Coupler des modèles épidémiologiques avec des données de santé publique : document pédagogique

Manon Costa, Fanny Delebecque, Grégory Faye, Jérôme Fehrenbach, Pascal Maillard, Sepideh Mirrahimi et Ariane Trescases[1]

[1]CNRS, UMR 5219, Institut de Mathématiques de Toulouse, 31062 Toulouse Cedex, France

7 avril 2021

Résumé

Ce document à vocation pédagogique présente des modélisations par compartiments de l'épidémie de COVID19, ainsi que le calibrage de ces modèles avec les données disponibles en Occitanie au cours de l'année 2020. Ce document est accompagné de ressources numériques (notebooks IPython) illustrant ces différents modèles, leur résolution numérique et leur couplage avec les données en Occitanie. L'objectif est de fournir des ressources, tant aux enseignants qu'aux étudiants, pour appréhender la modélisation épidémiologique et le couplage modèle-données. Et au passage on discute de processus de branchement et de résolution numérique de systèmes d'équations différentielles ordinaires. Les modèles par compartiments détaillés ici sont simplifiés par rapport aux modèles opérationnels (pas de structuration en âge ni en espace), mais leur introduction est étayée par des références bibliographiques permettant d'aller plus loin.

1 Introduction générale

La petite histoire

Au printemps 2020, pendant le confinement, les scientifiques se sont demandés comment ils pouvaient participer à l'effort contre la pandémie. Les mathématicien-ne-s spécialistes ou non de l'épidémiologie se sont penchés selon leurs compétences et leur disponibilité sur le sujet. Rapidement des modèles décrivant raisonnablement l'épidémie sont apparus, d'ailleurs des adaptations de modèles déjà existants se sont révélées pertinentes.

Cependant, on peut demander à un modèle de décrire les données passées et si tout va bien (et que le modèle est raisonnable) il y arrivera. Pour ce qui est de prédire l'avenir, ce qui est quand même le plus intéressant, les choses ne sont pas si simples. Au sein de la communauté mathématiques française, une étude menée par Josselin Garnier (non publiée) montrait qu'il fallait absolument disposer du paramètre f dénotant la proportion d'individus symptomatiques (ou $1-f$ qui est la proportion d'asymptomatiques) afin de pouvoir prédire l'avenir. Plus précisément : dans la famille de modèles compartimentaux, on arrivait à rendre compte des observations du début de l'épidémie en fixant la proportion $1-f$ d'asymptomatiques à une valeur arbitraire. Par contre l'évolution future de l'épidémie dépendait de façon cruciale de ce paramètre. Pour prendre deux exemples à l'extrême, imaginons que 50000 cas aient été déclarés à une certaine date. Si 1 personne contaminée sur 2 était asymptomatique alors au total 100000 personnes auraient été infectées, autrement dit une proportion négligeable de la population française. A l'inverse si 99 personnes sur 100 étaient asymptomatiques alors pour

50000 cas déclarés il y aurait en réalité 5 millions de personnes ayant contracté la maladie, donc supposément rendues immunisées. Et alors la population totale de personnes susceptibles de tomber malades se trouvait diminuée de cette valeur qui n'est pas négligeable par rapport aux 70 millions d'habitants de la France.

L'étude non publiée citée montrait de plus que le nombre f était mathématiquement impossible à estimer à partir des données recueillies (nombre de cas déclarées, nombre de décès). Et il était préconisé donc de déterminer ce paramètre f d'une autre façon, à l'aide par exemple de sondages randomisés sur la population.

De nombreuses personnes au sein de l'IMT ont utilisé leurs compétences pour appréhender la littérature sur l'épidémie de covid, et tâcher d'apporter une pierre à la connaissance générale. Parmi elles, un petit groupe s'est donné comme objectif d'appliquer une méthodologie inspirée par celle de Josselin Garnier sur des variantes du modèle, et de les calibrer en utilisant les données disponibles pour la région Occitanie (environ 6 millions d'habitants). Comme l'Occitanie a été moins touchée que d'autres régions par la première vague, certains paramètres notamment temporels pouvaient être différents de la valeur nationale.

Chemin faisant, les modèles se sont légèrement modifiés, des plages de valeurs pour certains paramètres sont apparues dans la littérature, ce qui fait que l'étude était sans cesse remise sur le métier. Dans le même temps la recommandation émise par les modélisateurs de procéder à des sondages randomisés pour estimer f afin de mieux connaître les paramètres de cette épidémie n'a pas été audible, ni en région parisienne ni en Occitanie. Cependant l'exercice se trouve quand même intéressant car il rassemble des morceaux de connaissance mathématique allant du début à la fin de la chaîne : du modèle théorique à une calibration des paramètres. Une envie de partager ces éléments a poussé quelques collègues à mettre au propre les travaux réalisés : documents scientifiques et codes de calcul.

Cette mise au propre contenant des pointeurs vers la littérature disponible (à une date donnée) constitue le présent document, qui se veut pédagogique. Aucun niveau n'est spécifiquement visé, même si le document tel quel serait lisible plutôt par des étudiants du niveau M1-M2. Les collègues intéressés pourront certainement en extraire de la substance pour illustrer des notions dans les cours de : équations différentielles ordinaires, processus stochastiques, analyse numérique, optimisation entre le L2 et le M1, ou encore servir de point de départ pour des projets ou exposés. Les notebooks IPython qui sont proposés peuvent également être adaptés à des niveaux et à des objectifs variés.

Qu'est-ce qu'un modèle ?

Modéliser un phénomène, c'est proposer des équations qui permettent de décrire et de prédire quantitativement des grandeurs qui lui sont associées. En physique et en mécanique, les lois de comportement sont assez reproductibles et un modèle décrivant le comportement d'une poutre en acier sera relativement précis pour décrire toutes les poutres en acier similaires.

En modélisation du vivant, les phénomènes en jeu sont extrêmement complexes et par exemple la transmission d'un virus fait intervenir tellement de processus distincts - pas forcément tous connus d'ailleurs - que l'on peut difficilement prédire si un individu donné va contracter la maladie. On peut s'en sortir de deux façons : établir des modèles probabilistes, ou bien établir des modèles sur une population nombreuse, ce qui revient à moyenner les effets sur un grand nombre d'individus. Les modèles que nous étudions sont dans la deuxième catégorie.

Quelles données peut-on utiliser ?

Lorsqu'un modèle est calibré au regard de données, il est clair que plus on dispose de données, et plus il sera facile de déterminer les paramètres du modèle. Cependant il faut prendre en compte les deux point suivants : 1) la plupart du temps on ne dispose pas d'observations de toutes les variables du modèles 2) les observations sont entachées de bruit de mesure. Le fait que toutes les variables ne soient pas observées n'est pas un souci en soi, l'étude discutée plus loin

de l'identifiabilité permet de savoir quels paramètres peuvent être estimés à partir de mesures partielles.

Le bruit non plus ne pose pas de problème en théorie à condition qu'on en connaisse la distribution. En règle générale on peut déterminer la distribution d'un bruit de mesure lorsque de nombreuses mesures sont disponibles, ce qui est le cas par exemple en météo lorsque les mêmes mesures sont collectées quotidiennement depuis des décennies. Ce n'est pas le cas pour la situation d'un début d'épidémie qui nous intéresse ici...

Dans notre cas, les données sont des statistiques de santé publique et sont constituées par des comptages d'effectifs (nombre de personnes infectées, hospitalisées, en réanimation, décédées). On considérera donc que notre jeu de données est affecté par un bruit de Poisson.

Identifiabilité des paramètres et identification effective.

On cherche à calibrer les paramètres du modèle avec les données mesurées, c'est-à-dire trouver des paramètres pour lesquels la prédiction du modèle est proche des mesures. On peut espérer y arriver lorsque le modèle est adapté au phénomène que l'on décrit. La technique présentée ici d'assimilation de données variationnelle consiste à minimiser une fonction-coût qui mesure l'écart entre les mesures et les prévisions du modèle.

Cependant certains paramètres ne peuvent pas être estimés : ce sont les paramètres qui influent pas (ou peu) sur les prévisions du modèle. Lorsque c'est le cas, plusieurs valeurs du paramètre peuvent fournir des prévisions satisfaisantes. Ce que l'on peut en déduire c'est que le paramètre en question n'est pas identifiable à partir des mesures. Mathématiquement nous fournirons une matrice de covariance d'erreur sur les paramètres estimés, en nous plaçant dans un cadre gaussien. Les directions dans lesquelles la covariance est grande sont les combinaisons linéaires de paramètres qui sont mal identifiables.

Organisation du document

Dans la première section, nous présentons les différentes données à disposition, puis nous introduisons dans la section suivante différents modèles épidémiologiques. Ensuite, nous décrivons la résolution numérique d'un modèle à compartiments et comparons les résultats obtenus aux données. Puis, nous présentons une analyse de sensibilité et expliquons comment optimiser certains paramètres de nos modèles à partir des données à notre disposition.

2 Des différentes données à disposition

Dans cette première section, nous présentons plusieurs jeux de données qui sont à notre disposition et en libre accès. Nous avons utilisé deux sources : l'agence régionale de santé d'Occitanie [1] et le site de santé publique France [2]. La plupart des données ont été collectées à partir de mi-Mars 2020 et sont mises à jour quotidiennement. Nous nous sommes arrêtés au 25 Octobre lors de la rédaction de la première version de ce document. Les données mises à disposition se décomposent en deux catégories principales : des données hospitalières et des données issues des laboratoires. Certaines d'entre elles sont accessibles par tranche d'âge ou encore par sexe. Nous avons fait le choix d'utiliser uniquement des données cumulées tout âge et tout sexe confondus puisque les modèles mathématiques que nous avons décidé de présenter sont à l'échelle (macroscopique) de la population totale en Occitanie et ils ne seront donc structurés par une variable d'âge ou de sexe.

1. https ://www.occitanie.ars.sante.fr/coronavirus-dernier-point-de-situation-en-occitanie-0
2. https ://geodes.santepubliquefrance.fr/#c=home

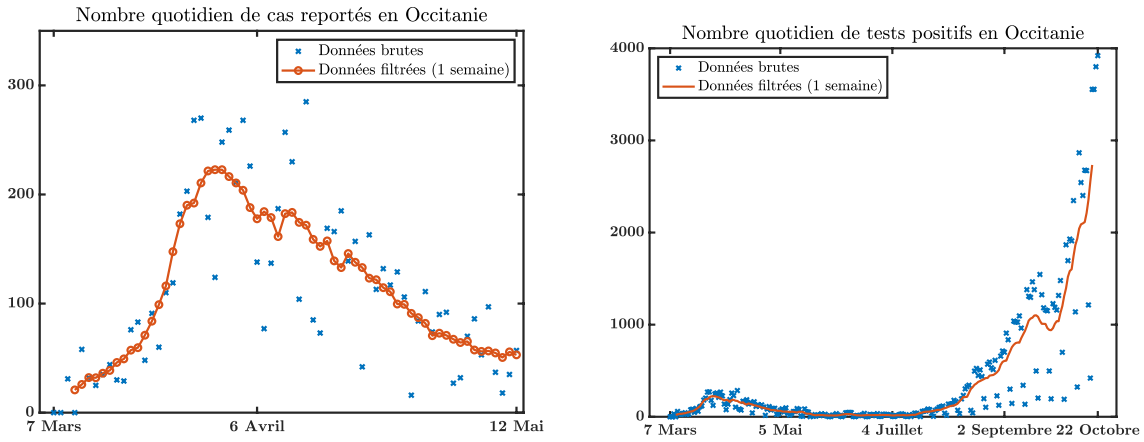

FIGURE 1 – Nombre quotidien de cas reportés à l'échelle de l'Occitanie sur deux échelles de temps différentes : au tout début de l'épidémie pendant la phase de confinement (à gauche) et du premier jour de l'épidémie jusqu'au 25 Octobre (à droite).

