Using hospital data for monitoring the dynamics of COVID 19 in France

Marc Lavielle

Using hospital data for monitoring the dynamics of COVID 19 in France

Marc Lavielle
Inria and Ecole Polytechnique

Abstract

The objective of this article is to show how daily hospital data can be used to monitor the evolution of the COVID-19 epidemic in France. A piecewise defined dynamic model allows to fit very well the available hospital admission, death and discharge data. The change-points detected correspond to moments when the dynamics of the epidemic changed abruptly. It is therefore a surveillance tool, not a forecasting tool. In other words, it can be used effectively to warn of a restart of epidemic activity, but it is not designed to assess the impact of a new lockdown or the emergence of a new variant. The model, data and fits are implemented in an interactive web application.

Keywords: COVID 19 data, change-points detection, statistical model, ODE model.

1. Introduction

After some early cases detected in China in late 2019, the COVID-19 outbreak spread very quickly around the world in early 2020 (Velavan and Meyer 2020).

This global pandemic quickly gave rise to numerous studies to try to understand the factors that could explain its spread, such as the effect of climate (Briz-Redón and Serrano-Aroca 2020; Wu et al. 2020) or of human mobility (Kraemer et al. 2020).

Many mathematical models have been developed to describe the dynamics of this pandemic and possibly to predict the future epidemiological situation. Among all these approaches, we can mention the agent-based models used, for instance, for simulating

the spread of COVID-19 among the inhabitants of a city (Silva et al. 2020).

But the most used approaches for the modeling of the dynamics of COVID-19 remain undoubtedly the SIR-type (or SEIR-type) epidemiological models (He et al. 2020).

Such models allow, among other things, to simulate different scenarios (Carcione et al. 2020), in order to predict what would be, for example, the effect on the epidemic of a public health intervention (Di Domenico et al. 2020; López and Rodo 2021; Yang et al. 2020). On the other hand, these compartmental epidemiological models offer the advantage of being able to take into account different subgroups in the population, such as asymptomatic individuals (Chen et al. 2020).

These different models that have been proposed claim to describe "the reality", i.e. how the pandemic evolves over time in the population. In order to get as close as possible to this reality, the models are necessarily complex, with many compartments, transfers between these compartments, and therefore many parameters. The use of these models to simulate the evolution of the epidemic or to evaluate the impact of a sanitary measure requires choosing the values of these parameters. Model calibration allows one to find empirically a set of parameters that provides a good fit between the model calculations and the observed data, but the complexity of the model makes it unidentifiable in practice. Indeed, the data available to fit the model are limited and do not allow to determine univocally the set of parameters. Then, although a set of parameters produces a good fit to the observed data, there is no guarantee that the predictions proposed by the model are reliable.

Our approach here is quite different. We do not claim to develop a model that accurately describes the dynamics of the epidemic, but rather a simple, robust model that fits the data very well. The objective of such a model is not to predict the evolution of the epidemic in the future, to determine the date of the next peak or to define the best strategy to implement to contain the epidemic. We will only try to describe the past dynamics, and predict what should happen in the near future if no change occurs in the dynamics of the epidemic, and above all to detect as soon as possible a change in these dynamics when it exists.

Consequently, the choice of data to be used is fundamental to our approach. The data we have chosen to use for this monitoring are the daily hospitalization and death data published daily by Santé Publique France, the French national public health agency. (Salje et al. 2020; Paireau et al. 2021).

We propose to describe these data by means of a statistical model that allows to combine several effects such as epidemic dynamics, a weekly pattern and irregular variations. The dynamics of hospital admissions (normal therapy and intensive care units) is described by considering an exponential dynamics, but for which the rate function is defined piecewise linearly, which allows to describe very well the different phases of growth and decrease of these admission numbers. Fitting this model to the data then consists in detecting change-points in the admissions data. The fitted model allows to identify the different epidemic waves observed in France since March 2020.

2. Which data to use?

Monitoring the dynamics of the pandemic in real time naturally requires reliable and

regularly updated data. The question is to determine which data can best describe these dynamics and also detect changes in these dynamics as quickly as possible.

A few weeks after the emergence of the virus, an online interactive dashboard was developed and hosted by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, Baltimore, MD, USA, to visualise and track reported cases of COVID-19 in real time (Dong et al. 2020). The data collected and made freely available include the number of confirmed COVID-19 cases, deaths, and recoveries for all affected countries. These data have been widely used to track and model the pandemic, whether through visual exploratory data analysis (Dey et al. 2020), random processes (Benvenuto et al. 2020), or epidemiological models (Lavielle et al. 2021).

French data are displayed Fig. 1. Although a general trend seems to be visible in these graphs, there are several problems with their use. First, the data appear to be extremely noisy and contain a significant number of outliers, both negative values and isolated abnormally high values. A second problem comes from the definition of some of these series. Indeed, the numbers of confirmed cases and recoveries that are reported represent only a fraction of the total infections and recoveries. These fractions are not homogeneous over time. For example, the number of confirmed cases in France during the first wave (between March and May 2020) corresponds to the number of patients whose infection was confirmed in hospital. From the second wave onwards (September to November 2020), confirmed cases include positive tests and are therefore much higher than during the first wave.

