Using early data to estimate the actual infection fatality ratio from COVID-19 in France

Lionel Roques[1,*], **Etienne K Klein**[1], **Julien Papaïx**[1], **Antoine Sar**[2], **and Samuel Soubeyrand**[1]

[1]INRAE, BioSP, 84914, Avignon, France
[2]Medicentre Moutier, 2740 Moutier, Suisse
[*]lionel.roques@inrae.fr

ABSTRACT

The first cases of COVID-19 in France were detected on January 24, 2020. The number of screening tests carried out and the methodology used to target the patients tested do not allow for a direct computation of the actual number of cases and the infection fatality ratio (IFR). We develop a 'mechanistic-statistical' approach coupling a SIR ODE model describing the unobserved epidemiological dynamics, a probabilistic model describing the data acquisition process and a statistical inference method. The objective of this model is not to make forecasts but to estimate the actual number of people infected with COVID-19 during the observation window in France and to deduce the IFR associated with the epidemic in France.
Main results. The actual number of infected cases in France is probably higher than the observations: we find here a factor $\times 5$ (95%-CI: 3.5-7.8) which leads to an IFR in France of 0.5% (95%-CI: $0.3-0.8$). Although accurate in France, the SIR model cannot capture the decline in the number of cases observed in South Korea: this points out the strong impact on the epidemic dynamics of the management strategy adopted in South Korea.

1 Introduction

The COVID-19 epidemic started in December 2019 in Hubei province, China. Since then, the disease has spread around the world reaching the pandemic stage, according to the WHO, on March 11. The first cases were detected in South Korea on January 19, 2020, and in France on January 24. The infection fatality ratio (IFR), defined as the number of deaths divided by the number of infected cases, is an important quantity that informs us on the expected number of casualties at the end of an epidemic, when a given proportion of the population has been infected. Although the data on the number of deaths from COVID-19 are probably accurate, the actual number of infected people in the population is not known. Thus, due to the relatively low number of screening tests that have been carried out in France (about 5 over $10,000$ people in France to be compared with 50 over $10,000$ in South Korea up to March 15, 2020; Sources: Santé Publique France and Korean Center for Disease Control) the direct computation of the IFR is not possible. Based on the PCR-confirmed cases in international residents repatriated from China on January 2020 Verity et al. (2020) obtained an estimate of the infection fatality ratio (IFR) of 0.66% in China, and, adjusting for non-uniform attack rates by age, an IFR of 0.9% was obtained in the UK (Ferguson et al., 2020).

Using the early data (up to March 17) available in France and South Korea, our objectives are: (1) to compute the IFR in France, (2) to estimate the number of people infected with COVID-19 in France, (3) to compute a basic reproduction rate R_0, (4) to compare the results obtained for France and South Korea.

2 Methods

Data. We obtained the number of positive cases and deaths in France an South Korea, day by day from Johns Hopkins University Center for Systems Science and Engineering (https://github.com/CSSEGISandData/COVID-19). The data on the number of tests carried out was obtained from Santé publique France, and the Korean Center for Disease Control. As some data (positive cases, deaths, number of tests) are not fully reliable (example: 0 new cases detected in France on March 12, 2020), we smoothed the data with a moving average over 5 days.

Mechanistic-statistical model. This formalism, which is becoming standard in ecology (Abboud et al., 2019; Roques and Bonnefon, 2016; Roques et al., 2011) allows the analyst to couple a mechanistic model, here an ordinary differential equation model (ODE) of the SIR type, and uncertain, non-exhaustive data. To bridge the gap between the mechanistic model and the

data, the approach uses a probabilistic model describing the data collection process. A statistical method is then used for the estimation of the parameters of the mechanistic model.

Mechanistic model. The dynamics of the epidemics are described by the following SIR compartmental model Murray (2002):

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

with S the susceptible population, I the infected population, R the recovered population (immune individuals) and $N = S+I+R$ the total population, supposed to be constant. The parameter α is the infection rate (to be estimated) and $1/\beta$ is the mean time until an infected becomes recovered. Based on the results in Zhou et al. (2020), the median period of viral shedding is 20 days, but the infectiousness tends to decay before the end of this period: the results in He et al. (2020) show 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 initial conditions are $S(t_0) = N-1$, $I(t_0) = 1$ and $R(t_0) = 0$, where N corresponds to the population size in the considered country ($67\,10^6$ in France, $62\,10^6$ in South Korea). The SIR model is started at some time $t = t_0$, which will be estimated and should approach the date of introduction of the virus in the considered country (this point is shortly discussed at the end of this paper). The ODE system (1) is solved thanks to a standard numerical algorithm, using Matlab® *ode45* solver.

