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

Lionel Roques [a,*], Etienne Klein [a], Julien Papaïx [a], Antoine Sar [b] and Samuel Soubeyrand [a]

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

[b] Medicentre Moutier, 2740 Moutier, Switzerland

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

Abstract

Background. The number of screening tests carried out in France 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). The main objective of this work is to estimate the actual number of people infected with COVID-19 and to deduce the IFR during the observation window in France.

Methods. We develop a 'mechanistic-statistical' approach coupling a SIR epidemiological model describing the unobserved epidemiological dynamics, a probabilistic model describing the data acquisition process and a statistical inference method.

Results. The actual number of infected cases in France is probably higher than the observations: we find here a factor $\times 8$ (95%-CI: 5-12) which leads to an IFR in France of 0.5% (95%-CI: $0.3-0.8$) based on hospital death counting data. Adjusting for the number of deaths in nursing homes, we obtain an IFR of 0.8% (95%-CI: $0.45-1.25$).

Conclusions. This IFR is consistent with previous findings in China (0.66%) and in the UK (0.9%) and lower than the value previously computed on the Diamond Princess cruse ship data (1.3%).

Keywords. COVID-19; infection fatality ratio; case fatality rate; SIR model; mechanistic-statistical model; Bayesian inference;

1 Background

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 [1], on March 11. The first cases were detected 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, [2] 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 [3]. Using data from the quarantined Diamond Princess cruise ship in Japan and correcting for delays between confirmation and death [4] obtained an IFR of 1.3%.

Using the early data (up to March 17) available in France, 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.

2 Methods

Data. We obtained the number of positive cases and deaths in France, day by day from Johns Hopkins University Center for Systems Science and Engineering [5]. The data on the number of tests carried out was obtained from Santé Publique France [6]. 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. Official data on the number of deaths by COVID-19 in France only take into account hospitalized people. About 728,000 people in France live in nursing homes (EHPAD, source: DREES [7]). Recent data in the Grand Est region (source: Agence Régionale de Santé Grand Est [8]), report a total of 570 deaths in these nursing homes, which have to be added to the official count (1015 deaths on March 31).

Mechanistic-statistical model. The mechanistic-statistical formalism, which is becoming standard in ecology [9, 10, 11] allows the analyst to couple a mechanistic model that describes a latent variable, 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 epidemic are described by the following SIR compartmental model:

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

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 [12], the median period of viral shedding is 20 days, but the infectiousness tends to decay before the end of this period: the results in [13] 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 $N = 67\,10^6$ corresponds to the population size. 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 France (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 $D(t)$ depends on $I(t)$ trough the 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 infected cases 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 $63 - 72\%$ [14]. We take here $\sigma = 0.7$ (70% sensitivity).

Statistical inference. 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 $D(t)$ (assumed to be exact) and the relationship (2). The likelihood $\mathcal{L}$ 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:

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

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

The maximum likelihood estimator (MLE, i.e., the parameters that maximize $\mathcal{L}$), is computed using the BFGS constrained minimization algorithm, applied to $-\ln(\mathcal{L})$, via the Matlab ® function *fmincon*. In order to find a global maximum of $\mathcal{L}$, we apply this method starting from random initial values for α, t_0, κ drawn uniformly in the following intervals: $\alpha \in (0,1)$, $t_0 \in (1,50)$, (January 1st - February 19th) and $\kappa \in (0,1)$. The minimization algorithm is applied to 10000 random initial values of the parameters.

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

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

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 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.

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.

3 Results

Model fit. To assess model fit, we compared the observations, i.e., the cumulated number of deaths $\Sigma_t := \Sigma_{i=1,\dots,t}\hat{\delta}_i$, with the 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 results are presented in Fig. 1. We observe a good match between the expectation $n_t\, p_t^*$ and the data.

Infection fatality ratio and actual number of infected cases. Using the posterior distribution of the model parameters (the pairwise distributions are presented in Appendix, see Fig. A1), we computed the daily distribution of the actual number of infected peoples. Using the relation (2) together with the data on $D(t) = \Sigma_t$, we deduce the distribution of the parameter $\gamma(t)$, at each date. The IFR corresponds to the fraction of the infected who die, that is:

$$IFR_t := \gamma(t)/(\gamma(t) + \beta).$$

We thus obtain, on March 17 an IFR of 0.5% (95%-CI: $0.3 - 0.8$), and the distribution of the IFR is relatively stable over time (see Fig. A2 in the Appendix).

