Modèle SIR mécanistico-statistique pour l'estimation du nombre d'infectés et du taux de mortalité par COVID-19

Lionel Roques, Etienne Klein, Julien Papaïx, Samuel Soubeyrand

Modèle SIR mécanistico-statistique pour l'estimation du nombre d'infectés et du taux de mortalité par COVID-19

Lionel Roques, Etienne Klein, Julien Papaïx et Samuel Soubeyrand
INRAE, BioSP, 84914, Avignon, France
Contact : lionel.roques@inrae.fr

Résumé

Les premiers cas de COVID-19 ont été détectés en France le 24 janvier 2020. Le nombre de tests de dépistage effectués et la méthodologie employée pour cibler les patients testés ne permettent pas de connaître avec certitude le nombre réel d'infectés et le taux de mortalité liés à l'épidémie. Nous développons dans cette note une approche dite 'mécanistico-statistique' couplant un modèle d'équations différentielles de type SIR décrivant la dynamique épidémiologique non observée, un modèle probabiliste décrivant le processus de recensement des infectés et une méthode statistique d'inférence. L'objectif de ce modèle n'est pas de faire de la prédiction mais d'estimer le nombre réel de personnes infectées par le COVID-19 en France durant la période d'observation et d'en déduire le taux de mortalité associé à l'épidémie.
Principaux résultats. Nous trouvons que le nombre réel d'infectés en France est bien supérieur aux observations, avec un facteur $\times 15$ (IC 95% 4-33), et que le taux de mortalité au 17 mars est de 5.2/1000 (IC 95% $1.5/1000 - 11.7/1000$). Nous trouvons un R_0 de 4.8, valeur élevée liée en partie à la période d'excrétion virale que nous supposons de 20 jours.

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 real number of cases and the mortality rate. In this report, 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 real number of people infected with COVID-19 during the observation window in France and to deduce the mortality rate associated with the epidemic.
Main results. The actual number of infected cases in France is probably much higher than the observations : we find here a factor $\times 15$ (95%-CI : $4 - 33$), which leads to a 5.2/1000 mortality rate (95%-CI : $1.5/1000 - 11.7/1000$) at the end of the observation period. We find a R_0 of 4.8, a high value which may be linked to the long viral shedding period of 20 days.

Introduction. L'épidémie de COVID-19 a démarré en décembre 2019 dans la province du Hubei, en Chine. Depuis, la maladie s'est propagée à travers le monde, avec notamment des premiers cas détectés en France le 24 janvier 2020, pour atteindre le stade de pandémie le 11 mars selon l'OMS. Le nombre de tests de dépistage effectués est très variable suivant les pays (36 747 en France vs 268 212 en Corée du Sud au 15 mars 2020, Sources : Santé Publique France et Korean Center for Disease Control) et ne permet pas de connaître avec certitude le nombre réel d'infectés dans la

population. Le nombre de décès liés au COVID-19 est connu avec plus de certitude ; néanmoins, le nombre de malades n'étant pas connu, il ne permet pas de calculer directement un taux de mortalité. En utilisant les données disponibles en France et en Corée du Sud, nos objectifs sont :

- d'estimer le nombre de personnes infectées par le COVID-19 en France ;
- de déduire de ce nombre le taux de mortalité associé ;
- de calculer les paramètres d'un modèle de type SIR associés à l'épidémie en France ;
- de comparer les résultats en France et en Corée du Sud.

Pour cela, nous nous basons sur un formalisme mécanistico-statistique. Ce formalisme permet de coupler un modèle mécaniste, ici un modèle d'équation différentielle ordinaire (EDO) de type SIR, et des données incertaines, non exhaustives et non nécessairement commensurables avec les solutions de l'EDO. Ce formalisme, que nous avons popularisé en l'appliquant à des invasions biologiques (Roques et al., 2011; Roques et Bonnefon, 2016; Abboud et al., 2019), repose sur un couplage entre (1) le modèle mécaniste, (2) un modèle probabiliste décrivant le processus de collecte des données conditionnellement à la solution du modèle mécaniste et (3) une méthode statistique d'estimation des paramètres du modèle mécaniste.

