Statistique et société
décembre 2016

Volume 4, Numéro 3

BigData, marketing, consommateurs, citoyens et entreprises

Sommaire
Statistique et Société

Volume 4, Numéro 3

Sommaire
Statistique et Société

Volume 4, Numéro 3

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.

Design graphique
fastboil.net

ISSN 2269-0271

Éditorial

Emmanuel DIDIER

Rédacteur en chef de Statistique et Société

Cher Lecteur,

Au nom du comité de rédaction de Statistique et Société, je vous souhaite une excellente année 2017. Etant donné les perspectives électorales qui s'annoncent à nous, elle ne manquera pas d'être hautement statistique !

Le dossier de ce numéro de la revue porte sur les usages marketing du Big Data par les entreprises privées, des acteurs sur lesquels nous ne nous penchons pas assez souvent. Le dossier a été réuni par Dominique. Une des questions qui ressort avec force de ce dossier est celle de savoir comment valoriser les masses de données qui sont accumulées par les entreprises. La valeur économique du savoir : une question utile et en même temps profonde. Ce sont les marketers qui s'en sont emparé les premiers, utilisant les données pour mieux connaître les clients et les consommateurs (Crié et Willart). Une autre façon de valoriser les données est de les monétiser pour les échanger (Willart et Crié). Puis nous proposons au lecteur de s'arrêter sur les intermédiaires devenus incontournables de l'extraction de la valeur des données. D'abord sur les algorithmes (Bénavent) puis sur différentes professions et métiers du Big Data (Calcio, Salerno, Moulins). Enfin nous donnons l'exemple concret de la valorisation de ces données par EDF, un acteur crucial du développement du Big Data en France.

A ce dossier, nous avons ajouté deux interventions passionnantes. D'abord, un article présentant les résultats d'une enquête statistique sur la confiance citoyenne dans les statistiques publiques (Chiche et Chanvril). Comme le dit le titre, la relation est contrastée, et de ce fait particulièrement intéressante. D'autre part, une comparaison menée par deux jeunes auteurs prometteurs, sur l'usage différencié des statistiques par Durkheim et Tarde, les deux piliers sur lesquels repose toute la sociologie française encore aujourd'hui (Oehmichen et Viedrov).

Bonne lecture, et n'hésitez pas à nous faire part de vos réactions.

Emmanuel Didier

Introduction au dossier Big Data, Marketing, Consommateurs, Citoyens et Entreprises

Dominique CRIÉ

Professeur des Universités, Université de Lille, Institut d'administration des entreprises[1], Lille Economie Management (UMR CNRS 9221)

Le Big Data envahit aujourd'hui une grande partie de notre quotidien, que nous en soyons conscient ou non. Sujet ô combien à la mode, tout le monde en parle dit-on, mais cependant comme disait Dan Ariely en 2013 : « *Big Data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it*[2] ». C'est finalement un objet toujours assez mal connu, y compris en marketing. Le Big Data doit révolutionner nos vies, nos métiers et façons de travailler, notre santé et nos modes de consommation… grâce entre autres aux individus et objets connectés et à l'exploitation des données qu'ils produisent.

Les cinq articles rassemblés dans ce dossier s'efforcent d'éclairer sous différents angles l'irruption du Big Data dans notre vie économique, et dans notre vie tout court !

Crié et Willart en attestent : c'est souvent par le marketing que le Big Data entre dans l'entreprise en permettant une vision client à 360°, qui plus est en temps réel, y compris en dehors de ses périodes de consommation, en ajoutant à l'analyse de données de nature transactionnelle, des données attitudinales glanées par le biais des médias sociaux. Nouveau capital immatériel, le BigData initie de nouvelles formes de marketing plus interactives, hyper-personnalisées, contextualisées et géolocalisées. On l'utilise pour la formulation stratégique, pour communiquer, re-cibler, personnaliser/individualiser, automatiser et prédire. Ces évolutions substantielles en cours concernent également la recherche marketing et les études où les objets d'attention et les méthodes d'analyse traditionnellement centrés sur les données transactionnelles puis le web, se focalisent maintenant sur l'individu, sa géolocalisation et son environnement. Cette nouvelle ingénierie des données induira inévitablement des changements fondamentaux dans la sphère sociale : les techniques de fixation des prix connaissent une révolution avec la notion de consentement à payer qui définit des prix fluctuant au gré d'une mesure instantanée de la demande ; le secteur de l'assurance se lance dans l'évaluation des risques individuels et au fil de l'eau ; le *data-driven behavior*[3] constate un effet réflexif des données sur les comportements ; les données physiologiques voire biologiques commencent à être mobilisées pour cerner encore plus précisément le consommateur…qui commence à mettre en place des stratégies d'offuscation pour se protéger.

1. Directeur adjoint du laboratoire Lille Economie Management – UMR 9221 CNRS
Dominique.crie@univ-lille1.fr

2. «Le Big Data, c'est comme le sexe chez les adolescents : tout le monde en parle, personne ne sait vraiment comment faire, tout le monde pense que tout le monde le fait, donc tout le monde prétend le faire…»
3. Comportement guidé par les données

Dans un second article, Willart et Crié s'interrogent sur la création de valeur par les données massives et la monétisation directe des données personnelles, sujet émergent mais sans doute appelé à devenir majeur dans un futur très proche. A partir d'un premier constat : l'exploitation des données massives n'est pas nécessairement synonyme de retour sur investissement, ils questionnent les éléments et lieux de création de valeur. Un second constat : la plupart de la valeur extraite des données vient d'usages secondaires, non envisagés au moment de leur collecte, ce qui pose problème dans la systématisation de la chaine de valeur de la data. Si la valorisation des données peut s'entendre de façon duale à travers la valeur d'usage et la valeur d'échange, les données peuvent sans doute être comptabilisées comme des actifs immatériels, essentiels à l'optimisation des décisions de l'entreprise. Bien que traditionnellement présente dans l'environnement marketing bases de données, une nouvelle valeur économique d'échange émerge de plus en plus sur un marché en structuration qui souffre aujourd'hui d'un certain manque de recul. De plus, une stratégie orientée données peut comporter des risques pour les entreprises, notamment ceux liés à l'utilisation de mauvaises données ; et la question de savoir si les Big Data sont effectivement les mieux à même de servir au pilotage de la relation client est posée par un nombre croissant d'acteurs. Leur usage provoque souvent un sentiment étrange d'intrusion et donc des phénomènes de réactance. L'avenir sera-t-il alors aux Small Data ?

De façon immanente, les algorithmes font peur ; peur qu'ils nous privent de notre libre arbitre, peur de leurs capacités prédictives et normatives, illustrées par le *yield management* des compagnies aériennes, les recommandations d'Amazon ou de Deezer, ou les résultats « personnalisés » d'une requête dans Google et qui forcent nos choix, nos goûts ou nos actions. De fait une certaine méfiance gagne les populations : 36% des Britanniques[4] et 22% des Américains[5] jugent que l'IA peut être une menace pour l'homme contre 65% des Français[6] qui craignent la compétition entre l'intelligence humaine et celle de la machine.

A cet égard, Bénavent questionne la redevabilité et la gouvernementalité algorithmique : les algorithmes doivent rendre des comptes. L'idée que les algorithmes ne sont pas neutres et agissent sur nos conduites se développe depuis plusieurs années. Ils peuvent produire des effets sociaux indésirables qu'il convient d'essayer de neutraliser dès la conception. L'algorithme doit être loyal à l'intention initiale. Si un vendeur ou un moteur de recherche recommande des produits ou services pour lesquels il a un intérêt particulier, l'algorithme est déloyal ; situation inacceptable qui plus est pour des produits ou services publics. Au travers d'exemples didactiques des éléments de réflexion et d'analyse pertinents nous sont proposés.

Plus techniques, Calciu, Salerno et Moulins soutiennent que « les Big Data sont trop importantes pour être laissées aux informaticiens » ! Les analystes marketing doivent investir le champ et contribuer aux nouvelles approches qui révolutionnent la science des données. En démystifiant et explicitant «la mécanique» Big Data, ils montrent que tout chercheur, ou toute entreprise même de taille modeste peut avoir accès à ces technologies de la donnée, pourvu qu'elle s'attache les compétences adéquates. En effet beaucoup d'outils récents sont de nature « open-source » ; ils peuvent s'organiser en écosystème autour du logiciel statistique R, et si l'on ajoute la possibilité de calcul distribué d'Hadoop, le traitement des Big Data se démocratise, offrant une opportunité exceptionnelle aux statisticiens et analystes marketing "d'opérer" de vraies usines à calcul, exemples et tests à l'appui. Le marketing repose de plus en plus sur les technologies de l'information tout en étant considéré comme l'un des moteurs des technologies Big Data à l'instar de la comptabilité analytique pour les bases de données dans les années 80. Les scientifiques du marketing semblent avoir quelques difficultés en matière de programmation moderne et de technologies de l'information, comme l'appréhension du contexte distribué,

4. Sondage YouGov pour la British Science Association, mars 2016
5. Sondage de l'Université de Chapman, avril 2015
6. Étude réalisée par l'IFOP pour l'Observatoire B2V des Mémoires, novembre 2015

de la richesse des données issues de l'Internet, du cloud computing ou encore du HPC (*High Performance Computing*). Démystifier vous dis-je ou plutôt nous disent-ils !

Enfin, de façon plus pragmatique, Anne Gayet et Jean-Michel Gautier, exposent toute la richesse d'une approche Big Data dans une grande entreprise comme EDF et ses enjeux pour le marketing client. La vélocité (temps réel) et la variété des sources de données apportent une autre vision, plus globale, de l'interaction client et permettent le développement de nouveaux services. Les parcours clients font l'objet d'une analyse interactive, et la facilité à acquérir de nouveaux flux de données a complètement changé l'appréhension de la gouvernance de la donnée, dans un environnement essentiellement open source. La « dataviz » complète le dispositif et rend l'information immédiatement intelligible. La transformation Big Data permet ainsi de mieux appréhender les mutations du secteur et de concevoir des services innovants répondant aux attentes des consommateurs.

Cette plongée au cœur des problématiques marketing liées au Big Data n'est qu'un aperçu, finalement assez sommaire, des bouleversements qu'engendrera sans nul doute sa pratique. Reste à chacun d'imaginer ou de rêver la (les) suite(s)…

Le marketing et les mégadonnées

Dominique CRIÉ
Professeur des Universités, Université de Lille,
Institut d'administration des entreprises,
Lille Economie Management (UMR CNRS 9221)

Sylvain WILLART
PhD, Risk analyst, Advanzia Bank ; Maître de Conférences
Université de Lille

Dans l'entreprise, c'est souvent le marketing qui ouvre la voie à l'adoption de l'analytique[1] puis à la révolution Big Data en raison de son appétence naturelle pour l'analyse et de sa déjà forte expérience dans le traitement statistique des données dites traditionnelles, stockées dans des entrepôts pouvant être déjà de taille conséquente (Erevelles, Fukawa & Swayne 2016). Certes l'analyse porte alors essentiellement sur des données structurées et de façon synchronique, afin d'offrir une vision comportementale généralement sur un événement précis. C'est l'exemple du ticket de caisse ou de la carte de fidélité qui photographie par instants le consommateur dans ce qu'il fait. Avec le Big Data on passe à une vision holistique, le fameux 360°, qui filme l'individu (et non plus le consommateur car aussi en dehors de ses périodes de consommation) pour saisir ce qu'il est, au sein des réseaux sociaux, dans la tonalité de son discours, dans les multiples traces qu'il laisse sur le web, dans les surfaces de vente ou call-centers, chez les concurrents ou encore dans sa mobilité ou ses horaires de vie…

De façon plus technique et toujours du point de vue marketing, ce qui différencie le plus l'approche « données traditionnelles » du Big Data, c'est la notion de temps réel et de vitesse de traitement. C'est cette caractéristique qui va influencer le plus la pratique marketing. Cette notion de temps réel sous-tend en fait trois types d'éléments : la production et la captation des données qui ne s'analysent plus en stock mais en flux continu, le traitement, l'analyse et l'interprétation instantanés de ces flux de données et enfin l'action marketing simultanée qui est déclenchée. Le second critère d'innovation est sans doute la variété. Une autre différence majeure entre Big Data et données traditionnelles est le passage de données de nature transactionnelle et structurée à des données surtout attitudinales et non structurées par le biais des médias sociaux, où les individus partagent des informations personnelles avec des amis et la famille.

Le Big Data peut également s'analyser comme **un nouveau capital ou une nouvelle valeur patrimoniale immatérielle** (Mayer-Schönberger & Cukier, 2013). Les données fournissent un *insight*[2] comportemental des consommateurs, le marketing traduit ces *insights* en avantages

1. L'analytique combine l'informatique, le traitement des données, la statistique et la modélisation pour adresser des problématiques managériales, industrielles ou sociales.
2. Un « consumer insight » est une attitude, une attente ou une façon d'envisager les choses par les consommateurs, utile pour la définition des politiques marketing et d'innovation.

marché (Bharadwaj et al., 2013). Par exemple, l'analyse en temps réel et de données en streaming (*Event Stream Processing*) ainsi que la prise en compte de la mobilité (terminaux et géolocalisation) sont des accélérateurs de valeur ajoutée (McAfee & Brynjolfsson, 2012). Dans ce sens le Big Data est bien une nouvelle forme de capital.

Quelques utilisations du Big Data en marketing

De toute évidence, le Big Data peut avoir un impact sur tous les domaines du marketing, que ce soit comme nouvelle source de génération d'idées pour le développement/différenciation des produits ou services, de stratégie prix ou de positionnement concurrentiel, de publicité communication ou encore en matière de distribution… Globalement tous les éléments du mix classique[3] peuvent bénéficier d'une approche par les Big Data (Fan, Lau & Zhao, 2015).

Le Big Data initie de nouvelles formes de marketing plus interactives (allant jusqu'à l'affichage dynamique interactif comme dans le film de S. Spielberg, *Minority Report*) ayant pour but d'accroitre la productivité marketing et le ROMI[4] … Le déluge de données disponibles sur le consommateur sonne la fin du marketing de masse. Il rend la segmentation, le ciblage et le profilage des clients plus pertinents, permettant d'opérationnaliser un marketing hyper-personnalisé et contextualisé allant jusqu'à l'anticipation des besoins et désirs. L'ensemble de ces techniques s'inscrivent dans le « *real-time marketing* » ou le marketing programmatique[5] ou encore l'*adtech*[6], qui grâce à une très grande variété de données, de provenance multiple, en temps réel et à l'ingénierie de l'écosystème Big Data (Demchenko, Ngo, & Membrey, 2013), autorise la proposition d'une offre ou d'une communication personnalisée instantanée. Cette personnalisation de plus en plus fine et tendant vers le marketing individualisé est automatisée par des algorithmes utilisant de nombreuses variables comme les données de navigation et celles relatives au contenu de la page web visitée afin de maximiser la probabilité que l'internaute clique sur la proposition mise en avant.

Le Big Data en marketing pour…

- La formulation stratégique[7] : L'utilisation du big data dans la formulation stratégique est évidente, même si elle n'est pas encore perçue comme essentielle par la majorité des entreprises à l'exception des GAFAM (Google, Apple, Facebook, Amazon, Microsoft) et autres pure players (Gantz & Reinsel, 2011). Nouveau capital, temps d'avance sur les concurrents, innovativité en matière de produits/services, proximité avec les clients, leurs besoins et désirs, accès à une information en temps réel et géolocalisée…., tous ces éléments influencent la structure organisationnelle dans l'objectif d'améliorer ses capacités d'adaptation et permettent à l'entreprise d'être beaucoup plus agile que ses concurrents.

- Communiquer : Les utilisations les plus visibles en matière de segmentation/ciblage/profilage des clients sont multiples (Chen et al., 2012) : moteurs de recommandation, *retargeting*[8] (e.g. Criteo), « *real-time bidding*[9] » qui permet à l'annonceur d'acheter, dans un système d'enchères automatisé et en temps réel, puis d'afficher des bannières publicitaires contextuellement à la navigation et au profil de l'internaute. L'efficacité apportée par la nature instantanée du ciblage est l'un des critères majeurs du développement de cette technique publicitaire, d'autant qu'elle permet un suivi de la réaction du client et favorise directement les ventes additionnelles. Cette

3. Le mix marketing de base comporte classiquement les 4P : le Produit, le Prix, la Place (circuit de distribution), la Promotion (publicité)
4. Return On Marketing Investment
5. Marketing dont les actions sont automatisées selon des algorithmes qui « réagissent » en fonction des comportements du consommateur
6. Pour advertising technologies, désigne le secteur technologique de la publicité digitale
7. Méthodologie d'élaboration du plan stratégique
8. Le retargeting est le ciblage secondaire de consommateurs ayant manifesté un comportement donné sur le net.
9. Real-Time Bidding ou RTB : Achat d'espaces publicitaires sur le web, par enchères en temps réel.

nouvelle temporalité des données autorise de nouveaux modes opératoires marketing. Par exemple, le *native advertising* (publicité native) est un mode de publicité en ligne qui se fond dans le media support consulté par l'internaute et qui enrichit l'expérience utilisateur. Le message, reprenant les codes et le design des contenus (souvent sous forme vidéo) est alors perçu comme moins intrusif, plus naturel et attire d'autant plus le clic de l'internaute. Cette technique s'intègre parfaitement à l'organisation en flux de l'information des media sociaux ce qui la rend encore plus naturelle vis-à-vis de l'utilisateur. Elle nécessite d'être, dans la forme et le fond, en cohérence avec les sujets d'intérêts et les attentes des consommateurs et ne pas perturber leur expérience utilisateur. L'affichage publicitaire plus traditionnel (*display*) sous forme de bannière et autres formats adaptés à l'internet (*banner, pop-up, rollover ads*….) bénéficie également du Big Data et de ses algorithmes qui agrègent les données personnelles et analysent les traces de navigation émanant de sites tiers afin de mieux cerner nos besoins et désirs et ainsi affiner le ciblage par la connaissance d'un comportement global en temps réel sur le web.

- Recibler : Un autre moyen de déclencher une action positive de l'internaute est le retargeting ou publicité comportementale car elle est déclenchée suite à certains comportements, comme l'abandon de panier par exemple. Grâce à une analyse précise de nombreux facteurs internes et externes autour d'une navigation sur un site, le Big Data, par une connaissance élargie des clients, permet de sélectionner les prospects les plus susceptibles de réagir à une nouvelle exposition à une offre donnée et donc d'accroître le taux de conversion. L'offre est donc reproposée en *display* au cours d'une connexion ultérieure quel que soit le site visité (e.g. Criteo).

- Personnaliser/individualiser : Les GAFAM, mais d'autres géants du net également, ont massivement investi dans les outils de caractérisation des clients et donc de ciblage et de personnalisation tant dans le contenu que dans l'expérience utilisateur. On aboutit alors à un univers commercial (et par là une navigation) complètement différencié avec un affichage, une expérience utilisateur, des recommandations, adaptés et différents pour chaque individu. Ces services ou produits ultra-personnalisés sont moteur d'innovation, de satisfaction client et conduisent à la réalisation de plus grandes marges de profits (Barton & Court, 2012).

- Automatiser : Les plateformes de données (Data Management Platforms –DMP-), véritables centres névralgiques de la *customer intelligence* qui permettent l'agrégation de multiples flux de données, forment le cœur du marketing automatisé, envoyant les relances ciblées par des algorithmes en fonction des interactions avec les clients ou prospects et par le media (canal) le plus adapté au moment et à la situation de l'individu. Ces techniques sont également le support d'un parcours client omnicanal fluide dit « sans couture » c'est-à-dire sans que le client ne soit gêné par le changement de canal[10] dans son parcours d'achat.

- Prédire : Les données internes aux entreprises sont souvent riches car elles peuvent être issues de l'activité des clients et prospects sur un ensemble de media (*owned media*) créés par/ou associés à l'entreprise (blogs et sites affiliés, pages facebook, comptes twitter, apps, YouTube, newsletters, e-mails…etc…). Cette ingénierie des données permet une connaissance affinée des segments et des individus autorisant une optimisation des contenus des messages, des recommandations, des canaux à utiliser et du tempo adaptés au client/prospect. Une modélisation de l'ensemble de ces données par des techniques prédictives renforce la capacité d'anticipation des besoins et désirs des consommateurs et accroit les taux de transformation (Banker, 2014 ; Ritson, 2014).

10. On entend ici les modalités de contact et de vente, généralement interactifs

Des évolutions substantielles en cours

Le Big Data induit une révolution dans de nombreuses activités marketing comme dans la chaine de création de valeur ou dans l'obtention d'avantages compétitifs mais aussi en termes de comportement des consommateurs.

- La **recherche marketing et les études** sont les premiers départements à être affectés par l'irruption du Big Data. Traditionnellement ce sont les études qui fournissaient aux entreprises une idée de l'opinion et du comportement de leurs clients vis-à-vis de leurs produits. Grâce au Big Data cette information est directement inférée par l'analytique des réseaux sociaux et autres données comportementales (Lycett, 2013). Le *machine learning* entraine la disparition des modèles, dans une logique de boîte noire, sans que l'on sache exactement quel a été le « chemin » qui a abouti au résultat, éloignant ainsi le chercheur de la compréhension profonde des mécanismes de prédiction et notamment de l'assurance que des variables essentielles au processus de décision ne soient pas oubliées. Un **renversement de paradigme**, bouleversant le raisonnement scientifique traditionnel d'ajustement des données à des théories préconçues du marché vers l'utilisation des données pour construire de nouvelles théories, est en train de se produire, à savoir une évolution du déductif vers l'inductif (Erevelles, Fukawa & Swayne, 2016). Cette démarche rend le chercheur moins dépendant des connaissances existantes et lui permet de se concentrer davantage sur ce qui est inconnu (Sammut & Sartaoui, 2012), mais l'éloigne de la compréhension intime des phénomènes.

- A un autre niveau, **les objets d'attention et les méthodes d'analyse** changent également de nature. En termes de Business Intelligence et d'analytique, l'évolution des pratiques peut se résumer en trois phases (Chen et al., 2012) : la phase 1.0 où les analyses sont ***centrées sur les données***, essentiellement *structurées*, collectées et stockées sur un mode ***statique*** dans des *entrepôts de données*, gérés dans un système de *bases de données relationnelles* et traitées au moyen d'analyses statistiques classiques et du datamining ; la phase 2.0 où les analyses sont ***centrées sur le Web*** à partir d'un contenu non structuré (95% des données du Big Data), avec des modes de collecte et de stockage ***dynamiques*** et traitées par des techniques avancées de *web-mining* et d'analyses spatio-temporelles ; la phase 3.0 où les analyses sont ***centrées sur les individus, la géolocalisation et la contextualisation***, les contenus issus des mobiles et objets/capteurs connectés et les analyses sont essentiellement de nature spatio-temporelle et contextuelle focalisées sur les interactions [homme-objets] pour rejoindre le paradigme P.O.S. (Personne-Objet-Situation) dans le cadre du marketing expérientiel, où le consommateur n'est pas que rationnel mais en quête d'émotions, d'expériences, de sensations ou encore de lien social.

- Le Big Data ne sert naturellement pas exclusivement à mieux satisfaire les attentes des consommateurs, l'objectif est d'augmenter le volume d'affaires avec eux dans les meilleures conditions possibles pour l'entreprise. C'est par exemple l'objet du ***yield management*** (tarification en temps réel) et de sa généralisation le ***dynamic pricing***[11] (tarification dynamique) qui trouvent avec le Big Data un instrument d'optimisation extraordinaire et font basculer la notion de prix vers celle de consentement individuel à payer… une autre révolution induite par le Big Data s'il en est et qui permet à une organisation de mettre en œuvre une stratégie de tarification flexible, réactive et différenciée et de maximiser les profits. Le prix consenti résulte alors d'un calcul complexe fait par le consommateur en fonction de ses besoins instantanés ou anticipés, d'influenceurs ou prescripteurs et de son interaction avec le vendeur. Uber utilise

11. Yield management ou gestion des capacités de service par le prix. Le dynamic pricing c'est la modulation du prix en fonction des conditions du marché et des caractéristiques du client. Le yield management prend en compte des données exclusivement liées au marché et à la demande, le dynamic pricing prend en compte des données liées au consommateur, notamment son consentement à payer qui dépend de la connaissance qu'il peut avoir du marché.

une forme de ce principe (*surge pricing*) pour ajuster parfaitement l'offre et la demande en cas de forte demande et ainsi satisfaire les acteurs des deux versants de son marché, clients et chauffeurs.

- Le marketing devient créateur de normes sociales. Par l'observation des comportements individuels, des secteurs entiers de l'économie vont réviser leurs standards et changer de paradigme pour aboutir à la création de nouvelles normes par le marketing. Par exemple, dans l'assurance, le Big Data d'une part et les objets connectés d'autre part permettent d'affiner de façon considérable le calcul du risque au niveau individuel. Allianz propose d'équiper les conducteurs de boitiers connectés qui permettent de les segmenter en «communautés comportementales» tarifaires (Grandin de l'Eprevier, 2016), ce qui remet fondamentalement en cause le principe de mutualisation des risques, mais qui joue un rôle d'éducation ou de norme pour la partie des automobilistes qui pensaient être de bons conducteurs mais qui n'ont obtenu aucune réduction de prime (e.g. youdrive pour les jeunes conducteurs).

- Le data driven behavior (comportement guidé par les données) suppose que le consommateur (ou l'individu) se comporte selon une norme implicite créée par les données et les métriques qu'elles sous-tendent, souvent différente des normes comportementales habituelles et participant à la génération d'autres normes. Dans une logique de comparaison et de feedback social, il existe un effet réflexif des données sur les comportements, un *datadriven behavior* qui mesure le réel pour mieux le transformer, la donnée devient alors un outil ou une extension de soi (e.g. quantified-self, programme MesInfos...).

- L'intimité biologique sollicitée. Des multi-capteurs portables permettent aux entreprises de recueillir une toujours plus grande variété de données (Hardy, 2012). Certaines entreprises (e.g. Proteus, Sano Intelligence, MC 10) ont commencé à collecter, stocker et analyser des données physiologiques en temps réel telles que la fréquence cardiaque, l'activité cérébrale et la température du corps, en utilisant des capteurs portables. Ces données physiologiques en temps réel, combinés avec des mesures comportementales telles que les visites en magasin dérivés de l'information géospatiale ou de *beacons*, offrent la possibilité d'inférer des *insights* consommateurs à la fois comportementaux et émotionnels. Jusqu'où aller dans la compréhension du consommateur et est-ce judicieux ? Pourra-t-on prédire les achats d'impulsions, les changements subits de marque ou de produit ? N'est-on pas dans une spirale qui peut aboutir au contraire à un contrôle plus important, de la part du consommateur, sur les données qu'il émet et qui aboutira à une stérilisation du processus?

Les réactions des clients...

Devant la pression publicitaire ou comportementale exercée par ces technologies, les consommateurs se rebellent en utilisant par exemple des outils permettant de bloquer la publicité (e.g. Adblock). Sans que les professionnels le reconnaissent vraiment, un grand nombre de leurs annonces ne sont pas affichées sur le terminal du client ciblé. Des stratégies plus générales d'offuscation consistent à produire de la fausse information de manière à masquer ou rendre ininterprétable l'information pertinente comme par exemple cliquer automatiquement sur toutes les annonces publicitaires ou générer des requêtes aléatoires sur Google (e.g. TrackMeNot) afin que celles réalisées par l'internaute n'aient aucun sens vis-à-vis de ses centres d'intérêts, ou encore effacer toute trace de navigation sur un site (Disconnect). D'autres stratégies sont possibles, le recours à des pseudonymes multiples, le brouillage d'IP (e.g. IP Hider) ou encore le recours aux solutions de VRM[12] (e.g. Onecub) qui constituent un écran entre le commerçant et le client.

12. Le VRM ou Vendor Relationship Management est le pendant du Customer Relationship Management (CRM). Ici le client prend le contrôle de la relation qu'il veut entretenir avec l'entreprise.

Un cadre juridique à écrire

A qui appartiennent les données ? À ceux qui les produisent ou à ceux qui les récoltent ? Derrière cette question se cache celle de la réelle valeur de ces données personnelles et donc celle de leur monétisation, notamment vis-à-vis des entreprises dont le business model est exclusivement adossé aux données et traces du consommateur et qui monétisent les données, les modèles, les actions commerciales et leurs résultats. Aucun dispositif législatif ou réglementaire n'encadre encore la collecte et l'utilisation des données, même si la récente Loi Numérique pose quelques jalons en France. Par contraste, le droit à l'image, (qui n'est autre qu'un ensemble de données ou pixels), permet à quiconque de s'opposer à l'utilisation, commerciale ou non, de son image, au nom du respect de la vie privée. Pour Russel W. Belk (2013) : « les données que nous possédons reflètent l'identité de leur propriétaire, elles contribuent à notre identité, elles sont un bien virtuel ».

Références

- Banker, S., (2014), "Amazon and anticipatory shipping: A dubious patent?" *Forbes* (consulté le 8 mai 2016)
- Barton, D. & Court, D. (2012), « Making Advanced Analytics Work for You », *Harvard Business Review*, Vol. 90, n° 10, p.78-83.
- Belk, R. W., (2013), "Extended self in a digital world", *Journal of Consumer Research*, 40, 3, 477–500
- Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., & Venkatraman, N., (2013), « Digital business strategy: Toward a next generation of insights », *Management Information Systems*, 37(2), 471–482
- Chen H., Chiang R. H. L., & Storey V. C., (2012), "Business Intelligence and Analytics: From Big Data to Big Impact", *MIS Quaterly*, Special Issue: Business Intelligence Research, 36, 4, 1165-88
- Demchenko Y., Ngo, C., & Membrey, P., (2013), Architecture framework and components for the big data ecosystem, *Journal of System and Network Engineering*, pp 1-31
- Erevelles S., Fukawa N. & Swayne L., (2016), "Big Data consumer analytics and the transformation of marketing", *Journal of Business Research*, 69, 897–904
- Fan S., Lau R.Y.K. & Zhao J. L., (2015), « Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix », *Big Data Research*, 2, 28–32
- Gantz J. & Reinsel D., (2011), "Extracting Value from Chaos", International Data Corporation (IDC) IVIEW, June 2011, disponible à : http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
- Hardy, Q., (2012), « Big data in your blood », *New York Times*, 7 September 2012
- Grandin de l'Eprevier Jade, (2016), « Trier les bons et les mauvais conducteurs, le pari connecté d'Allianz », *Le Monde*, 13.04.2016
- Lycett,M., (2013), « 'Datafication': Making sense of (big) data in a complex world », *European Journal of Information Systems*, 22(4), 381–386
- Mayer-Schönberger, V., & Cukier, K. (2013), *Big data: A revolution that will transform how we live, work, and think*, New York: Houghton Mifflin Harcourt
- McAfee, A. et Brynjolfsson, E. (2012), "Big Data: The Management Revolution", *Harvard Business Review*, Vol. 90, Issue 10, p60-68
- Ritson, M., (2014), « Amazon has seen the future of predictability », *MarketingWeek*, http://www.marketingweek.co.uk/opinion/ritson/amazon-has-seen-the-future-of-predictability/4009154.article, consulté le 8 Mai 2016
- Sammut, G., & Sartawi, M., (2012), « Perspective-taking and the attribution of ignorance », *Journal for the Theory of Social Behaviour*, 42(2), 181–200

Création de valeur par les données massives

Sylvain WILLART
PhD, Risk analyst, Advanzia Bank ; Maître de Conférences
Université de Lille

Dominique CRIÉ
Professeur des Universités, Université de Lille,
Institut d'administration des entreprises,
Lille Economie Management (UMR CNRS 9221)

L'information est source de pouvoir. La décennie qui suivra sera celle de l'émergence ou du renforcement de géants informationnels marquant le Web de leur empreinte et de leur pouvoir. Ces entreprises ont ouvert un chemin que beaucoup d'autres se pressent aujourd'hui d'emprunter, celui de l'exploitation des données massives, ou Big Data. Cependant, la mise en place d'un projet orienté données n'est pas nécessairement synonyme de retour sur investissement. Il peut être intéressant de s'interroger sur les leviers de création de valeur permis par les Big Data afin de maximiser la potentialité d'un retour sur investissement (ROI) positif.

