Histoire & Mesure

Méthodes de régression et analyse factorielle

Félicité des Nétumières

Citer ce document / Cite this document :

Nétumières Félicité des. Méthodes de régression et analyse factorielle. In: Histoire & Mesure, 1997 volume 12 - n°3-4. Penser et mesurer la structure. pp. 271-297;

doi : https://doi.org/10.3406/hism.1997.1547

https://www.persee.fr/doc/hism_0982-1783_1997_num_12_3_1547

Fichier pdf généré le 28/03/2019

Résumé

Résumé : Dans le concert des travaux menés en sociologie, il en est un certain nombre qui s'interrogent sur les méthodologies employées par les chercheurs en sciences sociales. Le présent article a précisément pour dessein d'analyser deux techniques utilisées dans les approches quantitatives en sociologie, à savoir l'analyse factorielle et la régression multiple. Partant du constat qu'il est bien rare que ces deux méthodes soient utilisées conjointement, il cherche à rompre avec le débat qui oppose les tenants de l'une et de l'autre. Après un exposé des principes généraux qui sous-tendent les deux outils, il tente d'en souligner les différences en montrant qu'ils ont pour vocation de répondre à des questions différentes, elles aussi. L'ensemble de l'argumentation est enfin illustré par un exemple, où les mêmes données sont traitées successivement à l'aide de l'une puis de l'autre de ces techniques.

Abstract

Abstract. : Methods in Multiple Regression and Factorial Analysis. One of the fields in sociological study considers the ways methodologies are used by researchers in the social sciences. This article is situated within that orientation and explores two techniques, factorial analysis and multiple regression, employed in the qualitative approach to sociology. Noting that the two techniques are rarely if ever used together, the present study seeks to resolve the argument that pits the proponents of the one against those of the other. Having summarized the theoretical underpinnings of the two techniques, the study then emphasizes the differences in order to demonstrate that they are meant to solve different problems. The argument is then illustrated by the use of the same data treated first by one then by the other of the two techniques.

BY: creative commons

Histoire & Mesure, 1997, XII-3/4, 271-298

Félicité des Nétumières*

Méthodes de régression et analyse factorielle

Résumé : Dans le concert des travaux menés en sociologie, il en est un certain nombre qui s'interrogent sur les méthodologies employées par les chercheurs en sciences sociales. Le présent article a précisément pour dessein d'analyser deux techniques utilisées dans les approches quantitatives en sociologie, à savoir l'analyse factorielle et la régression multiple. Partant du constat qu'il est bien rare que ces deux méthodes soient utilisées conjointement, il cherche à rompre avec le débat qui oppose les tenants de l'une et de l'autre. Après un exposé des principes généraux qui sous-tendent les deux outils, il tente d'en souligner les différences en montrant qu'ils ont pour vocation de répondre à des questions différentes, elles aussi. L'ensemble de l'argumentation est enfin illustré par un exemple, où les mêmes données sont traitées successivement à l'aide de l'une puis de l'autre de ces techniques.

Abstract. : Methods in Multiple Regression and Factorial Analysis. One of the fields in sociological study considers the ways methodologies are used by researchers in the social sciences. This article is situated within that orientation and explores two techniques, factorial analysis and multiple regression, employed in the qualitative approach to sociology. Noting that the two techniques are rarely if ever used together, the present study seeks to resolve the argument that pits the proponents of the one against those of the other. Having summarized the theoretical underpinnings of the two techniques, the study then emphasizes the differences in order to demonstrate that they are meant to solve different problems. The argument is then illustrated by the use of the same data treated first by one then by the other of the two techniques.

* CREST-INSEE, Laboratoire de Sociologie Quantitative, 3 avenue Pierre Larousse, Timbre J350, 92245 - Malakoff cedex.

Persée creative commons

L'analyse factorielle des correspondances a pendant longtemps été l'outil le plus couramment utilisé par les sociologues français pour traiter les enquêtes de grande taille. Depuis quelques années en France, se développe l'usage de méthodes fondées sur la régression linéaire, très employées par les économistes, les démographes, ainsi que par les sociologues anglo-saxons. L'introduction de cette technique dans le champ de la sociologie réactive, en partie, le débat déjà ancien qui avait opposé chez les économistes, les tenants de l'une et de l'autre de ces deux familles de méthodes statistiques, dans les années soixante-dix.

L'objet de cet article est de transposer les termes dans le champ sociologique, en cherchant à montrer à partir de quelques applications, que comme toute technique, ces méthodes s'inscrivent « dans des constructions argumentatives, scientifiques ou politiques » [1] qui sont différentes, et qu'elles permettent de répondre à des questions différentes, elles aussi. Il s'agit ainsi de se placer volontairement du côté des usages qui en sont faits, de chercher à montrer que loin d'être concurrentes, elles présentent des caractères de complémentarité et qu'elles ont ainsi vocation à être utilisées en parallèle.

Pour ce faire, il est apparu important, dans un premier temps, de rappeler les principes généraux de la régression multiple et de l'analyse de correspondances multiples. Nous avons ensuite cherché à montrer la proximité de ces principes avec les critères usuellement employés en statistique et/ou en sociologie, pour la formulation d'énoncés de type descriptif et explicatif. Enfin, la dernière partie tente d'illustrer ce propos en présentant un exemple, dans lequel, à partir des mêmes données, les deux techniques sont utilisées tour à tour. Ici, il s'agit d'envisager la question de la précarité sociale, sous l'angle d'un cumul de handicaps, d'une part, et de l'enchaînement dynamique de difficultés, d'autre part.

1. La régression multiple : principes de base

Le raisonnement expérimental

On ne peut comprendre le sens et l'intérêt des techniques de régression sans effectuer un détour par le raisonnement expérimental,

---

1. DESROSIÈRES, A., 1995.

tel qu'il est pratiqué dans les sciences de la nature, en médecine et dans certaines sciences humaines comme la psychologie.

Le raisonnement est le suivant. Lorsqu'un médecin veut tester l'effet d'un médicament sur l'évolution d'une maladie, il réalise une expérimentation. Pour cela, il constitue deux groupes de patients atteints de la maladie en question, puis il prescrit le médicament aux membres d'un des groupes, et un placebo à ceux de l'autre groupe. C'est en comparant, au bout d'un certain délai, le pourcentage de guérison dans chacun des groupes qu'il conclura à l'efficacité (ou à l'absence d'efficacité) du traitement.

Pour que cette méthode soit valide, il doit prendre un certain nombre de précautions. En effet, une multitude d'autres facteurs peuvent intervenir dans le processus de guérison et fausser les résultats de l'expérience. Afin d'en annihiler les effets, le médecin doit s'assurer que ces facteurs sont distribués *de la même façon* parmi les individus des deux groupes, de manière à ce que le seul élément qui les différencie soit la prise ou non du médicament. Ainsi, le médecin sera certain de bien mesurer l'« effet propre » de son traitement.

Pour s'assurer que les deux groupes sont bien équivalents, il suffit – à condition qu'ils soient de taille suffisante – :

— d'affecter *au hasard* les individus entre les deux groupes,

— qu'ils ne sachent pas dans quels groupes ils se trouvent,

— que le médecin lui-même, au moment de l'évaluation, ne puisse pas différencier les patients traités de ceux qui ne l'ont pas été.

« Toutes choses égales par ailleurs »

En sociologie, il est bien évident qu'il est, la plupart du temps, impossible de procéder de la sorte. Si l'on cherche à déterminer l'effet propre du sexe sur la détermination des salaires (si on se pose la question de savoir s'il y a ou non discrimination salariale selon le sexe), il faudrait affecter un sexe au hasard aux individus, ce qui n'a évidemment pas de sens. Il faut donc trouver un moyen de s'assurer que les groupes dont on va comparer la moyenne des salaires sont bien équivalents du point de vue de toutes les autres variables qui peuvent avoir un effet sur le salaire (diplôme, ancienneté, catégorie socio-professionelle, type d'entreprise, etc.). En d'autres termes, il faut chercher à éliminer tout effet de structure qui viendrait fausser les résultats de l'étude.

