Université de Montréal

Les mises en forme algorithmiques, ruptures et continuités dans la quantification du social

par
Justine Lareau

Département de sociologie
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de Maître ès sciences (M. Sc.)
en sociologie

Août 2021

© Justine Lareau, 2021

Université de Montréal
Département de sociologie, Faculté des études supérieures et postdoctorales

---

Ce mémoire intitulé

Les mises en forme algorithmiques, ruptures et continuités dans la quantification du social

Présenté par

Justine Lareau

A été évalué par un jury composé des personnes suivantes

Marianne Kempeneers
Présidente-rapporteure

Paul Sabourin
Directeur de recherche

Stéphane Moulin
Codirecteur

Jacques Hamel
Membre du jury

Résumé

Ce mémoire de maîtrise porte sur les algorithmes de « data mining » et de « machine learning », constitutifs d'un domaine que l'on appelle plus récemment la « science des données ». Pour essayer d'éclairer la portée et la spécificité des enjeux que leur usage soulève dans nos sociétés, il est proposé d'interroger le rapport qu'ils entretiennent avec les fondements et les limites des outils plus traditionnels de la statistique sociale/mathématique, bien documentés en sociologie, à l'égard notamment du « langage des variables » et du raisonnement expérimental « toutes choses égales par ailleurs ».

En inscrivant l'approche au croisement de la sociologie de la connaissance et de la quantification, le cadre conceptuel s'inspire de l'épistémologie comparative de Gilles-Gaston Granger, de la « méta-épistémologie historique » de Ian Hacking et de la sociohistoire de la statistique sociale d'Alain Desrosières. Par l'idée de *mises en forme algorithmique de la vie sociale*, les algorithmes de calcul sont envisagés comme modes d'investigation, partiellement ou complètement automatisés, procédant à des mises en forme et en ordre plurielles et différenciées du social et de ses propriétés.

À partir de données de Statistique Canada servant à étayer plus concrètement les *formes* de connaissances produites et les visées d'*objets* qu'elles délimitent en termes de possibilités et de contraintes d'expérience, la présente étude de cas entreprend d'examiner le clivage des méthodes « classiques » et « contemporaines » à l'intérieur du cadre supervisé de l'apprentissage. Pour ce faire, trois techniques/familles d'algorithmes sont comparées sous l'angle de leurs *opérations* d'analyse: 1) les méthodes de régression logistique, 2) les arbres de décision et 3) les forêts aléatoires. L'objectif de cette analyse sociologique théorique comme empirique est d'examiner comment ces approches opèrent certains modes de classification et facilitent ou défavorisent des représentations du monde et de l'individu.

Le travail conduit plus généralement à ouvrir quelques pistes de réflexion quant aux rapports de compatibilité et d'incompatibilité des formes de raisonnement du style statistique et probabiliste avec certains états du développement de la sociologie.

Mots-clés : algorithmes d'apprentissage, exploration de données, science des données, sociologie, épistémologie, statistique sociale, analyse de données, méthodes/techniques de recherche, sciences sociales, quantification.

Abstract

This master's thesis focuses on data mining and machine learning algorithms, constituting a field more recently called “data science”. To try to shed light on the specificity of the issues they raise in our societies, it is proposed to question the relationship they maintain with the foundations and the limits of the more “classic” tools of mathematical statistics in sociology, with regard in particular to the “language of variables” and to the experimental reasoning “all other things being equal” (*cetaris paribus*).

By placing the approach at the intersection of the sociology of knowledge and quantification, the conceptual framework is inspired by the comparative epistemology of Gilles-Gaston Granger (1920-2016), the historical meta-epistemology of Ian Hacking (1936-) and the sociohistory of social statistics by Alain Desrosières (1940-2013). Through the idea of “mises en forme algorithmique de la vie sociale”, computational algorithms are considered as partially or completely automated types of investigation, carrying out plural and differentiated of shaping and ordering of the social and its properties.

Using data from Statistics Canada used to more concretely support the forms of knowledge produced as well as the possibilities and experience constraints that they define, this case study sets out to examine the divide between “classical” and more “contemporary” methods of analysis within the framework of “supervised” learning. To do this, three algorithm techniques (or families of algorithms) are compared from the angle of their knowledge operations: 1) logistic regressions, 2) decision trees and 3) random forests. The objective of this theoretical as well as empirical work is to examine how these approaches operate certain modes of classification, facilitate or disadvantage representations of the world and can also be performative in social activities.

The research work more generally leads to opening up some avenues of reflection as to the compatibility and incompatibility relationships of the forms of reasoning of the statistical and probabilistic style with certain states of development in society and in sociology.

Keywords: learning algorithms, data mining, data science, sociology, epistemology, research methodology, social statistics, data analysis, quantification.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

BMS : Bulletin de Méthodologie Sociologique

CART : Classification and Regression Trees

CRSH : Conseil de recherches en sciences humaines

DM : Data Mining

DS : Data Science

DT : Decision Trees

FAS : Faculté des arts et des sciences

FRQSC : Fonds de Recherche du Québec – Société et Culture

GLM : Generalized Linear Models

IA : Intelligence artificielle

LR : Logistic Regression

KDD : Knowlegde Discovery in Databases

ML : Machine Learning

R : Environnement (ou langage de programmation)

RF : Random Forest

SHS : Sciences Humaines et Sociales

VD : Variable dépendante

VI : Variable indépendante

À la sociologie.

Remerciements

Je tiens d'abord à remercier mon directeur de recherche, Paul Sabourin pour l'autonomie et la liberté intellectuelle qu'il m'a offert, pour sa patience, sa générosité et sa bienveillance tout au long de l'avancement de mon projet de recherche. Je remercie aussi Stéphane Moulin d'avoir accepté de codiriger ce mémoire de maîtrise et de m'avoir initié au vaste champ de recherche de la sociologie de la quantification.

Je tiens également à exprimer mes remerciements à Eric Lacourse et à Stéphanie qui m'ont permis de découvrir l'univers des algorithmes de l'intelligence artificielle à leurs côtés. Le choix du sujet de cette recherche n'aurait probablement pas été envisagé sans tous nos échanges, et je leur en serai toujours reconnaissante. Je tiens également à remercier Marianne Kempenners, Estelle Carde, Pierre Hamel, Cécile Van de Velde ainsi que Céline Lafontaine qui, par leurs diverses formes d'encouragement à différents moments clés de mon parcours universitaire en sociologie, ont su me donner le courage de persévérer au sein de cette discipline. Merci à Claire Durand et à Anne Calvès qui, par leurs commentaires et leurs critiques constructives, m'ont amené à cheminer encore plus loin dans mes réflexions. Je remercie aussi Jacques Hamel pour avoir bien voulu que je participe au dernier séminaire qu'il donnait *Epistemologie et méthodologie qualitative*. Non seulement l'enseignement fut très stimulant, mais les discussions ont été très fécondes pour ce mémoire.

De plus, je tiens à exprimer ma gratitude envers Louise-Andrée et Lamyae qui, par leur empathie et leur écoute, m'ont grandement aidé à cheminer comme entité humaine. Des remerciements tout particuliers s'adressent également à Didier Fayon pour sa disponibilité, ses conseils pratiques et toutes les réflexions partagées avec moi.

Je remercie aussi le Conseil de recherche en sciences humaines du Canada (CRSH), les Fonds de recherche du Québec – Société et culture (FRQSC) et la Faculté des arts et des sciences de l'Université de Montréal pour le soutien financier dont a bénéficié ce projet de recherche.

Finalement, un profond merci à mes parents, Marie et Jean, qui ont été des piliers essentiels à l'aboutissement de ce travail. À ma sœur Ariane, merci d'avoir toujours su m'éclairer avec lucidité tant dans tes paroles que par tes grands silences.

INTRODUCTION GÉNÉRALE

> « A la tentation toujours renaissante de transformer les préceptes de la méthode en recettes de cuisine scientifique […] on ne peut opposer que l'entraînement constant à la vigilance épistémologique qui […] interdit les facilités d'une application automatique de procédés »
>
> Pierre BOURDIEU, Jean-Claude CHAMBERON et Jean-Claude PASSERON, 1968, *Le métier du sociologue.*

Ce projet de recherche porte sur les algorithmes de « data mining » (DM dans la suite), relevant du « machine learning » (ML), que l'on définit comme un ensemble de techniques statistiques et informatiques[1]. Ces méthodes de calcul sont au cœur d'un domaine apparu au début des années 2000 que l'on appelle la « science des données » (DS pour *data science*). Dans le cadre de ce mémoire de maîtrise, il est proposé de s'y intéresser sous un angle plus spécifique, celui des liens que ce domaine aujourd'hui en plein essor entretient avec celui de la « statistique sociale »[2].

Le fil conducteur de ce travail est la tension « méthodologique » entre certaines techniques d'analyse plus anciennes et d'autres plus récentes qui coexistent dans le développement de la DS et de ses applications. Comment expliquer l'émergence d'une DS dans ses rapports avec la statistique sociale comme tradition d'enquêtes bien connue et largement discutée en sciences sociales, en particulier dans une discipline comme la sociologie reconnue pour son « assise statistique »[3] ?

---

[1] L'« apprentissage machine » (ML) est l'un des sous-domaines de l'intelligence artificielle (IA), que l'on traduit souvent en français par « apprentissage automatique », parfois « apprentissage statistique » pour la référence plus explicite à la théorie mathématique sous-jacente. Les algorithmes dits d'apprentissage sont couramment désignés par d'autres expressions, notamment celle d'« exploration de données », de « fouille » ou de « forage » des données (DM) (chapitre I).

[2] Dans ce mémoire, par statistique sociale, on entend des moyens d'investigation statistique – d'observation et d'expérience empirique, ou encore d'analyse informatisée – appliqués pour appréhender des phénomènes sociaux, des comportements humains.

[3] HÉRAN François, « L'assise statistique de la sociologie », *Economie et Statistique*, vol. 168, n° 1, 1984, p. 23-35.

La plupart des questionnements intellectuels de cet écrit trouve leur origine dans un stage de recherche réalisé à l'été 2019 auprès du professeur titulaire Éric Lacourse. Les apprentissages développés pendant le stage[4] ont été déterminants dans le choix de l'orientation de ce mémoire, en ayant permis d'appréhender le vaste domaine que représente l'intelligence artificielle (IA) sous l'angle des méthodes algorithmiques qui s'y développent. Plus précisément, l'application concrète des algorithmes d'apprentissage sur des données portant sur le décrochage scolaire a constitué une source foisonnante de questionnements relatifs au statut des méthodes et des données dans la production de connaissance. Confrontée au problème de « l'opacité algorithmique » relativement courant dans la démarche en ML (chapitre I), diverses tentatives plus ou moins fructueuses ont été poursuivies après le stage pour reconstruire les modèles qui émergeaient des analyses (examen des sorties informatiques non pertinentes socialement, calculs à la mitaine, programmation des fonctions mathématiques, visualisations graphiques…). L'idée générale proposée dans ce mémoire de *mises en forme algorithmique de la vie sociale* est le résultat de l'ensemble de ces expériences « de terrain » et de ces exercices « pratiques » de décomposition des techniques d'algorithmes.

Par cette idée, les algorithmes de calcul sont considérés à la fois comme *méthode* d'analyse statistique et probabiliste et comme *objet* d'investigation sociologique. Nous voulons montrer *comment* (ou par quels processus) les algorithmes pourraient-ils faire « parler » des données empiriques du monde social. La première appréhension à l'origine du projet est que les usages opératoires des mathématiques[5] intégrés dans la conception et le développement des algorithmes peuvent être relatifs à certains états des sociétés. Cette vague intuition à peine plus développée est posée comme horizon (ou trame de fond) des réflexions, car diverses questions d'ordre théorique, pratique et empirique surgissent bien avant de pouvoir en formuler des idées plus précises : comment comparer des techniques d'analyse autrement qu'en termes de performances prédictives, selon des critères de précision usuels en ML (matrice de confusion, courbe ROC, etc.[6]) ? Est-ce

[4] Notamment la formation intensive organisée par l'Association américaine de psychologie (American Psychological Association) à l'Arizona State University intitulée « Big Data : Exploratory Data Mining in Behavioral Research » (août 2019).

[5] À l'instar d'Olivier Martin, « le terme « mathématique » désigne à la fois les formalismes non-quantitatifs (comme la théorie des groupes, celles des réseaux ou des jeux, l'algèbre booléenne…) et les outils de la statistique et du calcul des probabilités. » MARTIN Olivier, « Mathématiques et sciences sociales au XXème siècle », *Revue d'Histoire des Sciences Humaines*, vol. 1, nº 6, 2002, p. 3.

[6] Matrice de concordance, précision, sensibilité, spécificité, aire sous la courbe AUC, score F1, erreur Out-Of-Bag, etc.

concevable mentalement, théoriquement[7] ? Est-ce faisable techniquement, concrètement ? Serait-ce intelligible sociologiquement ? Quel en serait l'intérêt scientifiquement, cognitivement ? Dans quelle mesure une telle comparaison pourrait-elle être reconnue comme « sociologique » fondamentalement ?

Notre préoccupation est double. Tel que l'invitait le sociologue du numérique Dominique Cardon, nous croyons « nécessaire d'entrer dans les calculs, d'explorer leurs rouages […] [a]vant de réduire la logique calculatoire aux intérêts économiques de ceux qui la fabriquent »[8]. Sachant toutefois que des outils aussi anciens que la régression linéaire coexistent avec les nouveautés de la DS et font partie des premières approches enseignées en ML, nous nous demandons comment cerner plus finement la portée des changements qu'induirait les formes de quantification « algorithmique » contemporaines par rapport aux enjeux que soulevait déjà la « raison statistique » associée à l'État[9]. Notre étude permettra entre autres de montrer que la compatibilité des nouvelles formes d'analyse statistique et probabiliste avec certains modes d'individuation sociale dans les sociétés occidentales pluralistes peut difficilement s'expliquer par l'avènement d'une rationalité « immanente »[10], susceptible de véhiculer l'idée d'une existence dématérialisée, d'un monde éclaté où « l'individu calculé est un flux »[11].

Dans le cadre d'une méthodologie sociologique d'étude de cas, ce mémoire se bornera à trois techniques/familles d'algorithmes d'apprentissage supervisé[12], apparaissant particulièrement expressives de l'antinomie des méthodes « classiques » et « contemporaines » (chapitre 3) : 1) les régressions logistiques (LR pour *logistic regression*), 2) les arbres de décision (DT pour *decision*

---

[7] Des termes distincts ont été utilisés pour souligner que tout ce qui nous passe par la tête, toutes les activités de pensée ne sont pas nécessairement, « spontanément » d'ordre « théorique ».

[8] CARDON Dominique, *À quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Paris, Seuil, coll. « La République des idées », 2015, p. 13.

[9] Voir les travaux pionniers en sociologie d'Alain Desrosières (1940-2013), notamment son livre majeur *La politique des grands nombres. Histoire de la raison statistique* paru pour la première fois en 1993, enrichi d'une postface de l'auteur en 2000.

[10] ROUVROY Antoinette et Thomas BERNS, « Gouvernementalité algorithmique et perspectives d'émancipation », *Réseaux*, n° 177, nº 1, 2013, p. 163-196.

[11] CARDON Dominique, À quoi rêvent les algorithmes, op. cit., p. 77.

[12] Indiquons simplement que la classe d'algorithmes dite « supervisée » (ou « dirigée ») fait référence aux situations qui exigent de différencier préalablement le statut des variables de l'analyse, en désignant au moins une variable d'intérêt – dite « dépendante », aussi appelée à expliquer/prédire – parmi d'autres, qualifiées d'« indépendantes », susceptibles d'être potentiellement explicatives, prédictives ou descriptives. À l'inverse, l'apprentissage « non supervisé » désigne des formes d'analyse qui ne reposent pas sur ce formalisme asymétrique entre les données ou les variables (par exemple, les analyses factorielles).

trees) et 3) les forêts aléatoires (RF pour *random forests*)[13]. L'analyse comparée visera à examiner la manière dont leurs logiques d'analyse respectives opèrent certains modes de classification, facilitent ou défavorisent des représentations du monde et de l'individu. À cette fin, une base de données de Statistique Canada sera mobilisée à titre illustratif pour étayer plus concrètement, au besoin, les implications cognitives et pratiques de ces outils dans l'appréhension d'un phénomène social donné (quelconque). L'exemple de l'étude des perceptions de santé servira de problématique de référence empirique, « appliquée ». Afin de tenir compte « des *continuités* que les complexes formalisations mathématiques contemporaines pourraient dissimuler »[14], le travail revisitera les critiques internes et externes couramment formulées à l'égard de la méthodologie statistique/« quantitative » en sociologie, et notamment du « langage des variables » et du raisonnement expérimental « toutes choses égales par ailleurs ».

Notre réflexion à propos de l'articulation des formes mathématiques aux formes sociales conduira plus généralement à constater des possibles rapprochements avec certains états de la sociologie. Il importe d'indiquer d'emblée que, dans le cadre restreint de ce travail, il s'agira surtout d'ouvrir quelques pistes de réflexion en filigrane de l'analyse afin de montrer l'importance de garder certaines nuances à l'esprit face au « positivisme algorithmique » des dispositifs sociotechniques contemporains en sciences humaines et sociales (SHS). En cherchant moins à justifier les rapports de compatibilité, nous supposerons au contraire que le renouvellement des formes de quantification du monde social est sans rapport avec l'engouement que semblent susciter parallèlement certaines formes de recherches sociologiques en SHS depuis les quarante dernières années.

Nous pourrons dès lors constater que cet écrit se veut plus fondamentalement une contribution à la sociologie générale, un exercice de réflexion méthodologique et épistémologique interrogeant de front une série d'oppositions et d'associations « conventionnelles[15] », « ritualisées, institutionnalisées »[16], devenues « canoniques » à travers le débat quanti/quali en sociologie,

---

[13] Par ces termes, nous nous référons aux techniques telles qu'elles ont été proposées par le statisticien Léo Breiman (1928-2005).

[14] DESROSIÈRES Alain, « Masses, individus, moyennes: la statistique sociale au XIXe siècle », *Hermès*, vol. 2, nº 2, 1988, p. 61 (nous soulignons).

[15] Dans ce mémoire, les aspects « conventionnels » désignent des éléments ni complètement nécessaires, ni totalement arbitraires. Voir les trois auteurs mobilisés dans le chapitre II, G.-G. Granger, I. Hacking et A. Desrosières.

[16] BECKER Howard, *Faire preuve: Des faits aux théories*, Paris, la Découverte, 2020, p. 56.

« paraissant aujourd'hui trop évidentes pour être discutées »[17] (explication/compréhension, paradigmes positiviste/interprétatif, objectivisme/subjectivisme, macro-/micro-sociologies, etc.)[18]. Nous verrons ultimement que l'approche de la sociologie de la connaissance et de la quantification, ici privilégiée, conduit à poser davantage de questions à la sociologie, à son « objet » (la société) et à ses longues traditions d'enquêtes qu'aux nouvelles sciences sociales computationnelles. D'ailleurs, comment se fait-il que la sociologie semble éprouver plus de difficultés à faire reconnaitre son statut et son autonomie comme « science du social »[19] que l'éminente « science » des « données brutes parlantes » en pleine expansion aujourd'hui dans nos sociétés ?

Ce mémoire de recherche est divisé en cinq chapitres. Le premier chapitre se consacre d'abord à une recension des travaux dans la littérature récente portant sur les algorithmes de DM et de ML. Nous verrons en somme que les fondements théoriques et pratiques de la DS s'inscrivent dans l'histoire de la statistique, de ses enjeux et de ses usages en société. Afin de mieux situer la perspective théorique et méthodologique du présent projet, l'originalité de la contribution espérée et son éventuelle pertinence sociologique, il se conclut en dégageant une cartographie de quatre tendances de recherche idéales-typiques sur le sujet : *philosophique*, *professionnelle*, *ethnographique* et *statistique/méthodique*.

---

[17] HAMEL Jacques, « Pour la méthode de cas. Considérations méthodologiques et perspectives générales », *Anthropologie et Sociétés*, vol. 13, nº 3, 1989, p. 59.

[18] Rappelons que « la première moitié du XXe siècle » est souvent caractérisée par le règne du « positivisme », lequel est généralement reconnu par le recours massif aux mathématiques pour appréhender et pour gérer les sphères des activités humaines. Sous l'impulsion de P. F. Lazarsfeld, figure de proue de l'École de Colombia, cette domination fut intimement liée à des enjeux relatifs à la recherche appliquée et aux politiques « éclairées » des États dans la période d'après-guerre. Face à l'hégémonie de la recherche dite quantitative, certains ont vivement contesté les lacunes du formalisme statistique et de son langage décontextualisant par variables, parmi lesquels H. Blumer, fondateur de l'interactionnisme symbolique, considéré comme l'un des courants, avec la phénoménologie, à la base du paradigme « interprétatif ». La prolifération d'ouvrages et de manuels issus de cette tradition intellectuelle au tournant des années 70 aurait marqué un regain d'intérêt pour les études « de terrain » et les techniques d'observation participante dites qualitatives dans le paysage de la sociologie nord-américaine. Mais comme le rappelle Pirès, les « mouvements d'expansion et de développement [d'un type de recherche] *ne vont pas dans le sens d'une substitution* » de l'autre (Pirès, 1987 : 94). Voir PIRÈS Alvaro, « Deux thèses erronées sur les lettres et les chiffres », *Cahiers de recherche sociologique*, vol. 5, nº 2, 1987, p. 85-105 ; BRYMAN Alan, « Quantitativisme et qualitativisme: un faux débat ? », dans Jean-Michel BERTHELOT (dir.), *Sociologie, épistémologie d'une discipline : textes fondamentaux*, Paris, De Boeck Université, 2000, p. 209-220 ; ANADÓN Marta, « Les méthodes mixtes : implications pour la recherche « dite » qualitative », *Recherches qualitatives*, vol. 38, nº 1, 2019, p. 105-123.

[19] SABOURIN Paul, « Sociologie, éthique et politique : itinéraire d'une éthique dans la recherche pour une coopération sociologique élargie », *Sociologie et sociétés*, vol. 52, nº 1, 2020, p. 19-46.

Le deuxième chapitre entreprend ensuite d'esquisser les prémisses de ce que *pourrait* être une approche de sociologie de la connaissance et de la quantification cherchant à appréhender la constitution *sociale* de la construction mathématique (dynamique, cumulative et performative) des algorithmes de calcul. Sont succinctement exposés les repères théoriques fondamentaux du mémoire, inspirés plus particulièrement de l'épistémologie comparative de Gilles-Gaston Granger (1920-2016), de la méta-épistémologie historique de Ian Hacking (1936-) et de la sociohistoire de la statistique d'Alain Desrosières (1940-2013). Cette partie sera ensuite suivie de remarques supplémentaires à propos de la problématique de recherche fondamentale, ou de sociologie *générale* dans laquelle s'inscrit ce mémoire.

Dans un troisième temps, le chapitre méthodologique commence par préciser les objectifs de la recherche en regard du cadre théorique précédemment exposé, puis opérationnalise la conceptualisation des algorithmes de calcul. Sont présentés par la suite les données empiriques de l'*Enquête sociale générale* de 2016 (ESG2016) de Statistique Canada mobilisées pour construire les données sociologiques, puis les autres formes de « supports » servant à l'analyse comparée des algorithmes. Après avoir introduit sommairement les principes généraux des trois méthodes d'apprentissage supervisé étudiées, ce chapitre se termine par de brèves remarques concernant le processus et les modalités de la description et de l'analyse sémantique et pragmatique.

Enfin, le travail d'analyse et les résultats sont présentés en deux chapitres. Le quatrième interroge les limites des formes « classiques » de la statistique sociale/mathématique (LR), telles que certaines critiques épistémologiques et méthodologiques en sociologie les ont identifiées. Repérés en grande partie dans les débats de méthode, ces éléments serviront de référents analytiques pour examiner, au chapitre suivant, les deux techniques plus contemporaines choisies que sont les arbres (DT) et les forêts (RF). Dans le cinquième et dernier chapitre, nous essayerons ainsi d'y voir comment celles-ci contribuent notamment à réactualiser des enjeux classiques de catégorisation sociale dans une optique désormais toutes choses *différentes* ou *inégales* par ailleurs.

À l'issue de cet exercice, nous proposons une synthèse des éléments d'observation et de réflexion ouverte, collective et publique quant aux conclusions provisoires qui peuvent en être tirées, avant de brièvement discuter à nouveau du rapport entre statistique, société et sociologie. Face aux nombreuses limites de ce projet de sociologie *en train de se faire*, indiquons que le degré

d'approfondissement des liens qu'aura simplement cherché à dégager ce texte est essentiellement tributaire des nécessités physiques et biologiques de l'existence humaine[20].

[20] Je remercie Paul Sabourin pour ses judicieux conseils dès nos premières rencontres (ne pas associer la volonté de faire de la recherche à la « mort », s'inscrire dans une perspective de « longue durée »…)

CHAPITRE I – REVUE DES ÉCRITS

Entre rupture et continuité

Qu'ils soient utilisés à titre d'outils de prédiction, de classification ou d'aide à la décision, les algorithmes d'« apprentissage » sont de plus en plus utilisés en société, dans des pans toujours plus variés de la vie quotidienne, en passant par la santé, l'éducation, le droit, les assurances, l'emploi, la sécurité, la recherche scientifique, le marketing, l'établissement des relations amoureuses, et bien d'autres domaines.

La question de départ peut se formuler ainsi : comment les travaux récents ont-ils caractérisé le rapport entre les formes mathématiques contemporaines, dites « algorithmiques » et celles plus classiques en statistique ? Le premier chapitre procède en trois temps. D'abord, nous présentons une brève description historique des conditions matérielles et institutionnelles qui ont permis d'ériger la « science des données » (DS) comme discipline à prétention scientifique. Ensuite, il s'agit de survoler certaines contributions plus théoriques qui ont examiné en quoi l'intégration croissante des techniques d'analyse plus récentes apparaît comme un bouleversement « inédit » de nos sociétés par rapport aux outils statistiques plus anciens ou « conventionnels ». Enfin, nous abordons deux types d'enquêtes empiriques qui ont montré quelques continuités avec la tradition statistique, en prenant les algorithmes, soit comme objets d'étude (*dans la pratique*), soit comme méthodes de recherche alternative (*en pratique*).

Ce bilan des écrits sélectif se termine en discutant des limites de quatre grandes tendances de recherche portant sur ces nouveaux dispositifs sociotechniques dans l'état actuel des connaissances (*philosophique*, *professionnelle*, *ethnographique*, *statistique/méthodique*). Ce sera en fonction de cette typologie, que nous tenterons d'esquisser, au chapitre suivant, les fondements théoriques généraux de ce que pourrait être une approche sociologique des logiques de calcul des algorithmes statistiques et informatiques.

1.1 L'historique des algorithmes

1.1.1 La « mise en données » des sociétés

D'abord, plusieurs études en sciences sociales inscrivent l'engouement accru que suscite ces méthodes de calcul informatique dans une période historique du développement rapide des technologies de l'information et de la communication (TIC) – objets connectés, capteurs et applications mobiles – qui mettent à jour les flux d'information en continu. Couplé à l'augmentation des capacités de stockage et de calcul des systèmes informatiques, ce « déluge » de traces numérisées de la vie sociale, diversifiées et agrégées dans d'énormes « entrepôts » de données (*big data*) serait lié au vaste mouvement de la « quantification » des sociétés[21]. Aujourd'hui, plusieurs font référence au prolongement de cette tendance par le néologisme de « datafication »[22], entendu comme des processus de « mise en données » convoitant désormais l'appréhension de « nos relations, nos expériences et nos états d'âme »[23].

Outre les nombreuses critiques autour du phénomène *big data*[24], plusieurs recherches ont récemment insisté sur l'importance en SHS (*sciences humaines & sociales*) de se pencher sur les procédures algorithmiques qui automatisent la collecte, le traitement et le croisement des masses de données, et qui permettent l'« extraction » de leur « valeur » culturelle, politique et financière[25]. Concept ancien qui tire ses origines étymologiques du mathématicien perse Al-Khwârismî au IXe siècle[26], un « algorithme » consisterait au sens le plus large en un « procédé qui permet de résoudre un problème, sans avoir besoin d'inventer une solution à chaque fois »[27].

---

[21] À ne pas confondre avec la « numérisation » (*digitalization*). *Cf.* REY Olivier, *Quand le monde s'est fait nombre*, Paris, Stock, 2016 ; CHIAPELLO Ève, Corine EYRAUD, Philippe LORINO, Alain SUPIOT, Ève LAMENDOUR et Yannick LEMARCHAND, « À propos de l'emprise du chiffre », *Entreprises et histoire*, vol. 2, nº 79, 2015, p. 174-187.

[22] SADIN Éric, La vie algorithmique: critique de la raison numérique, Paris : Éditions l'Échappée, 2015.

[23] MAYER-SCHNBERGER Viktor et Kenneth CUKIER, *Big Data: A Revolution That Will Transform How We Live, Work and Think*, Londres, John Murray, 2013, p. 91.

[24] BOYD Danah et Kate CRAWFORD, « Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly phenomenon », *Information, Communication & Society*, vol. 15, nº 5, 2012, p. 662-679 ; KITCHIN Rob, « Big Data, New Epistemologies and Paradigm Shifts », *Big Data & Society*, vol. 1, nº 1, 2014, p. 1-12.

[25] ROUVROY Antoinette et Thomas BERNS, « Gouvernementalité algorithmique et perspectives d'émancipation », *op. cit.* ; CARDON Dominique, *À quoi rêvent les algorithmes*, *op. cit.* ; O'NEIL Cathy, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, Penguin Books, 2016.

[26] ABITEBOUL Serge et Gilles DOWEK, *Le temps des algorithmes*, Paris, Le Pommier, 2017, p. 14.

[27] *Ibid.*, p. 11.

Grâce aux avancées technologiques (ordinateurs, programmes informatiques) qui permettraient « d'exécuter sans réfléchir » diverses manières de *faire* des choses, indépendamment presque du domaine d'application[28], plusieurs ont suggéré une profonde altération dans nos manières de connaître, d'éprouver et d'interagir avec le monde[29]. Concernant la démarche scientifique elle-même, l'ancien rédacteur en chef de la revue *Wired*, Chris Anderson annonçait il y a plus de dix ans déjà le déclin des « théories sur le comportement humain, de la linguistique à la sociologie » face à l'ascension d'une « science » dirigée par l'abondance des données empiriques disponibles (*data-driven science*). Ce dernier écrivait dans son article[30] :

> Les pétaoctets nous permettent de dire: "La corrélation suffit." […] Nous pouvons analyser les données sans hypothèse sur ce que cela pourrait montrer. Nous pouvons jeter ou injecter les chiffres dans les plus grandes grappes informatiques que le monde n'ait jamais vu, et laisser les algorithmes trouver des schémas ou des configurations là où la science en est incapable […] La corrélation prime sur la causalité, et la science peut progresser même sans modèles cohérents, théories unifiées ou explications mécanistes. […] Il est temps de demander: Que peut apprendre la science de Google ?

Dans le mouvement incarné de manière emblématique par Anderson, il serait ainsi non seulement envisageable, mais même souhaitable de « laisser parler les chiffres par eux-mêmes » pour pouvoir produire des connaissances hypothétiques et opératoires d'autant plus objectives, neutres, impartiales et inusitées[31].

Au sein de la communauté universitaire, ce type de discours que certains désignent par l'« idéologie technique des big data »[32] ou encore la philosophie du « dataïsme »[33] se voit souvent dénoncé, sinon rapidement balayé, sous prétexte de n'être que la résurgence d'une idéologie « positiviste » naïve, radicale et provocatrice. On ne cessera de le répéter aux apprentis sociologues

---

[28] Ibid., p. 26. *Cf*. BERRY Gérard, *L'Hyperpuissance de l'informatique: Algorithmes, données, machines, réseaux*, Paris, Odile Jacob, 2017.

[29] AMOORE Louise et Volha PIOTUKH, « Life Beyond Big Data Governing With Little Analytics », *Economy and Society*, vol. 44, nº 3, 2015, p. 341-366 ; BASTIN Gilles et Paola TUBARO, « Le moment big data des sciences sociales », *Revue française de sociologie*, vol. 59, nº 3, 2018, p. 375-394.

[30] ANDERSON Chris, « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired Magazine*, vol. 16, nº 7, 23 juin 2008. En ligne au <https://www.wired.com/2008/06/pb-theory/> (traduction libre).

[31] WILLIAMS Graham, *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*, New York: Springer-Verlag, 2011 ; MAYER-SCHNBERGER Viktor et Kenneth CUKIER, Big Data: A Revolution That Will Transform How We Live, Work and Think, *op. cit*., p. 6.

[32] ROUVROY Antoinette et Thomas BERNS, « Gouvernementalité algorithmique et perspectives d'émancipation », *op. cit.*

[33] THÉVENOT Laurent, « Mesure pour mesure : formes d'enquête, d'évaluation et de gouvernement, depuis la statistique d'État jusqu'au « soi quantifié » », dans le cadre du colloque *Histoire aujourd'hui, statistiques demain : regards croisés sur la production et l'usage des statistiques* présenté à Paris, au centre de conférences Pierre-Mendès-France, 29 juin 2016.

dont je fais partie : « les données ne sont pas données »[34] et « parlent » encore moins d'elles-mêmes, mais par l'entremise de la lecture d'un-e chercheur-e.

Pourtant, si l'on regarde uniquement à l'Université de Montréal, on s'étonnera de la mise en place de nouveaux programmes consacrés à l'enseignement des rudiments de cette « science » issue des *big data* parmi les premiers qui seraient offerts au Québec[35]. C'est ainsi que, comme l'indiquaient les sociologues Éric Dagiral et Sylvain Parasie[36], le scepticisme initial des SHS face au virage inductiviste de la production de connaissance s'avère insuffisant pour expliquer l'utilisation progressive et continue de ces techniques de calcul spécialisées au sein de secteurs d'activité toujours plus nombreux et variés. Aux yeux de ces chercheurs d'inspiration STS (*Science & Technology Studies*), cette situation justifierait l'intérêt de se pencher sur ce qui compose la DS comme discipline institutionnalisée, sur ce qu'elle recouvre *concrètement* en société, matériellement (techniquement) et actuellement (historiquement), plutôt que de se buter à l'expression « big data » qu'ils considèrent comme inopérante analytiquement.

1.1.2 Les techniques de l'« analyse »

Selon Dagiral et Parasie, l'expression relativement récente[37] de « science des données » (DS) vise empiriquement « un assemblage de pratiques, de savoirs et de technologies situées au croisement de l'informatique et des statistiques »[38]. De nos jours, la notion de DS désigne une nouvelle profession, celle des « scientifiques de données » (*data scientists*)[39]. En 2012, la prestigieuse *Harvard Business Review* qualifiait d'ailleurs ce nouveau métier comme étant « le plus sexy du XXI siècle ». De plus en plus recrutés sur le marché du travail, ces praticiens sont appelés

---

[34] Voir HOULE Gilles et Nicole RAMOGNINO, « Présentation. La construction des données », *Sociologie et sociétés*, vol. 25, nº 2, 1993, p. 5-9 ; BERNARD Paul, « L'insignifiance des « données » : bref essai contre la stigmatisation positiviste », *Sociologie et sociétés*, vol. 14, nº 1, 1982, p. 65-82. « encore une fois, les données ne parlent pas d'elles-mêmes » (Gauthier, 2009 : 176).

[35] SAUVÉ Mathieu-Robert, « Une orientation Science des données s'ajoute au baccalauréat en mathématiques et informatique », *UdeMNouvelles*, 12 décembre 2018. En ligne au <https://nouvelles.umontreal.ca/article/2018/12/12/une-orientation-science-des-donnees-au-baccalaureat-en-mathematiques-et-informatique/>.

[36] DAGIRAL Éric et Sylvain PARASIE, « La "science des données" à la conquête des mondes sociaux: ce que le "Big Data" doit aux épistémologies locales », dans Pierre-Michel MENGER et Simon PAYE (dir.), *Big data et traçabilité numérique: Les sciences sociales face à la quantification massive des individus*, Collège de France, 2017, p. 85-104.

[37] Proposée en 2001 par le chercheur statisticien des Bell Labs William S. Cleveland (?). (Dagiral et Parasie, 2017: 91).

[38] DAGIRAL Éric et Sylvain PARASIE, « La «science des données» à la conquête des mondes sociaux », *op. cit.*, p. 86.

[39] ABITEBOUL Serge et Gilles DOWEK, *Le temps des algorithmes*, *op. cit.*

à concevoir des « produits » et des « services », que l'on appelle des « data products », et qui circulent dans une diversité de mondes sociaux[40]. Cette communauté se rassemble dans des challenges/concours *Kaggle*[41] et mobilise des logiciels et langages de programmation comme *R* ou *Python* qui donnent corps à la DS et participent à sa diffusion sur la scène internationale. En dépit de ces « nouveautés », certains ont néanmoins relevé la profonde ambiguïté autour de ce qu'est (ou serait) la DS en tant que « nouvelle discipline scientifique » et « sa relation avec les statistiques »[42]. Pour le chercheur Étienne Ollion qui s'est intéressé à la DS à la lumière de l'histoire de la sociologie quantitative, non seulement les SHS auraient fait de la DS depuis plusieurs décennies « sans le savoir », mais la DS leur aurait même emprunté « les méthodes et le raisonnement inductif qu'elles mettent de l'avant »[43] (*cf.* section 1.3.2).

D'un point de vue historique, l'avènement de la DS au tournant des années 2000 réactive le champ des statistiques « exploratoires » multidimensionnelles apparu quarante ans auparavant, dès le début des années 1960. Plus précisément, certains travaux ont indiqué que les techniques de calcul et les pratiques d'analyse constitutives de la DS renouent avec un mouvement en faveur de l'exploration des données promu aux États-Unis par John W. Tukey à travers la « data analysis »[44] et en France par le mathématicien Jean-Paul Benzécri qui écrivait jadis que « le modèle doit suivre les données, non l'inverse »[45].

S'étant répandu en réaction à la statistique mathématique dominante en Amérique du Nord centrée sur la recherche « confirmatoire » de modèles préexistants[46], ce courant des statistiques « exploratoires » longtemps minoritaire fut revalorisé dans les années 1990 avec l'émergence d'un domaine né d'intérêts commerciaux et industriels se consacrant surtout à l'exploitation

---

[40] DAGIRAL Éric et Sylvain PARASIE, « La «science des données» à la conquête des mondes sociaux », *op. cit.*, p. 86-87.

[41] *Cf.* BOULLIER Dominique et El Mahdi EL MHAMDI, « Le machine learning et les sciences sociales à l'épreuve des échelles de complexité algorithmique », *Revue d'anthropologie des connaissances [En ligne]*, vol. 14, nº 14-1, 2020, p. 1-33.

[42] DONOHO David, « 50 Years of Data Science », *Journal of Computational and Graphical Statistics*, vol. 26, nº 4, 2017, p. 746-748. Voir aussi Stéphane Mallat (2018), Chaire « Sciences des données » du Collège de France.

[43] OLLION Étienne, « Les sciences sociales, contre la data science ? », *Regards croisés sur l'économie*, n° 23, nº 2, 2018, p. 81.

[44] DONOHO David, « 50 Years of Data Science », *op. cit.* ; DAGIRAL Éric et Sylvain PARASIE, « La «science des données» à la conquête des mondes sociaux », *op. cit.*

[45] BENZÉCRI Jean Paul, *L'analyse des données*, Dunod, 1984, p. 6.

[46] CIBOIS Philippe, « Analyse des données et sociologie », *L'Année sociologique (1940/1948-)*, vol. 31, 1981, p. 333-348.

« secondaire » de grandes bases de données, le « data mining » (DM)[47]. Dans la littérature, ce domaine interdisciplinaire renvoie parfois à un courant plus vaste connu sous le nom de « processus d'extraction (ou de découverte automatisée) de connaissance dans des bases de données » (*knowledge discovery in databases*, KDD). Comme démarche d'analyse, le DM vise à « découvrir des idées à partir de données »[48] sous forme de structures, de règles ou de co-relations auparavant inconnues, qui seraient au socle de la « forme algorithmique de la connaissance »[49]. C'est la signification particulière que prend la notion de « modèle » qui différencierait les techniques de DM de la démarche hypothético-déductive traditionnelle, de la statistique « inférentielle » (observation-hypothèse-vérification[50]). En DM, les modèles de relations par lesquels s'expriment des hypothèses théoriques résultent en sortie, en étant « contingents » aux données fournies en entrée, à partir desquels ils « apprennent » (*cf.* section 1.3.2)[51].

Pour Dagiral et Parasie, qui élargissent au secteur privé la sociohistoire de la statistique « publique » (ou « officielle »), liée aux appareils de l'État, tel qu'initiée dans les travaux pionniers d'A. Desrosières[52], l'expansion des horizons d'application de la DS vient répondre aux instabilités professionnelles des statisticien-ne-s à la fin du 20e siècle[53].

1.1.3 Le raisonnement algorithmique

Dans une publication jugée « historique » au sein de la discipline parue en 2001, le statisticien américain Léo Breiman (1928-2005) invitait les membres de sa communauté des milieux académiques et universitaires à s'engager dans le monde des affaires pour résoudre des

---

[47] BESSE Philippe, Caroline LE GALL, Nathalie RAIMBAULT et Sophie SARPY, « Data mining et statistique », *Journal de la société française de statistique*, vol. 142, nº 1, 2001, p. 5-36.

[48] WILLIAMS Graham, Data Mining with Rattle and R, op. cit., p. 3.