Un renouvellement de la production et des usages des données

La production des Big Data est largement assurée par la « multitude » et ce de façon plus ou moins consciente. Les données de clics (*clickstream*) sont générées par les internautes presqu'à leur corps défendant, celles des réseaux sociaux relèvent d'une démarche plus pro-active mais sans doute non conscientisée ; certaines enfin représentent un réel investissement collaboratif dans une mission précise (OpenStreetMap). Mais la production des Big Data a également une dimension automatisée du fait de la multiplication des capteurs (GPS, RFID…) et des objets connectés (*Internet of Things*). Ces deux éléments se conjuguent avec le phénomène de « *datafication* » qui renvoie à cette propension toujours plus grande dans toutes les sphères de la société de mesurer tout ce qui peut l'être et de tenter de conserver une trace chiffrée de tous les évènements. Les données sont ainsi de plus en plus souvent récoltées en continu et concernent tout type d'activité.

La question de l'usage de ces données n'intervient souvent qu'après la récolte ! Et c'est là le second aspect important: la plupart de la valeur extraite des données vient d'usages secondaires, non envisagés au moment de leur collecte. Ces usages ex-post sont également contingents du croisement de plusieurs sources de données. Dans cette optique, le développement de l'open-data a une importance capitale : en mettant à disposition de nombreuses données, les possibilités de croisements et donc de découverte d'informations augmentent exponentiellement. Mais ces croisements et usages secondaires ne sont pas sans problème, notamment lorsqu'il s'agit de données à caractère personnel, qui ne peuvent normalement être recueillies sans un consentement relatif à un usage précisé et donc prédéfini.

Des données à la valeur ou de la valeur des données

Si la récolte de données ne nécessite pas de but prédéfini, il faut souligner toutefois que la valorisation requiert en revanche un travail important de raffinage et d'extraction. L'objectif est alors de transformer la donnée en information, laquelle pourra produire de la valeur si délivrée au bon moment dans un contexte adapté. La valorisation des données peut s'entendre de façon duale : valeur d'usage et valeur d'échange. La première apparaît lorsqu'un service est construit à partir des données, la valeur de ce service pour l'utilisateur peut alors être corrélée à la quantité des données et à la qualité des informations qu'il intègre. De tels services peuvent être destinés à des consommateurs (suivi de consommation, points de fidélité récoltés, recherche d'une offre optimale…) ou à des entreprises (aide à la décision, tableau de bord, tendances, prédictions…). Quant à la valeur d'échange, le traitement préalable des données est moins crucial. En revanche, l'échange nécessite la structuration d'un marché autour de vendeurs et d'acheteurs. Il est aujourd'hui largement oligopolistique.

Des contraintes techniques et méthodologiques nouvelles

Le traitement des Big Data dans l'optique de l'extraction d'une valeur d'usage suppose un changement radical des pratiques en termes d'équipement, de logiciels, et de méthodes d'analyse.

Au plan de l'équipement, « l'informatique dans les nuages » (Cloud computing) s'impose : la capacité de stockage et la puissance de calcul n'ont plus vocation à être hébergés au sein de l'entreprise mais consommés auprès d'un prestataire, variabilisant les coûts fixes de l'informatique.

Le stockage des données dans le cloud requiert l'utilisation d'outils spécifiques, capables de les traiter à partir de différents serveurs et d'effectuer des calculs en utilisant les nombreux processeurs de ces serveurs. Ce « parallélisme[1] » informatique nécessite des solutions logicielles capables de traiter le stockage distribué (e.g. Hadoop), d'effectuer de façon performante des tris sur les données (e.g. MapReduce) et de réaliser sur ces données des analyses statistiques (clustering, classification, moteur de recommandation e.g. Mahout).

Une dernière contrainte tient à la nature des données intégrant de plus en plus souvent des dimensions spatiales, temporelles, et de réseau. Les liens qui en découlent entre les observations (plus ou moins proches dans le temps, dans l'espace, et dans leurs relations sociales) complexifient les analyses notamment via la forme particulière de la matrice de variance-covariance : la mise en œuvre pratique est encore un défi en termes de puissance de calcul et de rapidité d'obtention des résultats.

Le processus de production, les nouveaux usages potentiels, les défis de traitement des données ou l'extraction des informations qu'elles contiennent peuvent modifier en profondeur la façon dont les entreprises créent une valeur pour elles-mêmes et leurs clients. De nouveaux rapports de pouvoir s'instaurent sur un marché qui se structure pour le moment autour de quelques majors. En conséquence, il est crucial de s'interroger sur la, ou les, valeur(s) que les données peuvent apporter à l'entreprise, ainsi que sur les conditions sous lesquelles cette valeur peut être capturée.

1. Utilisation de plusieurs ordinateurs/serveurs en même temps, en « parallèle », de façon « distribuée ».

Une approche économique et comptable de la valeur des données

En démultipliant les pistes d'exploitation des données, la révolution des Big Data positionne l'information comme une ressource stratégique. Mais elle soulève également la question de la valorisation de cette ressource.

Dans une certaine mesure, les données peuvent être comptabilisées comme des actifs immatériels, ce qui pose la question de leur valorisation dans le bilan de l'entreprise. De plus, les données peuvent être considérées comme un bien public, ou plus précisément collectif (non-rival et exclusif), dont la valeur ne s'épuise pas avec l'usage. En fait, la valeur des données ne s'érode qu'avec le temps, parfois très rapidement, notamment pour les moteurs de recommandation ou la publicité ciblée. La valorisation comptable se heurte donc à un double problème : actif immatériel et à amortir dans un temps court mais non dépendant des usages. Les partisans de l'orientation Big Data soulignent que la valeur des données réside dans l'optimisation des décisions de l'entreprise. Sur un échantillon de 179 entreprises cotées, Brynjolfsson et al. (2011) montrent que les entreprises engagées dans une démarche Big Data augmentent leurs ratios de productivité de 5 à 6% en moyenne.

La valorisation boursière de certaines entreprises valide cette thèse : Facebook a été introduit en bourse le 18 mai 2012 pour une capitalisation de 104 milliards de dollars alors que son total bilan de 2011 s'élevait à 6,3 milliards. Nonobstant les cash-flows espérés, ces chiffres valorisent les données de l'entreprise à 97,7 milliards, soit 102,3 dollars pour chacun de ses 955 millions de comptes actifs à la date de son introduction.

Les données ont, depuis longtemps, également une valeur économique d'échange. La révolution des Big Data tend à structurer et massifier ce marché en y incluant de nouvelles sources de données et en multipliant le nombre d'acteurs. Les prix constatés y sont extrêmement variables en fonction de la précision des données (tableau 1).

Tableau 1 : exemples de prix pour différents types de données sur le marché

Entreprise	Type de donnée	Valorisation
Axciom	Adresses email (accompagnées éventuellement d'un profil)	**2 à 5 cents par contact**
Plateformes RTB et AdExchange	Profils de navigation (sans identification)	**0 à 1 dollar pour 1000 affichages en Europe**
Facebook	Valorisation boursière d'un profil	**102,3 dollars**
Federico Zannier	Données de navigation, localisation du pointeur de souris, GPS, webcam, fichiers log[2]	**2 dollars pour un jour, 5 pour une semaine**

2. « A bit(e) of Me »: <http://www.kickstarter.com/projects/1461902402/a-bit-e-of-me>, accédé le 05 Mai 2016

Entreprise	Type de donnée	Valorisation
Datasift	Tweets et analyses (tarif fonction de la charge de calcul; prix du Data Processing Unit : 20 cents par heure)	**10 cents pour 1000 tweets plus un coût de traitement exprimé en DPU**
Datacoup	Données de profil agrégées (navigation, réseaux sociaux, transactions de carte bancaire) type panel	**8 dollars par mois offert à chaque membre du panel**
Appannie.com Intelligence	Données de téléchargement d'applications mobiles et ebooks par plateformes et par pays	**Gratuit pour l'éditeur d'une application. 15000 dollars par an pour un accès à l'ensemble des données.**
Datamarket.com	Données mises en vente par les entreprises qui les détiennent ou les créent	**Prix fixé par le vendeur**

La valorisation des données par le marché souffre aujourd'hui d'un certain manque de recul. Les exemples sont trop peu nombreux ou biaisés par le fait que les vendeurs proposent des produits trop spécifiques et peu comparables. Il est donc difficile d'extrapoler une tendance, un prix, ou même le contour exact du produit vendu. Ces données sont également rarement vendues par les personnes auxquelles elles se rapportent (les exemples de Facebook, mais également de DataSift, Appannie, ou du Real-Time Bidding concernent en effet des acteurs intermédiaires qui récoltent des données de tiers, les analysent, les agrègent, et les revendent sous forme d'information – profils de navigation par exemple).

En cas de vente de données personnelles par les personnes concernées, la valeur accordée à ces données diffère entre le déclaratif et le comportemental et elles sont souvent échangées contre la gratuité de certains services. On pourrait en conclure que, du point de vue du consommateur, ses données ont moins de valeur que le service qu'il reçoit en retour (dans le cas contraire, il refuserait le service). Mais il existe un biais cognitif réduisant l'importance perçue des conséquences de long terme (sollicitations commerciales) au profit des bénéfices de court terme (obtention immédiate du service). Enfin pour ne pas dévoiler ses informations, l'internaute peut également recourir à différents stratagèmes dont le mensonge. Une nouvelle option émerge pour la monétisation des données personnelles par l'intermédiaire de plateformes dédiées (coffres-forts numériques), les internautes pourront stocker et louer ou vendre leurs informations personnelles (e.g. enliken.com).

La valeur d'usage des données

Si les résultats de Brynjolfsson et al. (2011) valident l'importance de l'orientation données, ils précisent mal comment et à quels niveaux de l'entreprise les données peuvent délivrer leur valeur d'usage. Pourtant, l'ensemble des fonctions de l'entreprise semble devoir être impacté par les Big Data. Une analyse en termes de chaîne de valeur peut clarifier comment cette révolution se décline dans l'ensemble de l'entreprise, pour ensuite soulever la question de l'impact sur la prise de décision stratégique de l'usage des données et des informations qu'elles peuvent receler.

Certaines activités sont déjà largement impactées par les données. La logistique par exemple

a intégré le suivi des livraisons en temps réel par l'application de puces RFID sur les colis. Le marketing en ligne a profité de l'analyse des données de clic issues des fichiers LOG des sites internet pour développer des mesures d'audience ou les techniques de recommandation, de retargeting (publicité comportementale), et de *Real Time Bidding* (enchères en temps réel d'espaces publicitaires).

Pour la logistique comme pour le marketing, la datafication a permis l'amélioration des services existants (meilleur suivi des livraisons, optimisation du design des sites, personnalisation des contenus publicitaires), mais également une accélération du cycle de décision. La fonction logistique en intégrant le suivi des ventes à celui des stocks gagne en réactivité et automatise les réapprovisionnements. Certaines entreprises spécialisées utilisent leurs masses de données pour calculer avec précision le coût de chaque livraison et ajuster automatiquement leurs prix.

Au-delà de l'amélioration des solutions existantes, les produits vendus peuvent être assortis de divers capteurs permettant la récolte des données d'usage des consommateurs. Ces données peuvent ensuite être utilisées pour améliorer le produit, ou proposer un service après- vente personnalisé. L'objectif ici étant d'automatiser la récolte des retours d'expérience clients en prenant en compte l'ensemble de ces derniers et non seulement ceux qui prennent le temps de répondre aux sollicitations des entreprises ou de contacter le service après-vente.

Le stockage, la sécurisation, et l'analyse de ces données liées aux objets et/ou aux individus peut également représenter un marché important qu'investissent déjà les fournisseurs de « coffres-forts numériques » (adminium.fr, personal.com…).

Limites et risques de l'orientation données

La stratégie orientée données comporte également des risques pour les entreprises notamment ceux liés à l'utilisation de mauvaises données (en termes de qualité ou de nature) et aux mauvaises utilisations des données.

Le contrôle de la nature des données revient quant à lui à s'assurer que le type de données utilisé est le mieux adapté pour répondre à la question posée. Sur ce point, on peut citer notamment le point de vue hétérodoxe de la société CtrlShift (2011) qui soulève la question de savoir si les Big Data sont effectivement les mieux à même de servir au pilotage de la relation client. L'idée ici est que ces données, trop massives, génèrent des segmentations trop floues qui, du point de vue du consommateur final, sont toujours mal adaptées à ses intentions d'achats et provoquent un sentiment d'intrusion. A l'inverse, l'utilisation de « *small data* », émises directement par les consommateurs via leur coffre-fort numérique, pourrait permettre de mieux répondre à cette attente. Au plan statistique, il s'agit notamment d'une amélioration du rapport signal/bruit : plutôt que d'essayer de récolter un grand nombre d'informations, on se concentre sur les sources les plus pertinentes i.e. les consommateurs eux-mêmes. Cela étant, on peut souligner que ces SmallData, si elles ne sont plus aussi volumineuses, conservent les caractéristiques de variété et de vélocité de leurs grandes sœurs.

Pour conclure….

D'une part les données vont continuer de croître en volume, en diversité de sources, en rapidité de captation et de traitement, et en utilisations. Les entreprises mal préparées à ce déluge de données risquent de rencontrer des difficultés d'adaptation. D'autre part, et paradoxalement, les données disponibles risquent de ne jamais être complètes, la complexité des phénomènes humains dépassant, certainement encore pour longtemps, la capacité de mesure et d'analyse. Si la quantité d'information et la capacité de calcul peuvent repousser les frontières de la

rationalité limitée, elles ne peuvent la transformer en rationalité parfaite.

L'adaptation des entreprises est nécessaire afin de profiter des leviers de création de valeur que permettent les stratégies orientées données. De l'adaptation des produits et services existants à l'investissement de nouveaux marchés, de nombreuses opportunités restent ouvertes.

Référence

- Brynjolfsson, E., Hitt, L. et Kim, H., "Strength in Numbers: How does data-driven decision-making affect firm performance?", (6 décembre 2011), International Conference on Information Systems

Big Data, algorithmes et marketing : rendre des comptes

Christophe BENAVENT
Professeur à l'Université Paris Nanterre

Cet article s'intéresse à la question de la mise en œuvre à vaste échelle d'algorithmes utiles au marketing, et s'intégrant dans une logique de plateforme. Prenant en compte des observations répétées d'externalités négatives produites par les algorithmes : ségrégation, biais de sélection, polarisation, hétérogénéisation, mais aussi leurs faiblesses intrinsèques résultant de la dette technique, de dépendances des données, et du contexte adversial dans lequel ils s'exercent, nous aboutissons à la nécessité d'une redevabilité algorithmique et nous nous questionnons sur la manière dont les algorithmes doivent être gouvernés.

Le Big Data pour le marketing n'est pas simplement un élargissement du spectre des méthodes d'étude, ni le déploiement à vaste échelle du recueil des données et l'exploitation d'une grande variété de formats d'information. C'est une intégration de l'information à la décision si intime que de la mesure à l'action il n'y a quasiment plus d'espace donné à la délibération.

Ce qui est désormais géré est un flux de décisions continu et contenu par l'architecture des algorithmes. C'est une transformation de la nature même du cycle de décision marketing (Salerno et al, 2013) qui pose un problème de nature éthique et politique : que se passe-t-il dans les boîtes noires de la société (Pasquale, 2015) et particulièrement celles des dispositifs marketing ? Comment en rendre compte ?

Deux événements en témoignent. Le cas de Volkswagen et de la tromperie algorithmique qui permettait d'échapper aux tests de pollution, ne trouble pas tant par le constat de la volonté stratégique de tricher et de cacher, que par l'hypothèse qu'il pourrait être un patch apporté et oublié par des ingénieurs incapables de résoudre la contrainte des législations environnementales. Le cas de Tay, l'agent conversationnel de Microsoft perverti par des activistes d'extrême droite qui l'ont entraîné à formuler un discours raciste, sexiste et suprémaciste, illustre une faille possible des systèmes auto-apprenants : ils sont dépendants des données.

Le premier cas relève certainement de ce que les informaticiens appellent la dette technique. Celui de Tay relève de l'informatique « adversiale ». Ces deux notions clés permettent d'éclairer ce qui fait la faiblesse des algorithmes. Cette faiblesse peut venir aussi de la nature du recueil de données, massif et intrusif, qui conduit à une réaction stratégique des sujets dont le mensonge est une des formes. Il est d'ailleurs une préoccupation importante des spécialistes de management des systèmes d'information comme en témoigne le travail de Joey George (2008). Le mensonge n'est pas une nouveauté, ni l'omission ; si la stratégie de leurre comme le

promeuvent Brunton et Nissembaum (2015) sert à protéger la vie privée, et c'est bien pour cela que les pièces comptables ont été inventées, les comptes doivent être prouvés. Les Sumériens ont inventé la comptabilité en même temps que les tribunaux. Pas de royaume sans scribe scrupuleux, ni recensement, ni mémoire ni registre[1]. La vitesse de calcul et son échelle sont nouvelles, pas le principe de compter ni celui de d'assurer l'intégrité des comptes.

Dans un monde de plateformes où les données et les algorithmes façonnent les services délivrés et les conditions de leur livraison, le regard critique sur le Big Data ne doit pas se tenir au registre de la dénonciation, il doit envisager les processus par lesquelles les algorithmes produisent des effets qui diffèrent de ce qui est attendu et des externalités négatives dans l'environnement où ils agissent. Il conduit à faire de l'idée de redevabilité algorithmique un concept central de la mise en œuvre des méthodes de Big Data

Gouvernementalité algorithmique : une nouvelle politique marketing

L'idée que les algorithmes ne sont pas neutres et agissent sur nos conduites se développe depuis plusieurs années. Lessig (2002) sans doute est un des premiers à voir et à vouloir mettre de la politique dans les algorithmes avec sa formule « code is law ». C'est cette idée qui est au centre du travail de Rouvroy et Berns (2013) ou de celui de Dominique Cardon (2015) qui en propose une sociologie perspectiviste.

L'idée de gouvernementalité, terme proposé par Michel Foucault[2], se définit comme l'action qui agit sur les conduites individuelles par des dispositifs de connaissance et de contrainte, pour gouverner la population et la ressource qu'elle représente. L'originalité de cette conception réside certainement en ce qu'elle dissocie l'objet du gouvernement, la population et son lieu d'exercice, la psychologie des individus. Dans un univers de plateformes qui est celui des moteurs de recherche, des marketplaces, des réseaux sociaux, cette gouvernementalité peut être saisie au travers de trois grands types de dispositifs et de ce qui propage leurs actions dans la population - les algorithmes - (Benavent, 2016).

Le dispositif principal est celui qui règle la capacitation des sujets de la population et qui, par son architecture et ses interfaces, leur permet d'agir tout en définissant les limites de cette action (restriction). Il peut se différencier selon les populations, et évolue en fonction de leurs comportements et de leurs interactions. Sur cette architecture se greffent les « *policies* », ou règlements intérieurs qui régissent les droits relatifs aux données personnelles, à la propriété sur les contenus et limite la liberté d'expression. Le troisième type de dispositif a pour finalité de motiver l'action comme les mécanismes de fidélisation, les indicateurs de réputation, la conception de nudges, le design d'un système de gamification destiné à motiver l'action des utilisateurs[3]. C'est le terrain de jeux des technologies persuasives dont Fogg est un des principaux promoteurs (2002)

1. C'est la raison d'être de l'effervescence aujourd'hui des débats sur la Blockchain.
2. Michel Foucault (2004), Sécurité, territoire, population, éditions du Seuil, 2004, « Par gouvernementalité, j'entends l'ensemble constitué par les institutions, les procédures, analyses et réflexions, les calculs et les tactiques qui permettent d'exercer cette forme bien spécifique, quoique très complexe de pouvoir qui a pour cible principale la population, pour forme majeure de savoir l'économie politique, pour instrument essentiel les dispositifs de sécurité » pp.111-112.
3. Les nudges sont des dispositifs qui prennent avantage des biais cognitifs pour orienter, sans imposer, la réponse des sujets dans une direction favorable au bien-être du sujet et de l'environnement. La gamification est cette discipline nouvelle née dans l'industrie des jeux vidéo, qui en utilisant les notes, les statuts, des scores de performance « ludifie» les activités apportant une gratification à ce dont l'utilité n'est pas pleinement perçue ou qui réclame un effort trop élevé, maintenant ainsi un niveau élevé de motivation.

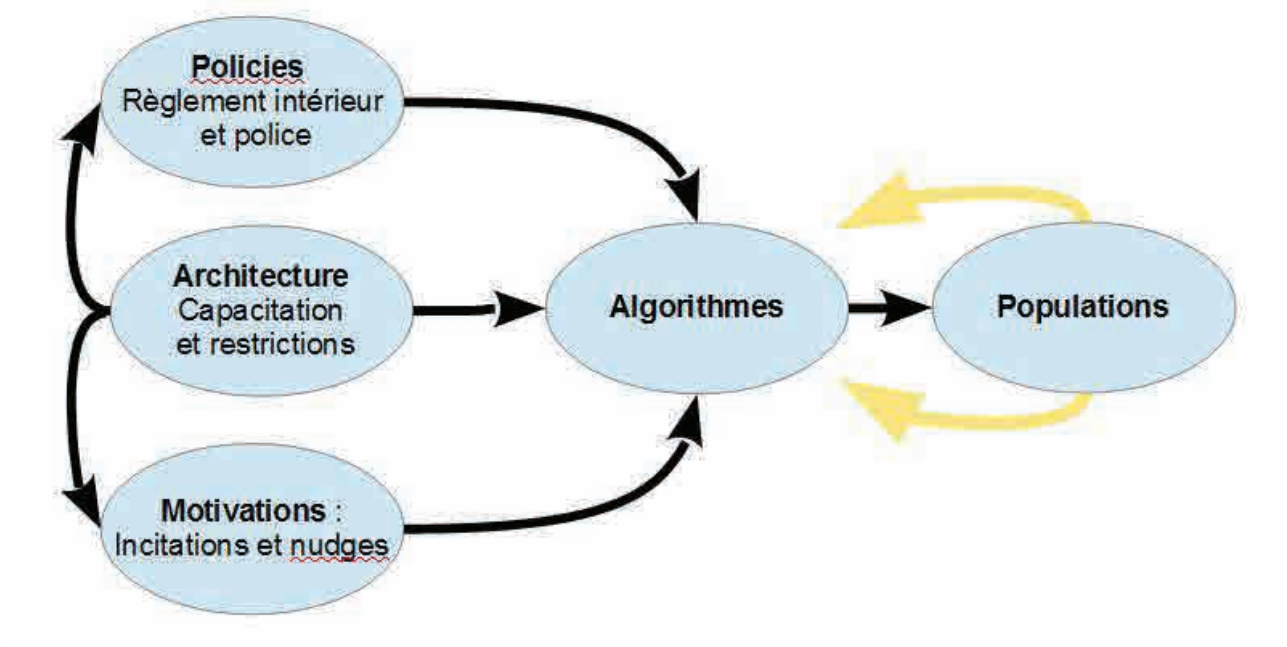


Figure 1 : les composantes de la gouvernementalité algorithmique

Les algorithmes ont un rôle d'abord évidemment de calcul dans l'instanciation du système complexe d'actions, de règles et de motivations, pour des situations particulières. Le système technique sur lequel ils s'appuient se caractérise par son volume (Big Data) et par sa granularité : l'échelle est celle de la minute et de quelques mètres. Ils peuvent être simples (un tri par ordre de prix) ou sophistiqués (le calcul d'un indice de sentiment à partir d'une méthode de machine-learning appliquée à des données textuelles). Mais plutôt que de mettre l'accent sur le calcul, c'est le rôle médiateur, la synchronicité du système qui en font un media équivoque.

Si on les examine sur un plan plus concret, comment les algorithmes exercent-ils leurs effets ? On s'aperçoit que pour le marketing, ils se concentrent sur un petit nombre de tâches. Ce à quoi servent les algorithmes s'inscrit dans la liste suivante :

- Filtrer et chercher : c'est bien le sens du Page Rank, mais aussi de très nombreuses méthodes, qui s'appuient sur l'historique des demandes dont on a montré depuis longtemps que sous-certaines conditions elles pouvaient présenter des résultats peu pertinents.
- Trier les fils de nouvelles : c'est l'outil privilégié par Facebook et qui fait l'objet d'une polémique récente à cause de l'obscurité de ses choix implicites.
- Recommander des profils, des produits : c'est le domaine d'application le plus important et un des plus anciens ; Amazon et Netflix en sont les porte-drapeaux. La question de la sérendipité en est un enjeu essentiel pour aller au-delà de recommandations trop évidentes.
- Indiquer des tendances et prédire des évolutions : une des armes de Twitter pour mettre en avant le contenu (trending topics). De manière plus générale c'est la prédiction des séries chronologiques : ventes, audience, à un niveau de plus en plus micro-économique.
- Calculer des scores de réputation, de risque, de qualité à des échelles jamais connues comme l'estimation des valeurs des maisons de Zillow (*Zestimate*), qui s'appuie sur un recensement de 185 millions de logements avec des milliers de caractéristiques.
- Catégoriser des images : le cas de Flickr, une plateforme de photos, est un très bon exemple d'application de deep learning pour taguer les images et ainsi améliorer leur « recherchabilité » et donc leur valorisation sur la marketplace.
- Produire et déclencher des alertes et notifications. C'est un enjeu du monde des objets connectés tels que les trackers, les balances connectées, les compteurs d'énergie, les

piluliers digitaux et la plupart des « applis » de nos smartphones.
Pour examiner avec plus de précision cette notion générale nous examinerons d'abord les effets sociaux indésirables que les algorithmes produisent : des effets de ségrégation et de biais de sélection, des effets de performativité, et au travers de certaines tentatives de correction, un accroissement potentiel de perte d'éléments de vie privée.

Ségrégation, polarisation et biais de sélection

Depuis quelques d'années de nombreux universitaires (les politiques sont bien moins nombreux) s'inquiètent du caractère obscur et ésotérique des algorithmes dont le fonctionnement peut même, à l'insu de leur concepteurs, produire des effets non-anticipés sur les populations. Tarleton Gillespie and Nick Seaver en maintiennent une bibliographie très dense sur *Social Media Collective*, un blog de recherche de *Microsoft*[4].

Un bon exemple de ce type d'effet est donné par l'étude de Eldelman et Luca (2014) sur les prix de location *Airbnb* à New-York. Ils observent une différence de 12$ entre les logements proposés par les Blancs et les Noirs. Ceci résulte principalement d'un mécanisme de réputation : l'exposition de la photo de l'hôte. L'algorithme est simplissime, il réside simplement dans le protocole de production d'une page d'offre. Ce n'est dans ce cas pas tant l'algorithme lui-même qui ségrégue, mais en « enactant » les décisions par la mise en évidence de la couleur de peau, il donne à cette dernière la valeur d'un attribut signalant un certain risque, conduisant ceux qui en sont victimes à s'ajuster en proposant des prix significativement plus bas.

Cet effet de ségrégation peut être généralisé à la notion de polarisation qui est au cœur de l'ouvrage : « The bubble society » (Pariser, 2011) mais semble aussi être une hypothèse forte sur la vie politique américaine. Comme l'indique un rapport de Pew Research (2014) même si le rôle des réseaux n'est pas démontré, on sait au moins que l'exposition aux opinions contraires ne concerne qu'un tiers des internautes. Il reste aussi à faire la part de l'effet purement algorithmique et de l'auto-sélection des individus (Bakshy et al,2015). Les réseaux sociaux amplifient-ils les liens homophiles en favorisant des phénomènes d'attachement préférentiels, ou ne sont-ils que le miroir de comportements soumis aux biais de confirmation ?

Les effets de biais de sélection peuvent s'observer sur l'agrégation des informations dans les réseaux sociaux. Un exemple simple illustre le phénomène : supposons que les individus qui postent le plus souvent, postent du contenu positif tandis que ceux qui postent moins souvent sont animés par la vengeance et seront donc négatifs. Supposons que ces deux groupes soient d'effectifs égaux et que les premiers postent 10 fois plus, la production collective de commentaires est composée à 80% de contenus positifs alors que la population qui les émet ne représente que 50%.

L'effet algorithmique est dans ce cas très simple : il résulte simplement d'une opération d'agrégation, mais il peut être volontairement amplifié comme l'a tenté Facebook dans une expérience contestée visant à filtrer les messages sur le fil de nouvelles des individus, en ôtant aléatoirement de 5% à 90% des contenus positifs et négatifs dans une population de 670 000 personnes. Le défaut éthique de l'expérience est que les sujets n'ont pas été informés conduisant les éditeurs à faire précéder la publication d'un avertissement. (Adam et Al, 2014). Le résultat de l'expérience est de démontrer que l'on peut propager l'émotion sociale en manipulant la composition du fil de nouvelles. Filtrer en renforçant les contenus positifs conduit les sujets à produire plus de posts positifs aussi.

4. Accessible sur https://socialmediacollective.org/reading-lists/critical-algorithm-studies/

Cet effet de sélection est amplifié par les retours d'information que constituent les notes et commentaires désormais employés de manière systématique. Après tout, chiffres et graphiques, alertes sonores ou ranking…tout élément qui indique une performance peut agir au moins par cette vieille idée des prophéties auto-réalisatrices. Il s'agit ici de la notion de performativité qui fait l'objet d'un regain d'intérêt dans les sciences de gestion (Berkowitz et Dumez, 2014). Elle propose qu'un acte de communication est aussi une action, le langage ne s'épuise pas en passant l'information de l'un à autre, il est aussi action sur ceux qui le reçoivent. Dans le domaine des études marketing, ce qui était réservé aux décideurs est désormais distribué à la foule, qui réagit en en connaissant le résultat. C'est ainsi que les systèmes de notation plutôt que de mesurer la satisfaction de manière fine et valide, présentent des distributions biaisées à droite (vers la satisfaction) et une variance faible : les usagers connaissant l'effet négatif sur la partie qu'ils évaluent (d'autant plus quant ils sont eux-mêmes évalués), préfèrent s'en tenir à une sorte de note de politesse.

Les développements récents en matière de Deep Learning font apparaître ce problème sous un visage différent. Il s'agit d'architectures de réseaux de neurones à plusieurs couches qui font l'objet d'un premier apprentissage non supervisé dont Yann Le Cun avec quelques autres chercheurs (Bengio et al. 2007) a relancé l'intérêt dans les années 2006-2007, et qui trouvent actuellement des terrains d'application importants dans le domaine de la reconnaissance d'objet dans les images, dans la reconnaissance vocale ou l'annotation de vidéos. Ils s'appuient sur des millions d'exemples qu'ils modélisent dans des espaces de plusieurs centaines de milliers, voire de millions de paramètres.

Le problème vient moins du calcul que de la manière dont sont présentés les objets dont on souhaite reconnaître des éléments de forme et pouvoir les associer à des catégories prédéterminées. Les catégories sont celles choisies pour l'algorithme par des humains. *Flickr* a entrepris de taguer automatiquement avec une centaine de catégories son stock considérable de 11 milliards d'images que les dépositaires répugnent à documenter par des mots clés. On choisit donc d'abord ce qu'il faut reconnaître[5]. Les ingénieurs de *Flickr* ont fait des choix, qui dans un second temps favoriseront certaines images plutôt que d'autre. S'ils définissent 12 éléments pour une catégorie « architecture », et seulement 4 pour décrire la catégorie « animaux », plus de détails sur l'une donne plus de chances à l'image d'être retrouvée, et ces choix sont redéfinis ensuite dans un processus d'apprentissage dont on ignore le protocole.

