

JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

CLAUDE THÉLOT

La statistique, science de la mesure

Journal de la société statistique de Paris, tome 127, n° 2 (1986), p. 67-85

http://www.numdam.org/item?id=JSFS_1986__127_2_67_0

© Société de statistique de Paris, 1986, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

COMMUNICATION

LA STATISTIQUE, SCIENCE DE LA MESURE (*)

Claude THÉLOT

chef de la division « Emploi » de l'INSEE

Toute statistique est une construction. L'article le montre à partir d'exemples tirés du chômage et de la mobilité sociale, et ceci en parcourant les deux phases qui conduisent à une statistique : celle qui produit une observation, puis celle qui analyse cette observation. L'observation est construite à partir des cadres généraux (questionnaire, nomenclatures), du type de recueil de l'information (interaction entre enquêteur et enquêté), enfin du mode concret d'enregistrement de cette information. Quant à l'analyse, elle suppose toujours un outil. Et du plus humble (indice, tableau) au plus sophistiqué (analyse des données, économétrie), tout outil exprime un point de vue sur le phénomène, point de vue d'où dérivent directement les conclusions.

All statistics are built. This article shows it from examples drawn from unemployment and social mobility, by examining the two phases which lead to a statistic : the first one which produces an observation, the second one which analyses this observation. The observation is built from general frameworks (questionnaires, codes), from the way of the information (interaction between the interviewer and the interviewed), and from the precise process of enregistrement of this information. On the other hand, the analysis always implies a means of analysis. The means can be very simple (index, table) or more or less complex (data analysis, econometrics). Whether simple or complex any means carries out a point of view on the phenomenon, from which the conclusions are directly drawn.

Monsieur le Président,
Mesdames, Messieurs,

Je suis très fier et flatté de recevoir le prix du statisticien d'expression française de 1985. Je le suis d'autant plus qu'à sa sortie, mon livre « Tel père, tel fils ? » [1] a principalement été commenté, loué ou critiqué par des sociologues ou des journalistes, dans des quotidiens, des hebdomadaires ou des revues de sociologie, mais guère par des statisticiens. Or c'est bien en statisticien que j'ai écrit cet ouvrage et, à aucun moment, je n'ai eu explicitement l'intention de faire œuvre de sociologue. Je me suis borné à présenter et étudier les éléments dont nous disposons pour éclairer la transmission du statut social d'une génération à la suivante, sans m'interroger, comme on a pu me le reprocher, « sur les ressorts possibles de la mobilité (concurrences, rapports de force, inégalités sociales...), les matériaux qu'elle utilise et transmet (atouts multiformes qui ne se réduisent pas à l'emploi, au revenu et au diplôme, (...)) et enfin les mécanismes qui l'animent (stratégie ou inertie, recherche de la conformité ou de la distinction...) » [2].

N'ayant donc pas vraiment fait œuvre de sociologue et m'adressant à des statisticiens, c'est en statisticien que j'ai choisi de faire cet exposé ce soir et je vous propose d'axer notre réflexion sur la mesure : à partir du moment où l'on cherche à quantifier les phénomènes économiques et sociaux et,

(*) Communication présentée le 15 janvier 1986 devant les Sociétés de statistique de Paris et de France

sur la base des données obtenues, à formuler un jugement ou émettre un diagnostic, la question de la mesure est centrale et la statistique est une discipline-reine.

Notre activité quotidienne peut se scinder, pour la commodité de l'exposé, en deux grandes phases : l'observation et l'analyse. La frontière entre ces deux aspects n'est pas très facile à tracer. On peut dire que l'observation va jusqu'à l'obtention d'un fichier, tandis que l'analyse débute à partir du moment où le fichier est obtenu. Je voudrais étayer, à partir d'exemples pris dans chacune de ces deux grandes phases, l'idée majeure de cet exposé : tant l'observation que l'analyse des phénomènes économiques et sociaux dépendent étroitement de ce qu'on peut appeler, selon le cas, une convention de mesure ou un point de vue. Il n'y a pas d'observation ou d'analyse en soi : elles s'inscrivent toujours dans des cadres plus ou moins généraux et formalisés, elles dérivent toujours de choix préalables plus ou moins conscients.

1. L'OBSERVATION

Décrire un protocole d'observation c'est décrire exhaustivement toute la chaîne qui aboutit à un fichier : principes théoriques et méthodes de mise en œuvre. Dans le cas des enquêtes par sondage, on insiste d'ordinaire sur l'échantillonnage, les estimateurs, l'estimation de leur variance, etc..., bref sur tout ce qui tourne autour de la théorie des sondages, et que vous connaissez bien. Je préfère insister sur d'autres aspects qui me paraissent au moins aussi importants et décisifs. Pour ce faire, et tout en suivant chronologiquement le processus qui va de la prise d'information au fichier, j'évoquerai successivement trois aspects :

- la définition des cadres généraux;
- la prise de l'information à la source;
- une fois les cadres généraux définis et l'information prise, les modalités d'enregistrement de cette information.

Concevoir des nomenclatures générales, on l'a dit souvent, est une tâche majeure du statisticien. La réforme de la nomenclature des catégories socioprofessionnelles a donné lieu à des études et des réflexions approfondies sur les différents aspects que cette conception soulève [3]. Je n'y reviens pas. Je me contenterai de rappeler deux points :

- en permanence, le statisticien est confronté à deux tendances fortes et contradictoires : maintenir des cadres stables pour élaborer des données temporellement comparables d'une part, faire évoluer ces cadres pour épouser au plus près l'évolution de la société d'autre part;
- en second lieu, certes une nomenclature rend visibles certains phénomènes mais, en même temps, elle en masque d'autres.

Pour illustrer ces deux réflexions, on peut retenir deux exemples : la structure sociale et le chômage.

L'évolution de la structure sociale

Au début de mon livre, j'ai présenté une série de long terme décrivant l'évolution de la structure sociale, plus exactement la structure en quelques groupes sociaux des hommes actifs depuis le début du siècle (tableau 1, issu de [1]). Les grands traits en sont connus : baisse du nombre de paysans, augmentation (puis stabilité entre 1975 et 1982 et diminution depuis) des ouvriers, explosion des couches moyennes et supérieures salariées. Une telle série avait pour ambition d'en préciser l'ampleur : en 70 ans, environ 3 600 000 paysans de moins, 2 500 000 ouvriers de plus, 3 000 000 employés, cadres, etc... de plus. Pour constituer une telle série, il faut s'appuyer sur les recensements et, comme depuis 80 ans la nomenclature des positions sociales n'a pas été stable, il faut « traduire » les résultats des

TABLEAU 1. — *Structure sociale des hommes actifs*

	Début du siècle	Trois quarts du siècle (1975)
Paysans	42-44 %	13 %
Ouvriers	30-31 %	45 %
Employés, cadres, etc	12-14 %	32 %
Artisans, commerçants industriels	11-13 %	8 %
Membres d'une profession libérale	(1 %)	(1 %)

Source : [1]

recensements d'avant-guerre dans la nomenclature des catégories sociales d'après-guerre. Une telle traduction peut être critiquée à deux titres :

- elle est imprécise, critique exacte mais qui n'en remet pas en question le principe;
- elle est a — ou même anti-sociologique dans la mesure où ce n'est pas un hasard si avant et après-guerre, les nomenclatures dans lesquelles la position des Français est repérée diffèrent : elles diffèrent parce que la structure de la société française a profondément évolué. Si « les découpages, et en particulier ceux qui sont le support de comptages statistiques, dont se dote une société, reflètent la conscience qu'elle a d'elle-même et les problèmes qu'elle se pose à un moment donné » [4], alors il n'est pas très correct de constituer des séries longues dans une nomenclature donnée. Répartir comme je l'ai fait les « isolés » — catégorie d'actifs qui apparaît dans les recensements d'avant-guerre — dans les positions d'après-guerre, par analogie, serait une sorte de trahison car cela revient à faire disparaître « une composante très originale et spécifique de la force de travail, une sorte d'étape dans l'intégration à la main-d'œuvre capitaliste, transition entre deux modes de production » (idem).

On ne peut pas être insensible à cet argument. Cependant, je ne crois pas qu'il doive conduire à renoncer à établir des séries aussi longues que possible (qui d'ailleurs ne seront pas très longues au regard de l'histoire). Or, à moins de se limiter à de simples comptages de population totale, cet objectif requiert soit de se fonder sur des nomenclatures stables soit de passer d'une nomenclature à une autre. Ces séries longues seront imprécises, fondées sur des regroupements grossiers, elles auront une valeur relative comme ces études sociologiques de genèse des nomenclatures nous le démontrent, mais elles sont les seules à pouvoir donner un ordre de grandeur des évolutions survenues. Et je pense que nous ne consacrons pas assez d'attention à l'INSEE à l'établissement de séries longues dans les domaines sociaux.

Une fois qu'une nomenclature aussi générale que celle des catégories socioprofessionnelles a été établie et est utilisée largement, elle joue à plein un rôle de filtre : elle fait voir certains phénomènes, dont elle permet ainsi l'étude; mais aussi, et c'est important, elle rend invisibles un certain nombre de dimensions.

