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 aim of this article is to show how daily hospital data can be used to track the evolution of the COVID-19 epidemic in France. A piecewise defined dynamic model allows a very good fit of the available data on hospital admissions, deaths and discharges. The change-points detected correspond to moments when the dynamics of the epidemic changed abruptly. Although the proposed model is relatively simple, it can serve several purposes: It is an analytical tool to better understand what has happened so far by relating observed changes to changes in health policy or the evolution of the virus. It is also a surveillance tool that can be used effectively to warn of a resurgence of epidemic activity, and finally a short-term forecasting tool if conditions remain unchanged. The model, data and fits are implemented in an interactive web application.

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

1. Introduction

After some early cases were discovered 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 trying to understand the factors that could explain its spread, such as the effects of climate (Briz-Redón and Serrano-Aroca 2020; Wu et al. 2020) or human mobility (Kraemer et al. 2020).

Many mathematical models have been developed to describe the dynamics of this pan-

demic and possibly predict the future epidemiological situation. Among all these approaches, we can mention the agent-based models used, for example, to simulate the spread of COVID-19 among the inhabitants of a city (Silva et al. 2020). But the most commonly used approaches for modeling the dynamics of COVID-19 undoubtedly remain 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) to predict how, for example, a public health intervention would affect the epidemic (Di Domenico et al. 2020; López and Rodo 2021; Yang et al. 2020). On the other hand, these compartmental epidemiological models have the advantage of being able to account for different subgroups in the population, such as asymptomatic individuals (Chen et al. 2020).

These various models that have been proposed claim to describe "reality", i.e., how the pandemic evolves over time in the population. 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 the choice of the values of these parameters. Model calibration makes it possible to find empirically a set of parameters that provides a good fit between the model calculations and the observed data. However, due to the complexity of the model, this set may not be identifiable in practice. Indeed, the data available to fit the model are limited and do not allow the parameter set to be uniquely identified (Hamelin et al. 2020). Nevertheless, there are methods to estimate the parameters of the model in this context, for example by introducing prior information about the values of these parameters or directly by fixing some of them to values from the literature.

Our approach here is quite different. We do not claim to develop a model that accurately mimics the dynamics of the epidemic, but rather a simple, robust model that fits the data very well. The goal 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 contain the epidemic. We will simply try to describe the dynamics of the past and predict what should happen in the near future if the dynamics of the epidemic do not change, and most importantly, to detect a change in those dynamics as soon as possible, if it does occur.

Consequently, the choice of data to use is fundamental to our approach. The data we use for this surveillance are the daily hospital admissions and deaths reported by Santé Publique France, the French national public health agency (Salje et al. 2020; Paireau et al. 2021).

We propose to describe these data using a statistical model that allows us to combine different effects such as epidemic dynamics, a weekly pattern and irregular fluctuations. The dynamics of hospital admissions (normal therapy and intensive care units) are described by assuming exponential dynamics, but for which the rate function is defined in a piecewise linear way, which allows a very good description of the different phases of growth and decline of these admission numbers. Fitting this model to the data then consists in detecting change-points in the admission data. The fitted model makes it possible 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 obviously requires reliable and regularly updated data. The question arises as to which data can best describe these dynamics and also detect changes as quickly as possible.

A few weeks after the virus emerged, an interactive online dashboard was developed and hosted by the Centre 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 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). The French data are shown in Fig. 1.

Although a general trend can be seen in these charts, there are several problems with their use. First, because these data are very noisy, and second, because the definition of certain data, such as the number of confirmed cases, is not homogeneous over time.

