

LA MODÉLISATION EN SCIENCES SOCIALES: INCERTITUDES ET DÉFIS

Guillaume Wunsch; Michel Mouchart and
Federica Russo

DISCUSSION PAPER | 2019 / 03

La modélisation en sciences sociales: incertitudes et défis

Guillaume WUNSCH^a, Michel MOUCHART^b, Federica RUSSO^c

^a Démographie, UCLouvain

^b Institut de Statistique, Biostatistique et Sciences Actuarielles
(ISBA), UCLouvain

^c Philosophie des sciences, Université d'Amsterdam

13/03/2019

Table des matières

- 1 Introduction
- 2 Le modèle comme représentation de la réalité
- 3 Expérimentation et Sciences sociales
- 4 Du modèle à l'explication scientifique
- 5 Cadre conceptuel, cadre opérationnel et modélisation statistique
- 6 Un modèle ou des modèles ?
- 7 "Big data" et causalité
- 8 Modèle, explication, intervention politique
- 9 Quelques remarques en guise de conclusion

1 Introduction

Comme toute science, les sciences sociales cherchent à expliquer et à comprendre les phénomènes relevant de leur objet, ici les phénomènes sociaux. Quelles sont les causes du terrorisme ? Pourquoi l'économie allemande est-elle plus performante que celle de la France ? Quelles sont les raisons des différences de mortalité observées entre les trois Régions de la Belgique ? ... Voilà une série de questions judicieuses que l'on peut se poser. Il s'agit chaque fois d'une *explanation-seeking why question*, selon Carl Hempel (1965), c'est-à-dire d'une question sur le 'pourquoi' des choses, requérant une explication. Pour tenter d'y répondre, les scientifiques ont mis au point diverses approches et méthodes quantitatives ou qualitatives. Quelle que soit la méthodologie, et elles sont nombreuses en sciences sociales, il s'agira toujours de confronter ses hypothèses, théories et modèles aux faits. L'observation des faits et l'évaluation de la pertinence de leur interprétation constituent ainsi un aspect important de la démarche scientifique.

L'objet de cet article n'est pas de présenter un tour d'horizon des différentes approches scientifiques adoptées par les sciences sociales. Il faudrait un très gros volume pour le faire : voir par exemple l'ouvrage bien connu de Madeleine Grawitz (2000) à ce sujet. Plus modestement, ce texte vise à montrer, dans une perspective causale, les particularités, les incertitudes et les défis que soulève la modélisation en sciences sociales. Pour saisir plus aisément nos propos, des exemples seront fournis tout au long du présent article. Le premier auteur étant démographe, les exemples seront principalement tirés des sciences de la population.

Nous traiterons d'abord brièvement, dans la section 2, du modèle en tant que représentation de la réalité. La section 3 examinera les différences entre l'approche expérimentale et l'approche non-expérimentale, sur le plan de la causalité. La section 4 portera ensuite sur la relation entre modèle et explication scientifique. La section 5 examinera un exemple pratique de modélisation en démographie, tandis que la section 6 posera la question de l'unicité ou non du modèle. La section 7 traitera d'un problème très actuel : du bon usage du *Big Data* dans la modélisation en sciences sociales. La section 8 examinera la relation entre explication, modèle et intervention politique. L'article se terminera par quelques remarques en guise de conclusion.

2 Le modèle comme représentation de la réalité

La réalité¹ sociale est complexe ; de plus, elle est variable dans le temps et l'espace. Pour appréhender les phénomènes sociaux, il est nécessaire de simplifier la réalité et de n'en retenir que les traits essentiels. On construit de ce fait un modèle censé la représenter, mais cette représentation est forcément imparfaite. Tout modèle est donc nécessairement incomplet et, dans ce sens, il est faux vis-à-vis de la réalité. Par exemple, divers modèles ou théories (les deux termes sont

¹ Nous n'ouvrons pas ici le débat sur la définition du concept de réalité.

considérés ici comme équivalents²) de la migration s'inspirent des lois de la gravitation : les migrants quitteraient leurs aires de départ, attirés par des aires présentant plus d'opportunités. Mais ce n'est qu'une cause de migration parmi d'autres, comme le montrent notamment les déplacements de population en cas de conflits ou de catastrophe naturelle. Par ailleurs, de quelles opportunités s'agit-il ? Et pourquoi certaines personnes migrent-elles et d'autres pas ? Le modèle soulève autant de questions que celles qu'il essaie de résoudre ! Par ailleurs, un modèle ou théorie des déterminants de la fécondité va différer selon qu'on étudie la natalité du 17^{ème} ou du 21^{ème} siècle, de l'Europe ou de l'Afrique. Les phénomènes sociaux sont souvent plus variables que les phénomènes étudiés par les sciences de la nature. Un modèle ou théorie, dans les sciences sociales, ne peut avoir de vocation universelle : il n'y a pas de lois ici, à la différence de la biologie des individus composant la société.³

La représentation des phénomènes sociaux peut être faite en termes mathématiques ; voir à ce sujet l'excellent petit livre de Jean Mahwin (2017). C'est ainsi qu'Alfred Lotka a établi son modèle de *population stable*⁴, à fécondité et mortalité par âge constantes et à migration nulle. Mais un modèle ou théorie peut aussi bien être exprimé sous une forme verbale (comme les théories de Marx ou de Weber) que sous une forme graphique (voir la section 5). Le mode de représentation peut varier d'un objet à l'autre, et une formalisation mathématique n'est pas toujours la meilleure solution.

L'approche *quantitative* en sciences sociales s'adresse souvent aux grands effectifs à nombreuses données. Cette approche n'éclaire toutefois pas suffisamment la dynamique interne des groupes ; pour ce faire, les approches *qualitatives* sur petits effectifs sont plus adaptées. Ces dernières ne seront pas examinées ici.