Données. Nous disposons de données de dépistages du COVID-19 en France et en Corée du Sud sur une période allant du 22 janvier 2020 au 17 mars 2020. Ces données décrivent le nombre de cas positifs et de décès, jour par jour (source : Johns Hopkins University Center for Systems Science and Engineering, https ://github.com/CSSEGISandData/COVID-19). Le nombre de tests effectués, qui lui n'est connu qu'à partir du 22 février (Sources : Santé publique France et Korean Center for Disease Control). Certaines données (cas positifs, décès) n'étant pas fiables (exemple : 0 nouveaux cas détectés en France le 12 mars 2020), nous avons procédé à un lissage des données via une moyenne mobile sur 5 jours.

Modèle mécaniste. Les modèles SIR sont les modèles d'EDO (équations différentielles ordinaires) les plus classiques en épidémiologie. Ce sont des modèles dits compartimentaux, qui divisent la population en plusieurs classes : les susceptibles, les infectés et les résistants (immunisés *recovered* en anglais), d'où le nom de modèle SIR. L'exemple le plus simple ne tient pas compte de la démographie des S :

$$\begin{cases} S' = -\dfrac{\alpha}{N} S\, I, \\ I' = \dfrac{\alpha}{N} S\, I - \beta\, I, \\ R' = \beta\, I, \end{cases} \tag{1}$$

avec $N = S + I + R$ la population totale, qui reste constante au cours du temps. On néglige donc ici l'impact du compartiment D (nombre de morts) sur la dynamique du système SIR. Ce compartiment D vérifie :

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

équation qui nous permettra de calculer le taux de mortalité. La donnée initiale $N - 1 = S(t_0)$ est la population totale Française ou Sud-Coréenne (respectivement $67 \cdot 10^6$ et $52 \cdot 10^6$ habitants), $I(t_0) = 1$, $R(t_0) = 0$. Le temps t_0 correspond au démarrage du modèle SIR, et devrait approcher la date d'introduction de l'épidémie.

On note que $I'(t) = \beta\, I\, (R_0\, S/N - 1)$, avec $R_0 = \alpha/\beta$ le taux de reproduction de base (Murray, 2002). Si $R_0 < 1$, on voit que $I' < 0$, donc l'épidémie ne peut se développer. Si $R_0 > 1$, le nombre d'infectés croît tant que $R_0\, S > N = S + I + R$.

Le modèle (1) peut être résolu analytiquement, via un changement de variable de temps, impliquant une intégration numérique. Nous lui préférons ici une résolution numérique standard, via le solveur Matlab® *ode23s*.

Modèle d'observation. Notons $\hat{\delta}_t$ le nombre de cas testés positifs le jour t. On suppose que ces incréments suivent une loi binomiale, conditionnellement au nombre de tests et à $I(t)$, $S(t)$:

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

où n_t correspond au nombre de tests effectués le jour t et p_t la probabilité d'être testé positif dans cet échantillon. La population testée est constituée d'une fraction des infectés et d'une fraction des sains : $n_t = \tau_1(t)\, I(t) + \tau_2(t)\, S(t)$. Ainsi,

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

avec $\kappa := \tau_2(t)/\tau_1(t)$, la probabilité relative de subir un test pour un individu de type S vs un individu de type I (probabilité d'être testé conditionnellement au fait d'être S/probabilité d'être testé conditionnellement au fait d'être I). Nous faisons l'hypothèse que le ratio κ ne dépend pas de t au début de l'épidémie c'est-à-dire sur la période que nous utilisons pour estimer les paramètres du modèle). Le nombre journalier de morts causées par le pathogène considéré est supposé connu de façon exacte.

Inférence. En se basant sur Zhou et al. (2020) (période médiane d'excrétion virale de 20 jours), on fixe $\beta = 1/20$. Les paramètres restant à estimer sont α, la date d'introduction t_0, et κ. En supposant les incréments $\hat{\delta}_t$ indépendants conditionnellement au processus, et n_t connu, la vraisemblance $\mathcal{L}$ associée aux paramètres correspond à la probabilité d'obtenir les observations (ici la famille $\{\hat{\delta}_t\}$) conditionnellement aux paramètres. En utilisant le modèle (3), nous obtenons :

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

avec t_i la date de la première observation et t_f la date de la dernière observation. Dans l'expression ci-dessus, la dépendance à α, t_0, κ se fait via p_t.