2.1 Nombre quotidien et cumulé de cas reportés

Nous commençons par présenter en Figure 1 le nombre quotidien de cas reportés à l'échelle de l'Occitanie sur deux échelles de temps différentes où les stratégies de dépistage ont été différentes. Au début de l'épidémie et jusqu'à la fin du confinement (11 Mai), les tests étaient principalement effectués en hôpital sur des patients dont on était presque certains qu'ils étaient infectés. Progressivement, le dépistage est devenu plus massif, et les laboratoires d'analyse ont pris le relai. Alors que l'on comptait 300 cas reportés au plus fort de l'épidémie pendant le confinement, nous sommes maintenant (au 25 Octobre) à des nombres records de presque 4000 cas reportés. Cela traduit simplement le fait que notre stratégie de test a évolué. Il est raisonnable de conjecturer que si l'accès aux tests avaient été plus systématique lors de la première vague nous aurions eu des résultats comparables à ce que l'on observe sur les mois de Septembre et d'Octobre. Une observation des données montrent les disparités de nombre de cas suivant les jours de la semaine, traduisant simplement le fait que moins de tests sont effectués en week-end. C'est pourquoi dans toutes les données que nous présenterons, nous afficherons toujours les données brutes et des données filtrées sur une semaine permettant de dégager une tendance plus claire de l'évolution au cours du temps des données étudiées.

Il est aussi intéressant de regarder le nombre de cas reportés en cumulé. Nous présentons en Figure 2 les données pour la période du 7 Mars au 12 Mai. Une première observation consiste à remarquer que les données sont naturellement plus régulières que pour les cas quotidiens, ce qui est normal puisque la courbe du nombre de cas reportés en cumulé est l'intégrale au cours du temps de la courbe du nombre quotidien de cas reportés. La deuxième observation, cruciale pour l'analyse des sections à venir, consiste à voir qu'en échelle logarithmique, le début de l'épidémie se traduit par une croissance linéaire, matérialisée par la droite bleue sombre sur la Figure 2 (droite). Cette croissance linéaire en échelle logarithmique correspond donc à une croissance exponentielle pour le nombre cumulé de cas reportés, typique de la phase initiale d'une épidémie. Cela permet notamment de donner un début à l'épidémie, à savoir donner une approximation à la date possible du premier cas en Occitanie. L'idée est la suivante. Si l'on note $\mathcal{R}(t)$, le nombre cumulé de cas reporté à l'instant, et que l'on suppose que dans les premiers jours de l'épidémie $\mathcal{R}(t)$ croît de façon exponentielle, alors on écrit

$$\mathcal{R}(t) = e^{\chi(t-t_0)},$$

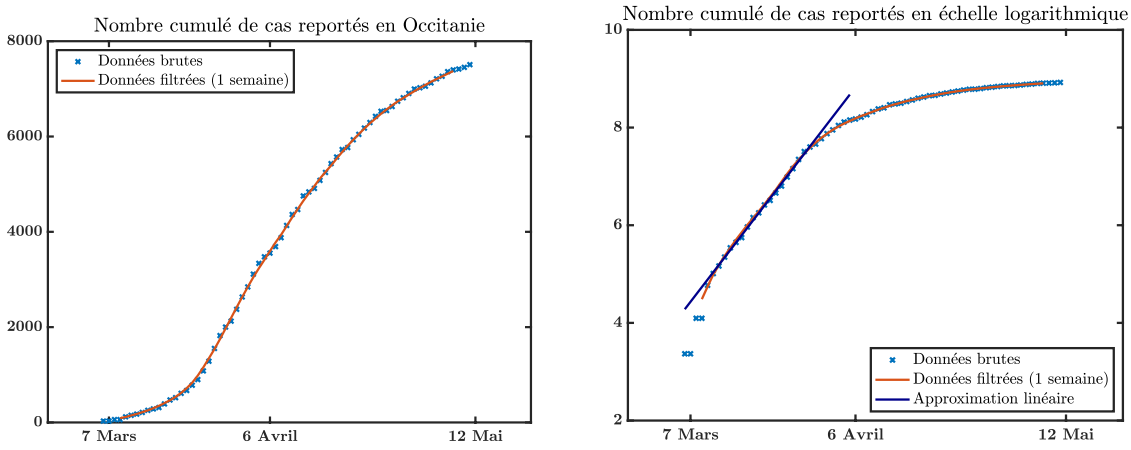

FIGURE 2 – Nombre cumulé de cas reportés à l'échelle de l'Occitanie du 7 Mars au 12 Mai. La figure de droite est en échelle logarithmique et permet de visualiser la propriété qu'au début d'une épidémie le nombre de cas reportés est exponentiel, puisque linéaire en échelle logarithmique. L'approximation linéaire (courbe en bleu foncé) a été faite entre les données du 13 et 30 Mars, donnant une pente de 0.1516 et une ordonnée à l'origine de 4.2796 permettant d'inférer une date approximative du premier cas en Occitanie autour du 8 Février.

avec pour normalisation qu'à l'instant initial de l'épidémie, noté t_0 ici, il y avait un seul individu d'infecté. C'est à dire $\mathcal{R}(t_0) = 1$. A l'aide de la Figure 2, on estime

$$\chi \simeq 0.1516 \text{ et } t_0 \simeq -28.2366 \text{ jours.}$$

Cela nous donne donc que l'épidémie aurait commencé environ 28 jours avant le 7 Mars, soit le 8 Février. C'est une date approximative, mais elle véhicule bien l'idée que les premiers cas de malades dus au virus ont dû être au tout début de l'année et sont restés non détectés par nos systèmes de santé.

2.2 Nombre de personnes hospitalisées ou en réanimation

Une des causes indirectes du virus est la nécessité pour certains malades de devoir être hospitalisés, parfois en réanimation. Et nous disposons donc des données hospitalières telles que le nombre de personnes hospitalisées par jour, le nombre de personnes en réanimation par jour, ainsi que les données des nouvelles admissions journalières à l'hôpital ou en chambre de réanimation, voir la Figure 3. Ces données sont des plus sensibles puisque nous savons que notre système de santé dispose d'un plafond sur le nombre maximum de lits possibles dans les hôpitaux mais aussi du nombre maximum de ventilateurs nécessaires aux personnes en réanimation. Et donc dans la prochaine section, nous proposerons un modèle qui permet de modéliser le taux d'occupation des lits en réanimation. Il est intéressant d'observer que la fraction de personnes en réanimation par rapport au nombre total de personnes hospitalisées varie entre 0.11 et 0.33 sur la période considérée avec une moyenne à 0.21 traduisant qu'environ 1/5 des patients admis à l'hôpital nécessitent une prise en charge en unité de réanimation.

2.3 Nombre cumulé de personnes décédées

Hélas, des décès sont à déplorer en lien direct avec l'épidémie. Et nous reportons dans la Figure 4 les données correspondantes : nombre cumulé de personnes décédées et nombre quotidien de nouvelles personnes décédées. Bien que les personnes décédées n'étaient pas toutes

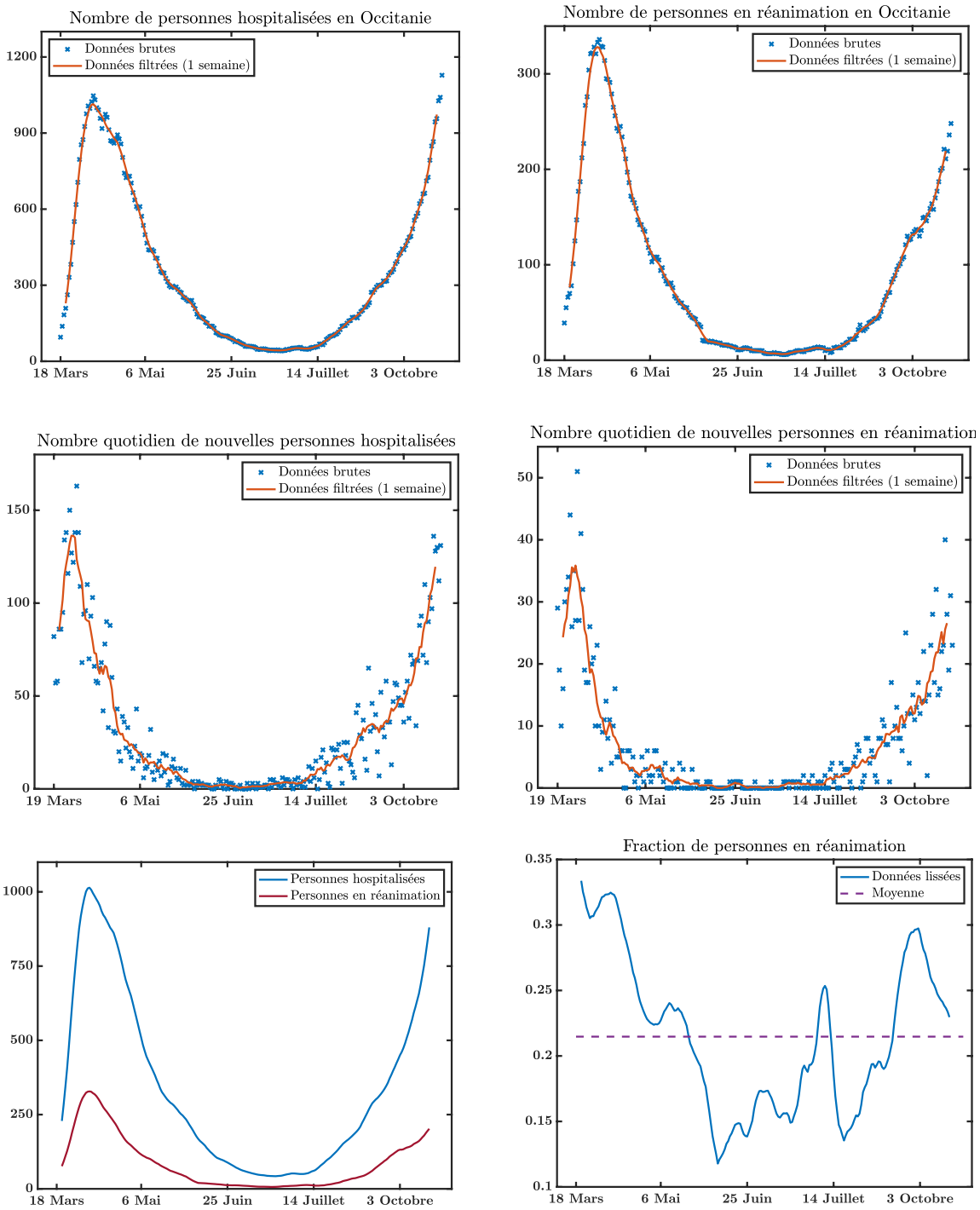

FIGURE 3 – En haut : nombre de personnes hospitalisées (gauche) et en réanimation (droite) en Occitanie. Au milieu : nombre quotidien de nouvelles personnes hospitalisées (gauche) et en réanimation (droite) en Occitanie. En bas : comparaison du nombre de personnes hospitalisées et en réanimation (gauche) et fraction du nombre de personnes en réanimation par rapport au nombre de personnes hospitalisées (droite).