[49] ABITEBOUL Serge et Gilles DOWEK, *Le temps des algorithmes*, *op. cit.*

[50] SAPORTA Gilbert, *Probabilités, analyse des données et statistique*, Paris, Technip, 2006, p. xxxi.

[51] MACÉ Yannick, « L'approche statistique : entre réalité(s) et subjectivité », *Journal de la société française de statistique*, vol. 147, nº 4, 2006, p. 85-102. 2016, p. 18

[52] DAGIRAL Éric et Sylvain PARASIE, « La «science des données» à la conquête des mondes sociaux », *op. cit.*, p. 87 et 101.

[53] Contexte de « crise » professionnelle des statistiques perceptible dès la fin des années 70, voir Desrosières (1993).

problèmes pratiques dans diverses entreprises et organisations confrontées à d'immenses quantités de données souvent peu structurées[54].

Selon Breiman, le traitement des données peut être regroupé en deux « cultures ». La première, « traditionnelle » en statistique, représentait à ses yeux la « modélisation des données » (*data modeling*), « générative »[55], basée sur la théorie classique de l'inférence mise au point dans l'entre-deux-guerres avec Ronald A. Fisher, Jerzy Neyman et Egon S. Pearson (*cf.* tests d'hypothèses et estimations). La deuxième pratique d'analyse dite « algorithmique », à laquelle Breiman adhérait**,** était quant à elle fondée sur la théorie statistique de l'« apprentissage » (*statstical learning*) rendue célèbre à la fin des années 1990 avec l'ouvrage fondateur de Vladimir Vapnik[56].

Suivant le cadre de l'apprentissage, Breiman soutenait que l'un des moyens d'obtenir de l'information « fiable » consiste à optimiser l'efficacité « prédictive » des modèles (*prédictive accuracy*) dans un versant « pragmatique » de la recherche propre à la pensée « computationnelle »[57]. Comme en témoigne l'aphorisme attribué au célèbre statisticien britannique George E. P. Box (1976) « tous les modèles sont faux, mais certains sont utiles », l'enjeu de la culture algorithmique[58] renvoie donc moins à l'exactitude des connaissances inférées en termes d'objectifs descriptifs et analytiques pour établir des faits scientifiques, qui seraient au final vaine et illusoire, qu'à leur pertinence pour l'action et pour la décision.

Dans son texte polémique, Breiman poursuivait en indiquant que la valorisation de ce critère de validation, fondé sur la capacité à prédire le réel (connu dans l'actuel), entre en conflit avec l'interprétabilité des modèles, chère à la culture traditionnelle[59]: à mesure que s'accroît les performances des modèles, ceux-ci se complexifient jusqu'à perdent en intelligibilité, d'où les controverses initialement scientifiques autour de l'opacité des modèles générés par certains outils

---

[54] BREIMAN Leo, « Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author) », *Statistical Science*, vol. 16, nº 3, 2001, p. 199-231. Voir aussi FRIEDMAN Jerome H., « Data Mining and Statistics : What's the Connection ? », 1997 ; HAND David J., « Data Mining: Statistics and More? », *The American Statistician*, vol. 52, nº 2, 1 mai 1998, p. 112-118.

[55] DONOHO David, « 50 Years of Data Science », *op. cit.*

[56] *Cf.* VAPNIK Vladimir N., *The Nature of Statistical Learning Theory*, New York, Springer-Verlag, 1995 ; VAPNIK Vladimir N., *Statistical Learning Theory*, New York, Wiley, 1998.

[57] WING Jeannette M., « Computational Thinking », *Communications of the ACM*, vol. 49, nº 3, mars 2006, p. 33-35 ; ABITEBOUL Serge et Gilles DOWEK, *Le temps des algorithmes*, *op. cit.*

[58] DONOHO David, « 50 Years of Data Science », *op. cit.*

[59] BREIMAN Leo, « Statistical Modeling », *op. cit.*, p. 206.

informatiques. Par ce « conflit de valeurs », Breiman faisait ainsi état d'un enjeu relativement nouveau ouvert par les pratiques de modélisation contemporaine, renvoyant à la possibilité de modéliser, d'anticiper et de prévoir des phénomènes sociaux ou des comportements humains sans nécessairement « comprendre » ou pouvoir rendre compte des processus sous-jacents[60].

Dans la configuration sociale actuelle de la division du travail intellectuel où la représentation des algorithmes comme « boîte noire » tend à devenir un phénomène de société, il apparaît utile de mentionner ici que ni l'enseignement de l'informatique, ni l'accès au code ou même aux données utilisées ne serait susceptible de rendre plus intelligible le type d'opacité auquel se réfère Breiman.

Dans la typologie proposée par la chercheure Jenna Burrell (2016)[61], ce rapport aux algorithmes ne constituerait en fait qu'une seule forme d'opacité, souvent confondue avec d'autres conceptions de la transparence dans les débats publics. Cette dernière mettait de l'avant que l'audit des systèmes automatisés recourant à des algorithmes d'apprentissage se confronte à deux autres problèmes importants. Le premier étant le « secret » professionnel/commercial ou l'aspect confidentiel lié au code lui-même[62] et le second étant les « compétences » nécessaires pour « lire » et interpréter adéquatement le code informatique, renvoyant donc à un enjeu de formation spécialisé[63]. Si chaque facteur d'opacité appelle à réclamer selon Burrell, certains modes de régulation, nous verrons que les écrits s'inscrivant dans la « politique des algorithmes » qui se sont profilés à peu près à la même période que l'article de Breiman, semblent principalement évoquer des enjeux qui découlent de ces deux autres sources d'opacité.

---

[60] ABITEBOUL Serge et Gilles DOWEK, *Le temps des algorithmes*, *op. cit.*, p. 44-45.

[61] BURRELL Jenna, « How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms », *Big Data & Society*, vol. 3, nº 1, 2016, p. 1-12.

[62] Par exemple PASQUALE Frank, *The Black Box Society: The Secret Algorithms that Control Money and Information*, Cambridge: Harvard University Press, 2015.

[63] Par exemple CARDON Dominique, À quoi rêvent les algorithmes, op. cit.

1.2 La politique des algorithmes

Suivant la mise en garde du professeur en droit à Harvard Lawrence Lessig (2000) alertant contre le danger pour nos libertés et nos droits fondamentaux lorsque « le code fait loi », une foulée de recherches multidisciplinaires s'est penchée sur les impacts sociaux, éthiques et politiques d'une « vie algorithmique » généralisée[64]. Partant, semble-t-il, des *Surveillance Studies*, jusqu'aux *Études des médias et du numérique*, ces critiques ont mis de l'avant les aspects décisionnel et prédictif du « pouvoir des algorithmes », en insistant sur l'importance d'étudier les implications, les « ramifications politiques » et les projets de société qu'ils véhiculent. Dans cette lignée, les algorithmes apparaissent eux-mêmes « pertinents publiquement » pour les informations (ou le contenu) qu'ils rendent (in)visible dans nos fils d'actualité[65], les comportements qu'ils orientent parfois à notre insu et les décisions politiques qu'ils tendent à autoriser au nom de l'uniformité de leurs procédures[66].

1.2.1 Gouvernementalité algorithmique

Pionniers de cette vague de travaux qui s'est élaborée autour de l'idée d'une « gouvernementalité algorithmique », s'apparentant à certains égards à la « gouvernance par les nombres » d'Alain Supiot (2015), les philosophes et spécialistes en droit Antoinette Rouvroy et Thomas Berns[67] ont contesté la performativité des outils de DM, et plus précisément les applications contemporaines de la statistique dite « décisionnelle ».

Dans une perspective de sensibilité foucaldienne, deleuzienne et simondienne, les auteurs soutiennent que l'extension de la traçabilité des sphères de l'existence s'accompagne d'une rationalité « apolitique » qui prétend capter, prédire et réguler le monde réel, *à même le réel*, en temps *réel,* de manière systématique et immédiate. Selon ces philosophes, le nouveau régime

---

[64] SADIN Éric, La vie algorithmique, op. cit.

[65] GILLESPIE Tarleton, « The Relevance of Algorithms », dans Tarleton GILLESPIE, Pablo J. BOCZKOWSKI et Kirsten A. FOOT (dir.), *Media Technologies: Essays on Communication, Materiality, and Society*, Cambridge: The MIT Press, 2014, p. 169.

[66] BEER David, « The Social Power of Algorithms », *Information Communication & Society*, vol. 20, nº 1, 2017, p. 1-13 ; CARDON Dominique, *À quoi rêvent les algorithmes*, *op. cit.* ; KITCHIN Rob, « Thinking Critically about and Researching algorithms », *Information, Communication & Society*, vol. 20, nº 1, 2017, p. 14-29.

[67] ROUVROY Antoinette et Thomas BERNS, « Gouvernementalité algorithmique et perspectives d'émancipation », *op. cit.*

d'anticipation institué par les dispositifs « autonomes » est « préemptif »[68] en ciblant la dimension *potentielle* des conduites humaines[69]. À la différence des pratiques statistiques traditionnelles que les auteurs ramènent au probabilisme fréquentiste de l'astronome belge Adolphe Quetelet (1796-1874) et de sa théorie de l'« Homme moyen », les processus de recherche « corrélationnelles » (DM) seraient au service d'un profilage comportemental qui tire profit de la singularité des cas « marginaux » ou « déviants » vus sous la courbe « normale » de Laplace-Gauss, en évitant d'interpeller la subjectivité et la réflexivité des sujets tout au long de son déploiement par des interventions subtiles et directes dans nos environnements personnels.

Pour Rouvroy et Berns, ce serait donc en classifiant les individus sans les rapporter à aucune norme ou catégorie générale (*p. ex*. le sexe/genre, l'origine ethnique…) qui pouvaient être débattues collectivement que les approches de DM seraient en rupture avec le fondement « conventionnaliste » du gouvernement statistique appréhendé par Desrosières. Comme bien d'autres juristes de formation, l'enjeu principal chez Rouvroy et Berns renvoie à la place des délibérations démocratiques dans « l'idéologie technique des *big data* »[70].

Dans cet élan de dénonciations, des chercheurs en informatique semblent aussi s'être donné la mission d'informer publiquement des choix politiques au « temps des algorithmes »[71]. Cathy O'Neil par exemple entend joindre sa voix à titre d'ancienne *data scientist* en finance à Wall Street pour déconstruire l'aura de neutralité dont semble hériter le formalisme des modèles mathématiques prédictifs, dissimulant et intensifiant les discriminations et inégalités sociales en son nom[72]. Pour cette mathématicienne et militante américaine, et bien d'autres expert-e-s en matière d'IA qui se sont mobilisé-e-s autour de thèmes comme celui « loyauté des décisions algorithmiques »[73], l'un des problèmes majeurs renvoie aux « données biaisées », ou aux préjugés

---

[68] ROUVROY Antoinette et Thomas BERNS, « Le nouveau pouvoir statistique », *Multitudes*, n° 40, nº 1, 2010, p. 159 et 172.

[69] Voir Pariser (2012) pour l'idée de « bulle de filtre ».

[70] *Cf*. FAYOLLE Jacky, « "La gouvernance par les nombres" est-elle la fin de l'histoire de la statistique ? », dans le cadre de la Journée européenne de la statistique, Luxembourg, 20 mai 2016.

[71] ABITEBOUL Serge et Gilles DOWEK, *Le temps des algorithmes*, *op. cit.*

[72] O'NEIL Cathy, Weapons of Math Destruction, op. cit.

[73] BESSE Philippe, Céline CASTETS-RENARD, Aurélien GARIVIER et Jean-Michel LOUBES, « L'IA du quotidien peut-elle être éthique ? », *Statistique et Société*, vol. 6, nº 3, 2018, p. 9-31.

« incrustés » dans le code, qui nourrissent des systèmes « apprenants », à leur tour « biaisés », en générant des boucles récursives vicieuses et néfastes dans la vie des gens.

En dressant un bilan de l'état des discussions depuis les dernières décennies sur le sujet, Malte Ziewitz (2016)[74] montre cependant l'ironie de ce qu'elle nomme le (double) « drame » de la montée en puissance des algorithmes, qui en viennent à incarner une figure mythique du monde contemporain, car toujours plus dominants et insaisissables à la fois[75]. En s'érigeant d'abord en véritables entités puissantes, omniprésentes et auto-apprenantes, les algorithmes posent ensuite beaucoup de soucis au plan méthodologique pour quiconque souhaite rendre compte plus finement de la façon dont ils exercent une emprise dans nos opportunités d'existence (d'emploi, de logement, etc.); l'obscurité des calculs étant bien souvent interprétée comme « un autre symbole » de leur autorité[76].

Rejoignant le constat que faisait Ziewitz à propos de la situation devenue ironique et doublement dramatique, d'autres spécialistes vont également remarquer que la plupart des chercheurs en SHS ont préféré théoriser ou spéculer sur les conséquences plus ou moins inéluctables de la place grandissante des algorithmes alors que leurs « logiques techniques » et mathématiques repoussent et demeurent nébuleuses[77].

1.2.2 Logiques du calcul algorithmique

Pour sortir de l'impasse du mythe algorithmique (Jaton, 2019), le sociologue Dominique Cardon a pour sa part défendu l'intérêt en SHS d'entrer dans les composantes, les rouages et le fonctionnement du calcul des algorithmes du web[78]. Pour lui, une critique « internaliste » de la

---

[74] ZIEWITZ Malte, « Governing Algorithms: Myth, Mess, and Methods », *Science, Technology, & Human Values*, vol. 41, nº 1, 2016, p. 3-16.

[75] *Ibid.*, p. 5-6.

[76] Voir les stratégies de rétro-ingénierie. DIAKOPOULOS Nicholas, « Algorithmic Accountability Reporting: On the Investigation of Black Boxes », *Digital Journalism*, 2014.

[77] BURRELL Jenna, « How the machine 'thinks' », *op. cit.* ; MACKENZIE Adrian, « The Production of Prediction: What Does Machine Learning Want? », *European Journal of Cultural Studies*, vol. 18, nº 4-5, 2015, p. 429-445 ; RIEDER Bernhard, « Scrutinizing an Algorithmic Technique: the Bayes Classifier as Interested Reading of Reality », *Information, Communication & Society*, vol. 20, nº 1, 2017, p. 100-117.

[78] CARDON Dominique, À quoi rêvent les algorithmes, op. cit., p. 13.

raison calculatoire permet de mieux saisir la manière dont s'articulent les logiques sociales et techniques qui animent, configurent et font exister les formes plurielles du pouvoir algorithmique.

Proposant une typologie de quatre « familles de calcul numérique », Cardon caractérise la spécificité des mesures en « machine learning » (ML), non seulement par la valeur centrale accordée à la prédiction, mais aussi par le fait que les personnes sont saisies « par le bas », à partir de leurs traces d'activité et de comportements, dont celles de navigation. D'après ce sociologue du numérique, ce type de données disparates, morcelées et décontextualisées auraient pour effet de nous éloigner les uns des autres, plutôt que d'être rassemblés « par le haut » au moyen de catégories d'appartenance, traditionnelles en statistique[79].

Si les nouvelles manières de quantifier le social – de personnaliser le calcul des prédictions – soulèvent des enjeux relatifs entres autres à l'affaiblissement de la solidarité collective (*p. ex.* au travers des systèmes assurantiels), Cardon explique que celles-ci résonnent néanmoins avec les processus historiques d'individuation des rapports sociaux, tel qu'ils s'expriment à travers la montée des revendications identitaires depuis les années 1980[80]. Dans sa thèse, le contexte spécifique du web dans lequel sont étudiés les usages des méthodes statistiques d'apprentissage conduit ainsi l'auteur à assimiler les formes de calculs relevant du ML aux traces et signaux numériques. Tel que le prétendaient Rouvroy et Berns et certains partisans des *big data*, Cardon conclut dès lors que les modes opératoires computationnels procèdent en s'affranchissant de toute catégorie analytique prédéterminée, englobante et totalisante.

Pour clore cette section, il est intéressant de constater que de nombreuses initiatives issues de la sociologie du numérique[81] ou de spécialistes en *études des médias et en communication*, ayant véhiculé la conception selon laquelle les algorithmes sont « impénétrables », « indiscutables », « hors de notre portée et conçus pour l'être »[82], se soient penché sur les algorithmes des plateformes et du web qui tiennent d'abord leur opacité du fait qu'ils sont contrôlés par des compagnies

---

[79] *Ibid.*, p. 16 et 39-42.

[80] *Ibid.*, p. 39-42.

[81] BOULLIER Dominique, « Vie et mort des sciences sociales avec le big data », *Socio*, , nº 4, 2015, p. 19-37.

[82] GILLESPIE Tarleton, « The Relevance of Algorithms », *op. cit.*, p. 169 et 194 ; CARDON Dominique, *À quoi rêvent les algorithmes*, *op. cit.*, p. 8.

privées[83]. Or, comme l'avançait Bernard Rieder[84], les fameux géants du numérique GAFAM[85] sont aussi des entreprises de statistiques bien ordinaires.

[83] BURRELL Jenna, « How the machine 'thinks' », *op. cit.*

[84] RIEDER Bernhard, « Scrutinizing an Algorithmic Technique », *op. cit.*

[85] Un acronyme désignant Google, Apple, Facebook, Amazon et Microsoft

1.3 La fabrique des algorithmes

Confrontées aux difficultés méthodologiques et épistémologiques pour *saisir* des entités supposées *insaisissables*, deux voies plus récentes en SHS semblent avoir cherché avant tout à mieux décrire les algorithmes statistiques et informatiques du point de vue des êtres sociaux « incarnés » qui les élaborent, les mettent en œuvre et en font appel (ou non) dans leurs pratiques.

1.3.1 Socialisation algorithmique

Dans une première perspective plutôt « culturelle », des chercheurs ont ainsi entrepris de retracer, en détail et *in situ*, les réseaux d'acteurs et d'agents engagés aux alentours des logiques de calcul. Étudiés *dans la pratique* effective*,* les algorithmes sont conçus comme des « artefacts » « enchâssés » dans des « assemblages sociotechniques » et des « collectifs humains » plus vastes[86]. Pour l'anthropologue Nick Seaver[87], il s'agirait moins de s'acharner à décoder ou à décrypter les calculs qu'exécutent les algorithmes en SHS qu'à interroger *concrètement* le sens que donnent les acteurs dans leurs activités. Dans cette lignée, reconnu comme l'un des principaux représentants de l'approche culturelle des *Software Studies*[88], Adrian Mackenzie (2015) insiste sur le fait que « la production de prédiction n'est pas automatique (ou autonome), bien qu'elle soit en cours d'automatisation »[89], d'où l'importance selon lui d'examiner les processus par lesquels des prédictions algorithmiques sont produites.

Dans une enquête ethnographique à visée essentiellement méthodologique[90], le sociologue Sébastien Vayre propose par exemple de décrire ce travail de conception comme un processus de

---

[86] MACKENZIE Adrian, « The Production of Prediction: What Does Machine Learning Want? », *op. cit.* ; MÉADEL Cécile et Guillaume SIRE, « Les sciences sociales orientées programmes », *Réseaux*, n° 206, nº 6, 2017, p. 9-34 ; DAGIRAL Éric et Sylvain PARASIE, « La «science des données» à la conquête des mondes sociaux », *op. cit.*

[87] SEAVER Nick, « Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems », *Big Data & Society*, vol. 4, nº 2, 2017, p. 1-12.

[88] MÉADEL Cécile et Guillaume SIRE, « Les sciences sociales orientées programmes », *op. cit.*

[89] MACKENZIE Adrian, « The Production of Prediction: What Does Machine Learning Want? », *op. cit.*, p. 444.

[90] VAYRE Jean-Sébastien, « Comment décrire les technologies d'apprentissage artificiel ? », *Réseaux*, n° 211, nº 5, 2018, p. 71.

socialisation des environnements « d'apprentissage, de traitement et politique »[91]. Selon lui, les algorithmes « prédictifs » constituent des technologies d'apprentissage « artificiel » dont « la liberté de calcul » est nécessairement le résultat d'inscriptions culturelles et matérielles[92].

À l'appui d'extraits d'entretiens et de ses observations, Vayre prend des distances avec l'hypothèse de Rouvroy et Berns (2013)[93] quant à l'absence de tout aspect conventionnel et « normatif » dans la production algorithmique du savoir statistique et probabiliste en montrant que les concepteurs et les implémenteurs négocient constamment les paramètres[94] des inférences réalisées par les systèmes en fonction des besoins et des attentes spécifiques de chaque organisation. En revanche, Vayre estime que les machines statistiques contemporaines se distinguent de celles traditionnelles en établissant un « pluralisme prédictif » qui offrirait des positions analytiques moins « structuralistes » et plus « interactionnistes »[95] :

> n'ayant aucun a priori sur l'importance qui doit être accordée aux données sociodémographiques, comportementales et environnementales, les *big data* permettent [aux systèmes] d'apprendre à prédire [...] en systématisant la mise à l'épreuve des perspectives qui sont associées à chaque catégorie de données[96]

Son étude de terrain dans des entreprises concrètes permet dès lors d'interroger l'image de l'individu social comme être émietté, voire pixellisé, que véhicule une version radicale du « comportementalisme numérique »[97]. Pour Vayre, la variété des types de données numériques constitue autant de points de vue que modulent les technologies pour induire des modèles prédictifs et anticiper des avenirs.

---

[91] *Ibid.*, p. 96.

[92] *Ibid.*, p. 98.

[93] ROUVROY Antoinette et Thomas BERNS, « Gouvernementalité algorithmique et perspectives d'émancipation », *op. cit.*

[94] *Cf.* Le problème de l'explicabilité (section 1.1)

[95] Vayre reprend à son compte l'idée de « pluralisme explicatif » tel qu'entendu par l'épistémologue J.-M.Berthelot (1990) et parle d'une « variation automatique des échelles d'analyse » en se référant à M. Grossetti (2006) (Vayre, 2018 : 77).

[96] VAYRE Jean-Sébastien, « Comment décrire les technologies d'apprentissage artificiel ? », *op. cit.*, p. 77.

[97] ROUVROY Antoinette et Thomas BERNS, « Gouvernementalité algorithmique et perspectives d'émancipation », *op. cit.* ; CARDON Dominique, *À quoi rêvent les algorithmes*, *op. cit.* ; Francis JAURÉGUIBERRY, « L'individu hypermoderne face aux big data », *Sociologie et sociétés*, vol. 49, nº 2, 2017, p. 33-58.

1.3.2 Méthodes algorithmiques

Dans une autre voie, celle de l'étude des algorithmes *en pratique*, certains spécialistes des sciences sociales plus familiers avec la méthodologie statistique vont entreprendre d'introduire les « nouvelles » techniques d'analyse dans leur domaine de recherche respectif, en les inscrivant dans le paysage des méthodes déjà existantes dans le milieu de la recherche sociale « quantitative ».

En distinguant par exemple la forme « supervisée » de l'apprentissage à partir des données, de celle « non supervisée », l'un des principaux intérêts de ces contributions est d'avoir mis en lumière la profonde ambivalence des frontières séparant les outils traditionnels issus de la statistique mathématique de ceux qui relèveraient à proprement parler de la « culture algorithmique » « prédictive »[98]. En effet, dans la classe dite « supervisée » (ou « dirigée »), qui serait la plus utilisée actuellement, des approches « paramétriques »[99] aussi usuelles que les régressions linéaires et logistiques figurent régulièrement parmi les premières méthodes enseignées en ML, à côté de techniques plus récentes, telles que les arbres de décisions, les forêts aléatoires et les machines à vecteurs supports par exemple. De la même façon, les analyses des correspondances multiples (ACM) et l'analyse en composantes principales (ACP), bien connues dans la sociologie française[100], se retrouvent parmi les méthodes d'apprentissage dites « non supervisées ». En sociologie, ces deux classes de l'apprentissage ont donc bien souvent été rattachées à deux traditions intellectuelles, nationales et philosophiques distinctes.

Dans une étude comparant les modèles de la statistique classique avec des modèles d'apprentissage supervisé qui a relativisé la portée révolutionnaire du ML en SHS, les chercheurs ont bien montré les implications concrètes qui découlent du rôle décisif joué par la théorie mathématique[101]. Très brièvement, 1) l'usage des méthodes statistiques en ML incite (ou contraint) à un ensemble de pratiques de recherche : 2) celle de séparer de manière aléatoire le corpus de données en au moins deux groupes, d'« entraîner » plusieurs modèles sur l'un des sous-ensembles,

---

[98] BREIMAN Leo, « Statistical Modeling », *op. cit.*

[99] Des méthodes qui « reposent sur une formulation explicite du modèle et sont interprétées à travers des tests de signification statistique. Elles sont paramétriques en ce sens que leur objectif principal est d'identifier la valeur de certains paramètres, censés jouer un rôle clé dans le phénomène à l'étude. » (traduction libre) (Boelaert et Ollion, 2018 : 477).

[100] *Cf.* Benzécri, Brigitte Cordier-Escoffier, Ludovic Lebart, Michel Volle, Henri Rouanet…

[101] BOELAERT Julien et Étienne OLLION, « The Great Regression Machine Learning, Econometrics, and the Future of Quantitative Social Sciences », *Revue française de sociologie*, vol. 59, nº 3, 2018, p. 481-485.

dit d'apprentissage, en faisant varier des paramètres et des hyperparamètres et 3) d'évaluer la qualité de leurs prédictions sur un jeu de données n'ayant pas servi à les ajuster[102]. Les divers procédés de « validation croisée » et techniques de « rééchantillonnage » visent à sélectionner un modèle prédictif « optimal » en cherchant à éviter à la fois un « sous- » et un « sur- » ajustement aux données d'entraînement. Cela signifie plus concrètement qu'en contraste à la démarche classique visant à minimiser les « erreurs » des modèles en utilisant l'ensemble de l'échantillon initial, la généralisation inductive des savoirs statistiques dans la modélisation « algorithmique » implique de prendre en compte, de façon simultanée, les erreurs de prédiction et la variabilité des représentations[103]. Du point de vue 4) de l'interprétation des résultats, ce processus de construction de modèles (DM), exploratoire et itératif, est ainsi révélateur d'un ensemble de nouvelles normes encadrant les façons légitimes de décrire le monde social; lesquelles reposent par exemple, de moins en moins sur la valeur-p, considérée comme l'un des symboles emblématiques de la statistique mathématique/inférentielle « classique », mais aujourd'hui « soupçonnée d'être au cœur de la crise de la reproductibilité » en sciences[104].

Bien souvent, en exposant méthodiquement certains de ces principes de méthodes/« théoriques » en ML, les enquêtes de type statistique aboutissent à la conclusion suivante : les algorithmes d'apprentissage s'ajoutent aux outils d'analyse déjà disponibles dans la recherche en SHS plutôt que de s'y substituer, de telle sorte que pour d'autres chercheurs, l'un des enjeux consiste à réfléchir aux différents contextes d'usage dans lesquels ces méthodes peuvent (ou pourraient) être utiles, pertinentes ou appropriées. C'est dans cet esprit de la « sociologie quantitative » que Molina et Garip estiment par exemple que la prédiction algorithmique qui fait tant jaser ces dernières années ne constituerait qu'une seule modalité de ces outils, susceptible d'être intégrée à d'autres finalités de recherche plus large (décrire, comprendre et expliquer)[105].

---

[102] Parfois trois groupes, apprentissage-évaluation-validation.

[103] LACOURSE Éric, Charles-Édouard GIGUÈRE et Véronique DUPÉRÉ, « Algorithmes d'apprentissage et modèles statistiques: Un exemple de régression logistique régularisée et de validation croisée pour prédire le décrochage scolaire », dans Marc CORBIÈRE et Nadine LARIVIÈRE (dir.), *Méthodes qualitatives, quantitatives et mixtes dans la recherche en sciences humaines, sociales et de la santé*, 2e éd., Québec, PUC, 2020, p. 581-612.

[104] *Cf.* Tests de signification statistique et estimations. CAPEL Roland, Denis MONOD et Jean-Pierre MÜLLER, « De l'usage perverti des tests inférentiels en sciences humaines », *Genèses. Sciences sociales et histoire*, vol. 26, nº 1, 1997, p. 123-142 ; WASSERSTEIN Ronald L. et Nicole A. LAZAR, « The ASA Statement on p-Values: Context, Process, and Purpose », *The American Statistician*, vol. 70, nº 2, 2016, p. 129-133.

[105] MOLINA Mario et Filiz GARIP, « Machine Learning for Sociology », *Annual Review of Sociology*, vol. 45, nº 1, 2019, p. 40.

1.4 Problème de recherche

En résumé, les recherches jusqu'à présent semblent avoir caractérisé les formes mathématiques contemporaines par rapport à celles plus traditionnelles par les éléments suivants :

- 1) une attitude de recherche tournée vers l'exploration et la découverte de modèles;
- 2) une démarche appliquée axée sur la capacité « prédictive »;
- 3) des techniques d'analyse fondées sur le concept de corrélation ;
- 4) des logiques de personnalisation du calcul ;
- 5) l'ouverture vers des perspectives heuristiques plus « interactionnistes » et;
- 6) des approches de modélisation plus « flexibles », non-paramétriques.

À travers ces divers éléments bien évidemment interreliés, le recours aux algorithmes dans la théorie statistique de l'apprentissage semble introduire une nouvelle façon de produire, de cumuler et de généraliser des connaissances sur le monde social, avec pour enjeu une forme particulière d'opacité qui résulte d'un compromis entre la valeur prédictive et explicative des modèles.

À mon sens, les différentes contributions recensées dans le présent bilan des écrits peuvent être regroupées schématiquement dans quatre approches idéales typiques (au sens wébérien), pouvant s'articuler les unes avec les autres du point de vue de leurs tenants et aboutissants respectifs.

1.4.1 Typologie des postures de recherche

Tableau 1. – Quatre approches idéales-typiques des algorithmes

↑ Considérations techniques / Propriétés théoriques (fondamentales) ↑		+ Quali : Objets (*Externe*) →	↑ Considérations pratiques / Propriétés empiriques (appliquées) ↑
	A) PHILOSOPHIQUE Les algorithmes (« apprentissage automatique ») comme régime de « gouvernementalité », *À quoi rêvent les algorithmes ?*		**C) ETHNOGRAPHIQUE** Les algorithmes (« apprentissage artificiel ») comme réseaux « concrets », « découvertes » *Comment « rendre visible » les algorithmes ?*
	Mots clés : suprématie, règne, tyrannie, soumission, surveillance, technocratie, « vie algorithmique », « société automatique », idéologie, démocratie, domination, mécanisme, « flux machinique », réseaux, performativité (Rouvroy et Berns, 2013; Gillepsie, 2014; Cardon, 2015; Sadin, 2015)		Mots clés : « cultures ou systèmes algorithmiques », terrain, artefacts, intérêts, usagers, négociations, infrastructures, assemblages, acteurs, automatisation, ethnométhodologie, plasticité, agentivité (Seyfert et Roberge, 2016; Seaver, 2017 ; Dagiral et Parasie, 2018; Vayres, 2018; Jaton, 2019).
	Registre : Droit et Société (*Normative* ?) **Limite** : Théoricisme (La place de la théorie ?) **Type d'enjeux** : Médiation (Dématérialisation ?)		**Registre** : Culture et Société (*Restitutive* ?) **Limite** : Empiricisme (Ses rapports à l'expérience ?) **Types d'enjeux** : Situation (Incarnation ?)
	← Enjeux « sociaux », éthiques et politiques *… Expliquer sans décrire ?*		**Enjeux épistémologiques et méthodologiques →** *… Décrire sans expliquer ?*
	B) PROFESSIONNELLE Les algorithmes (« apprentissage machine ») comme expertise, savoir(-faire) « savant » *À quoi servent les algorithmes ?*	↓ + Quanti : Méthodes (*Interne*)	**D) STATISTIQUE** Les algorithmes (« apprentissage statistique ») comme outils, « méthodes » en SHS *Comment utiliser les algorithmes ?*
	Mots clés : responsabilité, biais, données personnelles, discrimination, risque, loyauté algorithmique, décisions transparence, espace public, populations, inégalités, justice, contrôle, formalisme, programmes, calculs, modèles mathématiques, fonctionnalité (O'Neil, 2016; Abiteboul et Dowek, 2017; Berry, 2017)		Mots clés : outils, savoirs, recherches appliquées, avancement de la connaissance, méthodologie, modélisation, statistique, compromis biais/variance, utilisateurs, monde académique, applicabilité, utilité (Ollion et Boelart, 2018 ; Molina et Garip, 2019; Lacourse, Giguére et Dupéré, 2020).
	Registre : Informatique et Société (*Informative* ?) **Limite**: Formalisme (Le rôle de la logique ?) **Type d'enjeux:** Spécialisation (Légitimité ?)		**Registre**: Connaissance et Société (*Illustrative* ?) **Limite**: Méthodologisme (Ses applications pratiques ?) **Types d'enjeux**: Appropriation (Tradition ?)

Dans le continuum de la colonne de gauche, plusieurs recherches gravitant dans l'orbite de la *politique des algorithmes* (section 1.2) tendent à soulever des enjeux de légitimité relatifs au statut accordé à différentes formes de connaissance.

(A) D'une part, les théorisations sociologiques du mode de gouvernance algorithmique s'inspirant le plus des lectures *philosophiques* (Deuleuze, Guattari…) ont tendance à acquérir leur caractère transcendantal, à totaliser le sens de l'expérience, sinon à disqualifier le savoir de sens commun : les algorithmes « rêvent » d'un monde face aux individus (ou plutôt les « dividus »)

invités à se « réveiller avant qu'il ne soit trop tard »[106]. De nos jours, cette position apparait cependant difficilement soutenable, ou du moins discutable, si l'on veut bien admettre que le sociologue lui-même, en tant qu'être humain non omniscient, procède lui-même de ce savoir pratique pour s'inscrire dans le monde[107]. En faisant allusion à l'ironie du double « drame algorithmique »[108], sommes-nous réellement, concrètement devenus que « flux » et poussières, toujours plus insaisissables et « invisibles » au même titre que les algorithmes statistiques et informatiques ?

(B) D'autre part, si plusieurs chercheurs en SHS s'entendent sur l'importance d'avoir quelques notions de base en statistique et en informatique, l'unique maîtrise « technique » des logiques de calculs apparaît problématique. Comme le soulignait brillamment l'historien des sciences Yves Gingras lors d'une conférence intitulée *L'intelligence sociologique confrontée à l'intelligence artificielle* en avril 2019, les discours sur les réalités sociales tenus par certains professionnels spécialisés dans un domaine connexe à l'IA peuvent contribuer à discréditer implicitement la valeur du savoir sociologique, le travail de recherche en sciences sociales ainsi que la pertinence des formations universitaires dans ces disciplines[109]. On remarquera d'ailleurs que, dans la postface de l'édition française de son livre destiné au grand public paru en 2018, Cathy O'Neil revendique un rôle « élargi » des spécialistes en *data science* face aux enjeux sociétaux que soulèvent les modèles qu'elle qualifie d'« Armes de Destruction Mathématique » (ADM).

Dans le continuum de la colonne de droite, quel que soit le type d'enquête privilégié, ethnographique ou statistique, l'ancrage « empirique » des études ayant interrogé la *fabrication des algorithmes* (section 1.3) semble avoir permis d'apporter un regard plus nuancé à la critique, plus constructif et moins polémique. En revanche, le fossé qui sépare les chercheurs les plus « qualitativistes » des plus « quantitativistes » apparaît problématique pour articuler *conjointement* les usages sociaux des algorithmes avec leurs fondements mathématiques.

---

[106] CARDON Dominique, À quoi rêvent les algorithmes, op. cit., p. 8.

[107] Voir HOULE Gilles, « Le sens commun comme forme de connaissance », *Sociologie et sociétés*, vol. 19, nº 2, 1987, p. 77-86.

[108] ZIEWITZ Malte, « Governing Algorithms », *op. cit.*

[109] GINGRAS, Yves. « L'intelligence sociologique confrontée à l'"intelligence artificielle" » (Conférence inaugurale), *Ateliers SociologIA.* Montréal : Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), 10 avril 2019.

(C) D'un côté, dans les approches *ethnographiques* « culturelles », les « découvertes » mobilisées par les chercheurs qualitativistes pour plaidoyer en faveur des enquêtes de terrain peuvent parfois être étonnantes aux yeux de leurs homologues et collègues de travail quantitativistes du même département qui ont travaillé depuis toute leur carrière avec des méthodes statistiques – voir par exemple l'usage de notion de « modèle » plutôt que d'algorithme[110].

(D) De l'autre côté, les recherches appliquées de type *statistique* se limitent bien souvent à une conception *méthodologique* des méthodes – qu'elles soient décrites comme complémentaires ou rivales, utiles ou non – qui n'épuise pas le fond des questions (si existant soit-il), en concluant par exemple, que leur mise en concurrence pourrait tout ou plus s'avérer favorable à l'avancement de la « science dans son ensemble »[111]. Il est d'ailleurs intéressant de constater que la plupart des rares chercheurs quantitativistes en SHS ayant porté une attention aux propriétés théoriques des nouveaux algorithmes semblent l'avoir fait sous l'angle de la méthodologie de recherche, que ce soit pour en évaluer les répercussions *effectives* ou les bénéfices *potentiels*[112]. Or, peut-on également envisager les implications « sociales » de ces objets mathématiques « formels » et des choix *a priori* « techniques » que suppose la diffusion de leur usage ?

Plus généralement, est-il possible d'appréhender des questions de société relatives aux logiques de calculs des algorithmes statistiques et informatiques dans une enquête *empiriquement ancrée* en sociologie ? Est-il possible de le faire dans des termes proprement *théoriques* (pour inscrire la sociologie dans l'horizon des sciences) (Cas limite de C et de D), sans pour autant produire des savoirs totalisants de l'expérience vécue (Cas limite de A) ou susceptibles de délégitimer ultérieurement les connaissances produites au nom de la discipline sociologique (Cas limite de B) ?

---

[110] JATON Florian, « « Pardonnez cette platitude » : de l'intérêt des ethnographies de laboratoire pour l'étude des processus algorithmiques », *Zilsel*, vol. 1, nº 5, 2019, p. 315-339.

[111] BOELAERT Julien et Étienne OLLION, « The Great Regression Machine Learning, Econometrics, and the Future of Quantitative Social Sciences », *op. cit.*, p. 478.

[112] Dans une perspective que Desrosières, en tant qu'historien et praticien de la statistique, qualifierait probablement d'épistémologique/« normative » plutôt que de sociologique/« descriptive » ?

Est-il possible de comparer diverses techniques d'analyse autrement que sur la base des performances prédictives, comme le veulent les discours contemporains « conventionnels »[113] ? En prenant appui sur les acquis de la sociologie de la connaissance et de la quantification, *comment* les algorithmes pourraient-ils faire « parler » les données « sociales » du monde empirique, des constructions empiriques du monde social ?

[113] BOELAERT Julien et Étienne OLLION, « The Great Regression Machine Learning, Econometrics, and the Future of Quantitative Social Sciences », *op. cit.* ; MOLINA Mario et Filiz GARIP, « Machine Learning for Sociology », *op. cit.* ; LACOURSE Éric, Charles-Édouard GIGUÈRE et Véronique DUPÉRÉ, « Algorithmes d'apprentissage et modèles statistiques: Un exemple de régression logistique régularisée et de validation croisée pour prédire le décrochage scolaire », *op. cit.*

CHAPITRE II – CADRE THÉORIQUE ET PROBLÉMATIQUE

Sociologie de la connaissance et de la quantification

Ce deuxième chapitre tente d'esquisser les prémisses d'une approche sociologique visant à appréhender théoriquement et empiriquement la constitution *sociale* de la construction mathématique des algorithmes de calcul. Dans un premier temps, nous posons quelques repères d'épistémologie contemporaine formulés par Gilles-Gaston Granger (1920-2016) pour éclairer la dynamique opératoire de la production des analyses « informatisées » ou « formalisées ». Les grandes lignes directrices qui ont été retenues de son œuvre d'épistémologie comparée serviront à élaborer une perspective de sociologie des sciences, dont l'entreprise n'est pas politique (ou morale), et dont le propos n'est ni uniquement technique, mathématique ou historique. Dans un deuxième temps, il s'agit d'exposer succinctement la théorie générale des « styles de raisonnement » scientifiques tel que formulée par le philosophe canadien Ian Hacking (1936-), en nous concentrant sur celui probabiliste statistique. Son approche historico-épistémologique des usages du calcul des probabilités et des statistiques permettra d'inscrire dans un troisième temps, les processus de quantification des sociétés conceptualisés en sociologie par Alain Desrosières (1940-2013), qui invitait depuis bien longtemps à remonter en amont des « boîtes-noirisées ».

Nous pourrons finalement constater que le cadrage philosophique et historique de Ian Hacking se présente comme intermédiaire entre l'optique « rationaliste » de Granger et celle « conventionnaliste » de Desrosières[114], en permettant de considérer dans la longue durée, les processus de « façonnement » réciproque de toute classe d'*objets* (envisageables) et d'*opérations* (réalisables) du monde « social », nécessaire pour vivre ensemble dans le monde. La perspective plus pragmatique de la connaissance chez Desrosières permettra d'établir quelques distinctions conceptuelles fondamentales préliminaires qui guideront notre analyse sociologique comparée.

---

[114] Quatre ouvrages ont été principalement mobilisés : *Le Probable, le Possible et le Virtuel* (Granger, 1995), *Formes, opérations et objets* (Granger, 1994), *La politique des grands nombres. Histoire de la raison statistique* (Desrosières, 1993), et les essais du premier volume de l'Argument statistique, *La sociologie historique de la quantification* (Desrosières, 2008). Étant donné le rôle médiateur que joue Hacking entre Granger et Desrosières, mais tout autant important dans l'élaboration de notre cadre théorique « sociologique », les cours et séminaires qu'il a donné au Collège de France de 2001 à 2006 nous ont semblé plus pertinents à mobiliser que ses nombreux livres magistraux publiés avant les années 2000 (nous pensons ici notamment à *L'émergence de la probabilité*). Non seulement, ils constituent une synthèse commode de son œuvre dans le cadre d'un mémoire de maîtrise, mais plusieurs passages montrent l'auteur nuancer ses propres affirmations du passé.