Une première méthode consiste à composer autant de groupes qu'il y a de croisements possibles entre les diverses modalités des variables dont on cherche à annihiler l'effet (tris croisés de profondeur égale au nombre de variables). On voit immédiatement la limite que rencontre cette tentative : le nombre de groupes à constituer croît exponentiellement et, aussi grand que soit l'échantillon dont on dispose au départ, on risque de se trouver en face d'effectifs minuscules, voire nuls, à l'intérieur de chaque sous-population ; et, bien entendu, il n'est pas possible de prendre en compte l'effet de variables quantitatives continues (à moins d'effectuer des regroupements en classes ce qui entraîne une perte d'information).

Les techniques de régression multiple permettent de s'affranchir de ces difficultés. On n'en détaillera pas ici les fondements, qui nécessitent des connaissances statistiques poussées. Il suffit de savoir que l'idée générale consiste à rechercher l'effet propre d'une variable sur une autre, comme dans le raisonnement expérimental, et que cela est possible grâce à des hypothèses probabilistes.

Effectuer une régression multiple consiste à construire un modèle, en faisant un certain nombre d'hypothèses, parmi lesquelles certaines seront assumées, et d'autres testées. Reprenons l'exemple de la discrimination salariale. On se pose la question de savoir si le sexe, en tant que tel, a un effet propre sur le salaire, c'est-à-dire une fois que tous les autres effets sont contrôlés. À emploi égal, y a-t-il salaire égal entre les hommes et les femmes ? On imagine que parmi les déterminants du salaire, interviennent des éléments tels que le niveau de diplôme, la catégorie socio-professionelle, l'ancienneté, la fonction, le secteur (public ou privé) de l'entreprise, la taille de l'entreprise et bien d'autres choses encore, que l'on peut ou non identifier et mesurer. On écrit donc le modèle suivant, qui indique que le salaire est fixé en fonction de ces caractéristiques, et du sexe :

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + aS + u$$

avec :

Y : le salaire

$X_1, \ldots, X_n$: les variables explicatives du salaire

S : le sexe

u : un résidu, qui capture toute l'hétérogénéité qu'on n'a pas pu prendre en compte avec les variables explicatives du modèle.

Le logiciel de régression est alors capable de calculer la valeur des coefficients (a, a_1, ..., a_n). Un coefficient nul signifiera que la variable auquel il est associé n'a pas d'effet propre. S'il est différent de 0 au contraire, cela voudra dire que « toutes choses égales par ailleurs », c'est-à-dire une fois contrôlé l'effet de toutes les autres variables explicatives, la variable a un effet propre sur le salaire dont on cherche à comprendre les déterminants [2].

On voit bien que le but visé, ici, est de se rapprocher le plus possible des conditions expérimentales, sachant bien qu'on ne pourra jamais les atteindre. Il est, en effet, impossible de prendre en compte toutes les variables qui ont un effet, soit qu'elles ne soient pas mesurables, soit qu'on n'en imagine pas l'existence. L'erreur résiduelle u est là pour le rappeler. Bien évidemment, si elle est trop importante, les résultats seront ininterprétables et on dira que le modèle est mal spécifié. Il existe des moyens de tester la spécification du modèle, mais quels que soient le verdict du test, on ne pourra jamais être totalement sûr de la véracité de ce que l'on affirme. Ceci est un problème dont il faut être conscient, mais qui n'invalide pas pour autant la démarche dans son ensemble. Il faut, en effet, garder à l'esprit qu'aucune science n'est jamais à même de garantir la vérité de ces propositions. Rappelons-nous la position de Karl Popper : « La meilleure stratégie pour un scientifique est de formuler chaque proposition de telle sorte qu'elle survive aux tests les plus sévères qu'il pourra inventer et si son hypothèse se révèle fausse, il doit en énoncer une nouvelle qui survive à tous les tests précédents, recommençant ainsi le cycle hypothèse et réfutation » [3]. Ainsi, le sociologue introduit dans son modèle toutes les variables envisageables et qui sont disponibles dans son enquête, tout en sachant qu'il y aura sans doute un jour une nouvelle enquête, avec d'autres variables, et que ses résultats ne sont pas à l'abri d'une réfutation.

---

2. Selon que l'on a affaire à des variables qualitatives ou quantitatives, les mises en œuvre et les interprétations sont différentes, mais l'idée générale de base est celle que nous venons d'exposer. Il ne paraît pas indispensable ici de rentrer davantage dans le détail.

3. POPPER, K., 1959.

2. Application de l'analyse factorielle au dépouillement d'enquête : l'analyse de correspondances multiples (ACM)

L'analyse factorielle n'a pas du tout les mêmes visées. Son objectif premier est de permettre au chercheur d'appréhender le plus simplement possible la masse de données dont il dispose dans son enquête et d'en extraire les informations pertinentes. Pour reprendre une expression consacrée, l'analyse factorielle est « un aveu d'ignorance », « un radar tourné vers le brouillard » [4] qui « sert avant tout à dépeindre à grands traits les dimensions les plus importantes d'une variation dans un nouveau champ de recherche » [5].

C'est donc, en premier lieu, lors de la phase exploratoire des données, que l'analyse factorielle se présente comme un outil particulièrement utile. Au commencement de toute étude, le chercheur consacre toujours un temps qui peut se révéler très long à « sentir les données », c'est-à-dire tout d'abord à découvrir la population de son fichier, à la décrire à l'aide de ses principales caractéristiques, puis à sélectionner les variables dont il peut supposer qu'elles auront quelque chose à voir avec le sujet de son étude. Classiquement, cette étape s'effectue en construisant des tableaux statistiques (tris à plat et tris croisés), permettant de mettre en évidence la variabilité de l'échantillon, ainsi que les premiers liens entre les variables de l'enquête. Bien évidemment, plus il y a de variables, plus ce travail peut s'avérer pénible. En l'absence d'autres outils, le chercheur se borne alors à n'effectuer que quelques croisements, ceux qui lui semblent les plus pertinents, en fonction de suppositions qu'il aura pu formuler par ailleurs. L'ACM, en revanche, permet en quelque sorte d'automatiser cette étape, sans qu'il soit nécessaire d'émettre la moindre hypothèse préalable concernant les associations éventuelles entre les différentes variables.

D'un point de vue technique, les données brutes apparaissent sous la forme d'un nuage de points dans un espace qui a autant de dimensions qu'il y a de variables introduites dans l'analyse. L'ACM cherche alors à construire un nouvel espace sur lequel sont projetés les points du nuage initial. Ce nouvel espace est conçu de manière à concentrer le maximum de l'information contenue dans les données initiales, à partir d'un minimum de dimensions.

---

4. Cattell, R.-B., 1952.
5. Hirschi, T. & Selvin, H.-C., 1975.

Prenons comme exemple une cuisinière ayant à sa disposition toutes sortes d'ingrédients qu'elle ne connaît pas et à partir desquels elle cherche à confectionner un bon repas. Le chercheur est comme cette cuisinière, incapable de savoir quels sont les aliments qui vont s'accorder entre eux, ni d'imaginer la saveur finale du plat qu'elle proposera à ses convives. Bien entendu, elle a la possibilité de chercher à connaître le goût des aliments en les testant un par un, puis deux à deux, et si elle est un tant soit peu méthodique, on peut imaginer qu'elle finira par bien les connaître tous. Mais cette quête risque de lui prendre du temps (et de lui donner une bonne indigestion).