C'est ainsi, on vient de le voir, que les « isolés » disparaissent dans une série de long terme. Mais, à mes yeux, l'exemple le plus clair de cette fonction d'occultation porte sur un clivage que je crois majeur de la société française : entre le secteur public et le secteur privé. Cette distinction figurait dans la nomenclature à ses débuts et des résultats ou des études ont été alors produits, par exemple sur la fécondité [5]. Puis la distinction a disparu de la nomenclature et pendant vingt ans, pratiquement aucune analyse n'a été menée sur ce sujet. Ceci est très remarquable. Il est toujours loisible de dire ou de croire qu'on pourra croiser avec un autre critère — ici, le statut — et donc faire apparaître la distinction entre le public et le privé. En réalité, cette expérience montre qu'il n'en est rien, et il faut donc pour une nomenclature aussi essentielle que celle des professions et catégories socioprofessionnelles, bien choisir les lignes de clivage retenues. C'est ainsi que, dans le nouveau code établi en 1982, a été rétablie la distinction entre privé et fonction publique et figure d'autre part une distinction nouvelle entre ouvriers de type artisanal et ouvriers de type industriel. On peut alors s'attendre dans les années à venir à l'apparition d'études sur ces thèmes.

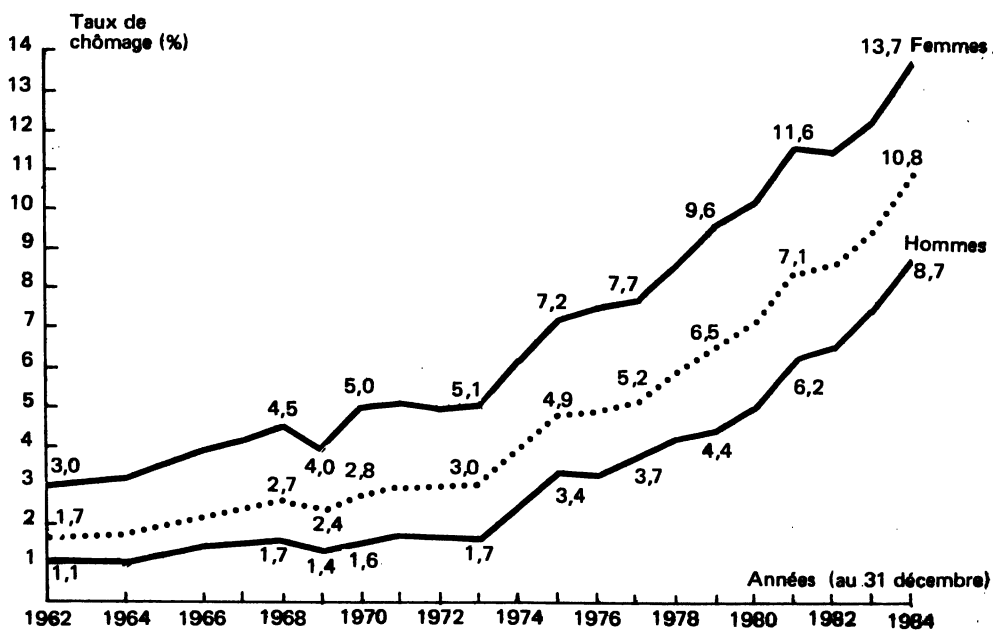
Le chômage

L'exemple du chômage illustre aussi les deux points évoqués à l'instant : difficulté et nécessité d'établir des séries longues; mise en évidence et occultation de certains phénomènes.

Au recensement de 1936, on a dénombré 864 000 chômeurs; à celui de 1982, 2 059 000. Contrairement à l'apparence, il n'est pas du tout certain qu'on puisse, de ces deux estimations, tirer une conclusion solide. Le phénomène du chômage est en effet trop lié à un état de la société, et sa mesure à un ensemble de conventions pour assurer, par exemple, que le chômage est plus important aujourd'hui qu'il y a cinquante ans, ou même pour être sûr qu'une telle comparaison a un sens. Dans un article récent, j'ai considéré qu'il était possible, en revanche, d'établir une série longue sur les vingt dernières années, parce qu'elles présentaient une certaine homogénéité. C'est au début ou au milieu des années soixante en effet qu'apparaissent ou s'amplifient quelques traits majeurs qui modèleront la société française : baisse de la fécondité, développement de l'activité féminine salariée, exode agricole, urbanisation, arrivée des générations d'après-guerre sur le marché du travail, ouverture de l'économie française sur l'extérieur par suite du traité de Rome. Sur un plan plus général, les institutions découlant de la Constitution de 1958, modifiée en 1962, sont stabilisées, et la France se réduit, aux départements et territoires d'outre-mer près, à la métropole : elle a cessé d'être une puissance coloniale, dans les derniers temps déchirée par la guerre.

D'où la constitution, sur cette période, d'une série de taux de chômage homogène, définis à partir des recommandations du Bureau International du Travail : le taux s'établissait à 1,7 % fin 1962 et à 10,8 fin 1984 (graphique I, tiré de [6]). Ce qui est remarquable, c'est l'impact des conventions : dans un autre concept, le taux de chômage que l'INSEE avait publié pour 1962 tournait autour de 1 %, et non 1,7 % comme ici (la différence porte essentiellement sur les femmes). L'écart est considérable.

GRAPHIQUE I

*Taux de chômage depuis vingt ans**

* Il s'agit ici d'une série homogène de taux au sens du BIT, non corrigés des variations saisonnières, au 31 décembre de chaque année.
Source : [6].

En second lieu, partager les personnes en trois catégories actifs occupés/chômeurs/inactifs — ce qui est nécessaire pour établir un taux de chômage — fait par définition voir le chômage — du moins : le chômage selon une convention précise —, mais tend à occulter ou à rendre difficilement perceptibles certains phénomènes qui ont pris de l'ampleur depuis 10 ans sur le marché du travail. Il est d'ailleurs significatif qu'on parle à ce sujet de « situations intermédiaires »; or, seules sont à peu près correctement « vues » celles qui sont institutionnalisées, les TUC par exemple, ou, depuis très longtemps, l'apprentissage. Mais l'espèce de no man's land qui se développe, en particulier lors de l'insertion des jeunes, et qui est un état constitué d'allers et retours fréquents entre un emploi court ou partiel, du chômage et des périodes d'inactivité courtes elles aussi, n'est pas repéré comme tel. Dans le même esprit, considérer que doit être exclue du chômage et incluse dans l'emploi toute personne même si elle n'a travaillé qu'une heure durant la semaine de référence aboutit à ne pas réellement mettre en lumière cette population exerçant des petits travaux, à temps très partiels (souvent multiples, successifs, discontinus) : en mars 1985, par exemple, selon l'enquête sur l'emploi, environ 400 000 personnes avaient travaillé moins de 15 heures dans la semaine précédant l'enquête, soit qu'il s'agisse de leur horaire habituel, soit qu'elles y aient été contraintes.

L'interaction enquêteur-enquêté

Après avoir défini les cadres généraux, nomenclatures et questionnaires, le statisticien prélève l'information et, si l'on se situe dans le cadre d'une enquête, il faut mentionner le rôle important que joue le type d'interaction entre enquêteur et enquêté. Le chômage constitue à nouveau un bon exemple. On connaît la différence entre les résultats des recensements et ceux des enquêtes sur l'emploi : les deux protocoles sont si éloignés — avec l'intervention d'un enquêteur dans le second cas — qu'il n'y a pas lieu de s'étonner ni de s'appesantir. Il est beaucoup plus instructif — et moins connu — de se placer au sein du même protocole général — l'enquête sur l'emploi — et de constater que la mesure diffère selon que c'est la première fois ou non que l'on interroge les ménages. En effet, dans le sous-échantillon entrant, la proportion des chômeurs dans la population totale est, très souvent, supérieure ou égale à la proportion mesurée sur l'ensemble de l'échantillon (tableau 2; voir aussi [7]). Le phénomène paraît assez général : chez les hommes et chez les femmes, dans toutes les tranches d'âge — notamment chez les 25-49 ans. Au total, en moyenne sur chaque série homogène d'enquêtes, le nombre de chômeurs dans le sous-échantillon entrant excède d'environ 20 000 à 50 000 le nombre mesuré sur l'échantillon complet. L'écart est du même ordre que celui qui, dans l'enquête, est attribuable à la différence de concept de chômage.

Ceci n'est pas spécifique à la France : aux États-Unis où l'enquête sur l'emploi, mensuelle, s'effectue elle aussi par rotation de sous-échantillons, la même constatation a été faite [8] et la même conclusion générale tirée : on ne sait pas exactement pourquoi il en est ainsi, mais on s'efforce d'interpréter ce résultat en termes d'interaction enquêteur — enquêté, interaction qui serait différente à la première interview et aux suivantes. Ainsi, par exemple :

— lors de la première enquête, les gens seraient plus mal à l'aise que lors des suivantes, ce qui inciterait plus qu'ensuite un certain nombre d'entre eux à se déclarer chômeur plutôt qu'inactif, parce que la première situation serait jugée socialement plus avouable devant l'enquêteur que la seconde;

— ou encore, une enquête venant de l'administration ferait penser que cela accroît la possibilité d'occuper un emploi, et, en conséquence, on se déclarerait un peu plus volontiers au chômage. A la seconde enquête, cette impression aurait été reconnue pour ce qu'elle est, une illusion.

