A parsimonious model for spatial transmission and heterogeneity in the COVID-19 propagation

Lionel Roques [a,*], Olivier Bonnefon [a], Virgile Baudrot [a], Samuel Soubeyrand [a] and Henri Berestycki [b]

[a] INRAE, BioSP, 84914, Avignon, France

[b] EHESS, CNRS, CAMS, France

[c] Senior Visiting fellow, HKUST Jockey Club Institute for Advanced Study, Hong Kong University of Science and Technology

[*] Contact : lionel.roques@inrae.fr

July 18, 2020

Abstract

Raw data on the cumulative number of deaths at a country level generally indicate a spatially variable distribution of the incidence of COVID-19 disease. An important issue is to determine whether this spatial pattern is a consequence of environmental heterogeneities, such as the climatic conditions, during the course of the outbreak. Another fundamental issue is to understand the spatial spreading of COVID-19. To address these questions, we consider four candidate epidemiological models with varying complexity in terms of initial conditions, contact rates and non-local transmissions, and we fit them to French mortality data with a mixed probabilistic-ODE approach. Using standard statistical criteria, we select the model with non-local transmission corresponding to a diffusion on the graph of counties that depends on the geographic proximity, with time-dependent contact rate and spatially constant parameters. This original spatially parsimonious model suggests that in a geographically middle size centralized country such as France, once the epidemic is established, the effect of global processes such as restriction policies, sanitary measures and social distancing overwhelms the effect of local factors. Additionally, this modeling approach reveals the latent epidemiological dynamics including the local level of immunity, and allows us to evaluate the role of non-local interactions on the future spread of the disease. In view of its theoretical and numerical simplicity and its ability to accurately track the COVID-19 epidemic curves, the framework we develop here, in particular the non-local model and the associated estimation procedure, is of general interest in studying spatial dynamics of epidemics.

Keywords. COVID-19 | Spatial diffusion | Mechanistic-statistical model | Non-local transmission | Immunity rate

1 Introduction

In France, the first cases of COVID-19 epidemic have been reported on 24 January 2020, although it appeared latter that some cases were already present in December 2019 [1]. Since then, the first

important clusters were observed in February in the Grand Est region and the Paris region. A few months later, at the beginning of June, the spatial pattern of the disease spread seems to have kept track of these first introductions. As this spatial pattern may also be correlated with covariates such as climate [2] (see also SI 1), a fundamental question is to assess whether this pattern is the consequence of a heterogeneous distribution of some covariates or if it can be explained by the heterogeneity of the initial introduction points. In the latter case, we want to know if the epidemic dynamics simply reflects this initial heterogeneity and can be modeled without taking into account any spatial heterogeneity in the local conditions.

SIR epidemiological models and their extensions have been proposed to study the spread of the COVID-19 epidemic at the country or state scale (e.g., [3] in France and [4] in three US states), and at the regional scale ([5, 6] in China, [7] in France and [8] in Italy). In all cases, a different set of parameters has been estimated for each considered region/province. One of the main goals of our study is to check whether the local mortality data at a thin spatial scale can still be well explained by a single set of parameters, at the country scale, but spatially varying initial conditions. In this study we consider SIR models with time-dependent contact rate, to track changes over time in the dynamic reproductive number as in the branching process considered in [4]. Using standard statistical criteria, we compare four types of models, whose parameters are either global at the country scale or spatially heterogeneous, and with non-local transmission or not. We work with mortality data only, as these data appear to be more reliable and less dependent on local testing strategies than confirmed cases, in settings where cause of death is accurately determined [9]. Additionally, it was already shown that for this type of data, SIR models outperform other classes of models (an SEIR models and a branching process) in the three US states considered in [4]. The approach we develop here is in line with the general principle of using parsimonious models eloquently emphasized in [4].

Another important issue that such models may help to address is the quantification of the relative effects of restrictions on inter-regional travels, *vs* reductions in the probability of infection per contact at the local scale. France went into lockdown on March 17, 2020, which was found to be very effective in reducing the spread of the disease. It divided the effective reproduction number (number $\mathcal{R}_t$ of secondary cases generated by an infectious individual) by a factor 5 to 7 at the country scale by May 11 [10, 7]. This is to be compared with the estimate of the basic reproduction number $\mathcal{R}_0$ carried out in France at the early onset of the epidemic, before the country went into lockdown (with values $\mathcal{R}_0 = 2.96$ in [7] and $\mathcal{R}_0 = 3.2$ in [3]). Contact-surveys data [11] for Wuhan and Shanghai, China, have found comparable estimates of the reduction factor. The national lockdown in France induced important restrictions on movement, with e.g. a mandatory home confinement except for essential journeys, leading to a reduction of the number of contacts. In parallel, generalized mask wearing and use of hydroalcoholic gels reduced the probability of infection per contact.

After the lockdown, restrictions policies were generally based on raw data rather than modeling. An efficient regulation at the local scale would need to know the current number of infectious and the level of immunity at the scale of counties. The French territory is divided into administrative units called 'départements', analogous to counties. We use 'counties' in the sequel for départements. These quantities cannot be observed directly in the absence of large scale testing campaigns or spatial random sampling. In particular, there is a large number of unreported cases. Previous studies developed a mixed mechanistic-probabilistic framework to estimate these quantities at the country scale. These involved estimating the relative probability of getting tested for an infected individual *vs* a healthy individual, leading to a factor x8 between the number of confirmed cases and

the actual number of cases before the lockdown [3], and x20 at the end of the lockdown period [10]. This type of framework – often referred to as 'mechanistic-statistical modeling' – aims at connecting the solution of continuous state models such as differential equations with complex data, such as noisy discrete data, and identifying latent processes such as the epidemiological process under consideration here. Initially introduced for physical models and data [12], it is becoming standard in ecology [13].

Using this framework the objectives of our study are (i) to assess whether the spatial pattern observed in France is due do some local covariates or is simply the consequence of the heterogeneity in the initial conditions together with global processes at the country scale; (ii) to evaluate the role of non-local interactions on the spread of the disease; (iii) to propose a tool for real-time monitoring the main components of the disease in France, with a particular focus on the local level of immunity.

Materials and Methods

Data

Mainland France (excluding Corsica island) is made of 94 counties called 'départements'. The daily number of hospital deaths – excluding nursing homes – at the county scale are available from Santé Publique France since 18 March 2020 (and available as Supplementary Material). The daily number of observed deaths (still excluding nursing homes) in county k during day t is denoted by $\hat{\mu}_{k,t}$.

We denote by $[t_i, t_f]$ the observation period and by n_d the number of considered counties. To avoid a too large number of counties with 0 deaths at initial time, the observation period ranges from $t_i =$ March 30 to $t_f =$ June 11, corresponding to $n_t = 74$ days of observation. All the counties of mainland France (excluding Corsica) are taken into account, but the Ile-de-France region, which is made of 8 counties with a small area is considered as a single geographic entity. This leads to $n_d = 87$.

