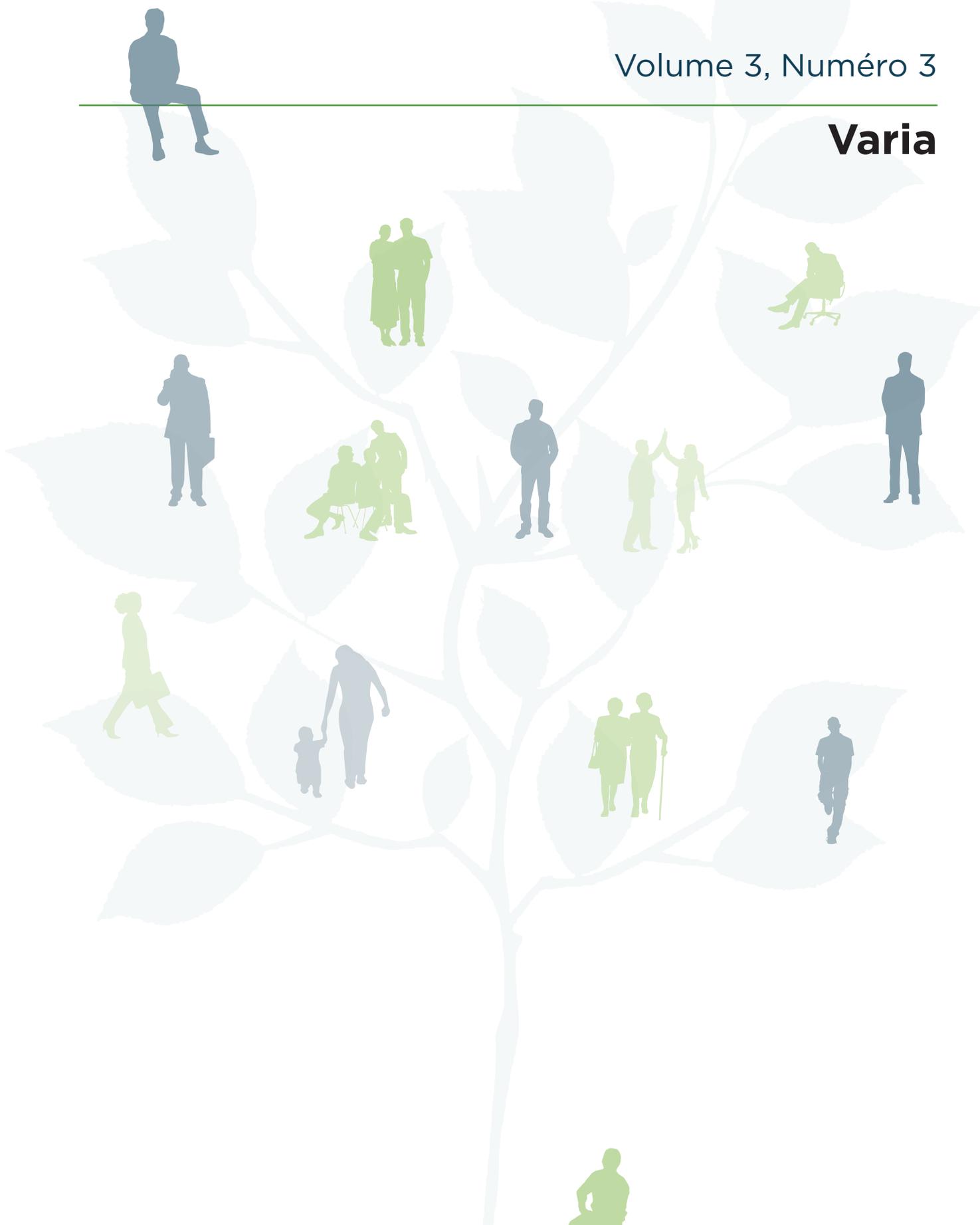


Statistique et société

décembre 2015

Volume 3, Numéro 3

Varia



Sommaire

Statistique et Société

Volume 3, Numéro 3

7 **Éditorial**

Emmanuel Didier

Rédacteur en chef de Statistique et Société

9 **Projections de populations : l'ONU adopte une méthode bayésienne**

Vianney Costemalle

Insee

15 **Quantifier la littérature, qualifier la quantification : la culturomique et le N-gramme de Google**

Sanna Alas

Étudiante diplômée

Université de Californie – Los Angeles

Vingt-sept ans plus tard : une réflexion sur la statistique publique

25 **Introduction**

Jean-Louis Bodin

Inspecteur général de l'Insee honoraire

27 **Le service public de la statistique en occident : le point actuel**

Edmond Malinvaud

Reproduction d'un article de 1988

37 **Commentaires concernant la communication de Edmond Malinvaud**

Carlo Malaguerra

Ancien directeur général de l'Office fédéral suisse de statistique

Sommaire

Statistique et Société

Volume 2, Numéro 3

Après un Café de la statistique

- 41 Compter le nombre de manifestants sur la voie publique : une problématique statistique mais aussi et surtout politique**
Pierre Muller
Inspecteur général de l'Insee en retraite
- 49 Nombre de manifestants : « Information partagée » ou « composante de la confrontation »**
Jan-Robert Suesser
Administrateur Insee honoraire, 48 années de pratique manifestante



Statistique et société

Magazine trimestriel 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.

Rédaction

Rédacteur en chef : **Emmanuel Didier**, CNRS, France

Rédacteurs en chef adjoints :

Jean-Jacques Droesbeke, Université Libre de Bruxelles, Belgique

François Husson, Agrocampus Ouest, France

Jean-François Royer, SFdS - groupe Statistique et enjeux publics, France

Jean-Christophe Thalabard, Université Paris-Descartes, pôle de recherche et d'enseignement supérieur Sorbonne Paris Cité, France

Comité éditorial

Représentants des groupes spécialisés de la SFdS :

Ahmadou Alioum, groupe Biopharmacie et santé

Christophe Biernacki, groupe Data mining et apprentissage

Alain Godinot, groupe Statistique et enjeux publics

Delphine Grancher, groupe Environnement

Marthe-Aline Jutand, groupe Enseignement

Elisabeth Morand, groupe Enquêtes

Alberto Pasanisi, groupe Industrie

Autres membres :

Jean Pierre Beaud, Département de Science politique, UQAM, Canada

Corine Eyraud, Département de sociologie, Université d'Aix en Provence, France

Michael Greenacre, Department of Economics and Business, Pompeu Fabra
Université de Barcelone, Espagne

François Heinderyckx, Département des sciences de l'information, Université
Libre de Bruxelles, Belgique

Dirk Jacobs, Département de sociologie, Université Libre de Bruxelles, Belgique

Gaël de Peretti, INSEE, France

Theodore Porter, Département d'histoire, UCLA, États-Unis

Carla Saglietti, INSEE, France

Patrick Simon, INED, France

Design graphique
fastboil.net

ISSN 2269-0271



Emmanuel DIDIER

Rédacteur en chef de *Statistique et Société*

La présente livraison de notre journal est remarquable par sa variété.

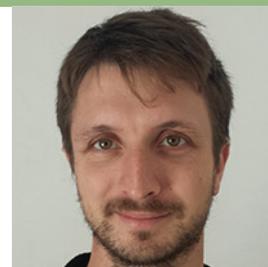
On y trouve un premier article démographique de Vianney Costemalle sur les projections de la population mondiale qui montre tous les avantages des méthodes bayésiennes. Vient ensuite un essai de Sanna Alas, étudiante à l'Université de Californie à Los Angeles, posant une question originale et pourtant essentielle : comment les études littéraires et les méthodes digitales les plus récentes se transforment-elles les unes les autres, autrement dit, comment qualités et quantités s'articulent-elles aujourd'hui.

Nous prolongeons ensuite le programme que nous avons annoncé au précédent numéro en rendant hommage à la mémoire d'Edmond Malinvaud. Avec la complicité de Jean Louis Bodin qui en a écrit l'introduction, nous republions un article de 1988 de l'ancien directeur de l'INSEE dans lequel il faisait rien de moins que le bilan de la statistique en occident ; puis Carlo Malaguerra, ancien directeur général de l'Office fédéral suisse de statistique, nous en offre un commentaire et évalue sa force de prédiction.

Pour finir, nous retournons au Café de la statistique où ont été discutées et analysées les différentes méthodes de décompte des manifestants, un puzzle que tous les amateurs de nombre connaissent bien.

Population, littérature, occident, expression politique... Assurément, la statistique se saisit de tous les sujets sociaux possibles et leur donne sa texture propre et nouvelle. Nous espérons, cher lecteur, que vous ressentirez aussi ce vent de fraîcheur à la lecture de ces textes. N'hésitez pas à nous faire part de vos réactions.

Projections de populations : l'ONU adopte une méthode bayésienne



Vianney COSTEMALLE

Insee¹

Des “projections de population probabilistes” qu’est-ce que ça veut dire ? Comment doit-on s’en servir ? Ces questions concernent désormais le grand public, et pas seulement les spécialistes, depuis que l’ONU a adopté en 2014 une méthodologie “bayésienne” pour ses projections démographiques officielles. Les méthodes statistiques bayésiennes sont encore entourées d’une aura de difficulté, voire d’une certaine suspicion chez certains. Adrian Raftery, statisticien et sociologue Irlandais ayant fait sa thèse à Paris et travaillant actuellement à l’Université de Washington, a joué un rôle majeur dans le développement théorique de ces méthodes et dans leur utilisation pratique sur de nombreux domaines d’application. Il a été la cheville ouvrière de l’élaboration des nouvelles projections de l’ONU, et il a aussi contribué à les expliquer au grand public anglophone, notamment dans la revue *Science*. Sur sa suggestion, *Statistique et Société* a demandé à un jeune statisticien français d’expliquer et de commenter ses travaux pour notre public. Nous remercions Vianney Costemalle d’avoir accepté cette tâche.

Les projections de population sont des prévisions démographiques à base statistique qui permettent d’éclairer le futur d’une région, d’un pays ou de l’humanité entière. Elles sont particulièrement utiles pour anticiper les besoins futurs, en ce qui concerne les infrastructures publiques (écoles, hôpitaux, logements, transports), l’économie (part des actifs, financement des retraites) ou encore la gestion des ressources naturelles.

Tous les deux ans, l’Organisation des Nations Unies (ONU) réalise des projections de populations sur plus de 200 pays et territoires ainsi que sur l’ensemble de la population humaine à l’horizon 2100. Elle a récemment mis en place une nouvelle méthode, reposant sur les statistiques bayésiennes, qui bouleverse notre appréciation des prévisions démographiques.

Des prévisions déterministes aux prévisions probabilistes

En 2014, la prestigieuse revue scientifique *Science* publiait un article dû à Gerland *et al.*, qui présentait les principaux résultats de la nouvelle méthode employée par l’ONU et qui déclarait peu probable la stabilisation de la population mondiale au 21^{ème} siècle contrairement à ce qu’on pensait jusqu’alors. « Peu probable » : c’est bien en termes de probabilité qu’est décrit l’avenir. On ne sait rien de lui avec certitude, mais on estime plus ou moins possibles certains scénarios.

1. Cet article n’engage que son auteur et non l’Institut national de la statistique et des études économiques

Alors que l'homme a longtemps cherché à prédire l'avenir de façon peu rationnelle, en cherchant des signes dans le hasard, des augures de Rome aux chiromanciennes de nos foires, de nouvelles techniques sont apparues dès le 17^{ème} siècle, fondées sur une représentation mathématique du monde. Les premières prévisions issues de ces modèles sont dites déterministes puisque l'état futur du système étudié est entièrement déterminé par l'état initial et les règles d'évolution de ce système ; il n'y a qu'un seul avenir possible : la ronde des planètes dans notre système solaire. Plus tard, avec le développement de la théorie des probabilités au 20^{ème} siècle, ont émergé les prévisions probabilistes où l'avenir apparaît incertain comme c'est le cas des prévisions météorologiques : on annonce par exemple qu'il va pleuvoir demain avec 70% de chance. Dans le cadre de la modélisation mathématique, cette incertitude provient fondamentalement d'une ignorance : soit l'état initial n'est pas connu parfaitement, soit cela provient d'une méconnaissance des mécanismes exacts qui sont à l'œuvre dans l'évolution du système étudié, ce qui est souvent le cas avec des systèmes complexes.

L'évolution démographique est un phénomène complexe

Prévoir la taille future de la population mondiale est en apparence simple ; cela revient à prévoir le nombre de naissances et de décès chaque année. Une méthode pour faire cela, la méthode "standard cohort-component projection" développée par Leslie en 1945, est d'appliquer des taux de fécondité et des taux de mortalité chaque année à la population étudiée. Ces taux dépendent de l'âge des personnes, des pays et aussi de l'époque. Le taux de mortalité à 80 ans n'est pas le même en 2015 qu'en 1915. Quelle sera sa valeur en 2100 ? Afin d'anticiper combien d'humains nous serons dans l'avenir, il faut prévoir comment évolueront la fécondité et l'espérance de vie. Or, à ce jour, aucune théorie, aucun mécanisme exact ne permet d'expliquer l'évolution de la fécondité ou de la mortalité. Ces évolutions dépendent de tellement de facteurs, de circonstances, d'aléas (les progrès de la médecine, l'évolution climatique, les décisions politiques, les catastrophes naturelles, les guerres, les comportements de chaque être humain,...) qu'il semble *a priori* impossible de les prévoir.

Une manière d'anticiper est d'observer comment la fécondité et l'espérance de vie ont varié dans les décennies passées. En détectant quelles sont les grandes tendances, les invariants, les motifs qui se répètent, on peut se faire une idée des évolutions futures. On a observé par exemple pour un grand nombre de pays une transition des taux de fécondité ayant toujours la même forme : au départ une fécondité élevée puis une diminution plus ou moins rapide jusqu'en-dessous du seuil de deux enfants par femme, puis dans certains cas un léger regain avec une convergence vers deux enfants par femmes. Cette « transition démographique » est achevée pour les pays développés et est encore en cours pour certains pays en voie de développement (*figure 1*). De même, on observe par exemple que l'espérance de vie augmente de façon régulière partout, et que les femmes vivent plus longtemps que les hommes en moyenne.

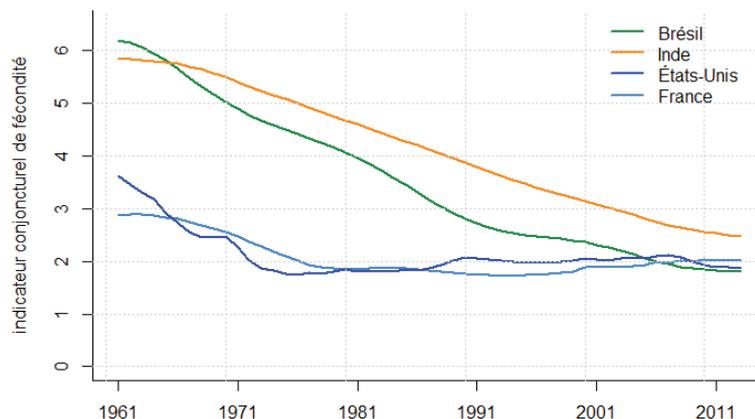


Figure 1. Évolution de l'indicateur conjoncturel de fécondité pour la France, les États-Unis, le Brésil et l'Inde. Source : Banque mondiale.

Il est possible de modéliser ces tendances, ces invariants, ces motifs, par des fonctions mathématiques en introduisant des paramètres qui, comme les rouages d'une horloge, permettent de régler la façon dont peut évoluer la fécondité et la mortalité. Certains paramètres contrôlent, par exemple, à quelle vitesse augmente l'espérance de vie chaque année.

La principale difficulté réside alors dans l'incertitude autour des valeurs de ces paramètres. Jusqu'en 2008, les projections de l'ONU faisaient appel à des savoirs d'experts pour estimer les valeurs les plus vraisemblables. À partir de ces estimations on envisageait alors un scénario médian, un scénario avec une fécondité haute et un scénario avec une fécondité basse correspondant respectivement à plus ou moins 0,5 enfant par femme par rapport au taux du scénario médian. Il s'agissait de projections déterministes : on ne sait pas avec quelle probabilité les scénarios extrêmes peuvent avoir lieu. À partir de 2014, la nouvelle méthode employée permet de réaliser des projections probabilistes à l'aide des méthodes bayésiennes.

Dans le cas de la projection pour un pays, il faut aussi prendre en compte les migrations qui peuvent modifier la taille de la population chaque année : l'émigration fait diminuer cette taille tandis que l'immigration la fait augmenter.

L'inférence bayésienne au service des projections probabilistes

Selon le paradigme bayésien, les paramètres dont on ne connaît pas la valeur exacte sont considérés comme des variables aléatoires. C'est-à-dire que pour chaque intervalle de valeur possible d'un tel paramètre on associe une probabilité. Par exemple, en simplifiant on peut imaginer qu'on ne sait pas de combien augmente exactement l'espérance de vie chaque année mais en revanche on peut savoir avec quelle probabilité cette espérance de vie augmente de 3 mois par an (plus ou moins 1 mois) ou avec quelle probabilité elle augmente de 5 mois par an (plus ou moins 1 mois). L'ensemble de ces probabilités associées à chaque intervalle de valeur possible des paramètres définit ce qu'on appelle la *distribution* de la variable aléatoire. Dans le cadre bayésien il s'agit de déterminer quelle est cette distribution pour l'ensemble des paramètres du modèle. Pour cela on utilise le théorème de Bayes (ecclésiastique - mathématicien britannique du 18^{ème} siècle) qui permet d'actualiser cette distribution à l'aide des données observées sur le passé : avant toute observation on a une distribution *a priori* qui reflète les suppositions minimales de la valeur des paramètres sur lesquelles les experts peuvent s'accorder, et une fois les observations faites et les données recueillies on en déduit une distribution *a posteriori* qui prend en compte toute l'information apportée par les données. L'inférence bayésienne permet de remonter des observations à leurs causes, en attribuant une probabilité à chacune d'elles. Plus le nombre de données observées est grand, plus la distribution *a posteriori* sera resserrée autour d'une seule valeur. Autrement dit, plus on a de données, plus l'incertitude sur la valeur des paramètres est faible.

Les distributions des paramètres contrôlant l'évolution de la fécondité et de la mortalité sont ainsi estimées pour chaque pays en prenant en compte les données de tous les pays. Ces distributions ne sont en effet pas indépendantes entre elles, mais sont reliées dans la modélisation, par ce qui s'appelle un modèle bayésien hiérarchique.