Il va de soi qu'elle préférerait pouvoir disposer de critères de classement, même grossiers, qui lui permettraient de se faire une idée des grandes catégories d'aliments. L'analyse factorielle, réalisée sur les caractéristiques des denrées, peut l'y aider : on imagine que celle-ci lui fournirait un espace de saveurs pour les deux premiers axes, opposant le sucré au salé, d'une part, et l'acide à l'amer, d'autre part. Un autre axe opposerait le liquide au solide, un autre encore le gras au maigre et pourquoi pas, le cher au bon marché, etc. En projetant ensuite le nom de chacun des ingrédients en variables supplémentaires sur l'espace créé, elle verrait apparaître des familles d'aliments, qu'elle aurait tout loisir ensuite de combiner et de cuire à sa guise. Elle saurait, par exemple, qu'une pomme est plutôt sucrée, acide, solide, dépourvue de matières grasses, etc. Sur l'axe des prix, elle verrait sans doute s'opposer le caviar et la soupe en sachet...

Les difficultés que rencontre le sociologue sont comparables à celles de notre cuisinière. Grâce à l'ACM, il sera en mesure de visualiser la diversité de son échantillon, à partir de classements automatiques, qui vont lui permettre de faire émerger une structure cohérente. En projetant *orthogonalement* le nuage de points initial sur un espace à deux dimensions (celles de sa feuille de papier), il obtiendra la meilleure approximation possible de ses données. Cependant, il est bien évident qu'il va perdre une partie de l'information initiale en projetant un espace multidimensionnel sur une simple surface, même si cette surface a été construite pour en conserver le maximum. Et, sans doute, aura-t-il besoin d'examiner également un ou deux axes supplémentaires, qui apparaîtront sur un second graphique, pour améliorer sa connaissance des données.

Avant de s'intéresser à la lecture des graphiques, quelques remarques s'imposent. Tout d'abord, il faut toujours être conscient que l'ACM est une technique fondée sur le tri à plat et le tri croisé

simple (c'est-à-dire de profondeur 2). L'ACM repère les axes principaux à partir de l'examen des écarts à l'indépendance dans le grand tableau rassemblant tous les tris croisés entre toutes les variables de l'analyse (tableau de Burt). D'une certaine manière, l'ACM nous renseigne automatiquement sur les croisements de variables qui sont les plus éloignés de la situation d'indépendance, c'est-à-dire sur les liaisons statistiques les plus fortes qui existent dans le fichier de départ. Nous ne sommes pas ici dans le domaine de l'analyse multivariée, même si l'on manipule un grand nombre de variables.

Seconde remarque, les variables dites « actives », introduites dans l'analyse, qui vont servir à la construction des axes, sont toutes traitées par l'ACM sur le même plan. On cherche, en effet, au cours de cette opération, à repérer des associations mais pas de liaison d'ordre causal, ce qui implique qu'il n'y a pas de variable dépendante ou indépendante. Pour cette raison, il ne faut introduire que des variables situées sur un même plan conceptuel (variables d'état *ou* pratiques, mais pas les deux en même temps). Dans le premier cas, on construit un espace de caractéristiques et dans le second, un espace de pratiques, mais la construction d'un espace mixte poserait de sérieux problèmes d'interprétation.

Ce n'est qu'une fois l'espace construit qu'il est possible d'ajouter des variables situées à un autre niveau, par le biais de la projection des variables dites « supplémentaires ». Ces variables supplémentaires servent à conforter l'interprétation des axes, à vérifier qu'ils ont bien un sens. On dit souvent qu'il faut réaliser l'analyse en prenant les pratiques comme variables actives et les caractéristiques comme variables supplémentaires (avec comme sous-entendu que ces dernières vont expliquer les premières). Ceci n'est pas forcément indispensable, dans la mesure où la notion de causalité, encore une fois, est étrangère à l'esprit même de l'ACM.

Que nous apprend la lecture des graphiques ? En premier lieu, on voit se dessiner un espace social structuré par les variables qui ont le plus contribué à sa construction. Par exemple, le premier axe opposera les jeunes aux moins jeunes et le second axe, les diplômés aux « sans diplôme ». Ou alors, si on a choisi de représenter un espace de pratiques, le premier axe opposera des pratiques éducatives visant à développer l'autonomie de l'enfant à des pratiques plus traditionnelles davantage axées sur la surveillance et le contrôle, etc. En second lieu, on verra « apparaître des proximités entre certains des individus et certaines de leurs pratiques, (...), selon une logique

statistique et probabiliste, et non pas déterministe » [6]. En d'autres termes, il sera possible d'affirmer, au vu de la faible distance angulaire qui sépare les points sur le graphique, que les mères, cherchant à exercer une surveillance forte sur leurs enfants, ont plutôt telle ou telle caractéristique. Ou encore, que ce sont les jeunes sans diplôme qui ont le plus tendance à regarder « Robocop 2 », quand ce film passe à la télévision. Mais rien ne permet de dire que c'est l'âge ou le niveau de diplôme qui détermine le choix des programmes télévisés, dans la mesure où on a seulement mis en évidence des cooccurrences et rien de plus.

3. Décrire ou expliquer

La régression multiple apparaît comme le moyen privilégié pour mettre en évidence des liaisons déterminantes entre variables. On dit d'ailleurs couramment à partir des résultats d'un modèle que telle variable « explique » telle autre et on parle de variables « explicatives » et de variables « expliquées ». En effet, la démarche mise en œuvre est la seule qui satisfasse aux critères habituellement avancés en statistique de l'existence d'une relation causale, à savoir que :

1) il existe une liaison statistique entre la variable A et la variable B,

2) A précède B dans l'ordre temporel,

3) la corrélation entre A et B subsiste lorsque l'on contrôle l'effet d'autres variables qui peuvent avoir un lien avec A ou B.

Ceci est, en fait, très proche de ce que Durkheim cherchait à conceptualiser dans *Les Règles de la Méthode Sociologique*. Après avoir affirmé au chapitre V que « la cause déterminante d'un fait social doit être cherchée parmi les faits sociaux antécédents » (critère 2), il annonce au début du chapitre VI : « Nous n'avons qu'un moyen de démontrer qu'un phénomène est cause d'un autre, c'est de comparer les cas où ils sont simultanément présents ou absents et de chercher si les variations qu'ils présentent dans ces différentes combinaisons de circonstances témoignent que l'un dépend de l'autre. » (critères 3 et 1) [7].

---

6. DESROSIÈRES, A., 1995.
7. DURKHEIM, É., 1895, p. 124.

Citons également Lazarsfeld qui, avec l'« Analyse multivariée » [8], pose les bases du raisonnement que permet de généraliser la régression multiple. Après avoir constaté que les jeunes auditeurs s'intéressaient moins aux programmes religieux à la radio que les plus âgés, il se demande si on peut y voir « le signe d'un moindre attachement des jeunes aux valeurs religieuses ». Mais, avant de conclure à un effet de l'âge, il introduit, comme variable supplémentaire dans l'analyse, le niveau d'instruction. Celui-ci est fortement lié à l'intérêt pour les programmes religieux et, en effectuant un contrôle avec cette nouvelle variable, la relation précédente disparaît. C'est donc parce que les jeunes ont un niveau d'éducation plus élevé que leurs aînés, qu'ils écoutent moins d'émissions à caractère religieux et non pas parce qu'ils sont jeunes. La tentative d'homogénéiser les groupes que l'on compare, par la prise en compte d'une variable-test, est tout à fait dans l'esprit de la démarche « toutes choses égales par ailleurs ».

Il faut cependant nuancer un peu le propos et comprendre de quel type de cause il est ici question. Dire que le sexe explique le salaire, parce qu'on a réussi à mettre en évidence l'existence d'un effet propre du sexe sur le niveau de salaire, « toutes choses égales par ailleurs », ne signifie pas qu'il existe une détermination d'ordre fonctionnel entre le sexe et le salaire (au sens où sont liés la vitesse, la distance parcourue et le temps qu'il faut pour la parcourir). Nous sommes ici dans un schéma probabiliste. Cela veut dire que le modèle ne permettra jamais de connaître, précisément et de manière absolue, la modalité prise par la variable dépendante pour un individu, même si on connaît toutes ses caractéristiques. En revanche, il sera à même de nous donner la loi de probabilité qui régira, pour cet individu, la variable que l'on a expliquée. Plus simplement, on ne pourra pas dire que tel individu qui a tel sexe, tel âge, tel niveau de diplôme, telle origine sociale, telle profession, etc., va passer tant d'heures par semaine devant la télévision. En effet, le modèle nous révélera que cet individu a telle probabilité de ne jamais la regarder, telle probabilité d'y consacrer une heure, deux heures, et ainsi de suite.