2.1 L'« épistémologie comparée » des *formes* de connaissance

Dans l'épistémologie comparée des *formes* de connaissance développée par le philosophe français Gilles-Gaston Granger (1920-2016), la catégorie de dualité est définie comme « *la condition de possibilité* […] *du symbolisme* »[115], sur laquelle reposerait tout « acte de connaissance ». Selon lui, ce principe de dualisation symbolique codétermine *simultanément* et *réciproquement* « tout système d'objets possibles et système d'opérations intellectuelles »[116]. En ce sens, les sciences comme formes de connaissance telles que le conçoit Granger « sont nécessairement des *formes d'objets* »[117], susceptibles d'engendrer dans les processus scientifiques des « *rapports* de forme à contenu »[118], qu'il nomme des « contenus formels[119] ». Comme « méta-concept[120] », la dualité « de l'opération et de l'objet » chez Granger fonde ainsi les rapports de connaissance :

> Par l'exercice du principe de dualité, la saisie perceptive d'un phénomène se dédouble en acte de position d'objet et en un système d'opérations implicitement, et peut-être virtuellement, établi, dont l'objet est à la fois le *support* – en tant qu'indéterminé ['d'un système d'opérations, de transformation', 'manipulable au moyen d'opérations formelles dans un espace de représentation'] – et le *produit* – en tant que détermination d'une expérience ['d'une possibilité d'objets', 'plus ou moins directement rattachable aux phénomènes par un système d'opérations matérielles']**[121]**

Concept de « modèle » : « Expliquer » (Représenter)

Dans la perspective de cet épistémologue comparatiste, le *travail* de connaissance se caractérise par ses « objets », entendus au sens de l'élaboration d'une visée. Pour Granger, un mode de connaissance « selon les visées [*objets*] et les méthodes [*opérations*] de la science »[122], « n'embrasse […] pas la totalité du réel »[123], mais procède par réduction, par « construction de

---

[115] GRANGER Gilles-Gaston, *Formes, Opérations, Objets*, Paris, Vrin, 1994, p. 57.

[116] *Ibid.*, p. 55 et 383.

[117] *Ibid.*, p. 383.

[118] *Ibid.*, p. 60.

[119] Les « contenus formels [des *objets*] se manifestent comme produits, corrélatifs de la dualité, dans le développement des concepts indépendamment de tout contenu empirique. » *Ibid.*, p. 53.

[120] Ceux-ci « port[e]nt sur l'organisation de la connaissance et sur son sens, non directement sur son contenu. » *Ibid.*, p. 69.

[121] *Ibid.*, p. 57. p. 383-384

[122] GRANGER Gilles-Gaston, *Le Probable, le Possible et le Virtuel*, Paris, Odile Jacob, 1995, p. 8.

[123] Ibid.

modèles » permettant d'expliquer (au moins provisoirement et partiellement) des phénomènes actuellement observés (ou éventuellement observables) que révèle l'expérience[124]. La notion de modèle désigne ici « une représentation *abstraite* d'un phénomène », qui a pour « résultat […] de rendre saisissable (explicitement) le rapport opératoire d'une forme à un contenu »[125]. Expliquer**,** tel que le conçoit Granger, signifie « insérer [d]es représentations dans des systèmes de virtualités […], et formuler des règles déterminant », avec plus ou moins de précision, des faits *actuels,* « effectivement réalisés »[126].

La « réalité » comme rapport actualité/virtualité

En montrant le rôle crucial que jouent trois figures du *non-actuel*, Granger distingue dès lors la connaissance « scientifique » du monde social de la connaissance « historique ». Plus précisément, selon lui, la « réalité » qu'atteignent les sciences comme la sociologie – qui « s'efforcent délibérément de neutraliser l'individuation de leurs objets concrets »[127] – n'est en aucun cas réductible à des actualités *concrètes*, effectives du monde, mais « comporte des facettes composites de *virtuel* et de *probable* »[128]. D'après Granger, « les objets qui constituent le[ur] réel »[129] « sont fondamentalement des virtualités »[130], des « faits » *virtuels* « doués de propriétés schématisant les faits actuels »[131]. Par *virtualité*, Granger entend « une représentation des choses et des faits, détachée, des conditions […] d'une saisie individuée, singulière, vécue »[132], des variantes empiriquement non actualisées, « éventuellement non actualisables »[133], qui s'élaborent par le raisonnement[134].

---

[124] *Ibid.*, p. 9 et 234.

[125] GRANGER Gilles-Gaston, « Modèles qualitatifs, modèles quantitatifs dans la connaissance scientifique », *Sociologie et sociétés*, vol. 14, nº 1, 1982, p. 7.

[126] GRANGER Gilles-Gaston, Le Probable, le Possible et le Virtuel, op. cit., p. 165, 233-236.

[127] *Ibid.*, p. 124-125. Le caractère *individuel* des événements incommensurables historiques.

[128] *Ibid.*, p. 81, 232-234.

[129] *Ibid.*, p. 234 ; GRANGER Gilles-Gaston, *Formes, Opérations, Objets*, *op. cit.*, p. 386.

[130] GRANGER Gilles-Gaston, Le Probable, le Possible et le Virtuel, op. cit., p. 99.

[131] *Ibid.*, p. 99, 233-234.

[132] *Ibid.*, p. 23.

[133] *Ibid.*, p. 81.

[134] GRANGER Gilles-Gaston, *Formes, Opérations, Objets*, *op. cit.*, p. 174-175.

Dans la conception de cet épistémologue du processus de construction des représentations du « social » à prétention scientifique, c'est « l'usage de la catégorie et du calcul du *probable* »[135] qui articule le virtuel à l'actuel, en introduisant (et supposant) la « grandeur », des degrés du *possible*[136]. Pour Granger, le *probable* « opère une organisation, une structuration du *possible* » – plus exactement, le rapport qu'il ouvre à l'actualité (« historique », *concrète*) – et en constitue la « mesure »[137], mais « en demeurant strictement une théorie mathématique »[138]. Dans son « réalisme modéré », les objets mathématiques, que constitue et qu'explore le calcul des probabilités[139], sont ainsi conçues comme des « formes d'*objets virtuels* » de symbolisation[140], qui n'ont pas de « matière » (sensible), mais des propriétés « formelles »[141].

En termes épistémologiques, les algorithmes de calcul opèrent des « mesures » du *possible* pour élaborer des points de vue sur le réel, en codéterminant le *rapport* entre des « états » du monde actualisés – de la saisie perceptive, observés ou observables empiriquement – et non-actualisés (*virtuels*).

---

[135] GRANGER Gilles-Gaston, Le Probable, le Possible et le Virtuel, op. cit., p. 233.

[136] *Ibid.*, p. 14 et 142-143.

[137] *Ibid.*, p. 130-131.

[138] *Ibid.*, p. 162.

[139] Ibid.

[140] *Ibid.*, p. 81 et 236.

[141] *Ibid.*, p. 82. Pour Granger, « les tentatives d'interprétation du calcul et les controverses non résolues qu'elles soulèvent montrent […] que l'information apportée par les mathématiques ne constitue nullement en soi une information empirique. Si les mathématiques […] n'étaient simplement qu'un langage, un cadre d'exposition arbitraire et vide de tout contenu, son usage ne soulèverait que des problèmes purement techniques. C'est parce que la connaissance mathématique a un contenu, et que ce contenu n'est pourtant pas empirique, que son application à une description et à une explication du monde est problématique, mais qu'elle est également féconde » (Granger, 1994 : 48).

2.2 La « méta-épistémologie historique » des *styles* de raisonnement

« Façonner » : Possibilités et modalités d'expérience ?

Titulaire de la chaire de *Philosophie et histoire des concepts scientifiques* au Collège de France de 2000 à 2006[142], Ian Hacking (1936-) s'est intéressé aux manières de constituer des *objets* intellectuels collectifs[143], aux processus en vertu desquels « des concepts [organisateurs, tels que la probabilité] et [des] pratiques variés, […] découvrent en même temps de nouvelles possibilités pour l'action et le choix humains »[144]. Outre des concepts, Hacking entend par la notion d'« objets » des « *types* d'entités, de personnes, de preuves factuelles, de phrases, de possibilités, de classifications et d'explications » (*sociales*)[145]. Cherchant à développer une « 'méta-épistémologie historique' » qu'il décrit comme un « type d'enquête »[146] portant sur « l'espace de possibilités [et de ses mutations] pour la formation du caractère qui entoure une personne et crée des potentiels pour l'"expérience individuelle" »[147], Hacking affirme que

> Nous sommes orientés vers ce qu'il est possible d'être et de faire […] nous sommes constitués par ce que nous faisons. Mais nos libres choix ne peuvent se faire qu'entre les actions qui nous sont ouvertes, les actions possibles. Et nos manières d'être, qu'elles soient librement choisies ou non, proviennent de manières d'être possibles[148]

Rapport entre méthodes et objets

En s'inspirant des six « styles de pensée scientifique dans la tradition européenne » documentés par l'historien A. C. Crombie (1978)[149], Hacking développe le concept de « style de

---

[142] Voir les *Cours au Collège de France*, en particulier les années 2002-2003 « Des styles de raisonnements scientifiques », 2005-2006 « Raison et véracité – Les choses, les gens, la raison » [En ligne].

[143] HACKING Ian, « L'ontologie historique », dans Laurence KAUFMANN et Jacques GUILHAUMOU (dir.), *L'invention de la société. Nominalisme politique et science sociale au XVIIIe siècle*, Paris, Éditions de l'EHESS, coll. « Raisons Pratiques. Épistémologie, sociologie, théorie sociale », 2003, p. 289.

[144] *Ibid.*, p. 289-293.

[145] HACKING Ian, « Style pour historiens et philosophes », dans Jean-François BRAUNSTEIN (dir.), *L'histoire des sciences. Méthodes, styles et controverses*, Paris, Vrin, 2008, p. 303 et 311.

[146] HACKING Ian, « L'ontologie historique », *op. cit.*, p. 294.

[147] *Ibid.*, p. 304.

[148] *Ibid.*, p. 303.

[149] « 1) la postulation mathématique et la démonstration axiomatique, 2) la mise en œuvre de l'expérimentation à la fois pour contrôler la postulation et pour explorer par l'observation et la mesure, 3) la construction par hypothèse de modèles analogiques, 4) la mise en ordre du divers par la comparaison et la taxinomie, 5) l'analyse statistique des régularités […] et le calcul des probabilités et 6) la dérivation historique propre au développement génétique » (Hacking, 2008 : 292).

raisonnement »[150], insistant sur les procédures et les pratiques de la recherche scientifique[151]. Selon Hacking, chaque style se caractérise par des « méthodes » d'investigation et des « *types* nouveaux d'objets »[152]. En établissant ses propres critères et conditions de validité[153] (ce que c'est que d'être *objectivement* orienté pour le dire en termes grangériens), un style scientifique de raisonnement vient à « s'autojustifier[154] ». Par cette théorie, ce philosophe des sciences entend ainsi « expliquer l'origine (et la persistance) des débats ontologiques »[155] comme forme de controverses que suscitent les objets qu'introduit un style tout au long de sa trajectoire[156].

Statistique et probabilités

À travers ses nombreux travaux sur le style probabiliste et statistique de raisonnement, qu'il caractérise par sa vocation à « domestiquer le hasard »[157], Hacking a bien montré le dualisme conceptuel originaire qu'encapsule l'idée moderne de probabilités – relative à des propriétés empiriques (observables) relativement stables du monde (en termes fréquentistes) ainsi qu'à des degrés variables (contrôlables) de connaissance (en termes épistémiques)[158].

Concernant les nouveautés apparues des applications de l'analyse statistique et du calcul des probabilités au monde « social » dès le milieu des années 1830, Hacking donne les exemples de la moyenne (Quetelet) et du coefficient de corrélation (Galton) pour lesquels Desrosières parle

---

[150] Ce concept permettrait selon Desrosières d'« historiciser des questions d'épistémologie sans pour autant les relativiser » (Desrosières, 2008 : 235) et discréditer la validité des connaissances antérieures : « [p]ar exemple, à partir de 1660 environ, des catégories d'énoncés (probabilistes) deviennent possibles, alors qu'ils n'avaient aucun sens clair auparavant. De même, à partir des années 1820, deviennent possibles des énoncés sur des déterminismes « macro », indépendants de tout déterminisme « micro » sous-jacent. Ou encore, à partir des années 1890, émergent et s'autonomisent de nouvelles façons de formuler une sorte de causalité partielle, exprimée par une régression linéaire dotée d'un « résidu non expliqué », ou par un coefficient de corrélation compris entre – 1 et + 1 : ce genre d'énoncé était inconcevable avant 1890 » (Desrosières, 1993 : 235).

[151] HACKING Ian, « L'ontologie historique », *op. cit.*, p. 288, 301-302.

[152] HACKING Ian, « Style pour historiens et philosophes », *op. cit.*, p. 303-311.

[153] HACKING Ian, « L'ontologie historique », *op. cit.*, p. 304-306.

[154] « deviennent non pas révélateurs de la vérité objective, mais plutôt les standards de l'objectivité » (Hacking, 2008 : 316).

[155] HACKING, Ian, *Raison et véracité. Cours au Collège de France*, *2005-2006*, leçon du 7 février 2006, « Objets » [En ligne], p. 4. Voir aussi les leçons du 4 mai 2006, « Cognition » et du 28 mars 2006, « La stabilité des styles de pensée scientifique ».

[156] Voir HACKING, Ian, « Des styles de raisonnement scientifique » (résumé annuel). *Cours au Collège de France*, *2002-2003*, p. 544-546. En ligne <https://www.college-de-france.fr/media/ian-hacking/UPL4445123752442236773_Hacking2002_2003.pdf>.

[157] HACKING Ian, « Comment faire l'Histoire de la statistique ? », *LINX*, vol. 1, nº 1, 1980, p. 181-191.

[158] Voir aussi Desrosières (1993 : 354).

respectivement d'un « réalisme des agrégats » et d'un « réalisme des causes »[159]. Plus récemment, nous pourrions également penser à l'« opinion publique », comme entité issue de la méthode de sondage que peut expliquer la théorie hackingienne, à la suite de P. Bourdieu qui affirmait pour la première fois en 1972 que « L'opinion publique n'existe pas »[160].

[159] Sont repris ici les titres des chapitres 3 et 4 de la *Politique des grands nombres* (Desrosières, 1993). Concernant le style statistique, les taux de chômage, les indices de prix, « sont-ils de pures fictions [artefacts de l'esprit], « des propriétés réelles des populations, 'de simples résumés d'observations ou […] les produits des structures institutionnelles de classification et de mesure ? » HACKING Ian, « Style pour historiens et philosophes », *op. cit.*, p. 304. Le même constat à propos du « statut de réalité ou de fiction » des constructions/objectivations statistiques se retrouve chez Desrosières.

[160] BOURDIEU, Pierre, « L'opinion publique n'existe pas », dans *Questions de sociologie*, Paris: Éditions de Minuit, 1984, p. 222-235.

2.3 La « sociohistoire » des *usages* de la statistique

Dans *La politique des grands nombres*, Alain Desrosières (1940-2013) définit la « raison statistique » comme un « espace cognitif d'équivalence et de comparabilité construit à des fins pratiques »[161], « lié aux nécessités de l'action et de la décision » dans les activités humaines[162]. Cherchant à comprendre le fait que le terme « statistique » évoque la fois « résultats, descriptions quantifiées » et résumées d'observations et « méthodes, formalisme mathématique et mode de raisonnement »[163], Desrosières retrace la « double origine » politico-administrative et logico-scientifique de la statistique et de ses usages, en s'intéressant à l'entrelacement des trajectoires qu'il qualifie d'internaliste (conceptuelle, *contenus* « techniques » des énoncés : outil de « preuve ») et d'externaliste (institutionnelle, *contextes* « sociaux » de leur insertion : outil de « gouvernement ») de son histoire[164].

De la commensurabilité à la discutabilité

L'une des notions centrales qu'il développe pour caractériser « l'*acte* fondateur » de la démarche statistique (au sens moderne)[165], depuis la théorie de la moyenne d'A. Quetelet, est celle de « mise en équivalence », de « faire tenir ensemble » des choses ou des gens. À l'instar de la conception grangérienne de la saisie/connaissance/visée scientifique du monde empirique, Desrosières indique que le travail statistique « vise à réduire la multiplicité des situations » « à un petit nombre de caractéristiques, qualifiées d'*attributs* de l'objet dans une perspective plutôt

---

[161] DESROSIÈRES Alain, La politique des grands nombres: histoire de la raison statistique, 2e éd., Paris, La Découverte, 2000, p. 399 ; DESROSIÈRES Alain, Pour une sociologie historique de la quantification : L'Argument statistique I, Paris, Presses des Mines, 2008, p. 79. [En particulier le chapitre 5 « Discuter l'indiscutable »]

[162] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 169.

[163] DESROSIÈRES Alain, La politique des grands nombres, op. cit., p. 398-399 ; DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 7-8 et 58-59.

[164] DESROSIÈRES Alain, *Pour une sociologie historique de la quantification*, *op. cit.*, p. 31 ; DESROSIÈRES Alain, *La politique des grands nombres*, *op. cit.*, p. 18-22. Dans la postface, il écrit que « [l']objectif principal […] était d'articuler une histoire de type internaliste des formalismes et des outils, avec celle, plus externaliste, des institutions et des usages sociaux des statistiques » (Desrosières, 2010 : 438). Voir aussi HACKING Ian, « Comment faire l'Histoire de la statistique ? », *op. cit.*

[165] DESROSIÈRES Alain, *La politique des grands nombres*, *op. cit.*, p. 34-35. Selon lui, ce n'est qu'au 20e siècle, « quand sont routinisées et diffusées les techniques de la régression et de la corrélation, à partir […] de Karl Pearson, puis celles de la statistique inférentielle (estimation, tests, analyse de variance) […] de Ronald Fisher », que la statistique « apparaît comme une « branche spécialisée des mathématiques » (Desrosières, 1993 : 22), « utilisée pour induire, tester et généraliser des connaissances à partir de faits observés, dans les sciences de la nature comme dans les sciences sociales. » (Desrosières, 2008 : 7-8).

fréquentiste, ou de *paramètres* d'un modèle dans une perspective plutôt épistémique »[166], en créant des « espaces de communes mesures »[167]. Dans son approche sociohistorique, Desrosières soutient que ce processus repose sur la combinaison de contraintes, de logique, de type *classificatoire* (orientées vers la description, s'exprimant à travers le « langage des groupes ») et *métrologique* (explication, « langage des variables »)[168] (*Cf.* Grille conceptuelle). Selon lui, c'est par ces deux formes de procédures que des objets statistiques viennent à exister, à circuler (1988 :53), à « éprouver leur consistance »[169].

En entendant ouvrir une épistémologie « conventionaliste », Desrosières redéfini la notion de « quantification » en lui conférant un sens proprement sociologique. Comme ce dernier le souligne à maintes reprises dans ses écrits, toute opération arithmétique conduisant à la « mise en nombre » implique un ensemble de « conventions » déterminant ce qu'il est possible *cognitivement* et *socialement* (pensable et convenable) de comparer, d'ordonner et de mesurer[170], d'où la dimension « politique » de cet acte de connaissance. Comme Hacking[171], Desrosières estime ainsi que les techniques statistiques contribuent à stabiliser, à reconfigurer et à transformer l'*ordre* social[172], « en créant [de] nouvelle[s] façon[s] [ou possibilités] de penser, [d'interroger, d'interpréter], de représenter, d'exprimer le monde et d'agir sur lui »[173] dans sa diversité et sa commune humanité.

Par cette conceptualisation, Desrosières relie de façon presque indissociable l'objectivation statistique à la notion d'« espace public ». Selon lui, le fait de quantifier des aspects du réel offre aux acteurs des référents langagiers communs, intersubjectifs, « mémorisables, transmissibles et utilisables comme points d'appui », qui doivent être à la fois *incontestables* et *discutables*[174]. Autrement dit, dans cette perspective, les contraintes opératoires qu'implique tout acte de

---

[166] Ibid., p. 20-22 ; DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 288.

[167] DESROSIÈRES Alain, La politique des grands nombres, op. cit., p. 405.

[168] Voir respectivement les chapitres 8 et 9. DESROSIÈRES Alain, *La politique des grands nombres*, *op. cit.*

[169] *Ibid.*, p. 19-20.

[170] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 10-13.

[171] HACKING Ian, « Comment faire l'Histoire de la statistique ? », *op. cit.*, p. 182 et 189.

[172] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 8-12.

[173] *Ibid.*, p. 288 ; DESROSIÈRES Alain, *La politique des grands nombres*, *op. cit.*, p. 9, 19-22, 79 et 138.

[174] DESROSIÈRES Alain, *La politique des grands nombres*, *op. cit.*, p. 22 et 413 ; DESROSIÈRES Alain, *Pour une sociologie historique de la quantification*, *op. cit.*, p. 100 et 288. [En particulier le chapitre 5 « Discuter de l'indiscutable » et le chapitre 11 « Refléter ou instituer : L'invention des indicateurs statistiques »]

quantification déterminent du même coup les conditions de possibilité de l'idée même d'une délibération collective, en fournissant précisément « réalité » (extériorité, solidité) à des facettes de la vie humaine susceptibles de faire *objet* (enjeu) « public » de débat[175].

Plus généralement, dans le champ de recherche de Desrosières, l'analyse statistique et le calcul des probabilités apparaissent comme un moyen en adéquation avec le contexte des sociétés occidentales démocratiques, permettant de concilier les idéaux de scientificité, d'impartialité et d'équité[176]. En référence aux travaux de l'historien des sciences Ted Porter que cite aussi Desrosières, comme série de règles explicitées, non-ambigües et impersonnelles, un algorithme tend à incarner la figure la plus aboutie du type « mécanique/procédural » d'objectivité, susceptible d'apparaitre comme « une façon de prendre des décisions sans avoir l'air de décider »[177] au nom d'une plus grande « justice » (sociale) et « justesse » (mathématique) à la fois.

Repenser le rapport entre « contexte » (ou *forme*) et « contenu »

Aux yeux de Desrosières, l'un des défis du programme de « la sociologie *historique* de la quantification » consisterait à saisir conjointement les dimensions politico-administratives (institutionnelles) et logico-mathématiques (conceptuelles) de la raison statistique[178], de manière à montrer la cohérence d'ensemble, à certains moments donnés, entre des idées d'une époque sur la société, « des modalités d'action en son sein et des modes de description »[179]. Comme l'écrivait Desrosières

> La question des formes de totalisation et de mise en équivalence est étroitement liée à celle de la gestion d'une société que l'on qualifie parfois « de masse », si l'on veut attirer l'attention sur la standardisation d'un certain type d'interactions entre les personnes. Mais cette façon de dire présente l'inconvénient de polariser le regard sur un des modes possibles de totalisation, avec lequel le travail statistique a, d'une certaine façon, partie

---

[175] *Cf*. Quetelet. DESROSIÈRES Alain, *La politique des grands nombres*, *op. cit.*, p. 401.

[176] « Le débat à propos d'une action suppose l'explicitation de relations entre des objets ou des événements a priori singuliers et incommensurables, dans un cadre de référence permettant de les penser en même temps. Ce débat peut se dérouler entre plusieurs personnes, ou bien, pour une même personne, entre des moments, ou des actions alternatives. La cohérence avec soi-même pose des problèmes de même type que la production d'un cadre d'objectivation commun à plusieurs sujets. […] La notion d'espérance [mathématique], antérieure à celle de probabilité, permet de construire de tels cadres de référence, et de rendre cohérents et commensurables, soit les décisions et les choix d'une personne, soit les arbitrages entre plusieurs personnes ». (Desrosières, 2008 : 83)

[177] PORTER Theodore M., *La confiance dans les chiffres: La recherche de l'objectivité dans la science et dans la vie publique*, Paris, Les Belles Lettres, 2017, p. 11.

[178] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 60-61.

[179] *Ibid.*, p. 68.

liée. En faisant cela, on méconnait le fait que les personnes ont toujours plus ou moins la possibilité […] de se présenter et d'être vues dans des systèmes de grandeur différents. L'assimilation fréquente des totalisations statistiques au processus de massification et d'anonymisation des sociétés bureaucratiques participe de la dénonciation plus générale des effets de la rationalisation de la gestion du monde social […]. [Or] [e]n tant que formes, [les statistiques] ne sont ni plus ni moins oppressives que ne le sont des formes cognitives antérieures (langage, écriture, imprimerie...) ou postérieures (informatique, intelligence artificielle…). Ce sont des moyens parmi d'autres pour affronter l'inquiétude que suscitent les masses sans formes[180].

[180] DESROSIÈRES Alain, « Masses, individus, moyennes: la statistique sociale au XIXe siècle », *op. cit.*, p. 63-65.

2.4 De la « sociologie historique » à la sociologie *générale*, « scientifique » ?

2.4.1 Quelques remarques à propos la problématique fondamentale

L'idée très générale que cherche à défendre ce travail est que l'engouement croissant que suscitent les « nouvelles » approches d'algorithmes ne s'explique pas *uniquement* en termes d'infrastructures, de puissance de calcul et de bases de données « massives » et hétérogènes, mais également par la compatibilité des formes de raisonnement du *style* statistique et probabiliste avec certaines configurations de la société et de la sociologie[181]. Cette manière de problématiser l'utilisation des algorithmes réactualise ainsi, celle des relations entre « formes cognitives et histoire sociale »[182], s'étant elle-même inscrite dans le prolongement de la problématique de l'École durkheimienne sur les fondements « sociaux » et « logiques » des « classements symboliques » en usage dans les groupements humains[183]. Comme l'avançait déjà Desrosières,

> L'hypothèse retenue pour analyser la place de l'information statistique dans l'espace du débat public, est que ce lang[age] prend, dans certains [espaces] et dans certaines périodes, une consistance originale, elle-même liée à la consistance d'une forme de régulation des rapports sociaux. C'est précisément ce langage qui fournit à ces rapports les points de repère et le sens commun […] pour mettre en forme les choses, pour dire les fins et les moyens de l'action, pour en discuter les résultats[184].

Expliquer le social par le social *et* par la sociologie ? Cependant, il importe de noter que notre hypothèse recouvre deux sous-propositions distinctes, quoiqu'intimement liées entre elles. La première, très classique en sociologie depuis sa formulation par Durkheim, est celle de chercher à « expliquer le *social* par le *social* », par la « matière », en supposant que les nouvelles *formes* de quantification « algorithmique » peuvent être éclairés à la lumière des changements dans l'organisation de la société, dans nos manières de « vivre ensemble », de « faire société », la nature du lien social.

---

[181] Plus exactement, d'une sociologie qui ne dépasse pas les idéologies d'une époque (Sabourin).

[182] DESROSIÈRES Alain, « Comment faire des choses qui tiennent : histoire sociale et statistique », *Histoire & Mesure*, vol. 4, nº 3, 1989, p. 225-242 ; THÉVENOT Laurent, « La politique des statistiques : les origines sociales des enquêtes de mobilité sociale », *Annales*, vol. 45, nº 6, 1990, p. 1275-1300.

[183] Voir notamment les commentaires de Desrosières (1993) à propos de l'article d'Émile Durkheim et de Marcel Mauss publié en 1901 intitulé « De quelques formes primitives de classification : contribution à l'étude des représentations collectives ».

[184] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 89-90 ; DESROSIÈRES Alain, La politique des grands nombres, op. cit., p. 406.

En apparence moins banale qu'elle ne pourrait apparaître, la seconde partie de l'énoncé voudrait suggérer d'expliquer la naissance de la DS et ses succès récents par le développement des « moyens » par lesquels on cherche à connaître la « matière » en sociologie, comme discipline empirique à prétention scientifique, reconnue parmi les SHS pour son « assise statistique ». Selon le démographe F. Héran, à qui nous empruntons l'expression, négliger cet ancrage spécifique reviendrait d'ailleurs à remettre en cause la capacité de la sociologie à s'instituer en tant que discipline scientifique autonome. Selon lui, le lien qui unit la statistique à la sociologie est non seulement pratique (utile) mais également théorique (nécessaire)[185].

Comme le précisait Desrosières, c'est le raisonnement probabiliste fréquentiste de Quetelet pour le calcul de moyenne, qui a permis de fonder l'idée selon laquelle « le groupe "social" a des structures spécifiques, des propriétés de régularité et de prévisibilité, dont sont dénués les individus, volatiles [libres] et imprévisibles », « non intrinsèquement déterminés au niveau élémentaire »[186]. Il y a lieu ici de rappeler ici que dans le modèle inventé par Quetelet, connu sous le nom de la « *cause constante* macro-déterministe », la notion de causalité renvoie simplement à la « consistance » (ou à l'existence) d'objets ou de faits *réels,* distincts des constatations individuelles, « du foisonnement sans limites des manifestations sensibles [et contingentes] des cas singuliers »[187]. Selon Desrosières, cette notion aurait donc non seulement permis à la sociologie de justifier sa raison d'exister comme activité de connaissance, mais également marqué ses développements ultérieurs de Durkheim à Bourdieu, en passant par Lazarsfeld[188].

Dans les termes d'épistémologie comparée de Granger, cela suggère ainsi que l'*usage* (et le *sens*) appliqué des mathématiques – de la loi des grands nombres notamment[189] – aux observations/constructions empiriques du monde social aurait permis de constituer l'objet *virtuel* de la sociologie comme étude scientifique des faits sociaux.

---

[185] HÉRAN François, « L'assise statistique de la sociologie », *op. cit.*

[186] DESROSIÈRES Alain, La politique des grands nombres, op. cit., p. 102 ; DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 49, 59, 154-156, 231-232 et 244.

[187] DESROSIÈRES Alain, La politique des grands nombres, op. cit., p. 288-289.

[188] Deux figures de proue généralement reconnue pour être en opposition du point de vue de leur conception de la méthodologie statistique. *Ibid.*, p. 262.

[189] Desrosières décrivait « cette « loi », fondement des formulations probabilistes et des théorèmes de convergence qui en résultent, […] [comme] une sorte d'opérateur de transformation et de passage entre le monde des observations et celui de la généralisation, de l'extrapolation et de la prévision » (Desrosières, 2008 : 211).

2.4.1.1 Grille conceptuelle préliminaire

De nos jours, on l'a vu, la différenciation de familles de méthodes statistiques basée sur des vertus épistémologiques intrinsèques (descriptive/explicative, exploratoire/confirmatoire), encore assimilées à des traditions nationales en sociologie quantitative, se révèle caduque au sein du cadre supervisé de l'apprentissage, où les analyses en vertu desquelles la prédiction devient un « principe d'intervention inédit *dans* et *sur* la société »[190] sont nécessairement fondées sur l'asymétrie des variables. Par conséquent, un ensemble d'oppositions proposé par Desrosières offre un point de départ pertinent pour analyser en termes sémantiques et pragmatiques divers « modes d'analyse » reposant à la fois sur le formalisme asymétrique des modélisations « économétriques » anglo-saxonnes, tout en présentant certaines similarités avec les principes de l'analyse « géométrique » des données « à la française »[191].

Dans ses réflexions à propos des débats de méthodes, Desrosières proposait de différencier analytiquement les traditions d'enquêtes monographiques et statistiques comme types idéaux d'investigation de la vie sociale, recoupant la distinction qu'il opérait entre les approches centrées sur les *variables* et celles centrées sur les *groupes,* et plus exactement, entre deux manières de « lire » un fichier/tableau statistique – soit en colonne, par caractéristique, soit en ligne, par individu – « déjà en germe dans la différence entre les outils, très proches formellement, que constituent la *régression* et la *corrélation* »[192]. À mon sens, plusieurs de ces dichotomies peuvent être vues comme des « dualités » symboliques (interprétatives) au sens de Granger, élémentaires à la saisie de tout « fait social », irréductibles à l'analyse statistique et probabiliste, au fondement même de sa « capacité à créer de nouvelles catégories sociales » comme dirait Hacking[193].

---

[190] BENBOUZID Bilel et Dominique CARDON, « Machines à prédire », *Réseaux*, n° 211, nº 5, 2018, p. 11.

[191] Comme le montre un bref survol de la littérature méthodologique (chapitre IV), ces deux traditions intellectuelles de la sociologie quantitative véhiculeraient non seulement des conceptions divergentes de la statistique et des probabilités, mais aussi de la recherche sociologique et de son raisonnement. DESROSIÈRES Alain, « Analyse des données et sciences humaines : comment cartographier le monde social ? », *Journ@l Electronique d'Histoire des Probabilités et de la Statistique*, vol. 4, nº 2, 2008, p. 11-18 ; CIBOIS Philippe, « Analyse des données et sociologie », *op. cit.*

[192] DESROSIÈRES Alain, « Masses, individus, moyennes: la statistique sociale au XIXe siècle », *op. cit.*, p. 61-62. Desrosières (2008), en particulier les chapitres 4, 7, 8 et 14, intitulés respectivement « Pour une politique des outils du savoir : le cas de la statistique », « Classer et mesurer : Les deux faces de l'argument statistique », « L'opposition entre deux formes d'enquête » et « Quetelet et la sociologie quantitative »). Voir aussi, du même auteur, Desrosières (1993), les chapitres 8 et 9, « Classer et coder » et « Modéliser et ajuster ».

[193] HACKING Ian, « Comment faire l'Histoire de la statistique ? », *op. cit.*, p. 189.

L'un des intérêts par exemple avec la distinction qu'il propose entre les « modes de connaissance » *monographique* et *statistique* est de montrer que ceux-ci ne s'opposent pas par leur capacité (ou non) de généraliser du savoir, mais plutôt par leur conception du « tout d'une société et des parties qui le composent » (de la totalité *sociale*), pouvant être fondée sur la valeur exemplaire (*type*) ou exhaustive (*limite*) des cas singuliers[194]. Renvoyant au moment de la description sous l'optique grangérienne[195], cette différence fut centrale dans son étude des conditions sociocognitives liées à l'invention et à la diffusion des enquêtes par « échantillonnage probabiliste », dite de sondage ou de « dénombrement partiel ». Selon Desrosières, ces méthodes auraient non seulement contribué à véhiculer les idées modernes de « représentativité » et d'« approximation », mais leur usage se serait uniquement répandu en parallèle de la formation de l'État-providence, montrant dès lors les imbrications étroites, presque indissociables, entre « modes de généralisation » du savoir (structure de pensée) et « modes de gestion » du pouvoir (structure d'action)[196].

Par conséquent, chercher à mettre à jour les façons dont se réarticulent, « se combinent continuellement » les doubles versants originaires du *travail* statistique permet à mon sens d'examiner « le[s] rapport[s] de connaissance définissant ce qu'est [le social], et comment on peut le connaître », d'en « saisir la relativité […] à travers l'émergence de nouveaux points de repères c'est-à-dire de nouvelles catégories cognitives relatives à des contenus d'expérience »[197]. Comment situer par exemple les nouvelles formes statistiques plutôt « exploratoires », issues du modèle de la corrélation, très loin d'être présentées comme de purs instruments descriptifs, dépourvus de toute logique d'action et de décision ?

---

[194] À ses yeux, les logiques de raisonnement propres à ces *formes* d'enquêtes, compatibles avec une certaine conception et utilisation des statistiques et du calcul des probabilités (les monographies étant compatibles avec celle de Quételet et ses recensements) peuvent se combiner et s'articuler de plusieurs façons, comme le montrent les enquêtes « représentatives ». DESROSIÈRES Alain, *Pour une sociologie historique de la quantification*, *op. cit.*, p. 49 et 143-151.

[195] « Expliquer, c'est-à-dire, ayant repéré un phénomène comme totalité et dissocié ses parties (c'est « décrire ») ayant établi les relations et les contraintes qui les associent (c'est « comprendre »), savoir insérer ce système dans un système plus vaste dont dépend sa genèse, sa stabilité et son déclin […] préciser dans un phénomène les relations du « local » et du « global ». » (Granger, 1992 : 11).

[196] *Cf.* DESROSIÈRES Alain, *La politique des grands nombres*, *op. cit.*, chapitre 7 (« La partie pour le tout : monographies ou sondages », p. 258-288). DESROSIÈRES Alain (2008), chapitre 8 (« L'opposition entre deux formes d'enquête : monographie et statistique »). Voir aussi DESROSIÈRES Alain, « La partie pour le tout : comment généraliser ? La préhistoire de la contrainte de représentativité », *Journal de la société statistique de Paris*, vol. 129, nº 1-2, 1988, p. 96-115.

[197] SABOURIN Paul, « La régionalisation du social. Une approche de l'étude de cas en sociologie », *Sociologie et sociétés*, vol. 25, nº 2, 1993, p. 73 et 76.

Tableau 2. – Repères théoriques fondamentaux préalables

GRILLE DE LECTURE (INITIALE) Dualités symboliques du *style* de raisonnement statistique et probabiliste	
Métrologique (« matérielle » : *comptage*) Volet « factuel » (INT) : Échelle ordonnée de grandeurs	**Classificatoire** (« formelle » : *repérage*) Volet « décisionnel » (EXT) : Liste exhaustive d'items
Langage des variables (« Calculateur ») Lecture en colonne : Par indicateur (ou critère)	**Langage des groupes** (« Organisateur ») Lecture en ligne : Par individu (ou unité)
Explication (« Confirmatoire ») : É/Prouver Orienté « vers l'action et le contrôle » (COG)	**Description** (« Exploratoire ») : Co/Ordonner Orienté « vers la visualisation et la narration » (EMP)
Modèle de régression (« Moyenne ») Propriétés algébriques : Traits « généraux »	**Modèle de corrélation** (« Dispersion ») Propriétés topologiques : Traits « particuliers »
Mise en série (Pôle « technique ») Valeur *limite* : Exhaustivité (Exactitude et Précision)	**Mise en récit** (Pôle « historique ») Valeur *type* : Exemplarité (Pertinence et Adéquation)
Monde de la nature (Inégalité, équivalence) « Positiviste »; « Scientisme rationaliste »	**Monde de la culture** (Identité, contingence) « Interprétativiste »; « Historicisme institutionnaliste »

Par les repères conceptuels préalables que fournit la grille d'analyse, le travail espère ainsi appréhender l'irréductibilité *sociale* – spécifique et générale – de la « raison statistique » comme ancrage temporel de la discipline sociologique. La posture conventionnaliste de Desrosières telle que conçue jusqu'au moment de rédiger ces lignes[198] parait justement résider dans la tension constante opposant le « réalisme métrologique, photographique ou cartographique » au « constructivisme relativiste »[199]. La même ambiguïté semblait également préoccuper l'épistémologue canadien Hacking lorsqu'il affirmait dans sa leçon inaugurale au Collège de France faite le 11 janvier 2001

> Il ne suffit pas de montrer du doigt des illusions pour en venir à bout, ou de se contenter de les tourner en ridicule. On n'échappe pas aux classifications en proclamant qu'elles sont des productions historiques, sociales, et mentales. Nous vivons dans un monde classifié, que l'on pourrait déconstruire pour s'amuser, mais nous aurons besoin de ces structures pour penser, en attendant qu'elles soient modifiées, non pas par déconstruction, mais par construction, par création[200]

---

[198] Plusieurs dualismes inspirés de ma lecture des trois auteurs que j'articule à partir de ma compréhension limitée et possiblement erronée. Ils seront abordés plus en profondeur aux chapitres IV et V.

[199] « Quel langage imaginer qui ne soit ni celui du réalisme métrologique naïf des sciences de la nature (qui est comme le rêve perdu impossible des sciences sociales quantitatives), ni celui d'un constructivisme relativiste, vu comme la négation de la dure réalité d'un monde social dont la description ne relèverait que de l'arbitraire de rapports sociaux contingents orientés par des intérêts particuliers ? » DESROSIÈRES Alain, *Pour une sociologie historique de la quantification*, *op. cit.*, p. 65 et 104. (Voir Chapitre 4 « Pour une politique des outils du savoir » et Chapitre 6 « Du singulier au général »). Pour le rapport entre le nécessaire et l'arbitraire dans les sciences sociales du social, voir aussi GRANGER Gilles Gaston, « Conventions, normes, axiomes dans la connaissance des faits humains », *Dialectica*, vol. 41, nº 1/2, 1987, p. 39-67.

[200] HACKING Ian, « Leçon inaugurale », *Chaire de philosophie et histoire des concepts scientifiques (2000-2006)*, Collège de France, 11 janvier 2001, p. 6. En ligne <https://www.college-de-france.fr/media/ian-hacking/UPL7027195376715508431_Le__on_inaugurale_Hacking.pdf.>

CHAPITRE III – MÉTHODOLOGIE
Démarche sociologique d'étude de cas

Dans une démarche méthodologique sociologique d'étude de cas[201], 1) les méthodes de régression multiple (logistique) (LR) plus « classiques » sont comparées à deux autres approches plus récentes que sont 2) les arbres de décision (DT) tels qu'implémentés dans l'algorithme CART[202] et 3) leur extension dans l'algorithme des forêts aléatoires (RF)[203]. Il convient de rappeler que bien qu'il s'agisse d'outils appartenant à la classe d'apprentissage « supervisée », la présente recherche ne les appliquera pas dans le cadre « méthodologique » – théorique, technique et pratique – usuel en machine learning (ML)[204].

