



## Modèles

17 juin 2020

### Résumé

Les modèles dits compartimentaux, décrits par des équations différentielles, ont été à la base de bien des analyses récentes de la pandémie de covid-19. Ils peuvent être précisés pour prendre en compte diverses données de la maladie (hétérogénéité des populations, possible saisonnalité). De manière générale, après traitement statistique des données médicales, la calibration des paramètres permet aux modèles mathématiques de prédire et de contrôler la propagation de l'épidémie. Pour dépasser les limitations de ces modèles qui supposent de grandes populations, des modèles qui prennent en compte des aspects géographiques et aléatoires ont été développés plus récemment.

Les modèles mathématiques sont largement utilisés dans les sciences de l'ingénieur, la mécanique, la physique ou la chimie. L'efficacité des simulations numériques en font un enjeu majeur dans ces sciences. Qu'en est-il du domaine de la propagation d'épidémies ?

### 1- Modèles et prédiction

L'idée de quantifier l'effet d'une épidémie n'est pas nouvelle ; dans un Mémoire célèbre de 1760, Daniel Bernoulli analyse "la mortalité causée par la petite vérole et les avantages de l'inoculation pour la prévenir". Le formalisme a progressé et a conduit aux modèles compartimentaux qui occupent aujourd'hui une place centrale dans la compréhension des données, la prévision et le contrôle de l'épidémie de covid-19.

**SIR : prédire avec peu de données.** De ces modèles compartimentaux, le plus simple, qui décrit en particulier l'initiation de l'épidémie, prend en compte l'évolution du nombre d'individus sains  $S(t)$ , non immunisés, les infectants  $I(t)$  (les infectés isolés ne comptent pas ici) et une population retirée  $R(t)$  (rétablis et immunisés ou décédés). On utilise alors deux paramètres :  $\beta$ , qui désigne le taux de contacts (infectant passant le virus à un individu sain) et  $\tau_I$ , la durée du stade infectant. Une fois le système adimensionné (on parle donc de proportions et  $S(0) = 1$ ) on obtient :

$$\begin{cases} \frac{dS}{dt} = -\beta IS & \text{(Sains)} \\ \frac{dI}{dt} = \beta IS - \frac{I}{\tau_I} & \text{(Infectants)} \\ \frac{dR}{dt} = \frac{1}{\tau_I} I & \text{(Retirés)} \end{cases}$$

Les paramètres  $\beta$  et  $\tau_I$  du système SIR se calibrent facilement à partir de deux nombres, le nombre de reproduction  $R_0 = \beta \tau_I$  représentant le nombre de malades qu'un infectant produira (il doit être inférieur à 1



pour que l'épidémie s'arrête) et une échelle de temps (souvent on donne le temps de doublement). Pour que le nombre d'infectants commence à diminuer ( $dI/dt \leq 0$ ), il devra rester une proportion  $S(t) \leq 1/R_0$ . La fraction de la population susceptible qui devra avoir été en contact avec le virus est donc :

$$f_{SI} = 1 - 1/R_0.$$

On parle alors d'immunité de groupe. Par exemple, pour SARS-CoV-2, si on utilise  $R_0 \approx 3$  on trouve que  $f_{SI} \approx 0.66$ . Mais si on ne prend pas de mesure, comme on peut le calculer et le voir sur la Figure 1, ce sera plus de 80% de la population saine qui aura finalement été infectée. Cette constatation a conduit les gouvernements à prendre des mesures drastiques.

Les limitations d'un système aussi simple ont très vite été pointées et dès 1927, W. Kermack and A. McKendrick proposaient un formalisme d'équations aux dérivées partielles pour prendre en compte la durée d'infection. Néanmoins, les modèles compartimentaux restent la base des prévisions actuelles et ils sont raffinés de multiples façons suivant les situations considérées.

**SEIR, latence et durée de contagiosité.** Il faut toutefois distinguer la phase de contagiosité d'une période de latence (exposée) de durée  $\tau_E$ , où la charge virale des infectés, dont le nombre est noté  $E$ , est très faible. On peut alors négliger la capacité à infecter, ce qui mène à un modèle SEIR s'écrivant :

$$\left\{ \begin{array}{l} \frac{dS}{dt} = -\beta SI, \\ \frac{dI}{dt} = \frac{1}{\tau_E} E - \frac{1}{\tau_I} I, \end{array} \right. \quad \boxed{\begin{array}{l} \frac{dE}{dt} = \beta SI - \frac{1}{\tau_E} E, \\ \frac{dR}{dt} = \frac{1}{\tau_I} I. \end{array}}$$

La phase de latence  $E$  ne change pas la valeur de  $R_0 = \beta\tau_I$ , mais elle affecte de façon significative l'échelle de temps du développement de l'épidémie et la prédiction du pic d'infection.

