Mesure, estimation et représentations de la Covid-19

Par Éric GUICHARD

Maître de conférences HdR à l'Université de Lyon et chercheur au laboratoire Triangle de l'ENS de Lyon et du CNRS, et de l'IXXI

Et Patrice ABRY

Directeur de recherche au CNRS et chercheur au laboratoire de physique de l'ENS de Lyon et de l'IXXI

Comme nombre de chercheurs, nous avons mis nos savoirs au service de la lutte contre la Covid-19. Dès février 2020, nous avons produit des graphiques quotidiens afin de comparer l'évolution de la pandémie selon les pays. Nous avons aussi fédéré un réseau scientifique qui élaborait et documentait des analyses pertinentes.

Ensuite, nous avons conçu un outil qui estime au mieux le taux de reproduction de la pandémie dans les pays du monde et dans les départements français, en n'utilisant que le nombre des infections quotidiennes observées dans chaque territoire. Disposant de données de qualité limitée, notre modèle s'est focalisé sur la correction des erreurs, *via* des méthodes dites de « problèmes inverses ». Bien qu'il ne permette d'estimer que le taux de reproduction du jour, notre outil permet une évaluation de tendance à court terme.

Enfin, pour analyser au mieux l'évolution spatiale et temporelle de la pandémie, nous avons réalisé une carte animée et interactive intégrant la production de graphiques permettant la comparaison entre deux pays. Nous concluons cet article en abordant quelques pistes épistémologiques.

Premiers journaux

Dès janvier 2020, circulait dans les milieux scientifiques l'hypothèse d'une pandémie mondiale à venir. Ce qu'officialisa l'OMS, le 11 mars 2020, puis la France, qui décida d'un confinement généralisé le 17 mars.

Les premières « données » relatives à la pandémie étaient présentées de façon anxiogène : le site de l'Université Johns Hopkins (https://systems.jhu.edu) affichait une carte du monde avec d'énormes cercles rouges sur fond noir pour signaler les morts de la Covid-19. Sur le site https://coronavirus.politologue.com, l'accroissement du nombre total des morts et des confirmés semblait exponentiel.

Pour lutter contre cette tendance, nous avons décidé de représenter sous forme d'un graphique le nombre quotidien (et non cumulé) de décès rapporté à la population du pays considéré. Les pays étaient regroupés par lots de quatre à six pays voisins pour faciliter les comparaisons entre eux. Ces graphiques étaient commentés dans un journal diffusé en ligne de façon automatisée, publié tous les matins, il intégrait aussi des notes méthodologiques et des références issues du Web. Le premier journal a été publié le 26 février 2020, son automatisation a été achevée le 27 mars.

Ces journaux[1] diffusent rapidement des résultats qui sont peu évoqués : l'Espagne s'avère plus menacée que l'Italie ou la France, et ce dès le 20 mars (voir la Figure 1 de la page suivante) ; ou encore, sont mises en avant les difficultés des pays à organiser un recensement rigoureux des décès, avec parfois des « morts négatifs ». Ces journaux nous ont aidé à construire un débat au sein de l'Institut rhônalpin des systèmes complexes – IXXI (http://www.ixxi.fr) – et une liste de discussion, theuth@listes.univ-rennes1.fr. Ils nous ont incités à participer au débat public[2] et à développer des contacts avec les Académies[3]. Ces échanges nous ont alors amenés à inventer des indicateurs permettant de prévoir l'évolution de la pandémie.

[1] Ils sont toujours fonctionnels. Voir : http://barthes.enssib.fr/coronavirus

[2] Voir l'article co-écrit avec Pierre-Antoine Chardel et Valérie Charolles, du 11 mai 2020, https://www.revuepolitique.fr/stopcovid-une-application-problematique-sur-le-plan-ethique-et-politique

[3] Voir l'article de Mireille Delmas-Marty, https://academiesciencesmoralesetpolitiques.fr/2020/05/14/mireille-delmas-marty-stopcovid-une-application-democratiquement-fragile