Thus, one can imagine using the results of the virological tests shown in Fig. 2 as markers, since they are directly related to the incidence rate of COVID-19 in the population. It should be noted that the definition of incidence rate commonly used by both authorities and the media is simply the number of positive tests in a week, per 100,000 population. This definition, of course, does not reflect the actual incidence rate, since not the entire population is tested (Pullano et al. 2021). Its evolution also does not necessarily reflect the evolution of the epidemic in France, as the number of tests performed each day changes over time. Thus, the sharp increase in positive tests in October 2020 and March 2021 is partly explained by a sharp increase in the number of tests performed during these periods. The positivity rate (i.e., the proportion of positive tests to tests performed) appears to be a better indicator because, by definition, it takes into account fluctuations in the number of tests performed. Unfortunately, although it provides relevant and complementary information, this positivity rate is 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 decline is likely not due to a sudden drop in infections, but rather a one-time increase in the number of people tested who, while not at risk, still wanted to get tested before year-end celebrations and family gatherings.

Finally, we will use hospital data from the SI-VIC database, the national inpatient surveillance system used during the pandemic. The data is transmitted daily to Santé Publique France, the French health authority responsible for publishing the data:
https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19.

These data are shown in Fig. 3. They are the daily number of patients *i)* newly admitted to normal therapeutic wards (NTW), *ii)* patients newly admitted to intensive care units (ICU), *iii)* patients who died in hospital, *iv)* patients who were allowed to leave hospital (hospital discharges).

There are several advantages to using these data. First, these data are regularly consolidated and can therefore be considered reliable. Moreover, apart from a clearly

discernible weekly pattern, the data are homogeneous over time: Values at different points in time are directly comparable. Finally, the dynamics of admissions are directly related to the dynamics of new infections, with a certain time lag: An increase in admissions necessarily reflects a previous increase in infections, and the same is true for 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 the temporal variations in the data due to several combined effects: a general trend (epidemic dynamics), a periodic component (weekly pattern), and irregular fluctuations.

Let $z_{1,j}$ and $z_{2,j}$, be the numbers of admissions to normal therapy ward and intensive care units, respectively, on day j. Let $z_{3,j}$ and $z_{4,j}$ be the numbers of deaths and discharges, respectively, on day j. 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 account for the fact that the amplitude of both periodic and irregular variations varies with the value of the trend. The exponent α_ℓ here allows us to control the link between these amplitudes.

We propose to represent the trends $(f_\ell, 1 \leq \ell \leq 4)$ using a dynamical system. The construction of this system and its estimation are 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. In the following, $\dot{f}$ and $\ddot{f}$ denote the first and second derivatives of f.

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

So let us start with the admissions. However, we will not model the total number of admissions $I_{\text{ntw}}(t)$ and $I_{\text{icu}}(t)$, but rather their fluctuations, since these functions by definition 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 where 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 that a rate is used that can vary with time then allows the transition between different regimes of exponential growth and decay. We will assume 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 number of deaths and the number of discharges between times t and $t+dt$ both 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, we need 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 algorithm

The objective is now to estimate the parameters of the 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 $(z_{\ell j})$ as shown Fig. 4. 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.

To avoid making the notation unnecessarily cumbersome, we may omit the index ℓ to describe the estimation procedure used, which is identical for all 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

$$f(t; \theta_K, T_K) = a + b\,t + c\,t^2 + \sum_{k=1}^{K} h_k \ \max(t - \tau_k \ , \ 0)^2. \tag{8}$$

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 into several steps. For a given series of change-points instants T_K, minimizing U with respect to θ_K is immediate, since it simply involves computing the least squares estimate in a linear model. For a given number of changes K and by setting

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

The estimator of T_K is then defined as

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

The number of changes K is therefore chosen as

$$\hat{K} = \arg\min_{K} \left\{ \sum_{j=1}^{n} \left(y_j - f(t_j; \hat{\theta}(\hat{T}_K), \hat{T}_K)\right)^2 + \lambda\, K \right\} \tag{11}$$

The tricky part is estimating the change-points as defined in (10). In fact, we cannot use a dynamic programming algorithm because the criterion to be minimized cannot be decomposed as a sum of independent criteria for each segment due to the continuity constraints on f and its derivative.

