Petit guide des statistiques exploratoires en sciences sociales

Michel Grossetti

▶ **To cite this version:**

Michel Grossetti. Petit guide des statistiques exploratoires en sciences sociales. 2023. halshs-03947774v2

HAL Id: halshs-03947774
https://shs.hal.science/halshs-03947774v2

Preprint submitted on 18 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Michel Grossetti

Petit guide des statistiques exploratoires en sciences sociales

Version 1 - Janvier 2023

Préambule

Je suis sociologue depuis très longtemps, mais j'ai été formé initialement en mathématiques appliquées (DEA de mathématiques appliquées en 1981, enseignant de mathématiques dans le secondaire de 1981 à 1983) et j'ai travaillé plusieurs années (de 1983 à 1987) comme ingénieur statisticien, d'abord dans une petite structure de service annexe d'un laboratoire, puis dans un centre interuniversitaire de calcul où j'étais chargé d'aider les chercheurs de toutes les disciplines à utiliser les logiciels qui étaient disponibles à l'époque. Devenu en 1988 chercheur au CNRS en sociologie, j'ai continué à faire des statistiques, pour des doctorants, des collègues, ou moi-même bien sûr. Formé au départ dans une équipe d'analyse de données « à la française » qui développait des variations diverses sur les techniques factorielles, j'ai appris à utiliser à peu près toutes les autres techniques et je me suis construit progressivement des habitudes de travail qui privilégient la compréhension des données et la robustesse des résultats. Ces habitudes consistent à effectuer un travail exploratoire important, au moyen de techniques simples de description ou de recherche de corrélations et aussi de recombinaisons diverses de variables. Il m'est arrivé souvent de former des chercheurs en sciences sociales à ces techniques et l'expérience m'a montré qu'avec quelques heures de pratique et des logiciels d'usage simple ils se montraient capables de traiter très efficacement leurs données. Je suis convaincu qu'un spécialiste des sciences sociales qui connait bien ses données est plus pertinent dans ses analyses qu'un ingénieur statisticien qui n'a pas en tête les éléments de problématique et la connaissance des objets, même si les conseils d'un spécialiste sont utiles à toutes les étapes. A mon avis, les ingénieurs devraient prendre en charge la mise en œuvre des techniques lorsque les chercheurs en sciences sociales ont déjà analysé leurs données au moyen de méthodes simples et qu'ils ont besoin de techniques plus sophistiquées pour résoudre certains problèmes.

Ce texte est destiné aux étudiants ou chercheurs en sciences sociales qui n'ont pas de formation particulière en mathématiques ou en statistiques. Si des enseignants chargés des cours d'initiation aux statistiques y trouvent un intérêt, qu'ils se sentent libres de l'utiliser comme bon leur semble, d'en reprendre des éléments ou la totalité, de l'adapter ou le modifier. Ou même de le compléter par des exemples plus documentés du point de vue de la pratique de tel ou tel logiciel. Je n'ai évidemment aucune revendication d'« auteur » sur ces techniques qui font partie du patrimoine scientifique général et je n'en ai pas non plus sur la façon dont je les présente dans ce texte. Si ce petit texte pouvait susciter la rédaction collective d'un corpus de documents décrivant les usages de telle ou telle technique ou discutant des questions plus générales sur les statistiques, j'en serais ravi.

Les lecteurs ou utilisateurs qui auraient des questions ou des remarques peuvent m'écrire (Michel.Grossetti(at)univ-tlse2.fr).

Résumé

L'objectif de ce petit livre est de présenter les techniques de statistique exploratoire et leurs usages pour des personnes qui connaissent les sciences sociales mais qui ne sont pas familières de ces techniques. Il commence par présenter la logique de mise en équivalence qui préside à la construction des variables, puis la conception du hasard comme forme particulière d'imprévisibilité, et la « loi des grands nombres », un phénomène de saturation de l'information lorsque celle-ci porte sur des activités suffisamment peu dépendantes les unes des autres. Ensuite, il traite des types de données, et de leur mise en forme par la construction de variables de différentes catégories. La première étape de l'exploration est un examen de chaque variable prise séparément pour identifier les valeurs extrêmes ou les répartitions déséquilibrées, et procéder aux recodages nécessaires. La deuxième étape consiste à combiner les variables par des techniques simples de recherche de corrélations, tableaux de contingences, coefficients de corrélation, comparaison de moyennes ou de médianes. Les analyses simples de corrélations peuvent déboucher sur, ou se combiner à, l'usage de méthodes plus globales comme les résumés graphiques (analyses de correspondances), les typologies (classifications), ou les modèles cherchant à évaluer la part prise par diverses variables explicatives dans les variations d'une variable à expliquer. Dans certains cas, les données portent sur des relations et les réseaux qu'elles constituent, ce qui implique des méthodes spécifiques, qui relèvent, soit d'un usage particulier des méthodes statistiques standard (en distinguant les personnes interrogées des relations qu'elles ont citées), soit de méthodes spécifiques d'analyse de graphes (qui ne sont pas détaillées dans ce texte). D'autres données portent sur des processus, qui peuvent être étudiés au moyen de méthodes qui tiennent compte des enchaînements de séquences. Enfin, il existe de nombreux cas où l'on a affaire à des corpus de données textuelles, ce qui amène à utiliser des méthodes statistiques d'une manière particulière. Chaque technique est présentée au moyen d'exemples issus de données d'enquêtes réelles.

Remerciements

J'ai fait lire ce texte à divers collègues, dont les suggestions m'ont été très utiles : Alain Degenne, Jean-Michel Chapoulie, Pierre Blavier, Julien Gros…. Je les remercie chaleureusement.

Table des matières

Introduction

Depuis les débuts des enquêtes empiriques aux XIXe siècle, les sciences sociales utilisent des analyses quantitatives. Elles ont en cela suivi les usages des administrations[1], utilisant les données collectées par ces dernières ou réalisant leurs propres enquêtes. L'analyse de ces données ne s'écarte pas fondamentalement des usages des méthodes statistiques dans les autres sciences, à ceci près que, sauf dans quelques cas, les sciences sociales ne procèdent pas à des expérimentations contrôlées, et que les données qu'elles utilisent sont issues soit d'observations ou d'archives qui ont été codées par la suite, soit de collectes de données organisées en milieu social ordinaire au moyen de questionnaires conçus par les administrations ou par les chercheurs eux-mêmes. Ces données sont toujours complexes, les informations rassemblées n'étant pas toujours disponibles pour toutes les entités étudiées (individus, organisations, textes, situations sociales, etc.) et les variables constituées comportant toujours une certaine part d'ambiguïté dans les significations auxquelles elles renvoient. Cela implique que l'analyse de ces données comporte toujours une part prépondérante dédiée à la compréhension de leur constitution et des informations qu'elles comportent, et des tendances que révèle l'analyse de ces informations. Cette part de l'analyse, que je qualifie ici d'**exploratoire** (on pourrait aussi utiliser le terme « heuristique »), constitue l'essentiel de la construction des résultats. Elle peut déboucher sur une analyse plus probatoire lorsqu'il s'agit de passer d'un résultat qui semble solide à une publication scientifique qui implique d'étayer l'argumentation par des tests statistiques ou d'autres méthodes acceptées comme éléments de preuve par les comités de lecture des revues. Comme nous le verrons, cela ne signifie pas que l'exploration n'utilise pas les statistiques de tests ou les modèles, car il existe un usage exploratoire de ceux-ci, malheureusement trop peu connu, mais que l'objectif est d'abord la compréhension des données et de ce qu'elles révèlent des phénomènes étudiés et non l'établissement de preuves ou d'inférences quantifiées destinées à convaincre des communautés en respectant des normes méthodologiques de ces communautés. Si l'exploration est bien conduite, les tendances identifiées résisteront aux tests et autres méthodes probatoires qu'il faudra réaliser pour certaines publications.

Le présent ouvrage est consacré à ce travail d'exploration des données. Il commence par présenter les fondements des analyses statistiques, la logique de mise en équivalence qui préside à la construction des variables, la conception du hasard comme forme particulière d'imprévisibilité, et la « loi des grands nombres », un phénomène de saturation de l'information lorsque celle-ci porte sur des activités suffisamment peu dépendantes les unes des autres. Ce phénomène, identifié par les mathématiciens au 17e siècle est la tendance de certaines mesures à se stabiliser lorsque l'on accumule les observations. Ensuite, j'évoquerai les types de données, de première ou de seconde main, et leur mise en forme par la construction de variables de différentes catégories. Ces catégories commandent les méthodes que l'on peut utiliser : cela n'a pas de sens de calculer une moyenne sur les catégories de situation professionnelle et pas non plus de comptabiliser les personnes selon leur salaire exprimé en centimes. Une fois les variables constituées, il faut les analyser. La première étape est un examen de chaque variable prise séparément pour identifier les valeurs extrêmes ou les répartitions déséquilibrées, et procéder aux recodages nécessaires. La deuxième étape de l'analyse est la plus importante à mes yeux. Elle consiste à combiner les variables par des techniques simples de recherche de corrélations, tableaux de contingences, coefficients de corrélation, comparaison de moyennes ou de médianes, pour tester des idées, comprendre où se situent les variations les plus intéressantes pour la problématique de l'étude. C'est l'occasion de rappeler que l'analyse de données implique toujours des problématiques et des questions de recherche, qu'elle n'est jamais totalement dénuée d'hypothèses, plus ou moins explicites. Par exemple, si l'on calcule des différences de salaire entre hommes et femmes c'est parce que l'on a des hypothèses, plus ou moins explicites et précises, sur l'existence de différences et sur ce qui peut expliquer éventuellement ces différences.

[1] Alain Desrosières, 1993, *La politique des grands nombres Histoire de la raison statistique*, Paris, La Découverte.

Les analyses simples de corrélations peuvent déboucher sur, ou se combiner à, l'usage de méthodes plus globales comme les résumés graphiques (analyses de correspondances), les typologies (classifications), ou les modèles cherchant à évaluer la part prise par diverses variables explicatives dans les variations d'une variable à expliquer. L'important est d'explorer les données en détail, de chercher à les comprendre, de ne pas se contenter de plans standardisés de traitement enchaînant mécaniquement les techniques. Dans certains cas, les données portent sur des relations et les réseaux qu'elles constituent, ce qui implique des méthodes spécifiques, qui relèvent, soit d'un usage particulier des méthodes statistiques standard (en distinguant les personnes interrogées des relations qu'elles ont citées), soit de méthodes spécifiques d'analyse de graphes. D'autres données portent sur des processus, qui peuvent être étudiés au moyen de méthodes qui tiennent compte des enchaînements de séquences. Enfin, il existe de nombreux cas où l'on a affaire à des corpus de données textuelles, ce qui amène à utiliser des méthodes statistiques d'une manière particulière.

Pour une bonne part des exemples qui suivent, notamment dans les premiers chapitres, j'utiliserai un fichier Insee sur un échantillon de plus de deux millions de salariés français en 2015 (https://www.insee.fr/fr/statistiques/3536754) qui correspond à un douzième de la population totale des salariés. Comme nous le verrons, la loi des grands nombres permet de considérer cet échantillon aléatoire de grande taille comme représentatif de l'ensemble dont il est extrait. J'ai pris le parti de ne pas expliquer comment on met en œuvre concrètement les techniques présentées sur tel ou tel logiciel (ou avec un papier et un crayon comme c'est parfois possible). Cette question qui angoisse en général les débutants est à mon sens secondaire. Les logiciels ont leur importance parce qu'ils permettent de mettre en œuvre plus ou moins facilement les différentes techniques, mais ce sont ces dernières et la logique de leur usage qui sont primordiales. Tous les pratiquants finissent par trouver les moyens de mettre en œuvre les techniques[2]. Il est bien plus important de comprendre à quoi celles-ci peuvent servir et dans quel cadre on peut les utiliser.

Dans cet esprit, j'ai choisi d'écrire ce texte en langage naturel, en évitant le plus possible le formalisme mathématique. L'objectif est de permettre à des personnes qui connaissent les sciences sociales mais qui sont peu familières avec ce formalisme de comprendre la logique de l'analyse exploratoire de données à l'aide des techniques statistiques usuelles.

[2] Pour celles et ceux qui paniqueraient face à la réalisation concrète des techniques, je peux rappeler la procédure la plus efficace : d'abord taper sur un moteur de recherche « comment faire une « nom de la technique » sur « nom du logiciel » », ensuite si aucune solution n'a été trouvée au bout d'un quart d'heure, consulter son carnet d'adresse, trouver quelqu'un qui connaît ce genre de logiciel et lui écrire un message poli et aimable pour demander conseil. Si c'est vraiment difficile, il est aussi possible de poster un message dans une liste de discussion (quanti@groupes.renater.fr par exemple).

1. Fondements de l'analyse statistique

L'analyse statistique se fonde sur plusieurs principes qui ne sont pas toujours clairement énoncés dans les ouvrages d'enseignement qui tendent à se centrer sur les aspects techniques. Le premier est la mise en équivalence, qui commande la formation des variables, que celles-ci prennent la forme de mesures ou de catégories. Compter, recenser, mesurer implique de définir des commensurabilités et des équivalences, ce qui ne va jamais de soi. Le deuxième fondement est la prise en compte du hasard, du fait que les régularités étudiées ne sont jamais mécaniques (ce qui prendrait la forme d'affirmations du type « si X vaut x alors Y vaut toujours immanquablement y ») mais se combinent dans les données avec des variations « aléatoires », c'est-à-dire dues à des causes considérées comme négligeables pour une analyse donnée. La notion de corrélation, centrale dans l'analyse statistique exploratoire, implique de faire la part entre les régularités recherchées et cette « part du hasard ». Enfin, troisième fondement, la loi des grands nombres, qui permet de faire le lien entre des échantillons d'observations et des populations plus vastes. Une section est également consacrée à la différence entre corrélation et causalité.

Mettre en équivalence

Compter c'est mettre en équivalence des activités ou des caractéristiques. Par exemple, à son niveau le plus agrégé, la nomenclature française des professions et catégories socio-professionnelles (PCS) comprend une catégorie « ouvriers » qui englobe de très nombreux métiers sur la base d'un type d'activité, d'un niveau de qualification et d'un niveau de rémunération. A un niveau un peu plus détaillé, la nomenclature distingue 7 catégories d'« ouvriers »[3]. A un niveau encore plus détaillé elle en comprend 126. La constitution de ces catégories de sorte à ce qu'elles soient le plus homogène possible mobilise des experts durant de longues périodes et fait l'objet de débats intenses. De la même façon, une mesure comme le salaire annuel par exemple, implique de faire de nombreux choix (faut-il prendre en compte les primes ? d'autres revenus annexes ?). Dans les deux cas, cette mise en équivalence est indispensable pour faire apparaître des régularités (par exemple le fait que les ouvriers ont des salaires plus faibles que ceux d'une catégorie telle que celle, tout aussi complexe, des « cadres »). Cette mise en équivalence doit se faire en évitant deux pièges. Le premier est celui du postulat de la non commensurabilité, qui consisterait à considérer qu'aucune mise en équivalence n'est possible, chaque situation étant unique. On s'interdit alors de faire apparaître des régularités et de voir qu'il y a, pour prendre un exemple qui a beaucoup occupé les sociologues, des différences de réussite scolaire selon la profession des parents. Le deuxième piège est la réification des variables, l'oubli de leur construction. On a ainsi pu voir parfois des discours transformant l'« espérance de vie », qui est une moyenne de durée de vie calculée sur l'ensemble de la population, en prédiction du nombre d'années restant à vivre pour des personnes particulières, ce qui est absurde. Le fait qu'une personne ait atteint l'âge moyen des décès dans la population à laquelle elle appartient ne peut pas être transformé en prédiction de sa mort très prochaine. On voit aussi parfois des discours transformant telle ou telle catégorie de population (par exemple « les immigrés ») en ensemble homogène, dont les comportements ou les caractéristiques seraient identiques, ce qui est également absurde. Il faut mettre en équivalence mais faire preuve d'une réflexivité permanente dans la construction et l'usage des mesures ou des catégories.

[3] « Ouvriers qualifiés de type industriel » ; « Ouvriers qualifiés de type artisanal » ; « Chauffeurs » ; « Ouvriers qualifiés de la manutention, du magasinage et du transport » ; « Ouvriers non qualifiés de type industriel » ; « Ouvriers non qualifiés de type artisanal » ; « Ouvriers agricoles et assimilés ».

Faire la part du hasard

Les régularités que les statistiques s'efforcent de détecter ne sont pas mécaniques, ce sont des corrélations. Cela signifie que le lien entre deux caractéristiques (la catégorie professionnelle et le salaire par exemple) n'est jamais une correspondance stricte (tous les membres d'une même catégorie recevraient un salaire identique) mais une régularité, une tendance. Cela implique de trouver des moyens pour distinguer cette régularité de variations dues à d'autres causes. Ce point sera développé dans des chapitres ultérieurs, mais il est utile de signaler dès à présent que, en statistiques, le « hasard » est en général considéré comme la combinaison de petites causes indépendantes les unes des autres (par exemple le fait que tel ouvrier ait été particulièrement augmenté parce qu'il a une compétence particulière, qui fait que son salaire est plus élevé que ceux de ses collègues) qui produisent des variations considérées comme négligeables pour une analyse donnée. C'est pourquoi une régularité statistique ne peut pas être infirmée par un contre-exemple. Par exemple le fait qu'il existe des enfants d'origine sociale favorisée qui connaissent des difficultés scolaires et à l'inverse des élèves d'origine modeste accédant à des diplômes socialement valorisés ne contredit en rien l'existence d'une corrélation entre l'origine sociale et la réussite scolaire.

Ce hasard des statistiques ne recouvre pas la totalité de ce que l'on peut désigner comme des imprévisibilités[4]. C'est un type particulier d'imprévisibilité, qui peut être isolé des régularités. Par exemple si l'on veut exprimer le lien entre le salaire S et la catégorie professionnelle PCS, on n'écrit pas $S = f(PCS)$, f étant une fonction mathématique, mais $S = f(PCS) + \varepsilon$, ce dernier terme exprimant « la part du hasard » qui est censée, premièrement, pouvoir être isolée de la régularité recherchée dans la variation des salaires selon la profession, et, deuxièmement, se distribuer d'une certaine façon (avec une fréquence faible des valeurs qui s'éloignent de la moyenne en général). Ce hasard est conceptualisé depuis les travaux du mathématicien Cournot comme la combinaison de causes indépendantes les unes des autres. L'analyse statistique implique toujours une comparaison entre ce qui est observé et ce qui résulterait du « hasard », afin d'identifier et de mettre en évidence des régularités. Je mets des guillemets à « hasard » parce qu'il s'agit d'une convention adoptée le temps d'une analyse, avant que l'on découvre d'autres facteurs explicatifs. On cherche à déterminer si des différences observées dans une variable par rapport aux valeurs prises par une autre variable pourraient être expliquées par des combinaisons de petites causes que l'on considère comme négligeables et que qu'on ne connait pas, mais que l'on représente par certaines distributions de valeurs (notamment la loi « normale » sur laquelle je reviendrai). C'est le cas lorsque l'on dispose de données exhaustives (toutes les mesures possibles sur une population donnée, par exemple tous les salariés français). Mais il arrive très souvent que l'on ne dispose pas de données exhaustives, mais seulement d'un échantillon, un ensemble d'observations censé représenter une population plus vaste. Les analystes peuvent alors s'appuyer sur un phénomène très important qui est la loi des grands nombres.

La loi des grands nombres

Les habitués des méthodes statistiques savent que dans beaucoup de phénomènes sociaux on observe ce que l'on appelle la loi des grands nombres, identifiée au XVII[e] siècle par des mathématiciens[5] : des régularités observées sur un échantillon d'activités se retrouvent à des niveaux plus amples de population. Supposons par exemple que l'on observe une corrélation entre l'origine sociale des élèves et leur réussite scolaire sur un échantillon suffisamment large (quelques centaines par exemple si l'on veut pouvoir utiliser des catégories assez précises) et représentatif,

[4] Michel Grossetti, 2016, « L'imprévisibilité dans le monde social », in Jean-Claude S. Levy, *Complexité et désordre : éléments de réflexion*, EDP Sciences, pp.97-112, 2016, Grenoble sciences-rencontres scientifiques, 978-2-7598-1777-1. hal-01390035.

[5] On crédite en général Jacques Bernoulli (1654-1705) pour sa formulation.

c'est-à-dire présentant les mêmes variations que la population dont il est extrait pour certaines caractéristiques, ici les professions des parents et les types de parcours scolaires. Il y a de fortes chances que cette corrélation se retrouve de façon presque identique dans des mesures qui seraient effectuées sur la totalité de la population scolaire. C'est ce qui justifie le recours à des échantillons pour de nombreuses études, y compris qualitatives, puisque le principe de saturation des variations utilisé dans la plupart des enquêtes par entretien[6] relève en fait du même principe. C'est ce phénomène qui explique que les instituts de sondage parviennent à prédire assez précisément des résultats électoraux à partir d'échantillons très restreints en comparaison de la population des électeurs lorsque les conditions sont réunies (les sondages à la « sortie des urnes » par exemple, qui évitent les changements d'avis des électeurs dans les derniers jours d'une campagne électorale et limitent la sous-déclaration de votes jugés transgressifs). Il est courant de moquer ces instituts pour leurs erreurs de prévision, ce qui est compréhensible au regard de l'assurance très excessive avec laquelle leurs représentants s'expriment parfois, mais cela n'enlève rien au fait que des mesures effectuées sur un ou plusieurs milliers de personnes permettent d'estimer assez précisément les résultats d'une consultation électorale impliquant des millions de votants.

Cependant la loi des grands nombres ne fonctionne pas toujours. Elle s'applique assez bien à toutes les situations dans lesquelles on peut raisonner comme si les phénomènes microsociologiques concernés (votes, parcours scolaires, problèmes de santé, etc.) étaient peu dépendants les uns des autres. Cette indépendance relative n'est pas une caractéristique intrinsèque des activités sociales, elle résulte dans bien des cas de dispositifs bien précis d'individuation : l'urne et l'isoloir qui sont censés empêcher les influences au moment du vote, l'examen anonyme, la supposée neutralité des professeurs par rapport aux caractéristiques sociales des élèves, etc. Dans les activités de consommation, il faut aussi de multiples dispositifs pour que celles-ci s'effectuent sur un mode individuel[7].

Mais il existe une grande variété de situations sociales dans lesquelles ces conditions ne sont pas réunies. Elles peuvent être simplement affaiblies, par exemple lorsque, comme c'est le cas le plus fréquent dans les activités sociales, il existe des phénomènes cumulatifs, du type « les riches s'enrichissent » ou « les personnes déjà connues ont plus de chances de l'être encore plus que de parfaits inconnus ». Ces effets résultent de multiples processus à priori peu dépendants les uns des autres mais dont les effets tendent à se cumuler : dans un contexte d'économie de marché, il y a de multiples façons pour les personnes disposant d'une fortune d'accroître celle-ci ; dans une période où les médias sont très présents, la notoriété a des sources très variées. Les effets cumulatifs expliquent que les distributions de probabilité de mesures effectuées sur des activités ou des caractéristiques sociales suivent rarement des lois dites « normales » qui s'équilibrent autour de la moyenne et où les mesures éloignées de celle-ci se raréfient en fonction de cet éloignement, ces distributions « en cloche » que l'on connaît dans bien des domaines[8]. Dans la mesure des activités sociales, on observe plutôt des distributions dites « lognormales »[9], qui ne sont pas symétriques autour de la moyenne et où les valeurs élevées sont très éloignées de celle-ci (il y a beaucoup plus de pauvres et moins de riches mais ceux-ci accaparent une part importante de la richesse). La loi

[6] C'est Daniel Bertaux qui a proposé ce principe d'arrêt des entretiens lorsque les tendances observées et les thèmes abordés se stabilisent (*Les récits de vie : perspective ethnosociologique*, Paris, Nathan, 1997).

[7] Les travaux de Franck Cochoy sur le marketing et le packaging sont éclairants sur ce point (*Une histoire du marketing. Discipliner l'économie de marché*, Paris, La Découverte, 1999, et *Sociologie du packaging ou l'âne de Buridan face au marché*, Paris, Presses Universitaires de France, 2002).

[8] Pour expliquer les distributions de ce type, on prend souvent l'exemple d'une série de personnes mesurant un objet avec un instrument rudimentaire. Il y aura à peu près autant de personnes qui sous-estiment la longueur que de personnes qui la surestiment et le nombre de mesures diminue lorsque l'on s'éloigne de la valeur réelle (normale). La distribution des mesures aura l'allure d'une cloche et la moyenne de ces mesures approximatives sera très proche de la valeur réelle.

[9] Ou parfois des distributions dites « de Pareto », du nom de l'économiste italien Vilfredo Pareto (1848-1923) qui observait que 20% de la population italienne possédait 80% des richesses. Les distributions « lognormales » et « de Pareto » ont des formulations mathématiques distinctes mais sont assez proches et sont utilisées l'une ou l'autre pour décrire des phénomènes cumulatifs.

des grands nombres s'applique mais le risque d'avoir un échantillon non représentatif est plus élevé puisqu'il suffit pour cela que celui-ci n'intègre pas suffisamment d'éléments qui ont des valeurs très élevées. Par exemple le calcul de la moyenne des patrimoines effectué sur un échantillon même large mais ne comprenant aucun des ménages faisant partie des 1% les plus riches risque d'être très décalé par rapport à un calcul fait sur les données exhaustives. Les effets cumulatifs relèvent à la fois d'une interdépendance diffuse et des effets d'inertie des avantages acquis par des personnes ou des organisations.

Dans d'autres cas, l'interdépendance entre les activités mesurées est plus forte. Les délibérations collectives, les mouvements sociaux, l'émergence de collectifs explicites et bien d'autres processus font intervenir des interactions ou au moins des ajustements réciproques qui rompent la condition d'indépendance des activités concernées et font que la loi des grands nombres ne s'applique pas et donc que l'on ne peut pas utiliser les techniques statistiques de la façon la plus courante. Si l'on veut continuer à utiliser des méthodes formelles, il faut alors faire intervenir des techniques comme les analyses de réseaux ou des modélisations de processus dans lesquelles les effets d'une séquence sont des conditions initiales pour la séquence suivante et ainsi de suite. Cela permet d'observer des systèmes « dynamiques » ou « complexes » et des effets non linéaires.

Une illustration de la loi des grands nombres

Dans le fichier DADS-2015, il existe une variable S_NET_TOT qui « correspond aux rémunérations nettes de toutes cotisations sociales salariales obligatoires et de CSG et CRDS, ce qui correspond également au net fiscal duquel ont été retranchées la CRDS et la CSG non déductible », donc le total des salaires nets perçus en une année. Le salaire est exprimé en euros, sans décimales. Pour les 1708819 salariés à temps complet, la moyenne des salaires (c'est-à-dire la somme de tous les salaires, divisée par le nombre de salariés) s'établit à 24850,17 euros. Divisons à présent les salariés en deux groupes selon un critère arbitraire qui ne devrait pas trop influer sur leur niveau de rémunération, le fait que le montant de leur salaire soit un nombre pair ou impair[10]. 853480 salariés ont un salaire total pair, 855339 un salaire total impair. La moyenne des salaires du premier groupe est de 24879,38 euros et celle du second groupe 24821,02 euros, donc des nombres très peu différents[11]. Chaque groupe peut être considéré comme un échantillon acceptable de l'ensemble.