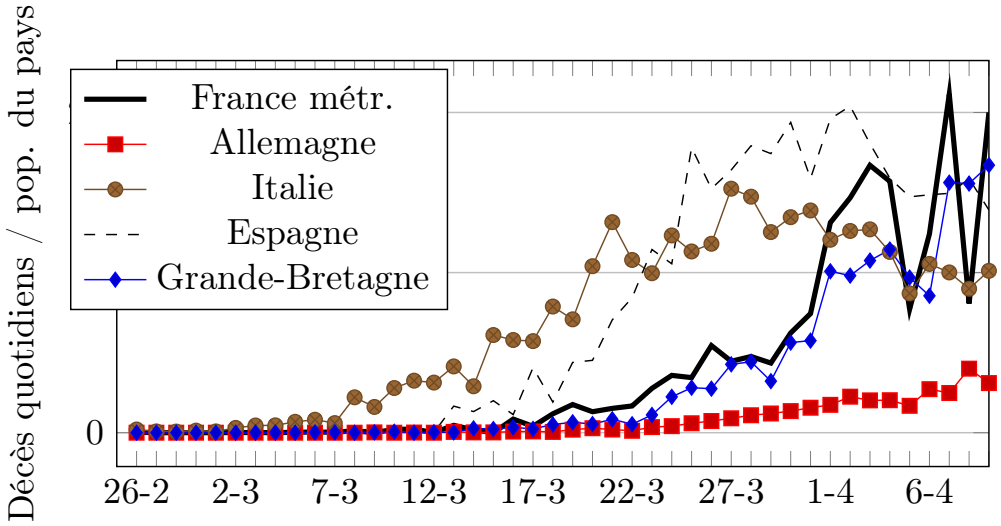

Figure 1 : Le taux de mortalité en Espagne dépasse celui de l'Italie, à la date du 20 mars. On constate les errances des recensements réalisés dans les différents pays (réalisée par É. Guichard).

Estimer le taux de reproduction de la pandémie

Pour analyser une pandémie, sont souvent utilisés des « modèles à compartiments », dont l'exemple SIR (individu Susceptible d'être sain, Infecté, guéri (correspondant à R pour Recovered) constitue la référence. Ces modèles sont surtout efficaces *a posteriori* et sont peu utilisables lorsque l'on dispose de données de qualité limitée, comme ce fut le cas avec la Covid-19.

Le taux de reproduction de la pandémie, qui permet d'estimer au moment t le nombre d'individus qu'une personne infectée peut contaminer, noté $R_{(t)}$, est certes plus fruste, mais il rend bien compte de l'intensité de la pandémie au fil du temps : ainsi, plus $R_{(t)}$ se situe au-dessus de 1, plus la pandémie accélère ; plus il se rapproche de 0, plus elle ralentit.

Pour l'estimer, une dizaine de chercheurs de l'ENS de Lyon, en lien étroit avec l'IXXI, ont partagé leurs savoirs. Outre des points théoriques ardus à résoudre, leur souci principal était lié à la faible qualité et au petit nombre des données. Ces « données » étaient en outre fortement corrompues (données manquantes, valeurs aberrantes, comptes négatifs, réajustements rétrospectifs), des défauts variables selon les pays. Aussi se sont-ils focalisés sur le traitement des données aberrantes.

Une explication destinée au public éclairé de la méthode employée est disponible à l'URL suivante : https://theconversation.com/comment-estimer-levolution-du-covid-19-malgre-des-donnees-de-contaminations-de-qualite-limitee-177777. Sa version « technique » est accessible à : https://hal.inria.fr/hal-02921836/document

Le casse-tête des données

Si la situation d'urgence initiale pouvait expliquer la faible qualité des « données Covid » recensées par les agences de santé publique, force est de constater que, plus de 24 mois après le début de la pandémie, cette qualité ne s'est guère améliorée. De plus, les organismes qui collectent ces données ne s'engagent pas tous à le faire sur le long terme. Par exemple, au début de notre travail, nous avons récupéré automatiquement auprès de l'European Centre for Disease Prevention and Control[4] le nombre des nouvelles infections quotidiennes de plus de 200 pays du monde. Or, le 7 décembre 2020, cet organisme a, sans information préalable, choisi de dégrader l'information sous sa forme hebdomadaire.

Désormais, nous utilisons deux sources de données pour produire des estimations du $R_{(t)}$ concernant plus de 200 pays du monde et les 101 départements français :

- La première source est celle qui nous a servi à produire nos journaux automatisés : elle émane de la Johns Hopkins University, qui est devenue le centre de référence de la collecte quotidienne des données de plus de 250 pays et territoires, et ce dès les premiers jours de la pandémie. Les données sont disponibles peu après minuit, heure de New York.
- La seconde correspond au site de Santé publique France, qui publie depuis le 19 mars 2020, chaque jour vers 19 heures (heure française), des données hospitalières pour les 101 départements français : nouvelles infections ayant induit une entrée à l'hôpital, transferts en réanimation et décès survenus à l'hôpital[5].

[4] Voir, par exemple : https://www.ecdc.europa.eu/sites/default/files/documents/covid-19-geographic-disbtribution-worldwide.xlsx

[5] https://www.data.gouv.fr/fr/datasets

Estimation de l'évolution spatio-temporelle du taux de reproduction de la Covid-19

Pour estimer le $R_{(t)}$ nous n'utilisons que le nombre de nouvelles infections quotidiennes.