Quelle qu'en soit l'interprétation — fragile et pas très convaincante en l'état actuel des connaissances (1)—, la constatation, elle, est très instructive parce qu'elle est reliée à une modification très fine du protocole expérimental : la seule chose qui change, c'est que c'est la première fois ou non que l'on enquête le ménage.

1. Il se peut qu'au moins dans le cas de la France, il faille faire intervenir des considérations de redressement des non-réponses : la proportion de non-réponses parmi les résidences principales interrogées pour la seconde fois est un peu plus élevée lorsqu'à la première enquête il y avait au moins un chômeur dans cette résidence que lorsqu'il n'y en avait pas [7].

TABLEAU 2. — *Mesure du chômage selon le sous-échantillon dans les enquêtes sur l'emploi*

	1 ^{er} sous-échantillon			2 ^e sous-échantillon			3 ^e sous-échantillon			TOTAL
	+	=	-	+	=	-	+	=	-	
PDRE (1971 à 1981, 16 points)										
Hommes										
< 25 ans	6	5	5	5	2	9	6	2	8	
25-49 ans	11	3	2	5	4	7	0	3	13	
≥ 50 ans	6	6	4	4	5	7	3	7	6	
Total	9	6	1	3	6	7	2	6	8	
Femmes										
< 25 ans	10	3	3	6	0	10	9	0	7	
25-49 ans	9	6	1	3	6	7	3	4	9	
≥ 50 ans	6	8	2	0	13	3	3	6	7	
Total	10	5	1	2	8	6	2	9	5	
Ensemble										
< 25 ans	8	4	4	5	2	9	9	0	7	
25-49 ans	11	5	0	4	5	7	1	2	13	
≥ 50 ans	5	8	3	1	9	6	3	7	7	
Total	12	4	0	4	4	8	1	9	6	
BIT (1975 à 1985, 16 points)										
Hommes										
< 25 ans	8	4	4	8	1	7	4	0	12	
25-49 ans	11	2	3	3	5	8	3	3	10	
≥ 50 ans	6	7	3	6	3	7	3	5	8	
Total	10	5	1	4	6	6	2	4	10	
Femmes										
< 25 ans	6	3	7	10	1	5	6	2	8	
25-49 ans	12	4	0	3	4	9	4	4	8	
≥ 50 ans	7	4	5	3	5	8	4	7	5	
Total	8	6	2	4	4	8	3	6	7	
Ensemble										
< 25 ans	9	1	6	9	2	5	4	4	8	
25-49 ans	11	5	0	1	6	9	2	4	10	
≥ 50 ans	8	5	3	3	7	6	2	7	7	
Total	11	5	0	4	6	6	0	10	6	
Moyennes de chômage										
Mars 71 à Mars 74 (PDRE)	439 522			419 793			421 081			426 799
Avril-Mai 75										
à Mars 81 PDRE	1 177 988			1 125 501			1 130 203			1 144 564
BIT	1 211 284			1 176 096			1 157 896			1 181 759
Octobre 77										
à Octobre 81 PDRE	1 558 735			1 485 025			1 475 899			1 506 553
BIT	1 540 301			1 483 078			1 476 814			1 500 065
Avril-Mai 82										
à Mars 85 BIT	2 132 398			2 083 661			2 078 693			2 098 250

On a établi pour chaque enquête emploi la proportion de chômeurs dans la population totale dans chaque sous-échantillon. Exprimée en % et arrondie à un chiffre après la virgule, chaque proportion est comparée à celle qui vient de l'ensemble de l'échantillon : elle est supérieure (+), égale (=) ou inférieure (-). La comparaison se fait par sexe et tranche d'âge.

Deux concepts du chômage sont retenus :

— la PDRE pour les enquêtes de mars 71 à mars 81 et d'octobre 77 à octobre 81;

— le chômage au sens du BIT pour les enquêtes d'avril-mai 75 à mars 85 et d'octobre 77 à octobre 81.

Dans les deux cas, il y a donc 16 observations.

D'autres part, sur chacune des 4 séries homogènes d'enquêtes des moyennes de chômage par sous-échantillon ont été effectuées.

Enfin, pour les années 1971 à 1976 il s'agit de fausses premières enquêtes dans le sous-échantillon entrant, dans la mesure où tous les sous-échantillons avaient été interrogés en 1968. Mais ceci n'affecte sans doute pas le résultat puisqu'on a commencé dans ce tableau l'observation en 1971, première année où l'on peut vraiment distinguer un 1^{er}, un 2^e et un 3^e sous-échantillon.

Source : enquêtes sur l'emploi.

Interaction avec la structure sociale?

Je crois qu'il peut y avoir aussi — et, d'une certaine façon, c'est encore plus intéressant — une interaction entre la réponse et l'état global de la société, et pas seulement une interaction interindividuelle entre enquêté et enquêteur. Je prends pour exemple l'origine sociale, c'est-à-dire la catégorie sociale à laquelle l'enquêté déclare que son père appartenait (à un moment précis : quand lui, l'enquêté, finissait ses études). Ayant effectué plusieurs enquêtes comparables sur ce sujet depuis trente ans, nous pouvons pour les hommes d'une génération donnée rapprocher la structure sociale qu'à différents moments du temps ils ont déclarée pour leur père (tableau 3, issu de [9]). Ce qui est très remarquable, c'est que ces déclarations non seulement ne sont pas stables, mais surtout évoluent dans le même sens que la société globale : à mesure qu'ils vieillissent et qu'ils vivent dans une société de moins en moins rurale ces hommes sont de moins en moins nombreux à déclarer que leur père était paysan, de plus en plus nombreux à déclarer que leur père était cadre ou technicien.

Certes les différentes enquêtes ne sont pas absolument comparables, notamment la première (faite en 1953) qui se distingue des trois autres. Cependant, on peut peut-être faire l'hypothèse d'une sorte de contagion de la structure sociale globale sur la façon dont le milieu d'origine est déclaré, contagion parfois facilitée par la relative imprécision de ce dernier, jointe à une défaillance de la mémoire : le père peut avoir changé de position au cours de sa propre vie active (et alors, dans ce cas, c'est bien en moyenne dans le sens de l'évolution globale de la société), sans que son fils se souvienne avec certitude si ce changement a eu lieu avant ou après que lui-même a achevé ses études.

Si cette hypothèse de contagion était confirmée, alors ce serait assez grave car, pour certaines informations, la possibilité d'élaborer des séries temporelles homogènes précises serait affectée. Cela étant, il faut à nouveau tenir ensemble les deux bouts de la chaîne : dans ce cas, reconnaître la dépendance de l'observation vis-à-vis de la structure sociale globale n'empêche pas de constater, comme prévu, qu'à un moment donné l'origine paysanne est moins fréquente et l'origine cadre plus fréquente parmi les jeunes générations que parmi les plus anciennes.

Le processus d'enregistrement

En continuant de parcourir le dispositif d'observation, on aboutit, après la définition des cadres généraux et le prélèvement de l'information, au processus d'enregistrement de cette information. Définir un protocole d'observation, ce n'est pas seulement en exposer les grandes lignes ou les principes, c'est, vu leur importance, décrire par le menu les modalités les plus concrètes de l'enregistrement, tout au long de la chaîne qui va de l'unité enquêtée au fichier.

Je voudrais illustrer le rôle très important des modalités les plus concrètes du recueil de l'information par l'exemple des changements d'emploi. Comparer la situation professionnelle d'une même personne à deux moments distincts d'un an, par exemple, ou de cinq ans, peut se faire par l'une ou l'autre méthode suivante :

— soit on interroge la personne en une fois sur ses deux situations, et l'on fait donc appel à sa mémoire (modalité a);

— soit on l'interroge à deux reprises, et c'est le statisticien qui opère le rapprochement. Et ce rapprochement peut s'opérer de deux façons différentes, soit directement sur la base des déclarations littérales (modalité b1) soit automatiquement, une fois ces déclarations transformées en codes (modalité b2).

Il se trouve qu'au cours de ces quinze dernières années l'INSEE a expérimenté ces trois modalités. Deux conclusions majeures se dégagent :

— Les résultats auxquels on aboutit sont souvent très différents. Comparons, par exemple, la solution a (mise en œuvre lors de l'échantillon entrant de l'enquête emploi) et la solution b1 (mise en

TABLEAU 3. — *Structure sociale des pères, selon les déclarations successives de leurs fils*

Génération	Enquêtes	En %			
		1953	1964	1970	1977
Génération 1					
(nés entre 1914 et 1923, en France avant 1946)					
Paysan		35,2		37,0	
Artisan, petit commerçant		15,3		13,2	
Cadre		6,2		11,0	
Employé		8,6		10,5	
Ouvrier		34,7		28,3	
	Ensemble	100,0		100,0	
Génération 2					
(né entre 1924 et 1933, en France avant 1946)					
Paysan		38,5		33,1	32,1
Artisan, petit commerçant		12,3		11,7	11,9
Cadre		8,5		11,3	12,2
Employé		11,0		11,2	10,3
Ouvrier		29,7		32,7	33,5
	Ensemble	100,0		100,0	100,0
Génération 3					
(nés entre 1925 et 1934, Français)					
Paysan			32,6	32,4	31,5
Artisan, petit commerçant			12,4	11,8	12,0
Cadre			11,3	11,9	12,5
Employé			9,7	11,2	10,3
Ouvrier			34,0	32,7	33,7
	Ensemble		100,0	100,0	100,0
Génération 4					
(nés entre 1935 et 1944, Français)					
Paysan			27,4	26,4	25,3
Artisan, petit commerçant			12,1	11,0	11,8
Cadre			9,5	14,2	15,7
Employé			12,0	13,3	11,2
Ouvrier			39,0	35,1	36,0
	Ensemble		100,0	100,0	100,0
Génération 5¹					
(nés entre 1941 et 1950)					
Paysan				23,9	22,2
Artisan, petit commerçant				10,0	10,1
Cadre				13,9	16,9
Employé				12,6	11,2
Ouvrier				39,6	39,6
	Ensemble			100,0	100,0

1. La comparaison n'est pas absolument pure puisqu'on ne s'est pas limité, en 1977, aux personnes de cette génération déjà en France en 1970. Mais la différence est négligeable.