On peut distinguer deux types de modèles, *prédictif* et *explicatif*, chacun ayant une utilité manifeste mais différente. Les scientifiques ont fait apparaître, dans les années '50, une forte association statistique entre l'usage de la cigarette et le cancer du poumon. L'explication biologique n'a été développée que plus tard. On a ainsi établi un modèle prédictif du cancer du poumon sans disposer pour autant d'une explication cohérente de sa relation avec le tabagisme, même si certains processus étaient considérés comme plausibles. Le modèle prédictif, basé sur cette simple association, a en effet conduit à décourager de plus en plus la consommation de tabac. Le modèle explicatif n'est apparu que plus tard, à la suite de recherches épidémiologiques (en analyse longitudinale⁵) et médicales.

² La littérature philosophique s'interroge sur le statut des théories et des modèles, et sur leurs différences, mais pour les sciences sociales la distinction n'est pas nécessairement marquée (voir notamment Russo, 2017). Ainsi, le prix Nobel d'économie Herbert Simon écrit que, dans l'usage contemporain, le terme « modèle » est simplement un synonyme de « théorie » (Simon et Newell, 1956). Pour d'autres, une théorie serait plus vaste qu'un modèle.

³ La stabilité des lois biologiques est toutefois relative : elles peuvent évoluer à long terme sous l'influence de la sélection naturelle et de l'environnement.

⁴ Mais non nécessairement *stationnaire*.

⁵ La célèbre *British Doctors Study* de Doll, Hill et Peto sur la période 1951-2001, auprès de quarante mille médecins.

Dans les années '70, des économètres ont constaté que des modèles explicatifs même élaborés, tels que des modèles à équations simultanées, donnaient de moins bonnes prévisions que des modèles purement prédictifs même grossiers. Un modèle explicatif en sciences sociales ne donne pas toujours de meilleures prévisions qu'un modèle prédictif. En contrepartie, un modèle explicatif permet de mieux comprendre le fonctionnement d'un système de variables et peut conduire ainsi à des interventions politiques plus efficaces pour son évolution. Un exemple de modèle explicatif sera esquissé à la section 5.

3 Expérimentation et Sciences sociales

Pour expliquer et comprendre un phénomène social, telle que la différence de fécondité entre groupes sociaux en Belgique, on peut partir tout d'abord des différences observées entre les groupes. On remarque ainsi que de nombreuses autres variables diffèrent aussi entre les groupes sociaux: niveau d'instruction, revenu, profession, origine ethnique, habitat, *etc.* Ceci suggère que de nombreux mécanismes et sous-mécanismes peuvent opérer pour engendrer les comportements de fécondité. Pour comprendre le phénomène étudié, nous devons simplifier le problème en privilégiant certaines variables au détriment d'autres. On peut penser, par exemple, que le revenu est un facteur-clé dans l'explication des différences de fécondité⁶ mais comment le montrer ? Dans une approche expérimentale appelée 'essai randomisé', telle que celle pratiquée pour tester de nouveaux médicaments, on prendrait un échantillon aléatoire de n personnes (par exemple mille individus) d'âges féconds. La moitié recevrait un revenu substantiel et l'autre moitié un salaire de misère, et on comparerait ensuite, à la longue, la fécondité des groupes. Cette approche n'est évidemment guère réalisable ici, tant pour des raisons éthiques que pratiques. Elle est toutefois utilisée en psychologie ou parfois en économie notamment.

Il faut donc choisir le plus souvent une voie non-expérimentale, basée sur l'observation des faits dans le monde réel. Sur un graphique, on peut d'abord voir si le revenu et la fécondité sont associés. Si c'est le cas, on peut ensuite ajuster une fonction mathématique à cette relation, appelée *équation de régression*, permettant d'estimer le niveau de fécondité à partir d'un niveau de revenu. Toutefois, l'association entre revenu et fécondité peut être partiellement due à d'autres variables. Par exemple, le niveau d'instruction ou l'origine ethnique influencent à la fois le niveau de revenu et celui de la fécondité, par d'autres voies que le revenu dans ce dernier cas. Il faut donc « contrôler » ces autres variables, appelées *variables de confusion*, pour déduire la relation correcte entre revenu et fécondité, c'est-à-dire étudier la relation revenu-fécondité indépendamment de l'effet de l'instruction et de l'origine ethnique. A chaque niveau d'instruction et dans chaque groupe ethnique on examinera la relation revenu-fécondité ; celle-ci ne dépendra donc plus des

⁶ D'où les allocations familiales et de naissance utilisées dans certains pays européens pour favoriser la natalité.

différences d'instruction ou d'ethnie⁷. Une variable de confusion est donc un facteur commun aux causes et à leurs effets éventuels.

Prenons un exemple plus simple, pour mieux faire apparaître le problème. La mortalité en Afrique, mesurée par le *taux brut de mortalité*, est nettement inférieure à celle de l'Europe. Ce taux (nombre de décès annuels pour mille habitants) est en effet de 9 ‰ en Afrique mais il est de 11 ‰ en Europe. Serait-ce une raison pour inverser les flux migratoires et partir en Afrique vivre plus longtemps ? Non, bien sûr, car la durée de vie est bien plus courte en moyenne en Afrique qu'en Europe : 64 ans (pour les femmes) contre 82 ans en Europe (PRB Data Sheet, 2018). Ce paradoxe s'explique par la structure par âge très différente entre les deux régions, qui agit ici comme variable de confusion : la population européenne est nettement plus vieille que la population africaine, et c'est aux âges élevés que la mortalité est la plus forte. Si l'on contrôle cette structure par âge⁸, c'est-à-dire si on neutralise l'effet des différences de structures, les taux de mortalité seraient respectivement de 16 ‰ pour l'Afrique et de 12 ‰ pour l'Europe, renversant donc le sens de l'inégalité.

Un modèle explicatif basé sur les données d'observation doit donc toujours contrôler les variables de confusion pour déduire d'éventuelles relations causales débarrassées de l'influence perturbatrice de ces variables. On se heurte toutefois ici à un problème majeur : on ne connaît souvent pas toutes les variables de confusion ou on ne dispose pas des données nécessaires à leur sujet. Dans cette situation, les associations observées entre causes possibles et effets éventuels peuvent être fallacieuses sur le plan de la causalité. Dans un tel cas, les économètres parlent d'une *perte d'exogénéité*.