Next we denote by $D(t)$ the number of deaths due to the epidemic. Note that the impact of the compartment $D(t)$ on the dynamics of the SIR system and on the total population is neglected here. The dynamics of $D(t)$ depend on $I(t)$ trough de differential equation:

$$D'(t) = \gamma(t) I(t), \tag{2}$$

with $\gamma(t)$ the mortality rate of the infecteds.

Observation model. We suppose that the number of cases tested positive on day t, denoted by $\hat{\delta}_t$, follow independent binomial laws, conditionally on the number of tests n_t carried out on day t, and on p_t the probability of being tested positive in this sample:

$$\hat{\delta}_t \sim Bi(n_t, p_t), \tag{3}$$

The tested population consists of a fraction of the infecteds and a fraction of the susceptibles: $n_t = \tau_1(t) I(t) + \tau_2(t) S(t)$. Thus,

$$p_t = \frac{\sigma\, \tau_1(t) I(t)}{\tau_1(t) I(t) + \tau_2(t) S(t)} = \frac{\sigma I(t)}{I(t) + \kappa_t S(t)},$$

with $\kappa_t := \tau_2(t)/\tau_1(t)$, the relative probability of undergoing a screening test for an individual of type S *vs* an individual of type I (probability of being tested conditionally on being S / probability of being tested conditionally on being I). We assume that the ratio κ does not depend on t at the beginning of the epidemic (i.e., over the period that we use to estimate the parameters of the model). The coefficient σ corresponds to the sensitivity of the test. In most cases, RT-PCR tests have been used and existing data indicate that the sensitivity of this test using pharyngeal and nasal swabs is about $63-72\%$ Wang et al. (2020). We take here $\sigma = 0.7$ (70% sensitivity).

Statistical inference. The data $\hat{\delta}_t$ used to compute the MLE and the posterior distribution are those corresponding to the period from February 29 to March 17. The unknown parameters are α, t_0 and κ. The parameter $\gamma(t)$ is computed indirectly, using the estimated value of $I(t)$, the data on the mortality rate (assumed to be exact) and the relationship (2).

The likelihood $\mathscr{L}$, is defined as the probability of the observations (here, the increments $\{\hat{\delta}_t\}$) conditionally on the parameters. Using the observation model (3), and assuming that the increments $\hat{\delta}_t$ are independent conditionally on the underlying SIR process and that the number of tests n_t is known, we get:

$$\mathscr{L}(\alpha, t_0, \kappa) := P(\{\hat{\delta}_t\} | \alpha, t_0, \kappa) = \prod_{t=t_i}^{t_f} \frac{n_t!}{(\hat{\delta}_t)!(n_t - \hat{\delta}_t)!} p_t^{\hat{\delta}_t} (1-p_t)^{n_t - \hat{\delta}_t},$$

with t_i date of the first observation and t_f the date of the last observation. In this expression $\mathscr{L}(\alpha, t_0, \kappa)$ depends on α, t_0, κ through p_t.

The maximum likelihood estimator (MLE, i.e., the parameters that maximize $\mathscr{L}$), is computed using the BFGS constrained minimization algorithm, applied to $-\ln(\mathscr{L})$, via the Matlab ® function *fmincon*. In order to find a global maximum of $\mathscr{L}$, we apply this method starting from random initial values for α, t_0, κ drawn uniformly in the following intervals:

$$\begin{cases} \alpha \in (0,1), \\ t_0 \in (1,50), \text{ (January 1st - February 19th)} \\ \kappa \in (0,1). \end{cases} \tag{4}$$

For each country, the minimization algorithm is applied to 10000 random initial values of the parameters.

To assess model fit, we compare the observations with expectation of the observation model associated with the MLE, $n_t\, p_t^*$ (expectation of a binomial) with

$$p_t^* = \frac{\sigma I^*(t)}{I^*(t) + \kappa^* S^*(t)},$$

and $I^*(t)$, $S^*(t)$ the solutions of the system (1) associated with the MLE.

