Dictionnaire incomplet d'analyse de données en SHS des ingénieurs de SO-MATé – regards disciplinaires et outils statistiques

Grégoire Le Campion, Solenne Roux, Gautier Debruyne, Gaëlle Deletraz, Laure Gayraud, Claire Kersuzan, Viviane Le Hay, Sandrine Lyser, Delphine Montagne, Karine Onfroy, et al.

Dictionnaire incomplet d’analyse de données en SHS des ingénieurs de SO-MATé – regards disciplinaires et outils statistiques

Qui sommes-nous ?

Notre équipe est constituée :

- **D'une équipe de rédacteurs / relecteurs et gentils penseurs** : Gautier Debruyne ; Gaëlle Deletraz ; Laure Gayraud ; Claire Kersuzan ; Grégoire Le Campion ; Viviane Le Hay ; Sandrine Lyser ; Delphine Montagne ; Karine Onfroy ; Solenne Roux ; Frédéric Santos
- **D'un comité scientifique :** Gaëlle Deletraz, Laure Gayraud, Grégoire Le Campion, Viviane Le Hay, Sandrine Lyser, Solenne Roux, Frédéric Santos
- **De 2 coordinateurs** : Grégoire Le Campion & Solenne Roux

Gautier Debruyne est ingénieur en Production, traitement, analyse de données et enquêtes à l'UMR COMPTRASEC – CNRS, Université de Bordeaux depuis 2010. Il est en charge de l'analyse de données qualitative et quantitative sur des projets de recherche juridique dans les domaines de la santé au travail, l'égalité professionnelle ou encore la traite des êtres humains.

Gaelle Deletraz est ingénieure à l'Université de Pau et des Pays de l'Adour depuis 2009. Docteure en géographie (2002), elle a d'abord été recrutée au sein de l'UMR SET (Société, Environnement, Territoires) où elle a travaillé sur le traitement et l'analyse des données spatiales. En 2016, lors de la création de l'UMR Passages, elle élargit ses compétences à la production, au traitement et à l'analyse de données et d'enquêtes. Depuis 2021, elle exerce à l'UMR TREE (Transitions Énergétiques et Environnementales), unité pluridisciplinaire où elle accompagne les programmes de recherche dans la mise en œuvre d'enquêtes quantitatives et qualitatives, de la conception à la diffusion et à l'analyse, en intégrant également les enjeux liés au RGPD et à la gestion des données. Elle organise régulièrement des formations et ateliers méthodologiques, et anime notamment des sessions dédiées aux logiciels textométriques et aux outils d'aide à l'analyse qualitative (CAQDAS).

Laure Gayraud est ingénieure en Production, traitement, analyse de données et enquêtes au Centre Régional associant le CEREQ au Centre Emile Durkheim (UMR 5116 CNRS), Institut d'Études Politiques de Bordeaux, Université de Bordeaux depuis juillet 1996. Elle réalise des études et des évaluations de politiques publiques sur la décentralisation de la formation professionnelle, les politiques du handicap, la professionnalisation dans l'enseignement supérieur ; la formation tout au long de la vie et pour les derniers thèmes en cours sur la transition écologique et jeunesse ; Territoires Zéro Chômeur Longue Durée.

Claire Kersuzan est ingénieure en Production, traitement, analyse de données et enquêtes. Après un doctorat en démographie (2012) et deux post-doctorats au sein d'équipes pluridisciplinaires (économie, histoire, sociologie, démographie, épidémiologie), elle est, depuis 2022, responsable de la Plateforme universitaire de données de Bordeaux (PUD-Bx), relais local de l'IR* Progedo, hébergée à la Maison des sciences de l'homme de Bordeaux (MSH-Bx). Elle est également ingénieure pour le volet « Formation » de l'Equipex+ LifeObs, coordonné par l'INED, dans le cadre duquel elle développe des kits pédagogiques (au format Quarto) visant à reproduire des résultats scientifiques afin de valoriser et faciliter la prise en main des données collectées dans l'Observatoire des parcours de vie. Elle est enfin ingénieure au sein de l'UMR COMPTRASEC – CNRS, Université de Bordeaux.

Grégoire Le Campion est ingénieur en production, traitement, analyse de données et enquêtes au sein du Laboratoire Passages depuis 2016. Passages est une UMR pluridisciplinaire en SHS avec une majorité de géographes. Il accompagne les chercheurs sur la production et le traitement de leurs données de recherche sur des sujets diversifiés. Pour ce faire il mobilise des méthodes d'analyses de données variées et réalise régulièrement des formations sur ces méthodes auprès de différents publics (étudiants, doctorants, enseignants-chercheurs, ingénieurs). Il est également impliqué dans la promotion et la diffusion du langage de programmation R en étant membre du comité éditorial Rzine.

Viviane Le Hay est ingénieure en Production, traitement, analyse de données et enquêtes au sein de l'UMR CED à l'IEP de Bordeaux. Docteurs en sociologie, elle accompagne chercheurs, enseignants-chercheurs et doctorants sur la production et le traitement de leurs données de recherche principalement en sociologie et science politique. Elle est spécialiste du traitement de données d'enquête par questionnaire et de l'analyse géométrique des données. Elle co-dirige actuellement la revue à comité de lecture bilingue Bulletin de méthodologie sociologique (BMS), qui porte sur les méthodes empiriques de recherche en sociologie, science politique et dans des disciplines proches, et est responsables des enseignements DECA (Données : enjeux, collecte et analyse) du premier cycle de Sciences Po Bordeaux.

Sandrine Lyser est ingénieure en statistique et analyse de données d'enquête au sein du laboratoire de recherche ETTIS - INRAE Nouvelle-Aquitaine Bordeaux depuis 2003. Elle est spécialisée dans la mise en œuvre d'enquêtes socio-économiques interrogeant les pratiques de consommation de biens et de services (loisirs, fréquentation des forêts, fréquentation des plages) et les perceptions de différentes catégories de populations (résidents, touristes, usagers, etc.).

Delphine Montagne est ingénieure d'études en sciences de l'information géographique à l'Université de Pau et des Pays de l'Adour. Elle traite et analyse des données géographiques et de réseau pour un laboratoire de sciences humaines.

Karine Onfroy est ingénieure en production, traitement, analyse de données et enquêtes au sein de l'UMR BSE - CNRS - INRAE, Université de Bordeaux depuis 2013. Elle conçoit et met en œuvre des enquêtes socio-économiques pour les projets de recherche de son unité dans des domaines variés comme les discriminations, les données de recherche, les dynamiques de développement, la précarité alimentaire, etc. Elle accompagne également les membres de son laboratoire vers une science plus ouverte et plus transparente.

Solenne Roux est ingénieure en production, traitement, analyse de données et enquêtes au sein du Laboratoire de Psychologie de l'Université de Bordeaux (LabPsy - UR4139) depuis 2011. Elle accompagne les chercheurs du LabPsy sur la production et le traitement de leurs données de recherche sur des sujets variés tels que : la construction d'outils psychométriques sur la régulation des émotions, la mise en place de programme clinique auprès de chômeurs de longue durée ou encore des recherches fondamentales en psychologie sociale, psychologie clinique et psychopathologie notamment. Pour ce faire elle mobilise des méthodes d'analyses de données

variées. Elle réalise régulièrement des formations sur ces méthodes auprès de différents publics (étudiants, doctorants, enseignants-chercheurs, ingénieurs).

Frédéric Santos est ingénieur d'études en statistique au sein laboratoire PACEA (Université de Bordeaux – CNRS – Ministère de la Culture) depuis 2010. Il effectue modélisations mathématiques et analyses de données pour diverses disciplines en sciences archéologiques ainsi qu'en anthropologie médico-légale. Il est également formateur en langage R et ambassadeur pour l'archive ouverte de code source Software Heritage.

Nous sommes tous et toutes ingénieurs membres du réseau métier SO-MATé, déclinaison locale du réseau métier du CNRS MATE-SHS (pour Méthodes, Analyses, Terrains et Enquêtes en SHS). Le SO de SO-MATé renvoie au Sud-Ouest (de la France), région d'où proviennent les auteurs du dictionnaire. Le réseau SO-MATé a été officialisé en février 2019 et compte actuellement une cinquantaine de membres (principalement des ingénieurs en Sciences Humaines et Sociales - SHS) issus de différentes institutions (Universités, CNRS, INRAE, CEREQ et PUD). Ils proviennent des différentes disciplines des SHS (Anthropologie, Archéologie, Démographie, Droit, Économie, Géographie, Histoire, Psychologie, Science politique et Sociologie) ainsi que de la Statistique.
C'est dans le cadre des échanges au sein de ce réseau qu'est née l'idée de réaliser un panorama des méthodes d'analyse employées en SHS. Cette idée a été motivée surtout par le constat que les méthodes d'analyse franchissaient difficilement les barrières disciplinaires, alors qu'elles pouvaient être utiles à d'autres données et problématiques posées que celles de la discipline au sein de laquelle elles étaient majoritairement employées.

S'il est incomplet, ce dictionnaire permet néanmoins une ouverture à des méthodes d'analyses, parfois peu ou pas connue dans certaines disciplines, uniquement pour des raisons d'habitudes disciplinaires et non pour une réelle inadéquation avec le sujet traité.

Préambule

« Dictionnaire : Recueil des mots d'une langue ou d'un domaine de l'activité humaine, réunis selon une nomenclature d'importance variable et présentés généralement par ordre alphabétique, fournissant sur chaque mot un certain nombre d'informations relatives à son sens et à son emploi et destiné à un public défini. » Cnrtl – Centre National de Ressources Textuelles et Lexicales[1]*.*

Le dictionnaire que vous êtes en train de lire porte sur l'analyse de données en SHS. N'étant pas une encyclopédie, il ne prétend pas à l'exhaustivité : le champ de l'analyse de données est en effet d'une très grande richesse et diversité. En revanche, ce dictionnaire présente les principales méthodes utilisées au quotidien par un ensemble d'ingénieurs en production, traitement et analyse de données en SHS. Cet ouvrage s'adresse prioritairement à des étudiants, ingénieurs, chercheurs et plus largement à des praticiens des sciences humaines et sociales mobilisant des outils statistiques dans leurs travaux, sans nécessairement disposer d'une formation approfondie en statistique théorique. Les différentes entrées du dictionnaire rendent compte de la diversité des approches et des disciplines dans lesquelles s'inscrivent les auteurs. L'objectif est de présenter ces méthodes dans un cadre dépassant les frontières disciplinaires, et de permettre au lecteur de découvrir et de s'approprier des méthodes encore peu utilisées dans son domaine, mais susceptibles de s'avérer très utiles. Pour chaque entrée, un point de vue plus personnel est proposé, afin d'apporter l'éclairage d'un praticien et utilisateur de la méthode considérée.

Cet ouvrage n'est pas non plus un ouvrage méthodologique : il n'a donc pas vocation à présenter l'intégralité du processus de traitement des données. En effet, les phases en amont et en aval du traitement de la donnée ne seront pas abordées. Vous n'y trouverez donc pas les réponses aux questions portant sur la collecte de données (par exemple la constitution d'un échantillon) et la valorisation des résultats (comme leur restitution sous forme graphique, par exemple). Il n'a pas non plus vocation à guider pas à pas le choix d'une méthode ou la construction d'une démarche d'analyse complète, mais plutôt à accompagner et éclairer des pratiques déjà engagées

Les méthodes d'analyses sont volontairement présentées de manière synthétique, avec un accent sur la méthode et les apports et inconvénients de celle-ci, en lieu et place d'un formalisme mathématique, peu présent dans l'ouvrage, mais accessible parmi les nombreuses ressources citées. Dans cette perspective, certaines formulations relèvent de simplifications pédagogiques, de facilités d'expression ou d'abus de langage assumés, fréquents dans les contextes de transmission appliquée des méthodes statistiques, afin de faciliter la compréhension de notions complexes**.** Ainsi, les différentes méthodes présentées permettent d'analyser une, deux ou plusieurs variables simultanément, de décrire ou de comprendre les relations entre ces variables, qui peuvent être de natures différentes. Les lecteurs souhaitant approfondir les fondements théoriques et les formulations rigoureuses sont invités à se référer aux ressources bibliographiques proposées dans chaque fiche.

Ce dictionnaire est donc incomplet, mais nous espérons qu'il sera fort utile à quiconque souhaite explorer le champ des possibles en analyse de données en SHS, qu'il soit étudiant, chercheur ou ingénieur. Il se veut avant tout un outil de travail, de repérage et d'ouverture, fondé sur le partage d'expériences et de pratiques professionnelles.

[1] CNRS et Nancy Université, *Centre National de Ressources Textuelles et Lexicales*, [https://www.cnrtl.fr/].

Comment se repérer dans l'ouvrage :

L'ouvrage se présente comme un dictionnaire. Il est donc organisé par ordre alphabétique. Ce choix implique une consultation non linéaire de l'ouvrage : il suppose soit une connaissance préalable du nom des méthodes, soit une exploration guidée par les outils de repérage proposés
Ce dictionnaire ne comprend pas de table des matières, mais nous avons choisi de représenter l'ensemble des entrées sous la forme de 4 quatre arbres originaux permettant de se repérer dans l'ouvrage et de découvrir de nouvelles méthodes. Ces arbres représentent chacun une catégorie que nous avons définie et qui répond à un grand enjeu de l'analyse et du traitement de donnée.
Les arbres ont été dessinés par l'illustrateur Tom Gauld, auteur et illustrateur britannique, connu pour ses livres humoristiques et ses collaborations avec le New Yorker et le Guardian[2]. Il a reçu de nombreuses récompenses et distinctions (prix Eisner, USA ; sélection officielle au festival d'Angoulême).

Connu pour son intérêt et ses productions dans le champ de la recherche scientifique sur un ton humoristique, nous avons souhaité faire appel à ses services pour offrir une porte d'entrée agréable et instructive à notre dictionnaire. En effet, réaliser un dictionnaire des méthodes d'analyses de données en SHS ne semblait pas renvoyer au livre que nous souhaiterions ouvrir avec un certain plaisir, mais grâce au travail de Tom Gauld et son apport poétique, cet ouvrage se montre dès le début abordable et agréable. Il a en effet choisi de représenter ces arbres selon le rythme de la journée afin d'illustrer la notion de cycle dans l'analyse et le traitement de la donnée.

Chaque entrée est représentée dans un ou plusieurs arbres. Les feuilles (en vert foncé) renvoient toutes à une entrée. Toutefois, elles ne sont pas systématiquement détaillées et les notions sont alors présentées dans une catégorie plus englobante (constituée par les feuilles en vert clair). Les termes spécifiés en **gras** dans chaque entrée renvoient à une autre entrée présente dans le dictionnaire. Les termes indiqués en *italiques* sont détaillés dans la partie notions-clefs – fiches roses.

Une partie *fiches roses*, clin d'œil aux pages dédiées aux locutions latines dans les dictionnaires traditionnels, rassemble des définitions de notions-clefs transversales aux différentes méthodes d'analyse des données abordées dans l'ensemble de l'ouvrage. Cette partie *fiches roses* est volontairement placée en début d'ouvrage afin de présenter les notions essentielles à la bonne compréhension des différentes entrées et de faciliter la lecture et la compréhension du Dictionnaire.

Les intitulés des différentes méthodes d'analyses présentées sont représentés comme dans un dictionnaire linguistique avec la mention de la classe grammaticale et la prononciation phonétique (en Alphabet Phonétique International, avec l'aide de l'intelligence artificielle Le Chat de MistralAI[3]).

[2] Gauld Tom, *Tomgauld.com*, [https://www.tomgauld.com].(site consulté le 04/07/2025)
[3] Mistral AI, *Frontier AI LLMs, assistants, agents, services | Mistral AI - Le Chat (September 2025 version) [Large language model]*, [https://mistral.ai/].[https://chat.mistral.ai/chat] (site consulté le 09/05/2025).

Les Arbres

Se familiariser et connaître

Dans cet arbre nous retrouverons principalement les méthodes permettant de se familiariser avec les données à traiter. Les méthodes d'analyses présentes permettent de réaliser une description de ces données et d'orienter les choix vers d'éventuelles analyses plus complexes afin de répondre de façon appropriée à l'hypothèse posée. Il s'agit très souvent de la première étape essentielle au traitement de données.

Explorer et investiguer

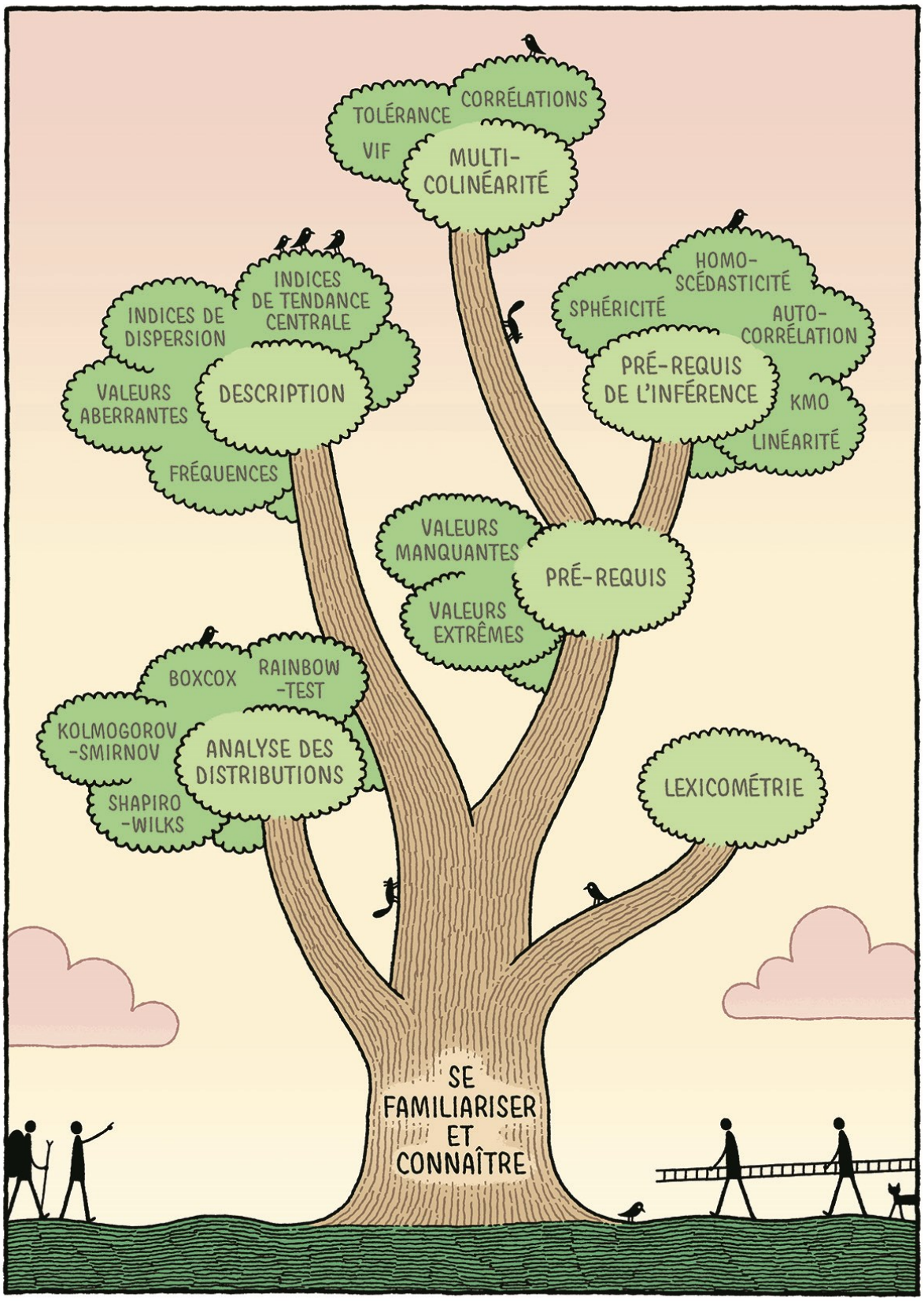
Cet arbre comporte des méthodes d'analyse permettant de réaliser des premiers liens entre des variables. Elles permettent de mettre en lumière des premières relations simples qui pourront être éprouvées par d'autres analyses intégrant des modélisations de relations plus complexes. Une première approche exploratoire permettant souvent d'affiner ses hypothèses.

Structurer et synthétiser

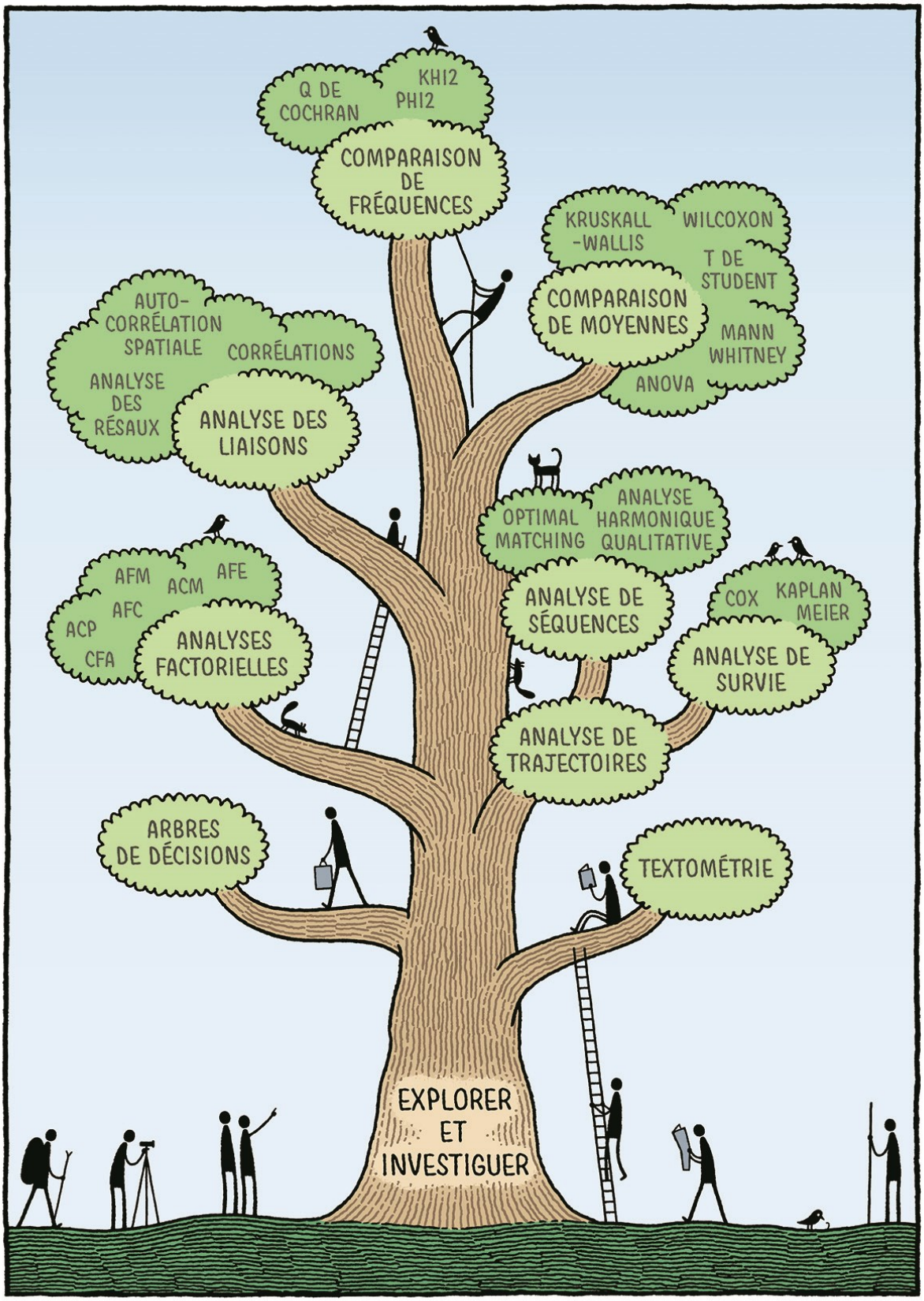
L'idée commune de l'ensemble des méthodes d'analyse présentées dans cet arbre est de réaliser des patterns de variables ou d'individus issus de nos données. Les profils ou patterns alors obtenus constituent des variables synthétiques, résumant un ensemble d'informations. Ces variables synthétiques peuvent être ensuite intégrées à des analyses inférentielles ou multidimensionnelles.

Modéliser et prévoir

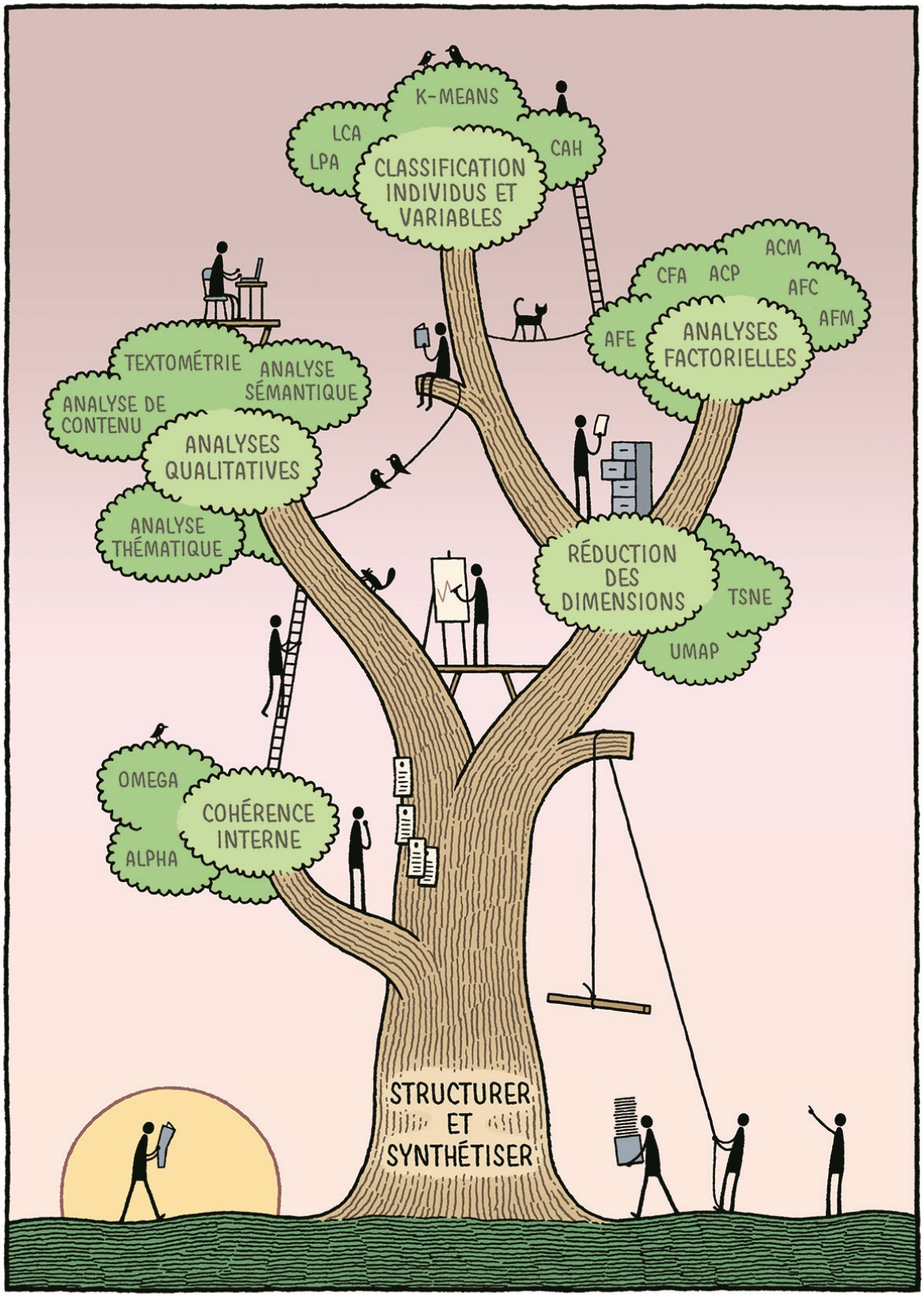
Les méthodes d'analyse présentées dans cet arbre relèvent de modélisations statistiques permettant d'intégrer des relations complexes entre les variables sélectionnées. Certaines d'entre-elles intègrent également les dimensions spatiales et temporelles transversales à l'ensemble des sciences humaines et sociales.



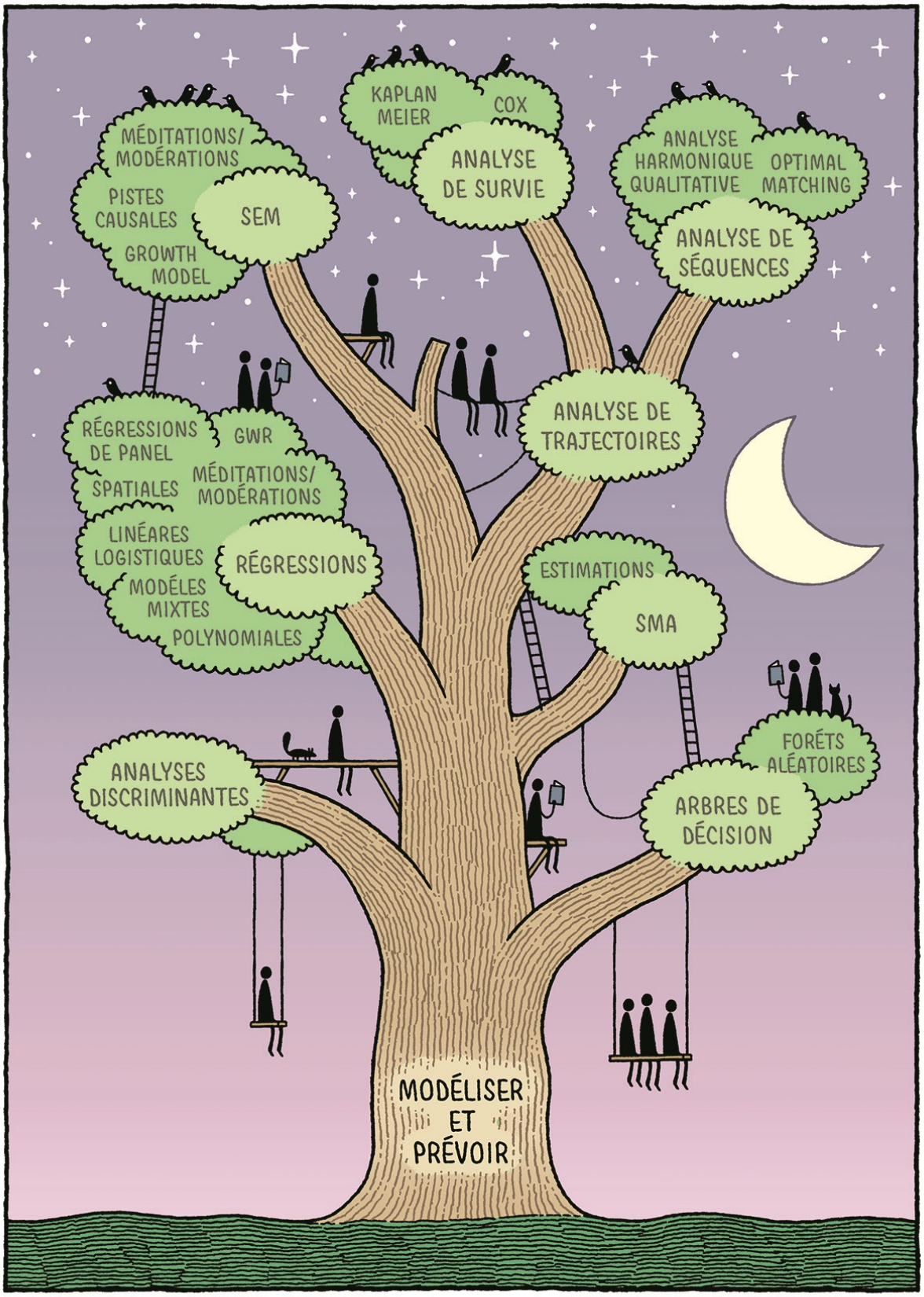
« Arriving at the three. Ready to start Work » Tom Gauld



« Climbing into the three. Exploring the environment » Tom Gauld



« Building structures in the three. Doing their work » Tom Gauld



« Thinking things over whilewatching the stars and the moon » Tom Gauld

Fiches roses

Notions - clefs

Dans cette partie vous retrouverez les notions qui nous paraissent essentielles à la bonne compréhension des différentes entrées du dictionnaire. Elles sont pour la plupart transversales à plusieurs entrées c'est pourquoi nous avons fait le choix de les détailler ici plutôt que dans chaque entrée concernée.

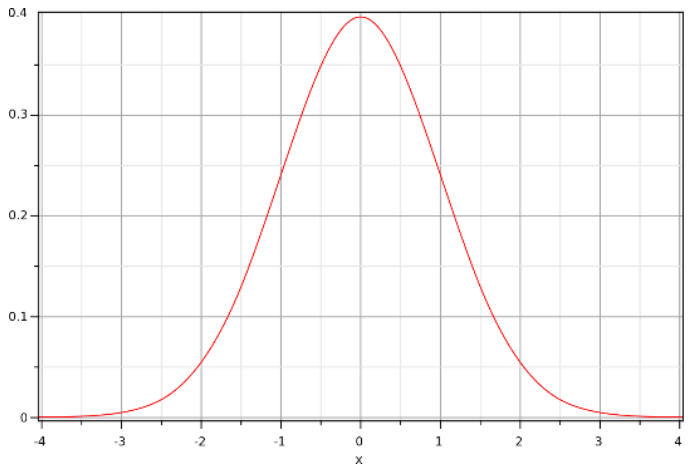
Définitions

Distribution des données et lois de probabilité

Les lois de probabilité (distribution de probabilité) sont utilisées en statistiques et selon les tests employés, ceux-ci ne seront pas basés sur la même loi. Il existe une infinité de distributions de probabilités possibles. Toutefois, certaines sont plus connues, car elles sont plus fréquemment employées ou revêtent une importance particulière dans la théorie. Nous pouvons notamment citer : la loi normale, la loi binomiale, la loi de Poisson, la loi logit-normale, la loi Gamma, la loi du Khi², etc. Différentes méthodes d'analyses de données présentées dans le Dictionnaire reposent sur le principe des lois que nous venons de citer.

Loi normale

La loi normale fait référence à la forme que suit une distribution de données quantitatives. Une distribution de données qui suit spécifiquement une loi normale, a ses données qui se répartissent majoritairement autour de la valeur moyenne et dans des proportions plus faibles à mesure que les valeurs s'éloignent de cette valeur moyenne.



Exemple de distribution suivant une loi normale avec une moyenne à 0.

Avant de réaliser certains tests statistiques qui reposent sur la valeur moyenne, il faut s'assurer que la moyenne de l'échantillon soit bien représentative de la distribution de la variable, qu'il sera alors nécessaire d'examiner. Si la distribution s'approche de la loi normale, alors la majorité des valeurs sont regroupées autour de la valeur moyenne : la moyenne est donc bien représentative de l'échantillon. Si la distribution de la variable s'écarte de la loi normale, il se peut que celle-ci soit altérée par une (ou plusieurs) valeur extrême qu'il faut corriger pour que la valeur moyenne devienne représentative de l'échantillon. Il est aussi possible que la moyenne ne soit pas l'indicateur le plus adapté pour rendre compte des données, et dans ce cas, il faut se tourner vers un autre type de test.

Pour vérifier si des données se distribuent selon une loi normale, il est possible de réaliser des tests statistiques, tels que le test de **Shapiro-Wilk** ou encore celui de **Kolmogorov-Smirnov**, tous deux présentés dans ce Dictionnaire. Il est également possible de procéder à une inspection graphique, procédure, qui n'est en revanche pas présentée dans cet ouvrage.

Modèles de mesures / modèles de structure

Il existe deux types de modèles quand nous effectuons des SEM (Structural Equation Model / Modèles en équations structurelles) :

- Modèle de mesure : Qui va rendre compte de la relation entre les *variables manifestes* et les *variables latentes* du modèle.
- Modèle de structure : Qui va tester la relation entre les différentes *variables latentes*.

Selon le modèle théorique testé, l'un de ces modèles ou les deux peuvent être réalisés.
Les SEM peuvent également permettre de tester des modèles causaux complexes entre variables manifestes (uniquement des variables observées, sans variable latentes donc) qui ne seraient pas possibles (ou plus compliqué) avec les modèles de régression classiques. C'est notamment le cas lors de modèles de **médiation** ou de **modération** qui ne comportent que des variables mesurées directement.

Non-paramétrique

Les tests non-paramétriques sont des méthodes d'analyse statistique qui ne reposent pas sur une hypothèse de distribution des données. Au contraire, les tests paramétriques sont basés sur une *loi de probabilité* induisant une hypothèse spécifique quant à la distribution des données. Les méthodes d'analyses non-paramétriques peuvent donc être employées lorsque la distribution des données ne suit pas celle d'une loi de probabilité sur laquelle repose le test initialement envisagé. Par exemple, afin de réaliser un test de comparaison de moyennes entre deux (ou plusieurs) groupes il est possible d'effectuer une **ANOVA** ou un **test t**. Toutefois, ces deux méthodes d'analyse reposent sur l'hypothèse que les données se distribuent selon une *loi normale*. Il s'agit donc d'un pré-requis à l'utilisation de ces méthodes. Si la distribution des données ne suit pas une loi normale, alors il sera envisageable de recourir aux tests non-paramétriques : dans notre exemple, de **Kruskall-Wallis** (pour comparer plus de deux groupes) ou de **Mann-Whitney** et de **Wilcoxon** (pour comparer deux groupes). Les tests non-paramétriques sont sélectionnés en deuxième choix, après vérification de la distribution des données car ceux-ci ont une *puissance statistique* inférieure aux tests paramétriques. C'est-à-dire, que nous aurons une probabilité inférieure de détecter un effet si celui-ci existe.