en réanimation avant leur mort, nous pouvons néanmoins estimer la du nombre de personnes nouvellement décédées sur le nombre de personnes en réanimation. Ce ratio permet de se donner

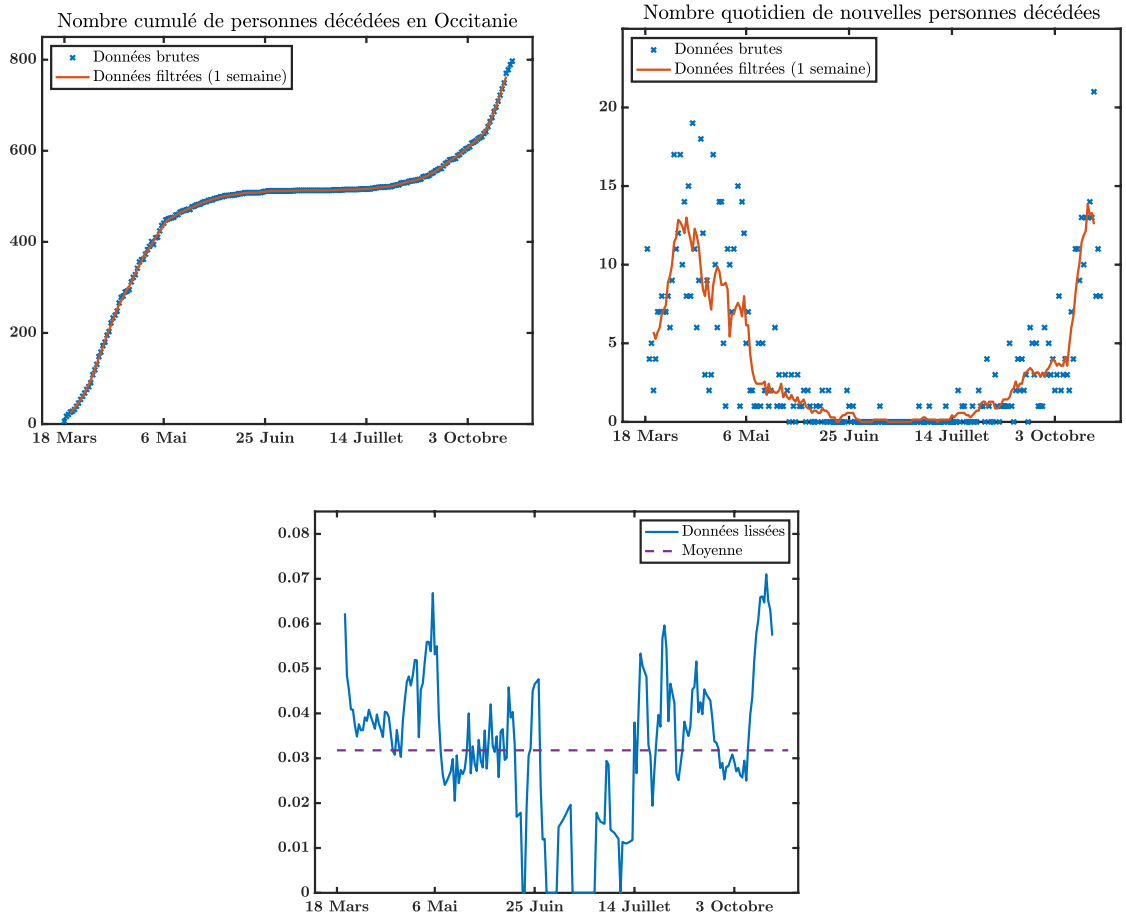

FIGURE 4 – En haut : nombre cumulé de personnes décédées (gauche) et nombre quotidien de nouvelles personnes décédées (droite). En bas : fraction du nombre de personnes nouvellement décédées sur le nombre de personnes en réanimation.

une certaine idée sur le taux de mortalité λ_m pour les personnes en réanimation. Nous trouvons que ce taux moyen est d'environ $\lambda_m \simeq 0.03$ sur la période considérée.

3 Différents modèles épidémiologiques

3.1 Modèles à compartiments

3.1.1 Modèle SIR

Un outil classique pour la modélisation des épidémies est les modèles dits à compartiments. Ce sont des systèmes d'équations différentielles où chaque inconnue représente un compartiment, c'est-à-dire une classe d'individus classés en fonction de leur statut par rapport à la maladie : infecté, guéri, etc. Le modèle à compartiments originel est le modèle SIR [5] :

$$\begin{cases} S'(t) = -\tau\, S(t)\, I(t), \\ I'(t) = \tau\, S(t)\, I(t) - \nu\, I(t), \\ R'(t) = \nu\, I(t). \end{cases} \tag{3.1}$$

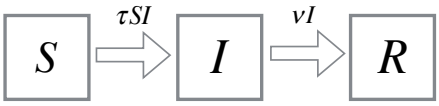

FIGURE 5 – Modèle à compartiments de type SIR (3.1).

La quantité S représente le nombre d'individus non encore infectés mais susceptibles de l'être, formant le compartiment des *Susceptibles* ; la quantité I représente le nombre d'individus infectieux, formant le compartiment des *Infectieux* ; et la quantité R représente le nombre des individus qui après infection ne sont plus infectieux (guéris ou morts), formant le compartiment des *Rétablis*. On suppose ici qu'une immunité est acquise, et donc ces individus n'interviennent plus dans la propagation de l'épidémie. Le paramètre $\tau > 0$ représente la force d'infection, que l'on peut aussi écrire $\tau = \beta/N$, où N est la taille de la population, et β correspond à la fréquence de nouvelle contamination d'un individu infectieux. Le paramètre $\nu > 0$ est le taux de guérison et/ou décès des infectieux.

Remarquons en sommant les trois équations que la masse du système est conservée : $S'(t) + I'(t) + R'(t) = 0$. En effet, on suppose que sur l'échelle de temps de l'épidémie la population est stable : on néglige les effets d'immigration ou d'émigration, de naissances et de décès autres que ceux de l'épidémie.

Un individu infecté reste infectieux pendant une durée moyenne $1/\nu$. Durant ce temps, la fréquence moyenne à laquelle il contamine de nouveaux individus est β. Le produit β/ν peut donc s'interpréter comme le nombre moyen d'individus contaminés par chaque personne infectieuse. Noté

$$\mathscr{R}_0 := \frac{\beta}{\nu} = \frac{N\tau}{\nu}, \tag{3.2}$$

on l'appelle le *nombre de reproduction de base*.

Propriété 3.1. *On suppose $S(0) > 0$ et $I(0) > 0$. Si $\mathscr{R}_0 < \frac{S(0)}{N}$, alors l'épidémie s'éteint au sens suivant :*

$$\text{pour } t > 0,\ I'(t) < 0; \qquad I(t) \xrightarrow[t\to\infty]{} 0; \qquad S(t) \xrightarrow[t\to\infty]{} S_\infty > 0.$$

Si $\mathscr{R}_0 > \frac{S(0)}{N}$, alors l'épidémie envahit la population jusqu'à atteindre un pic avant de s'éteindre : il existe un temps $t_{pic} > 0$ tel que $S(t_{pic}) = \frac{N}{\mathscr{R}_0}$, et

$$\text{pour } t < t_{pic},\ I'(t) > 0; \qquad \text{pour } t > t_{pic},\ I'(t) < 0;$$
$$I(t) \xrightarrow[t\to\infty]{} 0; \qquad S(t) \xrightarrow[t\to\infty]{} S_\infty > 0.$$

*On dit que l'*immunité collective *est atteinte quand on le pic est atteint, c'est-à-dire quand S atteint $\frac{N}{\mathscr{R}_0}$.*

Remarque 3.2. *Si on prend comme date initiale ($t = 0$) l'entrée dans le système du premier individu infectieux, alors la population susceptible initiale est quasiment l'ensemble de la population. Le seuil pour $\mathscr{R}_0$ est donc $\frac{S(0)}{N} \sim 1$.*

Pour la preuve, on vérifie d'abord que S et I restent strictement positives pour tout temps. On peut ensuite utiliser la monotonie de S grâce au signe de S' et étudier la monotonie de I en utilisant $I' = (\tau S - \nu)I$.

Le nombre de reproduction de base est donc un indicateur très important : en fonction de s'il est au-dessus ou en dessous du seuil critique on s'attend à ce que l'épidémie se propage ou au contraire s'éteigne ; il indique de plus la hauteur du pic.

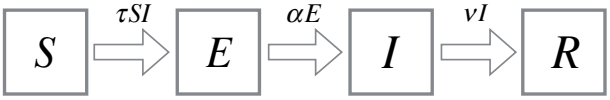

FIGURE 6 – Modèle à compartiments de type SEIR (3.3).

Par ailleurs, dans le cas où l'épidémie se propage, on peut calculer sa vitesse de propagation en tout début d'épidémie. En effet, supposons $\mathscr{R}_0 > 1$. Au début de l'épidémie on a $S(0) \approx N$, d'où

$$I'(t) \approx (\tau N - \nu) I(t) = (\mathscr{R}_0 - 1)\nu I(t).$$

De manière informelle, la croissance du nombre d'infectés en tout début d'épidémie est exponentielle :

$$I(t) \approx e^{(\mathscr{R}_0 - 1)\nu t} I(0).$$

3.1.2 Modèle SEIR

Une hypothèse sous-jacente du modèle (3.1) est que les individus infectés sont instantanément infectieux. En réalité, il faut prendre en compte un temps de latence pendant lequel un individu exposé à la maladie va l'incuber mais ne pas encore être infectieux. On considère alors le modèle SEIR, à quatre compartiments :

$$\begin{cases} S'(t) = -\tau\, S(t)\, I(t), \\ E'(t) = \tau\, S(t) I(t) - \alpha\, E(t), \\ I'(t) = \alpha\, E(t) - \nu\, I(t), \\ R'(t) = \nu\, I(t). \end{cases} \tag{3.3}$$

La quantité E représente les individus *Exposés* à la maladie qui sont en train de la développer mais ne sont pas encore infectieux. Ils passent du compartiment *Exposé* au compartiment *Infectieux* avec un taux $\alpha > 0$, qui représente l'inverse du temps moyen de latence de la maladie.