À partir de la modélisation mathématique et de la distribution *a posteriori* des paramètres on peut prévoir les différentes probabilités des évolutions futures de la taille de la population (*figure 2*). Pour une évolution possible, on tire au hasard un jeu de paramètres selon la distribution calculée. Partant de ce jeu de paramètres on fait évoluer jusqu'à l'horizon souhaité la fécondité et l'espérance de vie, de façon mécanique selon les règles retenues lors de la phase de modélisation. On en déduit alors les taux de fécondité et de mortalité par âge pour chaque période future, puis l'évolution de la taille de la population. On obtient ainsi un chemin possible d'évolution de la population mondiale. Si l'on répète ces opérations une deuxième fois, on aura

alors un autre chemin possible (car on aura tiré au départ un autre jeu de paramètres). La répétition de ces opérations un très grand nombre de fois donne toute une palette de chemins possibles. C'est ensuite à partir de l'ensemble de ces chemins qu'on détermine des intervalles de confiance : l'intervalle de confiance à 95% est un intervalle dans lequel 95% des chemins se situent. L'intervalle de confiance à 80% est un intervalle dans lequel huit chemins sur dix se situent. L'intervalle de confiance à 80% est donc lui-même à l'intérieur de l'intervalle de confiance à 95%. En fait, plus le niveau de confiance est élevé, plus l'intervalle, et donc l'incertitude sur le chemin futur, est grand.

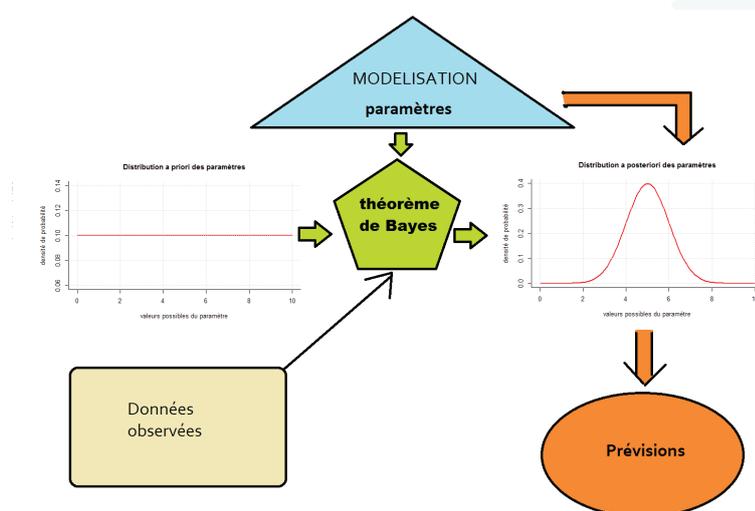


Figure 2. Schéma illustrant les prévisions probabilistes par une approche bayésienne

Très probablement entre 9,5 et 13,3 milliards d'êtres humains en 2100

La *figure 3* présente les résultats de la projection probabiliste pour la taille de la population mondiale, ainsi que les résultats de la projection déterministe classiquement effectuée par l'ONU. L'intervalle de confiance à 95% est beaucoup plus étroit que l'intervalle entre le scénario « haut » et le scénario « bas » de la projection déterministe. Le scénario médian se trouve quant à lui presque au milieu des intervalles de confiance à 95% et 80%. Selon ces projections probabilistes, le monde comptera entre 10 milliards et 12,5 milliards d'habitants en 2100 avec une probabilité de 80% et entre 9,5 milliards et 13,3 milliards avec une probabilité de 95%. Les intervalles de confiances obtenus avec la méthode probabiliste peuvent parfois être plus grands que l'intervalle entre le scénario haut et bas des projections déterministes, ce qui est le cas pour beaucoup de pays d'Afrique. Cela reflète que l'incertitude sur la fécondité dans ces pays est plus grande que plus ou moins 0.5 enfant par femme. Il y a en particulier une très grande incertitude sur l'évolution de la population au Nigéria qui, selon ces estimations, a 80% de chance d'atteindre un niveau entre 440 millions et 1,2 milliard d'habitants d'ici la fin du siècle, soit un écart de près de 800 millions de personnes ! Par ailleurs, le milieu de cet intervalle (820 millions) est bien supérieur au scénario médian de la projection déterministe pour ce pays (750 millions). Cela provient de l'incertitude sur la fécondité qui ne diminue pas aussi vite que cela avait été observé dans les pays d'Asie ou d'Amérique Latine il y a quelques années. Il reste néanmoins que la population du Nigéria va très probablement augmenter très fortement et contribuer pour beaucoup à la forte augmentation de la population de l'Afrique. L'ONU prévoit de plus que le rapport du nombre de personnes de 20 à 64 ans sur le nombre de personnes de 65 ans et plus va chuter dans beaucoup de pays : selon ces prévisions il va passer en Chine de 7,8 aujourd'hui à 1,8 en 2100 et de 8,6 à 1,5 pour le Brésil, atteignant ainsi des niveaux bien inférieurs au niveau actuel du Japon (2,6).

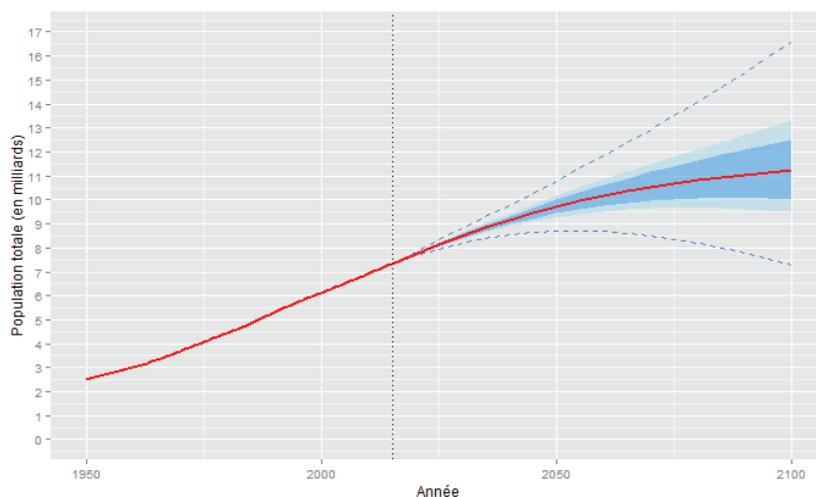


Figure 3. Évolution passée et future de la taille de la population mondiale. En rouge le scénario médian obtenu par projections déterministes et en pointillés les scénarios « haut » et « bas » correspondant à une fécondité de plus ou moins 0.5 enfant par femme. La zone en bleu très clair indique l'intervalle de confiance à 95% et celle en bleu plus foncé indique l'intervalle de confiance à 80% obtenu par projections probabilistes.

Source : Nations Unies, Département d'économie et des affaires sociales, Division de la population.

Le cas de la France : les prévisions de l'ONU et celles de l'Insee.

Les dernières projections démographiques réalisées en 2010 par l'Institut national de la statistique et des études économiques (Insee) vont jusqu'en 2060. Ce sont des projections déterministes déclinées selon plusieurs scénarios de fécondité, mortalité et soldes migratoires. Le scénario central retenu donne une population de 73,6 millions habitants en France métropolitaine au 1er janvier 2060. Selon le scénario de fécondité haute, la population pourrait atteindre 77,6 millions d'habitants d'ici 2060 et selon le scénario de fécondité basse, 69,9 millions (la différence de fécondité entre ces deux scénarios est de 0.3 enfant par femme). Les prévisions probabilistes de 2014 de l'ONU indiquent que la population de France métropolitaine se situera entre 68,4 et 76,1 millions d'habitants en 2060 avec une probabilité de 80% et entre 66,5 et 78,5 millions avec une probabilité de 95%.

Les progrès informatiques à la source des avancements méthodologiques

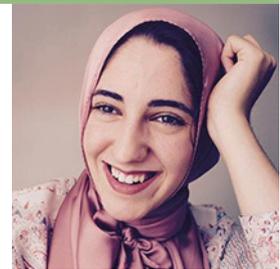
Les prévisions probabilistes fondées sur l'inférence bayésienne ont été rendues possibles grâce à l'essor exponentiel de la puissance de calcul informatique. Il faut en effet explorer l'ensemble du champ des possibles un très grand nombre de fois pour déterminer les probabilités d'occurrences des événements. Les prévisions probabilistes permettent de mieux appréhender l'incertitude inhérente aux phénomènes complexes pour lesquels d'innombrables facteurs entrent en jeu. Quantifier la probabilité de réalisation des différents avenir possibles apporte plus d'information sur nos connaissances réelles qu'une unique prévision. Dans le cas des projections de population de l'ONU, cela a permis de réduire considérablement l'incertitude. De plus, une des forces des méthodes bayésiennes est que l'ensemble des données est utilisé pour estimer un paramètre précis. Par exemple, l'évolution future de l'espérance de vie en

France est estimée à l'aide des données françaises mais aussi à l'aide des données de l'ensemble des autres pays. Enfin, avec l'inférence bayésienne, on peut intégrer des savoirs extérieurs à la modélisation, par le biais des distributions a priori. Ces distributions reflètent notre connaissance (ou méconnaissance) des paramètres avant toute observation. Le statisticien a en effet le choix de la distribution a priori qu'il va utiliser. Il peut alors prendre en compte des savoirs divers qui permettent d'avoir une idée de cette distribution. C'est souvent cette liberté qui est la source la plus vive des critiques émises à l'encontre des méthodes bayésiennes car, pour certains, liberté est alors synonyme d'arbitraire.

Références

- Alkema, L., Raftery, A.E., Gerland, P., Clark, S.J., Pelletier, F., Buettner, T. and Helig, G. (2011). *Probabilistic Projections of the Total Fertility Rate for All Countries*. *Demography* vol. 48, p.815-839.
- Attali J., « Peut-on prévoir l'avenir ? », Fayard, 2015.
- Blanpain N. et Chardon O., « Projections de population 2007-2060 pour la France métropolitaine : méthode et principaux résultats », *Document de travail*, Direction des Statistiques Démographiques et Sociales, Insee, 2010.
- Efron B., « Bayes' Theorem in the Twenty-First Century », *Science*, 2013.
- Gerland, P., Raftery, A.E. ; Ševčíková , H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B.K., Chunn, J.L., Lalic, N., Bay, G., Buettner, T., Heilig, G.K. and Wilmoth, J. (2014). World Population Stabilization Unlikely This Century., *Science*, 2014, n°346, p 234-237.
- Leslie P.H., « On the use of matrices in certain population mathematics », *Biometrika*, 1945, Vol. 33, p 183-212.
- Raftery A., « Use and communication of probabilistic forecasts », arXiv 1408.4812v1, 2014.
- Raftery A., Alkema L. et Gerland P., « Bayesian Population Projections for the United Nations », *Statistical Science*, 2014, Vol. 29, n°1, p 58-68.
- Raftery, A.E., Chunn, J.L., Gerland, P. and Ševčíková , H. (2013). *Bayesian Probabilistic Projections of Life Expectancy for All Countries*. *Demography*, vol. 50, p.777-801.
- Raftery A., Lalic N., Gerland P., « Joint probabilistic projection of female and male life expectancy », *Demographic Research*, 2014, Vol. 30, p 795-822.
- Robert C., *Le Choix Bayésien*, 2006, Springer.

Quantifier la littérature, qualifier la quantification : la culturomique et le N gramme de Google



Sanna ALAS

Étudiante diplômée – Université de Californie – Los Angeles

Introduction

Des rouleaux de papyrus poussiéreux des tombes de l'ancienne Égypte, jusqu'aux romans à quatre sous des rayonnages de librairies d'aujourd'hui, l'écriture a été une partie intégrante de l'expression culturelle humaine. La presse à imprimer à caractères mobiles de Johannes Gutenberg a largement accéléré le processus, et depuis, les livres sont publiés et diffusés à des rythmes étonnants. En 2010, on estimait à 129 864 880 le nombre de livres ayant été publiés, ce qui serait équivalent à 4 milliards de pages et 3000 milliards de mots (Taycher 2010).

Dénombrer ces livres ne représente pourtant qu'une partie de la mission plus ample que s'est fixée Google, laquelle consiste à digitaliser tous les livres du monde, une entreprise ambitieuse qui a déjà produit une bibliothèque digitale toujours en croissance estimée aujourd'hui à 30 millions de livres (Darnton 2013) dans 480 langues différentes (Jackson, 2010). Selon les mots d'Eric Schmidt, qui était alors le président-directeur général de Google, ce projet vise à favoriser un « égalitarisme de la dispersion de l'information » grâce auquel chaque individu pourra accéder à des textes qui auparavant pouvaient rester inaccessibles et les compulsurer (Schmidt., 2005). Google n'est pas seul dans cette entreprise ; son effort pour digitaliser a été soutenu par d'innombrables bibliothèques, des groupes de publication et des institutions académiques (y compris dans ma propre Université de Californie) partout dans le monde. Des millions de textes qui avaient été oubliés sont ajoutés à cette nouvelle bibliothèque digitale.

Lire à des distances différentes

Pour ceux qui étudient les choses de l'écrit, qu'ils soient étudiants en littérature, économistes, historiens ou linguistes, ce corpus a le potentiel de transformer leur façon de conduire leur recherche, ainsi que de soulever de nouvelles questions et de nouvelles méthodes d'analyse. Traditionnellement, l'analyse des textes a été dominée par une approche qualitative, procédant à une lecture rapprochée dans lequel un petit nombre de textes clefs sont analysés en grand détail. L'approche utilisée y est généralement inductive et exploratoire ; l'intention de l'auteur, le contexte historique, et les connotations de mots y jouent un grand rôle. Cette approche limite nécessairement la quantité de textes qui peuvent être analysés, et il en résulte un corpus réduit d'ouvrages canoniques sur lesquels la recherche se focalise. Cette approche qualitative est souvent considérée comme diamétralement opposée à une approche quantitative, dont le principe réside dans la déduction, le test d'hypothèses et l'expérimentation. Cette dernière méthode est généralement présentée comme scientifique, empirique et objective à cause de son usage de l'analyse numérique (Leeds-Hurwitz 1995).

Pourtant, avec la montée des données massives, beaucoup de gens testent des méthodes d'analyse quantitative dans des domaines qui ont été traditionnellement dominés par les méthodes qualitatives. Les chercheurs en littérature peuvent maintenant fouiller dans les milliers de livres qui pouvaient avoir été oubliés par ailleurs et poser des questions nouvelles au sujet de la culture ou du milieu littéraire d'une période. Ceci a engendré des tensions entre les tenants de la recherche qualitative et de la recherche quantitative. C'est pourquoi, étant donné les collections massives d'information littéraire qui émergent, on peut se demander quel est aujourd'hui le rôle de chacune de ces méthodes dans l'analyse des données ?

Selon la perspective de certains, comme Franco Moretti, les analyses quantitatives sont supérieures et devraient supplanter les analyses qualitatives. Le travail de ce dernier dans son laboratoire littéraire de Stanford en fait le champion de la lecture distante, dans laquelle les textes d'une période entière sont agrégés et analysés comme un tout, transcendant le point de vue jugé limité de la lecture rapprochée (Moretti, 2013). Parmi les études conduites par ce laboratoire on peut citer la différenciation quantitative des genres littéraires (Allison, 2011), la cartographie réticulaire des intrigues et des caractères (Moretti, 2011), ou encore les changements sémantiques dans le roman britannique (Heuser, 2012), toutes conduites au moyen de techniques de critique computationnelle. J'évoque ici Moretti non pour critiquer sa méthodologie, mais plutôt pour analyser sa rhétorique. Il présente un cas plutôt extrême dans sa façon de créer une fausse opposition entre recherches qualitatives et recherches quantitatives, alors qu'en fait ces méthodes s'appuient l'une sur l'autre. Ce cas révèle peut-être le sentiment inconscient que la recherche « scientifique » ou quantitative a plus de valeur, est plus précise et plus objective. Pour prendre un exemple moins extrême mais aussi significatif, le premier grand article publié à propos d'une recherche conduite en utilisant le corpus littéraire massif de Google Books a été publié non pas dans une revue consacrée aux humanités mais dans le journal *Science* (Michel, 2011). Ceci montre que la rhétorique scientifique a été, de façon générale, associée aux valeurs de la « précision » et de « l'objectivité ».

Dans le présent texte, j'espère déstabiliser le discours binaire qui entoure la recherche quantitative et la recherche qualitative en littérature et démontrer que ces deux méthodes ne sont pas opposées de façon inhérente. Jusque là j'ai cherché à fournir un arrière-plan de l'émergence d'une bibliothèque digitale, et des discussions critiques à son propos. Dans le reste de cette étude, je me focaliserai sur un corpus, le N-gramme de Google Books, et j'analyserai la manière dont il a été utilisé comme un supplément ou un remplacement de différentes méthodes de recherche à travers diverses disciplines.¹

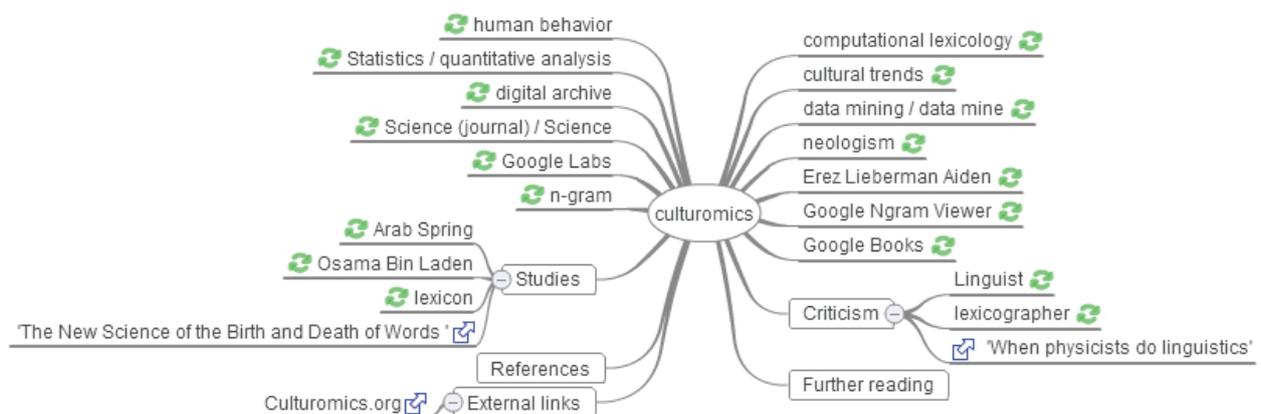


Figure 1. Carte Wikipedia représentant la « culturomique ». On y voit les différentes disciplines et champs de recherche qu'elle a influencés et ceux qui à leur tour l'ont influencée.

1. On le trouvera à cette adresse : <https://books.google.com/ngrams>

Le visualisateur N-gramme de Google Books fait partie d'un projet plus important appelé « culturomique » (figure 1). Fondée par Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden et d'autres de l'Observatoire culturel de Harvard, la culturomique cherche « à digitaliser et à analyser des données à propos de la culture sur de très grandes échelles : tous les livres, tous les journaux, tous les manuscrits, etc. » (FAQ Culturomics, 2015). Le mot a émergé comme un analogue de termes tels que « génomique », « protéomique » (à ne pas confondre avec la technique en microbiologie qui porte le même nom) et pointe vers la tendance actuelle pour le travail interdisciplinaire utilisant les nouvelles technologies et les méthodes d'interprétation quantitatives / computationnelles. Ce projet, qui tombe sous le parapluie des « humanités digitales », se présente lui-même comme un ajout aux formes traditionnelles de la recherche qualitative plutôt que comme un remplacement. Ce point de vue est incroyablement important parce qu'il subvertit la rhétorique de la supériorité pseudo scientifique et crée un aller-retour entre les méthodes quantitatives et qualitatives qui sont souvent considérées comme disparates. En analysant les types d'enquêtes conduites en utilisant le visualisateur N-gramme de Google Books, j'espère montrer comment l'émergence de la littérature comme une forme de données massives transforme les façons selon lesquelles les chercheurs analysent et comprennent des textes.

