NAKIAP

Zeste de savoir

Modéliser la propagation du SARS-CoV-2

13 avril 2020

Table des matières

Dans ce billet nous allons voir comment élaborer un **modèle simple** pour étudier la propagation d'un virus dans une population, puis nous utiliserons les données sur la propagation du SARS-CoV-2 pour **effectuer une regression** et **prédire** le nombre de cas futurs.

Le but de ce billet n'est pas de donner une **évaluation précise** mais plutot de présenter des méthodes et outils qui pourront être utilisés ailleurs et que j'utilise moi-même, je ne suis pas du tout un expert des épidémies et je suis donc incapable d'émettre un quelconque avis sur le sujet

Pré-requis : maths niveau L1 pour comprendre toutes les équations.

1. Modèle mathématique

1.0.1. Croissance exponentielle

Un virus se propage d'individus en individus : **plus il y aura de gens infectés, plus il y aura de gens pour transmettre le virus**.

Si l'on note $P(t)$ le nombre d'infectés à un instant t donné, alors à un instant $t + \mathrm{d}t$, le nombre de *nouveaux infectés* sera proportionel à $P(t)$. Le coefficient de proportionalité dépendra bien sûr de plusieurs facteurs :

- La durée d'exposition $\mathrm{d}t$
- La probabilité de transmettre le virus à quelqu'un lors d'un contact
- Le nombre de personnes que l'on peut croiser pendant $\mathrm{d}t$
- Le temps pendant lequel un individu est contagieux
  Pour le moment nous allons confondre les trois derniers paramètres en un seul produit que nous noterons a (les épidémiologistes parlent de R_0).
  Le nombre d'infectés à $t + \mathrm{d}t$ est donné par la relation suivante :

$$P(t + \mathrm{d}t) = \text{nouveaux infectés} + \text{anciens infectés} = a \times P(t) \times \mathrm{d}t + P(t)$$

D'ou l'on tire l'équation différentielle linéaire de premier ordre :

$$P' = aP$$

Qui a pour solutions :

