Statistique et société
avril 2017

Volume 5, Numéro 1

Interrogations sur la statistique publique

www.statistique-et-societe.fr

Sommaire
Statistique et société

Volume 5, Numéro 1

Sommaire
Statistique et Société

Volume 5, Numéro 1

Statistique et société

Magazine quadrimestriel publié par la Société française de statistique.
Le but de Statistique et société est de montrer d'une manière attrayante et qui invite à la réflexion l'utilisation pratique de la statistique dans tous les domaines de la vie, et 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.

ISSN 2269-0271

Éditorial

Emmanuel DIDIER

Rédacteur en chef de Statistique et Société

Cher Lecteur,

A l'heure où vous lirez ce texte, nous aurons un nouveau Président de la République, choisi après une campagne électorale très inhabituelle pendant laquelle se sont exprimés des bouleversements sociaux d'une profondeur rarement atteinte. La statistique a joué un rôle crucial dans ces événements, nous y reviendrons souvent.

Le numéro que nous vous présentons ici aborde déjà cette question. En effet, nous avons choisi de traduire en Français un article de William Davies initialement publié en Grande-Bretagne arguant du fait que les transformations actuelles des modes de quantification démontrent la profondeur de la révolution sociale que nous traversons. En effet, selon lui, les statistiques, et spécialement les statistiques publiques, laissent de plus en plus souvent le public dubitatif voire carrément soupçonneux quant à leur validité, alors que le *Big Data* parvient à capturer son attention et sa confiance. L'argument de l'auteur est que la statistique était adaptée à un monde ancien, stable et catégoriel, né à l'époque des Lumières, alors que le *Big Data* convient bien mieux au monde contemporain, plus fluide ou labile, et en tout cas radicalement différent du précédent. Le fait que la statistique soit en proie au doute aujourd'hui montre que le monde des Lumières est en train de disparaître et se voit remplacé par de nouveaux modes de quantification plus adaptés aux besoins sociaux actuels – dont l'auteur prévient que le « populisme » politique fait partie.

L'argument est assez puissant et osé pour que nous ayons voulu le discuter dans un dossier, d'autant plus que l'auteur nous a indiqué très vite qu'il se sentait redevable des analyses d'Alain Desrosières.

Nous avons traduit la réaction de John Pullinger, le directeur général de l'Autorité Statistique du Royaume Uni, et demandé à Dominique Bureau, le Président de l'Autorité de la Statistique publique française ce qu'il en pense. Nous avons ajouté à ces prises de paroles institutionnelles un compte rendu de la 100e séance du Café de la Statistique qui portait, en fait, sur le même sujet, et qui a permis au public varié qui le fréquente d'exprimer son opinion à son tour.

Ce dossier est suivi d'un article méthodologique de Gilbert Saporta, professeur émérite de statistique au Conservatoire national des arts et métiers (CNAM), qui explique quelles sont selon lui les spécificités des outils statistiques utilisés dans un contexte de *Big Data*. Puis d'une libre intervention de Francis Laloë, statisticien, ancien directeur de recherche à l'Institut de recherches pour le développement (IRD), qui présente les difficulté du métier de statisticien sur le terrain, selon qu'il travaille en tête à tête avec un seul chercheur d'une autre discipline, ou qu'il est confronté à plusieurs interlocuteurs et donc plusieurs demandes disciplinaires différentes.

Pour finir, nous proposons une interview de Judith Rochfeld, spécialiste du droit de l'économie numérique, concernant les conséquences de la loi « Pour une République numérique » (déjà abordée dans le numéro d'automne 2016) pour les données personnelles.

Bonne lecture !

Emmanuel Didier

DOSSIER

Interrogations sur la statistique publique Introduction

La Rédaction

Deux phénomènes sociaux majeurs viennent percuter le monde de la statistique publique. Le premier est un courant de défiance vis-à-vis de toutes les expertises et de toutes les autorités, présent dans beaucoup de sociétés du monde occidental. Le second est la « datafication » du monde, l'essor de la quantification dans des domaines où elle n'avait jusque là que peu d'importance ; les statistiques publiques sont désormais noyées au milieu d'un océan d'autres informations chiffrées. Face à ces deux révolutions, la statistique publique peut-elle persévérer tranquillement dans son être ?

Dans les démocraties, la statistique publique joue un rôle essentiel. Ce rôle est double : aider les décideurs publics, éclairer l'ensemble de la société.

Ce rôle n'est pas toujours reconnu, il est parfois contesté. L'histoire récente est riche en polémiques à propos de tel ou tel indicateur : l'indice des prix, le taux de chômage, le produit intérieur brut... Mais jusqu'à présent l'institution statistique publique est parvenue à traverser les épisodes tourmentés. Dans l'Union Européenne, le début des années 2000 a vu l'émergence d'un corps d'institutions statistiques communautaires, appuyées sur les instituts nationaux, le tout formant un édifice apparemment solide. C'est de cette période que datent le « code de bonnes pratiques »[1] et les différentes « autorités statistiques »[2] qui doivent asseoir la confiance des citoyens dans la statistique publique.

William Davies, sociologue anglais, pense que cette fois-ci c'est différent. Dans un long article publié en janvier dernier dans le journal « The Guardian », il retrace l'histoire de la statistique publique depuis la Renaissance, et dépeint crûment les défis auxquels elle est actuellement confrontée. Ses conclusions sont loin d'être optimistes. Son article, intégralement traduit, constitue le premier volet de notre dossier.

En face de cette analyse, le lecteur trouvera les réactions de deux responsables institutionnels de la statistique en Europe : le « Statisticien national » britannique, John Pullinger, et le président de l'Autorité de la statistique publique française, Dominique Bureau. L'un et l'autre reconnaissent les nouveautés et les périls que l'environnement actuel de la statistique publique comporte, mais tous deux soulignent en regard les opportunités nouvelles, et, fonction oblige, affirment leur détermination à agir pour qu'une nouvelle fois l'institution sorte renforcée des difficultés qu'elle traverse.

1. Le « code de bonnes pratiques de la statistique européenne » contient les principes qui doivent régir le développement, la production et la diffusion des statistiques européennes. Il a été adopté en 2005 et complété en 2011. Il est consultable à l'adresse : http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-32-11-955
2. En France, l'Autorité de la statistique publique a été créée en 2009. Son équivalent au Royaume-Uni, « U.K. Statistics Authority » a été établie en 2007. Pour l'Union Européenne, un rôle analogue est assigné à l' « European Statistical Governance Advisory Board » (ESGAB) institué en 2008.

ARTICLE DU DOSSIER

Comment la statistique a perdu son pouvoir - et pourquoi nous devrions craindre ce qui va suivre

William DAVIES

Sociologue, économiste politique[1]

La capacité de la statistique à représenter le monde avec précision baisse. Un âge nouveau, de données massives contrôlées par des sociétés privées, prend le dessus et menace la démocratie.

En théorie, la statistique devrait aider à trancher des débats. Elle doit fournir des références stables, dont chacun – quelles que soient ses opinions - peut convenir. Pourtant, ces dernières années, l'inégale confiance dans la statistique est un des schismes aigus qui se sont ouverts dans des démocraties libérales occidentales. Peu avant l'élection présidentielle de novembre, une étude aux États-Unis a révélé que 68 % des partisans de Trump se méfient des données économiques publiées par l'administration fédérale. Au Royaume-Uni, une étude de l'Université de Cambridge et de YouGov sur les théories du complot a révélé que 55 % de la population croient que le gouvernement « cache le véritable nombre d'immigrés vivant ici ».

Plutôt que dissiper la controverse et la polarisation, il semble que la statistique les attise en réalité. L'aversion pour la statistique est devenue un des traits de la droite populiste, mettant les statisticiens et les économistes au premier rang des divers « experts » qui ont été ostensiblement rejetés par les électeurs en 2016. Non seulement la statistique est vue par beaucoup comme douteuse, mais elle leur apparaît en quelque sorte insultante ou arrogante. Ramener les questions sociales et économiques à des agrégats numériques et des moyennes semble à beaucoup violer la décence politique.

Nulle part ceci ne se manifeste avec plus d'éclat que pour l'immigration. Le thinktank « British Future » a examiné comment on peut argumenter au mieux en faveur de l'immigration et du multiculturalisme. Un de ses principaux constats est que les gens réagissent souvent avec chaleur face à des témoignages qualitatifs, tels que des histoires de migrants ou des photographies de diverses communautés. Mais des statistiques - spécialement celles qui portent sur les avantages supposés de l'immigration pour l'économie britannique - suscitent la réaction tout à fait opposée. Les gens pensent que les chiffres sont manipulés et détestent l'élitisme consistant à se référer à une mesure quantitative. Devant les évaluations officielles du nombre d'immigrants clandestins, la réaction habituelle est d'en rire. « British Future » a montré que mentionner l'effet positif de l'immigration sur le PIB, loin de plaider en sa faveur, peut en réalité y rendre les gens encore plus hostiles. Le PIB lui-même en vient à être vu comme le cheval de Troie d'un projet libéral élitiste. Les politiciens l'ont maintenant bien perçu et ont en général abandonné les discours sur l'immigration formulés en termes économiques.

1. NDR : Article original publié le 17 janvier 2017 dans le journal The Guardian sous le titre « How statistics lost their power – and why we should fear what comes next ». Traduction publiée ici avec l'autorisation de l'auteur et de l'éditeur, que nous remercions. Les intertitres ont été ajoutés par notre revue.

Tout cela représente un sérieux défi pour une démocratie libérale. Pour faire court, le gouvernement britannique - fonctionnaires, experts, conseillers et beaucoup de politiciens - croit vraiment que l'immigration est tout compte fait bonne pour l'économie. Le gouvernement britannique était convaincu que le Brexit était un mauvais choix. Le problème est que maintenant le gouvernement s'autocensure, de crainte de provoquer davantage le peuple.

C'est un dilemme malvenu. Ou bien l'État continue de se réclamer de ce qu'il considère valide et les sceptiques l'accusent de propagande, ou bien les politiciens et officiels sont confinés à dire ce qui est considéré plausible, intuitivement vrai, mais peut en fin de compte être inexact. Dans tous les cas, la politique se retrouve embourbée dans l'accusation de mensonge et de camouflage.

Cette perte de crédibilité des statistiques - et des experts qui les analysent - est au cœur de la crise qui est désormais désignée comme politique de la « post-vérité ». Dans ce monde nouveau et incertain, les attitudes envers l'expertise quantitative deviennent de plus en plus divisées. D'un côté, fonder la politique sur la statistique est élitiste, antidémocratique et imperméable à l'investissement émotionnel des gens dans leur communauté ou leur nation. Ce n'est qu'une manière de plus pour favoriser les gens à Londres, Washington ou Bruxelles qui cherchent à imposer leur vision du monde à tous les autres. Vue d'un autre côté, la statistique est tout le contraire d'élitiste. Elle permet aux journalistes, aux citoyens et aux politiciens de saisir la société dans son ensemble, non sur la base de l'anecdote, du sentiment ou du préjugé mais bien d'une façon qu'il est possible de valider. L'alternative à cette expertise quantitative est vraisemblablement encore moins la démocratie que l'attitude qui consisterait à lâcher la bride aux rédacteurs de tabloïds et aux démagogues pour fournir leur propre « vérité » sur ce qui se passe dans la société.

Peut-on s'affranchir de cette polarisation ? Devons-nous simplement choisir entre une politique de faits et une politique d'émotions, ou y a-t-il une autre façon de considérer la situation ? Une approche consiste à regarder la statistique à la lumière de son histoire. Essayons de la voir pour ce qu'elle est : ni vérité incontestable ni conspiration d'élite, mais plutôt un outil conçu pour simplifier la tâche de ceux qui gouvernent, pour le meilleur et pour le pire. Voyons, historiquement, le rôle crucial que la statistique a joué pour notre compréhension de l'état et du progrès des nations. Cela soulève la question préoccupante de savoir comment nous continuerons si peu que ce soit à avoir des idées partagées entre tous sur la société et le progrès collectif, si la statistique fait naufrage.

Une histoire tri-séculaire

Dans la seconde moitié du 17ème siècle, à la suite de conflits prolongés et sanglants, les dirigeants européens ont adopté un point de vue entièrement nouveau sur le rôle de ceux qui gouvernent : une approche centrée sur les tendances démographiques, rendue possible par la naissance de la statistique moderne. Depuis des temps immémoriaux, les recensements avaient permis de suivre la taille de la population, mais ils étaient coûteux et laborieux à effectuer et centrés sur les citoyens considérés comme politiquement importants (les propriétaires), plutôt que sur l'ensemble de la société. La statistique a offert quelque chose de tout à fait différent, transformant par là même la nature de la politique.

La statistique a été conçue pour donner une compréhension globale d'une population, plutôt que de simplement cerner les sources à valeur stratégique du pouvoir et de la richesse. Dans les premiers temps, ceci n'impliquait pas toujours de produire des chiffres. En Allemagne, par exemple (d'où nous vient le terme Statistik) le défi était de dresser la carte des douanes, institutions et lois disparates dans un empire fait de centaines de micro-états. Ce qui caractérisait cette connaissance en tant que statistique était sa nature holistique : il s'agissait de produire une

image d'ensemble de la nation. La statistique ferait pour les populations ce que la cartographie faisait pour le territoire.

L'inspiration des sciences naturelles était également importante. Grâce aux mesures standardisées et aux techniques mathématiques, la connaissance statistique pouvait être présentée comme objective, assez largement comme l'astronomie. Des démographes anglais pionniers, comme William Petty et John Graunt, ont adapté des techniques mathématiques pour évaluer l'évolution des populations, travaux pour lesquels ils avaient été recrutés par Oliver Cromwell et Charles II.

L'apparition à la fin du 17e siècle de conseillers gouvernementaux se réclamant de l'autorité scientifique, plutôt que du sens politique ou militaire, marque les origines de la culture « experte » maintenant si dénigrée par les populistes. Ces individus novateurs n'étaient ni de purs universitaires, ni des fonctionnaires, mais planaient quelque part entre les deux. C'étaient des amateurs enthousiastes qui proposaient une nouvelle façon de penser les populations, qui prônaient des ensembles et des faits objectifs. Par leur prouesse mathématique, ils pensaient calculer ce qui autrement aurait exigé un vaste recensement.

Il n'y avait initialement qu'un seul client pour ce type d'expertise, ce que montre le mot « statistique ». Seuls des États nationaux centralisés avaient la capacité de récolter les données pour de grandes populations de façon normalisée et eux seuls avaient un besoin crucial de telles données. Durant la deuxième moitié du 18^{e} siècle, des états européens se sont mis à collecter plus de statistiques d'une nature qui nous semble aujourd'hui familière. Portant le regard sur des populations nationales, les états se sont intéressés à toute une gamme de quantités : les naissances, les décès, les baptêmes, les mariages, les récoltes, les importations, les exportations, les fluctuations de prix. Ce qui auparavant aurait été enregistré localement et différemment pour chaque paroisse fut dès lors agrégé au niveau national.

De nouvelles techniques ont été développées pour représenter ces indicateurs, qui ont exploité les deux dimensions de la page, verticale et horizontale – disposant les données en tableaux ou matrices, comme les marchands l'avaient fait en développant des techniques comptables standardisées à la fin du 15^{e} siècle. Organiser les nombres en rangées et colonnes était une façon nouvelle et puissante de montrer les caractéristiques d'une société donnée. De grandes questions complexes pourraient maintenant être étudiées rien qu'en examinant des données disposées géométriquement sur une simple page.

Ces innovations constituaient un extraordinaire potentiel pour les gouvernements. En ramenant des populations diverses à des indicateurs spécifiques et les disposant dans des tableaux appropriés, les gouvernements n'avaient plus besoin de prendre en considération un plus grand détail local et historique. Bien sûr, d'un autre point de vue, cette cécité à la variabilité culturelle locale est précisément ce qui déplait dans la statistique et la rend potentiellement offensante. Sans se soucier de savoir si une nation donnée a une quelconque identité culturelle, les statisticiens feraient l'hypothèse d'une uniformité standard ou, selon certains, imposeraient cette uniformité.

On ne saurait capter statistiquement toutes les facettes d'une population donnée. Il y a toujours un choix implicite dans ce qui est inclus et dans ce qui est écarté ; et ce choix peut devenir en soi une question politique. Le fait que le PIB ne comptabilise que la valeur de travail rémunéré, excluant ainsi celui qui est traditionnellement fait au foyer par les femmes, en a fait depuis les années 1960 une cible de la critique féministe. En France, il est interdit de recueillir dans les recensements l'appartenance ethnique depuis 1978, au motif que de telles données pourraient être utilisées à des fins de politiques racistes. (Ce qui a pour effet secondaire de rendre beaucoup plus difficile de quantifier le racisme ordinaire sur le marché du travail).

Malgré ces critiques, l'aspiration à dépeindre une société dans sa globalité et de façon objective a relié divers idéaux progressistes à la statistique. L'image de la statistique comme science impartiale de la société n'est qu'une partie de l'histoire. L'autre partie est que des idéaux politiques puissants sont venus investir ces techniques : ceux d'une « politique fondée sur des preuves », de rationalité, de progrès ou d'une idée de nation fondée sur les faits, plutôt que sur des récits idéalisés.

