
**Les chiffres au cœur de la crise du
coronavirus (2)**

Sommaire

Statistique et société

Volume 10, Numéro 2

7 **Éditorial**

Emmanuel DIDIER

Rédacteur en chef de Statistique et société

9 **La pandémie de Covid-19 : un moment charnière pour l'ouverture des données en France**

Samuel GOËTA

Co-fondateur de Dataactivist ; Maître de conférences associé en sociologie, Sciences Po Aix (Mesopolhis UMR 704)

31 **Biostatistics and the COVID-19 Pandemic in Belgium, in 2020 and 2021**

Geert MOLENBERGHS

Interuniversity Institute for Biostatistics and statistical Bioinformatics, Data Science Institute, Hasselt University and KU Leuven, Belgium

45 **La politique des grands nombres et les grands nombres de la politique dans la pandémie au Brésil**

Eugênia MOTTA

Professeure à l'Université de l'État de Rio de Janeiro et chercheure au Musée National, Université Fédérale de Rio de Janeiro

Victor ALVES MOURÃO

Professeur à l'Université Fédérale de Viçosa

Alexandre DE PAIVA RIO CAMARGO

Professeur à l'Université Candido Mendes et chercheur à l'Institut Universitaire de Recherche de Rio de Janeiro

65 **Pandémie, statistiques et fédéralisme : la COVID au Canada**

Jean-Guy PRÉVOST

Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal

Sommaire

Statistique et société

Volume 10, Numéro 2

- 81 Une conversation avec Yves Escoufier**
Yves ESCOUFIER
Université de Montpellier
Gilbert SAPORTA
Conservatoire national des arts et métiers, Paris
- 95 Sociologie de la quantification
de Anaïs HENNEGUELLE et Arthur JATTEAU (2021)**
Pauline ADAM
Université libre de Bruxelles (ULB), Recherche et études
en politique internationale (REPI)
École normale supérieure (ENS), Centre Maurice
Halbwachs (CMH)
- 99 The New Politics of Numbers:
Utopia, Evidence and Democracy
de Andrea MENNICKEN et Robert SALAIS (eds.)
(2022)**
Harro MAAS
Centre Walras-Pareto, UNIL

Statistique et société

Magazine quadrimestriel publié par la Société Française de Statistique. Le but de Statistique et société est de présenter, d'une manière attrayante et qui invite à la réflexion, l'utilisation pratique de la statistique dans tous les domaines de la vie. Il s'agit de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

Directrice de publication

Anne Philippe, Présidente de la SFdS

Rédaction

Rédacteur en chef : Emmanuel Didier, CNRS, France

Rédacteurs en chef adjoints :

Thomas Amossé, Conservatoire national des arts et métiers, France

Jean Chiche, Institut d'études politiques de Paris, France

Quentin Dufour, Ecole Normale Supérieure, France

Jean-Jacques Droesbeke, Université libre de Bruxelles, Belgique

Chloé Friguet, Université Bretagne-Sud, France

Pauline Hervois, Sorbonne Université & Muséum national d'histoire naturelle, France

Olivier Martin, Université Paris Cité, France

Antoine Rolland, Université Lyon 2, France

Jean-Christophe Thalabard, Université de Paris, France

Catherine Vermandele, Université libre de Bruxelles, Belgique

Comité éditorial

Représentants des groupes spécialisés de la SFdS :

AGRO : Nicolas Pineau (Nestlé)

Banque Finance Assurance : Idriss Tchabda-Djamen (BNP Paribas)

Biopharmacie et Santé : Emmanuel Pham (IPSEN)

Enquêtes : Alina Gabriela Matei (IRD Université de Neuchâtel)

Enseignement : Catherine Vermandele (Université Libre de Bruxelles)

Environnement : Thomas Opitz (INRAE)

Fiabilité-Incertitudes : Vlad Stefan Barbu (Univ. Rouen)

Histoire de la Statistique : Jean-Jacques Droesbeke (Université Libre de Bruxelles)

Jeunes Statisticiens : Vivien Goepp (CBIO, Mines ParisTech)

MALIA : Christine Keribin (Université Paris-Sud)

Stat&Sport : Christian Derquenne (EDF)

Statistique et Enjeux Publics : Chantal Cases (INSEE)

Autres membres :

Jose Maria Arribas Macho, revue Empiria (Espagne)

Assaël Adary (Occurrence)

Denise Britz do Nascimento Silva (IASS - International Association of Survey Statisticians)

Gwenaëlle Brihault (INSEE)

Yves Coppieters't Wallant (Ecole de santé publique ULB)

Christophe Ley (Société Luxembourgeoise de Statistique, Gent Universiteit)

Theodore M. Porter (UCLA)

Walter J. Radermacher (La Sapienza Università, Rome)

Design graphique

fastboil.net

ISSN 2269-0271

Éditorial



Emmanuel DIDIER

Rédacteur en chef de *Statistique et société*

Chère lectrice, cher lecteur,

Cette livraison de *Statistique et société* vous propose un deuxième dossier portant sur « les chiffres au cœur de la crise du coronavirus ». L'appel à communication que vous avons fait circuler a suscité beaucoup de réponses qu'il eut été regrettable de ne pas publier. Cela montre à quel point les nombres étaient, comme nous l'écrivions déjà dans le précédent numéro, au cœur de ce phénomène. Vous trouverez d'abord un article de Samuel Goëta qui montre que la crise a été l'occasion de faire avancer les propositions en faveur de l'ouverture des données. Viennent ensuite trois articles qui apportent un point de vue sur cette problématique depuis trois autres pays, la crise et la nécessité de quantification ayant été mondiales : Geert Molenberghs nous décrit la situation de la Belgique, et montre en particulier le rôle qu'a joué le contrôle de l'aviation internationale dans ce pays ; Eugenia Motta et ses co-auteurs s'intéressent au Brésil et insistent, entre autres, sur le fait que la présentation esthétique des chiffres dans la presse a eu une importance réelle ; enfin, Jean-Guy Prévost s'intéresse au Canada et exhibe le rôle crucial du fédéralisme dans la compréhension des chiffres dans ce pays.

Ce numéro contient aussi un entretien biographique d'Yves Escoufier par Gilbert Saporta – qui après avoir été lui-même interrogé en 2020 par notre revue, se retrouve donc ici dans la position du questionneur. Nous proposons ensuite deux recensions. L'une est rédigée par Pauline Adam à propos de l'ouvrage écrit par Anaïs Henneguelle et Arthur Jatteau intitulé *Sociologie de la quantification* ; l'autre est due à Harro Maas, qui vous parle du livre édité par Andrea Mennicken et Robert Salais intitulé *The New Politics of Numbers: Utopia, Evidence and Democracy*.

Bonne lecture !

Emmanuel Didier

La pandémie de Covid-19 : un moment charnière pour l'ouverture des données en France



Samuel GOËTA¹

Co-fondateur de Dataactivist ; Maître de conférences associé en sociologie, Sciences Po Aix (Mesopolhis UMR 704)

TITLE

The Covid-19 pandemic: a turning point for open data in France

RÉSUMÉ

Cet article revient sur l'ouverture des données liées à la pandémie de Covid-19 en France pour montrer en quoi elle peut être considérée comme un moment charnière pour le mouvement récent de l'*open data*. S'appuyant principalement sur des sources ouvertes et des observations personnelles, il revient d'abord sur l'émergence des premiers outils internationaux de suivi de la pandémie de Covid-19, pour ensuite étudier la mise en place des outils de surveillance de la pandémie en France. Il montre aussi comment les premières ouvertures ont été accompagnées par des initiatives citoyennes de collecte et de visualisation des données. Ces initiatives citoyennes ont abouti à l'émergence de médiateurs de données qui ont joué un rôle important dans la diffusion et la compréhension des données sur la pandémie.

Mots-clés : Covid-19, pandémie, médiation, open data, covid tracker, tableaux de bord.

ABSTRACT

This article looks back at the release of data related to the Covid-19 pandemic in France to explain how it can be considered as a turning point for the recent open data movement. Relying mainly on open sources through a documentary research and personal observations, it first looks back at the emergence of the first international tools for monitoring the Covid-19 pandemic, and then studies the implementation of tools for monitoring the pandemic in France. It also shows how the first openings were accompanied by citizen initiatives of data collection and visualization that led to the emergence of data mediators who played an essential role in the dissemination and understanding of data on the pandemic.

Keywords: Covid-19, pandemic, mediation, open data, covid tracker, dashboard.

1. samuel@dataactivist.coop

« *On peut débattre de tout, sauf des chiffres.* » C'est ainsi que se concluait une des campagnes de publicité du gouvernement en faveur de la vaccination mettant en avant que 8 personnes sur 10 hospitalisées à cause de la Covid-19 n'étaient pas vaccinées². Ce message ne fait pas qu'user des chiffres comme argument d'autorité, il témoigne de la place centrale qu'ont occupée les données dans les débats relatifs à la pandémie de Covid-19. Selon l'évolution de ces chiffres, nous pouvions envisager de revoir des proches, de nous rendre sur notre lieu de travail, de partir en voyage, d'aller dîner au restaurant ou de sortir de chez soi passée une certaine heure. Cette situation inédite a été résumée par Emmanuel Didier³ dans AOC le 16 avril 2020 : « *Toutes les mesures politiques, tous nos discours, toutes nos réflexions sont orientés et conformés à un degré rarement atteint, par des outils quantitatifs.* » Ce constat s'est étendu bien au-delà de la première vague et de son confinement drastique. Très vite après l'arrivée des premiers cas en France, l'ouverture des données permettant de suivre l'évolution de l'épidémie a constitué une demande essentielle pour une multitude d'acteurs.

Nous nous intéressons ici aux indicateurs locaux et nationaux qui ont servi à suivre l'évolution de la pandémie et qui ont été scrutés quotidiennement par une grande part de la population (contaminations, entrées à l'hôpital, passages en réanimation, décès, tests, injections de vaccins, état des stocks, part des variants, en passant par les taux d'incidence et de positivité, le niveau du R0...) De vastes systèmes d'information ont permis de collecter toutes ces données sur les individus et la circulation du virus. Conformément à la législation, les données mises à disposition du public ont été expurgées de données à caractère personnel, agrégées ou anonymisées pour éviter toute réidentification d'individus. Depuis 2016, la loi pour une République numérique prévoit que les bases de données et les données qui présentent un intérêt économique, social, sanitaire ou environnemental doivent être diffusées en ligne, mises à jour et « dans un standard ouvert, aisément réutilisable et exploitable par un système de traitement automatisé. » Cette exigence de lisibilité des données par les machines a été essentielle pour de nombreux acteurs qui ne souhaitaient pas seulement accéder aux chiffres. Obtenir des données exploitables par les machines avec un fort niveau de détail a permis de refaire certains calculs, de proposer de nouvelles analyses, de nouvelles formes de visualisations, de renforcer ou de contester une décision.

Cet article est ainsi centré sur l'ouverture des données liées à la pandémie de Covid-19 en France et vise à documenter ce moment charnière pour le mouvement récent de l'*open data*. Pour décrire les frictions dans l'ouverture des données, je m'intéresserai aussi aux infrastructures qui ont été mises en place pour faire remonter les données collectées localement (Ronai, 2021). L'article s'appuie principalement sur des sources ouvertes à travers une recherche documentaire d'articles de presse et de documents officiels. Il s'appuie aussi sur des observations issues de mes activités associatives d'administrateur bénévole du portail d'accès aux documents administratifs « madada.fr » et professionnelles ayant participé à l'organisation et l'animation du Hackathon Covid, un événement rassemblant producteurs et réutilisateurs de données en avril 2021.

L'article décrit d'abord les premiers outils internationaux de suivi de la pandémie de Covid-19 qui ont posé une exigence internationale d'ouverture et d'actualisation des données de veille sanitaire. Il nous conduit ensuite à étudier la mise en place des outils de surveillance de la pandémie en France et à comprendre comment les premières ouvertures ont été accompagnées par des initiatives citoyennes de collecte et de visualisation des données. Nous verrons ensuite que l'utilisation massive des données sur la pandémie a conduit le gouvernement à relancer la politique d'*open data* en France qui semblait s'essouffler. Les médiateurs de données ont

2. Gouvernement (2021), « #COVID19 | Aujourd'hui en France, 8 personnes sur 10 hospitalisées à cause de la Covid-19 ne sont pas vaccinées. On peut débattre de tout, sauf des chiffres », *Twitter*. Adresse : <https://twitter.com/gouvernementfr/status/1429145022427893760>

3. Didier E. (2020), « Politique du nombre de morts », *AOC media - Analyse Opinion Critique*. Adresse : <https://aoc.media/opinion/2020/04/15/politique-du-nombre-de-morts/>

joué un rôle essentiel dans l'ouverture de nouvelles données sur la pandémie, en particulier CovidTracker, un site web lancé par le *data scientist* Guillaume Rozier, sur lequel je me concentrerai ici. L'article se conclut en évoquant le Hackathon Covid, un événement qui s'est tenu en avril 2021 à l'initiative de la Direction interministérielle de la transformation publique (DITP) afin de répondre à la pandémie en renforçant les liens entre les acteurs de la société civile et les administrations.

1. Une pandémie instantanément mise en données

Il n'est pas excessif d'affirmer que la pandémie a débuté par une ouverture de données. Alors que le 31 décembre 2019 les autorités de la ville de Wuhan en Chine annonçaient publiquement les signes d'une épidémie de pneumonie, le département des zoonoses du Centre chinois de contrôle et de prévention des maladies soumettait le 5 janvier 2020 le génome du virus identifié dans le marché de fruits de mer de Wuhan⁴. Le 11 janvier, ces données étaient publiées sur Genbank⁵, une des plus anciennes et plus grandes bases de données en génétique au monde (Strasser, 2011). La publication du génome a donné le départ de la course internationale aux vaccins, plusieurs équipes déjà préparées à l'éventualité de l'émergence d'un nouveau virus se constituant dès le 11 janvier. Dans un documentaire⁶ de la BBC, la professeure Teresa Lambe de l'université d'Oxford a raconté qu'elle a reçu une notification de la publication du génome dans la nuit du 10 au 11 janvier et s'est lancée immédiatement dans la conception d'un nouveau vaccin : « *Au petit matin, toujours en pyjama dans ma chambre, je me suis collée devant mon ordinateur pour développer un vaccin.* » Le 16 janvier, sans avoir eu accès au virus, une équipe de l'hôpital universitaire de la Charité de Berlin menée par le professeur Christian Drosten publiait le protocole du premier test RT-PCR⁷.

Les données de veille épidémiologique ont circulé plus difficilement du fait de la difficulté à détecter les cas, mais aussi de l'opacité qui a régné en Chine sur l'émergence de la pandémie. Une initiative a toutefois donné une grande visibilité aux chiffres sur la circulation de l'épidémie. Le 21 janvier, Ensheng Dong, doctorant à l'école d'ingénieur de l'université Johns Hopkins, sur les conseils de Lauren Gardner, sa directrice de thèse, a développé un tableau de bord de suivi de la pandémie⁸. Composé d'une carte du monde avec des points rouges regroupant les cas, le tableau de bord a attiré en quelques semaines plusieurs millions de visiteurs.

-
4. « Timeline of the COVID-19 pandemic in January 2020 », 2021, *Wikipedia*. Adresse : https://en.wikipedia.org/w/index.php?title=Timeline_of_the_COVID-19_pandemic_in_January_2020
 5. « Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome », 2020. Adresse : <http://www.ncbi.nlm.nih.gov/nuccore/MN908947.1>
 6. Gale C. (2021), « Covid-19, la course aux vaccins », *ARTE*. Adresse : <https://www.arte.tv/fr/videos/104055-000-A/covid-19-la-course-aux-vaccins/>
 7. Corman V. M. et al. (2020), « Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR », *Eurosurveillance*, vol. 25, n° 3. Adresse : <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.3.2000045>
 8. « How the Johns Hopkins coronavirus tracker was built and what it tells us », *The Washington Post*. Adresse : https://www.washingtonpost.com/local/johns-hopkins-tracker/2020/06/29/daea7eea-a03f-11ea-9590-1858a893bd59_story.html
« Meet The Team Behind The Coronavirus Tracker Watched By Millions », 2020, *NPR*. Adresse : <https://text.npr.org/833073670>

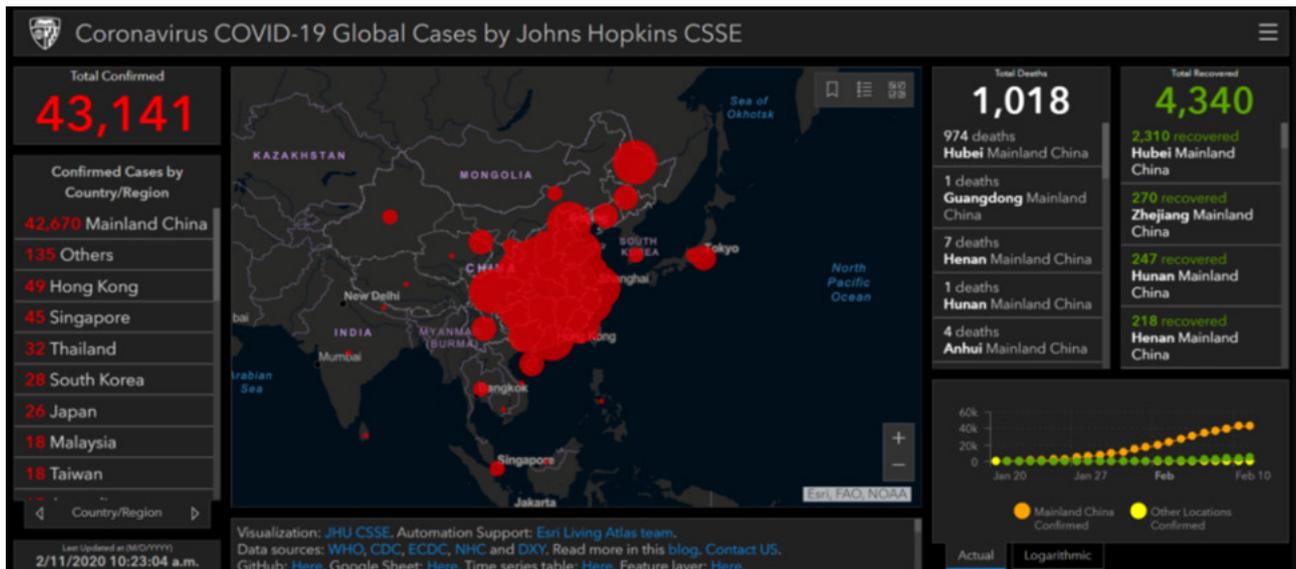


Figure 1 – Tableau de bord de l'université Johns Hopkins (février 2020)

Dans un premier temps, jusqu'au 31 janvier, toutes les données étaient collectées manuellement, puis le processus a été semi-automatisé (Dong, 2020). Pour la Chine, le tableau de bord mettait à jour toutes les 15 minutes DXY, une plateforme alimentée par des professionnels de santé qui agrégeait les chiffres transmis par les autorités à la presse locale dans chacune des provinces. Dans les autres pays, les données étaient actualisées principalement de manière manuelle à partir des sites des autorités de santé locales et nationales. Ces données sont devenues la source internationale de référence, au moins pendant la première vague de l'épidémie, pour connaître l'évolution des contaminations partout dans le monde. Le tableau de bord de l'université Johns Hopkins est resté une référence au-delà de la première vague même si d'autres acteurs internationaux ont pris le pas, comme Our World in Data qui agrège aujourd'hui les données sur les contaminations, les tests, les hospitalisations, les décès ou les vaccinations dans une centaine de pays dans le monde.

Pourquoi revenir sur cet outil dans un article centré sur la France ? Le tableau de bord de l'université Johns Hopkins a posé une exigence d'ouverture et de mise à jour quotidienne des données, devenues la norme internationale en matière de veille sanitaire sur la Covid-19. Cet outil a aussi montré l'importance de la médiation, au-delà des données brutes, pour aider le public à suivre l'évolution de l'épidémie.

2. Un dispositif de surveillance en France mis en place en urgence

En France, un dispositif de surveillance de la propagation de la Covid-19 a été mis en place à partir du 10 janvier 2020⁹. Conçu à la suite de la pandémie de grippe A (H1N1) en 2009, le plan national de prévention et de lutte « pandémie grippale » de 2011 prévoyait des outils de pilotage adaptés aux quatre stades de l'épidémie¹⁰. Le premier stade vise à freiner l'introduction du virus sur le territoire. Le 24 janvier 2020, les trois premiers cas de Covid-19 sont déclarés en France¹¹. Le deuxième stade qui vise à freiner la propagation du virus est déclenché le 29 février avec une centaine de cas déclarés sur le territoire. Les autorités ont tenté de retracer

9. Santé Publique France (2020), Livret de présentation « Le suivi de l'épidémie de COVID-19 », Adresse : https://solidarites-sante.gouv.fr/IMG/pdf/livret_suivi_epidemie_covid-19_-_270320.pdf

10. Secrétariat général de la défense et de la sécurité nationale (2011), Plan national de Prévention et de lutte « Pandémie grippale ». Adresse : https://solidarites-sante.gouv.fr/IMG/pdf/Plan_Pandemie_Grippale_2011.pdf

11. « Chronologie de la pandémie de Covid-19 en France », 2021, Wikipédia. Adresse : https://fr.wikipedia.org/w/index.php?title=Chronologie_de_la_pand%C3%A9mie_de_Covid-19_en_France&oldid=184735959

et d'isoler l'ensemble des cas contacts, mais très vite, l'épidémie est devenue hors de contrôle. Au troisième stade qui est déclaré par le gouvernement le 14 mars 2020, le plan national de prévention prévoit l'arrêt de la surveillance individuelle des cas avec l'augmentation rapide du nombre de cas. Le ministère des Solidarités et de la Santé et Santé Publique France étaient alors passés dans une stratégie de surveillance dite « syndromique » qui consiste à identifier la part de la population qui présente des symptômes de l'infection à la Covid-19 même si ces signes se retrouvent dans d'autres maladies¹². Pour assurer cette surveillance, Santé Publique France a comptabilisé les passages aux urgences hospitalières pour suspicion de Covid-19 en s'appuyant sur un réseau de 194 services de réanimation et le dispositif SI-VIC (Système d'information pour le suivi des victimes) conçu en 2016 suite aux attentats de Paris pour des événements et des catastrophes impliquant un grand nombre de victimes. Concrètement, ces systèmes d'information ont permis de collecter des indicateurs essentiels pour suivre l'engorgement des hôpitaux comme le nombre de patients hospitalisés pour Covid-19, placés en réanimation, guéris ou décédés. La médecine dite « de ville » a aussi fait l'objet d'une surveillance avec une remontée du nombre quotidien de patients présentant les signes d'une infection respiratoire aiguë. Ces données sont collectées parmi les praticiens membres du réseau Sentinelles ou d'une des associations SOS Médecins puis extrapolées à l'ensemble des généralistes de la même zone¹³. La prévalence des pathologies saisonnières comme la grippe est estimée avec une méthodologie similaire.

Avec l'accélération de la propagation du virus, les autorités ont continué à communiquer quotidiennement le nombre de cas confirmés malgré des avertissements sur son site : « *Du fait de la difficulté de l'identification et de la confirmation biologique de l'ensemble des cas de Covid-19, les données présentées sous-estiment le nombre réel de cas*¹⁴. » La France n'a pas procédé à un dépistage de grande envergure de la Covid-19, les tests ayant généralement été réservés à des publics particuliers : malades à l'hôpital, personnel soignant, personnes à risques... Malgré ces limites évidentes, l'indicateur du nombre de cas confirmés a pourtant été très scruté, à la fois en valeur absolue et en comparaison avec d'autres pays, y compris certains ayant adopté des politiques de tests massifs de leur population. Le nombre de tests pratiqués en France était lui-même à prendre avec précaution. Les tests remontaient des laboratoires hospitaliers et d'un système de surveillance appelé « 3labos » reposant sur trois laboratoires centralisateurs de prélèvements : Eurofins, Biomnis et Cerba. Ces réseaux fédèrent 1300 laboratoires représentant environ 30 % du total des 4 000 sites de prélèvements en France. Un article de Libération expliquait cette situation par l'urgence qui a conduit les autorités à s'appuyer sur les systèmes d'information déjà en place pour la surveillance épidémique avec ces laboratoires sous-traitants traditionnels des hôpitaux français¹⁵. Ces trous dans l'appareil statistique se sont retrouvés aussi sur le nombre de décès qui était probablement l'indicateur le plus scruté pour évaluer l'impact de l'épidémie. On le sait aujourd'hui, environ 44 % des décès liés à la Covid-19 ont eu lieu dans des établissements d'hébergement pour personnes âgées dépendantes (EHPAD)¹⁶ et d'autres établissements médicaux sociaux. À partir du mois de mars, plusieurs clusters sont apparus dans ces établissements, mais leur signalement ne faisait pas l'objet d'une collecte systématique, faute d'outil adapté. Le 28 mars, le ministère des Solidarités et de la Santé a déployé une application pour permettre à ces établissements de signaler les

12. Santé Publique France (2020), *op.cit.*

13. Réseau Sentinelles > France > Surveillance continue.
Adresse : <https://www.sentiweb.fr/france/fr/?page=methodes>

14. « Cas confirmés, hospitalisés, guéris... Comment les malades du coronavirus sont comptabilisés en France », 2020, *FranceInfo*. Adresse : https://www.francetvinfo.fr/sante/maladie/coronavirus/cas-confirmes-hospitalises-gueris-comment-les-malades-du-coronavirus-sont-comptabilises-en-france_3884155.html

15. Mathiot C., « Covid-19 : est-il vrai que la France est incapable de compter le vrai nombre de tests pratiqués chaque jour ? », *Libération*. Adresse : https://www.liberation.fr/checknews/2020/04/15/covid-19-est-il-vrai-que-la-france-est-incapable-de-compter-le-vrai-nombre-de-tests-pratiques-chaque_1785211/

16. « Les résidents d'Ehpad représentent 44 % des morts du Covid-19 », 2020, *Le Monde.fr*. Adresse : https://www.lemonde.fr/les-decodeurs/article/2020/12/03/les-residents-d-ehpad-representent-44-des-morts-du-covid-19_6062084_4355770.html

cas possibles (ceux présentant le tableau de la maladie) ou confirmés par un test PCR parmi les pensionnaires ou le personnel ainsi que les décès dans la structure ou à l'hôpital de ces patients¹⁷. Les établissements devaient réaliser un signalement dès le premier cas détecté afin de suivre les cas et les décès et de mettre en place des mesures de gestion des cas groupés. Ces indicateurs ont été intégrés dès le 2 avril dans les points épidémiologiques de Santé Publique France¹⁸.

3. Des premières ouvertures de données incitées par des initiatives citoyennes

Dès l'apparition des premiers cas, Santé Publique France et le ministère de la Santé ont communiqué ces indicateurs aux médias sans pour autant ouvrir les données. Ces indicateurs étaient principalement communiqués sous la forme de communiqués de presse au format PDF ou à l'oral lors de points presse : impossible de télécharger les données pour réaliser des analyses ou proposer de nouvelles visualisations. Au moment du confinement, Santé Publique France n'avait pas de compte sur data.gouv.fr (Chignard, 2021). Son outil d'observation cartographique Geodes n'a inclus certaines données sur la Covid-19 qu'à partir du 18 mars 2020¹⁹. Ce manque de données publiques actualisées et dans un format exploitable était déploré à de nombreuses reprises sur les réseaux sociaux à la fois par des chercheurs, mais aussi des acteurs variés, sans expertise sanitaire spécifique, qui ont voulu proposer leur aide face à la crise. Philippe Lemoine, chercheur en philosophie à l'université Cornell, se faisait leur porte-parole le 16 mars 2020 sur Twitter²⁰ déplorant la non-accessibilité des données et la comparait avec des pays comme la Corée du Sud ou l'Italie qui ont ouvert les données détaillées utilisées pour la gestion de la pandémie : « *Pour que les experts puissent donner des conseils de bonne qualité aux décideurs politiques, il faut que des données aussi complètes que possible soient publiées régulièrement et, à cet égard, la situation est absolument scandaleuse. Le site de Santé Publique France ne publie qu'un malheureux communiqué sur le nombre de cas et de morts en France [...] Les agences régionales de santé publient aussi des chiffres sur ce qui se passe sur leur territoire, mais ces derniers sont aussi très pauvres et les informations publiées varient beaucoup selon les régions.* »

Pour y remédier, des collectifs citoyens ont produit et actualisé des données dès le début du mois de mars 2020. Le 6 mars, Lior Perez, ingénieur en science des données chez Météo France, lançait un appel sur LinkedIn pour répondre à cette situation²¹ : « *Amis data scientists, avez-vous quelques minutes pour m'aider à monter un projet open data ? L'objectif est de consolider les données Covid-19 françaises dans un fichier csv, comme l'ont fait des data scientists coréens dans leur pays. Nous ne savons pas nous servir d'un stéthoscope, mais nous pouvons peut-être aider ceux qui savent en leur apportant de la data facilement exploitable et de la dataviz. Les contributions attendues sont simples : lire quelques communiqués de presse et reporter les infos dans un csv. Avec un contributeur par région, ça ne prendrait que quelques minutes.* »

17. Santé Publique France (2020), « Protocole de la surveillance active des cas de COVID-19 en EHPAD et EMS ». Adresse : https://www.preventioninfection.fr/wp-content/uploads/2020/03/2020_mars-28_COVID-19_protocole-SV-SGLT_ems_SPF.pdf

18. Santé Publique France (2020), « Point épidémiologique hebdomadaire du 2 avril ».

19. Santé Publique France (2020), « Covid-19 : Point épidémiologique - Situation au 15 mars 2020 à minuit ». Adresse : <https://www.santepubliquefrance.fr/content/download/237536/2535424>

20. Philippe Lemoine sur Twitter, 2020. Adresse : <https://twitter.com/phl43/status/1239590836947517442>

21. Perez L. (2020), « Amis data scientists, avez-vous quelques minutes pour m'aider à monter un projet open data ? », *LinkedIn*. Adresse : https://www.linkedin.com/posts/liorperez_opendata-covid19-dataset-activity-6642013973241704448-55L2/

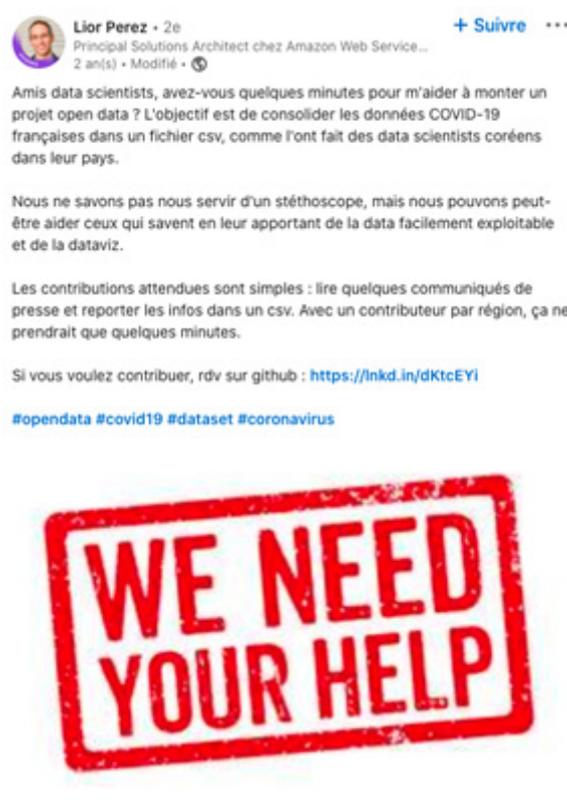


Figure 2 – Message de Lior Perez sur LinkedIn

En publiant ce message, il s'est inspiré d'un travail similaire réalisé en Corée du Sud et a créé un dépôt sur GitHub contenant un premier fichier pour six régions. Les données étaient collectées²² à partir des communiqués de presse des Agences régionales de santé (Ars). Le 14 mars 2020²³, ce projet a fusionné avec celui d'une autre équipe initiée à titre personnel par Jérôme Desboeufs et Bastien Guerry, deux agents d'Etalab, la mission interministérielle en charge de l'ouverture des données (Goeta, 2018). L'initiative Data Against Covid²⁴ se présentait comme « une organisation informelle issue de la société civile avec des développeurs, des data scientists, des ingénieurs, des chefs de projets et talents de tous corps de métiers qui s'organisent ensemble pour fournir des données consolidées sur l'épidémie de covid 19 et proposer des outils de visualisation » afin de « favoriser une information claire, précise et accessible pour tous sur la propagation du virus ». Une communauté en ligne intitulée opencovid19-fr s'est structurée autour de canaux de discussions sur Slack. 1 200 membres y ont échangé près de 90 000 messages (Chignard, 2021). Dans la continuité du projet de Lior Perez, une cinquantaine de personnes se sont affairées à collecter les données à partir des sites de Santé Publique France, des agences régionales de santé, le ministère des Solidarités et de la Santé et certaines préfectures. Un guide de la contribution a permis d'établir une méthodologie qui comprenait l'archivage systématique des sources et la production de métadonnées précises sur chacun des chiffres²⁵. Les contributeurs reportaient même les chiffres annoncés lors des points presse quotidiens du directeur général de la santé et du ministre. À partir du 13 mars 2020, cette communauté a produit un tableau de bord de suivi de l'évolution de l'épidémie en France intitulé veille-coronavirus.fr alimenté par

22. Perez L. (2020), « Update README.md · lperez31/coronavirus-france-dataset@5e87cb8 », *GitHub*. Adresse : https://github.com/lperez31/coronavirus-france-dataset/commit/5e87cb8abc46458639bd218254d3136e0ecc97d?short_path=b335630#diff-b335630551682c19a781afebcf4d07bf978fb1f8ac04c6bf87428ed5106870f5

23. « Ces citoyens luttent contre le coronavirus grâce à une arme : la data », *Les Echos Start*. Adresse : <https://start.lesechos.fr/societe/engagement-societal/ces-citoyens-luttent-contre-le-coronavirus-grace-a-une-arme-la-data-1191396>

24. « Data Against COVID-19 / FR ». Adresse : <https://opencovid19-fr.github.io/>

25. GitHub – opencovid19-fr/data : Consolidation des données de sources officielles concernant l'épidémie de COVID19. Adresse : <https://github.com/opencovid19-fr/data>

les données collectées par les contributeurs d'opencovid19-fr²⁶.

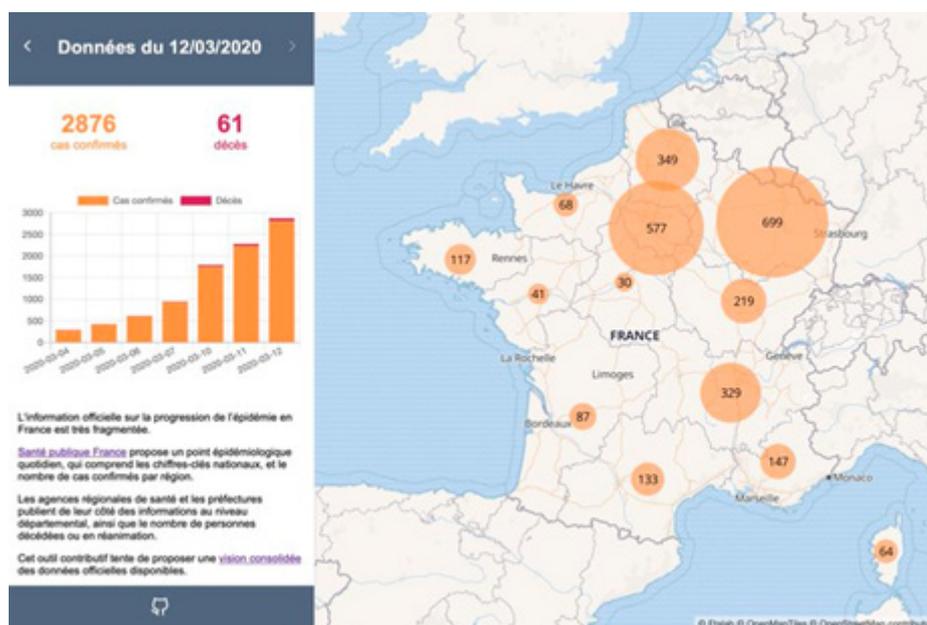


Figure 3 – Tableau de bord veille-coronavirus.fr

Les autorités sanitaires ont diffusé certaines données officielles en open data à partir du 19 mars 2020 quand Santé Publique France a publié sur data.gouv.fr les données sur la pression hospitalière (nombre de patients hospitalisés, en réanimation, retournés à domicile ou décédés), les urgences (nombre de passages pour suspicion de Covid-19) et la médecine de ville (actes médicaux au sein du réseau SOS Médecins pour suspicion de Covid-19)²⁷. Le 27 mars, Santé Publique France publiait en complément un jeu de données sur le nombre de tests de dépistage de la Covid-19 réalisés par département au sein du réseau « 3labos » évoqué précédemment. Ce système de surveillance a été remplacé en mai 2020 par un système d'information dédié au dépistage de la Covid-19, nommé SI-DEP. Opéré par l'Assistance publique hôpitaux de Paris (AP-HP), SI-DEP collecte à l'échelle nationale tous les résultats de tests de dépistage de la Covid-19²⁸. En complément de ces données, l'Insee a publié le 27 mars 2020 « à titre exceptionnel » des données quotidiennes sur l'évolution de la mortalité en France, toutes causes de décès réunies. Chaque vendredi, l'institut statistique publiait des données sur le nombre de décès constatés 11 jours plus tôt (7 jours pour les communes ayant dématérialisé la déclaration des décès). Des fichiers détaillés permettent de connaître la date et la commune du décès, l'âge et le sexe du défunt, sa commune de résidence et le lieu de décès (hôpital ou clinique, domicile, maison de retraite, autres)²⁹. Dans un billet accompagnant leur publication³⁰, les statisticiens de l'Insee ont averti que l'évaluation de la surmortalité due à la Covid-19 exige « beaucoup d'humilité ». La situation inédite du confinement a provoqué de nombreux effets qui perturbent l'interprétation des données : la transmission ralentie des déclarations par voie postale, la sous-mortalité pour d'autres causes que la Covid-19 (comme les accidents de la route) ou une surmortalité pour les accidents cardio-vasculaires ou domestiques. S'interrogeant sur la

26. Tableau de bord COVID-19, 2020.

Adresse : <https://web.archive.org/web/20200329222434/https://dashboard.covid19.data.gouv.fr>

27. Données relatives à l'épidémie du covid-19 – data.gouv.fr, 2020.

Adresse : <https://web.archive.org/web/20200323212238/https://www.data.gouv.fr/fr/datasets/donnees-relatives-a-lepidemie-du-covid-19/>

28. « Le Système d'Information de Dépistage Populationnel (SI-DEP) opéré par l'AP-HP évolue pour fournir des résultats de tests certifiés », APHP. Adresse : <https://www.aphp.fr/actualite/le-systeme-dinformation-de-depistage-populationnel-si-dep-opere-par-lap-hp-evolue-pour>

29. « Téléchargement des fichiers des décès quotidiens – Nombre de décès quotidiens », Insee. Adresse : <https://www.insee.fr/fr/statistiques/4487988?sommaire=4487854#documentation-sommaire>

30. « Mourir de la grippe ou du coronavirus : faire parler les chiffres de décès publiés par l'Insee... avec discernement », 2020, Le blog de l'Insee. Adresse : <https://blog.insee.fr/mourir-de-la-grippe-ou-du-coronavirus-faire-parler-les-chiffres-de-deces-publies-par-linsee-avec-discernement/>

très forte hausse des décès à domicile dans certaines communes, un journaliste de Médiapart a montré que le lieu du décès devait être pris avec grande précaution³¹ : beaucoup de mairies cochaient la case « à domicile » pour le lieu du décès dès lors que le défunt était domicilié dans un Ehpad.

Après ces premières ouvertures de données, fin mars 2020, le tableau de bord veille-coronavirus.fr a rapidement été repris au sein de l'État afin de devenir l'outil officiel du gouvernement pour le suivi de la pandémie. Cette internalisation a été facilitée par la présence de certains agents d'Etalab au sein du projet opencovid19-fr et d'une collaboration entre Etalab, le ministère des Solidarités et de la Santé, le Service d'information du gouvernement (SIG) et Santé Publique France³². Le 28 mars 2020, lors d'un point presse quotidien avec le Premier ministre, le directeur général de la santé, Jérôme Salomon, le 28 mars, annonçait l'ouverture des données liées à la pandémie et la publication du tableau de bord.

Son discours vantait les bénéfices de l'open data et soulignait l'importance de la collaboration avec les acteurs de la société civile, désignant vraisemblablement les bénévoles d'opencovid19-fr : « *Le libre accès aux données relatives à l'épidémie, ce qu'on appelle l'open data, assure la confiance des citoyens dans les éléments qui leur sont communiqués, favorise les actions de prévention contre la propagation du virus et facilite la prise de décision. La diffusion et l'exploitation de ces données ont été rendues possibles grâce à une collaboration étroite entre plusieurs administrations et la société civile.* » Ces données ont encore gagné en importance au moment du déconfinement. Le 28 avril 2020, le Premier ministre Edouard Philippe présentait la stratégie de déconfinement à l'Assemblée nationale, comprenant une différenciation selon la circulation du virus dans les territoires. La carte de vigilance distinguait pour chaque département un niveau rouge ou vert basé sur trois indicateurs : la proportion de passages aux urgences pour une suspicion de Covid-19, le taux d'occupation des lits de réanimation par des patients atteints de Covid-19 et le taux de couverture des besoins en tests estimés au 11 mai³³. À partir du 28 mai, ces critères ont été remplacés par quatre indicateurs : le taux d'incidence (tests positifs pour 100 000 habitants), le taux de positivité des tests, l'évolution du R0 (nombre de personnes contaminées par chaque malade) et le taux d'occupation des lits en réanimation. Ces indicateurs très scrutés ont été publiés le 30 avril sur data.gouv.fr et visualisés dans le tableau de bord du gouvernement. Fin juin 2021, des données au niveau infradépartemental (EPCI) ont été publiées par Santé Publique France pour permettre un suivi plus localisé de l'épidémie³⁴.