$$P(t) = K\exp(at), \qquad K \in \mathbb{R}$$

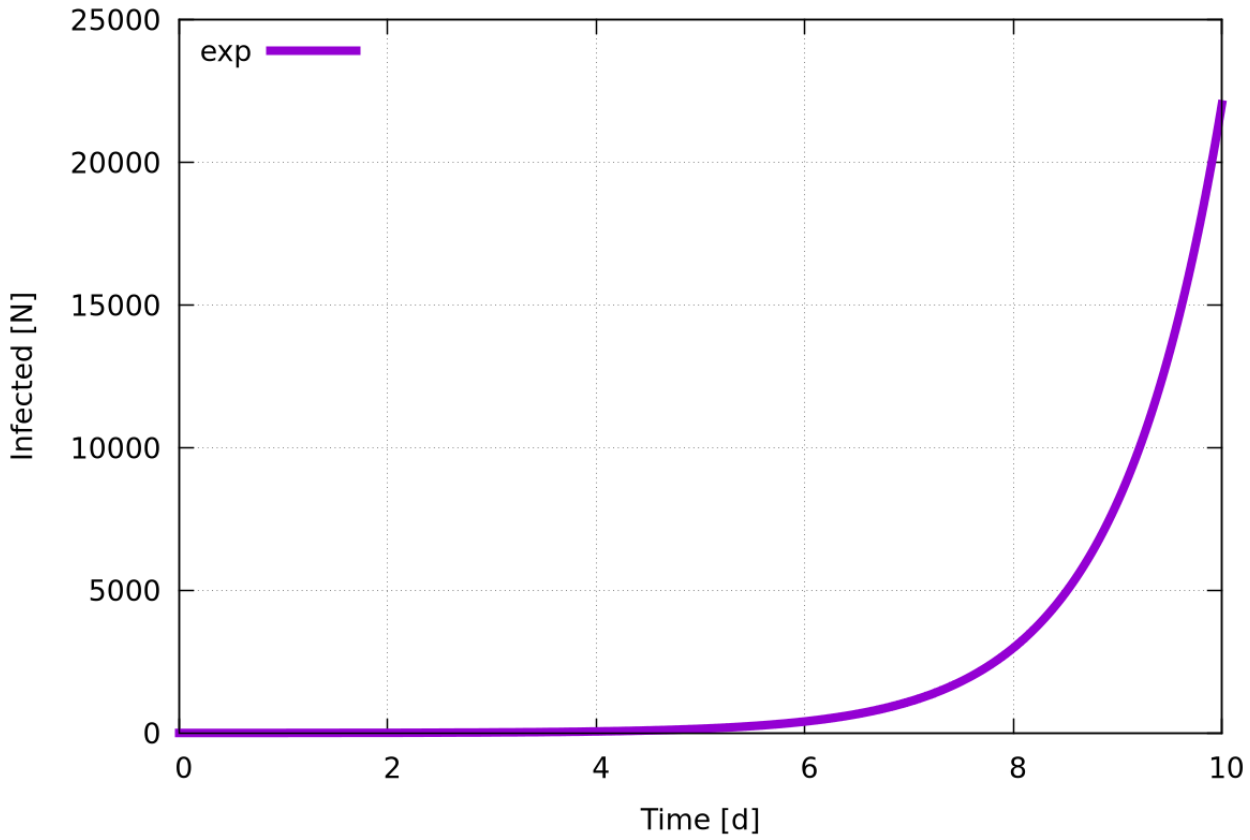


FIGURE 1.1. – Une exponentielle ça grandit vite !

On a donc affaire à une croissance exponentielle sans fin, dans quelques mois l'humanité sera donc complètement infectée ! Ou pas ...

1.0.2. Plafonnement logistique

Notre modèle ne colle pas à la réalité, en effet le nombre d'infectés ne peut pas grandir indéfiniment, déjà parce que nous ne sommes qu'un nombre fini de terriens. On pourrait donc supposer que la population mondiale P_W est une borne supérieure de P. Dans l'équation différentielle cela se traduit par le fait que l'on ne peut infecter que des gens qui n'ont pas *déjà été infectés*. **Plus il y a de contaminés, moins il y a de gens à contaminer** Le coefficient a dépend donc de P et a fortiori du temps :

$$a(P) = \alpha \times \text{Proportion de la population non infectée} = \frac{P_W - P}{P_W}$$

On a donc :

$$P' = \alpha \times \left(1 - \frac{P}{P_W}\right) \times P$$

En réalité P_∞ sera inférieur à P_W, ici on supposait un cas idéal (pour le virus, pas pour nous) où l'ensemble de la population est mélangé aléatoirement à chaque étape du procesus, en pratique ce n'est pas le cas certains pays ne jouent pas le jeu et ferment leurs frontières, se confinent bref... des barrières sont mises

Posons donc que le nombre final d'infectés est une inconnue que l'on nomme N. Il nous faut donc résoudre l'équation différentielle de premier ordre :

$$P' = \alpha \times \left(1 - \frac{P}{N}\right) \times P$$

Si vous êtes allergique aux maths, sautez les trois lignes suivantes

En séparant les variables on obtient :

$$\int \frac{\mathrm{d}P}{(1 - P/N)P} = \int \alpha \mathrm{d}t$$

On décompose en éléments simples à gauche, on intègre les fractions et finalement :