Mechanistic-statistical framework

The mechanistic-statistical framework is a combination of a mechanistic model that describes the epidemiological process, a probabilistic observation model and a statistical inference procedure. We begin with the description of four mechanistic models, whose main characteristics are given in Table 1.

Mechanistic models

Model $\mathcal{M}_0$: SIR model for the whole country. The first model is the standard mean field SIRD model that was used in [3, 10]:

$$\begin{cases} S'(t) = -\dfrac{\alpha(t)}{N}\, S\, I, \\ I'(t) = \dfrac{\alpha(t)}{N}\, S\, I - (\beta + \gamma)\, I, \\ R'(t) = \beta\, I, \\ D'(t) = \gamma\, I, \end{cases} \tag{1}$$

	Heterog. initial data	Heterog. contact rate	Intercounty transmission	Nb. parameters
$\mathcal{M}_0$	no	no	no	n_t
$\mathcal{M}_1$	yes	no	no	n_t
$\mathcal{M}_2$	yes	yes	no	$n_d \times n_t$
$\mathcal{M}_3$	yes	no	yes	$n_t + 2$

Table 1: Main characteristics of the four models. The quantity $n_t = 74$ corresponds to the number of days of the observation period and $n_d = 87$ corresponds to the number of administrative units.

with S the susceptible population, I the infectious population, R the recovered population, D the number of deaths due to the epidemic and N the total population, in the whole country. For simplicity, we assume that N is constant, equal to the current population in France, thereby neglecting the effect of the small variations of the population on the coefficient $\alpha(t)/N$. The parameter $\alpha(t)$ is the contact rate (to be estimated) and $1/\beta$ is the mean time until an infectious becomes recovered. The results in [14] show that the median period of viral shedding is 20 days, but the infectiousness tends to decay before the end of this period: the results in [15] indicate that infectiousness starts from 2.5 days before symptom onset and declines within 7 days of illness onset. Based on these observations we assume here that $1/\beta = 10$ days. The parameter γ corresponds to the death rate of the infectious. It was estimated independently in [3, 10] and [7], leading to a value $\gamma = 5 \cdot 10^{-4}$. This value only takes into account the deaths at hospital, and is therefore consistent with the data that we used here.

Model $\mathcal{M}_1$: SIR model at the county scale with globally constant contact rate and no spatial transmission. The model $\mathcal{M}_0$ is applied at the scale of each county k, leading to compartments S_k, I_k, R_k, D_k that satisfy an equation of the form (1), with N replaced by N_k, the total population in the county k. In this approach, the contact rate $\alpha(t)$ is assumed to be the same in all of the counties.

Model $\mathcal{M}_2$: SIR model at the county scale with spatially heterogeneous contact rate and no spatial transmission. With this approach, the model $\mathcal{M}_1$ is extended by assuming that the contact rate $\alpha_k(t)$ depends on the considered county.

Model $\mathcal{M}_3$: County scale model with globally constant contact rate and spatial transmission. The model $\mathcal{M}_1$ is extended to take into account disease transmission events between the counties:

$$\begin{cases} S_k'(t) = -\dfrac{\rho(t)}{N_k}\, S_k \displaystyle\sum_{j=1}^{n_d} w_{j,k}\, I_j, \\ I_k'(t) = \dfrac{\rho(t)}{N_k}\, S_k \displaystyle\sum_{j=1}^{n_d} w_{j,k}\, I_j - (\beta+\gamma)\, I_k, \\ R_k'(t) = \beta\, I_k, \\ D_k'(t) = \gamma\, I_k. \end{cases} \tag{2}$$

The weights $w_{j,k}$ describe the dependence of the contagion rates with respect to the distance between counties. We assume a power law decay with the distance:

$$w_{j,k} = \frac{1}{1 + (\mathrm{dist}(j,k)/d_0)^{\delta}}, \tag{3}$$

with $\mathrm{dist}(j,k)$ the geographic distance (in km) between the centroids of counties j and k, $d_0 > 0$ a proximity scale, and $\delta > 0$. Thus, the model involves two new global parameters, d_0 and δ, compared to model $\mathcal{M}_1$. This model extends the Kermack-McKendrick SIR model to take into account non-local spatial interactions. It was introduced by Kendall [16] in continuous variables. The model we adopt here is inspired from the study of [17] in a different context where the same types of weights have been used. We thus take into account diffusion on the weighted graph of counties in France. This amounts to considering that individuals in a given county are infected by individuals form other counties with a probability that decreases with distance as a power law, in addition to contagious individuals from their own county. This dependence of social spatial interactions with respect to the distance is supported, notably, by [18] that analyzed the short-time dispersal of bank notes in the US. We also refer to [19] for a thorough discussion on the various applications of power law dispersal kernels since they were introduced by Pareto [20].

With this non-local contagion model, in contradistinction to epidemiological models with dispersion such as reaction-diffusion epidemiological models [21], the movements of the individuals are not modeled explicitly. The model implicitly assumes that infectious individuals may transmit the disease to susceptible individuals in other counties, but eventually return to their county of origin. This has the advantage of avoiding unrealistic changes in the global population density.

Observation model

We denote by $\overline{D}_k(t)$ the expected cumulative number of deaths given by the model, in county k. With the mean-field model $\mathcal{M}_0$, we assume that it is proportional to the population size: $\overline{D}_k(t) = D(t)\, N_k/N$, with N_k the population in county k and N the total French population. With models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$, we simply have $\overline{D}_k(t) = D_k(t)$. The expected daily increment in the number of deaths given by the models in a county k is $\overline{D}_k(t) - \overline{D}_k(t-1)$.

The observation model assumes that the daily number of new observed deaths $\hat{\mu}_{k,t}$ in county k follows a Poisson distribution with mean value $\overline{D}_k(t) - \overline{D}_k(t-1)$:

$$\hat{\mu}_{k,t} \sim \mathrm{Poisson}(\overline{D}_k(t) - \overline{D}_k(t-1)). \tag{4}$$

Note that the time t in the mechanistic models is a continuous variable, while the observations $\hat{\mu}_{k,t}$ are reported at discrete times. For the sake of simplicity, we used the same notation t for the days in both the discrete and continuous cases. In the formula (4) $\overline{D}_k(t)$ (resp. $\overline{D}_k(t-1)$) is computed at the end of day t (resp. $t-1$).