Corrélations et causalités

Dans les situations où il est pertinent de les utiliser, les statistiques peuvent seulement établir des corrélations, dont l'interprétation fait intervenir bien d'autres informations et considérations que les seuls indicateurs quantifiés. L'interprétation causale des corrélations est une des plaies des sciences sociales, qui atteint peut-être son apogée lorsqu'il s'agit de corrélations spatiales ou historiques, où la coexistence dans l'espace ou le temps est la base d'inférences causales bien souvent sans fondement. Comment déceler des causes derrière des régularités statistiques ? Dans le cas du lien entre l'origine sociale et la réussite scolaire en France, les recherches empiriques semblent indiquer des causes multiples incluant la transmission de pratiques par les parents (pour la lecture par exemple), des stratégies des parents des couches sociales favorisées par rapport au système scolaire (choix des établissements, mobilisation des dispositifs pouvant éviter un étiquetage jugé néfaste à leurs enfants ou des orientations non recherchées), des habitudes des enseignants aussi bien dans leurs méthodes que dans la façon dont ils perçoivent les élèves, des dispositifs

[10] Cela revient à faire l'hypothèse qu'il n'existe pas de phénomène systématique qui favorise l'une ou l'autre des deux situations. Ce pourrait être le cas par exemple si dans certaines professions l'usage était d'arrondir le salaire à la dizaine d'euros, ce qui favoriserait les nombres pairs.

[11] Une analyse de variance (voir le chapitre 4) sur les logarithmes (idem) conclut à l'absence de différence significative.

institutionnels d'assignation des élèves à des catégories. Des tentatives pour évaluer l'importance relative des différentes causes[12] donnent plus d'importance à la transmission de pratiques (et de valeurs), mais montrent aussi le poids des stratégies d'orientation et les effets de la composition des établissements et des classes. Les processus dont la résultante est la corrélation entre l'origine sociale et le parcours scolaire sont donc multiples et ne se réduisent pas à une cause simple. Cela signifie que les techniques statistiques doivent toujours être multiples et combinées avec des hypothèses et des savoirs plus ou moins stabilisés pour rendre compte d'un phénomène social.

[12] Marie Duru-Bellat, Jean-Pierre Jarousse, Alain Mingat. « De l'orientation en fin de cinquième au fonctionnement du collège : 3 - Les inégalités sociales de carrières du cours préparatoire au second cycle universitaire ». IREDU, 156 p., 1992, *Les Cahiers de l'IREDU*, 2-85634-056-3. hal-02053732

2. Les données et les types de variables

Les sources

Les données analysées par les sciences sociales sont construites à partir des activités des personnes. En effet, tout ce que ces sciences peuvent affirmer sur des entités dérive de données portant sur les activités. Ces dernières se présentent sous la forme d'attitudes, de gestes, de paroles, de traces laissées dans des bâtiments, de textes, d'images … Selon les cas, les chercheurs en sciences sociales analysent les traces laissées « spontanément » par les activités ou ils organisent des situations de création de traces adaptées aux questions qu'ils se posent, en réalisant des observations, ou encore ils cherchent à orienter les activités par des entretiens, des questionnaires ou d'autres formes d'enquête, certains allant jusqu'à des formes d'expérimentation contrôlée, en psychologie sociale par exemple. Ces quatre types d'intervention sur les activités ont des liens avec les divisions disciplinaires, mais seulement partiellement. Ainsi, pour l'analyse des traces, on pense immédiatement à des archéologues s'efforçant de décrypter les mondes anciens à travers des traces matérielles ou à des historiens plongés dans des archives, mais aussi à des spécialistes des mondes contemporains utilisant des données administratives ou des enregistrements électroniques de plus en plus volumineux. L'observation est souvent associée à l'ethnologie, mais elle est de plus en plus largement pratiquée par les sociologues. Les entretiens (il en existe de nombreux types) sont pratiqués par les sociologues, les ethnologues, mais également par certains historiens des périodes contemporaines. Les questionnaires ont été souvent utilisés par les sociologues, mais ils sont très présents en science politique, économie, gestion … L'expérimentation contrôlée pratiquée par les psychologues, et depuis quelques années également par certains économistes, est souvent aux limites des sciences sociales et elle ne sera pas abordée dans ce texte. Toutes les données peuvent donner lieu à des formes de codage pour des analyses statistiques. Si les questionnaires sont conçus dès le départ pour cela, il est toujours possible de coder une partie des informations collectées au moyens d'observations ou d'entretiens, dans la logique des méthodes « mixtes » qui associent des approches qualitatives et quantitatives.

Codages

Un codage est une opération de mise en équivalence de situations sociales diverses et implique donc de nombreux choix. Même lorsque l'on dispose de données issues de questionnaires, on doit toujours s'efforcer de comprendre comment les personnes ont répondu. De multiples études ont montré qu'une même question peut donner lieu à des interprétations très différentes selon les personnes, dont les réponses ne signifient donc pas la même chose[13]. Le travail de mise en équivalence est encore plus important pour des observations ou des entretiens lorsque l'on passe de notes de terrain ou de discours décrivant des situations diverses à des catégories ou des mesures. Tous les travaux sur les catégories professionnelles montrent la difficulté de regrouper dans un même ensemble des métiers divers et l'établissement des nomenclatures est toujours un travail impliquant de nombreuses hypothèses et considérations théoriques.

La première étape d'une analyse consiste à comprendre **d'où proviennent les données, comment elles ont été construites et quelles sont leurs limites**. S'il faut a minima y consacrer du temps avant de commencer à utiliser des techniques, il faut souvent y revenir pour comprendre certains résultats. Par exemple, le fichier des salariés que j'utilise pour beaucoup des exemples pris dans ce texte, ne porte par définition que sur les salariés et ne peut fournir aucune information sur

[13] Ces variations dans la compréhension des questions peuvent être distribuées aléatoirement, mais parfois elles peuvent être corrélées à des caractéristiques sociales, ce qui est plus problématique (je remercie Julien Gros de m'avoir suggéré cette remarque).

les professions libérales et les personnes rémunérées en honoraires ou par les revenus d'un commerce. C'est une limite intrinsèque dont les interprétations doivent en permanence tenir compte.

Représentativité

La question de la **représentativité** est importante. En général les analyses portent sur des échantillons et ont pour ambition de faire apparaître des résultats qui valent pour les populations plus larges dont ces échantillons sont extraits[14]. Il faut d'abord insister sur le fait que **le rapport entre la taille d'un échantillon et celle de la population qu'il représente n'a aucun effet sur la précision des mesures** : que l'on étudie une population de 1000, 100000 ou 10 millions de personnes, le rapport entre la taille de l'échantillon et la précision des mesures est le même.

Il est plus important de prendre la mesure des **biais**, c'est-à-dire des décalages entre l'échantillon et la population de référence en ce qui concerne certaines variables que l'on juge importantes pour l'analyse. Par exemple, si un échantillon est très déséquilibré du point de vue des catégories professionnelles, avec une proportion très élevée de cadres par exemple, il ne permettra pas bien d'évaluer des pourcentages relatifs à des activités culturelles. Si l'on cherche à estimer la proportion de Français fréquentant les musées, ce décalage de l'échantillon peut être un problème puisque toutes les études montrent que cette pratique est plus fréquente pour les membres de cette catégorie professionnelle.

Cependant, l'estimation de pourcentages portant sur une population générale, même si elle prend parfois beaucoup d'importance dans la statistique publique (le taux de chômage par exemple) n'est pas centrale dans les sciences sociales. Le plus souvent ce sont les corrélations qui intéressent les analystes. Par exemple le fait que les cadres fréquentent plus les musées peut parfaitement être établi avec un échantillon biaisé à partir du moment où les autres catégories sont suffisamment représentées. Pour corriger les biais, les fournisseurs de données ou les analystes utilisent souvent des **pondérations** pour « redresser » les échantillons. Cela consiste à calculer des coefficients par lesquels on multiplie certaines observations pour rééquilibrer les proportions pour des variables jugées importantes. Dans le cas d'une surreprésentation des cadres, il faudrait un coefficient inférieur à 1 pour les membres de cette catégorie et supérieur à 1 pour les autres. Par exemple si la proportion de cadres est deux fois plus élevée que dans la population de référence, ce coefficient devrait être de 0,5 pour cette catégorie. Pour calculer ces coefficients, on peut procéder à un calcul analytique de ce type ou recourir à des algorithmes qui ajustent les coefficients par approximations successives par rapport à des proportions connues dans la population générale. Les algorithmes sont utiles lorsqu'il y a un nombre relativement important de variables à corriger. Les pondérations sont efficaces si on dispose de variables très prédictives de la proportion à estimer. Ainsi, dans les sondages politiques, les intentions de vote sont très corrélées au vote lors des élections précédentes, ce qui fait qu'un redressement calculé à partir de cette dernière variable permet en général de s'approcher de la représentativité. En revanche, dans les études plus générales, les variables que l'on cherche habituellement à redresser (âge, sexe, profession, niveau d'études) n'expliquent qu'une part minoritaire des variations des variables qui relèvent du thème étudié plus particulièrement. Les pondérations sont alors peu efficientes et peuvent avoir pour effet néfaste de faire croire à l'analyste qu'il dispose d'un échantillon représentatif alors qu'il ne l'est pas, ou de faire apparaître des corrélations significatives par l'augmentation du nombre apparent de cas, alors que les données brutes portent sur un nombre trop peu élevé pour tirer des conclusions. Les biais peuvent être gérés par des tableaux de contingence à plusieurs entrées ou des modèles de régression que nous verrons plus loin.

[14] Cette ambition est celle d'une généralisation partielle des résultats dont l'inférence statistique normée n'est qu'une forme parmi d'autres. Les enquêtes par observation ou entretien ont la même ambition.

Malgré les efforts des personnes concevant les enquêtes, il est rare que les échantillons soient parfaitement représentatifs. Par exemple, dans les questionnaires, il est très difficile, voire impossible, d'atteindre des personnes en grande difficulté ou à l'inverse des personnes très fortunées. Même l'usage de quotas laisse des effets liés à l'acceptation de l'enquête, acceptation qui n'est pas également distribuée socialement. Enfin, il faut être prudent dans l'analyse d'échantillons de petite taille (moins de 500 pour donner un ordre de grandeur) pour lesquels la combinaison de variables peut être délicate. Les problèmes de représentativité et de taille de l'échantillon rendent d'autant plus nécessaire le travail exploratoire.

Unités statistiques et variables

Un **tableau statistique** comporte en général deux entrées, les lignes qui correspondent aux **unités statistiques** et les colonnes qui correspondent aux informations disponibles sur ces unités, informations qui prennent la forme de **variables**.

Les **unités statistiques** peuvent être très diverses. Si dans beaucoup d'enquêtes elles correspondent à des personnes ayant répondu à des questionnaires ou des entretiens, ou sur lesquelles on dispose d'informations administratives, il y a aussi beaucoup de cas (en économie par exemple) où il s'agit d'entreprises ou d'autres organisations. Parfois ce sont des activités qui sont les unités statistiques. Par exemple si je vais utiliser un fichier Insee de salariés, il existe aussi un fichier « postes » qui correspond aux fiches de salaires prises séparément, une même personne pouvant être salariée de plusieurs employeurs. L'unité statistique est ici l'emploi et non le salarié. Dans les enquêtes sur les relations personnelles, dans l'optique des études de « réseaux personnels », on dispose en général d'un fichier pour les personnes ayant répondu à l'enquête et un autre pour les relations citées par ces personnes, plusieurs lignes de ce fichier correspondant aux relations d'une même personne. Comprendre les unités statistiques et savoir passer d'un type d'unité à un autre est essentiel dans le travail exploratoire.

Il existe trois types de variables : qualitatives non ordonnées, qualitatives ordonnées, quantitatives[15].

Une **variable qualitative non ordonnée** répartit les unités statistiques dans des catégories (qu'on appelle en général des **modalités**) entre lesquelles il n'existe pas à priori un ordre qui les hiérarchiserait. Le sexe, qu'il soit réduit aux deux catégories traditionnelles (femmes, hommes) ou qu'il intègre d'autres catégories, est typiquement une variable de ce type. Cela n'empêche évidemment pas l'existence d'inégalités et de dominations mais les catégories statistiques ne sont pas hiérarchisées à priori. Une autre variable classique de ce type est constituée par les catégories professionnelles avec leurs diverses nomenclatures. Le fait que les « cadres et professions intellectuelles supérieures » soient en général plus diplômés et mieux rémunérés que les « ouvriers » peut parfois conduire à utiliser cette variable sur le registre des variables ordonnées, mais la présence de catégories comme celles des artisans ou des agriculteurs, plus hétérogènes, fait qu'à priori on considère cette variable comme non ordonnée. Lorsque ces variables sont codées par des nombres, ce qui est le cas dans la plupart des logiciels, ces nombres sont arbitraires et **ne peuvent être utilisés pour eux-mêmes en calculant des moyennes ou d'autres indicateurs strictement numériques**. Il faut insister sur ce point qui est une source fréquente d'erreur de la part des débutants dans l'usage des techniques statistiques.

Les **variables qualitatives ordonnées** résultent soit d'un ordonnancement à priori des catégories, comme lorsque l'on demande dans un questionnaire si les personnes effectuent une pratique très fréquemment, fréquemment, rarement ou jamais, soit par la transformation d'une mesure en catégories en définissant des « tranches » (d'âge, de revenus, etc.). Dans ces variables il

[15] Je présente ici les termes que j'utilise, il en existe d'autres. Par exemple les variables qualitatives peuvent être appelées aussi « catégorielles », ce qui est peut-être plus juste. Mais sous les variations de vocabulaire on retrouve toujours les mêmes trois types.

existe un ordre strict des catégories. Lorsque ces variables sont codées par des nombres, la valeur de ceux-ci est arbitraire, mais leur ordre correspond à l'ordonnancement des catégories, ce qui permet l'usage de certaines méthodes quantitatives (celles fondées sur les rangs et non sur les valeurs par exemple). Il faut être prudent dans l'usage de ces méthodes. En effet si les codages reprennent l'ordre, ils ne disent rien sur les amplitudes, se contentant généralement d'une progression linéaire (1, 2, 3, etc.) alors que parfois une autre serait possible (par exemple 10, 100, 1000, etc.).

Les variables **quantitatives** sont des mesures chiffrées au moyen de nombres entiers ou décimaux : l'âge[16], le salaire annuel ou le nombre d'heures travaillées dans l'année, en sont des exemples. Ici il n'y a pas de code, le nombre est lui-même l'information.

[16] Même si l'âge est une mesure quantitative, il implique souvent comme nous le verrons d'être recodé en catégories construites en fonction de logiques analytiques portant sur les cycles de vie, les générations ou d'autres aspects.

3. Analyser les variables une par une et les recoder lorsque c'est nécessaire

Lorsqu'on a une idée assez précise des sources des données dont on dispose, une bonne façon de « faire connaissance » avec les données est de regarder les variables une par une. On dispose pour cela d'un certain nombre de techniques qui diffèrent selon les variables dont il est question.

Les variables qualitatives

Ici c'est assez simple, on compte simplement les occurrences de chaque catégorie. Je prends à titre d'exemple la variable « condition d'emploi »[17] du fichier salariés - DADS - 2015.

Tableau 1. Les conditions d'emploi des salariés français en 2015

	Fréquence	Pourcentage
non renseigné	9214	0,4
temps complet	1708819	76,1
travail à domicile	28289	1,3
faible temps partiel	22679	1,0
condition d'emploi mixte à dominante temps complet	24747	1,1
temps partiel	438309	19,5
condition d'emploi mixte à dominante temps non complet	14031	0,6
Total	2246088	100,0

Source : Insee, DADS-2015-Fichier salariés au 1/12e
Lecture : pour 9214 salariés de cet échantillon, soit 0,4% du total, les informations ne sont pas disponibles pour cette variable.

Que nous apprend ce recensement des catégories ? D'abord qu'il existe un petit pourcentage de cas pour lesquels l'information n'est pas disponible, ce qui va demander un recodage. Ensuite que deux catégories (temps complet et temps partiel) sont quasiment hégémoniques. D'ailleurs l'Insee suggère aimablement de regrouper les modalités puisque la documentation des variables spécifie : « On se limite généralement à distinguer les postes à temps complets (C) des autres postes salariés (autres modalités regroupées) »[18]. Sauf si l'on s'intéresse particulièrement aux catégories intermédiaires (« faible temps partiel », « condition d'emploi mixte à dominante temps complet », etc.), il semble logique, comme le suggère l'Insee, de regrouper les catégories qui diffèrent de l'emploi à plein temps. Que faire des cas non renseignés ? Deux options sont possibles. Dans la première option, on considère qu'il s'agit probablement de situations particulières distinctes de l'emploi à temps plein ou à temps partiel et on les regroupe avec les situations qui diffèrent de l'emploi à temps complet. Dans la seconde option, on considère qu'il s'agit d'un problème de collecte d'informations et qu'on ne peut pas faire d'hypothèses sur leur répartition dans les diverses catégories et, dans ce cas, il semble logique de recoder la variable en mettant en « valeurs manquantes » (qui ne seront pas pris en compte dans les calculs) les cas qui sont codés dans la modalité « non renseigné »[19]. C'est l'option que j'ai choisie.

[17] La variable CPFD du fichier, recodée pour faire apparaître en clair les significations des codes cryptés de l'Insee (le temps complet est codé « C » par exemple).
[18] https://www.insee.fr/fr/statistiques/3536754#dictionnaire
[19] D'une façon générale, on manque souvent d'informations pour savoir comment les données ont été constituées par des administrations publiques et cela rend d'autant plus nécessaire un travail exploratoire important. Je remercie Jean-Michel Chapoulie de m'avoir suggéré cette remarque.

Tableau 2. Les conditions d'emploi regroupées des salariés français en 2015

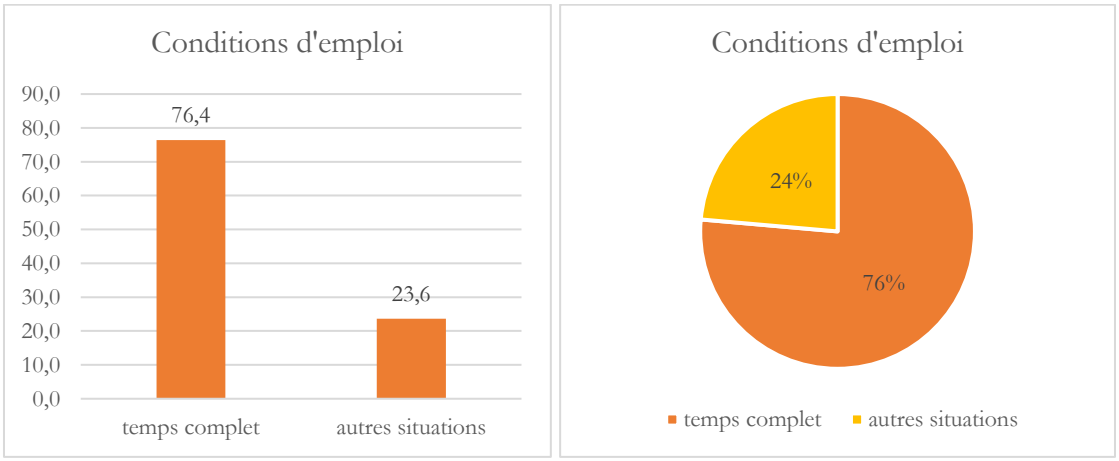
	Fréquence	Pourcentage	Pourcentage valide
temps complet	1708819	76,1	76,4
autres situations	528055	23,5	23,6
total renseigné	2236874	99,6	100,0
valeurs manquantes	9214	0,4	
total des observations	2246088	100,0	

Source : Insee, DADS-2015-Fichier salariés au 1/12e
Lecture : 1708819 salariés de cet échantillon, soit 76,4 % du total renseigné travaillent à temps complet.

Dans ce deuxième tableau apparaissent les valeurs manquantes et deux calculs de pourcentages, en comptant celles-ci ou non. Dans ce cas, les valeurs manquantes sont rares et cela ne change pas grand-chose aux répartitions. Comme il ne reste plus que deux catégories, ce tableau n'est pas forcément nécessaire et l'on pourrait se contenter d'écrire que « 76,4% des salariés sont à temps complet », ou, si l'on ne veut pas prendre de risque relativement aux valeurs manquantes et aux éventuels petits décalages de l'échantillon par rapport à la population totale, écrire que « environ les trois quarts des salariés sont à temps complet ».

On peut construire des représentations graphiques dans lesquelles les surfaces sont proportionnelles aux pourcentages, soit sous la forme de rectangles, soit sous la forme de sections d'un disque[20].

Graphique 1. Deux représentations des conditions d'emploi



Les variables qualitatives ordonnées

Dans l'échantillon de salariés que j'utilise pour les exemples, les âges ne sont pas répartis totalement à l'identique par rapport à la population générale[21]. L'Insee a donc inclus dans le fichier

[20] que les francophones appellent des camemberts et les anglophones des tartes, exemple de l'encastrement des termes techniques dans des ensembles culturels plus larges ... J'ai réalisé ces graphiques avec un tableur distinct du logiciel statistique. Il est souvent commode de travailler avec trois logiciels, l'un pour les analyses statistiques, un tableur pour mettre en formes des tableaux ou créer des graphiques et un éditeur de textes pour stocker les résultats intéressants en leur associant un commentaire.

[21] « L'échantillon est constitué des personnes nées 16 jours donnés chaque année ainsi que des personnes nées un mois donné les années paires. Les effectifs par âge détaillé sont impactés par ce nouveau mode d'échantillonnage, qui

une variable qui regroupe les âges par tranches de 4 ans au moins. Dans le tableau qui suit, j'ai transformé en valeurs manquantes les 77 individus pour lesquels l'âge n'est pas connu.

Tableau 3. Les tranches d'âge des salariés

	effectif
[0;15[ans	40
[15;19[ans	26647
[19;23[ans	148446
[23;27[ans	202713
[27;31[ans	215455
[31;35[ans	216632
[35;39[ans	214313
[39;43[ans	217383
[43;47[ans	227959
[47;51[ans	221230
[51;55[ans	210188
[55;59[ans	182274
[59;63[ans	114992
[63;67[ans	33271
[67;71[ans	10099
[71;+[ans	4369
ensemble	2246011

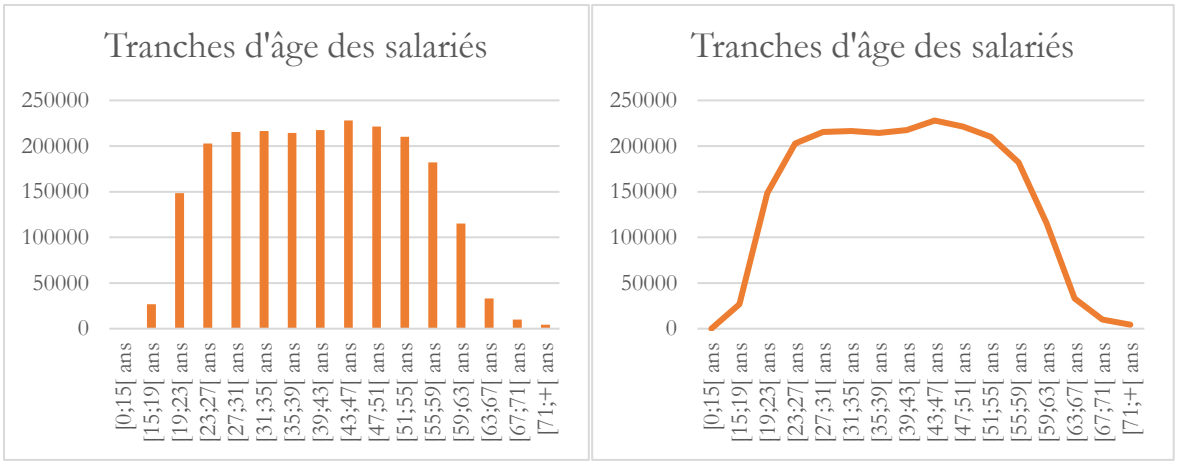
Source : Insee, DADS-2015-Fichier salariés au 1/12e
Lecture : 40 salariés de cet échantillon ont moins de 15 ans.

Le tableau permet de voir qu'à partir de la tranche 23-27 ans, l'effectif est assez stable jusqu'à celle des 51-55 ans, ce qui correspond assez bien à ce que l'on sait du marché du travail et des limites de la « vie active » avec l'entrée progressive des jeunes dans l'emploi et les départs en retraite à partir de 55 ans.

Comme pour les variables qualitatives on peut visualiser les résultats sous la forme de graphiques, mais en privilégiant cette fois-ci ceux qui intègrent l'ordre existant entre les modalités, donc plutôt des diagrammes en bâtons ou des courbes.

surreprésente les salariés nés une année paire. Les utilisateurs des données doivent donc veiller à tenir compte de cette surreprésentation dans l'interprétation des tabulations réalisées : les moyennes ou les répartitions par âge détaillé ne sont pas affectées. En revanche dès lors qu'on regroupe différents âges, les poids différenciés des générations doivent être pris en compte. Notamment pour effectuer des comparaisons par tranche d'âges entre différents millésimes du fichier, il convient de regrouper les données dans des tranches contenant un nombre pair d'âges, pour éviter des fluctuations d'effectifs. » (https://www.insee.fr/fr/statistiques/3536754#documentation).

Graphique 2. Deux représentations des tranches d'âge



Les variables quantitatives

Les variables quantitatives se prêtent à une très grande variété de calculs d'indicateurs. Je ne reprends dans le tableau 4 que quelques-uns qui font partie des plus courants.

Le premier indicateur est la **moyenne** qui est simplement la somme des valeurs divisée par le nombre d'observations. Ici la moyenne des salaires annuels de cet échantillon s'établit à 22124,69 euros. La **variance** est la somme des carrés des écarts à la moyenne divisée par le nombre d'observations (ici somme sur toutes les lignes de (salaire - 22124,69)^2 divisé par 2246088 (nombre de lignes du fichier)). C'est une mesure de la variabilité des valeurs (ici très grande). L'**écart-type** est la racine de la variance. C'est également une mesure de la variabilité des valeurs, qui peut s'intégrer à des calculs sur des distributions de probabilité, sur lesquelles je reviendrai plus loin. Le minimum et le maximum permettent de mesurer l'étendue des valeurs. Ici nous voyons qu'entre les salaires nuls (il y en a) et le plus élevé qui atteint presque 17 millions d'euros, l'étendue est considérable. Les lignes qui suivent (5%, 10%, etc.) correspondent aux valeurs qu'il faut pour atteindre, en partant du minimum, 5% des salariés, 10%, etc. La valeur correspondant à 50% est la **médiane**, ici 19156 euros, qui correspond à une division en deux de la population. Le fait que la médiane soit différente de la moyenne et inférieure à celle-ci est dû à l'existence d'un petit nombre de très hauts salaires qui contribuent fortement à la somme qui entre dans le calcul de la moyenne, alors qu'ils comptent beaucoup moins (un par salaire) dans celui de la médiane. Les valeurs correspondant à 25% et 75% complètent la moyenne pour former les **quartiles**, les valeurs qui divisent les effectifs en quatre parts égales (les déciles les divisent en dix, les centiles en cent, etc.). La dernière valeur, pour 95% montre que seuls 5% des salaires dépassent 48515 euros, ce qui est encore très loin de la valeur la plus élevée.