Par conséquent, ce troisième chapitre est consacré à la présentation de la méthodologie « sociologique ». D'abord, nous précisons l'objet de recherche et les questions plus spécifiques du présent travail, pour ensuite décrire succinctement la problématique de référence empirique, appliquée ou « concrète » (illustrative) – relative à la santé perçue au Canada – ayant permis de construire des données sociologiques. Après avoir abordé de manière précise et concise les opérations élémentaires et fondamentales constitutives des trois techniques d'analyse, la pertinence théorique et pratique du choix de ces méthodes est justifiée. Ce chapitre se conclut par quelques précisions concernant le processus de description et d'analyse sémantique et pragmatique, et plus exactement, le « langage » de présentation des résultats aux chapitres VI et V, croisant formules mathématiques « froides et déshumanisées »[205] avec formules sociologiques.

---

[201] Cette approche a été privilégiée puisqu'elle permet « d'analyser des réalités négligées par la science » (Roy, 2009 : 209).

[202] Un acronyme pour Classification And Regression Trees. Cf. BREIMAN Leo et al., Classification and Regression Trees, Chapman & Hall, 1984.

[203] BREIMAN Leo, « Random Forests », *Machine Learning*, vol. 45, nº 1, 2001, p. 5-32.

[204] Comme mentionné au chapitre I, cela reviendrait essentiellement à séparer l'ensemble de données en deux sous-échantillons, dont l'un est destiné à l'entrainement des modèles et l'autre à leur évaluation, de sorte à chercher à optimiser leur capacité à se généraliser sur des données (observations ou « individus ») n'ayant pas servi à les ajuster. C'est pour la même raison que les pratiques d'élagage pourtant centrales à la méthode CART et l'erreur Out-Of-Bag calculé pour les RF ne seront pas abordées. Pour plus de détails, voir HASTIE Trevor, TIBSHIRANI Robert et FRIEDMAN Jerome H., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, Springer, 2009 ; JAMES Gareth, WITTEN Daniela, HASTIE Trevor et TIBSHIRANI Robert, *An Introduction to Statistical Learning: with Applications in R*, New York, Springer, 2013.

[205] Un clin d'œil à la mise en garde de Selz et Maillochon (2009 : 85).

3.1 Objet de recherche

3.1.1 Les algorithmes comme « procédures », « modèles » et « rapports »

En s'inspirant des trois auteurs mobilisés dans le cadre théorique, la recherche propose de comparer des familles/techniques algorithmiques sous l'angle de leurs opérations de connaissance, des implications qui en découlent et de leur articulation avec certaines configurations sociales.

Par l'idée de *mises en forme algorithmiques de la vie sociale*, les algorithmes de calcul opèrent, de façon partiellement ou complètement automatisée, des mises en forme et en ordre plurielles et différenciées du « social » et de ses propriétés[206]. Comme moyens d'investigation inscrits dans la trajectoire du *style* statistique et probabiliste de raisonnement scientifique[207], les algorithmes de quantification déterminent des « espaces de commune mesure[208] » à des fins pratiques, produisent des référents symboliques susceptibles d'orienter « des possibilités » d'être, « de choix, et d'action »[209]. Conçus sous des perspectives opératoires différentes, les algorithmes se manifestent à la fois comme « procédures », « modèles » et « rapports ».

Comme procédure d'analyse explicite et structurée, un algorithme consiste en une série d'opérations de « calcul »[210], finie et non ambiguë en vue de résoudre un problème donné, de produire un résultat sous forme de « modèle » de relations. Comme formes *virtuelles* (possibles ou non) d'objets pensables (actualisables, réalisables ou non), provisoirement stabilisées, les modèles mathématiques constituent des traces empiriquement observables des processus opératoires, interprétables et utilisables comme « structures » ou rapports entre des *formes* et des *contenus*[211]. Que les algorithmes soient effectivement (ou non) utilisés à titre d'outils de prédiction, de classification ou d'aide à la décision, l'*usage* et le *sens* appliqué des mathématiques intégrées dans

---

[206] « Le social a une historicité et une spatialité propres, il est un phénomène vivant et signifiant, sa consistance n'est pas substantive, mais relationnelle et processuelle, [actuelle et virtuelle], etc. » (Sabourin, 2021 : 21-22).

[207] HACKING Ian, « Style pour historiens et philosophes », *op. cit.*

[208] DESROSIÈRES Alain, La politique des grands nombres, op. cit.

[209] HACKING Ian, « L'ontologie historique », *op. cit.*, p. 290.

[210] Par la notion de « calcul », on entend l'idée « d'opérations explicitement et univoquement définies et réglées » (Granger, 1988 : 13 dans Hamel, 1989 : 64), à la fois « classificatoires » (*formelles*) et « métrologiques » (*matérielles*). Dans ce mémoire, « mesurer » signifie « faire correspondre aux opérations de calcul effectuées (…) des opérations empiriques bien définies pour le phénomène considéré » (Granger, 1992 : 8), la « mise en œuvre réglée » de mises en équivalence (Desrosières, 2008 : 11).

[211] GRANGER Gilles-Gaston, Formes, Opérations, Objets, op. cit.

la conception et le développement des algorithmes renvoient à des rapports « historiques » (spécifiques) au monde empirique (générique). Ici, les états « futurs » (non effectivement réalisés) « se font historiques, en venant à exister[212] » et renvoient donc à l'aspect le plus « individuel » de la connaissance des faits humains[213].

3.1.2 Questions/Objectifs général et spécifiques

L'objectif de l'analyse comparée consistera alors à examiner comment certaines techniques/familles d'algorithmes opèrent certains modes de classification, facilitent ou défavorisent des représentations du monde et peuvent aussi être performatives dans des activités sociales. Trois questions de recherche plus spécifiques peuvent préciser cet objectif général : Comment se combinent différents procédés d'analyse pour construire des « modèles », des représentations (mathématiques, *virtuelles*) de phénomènes (sociaux, *actuels*) de l'expérience « toujours singulière » (historique, *personnelle*) ? Quelles pratiques et lectures *possibles* du social, plus ou moins compatibles avec certaines formes sociales, sont offertes et actualisées par les catégories auxquelles renvoient certains algorithmes statistiques et informatiques ? Que témoignent les dispositifs d'analyse algorithmiques à l'égard des modes de différenciation et de structuration sociales ?

Par ces sous-questions, la recherche s'intéresse davantage à l'articulation réciproque entre les formes mathématiques et les formes sociales, moins aux effets des unes sur les autres[214]. Par leurs propriétés, les formes mathématiques produisent des objets (ou « agrégats ») statistiques, des classifications sociales plus ou moins compatibles à des ontologies sociales, c'est-à-dire à des entités définies dans des conceptions de la vie sociale.

---

[212] HACKING Ian, « L'ontologie historique », *op. cit.*, p. 290.

[213] GRANGER Gilles-Gaston, Le Probable, le Possible et le Virtuel, op. cit.

[214] Types de question à poser à nos matériaux (*Cf.* Grille conceptuelle) : Comment sont classifiés les individus et reconstitués des groupes sociaux (types d'analyse) ? Dans quels « systèmes de grandeurs » (Desrosière, 1988 : 61) ? Comment procède-t-on pour déterminer l'état (le fait ou le comportement) virtuel le plus susceptible de s'actualiser (modes de calcul) ? En concevant le probable tel que le conçoit Granger, c'est-à-dire comme une « mesure » qui structure des possibilités (1995 : 130), comment les probabilités sont-elles calculées ? Comment peut-on interpréter les résultats des analyses (types d'énoncé) ? Quelles sont les principales entités mises en scène ? Comment « construit-on la généralité » (Desrosière, 2008: 15) (types de cumul et de validité du savoir) ? Quelles « structures d'action » sont plus ou moins adaptées (types de gestion) ? « Comment passer de la « partie » au « tout » ? » (Desrosière, 1993: 260) « Qu'est-ce qu'un individu ? Comment mettre en équivalence plusieurs individus pour constituer des catégories ? Qu'est-ce que le tout duquel on infère ? » (Desrosières, 2008 : 143-144). Comment les classifications sociales façonnent-elles certains espaces possibles de savoir et d'action (Hacking, 2003) ?

3.2 Construction des données sociologiques

> « La mesure elle-même est la recherche de l'explication. Elle n'est ni qualitative, ni quantitative […], elle est la recherche des propriétés de l'objet défini dont la construction théorique constitue l'explication. » (Houle, 1982 : 5)
>
> « La mesure n'a pas à susciter une opposition, mais bel et bien un *rapport* qualité-quantité pour donner lieu à l'objectivation d'une expérience au sein d'un savoir théorique. » (Hamel, 1997 : 63)
>
> « la mesure a la double fonction d'explorer en réfléchissant et de réfléchir en explorant. » (Pirès, 1997 : 61)

3.2.1 Analyse secondaire de données d'enquête (ESG2016)

À l'instar de A. Pirès citée en exergue, les formes mathématiques utilisées et leurs paramétrages ont été décomposés afin de réfléchir aux différentes façons par lesquelles sont classifiés les individus, reconstitués des groupes sociaux et modélisés des comportements. Pour retracer concrètement les opérations qu'effectuent les algorithmes de calcul, la recherche s'appuie sur les données issues de l'*Enquête sociale générale* de 2016 menée par Statistique Canada, et plus précisément le fichier de microdonnées à grande diffusion (FMGD) de l'ESG de 2016, accessible via la plateforme ODESI[215]. Notre échantillon est composé de 18 483 individus (ou observations), après exclusion des données manquantes, sur les 19 609 de cette enquête.

Pour mener la comparaison des méthodes, nous procédons en nous appuyant sur un exemple concret, tiré d'une problématique déjà étudiée au premier cycle, qui était celle des perceptions de santé[216]. Cette thématique a été choisie parce qu'elle offrait un cadre plus familier pour manipuler des données. Au total, un ensemble de variables a été pré-sélectionné à des fins d'analyse: une

---

[215] Les chercheur-e-s appliqué-e-s pourront sans doute s'étonner, à juste titre, des informations concernant l'enquête, l'échantillon et la transformation des variables qui brillent par leur absence. Notons simplement que les données de l'ESG2016 ont été choisies pour ne pas avoir à nous préoccuper du problème des classes « non balancées » (Lebart et al., 1995). D'ailleurs, des épreuves/tests empiriques ont aussi été menés sur des données issues d'autres plateformes et enquêtes, avec des échantillons de taille variable, en appliquant diverses variantes et extensions des méthodes (voir des exemples en annexe).

[216] Ce mémoire reprend la problématique de recherche que ma collègue Louise-Andrée Boudreault et moi avions élaborée au sujet des déterminants « sociaux » qui médient et qui interagissent dans la manière dont les autochtones perçoivent leur état de santé en général. Nos analyses corroboraient l'idée selon laquelle l'évaluation que les individus font de leur santé n'est pas qu'une question relative à l'état biologique ou médical, qui n'appellerait que le système de soins de santé, et qu'elle semble encore loin de reposer uniquement sur des modes de vie personnels.

variable d'intérêt, dite dépendante (VD) – la santé autoévaluée (*Y*) – ainsi qu'une quinzaine de caractéristiques individuelles appelées variables indépendantes (VI). En somme, il s'agira de mieux comprendre le fait qu'une personne se déclare être en « excellente » ou « très bonne » santé générale (Y=1), plutôt que « bonne », « passable » ou « mauvaise » (Y=0), à partir des caractéristiques synthétisées ci-dessous[217]:

Tableau 3. – Synthèse des variables indépendantes

Variables	**Modalités / Valeurs**
SEX (sexe des répondant-e-s)	0-Homme; 1-Femme
AGE (groupes d'âge)	1-15 à 24 ans; 2-25 à 34 ans; 3-35 à 44 ans; 4-45 à 54 ans; 5-55 à 64 ans; 6-65 à 74 ans; 7-75 ans et plus
MAT (état matrimonial)	1-Célibataire; 2-Marié; 3-Union libre; 4-Veuf/Séparé/Divorcé
IMM (statut d'immigrant)	0-Non; 1-Oui
MIN (minorité visible)	0-Non; 1-Oui
EDU (niveau de scolarité)	0-Primaire; 1-Secondaire; 2-Collégial; 3-Universitaire
REV (revenu personnel)	1-Moins de 25 000$; 2-25 000$-49 999$; 3-50 000$-74 999$; 4- 75 000$-99 999$; 5-100 000$-124 999$; 6-125 000$ et plus
TRA (situation professionnelle)	0-Emploi; 1-Pas d'emploi
PSC (classe sociale)	0-Inférieure; 1-Moyenne; 2-Supérieure
DOS (sent. appart. à la coll.)	0-Non; 1-Oui
ENV (qualité de l'env. local)	0-Non; 1-Oui
GEO (région de résidence)	0-Urbaine; 1-Rurale
ACT (activités de plein air)	0-Non; 1-Au moins une
FOO (habitudes alimentaires)	1-Pas dern. mois; 2-Quelques fois/mois; Plusieurs fois/sem.
DDR (consommation d'alcool)	0-Jamais; 1-Occasion; 2-Régulier
SKM (tabagisme)	0-Non-fumeur; 1-Fumeur
SMG (stress dans la vie)	1-Pas du tout; 2-Pas tellement; 3-Un peu; 4-Assez/Extrêmement
INC (incapacité)	0-Aucune; 1-Une

[217] Pour alléger parfois le texte, 0=« Négative » et 1=« Positive ». Certaines subtilités quant à l'univers des variables ne seront pas réitérées à chacune interprétation afin d'alléger également le texte. Pour plus de détails, voir le tableau récapitulatif en annexe.

3.2.1.1 Technique d'échantillonnage par cas multiples

L'analyse secondaire des données de l'ESG de 2016, on l'aura vu, est utilisée comme méthode pour construire des données sociologiques, et notamment un échantillon de modèles mathématiques. Cet échantillon a été diversifié, en faisant varier les paramètres des modèles (*p. ex.* le nombre de variables) et/ou les hyperparamètres des méthodes (*p. ex*. le nombre de feuilles, d'arbres…), de façon à faciliter la compréhension des logiques de fonctionnement et de raisonnement propres à chaque technique. Trois groupes de modèles allant d'un plus « simple » à un plus « complexe » seront présentés[218]. L'analyse des trois méthodes d'analyse a été menée à l'aide du logiciel d'analyse SPSS, du langage R[219] et de son interface RStudio, de façon à accroitre la validité interne de la recherche[220]. Comme principaux matériaux de la recherche, les sorties informatiques donnent accès aux modèles mathématiques ainsi qu'à certains paramétrages des procédés d'analyse participant à la « mise en modèle ».

3.2.1.2 Le statut des données d'enquête

Bien que cette problématique appliquée, empirique ou « concrète » n'ait pas de lien direct avec les préoccupations centrales d'ordre épistémologique de la recherche[221], celle-ci permet de délimiter un « cadre de référence », une situation dans laquelle les différences de méthode pourront difficilement s'expliquer par des différences dans les données[222]. « Contrôler » autrement dit les conditions concrètes des analyses de la recherche permet de mieux observer et de mieux cerner les implications cognitives et pratiques des dispositifs retenus dans l'appréhension d'un phénomène social donné[223]. Pour mieux illustrer les opérations automatisées dans les analyses, le travail de

---

[218] Par simplicité, les résultats de régression logistique binaire/dichotomique (plutôt que multinomiale ou ordinale) sans terme d'interaction ou polynomial modélisé seront uniquement présentés dans ce mémoire (données non pondérées normalisées).

[219] Les packages « glm », « rpart » et « randomForest » ont été les principaux utilisés.

[220] GAUTHIER Benoît. (dir.), *Recherche sociale: de la problématique à la collecte des données*, 5e éd., Sainte-Foy, Presses de l'Université du Québec, 2009.

[221] Bien que nous pourrions nous y intéresser et que nous avons déjà tenté de le faire avec plus ou moins de succès dans un autre travail. Dans le cadre d'une véritable recherche appliquée, les trois méthodes étudiées seraient utilisées conjointement dans une démarche plus compréhensive et explicative que prédictive, en cherchant *p. ex.* à approfondir les effets d'interactions repérés dans les arbres au moyen de régressions par entrée standard et/ou hiérarchique (voir en annexe des exemples plus détaillés).

[222] VAN CAMPENHOUDT Luc, Jacques MARQUET et Raymond QUIVY, *Manuel de recherche en sciences sociales*, 5e éd., Paris, Dunod, 2017.

[223] GAUTHIER Benoît. (dir.), Recherche sociale: de la problématique à la collecte des données, op. cit.

comparaison sera accompagné de cas de figure (annexe): Gisèle-Observation no 517, Lise-no 9 213 et Germain-no 19 245[224].

Enfin, il est important de retenir que les principes généraux des méthodes choisies resteraient sensiblement les mêmes[225], que celles-ci soient utilisées pour évaluer les inégalités de revenu, les discriminations à l'emploi, les risques de récidive ou de suicide, les comportements d'achat ou les demandes de prêts bancaires (« credit scoring »). Comme l'indiquait Desrosières, « [l]a quantification offre un langage spécifique, doté de propriétés remarquables de transférabilité, de possibilités de manipulations standardisées par le calcul, et de systèmes d'interprétations routinisées ».[226]

[224] ROY Simon N., « L'étude de cas », dans Benoît GAUTHIER (dir.), *Recherche sociale: De la problématique à la collecte des données*, 5e ed., Québec, Les Presses de l'Université du Québec, 2009, p. 199-225.

[225] Plusieurs variantes « techniques », non substantielles, même chose pour les logiciels et implémentations ?

[226] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 12.

3.2.2 Identification des autres « supports »

Outre les sorties informatiques, le travail s'appuie également sur divers textes[227] de méthodologie et d'épistémologie de la statistique et de la sociologie, mais de manière moins formalisée qu'initialement envisagé[228]. Un intérêt particulier a été porté aux pages comportant les contenus suivants : les règles relatives à l'application des méthodes et à l'interprétation des résultats ainsi que les conceptions du rapport théorie/expérience sous-jacentes. Deux groupes de documentation écrite peuvent être distingués.

1) Méthodologie/Épistémologie statistique : Une première catégorie rassemble des textes consacrés à l'enseignement des rudiments de la « science des données », à l'intention des étudiant-e-s, des chercheur-e-s universitaires et des professionnel-le-s de divers milieux sociaux (publics ou privés). Les documents de référence proposant des survols synthétiques d'un ensemble de techniques statistiques et informatiques ont été privilégiés, en particulier ceux discutant des différences entre méthodes, techniques, théories/approches, ceux plus proches du domaine de la statistique que de celui de l'informatique ainsi que ceux traitant de problématiques similaires à celles des sciences sociales appliquées. Ceux-ci renseignent sur les fonctions mathématiques, les formules et concepts de base constitutifs des règles implémentées dans les algorithmes.

2) Méthodologie/Épistémologie sociologique : Une deuxième catégorie de références comprend des textes produits par et pour des sociologues. Sans prétention d'exhaustivité, les discours s'étant explicitement positionnés sur la place des approches statistiques, parfois dites « quantitatives » dans la recherche sociologique[229] ainsi que ceux portant sur le moment et les

---

[227] Ces textes se présentent généralement sous forme d'articles de revues scientifiques, de ressources pédagogiques, de chapitres d'ouvrages collectifs, de livres ou manuels de méthodologie de recherche.

[228] Étant donné que je ne suis jamais parvenue à finaliser mon analyse de discours assistée par le logiciel Atlas.ti, j'ai décidé de trouver une fonction/utilité alternative à certains extraits repérés : présenter l'analyse et les résultats dans des sémantiques « sociologiques » (section 3.4.1.1).

[229] Parmi ces discours sociologiques se retrouvent les critiques issues des courants phénoménologiques, interactionnistes et ethnométhodologiques des années 60-70 (P. Sorokin, A. Cicourel, H. Blumer, E. Goffman, H. Becker, H. Garfinkel, C. Javeau…), les critiques épistémologiques et leurs dimensions « cognitives » (F. Simiand, M. Halbwachs, J.-C. Passeron…), mais aussi les considérations « internes » à la méthodologie statistique issues de la sociologie française (A. Degenne, L. Lebart, H. Rouanet, F. Lebaron…). Voir les chapitres 8 « L'opposition entre deux formes d'enquête » et 9 « Entre réalisme métrologique et conventions d'équivalence » de Desrosières (2008) qui constituent notre point de départ. Voir aussi SINGLY François de, « Les bons usages de la statistique dans la recherche sociologique », *Économie et statistique*, vol. 168, nº 1, 1984, p. 13-21 ; MARTIN Olivier, « Mathématiques et sciences sociales au XXème siècle », *op. cit.* ; MOULIN Stéphane et Jean-Pierre BEAUD, « Quantification et mesure », dans Frédéric BOUCHARD, Pierre DORAY et Julien PRUD'HOMME (dir.), *Sciences, technologies et sociétés de A à Z*, Presses de l'Université de Montréal, 2015, p. 186-188.

principes même de l'analyse de données ont été privilégiés. En permettant de repérer quelques éléments de critique saillants des outils « traditionnels », ces textes serviront d'ancrages supplémentaires aux repères de la grille conceptuelle pour mieux saisir la portée et la spécificité des enjeux que soulèvent certaines techniques d'analyse plus récentes[230].

Comme *forme* de matériau (ici, « support » additionnel, intermédiaire à la recherche), les discours de méthodes produisent et stabilisent des « conceptions » de la nature et de l'ordre de l'objet à connaître, établissent une relation sociale d'argumentation à la fois descriptive et prescriptive, en justifiant et/ou en dénonçant des rapports de connaissance, des « modes de construction et diffusion du savoir »[231]. Ce type de discours présente donc l'intérêt de faire ressortir des thèmes fondamentaux de la *pensée* et de la *pratique* « sociologique » dans ses développements et régularités « historiques »[232].

Il convient de souligner que le choix de ces textes a été grandement influencé par les lectures et échanges de six séminaires : SOL6941-H19 – Séminaire de projet de mémoire (Cécile Van de Velde); SOL6010-H19 – Statistiques sociales et politiques publiques (Stéphane Moulin); SOL6210-A19 – Analyse quantitative avancée (Claire Durand); SOL6205-H20 – Terrain et décloisonnement des méthodes (Anne Calvès); SOL6212-E20 – Analyse du discours (Paul Sabourin) et; SOL6447-A20 – Épistémologie et méthodologie qualitative (Jacques Hamel).

---

[230] À titre d'exemple, comment les principes théoriques des forêts d'arbres aléatoires se situent par rapport aux débats sur l'« Homme moyen » de Quetelet et aux effets de totalisation du « langage des variables », vivement dénoncés depuis Blumer en 1956 (Bryman, 2001) ?

[231] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 147.

[232] BERTHELOT Jean-Michel, « Les nouveaux défis épistémologiques de la sociologie », *Sociologie et sociétés*, vol. 30, nº 1, 1998, p. 23-38. « Les distinctions abusives entre théorie et méthode ['que recouvre l'opposition qualitatif/quantitatif'] nous ramènent aux problèmes fondamentaux de la démarche de recherche elle-même, à la démarche proprement méthodologique d'articulation de la théorie à la méthode, de définition opératoire qui fonde la construction d'un objet ». HOULE Gilles, « Présentation. La sociologie : une question de méthode ? », *Sociologie et sociétés*, vol. 14, nº 1, 1982, p. 4-5.

3.3 Trois techniques d'apprentissage, deux approches

Comme nous le verrons dans les prochains paragraphes, après avoir présenté « formellement » les trois « cas statistiques » de l'étude, le choix de ces méthodes d'apprentissage « supervisé » est justifiable d'un point de vue théorique (conceptuel) et pratique (technique).

3.3.1 La régression logistique (LR)

Développée à partir des années 1940, la régression logistique (LR pour *logistic regression*) constitue un cas particulier appartenant à la famille des modèles linéaires généralisés (*generalized linear models,* GLM)[233], qui « généralise la logique du modèle linéaire » au moyen d'une fonction de lien, nommée logit[234]. Inspirés de la démarche expérimentale, cette méthode paramétrique, couramment dite « économétrique », repose sur l'hypothèse qu'un phénomène social s'explique par plusieurs « facteurs » (ou dimensions) observables « indépendamment des autres »[235]. Comme « forme d'analyse tabulaire », aussi connue sous le nom d'« analyses multivariées »[236], le principe de base commun aux techniques de régression multiple (linéaire ou logistique) consiste en une « procédure de neutralisation » de l'effet des variables[237].

Estimés au moyen de la méthode de maximum de vraisemblance[238], les coefficients des paramètres du modèle se combinent linéairement, de façon additive ou multiplicative[239]. Cette équation de régression permet de prédire des probabilités individuelles pour toute configuration d'attributs[240]. L'effet de chaque facteur sur la probabilité de se percevoir en excellente ou très

---

[233] *Cf.* Nelder et Weddeburn (1972). Hosmer et Lemeshow (2000).

[234] DESROSIÈRES Alain, « Comparer l'incomparable. Essai sur les usages sociaux des probabilités et des statistiques », dans Jean-Philippe. TOUFFUT (dir.), *La Société du probable: les mathématiques sociales après Augustin Cournot*, Paris, Albin Michel, 2007, p. 181 ; DESROSIÈRES Alain, *Pour une sociologie historique de la quantification*, *op. cit.*, p. 165.

[235] CIBOIS Philippe, *Les méthodes d'analyse d'enquêtes*, Lyon, ENS Éditions, 2015, p. 83.

[236] Systématisées dans la sociologie américaine avec Lazarsfeld à la fin des années 60. *Cf.* CHAPOULIE Jean-Michel, « Un type d'explication en sociologie : les systèmes de variables en relations causales », *Revue française de sociologie*, vol. 10, nº 3, 1969, p. 333-351.

[237] CIBOIS Philippe, « Modèle Linéaire contre modèle logistique en régression sur données qualitatives », *BMS*, vol. 64, nº 1, 1999, p. 6 ; CIBOIS Philippe, « Les techniques d'analyse "toutes choses égales par ailleurs" », *Les méthodes d'analyse d'enquêtes*, Lyon, ENS Éditions, 2015, p. 83.

[238] *Cf.* CIBOIS (2015) « Annexe au chapitre V : algorithme du maximum de vraisemblance », p.101-108.

[239] CIBOIS Philippe, « Les techniques d'analyse "toutes choses égales par ailleurs" », *op. cit.*, p. 99-100.

[240] *Cf.* Riandey et al. (1991 : 83); Deauvieau (2010); Des Nétumières (1997 : 293).

bonne santé (Y=1) plutôt que mauvaise, passable ou bonne (Y=0) est interprété en rapportant toutes les valeurs des autres variables à leur modalité de référence[241].

3.3.2 Les arbres de décision (DT)

Proposée par Breiman *et al*. (1984), la procédure CART marquerait le point culminant des approches inductives fondées sur la représentation par arborescence[242] et demeurerait parmi les algorithmes les plus utilisés encore aujourd'hui pour produire des arbres[243].

Décrits du point de vue des opérations, les arbres de décision (DT) consistent en des méthodes de « segmentation », de « partitionnement » (itératif, récursif), rattachées au principe "diviser pour régner". L'idée générale consiste à séparer de façon successive les n observations de l'échantillon, par le biais d'attributs (Xi) qui rendent les « nœuds », les plus « purs » ou homogène possible en regard de la variable d'intérêt (Y). Appliquée à notre exemple, cette procédure sélectionne les variables les plus « liées » à la santé perçue, de manière à regrouper le plus répondant-e-s ayant le même état de santé ensemble[244] . Évaluée à partir d'un critère de « pureté » calculé pour tous les points de coupures possibles à chaque nœud, la découpe « optimale » correspond à l'attribut qui maximise la réduction d'hétérogénéité[245].

Un modèle d'arbre est composé d'un ensemble de règles/conditions logiques de type « Si… Alors… » qui relient toutes les « feuilles » de l'arbre, aussi appelées « nœuds terminaux », à la « racine »[246]. Les variables situées les plus près de la racine de l'arbre peuvent parfois être

---

[241] BILAND Émilie, Jean-Sébastien EIDELIMAN et Séverine GOJARD, « Ceteris (non) paribus ? », *Geneses*, n° 73, nº 4, 2008, p. 39 ; CIBOIS Philippe, « Modèle Linéaire contre modèle logistique en régression sur données qualitatives », *op. cit.*, p. 20 ; SELZ Marion et Florence MAILLOCHON, *Le raisonnement statistique en sociologie*, Paris, PUF, 2009, p. 63.

[242] Par exemple CHAID ou encore, ID3 et C4.5. *Cf*. Lebart et *al*. (1995); Falissard (2005 : 297); Gueguen et Nakache (1988); Genuer et Poggi (2018 : 295); Pour des liens plus approfondis entre LR et DT, voir Nakache et Confais (2003).

[243] GHATTAS Badih, « Prévisions par arbres de classification », *Mathématiques et sciences humaines*, vol. 37, nº 146, 1999, p. 31-49.

[244] Parfois les termes de « segments », de « partitions » sont aussi utilisés pour désigner les « sous-ensembles/groupes », résultant de chaque découpage. « La forme des modèles obtenus conduit à délimiter des régions rectangulaires de l'espace des variables. » (Tufféry, 2017 : 673)

[245] Pour des problèmes de classification, les arbres CART utilisent l'indice de Gini, mais plusieurs autres fonctions peuvent être utilisées pour mesurer le degré de « pureté » des groupes (entropie/gain d'information, variance, khi-2 d'écart à l'indépendance, etc.).

[246] Voir des exemples du *type* de lecture ou d'interprétation dans Saporta (2015 : 490) et Tufféry (2017 : 652).

interprétées comme étant les plus pertinentes ou « utiles » relativement à la variable d'intérêt, avec une diminution d'impureté des nœuds plus élevée[247]. L'estimation (ou l'attribution) des « probabilités conditionnelle d'appartenance »[248] s'effectue à partir des « fréquences relatives »[249] de chaque valeur/classe de la VD indiquée aux feuilles de l'arbre.

3.3.3 Les forêts aléatoires (RF)

Introduites au tournant des années 2000[250], les forêts d'arbres aléatoires (RF) incarnent un exemple typique de ce que l'on désigne par l'idée de « boîte noire » en ML et tendent désormais à se hisser parmi les méthodes les plus performantes en termes de précision des prédictions.

Les forêts aléatoires (RF) appartiennent à la « famille des méthodes d'ensemble », d'agrégation[251]. L'idée générale est de produire un « collectif » d'arbres décisionnels les plus diversifiés possibles, « décorrélés » ou indépendants les uns des autres[252]. Comme leur nom l'indique, les RF utilisent la notion d'aléatoire pour générer de la variété en perturbant la construction des sous-modèles que l'on agrège : chaque arbre est « construit sur un nombre restreint d'observations », à partir d'une sous-partie de l'échantillon de base (*bootstrap*), en « sélectionnant au hasard, à chaque division, un nombre restreint de variables ». Dans le cas d'un problème de classification (variables discrètes, nominales) qui est le nôtre, les résultats obtenus par RF sont fournis par le « vote majoritaire » des « réponses » (ou « prédictions individuelles ») données par chacun des arbres de la forêt (*bagging*)[253].

---

[247] GENUER Robin et Jean-Michel POGGI, « Arbres CART et Forêts aléatoires, Importance et sélection de variables », dans Myriam MAUMY-BERTRAND, Gilbert SAPORTA et Christine THOMAS-AGNAN (dir.), *Apprentissage statistique et données massives*, Paris, Technip, 2018, p. 313.

[248] Saporta (2015 : 489-490); Tufféry (2017 : 675-676).

[249] LEBART Ludovic, Marie PIRON et Alain MORINEAU, *Statistique exploratoire multidimensionnelle*, Paris, Dunod, 1995, p. 315.

[250] BREIMAN Leo, « Random Forests », *op. cit.*

[251] GENUER Robin et Jean-Michel POGGI, « Arbres CART et Forêts aléatoires, Importance et sélection de variables », *op. cit.*, p. 301, 303 et 317. « Une forêt aléatoire est l'agrégation d'une collection d'arbres aléatoires » (Grenier et Robin, 2018 : 317)

[252] STROBL Carolin, James MALLEY et Gerhard TUTZ, « An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests », *Psychological Methods*, vol. 14, nº 4, 2009, p. 324.

[253] GENUER Robin et Jean-Michel POGGI, « Arbres CART et Forêts aléatoires, Importance et sélection de variables », *op. cit.*, p. 300. « Le résultat d'une classification est alors celui du vote de chacun de ces modèles tandis qu'une simple moyenne fournit la prévision d'une variable quantitative » (Besse, 2001 : 11)

3.4 Justification des cas de l'étude

3.4.1 La tension méthodologique « typique » dans la « culture algorithmique »

Dans la littérature statistique/informatique (méthodologique), l'approche par arbres de décision (DT) est souvent présentée comme étant l'une des principales rivales aux méthodes de régression classiques (linéaire ou logistique). Contrairement à la régression logistique (LR) qui exige que les termes d'interaction soient postulés préalablement, l'approche par arborescence permettrait plutôt de les détecter « automatiquement »[254].

Au plan pratique, les DT présentent l'avantage de rendre les ambitions du présent travail un peu plus faisables. Outre les GLM dont fait partie la LR, cette approche constituerait l'une des rares techniques de référence en ML reconnue pour son interprétabilité, par opposition à leur extension très populaire par forêts aléatoires (RF). Comme nous le verrons, les forêts d'arbres décisionnels pallient la tendance au « surajustement » des arbres *individuels* (singuliers) aux données d'entraînement, le fait qu'ils soient très sensibles ou instables face à de légères variations (modifications) dans l'échantillon d'apprentissage (figure 4), mais perdent en revanche la « simplicité » et la « facilité d'interprétation » des arbres.

En outre, à ma connaissance, le *rapport* entre les approches par arbre et par forêt exprime de manière exemplaire le compromis entre l'efficacité prédictive et l'intelligibilité des modèles propre à la culture de modélisation algorithmique, tel que décrit par Breiman (chapitre I)[255]. Puisqu'un seul arbre correspond à l'unité de base des RF[256], proposer d'examiner de manière systématique les processus classificatoires et métrologiques de plusieurs arbres décisionnels « simples » (ou pris séparément) apparait judicieux pour essayer de réfléchir aux implications sociales relatives à l'opacité associée au fonctionnement des modèles de forêt, avant de reconduire systématiquement l'argument de la « boite noire » qui circulent au sujet des RF ou d'en théoriser les conséquences. Qu'implique entre autres la « flexibilité » croissante des modélisations dans la généralisation inductive par arbres décisionnels, en termes de catégories de pensée et d'action ?

---

[254] LEBART Ludovic, Marie PIRON et Alain MORINEAU, *Statistique exploratoire multidimensionnelle*, *op. cit.*

[255] BREIMAN Leo, « Statistical Modeling », *op. cit.*

[256] Voir Chapitre 15 TUFFÉRY Stéphane, *Data mining et statistique décisionnelle: la science des données*, 5e éd., Paris, Technip, 2017.

3.4.2 Critiques « conventionnelles » de la méthodologie « conventionnelle »

Comme on l'a vu, les controverses épistémologiques que suscitent les techniques d'apprentissage issues des *big data,* concernant par exemple le primat des données empiriques par rapport aux modèles/hypothèses théoriques, réactualisent de vieux débats de méthodes[257].

Prendre comme objet la régression logistique (LR) présente l'intérêt de pouvoir disposer d'une littérature abondante en sciences sociales, concernant tant les fondements de cette approche que ses limites dans l'appréhension de phénomènes sociaux. Encore souvent utilisée aujourd'hui comme technique d'analyse en SHS, mais aussi vivement remise en cause à travers la « querelle des méthodes » du siècle dernier et le « terrorisme quantitatif »[258] d'après-guerre, la LR offre un point d'appui intéressant à partir duquel il est possible de mettre en relation les outils de ML plus récents.

En effet, la démarche de recherche impliquant le recours à ce type de procédure se révèle emblématique de la « sociologie des variables » et de la démarche expérimentale « toutes choses égales par ailleurs » (en latin, « ceteris paribus »), « inhérent aux méthodes de régression multiple, […] antérieur en sociologie quantitative à la méthode elle-même »[259].

3.4.2.1 Vers un éventuel langage commun en sociologie ?

Comme relevé dans l'introduction générale, l'une des interrogations centrales que soulevaient les ambitions de ce projet de recherche était celle de communiquer efficacement certaines idées avec un vocabulaire s'approchant idéalement d'un langage commun. Par conséquent, en dépit du caractère inachevé de l'analyse de discours dans Atlas.ti, plusieurs textes sociologiques ayant très peu ou pas de lien apparent avec les méthodes algorithmiques m'ont semblé opportuns à utiliser pour deux raisons. La première : mettre des mots sur certains questionnements que j'ai encore moi-même parfois de la difficulté à exprimer au moyen de

---

[257] BOELAERT Julien et OLLION Étienne, « The Great Regression Machine Learning, Econometrics, and the Future of Quantitative Social Sciences », *op. cit.* ; DESROSIÈRES Alain, « Analyse des données et sciences humaines : comment cartographier le monde social ? », *op. cit.*

[258] VAN CAMPENHOUDT Luc, MARQUET Jacques et QUIVY Raymond, *Manuel de recherche en sciences sociales*, *op. cit.*, p. 296. Voir aussi JAVEAU, Claude, « Le terrorisme du nombre », *Revue de l'Institut de Sociologie,* 4, 1976, p. 371-383.

[259] DEAUVIEAU Jérôme, « Comment traduire sous forme de probabilités les résultats d'une modélisation logit ? », *Bulletin of Sociological Methodology*, , nº 105, 2010, p. 21. Ce que vise le concept de style de raisonnement de Ian Hacking.

symboles langagiers. La seconde : interpeller un tant soit peu la communauté des sociologues dans laquelle espère pouvoir s'inscrire ce mémoire de maîtrise. Comme nous le verrons, plusieurs expressions et formulations tirées de ces discours sociologiques[260] seront ainsi reprises dans le travail de description et d'analyse comme « stratégie » d'écriture.

On pourra sans nul doute me reprocher, avec raison, d'avoir détaché « arbitrairement » de leur contexte certains extraits de façon parfois hautement « réductrice » ou « simplificatrice » de la complexité des sémantiques sociologiques dans lesquels ils sont (et ont été) insérés socialement. Cependant, malgré toutes les lacunes dans ma compréhension limitée et située des notions mobilisées en tant qu'apprentie sociologue, l'objectif ultimement, et beaucoup plus large dans lequel s'inscrit cet exercice est d'essayer de tisser des liens explicités entre générations et « chapelles » de sociologues.

Rappelons en terminant que la comparaison d'ordre sociologique des modèles mathématiques plus classiques et plus contemporains vise à mettre en évidence leurs compatibilités ou leurs incompatibilités à certaines classifications sociales. En ce sens, ce présent travail s'insère dans une analyse sociosémantique des classifications sociales et de leurs élaborations, sachant que, selon les personnes et groupes sociaux, certaines façons de décrire le monde sont privilégiées par des catégories et des raisonnements sociaux[261], par l'entremise de divers dispositifs (matériaux, méthodes, démarches). C'est pourquoi, dans les prochains chapitres de ce mémoire, il s'agira d'étudier comment divers modèles mathématiques (et processus de mise en modèle) résument, distinguent, rassemblent et ordonnent des attributs sociaux généraux (*i.e.* les p « variables » en colonne)[262] et des observations spécifiques correspondantes (*i.e.* les n « individus » ou « unités statistiques » en ligne)[263] pour permettre certaines lectures par quiconque utilisant ce type de procédures.

---

[260] Traces empiriques de la vie sociale « dépersonnalisées » et plutôt « disparates » en regard de notre objet d'étude sociologique.

[261] SABOURIN Paul, « L'analyse de contenu », dans Benoît GAUTHIER (dir.), *Recherche sociale: de la problématique à la collecte des données*, 5e éd., Québec, Presses de l'Université du Québec, 2009, p. 415-444.

[262] Colonne ($j = 1, \dots, p$) où p est le nombre de caractéristiques observées/mesurées sur tous les individus.

[263] Ligne ($i = 1, \dots, n$) où n est la taille de l'échantillon de base.

CHAPITRE VI – ANALYSE ET RÉSULTATS

Les deux derniers chapitres du mémoire présentent le travail d'analyse et les résultats, en y intégrant quelques éléments de « discussion[264] ».

Le quatrième chapitre donne d'abord l'occasion de revenir sur certaines distinctions conceptuelles fondamentales préalables, en décrivant la manière dont Desrosières concevait ou « situait » lui-même le type de démarche qu'impliquait le recours aux méthodes de régression logistique (LR). Par la suite, sont revisités quelques-unes des limites du raisonnement statistique et probabiliste, communément admises aujourd'hui, telles que certaines discussions méthodologiques et épistémologiques « historiques » en sociologie les ont identifiées. Nous pourrons en somme constater comment et sur quels aspects ces critiques « sociologiques » – courantes et datées – ont progressivement contribué à lier toute forme d'abstraction mathématique et de descriptions « totalisantes, universalisantes et réifiantes » aux recherches dites « quantitatives », aux sciences pures et dures selon certaines expressions rituelles.

Finalement, le cinquième chapitre entreprendra d'examiner de manière relationnelle les arbres de décision (DT), puis les forêts aléatoires (RF). D'une part, nous essaierons de voir ensemble comment semblent s'articuler diverses tensions originaires de la raison statistique à travers certains croisements particuliers, « inusités » des axes de la grille conceptuelle. D'autre part, nous interrogerons certaines observations à la lumière des éléments de critique précédemment dégagés (chapitre IV). En cherchant à mieux comprendre comment et en quoi ces deux techniques algorithmiques contemporaines se différencieraient des logiques et des enjeux sociaux « classiques » associées aux méthodes statistiques plus anciennes, propres au raisonnement expérimental, notre tentative d'analyse sociologique des analyses statistiques conduira à « rendre visible » certains aspects de convergence entre formes mathématiques et formes « sociales »/sociologiques.