The posterior distribution of the parameters (α, t_0, κ) is computed with a Bayesian method, using uniform prior distributions in the intervals given by (4). This posterior distribution corresponds to the distribution of the parameters conditionally on the observations:

$$P(\alpha, t_0, \kappa | \{\hat{\delta}_t\}) = \frac{\mathscr{L}(\alpha, t_0, \kappa)\, \pi(\alpha, t_0, \kappa)}{C},$$

where $\pi(\alpha, t_0, \kappa)$ corresponds to the prior distribution of the parameters (therefore uniform) and C is a normalization constant independent of the parameters. The numerical computation of the posterior distribution (which is only carried out for French data) is performed with a Metropolis-Hastings (MCMC) algorithm, using 4 independent chains, each of which with 10^6 iterations, starting from random values close to the MLE.

Computation of the infection fatality ratio and of R_0. The IFR corresponds to the fraction of the infected who die, that is $\gamma(t)/(\gamma(t)+\beta)$. Given the (estimated) population I, the term $\gamma(t)$ is computed using the formula (2) and the mortality data. With SIR systems of the form (1), the basic reproduction rate R_0 can be computed directly, based on the formula $R_0 = \alpha/\beta$ (Murray, 2002). When $R_0 < 1$, the epidemic cannot spread in the population. When $R_0 > 1$, the infected compartment I increases as long as $R_0 S > N = S + I + R$.

3 Results

Model fit. Fig. 1 compares the expectation of the observation model associated with the MLE with the actual observations. In France, we get a good match between this expectation $n_t\, p_t^*$ and the data. In South Korea, on the other hand, the gap between the data and the model prediction is significant: the SIR model, which leads to an exponential trajectory of I at the beginning of the epidemic, cannot properly render the dynamics.

Infection fatality rate. Using the posterior distribution of the model parameters (described in Appendix, Fig. 3), we can compute the daily distribution of the actual number of infected peoples (see Fig. 4 in Appendix). Using this information we thus obtain, on March 17, an IFR in France of 0.5% (95%-CI: $0.3-0.8$). The estimated distribution of IFR is relatively stable over time. Its is depicted in Fig. 2. Additionally, we get that the average estimated ratio in France between the actual number of infected and observed cases ($I(t)/\Sigma\hat{\delta}_t$, with $\Sigma\hat{\delta}_t$ the sum of the observed infected cases at time t) is 5.3 (95%-CI: 3.5-7.8) over the considered period.

Basic reproduction rate. We computed the marginal posterior distribution of the basic reproduction rate R_0 (Fig. 5 in Appendix). This leads to a mean value of R_0 of 3.2 (95%-CI: 3.1-3.3).

4 Discussion.

On the IFR and the number of infecteds. The actual number of infected individuals in France is probably much higher than the observations (we find here a factor $\times 5$), which leads at a lower mortality rate than that calculated on the basis of the observed cases: we find here an IFR of 0.5%, to be compared with a case fatality rate (CFR, number of deaths over number of diagnosed cases) of 2%. However, if the virus led to contaminate 80% of the French population (Ferguson et al., 2020), the total number of deaths to deplore in the absence of variation in the mortality rate (increase induced for example by the saturation of hospital structures, or decrease linked to better patient care) would be $336,000$ (95%-CI : $192,000-537,000$)). This estimate could

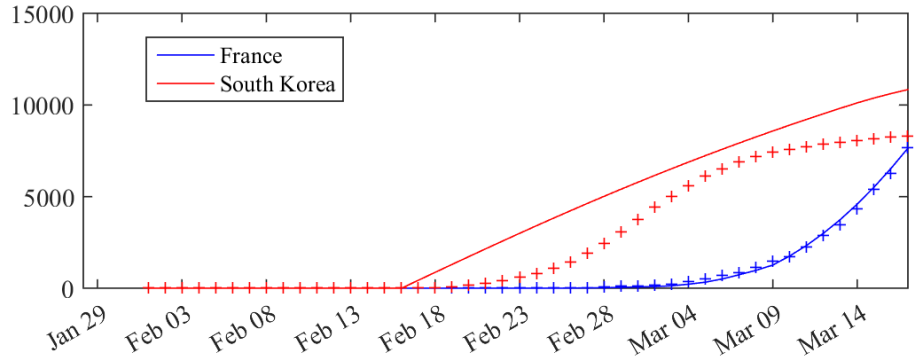


Figure 1. Expected number of observed cases associated with the MLE *vs* number of cases actually detected (total cases). The curves corresponds to the expected observation $n_t\, p_t^*$ given by the model, and the crosses correspond to the data (cumulated values of $\hat{\delta}_t$).

