Alain Desrosières

Entre réalisme métrologique et conventions d'équivalence : les ambiguïtés de la sociologie quantitative*

(article publié dans *Genèses*, 43, juin 2001, pp. 112-127)

La " méthodologie statistique " utilisée par les sciences sociales quantitatives est en général présentée en deux parties bien distinctes, rarement reliées entre elles : d'une part, la construction des données, puis, d'autre part, leur traitement et leur interprétation. Souvent, entre ces deux mondes, se dresse la " banque de données ", qui fonctionne comme un sas de passage de l'un à l'autre. Or le monde de la " construction " est lui-même tendu entre deux façons de rendre compte de ses pratiques : la *mesure*, issue du langage des sciences de la nature, le *codage conventionnel*, inspiré, selon les cas, du droit, des sciences politiques, ou des sciences cognitives. Cette ambiguïté, caractéristique des sciences sociales quantitatives, peut être analysée à travers une mise en perspective historique, qui permet aussi de relier les questions soulevées par la construction des données à celles qui résultent de la diversité des modes d'analyse : l'exemple des différences entre " régression logistique " et " analyse des données à la française " sera notamment présenté ici.

L'expression même de " méthodologie statistique " implique une division du travail entre, d'une part, des " experts " de l'outil statistique en tant que tel et, d'autre part, des " usagers " de celui-ci, économistes, sociologues, historiens ou psychologues. Dans cette perspective, les progrès et les innovations de la statistique auraient un impact sur chacune de ces sciences, perçues comme des domaines d'application de formalisations élaborées en dehors d'elles. Il est vrai que l'autonomisation croissante de la recherche et de l'enseignement des statistiques mathématiques tend à conforter cette impression. Pourtant l'histoire et la sociologie de la statistique offrent de nombreux exemples d'effets *de sens inverse*, où des questions posées par des sociologues, des économistes ou des spécialistes d'autres disciplines ont profondément transformé la méthodologie statistique. Un exemple important en est fourni, à la fin du XIX^{e} siècle, par les travaux de Francis Galton et Karl Pearson sur l'hérédité humaine et l'eugénisme, qui ont conduit à formaliser la régression linéaire, la corrélation et le test du chi-deux[i]. À leur suite, les recherches agronomiques de Ronald Fisher conduisent à la statistique inférentielle moderne et à l'analyse de la variance.

Ce sont bien sûr les développements de l'économétrie qui, depuis les années 1930 et 1940, offrent les exemples les plus connus des interactions réciproques entre une discipline scientifique et la théorie statistique. C'est même à l'occasion des controverses de cette période cruciale de la quantification des sciences sociales, qu'ont été distinguées trois façons différentes d'envisager les relations entre " théorie " et " données observées "[ii]. Pour la première, une théorie, de type hypothético-déductif (Léon Walras) peut, au mieux, être illustrée par des données, mais elle est logiquement supérieure à celles-ci. Pour la seconde en revanche, l'induction permet d'inférer, à partir de régularités statistiques observées, des " lois " générales (Adolphe Quételet), dont la nature est complètement différente des " lois " postulées par les théories déductives du premier type. Enfin, dans une troisième perspective, qui n'apparaît que dans les années 1930, des hypothèses théoriques sont *testées*, rejetées ou provisoirement acceptées, au moyen des outils de la statistique inférentielle alors naissante. Ces trois manières d'articuler " théorie " et " données " sont clairement visibles dans l'histoire de la pensée économique, où, jusqu'aux années 1940, se sont affrontés, d'une part, des économistes logiciens et mathématiciens, et, d'autre part, des économistes dits

“ historicistes ”, ou “ institutionnalistes ”, *grands utilisateurs de statistiques descriptives*, mais hostiles aux “ lois ” théoriques générales. L'approche probabiliste de l'économétrie[iii] transforme profondément ce paysage manichéen, en introduisant la notion de “ modèle stochastique ”.

Cette distinction nette entre trois façons de relier la théorie et les données permet de reconstruire l'histoire longue des relations entre économie et statistique : Mary Morgan[iv] ou Michel Armatte[v] offrent de bonnes synthèses de cette histoire. Mais, dans cette histoire, le pôle “ théorie ” est relativement dominant. Depuis Adam Smith, les controverses entre économistes ont plus souvent porté sur le corpus plus ou moins formalisé et cohérent des hypothèses théoriques que sur l'observation et la construction des données, laissées à des spécialistes souvent extérieurs à la discipline : les statisticiens, les comptables nationaux. La dernière période, où des modèles sont proposés et testés au vu de données empiriques n'a que partiellement modifié cette situation, même si, dans le cas français de l'Insee qui rassemble statisticiens et économistes, cette séparation est moins nette que dans les pays anglo-saxons.

Le cas de la sociologie est très différent. Celle-ci n'a jamais disposé d'un corps théorique dominant et formalisé, autour duquel se noueraient les controverses. Par contre, le pôle “ statistique ” y a joué, historiquement, un rôle essentiel. Parmi les sciences humaines, la sociologie se distingue par son hypothèse centrale : une société doit être étudiée en soi, et non comme la simple juxtaposition des individus qui la composent. Dans ses versions dites “ holistes ”, ou “ réalistes ”, cette hypothèse est forte : les propriétés du groupe sont radicalement distinctes de celles des individus. Le groupe a une réalité par lui-même. En revanche, dans des versions plus “ individualistes ”, ou “ nominalistes ”, l'individu reste premier, mais des effets de composition, d'interaction (tels que les “ effets pervers ”), d'imitation ou de panique, suffisent à donner à la sociologie un statut distinct et spécifique. Mais, qu'ils tiennent pour l'une ou l'autre de ces deux versions, les sociologues ont traditionnellement cherché à explorer cette hypothèse en suivant deux voies (non indépendantes) : la statistique et le droit. La statistique est du social observé in vivo, le droit est du social cristallisé, durci, déposé dans des institutions[vi]. Mais, comme les statisticiens et les sociologues quantitativistes ne l'observent pas toujours, la statistique est souvent tributaire du droit et des pratiques administratives, comme dans le cas des statistiques de la délinquance et du suicide, chères à A. Quételet et à Émile Durkheim. Le durcissement institutionnel est même une condition sine qua non de la “ fiabilité ” des statistiques : les polémiques récurrentes sur la mesure du chômage le démontrent tous les jours. Le développement du réseau statistique est lié à celui d'un système d'institutions. Cet investissement, analogue à celui d'un réseau routier ou ferroviaire, crée des catégories qui deviennent ensuite incontournables, le plus souvent différentes d'un pays à l'autre. De tout cela, le sociologue moderne qui recourt à d'inépuisables “ banques de données ” est quelquefois peu conscient.