Le modèle épidémiologique relatif au taux de reproduction, R

Le modèle épidémiologique utilisé se fonde sur celui développé par Cori, Ferguson, Fraser et Cauchemez (2013) et reposant sur deux arguments principaux : 1) conditionné à la connaissance des nombres de nouvelles infections des jours passés, le nombre Z_t de nouvelles infections du jour t suit une loi de Poisson ; 2) ce paramètre dépend donc du nombre des infections des jours passés, mais aussi de la fonction sérielle d'intervalles ϕ_t et du taux de reproduction courant R_t que l'on cherche à estimer.

La fonction sérielle d'intervalles ϕ_t modélise la distribution des délais (aléatoires) entre la survenue des symptômes chez un sujet infecté et celle des symptômes chez ceux qu'il a contaminés (Cori, Ferguson, Fraser et Cauchemez, 2013 ; Obadia, Haneef et Boëlle, 2012 ; Thompson *et al.*, 2019 ; Liu, Ajelli, Aleta, Merler, Moreno et Vespignani, 2018). Cette fonction ϕ est modélisée par une fonction Gamma, avec des paramètres correspondant à un délai moyen d'infection de 6,6 jours pour un écart-type de 3,5 jours et de forts risques d'une infection d'autrui dans un laps de temps de 3 à 10 jours après l'apparition des symptômes (Ma, Zhang, Zeng, Yun, Guo, Zheng, Zhao, Wang et Yang, 2020 ; Riccardo, Ajelli, Andrianou, Bella, Del Manso, Fabiani, Bellino, Boros, Urdiales, Marziano *et al.*, 2020 ; Guzzetta *et al.*, 2020). Ce modèle suggère de construire un estimateur du maximum de vraisemblance, qui se lit comme le rapport du nombre des infections du jour, Z_t, sur une moyenne des nouvelles infections observées au cours des jours passés, pondérées par la fonction ϕ.

Cette estimation obtenue à partir des données réelles relatives à la Covid-19 est illustrée par le graphique central de la Figure 2 (traits noirs) ci-dessous. Le taux de reproduction estimé est trop erratique pour être utilisable par qui souhaite surveiller l'évolution d'une épidémie.