Puissance statistique

La puissance statistique est la probabilité de détecter un effet selon la taille de celui-ci et la taille de l'échantillon testé. La puissance statistique peut être calculée a priori, afin d'évaluer la taille d'échantillon adéquate pour que l'effet recherché, s'il existe, puisse ressortir. Ceci nécessite de connaître la *taille d'effet* de l'effet recherché suite à une revue de littérature ou à sélectionner par défaut une *taille d'effet* moyenne, qui par conséquent donnera une estimation moins précise des conditions idéales pour que si effet il y a, il apparaisse. La puissance statistique peut également se calculer a posteriori pour évaluer la robustesse de l'effet obtenu.

Standardiser

Il s'agit de la transformation d'une variable faisant en sorte que sa moyenne soit égale à 0 et son écart-type à 1. C'est une pratique très courante, et souvent recommandée, qui permet ainsi de rendre plusieurs variables comparables malgré des unités initialement différentes tout en évitant de modifier leurs profils respectifs de variation.

Synonymes : Centrer-réduire, z-score

Taille d'effet

Plus un échantillon est grand, plus les moindres effets seront observés, mais ceux-ci peuvent être d'une taille insignifiante. Il est donc important d'évaluer la taille de l'effet observé en calculant un indice de taille d'effet. Lorsque nous calculons un test de la famille des **tests t**, l'indice de taille d'effet choisi est généralement le d de Cohen, mais il existe d'autres indices permettant de rendre compte de la taille d'un effet observé selon le test réalisé. Nous pouvons notamment citer le coefficient de corrélation *r* de Pearson, le coefficient de détermination R^2, l'eta-carré η^2, l'omega-carré ω^2 ou encore les odds-ratio.

Tests d'hypothèse

La plupart des tests présentés dans le dictionnaire reposent sur la notion de test d'hypothèse. Cette notion correspond à une comparaison entre une hypothèse dite *nulle*, aussi appelée H0 et une hypothèse *alternative*, également appelée H1. L'idée est de répondre de façon dichotomique à une question[4],[5] : *les résultats obtenus permettent-ils de rejeter ou de ne pas rejeter l'hypothèse nulle ?* La réponse à cette question est basée sur un seuil à ne pas franchir par une probabilité associée au test concerné. En SHS ce seuil est souvent de .05, ce qui implique que l'hypothèse nulle est considérée comme rejetée si la probabilité associée au test est inférieure à 0.05.

La pratique des tests d'hypothèse a été développée par Fisher (en 1925 puis 1941) en reprenant un ensemble de méthodes basées sur la formulation de Neyman et Pearson (1928) sur la puissance statistique, les erreurs de type I et de type II. Si cette méthode de lecture des résultats est critiquée dès ses origines (Yates, 1951, Bakan 1966 et Tukey 1969), elle reste majoritairement employée

[4] Cohen Jacob, « The earth is round (p < .05). », *American Psychologist*, n° 12, vol. 49, 1994, p. 997-1003, [https//doi.org/10.1037/0003-066X.49.12.997].
[5] Nuzzo Regina, « Scientific method: Statistical errors », *Nature*, n° 7487, vol. 506, 2014, p. 150-152, [https//doi.org/10.1038/506150a].

dans la littérature scientifique en SHS. Pourtant, aujourd'hui encore, son impression de facilité de lecture et d'interprétation conduisent à la production de résultats erronés et de contre-sens[6],[7].

Synonymes : Null hypothesis significance testing (NHST).

Variables dépendantes / Variables indépendantes

Le terme de variables indépendante renvoie à la variable que nous choisissons de faire varier. Nous pouvons également parler de variable manipulée. Elle est généralement notée *x*.

Le terme de variable dépendante fait référence à la variable que nous mesurons. Elle est susceptible de varier selon les variations de la variable indépendante. Elle est alors notée *y*.

Variables à expliquer / Variables explicatives

Les variables explicatives sont celles dont nous supposons qu'elles peuvent fournir une part d'explication d'un phénomène étudiée. Nous pouvons également parler de part de variance expliquée par ces variables lorsque nous mesurons la taille d'effet spécifique de chacune dans un traitement statistique. Les variables explicatives sont similaires aux variables indépendantes.

Les variables à expliquer sont celles dont nous souhaitons évaluer les variations et ce qui influence ces variations. Les variables à expliquer sont également appelées variables dépendantes.

Variables latentes / Variables manifestes

Les variables observées, manifestes ou encore exogènes sont mesurables directement auprès de la population. Si nous prenons par exemple la mesure de l'anxiété, nous ne pouvons pas la mesurer directement. Il est difficile de répondre simplement à la question : "*Etes-vous anxieux ?*". Afin de mesurer l'anxiété des patients, il faudra donc poser des questions qui relèvent de l'anxiété, auxquelles il est possible de répondre simplement telle que : "*Je me sens indécis(e)*". Il s'agit alors d'une variable directement mesurable, elle est donc manifeste. Ces variables sont supposées sous l'influence de variables latentes ou endogènes, qui ne sont pas mesurables directement (comme par exemple l'anxiété) mais qui influencent les variables observées. Ces variables sont inférées à partir d'un ensemble d'indicateurs observables, car elles ne sont pas appréhendables directement. Nous pouvons illustrer notre propos par un autre exemple, portant sur les croyances dans les phénomènes paranormaux. Ces derniers ne sont pas directement appréhendables, mais influencent différentes variables qui elles sont mesurables, telles que le degré d'accord avec les affirmations suivantes : "*Les chats noirs portent malheur*" ou encore "*il est possible de communiquer avec les morts*". Si les participants peuvent indiquer leur degré d'accord avec ces questions (car ce sont des variables manifestes) sous l'influence d'une variable latente (la croyance dans les phénomènes paranormaux), ils auraient été bien incapables de répondre directement, et de façon aussi précise, à cette variable latente, qui ne peut donc être que latente[8].

[6] Ioannidis John P. A., « Why Most Published Research Findings Are False », *PLoS Medicine*, nº 8, vol. 2, 2005, p. e124, [https//doi.org/10.1371/journal.pmed.0020124].
[7] Yaddanapudi LakshmiNarayana, « The American Statistical Association statement on *P* - values explained », *Journal of Anaesthesiology Clinical Pharmacology*, nº 4, vol. 32, 2016, p. 421, [https//doi.org/10.4103/0970-9185.194772].
[8] Cet exemple est tiré de l'échelle permettant de mesurer les croyances dans les phénomènes paranormaux développée par Tobacyk en 1983 puis révisée en 2004. Tobacyk Jerome J., « A Revised Paranormal Belief Scale », *International Journal of Transpersonal Studies*, nº 1, vol. 23, 2004, p. 94-98, [https//doi.org/10.24972/ijts.2004.23.1.94].

Synonymes : Variables endogènes / Variables exogènes

Variables qualitatives

Une variable qualitative est une variable composée de valeurs numériques ou textuelles non quantifiables et hiérarchisables. Les modalités des variables qualitatives sont des catégories pour lesquelles il est possible de calculer des fréquences ou des proportions. Par exemple : les PCS ; la commune de résidence ; la couleur des cheveux, etc.

Les variables ordinales, donc avec des modalités ordonnées sont considérées comme des variables qualitatives dans certaines disciplines. Car si les modalités sont ordonnées, l'écart entre les modalités n'est pas le même, elles ne sont donc pas situées sur un continuum et sont considérées comme qualitatives.

Variables quantitatives

Une variable quantitative est une variable composée de valeurs numériques quantifiables. Les modalités des variables quantitatives suivent un continuum et il est envisageable de calculer des indices de tendances centrales et des indices de dispersion.

Les variables ordinales, donc avec des modalités ordonnées sont considérées comme des variables quantitatives dans certaines disciplines. Car si les modalités sont ordonnées, l'écart entre les modalités est alors le même et se situent sur un continuum.

Fiches entrées

Par ordre alphabétique – aucune cohérence statistique n'a été retenue pour l'ordonnancement de l'ouvrage. Il ne tient sa cohérence qu'au modèle des dictionnaires

ACM n.f. [a.se.ɛm]

Synonyme : Analyse des correspondances multiples

A quoi ça sert ?

L'Analyse des Correspondances Multiples (ACM) est une des rares analyses statistiques que tous les chercheurs en SHS connaissent, même sans la pratiquer. L'ACM est une méthode très connue et particulièrement utilisée pour analyser un jeu de données composé de variables qualitatives (ou quantitatives discrétisées / regroupées en classes). Cette méthode est couramment employée pour l'analyse de jeux de données issues d'enquêtes par questionnaire. Il s'agit d'une méthode de réduction de dimensionnalité, comme l'ensemble des méthodes de la famille des analyses factorielles. C'est-à-dire qu'elle permet de résumer l'information contenue dans un grand nombre de variables nécessairement corrélées en un nombre réduit de dimensions non corrélées.

L'ACM est une généralisation de l'analyse factorielle des correspondances (AFC) au cas de tableaux de données croisant individus statistiques (en lignes) décrits par un ensemble de plusieurs variables qualitatives ou traitées comme telles (en colonnes). Toutefois, nous la considérons bien comme une méthode en soi du fait de ses propriétés spécifiques et des résultats qu'elle fournit.

Dans l'exemple d'une enquête par questionnaire, l'ACM est une méthode descriptive visant à résumer l'information contenue dans un grand nombre de variables. Elle permet de comprendre comment les individus se rapprochent ou s'opposent entre eux dans la façon dont ils combinent leurs réponses à l'ensemble des questions considérées. Nous cherchons à savoir quelles sont les modalités de réponses liées, autrement dit, qui font « pattern », entre elles. L'intérêt de l'ACM repose sur la possibilité d'étudier simultanément les données selon différents prismes : celui des variables ou des modalités (en étudiant les relations qui apparaissent entre elles et la variété des combinaisons existantes) ainsi que celui des individus (en explorant les proximités entre individus). L'étude de ces 3 projecteurs différents d'un même objet va de fait entraîner des éclairages différents, et souligner les principales polarisations/oppositions entre les individus et les variables.

Sans entrer dans le détail mathématique, l'ACM va synthétiser l'information contenue dans le tableau de données en recherchant des axes dits « factoriels » qui résument au mieux les relations entre modalités et entre individus. Ces axes (dimensions), qui sont 2 à 2 orthogonaux, maximisent la dispersion du nuage de points. Leur interprétation repose sur l'analyse des coordonnées, des contributions et des qualités de représentation des variables, des modalités et des individus sur ces axes. En pratique, on n'interprète souvent que le premier plan factoriel, constitué par les deux premiers axes factoriels.

D'où ça vient ?

L'ACM partage la même histoire que l'analyse factorielle des correspondances (AFC). Elle est l'héritière des premiers travaux de Spearman et elle est issue des travaux de « l'école française d'analyse des données » réunie autour du mathématicien Jean-Paul Benzecri. L'AFC et l'ACM sont nées à cette époque aussi grâce aux progrès de l'informatique qui permettaient dorénavant d'envisager la réalisation des calculs jusqu'à alors inenvisageables à la main. Avec ces méthodes Benzecri souhaite refonder la pratique des statistiques, notamment en sciences humaines. L'AFC et l'ACM lui paraissent préférables aux analyses des relations statistiques entre des variables prises deux par deux (en y projetant parfois des relations de causalité non justifiées).

Ces méthodes vont connaître un essor extrêmement important dans toutes les disciplines des SHS et restent encore aujourd'hui extrêmement utilisées. L'ACM en particulier doit aussi son succès à son utilisation par Pierre Bourdieu dans son ouvrage la Distinction, ce qui a permis aussi de diffuser son usage dans toutes les disciplines des SHS.

Exemple d'application

En géographie, l'ACM est très souvent utilisée, notamment pour traiter de manière statistique et non géographique la proximité de variables et modalités spatiales. Rarement utilisée seule, elle sert souvent d'étape préalable à une méthode de partitionnement des individus (**classification ascendante hiérarchique - CAH**, par exemple).

Comme exemple d'usage publié nous pouvons citer les travaux de Guyot et al. (2020) qui étudient les sites artistiques caractérisés par différentes variables comme éléments géographiques, financiers, artistiques, d'accueil du public. Dans cet article, les auteurs utilisent le couple ACM/CAH où l'ACM est utilisée dans sa dimension d'analyse des variables et des modalités.

L'ACM montre une structuration des sites en fonction (1) d'éléments liés à la mise en art et à la politique artistique des sites étudiés (axe 1) et (2) de politique territoriale (axe 2). Ce qui joue ce n'est pas tant les dimensions géographiques que le type de territoire et les politiques associées.

Pour des exemples plus typiques, l'ouvrage de Husson, Lê et Pagès (2009, rééd 2016) présente de nombreux cas illustrés, facilement compréhensibles.

Mot du praticien :

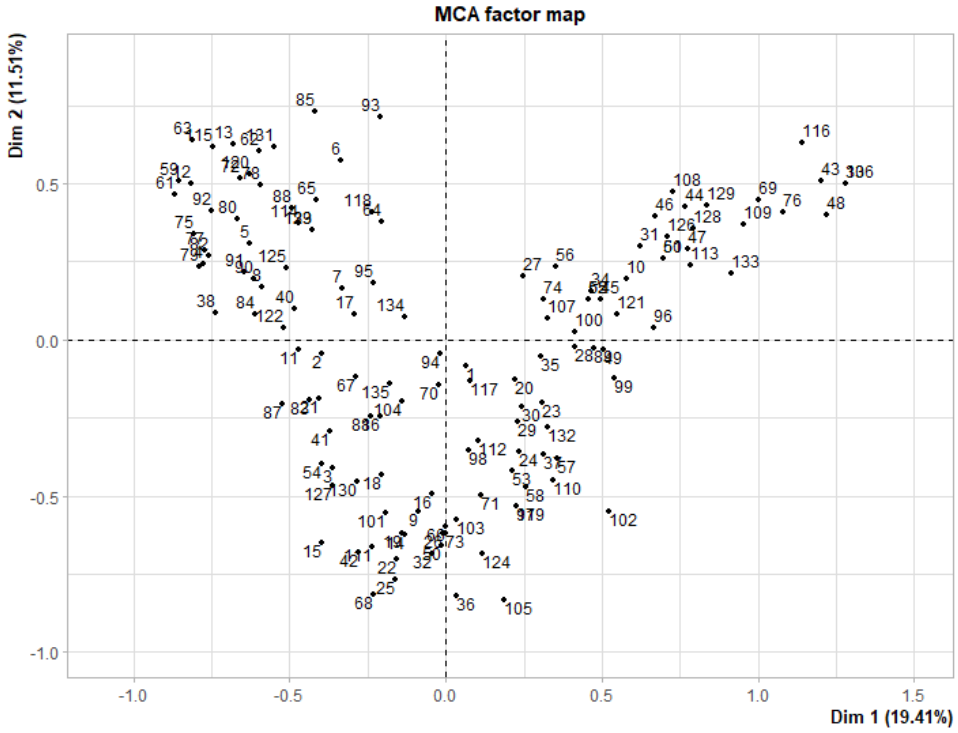
L'ACM est particulièrement sensible aux modalités rares qui contribuent significativement à l'inertie totale, ce qui peut facilement biaiser les analyses.

Il existe plusieurs solutions pour traiter de cette problématique des modalités rares :

- Soit opérer des regroupements naturels, si cela a du sens. Typiquement le cas de modalités ordonnées
- Soit ventiler aléatoirement ces modalités dans les autres modalités
- Soit supprimer les individus ayant ces modalités.

Dans le cas d'une ACM les nuages de points peuvent prendre une forme de « U » ou de « V ». Cette forme en parabole est appelée « effet Guttman » du nom du sociologue Louis Guttman (1916-1987) connu pour ses recherches méthodologiques, en particulier sur les échelles de réponses. Ce genre de configuration, que l'on retrouve souvent en SHS, se produit quand il y a des liens multiples entre les réponses, autrement dit lorsqu'il y a une redondance entre les variables qui ont servi à la construction du plan factoriel. Pour rendre compte de ces fortes liaisons, l'analyse des correspondances construit un premier axe d'opposition entre très en accord et très opposés, à un deuxième axe qui, artificiellement, oppose les positions extrêmes aux positions moyennes.

Graphique des individus en « V » caractéristique d'un effet Guttman (Husson, Lê et Pagès - 2009, rééd 2016)



L'analyse des correspondances multiples a beaucoup de succès car elle donne une impression de facilité d'utilisation et d'interprétation des résultats. C'est bien sur trompeur, l'analyse d'une ACM peut s'avérer très complexe et les données peuvent ne pas être adaptées à cette analyse.

Comme cela a été évoqué, l'ACM est une méthode qui est très sensible aux modalités rares, aux effectifs déséquilibrés ou aux données manquantes. De façon empirique, on considère que des modalités qui contiennent (peu ou prou) moins de 5% de l'échantillon sont rares. Par ailleurs, il est possible que nos données ne soient pas adaptées à ce genre de méthode. On utilisera alors les tests du KMO et de Bartlett pour le vérifier. Il est également très important d'étudier les liaisons entre les variables pour étudier le risque de multicolinéarité, en ayant par exemple recours à l'analyse du VIF (Variance Inflation Factor).

Ceci dit l'ACM est une méthode riche et particulièrement intéressante qui permet d'exploiter des données qualitatives et donne toute sa place à la connaissance du sujet et des données sur l'interprétation des axes factoriels. Elle peut également servir d'étape préliminaire pour des données qualitatives avant d'utiliser une méthode de classification, notamment la CAH.

L'ACM est très pratiquée en sociologie et science politique, plus spécifiquement en France, du fait en particulier des travaux de P. Bourdieu qui a eu recours à cette méthode dans plusieurs de ces ouvrages autour de son concept de champ (par exemple : La Distinction, Homo academicus, La Noblesse d'Etat, mais aussi dans ses études sur la Kabylie). Elle est employée depuis au-delà de la notion de champ dans de nombreux domaines de ces deux disciplines (Tel qu'en sociologie de la consommation (Ferry, M., 2025) ; de la culture (Coulangeon, P., 2021) ; des classes sociales (Savage, M., et al., 2013) ; ou encore du vote (Perrineau, P., et al., 2000)).

Elle s'est également développée grâce aux travaux des mathématiciens Henry Rouanet et Brigitte Le Roux (cette dernière ayant réalisé sa thèse avec Benzécri) et leur collaboration d'une part avec P. Bourdieu à la fin de sa vie (qui leur a ouvert des recherches avec d'autres sociologues en France et à l'étranger), d'autre part avec plusieurs politologues. Ils ont dans ce cadre développé une nouvelle méthode, l'analyse des correspondances multiples spécifique, qui propose un traitement statistique particulier (spécifique) aux 'modalités rares' (le plus souvent les non réponses, mais également d'autres modalités peu choisies par les répondants à un questionnaire). Avec cette méthode, l'ensemble de l'échantillon est conservé, ce qui est primordial dans une analyse sociologique ou politologique dans la mesure où les non-répondants présentent des profils particuliers et différents du reste de l'échantillon (les supprimer revient donc à biaiser considérablement la qualité/représentativité d'un échantillon ; à l'inverse les conserver revient à créer artificiellement des polarités autours de ces modalités peu attribuées et qui s'éloignent par construction des autres).

Cette méthode (l'ACM spécifique) est accessible via le logiciel SPAD, mais également en s'appuyant sur le package R GDAtools réalisé par Nicolas Robette.

Pour la légende, le terme 'Analyse des correspondances multiples' aurait été inspiré par Benzécri du sonnet « Correspondances » de C. Baudelaire (où « Les parfums, les couleurs et les sons se répondent », à l'instar des individus, des variables et des modalités).

Dans certaines pratiques, pour le traitement de données d'enquêtes par questionnaire, en économie ou sociologie, l'ACM est remplacée par une méthode de classification de variables, la méthode ClustOfVar, développée par Chavent et al. (2012). Cette méthode propose un algorithme de partitionnement de type k-means et un algorithme ascendant hiérarchique, qui visent à maximiser un critère d'homogénéité fondé sur le rapport de corrélation pour des variables qualitatives. Ces algorithmes n'imposent pas de contraintes d'orthogonalité entre les variables synthétiques, ce qui facilite l'interprétation par rapport aux axes factoriels. Un exemple d'application est présenté dans Kuentz-Simonet et al. (2013).

Ressources :

- Benzécri Jean-Paul, *L'Analyse des données*, Paris Bruxelles Montréal, Dunod, 1973.
- Bourdieu Pierre, *La distinction critique sociale du jugement*, Paris, Editions de Minuit : Maison des sciences de l'homme, coll. « Le Sens commun », 2012.
- Chavent Marie, Kuentz-Simonet Vanessa et Saracco Jérôme, « Orthogonal rotation in PCAMIX », *Advances in Data Analysis and Classification*, n° 2, vol. 6, 2012, p. 131-146, [https//doi.org/10.1007/s11634-012-0105-3].
- Cibois Philippe, *Les méthodes d'analyse d'enquêtes*, Lyon, ENS Éditions, 2014, [https//doi.org/10.4000/books.enseditions.1443].
- Cibois Philippe, « Les pièges de l'analyse des correspondances », *Histoire & Mesure*, n° 3, vol. 12, 1997, p. 299-320, [https//doi.org/10.3406/hism.1997.1549].
- Cibois Philippe, « L'analyse des correspondances : l'indispensable retour aux données », *Histoire & Mesure*, n° 3, vol. 1, 1986, p. 239-247, [https//doi.org/10.3406/hism.1986.1540].
- Coulangeon Philippe, « Chapitre 8 - La recomposition des structures sociales du goût et des attitudes culturelles », *Culture de masse et société de classes : Le goût de l'altérité*, Presses Universitaires de France., 2021, p. 241-272.
- Ferry Mathieu, « Le prix du végétarisme. Légitimité et autonomie culturelle de la caste en Inde contemporaine », *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, n° 1-2, vol. 165-166, 2025, p. 110-153, [https//doi.org/10.1177/07591063251317058].

- Guyot Sylvain, Le Campion Grégoire et Pissoat Olivier, « Diversité et enjeux territoriaux de la mise en art des espaces périphériques dans le monde », *Cybergeo*, , 2020, [https//doi.org/10.4000/cybergeo.35837].
- Husson François, Lê Sébastien et Pagès Jérôme, *Analyse de données avec R*, 2e éd. revue et Augmentée., Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2016.
- Kuentz Simonet V., Lyser Sandrine, Candau Jacqueline, Deuffic Philippe, Chavent Marie et Saracco Jérôme, « Une approche par classification de variables pour la typologie d'observations : le cas d'une enquête agriculture et environnement », *Journal de la Societe Française de Statistique*, nº 2, vol. 154, 2013, p. 37-63.
- Le Roux Brigitte et Rouanet Henry, *Multiple Correspondence Analysis*, 2455 Teller Road, Thousand Oaks California 91320 United States of America, SAGE Publications, Inc., 2010, [https//doi.org/10.4135/9781412993906].
- Lebart Ludovic, Piron Marie et Morineau Alain, *Statistique exploratoire multidimensionnelle: Visualisation et inférences en fouille de données*, 4e éd (2006)., Paris, Dunod, coll. « Sciences SUP », 1995.
- Makowski Dominique, « The psycho Package: an Efficient and Publishing-Oriented Workflow for Psychological Science », *The Journal of Open Source Software*, nº 22, vol. 3, 2018, p. 470, [https//doi.org/10.21105/joss.00470].
- Pages J. P., Cailliez F. et Escoufier Y., « Analyse factorielle : un peu d'histoire et de géométrie », *Revue de Statistique Appliquée*, nº 1, vol. 27, 1979, p. 5-28.
- Perrineau Pascal, Chiche Jean, Le Roux Brigitte et Rouanet Henry, « L'espace politique des électeurs français à la fin des années 1990. Nouveaux et anciens clivages, hétérogénéité des électoratsé », *Revue française de science politique*, nº 3, vol. 50, 2000, p. 463-488, [https//doi.org/10.3406/rfsp.2000.395484].
- Robette Nicolas, *GDAtools: Geometric Data Analysis in R. version 2.0.*, [https://nicolas-robette.github.io/GDAtools/].
- Savage Mike, Devine Fiona, Cunningham Niall, Taylor Mark, Li Yaojun, Hjellbrekke Johs, Le Roux Brigitte, Friedman Sam et Miles Andrew, « A New Model of Social Class? Findings from the BBC's Great British Class Survey Experiment », *Sociology*, nº 2, vol. 47, 2013, p. 219-250, [https//doi.org/10.1177/0038038513481128].
- Spearman C., « The Proof and Measurement of Association between Two Things », *The American Journal of Psychology*, nº 72-101, vol. 15, 1904, [https//doi.org/10.2307/1422689].

ACP n.f. [/a se pe/]

Synonyme : Analyse en Composantes Principales, Analyses Géométriques

A quoi ça sert ?

L'Analyse en Composantes Principales (ACP) est une méthode de réduction de dimensionnalité, c'est-à-dire qu'elle permet de résumer l'information contenue dans un grand nombre de variables nécessairement corrélées en un nombre réduit de dimensions non corrélées appelées dans le cadre de l'ACP : composantes principales. L'ACP est utilisée exclusivement sur des variables quantitatives.

A partir de ce concept de base, l'analyse en composantes principales va s'avérer très utile pour :

- Détecter des structures dans nos données en mettant en évidence ces fameuses composantes principales. Autrement dit, cette détection de structure va permettre de simplifier les données en transformant un grand nombre de variables en quelques composantes.
- Visualiser les données grâce à l'apport de graphiques qui viennent faciliter la compréhension et illustrer les différentes relations qui structurent notre jeu de données.

Par ailleurs, l'ACP peut aussi jouer un rôle dans le pré-traitement des données, notamment pour réduire le bruit ou diminuer la multicolinéarité de nos données en amont des méthodes de régression ou de classification par exemple.

D'où ça vient ?

Les premiers travaux en matière d'analyse factorielle remontent au tout début du xx^{e} siècle, notamment en 1901 avec les travaux du mathématicien anglais Karl Pearson qui développera sa réflexion sur les analyses en composantes principales (ACP) (Pearson F.R.S., K., 1901).
Le véritable essor de ces méthodes en SHS va notamment se faire avec les travaux du mathématicien français Jean-Paul Benzecri (Benzecri J.-P., 1973), dans les années 70 (Pages, J-P., et al., 1979). En particulier grâce à l'apport des représentations graphiques qui permettent de synthétiser et illustrer les résultats.

Exemple d'application

L'ACP peut être utilisée par exemple pour faire émerger des composantes principales qui viendraient synthétiser l'information entre des communes et constituer une étape préliminaire à une **CAH** comme par exemple dans l'article de Antolinos-Basso et al. (2020). Cette utilisation permet notamment de réduire un certain nombre de biais qui pourraient nuire à la CAH comme par exemple les risques de multicolinéarité.

En Sociologie, l'ACP peut par exemple être utilisée afin d'appréhender le rapport des français à la fiscalité selon le type de taxes et d'impôts abordés. Brouard et Le Hay (2012)[9] ont mobilisé une Analyse en Composantes Principales sur cette thématique et ont obtenu deux composantes. L'une renvoie au clivage sur la fiscalité redistributive et la seconde porte sur la question du désendettement.

[9] Sylvain Brouard, Viviane Le Hay. Les Français et la fiscalité. 2012, 12 p. ⟨halshs-00718416⟩

Mot du praticien

En Psychologie, nous n'utilisons que très rarement l'ACP, car celle-ci a une dimension exploratoire qui n'est pas souvent recherchée. Nous lui préférons **l'Analyse Factorielle Exploratoire (AFE)** qui va tester l'agrégation de variables manifestes en variables latentes, issues d'un modèle théorique défini au préalable.

L'AFE est une méthode d'analyse cousine de l'ACP, une des questions essentielles à se poser pour choisir entre ces deux méthodes, c'est de savoir que voulez-vous faire de la variance de vos données et si vous souhaitez étudier des facteurs latents ou des composantes principales.

Ressources

- Antolinos-Basso Diégo, Blanc Nathalie, Chiche Jean et Paddeu Flaminia, « S'engager pour l'environnement dans le Grand Paris : territoires, politiques et inégalités », *Cybergeo: European Journal of Geography*, , 2020, [https//doi.org/10.4000/cybergeo.34544].
- Benzécri Jean-Paul, *L'Analyse des données*, Paris Bruxelles Montréal, Dunod, 1973.
- Brouard Sylvain et Le Hay Viviane, « Les Français et la fiscalité », , 2012, p. 12 p.
- Husson François, Lê Sébastien et Pagès Jérôme, *Analyse de données avec R*, 2e éd. revue et Augmentée., Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2016.
- Pages J. P., Cailliez F. et Escoufier Y., « Analyse factorielle : un peu d'histoire et de géométrie », *Revue de Statistique Appliquée*, n° 1, vol. 27, 1979, p. 5-28.
- Pearson Karl, « LIII. *On lines and planes of closest fit to systems of points in space* », *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, n° 11, vol. 2, 1901, p. 559-572, [https//doi.org/10.1080/14786440109462720].

AFC n.f. [/a ɛf se/]

Synonyme : Analyse factorielle des correspondances

A quoi ça sert ?

L'Analyse Factorielle des Correspondances (AFC) est une technique très utile pour explorer et visualiser les relations entre deux variables catégorielles dans un tableau de contingence (tableau croisé). Elle s'appuie sur la statistique du Khi² pour mesurer sa dépendance entre ces variables. Elle est largement utilisée pour détecter des patterns et des associations dans des données qualitatives, en offrant une représentation graphique qui facilite l'interprétation des résultats.

Elle permet de représenter les lignes (les modalités d'une première variable) et les colonnes (les modalités d'une seconde variable) du tableau de contingence dans un même espace, facilitant l'interprétation des relations entre ces modalités.

L'AFC va ainsi permettre de répondre à des questions comme :

- Y a-t-il des lignes de mon tableau qui se ressemblent ou au contraire qui s'opposent ?

- Y a-t-il des colonnes de mon tableau qui se ressemblent ou au contraire qui s'opposent ?

- Y a-t-il des associations de modalités de mes variables qui s'attirent (effectif conjoint particulièrement élevé) ou qui se repoussent (effectif conjoint particulièrement faible) ?

Finalement la finalité de l'AFC est de faire une typologie des lignes et des colonnes de son tableau et analyser le lien entre ces deux typologies. En cela, les objectifs de l'AFC et de l'ACP sont similaires.

C'est une analyse statistique particulièrement utilisé en SHS car elle permet d'étudier des données qualitatives mais aussi de traduire des données ou des relations complexes entre les données en une série de tableaux plus simples et illustrés par des graphiques.

D'où ça vient ?

Les premières réflexions sur les analyses factorielles remontent au tout début du xx^{e} siècle avec Spearman (1904) Mais c'est notamment avec les travaux du mathématicien français Jean-Paul Benzecri (Benzecri J.-P., 1973), dans les années 70, que ces méthodes vont connaître leur essor en France (Pages, J-P., et al., 1979). En particulier grâce à l'apport des représentations graphiques qui permettent de synthétiser et illustrer les résultats. Benzecri va développer les méthodes sortant du modèle gaussien en prenant en compte les variables catégorielles, telles que les analyses factorielles des correspondances (AFC) et les analyses factorielles des correspondances multiples (ACM) (Pages, J-P., et al., 1979). Développée essentiellement par J.-P. Benzecri durant la période 1970-1990, les AFC et ACM ont également connu par la suite beaucoup de développement. Par ailleurs, l'usage par Pierre Bourdieu de l'analyse factorielle des correspondances multiples dans son ouvrage La Distinction publié en 1979 qui porte sur les goûts culturels des différents groupes sociaux en France va contribuer à la renommée des méthodes développé par Benzecri (AFC et ACM).

Exemple d'application :

Nous trouvons sur internet un très grand nombre de tutoriels et d'exemples d'usage de l'AFC. Au niveau académique nous recommandons l'excellent ouvrage de Husson et al. (2009) qui reprend,

de manière détaillée tous les éléments théoriques concernant les principales analyses factorielles à savoir (analyse en composantes principales, analyse factorielle des correspondances et analyses des correspondances multiples) ainsi que de nombreux exemples d'analyses avec le code R associés pour les reproduire.

Nous reprendrons ici de manière synthétique l'exemple détaillé dans cet ouvrage sur la cause de mortalités des français. Les données utilisées sont fournies par le centre d'épidémiologie sur les causes médicale de décès (Cépidc). Ici l'objectif de cet exemple est d'étudier la liaison entre l'âge et les causes de décès. Pour réaliser une AFC, la variable âge a été discrétisée et découpée en intervalles. Nous pouvons noter que l'AFC fonctionnera même si la relation entre les deux variables est non linéaire.

Cet exemple, illustre le fait que l'AFC peut s'appliquer sur des tableaux de contingences complexes et conçus à partir de variables construites. Ainsi en colonne nous avons les classes d'âges définis par les chercheurs. Nous avons 12 colonnes avec des classes d'âges de 10 ans excepté pour les enfants les plus jeunes où on a deux classes d'âge (0-1 ans et 1-4ans) ce choix est justifié par des causes de décès très spécifique comme par exemple la mort subite du nourrisson, et pour les personnes les plus âgés où nous avons la classe 95ans et plus afin de regrouper le très grand âge et éviter des classes avec des effectifs trop faibles.

En ligne, nous retrouvons les causes de décès pour les années 1979 et 2006, ainsi que les différentes causes de décès, et enfin les années de 1979 à 2006. Il y a donc un total de 222 lignes. Les lignes années, et causes de décès pour les années 1979 et 2006 seront désignées comme lignes supplémentaires, cela implique qu'elles ne participeront pas à la construction du plan factoriel mais permettrons une analyse plus fine et notamment voir s'il y a eu une évolution des profils typiquement une cause de décès caractéristique d'une tranche d'âge en 1979 qui le serai moins en 2006.

Les auteurs vont bien sur analyser les marges du tableau ce qui permet d'obtenir des informations sur les causes de décès les plus fréquentes ou les classes d'âges avec un plus grand nombre de décès. Si on voit bien sur un effet d'âge avec plus de décès chez les personnes âgées on note cependant que la classe d'âge des 0-1an est relativement plus importante par rapport aux tranches d'âge suivantes (1-4ans et 5-14ans).

La réalisation du chi² permet de mettre à jour une relation significative entre les variables. Aussi, pour étudier l'intensité de cette relation on pourra utiliser le V de Cramer.

Vient ensuite, l'enjeu autour de la sélection du nombre de dimensions à retenir pour l'analyse. Il existe un grand nombre de méthodes différentes, sans avoir vraiment de règles désignant une méthode plus efficace que d'autres. Makowski en 2018, propose de se reposer sur un consensus parmi les méthodes plutôt que sur une méthode en particulier.

Le mot du praticien :

Attention l'AFC comme d'ailleurs les autres méthodes de l'analyse factorielle, fournit des visualisations et des métriques concernant les variables étudiées et venant illustrer leur relation. En revanche, elle ne démontre pas cette relation.
Avant toute analyse factorielle, il est également indispensable de réaliser une analyse préliminaire de chaque variable, afin de voir si toutes les modalités sont aussi bien représentées ou s'il existe un déséquilibre. L'AFC comme toutes les analyses factorielles est sensible aux petits effectifs. Aussi il peut être préférable de regrouper les modalités peu représentées le cas échéant.

L'AFC, étant un croisement de deux variables qualitatives, il est important de réfléchir en amont aux questions que l'on veut poser et surtout vérifier l'existence d'une relation entre les deux variables que l'on souhaite étudier. L'AFC repose sur le test du chi², il est donc nécessaire en premier lieu de réaliser ce test pour confirmer l'absence d'indépendance entre les deux variables.

Nous rajouterons qu'il est également vivement conseillé d'étudier en amont les marges du tableau croisé. Ces étapes ne doivent pas être négligées car elles donnent des informations importantes sur nos données et l'analyse future.

Ressources :

- Benzécri Jean-Paul, *L'Analyse des données*, Paris Bruxelles Montréal, Dunod, 1973.
- Bourdieu Pierre, *La distinction critique sociale du jugement*, Paris, Editions de Minuit : Maison des sciences de l'homme, coll. « Le Sens commun », 2012.
- Husson François, Lê Sébastien et Pagès Jérôme, *Analyse de données avec R*, 2e éd. revue et Augmentée., Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2016.
- Makowski Dominique, « The psycho Package: an Efficient and Publishing-Oriented Workflow for Psychological Science », *The Journal of Open Source Software*, nº 22, vol. 3, 2018, p. 470, [https//doi.org/10.21105/joss.00470].
- Pages J. P., Cailliez F. et Escoufier Y., « Analyse factorielle : un peu d'histoire et de géométrie », *Revue de Statistique Appliquée*, nº 1, vol. 27, 1979, p. 5-28.
- Spearman C., « The Proof and Measurement of Association between Two Things », *The American Journal of Psychology*, nº 72-101, vol. 15, 1904, [https//doi.org/10.2307/1422689].

AFE n.f [/a.ɛf.ə/]

Synonymes : Analyse Factorielle Exploratoire, EFA, Exploratory Factorial Analysis

A quoi ça sert ?