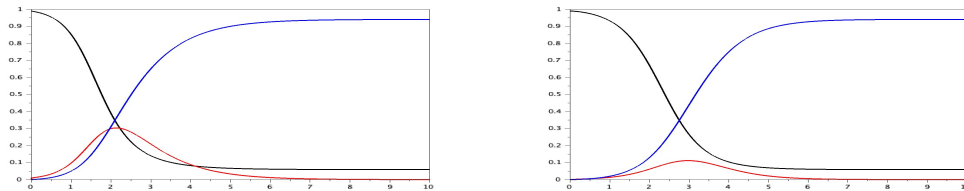


Figure 1 – Solutions du système SIR (gauche avec  $\tau_I = 6$  jours) et SEIR (droit avec  $\tau_E = 5$  jours,  $\tau_I = 2$  jours). On a pris  $R_0 = 3$ . Noir= sains, rouge= infectés, bleu= retirés. Échelle de temps= 1 semaine.

**Hétérogénéité et matrice de contact.** Considérer un nombre de contacts effectifs et une valeur  $R_0$  qui soient communs à tous est aussi trop grossier pour une population inhomogène. Pour améliorer le modèle, on peut ajouter des compartiments dans les modèles et introduire des classes d'âges (ou plus généralement de contacts) en utilisant une matrice de contact  $\beta_{ij}$  pour écrire :

$$\frac{d}{dt} S_i(t) = - \sum_j \beta_{ij} I_j S_i, \quad \frac{d}{dt} I_i(t) = S_i \left[ \sum_j \beta_{ij} I_j - \tau_I I_i \right]$$



L'accès aux paramètres de tels modèles a été un enjeu critique puisqu'ils ont été utilisés pour définir différents types de scénarios de sortie de confinement. Récemment, ils ont aussi été utilisés pour estimer l'impact de l'hétérogénéité sur l'immunité de groupe, M. G. M. [Gomes](#) et al.

**Saisonnalité.** La possible saisonnalité de SARS-CoV-2 est actuellement envisagée avec des crises jusqu'en 2022, voire pire jusqu'en 2024 (S. M. Kissler *et al.* Science 368, 2020). La prise en compte de la saisonnalité se traduit par un coefficient  $\beta(t)$  périodique de période un an.

## 2- Données et calibration : un enjeu majeur

Un modèle mathématique d'épidémie est une abstraction qui rend compte de la diffusion de l'infection en s'appuyant, à un moment donné, sur des hypothèses reflétant notre compréhension des principaux mécanismes en jeu. Une fois le modèle mathématique construit, la fiabilité des prédictions dépend de la qualité des coefficients estimés, eux-mêmes déduits des données de diffusion de l'épidémie. Deux difficultés se posent ici, l'accès aux données et la calibration du modèle.

**Donner du sens aux observations : l'exemple du nombre de reproduction.** L'accès aux données pose de nombreuses difficultés qui nécessitent diverses analyses expertes : accès restreint, données bruitées, partielles ou fausses, comme on l'a vu sur les EHPAD. Ici, l'[Institut Pasteur](#) et les instituts de Santé Publique [IPL](#) jouent évidemment un rôle majeur et réalisent les analyses statistiques nécessaires.

Un exemple est la détermination du nombre de reproduction introduit ci-dessus. Valant  $R_0$  dans la phase initiale, sa valeur évolue avec la diminution du réservoir des personnes contaminables  $S(t)$  et les mesures de limitations des contacts. L'estimation de sa valeur courante  $R(t)$  est très délicate et peut être envisagée de plusieurs façons. Partant d'un modèle compartimental, on peut inférer le paramètre  $\beta(t)$  instantané sur la base d'un échantillonnage généralement journalier. Plus directement, l'observation d'incidence  $z(t)$  des nouveaux cas peut être considérée comme résultant d'une distribution aléatoire poissonnienne de paramètre

$$R(t) \sum_{s=1}^{\infty} \Phi(s) z(t-s)$$

mettant en jeu les incidences passées et une fonction  $\Phi(t)$  représentant la distribution de l'*intervalle sériel* qui mesure le temps séparant l'apparition de symptômes entre infectant et infecté. En supposant que soit connue la distribution  $\Phi$ , l'estimation de  $R(t)$  peut alors être effectuée par toute approche, par exemple bayésienne, visant à maximiser la vraisemblance attachée au modèle.