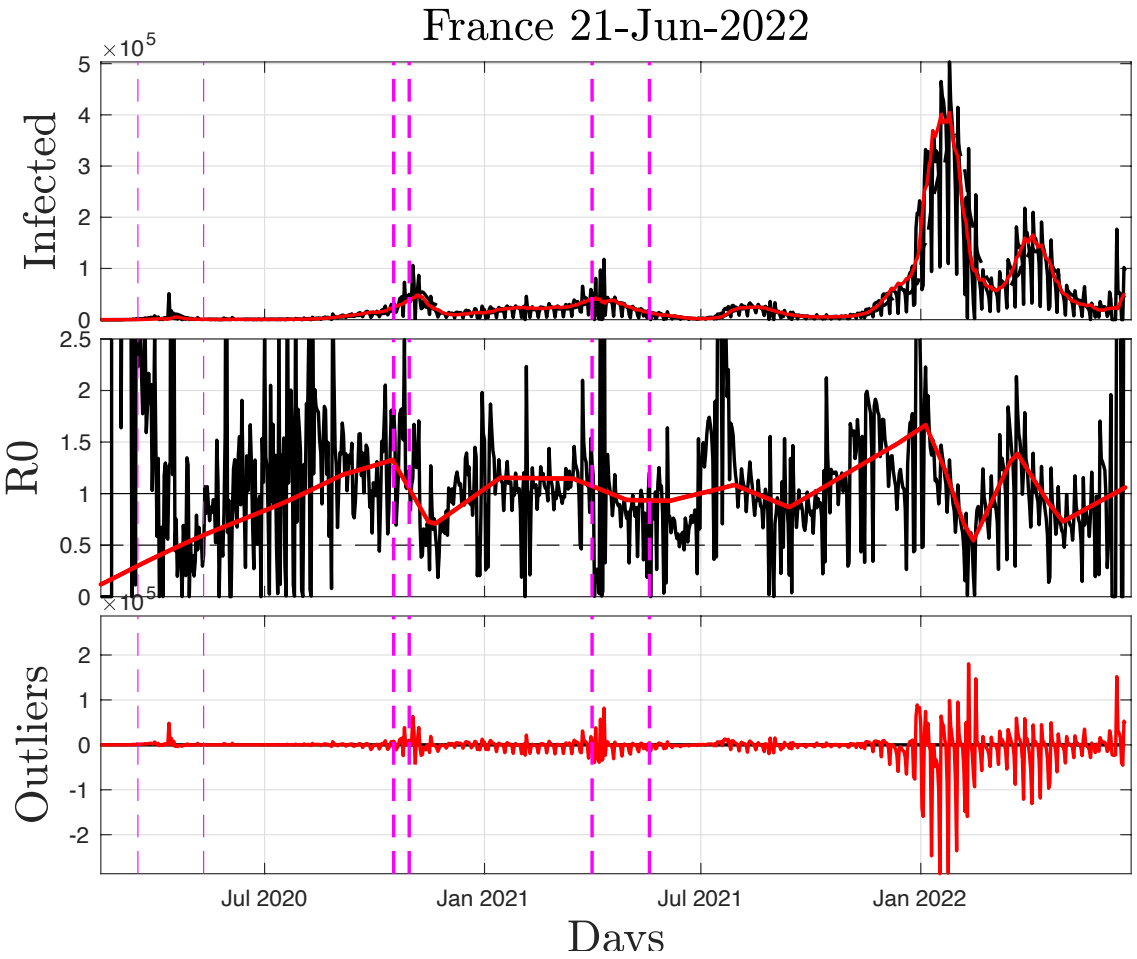

Figure 2 : Nouvelles infections et estimations du R_t pour la France au cours du temps – Source : Santé publique France.
Notes : Le graphique supérieur correspond aux nouvelles infections quotidiennes ; sur ce graphique, la ligne rouge pleine montre le décompte des nouvelles infections quotidiennes corrigé par application de notre méthode. Le graphique central restitue l'estimation par la méthode directe (en noir) et celle obtenue avec notre méthode (en rouge). Le graphique inférieur correspond à l'estimation de O_t, la corruption des données.

Estimation robuste du taux de reproduction pour un seul territoire

Nous avons alors décidé d'améliorer cette estimation (Abry, Pustelnik, Roux, Jensen, Flandrin, Gribonval, Lucas, Guichard, Borgnat et Garnier, 2020 ; Pascal, Abry, Pustelnik, Roux, Gribonval et Flandrin, 2021) à partir de la seule observation de Z_t. Pour cela, nous avons ajouté deux ingrédients clés.

Nous avons postulé que Z_t restait bien, conditionnellement au passé, modélisé par une loi de Poisson, dont les paramètres dépendaient en plus d'une quantité inconnue O_t, appelée outlier, qui modélisait une éventuelle corruption (quelle qu'en soit la nature) de l'observation Z_t des nouvelles infections enregistrées à la date t. Il s'agissait alors d'estimer chaque jour à la fois le taux de reproduction R_t et le nombre des données corrompues O_t.

Ensuite, en respectant le principe fondamental de la théorie de l'estimation, nous avons appliqué aux « données » le modèle épidémiologique, que nous avons assorti d'un ensemble de contraintes sur les estimées de R_t et O_t, notées $\hat{R}_t$ et $\hat{O}_t$.

Les contraintes imposées à $\hat{R}_t$ sont de deux ordres. D'abord, si R_t est positif ou nul, $\hat{R}_t$ doit lui aussi l'être. Ensuite, R_t évoluant faiblement ou régulièrement au cours du temps, nous avons soumis $\hat{R}_t$ à une évolution linéaire par morceaux, avec des points de raccord à des dates *a priori* inconnues.

L'absence de modèle commun à tous les pays nous a imposé d'adopter pour l'estimation $\hat{O}_t$ de O_t une structure parcimonieuse : la corruption des données se produit à des dates isolées ; ces données sont quelconques et sans structure prédéfinie.

L'estimation de R_t et celle de O_t reposent alors sur l'écriture d'une fonctionnelle qui met en compétition l'adéquation des données au modèle et les contraintes imposées à $\hat{R}_t$ et $\hat{O}_t$. Elle intègre deux hyperparamètres de régularisation qui permettent d'arbitrer l'importance relative de l'attache données-modèle et de celle des contraintes de régularité temporelle, de positivité et de parcimonie des outliers.

Nous avons montré dans deux études (Abry, Pustelnik, Roux, Jensen, Flandrin, Gribonval, Lucas, Guichard, Borgnat et Garnier, 2020 ; Pascal, Abry, Pustelnik, Roux, Gribonval et Flandrin, 2021) que notre fonctionnelle était convexe et non différentiable, qu'elle était efficacement minimisable grâce à des algorithmes itératifs à opérateurs proximaux (un outil enrichissant la traditionnelle descente de gradient) et qu'elle fournissait des estimées de R_t et de O_t robustes et fiables malgré la faible qualité des données disponibles.

L'estimation de R_t obtenue grâce à cette approche est illustrée (et comparée à celle obtenue par la méthode directe) par la Figure 2 de la page précédente, ainsi que celle obtenue pour O_t, laquelle représente la corruption quotidienne des données. On voit que l'estimation obtenue en recourant à cette approche est beaucoup plus régulière que l'estimée directe : elle évolue linéairement au cours du temps, sauf en quelques points de rupture ressortant des données et où la tendance change. Cette approche, bien plus compatible avec une appréciation de la variation épidémique, permet donc une véritable surveillance de l'évolution de l'intensité de la pandémie.