L'objectif de l'analyse factorielle exploratoire est de réduire un nombre important d'informations (c'est-à-dire les valeurs contenues dans différentes variables) à quelques grandes dimensions appelées *facteurs latents*. Il s'agit de synthétiser l'information en procédant par regroupement de variables. L'idée est d'explorer la structure sous-jacente d'une base de données en identifiant des *facteurs latents* qui vont expliquer les relations entre les variables de notre base de données.

L'analyse factorielle exploratoire est un type d'analyse factorielle au même titre que **l'Analyse en Composantes Principales (ACP)**, par exemple. Toutefois, la particularité de l'AFE est qu'elle se concentre sur la variance partagée entre les variables pour déterminer des facteurs latents et explorer une structure sous-jacente à nos données, alors que l'ACP va plutôt chercher à conserver la totalité de la variance présente dans le jeu de données afin de dégager des composantes. Une variable qui ne corrèlerait pas avec le reste du jeu de données que nous souhaiterions factoriser, serait alors sortie de l'analyse dans le cadre d'une AFE, mais pas nécessairement avec une **ACP**.

D'où ça vient ?

Les premiers à théoriser les méthodes d'analyses factorielles sont le mathématicien britannique Karl Pearson (1901) et le psychologue anglais Charles Spearman (1904). Pearson développe sa réflexion sur les **analyses en compostantes principales (ACP)** (Karl Pearson F.R.S., 1901)[10], tandis que Spearman se concentre sur l'analyse factorielle, afin de rendre compte de la variance commune partagée par les items d'un même outil psychométrique (Spearman, C., 1904)[11].

Si les travaux du mathématicien français Jean-Paul Benzecri dans les années 70 ont contribué à l'essor des **analyses en composantes principales** en France (Benzecri J.-P., 1973)[12], (Pages, J-P., et al., 1979)[13], en SHS, l'AFE méthode très utilisée en psychologie (discipline de Spearman), reste beaucoup plus rare dans les autres sciences humaines et sociales.

Exemple d'application

L'AFE est très utilisée en Psychologie et reste encore peu connue dans les autres disciplines des SHS. L'emploi de l'AFE en Psychologie est surtout réalisé pour valider, standardiser, des tests psychotechniques. En effet, afin de réaliser des tests standardisés pour évaluer des processus psychologiques les psychologues ont recours au couple d'analyses factorielles en réalisant une analyse factorielle exploratoire sur un premier échantillon et une analyse factorielle confirmatoire sur un second échantillon (ayant les mêmes caractéristiques). Il s'agit d'une démarche très largement

[10] Pearson Karl, « LIII. *On lines and planes of closest fit to systems of points in space* », *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, nº 11, vol. 2, 1901, p. 559-572, [https//doi.org/10.1080/14786440109462720].

[11] Spearman C., « The Proof and Measurement of Association between Two Things », *The American Journal of Psychology*, nº 72-101, vol. 15, 1904, [https//doi.org/10.2307/1422689].

[12] Benzécri Jean-Paul, *L'Analyse des données*, Paris Bruxelles Montréal, Dunod, 1973.

[13] Pages J. P., Cailliez F. et Escoufier Y., « Analyse factorielle : un peu d'histoire et de géométrie », *Revue de Statistique Appliquée*, nº 1, vol. 27, 1979, p. 5-28.

documentée dans la littérature que nous ne développerons pas ici. Pour en savoir plus vous pouvez consulter notamment l'article de Flora, D. B., & Flake, J. K., 2017[14].
Une application de l'AFE est par exemple présentée afin de valider un outil psychométrique permettant de mesurer les comportements personnels de déviance ou d'adhésion aux normes dans un contexte organisationnel. Dans cet exemple l'analyse factorielle exploratoire porte sur 12 variables qui se répartissent, selon l'AFE, en quatre facteurs latents : conformité normative, adéquation aux règles normatives, recherche de performance déviante et recherche de proactivité déviante (Déprez, G. R. M., et al., 2019)[15].

Mot du praticien

L'AFE est plus efficiente avec des variables continues. Il est tout-à-fait possible de l'utiliser avec des variables ordonnées : catégories qui suivent un ordre, (par ex. de 1. "Pas du tout d'accord" à 5. "Tout-à-fait d'accord") en traduisant chaque catégorie par un score. On peut également réaliser ce type d'analyse avec des variables catégorielles (catégories sans échelle de valeur entre elles par ex. : profession) mais il faudra alors être beaucoup plus prudent sur l'interprétation, car il n'y aura alors pas de garanties que les coefficients obtenus soient justes. Et selon les cas, il peut être plus pertinent d'utiliser une autre méthode de factorisation comme **l'analyse des correspondances multiples (ACM)** ou une **analyse factorielle des correspondances (AFC)**.

La taille de l'échantillon est également un élément contraignant pour ce type d'analyse. Si l'échantillon est trop faible en regard de la taille des informations à synthétiser, les résultats obtenus risquent de surreprésenter les spécificités de la population alors testée. La structure de l'analyse ne sera pas généralisable, mais influencée par les spécificités de l'échantillon. Il est donc important d'avoir une taille d'échantillon suffisamment importante afin de pallier ce type de biais.

Des corrélations trop fortes ou trop faibles entre les variables peuvent mettre en péril la mise en facteur des informations, mais pas pour les mêmes raisons. Si les variables soumises à la factorisation ne sont pas du tout corrélées, alors les éléments présentés ne partagent pas d'éléments communs et ne peuvent pas être résumés sous un même facteur. L'hétérogénéité des informations peut tout-à-fait empêcher son résumé statistique via l'utilisation des analyses factorielles. Il n'est pas possible de réaliser des analyses factorielles sur des ensembles de variables non-corrélés. A l'inverse, si les variables retenues pour l'analyse sont trop fortement corrélées alors ceci sous-tend qu'elles contiennent des informations tellement similaires qu'elles en sont redondantes. Cette redondance peut tout-à-fait biaiser la factorisation des informations car certains éléments seront surreprésentés, sans que cela ne reflète une quelconque réalité, par rapport à d'autres. Ce problème peut être résolu en sélectionnant uniquement l'une des deux variables représentatives d'un même phénomène. Par exemple, la Catégorie Socio-Professionnelle et le niveau de diplôme sont très souvent très corrélés, il faudra sélectionner uniquement l'une de ces deux variables dans le modèle d'analyse factorielle pour que celui-ci soit optimal.

[14] Flora David B. et Flake Jessica K., « The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. », *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, n° 2, vol. 49, 2017, p. 78-88, [https//doi.org/10.1037/cbs0000069].

[15] Déprez Guillaume Roland Michel, Battistelli Adalgisa et Antino Mirko, « Norm and Deviance-Seeking Personal Orientation Scale (NDPOS) Adapted to the Organisational Context », *Psychologica Belgica*, n° 1, vol. 59, 2019, [https//doi.org/10.5334/pb.462].

La subjectivité de l'interprétation peut aussi être une limite, notamment sur l'identification et l'interprétation des facteurs latents, qui dépendra beaucoup du chercheur et de son positionnement théorique.

Cette analyse gagnerait à être plus utilisée dans les disciplines autres que la psychologie car elle répond à un questionnement qu'on retrouve globalement dans toutes les SHS. L'influence de facteur latents sur les variables est en fait une réalité globalement des données en sciences humaines et sociales. La non utilisation de l'AFE relève plus bien souvent d'une méconnaissance de son existence plutôt qu'à un choix conscient.

Ressources

- Benzécri Jean-Paul, *L'Analyse des données*, Paris Bruxelles Montréal, Dunod, 1973.
- Déprez Guillaume Roland Michel, Battistelli Adalgisa et Antino Mirko, « Norm and Deviance-Seeking Personal Orientation Scale (NDPOS) Adapted to the Organisational Context », *Psychologica Belgica*, nº 1, vol. 59, 2019, [https//doi.org/10.5334/pb.462].
- Flora David B. et Flake Jessica K., « The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. », *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, nº 2, vol. 49, 2017, p. 78-88, [https//doi.org/10.1037/cbs0000069].
- Pages J. P., Cailliez F. et Escoufier Y., « Analyse factorielle : un peu d'histoire et de géométrie », *Revue de Statistique Appliquée*, nº 1, vol. 27, 1979, p. 5-28.
- Pearson Karl, « LIII. *On lines and planes of closest fit to systems of points in space* », *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, nº 11, vol. 2, 1901, p. 559-572, [https//doi.org/10.1080/14786440109462720].
- Spearman C., « The Proof and Measurement of Association between Two Things », *The American Journal of Psychology*, nº 72-101, vol. 15, 1904, [https//doi.org/10.2307/1422689].

Un bookdown :

- Lugu Benjamin, *Exploratory Factor Analysis in R*. [https://bookdown.org/luguben/EFA_in_R/] (consulté le 27/03/2025).

AFM n.f. [/a.ɛf.ɛm/]

Synonyme : Analyse factorielle Multiple

A quoi ça sert ?

L'Analyse Factorielle Multiple (AFM) au même titre que les autres méthodes d'analyses factorielles (telles que l'**ACP, l'AFE, l'AFC, l'ACM**), est une méthode de réduction de dimension. Elle permet donc de synthétiser une base de données. L'intérêt et la particularité de l'AFM c'est qu'elle permet d'analyser des données où les variables sont structurées en groupes ou thématiques. Ces groupes peuvent être soit quantitatifs, soit qualitatifs. Elle est donc particulièrement adaptée à des bases de données complexes composées de groupes de variables quantitatives ou qualitatives.
Cette analyse présente l'intérêt de pouvoir avoir une analyse au niveau des individus, des variables ou des groupes de variables. Comme les autres analyses factorielles, il y a une dimension graphique importante, il est possible de réaliser des représentations graphiques qui permettent de faciliter grandement sa lecture et son analyse.

D'où ça vient ?

L'AFM a été développée par Escoffier et Pagès (1983) initialement pour l'analyse de groupe de variables quantitatives. Puis elle a été élargie à l'étude de groupe de variables qualitatives (Escoffier et Pagès, 1998) et enfin une dernière extension proposée par Pagès en 2002 a permis d'avoir une approche mixte en intégrant l'analyse dans une même base de données de groupes de variables quantitatives et de groupes de variables qualitatives.

Exemple d'application

Comme présenté par Maëlle Amand dans un tutoriel vidéo en ligne en 2021[16] cette approche est utilisée dans les études de comportement alimentaire, en écologie, sensorimétrie ou pour analyser les résultats de questionnaires d'enquête de satisfaction (Bécue-Bertaut et al. 2008). Sinon nous pouvons renvoyer directement à la thèse de Maëlle Amand (2019) en sociolinguistique.

Mot du praticien

Nous pouvons noter que le champ de l'analyse factorielle continue d'évoluer, en 2012 Chavent et al. proposent la méthode PCAMIX qui une méthode d'analyse factorielle pour des données mixtes et en 2013, en s'appuyant sur ces travaux Labenne et al. proposent une approche complètement mixte de l'AFM où cette fois les groupes de variables eux même peuvent être mixtes et composés de variables quantitatives et qualitatives, ce que ne permet pas l'AFM, il s'agit de la MFAMIX (Multiple Factor Analysis of mixed groups of variables - Analyse factorielle multiple de groupes de variables mixtes).

Ressources

- Amand Maëlle, *Tuto@MATE - Les Analyses Factorielles Multiples (AFM)*, [https://mate-shs.cnrs.fr/actions/tutomate/tuto32-les-analyses-factorielles-multiples-afm-amand/].

[16] Tuto@Mate du 13 avril 2021 : Amand Maëlle, *Tuto@MATE - Les Analyses Factorielles Multiples (AFM)*, [https://mate-shs.cnrs.fr/actions/tutomate/tuto32-les-analyses-factorielles-multiples-afm-amand/].

- Amand Maelle, *A sociophonetic analysis of Tyneside English in the DECTE corpus : the case of FACE, GOAT, PRICE and MOUTH*, thèse de doctorat, Université Paris Cité, 2019.
- Bécue-Bertaut Mónica et Pagès Jérôme, « Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data », *Computational Statistics & Data Analysis*, nº 6, vol. 52, 2008, p. 3255-3268, [https//doi.org/10.1016/j.csda.2007.09.023].
- Chavent Marie, Kuentz-Simonet Vanessa et Saracco Jérôme, « Orthogonal rotation in PCAMIX », *Advances in Data Analysis and Classification*, nº 2, vol. 6, 2012, p. 131-146, [https//doi.org/10.1007/s11634-012-0105-3].
- Escofier B. et Pages J., « Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation de vins rouges du Val de Loire », *Revue de Statistique Appliquée*, nº 2, vol. 31, 1983, p. 43-59.
- Escofier Brigitte et Pagès Jérome, *Analyses factorielles simples et multiples. Objectifs méthodes et interprétation*, Dunod, coll. « Sciences Sup », 2008.
- Labenne Amaury, Chavent Marie, Kuentz Simonet V., Rambonilaza Mbolatiana et Saracco Jérôme, « Une extension de l'analyse factorielle multiple pour des groupes de variables mixtes: MFAmix », Toulouse, France.
- Pagès J., « Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes », *Revue de Statistique Appliquée*, nº 4, vol. 50, 2002, p. 5-37.

Alpha de Cronbach n.m. [/alfa də kʁɔnbax/]

Voir entrée **Cohérence interne**.

Analyse de contenu n.f. [/analiz də kɔ̃təny/]

Voir entrée **Analyses Qualitatives**.

Analyse Discriminante n.f. [/a.na.liz dis.kʁi.mi.nɑ̃t/]

Synonymes : Analyse factorielle discriminante, AFD

À quoi ça sert ?

Dans une analyse factorielle discriminante (AFD), l'objet est de prédire une variable qualitative Y en fonction d'un ensemble de variables continues $X_1, \dots, X_p$. Il s'agit donc d'une méthode de classification supervisée, permettant de prédire le groupe d'appartenance d'un individu en fonction de plusieurs variables numériques[17].

On peut usuellement présenter l'AFD selon deux points de vue différents : ou bien selon une approche faisant appel uniquement à l'algèbre linéaire, ou bien selon une approche probabiliste bayésienne. Ces deux approches sont en réalité équivalentes sous certaines conditions. Pour une présentation théorique complète, on pourra par exemple consulter Bardos (2001) ou Saporta (2011).

D'où ça vient ?

L'analyse factorielle discriminante est formalisée de manière rigoureuse pour la première fois dans un article de Ronald Fisher (1936), illustrant la méthode pour différencier trois espèces d'iris en fonction de mesures sur leurs pétales et sépales — le fameux jeux de données des "iris de Fisher". Fisher note que l'analyse discriminante pourra par exemple trouver des cas d'application en anthropologie biologique (estimation du sexe à partir de mesures craniométriques, étude de l'évolution de dimensions au cours du temps, etc.), ce qui a en effet été le cas de façon intensive depuis lors.

Avant l'avènement des méthodes modernes de classification supervisée nécessitant une grande puissance de calcul (arbres de décision, forêts aléatoires, SVM, réseaux de neurones, etc.), l'analyse factorielle discriminante est restée pendant des décennies la méthode la plus employée. Calculable à la main sur des jeux de données de petite dimension, elle fournit néanmoins des modèles prédictifs très efficaces dans la majorité des cas.

Exemple d'application

- En économie : dans un article intitulé "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy", Edward Altman (1968) utilise des ratios financiers et des modèles d'analyse discriminante pour prédire la faillite d'une entreprise. Altman y présente un modèle qui, en combinant des ratios financiers pertinents, permet de distinguer les entreprises susceptibles de faire faillite de celles qui sont durablement viables. Dans cette étude, la variable à prédire est donc binaire (faillite / non-faillite), et les variables numériques prédictives sont cinq ratios financiers impliquant par exemple le chiffre d'affaires, la valeur des capitaux propres, le total des actifs, etc. L'étude démontre l'efficacité de l'analyse discriminante dans la prédiction des défaillances d'entreprises, et met en lumière l'importance de certains ratios financiers dans cette démarche.
- En sciences politiques et criminologie : Tollenaar & Heijden (2012) se proposent de comparer plusieurs méthodes d'apprentissage automatique (l'analyse discriminante, les arbres de

[17] On notera donc qu'en présence de variables prédictives qualitatives, il existe des méthodes plus adaptées, comme les arbres de décision ou la régression logistique. Il existe certes des "contournements" permettant d'inclure des prédicteurs qualitatifs dans une AFD, mais il semble plus judicieux de s'orienter alors vers un autre algorithme gérant nativement ce type de données.

décision, la régression logistique, les SVM, etc.) pour prédire la récidive de criminels en fonction de différents indicateurs numériques et non-numériques. En dépit d'un cadre d'application qui n'est pas idéal pour bâtir des modèles d'analyse linéaire discriminante (présence de prédicteurs qualitatifs), l'étude prouve que l'analyse discriminante permet d'obtenir un modèle aussi précis, voire plus précis, que ceux fournis par des méthodes d'apprentissage automatique plus récents et beaucoup plus gourmands en temps de calcul.

Mot du praticien

L'analyse factorielle discriminante pose plusieurs hypothèses assez fortes sur les données, qu'il convient idéalement de vérifier avant d'appliquer le modèle, même si l'AFD est modérément robuste à la violation de certaines de ces hypothèses (Lachenbruch et al., 1973).

- L'analyse factorielle discriminante est un modèle paramétrique car elle suppose des distributions normales (multivariées) pour les données. Cette hypothèse peut se vérifier à l'aide de tests de normalité multivariée (il en existe un grand nombre).
- De plus, l'analyse discriminante suppose que les groupes comparés ont même matrice de covariance, c'est-à-dire même structure de dispersion. Cette hypothèse peut se vérifier à l'aide du test M de Box.
- Les individus sont supposés être tous indépendants.
- Il convient enfin de vérifier l'absence de forts outliers dans le jeu de données, en particulier lorsqu'on travaille sur de petits échantillons.

Extensions et alternatives :

- Avec une analyse factorielle discriminante, il n'est pas impératif (et d'ailleurs même pas usuel) de procéder à une sélection des meilleures variables, contrairement par exemple à la régression logistique. Toutefois, lorsque le nombre de variables est élevé par rapport au nombre d'individus, ou lorsque l'identification des variables les plus discriminantes est un objectif en soi, il peut être utile d'appliquer automatiquement une procédure de sélection de variables, par exemple par un algorithme pas-à-pas descendant. Le package R {klaR} implémente cette procédure pour l'AFD.
- Lorsque l'hypothèse d'homogénéité des matrices de covariance n'est pas satisfaite, il existe une alternative à l'analyse linéaire discriminante qui permet de relâcher cette hypothèse : l'analyse quadratique discriminante. Elle est elle aussi implémentée dans le package R {MASS}, via la fonction qda().
- Enfin, il existe des variantes robustes de l'analyse discriminante (Ghosh et al., 2020; Leys et al., 2018) afin de minimiser l'impact négatif de potentiels outliers.

Ressources

- Altman Edward I., « FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY », *The Journal of Finance*, n° 4, vol. 23, 1968, p. 589-609, [https//doi.org/10.1111/j.1540-6261.1968.tb00843.x].
- Bardos Mireille, *Analyse discriminante: application au risque et scoring financier*, Paris, Dunod, coll. « Collection Éco sup. Manuel », 2001.
- Fisher R. A., « THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS », *Annals of Eugenics*, n° 2, vol. 7, 1936, p. 179-188, [https//doi.org/10.1111/j.1469-1809.1936.tb02137.x].

- Ghosh Abhik, SahaRay Rita, Chakrabarty Sayan et Bhadra Sayan, « Robust Generalised Quadratic Discriminant Analysis », [https//doi.org/10.48550/arXiv.2004.06568].
- Lachenbruch Peter A., Sneeringer Cheryl et Revo Lawrence T., « Robustness of the linear and quadratic discriminant function to certain types of non-normality », *Communications in Statistics*, nº 1, vol. 1, 1973, p. 39-56, [https//doi.org/10.1080/03610927308827006].
- Leys Christophe, Klein Olivier, Dominicy Yves et Ley Christophe, « Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance », *Journal of Experimental Social Psychology*, vol. 74, 2018, p. 150-156, [https//doi.org/10.1016/j.jesp.2017.09.011].
- Saporta Gilbert, *Probabilités, analyse des données et statistique*, 3e éd. révisée., Paris, Éd. Technip, 2011.
- Tollenaar N. et Van Der Heijden P. G. M., « Which Method Predicts Recidivism Best?: A Comparison of Statistical, Machine Learning and Data Mining Predictive Models », *Journal of the Royal Statistical Society Series A: Statistics in Society*, nº 2, vol. 176, 2013, p. 565-584, [https//doi.org/10.1111/j.1467-985X.2012.01056.x].

Analyse Harmonique Qualitative n.f. [/a.na.liz aʁ.mɔ.nik ka.li.ta.tiv/]

Synonyme : AHQ

A quoi ça sert ?

L'analyse harmonique qualitative permet d'étudier des données temporelles. En ce sens nous pouvons la classer dans le champ des analyses longitudinales bien qu'il s'agisse en réalité d'une variante de l'analyse des correspondances. Son objectif est d'analyser des données qui décrivent le changement d'état d'une variable pour un ensemble d'individus. Typiquement nous utiliserons une analyse harmonique qualitative lorsque les individus statistiques de nos échantillons sont décrits par une chronologie de changements d'états parmi un nombre fini d'états possibles, par exemple : changement de commune, changement de statut matrimonial ou changement d'emploi, etc. L'AHQ permet d'analyser des données biographiques ou des histoires individuelles. Elle va nous permettre de structurer des informations sur les changements d'états successifs au cours du temps, d'identifier des traits dominants de différenciation entre les trajectoires individuelles et d'établir une typologie des trajectoires des individus. Nous pouvons également souligner que l'AHQ partage avec les **analyses factorielles** la possibilité d'intégrer des variables supplémentaires, ce qui permet de compléter l'analyse et d'éclairer les trajectoires étudiées à l'aide d'éléments contextuels ou caractéristiques supplémentaires des individus.

D'où ça vient ?

L'analyse Harmonique Qualitative a été développée par Jean-Claude Deville dans les années 1970 (Deville, 1974, 1977) pour introduire la dimension temporelle dans l'analyse des phénomènes sociaux en s'inspirant de l'analyse harmonique en mathématiques, alors utilisée en biologie et dans les sciences physiques. Elle a ensuite été adaptée pour la statistique exploratoire des trajectoires complexes et la constitution de typologies de parcours (Deville et Saporta,1980), puis à nouveau par Jean-Claude Deville, dans un article de 1982 dans les Annales de l'INSEE, où elle est appliquée à des données sur nuptialité d'une cohorte de femmes ayant connues plusieurs mariages. Son usage s'est développé avec les collectes biographiques modernes permettant de suivre des trajectoires complètes.

Mot du praticien :

L'analyse harmonique qualitative nécessite des trajectoires complètes et structurées, et son interprétation dépend des choix méthodologiques et du terrain étudié. La taille et l'amplitude des intervalles doivent être adaptées au rythme des changements. S'ils sont trop fins, ils génèrent des cases nulles et du bruit tandis que s'ils sont trop larges, ils entraînent une perte de précision. Enfin, la méthode résume l'information mais ne prend pas en compte la similarité exacte entre les séquences : elle repose sur la construction d'une matrice harmonique où l'on calcule les proportions de temps passées dans chaque état, puis sur une analyse factorielle des correspondances pour résumer l'information. L'information est ainsi agrégée. Elle est donc plus adaptée pour mettre en évidence des profils globaux et des motifs généraux. En comparaison, l'appariement optimal mesure la distance précise entre deux séquences à l'aide de coûts pour insertions, suppressions ou substitutions. Il est plus adapté pour analyser la ressemblance fine entre parcours.

Ressources

- Deville, « Méthodes statistiques et numériques de l'analyse harmonique », *Annales de l'inséé*, nº 15, 1974, p. 3, [https//doi.org/10.2307/20075177].

- Deville Jean-Claude, « Analyse harmonique du calendrier de constitution des familles en France. Disparités sociales et évolution de 1920 à 1960 », *Population (French Edition)*, n° 1, vol. 32, 1977, p. 17, [https//doi.org/10.2307/1531590].
- Deville, « Analyses de données chronologiques qualitatives: Comment analyser des calendriers? », *Annales de l'inséé*, n° 45, 1982, p. 45, [https//doi.org/10.2307/20076433].
- Deville Jean-Claude et Saporta Gilbert, « L'analyse harmonique qualitative », Versailles, France, North-Holland, coll. « Data Analysis and Informatics ».
- Diday E. et Institut National de Recherche en Informatique et en Automatique (dir.), *Data analysis and informatics, V: proceedings of the Fith International Symposium on Data Analysis and Informatics, organised by the Institut National de Recherche en Informatique et en Automatique, Versailles, September 29 - October 2, 1987*, Amsterdam, North-Holland, 1988.

Analyses Qualitatives n.f.p [/a.na.liz ka.li.ta.tiv/]

Synonymes : analyse du discours

À quoi ça sert ?

L'analyse qualitative permet d'exploiter des données non chiffrées tels que des entretiens, des corpus de textes ou d'articles, des professions de foi, des textes de lois, etc. Le matériel de départ va être le texte sous toutes ses formes. Elle vise à comprendre le sens des discours : ce que disent les personnes, leurs représentations, leurs valeurs et les catégories implicites qui structurent leurs propos. En cela, c'est donc un type d'analyse de données central en sciences humaines et sociales. Elle diffère de l'analyse quantitative de données qualitatives, qui met l'accent sur la fréquence des co-occurences telles que les approches développées en **textométrie**.

L'analyse qualitative pourra répondre à plusieurs objectifs principaux :

- Documenter un point de vue, des attitudes, des opinions ou des valeurs.
- Identifier des thématiques récurrentes ou divergentes.
- Construire des catégories, des modèles interprétatifs ou des théories.
- Mettre à jour le système de valeurs sous-jacent aux réponses.

Ainsi, selon les objectifs, l'analyse qualitative peut viser : soit à illustrer un propos par des extraits choisis (verbatim), soit à structurer systématiquement l'ensemble des réponses pour en faire émerger des tendances, des structures ou des modèles.

En fonction des objectifs visés, différentes méthodes pourront être utilisées, nous pouvons notamment citer :

- L'analyse de contenu : repérage systématique à partir d'une grille prédéfinie (plutôt déductive). A privilégier si l'objectif est de valider ou d'illustrer une hypothèse. Dans cette approche, les éléments ou codes spécifiques des données sont catégorisés et quantifiés de manière systématique et quantitative. L'analyse de la distribution et de l'occurrence de ces éléments est l'objectif principal.
- L'analyse thématique : émergence inductive de thèmes (plutôt inductive donc). A privilégier si vous voulez explorer de nouvelles pistes sans a priori. Dans cette approche, les thèmes récurrents, les modèles et les significations sont identifiés et analysés dans les données en utilisant une approche interprétative et qualitative. Les expériences et les points de vue des participants sont mis en avant afin de mieux les comprendre.
- La théorie ancrée (Grounded Theory) : construction itérative de concepts et d'une théorie à partir des données (totalement inductive). A privilégier si vous souhaitez faire émerger un modèle théorique. Cette approche vise à développer une théorie basée directement sur les données collectées sur le terrain. Elle se concentre sur l'élaboration de concepts et de catégories à partir des données brutes, sans hypothèses préexistantes.

L'analyse qualitative est un champ riche composé de nombreuses méthodes. Il existe donc d'autres méthodes que celles citées précédemment, telles que : l'analyse narrative, l'analyse phénoménologique interprétative, ou encore l'Analyse des Relations par Opposition (ARO).
Toutefois, pour être appliquées ces différentes méthodes ont un point commun, elles requièrent un investissement important de la part de celui qui souhaite l'utiliser, tant dans la préparation des données que dans la réalisation de l'analyse elle-même.

D'où ça vient ?

L'analyse qualitative s'est construite progressivement au fil du XXe siècle, d'abord aux Etats-Unis, et a connu différentes étapes importantes en fonction de l'émergence parallèle des différentes méthodes de ce champ d'analyse (Bardin, 2013).

L'analyse de contenu émerge dans la première moitié du XXe siècle, notamment dans le cadre de l'École de journalisme de Columbia. Cette méthode visait initialement à étudier la propagande et les médias de masse, en codant de manière systématique des éléments textuels. Berelson (1952) joue un rôle central dans la formalisation de cette approche en posant les bases méthodologiques : objectivation du codage, segmentation du corpus, quantification des occurrences, etc. Néanmoins, cette méthode, alors dominée par une visée descriptive et statistique, est rapidement critiquée pour son manque de profondeur interprétative.

Dès les années 1950-60, les limites de l'analyse de contenu traditionnelle conduisent à l'exploration de nouvelles approches. La notion de cooccurrence marque un tournant important. Elle permet de dépasser la simple fréquence des mots pour analyser leurs relations sémantiques dans le discours. Cette évolution accompagne l'arrivée des premiers ordinateurs dans les années 1970-80, qui facilitent l'automatisation partielle des traitements textuels. Les logiciels d'analyse lexicale permettent ainsi une approche plus fine des structures discursives, tout en posant de nouveaux défis méthodologiques et techniques.

Parallèlement, l'analyse thématique se développe à partir des années 1950-60, dans un contexte d'essor des sciences humaines et sociales. Elle met l'accent sur l'identification des thèmes récurrents et significatifs au sein des discours. Cette approche, davantage ancrée dans l'interprétation du sens, s'intéresse à la manière dont les individus expriment leurs représentations, leurs valeurs ou leurs stratégies narratives. Selon Mucchielli (1996), l'analyse de contenu est un terme générique englobant diverses techniques visant à expliciter « *le ou les sens qui sont contenus et/ou le ou les manières dont ils parviennent à faire effet de sens* » (p. 36).

La formalisation de la théorie ancrée (Grounded Theory) par Glaser et Strauss (1967), va constituer une étape majeure dans le domaine de l'analyse qualitative. Cette approche rompt avec les modèles hypothético-déductifs traditionnels en postulant que la théorie doit émerger inductivement des données empiriques (et s'appuie donc sur les acquis de l'analyse thématique). L'entretien est alors perçu comme une source de construction de sens, et non comme un simple outil de vérification d'hypothèses. La théorie ancrée introduit une logique itérative : la collecte des données, leur codage et l'élaboration théorique s'articulent dans un processus circulaire.

Enfin, l'introduction des outils informatiques dans les années 1980-90 marque une nouvelle étape dans l'analyse qualitative. Des logiciels comme NUD*IST, puis NVivo, Atlas.ti ou MAXQDA, vont fournir aux chercheurs des interfaces graphiques permettant de coder, d'organiser et de visualiser les données textuelles de manière systématique. Nous parlons alors des CAQDAS (Computer Assisted Qualitative Data Analysis Software), termes qui désignent ces logiciels dont l'objectif est de faciliter l'analyse qualitative.

Exemple d'application

L'ARO est une méthode d'analyse du discours (écrit ou parlé) qui permet de mettre à jour l'univers de référence des personnes interviewées. En d'autres termes, cette méthode permet de voir ce qui se cache derrière un discours et plus précisément les valeurs associées à un univers de référence.

C'est une technique qui révèle les relations de signification entre les signifiants (les objets spécifiés) et les signifiés – ce que le locuteur dit à propos de ces objets – qui s'opposent terme à terme.

Elle repose sur l'analyse structurale des récits et suppose l'existence d'une relation (correspondance) entre les éléments d'un système pratique et les éléments d'un système symbolique. La structuration de cette relation en opposition, étant constitutive de la fonction symbolique. Il faut cependant souligner que l'ARO est une technique qui n'est pas facile à mettre en œuvre car « *il arrive cependant que la relation par opposition exige, pour être complétée, de prendre en compte des énoncés dispersées sur plusieurs pages. Le lecteur, alerté par l'identification des premiers éléments, poursuit la lecture en attendant les éléments correspondants* » (Blanchet et Gotman, 1992). Cette étape est particulièrement importante et longue.

La Méthode consiste en une manipulation des objets de l'entretien et à repérer « un lien de signification » entre « un signifiant **»** (les 'objets' dont nous parlons) et « un signifié » (ce que nous pensons ou ressentons à propos de ces 'objets'). L'hypothèse est que c'est en opposant deux univers de références que l'interviewé donne un sens à ses actions, ses décisions et ses sentiments.

Elle est constituée de 5 phases :

- Le découpage des énoncés (proche d'une analyse de contenu) permet de regrouper les propos qui concernent les éléments recherchés, comme par exemple le travail et les loisirs ou la ville et la campagne. Cette étape permet de constituer un lexique, où l'objectif sera de grouper et de classer les significations associées à chacun des termes du lexique.

Exemple issu de l'ouvrage de Blanchet et Gotman, 1992.

Signifiants		Signifiés médiateurs	Signifiés
Etudes	Droit	Mes parents voulaient que j'aie un métier solide	Ils voulaient que j'aie un diplôme qui sur le marché du travail puisse être bénéfique
	Psychologie	Des psychologues on n'en a pas énormément besoin	(Mes parents) considéraient que c'était quelque chose qui était plus ou moins bouché

- Commentaire de la construction de l'énoncé : si les parents associent le droit à un diplôme bénéfique sur le marché du travail c'est parce qu'ils veulent que leur enfant ait un métier solide. À l'opposé, ils considèrent la psychologie comme quelque chose qui est bouché parce qu'on n'a pas besoin de psychologues.
- La réduction des énoncés : on ne garde que les mots clés par exemple : « j'aime » vs « ce que je n'aime pas » ; « facile » vs « difficile » ; « proche » vs « étranger » etc. On obtient des oppositions.

Signifiants		Signifiés médiateurs	Signifiés
Etudes	Droit	Métier solide	Diplôme bénéfique
	Psychologie	Pas besoin	Bouché

- Le classement du discours signifiant : ordonne un lexique de données saisies à partir des thèmes de l'entretien.
- Le classement du discours signifié : donne des axes sémantiques qui décrivent les ensembles d'évaluation et de représentation des données du lexique. Cela réorganise les verbatims de l'entretien en fonction du signifiant auquel ils sont rattachés. Le système de valeur recherché apparait.

Mots du praticien

L'analyse qualitative est une méthode reconnue et centrale en SHS, elle s'ancre dans un autre paradigme théorique et méthodologique que les différentes méthodes d'analyses quantitatives.
Un des enjeux fondamentaux est que l'interprétation doit rester réflexive : il est important d'être conscient de sa propre posture théorique et de ses biais possibles. Ainsi, la qualité d'une analyse repose autant sur la rigueur du codage que sur la capacité à rester fidèle au sens du discours.

Il est important de noter que cette méthode ne vise pas la « représentativité statistique » mais l'exhaustivité sémantique (saturation théorique).

L'expérience nous apprend qu'il y a plusieurs étapes cruciales à respecter pour disposer d'un bon matériau qui servira ensuite de base à l'analyse :

- Élaboration rigoureuse des grilles d'entretien (idéalement testées en pré-enquête). Grilles qui permettront de tester les questions/hypothèses et qui ont pour objectif le recueil d'un matériau riche et le plus possible non biaisé. Il faudra bien évidemment veiller à ce que la structure des questions n'induise pas les réponses.
- La conduite de l'entretien dans des conditions optimales.
- La retranscription fidèle du discours selon une méthode précise : retranscription totale ou partielle, manuelle ou via outils automatiques.

De fait, l'analyse qualitative va demander un investissement particulièrement important et sera très chronophage.

Les approches mixtes (quantification de résultats qualitatifs entre autres) sont bien sûr possibles, mais doivent être maniées avec prudence pour ne pas trahir la richesse des discours.
Concernant le choix de la méthode d'analyse, il dépend de la question de recherche mais aussi du positionnement initial. La méthode idéale n'existe pas et le choix entre les différentes méthodes dépendra moins d'une « hiérarchie » que de la question de recherche, du temps disponible, du nombre d'entretiens à traiter et de la familiarité avec l'interprétation qualitative.

Ressources :

- Bardin Laurence, *L'analyse de contenu*, Presses Universitaires de France, 2013, [https//doi.org/10.3917/puf.bard.2013.01].
- Berelson Bernard, *Content analysis in communication research*, New York, NY, US, Free Press, coll. « Content analysis in communication research », 1952.
- Blanchet Alain et Gotman Anne, *L'entretien*, 2e éd., nouv. Prés., Suite du tirage., Paris, A. Colin, coll. « Tout le savoir en 128 pages », 2017.
- Glaser Barney G. et Strauss Anselm L., *The discovery of grounded theory: strategies for qualitative research*, 11th printing., New York, Aldine, 1980.

- Lejeune Christophe, *Manuel d'analyse qualitative: Analyser sans compter ni classer*, De Boeck Supérieur, 2019, [https//doi.org/10.3917/dbu.lejeu.2019.01].
- Martin Angélique, *Traitement des entretiens par analyse de contenu thématique in. Les jeunes, l'insertion et les missions locales du pays d'Auge (Normandie) : les évolutions des représentations sociales entre 1982 et 2017*, thèse de doctorat, Conservatoire national des arts et metiers - CNAM, 2018.
- Messu Michel, *L'analyse des relations par opposition*, [https://www.credoc.fr/publications/lanalyse-des-relations-par-opposition].
- Messu Michel, *L'analyse de contenu : premiers éléments de réflexion*, [https://www.credoc.fr/publications/lanalyse-de-contenu-premiers-elements-de-reflexion].
- Mucchielli Alex, *Dictionnaire des méthodes qualitatives en sciences humaines et sociales*, Paris, A. Colin Masson, coll. « U », 1996.
- Negura Lilian, « L'analyse de contenu dans l'étude des représentations sociales », *SociologieS*, , 2006, [https//doi.org/10.4000/sociologies.993].
- Paillé Pierre, « L'analyse par théorisation ancrée », *Cahiers de recherche sociologique*, nº 23, 1994, p. 147-181, [https//doi.org/10.7202/1002253ar].
- Paillé Pierre et Mucchielli Alex, *L'analyse qualitative en sciences humaines et sociales*, Armand Colin, 2012, [https//doi.org/10.3917/arco.paill.2012.01].
- Raymond Henri, « Analyse de contenu et entretien non-directif : application au symbolisme de l'habitat », *Revue française de sociologie*, nº 2, vol. 9, 1968, p. 167-179, [https//doi.org/10.2307/3320589].