$$P(t) = \frac{K}{\exp(-\alpha t) + K/N} \qquad , \qquad K \in \mathbb{R}$$

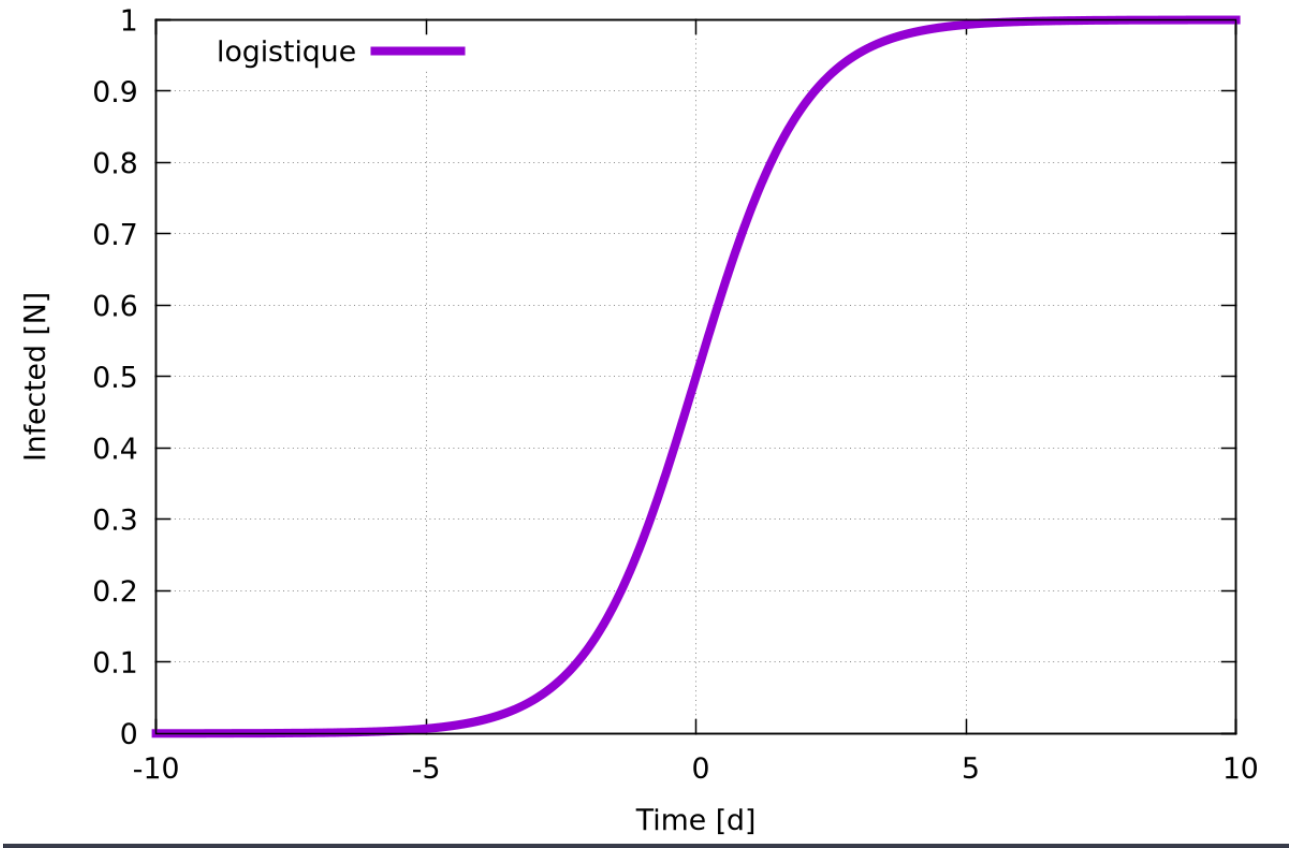


FIGURE 1.2. – Toute courbe exponentielle cache une courbe logistique

On a donc trois grandes inconnues qui sont K, N et α. Nous allons tenter de déterminer leurs valeurs pour effectuer une prédiction du nombre de cas total en France.

Mais avant intéressons nous sur le *sens physique* de ces paramètres. C'est bien beau de faire des modèles mais encore faut-il qu'ils aient un sens !

N a été posé comme étant le nombre de cas limite, et en effet on obtient bien :

$$\lim_{t \to +\infty} P(t) = N$$

a est notre coefficient de transmission et on peut voir qu'il ait bien homogène à l'inverse du temps comme nous l'avions défini au début.

? Et K alors késako ? à vous de trouver !

Contenu masqué n°1

2. Modèle informatique

Ici le but est donc d'effectuer une **regression** sur un jeu de données pour tenter de deviner ces paramètres, notre modèle étant *non linéaire et non linéarisable* on choisit d'utiliser l'algorithme de *Levenberg–Marquardt* pour minimiser le χ^2. (code disponible ici)