Since the data series are updated daily, the proposed algorithm is a sequential procedure that requires little computation. In fact, the configuration on day $j+1$ is obtained from local changes in the configuration obtained on day j. Suppose that $T_K^{(j)}$ is the optimal configuration obtained on day j. We then compute $T_K^{(j+1)}$ and $T_{K+1}^{(j+1)}$ as the best configurations with K and $K+1$ breaks, respectively, when there is a new observation at time t_{j+1}. These configurations are obtained by iterative optimization, changing the position of a single change point at each iteration. The best of these two configurations is then selected on the basis of the penalized criterion (11).

The value of the penalty parameter λ here is manually adjusted so that the result is a segmentation that visually "looks like" the segmentation one would create oneself

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

The results

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 $k_{\text{icu}} = \ddot{I}_{\text{icu}}/\dot{I}_{\text{icu}}$ which are also shown in the two lower graphs of the figure. From these graphs we can see very clearly that it is reasonable to consider piecewise linear functions for k_{ntw} and k_{icu}. It is ultimately the variations in these rate functions that provide 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. 7 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 in the second half of March 2020 corresponds to the implementation of the first, very strict lockdown. The drop in admissions was not immediate, of course, as 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 rise in the rate functions reflects a gradual slowdown in the decline in admissions before reaching a minimum in early July and slowly rising again. The rapid rise in admissions is then clearly visible in early September, but especially in early October. Both the authorities and the media have placed the start of the second wave at this time, when it was most visible, but the change in dynamics was much earlier!

A marked decline in rate functions around October 18 shows that the increase in the daily number of hospitalizations began to weaken around that date. It is interesting to note that the measures to curb this second wave were not put into effect until after the change in dynamics had occurred (general 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 hospital admissions began to fall and the relative variations became negative.

Between mid-November and mid-March there were a series of periods of relatively slow growth and decline in rate functions, which are difficult to associate with specific events. For several months, France was in a relatively stable state, as the measures

taken prevented a new explosion of contaminations - and therefore hospitalisations - but also did not allow a return to a normal situation.

The decrease in rate functions observed from the end of March led to negative values of these functions from mid-April to the end of June, resulting in a continuous decline in hospitalizations. It is likely that the increase in vaccination coverage from 10% to 50% (for at least one dose) during this period, as shown in Fig. 6, partly explains this marked decline in epidemic activity..

The appearance of the delta variant at the end of June led to a very significant change in the dynamics, as the rate function rapidly rebounded by the end of July. Again, we can only hypothesise to explain this new reversal of dynamics, such as the second round of vaccination observed in July. These observed links between vaccination and epidemic dynamics are clearly non-linear: although collective immunity no longer seems possible, critical vaccination coverage seems to control the epidemic. The resurgence of the epidemic observed at the end of September in France, as in other European countries, would then indicate a decline in individual immunity among those formerly vaccinated.

Confidence and prediction intervals

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, the point is not to evaluate the performance of the model by checking that the future observations are indeed within the constructed prediction interval, but by checking that the prediction interval does not include the observed data after a change in the dynamics. Examples of such intervals are shown in Fig. 8. Data were considered available through March 4, 2021 on the left and through March 24 on the right. Confidence intervals were then calculated for the functions$\dot{I}_{\text{ntw}}$ and k_{ntw} and 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 plotted with the data actually observed during these forecast periods. In the left figure, we can see that the prediction intervals contain the two observed series. Indeed, no change will be detected during this forecast period (March 5 - March 19) and the model provides predictions that are consistent with the observations. In contrast, the figures on the right show 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 began to decline from that date. This example shows that this change was detectable only a few days after it occurred.

Using a Poisson model

We use a penalized least squares criterion here to estimate the parameters of the model, which means that we implicitly compute the maximum likelihood estimator in a Gaussian model. This assumption is quite justified for the death and discharge rates, which are continuous variables. It may seem questionable for admissions, which are count data and for which one might prefer a Poisson model, for example.