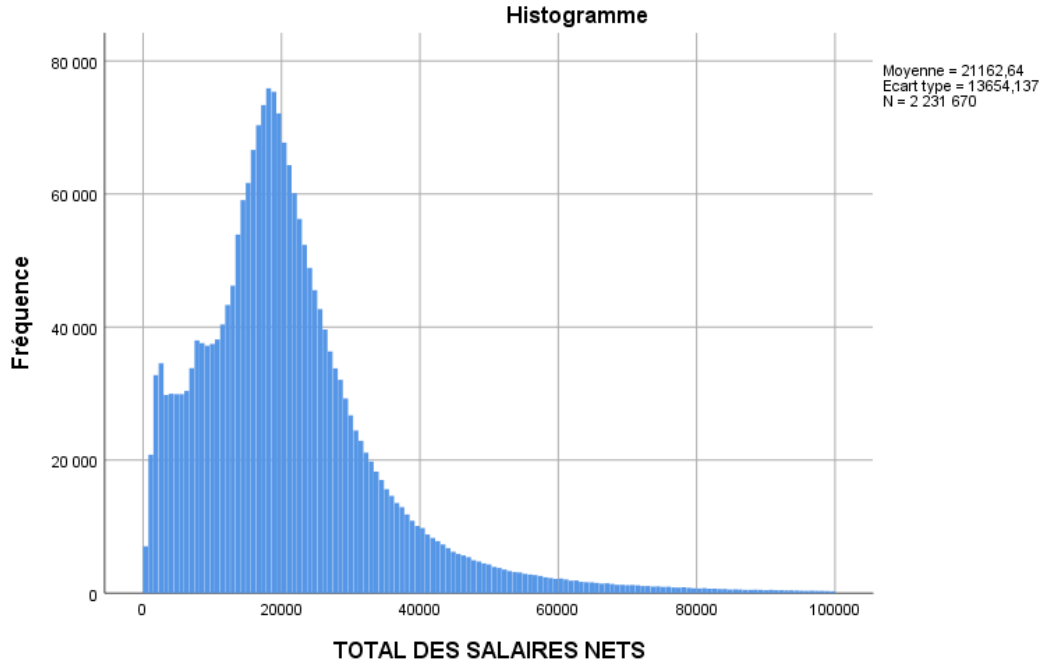
Tableau 4. Indicateurs de la variable portant sur le total des salaires

Moyenne	22124,69
Variance	669539311,556
Ecart type	25875,458
Minimum	0
Maximum	16916142
5%	3390,00
10%	6169,00
25%	12765,00
50% (médiane)	19156,00
75%	26695,00
90%	37772,00
95%	48515,00

Source : Insee, DADS-2015-Fichier salariés au 1/12e
Lecture : le salaire moyen est de 22124,69 euros.

Représenter graphiquement une telle étendue de valeurs est difficile. Dans le graphique qui suit, j'ai exclu les 14 418 valeurs supérieures à 100000 euros. C'est un **histogramme**, c'est-à-dire un ensemble de rectangles dont la surface représente la proportion d'une tranche de valeurs. En général les valeurs sont découpées selon des tranches de même amplitude.

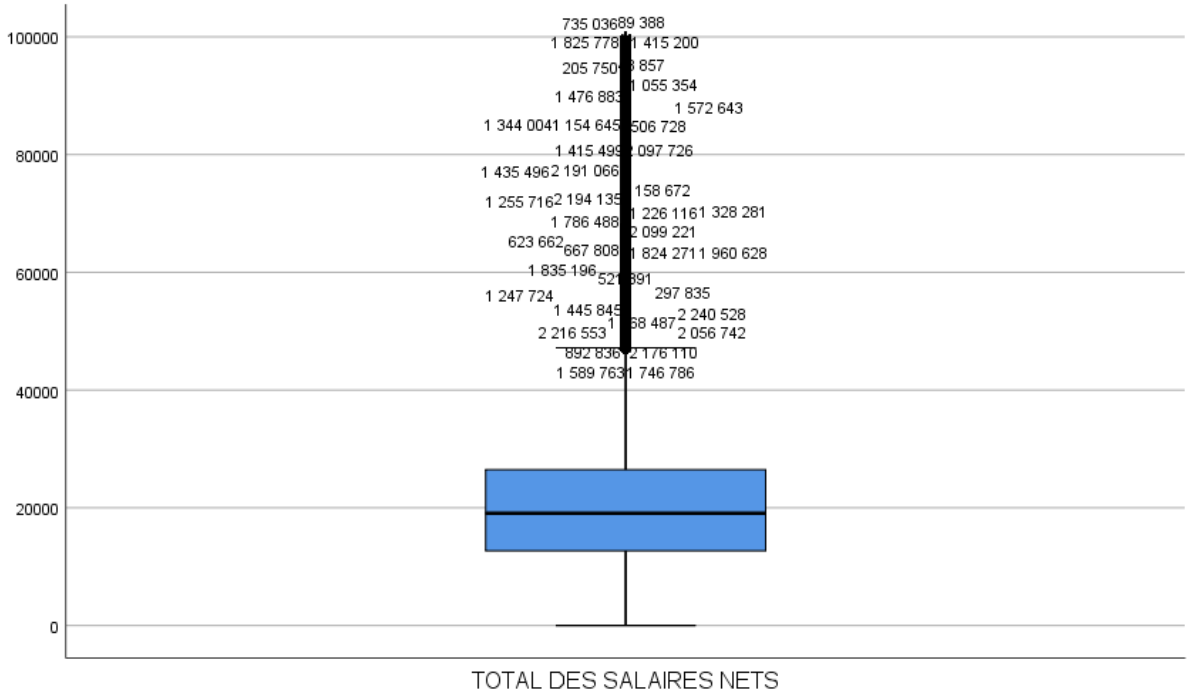
Graphique 3. Histogramme des salaires inférieurs ou égaux à 100000 euros



Ce graphique permet de vérifier que, même en excluant les très hauts salaires, la répartition est déséquilibrée avec beaucoup de salaires modestes ou moyens et de moins en moins de cas lorsque les valeurs s'élèvent, alors que ces valeurs élevées contribuent fortement à la somme totale des salaires.

Une autre représentation graphique de type « **boîte à moustaches** »[22], est construite autour de la médiane et d'une fonction de la distance entre les quartiles. La boîte du milieu présente la médiane (barre horizontale) et les quartiles (limites de la boîte), les « moustaches » étant calculés ne fonction des déciles. Les valeurs extrêmes apparaissant au-delà des moustaches, avec les numéros des lignes concernées.

Graphique 4. Boîte à moustaches des salaires inférieurs ou égaux à 100000 euros



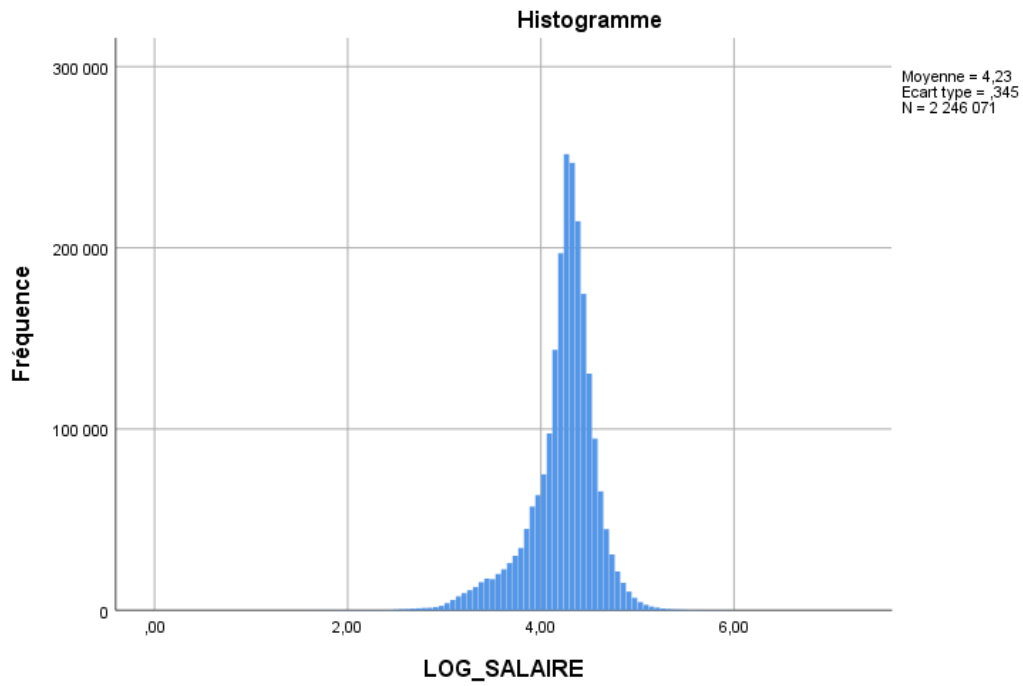
La répartition des salaires ne ressemble pas à une **distribution** « normale » ou « gaussienne » (d'après le nom du mathématicien Carl Friedrich Gauss) qui est symétrique autour de la moyenne et dont 95% des cas entrent dans un intervalle compris entre la moyenne moins deux écart-types et la moyenne plus deux écart-types[23]. Comme d'autres, ce type de distribution peut se calculer au moyen d'une fonction numérique (une exponentielle négative prenant la forme d'une sorte de cloche) dont les paramètres sont la moyenne et l'écart-type. Ce type de distribution correspond au cas où les variations résultent de multiples petits effets indépendants les uns des autres.

Les distributions normales s'observent souvent dans les sciences de la nature ou de la technique (par exemple la taille des animaux dans une population donnée ou le nombre de pièces défectueuses par an dans une ligne de production industrielle) mais sont beaucoup plus rares dans les sciences sociales où il y a de multiples facteurs qui renforcent des avantages acquis (les riches s'enrichissent, les personnes très médiatiques le deviennent encore plus, etc.). C'est le cas des salaires qui ne se répartissent pas de façon équilibrée autour de la moyenne et pour lesquels la moyenne est supérieure à la médiane. La répartition ressemble plutôt à une distribution de type « lognormal », c'est-à-dire que si l'on prend le logarithme[24] des valeurs, on se rapproche d'une distribution de type normal. On peut le vérifier en calculant le logarithme des salaires (en enlevant

[22] Appelées également « boîtes à pattes ».
[23] 1,96 écarts-types pour être plus précis.
[24] Le logarithme décimal est une fonction qui associe à un nombre la puissance de 10 qui lui correspond. Le logarithme de 10 est 1, celui de 100 est 2 et celui d'un nombre x positif quelconque la solution de l'équation $x = 10^{\log(x)}$. La logarithme népérien, préféré des mathématiciens, remplace 10 par le nombre e, qui vaut approximativement 2,71828. Pourquoi ? pour des raisons … mathématiques (le logarithme népérien de e est égal à 1).

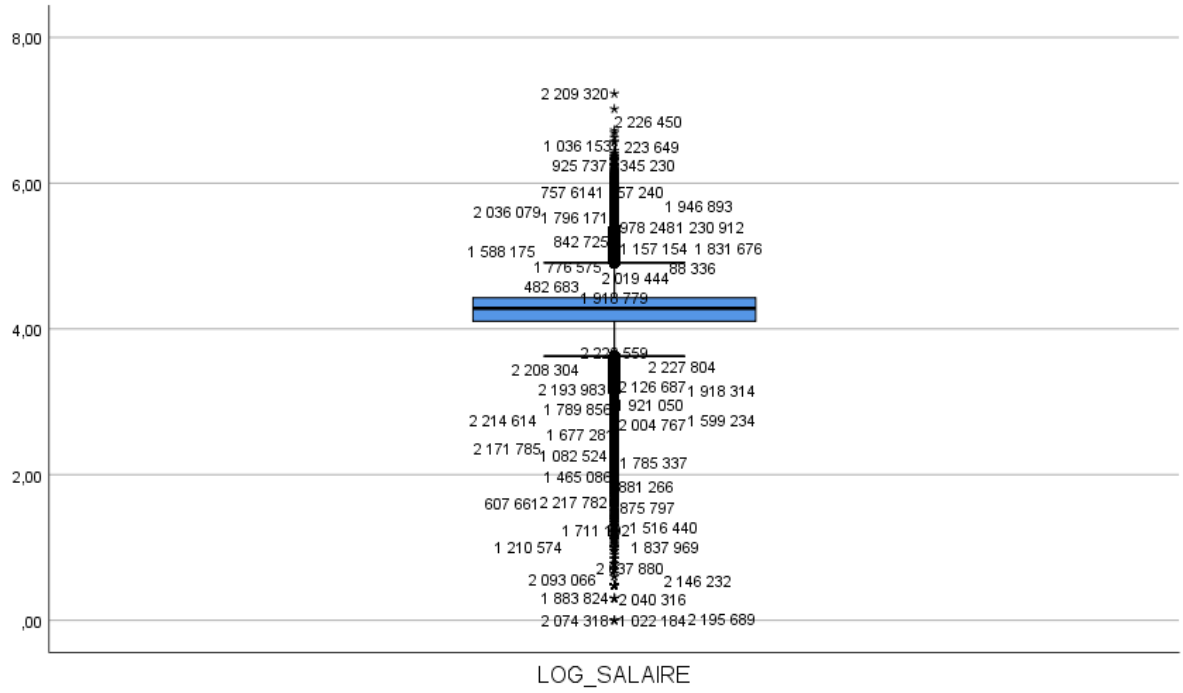
les 17 salaires nuls parce qu'on ne peut pas calculer un logarithme pour une valeur nulle), et en réintégrant les très gros salaires qui ne posent plus autant de problèmes à présent.

Graphique 5. Histogramme des logarithmes décimaux des salaires supérieurs à 0



La courbe est équilibrée, la moyenne (4,23) est proche de la médiane (4,28), on se rapproche d'une distribution normale. Il est assez fréquent de prendre le logarithme de variables qui sont réparties de cette façon afin d'utiliser des modèles qui présupposent une distribution de type normal. On retrouve cela dans la boîte à moustaches.

Graphique 6. Boîte à moustaches des salaires supérieurs à zéro



Remarquons que cette opération de « normalisation » n'est pas neutre : elle atténue les différences et masque l'existence de salaires très élevés par contraste à l'écrasante majorité ce ceux-ci. C'est pourquoi des auteurs de plus en plus nombreux préfèrent calculer des fractiles (médianes, déciles, centiles, etc.) et la part de la somme totale cumulée par les x% les mieux rémunérés. Pour le fichier que j'utilise ici, 5% des personnes dépassent 48515 euros par an. Ceux qui dépassent ce seuil cumulent 17,5% du total des salaires versés[25].

[25] Je remercie Alain Degenne de m'avoir suggéré cette remarque.

4. Croiser les variables

Une fois qu'on a une bonne idée des variables dont on dispose, qu'on a procédé à un premier recodage pour coder en valeurs manquantes les cas où l'information fait défaut, pour regrouper des modalités à faible effectif ou pour constituer des catégories plus robustes au regard de la problématique de l'étude, on peut commencer à explorer les corrélations. C'est la partie des analyses la plus riche et productrice de résultats. Qu'est-ce qu'une corrélation statistique ? C'est l'existence d'un lien dans les variations : si on modifie la valeur d'une variable, l'autre a tendance à changer également. Comme ce lien est statistique, il n'est pas systématique (comme le serait par exemple la relation entre le poids d'une bouteille de vin et le volume qui reste à boire[26]) et admet des contre-exemples. Cela implique que l'on cherche en général à faire la « part du hasard » c'est-à-dire de variations aléatoires provoquées par de multiples petites causes auxquelles on ne s'intéresse pas. Toutes les techniques de recherches de corrélations intègrent une forme de comparaison avec une situation considérée comme aléatoire. Avant de présenter les techniques les plus courantes, autant redire en criant (métaphoriquement) très fort qu'**une corrélation n'est pas une relation de cause à effet**[27]. Deux variables peuvent avoir un lien statistique sans que l'une mesure un phénomène qui a un effet sur l'autre. Il existe une corrélation historique (si l'on mesure chaque année) entre la vente de disques vinyles et le cours de l'or[28] mais l'écoute de musique n'enrichit pas les mélomanes. L'établissement d'un lien causal (y compris au sens faible de « influe sur », « favorise », etc.) demande à combiner des corrélations avec des hypothèses et des arguments logiques.

Les techniques ne sont pas les mêmes selon les types de variables. Trois cas de figure se présentent : des variables uniquement qualitatives, uniquement quantitatives, ou un mélange des deux.

Croiser des variables qualitatives

La technique la plus simple est de faire un **tableau croisé** ou tableau de contingence. Prenons l'exemple d'un croisement entre le sexe et la catégorie professionnelle regroupée[29]. Commençons par un tableau de contingence simple, celui des effectifs.

Tableau 5. Répartition des hommes et des femmes pour chaque catégorie professionnelle (effectifs)

	homme	femme	total
salariés agricoles, artisans, commerçants, chefs d'entreprises salariés	12865	3438	16303
cadres	217594	152035	369629
professions intermédiaires	206733	239450	446183
employés	246733	574376	821109
ouvriers	465639	124187	589826
ensemble	1149564	1093486	2243050

Source : Insee, DADS-2015-Fichier salariés au 1/12e

[26] La précision des mesures faites par le consommateur diminuant lorsque le volume diminue, ce qui est une autre corrélation, d'une nature un peu différente.

[27] Pour une discussion précise de ce point dans l'usage des régressions, voir David A. Freedman, 1991, « Statistical Models and Shoe Leather », *Sociological Methodology*, Vol. 21 (1991), pp. 291-31.

[28] https://www.courrierinternational.com/grand-format/statistiques-les-correlations-de-labsurde

[29] J'ai procédé à des regroupements à partir de la variable « CS » du fichier.

Lecture : parmi les salariés agricoles, artisans, commerçants, chefs de petites entreprises salariés, 12865 sont des hommes, 3438 sont des femmes[30].

C'est clair non ? Euh … non. Vous avez raison, ce n'est pas du tout clair, ça ne sert à rien, sinon à vérifier qu'il n'y a pas de très petits effectifs dans certaines cases, ce qui serait peu probable dans ce corpus de données et pour des variables aussi synthétiques (avec peu de modalités). Que faut-il faire ? Des pourcentages bien sûr. En ligne ou en colonne ? Ici les deux auraient du sens, mais il semble logique de privilégier la proportion d'hommes de femmes dans chaque catégorie donc les pourcentages lignes[31].

Tableau 6. Répartition des hommes et des femmes pour chaque catégorie professionnelle (pourcentages lignes)

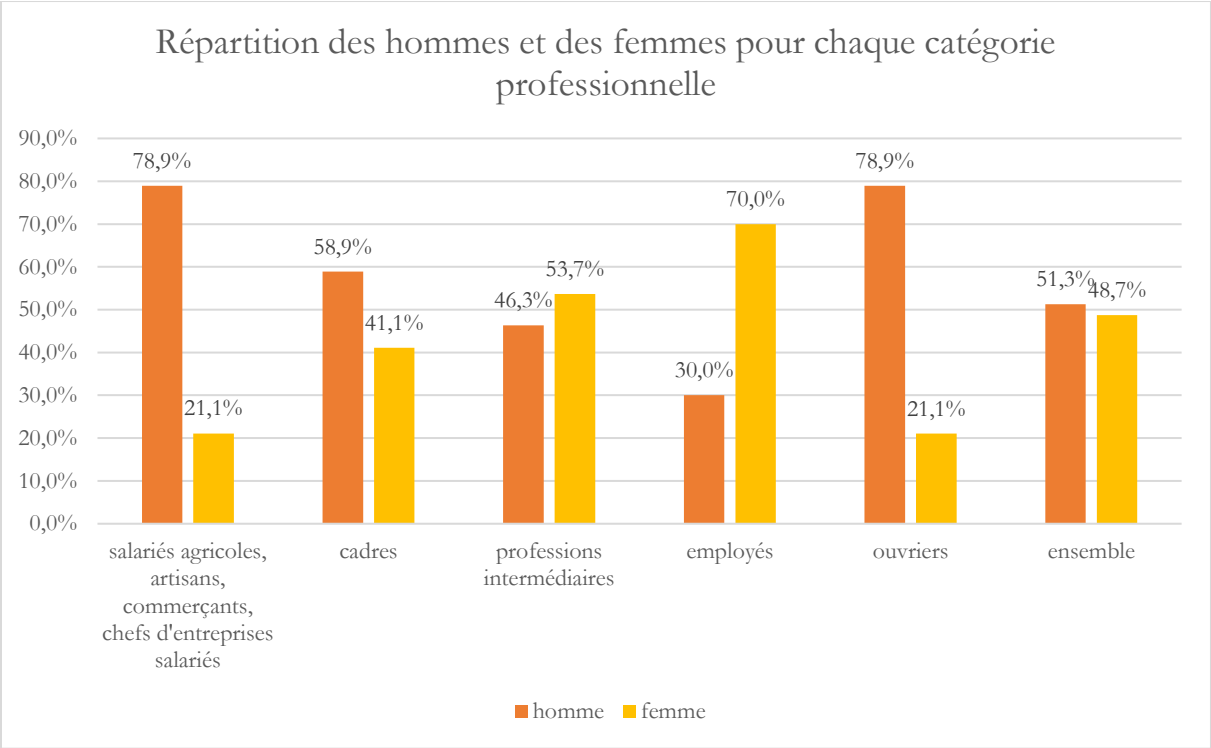
	homme	femme	total
salariés agricoles, artisans, commerçants, chefs d'entreprises salariés	78,9%	21,1%	100,0%
cadres	58,9%	41,1%	100,0%
professions intermédiaires	46,3%	53,7%	100,0%
employés	30,0%	70,0%	100,0%
ouvriers	78,9%	21,1%	100,0%
ensemble	51,3%	48,7%	100,0%

Source : Insee, DADS-2015-Fichier salariés au 1/12e
Lecture : parmi les salariés agricoles, artisans, commerçants, chefs de petites entreprises salariés, 78,9% sont des hommes, 21,1% sont des femmes.

C'est plus clair non ? Oui c'est très clair, certaines catégories comprennent nettement plus d'hommes (les salariés agricoles, artisans, commerçants, chefs de petites entreprises salariés ou les ouvriers), d'autres nettement plus de femmes (les employés). Tiens, puisque le résultat semble intéressant, pourquoi ne pas le représenter sous la forme d'un petit graphique ?

[30] Dans le fichier global, pour 3038 personnes, le sexe n'est pas renseigné.
[31] On peut bien sûr calculer les deux à la fois (et d'autres indicateurs comme la proportion représentée par une case par rapport à la population totale), mais dans le travail d'exploration il vaut mieux faires ces calculs séparément en testant des idées et en créant des tableaux faciles à lire.

Graphique 7. Répartition des hommes et des femmes pour chaque catégorie professionnelle



Lecture : dans la première catégorie, 78,9% des salariés sont des hommes.

Ces différences sont énormes, elles ont peu de chances d'être dues au hasard, même si c'est peut-être un peu moins net pour les cadres et les professions intermédiaires que pour les autres catégories.

Pour se faire une idée plus précise de **l'écart entre ce tableau et un tableau dans lequel les différences entre les modalités seraient aléatoires**, on recourt habituellement à une procédure qui consiste à calculer des **effectifs théoriques** dans chaque case. Le raisonnement est simple. S'il y a 48,7% de femmes dans la population des salariés (ayant déclaré leur sexe), alors on devrait retrouver cette proportion dans toutes les catégories sociales. Comme on peut le vérifier dans le tableau 5 il y a au total 16303 salariés agricoles, artisans, commerçants et chefs de petites entreprises salariés. Il devrait donc y avoir 48,7% de 16303 = 7939,561 femmes dans cette catégorie. On se rappelle que 48,7% est le résultat de 1093486 (nombre de femmes) divisé par 2243050 (total des salariés pour ce tableau). On peut donc éviter les erreurs d'arrondi (48,7 est en fait 48,749961) en faisant le calcul 1093486 (nombre de femmes) multiplié par 16303 (nombre de salariés de cette catégorie à l'intitulé très long) divisé par 2243050 (total des salariés pour lesquels on dispose des informations nécessaires). Le résultat est 7947,7, ce qui montre au passage que sur de gros chiffres comme ceux-ci les erreurs d'arrondi ont des effets non négligeables. Comment comparer pour cette case l'effectif observé (3438) à l'effectif théorique (7947,7) ? On cherche à savoir si la différence entre les deux est due à des variations aléatoires, c'est-à-dire **produites par des petites causes indépendantes les unes des autres**, ce qui produit en général une distribution de type normal[32]. Des mathématiciens se sont fatigués à démontrer que si les variations aléatoires se répartissent selon une distribution normale de moyenne nulle et de variance 1, alors la différence entre l'effectif

[32] C'est peut-être l'occasion de rappeler que l'usage des lois normales n'est pas nécessairement liée à une logique inférentielle. Si elles sont souvent utilisées pour modéliser les écarts entre les mesures effectuées sur un échantillon et celles qui seraient faites sur la population générale, cela n'est pas leur seul usage. Fondamentalement elles permettent de représenter une série de causes indépendantes d'amplitude équivalente.

observé et l'effectif théorique divisée par la racine carrée de l'effectif théorique ((eff.obs – eff.th)/ √ eff.th) se distribue aussi de cette façon. Cela signifie qu'elle n'a que 5% de chances de se situer en deçà de -2 ou au-delà de +2 (1,96 en réalité). C'est ce qui s'appelle le **résidu standardisé** dans la plupart des logiciels et parfois aussi **chi2 partiel**. Calculons les résidus standardisés du tableau précédent.

Tableau 7. Répartition des hommes et des femmes pour chaque catégorie professionnelle (résidus standardisés)

	homme	femme
salariés agricoles, artisans, commerçants, chefs d'entreprises salariés	49,3	-50,6
cadres	64,7	-66,3
professions intermédiaires	-45,9	47,0
employés	-268,4	275,2
ouvriers	297,1	-304,6

Source : Insee, DADS-2015-Fichier salariés au 1/12e

Lecture : le résidu standardisé de la case des hommes salariés agricoles, artisans, commerçants, chefs de petites entreprises salariés est de 49,3.

Les résidus sont tous très élevés, on peut considérer les différences comme significatives. Il est logique que les résidus soient élevés lorsque l'échantillon est important. En effet, il suffit de revenir à la formule du calcul des résidus pour vérifier que si l'échantillon est multiplié par 100 avec les mêmes proportions des catégories, la différence entre l'effectif observé et l'effectif théorique est aussi multipliée par 100 alors que le dénominateur, la racine de l'effectif théorique, est multiplié seulement par 10. Soit on trouve que cet indicateur est trop sensible et on cherche à le corriger par diverses formules, soit on considère qu'un échantillon plus important est plus précis et qu'il permet de mieux détecter les corrélations. Je ne vous inflige pas le tableau mais, dans ce cas, si l'on remplace le sexe par le fait que le salaire soit pair ou impair, ce que j'ai considéré plus haut comme purement aléatoire, alors les résidus ne sont pas significatifs, ce qui est rassurant. Pour ma part je n'utilise pas de correction.

Les résidus standardisés sont un outil très efficace d'exploration, pour repérer des différences importantes et donc des corrélations[33]. Ce n'est pas ce qui est le plus enseigné dans les cours de statistiques pour les sciences sociales, où l'on insiste en général plutôt sur le **test du chi2 (ou Khi2)**.

Ce test a pour objectif de vérifier si l'ensemble du tableau des effectifs observés s'écarte significativement de celui des effectifs théoriques. Pour cela on utilise un autre résultat mathématique, le théorème du chi2, qui dit que la somme des carrés de ce que j'ai appelé les résidus standardisés converge vers une distribution dite du chi2 dont on sait calculer le seuil en dessous duquel se situent 95% des effectifs. Ici le chi2 est énorme (346697,9). On ne prend pas beaucoup de risques à dire qu'il y a des différences entre les catégories professionnelles en ce qui concerne la répartition entre les hommes et les femmes.

Mais ces différences ne sont-elles pas liées à d'autres variables, par exemple le fait que l'emploi soit à temps complet ou partiel ? En effet, les femmes sont beaucoup plus nombreuses dans les emplois à temps partiels ou hybrides (69,3%) que dans ceux à temps complet (42,5%). Et bien, on ne se décourage pas et on se lance dans un tri à trois entrées.

[33] Je ne présente pas d'exemple de cette technique ici parce que le fichier utilisé jusque-là ne s'y prête pas, mais il est possible de faire usage des résidus standardisés (ou une autre statistique équivalente comme celle du test binomial) pour repérer les corrélations les plus importantes entre une caractéristique sociale correspondant à une modalité d'une variable qualitative (par exemple « être une femme ») et toutes les autres modalités figurant dans les données. Il suffit de transformer toutes les modalités des variables qualitatives en variables indicatrices (« présent » ; « absent ») et de calculer tous les résidus entre la modalité que l'on examine et toutes les autres, puis de les ordonner par importance. C'est utile pour explorer des corpus de données comportant beaucoup de variables qualitatives.