Pour calculer l'estimateur de maximum de vraisemblance (i.e., les paramètres qui maximisent $\mathcal{L}$), nous utilisons une méthode de minimisation sous contrainte de type BFGS, appliquée à $-\ln(\mathcal{L})$, via l'outil Matlab® *fmincon*. Afin de trouver un maximum global de $\mathcal{L}$, nous appliquons cette méthode à partir de valeurs initiales de α, t_0, κ tirées aléatoirement (uniformément) dans les intervalles suivants :

$$\begin{cases} \alpha \in (0, 1), \\ t_0 \in (1, 30), \text{ (du 1er au 30 janvier)} \\ \kappa \in (0, 1). \end{cases} \tag{4}$$

Pour chaque pays, l'algorithme de minimisation est appliqué à 2000 valeurs initiales des paramètres.

La distribution *a posteriori* des paramètres (α, t_0, κ) est calculée avec une méthode bayesienne, en utilisant des distributions *a priori* uniformes dans les intervalles (4). Cette distribution *a posteriori* correspond à la distribution des paramètres conditionnellement aux observations :

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

où $\pi(\alpha, t_0, \kappa)$ correspond à la distribution *a priori* des paramètres (donc uniforme) et C est une constante de normalisation indépendante des paramètres. Le calcul numérique de la distribution *a posteriori* (uniquement en France) est effectué avec un algorithme de Metropolis-Hastings (MCMC), en utilisant 5 chaînes indépendantes, avec chacune 10^6 itérations, et une valeur de départ proche du MLE.

Sauf mention contraire, les données $\hat{\delta}_t$ utilisées pour calculer le MLE et la distribution *a posteriori* sont celles correspondant à la période allant du 29 février au 17 mars.

Résultats.
Adéquation aux données. On note α^*, t_0^*, κ^* l'estimateur du maximum de vraisemblance (MLE), et $I^*(t)$, $S^*(t)$ les solutions du système (1) associées à ces valeurs. En France, nous obtenons $(\alpha^*, t_0^*, \kappa^*) = (0.24, 26, 2{\cdot}10^{-4})$. L'espérance des observations associées à ce MLE est $n_t\, p_t^*$ (espérance d'une binomiale) avec

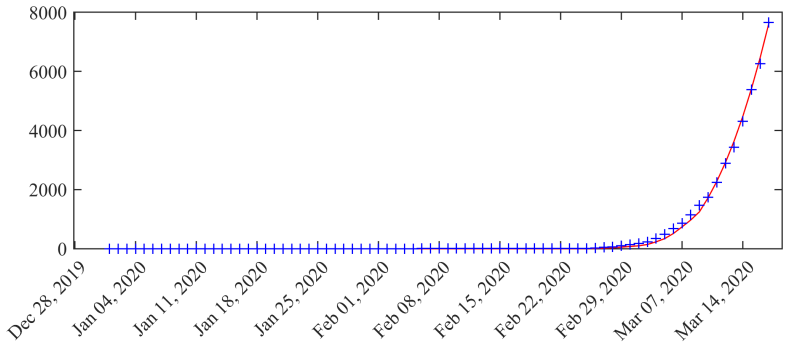
$$p_t^* = \frac{I^*(t)}{I^*(t) + \kappa^*\, S^*(t)}.$$

La Fig. 1 compare cette espérance avec les observations. En France, nous obtenons une bonne adéquation entre $n_t\, p_t^*$ et les données. En Corée du Sud, en revanche, l'écart aux données est important : le modèle SIR, qui conduit à une trajectoire exponentielle de I au début de l'épidémie, ne permet pas de décrire la dynamique. En utilisant des données obtenues à un stade plus précoce en Corée du Sud, l'adéquation aux données est meilleure (Fig. 2). Le MLE correspondant est : $(\alpha^*, t_0^*, \kappa^*) = (0.13, 3, 3 \cdot 10^{-5})$.