Statistiques et progrès national

Depuis l'apogée des Lumières à la fin du 18e siècle, les libéraux et les républicains ont nourri l'immense espoir qu'une métrologie nationale conduirait à une politique plus rationnelle, ordonnée par des améliorations démontrables dans la vie sociale et économique. Benedict Anderson, grand théoricien du nationalisme, est célèbre pour avoir décrit les nations comme des « communautés imaginées », mais la statistique offre la promesse d'ancrer cette imagination dans quelque chose de tangible. Elle promet également de montrer sur quel chemin historique la nation est engagée : quel genre de progrès s'accomplit et à quelle vitesse ? Pour les libéraux des Lumières, qui voyaient les nations se suivre dans un même sens de l'histoire, ceci était central.

La France postrévolutionnaire a bien saisi le pouvoir qu'a la statistique pour révéler l'état de la nation. L'État jacobin a imposé un cadre totalement nouveau pour effectuer des mesures et collecter des données au niveau national. Le premier bureau officiel de statistique du monde fut créé à Paris en 1800. Une collecte uniformisée des données, supervisée par un corps central d'experts hautement instruits, faisait partie intégrante de l'idéal d'une république dirigée centralement, qui visait à établir une société égalitaire et unifiée.

Depuis la période des Lumières, la statistique a joué un rôle de plus en plus important dans la sphère publique, nourrissant le débat dans les médias, fournissant aux mouvements sociaux les arguments qu'ils pourraient utiliser. Au fil du temps, la production et l'analyse de telles données furent moins dominées par l'État. Les spécialistes universitaires des sciences humaines se sont mis à analyser des données pour leurs propres travaux, souvent sans lien avec la politique gouvernementale. À la fin du 19e siècle, des réformateurs comme Charles Booth à Londres et W.E.B. Du Bois à Philadelphie menaient leurs propres enquêtes pour analyser la pauvreté urbaine.

Pour bien voir combien la statistique a été mêlée aux notions de progrès national, considérons le cas du produit intérieur brut. Le PIB est une évaluation de la somme totale des dépenses de consommation des ménages nationaux, des dépenses publiques, des investissements et de la balance commerciale (le solde des exportations et des importations), qui se ramènent à un simple nombre. Ceci est diaboliquement difficile à faire correctement et les efforts pour calculer ce chiffre ont commencé, comme tant de techniques mathématiques, par une recherche plutôt marginale et un peu byzantine dans les années 1930. Ce n'est devenu une priorité politique nationale qu'avec la seconde guerre mondiale, lorsque les gouvernements ont dû vérifier si la population nationale produisait assez pour entretenir l'effort de guerre. Dans les décennies qui ont suivi, cet indicateur simple, bien que non exempt de critiques, a acquis un statut politique sacralisé, baromètre suprême de la compétence du gouvernement. Que le PIB monte ou baisse est maintenant vu comme l'indice que la société avance ou recule.

Ou encore, prenons l'exemple des sondages d'opinion, premier exemple d'innovation statistique advenue dans le secteur privé. Dans les années 1920, les statisticiens ont développé des méthodes pour construire un échantillon de personnes interrogées qui soit représentatif des attitudes du public dans son ensemble. Cette percée, qui a été d'abord réalisée par des analystes du marketing, a bientôt donné naissance aux sondages d'opinion. Cette nouvelle technique

est immédiatement devenue un objet de fascination publique et politique, à mesure que les médias rapportaient ce que cette nouvelle science nous enseigne de ce que « les femmes » ou « les Américains » ou « les travailleurs manuels » pensent du monde.

De nos jours, les défaillances des sondages sont décortiquées à l'infini : ceci, en partie en raison des espoirs énormes investis en eux depuis leurs origines. C'est seulement dans la mesure où nous croyons à la démocratie de masse que nous sommes si fascinés ou inquiétés par ce que le public pense. Or, c'est essentiellement grâce à la statistique plutôt qu'aux institutions démocratiques en tant que telles, que nous pouvons savoir ce que le public pense sur des questions spécifiques. Nous ne soupçonnons pas combien notre sens de « l'intérêt public » est enraciné dans le calcul expert, bien plus que dans les institutions démocratiques.

À mesure que les indicateurs de santé, de prospérité, d'égalité, d'opinion et de qualité de la vie en sont venus à nous dire qui nous sommes collectivement et si les choses s'améliorent ou non, les politiciens se sont lourdement appuyés sur la statistique pour soutenir leur autorité. Souvent, trop lourdement, tirant trop loin la preuve, interprétant des données de façon trop peu rigoureuse, afin de servir leur cause. Mais ceci est une conséquence fâcheuse mais inévitable de la prévalence des chiffres dans la vie publique et ne justifie pas pour autant les rejets définitifs de l'expertise dont nous avons été récemment témoins.

À bien des égards, l'attaque populiste contemporaine envers « les experts » procède du même ressentiment que celle qui s'exerce à l'encontre des représentants élus. À force de parler de la société globalement, de diriger l'économie globalement, les politiciens tout comme les technocrates sont perçus comme ayant « perdu le contact » avec ce qu'éprouve un simple citoyen pour lui-même. Tant les statisticiens que des politiciens sont tombés dans le piège d' « un regard étatique », pour prendre une expression du penseur anarchiste James C Scott. Parler scientifiquement de la nation - par exemple en termes de macroéconomie - est une insulte à ceux qui préféreraient fonder leur sens de la nation sur la mémoire et sur le récit, et qui en ont assez d'entendre que leur « communauté imaginée » n'existe pas.

D'un autre côté, la statistique (tout comme les élus) s'est convenablement acquittée de la production d'un discours public crédible, pendant des décennies si ce n'est pendant des siècles. Qu'est-ce qui a changé ?

Crise de la statistique

Cette crise de la statistique n'est pas tout à fait aussi soudaine qu'on pourrait le croire. Pendant environ 450 ans, le grand exploit des statisticiens a été de réduire la complexité et la fluidité de populations nationales à des faits et chiffres utilisables, compréhensibles. Or, durant les décennies récentes, le monde a changé radicalement, grâce à la politique culturelle apparue dans les années 1960 et à la transformation de l'économie mondiale qui a suivi peu après. Les statisticiens ne semblent pas avoir suivi le rythme de ces changements. Les définitions et nomenclatures statistiques d'antan sont mises en question par des identités, des attitudes et des trajectoires économiques plus labiles. Les efforts pour représenter les changements démographiques, sociaux et économiques en termes d'indicateurs simples et bien reconnus perdent leur légitimité.

Regardez le changement dans la géographie politique et économique des États-nations ces 40 dernières années. Les statistiques qui dominent le débat politique sont essentiellement des statistiques de niveau national : pauvreté, chômage, PIB, balance migratoire. Mais la géographie du capitalisme a tiré dans des directions plutôt différentes. À l'évidence, la mondialisation n'a pas aboli la géographie. Souvent, la localisation de l'activité économique est même devenue plus importante encore, exacerbant l'inégalité entre les territoires qui réussissent (comme

Londres ou San Francisco) et ceux en déclin (comme le nord-est de l'Angleterre ou la « ceinture de rouille » des États-Unis). Les unités géographiques clés ne sont plus des États-nations. Ce sont plutôt les villes, les régions ou les périphéries urbaines qui montent ou dépérissent.

L'idéal cher aux Lumières, d'une nation définie comme communauté identifiée, réunie par un cadre commun pour le mesurer, est de moins en moins soutenable. Si vous vivez dans une de ces villes des vallées galloises où autrefois l'emploi reposait sur la fabrication de l'acier ou sur la mine, entendre les politiciens dire que « l'économie se porte bien » ne peut qu'accroître encore le ressentiment. De ce point de vue, le terme « le PIB » ne saurait rien porter d'intelligible ni de crédible.

Quand on utilise la macroéconomie comme argument politique, on sous-entend que les pertes dans une partie du pays sont compensées par des gains réalisés ailleurs. Des indicateurs nationaux fortement médiatisés, comme le PIB et l'inflation, dissimulent toutes sortes de gains et de pertes localisés qui sont moins couramment évoqués par les politiciens nationaux. L'immigration peut être bonne pour l'économie en général mais ceci ne signifie pas qu'il n'y ait aucun coût local. Ainsi quand les politiciens se réclament d'indicateurs nationaux pour défendre leur point de vue, ils supposent implicitement un certain esprit de sacrifice patriotique mutuel chez les électeurs : vous êtes peut-être perdant à présent, mais la prochaine fois vous pourriez y gagner. Mais que se passerait-il si la chance ne tournait jamais ? Si la même ville ou la même région gagnent encore et toujours, tandis que d'autres sont toujours perdantes ? Sur quel principe de donnant-donnant ceci est-il justifié ?

En Europe, l'union monétaire a exacerbé ce problème. Les indicateurs que la Banque centrale européenne prend en compte, par exemple, représentent un demi-milliard de personnes. La BCE s'occupe de l'inflation ou du taux de chômage dans la zone Euro comme si c'était un territoire unique et homogène, tandis que le destin économique des citoyens européens s'éparpille dans des directions différentes, selon la région, la ville ou le quartier où ils vivent. La connaissance officielle est de plus en plus étrangère à l'expérience vécue, au point de cesser tout simplement d'être appropriée et crédible.

Favoriser le niveau national comme échelle naturelle d'analyse est un point de vue statistique qu'ont érodé des années de changements économiques. Les classifications forment un autre parti-pris, de plus en plus problématique. Une bonne partie du travail du statisticien consiste à classer les gens en les rangeant dans toutes sortes de boîtes qu'il a créées : employé ou chômeur, marié ou non, favorable ou opposé à l'Europe. Tant que les gens peuvent être classés dans de telles catégories, il reste possible de cerner le poids d'une catégorie donnée dans la population.

Ceci peut impliquer des choix quelque peu réducteurs. Pour être compté comme chômeur, par exemple, un individu doit répondre à une enquête qu'il est involontairement sans emploi, même si cela peut être en réalité plus compliqué. À tout moment, beaucoup entrent ou sortent de l'emploi pour des raisons qui pourraient avoir autant à faire avec la santé ou avec des nécessités familiales qu'avec les conditions du marché du travail. Mais grâce à cette simplification, il devient possible d'estimer un taux de chômage pour l'ensemble de la population.

Or, voici un problème ! Que se passe-t-il si beaucoup de questions de notre époque ne trouvent pas leurs réponses dans le nombre de personnes correspondant à une définition, mais dans l'intensité avec laquelle elles y sont affectées ? Le chômage en est un exemple. Le fait que la Grande-Bretagne ait traversé la Grande Récession de 2008-13 sans que le chômage augmente sensiblement est généralement reconnu comme un succès. Mais mettre l'accent sur « le chômage » a masqué l'accroissement du sous-emploi, c'est-à-dire le fait que des gens n'aient pas assez de travail, en quantité, ou soient employés au-dessous de leur niveau de qualification.

Ceci représente actuellement autour de 6 % de la force de travail « employée ». Et puis, il y a la hausse de la main-d'œuvre indépendante, pour laquelle la distinction entre « employé » et « involontairement sans emploi » n'a pas grand sens.

Ce n'est pas ici une critique d'organismes tels que l'Office national de statistique (ONS), qui produit bien des données sur le sous-emploi. Mais tant que les politiciens continueront à répondre aux critiques en se référant au taux de chômage, la situation de ceux qui ont du mal à obtenir assez de travail pour en vivre sera sous-estimée dans le débat public. Il ne serait pas du tout étonnant que ces mêmes gens soient devenus méfiants envers les experts et les statistiques utilisées dans le débat politique, étant donné le désaccord entre ce que les politiciens disent du marché du travail et ce qu'ils vivent en réalité.

L'essor de politiques identitaires, depuis les années 1960, a mis une tension supplémentaire sur ces systèmes de classification. Les statistiques ne sont crédibles que si les gens se reconnaissent dans la gamme limitée de catégories démographiques disponibles, qui sont choisies par l'expert et non par la personne interrogée. Mais lorsque l'identité devient un enjeu politique – lorsqu'on parle de genre, de sexualité, de race ou de classe – les gens exigent de se définir selon leurs propres critères.

Les sondages d'opinion peuvent souffrir pareillement. Ces sondages saisissent traditionnellement les attitudes et les préférences des gens, sous l'hypothèse raisonnable qu'ils se comporteront en conséquence. Mais à une époque où la participation politique décline, il ne suffit pas de savoir simplement quelle case quelqu'un préfère cocher. Il importe de savoir s'ils le ressentent assez fortement pour se sentir concernés. Et lorsqu'il s'agit de saisir de telles fluctuations dans l'intensité émotionnelle, le sondage est un piètre outil.

La statistique a toujours subi la critique tout au long de son histoire. Les défis que lui posent les politiques identitaires et la mondialisation ne sont pas nouveaux non plus. Pourquoi dès lors les événements de l'année passée paraissent-ils si dommageables pour l'idéal d'expertise quantitative et pour son rôle dans le débat politique ?

Irruption du *Big Data*

Depuis quelques années, une nouvelle façon de quantifier et visualiser les populations apparaît, qui potentiellement marginalise la statistique et sonne l'avènement d'une ère toute différente. La statistique, collectée et compilée par des techniciens experts, laisse la place aux données qui s'accumulent automatiquement, du fait de la numérisation envahissante. Traditionnellement, les statisticiens savaient quelles questions ils voulaient poser et quelle était la population concernée, puis ils allaient chercher les réponses. En revanche, les données se produisent d'elles-mêmes lorsque nous utilisons une carte de fidélité, laissons un commentaire sur Facebook ou cherchons quelque chose sur Google. Comme nos villes, nos voitures, nos maisons et les objets du ménage sont dorénavant connectés, la masse des données que nous laissons dans notre sillage va devenir encore plus considérable. Dans ce monde nouveau, les données sont captées d'abord : les questions de recherche viennent ensuite.

À terme, ceci aura des implications probablement aussi profondes que l'invention de la statistique à la fin du 17e siècle. L'expansion des « données massives » fournit des occasions d'analyses quantitatives beaucoup plus abondantes qu'autant de sondages ou de modèles statistiques que vous voudrez. Et, ce n'est pas seulement la quantité des données qui diffère. C'est un type de connaissance entièrement différent, accompagné d'un nouveau mode d'expertise.

Tout d'abord, il n'y a aucun cadre d'analyse déterminé (comme la nation) ni aucune catégorie constituée (comme « chômeur »). Ces énormes nouveaux stocks de données peuvent être

explorés pour rechercher des motifs, des tendances, des corrélations et des tendances émergentes. Cela devient une façon de suivre à la trace les identités que les gens s'attribuent (comme « #JesuispourCorbyn » ou « entrepreneur ») plutôt que de leur assigner une catégorie. C'est là un mode d'agrégation approprié à une ère politique plus fluide, où tout ne peut être – comme avec les Lumières – relié de façon fiable à un quelconque idéal d'État-nation gardien de l'intérêt public.

En second lieu, la plupart d'entre nous oublions totalement ce que toutes ces données disent de nous, individuellement ou collectivement. Il n'y a aucun équivalent d'un Office National de Statistique pour les données de masse enregistrées par le commerce. Nous vivons à une époque où nos sentiments, identités ou affiliations peuvent être pistés et analysés à une vitesse et avec une précision sans précédent - mais rien ne rattache cette nouvelle capacité à l'intérêt public ou au débat public. Des « data-analystes » travaillent pour Google et Facebook, mais ce ne sont pas des « experts » du même genre que ceux qui produisent la statistique et qui sont maintenant si largement condamnés. L'anonymat et le secret des nouveaux analystes les rendent potentiellement plus puissants politiquement que tout spécialiste des sciences humaines.

Une entreprise comme Facebook est capable de procéder à une analyse sociale quantitative sur des centaines de millions de gens, à très bas prix. Mais elle est très peu incitée à en révéler les résultats. En 2014, quand les chercheurs de Facebook ont publié les résultats d'une étude « de la contagion émotionnelle » qu'ils avaient effectuée sur leurs usagers – où ils avaient modifié des fils d'actualités pour voir comment cela affectait les contenus que les usagers partageaient en réponse – ce fut un tollé : les gens avaient été soumis à une expérience à leur insu. Ainsi, du point de vue de Facebook, pourquoi, en publiant, aller au-devant de tracasseries ? Pourquoi ne pas plutôt faire l'étude et se taire ?

Un nouveau climat technique et politique

Ce qui est politiquement le plus significatif de ce passage d'une logique de statistiques à une logique de données est de voir combien il s'accorde au développement du populisme. Les leaders populistes peuvent accumuler le mépris sur les experts traditionnels, tels qu'économistes et sondeurs, tout en faisant confiance à une forme différente d'analyse numérique. Ces politiciens s'appuient sur une élite nouvelle, moins visible, qui recherche les régularités en explorant des banques de données énormes, mais font rarement état publiquement de conclusions, sans parler d'en apporter la moindre preuve. Ces analystes de données sont souvent des physiciens ou des mathématiciens, dont les compétences n'ont pas été développées pour l'étude de la société. Tel est le cas, par exemple, de la vision du monde propagée par Dominic Cummings, ancien conseiller de Michael Gove et directeur de campagne de du vote pour le Brexit. « La physique, les mathématiques et l'informatique sont des domaines où il y a de vrais experts, contrairement à la prédiction macro-économique », affirme Cummings.