One can then imagine using the results of the virological tests represented on Fig. 2 as a marker since they are directly related to the incidence rate of COVID-19 in the population. It should be noted that the definition of the incidence rate, widely used by the authorities as well as the media, is simply the number of positive tests over one week, related to 100,000 inhabitants. This definition obviously does not correspond to the real incidence rate since not all the population is tested (Pullano et al. 2021). Its evolution does not necessarily reflect the evolution of the epidemic in France either, since the number of tests performed daily changes over time. For example, the large increases in positive tests in October 2020 and March 2021 are partly explained by a large increase in the number of tests performed during these periods. The positivity rate (i.e. the proportion of positive tests among the tests performed) seems to be a better indicator, since it takes into account, by definition, variations in the number of tests performed. Although it provides relevant and complementary information, this positivity rate is unfortunately not homogeneous over time because the tested population is not homogeneous over time. We see a spectacular drop in the positivity rate in December 2020. This drop is probably not due to a sudden drop in infections, but rather to a one-time increase in the number of tests by people who are not at risk, but who nevertheless wished to be tested before the end-of-year festivities and family gatherings.

We will finally use hospital data obtained from the SI-VIC database, the national inpatient surveillance system used during the pandemic. Data are sent daily to Santé Publique France, the French national public health agency in charge of making them publicly available:

https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19.

These data are shown Fig. 3. These are the daily numbers of patients *i)* newly admitted to normal therapeutic wards, *ii)* newly admitted to intensive care units, *iii)* deceased in the hospital, *iv)* allowed to leave the hospital (hospital discharges).

There are several advantages to using this data. First of all, these data are regularly consolidated and can therefore be considered reliable. Furthermore, apart from a clearly visible weekly pattern, the data are homogeneous over time: values at different dates are directly comparable. Finally, the dynamics of admissions are directly linked, with a certain time lag, to the dynamics of new infections: an increase in admissions necessarily reflects a prior increase in infections, and the same applies to decreases. We can therefore reasonably expect to detect changes in the dynamics of the epidemic by detecting changes in the dynamics of admissions.

3. The model

3.1. The statistical model

Fig. 3 shows variations over time in the data due to several combined effects: an overall trend (epidemic dynamics), a periodic component (weekly pattern), and irregular variations.

For each of the four series $(z_{\ell j}, 1 \leq \ell \leq 4, 1 \leq j \leq n)$ observed at time $(t_j, 1 \leq j \leq n)$, we propose the following model

$$z_{\ell j} = f_\ell(t_j) + f_\ell^{\alpha_\ell}(t_j)(s_{\ell j} + \varepsilon_{\ell j}) \tag{1}$$

where f_ℓ is the trend for the ℓ-th series, $(s_{\ell j}, 1 \leq j \leq n)$ is a weekly periodic component such that $s_{\ell,j+7} = s_{\ell j}$ for any j and $(\varepsilon_{\ell j})$ is a sequence of residual errors. The multiplicative term $f_\ell^{\alpha_\ell}(t_j)$ allows us to take into account the fact that the amplitude of both the periodic and irregular variations varies with the value of the trend. The exponent α_ℓ allows here to control the link between these amplitudes.

We propose to represent the trends $(f_\ell, 1 \leq \ell \leq 4)$ by means of a dynamic system. It is the construction of this system and its estimation that represents the most delicate part of this modeling work.

3.2. The dynamical model

We consider that the study starts at a time t_0 and we will arbitrarily set $t_0 = 0$. We note $I_{\text{ntw}}(t)$ and $I_{\text{icu}}(t)$, the total numbers of patients admitted, respectively in normal therapy services and in intensive care units, between time t_0 and time t. We also note $D(t)$ and $O(t)$, the numbers of patients who died in hospital and were discharged recovered from hospital between time t_0 and t, respectively. Finally, we note $H(t)$ the number of patients present in the hospital (in normal care or in intensive care) at time t.

The variations of the number of hospitalized patients thus depend on the admissions and discharges according to the following dynamics:

$$\dot{H}(t) = \dot{I}_{\text{ntw}}(t) + \dot{I}_{\text{icu}}(t) - \dot{D}(t) - \dot{O}(t) \tag{2}$$

Our goal is now to build a model for each of these 4 terms.

So let's start with the admissions. But rather than modeling the total admissions $I_{\text{ntw}}(t)$ and $I_{\text{icu}}(t)$, it is their variations that we will model, since it is, by definition, these functions that directly describe the dynamics of admissions over time, i.e., how admissions increase at the beginning of an epidemic wave or decrease at the end of a wave. We propose to use exponential-type dynamics for each of these series, but for which the rate functions can vary over time:

$$\ddot{I}_{\text{ntw}}(t) = k_{\text{ntw}}(t)\,\dot{I}_{\text{ntw}}(t) \quad (3)$$

$$\ddot{I}_{\text{icu}}(t) = k_{\text{icu}}(t)\,\dot{I}_{\text{icu}}(t) \quad (4)$$

A constant and positive (resp. negative) rate function k_{ntw}, or k_{icu}, means that the number of admissions increases (resp. decreases) exponentially fast. The fact of using a rate that can vary with time allows then to transition between different regimes of exponential growth and decay. We will consider that these transitions are linear, using for k_{ntw} and k_{icu} piecewise linear functions. We therefore suppose that there exist K_{ntw} and K_{icu} instants, called change-points, $\tau_{\text{ntw},1}$, $\tau_{\text{ntw},1},\ldots$, $\tau_{\text{ntw},K_{\text{ntw}}}$ and $\tau_{\text{icu},1}$, $\tau_{\text{icu},1},\ldots$, $\tau_{\text{icu},K_{\text{icu}}}$ such that

$$k_{\text{ntw}}(t) = b_{\text{ntw}} + 2\,c_{\text{ntw}}t + 2\sum_{k=1}^{K_{\text{ntw}}} h_{\text{ntw},k}\ \max(t-\tau_{\text{ntw},k}\ ,\ 0)$$

$$k_{\text{icu}}(t) = b_{\text{icu}} + 2\,c_{\text{icu}}t + 2\sum_{k=1}^{K_{\text{icu}}} h_{\text{icu},k}\ \max(t-\tau_{\text{icu},k}\ ,\ 0)$$

