Une tentative de prédiction de l'évolution hebdomadaire du risque épidémique avec le Baromètre Covid 19

Par Mathieu Moslonka-Lefebvre , PhD en épidémiologie mathématique, président de DataCovid s'exprimant ici à titre personnel.

Version actualisée le 28 juin 2020 avec les données de la 8e vague.

Avant-propos

Ce billet, à visée scientifique et didactique, illustre un **cas d'usage des données du Baromètre** Covid 19 avec pour objectif de **mesurer l'évolution du risque épidémique dans l'espace et dans le temps, avec une portée de prédiction limitée, de l'ordre de la semaine ("nowcasting")**.

Les analyses mises en oeuvre n'engagent que leur auteur et n'ont pas fait l'objet d'une relecture par les pairs. Les éléments rapportés sont susceptibles d'évoluer en fonction des éventuels commentaires qui seront effectués par la communauté scientifique.

Pour rappel, l'intégralité de ce billet repose sur des données de sondages en population générale qui présentent un certain nombre d'avantages, mais également des sources de biais potentiels. A titre d'exemple, et sans chercher à être exhaustif, le caractère *déclaratif* des données peut être source de multiples biais :

- Un exemple de biais potentiel peut être lié à la possibilité que des répondants souhaitent communiquer des informations au-delà de la fenêtre temporelle fixée pour certaines questions, par exemple une fenêtre de 24 heures pour le nombre de contacts à moins d'un mètre et le temps passé dehors. Pour les répondants concernés, ce biais peut trouver sa source dans le sentiment de contribuer à la lutte contre l'épidémie en fournissant une information inexacte, mais qu'il pensent utile. Cette situation conduirait à surestimer les valeurs réelles ;
- Un autre exemple de biais potentiel serait lié à la possibilité que des répondants se méfient et ne souhaitent pas communiquer d'information ; cela conduirait au contraire

> du premier biais à sous-estimer les valeurs réelles. S'il est possible que ce biais ne soit pas très grand, il est aussi possible qu'il varie dans le temps, par exemple en fonction des signaux envoyés par le gouvernement sur la prise en charge des personnes infectées.

Ce billet a vocation à être actualisé chaque semaine et enrichi par des analyses plus élaborées, notamment dans le cadre de travaux scientifiques spécifiques ultérieurs qui seront résumés sur le blog.

En matière de santé publique, il est recommandé de suivre et consulter les instructions officielles disponibles sur https://www.gouvernement.fr/info-coronavirus.

Mesurer les risques de rebonds épidémiques avec le "R zéro"

Pour évaluer le risque épidémique associé à l'épidémie de Covid 19, les épidémiologistes calculent un indicateur stratégique, le ***R0*** (prononcé "R zéro" et aussi appelé nombre de reproduction de base). Ce nombre correspond au nombre moyen de personnes qui vont être contaminées par un individu infecté lors de sa période infectieuse. Le "zéro" du ***R0*** signifie que ce nombre se calcule dans une population encore très largement sensible au virus, c'est à dire au début de l'épidémie lorsque le niveau d'immunité demeure négligeable. [1]

Lorsque le ***R0*** est strictement supérieur à 1 l'épidémie se déclenche et se développe. A l'opposé, lorsque le ***R0*** à une valeur inférieure ou égale à 1, l'épidémie recule. Des mesures telles que la pratique des gestes barrières ou le confinement permettent de faire décroître le ***R0*** et ainsi de maîtriser l'épidémie s'il atteint une valeur inférieure ou égale à 1.

Avant le confinement, la valeur du **R0** a été estimée pour la France à 2,9 par une équipe de modélisateurs de l'Institut Pasteur [2]. Cette valeur est cohérente avec l'estimation antérieure proposée par autre équipe du CNRS, de l'Institut de Recherche sur le Développement et l'Université de Montpellier », avec une fourchette estimée comprise entre 2,5 et 3,5 [3].

D'après les données d'hospitalisation en France, le confinement décidé le 17 mars dernier a permis, après un certain délai, de rompre les chaînes de transmission de l'épidémie en réduisant le **R0** a une valeur inférieure au seuil critique de 1. Les valeurs les plus probables sont comprises entre 0,67 et 0,75 suivant les estimations des deux équipes précitées [4]. Pour la suite nous retiendrons une valeur indicative de "**R0** post-confinement" de 0,7.