Les personnages proches de Donald Trump, comme son stratège en chef Steve Bannon et le milliardaire de la Silicon Valley Peter Thiel, connaissent de près les techniques d'avant-garde pour l'analyse des données, via des entreprises comme Cambridge Analytica, au conseil duquel siège Bannon. Pendant la campagne présidentielle, Cambridge Analytica a rassemblé diverses sources de données pour établir les profils psychologiques de millions d'Américains, qu'il a utilisés pour aider Trump à cibler les électeurs avec des messages ajustés.

Cette capacité à développer et affiner une vision psychologique dans de grandes populations est une des caractéristiques les plus novatrices – et sujette à controverse – de la nouvelle analyse de données. À mesure que s'incorporent aux campagnes politiques ces techniques « d'analyse des sentiments », qui détectent l'humeur du grand nombre en suivant des indicateurs comme

les mots employés sur des réseaux sociaux, le charme émotionnel de personnages comme Trump devient l'objet possible d'un examen scientifique minutieux. Dans un monde où l'on peut à ce point détecter les sentiments politiques du grand public, qui a besoin de sondeurs ?

Peu de conclusions sociales résultant de cette sorte d'analyse des données aboutissent jamais dans le domaine public. On n'a dès lors pas grand-chose pour ancrer le discours politique dans une quelconque réalité partagée. Avec l'autorité déclinante de la statistique et sans que rien ne la remplace dans la sphère publique, les gens peuvent vivre dans n'importe quelle communauté imaginaire avec laquelle ils se sentent le plus en phase et à laquelle ils désirent croire. Dans les domaines où la statistique pourrait corriger les affirmations erronées sur l'économie, la société ou la population, à l'âge de l'analyse des données, il existe peu de mécanismes pour empêcher les gens de donner libre cours à leurs réactions instinctives ou leurs préjugés émotionnels. Au contraire, les entreprises telles que Cambridge Analytica traitent ces sentiments comme des traces à suivre.

Mais même s'il y avait un Bureau pour l'Analyse des Données, agissant pour le compte du public et de l'État, comme le fait l'ONS, il n'est pas clair qu'il offrirait la sorte de perspective neutre que les « libéraux »[2] défendent aujourd'hui. Le nouveau mode de calcul convient bien pour détecter les tendances, percevoir l'humeur et découvrir des choses comme le gonflement d'une bulle. Il sert très bien les directeurs de campagne et de marketing. Il convient moins à la formulation de constats sur la société non ambigus, objectifs, faisant consensus, ce pour quoi les statisticiens et économistes sont payés.

Dans ce nouveau climat technique et politique, il incombera à la nouvelle élite numérique d'identifier les faits, les projections et les vérités dans le flot bouillonnant des données. Il reste à voir si des indicateurs comme le PIB et le chômage conserveront un intérêt politique ; mais si ce n'est pas le cas, ce ne sera pas nécessairement la fin des experts et moins encore la fin de la vérité. La question à considérer plus sérieusement, maintenant que des chiffres sont constamment produits dans notre dos et hors de notre connaissance, est de savoir dans quel état la crise de la statistique laisse la démocratie représentative.

D'une part, la capacité de riposte des institutions politiques de vieille tradition mérite d'être reconnue. Tout comme « l'économie partagée » des plates-formes comme Uber et Airbnb a récemment été contrecarrée par des décisions légales (Uber étant contraint de reconnaître ses conducteurs comme salariés, Airbnb étant totalement interdit dans quelques municipalités), le droit de la vie privée et les droits de l'homme représentent un obstacle potentiel pour l'extension de l'analyse des données. Ce qui est moins clair, c'est comment les bénéfices de l'analyse numérique pourraient être offerts au public, comme beaucoup de données statistiques le sont. Des organisations comme l'Open Data Institute, cofondé par Tim Berners-Lee, font campagne pour rendre les données disponibles au public, mais ils ont peu de poids sur les entreprises où une si grande partie de nos données s'accumulent maintenant. La statistique a débuté comme un outil par lequel l'État pouvait avoir une vision de la société, mais s'est progressivement développée de façon que les universitaires, les réformateurs de la société et les entreprises en soient parties prenantes. Mais pour beaucoup de sociétés d'analyse de données, le secret entourant les méthodes et les sources de données est un avantage compétitif auquel elles ne renonceront pas volontiers.

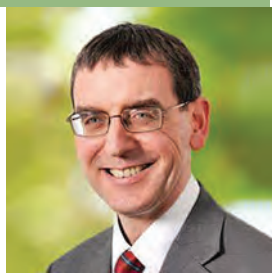
2. NDR : Dans le texte anglais : « liberals ». Attention, ce mot est un faux ami. Selon Guy Sorman : « Très éloigné du libéralisme européen, le « liberalism » aux États-Unis est en réalité un étatisme de gauche. Le libéral américain est keynésien, c'est-à-dire interventionniste sur le plan économique et libertaire sur le plan des mœurs. [...]. Malheureusement, rares sont les traducteurs qui font la nuance ».

Vers une société post-statistique ?

Une société post-statistique est une proposition potentiellement effrayante, non pas parce qu'elle ne contiendrait aucune forme de vérité ou d'expertise, mais parce qu'elle les privatiserait vigoureusement. La statistique est un des nombreux piliers du libéralisme et, en fait, des Lumières. Les experts qui la produisent et l'utilisent sont maintenant dépeints comme arrogants et oublieux des dimensions émotionnelles et locales de la politique. Il est sans doute possible de modifier la collecte des données pour mieux refléter les expériences vécues. Mais la bataille à mener n'est pas, à long terme, entre une politique élitiste de faits et une politique populiste de sentiments. Elle est entre ceux qui défendent encore la connaissance publique et le débat public et ceux qui profitent de leur désintégration en cours.

ARTICLE DU DOSSIER

Réponse à William Davies La statistique est encore plus importante dans un monde de « post-vérité »



John PULLINGER

Statisticien national, Chef du Service Statistique Gouvernemental[1] (GSS) et directeur général de l'Autorité Statistique du Royaume-Uni[2].

J'ai lu avec grand intérêt la contribution de William Davies, fascinante quoique très pessimiste, au débat sur un éventuel monde de « post-vérité » (La fin de la statistique, le 19 janvier). Je ne suis pas d'accord. Il n'y a jamais eu de temps plus passionnant pour la communauté des données. La demande d'une vision statistique pour aider à comprendre et traiter les questions qui se posent à la Grande-Bretagne et au monde n'a jamais été plus grande. Les gens veulent davantage de ce que nous fournissons : plus rapidement, plus détaillé, et abordant pleinement les questions de vie quotidienne.

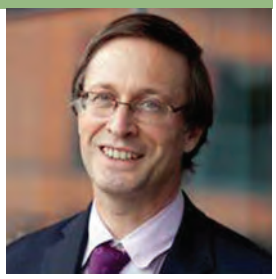
La gamme des statistiques disponibles s'étend à un rythme rapide, offrant plus que jamais la possibilité de comprendre, derrière les chiffres, vraiment ce qui arrive. Et l'on attend ardemment une analyse digne de confiance, qui conjure aussi bien les biais inhérents à beaucoup de sources de données que les intérêts institués de tous ceux qui essayent de déguiser leurs propres avis et préjugés en « faits irréfutables ».

Loin que « l'âge de statistique » soit derrière nous, voici le temps où nous pouvons apporter notre contribution à la société en fournissant la meilleure statistique qui permette de meilleures décisions.

1. NDR : « Government » désigne ici ce qu'en France on appellerait l'État.
2. John Pullinger nous a autorisés à publier ici la traduction de la réponse qu'il a faite à l'article de William Davies. Cette réponse est parue dans le courrier des lecteurs de The Guardian (24/1/2017).

ARTICLE DU DOSSIER

La statistique publique à l'ère du numérique : entre déclin, mission impossible et nouveau départ



Quatre questions à Dominique BUREAU

Président de l'Autorité de la statistique publique[1]

Statistique et société : Dans l'article du Guardian, William Davies déplore une « perte de crédibilité » des statistiques publiques. "Plutôt que de dissiper la controverse et la polarisation, il semble que la statistique les attise en réalité". Ce constat semble confirmé, en France, par certaines enquêtes d'opinion[2]. Nombreux sont ceux qui font même le constat d'une crise de la quantification en général. Faites-vous aussi ce constat ? Y apporteriez-vous des nuances ?

Dominique Bureau : Notre pays a la chance de disposer d'un système d'information statistique très fourni et qui s'enrichit continuellement. À cet égard, le rapport de l'Autorité pour l'année 2016 signale, par exemple, la publication, pour la première fois, des résultats infra-communaux sur le revenu disponible et la pauvreté monétaire, ainsi que, comme les années précédentes, de nombreuses enquêtes ou publications inédites sur des sujets touchant de près nos concitoyens.

Cependant, malgré la quantité considérable d'informations et les progrès réalisés, le débat public demeure difficile, notamment sur les questions sensibles que constituent l'emploi, le chômage et la précarité de l'emploi, comme l'ont montré les travaux de la Commission d'enquête du Sénat sur les chiffres du chômage. Plus généralement, le public attend des chiffres variés couvrant tous les domaines de l'économie et de la société, reflétant la diversité des situations, permettant les comparaisons internationales… Dans ce contexte, la statistique publique n'échappe pas au constat de défiance et d'affaiblissement général de « l'autorité des chiffres ».

Les prophètes du déclin de la statistique ont cependant tendance à faire « feu de tout bois ». En effet, il ne faut pas confondre les insatisfactions vis-à-vis de certaines politiques ou services publics, ou celles ressenties par rapport aux situations rencontrées, avec les questions sur la production des chiffres. Et il faut distinguer les interrogations sur la qualité de la statistique de celles sur les lacunes en matière d'évaluation des politiques publiques ou dans la mobilisation de l'expertise pour les concevoir, ou encore par rapport à la participation du public à leur élaboration.

Il faut aussi se garder d'embarquer sans discernement la production statistique dans les débats sur la pertinence du modèle de décision rationnel des agents économiques ou du critère utilitariste pour les choix publics. En effet, la statistique est fondamentalement observationnelle et à visée descriptive. De plus, pour éclairer les enjeux macroéconomiques, la statistique doit estimer des agrégats, construits sur des bases transparentes et aussi pertinentes que possible. La résurgence de ces critiques alors même que, par exemple, jamais les travaux empiriques

1. Si les vues exprimées ici s'appuient sur les travaux de l'Autorité (dont les rapports annuels sont consultables sur le site www.autorité-statistique-publique.fr), elles n'engagent cependant que leur auteur.
2. Voir par exemple, à propos d'une enquête récente du Centre d'études de la vie politique en France (Cevipof), l'article de Jean Chiche dans le numéro 2016-4 de Statistique et société.

en économie comportementale et le souci de comprendre les écarts de perception sur les situations économiques et sociales n'ont été aussi actifs, est un autre paradoxe des débats actuels.

Enfin, s'il est vrai que l'irruption du *Big Data* conduit à renouveler la réflexion méthodologique sur les potentialités d'analyses prédictives « sans modèle » par rapport à l'économétrie privilégiant l'énoncé préalable d'hypothèses précises, contrôlables par des tests, force est de noter que le sujet n'est pas totalement nouveau non plus. Il importe donc de distinguer les différents aspects de cette défiance et d'identifier ce qui met plus spécifiquement en cause la statistique. Ceci conduit à observer un tableau plus nuancé.

En effet, les enquêtes de satisfaction réalisées par l'Insee en 2016 montrent beaucoup d'éléments positifs : la notoriété de l'Insee est élevée, associée en général (80 %) à une bonne opinion ; alors que la confiance du public sur les chiffres et données publiés sur la situation économique et sociale ne dépasse pas 43 %, ce chiffre atteint 67 % pour ceux publiés par l'Insee, avec comme première raison pour cette confiance, celle dans l'organisme qui les produit. Toutefois, il est exact que le public se reconnaît plus ou moins dans ces chiffres. Surtout, la première raison de défiance mise en avant concerne plutôt la manière dont les chiffres sont utilisés.

Pour autant, ces constats plus rassurants ne doivent pas conduire à ignorer les menaces et défis pour la statistique publique. En effet, celle-ci se fixant pour objectif de fournir à tout un chacun à des fins de prise de décision, de recherche et de débat public, des informations de qualité, élaborées en toute indépendance, sur l'économie et la société car cela constitue un fondement des processus démocratiques et le progrès de la société, la défiance actuelle interpelle et met nécessairement en première ligne la statistique.

S&S : S'interrogeant sur les raisons d'une possible « crise », on peut rapprocher les critiques de William Davies de celles qui avaient été émises au moment du lancement de la commission Stiglitz – Sen – Fitoussi (SSF). Les recommandations de cette commission ont elles été suivies d'effet ?

DB : Le « Ramener les questions sociales et économiques à des agrégats numériques et des moyennes semble à beaucoup violer la décence politique» de Davies fait effectivement écho à la saisine de la Commission « SSF » en 2008, qui faisait état d'une « insatisfaction par rapport à l'état actuel des informations statistiques touchant à l'économie et la société » et lui donnait « mission de déterminer les limites du PIB comme indicateur des performances économiques et du progrès social, en soulignant les problèmes relatifs à sa mesure, d'identifier les informations complémentaires qui pourraient être nécessaires pour aboutir à des indicateurs plus pertinents du progrès social, d'évaluer la faisabilité de nouveaux instruments de mesure et enfin de débattre de la présentation la plus appropriée des informations statistiques ».

Face à ce qu'ils perçoivent comme des critiques excessives ou injustes, les statisticiens ont tendance à se mettre sur la défensive, et à rappeler que l'on n'a pas attendu la crise financière pour développer des indicateurs d'inégalités, cerner au plus près les caractéristiques de la pauvreté et l'exclusion ; ou pour essayer « de rendre compte des situations complexes et diversifiées du marché du travail par une batterie d'indicateurs sur l'emploi, le chômage, le sous-emploi et la précarité de l'emploi »[3] ; et, il y a plus de quarante ans, pour que Carré, Dubois et Malinvaud insistent sur la nécessité de comprendre les ressorts psychologiques et sociologiques de la croissance…

3. Cf. le rapport « Emploi, chômage, précarité – Mieux mesurer pour mieux débattre et mieux agir » du groupe de travail présidé par Jean-Baptiste de Foucauld, CNIS, 2008

Le rapport « SSF » reconnaissait d'ailleurs que bon nombre des questions abordées avaient été posées de longue date, par ceux-là même qui avaient contribué à élaborer nos systèmes actuels de comptabilité nationale. Toutefois - et c'est sans doute le point important à prendre en compte par les responsables statistiques - il considérait que, s'il avait été reconnu de longue date que le PIB posait problème en tant qu'outil de mesure des performances économiques, bon nombre des changements intervenus dans la structure de nos sociétés ont rendu ces déficiences plus criantes. Surtout, il mettait en avant que : « si la question de la mesure des performances économiques et du progrès social revêt de nos jours une importance particulière, c'est précisément parce que l'on craint que les mesures usuelles risquent d'encourager nos sociétés à évoluer dans une mauvaise direction, ce qui, la crise actuelle nous le montre, peut être générateur de détresse sociale et de dégradation du bien-être. »

Leur travail débouchait donc sur des recommandations opérationnelles, à partir de l'idée que les progrès de la recherche en de nombreuses disciplines rendent possible la conception de mesures plus larges du bien-être qui en intégreraient davantage d'aspects, notamment par rapport aux questions de soutenabilité de la croissance. En parallèle avec les réflexions sur les nouveaux indicateurs de richesse, notre système statistique s'est attaché à mettre en œuvre le programme de travail esquissé alors[4], ce qui s'est avéré fructueux et montre que, plus que de nourrir les débats de principe ou idéologiques, ce que nous avons à faire est d'abord de répondre aux attentes du public en matière d'information statistique.

Dans cette perspective, notre système statistique public est maintenant en ordre de marche pour assurer la coordination des travaux statistiques relatifs aux indicateurs de suivi pour les objectifs de développement durable approuvés par l'Assemblée générale de l'ONU en septembre 2015. Le recensement réalisé à cette fin montrait que : parmi les 229 indicateurs (sans les doublons), 198 d'entre eux relèvent effectivement du domaine de la statistique, 31 relevant plutôt de la mise en œuvre des politiques publiques ; les services producteurs sont bien identifiés pour 84 % des indicateurs statistiques, le SOeS, l'Insee et la Drees[5] étant les plus grands pourvoyeurs ; par ailleurs, 65 % des indicateurs existent déjà dans une version exacte ou approchée.

S&S : Êtes-vous d'accord avec l'idée selon laquelle le *Big Data* serait un nouveau « mode » de quantification qui remplacerait les anciennes méthodes légitimes comme les enquêtes ou les comptes nationaux ? Pensez-vous qu'il existe un risque réel que les analyses tirées de *Big Data* par des organismes privés concurrencent les résultats de la statistique publique ?