3.2 Quelques modèles pour l'épidémie de Covid-19

Si les systèmes SIR et SEIR constituent un modélisation théorique intéressante, ils sont en pratique parfois difficiles à appliquer à des cas concrets. En particulier dans notre cas, toute la difficulté réside dans le fait que nous n'observons pas directement la quantité I d'individus infectieux (compte tenu d'une part de la proportion non négligable d'individus asymptomatiques ou paucisymptomatiques, difficilement repérables, et d'autre part des différentes politiques de tests).

En revanche, nous disposons

1. de données partielles fournies par les tests : nombre de cas malades détectés et nombre d'individus testés. Ces données partielles peuvent nous aider à retrouver le nombre total d'infectés dans la population.
2. de données variées sur les individus malades sévères : nombres d'hospitalisations, nombre de malades en soins intensifs, nombre de guérisons (correspondant à la sortie de l'hôpital) et décès. Ces données sont une indication très utile sur l'état de l'épidémie dans la population.

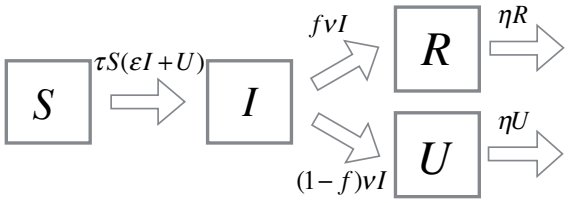

FIGURE 7 – Modèle à compartiments de type Susceptibles/Infectés avec cas reportés et non reportés.

3.2.1 Un modèle prenant en compte les cas reportés et non reportés

Nous commençons par un modèle qui exploite les données fournies par les tests [6, 7]. Partant d'un modèle SIR classique, nous supposons que les individus *Infectieux* peuvent évoluer vers deux compartiments, suivant si ils sont reportés ou non.

$$\begin{cases} S'(t) = -\tau\, S(t)\,[\epsilon I(t) + U(t)], \\ I'(t) = \tau\, S(t)\,[\epsilon I(t) + U(t)] - \nu\, I(t), \\ R'(t) = f\nu\, I(t) - \eta\, R(t), \\ U'(t) = (1-f)\nu\, I(t) - \eta\, U(t). \end{cases} \tag{3.4}$$

La quantité S représente les *Susceptibles*, qui une fois infectés passent dans le compartiment des *Infectieux*, notés I. Une proportion $f \in]0,1[$ des *Infectieux* est reportée : elle passe au bout d'un certain temps moyen $\nu^{-1} > 0$ dans le compartiment des *Reportés*, notés R. On fera bien attention que dans le modèle (3.4), la notation R ne fait plus référence aux personnes rétablies, mais aux personnes infectées dont les cas sont reportés. La proportion $(1-f)$ restante passe (en suivant la même loi de durée de transition) dans le compartiment des *Non Reportés*, notés U (pour *Unreported*). Les *Reportés* et *Non Reportés* restent dans leur compartiment jusqu'à leur guérison ou décès, survenant au bout d'un temps moyen $\eta^{-1} > 0$. Pour alléger l'écriture du système, nous n'écrivons pas l'équation d'évolution de la quantité des *Rétablis* (quantité qui par conservation de la taille de la population N s'obtient par le calcul $N - S - I - R - U$).

Dans ce modèle, on a supposé que les individus reportés ne sont plus source d'infections car ils se mettent en isolation. Les *Infectieux* et les *Non Reportés* sont sources de nouvelles infections avec une infectiosité potentiellement distincte : $\epsilon > 0$ correspond à leur infectiosité relative.

Pour exploiter ce modèle, il est crucial de connaître ou d'estimer le paramètre f qui correspond à la proportion d'individus malades qui sont détectés. En pratique, un moyen d'estimer ce paramètre est de pratiquer des tests aléatoires au sein de la population (sondages) à un instant t donné. Ces sondages doivent être intelligemment calibrés pour être suffisamment représentatifs de la population dans sa diversité.

3.2.2 Un modèle basé sur les hospitalisations

Nous proposons ici un modèle qui repose sur les données dont nous disposons sur les hospitalisations en unité de soins intensifs (ICU), les décès. Remarquons que nous pourrions imaginer plusieurs variantes faisant apparaître au choix les données dont nous disposons : les reportés et/ou les hospitalisations et/ou les hospitalisations en ICU et/ou les décès et/ou les guérisons (sortie d'ICU). À chaque fois que nous ajoutons une équation, nous ajoutons aussi un ou des paramètres à estimer. Il n'est donc pas forcément judicieux de maximiser le nombre d'équations. Ici, nous avons choisi de nous appuyer sur les deux variables les plus critiques à prévoir et réguler : le nombre de décès et le nombre d'hospitalisations en ICU (le nombre de places en ICU étant limité). Un autre avantage de ces deux variables est que comme elles concernent les

cas sévères on peut s'attendre à ce que - au moins hors des périodes de saturation du système sanitaire - les modifications des politiques sanitaires affectent peu leur comptabilisation.

Nous proposons le modèle suivant :

$$\begin{cases} S'(t) = -\tau S(t)\left[\epsilon I_s(t) + I_a(t)\right], \\ E'(t) = \tau S(t)\left[\epsilon I_s(t) + I_a(t)\right] - \alpha E(t), \\ I_s'(t) = f\alpha E(t) - \eta_s I_s(t), \\ I_a'(t) = (1-f)\alpha E(t) - \eta_a I_a(t), \\ M'(t) = f\alpha E(t) - \eta_m M(t), \\ H'(t) = \eta_m M(t) - \lambda_h H(t), \\ D'(t) = \lambda_d H(t) \end{cases} \tag{3.5}$$

Il s'agit d'une extension d'un modèle SEIR. Les individus *Susceptibles*, notés S, si infectés, passent dans le compartiment des *Exposés*, notés E. Pendant un temps de latence les *Exposés* développent la maladie sans être encore infectieux. Au moment où ils deviennent infectieux, on les sépare en deux catégories : une certaine proportion $f > 0$ d'entre eux sera amenée à être hospitalisée en ICU, et l'autre proportion $(1 - f)$ ne passera jamais en ICU.

La proportion $(1 - f)$ d'individus *Exposés* qui n'est pas amenée à être hospitalisée en ICU passe par le compartiment des *Infectieux Non Futurs ICU*, noté I_a, avant de sortir du système quand ils cessent d'être infectieux (guérison ou décès, compartiment non explicité).

La proportion $f > 0$ d'individus *Exposés* qui sera hospitalisée en ICU passe (au moment où ces individus deviennent infectieux) dans le compartiment des *Futurs ICU*, noté M. Ils y restent jusqu'au moment de leur hospitalisation en ICU, passant alors dans le compartiment des *Hospitalisés en ICU* noté H. Les *Hospitalisés en ICU* sortent de leur compartiment soit au moment de leur guérison (compartiment non explicité) soir au moment de leur décès (compartiment des *Décédés en ICU*, noté D).

Concernant les individus malades amenés à être hospitalisés en ICU, on remarque que le moment de l'hospitalisation ne coïncide pas nécessairement avec le moment de la fin de l'infectiosité. Souvent, les hospitalisations en ICU arrivent plutôt après la fin de l'infectiosité. On ne peut pas donc pas dire que le compartiment des *Futurs ICU* contient (uniquement) des individus infectieux. Pour prendre en compte l'infectiosité de ces cas, on introduit en parallèle le compartiment des *Infectieux Futurs ICU*, notés I_s. Ce compartiment est abreuvé par les mêmes individus que le compartiment *Futurs ICU*, mais alors que les *Futurs ICU* quittent leur compartiment au moment de leur hospitalisation (vers le compartiment des *Hospitalisés ICU*), les *Infectieux Futurs ICU* quittent leur compartiment au moment où ils cessent d'être infectieux (vers un compartiment non explicité). Sont donc pris en compte dans les sources d'infection uniquement les compartiments *Infectieux Futurs ICU* et *Infectieux Non Futurs ICU*. Leur infectiosité relative est notée $\epsilon > 0$.

Le paramètre $\alpha > 0$ correspond à l'inverse du temps moyen de latence (durée entre l'infection et le début de l'infectiosité). Les paramètres η_s et η_a correspondent respectivement à l'inverse de la durée d'infectiosité effective (c'est-à-dire prenant en compte les éventuelles isolations en quarantaine) des *Infectieux Futurs ICU* et *Infectieux Non Futurs ICU*. Le paramètre η_m est l'inverse de la durée moyenne entre le début de l'infectiosité et l'hospitalisation en ICU. Le paramètre $\lambda_h > 0$ correspond au taux quotidien de sortie de l'hôpital (décès ou guérison), des hospitalisés en ICU, et le paramètre $\lambda_d > 0$ est leur taux quotidien de décès.

3.2.3 Impact du confinement

Pour lutter contre l'épidémie, une mesure radicale consiste à confiner la population afin de limiter les interactions entre individus, et donc mécaniquement limiter les contaminations. Dans notre modèle, cela se traduit par une chute brutale de la force d'infection au moment de la mise en place du confinement. Dans le système (3.5), on considère désormais que le paramètre τ dépend du temps selon

$$\tau(t) = \tau_0 \text{ si } 0 \leq t < T, \qquad \tau(t) = \tau_0\theta \text{ si } t \geq T. \tag{3.6}$$

(En pratique, on pourra remplacer par une version lissée, plus stable numériquement). Ici, $\tau_0 > 0$ est la force d'infection intrinsèque dans la population (avant la mise en place des mesures sanitaires type confinement), et $\theta \in]0, 1[$ mesure l'efficacité du confinement. Le temps T est la date à laquelle les mesures commencent à faire effet. On n'impose pas que cette date T soit la date de mise en place officielle du confinement. D'une part cela permet de prendre un compte un éventuel décalage entre la mise en place des mesures et leur application réelle dans la population. D'autre part, dans notre modèle, le fait de modifier le paramètre τ à une date donnée à un effet immédiat sur, par exemple, le nombre de décès ; alors que dans la réalité on n'a pas cette immédiateté, on voit plutôt les effets plusieurs jours voire semaines plus tard. Considérer T comme un paramètre libre permet de compenser ce décalage.