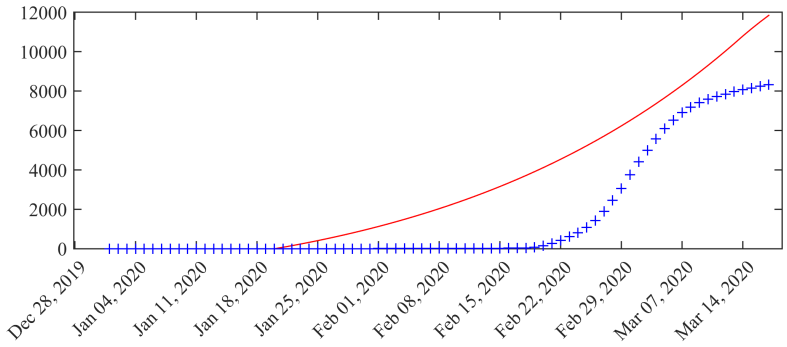
Distribution des paramètres. Les distributions jointes des trois couples de paramètres (α, κ), (t_0, α) et (t_0, κ) en France sont présentées dans l'Appendice A (Fig. 6). On note que des distributions très différentes de la distribution *a priori* uniforme. Néanmoins, les distributions de t_0 et κ sont assez étalées. La distribution jointe de (t_0, κ), présentée dans l'Appendice A montre une corrélation entre t_0 et κ. Ainsi, en supposant t_0 compris entre le 13 et le 30 janvier, nous diminuons l'incertitude sur κ.

La Fig. 3 présente la distribution *a posteriori* marginale du taux de reproduction de base R_0. La valeur de R_0 correspondant au MLE en France est $R_0^* = \alpha^*/\beta = 4.8$. Un calcul similaire en Corée du Sud, sur la base des données utilisées dans la Fig. 2 donne $R_0^* = 2.6$.

Nombre réel d'infectés. En utilisant la distribution *a posteriori* des paramètres du modèle, avec la contrainte 't_0 compris entre le 13 et le 30 janvier' nous en déduisons une distribution des infectés. Cette distribution est représentée en Fig. 4. Nous en déduisons les ratios suivants entre le nombre d'infectés réels et observations, $I(t)/\Sigma\hat{\delta}_t$ (avec $\Sigma\hat{\delta}_t$ le cumul des infectés observés au temps t). Ainsi, en France, le rapport entre nombre d'infectés réels et observations est de 15 (intervalle à 95% : $(4, 33)$).

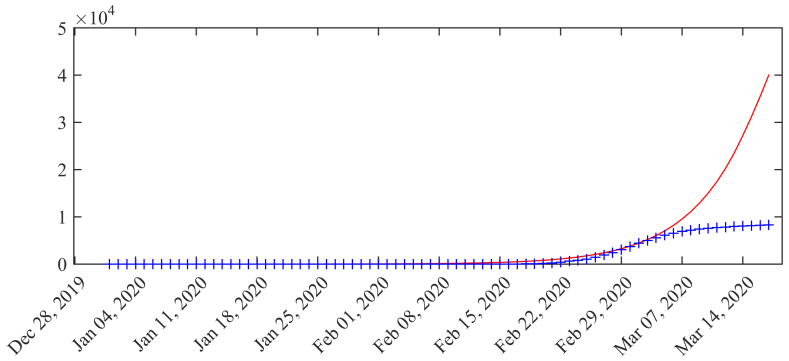


(a) France



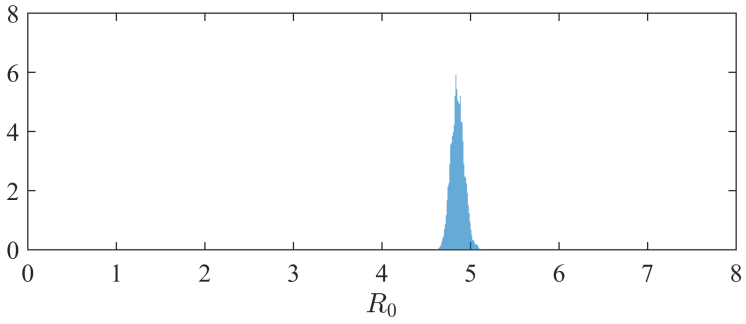
(b) Corée du Sud

FIGURE 1 – **Espérance du nombre de cas détectés associés au MLE vs nombre de cas réellement détectés (total des cas).** La courbe rouge correspond à l'espérance $n_t\, p_t^*$, les croix bleues aux données (cumul des $\hat{\delta}_t$). Calcul du MLE basé sur les données du 29 février au 17 mars.



(a) Corée du Sud

FIGURE 2 – **Espérance du nombre de cas détéctés associés au MLE en Corée du Sud.** La courbe rouge correspond à l'espérance $n_t\, p_t^*$, les croix bleues aux données (cumul des $\hat{\delta}_t$). Calcul du MLE basé sur les données du 18 février au 4 mars.