Il est vrai que, dès lors qu'elle est coupée de ses conditions sociales et administratives d'enregistrement, la statistique offre, par ses régularités de mieux en mieux explorées par les développements méthodologiques, des arguments “ clé en main ” pour justifier l'autonomie et la prééminence d'une science nettement distincte de la philosophie, du droit, de la psychologie, ou même de l'économie. Le groupe social a des structures spécifiques et des propriétés de régularité et de prévisibilité, dont sont dénués les individus, volatiles et imprévisibles. Cette version classique de la sociologie quantitative a été clairement formulée par A. Quételet, dans les années 1830, puis approfondie par É. Durkheim, Paul Félix Lazarsfeld, Pierre Bourdieu, l'Insee, l'Ined et beaucoup d'autres. Elle a été critiquée de plusieurs façons. Des sociologues inspirés par la théorie économique ont cherché à réintégrer

des hypothèses de rationalité des agents individuels. Identifiant la sociologie à l'analyse des effets de composition ou d'agrégation de ces comportements rationnels, ils la rapprochent ainsi de la démarche classique des économistes qui partent de la théorie, notamment celle de la maximisation de l'utilité individuelle.

Les apports de la critique ethno-méthodologique

D'autres sociologues, d'inspiration toute différente de celles des deux catégories précédentes, ont procédé, à partir des années 1960, à ce qu'ils percevaient alors comme une critique radicale de la sociologie quantitative, et particulièrement de la sociologie d'enquête par questionnaire, qui s'était répandue rapidement aux États-Unis, notamment sous l'influence de P. F. Lazarsfeld. Ils attaquaient la "méthodologie", selon eux réductrice et plaquée de l'extérieur, des sociologues statisticiens. Ils lui opposaient une " ethno-méthodologie "[vii], c'est-à-dire une analyse des façons dont les personnes enquêtées elles-mêmes, comprennent, décrivent et catégorisent leurs propres activités[viii]. La comparaison entre les schèmes de perception des acteurs et ceux des sociologues tournait souvent au désastre pour ces derniers, et ceci d'autant plus que l'écart social et culturel entre les premiers et les seconds est plus grand. Cette critique, qui se voulait ravageuse de la sociologie quantitative, a été ensuite intégrée en partie dans la méthodologie de fabrication et d'interprétation des questionnaires, notamment en matière d'opinion. Une attention plus grande a été portée aux questions de vocabulaire, et, plus profondément, de signification des questions, différente selon les milieux sociaux. Ainsi, par exemple, la question, posée à des femmes : " Êtes-vous favorable au travail des femmes ? " était encore, dans les années 1960 et 1970, entendue très différemment, par les femmes de milieux ouvrier et cadre. Les premières percevaient ce travail comme un pis-aller, inévitable si le salaire du mari était insuffisant. Les secondes, en revanche, y voyaient le signe de l'émancipation de la femme, et y étaient favorables.

Dans le climat américain des années 1960 et 1970, les critiques des ethno-méthodologues contre les sociologues statisticiens reflétaient un clivage profond dans le groupe des sociologues, perceptible même géographiquement. Les premiers étaient plutôt sur la côte Ouest, et les seconds sur la côte Est. Il était donc peu probable que de réels échanges aient lieu entre eux. Ainsi le riche article de Clifford Clogg[ix], publié dans *Statistical Science*, sur l'" impact de la méthodologie sociologique sur la méthodologie statistique ", ne mentionne même pas l'existence du travail d'Aaron Cicourel[x]. Cet article de C. Clogg explique en détail comment des innovations concernant les méthodes de *traitement* et *d'analyse* statistique des données ont été induites par des questions sociologiques, mais il n'insiste pas du tout sur la *construction* de ces " données ". Le partage entre les deux " méthodologies " semble complet.

En France, du fait de l'existence d'institutions comme l'Insee et l'Ined, qui concentrent la conception, la réalisation, l'analyse et l'interprétation des enquêtes, ce clivage est moins marqué qu'en Amérique. Dès les années 1950, les enseignements, à l'école de l'Insee, de Gabriel Chevry, Marcel Croze et Jacques Desabie contiennent de nombreuses indications pratiques manifestant une grande sensibilité aux dimensions sociologiques des formulations des questions, de l'interaction entre enquêteurs et enquêtés, et des problèmes de codage des cas ambigus, trois domaines de prédilection des critiques des ethno-méthodologues américains. À partir des années 1960, des sociologues universitaires, P. Bourdieu, C. Baudelot, François Héran et d'autres, enseignent à l'École nationale de la statistique et de l'administration économique (Ensae), et contribuent à construire théoriquement les conseils pratiques fournis dans les manuels écrits par la première génération des statisticiens-enquêteurs de l'Insee.

Ainsi, les définitions des variables et des catégories, auparavant perçues souvent comme des questions de logique formelle et de bonne organisation des instructions adressées aux ateliers de chiffrement, sont réinterrogées dans une perspective sociologique et historique[xi]. Ces définitions sont le plus souvent des *conventions d'équivalence* entre des cas hétérogènes. L'origine de ces conventions peut être externe et antérieure au travail du statisticien, par exemple inscrite dans du droit ou des coutumes. Elle peut aussi être produite par le statisticien lui-même au moment de la conception ou même de l'exploitation de l'enquête. La distinction de ces deux cas est essentielle pour la phase ultérieure *d'interprétation* des résultats obtenus. Ainsi, l'étude des nomenclatures, auparavant perçue comme austère et ingrate, devient, dans le contexte particulier de la France des années 1970 et 1980, un fécond thème de recherche sociologique et historique : sur les postes de dépense des enquêtes budgets de famille[xii], sur les branches industrielles[xiii], sur les catégories socioprofessionnelles[xiv], sur le chômage[xv], sur les causes de décès[xvi], sur la criminalité[xvii]. Dans toutes ces recherches, deux questions distinctes bien que très liées sont soulevées : celle de la définition théorique et pratique des classes, et celle du codage, c'est-à-dire du travail concret d'affectation d'un cas à une classe. Un des apports de ces recherches est précisément de montrer que la première phase (définition des classes) ne peut jamais être pensée indépendamment de la seconde. Celle-ci est, en définitive, une des plus suggestives du travail du sociologue statisticien, ce qu'ignore en général l'épistémologue théoricien et logicien : les " critères " les plus rigoureux sont souvent balayés par l'exploration d'une pile de questionnaires. Les critiques des ethno-méthodologues sont dès lors réintégrées dans la pratique du statisticien.