Le N-gramme de Google Books : une brève discussion de la « datafication »

Avant de présenter quelques applications de ce corpus, il est essentiel d'expliquer que les pages sont d'abord scannées puis ensuite transformées en lettres et en mots « discrets » appartenant à un texte en ligne. Le premier processus, le scan, s'appelle la *digitalisation* et doit être distingué du second processus, la *datafication*. La digitalisation se réfère à la conversion d'une information analogique en un code binaire utilisé par les ordinateurs, alors que la « datafication » se réfère à la quantification des informations binaires pour l'analyse et la tabulation (Mayer-Schonberger, 2013, p. 78). Si des livres ont été digitalisés depuis des années, ce qui a changé récemment est que chaque caractère est maintenant « datafié » à travers le processus de la reconnaissance optique des caractères (ROC). Ce processus utilise une série d'algorithmes qui permet de convertir une photographie du texte en un texte digital dans lequel on peut faire une recherche. Ceci permet aux chercheurs de parcourir des pages d'information et de différencier par période de temps, genre, langage, année de publication, etc.

Le N-gramme de Google Books dépend entièrement de ce processus. Il identifie la fréquence relative de différents mots figurant dans 22 différents corpus divisés par langue (anglais, espagnol, chinois, russe, allemand, hébreu, français et italien), par genre (anglais vs fiction anglaise), ou par dialecte (anglais américain et anglais britannique) (Karch, 2015). Le N-gramme lui-même est un outil statistique utilisé en probabilités et en linguistique computationnelle pour trouver le nombre de fois, N, où un mot spécifié ou une phrase apparaît dans un corps de textes plus étendu. Le N-gramme peut non seulement compter le nombre d'apparitions d'une phrase mais il peut aussi fournir de l'information au sujet de la vraisemblance de son apparition dans le futur. À titre d'exemple, la figure 2 ci-dessous est un N-gramme (insensible à la casse) des mots « qualitative » et « quantitative » utilisant le corpus anglais pour la période 1800 à 2008.

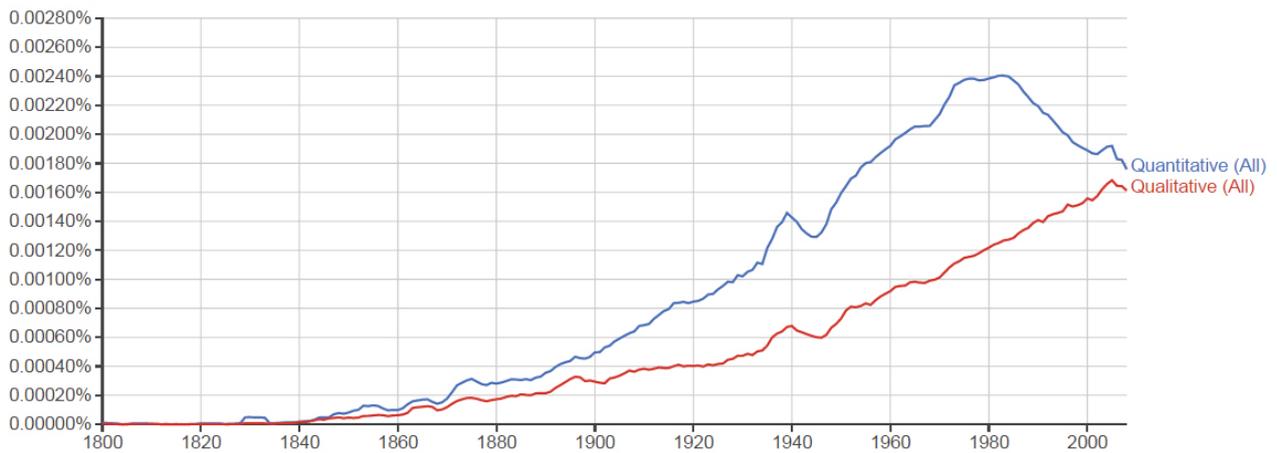


Figure 2. Ce N-gramme retrace les variations de l'usage des termes « quantitative » et « qualitative » dans le corpus anglais de Google Books. Les lois sur le copyright nous empêchent d'analyser les données brutes directement ; aussi, les données de fréquences fournissent une alternative pour chercher à travers la totalité de Google Books tout en protégeant les droits des éditeurs.

Bien que la collection de données soit gigantesque, il reste diverses limitations au visualisateur N-gramme de Google Books. L'une de ces limitations réside dans le processus ROC utilisé pour reconnaître les caractères, mis en défaut par le fait que les différents textes utilisent différentes polices et sont de qualité variable. De même, des erreurs dans les métadonnées d'un texte peuvent le dater faussement, et il en résulte des pics temporels factices qui ne représentent pas réellement la littérature d'une époque (FAQ Culturomics, 2015). De plus, seulement certaines langues sont représentées dans le corpus, et les textes de langues moins étudiées restent invisibles pour le N-gramme. Pour une recherche qui se propose de retracer les changements culturels à travers le temps, ces omissions risquent de réifier des interprétations hégémoniques de l'histoire, de la littérature et du langage, en continuant à ignorer des sources déjà marginalisées par les universitaires. Ceci étant dit, au fur et à mesure que la collection de livres de Google grandit et que les algorithmes qui dictent les processus de ce système sont développés, les ingénieurs du N-gramme espèrent le rendre de plus en plus précis.

Littérature et données : une étude de cas en « culturomique »

L'Observatoire culturel de Harvard a été le premier groupe à faire un usage original du visualisateur N-gramme de Google Books. Il a produit plusieurs études fascinantes qui ont montré tout l'intérêt des puissants outils de la culturomique. Beaucoup de questions qu'ils ont traitées l'auraient été traditionnellement par les chercheurs des humanités – de l'évolution du langage à l'incidence de la renommée.

Plutôt que de présenter une vue d'ensemble de leurs études, je vais me focaliser sur une seule question, d'ailleurs exprimée dans le sous-titre d'un article : comment « détecter la censure et la suppression » (Michel, 2011, p. 181) ? Dans cette étude, la suppression et la censure sont quantifiées en comparant la fréquence d'apparition du nom de l'artiste juif Marc Chagall dans le corpus littéraire anglais et dans le corpus littéraire allemand pendant le régime nazi. Une recherche par N-gramme sur son nom a montré une décroissance drastique du nombre de fois où il a été mentionné entre 1936 et 1944 dans le corpus Allemand, une période qui est corrélée directement à la censure et à la suppression de l'Allemagne nazie.

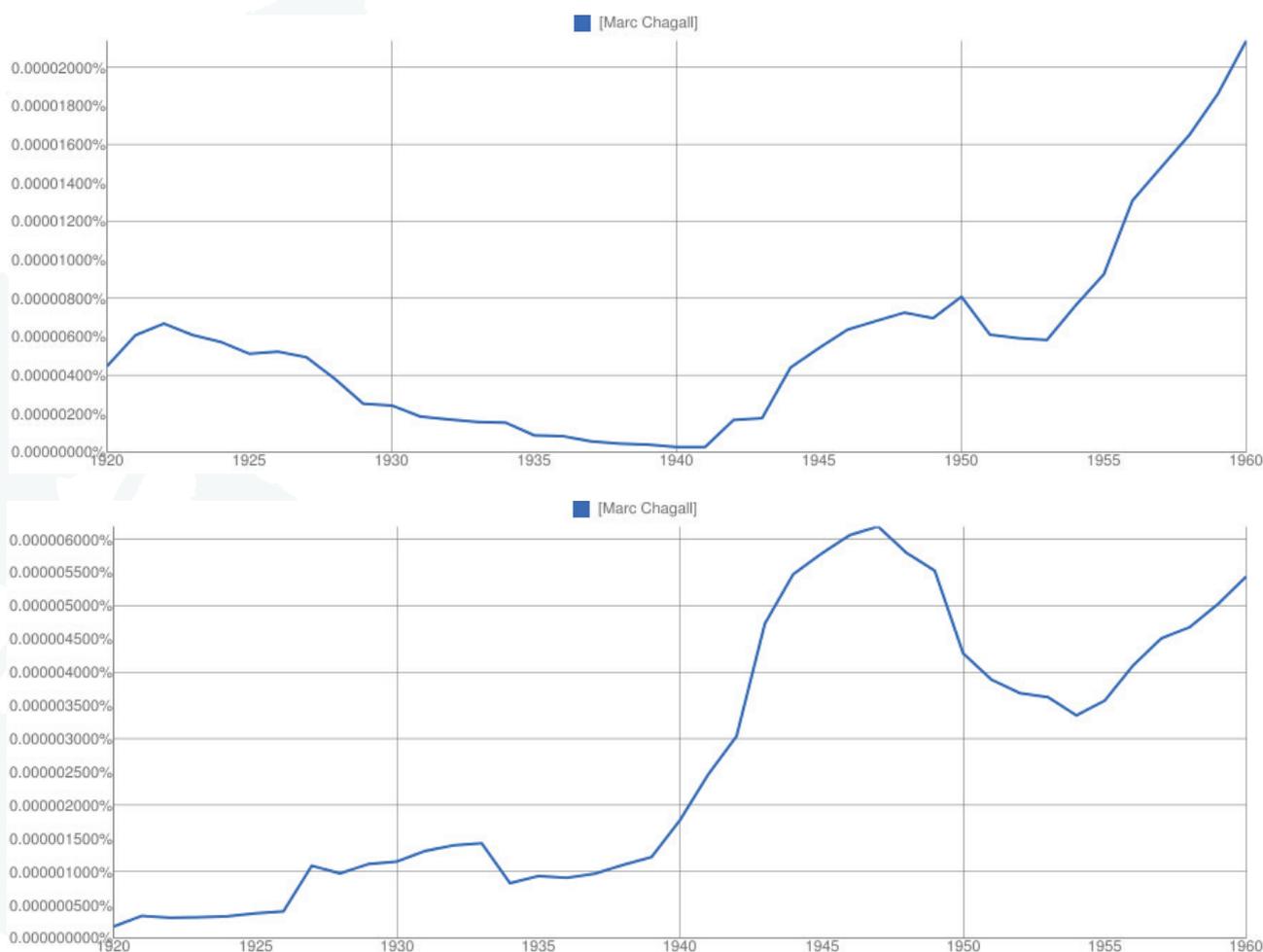


Figure 3. N-grammes de la requête « Marc Chagall » dans le corpus allemand (en haut) et dans le corpus anglais (en bas) entre les années 1920 et 1960

Cette étude est significative parce qu'elle montre que la censure subie par un individu est liée à un contexte socio-politique plus large. Cependant on doit prendre garde aux confusions, car le principe « corrélation n'est pas causalité » tient aussi lorsque les données sont littéraires. Les chercheurs de Harvard ont conduit une étude similaire sur la censure durant cette même période où les travaux d'auteurs divers et d'artistes étaient bannis à cause de leur nature « indésirable ». En analysant la fréquence d'apparition des écrivains dans différentes catégories comme l'histoire, la littérature, la politique, la philosophie et les arts, ils ont pu déterminer une décroissance marquée de la production en philosophie, en politique, et en art (p. 181). Cette étude, à son tour, a permis de créer un « indice de suppression », grâce auquel les victimes de la censure peuvent être identifiées de novo.

Dans le cas de cette étude particulière, les notions de suppression et de censure qui ont été longtemps analysées comme des idées abstraites peuvent maintenant être quantifiées et mesurées. Imaginez les applications à long terme de telles métriques : il serait possible d'identifier les victimes de la censure ou de mesurer quantitativement le totalitarisme. Cependant, de telles applications soulèvent aussi beaucoup de questions : pourquoi ces indices seraient-ils plus valides que les formes traditionnelles de preuves telles que les comptes-rendus d'archives, les comptes-rendus historiques ou les témoignages individuels ? L'utilisation de tels vastes ensembles de données fait implicitement l'hypothèse que les conclusions sont fondamentalement les moins faillibles et les plus concrètes, et par conséquent les plus convaincantes. Cependant, une telle supposition néglige les biais inhérents aux méthodes de recherche quantitative, dans lesquels des facteurs de confusion peuvent mener à des

conclusions fallacieuses. Au final, disons que les relations entre les différentes variables et les diverses corrélations identifiées ont un grand potentiel pour aider les chercheurs en sciences humaines et sociales à poser des questions différentes. Dans ce sens, il y a une rétroaction entre la recherche qualitative qui fournit un contexte et un cadre théorique et la recherche quantitative qui y applique ses méthodes.

La littérature comme données : une étude de cas sur l'indice de misère littéraire

En utilisant le N-gramme de Google Books, des chercheurs ont montré que la littérature et les livres représentent en moyenne la situation économique de la décade précédente. Bien que ce point puisse sembler intuitif, il s'agit d'un important exemple de la manière selon laquelle des méthodes empiriques quantitatives peuvent être utilisées pour explorer des questions qui étaient traditionnellement dévolues aux méthodes qualitatives. Les chercheurs ont développé un « indice de misère littéraire » (IML) en utilisant l'analyse des données pour détecter les mots associés avec un état d'esprit particulier (Bentley 2014).

Trois méthodes ont été utilisées. La première était l'outil d'analyse textuel «Affect Net du Mot» (ANM), un outil digital qui génère des listes de mots synonymes et les associe avec des humeurs. Cette méthodologie a été modélisée à partir des six émotions de base (colère, dégoût, peur, joie, tristesse, surprise) identifiées par Strapparava *et al.* dans leur article de 2008 sur l'identification des émotions dans des textes.

La deuxième méthode utilisait une collection préexistante de mots présents dans l'outil «Recherche linguistique et comptage de mots » (RLCM) qui a été développé par le linguiste computationnel James W. Pennebaker et d'autres pour calculer le degré selon lequel un texte utilise des émotions positives ou négatives, des références à soi-même et d'autres dimensions du langage (Pennebaker, 2007).

La troisième et dernière méthode a été l'approche hédonométrique créée par Peter Dodds *et al.* dans leur recherche de 2011 évaluant le "contenu en bonheur" des comptes twitter. Cette étude a été conduite en utilisant un corpus de plus de 10 000 mots évalués du point de vue de l'émotion en utilisant l'outil "intelligence artificielle artificielle" du "Mechanical Turk" d'Amazon. Ces trois méthodes se situent donc dans une zone grise entre le qualitatif et le quantitatif, au sens où les humeurs que chaque méthode associait aux mots isolés l'ont été de façon subjective par des personnes, qu'il s'agisse de travailleurs anonymes sur le "Mechanical Turk" d'Amazon ou d'un ensemble de chercheurs. Cependant, tout comme dans le cas des conclusions tirées du visualisateur N-gramme de Google, l'accroissement de la taille de l'échantillon rend les résultats plus objectifs.

Pour la méthode ANM, l'indice de misère littéraire a été finalement calculé comme la différence entre les fréquences des mots caractérisés comme "de joie" et des mots caractérisés comme "de tristesse". Pour la méthode RLCM, l'indice a été calculé comme la différence entre les "mots de tristesse" et les "mots d'émotions positives". Enfin, pour la méthode hédonométrique, les mots associés avec le bonheur ont été corrélés à l'inverse de l'indice de misère économique, une hypothèse qui aurait besoin d'être validée. Les mots ainsi identifiés ont alors été entrés dans le visualisateur N-gramme de Google Books pour explorer le taux de croissance ou de décroissance de leur fréquence au cours du temps. Ce taux a ensuite été mis en graphique par rapport à un indice économique établi par ailleurs officiellement. Cet indice est une mesure du taux de chômage et du taux d'inflation. Il est communément utilisé comme un indicateur de la performance économique courante et future d'un pays.

Une corrélation positive a été trouvée, non seulement dans le corpus littéraire des États-Unis

mais aussi dans celui de l'Allemagne et de la Grande-Bretagne. Cette corrélation entre l'indice de misère littéraire et l'indice de misère économique est meilleure que pour n'importe quel autre indicateur d'humeur, reflétant la notion que l'économie a un effet profond sur l'état d'esprit et la production littéraire.

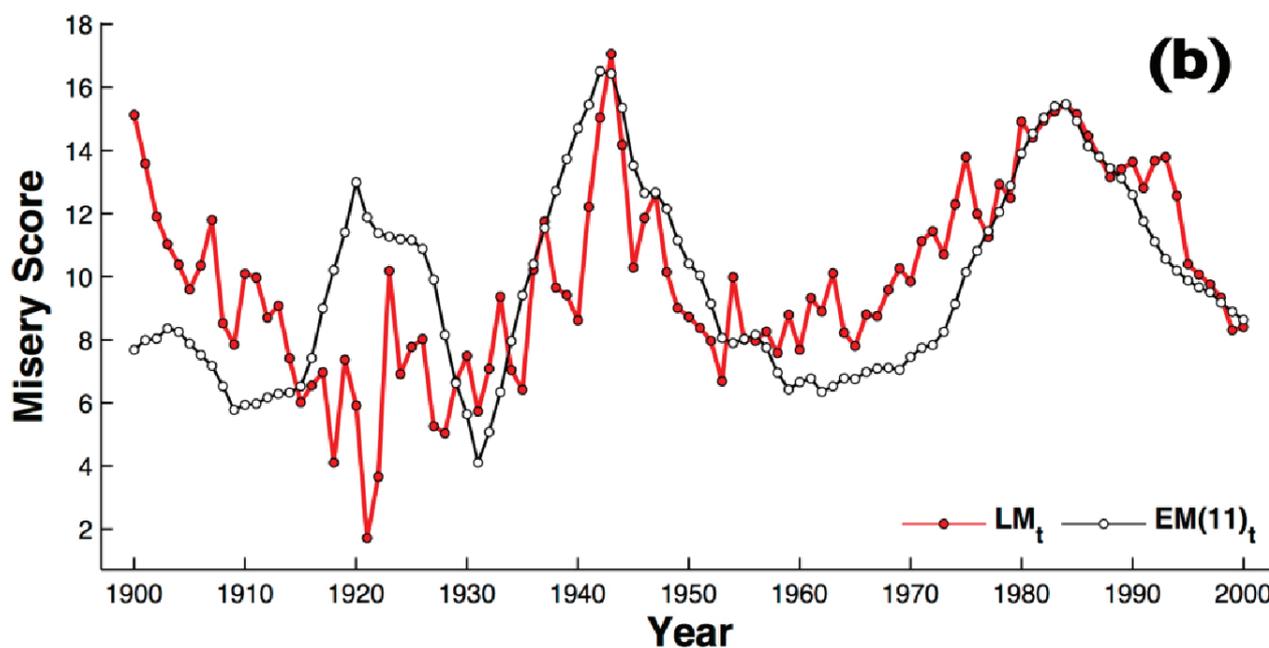


Figure 4. Comparaison entre, d'une part, l'indice de misère littéraire produit selon la méthode ANM (en rouge) pour tous les livres et, d'autre part, l'indice de misère économique aux États-Unis (en blanc). Les auteurs ont utilisé une moyenne mobile de l'indice économique sur 11 ans en arrière pour renforcer la corrélation entre les deux indices (Bentley, 2014).