DB : La transformation numérique bouleverse le fonctionnement des entreprises et des marchés, avec l'émergence, autour d'internet, de nouveaux canaux d'information pour mettre en relation les différents acteurs, et le développement par ceux-ci de stratégies mobilisant l'abondance de nouvelles données et leur traitement par les « data-sciences »[6].

L'économie « numérique » se caractérise ainsi par la production de flux importants de données reflétant l'activité économique, issues de l'internet ou de différents capteurs, stockées sous des formes variées. L'exploitation de ces données associées au phénomène *Big Data* pour la production statistique suscite un intérêt croissant, ces données étant susceptibles de fournir de nombreuses opportunités : pour réduire les délais de publication, compte-tenu de

4. Cf. l'article. « Les préconisations du rapport Stiglitz-Sen-Fitoussi : quelques illustrations », Clerc M., Gaini M. et Blanchet D. in « L'économie française, édition 2010 » Insee
5. L'auteur cite trois services statistiques publics français : le Service de l'observation et des statistiques (SOeS) du ministère de l'Environnement, de l'Énergie et de la Mer, l'Institut national de la statistique et des études économiques (Insee), et la Direction de la recherche, des études, de l'évaluation et des statistiques (DREES) du Ministère des Affaires sociales et de la Santé.
6. Pour une description générale de la rupture que représentent ces données en termes de volumes, d'instruments pour les traiter et d'applications potentielles, cf. « Analyse des big data. Quels usages, quels défis ? », Hamel et Marguerit, France stratégie, 2013

la disponibilité immédiate de l'information ; pour disposer d'observations à des échelles plus fines ; pour compléter les indicateurs existants… Le projet « données de caisse » a ainsi été lancé en 2015, après une phase expérimentale en 2011, avec objectif de l'intégrer en production à l'horizon 2019.

Plus généralement, le groupe Cnis-Insee (2015) sur les potentialités de développements statistiques à partir de systèmes d'information (SI) de gestion privés a enclenché une dynamique. Ce groupe avait identifié notamment trois secteurs particulièrement prometteurs : outre l'utilisation des « données de caisse » pour la production d'indices de prix, les données de téléphonie pour mesurer la population présente et celles des cartes bancaires pour la consommation.

D'autres applications cherchent à utiliser d'autres sources, telles que : les données « satellites » (accessibles grâce au projet Copernicus) ; les requêtes des internautes, pour enrichir ou développer des méthodes alternatives pour la prévision conjoncturelle de la consommation (cf. *Google Trends*[7]), dont l'Insee a analysé les potentialités, avec, en l'état, des résultats peu concluants ; d'autres encore, à utiliser les données de systèmes de réservation ou celles d'usage de sites internet pour enrichir, par exemple, les statistiques culturelles, ou, plus généralement, combler les lacunes dans la mesure du volume d'activité du secteur tertiaire… Ce n'est qu'en les testant que la valeur de ces opportunités nouvelles pourra être établie. A ce titre, les projets en cours seront précieux.

L'adoption de la loi sur la République numérique fournit un cadre pour leur réalisation, puisque la statistique publique pourra désormais, pour les besoins d'enquêtes statistiques obligatoires, se voir transmettre sous forme électronique sécurisée des informations issues de certaines bases de données des personnes de droit privé concernées. Les conditions de confidentialité des informations communiquées par ces fournisseurs de données, socle de la confiance entre ceux-ci et le système statistique, sont ainsi établies, mais elles devront se décliner dans des conventions signées ensuite avec ceux-ci pour construire des cadres de coopération.

Ces nouvelles données offrent donc de nouvelles opportunités pour la statistique publique, qui doit par ailleurs développer les méthodes appropriées pour mesurer la valeur créée par la numérisation de l'économie. Elles obligent aussi à anticiper l'évolution de son rôle, notamment par rapport à l'émergence de nouveaux producteurs d'indicateurs statistiques, même si ceux-ci ne répondent pas, ou seulement partiellement, aux besoins d'information fiable et de qualité pour mesurer l'économie et les transformations sociales.

Le système statistique public (SSP) doit donc s'y préparer activement, en se gardant de relativiser les enjeux, même lorsque leur maturité semble encore incertaine. Certes, l'offre de nouveaux producteurs de données se situe encore « à côté » de la statistique, un peu comme, par exemple dans le domaine de la santé, avec les applications « data-mobiles » qui ont d'abord - mais cela évolue maintenant - concerné le bien-être plutôt que les traitements médicaux. Plus spécifiquement, il y a encore beaucoup de grain à moudre dans le développement de la statistique administrative, grâce notamment à l'appariement de ses fichiers, qui permet actuellement de renouveler l'éclairage de nombreux sujets.

Mais ceci ne doit pas conduire à négliger les perspectives de développement de la statistique à partir de nouvelles sources et, potentiellement l'émergence de nouveaux acteurs. En effet, la rigidité de l'offre publique la rend toujours fragile lorsqu'émergent des producteurs potentiellement « concurrents », même si dans un premier temps, le recouvrement des champs

7. Cf. dossier de la note de conjoncture publiée par l'Insee en mars 2015 : « Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées » Clément Bortoli et Stéphanie Combes

d'activités peut sembler marginal. De plus, dans un monde où l'information est disponible de façon quasi-instantanée, il faut s'attendre à un renforcement des exigences du public en termes de réactivité, de capacité à qualifier les phénomènes et à publier des données fiables, avec aussi des difficultés pour en faire reconnaitre la qualité ou l'objectivité dans un contexte de prolifération de l'information.

S&S : Faisant preuve d'un optimisme volontariste, le « statisticien national » anglais, John Pullinger, a répondu au pessimisme de l'article de William Davies en se réjouissant de la demande croissante adressée aux statistiques publiques, mais sans citer de piste pour satisfaire cette demande. Pour votre part, quels remèdes envisagez-vous pour faire face à la situation actuelle ?

DB : Effectivement, si l'on adhère à l'idée exposée par John Pullinger selon laquelle c'est le moment pour la statistique d'apporter sa meilleure contribution à la société, encore faut-il établir les conditions pour cela. Une première réponse se situe au niveau de la définition des programmes de travail appropriés pour mieux satisfaire les attentes des utilisateurs, de la résolution des problèmes de méthodes rencontrés, de la qualité des publications…Mais il faut placer cela dans une stratégie plus globale car la gouvernance est cruciale pour cela, avec, évidemment, comme premier principe celui d'indépendance.

Fondamentalement, l'existence même de l'Autorité de la statistique publique s'inscrit dans cette perspective, sa mission étant de veiller « au respect du principe d'indépendance professionnelle dans la conception, la production et la diffusion de statistiques publiques ainsi que des principes d'objectivité, d'impartialité, de pertinence et de qualité des données produites ». L'objectif ultime est ainsi que le public puisse avoir confiance en son système statistique. Et c'est la perspective de contribuer à celle-ci qui guide son action.

Pour cela, le contrôle du respect exemplaire des principes du Code de bonnes pratiques de la statistique constitue le cœur de l'activité de l'Autorité, avec leurs trois dimensions : des facteurs institutionnels et organisationnels déterminants de la crédibilité, notamment l'indépendance professionnelle, l'engagement sur la qualité, le secret statistique, l'impartialité et l'objectivité ; des procédures pour organiser, collecter, traiter et diffuser les statistiques ; et des résultats statistiques en termes de pertinence, exactitude et fiabilité, mais aussi d'actualité, de cohérence et d'accès pour les utilisateurs.

Mais c'est évidemment l'ensemble du système statistique qui est concerné. À ce titre, le plan d'action de gestion de la qualité au sein des processus de production statistique qui fait suite à la création du Comité stratégique de la qualité à l'Insee fait partie des avancées importantes à porter au crédit de 2016. Il est aussi essentiel que les programmes statistiques répondent aux attentes du public et que les parties prenantes se les approprient, avec une vision partagée. Le CNIS est donc aussi une pièce maitresse.

Enfin, on ne peut laisser dans l'angle mort la question de l'utilisation qui est faite des statistiques, qui souvent nuit à l'image d'ensemble de la « marque » lorsqu'elle est inappropriée. Le rapport de Foucauld observait, par exemple, que, dans le débat public, le taux d'emploi est moins considéré que le taux de chômage, dont les variations pour ce qui concerne les chiffres mensuels de Pôle emploi, peu significatives, prennent une importance exagérée, par rapport aux chiffres sur l'ancienneté au chômage, par exemple. Plus généralement, l'enquête sur l'image de l'Insee auprès des « Inseenautes » pose une question sur les raisons de la défiance dans les données sur la situation économique et sociale de la France. La réponse principale met en cause non pas la valeur des chiffres mais ce qu'on leur fait dire. Ceci ne peut être négligé.

8. Op.cit. note 3.

ARTICLE DU DOSSIER

Le *Big Data*, au fond, qu'est-ce que ça change ?

Après le centième Café de la statistique

Jean-François ROYER
SFdS[1]

Le 27 mars 2017 s'est tenu le centième « Café de la statistique ». Il était consacré au thème omniprésent aujourd'hui : « Big Data : Big Science ? Big Brother ? ». Il a rassemblé 130 participants autour de l'invitée, Valérie Peugeot, sociologue et commissaire de la Cnil. J'étais l'un d'eux : voici un témoignage personnel.

En observatrice de la mutation numérique, Valérie Peugeot nous a d'abord appelés à reconnaître la nouveauté que le *Big Data* apporte dans la société. Certains, devant des promesses mirobolantes, parfois suivies de déceptions, peuvent se demander s'il ne s'agit pas d'une « bulle » médiatique. Non – des pratiques réellement nouvelles, réellement liées à l'exploitation des données massives sont attestées dans de nombreux secteurs, et pas seulement le marketing commercial ou politique. La médecine, l'agriculture, la finance et l'assurance, l'urbanisme ont été cités comme exemples. Dans le champ des connaissances, astronomie, écologie, géographie et même analyse littéraire obtiennent des résultats grâce aux données massives. Et des listes complètes seraient certainement plus longues.

S'agit-il d'une « révolution technologique » ? Valérie Peugeot est très méfiante devant le « déterminisme technologique ». Les nouvelles possibilités de la communication et du calcul ne contiennent pas dès le début leurs capacités « révolutionnaires » qui se déploieraient indépendamment des influences sociales. Il faut récuser toute « prédestination » et constater les allers-retours permanents entre technique et économie, entre conception et appropriation sociale. Les points de contact, ou de friction, entre techniques nouvelles et pratiques sociales méritent d'être observés pour eux-mêmes, et aussi pour agir dans la société nouvelle. L'invitée est impliquée dans une action « citoyenne » pour promouvoir les usages du numérique qui respectent les droits des personnes et pour combattre les menaces que les nouvelles technologies font courir aux libertés publiques[2].

Le débat qui a suivi l'exposé initial de Valérie Peugeot a pendant deux heures permis de développer ces différents points. Beaucoup d'interventions prolongeaient dans le sens de l'inquiétude les analyses de l'invitée quant à l'usage des données individuelles dans le monde qui vient. Les craintes pour soi-même sont vives ; les craintes pour la société le sont encore plus.

1. Les Cafés de la statistique sont organisés à Paris depuis 2005 par le groupe « Statistique et enjeux publics » de la SFdS, pour examiner « si la statistique éclaire les questions de société ». Chaque séance est introduite par un invité ou une invitée, et se poursuit par un débat. Des comptes rendus écrits sont rédigés ; depuis 2012, ces comptes rendus sont complétés par des vidéos. Textes et vidéos sont téléchargeables à partir du site de la SFdS, rubrique Cafés de la statistique.
2. Outre ses fonctions à la Cnil, elle préside l'association « Vecam – Citoyenneté dans la société numérique »

Les réglementations protectrices des individus ne cessent de se renforcer[3] ; et il est possible de donner à chacun de bons conseils pour assurer la sécurité de ses propres données[4]. Mais sur le plan collectif, la perspective d'une individualisation de plus en plus poussée des « traitements » appliqués aux personnes par des institutions de plus en plus gigantesques (les « GAFA[5] », les grands États) laisse craindre un fonctionnement de plus en plus atomisé de la société, voire oppressif, face auquel on discerne mal comment des « corps intermédiaires » pourraient émerger et devenir des recours. C'est sans doute là qu'on touche de plus près la menace de « révolution » contenue dans le *Big Data*.

Le public du Café comportant beaucoup d'étudiants et de professionnels de la statistique, la question épistémologique n'a pas été absente du débat. Rupture, ou pas rupture ? La tonalité dominante était là rassurante : les anciennes méthodes ne seraient pas supplantées par celles qui traitent des données massives, mais seulement appelées à se renouveler ; le discours selon lequel le *Big Data* signerait la fin des hypothèses et des théories serait déjà dépassé. Pourtant, dans bien des contextes scientifiques, la recherche des meilleures prédictions, facilitée par l'abondance des données, n'est-elle pas privilégiée par rapport à la recherche sur les mécanismes sous-jacents aux observations, recherche qui est plus aléatoire ? On peut s'en inquiéter. Mais le débat a surtout porté sur l'accompagnement nécessaire du développement des traitements de données massives. Le souci de la qualité des données doit rester présent même si elles sont recueillies « au vol ». La formation des étudiants en statistique se transforme presque partout, pour faire place aux nouvelles compétences théoriques et pratiques requises pour traiter les données massives. Et l'invitée a décrit de façon positive deux innovations dans le domaine de la connaissance scientifique : le recours à des collaborations ouvertes nombreuses (« science participative ») et les droits nouveaux accordés à la fouille des gisements de données pour la recherche, fût-ce au détriment des droits de propriété (droits d'auteurs, brevets).

La statistique publique peut profiter du *Big Data*. Elle est une possible « consommatrice » de gros fichiers produits par des acteurs privés, qui lui permettraient d'économiser sur ses modes de collecte traditionnels[6]. Encore faut-il qu'elle sache s'en emparer, et que ces nouvelles sources garantissent une production régulière et de qualité. Ce point de vue a été largement exprimé au cours du Café. Un autre angle a été moins développé : et si l'ère du *Big Data* venait délégitimer beaucoup de productions de la statistique publique, en proposant des informations concurrentes plus rapides, plus détaillées, moins « étatiques » ? La menace[7] d'une élimination par la concurrence ne semble guère ressentie : les statisticiens publics restent dans l'ensemble confiants dans la protection que leur donne la réputation actuelle de leurs productions. En revanche, beaucoup se préoccupent du trouble que de telles informations, si elles étaient de qualité insuffisante, pourraient jeter dans les débats publics. Certains participants n'hésitent pas à suggérer une « charte universelle » de la diffusion d'informations publiques issues de traitements de données. Une sorte de « label rouge » des traitements de données massives destinés à être versés à un débat public et irréprochables sur le plan méthodologique, voire éthique…

Oui, le *Big Data*, « ça change beaucoup de choses » ! Mais caractériser ce changement reste difficile, et on est loin d'en pouvoir mesurer « statistiquement » les effets. Le centième Café de la statistique a parcouru ce vaste domaine : puisse-t-il être suivi de beaucoup d'autres pour débattre des mesures et des solutions.

3. Voir dans ce numéro l'analyse par Judith Rochfeld des dispositions de la « loi numérique » française et du nouveau règlement européen sur la protection des données
4. A été cité l'ouvrage de Tristan Nitot « Surveillance:// Les libertés au défi du numérique » C&F Editions 2016 – 19€
5. Google – Amazon – Facebook – Apple
6. Voir dans ce numéro les exemples donnés par Dominique Bureau et par Gilbert Saporta
7. Voir dans ce numéro l'article de William Davies pour « The Guardian »

MÉTHODES

Quelle statistique pour le *Big Data* ?

Entretien avec Gilbert SAPORTA

Professeur émérite de statistique appliquée
Conservatoire National des Arts et Métiers

Tout le monde s'intéresse au *Big Data*. Le public est de mieux en mieux informé sur les potentialités que les données massives recèlent et sur les dangers que leur utilisation peut comporter. Mais très rares sont ceux qui savent ce qui se cache « sous le capot » des nouvelles méthodes. Statistique et Société a demandé à Gilbert Saporta, qui fait partie de ce petit nombre, d'éclairer autant que possible les non-spécialistes.

Statistique et société : Du point de vue des méthodes, qu'est-ce qui détermine si on est dans un contexte *Big Data* ou non ? Y a-t-il un seuil en nombre d'observations, de variables ? Un seuil par rapport aux capacités mémoire d'un ordinateur ? Que penser du critère de la « vélocité » si souvent avancé ?

Gilbert Saporta : Quand on évoque le *Big Data*, on pense en premier lieu au volume des données, en d'autres termes à la taille du fichier correspondant. Une des premières occurrences de l'expression *Big Data* dans la littérature scientifique est la communication de deux chercheurs de la NASA, Cox et Ellsworth, au congrès SIGGRAPH de 1997[1].