Initial conditions

In models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, at initial time t_i, we assume that the number of susceptible cases is equal to the number of inhabitants in county k: $S_k(t_i) = N_k$, the number of recovered is $R(t_i) = 0$ and the number of deaths is given by the data: $D_k(t_i) = \hat{\mu}_{k,t_i}$. To initialise the number of infectious, we use the equation $D'(t) = \gamma\, I(t)$, and we define $I(t_i)$ as $1/\gamma \times$ (mean number of deaths over the

period ranging from t_i to 29 days after t_i):

$$I_k(t_i) = \frac{1}{\gamma}\frac{1}{30}\sum_{s=t_i,\ldots,t_i+29}\hat{\mu}_{k,s}. \tag{5}$$

The 30-days window was chosen such that there was at least one infectious case in each county. In model $\mathcal{M}_0$, the initial conditions are obtained by adding the initial conditions of model $\mathcal{M}_1$ (or equivalently, $\mathcal{M}_2$) over all the counties.

Statistical inference

Real-time monitoring of the parameters and data assimilation procedure. To smooth out the effect of small variations in the data, and to avoid identifiability issues due to the large number of parameters, while keeping the temporal dependence of the parameters, the parameters $\alpha(t)$ and $\alpha_k(t)$ of the ODE models $\mathcal{M}_0$, $\mathcal{M}_1$, $\mathcal{M}_2$ are estimated by fitting auxiliary problems with time-constant parameters over moving windows $(t-\tau/2, t+\tau/2)$ of fixed duration equal to τ days. These auxiliary problems are denoted respectively by $\tilde{\mathcal{M}}_{0,t}$, $\tilde{\mathcal{M}}_{1,t}$, and $\tilde{\mathcal{M}}_{2,t}$ (see SI 2 for a precise formulation of these problems). The initial conditions associated with this system, at the date $t-\tau/2$ are computed iteratively from the solution of $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively.

Inference procedure. For simplicity, in all cases, we denote by

$$f_{\overline{D}_k,\hat{\mu}_k}(s) := \frac{(\overline{D}_k(s)-\overline{D}_k(s-1))^{\hat{\mu}_{k,s}}}{\hat{\mu}_{k,s}!}\, e^{-(\overline{D}_k(s)-\overline{D}_k(s-1))}$$

the probability mass function associated with the observation process (4) at date s in county k, given the expected cumulative number of deaths $\overline{D}_k$ given by the considered model in county k.

In models $\tilde{\mathcal{M}}_{0,t}$ and $\tilde{\mathcal{M}}_{1,t}$, the estimated parameter is $\tilde{\alpha}$. The likelihood $\mathcal{L}$ is defined as the probability of the observations (here, the increments $\{\hat{\mu}_{k,s}\}$) conditionally on the parameter. Using the assumption that the increments $\hat{\mu}_{k,s}$ are independent conditionally on the underlying SIRD process $\tilde{\mathcal{M}}_{0,t}$ (resp. $\tilde{\mathcal{M}}_{1,t}$), we get:

$$\begin{aligned}\mathcal{L}(\tilde{\alpha}) &:= P(\{\hat{\mu}_{k,s},\, k=1,\ldots,n_d,\, s=t-\tau/2,\ldots,t+\tau/2\}|\tilde{\alpha})\\ &= \prod_{k=1}^{n_d}\prod_{s=t-\tau/2}^{t+\tau/2} f_{\overline{D}_k,\hat{\mu}_k}(s).\end{aligned}$$

We denote by $\tilde{\alpha}^*_t$ the corresponding maximum likelihood estimator, and we set $\alpha(t)=\tilde{\alpha}^*_t$ in model $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).

For model $\tilde{\mathcal{M}}_{2,t}$, the inference of the parameters $\tilde{\alpha}_k$ is carried out independently in each county, leading to the likelihoods:

$$\begin{aligned}\mathcal{L}_k(\tilde{\alpha}_k) &:= P(\{\hat{\mu}_{k,s},\ s=t-\tau/2,\ldots,t+\tau/2\}|\tilde{\alpha}_k)\\ &= \prod_{s=t-\tau/2}^{t+\tau/2} f_{\overline{D}_k,\hat{\mu}_k}(s).\end{aligned}$$

We denote by $\tilde{\alpha}^*_{k,t}$ the corresponding maximum likelihood estimator, and we set $\alpha_k(t)=\tilde{\alpha}^*_{k,t}$ in model $\mathcal{M}_2$.

For model $\mathcal{M}_3$, we apply a two-stage estimation approach. We first use the estimate obtained with model $\mathcal{M}_1$ by setting $\rho(t) = C\,\alpha(t)$, where $\alpha(t)$ is the estimated contact rate of model $\mathcal{M}_1$ and C is a constant (to be estimated; note that estimating $\rho(t)$ means estimating n_t parameters). Thus, given $\alpha(t)$, the only parameters to be estimated are the constant C, the proximity scale d_0 and the exponent δ. They are estimated by maximizing:

$$\begin{aligned}\mathcal{L}(C, d_0, \delta) &:= P(\{\hat{\mu}_{k,s},\ k = 1, \ldots, n_d,\ s = t_i, t_f\}|C, d_0, \delta) \\ &= \prod_{k=1}^{n_d} \prod_{s=t_i}^{t_f} f_{\overline{D}_k, \hat{\mu}_k}(s).\end{aligned}$$

Model selection. We use the Akaike information criterion (AIC) [22] and the Bayesian information criterion (BIC) [23] to compare the models. For both criteria, we need to compute the likelihood function associated with the model, with the parameters determined with the above inference procedure:

$$\mathcal{L}(\mathcal{M}_m) = \prod_{k=1}^{n_d} \prod_{s=t_i}^{t_f} f_{\overline{D}_k, \hat{\mu}_k}(s), \tag{6}$$

with $m = 0\,, 1,\ 2,\ 3$ and $\overline{D}_k$ the expected (cumulative) number of deaths in county k given by model $(\mathcal{M}_m)$. Given the number of parameters $\sharp\mathcal{M}_m$ estimated in model $\mathcal{M}_m$, the AIC score is defined as follows:

$$AIC(\mathcal{M}_m) = 2\,\sharp\mathcal{M}_m - 2\,\ln(\mathcal{L}(\mathcal{M}_m)),$$

and the BIC score:

$$BIC(\mathcal{M}_m) = \sharp\mathcal{M}_m\,\ln(K) - 2\,\ln(\mathcal{L}(\mathcal{M}_m)),$$

with $K = n_d \times n_t = 6\,438$ the number of data points.

Numerical methods. To find the maximum likelihood estimator, we used a BFGS constrained minimization algorithm, applied to $-\ln(\mathcal{L})$ via the Matlab ® function *fmincon*. The ODEs were solved thanks to a standard numerical algorithm, using Matlab® *ode45* solver. The Matlab ® codes are available as Supplementary Material.