Soit, par exemple, la question extrêmement sensible, socialement et politiquement, du partage de la population totale en trois catégories : population active occupée, population active sans emploi (ou chômeurs), population inactive. Les frontières entre ces trois groupes peuvent être définies par des critères généraux (variables d'un pays à l'autre), ou avec le Bureau international du travail (BIT), qui est supposé fixer une norme internationale en la matière. Mais l'application concrète de ces règles générales multiplie les difficultés et les hésitations. Ainsi l'état de chômage est défini, selon le BIT, par trois critères : être sans emploi, faire des démarches pour en trouver un, être immédiatement disponible. Or la mise en œuvre concrète de chacun de ces trois critères se révèle problématique. Cette situation s'aggrave même en période de crise de l'emploi, où les critères juridiques sont de plus en plus tournés par des situations dites " informelles ", qui contraignent le statisticien à hésiter entre " le droit " et " le fait ". Ces situations et ces ambiguïtés mettent mal à l'aise le statisticien, sommé par la presse et les usagers de donner " le bon chiffre ", approchant au mieux " la réalité ". Il s'en tire souvent par des expressions comme " halo " ou " flou ", comparable à la situation de l'astronome visant une étoile avec un instrument mal règlé, ou à celle du myope qui aurait perdu ses lunettes. Dans la période plus récente (2000-2001),ces définitions se trouvent même menacées par de nouvelles problématiques du chômage, issues de la théorie microéconomique, qui, en noyant celui-ci dans la catégorie plus vaste de " non-emploi ", suggèrent qu'une partie des chômeurs est constituée de " volontaires ", dans la mesure où la différence entre les aides sociales et le Smic serait jugée par eux insuffisante pour les inciter à reprendre un travail : une telle formulation remet radicalement en cause la notion de " chômage " encore communément répandue, et définie par les critères du BIT.

La sociologie quantitative et ses trois modèles de réalité

Cette pression à fournir le " bon chiffre ", comme les métaphores sur le flou, résultent de ce que la statistique sociale a été construite, légitimée et diffusée à partir du modèle

métrologique réaliste des sciences de la nature. La réalité existe antérieurement à son observation, comme l'étoile polaire a existé bien avant tous les astronomes. Mais précisément la définition et la mesure de la population active et du chômage relèvent d'une autre épistémologie que celle de l'étoile polaire. Elles impliquent des *conventions* (analogues aux principes généraux des lois et des codes votés par les Parlements) et des *décisions* (analogues à celles d'un juge) d'affecter tel cas à telle classe. Pour certains domaines, comme la statistique criminelle, cela semble presque évident, bien que, même dans ce cas, la demande "réaliste" surgisse toujours.

Les questions soulevées ici relèvent de la sociologie, mais non pas au sens de la sociologie quantitative usuelle, qui utilise et interprète sociologiquement des données *auparavant* construites, mais au sens d'une sociologie réflexive, qui étudie les usages sociaux et les rhétoriques d'interprétation de ces données statistiques. Une des raisons pour lesquelles la sociologie quantitative a toujours eu des difficultés à trouver sa place dans une discipline sociologique (de toute façon très émiettée en paradigmes différents), tient à son hésitation sur le statut de réalité des objets "mesurés". On peut, schématiquement, distinguer trois façons distinctes d'interpréter ces données. Elles sont inspirées par trois grandes familles de disciplines scientifiques : les sciences de la nature, les sciences de la vie, les sciences juridiques et politiques. Chacune implique une conception différente de " la réalité ", entre lesquelles la sociologie oscille, non pas pour des raisons épistémologiques, mais en fonction des *contraintes spécifiques à ses divers usages.* Une étude sociologique des usages de la statistique, notamment en sociologie, implique une explicitation approfondie de ces contraintes de situation : qui va lire ? avec quelles notions a priori des objets manipulés et du réalisme de leur mesure ? pour faire quoi (argumenter, contester, décider, etc...) ?

Les sciences de la nature (astronomie, physique), ont imposé, dès le XVIII^{e} siècle une épistémologie de la mesure, enserrée par des schèmes probabilistes. La " loi des erreurs " a inspiré A. Quételet pour construire son " homme moyen ". Cette métrologie réaliste s'est imposée comme un modèle premier, dont les statisticiens ne peuvent jamais se défaire complètement, ne serait-ce que parce qu'elle leur est inlassablement demandée : " quel est le *vrai* chiffre du chômage ? de la hausse des prix ? ". Les commentaires d'accompagnement sur le caractère *conventionnel* de ces mesures ne peuvent rien contre une demande sociale, entretenue aussi en partie par les statisticiens eux-mêmes, soucieux de s'inspirer du modèle le plus achevé, celui des sciences de la nature. Le succès du mot " mesure ", attesté par une exposition récente et par des ouvrages nombreux, est un signe de la prédominance implicite de ce modèle, là ou d'autres mots, comme, d'une part, " indice " ou " symptôme ", ou, d'autre part, " convention d'équivalence " ou " domaine d'action ", typiques des deux autres modèles, pourraient être utilisés.

Le langage des *sciences de la vie,* fort différent du précédent a été utilisé très tôt par A. Quételet et par les hygiénistes, adeptes de la " statistique morale ". Les " moyennes subjectives ", les " propensions " au mariage, au crime ou au suicide, calculées à partir de statistiques administratives globales, étaient des indicateurs macrosociaux, révélés par les régularités statistiques et supposés consistants, *reflétant* des attributs de la société impossibles à atteindre directement. Plus tard, les " variables latentes " de la sociologie anglo-saxonne auront les mêmes propriétés que les " propensions " d'A. Quetelet. Impossibles à mesurer directement, elles apparaissent comme des résultats plus ou moins robustes de l'analyse statistique des variables " patentes ", et sont supposées refléter un contenu sociologique plus profond et plus généralisable que ces dernières. Les axes factoriels de l'analyse des données à la française ont des propriétés comparables. L'" intelligence générale " ou le " quotient

intellectuel " d'Alfred Binet et Théodore Simon en étaient des ancêtres. De même, un indice de prix, un indice de production industrielle ou un indice boursier (Dow Jones, CAC 40 ou Nikkei) sont à la fois des moyennes pondérées et des " variables latentes ", plus générales et explicatives que chacune de leurs composantes.