[264] Ce mot a été mis en guillemet puisque jusqu'à présent, le terme le plus adéquat, quoique moins usuel par convention pour décrire ce type d'entreprise, aurait été celui de « monologue ».

LE RAISONNEMENT EXPÉRIMENTAL « *CETARIS PARIBUS* »

4.1 Lecture en termes de déterminants sociaux : Indicateurs généraux

Concernant l'interprétation des analyses logistiques, référons-nous aux coefficients du modèle de régression parcimonieux (M_3). Le tableau 3 indique les rapports de cotes (ou *odds ratios*, OR)[265].

Tableau 4. – Résultats de la régression logistique (M_2 et M_3)

	Modèle 2		Modèle 3	
	EXP(B)		EXP(B)	
Sexe (réf: Homme)	11,959		15,096	
Femme	1,115	**	1,135	***
Groupe d'âge (réf: 75 ans et plus)	58,789		100,243	
15-24 ans	1,473	***	1,801	***
25-34 ans	1,218	**	1,509	***
35-44 ans	1,041		1,263	**
45-54 ans	1,074		1,262	**
55-64 ans	1,056		1,180	*
65-74 ans	1,036		1,066	
Minorité visible (réf: Pas minorité)	22,32		21,569	
Minorité visible	0,825	***	0,819	***
Plus haut niveau de scolarité atteint (réf. Primaires)	87,177		75,561	
Secondaires	1,081		1,092	*
Collégiales	1,301	***	1,287	***
Universitaires	1,533	***	1,527	***
Situation professionnelle (réf. Emploi/sem.dern.)			22,669	
Pas d'emploi			0,826	***
Classe sociale (réf: Inférieure)	196,483		165,529	
Moyenne	1,372	***	1,285	***
Supérieure	2,114	***	1,989	***
Sentiment d'appartenance à la collectivité (réf: Oui)	231,206		129,226	
Non	0,585	***	0,657	***
Qualité de l'environnement local (réf: Non)			52,159	
Oui			1,3	***
Participation à des activités de plein air (Réf. Non)	92,112		84,369	
Une ou plus	1,444	***	1,432	***
Manger à l'extérieur/acheter des plats à emporter (réf. Quelques fois/mois)			77,371	
Pas dern. Mois			1,053	
Plusieurs fois/sem.			0,734	***
Consommation de boissons alcoolisées (réf: Jamais)			34,998	
Occasion			0,939	
Régulier			1,174	**
Usage actuel du tabac (réf. Non-fumeur)			55,685	
Fumeur			0,712	***
Niveau de stress dans la vie (réf. Pas du tout)			80,182	
Pas tellement			0,853	**
Un peu			0,736	***
Assez/Extrêmement			0,591	***
Incapacité(s) (réf. Non)	687,365		520,742	
Oui	0,368	***	0,409	***
Constante	0,607	***	0,713	**
R deux Nagelkerke	14,0%		16,2%	
*:p<,05; **:p<,01; ***:p<,001				
Source : Statistique Canada, FMGD de l'Enquête sociale générale (ESG) de 2016				

Dans l'ordre, les trois variables les plus importantes sont les incapacités (Wald=520.74, p<0.001), la classe sociale (Wald=165.53, p<0.001) et le sentiment d'appartenance communautaire (Wald=129.23, p<0.001).

Effets homogènes et indépendants ? L'analyse de régression montre par exemple que les personnes insatisfaites de leur appartenance à leur collectivité sont moins susceptibles d'évaluer plus positivement leur santé (β=-0.42; OR=0.66, soit 1.52 fois moins), *toutes choses égales par ailleurs*. Comparées aux personnes de la classe inférieure, celles se décrivant des classes moyennes et supérieures ont une plus grande probabilité de s'estimer être en excellente ou en très bonne santé (de 1.3 à 2 fois plus de chances), lorsqu'on tient en compte de toutes les différences liées aux caractéristiques individuelles (démographiques, socio-économiques, habitudes de vie, etc.) considérées dans le modèle. Les personnes avec incapacité ont plus de la moitié moins de chances d'avoir une santé perçue excellente ou très bonne (0.41, soit 2.4 fois moins), une fois « neutralisés » les « effets » (actions) des autres variables retenues dans l'analyse.

[265] La source pour tous les tableaux et figures présentés aux chapitres IV et V est Statistique Canada.

4.1.1 Langage des variables, de l'« explication » : La question des « actions *possibles* »

Comme l'observait Desrosières, les « sujets des verbes » dans les phrases des comptes-rendus des modèles logistiques sont des variables « discrètes »[266], « découpant exhaustivement l'univers en classes disjointes », « au lieu de refléter des mesures »[267]. Celles-ci sont supposées « agir » de façon uniforme (ou « homogène ») dans « tout l'espace étudié », au sens de produire des « effets […] brouillés par ceux de variables concurrentes » « selon une logique déterministe »[268].

Il convient de souligner ici que la notion d'explication dans la sémantique statistique, dans l'idée du « langage de l'explication tourné vers l'action et le contrôle » (cognitif, intersubjectif) suggérée par Desrosières renvoie, dans l'optique épistémologique, aux formes de connaissance « techniques » plutôt que « scientifiques ». Selon Granger, la connaissance « technique » se distingue de celle « scientifique » en ce qu'elle « vise […] la construction de modèles abstraits conduisant à produire des effets » et « s'intéresse aux résultats des procédures »[269]. Par le formalisme asymétrique[270], ce type d'analyse exige de spécifier préalablement un sens précis, une direction aux relations observées/observables empiriquement, bien que rien ne garantisse l'univocité du sens, de la signification *sociale* « des liaisons déterminantes entre variables »[271], entre facteurs et effets.

C'est en cela que dans la littérature méthodologique (sociologique/statistique), les produits (ou résultats) de ces méthodes, en « liant des effets (recherchés) à des causes (actions

---

[266] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 139 et 144.

[267] *Ibid.*, p. 166-167.

[268] *Ibid.*, p. 166. L'indétermination « est enserrée par des urnes probabilistes constantes ou variant selon des lois constantes » (Desrosières, 2008, p. 141).

[269] GRANGER Gilles-Gaston, *Le Probable, le Possible et le Virtuel*, *op. cit.*, p. 234. La connaissance « scientifique » vise quant à elle « la construction de modèles abstraits explicatifs des phénomènes » dans la mesure où « le virtuel […] fonctionne […] comme ce que l'on peut penser » (Granger, 1995 : 234).

[270] Conduisant à distinguer des variables « dépendantes » et « indépendantes » (*Cf.* Introduction générale, p. 15)

[271] Par exemple, le fait que le revenu (i.e. variable dite « explicative ») « explique » l'état de santé (i.e. variable dite « à expliquer »); l'état de santé pourrait tout aussi bien « expliquer » inversement (réciproquement) le niveau de revenu. Comme l'indique Des Nétumières, la notion de « cause » statistique (« constante » dans un schème probabiliste) « ne signifie pas qu'il existe une détermination d'ordre fonctionnelle » (Des Nétumières, 1997 : 279-280) Voir aussi Boudon (1976), Foucart (2001 : 23) et Fouquet (2010).

possibles) »[272], sont généralement vues comme des moyens adaptés à des visées explicatives, prédictives[273], voire prescriptives, « au sens statistique, à des fins d'action »[274]. « Détachée[s] des personnes » concrètes, les entités conventionnelles que constituent les variables sont « attachées à des actions spécifiques »[275]. Pour Desrosières, « la question des effets de certaines variables sur d'autres […] ne trouve sens que dans une perspective d'action et de transformation du monde »[276], d'où la dimension pratique (ou « politique ») de l'explication « statistique », de la connaissance « technique » au sens de Granger.

> Les analyses fondées sur la modélisation [sur logique de la *régression* multiple], en repérant l'effet d'une variable sur une autre, donnent les moyens aux politiques d'agir sur tel ou tel phénomène par le biais de mesures spécifiques. Ainsi, si l'on constate qu'il y a discrimination salariale, l'État peut promulguer une loi visant à l'interdire (Cf. la loi de 1983 « À travail égal, salaire égal ») et veiller à son application[277].

Dans ce qui suit, trois formes de critiques[278] à l'égard de la place du « langage des variables » propre au raisonnement statistique « traditionnel » en sociologie sont examinées. Comme l'avait bien vu Desrosières, ces résistances classiques à la méthodologie des sciences sociales quantitatives accusent de perdre de vue la totalité sociale « d'une personne, d'une situation, d'un sens » plutôt que « celle d'une population, dotée de limites précises, définie comme une catégorie logique, un ensemble d'éléments distincts »[279]. En dénonçant les « grandes totalisations comme des artifices », de « purs artefacts statistiques »[280], ces discours ont progressivement

---

[272] DESROSIÈRES Alain, *Pour une sociologie historique de la quantification*, *op. cit.*, p. 140. « Les méthodes économétriques […] mettent en œuvre des formes grammaticales centrées sur le langage des variables. Le fichier statistique est lu du point de vue de ces dernières, mises en relation par des équations, les unes expliquant les autres selon un modèle comportant une partie déterministe et un résidu aléatoire non expliqué […] Dans l'analyse économétrique, des facteurs agissent, influencent d'autres variables, dont les valeurs résultent de ces déterminations. » *Ibid.*, p. 138.

[273] ROUANET Henry et Frédéric LEBARON, « La preuve statistique : examen critique de la régression ».

[274] DES NÉTUMIÈRES Félicité, « Méthodes de régression et analyse factorielle », *Histoire & Mesure*, vol. 12, nº 3, 1997, p. 297.

[275] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 63 et 108.

[276] *Ibid.*, p. 167.

[277] DES NÉTUMIÈRES Félicité, « Méthodes de régression et analyse factorielle », *op. cit.*, p. 283.

[278] Plus exactement, trois dimensions des critiques : cognitives (Passeron), pratiques (Bertaux) et techniques (Degenne)

[279] DESROSIÈRES Alain, *Pour une sociologie historique de la quantification*, *op. cit.*, p. 144-146. Aux yeux des tenants du mode de connaissance de type monographique, « non seulement le codage aplatit, fait perdre une partie des observations, mais en plus il fractionne, [mutile, tronçonne, réduit] isole, selon des critères, des aspects de situations, de personnes, de groupes, qui doivent être vus comme des totalités, perçues et décrites globalement » (Desrosières, 2008 : 144).

[280] BECKER Howard, *Faire preuve*, *op. cit.*, p. 74.

contribué à « cantonn[er] […] l'argument statistique du côté du général et de l'universel, caractéristiques des sciences de la nature »[281], dites « exactes »[282].

[281] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 102.

[282] DEGENNE Alain, « Une méthodologie « douce » en sociologie », *L'Année sociologique (1940/1948-)*, vol. 31, nº 3, 1981, p. 107.

4.2 Quelles limites ? Catégorisations (standardisées et universelles)

Considérée comme « extension et systématisation de l'idée d'élimination des effets de structure »[283], la régression logistique (LR) hérite en quelque sorte des critiques formulées depuis les années 1920 par les sociologues durkheimiens Maurice Halbwachs et son maître François Simiand[284]. Comme le rappellent plusieurs sociologues, méthodologues et épistémologues lorsqu'ils invoquent le « problème du renne au Sahara et du chameau au Pôle Nord, cette technique est en effet considérée, plus que d'autres, comme susceptible de produire des artefacts, c'est-à-dire des résultats dénués de toute signification sociologique »[285], des « abstraction[s] trop soigneusement détachée[s] de la réalité pour nous apprendre quoi que ce soit sur le réel »[286].

Plus précisément, la recherche « d'effets (relativement) purs » (propres ou « nets »[287]) à travers le « contrôle » des variables conduirait à créer ce qu'Halbwachs appelait des « populations fictives », des « univers factice[s] tout à fait éloigné[s] des configurations réelles du monde social »[288], « où la probabilité de figurer est très inégale socialement, les groupes sociaux ayant été inégalement "essorés" »[289]. Pour reprendre l'exemple, le nombre de cas/situations imaginables est de douze pour le modèle simple avec trois VI (M_1), dépasse les cinq mille pour le modèle réduit avec huit VI (M_2) et atteint près d'un million combinaisons possibles pour le modèle parcimonieux avec quatorze variables (M_3)[290]. Par conséquent, si un modèle de régression permet d'envisager *théoriquement* (logiquement, cognitivement) autant de configurations (« espaces ») de relations singulières qu'il y a de croisements possibles entres les variables et leurs diverses modalités,

---

[283] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 23 et 165.

[284] Voir MARTIN Olivier, « Raison statistique et raison sociologique chez Maurice Halbwachs », *Revue d'histoire des sciences humaines*, vol. 1, nº 1, 1999, p. p.69-101.

[285] VALLET Louis-André, « Sur l'analyse de régression en sociologie », Communication au RT20 (Méthodes) au congrès de l'Association Française de Sociologie, Bordeaux, 5-8 septembre 2006, p. 1-13.

[286] HALBWACHS Maurice, « La statistique en sociologie », *La statistique, ses applications, les problèmes qu'elle soulève*, Paris, Presses Universitaires de France, 1935, p. 122-123.

[287] Je remercie Stéphane Moulin d'avoir attiré mon attention sur l'historicité sociale ou la périodicité des termes utilisés.

[288] DES NÉTUMIÈRES Félicité, « Méthodes de régression et analyse factorielle », *op. cit.*, p. 281.

[289] SINGLY François de, « Les bons usages de la statistique dans la recherche sociologique », *op. cit.*, p. 20.

[290] M_1=3 (PSC)*2 (DOS)*2 (INC)=12. M_2=2 (SEX)*7 (AGE)*2 (MIN)*4 (EDU)*3 (PSC)* 2 (DOS) *2 (ODA) *2 (INC)=5 376. M_3=2 (SEX)*7 (AGE)*2 (MIN)*4 (EDU)*2 (TRA)*3 (PSC)*2 (DOS)*2 (DOS)*2 (ODA)*3 (ALI)*3 (DRR)*2 (SMK)*5 (SMG)*2 (INC)=967 680 (*Cf.* annexe).

certaines sont très rares ou n'existent pas *empiriquement* (actuellement, historiquement) au sein des groupements humains étudiés, « même sur de grands effectifs »[291].

4.2.1 Décomposition et configurations historiques improbables : Formes « artificielles »

Dans ses travaux d'épistémologie de la sociologie, Jean-Claude Passeron (1991) a en quelque sorte contribué à systématiser les remarques d'Halbwachs et le « paradoxe de Simiand ». Selon lui, raisonner en termes « d'effets *statistiquement* purs » (propres ou spécifiques) sur la base de « critères stratificateurs », consiste en un processus sans fin de « décomposition des interactions »[292] entre variables.

Passeron indique que l'expérimentation statistique en sociologie, ou le fait de recourir au « langage des variables dans l'analyse des variations sociales », « doit, pour pouvoir énoncer univoquement la différence des effets, supposer l'identité de sens d'une valeur donnée de chaque variable, quels que soient les croisements entre elles »[293]. Or, « naturalisées », « conceptualisées hors de toute indication du contexte pertinent »[294], c'est-à-dire « sous réserve de la constance ou du contrôle du contexte »[295], les variables « deviennent trans-historiques »[296]. Il s'ensuit que, chez Passeron, le problème que pose cette démarche réside dans la « présence inégale des divers groupes au sein de la population […] [qui] substantialise […] l'action des variables en figeant administrativement le sens des catégories »[297]. C'est en ce sens que pour cet épistémologue des sciences sociales, le travail sociologique serait voué à osciller entre une double exigence méthodologique – à la fois expérimentale (logique) de la comparaison et empirique (historique) de la description – pour parvenir à échapper à la double « illusion » nomologique et herméneutique[298].

---

[291] LEMERCIER Claire et Claire ZALC, « Des corrélations aux causalités ? », *Méthodes quantitatives pour l'historien*, La Découverte, 2008, p. 72.

[292] PASSERON Jean Claude, *Le Raisonnement sociologique: Un espace non poppérien de l'argumentation*, Paris, Albin Michel, 2006, p. 222. Voir aussi Desrosières pour les techniques « privilégiant les interactions entre variables […] dérivées de l'analyse de variance » (2008 : 151).

[293] *Ibid.*, p. 224.

[294] *Ibid.*, p. 219.

[295] *Ibid.*, p. 153.

[296] *Ibid.*, p. 166.

[297] *Ibid.*, p. 219. (ou l'improbabilité d'apparition, p. 229)

[298] En particulier les chapitres, « Histoire et sociologie. Identité sociale et identité logique d'une discipline » (p. 125-168) et « Ce que dit un tableau et ce qu'on en dit: Le langage des variables et l'interprétation dans les sciences sociales » (p. 199-232).

Pour illustrer les aboutissants du raisonnement faisant *comme si* toutes les choses étaient égales, certains auteurs évoquent ainsi des « croisements improbables », « absurdes, exceptionnels ou impossibles », des « cas vides », par exemple l'idée d'observer des « retraités de 30 ans[299] » ou des « jeunes noirs » très scolarisés issus de familles aisées dans les années 50 aux États-Unis[300]. D'autres affirment dans la même voie que la procédure fait « comme si jeunes et vieux avaient autant de chances d'avoir des [problèmes de santé] anciennes »[301] ou qu'elle revient à se demander si le salaire des femmes serait le même que celui des hommes, si restant femmes, elles avaient les caractéristiques identiques à celles des hommes ou vivaient exactement dans les mêmes conditions[302].

Par le recours à des formules telles que « toutes choses *inégales* par ailleurs », il s'agit dès lors d'insister sur l'hétérogénéité du social, « l'inégale probabilité des cooccurences, constitutive de l'objet *concret* »[303], sur le fait que « les attributs sociaux sont regroupés selon des configurations récurrentes »[304], que « la condition « toutes choses égales par ailleurs » ne peut jamais être totalement réalisée, parfois même par contradiction interne »[305]. Une comparaison valide « ne peut être réalisée qu'en tenant compte explicitement des différences réelles, d'échelle, de situation, d'histoire »[306], ce qui « impose la connaissance des configurations réelles comme systèmes [à chaque fois] singuliers et non-reproductibles de cooccurrences de propriétés »[307].

---

[299] RIANDEY Benoît, Laurent TOULEMON et Jacqueline FELDMAN, « L'utilisation de la régression logistique dans les enquêtes », *BMS*, nº 33, 1991, p. 83.

[300] LEMERCIER Claire et Claire ZALC, « Des corrélations aux causalités ? », *op. cit.*, p. 73.

[301] BILAND Émilie, Jean-Sébastien EIDELIMAN et Séverine GOJARD, « Ceteris (non) paribus ? », *op. cit.*, p. 44. « Ce procédé revient, de fait, à invisibiliser la possible variabilité de l'effet du SSE selon qu'il affecte la santé des hommes ou celle des femmes. Pour le dire en termes plus sociologiques, ce procédé tend à réifier, essentialiser les statuts (le SSE ou tout autre statut ainsi « neutralisé »), en présentant leurs effets sur la santé comme étant fixes, indépendants de leur contexte, c'est-à-dire de la combinaison de statuts qu'ils constituent ensemble. » CARDE Estelle, « Les inégalités sociales de santé au prisme de l'intersectionnalité », *Sciences sociales et sante*, vol. 39, nº 1, 2021, p. 18.

[302] DES NÉTUMIÈRES Félicité, « Méthodes de régression et analyse factorielle », *op. cit.* ; ROUANET Henry, Frédéric LEBARON, Viviane LE HAY, Werner ACKERMANN et Brigitte LE ROUX, « Régression et analyse géométrique des données : réflexions et suggestions », *Mathématiques et sciences humaines*, vol. 40, nº 160, 2002, p. 15 ; FOUCART Thierry, « L'interprétation des résultats statistiques », *Mathématiques et Sciences Humaines*, vol. 39, nº 153, 2001, p. 25.

[303] SINGLY François de, « Les bons usages de la statistique dans la recherche sociologique », *op. cit.*, p. 20.

[304] BILAND Émilie, Jean-Sébastien EIDELIMAN et Séverine GOJARD, « Ceteris (non) paribus ? », *op. cit.*, p. 44.

[305] FOUCART Thierry, « L'interprétation des résultats statistiques », *op. cit.*, p. 25.

[306] DURAND Claire, Mélanie DESLAURIERS et Gérard DUHAIME, « Quelles statistiques pour analyser les inégalités ? Le cas des Premières Nations au Québec », *SociologieS*, [En ligne], mai 2012, p. 36-37.

[307] PASSERON Jean Claude, Le raisonnement sociologique, op. cit., p. 228.

4.2.2 Totalisation et positivisme « abstrait » : Contenus « substantiels »

Dans une voie similaire à Passeron, quoique plus radicale, Daniel Bertaux s'est également attaqué au « langage des variables » qu'imposerait l'« idéologie méthodologique néopositiviste » dominante dans la sociologie empirique d'après-guerre (Stouffer, Lazarsfeld). Pour ce pionnier de l'approche biographique en sociologie, cette « pratique » « s'est dissimulée dans les techniques (probabilistes) »[308]. Dans sa critique de la « méthode » elle-même, Bertaux substitue la notion de « mobilité sociale » au « concept de distribution » pour saisir le mouvement réel/historique des rapports sociaux « selon les conditions locales et temporelles »[309], en déclarant ceci :

> aplatir toutes les différences sur ce lit de Procuste qu'est l'idée de « variable », c'est choisir de tourner le dos à l'étude attentive, concrète, des processus réels ['de transmission du statut social, de distribution des êtres humains *dans* et *par* les rapports sociaux']. La forme qui en résulte, relation mathématisable entre variables « universelles » (universelles parce qu'elles se mesurent sur tout individu) n'a de scientifique que l'apparence; la forme est creuse, le contenu en est absent. La question de la *scientificité* d'une approche est une question de *contenu*, et non une question de forme[310]

Selon Bertaux, penser le monde social en termes de « lois universelles », de relations statistiques « entre des propriétés observables au niveau individuel »[311] signifie d'englober « des contenus *absolument* hétérogènes »[312]. Invitant à faire de la « sociologie *concrète* »[313] en réaction aux « présupposés théorico-épistémologiques » de la sociologie empirique, Bertaux entend ainsi rejeter toute conception « probabiliste » du monde, « qui rend[rait] impossible l'étude des structures, bref la réalisation de l'idée fondamentale, matérialiste »[314].

---

[308] BERTAUX Daniel, « Pour sortir de l'ornière néo-positiviste », *Sociologie et sociétés*, vol. 8, nº 2, 1976, p. 122.

[309] *Ibid.*, p. 128.

[310] *Ibid.*, p. 125.

[311] *Ibid.*, p. 120.

[312] *Ibid.*, p. 124-125.

[313] Voir aussi LAHIRE Bernard, « La variation des contextes dans les sciences sociales. Remarques épistémologiques », *Annales. Histoire, Sciences Sociales*, vol. 51, nº 2, 1996, p. 381-407. Selon Lahire, « [i]l ne faut ainsi jamais perdre de vue le fait que ce sont des êtres sociaux *concrets* qui entrent dans des relations d'interdépendance spécifiques et non des variables ou des facteurs qui agissent dans la réalité sociale. » (*Ibid.*, p. 385.) « [L]a recherche de causes ou de facteurs déterminants, qui organise bien souvent l'interprétation de tableaux statistiques, devient de plus en plus […] inadéquate lorsqu'on s'attache à rendre compte finement de cas singuliers limités en nombre toujours plus complexes que les groupes ou les catégories réduits quelques propriétés sociales fondamentales. […] Ainsi, la saisie de configurations […] singulières n'a pas (et ne peut avoir) pour objectif de mettre en évidence des facteurs déterminants ou explicatifs, mais de saisir la manière dont les multiples contraintes simultanées et enchevêtrées orientent les comportements. » (*Ibid.*, p. 388-389.)

[314] BERTAUX Daniel, « Pour sortir de l'ornière néo-positiviste », *op. cit.*, p. 119-120. Voir le dialogue avec l'article de Paul Bernard (1982) dans les numéros de la revue *Sociologie et Sociétés*. BERNARD Paul, « L'insignifiance des « données » », *op. cit.*

4.3.3 Appropriation et interprétation « contextuelle » : Produits « factoriels »

Enfin, dans la lignée de l'Analyse des données dite « à la française », certains sociologues (statisticiens) méthodologues ont également cherché à se distinguer de l'approche quantitative dominante en Amérique du Nord, de la « méthodologie expérimentale »[315] et notamment de la modélisation dite « économétrique » qui repose sur des hypothèses probabilistes « techniques »[316], « contraignantes » et jugées « arbitraires » du point de vue de la sociologie[317].

En inscrivant les analyses de régression et « la comparaison de type expérimental » dans ce qu'il nommait le « raisonnement de type factoriel », Alain Degenne lui opposait l'idée (et l'application) d'un « raisonnement contextuel », comme « attitude » qui serait « beaucoup plus proche de la tradition et de l'intention sociologique »[318]. Selon lui, l'un des problèmes majeurs de l'approche factorielle serait la conception de l'individu, le « principe de passivité » du sujet statistique :

> Ce raisonnement […] s'appuie sur le présupposé que l'individu est agi [un support d'information passif] et non acteur et qu'il suffit de faire varier certains traits caractéristiques pour obtenir un effet. L'autonomie de l'individu est supposée négligeable parce que ses effets sont supposés aléatoires […] En plus […] il suppose donc que l'on peut isoler l'effet de certains facteurs des autres et d'une manière générale que ces effets sont largement indépendants et plus ou moins additifs[319]

Dans son texte, Degenne se rapproche ainsi de la critique de C. Wright Mills (1967[1959]) à propos de la standardisation qu'induit « l'enquête de type statistique »[320]. Dans ce type de critique plus « technique » (ou *interne* à la méthodologie statistique), il ne s'agirait non pas d'abandonner de manière catégorique les « techniques classiques de la statistique mathématique » comme celles

---

[315] ROUANET Henry, Frédéric LEBARON, Viviane LE HAY, Werner ACKERMANN et Brigitte LE ROUX, « Régression et analyse géométrique des données », *op. cit.*

[316] LEBART Ludovic, Marie PIRON et Alain MORINEAU, Statistique exploratoire multidimensionnelle, op. cit.

[317] DEGENNE Alain, « Une méthodologie « douce » en sociologie », *op. cit.*, p. 113.

[318] *Ibid.*, p. 107.

[319] *Ibid.*, p. 106.

[320] « L'individu statistique n'est pas un cas. Il cesse même d'être une personne sociale pour devenir un objet abstrait, un pur produit de l'intellect. Sa valeur pour l'analyste n'est que celle que lui confère le système conceptuel sur lequel est fondée l'enquête. […] Utiliser la statistique, c'est donc réduire l'individu social à un ensemble de traits qui sont sa seule description dans ce langage et qui représentent un appauvrissement volontaire de son « caractère » social. » (Degenne, 1981 :103)

de régression, mais d'en faire un « usage mesuré »[321], « croisé »[322], bref « raisonné »[323] en sociologie.

Souhaitant laisser la « parole » aux groupes sociaux pour reprendre la terminologie de la grille de lecture, l'enjeu consisterait à « *donner un sens* à la sociologie des variables [et à la vision individualiste que promeut la régression[324]], en la *replaçant* dans le cadre de l'espace social » au sens littéral ou « propre » de la représentation spatiale[325].

Selon F. Héran toutefois, « les sociologues qui accusent la statistique d'« atomiser » le monde social […] entretiennent encore une vision substantialiste des liens qui unissent l'individu aux groupes ». À ses yeux, « loin d'atomiser les relations sociales, l'enquête statistique offr[irait] [déjà] le moyen privilégié de définir l'identité d'un individu de façon structurale, c'est-à-dire en référence à l'identité des autres et non de façon intrinsèque »[326].

---

[321] BRY Xavier, Nicolas ROBETTE et Olivier ROUEFF, « Un dialogue de sourds dans le théâtre statistique ? Analyse géométrique des données et effets de structure », 2014.

[322] BILAND Émilie, Jean-Sébastien EIDELIMAN et Séverine GOJARD, « Ceteris (non) paribus ? », *op. cit.*, p. 38.

[323] VALLET Louis-André, « Sur l'analyse de régression en sociologie ».

[324] BILAND Émilie, Jean-Sébastien EIDELIMAN et Séverine GOJARD, « Ceteris (non) paribus ? », *op. cit.*, p. 39.

[325] ROUANET Henry et Frédéric LEBARON, « La preuve statistique : examen critique de la régression », *op. cit.*, p. 2, p. 2.

[326] HÉRAN François, « L'assise statistique de la sociologie », *op. cit.*, p. 32.

4.3 Fiction, artefact et vérité/illusion (*reflet*) statistiques

En résumé, dans le schéma d'analyse de la *régression* multiple[327], l'idée de la « localisation sociale » consiste à « déterminer une position à partir d'une somme de traits et d'attributs substantifs et substantialisés tels que les niveaux de scolarité et de revenu »[328], de « statuts dont [l]es individus sont porteurs »[329]. Décrites comme des entités « abstraites », « universelles »[330], « transhistoriques »[331], « fixes, indépendantes les unes des autres »[332], les variables décomposent les individus en « items standardisés » de façon dite « mécanique ». Face à cette « somme imprécise et incomplète de comportements »[333], l'un des aspects également remis en cause est le caractère jugé « arbitraire » du processus de sélection des variables « pertinentes », voire l'absence de règles de méthode précises et claires pour guider la démarche[334].

> Cette condition (toutes choses égales par ailleurs) est finalement une hypothèse définie par le sociologue et vérifiée approximativement, dont les conséquences ne peuvent être que des suppositions émises inévitablement en fonction de la personnalité de leur auteur.[335]

À travers certaines de ces critiques parfois virulentes de la quantification en sciences sociales, on voit dans les descriptions chiffrées et « représentations simplifiées » du monde social que produisent les méthodes statistiques, un caractère « déshumanisant », « froid » et « rigide », « plaqué de l'extérieur » qui fait fi de « la singularité irréductible » des expériences et des appartenances sociales de la personne (effets de « mise en boîte »)[336].

---

[327] DESROSIÈRES Alain, « Masses, individus, moyennes: la statistique sociale au XIXe siècle », *op. cit.*

[328] PARENT Frédéric et SABOURIN Paul, « Ethnographie et théorie de la description », *Cahiers de recherche sociologique*, , nº 61, 2016, p. 120. Les auteurs citent le sociologue québécois Léon Guérin selon lequel « la méthode statistique morcelle la réalité vivante agissante » (Guérin, 1932 : 244).

[329] CARDE Estelle, « Les inégalités sociales de santé au prisme de l'intersectionnalité », *op. cit.*, p. 20.

[330] BERTAUX Daniel, « Pour sortir de l'ornière néo-positiviste », *op. cit.*, p. 125.

[331] PASSERON Jean Claude, Le raisonnement sociologique, op. cit.

[332] CARDE Estelle, « Les inégalités sociales de santé au prisme de l'intersectionnalité », *op. cit.*, p. 18-19.

[333] CARDON Dominique, À quoi rêvent les algorithmes, op. cit.

[334] BLUMER Herbert, « Sociological Analysis and the "Variable" », *American Sociological Review*, vol. 21, nº 6, 1956, p. 683-690.

[335] FOUCART Thierry, « L'interprétation des résultats statistiques », *op. cit.*, p. 25. « La statistique appliquée ne donne qu'une image approximative de la réalité qui nous entoure, beaucoup trop complexe pour être contenue dans une liste de nombres aussi grande soit-elle. La modélisation n'est qu'un outil supplémentaire d'observation, et ne peut représenter un phénomène dans sa globalité » (*Ibid.*, p. 28.).

[336] JAVEAU, Claude, « De l'homme moyen à la moyenne des hommes : l'illusion statistique dans les sciences sociales », dans DE COOREBYTER Vincent (dir.), *Rhétoriques de la science*, Paris : Presses Universitaires de France, 1994, p. 53-67. PHILLIPSON M., « La méthodologie conventionnelle: critique phénoménologique », dans Jean PADIOLEAU (dir.), *L'opinion publique*, De

4.3.1 Entre processus (constant) et état (mouvant)

Dans cette configuration sociale-historique particulière, se posent donc les enjeux « classiques » de la catégorisation statistique[337], « artificiellement universalisante »[338], de « l'homogénéisation et de l'essentialisation des catégories sociales »[339] de l'expérience ou « de différence »[340].

Comme le notait cependant Desrosières, ce type de critique de la « réduction statistique » porte davantage sur les procédures (ou « l'espace cognitif ») de « qualification » que celles de « quantification » proprement dites à l'intérieur de la démarche statistique, de la recherche empirique en sociologie[341]. Sur cette lancée, rappelons également que sous l'optique épistémologique contemporaine, ces opérations de « réduction » n'ont nullement une connotation « péjorative », en « rel[evant] […] d'un "processus de sélection et de focalisation" (de Sardan, 2003, p. 13-39) [permettant] d'avoir face à l'objet un "contact précis et pénétrant" dans l'esprit de l'adage "distinguer pour mieux comprendre" »[342]. Par conséquent, comme règle nécessaire au travail d'objectivation scientifique/sociologique, cette réduction « n'a pas [nécessairement] pour but de le chosifier suivant les propriétés du social que nous avons définies, […], de l'amoindrir ou de le dévaloriser »[343].

Gruyter Mouton, 1981, p. 90 ; MERON Monique, « Statistiques ethniques : tabous et boutades », *Travail, genre et sociétés*, vol. 1, nº 21, 2009, p. 55-68 ; SIMON Patrick, « Les statistiques, les sciences sociales françaises et les rapports sociaux ethniques et de « race » », *Revue française de sociologie*, vol. 49, nº 1, 2008, p. 153-162.

[337] BLUM Alain et Maurizio GRIBAUDI, « Des catégories aux liens individuels : l'analyse statistique de l'espace social », *Annales. Histoire, Sciences Sociales*, vol. 45, nº 6, 1990, p. 1365-1402 ; GOSSELIN Gabriel, « Sociologie, classement et quantification », *Cahiers Internationaux de Sociologie*, vol. 93, 1992, p. 321-337 ; CHENU Alain, « La catégorisation statistique », *Sociétés Contemporaines*, vol. 26, nº 1, 1997, p. 5-9 ; SIRACUSA Jacques, « Justifier et préciser l'interprétation des données statistiques », *BMS*, , nº 118, 2013, p. 60-72 ; MOULIN Stéphane, « La statistique en action », *Sociologie et sociétés*, vol. 43, nº 2, 2011, p. 5-15 ; MOULIN Stéphane, « Classification », dans Frédéric BOUCHARD, Pierre DORAY et Julien PRUD'HOMME (dir.), *Sciences, technologies et sociétés de A è Z*, Presses de l'Université de Montréal, 2015, p. 43-46.

[338] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 295.

[339] PARENT Frédéric et SABOURIN Paul, « Présentation. Les espaces-temps de la production ethnographique », *Cahiers de recherche sociologique*, , nº 61, 2016, p. 8. SABOURIN Paul, « Sociologie, éthique et politique », *op. cit.*, p. 25.

[340] BILGE Sirma, « Théorisations féministes de l'intersectionnalité », *Diogene*, vol. 1, n°225, 2009, p. 73.

[341] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 123-124 ; DESROSIÈRES Alain, La politique des grands nombres, op. cit., p. 301.

[342] HAMEL Jacques, « Décrire, comprendre et expliquer. Réflexions et illustrations en sociologie », *SociologieS*, [En ligne], octobre 2006, p. 11.

[343] SABOURIN Paul, « Sociologie, éthique et politique », *op. cit.*, p. 33.

Dans le prochain chapitre, nous verrons comment ces enjeux de classification sociale et de définition analytique – à la fois théorique et pratique – des espaces d'équivalence se reconfigurent et se renouvellent avec les développements méthodologiques récents.

CHAPITRE V – DES ARBRES AUX FORÊTS ALÉATOIRES
Quelle forme *sociale* de raisonnement ?

> « Comment faire de l'un à partir du multiple ? Comment défaire cette unité pour refaire de la diversité ? Pour quoi faire ? Ces trois questions sont différentes, mais inséparables. Elles parcourent la lente élaboration des outils statistiques d'objectivation du monde social. »
>
> Alain DESROSIÈRES, 1993,
> *La Politique des grands nombres*.

5.1 Lecture en termes de croisements variables : Configurations particulières ?

Par souci de clarté et de synthèse, considérons deux arbres de décision représentés à la figure 1[344] : T_1 et T_2, où T_1 est un sous-arbre de T_2. Leur racine contient 18 483 individus. Parmi eux, 9 295 ont déclaré être en excellente ou très bonne santé (soit 50,29%). Selon ces modèles, la variable discriminant le mieux l'ensemble des répondants au départ, selon la perception qu'ils ont de leur état de santé, est le fait de vivre avec une incapacité. L'arbre T_1 montre que « les personnes ayant une incapacité ont évalué plus négativement leur santé » (n_2) alors que T_2 précise que dans cette sous-population, les personnes de classe supérieure satisfaites de leur appartenance communautaire sont les plus susceptibles d'avoir déclaré être en excellente ou très bonne santé (n_{23}).

Processus (« Effets ») différenciés et *interdépendants* ? Du côté des 13 276 personnes sans incapacité (71,8% de l'échantillon), le sentiment d'appartenance communautaire est sélectionné comme critère séparant le mieux les individus par rapport à leur santé perçue. Quelle que soit la réponse quant à leur appartenance, la classe sociale opère subséquemment le meilleur découpage. Cela suggère donc que le lien entre la classe sociale et l'état de santé diffère (ou varie) en fonction de l'appartenance communautaire. Un seul groupe sur quatre est classé avec un moins bon état de santé : les personnes qui jugent avoir un faible niveau d'appartenance à leur collectivité et ne pas faire partie de la classe supérieure (n_{12}). Encore une fois, l'arbre T_2 affine les contraintes « logiques » des énoncés, en prédisant un meilleur état de santé chez les gens sans incapacité insatisfaits de leur appartenance à leur collectivité s'ils sont de la classe supérieure (n_{13}) et qu'ils ne sont pas des minorités visibles (n_{55}).

---

[344] Ces deux exemples d'arbre ont été estimés sur l'ensemble de l'échantillon (n=18 483). T_1 (cp=0.004) et T_2 (cp=0.002) constituent des sous-modèles de l'arbre maximal, noté T_{max} (cp=0), composé de 6 731 règles d'affectation.

Figure 1. – Modèles d'arbre de classification T_1, T_{max} et T_2

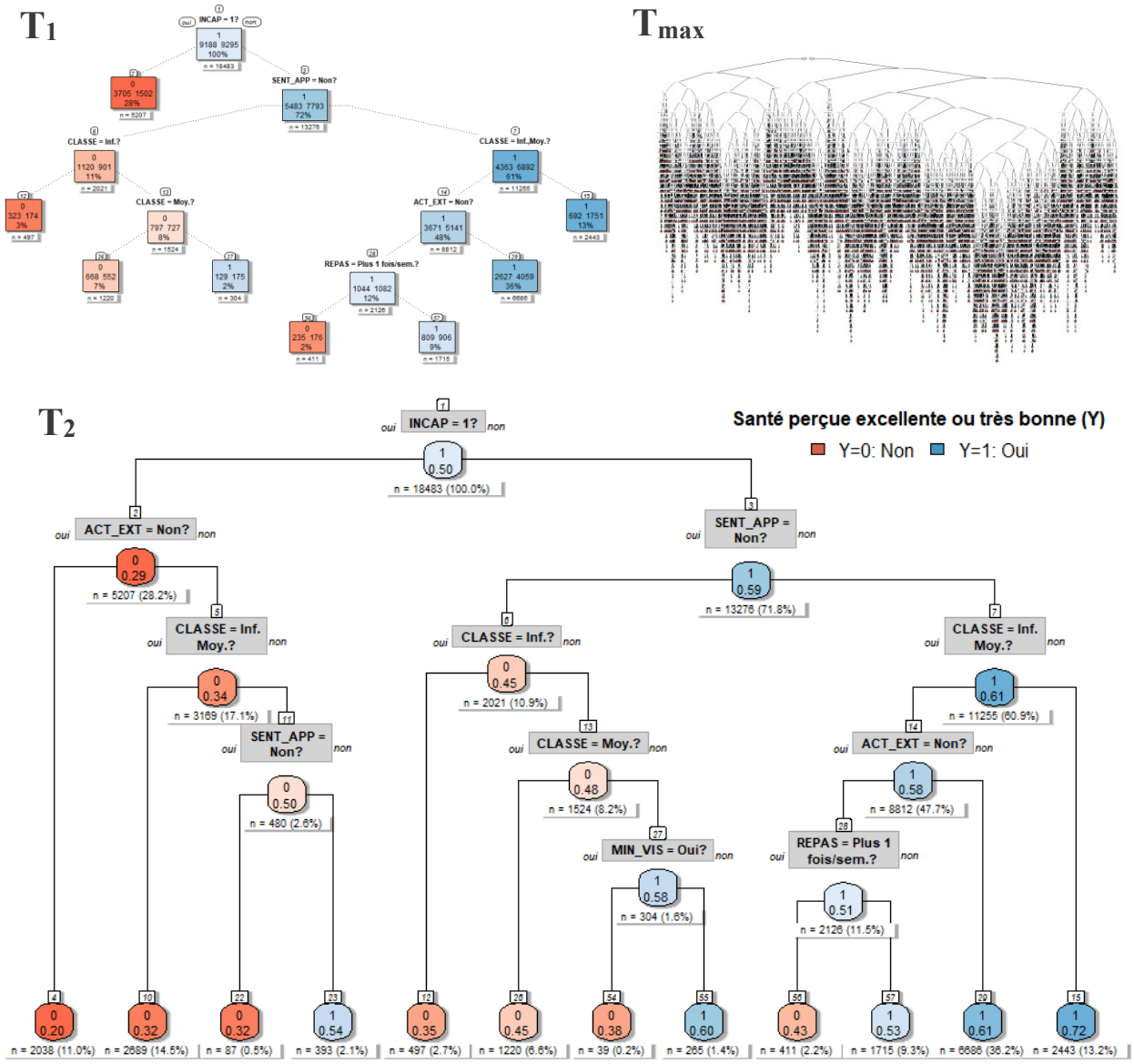


Tableau 5. – Arbre CART composé de douze règles (T_2)

R	No	Y	Prob(Y=1)	Effectif (n)	MIN	PSC	DOS	ODA	FOO	INC	X_p	Cas
1	4	0	20%	2 038	-	-	-	Non	-	1	…	
2	10	0	32%	2 689	-	Inf., Moy.		Oui	-	1	…	Germain
3	22	0	32%	87	-	Sup.	Non	Oui	-	1	…	
4	12	0	35%	497	-	Inf.	Non	-	-	0	…	Lise
5	54	0	38%	39	Oui	Sup.	Non	-	-	1	…	
6	56	0	43%	411	-	Inf., Moy.	Oui	Non	Rég.	0	…	
7	26	0	45%	1 220	-	Moy.	Non	-	-	0	…	
8	57	1	53%	1 715	-	Inf., Moy.	Oui	Non	Jam., Occ.	0	…	
9	23	1	54%	393	-	Sup.	Oui	Oui	-	1	…	Gisèle
10	55	1	60%	265	Non	Sup.	Non	-	-	0	…	
11	29	1	61%	6 686	-	Inf., Moy.	Oui	Oui	-	0	…	Roger
12	15	1	72%	2 443	-	Sup.	Oui	-	-	0	…	

Source: Statistique Canada, FMGD de l'Enquête sociale générale (ESG) de 2016.