On retrouve ici une critique traditionnelle de la sociologie qui considère que les catégories produites indépendamment de l'effort théorique peuvent conduire à des contresens ; et l'on pourra reprendre par exemple l'analyse de Dominique Merllié (in Vatin, 2009) sur la discordance entre les déclarations de relations sexuelles des hommes et des femmes (1,57 contre 1,11 relations) où l'analyse élimine différentes explications (biais de désirabilité, exclusion des prostituées, effets de distributions) jusqu'à considérer le fait que par « relations sexuelles » hommes et femmes n'entendent pas la même chose. La constitution des catégories ne répond pas au critère de conventions qui s'établissent par la délibération. Dans le fil des travaux de Derosières et Thévenot (1988), il est désormais accepté que la statistique n'organise qu'un rapport de correspondance avec le réel, que les catégories qu'elle emploie sont le résultat de négociations, de conflits, de débats qui s'objectivent dans le consensus de la convention (Vatin, Caillé, Favereau, 2010). Sa validité tient à un accord sur ce qu'est la réalité.

Le problème dans la société des algorithmes est qu'ils ne font pas l'objet d'un tel débat. Ce monde proliférant de statistiques, est un monde sans convention ni accord, ou du moins seulement avec des accords partiels et peut-être partiaux.

5. Ceci est poussé au paroxysme avec les algorithmes « psychotiques » de l'équipe de Google Deep Dream qui les a entraîné à reconnaître des formes imaginaires dans des images ordinaires, générant des tableaux à la Jérôme Bosch.

Un algorithme juste

L'ignorance de ce type de phénomène peut conduire à des problèmes sociaux importants notamment lorsqu'on soupçonne que les algorithmes (même simples) produisent des effets de discrimination. Le cadre légal aux Etats unis a incité des chercheurs à s'intéresser à des méthodes de *fair scoring* qui tentent d'effacer l'aspect discriminant des algorithmes en minimisant le coût que représente la perte de précision L'algorithme juste n'est pas seulement celui qui prédit précisément, c'est celui qui produit un résultat socialement acceptable.

Prenons le cas des algorithmes de scoring courants dans le domaine bancaire pour attribuer ou non un crédit à la consommation. Ces algorithmes s'appuient sur ce que les spécialistes du machine learning appellent un classificateur, autrement dit une équation dont la forme très générale est la suivante : $S=f(X,\Theta)$

S, le risque, est le score calculé, (c'est mieux quand il s'exprime sous la forme d'une probabilité, celle de ne pas rembourser le prêt), en fonction d'un vecteur X de caractéristiques qui décrivent ce que l'on sait sur l'individu : âge, revenu, amis sur facebook, historique des mouvements bancaires, et peut être des données de santé. Θ désigne les paramètres du modèle. Ces modèles peuvent prendre une large variété de formes : modèle de régression, arbre de décision et random forest, SVM (Support Vector Machine), analyse discriminante, réseaux de neurones. Un tel modèle fournit, au mieux, une probabilité que le risque advienne.

En réalité il est mis en œuvre au travers d'une structure de décision. Elle peut être primitive quand on indique un seuil, par exemple : « si S est supérieur à 3% alors ne pas prêter ». Elle peut être un peu plus sophistiquée en pondérant gains et pertes espérées. Par exemple si G est la marge gagnée sur le prêt dans le cas où il n'y a pas d'incident, et P la perte subie si le client n'est pas en mesure de rembourser, le critère devient « $P*S+G*(1-S)>0$ ». On peut imaginer plus compliqué.

Le problème posé est qu'un tel algorithme n'est pas forcément juste, au sens de la précision. On connaît le problème classique des faux positifs. La théorie de la décision fournit des moyens de réduire, du point de l'entreprise, cet impact et de mieux définir sa stratégie (minimiser les risques, optimiser le gain...) ; mais du point de vue des individus qui, bien qu'en droit d'obtenir le prêt, ne l'auront pas, il y a injustice.

Il est assez facile de mesurer l'importance de ce risque simplement au moment où l'on teste le modèle. Il suffit de comparer les résultats à la réalité. On pourrait parfaitement exiger de ceux qui emploient de tels algorithmes, sans donner les paramètres Θ (qui sont un secret de fabrication), de rendre compte de la précision de leur algorithme ; et donc du risque de produire des décisions injustes. La plupart des modèles ne prédisent qu'assez mal les résultats, sauf dans les cas triviaux. Il faut garder en tête que même si la performance est remarquable, elle est loin d'être parfaite. Des jeux de données tests permettent d'en réaliser l'importance. Un exemple pour le machine learning est le jeu de données *Minist*, pour la reconnaissance de caractères ou le jeu de données CIFAR 100 pour la catégorisation des images. Il serait utile d'en disposer d'équivalents pour les applications marketing, auxquels les algorithmes commerciaux devraient se confronter ; les résultats de ces tests devraient être publiés.

Allons plus loin avec Dwork et al (2011). Un autre critère de justice est introduit, partant de la condition de Lipschitz qu'on peut exprimer assez simplement : un algorithme sera juste si la distance entre deux individus (telle qu'on la mesure au travers des caractéristiques X) est plus grande que la distance entre les scores calculés à partir de ces profils. C'est un critère de parité statistique à partir duquel les chercheurs proposent des modèles qui permettent de gommer les effets de variables discriminatoires. Cependant même si l'on introduit des approches justes

(fair) cela a un coût qui est celui de la confidentialité. Pour s'assurer que l'algorithme soit juste, il faut selon l'auteur que l'on détienne des données "sensibles", relevant de l'intimité. Un algorithme pour être juste devrait ainsi violer la vie privée. Ce qui amène un commentateur à relever qu'il n'est pas possible de construire un algorithme intrinsèquement juste, car même s'il est exact c'est au prix d'une perte de confidentialité. C'est à l'éthique de trancher : violer l'intimité ou éviter les discriminations.

Dette technique, dépendance des données et le risque adversial

Les effets inattendus des algorithmes sur les populations, le caractère peu maîtrisé des catégories qu'ils emploient, les difficultés de méthodes pour construire des algorithmes justes qui ne renforcent pas les différences qu'ils produisent ne sont pas les seuls problèmes, et s'ils résultent de l'interaction entre la technique et le social, d'autres sont propres à la technique. Trois notions proposées par les informaticiens sont utiles pour penser ces difficultés.

La première est la « dette technique » qui traduit qu'avec le temps les défauts des logiciels deviennent de plus en plus handicapants. Certains chercheurs, Sculley D. et al (2015), désignent ces effets inattendus par le terme de dette technique comme le fait une équipe de Google IA en décrivant par le menu les risques du machine learning : l'érosion des limites entre les composants stables du logiciel et les données dont leurs comportements dépendent ; l'intrication des paramètres et des données qui rend l'amélioration plus difficile que la mise en œuvre initiale ; les boucles de feed back cachées qui résultent de l'interaction du système et du monde ; les utilisateurs non déclarés ; la dépendance aux données et aux changements de l'environnement. Autant de problèmes qu'ils suggèrent de résoudre par des solutions graduelles et partielles. La dette technique est en quelque sorte le pendant des processus d'apprentissage qui graduellement accroissent la connaissance, c'est aussi l'accumulation des défauts, des rustines (*patch*), des restructurations incomplètes qui progressivement alourdissent le système. Un de ces principaux éléments résulte du phénomène de « dépendance aux données» qui désigne en informatique le fait qu'une instance utilise un résultat calculé précédemment, ce qui peut poser des problèmes quant à l'identification des erreurs dont il devient difficile d'isoler la source : des données inappropriées ou la structure même du modèle. Cette dépendance est forte quand on utilise des techniques à base de réseaux de neurones, et notamment leurs dernières générations qui possèdent de nombreuses couches et s'appuient sur des phases intermédiaires de filtrage. Les paramètres de ces algorithmes dépendent de données dont la production n'est pas forcément contrôlée et qui ne représentent pas tous les états naturels. L'effet se traduit par des modèles précis mais pas forcément stables dans le temps et qui évoluent au fil de l'évolution des populations.

Cette dépendance aux données est d'autant plus élevée que le « risque adversial » est important. C'est celui que fait peser l'individu qui veut déjouer les défenses et éventuellement exercer une nuisance. Les situations adverses sont celles dans lesquels le produit de l'algorithme peut être contré par un adversaire : par exemple le spammer qui veut échapper au filtre du spam. Certains résultats récents montrent ainsi qu'il est possible de modifier de manière imperceptible une image pour produire un mauvais classement/ inférence, et inversement qu'un algorithme peut apprendre à reconnaître ce qui n'est pas reconnaissable (comme l'expérience Google DeepMind l'illustre).

C'est en considérant cet aspect du problème qu'on comprend la grande différence entre le Big Data et la statistique traditionnelle telle qu'elle est pratiquée par les cabinets d'étude traditionnels. Cette dernière collecte des données sans affecter le corps social. Il y a juste des points de sonde, et elle les traite dans l'enceinte close du laboratoire de statistique, ses valeurs sont étudiées dans des salles de réunion. Dans la perspective du Big Data la collecte à grand échelle a toutes les chances de faire réagir le corps social qui répond moins sincèrement et plus stratégiquement.

Plus encore, la production et la diffusion de ses résultats étant instantanées c'est une seconde possibilité de réaction qui surgit.... et donc la nécessité de réajuster l'algorithme. L'évolution du moteur de recherche Google étudié par Dominique Cardon est un magnifique exemple. Les spécialistes du Search Engine Optimisation (SEO) ont appris à construire des univers web vides, sans autre contenu que celui qui est syndiqué, pour améliorer le référencement, conduisant Google à aménager ses critères pour en réduire le poids, mais alourdissant la conception même de l'algorithme. De ce point de vue les concepteurs et propagateurs d'algorithmes auraient intérêt à s'inspirer de la réflexion des statistiques publiques.

En revenant à la question des catégories que nous avons déjà évoquée, la perspective adversiale pose une question relative aux méthodes d'entraînement des algorithmes. Les choix extrêmes sont d'un côté celui d'un petit groupe d'experts qui définit les catégories et conclut sur le pronostic de la machine, ou, à l'opposé, le recours à la foule qui multiplie les épreuves, mais fait peser un risque d'entraînement inadéquat. Ce problème se rencontre dans la reconnaissance d'image, où l'apprentissage se fait dans le choix implicite de catégories auxquelles le réseau de neurones est confronté et tente de s'ajuster. Comment et pourquoi ainsi jugeons-nous la justesse de l'identification d'un lion dans une image ? Un félin, le symbole du lion, un lion proprement dit ? L'expert ou la foule par paresse ou superficialité risquent de confondre allégrement les lions représentés (des photographies de lions singuliers) et les représentations de lion (ceux des dessins animés). Il n'y aura pas eu de discussion pour convenir que la catégorie se définit comme « images de lion ». Les catégories ne correspondent pas tant à une réalité naturelle qu'à la réalité de nos catégories, et les machines n'élaborent pas encore les catégories même si elles peuvent apprendre à les reconnaître.

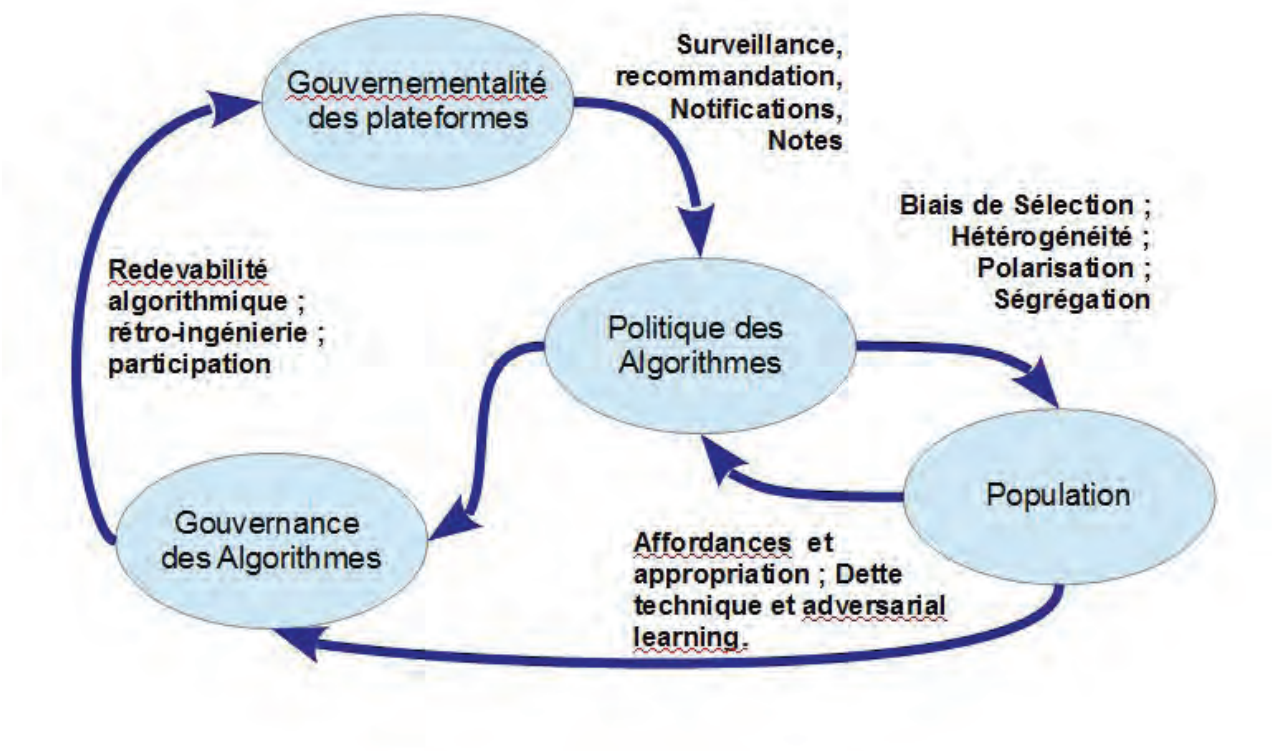
Dans la conception des algorithmes de cette espèce, ressurgit une tâche essentielle qui correspond à l'idée de validité de contenu, qui consiste à s'assurer que la catégorisation suit des protocoles particuliers qui assurent la légitimité des catégories et de la catégorisation. Ils sont probablement intermédiaires entre les experts et la foule et doivent présenter une qualité délibérative.

Les formes de la redevabilité algorithmique

L'ensemble de ces problèmes mobilise désormais les chercheurs, notamment en droit, autour de la notion d'*algorithmic accountability*[6]. Un vieux terme français « redevabilité » respecte mieux un esprit moins dominé par l'idée de l'agence et des défauts d'information inhérents, que par celle de l'intendance (théorie du *stewardship*) qui suggère que l'on peut être responsable aussi de ce que l'on ne possède pas : le bien public, le bien des autres, le royaume de Dieu... L'idée reste la même : celle d'une nécessité de rendre des comptes, de justifier des actions entreprises.

Si les algorithmes peuvent produire des effets inattendus et injustes, même s'ils ne sont pas le fruit d'une intention maligne, il est encore plus urgent de s'interroger sur la nécessité et les formes de leur obligation de rendre compte de leurs conséquences.

6. Comme l'a inauguré l' « Algorithms and Accountability Conference » tenue à la New York University le 28 février 2015.



Cette obligation se traduit aujourd'hui par la nécessité de déverrouiller les boites noires. C'est une exigence politique croissante, qui conduit à l'injonction que les plateformes doivent rendre compte des méthodes qu'elles emploient à la société dans son ensemble, en se cachant moins derrière la confidentialité des procédés.

C'est moins une question de droit que de politique, car c'est le politique qui aménage le droit. Cette exigence est d'autant plus naturelle que les algorithmes sont le plus souvent empruntés au domaine public comme le fut le page rank de Google ou la méthode de filtrage collaboratif d'Amazon.com. Mais ce n'est pas le problème principal. Par redevabilité on entend le fait que les effets de ces algorithmes doivent être considérés comme n'importe quelle externalité ; il est devenu évident aujourd'hui que dans les rapports d'activités, les effets environnementaux et sociaux de l'activité soient mentionnés.

On notera, avant d'aller plus loin, que dans la loi intitulée « Loi pour une République Numérique » (promulguée le 7 octobre 2016), cette dimension est relativement absente (sauf pour l'administration avec l'article 2).Si le droit à la vie privée semble être parfaitement reconnu et étendu (le droit de contrôle et de rectification va s'étendre à celui de la restitution des données sous une forme portable), rien par contre ne concerne ce que l'on fait de ces données, à l'exception des rapprochements autorisés ou d'éléments relatifs à la revente des données. Aucune exigence n'est formulée à l'égard des données transformées par les algorithmes, sauf un principe de loyauté des plateformes et de production d'indicateurs de transparence (titre II, section 3). Cette observation est un deuxième argument qui milite en faveur d'une obligation à rendre compte qui doit être imposée aux plateformes.

S'il faut formuler ces points d'obligation, la réflexion doit suivre leur nature. Un algorithme est une suite finie et non ambiguë d'instructions qui permettent d'obtenir un résultat ou de résoudre un problème. Il demande des entrées, il se caractérise par sa finitude, son exactitude et son rendement, et se traduit par ses sorties.

Les entrées soulèvent le premier problème. Ce problème est celui de la sincérité, de la fiabilité, de la validité et de la précision des informations qui sont introduites dans l'algorithme. Dans le monde qui est le nôtre ces entrées sont massives, entachées d'erreurs, de mensonges, d'omissions, souvent d'inexactitudes. Rendre compte de l'algorithme c'est donc très simplement préciser la qualité de ces informations entrantes et leur ôter l'évidence de leur véracité. Il s'agit aussi d'évaluer les effets de leur distribution incontrôlée sur les résultats.

Le traitement est un second problème. La finitude concerne notamment la nature des méthodes de calcul employées. À titre d'exemple dans les méthodes d'estimation d'algorithmes statistiques on sait qu'il y a des problèmes de minimum local et que, quel que soit le volume des données traitées, il n'est pas toujours assuré qu'une stabilité des paramètres soit obtenue. De manière plus sophistiquée il est indispensable que les algorithmes ne produisent pas dans leur décision des effets de faux positif ou du moins rendent compte de la balance de ces effets. Un bon exemple est celui de la lutte anti-terrorisme dont certains ont calculé les « effets secondaires » : pour identifier 3000 terroristes parmi 35 millions de personnes, un système précis (les identifiant à 99%) et faisant peu d'erreurs (accuser à tort 1% des innocents) générera près de 350 000 alertes, dont seules 2970 correspondront à de vrais « positifs ». Le rendement ici est particulièrement faible et l'injustice patente.

Quant aux sorties, il s'agit d'examiner leurs effets sociaux : compétition accrue, discrimination ou polarisation. L'obligation d'examiner ces effets pourrait s'inspirer des procédures imposées à l'industrie pharmaceutique pour limiter les effets secondaires, ou des obligations environnementales de tout un chacun.

La redevabilité algorithmique en ce sens s'approche d'une responsabilité particulière qui provient non pas d'un mandat que l'on a reçu d'un propriétaire, mais de celui qu'octroie la société dans son ensemble, celui d'un bien commun. Les algorithmes et le Big Data se nourrissent de ce qui est à la fois un commun (au sens où nulle propriété ne définit la valeur qu'ils produisent de manière agrégée) et de ce que leur mise en œuvre peut affecter l'environnement commun. Il reste à en définir les modalités, ce qu'il conviendra d'appeler gouvernance des algorithmes. Les GAFAs ne s'y trompent pas ! Conscients des problèmes de légitimité qui pourraient advenir, ils ont lancé le partenariat pour l'IA : « Partnership on Artificial Intelligence to Benefit People and Society ».

Conclusion

Les choses sont donc claires : ne comptons pas sur les machines pour réduire les problèmes sociaux, la responsabilité de ce que font les machines appartient entièrement aux humains, à leurs concepteurs et à leurs propriétaires.

Plutôt que d'interdire ou d'autoriser, la première étape de la régulation du monde des machines est d'imposer socialement l'exigence de rendre compte, ne serait-ce que par une approche d'ingénierie inversée comme le propose Diakopoulos (2014) qui vise à partir de l'étude des données d'entrée et de sortie à reconstituer le fonctionnement effectif des algorithmes, en démontant à rebours leur mécanique pour en retrouver les principes et identifier la source des effets pervers. Au-delà, la pression sociale et juridique va tendre à imposer aux gestionnaires des algorithmes la production d'études d'impacts, et dans certains cas, à engager des politiques compensatoires. C'est ainsi le cas de Airbnb en manière de discrimination qui se manifeste aujourd'hui par une invitation faite aux utilisateurs de signer une charte anti-discrimination.

La nécessité d'ouvrir les boites noires au moins partiellement s'impose. Les algorithmes doivent rendre des comptes et pas seulement sur leur efficacité. Quels biais de jugement véhiculent-ils ? Quels sont les risques de ségrégation et de discrimination ? Quelles injustices produisent-ils ?

Sur quelles conventions acceptables sont-ils conduits ?

La redevabilité algorithmique, s'impose moins comme solution à un conflit d'agents (des consommateurs qui prêtent leurs données en échange d'un service sans connaître ce qui en est fait) que comme responsabilité globale à l'égard de la société et de ses membres car, peu ou prou, les algorithmes façonnent l'univers où nous vivons. Cet univers est commun. On y démêle difficilement les contributions des uns et des autres. Dans un monde où les asymétries d'information sont majeures - certains contrôlent les algorithmes qui génèrent de la valeur et des millions d'autres n'en sont que les objets - on ne peut guère espérer que la distribution des incitations puisse renverser les choses. Les opérateurs prendront certainement soin de la légitimité des algorithmes, il n'est pas sûr qu'ils aillent plus loin. Il faut espérer au moins obtenir que les algorithmes n'œuvrent pas contre l'intérêt commun.

Pour les marketers, la principale conséquence est qu'il faudra prendre en compte dans la conception des algorithmes - ceux qui fixent des prix, ceux qui définissent un niveau d'assurance, ceux qui décident de l'octroi d'un crédit, ceux qui façonnent les environnements éditoriaux, ceux qui animent les places de marché - non seulement l'acceptation sociale de leurs technologies mais aussi les effets inattendus de leurs machines imparfaites. Ils ne devront pas ignorer la dimension politique de « leur machinerie ».

Références

- Adam et al. (2014)
- Bakshy, Eytian, Solomon Messing, Lada A. Adamic. (2015), "Exposure to ideologically diverse news and opinion on Facebook", Science. 2015 Jun 5;348(6239).
- Benavent Christophe (2016)
- Bengio Yoshua and Yann LeCun: Scaling learning algorithms towards AI, in Bottou, L. and Chapelle, O. and DeCoste, D. and Weston, J. (Eds), Large-Scale Kernel Machines, MIT Press, 2007,
- Berkowitz Héloïse, Dumez Hervé (2014) Un concept peut-il changer les choses ? Perspectives sur la performativité du management stratégique, Actes AIMS.
- Brunton, Finn et Helen Nissenbaum, Obfuscation : *A User's Guide for Privacy and Protest*, Cambridge, MIT Press, 2015,
- Cardon, Dominique (2015), *A quoi rêvent les algorithmes. Nos vies à l'heure des Big Data*, Seuil/La République des idées.
- Derosières, Alain et Laurent Thévenot (1988) « les catégories socio-professionnelles », La Découverte, Collection Repère, 1988.
- Diakopoulos, Nicholas (2014) Algorithmic-Accountability : the investigation of Black Boxes, Tow Center for Digital Journalism. June 2014.
- Dwork, Cynthia; Hardt, Moritz; Pitassi, Toniann; Reingold, Omer; Zemel, Rich (2011) « Fairness Through Awareness », arxiv, arXiv:1104.3913
- Edelman, Benjamin and Michael Luca (2014), « Digital Discrimination: The Case of Airbnb.com », Harvard Business School, Working Paper 14-054 January 10, 2014
- Fogg, B. J. (2002). Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann
- George, Joey F. and Robb, A. (2008) "Deception and Computer-Mediated Communication in Daily Life." Communication Reports 21(2), 2008, 92-103.
- Goodfellow,Ian J. , Yaroslav Bulatov ,Julian Ibarz , Sacha Arnoud, Vinay Shet(2014), Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks, (Submitted on 20 Dec 2013 (v1), last revised 14 Apr 2014 (this version, v4)), arXiv.org, arXiv:1312.6082
- Guérard, Stéphane, Ann Langley, and David Seidl 2013. "Rethinking the concept of performance in strategy research: towards a performativity perspective." M@N@Gement 16, no. 5: 566-578.
- Kramera, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock, (2014) « Experimental evidence of massive-scale emotional contagion through social networks », PNAS, vol. 111 no. 24
- Lessig (2002)
- Michel Foucault (2004), Sécurité, territoire, population, éditions du Seuil
- Pasquale, Frank (2014), *The Black Box Society*, Harvard University Press
- Pariser, E. 2011. *The Filter Bubble*. Penguin.
- Pew Research, (2014) « Political Polarization in the American Public » Pew Research, June 2014
- Rouvroy, Antoinette et Berns Thomas, (2013) « Gouvernementalité algorithmique et perspectives d'émancipation » Le disparate comme condition d'individuation par la relation ?, Réseaux, 2013/1 n° 177, p. 163-196.
- Salerno, Francis, Christophe Benavent, Pierre Volle, Delphine Manceau, Jean-François Trinquecoste, Eric Vernette et Elisabeth Tissier-Desbordes (2012), Eclairages sur le marketing de demain : prises de décisions, efficacité et légitimité, Décisions Marketing, Oct dec, n°72
- Sculley D. , Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young (2015), "Machine Learning: -The High-Interest Credit Card of Technical Debt", SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)
- Vatin François, Caillé Alain, Favereau Olivier, (2010) « Réflexions croisées sur la mesure et l'incertitude », Revue du MAUSS 1/2010 (n° 35) , p. 83-10
- Vatin, François (2009), Evaluer et Valoriser, Une sociologie économique de la mesure, Presse Universitaire du Mirail

Les calculs sur données-client massives sont-ils trop importants pour être laissés aux analystes marketing ?

Michel CALCIU
Maître de conférences, Université Lille 1, IAE, RIME Lab EA 7396

Francis SALERNO
Professeur des Universités, Université Lille 1,
IAE, Lille Economie Management (UMR CNRS 9221)

Jean-Louis MOULINS
Professeur des Universités, Cret-Log, Aix Marseille Université

Introduction

La loi de Moore et les limites du microcosme CPU poussent à explorer le macrocosme CPU. Après de nombreuses années d'augmentation de puissance de calcul grâce à la miniaturisation des circuits, les progrès à rythme exponentiel du microcosme CPU (Unité à Processeur Central), connus sous le nom de loi de More, semblent atteindre un goulot d'étranglement en raison des limites physiques. Tandis que les processeurs continuent d'être de plus en plus rapides nos ensembles de données ne cessent de grandir à un rythme encore plus rapide. L'intérêt semble par conséquent se déplacer progressivement vers l'exploration du macrocosme des processeurs par le calcul parallèle et distribué. Il devient habituel d'avoir des ordinateurs portables ayant 2 ou 4 cœurs au sein de la même CPU et les serveurs ayant 8, 32 ou plusieurs noyaux sont monnaie courante.

Le *High Performance Computing* (HPC) se réfère plus généralement à la pratique de l'agrégation de la puissance de calcul d'une manière offrant des performances beaucoup plus élevées que celles d'un ordinateur de bureau ou d'une station de travail typique pour résoudre de grands problèmes en science, en ingénierie ou en gestion. Le HPC utilise habituellement des superordinateurs et / ou des clusters (grappes) d'ordinateurs.

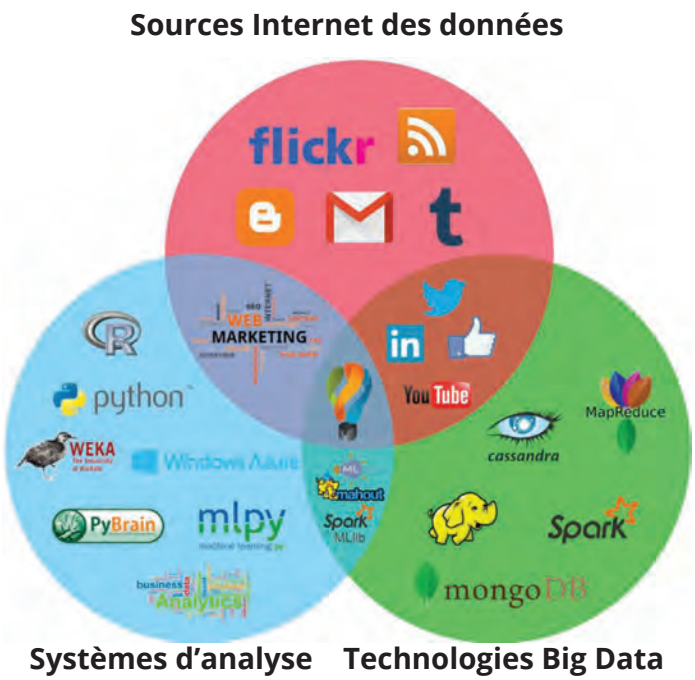
Bien qu'en développement continu, ce traitement d'énormes quantités de données à l'aide des supercalculateurs ou de grappes d'ordinateurs est considéré par la plupart des utilisateurs, mais aussi par la plupart des scientifiques de la donnée et notamment par certains analystes marketing, comme quelque chose dont il ne vaut pas la peine de se soucier, ou de trop coûteux, ou même comme un mythe, comme quelque chose d'inaccessible.

La disponibilité d'une énorme puissance de calcul ou la fascination d'un mythe se transformant en réalité. L'accès démocratisé à une immense puissance de calcul, souvent par cloud computing basé sur l'Internet, génère une nouvelle fascination, celle de pouvoir utiliser l'incroyable pouvoir de l'informatique ; l'impossible devenant possible, les mythes se transforment en réalité.

La Fascination avec le pouvoir. « De l'eau, de l'eau, partout de l'eau, Et pas une seule goutte à boire », cet extrait du poème Samuel Taylor Coleridge intitulé « La complainte du vieux marin » peut aussi être paraphrasée. On doit se sentir bien d'être en mesure d'aider les managers

"naufragés" qui voient « des données, des données, des données partout, et pas un seul octet à utiliser ». La sensation devient plus grande lorsqu'on est capable d'utiliser des grappes d'ordinateurs ou superordinateurs pour faire face au tsunami de données. La Figure 1 montre les principales sources Internet de données client massives ainsi que les systèmes d'analyse et technologies de traitement de Big Data.

Figure 1 - Big Data: Sources Internet, Technologies et Systèmes d'analyse



Source : adapté de Bello-Orgaz & al. (2016)

Les chercheurs et analystes en marketing, cantonnés dans les paradigmes du marketing transactionnel classique où la connaissance du client se limitait aux études à base d'échantillons et panels, se sont fait surprendre par l'avalanche de données comportementales issues des nouvelles techniques du marketing relationnel et digital au point que les informaticiens se sont emparés d'une partie de leur corps de métier. En paraphrasant une citation célèbre de Clemenceau "La guerre est une chose trop importante pour être laissée aux militaires", nous soutenons que "les Big Data sont trop importantes pour être laissées aux informaticiens". Les statisticiens et les analystes marketing devraient jouer un rôle actif et contribuer aux nouvelles approches qui conduisent à des changements révolutionnaires dans la science des données. Les utilisations sont déjà nombreuses en marketing pour le e-commerce et dans les réseaux sociaux (Tableau1).