Une approche expérimentale serait donc à préférer sur ce point car l'essai randomisé, par exemple, échappe à ce problème. En contrepartie, un essai randomisé ne modélise pas le mécanisme explicatif et souvent en présuppose l'existence. Par exemple, un essai randomisé peut indiquer que l'aspirine soulage généralement le mal de tête mais ne permet pas d'explicitier le mécanisme sous-jacent. Par ailleurs, toute expérimentation comporte une intervention sur le mécanisme générant les données. Ceci a soulevé une critique de la part de Robert Lucas (1976), qui a attiré l'attention sur le fait qu'une telle intervention peut modifier d'autres sous-mécanismes que celui visé par l'intervention.

Un bel exemple correspondant à la critique de Lucas est l'intervention de santé publique réalisée il y a quelques décennies en Carélie du nord (Finlande) pour réduire la très forte morbidité et mortalité par maladies cardio-vasculaires (MCV) dans cette région ; voir notamment E. Vartiainen (2018). Une série de mesures favorables à la santé (notamment dans le domaine alimentaire) furent

⁷ On 'conditionne' donc sur les variables de confusion. Pour un examen approfondi de cette question, voir Mouchart, Wunsch et Russo (2016).

⁸ En prenant par exemple les taux de mortalité des populations stationnaires correspondant aux tables de mortalité des deux régions, soit l'inverse des vies moyennes.

prises dans cette région-cible et l'évolution des MCV fut ensuite comparée à celle de la région-témoin de Kuopio, où une intervention similaire ne fut pas effectuée. On pouvait s'attendre à une amélioration de la santé dans la région-cible, à la suite des mesures politiques prises, et guère de progrès par contre dans la région témoin. La politique eut certes un effet, car les facteurs de risque se réduisirent considérablement dans la région-cible. Mais l'évolution fut parallèle dans la région-témoin ! La population de celle-ci, au courant de l'action entreprise dans la région-cible, modifia également son comportement dans le sens favorable à la santé.

On voit donc que l'expérimentation, à elle seule, ne permet pas de conclure quant aux relations de cause à effet. Elle doit toujours être accompagnée d'une démarche interprétative tout au long du processus de modélisation.

4 Du modèle à l'explication scientifique

Un modèle explicatif en sciences sociales cherche non seulement à fournir une description (simplifiée) du système social étudié mais aussi à examiner le mécanisme, sous forme probabiliste, censé engendrer les faits observés. L'accent est mis, en particulier, sur l'architecture ou structure de ce système, c'est-à-dire sur l'identification des sous-mécanismes structurant le mécanisme principal. Pour ce faire, un ingrédient essentiel est la prise en compte de l'état des connaissances, basé sur des observations situées dans un horizon temporel et un environnement spatial définis. Insistons sur le fait que, pour nous, l'identification du mécanisme a valeur épistémologique plutôt qu'ontologique : il s'agit de chercher à comprendre *comment* certaines causes conduisent à certains effets. Nous ne souscrivons donc pas à la vue mécaniciste du 17^{ème} siècle, selon laquelle le fonctionnement de la société (ou de ses composantes) se réduirait à des mécanismes qui ressembleraient à ceux d'une montre.

L'analyse du fonctionnement d'un système se base notamment sur l'étude des caractéristiques, ou paramètres, de chaque sous-mécanisme postulé. Bien que considérés comme fixés à un moment donné de l'analyse, ces paramètres sont eux-mêmes le résultat du fonctionnement de sous-mécanismes ayant œuvré dans le cadre d'un horizon temporel et spatial défini. Comme on l'a signalé à la section 2, les phénomènes sociaux sont instables dans le temps et l'espace, en fonction notamment de changements dans les normes et valeurs de la société et des individus qui la composent. La structure du modèle explicatif et la valeur des paramètres peuvent l'être également et l'horizon spatio-temporel doit être réévalué en conséquence. Cette question sera réexaminée à la section 6.

Considérons l'exemple relativement simple suivant. Quels sont les facteurs qui déterminent le fait qu'un immigrant soit perçu comme illégal ou légal, en dehors des critères purement juridiques ? Deux sociologues américains (Flores et Schachter, 2018) ont étudié la question en procédant d'abord à un inventaire des différents facteurs relevés dans la littérature scientifique. Des variables

comme le statut professionnel ou l'origine ethnique, parmi d'autres, semblent participer à la construction des stéréotypes sociaux à ce sujet. Ces chercheurs ont ensuite considéré que ces différents facteurs ne jouaient pas seulement indépendamment les uns des autres mais que leur action était renforcée lorsqu'ils agissaient en conjonction. En fonction de cette hypothèse explicative, ils ont dès lors établi une série de profils 'à risque' basé sur des variables telles que l'origine ethnique, le statut professionnel, le genre, l'âge, la connaissance de l'anglais, etc. Ces profils ont ensuite été soumis, par paires aléatoires, en 2017 à un échantillon représentatif de 1515 américains, blancs non-hispaniques. L'importance respective des diverses variables et profils a été estimée statistiquement. Il ressort de l'étude que les facteurs classiques, comme l'origine ethnique, contribuent à la formation des stéréotypes concernant l'illégalité du migrant : mieux vaut être italien que mexicain ! Mais l'étude montre aussi clairement que les stéréotypes sont renforcés par la conjonction des variables. Comme on pouvait s'y attendre, un homme mexicain de 40 ans, à faible niveau d'instruction, n'ayant pas d'activité formelle, et connaissant moyennement l'anglais, est très mal coté. Par contre, un homme italien du même âge, de bon niveau d'instruction, ayant une profession élevée, et parlant bien l'anglais, est très bien noté par l'échantillon. Ce modèle explicatif, basé sur la conjonction des caractéristiques, permet de mieux appréhender, pour l'époque et le pays considérés, la construction des stéréotypes sociaux relatifs à l'illégalité perçue des migrants.

