Vous avez dit « modèle » ?

Pierre de Buyl

Lors de l'épidémie de coronavirus, le comité fédéral belge d'experts scientifiques a été le centre de l'attention médiatique et politique. Mais… comment les scientifiques peuvent-ils comprendre l'évolution d'une épidémie ? En utilisant ce qu'on appelle un « modèle », c'est-à-dire une représentation mathématique d'un phénomène. En variant les paramètres du modèle, on peut tester de façon théorique des situations nouvelles. Petit décryptage…

La plupart des théories scientifiques sont des modèles qui acquièrent un rôle plus ou moins important en fonction de leur validation par l'expérience. Un des exemples les plus célèbres est probablement la mécanique classique : les équations et principes énoncés par Newton sur le mouvement des corps.

On sait aujourd'hui que cette théorie doit être remplacée, selon les situations, par la mécanique quantique ou la relativité générale (ces deux théories n'étant pas utilisables simultanément). Mais dans son domaine de validité, la mécanique classique conserve toute son importance comme pour le guidage de fusées et de satellites et l'étude de la structure des protéines, par exemple.

Dans cet article, nous verrons d'abord le concept de modèle, pour ensuite considérer une illustration de modèle de propagation virale que certains chercheurs ont utilisé pour le coronavirus… afin, en conclusion, de revenir sur l'actualité.

Un modèle jouet de modèle

Le caractère simplifié d'un modèle est souvent identifié par l'adjectif « jouet ». L'intérêt de cette démarche n'est alors plus d'estimer des résultats quantitatifs, mais de comprendre un mécanisme d'action à des fins de recherche ou d'illustration pédagogique. Je vais donc prendre un tel modèle jouet comme premier objet d'étude.

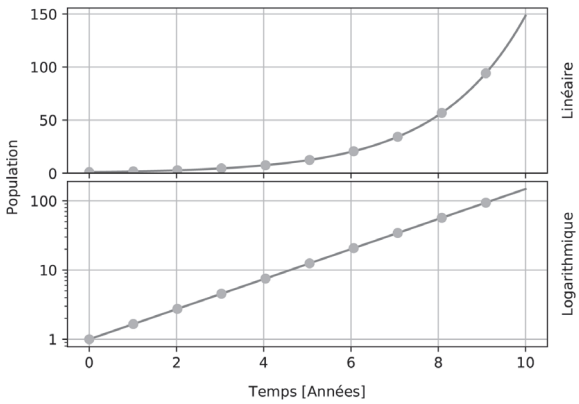
Concentrons-nous un moment sur le phénomène des naissances. Faisons l'hypothèse que le nombre de naissances à chaque instant soit proportionnel à la taille P de la population. On écrit alors la variation de la population comme le produit de P avec le nombre d'enfants de chaque individu. La solution de ce modèle est simple, il s'agit de la fameuse fonction exponentielle. On connait donc la taille de la population au cours du temps. On peut en déduire également le temps nécessaire au doublement de la population.

Quelles sont les limites de ce modèle ? *Primo*, il n'y a pas de valeur maximale : la population pourrait bien atteindre

un milliard de milliards de personnes ! *Secundo*, toute la population est considérée comme « identique », sans nuancer la variation de fertilité d'une région à une autre, par exemple.

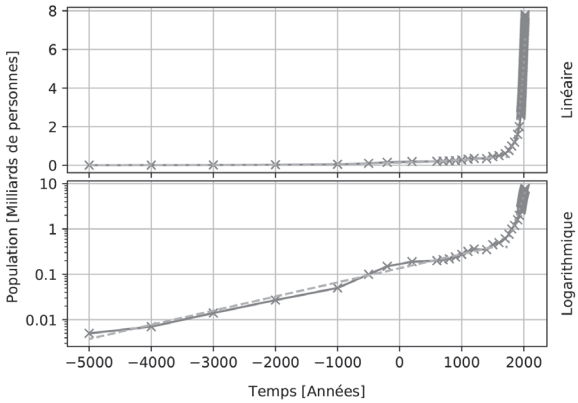
Certains enseignements du modèle sont transposables, par contre. D'une part, la définition même du taux de reproduction est utile pour construire des modèles plus élaborés. D'autre part, la formule mathématique pour la population se visualise clairement avec une ligne droite sur un graphique dit logarithmique. C'est pour cette raison que de nombreuses illustrations ont une échelle de données de type « 1, 10, 100, 1 000 » (logarithmique) plutôt que « 1, 2, 3, 4 » (linéaire). Je montre dans le graphique 1 la fonction exponentielle sur les deux types de graphiques.