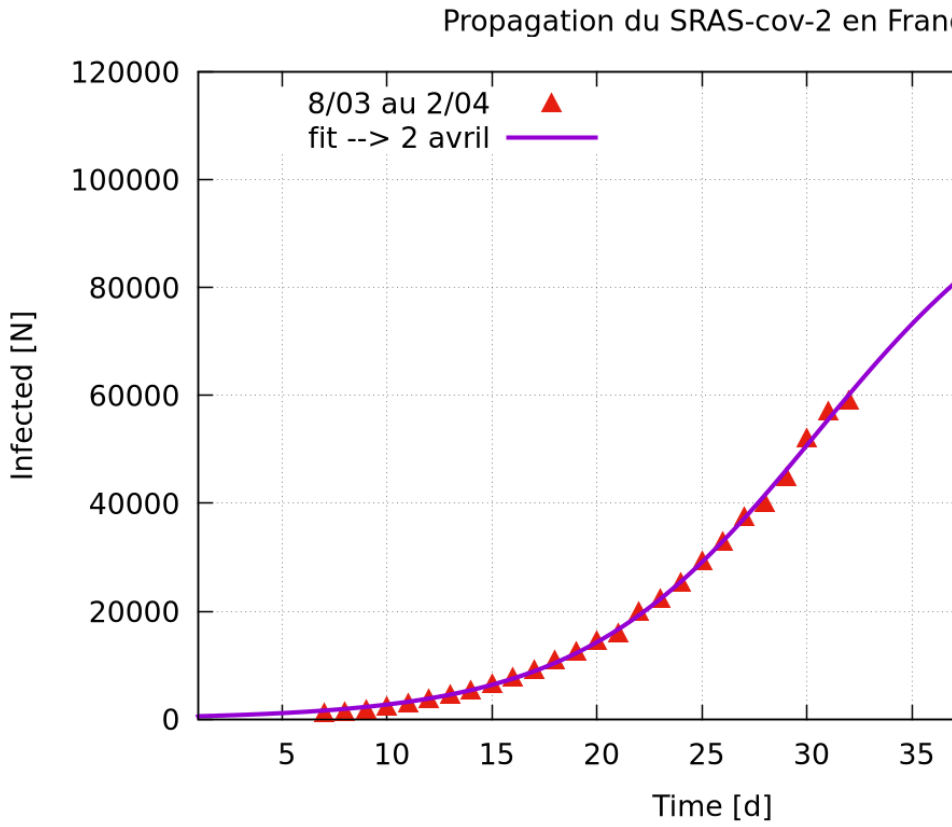
Je rappelle que le χ^2 désigne ici une mesure de l'erreur quadratique :

$$\chi^2 = \sum_i (y_i - f_p(x_i))^2$$

Où f_p représente la fonction que l'on suppose être solution du *fit* et p ses paramètres, dans notre pas f_p est la fonction logistique et p est le vecteur $(NK\alpha)$

> Si vous ne comprenez pas bien ce qui vient d'être dit, cela n'a aucune importance pour comprendre la suite.

Nous allons commencer par faire un premier *fit* en nous basant sur les données du site worldometers et nous utiliserons le logiciel gnuplot. L'implémentation de LM y est très puissante et permet de trouver facilement les paramètres de beaucoup de courbes ! Souvenez vous cependant que l'on cherche à minimiser le χ^2 et que pour cela l'algorithme va chercher à tâtons une solution, il faut donc avoir une bonne idée de la solution pour que cela fonctionne. En effet si vous démarrez avec un p initial trop éloigné de la réalité, vous risquez de ne jamais converger et de vous perdre dans les limbes des algorithmes de regression ...



On obtient alors la courbe suivante :

Et les résultats du fit sont :

À première vue le fit semble correct et on obtient moins de 10% d'erreurs pour chaque paramètre.

?

Mais pourquoi tu ignores les jours avant le 8 mars ?

Parceque qu'il faut bien commencer quelque part ! En traçant les données du tout début de l'épidémie on peut voir qu'elles sont très fluctuantes et semblent donc peu pertinentes. Cela ne se voit pas forcément en échelle linéaire mais en traçant en échelle logarithmique comme on doit toujours le faire d'ailleurs