Ce langage de la variable latente est très différent de celui de la métrologie des sciences de la nature – même si les progrès de la métrologie moderne font de plus en plus apparaître les définitions des unités de longueur, de poids et de temps, comme elles aussi des *conventions,* historiquement variables. La rhétorique réaliste *directe* reste de règle dans la plus grande partie des sciences de la nature (à l'exception peut-être de la physique relativiste et de la mécanique quantique) et sert de modèle à une statistique sociale qui, par ailleurs, recourt *aussi* au langage des sciences de la vie, avec le réalisme *indirect* de ses indicateurs et de ses symptômes. Une partie des débats politiques et scientifiques autour de la statistique sociale résulte de ce flottement entre deux formes de réalisme, direct ou indirect, centrées sur l'idée de *mesure.* Le référent de celle-ci est, selon les cas, visible, caché, ou postulé par l'opération même de construction de la variable latente.

Mais ces deux formes de réalisme se distinguent d'une troisième rhétorique, conventionnaliste, où la trace de l'acte initial de codage reste visible et importante, soit dans une perspective de dénonciation, soit parce que l'agrégat est directement articulé sur une forme d'action collective. Dans ces cas, l'*intention* de l'agrégation et de l'addition reste présente dans l'usage de la statistique présentée. Soit, par exemple, une statistique récente et largement commentée, celle de la maltraitance à enfant, problème social et politique aujourd'hui jugé essentiel. Le fait qu'une question devienne " socialement jugée sociale ", c'est-à-dire relevant d'une action publique, transforme son statut statistique. Des procédures de repérage (numéros verts), d'enregistrement et de comptage sont mises en place. Des définitions et des critères sont formulés. Quand cette opération est encore récente, les interprètes hésitent entre deux lectures : " le nombre des enfants maltraités a augmenté ", ou " les procédures d'observation se sont améliorées ". Ce flottement avait déjà été observé, à propos du chômage, dans les années 1960, avec la progressive mise en place de l'ANPE. Il met mal à l'aise les commentateurs, qui ne peuvent se résoudre à renoncer à une rhétorique réaliste, et critiquent les incertitudes du système d'observation qui ne peut leur fournir des " chiffres fiables "[xviii]. Mais c'est précisément parce que la question est devenue telle qu'on en parle dans les journaux, que ce système d'évaluation évolue vite, et est jugé " peu fiable ". La " fiabilité " est étroitement associée à la stabilité et à la routinisation de la chaîne d'enregistrement et de comptage, qui impliquent que le sujet est devenu moins brûlant.

Mesures, indices ou classes d'équivalence

La sociologie quantitative est hantée, depuis A. Quételet, par trois modèles, issus de trois types de sciences, mais elle l'ignore en général. Ces modèles impliquent des outils et des usages rhétoriques différents. Les *mesures* des sciences de la nature, comme les *indices* des sciences de la vie sont exprimés par des *variables continues* directement observables dans le premier cas, et " latentes " dans le second cas. Mais ces variables caractérisent tout l'univers étudié de façon uniforme. Elles peuvent être modélisées, ajustées à des lois de probabilité, comparées, corrélées, regressées, testées selon les méthodes de la statistique inférentielle. Le modèle initial de ces méthodes a été fourni par les recherches de R. Fisher, au laboratoire d'agronomie expérimentale de Rothamstead, dans les années 1920 et 1930. Les " variables "

sont des caractérisations homogènes interchangeables : la nature suit des lois générales et transposables d'une expérience à l'autre. Cette logique de la mesure a souvent été transférée telle quelle aux sociétés humaines par les premiers pionniers de la sociologie quantitative, notamment anglo-saxons. Par exemple, leur échelle sociale, continue et unidimensionnelle (utilisée pour étudier la mobilité sociale) est issue des travaux de F. Galton sur l'échelle des aptitudes (*abilities)*, reprise aussi par le psychologue[xix] pour étalonner une échelle d'intelligence générale[xx]. La recherche des " variables latentes ", statistiquement plus efficaces et interprétables de façon synthétique, à la façon d'une moyenne, est au cœur de cette démarche, où les sciences de la vie cherchent à coller au plus près des sciences de la nature. Dans cette perspective, où les variables interagissent de façon uniforme, on peut étudier l'" effet d'une variable ", comme l'expérimentateur agricole ajoute ou retire une quantité d'engrais.

Mais, dans le cas des sciences sociales, une troisième forme d'identification des objets de la statistique intervient. La *classe d'équivalence* est une *convention,* issue des sciences juridiques et politiques. C'est une construction humaine, affectant des droits et des devoirs communs à une classe d'hommes définis par des lois, des règlements, des conventions ou de simples usages : " les hommes naissent et demeurent libres et égaux en droits ". La société elle-même est une monumentale entreprise de taxinomie, dont la statistique enregistre les effets : le sexe, l'âge, le lieu de naissance, le diplôme, la catégorie socioprofessionnelle, le lieu de résidence, le statut matrimonial et familial, etc. F. Galton avait tenté de réduire le statut social à une échelle naturelle d'aptitudes innées, mais cet essai de naturalisation a fait long feu. La notion de " classe d'équivalence " est moins familière au statisticien, spontanément influencé par le modèle réaliste des sciences de la nature, où l'on procède à des *mesures*, et non à des opérations de *jugement*, visant à coder, c'est-à-dire à affecter, selon des conventions générales fixées a priori, des cas singuliers à des classes. Bien que les contraintes cognitives de ces opérations de jugement ne fassent pas partie de la culture du statisticien, celui-ci procède intensivement à ces opérations. Les discussions sur le " flou " ou le " halo " de tel ou tel agrégat statistique reflètent la complexité de ces contraintes, et surtout la tentation de les rabattre sur le modèle réaliste des sciences naturelles.