L'évolution à la hausse du "R zéro", avec un risque qui semble pour l'instant limité

Les résultats de la 8ème vague d'enquête du baromètre de Datacovid réalisée du 5 au 9 juin 2020 et rapportés par Ipsos indiquent que "les comportements de distanciation sociale continuent de se relâcher avec une nouvelle progression du nombre moyen de personnes avec lesquelles il y a eu un contact rapproché : 11,3 personnes en moyenne chez les personnes qui sortent dehors contre 8,5 la semaine dernière et 4,2 début avril."[5].

L'enjeu auquel nous tentons d'apporter de premières réponses dans ce billet est de déterminer dans quelle mesure **les changements de comportements constatés d'une vague à l'autre pourraient modifier le R0** à la hausse ou à la baisse.

Pour bien comprendre les facteurs qui vont jouer sur le **R0**, il est utile de comprendre ses composantes. Quel que soit la complexité du modèle épidémiologique employé pour le calculer, le **R0** fait presque systématiquement intervenir trois composantes [6] : la probabilité d'infection par contact notée *ß*, une composante notée *C* qui reflète l'influence de la structure, de la densité et de l'intensité des contacts sur la transmission et la durée pendant laquelle une personne porteuse du virus peut le transmettre (période infectieuse) notée *D*. Le **R0** se calcule typiquement de la façon suivante : **R0** = *ß x C x D* .

En lien avec les données du baromètre, on note alors que :

- L'adoption de mesures d'hygiène et de protection individuelle (se laver les mains plusieurs fois par jour, porter un masque efficace, etc.) va diminuer le **R0** en réduisant la probabilité d'infection par contact ß ;
- Le respect du confinement (éviter de sortir de chez soi sauf pour des motifs impérieux, respecter une distance de sécurité d'au moins un mètre, etc.) va réduire le **R0** en diminuant la composante relative aux contacts *C ;*
- La période infectieuse *D*, quant à elle, ne sera pas modifiée par des aspects comportementaux. Elle pourrait toutefois être réduite à l'avenir quand de nouveaux traitements efficaces seront proposés contre le Covid 19. Pour la suite nous prendrons comme hypothèse que *D = 7* jours [7].

Dans le présent billet, seule l'influence de la composante relative aux contacts *C* est explorée. On supposera, pour fixer les idées, que la probabilité d'infection par contact *ß* est stable à partir de la première vague du baromètre, soit après la date de mise en oeuvre du confinement[8]. Cette hypothèse permet de déduire la valeur de *ß* en prenant pour base un **R0**

post-confinement de 0,7 lors de la première vague du baromètre.

A la différences des méthodes de calcul du **R0** fondées sur des comptages de cas qui détectent l'épidémie de façon très fiable mais avec un retard d'au moins une semaine [9], les estimations du **R0** qui suivent constituent un exemple d'approche dite de « **signaux faibles** », c'est-à-dire un moyen qui pourrait permettre de détecter l'épidémie de façon précoce en se fondant sur le suivi temporel de certains mécanismes de l'épidémie, en l'occurrence les descripteurs de la structure de contacts.

Pour la suite, les nombres entre crochets représentent respectivement les estimations basses et hautes du R0 à 2,5 % et à 97,5 %. Les distributions du R0 sont obtenues par *bootstrap*[10] sur les observations de chaque vague (10 000 réplications par vague). Cette technique permet de "propager" les incertitudes pesant sur la composante relative aux contacts *C* dans le calcul **R0** et ainsi de fournir un intervalle de confiance dans l'estimation.

De la 1ère à la 8ème vague, le nombre moyen de contacts rapprochés au niveau national augmente de 2,3 par jour à 9,5 par jour (pour mémoire : 7,2 contacts par jour pour la 7ème vague), soit une augmentation significative[11] à hauteur de 82 % environ. Cet effet, dans un modèle simpliste où *C* est donné par le nombre moyen de contacts noté <k> où <.> désigne l'opérateur "moyenne" (voir Annexe), conduirait à une augmentation potentielle du **R0** dans les mêmes proportions, soit de 0,7 à 2,8 [2,7 à 3,0] ce qui serait inquiétant, car excédant le seuil épidémique.