Source : [9].

œuvre dans les deux sous-échantillons communs à deux enquêtes successives) (2). D'une manière générale, les changements d'emploi sont systématiquement plus faibles quand ils sont mesurés par la solution a (appel à la mémoire) que quand ils le sont par la solution b1 (rapprochement, sur documents, de deux déclarations indépendantes). Et l'écart est parfois considérable (tableau 4, issu de [10]). Ainsi 64,4 % des chômeurs de mars 1984 (a) ou 57 % (b1) sont chômeurs un an après (en mars 1985); ou, pire, parmi les hommes actifs occupés en mars 1984 et mars 1985 2,7 % (a) ou 6,4 % (b1) ont changé de catégorie socioprofessionnelle; parmi les femmes, les proportions correspondantes sont 1,9 % (a) et 4,3 % (b1); ou encore, parmi l'ensemble des actifs occupés 3,4 % (a) ou 10,3 % (b1) ont changé soit de statut soit de secteur d'activité économique. Les taux globaux de mobilité peuvent ainsi varier du simple au double, voire au triple, ce qui est considérable.

— En second lieu, il n'est pas toujours possible de dire avec certitude quelle est la « bonne » mesure, et ceci montre les limites du raisonnement qui ne parle que d'erreur de mesure. Certes, la solution b2 a été reconnue indiscutablement moins bonne que les deux autres, car le rapprochement automatique de deux codes chiffrés indépendamment l'un de l'autre génère une mobilité fictive très importante, notamment dans les cas où les frontières sont floues, entre cadres supérieurs et cadres moyens par exemple [11]. Mais entre les solutions a et b1 on peut hésiter : d'un côté rôle de la mémoire, de l'autre interprétation du personnel de chiffrage au vu des deux déclarations littérales. Dans nos publications, nous avons en réalité privilégié la solution b1 sans pour autant affirmer qu'il s'agissait de la « bonne » mesure; d'ailleurs, dans certains cas sont publiés les résultats des deux modalités, a et b1.

Ainsi, plutôt que raisonner exclusivement en termes d'erreur de mesure, il vaut mieux reconnaître que l'idée d'une bonne mesure absolue, c'est-à-dire entendue indépendamment d'un protocole d'observation, est dépourvue de sens. Et la dépendance à l'égard du protocole n'est pas toujours du second ordre, comme le montre cet exemple. La nécessité de se fonder sur un protocole stable ou à peu près stable pour apprécier des évolutions en est, dans ce cas, renforcée : on conclut alors que la mobilité sectorielle a sensiblement diminué depuis 1976 [12].

Les trois illustrations précédentes portent sur les enquêtes par sondage. Il est évident, même si ce peut être parfois dans un sens différent, que l'idée maîtresse s'applique aux données issues de fichiers administratifs : la dépendance de l'observation à l'égard des règles et conventions qui président à la constitution de ces fichiers est, dans ce cas, patente. Dans ce cas, en outre, le statisticien n'est pas souvent maître de ces règles et conventions.

Si toute observation est relative à un protocole, ce dernier étant défini par la description minutieuse et concrète de toutes les étapes qui conduisent au fichier, il paraît essentiel que dans tous les supports de diffusion — de la publication à la banque de données —, le protocole soit décrit aussi exhaustivement que possible.

2. L'ANALYSE

La phase d'analyse commence à partir du moment où on dispose d'un fichier. Elle recouvre la constitution de codes synthétiques — cette constitution est, en réalité, à cheval entre ce que j'ai appelé ici l'observation et l'analyse —, le redressement des non-réponses, le choix et l'analyse des indices et des tableaux qui seront tirés et publiés, enfin les méthodes autres que les tableaux qui seront mises en œuvre pour étudier les données.

2. L'enquête annuelle sur l'emploi est une enquête aréolaire au 1/300^e, dont l'échantillon est renouvelé par tiers : chaque année, l'échantillon est constitué d'un tiers d'aires (de 40 logements environ) où les ménages sont interrogés pour la première fois, et de deux tiers qui figuraient déjà l'année ou les deux années précédentes. Dans ces deux tiers d'aires communes à deux enquêtes successives, pour les ménages n'ayant pas déménagé et ayant répondu à l'enquête précédente (environ 85 %), c'est la solution b1 qui est utilisée; pour les autres, comme dans l'échantillon entrant, on utilise l'appel à la mémoire. La mesure de la mobilité professionnelle sur les deux sous-échantillons communs à deux enquêtes successives est donc mixte, mais très proche d'une mesure pure de type b1.

TABLEAU 4. — *Changements de situation observés dans l'échantillon entrant (e) et dans les sous-échantillons communs (c) de l'enquête emploi de mars 1985.*

Situation en mars 1985		Actifs occupés	Chômeurs	Inactifs	Ensemble
Situation en mars 1984					
Actifs occupés	{E	93,6	3,3	3,1	100
	{C	92,4	3,6	4,0	100
Chômeurs	{E	28,1	64,3	7,6	100
	{C	29,6	56,8	13,6	100
Inactifs	{E	4,1	2,2	93,7	100
	{C	5,2	2,5	92,3	100

Proportion d'actifs occupés en mars 84 et mars 85, ayant changé de catégorie socioprofessionnelle (%)

HOMMES			FEMMES		
Catégorie socioprofessionnelle (niveau agrégé, intermédiaire pour les ouvriers) l'année N-1	E	C	Catégorie socioprofessionnelle (niveau agrégé, intermédiaire pour les employées) l'année N-1	E	C
1. Agriculteurs exploitants	0,4	1,6	1. Agriculteur exploitants	-	1,2
2. Artisans, commerçants et chefs d'entreprise	2,8	4,6	2. Artisans, commerçants et chefs d'entreprises	1,9	3,3
3. Cadres et professions intellec- tuelles supérieures	1,2	2,5	3. Cadres et professions intellec- tuelles supérieures	0,8	3,4
4. Professions intermédiaires	2,4	5,6	4. Professions intermédiaires	1,8	4,3
5. Employés	4,2	8,8	51. Employés de la fonction publique	1,3	2,9
61. Ouvriers qualifiés	2,9	5,6	54. Employés administratifs d'entreprises	2,7	5,6
66. Ouvriers non qualifiés	3,3	15,7	55. Employés de commerce	4,0	5,8
69. Ouvriers agricoles	6,9	6,1	56. Personnel des services directs aux particuliers	3,4	6,5
TOTAL	2,7	6,4	6. Ouvriers	1,1	4,7
			TOTAL	1,9	4,3

Proportion d'actifs occupés en mars 84 et 85, ayant changé de secteur x statut (%)

Statut et secteur en mars 1984	E	C
Non-salariés de l'agriculture	0,2	1,4
Non-salariés de l'industrie (y. c. BTP)	2,1	6,2
Non-salariés du tertiaire	2,5	4,5
Salariés de l'agriculture	6,9	9,1
Salariés de l'industrie (y. c. BPT)	3,7	12,3
Salariés du tertiaire	3,8	11,4
Ensemble	3,4	10,3

E : sous-échantillon entrant C : sous échantillons communs.

Source : [10].

De nouveau, toute analyse (et donc tout résultat) est relative à un choix de mesure ou à ce qu'on pourrait aussi appeler un point de vue, un angle d'attaque, qui peut ou non être formalisé dans un modèle. Je vais illustrer cette affirmation générale à partir de quelques exemples, du plus humble — le taux ou le tableau — au plus sophistiqué — le modèle économétrique. Auparavant il faut évoquer le redressement des non-réponses.

Le redressement des non-réponses

On peut s'étonner que le redressement des non-réponses soit situé dans la phase d'analyse : même si la distinction entre observation et analyse n'est pas très nette, peut-être aurait-on préféré spontanément placer cette opération dans la phase d'observation. Je la range cependant sous le chapeau de l'analyse car, quelle que soit la façon dont on redresse les non-réponses, elle repose sur un modèle, explicite ou non. La démarche la plus usuelle, la stratification a posteriori, le montre bien : au sein d'une « case » constituée par un croisement de critères appropriés, on admet que les non-répondants ont, en espérance, des réponses identiques aux répondants, et on affecte donc à ces derniers un coefficient légèrement supérieur à 1. On suppose donc que, conditionnellement à des variables x , les variables d'intérêt y et le fait de répondre ou non sont indépendants.