Analyse des Réseaux n.f. [/a.na.liz de ʁe.zo/]

Synonymes : analyse de graphes

A quoi ça sert ?

L'analyse de réseau sert à étudier les relations et les interactions entre des entités reliées entre elles par des liens. Elle est utilisée pour comprendre la structure, la dynamique et le rôle des éléments dans un système interconnecté. Elle constitue une très bonne alternative à l'analyse statistique classique, dans le cas où l'hypothèse d'indépendance au niveau des données n'est pas respectée. C'est le cas lorsque les données sont relationnelles (reliées, interconnectées). En effet, l'interdépendance entre les individus est au cœur de cette analyse, qui est faite pour ce type de données.

L'analyse de réseau permet d'obtenir des mesures qui seront employées notamment afin de qualifier le réseau alors obtenu et de classer les individus.

L'analyse de réseau constitue un champ entier de l'analyse de données avec ses indicateurs et méthodes spécifiques. Ainsi, il existe un grand nombre de types de réseau différents, et donc d'approches différentes. Le type de réseau va conditionner les analyses et les objectifs de l'analyse de réseau, et c'est la forme des données et de leurs relations qui va définir le type de réseau et donc les analyses possibles. Ainsi l'étape de définition de ce qui fait réseau est fondamentale.

Une approche descriptive pour qualifier le réseau est très souvent le point de départ classique à toute analyse de réseau. Ce travail repose sur le calcul d'indicateurs classiques (nombre de liens, nombre d'individus…) et d'autres plus complexes (poids dans le réseau...). Il est à noter que certains indicateurs vont pouvoir qualifier le réseau dans son intégralité, d'autres quilifieront les nœuds et encore d'autres les liens.

D'où ça vient ?

Le mythe fondateur de l'analyse de réseau est très souvent associé à la réponse apportée par le mathématicien suisse Leonhard Euler à l'énigme des ponts de Königsberg (Kaliningrad aujourd'hui) au XVIIIème siècle. La question était : est-il possible de faire le tour complet des quartiers de Königsberg en passant une seule fois sur chacun des ponts de la ville ? Euler ayant symbolisé les quartiers par des points et les ponts par des lignes pour résoudre son problème, il est souvent présenté comme l'inventeur de la théorie des graphes.

Bien sûr l'histoire des sciences montre que cela est plus complexe et l'analyse de réseau emprunte finalement assez peu à la théorie des graphes (Beauguitte, 2022)[18].

En sciences sociales, les travaux du sociologue classique allemand Georg Simmel au XIX^{e} siècle ont contribué à la structuration de la méthode en mettant les relations au cœur de son analyse. C'est au XXème siècle que le début de l'analyse de réseau en SHS commence véritablement avec Jacob Moreno en 1934 et son ouvrage « Who shall Survive ? »[19]. Dans ce travail Moreno va étudier les

[18] Beauguitte Laurent, 2022, *Théorie des graphes et analyse de réseau en géographie : histoire d'un lien faible (1950-1963)*, [https://ouest-edel.univ-nantes.fr/passerelleshs/index.php?id=155].

[19] Moreno J. L., *Who shall survive?: A new approach to the problem of human interrelations.*, Washington, Nervous and Mental Disease Publishing Co, 1934, [https//doi.org/10.1037/10648-000].

structures relationnelles entre les individus, il va notamment inventer des sociogrammes qui sont des dessins des relations observées. Dans la réalisation des sociogrammes, Moreno propose certaines règles qui sont encore appliquées par les algorithmes d'aujourd'hui, comme notamment le fait de placer les individus les plus connectés au centre, éviter le chevauchement de liens, etc.

La méthode va connaître un essor important puis décliner et être mise de côté jusque dans les années 1950-60 où elle connaît un nouvel essor grâce à l'analyse des réseaux sociaux. L'anthropologie britannique avec John A. Barnes va jouer rôle important et mettre en avant les propriétés structurales (points, chaînes). Mais nous pouvons également citer les travaux d'Harrison White[20] à Harvard, qui va faire école et intégrer pleinement l'analyse des réseaux sociaux à la sociologie quantitative.

Dans le même temps, toujours en Amérique du nord, des géographes vont se saisir de l'analyse de réseaux pour, entre autres, étudier l'analyse des flux et des réseaux d'infrastructure notamment avec les travaux de William Garrison[2122]

A la toute fin des années 90, l'analyse de réseaux connaît une nouvelle révolution avec la publication de travaux issus de la physique qui ont fait émerger de nouveaux modèles qui ont bouleversé les travaux menés sur l'analyse de réseaux en SHS, il s'agit du modèle du réseau small-World (Watts et Strogatz en 1998[23]) et du modèle de réseau scale-free (Barabási et Albert en 1999[24]), modèles qui ont infusé avec plus ou moins de facilité dans les SHS.

Aujourd'hui, le champ de l'analyse de réseaux a su s'imposer dans les différentes sciences humaines et sociales.

Exemple d'application :

L'application de méthodes d'analyse de réseaux sous diverses formes a également démontré une réelle portée dans plusieurs travaux de recherche en droit ou parmi des disciplines mobilisant un terrain juridique[25]. Toutefois, le recours à ces méthodes demeure encore peu fréquent et assez

---

[20] Freeman Linton C., *The development of social network analysis: a study in the sociology of science*, Vancouver (B.C.), Empirical press, 2004.

[21]Garrison William L., « Connectivity of the interstate highway system », *Papers in Regional Science*, n° 1, vol. 6, 1960, p. 121-137, [https//doi.org/10.1111/j.1435-5597.1960.tb01707.x].

[22] Garrison, L. William, Beauguitte Laurent, Beauguitte Pierre et Gourdon Paul, « William L. Garrison, 1960, Connectivity of the Interstate Highway System.Version bilingue et commentée ».

[23] Watts Duncan et Strogatz Steven, « Collective dynamics of 'small-world' networks », *Nature*, n° 6684, vol. 393, 1998, p. 440-442, [https//doi.org/10.1038/30918].

[24] Barabási Albert-László, Albert Réka et Jeong Hawoong, « Mean-field theory for scale-free random networks », *Physica A: Statistical Mechanics and its Applications*, n° 1-2, vol. 272, 1999, p. 173-187, [https//doi.org/10.1016/S0378-4371(99)00291-5].

[25] François Axelle, Nolet Anne-Marie et Morselli Carlo, « Sociabilité carcérale et réinsertion »:, *Déviance et Société*, n° 2, Vol. 42, 2018, p. 389-419, [https//doi.org/10.3917/ds.422.0389].

Kreager Derek A., Schaefer David R., Bouchard Martin, Haynie Dana L., Wakefield Sara, Young Jacob et Zajac Gary, « Toward a Criminology of Inmate Networks », *Justice Quarterly*, n° 6, vol. 33, 2016, p. 1000-1028, [https//doi.org/10.1080/07418825.2015.1016090].

Dupont Benoît, « La gouvernance polycentrique du cybercrime : les réseaux fragmentés de la coopération internationale », *Cultures & conflits*, n° 102, 2016, p. 95-120, [https//doi.org/10.4000/conflits.19292].

récent dans ce champ disciplinaire. Les difficultés méthodologiques entourant l'application de ces techniques pour des chercheurs peu formés à ce type d'analyse expliquent en partie cette situation. Néanmoins, certaines initiatives ou productions mobilisant l'analyse de réseaux pour l'étude de la traite des êtres humains ou l'histoire de la pensée juridique ont déjà été réalisées par des chercheurs juristes[26]. En effet, à titre d'exemple, il est possible de détecter des communautés épistémiques à travers l'analyse et le recensement des bibliographies d'un corpus de textes. La mise en lien de différents textes à l'aide de logiciels permet de visualiser la structure de ces ensembles, ou des échanges, via le calcul d'indicateurs spécifiques, mais également de les comprendre. De telles méthodes constituent un atout indéniable dans le cadre d'analyses historiques de la pensée ou dans une démarche comparative.

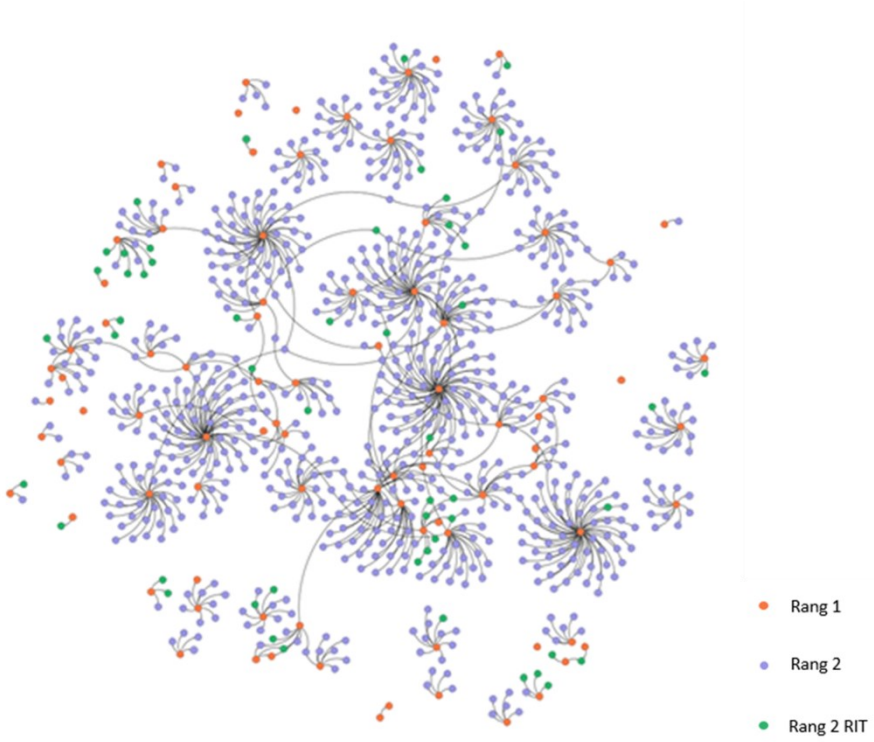
Notre exemple porte sur une recherche juridique mêlant droit comparé et histoire de la pensée sur le thème du travail. Néanmoins, la méthode proposée ici peut s'appliquer à d'autres disciplines.
Le but est de comprendre comment est décrit le travail en Afrique à partir d'une revue centrale en droit du travail : la Revue Internationale du Travail (RIT). Il s'agit dans notre exemple de recenser les articles de fonds sur l'Afrique de la RIT (les rapports sont exclus de l'étude) et de relever les références bibliographiques afin de détecter des éventuelles communautés épistémiques.
Pour les articles de fonds sur l'Afrique issus de la RIT, nous les qualifierons de textes de rang 1 et les références de textes (qui ne sont donc pas directement des articles) seront qualifiés de rang 2.

Le point central d'une visualisation en réseau est la mise en relation des nœuds entre eux à travers des liens. Dans notre exemple, les nœuds sont des textes et le lien (orienté) est le fait d'être cité pour un texte de rang 2 par un texte de rang 1.

Après un travail sur l'aspect et la spatialisation, nous obtenons le graphe suivant :

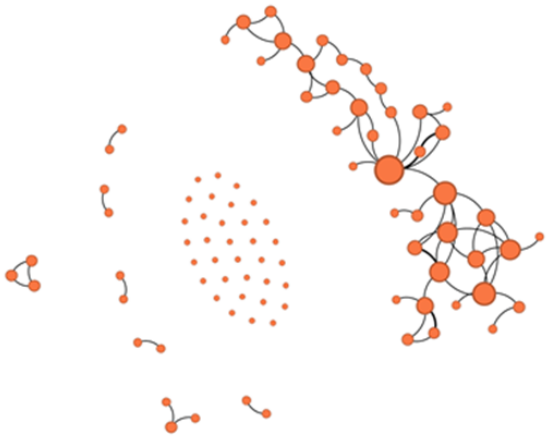
Benaze Maud Benichou Duhil de, *Modélisation hybride de réseaux dans un « champ criminel » : contribution des sciences sociales et d'outils logiciels au renseignement criminel*, thèse de doctorat, Université Michel de Montaigne - Bordeaux III, 2023.

[26] Hakim Nader et Monti Annamaria, « Histoire de la de la pensée juridique et analyse bibliométrique : l'exemple de la circulation des idées entre la France et l'Italie à la Belle Epoque », *Clio@Thémis : Revue électronique d'histoire du droit*, n° 14, 2018, coll. « L'histoire de la pensée juridique : historiographie, actualité et enjeux », [https//doi.org/10.35562/cliothemis.763].
Lavaud-Legendre Bénédicte, Melançon Guy, Pinaud Bruno, Plessard Cécile et Feron Norbert, *Analyse et visualisation des réseaux criminels*, COMPTRASEC - CNRS - UMR 5114 ; LABRI - UMR 5800.
Lavaud-Legendre Bénédicte, Plessard Cécile, Desquesnes Gillonne, Proia-Lelouey Nadine, Encrenaz Gaelle et Debruyne Gautier, *Prostitution de mineures - Parcours de vie des individus impliqués dans la prostitution par plans*, CNRS COMPTRASEC UMR 5114.



Représentation en réseau des textes qualifiés de rang 1 et de rang 2. (Réalisé à l'aide du logiciel Gephi)

Une représentation simple d'un réseau ne permet pas forcément de répondre aux questions alors posées. Dans notre cas, nous souhaitons voir comment il est possible de détecter des communautés épistémiques dans les articles sur l'Afrique de la RIT. Or, les liens qui apparaissent sont avant tout ceux des citations de chaque article. Ce que nous souhaitons faire, c'est créer un réseau d'articles de rang 1, reliés entre eux lorsque ces articles partagent une ou plusieurs citations communes. Cela équivaut à passer d'un graphe bipartite à un graphe monopartite.

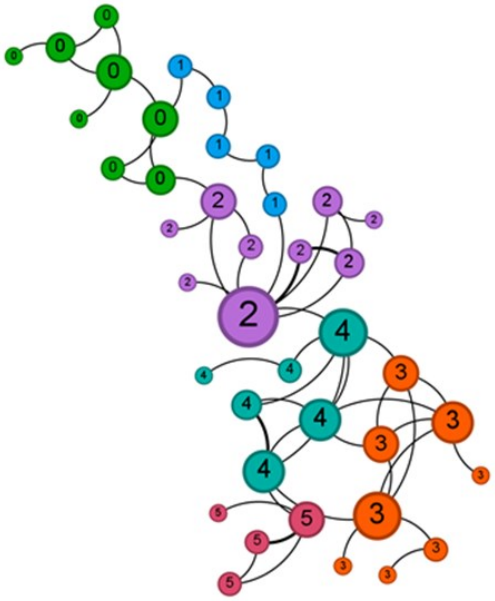


Représentation graphique du réseau monopartite regroupant les textes de rang 1 partageant une ou plusieurs citations. (Réalisé à l'aide du logiciel Gephi)

Le graphe qui découle de cette procédure nous permet d'observer ainsi des textes totalement isolés (au centre) et des textes plutôt isolés en dyades ou triades à gauche. C'est sur l'élément de droite que va porter notre attention désormais, comment catégoriser cet ensemble ? N'y a-t-il pas des sous-graphes ou communautés que nous pourrions faire ressortir ?

Il est possible de détecter ces communautés. En calculant la modularité[27], des classes vont être créées et le graphe sera alors scindé en plusieurs sous-graphes ayant le plus de liens possibles entre les nœuds du sous-graphe (ici des textes de rang 1) et le moins possible de liens entre chaque sous-graphe.

Nous obtenons le découpage suivant[28] :



Représentation graphique en sous-graphes suite au calcul de modularité (Réalisé à l'aide du logiciel Gephi)

Il est ainsi possible à ce stade de décrire les communautés avec les attributs des articles de chaque groupe.

Dans notre exemple :

Catégories	Caractéristiques
0	Textes des années 1980, début des années 1990. Afrique de l'ouest
1	Textes des années 1980. Afrique tropicale
2	Afrique subsaharienne
3	Textes d'après-guerre, années 50. Afrique subsaharienne
4	Textes des années 1950 et 1960. Afrique
5	Textes des années 1950 et 1960. Afrique francophone

Nous pouvons lire de la manière suivante : la communauté « 3 » regroupe principalement des textes d'après-guerre ayant pour périmètre géographique l'Afrique subsaharienne.
De manière générale, ces communautés peuvent être comparées aux classes générées grâce aux méthodes de classifications hiérarchiques, à la différence près que c'est la dimension relationnelle qui sera au cœur du découpage. Il est possible ensuite, comme nous l'avons fait, de caractériser ces groupes avec les caractéristiques des unités qui les composent.

[27] Blondel Vincent D., Guillaume Jean-Loup, Lambiotte Renaud et Lefebvre Etienne, « Fast unfolding of communities in large networks », *Journal of Statistical Mechanics: Theory and Experiment*, vol. P10008, 2008, p. 1-12, [https//doi.org/10.1088/1742-5468/2008/10/P10008]
[28] Notons que la numérotation de 0 à 5 ne renvoie pas à une quelconque échelle, il s'agit simplement d'étiquettes

Mot du praticien

L'analyse de réseaux est donc une alternative à l'approche fréquentiste « classique ». Elle a été conçue pour étudier la relation et l'interdépendance entre les données d'une même base et connaît à l'heure actuelle un essor très important. Néanmoins, la question fondamentale et qui prédétermine toute l'analyse de réseau c'est de savoir ce qui fait réseau. Ainsi, il est essentiel de prendre le temps de définir ce que sont nos individus et ce qui fait le lien.

Par ailleurs, l'analyse de réseaux est surtout connue pour sa représentation graphique et la possibilité de calculer un certain nombre d'indicateurs plus ou moins complexes, mais en tant que cadre alternatif, il est aussi possible de réaliser un certain nombre de modèles statistiques. Celui qui semble être le plus fréquent en SHS est le modèle ERGM. Il vise à expliquer la création des liens dans un réseau en fonction d'un certain nombre de variables explicatives. Statistiquement le modèle ERGM ressemble beaucoup à une régression logistique mais avec la différence fondamentale que dans son cas les données ne sont pas indépendantes. Nous pouvons également citer le modèle SIENA développé par Tom Snijders qui est une variante du modèle ERGM mais adapté à une analyse de réseaux longitudinale.

Si dans les différents travaux scientifiques nous notons assez peu de croisements entre les différentes disciplines qui utilisent l'analyse de réseaux, il existe malgré tout une communauté active et dynamique qui a mis en place un certain nombre de rendez-vous annuels permettant de structurer cette communauté interdisciplinaire et de diffuser les différentes innovations, tels que le colloque de la Sunbelt au niveau international et le colloque Frognet au niveau francophone.

Ressources

- Barabási Albert-László, Albert Réka et Jeong Hawoong, « Mean-field theory for scale-free random networks », *Physica A: Statistical Mechanics and its Applications*, nº 1-2, vol. 272, 1999, p. 173-187, [https//doi.org/10.1016/S0378-4371(99)00291-5].
- Barnes J. A., « Class and Committees in a Norwegian Island Parish », *Human Relations*, nº 1, vol. 7, 1954, p. 39-58, [https//doi.org/10.1177/001872675400700102].
- Beauguitte Laurent, 2023,« L'analyse de réseau en sciences sociales. Petit guide pratique ».
- Beauguitte Laurent, 2022, *Théorie des graphes et analyse de réseau en géographie : histoire d'un lien faible (1950-1963)*, [https://ouest-edel.univ-nantes.fr/passerelleshs/index.php?id=155].
- Beauguitte Laurent et Ognyanova Katherine, « Visualisation de réseaux avec R », , 2017, [https//doi.org/10.58079/CZE1].
- Benaze Maud Benichou Duhil de, *Modélisation hybride de réseaux dans un « champ criminel » : contribution des sciences sociales et d'outils logiciels au renseignement criminel*, thèse de doctorat, Université Michel de Montaigne - Bordeaux III, 2023.
- Bidart Claire, Degenne Alain et Grossetti Michel, *La vie en r*  *seau*, Presses Universitaires de France, 2011, [https//doi.org/10.3917/puf.bidar.2011.01].
- Blondel Vincent D., Guillaume Jean-Loup, Lambiotte Renaud et Lefebvre Etienne, « Fast unfolding of communities in large networks », *Journal of Statistical Mechanics: Theory and Experiment*, vol. P10008, 2008, p. 1-12, [https//doi.org/10.1088/1742-5468/2008/10/P10008].
- Dupont Benoît, « La gouvernance polycentrique du cybercrime : les réseaux fragmentés de la coopération internationale », *Cultures & conflits*, nº 102, 2016, p. 95-120, [https//doi.org/10.4000/conflits.19292].

- Forsé Michel et Degenne Alain, *Les réseaux sociaux:*, Armand Colin, 2004, [https//doi.org/10.3917/arco.forse.2004.01].
- François Axelle, Nolet Anne-Marie et Morselli Carlo, « Sociabilité carcérale et réinsertion »:, *Déviance et Société*, nº 2, Vol. 42, 2018, p. 389-419, [https//doi.org/10.3917/ds.422.0389].
- Freeman Linton C., *The development of social network analysis: a study in the sociology of science*, Vancouver (B.C.), Empirical press, 2004.
- Garrison, L. William, Beauguitte Laurent, Beauguitte Pierre et Gourdon Paul, « William L. Garrison, 1960, Connectivity of the Interstate Highway System.Version bilingue et commentée ».
- Garrison William L., « Connectivity of the interstate highway system », *Papers in Regional Science*, nº 1, vol. 6, 1960, p. 121-137, [https//doi.org/10.1111/j.1435-5597.1960.tb01707.x].
- Hakim Nader et Monti Annamaria, « Histoire de la de la pensée juridique et analyse bibliométrique : l'exemple de la circulation des idées entre la France et l'Italie à la Belle Epoque », *Clio@Thémis : Revue électronique d'histoire du droit*, nº 14, 2018, coll. « L'histoire de la pensée juridique : historiographie, actualité et enjeux », [https//doi.org/10.35562/cliothemis.763].
- Kreager Derek A., Schaefer David R., Bouchard Martin, Haynie Dana L., Wakefield Sara, Young Jacob et Zajac Gary, « Toward a Criminology of Inmate Networks », *Justice Quarterly*, nº 6, vol. 33, 2016, p. 1000-1028, [https//doi.org/10.1080/07418825.2015.1016090].
- Lavaud-Legendre Bénédicte, Melançon Guy, Pinaud Bruno, Plessard Cécile et Feron Norbert, *Analyse et visualisation des réseaux criminels*, COMPTRASEC - CNRS - UMR 5114 ; LABRI - UMR 5800.
- Lavaud-Legendre Bénédicte, Plessard Cécile, Desquesnes Gillonne, Proia-Lelouey Nadine, Encrenaz Gaelle et Debruyne Gautier, *Prostitution de mineures - Parcours de vie des individus impliqués dans la prostitution par plans*, CNRS COMPTRASEC UMR 5114.
- Lazega Emmanuel, *Réseaux sociaux et structures relationnelles:*, Presses Universitaires de France, coll. « Que sais-je ? », 2007, [https//doi.org/10.3917/puf.lazeg.2007.01].
- Lazega Emmanuel, « Analyse de réseaux d'une organisation collégiale : les avocats d'affaires », *Revue française de sociologie*, nº 4, vol. 33, 1992, p. 559-589, [https//doi.org/10.2307/3322226].
- Mercklé Pierre, *La sociologie des réseaux sociaux:*, La Découverte, coll. « Repères », 2011, [https//doi.org/10.3917/dec.merck.2011.01].
- Moreno J. L., *Who shall survive?: A new approach to the problem of human interrelations.*, Washington, Nervous and Mental Disease Publishing Co, 1934, [https//doi.org/10.1037/10648-000].
- Watts Duncan et Strogatz Steven, « Collective dynamics of 'small-world' networks », *Nature*, nº 6684, vol. 393, 1998, p. 440-442, [https//doi.org/10.1038/30918].
- White Harrison C., Grossetti Michel et Godart Frédéric, *Identité et contrôle: une théorie de l'émergence des formations sociales*, Nouvelle éd. révisée., Paris, Éd. de l'EHESS, coll. « EHESS translations », 2011.

Quelques sites Internet :

- Gourdon Paul, *Tuto@MATE - GEPHI*, [https://mate-shs.cnrs.fr/actions/tutomate/tuto08-gephi-gourdon/].
- Ognyanova Katya, « Static and dynamic network visualization with R ».[https://kateto.net/network-visualization]
- Organisation Internationale du Travail, *Revue internationale du Travail*, [https://my.visme.co/view/4dyy1m3y-revue-internationale-du-travail].

Analyse Sémantique n.f. [/a.na.liz se.mɑ̃.tik/]

Voir entrée **Analyses Qualitatives**.

Analyse Thématique n.f. [/a.na.liz te.ma.tik/]

Voir entrée **Analyses Qualitatives**.

ANOVA n.f [/əˈnoʊvə/]

Synonymes : Analyse de variance, Analysis of Variance, Comparaison de moyennes

A quoi ça sert ?

Dans la recherche en SHS, il peut s'avérer nécessaire de réaliser des comparaisons de **moyennes**. Par exemple, pour comparer le temps moyen passé sur les écrans selon le type de diplôme obtenu ou la profession exercée. Pour comparer les moyennes de deux groupes (indépendants ou appariés), nous utilisons les tests de la famille des **tests t.** Afin de comparer les moyennes entre plus de deux groupes nous ferons appel à l'ANOVA. Le test t n'est en réalité qu'un cas particulier de l'Analyse de variance (Howell, D., C., 2008)[29].

D'où ça vient ?

La parternité des méthodes d'analyse de variance est attribuée à Fisher. Il a en effet écrit différents articles annonçant les prémices du développement de ces techniques en utilisant pour la première fois le terme statistique de variance (Par exemple, Fisher, R., A. 1919) [30] et surtout rédigé un ouvrage où il détaille ces méthodes d'analyse des données (Fisher, R., A., 1925) [31].

Mot du praticien

L'ANOVA peut s'appliquer pour des mesures inter-sujet (Between), les moyennes sont alors comparées entre des groupes constitués d'individus différents. Par exemple, l'ANOVA peut être employée afin de comparer le score de confiance en la Science et le candidat pour lequel les participants ont voté lors des élections présidentielles américaines. Mais elle peut aussi s'appliquer sur des mesures intra-sujet (Within), les moyennes sont alors comparées sur des temps différents (ou tâches différentes) mais pour les mêmes sujets. Les mêmes sujets sont alors soumis aux mêmes mesures mais sur des temps différents. Par exemple, il sera possible de comparer le score de confiance en la Science sur différents temps de mesure comme avant, pendant et après la pandémie de COVID19. Il est également possible de réaliser des modèles d'ANOVA mélangeant ces deux approches inter et intra-sujets.

Les différences étant calculées sur plus de deux groupes lors d'une ANOVA, celle-ci doit être complétée par une analyse supplémentaire appelée post-hoc afin de déterminer entre quels groupes précisément se situe la différence observée, le cas échéant.

Pour quantifier la différence observée il sera nécessaire de faire appel à un indice de taille d'effet. Dans le cas de l'ANOVA nous n'emploierons pas le d de Cohen comme pour les **tests t**, mais traditionnellement nous mobiliserons l'Eta carré (ou l'oméga carré).

Avant de réaliser une ANOVA il faut s'assurer que la moyenne de notre échantillon soit bien représentative de la distribution de la variable, qu'il faudra alors examiner. Si la distribution au sein

---

[29] Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.

[30] Fisher R. A., « XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. », *Transactions of the Royal Society of Edinburgh*, nº 2, vol. 52, 1919, p. 399-433, [https//doi.org/10.1017/S0080456800012163].

[31] Fisher, R. A., *Statistical methods for research workers*, 1st éd., Oliver and Boyd, 1925.

de chaque groupe s'approche de la *loi normale*, alors la majorité des valeurs sont regroupées autour de la valeur moyenne : la moyenne est donc bien représentative de l'échantillon. Il est aussi possible que la moyenne ne soit pas l'indicateur le plus adapté pour rendre compte des données, et dans ce cas, il faut se tourner vers un autre type de test, tels que les *tests non-paramétriques* comme notamment le test de **Mann-Whitney**.

Pour les ANOVA inter-sujets, il est également nécessaire de s'assurer que les variances des groupes sont homogènes (la répartition des valeurs autour de la valeur moyenne est dans les mêmes proportions). Afin de vérifier si les variances sont homogènes (on parle aussi de variances égales ou d'**homoscédasticité**), nous employons **le test de Lévène**.

Pour les ANOVA intra-sujets, il faudra en outre vérifier la **sphéricité** des données grâce au test de Mauchly. Si la sphéricité des données n'est pas respectée, il existe des corrections telles que celle de Greenhouse-Geisser (1959)[32] ou encore celle de Huynh-Feldt (1976)[33].

Comme avant de réaliser tout test statistique, avant de réaliser une ANOVA il est donc nécessaire de vérifier les statistiques descriptives et les conditions d'application (citées précédemment). Dans le cas de l'ANOVA il s'agit aussi de vérifier que les différentes moyennes comparées ne sont pas strictement égales, auquel cas, l'intérêt du test disparaît bien évidemment.

Ressources

- Cohen Jacob, *Statistical Power Analysis for the Behavioral Sciences*, 0 éd., Routledge, 2013, [https//doi.org/10.4324/9780203771587].
- Fisher, R. A., *Statistical methods for research workers*, 1st éd., Oliver and Boyd, 1925.
- Fisher R. A., « XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. », *Transactions of the Royal Society of Edinburgh*, nº 2, vol. 52, 1919, p. 399-433, [https//doi.org/10.1017/S0080456800012163].
- Greenhouse Samuel W. et Geisser Seymour, « On Methods in the Analysis of Profile Data », *Psychometrika*, nº 2, vol. 24, 1959, p. 95-112, [https//doi.org/10.1007/BF02289823].
- Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.
- Huynh Huynh et Feldt Leonard S., « Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs », *Journal of Educational Statistics*, nº 1, vol. 1, 1976, p. 69, [https//doi.org/10.2307/1164736].

[32] Greenhouse Samuel W. et Geisser Seymour, « On Methods in the Analysis of Profile Data », *Psychometrika*, nº 2, vol. 24, 1959, p. 95-112, [https//doi.org/10.1007/BF02289823].

[33] Huynh Huynh et Feldt Leonard S., « Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs », *Journal of Educational Statistics*, nº 1, vol. 1, 1976, p. 69, [https//doi.org/10.2307/1164736].

Arbre de décision n.m [/aʁbʁ də de.si.zjɔ̃/]

Synonymes : Decision Tree

À quoi ça sert ?

Les arbres de décision sont un outil polyvalent d'apprentissage supervisé, permettant d'expliquer une variable réponse grâce à une suite de choix successifs. Leur représentation graphique sous forme d'arbre, où chaque choix est symbolisé par un embranchement, leur a donné leur nom. Ils peuvent être employés aussi bien à des fins exploratoires que prédictives, et sont utilisables avec tous les types de variables (catégorielles, continues et ordinales).

Plus formellement, un arbre de décision vise à prédire une variable qualitative (on parle alors d'arbre de classification) ou quantitative (arbre de régression) en fonction d'un ensemble de variables explicatives qui peuvent aussi bien être quantitatives que qualitatives, au sein du même modèle. Le principe de l'algorithme consiste à diviser successivement l'échantillon de données en sous-ensembles de plus en plus homogènes quant à la variable réponse. Chacune de ces divisions est déterminée en trouvant une "question" binaire portant sur les variables explicatives, qui permet de séparer les données en deux groupes maximalement différents. Le processus est itéré sur chacun des groupes ainsi créés, opérant ainsi un partitionnement récursif des données, jusqu'à rencontrer des critères d'arrêt définis par l'utilisateur.

Un arbre pousse depuis sa racine (l'échantillon complet), en formant des nœuds qui divisent récursivement l'échantillon en sous-ensembles, pour aboutir à des feuilles terminales donnant une valeur finale associée à des combinaisons précises de critères portant sur les variables explicatives. L'intérêt principal des arbres de décision réside dans leur simplicité d'interprétation. Les représentations visuelles sont faciles à comprendre[34], et permettent de capturer des relations et des interactions complexes entre les variables explicatives.

D'où ça vient ?

Le premier algorithme d'arbres de décision a été mis au point par Morgan & Sonquist (1963), et un travail très actif de recherche a exploré et étendu ce concept dans de nombreuses directions (Loh, 2014). Il existe aujourd'hui un grand nombre d'algorithmes d'arbres de décision : C4.5, CHAID, CART, Conditional Inference Tree, etc. Tous correspondent à des problématiques et des philosophies légèrement différentes, mais reprennent les idées fondamentales de l'algorithme historique CART (Classification And Regression Trees), publié par Breiman et al. (1984 – 1ère éd.)[35]. CART reste encore l'algorithme le plus utilisé aujourd'hui, en particulier grâce à sa flexibilité et sa polyvalence.

Pour un bon aperçu des aspects théoriques de l'algorithme CART ainsi que de ses alternatives les plus courantes, on pourra notamment consulter Genuer & Poggi (2019)[36].

---

[34] Toutefois, ne pas confondre les représentations des arbres de décision avec les dendrogrammes habituellement issus des méthodes de *clustering* (classification non supervisée) telles que la classification ascendante hiérarchique – ces dernières méthodes répondant à une problématique radicalement différente.

[35] Breiman Leo, Friedman Jerome H., Olshen Richard A. et Stone Charles J., *Classification And Regression Trees*, 1re éd., Routledge, 2017, [https//doi.org/10.1201/9781315139470].

[36] Genuer Robin et Poggi Jean-Michel, *Les forêts aléatoires avec R*, Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2019.

Le mot du praticien

Les arbres de décision présentent des avantages qui les distinguent de manière intéressante des autres méthodes d'apprentissage supervisé présentées dans ce dictionnaire.

- Leur simplicité et l'application immédiate des règles menant à la décision.
- Les variables explicatives peuvent être de différents types : un même arbre peut utiliser conjointement des variables qualitatives, ordinales et continues, en appliquant à chacune d'entre elles des critères appropriés.
- La gestion des données manquantes est particulièrement efficace, et est au coeur de l'algorithme. Aucune imputation ni suppression de données manquantes n'est nécessaire en amont : l'algorithme CART gère nativement les valeurs manquantes, en exploitant la corrélation entre les variables.
- Contrairement à la plupart des autres méthodes d'apprentissage supervisé, la multicolinéarité des variables explicatives ne pose donc aucun problème. De même, aucune sélection de variables n'est nécessaire en amont, puisque l'algorithme CART sélectionne lui-même les meilleures variables à chaque nœud.
- En particulier, cela signifie également que l'algorithme CART peut sans problème être utilisé dans des cas où le nombre de variables explicatives est très supérieur au nombre d'individus.

Les arbres de décision peuvent donc être utilisés dans des cas extrêmement défavorables où la plupart des autres méthodes seraient inapplicables. Ils peuvent également fournir une excellente compréhension d'un phénomène, en délivrant une représentation schématique des interactions entre variables explicatives, et de leurs liens avec la variable réponse. Toutefois, lorsque nous souhaitons réaliser des prédictions sur un nouvel échantillon, il est plus adapté de faire appel à des modèles de **régressions,** par exemple, en particulier lorsque les données sont de bonne qualité (c'est-à-dire lorsque tous les pré-requis à leurs applications sont remplis) (e.g., Galibourg et al., 2021 ; Maroco et al., 2011).

Exemples d'application

Un jeu de données fourni dans Fox & Weisberg (2019) vise à expliquer le salaire d'enseignants-chercheurs dans une université américaine en fonction de leur ancienneté, leur discipline, leur statut, leur genre, etc. Le but était notamment de rechercher de possibles disparités de salaire liés au genre. Un arbre de régression expliquant la variable continue "Salaire" en fonction de tous ces différents prédicteurs numériques (e.g., les années d'ancienneté) ou qualitatifs (e.g., la discipline) peut être un bon premier pas pour explorer cette problématique. La variable genre apparaît-elle dès le début sur cet arbre de décision, suggérant qu'il y a une inégalité de traitement globale liée au genre ? Ou peut-être n'apparaît-elle qu'en association avec un certain statut, ou une certaine discipline, ou un certain niveau d'ancienneté, suggérant ainsi plutôt des inégalités "localisées" ? Un arbre de régression pourrait permettre de démêler les interactions complexes entre variables prédictives permettant d'expliquer la variabilité observée sur les niveaux de salaire.

Il existe aussi des exemples en géographie, par exemple Marie Faulon (2020) qui a utilisé les arbres de décision pour étudier le niveau de disparités des équipement électriques et sanitaires des lodges situées dans le Khumbu au Népal.

Ces méthodes pourraient être tout-à-fait pertinentes afin de gérer des valeurs manquantes ou lorsqu'il y a une sélection de variables à opérer, pourtant elles sont encore assez peu utilisées car mal connues.

