Piecewise estimation of R0 by a simple SEIR model. Application to COVID-19 in French regions and departments until June 30, 2020

Stéphane Derrode, Romain Gauchon, Nicolas Ponthus, Christophe Rigotti, Catherine Pothier, Vitaly Volpert, Stéphane Loisel, Jean-Pierre Bertoglio, Pascal Roy

Piecewise estimation of R_0 by a simple SEIR model.

Application to COVID-19 in French regions and departments until June 30, 2020.

S. Derrode · R. Gauchon · N. Ponthus · C. Rigotti · C. Pothier · V. Volpert · S. Loisel · J.-P. Bertoglio · P. Roy · for the cofine-project research group

Version: V3 (date:September 2, 2020)

Abstract The estimation of R_0, the so-called "basic reproductive ratio", of the COVID-19 pandemic is of particular importance to help decision-makers take the necessary safeguard measures to protect the population. In this work, we examine a method based on the successive estimation of R_0 over 3 non-overlapping periods (beginning of lockdown, during lockdown and after). The approach is based on a variant of the, simple but flexible, SEIR compartmental model that allows to exploit the number of recovered individuals that are reported in the daily database published by national health agencies. The results of the approach is analysed *w.r.t.* data from France, at two levels of geographical sub-divisions, *i.e.* the 13 regions and 96 departments that make up the metropolitan territory.

Keywords R_0 estimation · Epidemiology · SEIR compartmental model · COVID-19 · France regions and departments

1 Introduction

The basic reproduction number, also called "basic reproductive ratio", denoted by R_0, is a measure of the potential for disease to spread in a population. R_0 is the number of individuals infected by an infectious person in a population of susceptible. If $R_0 < 1$, then a few infected individuals introduced into a completely susceptible population will, on average, fail to replace themselves, and the disease will not spread. If, on the other hand, $R_0 > 1$, then the number of infected individuals will increase with each generation and the disease will spread.

From the work of Diekmann, van den Driessche and Watmough (Diekmann et al., 1990; van den Driessche and Watmough, 2002; van den Driessche and Watmough, 2008), the so-called "second generation method" is a general method of deriving R_0 for a deterministic compartmental model of disease transmission. R_0 is the expected (*i.e.* averaged over many epidemics) numbers of secondary cases produced by generation zero. Heffernan et al. (2005) provides a nice readable introduction for calculating R_0 in structured population models. R_0 is rigorously defined by the largest eigenvalue (or spectral radius) of the matrix product $\boldsymbol{FV}^{-1}$ (called "next generation matrix"), where

$$\boldsymbol{F} = \left[\frac{\partial F_i(x_0)}{\partial x_j}\right], \boldsymbol{V} = \left[\frac{\partial V_i(x_0)}{\partial x_j}\right], \quad (1)$$

in which x_j denotes the number or proportion of individuals in the jth compartment and x_0 is the disease-free equilibrium state. The F_i are the new infections, while the V_i are the transfers of infections from one compartment to another. This principle has been applied in numerous studies to determine the R_0 associated with, *e.g.*, tuberculosis (Castillo-Chavez and Feng, 1997), malaria (Heffernan et al., 2005), scrapie (Matthews et al., 1999)... . See also (Brauer et al., 2019), for a detailed description of many other examples.

In this work, we deal with Bacaër, N.'s variant of the SEIR compartmental model ('S': number of susceptible individuals, 'E': number of exposed individuals, 'I': number of infectious individuals, 'R': number of recovered individuals), called SEIR^{1}R^2. In the SEIR^{1}R^2

Stéphane Derrode
Université de Lyon, CNRS, École Centrale de Lyon,
LIRIS, CNRS UMR 5205, France.
E-mail: stephane.derrode@ec-lyon.fr

Pascal Roy
Université de Lyon, CNRS,
LBBE, CNRS UMR 5558, France.
E-mail: pascal.roy@chu-lyon.fr

variant, R^1 and R^2 both include the number of individuals that returned home and the number of deaths due to the pandemic (the sum of the two populations will be called 'removed'). The difference between the two R compartments is that R^1 reports the removed that are accounted for at the hospital, and R^2 reports the others, those which are not accounted for at the hospital, in proportions f and $1-f$ respectively. So the meaning of $R(t) = R^1(t) + R^2(t)$ is quite different from the meaning of $R(t)$ in the SEIR model.