D'autres hypothèses sont naturellement envisageables pour redresser les non-réponses. Le modèle « tobit généralisé », par exemple, peut ici s'appliquer de façon intéressante [13]. On suppose :

- 1) qu'une variable latente l règle la non-réponse et qu'elle s'explique par des caractéristiques (connues) z de l'unité enquêtée;
- 2) que la variable d'intérêt y qu'on veut redresser s'explique, elle aussi, par des caractéristiques connues x de l'unité enquêtée (les mêmes que z ou d'autres);
- 3) qu'il y a corrélation entre la variable d'intérêt et la variable latente; on fait alors explicitement l'hypothèse que la non-réponse est liée à la variable d'intérêt (pour une discussion précise, voir [14]) (3).

Selon les variables z et x et selon la loi du couple des résidus des deux équations, on a ainsi toute une série de modèles explicites, dérivant d'une même architecture générale, à partir desquels les non-réponses seront redressées.

Naturellement, selon le modèle retenu, le résultat de l'enquête ne sera pas le même. Ceci peut être illustré à partir du salaire dans l'enquête emploi. L'observation sur les répondants et l'estimation

3. Le modèle s'exprime alors pour l'individu i , par les deux équations suivantes :

$$\begin{cases} \text{(a)} & l_i = z_i' \gamma + \eta_i \quad \text{où} \quad E(\eta_i) = 0, V(\eta_i) = \sigma_\eta^2, \text{cov}(\eta_i, \eta_j) = 0 \\ \text{(b)} & y_i = x_i' \beta + \varepsilon_i \quad \text{où} \quad E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma_\varepsilon^2, \text{cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ et } \text{cov}(\varepsilon_i, \eta_i) \neq 0 \end{cases}$$

Si $l_i > 0$ il y a réponse (c'est-à-dire que y_i est connu); si $l_i \leq 0$, il y a non-réponse

La méthode de redressement est la suivante : l'estimation de ce modèle — l'équation (a) sur tous, l'équation (b) sur les seuls répondants — conduit à des estimateurs β et γ et l'on affecte au non-répondant qui a les caractéristiques x_k : $y = E(y/x_k \text{ et } l_k \leq 0)$. Pour éviter d'affecter à chaque non-répondant l'espérance (sous condition qu'il n'a pas répondu) correspondant à ses caractéristiques, on peut le remplacer par n individus fictifs ayant les caractéristiques x_k à qui on affectera $y_k/l_k \leq 0$, où ε_k est obtenue par tirage dans la loi estimée et donner à chacun de ces individus un poids égal à $1/n$.

D'autre part, si on suppose $\text{cov}(\varepsilon, \eta) = 0$, l'introduction de l'équation (a) de non-réponse est inutile : on affectera dans ce cas au non-répondant $E(y_k/x_k) = x_k' \beta$.

Enfin, la stratification a posteriori est, dans son principe, très proche de ce dernier cas : dans les deux démarches, on suppose en effet que $E(y/x_k)$ est identique chez les répondants et chez les non-répondants. Pratiquement, le redressement est cependant différent puisque dans un cas on donne un poids supérieur à 1 aux répondants, dans l'autre on affecte une valeur y aux non-répondants. Cependant l'estimation de la moyenne y est bien la même, comme on le voit aisément :

— stratification a posteriori : on part des cases h définies par la variable x ; dans h il y a n_h individus : n_h^r ont répondu n_h^{nr} n'ont pas répondu. Alors $\bar{y} = \frac{1}{n} \sum_{h=1}^H \sum_{i=1}^{n_h} \left(1 + \frac{n_h^{nr}}{n_h^r}\right) y_{ih} = \frac{1}{n} \sum_{h=1}^H n_h \bar{y}_h$, où \bar{y}_h est la moyenne sur les répondants.

— affectation dans le cas où $\text{cov}(\varepsilon, \eta) = 0$: on part de l'équation $y_{ih} = \beta_h + \varepsilon_{ih}$. On en déduit $\beta_h = \bar{y}_h$; on affecte donc \bar{y}_h à tous les non-répondants ayant la caractéristique x_h . La moyenne s'écrit alors :

$$\bar{y} = \frac{1}{n} \sum_{h=1}^H \left[\sum_{i=1}^{n_h^r} y_{ih} + \sum_{j=1}^{n_h^{nr}} \bar{y}_h \right] = \frac{1}{n} \sum_{h=1}^H (n_h^r \bar{y}_h + n_h^{nr} \bar{y}_h), \text{ ce qui est la même expression que celle obtenue dans la}$$

stratification a posteriori.

TABLEAU 5. — *Médiane, 1^{er} et 9^e déciles observés et redressés des salaires annuels déclarés à l'enquête emploi d'avril-mai 1982 (en F)*

	Observés	Redressés (4 méthodes)			
		Stratification a posteriori (S)	régression linéaire (RL)	tobit généralisé (T)	logistique à deux dimensions (L)
HOMMES					
D1	3301	3306	3301	3323	3317
M	4713	4739	4742	4804	4774
D9	8751	8904	8952	9216	9066
FEMMES					
D1	2952	2953	2948	2955	2951
M	4111	4117	4119	4131	4124
D9	6624	6656	6654	6677	6662

Champ : Salariés à temps complet. Pour plus de détails, cf. [14]. On a retenu ici volontairement les résultats de 1982, car l'effet du redressement y est plus sensible qu'en 1983, 1984, 1985. Le taux de non-réponse au salaire a été en effet cette année-là de 5,0 %. Pour la définition précise des 4 façons de redresser les non-dépenses, cf. le texte ou [14].

obtenue par quatre redressements (S : stratification a postériori; RL : modèle précédent sans corrélation entre variable d'intérêt et variable latente; T : modèle précédent où le couple est une variable normale à deux dimensions; L : modèle précédent où le couple est une variable logistique à deux dimensions) figurent sur le tableau 5.

Les deux premières méthodes conduisent à des estimations proches, les deux dernières aussi. Mais entre ces deux groupes, la différence n'est pas négligeable, par exemple sur le neuvième décile chez les hommes : de l'ordre de 250 à 300 F d'écart, soit environ 3 %. Et, à nouveau, on ne peut guère garantir qu'un redressement est meilleur qu'un autre.

On dispose donc maintenant d'un fichier propre, redressé, et les nomenclatures, codes, regroupements, etc... ont été fixés. C'est alors le moment de l'analyse stricto sensu.

L'évolution temporelle des disparités de chômage

Un taux, ou une proportion, telle est la donnée la plus élémentaire qu'un fichier fournit. Dès que ce taux est obtenu à deux moments différents on va s'interroger sur l'évolution des disparités. Ainsi, dès le début des années soixante le taux de chômage des femmes était plus élevé que celui des hommes (graphique I). D'où la question : les disparités entre hommes et femmes se sont-elles atténuées ou, au contraire, accrues? La réponse est : on ne sait pas ou, plus précisément, cela dépend de la façon dont on choisit de mesurer les disparités.

En effet, en termes de disparités absolues, il faut conclure à une accentuation : fin 1962, la différence des taux de chômage féminin et masculin était de 1,9 point; fin 1984, elle est de 5 points. En termes de disparités relatives en revanche, il faut conclure à une atténuation : fin 1962, le taux de chômage des femmes était 2,7 fois plus élevé que celui des hommes; fin 1984, 1,6 fois seulement. Dans beaucoup de cas, cette situation se retrouve et, parfois, alimente sans fin le débat public. Par exemple, l'évolution respective de la proportion de fils de cadres supérieurs et d'ouvriers dans l'enseignement supérieur, à propos de laquelle la Revue française de sociologie s'est interrogée sur ces questions de mesure des inégalités [15], ou l'évolution comparée de la hausse des prix de détail en France et en République fédérale d'Allemagne au cours des quatre ou cinq dernières années relèvent de cette difficulté : ainsi en 1981, la hausse des prix (en glissement) a été de 13,9 % en France et de 6,3 % en RFA; en 1985, elle a été de 4,7 % en France et de 1,8 % en RFA. Dans quel pays la diminution de la hausse a-t-elle été la plus rapide?

On ne sait pas répondre, et ceci d'autant moins que d'autres mesures des disparités sont concevables, par exemple l'évolution logistique. A chaque mesure sa conclusion. Le statisticien n'a pas,

me semble-t-il, d'arguments décisifs en faveur de telle ou telle façon d'apprécier les disparités. Dans le cas des taux de chômage, j'ai plutôt choisi la mesure absolue (en points), auquel cas on conclut que la disparité hommes/femmes s'est accrue. Il est vrai que, parfois, disparités absolues et disparités relatives vont dans le même sens : taux de chômage par tranche d'âge, ou par catégorie sociale. Dans ces cas, la conclusion est évidemment plus assurée.

La table de mobilité sociale

Moins élémentaire, mais encore simple, le tableau à deux ou plusieurs entrées est le mode le plus fréquent de présentation des données et donc le fondement le plus général des analyses. Je voudrais montrer sur un exemple que le tableau procède toujours d'un point de vue. Cela ne signifie pas qu'il ne dit rien naturellement, mais cela signifie que ses enseignements doivent être appréhendés uniquement à la lumière des présupposés qui ont conduit à le construire.