La sociologie quantitative est donc vouée à manipuler en même temps plusieurs outillages, de statuts sociaux et techniques très différents, la *mesure* (sous sa forme directe ou "latente"), et la *classe d'équivalence*. Dans leur majorité, les sociologues anglo-saxons, proches historiquement des sciences expérimentales et de leurs outils, cherchent spontanément à construire des variables continues et mesurables. Celles-ci sont en petit nombre afin de pouvoir être confrontées, corrélées et testées dans des modèles probabilistes, issus précisément des sciences expérimentales. Beaucoup de sociologues français, en revanche, peut-être de culture plus historique et philosophique, ont résisté à cette idée de mesure continue. Ainsi, par exemple, les " échelles de prestige social ", construites à partir d'enquêtes d'opinion sur les métiers, ont eu grand succès en Amérique et en Grande-Bretagne mais ont longtemps été inconnues en France. En revanche, la nomenclature des catégories socioprofessionnelles a été particulièrement travaillée et utilisée en France, depuis Jean Porte dans les années 1950[xxi].

Presque dès son origine, la méthodologie statistique a été marquée par une controverse, entre Karl Pearson et son élève Udny Yule, entre 1900 et 1914, au point que leurs relations en furent affectées[xxii]. Le débat portait précisément sur la tension entre les notions de mesure et de classe d'équivalence : comment mesurer la forme de l'association, ou " corrélation " entre deux variables, quand celles-ci ne sont pas des mesures sur une échelle continue, mais des

classements, ou “ variables discrètes ” ? Le cas le plus simple est un tableau de *2x2=4 cases*, croisant deux tris dichotomiques. De tels cas sont usuels dans les sciences sociales, domaine où U. Yule tente de transférer les méthodes imaginées par K. Pearson pour les sciences biologiques. U. Yule cherche notamment à étudier les effets sur la pauvreté de deux modes d'assistance (à domicile ou en asile). Il propose un indicateur de “ corrélation ” facile à calculer. Si *a, b, c* et *d* dont les quatre cases du tableau, la force de l'association est mesurée par *Q=(ad-bc)/(ad+bc)*. Aucune hypothèse n'est nécessaire sur la distribution des 4 variables.

K. Pearson, en revanche travaille, dans le cadre de la biométrie, sur des mesures continues, physiques ou non, dont la distribution typique suit une loi normale. Si les variables ne sont pas continues mais discrètes, son réflexe est de les rendre continues, en utilisant les fréquences observées des catégories pour étalonner celles-ci sur une échelle continue, *en supposant la distribution normale*. Ce tour de passe-passe lui permet de revenir à la situation continue, en bâtissant une loi normale à deux dimensions dont les distributions marginales s'ajustent sur les deux distributions marginales observées dans le tableau initial. Il démontre qu'il y en a une et une seule, et l'un de ses paramètres fournit la corrélation souhaitée, baptisée “ coefficient de corrélation tétrachorique ”. K. Pearson n'a en revanche que dédain et sarcasmes pour la formule simple de Yule, *Q=(ad-bc)/(ad+bc)*. Elle est arbitraire et ne repose sur rien. Elle pourrait aussi bien être remplacée par ☒ ou ☒. La polémique, qui dure de longues années, porte sur les hypothèses de continuité, naturelles pour le biométricien, mais peu évidentes en sciences humaines. À un moment, U. Yule évoque une situation où des individus sont vivants ou morts. Que signifie donc, dans ce cas, une hypothèse de continuité ?

L'hybridation des outils et l'oubli de leurs origines

La recherche des généalogies et des implications sémantiques des deux notions de *mesure* et de *classe* ne doit pas laisser croire que, ensuite, elles ont mené des vies séparées. Toute la dynamique de la méthodologie statistique a poussé à les associer, les combiner, et à façonner des opérateurs d'échange entre elles. On en citera ici deux exemples, portant sur deux méthodes statistiques, souvent présentées comme concurrentes, et très utilisées aujourd'hui par les sociologues : la régression logistique et l'analyse des données. Les controverses récurrentes qui les opposent ne sont pas sans rapport avec la distinction suggérée ci-dessus entre, d'une part, les sciences de la nature et de la vie, et, d'autre part, celles de la société et du droit. Pourtant, elles se sont en partie échangé les idées de variables discrètes et continues.

La régression logistique est une extension et une systématisation de l'ancienne idée d'“ élimination des effets de structure ” ou “ une variable peut en cacher une autre ”. Cette question a été traitée par la régression multiple et les calculs de corrélation partielle, par U. Yule, depuis le début du siècle. Mais le problème vient de ce que, dans le cas de la sociologie, les variables dont on souhaite analyser les effets sont souvent “ discrètes ”, c'est-à-dire constituées de “ classes d'équivalence ”, au sens défini ci-dessus. Les modèles de régression logistique (du type logit) permettent d'utiliser des formules de régression linéaire classique par des transformations logarithmiques ad hoc. Mais, ce faisant, on revient dans la situation des sciences de la nature où, comme dans les expériences agronomiques de R. Fisher, on distingue des “ effets purs ” de variables agissant de façon homogène sur tout l'espace étudié. L'idée que les lois et leurs effets sont transportables et reproductibles, pourvu que soient respectées les conditions *ceteris paribus*, est sous-jacente à cette façon de traiter les variables sociologiques, et elle est issue des sciences de la nature.

Il ne s'agit pas ici de critiquer cet usage, comme cela a déjà été fait maintes fois, depuis les économistes historicistes allemands du XIX^e^ siècle, François Simiand, Maurice Halbwachs et, plus récemment, Jean-Claude Passeron[xxiii], qui revendique, pour la sociologie, la possibilité d'un " espace non-poppérien du raisonnement ", basé sur l'historicité des sociétés humaines. Mais cette nécessaire historicisation n'est pas appliquée par J.-C. Passeron lui-même aux usages des statistiques à la sociologie. " Historiciser " signifierait étudier, dans un contexte historique donné, la cohérence, formelle et sociale, et l'efficacité propre d'un montage de définitions, de tableaux, de graphiques et de calculs. Ces montages ne peuvent être compris que du point de vue de leur insertion dans un réseau plus vaste d'argumentation et d'action et non pas seulement en tant que porteur d'une connaissance supplémentaire, une brique dans l'édifice de la science. Un exemple : l'élimination des effets de structure a été pratiquée et discutée au moins depuis les années 1920. F. Simiand a formulé à leur sujet une critique spectaculaire : " cette méthode conduit à étudier et comparer les comportements d'un renne au Sahara et d'un chameau au Pôle Nord[xxiv] ".