**Calibration du modèle.** La calibration des paramètres sur les données est un problème qui se retrouve dans de nombreux domaines. Les méthodes de moindres carrés et maximum de vraisemblance sont d'un usage très répandu. Toutefois, il a été montré que, sur un modèle SEIR adapté à la covid-19, la donnée journalière des nouveaux cas dans la phase exponentielle de l'épidémie ne permet pas de prédire le pic épidémique. Même en mettant en œuvre les méthodes de l'état de l'art de la statistique, on pouvait trouver des variations d'un facteur 4 sur la prévision du nombre de décès. Plus précisément, il semble impossible d'estimer par une méthode inférentielle (qu'elle soit bayésienne ou fréquentiste) un paramètre important, la proportion de malades symptomatiques non-comptabilisés. Pour cela, il aurait fallu disposer de données supplémentaires,

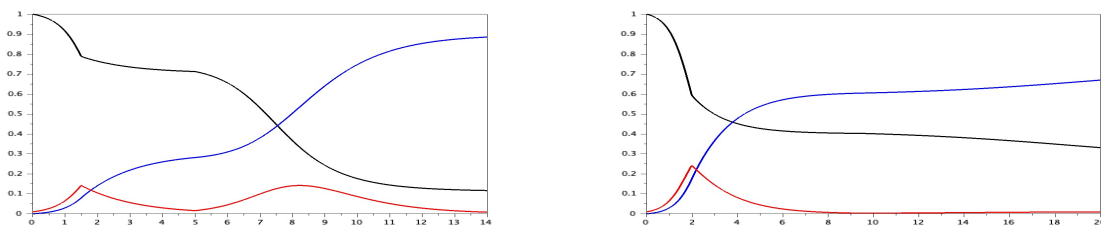
comme des tests sérologiques dans des sous-ensembles représentatifs de la population, ce qui n'a pas pu être fait en France au mois de mars. Devant la disparité des données journalières, certains auteurs ont jugé plus fiable d'utiliser le nombre d'hospitalisations ou de décès en hôpital.

**Sensibilité aux paramètres.** Si certains paramètres peuvent être obtenus par observation médicale, par exemple la durée de latence ou d'infectiosité, et d'autres par les méthodes statistiques, il demeure une difficulté cruciale. Les modèles d'épidémiologie décrivant des croissances exponentielles, les paramètres déterminant cette croissance sont très sensibles. La validation des conclusions nécessite donc que les paramètres utilisés, dans leur intervalle de valeurs possibles, ne changent pas significativement le résultat. Valider ces conclusions repose donc sur des analyses intensives de sensibilité, ce qui explique que l'utilisation des modèles compartimentaux nécessite finalement des temps de calcul importants.

### 3- Rebond, mouvements, modèles sur réseaux

**Contrôle de l'épidémie.** Un modèle permet aussi de définir les meilleurs moyens de contrôler le système. Par exemple, revenant au modèle SIR, les paramètres  $\tau_i$  sont purement physiologiques et les seuls *contrôles* se situent au niveau des taux de contacts  $\beta$  et des mouvements de populations. Les différentes mesures de confinement ou déconfinement se traduisent en valeurs des coefficients des matrices de contact  $\beta_{ij}(t)$ . Le contrôle optimal vise à choisir ces matrices pour maximiser, par exemple, les effets sur l'économie et le bien-être social avec la contrainte de garder un nombre d'infectés compatible avec les capacités hospitalières. Ces questionnements semblent nouveaux dans la théorie du contrôle des modèles SIR ou SEIR, les travaux portant jusqu'ici sur les stratégies de vaccination ou de quarantaine. Souvent évoqué, le risque de rebond montre l'importance du sujet, voir Fig. 2.