4. Des données réutilisées massivement, une relance de la politique d'ouverture des données

Les données publiées sur data.gouv.fr liées à la Covid-19 ont répondu à une très forte demande de la part d'utilisateurs très variés. De mars à octobre 2020, plus du tiers des visites (37 %) sur data.gouv.fr concernaient des jeux de données publiés par Santé Publique France, l'Insee (fichier des personnes décédées) ou opencovid19-fr³⁵. Si l'on consulte les statistiques disponibles en ligne sur stats.data.gouv.fr, la fréquentation du portail national data.gouv.fr a explosé en particulier pendant les périodes de confinement.

31. « Suivi de l'épidémie : le cafouillage de l'Insee sur les morts à domicile », *Mediapart*. Adresse : <https://www.mediapart.fr/journal/france/190520/suivi-de-l-epidemie-le-cafouillage-de-l-insee-sur-les-morts-domicile?onglet=full>

32. Etalab (2020), « Comment les administrations ont collaboré à l'ouverture des données du Coronavirus : le cas français », *Le blog d'Etalab*. Adresse : <https://www.etalab.gov.fr/comment-les-administrations-ont-collabore-a-louverture-des-donnees-du-coronavirus-le-cas-francais>

33. Données de la carte de vigilance COVID-19 – data.gouv.fr. Adresse : <https://www.data.gouv.fr/fr/datasets/donnees-de-la-carte-de-vigilance-covid-19/>

34. Samuel Goëta sur Twitter, 2020. Adresse : <https://twitter.com/samgoeta/status/1323341113965580289>

35. Ce chiffre a été calculé à partir de l'outil de statistiques de fréquentation, stats.data.gouv.fr, qui comptabilise le trafic sur l'ensemble des sites d'Etalab. Les données brutes ayant permis ce calcul sont publiées par Dataactivist sur data.gouv.fr : <https://www.data.gouv.fr/fr/datasets/nombre-de-visiteurs-par-jour-depuis-2014-des-url-de-data-gouv-fr-1/>

Data.gouv.fr : des pics de fréquentation liés aux confinements

Les périodes de confinement sont indiquées dans les zones grisées

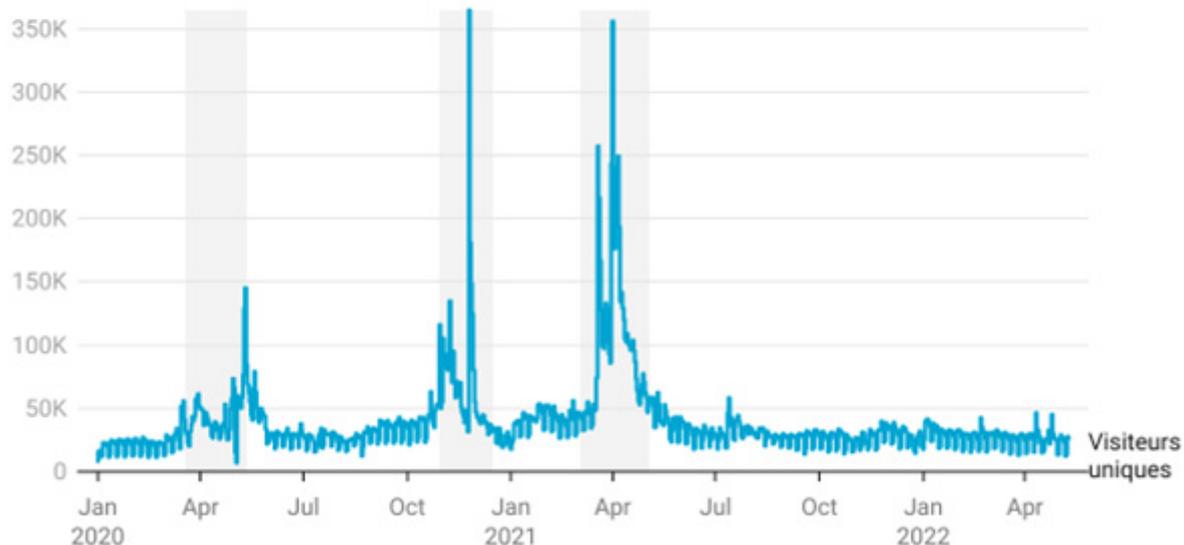


Chart: Samuel Goëta • Source: Stats.data.gouv.fr • Created with Datawrapper

Figure 4 – Évolution de la fréquentation de data.gouv.fr (Source des données : stats.data.gouv.fr)

En mai 2020, les données ont servi en particulier à calculer les indicateurs de la carte de vigilance qui déterminaient département par département les conditions de la reprise des activités. Le 24 novembre 2020, le président Emmanuel Macron déclarait lors d'un discours télévisé envisager le déconfinement d'ici au 15 décembre 2020 si la situation sanitaire continuait de s'améliorer avec des objectifs de 5000 nouveaux cas positifs quotidiens et moins de 3000 patients en réanimation. Les données liées à la propagation du virus ont alors été très suivies pour prévoir si le déconfinement allait se réaliser et si les réunions lors des fêtes de fin d'année étaient envisageables.

La forte croissance de la fréquentation de data.gouv.fr en 2020 et l'importance dans le débat public des visualisations issues de données ouvertes a contribué à relancer la politique d'open data du gouvernement. Avant la pandémie, l'ouverture des données semblait sur la sellette depuis la nomination en décembre 2018 de Nadi Bou-Hanna à la Direction interministérielle au numérique (DINUM), à laquelle Etalab est rattachée. S'inscrivant en rupture avec son prédécesseur Henri Verdier, Nadi Bou-Hanna avait placé l'ouverture des données en retrait face aux enjeux d'échange de données entre administrations. Cela s'est reflété en particulier dans la stratégie et la feuille de route 2019-2022 de la DINUM³⁶ intitulée « TECH.GOUV » qui, avant sa mise à jour en septembre 2021, ne contenait aucune mention de l'ouverture des données ni de data.gouv.fr. Le manque d'appui du directeur interministériel au numérique à l'open data pouvait être perçu par certains comme une remise en cause du gouvernement du principe d'open data par défaut prévu par la loi pour une République numérique de 2016.

36. TECH.GOUV : l'État lance un programme pour accélérer la transformation numérique du service public. Adresse : <https://www.numerique.gouv.fr/actualites/tech-gouv-accelerer-la-transformation-numerique-du-service-public/>

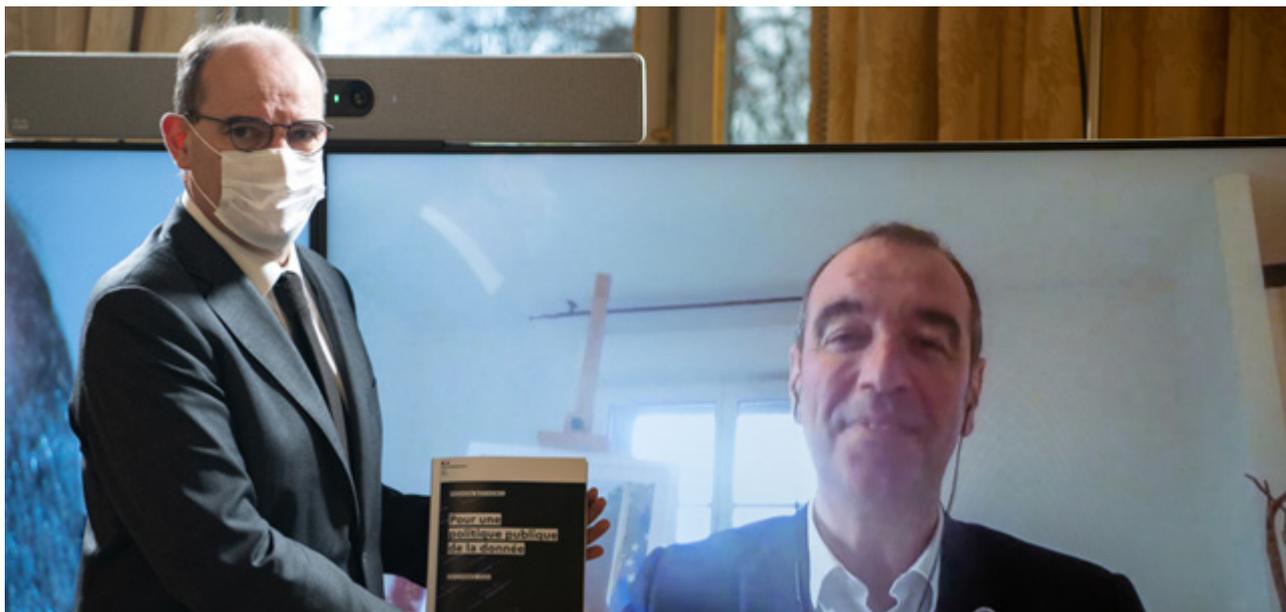


Figure 5 – Remise du rapport du député Eric Bothorel au Premier ministre Jean Castex
(Photo : Service d'information du Gouvernement)

Le rapport du député Éric Bothorel remis au Premier ministre le 23 décembre 2020, en tentant de tirer les leçons de la pandémie, a donné lieu à une relance vigoureuse de l'ouverture des données au sein de l'État. Le 22 juin 2020, le Premier ministre Édouard Philippe lui a confié une lettre de mission qui débute par une référence à la pandémie : « *La gestion de l'épidémie de Covid-19 l'a démontré [...] la prise de décision dépend plus que jamais de la maîtrise des données, publiques comme privées, donc de leur circulation et de leur partage.* » Le député de la 5^e circonscription des Côtes-d'Armor a été sélectionné pour guider cette mission du fait de ses nombreuses expériences professionnelles dans l'informatique. Renaud Vedel, préfet coordonnateur national pour l'intelligence artificielle, et Stéphanie Combes, directrice du Health Data Hub, ont aussi composé la mission à laquelle ont contribué plusieurs agents de l'administration. S'interrogeant sur « *notre capacité collective à toujours mieux organiser l'ouverture, le partage et l'exploitation des données et des codes sources* », le Premier ministre a chargé le député d'analyser les opportunités et les freins à lever pour favoriser l'ouverture par défaut des données. Le député a aussi eu pour mission d'accélérer le partage des données dites d'intérêt général collectées par des acteurs privés et l'ouverture des codes sources des applications et logiciels développés par les administrations. Au terme de six mois d'auditions et de concertation, le député a rendu son rapport le 23 décembre 2020³⁷. Le rapport a abouti à une circulaire le 28 avril 2021 par le Premier ministre imposant notamment la nomination d'un administrateur des données, des algorithmes et des codes sources (AMDAC) dans chaque ministère et la publication par chaque ministère d'une feuille de route sur l'ouverture des données et des codes sources le 28 septembre 2021³⁸. Les quinze feuilles de route publiées ont totalisé plus de 500 actions sur les enjeux des AGDAC. Certains acteurs de la société civile ont jugé pusillanimes les actions sur l'ouverture des données³⁹. Quoi qu'il en soit, la mobilisation des agents dans les ministères pour l'élaboration de ces feuilles de route a replacé le sujet de l'open data comme un sujet d'importance pour l'administration.

37. Remise du rapport sur la politique publique de la donnée, des algorithmes et des codes sources, *Gouvernement.fr*. Adresse : <https://www.gouvernement.fr/remise-du-rapport-sur-la-politique-publique-de-la-donnee-des-algorithmes-et-des-codes-sources>

38. « Données, algorithmes et codes sources : une mobilisation générale sans précédent, à travers 15 feuilles de route ministérielles », 2021, *numerique.gouv.fr*. Adresse : <https://www.numerique.gouv.fr/actualites/donnees-algorithmes-codes-sources-mobilisation-generale-sans-precedent-15-feuilles-de-route-ministerielles/>

39. « Data, code, algos : 15 feuilles de route et plus de 500 actions », 2021, #TeamOpenData. Adresse : <https://teamopendata.org/t/data-code-algos-15-feuilles-de-route-et-plus-de-500-actions/31297>

5. L'émergence des médiateurs et le rôle de CovidTracker dans l'ouverture des données de l'administration

Le rapport du député Bothorel a aussi contribué à mettre en lumière le rôle des initiatives de la société civile pour « accélérer » l'innovation de la sphère publique et renforcer la transparence des décisions. Deux initiatives font l'objet d'une étude de cas dans le rapport : OpenCovid19-fr, évoqué précédemment, et CovidTracker. Cette dernière initiative mérite qu'on s'y attarde tant sa visibilité auprès du grand public a été importante et son rôle central dans l'ouverture des données liées à la pandémie à partir de la fin de l'année 2020. CovidTracker est une initiative lancée par Guillaume Rozier, un ingénieur en sciences des données de 24 ans qui, au moment de la première vague, terminait ses études à Telecom Nancy⁴⁰. En mars 2020, stagiaire dans une banque luxembourgeoise, ne pouvant pas télétravailler ni rester au Luxembourg, il a décidé d'occuper son temps libre en analysant les données de la pandémie fournies par l'université Johns Hopkins. Début mars, il publiait sur Twitter un graphique pour montrer un décalage de neuf jours⁴¹ entre la situation en France et celle en Italie où la croissance du nombre de cas était exponentielle.

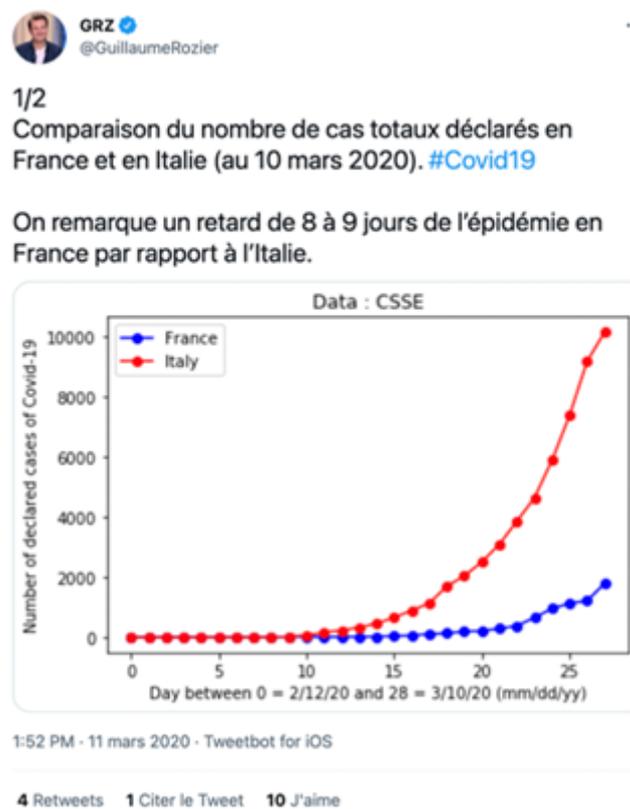


Figure 6 – Tweet de Guillaume Rozier le 11 mars 2020 comparant l'évolution du nombre de cas en France et en Italie

Au fur et à mesure de l'avancée de la première vague, ses tweets ont connu un grand succès aboutissant à la publication de son site intitulé CovidTracker en avril 2020. Le 13 avril 2020, 15 000 personnes l'avaient consulté⁴². De nombreux médiateurs, comme lui issus de formation

40. Bembaron E. et R. Enguérand (2020), « Comment est né le site très suivi CovidTracker », *Le Figaro*.

41. Rozier G. (2020), « Comparaison du nombre de cas totaux déclarés en France et en Italie (au 10 mars 2020). #Covid19 On remarque un retard de 8 à 9 jours de l'épidémie en France par rapport à l'Italie. », *Twitter*. Adresse : <https://twitter.com/GuillaumeRozier/status/1237723117419716608>

42. « Épidémie covid-19 : un étudiant de Télécom Nancy réalise un outil de visualisation de données », 2020. Adresse : <http://telecomnancy.univ-lorraine.fr/fr/big-data-0>

en sciences des données, ont contribué à visualiser et à expliquer les données de la pandémie. En novembre 2020, *Le Parisien* publiait un article sur « ces jeunes geeks passés maîtres en suivi de l'épidémie » dressant le portrait des « 3G » : Germain Forestier, Guillaume Saint-Quentin et Guillaume Rozier, trois spécialistes des données analysant quotidiennement sur Twitter les données de la pandémie⁴³. Selon une personne de la rédaction de *Quotidien – l'émission de grande écoute de TMC* animée par Yann Barthès – que j'ai interrogée, cet article a convaincu les journalistes d'inviter Guillaume Rozier sur le plateau pour présenter CovidTracker au million et demi de téléspectateurs de l'émission. L'ingénieur y a livré un plaidoyer en faveur de l'ouverture des données présentées comme une « mine d'or sous-exploitée » et un moyen d'améliorer la qualité de nos vies⁴⁴.

En plus de fournir des outils de visualisation des données, CovidTracker a joué un rôle essentiel dans l'ouverture de nouvelles données liées à la pandémie. En novembre⁴⁵ 2020, Elias Orphelin, un des premiers contributeurs de CovidTracker, relayait avec Guillaume Rozier sur Twitter une demande d'ouverture des données de détection de la concentration du SARS Cov-2 dans les eaux usées. Étudiant en dernière année d'école de commerce à l'ESCP Paris et analyste de données dans le cadre de son apprentissage, Elias Orphelin souhaitait intégrer ces données dans CovidTracker pour compléter les données de tests issues de SI-DEP qui sont très sensibles à l'évolution de la politique de tests et aux pratiques des laboratoires. Par exemple, les jours fériés réduisent la capacité de tests et créent artificiellement dans les données des baisses du taux de contamination, encore plus quand ils s'enchaînent en novembre ou en avril. Accompagné par l'équipe de madada dans la formulation et le ciblage de sa demande, Elias Orphelin a transmis sa demande au ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI) qui a consacré 3 millions d'euros au projet d'Observatoire épidémiologique dans les eaux usées (OBEPINE). Lancé en juillet, OBEPINE vise à réaliser un suivi en temps réel des traces de Covid-19 dans les eaux usées sur tout le territoire national afin d'anticiper les évolutions de la pandémie. Cette demande s'appuyait sur l'obligation d'accès ouvert pour les données issues de recherches financées sur projets prévue par le Plan national pour la science ouverte annoncé par la ministre Frédérique Vidal, le 4 juillet 2018. La demande formulée par Elias Orphelin le 20 novembre 2020 a reçu une réponse favorable le 9 janvier 2021 de la part d'Isabelle Blanc, l'administratrice ministérielle des données du MESRI⁴⁶, le ministère ayant aussi largement communiqué sur Twitter cette réponse. Les bénévoles de CovidTracker ont découvert les modalités d'exercice du droit d'accès et aussi ses lenteurs. La réponse favorable n'a conduit à l'ouverture des données que six mois après la demande initiale alors que le Code des relations entre le public et l'administration (CRPA) prévoit un délai d'un mois pour le traitement des demandes d'accès aux documents administratifs.

Les bénévoles de CovidTracker ont aussi joué un rôle essentiel dans l'ouverture des données liées aux variants. À la fin du mois de décembre, avec l'émergence du variant dit « anglais » (renommé au printemps « alpha » par l'OMS), les craintes étaient fortes d'une diffusion de cette souche plus contagieuse sur le territoire français, ce qui est confirmé le 25 décembre 2020. Pour suivre cette propagation, le 7 février 2021, la Direction générale de la santé (DGS) a exigé que tout test positif fasse l'objet d'un criblage pour déterminer s'il s'agit d'une contamination

43. Berrod N. (2020), « Covid-19 : ces jeunes "geeks" passés maîtres en suivi de l'épidémie », *Le Parisien*.

44. Quotidien sur Twitter : « "L'analyse de données peut sauver des vies !" @GuillaumeRozier, le fondateur de CovidTracker, dans #Quotidien » <https://t.co/iVjf3xbInj> » / Twitter.

Adresse <https://twitter.com/Qofficiel/status/1328790471531966464>

45. Orphelin E. (2020), « L'analyse des données des eaux usées pourrait devenir un indicateur essentiel dans le suivi de l'épidémie. Seul bémol : pas d'OpenData à court terme, pour des raisons de propriété des résultats, partagées entre de nombreux acteurs... Quel dommage ! », *Twitter*. Adresse : <https://twitter.com/eorphelin/status/1329366338058465280>

46. Données relatives à la concentration de SARS-CoV-2 dans les eaux usées – Une demande d'accès à l'information à Ministère de l'enseignement supérieur, de la recherche et de l'innovation, *Ma Dada*. Adresse : https://madada.fr/demande/donnees_relatives_a_la_concentra

par un variant d'intérêt⁴⁷. Le 19 février, Elias Orphelin sur Twitter⁴⁸ repérait un article de La Voix du Nord qui évoque un tableau Excel sur le bureau du Premier ministre Jean Castex avec la part de variant et l'évolution du taux d'incidence dans chaque département⁴⁹. Le contributeur de CovidTracker demandait sur Twitter la publication de ces données toujours pas accessibles au citoyen et formulait trois jours plus tard une demande d'accès aux documents administratifs⁵⁰ au ministère de la Santé via madada.fr. Guillaume Rozier s'est fait l'écho de la demande d'Elias Orphelin en lançant sur Twitter le hashtag #ExcelDeCastex repris une centaine de fois et en regrettant dans l'émission C à Vous sur France 5 la non-publication de ces données⁵¹. Le 8 mars 2021, les données sur la part des variants par département ont été publiées par Santé Publique France sur data.gouv.fr en réponse à cette mobilisation.

CovidTracker s'est aussi illustré dans l'ouverture des données liées à la vaccination. Le 27 décembre 2020, la première dose de vaccin Covid est administrée à une septuagénaire en Seine-Saint-Denis. La campagne de vaccination s'appuie sur un système d'information nommé VAC-SI (Vaccin Covid-SI) lancé le 4 janvier 2021. Le suivi de la campagne de vaccination était devenu essentiel, sa lenteur a beaucoup été critiquée en comparaison avec d'autres pays européens. Le 27 décembre 2020, Guillaume Rozier avait lancé VaccinTracker, un site présentant des graphiques et des cartes pour comprendre l'évolution de la campagne vaccinale⁵². Mais l'accès aux données du nombre de vaccinations en France a pris plusieurs semaines. Lors de la présentation de VAC-SI à la presse le 4 janvier 2021, le directeur général de l'Assurance maladie, Thomas Fatome, avait promis que les autorités sanitaires feraient un bilan quotidien de la campagne de vaccination dès le 5 janvier 2021⁵³. Pourtant, pendant la première semaine de janvier, seuls des chiffres arrondis ont été communiqués au public sur Twitter. Le ministère avait justifié ce choix par une volonté de privilégier le lancement de la vaccination à la remontée des chiffres. Le 30 décembre 2020, Guillaume Rozier avait été invité au ministère de la Santé à une réunion pour lui présenter la stratégie vaccinale en France et les difficultés rencontrées par les autorités sanitaires dans la remontée des chiffres⁵⁴. Quelques jours plus tard, le ministre de la Santé Olivier Véran s'est entretenu avec Guillaume Rozier le félicitant sur BFM TV pour son travail. À partir du 30 décembre, son cabinet a transmis à Guillaume Rozier, chaque jour, par SMS, les chiffres quotidiens de la vaccination. Le fondateur de CovidTracker publiait alors quotidiennement les chiffres de la campagne de vaccination, étant alors la seule source d'informations publiques sur ce sujet. Le 11 janvier, il mettait fin à cette exclusivité en publiant un message sur le site VaccinTracker relayé sur Twitter⁵⁵ : « *Le ministère de la Santé a contacté CovidTracker le 30 décembre afin de lui fournir des données officielles, plus exhaustives et à jour. Depuis, nous recevons régulièrement un nouveau chiffre du nombre de vaccinés de sa part, et nous le remercions pour cela. Cependant, cette situation n'est pas conforme avec nos principes d'OpenData. VaccinTracker ne sera désormais mis à jour qu'à partir de données publiques officielles, dès que celles-ci seront disponibles.* »

47. STRATEGIE DE FREINAGE DE LA PROPAGATION DES VARIANTES DU SARS-COV-2 : RENFORCEMENT SPECIFIQUE SUR LES VARIANTES D'INTERET 20H/501Y.V2 ET 20J/501Y.V3, 2021, DGS Urgent. Adresse : https://solidarites-sante.gouv.fr/IMG/pdf/dgs-urgent_no12_strategie_variant.pdf

48. Orphelin E. (2021), « Sur le bureau de @JeanCASTEX, un Excel comparant pour chaque département l'incidence et la part de variant. Monsieur le Premier Ministre, pouvez-vous svp transférer ce document aux citoyens ? Car ces chiffres ne sont toujours pas communiqués !! 😞 », *Twitter*.

49. « Covid-19 : face à "un ennemi qui change", l'exécutif obligé de naviguer à vue », 2021, *La Voix du Nord*.

50. « Circulation des variants de la Covid-19 sur le territoire - Une demande d'accès à l'information à Ministère des solidarités et de la santé », 2021, *Ma Dada*. Adresse : https://madada.fr/demande/circulation_des_variants_de_la_c

51. « C à vous sur Twitter », 2021, *Twitter*. Adresse : <https://twitter.com/cavousf5/status/1363916869414092802>

52. Rozier G. (2020), « 📊 Voici un tout nouvel outil, VaccinTracker ! Suivez l'évolution de la vaccination contre la #Covid19 en France. Nombre de personnes vaccinées, types de vaccins autorisés, phases de vaccination... », *Twitter*. Adresse : <https://twitter.com/GuillaumeRozier/status/1343209878064525312>

53. « Vaccination : un bilan quotidien à partir du mardi 5 janvier », 2021, *Le Figaro*. Adresse : <https://www.lefigaro.fr/sciences/vaccination-un-bilan-quotidien-a-partir-du-mardi-5-janvier-20210104>

54. « "Vaccintracker" : le site indépendant qui publie les chiffres de la vaccination en France », 2021, *Le Figaro*. Adresse : <https://www.lefigaro.fr/sciences/vaccintracker-le-site-independant-qui-publie-les-chiffres-de-la-vaccination-en-france-2021010>

55. Rozier G. (2021), « GRZ sur Twitter : "Message important à propos des données de vaccination. VaccinTracker ne sera désormais mis à jour qu'à partir de données publiques officielles #OpenData." », *Twitter*. Adresse : <https://twitter.com/GuillaumeRozier/status/1348576613928230912>



Figure 7 – Tweet de Guillaume Rozier annonçant la fin de la reprise des chiffres envoyés par le ministère des Solidarités et de la Santé

Se revendiquant des principes de l'open data, il refusait de mettre à jour le site tant que les données ne seraient pas publiques. Un premier pas est franchi par le ministère de la Santé le jour même, le 11 janvier, avec un communiqué de presse annonçant « plus de 138 000 » vaccinations en France depuis le début de la campagne⁵⁶ avec les chiffres pour chaque région. Les données sur le nombre quotidien de vaccinations en France et dans chaque région sont finalement publiées sur data.gouv.fr le 12 janvier 2021. Elles sont complétées le 23 janvier par des données sur les livraisons et les stocks de vaccins. À partir du 27 janvier, les données sont directement alimentées depuis VAC-SI alors qu'elles provenaient auparavant de remontées des ARS.

6. Le Hackathon Covid : renforcer et afficher la collaboration entre l'État et la société civile

Au début de l'année 2021, CovidTracker occupait une place croissante dans la médiation des données liées à l'évolution de la pandémie. Par exemple, le 25 février 2021 lors d'une conférence de presse, le Premier ministre Jean Castex s'appuyait sur un graphique représentant l'évolution du taux d'incidence des plus de 80 ans. La source affichée était CovidTracker⁵⁷. Les statistiques de recherches sur Google indiquent aussi la popularité de cette initiative, les recherches « covidtracker » ont dépassé par exemple celles sur le ministre de la Santé Olivier Véran entre mars et décembre 2021.

56. Ministère de la santé (2021), « Vaccination contre la COVID en France : au 11 janvier 2021, plus de 138 000 personnes ont été vaccinées en France ».

Adresse : https://solidarites-sante.gouv.fr/IMG/pdf/cp_dgs_chiffres_vaccination_110121_.pdf

57. Libeau D. (2021), « Pourquoi "CovidTracker" a été adopté par le gouvernement et les médias (entre autres) », *blog.DavidLibeau.fr*.

Adresse : <https://blog.davidlibeau.fr/pourquoi-covidtracker-a-ete-adopte-par-le-gouvernement-et-les-medias-entre-autres/>



Figure 8 – Évolution du nombre de recherches pour covidtracker, Olivier Véran et Oliver Veran entre mars 2020 et mai 2022 (Source : Google Trends)

À partir d'avril 2021, CovidTracker a étendu son champ d'action au-delà de la visualisation et de l'analyse de données en intervenant directement dans la campagne de vaccination avec le lancement de « Vite Ma Dose ! », un nouvel outil qui détecte automatiquement le prochain rendez-vous disponible à proximité sur Doctolib puis, progressivement, sur l'ensemble des plateformes de prise de rendez-vous. En proposant ce service répondant à la grande difficulté pour trouver un créneau à l'époque, la notoriété de CovidTracker a alors explosé. Au même moment, le site Covidliste est lancé à l'initiative de Martin Daniel, data scientist et Mathieu Ripert, ingénieur, en collaboration avec Dr Antoine Roux, pneumologue. Covidliste permet aux centres de vaccination de toucher rapidement les personnes à proximité inscrites sur une liste d'attente pour leur proposer les doses restantes à la fin de journée ou proches de leur expiration. Ces sites entièrement maintenus et animés par des bénévoles jouent progressivement un rôle important dans la campagne de vaccination. Ils ont aussi été critiqués pour renforcer les inégalités d'accès à la vaccination en favorisant les personnes les plus à l'aise avec les outils numériques.

L'émergence de ces nouveaux outils interrogeait au sein du gouvernement et de l'administration. D'une part, ils représentaient une forme d'idéal de l'open data dans lequel la société civile s'approprie rapidement les données ouvertes et propose de nouveaux services adaptés aux besoins de la population. D'autre part, Covidliste et Vite Ma Dose ont développé des services qui auraient pu être proposés par le service public. Ces plateformes ont aussi dû produire ou collecter des données privées (disponibilités des rendez-vous sur les plateformes) ou personnelles (candidats à la vaccination). Malgré sa proximité d'objet avec Covidliste, Vite Ma Dose s'est différencié en ne collectant pas de données à caractère personnel et en publiant en open data les données des rendez-vous disponibles. Face au succès de Vite Ma Dose, Guillaume Rozier a pu être présenté comme celui qui « ringardise le ministère de la Santé »⁵⁸. Plusieurs

58. « Guillaume Rozier : le CV du jeune ingénieur qui ringardise le ministère de la Santé », 2021, *L'Express.fr*. Adresse : https://www.lexpress.fr/actualite/societe/guillaume-rozier-le-cv-du-jeune-ingenieur-qui-ringardise-le-ministere-de-la-sante_2148411.html

articles opposaient l'agilité et l'innovation d'acteurs comme CovidTracker aux « lenteurs » de l'administration. Pourtant, Guillaume Rozier et les bénévoles de CovidTracker semblaient plutôt demandeurs de collaboration avec l'administration. Lors d'un échange en mars 2021 avec l'équipe de la Direction interministérielle de la transformation publique (DITP) en charge du plan d'action « gouvernement ouvert », portant sur la transparence de l'administration et la collaboration avec la société civile dans la mise en œuvre des politiques publiques, Elias Orphelin avait proposé l'idée d'un hackathon portant sur les données ouvertes liées à la pandémie. Un hackathon est un événement qui, sur une durée courte (souvent deux jours), rassemble un public divers qui contribue de manière bénévole à résoudre un certain nombre de défis, en développant de nouveaux outils (Goëta, 2018).

La proposition du hackathon a attiré l'attention de la DITP qui, début avril, apportait son soutien à l'initiative afin de faire émerger de nouveaux outils pour lutter contre la pandémie, mais aussi montrer le caractère innovant et collaboratif de l'administration et donner de la visibilité à la collaboration entre administration et société civile dans le cadre du programme « gouvernement ouvert. » Avec plusieurs collègues de Dataactivist, nous avons collaboré avec la DITP pour organiser et animer le hackathon Covid les 23 et 24 avril 2021, en lien avec de nombreuses administrations et acteurs de la société civile. Selon sa présentation officielle, ce hackathon visait à « unir [les] forces pour contribuer à mettre fin à la pandémie », « amplifier le mouvement d'ouverture des données et du code source déjà engagé par plusieurs administrations » et montrer que « l'ouverture et la collaboration répondent de manière plus efficace et rapide aux enjeux sans cesse renouvelés de la crise⁵⁹ ».

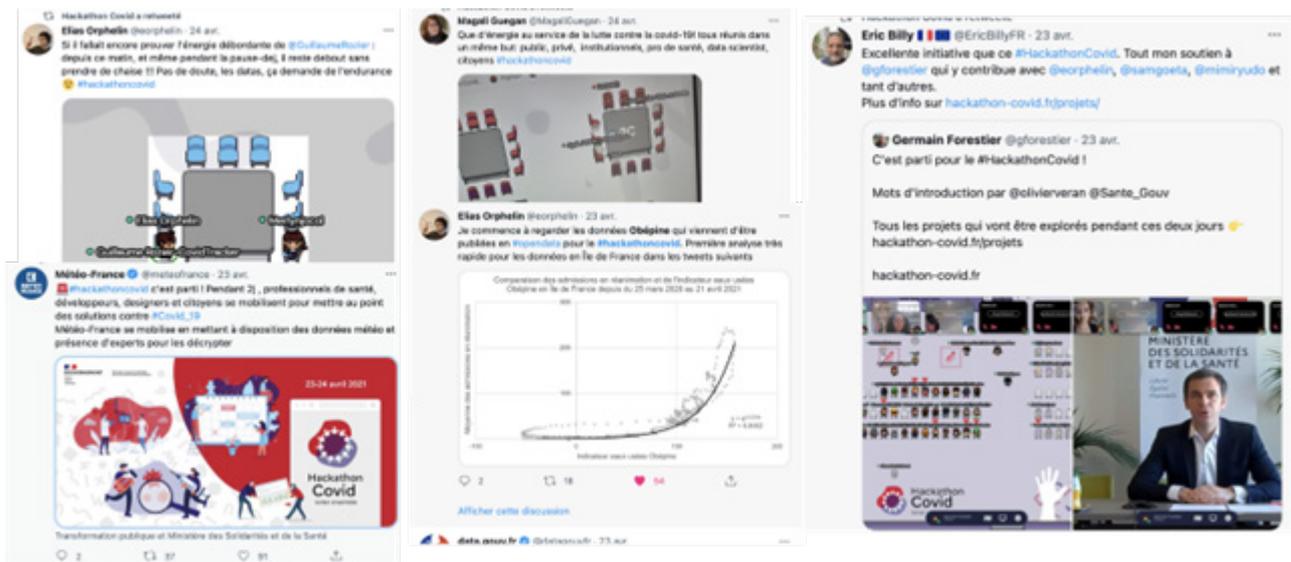


Figure 9 – Sélection de tweets publiés à l'occasion du hackathon Covid en mai 2021

Les projets des participants devaient répondre à un des quatre défis : mieux accompagner les victimes de la Covid à court et long terme, fluidifier la campagne de vaccination, suivre différemment l'évolution de la pandémie, mesurer les risques et favoriser les comportements vertueux. L'événement rassemblait de nombreuses administrations (Santé Publique France, Assurance Maladie, ARS Île-de-France...) et des organisations citoyennes (CovidTracker, Covidliste, collectif Du Côté de la Science...). Pendant deux jours dans un espace en ligne facilitant l'interaction entre les participants, les équipes ont fait émerger une dizaine de projets comme un mémorial des victimes de la Covid ou un outil d'analyse de l'effet des conditions

59. Hackathon Covid. Adresse : <https://hackathon-covid.fr/>

climatiques sur la propagation du virus. Certains projets ont été accompagnés par le ministère de la Santé par la suite comme un quizz sur les idées reçues liées à la vaccination qui a été relayé par le ministère de la Santé⁶⁰ ou un projet visant à faciliter l'échange de doses entre centres de vaccination qui a été soutenu dans son développement par une subvention visant à l'étendre potentiellement à d'autres produits sanitaires aux contraintes similaires.

Le hackathon a permis l'ouverture de nouvelles données comme celles d'OBEPINE qui, malgré l'appui du MESRI, n'étaient toujours pas publiées. Predict Services, une entreprise spécialisée dans la prévision des risques climatiques ayant Météo France pour actionnaire, a publié à l'occasion du hackathon l'Index PREDICT de Transmissivité Climatique de la Covid-19 (IPTCC). Fin mai 2021, l'Assurance Maladie a publié un portail intitulé Data vaccin Covid⁶¹ qui présente des données complémentaires de celles de Santé Publique France en indiquant les taux de vaccination par département de résidence et les taux de couverture vaccinale pour 50 pathologies graves et/ou chroniques⁶². À l'occasion du hackathon, avec Elias Orphelin de CovidTracker, nous avons formulé une demande d'ouverture des données sur le nombre de cas et d'hospitalisations selon le statut vaccinal des personnes. L'idée avait émergé pendant le hackathon que ces données puissent permettre d'évaluer l'impact de la vaccination sur les contaminations ou les hospitalisations, voire le nombre de vies « sauvées » par le vaccin. Ces données n'étaient pas collectées, elles nécessitaient de relier les systèmes VAC-SI, SI-DEP et SI-VIC qui n'avaient pas d'identifiants communs⁶³. De premiers résultats du croisement de ces bases ont été publiés en juillet par la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES) du ministère de la Santé⁶⁴, puis ouverts sur le portail de la DREES⁶⁵. Les données indiquent à l'échelle nationale la part des personnes vaccinées ou non selon leur classe d'âge parmi les dépistages, hospitalisations et placements en soins intensifs pour infection au coronavirus. Elles ont donné lieu à la création par CovidTracker de VaxImpact, un site qui vise l'impact de la vaccination sur l'épidémie et calcule la réduction des risques d'hospitalisation du fait de la vaccination. Consécration de ses travaux, le 20 juillet 2021, Guillaume Rozier était décoré Chevalier de l'ordre national du Mérite par le président de la République. Il a dédié sa médaille aux nombreux contributeurs qui l'ont aidé à développer CovidTracker et les services associés comme ViteMaDose ou VaccinTracker.

7. Conclusion

Que retenir de cette rétrospective de l'ouverture des données sur la pandémie de Covid-19 en France ? Bien que j'aie proposé un cadrage sur la France, à plusieurs reprises j'ai présenté des initiatives internationales qui ont servi de modèles aux acteurs investis dans l'ouverture des données. On a pu aussi voir l'émergence d'acteurs comme l'université Johns Hopkins ou Our World In Data qui a récupéré et visualisé les données d'un grand nombre de pays pour les comparer.

60. Ministère de la santé (2021), « Ministère des Solidarités et de la Santé sur Twitter : "#COVID19 | Le #HackathonCovid organisé en avril a permis de faire émerger des projets citoyens pour lutter contre la pandémie. Le Vaccino Quiz en est un ! Il permet de tester ses connaissances & de s'informer sur la vaccination Lancez-vous 📱 », *Twitter*. Adresse : https://mobile.twitter.com/sante_gouv/status/1408681433359699975

61. Assurance Maladie (2021), « Accueil - Data vaccin Covid ». Adresse : <https://datavaccin-covid.ameli.fr/pages/home/>

62. Assurance Maladie, « Covid-19 : lancement de Data vaccin Covid, un site pour visualiser la couverture vaccinale en France | ameli.fr | Assuré ». Adresse : <https://www.ameli.fr/bouches-du-rhone/assure/actualites/covid-19-lancement-de-data-vaccin-covid-un-site-pour-visualiser-la-couverture-vaccinale-en-france>

63. Gouthière F. (2021), « Sur quoi se base Olivier Véran pour affirmer que 93% des patients Covid admis dans 30 grands hôpitaux de France ne sont pas vaccinés ? », *Libération*. Adresse : https://www.liberation.fr/checknews/sur-quoi-se-base-olivier-veran-pour-affirmer-que-93-des-patients-covid-admis-dans-30-grands-hopitaux-de-france-ne-sont-pas-vaccines-20210719_L7T6UG4UGVE4FDCRHLJZ7P5XRY

64. Drees (2021), « Les appariements SI-VIC, SI-DEP et VAC-SI | Direction de la recherche, des études, de l'évaluation et des statistiques ». Adresse : <https://drees-site-v2.cegedim.cloud/sources-outils-et-enquetes/les-appariements-si-vic-si-dep-et-vac-si>

65. « Covid-19 : résultats par âge issus des appariements entre SI-VIC, SI-DEP et VAC-SI », 2021, *DATA.DREES*. Adresse : <https://data.drees.solidarites-sante.gouv.fr/explore/dataset/covid-19-resultats-par-age-issus-des-appariements-entre-si-vic-si-dep-et-vac-si/information/>

Cet article est centré sur les données de veille sanitaire, mais nous aurions pu aussi insister sur le fait que la pandémie a aussi joué un rôle essentiel dans le développement en France et à l'international de la science ouverte, en particulier de l'ouverture des données de la recherche. Plus largement, nous avons pu voir comment l'ouverture des données liées à la pandémie a constitué un tournant dans la politique d'open data du gouvernement et aussi dans la diffusion du concept, devenu parfois un sujet d'intérêt pour des émissions de télévision généralistes à grande écoute. L'accès aux données sur la pandémie a mis en lumière les faiblesses de la mise en œuvre du droit d'accès aux documents administratifs et du principe d'ouverture des données par défaut prévu par la loi pour une République numérique. Le rapport du député Bothorel au Premier ministre s'est appuyé sur ces failles révélées par la pandémie pour relancer la politique d'ouverture des données de l'État qui semblait subir une forme d'essoufflement. Dans *Les Échos*, Éric Bothorel présentait même le Covid-19 comme un décideur politique à part entière : « *La pandémie est le meilleur secrétaire d'État au Numérique qu'on ait eu depuis quinze ans, c'est un formidable accélérateur pour la coopération entre les administrations et les secteurs d'activité.* » Les initiatives de la société civile comme OpenCovid19-fr ou CovidTracker (il y en aurait bien d'autres à citer) ont aussi renouvelé le plaidoyer de l'ouverture des données en faisant de l'open data un outil essentiel de réponse à la crise sanitaire. L'émergence de médiateurs de données a été vue comme un « bouleversement » ou un « accélérateur » de changements dans certaines administrations⁶⁶. Elle a pu aussi, à bas bruit, faire l'objet de controverses comme on a pu le voir dès le début de la première vague avec les prévisions de propagation de l'épidémie et les préconisations réalisées par Tomas Pueyo, un entrepreneur français dont les billets de blog sur le Covid-19 publiés sur Medium ont cumulé plus de 40 millions de fois et été traduits dans plus de quarante langues⁶⁷.