Tableau 1 – Exemples d'analyses de données massives en e-commerce et pour les réseaux sociaux

Personnalisation

- Gestion du portefeuille de relations clients. Identification des clients ayant les plus grandes rentabilités et fidélités potentielles. Augmentation de la probabilité qu'ils souhaitent l'offre de produit ou service ; maintenir leur fidélité
- Marketing direct/interactif ; marketing relationnel. Machine de recommandation pour générer « Vous aimerez sans doute aussi » et développer ainsi du cross-selling
- Emails personnalisés. Amélioration des taux de réponse par une personnalisation fondée sur l'analyse de la BDD clients ou l'intégration de plusieurs BDD clients.
- Individualisation (customisation) des offres de produit ou service. Produire des profils détaillés de clients, micro-segmenter et personnaliser les offres-produit et ainsi renforcer la fidélité attitudinale et comportementale.

Prix dynamiques

- Identifier le prix qui maximisera la marge ou le profit
- Dériver les prix précis des produits et services avec de fins calculs de la rentabilité client
- Optimiser les prix de millions d'items en un temps record
- Programmer les réductions progressives de produits périssables avant qu'ils ne se gâtent

Gestion des produits et services

- Détecter plus tôt les problèmes de qualité et les minimiser
- Analyser les choix des clients et leurs innombrables avis
- Simuler les emplacements de nouveaux produits dans les linéaires pour tester les effets de conception afin d'améliorer l'acceptabilité des produits après lancement
- Analyse au niveau des unités de gestion des stocks (SKU : Stock Keeping Unit) pour s'assurer que les assortiments de produits sont déjà disponibles

Clusters, Analyses Prédictives

- Intégrer systématiquement les analyses et les connaissances clients en utilisant les données du programme de fidélisation afin de mieux segmenter et cibler
- Développer des segmentations comportementales et des programmes de récompenses multi-niveaux en analysant les profils des clients, les changements des comportements des clients en temps réel et la profitabilité des clients
- Fin réglage des promotions mondiales pour chaque media et chaque région
- Connaître les parts de marché des concurrents locaux dans différentes zones

Réseaux sociaux

- Ciblage publicitaire sur Facebook : Mesures en temps réel pour détecter les utilisateurs les plus valables
- Twitter comme mécanisme de e-BAO (bouche-à-oreille électronique) : Analyse des sentiments
- LinkedIn. Génération de millions de nouvelles pages vues en utilisant « Les gens que vous pouvez connaître »
- Twitter pour la prévision des recettes box-office pour les films : Détection des sujets de conversation, analyse des sentiments
- Marketing viral dans les réseaux sociaux : Analyses de réseau, modèles de diffusion de l'information

Source : d'après Akter et Wamda (2016) ; Bello-Orgaz et al. (2016)

Les solutions open-source relativement récentes qui forment un écosystème autour du plus élégant système statistique, R, et le système de calcul distribué Hadoop (sorte de Linux pour les clusters d'ordinateurs) démocratisent le traitement des Big Data et offrent une opportunité exceptionnelle aux statisticiens et analystes marketing "d'opérer" de vraies usines à calcul.

Quelles données sont "Big"? Solutions open-source sélectionnées

Certains auteurs prétendent qu'un fichier avec plus d'un million d'enregistrements peut être considéré comme Big Data. D'autres indiquent des tailles de plusieurs Tera ou même Peta octets. Une définition moins orthodoxe vient de Hadley Wickham (2015) qui considère que, dans l'analyse traditionnelle le temps de la cognition est plus long que le temps de calcul, tandis que pour les Big Data c'est l'inverse, le temps CPU de calcul prend plus de temps que le processus cognitif de conception d'un modèle.

Dans le choix des outils et des solutions utilisables par les scientifiques du marketing pour les calculs Big Data nous insistons sur des solutions open-source activement développées à la fois pour les tâches de calcul statistiques et parallèles et nous centrons nos illustrations et discussions autour du système statistique R et du système de calcul distribué Hadoop.

Le système statistique R et son écosystème Big Data

Le système statistique R, qui devient progressivement l'outil préféré par la majorité des analystes de données et par un nombre croissant de scientifiques du marketing, a eu, il y a quelques années la réputation de ne pas être en mesure de traiter des Big Data. Cela reste vrai pour beaucoup d'autres solutions logicielles statistiques mais plus pour R qui a beaucoup changé et qui propose aujourd'hui une série de packages spécialisés formant un écosystème pour traiter les Big Data.

Masquer la complexité du Calcul Parallèle Distribué avec MapReduce et Hadoop

Quelles données sont « Big » pour R et comment peuvent-elles être traitées? Wijffels (2013) suggère trois niveaux de taille des données. Le plus bas, est lorsque les données contiennent moins d'un million d'enregistrements et peuvent être traitées en utilisant le R standard. Le second s'entend lorsque les données contiennent entre un million et un milliard d'enregistrements. Elles peuvent également être traitées dans R mais ont besoin d'un effort supplémentaire en adoptant une ou plusieurs des cinq stratégies énumérées ci-dessous. Le troisième niveau, le plus élevé, est atteint lorsque les données contiennent plus d'un milliard d'enregistrements. Dans ce cas, les algorithmes peuvent être conçus avec R et traités avec des connecteurs pour Hadoop ou des solutions HPC alternatives. Les cinq stratégies qui peuvent être utilisées avec R pour faire face aux Big Data, sont : 1) l'échantillonnage, 2) des machines plus puissantes, 3) placer les objets mémoire sur disque, 4) l'intégration avec d'autres langages et 5) les interpréteurs alternatifs.

L'échantillonnage permet de réduire la taille des données. Bien que beaucoup de données soit préférable à peu de données, l'échantillonnage est acceptable si la taille des données franchit le seuil d'un milliard d'enregistrements.

Des machines plus puissantes peuvent être une solution lorsque les données deviennent importantes. Comme R conserve tous les objets en mémoire, une solution consiste à augmenter la mémoire de la machine. Sur les machines 64 bits, R peut traiter 8 To de RAM ; une énorme amélioration par rapport aux de 2 Go de mémoire vive adressables par des machines de 32 bits.

Stocker des objets sur le disque dur et les analyser par morceaux est une solution qui évite le stockage de données en mémoire. Le morcellement (*chunking*) conduit naturellement à la parallélisation et les algorithmes doivent être explicitement conçus pour traiter des types de données spécifiques au disque dur. "ff" et "ffbase" sont les packages CRAN les plus connus (open-source) qui suivent ce principe. Le package "scaleR" de Revolution R Enterprise est également une solution populaire.

L'intégration avec des langages de programmation performants comme le C ++ ou Java. Afin d'éviter les goulots d'étranglement et des procédures coûteuses en performance, des parties du programme sont déplacées de R à un autre langage. L'objectif est de combiner, lors du traitement des données, l'élégance de R avec la performance d'autres langages. Il est relativement facile d'externaliser le code de R vers ces langages en utilisant les packages Rcpp et Rjava.

Les interpréteurs alternatifs pourraient être plus rapides ou mieux adaptés dans certains cas que le R standard. Ceux-ci sont pqR (pretty quick R), Renjin qui réimplémente l'interpréteur R en Java et peut donc fonctionner sur la machine virtuelle Java (JVM) et TERR, un interpréteur basé sur C ++ créé par Tibco. Oracle R utilise la bibliothèque mathématique d'Intel pour obtenir de meilleures performances sans modifier le noyau de R.

Les cinq stratégies évoquées sont des palliatifs lorsque les données ne sont pas encore trop grandes pour R. Lorsque ce n'est pas le cas, des connecteurs à Hadoop ou des solutions HPC alternatives doivent être utilisés en conjonction avec R.

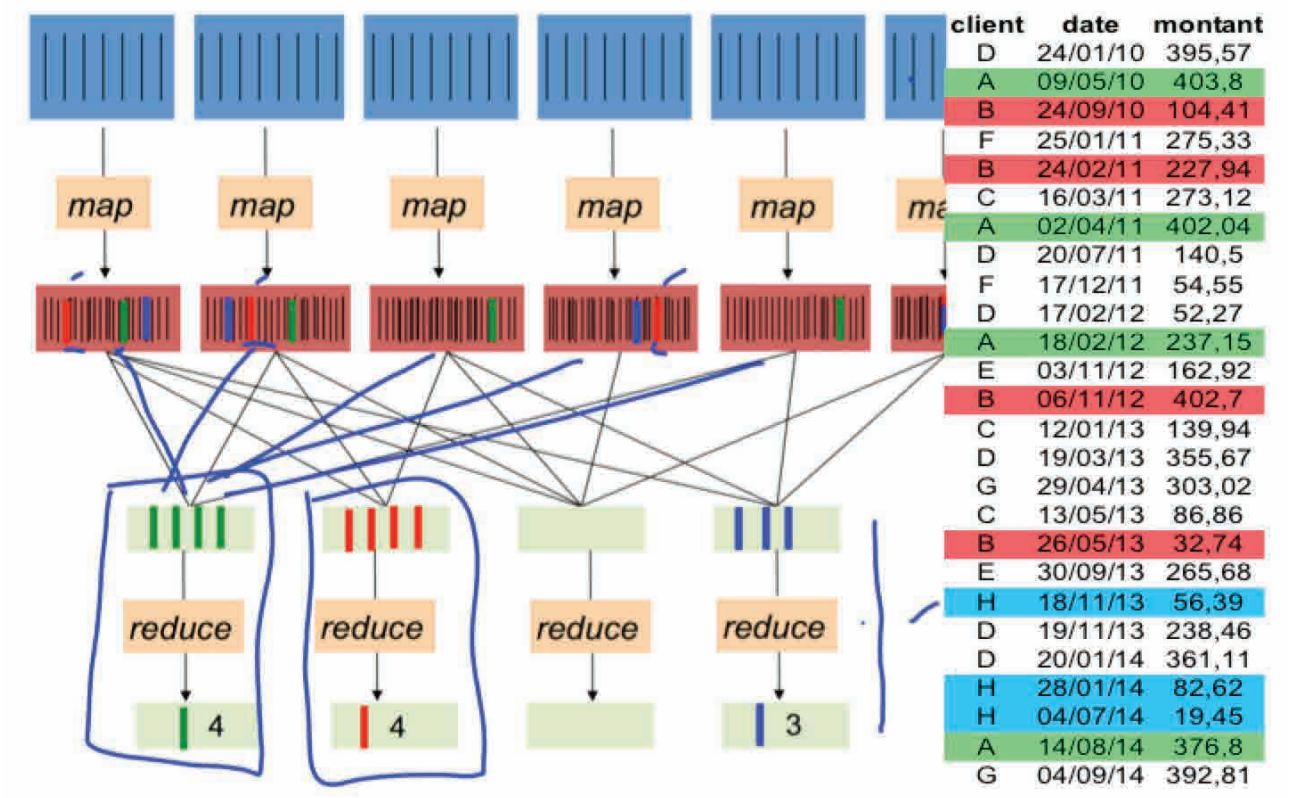
MapReduce est un modèle de programmation de haut niveau et une mise en œuvre associée pour le traitement de données en parallèle à grande échelle. Il a le mérite d'avoir fortement contribué à démocratiser le traitement des Big Data. A l'origine, le nom MapReduce faisait référence à une technologie propriétaire Google qui est devenue depuis une marque générique. Il est intégré dans Hadoop d'Apache, une plateforme logicielle open-source, écrite en Java, pour le stockage et le traitement distribué de très grands ensembles de données sur les grappes d'ordinateurs.

Il est utile pour les statisticiens et analystes marketing d'avoir une bonne compréhension de MapReduce afin d'être en mesure d'adapter leurs modèles et algorithmes au traitement parallèle distribué des ensembles de données volumineux. MapReduce est basé sur l'observation que la plupart des calculs peuvent être exprimés en termes d'une procédure Map, qui permet le filtrage et le tri, et d'une procédure Reduce, qui effectue une opération d'agrégation comme le comptage, la somme etc. Map et Reduce sont des fonctions d'ordre supérieur usuelles en programmation fonctionnelle à laquelle le système statistique R appartient.

Utilisation de MapReduce dans les calculs d'agrégation des données massives

L'approche de MapReduce peut être facilement appliquée aux calculs d'agrégation. Dans le marketing relationnel on agrège les données sur les transactions des clients pour calculer des variables fondamentales pour la segmentation comportementale et le ciblage : la fréquence (F), la récence (R) et le montant (M) des achats.

Figure 2 - Exemple MapReduce: calculs RFM sur une base de clientèle



Source: adapté de Howe B. eScience Institute, U. Washington, 2013

Dans l'exemple ci-dessus chaque transaction client, comme on peut le voir sur le côté droit de la Figure 2, est un enregistrement qui contient l'identifiant du client, la date de la transaction et le montant dépensé à cette occasion. Le fichier que nous utilisons peut être considéré comme Big Data car il contient 343.766.402 transactions (enregistrements) effectués au cours de 78 semaines par 6.326.658 clients.

Ainsi, pour la *fréquence* d'achat, la fonction *Map* retourne *une paire clé / valeur* composée de l'identifiant du client et la valeur un. La fonction *Reduce* sera, soit la somme soit la longueur du vecteur contenant ces valeurs groupées par la clé.

Pour la *récence* qui, dans ce cas, est la date de la dernière transaction d'un client, la fonction Map doit retourner, en tant que valeur de la paire clé / valeur, la date de la transaction, tandis que la fonction Reduce fusionnera la date maximale par client.

Pour le *montant* qui pourrait être le montant total ou moyen dépensé par chaque client, la fonction Map doit retourner en tant que valeur du couple clé / valeur le montant dépensé par transaction, tandis que la fonction Reduce fusionnerait par client, soit la somme soit la moyenne du montant dépensé.

Si les données étaient trop grandes et devaient être traitées sur un cluster d'ordinateurs, le processus de MapReduce (voir Figure 2) lirait le fichier d'entrée et le diviserait (*split*) en plusieurs morceaux (*chunks*). Chaque morceau se voit associé à un exemplaire du programme Map et ces derniers sont exécutés en parallèle sur les nœuds du cluster pour regrouper les données d'un morceau par client. Le système prend alors la sortie de chaque programme Map et fusionne (*shuffle / sort*) les résultats pour le programme Reduce, qui calcule la longueur (ou la somme des valeurs) du vecteur par client pour obtenir la *Fréquence*, la valeur maximale pour la *Récence* et

la somme ou la valeur moyenne pour le *Montant*.

Comme R est également un langage fonctionnel, les fonctions Map et Reduce et / ou des fonctions d'ordre supérieur équivalentes peuvent aussi être utilisées sur une seule machine. Dans ce cas, une fonction qui applique l'approche de MapReduce que nous avons utilisée est *tapply*. Sa syntaxe est tapply (valeur, clé, FUN) où la valeur et la clé peuvent être les sorties d'une fonction Map tandis que FUN est une fonction Reduce de fusion comme *somme, moyenne, longueur*, etc.

Par conséquent, en R, appliquer l'approche MapReduce (mais pas le processus MapReduce) aux données de la Figure 2 consiste à utiliser :
tapply(rep(1, length(client)),client, sum) pour le calcul de la fréquence d'achats par client, F
tapply(date,client, max) pour le calcul de la récence des achats par client, R
tapply(montant,client, sum) pour le calcul du montant des achats par client, M

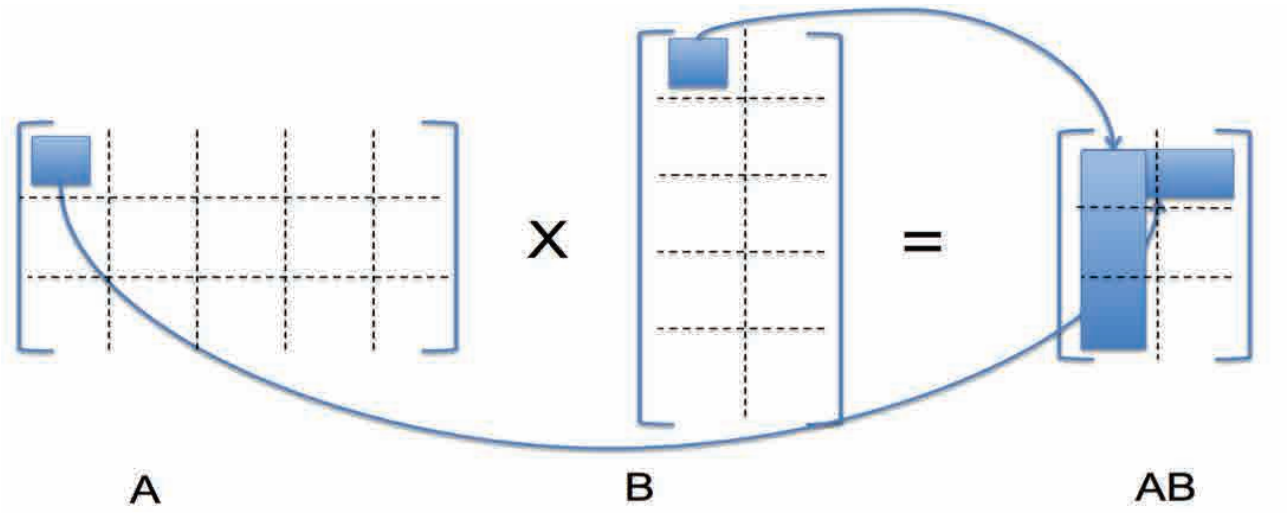
Mettre en œuvre un processus de MapReduce avec R et Hadoop. Lorsque les données sont trop grosses pour la mémoire d'un ordinateur, un processus MapReduce sur un cluster d'ordinateurs doit être mis en place. MapReduce par rapport aux approches HPC standard est un processus de type "partage rien".

Le processus de MapReduce peut être mis en œuvre après l'installation du système de Hadoop open-source sur chaque ordinateur de la grappe. Une série de packages R pour mettre en œuvre la solution RHadoop est disponible gratuitement auprès de *Revolution Analytics*. Ces packages gardent transparentes toutes les complexités du calcul parallèle pour les utilisateurs et sont donc hautement recommandables aux analystes marketing.

MapReduce dans des calculs statistiques avancés utilisés en marketing

Alors que les calculs d'agrégation pour obtenir des variables RFM qui résument le comportement transactionnel du client dans la BDD marketing sont une application évidente de l'approche MapReduce, ce n'est pas toujours le cas des modèles de marketing quantitatif plus sophistiqués, mais beaucoup peuvent être adaptés pour MapReduce. De nombreux modèles statistiques utilisés en marketing quantitatif et analyse des données utilisent des calculs d'algèbre linéaire. Parmi ceux-ci la multiplication de matrices est un calcul très important qui peut être adapté à l'approche MapReduce (voir Figure 3 et 4)

Figure 3 - La multiplication de matrices adaptée pour MapReduce

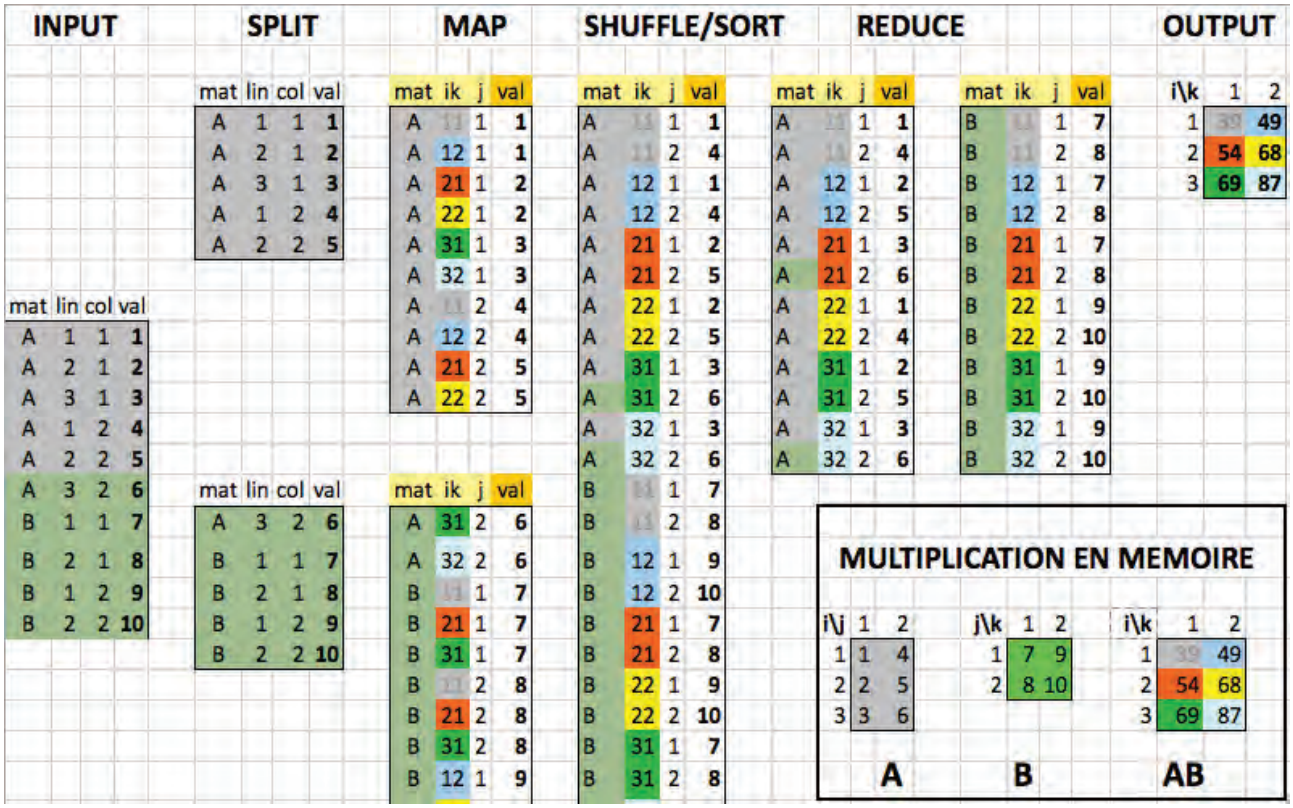


Il est bien connu que dans la multiplication matricielle la matrice résultat (AB) est la somme des produits des cellules de la ligne correspondante dans la première matrice (A) et des cellules de la colonne correspondante dans la deuxième matrice (B). Pour la phase "Reduce" toutes ces cellules doivent alors avoir la même clé, qui correspond à la position (ligne et colonne) de ce calcul dans la matrice résultat (AB).

La matrice résultat a les dimensions *IxK*, où I est le nombre de lignes de la première matrice et *K* est le nombre de colonnes de la deuxième matrice et *J*, le nombre de colonnes de la première matrice, doit être égal au nombre de lignes de la deuxième.

Comme le montre la Figure 3 chaque cellule (*i,j*) dans la matrice A est utilisée *K* fois pour le calcul des cellules de la ligne (*i*) dans la matrice résultat et chaque cellule (*j,k*) dans la matrice B est utilisée *I* fois pour le calcul des cellules de la colonne (*k*) de la même matrice (AB).
C'est la raison pour laquelle (voir Figure 4) dans la phase "Map" chaque valeur (cellule) de A doit être dupliquée *K* fois et se voit associée une clé (*i,k*) où *k=1..K* et chaque valeur (cellule) de B doit être dupliquée *I* fois et se voit associée une clé (*i,k*) où *i=1..I*.
La phase « reduce » calculera pour chaque cellule résultat (*i,k*) la somme de tous les produits des éléments des deux matrices associés à la même clé.

Figure 4 – Etapes du processus MapReduce de multiplication de matrices



La Figure 4 illustre le processus MapReduce comparé à la multiplication classique en mémoire. Dans la phase "*split*" le fichier qui représente les deux matrices regroupées est divisé en deux morceaux (*chunks*) et chaque morceau est attribué à un nœud (ordinateur). On remarque que la matrice A, étant "trop grande" pour le premier nœud, a vu son dernier élément passer dans le "*chunk*" destiné au deuxième ordinateur. La phase "*map*" se déroule séparément sur chaque nœud et les valeurs des éléments de chaque matrice sont dupliqués et associées à des clés, comme expliqué précédemment, pour pouvoir d'abord être triés sur chaque ordinateur. La phase "*shuffle/sort*" récupère ces données pour les regrouper et trier au niveau du "*cluster*".

Finalement la phase "*reduce*" calcule la somme des produits des valeurs des deux matrices qui ont la même clé.

Des modèles comme la régression linéaire, l'analyse factorielle ou l'analyse discriminante appliquent des algorithmes de calcul plus sophistiqués (inversion, diagonalisation) à une matrice symétrique d'assez petites dimensions qui dépendent du nombre de variables qu'elle résume et non du nombre d'observations. Cette petite matrice des sommes des produits croisés des variables est obtenue par la multiplication de deux matrices qui, elles, peuvent être des Big Data car elles contiennent les observations qui peuvent être très nombreuses.

Le modèle de programmation parallèle des données de MapReduce cache la complexité de la distribution et de la tolérance aux pannes. Sa philosophie principale est d'augmenter à l'échelle la puissance de calcul par l'ajout de matériel pour faire face à la taille des problèmes, mais aussi d'économiser les coûts de matériel, programmation et administration. MapReduce ne convient pas à tous les problèmes mais de nouveaux modèles et environnements de programmation sont encore en cours de création et renforcent ces idées. Comme nous avons pu le voir, les solutions de calcul doivent être adaptées à MapReduce car les programmes classiques ne peuvent faire le travail lorsque les données sont réparties entre plusieurs nœuds. Une solution prometteuse est l'approche DAG (graphes acycliques orientés), un style de programmation pour les systèmes distribués. Il peut être considéré comme une alternative à MapReduce. Alors que MapReduce comprend seulement deux étapes (Map et Reduce), DAG peut avoir plusieurs niveaux pouvant former une structure arborescente et DAG est donc plus souple avec davantage de fonctions comme map, filter, union, etc. Son exécution est aussi plus rapide grâce à la non écriture sur disque des résultats intermédiaires. Une des mises en œuvre les plus plébiscitées de DAG est le projet Spark de Apache. Son principal concept de RDD (Resilient Distributed Datasets) est expliqué dans Zaharia et al. (2012)[1] de l'Université de Berkley, à l'initiative de cette approche. Spark amène MapReduce à un niveau supérieur avec des remaniements moins couteux dans le traitement des données. Avec des fonctionnalités telles que le stockage de données en mémoire et le traitement en temps quasi-réel, la performance peut être plusieurs fois plus rapide qu'avec les autres technologies Big Data.

D'autres solutions HPC pour les scientifiques (analystes) marketing

Alors que MapReduce et ses descendants cachent complètement la complexité du calcul distribué et parallèle et sont donc facilement implémentables comme services et accessibles à tous par le biais du *cloud computing*, il existe tout un ensemble de méthodes et facilités HPC également disponibles pour les universitaires scientifiques du marketing. De nombreuses universités ont ou participent à des projets de *clusters* ou *grids* d'ordinateurs. Les auteurs, membres de l'université de Lille qui occupe en France le cinquième rang du point de vue de la capacité de calcul de son cluster d'ordinateurs[2], ont pu utiliser ces installations et une série de packages R afin de tester des gains de performances sur la modélisation prédictive basées sur des variables RFM en faisant varier le nombre de noyaux et d'ordinateurs de la grappe. Pour une liste détaillée des paquets disponibles R on peut lire la page Web officielle de la CRAN task view concernant le calcul de haute performance et parallèle avec R[3].

1. En bref, les RDD sont des ensembles de données distribués qui peuvent rester en mémoire et se replier sur disque. Le RDD en cas de perte peut être facilement reconstruits à l'aide d'un graphe qui indique comment le reconstruire. Les RDD sont très bien quand il est nécessaire de garder un ensemble de données en mémoire et lancer une série de requêtes - cela fonctionne mieux que l'extraction de données à partir du disque à chaque fois. Un autre concept important RDD est qu'il existe deux types de choses qui peuvent être faites sur un RDD : 1) des *transformations* comme, map, filter qui génèrent un autre RDD et 2) des *actions* comme count (compter) qui retournent des valeurs. Un job Spark se compose d'un DAG de tâches qui exécutent des transformations et des actions sur des RDDs.
2. Le cluster informatique utilisé se compose de machines hétérogènes qui incluent 2110 cœurs de processeurs (CPU) et 7168 cœurs de processeurs graphiques (GPU). Le cluster a une capacité de disque d'environ 156 Tb et un débit théorique de 50 Tflops. Toutes les machines sont reliées par un réseau Infiniband 40 Gbps.
3. https://cran.r-project.org/web/views/HighPerformanceComputing.html

Nous introduisons une approche brute d'enregistrement de l'analyse de la performance du calcul parallèle et l'appliquons à des calculs qui évaluent plusieurs modèles prédictifs basés sur des variables RFM du comportement d'achat. Les deux ensembles de données utilisés proviennent d'une chaîne de magasins et d'un catalogue de vente par correspondance. Ils reflètent le comportement d'achat répété pendant plusieurs saisons pour des cohortes de clients. Le premier ensemble de données a déjà été utilisé dans le présent document pour illustrer l'approche de MapReduce pour calculer des variables RFM et réduire un grand fichier de 344 millions de transactions à un fichier plus petit mais toujours assez grand avec 6,3 millions de clients. Afin de mesurer l'impact des variables RFM sur l'incidence d'achat, nous utilisons ce fichier pour estimer plusieurs modèles prédictifs : linéaire, logit, probit, réseaux de neurones ou les arbres de classification (CART). Comme le fichier est assez grand, l'objectif principal est ici de mesurer le temps qu'il faut pour calibrer un modèle en utilisant un noyau de la CPU, ici le microprocesseur Intel i7 d'un ordinateur MacBook Pro. Le temps le plus court, 4-5 secondes, a été atteint avec la régression linéaire et le plus long avec les réseaux neuronaux a pris plus de 820 secondes ou 14 minutes et 320 itérations pour converger.

Comme la régression linéaire multiple peut être résolue facilement en mimant une approche MapReduce nous l'avons appliquée sur le même ordinateur MacBookPro, pour comparer des calculs sériels à des calculs parallèles implicites en faisant varier le nombre de cœurs du processeur de 2 à 4. La régression linéaire peut être résolue par des multiplications simples de la matrice contenant les valeurs des variables indépendantes et une première colonne de uns, appelée **X** et d'un vecteur contenant les valeurs de la variable dépendante, appelée **y**. En supposant que la mémoire est limitée, ces grandes tables de 6,3 millions enregistrements ont été réparties en fonction du nombre de cœurs dans 2 à 4 morceaux. Des *fonctions Map* ont été définies pour calculer des multiplications par morceau (*chunkwise*) **X'X** et **X'y** et appliquées à la fois en série et en simultané en utilisant le parallélisme implicite. Pour la dernière, la fonction *mcapply* du package *parallel* de R a été utilisée.