The piecewise polynomial function f defined in (8) is now used to define the intensity

of the Poisson process:

$$e^{y_j} \sim \text{Poisson}(f(t_j; \theta_K, T_K))$$

In this context, θ_K, T_K and K are estimated by minimizing

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

The use of a linear Gaussian model offers several practical and algorithmic advantages, both for estimating the model parameters and for constructing confidence and prediction intervals. With this new objective function to minimize, things get a lot more complicated.

However, we have implemented a nonlinear optimization algorithm to compare the solutions obtained with the Poisson model and the Gaussian model with reduced portions of the data. One such comparison example is shown in Figure 10: It clearly shows that the results are indistinguishable. This finding confirms the idea of using a penalized least squares criterion to detect the instants of change and estimate the model parameters.

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 we have obtained the estimators $\hat{f}_1$ and $\hat{f}_2$ of the trends $f_1 = \dot{I}_{\text{ntw}}$ and $f_2 = \dot{I}_{\text{icu}}$, we can use the model (1) to estimate the other components of the model.

To simplify the notation, let us assume 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:

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

The exponent α_ℓ is chosen to produce a residual error series (ε_j) that is as uncorrelated as 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. 9 shows the estimated periodic component and the estimated residual errors for the two series. Not surprisingly, a decrease in admissions is observed on weekends, especially on Sundays. Examining the residuals allows us to highlight the impact of certain holidays on admissions that are difficult to see in the original data: For example,

we see "unusually" low values on Christmas and New Year's Day, Easter Monday (April 5), Ascension Day (May 13), and Whit Monday (May 24). These low values are usually compensated by "unusually" high values on the following days.

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/.

Some concluding remarks

First of all, it is important to remind that the role of the model proposed here is not to predict how the epidemic in France will develop in the coming weeks or months. It was not developed for this purpose, as it only uses data on hospitalizations and these data do not include information on possible behavioral changes, health interventions, vaccination policies, etc. Our main goal, then, is to propose a model that describes what has happened, not predicts what will happen. This study shows that relative variation in hospital admissions describes the dynamics of the epidemic very well, identifying both the moments when the epidemic starts again and those when it declines. Such an a posteriori analysis is very important to better assess future developments of the epidemic and thus make the right decisions as soon as possible.

However, to refine this analysis, various sources of heterogeneity should be considered. For example, it is well known that the risk of severe illness with COVID-19 increases with age, with older adults at highest risk. Analysis by age group would then be particularly interesting to determine whether the changes in dynamics observed over time may vary with age. Consideration of vaccination status would also be an extremely informative and useful extension for deciding which vaccination policy to pursue. Unfortunately, to our knowledge, the data to perform such strata analyses are not available. However, hospital data on admissions, deaths and discharges by region are available. Thus, the above Shiny app makes it possible to perform a separate analysis for each of the 12 French regions to highlight any regional variability.

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.

Hamelin, F., Iggidr, A., Rapaport, A., and Sallet, G. (2020). Observability, identifiability and epidemiology–a survey. *arXiv preprint arXiv:2011.12202*.

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.