Si l'on s'intéresse maintenant plus directement aux données, cette rétrospective souligne l'importance des systèmes d'information dans la collecte et la circulation des données liées à la pandémie. Dans les premiers temps de la crise sanitaire, les autorités se sont appuyées sur des outils existants (SI-VIC, Sentinelles, réseau « 3labos »...) qui ont été complétés progressivement par des systèmes d'information dédiés au Covid-19 comme SI-DEP, VAC-SI ou le réseau OBEPINE. Là où le plan de prévention de pandémie grippale de 2009 édictait des principes et donnait une marche à suivre, les autorités ont dû développer de véritables infrastructures de données pour faire face à la situation inédite d'extension de l'épidémie sur tout le territoire. Derrière une grande partie des frictions (Edwards *et al.*, 2011) ou des retards dénoncés par certains acteurs et journalistes comme des « couacs » ou des « cafouillages⁶⁸ », se cachent certainement des situations complexes et inédites où la circulation des données n'allait pas de soi et demandait un travail important en amont. La question de la fiabilité des données est aussi revenue régulièrement, les producteurs comme les usagers n'hésitant pas à avertir sur les incertitudes ou les limites des données utilisées. Cette situation s'est présentée par exemple pendant la première vague quand les tests n'étaient pas disponibles en nombre suffisant et que les chiffres de dépistage rendaient compte d'une réalité très éloignée de la situation épidémique. La fermeture des laboratoires lors des jours fériés a aussi donné l'impression de baisses factices de la circulation du virus. Les acteurs ont appris progressivement à composer avec ces incertitudes ou à s'adapter à une situation mouvante, par exemple avec l'émergence des variants ou des vaccins qui ont changé la donne à plusieurs reprises.

Enfin, on peut s'interroger sur les effets de l'ouverture de ces données. Guillaume Rozier a

66. Godeluck S. (2021), « L'open data, une révolution silencieuse dans l'administration », *Les Echos*, n° 23488.

67. Martin N. (2021), « Coronavirus : une épidémie de faux articles scientifiques », *France Culture*. Adresse : <https://www.franceculture.fr/emissions/radiographies-du-coronavirus-la-chronique/radiographies-du-coronavirus-du-lundi-16-mars-2020>

68. Le Borgne B. (2021), « Urgence, cafouillages, "mille-feuille" d'indicateurs... Dans les coulisses des données très stratégiques sur le Covid-19 », *France Info*.

Adresse : https://www.francetvinfo.fr/sante/maladie/coronavirus/urgence-cafouillages-mille-feuille-d-indicateurs-dans-les-coulisses-des-donnees-tres-strategiques-sur-le-covid-19_4099757.html

répété à plusieurs reprises que l'open data était une « arme anti-complotisme » expliquant que « si les autorités disent "il y a 300 morts aujourd'hui", vous aurez toujours des gens qui diront "ils mentent, ce n'est pas vrai !" Les données publiques, détaillées par département, par jour, par âge, par sexe, permettent de vérifier ce qu'il en est.⁶⁹ » Une affirmation qui rejoint celle d'Amélie de Montchalin, la ministre de la Transformation et de la Fonction publiques, sur le site de *L'Opinion* : « Les gens pensent que les chiffres mis en ligne par CovidTracker constituent la vérité. Si c'est nous qui les donnons, on pense que l'on ment. » Ouvrir des données à un haut niveau de précision et permettre à quiconque de les analyser renforce-t-il la confiance dans les faits ? Les premiers travaux parus sur la pandémie ont alerté sur la pollution cognitive que l'explosion des indicateurs et des visualisations de données génère, phénomène décrit sous le terme d'*infodemics* (Grusd et al., 2021). D'autres ont montré que la circulation des données sur la pandémie au contraire alimente l'ère dite de « post-vérité » qui réduit l'influence des faits objectifs par rapport aux appels à l'émotion ou aux opinions personnelles (Shelton, 2020). Dans un article traduit dans cette revue, William Davies (2017) avait déjà alerté contre le ressentiment que pouvait créer le gouvernement par les nombres. La période que nous traversons, où une série d'indicateurs sanitaires déterminent notre quotidien, a probablement entretenu ce phénomène. L'ouverture des données semble être une condition nécessaire à la confiance, je doute toutefois qu'elle soit suffisante.

Toutes ces questions seront au cœur du projet de recherche DoMeSCO (Données et Médiation scientifique : leçons du Covid-19) pensé autour d'un partenariat entre le Centre de sociologie de l'innovation (Mines ParisTech, i3 – UMR 9217) et Dataactivist. Soutenu par l'ANR dans le cadre de l'appel à projets « Science avec et pour la société : Médiation et communication scientifiques », ce projet de recherche a pour objectif de structurer et de diffuser les enseignements qui peuvent être tirés de cette période afin d'outiller les actions de médiation futures, de guider les modalités de production et d'ouverture des données au sein de l'administration publique et de faciliter les collaborations entre les deux⁷⁰.

Références

Chignard S. (2021), « L'open data de crise : entre mobilisation citoyenne et communication gouvernementale », *Annales des Mines*, n° 14, p. 73-77.

Davies W. (2017), « Comment la statistique a perdu son pouvoir - et pourquoi nous devrions craindre ce qui va suivre », *Statistique et Société*, vol. 5, n° 1, pp. 11-20.

Dong E., H. Du et L. Gardner (2020), « An interactive web-based dashboard to track COVID-19 in real time », *The Lancet Infectious Diseases*, vol. 20, n° 5, pp. 533-534.

Edwards P. N., M. S. Mayernik, A. L. Batcheller, G. C. Bowker, et al. (2011), « Science friction: Data, metadata, and collaboration », *Social Studies of Science*, vol. 41, n° 5, pp. 667-690.

Goëta S. (2018), « "Données recherchent publics" : les politiques d'open data à l'épreuve de la réutilisation », in G. Gourgues et A. Mazeaud (éds.), *L'action publique saisie par ses « publics » : Gouvernement et (dés)ordre politique*, Villeneuve d'Ascq, Presses universitaires du Septentrion, pp. 137-154.

69. Floc'h B. (2021), « Covid-19 : l'enjeu crucial des données publiques », *Le Monde*. Adresse : https://www.lemonde.fr/planete/article/2021/03/06/covid-19-l-enjeu-crucial-des-donnees-publiques_6072158_3244.html

70. Un carnet de recherche sur la plateforme Hypotheses permet de suivre les actualités du projet : <https://domesco.hypotheses.org/>

Gruzd A., M. De Domenico, P. L. Sacco, and S. Briand (2021), « Studying the COVID-19 infodemic at scale », *Big Data & Society*, vol. 8, n° 1, <https://doi.org/10.1177/20539517211021115>

Ronai M. (2021), « La numérisation à marche forcée du système de santé face à la Covid-19 », *Annales des Mines*, n° 14, p. 52-61.

Shelton T. (2020), « A post-truth pandemic? », *Big Data & Society*, vol. 7, n° 2, <https://doi.org/10.1177/2053951720965612>

Strasser B. J. (2011), « The Experimenter's Museum GenBank, Natural History, and the Moral Economies of Biomedicine », *Isis*, vol. 102, n° 1, pp. 60-96.

Biostatistics and the COVID-19 Pandemic in Belgium, in 2020 and 2021



Geert MOLENBERGHS¹

Interuniversity Institute for Biostatistics and statistical Bioinformatics, Data Science Institute, Hasselt University and KU Leuven, Belgium

TITRE

La biostatistique et la pandémie de COVID-19 en Belgique, en 2020 et 2021

ABSTRACT

From a statistician's standpoint, the first two years of the COVID-19 epidemic in Belgium are described. First, a narrative of the epidemic evolution is given, with attention for the regional differences within the country. Second, the policy advisory board structure is reviewed. The role of biostatistics is described next, both in terms of scientific research as well as regarding our profession's role in policy making and communication to media and general public. Two sectors are given specific attention: travel and education. Some general conclusions and recommendations are formulated.

Keywords: *data sharing, SARS-CoV-2, infection fatality rate, mathematical epidemiology, mathematical modeling, mortality, non-pharmaceutical intervention.*

RÉSUMÉ

Les deux premières années de l'épidémie de COVID-19 en Belgique sont décrites du point de vue d'un statisticien. Tout d'abord, un récit de l'évolution de l'épidémie est donné, en tenant compte des différences régionales au sein du pays. Deuxièmement, la structure des conseils consultatifs est examinée. Le rôle de la biostatistique est décrit ensuite, tant en termes de recherche scientifique que du rôle de notre profession dans l'élaboration des politiques et la communication aux médias ainsi qu'au grand public. Deux secteurs font l'objet d'une discussion particulière : les voyages et le tourisme d'une part et l'éducation d'autre part. Quelques conclusions et recommandations générales sont formulées.

Mots-clés : *intervention non-pharmaceutique, modélisation mathématique, mortalité, épidémiologie mathématique, partage de données, SARS-CoV-2, taux de létalité de l'infection.*

1. geert.molenberghs@uhasselt.be

1. Introduction

To encounter a health crisis of a scale comparable to the SARS-CoV-2 induced COVID-19 pandemic, one has to return to the 1918 H1N1 pandemic – the Spanish flu. For other crises of this magnitude, only world wars and perhaps financial meltdowns come to mind.

The extent of the pandemic is truly global in the sense that not a single country remains unaffected, albeit it to differing degrees.

When the virus spread from Wuhan China, notably in the early months of 2020, countries like Iran, the United States of America, and many European countries were among the earliest ones hit. The first blow on the European continent was dealt to the North of Italy, followed by Spain, France, Belgium, the Netherlands, the United Kingdom, and a range of other countries. Of note is that the more eastern European countries were relatively mildly hit in the early days.

Confronted with a respiratory infection caused by an hitherto novel virus, it is imperative to gauge its epidemiology as soon as possible, and devise a response strategy accordingly. In the early phase, and throughout most of 2020, we could not (yet) count on either of two important lines of defense, i.e., antiviral medication and vaccines. This only left so-called non-pharmaceutical interventions (NPI). Under such circumstances, there are essentially three response strategies.

The first is suppression of the epidemic, which is taken to mean that an extensive set of NPIs is implemented each time there is viral circulation, so as to keep incidence at or near zero. This strategy has been followed, for example, by China (apart from Hubei province in the early phase), Australia, and New Zealand. The advantage is that a society can function relatively normally for larger periods of time, and that measures taken are typically shorter in duration. The disadvantage is that such societies need to be sealed off from others that follow a different strategy. At the other end of the spectrum is so-called herd immunity, where the virus is allowed to circulate in a (nearly) unlimited fashion until the reservoir of susceptibles is so much depleted as to naturally end circulation. Several countries have considered this strategy, such as the United Kingdom, the Netherlands, and Sweden. However, it soon transpired that the hospitalization and mortality rates were such that this strategy is difficult to maintain from an ethical standpoint. In particular, in virtually no society can this strategy be implemented without overwhelming the health care system, thus potentially amplifying the death toll (Hens, Vranckx, and Molenberghs, 2020). The third strategy strikes a middle ground between the other two and is referred to as mitigation: the viral circulation is kept at a level low enough such as to ensure that the health care system is able to cope with the burden of disease induced by the virus.

The strategy of mitigation has been followed by most European countries, especially within the European Union. Given the interconnectedness of the European populations, and the legal framework imposed by the EU, it is very difficult to have radically divergent strategies. However, the time period under investigation here has shown that it is very difficult to regulate the burden of disease; rather, most European countries have had a sequence of waves of varying degrees of severity.

My personal role can be described as follows. From a scientific perspective, I have been involved in research, among others, on COVID-19 mortality (Molenberghs *et al.*, 2022), education (Bekkering *et al.*, 2020), the role of higher education (Natalia *et al.*, 2022), spatial aspects of the pandemic (Neyens *et al.*, 2020; Vranckx *et al.*, 2021), COVID-19 in the labor force sectors (Verbeeck *et al.*, 2021), and vaccine willingness (Delporte *et al.*, 2021). I was a member of CELEVAL and, at the time of writing, of both RAG and GEMS. Apart from frequent interaction with policy makers, I have taken up a role in communicating with both written and spoken mainstream media.

In Section 2, we will briefly describe the course of the epidemic in Belgium, and place it in a European perspective, with particular attention for the neighboring countries (France, the Netherlands, Germany, Luxembourg, but also the United Kingdom, given its proximity). Some attention is devoted to the typical regional differences within Belgium. The advisory board structure is reviewed in Sec-

tion 3. Section 4 focuses on the role played by biostatisticians from a research perspective. Their role in policy making, communication, and public debate is discussed in Section 5. Scientific government advisors often also provide advice to specific sectors. In Section 6, two examples are given. In Section 6.1, we address the specific challenge posed by transborder travel, as well as intra-European and international travel. One of the sectors of great importance is education, both compulsory and higher education. In Section 6.2 we turn to this sector and briefly discuss what measures were taken to essentially keep compulsory education open during the school year 2020–2021.

2. The Epidemic in Belgium, in 2020 and 2021

2.1. The Epidemic's Evolution

The first wave hit Belgium relatively soon after the virus arrived in Europe, facilitated by, among others, travel related to winter holidays and carnival festivities. The Belgian regions most severely hit were, broadly, situated in an east-west band in the middle of the country. The first confirmed corona death was registered on March 10, 2020. There was a rapid rise in cases, even though in the early days it was difficult to quantify because of very limited testing capacity. Hospitalization, ICU beds taken, and mortality were more reliable indicators. The ICU peak was reached on April 8, 2020, with 1286 beds taken, implying that normal capacity had to be scaled up and non-COVID care postponed.

Relatively stringent measures, of a lockdown type, were in place during this period, although less so than in, for example, Italy, Spain, or France. Because measures taken can take a variety of forms, the so-called *stringency index* was developed, for the purpose of the current pandemic, by the University of Oxford based Blavatnik School of Government (Hale *et al.*, 2020). Stringency is 0 when there are no measures taken, with 100 being the highest value. Table 1 shows the stringency indices for Belgium and selected countries, on the first day of April 2020, July 2020, and February 2021, respectively.

Table 1: *Stringency indices for Belgium, France, Italy, and Spain, on the first day of April 2020, July 2020, and February 2021.*

Country	01/04/2020	01/07/2020	01/02/2021
Belgium	81.5	50.0	63.0
France	88.0	51.9	63.9
Italy	91.7	58.3	78.7
Spain	85.2	41.2	71.3

It is clear that, apart from early summer 2020, Belgium had the mildest stringency among the four countries listed, whereas Italy maintained the most stringent measures throughout.

The epidemic situation improved in late April and May, and sectors started to gradually reopen. Like in many European countries, measures were drastically relaxed by the start of summer 2020, and international travel, in particular for tourist purposes, resumed. The number of cases in Belgium started to rise again from mid-July 2020 onwards, this time predominantly in the north of the country (Province of Antwerp) and Brussels. Of note is that particularly densely populated inner city areas were affected. The stringency of the NPIs was increased again in late July 2020, a curfew in the city of Antwerp. Because of these, and arguably supported by the seasonal effect, the rapid rise in the Province of Antwerp was countered. By the end of August 2020, roughly 400 cases per day were confirmed – as opposed to roughly 100 by the beginning of Summer.

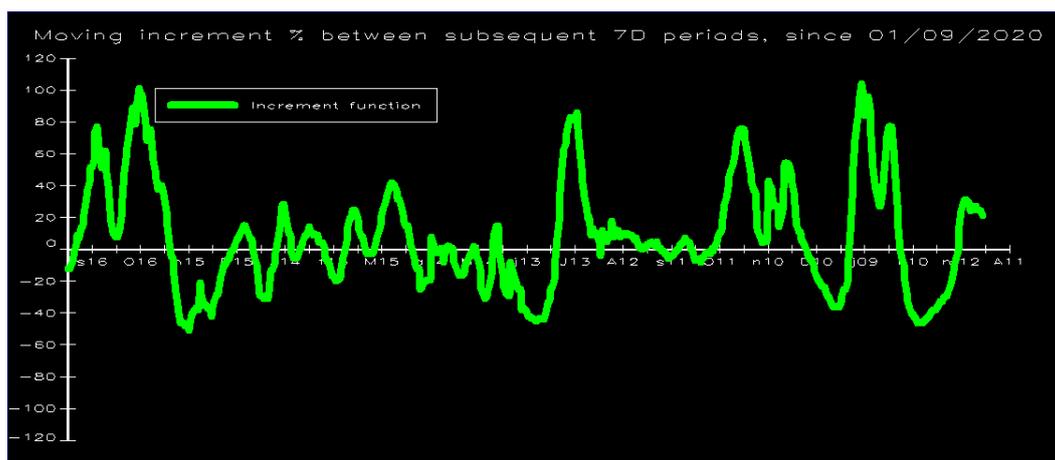


Figure 1: Increment function of the confirmed cases in Belgium, over the period September 1, 2020 – March 25, 2022. The increment is defined as the percentage change of the 7-day incidence at a given date relative to the immediately preceding period of seven days. Data source: Sciensano daily epidemiological bulletin (<https://covid-19.sciensano.be/fr/covid-19-situation-epidemiologique>).

The evolution of the epidemic in Belgium is represented by the increment function over the period September 1, 2020 – February 28, 2022. The increment is defined as the percentage increase or decrease of confirmed cases, accumulated over a seven-day time period, as opposed to the immediately preceding seven days. Precisely, let W_1 be the total number of confirmed cases over a given 7-day period and W_0 the total number of the immediately preceding week, then the increment function for the last day of W_1 is defined as $(W_1 - W_0)/W_0$. Working with 7-day periods mitigates day-of-the-week effects. The function supplements the information in the curve representing the cases itself, like a derivative function does; of course, the latter is very variable because of daily fluctuations. During the first half of September 2020, there was a rapid rise in cases, to a large extent provoked by seeding via returning travelers, as is corroborated by both phylogenetic (Hodcroft *et al.*, 2020) as well as epidemiological (Natalia *et al.*, 2022) evidence. During the second half of September 2020, the increase in cases slowed but did not revert to a decrease, only to pick up again in October 2020, this time catalyzed by, among others, higher education. A typical Belgian phenomenon is that students sojourn near the educational institution during weekdays, to then return to their home town over the weekend. Such “conveyer belts” of contacts led to a quick spread across the country, but this time with focus on Wallonia. The second wave that ensued reached its peak early November 2020, with maximal ICU occupancy of 1474 on November 9, 2020. Around the peak, roughly 20,000 cases per day were confirmed. The 14-day incidence per 100,000 inhabitants in Belgium peaked near 1800, whereas the incidence for the three regions at peak was about 1000 for Flanders, 2000 for Brussels, and 3000 for Wallonia.

Early December 2020, the number of confirmed cases per day had fallen to roughly 2000 per day. NPIs were made less stringent, and a relatively long “plateau phase” began. A resurgence around Christmas and New Year was avoided as no relaxations were implemented for this period, in part also with the emergence of the alpha variant in mind, and supported by mathematical modeling, which showed that relaxing measures before Christmas 2020 could have led to a considerable peak. Of note, such peaks appeared in Ireland and Portugal at that time. The gradual increase of the alpha variant necessitated frequent revision of measures over the winter and early Spring 2021 period, but eventually a new wave ensued, on top of the plateau phase. On April 13, 2021, the number of beds occupied in ICU equaled 947. The alpha variant provoked a higher burden of disease, even against the background of a well progressing vaccination campaign.

Over the months of May and June 2021, the numbers declined rapidly, and early July 2021, the number of confirmed cases per day was around 300–400, with about 20 hospital admissions per day, ICU occupancy below 100, and about 5 deaths per day.

From July 2021 onwards, the epidemic curves started rising again (as is also clear from Figure 1), a process that continued at a relatively gentle pace, until a fourth (delta) wave occurred in October–November 2021. At the end of 2021, the delta wave gave way to its omicron successor.

2.2. Some Key Regional Differences in Belgium

Belgium's three regions have considerable differences. The Northern Flanders is Dutch speaking, the Walloon south is French speaking, with the Brussels Capital Region bilingual, though with a vast majority of French speakers. Brussels being a city region with both a national and an international mission, it differs in various aspects from the other two regions. For example, Brussels welcomes more foreign travelers per capita than the other two regions, as is clear from Figure 2. It is also a very diverse region, with varying access to official government communication and mainstream media.

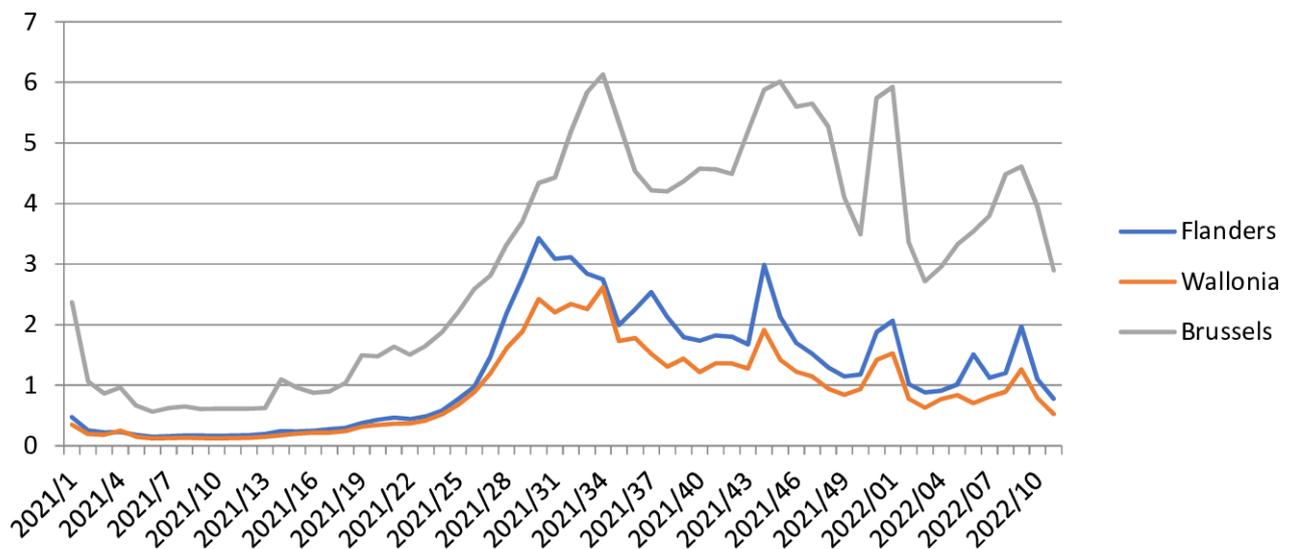


Figure 2: Travel rates (number of incoming travelers per 100 inhabitants) for each of the three Belgian Regions: Brussels, Flanders, and Wallonia, starting from 1 January 2021. Based on the Sciensano weekly epidemiological bulletins (<https://covid-19.sciensano.be/fr/covid-19-situation-epidemiologique>).

For various reasons, therefore, it is not surprising that the 14-day incidence curves for the three regions are not always synchronized, as is clear from Figure 3. During the Fall wave of 2020, the incidence in Wallonia rose to about 3000, whereas it topped off at about 2000 and 1000 for Brussels and Flanders, respectively. However, in the Fall wave of 2021, it was Flanders who reached the highest incidence, during the delta wave.

The latter fact is a bit surprising at first sight, given the high vaccination rate in Flanders. Indeed, on 31 December 2021, 60% of the entire Brussels population was fully vaccinated, whereas the number for Flanders is 81% and 72% for Wallonia. However, as vaccines provide good but not perfect protection, the high incidence in Flanders should be seen against the background of more extensive (high risk) contact behavior in the region in the Fall of 2021, as compared to the other two.

Also the booster campaign is evolving at a different in the three regions, with on 25 February 2022,

booster vaccination rates among the 18+ population of 46% in Brussels, 83% in Flanders, and 65% in Wallonia.

Because of the multilingual structure of Belgium, the regions consume different media, and have their own public opinion foci. This implies that the balance between the sanitary, economic, and well-being aspects may not always be drawn in the same way. For example, when NPIs were reinstated in Summer 2020, this was perceived in the South as primarily for the purpose of counteracting the re-emergence of the virus in the North (in particular in the Antwerp area). This might explain why there was a stronger reaction against the continuing measures from Southern actors in September-October 2020. In the Fall of 2021, in Flanders the impression existed that the epidemic was under control because of the high vaccination rate, whereas a more cautious attitude was adopted in Brussels and Wallonia. Thus, while at macro level all regions have been moving through the same set of waves, the relative ordering of the regions changed over time.

Region-specific 14-day incidence curves

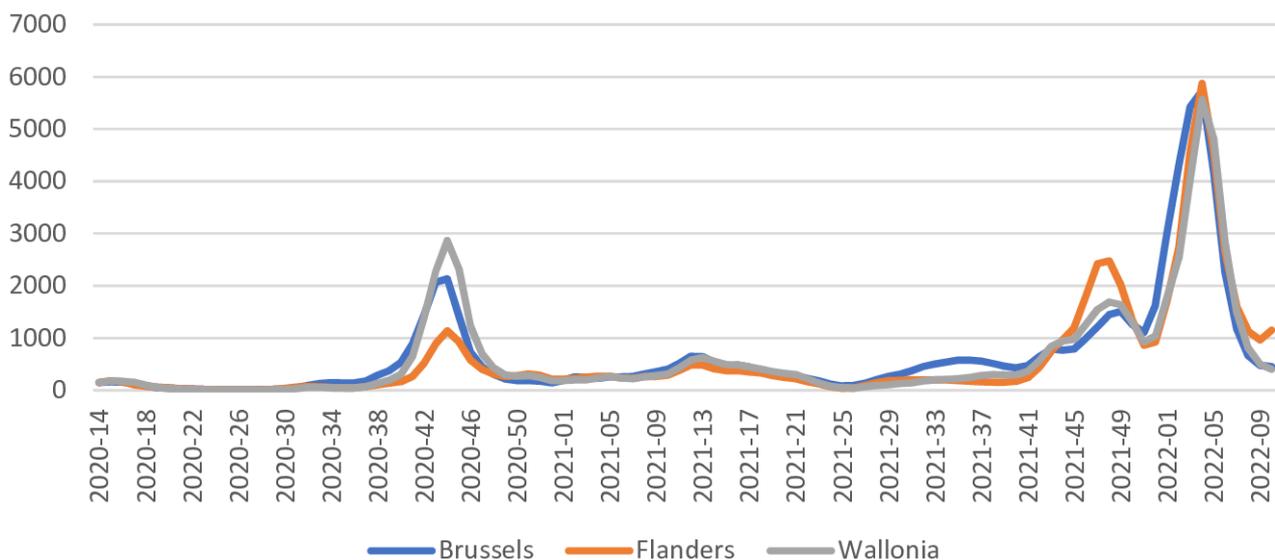


Figure 3: 14-day incidence curves from March 2020 until week 2022/10, for each of the three Belgian Regions: Brussels, Flanders, and Wallonia. Data source: ECDC (<https://www.ecdc.europa.eu/en/publications-data/weekly-subnational-14-day-notification-rate-covid-19>)

3. A Complex Advisory and Policy Making Structure

Belgium has a complex federal system. For example, the health care and public health competencies are split over the federal, regional (Flemish, Brussels, and Walloon Regions), and community levels (Flemish, Wallo-Brux, and German Communities). For simplicity, we will simply refer to the federal and regional levels. For example, hospitals are largely but not entirely federal matter, whereas nursing homes, playing such a key role in the pandemic, are a regional competency.

Thus, data that need to be consolidated, originate from various levels within the Belgian federal system. Hospital mortality is collected by the federal authorities, whereas nursing home mortality is collected by the three Regions. Data-related issues may arise for nursing home residents that are transferred to hospital and then die. It is crucial of course that every death is counted, and counted

only once. The national public health institute Sciensano had to set up data sharing protocols with both Regions and the federal authorities for this matter. In addition, this has to be done in real time, whereas in non-pandemic times, cause of mortality data become available with a couple of years of delay.

Likewise, the political, scientific, and administrative advisory structure has been complex. A number of factors are worth attention.

First, the onset of the crisis was set against the background of an acting rather than fully empowered federal government.

Second, because there is no natural hierarchy among the country's governments, dedicated coordination structures had to be invoked or put in place. Until October 2020, the principal ministers of the various governments met the National Security Council; thereafter, when the government-Wilmès was succeeded by the government-De Croo, this role was assumed by the so-called Coordination Committee.

Third, a large number of structures and boards, scientific and other, have been active in the management of the crisis. We discuss these in turn.

Sciensano. The national public health institute, responsible, among others, for surveillance, data collection, and reporting on the pandemic.

National Crisis Center. This permanent structure is responsible for both crisis communication as well as regular examination as to whether the country is prepared to face the challenges imposed by the pandemic (e.g., readiness of the health care system).

Risk Assessment Group (RAG). Sciensano hosts the RAG, consisting of scientific members of the administration, supplemented with scientists located in universities and other research institutes. The RAG's task is to carefully describe the state of the epidemic and its evolution, and to offer scientific advice to the authorities. The RAG is a permanent structure that can be invoked to handle various public health crises, not only the current pandemic.

Risk Management Group (RMG). The RAG's input is received by the RMG, at which level the scientific advice is used to generate policy advice. The RMG consists of members of the administration and governmental cabinets.

Group of Experts on the Exit Strategy (GEES). In the Spring and early Summer of 2020, the *ad hoc* GEES was charged with offering policy advice on successive relaxations of measures, directly to the Prime Minister and the key ministers involved. The board consisted of the disciplines: biostatistics, epidemiology, virology, infectiology, economy, social inequality, and European affairs.

Evaluation Cell (CELEVAL). This is a permanent but usually dormant cell, at the level of the Ministry of Public Health, but with representatives from various other ministries (such as Interior Affairs, Foreign Affairs, etc.). It was conceived at the time of the Tjernobyl nuclear crisis, and invoked again during the Summer of 2020. Like the RAG, it is populated according to the crisis at hand. Whereas RAG evaluates the epidemiological situation, CELEVAL formulated policy advice, e.g., proposed the establishment or lifting of NPIs.

In the Fall of 2020, GEES and CELEVAL have been discontinued and succeeded by new structures.

Corona Commissariat. A government Corona Commissioner has been appointed in the Fall of 2020, to coordinate all activities of all advisory boards.

Group on the Exit Management Strategy (GEMS). The GEMS can be seen as the successor of both GEES and CELEVAL; it reports to the Corona Commissariat. In spite of its name, GEMS has had to propose both increases as well as decreases in NPI stringency.

Task Forces. Finally, there are several Task Forces, charged with advice regarding specific aspects of pandemic response, such as on Testing and Tracing, on Ventilation, and on Vaccination. They advice on the policy related to their charge, and work with federal and regional authorities towards smooth implementation of decisions taken (e.g., the testing centers, the vaccination campaign).

Like the GEES, the other advisory boards have always been populated in a multi-disciplinary fashion, with members having a biomedical, exact sciences, or humanities background. These include virologists, infectiologists, vaccinologists, economists, health economists, labor force specialists, psychologists, sociologists, and pediatricians. Psychologists and sociologists have been instrumental in surveying the public regarding such key topics as mental wellbeing, motivation to follow the NPIs, vaccination willingness, etc. People working with disfavored groups have provided valuable advice to increase vaccination rates in target groups. Not surprisingly, the remit of the pediatricians is children's wellbeing, and providing advice around measures in schools and for extra-curricular activities. Because of the GEMS' larger size, it was easier to have a wider scientific representation than with, for example, GEES.

In spite of this relatively multi-disciplinary approach, there has been criticism, throughout the pandemic, that there was a certain dominance of biomedical and exact sciences on the board. This might have been provoked, at least in part, by certain constraints that applied. For example, for the GEES, the government had chosen for a board of not too large a size, 10 members to be precise. Given this small size, it was as multi-disciplinary as possible. Indeed, apart from the disciplines: biostatistics, epidemiology, virology, infectiology, there were members representing economy (Pierre Wunsch, Johnny Thys, and Mathias Dewatripont), social inequality (Céline Nieuwenhuys), and European affairs (Inge Bernaerts). It is fair to say that the balance could have been struck a bit differently, and that perhaps a slightly larger board would have been better. This is why GEMS has been composed of 25 rather than 10 people.

For the purpose of this article, it is important to mention that at several levels, statisticians and epidemiologists have been active; for example, in the capacity of members of GEES, CELEVAL, RAG, and GEMS.

4. Biostatistics Research

Biostatisticians have played various roles throughout the crisis. Biostatistics refers to the use of statistical and mathematical methods in the broad sense, including medicine, public health, biology, agriculture, and forestry. Evidently, in the current pandemic, the focus is on human health on the one hand, and the societal implications of the measures taken on the other.

In this section, we will focus on scientific research. A detailed description of the various research endeavors can be found in Molenberghs *et al.* (2020). A large fraction of the research efforts take place in a multi-disciplinary environment, encompassing one or more of the disciplines already mentioned.

A considerable part of the research efforts are geared towards modeling and monitoring the epidemic. For example, mathematical models are progressively tailored to the ever expanding and oftentimes changing epidemiological knowledge about the virus and the disease that it causes. The teams at Hasselt and Antwerp universities have over two decades of experience in mathematical modeling of infectious diseases, and could hence take part in rapid response from a biostatistical perspective.

Such elements include the existence of a pre-symptomatic period as well as fully asymptomatic infec-

tions, the length of the incubation period, the onset of infectiousness, the length of the serial interval, the phenomenon and extent of superspreading, the emergence of variants with quite different characteristics, and progressing vaccine uptake. This knowledge is, in turn, relevant towards optimally informing mathematical models (e.g., Abrams *et al.*, 2020).

Another strand of research is geared towards so-called nowcasting and early warning. To this effect, knowledge about the delay between symptom onset and hospitalization, hospital length of stay, etc. are needed, so that short-term predictions can be made. Important predictors for hospitalizations include, evidently, confirmed cases, the positivity rate, the workload of the first line of health care (general practitioners), etc. Many of these indicators come with their own issues. For example, the numbers of confirmed cases are influenced by the testing policy. In the first half of 2020, about 62,000 cases were confirmed; the number for the second half of 2020 is about 590,000, and for the first half of 2021 it is about 470,000. However, model-based estimates of the total number of cases are approximately 1 million, 1.5 million, and again 1 million, respectively. Without such an adjustment, the numbers of COVID-19 deaths would be implausible. They are, again roughly, 10,000; 10,000; and 5,000 for each of the three periods, respectively. In the Fall of 2020, the infection fatality rate (fraction of deaths among those infected, whether confirmed or not) had dropped somewhat (10,000 deaths, but among 50% more cases), due to somewhat better protection of the vulnerable population, and modest advance in antiviral treatment.

A perpetual strand of research for biostatisticians is towards the design and analysis of studies geared towards antiviral medication and vaccine development. Of course, the circumstances, including time-lines, were very different. Apart from COVID-19 related trials, biostatisticians have to be concerned with the pandemic's effect on ongoing trials in other therapeutic areas.

We also mention the estimation of important epidemiological indicators, such as seroprevalence, contribution towards testing and contact tracing efforts, and design an analysis of public opinion surveys. Such surveys are undertaken, for example, to gauge the state of the population's wellbeing in the face of a relatively long pandemic and the restrictions that come with it.

5. Biostatisticians in the Public Spotlight

Unlike in non-pandemic times, biostatisticians' role in research and policy making was often the subject of political and public debate. Moreover, they have frequently been called upon to explain various quantitative aspects of the Belgian epidemic and the pandemic at large.

While over the first months of the pandemic there was a general public consensus that appropriate measures (NPIs) were needed, and that they had been successful in flattening the curve, against the background of very severe conditions in Wuhan, China, in Bergamo, Italy, as well as in several other places, including in the south of the Belgian Limburg Province, there was a clear shift in the public opinion over the summer of 2020, in Belgium as well in a range of other countries. Arguably, this was due to the very unwelcome flare up that happened in Belgium and elsewhere in Europe, necessitating reinforced measures. Scientists, politicians, and the public realized that the pandemic would not be over with a single wave.

In Belgium, the fact that a large second wave could be avoided invoked the prevention paradox – because the intervention taken avoids an undesirable situation, it appears that this intervention was unnecessary.

5.1. Criticisms of Biostatisticians

Biostatisticians and epidemiologists and their expertise started to be called into question much more than before. Let us briefly consider three examples.

First, the death toll has been the subject of ongoing debate (Molenberghs *et al.*, 2021). For example, on July 31, 2020, Belgium had the highest number of deaths per million (849 DPM) on the planet, except for some ministates. Skeptics argued that the high death toll underscored the ineffectiveness of the NPIs. However, also here, counterfactual reasoning is important: What would the death toll have been, had no NPIs been implemented, or had less stringent measures been taken. A second point of criticism was towards Belgium's reporting of COVID-19 mortality. Indeed, both test confirmed as well as clinically confirmed cases were reported, whether they occurred in a hospital or nursing home setting. Given the very restrictive testing policy during the early weeks of the pandemic, not counting clinically confirmed cases would have drastically altered the apparent death toll – given the discrepancy between the 62,000 confirmed cases versus the estimated one million of total cases over the first half of 2020. Likewise, the death toll would have been drastically lower if only hospital mortality were reported. As Aron *et al.* (2020) showed, the reported Belgian COVID-19 mortality was close to reality, unlike in some countries such as the Netherlands and Spain, where only around 50–60% of COVID-19 deaths were reported. Upon completion of 2020, it turned out that Belgium's excess mortality and reported COVID-19 mortality were in agreement, in the sense that excess mortality, defined as the difference between the 2020 mortality and the average over the preceding five years, was about 85% of COVID-19 mortality.

Second, a debate, also well-known in other countries, that emerges when the number of confirmed cases increases is whether this is a genuine increase or rather a consequence of an expanding testing policy. To exacerbate the problem, initially increasing case numbers sometimes happen against the background of non-increasing or even decreasing hospitalization rates. Of course, this is a classical epidemiological issue. Commonly, the people being tested do not form a random sample from the population. Rather, the reason for being tested is related to the test outcome itself, such as a high risk contact or the emergence of symptoms. Nevertheless, when policies remain constant, the *evolution* of confirmed cases may be an important though impure indicator. This is why it is often considered jointly with other indicators, such as the positivity rate, or it is stratified over age and/or testing reason categories. Of course, ideally regular repeated samples should be taken according to proper survey design, such as in a sentinel network. This is one of the many important lessons that should be drawn from the pandemic, i.e., the need for properly designed, efficient surveillance networks.

Third, in Belgium as much as elsewhere, there have been continual debates over the sense and nonsense of NPIs, such as social distancing, face mask mandates, and curfews. Many opponents of NPIs argue that there is no randomized evidence for or against such measures and that there is no control group. This debate is similar to that in other areas of epidemiology, such as around the relationship between smoking and lung cancer. The problem is that gathering evidence about the impact of NPIs prior to their implementation and in roughly the same circumstances as present themselves in the field is virtually impossible. For example, there is evidence about the use of face masks from various angles: based on historic data in earlier pandemics, from routine use in clinical and industrial contexts, from dedicated laboratory experiments. But the evaluation of an NPI's impact arguably is possible only *after* implementing it. Even then, it is imperative to carefully correct for confounding factors. In addition, most NPIs are not taken in isolation, but rather in varying combinations (Liu *et al.*, 2021; Bo *et al.*, 2021).

5.2. Policy Advice and Communication

At first sight, policy advice towards government and administration on the one hand and general public communication on the other appear to be very different tasks. While they are, there is at the same time a strong connection.

A common factor is that, from a scientific and epidemiological standpoint, a pandemic like the current one is complex and multi-faceted. Many aspects are counterintuitive. We already referred to the difficulties with counterfactual thinking ("What if no NPIs had been implemented?") and the near impossibility to conduct controlled experiments, let alone in a timely fashion. In addition, statisticians will realize that outside their field and that of the exact sciences, exponential growth is poorly understood.

This implies that there is a constant need to warn against the risk of resurgence when NPIs are relaxed up to too low a stringency level. At a political level, it is important to have eye for public health, well-being, and the economy. While these are often seen as opposing forces, in fact a prudent management of the pandemic is the best way to ensure that simultaneously well-being and economy are optimally protected. Careful and frequent public opinion surveys are imperative in this regard (Neyens *et al.*, 2020; De Coninck *et al.*, 2021). Many countries have experienced the risks associated with relaxing NPIs too much or too quickly. Statisticians are well-positioned to gauge the impact of changing measures on human contact behavior and hence the spread of the virus, based on detailed data collection and analysis (Coletti *et al.*, 2020) as well as on insight in the general principles.

All of this implies that at the same time there is a great need for frequent communication on the one hand, while the political and societal context can be challenging, even tense at times. In Belgium, a relatively large number of scientists have taken the stage in conventional and social media to inform the public opinion. This includes scientists who also serve on advisory boards. This aspect alone requires a constant reflection on the equilibrium between academic freedom and loyalty as an advisor in a formal role. This is especially true when the public role provokes a strong response from certain factions within society, sometimes even taking the form of death threats. While it is not unnatural that a crisis of this magnitude and length leads to the search for scapegoats, it is a fate for the researcher uncommon in non-pandemic times.

6. Advice to Specific Sectors

Scientific government advisors, including statisticians, are often called upon to provide *ex officio* advice to certain sectors, such as the labor force, the cultural sector, travel, education, and higher education. We briefly expand on a few of these.

6.1. Travel

In the early phases of the epidemic, national borders quickly closed and (air) travel was reduced but did not stop. For example, Belgians undertook over 22 million trips in 2019, which reduced to 11 million in 2020. Brussels National Airport saw its volume of passengers reduce to roughly 25% over normal figures, over the pandemic months (excluding January and February 2020). For details, see <https://www.brusselsairport.be/fr/notre-aeroport/faits-chiffres/chiffres-de-traffic-mensuels>.