Graphique 1



La population mondiale, pour de longues périodes historiques, a suivi une croissance approximativement exponentielle. Je montre les données dans le graphique 2, accompagnées d'ajustements de courbes exponentielles pour deux périodes temporelles. De l'origine des données en 5000 avant notre ère jusqu'environ 1600, on obtient un temps de doublement d'environ 1 000 ans et un taux d'environ 0,07 % par année. Après 1600, le taux monte à 0,9 % et le temps de doublement est de septante-quatre

Graphique 2



ans. Même si la population ne suit pas une croissance exponentielle, l'utilisation du graphique logarithmique et l'analyse en termes de taux de doublement aident à la compréhension des données.

Un modèle épidémiologique

La propagation d'une épidémie virale peut souvent être décrite par un modèle. Des chercheurs chinois ont utilisé le modèle dit « SEIR » qui compartimentalise la population en personnes susceptibles (S), exposées (E), infectées (I) et remises (R). Le modèle mathématique définit la variation du nombre de personnes dans les différentes catégories en donnant une expression pour le taux de variation (la dérivée temporelle) des variables S, E, I et R. Il faut compléter ce modèle par des paramètres qui correspondent à l'infection étudiée (délai d'incubation et de guérison, taux de reproduction). Dans la version simple du modèle, on omet le taux de décès naturels de la population, ce qui est une hypothèse cohérente pour de courtes épidémies.

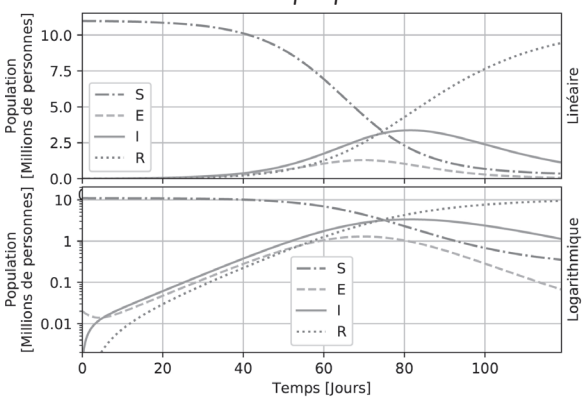
Au début de l'épidémie, la quasi-totalité de la population est considérée comme susceptible d'être exposée à la contamination venant des personnes infectées (l'infection prend place dans une deuxième étape). Le nombre de personnes exposées augmente donc proportion-

nellement à la taille du réservoir de personnes susceptibles d'être atteintes et au nombre de personnes infectées.

Une fois tout le modèle écrit, reste à résoudre les équations. Parfois on obtient une solution sous la forme d'une formule mathématique directe. Souvent, on utilise l'ordinateur pour obtenir une réponse numérique qu'on peut afficher et analyser en détail. Pour le modèle SEIR, j'ai réalisé différentes estimations de l'épidémie par résolution numérique. Motivé par la publication de Nicolas Vandewalle sur Twitter, j'ai repris la description dans l'article publié par l'équipe de l'école de santé publique de Shanghai Jiao (Tong University School of Medicine) dans lequel les scientifiques chinois estiment la valeur de plusieurs paramètres spécifiquement pour le Covid-19. Un des paramètres, le coefficient de reproduction R0, est fort dépendant des comportements humains (proximité, échanges sociaux, etc.) et est donc impossible à déterminer de façon universelle.

Je montre dans le graphique 3 les résultats du modèle à titre d'exemple, pour toutes les sous-populations.