Un des points faibles des arbres de décisions, par rapport notamment aux méthodes de **régressions**, c'est qu'ils ne fournissent pas de coefficients permettant de mesurer la force de l'impact des variables explicatives et des modalités sur le partitionnement de la variable à expliquer. Ceci dit, il est malgré tout possible de mesurer l'importance d'une variable explicative dans la construction de l'arbre. Plus une variable interviendra daens des divisions importantes plus son impact sera fort. On peut le mesurer par la capacité de la variable à réduire les erreurs dans le modèle.

Ressources

- Breiman Leo, Friedman Jerome H., Olshen Richard A. et Stone Charles J., *Classification And Regression Trees*, 1re éd., Routledge, 2017, [https//doi.org/10.1201/9781315139470].
- Faulon Marie et Sacareau Isabelle, « Tourisme, gestion sociale de l'eau et changement climatique dans un territoire de haute altitude : le massif de l'Everest au Népal », *Revue de géographie alpine*, nº 108-1, 2020, [https//doi.org/10.4000/rga.6759].
- Fox John et Weisberg Sanford, *An R companion to applied regression*, Third edition., Los Angeles London New Delhi Singapore Washington, DC Melbourne, SAGE, 2019.
- Galibourg Antoine, Cussat-Blanc Sylvain, Dumoncel Jean, Telmon Norbert, Monsarrat Paul et Maret Delphine, « Comparison of different machine learning approaches to predict dental age using Demirjian's staging approach », *International Journal of Legal Medicine*, nº 2, vol. 135, 2021, p. 665-675, [https//doi.org/10.1007/s00414-020-02489-5].
- Genuer Robin et Poggi Jean-Michel, *Les forêts aléatoires avec R*, Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2019.
- Loh Wei-Yin, « Fifty Years of Classification and Regression Trees », *International Statistical Review*, nº 3, vol. 82, 2014, p. 329-348, [https//doi.org/10.1111/insr.12016].
- Maroco João, Silva Dina, Rodrigues Ana, Guerreiro Manuela, Santana Isabel et De Mendonça Alexandre, « Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests », *BMC Research Notes*, nº 1, vol. 4, 2011, p. 299, [https//doi.org/10.1186/1756-0500-4-299].
- Morgan James N. et Sonquist John A., « Problems in the Analysis of Survey Data, and a Proposal », *Journal of the American Statistical Association*, nº 302, vol. 58, 1963, p. 415-434, [https//doi.org/10.1080/01621459.1963.10500855].

Autocorrélation n.f. [/otokɔʁelasjɔ̃/]

Synonymes : Autocorrelation, Non-indépendance

A quoi ça sert ?

Attention, la présentation ne porte que sur l'autocorrélation des résidus car elle est présentée ici comme étant un pré-requis de l'inférence.

Le terme d'autocorrélation renvoie à la répartition des résidus et au risque que ceux-ci ne soient pas distribués au hasard. Le terme de résidus exprime la part de variance non expliquée par le modèle statistique testé pour chaque individu. On parle également de la part d'erreur du modèle. Par exemple, si nous nous intéressons à la relation poids/taille de la population générale, nous obtiendrons une relation avec un modèle qui rendra compte de cette relation. Toutefois, si nous ne mesurons pas certains éléments tels que le fait que notre échantillon puisse contenir des femmes enceintes par exemple (ce qui influence nécessairement la relation poids/taille), alors la part résiduelle non-expliquée par le modèle pour ces personnes sera plus importante que pour les autres. L'autocorrélation des résidus, si elle est avérée, peut traduire la présence d'éléments non-mesurés, non pris en compte dans notre analyse et qui pourtant viendraient influencer les résultats obtenus. C'est pourquoi avant de réaliser des **modèles de régression** il est essentiel de vérifier si nous sommes en situation d'autocorrélation des résidus ou non.

D'où ça vient ?

L'autocorrélation est donc une mesure de la relation entre les différents résidus du modèle. La relation entre ces résidus ne peut être mesurés par les tests de **corrélations** classiques tel que celui de Pearson par exemple. L'un des tests les plus connu pour mesurer la présence ou non d'autocorrélation dans le cadre d'un modèle de régression est celui développé par James Durbin et Geoffrey Watson au début des années 1950 [37],[38]. D'autres tests ont été développé par la suite afin de prendre en compte davantage d'éléments liés à l'autocorrélation. Nous pouvons citer le test de Breusch-Godfrey dans les années 1970 [39],[40], développé afin de prendre en compte une mesure plus générale de l'autocorrélation que celui de Durbin-Watson. Le test de Liung-Box a été développé à la même période [41] et est actuellement très utilisé en économétrie ou dans le cadre de traitement de séries temporelles.

Il existe différentes formes d'autocorrélation. Elles peuvent être imputables à la structure des données, à la prise en compte de mesures temporelles (qui fait qu'un individu risque d'être fortement corrélé avec lui-même dans le temps) ou encore **l'autocorrélation spatiale** (un individu corrèle plus fortement avec ses voisins).

---

[37] Durbin J. et Watson G. S., « Testing for Serial Correlation in Least Squares Regression. II », *Biometrika*, nº 1/2, vol. 38, 1951, p. 159, [https//doi.org/10.2307/2332325].

[38] Durbin J. et Watson G. S., « Testing for Serial Correlation in Least Squares Regression. II », *Biometrika*, nº 1/2, vol. 38, 1951, p. 159, [https//doi.org/10.2307/2332325].

[39] Breusch T. S., « TESTING FOR AUTOCORRELATION IN DYNAMIC LINEAR MODELS* », *Australian Economic Papers*, nº 31, vol. 17, 1978, p. 334-355, [https//doi.org/10.1111/j.1467-8454.1978.tb00635.x].

[40] Godfrey L. G., « Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables », *Econometrica*, nº 6, vol. 46, 1978, p. 1293, [https//doi.org/10.2307/1913829].

[41] Ljung G. M. et Box G. E. P., « On a measure of lack of fit in time series models », *Biometrika*, nº 2, vol. 65, 1978, p. 297-303, [https//doi.org/10.1093/biomet/65.2.297].

Ressources

- Bressoux Pascal, *Modélisation statistique appliquée aux sciences sociales*, De Boeck Supérieur, 2010, [https//doi.org/10.3917/dbu.bress.2010.01].
- Breusch T. S., « TESTING FOR AUTOCORRELATION IN DYNAMIC LINEAR MODELS* », *Australian Economic Papers*, nº 31, vol. 17, 1978, p. 334-355, [https//doi.org/10.1111/j.1467-8454.1978.tb00635.x].
- Darlington Richard B. et Hayes Andrew F., *Regression analysis and linear models: concepts, applications, and implementation*, New York London, The Guilford Press, coll. « Methodology in the social sciences », 2017.
- Durbin J. et Watson G. S., « Testing for Serial Correlation in Least Squares Regression. II », *Biometrika*, nº 1/2, vol. 38, 1951, p. 159, [https//doi.org/10.2307/2332325].
- Godfrey L. G., « Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables », *Econometrica*, nº 6, vol. 46, 1978, p. 1293, [https//doi.org/10.2307/1913829].
- Ljung G. M. et Box G. E. P., « On a measure of lack of fit in time series models », *Biometrika*, nº 2, vol. 65, 1978, p. 297-303, [https//doi.org/10.1093/biomet/65.2.297].

Autocorrélation spatiale n.f [o.to.kɔ.ʁe.la.sjɔ̃ spa.sjal]

Synonymes : I de Moran, C de Geary

A quoi ça sert ?

L'auto-corrélation est la corrélation, positive ou négative, d'une variable avec elle-même. Lorsqu'elle est spatiale, elle se définit comme la corrélation, d'une variable avec elle-même du fait de la localisation géographique des observations.

L'auto-corrélation spatiale permet d'étudier la relation existante, ou non, entre une valeur d'une variable et la valeur de cette même variable pour l'individu voisin et/ou ses voisins. Cette méthode de statistique spatiale permet notamment d'étudier la dimension spatiale des données. Ainsi, l'auto-corrélation spatiale est positive lorsque les valeurs similaires de la variable à étudier se regroupent géographiquement. Elle est au contraire négative lorsque des valeurs différentes de cette variable se regroupent géographiquement, c'est à dire que les lieux proches sont plus différents que les lieux éloignés. L'absence d'auto-corrélation spatiale se traduit par une répartition spatiale aléatoire des observations.

L'usage de l'auto-corrélation est également utile pour éviter un biais important dans l'analyse statistique des données. De nombreuses analyses (analyse des corrélations, régressions linéaires, etc.) reposent sur l'hypothèse d'indépendance des observations et une absence de multicolinéarité entre les variables. Or, si une variable est spatialement auto-corrélée, le pré-requis d'indépendance n'est plus respecté. C'est l'hypothèse de la dépendance spatiale, une notion fondamentale en géographie, souvent évoquée comme la première loi de la géographie. Tobler (1970) la décrit sous ces termes : « Tout est lié à tout, mais les objets proches le sont davantage que les objets éloignés ».

L'auto-correlation permet ainsi de vérifier les conditions d'application d'un certain nombre de méthodes et dans le même temps permet l'étude de la structure spatiale, ce qui peut rendre ces mêmes méthodes inapplicables.

L'auto-corrélation spatiale repose sur un test d'hypothèse où H_0 renvoie à l'absence d'auto-corrélation spatiale et H_1 correspond à la présence d'auto-corrélation spatiale.

D'où ça vient ?

Les réflexions sur l'auto-corrélation spatiale débutent dans les années 1950. Il s'agit de mettre en évidence une relation plus marquée entre des voisins qu'avec le reste de la population. Le premier article fondateur est celui de Patrick Moran (1950), suivi par celui Roy Geary (1954). Ils ont donné naissance aux deux indices d'auto-corrélation spatiale les plus utilisés : le I de Moran et le c de Geary. C'est dans les années 1970 que la notion d'auto-corrélation sort du champ de la statistique et commence à se diffuser dans les disciplines des sciences humaines et sociales grâce à Andrew Cliff et Keith Ord (1970). Ces derniers remanient profondément les statistiques de Moran et Geary (Cliff A. et Ord K., 1969 et 1973). Leur version réécrite du I de Moran est aujourd'hui la plus utilisée pour mesurer l'auto-corrélation spatiale. S'il existe un grand nombre d'indices, le I de Moran est considéré comme le plus robuste (Cliff et Ord, 1981, p. 54-56 ; Upton et al. 1985) et présente également l'avantage de se lire comme un coefficient de corrélation classique.

Exemple d'application

L'auto-correlation spatiale est une méthode très utilisée en sciences humaines et sociales d'abord dans sa dimension de vérification de conditions d'applications des tests statistiques paramétriques telle que la régression linéaire, par exemple. En géographie, elle va être particulièrement utilisée pour analyser la structure spatiale d'un phénomène mesuré. Par exemple : « Y-a-t'il une distribution spatiale du revenu, de la croissance de population française, de l'équipement médical ? », etc.

Par exemple, Audard et al. en 2024 ont publié des travaux sur l'étude du prix médian de l'immobilier par EPCI (établissements publics de coopération intercommunale, découpage territorial correspondant globalement aux communautés de communes) en France métropolitaine. Dans leurs travaux les auteurs ont réalisé un test d'auto-corrélation spatiale afin de vérifier statistiquement l'existence, ou non, d'une structure spatiale des données.

L'article de Sébastien Oliveau et Yoann Doignon (2016) constitue également un très bon exemple d'application en géographie de l'autocorrélation spatiale.

Le mot du praticien

L'auto-corrélation spatiale repose sur des pré-requis impliquant la définition de la notion de voisinage d'une part et le type de données ainsi que leurs structurations d'autre part.

La définition du voisinage est essentielle car l'autocorrélation spatiale repose sur une comparaison de la valeur d'une variable prise par un individu et celle de ses voisins. Ainsi, la définition du voisinage, aussi bien de manière théorique (qui sont mes voisins ?) que pratique (construction d'une matrice de voisinage), est un pré-requis central. Sa définition va avoir un effet considérable sur le calcul et le résultat de l'autocorrélation. Etant un élément central à la réalisation d'autocorrélations spatiales la définition de voisinage a été traité et documenté dans la littérature scientifique (Oliveau, S., 2011 ; De Bellefon et al. 2018).

Il est important de savoir que le voisinage peut être défini en se basant sur 3 éléments différents :

- la contiguïté
- la distance
- la proximité

Le choix entre ces trois types de voisinage dépend de la manière dont nous souhaitons définir ce voisinage, mais également de la nature spatiale des données. Ces données sont-elles constituées de polygones ou de points ?

Le deuxième pré-requis, après la définition du voisinage, concerne les données, qui doivent bien entendu avoir une dimension spatiale mais, pour utiliser les principaux tests d'auto-corrélations spatiales (Moran et Geary), il est nécessaire que la variable d'intérêt soit continue. Dans le cas contraire, autrement dit, avec des variables catégorielles, il est malgré tout possible de réaliser une auto-corrélation spatiale. Dans ce cas, le degré d'association locale sera mesuré grâce à une analyse des « join count » (Zhukov, 2010). Pour plus d'informations sur le sujet, vous pouvez vous référer au chapitre 3 du Manuel d'analyse spatiale : Théorie et mise en oeuvre pratique avec R (De Bellefon et al. 2018).

Ressources :

- Anselin Luc, « The Moran scatterplot as an ESDA tool to assess local instability in spatial association », *Spatial Analytical Perspectives on GIS*, Routledge, 1996, .
- Anselin Luc, « Local Indicators of Spatial Association—LISA », *Geographical Analysis*, nº 2, vol. 27, 1995, p. 93-115, [https//doi.org/10.1111/j.1538-4632.1995.tb00338.x].
- Audard Frédéric, Le Campion Grégoire et Pierson Julie, « La régression géographiquement pondérée : GWR », , 2024, [https//doi.org/10.48645/WK1M-HG05].
- Cliff A. D. et Ord K. J., *Spatial autocorrelation*, Pion., London, 1973.
- Cliff A. D. et Ord K. J., « The Problem of Spatial Autocorrelation », in Allen John Scott (dir.), *Studies in regional science*, London, Pion, coll. « London papers in regional science », 1969, p. 25-55.
- Cliff Andrew D. et Ord Keith, « Spatial Autocorrelation: A Review of Existing and New Measures with Applications », *Economic Geography*, vol. 46, 1970, p. 269, [https//doi.org/10.2307/143144].
- Cliff Andrew David et Ord J. K., *Spatial processes: models and applications*, London, Pion, 1981.
- DE BELLEFON Marie-Pierre, LOONIS Vincent et LE GLEUT Renan, *Manuel d'analyse spatiale* | *Insee*, [https://www.insee.fr/fr/information/3635442].
- Geary R. C., « The Contiguity Ratio and Statistical Mapping », *The Incorporated Statistician*, nº 3, vol. 5, 1954, p. 115, [https//doi.org/10.2307/2986645].
- Getis Arthur et Ord J. K., « The Analysis of Spatial Association by Use of Distance Statistics », *Geographical Analysis*, nº 3, vol. 24, 1992, p. 189-206, [https//doi.org/10.1111/j.1538-4632.1992.tb00261.x].
- Moran P. A. P., « NOTES ON CONTINUOUS STOCHASTIC PHENOMENA », *Biometrika*, nº 1-2, vol. 37, 1950, p. 17-23, [https//doi.org/10.1093/biomet/37.1-2.17].
- Muller Jean-Claude, « Comparaison visuelle des cartes et groupements spatiaux », *L'Espace géographique*, nº 1, vol. 6, 1977, p. 59-72, [https//doi.org/10.3406/spgeo.1977.1694].
- Oliveau Sébastien, « Autocorrélation spatiale ».
- Oliveau Sébastien, *L'espace compte ! Mesurer les structures spatiales du changement social.*, thèse de doctorat, Université d'Aix-Marseille 1, 2011.
- Oliveau Sébastien, « Autocorrélation spatiale : leçons du changement d'échelle: », *L'Espace géographique*, nº 1, Vol. 39, 2010, p. 51-64, [https//doi.org/10.3917/eg.391.0051].
- Oliveau Sébastien et Doignon Yoann, « La diagonale se vide ? Analyse spatiale exploratoire des décroissances démographiques en France métropolitaine depuis 50 ans », *Cybergeo*, , 2016, [https//doi.org/10.4000/cybergeo.27439].
- Tobler W. R., « A Computer Movie Simulating Urban Growth in the Detroit Region », *Economic Geography*, vol. 46, 1970, p. 234, [https//doi.org/10.2307/143141].
- Upton Graham J. G., Fingleton Bernard et Upton Graham J. G., *Point pattern and quantitative data*, Reprinted., Chichester, Wiley, coll. « Spatial data analysis by example / Graham J. G. Upton; Bernard Fingleton », 1985.
- Wong David W., « Exploring Spatial Patterns Using an Expanded Spatial Autocorrelation Framework: Exploring Spatial Patterns », *Geographical Analysis*, nº 3, vol. 43, 2011, p. 327-338, [https//doi.org/10.1111/j.1538-4632.2011.00816.x].
- Zhukov Yuri. M., *Applied Spatial Statistics in R*, [https://zhukovyuri.github.io/teaching/].

BoxCox n.m. [/ˈbɒkskɒks/]

Synonyme : Transformée de Box-Cox

A quoi ça sert ?

La transformation de BoxCox est notamment utilisée dans le cadre de modèles de **régressions linéaires** afin de pallier à des conditions d'applications non remplies pour la réalisation de ce type de modèle, comme par exemple une répartition des résidus qui serait non-homogène.

D'où ça vient ?

Elle a été développée par Box et Cox en 1964[42]. L'idée étant d'appliquer une transformation non-linéaire à la variable dépendante afin de modifier la répartition des résidus du modèle. Cette transformation est très courante en statistique et en économétrie, mais elle est moins connue dans les autres disciplines.[43]

Il existe trois catégories de modèles avec application de la transformée de BoxCox :

- Une première classe de modèles où la transformée s'applique uniquement à la variable dépendante.
- Une deuxième classe de modèles avec une transformation dite "des deux côtés" car celle-ci s'applique à la fois à la variable dépendante et à la fonction de régression [44].
- Une troisième classe de modèles plus générale, avec la prise en compte d'observations sur la variable dépendante étant soumise à la transformée et dans le même temps non-soumise à cette transformée. Ce type de modèle de Box-Cox conventionnel est celui le plus utilisé (notamment en économétrie).[45]

La transformée de Box-Cox est une transformation non-linéaire indexée par le paramètre λ. Avec y la variable dépendante d'origine et $y^{(\lambda)}$ la variable transformée. Le paramètre λ doit être choisi pour stabiliser la variance et s'approcher d'une meilleure répartition des résidus. Différentes possibilités sont envisageables :

- Lorsque λ est égal à 1, alors il n'y a aucune transformation à faire. Les données restent inchangées.
- Lorsque λ est égal à 0, ça correspond à appliquer une transformation logarithmique.
- Lorsque λ est égal à 0,5, correspond à effectuer une transformation racine carrée.
- Lorsque λ est égal à -1, il s'agit d'une transformation réciproque.
- Lorsque λ est égal à -0,5, c'est une transformation de la réciproque de racine carrée.

---

[42] Box G. E. P. et Cox D. R., « An Analysis of Transformations », *Journal of the Royal Statistical Society Series B: Statistical Methodology*, nº 2, vol. 26, 1964, p. 211-243, [https//doi.org/10.1111/j.2517-6161.1964.tb00553.x].

[43] Davidson Russell et MacKinnon James G., *Estimation and inference in econometrics*, New York, Oxford University Press, 1993.

[44] Carroll Raymond et Ruppert David, *Transformation and Weighting in Regression*, 1st Ed. 1988., Routledge, 2017, [https//doi.org/10.1201/9780203735268].

[45] Davidson Russell et MacKinnon James G., *Estimation and inference in econometrics*, New York, Oxford University Press, 1993.

Le mot du praticien

Les transformations de BoxCox ne peuvent s'appliquer que pour des variables dépendantes ayant des valeurs positives (supérieures à 0).

La transformation de BoxCox peut s'appliquer dès qu'un modèle de **régression linéaire** comprend un problème de répartition de ses résidus, il s'agit donc d'une méthode transversale à toutes les disciplines des SHS. Nous pouvons, par exemple en psychologie, rechercher à appliquer une transformation de BoxCox dans le cadre d'un modèle de **régression linéaire** pour lequel la variable à expliquer serait le score de confiance envers la Science et les variables explicatives le score obtenu à une échelle de croyance dans les phénomènes paranormaux et le score obtenu à une échelle indiquant l'adhésion aux opinions conservatrices. Si ce modèle comprend un problème d'**homoscédasticité** alors il est pertinent d'appliquer une transformation de BoxCox à nos données afin de corriger ce problème et obtenir le modèle de régression le plus optimal.

Un autre exemple en économie, afin d'appréhender l'utilisation des cartes de fidélités par les clients de supermarchés du Nord de l'Italie, Atkinson, A. C., & Riani, M. (2006)[46] ont réalisé une **régression linéaire** avec comme variable à expliquer le montant dépensé en euros sur les 6 derniers mois et avec pour variables explicatives le nombre de visites en supermarché, l'âge des clients et la composition familiale. Afin d'étudier la transformation de BoxCox et de la comparer avec d'autres procédures ces mêmes auteurs ont appliqué une transformation de BoxCox à ce modèle de **régression linéaire** dans un article de 2023[47].

Ressources

- Atkinson Anthony C. et Riani Marco, « Distribution Theory and Simulations for Tests of Outliers in Regression », *Journal of Computational and Graphical Statistics*, nº 2, vol. 15, 2006, p. 460-476, [https//doi.org/10.1198/106186006X113593].
- Atkinson Anthony C., Riani Marco et Corbellini Aldo, « The Box–Cox Transformation: Review and Extensions », *Statistical Science*, nº 2, vol. 36, 2021, p. 239-255.
- Box G. E. P. et Cox D. R., « An Analysis of Transformations », *Journal of the Royal Statistical Society Series B: Statistical Methodology*, nº 2, vol. 26, 1964, p. 211-243, [https//doi.org/10.1111/j.2517-6161.1964.tb00553.x].
- Carroll Raymond et Ruppert David, *Transformation and Weighting in Regression*, 1st Ed. 1988., Routledge, 2017, [https//doi.org/10.1201/9780203735268].
- Davidson Russell et MacKinnon James G., *Estimation and inference in econometrics*, New York, Oxford University Press, 1993.
- Riani Marco, Atkinson Anthony C. et Corbellini Aldo, « Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression », *Statistical Methods & Applications*, nº 1, vol. 32, 2023, p. 75-102, [https//doi.org/10.1007/s10260-022-00640-7].

[46] Atkinson Anthony C. et Riani Marco, « Distribution Theory and Simulations for Tests of Outliers in Regression », *Journal of Computational and Graphical Statistics*, nº 2, vol. 15, 2006, p. 460-476, [https//doi.org/10.1198/106186006X113593].

[47] Riani Marco, Atkinson Anthony C. et Corbellini Aldo, « Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression », *Statistical Methods & Applications*, nº 1, vol. 32, 2023, p. 75-102, [https//doi.org/10.1007/s10260-022-00640-7].

CAH n.f. [/se.a.aʃ/]

Synonyme : Classification Ascendante Hiérarchique

A quoi ça sert ?

La classification ascendante hiérarchique (CAH) est une des méthodes de partitionnement/clustering les plus utilisées. Elle est globalement transversale à toutes les disciplines des sciences humaines et sociales. Cette méthode présente également l'avantage de s'appuyer sur des représentations graphiques qui sont des aides précieuses à l'interprétation.

L'objectif de la CAH va être de classer des individus statistiques (observations). Elle cherche à faire des partitions, des clusters au sein de notre population mais pas de n'importe qu'elle manière. Pour partitionner notre population la CAH va chercher à ce que les individus regroupés au sein d'une même classe soient le plus semblables possible entre eux (homogénéité intra-classe) tandis que les classes doivent être le plus dissemblables entre elles (hétérogénéité inter-classe).

Le principe de base de la CAH est donc un principe de ressemblance des individus. Cette ressemblance va se traduire par une matrice de distance qui exprime la distance entre les individus deux à deux. Plus 2 individus se ressemblent plus leur distance est faible, et plus ils sont dissemblables plus la distance est grande. La CAH va donc regrouper nos individus similaires en formant une hiérarchie des classes des plus petites à l'ensemble total des données.

Cette méthode de classification est dite ascendante car elle part des individus qui au départ sont seul dans leur propre cluster. Et hiérarchique car par fusion progressive les deux clusters les plus proches sont fusionnés pour obtenir une classe plus vaste, incluant des sous-groupes en son sein. En découpant cet arbre à une certaine hauteur choisie, on produira la partition désirée. Le découpage final doit permettre d'obtenir des classes équilibrées en termes de nombre d'observations et avoir du sens par rapport à l'hypothèse posée.

D'où ça vient ?

La question de la classification ou du clustering est depuis tout tempscentrale dans le questionnement des différentes disciplines scientifiques, elle est donc très largement étudiée. C'est d'ailleurs un champ encore très fécond qui poursuit son évolution avec de nouvelles techniques qui sont proposées encore régulièrement. Par exemple, une variante a été développée en géographie : la CAH avec contraintes de proximité géographique[48] ou encore en 2021 la CAH par compromis[49].

Sur la question spécifique de la CAH nous pouvons nous référer aux travaux de WARD (1963), et de Benzecri (1973) qui est l'auteur francophone incontournable sur cette question et plus largement sur les analyses factorielles. Un grand nombre d'ouvrages, y compris récents, présentent cette méthode de manière plus ou moins complexe. Nous pouvons notamment citer les travaux de Lebart et al. (1995), Husson et al. (2009), Cornillon et al., (2012), Pagès J. (2013). En géographie, l'ouvrage de Lena Sanders (1989) a eu un impact important pour la diffusion dans la discipline de cette

[48] Chavent Marie, Kuentz-Simonet Vanessa, Labenne Amaury et Saracco Jerome, « ClustGeo : Classification Ascendante Hiérarchique (CAH) avec contraintes de proximitté géographique », 2015, Lille, France.

[49] Bellanger Lise, Coulon Arthur et Husi Philippe, « Une méthode de classification ascendante hiérarchique par compromis : hclustcompro », 2021, Marseille, France, CIFSD, Mohamed QUAFAFOU.

méthode d'analyse.

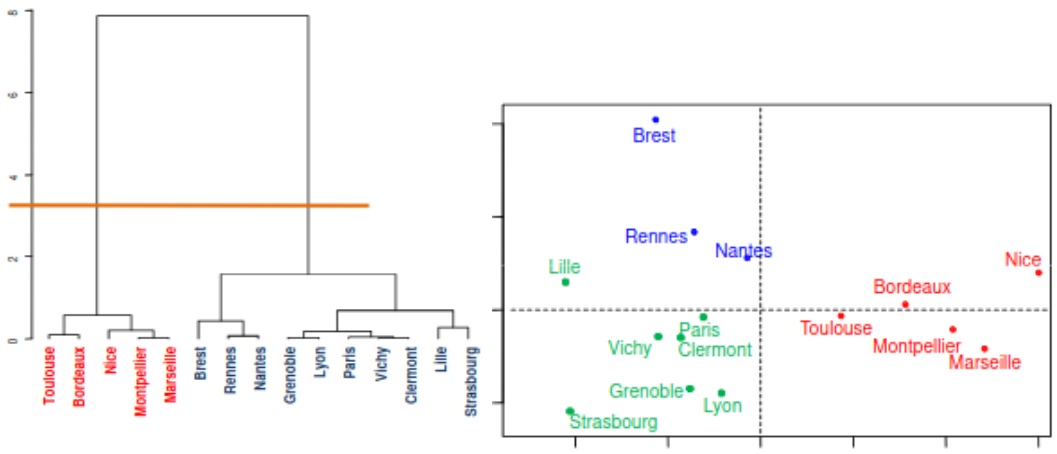
Exemple d'application

Les exemples d'application de la CAH en sciences humaines et sociales sont extrêmement nombreux. Pour une approche très pratique et pédagogique nous pouvons nous référer aux travaux de François Husson et aux exemples présents sur le site web de FactoMineR, le package qu'il a développé pour réaliser des CAH sur R. Ici, nous avons fait le choix de vous présenter un exemple en géographie. Dès 1976, des publications en géographie mobilisent la CAH. Denise Pumain (1976) proposait une typologie socioprofessionnelle des villes françaises à partir d'une analyse factorielle des correspondances (**AFC**) et d'une CAH.

Nous prendrons ici comme cas d'application l'article de Fleuret et Apparicio en 2011 qui appartient au champ de la géographie de la santé où la CAH a été employée pour obtenir une typologie des territoires de santé au Québec.
Dans cet article les auteurs souhaitent étudier la situation du système de santé de « première ligne » au Québec et pour ce faire catégoriser le territoire à l'échelle des centres de santé et des services sociaux du Québec. L'offre de services de santé de « première ligne » est définie localement à l'échelle de ces territoires québecois, en revanche l'allocation des ressources et la définition des orientations des politiques de santé se décident à une autre échelle qui est celle de la province. L'enjeu est donc de compléter la bonne connaissance du milieu local, par une vue d'ensemble du territoire. La base de données est composée d'une vingtaine de variables qui viennent caractériser chacun des territoires considérés au niveau du Québec. Il y a aura des variables telles que : la densité, la distance aux urgences, le décrochage scolaire, l'espérance de vie, etc.
La CAH a permis de faire ressortir 8 types de territoires différents qui peuvent être regroupés en profils différents. Ces profils rendent compte de réalités très disparates quant à l'accès au système de santé. Ces résultats permettent de réfléchir à des solutions pour corriger des inégalités tout en tenant compte des besoins et des disparités géographiques.

Une des grandes forces de la CAH est sa traduction systématique sur un dendrogramme et la représentation des données sur un plan factoriel. Ces représentations peuvent s'avérer particulièrement utiles pour aider à l'interprétation des classes proposées. Travail qui peut s'avérer ardu.

Exemple tiré de Husson (2009)



En Psychologie les CAH peuvent s'utiliser afin de dégager des profils de patients ayant un diagnostic de schizophrénie (Prouteau et al., 2017)[50]. Dans cet exemple, quatre profils différents ont été obtenu à partir du niveau de cognition subjectif et objectif des patients. Des analyses complémentaires ont ensuite été réalisées afin de comprendre si selon les profils dans lesquels ils étaient regroupés les patients étaient plus ou moins anxieux ou déprimés, ce qu'ils percevaient de leur qualité de vie et s'ils se sentaient stigmatisés. Aujourd'hui, ce type d'analyse est concurrencé par les classifications en classes latentes (ou en profils latents) qui reposent sur une structure théorique envisagée a priori.

Le mot du praticien :

La CAH est une des rares méthodes à n'avoir que très peu de pré-requis et de pouvoir être utilisée sur des données quantitatives et qualitatives. Ceci dit dans le cas où nous aurions des données qualitatives il sera nécessaire de réaliser auparavant une **analyse des correspondances multiples**.
Par ailleurs, deux paramètres vont être indispensables dans la réalisation de la CAH : le type de distance pour mesurer la ressemblance des individus et la méthode d'agrégation qui permet la construction du dendrogramme qui est la représentation classique de la CAH et illustre le regroupement successif des clusters.
Bien évidemment il existe un grand nombre de type de distances (Euclidienne, Manhattan, Φ^2, Gower,etc.) et de méthodes d'agrégation (Ward, saut minimum, moyenne, etc.).

Cependant, en SHS la méthode d'agrégation la plus souvent utilisée est celle de Ward qui cherche à minimiser l'inertie intra-classe et à maximiser l'inertie inter-classe afin d'obtenir des classes les plus homogènes possibles. Concernant les distances la méthode euclidienne est couramment utilisée pour les données quantitatives en SHS et la distance du Φ^2 pour les données qualitatives qui est d'ailleurs celle aussi utilisée dans l'**ACM**.

Ressources

- Bellanger Lise, Coulon Arthur et Husi Philippe, « Une méthode de classification ascendante hiérarchique par compromis : hclustcompro », 2021, Marseille, France, CIFSD, Mohamed QUAFAFOU.
- Benzécri Jean-Paul, « L' analyse des données. 1: La taxinomie », Paris, Dunod, 1973, .
- Chavent Marie, Kuentz-Simonet Vanessa, Labenne Amaury et Saracco Jerome, « ClustGeo : Classification Ascendante Hiérarchique (CAH) avec contraintes de proximitté géographique », 2015, Lille, France.
- Cornillon Pierre-André, Guyader Arnaud, Husson François, Jégou Nicolas, Josse Julie, Kloareg Maela, Matzner-Lober Eric et Rouvière Laurent, *Statistiques avec R*, 3e éd. revue et Augmentée., Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2012.
- Fleuret Sébastien et Apparicio Philippe, « Essai de typologie des centres de santé et de services sociaux au Québec », *Canadian Geographies / Géographies canadiennes*, nº 2, vol. 55, 2011, p. 143-157, [https//doi.org/10.1111/j.1541-0064.2010.00318.x].

[50] Prouteau Antoinette, Roux Solenne, Destaillats Jean-Marc et Bergua Valérie, « Profiles of Relationships Between Subjective and Objective Cognition in Schizophrenia: Associations With Quality of Life, Stigmatization, and Mood Factors », *Journal of Cognitive Education and Psychology*, nº 1, vol. 16, 2017, p. 64-76, [https//doi.org/10.1891/1945-8959.16.1.64].

- Husson François, Lê Sébastien et Pagès Jérôme, *Analyse de données avec R*, 2e éd. revue et Augmentée., Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2016.
- Lebart Ludovic, Piron Marie et Morineau Alain, *Statistique exploratoire multidimensionnelle: Visualisation et inférences en fouille de données*, 4e éd (2006)., Paris, Dunod, coll. « Sciences SUP », 1995.
- Pagès Jérôme, *Analyse factorielle multiple avec R*, Les Ulis, EDP sciences, coll. « Pratique R », 2013.
- Prouteau Antoinette, Roux Solenne, Destaillats Jean-Marc et Bergua Valérie, « Profiles of Relationships Between Subjective and Objective Cognition in Schizophrenia: Associations With Quality of Life, Stigmatization, and Mood Factors », *Journal of Cognitive Education and Psychology*, nº 1, vol. 16, 2017, p. 64-76, [https//doi.org/10.1891/1945-8959.16.1.64].
- Pumain Denise, « La composition socio-professionnelle des villes françaises : essai de typologie par analyse des correspondances et classification automatique », *L'Espace géographique*, nº 4, vol. 5, 1976, p. 227-238, [https//doi.org/10.3406/spgeo.1976.1663].
- Sanders Léna, *L'analyse statistique des données en géographie*, MONTPELLIER, DIFFUSION: LA DOCUMENTATION FRANCAISE,PARIS, ED.GIP RECLUS, coll. « ALIDADE », 1989.
- Ward Joe H., « Hierarchical Grouping to Optimize an Objective Function », *Journal of the American Statistical Association*, nº 301, vol. 58, 1963, p. 236-244, [https//doi.org/10.1080/01621459.1963.10500845].

CFA n.f [/se.ɛf.a/]

Synonymes : Analyse Factorielle Confirmatoire, AFC, Confirmatory Factorial Analysis

A quoi ça sert ?

L'analyse factorielle confirmatoire a pour objectif de tester l'adéquation d'un modèle théorique comprenant des *variables manifestes* (Observées) et des *variables latentes* (Non directement observables mais qui sont supposées influencer les variables observées) à un échantillon de données.

D'où ça vient ?

Les analyses factorielles confirmatoires font partie de la famille des **SEM Structural Equation Modeling - Modèles en équations structurelles**. La naissance des SEM est située au début du XX^{e} siècle avec les travaux de Sewall Wright en 1920 sur l'analyse de parcours, ayant pour objectif de rendre compte de relations causales à partir des variations de variables supposées a priori sous l'influence de variables latentes (Wright, S., 1923)[51]. Toutefois, il faudra attendre les années 1970 et les travaux de Jöreskog (Jöreskog, K., G., 1969)[52] pour voir la création des analyses factorielles confirmatoires. En effet, dans ses travaux Jöreskog intègre l'analyse factorielle à l'analyse de parcours et aux équations simultanées ce qui permet d'envisager des modèles de CFA (Confirmatory Factorial Analysis), tels que nous les présentons à présent.

Mot du praticien

Le principal pré-requis des analyses factorielles confirmatoires est d'avoir un modèle théorique à confronter à des données. Une fois ce modèle théorique défini, il sera alors nécessaire de disposer de données adaptées à ce type d'analyses.

Tout comme les **analyses factorielles exploratoires**, les analyses factorielles confirmatoires sont plus efficientes sur des données continues. Il est toutefois possible de les réaliser sur des variables ordinales ou catégorielles avec une méthode de factorisation adaptée.
La taille de l'échantillon est également un élément à prendre en compte. Dans la littérature il est fréquemment avancé le nombre de 10 observations par items, avec toutefois une diminution possible de ce ratio à partir de 300 observations (Kyriazos, T., A., 2018)[53].
La lecture de la matrice de corrélation est indispensable avant la réalisation de toute analyse factorielle (exploratoire et confirmatoire), afin de détecter les éventuelles corrélations trop fortes ou trop faibles, qui sont à éviter. Il est également nécessaire de s'assurer de l'absence de multicolinéarité entre les variables testées, dont la vérification est réalisable avec le test du **VIF** (Variance Inflation Factor).

---

[51] Wright Sewall, « THE THEORY OF PATH COEFFICIENTS A REPLY TO NILES'S CRITICISM », *Genetics*, n^{o} 3, vol. 8, 1923, p. 239-255, [https//doi.org/10.1093/genetics/8.3.239].

[52] Jöreskog K. G., « A General Approach to Confirmatory Maximum Likelihood Factor Analysis », *Psychometrika*, n^{o} 2, vol. 34, 1969, p. 183-202, [https//doi.org/10.1007/BF02289343].