3.3 Paramètres

On a donc construit le modèle suivant :

$$\begin{cases} S'(t) = -\tau(t)S(t)\left[\epsilon I_s(t) + I_a(t)\right], \\ E'(t) = \tau(t)S(t)\left[\epsilon I_s(t) + I_a(t)\right] - \alpha E(t), \\ I_s'(t) = f\alpha E(t) - \eta_s I_s(t), \\ I_a'(t) = (1-f)\alpha E(t) - \eta_a I_a(t), \\ M'(t) = f\alpha E(t) - \eta_m M(t), \\ H'(t) = \eta_m M(t) - \lambda_h H(t), \\ D'(t) = \lambda_d H(t) \end{cases} \tag{3.7}$$

avec

$$\tau(t) = \tau_0 \text{ si } 0 \leq t < T, \qquad \tau(t) = \tau_0\theta \text{ si } t \geq T. \tag{3.8}$$

Il faut à présent évaluer les valeurs des paramètres utilisés dans le modèle. On a onze paramètres :

— Le ratio f ;
— Autres paramètres liés à l'infectiosité : τ_0, θ, T, ϵ, α, η_s, η_a ;
— Autres paramètres liés à l'hospitalisation : η_m, λ_h, λ_d.

3.3.1 Le ratio f

Le ratio f représente la proportion d'individus infectés qui est amenée à être hospitalisée en ICU. Il est critique de connaître ce paramètre pour estimer le taux d'incidence de la maladie au sein de la population à partir de la donnée des hospitalisés en ICU. En effet, le nombre d'individus infectés ne fait pas partie de nos observables. En pratique, ce paramètre f sera choisi à partir de la littérature en épidémiologie basée sur des études de taux de sévérité de la maladie sur un échantillon de la population.

3.3.2 Infectiosité

Le paramètre ε en facteur de I_s dans le terme de transmission permet de rendre compte des différences relatives d'infectiosité entre les deux compartiments infectieux. Cela inclue notamment des effets des quarantaines/isolations lors de l'apparition de symptômes. Cependant, ce paramètre semble qualitativement faire double emploi avec le paramètre η_s. Il semble donc préférable de le fixer (=1) afin de réduire le nombre de paramètres du modèle. Il faut alors comprendre le paramètre $1/\eta_s$ comme un indicateur global de l'infectiosité du compartiment I_s, à la fois en termes de virulence et de durée (et qui prend aussi en compte la quarantaine ou l'isolation spontanée des malades). On choisit donc

$$\varepsilon = 1.$$

Les ordres de grandeur de α, η_s et η_a sont choisis à partir de la littérature en épidémiologie. Ces quantités peuvent être estimées à partir des mesures de ce qu'on appelle "le temps de génération" correspondant au temps entre le début de l'infection d'une personne infectée et du début de l'infection de son infectant, ou bien de ce qu'on appelle "l'intervalle sériel", correspondant au temps entre l'apparition des symptômes de l'infecté et de son infectant. Les distributions de ces quantités sont estimées en faisant des traçages de contacts [8, 10, 15]. Une autre information obtenue à partir des mesures cliniques, qui aide en particulier à donner un ordre de grandeur pour η_a est la durée d'infectiosité, entre le début de la maladie et la fin de l'infectiosité (test négatif). En se basant sur ces études nous avons pris en compte les contraintes suivantes sur nos choix des paramètres :

$$1/\alpha \in [1, 4],$$
$$1/\eta_s + 1/\alpha \in [4, 7],$$
$$1/\eta_s \geq 1,$$
$$\eta_a \in [0.05, \eta_s].$$

Les paramètres τ_0, θ et T sont estimés à partir des données (voir la section suivante).

3.3.3 Hospitalisation

On estime directement le taux λ_d à partir des données quotidiennes des décès et des hospitalisations en ICU. Les ordres de grandeurs des paramètres η_m et λ_h peuvent être également choisis à partir des données des hospitalisations, et en particulier en estimant le temps moyen depuis l'apparition des symptômes et l'hospitalisation en ICU, et le temps moyen passé en ICU. Nous avons en particulier utilisé des études faites dans [14] pour estimer ces paramètres. Une information supplémentaire qui est nécessaire pour pouvoir estimer η_m à partir de ces mesures est le temps d'incubation, correspondant au temps entre le début de l'infection et l'apparition des symptômes, estimé dans la littérature en épidémiologie entre 5 et 7 jours [10–12]. En se basant sur ces études nous avons pris en compte les contraintes suivantes sur nos choix des paramètres :

$$1/\eta_m + 1/\alpha \in [8, 20],$$
$$1/\lambda_h \in [10, 14].$$

Le ratio f représente la proportion d'individus infectés qui est amenée à être hospitalisée en ICU. Il est critique de connaître ce paramètre pour estimer le taux d'incidence de la maladie au sein de la population à partir de la donnée des hospitalisés en ICU. En effet, le nombre d'individus infectés ne fait pas partie de nos observables. En pratique, ce paramètre f sera choisi à partir de la littérature en épidémiologie basée sur des études de taux de sévérité de

la maladie sur un échantillon de la population [14] (voir aussi [13] pour un exemple d'étude sérologique sur une population aléatoirement sélectionnée). Nous prendrons

$$f \in [0.004, 0.01].$$

3.4 Limites et améliorations

Nous donnons ci-dessous quelques pistes pour améliorer ce travail de modélisation.

3.4.1 Choix des paramètres épidémiologiques

Les choix des paramètres épidémiologiques pourraient être améliorés au fur et à mesure avec une meilleure compréhension de la maladie. En effet, la maladie était peu connue au début de l'épidémie et les données pour estimer les différents paramètres épidémiologiques, comme le temps d'incubation et l'intervalle sériel restaient limitées. Très souvent elles provenaient d'autres pays que la France, en particulier non Européens, et surtout de la Chine, où le virus pouvait avoir d'autres propriétés (modifiées par la suite dû à une éventuelle mutation) et où les mesures de mise en quarantaine et la politique de traçage de contact étaient différentes. De telles mesures sont très différentes d'un pays à l'autre. Or il est important de prendre en compte dans une étude statistique, par exemple pour calculer l'intervalle sérielle, si de telles mesures étaient mises en place. Une autre difficulté pour estimer ce type de paramètres vient de la complexité et de l'hétérogénéité de la maladie. Par exemple l'estimation de l'intervalle sériel est naturellement basée sur les infectés symptomatiques ou symptomatiques sévères. Il est plus difficile d'estimer le temps d'infectiosité des personnes peu symptomatiques ou asymptomatiques. Enfin, les études consacrées à l'estimation des paramètres épidémiologiques étant très nombreuses il est important de s'assurer de la fiabilité des données traitées et de la prise en compte des biais d'échantillonage.

3.4.2 Choix de modèle à compartiments

Les modèles de type SIR supposent une loi exponentielle pour le temps passé dans chaque compartiment. Or les distributions des temps passés dans certains de ces compartiments, comme le temps de l'infectiosité ou le temps passé à l'hopital, s'approchent plus des lois gamma ou log-normal ou Weibull [1]. Pour s'approcher de ces distributions l'on peut utiliser un modèle intégro-différentiel au lieu du modèle de type SIR simple [9] (notons que ce type de modèle était déjà proposé dans l'article pionnier de Kermack et McKendrick [5]). Une autre option utilisée dans certains travaux de modélisation est de multiplier le nombres de compartiments pour avoir une somme des lois exponentielles [14].

3.4.3 Prise en compte de l'hétérogénéité

Les modèles présentés ci-dessus supposent une population homogène. Or la présence d'hétérogénéité, comme l'hétérogénéité spatiale ou en fonction des différentes classes de population, pourrait modifier la dynamique d'une épidémie. A titre d'exemple la prise en compte d'une hétérogénéité de taux de contact entre les différentes classes d'âge de population pourrait mener à une prédiction très différente du pic ou de la taille de l'épidémie [2]. Dans l'épidémie de COVID 19, il y a une hétérogénéité importante au niveau du taux de sévérité de la maladie pour les différentes classes d'âges, voir la Figure 8 pour une illustration. Beaucoup de travaux de modélisation de COVID 19 prennent alors en compte une structuration en classes d'âge [3, 14].

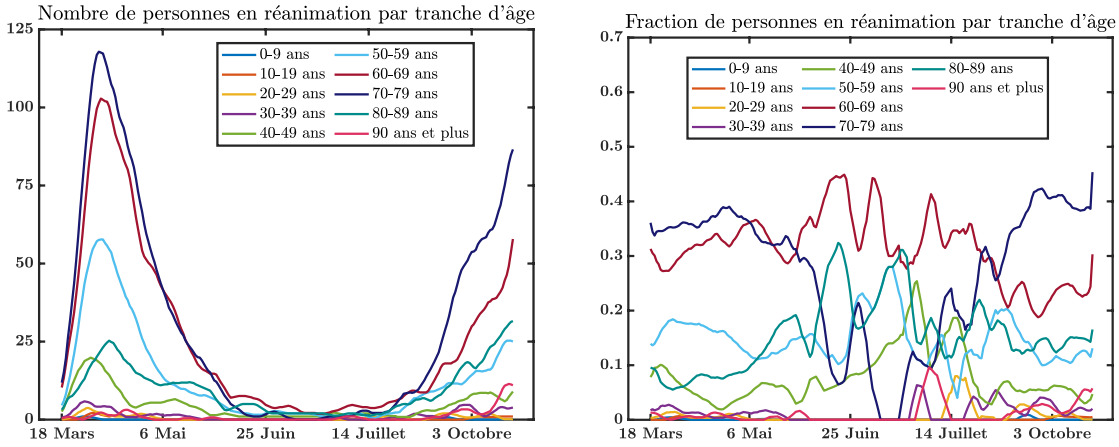

FIGURE 8 – Gauche : nombre de personnes en réanimation par tranche d'âge. Droite : fraction de personnes en réanimation par tranche d'âge. On observe une très forte hétérogénéité parmi les personnes en réanimation en fonction de leur âge. Environ 65% des personnes en réanimation ont entre 60 et 80 ans.

3.4.4 Modèles probabilistes

Quand le nombre de compartiments est grand ou dans le cas de forte hétérogénéité, la résolution numérique du modèle peut ne pas être aisée. Il convient alors de se tourner vers des modèles probabilistes et d'une approche Monte Carlo. Dans un tel modèle, chaque compartement C contient alors un nombre aléatoire $N_C(t)$ d'individus [3] au temps t. Quand un individu entre dans un compartiment au temps t, on tire alors une variable aléatoire T_C d'une certaine loi fixée en fonction du compartiment, et l'individu quittera alors ce compartiment au temps $t + T_C$. On adapte si plusieurs compartiments de sortie sont possibles.