This SEIR^{1}R^2 model (Bacaër, N., 2020) is defined by the following system of ODEs

$$\frac{dS}{dt} = -aS\frac{I}{N}, \frac{dE}{dt} = aS\frac{I}{N} - bE,$$
$$\frac{dI}{dt} = bE - cI, \frac{dR^1}{dt} = fcI, \frac{dR^2}{dt} = (1-f)cI, \qquad (2)$$

so $\frac{dR}{dt} = cI$, where

- a (or β): Rate of transmission (exposure),
- b (or σ): Rate of infection (upon exposure),
- c (or γ): Rate of recovery (upon infection),
- f : Fraction of recovered observed in hospital.

The total number of living individuals N is constant ($N = S(t)+E(t)+I(t)+R(t)$) and we assume that the re-susceptibility of the population is zero.

This simple variation of the SEIR model does not affect the well-known equation that relates R_0 to the model's parameters:

$$\boldsymbol{F} = \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{V} = \begin{pmatrix} b & 0 \\ -b & c \end{pmatrix}, \quad R_0 = \frac{a}{c}. \qquad (3)$$

Hence, the question of estimating R_0 in this model amounts to developing a method to estimate parameters a and c, based on the available data for R^1.

Although the methodology is general, our analysis focuses on the evolution of the COVID-19 pandemic in France, with data provided by "Santé Publique France" [1]. This database provides information on

- the number of people currently hospitalized;
- the number of people currently in intensive care unit (ICU);
- the cumulative number of people returned home;
- the cumulative number of people who died in hospital.

This information is available at the level of French departments[2] and according to gender and age group. The algorithm we have developed exploits the cumulative number of people returned home plus the cumulative number of people who died in hospital, which is exactly what represents R^1.

So, we propose an algorithm that divide the period of the COVID pandemic, from `2020-03-18` (the beginning of the database, which is also the lockdown date in France) to `2020-06-30`, in three non-overlapping periods. To take into account that a certain time is required before the effects of public health policies can be measured on the data, a delay of δ days is applied. The second period starts on the date of the lockdown plus δ days, and the third period starts on the date of the deconfinement plus δ days. The choice of the δ value is discussed in the text. An optimization algorithm, exposed in Section 2, is applied on each period, successively; the estimation results for a period are then used to initialize the optimization process for the next one. While we are specifically interested in the estimation of R_0 for the third period (to detect a possible resumption of the pandemic in the analysed territory), the algorithm also furnishes R_0 estimates for the first two periods as well as an estimated date for the first infected in the territory. The method is analysed at two different metropolitan France scales: scale of the 13 regions and scale of the 96 Departments in Section 3. Emphasis is put on graphical presentations of the results as maps, to get an overview of France situation as of June 30, 2020. The paper ends with the main conclusions of the study, and discusses future work directions. All the algorithms are available at `https://github.com/SDerrode/divoc`. An explanatory note makes it possible to regenerate the figures present in this document.

2 R_0 estimation methodology

The proposed methodology considers three non-overlapping periods for the COVID pandemic, from the beginning of the database to June 30:

- The first period (denoted by P_1) starts from the beginning of the data (*i.e.* `2020-03-18`) and ends on the date of lockdown plus a delay (in days), denoted by δ.
- The second period (P_2) begins the day after the previous period and ends on the date of the deconfinement (`2020-05-11` in France) plus the same delay of δ days.
- The third period (P_3) begins the day after the previous period and ends on the last day of the database (`2020-06-30` in France).

The delay δ is used to take into consideration the fact that public policies (such as national lockdown) have

[1] `https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/`, from the French national open data platform `data.gouv.fr`.

[2] In this database, the information concerning French overseas departments and territories is very incomplete, and will not be considered. The experiments only deal with the metropolitan France.