Dans notre minimisation de la fonctionnelle, le réglage des hyperparamètres est un élément crucial. Logiquement, il devrait être fait pays par pays. Mais une analyse dimensionnelle de la fonctionnelle nous a permis de proposer un réglage commun à tous les pays et indépendant de la population considérée ou de l'intensité de la pandémie. L'estimation peut ainsi être mise à jour automatiquement et quotidiennement pour environ 200 pays grâce à un seul réglage d'hyperparamètres, qui est effectué par les auteurs de l'article.

Estimation du taux de reproduction pour des territoires connectés entre eux

Dans certains cas (notamment la France et ses départements), les territoires sont administrés par une même autorité et sont parfois voisins, quand ils ne sont pas très interconnectés. De tels territoires ne peuvent présenter des R très différents au titre d'une même journée. Nous avons alors imposé une contrainte de régularité spatiale aux estimées $\hat{R}_t$ de ces territoires (typiquement *via* un graphe). Cette nouvelle fonctionnelle se minimise en recourant à la même stratégie que celle présentée précédemment. Elle facilite l'analyse spatiale de la pandémie.

Vers la prévision de l'évolution de la pandémie

Rappelons-le, notre approche permet de réaliser une estimation de R à une date courante et non une prévision de R dans le futur. Cependant, l'estimation linéaire par morceaux donne aussi la tendance, à la hausse ou à la baisse, de $\hat{R}_t$ autour de la date courante. Cela est manifeste lorsque $\hat{R}_t$ change de tendance, par exemple lorsqu'à une hausse succède à une baisse. Nous ne sommes donc pas dans une logique de *forecasting* (prévision), mais dans celle du *nowcasting* (estimation de tendance à court terme).

Cartographie de la Covid-19

Nous avons ensuite décidé de visualiser l'évolution spatiale et temporelle de la pandémie. Notre expérience tirée de l'exploitation des premiers journaux nous a aidés à concevoir des cartes animées, interactives et automatisées non seulement pour la France et ses départements, mais aussi pour les autres pays du monde. Nous présentons en page suivante la carte du monde, finalisée[6], visualisable à l'URL suivante : http://barthes.enssib.fr/coronavirus/cartes/prod/monde/monde.html (voir la Figure 3 qui correspond à une copie d'écran de celle-ci).

Trois variables peuvent être mentionnées pour illustrer cette carte : le $R_{(t)}$[7], le taux des infections confirmées du jour (pour 10 000 habitants) et le taux de mortalité Covid

[6] L'atlas animé de la France sera disponible avant la fin de l'été 2022.

[7] Pour le monde, il correspond à une estimation robuste issue de notre modèle général. Pour la France, le même taux est affiché, mais en intégrant la contiguïté territoriale.

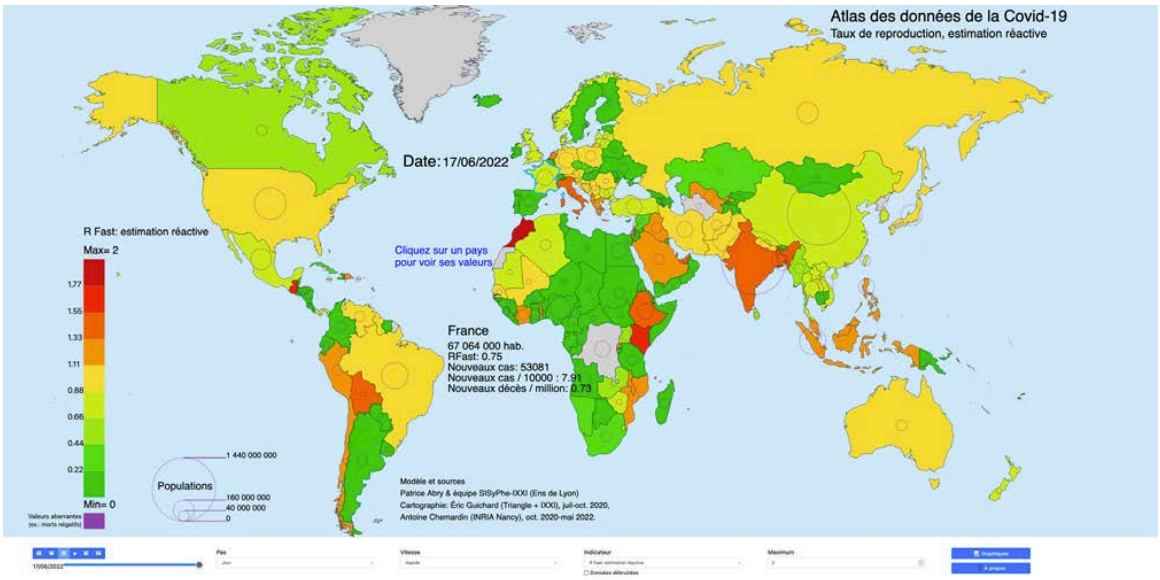

Figure 3 : Page d'entrée de l'atlas animé mondial de la Covid-19 (site géré par Antoine Chemardin).