Cet effet cependant peut être vu plus fortement dans le corpus onze ans après l'indice de misère économique. Cette période de 11 ans reflète la moyenne mobile qui a montré la plus grande corrélation avec l'indice de misère économique, plus qu'un simple retard ou que d'autres valeurs de la moyenne mobile. Ce calcul permet effectivement aux auteurs d'apercevoir des tendances plus importantes dans les données et de lisser les cycles de court terme qui semblent caractéristiques d'un corpus littéraire. Le graphique ci dessus présente des pics pendant des moments importants de l'histoire : la fin de la première guerre mondiale en 1918, les suites de la Grande Dépression vers 1938, et la crise de l'énergie en 1975. En fait, on a pu montrer que l'indice de misère littéraire avec la moyenne mobile sur 11 ans en arrière est plus robustement corrélé à l'indice de misère économique que les mesures économiques de l'inflation et du chômage qui sont les composantes de l'indice de misère économique lui-même.

C'est pourquoi cette étude a des conséquences fascinantes. Elle connecte deux formes disparates d'information, la production littéraire et la performance économique, d'une manière statistiquement mesurable. Elle montre tout le bénéfice qu'il peut y avoir à utiliser des sources de données non traditionnelles telles que la production littéraire et culturelle pour acquérir des informations "objectives" et des mesures quantitatives telles que des mesures de la performance économique. Elle change la conception de ce qui peut servir comme source de données dans la recherche quantitative ; les données massives peuvent être qualitatives si le corpus d'information est traité convenablement.

Ceci me conduit à retourner à notre point de discussion initial et à la rétroaction entre des méthodes de recherche quantitatives et qualitatives. L'article de Bentley est publié dans un

format qui est typique des articles scientifiques. Il présente une introduction qui fournit un survol du sujet en discussion et ensuite dessine une hypothèse de la manière suivante :

Nous supposons que les auteurs de livres, qui sont à la fois des producteurs et des consommateurs de connaissances communes, sont informés par les conditions économiques du passé selon une certaine échelle de temps que l'on peut définir ; ils transmettent la connaissance commune non seulement factuellement mais aussi indirectement, à travers un contenu émotionnel. Par conséquent nous pouvons nous attendre à trouver une corrélation positive entre l'état d'esprit tel qu'il est exprimé dans les livres et les conditions économiques du passé récent. Vu que les livres prennent du temps à être écrits et publiés, nous nous attendons à ce que chaque année de publications fasse la moyenne des influences économiques du passé sur un nombre défini d'années (Bentley, 2014, p.1).

Ce passage combine la méthodologie de la recherche quantitative, telle que l'indique l'usage des mots « nous supposons », « une échelle de temps que l'on peut définir », « corrélation positive » avec le savoir fondamental de la recherche qualitative dans ses descriptions de la position duale des auteurs comme producteurs et consommateurs de connaissances collective. Dans ce sens, l'information qualitative contextualise les méthodes quantitatives qui sont utilisées pour l'étudier.

Cette rétroaction est aussi observée dans la section Discussion de l'article, dans laquelle l'auteur cherche à aller au-delà de la corrélation et à rechercher la causalité. Il suggère plusieurs théories explicatives : l'une est celle de la construction de niches culturelles théoriques et l'assertion selon laquelle l'économie globale a été intégrée dans l'expérience émotionnelle de la société au 20e siècle (p.4) ; une seconde théorie repose sur « l'effet décade » selon lequel les souvenirs sont accumulés dans l'enfance et les individus se mettent à écrire à peu près 10 ans plus tard au sujet de leurs expériences (p.3). La première théorie est issue de la discussion des travaux en psychologie comportementale et évolutionniste de Steve Pinker sur les niches cognitives et sur les manières dont les humains ont évolué pour acquérir ces traits (Pinker, 2010). Cette théorie sert de pont entre la production culturelle et littéraire et le processus de l'évolution, déplaçant les discussions depuis un focus qualitatif sur le contenu (à partir des mots RLCM) vers un accent quantitatif sur la corrélation. De façon similaire, la seconde théorie formule des thèses sur le développement des enfants au moyen d'une étude longitudinale de 167 Californiens qui vivaient pendant la Grande Dépression (Elder, 1974) et cherche à prouver empiriquement à travers le processus de déduction ce que les méthodes quantitatives supposent en utilisant le processus d'induction : que les événements au cours de la vie influencent la mémoire dans le futur. Ce qui manque dans cette explication cependant, ce sont des preuves de l'argument implicite selon lequel tous les individus perçoivent les émotions de la même façon. Une autre faiblesse est la non-prise en compte du fait que, lorsqu'on mène des études rétroactives, il y a un grand potentiel de « biais de sagesse rétroactive » influençant la manière dont les données ont été perçues et présentées.

Cet usage du visualisateur N-gramme de Google Books montre que les méthodes qualitatives et quantitatives peuvent toutes les deux être utilisées simultanément dans un cadre de culturomique. On y trouve une rétroaction entre les méthodes quantitatives et qualitatives. Cependant, étant donné la nature scientifique de ce papier, les analyses quantitatives ont été privilégiées comme sources de preuves alors que les analyses qualitatives n'ont été utilisées que pour contextualiser l'argument. Ceci peut être un symptôme du rôle qu'y joue l'économie : comme la littérature est utilisée pour prédire (quoique rétroactivement) un indice économique, utiliser une preuve quantitative serait le choix logique. La discipline influence en définitive le type de questions que l'on pose et les méthodes que l'on choisit. Ici le processus de « datafication » est rendu clair puisque la littérature est découpée dans ses plus petits composants, dé-identifiée, et dé-contextualisée afin de tirer des conclusions empiriques au sujet de phénomènes sociaux, historiques et économiques. Dans cette étude, la littérature est devenue un autre fichier de

données qui, comme l'information statistique et démographique, est utilisée pour tirer des conclusions quantitatives.

Conclusion : des spectres rhétoriques

Le corpus de Google Books, quoiqu'impressionnant par sa taille, n'est qu'une collection parmi d'innombrables autres. Chaque corpus a son propre ensemble de textes, sa propre méthode pour le décrire, et son propre ensemble d'outils pour analyser les données qu'il contient. De plus, ce corpus n'inclut pas le vaste éventail d'autres textes à travers lesquels la culture se manifeste : les dépêches d'information, les notes de musique, les noms d'enfants, etc. Cette pluralité démontre à quel point le cœur même de ces efforts pour quantifier la culture peut être subjectif. J'ai commencé cet article en discutant la tendance vers la digitalisation d'un corpus massif de production textuelle. Cela m'a conduit à une discussion sur la signification même de la lecture, et sur les fausses oppositions binaires qui existent autour de la recherche qualitative et quantitative en littérature. La notion de lecture rapprochée est ce que la plupart des chercheurs dans les humanités pratiquent, un processus qui demande un ensemble d'outils analytiques reposant sur le raisonnement inductif. Diamétralement opposée à cela est la notion de lecture distante dans laquelle les textes eux-mêmes sont réduits à leurs composants unitaires et analysés comme un collectif. La collectivisation des textes apparaît comme parallèle à la « datafication », dans laquelle les caractères d'un texte scanné sont transformés en une information significative qui peut alors être analysée. Ces processus ont produit différentes manières de s'attaquer au corps des données, l'un d'entre eux étant le visualisateur N-gramme de Google Books qui a été utilisé comme un outil à l'intérieur des humanités et à l'extérieur de celles-ci. Les études de cas qui ont été analysées reflètent les différentes méthodes de recherche et la manière dans laquelle les données sont comprises. Plutôt que de faire une opposition binaire entre les méthodes de recherche quantitative et qualitative, je soutiens que ces études quantitatives utilisent les données massives collectives pour établir une rétroaction entre des disciplines et des méthodes traditionnellement disparates.

Tout au long de la discussion cependant, il y a eu un spectre qui a semblé hanter la rhétorique de ces textes. La tendance à quantifier, à décrire les humanités en termes scientifiques, reflète une tension innée entre différentes méthodes de recherche et différentes disciplines. Il y a de nombreuses raisons à ce changement mais même dans ma propre rhétorique et dans mes propres analyses je me suis retrouvée à glisser vers des descriptions réductrices de ces méthodes qualitatives et quantitatives. Finalement on doit se demander si ce sont des paradigmes qui peuvent être réconciliés ou s'ils sont incommensurables. Un fait cependant est certain : la transformation de la littérature en « données massives » n'est pas un processus neutre mais un processus qui a la charge de faire apparaître de nouvelles questions, de nouvelles idées, de nouvelles controverses.

Références

- Allison, S. D., Heuser, R., Jockers, M. L., Moretti, F., & Witmore, M. (2011). Quantitative formalism: an experiment. *Stanford Literary Lab*.
- Bentley, R. Alexander, et al. "Correction: Books Average Previous Decade of Economic Misery." *PloS one* 9.1 (2014).
- Darnton, R. (2013). The National Digital Public Library Is Launched!. *The New York Review of Books*, April 25, 2013 issue.
- Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one*, 6(12), e26752.
- Elder GH (1974) *Children of the Great Depression*. Westview Press
- FAQ - Culturomics. (n.d.). Retrieved March 21, 2015, from <http://www.culturomics.org/Resources/faq>
- Jackson, J. (2010). Google: 129 million different books have been published. *PC World*, 7.

- Karch, M. (n.d.). How to Use the Ngram Viewer - Google Books. Retrieved March 21, 2015, from <http://google.about.com/od/n/a/Google-Books-Ngram-Viewer.htm>
- Heuser, R., & Le-Khac, L. (2012). A quantitative literary history of 2,958 nineteenth-century British novels: The semantic cohort method.
- Leeds-Hurwitz, W. (Ed.). (1995). Social approaches to communication. Guilford Press.
- Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.
- Michel, Jean-Baptiste, et al. "Quantitative analysis of culture using millions of digitized books." *Science* 331.6014 (2011): 176-182.
- Moretti, F. (2011). Network theory, plot analysis. *New Left Review*.
- Moretti, F. (2013). Distant reading. Verso Books.
- Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences*, 107(Supplement 2), 8993-8999.
- Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007 [LIWC manual]. Austin, TX: LIWC.net.
- Schmidt, E. (2005). Books of revelation. *Wall Street Journal*, 18, A18.
- Strapparava, C., & Mihalcea, R. (2008, March). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1556-1560). ACM.
- Taycher, L. (2010). Books of the world, stand up and be counted! All 129,864,880 of you. Inside Google Books. Accessed December, 1, 2013.

Introduction à une réflexion sur la statistique publique



Jean-Louis BODIN

Inspecteur général de l'INSEE honoraire

Statistique et Société a décidé de rendre hommage à Edmond Malinvaud en évoquant différents aspects de sa riche carrière. Dans le précédent numéro, Pascal Mazodier a décrit son apport aux méthodes statistiques de l'économétrie. Aujourd'hui la revue reproduit un article d'Edmond Malinvaud paru fin 1988 dans le *Journal de la Société de Statistique de Paris* ; dans cet article, il dresse un bilan du service public de la statistique dans les pays occidentaux.

1988 est une année charnière aussi bien pour Edmond Malinvaud que pour la statistique mondiale. Fin 1988, voilà une année qu'Edmond Malinvaud n'est plus directeur général de l'INSEE et est devenu Professeur au Collège de France. Comme il l'indique, il s'interroge sur les grandes évolutions qu'ont connues les systèmes statistiques depuis la 2^{ème} Guerre Mondiale et sur les défis auxquels ils doivent faire face. Il dresse donc un bilan complet, lucide et sans langue de bois de l'état dans lequel il laisse l'INSEE et des problèmes auxquels il a été confronté, mais qui ne sont pas très différents de ceux rencontrés par ses homologues des autres pays occidentaux. Il est intéressant de constater que beaucoup de ces problèmes sont encore d'actualité.

Mais 1988 ce n'est qu'une seule année avant que ne s'effondre le Mur de Berlin et que ne débute la transition des pays d'Europe Centrale et Orientale vers des systèmes basés sur la démocratie représentative et l'économie de marché. Et l'article d'Edmond Malinvaud peut aussi être lu comme un ensemble de conseils qu'il va donner, sans le savoir, aux statisticiens de ces pays pour qu'ils puissent résoudre le principal défi qu'ils auraient à surmonter, à savoir gagner la confiance de leurs utilisateurs. Edmond Malinvaud a d'ailleurs été toute sa vie très conscient de la nécessité d'une étroite coopération entre statisticiens au niveau européen et mondial. Il a notamment présidé l'Institut International de Statistique de 1979 à 1981.

Pour tenir compte de cette coïncidence, *Statistique et Société* a demandé à Carlo Malaguerra de commenter l'article d'Edmond Malinvaud à la lumière des presque trente années écoulées depuis sa parution. Carlo Malaguerra est un haut-fonctionnaire fédéral suisse qui a accompli pratiquement toute sa carrière à l'Office Fédéral suisse de la Statistique. Il en a été le Directeur général de 1987 à 2002 et a présidé la Conférence des Statisticiens Européens, organe subsidiaire de la Commission de Statistique des Nations-Unies et de la Conférence Economique des Nations-Unies pour l'Europe, de 1989 à 1992, période qui a vu la chute du Mur de Berlin et de la désintégration de l'URSS. C'est sous sa présidence qu'ont été élaborés les Principes Fondamentaux de la Statistique Officielle, d'abord adoptés au niveau européen. Il a organisé en 2000 la Conférence de l'Association Internationale pour les Statistiques Officielles (section de l'Institut International de la Statistique) sur le thème « Statistiques et Droits de l'Homme ». Il est membre de l'Académie Suisse des Sciences Humaines et Sociales.

Le service public de la statistique en occident : le point actuel¹



Edmond MALINVAUD

Professeur au Collège de France
Ancien Président des Sociétés de statistique

Alors que la production statistique a beaucoup crû en quantité et en qualité, sa précision ne répond pas toujours aux attentes et des lacunes subsistent. Face aux difficultés budgétaires, à un accueil souvent moins favorable du public et à des freins entravant une utilisation efficace des données administratives, les statisticiens réussissent à s'adapter.

De par sa vocation et sa tradition, la Société de Statistique de Paris s'intéresse au service public de la statistique. Or, il se trouve qu'elle en a peu parlé récemment. Mon départ de l'INSEE fournit un prétexte pour combler ce qui pourrait apparaître comme une lacune ; prétexte légitime, car j'aurais pu tout aussi bien faire cet exposé l'an dernier mais l'éloignement risquerait de rendre plus difficile que je le retarde à l'an prochain.

Notre intérêt doit se porter sur les tendances de fond et les problèmes principaux, non sur les détails anecdotiques auxquels nos contemporains, fussent-ils intellectuels, sont trop enclins à prêter attention. Je voudrais donc m'interroger aujourd'hui sur les grandes évolutions qu'ont connu les services statistiques des pays occidentaux et sur les défis auxquels ils sont actuellement confrontés. Cet objectif, de même que la longueur raisonnable de mon exposé, imposent que je sois sélectif. Mes collègues statisticiens officiels trouveront sans doute que je passe trop sous silence des aspects qui leur tiennent à cœur ; effectivement mon choix peut être contesté.

Je parlerai d'abord des produits fournis par le service public de la statistique, puis du processus de production.²

I - LES PRODUITS

Bénéficiant d'investissements souvent anciens et des progrès techniques concernant le traitement des informations, la production statistique a beaucoup crû depuis quarante ans en quantité et en qualité. Cependant cette croissance ne suffit pas à répondre à la demande qui se développe à un rythme comparable. En particulier deux causes d'insatisfaction demeurent : la précision de beaucoup de statistiques ne répond pas à ce que l'on attend ; le service public de la statistique est incapable de fournir des informations fort demandées. Reprenons ces divers points.

1. Communication faite devant les Sociétés de statistique le 22 juin 1988

2. J'ai traité récemment d'un thème voisin, mais en insistant sur d'autres aspects du sujet dans E.Malinvaud, « Production statistique et progrès de la connaissance », ISTAT, Annali di Statistica, Anno 116, Serie IX, vol.7, 1987

I. Des statistiques plus nombreuses et de meilleure qualité

La comptabilité nationale ne nous fournit pas d'évaluation valable du volume des services rendus par les producteurs publics de statistiques. Je ne me hasarderai donc pas à avancer des chiffres pour le taux de croissance de ce volume. Je me contenterai d'évoquer une image, celle d'une courbe logistique. Plus ou moins tôt, les divers pays occidentaux ont décidé durant ce siècle de se doter d'une information statistique de bonne qualité, couvrant les domaines principaux de l'activité économique et de la vie sociale; les difficultés de mise en place des organismes et des instruments statistiques ont souvent fait que le démarrage a été lent ; mais après un temps la croissance de la quantité et de la qualité des statistiques fournies s'est accélérée ; puis les besoins les plus urgents paraissant satisfaits, les coûts de progrès supplémentaires devenant de plus en plus élevés, le rythme de développement s'est ralenti. Aujourd'hui la plupart des pays occidentaux en sont arrivés à cette troisième phase ; mais la croissance y est encore positive et appréciable.

Ce progrès se constate de quatre façons à propos desquelles je peux donner l'exemple de quelques réalisations françaises récentes, à titre illustratif.

Premièrement, des statistiques entièrement nouvelles apparaissent dans certains domaines non précédemment bien étudiés. C'est le cas pour les statistiques françaises sur les entreprises de service qui font depuis quelques années l'objet d'une enquête annuelle. C'est le cas aussi pour la statistique mensuelle des créations d'entreprises. Ce sera bientôt le cas pour la connaissance précise des retraites perçues par toute catégorie de retraités que l'on imaginera.

Deuxièmement des statistiques anciennes sont enrichies, tantôt ayant une couverture plus large, tantôt fournissant un plus grand nombre d'informations. C'est le cas pour les statistiques françaises de prix à la production, ou pour les évaluations régionales et locales de l'emploi.