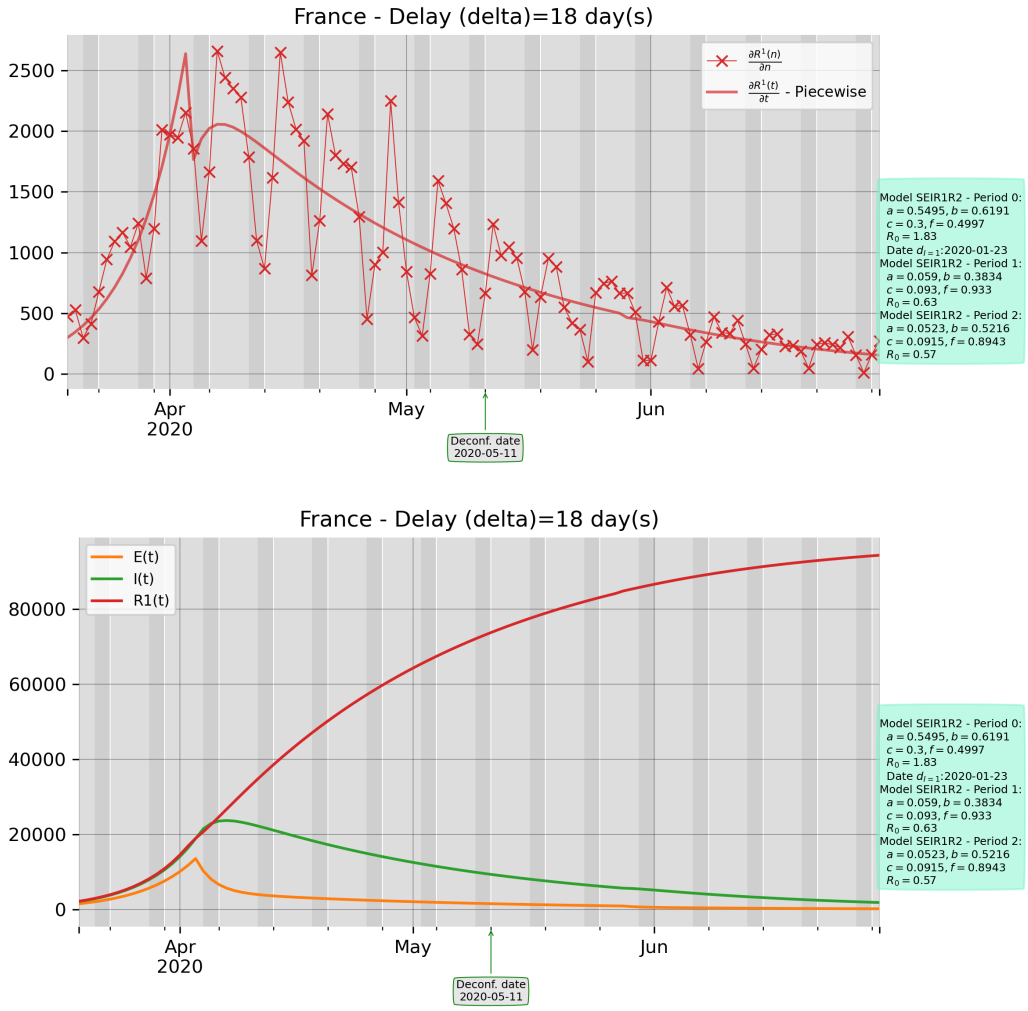


Fig. 1 Result of the three successive fits of SEIR^{1}R^2 model for the metropolitan France data. (Up) The red crosses denotes the number of (observed) new recovered and deaths, whereas the red curve represents the numerical derivative of $R^1(t)$. (Down) The fitting of the three SEIR^{1}R^2 models for compartments E, I and R^1. The delay *w.r.t.* the lock and unlock dates is set to $\delta = 18$ days.

not an immediate impact, but a certain time is required before their effects can me measured on the data. The choice for δ value is discussed below.

Each of the periods is processed the same way, by fitting a SEIR^{1}R^2 model specified by parameters a, b, c and f, using an optimization process, based on the match between the model's $R^1(t)$ –discretized at discrete instant n– and data $R^1(n)$. The fitting algorithm is based on the `lmfit`[3] library, which provides a high-level interface to non-linear optimization and curve fitting problems. We choose Powell's method for the optimization, because it does not require knowledge of derivatives. The parameters obtained for a period are used to initialize the optimizer for the next period, to help him to converge towards a workable solution. Also, all the values obtained for S, E, I, R^1 and R^2 at the end of a period are kept to initialize the next period.

From the first period, it is possible to forecast the date of the first infected individual. This date, denoted by $d_{I=1}$ in the remaining, is obtained by extrapolating I of the SEIR^{1}R^2 model obtained for period P_1 into the past.