5 Cadre conceptuel, cadre opérationnel et modélisation statistique

Voyons brièvement comment construire un modèle en sciences sociales. On consultera notamment avec intérêt, à ce sujet, le tableau historique des formes prises par la modélisation en sciences sociales présenté par Michel Armatte (2005), ainsi que le chapitre écrit par Federica Russo (2017). Il s'agira toutefois, dans le présent article, d'un modèle spécifique, à savoir le *modèle structurel causal* (MSC) utilisé pour vérifier, à partir de données d'observation, les relations causes-effets postulées entre des variables (voir par exemple Wunsch, Mouchart et Russo, 2015, chapitre 3).

Partant d'un inventaire des connaissances sur le sujet étudié, d'une analyse préliminaire des données et d'entretiens avec des experts, on construit un modèle des relations présumées entre les variables retenues, en distinguant les causes et les effets plausibles. Ce modèle peut notamment être représenté sous forme graphique ou mathématique. Il s'accompagne généralement d'une narration décrivant le mécanisme et les sous-mécanismes en jeu. Dans ce sens, les scientifiques sont aussi des narrateurs (Böhler, 2018) mais leurs narrations sont confrontées aux faits et ne sont pas de simples récits, comme on le verra dans l'étude ci-dessous. En d'autres termes, les données ne parlent pas d'elles-mêmes.

A titre d'exemple, Catherine Gourbin *et al.* (2017) ont étudié le recours à la contraception dans diverses grandes villes africaines. A partir d'une connaissance approfondie du domaine, dûment détaillée dans l'article cité, ils ont proposé un *modèle conceptuel* représenté par le graphe de la figure 1, où les flèches expriment de possibles influences causales entre les concepts concernés.

Pour tester le modèle, il faut disposer de variables traduisant en termes mesurables les concepts repris à la figure 1. A partir des données d'enquêtes réalisées dans les villes concernées par l'étude, les auteurs ont proposé un *modèle opérationnel* qui est construit cette fois à partir des variables mesurables. Ce modèle est également représenté graphiquement⁹ (figure 2). Dans ce graphe,

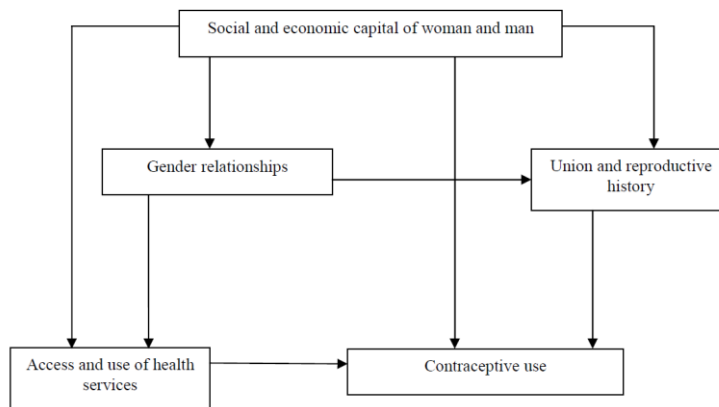


Fig. 1. *Modèle conceptuel de la contraception en Afrique (reproduit de Gourbin et al., 2017)*

chaque variable et ses causes directes représentent un sous-mécanisme spécifique au sein du mécanisme général reliant les variables entre elles. Ce modèle opérationnel a ensuite été testé à partir des données d'enquêtes, par régressions logistiques successives dans le cas présent. Celles-ci permettent de voir, dans une optique probabiliste, si les sous-mécanismes proposés sont statistiquement significatifs ou non. On a pu notamment vérifier si deux chemins indirects principaux proposés dans la littérature, allant du niveau d'instruction de la mère au recours à la contraception (notés en gras à la figure 2), étaient vérifiés ou non par les faits¹⁰, en contrôlant chaque fois les variables de confusion pertinentes.

Formalisons quelque peu la démarche esquissée ci-dessus. La modélisation statistique est basée sur une approche probabiliste, ou stochastique, des sous-mécanismes sous-jacents générant les données, appelés aussi processus générateur des données (PGD). Elle permet, entre autres, de mettre en relief la complexité des nombreux sous-mécanismes actifs dans le monde des sciences sociales.

⁹ Il s'agit d'un *graphe acyclique dirigé* où les variables sont ordonnées en fonction des sous-mécanismes postulés.

¹⁰ Un chemin (celui de droite à la figure 2) a été confirmé et l'autre (à gauche) infirmé par les faits.

Une seconde étape consiste à rechercher, à partir des connaissances existantes, un ordre entre ces variables d'intérêt, cet ordre représentant une succession de sous-mécanismes. Il s'agit, dans le jargon scientifique, d'une *décomposition récursive* du mécanisme en sous-mécanismes. Chaque variable est susceptible d'avoir été engendrée par un sous-mécanisme où les variables antérieures dans l'ordre seront considérées comme fixées. Plus formellement, chaque sous-mécanisme est représenté par une distribution conditionnelle aux variables antérieures à la variable engendrée par le sous-mécanisme¹¹.