Additionally, the distribution of the cumulated number of infected cases $(I(t) + R(t))$ across time is presented in Fig. 2. We observe that it is much higher than the total number of observed cases (compare with Fig. 1). The average estimated ratio between the actual number of individuals that have been infected and observed cases $(I(t) + R(t))/\Sigma_t$ is 8 (95%-CI: 5-12) over the considered period.

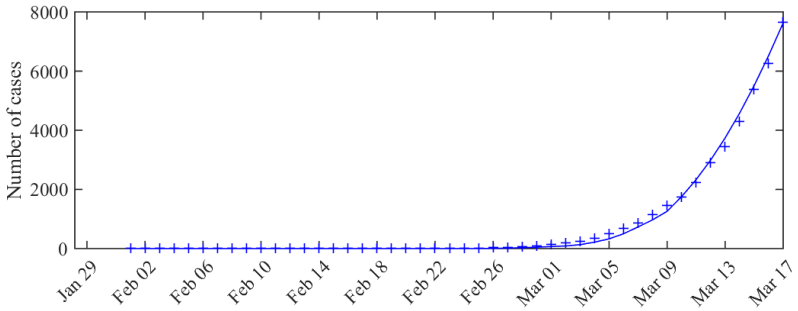

Figure 1: **Expected number of observed cases associated with the MLE** *vs* **number of cases actually detected (total cases).** The curves corresponds to the expected observation $n_t\, p_t^*$ given by the model, and the crosses correspond to the data (cumulated values of $\hat{\delta}_t$).

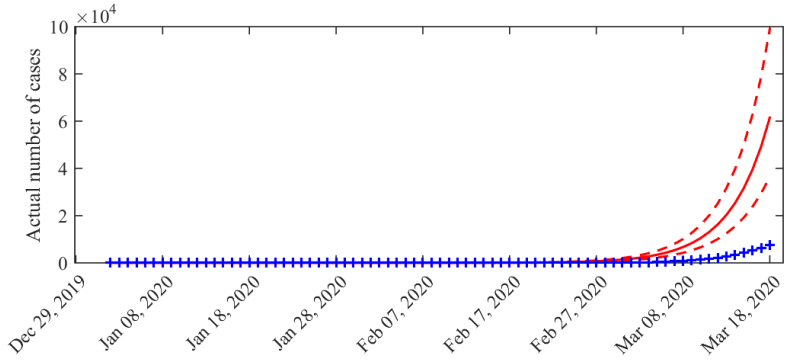

Figure 2: **Distribution of the cumulated number of infected cases ($I(t) + R(t)$) across time.** Solid line: average value obtained from the posterior distribution of the parameters. Dotted curves: 0.025 and 0.975 pointwise posterior quantiles. Blue crosses: data (cumulated values of $\hat{\delta}_t$).

Taking into account the data in the nursing homes. The above computation of the IFR is based on the official counting of deaths by COVID-19 in France, which does not take into account the number of deaths in nursing homes. Based on the local data in Grand Est region, we infer that the IFR that we computed has been underestimated by a factor about $(1015 + 570)/1015 \approx 1.6$, leading to an adjusted IFR of 0.8% (95%-CI: $0.45 - 1.25$).

Basic reproduction rate. With SIR systems of the form (1), the basic reproduction rate R_0 can be computed directly, based on the formula $R_0 = \alpha/\beta$ [15]. When $R_0 < 1$, the epidemic cannot spread in the population. When $R_0 > 1$, the infected compartment I increases as long as $R_0\, S > N = S + I + R$. We computed the marginal posterior distribution of the basic reproduction rate R_0. This leads to a mean value of R_0 of 3.2 (95%-CI: 3.1-3.3). The full distribution is available in the Appendix (Fig. A3).