Over certain time periods, there were travel bans even within a country, or other limitations of movement, such as a perimeter. A controversial situation arose when also intra-EU borders ended up being closed. When travel resumed in June 2020, Belgium assumed an exceptional position by retaining a ban on travel to red zones, against the background of an ever changing color coded map. This

exception was lifted in the Fall of 2020, only to be replaced by a full travel ban on non-essential travel in the early days of January 2021, and until the middle of April 2021. This implied, for example, that only about 20,000 travelers entered Belgium in, say, a week in February 2021, whereas this number approached 400,000 in the last week of July 2021. As stated before, seeding by incoming travelers has played an important role throughout, but especially during the high tourist season. The role of travel restrictions has always been controversial, but as more and more data accumulated, from a phylogenetic as well as epidemiological nature, it became clear that travel bans and/or strict hygienic measures on travelers (testing, tracing, quarantine, isolation) play an important role in controlling the local epidemic. Such measures are particularly relevant when variants of concern emerge that are not yet present in the country. This situation arose in December 2020 and the following months, with the advent of the alpha variant, a little later with the beta and gamma variants, and then in May–July 2021 with the coming of the delta variant.

From an epidemiological standpoint, it is important to minimize as much as possible introduction from incoming travelers. From a statistical point of view, it is crucial to have good data on border crossing, including background data, such as available in the Passenger Locator Form, as well as test result data, vaccination status data, and contact tracing follow up data.

6.2. Education and Higher Education

As soon as the pandemic struck Belgium, schools and higher education institutions reverted to distance learning. Policy advice, based on modeling, kept schools almost fully closed throughout the remainder of the school year 2019–2020, with the exception of a very limited number of face-to-face teaching days in primary and secondary schools. In the higher education system, there were on-site exams in Flanders, but not in the French speaking universities.

The school and academic years 2020–2021 opened against the background of an expanding epidemic. Color codes were developed for all educational systems. The compulsory education system opened at full capacity, but with a face mask mandate in secondary schools. When the epidemic worsened, it was decided to supplement the regular one-week Autumn break with an additional holiday week, after which time the second and third grades of secondary schools reverted to hybrid education (essentially 50% on-site and 50% online). In February 2021, when the delta variant started spreading, extra-scholarly activities were limited, a situation that continued until Easter 2021. Once again, the Easter holiday was extended by an extra week. The face mask mandate was extended to the fifth and sixth year of primary school. Schools returned to full-time on-site presence in June 2021 only. While the school year 2020–2021 was unusual, there no longer were extended and extensive closures, unlike in the preceding school year. See also Bekkering *et al.* (2020).

Also higher education started the academic year in so-called code yellow, meaning that lecture halls and classrooms could receive 50% of the students, effectively implementing a hybrid system. As stated earlier, one of the driving forces of the Fall 2020 wave were student activities, and after only a few weeks, universities reverted to code orange, further reducing on-campus teaching to 20%. After the Autumn break, this became code red, where in-person teaching was no longer allowed. Essential practical sessions that could not be organized online were still allowed. Apart from exam sessions, this situation continued almost throughout the entire academic year, apart from some slight relaxations for on-campus presence and exam sessions.

Statistical and epidemiological advice and modeling has played a role in policy advice to the education sectors throughout the epidemic.

7. Concluding Remarks

Biostatistics, epidemiological, and mathematical modeling research during a global pandemic serves the role, apart from furthering scientific knowledge, of offering advice to the national, regional, and local governments, to specific sectors. Furthermore, it is used as a solid basis for communication to science-oriented as well as general and even popular media, and hence to the general public.

Especially when there are neither vaccines nor anti-viral medicinal products, society has to revert to NPIs, of which the main effect is to keep the epidemic curves under control. The side effects are undesirable but unavoidable impact on well-being and economy.

As a comprehensive understanding of the global pandemic and national epidemics is complicated, especially from a mathematical and statistical point-of-view, quantitative scientists are continually challenged regarding the proportionality or even mere necessity of NPIs. This is especially true at times when numbers are low and/or decreasing, and even when they are mildly increasing. The characteristics of SARS-CoV-2, especially the long incubation period and serial interval, the time lapse between infection and hospitalization, combined with a high basic reproduction number, often make it difficult for scientists, politicians, and society, including key players from affected sectors, to properly assess the timing and stringency of measures to be taken. While models and the associated scenario analysis can help in communicating the potential risks for the immediate, medium, and longer term perspective, they themselves are surrounded with uncertainty. While naturally understood by statisticians, the communication of uncertainty takes specific skill.

Arguably, for these reasons combined, biostatisticians and scientists in related fields need to combine scientific with didactic skills.

References

- Abrams S., J. Wambua, E. Santermans, L. Willem, E. Kuylen, P. Coletti, P. Libin, C. Faes, O. Petrof, S.A. Herzog, The SIMID COVID-19 team, P. Beutels, N. Hens (2020), "Modeling the early phase of the Belgian COVID-19 epidemic using a stochastic compartmental model and studying its implied future trajectories", *medRxiv*, <https://doi.org/10.1101/2020.06.29.20142851>.
- Aron J., C. Giattino, J. Muellbauer, and H. Ritchie (2020), "A pandemic primer on excess mortality statistics and their comparability across countries", <https://ourworldindata.org/covid-excess-mortality>.
- Bekkering T., N. Delvaux, P. Vankrunkelsven, J. Toelen, S. Aertgeerts, S. Crommen, P. De Bruyckere, I. Devisch, T. Lernout, K. Masschalck, N. Milissen, G. Molenberghs, A. Pascal, O. Plomteux, M. Raes, L. Rans, A. Seghers, L. Sweldens, J. Vandenbussche, G. Vanham, E. Wollants, and B. Aertgeerts (2021), "Closing schools for SARS-CoV-2: A pragmatic rapid recommendation", *BMJ Paediatrics Open*, vol. 5, n° 1, e000971, <https://bmjpaedsopen.bmj.com/content/5/1/e000971>.
- Bo Y., C. Guo, C. Lin, Y. Zeng, H. Bi Li, Y. Zhang, M.S. Hossain, J.W.M. Chan, D.W. Yeung, K. On Kwok, S.Y.S. Wong, A.K.H. Lau, and X. Qian Lao (2021), "Effectiveness of non-pharmaceutical interventions on COVID-19 transmission in 190 countries from 23 January to 13 April 2020", *International Journal of Infectious Diseases*, vol. 102, pp. 247-253.
- Coletti P., J. Wambua, A. Gimma, L. Willem, S. Vercruysse, B. Vanhoutte, C.I. Jarvis, K. Van Zandvoort, J. Edmunds, P. Beutels, and N. Hens (2020), "CoMix: comparing mixing patterns in the Belgian population during and after lockdown", *Scientific Reports*, vol. 10, n° 21885.
- De Coninck D., L. d'Haenens, G. Molenberghs, A. Declercq, E. Van Roie, C. Delecluse, and K. Matthijs (2022), "Perceptions and opinions on the COVID-19 pandemic in Flanders, Belgium: Data from a five-wave longitudinal study", *Data in Brief*, vol. 42, n° 108010.

- Delporte M., M. Luyts, G. Molenberghs, G. Verbeke, S. Demarest, and V. Hoorens (2021), "Do optimism and moralization predict vaccination? A 5-wave longitudinal study", *Submitted for publication*.
- Hale T., N. Angrist, E. Cameron-Blake, L. Hallas, B. Kira, S. Majumdar, A. Petherick, T. Phillips, H. Tatlow, and S. Wester (2020), "Variation in government response to COVID-19", University of Oxford Blavatnik School of Government Working Paper BSG-WP-2020/032, Version 7.0.
- Hens N., P. Vranckx, and G. Molenberghs (2020), "The COVID-19 epidemic, its mortality, and the role of non-pharmaceutical interventions", *European Heart Journal: Acute Cardiovascular Care*, vol. 9, pp. 204-208.
- Hodcroft E.B., M. Zuber, S. Nadeau, I. Comas, F. González Candelas, SeqCOVID-SPAIN consortium, T. Stadler, and R. Neher (2020), "Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020", *medRxiv*, <https://www.medrxiv.org/content/10.1101/2020.10.25.202190>.
- Liu Y., C. Morgenstern, J. Kelly, R. Lowe, CMMID COVID-19 Working Group, and M. Jit (2021), "The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories", *BMC Medicine*, vol. 19, p. 40.
- Molenberghs G., M. Buyse, S. Abrams, N. Hens, P. Beutels, C. Faes, G. Verbeke, P. Van Damme, H. Goossens, T. Neyens, S. Herzog, H. Theeten, K. Pepermans, A. Alonso Abad, I. Van Keilegom, N. Speybroek, C. Legrand, S. De Buyser, and F. Hulstaert (2020), "Infectious diseases epidemiology, quantitative methodology, and clinical research in the midst of the COVID-19 pandemic: A Belgian perspective", *Controlled Clinical Trials*, vol. 99, pp. 106-189.
- Molenberghs G., C. Faes, J. Verbeek, P. Deboosere, S. Abrams, L. Willem, J. Aerts, H. Theeten, B. Devleeschauwer, N. Bustos Sierra, F. Renard, S. Herzog, P. Lusyne, J. Van der Heyden, H. Van Oyen, P. Van Damme, and N. Hens (2022), "COVID-19 mortality, excess mortality, deaths per million and infection fatality ratio, Belgium, 9 March 2020 to 28 June 2020", *Eurosurveillance*, vol. 27, n° 7, 2002060.
- Natalia Y.A., C. Faes, T. Neyens, and G. Molenberghs (2022), "The COVID-19 wave in Belgium during the Fall of 2020 and its association with higher education", *PLoS One*, vol. 17, n° 2, e0264516, <https://doi.org/10.1371/journal.pone.0264516>.
- Neyens T., C. Faes, M. Vranckx, K. Pepermans, N. Hens, P. Van Damme, G. Molenberghs, J. Aerts, and P. Beutels (2020), "Can COVID-19 symptoms as reported in a large-scale online survey be used to optimise spatial predictions of COVID-19 incidence risk in Belgium?", *Spatial and Spatio-temporal Epidemiology*, vol.35, 100379.
- Verbeek J., G. Vandersmissen, J. Peeters, S. Klamer, S. Hancart, T. Lernout, M. Dewatripont, L. Godderis, and G. Molenberghs (2021), "Confirmed COVID-19 cases per economic activity during Autumn wave in Belgium", *International Journal of Environmental Research and Public Health*, vol. 18, n° 23, 12489.
- Vranckx M., C. Faes, G. Molenberghs, N. Hens, P. Beutels, P. Van Damme, J. Aerts, O. Petrof, K. Pepermans, and T. Neyens (2022), "A spatial model to jointly analyse self-reported survey data of COVID-19 symptoms and official COVID-19 incidence data", *Biometrical Journal*, vol.11, 10.1002/bimj.202100186.

La politique des grands nombres et les grands nombres de la politique dans la pandémie au Brésil



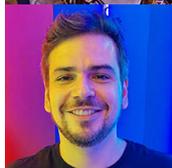
Eugênia MOTTA¹

Professeure à l'Université de l'État de Rio de Janeiro et chercheuse au Musée National, Université Fédérale de Rio de Janeiro



Victor ALVES MOURÃO²

Professeur à l'Université Fédérale de Viçosa



Alexandre DE PAIVA RIO CAMARGO³

Professeur à l'Université Candido Mendes et chercheur à l'Institut Universitaire de Recherche de Rio de Janeiro

TITLE

The politics of large numbers and the large numbers of politics during the pandemic

RÉSUMÉ

Cet article présente les résultats d'une recherche en cours basée sur les statistiques du covid-19 pendant les 18 premiers mois de pandémie au Brésil (de mars 2020 à octobre 2021). Nous avons étudié les relations entre les nombres produits à propos de la maladie et la pandémie elle-même en tant qu'expérience collective. Au cours de l'analyse, nous privilégions trois aspects : l'esthétique, les temporalités et les controverses. Le découpage chronologique proposé considère la forme sous laquelle les chiffres sont présentés et participent à la compréhension de la pandémie, en soulignant l'existence de quatre moments-clés séparant des périodes aux caractéristiques différentes. La politisation des nombres est l'un des aspects mis en évidence, aussi bien dans les rivalités politiques nationales que dans le processus d'attribution d'un sens à la pandémie. À partir du cas brésilien, nous revisitons les expressions « politique des nombres » et « grands nombres » proposées par Alain Desrosières. Au Brésil, les statistiques jouent aujourd'hui un rôle essentiel en tant que langage central du débat public. Ce rôle dépend notamment de la grandeur que les nombres doivent atteindre pour devenir politiquement pertinents.

Mots-clés : Covid-19, quantification, statistique, politique des nombres.

ABSTRACT

This paper presents the results of an ongoing research based on covid-19 statistics during the first 18 months of pandemic in Brazil (March 2020 to October 2021). We studied the relationships between the numbers produced about the disease and the pandemic itself as a collective experience. During the analysis, we focus on three aspects: aesthetics, temporalities and controversies. The proposed chronological breakdown considers the form

1. motta.eugenia@gmail.com

2. vmourao@ufv.br

3. alexandre.camargo.2009@gmail.com

in which numbers are presented and participate in the understanding of the pandemic, highlighting the existence of four key moments separating periods with different characteristics. The politicization of numbers is one of the aspects highlighted, both in national political rivalries and in the process of attributing meaning to the pandemic. From the Brazilian case, we revisit the expressions “politics of numbers” and “large numbers” proposed by Alain Desrosières. In Brazil, statistics now play an essential role as the central language of public debate. This role depends, among other things, on the magnitude that numbers must reach to become politically relevant.

Keywords: *Covid-19, quantification, statistics, politics of numbers.*

1. Introduction

Cet article propose d'utiliser les statistiques du covid-19 comme un outil de construction sociale de cette pandémie au Brésil. Nous soutenons la thèse que les différentes variations dans la production, la présentation et l'usage de ces « grands » nombres au cours des 18 premiers mois de la pandémie peuvent servir de laboratoire pour la compréhension des moments historiques qui ont vu s'articuler très étroitement les crises sanitaire et politique⁴.

Si nous vivons dans un monde où les nombres forment un langage central, l'énorme production de chiffres, graphiques, tableaux et classements pendant la pandémie s'explique par la multiplication des formes, acteurs et techniques de production et de la circulation des statistiques, encouragée par les technologies de communication. Cependant, l'actuelle crise sanitaire a ses particularités, s'agissant d'une épidémie aux proportions sans précédent. Dans la suite, nous abordons ces crises au sens de Koselleck (2006), comme une inflexion significative d'éléments déjà existants dans la production d'un nouveau temps historique.

Les deux crises, sanitaire et économique, que connaît le Brésil se sont abattues avec une intensité variable dans toutes les contrées du monde. Toutefois, le pays fait aussi face, depuis l'élection de Jair Bolsonaro, à une crise politique. Investi président en 2019, Bolsonaro met en œuvre des politiques qui affaiblissent délibérément les institutions de recherche, marquées par une brutale réduction de l'investissement dans ce domaine. Il souhaite également réduire la portée du prochain recensement, ce qui, selon les spécialistes, compromet la qualité des futures données. Et pourtant, le Brésil possède une tradition d'excellence en ce domaine. Tout cela s'accompagne d'un discours de méfiance envers les institutions et les experts, ce qui a conduit le président Bolsonaro à nier la gravité de la pandémie, à combattre les mesures de restriction de la circulation du virus et à contester l'efficacité des vaccins.

Le débat public et idéologique autour des experts, des universités et des institutions publiques qui produisent ces chiffres – déjà central depuis la campagne électorale de 2018 – l'est encore plus depuis l'éclatement de la pandémie. Comme l'indiquent les travaux de sciences sociales sur la quantification, l'examen de ces données, en constantes transformations, nous aide à mieux saisir des processus sociaux plus vastes (Hacking, 1990 ; Porter, 1995 ; Desrosières, 2008 ; Espeland et Stevens, 2008 ; Diaz-Bone et Didier, 2016 ; Camargo et Daniel, 2021). Au Brésil, sous la présidence de Bolsonaro et plus encore pendant la pandémie, c'est sur le terrain des statistiques que de nombreux conflits s'expriment. Plus que jamais, elles forment un laboratoire à partir duquel il est opportun d'observer et d'interpréter la réalité brésilienne.

Dans ce texte, nous proposons de nous pencher sur les statistiques du covid-19 pendant les 18 premiers mois de pandémie, de les discuter et de les mobiliser. Nous identifions quatre moments-clés qui séparent les périodes pendant lesquelles se présentent et s'articulent de manière distincte trois dimensions : l'esthétique, les temporalités et les controverses. Nous montrons comment les graphiques, les figures et les illustrations chiffrées construisent une esthétique de la pandémie (Motta, 2020) et comment s'y manifestent les notions d'avenir, de présent et de passé. Nous cartographions les principales controverses, ainsi que leur association à des conflits politiques plus larges. De cette façon, nous examinons comment la « politique des grands nombres » fonde les nouvelles modalités d'une critique de la réalité (Boltanski, 2014), pour enfin se transformer en « grands nombres de la politique ».

4. Ce texte est une version fortement révisée et augmentée d'une analyse déjà publiée en portugais (Camargo, Motta et Mourão, 2021) ; il fait partie du projet de recherche *Gouverner l'incertitude : pratiques et processus de quantification comme clé de lecture de la pandémie de covid-19*, toujours en cours.

Le découpage chronologique que nous proposons souligne les transformations qui ont eu lieu pendant la pandémie, ce qui permet une présentation plus systématique de nos résultats. Comme les lecteurs pourront s'en apercevoir, ce découpage ne constitue en aucun cas une division rigide ou ne suggère une rupture radicale entre les périodes choisies. Du point de vue méthodologique, notre analyse se fonde sur une compilation de reportages, images, graphiques, tableaux, tweets et autres productions médiatiques qui traitent les nombres de la pandémie. Ce corpus a été rassemblé au cours de la période allant de mars 2020 à octobre 2021.

Cette compilation concerne essentiellement l'expérience brésilienne de la pandémie, même si l'on a aussi pris en compte des matériaux internationaux qui nous semblent pertinents. Les données statistiques utilisées sont assez variées. On y trouve des bulletins d'hôpital indiquant le nombre de personnes contaminées, de décès et de lits occupés ; des enquêtes par sondage sur l'emploi et le revenu, effectuées (par téléphone) par des institutions régionales et fédérales ; des estimations issues de modèles épidémiologiques, qui discutent (et anticipent) plusieurs scénarios de lutte contre le virus ; des compilations de données réalisées par des agences publiques et des ONG pour estimer l'impact inégal de la pandémie sur des populations considérées comme vulnérables. En plus de ces productions déjà consolidées, nous avons examiné trois autres types de données chiffrées apparus pendant la pandémie et qui ont fourni un cadre à notre analyse de cette dernière : des « moyennes mobiles », calculées par de grands groupes de médias pour lisser les oscillations quotidiennes des cas positifs, des décès et du taux d'occupation des hôpitaux (sur 7, 15 et 30 jours) ; des taux de vaccination, qui permettent de comparer les performances de différents pays dans un espace mondial de gestion de la pandémie ; des estimations destinées à mesurer la responsabilité politique des gouvernements qui ont failli dans la maîtrise de la transmission.

En examinant les « nombres de la pandémie » à partir des trois critères mentionnés – esthétique, temporalités et controverses – nous suggérons que, depuis le début de la pandémie, quatre périodes se sont succédé, chacune marquée par une certaine circulation de données chiffrées et par des processus particuliers de quantification. La première période se distingue par la prolifération des modèles de prévision de la contagion, révélatrice de la perplexité initiale et qui constituait une tentative collective d'imaginer l'avenir. Par la suite, comme le nombre de malades et de décès n'a cessé d'augmenter, il s'agissait de savoir ce qu'il se passait. Le présent a donc remplacé l'avenir. Les experts et les agents publics se sont mobilisés sur l'adéquation des données à la réalité des décès et de la contagion. À l'occasion de la troisième période, marquée par le début de la vaccination, surgit l'espoir de surmonter la crise et s'opère une inflexion cognitive. Cependant les « bons nombres » de l'immunisation se mirent à représenter l'inégalité entre les pays : la pandémie n'était plus le destin commun de l'humanité. La quatrième est caractérisée par l'installation d'une commission d'enquête sénatoriale pour évaluer les actions du gouvernement, ce qui a introduit dans le débat les notions de « responsabilité » et de « culpabilité », tout en inaugurant une temporalité motivée par l'imagination de ce qui se serait passé si les décisions prises les derniers mois avaient été différentes.

2. Première période : des courbes, des modèles et un avenir à imaginer

Le début de l'épidémie de covid-19 – déclarée « pandémie » par l'Organisation mondiale de la santé, le 11 mars 2020 – est marqué par l'incertitude : aussi bien les individus que les États et organisations internationales s'inquiètent de cette maladie encore inconnue. À ce premier moment de perplexité et même d'effroi face à une possible catastrophe correspondent de multiples efforts de prédiction qui mobilisent aussi bien des spécialistes que des non-professionnels. Ils se caractérisent par la circulation de graphiques qui défendent le nécessaire « aplatissement de la courbe » – en anglais, *flattening the curve*. Ceux-ci deviennent un moyen

privilegié du débat sur la pandémie.

Au Brésil, plusieurs versions de ce graphique circulent, certaines purement illustratives, dépourvues de précision technique. Il s'agit, grosso modo, de deux courbes en cloche, dont les origines coïncident avec le début de l'épidémie (voir la figure 1). L'axe vertical correspond au nombre de malades, l'axe horizontal à celui du temps. La courbe qui enregistre la plus forte hausse est aussi celle qui baisse le plus vite, et inversement, l'autre courbe dilue cette hausse dans le temps. Les surfaces sous chaque courbe sont de la même taille, et correspondent au total de malades. Toutefois, elles se répartissent différemment dans le temps. On ajoute à l'image une droite horizontale qui représente la capacité fixe (ou légèrement croissante) d'accueil des systèmes de santé. La courbe la plus accentuée dépasse la ligne horizontale, tandis que la moins accentuée reste au-dessous. Ce graphique sert à représenter deux scénarios : dans le premier, l'augmentation rapide du nombre de malades excède la capacité des hôpitaux à les accueillir ; dans le deuxième, cette hausse reste contenue, mais dure plus longtemps. Ce graphique suggère donc qu'il faut « aplatir la courbe », c'est-à-dire réduire le rythme de la contagion, pour ne pas épuiser les services de santé.



Figure 1 – Graphique représentant un modèle épidémiologique
 (Source : Centers for Disease Control and Prevention, EUA (2007).
 Disponible sur : <https://stacks.cdc.gov/view/cdc/11425>)

Cette image accompagne la diffusion d'un ensemble de chiffres et de graphiques issus de modèles prédictifs⁵. À partir de ces données, surtout du taux de contagion des différents pays et des grandes régions et du monde, ces modèles indiquent plusieurs scénarios. Dans les graphiques qui illustrent « l'aplatissement de la courbe », les deux évolutions représentées correspondent à un choix : l'adoption ou non de politiques dites de « mesures non pharmacologiques » ou de « distanciation sociale ». Si de tels modèles exposent différents possibles et montrent que l'avenir n'est pas certain, ils permettent aussi d'imaginer *un* avenir, sur lequel aussi bien les individus que les gouvernements peuvent avoir une influence. Ces modèles sont donc utiles à la construction du monde sur lequel on veut agir, en tentant de limiter l'incertitude à quelques variables et de soumettre des forces incontrôlables (le virus et sa diffusion) à la volonté humaine.

5. Des chercheurs du *Data-driven Innovation Lab* de la *Singapore University of Technology & Design* ont produit, en avril 2020, une série de modèles prédictifs sur la fin de la pandémie dans plusieurs pays (Luo, 2020). En ce qui concerne le cas brésilien, les chercheurs ont estimé la fin de la pandémie au mois d'août de la même année. Les adresses électroniques des graphiques et l'article original ne sont plus disponibles (<https://ddi.sutd.edu.sg/when-will-covid-19-end/> et https://persi.or.id/images/2020/data/covid19_prediction_paper.pdf). Le chercheur responsable a publié un article sur les problèmes de prévision (Luo, 2021).

La temporalité sous-jacente à ces modèles construit une épidémie en quelque sorte prévisible, manipulable et homogène, qui s'intensifierait puis se relâcherait. Cette forme d'imagination, qui s'est avérée très optimiste, s'oppose à la peur de l'apocalypse.

Ces graphiques servent d'argument aux partisans des mesures préventives comme le fameux « restez chez vous ! » (« *fique em casa !* »). Ainsi, non seulement il serait possible d'aplatir la courbe, mais ceci se ferait grâce à l'action responsable de chaque individu. D'abord considérées comme modèles prédictifs, de telles courbes se sont converties en prescriptions, et cette transformation s'est produite entre les laboratoires spécialisés et le quotidien des individus. La quantification produit des effets similaires à ceux des actes de langage, qui changent le réel en le disant, le nommant ou le classant. Tout comme dans le langage, les effets de la quantification se manifestent à travers leurs usages (Espeland et Stevens, 2008, pp.404-405).

Tout au long de ce processus, ces graphiques ont perdu leurs échelles, en se détachant progressivement des débats techniques concernant l'acuité des modèles, et ont reçu des couleurs, des représentations de personnages ou se sont simplifiés au point de se muer en symboles (voir la figure 2). Cette transformation se rapproche des observations de Benedict Anderson (2006) sur les « cartes-logo » (« *maps-as-logo* »), ces images de périmètres géographiques qui servent à représenter une nation et une nationalité et détiennent un important pouvoir de mobilisation. Dans le graphique qui nous intéresse, deux courbes ont suffi pour communiquer la nécessité d'appliquer des mesures préventives. Elles ont d'ailleurs été reprises dans des illustrations en tout genre.

On est alors au début de la pandémie et le nombre de décès, même amplement diffusé, n'a pas encore atteint le niveau qui mobilisera la population et capturera l'imagination collective. Néanmoins, une série d'estimations occupe le débat public, avec pour objectif de limiter l'amplitude de la pandémie. À la mi-mars, nous nous retrouvons – pour le cas brésilien qui nous occupe – face à des prévisions qui débouchent sur des trajectoires assez différentes. L'Imperial College London estime, en l'absence de mesures qui limiteraient la propagation de la maladie, à 1,1 million le nombre de décès et, dans un scénario de distanciation sociale de toute la population, à 620 000⁶. Le député fédéral Osmar Terra, médecin et ancien ministre de la Citoyenneté du gouvernement Bolsonaro, rédige un tweet qui deviendra « fameux », affirmant que le coronavirus ne causera jamais plus de deux décès par jour au Brésil⁷. Le président fait ensuite des déclarations publiques en accord avec la prévision de son ministre⁸.

6. Cf. « Saiu a modelagem estatística do Imperial College London para os cenários do COVID-19 no Brasil ». *Portal da Associação Brasileira de Enfermagem*, le 29 mars 2020. Disponible sur : <http://www.abennacional.org.br/site/2020/03/29/saiu-a-modelagem-estatistica-do-imperial-college-london-para-os-cenarios-do-covid-19-no-brasil/>

7. Cf. Tweet du 18 mars 2020. Profil de Osmar Terra. Disponible sur : <https://twitter.com/OsmarTerra/status/1240232194259537920>

8. Cf. « Previsão é que coronavírus tenha menos mortes que a H1N1, diz Bolsonaro ». *Agência Estado*, le 22 mars 2020. Disponible sur : <https://noticias.uol.com.br/ultimas-noticias/agencia-estado/2020/03/22/previsao-e-que-coronavirus-tenha-menos-mortes-que-a-h1n1-diz-bolsonaro.htm>

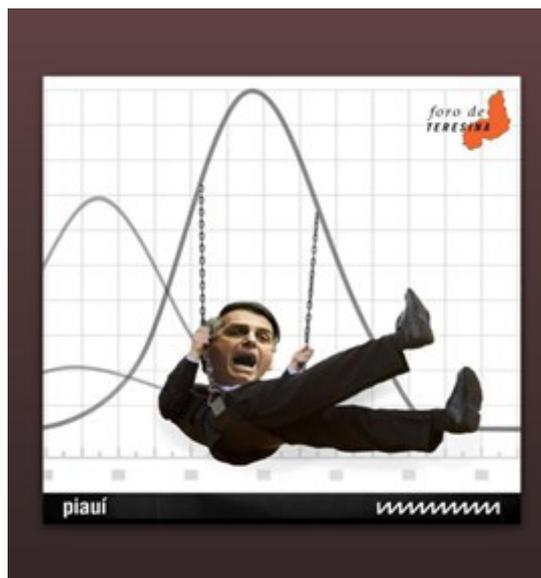


Figure 2 – Image publiée sur le compte Instagram de la Revista Piauí (Magazine Piauí) le 30 avril 2020

Très rapidement, on se met à douter des graphiques qui renvoient à l'objectif d'aplatir la courbe. Comme très souvent dans le cas des nombres rendus publics, des contestations et des résistances participent à leur constitution. Le premier registre de la critique allègue l'insuffisance des données pour décider si les mesures de restriction doivent ou non être appliquées. Il faut aussi prendre en compte « l'économie ». Des graphiques circulent qui ajoutent une courbe descendante décrivant son évolution, aux variations inversement proportionnelles à celles des cas de covid-19. L'activité économique serait pénalisée par la fermeture des commerces et l'interdiction de la libre circulation. La tension entre « économie » et « vie », qui a traversé toute la pandémie, relève à ce moment d'un choix dichotomique : préserver la vie ou s'occuper de l'économie. Ces deux pôles sont présentés comme des valeurs absolues et mutuellement exclusives et en viennent à incarner des positions politiques inconciliables. Comme nous allons le voir, la relativisation des composantes de cette dichotomie n'a eu lieu que plus tard, rendant possibles de nouvelles articulations.

3. Deuxième période : le présent, l'empirisme et la froideur des nombres

Il est impossible de poser un jalon, que ce soit un événement, une déclaration verbale ou un document, qui définisse précisément ce que nous entendons par « deuxième moment-clé de la pandémie ». Il concerne la collecte, la systématisation, la distribution et l'interprétation de données démographiques et épidémiologiques sur les événements présents, et non en la construction de scénarios. Les principales déclarations et polémiques proviennent de ces nombres et les modèles occupent une place secondaire.

Le début de la pandémie, avec la déclaration de l'OMS, entraîne l'adoption de nouvelles réglementations nationales (les circulaires 188 et 356 du ministère de la Santé, la loi 13 979/2020). Celles-ci produiront les nombres qui nous accompagneront les mois suivants. Désormais, d'innombrables bulletins épidémiologiques émis par les communes, les États et la fédération surveillent la propagation de l'épidémie. Leurs données indiquent le nombre de cas positifs et de décès, ainsi que la disponibilité en lits d'hôpitaux et de réanimation. D'autres informations – comme le nombre de cas suspects et les types de tests utilisés – et d'autres graphiques accompagnent ces documents, qui deviennent d'importants repères (voir la figure

3). Ceux-ci caractérisent une nouvelle période extraordinaire et produisent la perception d'une dynamique nationale de la pandémie. Les écarts régionaux et les spécificités métropolitaines ressortent : la maladie s'est propagée à partir des capitales, dotées d'aéroports internationaux, et s'est ensuite diffusée au reste du pays par les flux internes des déplacements individuels (Guimarães *et al.*, 2020). Cette production et cette consommation fédérative – décentralisées en l'occurrence – des nombres de la pandémie contribuent grandement à assurer la confiance dans les statistiques, alors que, comme nous allons le voir, une partie de l'opinion publique a douté des données publiées par le gouvernement fédéral.

La poursuite de la pandémie a conduit le public à s'immerger dans les données produites pendant cette période, qui ont été lues et interprétées par rapport à d'autres informations numériques, plus générales, reproduisant cette capture de l'imagination collective qui avait eu lieu au premier moment de la pandémie. Les nombres perdent leur caractère prédictif pour assumer une ambition plus empirique ; cette ambition, comme nous allons le voir, a suscité de nouvelles controverses.

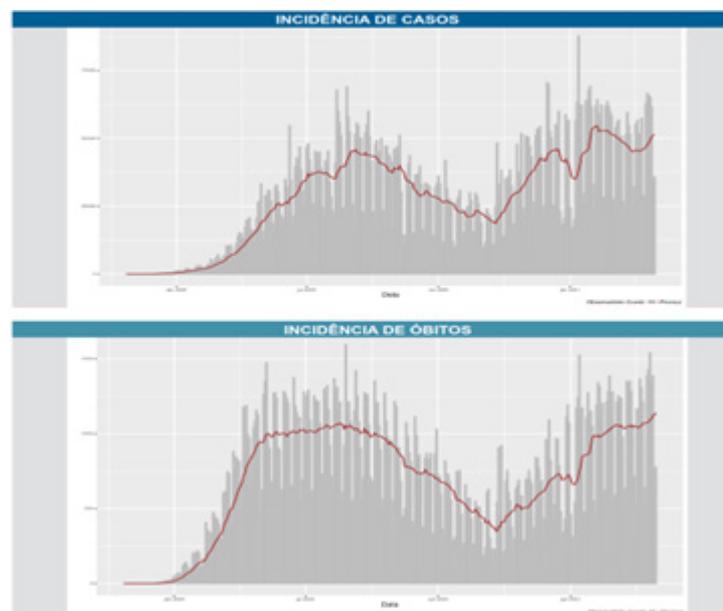


Figure 3 – Diagrammes représentant le nombre de malades et de décès du covid-19, en territoire brésilien, au cours de la pandémie (mars 2020/mars 2021), avec une courbe de moyennes mobiles (Source : Boletim Extraordinário do Observatório covid-19 da Fiocruz (02/03/2021).

Disponible sur :

https://agencia.fiocruz.br/sites/agencia.fiocruz.br/files/u35/boletim_extraordinario_2021-marco-03.pdf

L'interprétation de la pandémie en fonction de « vagues » reste implicite et les paramètres des modèles épidémiologiques demeurent disponibles. Cependant ces analyses se basent de plus en plus sur des tendances observées empiriquement et sur la comparaison entre pays et régions. Elles mobilisent des méthodes de comparaison des différentes expériences locales, régionales et nationales. Au Brésil, les principaux médias et les experts en santé publique encensent des exemples de succès international comme l'Allemagne et la Nouvelle-Zélande. Ils mobilisent les données fournies par ces deux pays par contraste avec l'expérience brésilienne. Celle-ci fait l'objet d'une évaluation basée sur le refus du diagnostic sanitaire et des mesures de restriction de circulation, caractéristique de ces pays qui connaissent des niveaux élevés de contamination et de décès, dont les leaders sont taxés de négationnistes⁹. Toutefois, le

9. Cf. *Covid Performance Index*, Lowy Institute : <https://interactives.lowyinstitute.org/features/covid-performance/>.

gouvernement brésilien – contrairement à ceux d'autres pays comme l'Angleterre et même les États-Unis – résiste particulièrement à changer d'avis et s'oppose aux mesures de restriction adoptées par les États et les communes jusqu'en mai 2021.

Ce nouveau moment-clé ajoute des dimensions sociales, économiques et politiques au premier évoqué ci-dessus. On observe des regroupements de données et la construction de critères et d'indicateurs indépendants concernant les Amérindiens¹⁰, les habitants des favelas¹¹, la population carcérale¹², en fonction de dimensions comme le métier, la race¹³, le revenu¹⁴, le genre et la grossesse (Freitas, 2020 ; Holanda, 2020 ; Mello, 2020 ; Takemoto *et al.*, 2021). Cette production de données et nombres pandémiques s'associe à des formes de stigmatisation (Bruno, Didier et Prévieux, 2014) qui rendent visibles ces groupes sociaux affectés différemment par la pandémie et soutiennent les revendications favorables aux mesures de contention de la contagion. Soulignons cet usage militant des nombres épidémiologiques par les populations indigènes qui s'en servent pour mettre en évidence le caractère intersectionnel de leur vulnérabilité sanitaire. Pour faire valoir ses revendications, *l'Articulation des peuples indigènes du Brésil (Articulação dos Povos Indígenas do Brasil ou APIB)* a recouru à la justice qui s'est prononcée en sa faveur. Pour autant, le gouvernement fédéral n'a pas fait appliquer ces décisions (Santos *et al.*, 2020).

En parallèle aux chiffres de la contagion et des décès, d'autres nombres circulent. La construction d'indicateurs à partir de bases de données privées – comme le taux d'isolement social par le traçage de dispositifs mobiles – sert à mesurer l'adhésion de la population aux mesures de restriction¹⁵. Plus récemment, des recherches par questionnaires auprès des usagers de réseaux sociaux, ont recueilli des informations sur leurs symptômes ; elles affirment pouvoir prédire, avec plusieurs semaines d'avance par rapport aux données officielles, les nouvelles vagues de contagion¹⁶. Ces recherches rappellent la prétention prédictive des premières modélisations, mais elles innovent dans leur méthodologie, en raison de la façon dont elles approchent les phénomènes collectifs : il ne s'agit plus d'agréger les résultats de tests positifs, mais d'organiser un corpus de récits de symptômes.

Le recours à ces nombres produits par et pour la pandémie, en tant que problème public, ritualise le deuil. Il leur fait ainsi une place dans la réponse émotionnelle collective à l'ampleur des décès. Ce travail symbolique et pratico-rituel s'effectue par l'affichage de grands *nombres ronds* : mille cas positifs, dix mille décès, un million de cas, deux mille décès par semaine. De tels nombres marquent, au cours de cette période, le passage à une temporalité apocalyptique¹⁷. Dans leurs premières pages, les journaux restaurent la singularité de chaque vie perdue de manière esthétique aussi bien que discursive en opposition à la soi-disant froideur des nombres, mais y ont tout de même recours pour construire et assurer le suivi de ces opérations

10. Cf. *Site Emergência Indígena*. Disponible sur : https://emergenciaindigena.apiboficial.org/dados_covid19/ ; *Covid Fora das Aldeias*. Disponible sur : <https://www.icv.org.br/covidforadasaldeias/>

11. Cf. « Favela communities made more resilient against COVID-19 with citizen data ». Disponible sur : https://warwick.ac.uk/newsandevents/pressreleases/favela_communities_made/

12. Cf. « Registros de Contágios e Óbitos do Conselho Nacional de Justiça ». Disponible sur : <https://www.cnj.jus.br/sistema-carcerario/covid-19/registros-de-contagios-obitos/>.

13. Cf. « Ausência de dados raciais nacionais compromete combate à Covid-19 nas periferias », *Ponte*, le 15 mai 2020. Disponible sur : <https://ponte.org/ausencia-de-dados-raciais-nacionais-compromete-combate-a-covid-19-nas-periferias/>

14. Cf. « Mortalidade por Covid-19 é maior entre os mais pobres, aponta pesquisa », *Lais*, le 8 juin 2020. Disponible sur : <https://lais.huol.ufrn.br/mortalidade-por-covid-19-e-maior-entre-os-mais-pobres-aponta-pesquisa/>

15. Cf. « Tendência de isolamento social desacelera no Brasil », *Journal GGN*, le 13 avril 2020. Disponible sur : <https://jornalgggn.com.br/a-grande-criese/tendencia-de-isolamento-social-desacelera-no-brasil/> ;

cf. Site : *Community Mobility Records*. Disponible sur : <https://www.google.com/covid19/mobility/>.

16. Cf. « Pesquisa que rastreia redes sociais aponta que próxima onda da pandemia será mais grave no Brasil », *El País*, le 28 mai 2021. Disponible sur : <https://brasil.elpais.com/brasil/2021-05-28/pesquisa-que-rastreia-redes-sociais-ja-aponta-que-a-proxima-onda-da-pandemia-sera-ainda-mais-grave-que-as-anteriores-no-brasil.html>

17. Cf. « Brasil ultrapassa marca de 10 milhões de casos registrados de Covid; foram 1.432 mortes nas últimas 24 hora », *G1*, le 18 février 2021. Disponible sur : <https://g1.globo.com/bemestar/coronavirus/noticia/2021/02/18/brasil-ultrapassa-marca-de-10-milhoes-de-casos-registrados-de-covid-foram-1432-mortes-nas-ultimas-24-horas.ghtml>

symboliques. Cet usage des nombres est particulièrement évident sur les unes des quotidiens considérés dans la figure 4, qui affichent respectivement dix mille, cent mille et quatre cent mille décès.



Figure 4 – Couvertures du Journal O Globo, les 10 mai et 9 août 2020, et du Journal Extra, le 30 avril 2021.

L'expérience brésilienne de la pandémie a été marquée par un processus de politisation qui, dans cette deuxième période, implique directement les nombres. Ce processus culmine avec la création d'un consortium de médias et d'institutions publiques régionales qui veut recueillir et systématiser les nombres de la pandémie. En avril 2021, le Bureau spécial de communication sociale de la présidence (*Secretaria Especial de Comunicação Social da Presidência*) divulgue un graphique nommé « Score de la vie » qui insiste sur le nombre de « Brésiliens sauvés » et « en rétablissement »¹⁸, mais ne dit rien des décès. Ce même *Score de la vie* adoptera ensuite des définitions différentes, dont les taux et classements de la « proportion de décès par million d'habitants ». L'objectif est de diluer les nombres de décès dans la masse et de produire une image positive de la pandémie à l'échelle nationale. Le secrétaire exécutif en vient même à célébrer la marque des « un million de personnes guéries », ce qui ferait du Brésil « le recordman Mondial du nombre de rétablis »¹⁹.