Assuming that the rate functions k_{ntw} and k_{icu} are piecewise linear functions allows to compute the solution of the equations (3) and (4) and verify that $\log(\dot{I}_{\text{ntw}})$ and $\log(\dot{I}_{\text{icu}})$ are piecewise quadratic functions:

$$\log(\dot{I}_{\text{ntw}}(t)) = a_{\text{ntw}} + b_{\text{ntw}}t + c_{\text{ntw}}t^2 + \sum_{k=1}^{K_{\text{ntw}}} h_{\text{ntw},k}\ \max(t-\tau_{\text{ntw},k}\ ,\ 0)^2$$

$$\log(\dot{I}_{\text{icu}}(t)) = a_{\text{icu}} + b_{\text{icu}}t + c_{\text{icu}}t^2 + \sum_{k=1}^{K_{\text{icu}}} h_{\text{icu},k}\ \max(t-\tau_{\text{icu},k}\ ,\ 0)^2$$

where $a_{\text{ntw}} = \log(\dot{I}_{\text{ntw}}(t_0))$ and $a_{\text{icu}} = \log(\dot{I}_{\text{icu}}(t_0))$

It is now assumed that the numbers of deaths and discharges between times t and $t+dt$ depend on the number of patients hospitalized at time t:

$$\dot{D}(t) = \gamma_{\text{deaths}}(t)H(t) \quad (5)$$

$$\dot{O}(t) = \gamma_{\text{out}}(t)H(t) \quad (6)$$

The mortality rate γ_{deaths} and the discharge rate γ_{out} are not constant over time. Again, we consider that the logarithms of these functions are piecewise quadratic functions:

$$\log(\gamma_{\text{deaths}}(t)) = a_{\text{deaths}} + b_{\text{deaths}}t + c_{\text{deaths}}t^2 + \sum_{k=1}^{K_{\text{deaths}}} h_{\text{deaths},k}\ \max(t-\tau_{\text{deaths},k}\ ,\ 0)^2$$

$$\log(\gamma_{\text{out}}(t)) = a_{\text{out}} + b_{\text{out}}t + c_{\text{out}}t^2 + \sum_{k=1}^{K_{\text{out}}} h_{\text{out},k}\ \max(t-\tau_{\text{out},k}\ ,\ 0)^2$$

Now that the model is defined, all that remains is to fit it to the data at our disposal.

4. Fitting the model to the French hospital data

4.1. Fitting the dynamical model

The objective is now to estimate the parameters of functions $\dot{I}_{\text{ntw}}$, $\dot{I}_{\text{icu}}$, γ_{deaths} and γ_{out}.

We first remove the weekly pattern and smooth the data using an unweighted 7-day moving average for the four series $(z_{\ell j})$. We then denote the four smoothed series obtained by $(q_{\text{ntw},j})$, $(q_{\text{icu},j})$, (d_j) and (o_j).

On the one hand, the daily series of admissions to the normal therapy wards $(q_{\text{ntw},j})$ and to intensive care units $(q_{\text{icu},j})$ will allow us to estimate the derivatives of the cumulative counts I_{ntw} and I_{icu} using the following model:

$$\begin{aligned} \log(q_{\text{ntw},j}) &= \log(\dot{I}_{\text{ntw}}(t_j)) + e_{\text{ntw},j} \\ \log(q_{\text{icu},j}) &= \log(\dot{I}_{\text{icu}}(t_j)) + e_{\text{icu},j} \end{aligned}$$

The mortality rate γ_{deaths} and the discharge rate γ_{out} can naturally be estimated using the observed daily rates (d_j/h_j) and (o_j/h_j) where h_j is the number of hospitalized patients (all units combined) at time t_j. We use for these series the model

$$\begin{aligned} \log(d_j/h_j) &= \log(\gamma_{\text{deaths}}(t_j)) + e_{\text{deaths},j} \\ \log(o_j/h_j) &= \log(\gamma_{\text{out}}(t_j)) + e_{\text{out},j} \end{aligned}$$

Let $y_{1,j} = \log(q_{\text{ntw},j})$, $y_{2,j} = \log(q_{\text{icu},j})$, $y_{3,j} = \log(d_j/h_j)$ and $y_{4,j} = \log(o_j/h_j)$. For $\ell = 1, \ldots, 4$, we then have the following model:

$$y_{\ell j} = a_\ell + b_\ell\, t + c_\ell\, t^2 + \sum_{k=1}^{K_\ell} h_{\ell,k}\ \max(t - \tau_{\ell,k}\ ,\ 0)^2 + e_{\ell j} \tag{7}$$

For each of the four series, the problem then becomes a problem of change-points detection:

- For a given number of change points K_ℓ,
  - Find the locations of the K_ℓ change points $\tau_{\ell,1}, \ldots, \tau_{\ell,K_\ell-1}$,
  - Estimate the parameters of the model $a_\ell, b_\ell, c_\ell, h_{\ell,1}, h_{\ell,2}, \ldots, h_{\ell,K_\ell}$,
- Select the "best" model, i.e. select the number of change points K_ℓ.

For each of the series, we propose here to use a penalized least squares criterion to estimate all the parameters of the model and the number of change-points.

In order not to make the notations unnecessarily heavy, we can remove the subscript ℓ to describe the estimation procedure used, which is identical for the four series.