2.1 Illustration at the scale of France

To illustrate the algorithm, let's look at the results for metropolitan France displayed in Figure 1. For the plot, we set $\delta = 18$ days and processed the data regardless of gender. We can observe that the red curve well-fits the noisy discrete data (red crosses). The "noise" here

[3] lmfit: non-linear least-squares minimization and curve-fitting for Python, `https://lmfit.github.io/lmfit-py/`.

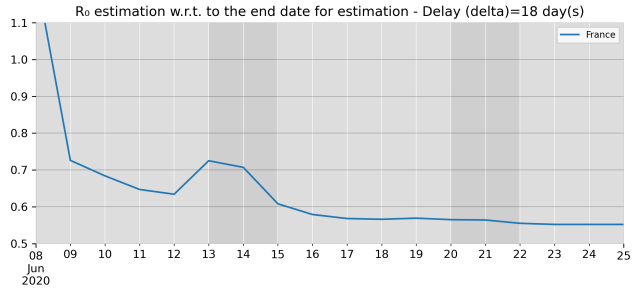


Fig. 2 R_0 estimation for France during period P_3, when data start $\delta = 18$ days after deconfinement and ends at the date specified in the x-axis. After an estimation period of 20 days, the R_0 estimation stabilizes at about 0.57.

refers to the irregularity of the feedback from hospitals, and the so-called "weekend effect" as observed in many countries. The green box to the right of the plots presents the estimated parameters for the 3 periods, as well as the R_0 estimates: $R_0 = 1.83$ for P_1, $R_0 = 0.63$ for P_2 and $R_0 = 0.57$ for P_3. According to the latter estimate, as of June 30, the pandemic has not resumed in France but its intensity is very weak. The first infected individual is estimated to be at $d_{I=1}$ =2020-01-23. This date seems consistent with the dates generally anticipated, as it is generally accepted that the first infected individual in France appears in the second half of January.

To evaluate the robustness of the R_0 estimation process during the third period, we estimated R_0 for increasing periods of time. Figure 2 shows the estimated values when the estimation period starts May 29 and ends in the interval [June 8, June 25]. It can be observed that R_0 stabilizes at about 0.57 after the 18th of June, which corresponds to an estimation period of 20 days [May 29, June 18].

To choose a value for δ –which corresponds to the estimated delay between a public decision and its possible impact on data–, we repeatedly estimated R_0 for P_3 and the means square error (MSE) between the model and the data for varying values of δ, see Figure 3. Plots show that the MSE and R_0 estimation present a minimum value when $\delta = 18$. It includes an incubation period of about 5 days. This time delay is to be taken with great care because it includes very different situations, with people who are doing a short stay in hospital and people who are doing a longer stay in ICU. Nevertheless, this value of $\delta = 18$ seems to be adapted to the simplicity of the model we are considering. This value is independent from the scale of analysis, and will be kept unchanged for regions as well as for departments.

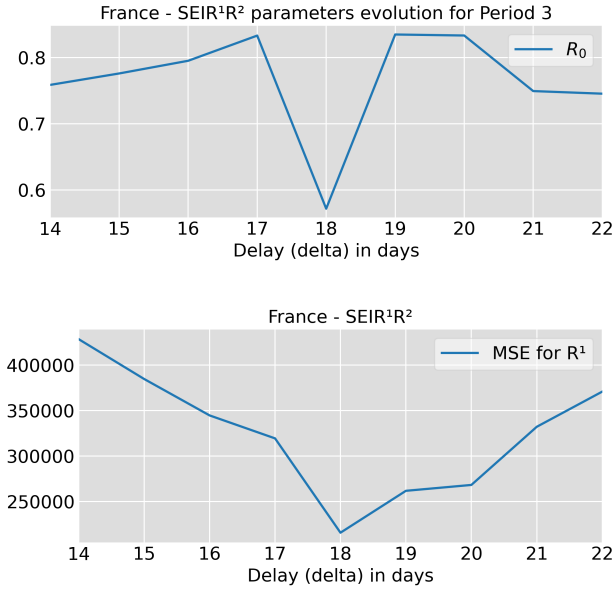


Fig. 3 (Up) Evolution of the estimation of R_0 (period P_3) for metropolitan France, according to the delay (δ) in days. (Down) Mean square error between the $R^1(n)$ data and the estimated $R^1(t)$ sampled at instant n.