(a) France

FIGURE 3 – **Distribution *a posteriori* du taux de reproduction de base R_0 en France.**

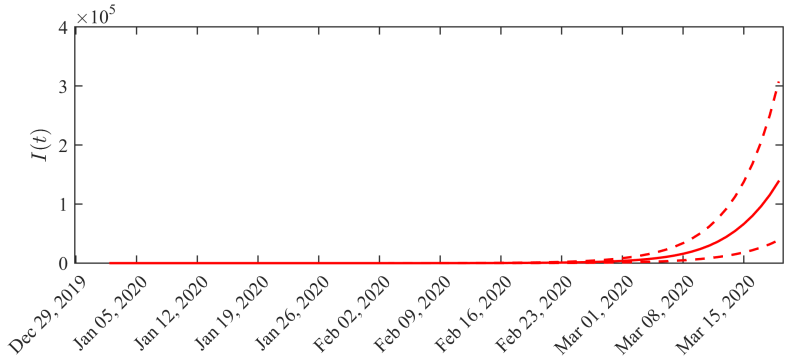


FIGURE 4 – **Distribution du nombre d'infectés en France.** Ligne pleine : valeur moyenne obtenue à partir de la distribution *a posteriori* des paramètres. Courbes pointillées : quantiles 0.025 et 0.975.

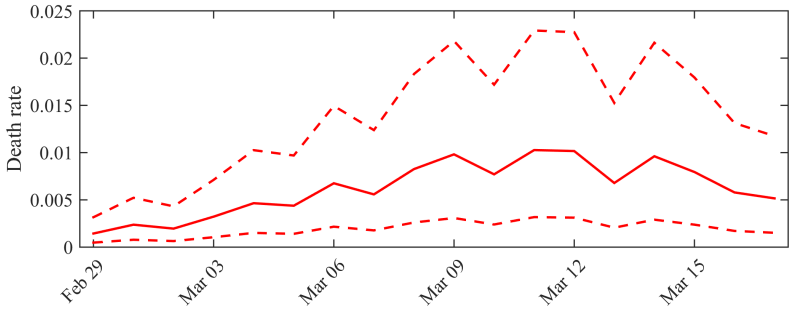


FIGURE 5 – **Evolution du taux de mortalité en France.** Ligne pleine : valeur moyenne obtenue à partir de la distribution *a posteriori* des paramètres. Courbes pointillées : quantiles 0.025 et 0.975.

Taux de mortalité réel. Le taux de mortalité correspond à la fraction des infectés qui meurent, soit $\gamma(t)/(\gamma(t)+\beta)$. Le terme $\gamma(t)$ est calculé via la formule (2) et les données de mortalité. Nous obtenons ainsi, au 17 mars un taux de mortalité en France de 5.2/1000 (IC-95% :$(1.5/1000, 11.7/1000)$). La dynamique temporelle du taux de mortalité est représentée en Fig. 5.

Discussion.
Sur le nombre d'infectés et le taux de mortalité. Le nombre réel d'individus infectés en France est sans doute bien supérieur aux observations (nous trouvons ici un facteur $\times 16$), ce qui conduit à un taux de mortalité plus faible que celui calculé sur la base des cas observés. Néanmoins, si le virus devait contaminer 80% de la population Française (Ferguson et al., 2020), le nombre total de décès à déplorer en l'absence de variation du taux de mortalité (augmentation induite par exemple par une saturation des structures hospitalières, ou diminution liée à une meilleure prise en charge des

malades) serait de 277 000 (IC-95% :(81 000,629 000)). Cette estimation pourra être corroborée ou invalidée lorsque 80% de la population aura été infectée, y compris sur plusieurs années, en supposant qu'un individu infecté est définitivement immunisé. Il est à noter que les mesures de confinement ou de distanciation sociale peuvent décroître à la fois le pourcentage d'individus infectés dans la population et le degré de saturation des structures hospitalières.

Sur les différences entre France et Corée du Sud. Le modèle SIR mécanistico-statistique décrit bien les données française, mais mal l'infléchissement rapide du nombre de cas observé en Corée du Sud. Si l'on se base uniquement sur la première phase de l'épidémie en Corée du Sud, la valeur du R_0 estimé reste deux fois plus faible, indiquant une dynamique épidémique d'emblée plus lente. La différence entre la dynamique prédite par le modèle SIR et les données Sud-Coréennes est probablement liée à une gestion différente de l'épidémie en Corée, ayant un fort impact sur la dynamique épidémique (dépistage, traçage, isolement plus importants en Corée du Sud).