Les tentatives du gouvernement fédéral de diminuer et de nier le nombre de morts – en mélangeant ces données avec d'autres qui sont perçues comme positives, ou en les occultant – se sont poursuivies les semaines suivantes. Au début du mois de juin 2021, les médias nationaux font régulièrement état des accusations de pressions du ministère sur la production des données officielles²⁰. Ces accusations sous-entendent que le gouvernement veut adopter un nouveau critère dans la divulgation quotidienne du nombre de décès, en comptant seulement ceux qui surviennent et sont confirmés le même jour, excluant ainsi les décès survenus auparavant mais confirmés à la date de divulgation. Ce changement de critère réduirait le nombre des décès à moins de 2000 par jour, en accord avec une requête du président²¹. Les chiffres ronds

18. Cf. Tweet du 30 avril 2020. Profil *Secretaria de Comunicação Social da Presidência*. Disponible sur : <https://twitter.com/secomvc/status/1255828872773488640>

19. Cf. Tweet du 9 juillet 2020. Profil *Ministério da Saúde*. Disponible sur : <https://twitter.com/minsaude/status/1281337365093584896>

20. Cf. « Cúpula da Saúde pressiona até Abin a maquiar dados », *Valor Econômico*, le 8 juin 2020. Disponible sur : <https://valor.globo.com/brasil/noticia/2020/06/08/cupula-da-saude-pressiona-ate-abin-a-maquiar-dados.ghtml>

21. Récemment, un changement du système de comptabilité a réduit, de manière radicale, le comptage de décès ; il a aussi été interprété comme une intervention du gouvernement fédéral dans les nombres de la pandémie. Cf. « Ministério da Saúde muda sistema e número de mortos por Covid-19 despenca artificialmente », *Folha de São Paulo*, le 24 mars 2021. Disponible sur : <https://www1.folha.uol.com.br/colunas/monicabergamo/2021/03/ministerio-da-saude-muda-sistema-e-numero-de-mortos-por-covid-19-despenca-artificialmente-em-sp.shtml>

sont donc importants non seulement pour la ritualisation du deuil et de l'empathie, mais aussi dans l'évaluation politique de la gestion sanitaire. Celle-ci devient la cible d'une contre-gestion statistique dont le but est de limiter les conséquences négatives sur l'image du gouvernement.

Le gouvernement cesse de communiquer sur les données agrégées et restreint l'accès à ses documents, dont les tableaux statistiques, pendant plusieurs jours. Allant de pair avec de telles modifications, il reporte la divulgation des données quotidiennes, qui avait traditionnellement lieu à 17 h, empêchant ainsi les journaux du soir de les mentionner²². Face à la crise de confiance potentielle envers les nombres produits par le gouvernement fédéral, un groupe d'acteurs publics et privés se mobilise pour occuper l'espace des données légitimes laissé vide, en fédérant leurs efforts pour permettre l'accès à l'information : les grands groupes de médias sollicitent désormais les données de la pandémie directement auprès de chaque État brésilien²³.

La presse se met à ignorer systématiquement les bulletins du gouvernement, rétablit l'horaire de diffusion et divulgue deux nombres concurrents mais similaires (c'est-à-dire non contradictoires) : la compilation des données réalisée par le consortium de médias et la statistique produite par le Conseil national des délégués régionaux à la santé (*Conselho Nacional de Secretários de Saúde*). Cette situation exceptionnelle dure encore aujourd'hui, rendant singulière l'expérience brésilienne. Même si les données officielles, certifiées par le pouvoir central, sont devenues délégitimées, les nombres gardent leur autorité, grâce à l'alliance inattendue d'une partie des pouvoirs publics et de la plupart des médias privés. Paradoxalement, le gouvernement fédéral cultive la méfiance sur la vraisemblance des statistiques qu'il produit, suscitant des controverses qui ne sont atténuées qu'avec la prise en charge de ces données par de nouveaux acteurs.

4. Troisième période : les vaccins et les « bons nombres »

Nous considérons que le début de la vaccination inaugure une nouvelle période pour l'analyse, marquée par l'apparition des « bons nombres », ceux qui donnent confiance en l'avenir et ouvrent la perspective de surmonter la crise. Cette période commence par une vraie « bataille des vaccins », avec la diffusion de leurs taux d'efficacité, mesurés et publiés tout au long des différentes phases de test. En même temps, la perception du présent continue d'évoluer avec l'épuisement des modèles prédictifs et l'abandon des mesures plus radicales d'isolement social. Les mouvements anti-confinement se renforcent ; ils revendiquent la fin des restrictions dans plusieurs pays d'Europe, ce qui tranche avec la capacité d'imposition dont bénéficiait le programme sanitaire des épidémiologistes au début de la pandémie.

La circulation des nombres se complexifie, en produisant une bifurcation. Les médias mettent quotidiennement à jour le nombre de vaccins comme s'il s'agissait d'un compte à rebours avant la fin de la pandémie. Le total de vaccins effectivement administrés – baptisés dans certaines régions de « vaccinomètre » – cohabite avec des projections et des prévisions, souvent dénigrées et continuellement revisitées, qui concernent maintenant le calendrier de vaccination – hebdomadaire, mensuel et annuel. Essentiellement positif et prédictif, ce rapport à l'avenir, basé sur la vaccination, est pétri d'incertitudes quant à sa mise en œuvre.

Dans leur ensemble, ces nombres contribuent à rendre toujours plus palpables les chiffres de la pandémie, une tendance de fond dans la perception de la pandémie depuis la deuxième période. À la moyenne mobile des décès et au taux d'occupation des lits en réanimation, qui quantifiaient déjà le rythme de la contagion, s'ajoute le nombre de vaccins et de vaccinés, ce qui

22. Cf. « Mudança de divulgação ocorreu após Bolsonaro exigir número de mortes abaixo de mil por dia », *Estadão*, le 8 juin 2020. Disponible sur : <https://saude.estadao.com.br/noticias/geral,mudanca-de-divulgacao-ocorreu-apos-bolsonaro-exigir-numero-de-mortes-abaixo-de-mil-por-dia,70003328525>

23. Cf. « Veículos de comunicação formam parceria por transparência em dados de covid », *UOL*, le 8 juin 2020. Disponible sur : <https://noticias.uol.com.br/saude/ultimas-noticias/redacao/2020/06/08/veiculos-de-comunicacao-formam-parceria-por-transparencia-a-dados-de-covid.htm>

réduit la distance entre ces chiffres et la vie quotidienne des gens. C'est peut-être pour cela, et en fonction de l'horizon ouvert par la vaccination, que la méfiance envers les statistiques s'avère maintenant beaucoup moins importante qu'aux premiers moments, quand le traitement précoce et non prouvé de la maladie semblait le seul remède contre le désespoir.

Ces nouveaux nombres révèlent aussi, très clairement, la différence entre les capacités des pays riches et pauvres à acquérir des vaccins et à les administrer, par-delà les postures idéologiques de leurs dirigeants. Il apparaît évident que les pays riches sont prioritaires dans les négociations avec les fabricants de vaccins ou dans l'achat des matières premières, pouvant même compter parfois sur la livraison anticipée de doses suffisantes pour vacciner plusieurs fois toute leur population.

Ces nombres, utilisés à des fins descriptives, nous mènent à un autre indicateur, dont l'efficacité performative nous intéresse de près. Prioritaires dans l'achat des vaccins, disposant de la logistique nécessaire au stockage et à la distribution des doses, un certain nombre de pays parmi les plus riches servent de cadre au surgissement d'un nouveau nombre : le taux d'efficacité des vaccins dans la contention de la contagion. L'expansion de la vaccination et l'accroissement des contrôles sur la transmission de la maladie génèrent surtout un nouveau consensus sur l'élimination du covid-19 qui prend d'abord forme dans les pays les plus riches. Différents experts, en grande partie des économistes, construisent des modèles qui impliquent une courbe à coût convexe, c'est-à-dire d'augmentation des coûts en l'absence d'éradication. De tels modèles soulignent que le coût des vies perdues est beaucoup plus élevé que celui des mesures les plus dures, ou que l'alternance, profondément critiquée, entre confinements et réouvertures²⁴. Moins d'un an après la guerre des nombres catastrophiques et des courbes descendantes, qui opposaient la protection de la vie à la performance de l'économie, la modélisation des économistes et des représentants du capital financier s'associe à celle de l'épidémiologie informatique.

Cette nouvelle stratégie de gestion de l'avenir engendre deux conséquences décisives. D'abord – comme l'a suggéré Simmel dans sa *Philosophie de l'argent* (2009) – depuis ses débuts, la science économique mesure la vie par sa valeur monétaire : le calcul du dédommagement des ménages d'une victime assassinée et celui de la valeur du travail humain en sont quelques exemples, qui ne créent aucun émoi. La métrique morale qui distribue les espérances de vie de manière inégale entre les classes sociales et les minorités raciales, ethniques et sexuelles ne choque pas non plus (Neiburg, 2020). Le conflit entre la vie et l'économie n'apparaît que lorsque l'on prend la première au singulier et dans son universalité, ce qui empêche de lui attribuer un prix déterminé. La troisième période de la pandémie se distingue par le fait qu'elle améliore l'échelle de la convention d'équivalence entre la valeur de la vie et le prix qu'on lui attribue, convention déjà courante dans les pratiques sociales et économiques. Bien que les vies ôtées perdent de leur singularité, en contrepartie, la société reconnaît davantage le coût de la mort, ce qui met en évidence la trame sociologique complexe des équivalences de valeurs, relatives à différents régimes de justification (Boltanski et Thévenot, 1991).

La métrique qui rapproche les pertes humaines d'une perte économique se consolide très vite entre les deuxième et troisième périodes. On en a un exemple dans la quantification de la responsabilité politique de Donald Trump et de Jair Bolsonaro dans la gestion de la pandémie. Plusieurs études ont calculé, avec plus ou moins de rigueur, l'impact du refus de l'ancien président nord-américain d'adopter des mesures de restrictions plus sévères. Nous mettons l'accent sur le rapport de la revue *Lancet*²⁵, en raison de sa vaste diffusion et aussi de son

24. Cf. « Has Covid changed the price of a life? », *The Guardian*, le 15 février 2021. Disponible sur : <https://www.theguardian.com/world/2021/feb/14/coronavirus-covid-19-cost-price-life>

25. Il s'agit du rapport « Public policy and health in the Trump era », disponible sur : [https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736\(18\)32171-8.pdf](https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(18)32171-8.pdf)

importance, s’agissant de l’une des revues de sciences biomédicales les plus anciennes et prestigieuses au monde. Évoquant la moyenne pondérée des décès causés par le covid-19 parmi les pays membres du G7, le rapport souligne que 40 % de ces morts auraient pu être évités, si « un plan national avait été adopté », si « on avait favorisé la coopération internationale », si « on avait testé en masse », si « on avait laissé les écoles fermées » et si « les attroupements sans port de masque n’avaient pas été stimulés » (Woolhandler *et al.*, 11 févr. 2021, p. 7). En gros, si les gouvernements avaient appliqué les mesures prônées par les modèles prédictifs de l’épidémiologie – et, maintenant aussi, de l’économie.

Il est important d’observer que de tels essais de quantification de la responsabilité politique, ainsi que d’autres, assez semblables, ne reposent toujours pas sur une méthode consensuelle et sur une valeur de référence commune. Quelle doit être la courbe « normale » : la moyenne du G7, la moyenne régionale, la moyenne mondiale ou une autre ? Comment mesurer des aspects intangibles, comme l’image et le discours des présidents qui dédaignent des protocoles d’hygiène, et leur impact sur la population ? De telles questions sont rarement formulées par les experts ainsi que par les journalistes qui reprennent ces nombres, face à la stratégie, pleinement justifiée, de mobiliser l’opinion internationale et de dénoncer des omissions que d’aucuns qualifient de crime contre l’humanité. Cet usage des nombres tranche par rapport à celui fortement biopolitique des première et troisième périodes, qui accompagnaient la normalisation de l’exception, la suppression de libertés et le renforcement du contrôle de l’État. En échange, on assiste à la définition d’une modalité pandémique de stactivisme, à une forme d’activisme qui s’en remet aux statistiques pour critiquer et s’opposer aux pouvoirs institutionnalisés et ainsi promouvoir l’équité sociale (Bruno, Didier et Prévieux, 2014, p. 5-30).

Il y a une autre conséquence qui résulte de la convergence entre les modélisations des économistes et les prévisions de l’épidémiologie. Si, d’une part, les « bons nombres » du vaccin dévoilent un horizon de dépassement de la crise, ils fournissent, d’autre part, les paramètres pour une nouvelle comparaison systématique entre les pays, donnant une forme institutionnelle à l’espace statistique de la gestion mondiale de la pandémie. La course à la vaccination devient elle-même un indicateur de performance de chaque gouvernement, qui sert aussi bien la propagande que l’opposition politique, comme dans les cas extrêmes d’Israël et du Brésil, respectivement. De manière plus subtile, ce nouveau *benchmarking* (Bruno et Didier, 2013) transfère aux pays les moins performants la responsabilité de devoir cohabiter avec le virus, tandis que se rapproche la perspective d’une vaccination totale de la population parmi les pays les plus riches.

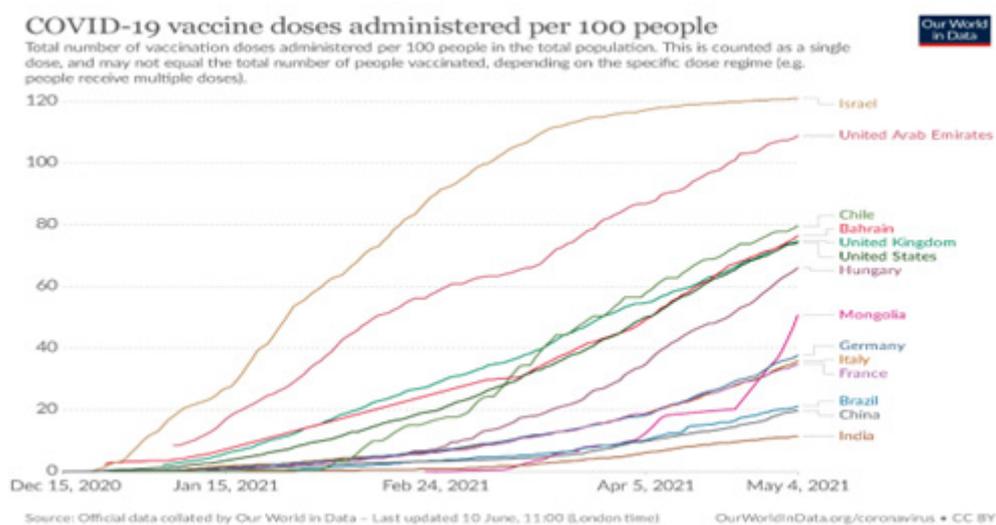


Figure 5 – La course aux vaccins (Source : Our World in Data, 10 juin 2021)

En conséquence de ces classements mis à jour quotidiennement, l'image d'un risque global permanent s'installe sur le front de la pandémie, résultat de l'impossibilité, à moyen terme, d'éradiquer le virus. L'attention des experts et des médias se retourne vers les taux de vaccination particulièrement bas de nombreux pays et vers l'apparition de nouveaux variants, potentiellement résistants aux vaccins. Il reste que l'aplatissement de la courbe appartient au passé, et, avec lui, le temps séquencé et partagé des modèles de prédiction. Le *benchmarking* établit une rupture radicale dans l'expérience et le rythme de la pandémie, en dilatant la distance entre les pays qui immunisent leurs populations pour s'extraire de la crise et ceux qui enchaînent les records de contaminations et de décès, malgré les campagnes de vaccination. Des temporalités multiples sont gouvernées par la dialectique des nombres, entre le ciel imaginaire et l'enfer de la réalité.

5. Quatrième période : les grands nombres de la politique

La « judiciarisation » de la gestion de la pandémie distingue la situation brésilienne. On utilise ce terme en sciences politiques pour faire référence à l'accroissement des moyens juridiques en tant qu'outil pour résoudre des problèmes et demandes d'intérêt public, aux niveaux législatif, exécutif ou sociétal. Comme le gouvernement fédéral faisait initialement preuve d'un certain scepticisme face à la gravité de la maladie, et qu'il a ensuite soutenu des traitements à base de médicaments inefficaces, d'autres acteurs de la société civile, voire des entités de la fédération (municipalités, États fédérés), vont aux tribunaux pour exiger des actions de contrôle sanitaire. Cette judiciarisation occupe une place de plus en plus centrale dans la gestion de la pandémie.

Il est intéressant de mentionner quelques décisions rendues par la Cour suprême (*Supremo Tribunal Federal*). Certaines sont très générales et typiquement constitutionnelles ; elles abordent l'autonomie des entités de la fédération, permettant aux États et aux villes de décider de sujets, comme déterminer si le commerce et les lieux de culte doivent ou non rester ouverts ou si le vaccin doit devenir obligatoire, voire inclure des sanctions et imposer des restrictions de mobilité. Il y a aussi des décisions sur la direction des politiques publiques, comme celle qui oblige l'exécutif à formuler un plan national de vaccination et celle qui proroge les dispositions de la loi 13979, créée pour traiter des mesures de lutte contre la pandémie. La Cour suprême prend également des décisions typiquement bureaucratiques et techniques, permettant d'inscrire ou de déplacer des groupes sociaux et des secteurs de métiers dans la liste prioritaire de vaccination, d'assurer l'approvisionnement en matières premières nécessaires à la lutte contre la pandémie et de créer des barrières sanitaires pour protéger les peuples amérindiens.

Le pouvoir judiciaire finit par se prononcer sur des questions à caractère essentiellement politique, relevant donc des domaines du législatif et de l'exécutif, selon son propre *modus operandi*. Cependant, de telles décisions ne dérivent pas que de préceptes constitutionnels ou juridiques, un point important pour notre argumentation : les juges (en particulier le ministre Ricardo Lewandowski qui assume la plupart de ces actions) combinent dans leurs avis des éléments juridiques et des nombres, c'est-à-dire des arguments statistiques. Les constats que « la population indigène présente un plus grand risque de contagion » et du « risque élevé de contamination lors d'activités religieuses collectives présentielles » alimentent des décisions prises « selon des critères techniques et scientifiques ». Les juges incorporent dans les actes du procès des preuves quantitatives construites à partir de recherches épidémiologiques.

Dans ce nouveau contexte, des chercheurs s'efforcent d'estimer combien de décès auraient pu être évités si le gouvernement avait négocié plus tôt l'achat de vaccins. Ce chiffre représenterait les décès attribuables à l'(in)action du gouvernement fédéral. Une des estimations qui circule le plus dans la presse est celle de Pedro Hallal, épidémiologiste et professeur de l'Université

fédérale de Pelotas, institution qui publie des bulletins quotidiens sur la pandémie au Brésil²⁶. Selon ce chercheur, 180 000 décès (les trois-quarts du nombre total de décès) pourraient être associés à la mauvaise gestion du gouvernement. Il s'agit de la même proportion, publiée en janvier 2021, dans son article de *Lancet* (Hallal, 2021). Si, à l'époque, on avait adopté la moyenne mondiale des décès comme valeur de référence pour le calcul, on dispose maintenant du nombre total de vaccins qui auraient pu être achetés et appliqués dans un temps politiquement habile. D'autres initiatives semblables adoptent le nombre de vaccins disponibles et le taux de vaccination, consacré par le nouveau *benchmarking*, comme critère de quantification de la responsabilité politique du chef de l'exécutif. L'article du sociologue Celso Rocha de Barros en est un des exemples les plus emblématiques : il défend l'hypothèse qu'au moins 100 000 décès doivent être attribués à la gestion Bolsonaro²⁷. La publication de ce texte a entraîné la demande, formulée par un sénateur proche du gouvernement, de mise en examen du sociologue, dans un nouvel épisode qui illustre le degré de politisation des nombres dans la gestion de la pandémie.

La décision juridique a une grande répercussion, aux conséquences politiques les plus profondes ; elle advient le 8 avril 2021, lorsque le ministre de la Cour suprême, Luis Roberto Barroso, ordonne la constitution, par le Sénat, d'une Commission d'enquête parlementaire (*Comissão Parlamentar de Inquérito* ou CPI) pour sonder l'action du gouvernement fédéral face à la pandémie. Alors que toutes les exigences formelles pour l'installation de la CPI semblent satisfaites, le président du Sénat, allié du président Bolsonaro, fait tout pour la retarder.

La CPI est une des institutions à travers lesquelles le pouvoir législatif brésilien exerce sa fonction de surveillance, notamment sur le pouvoir exécutif²⁸. Pour l'installer, il faut l'approbation d'au moins un tiers des députés ou des sénateurs²⁹. Les parlementaires choisis pour la composer sont investis d'un pouvoir d'enquête semblable à celui des autorités policières et juridiques. Leurs vastes attributions prévoient la convocation de témoins (y compris des ministres), la sollicitation de documents et d'informations de toutes les instances de l'administration publique, l'accès aux données bancaires et fiscales, confidentielles en temps normal, et même le pouvoir d'ordonner la mise en prison d'un déposant qui produirait un faux témoignage. Cependant, la commission ne peut ni juger ni condamner. Ses membres rendent aux autorités compétentes un rapport final pour qu'elles prennent les mesures face aux délits et infractions constatés.

Plusieurs CPI célèbres ont été constituées ces dernières années. Toutefois, elles ont exercé leurs effets davantage dans l'arène politique proprement dite que dans la sphère judiciaire, avec très peu de condamnations. Elles permettaient surtout de nuire à l'image publique de certaines forces politiques. Ladite « CPI du covid-19 » a porté sur les actions du gouvernement fédéral, ce qui a mis en échec la gestion Bolsonaro. La plupart des onze sénateurs qui la composent n'appartiennent pas à la majorité ; son rapporteur³⁰, Renan Calheiros, est une figure traditionnelle de la vie politique brésilienne et un fervent opposant au président. Les dépositions ont, durant de longues heures, été diffusées en direct à la télévision, à la radio et sur internet. Environ trois millions de personnes ont accompagné les travaux de la commission durant les premières semaines.

Aussi bien les nombres que leur interprétation se sont révélés centraux dans l'action de la commission, que ce soit dans la production des indices et des preuves, leur fonction manifeste,

26. Cf. <https://wp.ufpel.edu.br/covid19/>

27. Cf. « "Consultório do Crime" tenta salvar Bolsonaro na CPI da Covid », *Folha de São Paulo*, le 9 mai 2021. Disponible sur : <https://www1.folha.uol.com.br/colunas/celso-rocha-de-barros/2021/05/consultorio-do-crime-tenta-salvar-bolsonaro-na-cpi-da-covid.shtml?origin=folha>

28. La Commission parlementaire d'Enquête est prévue dans la Constitution brésilienne, au 3^e paragraphe de l'article 38, et son fonctionnement est régi par la loi 1.579 de 1952.

29. C'est-à-dire 171 députés ou 27 sénateurs.

30. Les deux principales fonctions au sein de la CPI sont celles de président et de rapporteur. Le président décide, en consultant bien sûr les autres membres de la commission, qui convoquer en tant que déposant, que ce soit comme suspect ou comme témoin. Le rapporteur de son côté joue un rôle tout aussi sinon plus important. Il s'agit du responsable de la rédaction du document final, qui présente les conclusions des enquêtes et indique qui doit être puni et pourquoi.

ou dans la production d'images et de récits sur la pandémie et le gouvernement. À ce moment de la pandémie au Brésil, où la CPI a orchestré le débat public, s'est dessinée une nouvelle temporalité, associée à la notion de responsabilité (ou de culpabilité) dans la gestion de la pandémie.

Les sessions de la CPI se sont déroulées dans une salle dont la disposition suivait de près celle coutumière aux institutions politiques, avec toutefois un détail qui nous intéresse beaucoup. La longue table où siègent le président et le rapporteur était fixée sur une plateforme qui place les autres rangées de la salle en contrebas. Parmi les écrans sur lesquels on pouvait lire le nom et la fonction des personnes assises à la table, une information attirait l'attention. En effet, devant le rapporteur Renan Calheiros, un nombre était quotidiennement mis à jour : le total de décès dans le pays, accompagné des mots « vies perdues à cause du covid-19 ».



Figure 6 – *Le rapporteur de la CPI, Renan Calheiros, devant le total des morts du covid-19*
(Source : Instagram, @renancalheiros³¹)

Les dépositions du ministre de la Santé, de ses prédécesseurs et de dizaines d'autres personnes ont servi à documenter les politiques mises en œuvre au cours des derniers mois et ainsi à établir l'adoption de conduites inappropriées ou l'existence de délits. Les enquêteurs ont cherché à mettre en évidence des liens objectifs entre les actions et les décisions (ou leur absence) des dirigeants et les effets de la pandémie sur la population.

L'enquête, qui porte sur les mesures et les possibles omissions du gouvernement fédéral, altère la temporalité du débat public, en convoquant, pour la première fois, le passé pandémique directement au centre de la scène, avec la notion de responsabilité. Il s'agit d'examiner si, en prenant en compte des critères partagés publiquement et des données objectives, les gestionnaires ont respecté l'obligation légale de veiller à la santé de la population. Au moyen de dates et de quantités, de séries statistiques et de séquences chronologiques d'événements, les nombres deviennent des éléments fondamentaux, protocoles d'objectivité dont l'évocation synthétise des positions morales et politiques.

Au fil des jours et en fonction des thèmes discutés dans la CPI, deux nombres retiennent l'attention, car ils condensent, d'une part, l'action (ou l'inaction) du gouvernement, et, d'autre part, leurs conséquences possibles : celui du total de personnes vaccinées et celui du total des décès. Les débats sur la responsabilité du gouvernement gravitent donc autour de deux indicateurs : la quantité de doses proposées avec une clause préférentielle par la compagnie

31. Image de la vidéo disponible sur : https://www.instagram.com/tv/CO0OJ7ajbEs/?utm_source=ig_web_copy_link

Pfizer et par l'Institut Butantan, offre rejetée en 2020³², et le nombre de morts de la maladie. C'est autour de ces deux séries temporelles que la CPI construit dans son rapport final sa notion de responsabilité politique (Senado Federal, 2021).

Dans la section nommée « Mise en danger de la vie d'autrui – décès qui auraient pu être évités », le rapport documente, à partir de la notion de « morts évitables » (*mortes evitáveis*), les décès causés par le retard et la lenteur de la vaccination. Adoptant les recommandations des chercheurs déjà mentionnés (Hallal, Barros et d'autres), le modèle statistique de vaccination mobilisé dans le rapport indique que, si la vaccination avait été conduite à un rythme soutenu et si le gouvernement avait répondu favorablement aux premières offres de Pfizer et Butantan, le pays aurait pu éviter 127 000 décès. Si l'on applique, à ce chiffre, la Valeur statistique d'une vie (VSL), le rapport estime à 418 milliards de réaux le coût du retard pris dans la vaccination, valeur qui dépasse largement celui des vaccins dans les propositions initiales, d'environ dix dollars par dose (*ibid.*, pp. 1015-1017). Ces retards dans l'achat, la distribution et l'application des vaccins, et la non-adoption de mesures non pharmacologiques pour limiter la transmission de la maladie, justifient la responsabilisation du gouvernement fédéral, d'autant qu'au Brésil, comparé à la population, le covid-19 a causé presque deux fois plus de morts (89 %) que dans le reste du monde. Le rapport conclut donc à « une gestion téméraire de la part du gouvernement dans le combat contre la pandémie », avec une « stratégie de propagation du virus conduite de manière systématique par le gouvernement fédéral, suivie par les tentatives de résistance des autres pouvoirs et des entités fédératives » (*ibid.*, p. 1021).

On observe ici l'effet accru du *benchmarking* de la vaccination sur le cadrage du débat public brésilien, puisque l'espace de comparaison des performances des gouvernements nationaux, établi par ce dispositif depuis le troisième moment, fonctionne comme un outil de preuve qui précède le débat (Desrosières, 1993), tout en fournissant les paramètres d'examen et de jugement des deux parties en dispute. Cette nouvelle temporalité est donc marquée par le rapport direct entre la politique des grands nombres et les grands nombres de la politique.

Aussi bien dans les enquêtes de la commission que dans les reportages de la presse et les commentaires des réseaux sociaux, nous pouvons observer la production des nombres pandémiques « en train de se faire ». La boîte noire que nous avons face à nous n'est déjà plus aussi hermétique. Au contraire, les processus techniques et politiques de la quantification se spectacularisent, avec des sénateurs, des ministres, des journalistes et des experts débattant publiquement des nombres, de leur pertinence, de leur fabrication et de leurs ordres de grandeur. De telles controverses présupposent que le nombre visé doit être reconnu par tous comme objectif, techniquement correct et qu'il doit revêtir l'autorité d'une preuve indiscutable, et ce en amont du débat. Combiné à cela, on assiste à un évident effort de produire des estimations plus modestes, appartenant au registre du « au moins » – « au moins tant de morts ont été provoquées par l'inaction du gouvernement ». Le chiffre doit être assez grand pour caractériser le délit de responsabilité et émouvoir l'opinion publique, mais assez petit pour qu'il soit crédible et politiquement acceptable. En dernière instance, il s'agit d'un nombre consensuel, qui résiste à la critique.

Antérieurement, la mesure de l'efficacité du gouvernement se faisait par comparaison avec d'autres pays. Dans ce quatrième moment, le *benchmarking* constitue déjà une réalité de premier ordre, qui jalonne l'avalanche de nombres en discussion dans la CPI et dans le débat public. Adossés aux grands classements internationaux, ces nombres mettent en évidence la singularité de l'expérience brésilienne et se réfèrent donc à des scénarios hypothétiques, dans lesquels le gouvernement aurait pris d'autres décisions. La question n'est plus de savoir ce

32. « Executivo da Pfizer revela à CPI seis ofertas de vacina e presença de Carlos Bolsonaro em reunião ». Disponible sur : <https://g1.globo.com/politica/noticia/2021/05/13/executivo-da-pfizer-revela-a-cpi-seis-ofertas-de-vacina-e-presenca-de-carlos-bolsonaro-em-reuniao.ghtml>

qui va se passer ou ce qui est en train de se passer, mais ce qui aurait pu se passer – le futur conditionnel. Le conditionnel présent se construit dans l'encadrement statistique du passé.

6. Considérations finales

Dans cet article, nous avons proposé d'organiser un contenu assez hétérogène d'informations relatives à des événements qui ont été au fondement d'interprétations chiffrées de la pandémie. Cela nous a permis de souligner le rôle de « circuits chiffrés » dans l'interprétation d'un phénomène extraordinaire comme celui de la pandémie. Puisqu'il s'agit d'une analyse exploratoire et immergée dans le temps atypique que nous vivons, l'article avait aussi pour but de formuler des questions et des pistes de recherches, qui devront faire l'objet de plus amples approfondissements.

C'est le cas de trois problématiques en particulier. Il y a d'abord l'observation ethnographique d'une chaîne de transformation qui prend en compte la production et la circulation de nombres, depuis les façons de les enregistrer et de compter des médecins et des infirmiers dans les hôpitaux, lorsqu'ils remplissent les rapports et les dossiers des patients, jusqu'à la compilation de ces données, réalisée par les agences publiques, sous pression suite à l'aggravation de la pandémie. La ritualisation, l'enregistrement, le classement et la comptabilisation des décès sont passés par une série d'acteurs qu'il faudrait davantage prendre en considération.

Deuxièmement, la pandémie a mis en évidence et renforcé le rôle central des grands groupes technologiques, comme Google et Facebook, qui ont identifié des cas suspects et mesuré l'adhésion de la population à la restriction de la mobilité. Nous avons étudié comment les informations produites par des entreprises privées se sont servies des statistiques publiques ou y ont été associées dans la gestion de la pandémie ; quels furent les intérêts et les justifications de ces entreprises lorsqu'elles se sont engagées dans une telle tâche ; enfin, comment l'expansion de la surveillance numérique pendant la crise du coronavirus a mis encore plus en péril le contrôle public de l'information. Dans ce cadre de complexification croissante du circuit de production de données, il convient de se pencher sur l'importance des chiffres créés par des sources privées, qui méritent une enquête systématique sur leur rayonnement et leurs conséquences.

Enfin, il faudra évaluer comment le rapport de la CPI, publié en octobre 2021, et, spécifiquement, le calcul des décès qui auraient pu être évités, ont impacté et impacteront le débat public et la mise en accusation du gouvernement Bolsonaro, aussi bien auprès de la Cour suprême que du Parlement. Cet état de fait présente une importance particulière, car le cas brésilien est vu par les observateurs internationaux comme un possible précédent ouvrant la voie à la responsabilisation d'autres dirigeants. En ce sens, la recherche en cours inclut une cartographie des différentes estimations des morts évitables en circulation au Brésil, de leurs méthodes et de leurs valeurs de référence, depuis leur apparition dans le milieu scientifique jusqu'à leur entrée dans l'arène politique, culminant dans les décisions juridiques et politiques basées sur elles.

Si la politique des grands nombres structure le champ d'actions possibles et les modalités de critique de la réalité, les grands nombres de la politique pendant la pandémie ont mis au premier plan les contingences et les négociations sur l'attribution de valeurs morales à ces chiffres. Ce n'est pas un hasard si le recensement de 2020 risque de répéter un tel scénario qu'il faudra étudier, marqué par l'usure du système politique et par la mise en spectacle des nombres publics. Tout comme les chiffres, les estimations, les graphiques et les classements de la pandémie, le recensement révèle lui aussi la centralité des formes de quantification, en tant que dimensions structurantes de la vie sociale – la politique des nombres – et enjeux du débat public – les nombres de la politique.

Références

- Anderson B. (2006), *Imagined communities: Reflections on the origin and spread of nationalism*, Verso books.
- Boltanski L. et L. Thévenot (1991), *De la justification : les économies de la grandeur*, Paris, Éditions Gallimard.
- Boltanski L. (2014), « Quelle statistique pour quelle critique ? », in I. Bruno, E. Didier et J. Prévieux (éds.), *Statactivisme : comment lutter avec des nombres*, Paris, La Découverte, pp. 33-50.
- Bruno I. et E. Didier (2013), *Benchmarking : l'État sous pression statistique*, Paris, La Découverte.
- Bruno I., E. Didier et J. Prévieux (éds.) (2014), *Statactivisme : comment lutter avec des nombres*, Paris, La Découverte.
- Camargo A. de P. R. and C. Daniel (2021), « Social studies of quantification and its implications in Sociology », *Sociologias*, vol. 23, no 56, p. 42-81.
- Camargo A. de P. R., E. Motta et V. A. Mourão (2021), « Números emergentes: temporalidade, métrica e estética da pandemia de Covid-19 », *Mediações*, vol. 26, n° 2, pp. 311-332.
- Cohen P. C. (1982), *A calculating people: the spread of numeracy in Early America*, Chicago, University of Chicago Press.
- Desrosières A. (1993), *La politique des grands nombres: histoire de la raison statistique*, Paris, La Découverte.
- Desrosières A. (2008), *Pour une sociologie historique de la quantification*, Paris, Presses de l'École des Mines.
- Diaz-Bone R. and E. Didier (2016), « The Sociology of Quantification: perspectives on an Emerging Field in the Social Science », *Historical Social Research*, vol. 41, no 2, pp. 7-26.
- Espeland W. et M. Stevens (2008), « The sociology of quantification », *Arch. europ. sociol.*, vol. 49, no 3, pp. 401-436.
- Freitas V. G. (2020), « As mulheres negras e a pandemia do coronavírus », *Boletim Ciências Sociais e Coronavírus*, n° 44. Disponible sur : http://anpocs.com/images/stories/boletim/boletim_CS/Boletim_n44.pdf (accès le 7 juin 2021).
- Guimarães R. B., R. C. Catão, O. S. Martinuci, E. A. Pugliesi et P. S. S. Matsumoto (2020), « O raciocínio geográfico e as chaves de leitura da Covid-19 no território brasileiro », *Estudos Avançados*, vol. 34, n° 99.
- Hacking I. (1990), *The taming of chance*, Cambridge, Cambridge University Press.
- Hallal P. (2021), « SOS Brazil: science under attack », *The Lancet*, vol. 397, n° 10272, pp. 373-374.
- Holanda V. M. de S. (2020), « Da Gripe Espanhola ao Coronavírus: notas etnográficas de uma visitante em Sevilha, Andaluzia », *Boletim Ciências Sociais e Coronavírus*, no 51. Disponible sur : http://anpocs.com/images/stories/boletim/boletim_CS/Boletim_n51.pdf (accès le 7 juin 2021).

- Koselleck R. (2006), « Crisis », *Journal of the History of Ideas*, vol. 67, n° 2, pp. 357-400.
- Luo J. (2020), « When Will covid-19 End: Data-Driven Predictions », Data-Driven Innovation Lab, Singapore University of Technology & Design. Disponible sur : https://www.persi.or.id/images/2020/data/covid19_prediction_paper.pdf (accès le 23 mars 2021).
- Luo J. (2021), « Forecasting covid-19 pandemic: Unknown unknowns and predictive monitoring », *Technological Forecasting & Social Change*, n° 166. Disponible sur : <https://doi.org/10.1016/j.techfore.2021.120602> (accès le 23 mars 2021).
- Mackenzie D. (2006), *An Engine, Not a Camera. How Financial Models Shape Markets*, Cambridge, MIT Press.
- Mello K. S. S. (2021), « covid-19 nas prisões brasileiras: seletividade penal e produção de corpos descartáveis », *Boletim Ciências Sociais e Coronavírus*, n° 44. Disponible sur : http://anpocs.com/images/stories/boletim/boletim_CS/Boletim_n44.pdf (accès le 7 juin 2021).
- Motta E. (2020), « “Achatar a curva”: estética, topografia e moralidade da pandemia », *Blog DADOS*. Disponible sur : <http://dados.iesp.uerj.br/estetica-da-pandemia/> (accès le 23 mars 2021).
- Neiburg F. (2020), « Vidas, economia e emergência », *Boletim Ciências Sociais e Coronavírus*, n° 22. Disponible sur : <http://www.anpocs.com/index.php/publicacoes-sp-2056165036/boletim-cientistas-sociais/2335-boletim-n-22-vidas-economia-e-emergencia> (accès le 23 mars 2021).
- Porter T. (1995), *Trust in numbers: the pursuit of objectivity in science and public life*, Princeton, Princeton University Press.
- Santos R. V., A. L. Pontes et C. E. A. Coimbra (2020), « Um “fato social total”: covid-19 e povos indígenas no Brasil », *Cadernos de Saúde Pública*, vol. 36, n° 10.
- Senado Federal (2021), *CPI da Pandemia. Relatório Final*, Brasília, Senado Federal. Disponible sur : <https://legis.senado.leg.br/sdleg-getter/documento/download/72c805d3-888b-4228-8682-260175471243>
- Simmel G. (2009), *Philosophie des Geldes*, Köln, Anaconda Verlag.
- Takemoto M., M. Nakamura-Pereira, M. Menezes, L. Katz, R. Knobel, M. M. R. Amorim et C. B. Andreucci (2021), « Higher case fatality rate among obstetric patients with covid-19 in the second year of pandemic in Brazil: do new genetic variants play a role? » (preprint) medRxiv. Disponible sur : <https://doi.org/10.1101/2021.05.06.21256651> (accès le 7 juin 2021).
- Woolhandler S., D. U. Himmelstein, S. Ahmed, Z. Bailey, M. T. Bassett, M. Bird, J. Bor, D. Bor, O. Carrasquillo, M. Chowkwanyun, S. L. Dickman, S. Fisher, A. Gaffney, S. Galea, R. N. Gottfried, K. Grumbach, G. Guyatt, H. Hansen, P. J. Landrigan, M. Lighty, M. McKee, D. McCormick, A. McGregor, R. Mirza, J. E. Morris, J. S. Mukherjee, M. Nestle, L. Prine, A. Saadi, D. Schiff, M. Shapiro, M. Tesema, A. Venkataramani (2021), « Public policy and health in the Trump era », *The Lancet*, vol. 397, n° 10275, p. 705-753. Disponible sur : [https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736\(18\)32171-8.pdf](https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(18)32171-8.pdf) (accès le 23 mars 2021).

Pandémie, statistiques et fédéralisme : la COVID au Canada



Jean-Guy PRÉVOST¹

Centre interuniversitaire de recherche sur la science et la technologie (CIRST),
Université du Québec à Montréal

TITLE

Pandemic, statistics, and federalism: COVID in Canada

RÉSUMÉ

La pandémie récente a suscité au Canada de nombreux appels à la mise sur pied d'un système national standardisé de données sur la santé en vue de lutter plus efficacement contre la contamination. En raison de la division constitutionnelle des pouvoirs, la santé est au Canada de juridiction provinciale et ce sont donc les provinces qui sont au premier rang de la lutte contre la COVID. Elles se trouvent par conséquent également en charge de recueillir toutes les données pertinentes. Or, les pratiques différentes des provinces en matière de collecte des données rendent plus difficile l'établissement d'un portrait national en temps réel de la situation épidémique. Dans ce contexte, les autorités fédérales en matière de statistique – Statistique Canada en partenariat avec l'Agence de la santé publique du Canada – ont été marginalisées et de nouveaux acteurs – médiatiques, scientifiques, entrepreneuriaux – ont suppléé en partie à cette carence, remettant *de facto* en cause le « monopole de l'autorité statistique légitime ». La publication récente des données sur la surmortalité a confirmé de manière nette l'incapacité du système de surveillance mis en place à offrir un portrait national cohérent et à jour de la mortalité due à la COVID.

Mots-clés : COVID, statistiques, données, Statistique Canada.

ABSTRACT

The recent pandemic has prompted numerous calls in Canada for the establishment of a national standardized health data system to combat contamination more effectively. Due to the constitutional division of powers, health in Canada falls under provincial jurisdiction and it is therefore the provinces that are at the forefront in the fight against COVID. They are therefore also responsible for collecting all the relevant data. However, the different practices of the provinces in terms of data collection make it more difficult to establish a national portrait in real time of the epidemic situation. In this context, federal statistical authorities – Statistics Canada in partnership with the Public Health Agency of Canada – have been marginalized, and new actors – the media, scientists, entrepreneurs – have partly made up for this deficiency, putting *de facto* into question the "monopoly of the legitimate statistical authority". Recently published data on excess mortality have clearly confirmed the inability of the surveillance system put in place to provide an update coherent national portrait of mortality due to COVID.

Keywords: COVID, statistics, data, Statistics Canada.

1. prevost.jean-guy@uqam.ca

Il existe au Canada une sorte de consensus vertueux suivant lequel un système national standardisé de données statistiques aurait permis de suivre quotidiennement l'évolution de la pandémie et constitué de ce fait un instrument efficace de coordination de la politique nationale. Ainsi, selon Michael Wolfson (2020a), ancien statisticien en chef adjoint du Canada (où il était responsable des statistiques de santé) : « *Dans l'urgence pandémique actuelle, des données de haute qualité, normalisées et en temps réel sur les décès supplémentaires, les cas de COVID et les hospitalisations, ainsi que des détails sur les milliers de foyers de soins infirmiers et de résidences pour retraités à travers le Canada sont essentiels* ». En août 2020, Wolfson appelait même le gouvernement fédéral à utiliser l'arme du financement pour contraindre les provinces et territoires à s'entendre avec les autorités fédérales pour la mise sur pied d'un système standardisé de collecte des données individuelles permettant d'avoir un portrait « en temps réel » de la situation (2020a, 2020b), appel qu'il réitérait en janvier 2021 (2021). Plus récemment, les politologues Matthew Bouldon et Daniel Béland (2021) abondaient dans le même sens, soutenant que le gouvernement canadien « *doit s'engager à mettre en place une infrastructure nationale de données intergouvernementale coordonnée* » et, à cette fin, « *établir et appliquer des normes nationales essentielles pour la collecte de données sur la santé publique* ».