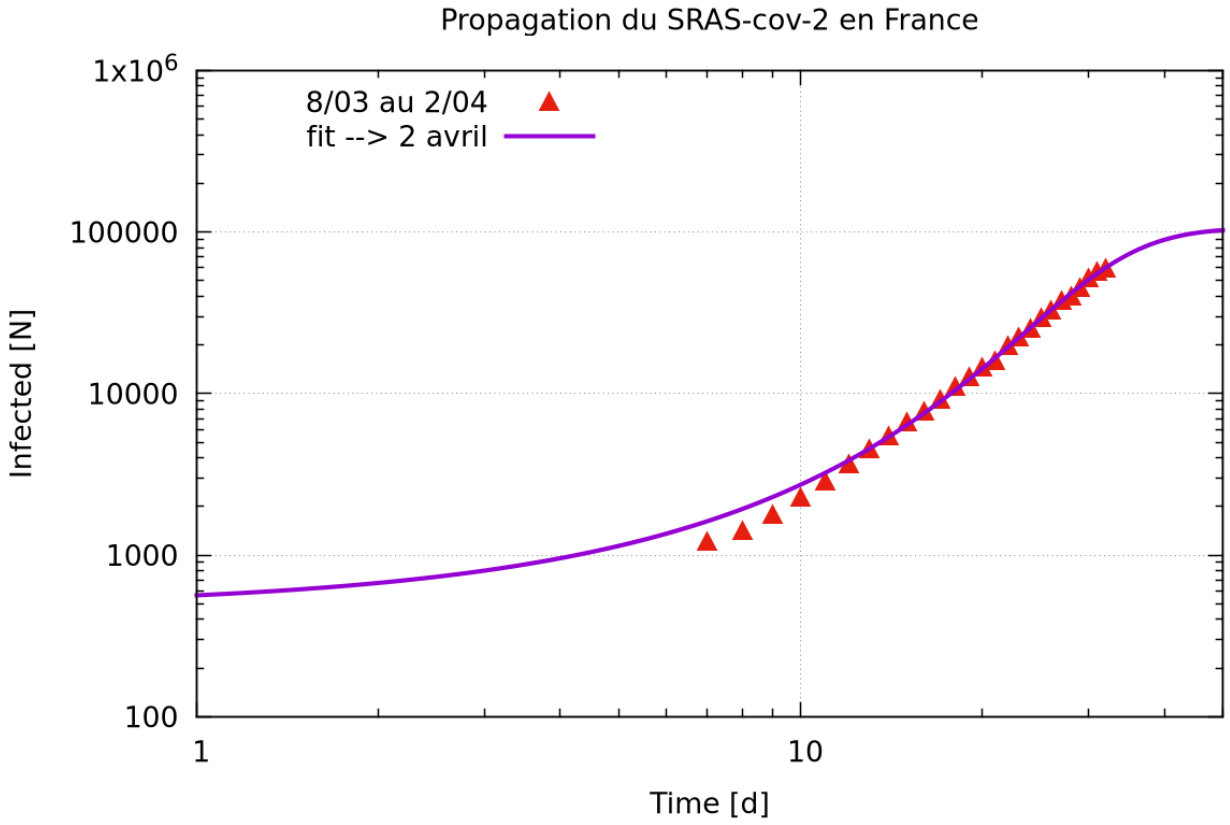


FIGURE 2.3. – Échelle logarithmique : la vraie, l'unique !

?

Ok mais tu te moques de nous ! Tu prédis 106 000 cas et là on est le 13 avril on a dépassé les 130 000 ! Ça marche pas ton truc !

Remarque judicieuse ! Et d'ailleurs si vous y avez prêté attention α vaut environ 0.18 s^{-1} alors que dans la réalité les épidémiologistes s'accordent pour dire que sa valeur est supérieur à deux pour une population non confinée. Où est l'arnaque ?

Alors bien sûr je n'ai aucun moyen d'en être sûr mais pour moi le problème vient de la sous-estimation des cas : en effet du à un manque de test en France il est impossible de savoir au jour le jour combien de personnes sont touchées. Nos données sont biaisés, notre résultats l'est donc aussi !

Et d'ailleurs si on s'intéresse aux données après le 2 avril on voit que tout change :