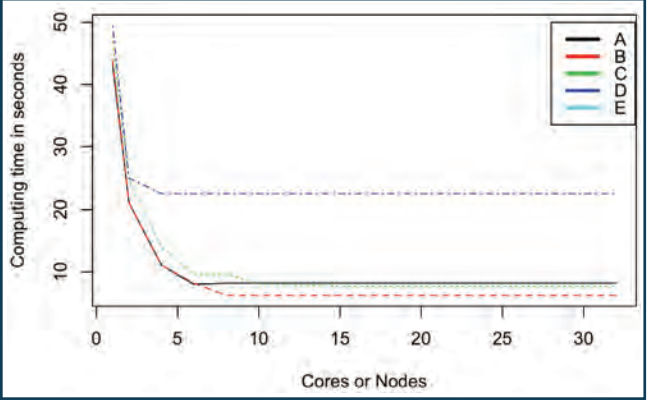
La *fonction Reduce* fait la somme des résultats par morceaux de ces multiplications particulières en exploitant simplement le fait que la somme des matrices est associative et commutative. Nous avons répété le calcul en série et en parallèle sur 2, 3 et 4 morceaux une centaine de fois. Le test t apparié a montré une différence positive importante en temps de calcul entre les calculs série et parallèle ($t = 40{,}2006$, $dl = 299$, p-valeur $<2{,}2e\text{-}16$). Alors que les calculs en série avaient une durée moyenne de 1,99 secondes, les calculs parallèles n'ont duré que 1,44 secondes et la différence moyenne était de 0,51 secondes. La même différence positive significative est observée en analysant les calculs avec coupure en 2 morceaux (0.60855s), 3 morceaux (0.21772s) et 4 morceaux (0.68903) séparément.

Le second fichier avec ses six mille enregistrements représentant l'incidence d'achat des clients en vente par correspondance est beaucoup plus petit et a été utilisé pour effectuer une validation croisée de type "*leave-one-out*" du modèle afin de limiter la surestimation et améliorer la validité prédictive. La validation croisée, comme le *boostrapping* est un problème dit « **embarrassingly parallel** » (aussi appelé *perfectly parallel* ou *pleasingly parallel*), c'est-à-dire facile à paralléliser. La validation croisée leave-one-out pour nos 6000 observations impliquait l'ajustement du modèle répété 6000 fois sur des ensembles obtenus en éliminant à chaque fois une observation. Ce calcul répétitif peut facilement être parallélisé aussi bien pour le parallélisme *implicite ou multicœurs*, signifiant le partage de la même mémoire par plusieurs noyaux, que pour le *parallélisme en cluster* qui regroupe plusieurs ordinateurs.

Tableau 2– Baisse de durée avec la parallélisation

Cores/Nodes	A	B	C	D
1	43.217	43.817	46.802	49.445
2	21.139	21.133	24.683	24.985
4	11.151	11.131	13.995	22.484
6	8.026	8.208	9.694	22.484
8	8.211	6.250	9.650	22.484
10	8.211	6.250	8.143	22.484
16	8.211	6.250	7.676	22.484
32	8.211	6.250	7.721	22.484

Figure 5: Gains de vitesse du calcul parallèle



A: linéaire, noeuds1/coeurs >0; B: logistic, noeuds1/coeurs >0; C: logistic, noeuds > 0/coeurs 1; D : logistic, interactive, noeuds 1/coeurs >0

Les trois premières colonnes du tableau 2 enregistrent des gains de performance de parallélisation à l'aide de calculs batch (par lots) sur des serveurs Dell PowerEdge ayant des CPU à 8 cœurs et appartenant au même cluster d'ordinateurs. La dernière colonne enregistre des gains de performances similaires sur un portable MacBook Pro avec CPU Intel I7 à 4 cœurs. Les colonnes A, B et D montrent des gains de performance obtenus en augmentant le nombre de noyaux, tandis que la colonne C analyse des gains de vitesse en augmentant le nombre de machines (nœuds) dans les clusters. On peut observer que lorsque le nombre maximum de noyaux est atteint, ici huit pour les serveurs et 4 pour l'ordinateur portable, aucune amélioration supplémentaire de la performance n'est possible. Les gains de performances tout en augmentant le nombre de machines ou noeuds (colonne C) peuvent être obtenus de façon constante mais, là aussi, il y a une limite lorsque les gains de performance deviennent moins importants.

Conclusion

Le marketing repose de plus en plus sur la technologie de l'information, qu'il s'agisse du choix de canal, de la personnalisation et des systèmes de recommandation, de l'analyse des contenus générés par les utilisateurs, des commentaires en ligne et de l'influence sociale dans les réseaux sociaux en ligne (Tableau 1). Le marketing est maintenant considéré comme l'un des moteurs (*driver*) des technologies Big Data, tout comme l'a été la comptabilité pour les bases de données dans les années 80. La technologie a toujours transformé la science du marketing en suivant une tendance assez systématique et prévisible. Cela implique des changements dans la formulation des problématiques, dans les méthodes, et de mettre davantage l'accent sur les aspects analytiques du marketing. Les scientifiques du marketing ont de bonnes connaissances en statistiques, en économétrie et en recherche opérationnelle, mais semblent en avoir beaucoup moins en matière de programmation moderne et de technologie de l'information. Beaucoup ignorent encore des fonctionnalités qui profitent du contexte distribué et riche en données fourni par l'Internet, le cloud computing et le HPC. Négliger les facteurs qui favorisent la facilité d'usage des modèles (convivialité) risque de rendre ces derniers non pertinents et de limiter leur utilisation. En essayant de démystifier les approches Big Data cet article invite les statisticiens et analystes marketing à accorder plus d'attention aux évolutions technologiques, à participer davantage à l'élaboration d'éléments analytiques spécifiques et à ne pas abandonner le champ aux informaticiens. Démystifier les approches et technologies Big Data ne signifie pas les banaliser mais, au contraire, insister sur la haute importance et les changements de rupture qu'elles engendrent pour la société en général, pour la science statistique et pour la science du marketing en particulier.

Références

- Akter S. et S.F. Wamda (2016), Big data analytics in E-commerce: a systematic review and agenda for future research, *Electronic Markets* 26:173–194
- Bello-Orgaza G., Jungb J. J., Camachoa D. (2016) Social big data: Recent achievements and new challenges, *Information Fusion* 28: 45–59
- Wickham H. (2015), R Packages. Sebastopol, CA: O'Reilly Media, Inc.
- Wijffels J., (2013), "ffbase: statistical functions for large datasets", The R User Conference, Albacete, Castilla-La Mancha
- Zaharia M., Chowdhury M., Das T., Dave A., Ma J., McCauley M., Franklin M.J., Shenker S., Stoica I. (2012) Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, NSDI 2012, April.

De la Data aux Big Data : enjeux pour le Marketing client – Illustration à EDF

ENTRETIEN AVEC ANNE GAYET ET JEAN-MICHEL GAUTIER

Jean-Michel GAUTIER

Professeur des départements Economie & Marketing à HEC, en charge de la Connaissance Client à EDF Commerce (commercialisation et gestion des contrats de fourniture des différents marchés : particuliers, professionnels, entreprises et collectivités locales).

Anne GAYET

Directrice associée chez AID (Add Intelligence to Data). AID accompagne les équipes de la Direction Numérique d'EDF-Commerce. Par ailleurs AID propose des services d'hébergement de CRM marketing, de datamining / data science, d'analyse des parcours clients et de conseil pour le marketing opérationnel.

Quand avez-vous débuté avec le "Big Data"?

Jean-Michel Gautier

Dès 2003, date de sa création, EDF Commerce a été confronté, à travers l'ouverture progressive des marchés, à l'arrivée de concurrents, et à la nécessité de passer d'une approche usager à une approche client. C'est dans ce contexte que se sont construites progressivement la connaissance client et les plateformes de stockage et d'analyse des données clients.

Après une première étape de développement d'entrepôts client et la mise en place d'une activité d'analyses au service du marketing et des opérations, s'est posée la question de l'adoption de nouvelles technologies capables de supporter de l'interaction en temps réel et d'exploiter les nouvelles sources de données que constituent le web, les applications mobiles, les objets connectés et en particulier les compteurs intelligents.

Nous avons décidé dès 2012 de migrer l'ensemble des infrastructures des entrepôts de données vers une architecture Hadoop, avec un double objectif fonctionnel et économique. Nous avons en parallèle migré les activités d'analyse vers l'open source R, en remplacement de SAS.

Anne Gayet

Chez AID nous avons pendant ce temps démarré une transformation complète de l'approche des "parcours clients".

Nous avons décidé en 2013 de développer une nouvelle façon d'analyser et de gérer les parcours clients, en développant une plateforme "Big Data" pour la centralisation de toutes les interactions clients quel que soit le canal, pour leur structuration, et pour l'analyse interactive des parcours. Entièrement élaborée à partir d'outils open source, notre solution a été initiée autour de deux ensembles d'outils : Hadoop et Hbase (une base de données NoSQL orientée colonnes) pour le stockage et l'analyse des données d'une part, Kafka pour la gestion de la file d'attente et Storm pour l'intégration et la transformation des données au fil de l'eau d'autre part. Une interface de visualisation totalement originale a été développée.

Qu'en est-il aujourd'hui? Les outils utilisés ont-ils évolué?

Jean-Michel Gautier

Cette rupture complète des technologies de stockage et d'exposition de données, de DB2 et Oracle vers Hadoop, Hbase et de SAS vers R, est un succès total avec un triple bénéfice :

- Réduction à périmètre égal des coûts de fonctionnement des entrepôts
- Capacité de stockage étendue à très bas coûts
- Passage progressif vers le traitement des interactions client en temps réel (en particulier

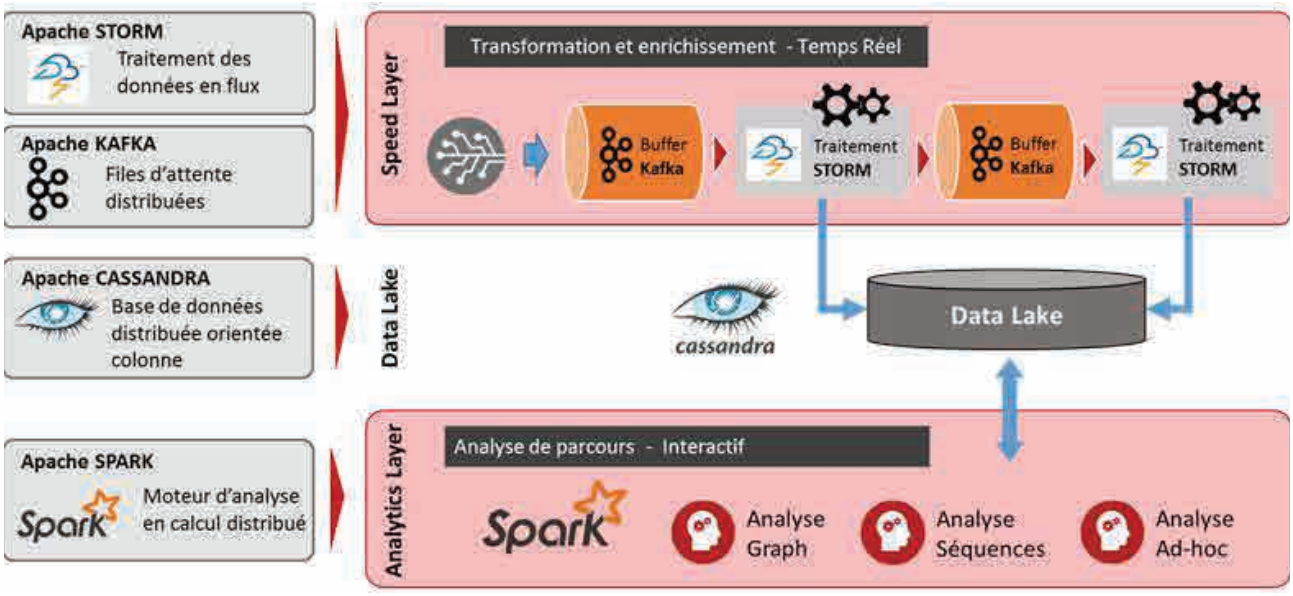
sur le web)

L'utilisation de Hbase pour l'exposition de données en temps réel et de Spark associé à R, est la suite logique pour travailler efficacement dans un environnement distribué.
Par ailleurs de nombreux tests d'outils autour d'Hadoop sont en cours.

Anne Gayet

Nous avons initialement pensé notre outil avec Hadoop et Hbase pour la partie analytique, mais nous avons rapidement migré vers Spark et Cassandra :

- Spark pour sa plus grande vélocité d'analyse réalisée en mémoire, et la richesse de sa bibliothèque d'algorithmes en calculs distribués (MLlib : Machine Learning library).
- Cassandra pour sa facilité d'utilisation, notamment grâce à son langage de requêtage CQL, qui ressemble au SQL.



Qu'est-ce qui a fait les premiers succès auprès des métiers et en particulier des marketeurs?

Jean-Michel Gautier

La réduction des coûts de stockage sur Hadoop et la facilité à acquérir de nouveaux flux de données ont complètement changé la vision de la gouvernance de la donnée.
Avant, il fallait démontrer la valeur d'usage pour justifier les dépenses de collecte et stockage. Aujourd'hui on peut capter et collecter tout ce qui passe sans le traiter. La modélisation objet, le traitement et l'analyse se font au fil des besoins selon les cas d'usage. L'investissement data se fait donc au fur et à mesure en fonction du business model des cas d'usages.
Chez EDF, je relèverai quatre facteurs clés de réussite :

- La rapidité à délivrer une solution isopérimètre (15 mois avec livraison au fil de l'eau par périmètre fonctionnel).
- Les premiers cas d'usage d'exposition de données en temps réel pour de nouveaux services au client : par exemple la possibilité pour les entreprises multi-sites d'accéder à leurs consommations et factures via le web et de regrouper à volonté leurs données selon des logiques d'agrégation diverses de leurs sites.
- La continuité, l'extension des activités décisionnelles classiques avec l'analyse de données non structurées (réclamations, commentaires conseillers, aides en ligne, mails, tweets), l'analyse fine des consommations (issues des compteurs Linky) à des fins d'accompagnement à la maîtrise énergétique.

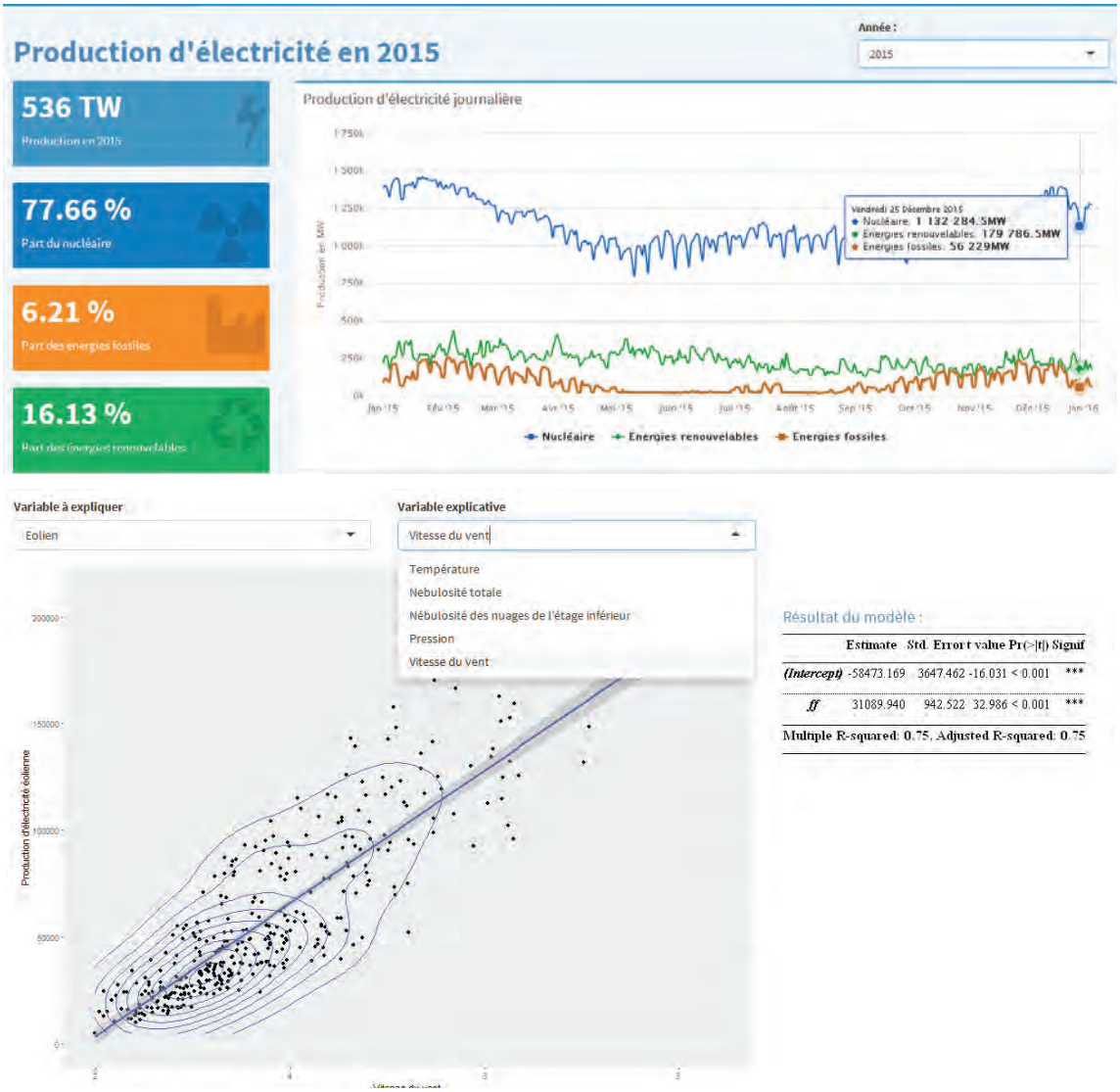
- La capacité à montrer en interne de nouvelles applications illustrant l'efficacité et la pertinence d'un traitement de bout en bout en temps réel. Par exemple : la collecte, le filtrage et la catégorisation en temps réel des tweets qui concernent EDF selon des axes prédéfinis visibles pour les managers concernés.

Anne Gayet

La dimension datavisualisation est extrêmement importante, tant chez EDF que dans notre plateforme datakili®. Cette dimension se concrétise de plusieurs façons :

- La restitution et l'analyse interactive avec des outils comme Tableau Software, mais aussi en open source avec R Shiny, ou D3.js.
- La synthèse des éléments de connaissance clients apportés par les analyses, sous forme d'infographies, de présentations très « design » et visuelles, ou encore par des mini-vidéos ou animations. Ces modes là n'ont rien de « Big Data » ...
- Et bien sûr par le développement d'interfaces orientées « dataviz » quand il s'agit d'outils ou d'applications.

A titre d'illustration ci-dessous un exemple d'application Shiny très simple permettant d'analyser la production d'électricité au cours du temps, et en fonction de paramètres météorologiques (sources de données ouvertes).



Nb : De nombreux exemples d'applications sont consultables dans ce catalogue en ligne http://www.showmeshiny.com

Quelles ont été les difficultés à surmonter ?

Jean-Michel Gautier
Dans l'exploitation opérationnelle des Big Data, le premier besoin est la capacité à exposer et traiter de l'information en temps réel, ce qui est relativement simple avec l'écosystème Hadoop. Même si, comme beaucoup d'entreprises, EDF est encore contraint par ses systèmes d'information traditionnels dont l'interface de publication est quotidienne, en nuit applicative. Une migration complète des politiques de publication de données par les SI traditionnels est nécessaire pour arriver au temps réel.
Cette contrainte aboutit aujourd'hui à une gestion du temps réel canal par canal, le plus facile étant les canaux web et mobile. La difficulté et l'enjeu à venir est bien la gestion client omnicanale en temps réel.

Anne Gayet
Une des plus grandes difficultés que nous rencontrons toujours chez nos clients est la lutte contre les « silos » d'informations. Par exemple il est encore fréquent que les comportements de navigation sur les sites web de l'entreprise ne soient pas ou peu accessibles par les analystes. Une solution consiste à poser un tag spécifique sur les sites web de l'entreprise pour alimenter les bases de données. Mais la maîtrise de la technique permettant de tracer la navigation sur un site web ou l'utilisation d'une application mobile est une compétence très particulière que n'ont pas les data managers et dataminers. La solution souvent pratiquée consiste à utiliser un traceur ou un gestionnaire de tag déjà implémenté, à condition que ce système donne accès aux données de navigation détaillées et pas uniquement à des comptages (par exemple Google Analytics Premium, Adobe Analytics, Tag Commander, ...).

C'est quoi le marketing chez EDF Commerce ?

Jean-Michel Gautier
C'est avant tout une dimension offre d'électricité et de services (conseil tarifaire, rénovation) couplée à un programme relationnel, une fidélisation par la relation client et une attention portée à tous les évènements clés de la relation client : déménagement, changement de puissance, difficultés de paiement, travaux. Ces prestations sont axées sur le confort dans l'habitat pour les particuliers et sur le service commercial pour les entreprises, mais aussi pour tous sur l'accompagnement de la maîtrise de l'énergie.

Pouvez-vous nous citer des cas d'usage en cours de développement ou à venir ?

Jean-Michel Gautier
Nous avons déjà évoqué l'analyse de tweets. Notre « Tweet Tracker » n'est pas uniquement destiné à prouver la possibilité de capter - analyser - restituer en temps réel. Il intègre des fonctions d'analyses linguistiques poussées non disponibles dans les plateformes de veille des réseaux sociaux. C'est le précurseur d'outils de relation client capables de comprendre la parole client.
De nouveaux services tirant parti d'objets connectés sont expérimentés : à partir de la consommation client et des températures (celle du logement, la température extérieure) on propose un service de régulation du chauffage.
Nous disposons de différentes zones d'expérimentation poussées, dont Smart Electric Lyon et Nice Grid, où le foisonnement technologique est de mise.
Le fait d'utiliser Hadoop permet d'ouvrir plus aisément les données à l'extérieur du SI. Les données sont ainsi présentées au client dans son espace client. Elles sont également poussées vers le conseiller en ligne qui bénéficie ainsi d'une « vision 360° » du client, pour une connaissance

parfaite de son interlocuteur et une relation commerciale optimisée.
Les pistes d'investigation actuelles s'appuient parfois sur des startups, par exemple en ce qui concerne la mise en place de nouveaux portails de visualisation pour les métiers ou partenaires.

Anne Gayet
L'analyse des parcours clients, ou le domaine appelé « expérience client », s'appuie habituellement sur une optimisation des processus de gestion de la relation client ou sur une mesure par des enquêtes. Une solution comme datakili® permettant de stocker, quantifier et analyser tous les parcours clients permet de les concrétiser, les analyser et faire de la prédiction. On peut ainsi par exemple identifier les événements précurseurs de la résiliation, mieux comprendre les parcours cross-canaux ou encore mettre en évidence les séquences d'interactions client – entreprise menant à un appel au call center.

Pour conclure qu'apporte ce nouvel écosystème Big Data?

Anne Gayet
En ce qui concerne les capacités analytiques pour le marketing, les solutions traditionnelles étaient adaptées à ce fonctionnement : compréhension du besoin marketing, compréhension des données, collecte et préparation des données, modélisation prédictive, évaluation, déploiement, surveillance et mises à jour. D'abord l'écosystème « Big Data » permet de satisfaire ces besoins classiques de façon plus performante en rapidité, en qualité statistique et en pertinence vis-à-vis du client. C'est à l'heure actuelle toujours le principal apport. Cette meilleure performance est obtenue grâce :

- à la baisse des temps de calcul,
- des mises à jour fréquentes des modèles,
- la prise en compte de données plus variées et détaillées pour une meilleure capacité explicative, cela inclut des données externes comme la météo,
- l'utilisation possible en temps réel pour être au plus près du « temps du client »,
- à l'application d'algorithmes très gourmands en calculs (parfois).

Par ailleurs, ce nouvel écosystème permet de diffuser l'utilisation des données au service des métiers, via la mise en place de « hubs de données » associés à des outils permettant aux utilisateurs d'explorer et visualiser les données en leur masquant la complexité sous-jacente.

Jean-Michel Gautier
Nous avons évoqué des développements internes à EDF, mais les startups de l'écosystème sont également des sources potentielles d'accélération.
Nous organisons donc une veille active et procédons à des tests de nouvelles technologies ou de nouveaux services, via des POCs (proof of concept) réalisés en interne ou en externe, i.e. des expérimentations limitées permettant de faire la preuve du concept avant le déploiement.
Le marché de l'énergie est en pleine mutation : énergie verte, nouveaux usages de l'électricité (voitures électriques), production locale (client fournisseur), équilibre de réseaux, etc. Et tout ceci produit un foisonnement de données qu'il faut savoir collecter et traiter pour des services à inventer. Seules les technologies du Big Data nous rendront capables d'aborder ces enjeux.

Confiance dans les statistiques publiques : une relation contrariée

Jean CHICHE
Flora CHANVRIL
Cevipof[1]

Cet article met en évidence la défiance du grand public dans les indicateurs de la statistique publique et cherche à en déterminer ses ressorts. La 7e vague du baromètre de la confiance du Cevipof, enquête administrée par l'institut OpinionWay en décembre 2015 auprès de 2064 personnes, en situe le niveau par rapport à d'autres dimensions ou institutions. Cette enquête, en utilisant l'analyse des données et des modèles de régression logistique, établit le lien très fort entre défiance dans les chiffres de la statistique publique et défiance dans les institutions politiques. Les indicateurs comme le taux de chômage, de croissance, d'inflation, de la délinquance ou de l'immigration sont considérés, pour la population, comme liés au pouvoir politique, même si l'Insee a une très bonne image. Cette défiance touche surtout la classe moyenne inférieure et s'explique aussi par l'orientation politique et les choix électoraux. Elle influe sur l'idée que la population française se fait du fonctionnement de la démocratie.

1 Confiance dans les statistiques publiques

La confiance est un concept qui fait l'objet d'une vaste littérature tant en sciences sociales qu'en économie ou en philosophie. L'une des définitions simples, tirée du Larousse, est « Sentiment d'assurance, de sécurité qu'inspire au public la stabilité des affaires, de la situation politique ». Les indicateurs statistiques établis très régulièrement par les instituts de statistique publique, l'Insee en premier lieu, sont très largement repris et diffusés par tous les media, y compris les réseaux sociaux. Ils sont souvent objets de débat et les questions qu'ils mesurent constituent depuis longtemps des enjeux politiques de premier plan. Rappelons que le taux d'inflation dans les années 70 était au cœur de la politique économique et sociale de l'exécutif de l'époque dirigé par V. Giscard d'Estaing et R. Barre, et sa mesure était un motif de conflit permanent entre majorité et opposition de l'époque. La progression du Front National est liée à son discours sur l'insécurité et sur l'immigration et les indicateurs statistiques qui les mesurent ne peuvent que faire débat. Depuis 2012, François Hollande a fait de l'inversion de la courbe du chômage l'alpha et l'oméga de sa politique avec pour conséquence personnelle pour lui son éventuelle candidature à un nouveau mandat en 2017. Ce sont des indicateurs qui influent sur le fonctionnement même de la vie sociale, économique et politique du pays. Dans son

1. Centre d'études de la vie politique en France – Fondation des Sciences Politiques.
Adresses email des auteurs : jean.chiche@sciencespo.fr ; flora.chanvril@sciencespo.fr

ouvrage de référence, Laurent (2012) propose une définition de la confiance qui correspond bien aux propos que nous souhaitons développer : « La confiance est une expérience de fiabilité dans des conditions humaines qui suppose un rapport à un autre être humain, rapport qui peut être médiatisé dans une norme collective éventuellement incarnée dans une institution auquel cas la confiance repose sur cette norme ». Les instituts de statistiques ne sont certes pas des institutions au même titre que l'Assemblée nationale, mais l'Insee est très connu et parfaitement considéré par le public qui en a une excellente image et reconnaît sa qualité en tant que producteur de données.

Selon le Littré : « La méfiance fait qu'on ne se fie pas du tout ; la défiance fait qu'on ne se fie qu'avec précaution. Le défiant craint d'être trompé ; le méfiant croit qu'il sera trompé. La méfiance ne permettrait pas à un homme de confier ses affaires à qui que ce soit ; la défiance peut lui faire faire un bon choix. ». La défiance correspond donc plus à une sorte de méfiance raisonnée, distanciée : le méfiant ne fait pas confiance ; le défiant peut accorder sa confiance à certaines conditions. Comme l'ont écrit Algan et Cahuc (2007) nous vivons dans une société de défiance et cette défiance, qui caractérise de plus en plus les relations sociales, menace aussi la statistique. C'est ce terme de défiance que nous privilégierons pour caractériser les non confiants en la statistique publique.

En ce qui concerne la statistique, Tassi (2015) pose la question fondamentale : « Comment établir et maintenir la confiance du grand public, partie prenante numéro 1, tout en respectant l'équilibre entre promesse de confidentialité et utilisation des données recueillies ? »

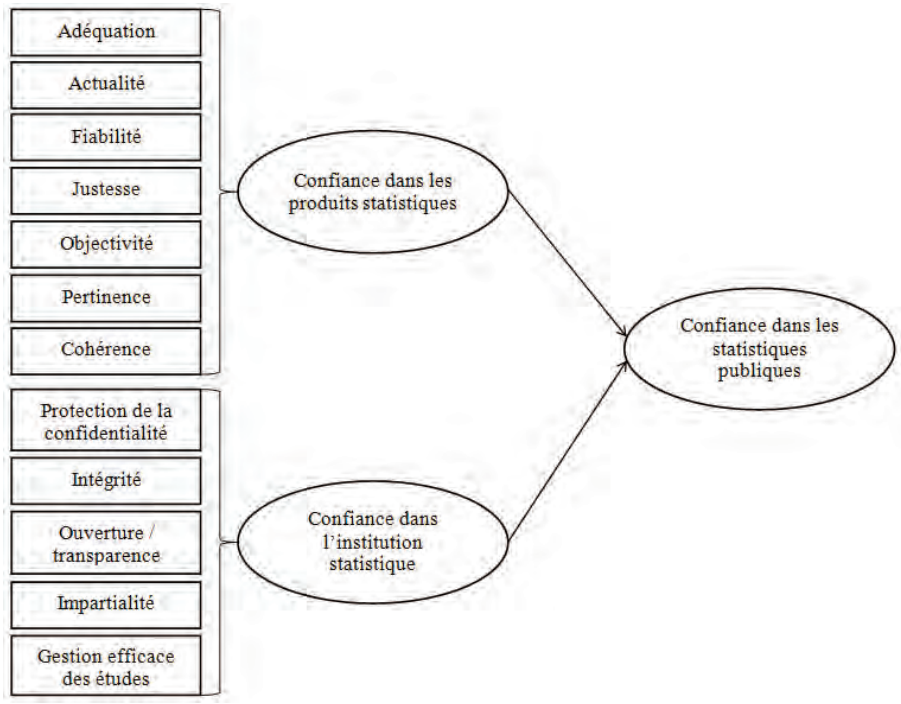
L'établissement de la confiance s'appuie sur deux conditions nécessaires et fondamentales : l'une est réglementaire, et montre que beaucoup d'États ont pris conscience depuis longtemps de la nécessité d'établir des garde-fous juridiques ; l'autre s'appuie sur la déontologie et l'éthique des acteurs. L'établissement de normes de qualité dans la production des données, notamment des enquêtes par sondages, est une condition nécessaire, mais pas suffisante, pour établir un lien de confiance avec le public. Pour le côté règlementaire et juridique, la Cnil doit assurer le respect des libertés individuelles face aux dangers de l'informatique. La déontologie et la qualité des données sont sous le contrôle en France du comité du label : « Le comité du label de la statistique publique régi par le chapitre IV du décret n° 2013-34 du 10 janvier 2013 examine les projets d'enquête que lui soumettent les services producteurs de la statistique publique ainsi que les projets d'exploitation, à des fins d'information générale, des données issues de l'activité des administrations, des organismes publics et des organismes privés chargés d'un service public ». En Europe, le « Code de bonnes pratiques de la statistique européenne » adopté par l'UE s'impose non seulement à Eurostat mais à tous les Instituts nationaux de statistique faisant partie du Système statistique Européen SSE. Pour les instituts privés il y a des organismes professionnels comme ESOMAR. Sans compter les organismes de contrôle comme le CESP, organisme indépendant qui audite les études concernant les mesures d'audience ou la commission des sondages pour tout ce qui est sondage politique en période électorale.