Troisièmement la qualité de statistiques anciennes est améliorée, et ceci de diverses façons. Il peut s'agir d'une plus grande rigueur conceptuelle dans la définition des grandeurs, par exemple pour les prélèvements obligatoires ou pour le sous-emploi. Il peut s'agir d'une plus grande précision de la mesure, comme pour le suivi mensuel des demandeurs d'emploi. Il peut s'agir de l'élaboration de données plus homogènes couvrant des unités géographiques diverses ou une longue durée, comme ceci se fait de mieux en mieux à l'occasion de chaque changement de base des comptes nationaux ou d'un indice important. Il peut s'agir enfin d'une meilleure lucidité et d'une meilleure information sur la précision effective des données élaborées, comme ceci ressort d'études méthodologiques ou d'un calcul plus fréquent des erreurs d'échantillonnage.

Quatrièmement la diffusion des statistiques est améliorée, en sorte que les utilisateurs puissent en disposer plus aisément. Ce peut être l'établissement de bureaux ayant cette vocation particulière, comme les observatoires économiques régionaux et leurs réseaux de correspondants. Ce peut être l'ouverture d'un service nouveau d'information comme le 3615 - INSEE sur Minitel. Ce peut être la constitution de banques de données comme la Banque de Données Locales regroupant sur chacune des communes françaises tout un ensemble de données les concernant, avec la possibilité de regroupements géographiques quelconques. Ce peut être enfin la mise au point de fichiers statistiques propres utilisables par des chercheurs en vue d'exploitations particulières, comme ceci se fait pour les grandes enquêtes par sondage.

2. Cependant une précision souvent encore médiocre

Ce n'est pas ici qu'il faut rappeler que le grand avantage de l'évaluation chiffrée est d'apporter la précision, mais que c'est aussi son danger quand la précision est illusoire. La lecture ou

l'audition d'un résultat chiffré donne toujours, même chez les plus avertis, une première impression d'exactitude. Or les statistiques sont rarement précises jusqu'à ce qu'il est convenu d'appeler leur dernier chiffre significatif. La situation commune est même celle dans laquelle ce dernier des chiffres dits significatifs n'a rien de significatif.

Sans doute dois-je faire ici une incidente afin de rappeler pourquoi une telle pratique est normale. C'est une vérité, et il est malheureux que l'éducation française évite de la mettre en valeur, que l'activité humaine opère dans un monde incertain, mais qu'il faut bien distinguer incertitude et ignorance. Dès lors qu'elle est objective, c'est-à-dire non intentionnellement biaisée, une information imprécise est meilleure que pas d'information du tout. L'utilisateur d'une statistique a ainsi intérêt à connaître le résultat obtenu tel quel, et ceci d'autant plus qu'il s'agit d'un utilisateur plus sérieux. De plus, une statistique est rarement isolée ; elle figure le plus souvent à côté d'autres avec lesquelles elle entretient des relations de cohérence, souvent apparentes dans les tableaux statistiques. Or les arrondis perturbent ces cohérences, ce qui interdit de les pousser jusqu'à la limite du significatif.

Donc une statistique imprécise est meilleure que pas de statistique du tout. Cependant, j'ajoute immédiatement que l'imprécision devrait être reconnue et même annoncée. Certes le statisticien ne doit pas pêcher par masochisme et dévaluer son travail ; mais renseigner sur le degré de précision d'une statistique est souvent aussi important que de fournir le résultat obtenu. Le renseignement peut rarement être lui-même précis puisque, à côté des erreurs d'échantillonnage, figurent d'autres erreurs de mesure non quantifiables ; il est néanmoins valable pour beaucoup d'utilisations. Je dois reconnaître avoir souvent été mal à l'aise à me dire que les statisticiens français faisaient des efforts peut-être insuffisants pour renseigner sur la précision de leurs résultats.

Quoi qu'il en soit, cette précision est souvent médiocre par rapport aux besoins d'information, et ceci surprend périodiquement ceux qui faisaient une confiance aveugle aux statistiques. En voici deux exemples révélés l'un et l'autre par l'incohérence entre des statistiques d'origines différentes.

Selon les définitions adoptées, l'effectif de la population active est rigoureusement égal à la somme du nombre des personnes ayant un emploi et du nombre des chômeurs. Cependant les évaluations, provenant pour ces trois nombres de sources différentes, ne respectent pas dans leurs variations temporelles l'identité en cause ; le calcul du chômage par différence entre population active et emploi donne, certaines années, une évolution nettement différente de celle fournie par la statistique des chômeurs. Le problème avait été discuté publiquement aux États-Unis il y a vingt ans. Il l'a été à nouveau en France il y deux ans.³

Autre exemple : « le trou noir » des balances des paiements, plus exactement le fait que, depuis un peu moins de dix ans, le total des soldes des balances des paiements courants de l'ensemble des pays du monde est fortement négatif, alors qu'il devrait être rigoureusement nul. Le trou a dépassé 100 milliards de dollars en 1982 et semble être actuellement de l'ordre de 60 milliards de dollars. L'étude d'un groupe d'experts réunis par le Fonds monétaire international a localisé la cause essentielle de cette « discordance statistique », ou tout au moins de sa croissance à partir de 1980 ; elle réside dans les revenus de certains investissements financiers qui, recensés dans les pays débiteurs, ne le sont pas dans les pays créditeurs.⁴

Dans un cas comme dans l'autre, l'étude de ces incohérences statistiques suggère des propositions constructives qui devraient, espère-t-on, améliorer les choses. Mais elle révèle

3. E. Malinvaud, Sur les statistiques de l'emploi et du chômage. Documentation française, Paris 1986.

4. Voir P. Esteva: « Paiements internationaux: le mystère de la discordance statistique est-il levé? », Courrier des statistiques, Octobre 1987.

aussi que le problème subsistera, car il provient des limitations inévitables des sources primaires d'information. Or ces exemples, où l'imprécision se manifeste au grand jour, ont évidemment bien des analogues moins visibles dans les domaines les plus divers.

3. De nombreuses demandes encore insatisfaites

Plus abondantes sont les informations diffusées, plus la soif d'informations s'exprime. A vrai dire, les demandes adressées au statisticien ont parfois un caractère abusif ou contradictoire ; le statisticien doit alors ne pas chercher à répondre à tout. On l'a bien vu il y a quelques années aux États-Unis quand l'information la plus attendue et commentée était celle relative à la masse monétaire hebdomadaire, comme si la situation économique et financière d'un pays pouvait changer d'une semaine à l'autre. On le voit de façon continuelle alors que quasiment les mêmes commentateurs demandent un jour aux statisticiens des évaluations précoces et détaillées, et se gaussent le lendemain des révisions apportées aux évaluations statistiques.

Mais il y a aussi des demandes apparemment raisonnables auxquelles le statisticien voudrait pouvoir répondre et auxquelles il ne le peut pas bien, compte-tenu de ce qu'est l'organisation de la société dans laquelle il vit et des moyens dont il peut raisonnablement espérer disposer. J'en donnerai trois exemples.

Là où existe une statistique nationale couramment utilisée, on imagine trop aisément que puisse aussi exister une statistique pour chaque région et même pour chaque localité ; mais le statisticien ne dispose pas de moyens réels pour fournir cette information localisée. La demande peut émaner d'un simple souci de connaissance désintéressé. Elle peut aussi avoir un objectif décisionnel direct. Ainsi, il arrive souvent que l'on ait à répartir des crédits ou des prélèvements obligatoires entre unités géographiques auprès desquelles la gestion est décentralisée. Pour asseoir cette répartition sur des bases objectives et pertinentes, on demande au statisticien des données souvent assez spécifiques sur les populations-cibles. Celui-ci se trouve alors confronté à l'alternative de soit refuser sa collaboration, soit fournir des données très imprécises ou très inadéquates aux objectifs poursuivis. Le problème s'est posé tel quel aux États-Unis pour la répartition des fonds fédéraux, et dans la Communauté Économique Européenne pour la politique régionale. Bien que les objectifs aient été plus diffus, les statisticiens français ont été aussi mis sur la sellette de façon fort désagréable en 1982 au moment de la décentralisation des pouvoirs.

De même existe de façon récurrente une demande pour une information précise sur des activités cachées, plus ou moins frauduleuses quoique souvent tolérées. Tantôt il s'agit des revenus réels de certaines professions non salariées, par opposition à leurs revenus fiscaux. Tantôt il s'agit de productions totalement irrégulières ou se situant en dehors de l'économie normalement organisée. Tantôt il s'agit de l'immigration clandestine ou des avortements clandestins, etc. Parfois le statisticien arrive à répondre à de telles demandes par des voies détournées, comme par exemple pour les revenus des agriculteurs ; mais il s'agit alors de cas concernant des activités non totalement cachées. Il subsiste des cas dans lesquels les bases manquent pour confirmer ou infirmer les informations fantaisistes qui ne manquent pas de circuler.

Mon troisième exemple concerne la demande d'un suivi statistique des techniques de production. On voudrait avoir une information qui décrive la transformation des techniques dans ses diverses composantes : procédés, degré d'automatisation, consommations de matières premières et énergie, etc. Il s'avère d'une part que les méthodes de production sont à notre époque si complexes, nombreuses et variées que la description statistique, nécessairement simple, se conçoit difficilement, d'autre part qu'aucun enregistrement systématique des techniques employées n'existe à l'intérieur des entreprises, de sorte que l'information disponible

à cet égard dans les ateliers ne remonte guère.

Que conclure vis-à-vis des cas où des statistiques ont une précision insuffisante et de ceux où des demandes légitimes d'information ne sont pas satisfaites, ou le sont mal?

Le statisticien doit évidemment éviter l'attitude qui consisterait à nier l'existence du problème, simplement parce qu'il se sent en paix avec sa conscience. D'abord il ne peut pas être certain d'avoir vraiment épuisé toutes les possibilités, toutes les voies d'approche ; il peut manquer d'imagination ou de pugnacité. Ensuite son attitude serait incomprise par les utilisateurs dont les besoins sont bien réels. On comprend dès lors que des groupes de travail, associant producteurs et utilisateurs de statistiques, se réunissent pour étudier les problèmes coriaces identifiés par les uns ou les autres ; on le comprend quoique la création et la vie de ces groupes fassent souvent penser à la célèbre phrase de Guillaume d'Orange: « Point n'est besoin d'espérer pour entreprendre ni de réussir pour persévérer »

Le rôle du statisticien dans de tels cas est en effet de faire comprendre la nature des obstacles qui s'opposent aux progrès de l'information. Certains de ces obstacles au moins pourraient être levés, mais à des coûts financiers ou autres que la société refuse. Le choix du statu quo est admissible mais le problème ne doit pas être dissimulé.

II. LES PROCESSUS DE PRODUCTION

Passant à la seconde partie de mon exposé, je ne vais pas m'étendre sur les gains considérables que la statistique publique a retiré de l'informatique au cours des dernières décennies. L'utilisation des modes modernes de traitement de l'information a fait disparaître de nos ateliers certaines des tâches parmi les plus fastidieuses qui soient. Elle a surtout permis de conserver à l'information au moment du dépouillement statistique toutes les dimensions de sa complexité. Alors qu'il fallait autrefois limiter assez étroitement le nombre des caractères que l'on relevait et arrêter, souvent une fois pour toutes, la liste des tableaux croisés que l'on produisait par le dépouillement, on peut aujourd'hui conserver et obtenir à coût assez faible des résultats sous des formes adaptées à telles demandes qui se manifesteront.

De plus la généralisation de l'informatique dans les administrations a ouvert la voie à des exploitations statistiques assez aisées des données rassemblées par les services dans le cadre de leur gestion courante. S'étant engagée tôt dans cette voie, la France est souvent apparue comme précurseur d'une évolution qui se manifeste un peu partout.

Je vais négliger aussi les progrès des techniques proprement statistiques. Ces progrès furent cependant considérables. Il y a quarante ans, les enquêtes par sondage aléatoire n'en étaient qu'à leurs débuts ; les concepts de la comptabilité nationale se cherchaient ; beaucoup de nos actuelles nomenclatures étaient inexistantes ou inadéquates ; la méthodologie pour le suivi de certains aspects de la vie économique et sociale était fruste, etc.

Soucieux d'insister ici plus sur les problèmes que sur les motifs de satisfaction, je vais concentrer mon attention sur trois questions: les difficultés budgétaires, l'accueil du public, enfin les freins à une utilisation efficace des données administratives.

1- Les budgets

Les déficits budgétaires, créés ou aggravés par le ralentissement de la croissance économique, ont partout amené les gouvernements à chercher à alléger le coût des charges collectives. Le service public de la statistique a souvent été visé dans la recherche des économies. Avec

l'expérience d'un certain nombre d'années et de nombreux pays, nous pouvons nous interroger sur les sacrifices qui en ont résulté et en résulteront.

Dans cette phase de l'histoire, les services statistiques ont souffert de deux handicaps et de trois avantages pour faire face.

Le premier handicap réside en ce que le coût budgétaire des statistiques est beaucoup plus aisément identifié que leur utilité et que l'efficacité des services chargés de les produire. Partout on a pu entendre des censeurs dire sans preuve que telles statistiques ne servaient à rien, ou qu'elles seraient obtenues à moindre coût par des organismes privés, ou qu'il y avait des doubles emplois, ce dernier langage étant d'autant plus facile que le système statistique était plus décentralisé et faisait donc intervenir un plus grand nombre d'unités. Parfois même le fait que la fourniture d'une information statistique ait la nature d'un service public a été perdu de vue.

Le second handicap tient au très long délai qui s'écoule entre la décision de construire une statistique nouvelle et le moment où elle a effectivement sa pleine utilité. Il serait facile de montrer que le délai moyen est de l'ordre de dix ans. Ce qui est à craindre est donc moins l'abandon de tout un pan de l'information statistique, que le tarissement des investissements dans un secteur important, pour lequel même les préoccupations de maintien de la qualité seraient abandonnées. La dégradation progressive de l'information statistique est ainsi probable sur de longues années dans les pays où les services statistiques ont fait l'objet des coupes les plus sombres.

Pour se protéger les statisticiens publics ont pu d'abord plaider. Au total l'information statistique ne coûte pas très cher par rapport à bien d'autres missions des administrations. A un âge où la collectivité dépense des sommes considérables en faveur des médias, il semblerait paradoxal d'abandonner le souci de la collecte régulière d'une information objective dont l'intérêt est reconnu, quoique non évaluable.

Puis les statisticiens étaient souvent prêts à réaliser des économies de personnel au fur et à mesure que l'informatique devenait plus performante, plus opérationnelle, plus généralement implantée dans les bureaux et les ateliers. De fait, une dégradation du service n'est probable que là où les économies ont imposé une réduction de l'effectif du personnel très qualifié, ce que dans d'autres pays on appelle « les professionnels ». Le cas s'est produit. Je ne pense pas qu'il ait été fréquent.

Enfin les statisticiens ont appris à réaliser quelques recettes en vendant des services qu'ils étaient capables de produire et pour lesquels une gestion efficace exigeait qu'il en soit ainsi. Tantôt il s'agit d'éviter le gâchis, comme lors de la vente des publications, de la tarification des services videotex ou de la cession de fichiers informatiques. Tantôt il s'agit de fournir à des entreprises ou organismes un service plus ou moins spécialisé dont il est normal que les bénéficiaires supportent le coût. Ces recettes ne constituent évidemment qu'un appoint, mais un appoint qui maintenant peut parfois atteindre 10 % des crédits budgétaires.

Au total j'estime quant à moi que, dans la plupart des pays, la rigueur budgétaire n'a pas gravement affecté le service public de la statistique. Il y a toutefois quelques exceptions.

La plus notable concerne les organismes internationaux, plus précisément même ceux appartenant à la famille des Nations Unies. Je n'ai pas besoin d'insister sur le double rôle que doivent jouer ces organismes, dans l'harmonisation des statistiques nationales, grâce à l'adoption de normes communes, et dans la conduite d'opérations statistiques internationales,

telle que celle concernant la comparaison du pouvoir d'achat des monnaies et des revenus réels⁵. Or la situation des services statistiques des Nations Unies est assez déplorable pour que l'on puisse avoir de sérieuses craintes sur la pérennité de ce rôle.

2. Accueil du public

Les statisticiens de pays occidentaux ont eu à faire face au cours des vingt dernières années à un autre groupe de difficultés concernant leurs rapports avec les enquêtés, auprès desquels ils collectent les informations de base. La multiplication des enquêtes et formulaires en tout genre provoque la lassitude et parfois le rejet. Les modes d'existence modernes rendent l'accès aux individus plus problématique qu'autrefois. Les craintes vis-à-vis du pouvoir d'inquisition de l'informatique compliquent nombre d'investigations. Considérons successivement ces trois sources de difficultés.

L'intérêt porté aux données statistiques a amené beaucoup d'administrations et d'organismes à lancer des enquêtes pour leur propre compte. Comme simultanément les programmes statistiques publics se développaient, la durée totale du temps consacré à répondre a beaucoup augmenté. Le phénomène n'a pas de gravité pour les ménages qui ne sont pas souvent individuellement concernés, les taux de sondage restant faibles pour eux. Il en va différemment avec les entreprises, qui ont aussi à répondre à beaucoup d'investigations émanant d'administrations ayant des pouvoirs réglementaires ou des fonctions de contrôle. Aussi les statisticiens se heurtent-ils parfois à la mauvaise humeur des entreprises enquêtées, ou même à l'exigence, imposée par les pouvoirs publics, d'avoir à alléger la charge des enquêtes. Comme précédemment à propos de l'observation des activités cachées, la véritable question est celle du bon arbitrage entre coût de collecte de l'information, y compris auprès des enquêtés, et utilité de cette information. Pour orienter cet arbitrage, il existe dans tous les pays avancés des organes consultatifs ayant reçu mission d'émettre des avis sur le programme des enquêtes statistiques publiques et sur ses modalités d'exécution. En France cet organe est le CNIS, Conseil National de l'Information Statistique, où siègent notamment les représentants des entreprises. L'expérience montre que de tels conseils peuvent bien fonctionner et éclairer utilement les décisions, surtout après quelques années, quand les contraintes et techniques de l'activité statistique ont été comprises par tous les membres. Mais ceci ne suffit pas toujours à aplanir les résistances à la base.

Bien que la charge effective de réponse aux enquêtes statistiques soit faible pour les ménages, on déplore aussi une certaine tendance à l'augmentation du taux de ceux que l'on n'arrive pas à enquêter, soit qu'ils refusent soit qu'on ne puisse pas les joindre. Le phénomène s'explique par l'importance croissante de la population urbaine, par la généralisation du modèle dans lequel tous les adultes du ménage travaillent en dehors du logement, par une certaine aggravation du sentiment d'insécurité etc. Pour y faire face, les statisticiens publics essaient d'adapter leurs méthodes d'enquête et de familiariser à l'avance le public avec leurs investigations.

Enfin, on a assisté à la diffusion d'une inquiétude, voire d'une psychose, vis-à-vis de la généralisation des fichiers informatiques de données individuelles. Or à certains stades au moins de leurs travaux les statisticiens utilisent de tels fichiers, qu'ils les aient constitués eux-mêmes ou qu'ils les aient obtenus en vue d'extraire l'information statistique latente dans des données collectées par les administrations pour leur gestion. Dans divers pays des lois ont été prises afin de régir la tenue des fichiers ; en France ce fut la loi de janvier 1978 sur l'informatique, les fichiers et les libertés.