Graphique 3

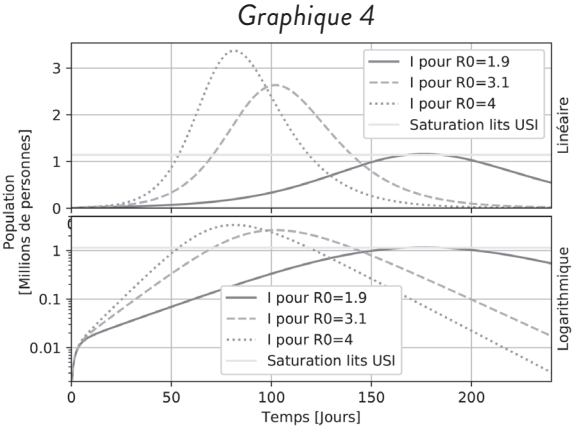


Le comportement général du modèle, visible dans ce graphique, est le suivant : le nombre de personnes exposées, puis infectées avec un délai, augmente fortement jusqu'à obtenir une valeur maximale de I, le pic de l'épidémie. L'augmentation de I est exponentielle du jour 10 au jour 50 environ et apparait donc comme une droite sur le graphique logarithmique. On peut donc définir un temps de doublement pour le nombre de personnes infectées, comme dans le modèle jouet de population. Après le pic, les infectés devenant guéris et supposés immunisés, le réservoir de personnes susceptibles diminue et les quantités E et I diminuent également. Le pire est passé. Le nombre de décès ou de personnes nécessitant des soins intensifs est généralement considéré comme une fraction de I. Dans les graphiques suivants, je montre uniquement les données pour I.

Je présente maintenant une utilisation typique du modèle : que se passe-t-il si on varie le paramètre de reproduction R0 ? Il suffit d'afficher la courbe I pour plusieurs valeurs. Je prends la gamme de valeurs estimées par l'article de référence pour le coronavirus, R0 = 1,9 et R0 = 3,1. Je rajoute la valeur R0 = 4 qui semble pertinente pour l'évolution initiale de la maladie en Belgique.

De cette étude, on peut tirer l'observation suivante : la valeur de R0 influence fortement la hauteur et le moment du pic de l'épidémie. Les conséquences pour la santé publique sont très importantes car la hauteur du pic détermine la

Graphique 4



surcharge du système de soins. En effet, une fraction des personnes infectées nécessite une hospitalisation. Dans ces patients, certains auront besoin d'un lit en unité de soins intensifs. Si l'épidémie dépasse ce seuil, la situation devient dramatique, comme cela a été le cas en Italie par exemple. Pour une situation idéalisée, je trace la barre horizontale de saturation du système de soins intensifs vers 1,14 million de cas infectés (une fraction seulement de ces cas nécessitant des soins intensifs). C'est le fameux effet d'aplatissement de la courbe dont les médias ont fait la présentation : en prenant des mesures qui réduisent le taux de contamination, le scénario d'épidémie peut devenir gérable.

Une fois l'épidémie entamée, on peut malheureusement comparer les résultats du modèle avec les données collectées. Cette phase est essentielle pour calibrer le modèle, c'est-à-dire pour vérifier et éventuellement mettre à jour les paramètres. Le modèle et l'exploration d'hypothèses deviennent des outils d'aide à la décision et permettent d'estimer si on réussit à « aplatir la courbe ».

Dans le cas du coronavirus, plusieurs observations ont été faites concernant la qualité des données. La première est le décalage entre le nombre de personnes effectivement infectées et le nombre de tests positifs obtenus, ce dernier étant probablement largement inférieur au premier. Une réponse possible est, considérant le nombre d'hospitalisations comme une fraction fixe de I, de déduire cette dernière valeur. La seconde objection est qu'une fraction importante des personnes infectées ne montre aucun symptôme (on parle des cas asymptomatiques), et qu'il est dès lors difficile de connaitre l'ampleur de la propagation du virus. En appliquant ces corrections aux données, en fonction des études dans d'autres pays, on obtient une vision plus réaliste de la situation.