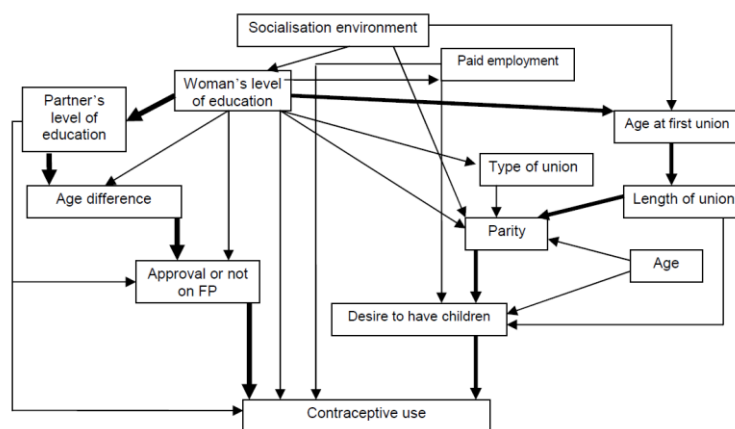


Fig. 2. Modèle opérationnel de la contraception en Afrique (reproduit de Gourbin et al., 2017)

Une troisième étape consiste à décider pour chaque sous-mécanisme quelles seraient les variables conditionnantes qui n'interviendraient pas effectivement. Les variables conditionnantes restantes pourront alors être considérées comme causales dans le sous-mécanisme concerné. Par exemple, à la figure 2, la variable 'désir d'avoir un enfant' est engendrée par un sous-mécanisme conditionné par les variables 'emploi rémunéré', 'parité', 'âge' et 'durée de l'union', les autres variables n'intervenant pas dans ce sous-mécanisme. Le modèle ainsi construit sera appelé un 'modèle structurel causal' (MSC).

6 Un modèle ou des modèles ?

Un bon modèle explicatif en sciences sociales devrait idéalement être unique en ce qui concerne la structure du mécanisme explicatif postulé, pour un contexte donné. Par exemple, un modèle censé représenter les déterminants de la fécondité en Belgique devrait être suffisamment général au

¹¹ Si on considère un vecteur de p variables d'intérêt, on pourra alors en décrire la distribution conjointe de la façon suivante : $P(X_1, \dots, X_p) = P(X_1) P(X_2 | X_1) \dots P(X_p | X_1, \dots, X_{p-1})$.

niveau de la population ; des modèles qui seraient spécifiques à chaque commune du pays, ou à la limite à chaque individu, n'auraient pas de portée explicative significative. Il faut en effet pouvoir distinguer l'essentiel de l'accessoire. Bien qu'il n'existe guère de lois en sciences sociales, on peut espérer néanmoins qu'un modèle s'applique à un large ensemble de cas au sein de la population de référence et qu'il nous permette aussi d'expliquer les différences éventuelles entre les sous-groupes, par exemple entre la Flandre et la Wallonie. Pour cela, il est nécessaire que l'état des connaissances soit tel qu'un seul modèle explicatif puisse être proposé. En se référant par exemple au modèle MSC élaboré à la section 5, il faudrait que la connaissance du domaine permette de formuler une décomposition unique en sous-mécanismes. Notons par ailleurs que la fidélité du modèle, par rapport à la réalité des mécanismes sous-jacents, est essentielle pour l'efficacité d'une intervention politique sur le système (voir section 8).

Malheureusement, ceci est rarement le cas. D'abord, l'état des connaissances est souvent imparfait (voir aussi à ce sujet la section 7) et les données nécessaires pour infirmer ou confirmer la théorie, manquantes ou déficientes. Ensuite, comme on l'a signalé antérieurement, les phénomènes sociaux sont variables dans le temps et l'espace à la suite de la modification de l'impact de certaines variables, ou de l'apparition ou la disparition d'autres. Un modèle explicatif doit donc être revu lorsque le contexte socio-historique se modifie. Il doit aussi fréquemment tenir compte de variables latentes, c'est-à-dire non-observées, et de données incomplètes. Dans ce cas, il n'y a pas nécessairement de « meilleur » modèle: il est possible que plusieurs modèles explicatifs soient non-infirmés par les faits et donc concurrents. Enfin, les comportements et attentes des individus sont souvent très hétérogènes, au sein de la population étudiée. Le modèle doit en tenir compte. C'est ainsi que les démographes s'intéressent aux femmes n'ayant pas eu d'enfants ou à celles ayant de grandes familles, et pas seulement aux femmes ayant donné naissance aux un ou deux enfants 'standards', lorsqu'ils étudient les déterminants de la fécondité européenne.

Les conditions sont moins sévères lorsqu'il s'agit d'un modèle purement prédictif, celui-ci étant basé sur la réalisation des effets des causes. Une connaissance précise du fonctionnement du mécanisme n'est pas indispensable dans ce cas. Malgré tout, son efficacité prédictive dépend crucialement de la stabilité des sous-mécanismes. Notons toutefois qu'un modèle prédictif donne parfois de meilleures prévisions qu'un (bon) modèle structurel pour des raisons de simplicité analytique et de stabilité sous-jacente. C'est ainsi qu'une décomposition de la mortalité selon les causes médicales de décès peut donner de moins bonnes prévisions de l'évolution de la mortalité qu'une projection 'toutes causes', l'évolution des décès par causes médicales étant plus volatile que celle pour l'ensemble des causes (Caselli, 2006).

Le contexte, qui limite la validité d'un modèle, est défini par un ensemble de variables souvent latentes. Une hypothèse implicite consiste à supposer que les variables définissant le contexte sont constantes pour la population de référence considérée ; pour plus de détails, voir Wunsch, Mouchart et Russo (2018). Le caractère latent rend la connaissance de la variation de ces variables (plus ou moins) impossible. Or une variation substantielle du contexte devrait impliquer une

modification du modèle explicatif. Il s'agit ici d'une des incertitudes importantes de la modélisation en sciences sociales. Une solution générale de ce problème ne semble guère possible; il faut donc être particulièrement prudent dans l'interprétation des résultats d'un modèle.