La régression a donc pour but d'expliquer la réalité. À l'inverse, l'analyse factorielle, parce qu'elle ne traite que de liaisons entre variables prises 2 à 2, ne peut avoir d'ambition explicative. Elle est particulièrement utile lors de la phase exploratoire et constitue l'outil privilégié de la description. Ce terme n'est en rien péjoratif, bien au

---

8. LAZARSFELD, P., 1966.

contraire, et il ne saurait être question ici d'établir une quelconque hiérarchie entre les deux familles statistiques. En effet, lorsqu'avec l'aide d'une régression on cherche à annihiler l'effet de variables pour mettre en évidence un effet pur, on crée un univers factice, tout à fait éloigné des configurations réelles du monde social. Or, « ce parasitage [9] est aussi un autre nom de la configuration historique, qui constitue (...) la seule réalité empirique », nous rappelle J.-C. Passeron. En cherchant à raisonner « toutes choses égales par ailleurs », on déconstruit la réalité, et ce faisant, on s'interdit toute considération sur la composition des groupes sociaux. Plus encore, ce type de raisonnement peut conduire, si l'on n'y prend pas garde, à des « non-sens historiques », à des interrogations absurdes, analogues à celle du paradoxe de Simiand, c'est-à-dire cherchant « 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. » Bref, on risque sans cesse d'oublier que « les cooccurrences ne sont données dans la réalité qu'*ainsi et pas autrement* » [10]. Le travail du sociologue consiste justement à les repérer et l'analyse factorielle est là pour l'y aider.

Pour illustrer tout ceci, nous allons reprendre comme premier exemple la question des salaires féminins. Une ACM nous révélera que les femmes occupent des emplois moins qualifiés que les hommes, qu'elles sont plus souvent employées et moins souvent cadres, ..., *et* qu'elles ont en moyenne des salaires inférieurs à ceux des hommes. Une régression multiple tentera de savoir si, « toutes choses égales par ailleurs », elles ont des rémunérations inférieures. On voit bien que l'on répond ici à deux questions différentes qui, à notre sens, méritent toutes deux d'être posées. Mais cet avis ne fait pas l'unanimité et le débat reste ouvert. Pour certains, il convient de remarquer que les femmes n'occupent pas les mêmes postes que les hommes, qu'elles n'obtiennent pas les mêmes avancements, et donc qu'elles ne sont pas en mesure d'être comparées, « toutes choses égales par ailleurs », aux hommes, employant même l'expression, « toutes choses *inégales* par ailleurs », pour bien montrer leurs doutes quant au repérage d'effet pur [11].

---

9. L'effet des autres variables.
10. PASSERON, J.-C., 1991.
11. Cahiers du MAGE, 1995.

Second exemple tiré d'une étude récente sur la réussite scolaire des enfants étrangers en France [12]. Là encore, une analyse factorielle va mettre en évidence que les enfants étrangers ou issus de l'immigration, appartiennent plutôt à des milieux sociaux défavorisés et qu'ils ont de moins bons résultats que leurs congénères français. Mais, dès que l'on cherche à contrôler les autres caractéristiques pouvant avoir de l'influence sur la réussite scolaire (catégorie socio-professionnelle du père, taille de la fratrie, statut de la mère vis-à-vis de l'activité professionnelle, ...), les résultats s'inversent, et force est alors de constater que les enfants étrangers connaissent une meilleure carrière au collège que leurs condisciples, « toutes choses égales par ailleurs ».

Que conclure de tout ceci ? J.-C. Passeron nous dit qu'« est sociologique tout raisonnement qui se tient sous la contrainte d'énoncer ses généralités en prenant appui sur des constats de base qui ne sont jamais comparables sous tous les rapports ». Mais il ajoute également que « l'exigence qui engendre (le raisonnement toutes choses égales par ailleurs) constitue un des deux pôles d'exigences entre lesquels se meut le raisonnement sociologique – l'autre étant le pôle historique ». Il nous semble, en effet, que les questions auxquelles les deux méthodes permettent de répondre sont plus complémentaires qu'antagonistes et que, une chose est de constater que les femmes occupent moins souvent que les hommes des postes élevés hiérarchiquement, qu'elles occupent moins souvent des postes de cadres etc., et une autre de se demander si, à position égale, leurs salaires sont moins élevés que ceux des hommes.

Alain Desrosières [13] apporte un éclairage tout à fait original et intéressant sur le sens à donner à ce débat. Il remarque que dans les commentaires énoncés autour des résultats issus de la mise en œuvre de l'une et l'autre méthodes, les sujets des verbes diffèrent. Autour d'une ACM, « les sujets des verbes sont des groupes sociaux, des classes d'individus, liés entre eux par une communauté probable de comportements, dans une perspective holiste de reconstitution de la globalité d'une personne, d'un groupe, ou d'une localité » (Les élèves étrangers ont de moins bons résultats...). En revanche, les méthodes fondées sur la régression multiple appellent une interprétation à partir de « formes grammaticales centrées sur le langage des variables » (Le fait d'être étranger conduit à de meilleurs résultats...). Ceci s'explique, selon lui, si l'on considère le but ultime des études que l'on peut

12. VALLET, L.-A. & CAILLE, J.-P., 1995.
13. DESROSIÈRES, A., 1995.

réaliser, avec l'aide de l'une et l'autre techniques. Les analyses fondées sur la modélisation, en repérant l'effet d'une variable sur une autre, donnent les moyens aux politiques d'agir sur tel ou tel phénomène par le biais de mesures spécifiques. Ainsi, si l'on constate qu'il y a discrimination salariale, l'État peut promulguer une loi visant à l'interdire (*Cf.* la loi de 1983 « À travail égal, salaire égal ») et veiller à son application. À l'inverse, une analyse réalisée à partir d'analyses factorielles, en décrivant la réalité sociale, pourra dénoncer les inégalités, mais sans être à même de proposer des solutions visant à les réduire.

4. Un exemple : la précarité professionnelle et le risque d'exclusion

Afin de rendre plus claires et plus tangibles les différences entre les deux méthodes présentées, nous allons examiner les résultats que l'on obtient lorsque l'on applique l'une et l'autre aux mêmes données. L'exemple que nous avons retenu ici est extrait d'un texte de Serge Paugam [14] dans lequel l'auteur cherche à comprendre le processus qui mène à l'exclusion.

Dans cet article, la « pauvreté » est considérée comme un phénomène multidimensionnel. Au critère économique bien évidemment pris en compte, s'ajoutent des facteurs de précarité « sociale », comme l'instabilité conjugale, la sociabilité réduite, l'absence de réseau d'aide privée, etc. Il s'agit, en effet, de tester l'hypothèse couramment avancée, du cumul des handicaps entraînant les individus vers la disqualification sociale. Mais il s'agit aussi d'analyser les liens entre ces différentes dimensions de la pauvreté, de manière à savoir si les difficultés sur le marché de l'emploi entretiennent un rapport de causalité avec des difficultés dans les autres domaines de la vie sociale. Nous nous trouvons donc, ici, face à deux questions fondamentalement différentes et dont le traitement nécessite de faire appel à l'analyse factorielle, pour la première, et aux méthodes de régression, pour la seconde.

Définition des variables et construction des indicateurs

Les données utilisées proviennent de l'enquête « Situations défavorisées » réalisée en 1986-1987 par l'INSEE. Elle rassemble

---

14. PAUGAM, S., ZOYEM, J.-P. & CHARBONNEL, J.-M., 1993.

18 700 individus (parmi lesquels 7 000 environ ont été utilisés ici) et surreprésente les individus vivant dans des logements sans confort. Nous ne fournirons donc pas les tris à plat qui n'ont que peu d'intérêt, puisqu'ils reflètent les choix d'échantillonnage et non la répartition des individus en fonction de leurs caractéristiques diverses. D'autre part, il convient de noter que les individus pris en compte ont tous un logement, ce qui élimine du champ de l'enquête les plus pauvres des plus pauvres, à savoir les « sans domicile fixe ».