5.1.1 Langage des groupes d'individus (singuliers et pluriels), de la « description »

Dans ces exemples, il est possible de constater que chaque personne « est conçue comme appartenant […] à un système de groupes sociaux »[345], « liés entre eux par une communauté probable de comportements »[346]. Chaque individu (observation) est inséré à travers des « réseaux » de relations, d'associations et de « connexions » plus ou moins complexes ou « élagués », selon des enchaînements ordonnés, « séquentiels » d'attributs sociaux. Ces agencements sont dès lors non réductibles à de simples opérations arithmétiques[347] tel qu'une « somm[ation] de traits et d'attributs substantifs et substantialisés »[348]. Moins *indépendants* des autres qu'*interdépendants* (mutuels, conjoints), les « effets » des variables, ou les « facteurs de variation », de « différenciation entre les individus »[349], « se conjuguent » diversement[350], dans une « perspective holiste [ou intégrée] de reconstitution de la globalité d'une personne, d'un groupe ou d'une situation »[351], caractéristique du mode de connaissance monographique.

Savoirs « socialement situés » ? Avec les méthodes d'induction par arbre comme l'algorithme CART, les êtres sociaux sont donc moins définissables par des appartenances catégorielles/personnelles en tant que telle (d'un point de vue *global*), pouvant être isolés des autres comme dans les équations de régression, mais plutôt par les liens qu'elles entretiennent entre elles (d'un point de vue *local*). À la lumière de notre grille de lecture, les groupes d'individus, caractérisés par des « constellations singulières de propriétés co-occurrentes » observées[352], variées et multiples, constituent donc les principaux « sujets des verbes » de l'analyse par segmentation, faisant l'objet des « commentaires écrits » et comptes-rendus.

---

[345] DEGENNE Alain, « Une méthodologie « douce » en sociologie », *op. cit.*, p. 107.

[346] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 139.

[347] BILGE Sirma, « De l'analogie à l'articulation : théoriser la différenciation sociale et l'inégalité complexe », *L'Homme la Societe*, vol. 2, n°176-177, 2010, p. 43-64.

[348] PARENT Frédéric et Paul SABOURIN, « Ethnographie et théorie de la description », *op. cit.*, p. 120.

[349] DURAND Claire, Mélanie DESLAURIERS et Gérard DUHAIME, « Quelles statistiques pour analyser les inégalités ? », *op. cit.*

[350] BRY Xavier, Nicolas ROBETTE et Olivier ROUEFF, « Un dialogue de sourds dans le théâtre statistique ? », *op. cit.*

[351] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 139.

[352] PASSERON Jean Claude, *Le raisonnement sociologique*, *op. cit.*, p. 222.

5.2 Quels enjeux ? Pensée relationnelle ou resubstantification du relationnel ?

5.2.1 Flexibilité interprétative (sémantique) : Production de réalités incomparables ?

Comme « technique graphique fondée sur les propriétés topologiques de l'espace », ce type de procédure « trouve […] des supports pour d'autres types de généralisation que celles qui s'appuient sur des catégories agrégeant des choses supposées équivalentes » (section 5.4)[353]. Plus précisément, en intervenant au moment du « codage[354] », la méthode des DT « engendre littéralement de nouvelles formules d'équivalence »[355], qui permettent de prendre en considération la « composition sociale » et d'appréhender la « diversité *interne* des catégories » préexistantes, les « interdépendances réelles » et « inégales »[356] des « co-occurrences entre propriétés »[357] décrivant l'espace de relations des personnes, « en amont de la catégorisation »[358]. À la différence des modèles paramétriques classiques (régression linéaire, logistique), les ensembles et sous-ensembles sociaux que les variables distinguent ont ainsi une existence relativement autonome par rapport aux définitions institutionnelles/intellectuelles préalables[359].

Par contraste au postulat d'univocité (et d'universalité) de sens[360] qu'impliquent les formes d'analyses multivariées, la signification de chaque valeur des variables n'est pas maintenue virtuellement constante[361], en fonction des divers classements et regroupements opérés. Puisque la « localisation sociale » se présente plutôt comme « *un espace de rapport entre des relations sociales*, dont certaines […] sont inégales dans leur intensité »[362], ces outils d'analyse apparaissent

---

[353] DESROSIÈRES Alain, *Pour une sociologie historique de la quantification*, *op. cit.*, p. 124. Voir « Les contraintes de la feuille de papier » dans le chapitre 7 « Classer et mesurer ».

[354] L'opération de « codage » (social, administratif) est défini par Desrosières comme étant « le fait (ou la décision) d'attribuer des cas singuliers à des classes » (Desrosières, 1993 : 16 et 290), le « travail concret d'affectation d'un cas à une classe » (Desrosières, 2008 : 132 et 158), « l'inscription des enregistrements (*individuels*) dans une grande variété de systèmes de classes d'équivalence » (Desrosières, 1988 : 62), faisant partie des « procédures *institutionnelles* d'identification des objets » (Desrosières, 1993 :406).

[355] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 169.

[356] BILGE Sirma, « De l'analogie à l'articulation », *op. cit.*, p. 62.

[357] PASSERON Jean Claude, Le raisonnement sociologique, op. cit., p. 227.

[358] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 124.

[359] *Ibid.*, p. 169.

[360] Postulat de stabilité (ou de fixité) des unités de référence ? *Cf.* ERMAKOFF Ivan, « La causalité linéaire. Avatars et critiques », *Andrew Abbott et l'héritage de l'école de Chicago*, 2016, p. 397-417.

[361] « la condition (Ai) n'a pas de signification à elle seule, indépendamment des autres » (Tufféry, 2017 :673).

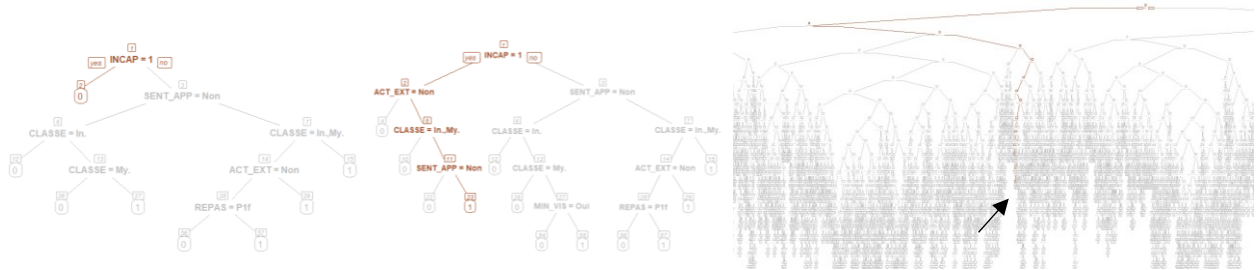
[362] PARENT Frédéric et Paul SABOURIN, « Ethnographie et théorie de la description », *op. cit.*, p. 120.

ainsi adaptés à l'idée de la détermination « contextuelle », circonstancielle des conduites des acteurs sociaux que défendait Degenne (1981). Chaque modèle d'arbre permet de formuler un nombre précis et limité d'énoncés, sans détacher « artificiellement » – par des lois de probabilité – des représentations « abstraites » les êtres sociaux « concrets ».

Comme on le voit à travers l'augmentation de la taille des modèles (de T_1 à T_2 vers T_{max}), cette variabilité sémantique des attributs (ou le caractère polysémique des référents) a pour corollaire l'incommensurabilité des classes d'équivalence. En d'autres termes, plus l'arbre est « dense », « complexe » ou développé, plus de personnes et de groupes sociaux sont représentés sans commune figure ou mesure « générale » susceptible de les « faire tenir ensemble ». Chaque règle donne forme à des sous-groupes « *complètement* différents »[363], des réalités/unités *substantielles* irréductibles (singulières et hétérogènes).

Par la prise en compte systématisée, « naturelle » des interactions dans les données, qui était l'une des limites des méthodes classiques, c'est donc le caractère proprement « relationnel » du social qui est « resubstantifié » ou réifié. Il ne s'agit plus de raisonner *par rapport* « à une situation de référence donnée »[364], « sorte d'individu-type », pour observer des « variations de probabilité »[365] et « estimer matériellement »[366] des « différences de moyenne »[367], comme le font les outils de régression multiple.

Figure 2. – Règles d'affectation de Gisèle selon différents arbres (T_1, T_2 et T_{max})



[363] BERTAUX Daniel, « Pour sortir de l'ornière néo-positiviste », *op. cit.*, p. 124-125.

[364] CIBOIS Philippe, « Modèle Linéaire contre modèle logistique en régression sur données qualitatives », *op. cit.*, p. 20.

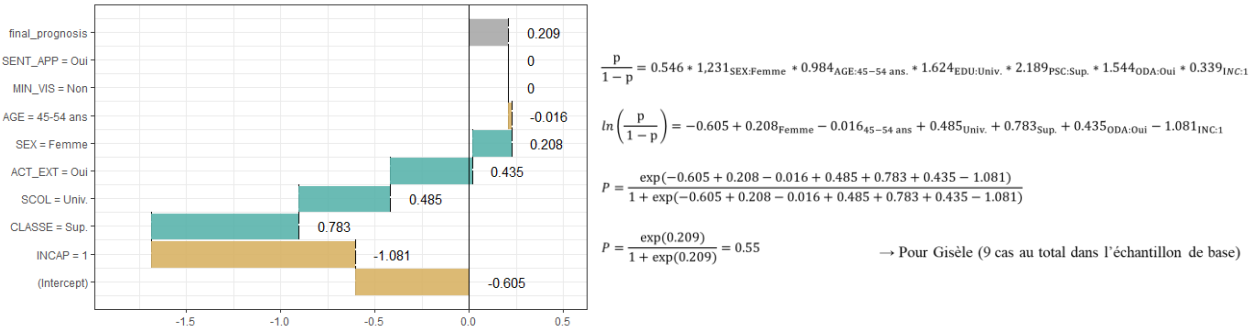
[365] DES NÉTUMIÈRES Félicité, « Méthodes de régression et analyse factorielle », *op. cit.*, p. 293.

[366] RIANDEY Benoît, Laurent TOULEMON et Jacqueline FELDMAN, « L'utilisation de la régression logistique dans les enquêtes », *op. cit.*, p. 83.

[367] DURAND Claire, Mélanie DESLAURIERS et Gérard DUHAIME, « Quelles statistiques pour analyser les inégalités ? », *op. cit.*, p. 7.

Dans l'arbre le plus simple T_1, Gisèle est affectée à la deuxième règle en vertu de laquelle il est très rare de s'être déclaré être en excellente ou en très bonne santé (28,2%)[368]. Outre l'incapacité que Gisèle a déclarée, la classification ne s'appuie sur aucun autre de ses attributs, tels que ses études universitaires ou son fort sentiment d'appartenance à sa collectivité qui contribuaient à faire augmenter sa probabilité dans les équations logistiques, par les effets cumulatifs postulés.

Figure 3. – Équation de régression (M_2) pour Gisèle



En considérant presque toutes les caractéristiques de Gisèle comme dans les analyses de régression multiple (M_3), l'arbre maximal (T_{max}) prédit que Gisèle évalue positivement sa santé (100%), mais celle-ci se retrouve finalement la seule représentée/représentante du groupe auquel s'applique la règle n°11 847[369]. Bien sûr, l'objectif de la méthode CART n'est pas de construire un arbre « le plus grand possible » à l'aide de l'entièreté des données d'apprentissage, mais de trouver le niveau d'embranchement « optimal » par validation croisée, en utilisant un échantillon « test » (section 1.3.2)[370]. Cet exemple d'arbre sera néanmoins utile ultérieurement dans le texte dans la mesure où la forme non élaguée de ce modèle sert d'unité de référence élémentaire à la méthode des forêts (section 5.3).

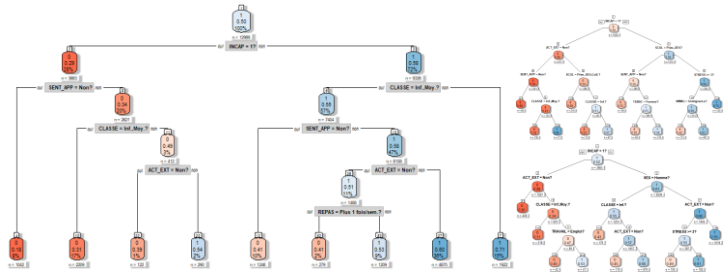
---

[368] Instabilité de ce type de procédure, en raison notamment des premières divisions (Gisèle *p. ex*. possédait pourtant tous les attributs de la règle n°15 : COM=Oui & PSC=Sup). *Cf.* Tufféry (2017 : 673).

[369] Nœud no 11 847 : INC=1 & ODA=Oui & PSC=Sup & DOS=Oui & REV>=1.5 & AGE=>=3.5 & MIN=Non & SMG>=2.5 & GEO=Centre & FOO=>1 fois/sem. & DDR=Jam./Rég. & ENV=Oui & EDU=Univ. & MAT=Marié & SEX=Femme.

[370] Voir Gueguen et Nakache (1988 : 25 et 33). Comme le montre la figure 4, des résultats différents (exprimés sous forme de combinaisons « booléennes » d'attributs) sont obtenus en fonction des parties (ou proportions) de l'échantillon total considéré. Dans l'arbre construit sur un sous-échantillon aléatoire de 70% de l'ESG2016, le fait d'avoir considéré la classe sociale pour distinguer les personnes sans incapacité conduit le modèle à affecter une santé positive aux personnes comme Dora (avec 70% par comparaison aux 38% de l'arbre T2). Voir les limites sur « l'aspect séquentiel » (Lebart et *al*., 1995 : 317; Tufféry, 2017 : 673).

Figure 4. – Trois arbres construits sur différentes parties de l'échantillon de base[371]



5.2.3 Différenciation des sous-populations par arbre décisionnel

En termes épistémologiques, les paramètres de ce type d'algorithmes permettent simplement de moduler ce que l'on pourrait appeler le niveau de « différenciation sociale »[372] autorisé dans la construction *technique* de modèles abstraits, ou encore la longueur des énoncés descriptifs, « classificatoires » écrits en langage naturel. À mon sens, l'enjeu de paramétrage typique en ML[373] pourrait se poser comme suit : jusqu'à quel point tenir compte de la spécificité de nos observations « de terrain » dans l'élaboration de nos constructions sociologiques à valeur générale, sans compromettre les autres éléments observables qui en sont exclus ? Comment éviter à la fois les risques de sur- et de sous- interprétation des données ?[374]

De ce point de vue, la théorie statistique de l'apprentissage renvoie à une question loin d'être nouvelle, celle de la généralisation des connaissances produites sous forme de modèles de « relations entre variables *observées* dans un contexte historique et culturel précis »[375]. Or, la manière d'y parvenir est toute autre, puisqu'il s'agit de « théoriser, formaliser[376], généraliser » des

---

[371] *À droite* : T_3 (n=70%). *À gauche* (de haut en bas) : T_4 (n=20%) et T_5 (n=10%).

[372] DEGENNE Alain, « Une méthodologie « douce » en sociologie », *op. cit.*, p. 107.

[373] *Cf.* Compromis biais/variance. LACOURSE Éric, Charles-Édouard GIGUÈRE et Véronique DUPÉRÉ, « Algorithmes d'apprentissage et modèles statistiques: Un exemple de régression logistique régularisée et de validation croisée pour prédire le décrochage scolaire », *op. cit.*

[374] Comme le suggérait déjà Héran, « [l]a statistique est toujours vouée à se voir accusée d'objectiver par excès ou par défaut. » HÉRAN François, « L'assise statistique de la sociologie », *op. cit.*, p. 32.

[375] SINGLY François de, « Les bons usages de la statistique dans la recherche sociologique », *op. cit.*, p. 13. (nous soulignons)

[376] « Formaliser consiste à décrire systématiquement les régularités dans un contenu » (Sabourin, 2009 : 432).

observations empiriques dans une optique « toutes choses *différentes* (ou *inégales*) par ailleurs »[377], sans chercher à « "rendre équivalents le plus de [situations] possibles" »[378] (section 5.4).

Modalité de l'action (échelle collective ou individuelle) ?

Dans ses travaux, Desrosières défendait l'intérêt d'examiner « l'interdépendance entre façon de penser, de gérer et de décrire *statistiquement* »[379]. Ici, puisque certaines variables seulement sont retenues et qu'elles peuvent être utilisées plus d'une fois dans la reconstitution progressive et provisoirement stabilisée de sous-groupes distincts (deux ou trois fois dans le cas de PSC), il devient difficile, sinon presque impossible, d'identifier des « facteurs » déterminants sociaux « globaux » de la santé, des rapports sociaux « déterminants » au niveau de la société dans son ensemble, « valables pour la collectivité entière »[380]. La santé perçue se conçoit plutôt comme étant le *produit* d'une « imbrication complexe [et singulière] de rapports sociaux » « de sexe, de genre et de race », « enchevêtrés », sans cesse différenciables socialement, de manière relationnelle. Par conséquent, les règles de classification d'un modèle d'arbre peuvent-elles favoriser des interventions ciblées ou personnalisées en devenant applicables à certaines populations exclusivement alors que les formes de raisonnement statistique plus anciennes ne permettaient pas l'identification des personnes singulières, de faire des prédictions à l'échelle individuelle, personnelle, « microsociale » (section 5.3.3)[381] ?

Pour reprendre l'exemple de la santé perçue, examinons les proportions des perceptions positives et négatives, l'estimation des probabilités de classe prédite par les modèles d'arbre (ou les « scores »)[382]. Dans les modélisations de type logit (M_2 et M_3), il était possible (logiquement, cognitivement) de soutenir l'idée selon laquelle la participation à une activité de plein air augmente les chances de s'estimer en excellente ou en très bonne santé (ou « le fait d'avoir fait une activité extérieure conduit, en règle générale, une meilleure évaluation de santé ») *toutes choses égales par*

---

[377] TREMBLAY André, « Feu la société globale et les méthodes quantitatives : de nouveaux termes pour un ancien débat? », *Cahiers de recherche sociologique*, , nº 28, 1997, p. 63-88.

[378] LAHIRE Bernard, « La variation des contextes dans les sciences sociales. Remarques épistémologiques », *op. cit.*, p. 405.

[379] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 196.

[380] TUFFÉRY Stéphane, Data mining et statistique décisionnelle, op. cit.

[381] Voir DES NÉTUMIÈRES Félicité, « Méthodes de régression et analyse factorielle », *op. cit.*, p. 280.

[382] *Cf.* Gueguen et Nakache (1988); Tufféry (2017 : 675); Strobl, Malley et Tutz (2015 : 328).

ailleurs, i.e. « conditionnellement aux variables retenues dans le modèle-cadre »[383] (le rapport de cote était de 1.4).

Figure 5. – Interactions PSC, DOS et ODA (Modèle de régression *vs* Arbre de décision)

LR **DT**

En revanche, dans les représentations par arbre, il est seulement possible de formuler des énoncés « localisés socialement » (« indexés »). Dans l'arbre T_1, le fait d'avoir fait une activité extérieure n'est *visible* qu'au sein des répondants sans incapacité, se sentant appartenir à leur collectivité et issus des classes inférieures ou moyennes (n_{14}), conditionnel donc aux valeurs prises par trois variables (INC, DOS et PSC). Dans l'arbre T_2, parmi les 3 169 répondants avec incapacité qui ont fait des activités de plein air, 2 776 d'entre eux sont malgré tout classés avec une santé perçue négative étant donné leur classe sociale inférieure ou moyenne (n_{10}, le cas de Germain), sinon leur insatisfaction à l'égard de leur appartenance communautaire (n_{22}).

Ces arbres tendent donc à suggérer qu'il n'y aurait pas de différence de santé perçue pour le fait d'avoir participé à une activité extérieure *entre* les classes moyennes et inférieures, *si* les personnes font partie de la classe sociale supérieure ou *lorsque* les personnes sont insatisfaites de leur appartenance communautaire, de telle sorte par exemple qu'au sein des gens de la classe supérieure, ce facteur potentiellement bénéfique pour la santé ne « s'additionne » pas « mécaniquement » (nécessairement, dans *tous* les « cas ») à leur statut socio-économique supposément déjà favorable.

5.2.4 Lecture « intersectionnelle », multiple et unique ?

Par « la présomption d'une certaine "liberté" dont jouissent les rapports sociaux pour se reconfigurer et jouer une partition autant inédite qu'éphémère, que leur inspire chacune des

[383] ROUANET Henry, LEBARON Frédéric, LE HAY Viviane, ACKERMANN Werner et LE ROUX Brigitte, « Régression et analyse géométrique des données », *op. cit.*, p. 17-20.

combinaisons qu'ils forment tous ensemble, au gré des circonstances »[384], nous pourrions dès lors dire que les arbres permettent de « quitte[r] le terrain de l'arithmétique »[385], de sortir des approches « pluralistes », « additives » ou « multiplicatives »[386] propres au raisonnement expérimental.

Dans ce type de modèle effectivement, « il parait difficile d'*imaginer* la possibilité » de représenter *abstraitement* la réalité vécue différemment personnellement d'individus présentant *concrètement* certaines caractéristiques communes (qui les *unit* sur certains plans), mais rassemblées *analytiquement* dans des classes distinctes (singulières/uniques). En modélisant à partir des « situations [d'interaction] concrètes »[387] (*effectives*), les arbres évitent d'« hypostasier *préalablement* des croisements de rapports sociaux », « d'affirme[r], […] sans démonstration, que les femmes immigrantes sont doublement discriminées ou minorisées, de par le sexe/genre et leur statut d'immigrant »[388].

À la lumière de ces observations, nous pourrions ainsi nous demander en quoi les principes de construction des arbres – comme règles d'élaboration du discours statistique/probabiliste contemporain – seraient incompatibles avec certaines perspectives dites « intersectionnelles ». De manière analogue aux discours militants de l'afro-américaine afro-féministe Kimberlé Crenshaw, la mise en œuvre des méthodes d'arbre permet de « remettre en cause l'idée selon laquelle les femmes [par exemple] formeraient un groupe homogène, désigné par une catégorie universelle « femme » qui pourrait traduire l'oppression qu'elles vivent toutes »[389].

Ce type d'analyse, on l'a vu, permet plutôt de voir comment des « croisements (variables) produisent [et dessinent] des configurations *uniques* »[390](multiples), « intégrées et fluides » selon les cas singuliers considérés, « qui débouchent sur des [évaluations personnelles de santé] différencié[e]s »[391]. Si l'utilisation d'algorithmes d'apprentissage dits « sexistes » ou « racistes »

---

[384] CARDE Estelle, « Les inégalités sociales de santé au prisme de l'intersectionnalité », *op. cit.*, p. 20.

[385] Ibid.

[386] BILGE Sirma, « De l'analogie à l'articulation », *op. cit.*, p. 61.

[387] JAUNAIT Alexandre et CHAUVIN Sébastien, « Représenter l'intersection », *Revue française de science politique*, vol. 62, nº 1, 2012, p. 13.

[388] BILGE Sirma, « De l'analogie à l'articulation », *op. cit.*, p. 45.

[389] CARDE Estelle, « Les inégalités sociales de santé au prisme de l'intersectionnalité », *op. cit.*, p. 7.

[390] BILGE Sirma, « Théorisations féministes de l'intersectionnalité », *op. cit.*, p. 73.

[391] JAUNAIT Alexandre et CHAUVIN Sébastien, « Représenter l'intersection », *op. cit.*, p. 16.

est fermement décriée aujourd'hui dans l'espace public/médiatique pour la reconfiguration des inégalités sociales dans diverses sphères d'activité de la société (chapitre I), force est toutefois de constater que ces outils permettent de repérer des « effets intersectionnels »[392], « contextuels », de produire des connaissances « situées »/« indexées » à des configurations sociales-historiques, en opérant selon des principes également à la mode dernièrement dans la recherche en sciences sociales[393].

Comme nous le verrons dans la prochaine partie de l'analyse, si un modèle d'arbre autorise dans ses principes « la flexibilité de la définition empirique des groupes sociaux, flexibilité reflétant la multiplicité "des systèmes sociaux enchevêtrés" », la construction de forêts « rend possible la » confrontation entre une « pluralité des points de vue » sur le monde[394].

[392] BILGE Sirma, « De l'analogie à l'articulation », *op. cit.*, p. 60-61.

[393] Explorer ces compatibilités dépasse la recherche que nous envisageons ici.

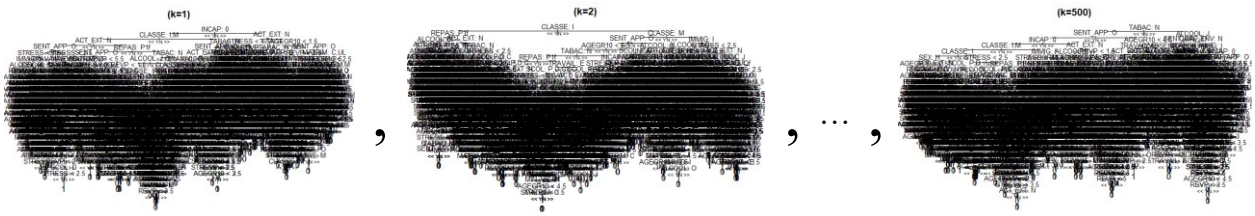
[394] Schweisguth (1982) cité dans SINGLY François de, « Les bons usages de la statistique dans la recherche sociologique », *op. cit.*, p. 17.

5.3 Recomposition « monographique » par forêts : Des récits « en série » ?

> « L'excès de précision, dans la règle de la quantité, correspond très exactement à l'excès du pittoresque, dans celui de la qualité. La précision numérique est souvent une émeute de chiffres, comme le pittoresque est […] une émeute de détails »
>
> Gaston BACHELARD, 1938,
> *La formation de l'esprit scientifique.*

À titre indicatif, le nombre d'arbres par défaut généré par le package « randomForest » dans le logiciel R est de cinq cents, sans contrainte quant à leur profondeur maximale, à la « densité des liens ». Or, il n'est guère difficile d'imaginer l'entreprise colossale que représenterait la mise à jour de toutes les règles de définition ou de composition des sous-groupes sociaux, auxquelles correspondrait chacune des feuilles dans chacun des arbres de la forêt. Selon certaines observations, pour une forêt conventionnelle, il aurait fallu expliciter plus de deux millions de règles différentes, extrêmement détaillées (pensons par exemple à la classe n°11 847 formée de Gisèle uniquement).

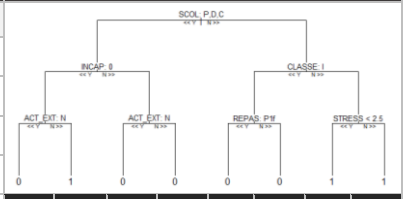
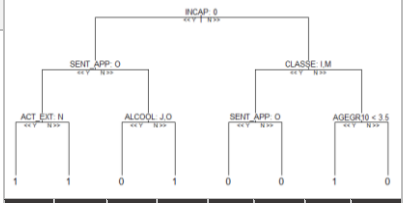
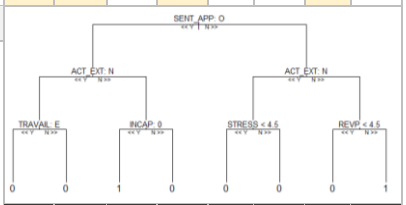
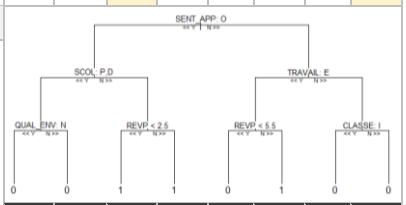
Figure 6. – Trois arbres d'une forêt par défaut (aperçu des deux premiers et du dernier)



En reprenant les règles d'interprétation pour un arbre « simple » (unique), nous avons donc essayé de proposer une version simplifiée de ce que représenterait la présentation des résultats d'une forêt miniature[395]. Le but de cette tentative de lecture des arbres individuels (singuliers) agrégés est surtout de chercher à en dégager la logique d'ensemble, les « formes grammaticales » qu'appellerait la mise en œuvre des méthodes de forêt pour reprendre le vocabulaire de Desrosières.

[395] Cette forêt est composée de cinq arbres simples, comprenant chacun huit feuilles, pour un total de quarante énoncés correspondants (mtry=4, ntree=4, maxnodes=10). Ici, les valeurs prédites de la santé par la forêt sont donc estimées par le classement majoritaire des cinq sous-modèles.

Tableau 6. – Forêt aléatoire simplifiée comportant cinq arbres/quarante règles

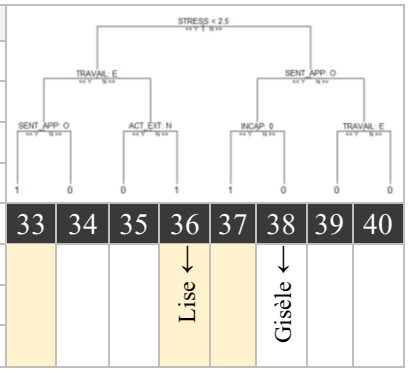
	R	Y	Lecture individuelle (forme « textuelle » ou « tabulaire »)							Graphique
			EDU	INC	ODA	PSC	FOO	SMG		
k=1	01	0	P,S,C	0	Non					
	02	1	P,S,C	0	Oui					
	03	0	P,S,C	1	Non					
	04	0	P,S,C	1	Oui					
	05	0	Univ.			Inf.	1,2			01 02 03 04 05 06 07 08
	06	0	Univ.			Inf.	3			Lise ↓ Gisèle ↓
	07	1	Univ.			Moy.,Sup.		1,2		
	08	1	Univ.			Moy.,Sup.		3,4		
			INC	DOS	ODA	DDR	PSC	AGE		
k=2	09	1	Selon le deuxième arbre, lorsque les personnes n'ont pas d'incapacité et qu'elles sont satisfaites de leur appartenance communautaire, celles-ci sont classées avec une santé positive[396]; sinon, celles-ci doivent avoir consommé régulièrement des boissons alcoolisées pour que le modèle estime qu'elles ont déclaré être en excellente ou très bonne santé. Parmi les gens avec incapacité, dès qu'ils ne sont pas de la classe supérieure, ce modèle prédit une moins bonne santé[397]. S'ils le sont, seulement les plus jeunes (de 44 ans et moins) sont classés avec une évaluation positive de santé.							09 10 11 12 13 14 15 16
	10	1								Lise ↓ Gisèle ↓
	11	0								
	12	1								
	13	0								
	14	0								
	15	1								
	16	0								
			DOS	ODA	TRA	INC	ODA	SMG	REV	
k=3	17	0	Le troisième arbre prédit une santé négative pour les gens satisfaits de leur appartenance communautaire qui n'ont pas fait d'activité extérieure; avoir participé à des activités de plein air est associé à une meilleure perception de santé uniquement lorsqu'aucune incapacité n'est déclarée. Les personnes insatisfaites de leur appartenance communautaire sont classées comme n'ayant pas une santé excellente ou très bonne (quel que soit leur niveau de stress), à l'exception de celles ayant un revenu supérieur à 75 000$ qui ont fait une activité extérieure.							17 18 19 20 21 22 23 24
	18	0								Lise ↓ Gisèle ↓
	19	1								
	20	0								
	21	0								
	22	0								
	23	0								
	24	1								
			DOS	EDU	ENV	REV	TRA	PSC		
k=4	25	0	Dans le quatrième arbre, parmi les personnes satisfaites de leur appartenance communautaire, ne pas avoir fait d'études postsecondaires est associé à une perception de santé plus négative[398]. Chez le groupe insatisfait de leur appartenance communautaire, le modèle classe les personnes sans emploi avec une santé négative (même après avoir pris en compte la classe sociale); advenant que celles-ci aient un emploi, seules celles ayant un revenu supérieur à 120 000$ sont prédites comme ayant déclaré une santé excellente ou très bonne.							25 26 27 28 29 30 31 32
	26	0								Lise ↓ Gisèle ↓
	27	1								
	28	1								
	29	0								
	30	1								
	31	0								
	32	0								

[396] Pas d'« effet » de la participation à une activité de plein air au sein de ce sous-groupe ?

[397] Que tu sois ou non satisfait.e.s de ton appartenance communautaire ?

[398] Dans cette configuration, le revenu ne distingue pas le classement des personnes relatif à leur santé perçue ?

			SMG	TRA	DOS	ODA	INC		
k=5	33	1	1,2	Emploi	Oui				
	34	0	1,2	Emploi	Non				
	35	0	1,2	Pas emp.		Non			
	36	1	1,2	Pas emp.		Oui			
	37	1	3,4		Oui		0		
	38	0	3,4		Oui		1		
	39	0	3,4	Emploi	Non				
	40	0	3,4	Pas emp.	Non				



33	34	35	36	37	38	39	40
			Lise ←		Gisèle ←		

Source: Statistique Canada, FMGD de l'Enquête sociale générale (ESG) de 2016.

Exemples de Gisèle (Obs. no 517) et de Lise (Obs. no 9 213)

En prenant un de nos cas, nous pouvons remarquer que la forêt simplifiée estime une santé perçue négative pour Gisèle (Y=0) dans trois configurations sur cinq. En considérant les études universitaires de Gisèle conjuguées à sa classe sociale (k1), ou le fait qu'elle se sente satisfaite de son appartenance communautaire combiné à son niveau de scolarité (k4), celle-ci reçoit deux cinquièmes des votes en faveur d'une santé excellente ou très bonne (40%). Dans trois sous-modèles en revanche, Gisèle est classée avec une perception « négative » de sa santé (60%) : les descriptions respectives tiennent compte, soit de son incapacité, de sa classe sociale et de son âge (k2), soit de son appartenance à sa collectivité, des activités extérieures réalisées et de son incapacité (k3), sinon de son appartenance communautaire, de son incapacité et de son niveau de stress élevé (k5). Dans le même ordre d'idées, la forêt simplifiée prédit que l'état de santé perçu par Lise (Obs. n°9 213) est « positif » (Y=1), avec quatre « votes » sur cinq (80%).

Dans un cadre moins expérimental et plus réaliste que celui-ci, le raisonnement serait, à ma connaissance, le même : la forêt « standard » (conventionnelle) présentée partiellement à la figure 6 conclut par exemple en attribuant une valeur positive à l'état de santé autoévalué par Gisèle (332/500=66,4%)[399], avec cette fois-ci, 332 sous-modèles qui prédisent une perception de santé excellente ou très bonne (Y=1) contre 168 (Y=0).

[399] Comme observation donnée (no 517 dans la liste), Gisèle fut utilisée pour construire le 199e arbre, lequel classait Gisèle avec une perception de santé négative (35,18%).

5.3.1 Les contraintes pragmatiques de l'intelligibilité « sociale » du social ?

Comme on le voit, en prolongeant la perspective « interactionniste » des arbres, à travers la production de propositions localement situées, les RF permettent de structurer, d'organiser et d'ordonner un système de représentations *virtuelles* également possibles du monde *actuel* « historique » exhaustivement diversifiées. Chaque règle de composition de sous-populations dépend de la permutation aléatoire des « individus » (ou données/observations d'apprentissage) à chaque arbre et des variables (ou attributs) sélectionnés à chaque découpe. Par conséquent, nous pourrions nous demander si des énoncés opposés, vus comme « interprétations socialement situées »[400], pourraient coexister au sein d'une même forêt parmi les centaines, voire les milliers, qui peuvent « voter » pour classer n'importe quel cas singulier (*concrets*)[401]. Autrement dit, les RF permettent-elles d'intégrer une vision pluraliste et conflictuelle de la réalité sociale, en « n'excluant pas [*a priori* du moins] la contradiction » dans les processus de détermination mathématique, probabiliste, contrairement à ce que pensait Bertaux[402] ?

À partir de notre grille d'analyse, comment décrire la façon dont les processus de généralisation (d'argumentation) en ML, qui fondent les performances des RF en particulier, intègrent les tensions conceptuelles, fondamentales et récurrentes de la raison statistique, entre sa double contrainte classificatoire et métrologique notamment ? D'une part, comme nous l'avons vu, par « segmentation progressive », les méthodes des arbres de décision (DT) visent « à suggérer des cohérences globales centrées sur des personnes ou des groupes, et non sur des variables »[403]. C'est à ce titre que ces outils constituent des formes statistiques « narratives », approchant la « richesse sémantique d'une description littéraire »[404], « idéographique »[405]. Sous l'optique épistémologique

---

[400] CARDE Estelle, « Les inégalités sociales de santé au prisme de l'intersectionnalité », *op. cit.*, p. 11. « Si, dans un contexte donné, un axe de domination [ou de différenciation] prévaut sur les autres, on ne peut donc pas en déduire qu'il en serait *nécessairement* de même en d'autres circonstances. » (Carde, 2021 : 8).

[401] Mais en même temps, les règles deviennent tellement précises, spécifiques, allant jusqu'à avoir un seul individu par règle, qu'il serait peu probable que des règles complètement contradictoires soient produites ?

[402] « L'étude de ce processus [de *distribution* des êtres dans], des rapports sociaux-historiques qui le déterminent (et la détermination n'exclut pas la *contradiction* dans la détermination!), remplace […] celle de la « mobilité » ». BERTAUX Daniel, « Pour sortir de l'ornière néo-positiviste », *op. cit.*, p. 123. Il serait intéressant d'approfondir ce que Bertaux entendait par « l'analyse par différenciation des sous-populations ».

[403] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 151.

[404] *Ibid.*, p. 139.

[405] LAHIRE Bernard, « La variation des contextes dans les sciences sociales. Remarques épistémologiques », *op. cit.*

grangérienne, nous pourrions également dire qu'un modèle d'arbre s'apparente davantage à une forme de connaissance pratique, « classificatoire », d'ordre « idéologique »[406] (au sens large) dans la mesure où il s'agit de dégager « un ensemble d'éléments concrets, organisés en un récit, visant à *présenter* la signification des phénomènes »[407].

D'autre part, l'extension de la méthode CART par forêts aléatoires permet quant à elle de « mettre en valeur la cohérence interne [d'une série de] constellation[s] de traits enregistrés pour un cas singulier, traité comme exemple »[408], comme résultat statistique (ou « totalité/cas de type monographique »). Dans les approches par forêt, ce sont les différentes lectures, et plus exactement les sous-modèles de comportement, qui font l'objet de la comparaison statistique, de la construction d'un « espace d'équivalence et de comparabilité » pour établir la représentation *virtuelle* (possible ou non) la plus proche ou « plausible » des phénomènes sociaux[409]. En introduisant de l'aléa dans la construction des modèles, cette approche produit et assemble une multiplicité « de descriptions historiques de la complexité et des dimensions d'un univers social »[410], sous forme d'agrégat plus vaste.

L'interprétation sous contraintes ? Bien que les forêts soient assimilées à de véritables « boites noires » dans la littérature en ML, il semble théoriquement (techniquement) possible de « lire » les cinq cents sous-modèles d'arbre constitutifs d'une forêt sur le plan sémantique. Or, cette tâche est rapidement apparue inefficace ou répétitive pour transmettre les connaissances produites d'un point de vue pratique et cognitif avec uniquement cinq arbres simplifiés. Ce résultat suggère donc que ce n'est pas parce que des « procédés de représentation schématique, d'analyse »[411] peuvent être descriptibles que celles-ci sont nécessairement transmissibles, compréhensibles et intelligibles « socialement » (individuellement et collectivement), d'où l'intérêt peut-être de distinguer la *lisibilité* de l'*interprétabilité* des formes de description du monde social (voir les

---

[406] HOULE Gilles, « L'idéologie : un mode de connaissance », *Sociologie et sociétés*, vol. 11, nº 1, 1979, p. 123-145.

[407] Granger (1967 : 771-772) cité dans PARENT Frédéric et Paul SABOURIN, « Présentation. Les espaces-temps de la production ethnographique », *op. cit.*, p. 18.

[408] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 108.

[409] GRANGER Gilles-Gaston, Le Probable, le Possible et le Virtuel, op. cit.

