Prévision de l'épidémie de covid19 en France

Loïc Pottier

Prévision de l'épidémie de covid19 en France

Loïc Pottier
loic.pottier@gmail.com

28 mars 2021

Résumé

Avec une méthode mathématique fondée sur l'algèbre linéaire, à partir de données en accès libre (data.gouv.fr, google, apple) on produit des prévisions pour le nombre de patients en réanimation en France avec une erreur moyenne de 4% à 7 jours, 7% à 14 jours, 8% à 21 jours, 10% à un mois, 17% à 2 mois, et 31% à 3 mois. Pour les autres indicateurs de l'épidémie l'erreur est en moyenne de 6% à 7 jours et 25% à 2 mois.

1 Introduction

La méthode employée débute par le calcul de corrélations maximales entre les données des indicateurs quotidiens de l'épidémie (hospitalisations, réanimations, nombre de cas, décès, etc) et celles, décalées arbitrairement dans le temps, des données de son contexte (vacances, météo, mobilité des personnes, couvre-feu, etc), pour chaque jour et chaque département français.
On en déduit des décalages temporels entre contextes et indicateurs, par exemple 19 jours entre la fréquentation des lieux de travail (contexte) et le taux de reproduction [1] associé au nombre de patients en réanimation (indicateur).
Puis, à partir de ces décalages, on calcule une approximation et une prévision linéaires par optimisation quadratique des taux de reproduction. Les approximations des indicateurs dont le taux de reproduction effectif est approximé avec une erreur moyenne supérieure à 5% sont rejetées. Pour les autres, on déduit des approximations et des prévisions des indicateurs correspondants.
Ceci est fait pour chaque département (pour les départements où les données complètes du jour sont présentes au moment du calcul, i.e. en général autour de 88).
On présente à la fin de cet article les prévisions actuelles pour le nombre de patients en réanimation. Des résultats détaillés et actualisés chaque jour se trouvent à cette adresse : `https://cp.lpmib.fr/medias/covid19/_synthese.html`.
Les seules hypothèses qui sont faites sont que les données du contexte gardent dans l'avenir les valeurs qu'elles ont au jour présent, sauf pour la météo, où l'on reprend les valeurs de l'année passée au même moment, et pour les vacances scolaires, qui sont prévues de longue date.

2 Corrélations et décalages

Les données concernent les indicateurs de l'épidémie (urgences, réanimations, décès, tests positifs, etc) et les contextes (données météo : température, pression, données de mobilité fournies par google : fréquentation des commerces et lieux de loisir, des lieux de travail,

1. le taux de reproduction effectif est le nombre moyen de personnes que contamine un malade (en dessous de 1, l'épidémie régresse, au-dessus elle augmente). Il est noté R effectif, Reff, R0 ou R dans la littérature.

etc). Ce sont des données quotidiennes (non cumulées). Elles se présentent, pour chaque donnée $x \in \{1, \ldots, N\}$ (indicateur ou contexte), et chaque département $d \in \{1, \ldots, D\}$, comme un vecteur de valeurs : $x_d = (x_{d1}, \ldots, x_{dn_x}) \in \mathbb{R}^{n_x}$, correspondant à un intervalle de n_x jours $[j_0(x), j_0(x) + n_x[$.
En tout on utilise $N = 48$ données, qui concernent $D = 86$ départements et plus de 464 jours.
On commence par calculer les coefficients de corrélation entre deux données x et y, pour tous les décalages temporels t d'au plus $t_{max} = 40$ jours.
Pour chaque département d, on calcule la valeur moyenne de x_d

$$E(x_d) = \frac{1}{n_x} \sum_{i=1}^{n_x} x_{d,i}$$

puis on complète par la valeur 0 le vecteur $x_d - E(x_d)$ pour les jours où il n'est pas défini dans l'intervalle

$$[j_{0,x,y}, j_{1,x,y}[= [\min(j_0(x), j_0(y)), \max(j_0(x) + n_x, j_0(y) + n_y)[$$