Avant de commencer le compte rendu de l'analyse proprement dite, nous présentons rapidement les principales variables de l'enquête, ainsi que les indicateurs construits pour l'étude.

Tableau 1. *Variables de l'enquête*

Variables	*Modalités*	*Définitions*
Situation sur le marché de l'emploi	Emploi stable non menacé	Individus qui ont un emploi et qui considèrent qu'ils ne risquent pas de le perdre au cours des deux années à venir
	Emploi stable menacé	Individus qui occupent le même emploi depuis au moins un an, mais qui considèrent qu'ils risquent de le perdre au cours des deux années à venir
	Emploi instable	Individus qui occupent un emploi depuis moins d'un an et qui considèrent qu'ils risquent de le perdre au cours des deux années à venir
	Chômage de moins de deux ans	
	Chômage depuis deux ans ou plus	
Pauvreté économique	Très pauvre	Individus vivant dans un ménage dont le revenu par unité de consommation est inférieur à 1 800 F. par mois
	Pauvre	Individus vivant dans un ménage dont le revenu par unité de consommation est compris entre 1 800 et 2 700 F. par mois
	Non pauvre	
Trajectoire conjugale	Couple sans rupture	
	Couple recomposé Seul après rupture Ayant toujours vécu seul	

Indicateur de sociabilité familiale (hors membres du ménage)	Forte	Individus ayant rencontré plusieurs personnes de leur famille au cours des trois derniers mois
	Moyenne	Individus n'ayant rencontré qu'une ou deux personnes de leur famille au cours des trois derniers mois
	Faible	Individus n'ayant rencontré aucun membre de leur famille au cours des trois derniers moise
Indicateur de réseau d'aide privée potentielle*	Support potentiel fort	
	Moyen	
	Faible	
Indicateur de la participation à la vie associative	Adhérent à au moins une association	
	Non adhérent	
Indicateur de pauvreté relationnelle	Très pauvre	Sociabilité familiale, supports relationnels faibles et absence de participation à une association
	Pauvre	Sociabilité familiale, supports relationnels moyens, quelle que soit la participation à la vie associative **ou** sociabilité familiale et supports relationnels faibles mais adhésion à une association
	Non pauvre	Sociabilité familiale et supports relationnels forts, quelle que soit la participation à la vie associative
Problèmes de jeunesse	Pas de problème	
	Problème familial	
	Problème d'argent	
	Problème familial et d'argent	
Santé	Très bonne	
	Médiocre	
	Mauvaise	
Sexe	Homme	
	Femme	
Âge	18 à 24 ans	
	25 à 29 ans	
	30 à 34 ans	
	35 à 49 ans	
	50 à 64 ans	

Nationalité du père	Français Autre pays de la CEE Hors CEE Inconnue	
Nombre d'enfants	Aucun Un Deux Trois ou plus	
Commune de résidence	Commune rurale Moins de 100 000 habitants Plus de 100 000 habitants Banlieue parisienne Paris	 (hors banlieue parisienne) (hors banlieue parisienne)
Statut d'occupation du logement	Logé gratuitement Locataire HLM Locataire privé Propriétaire	 (ou accédant à la propriété)
Catégorie socio-professionelle	Agriculteurs Art., com., chefs d'entreprise Cadres supérieurs Prof. intermédiaires Employés Ouvriers qualifiés Ouvriers non qualifiés Non déclaré	
Diplôme	Aucun diplôme CEP BEPC CAP-BEP BAC Etudes supérieures	

* Nous ne détaillons pas la construction de cet indicateur élaboré à partir de questions portant sur l'existence de possibilités d'hébergement, d'aide financière, de services divers et de soutien moral.

« Les inégalités face au risque d'exclusion » [15]

La première préoccupation de l'auteur dans cet article est d'établir à quel point la précarité professionnelle s'accompagne de tout un ensemble de difficultés d'ordre économique, familial ou relationnel, pouvant conduire les individus qui en sont victimes, dans une spirale de disqualification sociale.

Les premiers résultats établis à l'aide de simples tris croisés sont d'ores et déjà très parlants. En premier lieu, l'auteur étudie le lien

15. Nous reprenons ici un des titres de paragraphe de Serge Paugam.

entre la situation sur le marché de l'emploi et la pauvreté économique. Comme on le voit dans le tableau ci-dessous, ces deux variables sont fortement liées, mais elles ne sont pas complètement corrélées : le chômage de longue durée n'est pas forcément synonyme de misère (tout dépend des revenus des autres membres du ménage) et « la pauvreté économique accompagne aussi les situations d'emploi stable menacé ou encore davantage d'emploi instable ».

Tableau 2. *Situation sur le marché de l'emploi et pauvreté économique*

Pauvreté économique	*Pauvre et très pauvre en %*	*Non pauvre en %*
Emploi stable non menacé	5,6	94,4
Emploi stable menacé	12,4	87,6
Emploi instable	19,4	80,6
Chômage < 2 ans	26,6	73,4
Chômage > 2 ans	40,2	59,8
Ensemble	11,8	88,2

En second lieu, il apparaît nettement que plus les individus se trouvent dans une situation professionnelle difficile, plus leur *sociabilité familiale* est faible.

Tableau 3. *Situation sur le marché de l'emploi et sociabilité familiale*

Sociabilité familiale	*Faible %*	*Moyenne %*	*Forte %*
Emploi stable non menacé	7,0	48,7	44,3
Emploi stable menacé	8,6	53,7	37,7
Emploi instable	12,5	52,8	34,7
Chômage < 2 ans	11,6	58,6	29,8
Chômage > 2 ans	17,7	55,7	26,6
Ensemble	8,8	51,5	39,7

De même, les possibilités d'avoir recours à l'entourage (aide financière, de logement ou support affectif) s'amenuisent avec la dégradation de la situation professionnelle.

Tableau 4. *Situation sur le marché de l'emploi et supports relationnels*

Supports relationnels	*Faibles* %	*Moyens* %	*Forts* %
Emploi stable non menacé	4,9	27,7	67,4
Emploi stable menacé	8,2	36,5	55,3
Emploi instable	10,0	34,3	55,7
Chômage < 2 ans	10,1	32,8	57,1
Chômage > 2 ans	16,2	37,2	46,6
Ensemble	7,2	31,6	61,2

Enfin, et nous nous arrêterons là pour les tris croisés, tout en sachant bien qu'il serait possible d'en examiner d'autres, le tableau suivant confirme les liens entre précarité professionnelle et retrait de la vie associative :

Tableau 5. *Situation sur le marché de l'emploi et vie associative*

Participation à la vie associative	*Absence de participation* %	*Participation* %
Emploi stable non menacé	53,9	46,1
Emploi stable menacé	63,5	36,5
Emploi instable	69,7	30,3
Chômage < 2 ans	78,8	21,2
Chômage > 2 ans	76,4	23,6
Ensemble	60,7	39,3

Ces quelques rapides constats invitent à considérer tous ces aspects en même temps, de manière à repérer différents « types de population selon l'intensité du cumul de leurs handicaps », et à chercher à les caractériser à l'aide de variables socio-démographiques. Pour cela, l'outil privilégié est bien évidemment l'Analyse de Correspondances Multiples.

Les variables actives utilisées pour l'analyse sont toutes liées à la « pauvreté » (ou à « l'aisance ») au sens large, c'est-à-dire prenant en compte les dimensions professionnelle (situation sur le marché de l'emploi), économique (pauvreté en termes de revenu, statut d'occupation du logement), familiale (trajectoire conjugale, sociabilité

familiale) et relationnelle (réseau d'aide privée et participation à la vie associative). D'autre part, ont été ajoutés l'état de santé et un indicateur de problèmes de jeunesse. C'est donc l'espace de la précarité qui se construit ici grâce à l'ACM et qui est structuré par l'ensemble de ces variables.