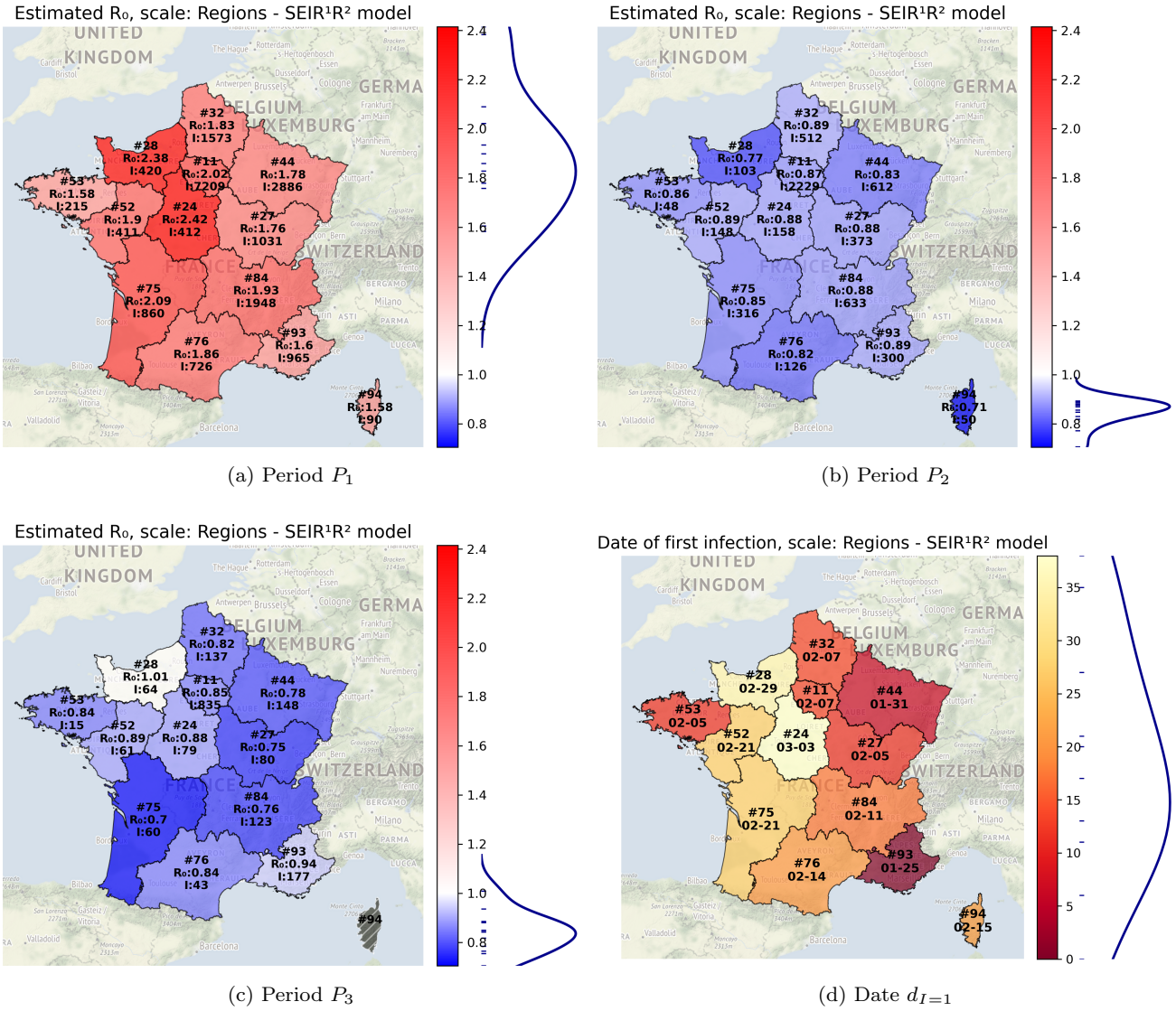
3 French region and department scales

Metropolitan France is divided in 13 regions, which are themselves organized into 96 departments. On a daily basis, the "Santé Publique France" database lists, among other data, the total amount of patients that returned home and the total number of deaths at the scale of the French departments. Hence, the estimation strategy we present can also be applied to the geographical subdivisions into regions and departments. To identify regions and departments, we use INSEE numbering[4], which associates sometimes the same number for a region or for a department, *e.g.* the *Auvergne-Rhône-Alpes* region and the *Vaucluse* department share the number #84!