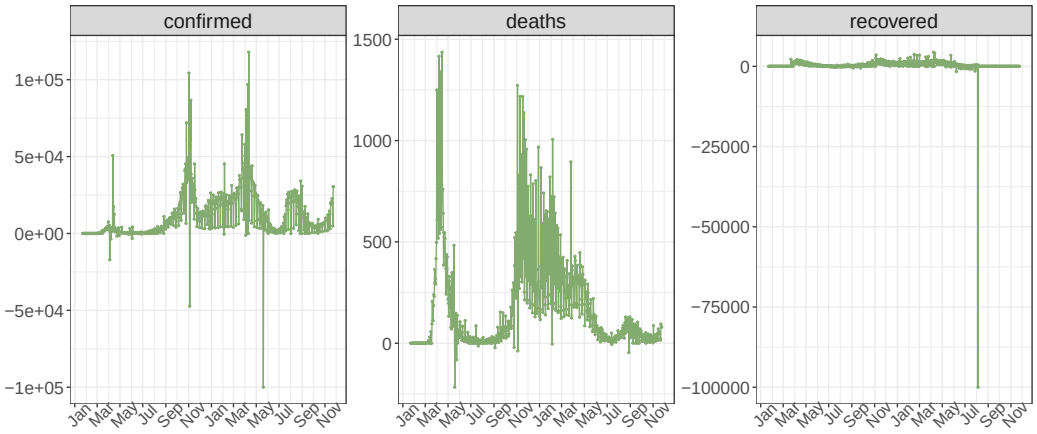

Figure 1: French COVID-19 data collected by the Center for Systems Science and Engineering at Johns Hopkins University: daily numbers of confirmed cases, deaths and recoveries.

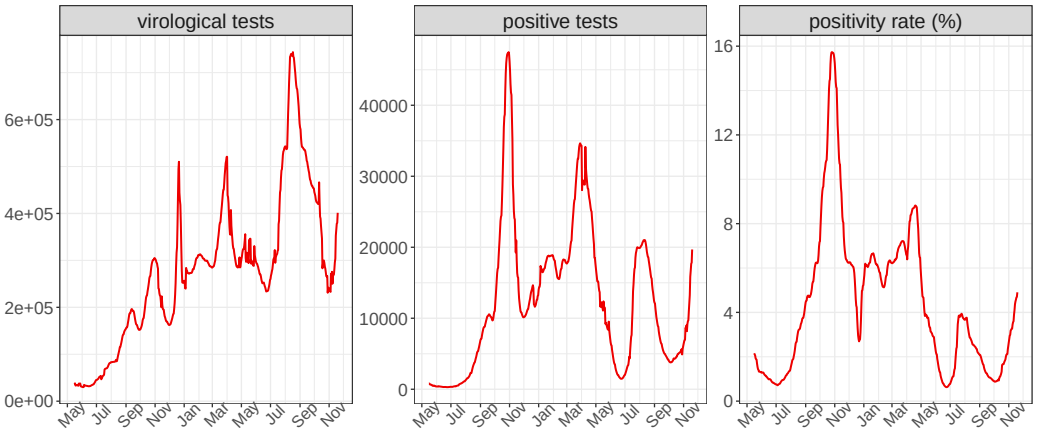

Figure 2: Data on COVID-19 virological test results in France, produced by Santé Publique France: daily numbers of tests, positive tests and positivity rate.

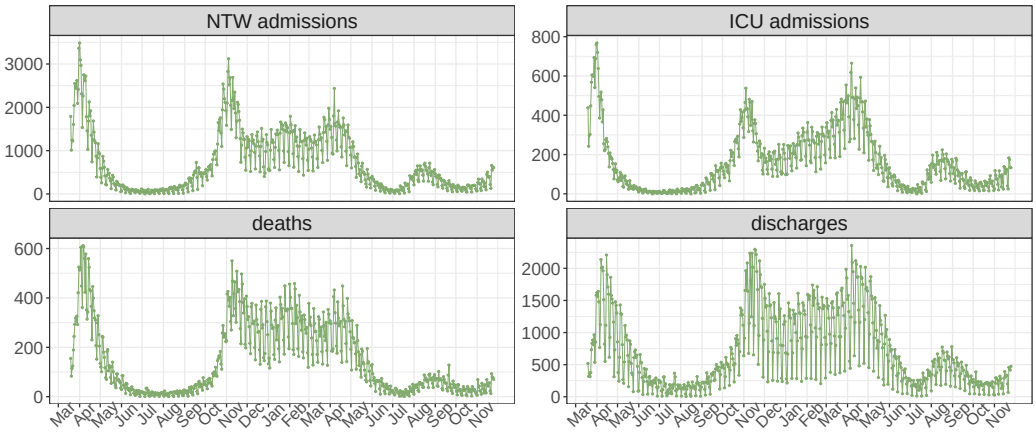

Figure 3: French hospital data for the COVID-19 produced by Santé Publique France: daily numbers of patients newly admitted to normal therapeutic wards, admitted to intensive care units, deceased in the hospital, allowed to leave the hospital.