La littérature internationale portant sur la mesure de la confiance de la population dans les statistiques publiques est assez peu fournie. Toutefois en juin 2008 l'OCDE, lors de son comité annuel consacré aux statistiques, a débattu de la question : « comment mesurer la confiance dans les statistiques officielles ». Et à la suite de ce débat cette organisation a été à l'initiative d'un questionnaire international (2011) qui doit être, ou a déjà été administré dans plusieurs pays (Italie, Norvège, Luxembourg, etc.) et a débouché sur quelques publications dont l'article de Alegrezza (2014) portant sur le Luxembourg.

L'OCDE a produit un schéma d'hypothèses explicitant les différents critères amenant à la confiance dans les statistiques publiques. Ce schéma constitue la base théorique d'un

questionnaire construit autour des rubriques décrites dans le graphique 1.
Il préconise de distinguer les qualités attendues lors d'une production de données et celles d'indépendance et d'intégrité rattachées à l'institution elle-même.

Graphique 1 : Schéma de mesure de la confiance dans les statistiques publiques



Il existe aussi dans quelques pays, dont la France, des enquêtes qui mesurent l'image de l'institut public national. Le baromètre de l'image de l'Insee administré par CSA (2015) en est un exemple. Nous y reviendrons plus loin.

Nous avons souhaité mesurer le niveau de confiance que les Français accordent aux principaux indicateurs de la statistique publique et avons sélectionné à cet effet des questions qui font débat, qui sont dans le champ de la controverse économique et politique, et qui sont censées orienter les politiques publiques. Nous chercherons à dessiner les profils socioéconomiques et culturels des confiants/défiants, puis à saisir ce qui produit de la confiance/défiance dans ces indicateurs.

2 Enquête et premiers résultats

Nous avons introduit dans la 7e vague du baromètre de la confiance du Cevipof (2015) six items spécifiques mesurant la confiance dans les chiffres de la statistique publique.

Chiche et Chanvril (2014) avaient détaillé le mode de collecte et l'échantillonnage des précédentes vagues de ce baromètre. Cette 7e vague a été administrée auprès d'un échantillon de 2064 personnes (2069 après redressement sociodémographique et politique), représentatif de la population française âgée de 18 ans et plus et inscrite sur les listes électorales. L'échantillon a été constitué selon la méthode des quotas, au regard des critères de sexe, d'âge, de catégorie socioprofessionnelle, après stratification par région de résidence et taille de commune. L'échantillon a été interrogé en ligne sur système Cawi (Computer Assisted Web Interview). Les interviews ont été réalisées du 17 au 28 décembre 2015. Les dates de terrain ont dû être décalées de quelques semaines en raison des attentats terroristes qui ont touché Paris le 13 novembre, alors que le terrain devait démarrer le soir même. Le terrain a été confié à l'institut de

sondages privé OpinionWay et a été réalisé en appliquant les procédures et règles de la norme ISO 20252. Nous avons ajouté au questionnaire socio politique habituel, dont les questions et les principaux résultats sont disponibles sur le site du Cevipof (2015), la question suivante : « *Personnellement, avez-vous très confiance, plutôt confiance, plutôt pas confiance ou pas du tout confiance dans les statistiques officielles en France, par exemple dans… ?* ». S'en suivait une batterie de six items. Les résultats figurent dans le tableau 1.

Tous les indicateurs recueillent un niveau de défiance très élevé. Et si les chiffres de la hausse des prix sont ressentis un peu moins mal que ceux des autres indicateurs, il n'en reste pas moins que la confiance que les Français lui accordent est très basse (38% de confiance). La défiance maximale s'exprime sur les chiffres du chômage, de la délinquance et de l'immigration (70% de défiance). Tous des thèmes qui clivent aujourd'hui très douloureusement la société française.

Si on compare ces taux de défiance aux autres questions portant sur les institutions politiques, publiques, qu'elles soient régaliennes ou pas, privées ou internationales, il faut bien se rendre à l'évidence : les indicateurs statistiques obtiennent des résultats parmi les pires. Bien entendu comparer des taux de confiance dans des résultats statistiques à des taux de confiance dans des institutions doit être fait avec prudence. Il aurait été plus correct de comparer la confiance dans l'institut public de statistiques – ici l'Insee – avec les autres items mesurés. Mais cela n'a pas été possible pour des raisons de longueur du questionnaire : la négociation qui précède tout allongement du questionnaire barométrique ne se fait pas sans douleur et compromis. L'OCDE (2009) avait pourtant préconisé d'intégrer cette question à la mesure de la confiance des statistiques publiques. Pour nous rassurer, l'étude Luxembourgeoise déjà citée montre qu'il n'y a pas de corrélation entre confiance dans STATEC (l'institut public local) et confiance dans les indicateurs socioéconomiques.

En France, afin d'améliorer le service qu'il rend à ses utilisateurs, l'INSEE effectue régulièrement des enquêtes de satisfaction sur son image et les données qu'il produit. La dernière enquête a été effectuée en mai 2015 auprès du « Grand public » : l'institut français est parfaitement connu (91%), a une excellente image (71%), ses missions sont bien comprises à 63%, mais 55% des personnes interrogées n'ont pas confiance dans les chiffres et données publiés.

Si les institutions régaliennes ou de santé sont toujours très appréciées par la population, les institutions politiques, syndicales, médiatiques sont, comme les indicateurs statistiques, déconsidérés. Dans une société en crise, le tableau 1 montre bien qu'après les attentats du 13 novembre 2015, le pays est en recherche de protection et de soin. Hôpitaux, armée, police, sécurité sociale, associations ont un niveau de confiance supérieur à 50%.

Il y a de très fortes corrélations entre les 6 items mesurés (voir le tableau 2). Déficit et croissance sont liés à 0,71, comme la confiance dans les indicateurs d'immigration et de délinquance. Une analyse en composantes principales nous a permis de construire une variable globale de confiance dans la statistique publique. Nous l'avons dichotomisée en une modalité de confiants (39% de notre échantillon) et une modalité de défiants (61%). Cette nouvelle variable, résumé des 6 indicateurs initiaux, servira de variable dépendante de synthèse dans les modèles logistiques que nous présentons.

La statistique n'est pas en dehors du champ social. Le tableau 3 des corrélations en atteste. La défiance dans les indicateurs statistiques est fortement corrélée aux institutions politiques (assemblée, président, gouvernement) et un peu moins, voire pas du tout, aux institutions non régaliennes. Elle est un des éléments qui permettent la construction d'un espace cognitif. Les media jouent un rôle prépondérant dans la perception que le public a de la fiabilité des résultats qu'elle produit mais la corrélation entre confiance dans les media et échelle de défiance dans

les statistiques publiques vaut 0,25. La défiance dans les media ne suffit donc pas à expliquer la défiance dans les statistiques.

Tableau 1 : La confiance dans les institutions comparée à la confiance dans les indicateurs statistiques

	BASE	ST Confiance	ST Pas confiance	Très confiance	Plutôt confiance	Plutôt pas confiance	Pas du tout confiance	No rep.
Les hôpitaux	2069	82%	17%	16%	66%	13%	4%	1%
L'armée	2069	81%	18%	24%	57%	12%	6%	1%
Les petites et moyennes entreprises	2069	80%	19%	12%	68%	15%	4%	1%
La police	2069	75%	24%	15%	60%	18%	6%	1%
L'école	2069	69%	30%	11%	58%	23%	7%	1%
Les associations	2069	66%	33%	8%	58%	23%	10%	1%
La sécurité sociale	2069	62%	37%	6%	56%	29%	8%	1%
L'Eglise catholique	2069	49%	48%	8%	41%	31%	17%	3%
Les grandes entreprises publiques	2069	46%	52%	3%	43%	39%	13%	2%
La justice	2069	44%	55%	4%	40%	38%	17%	1%
Les grandes entreprises privées	2069	43%	55%	4%	39%	38%	17%	2%
L'Assemblée Nationale	2069	41%	57%	3%	38%	41%	16%	2%
L'Union Européenne	2069	38%	61%	3%	35%	39%	22%	1%
L'institution présidentielle	2069	35%	63%	3%	32%	38%	25%	2%
Les chiffres de la hausse des prix	2069	38%	60%	4%	34%	40%	20%	2%
L'institution présidentielle	2069	35%	63%	3%	32%	38%	25%	2%
Les chiffres de la croissance économique	2069	36%	63%	3%	33%	44%	19%	1%
Les chiffres des déficits publics	2069	32%	66%	3%	29%	43%	23%	2%
Les chiffres de l'immigration	2069	29%	69%	3%	26%	40%	29%	2%
Les chiffres du chômage	2069	28%	70%	3%	25%	43%	27%	2%
Les chiffres de la délinquance	2069	29%	70%	3%	26%	44%	26%	1%
Les banques	2069	29%	70%	2%	27%	43%	27%	1%
Le gouvernement	2069	29%	70%	2%	27%	38%	32%	1%
L'Organisation Mondiale du Commerce	2069	26%	71%	2%	24%	48%	23%	3%
Les grandes conférences internationales,	2069	26%	72%	1%	25%	45%	27%	2%
Les syndicats	2069	27%	71%	2%	25%	38%	33%	2%
Les medias	2069	24%	75%	1%	23%	48%	27%	1%
Les partis politiques	2069	12%	87%	1%	11%	47%	40%	1%

Source : Vague 7 du Baromètre de la confiance. Cevipof 2015

Tableau 2 : Matrice de corrélations entre les toutes les mesures de la confiance

	Chômage	Croissance économique	Immigration	Déficits publics	Délinquance	Hausse des prix
Chômage	1					
Croissance économique	0.66	1				
Immigration	0.65	0.62	1			
Déficits publics	0.66	0.71	0.62	1		
Délinquance	0.68	0.68	0.71	0.63	1	
Hausse des prix	0.61	0.6	0.57	0.61	0.62	1

Source : Vague 7 du Baromètre de la confiance. Cevipof 2015

Tableau 3 : Matrice de corrélations entre les confiances dans les indicateurs

	Assemblée	Gouvernement	Président	UE	OMC	Partis	Hôpitaux	Medias	Banques	Police
Défiance dans les statistiques	0.368**	0.421**	0.419**	0.316**	0.209**	0.287**	0.119**	0.248**	0.178**	0.093**
N	2023	2031	2028	2040	1997	2027	2042	2039	2042	2035

*Note : * p < 0.05, ** p < 0.01, *** p < 0.001*
Source : Vague 7 du Baromètre de la confiance. Cevipof 2015

3 Modèles de la défiance

Nous avons construit une première série de six modèles LOGIT, à partir des données de l'enquête, en testant la confiance/défiance dans les six facettes des statistiques publiques (chômage, croissance économique, immigration, déficits publics, délinquance, hausse des prix) avec pour variables explicatives des éléments socioéconomiques et culturels (le genre, l'âge, le niveau de diplôme, les revenus par unité de consommation) et politiques (le positionnement avec l'échelle gauche / droite et le choix électoral en utilisant la reconstitution du vote au second tour de l'élection présidentielle de 2012). Les résultats des modèles sont présentés dans le tableau 4.

Les quatre indicateurs d'ordre économique présentent des logiques explicatives en partie comparables. On peut distinguer les chiffres mesurant les taux de chômage et de la hausse des prix d'une part, avec des explications essentiellement socioéconomiques, des chiffres de la croissance et des déficits publics d'autre part avec des profils à la fois socioéconomiques, liés au sexe, et politiques. Hausse des prix et chômage peuvent toucher chacun (enjeux egotropiques), alors que déficits publics et taux de croissance sont d'ordre plus général. Ils touchent l'ensemble de la société (enjeux sociotropiques).

Du point de vue socio culturel, l'âge et le niveau de diplôme ont des effets significatifs relativement proches quant à la défiance envers ces quatre indicateurs : les diplômés du baccalauréat, ainsi que des filières professionnelles pour la croissance économique, les déficits publics et la hausse des prix, sont davantage défiants ; les plus jeunes (gap à 35 ou 50 ans selon la mesure considérée) le sont moins. De faibles revenus (calculés par unité de consommation du ménage) renforcent également la défiance envers ces chiffres de manière significative, hormis pour la croissance économique où cette variable n'a pas d'effet. Le fait d'être un homme plutôt qu'une femme diminue cette défiance pour les deux chiffres se rapportant à des enjeux économiques plus sociotropiques. Sur le plan politique, le fait de se situer à gauche pour le chômage, au centre pour la hausse des prix, diminue la défiance envers les chiffres considérés.

En ce qui concerne la croissance économique et les déficits publics, le positionnement sur l'échelle gauche / droite explique de manière plus complète la défiance envers les indicateurs concernés : à gauche, voire au centre, on est moins défiant ; très à droite on est plus défiant. Peut-être trouve-t-on ici un effet thermostatique, c'est-à-dire que les partisans de la majorité politique en place sont toujours moins pessimistes ou défiants que les tenants de l'opposition. Le modèle d'explicitation de la défiance dans l'indicateur du taux de délinquance fait émerger trois variables explicatives : l'âge (les 25-34 ans sont moins défiants), le niveau de diplôme (les diplômés du baccalauréat et des filières professionnelles sont plus défiants) et le positionnement sur l'échelle gauche / droite (se situer du centre au très à gauche diminue la défiance).

Enfin le modèle sur les données de l'immigration est le plus politisé des six. Parmi les variables socioculturelles, seul le niveau de diplôme présente des effets significatifs : les moins diplômés sont plus défiants. L'orientation politique joue un rôle important : se dire très à gauche, à gauche ou au centre rend moins défiant ; par contre être très à droite rend plus défiant. Le vote au second tour de l'élection présidentielle de 2012 trouve ici son seul effet significatif : avoir voté pour Nicolas Sarkozy augmente la défiance envers les chiffres de l'immigration. Au second tour N. Sarkozy avait reçu une part importante du vote en faveur de Marine Le Pen du premier tour. Les six modèles explicatifs présentent des différences pouvant être caractérisées ainsi : des variables socioéconomiques prédominantes pour les deux indicateurs économiques egotropiques que sont le chômage et la hausse des prix ; des modèles plus équilibrés entre explications sociodémographiques et politiques pour les indicateurs sociotropiques que sont la croissance économique, les déficits publics ou la délinquance ; un modèle très politisé pour les chiffres de l'immigration.

Malgré ces différences, une trame de fond se dessine : lorsqu'ils sont significatifs, les effets des variables explicatives dans les différents modèles vont dans le même sens, vers toujours plus de défiance. Ce point d'ancrage des six modèles ainsi que l'homogénéité des six items mesurés plus haut avec l'ACP justifient d'établir un modèle d'explication plus global pour la défiance envers les statistiques publiques. Dans ce modèle global : les 25-24 ans ainsi que les personnes se situant politiquement de l'extrême gauche au centre sont moins défiantes envers les statistiques publiques ; les diplômés de filières professionnelles ou du baccalauréat, les revenus plutôt bas ainsi que ceux se situant très à droite sur l'échelle gauche / droite sont plus défiants envers ces mêmes statistiques publiques.

Tableau 4 : Modèles de régressions logistiques de la défiance dans les statistiques publiques

		Chômage	Croissance économique	Immigration	Déficits publics	Délinquance	Hausse des prix	Défiance stats
Genre	Un homme	-0,092	-0.237*	-0.058	-0.304**	0,085	0,040	-0.162
	Une femme	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Âge	18-24 ans	-0.697***	-0.433*	-0.378	-0.343	-0.317	-0.785***	-0.366
	25-34 ans	-0.464*	-0.372*	-0.286	-0.600***	-0.664***	-0.778***	-0.655***
	35-49 ans	-0.245	-0.266	-0.188	-0.346*	-0.311	-0.539***	-0.28
	50-64 ans	-0.06	-0.111	0.062	-0.148	-0.052	-0.171	-0.137
	65 ans en +	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Niveau de diplôme	Sans diplôme / CEP / BEPC	0.077	0.107	0.471*	0.195	0.129	0.306	0.279
	CAP / BEP	0.320	0.588***	0.662***	0.658***	0.518**	0.539***	0.571***
	Baccalauréat	0.348*	0.405**	0.545***	0.355*	0.412**	0.435**	0.358*
	Bac+2 (DEUG, DUT, BTS)	0.146	0.030	0.168	0.104	0.151	0.259	0.102
	Supérieur à Bac +2	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Revenus par unité de consommation	<= 1027,78	0.243	0.316	0.321	0.417*	0.113	0.444**	0.276
	1027,79 - 1472,22	0.454**	0.240	0.309	0.619***	0.162	0.427**	0.33*
	1472,23 - 2150,00	0.185	0.222	0.260	0.425**	0.121	0.211	0.177
	2150,01+	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Échelle gauche / droite	Très à gauche	-0.212	-0.642**	-0.769**	-0.569*	-0.619*	-0.341	-0.519*
	A gauche	-0.992***	-1.174***	-1.085***	-0.962***	-1.040***	-0.783***	-1.15***
	Au centre	-0.298	-0.578***	-0.697***	-0.302	-0.522**	-0.550**	-0.65***
	A droite	0.211	-0.144	0.056	0.056	-0.075	-0.127	-0.031
	Très à droite	0.407	0.546*	0.628*	0.453*	0.370	0.199	0.581*
	Ni à gauche ni à droite	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Vote T2 présidentielle 2012	François Hollande	-0,049	-0,152	-0.306	-0.292	-0,268	-0.184	-0.263
	Nicolas Sarkozy	0.197	0.017	0.499**	0.014	0,183	0.061	0.214
	Blancs / Nuls / Abstention	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Constante		0.975***	0.928***	0.866***	0.923***	1.228***	0.695***	1.084***
R^2 de Nagelkerke		0.113	0.137	0.199	0,144	0.129	0.100	0.168

*Note : base=2064, * p < 0.05, ** p < 0.01, *** p < 0.001*
Source : Vague 7 du Baromètre de la confiance. Cevipof 2015

4 Comparaison internationale

Pour situer la France par rapport aux autres pays européens, nous disposons des données des Eurobaromètres de 2009 et 2015. Ces deux enquêtes posent la question suivante : « Personnellement, dans quelle mesure faites-vous confiance aux statistiques officielles en (NOTRE PAYS), par exemple les statistiques sur le chômage, l'inflation ou la croissance économique ? Diriez-vous que vous avez plutôt confiance ou plutôt pas confiance en ces statistiques officielles ? ».

Nous constatons que la France reste parmi les pays les plus défiants envers la statistique

officielle (26ème en 2009, 27ème en 2015). Les pays les plus confiants sont les pays nordiques et les Pays-Bas ; les plus défiants sont les pays du Sud et de l'Est, ainsi que l'Allemagne. En moyenne la défiance reste majoritaire (44% de confiance dans l'Europe des 27 puis des 28) mais cache certaines disparités entre les pays. Les plus confiants dans les statistiques publiques se situent entre 20 et 30 points au-dessus de cette moyenne tandis que les plus défiants, en dehors de l'Espagne, sont proches de cette moyenne. Notons un recul de la confiance particulièrement marqué à Chypre, en Espagne, en Slovaquie, en Grèce, en Autriche et en Belgique, pays dans lesquels la crise économique et/ou politique a été particulièrement forte. Inversement Malte et le Royaume-Uni (avant le Brexit) voient les niveaux de confiance augmenter fortement.

Tableau 5 : Évolution de la confiance dans la statistique officielle entre 2009 et 2015

	2015	2009	Écart 2015-2009
Espagne	27	43	-16
France	38	40	-2
Allemagne	39	36	3
Hongrie	39	37	2
Slovénie	39	45	-6
Chypre	39	61	-22
Autriche	42	53	-11
Slovaquie	42	55	-13
Royaume uni	44	33	11
Europe 27 (28 en 2015)	44	44	0
Pologne	44	45	-1
Croatie	44	NP	-
Italie	45	42	3
Grèce	45	56	-11
Lettonie	46	47	-1
Belgique	48	58	-10
Rep.Tchèque	49	54	-5
Bulgarie	51	47	4
Lituanie	51	51	0
Portugal	51	52	-1
Roumanie	52	50	2
Estonie	54	51	3
Irlande	58	57	1
Luxembourg	59	68	-9
Malte	62	43	19
Danemark	68	70	-2
Finlande	72	67	5
Pays-Bas	72	69	3
Suède	73	70	3

Source : EB standard 83 (2015) et EB special 323 (2009)

5 Confiance et démocratie

Un modèle classique en économie ou en sociologie politique, Boy et Chiche (2010), montre que l'état de l'économie, le bien-être, le niveau culturel produisent dans la société un niveau de confiance qui permet l'épanouissement de la démocratie et par là même son bon fonctionnement. A contrario la défiance nuit à son bon fonctionnement. Les indicateurs de la statistique publique jouent un rôle important dans le débat public et par là même la défiance ressentie à leur égard influe sur ce que la population pense de l'état de la démocratie en France. Le tableau 6 illustre parfaitement cet état de fait. Nous obtenons ainsi 31 points d'écart entre les confiants qui pensent que la démocratie fonctionne bien et les défiants mécontents de son fonctionnement.

La France est en crise. Ce tableau en est une preuve de plus.

Tableau 6 : Confiance dans les statistiques publiques et fonctionnement de la démocratie en France

	BASE	ST Bien	ST Mal	Très bien	Assez bien	Pas très bien	Pas bien du tout	NSP
Confiance	2069	54%	45%	4%	49%	36%	9%	1%
Défiance	2069	22%	76%	1%	21%	46%	30%	1%
Total	2069	32%	66%	2%	30%	43%	23%	1%

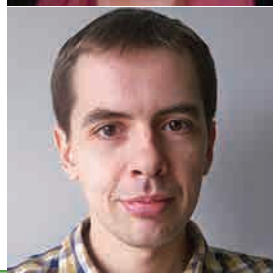
Source : Vague 7 du Baromètre de la confiance. Cevipof 2015

Notre étude ne remet pas en question le travail produit et diffusé par l'INSEE, mais montre que la population en état de crise généralisée de confiance est sceptique quant aux indicateurs et attribue à une forme de communication politique les commentaires qui en sont faits. Le Baromètre de l'image de l'INSEE (2015) montre que 52% des Français n'ayant pas confiance dans les chiffres et les données publiés sur la situation économique et sociale de la France citent pour première raison que « Les hommes politiques leur font dire ce qu'ils veulent ».
François Hollande, en ne se présentant pas pour un nouveau mandat présidentiel, honore sa promesse basée sur l'inversion durable de la courbe du chômage. Il peut peut-être à travers cette décision participer à rendre les Français moins défiants envers la statistique publique.
Les comparaisons internationales montrent également que si la France se situe parmi les pays les plus défiants, la confiance dans les statistiques publiques est forte là où l'économie est florissante et/ou le pouvoir politique est stable.

Références

[1] Laurent, E. (2012), Économie de la confiance, La Découverte, Paris.
[2] Algan, Y. et Cahuc, P. (2007), La société de défiance : comment le modèle social s'autodétruit, Éditions de la rue d'Ulm « Cepremap », Paris.
[3] Tassi, P. (2015), La technologie au service de la confiance ?, Colloque Individus, Données et Société connectée, Variances, 53, Paris.
[4] OCDE (2011), Measuring trust in Official Statistics ; Cognitive Testing, Report to the OEDC of the electronic working groupon measuring trust in official statistics. http://www.oecd.org/std/50027008.pdf
[5] OCDE Directorate (2012), Measuring trust in official statistics : New model on OEDC questionnaire. http://www.oecd.org/std/Stat%20newsletter%20final%20for%20the%20web.pdf
[6] Allegrezza, S. (2014), La confiance dans les statistiques publiques : les déterminants de la confiance, de l'indépendance politique et de l'utilisation des données du STATEC, Économie et Statistiques : Working papers in STATEC, 74, Luxembourg.
[7] CSA (2015), Baromètre de l'image de l'Insee : Enquête auprès du grand public. http://www.insee.fr/fr/insee-statistique-publique/default.asp?page=connaitre/enquetes/enquetes-satisfaction-grand-public-mai-2015.htm
[8] Chiche, J. et Chanvril, F. (2014), Confiance en politique ; méthode et analyse, 8e colloque francophone sur les sondages, Dijon. http://paperssondages14.sfds.asso.fr/submission_58.pdf
[9] CEVIPOF (2015), Le baromètre de la confiance politique. http://www.cevipof.com/fr/le-barometre-de-la-confiance-politique-du-cevipof/
[10] Boy, D. et Chiche, J. (2010), Confiances dans Boy, D. et Cautres, B. et Sauger, N. (2010), Français des européens comme les autres, Presses de Sciences Po, 45-71, Paris.

L'usage comparé des statistiques par Gabriel Tarde et Emile Durkheim



Hélène ŒHMICHEN
Oleksii VIEDROV
Étudiants en master 2 « Sociologie et Statistique »
(EHESS/ENS/ENSAE)

Le débat entre Gabriel Tarde et Emile Durkheim à la fin du XIX^{ème} siècle est souvent présenté comme fondamental pour la création de la sociologie en France. C'est au terme de ce débat que Durkheim, en sortant vainqueur, a pu se positionner pour la génération suivante comme père fondateur de la sociologie, puis qu'il a été redécouvert comme tel, dans les années 1970. Cette victoire de l'un sur l'autre se ressent encore aujourd'hui dans notre usage des statistiques dans la discipline : comparaison de phénomènes statiques, « variations concomitantes », moyennes et opérations qui en découlent. Or la redécouverte plus ou moins récente de Tarde et de son usage différent des statistiques réactualise depuis une quinzaine d'années les termes de ce débat fondateur. Retracer les différences, les ressemblances et les oppositions entre les deux théoriciens dans le domaine de la statistique présente donc un intérêt historique pour la discipline – qui nous permet également de comprendre la sociologie actuelle, ses différentes mouvances et ses enjeux, ainsi qu'un intérêt épistémologique. Nous exposerons tout d'abord le parcours des deux auteurs et tenterons de replacer leur débat dans le contexte historique et universitaire de l'époque, afin d'en comprendre les origines, les logiques et les conséquences. Nous examinerons ensuite plus en détail les différences dans l'usage et la définition des outils statistiques, et les conséquences épistémologiques que cela entraîne pour chacun d'eux, avant enfin de dresser un bref aperçu de la réception de ces usages débattus des statistiques.

I.Termes et lieux du débat Tarde-Durkheim

G. Tarde et E. Durkheim, et le débat dont Durkheim est sorti vainqueur pour être souvent présenté comme fondateur de la sociologie moderne en France, ne doivent pas être appréhendés comme un commencement absolu, dont la création intellectuelle est apparue *ex nihilo*. Deux séries de recherches ont permis de lever l'illusion d'isolement des « pères fondateurs » : une étude des sources et des influences des deux universitaires, et une étude du champ intellectuel dans lequel chacun tentait d'imposer ses problématiques. Afin de comprendre au mieux les ressorts de leur controverse et sa traduction dans les statistiques, nous commencerons par dresser un rapide tableau du champ intellectuel de l'époque, puis nous retracerons leurs parcours

et leurs influences, avant de présenter leurs œuvres et leurs théories générales, ainsi que les fondements de leurs oppositions.

1. Contexte universitaire de la naissance de la sociologie

La sociologie comme science du social s'institutionnalise à la fin du XIX^{ème} siècle, en France comme en Allemagne, ainsi que dans de nombreux autres pays, et ce via des ouvrages majeurs qui s'en réclament, des revues, des chaires de faculté, des écoles et autres institutions. En plus de ces mouvements institutionnels, il faut prendre en considération que "*les sciences sociales sont terriblement à la mode [...]. C'est la tarte à la crème de toutes les réunions mondaines, de tous les discours, et nul n'a de l'esprit s'il n'est sociologue*" (Henri Hauser, historien, 1903). L'institutionnalisation de la IIIème République et les évolutions sociales qui l'accompagnent ont un impact fort sur la création et les mutations en cours de la discipline, et donc sur les formes, les enjeux et la réception du débat entre les deux sociologues. Ainsi, de lourds efforts sont faits pour la consolidation de l'Instruction publique, avec l'idée qu'elle permettrait l'émergence d'une communauté morale par la transmission d'un savoir irrigué par la recherche. A ce titre, les sciences sociales ne sont pas négligées (forte présence au ministère de l'Instruction publique[1]). Elles sont par ailleurs portées par l'initiative privée de certains mécènes et grands bourgeois (Cuin et Gresle, 1992)

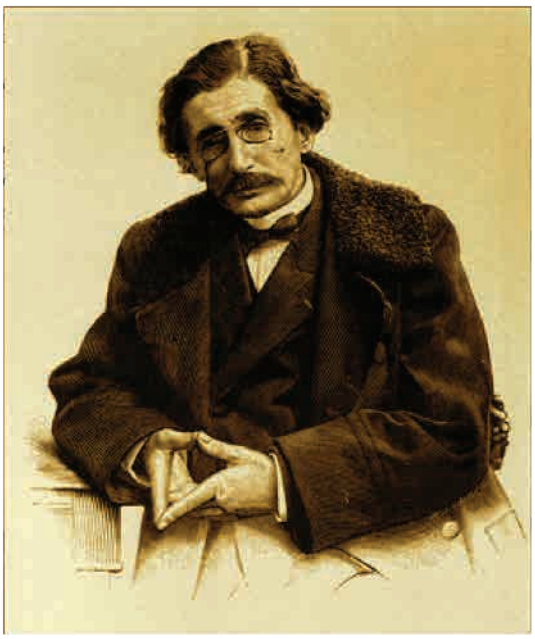
Dans les *Règles de la Méthode sociologique* (1894), Durkheim identifie plusieurs groupes qui dominent le champ intellectuel français en sociologie dans les années 1870. Pour présenter ce contexte, nous pouvons partir de ses observations directes et de ses catégorisations, à la lumière de travaux d'historiens et de sociologues sur le sujet (Mucchielli et Borlandi, 1995, Borlandi et al, 2000, Cuin et Gresle, 1992).

Le premier groupe est **l'anthropologie physique** dont Paul Broca (1824-1880) est la figure de proue. Pour le fondateur de la Société d'anthropologie de Paris, relayée par la *Revue d'anthropologie* (1872), les différences et les inégalités entre les peuples s'expliquent par le déterminisme de chacune des grandes « races » (Blancs, Jaunes, Noirs) de la planète. Sur le plan individuel, les différences de comportements sont reliées aux configurations du cerveau (« crâniologie »). Intellectuellement, le paradigme dominant de ce groupe est donc la complète détermination du social par le biologique. A la même époque, dans le « **groupe criminologiste** », l'Italien Cesare Lombroso (1835-1909) défend la thèse du « criminel né » selon laquelle la criminalité s'explique par des causes héréditaires. Le « milieu social » n'est qu'un révélateur de ces prédispositions biologiques. De son côté, Gustave Le Bon (1841-1931) professe une théorie raciste et inégalitaire de la psychologie des peuples. Le « **groupe universitaire** » enfin, identifié par Durkheim, « comprend les sociologues qui appartiennent à l'université », à savoir « des professeurs de philosophie ». En France, la sociologie cherche notamment à se démarquer de l'emprise universitaire de la philosophie et de l'Histoire : le cadre universitaire et ses modalités de fonctionnement ont été établis bien avant que ne cherche à s'institutionnaliser la nouvelle discipline. Le climat politique lui est certes favorable, puisque dans ces débuts de la Troisième République, elle a bien les préoccupations morales et civiques qui peuvent avoir des retombées psychologiques. Mais l'organisation de l'appareil scolaire et universitaire a pour conséquence que, comme la géographie ou la science de l'éducation, elle s'implante d'abord sous le couvert de la philosophie et de l'histoire traditionnelle.