Pour mieux comprendre quelles sont les populations qui sont représentées ici et pour conforter la cohérence des axes, l'auteur a projeté sur cet espace un certain nombre de variables supplémentaires (ou illustratives), à savoir, le niveau d'études, la catégorie socio-professionnelle, l'âge, et le nombre d'enfants.

Examinons les résultats de cette ACM à partir du graphique représentant les deux premiers axes factoriels :

L'axe 1 oppose les individus « riches », aux individus se trouvant dans une situation de précarité. En effet, à la droite du graphique se trouvent surreprésentées :

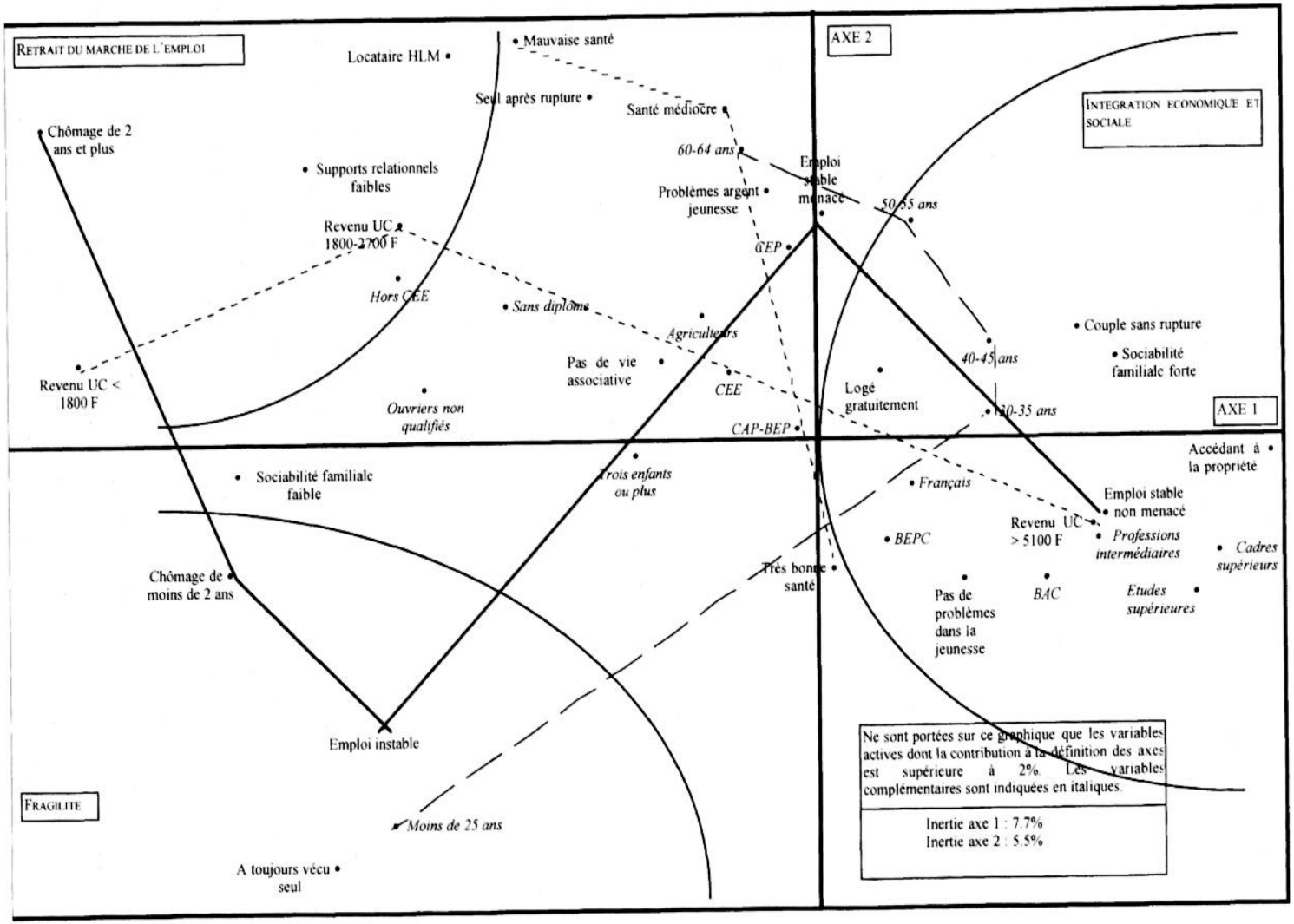
— des personnes occupant des emplois de cadres supérieurs ou exerçant des professions intermédiaires,
— des personnes ayant au moins le BAC,
— des personnes occupant un emploi stable non menacé,
— des personnes vivant en couple sans avoir connu de rupture,
— des personnes ayant une sociabilité familiale forte,
— des personnes vivant dans une certaine aisance économique,
— des personnes propriétaires de leur logement,
— des personnes participant à la vie associative, etc.

À gauche, en revanche, sont rassemblés les individus qui ont des difficultés. « Cet axe peut être défini comme l'axe de la participation à la vie économique et sociale ».

L'axe 2 se définit par une segmentation liée à l'âge avec en bas du graphique :

— des jeunes,
— des individus en bonne santé,
— des individus en emploi instable,
— des individus qui n'ont pas connu de vie de couple, et dans la partie haute :
— des personnes plus âgées,
— des chômeurs de longue durée,
— des personnes connaissant des problèmes de santé,
— des individus vivant seuls après une rupture, etc.

Figure 1. *Analyse factorielle des correspondances sur l'ensemble des actifs de 18 à 64 ans*



Attention, rappelons qu'avec l'ACM, nous sommes dans un contexte de tris croisés simples. Il serait donc abusif de parler pour le cadran sud-ouest du graphique de jeunes en phase d'insertion, vivant sans doute chez leurs parents, en attendant de stabiliser leur situation professionnelle et d'envisager une vie de couple actuellement impossible. C'est peut-être vrai. Mais l'ACM ne nous dit pas que les individus de cette zone présentent *à la fois* toutes ces caractéristiques ; elle montre que les détenteurs de chacune de ces modalités se trouvent probablement autour du point qui la représente. De la même façon, il serait tentant d'affirmer que la droite du graphique regroupe les catégories les plus favorisées ainsi que les classes moyennes, possédant tout à la fois des capitaux économique, culturel et social importants ; et de terminer le commentaire en assurant que ce sont les membres du cadran nord-ouest qui sont les plus susceptibles de basculer du côté de l'exclusion la plus radicale, puisqu'ils connaissent une situation de pauvreté extrême, à la fois économique et relationnelle, sans soutien familial, et puisque leur absence de diplôme et leur âge relativement avancé, leur laissent peu d'espoir de retrouver un jour un emploi. Mais seule une procédure de classification automatique peut permettre d'avancer des conclusions de ce genre (ou un recours aux tris de profondeur 5 ou 6). Encore une fois, il est très possible qu'elles soient vraies, mais il faut le vérifier. Notons à cet effet que bien des logiciels permettent de projeter les individus sur le plan factoriel, ce qui permet de voir à quel point ceux-ci forment ou non un groupe homogène autour des caractéristiques que l'on envisage.

La situation professionnelle face au risque d'exclusion

L'ACM a permis d'identifier des faisceaux de corrélations dont l'existence pourrait laisser entendre que les difficultés sur le marché de l'emploi provoquent la mise en route d'un processus de disqualification sociale. Toutefois, les tris croisés effectués en début d'analyse nous ont mis en garde contre des conclusions de cette sorte, un peu hâtives. Le type particulier de régression multiple que constitue le modèle LOGIT va nous aider à séparer ce qui relève véritablement des difficultés professionnelles de ce qui provient d'autres caractéristiques.