For a given number of change-points K, for a set of parameters $\theta_K = (a, b, c, h_1, \ldots, h_K)$ and a sequence of change-point instants $T_K = (\tau_1, \ldots, \tau_K)$, we write down

$$f(t; \theta_K, T_K) = a + b\,t + c\,t^2 + \sum_{k=1}^{K} h_k \ \max(t - \tau_k \ , \ 0)^2$$

We then estimate θ_K, T_K and K by minimizing

$$U(\theta_K, T_K, K) = \sum_{j=1}^{n} \left(y_j - f(t_j; \theta_K, T_K)\right)^2 + \lambda\, K$$

A high value of the penalty parameter λ favors configurations with few change-points while a lower value of λ allows a higher number of changes.

The minimization of the penalized criterion U can be decomposed in several steps. Indeed, for a given series of change-points instants T_K, the minimization of U with respect to θ_K is immediate since it is simply a matter of computing the least squares estimate in a linear model. For a given number of changes K and by setting

$$\hat{\theta}(T_K) = \arg\min_{\theta_K} \left\{ \sum_{j=1}^{n} \left(y_j - f(t_j; \theta_K, T_K)\right)^2 \right\} \tag{8}$$

The estimator of T_K is then defined as

$$\hat{T}_K = \arg\min_{T_K} \left\{ \sum_{j=1}^{n} \Big(y_j - f(t_j; \hat{\theta}(T_K), T_K)\Big)^2 \right\} \tag{9}$$

The number of changes K is therefore chosen as

$$\hat{K} = \arg\min_{K} \left\{ \sum_{j=1}^{n} \Big(y_j - f(t_j; \hat{\theta}(\hat{T}_K), \hat{T}_K)\Big)^2 + \lambda\, K \right\} \tag{10}$$

The tricky part is the estimation of the change-points as defined in (9). Indeed, we cannot use a dynamic programming algorithm because, due to the continuity constraints imposed on f and its derivative, the criterion to be minimized cannot be decomposed as a sum of independent criteria on each segment.

As the data series are updated daily, the proposed algorithm is a sequential procedure that requires little computation. Indeed, the configuration at day $j+1$ is obtained from local modifications of the configuration obtained at day j. Let us imagine that $T_K^{(j)}$ is the optimal configuration obtained on day j. We then compute $T_K^{(j+1)}$ and $T_{K+1}^{(j+1)}$ the best configurations with, respectively, K and $K+1$ breaks when a new observation y_{j+1} is available at time t_{j+1}. These configurations are obtained by iterative optimization, the position of a single break being modified at each iteration.

The best of these two configurations is then selected using the penalized criterion (10).

The value of the penalty parameter λ is here manually adjusted in such a way as to obtain as a result a segmentation that visually "looks like" the segmentation that one would build oneself by looking at the data. In other words, we make sure that all the

changes that we consider significant are well detected, while the smaller, more irregular variations are not associated with the signal but considered as random fluctuations. The results proposed below were all obtained by choosing $\lambda = 10^{-4}$.

Fig. 5 represents the fits obtained for the series $(q_{\text{ntw},j})$ and $(q_{\text{icu},j})$. We have also represented on this figure the relative variations $(r_{\text{ntw},j})$ and $(r_{\text{icu},j})$ where

$$r_{\text{ntw},j} = \frac{q_{\text{ntw},j} - q_{\text{ntw},j-1}}{q_{\text{ntw},j-1}} \quad ; \quad r_{\text{icu},j} = \frac{q_{\text{icu},j} - q_{\text{icu},j-1}}{q_{\text{icu},j-1}}$$

By construction, while the series $(q_{\text{ntw},j})$ and $(q_{\text{icu},j})$ fluctuate around the functions $\dot{I}_{\text{ntw}}$ and $\dot{I}_{\text{icu}}$, the series $(r_{\text{ntw},j})$ and $(r_{\text{icu},j})$ fluctuate around the rate functions $k_{\text{ntw}} = \ddot{I}_{\text{ntw}}/\dot{I}_{\text{ntw}}$ and $k_{\text{icu}} = \ddot{I}_{\text{icu}}/\dot{I}_{\text{icu}}$ also shown on the two bottom graphs of the figure.

We can see very well on these graphs that it is reasonable to consider for k_{ntw} and k_{icu} piecewise linear functions. It is finally the variations of these rate functions that give a synthetic picture of the dynamics of the epidemic in France.

Once the γ_{deaths} and γ_{out} functions have been estimated, equations (2), (5) and (6) allow us to obtain the D and O functions. The mortality and discharge rates are plotted Fig. 6 as well as the daily numbers of deaths and discharges. These graphs confirm that mortality rates vary over time and that these variations must be taken into account in order to correctly model deaths and discharges.

The sudden drop observed during the second half of March 2020 corresponds to the implementation of the first lockdown which was very strict. The decrease in admissions was of course not instantaneous since it took several days for the rate functions to become negative. This was followed by a period of more than two months during which admissions continued to decline, until about mid-June while the lockdown had ended in mid-May.

Although admissions remained at a very low level until early September, the rate functions clearly show a change in dynamics from mid-June onwards: the increase in the rate functions reflects a gradual slowdown in the decline of admissions, before reaching a minimum in early July and slowly rising again. The rapid increase in admissions is then very visible in early September, but especially in early October. Both the authorities and the media placed the start of the second wave at this time, when it was most visible, but the change in dynamics was much earlier!

A sudden decrease in the rate functions around October 18 shows that the increase in the daily number of hospitalizations started to weaken around this date. It is interesting to note that the measures to contain this second wave were put in place after the change in dynamics had occurred (generalized curfew on October 24 and then new confinement on October 30). This slowdown continued until the first days of November, when the number of new hospitalizations started to decrease and the variations became negative.