[53] Kyriazos Theodoros A., « Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General », *Psychology*, n^{o} 08, vol. 09, 2018, p. 2207-2230, [https//doi.org/10.4236/psych.2018.98126].

La CFA est très utilisée en Psychologie et reste encore peu employée dans les autres disciplines des SHS. La CFA est mobilisée en Psychologie afin de valider des tests psychotechniques en cours de construction, suite à une **AFE** (Flora, D. B., & Flake, J. K., 2017[54]**)**, ou afin de vérifier que la structure d'un test validé précédemment s'ajuste aux données collectées dans un autre cadre. Les outils psychométriques déjà validés ne nécessitent pas que leur structure soit questionnée. Toutefois, il est nécessaire de s'assurer que leur structure ne varie pas sur un nouvel échantillon avec une analyse factorielle confirmatoire.

Un exemple d'application d'une analyse factorielle confirmatoire est présenté dans l'article de Déprez, G. R. M., et al., (2019)[55] où les auteurs présentent la construction d'un outil psychométrique permettant de mesurer les comportements personnels de déviance ou d'adhésion aux normes dans un contexte organisationnel. L'analyse factorielle confirmatoire a permis de démontrer la stabilité de la structure en 4 variables latentes, de la répartition des 12 variables, observée une première fois via une **AFE**. Les quatre variables latentes étant : la conformité normative, l'adéquation aux règles normatives, la recherche de performance déviante et la recherche de proactivité déviante.

Ressources

Pour débuter :

- Lin Johnny, *Introduction to Structural Equation Modeling (SEM) in R with lavaan*, [https://stats.oarc.ucla.edu/r/seminars/rsem/].
- Mercklé Pierre , « Les méthodes d'équations structurelles (MES) : Pour qui ? Pour quoi faire ? Comment ça marche ? par Alain Lacroux (Vendredis Quanti, 31 janvier 2020) », , 2020, [https//doi.org/10.58079/T4CF].
- Rosseel Yves, « Lavaan: An R Package for Structural Equation Modeling », *Journal of Statistical Software*, vol. 48, 2012, p. 1-36, [https//doi.org/10.18637/jss.v048.i02].
- Whalley Ben, *Just Enough R*. [https://benwhalley.github.io/just-enough-r/]

Pour aller plus loin :

- Pornprasertmanit Sunthud, Miller Patrick et Schoemann Alexander, *Simsem*, [https://simsem.org/].

Bibliographie :

- Déprez Guillaume Roland Michel, Battistelli Adalgisa et Antino Mirko, « Norm and Deviance-Seeking Personal Orientation Scale (NDPOS) Adapted to the Organisational Context », *Psychologica Belgica*, nº 1, vol. 59, 2019, [https//doi.org/10.5334/pb.462].
- Flora David B. et Flake Jessica K., « The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. », *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, nº 2, vol. 49, 2017, p. 78-88, [https//doi.org/10.1037/cbs0000069].

[54]Flora David B. et Flake Jessica K., « The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. », *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, nº 2, vol. 49, 2017, p. 78-88, [https//doi.org/10.1037/cbs0000069].

[55] Déprez Guillaume Roland Michel, Battistelli Adalgisa et Antino Mirko, « Norm and Deviance-Seeking Personal Orientation Scale (NDPOS) Adapted to the Organisational Context », *Psychologica Belgica*, nº 1, vol. 59, 2019, [https//doi.org/10.5334/pb.462].

- Jöreskog K. G., « A General Approach to Confirmatory Maximum Likelihood Factor Analysis », *Psychometrika*, nº 2, vol. 34, 1969, p. 183-202, [https//doi.org/10.1007/BF02289343].
- Kyriazos Theodoros A., « Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General », *Psychology*, nº 08, vol. 09, 2018, p. 2207-2230, [https//doi.org/10.4236/psych.2018.98126].
- Rosseel Yves, « Lavaan: An R Package for Structural Equation Modeling », *Journal of Statistical Software*, vol. 48, 2012, p. 1-36, [https//doi.org/10.18637/jss.v048.i02].
- Wright Sewall, « THE THEORY OF PATH COEFFICIENTS A REPLY TO NILES'S CRITICISM », *Genetics*, nº 3, vol. 8, 1923, p. 239-255, [https//doi.org/10.1093/genetics/8.3.239].

Cohérence interne n.f [/kɔ.e.ʁɑ̃s ɛ̃.tɛʁn]

Synonymes : Mesure de fiabilité, Internal consistency, Reliability

A quoi ça sert ?

L'analyse de fiabilité (de cohérence interne) est une méthode statistique permettant de vérifier que l'ensemble des items proposés dans un test/questionnaire (psychométrique en psychologie) mesure bien "la même chose", c'est-à-dire qu'ils renvoient à un même construit/diagnostic théorique. En psychologie par exemple, il s'agit d'identifier les éléments constitutifs d'une pathologie particulière comme le syndrome dépressif afin de procéder à son diagnostic. En science politique, l'objectif est de passer d'une notion théorique conceptualisée dans la discipline à une opérationnalisation empirique, c'est-à-dire qui peut s'observer dans une population donnée en l'interrogeant par questionnaire ; c'est par exemple le cas de ce qu'on appelle une attitude, "concept purement opératoire" pour Alain Lancelot, qu'il faut "construire à partir de régularités observées dans les comportements (...) l'induire de ces comportements" (Lancelot, A., 1985)[56]. Cette opérationnalisation revient en quelque sorte à élaborer, à partir d'un concept le plus souvent abstrait et non mesurable directement, un ensemble de questions plus concrètes mais qui en réfèrent, à poser aux personnes interrogées : le concept est décomposé en unités mesurables rudimentaires, mais dont nous faisons l'hypothèse que considérées ensemble, elles vont constituer un faisceau d'indices convergents approchant la notion scientifique.

D'où ça vient ?

Le terme de cohérence interne fait son apparition dans la littérature en psychologie dans la première moitié du XX^{e} siècle. Nous pouvons notamment citer les travaux de Mosier (1936)[57], Kuder & Rirchardson (1937)[58], puis Guttman (1945)[59] et Hoyt (1941)[60] dans les années 1940 et ensuite Cronbach dans les années 1950[61]. Ces différents auteurs cherchent une façon de s'assurer que les outils psychométriques construits, les questions passées aux particpants, renvoient bien à la même idée, à la même chose.

Différentes méthodes de calcul voient donc le jour. La méthode la plus élémentaire, celle du « split-half » (développée au début du XX^{e} siècle[62],[63]), reste incontournable dans la mesure où les

---

[56] Lancelot A., « L'Orientation du comportement politique », *Traité de science politique*, Paris, Presses universitaires de France, 1985.

[57] Mosier Charles I., « A Note on Item Analysis and the Criterion of Internal Consistency », *Psychometrika*, n^{o} 4, vol. 1, 1936, p. 275-282, [https//doi.org/10.1007/BF02287879].

[58] Kuder G. F. et Richardson M. W., « The Theory of the Estimation of Test Reliability », *Psychometrika*, n^{o} 3, vol. 2, 1937, p. 151-160, [https//doi.org/10.1007/BF02288391].

[59] Guttman Louis, « A Basis for Analyzing Test-Retest Reliability », *Psychometrika*, n^{o} 4, vol. 10, 1945, p. 255-282, [https//doi.org/10.1007/BF02288892].

[60] Hoyt Cyril, « Test Reliability Estimated by Analysis of Variance », *Psychometrika*, n^{o} 3, vol. 6, 1941, p. 153-160, [https//doi.org/10.1007/BF02289270].

[61] Cronbach Lee J., « Coefficient Alpha and the Internal Structure of Tests », *Psychometrika*, n^{o} 3, vol. 16, 1951, p. 297-334, [https//doi.org/10.1007/BF02310555].

[62] Spearman C., « CORRELATION CALCULATED FROM FAULTY DATA », *British Journal of Psychology, 1904-1920*, n^{o} 3, vol. 3, 1910, p. 271-295, [https//doi.org/10.1111/j.2044-8295.1910.tb00206.x].

[63] Brown William, « SOME EXPERIMENTAL RESULTS IN THE CORRELATION OF MENTAL ABILITIES[1] », *British Journal of Psychology, 1904-1920*, n^{o} 3, vol. 3, 1910, p. 296-322, [https//doi.org/10.1111/j.2044-8295.1910.tb00207.x].

différentes méthodes développées ultérieurement s'appuient sur ses fondements. Le split half consiste à calculer la **corrélation** entre 2 groupes d'items piochés aléatoirement (ou en sélectionnant les items pairs d'une part et les items impairs d'autre part). Si ce coefficient est élevé, alors l'expérience psychotechnique est considérée comme fiable : les 2 groupes d'items envisagés apparaissent, tous ensemble, disposer d'une bonne cohérence interne et renvoient donc au même construit théorique.

Depuis les années 1950 de nombreux autres tests ont été développés afin de mesurer la fiabilité d'un questionnaire en tenant compte de ses spécificités, comme par exemple le fait d'avoir une seule ou plusieurs dimensions (que les questions renvoient à une ou plusieurs idées, proches entre-elles et donc rassemblées sous une idée générale). Par exemple Revelle et al. en 2009 comparent 13 coefficients de fiabilité[64]. Deux coefficients sont très utilisés actuellement en SHS : l'Alpha de Cronbach (1951)[65] et l'Omega de McDonald (1978)[66]. Aujourd'hui, ceux-ci font l'objet d'importants débats sur la pertinence de leur utilisation afin d'évaluer la fiabilité des tests (questionnaires) selon leurs caractéristiques (nombres d'items, de dimensions, taille d'échantillon, etc.). En effet, la littérature est foisonnante sur ces débats, nous pouvons notamment mentionner : Cho, E., & Kim, S., 2015[67] ; Hayes, F., A., & Coutts, J., J., 2020[68] ou encore Edwards, A. A., Joyner, K. J., & Schatschneider, C., 2021[69].

Le mot du praticien

Les mesures de fiabilité s'appliquent aux variables quantitatives, continues ou ordinales. Afin d'utiliser ces tests, il faut s'assurer que les questions posées ont bien été comprises et que les réponses apportées comportent suffisamment de variabilité pour être mises en perspectives. Les questions ne permettant pas de discriminer les répondants (valeur unique) ou ne faisant ressortir que des valeurs extrêmes (effets plancher ou plafond) ne sont pas pertinentes à conserver dans une analyse de mesure de fiabilité.

Depuis leurs premiers développements dans les années 1950, beaucoup d'indices permettant de mesurer la fiabilité des questionnaires ont été développés (Edwards, A. A., et al., 2021)[70]. Dans cette

---

[64] Revelle William et Zinbarg Richard E., « Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma », *Psychometrika*, n° 1, vol. 74, 2009, p. 145-154, [https//doi.org/10.1007/s11336-008-9102-z].

[65] Cronbach Lee J., « Coefficient Alpha and the Internal Structure of Tests », *Psychometrika*, n° 3, vol. 16, 1951, p. 297-334, [https//doi.org/10.1007/BF02310555].

[66] McDonald Roderick P., « Generalizability in Factorable Domains: "Domain Validity and Generalizability"1 », *Educational and Psychological Measurement*, n° 1, vol. 38, 1978, p. 75-79, [https//doi.org/10.1177/001316447803800111].

[67] Cho Eunseong et Kim Seonghoon, « Cronbach's Coefficient Alpha: Well Known but Poorly Understood », *Organizational Research Methods*, n° 2, vol. 18, 2015, p. 207-230, [https//doi.org/10.1177/1094428114555994].

[68] Hayes Andrew F. et Coutts Jacob J., « Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But… », *Communication Methods and Measures*, n° 1, vol. 14, 2020, p. 1-24, [https//doi.org/10.1080/19312458.2020.1718629].

[69] Edwards Ashley A., Joyner Keanan J. et Schatschneider Christopher, « A Simulation Study on the Performance of Different Reliability Estimation Methods », *Educational and Psychological Measurement*, n° 6, vol. 81, 2021, p. 1089-1117, [https//doi.org/10.1177/0013164421994184].

[70] Edwards Ashley A., Joyner Keanan J. et Schatschneider Christopher, « A Simulation Study on the Performance of Different Reliability Estimation Methods », *Educational and Psychological Measurement*, n° 6, vol. 81, 2021, p. 1089-1117, [https//doi.org/10.1177/0013164421994184].

entrée nous ne présenterons que deux indices, qui s'ils font l'objet de controverses, restent très utilisés en SHS. Il s'agit de l'Alpha de Cronbach et de l'Omega de McDonald.

Alpha de Cronbach

L'Alpha de Cronbach est notamment utilisé en Sciences Politiques afin d'évaluer la fiabilité de questionnaires unidimensionnels (qui renvoient donc à une seule dimension). En science politique, et plus particulièrement quand il s'agit de mesurer empiriquement des valeurs ou des attitudes (sociologie de l'opinion), il est fréquent de construire des indicateurs de mesure permettant de compiler plusieurs questions d'un questionnaire sur un continuum.

Omega de McDonald

L'Omega de McDonald est très utilisé en Psychologie depuis le début des années 2000, notamment suite à la publication des travaux de Zinbarg[71] et de Revelle[51], en lieu et place de l'Alpha de Cronbach, pour la validation des tests psychométriques et vérifier leur fiabilité. Tout comme les **Analyses Factorielles Confirmatoires** l'Omega peut être employé lors de la construction d'un outil psychométrique ou lors de l'application d'un outil psychométrique, déjà validé dans la littérature, à notre échantillon. Par exemple, afin de valider un outil psychométrique permettant d'évaluer les émotions positives et négatives des participants, différents outils déjà validés (tels que pour mesurer les symptômes dépressifs, etc.) ont été employés pour vérifier la convergence et la divergence avec ce nouveau test (Echegaray et al. 2024)[72]. Mais dans cet article, les auteurs mobilisent également l'Omega de McDonald afin de vérifier la fiabilité de l'outil qu'ils viennent de construire.

En Psychologie les tests de cohérence interne sont très fréquemment utilisés afin de vérifier la fiabilité d'outils psychométriques existants sur un nouvel échantillon de population ou lors de la création d'un nouvel outil psychométrique. Différentes mesures de cohérence interne ont été utilisées lors de leurs développements au début du XX^{e} siècle, mais rapidement l'alpha de Cronbach (Cronbach, 1951)[73] s'est imposé comme mesure incontournable de la fiabilité des outils psychométriques. Toutefois, après plusieurs décennies d'utilisation, à partir des années 1990, celui-ci commence à être décrié dans la littérature pour son utilisation inadaptée et ses interprétations abusives, notamment car celui-ci a souvent été employé, à tort, sur des questionnaires renvoyant à différentes dimensions, alors même que l'Alpha n'est pas adapté à ce type de structure de données. En Psychologie, les critiques ne s'estompent pas et les recommandations portent sur l'utilisation d'autres tests de mesure de la fiabilité (Sijtma, 2009[74] ; Cho & Kim, 2014[75]). L'omega de McDonald est retenu comme étant une alternative intéressante à l'alpha de Cronbach et fait consensus dans

[71] Zinbarg Richard E., Revelle William, Yovel Iftah et Li Wen, « Cronbach's α , Revelle's β , and Mcdonald's ω_H : their Relations with Each Other and Two Alternative Conceptualizations of Reliability », *Psychometrika*, n° 1, vol. 70, 2005, p. 123-133, [https//doi.org/10.1007/s11336-003-0974-7].

[72] Echegaray Fanny, Roux Solenne, Koleck Michèle, Jouvie Jessika, Artheix-Althabegoity Yvane, Lebourleux Paul, Munuera Caroline et M'bailara Katia, « Development and Preliminary Validation of the Unpleasant and Pleasant Emotion Regulation Assessment (UPER-A) », *European Journal of Psychological Assessment*, , 2024, p. 1015-5759/a000851, [https//doi.org/10.1027/1015-5759/a000851].

[73] Cronbach Lee J., « Coefficient Alpha and the Internal Structure of Tests », *Psychometrika*, n° 3, vol. 16, 1951, p. 297-334, [https//doi.org/10.1007/BF02310555].

[74] Sijtsma Klaas, « On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha », *Psychometrika*, n° 1, vol. 74, 2009, p. 107-120, [https//doi.org/10.1007/s11336-008-9101-0].

[75] Cho Eunseong et Kim Seonghoon, « Cronbach's Coefficient Alpha: Well Known but Poorly Understood », *Organizational Research Methods*, n° 2, vol. 18, 2015, p. 207-230, [https//doi.org/10.1177/1094428114555994].

la littérature (Zinbarg et al., 2005[76] ; Padilla & Divers, 2015[77] ; Hayes & Coutts, 2020[78]). En Science Politique les critiques de l'alpha se développent également (DeSante, 2011)[79].

Ressources

- Brown William, « SOME EXPERIMENTAL RESULTS IN THE CORRELATION OF MENTAL ABILITIES[1] », *British Journal of Psychology, 1904-1920*, nº 3, vol. 3, 1910, p. 296-322, [https//doi.org/10.1111/j.2044-8295.1910.tb00207.x].
- Cho Eunseong et Kim Seonghoon, « Cronbach's Coefficient Alpha: Well Known but Poorly Understood », *Organizational Research Methods*, nº 2, vol. 18, 2015, p. 207-230, [https//doi.org/10.1177/1094428114555994].
- Cronbach Lee J., « Coefficient Alpha and the Internal Structure of Tests », *Psychometrika*, nº 3, vol. 16, 1951, p. 297-334, [https//doi.org/10.1007/BF02310555].
- DeSante Christopher D., « Revisiting Reliability: The Misuse of Cronbach's Alpha in Political Science ».
- Echegaray Fanny, Roux Solenne, Koleck Michèle, Jouvie Jessika, Artheix-Althabegoity Yvane, Lebourleux Paul, Munuera Caroline et M'bailara Katia, « Development and Preliminary Validation of the Unpleasant and Pleasant Emotion Regulation Assessment (UPER-A) », *European Journal of Psychological Assessment*, , 2024, p. 1015-5759/a000851, [https//doi.org/10.1027/1015-5759/a000851].
- Edwards Ashley A., Joyner Keanan J. et Schatschneider Christopher, « A Simulation Study on the Performance of Different Reliability Estimation Methods », *Educational and Psychological Measurement*, nº 6, vol. 81, 2021, p. 1089-1117, [https//doi.org/10.1177/0013164421994184].
- Flora David B., « Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates », *Advances in Methods and Practices in Psychological Science*, nº 4, vol. 3, 2020, p. 484-501, [https//doi.org/10.1177/2515245920951747].
- Guttman Louis, « A Basis for Analyzing Test-Retest Reliability », *Psychometrika*, nº 4, vol. 10, 1945, p. 255-282, [https//doi.org/10.1007/BF02288892].
- Hayes Andrew F. et Coutts Jacob J., « Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But… », *Communication Methods and Measures*, nº 1, vol. 14, 2020, p. 1-24, [https//doi.org/10.1080/19312458.2020.1718629].
- Hoyt Cyril, « Test Reliability Estimated by Analysis of Variance », *Psychometrika*, nº 3, vol. 6, 1941, p. 153-160, [https//doi.org/10.1007/BF02289270].
- Kuder G. F. et Richardson M. W., « The Theory of the Estimation of Test Reliability », *Psychometrika*, nº 3, vol. 2, 1937, p. 151-160, [https//doi.org/10.1007/BF02288391].
- Lancelot A., « L'Orientation du comportement politique », *Traité de science politique*, Paris, Presses universitaires de France, 1985.

[76] Zinbarg Richard E., Revelle William, Yovel Iftah et Li Wen, « Cronbach's α , Revelle's β , and Mcdonald's ω_H : their Relations with Each Other and Two Alternative Conceptualizations of Reliability », *Psychometrika*, nº 1, vol. 70, 2005, p. 123-133, [https//doi.org/10.1007/s11336-003-0974-7].

[77] Padilla Miguel A. et Divers Jasmin, « A Comparison of Composite Reliability Estimators: Coefficient Omega Confidence Intervals in the Current Literature », *Educational and Psychological Measurement*, nº 3, vol. 76, 2016, p. 436-453, [https//doi.org/10.1177/0013164415593776].

[78] Hayes Andrew F. et Coutts Jacob J., « Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But… », *Communication Methods and Measures*, nº 1, vol. 14, 2020, p. 1-24, [https//doi.org/10.1080/19312458.2020.1718629].

[79] DeSante Christopher D., « Revisiting Reliability: The Misuse of Cronbach's Alpha in Political Science ».

- McDonald Roderick P., « Generalizability in Factorable Domains: “Domain Validity and Generalizability”1 », *Educational and Psychological Measurement*, n° 1, vol. 38, 1978, p. 75-79, [https//doi.org/10.1177/001316447803800111].
- Mosier Charles I., « A Note on Item Analysis and the Criterion of Internal Consistency », *Psychometrika*, n° 4, vol. 1, 1936, p. 275-282, [https//doi.org/10.1007/BF02287879].
- Padilla Miguel A. et Divers Jasmin, « A Comparison of Composite Reliability Estimators: Coefficient Omega Confidence Intervals in the Current Literature », *Educational and Psychological Measurement*, n° 3, vol. 76, 2016, p. 436-453, [https//doi.org/10.1177/0013164415593776].
- Revelle William et Zinbarg Richard E., « Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma », *Psychometrika*, n° 1, vol. 74, 2009, p. 145-154, [https//doi.org/10.1007/s11336-008-9102-z].
- Sijtsma Klaas, « On the Use, the Misuse, and the Very Limited Usefulness of Cronbach’s Alpha », *Psychometrika*, n° 1, vol. 74, 2009, p. 107-120, [https//doi.org/10.1007/s11336-008-9101-0].
- Spearman C., « CORRELATION CALCULATED FROM FAULTY DATA », *British Journal of Psychology, 1904-1920*, n° 3, vol. 3, 1910, p. 271-295, [https//doi.org/10.1111/j.2044-8295.1910.tb00206.x].
- Watkins Marley W., « The reliability of multidimensional neuropsychological measures: from alpha to omega », *The Clinical Neuropsychologist*, n° 6-7, vol. 31, 2017, p. 1113-1126, [https//doi.org/10.1080/13854046.2017.1317364].
- Zinbarg Richard E., Revelle William, Yovel Iftah et Li Wen, « Cronbach’s α , Revelle’s β , and Mcdonald’s ω_H : their Relations with Each Other and Two Alternative Conceptualizations of Reliability », *Psychometrika*, n° 1, vol. 70, 2005, p. 123-133, [https//doi.org/10.1007/s11336-003-0974-7].

Ressources en ligne :

Packages R :

- Revelle William, *Psych: Procedures for Psychological, Psychometric, and Personality Research*, 2025. ; R package version 2.5.3, https://CRAN.R-project.org/package=psych. https://search.r-project.org/CRAN/refmans/psych/html/reliability.html (consulté le 07/04/2025)
- Steiner Markus et Grieder Silvia, *EFAtools: Fast and Flexible Implementations of Exploratory Factor Analysis Tools*, 2020. ; https://search.r-project.org/CRAN/refmans/EFAtools/html/OMEGA.html (consulté le 07/04/2025) doi : https://doi.org/10.32614/CRAN.package.EFAtools

Corrélation n.f [/ko.ʁe.la.sjɔ̃/]

Synonyme : Analyse des liens

A quoi ça sert ?

La corrélation est un indicateur permettant d'établir (ou non) un lien entre deux variables traditionnellement quantitatives et continues. Le test de la corrélation permet d'obtenir une mesure (l'indice de la corrélation alors compris entre -1 et 1) et repose sur la logique des *tests d'hypothèses*. Les tests généralement utilisés en SHS sur données continues sont les coefficients de corrélation de Pearson, de Spearman et de Kendall. Les variables peuvent également être ordinales, qualitatives ou dichotomiques, mais dans ce cas, le coefficient de corrélation mobilisé ne sera pas l'un des trois cités précédemment. Il existe des coefficients de corrélation permettant d'établir un lien entre une variable quantitative et une variable dichotomique, telle que la corrélation bisérielle de point (rpb)[80]. Dans le cas où les deux variables sont dichtomiques il sera approprié d'employer le coefficient **phi**. Il est à noter qu'il existe une relation linéaire entre le coefficient **phi** et le **Chi2**, que nous ne détaillerons pas ici[81]. Il existe également le coefficient de corrélation bisérielle, qui est équivalent du coefficient bisérielle de point, mais avec pour pré-supposé qu'une distribution normale sous-tend la variable dichotomique. Dans la même idée le coefficient de corrélation tetrachorique est l'équivalent du coefficient **phi**, mais lui aussi avec le pré-supposé qu'une distribution normale sous-tend les deux variables dichotomiques.

Dans le cadre de cette entrée sur la corrélation nous présenterons uniquement les coefficients de corrélation adaptés aux données quantitatives, à savoir les coefficients de Pearson, Spearman et Kendall, et ce qui les différencie.

D'où ça vient ?

Les mesures de corrélation entre variables sont relativement anciennes puisque les premiers développements sur le sujet remontent à la fin du XIX^{e} siècle, début du XX^{e}. L'un des pionniers en la matière est Karl Pearson, qui s'est inspiré des travaux de Galton, afin de développer le coefficient de corrélation de Pearson r, en 1896, (Pearson, K., 1896, Stanton, 2001)[82],[83]. Ce coefficient de corrélation et la lecture de sa probabilité associée (test d'hypothèse) impliquent que les variables mises en relation aient une distribution qui se rapprochent le plus d'une distribution normale. En effet, ce coefficient de corrélation est particulièrement vulnérable aux distributions asymétriques et aux valeurs extrêmes[84]. Le pré-requis de normalité n'étant pas toujours respecté, Spearman et Kendall vont développer des coefficients de corrélation sur données rangées. Dans ce cas la

[80] Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.

[81] Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.

[82] Pearson Karl, « VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia », *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, 1896, p. 253-318, [https//doi.org/10.1098/rsta.1896.0007].

[83] Stanton Jeffrey M., « Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors », *Journal of Statistics Education*, n° 3, vol. 9, 2001, p. 3, [https//doi.org/10.1080/10691898.2001.11910537].

[84] Kowalski Charles J., « On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient », *Applied Statistics*, n° 1, vol. 21, 1972, p. 1, [https//doi.org/10.2307/2346598].

corrélation ne tient pas compte des paramètres (tels que la **moyenne**) de la distribution, mais va se baser sur le nombre de paires présentes sur des données alors classées par rang. Le rho de Spearman a été présenté dès 1904 (Speaman, C., 1904)[85] et le tau de Kendall en 1938 afin de proposer une alternative au coefficient proposé par Spearman [86].

Le mot du praticien

La corrélation de Pearson est très sensible aux **valeurs extrêmes** et il est important que les distributions des observations suivent une *loi normale*. Les corrélations de Spearman et de Kendall étant basées sur des calculs de rangs et non d'**indices de tendance centrale**, aucune préconnisation particulière n'est nécessaire à leur application.

Lorsqu'il s'agit de rechercher à mettre en évidence des liens entre variables quantitatives, nous pouvons employer la corrélation de Pearson ou sa version généralisée : **le modèle de régression linéaire**. Les trois coefficients de corrélation présentés précédemment peuvent être utilisés avec pour finalité de quantifier et de tester des liens entre deux variables, mais souvent il s'agit d'une étape préliminaire à une analyse multivariée, que ce soit afin de factoriser des données (dans le cas d'**analyses factorielles**) ou de les modéliser (dans le cas de **régressions**). En effet, les analyses multivariées impliquent que les variables sélectionnées soient liées entre elles, mais pas trop non plus ce qui témoignerait d'une redondance d'information.

La corrélation peut donc être utilisée afin de vérifier la présence de liens entre deux variables. Par exemple, en Psychologie Sulik et al. (2021) utilisent le coefficient de corrélation de Pearson afin de vérifier la présence de liens observés précédemment entre la confiance en la Science et l'idéologie politique notamment[87]. La corrélation est également employée lors de la construction d'outils psychométriques afin de vérifier si l'outil à venir converge bien avec des éléments similaires (par exemple entre le conservatisme et une police autoritaire), mais également qu'il diverge d'éléments opposés (par exemple entre le conservatisme et le mariage entre deux personnes de même sexe)[88]. L'article de Pennycook et al, (2020) sur le sujet présente un exemple d'application de ces corrélations[89]. Autre exemple, afin d'appréhender la consommation de fake news et ses liens éventuels avec l'anxiété ou encore certains biais cognitifs, Escolà-Gascón, Á., et al. (2023)[90] ont réalisé des **régressions multiples** avec en amont des corrélations de Pearson et de Spearman.

---

[85] Spearman C., « The Proof and Measurement of Association between Two Things », *The American Journal of Psychology*, nº 72-101, vol. 15, 1904, [https//doi.org/10.2307/1422689].

[86] Kendall M. G., « A NEW MEASURE OF RANK CORRELATION », *Biometrika*, nº 1-2, vol. 30, 1938, p. 81-93, [https//doi.org/10.1093/biomet/30.1-2.81].

[87] Sulik Justin, Deroy Ophelia, Dezecache Guillaume, Newson Martha, Zhao Yi, El Zein Marwa et Tunçgenç Bahar, « Facing the pandemic with trust in science », *Humanities and Social Sciences Communications*, nº 1, vol. 8, 2021, p. 301, [https//doi.org/10.1057/s41599-021-00982-9].

[88] Pennycook Gordon, Cheyne James Allan, Koehler Derek J. et Fugelsang Jonathan A., « On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs », *Judgment and Decision Making*, nº 4, vol. 15, 2020, p. 476-498, [https//doi.org/10.1017/S1930297500007439].

[89] Pennycook et al, (2020) Idem

[90] Escolà-Gascón Álex, Dagnall Neil, Denovan Andrew, Drinkwater Kenneth et Diez-Bosch Miriam, « Who falls for fake news? Psychological and clinical profiling evidence of fake news consumers », *Personality and Individual Differences*, vol. 200, 2023, p. 111893, [https//doi.org/10.1016/j.paid.2022.111893].

Ressources

- Escolà-Gascón Álex, Dagnall Neil, Denovan Andrew, Drinkwater Kenneth et Diez-Bosch Miriam, « Who falls for fake news? Psychological and clinical profiling evidence of fake news consumers », *Personality and Individual Differences*, vol. 200, 2023, p. 111893, [https//doi.org/10.1016/j.paid.2022.111893].
- Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.
- Kendall M. G., « A NEW MEASURE OF RANK CORRELATION », *Biometrika*, nº 1-2, vol. 30, 1938, p. 81-93, [https//doi.org/10.1093/biomet/30.1-2.81].
- Kowalski Charles J., « On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient », *Applied Statistics*, nº 1, vol. 21, 1972, p. 1, [https//doi.org/10.2307/2346598].
- Le Campion Grégoire, « Analyse des corrélations avec easystats: Guide pratique avec R », , 2021, [https//doi.org/10.48645/QHAV-CB52].
- Pearson Karl, « VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia », *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, 1896, p. 253-318, [https//doi.org/10.1098/rsta.1896.0007].
- Pennycook Gordon, Cheyne James Allan, Koehler Derek J. et Fugelsang Jonathan A., « On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs », *Judgment and Decision Making*, nº 4, vol. 15, 2020, p. 476-498, [https//doi.org/10.1017/S1930297500007439].
- Spearman C., « The Proof and Measurement of Association between Two Things », *The American Journal of Psychology*, nº 72-101, vol. 15, 1904, [https//doi.org/10.2307/1422689].
- Stanton Jeffrey M., « Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors », *Journal of Statistics Education*, nº 3, vol. 9, 2001, p. 3, [https//doi.org/10.1080/10691898.2001.11910537].
- Sulik Justin, Deroy Ophelia, Dezecache Guillaume, Newson Martha, Zhao Yi, El Zein Marwa et Tunçgenç Bahar, « Facing the pandemic with trust in science », *Humanities and Social Sciences Communications*, nº 1, vol. 8, 2021, p. 301, [https//doi.org/10.1057/s41599-021-00982-9].

Cox n.m. [/kɔks/]

Synonymes : modèle des risques proportionnels de Cox, modèle des risques proportionnels, Régression de Cox

A quoi ça sert ?

Le modèle de Cox est une méthode statistique utilisée pour analyser le temps (ou la durée) qu'il faut avant qu'un événement se produise. Ce temps est appelé le « temps de survie ». Le modèle de Cox est un outil fondamental en analyse de survie qui a pour objectif d'aider à comprendre quels facteurs influencent ce "temps de survie" et dans quelle mesure. Ce modèle est particulièrement utile lorsque nous souhaitons :

- Ne pas seulement s'intéresser à la probabilité qu'un événement survienne, mais aussi au moment où il a lieu (ex. : durée du chômage, délai avant la naissance du premier enfant, temps avant l'entrée en union, délai d'accès à un diplôme, etc.) ;
- Modéliser l'influence de variables explicatives (âge, genre, origine sociale, niveau d'études…) sur le moment de survenue d'un événement ;
- Travailler avec l'ensemble des observations disponibles, y compris celles dites censurées, c'est-à-dire les individus qui sortent de l'observation avant d'avoir connu l'événement étudié, une situation fréquente dans les enquêtes en coupe transversale menées auprès de personnes d'âges variés.

Le modèle se base sur le concept de "risque instantané". À tout moment donné, il calcule la probabilité qu'un événement binaire (oui/non), par exemple la survenue d'un événement, se produise à un instant t, sachant que l'individu a "survécu" jusqu'à maintenant.

Le modèle de Cox a une particularité c'est ce qu'on appelle une approche "semi-paramétrique". Vous pouvez analyser l'effet des variables explicatives sans faire d'hypothèses sur la distribution du temps de survie. On parle aussi de modèle des risques proportionnels car les modèles de Cox obéissent à l'hypothèse des risques proportionnels**.** Le modèle suppose que le rapport des risques entre deux individus reste constant dans le temps. Si un patient a deux fois plus de risque qu'un autre au début, ce rapport de 2 reste le même tout au long du suivi.

En bref, le modèle de Cox répond à la question : "Comment différents facteurs, considérés simultanément, influencent-ils le risque d'occurrence d'un événement ?" Il quantifie l'effet de variables explicatives sur la survie et permet d'inclure plusieurs covariables à la fois en ajustant leurs effets respectifs, voire leurs effets mutuels lorsque des interactions sont spécifiées. Il peut aussi intégrer des co-variables dépendantes du temps, dont la valeur change au fil de l'observation. Il peut quantifier l'effet propre de chaque variable en contrôlant les autres, à la manière de ce que peut faire une régression multiple.

D'où ça vient ?

Le modèle de Cox a été développé par Sir David Cox, statisticien britannique, en 1972. Cox publie son modèle dans un article fondateur intitulé "Regression Models and Life-Tables" dasn le Journal of the Royal Statistical Society.

Le développement de ce modèle représente une vraie avancée et répond à une problématique importante. Les statisticiens cherchaient des moyens d'analyser les données de survie tout en tenant compte de multiples facteurs explicatifs. Les méthodes existantes étaient soit trop rigides (modèles paramétriques classiques), soit trop limitées (méthodes non-paramétriques simples).

Cox a eu l'idée de créer une approche "semi-paramétrique" : estimer l'effet des covariables sans spécifier complètement la distribution du temps de survie. Aujourd'hui, les modèles de durée sont utilisés dans de nombreuses disciplines (démographie, économie, épidémiologie, sociologie, science politique, psychologie, etc.). D'abord centrés sur l'analyse de la survie d'où le nom initial d'« analyse de survie », ils se sont progressivement étendus à tout type d'événements sociaux ou comportementaux.

Exemple d'application

Les exemples d'application des modèles de Cox sont particulièrement nombreux, notamment en médecine, où ils sont très utilisés, par exemple, dans le cadre d'études pour analyser l'efficacité d'un traitement. Ils sont également utilisés en sciences humaines et sociales pour identifier quels facteurs accélèrent ou ralentissent la survenue d'un événement social étudié, comme par exemple la durée d'un épisode de chômage.

Le modèle de Cox constitue un outil aujourd'hui utilisé dans de nombreux domaines des sciences humaines et sociales :

- En démographie, pour l'étude du délai d'apparition d'une naissance (Setu SP, Kabir R et *al*., 2024; Ponkilainen M., Einiö E. et *al*., 2024), ou encore pour l'étude de la mortalité et de ses facteurs au sein de différentes populations (Guillot M., Khlat M. et *al*., 2023).
- En économie, pour l'étude du temps nécessaire à la sortie d'une crise locale (de Cezaro Eberhardt et Fochezatto, 2024), pour l'analyse de la durée d'inactivité (Trang, Hal et *al*., 2024), ou encore pour l'examen de la durée de vie des entreprises (Afin, Tibor et *al*., 2025).
- En géographie, pour l'analyse de la durée de construction des logements (Zhang et Miller, 2022), ou encore pour l'étude de la mobilité ou de l'accessibilité à différents services (Montero, Mejia-Dorantes et *al*., 2023; Jiao et Azimian, 2021).
- En psychologie, pour l'étude de la sortie d'un traitement (Arntz, Mensik et *al*., 2023), ou pour l'analyse de l'occurrence d'un trouble mental ou psychologique (McGuire, Huffhines et *al.*, 2021).

Les modèles de durée sont également mobilisés dans bien d'autres disciplines des sciences humaines et sociales, comme l'histoire, les sciences politiques, la sociologie ou encore l'anthropologie, dès lors que l'on s'intéresse au « temps d'occurrence » d'un événement.