5. M. Aufrant et H. Picard, « Les parités du pouvoir d'achat entre pays », *Courrier des statistiques*. Octobre 1987.

La législation a habituellement créé un organe chargé du suivi de son exécution, en France la CNIL, Commission Nationale de l'Informatique et des Libertés. Les statisticiens ont eu à exposer précisément l'objet et les modalités de leurs investigations et à tenir compte d'un certain nombre de contraintes qui leur ont été imposées. Celles-ci ont concerné tantôt la constitution et la conservation des fichiers statistiques, tantôt l'accès des statisticiens aux fichiers administratifs, tantôt les fusions de fichiers qui peuvent s'avérer utiles pour la richesse de l'information statistique⁶.

Suivant les pays l'adaptation fut plus ou moins facile ; mais opérer en accord avec un organe indépendant chargé de la protection des libertés a permis d'éviter le rejet massif de certaines investigations statistiques par la population, comme on l'a vu aussi parfois là où aucune législation spécifique n'existait, l'exemple le plus flagrant ayant été l'annulation en dernière minute du recensement de la population de l'Allemagne Fédérale en 1983.

3. Statistiques d'origine administrative

Le dernier point sur lequel je voudrais insister n'est la source de déception que dans la mesure où l'on avait pu entretenir à son sujet des illusions. Conscient de l'abondance des informations collectées par les administrations modernes dans le cadre de leur gestion, conscient aussi de la puissance de l'informatique, on a souvent imaginé que des statistiques toutes prêtes sortiraient aisément des données administratives. Comme il arrive souvent, l'imagination n'était pas foncièrement fautive, mais trop simpliste : entre son rêve et toute réalisation pratique il y avait bien des obstacles à franchir. Aujourd'hui, avec l'expérience acquise, notamment en France, nous pouvons mieux apprécier les contraintes inhérentes à cette voie pour l'obtention de statistiques⁷.

L'avantage en termes de coûts est évident. Pour les personnes qui ont déjà eu à répondre aux administrations comme pour les services statistiques, l'économie d'une nouvelle collecte est très importante, bien que la mobilisation statistique des données administratives comporte aussi des coûts qu'on avait toujours sous-estimés dans les débuts. Dans l'ensemble, la qualité des données élémentaires, si on les utilise avec discernement, est meilleure que dans les réponses fournies aux statisticiens. Il ne s'agit pas par exemple d'évaluer les revenus agricoles à partir des forfaits retenus par les administrations fiscales. Mais la réponse donnée à une administration est fournie avec plus de soin que celle donnée à une enquête purement statistique parce qu'une réponse erronée peut entraîner plus d'ennuis ; le plus souvent c'est donc aussi une réponse plus fiable.

Il reste que la production de statistiques a la nature d'un sous-produit de l'activité administrative. Ceci entraîne une certaine fragilité et une certaine inadéquation.

Une certaine fragilité, car les pratiques administratives peuvent varier. Après qu'ait été organisée une exploitation de fichiers administratifs d'un certain type, il se peut que les procédures de constitution de ces fichiers soient modifiées dans un souci d'efficacité propre à l'administration gestionnaire, que la définition des données change plus ou moins à la suite d'une législation ou réglementation nouvelle, que même les fichiers disparaissent totalement avec la suppression légale des contrôles administratifs qu'ils permettaient. Comme l'information statistique tire souvent sa valeur de sa continuité, tous ces avatars sont fort dommageables aux statisticiens qui doivent chercher à en corriger les effets mais ne le peuvent pas toujours.

6. Pour les règles élaborées en France, voir le chapitre 9, « Recherche et statistiques » dans CNIL, *Dix ans: d'informatique et libertés*, Economica, Paris 1988.

7. Pour un traitement plus complet, voir E. Malinvaud, « From statistics to data management - the French difficulties », *Statistical Journal of the United Nations*, vol. I, n° 3, p. 285-290, 1983.

Les données d'origine administrative concernent rarement exactement ce que le statisticien a pour mission d'observer. Elles sont conçues essentiellement pour les besoins des administrations. Elles sont donc plus ou moins adéquates. Ceci est vrai même en France où, à leur demande, les statisticiens interviennent souvent de façon précoce en vue de manifester leurs besoins quand une opération administrative est conçue ou réformée. Le champ couvert concerne celui de la compétence de l'administration et celui visé par la loi ou le règlement, non celui sur lequel on souhaite avoir une information économique ou sociale ; la liste des grandeurs observées comporte des lacunes par rapport à ce que le statisticien doit faire connaître ; la définition des grandeurs est conforme à l'objectif de l'administration, non à celui de la statistique, etc.

Vis-à-vis de ces manques d'adéquation, qu'il ne faudrait d'ailleurs pas dramatiser, la bonne solution est évidemment de concevoir les statistiques d'origine administrative comme complémentaires de celles obtenues directement par enquête statistique. Ainsi l'enquête annuelle sur l'emploi permet de cadrer et de mieux interpréter les résultats de la statistique mensuelle des demandeurs d'emploi. Ainsi encore l'enquête annuelle d'entreprise est conçue comme devant être complémentaire de l'exploitation systématique des déclarations que les entreprises font chaque année au fisc sur leurs résultats comptables.

*
* *

Ce tour d'horizon, trop rapide, a visé à décrire la situation présente et ses problèmes. Il témoigne de ce que le service public de la statistique n'a eu ni le goût ni le loisir de se complaire dans la routine. C'est une activité vivante en continuelle évolution, dont les progrès ne sont pas toujours faciles mais sont réels.

Commentaires concernant la communication de Edmond Malinvaud : Le service public de la statistique en occident : le point actuel¹



Carlo MALAGUERRA

Ancien directeur général de l'Office fédéral suisse de statistique

C'est sur les bancs d'une petite salle de cours de l'Université de Fribourg (Suisse), vers la fin des années 1960, que j'entendis pour la première fois le nom d'Edmond Malinvaud. Je suivais le cours d'économétrie, nouvellement introduit dans le programme d'études du Professeur Pietro Balestra : il nous avait vivement conseillé la lecture d'un livre qui venait de paraître (1964) aux éditions Dunod, rédigé justement par le Professeur Malinvaud : « Méthodes statistiques de l'Econométrie ». Nous étions tous à la fois soulagés et étonnés de pouvoir consulter un ouvrage en langue française, étant donné que la quasi-totalité de la bibliographie relative à l'économétrie n'était disponible qu'en langue anglaise.

Une bonne douzaine d'années plus tard, j'eus le privilège de faire la connaissance de Monsieur Malinvaud à Genève, au Palais des Nations, dans le cadre de la Conférence des Statisticiens Européens. Je me rappelle très bien cette première rencontre, qui se déroula dans une ambiance que j'avais perçue comme glaciale et qui m'avait ôté toute capacité de m'exprimer correctement en sa présence. J'avais devant moi non seulement l'auteur du livre de mes jeunes années d'étudiant mais aussi une personnalité susceptible de figurer sur les listes de nominations au Prix Nobel et, surtout, le grand patron de l'INSEE, une institution prestigieuse et à la pointe de la science et de la technique statistique publique en Europe et dans le monde. Monsieur Malinvaud rayonnait d'une autorité scientifique et professionnelle incontestée. Il aura fallu plusieurs rencontres à Paris (au Ministère des Finances, pour la présentation de la Banque de Données Macro-économiques BDM), à Neuchâtel (dans le cadre du congrès annuel de la Société Suisse d'Economie et de Statistique – référence 1), à Ascona (lors du colloque consacré au bicentenaire de la naissance de Stefano Franscini – référence 2) et à Paris (lors de la célébration des 50 ans de la Conférence des Statisticiens Européens - référence 3) pour que je découvre - à côté de sa rigueur scientifique et de son engagement sans concession pour la chose publique - la profonde humanité, la générosité et la modestie de Monsieur Malinvaud. J'éprouve pour lui un grand respect et une profonde gratitude.

La statistique et l'État

L'article qu'il me revient de commenter reflète bien les multiples facettes de la personnalité de l'auteur : rigueur de l'approche, capacité de synthèse, clarté des propos, élégance de l'expression,

1. Journal de la Société statistique de Paris, tome 129, n. 4 (1988), p. 227-243 ; reproduit dans le présent numéro.

recherche de l'essentiel. Cet article a été rédigé et présenté juste une année avant la chute du mur de Berlin. D'un point de vue historique, la date de rédaction de la contribution est très intéressante, car elle nous remet en mémoire la situation de la statistique européenne à la veille du grand changement. Deux systèmes et deux conceptions de la statistique publique, fort différents sinon opposés, coexistaient alors : à l'Est, la statistique officielle destinée avant tout au cercle restreint des organes dirigeants de l'État et, à l'Ouest, la statistique publique comme une information accessible à tous les citoyens, c'est-à-dire comme un des piliers essentiels de l'État démocratique. On connaît le vieil adage suivant lequel détenir l'information c'est détenir le pouvoir. La statistique publique est incontestablement un instrument de pouvoir, mais à l'Est elle était un instrument du pouvoir. Entre Est et Ouest, on avait trouvé bon gré mal gré un *modus vivendi*, en ce sens que l'on reconnaissait les différences de conception des deux systèmes sans pour autant mettre en discussion leurs principes de base respectifs.

Dans le titre de la communication déjà, Monsieur Malinvaud postule l'existence de ces deux mondes : il s'agit uniquement de la situation de la statistique publique dans les pays occidentaux, ce qui laisse supposer qu'une autre communication serait nécessaire pour exposer la situation de la statistique d'État dans les pays de l'Est.

Ce qui frappe également dans ce titre, c'est que Monsieur Malinvaud parle du « service public de la statistique ». Cette expression non seulement caractérise la conception de base de la statistique en Occident, mais emploie de façon pionnière le terme de statistique publique. Je rappelle que, jusqu'au début des années 1990, on n'utilisait pratiquement que le terme de « statistique officielle » pour désigner les services statistiques de l'État. C'est un des apports de la réflexion sur les « Principes fondamentaux de la statistique officielle² » que d'avoir suscité un changement de terminologie : en fait, l'État produit de la « statistique publique³ » - au même titre qu'il s'engage pour l'école publique ou les transports publics. Cela constitue bien plus qu'un simple changement d'étiquette : cette nouvelle appellation de la statistique devrait conduire à une nouvelle interprétation de l'activité statistique de l'État. Monsieur Malinvaud avait vu juste : ne serait-il pas temps de suivre son message ?

Un bilan plutôt positif

Le bilan que tire Monsieur Malinvaud de l'analyse de la situation à la fin des années 1980 en matière de statistique publique en Occident, en partant de l'exemple de la statistique française, est plutôt positif. La plupart des pays ont consenti des investissements considérables en matière de production statistique et la quantité d'informations disponibles s'est sensiblement accrue durant les dernières décennies. Des statistiques « entièrement nouvelles » sont apparues, des « statistiques anciennes » ont été enrichies et améliorées, et la diffusion de l'information a été considérablement développée. Toutefois, d'après Monsieur Malinvaud, nombre de problèmes perduraient : la qualité et la précision de l'information sont souvent insatisfaisantes, les difficultés de répondre aux demandes d'information spécifiques des utilisateurs sont évidentes et la charge pesant sur les répondants - malgré l'utilisation de nouvelles techniques de sondage - s'est accrue. Les statisticiens publics n'étaient pourtant pas restés à la traîne : ils avaient su tirer profit de l'utilisation des techniques modernes du traitement électronique de l'information et avaient lancé un vaste programme pour l'utilisation des sources administratives à des fins statistiques. Les statisticiens avaient aussi régulièrement plaidé, notamment dans la phase de l'élaboration des plans pluriannuels de la statistique, pour une meilleure assise budgétaire

-
2. Ces principes ont été élaborés par la Conférence des Statisticiens Européens en 1991 à la demande des pays en transition de l'Europe Centrale et Orientale, suite à la chute du Mur de Berlin. Puis ils ont été adoptés par la Conférence des ministres de la Commission Economique des Nations Unies pour l'Europe en avril 1992. Deux années plus tard, ils ont été adoptés au niveau mondial par la Commission de Statistique des Nations Unies.
 3. Une difficulté de vocabulaire réside toutefois dans l'impossibilité de traduire correctement en anglais l'expression « statistique publique ». De ce fait, la résolution des Nations unies fait référence aux « official statistics » traduit dans la version française par « statistique officielle ».

– non sans succès. Monsieur Malinvaud relève, entre autres, trois défis que le statisticien sera appelé à affronter à l'avenir. Il s'agit tout d'abord du renforcement de la protection des données personnelles, puis de l'exploitation des fichiers administratifs à des fins statistiques et enfin de l'approfondissement de l'utilisation des nouvelles technologies du traitement de l'information. Le public, constate l'auteur, est de plus en plus sensibilisé au stockage de données personnelles dans les divers fichiers administratifs, tout spécialement à cause de l'utilisation généralisée des nouvelles technologies pour le traitement des informations. Il est nécessaire que le statisticien consacre toute l'attention requise à ce problème, d'autant que la statistique publique devra exploiter de façon systématique tous les fichiers administratifs susceptibles de livrer une information valable pour le système statistique. L'auteur met toutefois en garde ceux qui auraient des attentes excessives vis-à-vis de l'exploitation de ces fichiers et ceci pour de nombreuses raisons.

Se garder des espoirs démesurés

Aujourd'hui, à presque 30 ans de distance, les remarques de Monsieur Malinvaud apparaissent prémonitoires. Elles revêtent encore et toujours une actualité indiscutable. Certes, bien des progrès ont été réalisés durant ce laps de temps, mais la sensibilité du public envers tout ce qui est donnée personnelle reste exacerbée. Ainsi, pour ce qui est des informations contenues dans les fichiers administratifs, le statisticien a dû souvent déchanter. Peut-on voir une analogie entre ce thème soulevé par Monsieur Malinvaud en 1988 et le problème actuel des BigData ? Seul le futur nous le dira. A juste titre, la communauté des statisticiens, qui se préoccupe depuis peu de ce phénomène, s'est mobilisée pour pouvoir faire face aux conséquences de l'accumulation vertigineuse de données, structurées ou non, sur les supports informatiques les plus variés et concernant à peu près tous les domaines de l'activité humaine. Sans nul doute, cette nouvelle dimension tant quantitative que qualitative des informations disponibles dépasse largement la problématique de l'utilisation des registres administratifs à des buts statistiques. Il n'y a qu'à relever les défis technologiques, informatiques, de modélisation statistique et mathématique nouveaux qui se présentent au statisticien lorsqu'il veut exploiter les BigData. Mais, conceptuellement, il existe des ressemblances entre les deux grandes catégories de données, c'est-à-dire les données des fichiers administratifs et les BigData : comment relier les « nouvelles données » avec les informations existantes (par exemple les données d'enquête), comment définir la qualité des résultats statistiques, quelles métadonnées sont disponibles, quel niveau de qualité des agrégats peut être accepté, quels problèmes de cohérence se posent avec les nomenclatures, définitions ou classifications existantes, quelle place donner à la protection des données personnelles, etc. Mais il existe aussi des différences liées à la propriété de ces données de masse ; celles-ci sont le plus souvent la propriété d'entreprises multinationales comme Google ou d'exploitants de réseaux de téléphonie et dans ces conditions il est très difficile pour les statisticiens publics de veiller à la qualité des données. Comme dans le cas de l'exploitation à des fins statistiques des registres administratifs, Monsieur Malinvaud estimerait probablement aujourd'hui que l'utilisation des BigData pour la statistique publique ne devrait pas susciter des espoirs démesurés de la part du statisticien.

Quant à l'informatique, son utilisation à la fin des années 1980 et au début des années 1990 était, si l'on peut dire, encore largement limitée à la production de l'information statistique. Pourtant, Monsieur Malinvaud évoque déjà la possibilité d'utiliser ces nouveaux moyens pour le stockage et la diffusion de l'information statistique en évoquant les banques de données, dont l'essor depuis cette date n'a cessé de modeler les systèmes de la statistique publique dans les différents pays.

Un chaînon de l'histoire de la statistique publique

S'agissant des organismes internationaux actifs en matière de statistique publique, notamment ceux appartenant à la famille des Nations-Unies, dont le rôle est essentiel en matière de définitions, de nomenclatures, de classifications etc., Monsieur Malinvaud exprime ouvertement la préoccupation que « la situation des services statistiques des Nations-Unies est assez déplorable pour que l'on puisse avoir de sérieuses craintes sur la pérennité de ce rôle ». Vingt-huit ans plus tard, même si la dotation en moyens financiers et en personnels de la Division de la statistique des Nations-Unies est restée précaire, les craintes de l'auteur ne se sont heureusement pas avérées. J'aurais bien aimé connaître l'opinion de Monsieur Malinvaud sur le rôle d'Eurostat pour la statistique européenne et, pourquoi pas, mondiale. Mais aucune mention n'est faite à ce sujet dans la publication.

En conclusion, la communication de Monsieur Malinvaud est une source d'information très riche pour tous ceux qui s'intéressent au développement dans le temps de la statistique publique. Elle représente un chaînon de l'histoire ô combien captivante de la mission confiée à la statistique publique dans nos pays. Cet apport fait le point sur une période où l'Europe était partagée en deux blocs, chacun avec ses idéologies, ses principes et ses prérogatives. Grâce aux organes des Nations-Unies et tout spécialement de la Conférence des Statisticiens Européens, le dialogue entre statisticiens des deux bords ne s'est jamais interrompu, ce qui a facilité, du moins en partie, le processus des grandes réformes – la transition – dans les pays d'Europe Centrale et Orientale à partir de 1990.

Monsieur Malinvaud aura profondément marqué les destinées de la statistique publique en Europe et dans le monde : il restera un modèle de rigueur scientifique et l'incarnation même d'un grand serviteur de l'État.

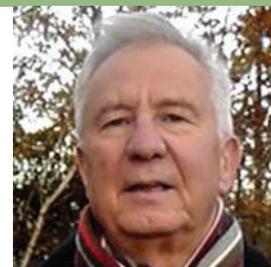
Références

Edmond Malinvaud : Statistiques, analyse économique et économétrique pour la connaissance de l'emploi, in *Revue suisse d'économie et de statistique*, 1994-III, p. 251-265

Edmond Malinvaud : Evolution of Economic and Social Statistics over two Centuries, in *Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini's Birth*. Ed by C. Malaguerra, S. Morgenthaler and E. Ronchetti, Birkhäuser Verlag, Basel-Boston-Berlin, 1997, p. 49-62

Edmond Malinvaud : The development and Achievements of the Conference of European Statisticians up to 1990s – The Point of View of Western Countries, in *50 Years of European Statisticians*. Ed. By Carlo Malaguerra, United Nations, Geneva, 2003, p. 41-51

Compter le nombre de manifestants sur la voie publique : une problématique statistique mais aussi et surtout politique



Thème d'un Café de la Statistique

Pierre MULLER

Inspecteur général de l'Insee à la retraite

Le Café de la Statistique du 10 novembre 2015 a été consacré au thème du comptage des manifestants. Pierre Muller en était l'invité : il a bien voulu mettre son exposé sous forme d'article pour notre revue. Son intervention a donné lieu à des débats animés, dont on trouvera le compte-rendu sur le site de la SFdS à l'adresse <http://www.sfds.asso.fr/ressource.php?fct=ddoc&i=2335> Nous avons sollicité un des participants à ce débat, Jan-Robert Suesser, pour qu'il développe son point de vue : ce qu'il a accepté de faire, avec l'accord de l'invité : cf. son article page 49.