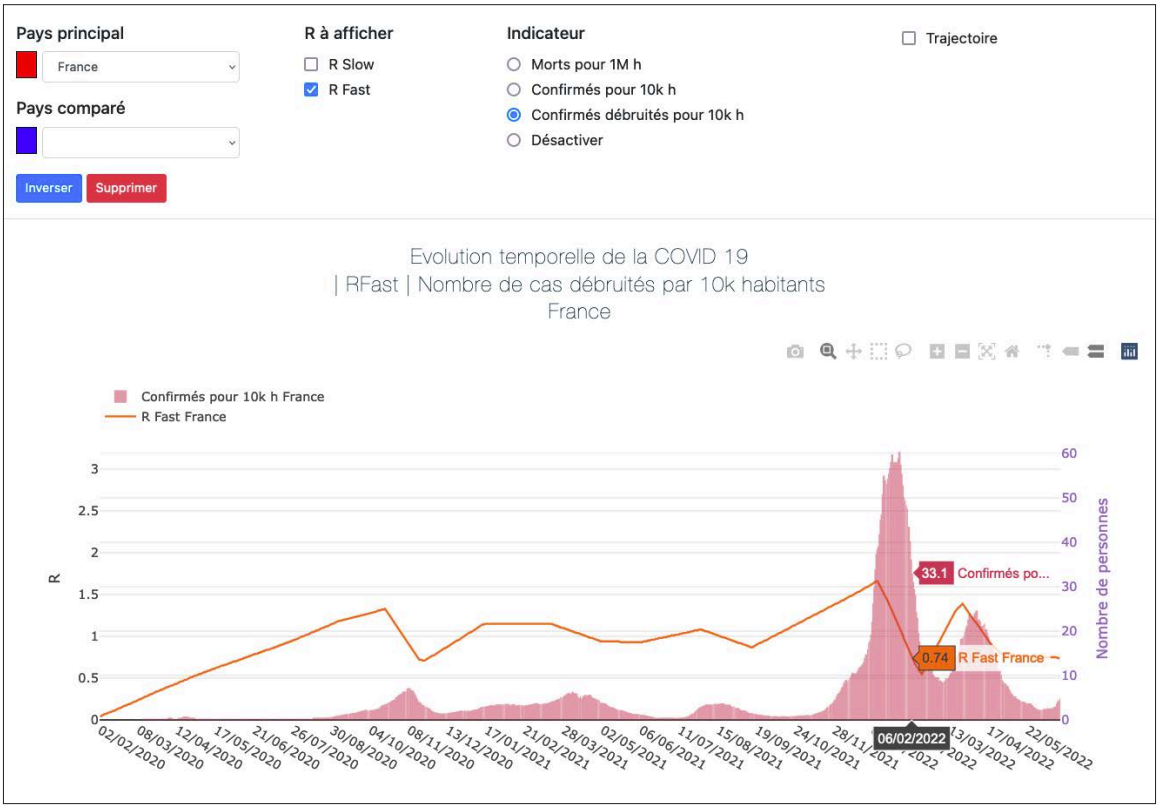

Figure 4 : Dans ce graphique, le lien entre évolution du *R(t)* et celle du taux de confirmés est bien visible. Le survol d'un point du graphique fait s'afficher les informations relatives à ce point (création de É. Guichard, P. Abry et A. Chemardin).

du jour (pour un million d'habitants). Lorsque le curseur passe sur un pays, les valeurs des trois variables précitées sont affichées, ainsi que la population totale du pays et le nombre de nouveaux cas (absolus) détectés du jour.

L'animation de la carte se fait en déplaçant le curseur bleu jusqu'à la date désirée, puis en cliquant sur l'icône de lancement de la vidéo. On peut en choisir la vitesse, le pas (1 jour, 1 semaine ou 1 mois), l'arrêter à tout moment, etc.

Comme la valeur maximale de chaque variable est initialement prédéfinie (2 pour le $R_{(t)}$, 6 pour le taux de confirmés et 10 pour celui de la mortalité), nous l'avons rendu modifiable, ce qui peut apporter un confort de visualisation en cas de vague pandémique (trop de pays apparaissant en rouges pour le taux d'infections confirmées (la vague de janvier-février 2022) ou d'une décrue généralisée (trop de pays apparaissant en vert). La légende s'adapte alors au maximum choisi par l'utilisateur.

Avec une telle carte et les paramètres associés, il est alors aisé de suivre pas à pas l'évolution mondiale de la pandémie, de repérer ses déplacements d'un continent à l'autre, les pays fortement menacés, etc. Un bouton « À propos » précise les sources, les méthodes et les liens avec les sites de données qui sont à l'origine de la réalisation de notre modèle et de la carte.