Rappelons tout d'abord que selon les époques la notion de « Big » a beaucoup varié. Qui ne se souvient que la grande taille pour les estimations et les tests commençait à n=30 ? Avec le traitement des recensements, les statisticiens sont confrontés depuis longtemps à des données massives. Comme le rappelle David Donoho[2], cela a conduit Herman Hollerith à inventer la carte perforée et à créer IBM. C'est la technologie qui fixe les limites. Comme le disait déjà à peu près en ces termes John Tukey : « Big », c'est quand cela ne rentre pas dans ma machine. Il n'est donc pas possible de fixer des seuils. Toutefois on parle de données de grande dimension quand le nombre de variables dépasse de beaucoup le nombre d'observations, mais c'est un sujet à part. On parlera donc de *Big Data* quand les données ne peuvent être stockées sur un seul ordinateur (données réparties) et quand les traitements vont nécessiter plusieurs machines (calculs distribués). Le volume est le premier V, d'une série de trois, introduits par le Gartner Group en 2008. Le deuxième V est en effet la vélocité qui renvoie aux flux de données recueillies en temps réel sur le web, ou par des capteurs, tels les objets connectés, les compteurs électriques « intelligents », etc. Les problèmes de stockage et d'algorithmique deviennent alors essentiels car on ne peut tout conserver, et il faut recourir à des méthodes incrémentales qui actualisent les résultats au fur et à mesure de l'acquisition de nouvelles données.

1. Cox M. and Ellsworth D. (1997), Managing Big Data for Scientific Visualization, Exploring Gigabyte Datasets in Real-Time: Algorithms, Data Management, and Time-Critical Design, *Siggraph* 97, Course Notes 4, New York, ACM Press.
2. Donoho D (2015), 50 years of Data Science, *Tukey Centennial workshop* http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Le troisième V, souvent mis en avant, est la variété des types de données recueillies : numériques et qualitatives comme d'habitude, mais aussi des images, des données issues de réseaux sociaux (qui est lié avec qui ?) et non structurées comme des textes.

S&S : Quels sont les traitements qu'on sait faire dans des temps raisonnables sur des bases de données de très grande taille ? Réciproquement, y a-t-il des traitements qu'on ne sait pas faire sur une base de données qui ne tient pas en mémoire d'un ordinateur ?

GS : Tout est une question de moyens et d'approches spécifiques. Quand on peut utiliser par exemple le cloud d'Amazon, ou de puissants réseaux d'ordinateurs, et des programmes conçus pour travailler en parallèle, suivant par exemple le modèle de programmation MapReduce inventé par Google, on peut mettre en œuvre la plupart des traitements statistiques standard. Des bibliothèques libres de programmes dédiées au *Big Data* comme Spark, Scikit-Learn, MLLib contiennent les méthodes favorites des statisticiens : régression linéaire et logistique, classification supervisée et non supervisée, réduction de dimension, mais aussi des algorithmes d'apprentissage comme les forêts aléatoires et les séparateurs à vaste marge (SVM). Le catalogue s'enrichit sans cesse grâce aux travaux de communautés d'utilisateurs car ce sont des systèmes ouverts, tout comme l'environnement R. Par contre, certaines méthodes comme l'estimation de densité multidimensionnelle sont mal adaptées aux données massives.

S&S : Dans l'univers *Big Data*, le nombre de données étant tellement important, tout test statistique devient significatif. Faut-il oublier les tests statistiques et les intervalles de confiance ? Et qu'utilisera-t-on pour les remplacer ?

GS : En effet. Tout écart à une hypothèse nulle devient significatif : déjà avec 10 000 observations, ce qui n'est pas du *Big Data*, un coefficient de corrélation égal à 0.02 est déclaré significativement non nul au risque 5 % bilatéral. Est-ce utile ? Bien sûr que non, et ne parlons pas de millions de données ! Tous les coefficients d'un modèle deviennent alors « significatifs », mais de quoi ? En plus, et ce n'est paradoxal qu'en apparence, les tests d'adéquation rejettent tous les modèles usuels, car ils sont trop simples pour exprimer de grandes masses de données. Une autre forme d'inférence doit être mise en œuvre, liée à la notion de reproductibilité ou de généralisabilité des résultats.

La théorie de l'apprentissage statistique, développée par Vladimir Vapnik et Alexei Cervonenkis, donne des bornes pour la différence entre la performance en apprentissage et la performance sur de nouvelles données, dans le cadre prédictif. Cette théorie permet de qualifier, en fonction d'une mesure de complexité, les modèles qui généralisent bien.

L'application de cette théorie n'étant pas toujours aisée, on recourt souvent au procédé suivant : ayant séparé aléatoirement les données disponibles en deux sous-ensembles, on vérifie sur le deuxième sous-ensemble si les résultats obtenus sur le premier restent valables. On pourra procéder à plusieurs séparations pour étudier la variabilité et éviter des cas trop particuliers. Cette pratique, voisine de la validation croisée, est particulièrement bien adaptée aux données massives. On l'attribue généralement au machine learning qui l'a systématisée pour éviter les phénomènes de surajustement, mais elle est bien plus ancienne. Dès 1941, le psychométricien Paul Horst dans un chapitre intitulé

> « *L'utilité d'une procédure de prédiction n'est pas établie lorsqu'on a trouvé qu'elle prédisait correctement sur l'échantillon original ; l'étape suivante nécessaire doit être son application à au moins un second groupe. Ce n'est que si elle prédit correctement sur des échantillons ultérieurs que la valeur de la procédure peut être considérée comme établie* »[3].

S&S : Y a-t-il toujours un modèle dans un traitement Big Data ? Explicite ou sous-jacent ? Que penser des déclarations de ceux qui disent qu'on doit « arrêter de modéliser » ?

GS : Il faut s'entendre sur ce que l'on appelle modéliser ! S'il s'agit de modèles génératifs, c'est-à-dire de modèles en général simples, et interprétables dans le langage du champ d'application, censés décrire le mécanisme qui a engendré les données, la réponse est clairement non. Aucun modèle simple ne peut représenter de grandes masses de données. Le célèbre aphorisme de George Box « tous les modèles sont faux, certains sont utiles » s'applique parfaitement.

En *Big Data*, on utilise abondamment des modèles prédictifs, mais au sens d'algorithmes, sans chercher à mimer le processus génératif que l'on considérera inconnu. Le seul critère est la capacité de prédire de nouvelles observations. On peut d'ailleurs prouver certains résultats en apparence paradoxaux : ainsi de la remarque de V.Vapnik : « On obtient parfois de meilleurs modèles en évitant délibérément de reproduire les vrais mécanismes »[4]. Dans le même ordre d'idée, Shmueli indique :

> « *La significativité statistique joue un rôle mineur, ou pas de rôle du tout, pour établir la performance en prédiction. En fait, il arrive parfois qu'on obtienne une meilleure précision de la prédiction en retirant des variables en entrée ayant de petits coefficients, même s'ils sont statistiquement significatifs* »[5]

J'avais abordé en 2008 dans une conférence invitée au congrès Compstat[6], la distinction entre *modèles pour comprendre et modèles pour prédire*, sans avoir connaissance de l'article exceptionnel de Leo Breiman de 2001[7] sur les deux cultures de la modélisation statistique dont je recommande vivement la lecture, ainsi que celle de la conférence de David Donoho citée plus haut, à l'occasion du centenaire de la naissance de John Tukey, qui reprend largement l'article de Breiman.

Certains modèles sont d'interprétation aisée, comme les arbres de décision, d'autres sont plutôt des boîtes noires, comme les réseaux de neurones, les forêts aléatoires, le boosting etc. Il est courant de combiner les prévisions de différents modèles (linéairement ou non) plutôt que de rechercher le meilleur modèle : ce sont les meta-modèles ou modèles d'ensemble qui remportent souvent les compétitions.

Arrêter de modéliser (au sens des modèles génératifs) renvoie à la tribune provocatrice de Chris Anderson (2008) sur la fin de la théorie qui prétendait :

> « *Les péta-octets nous permettent de dire : « la corrélation suffit ». On peut s'arrêter de rechercher des modèles. On peut analyser les données sans hypothèses sur ce que cela pourrait montrer. On peut injecter les chiffres dans les plus grandes grappes d'ordinateurs que le monde ait jamais vues, et laisser les algorithmes statistiques trouver des configurations là où la science en est incapable* »[8]

3. « *The usefulness of a prediction procedure is not established when it is found to predict adequately on the original sample; the necessary next step must be its application to at least a second group. Only if it predicts adequately on subsequent samples can the value of the procedure be regarded as established* ». Horst, P., Wallin, P. C., Guttman, L. C., Wallin, F. B. C., Clausen, J. A., Reed, R. C. et Rosenthal, E. C. (1941), The prediction of personal adjustment : A survey of logical problems and research techniques, with illustrative application to problems of vocational selection, school success, marriage, and crime. *Social science research council*.
4. « *Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms* ». V.Vapnik (2006), *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer
5. « *Statistical significance plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, even if they are statistically significant, results in improved prediction accuracy* » Shmueli G. (2010), To explain or to predict? *Statistical Science*, 25, 289–310
6. Saporta G. (2008) Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings*, Physica Verlag, 315-322
7. Breiman L. (2001) Statistical modeling: The two cultures. *Statistical Science*, 16 199–215
8. « *Petabytes allow us to say: « Correlation is enough. » We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot* ». C.Anderson (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired*, http://www.wired.com/2008/06/pb-theory/

La suite lui a donné tort et renvoie à la confusion entre corrélation et causalité. Ainsi de l'exemple souvent cité de la prévision de la grippe saisonnière avec Google Flu Trends. Vers 2010 des chercheurs de Google eurent l'idée de corréler les requêtes concernant des mots clés liés à l'apparition de la grippe (fièvre, etc.). Ils obtinrent ainsi un indicateur précurseur de l'épidémie, en avance sur les déclarations des médecins. Les prévisions furent excellentes jusqu'en 2012 mais le système se détraqua ensuite (surestimation) et fut abandonné dans sa version d'origine.

La moralité de l'histoire est que si on peut prédire sans comprendre, il faut prendre garde aux changements : toute méthode de prévision suppose que le futur ressemblera à ce que l'on connait déjà… Comprendre pour mieux prédire a de l'avenir !

S&S : Au fait, le *machine learning*, c'est quoi ?

GS : Pour faire simple, le *machine learning* désigne pour l'essentiel une famille d'algorithmes d'apprentissage supervisé, c'est à dire où on cherche à prédire une réponse en procédant par amélioration successive du prédicteur au fur et mesure qu'il traite de nouvelles données, en comparant valeur prévue et valeur vraie. L'ancêtre de ces méthodes est le perceptron de Rosenblatt né en 1958 des travaux en 1943 de McCulloch et Pitts sur les neurones artificiels. Les modèles dépendent d'un grand nombre de paramètres et pour bien apprendre, il leur faut en effet de très nombreuses données, d'où le lien avec les Big Data. Le *machine learning* est aussi un domaine de recherches très actif regroupant informaticiens, mathématiciens et statisticiens.

S&S : Analyser beaucoup de données ne signifie pas que l'on est exhaustif. Les données massives semblent souvent recueillies sans plans d'échantillonnage ou de sondage. Ne risque-t-on pas d'avoir des données massives mais biaisées car recueillies sur une sous-population uniquement (les utilisateurs de smartphone par exemple) ?

GS : Tout à fait ! Quantité ne veut pas dire forcément qualité et les précautions usuelles s'imposent. Il faut par exemple disposer de variables de calage pour redresser les résultats si on veut qu'ils soient en accord avec des études préalables.

J'évoquerai également un problème connexe : les données massives sont souvent récoltées automatiquement dans un but qui n'est pas nécessairement statistique : je pense ici aux capteurs installés dans de plus en plus d'équipements comme des véhicules, des fauteuils, les caméras de surveillance. Ces données sont souvent « sales » et les prétraitements sont essentiels.

S&S : Les données massives posent de façon cruciale la question de leur ouverture. Il y a *Big Data* aussi parce qu'on peut récolter des données facilement un peu partout sur le web, qu'elles soient publiques ou privées et les articuler les unes aux autres. Cela nous semble transformer assez profondément le métier de statisticien public dans la mesure où cela rend la frontière entre les données publiques et privées plus floue. Qu'en pensez-vous ?

GS : Bien qu'universitaire, je suis avec attention les évolutions de la statistique publique. A la suite du mémorandum de Scheveningen adopté en 2013 par l'assemblée des directeurs généraux des instituts nationaux de statistique européen, un plan *Big Data* a été adopté par les acteurs du système statistique européen et en particulier d'Eurostat. Son ancien directeur général, Walter Radermacher, a souvent évoqué la « Statistics 4.0 » et la fin de l'usine à enquêtes comme modèle des INS. Le plan *Big Data* a pour but de préparer le Système Statistique Européen à intégrer des sources de données massives dans la production des statistiques officielles. Ces sources nouvelles viennent compléter les registres et données administratives qui peuvent être déjà très volumineuses mais sont mises à jour lentement. Des expériences concluantes ont été menées concernant l'utilisation des données d'opérateurs de téléphonie mobile pour

améliorer les statistiques du tourisme et de la mobilité, la transmission des données de caisses de supermarché pour l'indice des prix à la consommation, l'utilisation des offres d'emploi sur Internet pour actualiser les enquêtes emploi etc. Il est clair que la mesure du commerce électronique passe par l'accès à des données issues des sites de vente en ligne. La rapidité de mise à jour, le volume traité et les économies réalisées sont des arguments essentiels, mais il ne faut pas sous-estimer la difficulté d'utiliser des données externes.

L'utilisation de sources de données privées soulève différentes questions : quelle en est la représentativité quand il y a plusieurs opérateurs de téléphonie ? comment s'assurer de la véracité des données (un quatrième V souvent évoqué) quand les statisticiens publics n'en contrôlent pas la production, et qu'il n'y a pas de vérité de terrain ? Quelle est la pérennité de ces sources privées qui peuvent se tarir selon le bon vouloir des entreprises ou leur propre pérennité ? La qualité des sources est également très variable selon la précision des recueils : caméras, réseaux sociaux, commerce électronique. Tout cela pose des questions contractuelles : quel modèle économique pour une coopération entre des services publics et des entreprises privées motivées par le profit, et légales (obligation de transmission).

L'accréditation de sources de données *Big Data* pour leur utilisation en statistique officielle est d'ailleurs documentée par Eurostat.

Le métier de statisticien public va en effet être amené à changer profondément, tout d'abord pour des raisons essentiellement technologiques : les statisticiens publics devront monter en compétences sur les aspects informatiques, s'approprier de nouvelles méthodologies d'analyse (*machine learning*) et être encore plus vigilants sur la protection des données et l'éthique. Vu la rapidité avec laquelle les données et les outils évoluent, les statisticiens publics doivent sortir de leur tour d'ivoire : tous ne deviendront pas des data scientists, mais ils devront collaborer avec les data scientists et ingénieurs du privé et les chercheurs universitaires. Le travail en équipe multidisciplinaire sera une nécessité car un même individu ne peut cumuler toutes les compétences. On qualifie de « *iStatisticien* » ce nouveau métier[9].

S&S : En quelques mots, quelles principales innovations dans la formation des jeunes pensez-vous devoir être introduites pour répondre à la nouveauté du *Big Data* ?

GS : Les jeunes statisticiens doivent être mieux formés aux nouvelles technologies informatiques évoquées dans la deuxième question de cet entretien, connaitre les principes des systèmes de gestion NoSQL (Not only SQL) et savoir coder dans des langages comme Python. Sur le plan méthodologique, les méthodes computationnelles et d'apprentissage doivent être enseignées avec les méthodes de l'analyse statistique multivariée. Dans le cadre de la grande dimension, les méthodes de régularisation doivent trouver leur place à côté des grands classiques que sont les moindres carrés et le maximum de vraisemblance.

Il faut que les étudiants aient accès à de grandes bases de données pour s'exercer. La participation à des compétitions comme Kaggle (https://www.kaggle.com/), Datascience (https://www.datascience.net/fr/home/), Challenge data (https://challengedata.ens.fr/fr/home) doit être encouragée et mieux, faire partie intégrante des cursus, à l'instar de la formation continue « Data Science pour l'actuariat » de l'Institut des actuaires.

Je suis optimiste sur l'évolution en cours des masters et des écoles de statistique en France car les enseignants-chercheurs et les directions ont bien saisi les enjeux et ont les compétences nécessaires. Par contre d'autres formations devraient faire leur *aggiornamento*, telles les

9. http://tietotrendit.stat.fi/mag/article/153/

licences d'économie et gestion où les cours de statistique datent souvent d'avant la révolution numérique. Je recommande à ce sujet la lecture de l'article de Hal Varian, célèbre professeur de microéconomie, devenu chef économiste de Google, qui encourage ses collègues économistes à aller voir du côté des algorithmes de *machine learning* et ne pas se contenter des classiques régressions linéaire et logistique[10].

S&S : Le grand public connait déjà le terme *Big Data* et certaines utilisations pratiques du Big Data. Comment expliquez-vous cela alors que la recherche en *Big Data* semble en être à ses débuts ?

GS : On ne peut pas dire cela. Même si « *Big Data* » n'est pas toujours dans les mots-clés, la recherche en *Big Data* est active depuis plus de 10 ans tant du côté des informaticiens que des statisticiens, et ce n'est pas que de la technologie.