Between mid-November and mid-March, there was a succession of periods of fairly slow growth and decline in the rate functions, which are difficult to associate with particular events. For several months, France was in a relatively stable regime, since the measures in place prevented a new explosion of contaminations - and therefore of hospitalizations - but did not allow a return to a normal situation either.

The decrease in the rate functions observed from the end of March onwards led to negative values of these functions from mid-April onwards, and thus a continuous decrease

in hospitalizations until today (i.e. end of June). It is reasonable to believe that the increase in vaccination coverage from 10% to 50% (for at least one dose) during this period explains in part this significant decrease in epidemic activity.

It is important to remember that the role of the model proposed here is not to predict how the epidemic will evolve in France in the coming weeks or months. It has not been developed for this purpose since it only uses hospitalization data and these data do not contain any information related to possible changes in behavior, health measures, vaccination policy, etc. The model simply predicts what would happen in the absence of change: the different estimated functions can be used and evaluated at times after the last observation time.

By using a linear Gaussian model for the model (7), it is also possible to construct a confidence interval for the estimated regression function and a prediction interval for future observations, in the absence of new changes. But again, it is not a question of evaluating the performance of the model by checking that future observations are indeed within the constructed prediction interval, but rather by checking that the prediction interval does not contain the observed data after a change in dynamics. Examples of such intervals are shown in Fig. 7. Data were considered to be available until March 4, 2021 on the left and until March 24 on the right. Confidence intervals were then calculated for the functions $\dot{I}_{\text{ntw}}$ and k_{ntw} as well as prediction intervals for the series $(q_{\text{ntw},j})$ and $(r_{\text{ntw},j})$ for the next 14 days, i.e. after the last observation. The intervals are represented with the data that were actually observed during these forecast periods. We can see in the left figure that the prediction intervals do contain the two observed series. Indeed, no change will be detected during this forecast period (March 5 - March 19) and the model provides predictions consistent with the observations. On the contrary, the figures on the right show an inconsistency between the predictions and the observations: while the model assumes that the rate function continues to increase linearly, a change in dynamics occurred around March 25, 2021 and the rate function started to decrease from that date. It is clear from this example that this change was detectable only a few days after it occurred.

4.2. Fitting the statistical model

Let us now return to the original series of daily admissions to conventional therapy $(z_{\text{ntw},j})$ and intensive care unit $(z_{\text{icu},j})$.

The regression model (1) suggests that these series decompose into a trend, a periodic component related to the day of the week, and a series of residual errors. Now that estimators $\hat{f}_1$ and $\hat{f}_2$ of the trends $f_1 = \dot{I}_{\text{ntw}}$ and $f_2 = \dot{I}_{\text{icu}}$ have been obtained, we will be able to use the model (1) to estimate the other components of the model.

Let us assume for simplicity of notation that $n = 7h$. Then, for $\ell = 1, 2$ and for a given value of α_ℓ, the periodic series $(s_{\ell j})$ and the series of residuals $(e_{\ell j})$ can easily be estimated:

$$
\begin{aligned}
w_{\ell,j} &= \frac{z_{\ell,j} - \hat{f}_\ell(t_j))}{\hat{f}_\ell^{\alpha_\ell}(t_j)} \qquad ; \qquad j = 1, 2, \ldots, n \\
\hat{s}_{\ell,m} &= \frac{1}{h} \sum_{k=0}^{h-1} w_{\ell,m+7k} \qquad ; \qquad m = 1, 2, \ldots, 7
\end{aligned}
$$

$$\hat{e}_{\ell,j} = \frac{z_{\ell,j} - \hat{f}_\ell(t_j)) - \hat{f}_\ell^{\alpha_\ell}(t_j)\hat{s}_{\ell,j}}{\hat{f}_\ell^{\alpha_\ell}(t_j)} \quad ; \quad j = 1, 2, \ldots, n$$

The exponent α_ℓ is chosen so as to obtain a residual error series (ε_j) that is the least correlated possible, or, more precisely, such that the empirical correlation between the series $(\hat{e}_{\ell,j})$ and $(\hat{e}_{\ell,j+7})$ is as close to 0 as possible. This criterion leads us to choose $\alpha_1 = \alpha_2 = 0.8$.

Fig. 8 shows the estimated periodic component and the estimated residual errors for the two series. Not surprisingly, a drop in admissions is evident on week-ends, mainly on Sundays. The examination of the residuals allows us to highlight the effect of certain public holidays on admissions, which are difficult to identify in the original data: we thus see "abnormally" low values on Christmas and New Year's Day, Easter Monday (April 5), Ascension Thursday (May 13) and Whit Monday (May 24). These low values are generally compensated on the following days by "abnormally" high values.

Computational Details

The results in this paper were obtained using R 4.0.3. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

The model and various data related to COVID-19 are implemented in the interactive Shiny app http://shiny.webpopix.org/covidix/app3en/.

References

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., and Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in brief*, 29:105340.

Briz-Redón, Á. and Serrano-Aroca, Á. (2020). The effect of climate on the spread of the COVID-19 pandemic: A review of findings, and statistical and modelling techniques. *Progress in Physical Geography: Earth and Environment*, 44(5):591–604.

Carcione, J. M., Santos, J. E., Bagaini, C., and Ba, J. (2020). A simulation of a COVID-19 epidemic based on a deterministic SEIR model. *Frontiers in public health*, 8:230.

Chen, Y.-C., Lu, P.-E., Chang, C.-S., and Liu, T.-H. (2020). A time-dependent SIR model for COVID-19 with undetectable infected persons. *IEEE Transactions on Network Science and Engineering*, 7(4):3279–3294.