4 Discussion

On the IFR and the number of infected cases. The actual number of infected individuals in France is probably much higher than the observations (we find here a factor $\times 8$), which leads at a lower mortality rate than that calculated on the basis of the observed cases: we found here an IFR of 0.5% based on hospital death counting data, to be compared with a case fatality rate (CFR, number of deaths over number of diagnosed cases) of 2% on March 17. Adjusting for the number of deaths in the nursing homes, we obtained an IFR of 0.8%. These values for the IFR are consistent with the findings of [2] (0.66% in China) and [3] (0.9% in the UK). The value of 1.3% estimated on the Diamond Princess cruise ship [4] falls above the top end of our 95% CI. This reflects the age distribution on the ship, which was skewed towards older individuals (mean age: 58 years), among whom the IFR is higher [3, 4].

If the virus led to contaminate 80% of the French population [3], 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 : $192,000 - 537,000$), excluding the number of deaths in the nursing homes. This estimate could be corroborated or invalidated when 80% of the population will be infected, eventually over several years, assuming that an infected individual is definitively immunised. 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.

On the value of R_0. The estimated distribution in France is high compared to recent estimates (2.0-2.6, see [3]) but consistent with the findings in [16] (2.24-3.58). 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 our distribution of R_0. 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 IFR 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.

A more complex epidemiological model of the COVID-19 epidemic in China has been proposed in [17], with an infectious class divided into several compartments (asymptomatic individuals, unobserved symptomatic infectious and observed symptomatic infectious). The authors use this model in [18] to make forecasts on the cumulative number of cases in China, while taking into account management strategies. In these two studies the authors emphasise the importance of being able to estimate the fraction of infectious cases that are not observed in order to forecast the dynamics of the epidemic. Our study, though based on a simpler SIR model, shows that this fraction can be estimated based on early data.

Acknowledgements

This work was funded by INRAE: MEDIA Network.

Author contributions statement

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

Competing interests

All authors report no conflict of interest relevant to this article.

Appendix

- The joint posterior distributions of the three pairs of parameters (α, κ), (t_0, α) and (t_0, κ) are depicted in Fig. A1.
- The dynamics of the estimated distribution of the IFR are depicted in Fig. A2.
- The marginal posterior distribution of R_0 is depicted in Fig. A3.

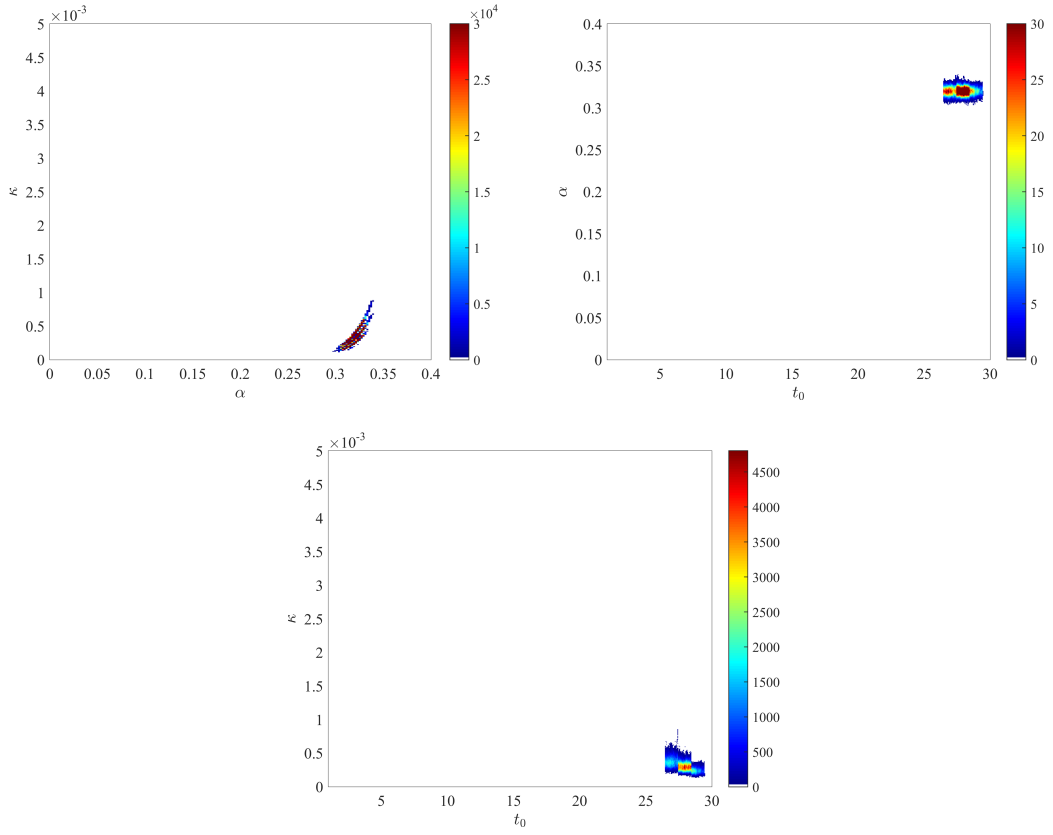

Figure A1: **Joint posterior distributions of** (α, κ)**,** (t_0, α) **and** (t_0, κ)**.**

References

[1] World Health Organization. Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020.