Le traitement des flux de données, de la très grande dimension suscite des travaux très pointus, publiés dans de grandes revues. L'analyse des réseaux sociaux avec leurs immenses graphes a ressuscité l'intérêt pour la théorie des graphes et conduit à des travaux originaux.

L'apprentissage profond (*deep learning*) dont la dénomination date de 2006, a renouvelé les problématiques des réseaux de neurones et conduit à des réalisations spectaculaires en apprentissage supervisé (reconnaissance d'images, intelligence artificielle) car on peut maintenant entrainer des modèles comportant des milliers de paramètres grâce aux bases de données massives.

Le Big Data permet de poser de manière nouvelle le problème de la causalité : la *National Academy of Sciences* américaine a organisé en 2015 un colloque passionnant intitulé « Drawing Causal Inference from Big Data » réunissant entre autres : Léon Bottou, Peter Bühlmann, Michael Jordan, Judea Pearl, Bernhard Schölkopf, Hal Varian[11]. La même année, la revue Political Science and Politics publiait un cahier spécial avec 8 articles sur « Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science? ».

Du côté de la statistique officielle, cela bouge aussi très vite et de nombreux travaux de recherche appliquée sont présentés dans les conférences NTTS (New Techniques and Technologies for Statistics) organisées par Eurostat, qui rassemblent tous les deux ans plusieurs centaines de participants.

Je conclurai sur deux citations. La première est de George E.P. Box :

> « *Les statisticiens, comme les artistes, ont la mauvaise habitude de tomber amoureux de leurs modèles* »[12]

Et la seconde de W.Edwards Deming :

> « *Les données scientifiques ne sont pas collectées pour des musées ; elles sont collectées comme une base pour une action. S'il n'y a rien à faire avec des données, il n'y a aucune utilité à les collecter. Le but ultime de la collecte des données est de fournir une base pour l'action, ou une recommandation pour une action. L'étape intermédiaire entre la collecte des données et l'action est la prédiction.* »[13]

10. H.Varian (2014) Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28, 2, 3–28
11. Le site http://www.nasonline.org/programs/sackler-colloquia/completed_colloquia/Big-data.html contient les vidéos des conférences dont certaines ont été publiés en 2016 dans les PNAS (volume 113, n°27)
12. « *Statisticians, like artists, have the bad habit of falling in love with their models.* »
13. « *Scientific data are not taken for museum purposes; they are taken as a basis for doing something. If nothing is to be done with the data, then there is no use in collecting any. The ultimate purpose of taking data is to provide a basis for action or a recommendation for action. The step intermediate between the collection of data and the action is prediction.* » Deming W.-E.(1942), On a Classification of the Problems of Statistical Inference, *Journal of the American Statistical Association*. Le Dr. Deming a écrit cet article alors qu'il travaillait pour le Bureau du Census des États-Unis.

LIBRE OPINION

Statistique et recherche interdisciplinaire - Implication d'une discipline sans objet

Francis LALOË

Statisticien – Ancien directeur de recherches à l'Institut de recherches pour le développement (IRD)[1]

L'application de la statistique consiste en la recherche de réponses, sous forme de statistiques, à des questions posées en dehors de la discipline. La statistique appliquée est de ce point de vue sans objet propre, exerçant une activité de service. Mais cette « définition » ne suffit pas à caractériser la statistique appliquée. En effet, le champ d'application peut éventuellement s'élargir lorsque le service conduit à produire des résultats relatifs à un domaine plus vaste que celui défini par les questions initiales. La statistique participe alors à la dynamique du questionnement et son rôle dépasse le cadre de l'application pour entrer dans celui de l'implication.

Cet élargissement peut être légitime ou non, selon la nature et le contexte du service initialement demandé, service que le statisticien a accepté de rendre. Les choses peuvent ainsi être très différentes selon qu'il est rendu au sein d'un projet mono disciplinaire où la question relève du point de vue d'une discipline (ou d'une combinaison donnée a priori de disciplines), ou bien qu'il est rendu dans le cadre d'un projet interdisciplinaire dans lequel les points de vues de plusieurs disciplines sur un objet commun peuvent être confrontés.

Dans tous les cas, au départ, le service consiste en la production d'une synthèse, selon une statistique de dimension réduite par rapport à celle des données dont elle est une fonction. C'est ce que le grand statisticien Ronald Fisher (1890-1962) a résumé en des termes très généraux en précisant les qualités d'une telle synthèse :

> « *L'objet des méthodes statistiques est la réduction des données. Une certaine quantité de données, qui en général du simple fait de sa masse ne peut pas entrer dans l'esprit, doit être remplacée par un nombre relativement petit de quantités qui représenteront adéquatement l'ensemble, ou qui, en d'autres termes, contiendront la plus grande part possible de l'information contenue dans les données d'origine – idéalement, la totalité.* »[2]

1. Ce texte présente, en reprenant son titre, mon livre publié en 2016 dans la collection « Indisciplines » des éditions Quae. Voir : http://www.quae.com/fr/r4969-statistique-et-recherches-interdisciplinaires.html
2. « *[…] the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data* » Fisher R. A., (1922), On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, A, 222, p. 309-368.

L'objet de la statistique est donc de fournir une représentation d'information et non la représentation d'un objet ou d'un phénomène[3]. Cette représentation doit être la plus courte (synthétique) possible et contenir le plus possible de l'information, voire sa totalité[4]. Ces deux qualités sont en première analyse contradictoires dès lors qu'une synthèse ne permet pas la reconstitution exacte de l'observation en « gommant » les différences entre les jeux possibles de données dont les synthèses sont identiques[5]. D'une manière générale, ces différences sont d'autant plus nombreuses que la synthèse est courte.

Une situation optimale est celle dans laquelle une telle statistique est en lien direct avec la question posée. C'est ce qu'on recherche en construisant un protocole d'observation[6] exclusivement dédié à cette question, auquel cas l'objet et la question sont en quelque sorte confondus. Par exemple, dans le cas d'une recherche relative à l'effet de diverses pratiques de culture sur le rendement de diverses variétés d'une plante, deux points de vue, sur les pratiques et sur les variétés, sont combinés a priori à l'intérieur d'un « plan d'expérience ». La théorie enseigne un résultat essentiel : si le plan est bien conçu, on peut répondre aux questions relatives aux sources de variation indépendamment les unes des autres et l'ensemble des estimations est bien relatif au questionnement qui est à l'origine de la construction de l'expérience. La qualité du protocole d'observation est donc essentielle.

Si le plan d'expérience est « déséquilibré », on ne peut pas estimer les effets des sources indépendamment les uns des autres. Cela peut parfois conduire à des malentendus lorsque la personne en charge du traitement des données doit conclure que la qualité du protocole d'observation ne permet pas d'estimer les effets d'une source sans faire d'hypothèse sur les effets des autres. C'est ce que Fisher a parfaitement résumé ainsi :

> « *Consulter le statisticien après la fin d'une expérience, cela revient souvent à lui demander seulement une autopsie. Il peut peut-être dire de quoi l'expérience est morte* »[7]

Ici, la personne en charge du traitement peut être un statisticien qui a accepté de rendre un service ou bien la personne qui a posé la question et qui maîtrise, et parfois même conçoit et développe, l'outil statistique. Cette double compétence est ainsi affichée par certaines disciplines « hybrides » telles que la biostatistique ou encore l'économétrie… au prix d'une difficulté d'identification comme l'indique à propos de Fisher son collègue L.J.Savage :

> « *Je rencontre de temps en temps des généticiens qui me demandent s'il est vrai que le grand généticien R.A.Fisher était aussi un statistician important* »[8].

3. Ce point de vue est développé dans : Varenne F., (2010), *Formaliser le vivant : Lois, Théories, Modèles ?* Visions des sciences, Hermann
4. R. Fisher a identifié des situations dans lesquelles une synthèse de dimension réduite peut contenir toute l'information. Il définit ainsi la qualité d'exhaustivité (traduction de « sufficiency ») pour une statistique :
« *[a sufficient statistic] is equivalent for all subsequent purpose of estimation to the original data from which it was derived* » Fisher R. A. (1925), Theory of statistical estimation, Proc. Camb. Philos. Soc., 22, p. 700-725. Cette qualité ne fait pas référence à la question initiale à laquelle on cherche à répondre mais elle établit que toute l'information contenue dans les données et relative à cette question pourra être exprimée comme une fonction d'une statistique exhaustive
5. Cette qualité peut s'exprimer naturellement en référence à la définition d'une information selon Bateson : « une différence qui fait une différence ». Une synthèse gomme des différences qui ne font pas de différences ou, en d'autres termes, qui laissent indifférents. Voir : Bateson G. (1972), *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*, University of Chicago Press
6. L'histoire et l'étymologie du mot observation (Dictionnaire historique de la langue française, Dictionnaires Le Robert, Paris, 2000) rendent bien compte de l'idée de service ; le mot peut tout autant être relatif à ce qui est observé (dans le sens d'une règle qu'on respecte) qu'à l'action même de l'observer (en le respectant ou le décrivant). Au delà, la construction du mot renvoie à ce qui est « au devant de », « au vu de » (ob) et ce à quoi on est attentif (servare : préserver, sauver…)
7. « *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of* » Citation extraite de : Fisher R.A. (1938) Presidential address to the first indian statistical congress *Sankhya* 4
8. « *I occasionally meet geneticists who ask me whether it is true that the great geneticist R.A. Fisher was also an important statistician* » Savage L. J. (1976), On rereading R. A. Fisher, The Annals of Statistics, 4(3), 441-500

La collaboration au sein d'un projet interdisciplinaire

Dans ce qui précède, il s'agissait de répondre à une question explicite et de construire un objet (une expérience) dédié à cette question. La situation est très différente dans le cas d'une participation à un projet interdisciplinaire dans lequel les points de vue de plusieurs disciplines sur un objet commun peuvent être confrontés. Cet objet a une existence propre, souvent attestée sous forme d'un enjeu ou d'un domaine général au sein même du titre des programmes interdisciplinaires.

Dans ce contexte, la référence explicite à un tel objet commun est nécessaire et l'exemple d'une exploitation halieutique sera utilisé dans les propos qui vont suivre : Il s'agit d'un objet réel dont on ne peut pas faire à ce titre une construction définitive et unique. On peut, par contre, en faire une multitude de représentations, chacune étant une construction. Cette multiplicité est inéluctable dès lors que l'objet commun peut être observé depuis plusieurs points de vue, chacun privilégiant un ou plusieurs éléments particuliers. Les constructions sont autant de combinaisons de ces éléments ; elles sont nécessairement relatives à un système complexe selon la définition proposée par Jean-Marie Legay[9] :

> « *Est complexe un système que la perte d'un de ses éléments fait changer de nature* »[10]

Une procédure d'observation peut être associée à chaque construction et une observation peut par ailleurs être restituée selon plusieurs d'entre-elles[11] ; l'observation est une confrontation à la réalité. Il s'agit d'une expérience, selon la définition également proposée par Legay :

> « *[On appelle expérience] toute procédure organisée d'acquisition d'information qui comporte, dans la perspective d'un objectif exprimé, une confrontation avec la réalité* »[12]

Dans ce cas, la confusion entre sources de variation ne résulte plus nécessairement d'une erreur de protocole. Dans le domaine halieutique on peut ainsi supposer, pour une espèce donnée, que les captures obtenues lors d'actions de pêche sont des réalisations de variables aléatoires dont l'espérance est positivement corrélée avec l'abondance. On peut même faire l'hypothèse que les rendements sont proportionnels à cette abondance en supposant que chaque poisson a la même probabilité d'être capturé lors d'une action de pêche (capturabilité constante). Mais il est possible que ces variations de rendements soient (aussi) liées au comportement des poissons qui peuvent être plus ou moins accessibles… et bien sûr les variations d'accessibilité peuvent conduire les pêcheurs à adapter leur pratique, soit pour maintenir le contact avec l'espèce qu'ils recherchent – auquel cas ils réduisent les variations d'accessibilité – soit pour rechercher une autre espèce plus accessible – auquel cas ils exacerbent ces variations.

Une série chronologique de rendements de pêche peut donc faire l'objet de synthèses selon différents points de vue, chacun offrant une interprétation cohérente[13]. On peut y voir des variations d'abondance ou des variations d'accessibilité de la ressource ou encore le reflet des décisions des pêcheurs dans le déroulement de leurs actions. Le problème est alors qu'en privilégiant un de ces points de vue on fait une hypothèse, généralement implicite, sur ce qui pourrait être vu à partir des autres : assimiler des variations de rendements à des variations d'abondance implique qu'on suppose que la capturabilité reste constante et que les pêcheurs font toujours la même chose… La confusion d'effets peut être masquée par le discours et être

9. Jean-Marie Legay biométricien français décédé en 2012
10. Legay J.-M. (1997), *L'expérience et le modèle. Un discours sur la méthode*, coll. « Sciences en questions », Inra.
11. La possibilité d'utiliser une observation selon plusieurs points de vue est sans doute une qualité propre à un observatoire.
12. Legay J.-M. (1993), Une expérience est-elle possible ? *In* J.-D. Lebreton et B. Asselain (eds), *Biométrie et environnement*, Masson, Paris, pp. 1-14
13. Legay (1997) considérait ainsi que la « complexité est d'une décision » dont une des conséquences est la perte des critères d'évidence » au profit de « réseaux de cohérence ».

source de malentendus, de manipulations, de négation de l'intérêt d'autres points de vue et d'incapacité de communication entre disciplines[14] lorsqu'un point de vue disciplinaire conduit à masquer les variations perceptibles selon les points de vue d'autres disciplines.

Si la question n'était qu'une de celles évoquées ci-dessus, être confronté à cette difficulté indique que l'expérience est morte et qu'il faut reconsidérer le protocole d'observation. Ainsi pour estimer un indice d'abondance selon une moyenne de captures réalisées lors d'actions entreprises par des unités de pêche[15], il faut utiliser les captures dont l'espérance est égale, à un facteur de proportionnalité près, à cette abondance (capturabilité constante). On doit donc privilégier l'observation des unités qui recherchent toujours la même espèce, avec la même efficacité[16]. Il se peut que cette qualité soit associée aux caractéristiques des unités de pêche selon, par exemple, qu'elles peuvent ou non poursuivre une espèce dans ses déplacements. Il sera alors légitime de privilégier l'observation des unités industrielles qui ont un rayon d'action supérieur à celui des unités artisanales[17]. Il en découle un problème de déontologie[18] si l'objet de référence est l'exploitation halieutique et qu'on ne peut exclure a priori que toutes les unités de pêche puissent faire l'objet de questions légitimes. En effet, si un protocole construit en référence à une question particulière conduit à négliger l'observation de l'activité et des résultats de certaines unités de pêche, il en découle une moindre capacité de réponse aux questions relatives à ces unités. Ce problème traduit un paradoxe quant au profil professionnel de la personne qui construit le protocole et « produit » la synthèse. S'il s'agit de celle qui a posé la question et qui maîtrise l'outil statistique, elle aura mis son expertise en œuvre de façon à ce que le protocole d'observation qu'elle aura mis en place – ou au moins validé – produise un résultat dont la synthèse soit sans ambigüité en relation avec la question qu'elle a posée ; il ne faut pas que d'autres questions soient soulevées par l'analyse des données… La situation est évidemment totalement différente si cette personne est un statisticien qui accepte de répondre à des questions, posées par d'autres, relatives à un objet ayant une existence propre.

14. Cette difficulté apparaît dans le titre même d'un article de Myers et Worm : « Rapid worldwide depletion of predatory fish communities. » paru dans la revue Nature (numéro 423 - 2003) où des données de rendements de pêche sont implicitement supposées parfaitement refléter des abondances. Un site web réunit un nombre important de réactions critiques http://www.soest.hawaii.edu/PFRP/large_pelagics/large_pelagic_predators.html
15. « Unité de pêche » consiste ici en *une « entreprise » qui pratique la pêche dans le site considéré, avec un centre de décision, des moyens et des connaissances (pirogues, bateaux, outils, savoir faire…).*
16. Plus précisément, ces moyennes peuvent être pondérées en admettant que les unités de pêche peuvent être caractérisées par des capturabilités différentes selon la puissance de leurs moteurs, la taille ou le type d'engin de pêche etc. On suppose alors que la capturabilité d'une action dépend de l'unité de pêche qui l'entreprend, mais que cette capturabilité reste constante pour une unité de pêche donnée.
17. Dans de nombreux cas aucune des unités de pêche n'est satisfaisante et le recours à des campagnes scientifique est privilégié, posant des difficultés liées aux coûts de ces campagnes…
18. Au sens d'un ensemble de règles qui régissent les rapports entre un professionnel et son ou ses clients

Un exemple

Ceci peut être illustré par un exemple « simple » issu d'une enquête sur les résultats obtenus par des unités de pêche lors de sorties quotidiennes réalisées au cours de six jours consécutifs en avril 1978 à partir du village de Kayar, un site majeur de pêche artisanale au Sénégal situé à une cinquantaine de kilomètres au nord de Dakar.