Pour autant, une fois la pandémie « domestiquée » (meilleure connaissance de celle-ci, découverte et diffusion de vaccins, etc.), cette carte s'avère plus informative qu'analytique. Par exemple, elle n'affiche pas le nombre des morts enregistrés au début de l'épidémie quand nous nous focalisons sur des temps plus récents. Pour remédier à ce type d'inconvénients et faciliter les comparaisons entre les pays, nous l'avons complétée de graphiques *ad hoc*, visualisables *via* le bouton « Graphiques ». S'affichent alors directement le $R_{(t)}$ et la variable choisie, par exemple les taux d'infections confirmées ; ce que montre la copie d'écran correspondant à la Figure 4 située en page précédente.

Nous espérons que ces cartes et les documents qui les complètent aideront le public éclairé à mieux comprendre la pandémie, à en repérer les moments marquants et à se familiariser avec le $R_{(t)}$ pour construire des raisonnements étayés. Pour les experts, nous avons ajouté un graphe de l'espace des phases, qui apporte un éclairage sur la structure des « vagues » (voir la Figure 5 ci-dessous).

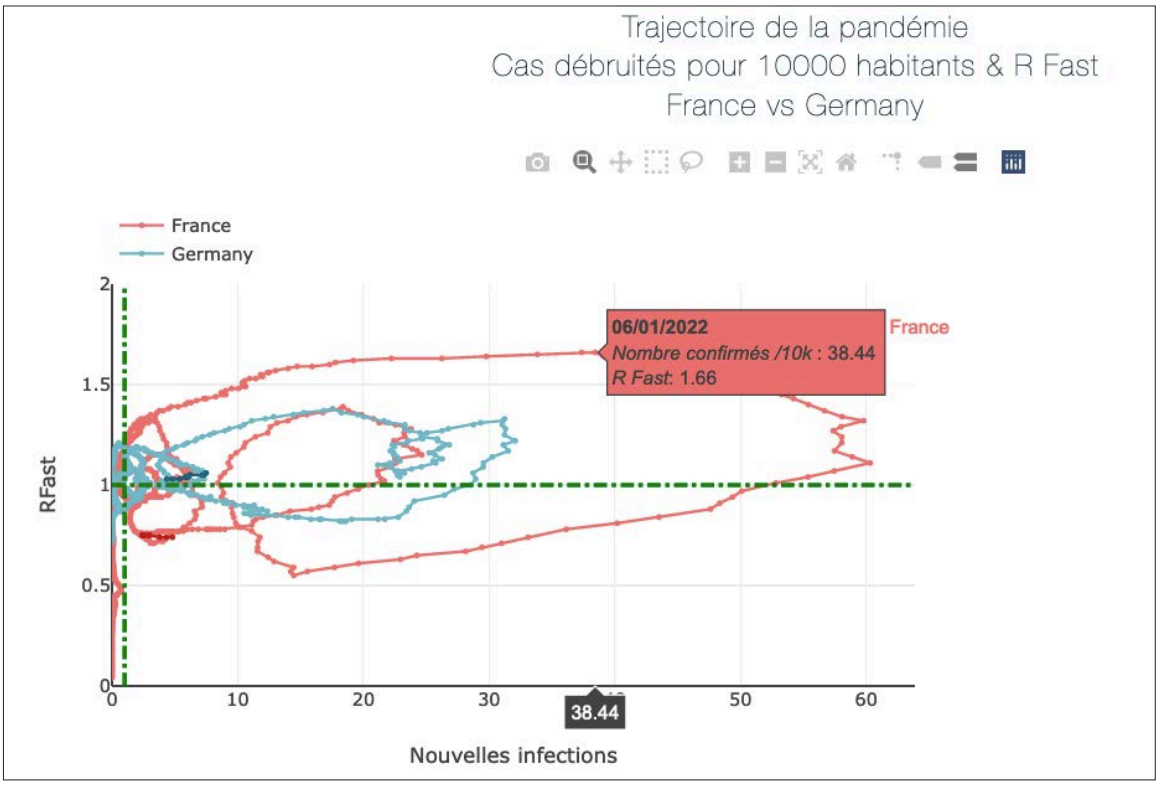

Figure 5 : L'option Trajectoire permet de visualiser les vagues de la pandémie. Les pays ici décrits sont la France (en rouge) et l'Allemagne (en bleu) (création de É. Guichard, P. Abry et A. Chemardin).

Conclusion

À l'instar des journaux, ces cartes traduisent des préoccupations sociales. Elles sollicitent aussi le *design* et l'informatique, sans oublier les compétences mathématiques et physiques pour modéliser le $R_{(t)}$. Nous sommes donc ici dans une logique d'interdisciplinarité, ce dont nous avons pu témoigner en participant à diverses tables rondes[8], conférences, séminaires[9]…, et au travers de publications (Abry, 2022).