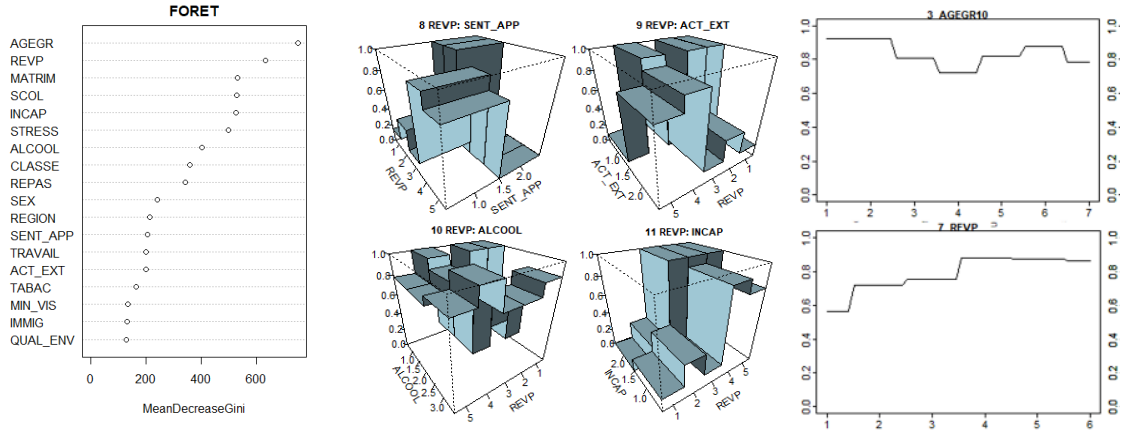
[410] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 170. 259

[411] HALBWACHS Maurice, « La statistique en sociologie », *op. cit.*, p. 123.

débats en sociologie entourant la description plus ou moins « dense », « riche » ou « épaisse » de la réalité, tel qu'entendu depuis l'anthropologue C. Geertz (1973)[412].

D'ailleurs, il est fort intéressant de voir resurgir le fameux « langage des variables » standardisées et des indicateurs globaux (synthétiques) parmi les efforts qui sont faits pour répondre aux controverses scientifiques/publiques à propos de l'opacité algorithmique. Pour pallier les enjeux interprétatifs que posent ces dispositifs, il s'agirait ni plus ni moins de fournir des « mesures », des indices ou des scores permettant de quantifier l'importance relative des variables indépendantes[413]. Comme Desrosières le constatait à propos de la « version cartographique de l'analyse des correspondances », la méthode des forêts « retrouve ainsi la perspective métrologique »[414] du style statistique et probabiliste.

Figure 7. – Importance des variables (avec exemples d'interactions)



Il y a donc une articulation beaucoup plus subtile entre le langage des groupes et celui des variables[415], d'où l'intérêt d'avoir « essayer d'intégrer dans l'analyse le contenu [technique] même des schèmes cognitifs et des savoirs mobilisés par les experts »[416].

---

[412] HAMEL Jacques, « Décrire, comprendre et expliquer. Réflexions et illustrations en sociologie », *op. cit*.

[413] GENUER Robin et Jean-Michel POGGI, « Arbres CART et Forêts aléatoires, Importance et sélection de variables », *op. cit.* Pour des exemples appliqués en sociologie, voir l'article de Boealart et Ollion (2018) qui présente également des techniques/mesures/graphiques de « dépendance partielle ».

[414] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 169.

[415] *Ibid.*, p. 151.

[416] *Ibid.*, p. 59.

5.3.2 Structuration des espaces (*possibles*) de la pensée (*formelle*) et de l'action (*sociale*)

Depuis les critiques épistémologiques (Halbwach, Passeron), il est désormais courant d'entendre dire que les méthodes de régression paramétrique comme celle logistique conduisent à des « non-sens historiques », à « postuler l'invraisemblable »[417]. Or, avec un modèle d'arbre, le nombre de cas de figure distincts (singuliers) qu'il est possible de décrire et/ou de prédire équivaut au nombre de chemins total, de « feuilles ». Pour reprendre l'exemple des arbres T_1 et T_2, ceux-ci peuvent seulement envisager de sept à douze combinaisons, ce qui est nettement inférieur au modèle logit réduit ($M_2 \approx 5\ 000$)[418]. Puisque la méthode CART restreint l'étendue « des combinaisons possibles aux partitions pouvant être dérivées par fractionnement récursif »[419], celle-ci permet donc plus difficilement d'envisager des « improbabilités sociales », des combinaisons éloignées de la « composition réelle » des groupements humains[420].

Concernant l'approche ensembliste, la présente analyse suggère que la construction de modèles de forêt répond également à la préoccupation liée à la création d'entités sociales « fictives », « absurdes » en renforçant « les éventualités [ou situations] les moins invraisemblables »[421], les plus probables actuellement, « mesurées par des fréquences ». Pour le dire plus simplement, en tirant profit de l'instabilité intrinsèque (ou de la « variabilité naturelle ») des arbres isolés qui génèrent un « inventaire exhaustif [de modèles] *possibles* » de la réalité (i.e. *virtualités*), l'algorithme des RF permet de les « hiérarchis[er] et de les sélectionn[er] […] en fonction de leurs probabilités de réalisation »[422]. Cette méthode permet donc de « stabiliser des représentations »[423], en faisant ressortir les configurations sociales « non reproductibles »

---

[417] BILAND Émilie, Jean-Sébastien EIDELIMAN et Séverine GOJARD, « Ceteris (non) paribus ? », *op. cit.*, p. 56.

[418] Avec un peu plus de 6 700 énoncés descriptifs distincts, même l'arbre maximal (T_{max}) comporte dix fois moins de possibilités que le modèle de régression parcimonieux ($M_p \approx 100\ 000$). Voir l'exemple plus détaillé en annexe.

[419] STROBL Carolin, James MALLEY et Gerhard TUTZ, « An Introduction to Recursive Partitioning », *op. cit.*, p. 325. « la plage de combinaisons possibles inclut ici toutes les partitions rectangulaires pouvant être dérivées par fractionnement récursif - y compris plusieurs fractionnements dans la même variable. Cela inclut notamment les règles d'association non linéaires et même non monotones, qu'il n'est pas nécessaire de spécifier à l'avance, mais qui sont déterminées en fonction des données » (Strobl, Malley et Tutz, 2009 : 325). Voir aussi Saporta (2015 : 489-490).

[420] Tous les groupements reconstitués (ou classements opérés) sont observables, « effectifs » dans l'un ou l'autre des sous-échantillons aléatoires : totalités *partielles* « concrètes ».

[421] GRIGNON Claude, « Prédiction et rétrodiction », *Revue européenne des sciences sociales*, vol. 46, n° 142, 2008, p. 85.

[422] Ibid.

[423] GHATTAS Badih, « Prévisions par arbres de classification », *op. cit.*

(singulières)[424] les plus récurrentes dans leur rapport à l'actualité concrète, effective, « historique » [425]. Pour reprendre les termes des philosophes Rouvroy et Berns, en rupture avec les modalités de la gouvernementalité statistique, « [c]ette production de savoir ne vise plus à maitriser l'actuel [...], mais à structurer le possible, à éradiquer le 'virtuel' »[426].

Du point de vue de l'épistémologie contemporaine (*interne* aux sciences), ce principe de méthode s'éloigne toutefois de la conception du *travail* scientifique chez Granger, qui soutenait bien au contraire l'importance dans les constructions intellectuelles à visée scientifique de décrire ce qui est – comme *image* du monde évocatrice – tout en intégrant ce qui pourrait être. Ici, ce sont précisément les états réels *virtuels* les moins probables que « contrôlent statistiquement » les RF par le « principe de compensation probabiliste par la loi des grands nombres »[427], qui permettraient possiblement, éventuellement de formuler des propositions « explicatives », d'ordre théorique en science : « c'est *ce qui n'a pas lieu* qui explique *ce qui a lieu* »[428]. Les modèles produits par les méthodes ensemblistes relèvent ainsi toujours de formes « techniques » de connaissance, puisque le virtuel fonctionne « comme possibilité, […] comme ce que l'on pourrait *faire* »[429], non comme virtualité.

Mais d'un point de vue *externe* à la visée scientifique (*i.e.* sociétés), n'y a-t-il pas des enjeux aussi à « contrôler » statistiquement, à « manipuler » formellement le monde tel qu'il peut se manifester, être observé et éprouvé concrètement, en regard de configurations sociales-historiques actuelles ? Si cette question renvoie à l'idée répandue dans maintes études sociales des algorithmes à propos de l'enfermement des êtres sociaux dans des « bulles de filtre »[430], dans leurs routines de perception et de comportement, nous voyons ici par quels processus ce type d'algorithmes y parviendrait; et étonnamment, ces processus, régulièrement associés à « la froide rationalité des

---

[424] PASSERON Jean Claude, Le raisonnement sociologique, op. cit., p. 145.

[425] GRANGER Gilles-Gaston, Le Probable, le Possible et le Virtuel, op. cit.

[426] ROUVROY Antoinette et Thomas BERNS, « Le nouveau pouvoir statistique », *op. cit.*, p. 159.

[427] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 195.

[428] GRANGER Gilles-Gaston, Le Probable, le Possible et le Virtuel, op. cit., p. 9.

[429] *Ibid.*, p. 234.

[430] PARISER Eli, The Filter Bubble: What The Internet Is Hiding From You, New York: Penguin Press, 2011.

calculs »[431], ne s'opposent pas, comme on le croit parfois, à certaines valeurs fondamentales des sociétés occidentales démocratiques et pluralistes (section 5.4).

Morphologie *sociale* de la causalité *statistique*, du « déterminisme social »

Comme nous l'avons vu, c'est la question de l'incommensurabilité des espaces de représentation par resubstantification du relationnel qui entrent en jeu avec la mise en œuvre de méthodes produisant des arbres. Avec leur extension par forêt aléatoire, il est possible de constater que chaque individu « concret » dans son unicité (non « interchangeable ») semble désormais devenir l'unique référent analytique, qui ne pourrait être comparable qu'en rapport avec lui-même. Or, qu'appelle ce référentiel en termes de changements sociétaux si, comme le soulignait Desrosières, c'est en rendant « comparable », ce qui était autrefois conçu comme « incomparable », que les sociétés se transforment – « comparer les Noirs et les Blancs appelle l'abolition de l'esclavage, comparer les femmes et les hommes appelle le suffrage [...] universel »[432] ? En contraste aux outils classiques de la statistique mathématique/inférentielle, l'algorithme de RF favorise-t-il la construction de « modèles d'action » et de gestion des populations efficaces « en masse, *en probabilité* », mais en « agissant cas par cas » sur des « personnes singulières »[433] ?

Modèle de la corrélation combiné à un usage *différencié* de la moyenne ?

Comme l'avançaient les philosophes et juristes Rouvroy et Berns, les forêts semblent faire partie des outils d'analyse exploratoire/de DM, fondés sur le modèle de la *corrélation* décrit par Desrosières, « laiss[ant] la possibilité de simuler une reconstruction du cas singulier brisé par l'émiettement des codages »[434]. Contrairement néanmoins à leur idée d'une forme « a-normative » de rationalité, affranchie de « toute *forme* de moyenne »[435], les RF semblent bien au contraire fonder expressément la robustesse de leurs performances, de leur efficacité technique (empirique), sur les propriétés formelles de la moyenne[436]. En continuité donc avec les « techniques empiriques

---

[431] CARDON Dominique, À quoi rêvent les algorithmes, op. cit., p. 13.

[432] DESROSIÈRES Alain, « Comparer l'incomparable. Essai sur les usages sociaux des probabilités et des statistiques », *op. cit.*, p. 175.

[433] DESROSIÈRES Alain, « Masses, individus, moyennes: la statistique sociale au XIXe siècle », *op. cit.*, p. 61-62.

[434] *Ibid.*, p. 62.

[435] ROUVROY Antoinette et Thomas BERNS, « Gouvernementalité algorithmique et perspectives d'émancipation », *op. cit.*, p. 165 et 192.

[436] DESROSIÈRES Alain, La politique des grands nombres, op. cit., p. 20.

probabilistes pour observer des faits »[437], il s'agit d'anticiper des comportements humains, des actes libres et individuels ou des préférences personnelles en faisant « émerger » des régularités empiriques.

Des formes « micro-sociales » de détermination ?

Or, si ces formes contemporaines d'analyse permettent toujours de « reconnaître des régularités sociales »[438], il n'empêche que celles-ci semblent bien loin de constituer « des caractères collectifs au-delà des destins individuels »[439]. À travers l'exploration de la « variabilité interne »[440] des personnes, des situations particulières potentiellement ou virtuellement infinies, c'est plutôt la mise à jour de régularités « centrées sur les idées d'échelles et de distributions *individuelles* »[441], d'où la question d'un déterminisme d'ordre microsocial. Ce type d'approche paraît donc rompre avec l'ancienne idée du « déterminisme » par la statistique sociale du 19e siècle, basée sur des régularités macrosociales, des « réalités macro-structurelles »[442] ayant permis de rendre pensable et possible la mise en place de « systèmes d'assurance et de protection sociale »[443]. Ce renversement conceptuel conforterait dès lors la thèse de Ian Hacking lorsqu'il avançait que « moins de déterminisme [par la statistique], plus de possibilités de contrainte » d'expérience[444].

---

[437] BERTAUX Daniel, « Pour sortir de l'ornière néo-positiviste », *op. cit.*, p. 122.

[438] « Ce que l'on appelle les régularités sociales des comportements humains ne sont pas autre chose qu'une connaissance nous assurant une représentation stabilisée du monde dans lequel nous agissons. » (Sabourin, 2009 : 423).

[439] LECLERC Olivier, « Statistiques et normes : jalons pour une rencontre interdisciplinaire », *Cahiers Droit, Sciences & Technologies*, nº 4, 2014, p. 43.

[440] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 138.

[441] *Ibid.*, p. 83.

[442] *Ibid.*, p. 235.

[443] *Ibid.*, p. 84 et 108.

[444] HACKING Ian, « Comment faire l'Histoire de la statistique ? », *op. cit.*, p. 190. Il y a donc des choix opérés de réduction des possibilités inscrits dans la conception même des modèles des algorithmes produisant les DT et les RF.

5.3.3 Personnalisation algorithmique: La place des traces et signaux numériques ?

Comme l'affirmait plusieurs auteurs cités dans le bilan des écrits, l'approche « relationnelle » ou « contextualiste » des méthodes d'arbres, et de forêts par extension, renvoie à une manière particulière de « calculer » le social qui apparait compatible avec l'individualisation des rapports sociaux « de classe, de race et de genre ». Cependant, dans la mesure où le présent travail s'appuie sur des catégories d'analyse « traditionnelles » issues d'une enquête nationale, la dynamique de personnalisation du calcul ne semble pas pouvoir être réduite à une question de données personnelles. Pour le dire autrement, cet enjeu parait difficilement pouvoir s'expliquer *exclusivement* par une saisie « par le bas », par les « traces d'activités » disparates et hétérogènes des internautes comme le prétendait par exemple Cardon avec sa typologie[445].

Au-delà de la construction *sociale* des données *sociales* ?

Accroître le nombre de données empiriques ou de « traces mortes de processus vivants »[446], ou même en diversifier les facettes sensibles, ne multiplierait que les « points de vue situés » (individuels et singuliers) sur le monde systématiquement explorés/explorables par ce type de procédure. Pour le dire concrètement, des caractéristiques plus banales ou anodines *sociologiquement* (théoriquement) telles que le fait d'être passé en voiture sur l'autoroute 20 un mardi après-midi ou le nombre de « likes » sur les photos de profil des répondants de l'ESG2016 auraient pu être intégrées comme éléments empiriques susceptibles d'être *statistiquement* (techniquement) liés ou corrélés à l'état de santé, parmi les grandes catégorisations « sociales »/statistiques « artificiellement universalisantes » retenues (familiales, scolaires, professionnelles, démographiques, géographiques, etc.). Cela aurait-il pour autant changé la forme *générale* de raisonnement ?

Par conséquent, du point de vue des opérations de connaissance, nos résultats rejoignent davantage l'hypothèse de Vayre au sujet de l'institution d'un « pluralisme prédictif », en référence

---

[445] ROUVROY Antoinette et BERNS Thomas, « Gouvernementalité algorithmique et perspectives d'émancipation », *op. cit.* ; CARDON Dominique, *À quoi rêvent les algorithmes*, *op. cit.* ; BOULLIER Dominique, « Les sciences sociales face aux traces du big data », *Revue francaise de science politique*, Vol. 65, nº 5, 2015, p. 805-828.

[446] Expression empruntée à Paul Sabourin.

aux travaux de J.-M. Berthelot[447]. En effet, la méthode de forêt permet de fournir des « versions[448] », des « conjectures plausibles »[449] du monde social, où chaque ensemble de règles extrait de chacun des arbres se présente comme autant d'énoncés « localisés socialement » dans les processus de représentations « abstraites » pour obtenir la prédiction finale (ou le classement) d'une *seule* et *unique* personne « concrète » (comme Lise et Germain). Ce que tend à suggérer la lecture de la forêt simplifiée (fictive) est donc une sorte de pluralisme « relativiste », « *réduisant la signification d'un énoncé à l'expression de son contexte singulier d'énonciation* »[450].

Pourrions-nous dire, en reprenant les mots de Bertaux, que les forêts permettent de « saisir ce qu'il y a de *structurel* dans les processus sociaux et historiques », « profondément diversifiés », en tenant compte de ce que l'on appelle parfois en sociologie les « conflits d'interprétation »[451] ?

---

[447] VAYRE Jean-Sébastien, « Comment décrire les technologies d'apprentissage artificiel ? », *op. cit.* Voir aussi l'idée d'un mode d'appréhension « ouvert » et « localement » situé évoqué par Dagiral et Parasie (2018). *Cf.* Berthelot (1990 et 1998) sur *L'intelligence du social*. Il est intéressant de remarquer que le sous-titre de la cinquième édition de l'ouvrage de Tufféry paru en 2017, *La science des données,* était, sept ans auparavant *L'intelligence des données* (2010, 4e ed.).

[448] LAHIRE Bernard, « La variation des contextes dans les sciences sociales. Remarques épistémologiques », *op. cit.*, p. 398.

[449] BECKER Howard, *Faire preuve*, *op. cit.*, p. 17.

[450] BERTHELOT Jean-Michel, « Les nouveaux défis épistémologiques de la sociologie », *op. cit.*, p. 5.

[451] BERTAUX Daniel, « Pour sortir de l'ornière néo-positiviste », *op. cit.*, p. 120-128.

5.4 Des principes de méthode « scientifiques » et « politiques » convergents ?

Comme nous l'avons vu, le cadre statistique de l'apprentissage dans lequel s'inscrivent la construction et l'évaluation des modèles mathématiques contemporains[452] témoigne d'une façon particulière d'articuler ce qui est connu dans l'état *actuel* des connaissances avec ce qui ne l'est pas encore, de combiner « données, preuves et idées » comme dirait H. Becker (2020). En ayant cherché à expliciter les dualismes constitutifs de la raison statistique, que pouvons-nous tirer des méthodes de raisonnement par forêts comme moyen d'administration de la preuve fournissant des connaissances stables et générales[453], fondé sur une certaine conception de la totalité sociale pertinente[454] ?

5.4.1 Généralisation par approximation ou par saturation ? Vérification « historique »

Pour les sociologues et méthodologues quantitativistes, le développement des modèles en ML renvoie à une autre école de pensée, celle de Benzécri, qui avançait que « le modèle doit suivre les données, non l'inverse »[455], voire celle de Quetelet[456]. Mais d'un autre point de vue, ces développements méthodologiques ne sont pas sans faire écho au principe de « saturation » dans la recherche qualitative (Glaser et Strauss, 1967)[457], comme manière alternative « de construire de la généralité »[458], d'enraciner le travail d'abstraction théorique. Pour Bertaux, le « concept » de saturation, couplé à la « diversification » de l'échantillonnage, justifierait la validité de l'approche biographique et des récits de vie en sociologie, en permettant de « stabiliser » les représentations dites « théoriques » :

---

[452] En particulier les procédures d'élagage, de validation croisée et de ré-échantillonnage orientant le choix des hyperparamètres « optimaux » (section 1.3.2).

[453] HACKING Ian, « Style pour historiens et philosophes », *op. cit.*, p. 316.

[454] DESROSIÈRES Alain, La politique des grands nombres, op. cit., p. 260 ; DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 143.

[455] BENZÉCRI Jean Paul, *L'analyse des données*, *op. cit.*, p. 6.

[456] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 69.

[457] GLASER Barney G. et Anselm L. STRAUSS, *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Chicago, Aldine, 1967 ; SAVOIE-ZAJC Lorraine, « Comment peut-on construire un échantillonnage scientifiquement valide? », *Recherches Qualitatives*, Hors Série, nº 5, 2007, p. 104 et 109.

[458] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 151.

> lorsque [la saturation] est atteinte, elle confère une base très solide à la *généralisation* : à cet égard elle remplit pour l'approche biographique très exactement la même fonction que la représentativité de l'échantillon pour l'enquête par questionnaires[459]

À mon sens, la mise en œuvre d'approche par forêts constitue un cas exemplaire de théorisation enracinée dans la totalité concrète. Dans le cadre statistique de l'apprentissage comme dans celui de la saturation en sociologie qualitative, la validation des modèles et connaissances inférés repose moins sur des critères de nature théorique postulés et justifiés *préalablement*[460], ou d'une « valeur fondée sur les motifs et procédés du raisonnement réfléchis par l'imagination méthodologique »[461] que sur des bases proprement « historiques », empiriquement déterminées: le degré de correspondance (ou de conformité) des modèles « abstraits » avec les états effectivement réalisés, qui viennent (et sont venus) à exister dans l'histoire, dans les espaces-temps sociohistoriques[462]. En reprenant la distinction proposée par J.-C. Combessie, les modèles de forêt sont ainsi plus proches de l'approche « compréhensive » (*intensive*) « intrinsèquement interprétative » (herméneutique), qui « sature de tous les sens possibles un cas isolé dans l'espace et le temps »[463] que celle « explicative » :

> L'approche compréhensive multiplie les points de vue sur l'objet. Rétive à l'homogénéisation, elle privilégie la diversité des aspects, la densité des relations qui constituent son objet. A ce titre elle est portée à le saisir comme ensemble spécifique de relations différenciées. […] Multiples et polymorphes, [ces relations] peuvent apparaître, disparaître ou se modifier selon les points de vue, les lieux et les moments de l'observation […] L'objet est contradictoire et conflictuel - et en même temps singulier[464].

Dès lors, l'un des enjeux que posent les forêts comme outil de description du monde social et de coordination (d'action, de décision, de prévision) en son sein est « d'oublier que ce qui s'est réalisé ne fut jamais qu'un possible, plus ou moins probable, parmi d'autres »[465], ou encore, la difficulté de pouvoir *voir*, *penser* et *faire* (savoir ?) « que l'effet attendu peut ne pas se produire,

---

[459] BERTAUX Daniel, « L'approche biographique: sa validité méthodologique, ses potentialités », *Cahiers Internationaux de Sociologie*, vol. 69, 1980, p. 208.

[460] FOUCART Thierry, « L'interprétation des résultats statistiques », *op. cit.*

[461] HAMEL Jacques, « A propos de l'échantillon, de l'utilité de quelques mises au point », *BMS*, nº 67, 2000, p. 16.

[462] HACKING Ian, « L'ontologie historique », *op. cit.*, p. 290.

[463] COMBESSIE Jean-Claude, « A propos de méthodes: effets d'optique, heuristique et objectivation », *BMS*, nº 10, 1986, p. 7.

[464] Ibid.

[465] À l'inverse de ce que disait Grignon à propos des méthodes statistiques. Voir GRIGNON Claude, « Prédiction et rétrodiction », *op. cit.*, p. 90.

[et que] l'inattendu, l'improbable peut se réaliser »[466]. Contrairement à ce que suggèrent certaines études sociales récentes des algorithmes, les enjeux à propos du « pouvoir prédictif » de ces dispositifs ne se situent pas, en ce qui concerne les arbres et les forêts du moins, au niveau des abstractions mathématiques « grossières », mais au niveau le plus « concret » de la connaissance des faits sociaux[467], de la « saisie descriptive » ou des conditions de l'« observation historique » pour reprendre l'expression de Passeron.

Face à la densité des multiples descriptions empiriques, hypothétiques (équivalentes, concurrentes, « plausibles ») du monde social, aucune forêt ne pourrait prétendre « épuiser » intégralement le réel. Cohérentes avec la multiplication des régimes d'inégalités caractérisant les sociétés postmodernes, « liquides » et fragmentées que décrit F. Dubet, les forêts aléatoires sont susceptibles de modéliser le « gouffre sans fin » de la sociologie des inégalités, en étant elles-mêmes, le résultat (provisoire) d'une agrégation de « *sortes* de théories pratiques, en actes, de la justice (relativement contradictoires) »[468]. En ce sens, la recherche de la « justesse mathématique » des prédictions via le développement de l'analyse statistique et du calcul des probabilités se révèle encore compatible avec des principes et des idéaux de « justice sociale » [469]. Comme le montre l'analyse, c'est en vertu d'une voie « démocratique », de règles basées sur la majorité[470], que les méthodes de forêt produisent des savoirs susceptibles de façonner des opportunités de vie et d'intervenir dans les espaces de possibilités de l'expérience personnelle, de l'observation historique, voire de l'« imagination sociologique ».

En décalage cependant avec les logiques d'analyse plus traditionnelles, les « relations d'interdépendance spécifiques »[471] entre traces observables variables du social sont prises « dans leur contingence historique »[472] plutôt que dans leur équivalence technique, suggérant dès lors une

---

[466] *Ibid.*, p. 82.

[467] GRANGER Gilles-Gaston, « Modèles qualitatifs, modèles quantitatifs dans la connaissance scientifique », *op. cit.*, p. 11.

[468] DUBET François, « Régimes d'inégalité et injustices sociales », *SociologieS*, [En ligne], novembre 2011, p. 62 et 67.

[469] Principes « scientifiques » = Règles de méthode « démocratiques » = Principes « politiques » ? Voir DESROSIÈRES Alain, *La politique des grands nombres*, *op. cit.*, p. 45 ; PORTER Theodore M., *La confiance dans les chiffres*, *op. cit.*

[470] (ou sinon d'un calcul de moyenne pour une variable continue)

[471] LAHIRE Bernard, « La variation des contextes dans les sciences sociales. Remarques épistémologiques », *op. cit.*

[472] BILGE Sirma, « Théorisations féministes de l'intersectionnalité », *op. cit.*, p. 16.

conception du « social » de plus en plus « réduit[e] à la singularité *actuelle* d'expériences »[473], aux aléas individuels de la vie quotidienne[474].

[473] GRANGER Gilles-Gaston, Le Probable, le Possible et le Virtuel, op. cit., p. 232.

[474] Quelle conception de l'empirie, de l'expérience pratique et du « terrain » a-t-on contribué implicitement à façonner, en s'opposant à la sociologie empirique, dite « quantitative » ? Est-ce qu'il y a des liens avec le débat Durkheim-Tarde ? et B. Latour ? Quelles en sont les implications du point de vue de notre conception évolutive des faits sociaux ? *Cf*. LATOUR Bruno, JENSEN Pablo, VENTURINI Tommaso, GRAUWIN Sebastian et BOULLIER Dominique, « "Le tout est toujours plus petit que ses parties" », *Réseaux*, n° 177, nº 1, 2013, p. 197-232 ; ŒHMICHEN Hélène et VIEDROV Oleksii, « L'usage comparé des statistiques par Gabriel Tarde et Emile Durkheim », *Statistique et Société*, vol. 4, nº 3, 2016, p. 65-82.

EN GUISE DE CONCLUSION

En résumé, le travail d'analyse présenté dans les deux derniers chapitres représente une synthèse non exhaustive des critiques formulées aux formes classiques de la statistique sociale/mathématique, appliquée dans ses différences aux techniques d'analyse plus récentes. Plusieurs de ces critiques relativement anciennes, mais d'actualité sont apparues pertinentes sociologiquement pour interroger les opérations de connaissance qu'implique la mise en œuvre de deux méthodes contemporaines que sont les arbres de décision (DT) et les forêts aléatoires (RF).

Portée et limites des critiques antérieures : En effet, les reproches antérieurs semblent avoir substantifié le concept de « variable » dans la sémantique mathématique[475], ou l'usage même de toute catégorie sociale d'analyse dans la sémantique statistique. Bien souvent, les variables sont assimilées à des « mesures » empiriques, à des « indicateurs » sociaux[476] ou à des « items standardisés » de nomenclatures renvoyant à des facettes sensibles de l'expérience vécue/perçue, qui peuvent être plus ou moins précises (au plan *quantitatif*) et/ou adéquates et adaptées (au plan *qualitatif*).

De plus, bien que différentes catégories « englobantes », « homogénéisantes » et « réductrices » puissent toujours être mobilisées dans les analyses, avec les développements récents en ML, ce sont moins les êtres sociaux et leurs expériences singulières – des contenus « concrets, substantiels, manifestes » de l'expérience – qui font l'objet d'un processus de standardisation et d'abstraction logico-formelle que les méthodes et procédures opératoires elles-mêmes, les processus d'analyse permettant de « contrôler statistiquement »[477], de « manipuler formellement »[478] les contenus/construits empiriques, effectifs/existants[479].

---

[475] Voir DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 66.

[476] Des « opérations de représentation » (Selz et Maillochon, 2009 : 143)

[477] DURAND Claire, DESLAURIERS Mélanie et DUHAIME Gérard, « Quelles statistiques pour analyser les inégalités ? », *op. cit.*, p. 7.

[478] Certaines confusions persistantes entre « la *qualité vécue* de la saisie immédiate, et la *qualité pensée* et manipulée comme forme » (Granger 1988 : 120 cité dans Hamel, 1989 : 66) ? Voir aussi Granger (1992).

[479] Selon Desrosières, « [l]a systématisation et l'automatisation des procédures offrent de grands avantages, non seulement du point de vue économique, en termes de coûts, par exemple pour le codage, mais aussi dans la perspective de la recherche de l'accord, de l'objectivation d'un sens commun aux divers acteurs, en dégageant ces procédures de l'intervention des personnes, et en les inscrivant dans des machines. » (Desrosières, 1993 : 340)

Statistique, société et sociologie : quels rapports ?

Tel qu'il prend forme à travers les arbres et les forêts, le raisonnement statistique et probabiliste peut difficilement être accusé de « totalisateur et réducteur de la diversité, des singularités individuelles »[480], culturelles ou historiques. Il s'agit moins de *réduire* la « singularité spécifique des personnes » par des modélisations universalistes ou « transhistoriques » que de la *restituer* dans toute sa complexité, ses ambiguïtés et ses variations « plausibles », « de façon de plus en plus détaillée », par « souci de coller à l'expérience immédiate »[481].

Par contraste à la logique expérimentale « toutes choses égales par ailleurs », « arithmétique », ces méthodes permettent plutôt d'appréhender de façon « relationnelle » le social, par une certaine forme de ré-indexation des descriptions statistiques/empiriques « littéraires[482] ». Les arbres de décision (uniques, « simples ») semblent procéder d'un mode de cumul et de généralisation du savoir « contextuel », « interactionniste », voire « intersectionnel »[483] que prolongent les forêts aléatoires, tout en intégrant une vision pluraliste, « relativiste » et potentiellement conflictuelle du réel[484].

Bien que l'on puisse supposer la permanence du passé en évaluant les méthodes et leurs modèles selon leur capacité à prédire efficacement la vie sociale déjà actualisée, le type d'interprétation que suscitent ces approches de quantification s'apparente davantage à des « mises en récit », plus proche du versant narratif et littéraire de la statistique, du mode de connaissance de type monographique. Pour le dire autrement, on semble bien loin des lois « figées » et « rigides », des entités désincarnées du monde « concret » (Halbwachs, Passeron), auxquels les outils statistiques sont encore aujourd'hui si fréquemment associés, à tort ou à raison, en sciences sociales. Comme le montre l'examen des forêts, la forme que tend à prendre l'analyse statistique et le calcul des probabilités répond toujours à des problèmes d'arbitrage entre une pluralité de

---

[480] DESROSIÈRES Alain, Pour une sociologie historique de la quantification, op. cit., p. 113 et 248.

[481] *Ibid.*, p. 144.

[482] *Ibid.*, p. 139.

[483] DEGENNE Alain, « Une méthodologie « douce » en sociologie », *op. cit.* ; BILGE Sirma, « Théorisations féministes de l'intersectionnalité », *op. cit.* ; CARDE Estelle, « Les inégalités sociales de santé au prisme de l'intersectionnalité », *op. cit.*

[484] BERTAUX Daniel, « Pour sortir de l'ornière néo-positiviste », *op. cit.*

« points de vue » socialement situés, nécessaires à la vie en société, à l'existence d'un monde « social » (commun et singulier), et qui étaient à l'origine de ce style scientifique de raisonnement[485].

Bien que les processus de vastes « totalisations statistiques » et de « montée en généralité » aient été finement étudiés par Desrosières, la recherche espère avoir pu contribuer à ouvrir quelques pistes de réflexion quant au procès/mouvement inverse : la « redescente dans le particulier »[486]. En effet, les arbres et les forêts soulèvent des questions relatives à l'articulation des catégories sociales (spécifiques et générales) de l'expérience dans une optique toutes choses *différentes* (ou *inégales*) par ailleurs[487], dépassant dès lors les termes « traditionnels » dans lesquels sont posés les enjeux entourant la dé-construction « sociale » de la réalité empirique des données[488] et le caractère strictement arbitraire des « conventions » et définitions plus ou moins explicitées.

D'ailleurs, le travail d'analyse réalisé jusqu'à ce jour tend à montrer que les logiques de connaissance et les paramètres des approches algorithmiques analysées peuvent véhiculer des conceptions et des principes de représentation qui ne semblent pas totalement incompatibles avec certaines tendances et courants en vogue ces dernières décennies dans la recherche en sciences sociales. Pour ne donner que quelques exemples, exigeant chacun des examens ultérieurs plus approfondis, les approches par arbre de décision et par forêt aléatoire paraissent étroitement liées aux processus d'identification « intersectionnelle », « multiple », « plurielle » et « fluide » dans les sociétés[489]. En ce sens, les types d'enjeux que nous avons cherché à éclairer à partir des éléments de critique usuels de la méthodologie « conventionnelle » entretiennent d'intrigantes relations avec

---

[485] DESROSIÈRES Alain, La politique des grands nombres, op. cit.

[486] À noter que Desrosières voyait déjà à l'œuvre la tendance inverse lorsqu'il indiquait que « [l]es développements récents [de la raison statistique] l'ont conduite aussi dans une direction apparemment opposée [à son caractère « galiléen »], celle d'une identification et d'une catégorisation fine des personnes, à des fins marchandes. […] L'intelligence artificielle et les systèmes experts permettent d'automatiser des choix, des décisions, des affectations d'individus à des classes de traitements variés. » (Desrosières, 2008 : 113). DODIER Nicolas, « Les sciences sociales face à la raison statistique (note critique) », *Annales*, vol. 51, nº 2, 1996, p. 408-428.

[487] TREMBLAY André, « Feu la société globale et les méthodes quantitatives », *op. cit.*

[488] HOULE Gilles et RAMOGNINO Nicole, « Présentation. La construction des données », *op. cit.* ; PARENT Frédéric et SABOURIN Paul, « Présentation. Les espaces-temps de la production ethnographique », *op. cit.*

[489] LAHIRE Bernard, « La variation des contextes dans les sciences sociales. Remarques épistémologiques », *op. cit.* DUBET François, « Régimes d'inégalité et injustices sociales », *op. cit.*

certains débats épistémologiques, méthodologiques et théoriques/ontologiques[490], entourant la « standpoint theory », la « thick description », les récits de vie, les tournants biographique et narratif, la (ou les) sociologie(s) de l'individu[491], etc. Existe-t-il des liens par exemple entre l'*Homme moyen* (Quetelet, 1844-1848), l'*Homme sans qualité* (Musil, 1930-1932), l'*Homme unidimensionnel* (Marcuse, [1964]1968) et l'*Homme pluriel* (Lahire, 1998) ?

Élargir la présente étude sociologique de « cas statistiques » à d'autres méthodes contemporaines en vogue dans la « science des données » (*p. ex.* les machines à vecteur support ou les réseaux de neurones) permettrait-il d'invalider notre hypothèse au sujet de la compatibilité des formes actuelles/historiques de raisonnement statistique et sociologique ? Existe-t-il d'autres parallèles entre le renouvellement des statistiques mathématiques/sociales et le développement des traditions de pensée (cognitives) et d'enquête (pratiques) en sociologie[492] comme discipline scientifique à part entière se caractérisant par son « assise statistique »? Dans les études « sociales » et critiques contemporaines des algorithmes de l'IA, il est en ce sens curieux de voir parfois ressasser les bons vieux reproches à l'égard des formalisations mathématiques *dans* et *par* des démarches de recherche qui partagent pourtant d'étonnantes proximités et similarités avec ce qu'elles entreprennent de dénoncer. Comme le disait Héran,

> Pour récuser la « mécanique du langage des variables », il semble […] que rien ne vaut le langage des variables. La question préjudicielle sur les conditions de possibilité de la statistique ne pourrait-elle se formuler que sous une forme statistique et probabiliste? Si oui, cela veut dire qu'une telle question, loin de réduire le rôle de la statistique, appelle à son tour de nouvelles mesures. La critique de la forme au nom des contenus ou des contextes de prélèvement devient inévitablement une critique de la forme par le recours à plus de formes[493].

---

[490] Voir notamment HOULE Gilles, « Présentation. La sociologie : une question de méthode ? », *op. cit.* HAMEL Jacques, « Les théories sociologiques comme expression de l'appropriation culturelle, un point de vue critique », *Canadian Review of Sociology/Revue canadienne de sociologie*, vol. 58, nº 2, 2021, p. 271-280.

[491] MARTUCCELLI Danilo, « Qu'est-ce qu'une sociologie de l'individu moderne ? Pour quoi, pour qui, comment ? », *Sociologie et sociétés*, vol. 41, nº 1, 2009, p. 15-33.

[492] BERNARD Paul, « Présentation : les nouvelles statistiques sociales », *Sociologie et sociétés*, vol. 35, nº 1, 2003, p. 3-18 ; HORVATH Robert A., « Épistémologie et méthodologie de la statistique en relation avec la sociologie », *Journal de la société statistique de Paris*, vol. 130, nº 4, 1989, p. 209-218.

[493] HÉRAN François, « L'assise statistique de la sociologie », *op. cit.*, p. 34.

Et si le « substrat empirique » de l'objet de la sociologie était de nature « statistique »[494]? Sinon, devrait-on plutôt conclure avec Passeron que « la *configuration historique singulière* constitue la seule réalité empirique »[495] existante ? Si le réel sociologique se résume au réel historique, alors il conviendrait d'adopter le pessimisme de Granger quant à la possibilité de voir reconnaître un jour le caractère « scientifique » de la discipline sociologique. La sociologie est-elle simplement, uniquement une science « historique » (Passeron, 1991), « concrète » (Bertaux, 1987) ou « contextuelle » (Degenne, 1981) ? Quelles formes (d'*objets*) de connaissance produisons-nous en sociologie ? Le savoir produit au nom de la sociologie se distingue-t-il de celui produit par les spécialistes de la « science des données » ? Si oui, en quoi ?

Personnellement, à titre d'apprentie sociologue parmi bien d'autres, pourquoi m'entêter à poursuivre des études au doctorat en sociologie alors que d'autres programmes me permettraient de m'intéresser à un large éventail de phénomènes sociaux à partir des *mêmes* méthodes d'analyse statistique (tout en faisant même partie de la nouvelle génération de « scientifiques » des temps modernes!) ? Le présent travail aurait-il même plus de chance d'être qualifié de *sociologique* au sens « traditionnel » du terme, tant par les quantitativistes que par les qualitativistes en sciences sociales ?

Dans la mesure où la scientificité de la « science des données » apparaît présentement plus évidente que celle de la sociologie, c'est en ce sens que l'on croit pouvoir voir la pertinence proprement « sociologique » de notre projet de recherche. Au bout du compte, ce dernier pose davantage de questions à la sociologie, qui reste dans un état d'une science émergente encore aujourd'hui, qu'à la « science des données » du XXIe siècle[496].

Avant de chercher à établir la « réciprocité des perspectives » d'acteurs sociaux externes à la sociologie, quel pourrait être l'intérêt d'établir la réciprocité des perspectives au sein même des sociologues, des traditions « conventionnelles » et des grands clivages qui ont animé le

---

[494] *Cf.* Table-ronde « Ethnographie et épistémologie. Sur la division du travail sociologique », organisée par le Laboratoire de recherches ethnographiques du Québec (LABREQ) à Montréal, en présence de Nicole RAMOGNINO, de Frédéric PARENT, de Jean-François CÔTÉ et de Paul SABOURIN, Département de sociologie, Université du Québec à Montréal, 11 octobre 2017.

[495] PASSERON Jean Claude, *Le raisonnement sociologique*, *op. cit.*, p. 222.

[496] Revoir ici les limites des approches *philosophiques*, *professionnelles*, *ethnographiques* et *statistiques* des algorithmes.

développement de la discipline, mais que l'on s'empresse si souvent de déclarer « dépassés », « stériles » ou « faux »[497] ? Depuis la parution du *Métier du sociologue* en 1968, il est courant d'entendre dire que la « malédiction de la sociologie est d'avoir affaire à un objet qui parle »[498]. Cette situation « problématique » à l'époque devient-elle une « bénédiction » (pour rester dans le registre religieux des auteurs) à l'heure des « données brutes » parlantes[499] ?