Il faut noter que l'analyse que l'on met ici en œuvre est statique. En effet, les données utilisées apportent des informations essentiellement sur la situation des individus au moment de l'enquête. On sait qu'aujourd'hui, tel individu chômeur a des liens distendus avec sa famille, mais on ne sait pas depuis quand ces liens sont distendus. En particulier, même si des indications sont données quant à l'ancienneté de sa situation de chômeur, il n'y a aucun moyen de savoir si son

éloignement familial est intervenu avant ou après la perte de son emploi. De même, on peut savoir qu'il a, dans le passé, connu une rupture conjugale, mais on ne sait pas si celle-ci a eu lieu avant ou après qu'il devienne chômeur. En conséquence, il est impossible ici, en l'absence de vraies données longitudinales, permettant de saisir l'ordre temporel entre les événements, de mettre en évidence des liens de causalité entre la situation professionnelle et l'absence ou l'intensité de tel ou tel lien social.

En revanche, il est connu, d'après les résultats de nombreuses études antérieures, que la sociabilité au sens large est différente, aussi bien en intensité que par les formes privilégiées qu'elle emprunte (sociabilité familiale, amicale ou par le biais d'adhésion à des associations), selon les classes sociales, les lieux de résidence, l'âge, la taille de la famille, etc. Dans la mesure où la situation professionnelle est également liée à ces mêmes variables, il est tout à fait possible que les corrélations mises en évidence dans les tris croisés et l'ACM ne soient que le reflet d'un effet de structure. C'est ce que les modélisations effectuées vont chercher à établir. Nous limiterons ici notre présentation de l'étude de Serge Paugam, à l'analyse de la sociabilité familiale. La question à laquelle il est possible de répondre est la suivante : une fois que l'on a pris en compte l'effet du sexe, de l'âge, de la nationalité du père, de la commune d'habitation, de la catégorie socio-professionnelle, du diplôme, du revenu, du nombre d'enfants, de la trajectoire conjugale, et de l'existence de problèmes de jeunesse, subsiste-t-il des écarts dans l'intensité de la sociabilité familiale, selon la situation professionnelle ?

Si la réponse à cette question s'avère négative, alors cela signifiera que les écarts observés en croisant la situation professionnelle avec la sociabilité familiale sont entièrement imputables au fait que ces deux variables sont toutes deux déterminées par les mêmes autres variables. En d'autres termes et, en citant Durkheim, on aura la preuve que « la concomitance (observée) est due non à ce qu'un des phénomènes est la cause de l'autre, mais à ce qu'ils sont tous deux effets d'une même cause » [16]

Si des écarts subsistent, en revanche, on comprendra que les écarts bruts (observés) ne sont pas imputables aux seules variables de contexte, mais que « quelque chose en plus », qui a un rapport avec la situation professionnelle, entretient des liens avec la sociabilité

16. DURKHEIM, É., 1895.

familiale. Il peut s'agir d'une variable non prise en compte (car non mesurable, ou à laquelle on n'a pas pensé) qui :

— soit, détermine à la fois la situation professionnelle et la sociabilité,

— soit, est déterminée par la situation professionnelle et détermine la sociabilité,

— soit, est déterminée par la sociabilité familiale et détermine la situation professionnelle.

Avant d'examiner les résultats obtenus, précisons rapidement les particularités de la régression logistique (ou modèle LOGIT) par rapport au schéma général de la régression que nous avons évoqué dans la première partie. Dans une modélisation logistique, la variable dépendante est une variable qualitative, qui comporte dans les cas les plus simples deux modalités, voire trois comme dans l'exemple étudié ici. Les variables explicatives sont le plus souvent également qualitatives. Mettre en évidence l'effet d'une variable indépendante X_i sur la variable dépendante Y, revient alors à regarder si la probabilité de prendre telle modalité de Y plutôt que telle autre pour un individu, varie selon que cet individu a telle ou telle caractéristique pour la variable X_i. On raisonne ainsi par rapport à une situation de référence, sorte d'individu-type dont les caractéristiques sont repérées en italique dans le tableau, et on regarde les variations de probabilité lorsque l'on s'écarte de cette situation donnée.

Tableau 6. *Effet, toutes choses égales par ailleurs, des caractéristiques sociodémographiques sur la sociabilité familiale**

Modalité de référence	*Modalité active*	*Coefficients*	*Test statistique*
Sexe	Femme	0,15	$p < 0,003$
Homme			
Âge	18 à 24 ans	0,65	$p < 0,001$
35 à 49 ans	25 à 29 ans	0,87	$p < 0,001$
	30 à 34 ans	0,46	$p < 0,001$
	50 à 64 ans	0,68	$p < 0,001$
Nationalité du père	CEE	- 0,49	$p < 0,001$
Française	Hors CEE	- 1,45	$p < 0,001$
	Inconnue	- 1,22	$p < 0,001$

Commune	Commune rurale	0,40	p < 0,001
Paris	Commune < 100 000	0,44	p < 0,001
	Commune > 100 000	0,31	p < 0,003
	Agglom. parisienne	0,03	n.s.
Catégorie socio-professionelle	Agriculteurs	0,00	n.s.
Employés	Artisans, Commerçants, Chefs d'entreprise	0.27	p < 0,01
	Cadres supérieurs	0,27	p < 0,01
	Prof. intermédiaires	0,20	p < 0,01
	Ouvriers qualifiés	0,09	n.s.
	Ouvriers non qualifiés	- 0,24	p < 0,005
	Non déclaré	- 0,08	n.s.
Diplôme	Aucun diplôme	- 0,24	p < 0,01
Baccalauréat	CEP	- 0,25	p < 0,01
	BEPC	- 0,01	n.s.
	CAP/BEP	- 0,05	n.s.
	Études supérieures	0,02	n.s.
Situation par rapport à l'emploi	Emploi stable menacé	- 0,06	n.s.
	Emploi instable	- 0,31	p < 0,001
Emploi stable	Chômage < 2 ans	- 0,36	p < 0,001
	Chômage > 2 ans	- 0,35	p < 0,002
Revenu par u.c.	< 1800 F	- 0,02	n.s.
< 5 100 F	1 800 F < 2 700 F	0,04	n.s.
	2 700 F < 5 100 F	0,03	n.s.
Taille du ménage	Sans enfant	0,62	p < 0,001
2 enfants	1 enfant	0,08	n.s.
	3 enfants et plus	- 0,30	p < 0,001
Trajectoire conjugale	Couple recomposé	0,10	n.s.
Couple sans rupture	Seul après rupture	- 0,27	p < 0,001
	Toujours vécu seul	- 1,38	p < 0,001
Problèmes de jeunesse	Problème familial	- 0,55	p < 0,001
Pas de problème	Problème d'argent	- 0,15	p < 0,006
	Prob. fam. et d'argent	- 0,55	p < 0,001

* Variable hiérarchisée comportant trois modalités (sociabilité forte ; sociabilité moyenne ; sociabilité faible). Le coefficient estimé pour les individus définis par la modalité active indique l'intensité de leur sociabilité familiale par rapport aux individus définis par la modalité de référence.

N = 7 517.
Source : Enquête INSEE « Situations défavorisées » 1986-1987.
Champ : Ensemble des actifs de 18 à 64 ans (PAUGAM, S., 1994 ; PAUGAM, S. & *alii*, 1993).

De cette manière, on constate, en premier lieu, que la sociabilité familiale est liée au sexe : les femmes rencontrent plus de membres de leur famille, « toutes choses égales par ailleurs », que les hommes. En effet, le coefficient associé à la modalité « femme » est positif (il vaut 0,15) et il est significativement différent de 0 (la probabilité que sa « non nullité » soit due au hasard de l'échantillonnage est inférieure à 0,003, ce qui est très peu).