7 "Big data" et causalité

Dans cette section, nous verrons dans quelle mesure les *Big Data*, données massives ou mégadonnées, peuvent être utilisées avantageusement dans le cadre de la modélisation causale des faits sociaux. Les grands ensembles de données ont pris une place importante dans l'analyse des données au cours des dernières années, à la suite des possibilités croissantes de stockage et de traitement des larges bases de données par des ordinateurs de plus en plus puissants. Les sciences sociales, et la démographie entre autre, ont depuis longtemps analysé de grandes bases de données. Par exemple, les données des recensements, portant sur toute la population belge, ont fait l'objet de nombreuses analyses statistiques au cours des décennies passées, grâce au progrès des ordinateurs. Ce qui est neuf, désormais, est la grande variété des sources de données accessibles et celle de leur contenu, les données pouvant être structurées ou non. Par exemple, les archives hospitalières contiennent notamment des données numériques provenant d'examens de laboratoire effectués auprès des patients mais aussi des rapports médicaux sur leur état de santé. De plus en plus d'informations sont également reprises des réseaux sociaux et de l'usage de l'internet, Google ayant par exemple essayé de prévoir les épisodes de grippe à partir des questions posées sur le web (Lazer *et al.*, 2014). De nouveaux programmes informatiques peuvent désormais traiter ces divers types de données.

Deux grandes catégories de données massives doivent être distinguées : soit les données peuvent être appariées, par exemple par un numéro d'identification personnel (NIP), soit les données ne peuvent pas être appariées faute d'identifiant individuel. Commençons par ce dernier cas. De nombreuses données démographiques individuelles se trouvent dans des sources diverses : recensements, registre national, registres spécifiques (tel un registre du cancer), bases de données médicales (des médecins, des hôpitaux, de la consommation de médicaments, etc.). Même sans appariement, ces données peuvent être mises en parallèle et leurs évolutions, temporelles ou spatiales, peuvent être comparées. Ces dernières permettent, notamment, d'étudier l'évolution de la consommation d'un médicament en relation avec celle des indicateurs de santé, ou de faire apparaître un nouveau problème de santé dans tel arrondissement. Même si les variables peuvent être reliées aux individus (rendus anonymes) au sein d'une même source (un recensement), les données de sources multiples ne peuvent être rapportées aux mêmes individus en absence d'un numéro commun d'identification personnel.

Lorsque les individus disposent d'un même identifiant dans les diverses sources, un appariement des sources peut désormais être effectué. On peut suivre le même individu d'une source à l'autre et disposer ainsi d'une vue longitudinale de son état de santé, les données provenant de l'état civil,

de multiples registres nationaux ou régionaux, de données médicales diverses, etc. De nouvelles technologies, telle la montre du fabricant ‘à la pomme’, fournissent aussi des données médicales individuelles¹². On dispose ainsi de l’ordre temporel des événements, cet ordre étant une condition nécessaire (mais non suffisante) pour déterminer les relations causes-effets (les premières précédant les seconds) et également pour ordonner la décomposition en sous-mécanismes du modèle structurel causal examiné brièvement à la section 5.

Le recours aux données massives, analysées au moyen de méthodes d’exploration des données, permet de fournir un modèle descriptif des données, basé sur les associations entre variables mais, utilisé tel quel, ce modèle est non-structurel au sens de la Section 5. Le mécanisme sous-jacent demeure une ‘boîte noire’. En effet, les ‘Big Data’ sont généralement caractérisées par des séries de grande longueur *et* par un nombre important de variables conjointes. Ceci met donc en jeu de très nombreux mécanismes et sous-mécanismes : une situation trop compliquée pour permettre une modélisation structurelle adéquate.

Ces données conservent donc leur intérêt pour explorer, par voie inductive¹³, des mécanismes complexes, dont la complexité rend une décomposition récursive (quasi) impossible. Elles peuvent conduire aussi à améliorer les modèles prédictifs. Lorsque les fichiers peuvent être appariés entre eux, au moyen d’un NIP par exemple, l’utilisation de sources multiples reliées entre elles permet, de plus, d’enrichir considérablement le cadre opérationnel du modèle structurel causal. En contrepartie, les *Big Data* posent le problème de la confidentialité des données personnelles¹⁴, de la masse énorme (et croissante) de données à stocker, traiter et analyser, de la rapidité d’obtention des résultats et du coût.

8 Modèle, explication, intervention politique

Nous avons vu qu’il existe, en sciences sociales, une classe de modèles quantitatifs qui nous permettent de décerner les causes et les effets des phénomènes qui nous intéressent. En particulier, les modèles structurels modélisent les mécanismes sous-jacents aux phénomènes. Ces modèles expliquent donc les phénomènes sociaux dans une optique causale.

Relier la modélisation à l’explication nous paraît important car, à notre sens, le scientifique ne doit pas se contenter de trouver des corrélations entre variables mais doit essayer, autant que possible, d’identifier le « pourquoi » des choses, c’est-à-dire leurs « parce que ». Cela n’implique pas qu’une explication est nécessairement causale (voir notamment A. Bokulich, 2017). A titre d’exemple, une

¹² L’*Apple Watch Series 4* relève la fréquence cardiaque du porteur et peut procéder à un ECG simplifié. Les données sont sauvegardées et peuvent être envoyées à un médecin.

¹³ On parle souvent ici d’une « data-driven approach », c. à d. d’une approche guidée par les données.

¹⁴ On se souvient du scandale *Facebook – Cambridge Analytica* (CA), où les données de Facebook ont été fournies à CA et utilisées à des fins électorales.

étude de la transition démographique peut avoir recours à une *modélisation systémique* examinant les interrelations entre variables démographiques et sociétales (Loriaux, 1994). Pour donner un autre exemple, Billari *et al.* (2007) ont recouru à une *modélisation multi-agents*¹⁵ pour simuler les trajectoires individuelles en matière de choix du conjoint, tenant compte des interactions sociales. Il existe dès lors différentes sortes d'explications utilisées en sciences, mais ici c'est bien l'explication causale qui est en jeu.