Tableau 8. Répartition des hommes et des femmes pour chaque catégorie professionnelle par catégories d'emploi (pourcentages lignes).

		homme	femme	total
temps complet	salariés agricoles, artisans, commerçants, chefs d'entreprises salariés	80,4%	19,6%	100,0%
	cadres	61,1%	38,9%	100,0%
	professions intermédiaires	51,3%	48,7%	100,0%
	employés	36,2%	63,8%	100,0%
	ouvriers	83,9%	16,1%	100,0%
	ensemble	57,5%	42,5%	100,0%
autres situations	salariés agricoles, artisans, commerçants, chefs de petites entreprises salariés	74,3%	25,7%	100,0%
	cadres	47,3%	52,7%	100,0%
	professions intermédiaires	25,3%	74,7%	100,0%
	employés	17,1%	82,9%	100,0%
	ouvriers	56,6%	43,4%	100,0%
	ensemble	30,7%	69,3%	100,0%
Total	salariés agricoles, artisans, commerçants, chefs de petites entreprises salariés	78,9%	21,1%	100,0%
	cadres	58,6%	41,4%	100,0%
	professions intermédiaires	46,3%	53,7%	100,0%
	employés	30,0%	70,0%	100,0%
	ouvriers	78,9%	21,1%	100,0%
	ensemble	51,2%	48,8%	100,0%

Source : Insee, DADS-2015-Fichier salariés au 1/12e

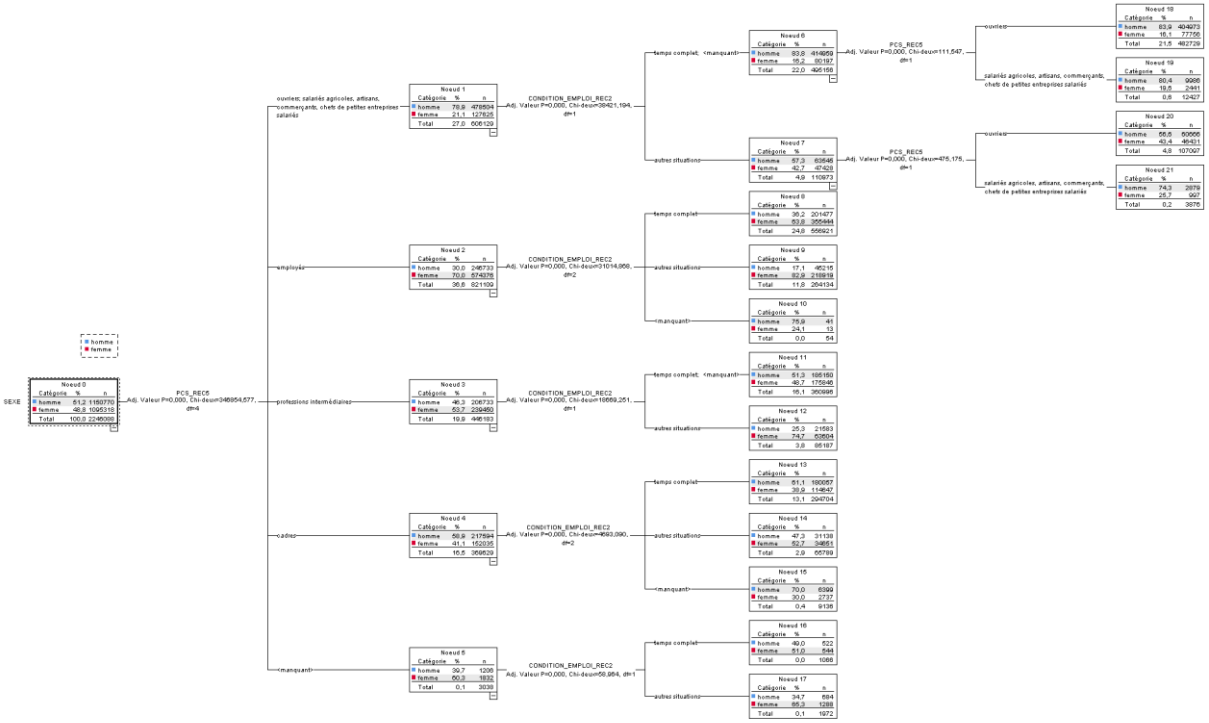
Lecture : La proportion d'hommes parmi les salariés agricoles, artisans, commerçants, chefs de petites entreprises salariés à temps complet est de 80,4%.

La différence se maintient (je ne les ai pas mis mais tous les résidus sont importants, le chi2 de chaque sous-tableau aussi évidemment). Et l'on pourrait refaire l'exercice avec d'autres variables en remplaçant les conditions d'emploi par les tranches d'âge par exemple, ou même en effectuant un tri à quatre entrées en calculant les ratios hommes / femmes par condition d'emploi et tranches d'âge.

Il existe une autre technique, utile pour explorer des corrélations au sein d'un vaste ensemble de variables (ce qui n'est pas vraiment le cas du fichier des salariés) en cherchant à maximiser les différences sur une variable précise. Il s'agit des **arbres de décision** (parfois appelés **segmentations**) A titre d'exemple, je reprends les variables que je viens d'utiliser. Nous cherchons des combinaisons de variables (ici, pour éviter que le graphique soit trop compliqué, j'en ai pris deux seulement, la catégorie professionnelle et la condition d'emploi) qui aboutissent à des ensembles comportant beaucoup plus d'hommes que de femmes ou l'inverse. La procédure consiste à chercher la variable la plus corrélée avec le sexe (au sens du test du chi2), à regrouper ses modalités de sorte à maximiser la corrélation puis à refaire fonctionner la procédure sur chaque sous-ensemble ainsi constitué. Le résultat prend la forme d'un arbre comme celui du graphique suivant. Il apparaît que la catégorie dans laquelle le ratio hommes / femmes est le plus en faveur des hommes est constituée par les ouvriers à temps complet, où ce ratio atteint 83,9%[34].

[34] Je me suis limité à deux variables pour cet exemple mais le résultat est tout de même mal adapté à la mise en page. Le rectangle du haut à droite, qui exprime cette proportion élevée d'hommes chez les ouvriers à temps complet est la dernière étape d'une décomposition d'abord sur des grandes catégories de salariés, regroupés par l'algorithme pour maximiser la corrélation, puis par type de condition d'emploi, puis à nouveau par catégorie professionnelle.

Graphique 8. Arbre de décision sur le ratio entre hommes et femmes selon la catégorie professionnelle et la condition d'emploi.



Deux variables quantitatives

Comment faire lorsque l'on a deux variables quantitatives ? Une solution est de les découper en tranches et de les traiter comme des variables qualitatives selon les techniques présentées dans la section précédente. Les fichiers de données en sciences sociales comportant très souvent beaucoup plus de variables qualitatives que quantitatives, cela peut simplifier les analyses.

Mais on dispose évidemment de nombreuses techniques dédiées aux variables quantitatives. L'une d'entre elle est le **coefficient de corrélation linéaire**. Il est fondé sur la covariance, c'est la dire la somme divisée par le nombre d'observations des produits terme à terme des écarts à la moyenne des valeurs des deux variables considérées. Par exemple dans notre fichier de salariés la covariance entre l'âge et le salaire annuel est la division par le nombre de lignes de la somme sur toutes les lignes des produits (âge – âge moyen) x (salaire – salaire moyen). Plus cette somme a une valeur élevée, plus les deux variables varient de façon conjointe. Si vous en doutez, examinez la distance euclidienne entre les deux vecteurs correspondant aux variables : d(âge, salaire) = racine carrée de somme pour tous les individus i de (âge pour l'individu i moins salaire pour l'individu i au carré). En relisant vos vieux cours de mathématiques, vous vous rappellerez que $(a - b)^2 = a^2 + b^2 - 2\,ab$, et que donc plus ab est grand, plus $(a-b)^2$ est petit. La covariance est difficile à comparer lorsque l'on change les variables et les échantillons, aussi utilise-t-on une autre formule qui au lieu de diviser par le nombre de lignes, divise par le nombre de lignes moins 1 multiplié par les écarts types respectifs des deux variables considérées. Ce coefficient varie entre – 1 (les deux variables varient en sens inverse) et +1 (les deux variables varient de la même façon) en passant par 0 (pas de corrélation). Il détecte les relations linéaires, c'est-à-dire les cas où les deux variables entretiennent une relation linéaire (leurs valeurs progressent ou régressent ensemble régulièrement). Pour l'âge et le salaire il est de 0,179. Le **coefficient de corrélation linéaire** ne fonctionne bien que si les deux variables sont gaussiennes et nous avons vu que ce n'est pas le cas

des salaires. Pour se rapprocher d'une situation gaussienne, nous pouvons prendre le logarithme du salaire. Le coefficient est alors de 0,291. En général les logiciels proposent des seuils de significativité des coefficients[35]. Ici ils sont considérés dans les deux cas comme significativement différents de zéro.

Une autre technique utilisable est la **régression linéaire**. Il existe des techniques équivalentes pour les variables qualitatives (les régressions logistiques) mais je ne les ai pas présentées pour le moment parce qu'elles sont un peu compliquées. Il suffit de savoir qu'elles ont pour objectif de produire des résultats similaires à ceux de la régression linéaire et que les logiciels procurent des indicateurs qui se rapprochent de ceux de la régression linéaire. Il est donc utile de présenter celle-ci. L'idée est très simple. Si Y est la variable à expliquer et X la variable explicative, il s'agit de trouver des coefficients a et b tels que l'on puisse écrire pour chaque unité statistique $Y = aX + b + \varepsilon$ avec ε le plus petit possible. Le lecteur ayant quelques souvenirs des cours de mathématiques du collège se rappellera que $Y = aX + b$ est l'équation d'une droite. Donc l'idée est de trouver une droite qui résume au mieux les liens entre les deux variables. ε est la différence entre le modèle aX+b (que l'on appelle également valeur prédite) et la valeur observée. C'est le « résidu » brut. Si le modèle s'ajuste bien, les valeurs résiduelles sont censées se répartir selon une distribution normale. C'est rarement le cas dans la pratique. Mais la régression peut être utile pour estimer la part d'explication que l'on peut attribuer à X dans les variations de Y. Pour cela il existe un indicateur $\mathbf{R^2}$ qui représente la proportion de la variance de la variable à expliquer qui s'explique par le modèle. Il varie entre 0 et 1 et se présente comme un pourcentage. Il se calcule en faisant 1 moins la somme pour chaque unité du carré de la différence entre la valeur observée de Y et celle prédite par le modèle, le tout divisé par la somme des différences entre les valeurs observées et la moyenne de la variable Y. Cette formule est compliquée ? Oui mais elle est égale au carré du coefficient de corrélation linéaire (d'où sa désignation qui comporte une élévation au carré). La part de l'âge dans l'explication des variations de salaire est donc $(0{,}179)^2 = 3{,}2\%$. Ce n'est pas beaucoup. Pour le logarithme du salaire c'est un peu mieux, avec 8,5%. Cet exemple permet de rappeler **qu'une variable peut être corrélée significativement avec une autre sans expliquer une part importante de ses variations**. S'il suffit d'élever au carré le coefficient de corrélation linéaire, pourquoi raconter cette histoire de régression ? Pour deux raisons. D'abord, il arrive que le modèle prédise très bien la variable à expliquer[36], ce qui permet d'écarter d'autres facteurs. Ensuite parce que ce cas très simple sert de base à toutes les modélisations (linéaires à plusieurs variables, logistiques, etc.) qui produisent toutes des équations, des résidus et des équivalents du R^2, moins simples à calculer, mais fonctionnant sur le même principe. Dans le modèle linéaire général, qui est donc une généralisation de la régression, il y a toujours une variable à expliquer Y, mais un ensemble de variables explicatives X_1, X_2, etc., la fonction qui constitue le modèle n'est plus nécessairement une droite et la distribution de référence peut ne pas être normale pour peu qu'elle fasse partie d'une famille dite « exponentielle » (qui comprend les distributions de Poisson, binomiale, etc.). Le modèle général s'écrit $Y = f(X_1, X_2, \text{etc.}) + b + \varepsilon$, ε se répartissant selon l'une des distribution de la famille exponentielle. Il n'entre pas dans le projet de ce texte de développer tout cela, il existe des ouvrages ou des cours sur ces modèles, mais il est utile de présenter la logique de base, même pour un travail exploratoire.

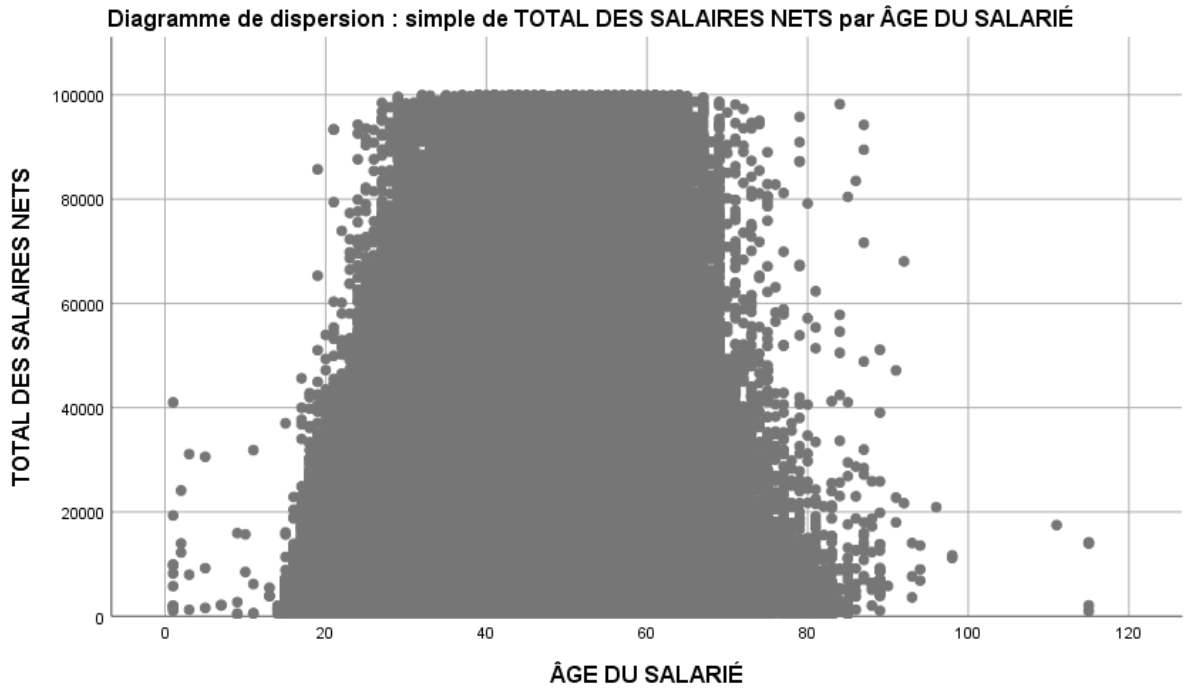
On peut aussi représenter graphiquement le lien entre les deux variables.

---

[35] Pour ces tests, on considère que la distribution est normale, de moyenne nulle et de variance $1-R^2$, R étant le coefficient de corrélation linéaire.

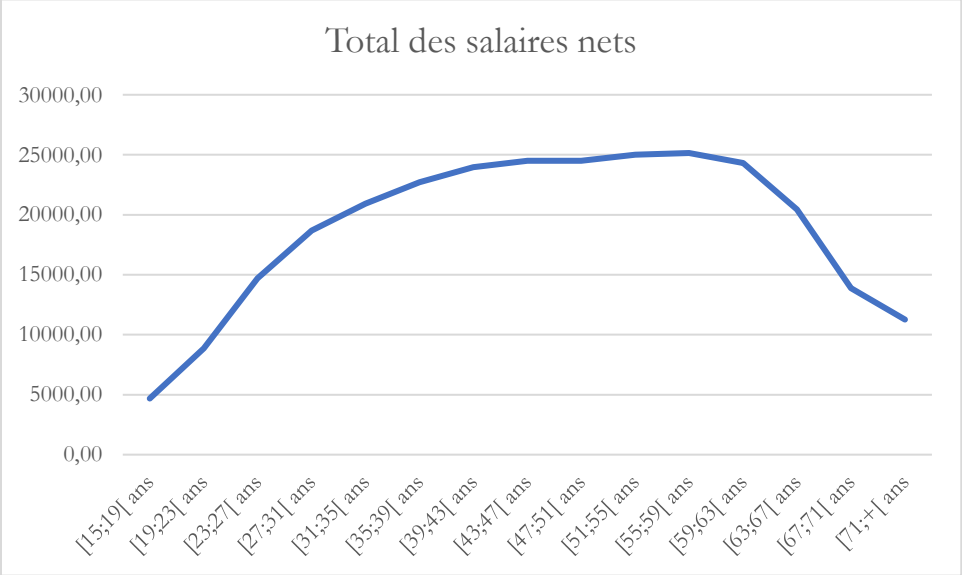
[36] C'est le cas pour la corrélation entre le nombre de chercheurs d'une ville et le nombre de publications comportant des auteurs de cette ville (Michel Grossetti, Marion Maisonobe, Laurent Jégou, Béatrice Milard, Guillaume Cabanac, « L'organisation spatiale de la recherche française à travers les publications savantes : régularité des tendances de long terme et désordre des politiques publiques (1999-2017) », 2020. hal-02627291)

Graphique 9. Digramme de dispersion des salaires selon l'âge des salariés



Pas très clair … Plutôt que l'âge et le salaire de chaque salarié, prenons les tranches d'âge et la moyenne des salaires.

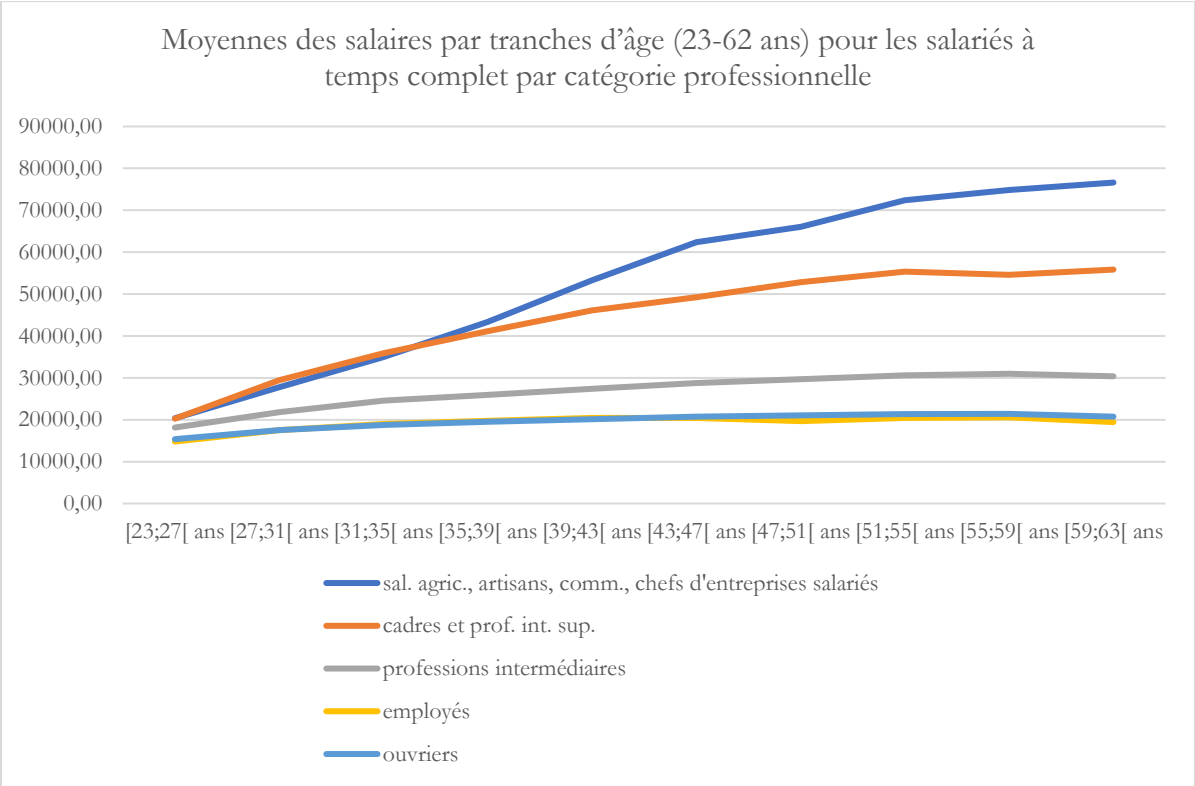
Graphique 10. Moyennes des salaires par tranches d'âge



On y voit un peu plus clair. Les salaires tendent à progresser avec l'âge (j'ai laissé de côté la première catégorie des moins de 15 ans, avec un petit effectif et une valeur extrême qui est probablement une erreur) et à régresser après l'âge du début des départs à la retraite. Mais cela reste confus : il y a des temps complets et partiels, toutes les catégories professionnelles. Pour mieux comprendre, nous pouvons examiner un sous ensemble de cas constitué des tranches d'âge entre

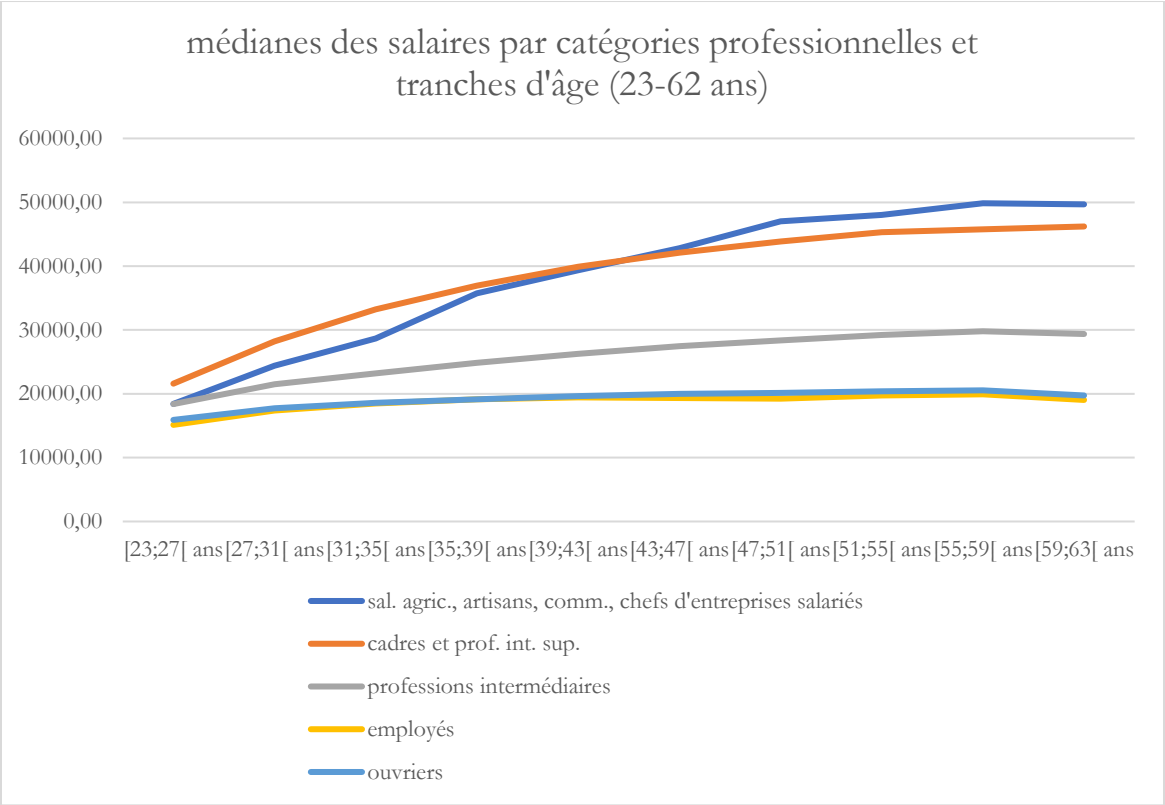
23 et 62 ans et des salariés à temps complet et distinguer les catégories professionnelles (en en restant aux 5 très grossières définies plus haut).

Graphique 11. Moyennes des salaires par tranches d'âge (23-62 ans) pour les salariés à temps complet par catégorie professionnelle



La première catégorie est hétérogène et comprend des valeurs extrêmes rares (par exemple elle intègre les chefs de grandes entreprises salariés, peu nombreux mais dont les salaires sont très élevés). La moyenne est sensible aux valeurs extrêmes, ce n'est pas un indicateur **robuste**. La **médiane** est plus robuste. Voici ce que donne le graphique lorsque l'on prend la médiane plutôt que la moyenne pour l'ensemble des salaires.

Graphique 12. Médianes par catégorie professionnelle des salaires par tranches d'âge (23-62 ans) pour les salariés à temps complet



Les catégories des cadres, professions intermédiaires, employés et ouvriers n'ont pas beaucoup bougé, alors que les valeurs se sont fortement modifiées pour la première catégorie, ce qui confirme que ces moyennes étaient tirées vers le haut pas nombre restreint de très gros salaires. Ce graphique confirme à la fois l'accroissement des salaires au cours de la carrière et les différences entre catégories professionnelles, ce qui introduit la question des croisements entre variables quantitatives et qualitatives.

Croiser des variables quantitatives et qualitatives

Comparer des moyennes ou des médianes est la meilleure façon de procéder pour comprendre comment une variable quantitative est liée à une variable qualitative. Nous venons de voir que les catégories professionnelles correspondent à des niveaux différents de salaire moyen ou médian. Pour changer, nous allons à présent nous intéresser aux rapports entre le salaire et le sexe. Nous avons vu que les femmes sont plus souvent à temps partiel. Nous allons donc conserver les critères des graphiques précédents : des salariés à temps complet entre 23 et 62 ans.

Tableau 9 : médianes et moyennes des salaires des hommes et des femmes (salariés à temps complet entre 23 et 62 ans)

	Médiane	Moyenne
homme	22234,00	27140,28
femme	19954,00	22322,41
Total	21215,00	25085,22

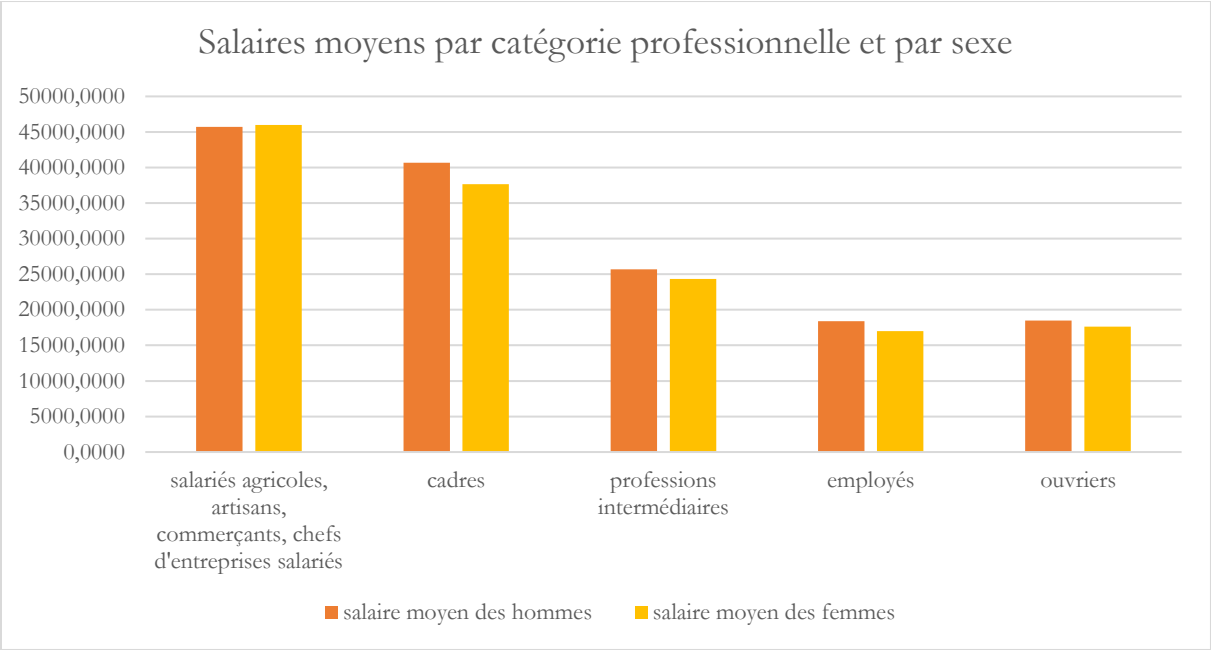
Comme le résidu standardisé ou le chi2 pour les variables qualitatives, ou le coefficient de corrélation linéaire pour les variables quantitatives, il existe une technique pour évaluer l'écart entre les valeurs observées et une situation aléatoire. C'est **l'analyse de variance**, qui fait partie de la vaste famille du modèle linéaire généralisé. Comme la plupart des techniques de cette famille, l'option de base suppose une distribution normale. Nous allons donc prendre le logarithme des salaires pour nous rapprocher de cette situation et calculer (enfin, demander au logiciel de calculer[37]) le test de Fisher (du nom de Ronald Aylmer Fisher, un biologiste et statisticien qui a proposé cette méthode de même que bien d'autres qui font partie à présent du corpus des méthodes statistiques standard). Je ne vous inflige pas le calcul du test mais celui-ci conclut sans équivoque que les moyennes des logarithmes des salaires sont différentes. La méthode utilisée dans le logiciel que j'ai utilisé estime que probabilité de se tromper en disant que les moyennes sont différentes (le **risque de première espèce**) est nulle. C'est l'occasion de dire que tous les tests comportent cette valeur, généralement notée sig. (pour « significativité »), ou p, ou encore alpha, qui donne une indication sur la force de la corrélation et sa « consistance » (le fait que l'on s'écarte d'une situation dans laquelle la différence serait due au hasard).