Cette boutade a été souvent reprise, jusqu'à nos jours, par ceux qui souhaitent critiquer la transposition du modèle des sciences de la nature aux sociétés humaines. Or cette élimination des effets de structure a été considérablement approfondie et sophistiquée, depuis 1980, par l'usage des modèles de régression logistique, qui permettent précisément de séparer et de quantifier finement les " effets purs " des diverses variables " explicatives ". La question n'est donc plus de savoir si ceux qui le font ont raison ou tort, mais *pourquoi* ils le font ? Comment la régression logistique est-elle intégrée dans une plus longue chaîne d'arguments, dans laquelle on peut conjecturer que le *jugement* et *l'action* (et non pas la description) occupent une place centrale. Les débats des épistémologues portent sur ce qu'*il faut faire* pour faire de la " vraie science ". Ceux des sociologues des sciences sont différents. Ils portent sur *ce que font* les scientifiques et les objets qu'ils construisent, et pourquoi, sans chercher d'abord à séparer le bon grain de l'ivraie.

Le modèle de la régression logistique est hybride en ce qu'il met en œuvre des *variables* dites " discrètes ", c'est-à-dire découpant exhaustivement l'univers en *classes* disjointes. Les acteurs de son théâtre sont ces variables : ce sont elles qui agissent, ont des effets, purs ou brouillés par ceux de variables concurrentes. Dans les comptes rendus, elles constituent les *sujets des verbes*, et, à ce titre, elles se rattachent au langage des sciences de la nature. Pourtant, au lieu de refléter des mesures, elles rassemblent des classes, constituées sur le modèle des sciences juridiques ou politiques. Mais ces classes ne parlent pas en tant que telles ; elles laissent la parole aux variables : le sexe, l'âge, le diplôme, le revenu, la CSP, la région, la taille de la commune. Ceux qui, à l'image de K. Pearson et de sa biométrie, sont les plus attirés par le modèle des sciences de la nature, sont gênés par ces variables discontinues. L'âge et le revenu pourraient, à la rigueur, être rapatriés dans le camp des " vraies " variables, mais les autres[xxv] sont toujours un peu suspectes d'arbitraire et de " conventionnel " : que se passerait-il si on " changeait de nomenclature " ?

Mais le cœur de ces méthodes reste la question des *effets* de certaines variables sur d'autres. Cette interrogation ne trouve sens que dans une perspective d'*action* et de transformation du monde. Sur quoi faut-il agir pour atteindre tel but ? La variable résume alors un objectif (un indicateur social, un critère de convergence fixé par un traité), ou un moyen d'action *de portée générale*. La variable est faite pour être inscrite sur un cadran du tableau de bord de l'homme d'action. La science sociale est une science expérimentale appliquée. Mais elle doit *composer* avec les classes d'équivalence produites historiquement par les États de droit : catégories administratives, salariales, scolaires, familiales, fiscales (différentes d'un pays à l'autre, pour

le malheur de la construction d'une statistique européenne). C'est pour cette raison que les critiques qui, de F. Simiand à J.-C. Passeron, ont visé ces méthodes, ont en partie manqué leur but, et n'ont eu aucun effet. Elles ne s'en prenaient qu'à leur dimension *cognitive*, au lieu de décrire leurs *usages et leurs effets sociaux*, qui ne sont intelligibles que dans une sociologie beaucoup plus vaste des moyens dont dispose une société pour se représenter et agir sur elle-même.

Analyse des données et Data analysis

L'analyse des données dite " à la française ", c'est-à-dire issue des travaux de Jean-Paul Benzécri et Brigitte Escoffier, combine, elle aussi, des aspects classificatoires et métrologiques. Elle est d'ailleurs dans la suite directe de l'" analyse factorielle " des psychomètres, qui poursuivaient une démarche typique de la métrologie "symptomatique" des sciences de la vie[xxvi]. L'intelligence générale (ou " facteur *g* ") de Charles Spearman[xxvii] était une variable latente, " moyenne " des résultats de *n* épreuves scolaires subies par *p* élèves. Elle était déterminée comme l'axe principal d'inertie du nuage des *p* points représentant les performances des élèves dans l'espace à *n* dimensions des épreuves. L'unidimensionnalité de ce nuage a été ensuite discutée et critiquée par Cyril Burt, puis Louis léon Thurstone, qui cherchent à explorer des axes orthogonaux, décrivant plus fidèlement la complexité de l'espace des " aptitudes ". Sans ordinateurs, les psychomètres acquièrent une grande dextérité pour opérer des " rotations d'axes ", dans des espaces à beaucoup de dimensions. Surtout utilisée par les psychologues, cette technique est peu connue des sociologues, du moins en France, jusqu'aux années 1960.

Une expérience remarquable resta pourtant isolée et sans suites. En 1954, J. Porte, le créateur des CSP, effectue à l'Insee une " enquête par sondage sur l'auditoire radiophonique ", ancêtre de l'audimat. Dans une " analyse factorielle des goûts " préfigurant, vingt-cinq plus tôt, " *La distinction* " de P. Bourdieu, il effectue une analyse factorielle d'un tableau des corrélations entre les préférences pour les divers types d'émissions. Il utilise pour cela la méthode de L. L. Thurstone dite " centroïde " : " Une telle opération ne peut guère être justifiée que par son succès, c'est-à-dire la possibilité d'interpréter les résultats[xxviii]. " Il interprète le premier facteur comme opposant les " émissions de qualité " aux " émissions légères ", puis après une habile rotation d'axes (effectuée graphiquement), un second facteur oppose les " émissions musicales " aux " émissions parlées ". Mais l'analyse porte seulement sur les proximités entre émissions, et non sur leurs préférences par les diverses CSP (analysées par des méthodes plus classiques), et surtout l'analyse factorielle ne conduit pas encore à une cartographie, qui sera le propre de l'analyse des correspondances.