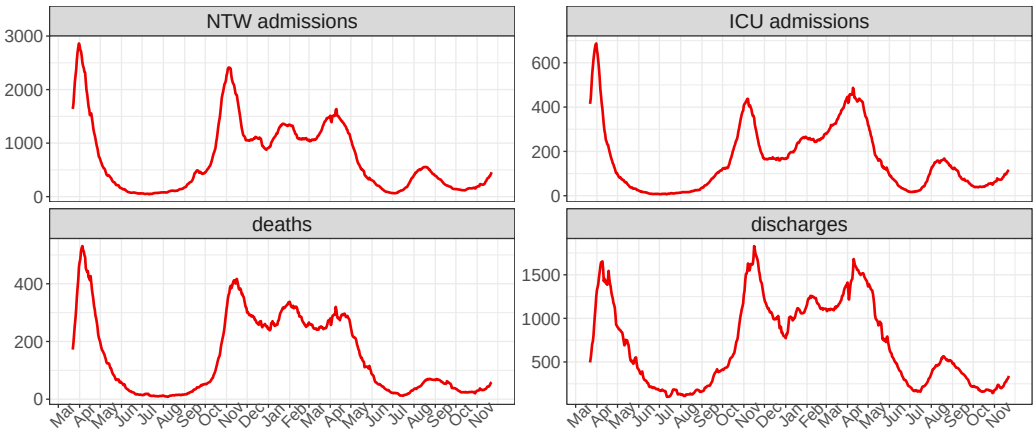

Figure 4: Unweighted 7-day moving averages for the four series displayed Fig. 3.

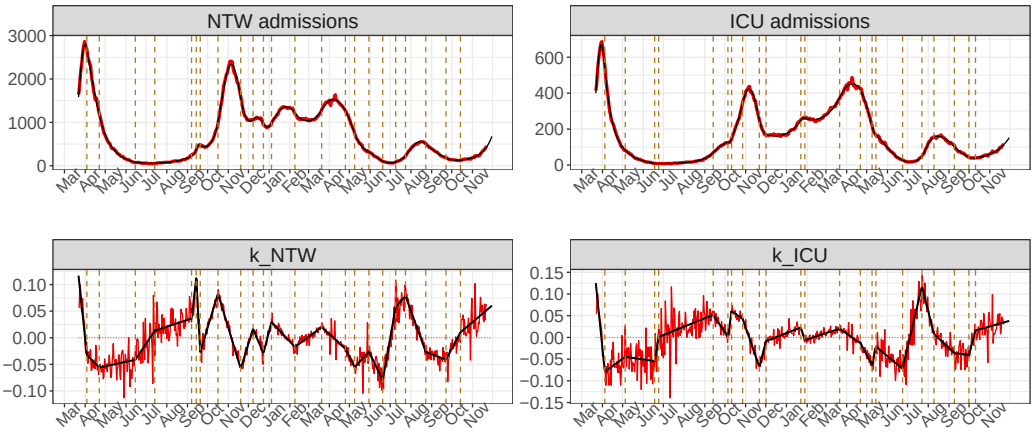

Figure 5: Top: smoothed series of admissions ; bottom: relative variations of these series. The grey lines are the fits obtained and the vertical dashed lines are the estimated change-points.