When the volume of data is very small, estimates become less robust. Indeed, some subdivisions do not present a sufficient number of patients that returned home over the period considered for the R_0 estimate to be assumed reliable. This is particularly true for regions or departments not dense in terms of population. Therefore, we have decided to exclude subdivisions according to the following rule

$$\frac{\#\text{ of patients that returned home over the period}}{\#\text{ of days in the period} \times \text{size of the population}} < 10^{-6}.$$

[4] Institut National de la Statistique et des Études Économiques, `https://insee.fr/en/accueil`. See files at `https://www.insee.fr/fr/information/3720946`.



(a) Period P_1

(b) Period P_2

(c) Period P_3

(d) Date $d_{I=1}$

Fig. 4 Estimation of R_0 for the 3 periods and first infected individual date at region scale. For the subfigures (a), (b) and (c) we added the estimated number of infected at the end of the period, to get an idea on the intensity of the pandemic. The right side of each map is a smoothed representation of the histogram of values appearing in the corresponding map. The regions with hatches have been excluded according to a threshold explained in the text. For (d), value '0' in the legend corresponds to February 3, and value '35' corresponds to March 8, *i.e.* 35 days after February 3.

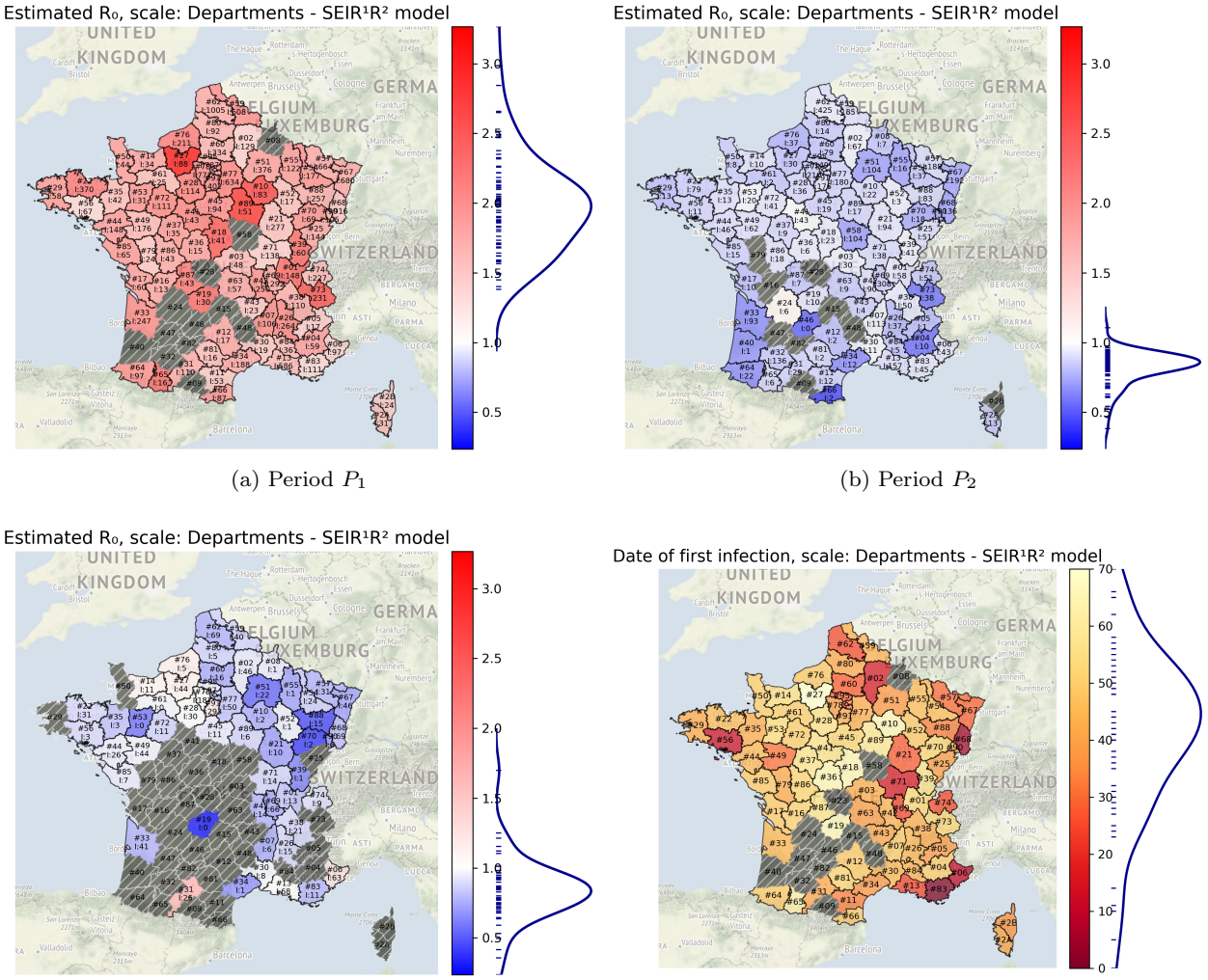
(4)