Tableau 1 : Effectifs des échantillons et captures moyennes par sortie (Kilogrammes) selon diverses espèces en fonction du jour

Jour	Effectif des échantillons	Captures moyennes par sortie (kg) selon les espèces						
		tiof	chinchard	tassergal	pageot	sarda	autres	Total
1	96	1.25	2.66	25.07	3.28	3.18	6.71	42.15
2	141	2.37	3.61	21.05	2.07	0.09	2.56	40.75
3	143	4.40	2.66	5.11	3.12	10.39	3.35	29.03
4	123	3.22	4.42	0.24	9.05	14.53	4.31	35.77
5	123	2.76	5.02	1.47	8.80	10.62	6.35	34.94
6	105	2.17	8.34	1.32	12.64	3.11	8.27	35.84

Source : Laloë F., Bergerard P., Samba A. (1981), Contribution à l'étude de la pêcherie de Kayar : étude d'une partie des résultats du sur-échantillonnage de 1978 concernant les pirogues motorisées pêchant à la ligne, *Documents Scientifiques* - CRODT, 79, p. 45 p. multigr.

Ces observations étaient faites dans le cadre de la conception d'un système d'enquêtes en vue d'estimer les captures réalisée[19] par la pêche artisanale. Dans le contexte d'un système à deux niveaux (sélection de jours d'enquêtes et sélection d'actions de pêche lors de ces jours), la présence d'effets jours importants (variance inter jours élevée) militait pour que les enquêtes aient lieu le plus grand nombre de jours possible. La question portait donc sur l'existence et l'importance de ces effets.

L'examen du tableau montre que ces différences sont flagrantes et considérables. Pour l'espèce tassergal le rendement observé est divisé par 100 entre le premier et le quatrième jour… Une interprétation cohérente (confortée par d'autres éléments tels que l'évolution de la fréquentation des lieux de pêche) met en relation cette chute avec l'augmentation des rendements en pageot et chinchard : une diminution de l'accessibilité du tassergal, espèce de pleine eau pêchée en poursuivant les bancs de poissons, a conduit les pêcheurs à rechercher, en ancrant leurs pirogues, des espèces vivant sur le fond. Le choix de l'une ou de l'autre de ces méthodes est une décision et les captures peuvent donc être des réalisations de variables aléatoires de distributions différentes. Si on suppose que les pêcheurs ne disposent pas d'un tel choix, leurs résultats sont des réalisations de variables ayant toutes la même distribution, dont les espérances peuvent éventuellement dépendre du jour de pêche. Dans ce cas les moyennes par jour sont des synthèses qui contiennent beaucoup d'information : si les captures réalisées pour une espèce lors des sorties faites le jour i sont des réalisations de lois normales indépendantes[20] d'espérance m_i et de variance S_i^2, les moyennes présentées dans le tableau 1 sont les meilleures estimations possibles des mi et elles constituent une partie importante d'une statistique exhaustive minimale ; il y manque les éléments relatifs à l'estimation des variances… Il y a donc une adéquation entre les trois éléments que sont la question de l'existence d'effets

19. Ce système est stratifié, chaque strate réunissant les actions de pêche réalisées à l'aide d'un engin donné, pendant une période de temps donnée et à partir d'un site donné. Les données utilisées ici relèvent toutes d'une même strate (lignes à main à partir de Kayar au cours d'une même semaine) et on « aimerait » à ce titre pouvoir les considérer comme un échantillon aléatoire simple… La présence d'effets jours met à mal cette hypothèse.
20. L'indépendance peut être assurée par le respect des règles de l'échantillonnage aléatoire simple lors de la sélection des échantillons quotidiens d'actions de pêche… Ce qui est loin d'être évident au vu des contraintes de terrain.

jours dans le contexte de la mise en place d'un système d'enquêtes, la forme des synthèses par espèce présentées au tableau 1 et une hypothèse de distribution des observations dont rend compte l'équation :

$$Y_{ik} = m_i + e_{ik}$$

(e_{ik} est l'écart entre la $k^{ième}$ observation du jour i et l'espérance m_i). En ce sens l'expérience conduisait à conclure en faveur de l'affectation d'un enquêteur à plein temps pour estimer les captures, et à estimer les indices d'abondance à partir des résultats d'autres composantes de la pêcherie.

Mais la discussion sur les moyennes observées, avec l'hypothèse cohérente d'un changement rapide d'accessibilités, auquel les pêcheurs s'adaptent en changeant d'espèces cibles, soulève une question « délicate ».

S'il s'agit de bien connaître la dynamique de la ressource exploitée, entre autres pour permettre sa gestion « rationnelle », il convient, comme évoqué plus haut, de privilégier l'observation des unités de pêche dont le comportement conduit à stabiliser la capturabilité. Ceci est d'autant plus utile que l'activité de ces unités peut être directement liée à la mortalité qu'elles provoquent et qu'elle peut donc être une variable de contrôle efficace de cette mortalité.

Par contre, si l'objet est l'exploitation halieutique, la présence de différentes flottes de pêche ne peut plus a priori être discutée du seul point de vue de la dynamique de la ressource qui amène à porter des jugements de valeurs négatifs sur les unités dont les pratiques les conduisent à générer un impact variable. Comprendre cette variabilité conduit à poser la question de la variabilité de la capturabilité causée par les décisions des pêcheurs. Cette question est orpheline dans la mesure où elle conduit pour les sciences de la vie à des difficultés de description de leur objet et où, en portant sur la dynamique de la ressource biologique, elle n'est pas directement relative à l'objet des sciences humaines.

Mais si cette question est légitime, il faut l'adopter et traiter « en ce sens » les données qui ont conduit à la soulever. Elle est légitime ici parce que relative à des unités de pêche dont la présence est essentielle pour de nombreuses raisons sociales et économiques et parce que l'activité de ces unités peut être viable parce qu'elles peuvent à tout moment rechercher une espèce à leur portée en choisissant la méthode qu'elles estiment être la plus utile parmi celles dont elles disposent. Chacune de ces méthodes j peut être caractérisée par une distribution d'espérance m_{ij} le jour i. En étant choisie avec une probabilité p_{ij} le modèle de distribution peut s'écrire :

$$Y_{ik} = \sum_{j=1}^{J} p_{ij} m_{ij} + e_{ik}$$

Les p_{ij} résultent des décisions des pêcheurs, engendrant ainsi la variabilité de leur impact.

Selon ce modèle de distribution, les moyennes simples du tableau 1 sont de qualité nettement moindre en termes de contenu d'information et la construction d'un modèle[21] est proposée, permettant d'exprimer les probabilités p_{ij} sous forme de fonctions de ses paramètres. Il articule la dynamique d'une ressource plurispécifique, à l'aide de modèles de production, avec la dynamique de l'exploitation menée par des unités de pêche réunies selon des flottes de pêches caractérisées par les ensembles de méthodes disponibles. La probabilité de choisir une méthode est estimée à l'aide d'un « modèle d'utilité aléatoire »[22]. Les paramètres du modèle

21. Le terme de modèle peut être appliqué à plusieurs choses différentes. Peut-être le terme de cadre serait-il préférable. Cette construction est décrite dans le livre que ce texte présente.

sont estimés en recherchant les valeurs qui conduisent à reconstituer des données d'activités et de rendements de pêche les plus proches possibles (selon un critère de moindres carrés) de celles résultant des observations collectées dans le cadre du système d'enquêtes.

En guise de conclusion

A l'issue de cette opération, il est possible de répondre, sous la forme de fonctions des estimations des paramètres, à des questions faisant intervenir les décisions des pêcheurs. Quelques exemples sont décrits dans le livre présenté ici.

En termes statistiques, on peut considérer que les estimations de l'ensemble des paramètres du « modèle » constituent une statistique contenant le plus possible de l'information présente dans les données « traitées ». Si certains des paramètres peuvent avoir une interprétation directe (des prix, des coûts, des capturabilités, des effectifs de flottes de pêche…), l'intérêt de cette statistique est de pouvoir être utilisée pour répondre, sous forme de fonctions de ses éléments, à des questions qui peuvent ne pas être toutes identifiées au départ. En ce sens, cette statistique peut être vue comme une première restitution de l'information collectée dans le cadre d'un observatoire. L'idéal serait de produire une statistique exhaustive minimale, en étant alors certain que toute meilleure réponse à une quelconque question serait une fonction de ses éléments… Cet idéal constitue une référence et un outil critique très utiles même si, dès lors qu'il n'y a pas de représentation unique et définitive, une telle statistique n'existe pas.

Une analogie avec le système judiciaire peut être utile : Le juge d'instruction instruit à charge et à décharge. Il est censé réunir, organiser et mettre à disposition des différentes parties toute l'information utile relative à l'affaire qui doit être jugée. Les avocats vont disposer de ces informations pour les restituer, chacun selon l'intérêt de la partie qu'il représente (il s'agit bien d'une représentation !). Le statisticien doit au moins assumer le rôle du juge d'instruction… mais il peut aussi restituer l'information selon un point de vue particulier. Dans le cadre d'un projet monodisciplinaire, ces rôles doivent pouvoir être confondus grâce à la mise en place d'un protocole d'observation adéquat. A l'inverse, dans celui d'un projet interdisciplinaire, ces deux rôles ne peuvent être confondus ; et si le statisticien est amené à les tenir, il est déontologiquement nécessaire qu'il précise en toute circonstance lequel de ces deux rôles il tient, et au service de qui.

22. Random Utility Model (RUM) dont une description a été proposée par Mac Fadden. On considère que chaque décideur réalise une estimation de l'utilité de chacune des options à sa disposition et qu'il choisit celle dont l'estimation est la plus élevée. Voir : McFadden D. (1973) Conditional logit analysis of qualitative choice behavior. In Zarembka P., éditeur *Frontiers in econometrics*, pages 105 – 142. Academic Press, New-York.

ACTUALITÉ

Données personnelles : quels nouveaux droits ?

Entretien avec

Judith ROCHFELD

Professeure à l'École de droit de la Sorbonne, directrice du Master 2 « Droit du commerce électronique et de l'économie numérique », co-directrice de l'Institut de recherche juridique de la Sorbonne[1]

Statistique et société a présenté dans son numéro d'automne de 2016 les principales dispositions de la loi « Pour une République numérique » concernant les *données publiques* (titre I de la loi). La question des *données personnelles* est au moins d'égale importance. Elle concerne tous les utilisateurs de services informatiques en ligne, c'est-à-dire tout le monde ; elle concerne aussi particulièrement les producteurs et les utilisateurs de travaux statistiques, car ceux-ci ont de plus en plus affaire à des données sur des personnes, et doivent respecter dans leur pratique professionnelle le droit en vigueur. Le titre II de la loi République numérique est consacré à cette question, qui vient également de faire l'objet d'un règlement européen. Celui-ci va se substituer à la plupart des dispositions de la loi Informatique et Libertés que la France connait depuis 1978. Nous avons demandé à Judith Rochfeld, spécialiste du droit de l'économie numérique, d'éclairer nos lecteurs sur les nouveautés contenues dans ces deux textes.

Statistique et société : La loi « Pour une République numérique »[2] promulguée en octobre 2016 contient des dispositions sur « la protection des droits » dans son titre II. Cette même année, l'Union Européenne a adopté le règlement 2016/679 « Protection des personnes physiques à l'égard du traitement des données à caractère personnel ». Comment ces deux textes s'articulent-ils ? Qu'est-ce qui va rester original dans le droit français ?

Judith Rochfeld : Le texte européen sera applicable partout dans l'UE le 25 mai 2018 : à partir de ce moment, conformément aux principes généraux du droit, ce texte, dans son champ d'application, se substituera aux lois nationales[3], et sera le seul texte applicable. Une discussion reste possible sur des questions qui seraient à la frontière du champ d'application de ce règlement. Par exemple, la loi française pour une république numérique contient des dispositions sur « la mort numérique » différentes de celles qui figurent dans le règlement européen. Si l'on considère que ce sujet était dans le champ couvert par le règlement européen,

1. Judith Rochfeld est l'auteur de nombreux livres et articles, parmi lesquels *A qui profite le clic ? Le partage de la valeur à l'ère numérique*, co-écrit avec Valérie-Laure Bénabou - O. Jacob, 2015 ; et Le projet de loi pour une République numérique : entre espoirs et regrets - *Dalloz IP/IT* n°1 janvier 2016
2. Loi 2016-1321 du 7 octobre 2016
3. Notamment plusieurs articles de la loi « Informatique et Libertés » de 1978. Cela dit, de nombreuses dispositions de la loi Informatique et libertés resteront en vigueur : par exemple, toutes celles qui portent sur le rôle et l'organisation de la Commission nationale de l'informatique et des libertés (Cnil).

les dispositions de la loi française cesseront en 2018 d'avoir force juridique. Mais on peut aussi considérer que c'est un sujet qui relève du droit des successions, et se trouve donc en dehors de la compétence européenne : auquel cas la loi française restera applicable.

Par ailleurs, ce règlement a la particularité de comporter beaucoup d'options constituant des marges de manœuvre nationales. C'est le cas en ce qui concerne les éventuelles « actions collectives » pour faire respecter le droit à la protection des données personnelles. De telles actions pourraient par exemple être intentées par des individus touchés par une faille de sécurité du système informatique d'une entreprise dont ils sont clients (en l'état du droit actuel, cette action est possible pour obtenir la cessation d'un manquement à la loi du 6 janvier 1978). Le règlement européen évoque cette possibilité, mais ne va pas plus loin : il appartiendra aux lois nationales de préciser la portée de cette disposition.

S&S : Est-ce que la loi nationale pourrait aller au-delà du règlement européen, être plus protectrice ?

JR : Attention ! Rappelez-vous le titre complet du règlement européen : c'est un règlement relatif « à la protection des personnes physiques à l'égard du traitement des données à caractère personnel *et à la libre circulation de ces données* ». Le deuxième membre de phrase ne doit pas être oublié : le règlement n'est pas seulement un outil de protection des personnes, c'est aussi un instrument de régulation dans le sens du marché unique. Si une loi nationale protégeait davantage que le règlement, elle créerait une distorsion de marché, en défavorisant les entreprises qui y seraient soumises par rapport aux autres. Par exemple, témoignage de cette protection des utilisations des données, les « intérêts légitimes des responsables de traitements » sont reconnus, par ce règlement comme par la directive qui le précédait, comme une cause légitime de traitement des données des usagers (d'une plateforme par exemple).

S&S : Que recouvre cette expression « intérêts légitimes des responsables de traitements » ?

JR : C'est un des six fondements qui peuvent rendre légitime un traitement de données personnelles. Ces fondements sont : 1° Le consentement des personnes concernées ; 2° La nécessité pour l'exécution du contrat auquel la personne concernée est partie ; 3° La nécessité pour le respect d'une obligation légale à laquelle le responsable du traitement est soumis ; 4° la sauvegarde des intérêts vitaux de la personne concernée ou d'une autre personne physique ; 5° La nécessité pour l'exécution d'une mission d'intérêt service public ; 6° La nécessité du traitement pour les intérêts légitimes du responsable du traitement (en balance avec les droits fondamentaux de la personne concernée par le traitement). D'après ce dernier motif et selon l'interprétation qui en est faite, un opérateur pourrait invoquer son modèle économique pour justifier de traiter des données personnelles. Dans quelle mesure cette invocation est-t-elle acceptable ? C'est typiquement une question d'interprétation, susceptible d'être différente selon les pays, et arbitrable en dernier ressort par la Cour de Justice de l'UE.

S&S : La définition de ce qu'on considère comme des « données à caractère personnel »[4] dans le champ juridique a-t-elle évolué avec ces nouveaux textes ? Toutes les données fournies par une personne physique à des fournisseurs de services informatiques sont-elles des données à caractère personnel au sens juridique ?

JR : Dès lors qu'une donnée est reliée à une personne identifiée ou identifiable, elle est considérée comme « personnelle » en droit européen : « porter une cravate grise » peut être

4. « Données à caractère personnel » est le terme exact, retenu aussi bien dans la loi française que dans le règlement européen. Dans le langage courant, on utilise plus souvent « Données personnelles ».

une donnée à caractère personnel… Cette définition, bien stabilisée désormais, est considérée comme très large par les Américains. L'une des nouveautés du texte européen est de parler, à l'inverse, de « donnée anonyme » lorsque tout lien avec la personne a été rompu, si tant est que cela soit possible, et de « donnée pseudonymisée »[5]. Cela dit, les données personnelles ne sont pas toujours « fournies » par les personnes concernées : elles peuvent aussi être captées à l'insu de ces personnes, notamment lorsqu'une telle captation est la contrepartie tacite de services gratuits. La définition juridique de ce qui est « collecte » et de ce qui est « traitement » de données personnelles est suffisamment large pour inclure ces cas.

La petite divergence qui existait entre les définitions de la directive européenne et de la loi française à propos de ce qu'est une donnée à caractère personnel subsiste : une information qui nécessiterait, pour identifier la personne à laquelle elle se rapporte, la mise en œuvre de moyens « déraisonnables » n'est pas considérée comme une donnée à caractère personnel par le règlement européen, mais elle l'est, selon la loi française. La référence au caractère « raisonnable » des moyens à mettre en œuvre ne figure pas dans le texte lui-même du règlement, mais dans ses « considérants ». Il est à noter que la Cour européenne de Justice se réfère parfois à ces « considérants ».