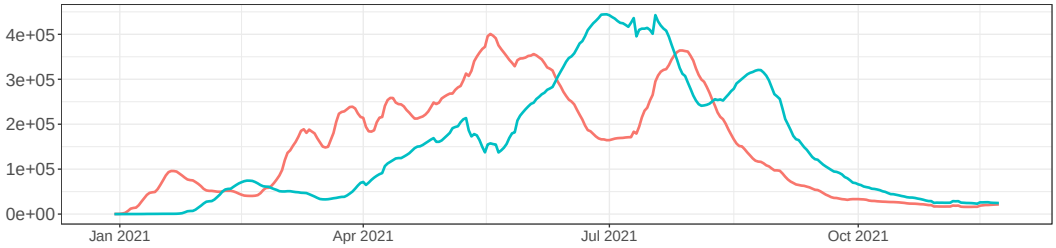

Figure 6: Daily COVID-19 vaccinations. The number of people who received at least one dose is in red and those who received two doses is in blue.

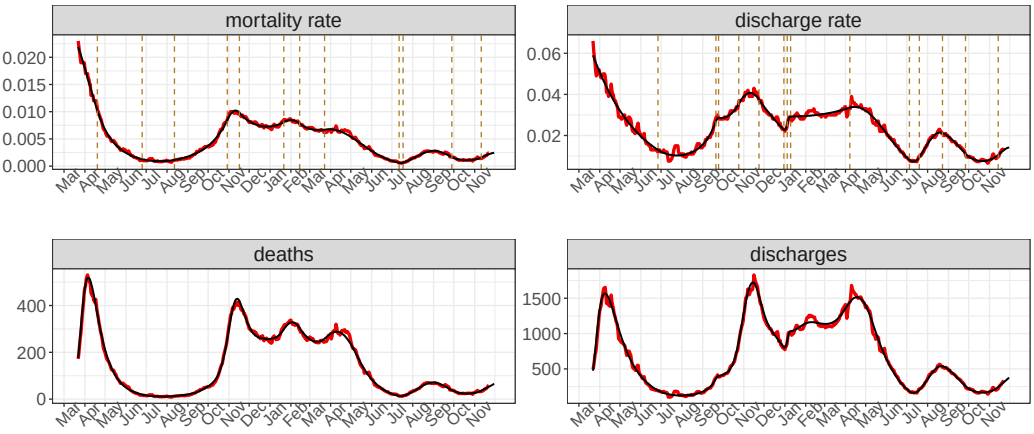

Figure 7: Top: smoothed series of mortality and discharge rates ; bottom: smoothed series of daily deaths and discharges. The grey lines are the fits obtained and the vertical dashed lines are the estimated change-points.

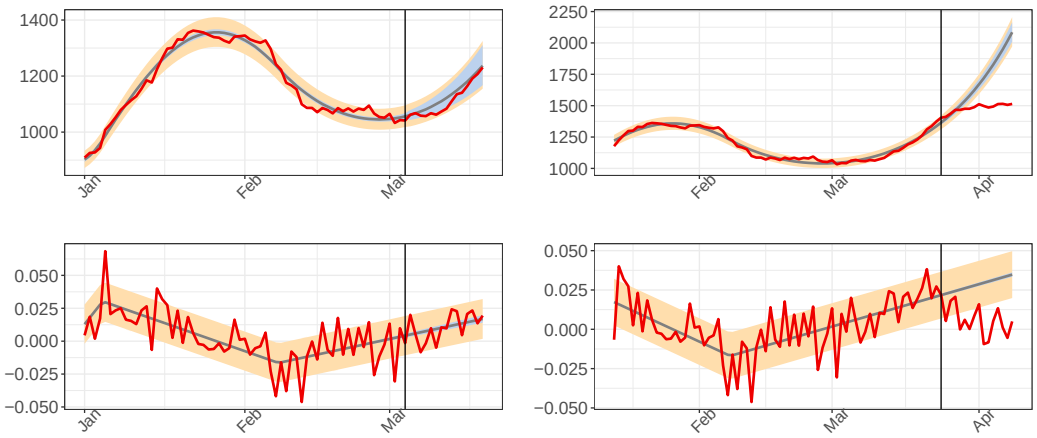

Figure 8: Confidence intervals for the estimated regression function in blue, prediction intervals for future observations in yellow and observed series in red. Top: admissions in normal therapy wards ; bottom: relative variations of these series. The intervals were constructed using data available until March 4, 2021 on the left and until March 24 on the right.