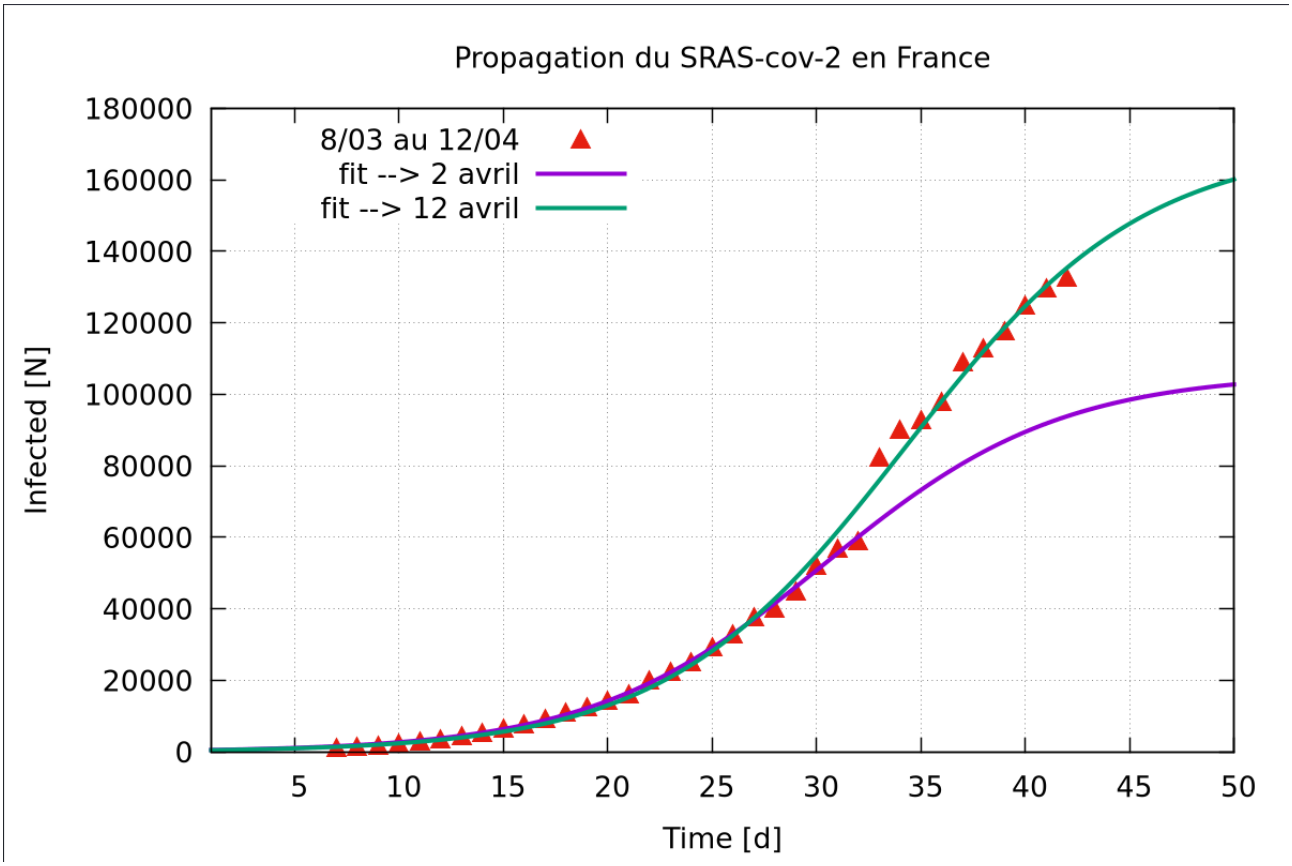


FIGURE 2.4. – Changement radical de direction

La faute à quoi ? Et bien là encore difficile d'être sûr mais apparemment cette augmentation brusque correspond à la prise en compte de nombreux cas comme ceux détectés dans les EHPAD ou dépistés par les médecins généralistes.

3. Critique du modèle / de la regression

Tentons de prendre un peu de recul pour analyser ce qui représente la réalité et ce qui est à côté de la plaque :

— Notre fit converge, il y a donc cohérence entre les données de terrain et un modèle purement mathématiques
— Le R_0 est très sous-évalué, probablement du à une mauvaise estimation des cas contaminés. On pourrait essayer de régler le problème en gonflant artificiellement les données (on peut par exemple inférer le nombre de cas à partir du nombde de morts et du taux de mortalité) mais on risque d'introduire de nouveaux paramètres peu ou mal connus donc pas sûr que ça soit beaucoup mieux ... Voir par exemple cet article pour en savoir plus sur la difficulté de déterminer le taux de mortalité d'une épidémie en cours.
— Il en résulte que le N est probablement lui aussi sous-évalué
— Notre modèle reste très simpliste par rapport au modèle SIR par exemple qui lui prend en compte l'évolution du nombre de décés, de cas rétablis etc ... mais cela est voulu : dans le cas du modèle SIR faire des regression sur un jeu de données en cours devient beaucoup plus compliqué étant donné que l'on a plus de solution explicite. Voir ce lien pour plus de détail.
— J'ai aussi tenté de créer un modèle qui prenait en compte le confinement pour obtenir deux R_0 différents, un avant et un après le confinement mais la différence n'était pas parlante alors je n'en ai pas parlé.

En conclusion : on a vu comment à parti d'outils simples et facilement accessible on peut modéliser une situation physique et effectuer des prédictions plus ou moins réalistes Ici le problème est que nous n'avons pas de données très fiables pour le moment sur lesquelles baser notre esimation, comme c'est souvent le cas en science : il faut faire avec ce que l'on a ! On a rarement un chemin tout tracé vers une solution merveilleuse ... J'espère que ma démarche pourra vous être utile !

Contenu masqué

Contenu masqué n°1

K est homogène à une population comme N et représente en fait la proportion de gens infectés par rapport aux gens non infectés au début de l'épidémie. En effet on peut vérifier que $K = NP_0/(N - P_0)$. Ce qui est un peu flou c'est la définition que l'on prend de *début de l'épidémie* ici c'est quand $t = 0$, en pratique cela ne correspond à rien de physique, il n'y a pas un moment précis ou l'épidémie commence... Retourner au texte.