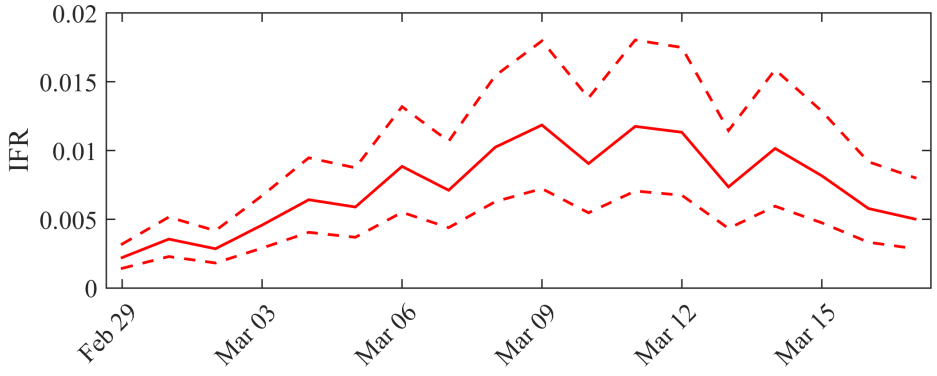


Figure 2. Dynamics of the IFR in France. Solid line: average value obtained from the posterior distribution of the parameters. Dotted curves: 0.025 and 0.975 pointwise quantiles.

be corroborated or invalidated when 80% of the population will be infected, eventually over several years, assuming that an infected individual is definitively immunized. It has to be noted that measures of confinement or social distancing can decrease both the percentage of infected individuals in the population and the degree of saturation of hospital structures. Note that an IFR of 0.5% (95%-CI: $0.3-0.8$) is consistent with the findings of Verity et al. (2020) (0.66% in China), and slightly lower than the value of 0.9% previously obtained in the UK (Ferguson et al., 2020).

On the differences between France and South Korea. The mechanistic-statistical SIR model achieves a satisfactory goodness-of-fit for French data, but does not capture the decline in the number of cases observed in South Korea. The difference between the dynamics predicted by the SIR model and the South Korean data is probably linked to a different management of the epidemic in Korea, having a strong impact on the epidemic dynamics (more important screening, tracing, social distancing in South Korea).

On the value of R_0. The estimated distribution in France is high compared to recent estimates (2.0-2.6, see Ferguson et al. (2020)). This difference could be due to a different definition of R_0 depending on the type of model used to calculate it. A direct estimate, by a non-mechanistic method, of the parameters (ρ, t_0) of a model of the form $\hat{\delta}_t = e^{\rho\,(t-t_0)}$ gives $t_0 = 36$ (February 5) and $\rho = 0.22$. With the SIR model, $I'(t) \approx I(\alpha - \beta)$ for small times ($S \approx N$), which leads to a growth rate equal to $\rho \approx \alpha - \beta$, and a value of $\alpha \approx 0.32$, that is to say $R_0 = 3.2$, which is consistent with the distribution presented in Fig. 5. Note that we have assumed here a infectiousness period of 10 days. A shorter period would lead to a lower value of R_0.

On the uncertainty linked to the data. The uncertainty on the actual number of infected and therefore the mortality rate are very high. We must therefore interpret with caution the inferences that can be made based on the data we currently have in France. In addition, we do not draw forecasts here: the future dynamics will be strongly influenced by the containment measures that will be taken and should be modeled accordingly.

On the hypotheses underlying the model. The data used here contain a limited amount of information, especially since the observation period considered is short and corresponds to the initial phase of the epidemic dynamics, which can be strongly influenced by discrete events. This limit led us to use a particularly parsimonious model in order to avoid problems of identifiability for the parameters. The assumptions underlying the model are therefore relatively simple and the results must be interpreted with regard to these assumptions. For instance, the date of the introduction t_0 must be seen as an *efficient* date of introduction for a dynamics where a single introduction would be decisive for the outbreak and the other (anterior and posterior) introductions would have an insignificant effect on the dynamics.

