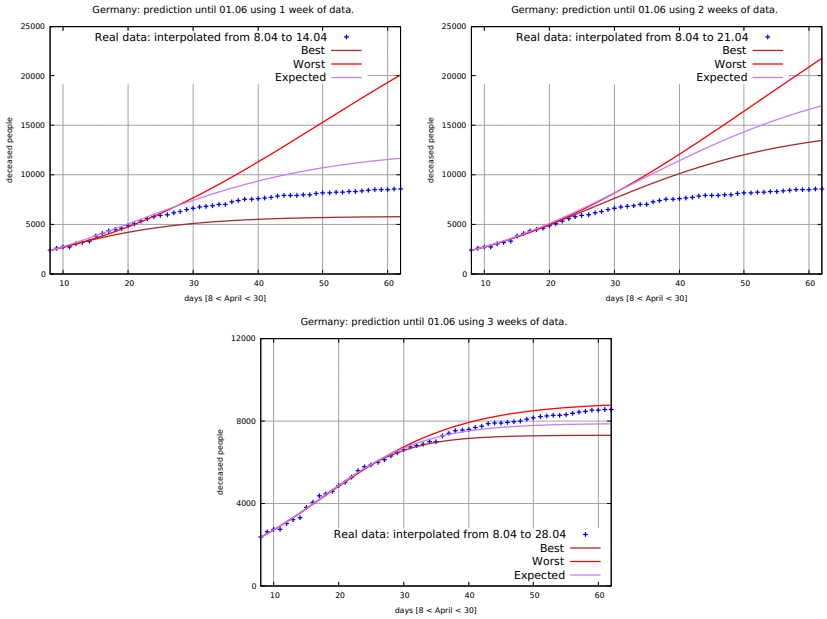

Figure 5: Germany - plot with 100% of tolerance "error". If using 1 or 2 weeks the predictions remain unreliable, but the surprise comes when interpolating 3 weeks. Now the future real data fit very well into the predicted pattern. An increasing trend is registered in the recent period, possibly associated with the weakening of lockdown measures.

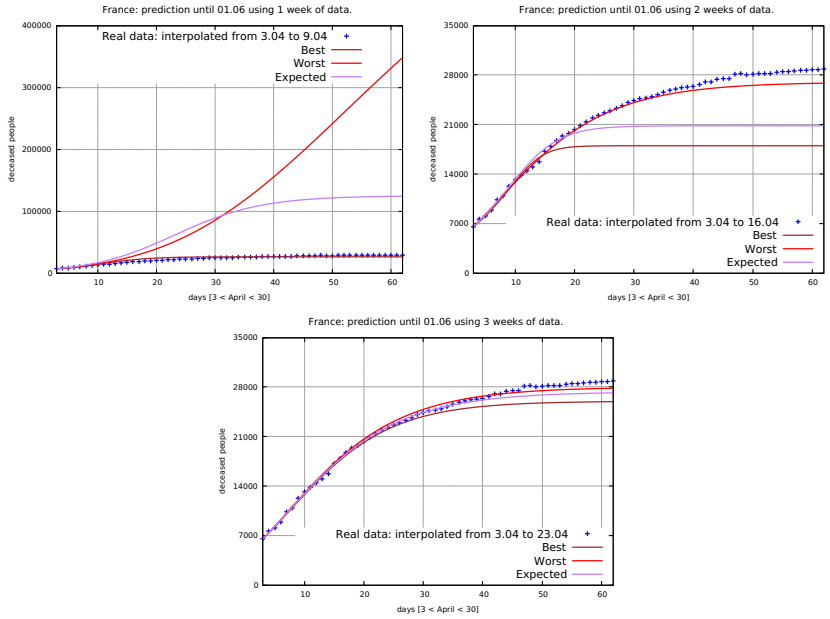

Figure 6: France - plot with "error" tolerance to 100%. Precisely as in the case of Italy, the plots are similar to the 10% case with generic sign of improvements. The best (and working) results are obtained with 3 weeks of data.