Le modèle SEIR présenté ci-dessus considère uniquement la taille des sous-populations S, E, I et R. En l'utilisant, on néglige donc la structure de la société des points de vue géographique et d'âge, entre autres. La théorie des réseaux permet de combler ce manque : on représente alors les individus comme élément de base du modèle en réseau et la transmission de la maladie peut alors prendre en compte les contacts entre personnes, en fonction de leur lieu d'habitation et de leur âge.

Au fur et à mesure qu'on augmente le détail du modèle, on oublie facilement la question des hypothèses sous-jacentes et celle du domaine de validité des équations. Il est alors tentant de confondre modèle et réalité et de donner une force prédictive exagérée au modèle. C'est à ce moment que l'expert·e (soit l'expert·e du domaine, soit l'expert·e en modélisation, idéalement les deux ensemble) doit exercer son esprit critique et rappeler ces limites.

Tous les modèles sont faux, certains sont utiles

Le titre de cette section fait référence à un dicton attribué au statisticien George Box. En établissant une description mathématique de la réalité, les scientifiques se fondent sur des hypothèses et acceptent des approximations, ce qui suffirait à les qualifier de faux. On peut cependant bâtir une compréhension fondamentale de certains phénomènes grâce à ces simplifications. Le dicton sert donc de rappel aux utilisateurs de modèle plus que de critique aveugle. L'utilisation d'un modèle pour influencer des choix politiques ne peut négliger les limites énoncées ci-dessus, particulièrement lorsqu'il s'agit de santé publique. Ignorer les apprentissages de ces modèles serait tout aussi malheureux.

Appliquons cette stratégie au modèle d'épidémie : quels sont les apprentissages ? Quelles sont les limites ? En premier lieu, on peut conclure que diminuer la vitesse de propagation de l'épidémie diminue le nombre maximum de personnes infectées à un moment donné. En conséquence, la durée d'application des mesures de ralentissement doit aussi augmenter. Le nombre de personnes qui auront été infectées à un moment ou un autre est d'au moins 90 % dans la plupart des scénarios, confirmant le message des médias qu'aplatir la courbe sert principalement à ne pas saturer le système hospitalier. L'hypothèse de base sur laquelle se basent ces conclusions est qu'on peut influencer le taux de reproduction R0 en prenant des mesures de type confinement et fermeture d'écoles, suivant en ça la littérature scientifique. L'interprétation des données journalières de l'épidémie se fera par ailleurs à la lumière du modèle, soulignant son utilité comme point de repère.

Du côté des limites, il faut rappeler que le modèle présenté ici est une approximation qui réduit les relations entre des millions d'individus à quatre quantités mathématiques et trois paramètres. Parmi ceux-ci, R0 varie en fonction de la population et doit être calibré pour chaque pays. Les différences de pyramide des âges, d'immunité et de comportement social sont, en effet, négligées et résumés de façon globale dans cette valeur. Une autre critique est liée à l'utilisation des résultats : là où l'évidence mathématique devrait être non ambigüe, différents pays ont tiré des conclusions complètement différentes sur les politiques de santé publique à adopter. La chaine d'analyse modèle-scientifique-politique est donc fort sensible aux variations.

En guise de conclusion, il nous faut insister : disposer d'un modèle n'est *pas* suffisant. Ce qui compte c'est l'expérience combinée des modélisateur·rice·s qui, avec des spécialistes du domaine concerné, pourront mettre les résultats en perspective. Construire un modèle n'est pas construire une politique publique.

Références

- Données historiques de la population mondiale : https://cutt.ly/ZtUR9du
- Utilisation du modèle SEIR pour le cas de Wuhan : H. Wang, Z. Wang, et co-auteurs, « Phase-adjusted estimation of the number of Coronavirus Disease 2019 cases in Wuhan, China », Cell Discovery (volume 6, article 10) 2020, https://cutt.ly/3tUR67q.
- Modèles compartimentaux en épidémiologie, Wikipedia.