Que l'on prenne la médiane ou la moyenne, la tendance est la même : les salaires des femmes sont plus faibles. Mais peut-être est-ce dû aux types de profession ? Une analyse de variance à deux facteurs sur le logarithme des salaires conclut que la différence entre hommes et femmes reste significative, mais des calculs de R^2, avec l'une ou l'autre des variables ou les deux, montrent que la profession « explique » plus fortement les variations de salaire. Toujours sur le logarithme, le R^2 vaut 1,6% pour le sexe seul, 25,4% pour la catégorie professionnelle en cinq grands types, et 26,7% lorsque l'on prend les deux variables[38]. Les R^2 ne s'ajoutent pas parce que les variables explicatives peuvent être corrélées entre elles (donc que leurs pouvoirs explicatifs se recouvrent). On grimpe à 32,8% en ajoutant l'âge. Et même à 38,5% en prenant les PCS en 29 catégories et 45,1% avec la PCS en 428 catégories. Et, emportés par notre élan, nous pourrions encore gagner quelques pourcents avec le secteur d'activité, la région, les interactions entre les variables, etc. Mais il faut revenir à notre problème initial et représenter graphiquement les effets du sexe.

Bon, on peut faire simple et faire apparaître les moyennes par sexe et par catégorie professionnelle :

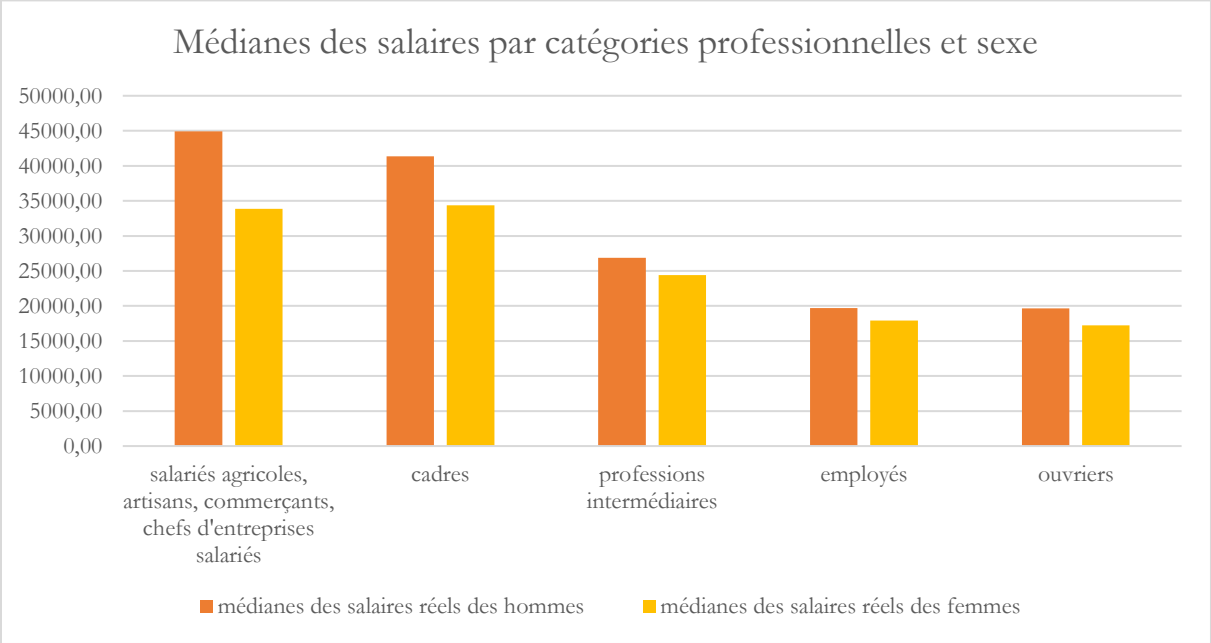
[37] J'ai utilisé sur Spss la comparaison de moyennes ainsi que le modèle linéaire général univarié pour les valeurs du R^2.
[38] Sans intégrer les interactions entre les variables explicatives (en ne gardant que les effets principaux). Pourquoi ? Pour aller plus vite dans les calculs sur ce gros fichier avec mon petit ordinateur. La prise en compte des interactions ajoute 2 ou 3% au R^2 sur les premiers modèles, mais je ne l'ai pas calculé pour le modèle avec les 428 professions…

Graphique 13. Salaires moyens par catégorie professionnelle et par sexe (travail à temps complet entre 23 et 62 ans)



Dans toutes les catégories sauf la première, les femmes ont des salaires inférieurs à ceux des hommes. On peut aussi regarder les médianes.

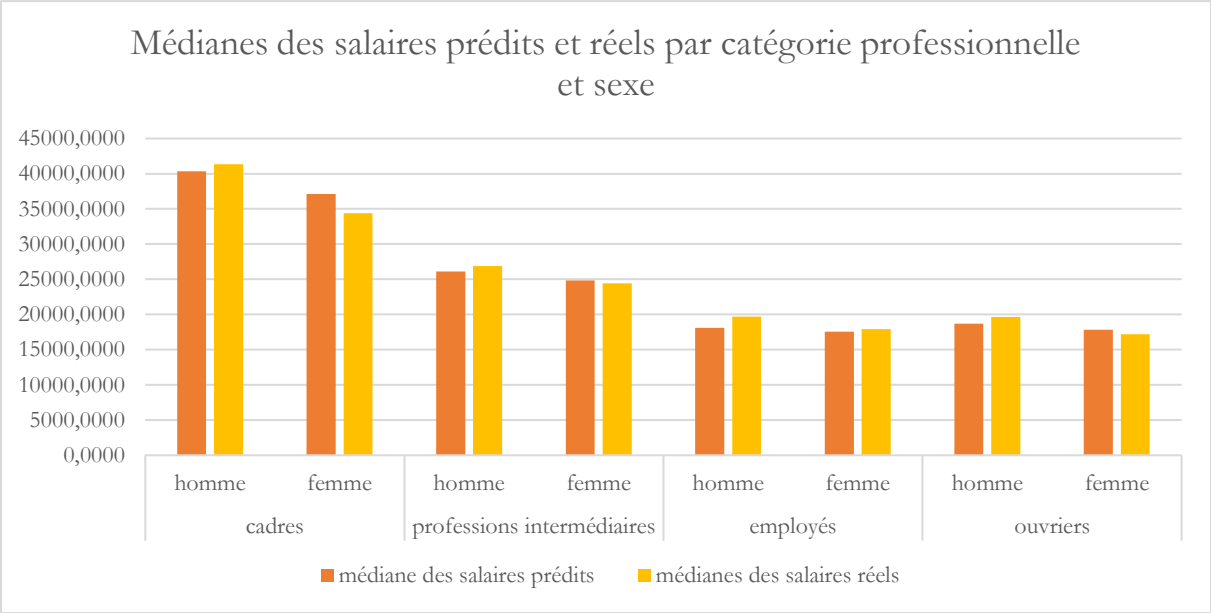
Graphique 13. Salaires médians par catégorie professionnelle et par sexe (travail à temps complet entre 23 et 62 ans)



Les différences sont plus nettes pour les deux premières catégories, ce qui suggère que les différences de salaires entre les femmes d'une même catégorie professionnelle sont plus marquées que celles qui concernent les hommes (une minorité de femmes à gros salaires fait croître les moyennes).

Enfin, on peut faire quelque chose de plus compliqué. J'ai fait un modèle à partir de la PCS en 428 catégories et de l'âge en tranches quadriennales et calculé les écarts entre la valeur prédite par le modèle et la valeur observée. Puis j'ai calculé la moyenne par grande catégorie professionnelle et sexe.

Graphique 14. Salaires médians prédits par catégorie professionnelle et par sexe (travail à temps complet entre 23 et 62 ans)



Dans toutes les catégories, les valeurs prédites pour les hommes sont inférieures aux valeurs observées et c'est l'inverse pour les femmes. Les salaires des hommes sont donc bien supérieurs à ceux des femmes en contrôlant l'âge et la catégorie professionnelle.

5. Résumés des corrélations de plus de deux variables

Lorsque l'on comprend bien les variables et les informations qu'elles contiennent, il peut parfois être utile d'avoir une vue globale sur un groupe de variables portant sur un même ensemble informationnel. Evidemment on peut également commencer par la vue globale si l'on connaît déjà bien les données, mais il est en général plus prudent de commencer par des analyses simples parce que l'interprétation des méthodes de résumé n'est pas toujours évidente.

Les résumés graphiques

Ces techniques, qu'on appelle parfois en France les **méthodes géométriques**, consistent à chercher à réaliser un ou plusieurs graphiques à deux dimensions qui résument le mieux possible l'information portée par un ensemble de variables.

L'analyse en composantes principales

La technique le plus ancienne et la plus simple est l'**analyse en composantes principales**, qui porte sur des variables quantitatives. Pour la présenter on prend souvent l'exemple d'un objet à trois dimensions, une baguette de pain par exemple, que l'on voudrait représenter à deux dimensions. Comment choisir ces dimensions et la façon de réaliser la représentation ? Il semble logique de chercher d'abord la plus grande longueur, puis la plus grande qui soit orthogonale avec la première. Si les coordonnées de différents points de la baguette dans les trois dimensions de l'espace usuel sont des variables, alors il existe une solution mathématique au problème. Cette solution consiste à partir de la matrice des covariances, ou, dans le cas le plus fréquent, de la matrice des corrélations, c'est-à-dire la matrice présentant dans chaque case le coefficient de corrélation linéaire entre la variable de la ligne et celle de la colonne. On cherche à diagonaliser cette matrice, ce qui signifie que l'on cherche une matrice équivalente sous certains aspects qui ne contienne des valeurs non nulles que sur la diagonale. Il existe des techniques mathématiques pour trouver les valeurs de la diagonale, que l'on appelle les « **valeurs propres** ». Si l'on dispose de cette matrice diagonalisée, alors il est possible d'exprimer chacune des variables de base comme une combinaison linéaire de ces valeurs et donc de représenter les points dans un espace correspondant aux axes définis par ces valeurs, axes que l'on appelle les composantes principales. On peut alors dessiner un graphique plaçant les points dans ce nouvel espace. L'intérêt est que le nombre de variables de départ peut être très grand et qu'il est donc très intéressant de les résumer par des graphiques simples.

Pour donner un exemple de ce type de technique, je vais changer de jeu de données. J'ai choisi les publications des 20 principales agglomérations scientifiques françaises en 2011 dans 10 grands domaines disciplinaires[39]. Une publication est attribuée à une agglomération si l'un des auteurs au moins a mentionné dans son adresse une commune de cette agglomération, chaque publication étant fractionnée en fonction du nombre d'agglomérations (dans le monde) auxquelles les auteurs sont rattachés. Comme pour les salaires, la distribution du nombre de publications par ville est de type lognormal. Il faut donc prendre le logarithme des variables pour utiliser des techniques qui présupposent la normalité des distributions, ce qui est le cas pour la version standard de l'analyse en composantes principales, qui se fonde sur les coefficients de corrélation.

[39] La source est le Web of Science, une base de données très utilisée bien que non exhaustive (les sciences sociales y sont notoirement sous-représentées). Merci à Laurent Jégou, Marion Maisonobe et Denis Eckert qui ont mis en forme ces données.

Les deux tableaux qui suivent présentent les trois premières lignes des données de base et de des variables résultant du calcul des logarithmes (népériens).

Tableau 10. Nombre fractionné de publications scientifiques 2011 des agglomérations françaises par grands domaines disciplinaires (trois premières lignes)

agglomération	bio_fondamentale	médecine	bio_appliquée	chimie	physique	sciences_univers	ingénierie	mathématiques	humanités	sciences_sociales
PARIS	2216,2044	4782,1876	838,2585	1590,7343	2403,1567	1357,6890	2356,7820	1056,3696	1135,6406	807,0165
LYON	410,3900	927,9924	133,9567	443,4591	314,2598	174,4685	463,2546	124,1194	106,1222	73,8714
TOULOUSE	260,4457	523,5604	174,6042	323,6017	290,8188	360,6055	539,4907	113,4500	120,3319	82,2524

Tableau 11. Logarithmes du nombre fractionné de publications scientifiques 2011 des agglomérations françaises par grands domaines disciplinaires (trois premières lignes)

agglomération	log_bio_fondamentale	log_médecine	log_bio_appliquée	log_chimie	log_physique	log_sciences_univers	log_ingénierie	log_mathématiques	log_humanités	log_sciences_sociales
PARIS	7,70	8,47	6,73	7,37	7,78	7,21	7,77	6,96	7,03	6,69
LYON	6,02	6,83	4,90	6,09	5,75	5,16	6,14	4,82	4,66	4,30
TOULOUSE	5,56	6,26	5,16	5,78	5,67	5,89	6,29	4,73	4,79	4,41

On calcule les coefficients de corrélation linéaires. Ici ces coefficients sont très élevés, ce qui signifie que l'information portée par chaque variable est très redondante avec celle portée par les autres, ce qui s'explique par le fait que le volume moyen de publications est très corrélé au nombre de chercheurs travaillant dans chaque agglomération.

Tableau 12. Coefficients de corrélations entre les logarithmes du nombre fractionné de publications scientifiques 2011 des agglomérations françaises par grands domaines disciplinaires

	log_bio_fondamentale	log_médecine	log_bio_appliquée	log_chimie	log_physique	log_sciences_univers	log_ingénierie	log_mathématiques	log_humanités	log_sciences_sociales
log_bio_fondamentale	1	0,938	0,882	0,876	0,864	0,865	0,796	0,831	0,918	0,929
log_médecine	0,938	1	0,781	0,821	0,797	0,769	0,774	0,833	0,954	0,943
log_bio_appliquée	0,882	0,781	1	0,732	0,716	0,831	0,739	0,792	0,768	0,807
log_chimie	0,876	0,821	0,732	1	0,877	0,851	0,884	0,836	0,827	0,890
log_physique	0,864	0,797	0,716	0,877	1	0,886	0,901	0,898	0,809	0,844
log_sciences_univers	0,865	0,769	0,831	0,851	0,886	1	0,918	0,854	0,767	0,886
log_ingénierie	0,796	0,774	0,739	0,884	0,901	0,918	1	0,924	0,791	0,875
log_mathématiques	0,831	0,833	0,792	0,836	0,898	0,854	0,924	1	0,841	0,876
log_humanités	0,918	0,954	0,768	0,827	0,809	0,767	0,791	0,841	1	0,935
log_sciences_sociales	0,929	0,943	0,807	0,890	0,844	0,886	0,875	0,876	0,935	1

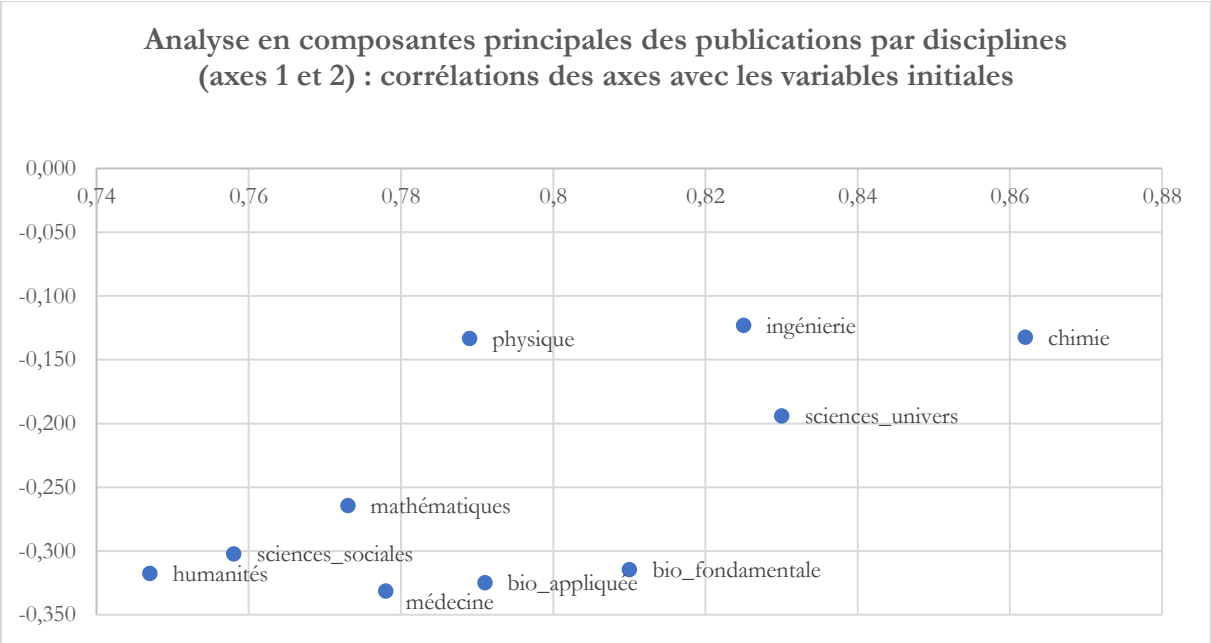
On calcule les valeurs propres. Le tableau suivant donne ces valeurs et la part de variance qu'elles représentent. On peut interpréter ces pourcentages comme la part d'information que porte chaque composante. On voir immédiatement que la première valeur est très importante et qu'elle capte une grosse part de l'information.

Tableau 13. Valeurs propres

Variance totale expliquée			
	Valeurs propres initiales		
Composante	Total	% de la variance	% cumulé
1	8,631	86,314	86,314
2	0,491	4,905	91,219
3	0,340	3,398	94,617
4	0,187	1,870	96,487
5	0,135	1,346	97,833
6	0,112	1,118	98,951
7	0,049	0,492	99,443
8	0,032	0,321	99,764
9	0,016	0,158	99,922
10	0,008	0,078	100,000

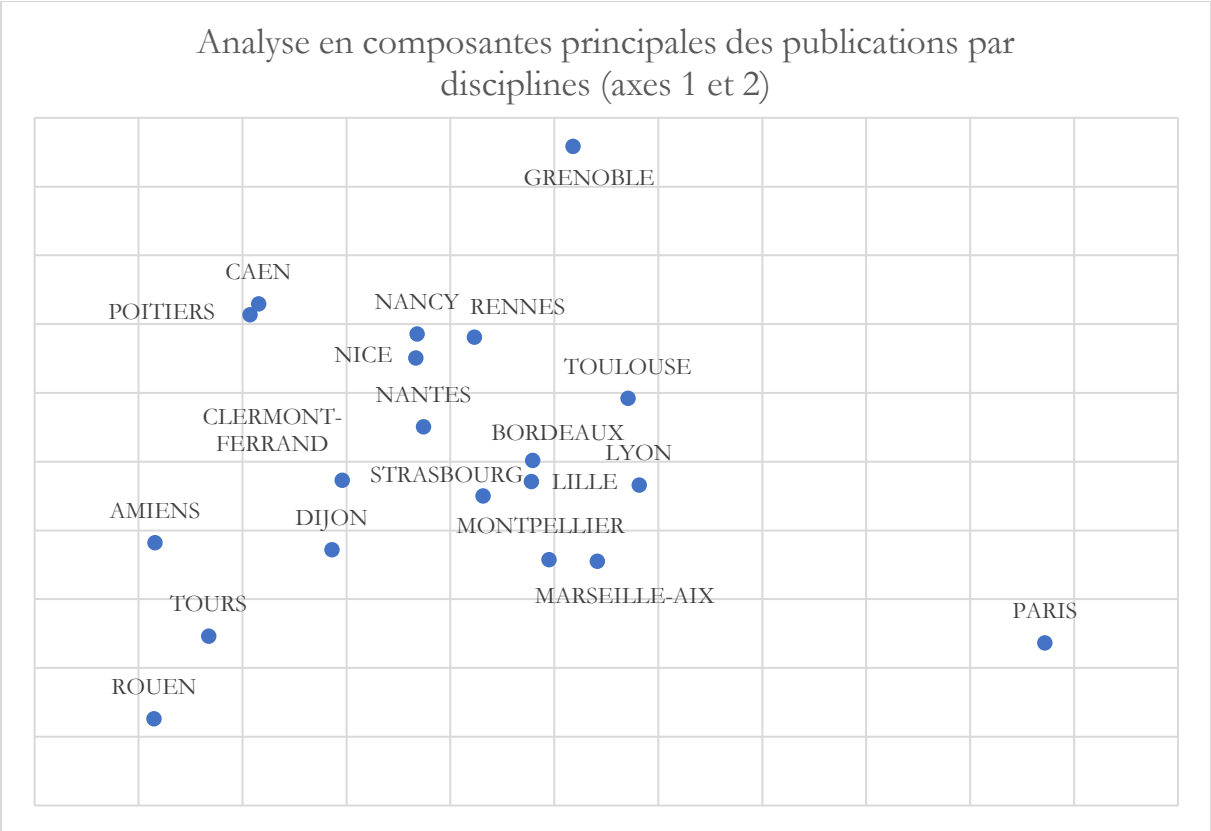
Habituellement on représente les variables de base en utilisant les coefficients de corrélation avec les composantes comme coordonnées. Pour chaque point, il faut imaginer une flèche qui part du centre du graphique. La direction de la flèche informe sur le sens des axes et l'éloignement du centre (le point de coordonnées 0,0) indique dans quelle mesure les variables initiales sont bien représentées. On peut imaginer l'image d'un bâton dont la représentation sur le plan serait l'ombre. Dans le graphique suivant, la chimie est mieux représentée dans le plan que les humanités. Globalement toutes les disciplines vont dans le même sens, ce qui fait du premier axe une représentation du nombre global de publications. C'est ce que l'on appelle un **axe « taille »** que l'on retrouve dans la plupart des analyses, les autres axes étant des **axes « forme »** qui différencient les unités d'analyse à niveau équivalent sur cet axe. Ici le deuxième axe opposera les agglomérations plutôt spécialisées dans les sciences de la matière et de la technique (physique, chimie, sciences de l'univers, ingénierie) à celles dans lesquelles les sciences du vivant sont plus présentes, les humanités, les sciences sociales et les mathématiques étant mal représentées sur ce plan (elles apparaissent plus nettement respectivement sur les axes 3, 5 et 4 que je ne présenterai pas pour cet exemple) et elles sont situées plutôt entre les deux.

Graphique 15. Représentation des variables dans le plan des deux premiers axes de l'analyse en composantes principales



On peut ensuite représenter les unités statistiques par leurs coordonnées sur les nouvelles variables. Paris est nettement plus à droite sur l'axe 1, ce qui exprime le décalage entre la capitale et les autres grandes agglomérations scientifiques. Grenoble est un site connu pour sa spécialisation dans les sciences de la matière et de la technique et se retrouve logiquement en haut. Paris, Montpellier, Lyon ou Aix-Marseille ont une part plus élevée de recherche dans les sciences de la vie, ces agglomérations sont donc plus bas sur le graphique.

Graphique 16. Représentation des unités statistiques dans le plan des deux premiers axes de l'analyse en composantes principales



Les analyses factorielles des correspondances

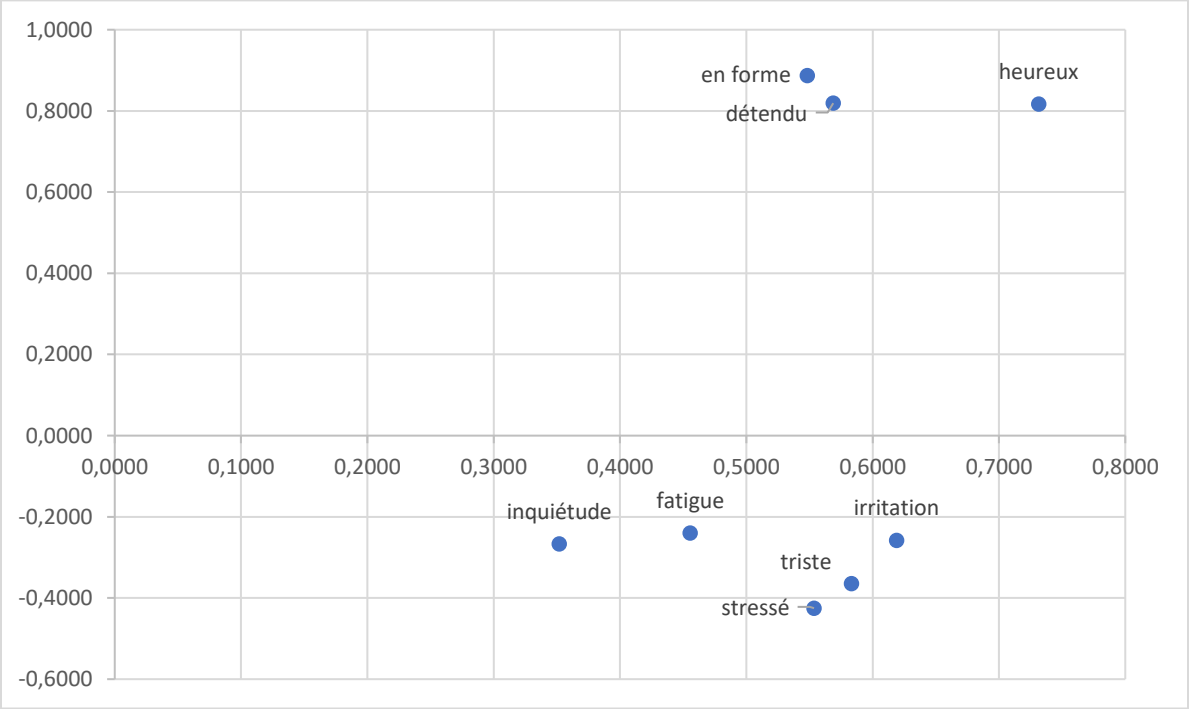
Lorsque l'on dispose de variables qualitatives, on ne peut pas utiliser l'analyse en composantes principales. Des mathématiciens ont imaginé une transposition consistant à faire des analyses en composantes principales sur des tableaux de pourcentages, qu'il s'agisse d'un simple tableau de contingence entre deux variables ou de tableaux plus complexes. Pour un tableau de contingence à deux variables, on a pu démontrer que les analyses en composantes principales des pourcentages lignes et des pourcentages colonnes peuvent se superposer (moyennant un coefficient permettant de créer une échelle commune), ce qui permet de faire apparaître les modalités des deux variables dans un même graphique. C'est **l'analyse factorielle des correspondances simple**. En tenant compte des qualités de représentation, pour lesquelles il existe des indicateurs (par exemple les cosinus carrés des vecteurs correspondants, avec le même principe de « l'ombre » que pour les analyses en composantes principales), la proximité sur les graphiques indique une corrélation (qu'il vaut mieux vérifier parce qu'il existe des effets qui font que deux modalités proches sur un graphique, mais mal représentées, sont en fait moins corrélées que ce qu'il paraît). L'intérêt est que cela procure une vue d'ensemble, utile lorsque les deux variables considérées ont un nombre relativement élevé de modalités. Mais ce qui est le plus utilisé, notamment par les sociologues français, c'est **l'analyse (factorielle) des correspondances multiples**, qui est une généralisation de la technique de base en agrégeant des tableaux de contingences multiples ou en transformant chaque modalité en variable indicatrice (1 si la modalité est présente, 0 sinon). Evidemment ces analyses sont effectuées au moyen de logiciels. Il suffit de comprendre leur fonctionnement général en se rappelant ce qui a été présenté plus haut pour l'analyse en composantes principales.