Prenons ainsi une table de mobilité sociale, c'est-à-dire un tableau à double entrée qui croise le milieu d'origine et la position actuelle. Il est inutile de revenir sur les questions de nomenclature et de précisions, pourtant essentielles ici. Une telle table peut se lire de deux façons (tableau 6, issu de [1]) :

— soit en termes de *destinée* : en 1977, sur 100 hommes de 40-59 ans fils de cadres supérieurs, 43 étaient cadres supérieurs eux-mêmes;

— soit en termes de *recrutement* : en 1977, sur 100 cadres supérieurs de 40-59 ans, 16 avaient un père cadre supérieur.

TABLEAU 6. — Destinées et recrutements

Destinées sociales en 1953, 1970, 1977
(Fils : hommes actifs occupés de 40 à 59 ans)

Catégorie sociale du fils		Agriculteurs exploitants	Salariés agricoles	Artisans petits commerçants	Industriels gros com. prof. libérales	Cadres supérieurs	Cadres moyens	Employés, pers. de service	Ouvriers	Ensemble
Agriculteurs exploitants	1953	60	6	8	(1)	(1)	2	5	17	100
	1970	42	6	6	1	2	4	8	31	100
	1977	38	3	6	2	3	4	9	35	100
Salariés agricoles	1953	14	26	13	(-)	(-)	(-)	7	40	100
	1970	9	16	7	(1)	(1)	4	9	53	100
	1977	4	13	6	(1)	(1)	4	10	61	100
Artisans petits commerçants	1953	4	(2)	48	3	3	6	8	26	100
	1970	3	2	26	7	9	10	12	31	100
	1977	3	(1)	22	6	12	10	14	32	100
Industriels gros commerçants, profession libérales	1953	(3)	(1)	(10)	43	(12)	(7)	(7)	17	100
	1970	(1)	(1)	10	26	23	11	14	14	100
	1977	(1)	-	14	21	26	12	10	16	100
Cadres supérieurs	1953	(4)	(-)	(16)	(6)	41	(17)	(14)	(2)	100
	1970	(3)	(1)	(5)	9	44	18	9	11	100
	1977	(2)	-	(4)	10	43	22	9	10	100
Cadres moyens	1953	(-)	(-)	(16)	(12)	(16)	(16)	(19)	(21)	100
	1970	(-)	(-)	(4)	(5)	28	29	15	19	100
	1977	(1)	(-)	(4)	6	30	30	11	18	100
Employés, personnels de service, autres actifs	1953	(5)	(-)	10	(4)	12	14	17	38	100
	1970	(1)	(1)	10	(3)	13	18	17	37	100
	1977	(1)	(1)	7	2	16	21	17	35	100
Ouvriers	1953	3	2	11	(1)	1	9	11	62	100
	1970	2	1	7	1	5	12	15	57	100
	1977	1	1	7	2	6	11	14	58	100
Ensemble	1953	23	6	16	3	3	6	9	34	100
	1970	14	4	9	4	8	10	12	39	100
	1977	12	2	8	4	10	11	12	41	100

Source : Enquêtes sur l'emploi de 1953 et FQP de 1970 et 1977. Les proportions mises entre parenthèses sont incertaines.

*Recrutement des catégories sociales en 1953, 1970, 1977
(Fils : hommes actifs occupés de 40 à 59 ans)*

Catégorie sociale du fils		Agriculteurs exploitants	Salariés agricoles	Artisans petits commerçants	Industriels gros com. prof. libérales	Cadres supérieurs	Cadres moyens	Employés, pers. de service	Ouvriers	Ensemble
Agriculteurs exploitants	1953	85	36	16	(12)	(7)	12	19	17	33
	1970	88	48	18	11	7	11	20	23	30
	1977	89	42	20	11	8	9	20	23	27
Salariés agricoles	1953	6	48	8	(-)	(1)	(1)	9	12	10
	1970	5	32	5	(2)	(1)	3	6	10	7
	1977	2	34	4	(2)	(1)	2	5	9	6
Artisans petits commerçants	1953	3	(4)	47	17	15	17	15	12	16
	1970	2	6	36	24	15	13	12	10	13
	1977	3	(6)	32	22	15	12	14	10	13
Industriels gros commerçants, profession libérales	1953	(-)	(1)	(2)	44	(9)	(3)	(2)	1	3
	1970	(-)	(1)	4	28	11	4	4	1	4
	1977	(1)	(-)	7	24	11	5	3	1	4
Cadres supérieurs	1953	(-)	(-)	(2)	(4)	24	(6)	(3)	(-)	2
	1970	(1)	(1)	(2)	(9)	19	6	3	1	3
	1977	(1)	(-)	(2)	11	16	8	3	1	4
Cadres moyens	1953	(-)	(-)	(1)	(7)	(8)	(5)	(4)	(1)	2
	1970	(-)	(-)	(2)	(5)	12	10	4	2	4
	1977	(-)	(1)	(2)	7	13	12	4	2	4
Employés, personnels de service, autres actifs	1953	(2)	(-)	4	(10)	24	16	14	8	7
	1970	(1)	(2)	11	(9)	17	19	15	10	10
	1977	(1)	(3)	8	6	16	20	15	9	10
Ouvriers	1953	4	11	20	(6)	12	40	34	49	27
	1970	3	10	22	12	18	34	36	43	29
	1977	3	14	25	17	20	32	36	45	32
Ensemble	1953	100	100	100	100	100	100	100	100	100
	1970	100	100	100	100	100	100	100	100	100
	1977	100	100	100	100	100	100	100	100	100

Source : Enquête sur l'emploi de 1953 et FQP de 1970 et 1977. Les proportions mises entre parenthèses sont incertaines.

Source : [1].

Dès que la structure sociale évolue, donc est différente chez les pères et chez les fils, ces deux modes de lecture ne sont pas équivalents : déjà à ce niveau le plus élémentaire, le point de vue choisi joue donc un rôle déterminant. On pourra conclure que la classe des cadres supérieurs est assez fermée (puisque 43 % de ses descendants en font partie, contre 10 % seulement de l'ensemble des fils) ou assez ouverte puisque son recrutement est très différencié (16 % sont fils de cadres supérieurs, 20 % sont fils d'ouvriers, etc...). De même pour la classe ouvrière : fermée (58 % des fils d'ouvriers le sont eux-mêmes à 40-59 ans), ou ouverte (32 % des ouvriers de cet âge sont fils de paysans). En poussant à l'extrême limite, la structure sociale se reproduit, ou bien elle est très fluide.

En second lieu, il faut admettre, plus ou moins explicitement, une certaine stabilité des nomenclatures : on nomme ainsi « immobiles » ceux qui sont dans la même catégorie sociale que leur père, mobiles les autres. Rien ne dit évidemment que les groupes sociaux son situés aux mêmes endroits respectifs de l'espace social qu'il y a trente ans; tout porte à penser en réalité le contraire. Mais admettons qu'il en est bien ainsi. On conclut alors que la proportion d'immobiles dans la société française a beaucoup diminué en trois décennies. Se pose alors la question du motif, et donc de l'analyse elle-même de la table de mobilité. Et là, il faut prendre l'un des deux points de vue suivants :

— soit on admet que la transmission du statut social est le moteur de la transformation de la structure sociale : par exemple, c'est parce que les cadres et employés font tout pour que leurs descendants — même s'ils sont peu diplômés — ne « tombent » pas dans la classe ouvrière que les positions d'employés et de cadres prennent de plus en plus d'importance;

— soit on postule le contraire : la structure sociale globale évolue sous l'effet du développement et c'est dans ce cadre que se fait la transmission du statut social. Ainsi, dans cette perspective, les fils de paysans sont contraints à quitter la terre.

Dans mon livre tout en reconnaissant que les deux effets peuvent jouer (et différemment tout au long de la structure sociale), j'ai privilégié le second point de vue. On conclut alors que la cause principale de l'augmentation de la mobilité sociale depuis trente ans réside dans l'évolution de la structure sociale elle-même. Si l'on met à part cette cause majeure, la fluidité de la société française s'est peu accrue [1]. Il me paraît plausible, sur cette base, de prévoir que dans les décennies futures la mobilité sociale devrait évoluer moins vite que depuis la guerre : il me semble en effet peu probable que la structure sociale connaisse un bouleversement aussi considérable que celui des dernières décennies.

Quoi qu'il en soit, une table de mobilité conduit donc à raisonner et à analyser le phénomène en termes de transition des individus, ou des familles entre positions sociales. Et l'on voit ainsi émerger un présupposé essentiel de la table : pour conclure que le recrutement des cadres supérieurs est ouvert, il faut avoir admis que les cadres supérieurs fils d'ouvriers et les cadres supérieurs fils de cadres supérieurs sont comparables, donc que dans la définition de la position sociale, n'intervenait pas l'origine sociale. Or, on sait que, sur nombre de dimensions (vie scolaire, vie familiale, insertion dans le système productif) les ouvriers diffèrent nettement selon leur milieu d'origine, la distinction majeure passant entre les fils de paysans et les autres; de même pour les cadres supérieurs, où le clivage le plus important sépare les fils de cadres supérieurs des autres [16]. Ces résultats ne permettent pas forcément de rejeter la pertinence du point de vue qu'exprime une table de mobilité. Mais ils illustrent à quel point il s'agit bien d'une sorte de coup de force théorique, préalable au tableau, aussi simple soit-il.