**Modèles géographiques et mouvement de populations.** L'épidémie de SARS de 2002 a montré l'importance des transports internationaux dans la propagation de l'épidémie. Les modèles régionaux de type SEIR interconnectés prennent en compte ces migrations (D. Bichara, Bull. Math. Biol. 77, 2015). Lorsque l'épidémie se propage par un processus continu de convection-diffusion, on est amené à considérer des systèmes d'équations aux dérivées partielles.



**Figure 2** – Deux exemples extrêmes de contrôles du système SIR en diminuant le paramètre  $\beta$ . Gauche : on observe un rebond pour un contrôle trop court et la proportion de population infectée reste de l'ordre de 90%. Droite : un contrôle plus long mène à l'immunité de groupe. Nous utilisons les mêmes codes couleurs et échelles que la pour la Figure 1.

Dans la phase initiale de l'épidémie, la forte hétérogénéité spatiale est causée par des importations aléatoires de cas infectieux liées aux déplacements et par des événements locaux engendrant des contacts accrus dans des environnements confinés, très favorables à la dissémination virale. Pour prendre en compte cela, l'utilisation de modèles stochastiques dynamiques régionaux couplés semble être une voie plus prometteuse.



En effet, lorsque l'on modélise une région avec une taille de population ( $N \sim 10^5$  à  $10^6$ ) qui reste petite par rapport à la population totale d'un pays ( $\sim 10^7$  à  $10^9$ ), la population infectée est faible et les effets stochastiques apparaissent alors que les modèles compartimentaux ci-dessus supposent de grandes populations.

La combinaison de modèles présentant des fluctuations importantes requiert le plus souvent d'utiliser des méthodes d'assimilation séquentielle des données pour l'inférence. Les données, agrégées à une échelle régionale, peuvent être affectées d'erreurs dues par exemple à des changements de protocole de test, ce qui donne lieu *a posteriori* à des corrections ou redressements.

**Modèles probabilistes sur réseaux.** Les processus épidémiques sur des graphes infinis sont actuellement étudiés dans le cadre de la théorie des processus de contact initialisée par T.E. Harris en 1974. Chaque nœud du graphe est dans l'un des deux états infecté ou sain et passe de manière aléatoire du premier au second avec un taux constant, et du second au premier avec un taux proportionnel au nombre de voisins infectés. À côté des simulations informatiques, il existe des résultats théoriques sur l'existence d'une mesure invariante maximale (directement liée à la probabilité de survie de l'épidémie) et sur l'existence de transitions de phases, ce qui signifie, dans certains cas, qu'il faut un taux d'infection suffisamment grand pour que l'épidémie survive. Sur des graphes aléatoires hétérogènes infinis, ces études conduisent à des résultats surprenants ; par exemple qu'il est possible que l'épidémie survive avec une probabilité positive pour des valeurs arbitrairement faibles du taux d'infection, même à partir d'un seul nœud infecté. Les hétérogénéités du graphe ont donc un impact majeur sur la propagation de l'épidémie.

Cette approche se développe dans d'autres domaines, comme les réseaux informatiques pair-à-pair où les interactions entre deux agents jouent un rôle similaire à une rencontre sain-infecté.

## Conclusion

La gestion de la crise de covid-19 a montré l'importance des méthodes statistiques pour l'analyse de données et des modèles compartimentaux pour comprendre l'état de l'épidémie et prédire son évolution, en particulier sur le court terme. Malgré leurs défauts et parfois leur simplicité, ces approches ont permis de fournir des informations quantifiées visant à déterminer l'impact des scénarios de sortie de confinement. Des modélisations bien plus riches existent toutefois qui permettront sans doute, lorsque toutes les données seront analysées, de mieux expliquer cette crise et d'être mieux armés pour la prochaine.

*Cette fiche a été conçue et rédigée par la cellule de crise Coronavirus de l'Académie des sciences. Créée à l'initiative de Pascale Cossart, Secrétaire perpétuel de l'Académie, celle-ci réunit des académiciens experts du domaine : Jean-François Bach, Pierre Corvol, Dominique Costagliola, Pascale Cossart (coordinatrice), Patrick Couvreur, Olivier Faugeras, Olivier Gascuel, Daniel Louvard, Eric Moulines, Georges Pelletier, Benoît Perthame, Félix Rey, Philippe Sansonetti, Alain-Jacques Valleron.*

*Les informations qui figurent sur cette fiche ont été produites collectivement et sont susceptibles d'évoluer. Elles seront éventuellement réactualisées en fonction des avancées des connaissances scientifiques.*