The threshold value was set to 10^{-6} by trials, to exclude areas only slightly affected by the pandemic, without excluding too many. The rule is applied for each period independently, so that an area can be excluded in P_2 and reintegrated in P_3. Indeed, even if model's parameters are not satisfactorily estimated at P_2 (so R_0), the obtained $R^1(t)$ function however gives always a good fit of the data (with bad parameters), so the estimated values for S, E, I, R^1 and R^2 are acceptable for processing P_3. The optimization method is robust enough to give a good estimate at P_3 whereas it could not at P_2. Those excluded areas are drawn with hatches in following maps.

Let us now comment on the results at the scale of regions and at the scale of departments. In all coming results, we set $\delta = 18$ and the end date for parameter estimation to 2020-06-30.

3.1 Coarse scale: Regions

Figure 4 presents France maps of estimated R_0 for the 3 periods –subfigures (a)-(c)– and the estimation of $d_{I=1}$ –subfigure (d)–, at the scale of regions. Since the maps



(a) Period P_1

(b) Period P_2

(c) Period P_3

(d) Date $d_{I=1}$

Fig. 5 Estimation of R_0 for the 3 periods and first infected individual date at department scale. The right side of each map is a smoothed representation of the histogram of values appearing in the corresponding map. The departments hatched have been excluded according to a threshold explained in the text. For (d), value '0' in the legend corresponds to January 6 (*Var* department, #06), and value '80' corresponds to March 24 (*Lozère* department, #06), 80 days after January 6.

only show the dynamic of the pandemic, not its intensity, we added the estimated number of infected at the end of the period for each region. Obviously, prior to lockdown (a), the values of R_0 all exceeded 1.58 (*Bretagne* region). Then, in the second period (b), the estimates all fell below the threshold of 1.0 (from 0.71 for *Corse* region #94 to 0.93 for both *Provence-Alpes-Côte d'Azur* region #93 and *Pays de la Loire* region #52). Finally, in the third period (c), still according to the proposed estimation method, the values of R_0 remains well below 1.0, except for *Normandie* region #28 with $R_0 = 1.01$. It should be stressed that colours do not give an indication of the intensity of the pandemic in a region, only its dynamics through R_0 rate. The intensity at the end of the period can be assessed through the number of infected (*e.g.* 2886 infected for *Alsace Champagne-Ardenne Lorraine* region, #44, at P_1).

Regarding the date of the first infected individual, Figure (d), the forecasting at the region scale shows dates from January 25 (first infected region: *Provence-Alpes-Côte d'Azur*, #93) to March 3 (last infected region: *Centre Val de Loire*, #24). It is interesting to note that the date of the first infected at the region scale is very close to the estimate produced at the country scale (recall: `2020-01-23`).

3.2 Finer scale: Departments

Likewise regional level, Figure 5 displays France maps representing the estimations of R_0 for the three periods

and the estimated date of the first infected, at the department scale. For period P_1, it can be observed that a number of departments was excluded according to the criterion, mainly in the centre-west part of the country. The distribution of R_0 spreads from 1.4 (*Morbihan* department, #56) to 3.3 (*Eure* department, #27) and presents a mode at about 2.0. For P_2, we can observed that some departments have been excluded according to the criterion specified above. Else, all other departments have a low diffusion rate, from 0.46 to 1.1. The distribution of R_0 values is sharp and presents a mode at about 0.85. The distribution of R_0 values continues to be sharp in P_3, albeit somewhat broader than in the previous period, with a spread from 0.25 to 1.4 and a mode at about 0.85. Most of the south west part of the territory has been excluded (plus some departments in the west coast and in the border with Italy), showing a significant withdrawal from the pandemic. It should be pointed out the particular situation of *Alpes-Maritimes* (#06) and *Haute-Garonne* (#31) for which the pandemic seems to be picking up, with $R_0 = 1.25$ and $R_0 = 1.7$ respectively. For this two departments, the intensity remains relatively small ($I = 63$ and $I = 26$ respectively). It should be noted that the summer holiday period starts in July in France, and that it can reshape this geographical distribution all over the territory.