Compter ou pas le nombre de manifestants : plusieurs positions possibles

Le droit de manifester est un droit fondamental, pas seulement d'ailleurs dans les démocraties. Dans ce contexte, s'agissant du comptage du nombre de personnes participant à une manifestation, trois positions (au moins) peuvent être défendues : on accepte les divergences, considérées comme inévitables, entre le comptage proposé par les organisateurs d'une part, par les autorités d'autre part, on cherche un « terrain d'entente » acceptable pour les deux parties en mettant en œuvre des méthodes reconnues (ou acceptées) par elles, enfin on considère que le comptage n'est pas (ou plus) une question essentielle du droit à manifester ou dit autrement, le fait de ne pas compter un nombre de manifestants ne remettrait aucunement en cause la portée et la légitimité de la manifestation.

Ces trois positions ont chacune leur défenseur et leur légitimité respective. La première est celle que l'on rencontre traditionnellement en France. Elle s'accompagne le plus souvent de la pratique consistant à procéder à la moyenne arithmétique entre les deux comptages, constat de l'impossibilité à « réconcilier » deux mesures considérées comme irréconciliables sur le fond. Faire cette moyenne peut éventuellement être admis lorsque l'écart entre les deux comptages est au départ dans une proportion de 1,5 à 2. Elle n'a plus aucun sens lorsque cette proportion s'élève à 3 ou 4, voire plus ! Refuser de compter est une position que l'on rencontre dans certains pays, notamment anglo-saxons, et son écho tend à se développer en France, y compris chez certaines organisations syndicales. Reste que c'est bien la deuxième position, celle qui consiste à rechercher, dans la mesure du possible, un « terrain d'entente » acceptable pour les deux parties grâce à une méthode adaptée, qui paraît (encore?) la plus satisfaisante, s'agissant en particulier d'appréhender la représentativité réelle des groupes exprimant leur opinion sur la voie publique. Cette ambition est d'autant plus légitime que les écarts entre les deux comptages tendent à s'accroître, comme on a pu le constater récemment avec certaines

manifestations « sociétales ». En outre, d'une manière ou d'une autre, la définition et la mise en œuvre d'un « terrain d'entente » de ce type renvoient au rôle de l'Etat, tel en tout cas que l'on peut le concevoir en France. A noter immédiatement que « terrain d'entente » ne signifie pas, à mon sens, recherche d'un compromis entre deux mesures indépendantes. On revient sur ce point en fin d'article.

La méthode mise en œuvre par la Préfecture de Police de Paris

Au cours des dernières années, la Direction du renseignement de la Préfecture de Police de Paris a cherché à s'inscrire dans la problématique (la recherche d'un terrain d'entente par mise en œuvre d'une méthode pertinente) qui vient d'être rappelée. Telle est en tout cas sa volonté affichée. Elle a ainsi mis au point une pratique de comptage qui s'est affinée et rodée au cours du temps.

Peu de temps avant la manifestation, la Préfecture réalise une prévision (sous forme de fourchette) du nombre de manifestants, ce qui lui permet de mettre en place les conditions logistiques de son dispositif de comptage. L'expérience montre que cette fourchette anticipe assez bien le décompte qui sera opéré par la suite. S'agissant des manifestations sur la voie publique d'au moins 3 000 personnes, ce dispositif repose sur un comptage en temps réel dont les résultats font l'objet d'une validation deux jours plus tard grâce à un visionnage en différé. Le comptage en temps réel s'appuie en général sur deux points d'observation de la manifestation, pris en charge par des équipes indépendantes l'une de l'autre et composées de fonctionnaires ayant bénéficié d'une formation adéquate et expérimentés. Les points sont situés en hauteur et non au ras du sol, ce qui permet d'obtenir a priori une vue suffisamment large du cortège et de tenir compte, en particulier, des personnes marchant sur les trottoirs. L'objectif est d'obtenir un visionnage exhaustif de la manifestation, tant dans son trajet que dans sa durée. Les résultats obtenus grâce aux points d'observation sont confrontés entre eux (ils sont proches le plus souvent) mais c'est le chiffre le plus élevé qui est systématiquement retenu. Un redressement, de l'ordre de 10 %, est effectué ensuite par le Directeur du renseignement afin de tenir compte de l'incertitude inévitable affectant les observations, étant entendu que cette incertitude est considérée, le plus souvent, comme un facteur de sous-estimation du nombre de manifestants. C'est ce chiffre redressé qui est communiqué à l'issue de la manifestation. Quant à lui, le comptage en différé est réalisé dans les locaux de la Préfecture par une équipe différente de celles ayant eu en charge le comptage en temps réel. La technique utilisée est la même, si ce n'est que le comptage repose ici sur la rétro-projection de la manifestation dans sa totalité. A priori, la qualité du comptage en différé est meilleure que celle du comptage en temps réel mais ses résultats ne sont pas communiqués au public par la Préfecture.

L'annexe consacrée à la dimension statistique du problème du comptage permet de mettre en perspective la méthode utilisée par la Préfecture de Police. A noter que les médias, comme les organisateurs de la manifestation, ont la possibilité de se rendre sur les points de comptage ou encore de participer à son visionnage en différé. C'est un aspect important de la volonté de transparence affichée par la Préfecture de Police de Paris.

Pourquoi une Commission de réflexion ?

Les améliorations apportées par la Préfecture de Police de Paris à sa méthode de comptage du nombre de manifestants n'ont cependant pas modifié la situation sur le fond. D'une part, les écarts avec les comptages effectués par les organisateurs ne se sont pas réduits. Il faut dire que les manifestations les plus importantes des dernières années (« manifestations sociétales ») ne se prêtent pas naturellement, pour diverses raisons, à la recherche d'un « terrain d'entente » entre Etat et organisateurs quant au nombre de manifestants. D'autre part, la volonté de transparence

de la Préfecture n'a pas, semble-t-il, rencontré un très large écho, tant du côté des médias que des différents organisateurs de manifestations. Chacun est resté plus ou moins campé sur ses positions, tant en matière de méthodes de comptage que de communication des résultats.

C'est dans ce contexte que le Préfet de Police de Paris (Bernard BOUCAULT) a décidé, au printemps 2014, la mise en place d'une Commission ayant un rôle double, à savoir :

- évaluer la pertinence du dispositif de comptage actuellement en vigueur et faire toute nouvelle proposition pour, le cas échéant, en renforcer l'exactitude et la transparence
- vérifier, à l'occasion de chaque comptage d'une certaine importance (au moins 3 000 personnes) la bonne mise en œuvre du dispositif

Le Préfet a demandé à Mme Dominique SCHNAPPER, ancienne membre du Conseil constitutionnel et Directrice d'études à l'école des hautes études en sciences sociales, de présider la Commission. Il a également souhaité la participation de M. Daniel GAXIE, Professeur de sciences politiques à l'Université Paris1 et spécialiste reconnu des questions de démocratie politique. Enfin, le Préfet a contacté le Directeur général de l'Insee afin d'avoir l'appui technique de l'Institut mais sans que cela n'engage officiellement l'Insee dans les travaux de la Commission. Ce faisant, le choix s'est porté sur un inspecteur général de l'Insee à la retraite (Pierre MULLER). Dès le départ, le Préfet a tenu à insister sur l'indépendance des trois membres de la Commission vis-à-vis de la Préfecture même si c'est cette dernière qui a défini le mandat de la Commission. Je peux témoigner que cette indépendance a été rigoureusement respectée.

La désignation de la Commission a quelque peu évolué au cours du temps. Baptisée au départ « Commission de sages », elle s'est ensuite désignée comme « Commission sur l'évaluation du nombre de manifestants ». Le rapport final du printemps 2015 utilise quant à lui l'expression de « Commission de réflexion sur la mesure du nombre des participants aux manifestations de rue », expression qui traduit bien la finalité des travaux de la Commission.

Les membres de la Commission, de même que les objectifs, la méthode de travail et le calendrier des travaux, ont été présentés aux médias lors d'une conférence de presse, le 22 mai 2014. La participation (presse écrite et radios-télévisions) a été conséquente, de même que le nombre d'articles rédigés à la suite de cette manifestation.

Les travaux et les propositions de la Commission d'expertise

La Commission d'expertise a fonctionné pendant un an environ, du printemps 2014 au printemps 2015. Elle a démarré ses travaux par une appropriation de la pratique mise en œuvre par la Préfecture de Police. Cette pratique ne fait pas l'objet d'une documentation écrite très conséquente si bien que l'essentiel des explications a été donné par voie orale, le souci de transparence des personnels de la Préfecture devant, là aussi, être souligné. Certains aspects de la méthode ont fait débat entre les membres de la Commission, par exemple l'existence d'un redressement forfaitaire, de l'ordre de 10 %, des résultats du comptage en temps réel. On notera également ici que le mandat de la Commission ne concernait que les manifestations dont la responsabilité incombe à la Préfecture de Police de la Ville de Paris, à l'exclusion donc des manifestations en Province. La Commission a également pris connaissance des pratiques mises en œuvre à l'étranger, dont certaines s'apparentent peu ou prou à celle de la Préfecture mais sans une formalisation aussi systématique. En règle générale, les pays qui « comptent » le font soit par une méthode globale (produit de la durée et de l'espace occupé par la manifestation, compte tenu d'une hypothèse sur la densité moyenne d'occupation), soit sur la base d'un visionnage plus ou moins « échantillonné », en mobilisant pour cela des techniques et outils

divers, y compris aériens D'autres pays ne comptent pas du tout. Au total cependant, cet examen s'est avéré quelque peu décevant et peu fructueux.

Les membres de la Commission ont assisté au comptage en temps réel de quelques manifestations, celles en particulier du 1er mai 2014 (rassemblement du Front national le matin, manifestation syndicale l'après-midi). Ils ont également procédé eux-mêmes au comptage en différé de ces manifestations et ont testé le comptage réalisé en différé de la « manifestation pour tous » de février 2014. Les tests ont confirmé, dans tous les cas, l'ordre de grandeur des comptages réalisés par la Préfecture de Police.

Des rencontres avec les organisations syndicales nationales ont été organisées à l'automne 2014. Deux organisations ont répondu favorablement à l'invitation de la Commission, à savoir la CFDT et FO. Quant à elle, la CGT n'a pas souhaité donner suite à cette invitation, probablement pour des raisons conjoncturelles sans rapport avec l'objet de la Commission d'expertise. En tout cas, à l'issue de ces deux entretiens, la Commission a pu constater que les syndicats rencontrés n'ont pas (ou plus) réellement de méthode de comptage du nombre de manifestants. En fait, ils considèrent que ce nombre peut être connu ex-ante, du moins le nombre de manifestants mobilisés par leur organisation respective. Inutile dans ces conditions, d'après eux, de mettre en œuvre des méthodes très sophistiquées, cela d'autant plus que le comptage affiché par la Préfecture à l'issue de la manifestation serait nécessairement entaché d'une volonté politique (aspect fortement souligné par FO, moins par la CFDT). Reste que les deux organisations sont sensibles (à un degré quelque peu inégal toutefois) aux problèmes que pose le creusement des écarts entre le comptage par la Préfecture et celui par les organisateurs. Dans ces conditions, elles seraient prêtes à engager une démarche de « réconciliation » avec la Préfecture mais à condition que les autres organisations en fassent de même !

La Commission a procédé par ailleurs à une investigation des différents aspects statistiques de la mesure d'un nombre de manifestants. Elle figure en annexe. Cette investigation, qui mériterait certainement d'être approfondie, met en évidence qu'en règle générale, c'est le visionnage de la manifestation dans son ensemble (visionnage exhaustif) qui constitue la méthode pertinente pour compter un nombre de manifestants (des méthodes moins lourdes peuvent être envisagées mais pour des manifestations d'ampleur limitée ou bien typées) mais également que plusieurs paramètres techniques et/ou organisationnels du visionnage méritent d'être adaptés aux caractéristiques de telle ou telle manifestation. Sont aussi abordées les questions tenant à la précision des résultats et à la communication.

La Commission a remis son rapport final au Préfet de Police en mars 2015. Le rapport a fait l'objet d'une présentation à la presse en mai de cette année. Il valide globalement la méthode mise en œuvre par la Préfecture de Police de Paris. Cette méthode fournit en effet les garanties d'une estimation fiable de l'ordre de grandeur du nombre de personnes participant à la plupart des manifestations ou rassemblements de rue, étant entendu que le visionnage en temps réel doit être couplé, dans la plupart des cas, à un visionnage en différé. Il s'agit bien de la « plupart des manifestations », certaines, particulièrement importantes et/ou complexes dans leur organisation, risquant en effet de poser des difficultés de comptage particulièrement délicates. Toutefois, en s'appuyant sur les différents travaux réalisés dans le cadre de son expertise, la Commission a fait plusieurs suggestions visant à perfectionner la méthode. Deux suggestions portent sur la procédure de visionnage en temps réel (nombre et localisation des points d'observation, outil de comptage) et trois autres sur les conditions de diffusion et de communication des résultats du comptage, avec une insistance particulière sur la transparence de la méthode utilisée et sur la communication de l'ensemble des résultats, y compris ceux résultant du comptage en différé. La Préfecture de Police s'est engagée à mettre en œuvre rapidement ces suggestions.

Quelques remarques pour terminer

La Commission de réflexion sur la mesure du nombre de manifestants a fait œuvre utile. Elle a notamment permis de mettre en évidence le caractère statistique de la pratique de comptage de la Préfecture de Police. C'est dans ce cadre que la Commission a pu pointer quelques insuffisances, s'agissant par exemple d'un manque d'adaptation du processus de comptage aux caractéristiques de telle ou telle manifestation ou encore d'une prudence excessive quant à la communication des résultats obtenus par le comptage en différé.

Par ailleurs, la Commission a pu constater que la méthode de comptage mise au point par la Préfecture est appliquée avec rigueur et conscience professionnelle. En outre, elle n'a pas constaté l'existence de pression politique afin « d'arranger » les résultats diffusés à l'issue du comptage en temps réel. Certes, elle n'a expertisé qu'un nombre limité de manifestations et n'a participé directement qu'au comptage de quelques unes d'entre elles. Il convient donc de rester prudent en la matière. De même, le souci de transparence et de dialogue, aspect souligné fortement à plusieurs reprises par le Préfet de Police lui-même, mérite d'être mis en exergue tout particulièrement même s'il convient de l'améliorer encore à l'avenir.

Pour autant, la Commission ne prétend pas avoir épuisé le sujet, loin de là serais-je tenté de dire ! En premier lieu, il conviendrait de s'assurer de la bonne mise en œuvre de ses suggestions par la Préfecture de Police de Paris. Cela concerne en particulier les propositions relatives à l'outil de comptage et celles visant à améliorer la transparence dans la communication des résultats du comptage. En ce sens, je pense (opinion purement personnelle!) qu'un bilan avant les échéances électorales de 2017 serait utile.

En deuxième lieu, pour être complet, les réflexions de la Commission pourraient être prolongées par un travail de même nature sur les conditions de comptage du nombre de manifestants dans les grandes agglomérations de Province. C'est d'autant plus intéressant que pour des raisons financières (limitation des coûts de déplacement), les organisations syndicales tendent à essaimer leurs manifestations sur l'ensemble du territoire. Pour l'instant toutefois, aucune demande en ce sens ne s'est exprimée.

En troisième lieu, l'analyse statistique menée dans le cadre de la Commission reste largement embryonnaire. Compte tenu de la composition de la Commission, cette analyse me revenait naturellement mais je n'ai pas eu le temps de procéder à une véritable recherche documentaire de la littérature statistique pouvant exister au niveau international sur ce thème. Cela étant, au vu notamment du nombre très limité d'expériences de comptage à l'étranger, on peut penser que cette littérature risque d'être peu abondante !

Reste enfin une dimension essentielle, à savoir l'organisation de la discussion, voire de l'association, avec les organisateurs, syndicaux ou non, de telle ou telle manifestation afin d'arriver avec eux à un « terrain d'entente » sur l'ordre de grandeur du nombre de manifestants. Cet aspect a été peu étudié par la Commission. Certes, il ne faisait pas directement partie de son mandat mais on perçoit bien que l'amélioration du dispositif de comptage, dans ses différents aspects, n'aura réellement de portée que si cette amélioration s'accompagne d'avancées quant à son acceptabilité du côté des organisateurs. Les termes utilisés sont importants : cette discussion ne doit pas avoir pour objectif de confronter deux comptages indépendants (celui de la Préfecture, donc de l'Etat, et celui des organisateurs) mais bien de définir en commun les conditions devant permettre aux organisateurs de reconnaître, d'une manière ou d'une autre, la pertinence du comptage réalisé par la Préfecture, c'est-à-dire par l'Etat. Ces conditions sont susceptibles de concerner l'ensemble du processus de comptage, en amont et en aval de la manifestation : mise en place du dispositif, déroulement du comptage, nombre et emplacement des points d'observation, communication des résultats.

Annexe statistique

Position générale du problème

Sur le plan statistique, compter le nombre de manifestants peut être comparé à une opération de recensement de la population. Il s'agit en effet de déterminer un nombre de personnes en un lieu donné et à un certain moment. Toutefois, contrairement à un recensement classique, il ne s'agit que de cela : on ne cherche pas, par exemple, à décrire les caractéristiques (âge, sexe, diplôme, catégorie sociale, activité...) de la population en cause. En outre, le comptage doit être réalisé en un temps très court, qui correspond à la durée de la manifestation, sans possibilité aucune de reporter l'opération. Cependant, la mesure peut intervenir de façon décalée en regard de l'événement lui-même (voir plus loin la question du visionnage en différé). Dernier point, à savoir que les personnes à compter sont généralement en mouvement, ce qui ajoute à la difficulté de l'exercice en raison du risque de double-compte dans le comptage.

Le plus souvent, il s'agit de compter un flux de personnes allant d'un endroit à un autre à l'intérieur d'une plage horaire plus ou moins bien déterminée. In fine, ce sont là deux paramètres déterminants de toute manifestation, à savoir son trajet et sa durée. Le trajet peut être limité, voire d'une distance nulle, dans le cas d'un rassemblement. Par ailleurs, il peut ne pas être unique si la manifestation emprunte plusieurs voies. De plus, le trajet dépend non seulement de la distance mais également de la largeur des voies. Au total, le trajet détermine l'espace qui sera occupé par la manifestation. Quant à elle, la durée peut être plus ou moins approximative, s'agissant en particulier de manifestations importantes. Outre les paramètres de trajet (espace) et de durée, une manifestation se caractérise également par la façon dont les gens occupent l'espace imparti, ce que l'on peut résumer par densité moyenne d'occupation de l'espace.