Selon la constitution canadienne (l'Acte de l'Amérique du nord britannique de 1867, devenu la Loi constitutionnelle de 1982), les *statistiques* sont en effet de juridiction fédérale² : et de fait, alors qu'existe Statistique Canada, une agence du gouvernement fédéral dont l'expérience, les ressources et la réputation sont considérables, les provinces, à l'exception du Québec, se sont peu investies, historiquement, dans le champ de la statistique publique, précisément parce que l'offre fédérale à cet égard était abondante et que, par conséquent, elles en ressentaient moins le besoin. Pourtant, sur le plan de la description statistique de la pandémie, on observe en fait ce que l'on pourrait décrire comme une sorte de « polyphonie » (certains diraient : une cacophonie) résultant du fait que, dans cette division du travail statistique où, normalement, le gouvernement central a le premier rôle, celui-ci et ses agences directement en cause, soit l'Agence de la santé publique du Canada (ASPC) et Statistique Canada (StatCan), sont en fait entièrement dépendants des choix faits par les autorités provinciales en première ligne dans la gestion de la crise sanitaire, ce qui rend délicates les comparaisons à l'intérieur même du pays aussi bien que l'évaluation des décisions prises dans les différentes juridictions. Déjà, au début de la pandémie, Wolfson (2020b) avait attiré l'attention sur les facteurs qui faisaient obstacle à une description statistique adéquate de la pandémie à l'échelle nationale : le conflit juridictionnel – les provinces protégeant jalousement leurs pouvoirs en matière de santé –, la crainte de transparence de la part des autorités fédérales comme provinciales et les obstacles posés par les autorités en charge de la protection de la vie privée.

En fait, depuis le début de la pandémie, les données produites sur l'évolution de la pandémie l'ont été essentiellement par les autorités de santé des provinces et territoires, et les usages qu'on en a faits ont été multiples. Ces usages peuvent être commodément regroupés sous trois rubriques³ :

- (1) D'abord, on a utilisé les données comme outil de *gouvernement*, ou, plus précisément, comme outil de *gestion* visant à orienter et à soutenir la prise de décision administrative et politique. On parle ici en premier lieu de tableaux de bord très simples construits autour d'observations portant sur quelques indicateurs faciles à comprendre – nombre de cas, d'hospitalisations, de patients aux soins intensifs, de décès, répartition par sexe et par âge –, mais aussi de modèles prévisionnels plus

2. L'article 91 de la loi de 1867 énumère les compétences du Parlement fédéral, parmi lesquelles on trouve « le recensement et les statistiques » (<https://www.justice.gc.ca/fra/pr-rp/sjc-csj/constitution/loireg-lawreg/p1t13.html>).

3. Cette tripartition s'inspire notamment de Desrosières (2008).

complexes reposant sur des hypothèses plus difficiles à interpréter.⁴

- (2) Dans une perspective plus rétrospective que prospective, les données statistiques ont également été utilisées comme outil de *preuve*. On peut mentionner ici, à côté de la littérature purement médicale ou épidémiologique, les nombreux articles qui ont cherché à évaluer l'impact des mesures prises par les gouvernements sur la diffusion du virus (ou sur le fonctionnement de l'économie), mais aussi les études de surmortalité, qui, en plus d'aider à évaluer l'impact de la pandémie, ont pu servir au contrôle de la qualité des données recueillies sur les décès au fil des mois.
- (3) Enfin, les données ont également et largement été utilisées comme outil de *communication* et de *débat*. Lors des conférences de presse fréquentes des autorités gouvernementales (ou des communications des autorités de la santé publique), les chiffres les plus faciles à comprendre ont été convoqués systématiquement à l'appui des recommandations et des décisions ; et la publication d'un tsunami de données dans les médias a rapidement suscité, par exemple, les comparaisons entre juridictions, notamment entre le Québec et sa voisine, l'Ontario.

Dans ce texte, nous commencerons par décrire rapidement le cadre constitutionnel dans lequel s'inscrivent les responsabilités et la marge d'action des gouvernements fédéral et provinciaux ; en d'autres termes, les contraintes générales découlant de la structure dans laquelle les « producteurs » de données statistiques évoluent. Nous situant ensuite au niveau de la production des données élémentaires, nous examinerons les problèmes pratiques qu'a pu poser leur agrégation, d'abord à l'échelle provinciale, puis à l'échelle du pays ; en d'autres termes, le jeu d'obstacles et de difficultés auxquels s'est heurtée l'ambition d'offrir un portrait statistique cohérent de la pandémie à *l'échelle canadienne*. Ensuite, en lien avec les trois types d'usages mentionnés plus haut, nous examinerons comment, en dépit de ces difficultés, s'est organisé l'effort d'harmonisation et de publicité des données statistiques à l'échelle canadienne, lequel a vu se développer, à côté de ce qu'on pourrait appeler « l'autorité statistique légitime », l'initiative de tiers comme les médias, des regroupements de chercheurs, voire des entreprises commerciales, engagés dans la mise à disposition de « données ouvertes ». Enfin, à la lumière des informations dont nous disposons après deux ans de pandémie, nous essaierons de tirer quelques conclusions quant à la qualité des données produites au fil de la pandémie et quant au poids respectif des divers obstacles évoqués.

1. Le cadre constitutionnel

Le Canada est une fédération dans laquelle les pouvoirs respectifs du gouvernement fédéral et des provinces sont à l'origine fixés par l'Acte de l'Amérique du Nord britannique. La santé est souvent présentée comme un pouvoir exclusif des provinces et, de fait, la santé et les services sociaux constituent aujourd'hui le poste budgétaire le plus important des provinces⁵. Toutefois, à partir de la fin des années 1950, le gouvernement fédéral s'est engagé massivement dans le financement des dépenses de santé, en remboursant aux provinces une partie des coûts liés à cette mission. Au fil des difficultés budgétaires, la part des coûts assumée par le gouvernement fédéral a fluctué : à l'origine de 50%, elle est tombée à moins de 20% à la fin des années 1990 – austérité oblige – et se situe aujourd'hui à un peu plus de 23% (Norris, 2020). La *Loi canadienne*

4. On a beaucoup parlé au début de la pandémie du modèle de l'*Imperial College* qui aurait convaincu le premier ministre britannique Boris Johnson de prendre des mesures plus robustes qu'il ne l'avait initialement souhaité ; au Québec, ce sont les travaux du Groupe de modélisation en mathématique et en économie de la santé liée aux maladies infectieuses, de l'Université Laval, en lien étroit avec la Direction de la santé publique, qui ont été les plus souvent mis de l'avant.

5. L'article 92 de la loi de 1867 énumère les compétences des législatures provinciales, parmi lesquelles on trouve « la création, l'entretien et la gestion d'hôpitaux, d'asiles et d'institutions ou établissements de bienfaisance dans les limites et pour les besoins de la province, à l'exclusion des hôpitaux maritimes » (<https://www.justice.gc.ca/fra/pr-rp/sjc-csj/constitution/loireg-lawreg/p1t13.html>).

sur la santé de 1984 énonce de son côté des principes qui lient ce financement à un certain nombre de principes (transférabilité, accessibilité, universalité, intégralité et gestion publique) auxquels les provinces doivent se conformer (*ibid.*, p. 2). En fait, le système de santé canadien est devenu au fil des ans l'un des piliers de l'État-Providence canadien ainsi qu'un marqueur significatif de l'identité canadienne – surtout canadienne-anglaise –, par rapport au voisin du sud. Mais il faut rappeler que l'assurance-santé publique universelle est en fait née en 1962 dans une province, la Saskatchewan, avant d'être répliquée d'une province à l'autre – le Québec l'a adoptée en 1970 – avec l'appui financier du gouvernement fédéral.

La Cour Suprême du Canada a toutefois statué en 1982 que la santé « *n'est pas l'objet d'une attribution constitutionnelle spécifique, mais constitue plutôt un sujet indéterminé que les lois fédérales ou provinciales valides abordent selon la nature ou la portée du problème de santé en cause dans chaque cas* » (cité dans Butler et Tiedemann, 2013, p. 1). Le gouvernement fédéral exerce des responsabilités très claires dans deux domaines pertinents au regard de la situation présente : la réglementation des produits pharmaceutiques ainsi que la surveillance et la prévention des maladies. Sur le plan administratif, c'est Santé Canada (le ministère de la santé) qui autorise les vaccins et c'est l'Agence de la santé publique du Canada (ASPC) qui est chargée de « surveiller » les maladies infectieuses. Depuis le début de la pandémie, le gouvernement fédéral est intervenu de plusieurs manières. Il a bloqué les frontières et fixé des contraintes aux voyages en avion – mais seulement après que les gouvernements provinciaux et les médias fussent intervenus vigoureusement pour lui demander de le faire. Il a autorisé, à la demande de ces provinces, le déploiement de militaires pour venir prêter main-forte au personnel des hôpitaux ou des centres de soins de personnes âgées dans les cinq provinces les plus touchées (Québec, Ontario, Manitoba, Saskatchewan et Alberta) ainsi qu'auprès de communautés autochtones. Il a engagé des dépenses colossales dans la mise sur pied de divers programmes pour soutenir l'économie, les entreprises et les particuliers. Enfin, il a procédé à l'achat et à l'autorisation des vaccins et veillé à la distribution des doses entre les provinces et territoires.

Les provinces étant responsables de l'administration des soins de santé, ce sont elles qui sont au front dans la lutte contre la COVID-19. Chacune dispose d'ailleurs d'une autorité de santé publique et de pouvoirs considérables pour mettre en œuvre les recommandations de celle-ci. Cela dit, l'organisation de la santé publique n'est pas la même d'une province ou territoire à l'autre (il y en a 13 au total) et les pratiques en matière de surveillance de la pandémie varient considérablement d'une juridiction à l'autre. L'intervention des provinces couvre donc la définition et la dissémination des mesures visant à éviter la propagation de la COVID, toutes les décisions touchant aux règles de confinement, l'organisation et la gestion du dépistage, la prise en charge des cas graves (hospitalisation) et leur traitement, l'établissement des constats de décès et l'administration des vaccins. En pratique, ce sont donc les provinces qui effectuent la surveillance au jour le jour de la progression de la pandémie ; ce sont elles qui *produisent* les données statistiques pertinentes, l'Agence de santé publique du Canada ne faisant que colliger les données que lui procurent les provinces.

La responsabilité des provinces et territoires en cette matière se traduit par le fait qu'elles « signalent *volontairement* les cas confirmés et probables de COVID-19 à l'Agence de santé publique » (mis en italique par nous). À cet effet, l'ASPC et Statistique Canada ont préparé un formulaire qui sollicite des renseignements sur les aspects suivants : (1) province ou territoire ; (2) classification des cas aux fins de surveillance (confirmés ou probables) ; (3) détails du cas (âge, sexe, genre, race, etc.) ; (4) profession ; (5) symptômes (date d'apparition) ; (6) conditions préexistantes et facteurs de risque ; (7) évolution clinique et résultats (hospitalisation, admission en soins intensifs, cas résolu, décès) ; (8) expositions (voyages, contact étroit, lieux d'exposition, etc.) ; (9) données de laboratoire (date et type de prélèvement, type de test, séquençage) ; (10) informations sur l'historique des vaccinations (statut de la vaccination, vaccin reçu, etc.) (Gouvernement du Canada, 2021a).

En pratique, comme on le verra plus loin, certaines de ces informations sont accessibles à un niveau très fin, mais seulement dans certaines juridictions ; certaines sont rendues publiques, mais seulement sous forme agrégée ; d'autres ne sont pas rendues publiques, d'autres enfin ne sont tout simplement pas recueillies (c'est le cas notamment de la « race »).

2. Quantifier la pandémie

Les problèmes qui surviennent lorsque l'on tente de combiner les informations produites par les provinces afin d'offrir un portrait en temps réel de l'évolution de la pandémie à l'échelle du pays ou encore de faire des comparaisons entre juridictions sont de plusieurs ordres. Ces difficultés tiennent d'abord (1) à la capacité des personnes et des institutions qui sont sur le terrain – médecins, hôpitaux, centres de soins de longue durée, laboratoires – d'acheminer avec la périodicité voulue (le plus souvent quotidienne) les informations requises. Elles tiennent également (2) aux différences entre juridictions (a) quant aux informations recueillies, (b) quant à la définition de certaines variables et (c) quant à la périodicité suivant laquelle elles sont publiées⁶. Elles tiennent enfin, et pour beaucoup, (3) à l'absence, constatée depuis bien avant 2020, de normes et de procédures communes pour la communication des données en santé à partir des provinces et territoires vers l'ASPC.

Au niveau local, les problèmes liés à la communication des informations essentielles à la gestion quotidienne de la pandémie ont été cruciaux et abondamment soulignés par les médias, surtout lors de la première vague. La vétusté du système informatique dans plusieurs établissements de santé, l'absence de compatibilité des formats dans lesquels étaient saisies les données, le recours au télécopieur (fax), voire au courrier postal (!) pour acheminer les données vers le centre ont nui sérieusement à l'actualité des données rendues publiques chaque jour (pour ne pas parler des données destinées aux seuls gestionnaires du système de santé, par exemple sur le personnel contaminé)⁷.

Plus complexe est le problème des différences entre juridictions. Ainsi, certaines variables (nouveaux cas, cas actifs, rétablis, décès) ont été généralement disponibles pour toutes les juridictions, d'autres (hospitalisations, patients aux soins intensifs) ne l'ont été qu'à partir d'un certain moment et, surtout, l'une d'entre elles, le taux de positivité, n'était publiée quotidiennement que par une seule juridiction, le Québec. Cela ne signifie pas qu'on ne calculait pas ces données ailleurs, mais seulement qu'on ne les publiait pas sur une base quotidienne. La Colombie-Britannique, par exemple, publiait la moyenne mobile sur 7 jours de ses taux de positivité. L'explication de cette situation n'est pas la même partout, mais elle tient sans doute au fait qu'il s'agit d'un indicateur très sensible aux protocoles de dépistage ; or, ces protocoles ont varié considérablement au fil de la pandémie.

Les définitions utilisées pour définir ces variables sont par ailleurs quelque peu différentes d'une juridiction à l'autre⁸. On pourrait parler ici de différences entre pratiques de quantification : si quantifier, c'est « *exprimer et faire exister sous une forme numérique ce qui, auparavant, était exprimé par des mots et non par des nombres* » et si cette opération consiste à convenir puis à mesurer (Desrosières, 2008, p. 10), on verra très bien, dans les exemples qui suivent, comment la comparaison entre mesures sera d'autant plus problématique que les conventions varieront d'une juridiction à l'autre. Ainsi, si toutes les juridictions s'entendent assez bien sur ce qu'est un « cas confirmé par laboratoire » (*i.e.* présence du virus confirmée par un test réalisé dans

6. Ajoutons à cela que les heures de publication des données quotidiennes varient d'une juridiction à l'autre et que le Canada couvre 5 fuseaux horaires.

7. Voir, par exemple, les passages à ce sujet dans Castonguay (2021), pp. 272-275. Ce problème ne semble pas avoir affecté le Québec seulement : voir Alba (2022), à propos de l'Italie.

8. L'examen le plus systématique de ces différences se trouve dans COVID-19 Canada Open date Working Group, 2020.

un laboratoire agréé), on observe que le Québec – seul – inclut dans les cas confirmés (par « lien épidémiologique ») les personnes dont les symptômes étaient compatibles avec ceux de la COVID et qui avaient subi une exposition au virus, mais n'avaient pas subi de test et n'étaient pas appelées à en subir. La notion de « cas probable par laboratoire », à distinguer du « cas confirmé », existe dans toutes les autres juridictions lorsqu'on est devant le cas d'une personne dont les symptômes sont compatibles avec ceux de la COVID mais dont le test n'est pas concluant. Cinq juridictions parlent de « cas probable par lien épidémiologique » lorsqu'il y a symptômes, test non concluant, mais exposition à une personne contaminée. Quelques juridictions ont développé la notion de cas « suspect » pour désigner les personnes ayant des symptômes et pour qui existe un lien épidémiologique (l'équivalent de ce qu'au Québec, on désigne comme un cas *confirmé* par lien épidémiologique). Seul le Québec, par ailleurs, a développé la notion de « cas clinique », correspondant à celui de personnes présentant des symptômes, mais pas de test concluant ni de lien épidémiologique. Il est difficile de mesurer précisément l'effet de telles différences de définition. Au Québec, seule province à inclure parmi les cas confirmés ceux établis par lien épidémiologique, ces derniers représentaient entre 10% et 13% durant les mois d'avril et de mai 2020, mais seulement 1% en décembre 2020 et janvier 2021, mois les plus lourds – en termes de cas – de toute la pandémie, pour remonter à des seuils entre 1% et 4% en février et mars 2021. Par ailleurs, l'ampleur de la vague liée au variant Omicron a contraint à restreindre très sévèrement le dépistage (mais pas au même rythme ni suivant les mêmes règles d'une juridiction à l'autre), ce qui a rendu très difficile l'estimation du nombre de cas et l'interprétation des taux de positivité.

Des différences sont également présentes quant à la définition de ce qu'est un décès relié à la COVID. Ainsi, une majorité des juridictions s'en tient à la définition de l'Organisation mondiale de la santé (OMS), elle-même recommandée par le gouvernement fédéral. Deux provinces toutefois – et parmi les plus importantes, soit la Colombie-Britannique et l'Ontario, qui représentent ensemble 52% de la population canadienne – ont choisi d'exclure des décès COVID les cas où seul un lien épidémiologique pouvait être établi. En revanche, ces deux provinces et trois autres ont choisi d'inclure parmi les décès les cas où les personnes n'avaient pas récupéré entre la maladie et le décès, même si le certificat de décès ne mentionnait pas la présence de la COVID. Un rapport sur la surmortalité observée depuis le début de la pandémie suggère que les différences quant aux pratiques de constat des décès survenus à domicile ont globalement eu pour effet probable de sous-estimer significativement les décès dus à la COVID en dehors des hôpitaux et des centres de soins pour personnes âgées, et ce, dans plusieurs provinces (Royal Society of Canada, 2021).

Dernier exemple : les « tests ». Les diverses provinces désignent ainsi quatre réalités différentes : il peut s'agir du nombre de tests « complétés », c'est-à-dire dont on a obtenu le résultat la veille, du nombre de tests « réalisés », en pratique le nombre des prélèvements effectués, du nombre de « personnes » testées et, dans un cas, du nombre de personnes testées et de celles « sous investigation ». La variété que l'on observe dans ces définitions tient sans doute à plusieurs facteurs, dont les ressources à la disposition des juridictions, mais aussi à des choix faits par les autorités de santé publique.

Mais ces différences témoignent d'un problème structurel plus profond, bien documenté par un rapport du Bureau du vérificateur général du Canada (BVG), rendu public le 25 mars 2021, et portant sur le travail accompli par l'Agence de la santé publique du Canada face à la pandémie de COVID-19. On ne parlera pas ici des critiques ayant trait à l'impréparation – qui ressemblent à celles qu'on entend dans bien des pays – ou encore à la gestion des frontières. Mais le rapport contient une section consistante sur l'information de surveillance, autrement dit les données sur la pandémie. Malgré une série de recommandations réitérées depuis 1999, en effet, l'établissement de normes et de procédures communes pour la communication des données en santé à partir des provinces et territoires vers l'ASPC demeure un processus inachevé. Une

entente multilatérale fut convenue en 2016, mais les annexes concernant l'information sur les maladies infectieuses n'ont jamais été complétées. Les différences que l'on a observées quant à l'information produite par les provinces et territoires n'ont donc rien d'étonnant. Dans les faits, selon le rapport du BVGC, entre février et juin 2020, seulement 4% des cas de COVID-19 avaient été signalés par les autorités provinciales et territoriales à l'ASPC dans les 24 heures suivant leur constatation. Et si l'ASPC déclarait en juillet 2020 avoir « obtenu 99% des formulaires de déclaration de cas attendus de la part des provinces et territoires », il n'en reste pas moins que « seulement 67% de ceux-ci contenaient des données sur l'hospitalisation » et « seulement 10% contenaient des données sur les symptômes, leur date d'apparition et les facteurs de risque » (Bureau du Vérificateur général du Canada, 2021, p. 22). Enfin, l'infrastructure technologique de la surveillance s'est également révélée déficiente. Comme on l'a dit, les données envoyées n'étaient pas toujours « lisibles par machine », obligeant à des interventions manuelles qui pouvaient être source d'erreurs. Le formatage des données différait d'une juridiction à l'autre et – mais c'est un problème qui ne concerne cette fois que l'ASPC – la capacité de stockage des données était insuffisante (*ibid.*, *passim*).

3. L'agrégation des données : peut-on harmoniser ?

Observons maintenant comment, dans ces conditions, divers acteurs ont tenté, à partir des données fournies par les provinces et en dépit des différences dans les pratiques et dans les choix, de broser un portrait statistique de la situation à l'échelle du Canada tout entier. Du côté fédéral, les deux agences publiques prioritairement impliquées dans ce travail sont, on l'a dit, l'ASPC et Statistique Canada. Chaque soir, à 19 heures (heure de l'Est), le gouvernement du Canada (en fait, l'ASPC) publie un sommaire épidémiologique national. Le tableau qui ouvre ce sommaire est présenté dans la Figure 1.

Faits saillants au 2 mars 2022, à 9 h HNE

Cas aujourd'hui	Cas totaux	Cas actifs	Personnes rétablies	Décès aujourd'hui	Décès totaux
3 696	3 296 503	111 663	3 148 202	43	36 638
Total des tests effectués	Pourcentage positif (derniers 7 jours)	Tests pour 100 000 personnes (derniers 7 jours)			
58 276 405	10,8%	127			

Figure 1 – *Bulletin quotidien du gouvernement du Canada* [Source : <https://sante-infobase.canada.ca/src/data/covidLive/Resume-epidemiologique-des-cas-de-COVID-19-au-Canada.pdf>]

Suivent plusieurs pages de tableaux et de graphiques fondés sur les formulaires de déclaration envoyés à l'Agence de santé publique du Canada par ses partenaires provinciaux et territoriaux. Ceux-ci permettent d'offrir une représentation « nationale » de l'évolution de la pandémie, mais aussi une image de son évolution différenciée selon les provinces et territoires. Une mise à jour hebdomadaire est également produite. S'appuyant sur certaines des données requises dans le formulaire de déclaration, ces rapports présentent une répartition des contaminations, hospitalisations et décès selon l'âge et le sexe, des données relatives aux milieux d'éclosion, aux

types d'exposition aux virus, etc. Enfin, une galerie de données visuelles interactives est tenue à jour (Gouvernement du Canada, 2021b).

Statistique Canada a évidemment collaboré étroitement avec l'ASPC pour le développement des produits présentés plus haut et développé un portail intitulé « La COVID-19 sous l'angle des données ». Les produits spécifiquement en lien avec la pandémie sont toutefois limités et ils portent sur ses effets – économiques et sociaux – plutôt que sur son suivi au jour le jour, laissé à l'ASPC. En fait, la collecte des données de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) a elle-même été interrompue pendant la pandémie de COVID-19 à la fin de mars 2020 et n'a repris qu'en septembre. Le produit le plus significatif – au regard du besoin de dresser un tableau national (rétrospectif) de la pandémie – est peut-être l'outil interactif portant sur les estimations provisoires hebdomadaires du nombre de décès, du nombre de décès attendus et de la surmortalité. Mais les renseignements permettant d'établir cet outil proviennent eux aussi des autorités provinciales, de l'état civil desquelles relève l'enregistrement des décès, et le délai dans la transmission des renseignements varie d'une juridiction à l'autre.

Il est clair que Statistique Canada – plus précisément : le tandem Statistique Canada/ASPC – n'a pas réussi à s'imposer comme l'autorité en mesure d'offrir un suivi ponctuel de la pandémie. L'architecture de la division des pouvoirs – le centre ne pouvant que « demander » des informations – et la diversité des conditions dans lesquelles s'opérait la production des données dans les provinces et territoires ont largement empêché l'autorité centrale de mener à bien la tâche de produire une représentation quantitative de la pandémie en temps réel visant à éclairer les décideurs et à orienter leur action, en somme de produire les statistiques susceptibles de constituer un outil actualisé de gestion ou de gouvernement.

En revanche, Statistique Canada, en tandem avec l'ASPC, a produit un certain nombre de travaux analytiques – les statistiques comme outil de preuve – sur les effets économiques de la pandémie. Les analyses à caractère plus démographique ou épidémiologique, elles, demeurent rares. Notons l'enquête sur la séroprévalence (Statistique Canada, 2021) dont les résultats furent quelque peu étonnants : comme l'ont souligné certains journalistes et médecins, le taux de séroprévalence observé au début 2021 (3,6% pour l'ensemble des Canadiens, soit 2,6% en raison d'une infection antérieure et 1% en raison de la vaccination – on était alors au tout début de cette opération) était beaucoup plus faible que celui trouvé dans d'autres pays. Ce taux faible rendait par ailleurs le croisement avec d'autres variables périlleux, les différences entre sous-groupes se situant souvent dans les marges d'erreur. Plus récemment, Statistique Canada, en s'appuyant sur des données administratives et de recensement, a produit une analyse croisant, d'une part, les décès dus à la COVID, et, de l'autre, le niveau de revenu des personnes décédées et le type de logement dans lequel elles vivaient (Statistique Canada, 2022)⁹. Mais le bilan reste maigre, si on le compare à ceux de l'Office for National Statistics (ONS) au Royaume-Uni ou de l'Institut national de la statistique et des études économiques (INSEE) en France, et il s'inscrit dans une tendance plus longue de diminution des travaux analytiques, Statistique Canada se limitant de plus en plus à la production d'ensembles de données dans les domaines relevant des responsabilités fédérales.

D'autres intervenants se sont toutefois révélés capables d'investir des ressources et de procéder eux aussi à une agrégation des données produites par les provinces et territoires. Ainsi, Radio-Canada, le diffuseur public, a conçu un outil qui offrait un portrait de la situation de chaque province ou territoire, mais aussi un portrait pour l'ensemble du pays. Le site permet également de créer des graphiques en fonction de diverses variables et ce, dans une interface nettement

9. De manière quelque peu surprenante, cette analyse révèle que le niveau de revenu n'est pas associé à une différence importante dans les taux de décès pour les personnes qui habitaient dans des maisons ; il l'est en revanche pour celles vivant dans des appartements et même de manière nettement plus forte lorsque l'édifice dépasse quatre étages.

plus « user-friendly » que celle de Statistique Canada. Les données présentées offraient ainsi un appui à qui souhaitait utiliser les statistiques comme outil de communication et de débat. Il est par exemple devenu possible de suivre en continu la campagne de vaccination à travers le pays et d'aiguillonner les gouvernements provinciaux qui traînaient de l'arrière dans cette course (Radio-Canada, 2020a, 2020b et 2020c). Ce rôle actif des médias, leur capacité à répondre à des impératifs de ponctualité et à offrir des produits visuellement attrayants au regard de la pandémie a été relevé pour d'autres pays dans le monde (voir, par exemple, Alba, 2022, à propos des journaux *Il Sole 24 Ore* et *Financial Times* ou, pour un tour d'horizon plus hâtif mais plus complet, Cahen, Cavalin et Ruiz, 2020).

Le *COVID-19 Canada Open Data Working Group* constitue pour sa part un regroupement de chercheurs qui s'est donné pour mission d'analyser et de visualiser en temps réel la progression de l'épidémie à l'échelle du pays en même temps que de rendre disponibles le plus grand nombre de données, au niveau le plus fin possible. Les données publiées par ce groupe proviennent des sites officiels des provinces et territoires (dans le cas de l'Ontario, directement des sites des autorités sanitaires locales), mais elles sont vérifiées et entrées manuellement, beaucoup des données publiées par ces sites l'étant dans des formats différents et pas toujours « lisibles par machine ». C'est le CCODWG qui, pour les besoins de son propre travail, a le plus documenté les variations que l'on peut observer dans les pratiques de collecte et de mise en forme des données entre les provinces et territoires mais aussi dans une même juridiction au fil du temps. Trois rapports techniques/méthodologiques présentent de manière détaillée les procédures suivies par le CCODWG dans la mise en ligne des données. Selon les fondateurs du groupe, leur ensemble de données est « plus détaillé, plus actuel et plus propice à l'analyse que les données agrégées » publiées au niveau fédéral (Berry, Soucy, Tuite, and Fisman, 2020). Certains outils interactifs développés par le CCODWG ont de fait été repris par les autorités publiques. C'est le cas de la carte du Canada animée mise en ligne par l'ASPC et du tableau de bord des interventions gouvernementales utilisé par l'Institut canadien d'information sur la santé (Canadian Institute for Health Information, 2021). Les données du CCODWG ont été utilisées également pour alimenter des sites comme *How's my Flattening?* qui concerne l'Ontario (<https://howsmyleftening.ca/#/home>), ainsi que celui de Radio-Canada.

Une autre contribution notable provient d'une entreprise privée, ESRI Canada, spécialisée dans les systèmes d'information géographique. Son portail sur la COVID-19 comprend des tableaux de bord pour le Canada, pour les provinces, ainsi que pour certaines régions et municipalités plus denses (ESRI, 2021a). On y trouve également diverses cartes liées aux données de recensement et de l'enquête nationale sur la santé. À côté de cela, ESRI rend disponibles des données sur les cas individuels : mais ici, comme c'était le cas pour les données agrégées, on observe une forte variation d'une province ou territoire à l'autre quant aux détails rendus publics (ESRI, 2021b)¹⁰. ESRI offre enfin un outil de visualisation qui met en rapport, pour des unités géographiques fines, les données relatives à l'évolution de la pandémie et celles de l'Enquête sur la santé de Statistique Canada (ESRI, 2021c).

Finalement, les tableaux de bord tenus à jour par le CCODWG et, surtout, par ESRI apparaissent plus attrayants, sur le plan de la forme, et plus riches, sur celui de la quantité des informations réunies que les visualisations produites par Statistique Canada et l'ASPC. Sur le plan analytique, les meilleurs produits sont également dus à des tiers : ainsi, à propos de la surmortalité due à la COVID, le rapport publié par la *Royal Society of Canada* (2021) ou, pour le Québec, les travaux du démographe Robert Choinière.

10. Seules la date et la région sanitaire dans la province sont prises en compte par toutes. Le sexe est fourni par cinq provinces, l'âge par quatre, le type d'exposition par trois et le statut (actif/rétabli) par deux seulement. Les autres informations ne sont fournies de manière systématique que par l'Ontario. Bien sûr, on trouve des données par région sanitaire sur ces variables, mais seul l'Ontario rend disponibles des données d'une telle granularité.

Les difficultés éprouvées par l'ASPC et Statistique Canada ont donc permis à d'autres joueurs – médias, groupes de scientifiques, entreprises de géolocalisation –, mus par leurs propres intérêts, de se donner pour mission de nettoyer, d'harmoniser et de tenir à jour les informations, relativement disparates, produites par les provinces. Ainsi, dans un pays caractérisé à ce jour par une centralisation particulièrement forte en matière de statistiques, assiste-t-on à l'avènement d'un écosystème plus complexe, où ces tiers ne pourraient certes exister sans la statistique publique, mais où ils suppléent à ses carences en offrant une indéniable « valeur ajoutée ».

4. Un jugement rétrospectif

À deux années du début de la pandémie, on peut s'interroger sur la solidité – ou la fragilité – de ces dispositifs. Les changements dans la stratégie de dépistage imposés par la vitesse à laquelle s'est propagé le variant Omicron ont amené à relativiser la valeur des cas et du taux de positivité comme indicateurs du développement de la pandémie. Le nombre d'hospitalisations est devenu une variable clé, la protection du système de santé constituant un objectif central au Canada comme ailleurs. Mais cet indicateur est plus complexe à interpréter qu'on ne l'aurait pensé au premier abord et on a vu des discussions surgir quant à la pertinence de distinguer entre les personnes hospitalisées en raison de la COVID et celles, hospitalisées pour d'autres raisons, mais qui contractaient le virus à la faveur de leur séjour à l'hôpital¹¹. La même distinction – avec/en raison de – a été soulevée à propos des décès. Et le recours à des méthodes indirectes d'estimation de la contamination comme la surveillance des eaux usées n'a pas été mené de manière systématique et durable.

Les données sur la surmortalité, maintenant disponibles pour une longue période, nous offrent un étalon à partir duquel on peut mieux évaluer non seulement les effets de la pandémie mais aussi la qualité de la mesure qui en a été faite. Au Québec, le démographe Robert Choinière a produit des travaux intéressants sur le sujet, mais un article récemment paru dans *The Lancet*, portant sur 74 pays et 226 juridictions subnationales (dont l'ensemble des provinces et territoires canadiens, moins l'Île-du-Prince-Édouard), compare le nombre de morts déclarées comme dues à la COVID à une estimation de la surmortalité pour les années 2020 et 2021. Le rapport entre les taux pour 100 000 correspondant à l'un et à l'autre permet du coup de porter un jugement sur la qualité des pratiques d'enregistrement déployées en cours de pandémie¹².

11. On a pu observer également que, même si le nombre de personnes hospitalisées était généralement plus élevé au Québec qu'en Ontario, la proportion des patients aux soins intensifs était généralement plus élevée en Ontario. On peut se demander si ces différences reflètent des pratiques diversifiées en matière d'admission ou de traitement des patients.

12. Les taux de surmortalité et le ratio entre ce dernier et la mortalité déclarée sont des estimations comportant une marge d'erreur. Nous avons choisi de ne retenir ici que le point médian de cette marge.

**Tableau 1 – Mortalité COVID et surmortalité, 2020-2021 :
Canada, provinces et territoires**

	Mortalité COVID/ 100 000 hab. (1)	Surmortalité/ 100 000 hab. (2)	Ratio 1/2
Canada	41.9	60.5	1.44
Québec	73.0	77.0	1.05
Manitoba	54.5	95.8	1.76
Saskatchewan	43.6	75,8	1.74
Alberta	40.3	67.7	1.68
Ontario	35.3	53.9	1.53
Colombie-Britannique	24.2	52.5	2.17
Yukon	19.7	42.4	2.15
Territoires du nord-ouest	14.4	38.1	2.64
Nouveau-Brunswick	11.1	27.0	2.44
Nunavut	8.3	27.7	3.35
Nouvelle-Écosse	6.1	18.0	2.95
Terre-Neuve et Labrador	2.0	11.0	5.59

Source : données tirées de COVID-19 Excess Mortality Collaborators 2022.

On observe d'abord des différences non négligeables entre les deux taux : ainsi, le Québec se retrouve en tête et loin devant la moyenne canadienne pour les décès COVID déclarés ; pour ce qui est de la surmortalité, il se retrouve plutôt dans un peloton, avec cette fois le Manitoba en tête. Alors que seulement deux provinces se trouvaient au-dessus du taux canadien pour les décès déclarés, elles sont quatre pour la surmortalité. Pour ce qui nous intéresse plus particulièrement, à savoir la qualité des données sur les décès publiées au fur et à mesure du déroulement de l'épidémie, on observe cette fois une différence très marquée entre le Québec (où le ratio des deux taux se situe tout près de 1) et le reste des provinces et territoires. Cette différence à elle seule démontre que les taux de décès agrégés pour l'ensemble du pays ne constituent pas un indicateur fiable des effets de la pandémie, le ratio du Québec (1.05) étant le seul à se situer sous le ratio canadien (1.44)¹³.

On peut comparer cette situation à celle d'autres pays où les données nationales résultent de l'agrégation de données au niveau subnational. Comme on peut le voir dans le tableau 2, les États-Unis et surtout le Royaume-Uni offrent, au niveau national, une meilleure congruence que le Canada entre taux de décès déclarés et taux de surmortalité, ce qui n'est pas le cas de l'Allemagne, de l'Italie, ni de l'Espagne.

13. Ce qui n'est pas sans poser problème pour les études comparatives qui ont cherché à évaluer l'effet des mesures prises par les autorités dans diverses juridictions en retenant le taux de décès déclarés comme variable dépendante.

Tableau 2 – Ratios entre taux de décès COVID déclarés et taux de surmortalité 2020-2021 (divers pays)

	Ratio	Juridictions	Écart-type
Royaume-Uni	0.97	4	0.13
États-Unis	1.37	51	0.56
Espagne	1.64	19	0.50
Allemagne	1.82	16	0.32
Italie	1.89	21	0.78
Canada	1.44	12	1.14

Source : données calculées d'après COVID-19 Excess Mortality Collaborators 2022.

Mais on peut également évaluer la cohérence des pratiques d'enregistrement des décès dus à la COVID entre juridictions de chacun de ces pays en calculant l'écart-type des ratios entre taux de décès déclarés et taux de surmortalité de chacune des juridictions. On relève ainsi pour le Royaume-Uni un ratio très proche de 1 et un écart-type faible (mais il y a peu de juridictions). L'Allemagne a fortement sous-estimé les décès dus à la COVID (1.82), mais de manière assez systématique d'un État à l'autre. On trouve une variation assez forte dans les autres pays, mais le Canada se démarque par un écart-type nettement plus élevé (1.14) que tous les autres¹⁴.

5. Conclusions

On peut tirer de tout cela un certain nombre de conclusions. En premier lieu, la difficulté du gouvernement fédéral et de ses agences, Statistique Canada et l'ASPC, à offrir un suivi, sous forme de tableau de bord, en temps réel de la pandémie est indéniable. Les données publiées quotidiennement par l'ASPC – c'est le cas pour les décès, variable pour laquelle les données sur la surmortalité fournissent *a posteriori* un étalon – ont clairement sous-estimé la situation réelle. Sur ce plan, le diagnostic des Wolfson, Béland et Bouldon paraît juste. Toutefois, ce problème de sous-estimation ne se manifeste pas à la même échelle d'une juridiction à l'autre. Le Québec se distingue positivement et de manière très nette des autres provinces et territoires à cet égard.

Pour ce qui est des obstacles ayant contribué au problème et que l'on a décrit dans la section 2, leur importance et leurs effets réels peuvent être caractérisés comme suit. (1) Les problèmes d'informatisation des hôpitaux et des résidences de personnes âgées dont on a abondamment parlé dans les médias ont pu entraîner des retards dans l'acheminement des informations et donc affecter l'actualité des données publiées au jour le jour. Mais les erreurs résultant de ces retards pouvaient être rapidement corrigées et n'affectaient pas durablement la qualité des données. (2) De la même manière, les incompatibilités entre provinces en matière d'infrastructure technologique ou de formatage des données ont entraîné des retards et engendré des coûts pour assurer leur harmonisation et leur « nettoyage », un travail dans lequel s'est grandement impliqué le CCODWG. (3) Mais ce sont essentiellement les différences entre juridictions quant aux définitions de variables comme les décès et aux protocoles mis en place pour les désigner

14. Si l'on retire la valeur extrême qu'offre Terre-Neuve et Labrador (5.59), on obtient un écart-type de 0.64. Et si l'on retire la valeur supérieure dans chacune des autres distributions, on obtient le même ordre : c'est toujours le Canada qui présente la plus forte variation entre juridictions. On pourrait répéter l'exercice pour l'ensemble de l'Europe de l'Ouest qui présente un taux de décès déclarés de 94.7/100 000, un taux de surmortalité de 140/100 000 et un ratio entre les deux de 1.48 : l'écart-type entre les ratios des (24) pays s'élève à 0.97 (une fois retirée la valeur extrême de l'Islande (-8,49), due à une forte sous-mortalité).

comme dus à la COVID ou non – la variété des conventions – qui ont sans doute joué le plus, comme le démontrent les écarts entre juridictions pour ce qui est des différences entre décès déclarés COVID et surmortalité.

En ce qui concerne les rôles qu'ont pu jouer et que jouent encore les statistiques depuis le début de la pandémie, on peut avancer les constats suivants. Les statistiques ont été utilisées comme outil de preuve par les chercheurs, parfois de manière qui a pu paraître un peu rapide (Institut canadien d'information sur la santé, 2020 et Société Royale du Canada, 2020)¹⁵, puis de façon plus sûre une fois que des données recueillies par d'autres voies que l'appareil de surveillance de la pandémie – sur la surmortalité, notamment – fussent rendues disponibles (voir notamment Royal Society of Canada, 2021, qui attirait l'attention sur les anomalies dans le décompte des décès survenus à domicile – par contraste avec ceux survenus dans les institutions). Les statistiques publiées au jour le jour ont largement alimenté la discussion publique, notamment en ce qui concerne les jugements sur la performance respective des divers gouvernements provinciaux et territoriaux à contenir les effets de la pandémie. Force est de constater, avec le recul, qu'en raison des différences entre provinces dans leurs pratiques de quantification, nombre de ces jugements se sont fondés sur des données difficilement comparables et devront être révisés à la lumière de résultats pour lesquels les conventions sont mieux établies. Enfin, comme outil de gouvernement ou de gestion, les statistiques ont indubitablement joué un rôle *au niveau des provinces et territoires*, quelles qu'aient été leurs imperfections. Le lien entre données et décisions suppose certes de nombreuses médiations, qui vont de différences dans l'analyse et l'interprétation de ces données au poids accordé aux autres considérations, qu'elles soient économiques ou politiques. Mais on peut dire qu'au niveau du gouvernement *fédéral*, ce rôle a été de toute façon moins important, les principales décisions qu'il a prises l'étant soit en réaction aux demandes des provinces (fermeture des frontières, envoi de militaires), soit non directement liées aux données de surveillance de la pandémie (l'achat et l'approbation des vaccins). Croire qu'un système national standardisé de collecte des données individuelles offrant un portrait en temps réel de la situation à l'échelle du pays aurait permis « *d'assouplir le confinement au moyen d'une 'quarantaine intelligente' et de récolter les principaux avantages d'un retour à la normale de l'économie à un rythme plus rapide* » (Wolfson, 2020b) ne tient aucunement compte de la réalité politico-administrative. En fait, compte tenu des dimensions du territoire canadien, de la division des responsabilités en matière de lutte contre la pandémie et de la diversité des pratiques sur le terrain, des outils comme le taux de positivité ou le taux de reproduction et les données brutes dont ils dérivent ont un sens à une certaine échelle – celle des provinces, voire celle de sous-régions dans les provinces les plus étendues. Combinés en vue d'offrir un portrait national de la pandémie, ils ne constituent que des agrégats artificiels et n'offrent en aucune manière un guide pour des décisions prises de toute façon à une autre échelle.