En réalité, il convient de prendre en compte le fait que les réseaux de contacts humains sont hautement hétérogènes en termes de nombre de contacts par personne. Dans ce cas de figure plus réaliste, *C* est mieux capturé par le ratio <k2> / <k> (voir Annexe), aussi le **R0** augmenterait de 0,7 à 1,0 [0,8 à 1,1], (avec 1,0 qui est un arrondi à 0,1 de 0,95), soit une augmentation plus modeste, inférieure à la valeur seuil de 1, avec donc une épidémie qui resterait sous contrôle.

En outre, les interactions sont non seulement hétérogènes en nombre de contacts, mais aussi en termes de durée des interactions. Dans ce cas de figure, *C* peut être capturé par un ratio plus complexe en faisant l'hypothèse que le réseau de contacts sous-jacent, appelé réseau pondéré en théorie des réseaux, est statique (voir Annexe). Nous prendrons ici comme base de calcul de durée des interactions le temps passé dehors par les répondants dans les dernières 24 heures. Cette quantité est passée de 66 minutes à 246 min entre la 1ère et la 8ème vague (pour mémoire : 208 minutes pour la 7ème vague), soit une augmentation significative [12] à hauteur de 270 % environ. Si on prend en compte le poids des interactions en sus de l'hétérogénéité des contacts, l'estimation du **R0** à la 8ème vague

est de 1,2 [1,1 à 1,3], avec donc un risque estimé qui devient supérieur au modèle hétérogène non-pondéré en raison de la forte augmentation intervenue depuis le déconfinement sur le temps passé dehors (voir *supra*). Si le risque ainsi estimé est légèrement supérieur au seuil épidémique situé à 1,0, cette estimation du **R0** constitue une estimation haute du risque car le temps passé dehors (proxy data ici employée) et le temps effectif passés avec les contacts prochent sont sans doute imparfaitement corrélés, avec un possible effet de saturation qui, s'il était avéré, viendrait réduire la valeur du **R0** ainsi prédite.

Tableau. L'évolution hebdomadaire du risque épidémique suivant les informations prises en compte dans le baromètre.

Vague du baromètre	**R0 (hypothèse des contacts homogènes)**	**R0 plus réaliste (avec contacts hétérogènes pris en compte)**	**R0 plus réaliste (avec contacts hétérogènes et durées de sorties prises en compte)**	**Prévision du statut épidémique *sous réserve des hypothèses du présent billet***
Vague 1 du au 7 au 14 avril)	0,7	0,7	0,7	Épidémie qui serait sous contrôle au niveau national quel que soit l'indicateur et le modèle employé
Vague 2 du 15 au 21 avril	1,0 [0,9 à 1,1]	0,8 [0,7 à 1,0]	0,8 [0,7 à 1,0]	Épidémie qui serait sous contrôle au niveau national pour les indicateurs les plus réalistes
Vague 3 du 22 au 27 avril	1,0 [0,9 à 1,0]	0,6 [0,5 à 0,8]	0,6 [0,5 à 0,7]	Épidémie qui serait sous contrôle au niveau national quel que soit l'indicateur et le modèle employé
Vague 4 du 28 avril au 4 mai	1,1 [1,0 à 1,3]	0,9 [0,7 à 0,9]	0,8 [0,7 à 0,9]	Épidémie qui serait sous contrôle au niveau national pour les indicateurs les plus réalistes

Vague 5 du 6 au 11 mai	1,3 [1,2 à 1,4]	0,9 [0,7 à 1,0]	0,8 [0,7 à 1,0]	Épidémie qui serait sous contrôle au niveau national pour les indicateurs les plus réalistes
Vague 6 du 6 au 11 mai	1,7 [1,6 à 1,8]	0,9 [0,7 à 1,0]	0,9 [0,8 à 1,0]	Épidémie qui serait sous contrôle au niveau national pour les indicateurs les plus réalistes
Vague 7 du 26 au 31 mai	2,1 [2,0 à 2,3]	0,8 [0,7 à 1,0]	1,1 [1,0 à 1,2]	Épidémie qui serait sous contrôle d'après l'un des deux indicateurs les plus réalistes
Vague 8 du 5 au 9 juin	2,8 [2,7 à 3,0]	1,0 [0,8 à 1,1] avec 1,0 qui est un arrondi à 0,1 de 0,95, soit une valeur inférieure à 1.	1,2 [1,1 à 1,3]	Épidémie qui serait sous contrôle d'après l'un des deux indicateurs les plus réalistes