Acknowledgements

This work was funded by INRAE.

Author contributions statement

L.R., E.K.K., J.P. and S.S. conceived the model. L.R. and S.S. wrote the paper, L.R. carried out the numerical computations. All authors reviewed the manuscript.

Competing interests

The authors declare no competing financial interests.

Appendix

The joint posterior distributions of the three pairs of parameters (α, κ), (t_0, α) and (t_0, κ) are depicted in Fig. 3. The distribution of the actual number of infected cases obtained from the posterior distribution of the parameters is depicted in Fig. 4. The marginal posterior distribution of R_0 is depicted in Fig. 5.

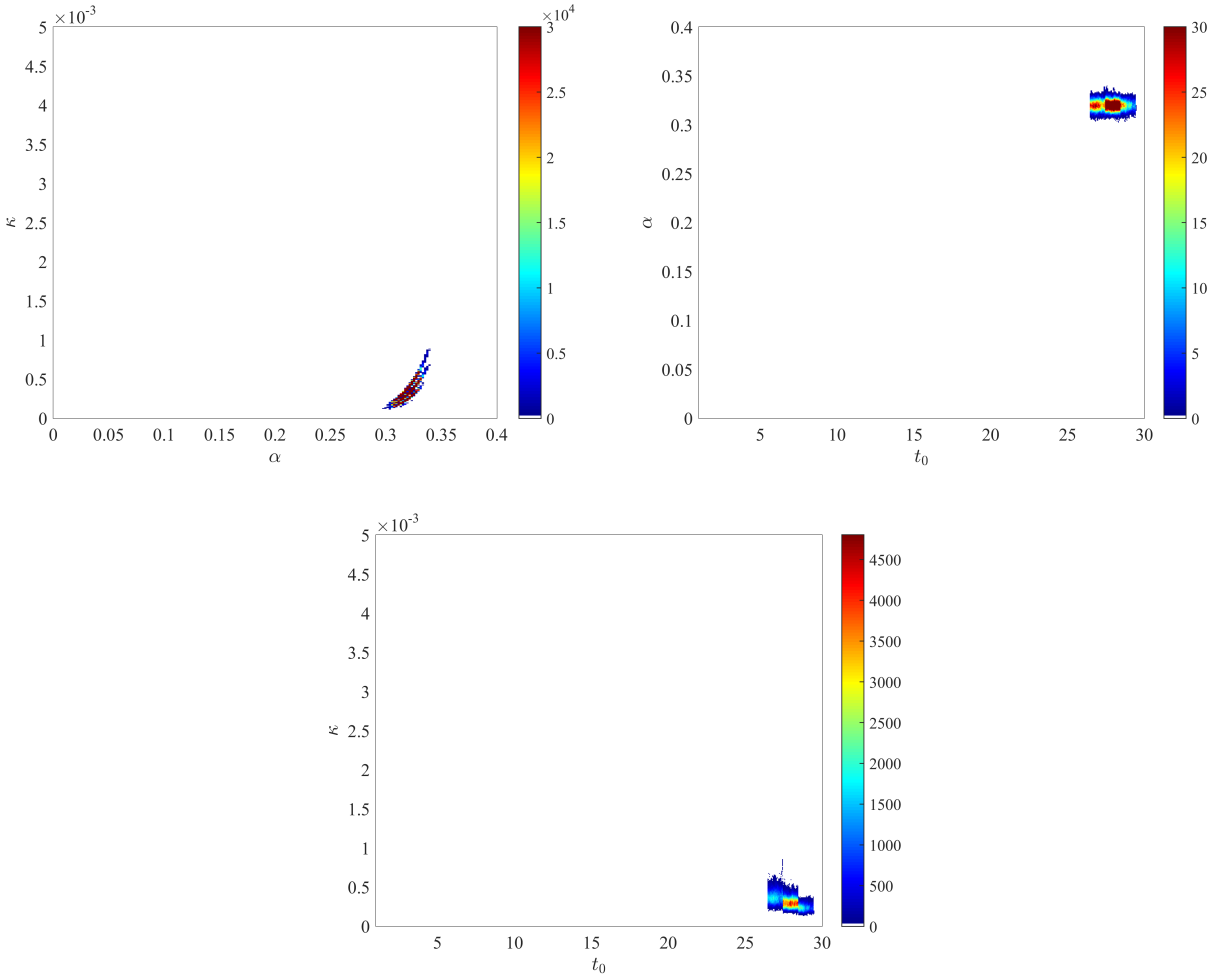


Figure 3. Joint posterior distributions of (α, κ), (t_0, α) and (t_0, κ) in France.