Sur la valeur de R_0. La valeur de R_0 obtenue en Corée du Sud est cohérente avec les estimations admises pour le COVID-19 (2.0,2.6) ; voir Ferguson et al. (2020). La distribution estimée en France est donc étonnamment élevée. Cette différence pourrait être due à une définition différente du R_0 suivant le type de modèle utilisé pour le calculer. Une estimation directe, par une méthode non-mécaniste, des paramètres (ρ, t_0) d'un modèle de la forme $\hat{\delta}_t = e^{\rho\,(t-t_0)}$ donne $t_0 = 30$ (30 janvier) et $\rho = 0.19$. Avec le modèle SIR, $I'(t) \approx I\,(\alpha - \beta)$ pour des temps petits ($S \approx N$), ce qui conduit à un taux de croissance égal à $\rho \approx \alpha - \beta$, et une valeur de $\alpha \approx 0.24$, soit $R_0 = 4.8$ ce qui est cohérent avec la distribution présentée en Fig. 3. Notons que $\beta = 1/20$ correspond à la période médiane d'excrétion virale de 20 jours décrite par Zhou et al. (2020). Une période plus courte conduirait à une valeur de R_0 plus faible.

Sur l'incertitude liée aux données. L'incertitude sur le nombre d'infectés et donc le taux de mortalité sont très élevés. Il faut donc interpréter avec prudence les prédictions pouvant être faites sur la base des données dont nous disposons actuellement en France. Nous ne proposons ici pas de prédiction, la dynamique future sera fortement influencée par les mesures de confinement qui seront prises.

Sur les hypothèses sous-jacentes au modèle. Les données utilisées contiennent une information limitée, d'autant plus que la période d'observation considérée est courte et correspond à la phase initiale de la dynamique épidémique qui peut être fortement influencée par des évènements discrets. Cette limite nous a conduit à utiliser un modèle particulièrement parcimonieux afin d'éviter des problèmes d'identifiabilité des paramètres. Les hypothèses sous-jacentes au modèle sont donc relativement simples et les résultats doivent être interprétés au regard de ces hypothèses. Ainsi, la date d'introduction t_0 doit être vue comme une date *équivalente* d'introduction dans une dynamique où une seule introduction serait déterminante pour le déclenchement de l'épidémie et les autres introductions (antérieures ou postérieures) auraient un effet non significatif sur la dynamique.

Appendice A : distributions jointes *a posteriori* Les distributions *a posteriori* jointes des trois couples de paramètres sont présentées en Fig. 6.