Pour illustrer cette technique, je vais m'appuyer sur un autre jeu de données, celui issu de la première vague de l'enquête « La vie en confinement » (Vico). Il s'agit d'une enquête en ligne,

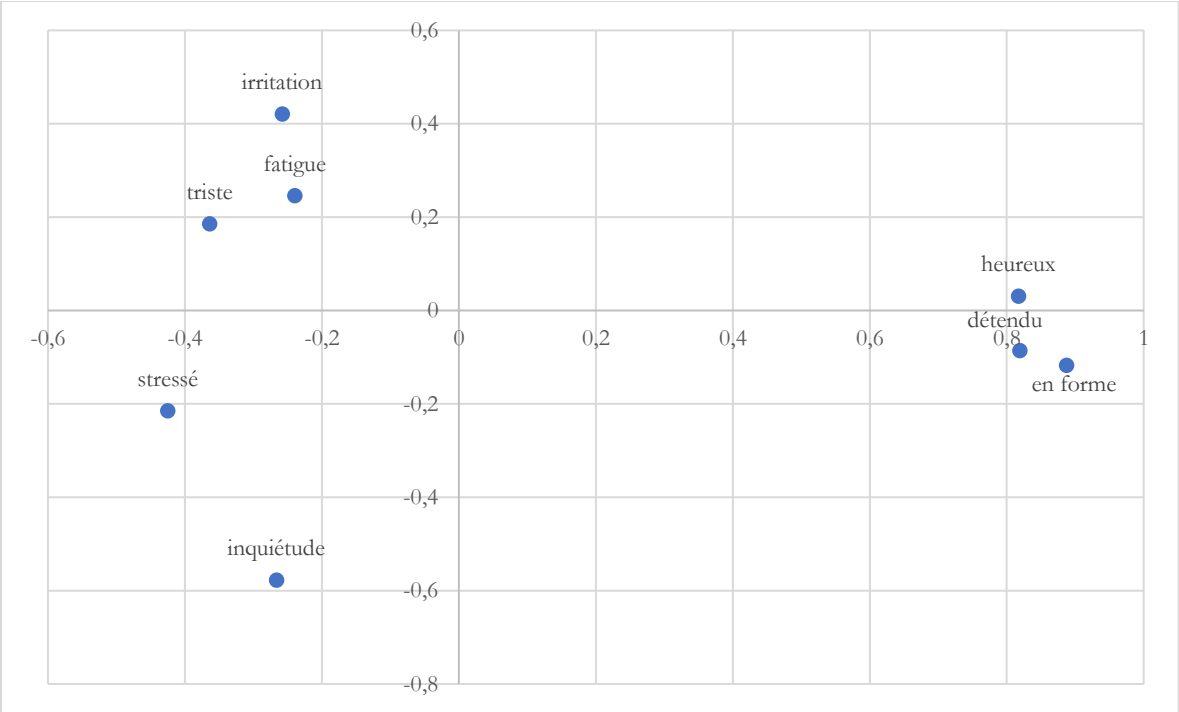
réalisée pendant le premier confinement de 2020 en France, qui a permis de collecter 16224 questionnaires complets utilisables. Parmi les thèmes abordés dans le questionnaire figurait une question sur les sentiments suscités par la situation : « Depuis le début du confinement, y a-t-il eu des moments où vous vous êtes senti(e) ? », suivie de la possibilité de cocher autant de réponses que l'on voulait parmi les catégories « fatigué(e) », « irrité(e) », « détendue(e) », « en forme », « stressé(e) », « triste », « heureux(se) », « inquiet(e) », « rien de tout cela », « autre » (avec la possibilité de préciser sous la forme d'un texte). La fréquence de la réponse « rien de tout cela » est nettement plus élevée chez les hommes, s'accroît avec l'âge (elle est surtout élevée chez les plus âgés) mais diminue avec l'élévation du niveau d'études : 0,7% chez les femmes les plus diplômées de 18-30 ans et 21% chez les hommes les moins diplômés de plus de 60 ans. Les réponses « autres » semblent surtout ajouter une explication par rapport aux réponses proposées, souvent la colère contre les autorités, ou l'angoisse, mais elles ne sont pas corrélées très significativement à l'une ou l'autre des autres réponses en particulier. Elles se font légèrement plus fréquentes avec l'avancée en âge. L'analyse des croisements entre les variables permet de vérifier que la réponse « rien de tout cela » correspond au cas où aucune des autres réponses n'est cochée. Cela conduit à centrer l'analyse sur les 8 premières réponses.

On réalise donc une analyse des correspondances multiples sur les 8 expressions de sentiments. Le premier axe (26,9% de la variance), comme attendu (un axe « taille » classique), oppose les personnes ayant coché peu de réponses, voire aucune, à celles qui en ont coché une ou plusieurs. Le deuxième axe (22,6%) oppose les expressions de sérénité (« détendue(e) », « en forme », « heureux(se) ») aux autres et le troisième (11,1%) oppose les expressions de souffrance (« fatigué(e) », « irrité(e) », « triste ») » à celle de stress et d'inquiétude.

Graphique 17. Axes factoriels 1 et 2 de l'analyse des correspondances multiples sur l'expression des sentiments durant le premier confinement français (variables actives)



Graphique 18. Axes factoriels 2 et 3 de l'analyse des correspondances multiples sur l'expression des sentiments durant le premier confinement français (variables actives)

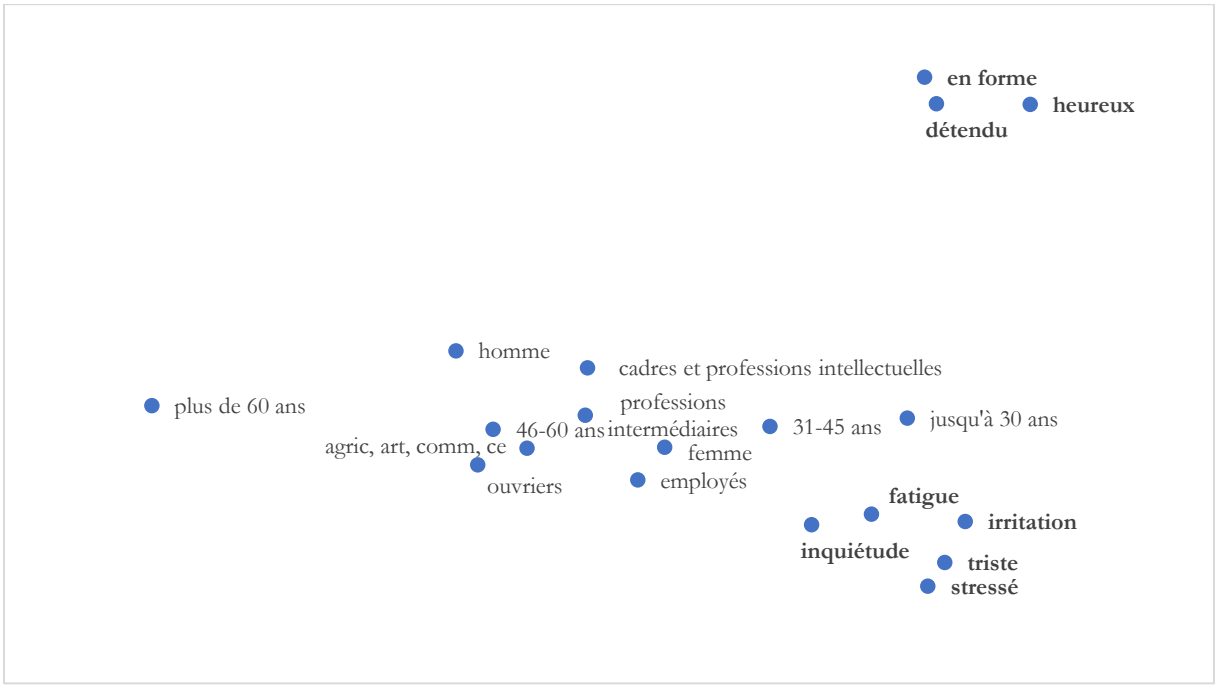


J'ai interprété rapidement ces graphiques, parce que je connais bien ces données et que le cas est assez simple, avec des variables relevant d'un même univers de sens, mais une analyse rigoureuse impliquerait d'évaluer la qualité de représentation des différentes modalités sur les plans factoriels au moyen d'indicateurs comme la contribution aux axes de ces modalités ou le carré du cosinus de l'angle entre le vecteur qui les représente dans l'espace comprenant toutes les dimensions correspondant à l'ensemble des données et la projection de ce vecteur dans le plan factoriel considéré. Cette phrase pouvant paraître un peu sibylline, je suggère aux personnes qui voudraient utiliser ces techniques de se référer à des cours ou des ouvrages plus spécialisés pour détailler ces différents indices. L'important est de comprendre que les plans factoriels sont une représentation partielle, une projection, comme lorsque l'on réalise un plan en deux dimensions d'un bâtiment, et qu'il faut toujours se préoccuper de la qualité de représentation des différents points.

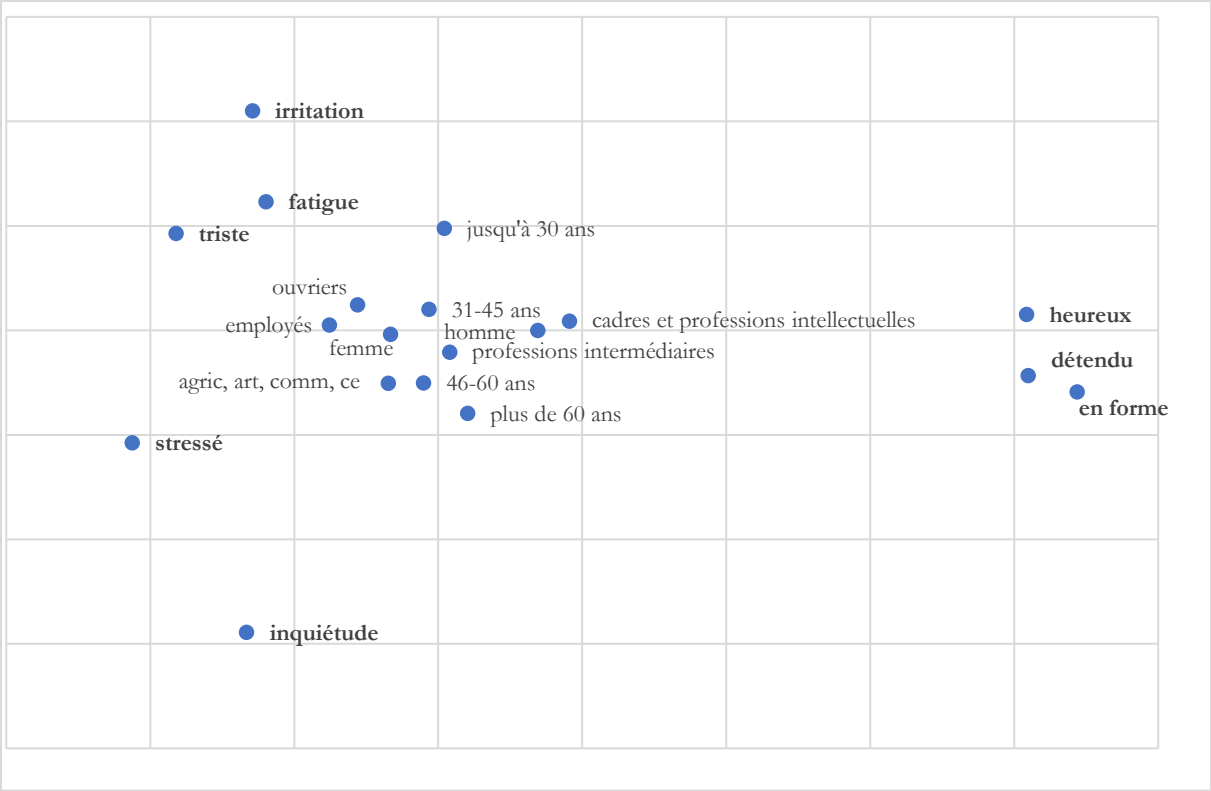
Il faut aussi signaler une autre caractéristique des techniques de cette famille. Quelles que soient les données, elles produisent toujours un résultat qui représente d'une façon globale les corrélations les plus importantes de l'ensemble des variable examinées, ce que les rend particulièrement précieuses pour avoir une vue d'ensemble. Cependant, si elles hiérarchisent les corrélations les unes par rapport aux autres, elles ne donnent guère d'information sur l'importance intrinsèque de ces corrélations. Par exemple dans les deux graphiques, « détendu » et « en forme » sont très proches, ce qui suggère une corrélation forte. C'est le cas puisque 64,5% des personnes se déclarant détendues se sont aussi dites en forme (contre 19,1% pour celles qui n'ont pas coché la réponse « détendu·e »), le pourcentage réciproque étant de 68,5% (contre 31,5%). Bien que la corrélation soit forte, elle ne va pas jusqu'à une identité des deux variables et elles finissent par se séparer sur le cinquième axe. Si les corrélations étaient globalement plus faibles dans ces données, l'algorithme produirait tout de même des graphiques avec des proximités entre les modalités les plus corrélées. Autrement dit, si l'on veut dire quelque chose sur la corrélation précise entre deux modalités, il vaut mieux effectuer un tri croisé simple.

Dans toutes les techniques de résumé graphique, il est possible représenter sur les axes factoriels des variables qui n'ont pas participé à la détermination de ces axes. Ces variables « **supplémentaires** » (les variables utilisées pour calculer les axes étant désignées comme « **actives** ») sont représentées à partir de leur corrélation avec les composantes principales pour cette technique et pour des variables quantitatives, et, pour l'analyse des correspondances, par les moyennes des modalités des variables supplémentaires sur les différents axes. Dans les graphiques qui suivent ont été projetées sur les plans factoriels des caractéristiques sociodémographiques (tranches d'âge, sexe, catégories professionnelles).

Graphique 19. Axes factoriels 1 et 2 de l'analyse des correspondances multiples sur l'expression des sentiments durant le premier confinement français (variables actives et supplémentaires)



Graphique 20. Axes factoriels 2 et 3 de l'analyse des correspondances multiples sur l'expression des sentiments durant le premier confinement français (variables actives et supplémentaires)



Le premier plan permet de voir que l'expression de sentiments divers est plus fréquente chez les jeunes et les femmes, le deuxième montre que les expressions de sérénité sont corrélées avec les positions de cadre et sont plus souvent présentes chez les hommes. Les plus jeunes expriment plus de fatigue et les plus âgés de l'inquiétude. Ce n'est là qu'un exemple pour illustrer cette technique et l'on pourrait analyser bien plus en détail ces données. Je les réutiliserai dans la section suivante sur les classifications automatiques. En effet, les résumés graphiques sont souvent combinés avec des méthodes de classification, qui ont pour objectif de regrouper des unités statistiques en fonction de leur similarité.

Les classifications automatiques

Il existe de nombreuses méthodes de classification automatique. Elles combinent nécessairement deux éléments essentiels : 1) une mesure de similarité ou de dissimilarité et 2) un algorithme de tri et d'affectation des unités statistiques à des classes.

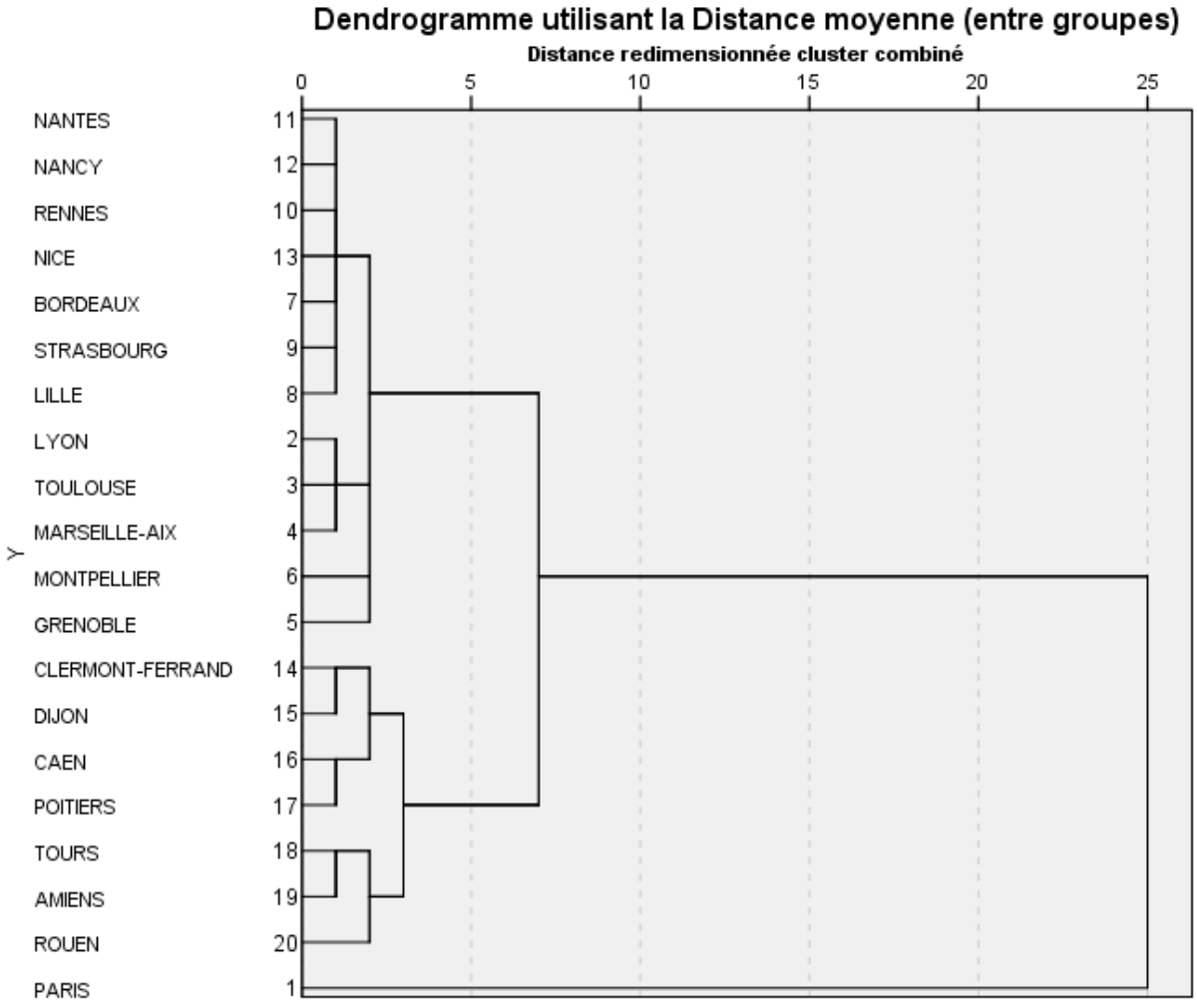
Pour les variables quantitatives, la mesure de similarité la plus fréquente est la simple distance euclidienne. Si deux individus A et B sont mesurés par les variables X, Y et Z (chaque individu a une mesure pour chacune des trois variables), alors la distance entre A et B est la racine carrée de la somme des carrés des différences. Il existe bien d'autres mesures de distance ou de proximité. Je ne vais pas les inventorier, mais le principe est toujours le même. Pour les variables qualitatives, il existe la distance dite du Chi2 (ou une variante appelée Phi2) qui est une fonction du nombre de modalités pour lesquelles deux unités statistiques prennent des valeurs différentes, du nombre de variables et des effectifs des modalités pour lesquelles les deux unités prennent des valeurs différentes. Une autre solution est de faire d'abord une analyse factorielle et d'utiliser les axes, qui sont des variables quantitatives, pour effectuer la classification. Si l'on prend tous les axes, cette

opération revient à transformer une distance du Chi2 (qui est au fondement de l'analyse factorielle) en distance euclidienne.

Les algorithmes de classification sont également nombreux. Une méthode très fréquemment utilisée consiste à procéder de façon hiérarchique, soit par regroupements successifs (**classification hiérarchique ascendante**) soit par divisions en classes de plus en plus petites (**classification hiérarchique descendante**), avec dans les deux cas de nombreuses variantes. Il existe aussi des méthodes de **centres mobiles** (dont l'une des variantes est la méthode dite des « **nuées dynamiques** ») consistant à tirer aléatoirement un nombre donné de centres de classes auxquels on agrège les éléments les plus proches, puis à recalculer le centre de chaque classe (le point le plus près de tous les autres), affecter à nouveau les éléments, et ainsi de suite jusqu'à stabilisation de la classification.

A titre d'exemple de classification sur des variables quantitatives, on peut reprendre les données sur les publications scientifiques par disciplines des 20 plus grandes agglomérations scientifiques françaises. Si l'on effectue une classification hiérarchique ascendante sur les logarithmes des totaux de publications par grands domaines disciplinaires (avec comme distance le carré de la distance euclidienne), on obtient les regroupements suivants :

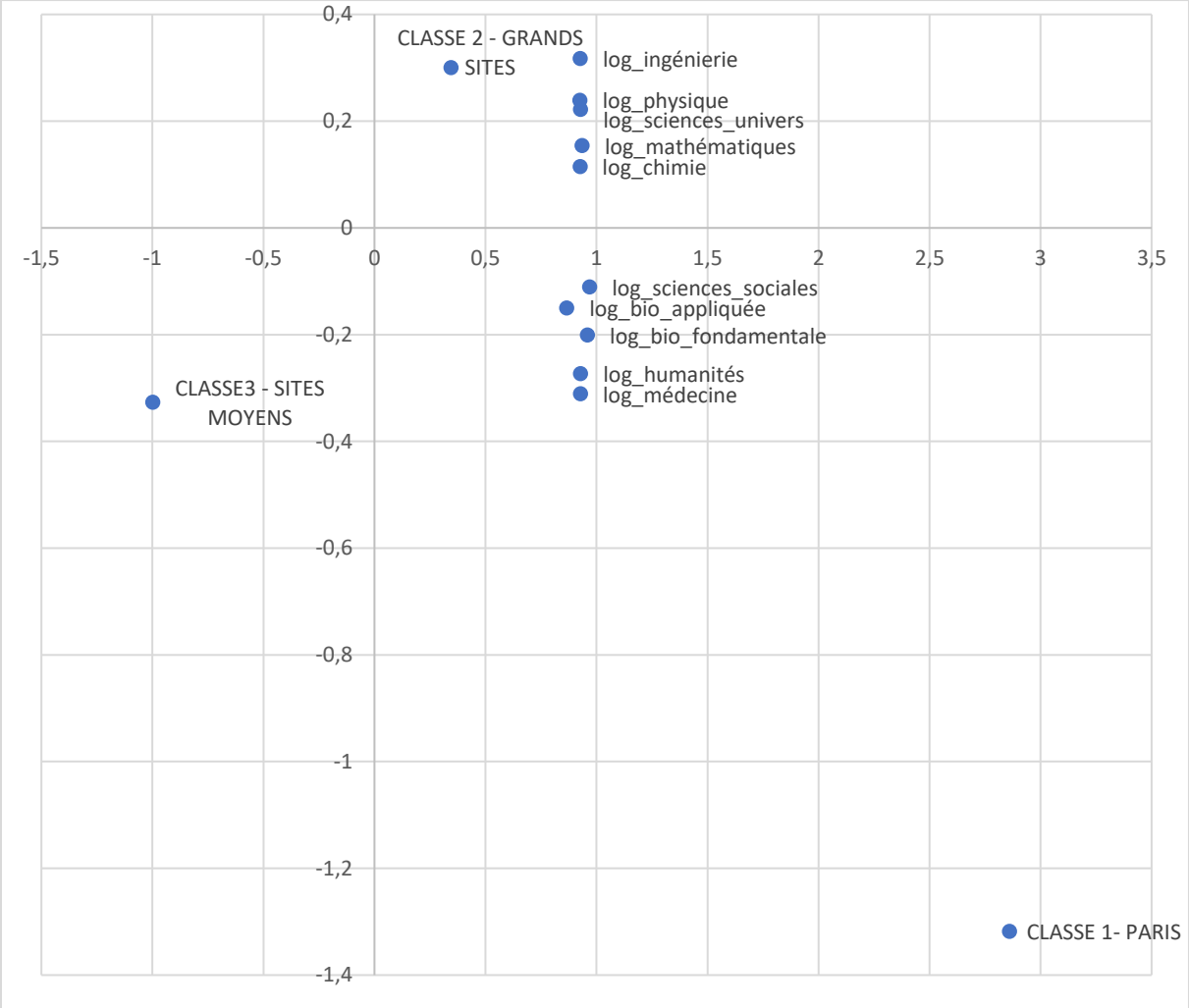
Graphique 21. Arbre de regroupement des agglomérations (distance euclidienne au carré moyenne entre groupes)



On voit que Paris, par sa taille, reste constamment isolé et que l'avant-dernier regroupement agrège les plus grands sites aux sites moyens. Le plus simple semble être de couper cet arbre en

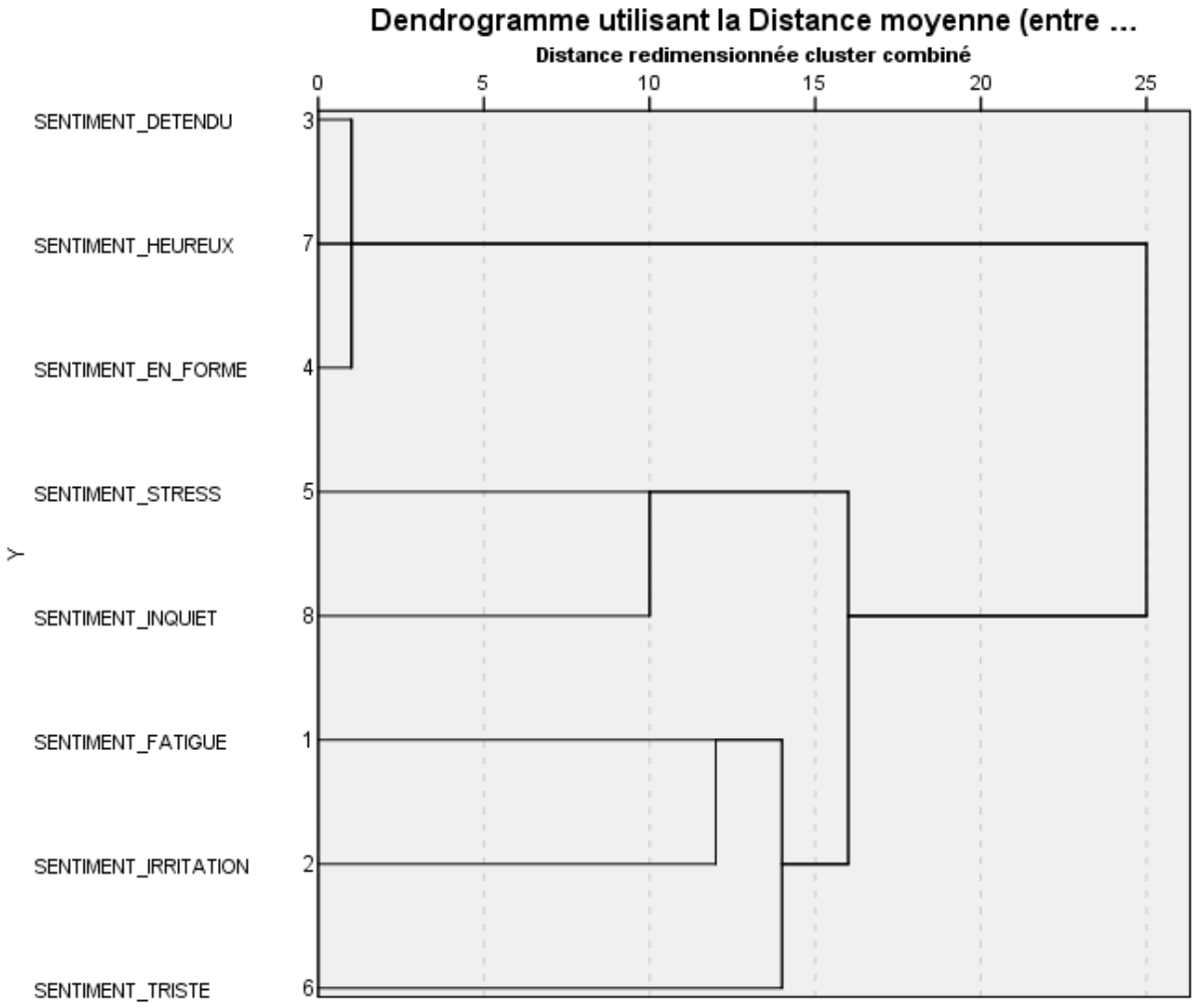
faisant trois groupes, que l'on peut représenter sur le premier plan de l'analyse en composantes principales que nous avions réalisée. Les classes sont représentées en même temps que les variables.