Les limites et les perspectives : ajouter de nouveaux prédicteurs et comparer les prédictions aux valeurs "réelles"

Sur la seule base de la dernière estimation nationale robuste connue correspondant grossièrement, en prenant en compte le décalage temporel, à la 8ème vague de DataCovid, et pour lequel le **R0** de référence a été estimé 0,92 [13], on constate que la prédiction du **R0** avec contacts hétérogènes pris en compte, dont la valeur prédite une semaine à l'avance est de 0,95, **se révèle très proche sans qu'il ne soit possible d'exclure à ce stade un simple effet du hasard**, étant donné qu'un seul point de comparaison est pris pour référence.

Comme la taille de l'échantillon est élevée (5 000 répondants), il est possible de décliner ces analyses **au niveau régional** pour dégager des tendances géographiques. Il sera ensuite possible de comparer les différents **R0** ainsi prédits au niveau régional aux estimations de référence mais avec un "effet retard", par exemple celles rapportées dans les points épidémiologiques de Santé Publique France. Le nombre de points de comparaison ainsi

obtenu pourrait se révéler suffisamment élevé pour conclure sur la fiabilité des prédiction. Ce point fera l'objet d'un billet spécifique.

Les méthodes d'estimation ici présentées sont particulièrement réductionnistes et leurs prédictions doivent être prises avec précaution dans l'attente de comparaisons robustes avec des valeurs de référence calculées sur la base de comptages de cas.

En particulier, les méthodes de "signaux faibles" ici présentées **ne prennent pas en compte une éventuelle sensibilité saisonnière du Covid ainsi que des mesures telles que le port du masque** qui vont réduire la probabilité d'infection par contact ß.

Un moyen pour tenter de dépasser ce dernier problème pourrait être de prendre en compte la **« directionalité » des contacts** induite par le port de masques "grand public" dans la population générale qui protègeraient davantage la contamination des tiers que les porteurs eux-même. La formule du **R0** est en effet connue pour les réseaux dirigés[14].

Annexe. Modèles et indicateurs employés.

Ce billet se focalise sur les modèles les plus simples permettant de décrire l'épidémie de COVID-19 : les modèles dits "SIR". Par souci de simplicité, de nombreuses caractéristiques ne sont pas ici prises en compte et sont susceptibles d'influencer les résultats sur un plan quantitatif : la structure d'âge de la population ; le fait que les personnes infectées peuvent être infectieuses un à plusieurs jours avant de présenter des symptômes etc.

Les formules ici employées pour calculer les **R0** sont disponibles aux équations (1a ; pour le réseau hétérogène avec mélanges) et au Tableau 2 (pour le réseau hétérogène pondéré supposé fixé) de la référence [15].

Notes de bas de page et références

1. Lorsque l'épidémie est installée, on ne parle plus de ***R0*** mais de nombre de reproduction efficace noté ***R***. Dans la mesure où la séroprévalence en France serait extrêmement faible, (cf. Salje et al. (2020) Estimating the burden of SARS-CoV-2 in France. *Science*), et par souci de simplicité, les deux notions ne sont pas distinguées. Pour une présentation plus large des enjeux associés à la modélisation de l'épidémie de Covid-19 avec une formalisation plus complète de ces concepts, des rapports en français d'excellente qualité sont proposés par l'équipe de Samuel Alizon (CNRS, IRD, Université de Montpellier). Voir en particulier cette synthèse :