L'explication est un thème classique en philosophie des sciences. A partir de la philosophie néopositiviste, l'explication a joué un rôle central dans le débat en philosophie des sciences. La théorie la plus influente, au cours des années 1960-70, a été le « modèle des lois de couverture » (*covering-law model*). Expliquer revient ici à donner un bon argument (déductif ou inductif), soutenu par un recours à des lois générales (Hempel, 1965 ; Hempel et Oppenheim, 1948). Ensuite, « expliquer » a été examiné par rapport à la possibilité d'identifier les processus causaux, en particulier les causes et les mécanismes. Ce tournant dans le débat est dû à Wesley Salmon pour qui « we have to put the cause into because » (Salmon, 1984, p. 96). Depuis Salmon, l'explication causale s'est engagée dans une voie importante en développant des notions d'explication qui rendent compte de ses spécificités dans différentes disciplines – de la biologie à la sociologie analytique. La « philosophie des mécanismes » a ainsi contribué à étudier comment les différentes disciplines scientifiques expliquent. (Glennan et Illari, 2017).

Ce qui a manqué, toutefois, est l'examen des différentes pratiques explicatives dans le contexte de la production de l'explication. En d'autres termes, expliquer n'est pas seulement justifier les choix de modélisation, élaborer la décomposition récursive et en estimer les paramètres, ou identifier un mécanisme par ses entités et fonctions. Expliquer, c'est aussi présenter les résultats de la recherche, les interpréter et les situer dans leur contexte. Cet aspect devient d'autant plus pertinent lorsqu'on considère une autre question : la connaissance issue de la modélisation doit-elle être traduite en action politique ? La question est importante et la réponse n'est pas manifeste. D'une part, ce n'est pas toujours évident de pouvoir déduire une politique des résultats de la modélisation. D'autre part, considérer que le rôle du scientifique est de donner des recommandations de nature politique est sujet à controverse. Voyons ces deux problèmes plus en détails.

Même si nous avons une bonne compréhension du mécanisme explicatif entre une cause et son effet, il n'est pas garanti que pour modifier l'effet il faille agir sur la cause. En d'autres termes, les mécanismes explicatifs ne sont pas nécessairement les mécanismes les plus appropriés pour une intervention. Ce raccourci dans le raisonnement a bien fonctionné, par exemple, pour de nombreuses maladies infectieuses, mais cette voie n'est pas du tout garantie pour les maladies chroniques causées par des facteurs multiples. Kelly et Russo (2017) montrent, en effet, que l'explication et la conception de la politique suivent des logiques différentes. Ceci est d'autant plus vrai si l'on considère que l'action sur des facteurs sociaux est soumise à une forte variabilité et

¹⁵ Plus connue sous son nom en anglais, i.e. *agent-based modelling*.

instabilité. Les politiques anti-tabac sont intéressantes à ce sujet car la relation entre tabac et maladie est claire au niveau biologique, mais les interventions varient fortement selon les comportements de la population cible. Par exemple, il peut s'agir d'une augmentation des prix pour toute la population, d'une campagne anti-tabac dans les écoles pour les jeunes, ou encore d'une campagne pour les femmes enceintes, etc. En effet, le tabagisme pose des questions à la fois sur la diversité des comportements et sur celle des effets. Ces brefs exemples montrent que l'action politique doit aussi tenir compte des spécificités des sous-populations. En résumé, la connaissance des mécanismes sociaux à divers niveaux est nécessaire pour élaborer des interventions politiques efficaces.

Venons-en maintenant à la deuxième question, sur le rôle du scientifique par rapport à l'action politique. Nous avons examiné pourquoi la traduction de la connaissance en action n'est pas automatique ; cela est d'autant plus vrai si l'on considère le processus décisionnel. Qui s'occupe de l'élaboration des politiques, et comment ? Et quelle est la place du scientifique dans ce processus ? Tout d'abord, le fait de savoir si le but d'une étude empirique est d'aboutir à des recommandations politiques explicites est sujet à controverse. En épidémiologie, par exemple, c'est un débat ouvert sur la nature même de la discipline. Admettons que la recommandation politique fasse partie de l'activité scientifique. Comment faut-il envisager, dans ce cas, les échanges entre scientifiques et décideurs politiques ? Et comment faudrait-il inclure les citoyens dans le débat ? Faut-il que les scientifiques fournissent plusieurs scénarios aux décideurs, au lieu d'une seule recommandation spécifique ? En bref, même s'il semble raisonnable et plausible de relier l'explication à l'action, nombre de questions se posent immédiatement ; ce sont des aspects que les sociologues des sciences ont étudiés depuis longtemps. Toutefois, la question de l'aspect normatif de la science reste un point délicat ; il suffit de penser aux débats sur le changement climatique aux Etats Unis, ou sur les mouvements « no-vax (vaccination) » en Europe.

9 Quelques remarques en guise de conclusion

Comme on l'a noté plusieurs fois au cours de cet article, les phénomènes sociaux sont fréquemment très complexes et variables dans le temps et l'espace. Pour les expliquer, on construit souvent un modèle simplifiant la réalité, c'est-à-dire ne retenant que les traits considérés comme essentiels. Un modèle explicatif ne peut donc être identifié à la réalité ; son but est de fournir les caractéristiques principales du processus générateur de données. Les modèles statistiques ont une composante aléatoire qui représente, précisément, la partie non explicative du modèle. La confrontation du modèle aux observations permettra ensuite de confirmer ou d'infirmer le modèle. Ce processus de vérification est fondamental dans toutes les sciences mais la conclusion est toujours temporaire. En effet, un modèle confirmé peut être infirmé par après, à la suite de l'obtention de nouvelles données ou moyens d'observation, ou de changements structurels dans la société.

Lors de l'établissement des sous-mécanismes d'un modèle causal récursif, il faut être très vigilant concernant des pertes possibles d'exogénéité. La présence de variables de confusion, manifestes

ou latentes, peut induire des conclusions fallacieuses concernant les relations causes-effets. Il faut donc contrôler, dans la mesure du possible, les éventuelles variables de confusion ou, au minimum, être conscient de leurs effets probables et en tenir compte dans l'interprétation des résultats.

Le modèle causal, construit pour une population de référence donnée, doit pouvoir bénéficier d'une certaine invariance au sein de cette population. Un modèle qui serait propre à chaque individu n'aurait aucune portée générale. A ce titre, les sous-mécanismes biologiques sont typiquement plus stables que les sous-mécanismes sociaux, comme l'exemple de la figure 2 l'a montré. Une définition précise de la population de référence et de ses coordonnées spatio-temporelles est essentielle à cette fin.