Regarding the dates of the first infected, we found that the first department was *Var* (#83) as of `2020-01-18`, and the last one was *Eure* (#48) as of `2020-03-28`. This date of January 18 is not so far from the dates we found at France scale (recall: `2020-01-23`), and at the regional scale (recall: `2020-01-25`). This estimation, obtained at the departmental level, is probably the least reliable of the three, since it has been processed with the smallest data set.

4 Conclusion

The R_0 rate (basic reproduction number), is an important indicator for analysing the evolution of a pandemic over time, and for predicting its spread in the population in the near future. In this work, we proposed an estimation method based on the successive fit of the same $SEIR^1R^2$ model, proposed by N. Bacaër (Bacaër, N., 2020), over 3 successive and non-overlapping periods. This approach is intended for countries that have observed a period of lockdown and a period of deconfinement. The 3 periods take into account a delay related to the fact that the impact of public policy measures take some time to be observed in the data.

The methodology proposed is tested on the COVID-19 pandemic spread in France. The three periods are defined by the lockdown date (March 18, 2020) and the deconfinement date (May 11, 2020), and a delay estimated to $\delta = 18$ days. We were able to apply the algorithm at two different geographical scales: regions and departments. Emphasis was put on the graphical presentations of the results to get an overview of France situation as of June 30, 2020. The results of the analysis at these two levels seem to confirm that France have contained the pandemic for the time being. But the summer period that is coming in July and August could favour a local resurgence, in the form of clusters, with a re-breeding of the infected in all the French departments, in particular the coastal regions of the Mediterranean and the Atlantic Ocean. It should be stressed that the methodology is not specific to France, and can be applied to other countries (*e.g.* USA states) or at the scale of (sub)-continent (Latin America and Africa).

In this work, we have decided to deal with a very simple model –which is far from capturing the complexity of the epidemic spread–, with relatively few parameters, so that R_0 only depends on the estimation of a (rate of transmission), and c (rate of recovery). From a methodological point of view, the algorithm developed in this work is not specific to the SEIR model and can be extended to more sophisticated compartmental models (*e.g.* SEIRAH Prague et al. (2020)), which are a priori able to better model the COVID-19 pandemic, by integrating ICU admissions, age and spatial structure of the population... The fitting will therefore relies on all the available data from the national public health databases, which can improve the accuracy, especially for areas with a low density of population. This extension is one of the main perspectives to this work.

References

Bacaër, N (2020) Un modèle mathématique des débuts de l´épidémie de coronavirus en france. Math Model Nat Phenom 15:29, URL `https://doi.org/10.1051/mmnp/2020015`

Brauer F, Castillo-Chavez C, Feng Z (2019) Mathematical Models in Epidemiology. Texts in Applied Mathematics, Springer-Verlag New York

Castillo-Chavez C, Feng Z (1997) To treat or not to treat: the case of tuberculosis. J Math Biol 35:629–656

Diekmann O, Heesterbeek JAP, Metz JAJ (1990) On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. Journal of Mathematical Biology 28(4):365–382

Heffernan JM, Smith RJ, Wahl LM (2005) Perspectives on the basic reproductive ratio. Journal of the Royal Society, Interface 2(4):281–293

Matthews L, Woolhouse ME, Hunter N (1999) The basic reproduction number for scrapie. Proceedings Biological sciences 266(1423):1085–1090

Prague M, Wittkop L, Clairon Q, Dutartre D, Thiebaut R, B-P H (2020) Population modeling of early COVID-19 epidemic dynamics in French regions and estimation of the lockdown impact on infection rate, DOI 10.1101/2020.04.21.20073536, unpublished

van den Driessche P, Watmough J (2002) Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. Math Biosci 180:9–48

van den Driessche P, Watmough J (2008) Further notes on the basic reproduction number. In: Brauer F, van den Driessche P, Wu J (eds) Mathematical Epidemiology, Lecture Notes in Mathematics, Springer Berlin Heidelberg, pp 159–178