Dans ce qui suit, nous allons prendre comme exemple un modèle probabiliste simple. On souhaite étudier l'évolution du nombre de personnes infectées partant d'une seule personne infectée au temps 0. On suppose qu'on observe l'épidémie pendant un temps relativement court, pendant lequel le nombre d'infectés est négligeable devant la taille de la population. Par conséquent, on peut supposer que le nombre de susceptibles est constant. On considère alors le modèle suivant : un individu qui contracte le virus à un temps t le transmets à un nombre aléatoire K de personnes aux temps aléatoires $t + \tau_1, t + \tau_2, \ldots, t + \tau_K$. Le vecteur aléatoire (de longueur aléatoire) $(K, \tau_1, \tau_2, \ldots, \tau_K)$ suit la même loi pour tous les individus, c'est cette loi qui est le paramètre du modèle. Un tel modèle est connu sous le nom de *processus de branchement général* ou encore *processus Crump–Mode–Jagers* [4]. Bien sûr, ce modèle peut être étendu de plusieurs façons, par exemple en ajoutant des « compartiments » ou en prenant en compte une population hétérogène : cela mène à des processus de branchement *multitypes*. Dans ce qui suit, nous allons rester dans le cadre le plus simple.

On définit la fonction

$$F(t) = \mathbb{E}\left[\sum_{i=1}^{K} \mathbf{1}_{\tau_i \leq t}\right],$$

où le symbole $\mathbb{E}$ note l'espérance. La fonction F est la fonction de répartition d'une mesure

3. Ces individus ne représentent pas forcément de vrais individus de la population, mais peuvent représenter un « paquet » d'individus ; le but étant simplement de discrétiser la population au lieu la considérer comme une masse homogène.

sur $\mathbb{R}_+$ et on suppose pour simplifier que cette mesure est absolument continue par rapport à Lebesgue, c'est-à-dire qu'il existe une fonction f telle que

$$F(t) = \int_0^t f(s)\,ds.$$

Notons que

$$F(\infty) = \int_0^\infty f(t)\,dt = \mathbb{E}[K],$$

c'est le fameux paramètre $\mathscr{R}_0$. Cette notation vient du fait que $\mathscr{R}_0$ est un cas particulier du paramètre $\mathscr{R}_a$ définit ainsi :

$$\mathscr{R}_a = \int_0^\infty t^a f(t)\,dt.$$

Pour ce modèle, dès lors que $\mathscr{R}_0 > 1$, on s'attend à ce que la taille du nombre d'infectés dans la population tende exponentiellement vite vers l'infini, dans le cas où l'épidémie ne s'éteint pas au tout début :

$$\exists \alpha > 0 : \frac{\log X(t)}{t} \to \alpha, \quad t \to \infty,$$

où $X(t)$ note le nombre de personnes ayant contracté le virus avant le temps t. Ceci est en effet vrai en grande généralité. Le paramètre α est appelé le paramètre *malthusien* du modèle et décrit la vitesse de croissance exponentielle. Il se calcule de la façon suivante : c'est le nombre qui satisfait à l'identité

$$\int_0^\infty e^{-\alpha t} f(t)\,dt = 1.$$

Notons que α est également défini si $\mathscr{R}_0 < 1$, dans quel cas $\alpha < 0$: le nombre de personnes infectées décroit exponentiellement vite et l'épidémie s'éteint au bout d'un moment.

Exemple. Prenons f une fonction exponentielle,

$$f(t) = \beta e^{-\nu t},$$

Cet exemple permet de faire le lien avec le modèle SIR de la section 3.1.1. Car supposons qu'un infecté reste infecté pendant un temps aléatoire suivant la loi exponentielle de paramètre ν, et que pendant ce temps-là il infecte d'autres personnes à taux β. Formellement, on construit donc $\tau_1, \ldots, \tau_K$ de la façon suivante : on part d'un *processus de Poisson* d'intensité β et on l'arrête au bout d'un temps aléatoire de loi exponentielle de paramètre ν. On définit $\tau_1, \ldots, \tau_K$ comme étant les instants de sauts de ce processus. Alors on peut montrer que f est exactement de la forme ci-dessus. De plus, l'espérance du nombre d'infectés au temps t suit exactement l'équation différentielle satisfaite par $I(t)$ dans le modèle SIR (avec $S(t)$ remplacé par N, ce qui est une approximation valable au début de l'épidémie). Il s'agit donc d'un vrai analogue probabiliste du modèle SIR.

On calcule aisément $\mathscr{R}_0 = \beta/\nu$; c'est à quoi nous nous attendions en vue de (3.2). On peut également calculer le paramètre malthusien, car

$$\int_0^\infty e^{-\alpha t} f(t)\,dt = \frac{\beta}{\nu + \alpha},$$

et donc $\alpha = \beta - \nu$.

4 Résolution numérique

Dans cette section, nous allons présenter quelques résultats numériques concernant le modèle suivant :

$$\begin{cases} S'(t) = -\tau(t)S(t)\left[I_s(t) + I_a(t)\right], \\ E'(t) = \tau(t)S(t)\left[I_s(t) + I_a(t)\right] - \alpha E(t), \\ I_s'(t) = f\alpha E(t) - \eta_s I_s(t), \\ I_a'(t) = (1-f)\alpha E(t) - \eta_a I_a(t), \\ M'(t) = f\alpha E(t) - \eta_m M(t), \\ H'(t) = \eta_m M(t) - \lambda_h H(t), \\ D'(t) = \lambda_d H(t) \end{cases} \tag{4.1}$$

où l'on rappelle que $S(t)$ est la population de personnes susceptibles d'être infectées, $E(t)$ est la population de personnes exposées, $I_s(t)$ est la population de personnes infectées qui nécessitera une prise en charge en soins intensifs tandis que $I_s(t)$ représente la population de toutes les autres personnes infectées. Ensuite, $M(t)$ consiste en la population de personnes malades qui nécessiteront d'être hospitalisées en soins intensifs et $H(t)$ est la population de personnes hospitalisées en soins intensifs. Enfin, $D(t)$ représente le nombre de personnes décédées à l'hôpital. On rappelle aussi que $\tau(t)$ varie au cours du temps suivant les stratégies publiques mises en place. Néanmoins, au tout début de l'épidémie (avant le confinement notamment), on peut partir du principe que $\tau(t) = \tau_0$ est constante.

4.1 Choix des conditions initiales

Le système (4.1) est accompagné d'une condition initiale de la forme

$$(S(t_0), E(t_0), I_s(t_0), I_a(t_0), M(t_0), H(t_0), D(t_0)) = (S^0, E^0, I_s^0, I_a^0, M^0, H^0, D^0),$$

pour un temps initial t_0 que l'on doit se donner. Nous allons travailler avec l'hypothèse que t_0 correspond à la date à laquelle il y a une première personne prise en charge en soins intensifs, c'est à dire :

$$H(t_0) = H^0 = 1.$$

Ensuite, nous allons supposer que S^0 correspond à la population totale en Occitanie, soit

$$S(t_0) = S^0 = 5.893 \times 10^6,$$

selon les chiffres du dernier recensement de 2019. Et enfin, on prendra $D(t_0) = D^0 = 0$. Il reste donc à estimer (E^0, I_s^0, I_a^0, M^0) ainsi que t_0. Nous allons partir du principe qu'au tout début de l'épidémie, les effets nonlinéaires du système (4.1) sont suffisamment faibles pour être négligés en première approximation et que l'on suit un comportement exponentiel qui s'observe aussi dans les données. Et donc, si l'on suppose que $H(t) = H^0 e^{\chi(t-t_0)} = e^{\chi(t-t_0)}$ puisque $H(t_0) = H^0 = 1$, on pourra estimer χ et t_0 directement à partir des données. Si $H(t)$ a un comportement exponentiel, alors à partir de (4.1), on en déduit que

$$M(t) = \frac{\chi + \lambda_h}{\eta_m} e^{\chi(t-t_0)},$$

et donc par identification on aura

$$M(t_0) = M^0 = \frac{\chi + \lambda_h}{\eta_m}.$$

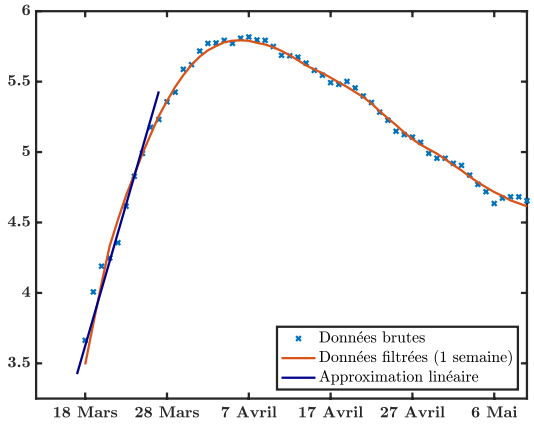

FIGURE 9 – Nombre de personnes hospitalisées en soins intensifs en échelle logarithmique. L'approximation linéaire (courbe bleu foncé) a été faite entre les données du 18 et 27 Mars, donnant une pente de 0.2007 et une ordonnée à l'origine de 30.6736 permettant d'inférer une date approximative du premier cas de personne hospitalisée en soins intensifs autour du 1er Mars.

En reproduisant en cascade le même argument, on obtient

$$E(t_0) = E^0 = \frac{\chi + \eta_m}{f\alpha} M^0, \quad I_a(t_0) = I_a^0 = \frac{(1-f)\alpha}{\chi + \eta_a} E^0, \quad \text{et} \quad I_s(t_0) = I_s^0 = \frac{f\alpha}{\chi + \eta_s} E^0.$$

Pour les données à notre disposition (voir la Figure 9), nous arrivons à estimer χ et t_0 comme étant :

$$\chi \simeq 0.2007 \quad \text{et} \quad t_0 \simeq -17.0573.$$

Ce qui veut dire qu'on peut estimer que le premier patient en réanimation est arrivé le 1er Mars. Dans la pratique, nous allons résoudre (4.1) sur un intervalle de temps $[0, T_f]$ où $T_f > 0$ est le temps final, en partant du principe que l'instant initial $t = 0$ correspond au 1er Mars.

4.2 Choix des valeurs des paramètres

Il nous faut maintenant choisir les valeurs des différents paramètres de notre système (4.1). La première étape consiste à évaluer τ_0 de la façon suivante. En reprenant l'approximation exponentielle faite précédemment, nous avons vu que $E(t)$, $I_{s,a}(t)$ étaient notamment approchées par des exponentielles avec

$$E(t) = E^0 e^{\chi(t-t_0)}, \quad I_s(t) = I_s^0 e^{\chi(t-t_0)}, \quad \text{et} \quad I_a(t) = I_a^0 e^{\chi(t-t_0)}.$$

En injectant ces expressions dans la seconde équation de notre système (4.1) et en approchant $S(t)$ par S^0, on trouve que

$$\tau_0 = \frac{\chi + \alpha}{S^0(I_s^0 + I_a^0)} E_0.$$

Il nous reste donc à donc à fixer les paramètres suivants : α, f, $\eta_{s,a,m}$ et $\lambda_{h,d}$. Certains paramètres sont des données biologiques de la maladie, d'autres des données cliniques.