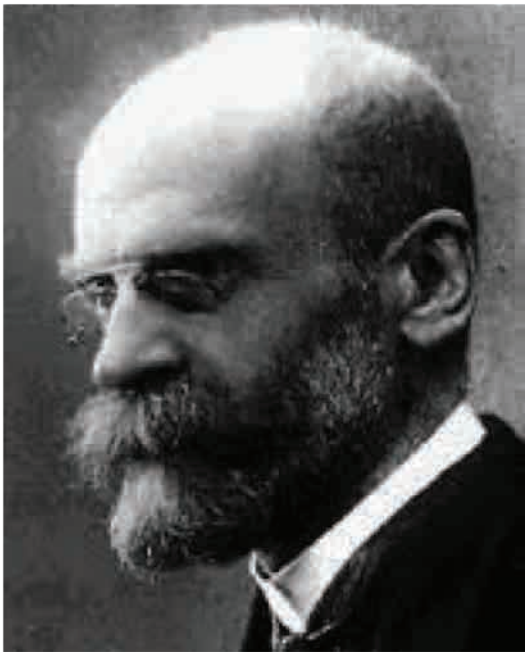
A partir des années 1880, s'éloignant de la philosophie mais aussi du « scientisme positiviste », Gabriel Tarde (1843-1904), puis René Worms (1867-1926), et enfin Emile Durkheim (1958-1917) contestent l'hégémonie du naturalisme, s'attachent à montrer la part du social dans les

1. Pour en savoir plus, voir A. Prost, *L'Enseignement en France*, 1968.

conduites humaines, parlent au nom de la sociologie naissante et prétendent la représenter. Seul le dernier a fait école, suite à un débat qui l'opposa principalement à Tarde.



(1) Gabriel Tarde (1843-1904)



(2) Emile Durkheim (1858-1917)

2. Parcours biographiques et universitaires

Gabriel Tarde a une formation de magistrat – et exerce jusqu'à sa mort sa fonction de juge, qu'il commence à Sarlat. Né dans une famille de juristes et de savants, il a été formé par les jésuites. Parallèlement à cette carrière, il est en relation avec l'école criminaliste de Cesare Lombroso, mais s'en démarque rapidement, préférant une explication psychologique et sociologique de la criminalité à une explication physique. En 1887, il débute sa collaboration aux Archives de l'anthropologie criminelle, revue fondée par son ami, le docteur Alexandre Lacassagne, médecin-légiste à Lyon. Il publie en 1890 *Les Lois de l'imitation*, qui rend compte des comportements sociaux par des tendances psychologiques individuelles, et *La Philosophie pénale*, ce qui le rend célèbre, puis d'autres ouvrages à succès tels que *La Logique sociale* (1895) ou *L'Opinion et la foule* (1901). Fin 1893, le ministre de la justice demande à Gabriel Tarde de rédiger un mémoire sur l'organisation de la statistique criminelle en France. L'année suivante, en janvier, il quitte Sarlat pour s'installer à Paris, au poste de directeur de la statistique judiciaire au ministère de la justice, où il est chargé de faire la synthèse des chiffres de la justice criminelle, civile et commerciale française depuis 1891. Gabriel Tarde publie plusieurs ouvrages et de nombreux articles, collabore à la Société Générale des prisons, enseigne la sociologie dans différentes écoles, en particulier à l'Ecole libre des sciences politiques et au Collège libre des sciences sociales. Il bénéficie de son vivant d'une audience internationale et de grands succès éditoriaux, ce qui lui vaut d'être élu en 1900 au Collège de France, à la chaire de philosophie moderne. La même année, il devient membre de l'Institut (Académie des sciences morales et politiques). Mais contrairement à son adversaire, il n'a élaboré aucune réelle doctrine universitaire ; n'ayant pas constitué d'école, son œuvre n'est poursuivie par aucun successeur après sa mort.

G. Tarde s'oppose fortement au positivisme de l'époque, et aux théories biologisantes de Lombroso. Sociologue autodidacte, membre d'un groupe marginal composé de personnalités isolées regroupées autour de R. Worms et de sa *Revue internationale de sociologie* (1893), il est influencé par l'idéalisme épistémologique et critique du philosophe Charles Renouvier, qui débouche sur la « critique de la science », dans la ligne de Claude Bernard à Augustin Cournot.

Issu d'une famille de rabbins, Emile Durkheim intègre en 1879 l'Ecole Normale Supérieure, où il côtoie Henri Bergson et Jean Jaurès. Il reçoit une formation de philosophe (agrégation de

philosophie en 1882) mais se lance très tôt dans la sociologie, en publiant par exemple à partir de 1885 des comptes rendus dans la *Revue philosophique* qui concernent tous exclusivement la « science sociale » ou la « sociologie ». Dès cette époque, il souhaitait se consacrer à la constitution d'une science sociale capable de répondre au malaise social et surtout moral des sociétés européennes de la fin du XIXème siècle. Sa thèse, qui porte sur les rapports entre individualisme et socialisme (1893), lui assure d'emblée une autorité intellectuelle le plaçant comme principal représentant de la sociologie en France. Il obtient en 1887 une chaire de pédagogie et de science sociale à l'Université de Bordeaux, puis bénéficie de la première chaire de sociologie française, qui est créée en 1895. Il est désigné en 1902 pour occuper la chaire de science de l'éducation à la Sorbonne, poste sur lequel il sera titularisé en 1906. Il rédige en 1912 *Les Formes élémentaires de la vie religieuse*. Il réussit en 1913 à transformer le titre du poste qu'il occupe en chaire de sociologie. C'est à Bordeaux qu'il rédige *La Division du travail social* (sa thèse, 1893), *Les Règles de la méthode sociologique* (1894 et 1895) et *Le Suicide* (1897). C'est également à Bordeaux qu'il conçoit les cinq premiers volumes de *L'Année sociologique*, revue créée en 1898 qui lui permet de rassembler autour de lui plusieurs jeunes disciples qui assureront sa postérité, tels que Marcel Mauss, Maurice Halbwachs, Célestin Bouglé, François Simiand ou Paul Fauconnet. Comme le fait remarquer P. Besnard (1983), il eut très tôt conscience de la nécessité de travailler en équipe pour sortir la sociologie du stade de l'amateurisme préscientifique dans laquelle elle stagnait.

Durkheim est notamment influencé par deux de ses prédécesseurs que sont Auguste Comte et Herbert Spencer. Le premier voulait appliquer la méthode scientifique des sciences naturelles aux sciences sociales, et le second développa une approche utilitariste évolutionniste pour étudier la société humaine. Durkheim fut influencé par le positivisme de Comte. Toutefois, il en réclame une lecture nuancée, loin de l'orthodoxie de ses disciples antispiritualistes, et se sépare notamment de sa philosophie de l'Histoire et de sa doctrine du progrès. Durkheim est également influencé par Spencer, référence philosophique principale des années 1875-1895, qui, par ses analyses fonctionnalistes et ses analogies organiques, prenait ses distances par rapport au positivisme et a su réhabiliter la sociologie dans le champ philosophique universitaire. Mais il s'en distancie également à cause de ce qu'il considère comme un manque de méthode et une vision unilinéaire de l'évolution sociale.

(3) Durkheim chassant Tarde de l'Université (caricature de http://socio-bd.blogspot.fr/)

En deçà des oppositions théoriques, Tarde et Durkheim n'occupent donc pas la même place dans le champ intellectuel français : l'un cherche à institutionnaliser et à autonomiser la sociologie autour d'une « école » tandis que l'autre reste en lien avec la psychologie et la philosophie. L'un s'organise autour d'une revue avec des disciples tandis que l'autre jouit d'une grande popularité hors du champ universitaire, et acquiert par là sa chaire au collège de France.

3. Des épistémologies différentes : des conceptions différentes de la science qui doit être fondée

Nous reviendrons ici rapidement sur les termes du débat qui opposa les deux chercheurs, puis nous exposerons la place que cela a pris dans leurs usages et conceptions des statistiques.

Leurs oppositions (principalement entre 1893 et 1904) s'exprimaient notamment par articles interposés ; on résumera les différends et les rapprochements entre les deux auteurs sur deux points : holisme/individualisme méthodologique et définition du fait social.

Tarde est souvent présenté par opposition à Durkheim comme un pur individualiste qui ne conçoit pas la société dans son ensemble. Or la position de Tarde a évolué depuis ses premiers écrits où il mettait en avant la primauté de l'individu (Lubek, Apfelbaum, 1981). Sa perspective s'oriente très tôt vers un point de vue psycho-social. Dans Les Lois de l'imitation, livre considéré comme fondateur de la psychologie sociale, il présente la vie sociale comme une alternance entre les inventions et les imitations des individus. Un fait social n'a aucune réalité transcendante en ce sens qu'il n'existe que par les individus qui l'inventent ou qui l'imitent. La différence avec Durkheim se situe donc dans le fait que Tarde choisit d'appréhender les processus sociaux à partir des individus ; ce sont eux qui donnent naissance à la réalité sociale. Cette dernière est donc essentiellement une "interpsychologie". Ainsi, l'imitation agit comme une onde ou un courant magnétique, qui se propage d'individu en individu et qui se renouvelle à chaque fois à travers chacun. L'imitation est au cœur de toute vie sociale et explique aussi bien les relations humaines que l'histoire, celle-ci n'étant rien d'autre que le processus par lequel les individus inventent en s'imitant d'une civilisation à l'autre. Dans *L'Opinion et la Foule* (1901), G. Tarde mobilise également sa théorie de l'imitation pour expliquer la naissance de l'opinion publique. Ainsi, selon lui, si les processus sociaux ne sont pas forcément biologiques ou physiques, ils présentent des régularités sur le long terme qui suivent des lois qui ressemblent aux lois physiques. En ce sens, il est d'accord avec son adversaire (accord qui pour nous apparaît comme une proximité épistémologique forte, mais qui était très répandu dans le contexte positiviste de l'époque) : comme les phénomènes physiques, les phénomènes sociaux suivent des lois. Tarde emprunte à Leibniz l'idée des monades : les forces psychologiques du désir et des croyances permettent aux monades, sujets individuels ou groupes sociaux, d'agir et de s'imiter les uns les autres, formant ainsi l'agrégat qu'est la société. Le psychologisme ici relève plus d'une contrainte interne de la même manière que le supposé holisme durkheimien renvoie à un externalisme réaliste. On peut ainsi comprendre que pour Tarde, de même que la séparation nature/société n'est pas importante pour comprendre les interactions humaines, l'opposition entre l'individu et la société ou entre les niveaux micro et macro empêche toute compréhension de la constitution des sociétés.

Pour Durkheim au contraire, l'idée d'une opposition entre l'individu et la société est nécessaire à la sociologie, le fait social étant indépendant de l'individu, puisqu'il lui est extérieur, coercitif et général. Le choix méthodologique de Durkheim est de faire du fait social une entité *sui generis*, qui émerge de la fusion des consciences individuelles et qui leur est irréductible. Cependant, de même que pour Tarde, il serait faux de réduire Durkheim à un holisme généralisé, analyse qui peut ressortir du débat, alors que Durkheim a eu tendance à accentuer ses positions pour s'affirmer face à Tarde (Borlandi et Mucchielli, 1995) : « Bien qu'il y ait une morale du groupe… chaque homme a sa morale à soi : même là où le conformisme est le plus complet, chaque individu se fait en partie sa morale. Il y a en chacun de nous une vie morale intérieure, et il n'est pas de conscience individuelle qui traduise exactement la conscience morale commune, qui ne lui soit partiellement inadéquate » (Durkheim 2014, p. 115).

Le principal désaccord entre les deux chercheurs dans ce domaine se situe donc dans l'interaction entre le tout et les parties : contrairement à Durkheim, Tarde pense qu'une explication du tout passe nécessairement par ses parties. Il en résulte qu'au cœur de leurs débats, il oppose son « nominalisme » au « réalisme scholastique » de Durkheim, dont il craint qu'il ne débouche sur la métaphysique.

Contrairement à Durkheim, Tarde ne souhaite pas autonomiser catégoriquement la sociologie des autres disciplines. Il souhaite une grande science unifiée, où les fondements

épistémologiques sont les mêmes, et cohérents, que ce soit en psychologie, en biologie ou en sciences sociales. L'une des idées maîtresses qu'il revendique comme moteur de la cohésion est le double mouvement de l'impulsion et de la diffusion, que l'on retrouve dans l'ensemble des sciences. Ainsi, la seule réalité sociale est pour lui l'existence de consciences individuelles liées les unes aux autres par les lois de l'imitation. Et parce que les sciences étudient les différentes formes que prennent les agrégats, elles sont toutes, selon lui, des formes de sociologies. La sociologie se distingue des autres sciences non par la nature de son objet, mais par notre connaissance intime de ses constituantes.

(4) Bruno Latour et Bruno Karsenti mettent en scène le débat de 1903 entre Tarde et Durkheim, 2008

Source: http://www.bruno-latour.fr/

Selon Durkheim, la sociologie est possible parce qu'il y a des lois qui sous-tendent la vie sociale. La société est complexe et réfractaire aux formes simples d'explications, mais elle n'est pas opaque. Dans la tradition positiviste, il propose une analogie entre les faits étudiés par la sociologie (comme des choses) et ceux de la biologie. Ces derniers se manifestent grâce à l'union de différents éléments chimiques qui ne contiennent aucune parcelle de vie en elles-mêmes. Les tendances collectives ont une existence propre et sont l'expression des forces qui guident l'action. Il propose également une analogie entre les méthodes des deux sciences, qu'il veut rigoureuse et scientifique. Tarde critique la notion et la traite d'ontologisante, relevant davantage du monde des idées platoniciennes que de la science. Le fait social étant distinct des actes individuels, il s'imposerait aux individus de l'extérieur vers l'intérieur grâce à sa propriété coercitive. Il se trouverait dans une espèce de ciel des intelligibles que personne n'a jamais pu observer. De son côté, Durkheim reproche à Tarde de définir trop rapidement le « fait social », en lui donnant pour seul fondement la communication. Si le taux de suicide est un fait social, il doit être externe et coercitif. Il possède ces caractéristiques dans la mesure où Durkheim s'intéresse à son caractère statistique plutôt qu'aux causes de ses occurrences chez un individu en particulier. Il n'y a pas de cause individuelle qui permette d'imaginer la stabilité statistique qu'observe Durkheim. La stabilité, comme la société, émerge de la simple union de ce qui ne la contient pas. S'il n'est pas dans les cas particuliers, il doit être ailleurs. Il est aussi coercitif puisque s'il ne l'était pas, la stabilité du phénomène n'existerait pas. La stabilité statistique d'un fait social est un effet de son caractère coercitif. C'est donc dans l'approche statistique que se distinguent radicalement Tarde et Durkheim.

II.Les usages des statistiques par Tarde et Durkheim

C'est Durkheim, le plus jeune parmi les deux sociologues, qui a commencé la discussion entre

les deux, en attaquant la théorie de l'imitation (formulée par Tarde en 1890 dans *Les Lois de l'imitation*) appliquée sur le problème du suicide (*Le suicide*, 1897). Nous allons partir de là pour montrer ensuite la différence profonde entre ces visions des sciences sociales et, ce qui en découle, de l'usage des statistiques.

1. Statistiques et imitation

Selon Durkheim, le problème avec les tentatives d'expliquer des phénomènes sociaux par l'imitation est l'ouverture de la notion, qui ne permet pas de formuler des hypothèses claires. Par imitation, on peut comprendre soit une action en accord avec un groupe (1), soit l'obéissance à l'autorité (2), soit l'imitation au sens propre, quand une personne répète l'action d'une autre personne sans réflexion ni d'autres motifs (3) (Durkheim 1976 : 107-115). Tandis que Tarde réunit des phénomènes de tous ces types sous la notion d'imitation et voit les sources de toute action imitative plutôt dans le deuxième type (obéissance d'un « somnambule » au commandeur, Tarde 2001 : 138-139), pour Durkheim c'est seulement dans son troisième sens que l'utilisation du concept est justifiée. En effet, l'action en accord avec les attentes d'un groupe est tout à fait créative, et nécessite la réflexion. L'obéissance à l'autorité témoigne du respect ou de la crainte devant une force sociale, ce qui présente déjà un motif allant au-delà de l'imitation irréfléchie (Durkheim 1976 : 107-115). Tarde, de son côté, critique Durkheim pour l'étroitesse de cette nouvelle définition (Tarde 2000 : 17). On pourrait dire qu'on a à faire avec une différence qui pourrait être résolue par une convention. Seulement, Durkheim avait pour but l'opérationnalisation de la notion qui permettrait de formuler et de vérifier une hypothèse. Avec une catégorie tardienne qui comprend beaucoup d'actions hétérogènes, cela ne serait guère possible.

Pour vérifier l'hypothèse des suicides imitatifs, Durkheim compare la distribution réelle des suicides sur le territoire français avec une distribution théorique, qui devrait consister en une propagation ondulatoire de ce phénomène à partir de quelques foyers (grandes villes et chefs-lieux des départements). Or sur la carte, on ne distingue ni les ondes ni les foyers d'où, de toute évidence, proviennent ces ondes, mais plutôt de grandes régions avec tel ou tel taux de suicide (Durkheim 1976 : 126-129). Par conséquent, même s'il existe des cas individuels de suicides imitatifs, l'effet de l'imitation ne se voit pas dans les données agrégées. Bien que la critique de Tarde (selon laquelle d'une part, le phénomène imité peut devenir plus répandu dans les lieux où il est copié qu'aux lieux-sources, et d'autre part, dans un moment donné on ne doit pas nécessairement pouvoir saisir les « ondes ») soit rationnelle (Tarde 2000 : 19-21), elle ne présente qu'un argument contre la distribution théorique formulée par Durkheim, mais pas contre le besoin même de l'opérationnalisation, qui pose un véritable problème pour la conception de Tarde.

En rejetant l'explication des suicides par l'imitation dans son sens étroit, Durkheim croit avoir infirmé la dernière explication psychologique disponible : les deux premiers sens mènent à l'étude des causes sociales du phénomène. En effet, le problème posé par Durkheim dans son étude sur le suicide est incompatible avec la tentative de fonder les sciences sociales sur la psychologie, idée très répandue à l'époque, partagée par Tarde ou défendue en Allemagne par Wilhelm Wundt. C'est une observation statistique qui pousse Durkheim à formuler sa question principale : il constate que les rapports entre les taux de suicide des différentes périodes, pays et groupes sociaux restent constants. Or une telle stabilité ne peut pas être expliquée par les phénomènes psychologiques ou par le passage d'idée du suicide d'un individu à l'autre, puisqu'il n'y a que rarement des relations directes entre les personnes qui se tuent : « L'état d'esprit qui détermine les hommes à se tuer n'est pas transmis purement et simplement, mais, ce qui est beaucoup plus remarquable, il est transmis à un égal nombre de sujets placés tous dans les conditions nécessaires pour qu'ils passent à l'acte. Comment est-ce possible s'il n'y a que des individus en présence ? En lui-même, le nombre ne peut être l'objet d'aucune

transmission directe » (Durkheim 1976 : 347). Si l'on prend en compte que des constantes statistiques similaires s'observent aussi pour d'autres phénomènes, on comprend pourquoi Durkheim pouvait revendiquer une réfutation de fondation des sciences sociales sur des lois psychologiques ou quasi-psychologiques (Durkheim 1895 : 110-152). Les choses sociales, selon lui, présentent une catégorie bien à part et ne sont pas enracinées dans la vie des individus (Durkheim 1976 : 350-351). D'où l'importance de la statistique : c'est uniquement à l'aide des données agrégées, de l'étude de la distribution et des variations concomitantes qu'on peut observer et expliquer les phénomènes sociaux ; de même qu'on ne peut étudier la vie d'un organisme si l'on n'étudie que des cellules particulières, on n'arrive pas à expliquer, ni même à décrire la société en se concentrant sur des individus.

Si le rejet du psychologisme tardien est si étroitement lié à l'utilisation des données statistiques, constate-t-on une négligence des statistiques par Tarde ? Non, puisque Tarde insiste également sur l'importance de la statistique. Mais le rôle de la statistique est autre que chez Durkheim, et dans l'ouvrage principal où Tarde présente sa vision de ce qu'est la statistique, il n'en fait pas usage. Cela constitue un des nombreux arguments qui ont permis à Durkheim de critiquer le travail de Tarde comme un mélange de métaphysique et littérature sans rigueur scientifique (Vargas et al. : 12). Cette différence entraîne une certaine asymétrie de notre présentation : tandis que Tarde, dans son ouvrage sur les lois de l'imitation, expose volontiers ses réflexions sur la nature, l'utilisation et même le futur de la statistique sans pour autant faire de recherches quantitatives lui-même (Tarde travaille avec les données empiriques dans quelques autres ouvrages, qui portent en majorité sur les questions de la criminalité), l'utilisation des données par Durkheim est très intensive, mais ses réflexions théoriques sur le rôle de la statistique sont plus rares.

2. Des outils statistiques utilisés différemment

On a vu que pour Durkheim la statistique est un moyen d'étudier les choses sociales par excellence. Tarde, quant à lui, prétend que la statistique sera pour la société ce que les perceptions sont pour les individus : elle traduit ou symbolise la réalité extérieure de sorte qu'on peut la saisir, comme des yeux ou des oreilles collectifs. La différence actuelle est une différence de degré et de vitesse, mais un jour, quand les données seront recueillies et présentées quasi automatiquement, l'analogie deviendra parfaite (Tarde 2001 : 192-195). L'existence même des statistiques est due à l'existence de l'imitation, elle est fonction des évènements particuliers et n'est rien d'autre que l'addition des observations faites sur ces évènements. Pour Durkheim, la nature de la statistique est tout autre et elle n'est pas à confondre avec le processus de recueil des données (ce dont Durkheim se méfie souvent) : elle permet de mesurer les phénomènes sociaux qui ne peuvent pas être réduits aux cumuls des observations individuelles (Durkheim 1895 : 57) ; les grandes quantités montrent des choses qualitativement différentes de celles qui sont observables pour les individus pris à part. Les rapports constants entre les taux de suicide en sont un bon exemple.

(5) Statistiques des suicides en Frances en 1892

JUSTICE CRIMINELLE. — TABLEAU N° 45. — *Suicides d'après le mode de perpétration.* — *Suicides par mois.* — ANNÉE 1892.

MODE DE PERPÉTRATION.	NOMBRE TOTAL des suicidés.	SEXE DES SUICIDÉS. Hommes.	Femmes.
Submersion	2,452	1,704	748
Pendaison	3,832	3,299	533
Armes à feu	1,225	1,162	63
Asphyxie par le charbon[1]	829	461	368
Instruments tranchants et aigus	237	194	43
Poison	172	100	72
Chute volontaire d'un lieu élevé, sous un train en marche ou sous une voiture	387	276	111
Abus de liqueurs alcooliques	72	65	7
Feu	79	57	22
Autres (non spécifiés)	"	"	"
Hommes		7,318	"
Femmes		"	1,967
TOTAUX (1892)	9,285	7,318	1,967
Rappel des années 1891	8,884	6,937	1,947
Rappel des années 1890	8,410	6,576	1,834
Rappel des années 1889	8,180	6,381	1,799
Rappel des années 1888	8,541	6,663	1,788

[1] Sur ce nombre de 829 suicides par asphyxie à l'aide du charbon, 331 appartiennent au département de la Seine (187 hommes et 144 femmes).

SUICIDES. — TOTAUX DES INDIVIDUS QUI SE SONT SUICIDÉS	en janvier.	en février.	en mars.	en avril.	en mai.	en juin.	en juillet.	en août.	en septembre.	en octobre.	en novembre.	en décembre.
Hommes	602	558	620	705	686	751	769	668	546	535	449	429
Femmes	172	145	163	199	198	184	185	169	164	126	134	128
TOTAL (1892)	774	703	783	904	884	935	954	837	710	661	583	557
Rappel des années 1891	540	589	772	833	919	988	887	737	683	764	583	589
Rappel des années 1890	707	544	721	734	819	822	888	734	720	675	571	475
Rappel des années 1889	625	512	643	795	919	829	818	604	597	648	618	482
Rappel des années 1888	604	487	715	820	924	851	825	786	673	603	580	574

Source : Compte de l'Administration de la justice criminelle en 1892. Ministère de la Justice, Paris, I. N., 1895.

Source: Annuaire statistique, vol. 16, 1895, p. 46

Formé à l'école positiviste, Durkheim croit en la détermination des phénomènes sociaux par d'autres phénomènes sociaux, il cherche en eux les causes et les effets de chacun. Tout en étant conscient de la différence entre la corrélation et la causalité, il cherche en premier lieu des explications causales, et conclut parfois très vite à la causalité. En opposition à cette vision qu'il juge déterministe, Tarde donne plus de place au hasard et à la probabilité, en premier lieu grâce à sa conception de l'invention. C'est pour cela qu'il est accusé par Durkheim de fonder sa théorie sur le miracle et l'indétermination (Vargas et al. : 12). S'il existe des lois de l'imitation, l'invention, elle, est presque entièrement accidentelle et imprévisible. Bien sûr, il y a des limites de ce qui est possible, données par l'état courant du développement de la civilisation, par les théories dominantes, etc. Mais à l'intérieur d'un tel champ du possible peuvent apparaître beaucoup d'inventions, grâce à l'intervention d'un « génie ». Comme chaque invention donne naissance à une onde d'imitation qui, à son tour, influe sur les autres ondes, cet élément important dû au hasard empêche la décomposition de la réalité sociale en un nombre soumis à des régularités et, par conséquent, empêche la prédiction. Pourtant, au fil du temps les

inventions deviennent plus rares et leur part par rapport aux propagations imitatives diminue, ce qui rend les prédictions faites à partir des lois de l'imitation plus probables (Tarde 2001 : 197).

Pour Tarde, la forme principale de l'usage de la statistique sont les tableaux chronologiques et les courbes dessinés à partir d'eux, qui permettent de suivre l'évolution d'une invention de sa naissance jusqu'à son déclin (Tarde 2001 : 163-164). Elle reflète le développement du nombre des imitations d'un comportement ou d'une croyance par les individus. Comme Durkheim, Tarde insiste sur l'importance du poids relatif d'un phénomène qui se propage dans une société, c'est-à-dire la part du nombre d'actes imitatifs dans le nombre possible de tels actes (Tarde 2001 : 165-166). Il y a une forme générale de courbe, qui représente une propagation normale : elle commence par une montée, d'abord lente, puis brusque, passe à un plateau quand la propagation est arrêtée par la propagation d'un phénomène concurrent et s'abaisse après que le concurrent a gagné. Comme l'imitation est un concept qui englobe à peu près tout ce qu'il y a dans le monde social – des biens de consommation, comme le café ou le tabac, aux moyens techniques, comme la locomotion à vapeur, ou encore aux institutions, comme l'égalité ou la propriété individuelle, – une telle courbe décrit un phénomène social de n'importe quel type. Les courbes, selon Tarde, sont comparables au trajet des oiseaux, qui montent et qui descendent ; cela appuie l'idée que selon lui, les statistiques doivent être utilisées pour montrer un caractère dynamique, et non statique, d'un phénomène.

Les exceptions sont-elles possibles ? En fait, selon Tarde, la courbe de cette forme générale n'est pas nécessairement une description empirique mais plutôt une construction théorique, un type idéal qui décrit comment un phénomène se propage en vacuum, où il ne rencontre pas l'action parallèle des autres phénomènes. Mais en réalité, une propagation est souvent empêchée ou renforcée par la propagation des autres inventions, ce qui peut changer la forme des courbes et nécessite l'étude de leurs interactions (Tarde 2001 : 174-188).

Durkheim quant à lui ne se limite pas à une forme principale de l'usage des statistiques, et il se sert de nombreux moyens en fonction des questions et des hypothèses formulées. Il utilise surtout des tableaux présentant la distribution des suicides dans les différents groupes sociaux, pays et périodes, il fait des calculs des moyennes et des différences entre les moyennes, des taux de suicide et leurs rapports, ainsi que des cartes géographiques. Tous ces outils servent un même but, celui de représenter les « variations concomitantes » du phénomène étudié selon le milieu social. L'étude des variations concomitantes est également un devoir de la statistique selon Tarde (Tarde 2001 : 170), mais dans son cas il s'agit de l'interaction des courbes.

Conformément à la place qu'occupe dans la philosophie sociale tardienne la propagation des phénomènes sociaux par l'imitation, la plupart des statistiques analysées sont des séries temporelles. En observant les dynamiques de développement de certains phénomènes à travers les données chronologiques (dont le pendant visuel sont les courbes si fréquemment adorées par l'auteur[2]), Tarde essaie de saisir les relations entre les différentes tendances longitudinales, ainsi que l'influence des événements singuliers, tels que les bouleversements politiques, sur ces développements. La comparaison des courbes lui permet de traiter les hypothèses sur les liaisons entre les phénomènes. Par exemple, il refusait la « loi » du développement inverse du suicide et de l'homicide partagée à l'époque par certains auteurs en comparant les deux courbes correspondantes (Tarde 1924 : 166-167). Les développements tant des effectifs que des pourcentages sont soumis à l'analyse, et les différences entre eux deviennent également l'objet de réflexions (Tarde 1924 :71). Même des questions scientifiques traditionnelles ont souvent

2. En dehors des contraintes techniques de l'époque, il y a une raison substantive pour laquelle Tarde présente les tendances étudiées sous la forme des tableaux chronologiques ou de dénombrements textuels plutôt que sous la forme des courbes tracées : « Il est bon cependant de prévenir que la vue des courbes, si on ne la complète et ne la corrige par la lecture du rapport et des tableaux, est très propre à égarer l'esprit » (Tarde 1924 : 63).

une signification plus spécifique et sont « traduits » vers sa théorie de la société, vers la langue des courbes imitatives. Ainsi, la question des causes de développement social devient celle de la mécanique de l'interaction des tendances et contre-tendances : « Pourquoi, en termes plus compréhensifs, ce genre d'exemple triomphe-t-il plus ou moins, suivant les lieux et les temps, dans sa lutte ou son concours avec d'autres genres d'exemples? » (Tarde 1892 : 57).

(6) Statistiques qui inspirent Tarde : tableau chronologique de l'importation des alcools en Franc

ALCOOLS. — ANNÉES 1850-1895.

Importation et exportation des alcools depuis 1850.