On obtient alors le vecteur x'_d. De même pour y. On définit alors x_f par concaténation des valeurs pour chaque département :

$$x_f = (x'_1, \ldots, x'_D) \in \mathbb{R}^{D(j_{1,x,y} - j_{0,x,y})}$$

et y_f de même.
On définit le décalage de t d'un vecteur $z \in \mathbb{R}^n$ par

$$\Delta(z, t) = (z_{t+1}, \ldots, z_n, 0, \ldots, 0) \in \mathbb{R}^n$$

où on complète donc par t valeurs nulles.
Le coefficient de corrélation entre x et y décalé de t est alors défini par

$$cc(x, y, t) = \frac{(x_f | \Delta(y_f, t))}{||x_f|| ||\Delta(y_f, t)||}$$

où $(.|.)$ et $||.||$ dénotent le produit scalaire et la norme euclidiens.
Le coefficient de corrélation de x à y est alors

$$cc(x, y) = \max_{t \in \{0, \ldots, t_{max}\}} \{|cc(x, y, t)|\}$$

et on définit le décalage de x à y par

$$\Delta(x, y) = \min\{t \text{ tel que } |cc(x, y, t)| = cc(x, y)\}$$

Par exemple, si on note *travail* la fréquentation des lieux de travail (donnée de Google) et *Rréanimations* le taux de reproduction associé au nombre de patients en réanimation (donnée de Santé Publique France), on obtient les décalages en jours suivants :

$$\Delta(\text{travail}, \text{Rréanimations}) = 19$$

et

$$\Delta(\text{vacances}, \text{Rréanimations}) = 13$$

Et on a les coefficients de corrélation

$$cc(\text{travail}, \text{Rréanimations}) = 0,38 > 0$$

et

$$cc(\text{vacances}, \text{Rréanimations}) = -0,26 < 0$$

ce qui suggère des causalités possibles : une hausse de la fréquentation des lieux de travail semble provoquer une accélération des réanimations 19 jours plus tard, et une période de vacances scolaires semble provoquer un ralentissement des réanimations 13 jours plus tard.
On définit alors les dépendances d'une donnée y comme

$$dep(y) = \{x \text{ tel que } \Delta(x,y) \geq 1 \text{ et } cc(x,y) \geq 0,03\}$$

La valeur $0,03$ est faible et minimise a posteriori l'erreur des prévisions, mais influe peu sur celles-ci.
Ce sont ces dépendances avec leurs décalages qui vont permettre de prévoir un indicateur de l'épidémie, d'abord en prévoyant son taux de reproduction effectif, puis en en déduisant ses valeurs, comme on l'explique dans la suite.

3 Coefficients de prévision linéaire

A présent on dispose de décalages de jours entre certaines des données. On dira qu'une donnée y dépend d'une donnée x si $x \in dep(y)$, i.e. si on a obtenu, à l'étape précédente, un décalage $\Delta(x,y) > 0$ de x à y et une corrélation $cc(x,y)$ suffisante. On considère alors qu'une donnée y sur une période de temps $[j_0, j_1]$ va dépendre des valeurs des données $x_i \in dep(y)$ sur les périodes $[j_0 - \Delta(x_i,y), j_1 - \Delta(x_i,y)]$.
Fixons un département. Appelons A la matrice dont la colonne i est formée des valeurs de x_i dans ce département sur la période $[j_0 - \Delta(x_i,y), j_1 - \Delta(x_i,y)]$, et B le vecteur colonne formé des valeurs de y dans ce département sur la période $[j_0, j_1]$.
On aimerait trouver une famille de coefficients C telle que $AC = B$. Mais il n'y a pas en général de solution à cette équation, car A a plus de lignes (les jours) que de colonnes (les données x_i dont y dépend). On cherche alors à minimiser

$$||AC - B||$$

C'est un problème quadratique convexe, dont la solution s'obtient simplement avec

$$C = ({}^tAA)^{-1}({}^tAB)$$

(si ${}^tAA$ est inversible, ce qui est le cas en pratique).
Avec C on peut prévoir une valeur pour la donnée y le jour $j_1 + 1$, simplement en calculant XC, où X est le vecteur ligne des valeurs des x_i pour les jours $j_1 - \Delta(x_i,y) + 1$:

$$X = (x_{i,j_1-\Delta(x_i,y)+1})_{x_i \in dep(y)}$$

On prévoit alors toutes les données d'un jour en parallèle, puis le jour suivant, etc. Si une donnée n'est pas prévisible (car elle n'a pas de dépendance, ou bien ${}^tAA$ n'est pas inversible), on garde sa valeur précédente. On fait cela pour chaque département.
Cela permet de prévoir des valeurs pour des données dans l'avenir, si on suppose les données des contextes constantes à partir du présent pour l'avenir, sauf pour les données météo que l'on reprend de l'année passée à la même date, et les dates de vacances, qui sont connues pour l'avenir.

4 Taux de reproduction effectif

Les effets visibles de l'épidémie se mesurent avec les indicateurs du système de santé, et sont le résultat essentiellement de contaminations entre un malade et une personne saine. Le taux de reproduction R est difficile à déterminer, car il change chaque jour, et on ne connaît pas tous les malades et qui les contamine.
Pour l'approcher, on déterminera pour chaque indicateur et chaque jour de l'épidémie un taux de reproduction effectif que l'on notera encore R, en utilisant un intervalle sériel

estimé $s = 4,11$ (c'est le nombre moyen de jours entre deux contaminations successives dans une chaîne de contamination). Si l'indicateur de l'épidémie choisi est donné chaque jour par une fonction f (par exemple le nombre quotidien de nouvelles hospitalisations), alors son R vérifie

$$R = e^{s\frac{f'}{f}}$$

de sorte que $f(x+s) = R(x)f(x)$.
On obtient cette expression simplement en considérant f comme localement exponentielle, ce qui est le cas lors d'une épidémie : en supposant qu'un malade contamine α personnes en moyenne en un jour, on a, pour tout instant x,
$f(x+1) = \alpha f(x)$, donc $f(x+1) - f(x) = (\alpha - 1)f(x)$,
donc, en identifiant la dérivée et dérivée discrète, $f' = (\alpha - 1)f$, donc $\dfrac{f'}{f} = \alpha - 1$,
puis en intégrant on obtient $\ln f(x) = (\alpha - 1)x + c$, donc $f(x) = e^{(\alpha-1)x}f(0)$, donc $f(x+s) = e^{(\alpha-1)s}f(x)$ et enfin $f(x+s) = e^{s\frac{f'(x)}{f(x)}}f(x)$,
d'où $R = e^{s\frac{f'}{f}}$.

La valeur $R(x)$ du taux de reproduction varie en effet chaque jour x : elle est maximale au début de l'épidémie, et décroît ensuite jusqu'à 0, quand la population est immunisée et que l'épidémie disparaît.
Notons que la dérivée discrète

$$f' : x \mapsto f(x+1) - f(x)$$

est en pratique calculée sur les valeurs lissées 2 fois sur 7 jours.
Dans le cas continu, on peut retrouver la valeur de l'indicateur au jour j à partir de la fonction R ainsi :

$$f(j) = e^{\int_{j_0}^{j} \frac{1}{s}\ln R(t)\mathrm{d}t} f(j_0)$$

On adapte cette formule au cas discret par intégration discrète et correction globale ensuite pour la partie correspondant aux données réelles (essentiellement on normalise pour obtenir l'intégrale réelle de f sur le passé et sa valeur correcte au présent).