Dey, S. K., Rahman, M. M., Siddiqi, U. R., and Howlader, A. (2020). Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. *Journal of medical virology*, 92(6):632–638.

Di Domenico, L., Pullano, G., Sabbatini, C. E., Boëlle, P.-Y., and Colizza, V. (2020). Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. *BMC medicine*, 18(1):1–13.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.

He, S., Peng, Y., and Sun, K. (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics*, 101(3):1667–1680.

Kraemer, M. U., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., et al. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490):493–497.

Lavielle, M., Faron, M., Lefevre, J. H., and Zeitoun, J.-D. (2021). Predicting the propagation of COVID-19 at an international scale: extension of an SIR model. *BMJ open*, 11(5):e041472.

López, L. and Rodo, X. (2021). A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: simulating control scenarios and multi-scale epidemics. *Results in Physics*, 21:103746.

Paireau, J., Andronico, A., Hozé, N., Layan, M., Crepey, P., Roumagnac, A., Lavielle, M., Boëlle, P.-Y., and Cauchemez, S. (2021). An ensemble model based on early predictors to forecast COVID-19 healthcare demand in France, https://hal-pasteur.archives-ouvertes.fr/pasteur-03149082. working paper or preprint.

Pullano, G., Di Domenico, L., Sabbatini, C. E., Valdano, E., Turbelin, C., Debin, M., Guerrisi, C., Kengne-Kuetche, C., Souty, C., Hanslik, T., et al. (2021). Underdetection of cases of COVID-19 in France threatens epidemic control. *Nature*, 590(7844):134–139.

Salje, H., Kiem, C. T., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., Andronico, A., Hozé, N., Richet, J., Dubost, C.-L., et al. (2020). Estimating the burden of SARS-CoV-2 in France. *Science*, 369(6500):208–211.

Silva, P. C., Batista, P. V., Lima, H. S., Alves, M. A., Guimarães, F. G., and Silva, R. C. (2020). COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals*, 139:110088.

Velavan, T. P. and Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical medicine & international health*, 25(3):278.

Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M., and Liu, M. (2020). Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. *Science of the Total Environment*, 729:139051.

Yang, Z., Zeng, Z., Wang, K., Wong, S.-S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., et al. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of thoracic disease*, 12(3):165.

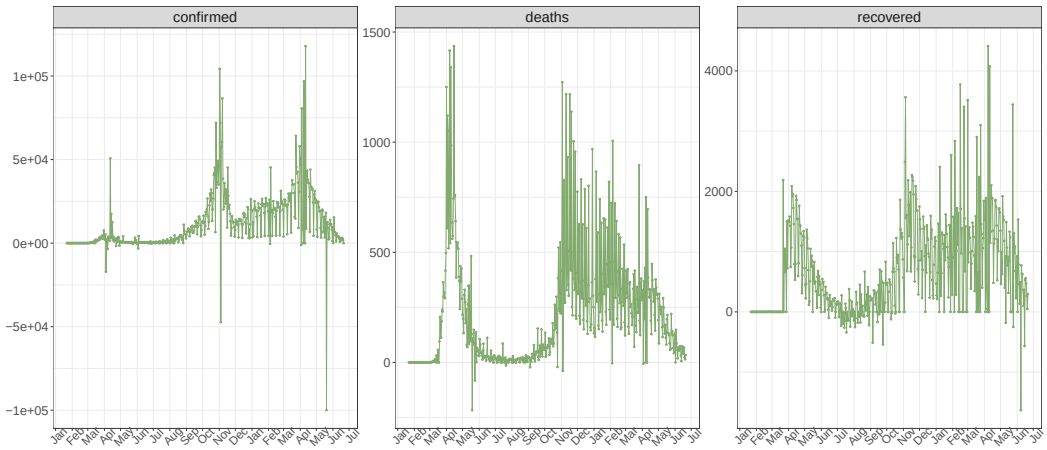

Figure 1: French COVID-19 data collected by the Center for Systems Science and Engineering at Johns Hopkins University: daily numbers of confirmed cases, deaths and recoveries.

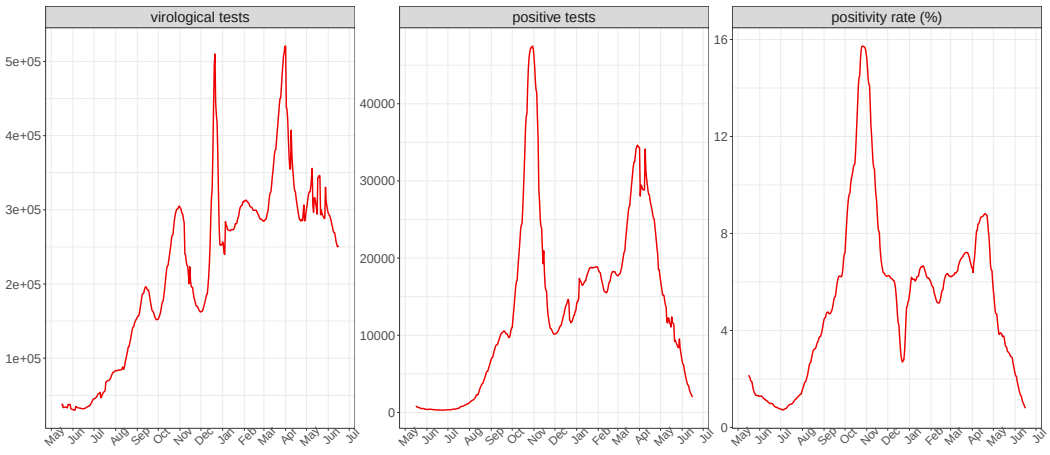

Figure 2: Data on COVID-19 virological test results in France, produced by Santé Publique France: daily numbers of tests, positive tests and positivity rate.