[497] Qu'est-ce qui est « faux » ? Le débat ou les clivages ? Ou les deux ? Ou aucun des deux ?

[498] BOURDIEU Pierre, PASSERON Jean-Claude et CHAMBOREDON Jean Claude, *Le métier de sociologue. Préalabes épistémologiques*, La Haye-Paris, Mouton-Bordas, 1968, p. 56-57.

[499] Un examen plus attentif aux étonnantes coïncidences entre formes mathématiques et formes sociologiques pourrait-il permettre de démontrer la cumulativité du savoir sociologique (non seulement la postuler), en appréhendant ce que Granger désigne par la notion de « contenu formel » des objets scientifiques et dont font parties à part entière les objets sociologiques ?

Références bibliographiques

ABITEBOUL Serge et Gilles DOWEK, *Le temps des algorithmes*, Paris : Éditions Le Pommier, 2017.

AMOORE Louise et Volha PIOTUKH, « Life Beyond Big Data Governing With Little Analytics », *Economy and Society*, vol. 44, nº 3, 2015, p. 341-366.

ANADÓN Marta, « Les méthodes mixtes : implications pour la recherche « dite » qualitative », *Recherches qualitatives*, vol. 38, nº 1, 2019, p. 105-123.

ANDERSON Chris, « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired Magazine*, vol. 16, nº 7, 23 juin 2008. En ligne au <https://www.wired.com/2008/06/pb-theory/>.

BASTIN Gilles et Paola TUBARO, « Le moment big data des sciences sociales », *Revue française de sociologie*, vol. 59, nº 3, 2018, p. 375-394.

BECKER Howard, *Faire preuve: Des faits aux théories*, Paris, la Découverte, 2020. [En particulier, « Première partie. Données, preuves, idées », pp. 9-88]

BEER David, « The Social Power of Algorithms », *Information Communication & Society*, vol. 20, nº 1, 2017, p. 1-13.

BENBOUZID Bilel et Dominique CARDON, « Machines à prédire », *Réseaux*, n° 211, nº 5, 2018, p. 9-33.

BENZÉCRI Jean Paul, *L'analyse des données*, Tome 2, Paris, Dunod, 1984, p. 3-17.

BERNARD Paul, « Présentation : les nouvelles statistiques sociales », *Sociologie et sociétés*, vol. 35, nº 1, 2003, p. 3-18.

BERTAUX Daniel, « L'approche biographique: sa validité méthodologique, ses potentialités », *Cahiers Internationaux de Sociologie*, vol. 69, 1980, p. 197-225.

———. « Pour sortir de l'ornière néo-positiviste », *Sociologie et sociétés*, vol. 8, nº 2, 1976, p. 119-134.

BERTHELOT Jean-Michel, « Les nouveaux défis épistémologiques de la sociologie », *Sociologie et sociétés*, vol. 30, nº 1, 1998, p. 23-38.

BESSE Philippe, Caroline LE GALL, Nathalie RAIMBAULT et Sophie SARPY, « Data mining et statistique », *Journal de la société française de statistique*, vol. 142, nº 1, 2001, p. 5-36.

BESSE Philippe, Céline CASTETS-RENARD, Aurélien GARIVIER et Jean-Michel LOUBES, « L'IA du quotidien peut-elle être éthique ? », *Statistique et Société*, vol. 6, no 3, 2018, p. 9-31.

BILAND Émilie, Jean-Sébastien EIDELIMAN et Séverine GOJARD, « Ceteris (non) paribus ? », *Geneses*, n° 73, nº 4, 2008, p. 37-56.

BILGE Sirma, « De l'analogie à l'articulation : théoriser la différenciation sociale et l'inégalité complexe », *L'Homme la Societe*, vol. 2, n°176-177, 2010, p. 43-64.

———. « Théorisations féministes de l'intersectionnalité », *Diogene*, vol. 1, n° 225, 2009, p. 70-88.

BLUM Alain et Maurizio GRIBAUDI, « Des catégories aux liens individuels : l'analyse statistique de l'espace social », *Annales. Histoire, Sciences Sociales*, vol. 45, nº 6, 1990, p. 1365-1402.

BLUMER Herbert, « Sociological Analysis and the "Variable" », *American Sociological Review*, vol. 21, nº 6, 1956, p. 683-690.

BOELAERT Julien et Étienne OLLION, « The Great Regression Machine Learning, Econometrics, and the Future of Quantitative Social Sciences », *Revue française de sociologie*, vol. 59, nº 3, 2018, p. 475-506.

BOUDON Raymond, « Les statistiques peuvent-elles donner une image réelle de la réalité sociale? », *Sociologie et sociétés*, vol. 8, nº 2, 1976, p. 141-156.

BOULLIER Dominique, « Vie et mort des sciences sociales avec le big data », *Socio*, nº 4, 2015, p. 19-37.

BOURDIEU Pierre, Jean-Claude PASSERON et Jean Claude CHAMBERON, *Le métier de sociologue. Préalabes épistémologiques,* La Haye-Paris, Mouton-Bordas, 1968.

BOYD Danah et Kate CRAWFORD, « Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon », *Information, Communication & Society*, vol. 15, nº 5, 2012, p. 662-679.

BREIMAN Leo, « Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author) », *Statistical Science*, vol. 16, nº 3, 2001, p. 199-231.

———. « Random Forests », *Maching Learning*, vol. 45, nº 1, 2001, p. 5-32.

BRY Xavier, Nicolas ROBETTE et Olivier ROUEFF, « Un dialogue de sourds dans le théâtre statistique ? Analyse géométrique des données et effets de structure », 2014. En ligne au < https://halshs.archives-ouvertes.fr/halshs-01018778/document>.

BRYMAN, Alan. « Quantitativisme et qualitativisme : un faux débat ? » dans Jean-Michel BERTHELOT (dir.), *Sociologie, épistémologie d'une discipline. Textes fondamentaux*, Paris, De Boeck Université, 2000, p. 209-220.

BURRELL Jenna, « How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms », *Big Data & Society*, vol. 3, nº 1, 2016, p. 1-12.

CAPEL Roland, Denis MONOD et Jean-Pierre MÜLLER, « De l'usage perverti des tests inférentiels en sciences humaines », *Genèses. Sciences sociales et histoire*, vol. 26, nº 1, 1997, p. 123-142.

CARDE Estelle, « Les inégalités sociales de santé au prisme de l'intersectionnalité », *Sciences sociales et sante*, vol. 39, nº 1, 2021, p. 5-30.

CARDON Dominique, *À quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Paris, Le Seuil, coll. « La République des idées », 2015.

CHAPOULIE Jean-Michel, « Un type d'explication en sociologie : les systèmes de variables en relations causales », *Revue française de sociologie*, vol. 10, nº 3, 1969, p. 333-351.

CHENU Alain, « La catégorisation statistique », *Sociétés Contemporaines*, vol. 26, nº 1, 1997, p. 5-9.

CHIAPELLO Ève, Corine EYRAUD, Philippe LORINO, Alain SUPIOT, Ève LAMENDOUR et Yannick LEMARCHAND, « À propos de l'emprise du chiffre », *Entreprises et histoire*, vol. 2, nº 79, 2015, p. 174-187.

CIBOIS Philippe, *Les méthodes d'analyse d'enquêtes*, Lyon, ENS Éditions, 2015.

———. « Modèle Linéaire contre modèle logistique en régression sur données qualitatives », *BMS : Bulletin de Méthodologie Sociologique*, vol. 64, nº 1, 1999, p. 5-24.

———. « Analyse des données et sociologie », *L'Année sociologique (1940/1948-)*, vol. 31, 1981, p. 333-348.

COMBESSIE Jean-Claude, « A propos de méthodes: effets d'optique, heuristique et objectivation », *Bulletin de Méthodologie Sociologique*, nº 10, 1986, p. 4-24.

DAGIRAL Éric et Sylvain PARASIE, « La «science des données» à la conquête des mondes sociaux: ce que le «Big Data» doit aux épistémologies locales », dans Pierre-Michel MENGER et Simon PAYE (dir.), *Big data et traçabilité numérique: Les sciences sociales face à la quantification massive des individus*, Collège de France, 2017, p. 85-104.

DEAUVIEAU Jérôme, « Comment traduire sous forme de probabilités les résultats d'une modélisation logit ? », *BMS Bulletin de Méthodologie Sociologique*, n°105, 2010, p. 5-23.

DEGENNE Alain, « Une méthodologie « douce » en sociologie », *L'Année sociologique (1940/1948-)*, vol. 31, nº 3, 1981, p. 97-124.

DES NÉTUMIÈRES Félicité, « Méthodes de régression et analyse factorielle », *Histoire & Mesure*, vol. 12, nº 3, 1997, p. 269-297.

DESROSIÈRES Alain, *Pour une sociologie historique de la quantification : L'Argument statistique I*, Paris, Presses de l'École des Mines, coll. « Sciences sociales », 2008.

———. « Analyse des données et sciences humaines : comment cartographier le monde social ? », *Journal Électronique d'Histoire des Probabilités et de la Statistique*, vol. 4, nº 2, 2008, p. 11-18.

———. « Comparer l'incomparable. Essai sur les usages sociaux des probabilités et des statistiques », dans Jean-Philippe TOUFFUT (dir.), *La Société du probable: les mathématiques sociales après Augustin Cournot*, France Paris, Albin Michel, 2007, p. 163-200.

———. *La politique des grands nombres: histoire de la raison statistique*, 2e éd., Paris : La Découverte, 2000.

———. « Comment faire des choses qui tiennent : histoire sociale et statistique », *Histoire & Mesure*, vol. 4, nº 3, 1989, p. 225-242.

———. « Masses, individus, moyennes: la statistique sociale au XIXe siècle », *Hermès*, vol. 2, nº 2, 1988, p. 41-66.

———. « Histoires de formes : statistiques et sciences sociales avant 1940 ». *Revue française de sociologie,* vol. 26, n° 2, 1985, p. 277-310.

DIAKOPOULOS Nicholas, « Algorithmic Accountability Reporting: On the Investigation of Black Boxes », *Digital Journalism*, 2014. <http://www.nickdiakopoulos.com/wp-content/uploads/2011/07/Algorithmic-Accountability-Reporting_final.pdf>.

DODIER Nicolas, « Les sciences sociales face à la raison statistique (note critique) », *Annales*, vol. 51, nº 2, 1996, p. 408-428.

DONOHO David, « 50 Years of Data Science », *Journal of Computational and Graphical Statistics*, vol. 26, nº 4, 2017, p. 745-766.

DUBET François, « Régimes d'inégalité et injustices sociales », *SociologieS*, Débats, Penser les inégalités, octobre 2011. En ligne <https://journals.openedition.org/sociologies/3643>.

DURAND, Claire. « La régression logistique, quelques notes » [Notes de cours], dans SOL2020 : *Statistiques sociales avancées*, Département de Sociologie, 2016, Université de Montréal.

DURAND Claire, Mélanie DESLAURIERS et Gérard DUHAIME, « Quelles statistiques pour analyser les inégalités ? Le cas des Premières Nations au Québec », *SociologieS*, Débats, Penser les inégalités, mai 2012. En ligne au< https://journals.openedition.org/sociologies/3914?lang=en>.

FALISSARD Bruno, « Méthodes de segmentation, CART » et « Modèles linéaires généralisés: régressions logistiques et de Poisson » dans *Comprendre et utiliser les statistiques dans les sciences de la vie*, Paris, Masson, 2005, p. 297-302 et 145-178.

FAYOLLE Jacky, « "La gouvernance par les nombres" est-elle la fin de l'histoire de la statistique ? », dans le cadre de la Journée européenne de la statistique, Luxembourg, 20 mai 2016.

FOUCART Thierry, « L'interprétation des résultats statistiques », *Mathématiques et Sciences Humaines*, vol. 39, nº 153, 2001, p. 21-28.

FRIEDMAN Jerome H., « Data Mining and Statistics : What's the Connection ? », *Computing Science and Statistics*, vol. 29, n° 1, 1998, p. 3-9.

GAUTHIER Benoît (dir.), *Recherche sociale: de la problématique à la collecte des données*, 5e éd., Sainte-Foy, Presses de l'Université du Québec, 2009.

GENUER Robin et Jean-Michel POGGI, « Arbres CART et Forêts aléatoires, Importance et sélection de variables », dans Myriam MAUMY-BERTRAND, Gilbert SAPORTA et Christine THOMAS-

AGNAN (dir.), *Apprentissage statistique et données massives*, Paris: Technip, 2018, p. 295-342.
GHATTAS Badih, « Prévisions par arbres de classification », *Mathématiques et sciences humaines*, vol. 37, no 146, 1999, p. 31-49.
GILLESPIE Tarleton, « The Relevance of Algorithms », dans Tarleton GILLESPIE, Pablo J. BOCZKOWSKI et Kirsten A. FOOT (dir.), *Media Technologies: Essays on Communication, Materiality, and Society*, Cambridge: MIT Press, 2014, p. 167-193.
GINGRAS, Yves. « L'intelligence sociologique confrontée à l'"intelligence artificielle" » (Conférence inaugurale), *Ateliers SociologIA*. Montréal : Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), 10 avril 2019.
GOSSELIN Gabriel, « Sociologie, classement et quantification », *Cahiers Internationaux de Sociologie*, vol. 93, 1992, p. 321-337.
GOLLAC Michel. « Des chiffres insensés ? Pourquoi et comment on donne un sens aux données statistiques ». *Revue française de sociologie*, vol. 38, no 1, 1997, p. 5-36.
GRANGER Gilles-Gaston, *Le Probable, le Possible et le Virtuel*, Paris, Éditions Odile Jacob, 1995.
———. *Formes, Opérations, Objets*, Paris : Vrin, 1994.
———. « Modèles qualitatifs, modèles quantitatifs dans la connaissance scientifique », *Sociologie et sociétés*, vol. 14, nº 1, 1982, p. 7-13.
GRIGNON Claude, « Prédiction et rétrodiction », *Revue européenne des sciences sociales,* vol. 46, nº 142, « Sociologie et idéologie », 2008, p. 75-90.
GUEGUEN A. et NAKACHE J. P., « Méthode de discrimination basée sur la construction d'un arbre de décision binaire », *Revue de Statistique Appliquée*, vol. 36, nº 1, 1988, p. 19-37.
HACKING Ian, « Style pour historiens et philosophes », dans Jean-François BRAUNSTEIN (dir.), *L'histoire des sciences. Méthodes, styles et controverses*, Paris, Vrin, 2008, p. 287-320.
———. « L'ontologie historique », dans Laurence KAUFMANN et Jacques GUILHAUMOU (dir.), *L'invention de la société. Nominalisme politique et science sociale au XVIIIe siècle*, Paris, Éditions de l'EHESS, coll. « Raisons Pratiques », n°14, 2003, p. 287-308.
———. « Leçon inaugurale » et « Cours au Collège de France », *Chaire de Philosophie et histoire des concepts scientifiques (2001-2006)*. Versions électroniques disponibles en ligne sur le site internet du Collège de France <https://www.college-de-france.fr/site/ian-hacking/index.htm>.
———. « Comment faire l'Histoire de la statistique ? », *LINX*, vol. 1, nº 1, 1980, p. 181-191.
HALBWACHS Maurice, « La statistique en sociologie », dans Centre international de synthèse, *La statistique, ses applications, les problèmes qu'elle soulève*, Paris: Presses Universitaires de France, 1944, Septième semaine internationale de synthèse, 3-8 juin 1935, p. 113-134.
HAMEL Jacques, « Les théories sociologiques comme expression de l'appropriation culturelle, un point de vue critique », *Canadian Review of Sociology/Revue canadienne de sociologie*, vol. 58, nº 2, 2021, p. 271-280.
———. « Décrire, comprendre et expliquer. Réflexions et illustrations en sociologie », *SociologieS*, Théories et recherches, octobre 2006, p. 1-14. En ligne au< https://journals.openedition.org/sociologies/132>.
———. « A propos de l'échantillon, de l'utilité de quelques mises au point », *BMS: Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, nº 67, 2000, p. 25-41.
———. « Pour la méthode de cas. Considérations méthodologiques et perspectives générales », *Anthropologie et Sociétés*, vol. 13, nº 3, 1989, p. 59-72.
HAND David J., « Data Mining: Statistics and More? », *The American Statistician*, vol. 52, n° 2, 1998, p. 112-118.

HÉRAN François, « L'assise statistique de la sociologie », *Economie et Statistique*, vol. 168, nº 1, 1984, p. 23-35.
HASTIE Trevor, TIBSHIRANI Robert et Jerome H. FRIEDMAN, *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, 2nd ed., New York, Springer, 2009.
HORVATH Robert A., « Épistémologie et méthodologie de la statistique en relation avec la sociologie », *Journal de la société statistique de Paris*, vol. 130, n°4, 1989, p. 209-218.
HOULE Gilles, « Le sens commun comme forme de connaissance », *Sociologie et sociétés*, vol. 19, nº 2, 1987, p. 77-86.
———. « La sociologie : une question de méthode ? », *Sociologie et sociétés*, vol. 14, nº 1, 1982, p. 3-6.
———. « L'idéologie : un mode de connaissance », *Sociologie et sociétés*, vol. 11, nº 1, 1979, p. 123-145.
HOULE Gilles et Nicole RAMOGNINO, « La construction des données », *Sociologie et sociétés*, vol. 25, nº 2, 1993, p. 5-9.
JAMES Gareth, Daniela WITTEN, Trevor HASTIE et Robert TIBSHIRANI, *An Introduction to Statistical Learning: with Applications in R*, New York, Springer, 2013.
JATON Florian, « "Pardonnez cette platitude" : de l'intérêt des ethnographies de laboratoire pour l'étude des processus algorithmiques », *Zilsel*, vol. 1, nº 5, 2019, p. 315-339.
JAUNAIT Alexandre et Sébastien CHAUVIN, « Représenter l'intersection », *Revue francaise de science politique*, vol. 62, nº 1, 2012, p. 5-20.
JAVEAU, Claude, « De l'homme moyen à la moyenne des hommes : l'illusion statistique dans les sciences sociales », dans DE COOREBYTER Vincent (dir.), *Rhétoriques de la science*, Paris : Presses Universitaires de France, 1994, p. 53-67.
KITCHIN Rob, « Thinking Critically about and Researching algorithms », *Information, Communication & Society*, vol. 20, nº 1, 2017, p. 14-29.
———. « Big Data, New Epistemologies and Paradigm Shifts », *Big Data & Society*, vol. 1, nº 1, 2014, p. 1-12.
LACOURSE Éric, Charles-Édouard GIGUÈRE et Véronique DUPÉRÉ, « Algorithmes d'apprentissage et modèles statistiques: Un exemple de régression logistique régularisée et de validation croisée pour prédire le décrochage scolaire », dans Marc CORBIÈRE et Nadine LARIVIÈRE (dir.), *Méthodes qualitatives, quantitatives et mixtes dans la recherche en sciences humaines, sociales et de la santé*, 2e éd., Québec, PUC, 2020, p. 581-612.
LAHIRE Bernard, « La variation des contextes dans les sciences sociales. Remarques épistémologiques », *Annales. Histoire, Sciences Sociales*, vol. 51, nº 2, 1996, p. 381-407.
LEBART Ludovic, Marie PIRON et Alain MORINEAU, *Statistique exploratoire multidimensionnelle*, Paris, Dunod, 1995.
LECLERC Olivier, « Statistiques et normes : jalons pour une rencontre interdisciplinaire », *Cahiers Droit, Sciences & Technologies*, n° 4, 2014, p. 37-44.
LÉGER Christian, « La valeur-p sous surveillance », *Bulletin AMQ*, LVI, nº 4, 2016, p. 77-85.
LEMERCIER Claire et Claire ZALC, « Des corrélations aux causalités ? » (chapitre V), dans *Méthodes quantitatives pour l'historien*, Paris : La Découverte, 2008, p. 58-79.
MACÉ Yannick, « L'approche statistique : entre réalité(s) et subjectivité », *Journal de la société française de statistique*, vol. 147, nº 4, 2006, p. 85-102.
MACKENZIE Adrian, « The Production of Prediction: What Does Machine Learning Want? », *European Journal of Cultural Studies*, vol. 18, nº 4-5, 2015, p. 429-445.
MARTIN Olivier, « Mathématiques et sciences sociales au XXème siècle », *Revue d'Histoire des Sciences Humaines*, vol. 1, nº 6, 2002, p. 3-13.

———. « Raison statistique et raison sociologique chez Maurice Halbwachs », *Revue d'histoire des sciences humaines*, vol. 1, nº 1, 1999, p. p.69-101.

MAYER-SCHNBERGER Viktor et Kenneth CUKIER, *Big Data: A Revolution That Will Transform How We Live, Work and Think,* Londres, John Murray, 2013.

MÉADEL Cécile et Guillaume SIRE, « Les sciences sociales orientées programmes », *Réseaux*, n° 206, nº6, 2017, p. 9-34.

MERLLIÉ, Dominique. « Le travail des catégories statistiques », *Sociétés Contemporaines*, vol. 14, nº 1, 1993, p. 149-63.

MERON Monique, « Statistiques ethniques : tabous et boutades », *Travail, genre et sociétés*, vol. 1, n° 21, 2009, p. 55-68.

MOLINA Mario et Filiz GARIP, « Machine Learning for Sociology », *Annual Review of Sociology*, vol. 45, nº 1, 2019, p. 27-45.

MOULIN Stéphane, « Classification », dans Frédéric BOUCHARD, Pierre DORAY et Julien PRUD'HOMME (dir.), *Sciences, technologies et sociétés de A è Z*, Presses de l'Université de Montréal, 2015, p. 43-46.

———. « Présentation : la statistique en action », *Sociologie et sociétés*, vol. 43, nº 2, 2011, p. 5-15.

MOULIN Stéphane et Jean-Pierre BEAUD, « Quantification et mesure », dans Frédéric BOUCHARD, Pierre DORAY et Julien PRUD'HOMME (dir.), *Sciences, technologies et sociétés de A à Z*, Presses de l'Université de Montréal, 2015, p. 186-188.

NAKACHE Jean-Pierre et Josiane CONFAIS, *Statistique explicative appliquée*, Paris: Technip, 2003.

OLLION Étienne, « Les sciences sociales, contre la data science ? », *Regards croisés sur l'économie*, n° 23, nº 2, 2018, p. 77-86.

O'NEIL Cathy, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Crown, 2017. [*Algorithmes: la bombe à retardement*, Paris : Les Arènes, 2018.]

PARENT Frédéric et Paul SABOURIN, « Ethnographie et théorie de la description – La construction des données sociologiques », *Cahiers de recherche sociologique*, nº 61, 2016, p. 109-126.

———. « Présentation. Les espaces-temps de la production ethnographique », *Cahiers de recherche sociologique*, nº 61, 2016, p. 20.

PASSERON Jean-Claude, *Le Raisonnement sociologique: Un espace non poppérien de l'argumentation,* Paris : Albin Michel, 2006 [Nathan, 1991].

PHILLIPSON M., « La méthodologie conventionnelle: critique phénoménologique », dans Jean PADIOLEAU (dir.), *L'opinion publique*, De Gruyter Mouton, 1981, p. 84-96.

PIRÈS Alvaro P., « De quelques enjeux épistémologiques d'une méthodologie générale pour les sciences sociales », dans Jean POUPART *et al.* (dir.), *La recherche qualitative. Enjeux épistémologiques et méthodologiques*, Montréal, Gaëtan Morin, 1997, p. 3-54 et 50-68.

———. « Deux thèses erronées sur les lettres et les chiffres », *Cahiers de recherche sociologique*, vol. 5, nº 2, 1987, p. 85-105.

PORTER Theodore M., *La confiance dans les chiffres: La recherche de l'objectivité dans la science et dans la vie publique*, Paris: Les Belles Lettres, 2017 [1995].

RAKOTOMALALA Ricco, « Arbres de Décision », *Revue MODULAD*, nº 33, 2005, p. 163-187.

REY Olivier, *Quand le monde s'est fait nombre*, Paris : Stock, 2016.

RIANDEY Benoît, Laurent TOULEMON et Jacqueline FELDMAN, « L'utilisation de la regression logistique dans les enquêtes », *BMS: Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, nº 33, 1991, p. 79-93.

RIEDER Bernhard, « Scrutinizing an Algorithmic Technique: the Bayes Classifier as Interested Reading of Reality », *Information, Communication & Society*, vol. 20, nº 1, 2017, p. 100-117.
ROBERGE, Jonathan et Robert SEYFERT (dir.). « What Are Algorithmic Cultures? », dans *Algorithmic Cultures. Essays on Meaning, Performance and New Technologies*, London: Routledge, 2016, p. 1-25.
ROUANET Henry et Frédéric LEBARON, « La preuve statistique : examen critique de la régression », Séminaire « Qu'est-ce que *Faire preuve* ? », Curapp, 5 mai 2006, p.1-21.
ROUANET Henry, Frédéric LEBARON, Viviane LE HAY, Werner ACKERMANN et Brigitte LE ROUX, « Régression et analyse géométrique des données : réflexions et suggestions », *Mathématiques et sciences humaines*, vol. 40, nº 160, 2002, p. 13-45.
ROUVROY Antoinette et Thomas BERNS, « Gouvernementalité algorithmique et perspectives d'émancipation », *Réseaux*, n° 177, nº 1, 2013, p. 163-196.
———. « Le nouveau pouvoir statistique », *Multitudes*, n° 40, nº 1, 2010, p. 88-103.
ROY Simon N., « L'étude de cas », dans Benoît GAUTHIER (dir.), *Recherche sociale: De la problématique à la collecte des données*, 5e ed., Québec, Les Presses de l'Université du Québec, 2009, p. 199-225.
SABOURIN Paul, « Sociologie, éthique et politique : itinéraire d'une éthique dans la recherche pour une coopération sociologique élargie », *Sociologie et sociétés*, vol. 52, nº 1, 2020, p. 19-46.
———. « L'analyse de contenu », dans Benoît GAUTHIER (dir.), *Recherche sociale: de la problématique à la collecte des données*, 5e éd., Québec: Presses de l'Université du Québec, 2009, p. 415-444.
———. « La régionalisation du social. Une approche de l'étude de cas en sociologie », *Sociologie et sociétés*, vol. 25, nº 2, 1993, p. 69-91.
SAPORTA Gilbert, *Probabilités, analyse des données et statistique*, Paris: Editions Technip, 2006.
SADIN, Éric, *La Vie algorithmique : critique de la raison numérique*, Paris : L'Échappée, 2015.
SAUVÉ Mathieu-Robert, « Une orientation Science des données s'ajoute au baccalauréat en mathématiques et informatique », *UdeMNouvelles* [Forum], 12 décembre 2018. En ligne au < https://nouvelles.umontreal.ca/article/2018/12/12/une-orientation-science-des-donnees-au-baccalaureat-en-mathematiques-et-informatique/>.
SAVOIE-ZAJC Lorraine, « Comment peut-on construire un échantillonnage scientifiquement valide? », *Recherches Qualitatives - Hors Série*, nº 5, 2007, p. 99-111.
SEAVER Nick, « Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems », *Big Data & Society*, vol. 4, nº 2, 2017, p. 1-12.
SELZ Marion et Florence MAILLOCHON, *Le raisonnement statistique en sociologie*, Paris : Presses universitaires de France (PUF), 2009.
SIRACUSA Jacques, « Justifier et préciser l'interprétation des données statistiques », *BMS*, nº 118, 2013, p. 60-72.
SIMON Patrick, « Les statistiques, les sciences sociales françaises et les rapports sociaux ethniques et de « race » », *Revue française de sociologie*, vol. 49, nº 1, 2008, p. 153-162.
SINGLY, François de, « Les bons usages de la statistique dans la recherche sociologique », *Économie et statistique*, vol. 168, nº 1, 1984, coll. « Sociologie et statistique », p. 13-21.
STATISTIQUE CANADA. Enquête sociale générale de 2016 – Les Canadiens au travail et à la maison (cycle 30): Guide de l'utilisation et Dictionnaire des données - Fichier de microdonnées à grande diffusion, Canada, Statistique Canada, 2016.
STROBL Carolin, James MALLEY et Gerhard TUTZ, « An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests », *Psychological Methods*, vol. 14, nº 4, 2009, p. 323-348.

SUPIOT Alain, *La Gouvernance par les nombres: cours au Collège de France*, 2012-2014, Paris, Fayard, 2015.
THÉVENOT Laurent, « Mesure pour mesure : formes d'enquête, d'évaluation et de gouvernement, depuis la statistique d'État jusqu'au « soi quantifié » », dans le cadre du colloque « Histoire aujourd'hui, statistiques demain : regards croisés sur la production et l'usage des statistiques », Centre de conférences Pierre-Mendès-France à Paris, 29 juin 2016, p. 1-26.
———. « Statistique et politique. La normalité du collectif », *Politix*, vol. 1, n° 25, 1994, p. 5-20.
TREMBLAY André, « Feu la société globale et les méthodes quantitatives : de nouveaux termes pour un ancien débat? », *Cahiers de recherche sociologique*, , nº 28, 1997, p. 63-88.
TUFFÉRY Stéphane, *Data mining et statistique décisionnelle: la science des données*, 5e éd., Paris, Éditions Technip, 2017.
VALLET Louis-André, « Sur l'analyse de régression en sociologie », Communication au RT20 (Méthodes) au congrès de l'Association Française de Sociologie, Bordeaux, 5-8 septembre 2006, p. 1-13.
VAN CAMPENHOUDT Luc, Jacques MARQUET et Raymond QUIVY, *Manuel de recherche en sciences sociales*, 5e éd., Paris: Dunod, 2017.
VAYRE Jean-Sébastien, « Comment décrire les technologies d'apprentissage artificiel ? », *Réseaux*, n° 211, nº 5, 2018, p. 69-104.
VOLLE, Michel. « Enjeux de la statistique », Etudes, vol. 356, no 1, 1982, p. 45-60.
WASSERSTEIN Ronald L. et Nicole A. LAZAR, « The ASA Statement on p-Values: Context, Process, and Purpose », *The American Statistician*, vol. 70, nº 2, 2016, p. 129-133.
WING Jeannette M., « Computational Thinking », *Communications of the ACM*, vol. 49, nº 3, mars 2006, p. 33-35.
ZIEWITZ Malte, « Governing Algorithms: Myth, Mess, and Methods », *Science, Technology, & Human Values*, vol. 41, nº 1, 2016, p. 3-16.

La source pour tous les tableaux et figures est Statistique Canada.

Annexes A

A1 – Liste des variables (Enquête sociale générale de 2016[500])

	Individu (Nom fictif)				Gisèle	Lise	Germain
	No de l'observation	i		avec $i = 1, \ldots, N$	517	9 213	19 245
X_p	Étiquette (Variable)			Valeurs / Modalités			
00	Santé perçue générale[501]	Y					
			0	Négative		X	
			1	Positive	X		X
01	Sexe du répondant	SEX					
			1	Homme			X
			2	Femme	X	X	
02	Groupe d'âge	AGE					
			1	15-24 ans			
			2	25-34 ans		X	
			3	35-44 ans			X
			4	45-54 ans	X		
			5	55-64 ans			
			6	65-74 ans			
			7	75 ans et plus			
03	État matrimonial	MAR					
			1	Célibataire		X	
			2	Marié	X		X
			3	Union libre			
			4	Veuf/Séparé/Divorcé			
04	Statut d'immigrant	IMM					
			1	Né au Canada	X	X	X
			2	Né à l'extérieur			
05	Minorité visible	MIN					
			1	Pas minorité	X		X
			2	Minorité visible		X	
06	Niveau de scolarité	EDU					
	Plus haut niveau de scolarité atteint (certificat, diplôme ou grade)		1	Primaires			
			2	Secondaires		X	
			3	Collégiales			
			4	Universitaires	X		X

[500] L'Enquête sociale générale de 2016 – Les Canadiens au travail et à la maison (cycle 30) menée par Statistique Canada. La population de référence n'a pas été limitée à certains groupes particuliers dans la présente recherche, en sélectionnant des questions posées à tous les répondants de l'enquête.

[501] La variable dépendante – l'état de santé autoévalué en général (SRH_110) – fut dichotomisée en regroupant les états de santé perçus « excellent » et « très bon » d'un côté (Y=1) et ceux « bon », « passage » et « mauvais » de l'autre (Y=0). « En général, diriez-vous que votre santé est : « excellente », « très bonne », « bonne », « passable » ou « mauvaise » ? [échelle de mesure de cinq points] » (Questionnaire, 2016)

07	Revenu personnel total	REV					
	Revenu annuel - Groupé (avant impôt)		1	Moins de 25 000$		X	
			2	25 000$-49 999$			
			3	50 000$-74 999$	X		
			4	75 000$-99 999$			X
			5	100 000$ et plus			
08	Situation professionnelle	TRA					
	Emploi/sem. dern.		1	Emploi	X		X
			2	Pas d'emploi		X	
09	Classe sociale	PSC					
	Distinction sociale perçue		1	Inférieure		X	
			2	Moyenne			X
			3	Supérieure	X		
10	Liens sociaux	COM					
	Sentiment d'appartenance à la collectivité		1	Non			
			2	Oui	X	X	X
11	Environnement	ENV					
	Satisfaction - Qualité de l'environnement local		1	Non		X	
			2	Oui	X		X
12	Activités de loisir	ODA					
	Participation à des activités de plein air/12 dern. mois		1	N'a pas participé			
			2	Au moins une	X	X	X
13	Habitudes alimentaires	FOO					
	Manger à l'extérieur/ acheter des plats à emporter - Dernier mois		1	Pas dern. mois			
			2	Quelques fois/mois		X	X
			3	Plusieurs fois/sem.	X		
14	Consommation d'alcool	DDR					
	(de boissons alcoolisées)		1	Jamais	X		
			2	Occasion		X	
			3	Régulier			X
15	Usage actuel du tabac	SMK					
			1	Non-fumeur	X	X	X
			2	Fumeur			
16	Niveau de stress dans la vie	SMG					
			1	Pas du tout			
			2	Pas tellement		X	
			3	Un peu	X		X
			4	Assez/Extrêmement			
17	Incapacité(s)	INC					
	(apprentissage, ouïe, mental/psycho., physique, vision)		1	Pas d'incapacité		X	
			2	Une ou plus	X		X
18	Indicateur de région	GEO					
			1	Grands centres urbains	X	X	X
			2	Régions rurales			

A2 – Modèle de régression multiple, « simple »

Modèle simple comprenant trois variables indépendantes (M_1) : Équation de régression permettant « de reconstruire une approximation des données »[502].

$$\frac{\mathrm{p}}{1-\mathrm{p}} = 1 * 1{,}48_{\text{PSC:Moy.}} * 2{,}57_{\text{PSC:Sup.}} * 0{,}53_{\text{DOS:Non}} * 0{,}33_{\text{INC:1}}$$

$$ln\left(\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right) = -0{,}0035 + 0{,}3928_{\text{PSC:Moy.}} + 0{,}9433_{\text{PSC:Sup.}} - 0{,}6261_{\text{DOS:Non}} - 1{,}1215_{\text{INC:1}}$$

$$Pr(Y=1) = 1/(1+exp^{-(-0{,}0035+0{,}3928*\text{PSC:Moy.}+0{,}9433*\text{PSC:Sup.}-0{,}6261*DOS:Non-1{,}1215*\text{INC:1})})$$

Combinaisons possibles : 3 (modalités pour la variable « PSC ») * 2 (modalités pour la variable « DOS ») * 2 (modalités pour la variable « INC ») = 12 combinaisons/cas de figure (« types idéaux ») possibles

Situation de référence : Si PSC : Inf. et DOS : Oui et INC : 0,

p = 0,9965 / 1,9965 = 0,5 soit 50 %

$$\Pr(Y=1) = \frac{exp(-0{,}0035)}{1+exp(-0{,}0035)} = 0{,}50009$$

Variable		N	Odds ratio	p
CLASSE	Inf.	3188		Reference
	Moy.	11783		1.48 (1.36, 1.61) <0.001
	Sup.	3441		2.57 (2.31, 2.86) <0.001
SENT_APP	Oui	14915		Reference
	Non	3497		0.53 (0.49, 0.58) <0.001
INCAP	0	13239		Reference
	1	5173		0.33 (0.30, 0.35) <0.001
(Intercept)				1.00 (0.92, 1.09) 0.9

0.5 1 1.5 2 2.5

Le tableau ci-dessous compare LR et DT

→ Exemple simple pour comprendre l'idée de configurations, de croisements « improbables »

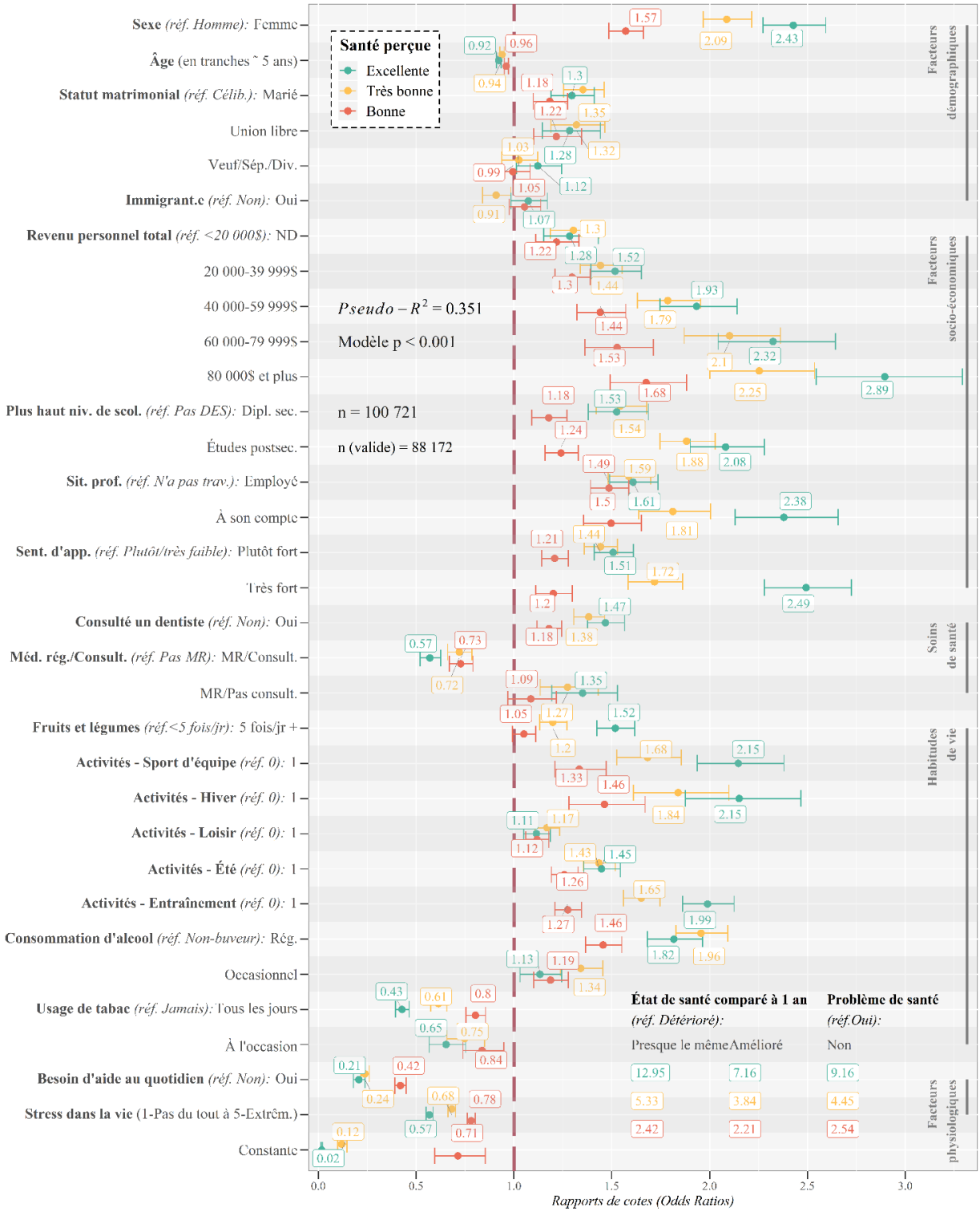
	PSC	DOS	INC	MS P(Y=1)	Effectif (n)	Cas	DT P(Y=1)	Règle	Y	Effectif (n)
1	Inf.	Non	1	14,88%	620		29%	2	0	5 207
2	Moy.	Non	1	20,57%	738					
3	Sup.	Non	1	30,99%	139					
4	Inf.	Oui	1	24,64%	834					
5	Moy.	Oui	1	32,62%	2310	Gilbert				
6	Sup.	Oui	1	45,64%	566	Gisèle				
7	Inf.	Non	0	34,92%	497		35%	12	0	497
8	Moy.	Non	0	44,28%	1220		45%	26	0	1 220
9	Sup.	Non	0	57,95%	304	Dora	58%	27	1	304
10	Inf.	Oui	0	50,09%	1261	Lise (réf.)	58%	14	1	8 812
11	Moy.	Oui	0	59,78%	7551					
12	Sup.	Oui	0	72,05%	2443		72%	15	1	2 443

[502] CIBOIS Philippe, Les méthodes d'analyse d'enquêtes, op. cit., p. 86.

A3 – Exemples de l'étude des perceptions de santé, « multi-classes »

Figure 8. – Régression logistique multinomiale : Modèle final

Catégorie de réf. « Passable ou mauvaise »



Source: Statistique Canada. Enquête sur la santé dans les collectivités canadiennes (ESCC), 2013-2014

Figure 9. – Arbre de décision avec données manquantes (n=100 597)

L'un des modèles CART pour prédire la santé perçue des Canadien.ne.s âgé.e.s de 18 à 74 ans