[2] R Verity, L C Okell, I Dorigatti, P Winskill, C Whittaker, N Imai, G Cuomo-Dannenburg, H Thompson, P Walker, H Fu, et al. Estimates of the severity of COVID-19 disease. *medRxiv*, 2020.

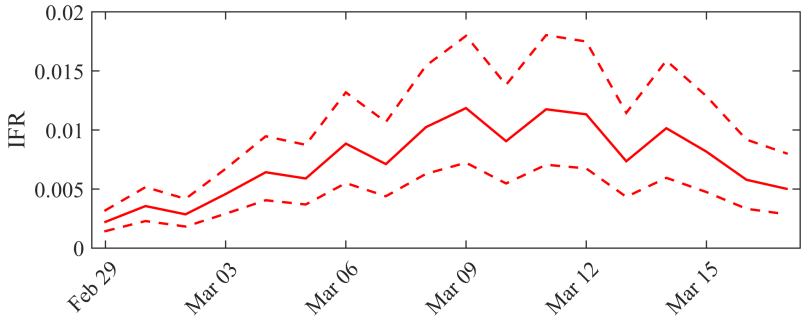

Figure A2: **Dynamics of the IFR in France.** Solid line: average value obtained from the posterior distribution of the parameters. Dotted curves: 0.025 and 0.975 pointwise quantiles.

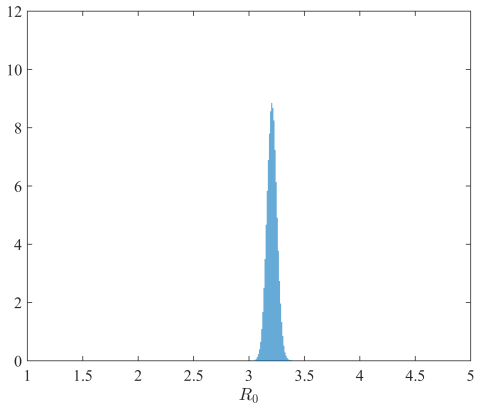

Figure A3: **Posterior distribution of the basic reproduction rate R_0 in France.**

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

[4] T W Russell, J Hellewell, C I Jarvis, K van Zandvoort, S Abbott, R Ratnayake, S Flasche, R M Eggo, W J Edmunds, A J Kucharski, et al. Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Eurosurveillance*, 25(12), 2020.

[5] E Dong, H Du, and L Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 2020.

[6] Santé Pulique France. COVID-19 : point épidémiologique du 24 mars 2020.

[7] DREES. 728 000 résidents en établissements d'hébergement pour personnes âgées en 2015. *https://drees.solidarites-sante.gouv.fr/IMG/pdf/er1015.pdf*, 2020.

[8] Agence Régionale de Santé Grand Est. Dossier de presse - covid 19 : point de situation dans le grand est, 2 avril 2020.

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

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

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

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

[13] X He, E HY Lau, P Wu, X Deng, J Wang, X Hao, Y Lau, J Y Wong, Y Guan, X Tan, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *medRxiv*, 2020.

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

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

[16] S Zhao, Q Lin, J Ran, S S Musa, G Yang, W Wang, Y Lou, D Gao, L Yang, Daihai He, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases*, 92:214–217, 2020.

[17] Z Liu, P Magal, O Seydi, and G Webb. Understanding unreported cases in the 2019-nCov epidemic outbreak in Wuhan, China, and the importance of major public health interventions. *MPDI Biology*, 9(3):50, 2020.

[18] Z Liu, P Magal, O Seydi, and G Webb. Predicting the cumulative number of cases for the COVID-19 epidemic in China from early data. *Mathematical Biosciences and Engineering*, 2020.