Malgré leur homonymie, les méthodes françaises d'*analyse des données*, et les méthodes anglo-saxonnes dites de *data analysis,* popularisées par John Tukey et Eugène Horber, n'ont pas les mêmes philosophies[xxix]. Les méthodes anglo-saxonnes distinguent nettement l'analyse *exploratoire,* qui, par des méthodes d'examen et de visualisation très simple d'un fichier, permettent de formuler de premières hypothèses ou des esquisses de modèles probabilistes, testées ensuite par l'*analyse confirmatoire* qui retrouve alors les techniques classiques de la statistique mathématique. En revanche, l'analyse des données à la française se présente comme une fin en soi, en poussant très loin le rejet de tout modèle probabiliste. Elle est avant tout une technique descriptive. Elle ne vise pas à confirmer ou infirmer une théorie préalablement formulée. De ce point de vue, elle renoue avec l'ancienne tradition des sociologues et des économistes historicistes du XIX^{e} siècle, qui bâtissaient des lois " générales " à partir des données observées.

Portant sur des " tableaux de contingence " distribuant des individus selon des classifications variées, l'analyse des correspondances est adaptée à la conception " conventionnelle ", issue des sciences politiques et du droit. Elle distribue ces classes selon des systèmes de proximités, possédant des configurations de propriétés voisines. Dans ce cas, les acteurs du théâtre ainsi mis en scène sont des *groupes* (ou même des *individus*), et non plus des *variables*. Les sujets des verbes sont, dans les phrases des interprétations, ces groupes (qui peuvent être définis par le sexe, l'âge, la CSP, etc.). Ceux-ci ont une existence autonome par rapport à la nomenclature exhaustive (à la différence des méthodes de régression logistique). Ces méthodes peuvent servir de façon classificatoire a posteriori, en regroupant (de façon ascendante) des individus, ou en découpant (de façon descendante) l'ensemble initial, après définition d'une " distance ", minimisée à l'intérieur des classes et maximisée entre les classes. Dans ce cas, l'analyse statistique engendre littéralement de nouvelles formules *d'équivalences conventionnelles*, réutilisables pour l'action, et n'ayant d'autre portée que dans l'usage qui en est fait.

Mais, dans sa version " cartographique ", très utilisée, l'analyse des correspondances retrouve la perspective métrologique et les variables latentes. Les " axes d'inertie ", déterminés par diagonalisation des matrices de variance covariance, engendrent un nouvel espace, dans lequel les individus et les groupes ont des " coordonnées ". Il est tentant d'interpréter celles-ci, c'est-à-dire de les traiter comme des mesures continues de " quelque chose " qui, bien que non directement visible, existerait dans la nature. Certaines interprétations de J.-P. Benzécri, associant parfois la structure des axes à un dessein divin, rappellent irrésistiblement celles d'A. Quételet, pour qui l'" homme moyen " ne pouvait être que le produit de la volonté divine. Qu'il s'agisse simplement de la nature, ou de Dieu, une statistique réaliste peut toujours contribuer à engendrer du réel, par la seule efficacité de ses procédures de calcul et d'objectivation.

Ainsi, chacune à leur façon, la régression logistique et l'analyse des correspondances opèrent une hybridation entre les optiques métrologiques et classificatoires. Elles constituent aujourd'hui deux des méthodologies statistiques les plus utilisées par les sociologues. On ne peut cependant les comparer, tant leurs langages et leurs usages sont différents[xxx]. Surtout, elles sont utilisées par des sociologues et dans des contextes institutionnels très distincts, ce qui rend difficile une confrontation *sociologique* de leurs différences d'usages. Les produits des régressions logistiques sont présentés comme des *résultats*, associant des effets à des causes, portant sur des variables décontextualisées et supposées de portée générale, à la façon dont les sciences expérimentales déroulent les étapes de leurs investigations[xxxi]. De ce point de vue, ils semblent au cœur de la démarche scientifique d'une sociologie qui progresse en accumulant de tels résultats.

En revanche, l'analyse des données à la française est rarement présentée (à la différence de la *data analysis* anglaise) comme préalable à une " analyse confirmatoire ", vérifiant des hypothèses théoriques dont elle serait une des sources. Elle est plutôt un élément parmi d'autres d'un ensemble de descriptions historiques de la complexité et des dimensions d'un univers social. Les " variables " ne figurent pas en tant que telles, mais à travers les classes qu'elles distinguent. Ce sont les configurations singulières de ces classes et de leurs propriétés qui font l'objet du commentaire du sociologue. La généralisation éventuelle procède d'une rhétorique différente de celle des sciences de la nature ou de la vie. C'est la juxtaposition de configurations similaires qui fournit un argument. Ainsi, la structure bi-dimensionnelle de l'espace des catégories sociales françaises a été suggérée et confirmée par une succession de travaux analysant les comportements de ces catégories à divers points de vue : structure des

consommations, pratiques culturelles, distribution spatiale dans les quartiers urbains, intermariages, comportements électoraux[xxxii]. Ces configurations sont historiques en ce qu'elles dépendent de taxinomies, plus ou moins durcies et elles-mêmes historiques, et de pratiques dont le sens évolue.

Ces différences d'usage reflètent le relatif émiettement d'une discipline, la sociologie, qui tire sa légitimité (mais peut-être aussi son originalité), d'un patchwork de modèles de scientificité. Elle aurait peut-être à gagner à *expliciter* ce mélange et sa portée sociologique, en termes d'insertion de son discours dans des pratiques sociales différentes, plutôt qu'à chercher à faire triompher l'un ou l'autre de ces modèles. L'histoire montre que ces combats en apparence " épistémologiques ", sont en général sans issue, car chacun de ces modèles a un usage social déterminé. Les remarques qui précèdent ne sont d'ailleurs que des hypothèses, qui demanderaient à être validées par une recherche détaillée et comparative sur les usages sociaux des méthodologies statistiques en sociologie, selon les institutions et selon les pays.