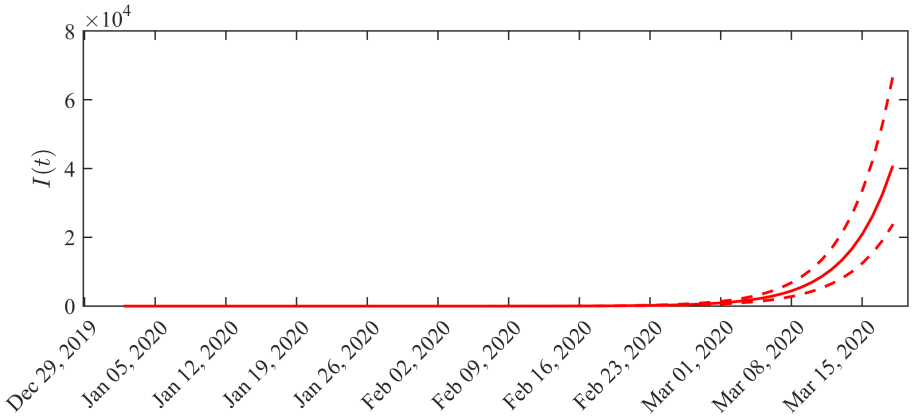
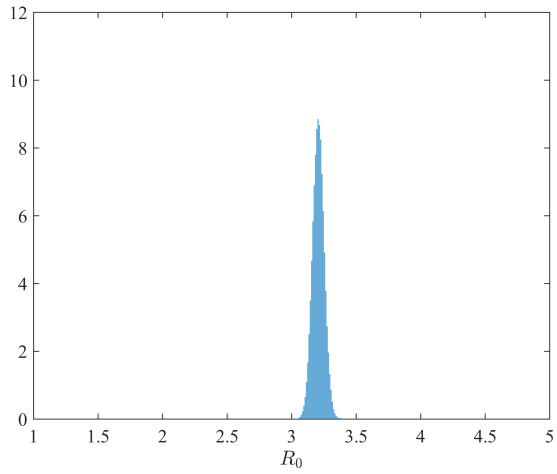


Figure 4. Distribution of the actual number of infected cases in France across time. Solid line: average value obtained from the posterior distribution of the parameters. Dotted curves: 0.025 and 0.975 pointwise posterior quantiles.



(a) France

Figure 5. Posterior distribution of the basic reproduction rate R_0 in France.

References

Abboud, C., O. Bonnefon, E. Parent, and S. Soubeyrand (2019). Dating and localizing an invasion from post-introduction data and a coupled reaction–diffusion–absorption model. *Journal of mathematical biology 79*(2), 765–789.

Ferguson, N. M., D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, et al. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. *Imperial College, London*. DOI: https://doi.org/10.25561/77482.

He, X., E. H. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. Lau, J. Y. Wong, Y. Guan, X. Tan, et al. (2020). Temporal dynamics in viral shedding and transmissibility of covid-19. *medRxiv*. DOI: https://doi.org/10.1101/2020.03.15.20036707.

Murray, J. D. (2002). *Mathematical Biology*. Third edition, Interdisciplinary Applied Mathematics 17, Springer-Verlag, New York.

Roques, L. and O. Bonnefon (2016). Modelling population dynamics in realistic landscapes with linear elements: A mechanistic-statistical reaction-diffusion approach. *PloS one 11*(3), e0151217.

Roques, L., S. Soubeyrand, and J. Rousselet (2011). A statistical-reaction-diffusion approach for analyzing expansion processes. *J Theor Biol 274*, 43–51.

Verity, R., L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. Walker, H. Fu, et al. (2020). Estimates of the severity of covid-19 disease. *medRxiv*. DOI: https://doi.org/10.1101/2020.03.09.20033357.

Wang, W., Y. Xu, R. Gao, R. Lu, K. Han, G. Wu, and W. Tan (2020). Detection of SARS-CoV-2 in different types of clinical specimens. *Jama*.

Zhou, F., T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*.