TABLEAU N° 371. — *Importation (commerce spécial).*

ANNÉES.	ALCOOL PUR PROVENANT de l'Allemagne.	ALCOOL PUR PROVENANT de l'Angleterre.	ALCOOL PUR PROVENANT d'autres pays.	TOTAL de l'alcool pur importé.	LIQUEURS. (Volume total.)	TOTAL GÉNÉRAL
	hectol.	hectol.	hectol.	hectol.	hectol.	hectol.
1850.....	15	110	5,430	5,555	99	5,654
1851.....	10	89	7,267	7,366	103	7,469
1852.....	21	121	12,857	12,999	111	13,110
1853.....	7	118	12,616	12,741	158	12,899
1854.....	3,818	14,834	46,480	65,132	145	65,277
1855.....	28,560	86,043	88,276	202,888	198	203,086
1856.....	6,065	80,600	90,981	177,646	208	177,854
1857.....	133,145	79,286	164,639	377,070	228	377,298
1858.....	1,694	2,360	34,812	38,852	215	39,067
1859.....	4,312	12,593	29,896	46,801	213	47,014
1860.....	29,893	21,505	37,252	88,650	257	88,907
1861.....	22,671	48,169	63,504	134,344	266	134,610
1862.....	9,149	17,090	39,745	65,984	309	66,293
1863.....	12,588	9,623	42,452	64,663	295	64,958
1864.....	13,109	21,445	35,018	69,572	258	69,830
1865.....	15,260	2,546	27,865	45,671	388	46,059
1866.....	20,808	2,315	41,298	64,421	558	64,979
1867.....	16,816	1,699	31,020	49,535	578	50,113
1868.....	43,412	1,847	47,716	92,975	633	93,608
1869.....	82,525	2,312	44,910	129,747	673	130,420
1870.....	29,085	1,303	32,867	63,255	566	63,821
1871.....	27,679	5,078	52,840	85,597	585	86,182
1872.....	2,782	1,438	43,006	47,226	754	47,980

ANNÉES.	ALCOOL PUR PROVENANT de l'Allemagne.	ALCOOL PUR PROVENANT de l'Angleterre.	ALCOOL PUR PROVENANT d'autres pays.	TOTAL de l'alcool pur importé.	LIQUEURS. (Volume total.)	TOTAL GÉNÉRAL.
	hectol.	hectol.	hectol.	hectol.	hectol.	hectol.
1873.....	8,276	2,038	36,932	47,246	809	48,055
1874.....	10,051	2,109	48,435	60,595	989	61,584
1875.....	5,473	1,854	55,901	63,228	1,205	64,433
1876.....	15,379	1,020	47,583	63,982	1,573	65,555
1877.....	35,362	3,173	56,824	95,359	1,514	96,873
1878.....	62,243	8,436	62,442	133,121	1,760	134,881
1879.....	102,211	21,849	74,214	198,274	1,871	200,145
1880.....	121,720	1,352	136,932	260,004	2,000	262,004
1881.....	122,863	45,266	68,307	236,436	2,483	238,919
1882.....	155,470	27,248	101,331	284,049	2,452	286,501
1883.....	46,537	15,966	104,476	164,979	2,623	167,602
1884.....	55,749	7,125	126,736	189,610	2,462	192,072
1885.....	48,911	11,931	142,858	203,700	2,424	206,124
1886.....	63,645	27,291	134,324	225,260	2,553	227,813
1887.....	33,152	18,492	158,925	210,569	1,944	212,513
1888.....	1,917	3,208	140,965	146,090	1,785	147,875
1889.....	977	3,676	123,069	127,782	1,660	129,382
1890.....	988	4,308	131,610	136,906	1,900	138,806
1891.....	795	5,595	127,223	131,413	2,311	133,724
1892.....	847	3,033	148,726	153,506	1,632	155,138
1893.....	706	3,382	139,917	143,905	1,490	144,239
1894.....	1,210	2,047	150,536	154,693	1,500	156,193
1895.....	540	3,050	133,495	137,085	1,614	138,699

Source : Annuaire statistique, vol. 16, 1895, p. 306

Une autre considération sur la place de la courbe idéal-typique, qui a chez Tarde le statut d'une loi, permet de questionner la construction théorique qu'il propose. Comme vu ci-dessus, si la forme empirique ne correspond pas à cette loi, c'est probablement parce que d'autres propagations entrent en interaction. Mis à part le hasard des inventions, on peut également formuler des lois de l'interaction des ondes de propagation. Mais, en fait, Tarde admet que même cette précision ne permet pas d'expliquer la forme de toutes les courbes qui existent : « *quand, par exception, une courbe irrégulière de statistique est réfractaire à l'analyse précédente et refuse de se résoudre en courbes normales, c'est qu'elle est insignifiante en soi, fondée sur des dénombrements peut-être curieux, mais nullement instructifs, d'unités dissemblables, d'actes ou d'objets arbitrairement groupés, à travers lesquels cependant un ordre soudain apparaît si la présence d'un désir ou d'une croyance déterminés vient à s'y révéler au fond* » (Tarde 2001 : 189-190). Si l'on prend cette remarque au sérieux, elle rend toute la construction théorique de Tarde infalsifiable, puisque toute courbe qui ne correspond pas à la théorie est déclarée « insignifiante ». Durkheim choisit une stratégie plus délicate du traitement des observations qui semblent, à première vue, ne pas correspondre à sa théorie. A savoir, il essaie de trouver une classification du phénomène étudié qui permettrait d'expliquer séparément les variations concomitantes de chaque type. Ainsi, grâce à la division des suicides en égoïstes, altruistes, fatalistes et anomiques, il arrive à rendre compte de tendances apparemment contradictoires (l'influence différente du mariage sur les suicides des époux et des épouses, l'augmentation

des suicides civils accompagnée d'une baisse des suicides à l'armée, etc.). Dans certains cas, Durkheim déclare ne pas avoir trouvé de régularités (Durkheim 1976 : 385). Parfois il cherche à améliorer sa démarche, par exemple en tenant compte de la médiation de l'âge qui intervient dans la relation entre l'état civil et les suicides (Durkheim 1976 : 175) ou en réfléchissant sur le mode de recueil des données (Durkheim 1976 : 144, 256). En tout cas, il est difficile de trouver des tendances visibles qu'il ne prend pas en compte et ne cherche pas à expliquer. Enfin, en ce qui concerne le contre-exemple mentionné dans la réponse de Tarde du monastère où il y avait apparemment moins de suicides malgré le caractère coercitif et dépersonnalisant de la communauté monastique (Tarde 2000 : 38), on ne peut pas savoir s'il n'était pas connu de Durkheim ou s'il l'a volontairement omis.

3. Définitions de la statistique

Mais quelle est cette statistique à partir de laquelle les deux sociologues appliquent ou espèrent appliquer leurs instruments scientifiques ? Qu'est-ce qui est ou doit être quantifié, mesuré ou comparé ? On a vu que les courbes tardiennes présentent des agrégats des actes imitatifs. Au fond, ces actes sont les manifestations soit des besoins ou des désirs, soit des croyances. Cette thèse est ancrée dans l'épistémologie de Tarde : la croyance et le désir sont les plus petites unités psychologiques qui sont mesurables : ce sont les quantités de base. Elles entrent en interaction avec les sensations, qui sont qualitativement différentes et ne peuvent être quantifiées en soi, sans que les croyances et les désirs s'appliquent à eux. Ces interactions sont le fondement de toute la vie psychologique et sociale (Tarde 1900 : 235-308). C'est la quantification des sensations dans les désirs et les croyances qui permet de comparer les choses psychologiques et sociales. Si les données quantitatives que Tarde traite dans les ouvrages empiriques portent le plus souvent sur la criminalité et les délits (ce qui est plus facile pour Tarde, compte tenu du poste qu'il occupe), il attribue dans sa théorie une priorité aux statistiques commerciales et industrielles, parce qu'elles reflètent le développement des besoins (Tarde 2001 : 171-172). En effet, à travers le nombre des marchandises produites ou vendues, on peut suivre la dynamique d'un besoin, et son degré de présence dans une société. Bien sûr, la quantification de telles unités est plus facile que dans le cas des croyances. Un exemple de la statistique des croyances seraient les résultats des votes (dans cette production des statistiques sur la société Tarde voit une importante fonction du suffrage universel : Tarde 2001 : 167 ; Tarde 1892 : 439-449).[3]

Contrairement à ce qu'on pourrait attendre en partant de la définition durkheimienne des faits sociaux, les besoins peuvent être pour Durkheim un objet statistique pertinent. Il est vrai que les régularités établies grâce à l'étude des données statistiques sont plus que de simples ensembles des actes individuels ; seulement, les faits sociaux auxquels on s'intéresse concernent aussi la manière dont la société et les règles de son fonctionnement influencent les comportements des individus. Quand Durkheim définit les trois types du suicide (si l'on exclut le suicide fataliste, qu'il ne traite que sur quelques pages), il recourt à une théorie qui décrit comment le contexte social limite, forme ou contient les besoins et les désirs d'une conscience individuelle. Par exemple, dans le cas des suicides anomiques, l'équilibre entre les besoins et les normes sociales qui les limitent est rompu (Durkheim 1976 : 271-285). De manière générale, les suicides dépendent de la manière « dont les individus sont attachés à la société » et « de la façon dont elle les réglemente » (Durkheim 1976 : 288). Il serait alors justifié de dire qu'une tâche importante de la statistique pour les deux auteurs est de rendre compte de l'état social d'un besoin. Quant à la croyance, elle n'a guère de place dans la théorie durkheimienne et dans son utilisation de

3. L'extension des droits électoraux aux femmes et même aux enfants doit servir, selon Tarde, à une représentation plus correcte de la société, ce qui serait préférable non seulement pour les enjeux de l'administration et de la science, mais aussi pour la politique, puisque les « pères de familles » obtiendraient plus de poids et on pourrait empêcher le danger d'une « éphébocratie » ou « célébocratie ». – L'argument apparemment progressiste est, en fait, utilisé par Tarde pour doter les hommes mariés de l'âge mûr d'une influence politique plus grande, parce que ce sont eux qui devraient exercer les droits électoraux de sa femme et de ses enfants (Tarde 1892 : 439-449).

la statistique, si ce n'est dans sa fonction d'intégration sociale, de l'attachement d'un individu à sa communauté. Les croyances subjectives ne sont pas théoriquement intéressantes, tant qu'elles n'influencent pas de manière importante les faits sociaux ; plus que ça, les consciences ne sont pas tout à fait transparentes à elles-mêmes et les individus peuvent se tromper sur leurs propres motifs.

Les deux théoriciens pensent que les résultats des recherches quantitatives ont une application immédiate pour les questions pratiques. Ainsi, Tarde espère, par l'étude des effets favorables ou nuisibles des courbes, influencer « le penchant qu'ils auraient à suivre ou à ne pas suivre tels ou tels exemples » (Tarde 2000 : 170). Quant à Durkheim, sa compréhension du rôle pratique des sciences sociales est proche de l'idée commune du rôle de la science en général : en trouvant les causes des phénomènes étudiés, on peut espérer agir sur ces causes pour éviter les effets négatifs et pour maximiser les effets souhaitables (voir ainsi ses préconisations en matière de consentement mutuel à partir des statistiques sur le suicide, « Le divorce par consentement mutuel », 1906).

Pour rendre compte de l'immense influence que la démarche scientifique de Durkheim a eue sur les sciences sociales, voici un court dénombrement des outils statistiques qu'il utilise :

- les calculs des taux de suicide selon les différents milieux sociaux (sexe, âge, statut familial, régions, religions, professions…) ;
- l'étude de corrélation de deux phénomènes à l'aide des « variations concomitantes ». Durkheim rend compte de la différence entre la causation et la corrélation et d'une possible influence d'un facteur perturbateur, tel qu'un facteur qui a un effet sur les deux variables étudiées (Durkheim 1976 : 36, 171 ; Durkheim 1895: 153-171) ;
- la comparaison des distributions empiriques avec les distributions théoriques sur lesquelles Durkheim formule des hypothèses ad hoc (Durkheim 1976 : 94-95, 188-189). Dans ce cas comme dans le précédent, Durkheim ne recourt pas aux tests statistiques dont on ne faisait pas encore usage à l'époque et définit la significativité des différences « à l'œil » : « Une correspondance aussi régulière et aussi précise ne peut être fortuite » (Durkheim 1976 : 98) ;
- comparaison des variations à l'aide des cartes géographiques ;
- recherches des variables cachées (le cas de l'ethnie et de la religion, Durkheim 1976 : 62, le cas du mariage et l'âge, Durkheim 1976 : 175) ;
- élimination de l'influence des valeurs extrêmes qui « peuvent élever ou abaisser artificiellement la moyenne » (Durkheim 1976 : 122). Cependant, Durkheim n'utilise pas les médianes ;
- classification du phénomène étudié en aval de l'étude, selon les différences de sa distribution dans les différentes conditions sociales ;
- étude des différences entre les rapports des moyennes (Durkheim 1976 : 169) ;
- étude des odds ratios, des rapports entre les taux des suicides. Pour autant, Durkheim ne les interprète pas dans un esprit probabiliste comme les chances de se suicider selon des caractéristiques sociales. Appliqué au problème de suicide, cet outil obtient le nom de « coefficients de préservation » (Durkheim 1976 : 181-183). L'utilisation de cette méthode nécessite le choix des modalités de référence ;
- recherche des indicateurs empiriques qui permettent de vérifier les hypothèses formulées ;
- tentatives de partager les variations établies en une partie expliquée par les causes trouvées et une partie expliquée par les particularités personnelles des suicidés (Durkheim 1976 : 312-313)

Cette liste n'est pas exhaustive, mais elle donne une idée du rôle de l'utilisation de la statistique par Durkheim dans la diffusion des procédures de recherche quantitative qui sont établies aujourd'hui.

(7) L'usage des odds ratios ("coefficients de préservation") par Durkheim

TABLEAU XXII

Comparaison du taux des suicides par million d'habitants de chaque groupe d'âge et d'état civil dans la Seine et en province (1889-1891).

Ages.	HOMMES (Province). Célibataires.	Époux.	Veufs.	COEFFICIENTS de préservation par rapport aux célibataires. des époux.	des veufs.	FEMMES (Province). Célibataires	Épouses.	Veuves.	COEFFICIENTS de préservation par rapport aux célibataires. des épouses.	des veuves.
15-20.....	100	400		0,25		67	36	375	1,86	0,17
20-25.....	214	95	153	2,25	1,39	95	52	76	1,82	1,25
25-30.....	365	103	373	3,54	0,97	122	64	156	1,90	0,78
30-40.....	590	202	511	2,92	1,15	101	74	174	1,36	0,58
40-50.....	976	295	633	3,30	1,54	147	95	149	1,54	0,98
50-60.....	1.445	470	852	3,07	1,69	178	136	174	1,30	1,02
60-70.....	1.790	582	1.047	3,07	1,70	163	142	221	1,14	0,73
70-80.....	2.000	664	1.252	3,01	1,59	200	191	233	1,04	0,85
Au delà.....	1.458	762	1.129	1,91	1,29	160	108	221	1,48	0,72
Moyennes des coefficients de préservation.....				2,88	1,45	Moyennes des coefficients de préservation ...			1,49	0,78
	HOMMES (Seine).					FEMMES (Seine).				
15-20.....	280	2.000		0,14		224				
20-25.....	487	128		3,80		196	64		3,06	
25-30.....	599	298	714	2,01	0,83	328	103	296	3,18	1,10
30-40.....	869	436	912	1,99	0,95	281	156	373	1,80	0,75
40-50.....	985	808	1.459	1,21	0,67	357	217	289	1,64	1,23
50-60.....	1.367	1.152	2.321	1,18	0,58	456	353	410	1,29	1,11
60-70.....	1.500	1.559	2.902	0,96	0,51	515	471	637	1,09	0,80
70-80.....	1.783	1.741	2.082	1,02	0,85	326	677	464	0,48	0,70
Au delà.....	1.923	1.111	2.089	1,73	0,92	508	277	591	1,83	0,85
Moyennes des coefficients de préservation.....				1,56	0,75	Moyennes des coefficients de préservation ...			1,79	0,93

Source: Durkheim, Le suicide, 1897

Quant à Tarde, même si la vision large de l'essence, du rôle et du futur de la statistique, ainsi que sa théorie de la quantification, présentent la source principale de l'inspiration des auteurs contemporains pour lui, certains de ses ouvrages comportent des études statistiques empiriques, des présentations des données et des raisonnements statistiques. On a mentionné ci-dessus quelques exemples portant sur l'étude des courbes : conformément à la place qu'occupe dans la théorie tardienne la propagation des phénomènes sociaux par l'imitation, la plupart des statistiques analysées sont des séries temporelles. Malgré toutes les différences épistémologiques, certains types de raisonnement que l'on trouve dans l'ouvrage de Durkheim sont présents également dans les œuvres de Tarde. Ainsi, Tarde formule des hypothèses contrefactuelles sur la distribution d'un phénomène dans le cas de présence d'un facteur influant, et il examine ensuite les distributions empiriques afin de confirmer ou d'infirmer ses hypothèses. Bien que ce ne soit pas exactement le raisonnement explicite par hypothèse nulle et par distribution due au hasard, le raisonnement de type « si un tel facteur était présent,

alors telle distribution devrait avoir telle forme ; cela n'est pas le cas, alors ce facteur n'a pas d'influence sur cette tendance » en est proche (Tarde 1924 : 95-96 et Tarde 1900 : 214 pour les exemples). L'analyse de la corrélation entre deux phénomènes ne le mène pas nécessairement à la conclusion hâtive sur la causalité de l'un sur l'autre. La corrélation forte et positive entre les divorces et les suicides pousse Tarde à conclure que ces deux phénomènes ont d'autres causes en commun (Tarde 1924 :176). On rencontre chez Tarde également le raisonnement « à tendance X donnée », qui permet de mieux discerner l'impact d'une tendance à l'autre, en neutralisant l'action modifiante d'une troisième tendance (Tarde 1924 : 79). Tarde calcule les odds ratio (Tarde 1900 : 274) et utilise des simples mesures de la dispersion (Tarde 1924 : 169). Il réfléchit sur le mode de collecte de données et sur le rapport entre les statistiques et la réalité. Par exemple, il se demande si le nombre des méfaits poursuivis reflète le nombre des méfaits commis (avec une réponse positive, dans l'esprit de sa vision optimiste de la statistique) (Tarde 1924 : 69).

L'usage de la statistique dans les ouvrages de Tarde et Durkheim : un résumé

Tarde	Durkheim
La statistique part des évènements individuels de caractère psychologique (des imitations) et se produit par une démarche additive	Les observations sur les données agrégées permettent de voir les phénomènes sociaux qualitativement différents des évènements individuels
L'unité de base se trouve au niveau sous-individuel (croyances et désirs)	L'unité de base se trouve au niveau supra-individuel (fait social)
La transition entre l'individu et l'agrégat est possible par une simple démarche arithmétique	La transition de l'individu à la totalité de société est impossible, le comportement de l'individu "incarne" des faits sociaux
Critique de Durkheim : caractère métaphysique du concept de la société qui transcende les individus	Critique de Tarde : absence de méthode scientifique
Le hasard et la probabilité sont introduits par le concept de l'invention	Les études portent sur les régularités et les rapports constants et cherchent à les expliquer
Optimiste pour le futur de la statistique (une sorte de développement des organes collectifs de sens), modéré pour le développement présent	Usage intensif des procédures disponibles, optimiste pour le rôle et les possibilités des sciences sociales
Usage préféré : les tableaux chronologiques et les courbes de diffusion des phénomènes	Usage préféré : calculs des rapports, variations concomitantes…
Utilité pratique : en étudiant les interactions des courbes, trouver quelles imitations peuvent aider à lutter contre l'imitation des phénomènes nuisibles	Utilité pratique : trouver les causes des phénomènes et agir sur ces causes pour minimiser les effets qui ne sont pas souhaitables

III.Usages et mésusages des statistiques de Tarde et de Durkheim

A l'époque du débat entre les deux auteurs, le vent est largement favorable à Tarde : la psychosociologie est à la mode pour les « sciences de l'esprit », pour reprendre la distinction proposée par Dilthey dans le cadre de sa théorie de la connaissance, distinction reprise justement par les psychosociologues. On retiendra Wilhelm Wundt et la psychologie scientifique

ou le communautarisme de Ferdinand Tönnies. De façon générale, pour l'immense majorité des intellectuels de l'époque, tout acte humain doit d'abord être envisagé dans son aspect individuel. Le social étant inscrit au plus profond de l'homme, il serait absurde de le considérer comme un fait extérieur à l'homme lui-même.

(8) Mentions de Durkheim et Tarde dans la littérature francophone, 1880-2008

Source : Google NGram Viewer

Il est intéressant de constater que la courbe des mentions de Tarde dans la littérature francophone présente exactement le développement mentionné par Tarde : elle monte lentement, puis brusquement, s'aplatit pour redescendre enfin au moment-même où le concurrent principal de Tarde dans le monde académique, Durkheim, gagne en popularité.

Les raisons institutionnelles et théoriques évoquées ci-dessus ont renversé le rapport de force après la mort des deux sociologues : Tarde n'a pas de successeur, tandis que la sociologie française est quasi uniquement durkheimienne jusqu'à la Deuxième Guerre mondiale. Ainsi, jusqu'aux années 1970, Tarde était très peu cité, et de façon partielle aux Etats-Unis pour ses théories sur l'imitation, et en France pour ses apports à la criminologie. Certains voient cependant dans son imitation, présentée comme un acte spontané, en quelque sorte consubstantiel au lien social et qui pousserait des êtres inférieurs à imiter des êtres supérieurs, l'embryon des théories diffusionnistes des traits culturels, dont on trouve de nombreux exemples dans la sociologie du XX^{ème} siècle. En ce qui concerne Durkheim, l'essor de *L'Année sociologique* met au premier plan ses héritiers ; tout d'abord les membres de son équipe de recherche (Marcel Mauss, Paul Fauconnet, Célestin Bouglé, ou Lucien Lévy-Bruhl) mais aussi les principales figures de la sociologie et de l'anthropologie de l'époque, telles que Maurice Halbwachs, Talcott Parsons, Alfred Radcliffe-Brown, ou encore Claude Levi-Strauss. Les positions universitaires étaient en majorité occupées par des durkheimiens (Georges Davy, Célestin Bouglé,...). L'immédiat après-guerre est dominé par une sociologie encore précaire, confrontant son héritage durkheimien aux influences allemandes (marxisme et phénoménologie) et américaines (les techniques d'enquête). De 1960 à 1975, la sociologie s'institutionnalise en tant que telle (licence de sociologie, ouvrages didactiques, élargissement éditorial), et prend une véritable ampleur après 1975 ; l'intérêt pour une histoire renouvelée de la discipline à cette époque se confond à une redécouverte de Durkheim comme "père fondateur". Parallèlement, la renaissance de diverses sociologies de l'acteur (dans des paradigmes qui accordent à cet acteur une place plus ou moins importante, en allant de l'individualisme méthodologique de Raymond Boudon à l'actionnalisme d'Alain Touraine, au modèle stratégique de Michel Crozier ou encore au

modèle de l'acteur-réseau de Bruno Latour) tend à donner à nouveau à G. Tarde un rôle central dans la sociologie, en tant précurseur revendiqué par les chefs de file de ces différents paradigmes sociologiques. Cette redécouverte s'accompagne d'une réédition des œuvres de Tarde dans la collection « les empêcheurs de tourner en rond », filiale de La Découverte (2001). Dans « Tardomania ? Réflexions sur les usages contemporains de Tarde », Laurent Mucchielli retrace l'histoire des quelques réactualisations qu'a subie l'œuvre de Tarde depuis les années soixante. Selon lui, la redécouverte de Tarde est généralement liée à une opposition aux thèses durkheimiennes, voire à un usage plus tactique que théorique visant à légitimer une école de pensée par rapport à une autre.

Dans le cas de Bruno Latour, « les rayons imitatifs » tardiens seraient des précurseurs des théories modernes de réseau social[4]. Latour insiste surtout sur l'idée tardienne de quantification dans le sens que tant la société que les individus sont animés de croyances et de désirs. Il est, en principe, possible de retracer la formation de l'agrégat par l'interaction de ces quanta, de voir comment les quanta se sont assemblés dans un tel ou tel agrégat. Déjà à l'époque de Tarde, on disposait de moyens pour le faire dans le domaine de la science, parce que les interactions des croyances des savants sont très bien documentées. Aujourd'hui, on dispose des moyens qui le permettent aussi dans d'autres domaines, grâce aux outils numériques de recueil et de visualisation des données. Ainsi, les descriptions visionnaires de Tarde sur le futur de la statistique seraient confirmées (Latour 2010).

Une des premières reprises de l'œuvre de Durkheim et de son usage des statistiques dans *Le Suicide* a été réalisée par son disciple M. Halbwachs dans *Les Causes du suicide* (1930). Halbwachs reprend également les variations concomitantes, avec des données plus fines et plus nombreuses que celles dont disposait Durkheim. Mieux formé en statistiques, il raisonne davantage sur les dispersions autour des moyennes. Il s'intéresse également à la qualité des sources statistiques, aux modes de suicide et aux tentatives. Il y confirme notamment que les données statistiques sont fausses en niveau mais justes dans la mesure des écarts et des variations. Par ailleurs, il accorde une plus grande importance à l'interaction des phénomènes, sans les isoler, ce qui fait qu'il ne dissocie pas la religion et l'environnement (la nation, le degré d'urbanité). Cela lui permet de remettre en cause les conclusions de son maître : après avoir exploré deux solutions (simplification de l'équation et substitution de la seconde variable – l'environnement – à la première – la religion -, et décomposition de l'effet pur de la religion et l'effet pur de l'environnement), il en propose une troisième : il affirme l'indissociable interaction entre ces variables dont les actions conjuguées portent la marque d'un « milieu ». Ramenée à ses dimensions culturelles plus que cultuelles, la religion fait corps avec le milieu et le genre de vie. Ainsi, loin de rompre avec des explications de type culturaliste, il leur ouvre au contraire la voie : la mise en évidence de relations statistiques régulières, robustes et complexes entre le suicide, la religion, le pays et le mode de vie.

Nous ne reviendrons pas en détail sur les discussions qui ont été faites autour du *Suicide* et de ses statistiques. Après un héritage très fort aux Etats-Unis, avec Parsons et les fonctionnalistes, après des critiques méthodologiques telles que celles de J. Douglas dans *The Social Meanings of Suicide* (1967), la France a réactualisé ses critiques, et s'en sont suivis d'importants débats sur la construction des statistiques du suicide (P. Besnard, « Anti- ou anté-durkheimisme ? Contribution au débat sur les statistiques officielles du suicide », Revue française de sociologie, 1976). Ces réinterprétations se basent souvent sur des données actualisées[5].

Une interprétation donnée par P. Besnard (1973), qui a donné lieu à une vaste analyse du

4. L'analyse des réseaux sociaux suppose différents types de liens entre les nœuds d'un réseau. Pour Tarde, le seul lien possible est l'imitation. Il n'y a donc pas d'analyse à faire au-delà de la forme du schéma.
5. En ce qui concerne la France, voir la relecture de Christian Baudelot & Roger Establet : *Durkheim et le suicide* (1984) : travail de synthèse et de vulgarisation, qui comporte des données françaises récentes manifestant une remarquable stabilité des relations entre famille et suicide, mises en évidence par Durkheim.

suicide féminin et du sens à lui donner, avait pour but de corriger à la fois la construction des tableaux et leur interprétation sociologique. Besnard parle de la "courbe en U", modélisation de ce que pensait Durkheim avec l'opposition entre les deux types de suicide pour régulation / intégration, mais modélisation que Durkheim n'a pas faite, ce qui l'a conduit à abandonner progressivement cette idée (d'où découle celle du juste milieu comme équilibre), en négligeant par exemple le suicide fataliste, en passant de références à des variables à des références à des courants. Cela a donné lieu à une réponse de C. Dubar (Dubar 2004), puis de P. Besnard lui-même (Besnard 1987). L'actualité de ces discussions montre l'importance toujours revisitée des statistiques durkheimiennes.

Conclusion

Gabriel Tarde et Emile Durkheim travaillent dans une époque marquée d'un côté par le psychologisme et les tentatives de fonder les sciences de la société sur l'étude des phénomènes psychologiques (tentatives soutenues par Tarde et rejetées par Durkheim), et d'un autre côté par un rapide développement des outils statistiques qu'ils théorisent et dont ils font usage. La compréhension et l'usage de la statistique par Tarde et Durkheim reflètent leurs conceptions respectives de la société et des sciences sociales. Tarde croit que les processus sociaux, grâce à leur nature de rassemblements de quanta, peuvent être quantifiés. A l'aide des tableaux chronologiques et des courbes, on peut voir comment les croyances et les désirs se propagent et interagissent en formant des sujets et des sociétés. Durkheim, au contraire, part du niveau supra-individuel, constate que les faits sociaux sont indépendants des variations individuelles et cherche à les expliquer en utilisant intensivement des moyens statistiques disponibles. Si le raisonnement et la démarche de recherche quantitative de Durkheim ont fortement marqué la sociologie du XX^{e} siècle, certains philosophes et sociologues ont récemment redécouvert Tarde qu'ils apprécient surtout en tant qu'un visionnaire heureux du futur de la statistique et des sciences sociales.

Références

- Bélanger Pierre-Luc, « La construction sociale de l'individu chez Tarde et Durkheim », Mémoire réalisé en 2010 à l'université de Montréal
- Besnard, Philippe. « Anti- ou anté-durkheimisme ? Contribution au débat sur les statistiques officielles du suicide », Revue française de sociologie, XVII, n° 2, 313-341, 1976
- Besnard, Philippe (1987). Les sociologistes et le sexe. Réponse à Claude Dubar. In: Revue française de sociologie, 28-1. pp. 137-144.
- Borlandi M. et al. Gabriel Tarde et la criminologie au tournant du siècle. Presses Universitaires du Septentrion, 2000.
- Borlandi M. et Cherkaoui M. (éd.). Le Suicide un siècle après Durkheim. Paris : PUF, 2000.
- Candea M. The Social after Gabriel Tarde: Debates and Assessments (2010). Routledge.
- Cuin Charles-Henry et Gresle François, Histoire de la sociologie 1, Editions La Découverte, Collection Repères, 1992
- Didier, Emmanuel (2010) "Gabriel Tarde and Statistical Movement," Social After Gabriel Tarde, London, Routledge.
- Douglas Jack D. The Social Meanings of Suicide, Princeton, Princeton University Press, 1967.
- Dubar, Claude (2004). « À propos de l'interprétation du Suicide de Durkheim par Philippe Besnard », Revue européenne des sciences sociales, XLII-129, pp. 365-373.
- Durkheim E. Sociologie et philosophie. Paris, PUF, 2014.
- Durkheim E. Les Règles de la méthode sociologique, 1895
- Durkheim E. Le Suicide, Paris, Alcan, 1897
- Latour B. (2005). Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford: Oxford University Press.
- Latour, B. 2010. Tarde's idea of quantification, in M. Candei (ed.), The social after Gabriel Tarde : debatesand assessments, 145–62. London : Routledge.
- Lubek Ian, Apfelbaum Erika (1981). Histoire de psychologies sociales perdues. Le cas de Gabriel Tarde. In: Revue française de sociologie, 22-3. Sociologies françaises au tournant du siècle. Les concurrents du groupe durkheimien. Etudes réunies par Philippe Besnard., sous la direction de Philippe Besnard. pp. 361-395.
- Merllié D. (2004). « Pistes de recherche pour une sociologie des statistiques du suicide. Note sur « Anti- ou anté-durkheimisme », Revue européenne des sciences sociales, XLII-129, 249-259.
- Mucchielli Laurent (dir.) et Massimo Borlandi (dir.), La Sociologie et sa méthode : les règles de Durkheim un siècle après, Paris, L'Harmattan, 1995, 415 p
- Mucchielli Laurent, « Tardomania ? Réflexions sur les usages contemporains de Tarde », Revue d'Histoire des Sciences Humaines, 2/2000 (no 3), p. 161-184.
- Paugam Serge, préface d'une édition du Suicide
- Tarde G. Essais et mélanges sociologiques. Lyon : A. Storck, Paris : G. Masson, 1900.
- Tarde G. Etudes pénales et sociales. Lyon : A. Storck, Paris : G. Masson, 1892.
- Tarde G. La criminalité comparée. Paris : Libraire Félix Alcan, 1924.
- Tarde G. Les lois de l'imitation, 1890
- Tarde G. La logique sociale, 1895
- Tarde, G. « Contre Durkheim à propos de son Suicide », 1897
- Taylor S. Durkheim and the study of suicide. London : Macmillan, 1982.
- Vargas E.V. Latour B., Karsenti B. et Aït-Touati F. Reprise du débat Tarde-Durkheim décembre 1903 [en ligne], 2008, http://www.bruno-latour.fr/sites/default/files/downloads/TARDE-DURKHEIM-GB.pdf.