Analyse factorielle et classification

A côté du tableau ont pris une place grandissante ces 15 dernières années, d'autres méthodes d'analyse des données, qui sont maintenant bien connues : analyses factorielles et classifications automatiques. Ces méthodes sont séduisantes pour plusieurs raisons :

- elles possèdent des propriétés d'optimalité. Ainsi l'analyse factorielle est la meilleure approximation du nuage de points dans un espace de dimension donnée [17];
- elles sont particulièrement adaptées à des fichiers de grande taille, par le nombre d'observations ou de variables;
- désormais « presse-bouton », elles aboutissent à des résultats d'abord facile : des graphiques ou des regroupements.

Elles sont donc, avec le développement de l'informatique, de plus en plus utilisées, plus par les sociologues quantitatifs et les statisticiens d'enquête que par les économistes, et plus sur données individuelles que sur données temporelles.

En un sens, elles paraissent plus à l'abri d'un choix préalable qu'un tableau : on met « dans le même sac » beaucoup de variables au lieu de deux ou trois, on en fait la synthèse — synthèse optimale, qui plus est — et on regarde. En réalité, elles aussi procèdent d'un point de vue et c'est à partir de là qu'elles sont optimales. Ce point de vue s'exprime dans la métrique choisie d'une part, dans les variables retenues d'autre part. Mettre par exemple dans une analyse des correspondances multiples n variables de capital culturel (âge de fin d'études, niveau de diplôme, niveau de formation,...) et une variable de capital économique conduit à « faire sortir » le premier effet plutôt que le second : le premier axe aura plus de chances d'être un axe de capital culturel. Or il n'est pas toujours facile, en présence des multiples variables que l'on met d'ordinaire dans ces analyses de reconnaître si on a, implicitement et involontairement, privilégié tel ou tel aspect.

Plus grave est la dépendance des résultats obtenus vis-à-vis des observations entrant dans l'analyse : en particulier, la classification automatique est peu robuste, en ce sens que les regroupements obtenus peuvent être assez différents selon qu'on inclut ou non telle ou telle observation. Ainsi, dans

une classification ascendante hiérarchique de 44 lignées de cadres supérieurs sur la base de 42 variables décrivant leur vie scolaire, familiale et professionnelle [16], il suffisait de retirer une lignée — il est vrai assez particulière — pour obtenir des regroupements différents. Dans le même esprit, les axes factoriels et l'interprétation qu'on est amené à leur donner peuvent varier suivant que tel groupe d'observations qui tire à lui tout seul un axe, par exemple, est retenu ou non : les personnes n'ayant pas répondu à tout ou partie des variables de l'analyse constituent un exemple connu de ce phénomène, du moins s'il n'y a pas indépendance entre la non-réponse et les autres variables de l'analyse.

Même si ces méthodes sont souvent utilisées comme un aboutissement alors qu'elles devraient être situées en amont de l'analyse, en tant que technique de statistique descriptive préalable à la modélisation (même s'il est possible, en fait, de procéder à des tests), elles constituent un moyen puissant pour débroussailler les données, pour en extraire une « substantifique moëlle ». Mais, pas plus que les autres moyens d'étude, elles ne peuvent être regardées comme une analyse en soi, indépendante de tout pré-supposé.

Modèles d'analyse

Après la description plus ou moins synthétique des données, vient l'étape de la modélisation. Ici, c'est par définition qu'un modèle procède d'un point de vue. Chacune des équations économétriques de nos modèles postule et estime des influences causales et leur enchaînement — c'est-à-dire le modèle lui-même — constitue une maquette exprimant une théorie sur la réalité économique et sociale. Il n'est donc pas nécessaire de s'appesantir. Peut-être cependant faut-il faire une place particulière à l'analyse de variance. En effet, mettre par exemple dans une équation de salaire les variables âge et âge au carré dérive bien de l'idée que la productivité augmente, et de moins en moins (car le coefficient de l'âge au carré est négatif), avec l'âge. Mais introduire 10 variables dichotomiques fabriquées à partir de l'âge quinquennal paraît un moyen de s'affranchir de toute hypothèse a priori. Plus généralement, l'analyse de variance comme les modèles qualitatifs (modèles « probit » et « logit ») dont l'usage se répand rapidement depuis quelques années ont un objectif commun : mesurer des effets. C'est la question classique en statistique : y a-t-il, toutes choses égales d'ailleurs, c'est-à-dire en ayant éliminé les effets de structure, un effet d'une variable x sur la variable d'intérêt y ? Je pense que cette question est pertinente et que les moyens utilisés pour y répondre (analyse de variance, modèles « probit » et « logit ») sont appropriés, mais ils soulèvent deux questions directement liées à notre sujet.

En premier lieu, l'estimation de l'effet dû à une variable va dépendre, outre du degré de finesse retenu pour caractériser la variable, de l'ensemble des variables prises en compte : l'expression « toutes choses égales d'ailleurs » ne veut, en fait, rien dire tant qu'on n'a pas précisé comment on traduisait l'expression « toutes choses ». Et on ne mesurera ce qu'à tort peut-être on appelle un « effet pur » que dans un modèle donné, celui que traduit l'ensemble des variables effectivement retenues dans l'analyse.

Cette remarque ne porte pas à conséquence dans deux situations précises :

— la première situation, c'est celle où l'on pense qu'il y a, qu'on connaît et qu'on peut mettre en œuvre le « vrai » modèle. Dans ce cas, l'effet pur sera celui qui se dégagera dans ce cadre, et il importe peu que dans un autre cadre on obtienne quelque chose de différent;

— la seconde situation, c'est celle où il y a indépendance. Ainsi l'estimation obtenue dans l'analyse de variance à un facteur est la même que celle obtenue dans l'analyse à deux facteurs si, dans chaque case, les effectifs sont égaux à 1 ou au produit des effectifs des marges [18].

Cependant, dans les analyses concrètes de la réalité économique et sociale on ne se trouve jamais ni dans l'une ni dans l'autre situation, et il faut alors être conscient que l'effet « pur » obtenu est relatif au cadre général dans lequel on s'est situé.

Pour illustrer cette dépendance, on peut donner l'exemple des salaires dans le secteur privé. Deux analyses sont menées :

— l'une qui contient toutes les variables individuelles (sexe, âge, diplôme, catégorie sociale, nationalité, type de famille, origine sociale,...), collectives (région, tranche d'unité urbaine, taille de

l'entreprise,...) ou mixte (ancienneté) dont on dispose dans l'enquête emploi, soit environ 200 modalités dichotomiques;

— l'autre retient les mêmes variables, moins trois : le secteur d'activité, la taille et la fonction précise des cadres et techniciens.

Quel est alors l'« effet pur » d'un diplôme d'une grande école sur le salaire des hommes du secteur privé? + 26,3 % dans la première analyse, mais + 37,6 % dans la seconde, par rapport aux salariés qui n'ont pas de diplôme (enquête emploi de 1984). Cette différence importante s'explique mécaniquement par le fait que les trois variables supplémentaires — et notamment la fonction — « pompent » une partie de l'effet qui, si elles ne sont pas introduites, se « reporte » sur la variable diplôme. Si l'on pense — ce qui est raisonnable — que dans un « vrai » modèle d'explication des salaires le secteur d'activité et la taille de l'établissement d'une part, la fonction précise des salariés d'autre part doivent être introduits, car ces variables jouent effectivement un rôle, alors c'est la première mesure qui serait plus plausible.

Cette dépendance de l'effet mesuré vis-à-vis du modèle postulé impose de procéder à de multiples essais avant de conclure avec certitude à l'ampleur — voire à la réalité — de l'effet. Même ainsi, c'est-à-dire même si l'effet se confirme dans plusieurs cadres différents, il faut sans doute ensuite s'assurer de sa réalité en fournissant une interprétation ou une élucidation des mécanismes qui conduisent à l'observer. Ainsi, constater dans une analyse de variance que l'origine sociale influe (modestement, de l'ordre de 1 % à 2 % toutes choses égales d'ailleurs) sur le salaire requiert que l'économiste ou le sociologue dégage les voies par lesquelles s'exerce cette influence. Dans le même esprit, le statut d'occupation du logement a , semble-t-il, une influence importante, toutes choses égales d'ailleurs, sur le comportement d'équipement [19] ou d'assurance [20] des ménages. Ceci se comprend bien. Mais, et c'est plus curieux, il jouerait aussi en matière de chômage : non seulement le risque d'être chômeur, toutes choses égales d'ailleurs, serait plus grand pour les hommes locataires que pour les accédants ou les propriétaires, mais il semblerait aussi — bien que ce ne soit significatif qu'à un écart-type et non à deux — que les locataires ayant conservé un emploi auraient consenti un sacrifice sur leur salaire plus grand que les accédants ou propriétaires [21]. Comme pour l'influence du milieu d'origine sur le salaire, il paraît nécessaire de conforter ce genre de résultat, d'une part en multipliant les cadres d'analyse, d'autre part en mettant en lumière les mécanismes réels qui y conduisent.

La seconde question soulevée par la recherche d'effets purs peut être introduite par la boutade de Simiand citée par Halbwachs [22] : cela ne revient-il pas « à se demander comment vivrait un chameau si, restant chameau, il était transporté dans les régions polaires, et comment vivrait un renne si, restant renne, il était transporté dans le Sahara »?

Halbwachs fait cette remarque à propos de la mesure et de l'élimination des effets de structure (plus précisément, au sujet d'un taux comparatif de mortalité) mais elle s'applique évidemment ici (4).