5 Résumé de la méthode

Pour résumer, le processus de prédiction des indicateurs est donc le suivant :

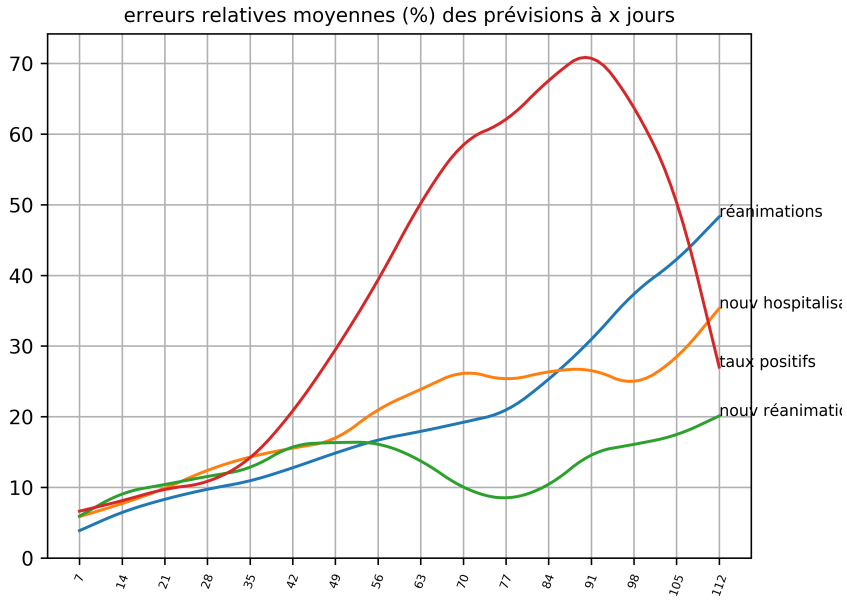
1. calcul des taux de reproduction effectif R quotidiens des indicateurs.
2. calcul des corrélations et décalages entre contextes et taux de reproduction effectifs des indicateurs
3. on en déduit les contextes dont dépendent ces taux.
4. à partir de ces contextes, calcul des coefficients de prévision linéaire des taux R.
5. avec les coefficients de prévision linéaire, calcul des prévisions des taux R dans l'avenir.
6. par intégration discrète et normalisation sur le passé, calcul des prévisions des indicateurs dans l'avenir.

6 Résultats et évaluation

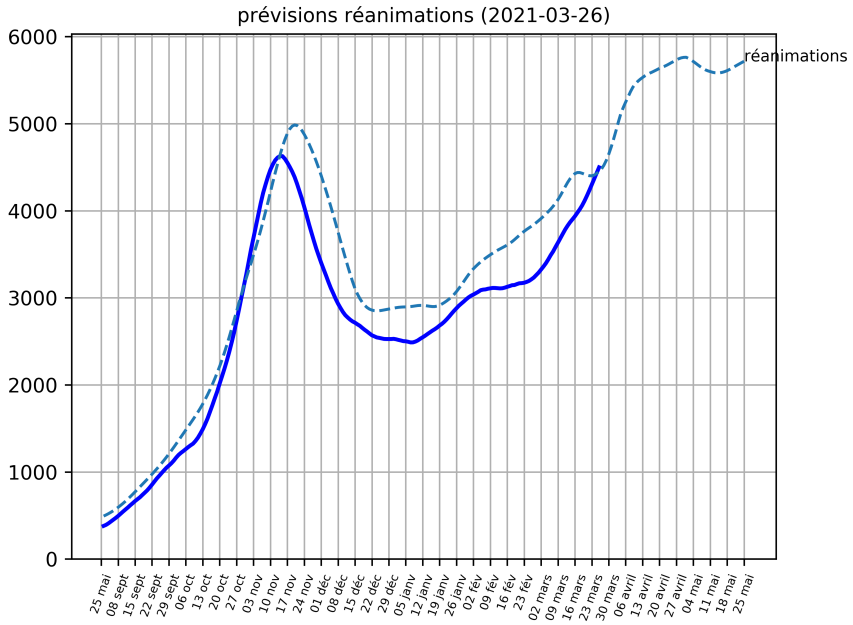
Pour évaluer la méthode, on l'applique dans le passé. On utilise les coefficients de prévision linéaire calculés sur la période de l'épidémie (d'avril 2020 à mars 2021) pour prévoir les valeurs des indicateurs 7 jours après, par exemple, le 1er décembre 2020, puis 14 jours

après, etc, jusqu'à 3 mois après (1er mars). Puis on calcule les erreurs relatives entre les valeurs prévues et les valeurs réelles.
Au 23 mars 2021, l'erreur relative moyenne entre les valeurs réelles et les valeurs prévues depuis 4 mois est

— à 7 jours : 4% pour les réanimations (6% pour l'ensemble des indicateurs)
— 14 jours : 7% (8%)
— 28 jours : 10% (11%)
— 2 mois : 18% (27%)
— 3 mois : 31% (36%).