On voit ensuite que le fait d'appartenir à la classe d'âge 18-24 ans, augmente la probabilité d'avoir une sociabilité familiale importante, par rapport aux individus de la modalité de référence, c'est-à-dire ceux qui ont entre 35 et 49 ans (coefficient égal à 0,65, donc positif et significativement différent de 0 car la probabilité qu'il soit nul est inférieure à 0,001). Et cela est vrai pour toutes les autres classes d'âge considérées, ce qui signifie que la tranche d'âge 35-49 ans est celle pendant laquelle on est le moins susceptible d'avoir des contacts familiaux (en-dehors du ménage).

Le lieu de résidence joue également un rôle. Les individus vivant en province ont une sociabilité plus forte que les Parisiens, à autres caractéristiques contrôlées. En revanche, le coefficient associé à la modalité « Agglomération parisienne » (en fait la banlieue parisienne) est certes positif, mais non significativement différent de 0. Ainsi, le fait de résider en banlieue parisienne n'induit pas de comportement différent de celui des Parisiens.

Sans passer en revue toutes les variables, notons que le revenu en soi n'a pas d'influence sur la sociabilité familiale, qu'en revanche celle-ci est plus forte pour un individu vivant en couple (recomposé ou non) que pour un individu seul, alors qu'elle diminue en intensité avec le nombre d'enfants. Enfin, le fait d'avoir eu des problèmes dans sa jeunesse conduit à une attitude d'éloignement vis-à-vis de la famille. Notons que cette dernière variable est une des rares (avec les variables d'état) pour laquelle il est possible de tenir un discours en termes de causalité, puisque les problèmes éventuels ont eu lieu dans la jeunesse, donc avant que ne se pose pour l'individu la question de la sociabilité familiale.

Enfin, on remarque que certains des coefficients associés à la variable « Situation par rapport à l'emploi » sont significativement différents de 0. Ainsi donc, nous avons là, la réponse à notre question, : il y a effectivement un lien entre situation professionnelle et sociabilité familiale, qu'on ne peut imputer exclusivement à l'effet de structure. Pour connaître la nature exacte de ce lien, il faudrait disposer de données supplémentaires. L'hypothèse que l'on peut formuler est bien évidemment que le fait de perdre un emploi, et plus encore de ne pas parvenir à en retrouver, provoque un repli des individus sur eux-mêmes, ce qui leur fait abandonner toute velléité de contact, même avec leur propre famille. Pour valider cette hypothèse, il faudrait pouvoir prouver qu'avant de perdre leur emploi, les individus avaient une sociabilité familiale importante et que, depuis qu'ils sont au chômage, les contacts familiaux se sont raréfiés, voire

ont disparu. Certes, on sait que les individus qui ont un emploi ont une sociabilité forte en moyenne, mais cela ne prouve rien, dans la mesure où il ne s'agit pas des mêmes individus. En poussant le raisonnement jusqu'au bout de sa logique, on pourrait imaginer un scénario inverse, où des individus acariâtres seraient complètement séparés de leur famille qui ne supporterait plus leur mauvaise humeur permanente et, dans le même temps, inemployables car incapables de s'entendre avec leurs collègues. Bien sûr, cette hypothèse est absurde et selon toute vraisemblance la première thèse est la bonne, comme d'autres travaux qualitatifs l'ont suggéré. Mais, du point de vue logique, elle ne peut être éliminée.

S. Paugam a, dans la même étude, envisagé les liens de la situation par rapport à l'emploi avec la rupture conjugale. Comme on ne sait pas qui, de la rupture ou de la perte d'emploi, est intervenue en premier, on ne peut pas interpréter la liaison en terme de causalité. Il est tout aussi plausible d'imaginer qu'un individu fragilisé par son divorce ne soit plus aussi efficace professionnellement que par le passé, et donc se retrouve dans la vague de licenciement de son entreprise, que l'inverse, à savoir que la situation de chômage de l'individu finisse par provoquer une mésentente conjugale qui se solde par une séparation. Non seulement les deux hypothèses sont plausibles, mais selon toute vraisemblance, elles sont toutes les deux vraies. On est alors en face de processus interdépendants, l'un ayant une influence sur l'autre et inversement. Rechercher le sens de la causalité perd alors toute signification.

L'utilisation des méthodes de régression en sciences sociales réactive, nous espérons l'avoir montré, la réflexion sur les liens entre les phénomènes, les relations de dépendance entre variables et plus généralement la causalité. Il est très difficile d'échapper au vocabulaire causal dès lors que l'on cherche à mettre en œuvre ce type de techniques, vraisemblablement parce que les sujets des verbes sont des variables et non des groupes sociaux. Mais l'utilisation peu scrupuleuse de l'analyse factorielle peut conduire au même type de dérive si l'on n'y prend pas garde. Si l'une et l'autre de ces techniques sont des outils d'une grande puissance pour l'analyse des faits sociaux, elles peuvent également se révéler extrêmement dangereuses, lorsqu'elles sont utilisées sans précaution. C'est pourquoi, nous avons voulu avant tout insister sur ce qu'on ne peut pas dire, sur les tentations d'interprétations abusives et appeler les futurs utilisateurs à beaucoup d'humilité dans leurs démonstrations, tout en leur souhaitant de ne pas se laisser envoûter par la magie des logiciels.

Tableau récapitulatif

	Analyse factorielle	*Régression multiple*
But recherché	« Description »	« Explication » (au sens statistique) à des fins d'action
	Appréhender l'information pertinente contenue dans les données de départ, par le classement automatique	Construction de modèles « explicatifs »
Statut des variables	Toutes les variables sont sur le même plan conceptuel	Distinction entre les variables explicatives (exogènes) et les variables à expliquer (endogènes)
Éléments servant de base à l'interprétation des résultats	Juxtaposition ou proximité entre les caractéristiques de groupes sociaux et leurs pratiques	Liaison « causale » entre facteurs et effets
Sujets des verbes dans l'interprétation	Groupes sociaux	Variables

BIBLIOGRAPHIE

Cahiers du MAGE, « Salaires : Toutes choses inégales par ailleurs ? », *Temps partiels, Salaires inégaux*,, n° 2, 1995, pp. 3-37.

CATTELL, Raymond B., *Factor Analysis*, New York, Harper, 1952, 21 p.

CIBOIS, Phillippe, *L'analyse des données en sociologie*, PUF, Le Sociologue, Paris 1984, 220 p. (pour une description complète de la méthode factorielle et de sa mise en œuvre).

DESROSIÈRES, Alain, « Classer et mesurer : les deux faces de l'argument statistique », *Réseaux*, n° 71, mai-juin 1995, pp. 11-29.

DURKHEIM, Émile, *Les règles de la méthode sociologique*, 1895, (réed. Paris, PUF, 1937), 149 p.

HIRSCHI, Travis, SELVIN, Hanan C., *Recherches en délinquances. Principes de l'analyse quantitative*, Paris, (édition française) Mouton, 1975, 294 p.

LAZARSFELD, Paul, « L'interprétation des relations statistiques comme procédure de recherches », *in* R. BOUDON & P. LAZARSFELD (éds.) *L'analyse empirique de la causalité*, Paris, Mouton, 1966, pp. 19-27.

PASSERON Jean-Claude, « Ce que dit un tableau et ce qu'on en dit », *Le raisonnement sociologique*, Nathan, 1991, pp. 111-133.

PAUGAM, Serge, « L'espace de la précarité. Éléments pour une analyse des inégalités face au risque d'exclusion », Présentation au Séminaire de Stratification Sociale, CREST-INSEE, 19 décembre 1994, 33 p.

PAUGAM, Serge, ZOYEM, Jean-Paul & CHARBONNEL, Jean-Michel, *Précarité et risque d'exclusion en France*, Document du Centre d'Étude des Revenus et des Coûts, n° 109, Paris, La Documentation Française, 1993, 169 p.

POPPER Karl, *The logic of scientific discovery*, New York, Basic Books, 1959, 480 p.

VALLET, Louis-André & CAILLE, Jean-Paul, « Les carrières scolaires au collège des élèves étrangers ou issus de l'immigration », *Éducation et Formations*, n° 40, 1995, pp. 5-14.