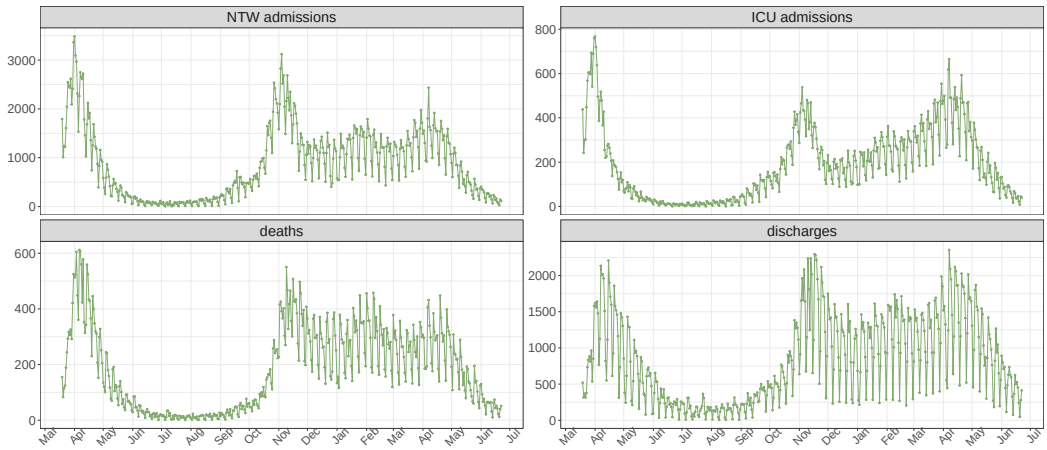

Figure 3: French hospital data for the COVID-19 produced by Santé Publique France: daily numbers of patients newly admitted to normal therapeutic wards, admitted to intensive care units, deceased in the hospital, allowed to leave the hospital.

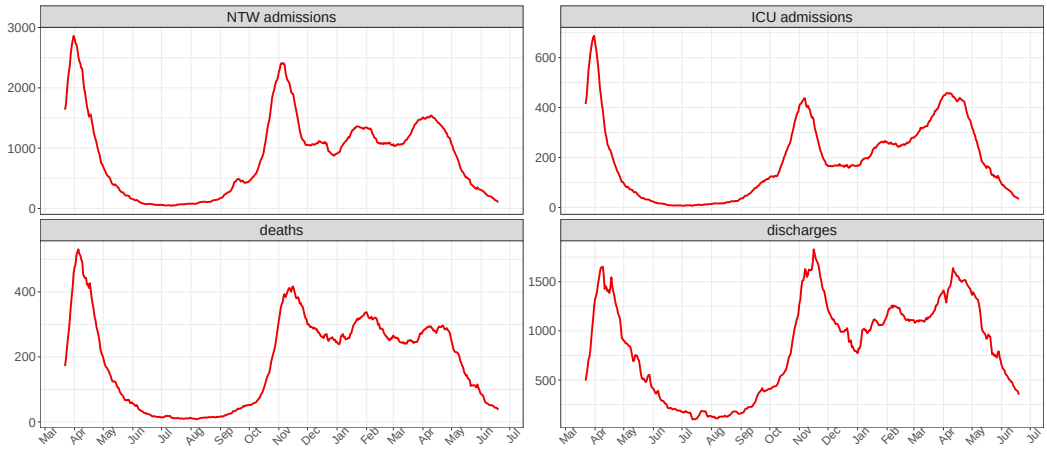

Figure 4: Unweighted 7-day moving averages for the four series displayed Fig. 3.

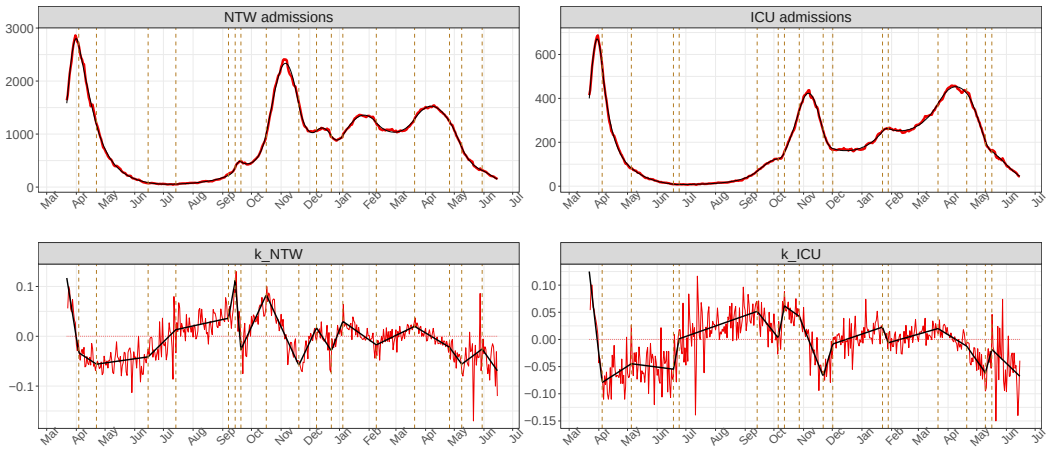

Figure 5: Top: smoothed series of admissions ; bottom: relative variations of these series. The grey lines are the fits obtained and the vertical dashed lines are the estimated change-points.