S&S : Quels sont les droits nouveaux dont disposent les individus par rapport à leurs données personnelles ?

JR : Rappelons d'abord les droits déjà formulés antérieurement : droit des personnes à être informées sur l'existence d'un traitement de données personnelles la concernant, d'y accéder, de les faire corriger si elles sont inexactes, de les faire supprimer si la durée de conservation correspondant à la finalité du recueil est dépassée. Tous ces droits figuraient déjà dans la directive européenne de 1995 et dans la loi Informatique et Libertés en France. Deux autres droits sont apparus dans la jurisprudence de la Cour de justice de l'Union Européenne, et sont inscrits dans le nouveau règlement : le droit à la portabilité de ses données et le droit au déréférencement.

S&S : Portabilité, comme pour le numéro de téléphone lorsqu'on change d'opérateur téléphonique ?

JR : Non, pas exactement. Dans le cas du numéro de téléphone, il s'agissait, dans un esprit de respect de la concurrence, de protéger les droits économiques des consommateurs qui risquaient d'être prisonniers de leurs opérateurs. Dans le cas des données personnelles, il s'agit surtout d'appliquer un autre principe, désormais inscrit dans l'article premier de la loi Informatique et Libertés, selon lequel « Toute personne dispose du droit de décider et de contrôler les usages qui sont faits des données à caractère personnel la concernant », autrement appelé le principe « d'autodétermination informationnelle »[6]. Selon ce principe, qui instaure une nouvelle conception du rapport de la personne avec ses données, chaque individu se voit reconnaître une certaine autonomie dans la gestion de ses données personnelles. En particulier, s'il souhaite changer d'opérateur pour un service particulier, il doit pouvoir emporter avec lui les données personnelles qu'il avait constituées avec son ancien opérateur[7]. L'application de ce

5. Article 4, 5° du règlement européen : la pseudonymisation fait que les données ne peuvent plus être attribuées à une personne concernée précise « sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable ». Par exemple, il ne suffit pas de remplacer les identifiants personnels – nom, prénom, ou numéro d'inscription au répertoire national d'identification des personnes physiques – par des numéros sans signification pour obtenir un fichier « pseudonymisé » ; encore faut-il qu'aucune combinaison des variables présentes dans le fichier ne permette une identification d'un individu. De telles variables, si elles existent, doivent être conservées séparément, dans un fichier distinct du fichier « pseudonymisé ».
6. Notamment par la Cour constitutionnelle allemande et par le Conseil d'État français.
7. Par exemple, sa propre liste de morceaux de musique préférés.

droit nécessite évidemment que les données en question puissent être séparées des données d'autres personnes, ce qui n'est pas toujours facile, ni même concevable, dans un contexte où les interactions entre utilisateurs d'un service sont intenses, et souvent constitutives de la valeur des données (commentaires, notations, etc.).

S&S : Et le droit au déréférencement ? S'agit-il d'un « droit à l'oubli » sur Internet ?

JR : Certains l'appellent ainsi, mais c'est abusif. Une personne qui s'estime lésée parce qu'une information la concernant est systématiquement associée à une recherche par son nom sur un moteur de recherche peut obtenir que cette information soit « déréférencée » par ce moteur ; mais pas qu'elle soit effacée des sites Internet où elle figure. En quelque sorte, il s'agit de diminuer la visibilité d'une information, non de la supprimer. Ce droit, qui a émergé en 2014 lors d'un arrêt célèbre de la Cour de justice de l'UE[8], a en partie été inscrit dans le nouveau règlement européen. La discussion autour de ce droit reste néanmoins intense : son application peut entrer en conflit avec la liberté d'expression, et fait naître le risque que la mémoire collective soit « subjectivée » par les désirs individuels (que l'on efface une information parce qu'elle est gênante pour un individu)[9]. Des exceptions sont d'ailleurs prévues pour la défense des libertés d'expression et d'information ainsi que pour les besoins des archives, et du travail journalistique ou scientifique ; des critères d'articulation s'élaborent. À cet égard, se pose la question : « qui doit arbitrer ? ». Actuellement, c'est Google qui reçoit des particuliers les demandes de déréférencement et décide, ou non, d'y donner suite : cela revient à donner à un opérateur privé un rôle d'arbitre entre des libertés fondamentales. Ce droit, qui pouvait déjà être extrapolé en France à partir de la loi de 1978 dans une version plus limitée[10], a été « poussé » par la décision de la Cour de justice de l'UE, et cette version plus forte a été confirmée, avec des nuances, dans le règlement européen. En pratique, pour l'instant, il s'applique principalement à l'égard des moteurs de recherche, moins des éditeurs de contenus.

S&S : N'y a-t-il pas une disposition spécifique sur le déréférencement pour les mineurs dans la loi République numérique ?

JR : Vous avez raison : cette loi française prévoit un « droit au déréférencement » pour les mineurs. De façon générale, parmi les critères permettant de juger de l'opportunité d'un déréférencement[11], la sensibilité des données révélées pour les personnes concernées figure en bonne place, à côté du caractère public ou non public de la personne et de l'importance de l'information en question pour le débat général. Pour un mineur, cette sensibilité est forte : le risque d'un préjudice important quant à la vie privée pèse pour eux plus lourd que les autres critères. D'ailleurs aux États-Unis, où il n'existe guère de lois de protection générale en matière de données personnelles, la Californie a instauré un droit de ce genre pour les mineurs.

S&S : La finalité des traitements joue un rôle important dans la loi de 1978 : avant de collecter des données personnelles auprès de quelqu'un, tout opérateur doit l'informer du « pourquoi » de cette collecte, et il doit exister une proportionnalité entre les données collectées et le but poursuivi. Est-ce toujours le cas, à l'heure où les données sont souvent collectées de façon invisible pour l'utilisateur d'un service ou d'une application ?

JR : Oui, c'est toujours le cas. Lorsqu'on vous avertit que des « cookies » vont être installés

8. Arrêt « Google Spain » qui a imposé à la filiale espagnole de Google de déréférencer une information concernant le passé judiciaire d'un citoyen espagnol
9. Vallespi E.et Wales J. (2016) « Le « droit à l'oubli » ne doit pas tourner à la censure » *Le Monde* 26 novembre 2016
10. Lorsque la durée de conservation d'une information apparaissait comme disproportionnée au regard de la finalité du traitement, la personne concernée par cette information pouvait déjà demander qu'elle soit effacée.
11. Critères mis au point entre le « groupe des autorités européennes » (CNIL et ses équivalentes dans les autres pays) et Google après l'arrêt de 2014

sur votre ordinateur, on vous explique que c'est nécessaire pour assurer le service dont vous désirez bénéficier, voire pour l'améliorer. Le fournisseur se met ainsi en conformité avec la loi qui exige de lui d'être transparent sur la finalité de la collecte et sur l'usage qui sera fait des données.

S&S : Est-ce que les droits des individus par rapport à leurs données personnelles ne seraient pas mieux protégés si ces données étaient considérées comme leur propriété, comme faisant partie de leur patrimoine ? N'est-on pas « propriétaire de ses données » ?

JR : C'est une question très importante. La réponse du législateur européen et français est négative : selon lui, la notion « d'autonomie informationnelle » (voir plus haut) est susceptible d'assurer une meilleure protection que la notion de propriété. Et je suis de cet avis. Mais cela mérite une explication détaillée.

Lors des débats internationaux sur ce sujet, l'option « propriété » a été soutenue de plusieurs côtés. Les opérateurs américains, publics et privés[12], ont soutenu être propriétaires des données qu'ils traitent, apurent, enrichissent… Et du côté des défenseurs des droits des personnes, certains estimaient et estiment toujours que la meilleure protection consiste à se prévaloir d'un droit de propriété de chacun sur ses données, permettant d'en réclamer la valeur et d'être associés aux bénéfices de leur utilisation. Pour ces personnes, le rapport aux données est de l'ordre de « l'avoir ».

L'autre option, qui l'a finalement emporté au niveau européen, considère les données personnelles comme liées à « l'être » lui-même, à son identité, dont elles constituent un « reflet numérique ». Et chaque personne doit pouvoir décider comment elle se rend visible par ses données : c'est le fondement du concept d'autonomie informationnelle. Dans cette conception, les données personnelles ne relèvent pas du droit patrimonial, mais du droit de la personne (de la vie privée, de la non-discrimination, etc.). D'ailleurs, les risques principaux que la personne encourt du fait de ses données sont-ils des risques de perte de valeur ? Non, car la valorisation des données d'une seule personne ne représente presque rien, « epsilon » dans un ensemble où c'est la masse globale des données et des liens entre elles qui peut faire l'objet d'une valorisation économique. Les risques réels ne sont pas là : ils concernent les libertés fondamentales des personnes, qui peuvent faire l'objet, sur le fondement de ces données personnelles, de profilages, de discriminations[13], de révélations de leur vie privée, de harcèlement, etc. Toutes choses qui sont plus de l'ordre de « l'être » que de « l'avoir ». La « charte des droits fondamentaux » de l'Union Européenne, dans son article 8, protège d'ailleurs les données comme un droit fondamental, dissocié de la vie privée, mais lié à la personne.

Dès 2014, le Conseil d'État français s'était également prononcé dans ce sens, en soulignant que la protection des données par leur valeur économique individuelle risquait d'être inopérante[14].

Mais le débat reste ouvert au niveau mondial.

S&S : Qu'en est-il dans ces conditions de la « mort numérique » ? Que deviennent nos droits sur nos données personnelles à notre mort ?

JR : La loi République numérique contient à ce sujet une disposition qui s'inscrit bien dans le principe d'autonomie informationnelle : elle prévoit la possibilité pour une personne de

12. Pour faire image : « Obama et la Silicon Valley » !
13. À l'embauche, ou pour l'attribution de prêts, d'assurances…
14. Étude 2014 du Conseil d'État sur le numérique et les droits fondamentaux – disponible sur le site http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/144000541.pdf

désigner quelqu'un pour décider du sort de ses données personnelles après sa mort. Mais la question du traitement successoral de cet ensemble de données reste ouverte et l'on peut également voir là la reconnaissance des données comme « patrimoine » (« avoir », donc).

S&S : On comprend bien ce principe d'autonomie informationnelle, on peut y adhérer, mais est-il susceptible de passer réellement dans les faits ? Comment le rendre effectif ?

JR : La première affirmation de ce principe remonte à 1983, et a été faite par la Cour Constitutionnelle d'Allemagne fédérale à propos du recensement de la population ![15] S'agissant de son effectivité, je crois qu'il faut considérer ce principe comme un « chapeau » pour les droits plus précis déclinés en vertu de lui : droit d'être informé sur les finalités, droit d'accès, de déréférencement, etc. C'est chacun de ces droits particuliers qui doit être rendu effectif, pour donner force au principe général.

S&S : Il n'y a pas que le sort de leurs données personnelles qui peut préoccuper les individus, il y a aussi la question des algorithmes. Les algorithmes prennent des décisions, ou contribuent à des décisions qui influent sur la vie de tous. Faut-il nous protéger contre les algorithmes ? Est-ce que le droit s'est emparé de cette question ?

JR : Cette préoccupation ne date pas d'hier, et la loi de 1978 prévoyait déjà qu'aucune décision affectant une personne ne devait être prise uniquement sur la base d'un traitement automatisé. La loi pour une République numérique a, un moment, eu pour ambition, au titre de l'exigence de « loyauté des plateformes »[16], que celles-ci rendent transparents pour l'utilisateur les modes de traitements des données de ces derniers. Néanmoins, cette « transparence » n'a finalement été imposée que dans les relations entre les administrations publiques et les usagers : l'article 4 de cette loi introduit une obligation pour l'administration, d'expliquer les critères mis en œuvre par l'algorithme pour prendre une décision à l'égard de l'usager si ce dernier en fait la demande (une mention doit le prévenir qu'un algorithme est intervenu). À ce propos, le contentieux de l'application « Admission post-bac » (APB) va être très intéressant à observer : de nombreux parents contestent les affectations proposées à leurs enfants et demandent la justification des calculs faits par cet algorithme. Néanmoins, la loi n'a pas été plus loin et n'a pas cherché à définir de façon générale ce que serait un algorithme « loyal » ou « neutre »… ce qui est d'ailleurs très difficile à concevoir ! À l'occasion des élections présidentielles américaines récentes, de nouvelles préoccupations sont également apparues. Les algorithmes qui filtrent les flots d'actualités pour choisir celles qui seront présentées à une personne particulière (sur son réseau social, sur un moteur de recherche, etc.) se fondent sur les préférences passées de cette personne, telles qu'elle les a révélées par ses clics, ses choix… Il en résulte une « bulle de filtrage » : vous n'êtes plus exposé *in fine* qu'aux informations émanant de personnes qui sont d'accord avec vous ; vous ne voyez même plus les arguments du camp adverse. Faudrait-il, pour empêcher cela, imposer à ces algorithmes des obligations de neutralité, voire de pluralisme, comme celles que le Conseil supérieur de l'audiovisuel impose aux chaines de télévision ? Il y va peut-être du fonctionnement de la démocratie dans une société où l'information obtenue sur Internet et par les réseaux sociaux prend une place de plus en plus grande. La réflexion juridique ne fait que commencer à ce sujet.

15. La RFA avait dû suspendre le recensement de 1983 à la suite d'une campagne de protestations contre le transfert de données sensibles (comme la religion) vers d'autres fichiers. En l'absence d'un équivalent de la loi « Informatique et libertés », la Cour constitutionnelle de Karlsruhe prononça le 15 décembre 1983 un « jugement sur le recensement » (Volkszählungsurteil) qui fit date : tout membre d'une société de droit devait pouvoir déterminer l'usage et la divulgation des données statistiques le concernant. Ce principe d'auto-détermination en matière d'information (informationnelle Selbstbestimmung) imposait de refondre les procédures de collecte et de traitement. [Source : Institut National d'études démographiques 2013 Note d'analyse de François Héran « L'Allemagne et ses minorités ou les surprises du recensement de 2011 »]
16. On pourra consulter : Rochfeld J. et Zolynski C. (2016) La « loyauté » des « plateformes ». Quelles plateformes ? Quelle loyauté ? *Dalloz IP/IT* n°11 novembre 2016

S&S : Le titre I de la loi pour une République Numérique instaure une nouvelle catégorie de données publiques, les « données d'intérêt général », dont l'usage est considéré comme « commun » à tous les membres de la société. Faut-il voir là une innovation juridique porteuse d'avenir ?

JR : Certainement : la question des « communs » se pose de plus en plus dans le droit. Deux types de problèmes ont amené cette question au premier rang des préoccupations de certains juristes, dont je fais partie. Il y a d'abord la crise environnementale internationale : elle a fait naître l'idée que certaines ressources communes de l'humanité devaient être mises « hors marché » pour faire l'objet d'une protection spécifique, qu'elles soient ou non sous la souveraineté des États. Et puis il y a la poussée numérique, l'émergence de l'information, bien « non rival »[17] par excellence, qui appelle le partage des usages et la création en commun. Ce sont là deux sources de renouvellement possible pour le droit, qui appellent à sortir du paradigme de la propriété « absolue », qu'elle soit privée ou publique. Vis-à-vis de la nature, il faut limiter le « pouvoir de détruire » ; s'agissant des usages numériques il faut construire des modèles qui permettent un partage. Plusieurs « modèles de communs » ont été avancés dans la période récente : Élinor Ostrom, prix « Nobel » d'économie en 2009, a décrit la gestion de ressources communes par différentes collectivités, et dégagé des conditions pour que cette gestion soit durable ; Pierre Dardot et Christian Laval, philosophes, mettent plutôt l'accent sur « l'agir commun » pour fonder la notion ; on peut citer aussi le projet (non abouti) de réforme du code civil italien en 2007, qui distinguait les « biens communs » comme ceux qui servent les droits fondamentaux des personnes, biens sur lesquels la communauté serait titulaire de droits même s'ils sont par ailleurs appropriés.

Pour en revenir aux données, la mention des données d'intérêt général dans la loi République Numérique va dans le sens de ce mouvement, de même que la disposition selon laquelle l'auteur d'une recherche scientifique financée à plus de moitié sur fonds publics peut mettre ses résultats en accès libre au terme d'une période d'embargo (6 ou 12 mois), nonobstant les conventions passées avec des éditeurs. On est là dans du « savoir commun » ! En revanche, selon moi, cette problématique ne peut pas s'appliquer aux données personnelles, pour fonder un usage collectif : on ne peut pas forcer quelqu'un à mettre en commun son intimité.

S&S : Pour des juristes, le domaine du numérique est en pleine évolution : ce doit être passionnant ?

JR : C'est passionnant, et en prise avec l'actualité : par exemple, l'arrivée explosive des « objets connectés » de la vie quotidienne va étendre encore le champ de nos préoccupations. Ce secteur des études juridiques où « ça bouge tout le temps » attire des étudiants dynamiques et curieux, qui nous aident beaucoup !

17. « un bien non rival » peut être utilisé par certains sans que cela diminue sa disponibilité pour les autres.