Results

Model fit and model selection. To assess model fit, we compared the daily increments in the number of deaths given by each of the four models with the data. For each model, we used the parameters corresponding to the maximum likelihood estimators. The comparisons are carried out at the regional scale: mainland France is divided into 12 administrative regions, each of which is made of several counties. The results are presented in Fig. 1. In all the regions, the models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ lead to a satisfactory visual fit of the data, whereas the mean field model $\mathcal{M}_0$ does not manage to reproduce the variability of the dynamics among the regions.

The log-likelihood, AIC and BIC values are given in Table 2. Models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ lead to significantly higher likelihood values than model $\mathcal{M}_0$. This reflects the better fit obtained with these three models, compared to model $\mathcal{M}_0$ and shows the importance of taking into account the spatial

Model	AIC	BIC	Log-likelihood	ΔAIC
$\mathcal{M}_0$	$2.68 \cdot 10^4$	$2.73 \cdot 10^4$	$-13.4 \cdot 10^3$	$-9.57 \cdot 10^3$
$\mathcal{M}_1$	$1.74 \cdot 10^4$	$1.79 \cdot 10^4$	$-8.62 \cdot 10^3$	-220
$\mathcal{M}_2$	$2.97 \cdot 10^4$	$7.36 \cdot 10^4$	$-8.41 \cdot 10^3$	$-1.25 \cdot 10^4$
$\mathcal{M}_3$	$1.72 \cdot 10^4$	$1.77 \cdot 10^4$	$-8.52 \cdot 10^3$	0

Table 2: Log-likelihood, AIC and BIC values for the four models. The last column ΔAIC corresponds to the difference with the AIC value of the best model (here $\mathcal{M}_3$).

heterogeneities in the initial densities of infectious cases. On the other hand, the log-likelihood, though higher with model $\mathcal{M}_2$ is close to that obtained with $\mathcal{M}_1$, and the model selection criteria are both strongly in favor of model $\mathcal{M}_1$. This shows that the spatial heterogeneity in the contact rate does not have a significant effect on the epidemic dynamics within mainland France.

Model $\mathcal{M}_3$ with spatial transmission leads to an intermediate likelihood value, between those of models $\mathcal{M}_1$ and $\mathcal{M}_2$, with only 2 additional parameters with respect to model $\mathcal{M}_1$. As a consequence, the model selection criteria exhibit a strong evidence in favor of the selection of model $\mathcal{M}_3$. This means a large part of the difference between models $\mathcal{M}_1$ and $\mathcal{M}_2$ can be captured by taking into account the spatial transmission, which therefore seems to have a significant effect on the epidemic dynamics.

As a byproduct of the estimation of the parameter $\alpha(t)$ (resp. $\alpha_k(t)$) of model $\mathcal{M}_1$ (resp. $\mathcal{M}_2$), we get an estimate of the effective reproduction number in each county, which is given by the formula [24]:

$$\mathcal{R}_t^k = \frac{\alpha(t)}{\beta+\gamma}\frac{S_k(t)}{N_k},$$

so that $I_k'(t) < 0$ (the epidemic tends to vanish) whenever $\mathcal{R}_t^k < 1$, whereas $I_k'(t) > 0$ whenever $\mathcal{R}_t^k > 1$ (the number of infectious cases in the population follows an increasing trend). The dynamics of $\mathcal{R}_t^k$ obtained with model $\mathcal{M}_1$ are depicted in Fig. 2, clearly showing a decline in $\mathcal{R}_t^k$, as already observed in [10, 7], but there the computation was at a fixed date.

For model $\mathcal{M}_3$, the maximum likelihood estimation gives $C = 0.87$, $d_0 = 2.16$ km and $\delta = 1.85$, which yields a nearly quadratic decay of the weights with the distance. The value of d_0, indicates that non-local contagion plays a secondary role compared to within-county contagion: the minimum distance between two counties is 36 km, leading to a weight of 5.5/1000, to be compared with the weight 1 for within-county contagion. However, the fact that the parameter C is significantly smaller than 1 (recall that the contact rate in model $\mathcal{M}_3$ is $\rho(t) = C\,\alpha(t)$ with $\alpha(t)$ the contact rate in model $\mathcal{M}_2$) shows that the non-local contagion term plays an important role on the spreading of the epidemic.

Immunity rate. Using model $\mathcal{M}_3$, which leads to the best fit, we derive the number of recovered individuals (considered here as immune) at a date t, in each county, and the immunity rate R_k/N_k. It is presented at time t_f in Fig. 3. The full timeline of the dynamics of immunity obtained with model $\mathcal{M}_3$ since the beginning of April is available as Supplementary Material (see SI 1).

Limiting movement vs limiting the probability of transmission per contact. Before the lockdown, the basic reproduction number $\mathcal{R}_0$ in France was about 3 [7, 3], and was then reduced by a factor