Graphique 22. Trois classes représentées sur le premier plan factoriel



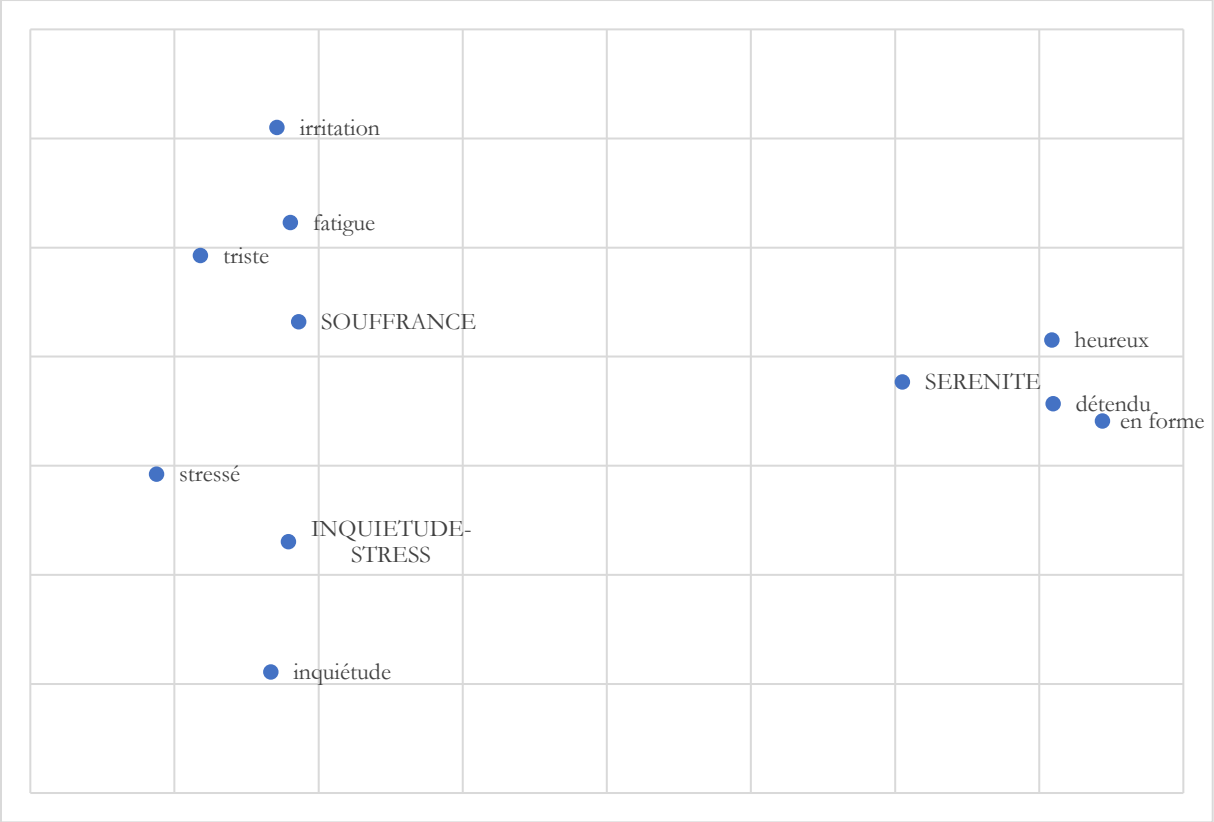
Prenons à présent des variables qualitatives, avec les données déjà utilisées de l'enquête « La vie en confinement ». La distance est celle du chi2 pour à nouveau une classification ascendante. Comme nous avons cette fois-ci plus de 16000 unités statistiques, c'est le dendogramme des variables qu'il est intéressant de regarder.

Graphique 23. Dendogramme de la classification hiérarchique sur les variables d'expression de sentiments



Les regroupements semblent logiques, d'abord les réponses traduisant une certaine sérénité puis celles qui signalent une inquiétude, puis enfin celles qui expriment une forme de souffrance. Il semble donc logique de regrouper les variables en 3 groupes. Je le fais en créant les variables synthétiques SERENITE, INQUIETUDE_STRESS et SOUFFRANCE. Projetons-les sur le deuxième plan factoriel de l'analyse des correspondances multiples présentée plus haut. Rappelons comment cela peut se faire : simplement en calculant les moyennes des axes concernés pour les modalités correspondant aux trois variables (chaque variable est codée sur deux modalités : présence ou absence).

Graphique 24. Projection des trois classes sur le plan factoriel des axes 2 et 3 de l'analyse des correspondances multiples sur les 8 variables d'expression de sentiments



Chacune de ces variables synthétiques peut ensuite être corrélée avec d'autres variables. En utilisant des arbres de décision (voir plus haut), on se rend compte que les expressions de sérénité sont particulièrement présentes chez les 18-30 ans de niveau d'étude égal ou supérieur à quatre années après le baccalauréat (76,7% contre 56,2% dans la population générale), que les expressions de souffrance sont plus fréquentes chez les femmes de 18 à 30 ans (90,9% contre 75,8%), et que les expressions d'inquiétude ou de stress sont également mentionnées plus souvent par les femmes de 18 à 45 ans ayant un niveau d'étude égal ou supérieur à quatre années après le baccalauréat (83,7% contre 73,4%). Comme les personnes expriment souvent des sentiments divers et contradictoires, particulièrement chez les 18-30 ans, on peut aussi construire des variables exclusives. Par exemple l'expression de sérénité sans présence des deux autres catégories ne concerne que 8% des réponses, mais le pourcentage le plus élevé est de 14,7% chez les hommes de plus de 45 ans.

6. Décomposer les corrélations : les régressions

Les régressions sont souvent associées à une logique probatoire dans laquelle on veut démontrer l'influence d'une variable sur une autre « toutes choses égales par ailleurs », en ayant auparavant formulé des hypothèses sur les raisons de cette influence. Mais ces techniques peuvent parfaitement être utilisées dans une optique exploratoire. Au lieu de faire une série de modèles destinés à mettre en évidence un effet précis, on utilise les modèles pour tester des idées, voir si une corrélation se maintient lorsque l'on neutralise les effets de telle ou telle variable, se faire une idée de l'importance de la corrélation par la part de variance expliquée plutôt que par des coefficients de corrélation linéaire ou des chi2 partiels.

Dans la partie consacrée aux corrélations entre variables quantitatives j'ai présenté le principe de la régression linéaire et j'ai aussi utilisé des modèles linéaires (des analyses de variance) dans celle portant sur les corrélations entre variables quantitatives et qualitatives en testant différentes variables pour expliquer les variations de salaire. Nous avions vu que la prise en compte des catégories professionnelles au niveau le plus fin, combinées avec l'âge et le sexe, permettait de rendre compte d'une part importante des variations de salaire.

Plutôt que de passer en revue à nouveau toutes ces techniques, je vais prendre un exemple de régression adaptée aux variables qualitatives, les régressions logistiques. Dans ce type de technique, le critère d'ajustement du modèle n'est pas une distance, mais la « vraisemblance », que l'on peut présenter comme la probabilité que les valeurs effectivement observées apparaissent sous l'hypothèse où le modèle les prédit. L'algorithme calcule les paramètres du modèle pour maximiser la vraisemblance. Le terme « logistique » renvoie à un type de distribution (les courbes en « S » qui rendent compte par exemple de la proportion de personnes ayant adopté une innovation par unité de temps à partir de l'introduction de celle-ci), parce que la probabilité de présence d'une modalité en fonction d'une configuration de variables explicatives se distribue de cette façon.

Pour illustrer la logique d'usage de ces régressions, nous allons rester sur les données utilisées précédemment sur les sentiments exprimés pendant le premier confinement français. La variable à expliquer sera l'expression d'un sentiment de sérénité (« heureux », « détendu » ou « en forme ») à l'exclusion des autres types de sentiment. Nous avions vu que ces expressions de sérénité sont particulièrement présentes chez les hommes de plus de 45 ans. Mais quelle part des variations ces deux variables (tranche d'âge et sexe) expliquent-elles ? Il n'existe pas pour les régressions logistiques de mesure aussi simple que le R^2 des régressions linéaires. Divers auteurs ont proposé des équivalents, généralement fondés sur le rapport entre la vraisemblance du modèle calculé et celle d'un modèle « nul » sans variables explicatives.

Tableau 14. Paramètres estimées de la régression logistique de la variable rendant compte de la présence d'une expression de sérénité sans autres type de sentiment avec les tranches d'âge comme variable explicative.

SENTI_SERENITE_SEULT[a]		B	Erreur standard	Wald	ddl	Sig.	Exp(B)	Intervalle de confiance à 95 % pour Exp(B)	
								Borne inférieure	Borne supérieure
,00	Constante	1,940	0,055	1262,999	1	0,000			
	18-30 ans	1,069	0,095	126,776	1	0,000	2,912	2,417	3,507
	31-45 ans	0,727	0,079	84,542	1	0,000	2,069	1,772	2,415
	46-60 ans	0,320	0,075	18,280	1	0,000	1,378	1,189	1,596
	Plus de 60 ans	0[b]			0				
a. La catégorie de référence est : 1,00.									
b. Ce paramètre est défini sur 0, car il est redondant.									

Il y a plein de choses dans ce tableau, mais la plupart des informations qu'il contient sont redondantes. La note « a » dit que la catégorie de référence est 1, c'est-à-dire l'expression d'un sentiment de sérénité exclusif (sans que les personnes aient mentionné des sentiments d'un autre type). Le modèle est donc censé prédire le contraire, l'absence de ce type d'expression. La note « b » dit que pour la variable explicative, c'est la dernière modalité qui fait référence, ce qui signifie que les paramètres expriment des variations par rapport à cette modalité. La constante et les valeurs B sont censées représenter une équation linéaire (Constante + B_1X_1 + B_2X_2, etc.[40]) qui prédit le rapport entre la probabilité de réalisation de la modalité à expliquer (ici ne pas avoir exprimé de sentiment de sérénité exclusif) et la probabilité de non réalisation. C'est un peu compliqué mais ici, on voit bien que, par rapport aux plus âgés, la probabilité de ne pas citer exclusivement des sentiments de sérénité augmente lorsque l'âge diminue, et donc que la probabilité de citer exclusivement ce type de sentiment augmente avec l'âge. La statistique de Wald se lit comme un test du Chi2, qui est là significatif pour toutes les tranches d'âge (sig. est la probabilité de se tromper en concluant qu'il y a un effet de la modalité sur la variable à expliquer). Exp(B) est un « rapport de chances » (odds ratio) qui exprime la proportion de réalisation de la modalité à expliquer (ne pas avoir cité, etc.) pour la modalité explicative par rapport à la modalité de référence. Ici les plus jeunes sont censés avoir 2,417 fois plus chances de ne pas avoir exprimé le sentiment exclusif de sérénité que les plus âgés. Enfin, il y a une estimation de l'« intervalle de confiance » (les valeurs entre lesquelles la valeur réelle a 95% de chances de se situer) pour ces rapports de chance.

Cette régression très simple ne nous apprend rien par rapport à un simple tri croisé.

[40] X_1, X_2, etc. prennent les valeurs 0 ou 1 selon que la modalité est absente ou présente.

Tableau 15. Expression exclusive de sérénité par tranche d'âge.

	Proportion de personnes expriment seulement de la sérénité
jusqu'à 30 ans	4,7%
31-45 ans	6,5%
46-60 ans	9,4%
plus de 60 ans	12,6%
Ensemble	8,0%

Source : enquête VICO
Lecture : 4,7% des 18-30 ans ont exprimé des sentiments de sérénité sans citer d'autres types de sentiments.

Elle nous donne toutefois une information supplémentaire par les estimations de la part de variance expliquée. Comme je l'ai expliqué plus haut, il n'y a pas de consensus sur la façon de calculer cette valeur, divers auteurs ayant proposé des solutions. Le logiciel que j'ai utilisé en présente trois.

Tableau 16. Part de variance expliquée (pseudo R^2) de la régression sur l'âge

Pseudo R-deux	
Cox et Snell	0,010
Nagelkerke	0,024
McFadden	0,018

Selon les modes de calcul, les tranches d'âge expliquent entre 1% et 2,4% de l'expression exclusive d'un sentiment se sérénité. D'un point de vue exploratoire, la valeur précise n'a pas d'importance. **Ce qui compte c'est l'ordre de grandeur et la façon dont il évolue lorsque l'on introduit d'autres variables**. Dans ce cas par exemple, l'ajout du niveau d'études n'augmente pas beaucoup l'explication.

Tableau 17. Part de variance expliquée (pseudo R^2) de la régression sur l'âge et le niveau d'études

Pseudo R-deux	
Cox et Snell	0,011
Nagelkerke	0,025
McFadden	0,019

Si l'on rajoute le sexe, cela augmente un peu (le premier coefficient passe à 0,017 et d'autres estimations montrent que, une fois l'âge pris en compte, le niveau d'études est peu significatif). On peut ainsi ajouter d'autres variables, la configuration de confinement par exemple, en faisant un peu augmenter les pseudo R^2, mais ils restent faibles (moins de 5%), ce qui permet de rappeler à nouveau que **l'existence d'une corrélation statistiquement significative ne présume pas de l'importance de l'effet de la variable explicative sur celle que l'on cherche à expliquer**.

On peut ainsi essayer divers modèles, les combiner avec des tableaux de contingence à plusieurs entrées et des arbres de décision pour se faire progressivement une idée de ce révèlent les données.

7. Analyses de réseaux

L'analyse de réseaux sociaux est un courant de recherche à présent relativement ancien puisque l'on crédite habituellement l'anthropologue John Barnes de la première définition explicite de la notion de réseau social[41] mais que des auteurs qui se sont intéressés à l'histoire de cette tradition en font remonter les prémices à des périodes bien antérieures[42]. Dans ce type d'étude on s'intéresse aux relations « dyadiques » (entre deux entités, que celles-ci soient des personnes, des organisations ou tout autre type de forme sociale) et aux réseaux constitués par l'agrégation de ces relations.

Comme on ne peut pas inventorier toutes les relations, les chercheurs ont progressivement constitué trois types de méthodes pour collecter des informations sur les relations.

La première méthode est celle des **réseaux dits « complets »** qui consiste à délimiter un ensemble social (les membres d'une organisation, les habitués d'un lieu donné, etc.) et d'essayer d'obtenir des informations sur toutes les relations d'un certain type au sein de cet ensemble. On obtient alors non seulement des informations sur chaque entité et sur les relations dans lesquelles elle est impliquée, mais aussi sur le réseau d'ensemble, que l'on peut représenter sous la forme d'une matrice dont les entités sont à la fois les lignes et les colonnes, chaque case comprenant une information sur la relation qui existe entre la ligne et la colonne. A partir d'une telle matrice on peut dessiner des graphes dans lesquels les unités sont représentées par des points (ou des cercles ou tout autre symbole graphique) et les relations par des traits. Ces traits peuvent être simples si l'information disponible est seulement l'existence ou non d'une relation. Ils peuvent prendre la forme d'une flèche si les relations sont orientées (par exemple si elles concernent des personnes qui en conseillent d'autres) et des traits plus ou moins épais si en plus ces relations sont valuées c'est-à-dire que l'on dispose d'informations sur leur importance (par exemple les flux d'informations). L'analyse de ces structures ne relève pas seulement des statistiques, elle fait intervenir des notions issues de la théorie des graphes ou de l'analyse combinatoire : nœuds, arrêtes, cliques, densité, centralité, etc. L'objectif du présent livre n'est pas de présenter toutes ces notions et les méthodes d'analyse des graphes. Je m'en tiendrai à ce qu'il me semble utile de savoir pour procéder à des analyses statistiques. Les analyses statistiques sont particulièrement utiles pour les deux autres méthodes, dans lesquelles la dimension structurelle est moins centrale.

La deuxième méthode de constitution de données relationnelles est celle des **réseaux dit « personnels »**. Elle part d'un échantillon de personnes construit de la même façon que dans les enquêtes habituelles en sciences sociales. A chaque personne interrogée par questionnaire ou entretien, on pose des questions en forme de « générateur de noms », des questions qui incitent les personnes à citer des relations. Par exemple, un générateur très utilisé à des fins de comparaison et qui a fait l'objet de beaucoup de discussions consiste à demander à qui la personne interrogée pourrait parler de choses importantes et à l'inciter à citer un nombre déterminé de noms (5 généralement). Pour chaque relation, on pose un certain nombre de questions sur la personne citée (âge, sexe, profession, etc.) et sur la relation (type de relation, contexte de rencontre, ancienneté du lien, etc.). Certaines enquêtes de ce type documentent beaucoup plus de relations (jusqu'à 50 ou 100), soit à partir d générateurs divers, soit en inventoriant des contextes d'activité et des personnes qui y sont fréquentées[43]. Dans de nombreux cas, on pose aussi des questions sur les relations entre les personnes citées, telles qu'elles sont perçues par la personne interrogée. Dans ce type d'enquête on dispose donc d'informations sur plusieurs unités d'analyse : les personnes interrogées (egos), les personnes citées (les alters) et leurs relations avec les egos, les relations entre alters. On a donc a minima un fichier sur les egos et un autre sur les alters, parfois plusieurs fichiers d'alters si une

[41] « Class and Committees in a Norwegian Island Parish », *Human Relations*, VII, 1954, pp. 39-58

[42] Linton C. Freeman, 2004, The Development of Social Network Analysis: A Study in the Sociology of Science, Empirical Press, Vancouver, BC

[43] Claire Bidart. Panel de Caen, 2016, « Note méthodologique. : Hypothèses, élaboration de l'enquête et suivi du panel » ⟨halshs-00118258v2⟩.

partie des alters seulement est documentée plus en détail[44]. Les informations sur les alters et sur les liens entre alters permettent de calculer des indicateurs sur le réseau (taille, densité, etc.), que l'on peut faire « remonter » sur le fichier des egos.

Les indicateurs permettant de caractériser les relations et les réseaux sont très nombreux. Les plus courants sont les suivants. Pour les relations, on essaie en général d'obtenir des informations sur le type de relation (famille, ami, collègue …), l'ancienneté (nombre d'années d'interconnaissance), le contexte de rencontre (durant les études, au travail, etc.), la similarité de caractéristiques (âge, sexe, niveau d'études, etc.) entre ego et chaque alter, la force du lien (par exemple à travers des questions sur le fait pour ego de se sentir proche de l'alter). Pour le réseau, on cherche en général à calculer la taille (le nombre d'alters), la densité (le nombre de relations observées entre les alters divisé par le nombre de relations possibles[45]), la composition (part des relations familiales, des relations professionnelles, etc.). Lorsque le nombre d'alters est élevé et que l'on dispose d'informations sur les relations entre eux, ce qui est rare, il est possible d'utiliser de nombreux autres indicateurs et de mettre en œuvre des méthodes spécifiques d'analyse que je ne détaille pas ici[46].

La troisième méthode, les **chaînes relationnelles**, consiste à documenter des processus d'accès à des ressources (emploi, services divers) ou de mise en contact de personnes qui mettent en jeu des relations personnelles. Ici, on ne cherche pas à cartographier un réseau de façon statique, mais à capter des relations activées dans un type particulier de situation. Par exemple, on peut demander aux personnes interrogées de donner des informations sur la façon dont ils ont obtenu leur dernier emploi, ou plusieurs emplois au cours de leur carrière et, lorsque le processus a impliqué des relations personnelles, demander des informations sur ces relations, dans une logique similaire à celle qui prévaut dans les études de réseaux personnels. Le questionnement sur le processus fonctionne comme un « générateur par événements ». Mais il y a des chaînes impliquant d'autres relations. Souvent les informations sur les autres relations, les personnes intervenant dans le processus qui ne sont pas connues directement de la personne qui répond, sont plus difficiles à obtenir, mais on parvient généralement à estimer le nombre d'intermédiaires, donc la longueur de la chaîne (le nombre de relations est égal au nombre d'intermédiaires plus un). On est en général conduit à construire plusieurs fichiers, un pour les personnes interrogées, un autre pour les situations d'accès aux ressources, un autre éventuellement sur les relations citées, etc.

Dans toutes ces approches, il faut pouvoir **passer facilement d'un niveau d'analyse à un autre**. Les logiciels d'analyse statistique permettent en général d'agréger des informations pour basculer sur d'autres unités d'analyse. Par exemple si l'on a une variable sur les types de relation (amis, voisins, famille, etc.) dans le fichier des alters d'une étude de réseaux personnels, on peut comptabiliser la part de chaque type de relation pour chaque identifiant des egos et faire « remonter » l'information dans le fichier des egos. Si tous les egos ont cité au moins un alter, on peut aussi utiliser le fichier des alters pour faire des statistiques sur les egos en pondérant les unités statistiques par 1/NbAlt où NbAlt est le nombre d'alters cités par chaque ego.

Pour illustrer la spécificité de l'usage des statistiques exploratoires dans le cas des études sur les réseaux, je vais prendre l'exemple d'une enquête sur les réseaux personnels effectuée en 2001 dans la région de Toulouse[47]. A des fins de comparaison, cette enquête reprenait en partie une méthode utilisée dans une enquête effectuée en 1977 en Californie[48]. Cette enquête comporte une dizaine de générateurs de noms (voir encadré).

---

[44] Voir par exemple Claude Fischer, 1982, *To Dwell among Friends: Personal Networks in Town and City*, University of Chicago Press.

[45] Quelques souvenirs des maths du lycée devraient vous convaincre que s'il y a n alters, alors le nombre de paires possibles est n(n-1)/2.

[46] Voir Alain Degenne et Michel Forsé, 2004, *Les réseaux sociaux*, Armand Colin ou Stanley Wasserman et Katherine Faust, 1994, *Social network analysis : methods and applications*, Cambridge, Cambridge University Press.

[47] C'est l'une des enquêtes présentées et utilisées dans l'ouvrage Claire Bidart, Alain Degenne, Michel Grossetti, 2011, *La vie en réseau. Dynamique des relations sociales*, Presses Universitaires de France.

[48] Claude Fischer, 1982, *To Dwell among Friends* … op.cit.

Enquête de Toulouse sur les réseaux personnels en 2001

Voici les générateurs de noms utilisés dans l'enquête de Toulouse :

1. « Lorsque des personnes s'absentent de la ville pour un moment, elles demandent parfois à quelqu'un de prendre soin de leur maison – par exemple, d'arroser les plantes, de ramasser le courrier, de nourrir les animaux, ou juste de surveiller. Si vous vous absentiez de la ville, demanderiez-vous à quelqu'un de prendre soin de votre maison pendant ce temps ? »

2. « Certaines personnes ne parlent jamais de leur travail ou de leurs études avec d'autres, ni au travail (ou à la fac) ni à l'extérieur. D'autres personnes discutent de choses comme des décisions qu'ils ont à prendre, des problèmes professionnels qu'ils ont à résoudre, et des manières d'améliorer leur façon de travailler. Y a t-il quelqu'un avec qui vous parlez de votre travail ? »

3. « Durant les trois derniers mois, des ami(e)s vous ont-ils aidé pour des tâches domestiques, comme peindre, déplacer des meubles, cuisiner, laver, ou des réparations majeures ou mineures ? »

4. « Parmi ces activités, quelles sont celles que vous avez pratiquées durant les trois derniers mois ?
- Recevoir quelqu'un chez vous à dîner ou souper
- Aller chez quelqu'un pour dîner ou souper
- Recevoir la visite de quelqu'un
- Aller rendre visite à quelqu'un
- Rencontrer une connaissance hors de la maison (ex: restaurant, bar, parc, club…)
- Autres activités : ……

Si oui, pouvez-vous me dire avec qui vous avez partagé ces activités ? »

5. « Parfois les gens discutent avec d'autres des loisirs ou des passe-temps qu'ils ont en commun. Discutez-vous de ce genre de choses ? Si oui, avec qui le faites-vous régulièrement ? ».

6. « Avez-vous un(e) fiancé(e) ou un(e) meilleur(e) ami(e) que vous voyez très souvent (en dehors du foyer) ».

7. « Lorsque vous avez des problèmes personnels – par exemple, concernant une personne proche ou quelque chose qui vous importe – (…) avec qui en parlez-vous ? ».

8. « Souvent les gens s'appuient sur les conseils de quelqu'un qu'ils connaissent pour prendre des décisions importantes – par exemple, des décisions concernant la famille ou le travail. Y a t-il quelqu'un dont vous considérez sérieusement l'avis pour prendre des décisions importantes ? Si oui, l'avis de qui considérez-vous ? ».

9. « Si vous aviez besoin d'une importante somme d'argent, que feriez-vous – demanderiez-vous à une connaissance de vous la prêter ; iriez-vous demander un crédit à la banque ; ou feriez-vous autre chose ? Que feriez-vous en situation d'urgence – y a-t-il quelqu'un (d'autre) à qui vous pourriez probablement demander de vous prêter une partie ou toute la somme d'argent ? ».

La liste des noms (ou prénoms ou pseudonymes), une fois établie, était soumise aux enquêtés avec la question : « manque-t-il quelqu'un d'important pour vous ? ». 24,1 % des enquêtés n'ont pas rajouté de noms, les autres en ajoutant de 1 à 47, ce qui fait une moyenne globale de 4 noms rajoutés. Ensuite nous soumettions à nouveau la liste complète aux enquêtés en leur demandant de qualifier les relations (« famille », « amis », « voisins », etc.). Enfin, nous construisions pour chaque enquêté un sous-échantillon de 5 personnes citées au maximum (les premiers noms cités en réponse à certains des générateurs), pour lesquelles nous posions des questions complémentaires.

Les 399 personnes interrogées ont cité 10 932 personnes dont 1 624 traitées dans le sous-échantillon ont fait l'objet de questions complémentaires. Les enquêtés ont aussi cité 305 personnes avec lesquelles ils ne sont plus en relation et 249 d'entre elles ont fait l'objet aussi de questions complémentaires. Nous disposons donc de plusieurs ensembles de données : une population de 399 enquêtés, une population de 10 932 relations sociales « actives », dont 1 624 davantage renseignées, et une population de 249 relations « disparues » renseignées elles aussi.