4.3 Choix de schémas numériques

Nous avons à notre disposition toute la panoplie des schémas numériques pour résoudre des systèmes d'équations différentielles, et nous avons fait le choix de présenter ici un schéma

numérique semi-implicite qui assure la positivité des solutions pour chaque itération et ce, indépendamment du pas de temps choisi. Soit donc $\delta t > 0$ un pas de temps fixé et notons $t_n = t_0 + n\delta t$ pour chaque $n \geq 0$. Nous noterons également $(S^n, E^n, I_s^n, I_a^n, M^n, H^n, D^n)$ l'approximation numérique au temps t_n de la solution du système (4.1) à partir de la donnée initiale $(S^0, E^0, I_s^0, I_a^0, M^0, H^0, D^0)$.

Le schéma numérique s'écrit comme suit :

$$\begin{cases} S^{n+1} = S^n - \delta t \tau(t_n) S^{n+1} \left[I_s^n + I_a^n\right], \\ E^{n+1} = E^n + \delta t \tau(t_n) S^{n+1} \left[I_s^n + I_a^n\right] - \delta t \alpha E^{n+1}, \\ I_s^{n+1} = I_s^n + \delta t f \alpha E^{n+1} - \delta t \eta_s I_s^{n+1}, \\ I_a^{n+1} = I_a^n + \delta t (1-f) \alpha E^{n+1} - \delta t \eta_a I_a^{n+1}, \\ M^{n+1} = M^n + \delta t f \alpha E^{n+1} - \delta t \eta_m M^{n+1}, \\ H^{n+1} = H^n + \delta t \eta_m M^{n+1} - \delta t \lambda_h H^{n+1}, \\ D'^{n+1} = D^n + \delta t \lambda_d H^{n+1} \end{cases} \tag{4.2}$$

pour chaque $n \geq 0$. Il est facile de voir que la première équation donne

$$S^{n+1} = \frac{S^n}{1 + \delta t \tau(t_n) \left[I_s^n + I_a^n\right]},$$

qui une fois reportée dans la seconde donne

$$E^{n+1} = \frac{E_n}{1 + \delta t \alpha} + \frac{\delta t \tau(t_n) S^n \left[I_s^n + I_a^n\right]}{1 + \delta t \tau(t_n) \left[I_s^n + I_a^n\right]},$$

et ainsi de suite. On obtient donc facilement la positivité des solutions du schéma numérique (4.2) en partant d'une donnée initiale $(S^0, E^0, I_s^0, I_a^0, M^0, H^0, D^0)$ à composantes positives, et ce indépendamment du choix de $\delta t > 0$.

4.4 Résultats et comparaison aux données

Pour simplifier la présentation, on choisit de travailler avec une fonction $\tau(t)$ seulement continue par morceaux de la forme

$$\tau(t) = \begin{cases} \tau_0, & t \in [0, T_c), \\ \theta_c \tau_0, & t \in [T_c, T_d), \\ \theta_d \tau_0, & t \geq T_d, \end{cases}$$

où $0 < \theta_c < \theta_d < 1$ et $0 < T_c < T_d$. Ici, θ_c et θ_d représente l'efficacité des interventions des pouvoirs publics sur le taux de contact durant et après le confinement, et $T_{c,d}$ est le jour où ces mesures deviennent effectives. Il est important de directement remarquer que $T_{c,d}$ est un temps *effectif*. Le confinement est entré en vigueur en France le 16 Mars, et l'on pourrait donc légitimement prendre $T_c = 15$. En pratique cela ne fonctionne pas et pour reproduire les données (c'est à dire la bosse épidémique) on remarque que $T_c \sim 24.5$. Inversement, le déconfinement a eu lieu le 11 Mai soit 71 jours après le 1er Mars, et l'on se rend compte que $T_d \sim 65$ traduisant le fait que la population a certainement dû relâcher ses efforts un peu avant.

Pour le jeu de paramètres donnés dans le Tableau 1, on obtient des résultats qualitativement comparables aux jeux de données, et l'on arrive à reproduire plutôt fidèlement l'évolution au cours du temps du nombre de personnes en réanimation ainsi que le nombre de personnes décédées (voir la Figure 10). En fait, beaucoup de jeux de paramètres permettraient d'obtenir les mêmes résultats qualitatifs, et il est donc important de comprendre quels paramètres peuvent être identifiés (ou pas) à l'aide des données et comment notre modèle est sensible aux variations des paramètres. C'est tout l'enjeu de la section suivante.

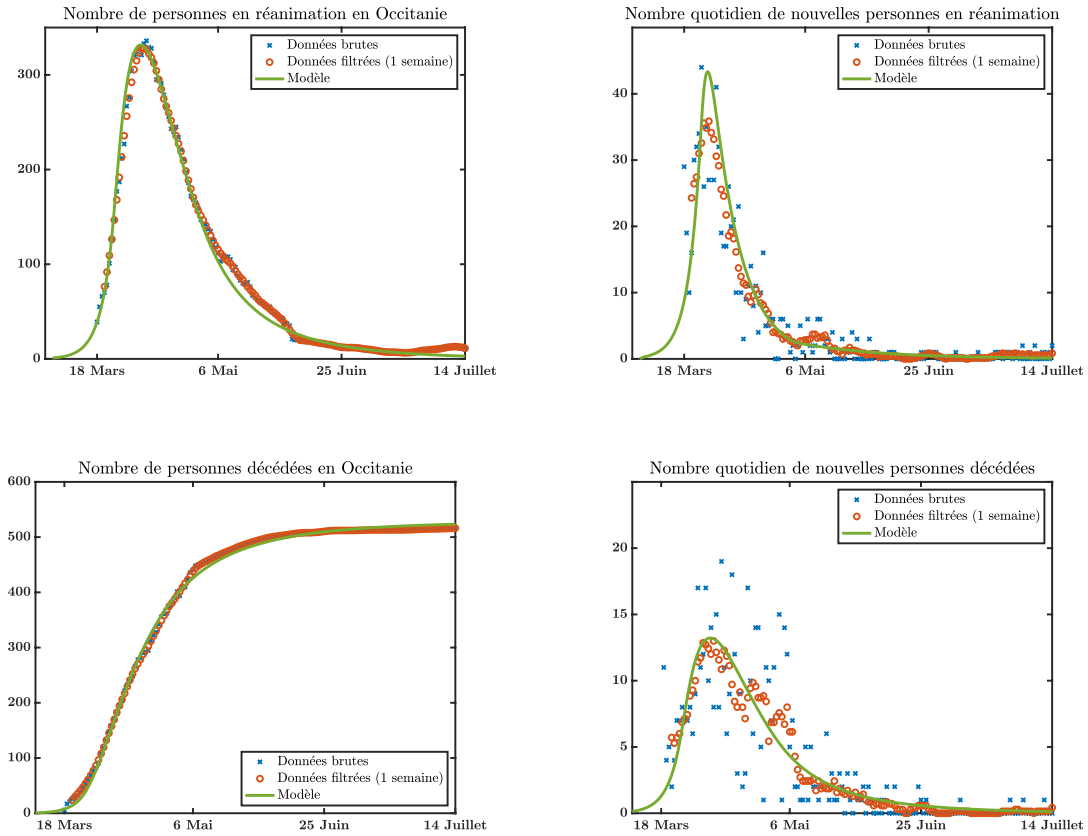

FIGURE 10 – Comparaison entre le modèle (4.1) et les données pour le jeu de paramètres du Tableau 1.

θ_c	θ_d	T_c	T_d	α	f	η_s	η_a	η_m	λ_h	λ_d
0.19	0.3	24.5	65	0.3649	0.1	0.2165	0.2718	0.2843	0.0694	0.04

TABLE 1 – Valeurs des paramètres utilisées pour produire la Figure 10.

5 Analyse de sensibilité et optimisation

L'assimilation de données consiste à exploiter des données mesurées pour estimer les paramètres d'un modèle constitué d'équations différentielles ordinaires, ou d'équations aux dérivées partielles. Les paramètres à estimer peuvent être des coefficients intervenant dans l'EDO/EDP, ou bien une condition initiale. Les résultats les plus connus d'assimilation de données sont les prévisions météo, mais de nombreuses autres applications existent. Parmi les méthodes employées, nous pouvons citer

- l'assimilation séquentielle (filtre de Kalman, théorie des observateurs de Luenberger)
- l'assimilation variationnelle, initié par Le Dimet-Talagrand. Nous allons détailler un peu plus les méthodes et les outils de cette dernière car c'est elle qui est mise en œuvre ici.

5.1 Assimilation de données variationnelle

Considérons une équation dépendant d'un certain nombre de paramètres, notés θ, qui peuvent être des conditions initiales et/ou des coefficients. La solution de l'équation pour la

valeur θ des paramètres est notée $u(\theta)$. On dispose par ailleurs d'observation pour une certaine valeur $\theta^\star$ des paramètres - c'est la "vraie" valeur que l'on souhaite estimer. Notons que l'on ne dispose pas nécessairement d'observations complètes de la solution, on dispose seulement de mesures obtenues à l'aide d'un opérateur d'observation que nous noterons L. On supposera pour simplifier que L est connu et linéaire. Les mesures dont on dispose sont par ailleurs entachées de bruit. Les observations dont on dispose y^{OBS} sont donc

$$y^{\text{OBS}} = Lu(\theta) + \eta,$$

où η désigne la réalisation du bruit de mesure.

Plaçons-nous dans le cas où le bruit de mesure est une réalisation d'une loi gaussienne $\mathcal{N}(0, \Sigma)$ centrée de matrice de covariance connue Σ. On note $p(\eta)$ la densité de cette loi [4], si bien que $p(\eta) \sim \exp(-\langle \eta | \Sigma^{-1}\eta \rangle / 2)$ avec $\langle \cdot | \cdot \rangle$ le produit scalaire euclidien et où le symbol « $\sim$ » signifie « proportionnel à » .

On adopte également le point de vue *bayésien*, selon lequel les paramètres θ sont elles-mêmes des variables aléatoires, avec comme interprétation que ceux-ci sont plus ou moins connus avec une certaine erreur. Nous supposons que cette erreur est également gaussienne, autrement dit que $\theta = \theta^0 + \rho$, où ρ est la réalisation d'une gaussienne centrée de matrice de covariance Σ_P.