Mot du praticien

Comme tous modèles, il existe un certain nombre de pré-requis d'utilisation. Il y a ceux concernant les hypothèses fondamentales du modèle de Cox :

1- L'hypothèse des risques proportionnels doit être respecté : c'est l'hypothèse au coeur de ce modèle. Le rapport des risques entre deux individus doit rester constant dans le temps et ce rapport ne doit pas changer. Si ce rapport varie, et donc si l'hypothèse des risques proportionnels est violée, il est possible d'envisager d'ajouter des éléments permettant de corriger le modèle tels que : des

interactions avec le temps, des co-variables dépendantes du temps ou le recours à d'autres types de modèles.
2- La linéarité : La relation entre les covariables et le logarithme du risque doit être linéaire. Pour les relations non-linéaires, il faut transformer les variables ou utiliser des splines.
3- Indépendance des observations : Chaque individu doit être indépendant des autres. Pour les données groupées, des extensions spécifiques sont nécessaires.

Et ceux concernant plutôt des conditions pratiques :

1- Disposer des données adaptées à l'analyse de survie.
2- Les covariables peuvent être continues ou catégorielles, mais doivent être mesurées au début de l'étude (ou à l'entrée dans l'étude pour chaque individu).
3- La taille d'échantillon, il faut suffisamment d'événements observés. Par convention on attend au moins 10 événements par variable explicative incluse dans le modèle.
4- Variabilité des covariables : Les variables explicatives doivent présenter une variabilité suffisante. Une variable constante ou quasi-constante n'apportera aucune information.
5 - Absence de colinéarité sévère entre nos prédicteurs.

Comme dans un modèle de régression le modèle de Cox va fournir des coefficients qui vont indiquer l'effet des prédicteurs sur la survie, c'est-à-dire s'ils augmentent ou diminuent le risque de survenue de l'événement. L'interprétation directe des coefficients n'est en revanche pas intuitive car elle s'effectue sur une échelle logarithmique. On interprétera plutôt les rapports de risques (Hazard Ratios) obtenus en prenant l'exponentiel des coefficients, qui permettent de traduire ces effets en valeurs relatives (par exemple, les coefficients en pourcentages d'augmentation ou de diminution du risque).

Attention les rapports de risques n'impliquent pas forcément une relation de causalité. Comme dans une régression il y aura également des intervalles de confiances et des p-values qui nous indiquent la précision des estimations et la significativité des effets de la variable.

Dans le cadre d'une analyse de survie le modèle de Cox est complémentaire du modèle de **Kaplan-Meier**. Ce dernier pourra être utilisé dans une première approche exploratoire et descriptive tandis que le modèle de Cox sera plutôt employé dans une approche explicative et prédictive. En effet, Kaplan-Meier ne peut analyser qu'une variable à la fois alors que Cox permet d'inclure plusieurs covariables simultanément.

Dans les deux cas, les résultats seront représentés par des courbes de survie.

Enfin, le modèle de Cox s'applique généralement à un seul événement, mais de nombreux phénomènes sociaux impliquent des successions d'états (par exemple : études, emploi, chômage, emploi, etc.). Dans ce cas, des modèles multi-états peuvent être considérés, car ils permettent de modéliser des trajectoires complexes en considérant plusieurs transitions possibles entre états. Chaque transition est alors modélisée par une fonction de risque spécifique (souvent de type Cox) (Le-Rademacher et al., 2022).

Ressources

- Afin Rifai, Tibor Keresztély et Ilona Cserháti, « Firm performance and markets: survival analysis of medium and large manufacturing enterprises in Indonesia », *Journal of Industrial*

and Business Economics, nº 1, vol. 52, 2025, p. 107-151, [https//doi.org/10.1007/s40812-024-00302-7].

- Arntz Arnoud, Mensink Kyra, Cox Wouter R., Verhoef Rogier E. J., Van Emmerik Arnold A. P., Rameckers Sophie A., Badenbach Theresa et Grasman Raoul P. P. P., « Dropout from psychological treatment for borderline personality disorder: a multilevel survival meta-analysis », *Psychological Medicine*, nº 3, vol. 53, 2023, p. 668-686, [https//doi.org/10.1017/S0033291722003634].
- Cox D. R., « Regression Models and Life-Tables », *Journal of the Royal Statistical Society Series B: Statistical Methodology*, nº 2, vol. 34, 1972, p. 187-202, [https//doi.org/10.1111/j.2517-6161.1972.tb00899.x].
- De Cezaro Eberhardt Paulo Henrique et Fochezatto Adelar, « Regional Resilience and the Asymmetric Effects of the 2008 Crisis in Brazil: A Survival Model Analysis », *Networks and Spatial Economics*, nº 3, vol. 24, 2024, p. 743-762, [https//doi.org/10.1007/s11067-024-09640-4].
- Guillot Michel, Khlat Myriam, Gansey Romeo, Solignac Matthieu et Elo Irma, « Return Migration Selection and Its Impact on the Migrant Mortality Advantage: New Evidence Using French Pension Data », *Demography*, nº 5, vol. 60, 2023, p. 1335-1357, [https//doi.org/10.1215/00703370-10938784].
- Jiao Junfeng et Azimian Amin, « Measuring accessibility to grocery stores using radiation model and survival analysis », *Journal of Transport Geography*, vol. 94, 2021, p. 103107, [https//doi.org/10.1016/j.jtrangeo.2021.103107].
- McFadden D., *Conditional Logit Analysis of Qualitative Choice Behavior*, Institute of Urban and Regional Development, University of California, 1973.
- McGuire Austen, Huffhines Lindsay et Jackson Yo, « The trajectory of PTSD among youth in foster care: A survival analysis examining maltreatment experiences prior to entry into care », *Child Abuse & Neglect*, vol. 115, 2021, p. 105026, [https//doi.org/10.1016/j.chiabu.2021.105026].
- Montero Lídia, Mejía-Dorantes Lucía et Barceló Jaume, « The role of life course and gender in mobility patterns: a spatiotemporal sequence analysis in Barcelona », *European Transport Research Review*, nº 1, vol. 15, 2023, p. 44, [https//doi.org/10.1186/s12544-023-00621-1].
- Ponkilainen Maria, Einiö Elina, Pietiläinen Marjut et Myrskylä Mikko, « Educational Differences in Fertility Among Female Same-Sex Couples in Finland », *Demography*, nº 6, vol. 61, 2024, p. 2053-2079, [https//doi.org/10.1215/00703370-11687583].
- Setu Sarmistha Paul, Kabir Rasel, Islam Md. Akhtarul, Alauddin Sharlene et Nahar Mst. Tanmin, « Factors associated with time to first birth interval among ever married Bangladeshi women: A comparative analysis on Cox-PH model and parametric models », *PLOS Global Public Health*, nº 12, vol. 4, 2024, p. e0004062, [https//doi.org/10.1371/journal.pgph.0004062].
- Trang Le Mai, Ha Dinh Thi, Son Dao The, Lan Ninh Thi Hoang et Anh Tran Kim, « Survival analysis of unemployment duration: a case study of Vietnam », *Journal of Education and Work*, nº 1-4, vol. 37, 2024, p. 216-233, [https//doi.org/10.1080/13639080.2024.2383562].
- Zhang Yu et Miller Eric J., « Predicting housing construction period based on a cox proportional hazard model—an empirical study of housing completions in the greater Toronto and Hamilton area », *Environment and Planning B: Urban Analytics and City Science*, nº 6, vol. 50, 2023, p. 1624-1644, [https//doi.org/10.1177/23998083221143386].

Description n.f [/dɛs.kʁip.sjɔ̃/]

Synonymes : Analyses descriptives, Statistiques descriptives

A quoi ça sert ?

Pour avoir une bonne connaissance de nos données, avant de réaliser une analyse statistique univariée, bivariée ou multivariée, il est nécessaire de réaliser des analyses statistiques descriptives. L'idée est de découvrir leurs particularités, par exemple leur niveau de variabilité, leur échelle, les relations entre elles, etc. Cela permet aussi d'identifier des observations ou des variables aberrantes, atypiques ou extrêmes qui ne suivent pas les caractéristiques du reste des données.

Pour résumer les propriétés de nos variables, il existe différents types d'indicateurs :

- Les indices de tendance centrale
- Les indices de dispersion
- Les indices de forme (des distributions)
- Les fréquences

a. Les **indices de tendance centrale**

Ils ont pour fonction principale de caractériser la tendance centrale d'une variable et de déterminer sa valeur la plus probable. Les indices tels que la moyenne, la médiane ou encore le mode sont des indices de tendance centrale.

b. Les **indices de dispersion**

Ils indiquent le niveau de dispersion des valeurs d'une variable autour d'une valeur centrale (moyenne ou médiane par exemple). Les indices tels que la variance, l'écart-type, l'écart interquartile, l'écart moyen, l'étendue ou encore le coefficient de variation sont des indices de dispersion.

c. Les indices de forme

Ils donnent des indications sur la forme de la distribution d'une variable (symétrie, aplatissement, etc.). Le **kolmogorov-smirnov** ou encore le **shapiro-wilks** sont des indices de forme. Ils permettent de renseigner sur la forme de la distribution des données, si par exemple celles-ci suivent une loi normale ou non.

d. Les **fréquences**

Les fréquences se calculent pour des variables catégorielles et permettent de connaître la répartition des observations dans les différentes catégories d'une variable ou d'un croisement de deux variables.

Par ailleurs, il est primordial, avant de réaliser une analyse de données de vérifier la présence de **valeurs aberrantes** dans l'échantillon. Les **valeurs aberrantes** étant des valeurs qui ne sont pas dans le domaine du possible et qui ne peuvent pas exister.

Enfin il est toujours intéressant de compléter l'analyse descriptive par des visualisations pertinentes en fonction du type de variables. Un histogramme par exemple peut permettre de détecter une bi-modalité.

Pour apprendre à construire des graphiques efficaces, faciles à lire et qui racontent une histoire, Leland Wilkinson a introduit en 1999 le concept de grammaire des graphiques qui a ensuite été popularisée par la library R ggplot2.

Ressources

Vous êtes plutôt visuel ?

- Healy Conor et Holtz Yan, *From data to Viz | Find the graphic you need*, [https://www.data-to-viz.com/data-to-viz.com].
- Holtz Yan, *Python Graph Gallery*, [https://python-graph-gallery.com/].
- Holtz Yan, *The R Graph Gallery – Help and inspiration for R charts*, [https://r-graph-gallery.com/].
- Kabacoff Robert, *Modern Data Visualization with R*.
- Wilkinson Leland, *The Grammar of Graphics*, New York, Springer-Verlag, coll. « Statistics and Computing », 2005, [https//doi.org/10.1007/0-387-28695-0].

Vous êtes plutôt scolaire ?

- Gouvernement du Canada Statistique Canada, *Les statistiques : le pouvoir des données!*, [https://www150.statcan.gc.ca/n1/edu/power-pouvoir/toc-tdm/5214718-fra.htm].

Outil de formation destiné principalement aux étudiants, mais également aux enseignants et à tous ceux qui désirent tirer pleinement parti des statistiques.

Vous êtes plutôt manuel ?

- Bertrand Frédéric et Maumy-Bertrand Myriam, *Initiation à la statistique avec R*, 4e éd., Malakoff, Dunod, coll. « Sciences sup », 2023.
- Jubénot Marie-Noëlle et Eudes Daniel, *Analyse des données sous R pour les sciences humaines: théories et exemples commentés*, Paris, Ellipses, 2022.

Vous êtes plutôt BD ?

- Klein Grady et Dabney Alan, *The cartoon introduction to statistics*, First edition., New York, Hill and Wang, a Division of Farrar, Straus and Giroux, 2013.

Forêt aléatoire n.f. [/fɔ.ʁɛ̃ a.le.a.twaʁ /]

Synonyme : random forest

À quoi ça sert ?

Les forêts aléatoires constituent une méthode dérivée des arbres de décision. Comme leur nom l'indique, les forêts aléatoires sont simplement l'agrégation de plusieurs arbres de décision.
Comme précédemment mentionné dans l'entrée correspondante, les arbres de décision sont un outil d'exploration de données très efficaces, mais ne sont pas toujours optimaux en tant que modèles prédictifs. Les forêts aléatoires ont été conçues pour améliorer cet aspect prédictif, mais en sacrifiant presque totalement l'aspect exploratoire.

Une forêt aléatoire est construite à partir de nombreux arbres de décision (typiquement quelques centaines à quelques milliers), et fournit un résultat moyenné à partir de tous ces arbres : c'est une forme de model averaging. Plus précisément :

- pour chaque arbre de la forêt, on n'utilise qu'une fraction d'individus (par exemple 1/5 ou 1/4, si le nombre total d'individus est élevé) piochée aléatoirement dans les données ;
- pour chaque nœud des arbres, on peut également n'utiliser qu'une fraction (par exemple 1/3) des variables possibles, choisies aléatoirement également ;
- la prédiction finale à l'issue de la procédure est obtenue (dans le cas d'un arbre de régression) en moyennant les prédictions de tous les arbres de la forêt, ou bien (dans le cas d'un arbre de classification) en retenant la prédiction la plus fréquente.

Techniquement, on parle de bagging, c'est-à-dire de bootstrap aggregating – l'agrégation de modèles combinée à des procédures de rééchantillonnage.

D'où ça vient ?

L'algorithme des forêts aléatoires a été conçu par les mêmes auteurs que l'algorithme CART pour les arbres de décision (Breiman, 2001)[91]. Diverses extensions et améliorations ont néanmoins été apportées à la méthode au fil des ans. Par exemple :

- une adaptation spécifiquement conçue pour les variables réponses ordinales (Hornung, 2020) ;
- le calcul d'intervalles de prédiction pour des variables réponses numériques (Roy & Larocque, 2020) ;
- des adaptations dans le cas d'un très fort déséquilibre de classes dans les données d'apprentissage, pour des variables réponses qualitatives (Chen et al., 2004).

Le mot du praticien :

Pour prédire une variable réponse qualitative ou quantitative, plusieurs arguments peuvent plaider en faveur de l'utilisation d'un modèle de forêt aléatoire plutôt que des modèles plus classiques (analyse factorielle discriminante, **régression logistique**, **régression linéaire**, etc.).

[91] Breiman Leo, « Random Forests », *Machine Learning*, n° 1, vol. 45, 2001, p. 5-32, [https//doi.org/10.1023/A:1010933404324].

- Il est possible de mélanger tous les types de variables prédictives (quantitatives, qualitatives, ordinales).
- Les forêts aléatoires, par construction, sont bien adaptées pour saisir des interactions complexes entre les variables prédictives, ou des liaisons non-linéaires avec la variable réponse.
- Contrairement aux méthodes statistiques classiques, les forêts aléatoires ne nécessitent aucune procédure de sélection de variables, car elle est au coeur de l'algorithme (seules les variables réellement utiles interviennent, les autres seront pratiquement ignorées). De plus, un score d'importance relative peut être obtenu pour classer les variables de la plus utile à la moins utile.
- Par conséquent, il est notamment possible d'avoir beaucoup plus de variables que d'individus dans l'échantillon d'apprentissage.

Ressources

- Breiman Leo, « Random Forests », *Machine Learning*, n° 1, vol. 45, 2001, p. 5-32, [https//doi.org/10.1023/A:1010933404324].
- Chen Chao, Liaw Andy et Breiman Leo, *Using Random Forest to Learn Imbalanced Data*, [https://digicoll.lib.berkeley.edu/record/85556].
- Hornung Roman, « Ordinal Forests », *Journal of Classification*, n° 1, vol. 37, 2020, p. 4-17, [https//doi.org/10.1007/s00357-018-9302-x].
- Roy Marie-Hélène et Larocque Denis, « Prediction intervals with random forests », *Statistical Methods in Medical Research*, n° 1, vol. 29, 2020, p. 205-229, [https//doi.org/10.1177/0962280219829885].

Fréquences n.m [/fʁe.kɑ̃ /]

Synonymes : Frequencies tables, Tris à plat, Distribution de fréquences

A quoi ça sert ?

Les tableaux de fréquences sont des outils pour organiser et résumer des ensembles de données. Ils indiquent la fréquence d'apparition de chaque valeur prise par une variable. Dans un échantillon, un tableau de fréquence pour la variable sexe indique, par exemple, la fréquence d'apparition des modalités homme et femme mais aussi de ceux qui n'ont pas répondu. Les tableaux de fréquences sont souvent utilisés pour obtenir un premier aperçu des données, mais aussi pour présenter des résultats.

D'où ça vient ?

Parmi les premiers exemples marquant de l'utilisation des tableaux de fréquence dans un contexte statistique, on notera le travail de John Graunt au 17ème siècle (un des premiers démographes avec William Petty)[92] qui a collecté et analysé les registres des décès à Londres.

Le mot du praticien

Les tableaux de fréquences peuvent se composer des colonnes suivantes :

- les effectifs ou les effectifs pondérés (auxquels on attribue un poids selon des critères définis) qui indiquent le nombre d'observations de la valeur.
- les pourcentages (avec les non-réponses) (= fréquence relative) qui indiquent la fréquence d'apparition de la valeur par rapport à l'ensemble des cas.
- les pourcentages valides (= pourcentage exprimé) qui indiquent la fréquence d'apparition de la valeur par rapport à l'ensemble des cas hors valeurs manquantes (colonne conseillée en cas de valeurs manquantes).
- les pourcentages cumulés. C'est la somme des pourcentages des valeurs jusqu'à une certaine valeur donnée (variables numériques uniquement).
- les pourcentages cumulés valides. C'est la somme des pourcentages valides des valeurs jusqu'à une certaine valeur donnée (variables numériques uniquement).

Pour améliorer la lisibilité et la compréhension des chiffres dans les tableaux, il est préférable de les classer en fonction des effectifs (par ex. dans le classement des médailles des Jeux de Paris 2024, les pays seront classés par ordre décroissant du nombre de titres) ou selon les valeurs prises par la variable (par ex. si les modalités sont ordonnées, il est judicieux de garder l'ordre logique comme pour la couleur des médailles ou les classes d'âges).

Les représentations classiques d'un tableau de fréquences sont le graphique en barres (Barplot ou Lollipop – au-delà de 3 modalités notamment) et le diagramme circulaire (Pie chart (autrement

[92] Graunt John, *Natural and political observations mentioned in a following index, and made upon the bills of mortality by John Graunt ... ; with reference to the government, religion, trade, growth, ayre, diseases, and the several changes of the said city.*, 1662.

appelés Camembert) ou Doughnut). Les histogrammes cumulatifs et les courbes de fréquence cumulée peuvent être tout aussi pertinents[93].

Les tableaux croisés (tableaux de contingence) sont une extension des tableaux de fréquences. Dans les tableaux croisés, on ne considère pas seulement une mais deux variables.

Les tableaux de fréquences se calculent pour des variables nominales (ex : le sexe), ordonnées (ex. les classes d'âges) ou discrètes (ex. le nombre de pièces du logement). Le calcul est possible pour une variable continue (ex. le poids) si elle est discrétisée.

Avant de vous lancer dans les calculs, vérifier la pondération des données et si celle-ci est nécessaire (ou pas).

Les tableaux de fréquence sont souvent utilisés notamment lorsque nous souhaitons réaliser une première description des données que nous souhaitons analyser afin de répondre à une hypothèse de recherche. Ceci permet de mettre en évidence des biais d'échantillonnage, si des groupes sont fortement déséquilibrés (alors qu'ils n'ont aucune raison de l'être, car ne le sont pas dans la population générale) ou au contraire si l'échantillon collecté reprend la structure de la population générale. Dans une approche expérimentale cela permet de s'assurer que les groupes contiennent bien des tailles d'échantillons comparables. Pour ce faire il est pertinent de réaliser des tableaux de fréquence ou de contingence et d'appliquer un test du **Khi2**.

Il semble difficile de se passer de la réalisation de tableaux de fréquences, tant ils permettent facilement d'avoir une vue d'ensemble de la répartition des observations sur les variables considérées. Celles-ci pouvant ensuite faire l'objet de différents traitements, comme des regroupements de modalités dans le cas de catégories proches avec des effectifs faibles pour la réalisation d'analyses ultérieures. Les variables alors décrites peuvent aussi être intégrées à des modèles de **régression** comme des variables contrôles, afin de contrôler les déséquilibres observés. Ou elles peuvent être exclues des analyses ultérieures, si la répartition attendue n'a pas été collectée, par exemple : l'échantillon collecté ne contient que des femmes, or celui-ci devait être équilibré en terme homme/femme. Le sexe ne pourra plus être un critère d'analyse et les analyses finales ne porteront que sur les femmes.

Un exemple de présentation de tableau de fréquences est disponible dans le dictionnaire des variables du Baromètre du Numérique 2023 (Crédoc). Il s'agit d'une étude de référence sur l'adoption par les Français des équipements et des usages numériques[94]. Les grandes enquêtes de l'INSEE ou de l'INED présentent généralement leurs dictionnaires de variables avec les tableaux de fréquences correspondant aux différentes variables présentes dans le jeu de données, telle que l'enquête Teo (Trajectoires et Origines) par exemple et son Dictionnaire des codes[95].

[93] Healy Conor et Holtz Yan, *From data to Viz | Find the graphic you need*, [https://www.data-to-viz.com/data-to-viz.com] ; site visité le 10 juin 2025.

[94] ARCEP, *Baromètre du Numérique*, [https://www.data.gouv.fr/fr/datasets/barometre-du-numerique/informations/].

[95] INED, *Enquête Trajectoires et Origines 1 - Questionnaire principal TeO*, [https://teo1.site.ined.fr/fr/le_contenu_de_l_enquete/les_grands_themes_traites_dans_le_quetsionnaire/]. (site consulté le 09/07/2025)

Le tableau de fréquences pondérées est une généralisation du tableau de fréquences lorsque tous les individus interrogés n'ont pas la même importance, le même poids. Dans l'enquête du baromètre du numérique, c'est la variable POND qui permet d'extrapoler les résultats à la population française.

Ressources

- ARCEP, *Baromètre du Numérique*, [https://www.data.gouv.fr/fr/datasets/barometre-du-numerique/informations/].
- De Belsunce Clément, *Tutoriel #6 Statistiques descriptives et tris de données avec R Studio*, PROGEDO.
- Gouvernement du Canada Statistique Canada, *Les statistiques : le pouvoir des données!*, [https://www150.statcan.gc.ca/n1/edu/power-pouvoir/toc-tdm/5214718-fra.htm].
- Graunt John, *Natural and political observations mentioned in a following index, and made upon the bills of mortality by John Graunt ... ; with reference to the government, religion, trade, growth, ayre, diseases, and the several changes of the said city.*, 1662.
- Healy Conor et Holtz Yan, *From data to Viz | Find the graphic you need*, [https://www.data-to-viz.com/data-to-viz.com].
- Iannone Richard, Cheng Joe, Schloerke Barret, Hughes Ellis, Lauer Alexandra, Seo JooYoung, Brevoort Ken et Roy Olivier, *Introduction to Creating gt Tables*, [https://gt.rstudio.com/articles/gt.html].
- INED, *Enquête Trajectoires et Origines 1 - Questionnaire principal TeO*, [https://teo1.site.ined.fr/fr/le_contenu_de_l_enquete/les_grands_themes_traites_dans_le_quetsionnaire/]. (site consulté le 09/07/2025)
- Klein Grady et Dabney Alan, *The cartoon introduction to statistics*, First edition., New York, Hill and Wang, a Division of Farrar, Straus and Giroux, 2013.
- Playfair (1759–1823) William, *The Commercial and Political Atlas, Representing, By Means of Stained Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, During the Whole of the Eighteenth Century.*, [https://fisherdigitus.library.utoronto.ca/document/7750].

Growth model n.m. [/grəʊθ ˈmɒdəl/]

Synonymes : Modèles en développement latents, modèles de croissance

A quoi ça sert ?

Les growth model ou modèle de croissance, ou encore modèle en développement latent, font partie de la famille des **modèles en équation structurelles**. Il s'agit de modèles longitudinaux mais qui ne s'apparentent pas à des modèles de **panel**. Dans le cadre de cette entrée nous ne présenterons que les modèles en développement latent, les growth models peuvent en effet renvoyer à différentes analyses incluant des modèles de croissance sans pour autant intégrer cette notion de variable latente. Recueillir et analyser des données longitudinales (donc sur plusieurs temps de mesures) est plus coûteux que des recherches transversales. En effet, il faut pouvoir suivre les mêmes personnes sur un certain temps, en s'assurant qu'elles répondent toujours à l'ensemble des questions posées. Puis, les analyses de données longitudinales sont toujours relativement complexes à réaliser à cause de la structure même des données induisant, par exemple, de **l'autocorrélation**. Il est donc important d'avoir de bonnes raisons de mener ce type de recherche et ce type d'analyse. Baltes et Nesselroade préconisent cinq bonnes raisons de conduire des recherches longitudinales (Baltes & Nesselroade, 1979)[96] :

- Les différences intra-individuelles dans le temps
- Les différences inter-individuelles lors de changements intra-individuels dans le temps
- L'analyse des interrelations lors des changements de comportement.
- L'analyse des déterminants des changements intra-individuels.
- L'analyse des déterminants des différences inter-individuelles lors de changements intra-individuels.

Les growth models vont pouvoir être mobilisés afin de répondre à ces différentes questions. Ceci va nécessiter la création de modèles plus ou moins complexes impliquant des effectifs plus ou moins importants et des conditions spécifiques à chacun d'entre-eux. La particularité des growth models par rapport aux analyses longitudinales classiques va être de pouvoir tester des patterns de relation entre variables (comme le permettent les **modèles SEM**) sur plusieurs temps de mesure. Ils intègrent donc la possibilité de prendre en compte des variables latentes dans les modèles testés.

D'où ça vient ?

Le développement des growth models remonte aux années 1930 avec notamment les travaux de Wishart[97]. Celui-ci menait une étude avec des collègues afin de comparer la prise de poids de cochons suivant trois types de régimes différents, sur plusieurs semaines (16 très exactement). Les analyses statistiques, menées initialement, portaient sur la différence entre le poids enregistré à la 16ème semaine par rapport au poids initial, selon le type de régime suivi. Les résultats n'étant pas concluant, Wishart a cherché à développer des analyses statistiques prenant en compte l'ensemble des mesures (relevées chacune des 16 semaines). Il commençait donc à développer ce qui allait

[96] Nesselroade John R. et Baltes Paul B. (dir.), *Longitudinal research in the study of behavior and development*, New York, NY, Academic Press, 1979.

[97] Grimm Kevin J., Ram Nilam et Estabrook Ryne, *Growth modeling: structural equation and multilevel modeling approaches*, New York, NY, Guilford Press, coll. « Methodology in the social sciences », 2017.

devenir les modèles de croissance (Wishart, 1938)[98]. En 1958 Tucker et Rao ont généralisé l'approche de Wishart en intégrant les **analyses en composantes principaples (ACP)** dans leurs analyses afin de réduire la dimensionalité des données à mesures répétées, provenant de plusieurs individus, en quelques courbes permettant d'étudier les différences inter-individuelles selon le poids de ces courbes (composantes)[99],[100]. Grâce à cette généralisation, Tucker et Rao sont considérés comme les pères de l'analyse de croissance via les modèles en équations structurelles.[101] Ces modèles n'ont depuis cessé de se développer afin d'intégrer des modélisations de plus en plus complexes et de pouvoir les réaliser assez simplement à partir de différents logiciels d'analyse de données.

Le mot du praticien

La réalisation des growth models repose sur les mêmes conditions que celles des **modèles en équation structurelles**. La taille de l'échantillon soit être suffisante pour bénéficier d'une puissance statistique satisfaisante.

Les variables peuvent être continues, ordinales ou catégorielles. Ces modèles intègrent la possibilité de relations linéaires ou non-linaires entre variables mesurées ou latentes. Ils peuvent également porter sur différents groupes qui seront alors comparés. Et bien évidemment ils doivent reposer sur différents temps de mesures afin d'intégrer la dimension longitudinale intrinsèque à ces modèles.
En psychologie, les modèles en développement latents sont utilisés afin d'étudier la persistance dans le temps de relation entre différents éléments. Ce peut être par exemple l'étude de la régulation des émotions des adolescents selon différents facteurs externes (socio-économiques, etc.) et leurs évolutions dans le temps. Ou encore l'évolution dans le temps de la relation entre les activités (extra-scolaires) et la consommation de substances (tabac, cannabis, alcool, etc.) chez les adolescents[102].

Ressources

- Bone Jessica K., Fancourt Daisy, Sonke Jill K. et Bu Feifei, « The Changing Relationship Between Hobby Engagement and Substance Use in Young People: Latent Growth Modelling of the Add Health Cohort », *Journal of Youth and Adolescence*, n° 1, vol. 54, 2025, p. 133-145, [https//doi.org/10.1007/s10964-024-02047-x].
- Grimm Kevin J., Ram Nilam et Estabrook Ryne, *Growth modeling: structural equation and multilevel modeling approaches*, New York, NY, Guilford Press, coll. « Methodology in the social sciences », 2017.
- Nesselroade John R. et Baltes Paul B. (dir.), *Longitudinal research in the study of behavior and development*, New York, NY, Academic Press, 1979.

[98] Wishart J., « GROWTH-RATE DETERMINATIONS IN NUTRITION STUDIES WITH THE BACON PIG, AND THEIR ANALYSIS », *Biometrika*, n° 1-2, vol. 30, 1938, p. 16-28, [https//doi.org/10.1093/biomet/30.1-2.16].

[99] Tucker Ledyard R., « Determination of Parameters of a Functional Relation by Factor Analysis », *Psychometrika*, n° 1, vol. 23, 1958, p. 19-23, [https//doi.org/10.1007/BF02288975].

[100] Rao C. Radhakrishna, « Some Statistical Methods for Comparison of Growth Curves », *Biometrics*, n° 1, vol. 14, 1958, p. 1, [https//doi.org/10.2307/2527726].

[101] Grimm Kevin J., Ram Nilam et Estabrook Ryne, *Growth modeling: structural equation and multilevel modeling approaches*, New York, NY, Guilford Press, coll. « Methodology in the social sciences », 2017.

[102] Bone Jessica K., Fancourt Daisy, Sonke Jill K. et Bu Feifei, « The Changing Relationship Between Hobby Engagement and Substance Use in Young People: Latent Growth Modelling of the Add Health Cohort », *Journal of Youth and Adolescence*, n° 1, vol. 54, 2025, p. 133-145, [https//doi.org/10.1007/s10964-024-02047-x].

- Ram Nilam et Grimm Kevin, « Using simple and complex growth models to articulate developmental change: Matching theory to method », *International Journal of Behavioral Development*, nº 4, vol. 31, 2007, p. 303-316, [https//doi.org/10.1177/0165025407077751].
- Rao C. Radhakrishna, « Some Statistical Methods for Comparison of Growth Curves », *Biometrics*, nº 1, vol. 14, 1958, p. 1, [https//doi.org/10.2307/2527726].
- Rosseel Yves, « Lavaan: An R Package for Structural Equation Modeling », *Journal of Statistical Software*, vol. 48, 2012, p. 1-36, [https//doi.org/10.18637/jss.v048.i02].
- Tucker Ledyard R., « Determination of Parameters of a Functional Relation by Factor Analysis », *Psychometrika*, nº 1, vol. 23, 1958, p. 19-23, [https//doi.org/10.1007/BF02288975].
- Wishart J., « GROWTH-RATE DETERMINATIONS IN NUTRITION STUDIES WITH THE BACON PIG, AND THEIR ANALYSIS », *Biometrika*, nº 1-2, vol. 30, 1938, p. 16-28, [https//doi.org/10.1093/biomet/30.1-2.16].

GWR n.f [/ʒe dubləve ɛʁ/]

Synonymes : Régression géographiquement pondérée, Geographically weighted regression

A quoi ça sert ?

Les données spatiales ont des particularités qui nécessitent, lorsque que nous souhaitons utiliser des modèles de régressions, de les traiter de manière particulière. Une de ces spécificités est l'hétérogénéité spatiale. Il s'agit du fait que les observations vont varier en fonction des lieux, c'est-à-dire de la variabilité des caractéristiques structurelles de l'espace étudié. La conséquence de cette particularité, qui est que l'hypothèse d'un coefficient de régression serait uniforme sur toute une zone d'étude, est souvent irréaliste (Brundson et al., 1996).

Ce concept d'hétérogénéité va se traduire statiquement par le concept de non stationnarité. L'effet de l'hétérogénéité est donc une variation en fonction des lieux (de la répartition dans l'espace) des paramètres statistiques. En d'autres termes, les moyennes, variances, etc. vont varier plus ou moins en fonction des lieux, et donc par voie de conséquence, il en ira de même pour les covariances, les corrélations, etc. La non-stationnarité c'est donc cette instabilité structurelle des paramètres statistiques dans l'espace. Par exemple en géographie de la santé, une étude des parcours de soins ne saurait être envisagée sans tenir compte des disparités territoriales en termes d'offre de soins et d'établissements de santé. Concrètement, dans le cas de la régression, cela signifie que le signe, l'intensité ou même la significativité des coefficients vont varier en fonction du lieu.

Au-delà du fait qu'avec les données spatiales les conditions d'application de la régression ne sont généralement pas respectées, notamment le principe d'indépendance des observations et l'absence d'**autocorrélation** des résidus, le risque d'une approche « globale » serait de lisser les différences locales et de simplifier trop fortement la complexité des phénomènes étudiés. Le risque majeur étant qu'une approche classique/globale amène à un paradoxe de Simpson (Simpson, 1951) qui stipule qu'une relation statistique mesurée dans deux groupes peut s'inverser lorsque les deux groupes sont réunis, les groupes étant ici nos zones spatiales.

La régression géographiquement pondérée (GWR) va donc rendre possible le fait de tenir compte de cette hétérogénéité. Elle permet d'explorer la non-stationnarité d'un phénomène en faisant varier les coefficients dans l'espace. Pour ce faire, cette méthode va estimer un modèle de régression par observation en tenant compte de son voisinage spatial.

D'où ça vient ?

Les premiers à proposer et à formaliser cette méthode de la régression géographiquement pondérée sont Brundson et al. (1996) et Fotheringham et al. (2002). Il s'agit donc d'une méthode relativement récente. Il faut toutefois souligner qu'il ne s'agit que d'une méthode parmi d'autres pour étudier la question de l'hétérogénéité spatiale qui préoccupe les sciences humaines et sociale depuis bien plus longtemps.

Exemples d'application

Les exemples d'application en sciences humaines et sociales sont nombreux. Il y a par exemple des publications récentes en géographie de la santé (Feuillet et al., 2018 ; Pilkington et al., 2021), en géographie de l'environnement (Feuillet et al. 2014) ou encore en géographie économique (Bulteau

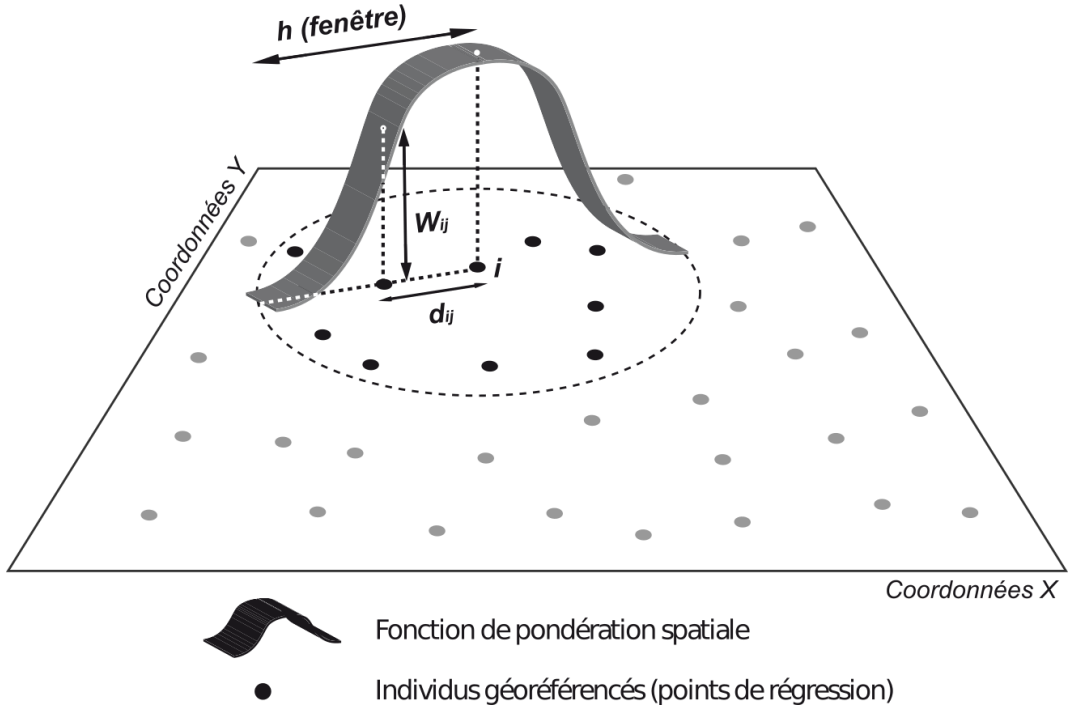
et al., 2018). Nous pouvons également citer les publications de Feuillet (2021), Comber et al. (2023), de Hilal et Le Gallo (2023), et Audard et al. (2024) qui reprennent des cas d'application de GWR et indiquent toute la marche à suivre. L'article de Comber et al. (2023) propose même une feuille de route pour réaliser une GWR avec une ouverture vers des modèles encore plus avancés (MX-GWR, MS-GWR).

Mot du praticien

Toutes les méthodes impliquant des statistiques spatiales nécessitent de définir ce qu'est le voisinage. Nous ne développerons pas ici ce qu'implique la définition d'un voisinage. Pour cela, nous vous renvoyons vers les travaux de Sébastien Oliveau (Oliveau 2011).