Références

Alba S. (2022), « Blurring the lines. Journalism and epidemiology in the time of corona », *Significance*, February, pp. 31-33, <https://rss.onlinelibrary.wiley.com/doi/10.1111/1740-9713.01611>.

Berry I., J.-P. R. Soucy, A. Tuite, and D. Fisman (2020), « Open access epidemiologic data and an interactive dashboard to monitor the COVID-19 outbreak in Canada », *CMAJ*, April 14, 2020, vol. 192, n° 15, E420, <https://doi.org/10.1503/cmaj.75262>.

Bouldon M. and D. Béland (2021), « Statistics Canada plays an important role in COVID », *Policy Options*, 10 février 2021, <https://policyoptions.irpp.org/magazines/february-2021/statistics-canada-plays-important-role-in-covid/>.

15. Voir la critique de ce rapport dans Prévost (2020).

Bureau du vérificateur général du Canada (2021), « Préparation en cas de pandémie, surveillance et contrôle des frontières », https://www.oag-bvg.gc.ca/internet/Francais/parl_oag_202103_03_f_43785.html.

Cahen F., C. Cavalin et É. Ruiz (2020), « Des chiffres sans qualités ? Gouvernement et quantification », <https://halshs.archives-ouvertes.fr/halshs-02659791>.

Canadian Institute for Health Information (2021), « COVID-19 Intervention Timeline in Canada », <https://www.cihi.ca/en/covid-19-intervention-timeline-in-canada>.

Castonguay A. (2021), *Le printemps le plus long. Au cœur des batailles politiques de la COVID-19*, Montréal, Québec-Amérique.

Choinière R. (2022), « Surmortalité hebdomadaire en 2020 et 2021 en lien avec la COVID-19 au Québec et ailleurs dans le monde », <https://comparaisons-sante-quebec.ca/exces-hebdomadaires-de-deces-en-2020-en-lien-avec-la-covid-19-au-quebec-et-ailleurs-dans-le-monde/#provinces>.

COVID-19 Canada Open date Working Group (2020), « Epidemiological Data from the COVID-19 Outbreak in Canada », Technical Report 2.0. Methodology & Codebook, https://drive.google.com/file/d/1o2L_Y5KtA85oOz-4FZ_w7ID6vrpID61p/view.

COVID-19 Excess Mortality Collaborators (2020), « Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020-21 », *The Lancet*, 10 March, [https://doi.org/10.1016/S0140-6736\(21\)02796-3](https://doi.org/10.1016/S0140-6736(21)02796-3).

Desrosières A. (2008), *L'argument statistique I. Pour une sociologie historique de la quantification*, Paris, Presses de l'École des Mines, <https://books.openedition.org/pressesmines/909?lang=fr>.

ESRI (2021a), *Ressources sur la COVID-19*, <https://ressources-fr-covid19canada.hub.arcgis.com/?locale=fr>.

ESRI (2021b), *Compiled COVID-19 Case Details (Canada)*, <https://resources-covid19canada.hub.arcgis.com/datasets/compiled-covid-19-case-details-canada/data>.

ESRI (2021c), *COVID-19 Health Dashboard*, <https://ressources-fr-covid19canada.hub.arcgis.com/apps/90fdd2da4bba4c79a33c2202760b3c5d/explore>.

Institut canadien d'information sur la santé (2020), « La pandémie dans le secteur des soins de longue durée : où se situe le Canada par rapport aux autres pays ? », <https://www.cihi.ca/sites/default/files/document/covid-19-rapid-response-long-term-care-snapshot-fr.pdf>.

Butler M. et M. Tiedemann (2013), « Le rôle fédéral dans le domaine de la santé et des soins de santé », Publication n° 2011-91-F, Ottawa, Service d'information et de recherche parlementaires, Bibliothèque du Parlement, <https://lop.parl.ca/staticfiles/PublicWebsite/Home/ResearchPublications/InBriefs/PDF/2011-91-f.pdf>.

Gouvernement du Canada, *Loi constitutionnelle de 1982*, <https://www.justice.gc.ca/fra/pr-rp/sjc-csj/constitution/loireg-lawreg/p1t13.html>.

Gouvernement du Canada (2021a), *Surveillance nationale concernant la maladie à coronavirus (COVID-19)*,

<https://www.canada.ca/fr/sante-publique/services/maladies/2019-nouveau-coronavirus/directives-provisoires-surveillance-infection-humaine.html#enc1>.

Gouvernement du Canada (2021b), *Visualisations actives de données de COVID-19*, <https://sante-infobase.canada.ca/covid-19/galerie-donnees-visuelles/>.

Norris S. (2020), « Le financement fédéral des soins de santé. Service d'information et de recherche parlementaires », Publication n° 2018-45-F,

<https://lop.parl.ca/staticfiles/PublicWebsite/Home/ResearchPublications/BackgroundPapers/PDF/2018-45-f.pdf>.

Prévost J.-G. (2020), « COVID, statistiques et effet de perspective : les décès survenus en soins de longue durée », Blogue du CIRST, 14 juillet 2020,

<https://www.cirst.uqam.ca/nouvelles/2020/covid-statistiques-et-effet-de-perspective-les-deces-survenus-en-soins-de-longue-duree/>.

Radio-Canada (2020a), <https://ici.radio-canada.ca/info/2020/coronavirus-covid-19-pandemie-cas-carte-maladie-symptomes-propagation/>.

Radio-Canada (2020b), <https://ici.radio-canada.ca/info/2020/09/covid-19-pandemie-cas-deces-propagation-vague-maladie-coronavirus/>.

Radio-Canada (2020c), <https://ici.radio-canada.ca/info/2021/02/vaccination-doses-immunite-collective-covid-19/>.

Royal Society of Canada (2021), « Excess All-Cause Mortality During the COVID-19 Epidemic in Canada », <https://rsc-src.ca/en/news/excess-all-cause-mortality-during-covid-19-epidemic-in-canada>.

Société royale du Canada (2020), « Rétablir la confiance : la COVID-19 et l'avenir des soins de longue durée », https://rsc-src.ca/fr/r%C3%A9tablir_la_confiance_la_covid-19_et_l_avenir_des_soins_de_longue_dur%C3%A9e.

Statistique Canada (2021), « Peu de Canadiens avaient des anticorps contre le SRAS-CoV-2 au début de 2021 », <https://www150.statcan.gc.ca/n1/daily-quotidien/210706/dq210706a-fra.htm>.

Statistique Canada (2022), « Les facteurs sociodémographiques et socioéconomiques qui sont liés aux taux de mortalité attribuables à la COVID-19, 2020-2021 », <https://www150.statcan.gc.ca/n1/fr/daily-quotidien/220308/dq220308d-fra.pdf?st=UJmgivky>.

Wolfson M. (2020a), « Ottawa must put data first and tie to health funding », *The Toronto Star*, 7 août 2020, <https://www.thestar.com/opinion/contributors/2020/08/07/ottawa-must-put-data-first-and-tie-to-health-funding.html>.

Wolfson M. (2020b), « During the pandemic, why has Canada's data collection lagged behind », *The Globe and Mail*, 14 avril, <https://www.theglobeandmail.com/opinion/article-during-the-pandemic-why-has-canadas-data-collection-lagged-so-far/>.

Wolfson M. (2021), « Canada needs a national database to track COVID-19 vaccination », *QUO*, 11 janvier 2021,

<https://quoimedia.com/canada-needs-a-national-database-to-track-covid-19-vaccination/>.

Une conversation avec Yves Escoufier



Yves ESCOUFIER¹
Université de Montpellier

Gilbert SAPORTA²
Conservatoire national des arts et métiers, Paris

TITLE

A conversation with Yves Escoufier

RÉSUMÉ

Yves Escoufier est un des fondateurs de l'école française d'analyse des données. Son influence a amplement dépassé les frontières de l'hexagone avec ses travaux sur les tableaux à trois indices, l'opérateur qui porte son nom et le coefficient RV de corrélation entre deux vecteurs.

Très investi dans la communauté des statisticiens, il a été président de l'Association pour la Statistique et ses Utilisations — ancêtre de la Société Française de Statistique — de 1984 à 1986. Par ailleurs, Yves Escoufier a été vice-président de 1989 à 1994, puis président de 1994 à 1999 de l'université de Montpellier 2. Il a été vice-président de l'Institut international de statistique de 1991 à 1993. Il est docteur honoris causa de l'université La Sapienza de Rome (1996).

Au cours de cet entretien réalisé en septembre 2021, Yves Escoufier revient sur sa carrière et nous fait part de ses réflexions.

Mots-clés : analyse factorielle, opérateurs, coefficient RV, STATIS.

ABSTRACT

Yves Escoufier is one of the founders of the French school of data analysis. His influence has extended far beyond France with his work on three-index tables, the operator that bears his name and the RV coefficient for correlation between two vectors.

Very involved in the community of statisticians, he was president of the Association pour la Statistique et ses Utilisations - ancestor of the Société Française de Statistique - from 1984 to 1986. In addition, Yves Escoufier was vice-president from 1989 to 1994, then president from 1994 to 1999 of the University of Montpellier 2. He was Vice-President of the International Statistical Institute from 1991 to 1993. He holds an honorary doctorate from La Sapienza University in Rome (1996).

In this interview conducted in September 2021, Yves Escoufier looks back on his career and shares his thoughts with us.

Keywords: factor analysis, operators, RV coefficient, STATIS.

1. escoufieryves@gmail.com / désigné par les initiales YE dans cet entretien.
2. gilbert.saporta@cnam.fr / désigné par les initiales GS dans cet entretien.

GS : Comment es-tu devenu statisticien ?

YE : Quand je me suis inscrit en licence de mathématiques, mon idée était d'aller ensuite dans l'industrie. J'avais d'ailleurs écrit à EDF et à la SNCF qui m'avaient répondu : « Quand vous aurez votre licence... ».

Au cours de la licence, j'ai suivi un certificat semestriel de probabilités et statistique qui était le seul certificat à caractère appliqué. Cela m'avait intéressé et j'avais même envisagé de m'inscrire à l'ISUP.

À la fin de la licence, Monique Lafont qui était la professeure de probabilités et statistique, a fait savoir que le centre d'études phytoécologiques (CEPE) que le professeur Louis Emberger venait de créer au CNRS, cherchait un calculateur³ pour faire des analyses de variance. D'autre part le département de maths créait un DEA et comme j'avais rencontré mon épouse, j'ai donc décidé de rester à Montpellier à la fois pour travailler comme calculateur à mi-temps au CNRS, et pour faire le DEA. C'était en 1964-65.

À la rentrée 1965, on nous a informés qu'il y avait des postes d'assistant à prendre. C'était le moment où l'université rentrait dans ses nouveaux murs, où le nombre d'étudiants augmentait, et il y avait plus de postes d'assistants que de candidats. Pendant un an j'ai été assistant au département de mathématiques.

J'avais demandé à pouvoir continuer ma recherche au CEPE parce que j'y avais découvert les méthodes factorielles et l'ACP, et aussi appris le Fortran. À la rentrée 1966, Jean Falguerettes, qui était l'assesseur du doyen Bernard Charles, créait le département informatique de l'IUT et cherchait des gens qui connaissaient un peu d'informatique. Et puisque je connaissais Fortran, j'étais considéré comme un informaticien. Et c'est comme ça que j'ai débuté ma carrière.

Pour revenir au CEPE, j'y avais donc été embauché pour faire des analyses de variance. Un jour, des chercheurs sont arrivés avec la thèse de Pierre Dagnelie⁴ sur l'analyse factorielle en écologie et m'ont demandé de leur expliquer ce qu'il faisait. Et c'est comme ça que j'ai été immergé dans ce type de méthodes. On avait – je crois que ça a été un gros avantage des Français, on en reparlera – fait beaucoup d'algèbre linéaire dans notre formation, si bien que toutes ces méthodes multivariées étaient évidentes pour nous, en tout cas très faciles à aborder. Et je me suis plongé dans le livre de H. Harman⁵, puis dans ceux de C. R. Rao⁶ et de T. W. Anderson⁷.

Mes premières analyses en composantes principales, je les ai faites à la main, sur la machine de bureau avec laquelle d'habitude je faisais des analyses de variance. Ça prenait du temps. J'ai alors appris qu'à Nancy, Richard Tomassone, qui était à ce moment-là à l'École des eaux et forêts, proposait des stages dans lesquels il présentait ces méthodes d'analyse statistique multivariée. Parallèlement, Claude Pair apprenait à les programmer en Fortran. Et c'est là que j'ai commencé à voir l'ensemble des méthodes multivariées et à les programmer. Richard Tomassone – que l'on retrouvera au moment de la création de l'unité de biométrie – a été mon premier contact avec quelqu'un qui connaissait ces méthodes et les enseignait. Sinon, c'étaient les livres. Il n'y avait pas de compétences sur Montpellier.

Côté recherche, j'ai toujours travaillé avec les sciences du vivant. Dès le début, j'ai fait de la statistique avec des chercheurs qui étaient en écologie, en biologie : pour eux la statistique,

3. Un emploi de technicien à l'époque, pas une machine

4. Professeur à la Faculté des Sciences agronomiques de Gembloux, Belgique

5. Harman H. H. (1960), *Modern factor analysis*, Univ. of Chicago Press.

6. Rao C. R. (1965), *Linear statistical inference and its applications*, John Wiley & Sons Inc.

7. Anderson T. W. (1958), *An introduction to multivariate statistical analysis*, John Wiley & Sons Inc.

c'était les analyses de variance ; ils comparaient des parcelles selon leur production et on faisait ça à la main, enfin à la machine. Ayant vu qu'ailleurs des chercheurs faisaient de l'analyse multivariante, les gens du CEPE se sont demandé pourquoi on ne ferait pas pareil. Une fois en poste à l'IUT, on a su que j'étais capable de faire des analyses statistiques et il y a eu énormément de demandes de biologistes et d'entreprises locales.

On a eu par exemple un contrat de recherche avec la direction environnement d'Elf Aquitaine. Elf Aquitaine extrayait à Lacq du gaz qui était chargé en soufre. Les paysans autour s'étaient plaints des rejets de soufre. Il y avait un réseau de mesures autour de Lacq pour mesurer la quantité de soufre dans l'air toutes les heures ou les deux heures. La direction est venue nous voir en nous demandant si on pouvait les aider à rationaliser ce réseau parce qu'ils étaient bien conscients qu'on avait mis un capteur ici parce que le maire avait crié fort, un autre là parce que c'était à côté de la maison de monsieur untel, etc. Mais les premières données qu'ils nous ont apportées étaient sous forme de listings car à cette époque-là – c'était le tout début – ce qu'on savait faire, c'était enregistrer sur des *listings*. Le premier travail a été de tout ressaisir. Plus tard on a vu arriver les disques et les bandes magnétiques. C'était pareil avec les Salins du Midi : ils enregistraient des données de météo depuis 100 ans. Quand ils ont eu un ordinateur, au lieu de les mettre dans un cahier, ils les ont mis sur l'ordinateur. Et puis un jour, il y a quelqu'un qui a dit : « Mais ça sert à quoi ? »

Rationaliser le réseau conduisait à un problème de sélection de variables, sujet que je développais. Cela a permis à plusieurs étudiants d'avoir des bourses financées par Elf Aquitaine.

C'étaient les débuts de l'informatique : quand on voulait faire un calcul, on écrivait son programme sur du papier, il y avait des personnels qui le perforaient dans des cartes. On allait porter son paquet de cartes perforées aux techniciens et on retournait une demi-journée après ou une journée après, on voyait le résultat sur du papier, quand il y en avait un.

Jusqu'aux années 1975, quand un biologiste par exemple faisait une régression, il mettait le programme de régression dans sa thèse. Moi, quand j'ai écrit ma thèse de troisième cycle sur l'ACP, il y avait le programme d'ACP dans ma thèse. C'est impensable à l'heure actuelle sous cette forme papier mais à ce moment-là, c'était nécessaire parce que si tu n'avais pas le programme, tu ne pouvais pas utiliser la méthode et cela faisait partie de la preuve que tu maîtrisais la méthode. Cela étant, on demande maintenant de fournir le code R pour s'assurer de la reproductibilité des résultats.

Le département informatique de l'IUT a été un des premiers créés avec ceux de Grenoble et de Paris, parce qu'il y avait à Montpellier l'usine IBM qui fabriquait les ordinateurs 360. Au début quelques ingénieurs d'IBM participaient à l'enseignement. Ensuite s'est créé un centre associé du CNAM avec une filière informatique où les diplômés de l'IUT qui étaient embauchés chez IBM ou dans des entreprises à Montpellier, venaient préparer le titre d'ingénieur du CNAM. Il n'y avait pas de visioconférence à ce moment-là. Paul Namian⁸ faisait un cours au CNAM à Paris qui était filmé, on recevait les bobines, on projetait les cours et on discutait avec les étudiants. C'est comme cela que j'ai vécu la naissance de l'informatique à Montpellier.

Au début, la théorie c'était que quand on créait un département d'IUT, on fermait un BTS. Par exemple, quand on a créé le département informatique, on a fermé un BTS mécanographie. Quand on a créé le département gestion des entreprises, on a fermé un BTS de secrétariat. Mais après, ça s'est arrêté. Les lycées techniques ont voulu garder leurs BTS.

8. Premier professeur d'informatique au CNAM

Pendant un ou deux ans, je suis intervenu dans un certificat de licence qui s'appelait « Analyse numérique », où je présentais des méthodes statistiques aux côtés d'un professeur de classes préparatoires qui présentait les méthodes de calcul numérique. Il y a eu une volonté portée par la direction de la faculté (Bernard Charles et Jean Falguerettes) de développer les mathématiques appliquées qui n'existaient pas à Montpellier. Comme il n'y avait personne du côté analyse numérique, on a développé la statistique.

Mais j'étais très isolé à Montpellier. Bernard Charles et Jean Falguerettes me disaient : « Il faut que tu fasses des stats », mais j'étais seul. Il n'y avait pas d'équipe, ni de séminaire ; chacun travaillait dans son coin.

L'AFCEP (Association Française pour la Cybernétique Économique et Technique)⁹ a joué un rôle important en me permettant de rencontrer des statisticiens, des numériciens et des électroniciens : c'était le seul endroit où on entendait de tout, il y avait des informaticiens d'entreprise, qui parlaient de Cobol, d'analyse au sens de l'informatique de gestion, d'aide à la décision, de recherche opérationnelle, etc. Tous ceux et celles qui commençaient à se confronter à l'ordinateur, que ce soit pour le construire ou l'utiliser, s'y retrouvaient¹⁰. J'y ai également rencontré Paul-Louis Hennequin¹¹.

GS : Ton nom est associé à un opérateur et à un coefficient. Quelles en furent la genèse ? Peux-tu en rappeler l'usage ?

YE : Mes interlocuteurs savaient que j'étais capable de faire des ACP et ils arrivaient avec leurs données. Les expérimentateurs avaient l'habitude de travailler avec certaines variables, mais ils disaient : « On pourrait en créer d'autres » ou bien « Celle-là, je pense qu'elle apporte la même information que l'autre ». Or, comment répondre mathématiquement à cette interrogation ? La seule solution, c'est que tel paquet de variables amène à telle représentation d'individus. Tel autre paquet de variables amène à telle autre représentation d'individus.

L'idée a été de dire : « Ce qui est caractéristique d'un ensemble de variables, c'est la manière dont elles vont nous permettre de représenter les individus ». Or, qu'est-ce qui fait la représentation des individus ? Ce sont les composantes principales qui sont les vecteurs propres de l'opérateur des produits scalaires entre les individus, noté \mathbf{W} . On calculait toujours la matrice de variance parce qu'elle a une dimension plus restreinte. Mais ce qui fait la représentation des individus, ce sont les valeurs et vecteurs propres de \mathbf{W} .

On connaissait déjà un autre opérateur : l'opérateur de projection de la régression. Mais l'ennui de l'opérateur de projection, c'est que s'il engendre l'espace vectoriel, il détruit les distances et ne génère pas la représentation des individus. C'est comme cela que j'ai été amené à travailler sur ce nouvel opérateur.

On retrouve le fait qu'en analyse des données on s'intéresse plus aux individus qu'aux relations entre variables. En statistique classique les individus ne sont là que pour apporter une information sur les variables. Alors que là, la perspective est différente : ce sont les variables qui apportent de l'information sur les individus.

Les types de problèmes étaient également différents : les chercheurs du CEPE étaient toujours capables d'inventer une variable de plus pour décrire le sol. « On a pris l'humidité à 10 cm, est-

9. Créée en 1968, l'AFCEP fut très active et influente mais fit faillite et disparut en 1998 après des investissements hasardeux.

10. C'est seulement depuis 1984 qu'existe une section informatique au CNU.

11. Professeur à Clermont-Ferrand (1930-...) ; joua un grand rôle dans la diffusion des probabilités et de la statistique. Il est un des fondateurs de l'école d'été de Saint-Flour.

ce que tu veux que je mette aussi l'humidité à 20 cm ?, etc. ». Il s'agissait de faire des tirages dans l'espace des variables, mais comment faire ?

J'avais rencontré dans mes lectures un texte dans lequel H. Hotelling s'interrogeait sur la convergence des valeurs propres et des vecteurs propres de la matrice de variance de groupes de variables tirées au hasard dans une population infinie de variables¹². Plus tard Mahalanobis, qui mesurait des crânes, du temps où il faisait de la comparaison des castes, écrivait en substance : « Je prends une mesure, puis je tourne d'un degré, il y a bien un moment où il y a trop de variables, où les nouvelles variables n'apportent rien ».

Ma thèse d'état a consisté à essayer de donner un formalisme qui permettait de parler de ce problème et de trouver une solution : comment tirer dans un espace qui est le produit d'un espace aléatoire sur les variables et d'un espace aléatoire sur les individus ? Après, il fallait montrer qu'il y avait des convergences, ce qu'on fait classiquement en mathématiques. Et l'opérateur est venu comme ça. Mais vraiment, ma motivation c'était la comparaison d'ensembles de variables.

Je suis persuadé qu'il y a encore une réflexion à faire sur les familles de variables. Pour les individus, on dit : « Ils sont tirés de manière indépendante », mais ça n'empêche pas qu'ils peuvent se ressembler. L'idée c'est que les variables, c'est pareil, elles peuvent être tirées de manière indépendante, mais elles peuvent se ressembler, elles peuvent être corrélées. Dans sa thèse Paulo Gomes¹³ se demandait : « Quand on norme les variables, elles sont toutes sur la sphère unité de \mathbb{R}^n , quelle est leur distribution ? ». El Mostafa Qannari¹⁴ a travaillé là-dessus aussi : au lieu de prendre une variable, il prend le projecteur associé à la variable et l'ensemble des projecteurs se trouve sur une autre sphère. Il y a toujours des gens qui s'y intéressent comme Xavier Bry¹⁵ avec la distribution de von Mises-Fisher.

Puisque toutes les méthodes d'analyse statistique multivariées consistent à calculer des valeurs propres et des vecteurs propres, cela veut dire que l'on peut expliciter l'opérateur associé à chaque méthode : l'analyse discriminante, l'analyse canonique et l'analyse des correspondances. Dans chaque cas, ce sont des choix de données et une manière de les utiliser, c'est-à-dire de pondérer les individus et de prendre une métrique sur les variables.

Le coefficient RV est venu un peu après avec l'idée de pouvoir comparer ces opérateurs. Mathématiquement ils sont dans des espaces de Hilbert ; si on prend un produit scalaire, on tombe automatiquement sur la cov (ou covariance vectorielle) qui est le produit scalaire des deux opérateurs. Et en divisant par les normes, on a le RV.

La première étape a donc été de définir l'opérateur, parce que pour comparer des groupes de variables, on ne peut que comparer la manière dont on représente les individus. Le produit scalaire entre opérateurs quantifie la ressemblance des représentations des individus. STATIS¹⁶ puis le triplet **X**, **Q**, **D** sont des enrichissements ultérieurs. Cela paraît simple maintenant et on se demande pourquoi on n'y avait pas pensé plus tôt ! Notons qu'on a montré plus tard qu'en prenant des matrices **X**, des métriques **Q** et des poids **D** particuliers, RV s'identifie aux différentes mesures de liaisons entre variables.

12. Hotelling H. (1933), Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, vol. 24, n° 7, pp. 498-520, <https://psycnet.apa.org/doi/10.1037/h0070888>

13. Gomes P. (1987), *Distribution de Bingham sur la n-sphère : une nouvelle approche de l'analyse factorielle*, Thèse, Montpellier.

14. Qannari E. M., E. Vigneau et Ph. Courcoux (1998), Une nouvelle distance entre variables. Application en classification, *Revue de Statistique Appliquée*, vol. 46, n° 2, pp. 21-32, http://www.numdam.org/item/RSA_1998__46_2_21_0/

15. Bry X. and L. Cucala (2018), Classifying variable-structures: a general framework, *arXiv preprint*, arXiv:1804.08901.

16. Structuration des Tableaux À Trois Indices de la Statistique

GS : Parlons donc de STATIS.

YE : À cette époque-là, il y avait tout un courant sur la représentation des matrices de similarité ou de dissimilarité. Joseph Kruskal et Douglas Carroll parlaient de *multidimensional scaling*. Une fois obtenus les opérateurs et le coefficient RV, on pouvait définir une matrice de similarité entre plusieurs tableaux par la matrice des RV entre opérateurs. Elle avait toutes les bonnes qualités, elle était définie positive. Ce que l'on a appelé *l'interstructure* – c'est-à-dire la comparaison des opérateurs entre eux – consistait à calculer les valeurs propres et les vecteurs propres de cette matrice de similarité.

Comme tous les coefficients RV sont positifs, le premier vecteur propre de la matrice de similarité pouvait être choisi avec tous ses coefficients positifs. Prendre la combinaison linéaire des opérateurs définis par ce premier vecteur propre donnait encore un opérateur positif. On pouvait le diagonaliser, ce qui donne le *compromis*. Par des méthodes tout à fait analogues à ce qu'on fait en ACP en projetant des points supplémentaires, on projetait les opérateurs et on avait ce que l'on a appelé *l'intrastructure*. La démarche était assez naturelle : finalement, c'est la même chose que l'ACP, mis à part qu'au lieu de parler de variables, on parle d'opérateurs ; et au lieu de parler de corrélations, on parle de RV.

Toute cette construction était prête quand je suis revenu du Canada, ce qui a conduit à la thèse de Henri l'Hermier des Plantes¹⁷.

GS : Et le schéma de dualité, si typique d'une certaine approche française¹⁸ ?

YE : Quand j'ai rencontré Jean-Pierre Pagès et Francis Cailliez en 1972, ils travaillaient sur ce qui deviendra ensuite leur livre¹⁹. Ils dessinaient déjà leur schéma de dualité. Ils notaient « E » pour l'espace des individus et « F » pour l'espace des variables. Les variables sont représentées dans F, mais peuvent aussi être représentées comme une application dans E, c'est-à-dire des éléments du dual E^* et la matrice X permet de passer de E^* à F. Aux individus qui sont des points de E, on peut aussi associer des éléments du dual F^* . La matrice transposée X' permet de passer de F^* à E. Du côté de E, il y avait la matrice de variance et la notion de choix d'une métrique Q , donc un opérateur VQ de E dans lui-même. Mais de l'autre côté, ils ne savaient pas quoi mettre.

Je suis arrivé en disant : « Mais ce qu'on fait avec VQ d'un côté, c'est la même chose de l'autre côté avec WD », ce qui a donc trouvé automatiquement sa place dans leur schéma.

On peut toujours se demander si cette lecture très algébrique est vraiment utile, mais l'intérêt que j'y trouvais, était de faire apparaître vraiment les choix effectués. Il y a le tableau de données X , avec la métrique Q que l'on utilise pour calculer les distances entre individus, et D , la matrice des poids que l'on donne aux individus pour calculer les variances et les covariances. Tout est explicite. Alors qu'en général, D c'est $(1/n)I_n$ et on n'en parle plus.

Henri Caussinus et Yves Aragon²⁰ ont traité du cas où les individus étaient auto corrélés – dans l'espace ou dans le temps – avec une métrique N à la place de D qui décorrélait les individus.

Les utilisateurs, par exemple, réduisaient leurs variables, donc les données changent, ce ne sont plus celles qui ont été observées, ce qui signifie que l'on utilise une métrique particulière

17. L'Hermier des Plantes, H. (1976), *Structuration des tableaux à trois indices de la statistique*, Thèse, Montpellier.

18. Pour une présentation récente en anglais : De la Cruz O. and S. Holmes (2011), The duality diagram in data analysis: Examples of modern applications, *The Annals of Applied Statistics*, vol. 5, n° 4, pp. 2266-2277.

19. Cailliez F. et J.-P. Pagès (1976), *Introduction à l'analyse des données*, SMASH, Paris.

20. Aragon Y. et H. Caussinus (1980), Une analyse en composantes principales pour des unités statistiques corrélées, in E. Diday et al. (eds.), *Data Analysis and Informatics*, pp. 121-131, North Holland.

sur le tableau d'origine.

Au début pour moi, c'était purement un outil de dialogue avec les expérimentateurs qui m'apportaient des données, pour bien expliciter les choix qu'ils faisaient. Et, en fait, ça s'est avéré beaucoup plus riche puisqu'après, on peut voir l'analyse discriminante, l'analyse des correspondances simple ou multiple, comme des analyses en composantes principales de triplets particuliers, etc.

Et ça a été à l'origine de l'ACPVI (Analyse en composantes principales avec variables instrumentales). Si j'ai deux tableaux X et Y décrivant les mêmes individus munis des mêmes poids, quelle métrique mettre sur X pour que les ACP que l'on peut faire sur X ressemblent autant que possible à celles que l'on a faites sur Y ? On s'aperçoit alors que c'est une autre lecture de la régression, obtenue par la représentation des produits scalaires entre individus et non par la reconstruction de chaque donnée.

J'avais suggéré à John Gower de présenter l'analyse procustéenne dans ce schéma mais je n'ai pas continué ma réflexion là-dessus. Pour lui, c'étaient plutôt des successions de rotations, d'homothéties et de translations.

J'ai repensé récemment à l'analyse canonique : si on prend le projecteur associé à Y , que l'on cherche la métrique à mettre sur X pour reconstruire ce projecteur, et que l'on fait de même dans l'autre sens, on arrive aux équations de l'analyse canonique.

GS : Comment définis-tu la spécificité de l'école française d'analyse des données ?

YE : Dans cette période de 1965 à 1971, les statisticiens, où qu'ils soient, étaient confrontés au fait qu'il y avait des ordinateurs, que des expérimentateurs, des entreprises et des laboratoires de recherche engrangeaient des données et qu'à un moment donné, on leur a dit : « Sur ces données, il faudrait quand même que vous nous sortiez de l'information ».

Ce qui a fait la spécificité des Français – et le livre de Pagès et Cailliez me paraît un bon exemple – c'est qu'on avait une formation en algèbre qui nous permettait d'écrire de manière précise les manipulations mathématiques qu'on faisait. Dans le livre de Harman déjà cité on ne trouve aucune matrice. Ce sont toujours des sommes de a_i , a_{ij} , x_{ij} ...

De l'autre côté de l'Atlantique, John Tukey proposait plutôt des extensions de ce qu'on appelle les méthodes descriptives. Au lieu de faire un histogramme, on faisait un histogramme avec des couleurs différentes. Les Américains jouaient aussi avec les possibilités que donnait l'ordinateur pour enrichir les analyses avec le Bootstrap ou le Jackknife : au lieu de faire l'analyse une seule fois avec l'échantillon, on ré-échantillonne dans l'échantillon, on fait des permutations dans l'échantillon. Ils n'hésitaient pas à faire des calculs multiples que nous n'osions pas faire encore. Mais cela restait très unidimensionnel alors que les Français ont essayé de voir ça globalement avec toujours ce souci de formaliser mathématiquement et d'arriver à quelque chose qui se calcule.

Il y a un article de C. R. Rao²¹ dans Sankhya que j'ai cité plusieurs fois, qui explique l'importance de faire du multidimensionnel. Si C. R. Rao en a vu la nécessité, cela veut dire qu'à cette époque-là, ce n'était pas si évident. Cet article prend la peine d'expliquer qu'on doit faire du multidimensionnel quand on a plusieurs variables, et que l'on n'a pas le droit de les étudier les

21. Rao C. R. (1964), The use and interpretation of principal component analysis in applied research, *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 26, pp. 329-358.

unes à côté des autres.

J'ai connu beaucoup plus tard les travaux japonais qui étaient complètement indépendants, à l'occasion des congrès qu'Edwin Diday organisait à Versailles et qui ont eu un rôle très important. C'est là qu'on a vu arriver pour la première fois Chikio Hayashi et Kameo Matusita, et, plus tard, Noboru Ohsumi. Ces congrès m'ont également permis de découvrir John Gower, John Nelder, Alfredo Rizzi et d'établir des liens forts avec beaucoup d'échanges, en particulier avec les Italiens qui ont invité certains Français à leur réunion annuelle en Italie et l'année suivante on invitait des Italiens.

GS : Et Jean-Paul Benzécri ?

YE : J'ai croisé une fois Jean-Paul Benzécri à Toulouse dans un congrès de l'AFCEC où j'avais présenté une analyse discriminante que j'avais faite avec des médecins de Montpellier. Je crois que par sa personnalité, par le fait qu'il enseignait à Paris, son aura, ses travaux, il a vraiment ouvert la porte à l'analyse des données en France. En plus, il a eu de nombreux étudiants, certains très brillants. Ils avaient les programmes qui permettaient de mettre en œuvre les méthodes, l'analyse des correspondances ou les méthodes de classification.

La difficulté, c'est qu'il n'a jamais fait partie vraiment d'un groupe, il ne s'est jamais intégré dans une communauté ; je pense qu'on ne l'a jamais vu aux réunions de l'ASU. Il était vraiment à part. Son comportement ou sa personnalité ne facilitaient pas les contacts.

GS : Quel rôle jouent les modèles en statistique en général et dans ta conception de la statistique ?

YE : Il faut préciser d'abord de quel modèle on parle : si on parle des lois comme la loi normale ou comme la loi binomiale, comme des modèles possibles d'aléatoire, c'est utile dans la mesure où si on a de bonnes raisons de penser qu'un phénomène peut être décrit par ce genre de modèles, cela donne beaucoup d'informations sur le phénomène. Je dirais que ces modèles sont des outils à la fois de dialogue, de description. Si tu sais qu'un modèle marche bien, c'est un élément de langage. Si tu dis à quelqu'un : « Ma loi suit une loi normale », tous les statisticiens ont compris.

Après, il y a les modèles du type de ceux qu'on a en biologie, comme les modèles de croissance. Jean-Marie Legay à Lyon a pas mal écrit sur ces sujets-là²². Il peut y avoir des ambiguïtés sur les objectifs : veut-on reconstruire les résultats, ou veut-on reconstruire le phénomène ? Ce n'est pas toujours clair chez les expérimentateurs. On peut vouloir faire un modèle qui va effectivement reconstruire le résultat, mais sans avoir l'ambition de dire : « Ça explique que le phénomène fonctionne comme ça »²³.

Interpréter des modèles à des niveaux très simples n'est pas toujours évident. Par exemple, après une régression multiple, certains utilisateurs constatent que « telle variable explicative a un signe négatif » et vont expliquer que quand la variable va croître alors la variable expliquée va décroître, sans tenir compte que cette variable est corrélée avec d'autres et que donc son effet ne peut pas être interprété indépendamment des autres variables.

Dans certains modèles, leurs concepteurs ont mis beaucoup d'*a priori* c'est-à-dire de savoirs sur le sujet, sans s'interroger sur la robustesse justement de l'impact de ce qu'ils ont mis dans

22. Legay J.-M. (1986), Méthodes et modèles dans l'étude des systèmes complexes, *Cahiers de la recherche-développement*, vol. 11, pp. 1-6.

23. Cette démarche se retrouve en apprentissage : « *Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms* » (Vapnik, 2006).

le modèle. Je pense à certains énormes modèles qui me font un peu peur comme les modèles climatiques, les modèles de production agricole, etc. Je m'interroge si tous leurs utilisateurs sont bien conscients de la fragilité liée au fait qu'ils y ont mis beaucoup de leur propre savoir ?

Par ailleurs, les prévisions météorologiques marchent relativement bien. Mais elles reposent sur des données énormes. Guy Der Mègréditchian²⁴ n'est plus là pour nous dire comment ils font, mais je pense qu'ils reconstruisent plus les résultats que les phénomènes. En plus, ils travaillent beaucoup par comparaison avec des situations semblables.

Finalement, je suis un peu mal placé pour parler des modèles dans la mesure où je me suis surtout intéressé à la statistique exploratoire !

GS : Peux-tu nous parler de tes années canadiennes ?

YE : Comme je le disais plus haut, j'étais très seul à faire de la statistique à l'université de Montpellier, et c'était un souci pour moi. À Montréal, il y avait tous les ans une école d'été de mathématique. En 1968, elle était consacrée à la statistique. Et donc j'y suis allé. Je n'en ai pas un grand souvenir : Lucien Le Cam avait fait un cours, et je ne me souviens plus de qui étaient les autres. Mais j'ai rencontré pour la première fois Pierre Robert qui était alors le directeur du département informatique et recherche opérationnelle de l'Université de Montréal.

Je suis revenu à Montpellier, j'ai continué, j'ai passé ma thèse d'État et je voulais aller ailleurs, pour respirer un peu et pour rencontrer des statisticiens. Et il se trouve que dans l'hiver 1970-1971, Pierre Robert est passé en France et son message était : « J'ai des postes au département IRO à Montréal et je cherche des gens qui voudraient venir quelques années ». J'ai dit « banco », et nous sommes partis avec nos trois enfants dont le plus jeune qui avait un mois.

Partis pour un an, nous sommes restés en fait deux années scolaires.

J'étais donc au département « Informatique et recherche opérationnelle » parce que les probabilistes se trouvaient au département de mathématiques, et les statisticiens au département IRO avec Pierre Robillard, Robert Cléroux, Roch Roy, plus des numériciens et des informaticiens.

Les probabilistes étaient des mathématiciens qui n'étaient pas intéressés par le traitement des données, mais par le développement de modèles. De l'autre côté, les statisticiens étaient confrontés à des données réelles, à la nécessité de les stocker, de les manipuler et d'en dire quelque chose d'opérationnel, de faire des prévisions.

Chaque semaine il y avait un séminaire au département et chaque mois un séminaire commun entre l'Université McGill et l'Université de Montréal. J'ai vu passer tous les grands noms de la statistique et tous les chapitres de la statistique que je ne connaissais pas. Pour moi, ça a été une révélation de découvrir des aspects de la discipline auxquels je ne m'étais jamais confronté et de pouvoir en parler facilement avec des collègues. Du point de vue recherche, ça a été très important, un grand moment d'ouverture.

C'est à Montréal que j'ai rédigé l'article pour *Biometrics* à partir de mes travaux de thèse : les opérateurs et le RV. Et j'ai eu la chance qu'un des lecteurs, J. Mosiman du National Health Center à Washington, se soit intéressé à cet article ; il a demandé à me voir et l'article a été publié.

24. Statisticien à Météo France (1933-1990)

Outre cet aspect recherche et ouverture sur l'ensemble de la statistique, je me souviens de Pierre Robillard qui m'a dit : « Tu vas certainement faire un polycopié comme tous les Français ». Et je lui ai dit : « Oui, et comment tu fais, toi ? » Et lui : « Moi, je repère le dernier livre paru sur le sujet que je dois enseigner, je dis aux étudiants de l'acheter et puis on le lit ensemble ».

Ça voulait dire deux choses :

- En France, les étudiants faisaient grève pour avoir des polycopiés gratuits, et au Canada on leur demandait d'acheter des livres et ils trouvaient cela normal.
- Mais en même temps, cela montrait une position très différente de l'enseignant vis-à-vis du savoir. En France, l'enseignant apporte le savoir. Là-bas, le savoir est dans le livre, l'enseignant est simplement celui qui a quelques années d'avance sur les étudiants et qui peut les aider à mieux le comprendre.

Pierre Robillard ajouta : « Quand il y a un chapitre qui ne me plaît pas ou que je trouve incomplet, à ce moment-là, je complète ». Donc, c'était vraiment une autre manière de vivre le rapport enseignant/enseigné. J'avais trouvé ça intéressant du point de vue pédagogique.

Je crains que cela n'ait pas beaucoup changé en France où chaque enseignant a toujours envie d'avoir son propre document et n'aime pas utiliser les documents des autres. Alors que là-bas, c'était le contraire.

L'attitude des étudiants était différente également. Avec le système majeure/mineure, certains venaient suivre un cours dans leur mineure pour avoir un éclairage sur des aspects particuliers. Ils me demandaient par exemple : « Allez-vous parler de tel modèle d'analyse de variance ? Parce que là où je travaille, on l'utilise et je veux le comprendre ». Ce n'était pas du tout du bachotage, alors que les étudiants français voulaient juste réussir l'examen à la fin de l'année, quel que soit le contenu du cours.

Il y avait une autre différence : les étudiants qui arrivaient à l'université étaient passés par le Cégep²⁵. Ils étaient à peu près du niveau deuxième année de licence avec un peu plus de maturité que ceux qui sortent du bac. Et puis peut-être que tout leur système éducatif est plus axé sur l'autonomie de l'individu.

GS : Comment s'est réalisé le passage du CRIG à l'unité de biométrie ?