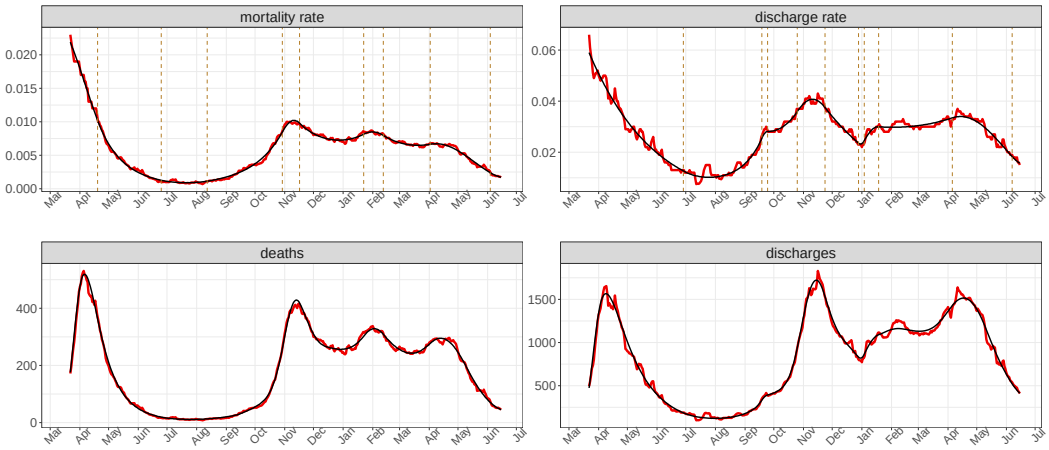

Figure 6: Top: smoothed series of mortality and discharge rates ; bottom: smoothed series of daily deaths and discharges. The grey lines are the fits obtained and the vertical dashed lines are the estimated change-points.

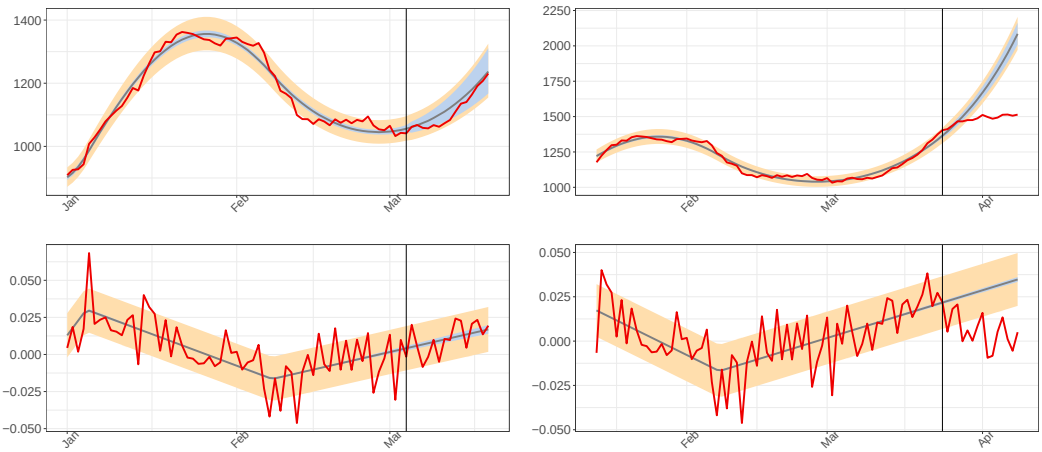

Figure 7: Confidence intervals for the estimated regression function in blue, prediction intervals for future observations in yellow and observed series in red. Top: admissions in normal therapy wards ; bottom: relative variations of these series. The intervals were constructed using data available until March 4, 2021 on the left and until March 24 on the right.

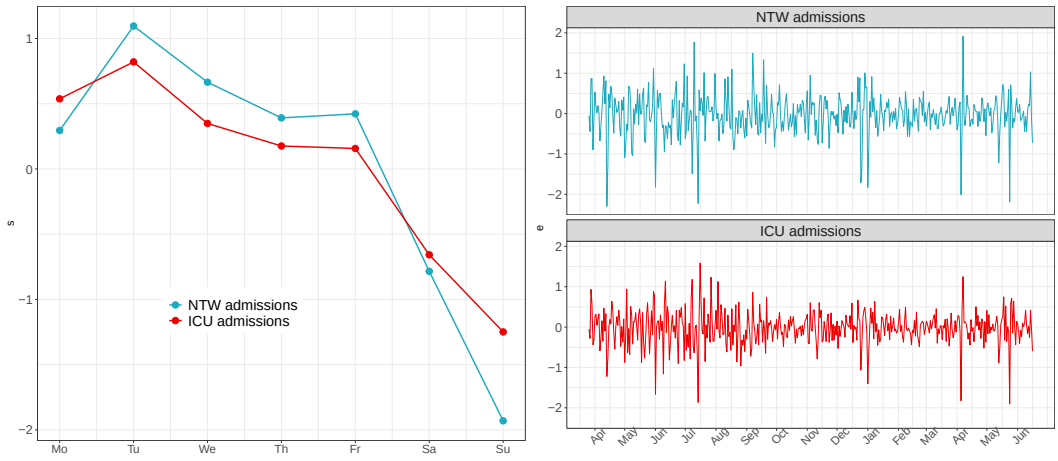

Figure 8: Components of the statistical model built for the admissions series. Left: periodic weekly pattern ; right: residual errors.

Affiliation:

Marc Lavielle
Inria, Saclay, France
and
CMAP, Ecole Polytechnique, CNRS, Institut Polytechnique de Paris
Route de Saclay
91128 Palaiseau Cedex, France
E-mail: Marc.Lavielle@inria.fr
URL: http://www.cmap.polytechnique.fr/~lavielle/