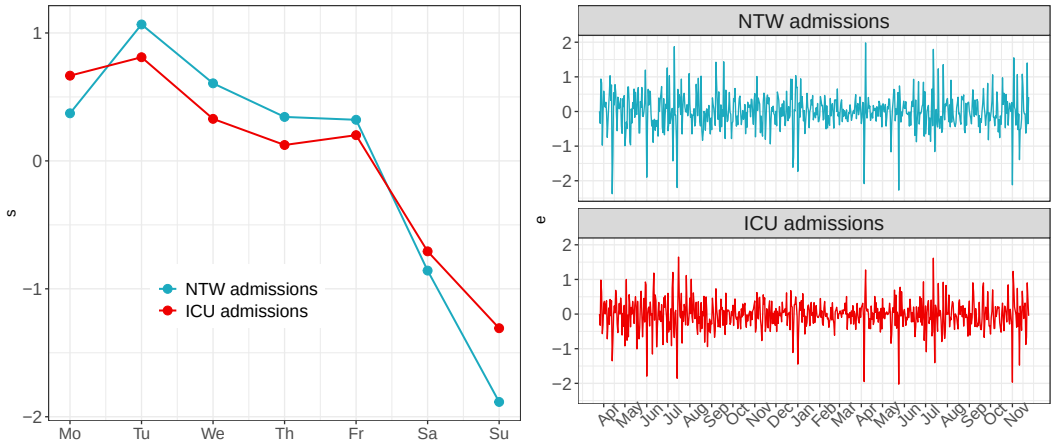

Figure 9: Components of the statistical model built for the admissions series. Left: periodic weekly pattern ; right: residual errors.

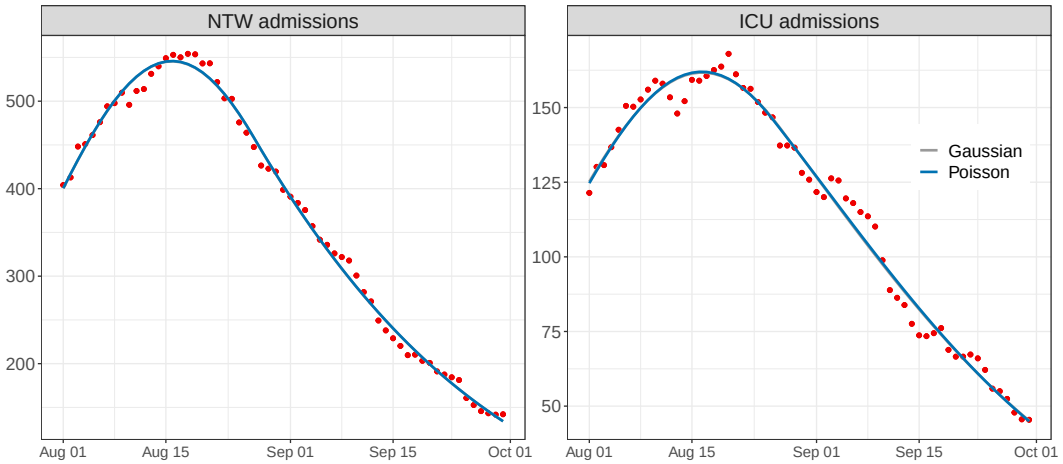

Figure 10: NTW and ICU admissions between 2021-08-01 and 2021-09-30 with the (indistinguishable) fits obtained using a Gaussian model and a Poisson model.

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/