YE : Après être revenu de Montréal, nous avons créé en 1973 avec deux collègues de l'IUT, Jean Ferrier qui était informaticien et Robert Reix qui était gestionnaire, le « Centre de recherche en informatique et gestion ». On nous a donné une salle dans le bâtiment K de l'IUT. C'était notre laboratoire ! J'avais aussi des étudiants de l'option Analyse des données du DEA de mathématiques où j'intervenais et quelques maîtres-assistants d'alors ont commencé à travailler avec moi, comme Christine Lavit, Alain Delcamp, Jean-François Durand, Henri Aris.

À Montréal, le département IRO avait institué la pratique suivante : chaque année un professeur de statistique était dispensé d'une partie de son enseignement pour être disponible, pour travailler avec les chercheurs des autres disciplines. Je suis allé voir le président d'alors à Montpellier pour lui proposer cette idée et il m'a répondu : « Ce n'est pas possible chez nous. En plus, vous êtes en poste à l'IUT ».

J'ai malgré tout mis en place avec le centre de calcul un système où le jeudi matin, je présentais

25. Collège d'enseignement général et professionnel, établissement d'enseignement public du Québec où est dispensé le premier niveau de l'enseignement supérieur.

une méthode, ensuite un technicien du centre de calcul montrait comment on pouvait la mettre en œuvre ; les participants avaient 15 jours pour utiliser la méthode sur leurs données et on discutait des résultats qu'ils avaient obtenus. Cela a duré un an ou deux à titre tout à fait gratuit pour moi, mais il y avait une forte demande de la part des biologistes de la Faculté des sciences, mais aussi des géographes de Paul Valéry, de chercheurs de l'INRA et enfin de l'environnement qui savaient qu'ils pouvaient venir nous demander de l'aide et on essayait de les former un peu.

C'est alors qu'il y eut la volonté de voir s'installer en un seul lieu le CIRAD, l'IRD, le CEMAGREF pour constituer un centre agronomique méditerranéen appelé ultérieurement « Agropolis ». L'INRA, avec Richard Tomassone qui était alors le directeur du département biométrie, a voulu accompagner ce mouvement et créer à Montpellier un laboratoire de biométrie et m'a proposé d'en prendre la direction pendant au moins huit ans. L'unité de biométrie est née en 1982.

Tomassone disait : « Je mettrai progressivement du personnel » ; l'école d'Agronomie disait : « Je donne les locaux » et les enseignants de mathématiques de l'école en feront partie. Moi, j'avais négocié avec la Faculté des sciences que les gens qui travaillaient avec moi seraient logés dans les locaux donnés par l'école d'Agronomie. Et c'est vrai qu'à ce moment-là, ça nous a donné des bourses, des moyens, une grande visibilité et des possibilités d'action importantes.

Ça a été une grande chance pour moi et pour la statistique à Montpellier. Il y a eu le recrutement de Xavier Millau venu de Toulouse, mais j'étais toujours en poste à l'IUT. En 1987, un poste s'est libéré au département de Mathématiques à la Faculté et j'ai été candidat avec le projet de développer des enseignements de statistique pour les biologistes. Il y avait quelques enseignements de statistique dans les filières de biologie et dans les DEA de biologie. Soit on me sollicitait pour intervenir deux heures dans un DEA de biologie – c'était ridicule –, soit l'enseignement était fait par des gens dont la statistique n'était pas la spécialité. J'ai expliqué au département de mathématiques qu'il fallait à tout prix investir ce domaine, qu'il y avait des postes à prendre ou à gagner, et que la formation des étudiants était en cause. On a alors pu recruter sur l'université parce que dès la création de l'unité de biométrie, il y avait l'idée de créer un DEA de statistique pour des mathématiciens et un DESS de méthodes statistiques appliquées à l'agronomie, l'agroalimentaire et la pharmacie, destiné plutôt à des biologistes.

L'unité de biométrie a donné une visibilité et une crédibilité à ce qu'on faisait, a entraîné des créations de postes, des bourses. On a géré beaucoup de bourses CIFRE dans le domaine de l'agroalimentaire, le champagne, la brasserie entre autres. Cela a ouvert beaucoup de débouchés aux étudiants qui sont passés par là et certains sont rentrés au CIRAD, à l'IRD, à l'INRAE.

Parti pour huit ans, je suis resté douze ans. Alain Berlinet m'a succédé, puis Jean-Pierre Vila et quelques années plus tard l'association avec l'université a cessé. L'unité de biométrie s'est restreinte à l'INRA et à l'école d'Agronomie, dénommée maintenant Institut Agro Montpellier²⁶. Mais les liens restent forts, les gens se connaissent, se retrouvent dans les séminaires.

GS : Le colloque organisé en 2019 en ton honneur s'intitulait, « Montpellier, berceau de la Data Science »²⁷. Quel regard portes-tu sur les relations entre l'informatique, la science des données et la statistique ? Plus particulièrement, penses-tu que l'informatique fasse progresser la statistique ?

YE : Les moyens de calcul permettent d'imaginer des choses qu'on ne pouvait pas imaginer auparavant. Ne serait-ce que l'estimation par noyau par exemple. Aujourd'hui, c'est courant,

26. L'unité de biométrie est devenue en 2009 l'UMR Mathématiques Informatique et STatistique pour l'Environnement et l'Agronomie (MISTEA).

27. C'est à l'occasion du deuxième séminaire franco-japonais organisé à Montpellier par Yves Escoufier en 1992 qu'est apparue la première occurrence de l'expression Data Science. Les actes du colloque en témoignent : Escoufier Y., B. Fichet, L. Lebart, C. Hayashi, N. Ohsumi, and Y. Baba (Eds.) (1995), *Data Science and Its Applications*, Academic Press, Tokyo, Japan.

mais on ne pouvait pas l'imaginer dans les années 1965-1970. Même si la théorie existait, les temps de calcul auraient été rédhibitoires.

C'est comme l'analyse factorielle au sens de Harman. Une des raisons pour lesquelles on mettait les spécificités dans la diagonale, c'était parce que diagonaliser une matrice était très compliqué. On cherchait un premier axe, un deuxième axe et pas plus.

Il est clair que les outils modifient la façon de penser et de réfléchir : le Bootstrap, le Jackknife, c'était inimaginable. Quand on mettait une après-midi pour faire une ACP, on n'allait pas la refaire 50 fois !

Il y a 30 ou 40 ans, on n'avait pas les outils qui nous permettaient d'imaginer de faire ces calculs.

Des chercheurs comme Kruskal, Carroll dont on parlait plus haut, avaient une pratique computationnelle bien plus performante que la nôtre, ce qui les a fait avancer, mais on s'est retrouvé sur le *multidimensional scaling* car en réalité on faisait la même chose.

En fait, c'est le problème de la place des données en statistique. J'ai toujours pensé que la statistique avançait en se confrontant à des données nouvelles qui nous amènent à créer des outils nouveaux. Et aujourd'hui, il n'en manque pas des données nouvelles et on a besoin de beaucoup de mathématiques et de mathématiques difficiles avec des algorithmes énormes que l'on ne pouvait même pas imaginer.

Après, il est assez naturel de prendre du recul pour justifier mathématiquement et pour construire informatiquement les solutions. C'est peut-être plus dur d'ailleurs.

GS : Comment réagis-tu au fait que des informaticiens spécialisés en apprentissage redécouvrent les vertus de l'ACP et d'autres méthodes familières aux statisticiens ?

YE : Je commencerai par dire que s'ils redécouvrent la méthode, c'est que la méthode est utile. Ils y arrivent par d'autres chemins, et cela signifie que ce que l'on cherche à obtenir dans une ACP les intéresse aussi. C'est donc un bon point pour la méthode. Il reste que « la redécouverte » peut interroger sur la formation donnée et sur les équipes de recherche constituées.

GS : La statistique se dissout-elle dans la *data science* ?

YE : Tout dépend de ce qu'on appelle la *data science*. Je pense que dans la *data science*, il y a à la fois la collecte des données, leur stockage, leur manipulation. Ensuite, il y a les traitements, donc de la statistique et aussi des approches « intelligence artificielle ».

Ma crainte serait que la *data science* ne devienne le domaine d'exercice de personnes qui n'ont aucune idée en statistique, qui pourraient seulement manipuler des données car on peut toujours en sortir quelque chose. La statistique, parce que c'est le domaine de l'aléatoire, fait germer le doute : « J'ai trouvé ce résultat, quelle confiance je peux lui faire ? ». Et c'est là que la statistique est importante.

Un *data scientist* pourrait être quelqu'un qui connaît la statistique et qui a une bonne pratique informatique. Alors, est-ce que l'inverse existe ? Quelqu'un qui connaît bien l'informatique et qui a une bonne pratique statistique ?

Cela pose la question de la formation. L'image que j'ai, c'est que dans les métiers du traitement de données, les gens arrivent soit par des filières où ils ont fait des mathématiques, des statistiques, des probabilités, soit par des filières où ils ont fait de l'informatique, et qu'il y a des

filières informatiques – du moins certaines – qui sont très loin des filières de mathématiques, avec des contenus très différents. Qu'enseigne-t-on de la statistique dans des filières *data science* où on parle surtout informatique ?

Cette diversité d'approches peut être un enrichissement mais il faut que ces approches-là se parlent. L'avantage d'une formation statistique, c'est quand même d'avoir une notion de l'aléatoire. Il est important de pouvoir dire : « Attention, cette méthode, sur ces données, elle nous donne ces résultats, mais sur des données un peu différentes elle peut donner des résultats un peu différents, voire très différents ».

GS : Tu as fait de l'ASU (l'ancêtre de la SFdS) une société à la fois savante et conviviale. Comment cela s'est-il passé ?

YE : Juste avant que je ne parte à Montréal, Bernard Charles avait fait en sorte que se crée à Montpellier une MIAGE – Maîtrise informatique appliquée à la gestion – dans laquelle je donnais un cours de statistique. Si bien qu'en février 1972, je suis revenu 15 jours pour faire mon cours de statistique à la MIAGE.

Pendant ce séjour, Richard Tomassone m'a invité à donner un séminaire à Orsay. J'y ai fait la connaissance de Jean-Pierre Pagès et de son schéma de dualité, comme je le disais plus haut. Jean-Pierre Pagès m'a fait connaître Georges Morlat qui animait les premières réunions de l'ASU avec des universitaires de province comme Marie-Jeanne Laurent-Duhamel, Henri Caussin et Michel Depaix qui cherchaient à faire savoir que la Statistique universitaire française ne se limitait pas à l'ISUP.

C'est ainsi que j'ai découvert cette association qui était en train de naître²⁸. Deux ans après mon retour de Montréal, j'ai organisé en 1975 une réunion de l'ASU à Montpellier où pour la première fois il y a eu des exposés scientifiques²⁹. Nous étions à peine 40.

Plusieurs congrès de l'ASU ont eu lieu ensuite à Montpellier, dont un à la Grande Motte en 1984.

GS : La crise sanitaire a révélé l'inculture scientifique de bon nombre de nos gouvernants, communicants, citoyens. Aurais-tu des pistes d'amélioration de ce qu'on appelle la littératie scientifique et plus particulièrement en statistique ?

YE : Quand j'étais président d'université, j'ai eu l'occasion de rencontrer un certain nombre d'hommes politiques. On se présente et quand je dis : « Je suis professeur de mathématiques », le plus souvent, la réponse était : « Oh ! Moi, vous savez, les mathématiques et les sciences, ce n'est pas mon fort ».

Que faut-il en penser ? Est-ce qu'en France, on a trop fait des mathématiques l'objet de sélection pour les écoles ? Je ne sais pas. Mais il y a manifestement une attitude vis-à-vis des mathématiques qui est regrettable. On considère que c'est la discipline clé et en même temps, ils disent presque en s'en vantant : « Moi, ça n'a jamais été mon fort ».

Qu'est-ce que ça veut dire ? Je crois qu'aujourd'hui, on ne peut plus vouloir dominer l'ensemble des disciplines, c'est très compliqué. Alors, à quel niveau, jusqu'où faut-il développer la culture scientifique ?

Récemment, je parlais avec un ami qui s'élevait contre le fait que des élèves de lycée, en ce

28. Les statuts de l'Association des Statisticiens Universitaires ont été déposés à Paris en février 1971.

29. C'est aussi à l'occasion de ces 7èmes Journées de Statistique que fut établie la tradition du banquet avec animation musicale.

moment, peuvent ne plus faire de maths en terminale. Il disait : « Mais tous les ingénieurs ont besoin de connaître les mathématiques, l'industrie a besoin de gens qui connaissent les mathématiques ». Bien sûr, mais à côté de cela, il y a la psycho, la sociologie, la communication, les relations humaines... Je m'interroge beaucoup sur le niveau de mathématiques qu'il est nécessaire d'y avoir.

Le problème essentiel, me semble-t-il, c'est que chacun sache où s'arrête son propre savoir et soit respectueux du savoir des autres. Après, est-ce qu'on pourrait me reprocher à moi par exemple, de ne pas avoir une culture suffisante en chimie ?

Je trouve que c'est une question très difficile. Quand je participais de près à l'Institut International de Statistique (IIS), cette question de littératie statistique est revenue plusieurs fois.

J'avais fait un exposé dans une réunion de l'IIS que j'avais intitulé : « Que faut-il enseigner et à qui ? ». J'avais présenté l'image ci-dessous qui donne différentes écritures de la moyenne telles qu'on peut les trouver dans les livres.

$$\frac{2+4+7}{3} \quad \sum_{i=1}^n \frac{x_i}{n} \quad \sum_{i=1}^k p_i x_i$$

$$\langle X_{1..n} \rangle_D \quad \int_{-\infty}^{+\infty} x f(x) dx \quad \int x dP$$

Figure 1 – Différentes écritures de la moyenne

On utilise l'une ou l'autre de ces écritures selon à qui on s'adresse. Et donc, cela veut dire qu'il y a des accès différents à des niveaux différents de compétences. Donc quel est le savoir minimal qu'il faut avoir ? Et ce savoir minimal est-il stable ? C'est vrai qu'aujourd'hui, avec tous les moyens de communication, on est abreuvé de chiffres : faut-il insister sur leur acquisition, leur interprétation ? Qu'y a-t-il derrière ? Qui valide une information ?

Comment faire pour intéresser les jeunes à la statistique ? Sylvette Maury qui dirigeait l'Institut de recherche sur l'enseignement des mathématiques (IREM) à Montpellier avait organisé en 1989 un concours pour les lycéens intitulé « À vos Stats »³⁰ parrainé par l'Insee. L'idée était de faire découvrir l'intérêt qu'il pouvait y avoir à savoir manipuler un peu les outils statistiques.

Il me semble qu'au-delà des techniques, ce qui est important avec la statistique, c'est de sensibiliser les gens à l'aléatoire. Qu'ils sachent qu'un phénomène va avoir plusieurs résultats possibles et que c'est naturel. Quand j'enseignais dans un certificat de biologie, en guise d'introduction, je prenais une craie et je leur disais : « Si je la lâche, elle tombe, c'est un phénomène qui n'est pas aléatoire. Mais en combien de morceaux va-t-elle se casser ? Ça, c'est aléatoire et vous ne pouvez pas me le dire avant, n'est-ce pas ? ».

Il faut prendre conscience de l'aléatoire, puis le décrire et le modéliser, et ensuite selon le nombre d'années d'étude, on va enrichir les moyens de description et de modélisation. Mais cette notion qu'il n'y a pas qu'une issue possible, c'est fondamental.

Note de la rédaction : L'ensemble des publications d'Yves Escoufier est disponible à l'adresse : <https://imag.umontpellier.fr/YvesEscoufier/>

30. Le concours *À vos Stats* a eu lieu jusqu'en 2004. Sous une forme différente, Eurostat et l'INSEE en France organisent tous les ans depuis 2018 une compétition européenne pour les jeunes collégiens et lycéens.

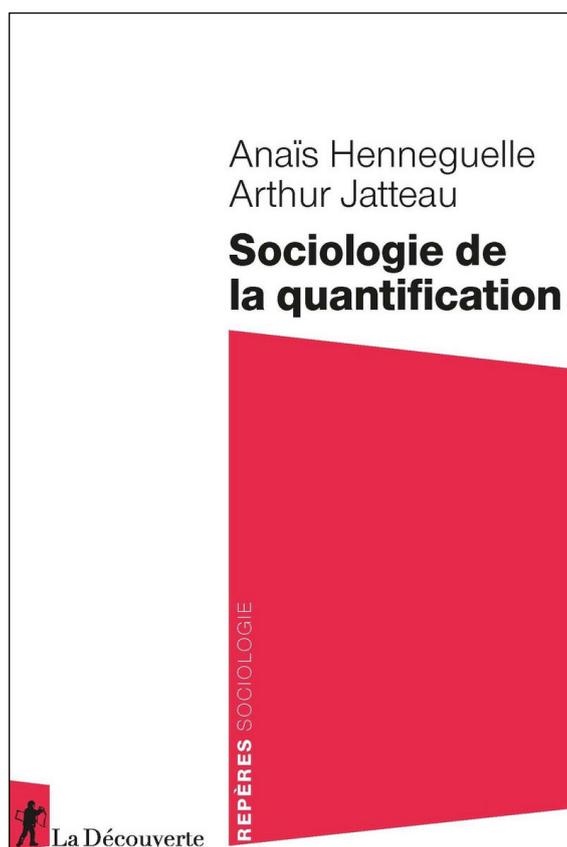
Sociologie de la quantification

de
Anaïs HENNEGUELLE et Arthur JATTEAU (2021)



Pauline ADAM¹

Université libre de Bruxelles (ULB), Recherche et études en politique internationale (REPI)
École normale supérieure (ENS), Centre Maurice Halbwachs (CMH)



Livre (128 pages)

Auteurs : Anaïs HENNEGUELLE et Arthur JATTEAU

Édition : La Découverte (Collection : Repères n°762) – 2021

ISBN : 9782348041853

1. pauline.adam@ulb.be

L'ouvrage *Sociologie de la quantification* relève largement le défi de présentation synthétique de la collection Repères. Ses auteures, Anaïs Henneguelle et Arthur Jatteau, sont à la fois économistes et sociologues et ont respectivement travaillé sur les statistiques pénitentiaires et les expérimentations aléatoires en économie. Iels proposent ici un guide pratique pour découvrir la sociologie de la quantification. Les différents chapitres de l'ouvrage reviennent sur l'histoire de la quantification (chapitre 1), le cadre théorique de l'analyse de la quantification (chapitre 2 et 5) et des études de cas (chapitre 3 et 4). Plusieurs encadrés didactiques parsèment la lecture et fournissent des éclairages bienvenus sur certaines notions, typologies, institutions ou encore types de statistiques. En complément, une annexe est proposée en ligne, donnant un peu plus de corps à l'exposé grâce à la présentation des différents « acteurs de la quantification ». L'ouvrage est dense mais pédagogique. Ainsi, les novices y trouveront des références incontournables et les initiés un grand résumé.

Le point de départ est le suivant : « les chiffres sont partout » et « occupent aujourd'hui une place prépondérante dans la société » (p. 4). De plus, ils sont bien trop souvent présentés comme des produits finis et leur construction n'est nullement interrogée. C'est ici tout l'intérêt de la sociologie de la quantification. Appréhendée telle une « boîte à outils » (p. 8), celle-ci est définie comme « l'étude des activités sociales de construction, d'élaboration et d'utilisation des chiffres » (p. 5).

La démarche d'Anaïs Henneguelle et d'Arthur Jatteau trahit de surcroît l'ambition d'institutionnaliser ce champ d'études. En effet, la publication même d'un ouvrage synthétique sur la sociologie de la quantification et la présentation de ses principales figures, entrées théoriques et objets d'études participent à rassembler différents travaux jusque-là dispersés entre différentes thématiques de recherche.

Les principales clés théoriques de cette « boîte à outils » proposées dans l'ouvrage sont directement empruntées à Alain Desrosières. Ingénieur statisticien et administrateur de l'Insee, il est présenté comme le fondateur de la sociohistoire de la quantification pour son travail réflexif qui a permis de repenser la production statistique (encadré p. 7).

Première clé théorique empruntée à Alain Desrosières et développée au fil de l'ouvrage : le lien intime entre **statistiques et construction de l'État**. L'idée étant que « les statistiques sont produites par l'appareil étatique autant qu'elles participent à sa construction » (p. 13). Au XVI^e siècle, le début de la centralisation et de l'unification des États accélère la systématisation et la légitimation du travail de quantification (p. 11). La statistique devient par la suite une science de l'État, fidèle à son étymologie (chapitre 1). Elle joue un rôle différent en fonction des époques (à des fins fiscales et militaires dès l'Antiquité, pour lutter contre le chômage dans les années 1930) et des types d'État (pour développer le libre-échange sous un État libéral ou mettre en place des lois et assurances sociales sous un État-providence, voir encadré pp. 14-15).

Deuxième clé théorique de la « boîte à outils » et directement reliée à la première : celle du chiffre comme **outil de preuve et de gouvernement** (chapitre 2). Discipline scientifique, la statistique est un outil de savoir, et en même temps une science de l'État et donc un outil de pouvoir. Les chiffres nous fournissent des informations sur notre environnement et ont également un effet direct sur l'organisation de nos sociétés. Ainsi, les chiffres n'établissent pas simplement des constats objectifs, ils transforment le monde social. L'approche d'Alain Desrosières est associée à celle d'Alain Supiot qui qualifie notre ère de « gouvernance par les nombres » (chapitre 5). Cette thèse repose tout d'abord sur la valeur et le pouvoir construit des chiffres. L'argument chiffré a progressivement été présenté comme indiscutable par sa rigueur, sa scientificité et son objectivité. Depuis la seconde moitié du XX^e siècle, le chiffre est également devenu indispensable dans l'organisation des institutions alors soumises à une « obligation de résultat » (p. 81). Les indicateurs servent à piloter l'action publique aux côtés des nouvelles

modalités du *New Public Management* qui importent des techniques de gestion du monde privé dans les administrations publiques. Les techniques de mise en concurrence du *benchmarking* et les nouvelles *big data*, ressources inépuisables, permettent de « contrôler et transformer les conduites des individus » (p. 83).

Enfin, l'ouvrage ne serait pas un guide pratique de la sociologie de la quantification sans la formule indispensable d'Alain Desrosières : « **quantifier c'est convenir puis mesurer** ». Penser les chiffres, c'est penser l'étape de la catégorisation, de la classification, du codage, de la collecte et de la comparaison. Pour compter encore faut-il savoir quoi. Interroger la définition retenue de l'individu qui doit être comptabilisé nous amène à repenser les catégories, à revenir au processus de définition des conventions d'équivalence, aux rapports de force en présence, mais également à la compréhension qui en est faite par les différents acteurs impliqués dans la collecte des données. Si la quantification appelle généralement à la comparaison, les définitions évoluent en fonction des cadres et des échelles dans lesquelles elles s'appliquent. Compter revient également à mesurer. Or, les opérations de codage et de collecte ne vont pas de soi. Elles s'inscrivent dans des contextes particuliers et sont réalisées par des acteurs bien souvent contraints au « bricolage » (p. 45).

Ces clés théoriques sont directement mobilisées dans l'ouvrage pour étudier des indicateurs médiatisés et remettre en cause leur caractère naturel et intemporel. Le Produit intérieur brut (PIB), l'indice des prix à la consommation et la mesure du chômage sur le plan économique (chapitre 3), la mesure de la population, la mesure de l'opinion et les crimes et délits sur le plan démographique et social (chapitre 4). À titre d'exemple, les chiffres de la délinquance sont interrogés via leurs sources. Celles-ci ont progressivement changé pour passer des institutions judiciaires au Ministère de l'intérieur. Cette évolution témoigne d'une acceptation différente de la délinquance, de ce qu'il s'agit de mesurer, de l'utilisation des chiffres et de leurs effets. Ainsi, les chiffres de la délinquance sont révélateurs des activités des institutions qui les produisent plus que d'une « réalité criminelle ».

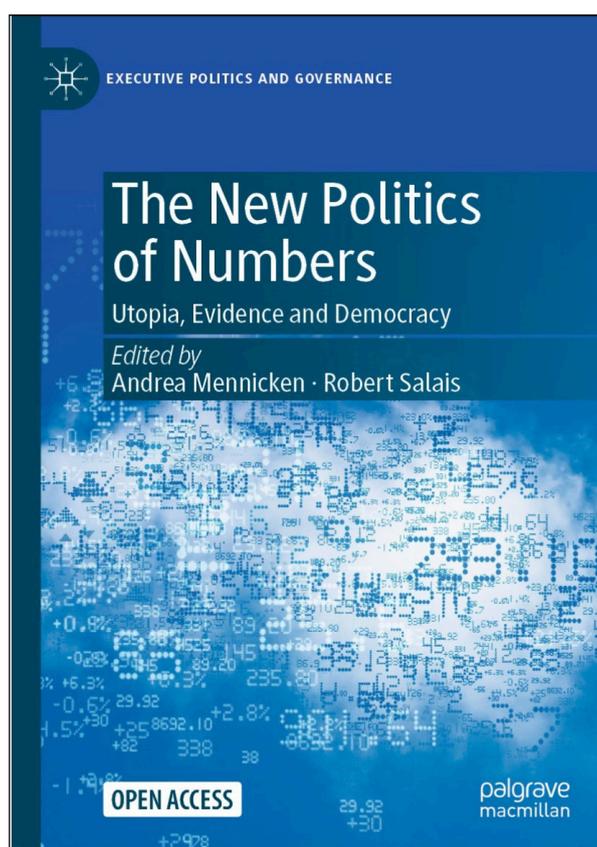
Il ne s'agit pas pour autant de renoncer à la mise en nombre et au pouvoir « émancipateur » des chiffres (p. 105). L'ouvrage se termine par la proposition d'un programme militant. Il appelle les acteurs sociaux à se réappropriier eux-mêmes les chiffres sur la base des idées lancées par le « statactivisme », néologisme d'Isabelle Bruno, d'Emmanuel Didier et de Julien Prévieux qui défendent l'idée d'une statistique comme arme critique. Ainsi, « ce qui n'est pas compté ne compte pas : dès lors, si quantifier revient bien à visibiliser, il ne faut pas hésiter à « faire du chiffre » ... pour plaider une juste cause ! » (p. 107). Il ne nous reste alors plus qu'à faire nôtre cette proposition.

The New Politics of Numbers: Utopia, Evidence and Democracy

de
Andrea MENNICKEN et Robert SALAIS (eds.) (2022)



Harro MAAS¹
Centre Walras-Pareto, UNIL



Livre (497 pages)

Éditeurs : Andrea MENNICKEN et Robert SALAIS

Édition : Palgrave Macmillan – 2022

ISBN : 978-3-030-78200-9

DOI : <https://doi.org/10.1007/978-3-030-78201-6>

1. harro.maas@unil.ch

After reading *The New Politics of Numbers: Utopia, Evidence, and Democracy*, edited by Andrea Mennicken and Robert Salais, it is just impossible to open a newspaper or news site without being reminded of the themes addressed in this volume. This can be about an upward adjustment of the number of unemployed from about 250,000 to 350,000 by the Dutch statistical office CBS to comply with new EU regulations.² Or a plea by (again Dutch) epidemiologists for competition between models to forecast the course of the covid pandemic to thus try to look “under the hood” of the official and for privacy reasons unpublished model and to promote an open debate about its assumptions and robustness.³ Or about a very interesting longread in *The Guardian* in which Ben Rawlence examined the damaging effects of the current climate crisis on the way of living of the Sami in the most northern province of Norway, Finnmark, where the changing natural conditions pit the state’s vision of reindeer as “a useful export” of a region otherwise considered as unproductive against the significance of reindeer to the Sami for whom reindeer represent not only economic, but also cultural and symbolic value. As Rawlence quotes one of his interviewees: “Reindeer are life. They are everything. Without reindeer, we die.”⁴

The fourteen essays in this volume show the reader that the making of numerical, quantified knowledge is not just a problem of objective knowledge, but—as famously phrased by Shapin and Schaffer in *Leviathan and the Airpump* – a problem of social order. Numbers, the quantification of the social, deal with problems of power and control, the imposition and negotiation of classifications, the social valuation and evaluation of past, present and future, and with questions of agency which affect, indeed, life and death.

The impetus for this volume came from a working group on social quantification at the *Wissenschaftskolleg* zu Berlin in 2014 that was continued at various conferences and at the Nantes Institute for Advanced Study (IAS-Nantes). In their introduction Mennicken and Salais present this collaboration as a growing mutual curiosity of two strands of research into the processes and consequences of quantification of the social, the one motivated by Foucauldian themes of power and (self)control that were taken up in Britain by historians and sociologists of accounting at the London School of Economics and elsewhere, most notably the late Anthony Hopwood, Peter Miller, and Michael Power, and that is associated with the journal *Accounting, Organizations, and Society* of which Mennicken is the current editor. The second strand of research will be better known to readers of this journal and is associated with sociologists and economists such as Luc Boltanski, Laurent Thévenot, and Mennicken’s co-editor, Robert Salais. While this strand of research stands in a complex relation with the sociological tradition of Pierre Bourdieu and the actor-network theory of Bruno Latour, Michel Callon and Madeleine Akrich at the *École des Mines*, the details of which need not concern us here, this French-based strand of research took its inspiration from the work of the late sociologist and INSEE statistician Alain Desrosières, whose monumental historical work on social quantification, *La politique des grands nombres* (1990), opened up an extremely fruitful field of research that examined, as Lorraine Daston wrote in her essay review of the English translation of 1998, “why statistics tend not only to describe the world, but to change it” (Daston 2000). In France, Desrosières canvassed this field in collaboration with Boltanski, Thévenot, Salais, but importantly also with historians of science who had been part of the highly successful Bielefeld working group of the early 1980s on the probabilistic revolution of which Daston, Theodore Porter and Mary Morgan, amongst others were part. Desrosières’ emphasis on the conventional basis of social quantification with real consequences aligned well with Boltanski’s earlier work on the French statistical classification of *cadres* in which he showed how a statistical classification transformed into a social class.⁵

2. <https://nos.nl/artikel/2413776-het-cbs-heeft-meer-dan-100-000-nieuwe-werklozen-gevonden-hoe-kan-dat>, accessed 20-1-2022.

3. <https://www.nrc.nl/nieuws/2022/01/19/oproep-modelleurs-nederland-heeft-recht-op-een-second-opinion-a4081121>, accessed 20-1-2022

4. <https://www.theguardian.com/news/2022/jan/20/norway-arctic-circle-trees-sami-reindeer-global-heating>, accessed 20-1-2022

5. Boltanski published his work in the same period that sociologists and political scientists in Eastern-Europe questioned to what extent the soviet administrative category of *nomenklatura* similarly evolved into a separate social class. See especially György Konrád and Iván Szelényi, *The intellectuals on the road to class power*, Harcourt, 1979.

The French strand of research that emerged from this collaboration is currently referred to as the economics or theory of conventions. It stresses the investment in form (as in the example of the *cadres*), the material differences between different spheres of valuation, and the agency of actors in the construction, use and counter-use of numbers. This agency of actors, and its concomitant possibility of resistance to social quantification, is an important distinguishing characteristic with Foucauldian approaches.

Against this background, it seems fit that the book is dedicated to the memory of the two founding representatives of the Foucauldian and convention approach to numbers, Anthony Hopwood and Alain Desrosières. But the explanation of the title of this volume by the editors may come as a surprise as it echoes William Alonzo and Paul Starr's 1989 edited volume *The Politics of Numbers* (without subtitle) in the Census series of the Russell Sage Foundation. Built around the history and politics of the American Census, the book was completely centered on the American situation and on the peculiar status of the American Census in American politics (the assignments of seats in the House of Representatives and the Electoral College) which to this day is still enmeshed with questions of racism and exclusion, as brought out so well in Gunnar Myrdall's *An American Dilemma: the Negro Problem and American Democracy of 1944*.⁶ Far from concealing the entanglements of statistics and politics, the contributions to Alonzo and Starr's edited volume presented the assumptions, choices, and resulting tensions in American (census) statistics as issues which at the end of the day could be resolved by an ordered technical discussion by statisticians in the service of the common good. Or, as Mennicken and Salais note in their introduction, the *Politics of Numbers* showed how vital government statistics is for a democratic polity. In this new collection, the optimistic tone of Alonzo and Starr's volume has vanished. While Alonzo and Starr viewed numbers as a support of governing by democracy, Mennicken and Salais emphasize in contrast that the current politics of numbers replaces government by democracy with government by numbers.

The meaning of this shift is explored in fourteen essays grouped under the headings of utopia, evidence, and democracy. The first group comprises essays by Martine Mespoulet, Tong Lam, Uwe Vormbusch, and Ota de Leonardis that all one way or another address utopian dreams to improve society through quantification. In a sense this theme runs as a basso continuo through the essays of the two other groups of essays as well. The four essays in the section *The Politics of Evidence* discuss the role of official statistical infrastructures in France in terms of knowledge and representation of society (Thomas Amossé) and in the transformation of education politics (Corine Eyraud), and on certification standards (Laurent Thévenot) and statistical indicators (Ousmane Oumarou Sidibé) as instruments of international politics and 'global world making'. The section *Voicing for Democracy* contains essays on British prison ratings and rankings (Andrei Guter-Sandu and Andrea Mennicken), price and (un)employment statistics (Boris Samuel and Robert Salais respectively), and the fiction of a history of purely qualitative sociology (Emmanuel Didier). With a clear nod to Albert Hirschman's magnificent *Exit, Voice, Loyalty* these essays deconstruct received tenets on the immutability of statistics for policy purposes, and exemplify citizens' and scientists' responses to declining trust in the statistics on offer from, as Hirschman had it, firms, organizations, and states. Peter Miller closes off this collection of essays with a reflection on the history and political function of a number that has become *the* most discussed number since the outbreak of the current pandemic: R, the reproduction ratio of the covid virus. Miller shows how this number emerged from historical debates on demographic modeling in the interwar period and then traveled to the field of epidemiology. Miller argues that R is not so much a number but a focal point in a modeling exercise that made R a mediating instrument in the machinery of current political decision making, with serious consequences on the health, wealth and wellbeing of the polity as we experience presently on a daily basis.

6. How insulated this book was on the American situation can also be seen from its remark in the introduction that "no social scientist has yet attempted any systematic comparative analysis of the organization and politics of the statistical process," (p. 25) whereas the Bielefeld project had already published some of its major results, and Theodore Porter's *The Rise of Statistical Thinking* was in the process of being published.

The comparison between a Foucauldian approach to quantification and the economics of conventions works well through the volume, though for some of the stories told they read a bit like overkill. Statistical classifications and indicators may make appear, disappear or redefine social phenomena such as poverty and unemployment, as Desrosières has shown so well, but a conventions 'light' approach seems sufficient to this reader to argue such cases. Salais beautifully shows how the shift from the measurement of unemployment to that of employment is embedded in shifting theoretical discourses in economics that moved away from Keynesian politics. The adverse effect of development politics by indicators comes out well from Sidibé's very interesting inside account of the measurement of the effects of development aid on socio-economic development of Mali by indicators, which leads to the strategic gaming of the indicators just as in the case of French education politics discussed by Corine Eyraud and British prison ratings and rankings discussed by Guter-Sandu and Mennicken. Thévenot mobilizes both the Foucauldian and conventions approach in his discussion of certification in the international palm oil industry to argue for the reductive forces of such technologies of standardization that straightjacket democratic discourses about different value systems into a measuring system fit for the market. In Thévenot's case study, little is left from Montesquieu's utopian *doux commerce* thesis, in which the standardization of measures was seen as a liberating force from vested interests and power relations (think of Ken Alders' book on the standard meter), and international certification becomes a destructive instrument used by international Palm Oil Companies against smallholders. In a similar vein does Boris Samuel discuss the social movement in Guadeloupe starting with local businesses against rent-seeking (or *pwofitasyon* in creole – outrageous exploitation) by businesses, a phenomenon that did not appear in the price statistics provided by the local INSEE, but that via a combination of social pressure (strikes) and local expertise became the subject of an official fact-finding mission. The chapter shows how different styles of calculation unveil different realities and as in Thévenot's case, points to the difficulties local actors in a minority position experience in using calculative strategies against vested institutional and business interests.

How utopian trust in numbers turned dystopic comes out strongest, perhaps, in the first three contributions to this volume that I will discuss at some greater length. Mespoulet starts with the Bolsheviks' high hopes for a new type of quantification, distinct from the 'bourgeois statistics' of the nineteenth century, that would fit to a new form of ordering society via planning instead of the market. Her reference to the experiences of the Great War reminds one of Otto Neurath's similar high hopes in planning, as witnessed from his pictorial statistics that showed the vanishing of business cycles during the war economy, hopes that led to the notorious planning debate between Neurath and Ludwig von Mises. Lenin pitched two different styles of quantification against one another: statistics and accounting. No longer should statistics merely register production decisions in the market, as bourgeois statisticians were doing, but the administrative scaffolding of the production processes – that is planning through accounting – should make markets oblivious. Controlling business accounts should produce the data that, Mespoulet quotes Lenin, "*the state needs to know now, today.*" Mespoulet explores how prerevolutionary statisticians tried to survive by accommodating their received style of quantification with accounting practices that promised planned historical change, but in the end became fully dependent on the whims of Stalin and the communist party.

Mespoulet's picture of Bolshevik utopian dreams turned oppressive through state controlled administrative procedures is contrasted with two contemporary cases, the first, the nowadays much discussed, and feared, Chinese Social Credit System, the second, the self-quantification movement known as the Quantified Self (QS). Lam investigates this Social Credit System as a centralized apparatus of authoritarian control of individual behavior, Vormbusch the Quantified Self Movement as a self-proclaimed collective movement of self-control through self-measurement, that seemingly find themselves at opposite ends of a spectrum running from state-control to self-governance. Both contributions show that the distinctions and similarities

between both forms of governance are more complex, nuanced and (unfortunately?) closer to one another than one might think.

Lam explains that the Social Credit System is in fact a public-private collaboration between the Chinese State (or rather the Communist Party) and private companies such as Alibaba and Tencent. Individual data are harvested to control, regulate, and engineer individual conduct through a system of merit and demerit points that affect people's access to public and private resources and services such as bullet-trains, mortgages, education, but also matching proposals in dating apps. Lam situates this Social Credit System against systems of individual surveillance that were in fact put in place by the National Party in the interwar period to control the conduct of party members. After the Communist Party came to power, this system was extended under Mao to cover the conduct of all urban citizens. Starting at child age, a system of personal files aimed to track and trace a person's character, attitudes, performance and social relationships. There is a limit of course to what Lam can cover in the space of one article, but an even longer durée would bring us to the early Buddhist and Confucian Ledgers of Merit and Demerit, so well discussed for the late Ming period by Cynthia Brokaw.⁷ Under the banner of the Social Credit System, the Communist Party now aspires to put a similar system in place to register and monitor the behavior of all Chinese citizens and like a FICO score to aggregate the merit and demerit points of an individual in one number. Lam rightly notes that the Credit systems developed by companies like Tencent and Alibaba (Sesame Credit) are more like private company fidelity point systems to nudge behavior with a much smaller social reach and that a unified state-led credit system is so far only a (utopian/dystopian) promise. Instead of one Social Credit System, China is rather implementing a convolute of different systems. Lam argues that the Social Credit System should be seen as "*part and parcel of a state-led neoliberal model of development and governance*" that "*seeks to economize and financialize*" the social world. Characterizing China's totalitarian aspirations as neoliberal seems to this reader however a missed opportunity to discuss the Social Credit System from the perspective of the economy of conventions in which a totalitarian state aims to capture different value systems in one number.⁸ The more so, because Lam situates the Social Credit System in a long history of Chinese social and individual state and party control which sits difficult with the discourse of neoliberalism. The joint circulation of monetary and moral (celestial) worth was already part and parcel of the Ledger system.

The contribution of Vormbusch explores the seemingly opposite of state or large enterprise led surveillance projects of individual citizens. He takes the Quantified Self movement, linked to the Californian based journal *Wired* and its former contributing editor Gary Wolf, as a laboratory that explores new ways of self-making by self-measurement, or as the movement's slogan goes "self-knowledge through numbers." The movement meets regularly, amongst others in show and tell sessions in which self-measurers show how they use and personalize technical artifacts, ranging from the good old cycle to measure Watts, to wristbands like Fitbit, and a plethora of performance measuring apps to fit to their personal needs of self-measurement or as Vormbusch says, *leibschreiben*: 'writing the body'. Vormbusch usefully problematizes the question of agency in the use of apps which one way or another rely on standardized practices of recording and visualization, even when these practices may not be as fixed as in other domains. Clearly, and as examined at length in this article, self-quantification through technical artifacts changes the attitude of an individual to his or her body. But technical devices as fitbit or moodtracker apps such as *Mood 24/7 of Track Your Happiness*, also show the dependence

7. The Ledgers (or tables) of Merit and Demerit of the late Ming period was firmly rooted in Buddhist and Daoist traditions. They became a popular means to measure an individual's karma and implied socio-economic standing. See Cynthia Joanne Brokaw, *The Ledgers of Merit and Demerit: Social Change and Moral Order in Late Imperial China* (Princeton: Princeton University Press, 1991); Cynthia Brokaw, "Yüan Huang (1533-1606) and The Ledgers of Merit and Demerit", *Harvard Journal of Asiatic Studies*, 1987, pp. 137-95.

8. The design of the Social Credit System is inspired by the US based system of FICO scores by credit companies but aims to expand these scores beyond the sphere of the market. On the US, see Josh Lauer, *Creditworthy: A History of Consumer Surveillance and Financial Identity in America*, Columbia Studies in the History of U.S. Capitalism (New York: Columbia University Press, 2017).

of individual experiences on business classifications and algorithms as most visible in his discussion of apps that monitor the sleeping behavior of babies to uncertain young parents and provide them preformed diagnoses. To quote Vormbusch: *"quantifying the self is as much about the self as a subject competing in markets, as it is about the cultural indeterminacy of today's forms of living."* If anything counts as the construction of the neoliberal self, we can find it here.

In all, this collection of essays makes a very useful contribution to the existing literature on social quantification that will be of use to sociologists, historians and philosophers of quantification. It creates a largely successful dialogue between the British Foucauldian and French convention approach to social quantification. The development of the French convention approach from Desrosières's work has been largely discussed in a francophone environment, and with this volume, as Wendy Espeland rightly notes in her preface to this collection, received the attention in the Anglosaxon world it deserves.