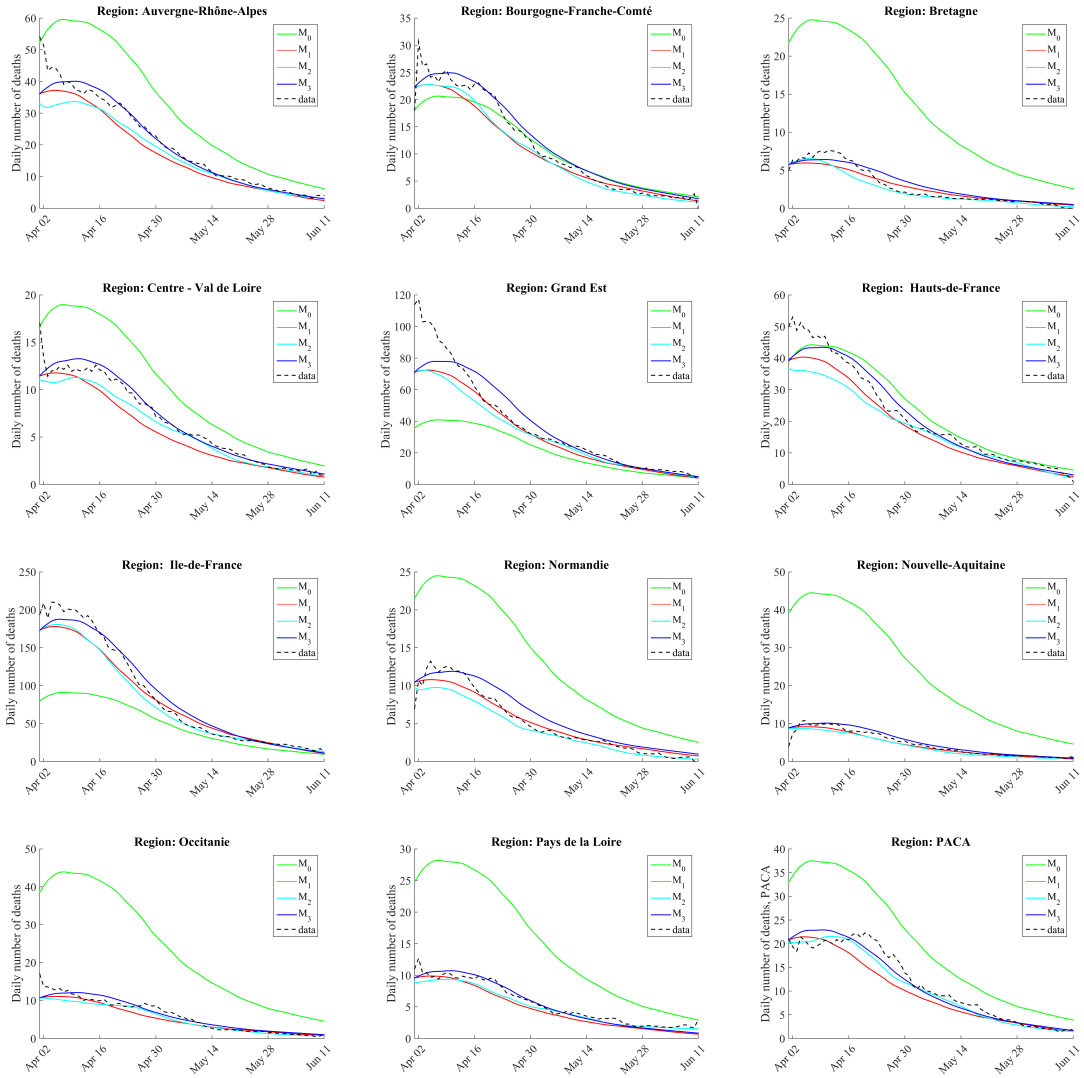

Figure 1: Model fit at the regional scale. The data have been smoothed (moving average over 15 days), for easier graphical comparison with the models.

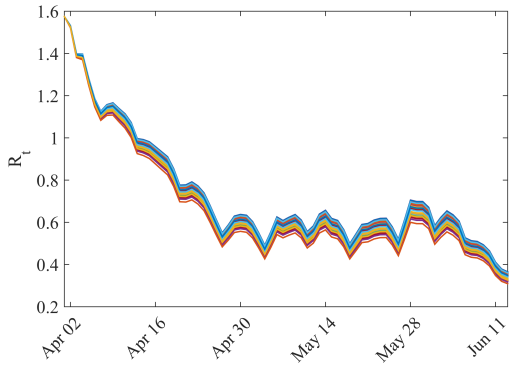

Figure 2: Dynamics of the effective reproduction rate $\mathcal{R}_t^k$ given by model $\mathcal{M}_1$ over the 87 considered counties.

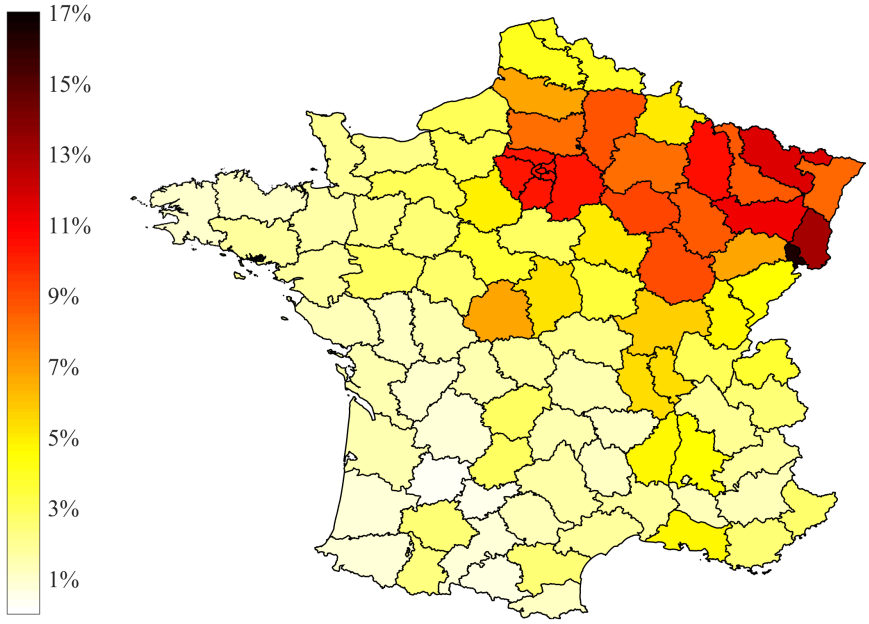

Figure 3: Estimated immunity rate, at the county scale, by June 11, 2020, using model $\mathcal{M}_3$.

5 to 7, leading to values around 0.5 (see [10, 7] and Fig. 2). This corresponds to a contact rate $\alpha(t) \approx \beta \mathcal{R}_t \approx 0.3$ before the lockdown and $\alpha(t) \approx 0.05$ after the lockdown. Let us now consider a hypothetical scenario of a new outbreak with an effective reproduction number that rises again to reach values above 1. Due to the higher awareness of the population with respect to epidemic diseases and the new sanitary behaviors induced by the first COVID-19 wave, the reproduction number will probably not reach again values as high as 3.

The new outbreak scenario is described as follows: we start from the state of the epidemic at June 11, and we assume a 'local' contact rate $\rho(t)$ jumping to the value 0.11 in model $\mathcal{M}_3$ (corresponding to a 2-fold increase compared to the previous 30 days). In parallel, to describe the lifting of restrictions on individual movements, we set $d_0 = 20$ km for the proximity scale in model $\mathcal{M}_3$. This new outbreak runs during 10 days, and then, we test four strategies:

- Strategy 1: no restriction. The parameters remain unchanged: $\rho(t) = 0.11$ and $d_0 = 20$ km;
- Strategy 2: restriction on intercounty movement. The parameter $\rho(t) = 0.11$ is unchanged, but $d_0 = 2.16$ km, corresponding to its estimated value during the period (t_i, t_f);
- Strategy 3: reduction of the contact rate within each county (e.g., by wearing masks), but no restriction on intercounty movement: $\rho(t) = 0.11$ and $d_0 = 20$ km;
- Strategy 4: reduction of the contact rate within each county and restriction on intercounty movement: $\rho(t) = 0.11$ and $d_0 = 2.16$ km.

The daily number of deaths corresponding to each scenario is presented in Fig. 4. We estimate in each case a value of the effective reproduction number $\mathcal{R}_t$ over the whole country by fitting the global number of infectious cases with an exponential function over the last 30 days. As expected, the more restrictive the strategy, the less the number of deaths. After 30 days, the cumulative number of deaths with the first strategy is 17 271, and $\mathcal{R}_t \approx 2$. Restriction on intercounty movement (strategy 2) leads to a 81% decrease in the cumulative number of deaths (3 281 deaths) and $\mathcal{R}_t \approx 1.2$; reducing the contact rate within each county leads to a 88% decrease (strategy 3, 2 139 deaths) and $\mathcal{R}_t \approx 0.8$; finally, control strategy 4, which combines both types of restrictions leads to a 91% decrease (1 503 deaths) and $\mathcal{R}_t \approx 0.4$.