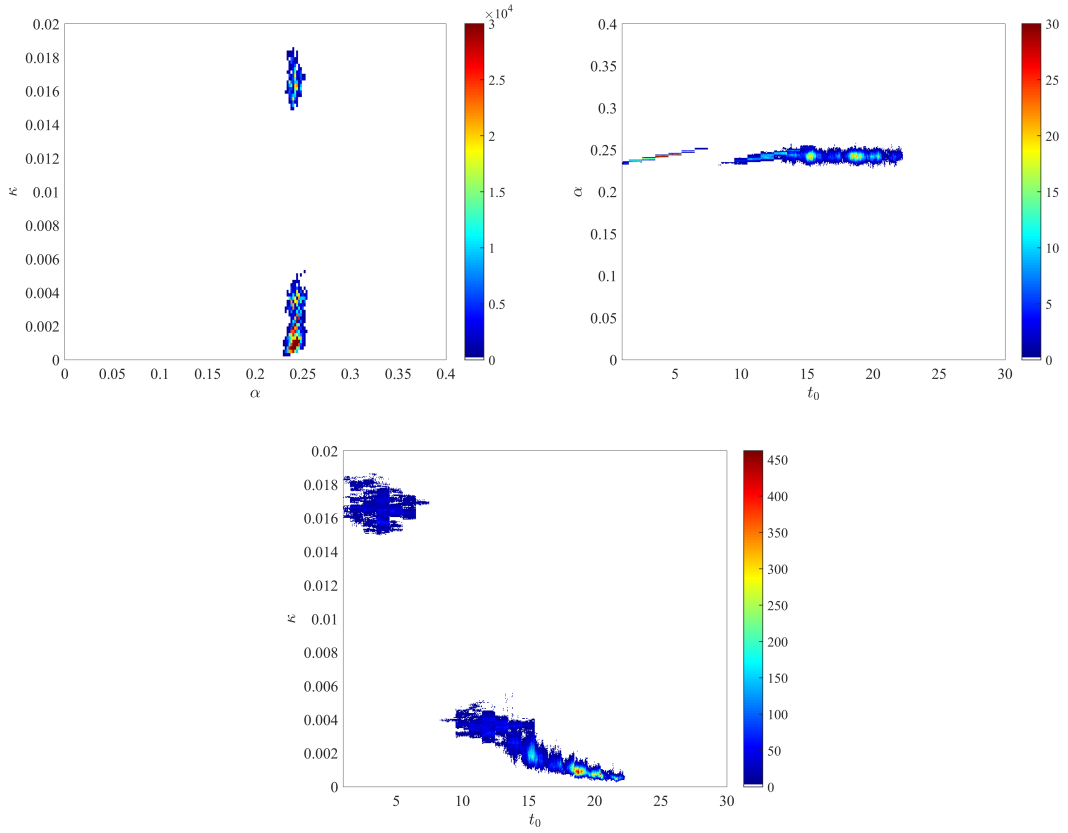


FIGURE 6 – **Distributions jointes *a posteriori* des paramètres (α, κ), (t_0, α) et (t_0, κ) en France.**

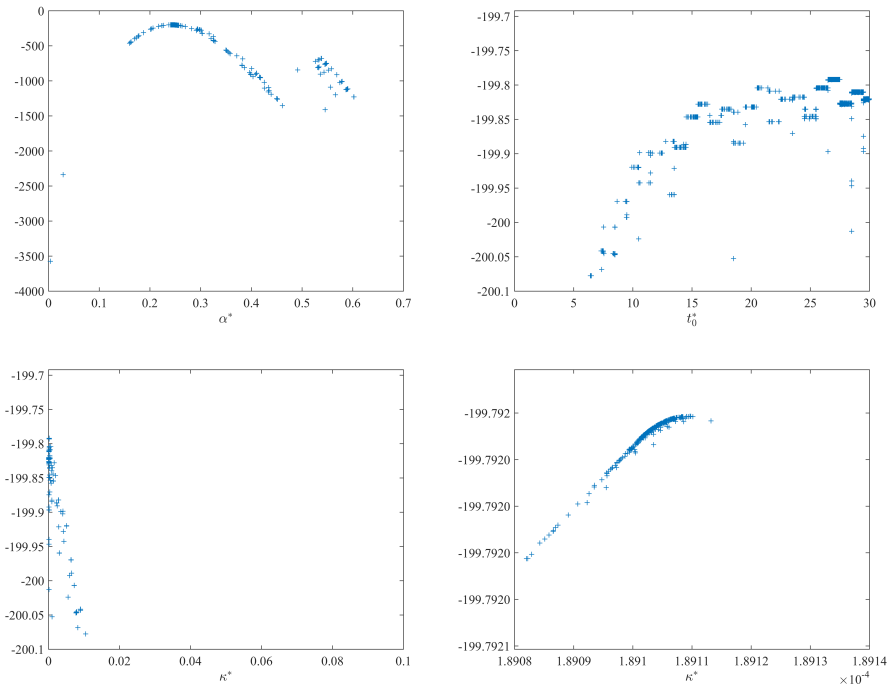


FIGURE 7 – **Estimateurs du maximum de vraisemblance.** Calculés à partir de 2000 valeurs initiales des paramètres, pour les données françaises.

Appendice B : estimateurs du maximum de vraisemblance Les calculs du MLE se font en utilisant une méthode de minimisation sous contrainte de type BFGS, appliquée à $-\ln(\mathcal{L})$, via l'outil Matlab® *fmincon*, à partir de 2000 valeurs initiales des paramètres. Cela conduit à 2000 valeurs de $(\alpha^*, t_0^*, \kappa^*)$. Nous n'avons retenu que la valeur conduisant à la plus forte vraisemblance. Les autres valeurs sont présentées en Fig. 7.

Références

Abboud, C., O. Bonnefon, E. Parent, et 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*.

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

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

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*.