Pour cette enquête, on utilise principalement deux fichiers. Le fichier « egos » comporte 399 lignes comportant 1) des informations sur les personnes interrogées, 2) des informations sur l'ensemble des alters cités (nombre relations familiales, avec des collègues, des voisins, etc., nombre total de relations, densité …) et 3) des informations détaillées sur les relations documentées plus précisément avec un groupe de variables pour la première, un autre groupe pour la deuxième ; etc. A partir de ce fichier, on a fabriqué un fichier d'alters qui comprend une ligne par relation avec 1) les informations sur ego (y compris celles portant sur le réseau dans son ensemble qui avaient été codées directement et 2) les informations sur la relation et sur l'alter concernés.

A partir de ces fichiers on peut s'intéresser aux similarités de niveau d'études entre les personnes interrogées et celles qu'elles ont cité.

Tableau 18. Niveau d'études d'ego et alter (fichier alters)

	niveau d'études d'alter				
niveau d'étude d'ego	inférieur au bac	bac	bac + 2	bac + 4	Total
inférieur au bac	55,7%	21,8%	14,2%	8,3%	100,0%
bac	30,1%	24,7%	26,3%	18,9%	100,0%
bac + 2	25,0%	23,0%	32,0%	20,0%	100,0%
bac + 4	16,8%	12,9%	24,2%	46,1%	100,0%
ensemble	32,7%	20,4%	24,0%	23,0%	100,0%

Source : Enquête RESTIC2001, 1515 relations renseignées.
Lecture 55,7% des relations citées par des personnes ayant un niveau d'études inférieur au bac correspondent à des personnes ayant le même niveau d'études.

Avec un œil un peu exercé, on voit que les proportions situées sur la diagonale sont toujours assez nettement plus élevées que celles de la dernière ligne, donc qu'il y a une corrélation entre le niveau d'études des personnes interrogées et celui de celles qu'elles ont citées. Pour les angoissés, le chi2 de 285,9 est largement significatif.

Il est donc intéressant de calculer une variable indiquant si les niveaux d'études sont similaires ou non.

On peut alors examiner les variations de cette similarité selon le niveau d'études d'ego, ce qui reprend les chiffres de la diagonale du tableau précédent.

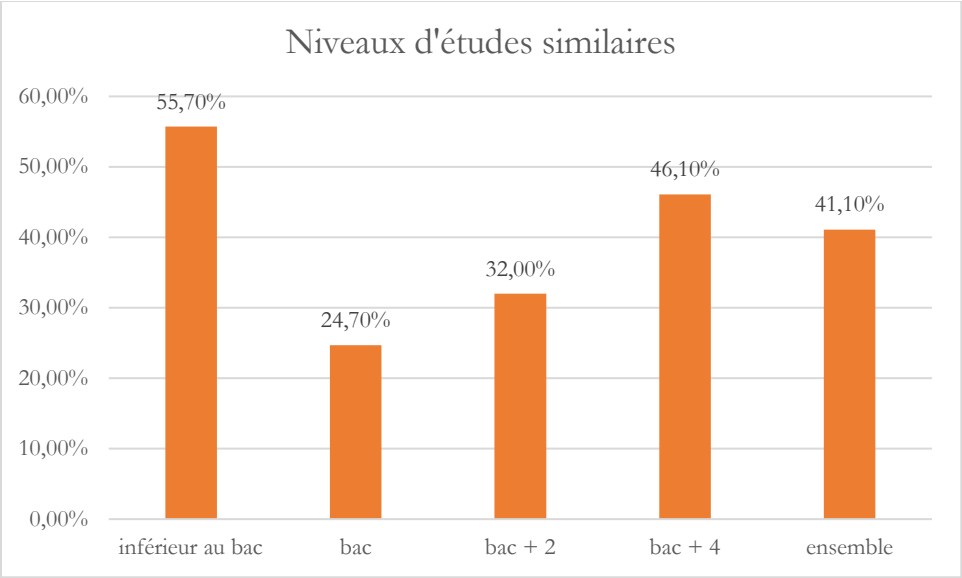
Tableau 19. Similarité de niveau d'études d'ego et d'alter selon le niveau d'études d'ego (fichier alters)

	Niveaux d'études similaires
inférieur au bac	55,7%
bac	24,7%
bac + 2	32,0%
bac + 4	46,1%
ensemble	41,1%

Source : Enquête RESTIC2001, 1515 relations renseignées.
Lecture 55,7% des relations citées par des personnes ayant un niveau d'études inférieur au bac correspondent à des personnes ayant le même niveau d'études.

Ce tableau peut se traduire par un graphique intéressant.

Graphique 25. Similarité de niveaux d'études entre ego et alter selon le niveau d'études d'ego (fichier alters).



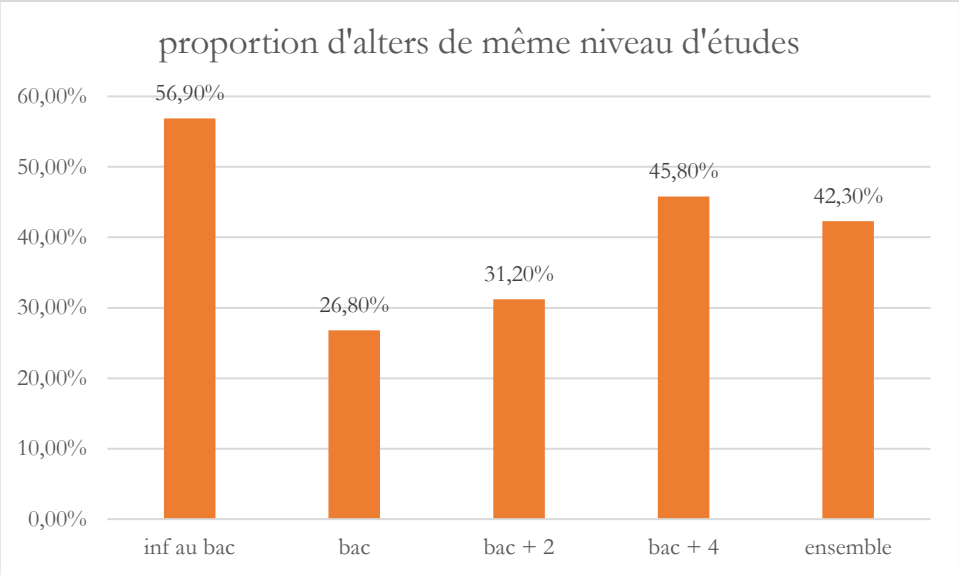
On peut aussi calculer la proportion de relations ayant le même niveau d'études parmi celles citées par chaque ego et documentées pour cette variable et le pourcentage moyen de relations de même niveau d'études dans le fichier des egos.

Tableau 20. Pourcentage moyen de relations avec même niveau d'études par niveaux d'études (fichier egos)

Niveau d'études	pourcentage de relations avec même niveau d'études
inférieur au bac	56,9
bac	26,8
bac + 2	31,2
bac + 4	45,8
Total	42,3

Le nombre de relations citées par personne interrogée et pour lesquelles on dispose des informations détaillées étant peu variable, cela donne à peu près les mêmes résultats. Ils seraient plus différents si le nombre de relations citées variait plus.

Graphique 26. Similarité de niveaux d'études entre ego et alter selon le niveau d'études d'ego (fichier egos).



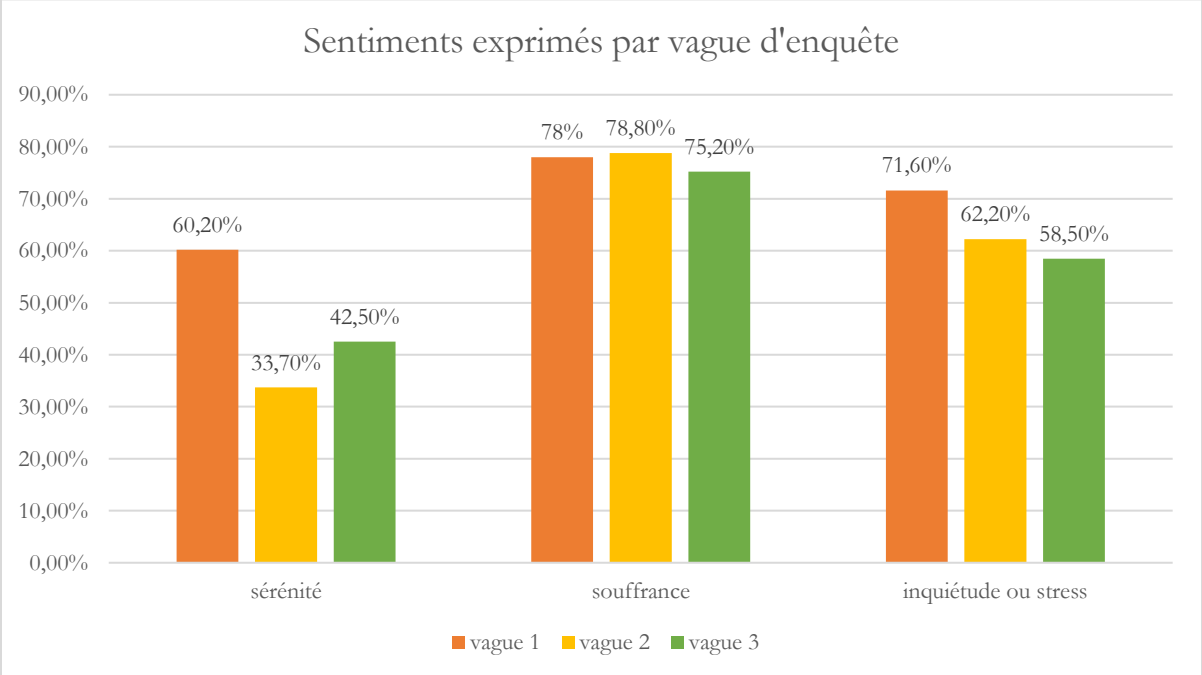
Les deux analyses, à l'échelle des relations (le fichier alters) ou à celle des personnes interrogées (le fichier ego) se complètent et il faut pouvoir passer facilement d'un niveau à l'autre.

8. Analyses de processus

Un processus est un ensemble d'activités auquel est associé une intrigue qui relie ces activités entre elles et avec diverses entités qui y sont impliquées. Cela peut donner lieu à des formes très variées d'enquête et de quantification. Il y a par exemple les enquêtes longitudinales, qui consistent à interroger les mêmes personnes à plusieurs reprises, souvent en posant à chaque fois certaines questions. Il existe aussi des collectes d'informations plus directement sur des processus comme les études de type « chaînes relationnelles » évoquées précédemment. Certains chercheurs tentent de modéliser des processus historiques comme un réseau d'événements[49]. Dans tous les cas, il faut disposer d'informations qui soient situées dans un ordre temporel, que celui-ci prenne la forme d'une inscription dans le calendrier ordinaire ou seulement d'une logique de succession et d'antériorité. Les analystes doivent en général choisir entre des successions de mesures statiques ou des mesures directes du changement, le mieux étant bien sûr de combiner les deux lorsque c'est possible.

Pour s'en faire une idée, je vais revenir aux données de l'enquête « La vie en confinement ». J'ai utilisé dans les chapitres précédents les données du premier questionnaire, rempli durant le premier confinement français. Deux autres questionnaires ont été renvoyés, l'un à la fin de l'année 2020, l'autre un an plus tard, fin 2021 donc, aux personnes qui avaient laissé leur adresse à la fin du premier questionnaire. 2553 personnes ont répondu aux trois questionnaires successifs. Chacun des questionnaires comportait la question sur les sentiments éprouvés que j'ai déjà utilisée comme exemple. J'ai regroupé les réponses en trois grands types que j'ai déjà évoqués : sérénité (« en forme », « détendu », « heureux ») ; souffrance (« fatigué », « irrité », « triste ») ; et enfin inquiétude ou stress.

Graphique 27. Trois types de sentiments selon les vagues

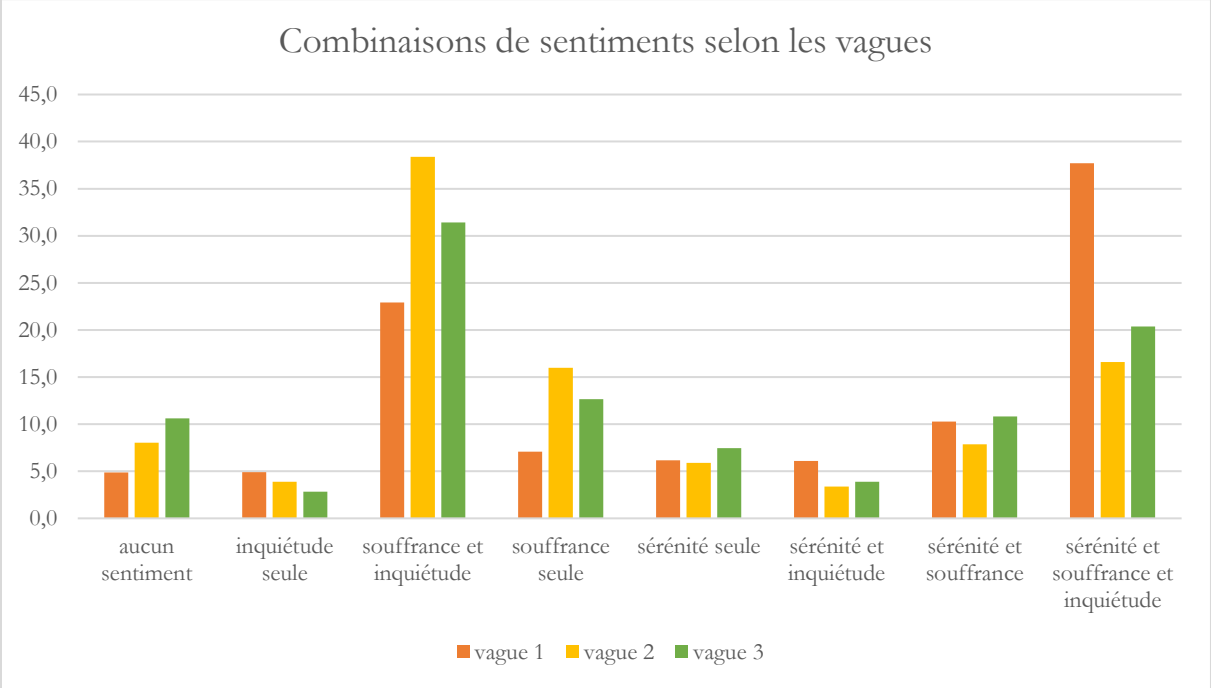


[49] Bearman Peter, Faris Robert, and Moody James, 1999, "Blocking the Future: New Solutions for Old Problems in Historical Social Science.", *Social Science History*, vol. 23, n°4, p. 501-533.

Alors que les expressions de souffrance (essentiellement la fatigue) sont assez stables, celles qui relèvent de la sérénité régressent fortement après le premier confinement avant de remonter un peu. L'inquiétude et le stress régressent régulièrement.

Mais, comme nous l'avons vu, les personnes pouvaient cocher plusieurs réponses. Pour y voir plus clair, on peut donc calculer toutes les combinaisons possibles des trois types (il y en a 2 à la puissance 3, soit 8) et définir ainsi pour chaque personne un « état d'esprit » unique pour chaque vague. On peut ensuite comment cet état d'esprit évolue.

Graphique 28. Combinaisons de sentiments selon les vagues



Ce sont les expressions mélangeant la sérénité avec d'autres types de sentiments qui ont diminué entre la première vague et la deuxième.

Pour avancer dans l'analyse, on peut essayer de voir qui sont les personnes qui ont avaient exprimé de la sérénité durant le premier confinement et ne l'ont plus fait pas la suite. On crée donc une variable selon ce critère. Après examen des corrélations avec diverses autres variables (au moyen d'arbres de décision et de tris à plusieurs entrées), il s'avère que ce sont les personnes qui ont été confinées avec des enfants jeunes dont l'état d'esprit a ainsi évolué d'un mélange de sentiments à l'expression d'une certaine souffrance (essentiellement de la fatigue) : parmi les 31-45 ans, qui est la tranche d'âge la plus concernée par cette évolution (25% contre 12,2% pour les plus de 60 ans), les personnes ayant des enfants sont 28,1% à être dans cette situation contre 20,1% pour celles qui n'en ont pas.

Il est aussi possible d'utiliser la technique de l'« **appariement optimal** » (« *optimal matching* ») que je ne vais pas présenter en détail ici. Cette technique est très bien expliquée dans un article de Laurent Lesnard et Thibaut de SaintPol[50]. Elle consiste à mesurer une distance entre deux séries d'états par une fonction du nombre de changements à effectuer pour passer de l'une à l'autre, puis à combiner cette distance avec un algorithme de classification automatique pour construire des classes. Je l'ai testée sur ces données mais elle ne donne pas des résultats plus intéressants qu'une

[50] Laurent Lesnard et Thibaut de Saint Pol, « Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis) », *Bulletin de méthodologie sociologique*, 90 | 2006, 5-25. Voir également Gabadinho, A., G. Ritschard, N.S. Müller and M. Studer, 2011, « Analyzing and Visualizing State Sequences in R with TraMineR », *Journal of Statistical Software*, 40(4), 1Ű37. DOI 10.18637/jss.v040.i04.

analyse des changements par divers tableaux croisés, ou même que des classifications automatiques plus classiques, probablement parce que trois vagues seulement ne permettent pas de tester vraiment la puissance de l'algorithme.

9. Analyses textuelles

Les textes constituent un matériau important pour les sciences sociales, qu'il s'agisse de ceux qui sont issus d'archives, de transcriptions d'entretiens, de questions ouvertes de questionnaires, ou de commentaires postés en ligne (cette liste n'est pas exhaustive). L'application de méthodes statistiques à ces textes peut donner lieu à des techniques très variées. L'une des plus courantes consiste à construire des fragments de texte (un nombre déterminé de phrases ou de lignes) puis, éventuellement, à réduire les mots à leur racine (« baigneur », « baignade », « bain » se retrouvent dans une catégorie générale « bain+ »). On obtient alors un tableau avec les fragments de textes comme unités statistiques et les racines lexicales comme variables. On peut alors utiliser des techniques de classification automatique pour détecter des registres discursifs.

Pour montrer un exemple, je vais utiliser les textes rédigés par des personnes ayant répondu à l'enquête sur la vie durant le confinement. A la fin du questionnaire, une question ouverte était rédigée ainsi : « Et pour finir, y a-t-il des remarques ou des commentaires que vous voudriez ajouter, au sujet de vos conditions de logement, votre situation de travail, vos activités et vos relations personnels pendant la crise épidémique et le confinement ? N'hésitez pas à utiliser l'espace ci-dessous pour nous en faire part ... ». Mon collègue et ami Julien Figeac a réalisé une première analyse textuelle[51] que je l'ai aidé à compléter. J'en reprends ici quelques éléments.

Le fichier initial contient 3915 réponses (sur 16224 personnes ayant répondu aux autres questions). Après un premier nettoyage (suppression des smileys, des signes de ponctuation (. ; - ; _), des R.A.S, des "Aucun(e)"), le corpus analysé regroupe 3 870 réponses. Le fait de rédiger une réponse est plus fréquent chez les hommes (26% contre 23,4% dans la population totale des 16224), les plus diplômés (25,8%) les retraités (32,2%) et les inactifs (32,2%), les agriculteurs (28,3%), les artisans et commerçants (28,6%), les cadres (27%). Il augmente avec l'âge (de 17,6% à 32,2%). On retrouve certains biais de l'échantillon constitué dans cette enquête en ligne (diplômés, cadres, etc.) qui se trouvent donc renforcés, et d'autres qui vont dans l'autre sens (hommes, agriculteurs …).

Julien a choisi de découper ces réponses en segments (8 112) en se basant sur la ponctuation (, et .). Puis, l'algorithme a analysé et agrégé en groupe (en classes lexicales) les formes (16 162) contenues dans ces segments. Les formes sont issues du processus lemmatisation : une action consistant regrouper les mots d'une même famille (On donne la « forme » canonique d'un mot ou d'un ensemble de mots qui se trouve réduit en une entité appelée en lexicologie lemme ou encore « forme » canonique d'un mot). Les segments de textes contiennent en moyenne 31,9 formes.

La mise en œuvre d'un algorithme de classification hiérarchique descendante[52] conduit à distinguer 18 classes lexicales. Le résultat prend la forme d'un dendrogramme "général" (cf. Figure 1).

[51] Au moyen du logiciel Iramuteq

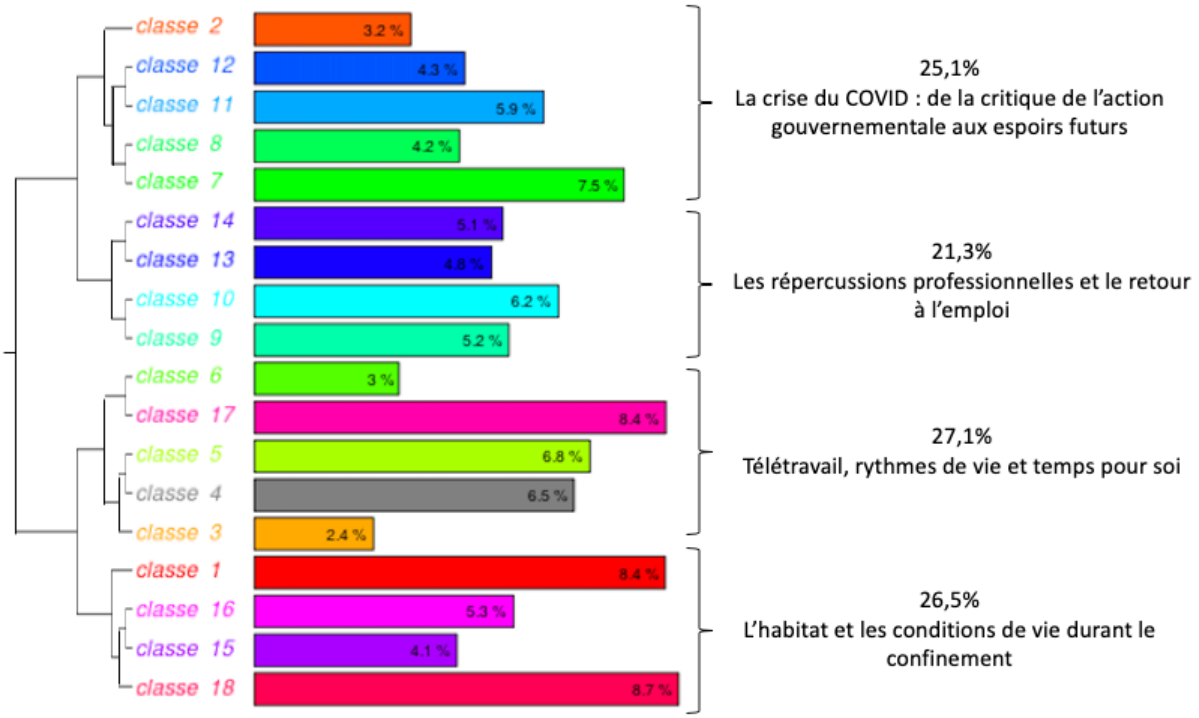
[52] Concrètement, le tableau initial comprend autant de lignes que de segments de textes et autant de colonnes que de formes lexicales. L'algorithme utilise le premier axe d'une analyse des correspondances pour diviser le corpus en deux, puis réitère cette procédure jusqu'à un stabilisation sur le nombre de classes prédéfini.

Figure 3. Dendrogramme général regroupant 18 classes lexicales

Si l'on se fonde sur le dendrogramme, ces classes s'organisent effectivement en 4 groupes principaux structurés autour des classes 1, 2, 3 puis 9 (Cf. la figure 2 ci-dessous).

Ces quatre groupes contiennent un nombre de formes presque équivalent (environ 25% du total ; 16 162). Dans la figure 2 ci-dessous, Julien Figeac a nommé ces quatre groupes de classes.

Figure4. Dendrogramme synthétique hiérarchisant les 18 classes lexicales



La typologie a été ensuite intégrée à un logiciel statistique standard, ce qui permet de calculer des corrélations entre les classes de discours et les caractéristiques des personnes qui ont répondu (et dont les réponses peuvent relever de plusieurs classes, celles-ci étant codées comme autant de variables binaires — présence ou absence). Une analyse rapide des quatre grandes classes regroupées montre les tendances suivantes.

La première classe (critique de l'action gouvernementale) est plus fréquente pour les 46-60 ans (36,4% contre 32,2% dans l'ensemble des personnes ayant laissé un commentaire), d'autant plus si elles se classent politiquement à droite à l'extrême droite (43%).

La deuxième classe (préoccupations relatives à l'emploi) apparaît particulièrement chez les 18-30 ans (42,3% contre 30,7% pour l'ensemble des personnes ayant laissé un commentaire).

La troisième classe (télétravail et rythmes) est particulièrement fréquente chez les diplômés du supérieur (39 % contre 35,9% parmi les personnes ayant laissé un commentaire), d'autant plus si ces personnes ont des enfants (45,3%) et si ce sont des femmes (47,2%).

La quatrième classe (conditions de vie) est quant à elle très fréquente chez les plus 60 ans (46,7 % contre 39,8 % pour l'ensemble des personnes ayant rédigé un texte).

Cette analyse sommaire n'a ici qu'une valeur d'exemple de ce qui est possible avec ce type de corpus.

Conclusion

Cet ouvrage est seulement un guide pour les statistiques exploratoires. L'apprentissage de celles-ci implique de travailler sur des données concrètes. C'est en effet par l'expérience pratique que se construit le savoir-faire dans ce domaine. C'est un savoir-faire d'artisan, qui demande un peu d'imagination et de créativité, et qui est orienté vers la compréhension des données et de leur construction.

La progression que j'ai utilisée, qui commence par une critique des sources et enchaîne ensuite les techniques selon la logique d'une complexification progressive, me semble la plus sûre, mais il est aussi tout à fait possible de commencer par des méthodes multivariées pour resserrer progressivement la focale sur des variables particulières, ou encore jongler avec les différents niveaux d'analyse. Il me semble que le plus important est de multiplier les angles, de tester librement des idées, de prendre les données à bras le corps. Lorsque j'étais étudiant en mathématiques appliquées, nous répétions des propos attribués à Jean-Pierre Benzékri, le créateur de l'analyse factorielle des correspondances, qui aurait dit que les analystes devraient « dormir avec leurs données ». Malgré la différence des méthodes, c'est une attitude qui se rapproche de celle prônée par John Tukey, inventeur entre bien d'autres techniques des « boîtes à moustaches » présentées dans le deuxième chapitre, qui fait partie de ce qu'il appelait l'analyse exploratoire des données. J'ai essayé de montrer aussi que les modèles utilisés le plus souvent dans une logique hypothético-déductive peuvent parfaitement être utilisés dans une logique exploratoire. Bien sûr la logique exploratoire n'exclut nullement de basculer vers une logique probatoire, avec des hypothèses plus formalisées, lorsque les résultats semblent robustes et qu'il faut s'engager dans le débat scientifique.

Je n'ai présenté que les techniques les plus courantes et qui me semblent les plus génériques, il en existe bien d'autres adaptées à des données ou des problématiques plus spécifiques. La lecture de ce guide devra donc certainement être complété des lectures ou des visionnages de présentations de techniques plus spécialisées. De même, je n'ai pas présenté des scripts ou des commandes de logiciels, ce qui présente l'avantage de laisser le lecteur libre de ses choix sur ce plan, mais rendra nécessaire l'apprentissage d'un outil d'analyse statistique. Chacun pourra trouver les compléments nécessaires en fonction de ses choix.

Mon souhait était d'exposer une logique intellectuelle plus que de donner des recettes. J'espère y être parvenu mais il faut évidemment des recettes. Ce livre peut aider les personnes à être réflexives vis-à-vis de celles qui existent et à être capables de créer les leurs propres.