Discussion

We show here that a parsimonious model can reproduce the local dynamics of the COVID-19 epidemic in France with a relatively high goodness of fit. This is achieved despite the spatial heterogeneity across French counties of some environmental factors potentially influencing the disease propagation. Indeed, our model only involves the initial spatial distribution of the infectious cases and spatially-homogeneous (i.e. countrywide) parameters. For instance, the mean temperature during the considered period ranges from 12.0°C to 18.4°C depending on the region. We do observe a negative correlation between the mean temperature and the immunity rate (see Fig. S1), however it does not reflect a causality. Actually, our study shows that if there is an effect of local covariates such as the mean temperature on the spread of the disease once it emerged, its effect is of lower importance compared to the global processes at work at the country scale. Such covariates might play a major role in the emergence of the disease, but our work focuses on the disease dynamics after the emergence.

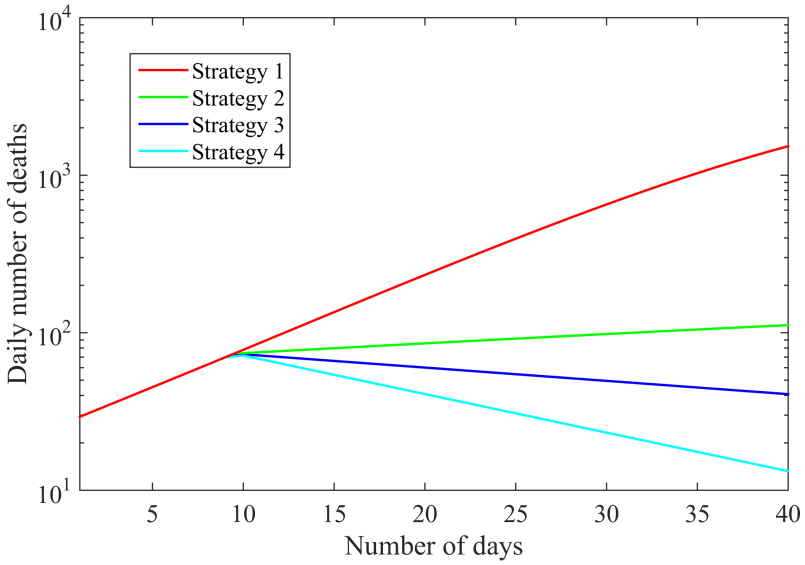

Figure 4: Daily number of deaths due to a new outbreak in logarithmic scale; comparison between four management strategies. The number of deaths is computed over the whole country.

Hence, we find that initial conditions and spatial diffusion are the main drivers of the spatial pattern of the COVID-19 epidemic. This result may rely on specific circumstances: e.g., mainland France covers a relatively middle-sized area, with mixed urban and countryside populations across the territory, a relatively homogeneous population age distribution, and a high level of centralism for public decision (in particular regarding the disease-control strategy). Of course, these features are not universal. In other countries with more socio-environmental diversity within which environmental drivers and state decentralization could significantly induce spatial variations in disease spread and, consequently, in which countrywide parameters would not be appropriate. Moreover, at the global scale, the COVID-19 dynamics in different countries seem highly contrasted as illustrated by the data of Johns Hopkins University Center for Systems Science and Engineering [25] —see also `http://covid19-forecast.biosp.org/` [26]— and could probably not be explained with a unique time-varying contact rate parameter. Nonetheless, this model could be adapted to many other situations at an appropriate geographical level, with a single well defined political decision process.

Herd immunity requires that a fraction $1 - 1/\mathcal{R}_0 \approx 70\%$ of the population has been infected. It is far from being reached at the country scale in France, but we observe that the fraction of immune individuals strongly varies across the territory, with possible local immunity effects. For instance, in the most impacted county the immunity rate is 16%, whereas it is less than 1% in less affected counties. At a thinner grain scale, even higher rates may be observed, for instance, by April 4 the proportion of people with confirmed SARS-CoV-2 infection based on antibody detection was 41% in a high-school located in Northern France [27].

Real-time monitoring of the immunity level will be crucial to define efficient management policies, if a new outbreak occurs. We propose such a tool which is based on the modeling approach $\mathcal{M}_3$ of this paper (see Immunity tab in `http://covid19-forecast.biosp.org/`). Remarkably, the estimated levels of immunity are comparable to those observed in Spain by population-based serosurveys [28], with values ranging from 0.5% to 13.0% at the beginning of May at the provincial resolution. Such a large-scale serological testing campaign has not yet been carried out in France. However, if such data become available in France, our predictions could be evaluated and our model updated accordingly by including this new dataset in our estimation procedure. The mechanistic-statistical approach that we followed here can indeed be easily adapted to multiple observation processes.

Our results indicate that —at the country scale— travel restriction alone, although they may have a significant effect on the cumulative number of deaths and the reproduction number over a definite period, are less efficient than social distancing and other sanitary measures. Obviously, these results may strongly depend on the parameter values, which have been chosen here on the basis of values estimated during the lockdown period. This is consistent with results for China [29], where travel restrictions to and from Wuhan have been shown to have a modest effect unless paired with other public health interventions and behavioral changes.

The model selection criteria led to a strong evidence in favor of the selection of model $\mathcal{M}_3$ with non-local transmission and spatially-constant contact rate. It is much more parsimonious than the fully heterogeneous model $\mathcal{M}_2$ and is therefore better suited to isolating key features of the epidemiological dynamics [4]. Despite important restrictions on movement during the considered period (mandatory home confinement except for essential journeys until 11 May and a 100 km travel restriction until 2 June), the model $\mathcal{M}_3$ was also selected against the model $\mathcal{M}_1$ which does not take into account non-local transmission. This shows that intercounty transmission is one of these key-features that the non-local model manages to take into account.

More generally, in France just as in Italy [8], the spatial pattern of COVID-19 incidence indicates that spatial processes play a key role. At this stage, only a few models can address this aspect. Some have adopted a detailed spatio-temporal modeling approach and use mobility data (see [8] for an SEIR-like model with 9 compartments). The framework we develop here, including the non-local model and the associated estimation procedure, should be of broad interest in studying the spatial dynamics of epidemics, due to its theoretical and numerical simplicity and its ability to accurately track the epidemics. This approach applies when geographic distance matters, which may not be the case at the scale of countries like the US. However, it does at more regional scales. Furthermore, we can envision natural extensions of the approach that would take into account long range dispersal events in the interaction term.