[8] Se référer au colloque « Philosophies des numériques, des techniques et des sciences », Lyon, 30 et 31 août 2021, http://barthes.enssib.fr/Phi-NTS

[9] Voir « La Covid-19 : regards et questions interdisciplinaires », IXXI, Lyon, printemps 2021, http://www.ixxi.fr/agenda/seminaires/la-covid-19-regards-et-questions-interdisciplinaires.

Nous avons été aussi contraints de faire de la science dans l'urgence, usant de tous les moyens à notre disposition. Nous pensons l'avoir produite dans une logique typique de ce que Dewey appelle l'enquête philosophique, sans pour autant oublier la critique épistémologique et sociologique sur la façon dont la science est conduite et organisée, ni même la critique de cette critique. Cette pratique scientifique intègre de « formidables méthodes d'observation, d'expérimentation, de réflexion et de raisonnement », ce que Dewey appelle l'intelligence (Dewey, 2014), pour rappeler que l'intellection du monde contemporain n'est pas qu'affaire d'esprits purs. Ce que précisaient également Gilles Gaston Granger et Jack Goody, qui insistaient sur l'importance de la technique dans la construction de la pensée.

Cette forme de réflexivité inséparable de la démarche scientifique est bien connue des sciences sociales (Bourdieu, 2001) ; elle favorise l'interdisciplinarité. Reste un point parfois passé sous silence par une sociologie des sciences qui insiste sur l'intérêt personnel des chercheurs : la possibilité de produire de façon pragmatique une science citoyenne, une science en société. Nous nous réjouirions que notre témoignage y contribue.

Bibliographie

ABRY P. (2022), « La cartographie de la Covid-19 vue par un physicien », *in* GUICHARD Éric (dir.), *Études digitales : cartographie et visualisation. Regards d'épistémologues et de concepteurs*, n°10. Garnier, pp. 145-153.

ABRY P., PUSTELNIK N., ROUX S., JENSEN P., FLANDRIN P., GRIBONVAL R., LUCAS CH.-G., GUICHARD E., BORGNAT P. & GARNIER N. (2020), "Spatial and temporal regularization to estimate Covid-19 reproduction number R(t): Promoting piecewise smoothness via convex optimization", *PLOS ONE* 15 (8), Public Library of Science San Francisco, CA USA, e0237901.

BOURDIEU P. (2001), *Science de la science et réflexivité*, Paris, Raisons d'agir.

CORI A., FERGUSON N. M., FRASER C. & CAUCHEMEZ S. (2013), "A new framework and software to estimate time-varying reproduction numbers during epidemics", *American Journal of Epidemiology* 178 (9), Oxford University Press, pp. 1505-1512.

DEWEY J. (2014), *Reconstruction en philosophie*, Gallimard/Folio.

GUZZETTA *et al.* (2020), "The impact of a nation-wide lockdown on COVID-19 transmissibility in Italy", arXiv :2004.12338 (q-bio.PE).

LIU Q.-H., AJELLI M., ALETA A., MERLER S., MORENO Y. & VESPIGNANI A. (2018), "Measurability of the epidemic reproduction number in data-driven contact networks", *Proceedings of the National Academy of Sciences* 115 (50), National Academy of Sciences, pp. 12680-12685, doi :10.1073/pnas.1811115115,https://www.pnas.org/content/115/50/12680

MA Shujuan, ZHANG Jiayue, ZENG Minyan, YUN Qingping, GUO Wei, ZHENG Yixiang, ZHAO Shi, WANG Maggie H. & YANG Zuyao (2020), "Epidemiological parameters of coronavirus disease 2019: A pooled analysis of publicly reported individual data of 1155 cases from seven countries", *American Journal of Epidemiology* 178 (9), Oxford University Press, pp. 1505-1512.

OBADIA T., HANEEF R. & BOËLLE P.-Y. (2012), "The R0 package: A toolbox to estimate reproduction numbers for epidemic outbreaks", *BMC Medical Informatics and Decision Making* 12 (1), 147.

PASCAL B., ABRY P., PUSTELNIK N., ROUX S., GRIBONVAL R. & FLANDRIN P. (2021), "Nonsmooth convex optimization to estimate the Covid-19 reproduction number space-time evolution with robustness against low quality data", arXiv 2109.09595.

RICCARDO Flavia, AJELLI Marco, ANDRIANOU Xanthi D., BELLA Antonino *et al.* (2020), "Epidemiological characteristics of COVID-19 cases in Italy and estimates of the reproductive numbers one month into the epidemic", medRxiv :2020.04.08.20056861, Cold Spring Harbor Laboratory Press, doi :10.1101/2020.04.08.20056861, https://www.medrxiv.org/content/10.1101/2020.04.08.20056861v1

THOMPSON R. N. *et al.* (2019), "Improved inference of time-varying reproduction numbers during infectious disease outbreaks", *Epidemics* 29, 100356, https//doi.org/10.1016/j.epidem.2019.100356, http://www.sciencedirect.com/science/article/pii/S1755436519300350