***** Ce texte est une version révisée d'une communication aux Journées de méthodologie statistique organisées à l'Insee en décembre1996.

i. Stephen Stigler,*The History of Statistics. The Measurement of Uncertainty before 1900*, Cambridge (Mass.), Harvard University Press, 1986.

ii. Mary Morgan, *The History of Econometric Ideas,* Cambridge, Cambridge University Press, 1990.

iii. Trygve Haavelmo, " The Probability Approach in Econometrics ", *Econometrica,* n° 12, 1944, pp. 1-118.

iv. M. Morgan, *The History ...,op. cit.*

v. Michel Armatte, " Histoire du modèle linéaire. Formes et usages en économie et économétrie jusqu'en 1945 ", Paris, thèse de doctorat, EHESS, 1995.

vi. François Héran, "L'assise statistique de la sociologie", *Économie et Statistique*, n° 168, 1984, pp. 23-35.

vii. On pourrait dire aussi : "méthodologie indigène".

viii. Aaron Cicourel, *Method and Measurement in Sociology*, New York, The Free Press of Glencoe, 1964.

ix. Clifford Clogg, "The Impact of Sociological Methodology on Statistical Methodology", *Statistical Science*, vol. 7, n° 2, 1992, pp. 183-207.

x. A. Cicourel, *Method and Measurement..., op. cit.*

xi. Alain Desrosières, *La politique des grands nombres. Histoire de la raison statistique*, Paris, La Découverte, 2000.

xii. Luc Boltanski, "Taxinomies populaires, taxinomies savantes : les objets de consommation et leur classement", *Revue française de sociologie*, vol. 11, n° 1, 1970, pp. 34-44.

xiii. Bernard Guibert, Jean Laganier, Michel Volle, "Essai sur les nomenclatures industrielles", *Économie et Statistique,*n° 20, 1971, pp. 23-36.

xiv. A. Desrosières, "Éléments pour l'histoire des nomenclatures socioprofessionnelles", in *Pour une histoire de la statistique, tome 1 : contributions*, Paris, Economica-Insee, 1987 (1re éd. 1976), pp. 155-231; L. Boltanski et Laurent Thévenot, "Finding One's Way in Social Space ; a Study Based on Games", *Social Science Information*, vol. 22, n° 4-5, 1983, pp. 631-679; Dominique Merllie, "Les classements professionnels dans les enquêtes de mobilité", *Annales ESC*, vol. 45, n° 6, 1990, pp. 1317-1333; Francis Kramarz, "Déclarer sa profession", *Revue française de Sociologie,* vol. 32, n° 1, 1991, pp. 3-27.

xv. Robert Salais, Nicolas Baverez, Bénédicte Reynaud EY, *L'invention du chômage*, Paris, Puf, 1986.
xvi. Anne Fagot-Largeault, *Les causes de la mort. Histoire naturelle et facteur de risque*, Paris, Vrin, 1989.
xvii. Philippe Robert *et al., Les comptes du crime. Les délinquances en France et leurs mesures*. Paris, L'Harmattan, 1994.
xviii. Des discussions analogues ont lieu chaque année, lors de la publication des statistiques de la délinquance : celles-ci reflètent-elles l'" insécurité ", comme le suggèrent presque tous les commentateurs, ou l'intensité de l'activité de la police ?
xix. Charles Spearman, "General Intelligence Objectively Determined and Measured", *American Journal of Psychology*, n° 15, 1904, pp. 201-293.
xx. Stephen J. Gould, *La Mal-mesure de l'homme*, Paris, Ramsay, 1983.
xxi. Cette comparaison franco-anglaise sur les usages respectifs des *classes* et des *mesures* dans les deux sociologies a été étudiée en détail, à partir d'enquêtes anglaises et françaises, par Marie-Ange Schiltz, " Influence du choix des traitements statistiques sur les opérations élémentaires dans un dépouillement d'enquête : hypothèses, codage et sélection des variables ", communication au congrès de l'Association internationale de sociologie, juillet 1990, Madrid, étude reprise, sous forme modifiée, *in* Michael Greenacre et Jorg Blasius (éd.), *Correspondance Analysis in the Social Sciences,* New York, London, Academic Press, 1994.
xxii. Donald Mac Kenzie, *Statistics in Britain, 1865-1930. The Social Construction of Scientific Knowledge*, Edinburgh, Edinburgh University Press, 1981.
xxiii. Jean-Claude Passeron, *Le raisonnement sociologique. L'espace non-poppérien du raisonnement naturel*, Paris, Nathan, 1991.
xxiv. Cette critique de l' "élimination des effets de structure", souvent citée, est attribuée à François Simiand par son ami Maurice Halbwachs, dans : "La statistique en sociologie", écrit en 1935 et publié en 1944 *in* Centre internationale de synthèse, *La statistique. Ses applications. Les problèmes qu'elles soulèvent*, Paris, Puf, pp. 113-160. Cette citation de F. Simiand, est, selon Olivier Martin (" Raison statistique et raison sociologique chez Maurice Halbwachs", *Revue d'histoire des sciences humaines*, n° 1, 1999, pp. 69-101), inspirée de : F. Simiand, *Cours d'économie politique*, Paris, Domat-Montchestien, 1930, p. 288.
xxv. À l'exception peut-être du sexe, mais des essais pour rendre continue sa définition ont été faits par la psycho-sociologie américaine.
xxvi. Jean-Paul Benzécri, *Histoire et préhistoire de l'analyse des données*, Paris, Dunod, 1982.
xxvii. Ch. Spearman, " General Intelligence... ", *op. cit*.
xxviii. Jean Porte, "Une enquête par sondage sur l'auditoire radiophonique", *Bulletin mensuel de statistique*, supplément janvier-mars 1954, Insee, 1954, p. 53.
xxix. Jean-Claude Deville et Edmond Malinvaud, "Data Analysis in Official Socio-economic Statistics", *Journal of the Royal statistical Society,* A, vol. 146, Part 4, 1983, pp. 335-361; Eugène Horber, " Analyse exploratoire des données et sciences sociales. Vers une approche méthodologique pragmatique ", thèse de doctorat, Université de Genève, 1990 ; M.-A. Schiltz, " Influence du choix ... ", *op. cit.*
xxx. Une discussion de ces différences est proposée par Félicité des Netumières, "Méthodes de régression et analyse factorielle", *Histoire et Mesure,* CNRS, 1997, vol.XII, n°3/4, pp. 271-298..
xxxi. Christian Licoppe, *La formation de la pratique scientifique. Le discours de l'expérience en France et en Angleterre (1630-1820),* Paris, La Découverte, 1996.
xxxii. A. Desrosières et L. Thévenot, *Les catégories socioprofessionnelles,* Paris, La Découverte, 2000.