http://covid-ete.ouvaton.org/Rapport7_resume.html.[↩]
2. Cf. Salje et al. (2020) Estimating the burden of SARS-CoV-2 in France. *Science*[↩]
3. Cf. http://covid-ete.ouvaton.org/Rapport7_resume.html[↩]
4. Cf. respectivement Salje et al. (2020) et http://covid-ete.ouvaton.org/Rapport5_R.html[↩]
5. Cf. note d'analyse des résultats de la 8ème vague d'enquête du Baromètre Covid-19 rédigée par Ipsos datée du 17 juin 2020 et disponible à l'adresse : https://www.ipsos.com/fr-fr/barometre-covid-19-la-majorite-des-francais-approuve-le-calendrier-de-deconfinement. Pour ses calculs, Ipsos se fonde sur l'assiette des personnes sorties de leur foyer (hors jardin). L'assiette du présent billet est l'ensemble des personnes mais la tendance qualitative reste analogue.[↩]
6. Cf. Keeling and Rohani 2008 *Modeling infectious diseases in humans and animals* (2018) Princeton University Press.[↩]
7. Cf. https://www.ecdc.europa.eu/en/covid-19/questions-answers : "The infectious period may begin one to two days before symptoms appear, but people are likely most infectious during the symptomatic period, even if symptoms are mild and very non-specific. The infectious period is now estimated to last for 7-12 days in moderate cases and up to two weeks on average in severe cases.".[↩]
8. Les données du baromètre Covid 19 pourraient utilement être mobilisées pour relier de façon fine les comportements des Français vis à vis de certains gestes barrière à la probabilité d'infection par contact.[↩]
9. Cf. "COVID-19 : point épidémiologique du 25 juin 2020" de Santé Publique France, section relative au nombre de reproduction effectif «R effectif» en page 13 : "Le R effectif estimé à partir de ces données est un indicateur de la dynamique de transmission du virus environ 1 à 2 semaines auparavant (intégrant le délai entre la contamination et le test, et le fait que le calcul est effectué sur une période de 7 jours). […] Les estimations du nombre de reproduction sont basées sur les nombres de tests PCR positifs au COVID-19 réalisés entre le 14 juin et le 20 juin 2020 [ce qui correspond donc aux infections qui se sont produites entre le 7 juin et le 13 juin; des dates qui chevauchent pour partie celles de la 8ème vague de DataCovid.]]"[↩]
10. Les techniques de *bootstrap* sont des méthodes fondées sur une réplication des données obtenue par rééchantillonnage à partir du jeu de données étudié. En l'occurrence le tirage avec remise des observations, ici les répondants de chaque vague, permet de "propager" les incertitudes associées au R0 sur la base des valeurs de ses composantes calculées à partir du baromètre (distribution des contacts et des durées de sortie en l'occurrence).[↩]
11. La significativité est ici appréciée dans un sens statistique, à travers un test de Wilcoxon. La probabilité que la différence ici observée est imputable au hasard (notion

de "p-valeur") est négligeable (inférieure à 2.2×10^-16).[↩]

12. La significativité est ici appréciée dans un sens statistique, à travers un test de Wilcoxon. La probabilité que la différence ici observée est imputable au hasard (notion de "p-valeur") est négligeable (inférieure à 2.2×10-16).[↩]
13. Cf. le point épidémiologique du 25 juin 2020 de Santé Publique France, section relative au nombre de reproduction effectif «R effectif» en page 13 : "Le R effectif estimé à partir de ces données est un indicateur de la dynamique de transmission du virus environ 1 à 2 semaines auparavant (intégrant le délai entre la contamination et le test, et le fait que le calcul est effectué sur une période de 7 jours). […] Les estimations du nombre de reproduction sont basées sur les nombres de tests PCR positifs au COVID-19 réalisés entre le 14 juin et le 20 juin 2020 [ce qui correspond donc aux infections qui se sont produites entre le 7 juin et le 13 juin; des dates qui chevauchent pour partie celles de la 8ème vague de DataCovid. Nous pouvons donc utiliser le R national, estimé à 0,92, et les R régionaux rapporté dans ce rapport et les précédents de Santé Publique France et le comparer à nos estimations.]"[↩]
14. Allard et al. (2020) *The role of directionality, heterogeneity and correlations in epidemic risk and spread*, pré-publication, https://arxiv.org/pdf/2005.11283.pdf. Je remercie la chercheuse Elisabeta Vergu de m'avoir indiqué cet article.[↩]
15. Kamp et al. (2013) Epidemic spread on weighted networks. *PLOS Computational biology*.[↩]