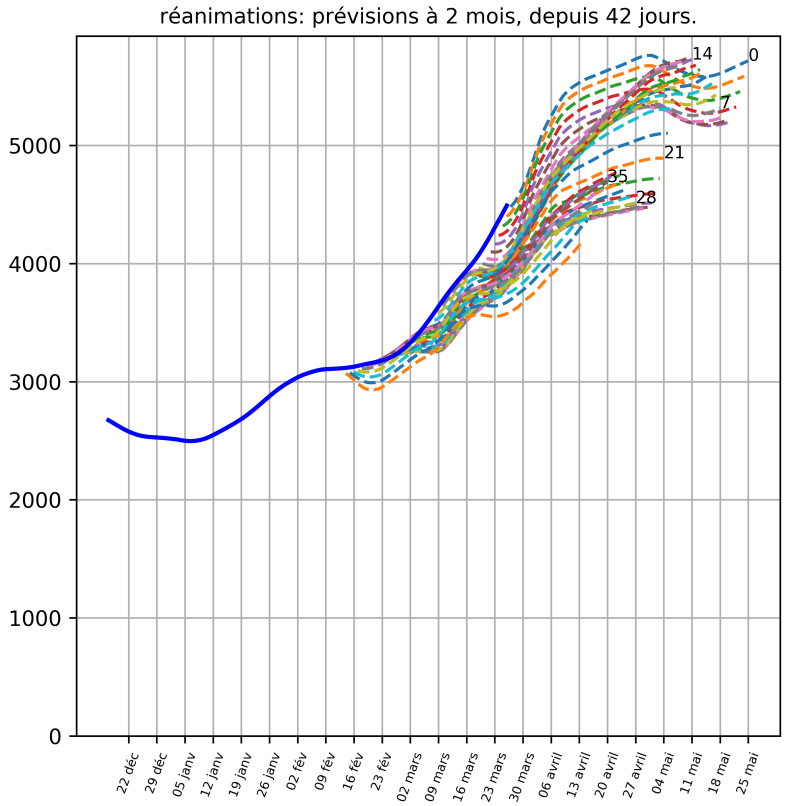


Le pic du nombre de réanimations de la vague actuelle est ainsi prévisible pour la fin avril, avec un plateau, sans que l'on puisse savoir ce qui se passera ensuite :



Données réelles : trait plein, données prévues : trait pointillé, pour 86 départements

En utilisant les coefficients de prévision calculés à partir des données avant le début de la prévision, on obtient, pour des prévisions sur 2 mois depuis les 42 derniers jours :



Prévisions sur 2 mois depuis 42 jours,
coefficients de prévision limités au passé.

On peut aussi évaluer la méthode en la comparant avec une approximation linéaire (par la tangente à la courbe) ou quadratique (utilisation des dérivées premières et seconde au présent pour approximer la courbe par un polynôme de degré 2). Ces approximations sont toujours moins bonnes, avec des erreurs de 8% à 7 jours, 17 et 21% à 14 jours, et plus de 50% à partir de 50 jours.
On peut aussi comparer avec les prévisions à court terme de Paireau et al [2] : elles donnent une erreur de 6% à 7 jours et 11% à 14 jours pour les lits de soins critiques.
Les prévisions complètes sont mises à jour quotidiennement sur la page web `https://cp.lpmib.fr/medias/covid19/_synthese.html`

2. Projection à court terme des besoins hospitaliers pour les patients COVID-19 `https://modelisation-covid19.pasteur.fr/realtime-analysis/hospital/` et An ensemble model based on early predictors to forecast COVID-19 healthcare demand in France `https://hal-pasteur.archives-ouvertes.fr/pasteur-03149082`