Une intervention politique sur un mécanisme implique idéalement la connaissance des causes produisant les effets. Deux remarques sont toutefois à retenir ici. D'abord, comme Lucas l'a signalé, une intervention sur un des sous-mécanismes du système peut entraîner des effets sur l'ensemble du mécanisme et modifier d'autres relations entre variables ou la valeur de certains paramètres. Ensuite, il n'est pas sûr que les mécanismes explicatifs les plus pertinents soient nécessairement les voies les plus appropriées pour effectuer une intervention politique.

Enfin, en tant que scientifiques, nous nous interrogeons sur l'utilité sociale de nos recherches, tout en restant fermes sur un principe fondamental : celui de la liberté et de l'autonomie de la science. En philosophie des sciences, par exemple, nombre de philosophes commencent à pratiquer une philosophie des sciences socialement impliquée, c'est-à-dire une philosophie qui, tout en restant fidèle à sa nature conceptuelle et théorique, utilise ses ressources pour aborder des problèmes réels. Sans doute devrions-nous développer une éthique de la recherche qui va, par exemple, au-delà des questions sur la gestion des données ou sur la manipulation des embryons. Les scientifiques ont en effet une responsabilité envers la société et devraient essayer d'avoir pour objectif de laisser le monde dans de meilleures conditions que celles où ils l'ont trouvé. Ceci est vrai pour nous en tant que chercheurs, en tant qu'enseignants, et certainement en tant que citoyens.

Remerciements

Ce texte a été rédigé à l'incitation de Jean-Louis Migeot, membre de l'Académie royale de Belgique, en vue du colloque sur les modèles en sciences qui s'est tenu à l'Académie le 8 mars 2019 à Bruxelles.

Les auteurs remercient particulièrement Catherine Gourbin et André Schreiber pour leurs remarques et commentaires très pertinents, basés sur une lecture attentive d'une version antérieure.

Bibliographie

Armatte M. (2005). La notion de modèle dans les sciences sociales : anciennes et nouvelles significations, *Mathematics and Social Sciences*, 172(4), 91-123.

Billari F.C., Prskawetz A., Aparicio Diaz B., Fent T. (2007). The "Wedding-Ring": An agent-based marriage model based on social interaction, *Demographic Research*, 17/3, 59-82, doi: 10.4054/DemRes.2007.17.3.

Böhler M. (2018). Scientific writing between tabloid storytelling, arcane formulaic hermetism, and narrative knowledge, *Journal for General Philosophy of Science*, 49(4), 551–567.

Bokulich A. (2017). Models and explanation, chap. 4 in L. Magnani et T. Bertolotti (Eds), *Springer Handbook on Model-based Science*, Springer, 103-118.

Caselli G. (2006). Mortality forecasts. Hypotheses and methods, in G. Caselli, J. Vallin, et G. Wunsch (Eds.) : *Demography - Analysis and Synthesis*, Volume III, Chapter 75, Academic Press, 177-188.

Flores R.D., Schachter A. (2018). Who are the "Illegals"? The social construction of illegality in the United States, *American Sociological Review*, 83(5), 839-868.

Glennan S., Illari P. (Eds.) (2017). *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, Routledge.

Gourbin C., Wunsch G., et al. (2017). Direct and indirect paths leading to contraceptive use in urban Africa, *Revue Quetelet / Quetelet Journal*, 5(1), 33-71.

Grawitz M. (2000). *Méthodes des sciences sociales*, 11^{ème} édition, Dalloz, Paris, 1040 p.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York, The Free Press.

Hempel, C. G., Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2), 135–175. Reprinted in C. G. Hempel, *Aspects of scientific explanation and other essays* (pp. 245–282). New York, The Free Press.

Kelly M., Russo F. (2017). Causal narratives in public health: the difference between mechanisms of aetiology and mechanisms of prevention in non-communicable diseases. *Sociology of Health & Illness*, 40(1), 82-99, doi: 10.1111/1467-9566.12621.

Lazer D., Kennedy R., King G., Vespignani A. (2014). The Parable of Google Flu: Traps in Big Data Analysis, *Science*, **343** (6176), 1203-1205.

Loriaux M. (1994). Des causes aux systèmes : la causalité en question, chap. 2 in R. Franck (Ed.), *Faut-il chercher aux causes une raison ?*, Vrin, Paris, 41-86.

Lucas R. (1976). Econometric policy evaluation: A critique, *Carnegie-Rochester Conference Series on Public Policy*, Volume 1, pp. 19-46.

Mahwin J. (2017). *Les modèles mathématiques sont-ils des modèles à suivre ?*, L'Académie en Poche N° 97, Académie royale de Belgique.

Mouchart M., Wunsch G., Russo F. (2016). Controlling Variables in Social Systems - A Structural Modelling Approach, *Bulletin of Sociological Methodology/ Bulletin de Méthodologie Sociologique*, 132, 5-2

Russo F. (2017). Model-based reasoning in the social sciences, chap. 44 in L. Magnani et T. Bertolotti (Eds.), *Springer Handbook on Model-based Science*, Springer, 953-970.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Simon, H.A., Newell, A. (1956). Models: Their uses and limitations, in L.D. White (Ed.), *The state of the social sciences*, Chicago, University of Chicago Press, 66-83.

Vartiainen E. (2018). The North Karelia Project: Cardiovascular disease prevention in Finland, *Global Cardiology Science & Practice*, 2018(2): 13, doi: 10.21542/gcsp.2018.13.

Wunsch G., Mouchart M., Russo F. (2015). *Les limites de la connaissance en sciences sociales*, L'Académie en Poche N° 62, Académie royale de Belgique.

Wunsch G., Mouchart M., Russo F. (2018), Causal attribution in block-recursive social systems - A structural modeling perspective, *Methodological Innovations*, 11(1), 1-11.