Data availability

All data used in this manuscript are publicly available. French mortality data at the county scale are available at https://www.gouvernement.fr/info-coronavirus/carte-et-donnees and are also available as Supplementary Material.

Acknowledgments

This work was funded by INRAE (MEDIA network) and EHESS. We thank Jean-François Rey for assistance in developing the web app.

References

[1] A Deslandes, et al., SARS-COV-2 was already spreading in France in late December 2019. *International Journal of Antimicrobial Agents*, 106006 (2020).

[2] J Demongeot, Y Flet-Berliac, H Seligmann, Temperature decreases spread parameters of the new Covid-19 case dynamics. *Biology* **9**, 94 (2020).

[3] L Roques, E Klein, J Papaix, A Sar, S Soubeyrand, Using early data to estimate the actual infection fatality ratio from COVID-19 in France. *Biology* **9**, 97 (2020).

[4] AL Bertozzi, E Franco, G Mohler, MB Short, D Sledge, The challenges of modeling and forecasting the spread of COVID-19. *Proceedings of the National Academy of Sciences* https://doi.org/10.1073/pnas.2006520117 (2 July 2020).

[5] BF Maier, D Brockmann, Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science* **368**, Issue 6492, 742–746 (2020).

[6] K Prem, et al., The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *The Lancet Public Health* **5**, Issue 5, 261–270 (2020).

[7] H Salje, et al., Estimating the burden of SARS-CoV-2 in France. *Science* **369**, Issue 6500, 208–211 (2020).

[8] M Gatto, et al., Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences* **117**, 10484–10491 (2020).

[9] AL García-Basteiro, et al., Monitoring the COVID-19 epidemic in the context of widespread local transmission. *The Lancet Respiratory Medicine* **8**, 440–442 (2020).

[10] L Roques, EK Klein, J Papaix, A Sar, S Soubeyrand, Impact of lockdown on the epidemic dynamics of COVID-19 in France. *Frontiers in Medicine* https://doi.org/10.3389/fmed.2020.00274 (5 June 2020).

[11] J Zhang, et al., Age profile of susceptibility, mixing, and social distancing shape the dynamics of the novel coronavirus disease 2019 outbreak in China. *medRxiv* https://dx.doi.org/10.1101%2F2020.03.19.20039107 (20 March 2020).

[12] LM Berliner, Physical-statistical modeling in geophysics. *J Geophys Res* **108**, 8776 (2003).

[13] S Soubeyrand, AL Laine, I Hanski, A Penttinen, Spatiotemporal structure of host-pathogen interactions in a metapopulation. *American Naturalist* **174**, 308–320 (2009).

[14] F Zhou, et al., Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* **395**, Issue 10229, 1054–1062 (2020).

[15] X He, et al., Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, **26**, Issue 5, 672–675 (2020).

[16] D Kendall, Discussion of 'Measles periodicity and community size' by MS Bartlett. *J. Roy. Stat. Soc. A* **120**, 64–76 (1957).

[17] L Bonnasse-Gahot, H Berestycki, M-A Depuiset, M B Gordon, S Roché, N Rodriguez, J-P Nadal, Epidemiological modelling of the 2005 French riots: a spreading wave and the role of contagion. *Scientific Reports* **8**, 1–20 (2018).

[18] D Brockmann, L Hufnagel, T Geisel, The scaling laws of human travel. *Nature* **439**, 462–465 (2006).

[19] S Meyer, L Held, , et al., Power-law models for infectious disease spread. *The Annals of Applied Statistics* **8**, 1612–1639 (2014).

[20] V Pareto, Cours d'Economie Politique, (Lausanne: F. Rouge, 1896).

[21] J Gaudart, et al., Demography and diffusion in epidemics: malaria and black death spread. *Acta Biotheoretica* **58**, 277–305 (2010).

[22] H Akaike, A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716–723 (1974).

[23] G Schwarz, Estimating the dimension of a model. *Ann Stat* **6**, 461–464 (1978).

[24] H Nishiura, G Chowell, The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends in *Mathematical and statistical estimation approaches in epidemiology.* (Springer), pp. 103–121 (2009).

[25] E Dong, H Du, L Gardner, An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20**, Issue 5, 533–534 (2020).

[26] S Soubeyrand, et al., The current COVID-19 wave will likely be mitigated in the second-line European countries. *medRxiv* https://doi.org/10.1101/2020.04.17.20069179 (22 April 2020).

[27] A Fontanet, et al., Cluster of COVID-19 in northern France: A retrospective closed cohort study. *medRxiv* https://doi.org/10.1101/2020.04.18.20071134 (23 April 2020).

[28] M Pollán, et al., Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. *The Lancet* https://doi.org/10.1016/S0140-6736(20)31483-5 (6 July 2020).

[29] M Chinazzi, et al., The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).

[30] X Liu, S Zhang, Covid-19: Face masks and human-to-human transmission. *Influenza and Other Respiratory Viruses* (2020).

Supplementary Material

SI 1. Supplementary figures

- The mean temperatures during the observation period are presented in Fig. S1, together with a scatter plot of the mean temperature vs immunity rate.

  (source: https://www.data.gouv.fr/fr/datasets/temperature-quotidienne-departementale-depuis-janvier-2018/).

- In Fig. S2, we describe the timeline of the spatio-temporal dynamics of the immunity rate during the observation period.

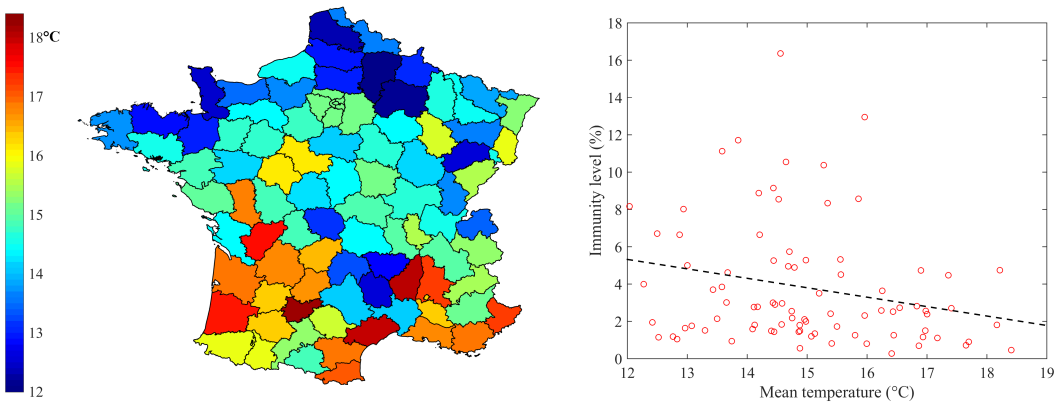

Figure S1: (a) Mean temperature in each county over the period ranging from 30 March 2020 to 11 June 2020. (b) Immunity rate vs mean temperature: we observe a negative correlation between the mean temperature and the immunity rate (Pearson correlation coefficient: -0.24).