4. Cette proximité peut être illustrée à partir de l'exemple de Halbwachs. Soit les deux pays 1 et 2, leur taux de mortalité m_a^1 et m_a^2 par âge, leur population P_a^1 et P_a^2 par âge. Leur taux de mortalité sont $m^i = \sum_a P_a m_a^i$, $i = 1, 2$, et m est le taux global sur les deux pays.

Au lieu de comparer m^1 et m^2 on compare les taux comparatifs $\sum_a P_a m_a^1$ et $\sum_a P_a m_a^2$. Ceci revient à dire que l'effet du pays 1 n'est pas $m^1 - m$ mais $\sum_a P_a m^1 - m$, soit $\sum_a P_a (m_a^1 - m_a)$. Et en faisant cela, on a bien éliminé l'effet de structure par âge, puisqu'on peut écrire $m^1 - m = \sum_a P_a (m_a^1 - m_a) + \sum_a m_a^1 (P_a^1 - P_a)$, où le second terme est bien une mesure de l'effet de structure par âge.

L'analyse de la variance s'introduit alors comme cela : si le pays 1 n'avait pas d'effet son taux m_a^1 serait une constante k_a . Le taux de mortalité global de 1 serait $\sum_a P_a^1 k_a$. La différence $m^1 - \sum_a P_a^1 k_a$ est alors bien une mesure de l'effet du pays 1. Or si on suppose $m_a = k_a$ quel est le meilleur estimateur de k_a ? C'est, en faisant l'analyse de variance $m^i = k_a + \varepsilon_{ai}$, $i = 1, 2$ où $V(\varepsilon_{ai}) = \frac{k_a(1-k_a)}{P_a^i}$, $\hat{k}_a = m_a$. L'effet de 1 est donc bien $\sum_a P_a^1 (m_a^1 - m_a)$. Sur ces différents points, voir [23].

Elle porte bien sur le sens de ce qu'on a estimé : l'analyse de variance conduit à casser les groupes sociaux ou humains réels pour aboutir à une abstraction dont, finalement, on ne sait plus très bien quelle signification elle a. Reprenons le cas du diplôme de grande école : + 26,3 %, cela veut dire par exemple qu'entre deux OS ne différant que par le diplôme — l'un n'en aurait aucun et l'autre serait sorti de Polytechnique —, il y a cet écart de salaire (5). Interprétation techniquement vraie mais difficile à défendre, non seulement parce qu'il n'est pas exact que dans la réalité un OS diplômé de grande école — après 1968, il y en eut — a ou aurait un tel avantage, mais surtout parce que ce diplôme de grande école est en fait incorporé ou possédé par d'autres personnes que par des OS. Le fonctionnement de la société n'est pas, comme on le suppose dans l'analyse de variance, analytique, mais synthétique ou structurel : un diplôme de grande école exclut d'être OS et c'est bien du même mouvement qu'il conduit à une autre position et à un autre salaire (et à bien d'autres choses différentes) que ceux d'un OS. À vrai dire, cette remarque porte plus à réfléchir sur le type de variable à introduire et le type d'utilisation des résultats du modèle que sur le modèle lui-même. L'exemple précédent est « idiot » parce qu'il porte sur une case quasi vide. D'ailleurs l'estimation elle-même (26,3 %) ne doit pratiquement rien à cette case mais plutôt aux populations où il y a à la fois des non diplômés et des diplômés de grandes écoles (les cadres supérieurs par exemple). De façon analogue, mesurer un effet de la taille de l'agglomération et un effet de la catégorie sociale ne doit pas conduire dans l'utilisation à comparer les paysans ruraux et les paysans urbains. Ici, la mesure de l'effet taille de l'agglomération s'effectue en fait sur les populations présentes à la fois en campagne et en ville, les ouvriers en particulier. Et si l'on veut prendre en compte le fait qu'il n'y a de paysans qu'à la campagne, plutôt qu'introduire deux variables séparées — catégorie sociale, taille de l'agglomération — il serait sans doute préférable de n'en introduire qu'une, judicieusement construite à partir d'un croisement.

Quoi qu'il en soit, ces deux remarques qu'appellent l'analyse de variance (et les modèles qualitatifs) soulignent l'extrême importance des variables qu'on y met sur l'ampleur des estimations obtenues.

**

Ainsi, toute statistique est une construction. Comme dans toute science, le résultat ne prend sens que rapporté au protocole d'observation et au cadre d'analyse qui l'ont produit. Pour autant, tous les protocoles ou tous les cadres ne sont pas équivalents : j'ai cité un cas où un protocole a été abandonné, parce que sa qualité a été jugée incontestablement moins bonne. Il y donc de bonnes statistiques, et il y en a de mauvaises; il y en a même de fausses. Mais y en a-t-il de vraies? Je pense que oui, à condition de ne pas considérer qu'une statistique est vraie parce qu'elle serait indépendante de tout présumé. Mais ceci nous fait entrer de plain-pied dans l'épistémologie de notre discipline, ce à quoi l'exposé précédent n'a voulu être qu'une introduction.

5. Il faut, en plus, dans le cas présent signaler une difficulté supplémentaire, technique elle, d'interprétation qui provient du fait que figurent dans l'analyse des effets croisés : âge x diplôme et catégorie sociale x diplôme. 26,3 % c'est l'effet pur de la variable diplôme elle-même. Pour être plus précis, il faut par exemple comparer deux ouvriers non qualifiés de type artisanal de moins de 30 ans, dont l'un serait diplômé d'une grande école et l'autre n'aurait pas de diplôme. L'effet est alors de :

26,3 %	+ (3,6 %)	= 29,9 %
<i>effet pur du diplôme</i>	<i>effet spécifique supplémentaire d'un diplôme de niveau I chez les ouvriers non qualifiés de type artisanal</i>	

Il est vrai que le second effet n'est pas significatif (à 5 %).

BIBLIOGRAPHIE

- [1] C. THÉLOT. « Tel père, tel fils? Origine familiale et position sociale », Dunod, 1982.
- [2] F. HERAN. « Compte-rendu de « Tel père, tel fils? », Population n° 1, janvier-février 1983
- [3] A. DESROSIÈRES, A. GOY, L. THEVENOT. « L'identité sociale dans le travail statistique : la nouvelle nomenclature des professions et catégories socioprofessionnelles », Économie et statistique n° 152, février 1983
- [4] A. DESROSIÈRES. « Éléments pour l'histoire des nomenclatures socioprofessionnelles », in « Pour une histoire de la statistique », I.N.S.E.E., 1977.
- [5] G. CALOT et J.C. DEVILLE. « Nuptialité et fécondité selon le milieu socioculturel », Économie et statistique n° 27, octobre 1971
- [6] C. THÉLOT. « Les traits majeurs du chômage depuis vingt ans », Économie et statistique n° 183, décembre 1985
- [7] P. LAULHÉ. « A propos de la variation du chômage dans les enquêtes emploi », Note 2671/432 du 10/11/1981, I.N.S.E.E., division Emploi
- [8] B.A. BAILAR. « The Effects of Rotation Group Bias on Estimates from Panel Surveys », JASA, volume 70, numéro 349, mars 1975
- [9] C. THÉLOT. « L'évolution de la mobilité sociale dans chaque génération », Économie et statistique n° 161, décembre 1983
- [10] « Enquête sur l'emploi de 1985 — Résultats détaillés », Les Collections de l'INSEE, n° D107, octobre 1985
- [11] P. LAULHÉ. « La mobilité professionnelle dans l'enquête Emploi : les effets de la méthode sur la mesure », Archives et Documents n° 38, décembre 1981
- [12] M. CEZARD et D. RAULT. « La crise a freiné la mobilité sectorielle », Économie et Statistique n° 184, janvier 1986.
- [13] C. GOURIEROUX. « Économétrie des variables qualitatives », Economica, 1984.
- [14] F. CHICOINEAU, J.F. PAYEN, C. THÉLOT. « Modélisation et redressement des non-réponses : le cas du salaire », Communication au congrès de l'Institut international de statistique, Amsterdam, août 1985.
- [15] J.C. COMBESSIE. « L'évolution comparée des inégalités : problèmes statistiques », Revue française de sociologie, XXVV (2), avril-juin 1984, et tous les articles qui, ensuite, ont alimenté dans cette revue le débat.
- [16] F. de SINGLY et C. THÉLOT. « Racines et profils des ouvriers et des cadres supérieurs », Revue française de sociologie XXVII (1), janvier-mars 1986.
- [17] M. VOLLE. « Analyse des données », Economica, 1977.
- [18] A. MONTFORT. « Cours de statistique mathématique », Economica, 1982.
- [19] D. VERGER. « L'achat d'un logement ne va pas sans achats d'équipement ». Économie et statistique n° 161, décembre 1983.
- [20] D. VERGER. « Les comportements des Français en matière d'assurances », Économie et statistique n° 174, février 1985
- [21] F. AMAND, F. MAUREL. et alii. « Rapport non publié d'un groupe de travail de l'ENSAE », animé par MM. LOLLIVIER et PAYEN, juin 1985.
- [22] M. HALBWACHS. « La statistique en sociologie » 1935, repris dans « Classes sociales et morphologie » Éditions de Minuit, 1972.
- [23] Y. LEMEL et A. VILLENEUVE. « Les consommations médicales des Français », Les Collections de l'INSEE n° M57, mars 1977.