Au total, le nombre de personnes à compter pour une manifestation résulte du produit de sa durée, de son espace et de la densité moyenne d'occupation.

Des méthodes de comptage envisageables mais largement approximatives

Une première méthode de mesure peut apparaître évidente, à savoir chercher à évaluer ex-ante chacun des termes de la multiplication mentionnée précédemment. En fait, le problème qui se pose est qu'au moins deux des paramètres (la durée et surtout la densité moyenne) ne peuvent être connus indépendamment du processus de comptage lui-même. Dit autrement, c'est le processus de comptage qui permet de connaître la densité moyenne (c'est vrai également pour la durée, bien qu'un peu moins) et non l'inverse. Reste qu'en faisant une hypothèse a priori réaliste de la densité moyenne d'occupation de l'espace par une manifestation, on peut espérer évaluer ex-ante un premier ordre de grandeur du flux attendu de manifestants. Pour autant, ce type de mesure ne saurait se substituer à un comptage ex-post, étant entaché de trop d'aléas. On pourrait faire une remarque de même type pour les calculs fondés sur l'importance des moyens de communication (cars, trains...) mobilisés par les organisateurs de la manifestation : ce type de calcul est encore plus approximatif que celui reposant sur une évaluation ex-ante des trois paramètres.

On peut également songer à procéder par échantillonnage, par exemple en comptant un nombre de manifestants sur un espace et un temps limités. La question qui se pose ensuite est celle de l'extrapolation de cet échantillon alors que rien ne permet d'affirmer qu'il représente correctement l'ensemble de la manifestation, s'agissant en particulier de la densité moyenne d'occupation, sans parler de l'incertitude affectant la durée effective de la manifestation et les mouvements d'entrées-sorties tout au long du cortège. Dit autrement et en utilisant une

référence statistique, le biais lié à l'échantillonnage est généralement trop important en regard des avantages (de coût notamment) qu'il est susceptible d'apporter. Toutefois, ce type de mesure peut être envisagé dans le cas de manifestations d'ampleur limitée avec un rythme suffisamment homogène.

Le visionnage de la manifestation dans son ensemble : la seule méthode de comptage réellement pertinente

La seule méthode de comptage applicable à tous les types de manifestations est ainsi, en visionnant la manifestation sur l'ensemble de son parcours et de sa durée, de compter le flux tout au long de son déroulement. On peut parler ainsi de recensement exhaustif par opposition à la notion d'échantillonnage. Pour cela, il faut se donner un ou plusieurs points d'observation, ces points devant être soigneusement situés tout au long de la manifestation afin d'être suffisamment représentatifs des trois paramètres de trajet, de durée et de densité moyenne d'occupation. Le nombre de points pourra être fonction du type de manifestation : plus celle-ci sera complexe et/ou importante, plus le nombre de points devra être conséquent.

Les points devront permettre le meilleur visionnage possible des manifestants, afin en particulier de permettre une mesure suffisamment fiable du paramètre de densité moyenne d'occupation. En ce sens, il convient de privilégier un comptage en hauteur plutôt qu'au niveau de la rue, ce dernier n'offrant généralement pas la visibilité suffisante. Toutefois, le visionnage ne doit pas non plus être trop en hauteur (hélicoptère...), ce qui pose, à l'inverse, d'autres types de difficultés. Enfin, les points d'observation (s'il y en a plusieurs) devront être autonomes les uns vis-à-vis des autres afin d'obtenir le nombre maximum d'observations indépendantes. En particulier, les équipes en charge du comptage ne doivent pas communiquer entre elles au moment du comptage.

Le visionnage peut être réalisé en temps réel, c'est-à-dire au cours du déroulement même de la manifestation, mais il peut également être opéré en différé grâce, par exemple, à l'enregistrement de la manifestation. Il importe que le comptage en différé soit réalisé par une équipe distincte des équipes en charge des points d'observation en temps réel. Les deux types de visionnages ne sont pas sans conséquence quant à la qualité du comptage, celle ressortant du visionnage en différé étant meilleure a priori que la qualité obtenue par le visionnage en temps réel.

La qualité de l'observation dépend aussi de l'expérience des « compteurs » et de l'outil utilisé

Le comptage à partir d'un certain nombre de points d'observation procède d'un processus largement manuel. Il faut en effet se donner une référence visuelle et repérer le flux de manifestants traversant cette référence, et cela sur toute la durée de la manifestation. Un comptage personne par personne étant souvent impossible, du moins dans le cas du visionnage en temps réel, il faut procéder en regroupant les manifestants par paquets de volume constant tout au long de la manifestation, opération d'autant plus complexe que la manifestation est irrégulière. Tout cela exige une réelle expérience de la part des personnes en charge du comptage, de même que des procédures rodées de passage de témoin afin de limiter les effets dus à la lassitude. Les compteurs doivent également faire preuve d'une véritable « déontologie », sans aller jusqu'à parler d'indépendance par rapport à leur hiérarchie.

La qualité de l'observation sera d'autant meilleure que l'on disposera d'un outil performant d'enregistrement des paquets, aussi bien dans le cas du visionnage en temps réel que du visionnage en différé.

Sur la précision des résultats du comptage

Même avec un nombre adapté de points d'observation, une expérience suffisante des équipes de compteurs et la disponibilité d'un outil performant, le visionnage en temps réel d'une manifestation ne permettra d'obtenir, au mieux, qu'un ordre de grandeur du nombre de manifestants. La confrontation entre les résultats obtenus par les différents points d'observation ouvre la possibilité de réduire l'incertitude mais elle ne la supprime pas, loin de là. Au demeurant, les écarts de mesure entre les points ne constituent qu'une appréhension partielle du niveau d'incertitude affectant les résultats, un certain nombre de facteurs d'incertitude n'étant pas directement observables à travers la simple existence de plusieurs points d'observation.

On peut inférer deux conséquences des remarques précédentes. D'une part, un redressement des résultats obtenus (généralement positif) par les points d'observation du visionnage en temps réel pourra s'imposer, son ampleur dépendant d'ailleurs des caractéristiques de la manifestation : plus celle-ci sera irrégulière dans son rythme et sa densité, avec un espace distendu, une durée qui s'effiloche..., plus un redressement important s'imposera. Les mieux placés pour l'évaluer seront certainement les équipes de compteurs et leurs responsables. D'autre part, toute communication des résultats obtenus par le visionnage en temps réel ne devrait pas masquer le fait qu'une incertitude plus ou moins grande affecte les résultats du comptage. Une bonne façon de procéder serait ainsi de communiquer non pas un chiffre absolu du nombre de manifestants mais plutôt un intervalle (« fourchette »), en s'étant assuré (dans la mesure du possible) que les bornes de l'intervalle comprennent bien le « véritable » résultat. Cet intervalle pourra s'appuyer sur les écarts prévalant entre les points d'observation mais il ne se limitera pas nécessairement à ces écarts.

La situation devrait être quelque peu différente avec le visionnage en différé. Celui-ci offre en effet des potentialités nettement supérieures au visionnage en temps réel, qui devraient réduire sensiblement (sans évidemment la faire disparaître en totalité) l'incertitude affectant les résultats obtenus à partir des points d'observation. Ce faisant, on peut envisager de communiquer un chiffre absolu suite au visionnage en différé. Pour autant, ce chiffre ne doit pas être présenté comme constituant une évaluation précise du nombre de personnes ayant participé à la manifestation mais bien plutôt, là aussi, comme une confirmation (ou pas) de l'ordre de grandeur du nombre de manifestants.

Nombre de manifestants : « Information partagée » ou « composante de la confrontation »



Jan Robert SUESSER

Administrateur Insee honoraire, 48 années de pratique manifestante

Cette contribution répond à une demande. Avoir dans cette publication une approche du « comptage des manifestants » venant du côté « manifestant », par quelqu'un qui affirme essayer de compter lors des cortèges auquel il participe. Elle n'est pas un apport scientifique. Elle ne résulte pas d'un travail collectif. Elle traduit une préoccupation : est-il raisonnable de penser disposer d'une information partagée sans préjuger de ce que seront ses utilisations ?

Après lecture de l'article de Pierre Muller (pas du rapport complet auquel il fait référence), cette contribution tente d'apporter des éléments de contexte, des compléments, des questions, des propositions critiques.

Manifester ?

Tout au long des décennies, les organisateurs de manifestations ont eu de multiples raisons d'occuper la rue : défendre une cause, affirmer une revendication, exprimer une solidarité, protester par indignation, construire un rapport de force, se faire entendre,... Les thèmes sont tout aussi divers : revendication sociale, environnement, droits des minorités, égalité des droits (y compris pour la majorité – les femmes), discriminations, solidarité internationale.

Compter ? Donner un chiffre ?

Cette diversité des raisons de manifester a des conséquences quant à l'enjeu du comptage. Dans le champ international, par exemple, exprimer une solidarité avec un peuple d'Amazonie luttant contre les accapareurs de leurs terres, protester contre les bombardements à Gaza, manifester contre l'invasion de l'Irak par les États-Unis relèvent d'enjeux différents pour les organisateurs de ces manifestations en France.

Dans le champ national, de même, l'importance du chiffre des manifestants varie pour la succession des manifestations sur la réforme des retraites de 2010 ou de refus du mariage des homosexuels, pour les Gay Pride annuelles, pour la manifestation contre Jean-Marie Le Pen en 2002 ou le 11 janvier 2015 après les attentats contre Charlie et l'Hyper Cacher.

Du point de vue des organisateurs, le chiffre peut n'avoir guère d'intérêt par rapport à d'autres éléments de valorisation de leur initiative. Le nombre peut aussi être central, et en fait un des éléments à la confrontation. Il en est de même pour les autorités politiques. Il en résulte les annonces souvent constatées de chiffres fantaisistes, outrés à la hausse ou à la baisse.

À qui s'adresse le chiffre ? À quoi sert-il ?

Du côté « manifestant », le chiffre peut avoir un usage surtout interne en direction des participants (nous sommes nombreux, la mobilisation progresse, ...) ou plutôt des usages externes vis-à-vis de l'opinion publique et des autorités, en France ou dans d'autres pays.

Le chiffre est d'autant plus important pour les organisateurs qu'il sert à montrer un rapport de force et qu'il contribue à sa construction.

En soi, les chiffres ne sont qu'un élément pour cela, qui interfère avec beaucoup d'autres. Lorsque le chiffre exprime la force de la mobilisation, qu'il fait comprendre à l'autre camp ce rapport des forces, alors son usage « externe » acquiert un positionnement central dans l'affrontement que la manifestation a été appelée à traduire.

Le chiffre des manifestants est aussi un élément d'information pour des décisions du côté des organisateurs, pour son usage « interne ». Il est un indicateur pour juger des opportunités que donne (ou ne donne pas) le processus d'affrontement. En ce sens, le côté « manifestant » a besoin d'avoir de « vrais chiffres » pour son action.

Cette tension entre les usages « interne » et « externe » du chiffre concerne chaque acteur de la confrontation sur laquelle il informe.

Premières conclusions

Un consensus concernant la technique du comptage et son contrôle citoyen se heurte à une réalité complexe : la double nature du chiffre, avec ses usages « externe » et « interne » pour chaque camp en présence ; son insertion dans une série d'éléments qui concourent à la « confrontation » ; son importance, variable selon les objectifs poursuivis.

Dans la réalité, le chiffre n'est pas une information partagée. Il est une composante de la « confrontation », jusqu'à en être un outil essentiel. La question sous-jacente à l'objectif de l'étude demandée par la Préfecture de Paris est « cela peut-il changer ? ». Mais aussi « qui est intéressé par un changement du rôle que joue le chiffre ? ».

Si comparaison n'est pas raison, on se souvient des débats des années 1970 sur les chiffres de l'indice des prix. Pour les contestataires de l'indice Insee, les enjeux étaient forts car ils touchaient à l'évolution du pouvoir d'achat. Les critiques de l'indice officiel étaient virulentes. Avec le recul du temps, on peut considérer que ces critiques ont permis d'importants progrès du côté de l'institution statistique. Quant aux revendications de pouvoir d'achat, elles sont dorénavant argumentées en termes de reconnaissance de la valeur du travail.

Entre-t-on avec le « nombre de manifestants » dans un processus similaire où la contestation conduirait à améliorer les chiffres produits ? Et où le chiffre ne serait plus objet de la confrontation que traduit la manifestation ?

Parler méthodologie

Pierre Muller dit dans son article que les diverses caractéristiques des manifestations ont un effet direct sur le comptage, et donc que la méthodologie doit en tenir compte. Il affirme également, qu'en règle générale, les comptages ne peuvent produire que des ordres de grandeurs. Il constate une meilleure précision des comptages réalisés au fil des ans. De mon expérience, je tire des conclusions analogues, qui cependant laissent des interrogations.

Mais, revenons sur les cas-types de manifestations. Une manifestation « usuelle » consiste à faire converger des manifestants en un lieu donné et de les conduire en un autre lieu en « battant le pavé ». Une mesure « en flux » semble a priori plus adéquate qu'une mesure « en stock ».

La mesure « en flux » consiste à compter ceux qui passent en un lieu du parcours au fur et à mesure du défilé. Cette mesure inclut des réalités différentes : les manifestants occupent la

rue ou aussi les trottoirs, la densité varie, y compris avec de grands « trous » dans le cortège, certains manifestants restent statiques, sans défiler à aucun moment, comme les badauds. J'ai remarqué depuis pas mal d'années, comptant tout seul en choisissant un point du parcours que je trouvais adéquat, que je « tombais » souvent sur le même ordre de grandeur que le chiffre de la Préfecture. Plus souvent pour les manifestations moyennes et grandes (quelques milliers à quelques dizaines de milliers), moins souvent pour les petites et les très grandes manifestations (au-delà de quelques dizaines de milliers). Je parle là du comptage in situ, celui que je pratique.

La mesure « en stock » consiste à compter une foule statique. Ce n'est pas le cas classique pour une manifestation. Mais cela peut être adéquat, et se pratiquer in situ. Qu'on pense aux rassemblements du 11 janvier 2015 dans des villes de province, à une manifestation en 2014 partant de Nation et se terminant en rassemblement à la Bastille, ou à des meetings de fin de campagne présidentielle de 2012 sur l'esplanade du château de Vincennes et à la Concorde.

Combien contiennent les 34 000 m² de la place de la République, les 50 000 m² de la place de la Nation ou encore les 18 000 m² de la place de la Bastille ? Et, combien peut accueillir la place des Quinconces à Bordeaux, la plus vaste de France avec ses 120 000 m² ? Y a-t-il 2, 3, 4 personnes au m², sur toute la surface ou seulement une partie ? Faut-il ajouter des rues adjacentes ? Lors de tels rassemblement qui durent plusieurs heures, cela doit pouvoir s'estimer à notre époque de drones. Pas par moi, le piéton, mais par une Préfecture de police, sûrement.

Lors de manifestations immenses, il y a la question des découragés-de-manifester. Ceux qui sont venus, présents un moment dans le stock mais jamais dans le flux. Ils peuvent quitter le stock et se trouver remplacés par quelqu'un-e d'autre. Là, cela devient vraiment difficile. Il faut penser à des sources d'information complémentaire utiles à valider des comptages avec, par exemple, le nombre de rames de métro qui auront convergé vers les stations où les manifestants débarquent (combien de lignes de métro - dans les deux sens, avec combien de passagers). Éviter de faire des double comptes... Mais ces compléments ne contribuent pas directement pour des chiffrages précis. Il y a aussi l'outil de la comparaison. En 2002 et 2015, deux manifestations géantes se déroulent sur les trois parcours entre République et Nation. Combien d'heures de défilé, mesuré « en flux », chacun ? Quelle différence ? Pour les grands comptes, décidément, l'ordre de grandeur est le seul objectif raisonnable. Et certains chiffres donnés dans le débat public semblent avoir été déraisonnables.

Quelle utilisation du chiffre ?

Ne pouvant traiter là toutes les dimensions de cette question, je n'illustrerais qu'une facette du sujet avec le rôle des médias.

Pour les organisateurs de manifestations, la façon dont le chiffre sera traité par les médias peut être importante. Pas tant s'ils soulignent « ...l'écart des estimations entre la police et les organisations qui appelaient ». Mais sûrement s'ils commentent « Après être passé, entre mars et avril de 1,2 à 2,7 millions de manifestants sur toute la France, les 2 millions d'aujourd'hui semblent démontrer que la mobilisation est sur le recul ».

Dans un autre ordre d'idée : paradoxalement, le débat public peut être phagocyté par un chiffre sorti de nulle part puisque « la Préfecture de Police n'a pas donné d'estimation ». Le 11 janvier 2015, les médias reprennent en cœur « aujourd'hui, nous avons vécu la plus grande mobilisation de rue de tout temps, plus qu'à la libération de Paris, en 1968, pour l'école libre ». Ramenant la légitimité de l'élan populaire au seul fait d'avoir été plus nombreux que jamais, au prix de présenter des chiffres peu vraisemblables... Pourtant, la Préfecture a fait une mesure. Mais, manifestement, les journalistes n'ont pas été intéressés à l'obtenir.

Secondes conclusions

Ces deux exemples, très différents, illustrent une même préoccupation des parties prenantes à l'annonce de chiffres : faire entendre son interprétation de l'évènement de rue, aux participants à la manifestation, à l'opinion publique, aux responsables institutionnels, aux organisateurs,... Cette préoccupation reflète un état de la société, du dialogue sociétal. Les chiffres ne sont pas considérés comme étant d'abord et surtout une information pour un dialogue constructif, un outil d'éducation et de réflexion. Ils sont traités le plus souvent comme des munitions que les parties cherchent à mettre à leur service.

Cela dans un contexte où, lorsqu'ils peuvent être rendus spectaculaires par les médias, ils deviennent un appât jeté dans le débat public, reflétant l'état de la pratique médiatique dominante d'aujourd'hui.

Au-delà de cela, l'utilisation (la més-utilisation) des chiffres de manifestants reflète la qualité des relations de confiance et de considération entre acteurs. Peut-on alors faire converger autour d'approches techniques des parties prenantes à l'utilisation des chiffres de manifestants qui ont des intérêts divergents ? Cela pose des défis considérables. Se concentrent là beaucoup des débats récurrents sur la place des chiffres dans notre société, sur la façon de constituer une information partagée par les acteurs concernés. La perspective d'un consensus sur les techniques de comptage ne pourra se penser indépendamment de la façon dont les dialogues fonctionnent dans la société.

Cela ne doit exonérer en rien les institutions qui fournissent un chiffre de donner une information fiable, construite selon une méthodologie robuste et publique, et contrôlable par ceux qui voudront le faire. Une contribution à une société plus démocratique en quelque sorte.