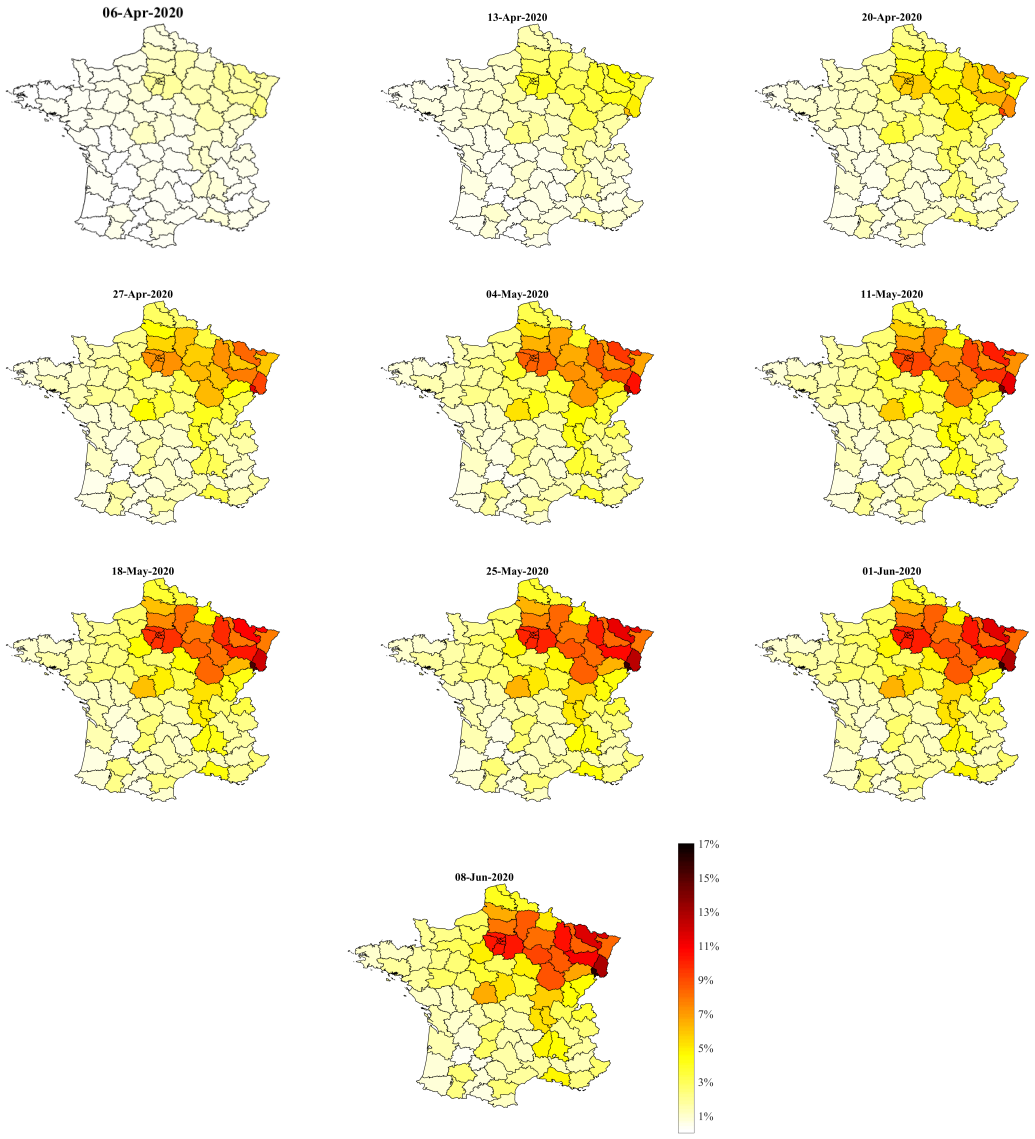

Figure S2: Dynamics of the immunity rate given by model $\mathcal{M}_3$, from 6 April 6 to 8 June.

SI 2. Auxiliary models

At each date t, the time-dependent parameters $\alpha(t)$ in $\mathcal{M}_0$ and $\mathcal{M}_1$ and $\alpha_k(t)$ in $\mathcal{M}_2$ maximum likelihood estimators (MLEs) in the window $(t-\tau/2, t+\tau/2)$ of the following auxiliary models:

$$\begin{cases} \tilde{S}'(s) = -\dfrac{\tilde{\alpha}}{N}\,\tilde{S}\,\tilde{I}, \\ \tilde{I}'(s) = \dfrac{\tilde{\alpha}}{N}\,\tilde{S}\,\tilde{I} - (\beta+\gamma)\,\tilde{I}, \quad \text{for } s \in (t-\tau/2, t+\tau/2), \\ \tilde{R}'(s) = \beta\,\tilde{I}, \\ \tilde{D}'(s) = \gamma\,\tilde{I}, \end{cases} \tag{$\tilde{\mathcal{M}}_{0,t}$}$$

and

$$\begin{cases} \tilde{S}_k'(s) = -\dfrac{\tilde{\alpha}}{N_k}\,\tilde{S}_k\,\tilde{I}_k, \\ \tilde{I}_k'(s) = \dfrac{\tilde{\alpha}}{N_k}\,\tilde{S}_k\,\tilde{I}_k - (\beta+\gamma)\,\tilde{I}_k, \quad \text{for } s \in (t-\tau/2, t+\tau/2), \\ \tilde{R}_k'(s) = \beta\,\tilde{I}_k, \\ \tilde{D}_k'(s) = \gamma\,\tilde{I}_k, \end{cases} \tag{$\tilde{\mathcal{M}}_{1,t}$}$$

and

$$\begin{cases} \tilde{S}_k'(s) = -\dfrac{\tilde{\alpha}_k}{N_k}\,\tilde{S}_k\,\tilde{I}_k, \\ \tilde{I}_k'(s) = \dfrac{\tilde{\alpha}_k}{N_k}\,\tilde{S}_k\,\tilde{I}_k - (\beta+\gamma)\,\tilde{I}_k, \quad \text{for } s \in (t-\tau/2, t+\tau/2). \\ \tilde{R}_k'(s) = \beta\,\tilde{I}_k, \\ \tilde{D}_k'(s) = \gamma\,\tilde{I}_k, \end{cases} \tag{$\tilde{\mathcal{M}}_{2,t}$}$$

The initial condition in these models is computed iteratively from the solutions of $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, over the period $[t_i, t-\tau/2]$.