On cherche la valeur la plus probable du paramètre θ, connaissant les observations. Pour cela utilisons la formule de Bayes :

$$p(\theta | y^{\text{OBS}}) = \frac{p(y^{\text{OBS}}|\theta)p(\theta)}{p(y^{\text{OBS}})}.$$

Pour trouver θ qui maximise $p(\theta|y^{\text{OBS}})$, on va minimiser le log du membre de droite. Comme y^{OBS} est donné, $p(y^{\text{OBS}})$ est une constante. On doit donc minimiser

$$\begin{aligned} -\log(p(y^{\text{OBS}}|\theta)p(\theta)) &= -\log p(y^{\text{OBS}}|\theta) - \log p(\theta) \\ &= \frac{1}{2}\langle Lu(\theta) - y^{\text{OBS}} | \Sigma^{-1}(Lu(\theta) - y^{\text{OBS}}) \rangle + \frac{1}{2}\langle \theta - \theta^0 | \Sigma_P^{-1}(\theta - \theta^0) \rangle + C. \end{aligned}$$

Dans le cas qui nous intéresse où l'on n'a pas d'a-priori sur θ, on considère la limite lorsque la matrice Σ_P^{-1} est nulle. On doit donc minimiser

$$j(\theta) = \frac{1}{2}\langle Lu(\theta) - y^{\text{OBS}} | \Sigma^{-1}(Lu(\theta) - y^{\text{OBS}}) \rangle. \tag{5.1}$$

Ceci n'est autre que la méthode des moindres carrés, mais le point de vue bayésien plus général sera utile par la suite.

Cette approche suppose que les bruits de mesure sont gaussiens et nécessite de connaître la matrice de covariance Σ. Si les mesures sont indépendantes, et la mesure i possède une variance σ_i^2, alors la matrice Σ est diagonale avec les coefficients σ_i^2. La formule (5.1) s'écrit alors

$$j(\theta) = \frac{1}{2}\sum_i \frac{((Lu(\theta)_i - y_i^{\text{OBS}})^2}{\sigma_i^2}.$$

En pratique on a affaire à un comptage (de cas, de décès, etc...) et un bruit de comptage est généralement modélisé par un bruit de Poisson. Comme un bruit de Poisson d'espérance σ possède une variance égale à σ - et donc un écart-type égal à $\sqrt{\sigma}$, nous estimons la matrice Σ par une matrice diagonale avec les coefficients y_i^{OBS}. La fonction coût à minimiser est donc

$$j(\theta) = \frac{1}{2}\sum_i \frac{((Lu(\theta)_i - y_i^{\text{OBS}})^2}{y_i^{\text{OBS}}}. \tag{5.2}$$

4. On adopte ici la convention d'utiliser le même symbole pour la variable aléatoire et l'argument de la fonction de densité.

5.2 Résolution du problème d'optimisation

On cherche à minimiser une fonction objectif du type moindre carré, qui peut s'écrire

$$j(\theta) = \frac{1}{2}\|F(\theta)\|^2,$$

avec

$$F(\theta)_i = \frac{Lu(\theta)_i - y_i^{\mathrm{OBS}}}{\sqrt{y_i^{\mathrm{OBS}}}}. \tag{5.3}$$

En général, le terme $Lu(\theta)$ est non-linéaire en θ, si bien que la solution n'est pas explicite. Un algorithme numérique efficace est la méthode de Gauss-Newton. C'est une méthode itérative donnée par l'algorithme ci-dessous.

Algorithme de Gauss-Newton pour minimiser $\frac{1}{2}\|F(\theta)\|^2$

1. initialisation : θ^0 donné
2. itérations : $\theta^{k+1} = \theta^k + d^k$, où d^k est solution de

$$DF^T DF d^k = -(DF^T)F.$$

NB : dans cette formule il faut comprendre $DF = DF(\theta^k)$.

On doit donc calculer la jacobienne DF pour F donné par (5.3). On détermine cette jacobienne colonne par colonne, en déterminant la dérivée de $u(\theta)$ par rapport à chaque paramètre θ_j, ceci est décrit plus bas. On a donc formellement

$$DF_{i,j} = \frac{1}{\sqrt{y_i^{\mathrm{OBS}}}} L\partial_j u(\theta).$$

La méthode de Gauss-Newton possède l'avantage d'avoir une convergence rapide (quadratique) au voisinage de l'optimum, sous certaines hypothèses de régularité. Par contre elle peut ne pas converger si le point de départ se situe trop loin du minimum. Pour éviter cet inconvénient, nous avons implémenté une variante, la méthode de Levenberg-Marquardt, pour laquelle le vecteur de mise à jour est solution de

$$(DF^T DF + \lambda Id)d^k = -(DF^T)F.$$

Le paramètre λ est choisi selon une heuristique mais d'autres choix sont possibles.

5.3 Modèle tangent

On souhaite évaluer la dérivée de u par rapport au j-ième paramètre θ_j, que nous noterons $\partial_j u$. Dans notre cas u est solution du problème

$$\partial_t u = \Phi(\theta, u), \qquad u(t=0) = u_0.$$

La fonction Φ décrit le modèle EDO de l'équation (4.1). Dans notre cas Φ est dérivable par rapport au j-ième paramètre θ_j. Comme l'égalité ci-dessus est valable pour tout t on peut dériver par rapport à θ_j et on obtient le modèle tangent dont la solution est $v = \partial_j u$:

$$\partial_t v = D_1\Phi(\theta, u)e_j + D_1\Phi(\theta, u)v, \qquad v(t=0) = 0.$$

Dans cette écriture, $D_1\Phi(\theta, u)e_j$ désigne la dérivée de $\Phi(\theta, u)$ par rapport au paramètre θ_j. Notons que le modèle tangent est linéaire.

5.4 Sensibilité

Notons $\theta^\dagger$ la valeur du paramètre qui réalise le minimum de (5.2), et S la matrice de sensibilité dont la colonne j est le vecteur $S_j = L\partial_j u(\theta)$.

Nous allons vérifier que les paramètres θ suivent une loi gaussienne centrée en $\theta^\dagger$ et dont nous allons déterminer la matrice de covariance. La condition d'optimalité au premier ordre de $\theta^\dagger$ dans (5.1) peut s'exprimer par le fait que pour tout j :

$$\langle S_j | \Sigma^{-1}(Lu(\theta^\dagger) - y^{\mathrm{OBS}})\rangle = 0.$$

Par ailleurs nous avons vu que pour une certaine constante K, dans le cas où l'on ne dispose pas d'a-priori sur le paramètre θ on a

$$-\log p(\theta) = K + \frac{1}{2}\|Lu(\theta) - y^{\mathrm{OBS}}\|^2_{\Sigma^{-1}}.$$

On approxime $Lu(\theta) - y^{\mathrm{OBS}}$ par son développement d'ordre 1 au voisinage de $\theta = \theta^\dagger$:

$$Lu(\theta) - y^{\mathrm{OBS}} = Lu(\theta^\dagger) - y^{\mathrm{OBS}} + S(\theta - \theta^\dagger) + o(\|\theta - \theta^\dagger\|).$$

$$\begin{aligned}\frac{1}{2}\|Lu(\theta^\dagger) - y^{\mathrm{OBS}} + S(\theta - \theta^\dagger)\|^2_{\Sigma^{-1}} &= \frac{1}{2}\|Lu(\theta^\dagger) - y^{\mathrm{OBS}}\|^2_{\Sigma^{-1}} + \frac{1}{2}\langle S(\theta - \theta^\dagger)|\Sigma^{-1}(Lu(\theta) - y^{\mathrm{OBS}})\rangle + \\ &\quad \frac{1}{2}\|S(\theta - \theta^\dagger)\|^2_{\Sigma^{-1}}.\end{aligned}$$

Le premier terme est constant, et le deuxième terme est nul par optimalité de $\theta^\dagger$. On considère l'approximation quadratique de $-\log p(\theta)$ donnée par le troisième terme, qui est

$$\frac{1}{2}\|S(\theta - \theta^\dagger)\|^2_{\Sigma^{-1}} = \frac{1}{2}\|\theta - \theta^\dagger\|^2_{S^T\Sigma^{-1}S}.$$

Cette approximation exprime que θ suit approximativement une gaussienne centrée en $\theta^\dagger$ dont l'inverse de la matrice de covariance est

$$S^T\Sigma^{-1}S.$$

Les grandes valeurs propres de cette matrice correspondent aux combinaisons linéaires de paramètres qui peuvent être estimées avec une "faible incertitude" par les mesures disponibles. Et inversement, les petites valeurs propres de cette matrice correspondent aux combinaisons linéaires de paramètres qui peuvent être estimées de façon peu fiable par les mesures disponibles.

6 Lien vers les ressources numériques

Références

[1] Yuhao Deng, Chong You, Yukun Liu, Jing Qin, and Xiao Hua Zhou. Estimation of incubation period and generation time based on observed length-biased epidemic cohort with censoring for COVID-19 outbreak in China. *Biometrics*, (June) :1–13, 2020.

[2] J. Dolbeault and G. Turinici. Heterogeneous social interactions and the covid-19 lockdown outcome in a multi-group seir model. *Math. Model. Nat. Phenom.*, 15 :36, 2020.

[3] L. D. Domenico, G. Pullano, C. E. Sabbatini, P.-Y. Boëlle, and V. Colizza. Expected impact of lockdown in île-de-france and possible exit strategies. *BMC Med.*, 18(1) :240, 2020.

[4] Theodore E Harris. *The theory of branching processes*, volume 119 of *Die Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1963.

[5] W. Kermack and A. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772) :700–721, 1927.

[6] Zhihua Liu, Pierre Magal, Ousmane Seydi, and Glenn Webb. Understanding unreported cases in the covid-19 epidemic outbreak in wuhan, china, and the importance of major public health interventions. *Biology*, 9(3) :50, 2020.

[7] Zhihua Liu, Pierre Magal, and Glenn Webb. Predicting the number of reported and unreported cases for the covid-19 epidemics in china, south korea, italy, france, germany and united kingdom. *Journal of Theoretical Biology*, 509 :110501, 2021.

[8] H. Nishiura, N. M. Linton, and R. A. Andrei. Serial interval of novel coronavirus (covid-19) infections. *International Journal of Infectious Diseases*, 93 :284–286, 2020.

[9] Q. Richard, S. Alizon, M. Choisy, M. T. Sofonea, and Djidjou-Demasse R. Age-structured non-pharmaceutical interventions for optimal control of covid-19 epidemic. 2020.

[10] Li, Q., et al. Early transmission dynamics in wuhan, china, of of novel coronavirus ?infected pneumonia. *New England Journal of Medicine*, 382(13) :1199–1207, 2020.

[11] N. M. Linton, et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation : A statistical analysis of publicly available case data. *J Clin Med.*, 9(2) :538, 2020.

[12] S. A. Lauer, et al. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases : Estimation and application. *Annals of internal medicine*, 172(9) :577–582, 2020.

[13] S. Stringhini, et al. Seroprevalence of anti-sars-cov-2 igg antibodies in geneva, switzerland (serocov-pop) : a population-based study. *The Lancet*, 396 :313–319, 2020.

[14] Salje, H. et al. Estimating the burden of sars-cov-2 in france. *Science*, 369(6500) :208–211, 2020.

[15] Z. Du, et al. Serial interval of covid-19 among publicly reported confirmed cases. *Emerg Infect Dis.*, 26(6) :1341–1343, 2020.