Ensuite, il sera nécessaire de définir la fonction de pondération spatiale attribuée aux voisins. Pour ce faire, on définit un rayon de voisinage (on parle de fenêtre ou bande passante ou encore de bandwidth) autour de notre observation (ce rayon peut être une distance ou un nombre de voisin). Au-delà de cette bande passante, les individus ne sont plus considérés comme voisins. Un poids va être attribué aux individus à l'intérieur de cette bande passante. Il s'agit donc des voisins de notre observation.

Figure 1 : Représentation schématique de la bande passante (bandwidth). Feuillet et al. 2019



Après ces différentes étapes nous pouvons donc procéder à la régression géographiquement pondérée et analyser dans quelle mesure les effets globaux observés sont instables spatialement.

Un exemple de régression géographiquement pondéré a été proposé par Audard et al. en 2024. Dans cet exemple, les auteurs étudient notamment l'influence de la densité de population sur le prix médian de l'immobilier au m². Les données concernent l'ensemble du territoire français métropolitain (l'hexagone) et comprennent 1223 unités administratives appelées EPCI (Établissements Publics de Coopération Intercommunale) qui regroupent les communautés de communes, les communautés d'agglomération et les communautés urbaines. La fonction de pondération choisie est gaussienne, et la bande passante définie indique que le nombre optimal de voisin est de 19.

Une approche globale montre un effet significatif et positif de la densité de population sur le prix median de l'immobilier par EPCI. Cela signifie donc, qu'une approche de **régression** classique (sans tenir compte de l'aspect spatial) indique que le prix médian de l'immobilier tend à être plus élevé dans les lieux les plus denses au sein de l'hexagone. Mais est-ce le cas sur tout le territoire de la France Hexagonale ? Les auteurs ont employé la GWR afin de répondre précisément à cette question. La GWR va donc générer 1223 coefficients de régression associés à la variable indépendante « densité de population », mais aussi autant de p-value et de R^2 locaux. Une des forces de la GWR c'est que tous ces indicateurs peuvent être cartographiés. Le premier constat c'est qu'effectivement il y a une non stationnarité des coefficients, ceux-ci vont varier de -411 à 673. Cela veut dire qu'au minimum il existe une relation négative. Dans ces espaces, quand la densité augmente d'un écart-type, le prix médian baisse de 411€. A l'inverse, dans les lieux où le coefficient est à son maximum, quand la densité augmente d'un écart-type, le prix médian augmente de 673€. Ainsi, en fonction des lieux et des observations, il se produit une inversion de signe de notre coefficient et donc un effet de la densité de la population qui n'est pas du tout le même.

Le second constat évident est la structuration spatiale mise en évidence par la cartographie des coefficients. L'interprétation des cartes GWR est une partie très intéressante, mais également très délicate. Tout l'enjeu va être d'interpréter les divergences spatiales dans le cadre de l'intensité de la relation statistique. Sont-elles dues à des effets exogènes ou endogènes ? C'est la connaissance du sujet et du territoire (ainsi que l'usage complémentaire d'autres analyses) qui va permettre de répondre à ces questions.

La GWR est une méthode riche qui permet d'explorer des phénomènes complexes en tenant compte d'une variabilité des effets dans l'espace. Toutefois, la GWR n'est bien sûr pas la seule approche pour s'intéresser à l'aspect spatial de phénomènes et de variables sociales. D'autant plus que ces dernières années de nombreuses publications ont exploré de nouvelles méthodes dérivées de la GWR pour étudier l'hétérogénéité spatiale et la non stationarité des effets adaptés à des modèles plus complexes ou des hypothèses différentes (tels que la GWR Lasso (Wheeler 2009) ou la GWR Multiscalaire (MGWR) (Fotheringham et al. 2017)).

Par ailleurs, la GWR comporte également quelques limites. Les nombreux paramétrages « manuels » (bande passante, voisinage), qui peuvent également être une force, constituent malgré tout aussi une faiblesse rendant la comparaison entre les études ou projets difficile. Il est donc important de pouvoir justifier l'ensemble des choix qui sont fait et de les faire varier afin de tester la stabilité des résultats et de trouver le meilleur modèle. Un autre problème a été relevé dans la littérature scientifique, il s'agit de la multicolinéarité locale des variables indépendantes dans le cadre des GWR multiples. (Wheeler et Tiefeldorf, 2005). Ce problème est bien connu dans le cadre des régressions et son risque est de causer notamment une variabilité non maitrisée des coefficients, des R^2 et une surestimation des erreurs types. Bien que cette question soit en débat (Fotheringham et Oshan, 2016) afin d'éviter toute difficulté il est préférable de s'assurer d'un nombre minimum d'observations dans chaque régression locale, selon les mêmes conditions que pour la **régression classique**. Enfin, l'analyse des résultats peut s'avérer complexe, il est fortement conseillé d'avoir une très bonne connaissance du sujet et du territoire de l'étude.

Ressources

- Audard Frédéric, Le Campion Grégoire et Pierson Julie, « La régression géographiquement pondérée : GWR », , 2024, [https//doi.org/10.48645/WK1M-HG05].

- Brunsdon Chris, Fotheringham A. Stewart et Charlton Martin E., « Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity », *Geographical Analysis*, n° 4, vol. 28, 1996, p. 281-298, [https//doi.org/10.1111/j.1538-4632.1996.tb00936.x].
- Bulteau Julie, Feuillet Thierry et Le Boennec Rémy, « Spatial Heterogeneity of Sustainable Transportation Offer Values: A Comparative Analysis of Nantes Urban and Periurban/Rural Areas (France) », *Urban Science*, n° 1, vol. 2, 2018, p. 14, [https//doi.org/10.3390/urbansci2010014].
- Comber Alexis, Brunsdon Christopher, Charlton Martin, Dong Guanpeng, Harris Richard, Lu Binbin, Lü Yihe, Murakami Daisuke, Nakaya Tomoki, Wang Yunqiang et Harris Paul, « A Route Map for Successful Applications of Geographically Weighted Regression », *Geographical Analysis*, n° 1, vol. 55, 2023, p. 155-178, [https//doi.org/10.1111/gean.12316].
- Feuillet Thierry, *SIGR 2021 - Atelier analyse spatiale (GWR)*, [https://sigr2021.github.io/gwr/].
- Feuillet Thierry, Coquin Julien, Mercier Denis, Cossart Etienne, Decaulne Armelle, Jónsson Helgi Páll et Sæmundsson Þorsteinn, « Focusing on the spatial non-stationarity of landslide predisposing factors in northern Iceland: Do paraglacial factors vary over space? », *Progress in Physical Geography: Earth and Environment*, n° 3, vol. 38, 2014, p. 354-377, [https//doi.org/10.1177/0309133314528944].
- Feuillet Thierry, Cossart Etienne et Commenges Hadrien, *Manuel de géographie quantitative: Concepts, outils, méthodes*, Armand Colin, 2019, [https//doi.org/10.3917/arco.illet.2019.01].
- Fotheringham A. Stewart et Oshan Taylor M., « Geographically weighted regression and multicollinearity: dispelling the myth », *Journal of Geographical Systems*, n° 4, vol. 18, 2016, p. 303-329, [https//doi.org/10.1007/s10109-016-0239-5].
- Fotheringham A. Stewart, Yang Wenbai et Kang Wei, « Multiscale Geographically Weighted Regression (MGWR) », *Annals of the American Association of Geographers*, n° 6, vol. 107, 2017, p. 1247-1265, [https//doi.org/10.1080/24694452.2017.1352480].
- Fotheringham Alexander Stewart, Brunsdon Chris et Charlton Martin, *Geographically weighted regression: the analysis of spatially varying relationships*, Nachdr. der Ausg. 2002., Chichester, Wiley, 2010.
- Hilal Mohamed et Le Gallo Julie, « Carte et modèle statistique pour explorer l'hétérogénéité spatiale », *Traitements et cartographie de l'information géographique*, Londres, Iste éditions, coll. « Géographie et démographie », 2023.
- Oliveau Sébastien, *L'espace compte ! Mesurer les structures spatiales du changement social.*, thèse de doctorat, Université d'Aix-Marseille 1, 2011.
- Pilkington Hugo, Feuillet Thierry, Rican Stéphane, Goupil De Bouillé Jeanne, Bouchaud Olivier, Cailhol Johann, Bihan Hélène, Lombrail Pierre et Julia Chantal, « Spatial determinants of excess all-cause mortality during the first wave of the COVID-19 epidemic in France », *BMC Public Health*, n° 1, vol. 21, 2021, p. 2157, [https//doi.org/10.1186/s12889-021-12203-8].
- Simpson E. H., « The Interpretation of Interaction in Contingency Tables », *Journal of the Royal Statistical Society Series B: Statistical Methodology*, n° 2, vol. 13, 1951, p. 238-241, [https//doi.org/10.1111/j.2517-6161.1951.tb00088.x].
- Wheeler David C., « Simultaneous Coefficient Penalization and Model Selection in Geographically Weighted Regression: The Geographically Weighted Lasso », *Environment and Planning A: Economy and Space*, n° 3, vol. 41, 2009, p. 722-742, [https//doi.org/10.1068/a40256].
- Wheeler David C., « Diagnostic Tools and a Remedial Method for Collinearity in Geographically Weighted Regression », *Environment and Planning A: Economy and Space*, n° 10, vol. 39, 2007, p. 2464-2481, [https//doi.org/10.1068/a38325].

- Wheeler David et Tiefelsdorf Michael, « Multicollinearity and correlation among local regression coefficients in geographically weighted regression », *Journal of Geographical Systems*, n° 2, vol. 7, 2005, p. 161-187, [https//doi.org/10.1007/s10109-005-0155-6].

Homoscédasticité n.f. [/ɔ.mo.ske.das.ti.site/]

Synonymes : homogénéité des variances, égalité des variances

A quoi ça sert ?

L'homogénéité des variances ou encore l'homoscédasticité est employée afin de vérifier si les valeurs des différents groupes que nous souhaitons comparer s'éloignent de la moyenne dans des proportions similaires. Il existe différents tests permettant de vérifier l'homogénéité des variances : le test de Breusch-Pagan (afin de vérifier l'homogénéité des variances d'erreur dans le cas de **régressions linéaires**) ou encore le test de Lévène (utilisé surtout lors de tests de comparaisons de moyennes tels que des **ANOVA** ou des **tests t**). Il existe d'autres tests que nous ne présenterons pas ici. L'homoscédasticité permet de savoir si la répartition des valeurs autour de la valeur moyenne est similaire dans l'ensemble des groupes ou non. Et par conséquent elle permet de savoir si la répartition de l'erreur autour de chaque valeur est la même ou non.

D'où ça vient ?

Le test de Breusch-Pagan a été développé par Breusch et Pagan en 1979 [103] afin de tester la constance du terme d'erreur dans le cadre d'un modèle de **régression** linéaire. Si celui-ci est constant alors il y a bien homoscédasticité et le modèle testé ne sera pas biaisé.
Le test de Lévène a été conçu par Howard Lévène en 1960 [104] afin de comparer les variances d'une variable pour au moins deux groupes. Si la répartition des valeurs autour de la valeur moyenne est équivalente dans l'ensemble des groupes testés alors les variances sont dites égales ou homogènes. Il s'agit d'un des pré-requis des tests de comparaisons de moyennes.

Le mot du praticien

Le test de Breusch-Pagan est généralement utilisé lors de la réalisation de modèles de **régression** (linéaires, de panel, etc.). Le test de Lévène est souvent employé avant de réaliser des **test t** pour échantillons indépendants ou encore sur des **ANOVA** inter-sujets.

Ressources

- Breusch T. S. et Pagan A. R., « A Simple Test for Heteroscedasticity and Random Coefficient Variation », *Econometrica*, n° 5, vol. 47, 1979, p. 1287, [https//doi.org/10.2307/1911963].
- Lévène H., « Robust Tests for Equality of Variances », *Contributions to probability and statistics; essays in honor of Harold Hotelling*, Stanford, Calif., Stanford University Press, coll. « Stanford studies in mathematics and statistics; 2 », 1960, .
- Nguyen Mike, *A Guide on Data Analysis*, 2020.

[103] Breusch T. S. et Pagan A. R., « A Simple Test for Heteroscedasticity and Random Coefficient Variation », *Econometrica*, n° 5, vol. 47, 1979, p. 1287, [https//doi.org/10.2307/1911963].
[104] Lévène H., « Robust Tests for Equality of Variances », *Contributions to probability and statistics; essays in honor of Harold Hotelling*, Stanford, Calif., Stanford University Press, coll. « Stanford studies in mathematics and statistics; 2 », 1960.

Indices de dispersion n.m.p. [/ɛ̃.dis də dis.pɛʁ.sjɔ̃/]

Synonymes : Variance, Standard deviation, Ecart-type, Etendue, *quartiles, déciles, centiles, percentiles, quantiles*

A quoi ça sert ?

Les indices de dispersion informent sur la répartition des valeurs autour d'un **indice de tendance centrale**, telles que la moyenne ou la médiane. Il existe différents indices de dispersion, comme par exemple la variance ou l'écart-type qui rendent compte d'un écart par rapport à une valeur moyenne. Il existe également les quantiles ou encore la MAD (pour Median Absolute Deviation) qui renseignent sur la dispersion des valeurs autour d'une valeur médiane[105]. Dans cette entrée uniquement la variance et l'écart-type seront présentés pour les indices de dispersion autour de la moyenne, et les quantiles pour les indices de dispersion autour de la médiane, car ce sont des indices de dispersion que nous rencontrons fréquemment en SHS.

Variance et écart-type

La variance et sa racine carrée (l'écart-type) font parties des statistiques de base que nous calculons pour les variables quantitatives avec le minimum, le maximum, la médiane, la moyenne et les quantiles. Malgré sa complexité apparente, la variance est très souvent calculée lorsque nous étudions la dispersion d'une série statistique. Cet indicateur mesure l'éparpillement des observations autour de la moyenne et permet d'effectuer des tests statistiques.

Quantiles

Les quantiles sont les valeurs qui découpent une variable quantitative en plusieurs groupes d'effectifs égaux. La valeur d'un quantile correspond à la valeur en-dessous de laquelle se situe un pourcentage de données. On distingue ainsi plusieurs types de quantiles :

- les quantiles d'ordre 4, appelés quartiles et notés Q, qui divisent la population en 4 groupes d'effectifs égaux ;
- les quantiles d'ordre 5, appelés quintiles, qui divisent la population en 5 groupes d'effectifs égaux ;
- les quantiles d'ordre 10, appelés déciles et notés D, qui divisent la population en 10 groupes d'effectifs égaux ;
- les quantiles d'ordre 100, appelés centiles et notés C, qui divisent la population en 100 groupes d'effectifs égaux.

Nous parlons ainsi de quantiles d'ordre n, pour désigner chacune des $n-1$ valeurs d'une variable quantitative qui partagent cette variable en n sous-ensembles d'effectifs égaux.

[105] Leys Christophe, Ley Christophe, Klein Olivier, Bernard Philippe et Licata Laurent, « Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median », *Journal of Experimental Social Psychology*, nº 4, vol. 49, 2013, p. 764-766, [https//doi.org/10.1016/j.jesp.2013.03.013].

Les quantiles les plus couramment utilisés sont les quartiles (utilisés par exemple dans les graphiques de type “Boîte à moustaches” popularisés par Tukey en 1977[106])[107]. Le calcul des quartiles permet d’obtenir trois valeurs, qui s’interprète de la façon suivante :

- $Q1$, le premier quartile, est la valeur au-dessous de laquelle se situent 25 % des valeurs ;
- $Q2$, le deuxième quartile ou médiane [voir la fiche spécifique], est la valeur au-dessous de laquelle se situent 50 % des valeurs ;
- $Q3$, le troisième quartile, est la valeur au-dessous de laquelle se situent 75 % des valeurs.
-

Les quantiles sont des indicateurs qui sont largement utilisés pour comprendre la répartition et la dispersion des données. Ils permettent d’identifier les valeurs atypiques dans la distribution d’une variable.

À partir de ces indicateurs, nous pouvons calculer l’écart ou l’intervalle interquartile, qui permet de mesurer une dispersion absolue. Il est égal à la différence entre Q3 et Q1 et qui est un indicateur très robuste aux valeurs extrêmes notamment dans le cas de distributions normales. Ces indicateurs permettent également de calculer le rapport inter-quantile qui permet de mesurer une dispersion relative. Il est obtenu en divisant la valeur du quantile le plus élevé par la valeur du quantile le moins élevé. Par exemple, les déciles sont souvent utilisés pour calculer le rapport interdécile $D9/D1$ et mesurer les inégalités de revenus.

D’où ça vient ?

C’est Ronald Aylmer Fisher (1890-1962), biologiste et statisticien britannique (aussi un des fondateurs de la génétique moderne), qui a employé le premier le terme de variance dans un article de 1918 « *The Correlation between Relatives on the Supposition of Mendelian Inheritance* » en tant que mesure de la variabilité d’un phénomène observé. Il a ensuite défini l’analyse de la variance telle qu’on la pratique aujourd’hui dans son livre « *Statistical methods for research workers* » paru en 1925.

Mot du praticien

Le concept de dispersion n’a de sens que dans le cas d’une variable quantitative. La variance, l’écart-type et les quantiles se calculent pour des variables continues ou discrètes.

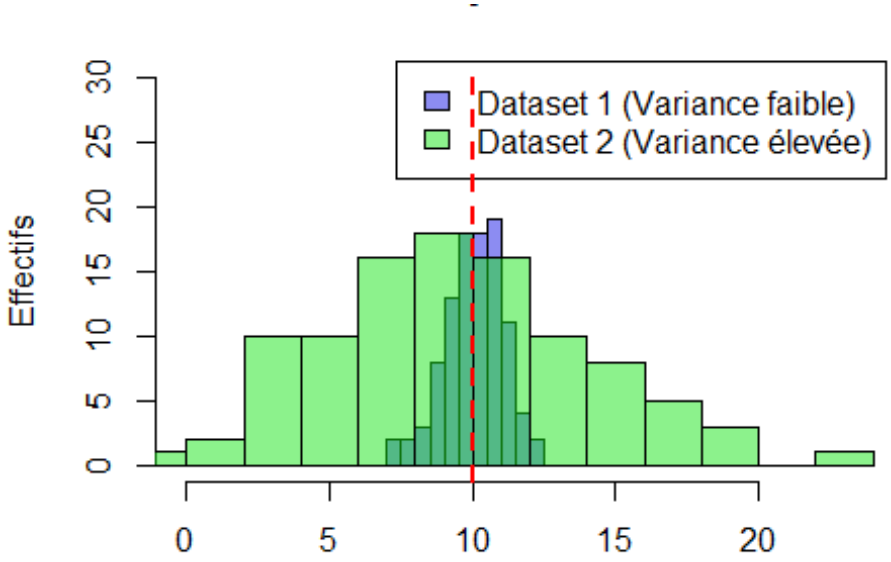
Exemple d’application

Les indices de dispersion sont utilisés dans toutes les SHS. Nous pouvons prendre l’exemple de la distance parcourue par les coureuses du Tour de France en 2024. Chaque jour les coureuses parcourent un certain nombre de kilomètres. Or, un jour elles parcourent moins de kilomètres que d’habitude, or cette petite valeur « attire à elle » la moyenne et « fait exploser » la variance. Celle-ci surévalue alors très fortement la dispersion réelle de la série observée. Il est donc important de traiter les valeurs extrêmes avant de calculer des indices de dispersion autour d’une valeur moyenne, sinon ceux-ci risquent de donner des résultats déformés et peu informatifs.
Quelques remarques importantes

[106] Tukey John Wilder, *Exploratory data analysis*, Springer, 1977.
[107] Mcgill Robert, Tukey John W. et Larsen Wayne A., « Variations of Box Plots », *The American Statistician*, n° 1, vol. 32, 1978, p. 12-16, [https//doi.org/10.1080/00031305.1978.10479236].

- La variance est nulle si et seulement si toutes les observations ont la même valeur (aucune dispersion).
- Nous divisons par n quand nous calculons la moyenne à partir d'une population complète. Sinon nous diviserons par n-1 !
- L'unité dans laquelle s'exprime la variance vaut le carré de l'unité utilisée pour les valeurs observées ce qui n'est pas parlant. Ainsi, dans l'exemple précédent, la série de distances exprimées en km possède une variance qui, elle, doit s'interpréter en "km-carré" et rien à voir avec une superficie. C'est cette difficulté dans l'interprétation de la valeur de la variance qui a incité à compléter cette mesure de dispersion en calculant l'écart-type, qui est donc la racine carrée de la variance.

L'illustration ci-dessous présente un exemple d'échantillons pour deux populations ayant la même moyenne mais des variances différentes.



moy1=moy2=10, var1=1 et var2=25

Le mot du praticien sur la variance

La variance corrigée ou non biaisée est une meilleure estimation de la population à partir d'un échantillon. L'estimation s2 (basée sur un échantillon) de la variance réelle (inconnue) d'une population utilise (n – 1) et non pas n au dénominateur. La moyenne de l'échantillon est une estimation de la moyenne de la population et cela introduit un certain biais. Diviser par n−1 permet de compenser ce biais et donne une estimation non biaisée de la variance de la population.
Par exemple, dans le contexte de **la régression linéaire**, diviser par n-2 pour calculer la variance permet au modèle d'ajuster deux paramètres (intercept et pente).

Il est à noter que lorsque nous étudions des sous-populations, la variance totale des observations est la somme de 2 quantités :

- la variance intra-classes ou variance résiduelle, qui résume la variabilité à l'intérieur des classes
- la variance inter ou variance expliquée, qui décrit les différences entre classes

Les méthodes de classification et de clustering, notamment, tendent à minimiser la variance intra-classes et à maximiser la variance inter-classes.

Le mot du praticien sur l'écart-type

L'écart-type est également très employé en SHS, car celui-ci se lit plus facilement que la variance. Il permet notamment de construire des intervalles qui contiennent un certain pourcentage des observations. Ainsi, lorsque la distribution observée est en forme de cloche et pas trop dissymétrique, nous obtenons ce qui s'appellent des « intervalles remarquables » :

- l'intervalle [moyenne - 1 écart-type; moyenne + 1 écart-type] contient à peu près 2/3 des observations ;
- l'intervalle [moyenne - 2 écart-types; moyenne + 2 écart-types] contient à peu près 95% des observations.

Le mot du praticien sur l'erreur standard

La racine carrée de la variance permet de calculer l'écart-type des données autour de leur moyenne. Mais il est ausi possible de calculer l'écart-type de la moyenne elle-même, que nous appelons l'erreur standard de la moyenne *(abréviation : e.s. en Français et s.e. en Anglais)*. Il s'agit de la racine carrée de la division de la variance par le nombre d'observations de la population considérée. Cet indicateur est très utile pour remettre en contexte la valeur moyenne d'une population. Car plus la variance des données est importante, plus l'erreur standard est grande, et moins la moyenne présentée est fiable. Au contraire une variance plus petite, renverra une erreur standard plus faible et une plus grande homogénéité des valeurs autour de la valeur moyenne, qui sera donc beaucoup plus représentative du phénomène observé.

Le mot du praticien sur les quantiles

Les quantiles sont très utilisés en cartographie pour représenter les données quantitatives sur une carte. Une méthode de discrétisation s'appuie sur la méthode des quantiles, avec 4 à 5 classes qui regroupent toutes le même nombre d'individus. Elle est particulièrement conseillée quand un jeu de données comporte quelques valeurs extrêmes.

Ressources

- Bertrand Frédéric et Maumy-Bertrand Myriam, *Initiation à la statistique avec R*, 4e éd., Malakoff, Dunod, coll. « Sciences sup », 2023.
- CartONG, *Information Management Resource Portal - Learning Corner*, [https://cartong.pages.gitlab.cartong.org/learning-corner/fr/intro_data_analysis].
- Fisher, R. A., *Statistical methods for research workers*, 1st éd., Oliver and Boyd, 1925.
- Fisher R. A., « XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. », *Transactions of the Royal Society of Edinburgh*, n° 2, vol. 52, 1919, p. 399-433, [https//doi.org/10.1017/S0080456800012163].
- Leys Christophe, Ley Christophe, Klein Olivier, Bernard Philippe et Licata Laurent, « Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median », *Journal of Experimental Social Psychology*, n° 4, vol. 49, 2013, p. 764-766, [https//doi.org/10.1016/j.jesp.2013.03.013].
- Massé Antoine, *Aide à l'utilisation de R - Se former à R*, [https://sites.google.com/site/rgraphiques/se-former-a-r].
- Mcgill Robert, Tukey John W. et Larsen Wayne A., « Variations of Box Plots », *The American Statistician*, n° 1, vol. 32, 1978, p. 12-16, [https//doi.org/10.1080/00031305.1978.10479236].

- Simon-Bouhet Benoît, *4 Visualiser des données avec ggplot2 | Travaux Pratiques de Biométrie 2*, 2022.
- Tukey John Wilder, *Exploratory data analysis*, Springer, 1977.

Indices de tendance centrale n.m.p [/ɛ̃.dis də tɑ̃.dɑ̃s sɑ̃.tʁal/]

Synonymes : deuxième quartile, valeur médiane, median, second quartile

A quoi ça sert ?

Les indices de tendance centrale présentés ici seront la médiane et la moyenne. Il en existe d'autres tel que le mode, qui ne seront pas présentés, mais dont vous pouvez facilement trouver des ressources sur le sujet dans la bibliographie proposée dans cette entrée.
Il est à noter que les indices de tendance centrale se calculent uniquement pour des variables quantitatives, qu'elles soient continues ou ordinales.

Médiane

La médiane fait partie des indicateurs de tendance centrale qui résument une variable, au même titre que la moyenne. La médiane est la valeur qui partage la distribution en deux classes d'effectifs égaux. En d'autres termes, cela signifie que 50 % des valeurs sont inférieures à la valeur médiane et 50% ont des valeurs supérieures à la valeur médiane.
La médiane est très utile pour résumer et comprendre la répartition des données, en particulier en présence de valeurs extrêmes (éloignées du reste de l'échantillon), qui influencent la valeur de la moyenne.

Graphiquement, nous pouvons visualiser la médiane, entre autres, sur les graphiques appelés boîtes à moustaches (boxplot) : elle est représentée par la ligne centrale dans ces graphiques.

Pour calculer cet indicateur, les valeurs de la variable doivent pouvoir être classées de la plus petite à la plus grande. Une fois que les valeurs sont ordonnées par ordre croissant, on a deux possibilités pour trouver la médiane :

1. Si le nombre d'observations est impair, la médiane est la valeur qui est située au milieu de l'ensemble. Elle se trouve donc de manière assez simple, puisque c'est une valeur qui existe réellement.
2. En revanche, dans le cas où le nombre d'observations est pair, trouver la médiane nécessite d'effectuer un calcul qui reste assez simple. La valeur médiane est alors la moyenne arithmétique entre les deux valeurs situées au centre de la distribution.

Moyenne

La moyenne fait partie des statistiques descriptives que nous calculons pour les variables quantitatives avec le minimum, le maximum, la médiane, les quantiles et l'écart-type. C'est un indicateur de tendance centrale d'un ensemble de données au même titre que la médiane, le mode et la valeur centrale (moyenne des valeurs minimale et maximale). Toutefois, la moyenne peut être sensible aux valeurs extrêmes contrairement à la médiane.

D'où ça vient ?

C'est Pythagore, philosophe et mathématicien grec du VIe siècle av.J.-C., qui a fourni les quatre définitions de la moyenne, présentées précédemment[108].

Concernant la médiane, ses origines remontraient à 1599 et sont attribuées à Edward Wright dans un ouvrage sur la navigation (Bakker, A., Gravemeijer, K.P.E., 2006)[109]. Toutefois, la présentation de la médiane telle qu'elle est réalisée dans cette ouvrage ne s'entend pas de la même manière que la façon dont elle est employée actuellement. Legendre, (1805) et Laplace, (1812/1891)[110] ont contribué au développement de la médiane, telle que nous l'employons aujourd'hui, en calculant ce qu'ils appelaient "le milieu de probabilité" (Bakker, A., Gravemeijer, K.P.E., 2006). Mais c'est Cournot (1843/1984) qui employa le terme de "médiane" pour la première fois (Bakker, A., Gravemeijer, K.P.E., 2006).

Le mot du praticien sur la médiane

La médiane est utilisée pour résumer la tendance centrale des données. Elle n'est pas sensible aux valeurs extrêmes (ces valeurs n'ont pas d'influence sur la médiane). Lorsque l'on est en présence de valeurs extrêmes ou lorsque la distribution est asymétrique, on préfèrera donc résumer la distribution de la variable par la médiane plutôt que la valeur moyenne. Cet indicateur est donc plus robuste que la moyenne. Comme pour cette dernière, il est possible d'associer un indicateur de dispersion à la médiane, qui va permettre de mesurer la répartition des individus autour de cet indicateur de tendance centrale. Pour plus d'information sur l'écart interquartile, l'indicateur de dispersion lié à la médiane, se référer à l'entrée **indices de dispersion**. Il existe également un autre indice de dispersion lié à la médiane appelé MAD - Median Absolute Deviation popularisé par Leys et al. en 2013 afin de détecter les valeurs extrêmes[111]. Cet indice de dispersion permet notamment de détecter des valeurs extrêmes à partir d'une valeur seuil qui n'est pas influencée par les valeurs extrêmes comme ce peut être le cas pour les écart-types.

La médiane se révèle donc être un indicateur pertinent lorsque nous sommes face à des données qui comportent de nombreuses valeurs aux extrémités de la distribution et que celle-ci ne suit pas de loi normale. On peut par exemple citer la façon de rendre compte des revenus des français. En 2018, le revenu moyen disponible des ménages français est de 37670€, alors que le revenu médian des ménages français s'élève à 30620€[112]. La différence observée est due à la présence de très hauts revenus qui impactent la valeur moyenne, alors que la valeur médiane reste robuste face à

[108] Heath Thomas Little, *A history of Greek mathematics*, Oxford, The Clarendon press, 1921.

[109] Bakker Arthur et Gravemeijer Koeno P. E., « An Historical Phenomenology of Mean and Median », *Educational Studies in Mathematics*, n° 2, vol. 62, 2006, p. 149-168, [https//doi.org/10.1007/s10649-006-7099-8].

[110] Stigler Stephen M., « Studies in the History of Probability and Statistics. XXXII: Laplace, Fisher, and the discovery of the concept of sufficiency », *Biometrika*, n° 3, vol. 60, 1973, p. 439-445, [https//doi.org/10.1093/biomet/60.3.439].

[111] Leys Christophe, Ley Christophe, Klein Olivier, Bernard Philippe et Licata Laurent, « Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median », *Journal of Experimental Social Psychology*, n° 4, vol. 49, 2013, p. 764-766, [https//doi.org/10.1016/j.jesp.2013.03.013].

[112] Source : INSEE, *Revenu disponible des ménages − Revenus et patrimoine des ménages*, [https://www.insee.fr/fr/statistiques/5371205?sommaire=5371304] ; INSEE, *En 2018, les inégalités de niveau de vie augmentent - Insee Première - N°1813*, [https://www.insee.fr/fr/statistiques/4659174].

ces valeurs extrêmes et renvoie une valeur plus proche de celle connue par la majorité de la population.

Cette mesure de centralité d'un jeu de données est un indicateur également utilisé pour représenter les données géographiques. La discrétisation dite "Q6" qui divise un jeu de données en six classes s'appuie sur plusieurs indicateurs, dont la médiane pour cet aspect central. La médiane est en effet très utilisée en SHS afin de décrire des données. Il est notamment utile de comparer sa valeur à la valeur moyenne pour résumer au mieux l'information donnée par les valeurs d'une ou plusieurs variables.

Le mot du praticien sur la moyenne

La moyenne est une notion très familière que nous utilisons sans se poser trop de questions et pour résumer une série statistique, les logiciels que nous utilisons font souvent le calcul par défaut. Pourtant, il existe différentes moyennes. Mais à quoi servent-elles ? Laquelle utiliser ?

Il existe en effet 4 définitions de la moyenne, à adapter selon la nature des données :

a, la moyenne arithmétique (ou empirique)
La moyenne arithmétique reflète la valeur centrale d'un ensemble de données. Elle se calcule en additionnant toutes les valeurs puis en divisant cette somme par le nombre total de valeurs. Elle est sensible à chaque valeur d'un ensemble de données, ce qui signifie que tout changement dans l'ensemble de données affectera la valeur moyenne. Elle est particulièrement sensible aux **valeurs extrêmes**. La moyenne arithmétique est souvent employée en économie pour traiter des questions relatives aux revenus et aux salaires. Elle est alors utilisée pour calculer le revenu moyen ou le salaire moyen des individus dans un secteur ou une région donnée.

g, la moyenne géométrique
Cette moyenne est moins sensible que la moyenne arithmétique aux valeurs les plus élevées. Elle ne s'applique pas aux distributions qui suivent la loi normale c'est à dire où les fréquences décroissent rapidement de façon exponentielle. Nous l'utilisons pour des ensembles de données positives (pas de valeur négatives ni égales à 0). Elle est donc particulièrement adaptée lorsque les données sont multiplicatives (par exemple, taux de croissance ou rendements), lorsque les valeurs sont sur des échelles logarithmiques ou si elles sont exprimées en termes de ratios ou pourcentages. La moyenne géométrique est également employée en économie. Elle est par exemple employée pour calculer le rendement moyen d'un investissement lorsque les rendements d'une période se basent sur la valeur finale d'une période précédente. Contrairement à la moyenne arithmétique, elle tient compte des effets de la capitalisation des rendements.

h, la moyenne harmonique
La moyenne harmonique est particulièrement sensible aux petites valeurs de l'ensemble de données. Si une ou plusieurs des valeurs sont très petites par rapport aux autres, elles auront un fort impact sur le résultat final. À l'inverse, la moyenne harmonique réagit peu aux grandes valeurs. Elle est utilisée pour des ensembles de données positives (pas de valeur négatives ni égales à 0). Comparée à la moyenne arithmétique ou géométrique, la moyenne harmonique est plus adaptée aux situations où nous souhaitons modéliser des phénomènes où il existe des liens de proportionnalité inverse (taux, rendements ou ratios comme des vitesses).
Dans les enquêtes où les temps de réponse sont mesurés, la moyenne harmonique est souvent utilisée pour calculer la moyenne des taux de réponse lorsque les données sont des inverses de

temps (Taux de réponse=réponses par minute=1/Temps de réponse). Autre exemple, pour calculer la vitesse moyenne lorsque les distances parcourues à différentes vitesses sont impliquées, la moyenne harmonique peut être appropriée.

q, la moyenne quadratique
La moyenne quadratique est très sensible aux grandes valeurs car elles sont élevées au carré. Elle reflète l'amplitude des variations dans un ensemble de données. En statistique, la moyenne quadratique s'utilise principalement pour estimer l'écart moyen entre les éléments et le centre du nuage (défini par la moyenne arithmétique). Elle est utile dans les cas où nous voulons éviter que les écarts négatifs et positifs ne se compensent (comme avec la moyenne arithmétique). Les applications les plus courantes de la moyenne quadratique sont dans les analyses où nous nous s'intéressons à la dispersion des données, aux écarts-types et aux évaluations de la performance des modèles prédictifs tels que : l'**analyse de la variance**, la **régression linéaire**, les écarts-types des rendements, les **k-means**, etc.

Entre ces 4 moyennes, nous observons ces inégalités : h<=g<=a<=q

Il est important de préciser qu'en amont du calcul des statistiques descriptives pour les variables numériques (dont la moyenne) il est fortement recommandé de visualiser, au préalable à tous calculs statistiques visant à résumer une série de données, les données qui vous sont confiées (ex. sous forme de courbes, d'histogramme ou de nuages de points pour une ou plusieurs variables quantitatives). Deux variables avec des statistiques d'aspect normal (moyenne, écart-type et corrélation de Pearson) peuvent révéler l'image d'un dinosaure en les croisant ![113]

Les moyennes sont toujours comprises entre les valeurs minimales et maximales de nos données. Il est nécessaire de le vérifier à chaque fois.

La moyenne arithmétique est la plus utilisée en SHS, toutefois il est pertinent de ne pas oublier que d'autres moyennes existent et qu'elles peuvent se révéler utiles dans des cas rencontrés dans l'ensemble des disciplines des SHS. Les calculs de moyennes sont généralement réalisés lors de la phase de description des données et constituent un préalable important en amont des analyses inférentielles, au même titre que d'autres indices de tendance centrale, telle que la médiane par exemple.

Ressources

- Antoine Charles, *Les moyennes*, 1. éd., Paris, Presses Universitaires de France, coll. « Que sais-je ? », 1998.
- Bakker Arthur et Gravemeijer Koeno P. E., « An Historical Phenomenology of Mean and Median », *Educational Studies in Mathematics*, n° 2, vol. 62, 2006, p. 149-168, [https//doi.org/10.1007/s10649-006-7099-8].
- Heath Thomas Little, *A history of Greek mathematics*, Oxford, The Clarendon press, 1921.
- Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.

[113] cf. le datasaurus Dozen: Matejka Justin et Fitzmaurice George, *The Datasaurus data package*, 2025.
Matejka Justin et Fitzmaurice George, « Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing », Denver Colorado USA, ACM.

- INSEE, *Revenu disponible des ménages − Revenus et patrimoine des ménages*, [https://www.insee.fr/fr/statistiques/5371205?sommaire=5371304].
- INSEE, *En 2018, les inégalités de niveau de vie augmentent - Insee Première - N°1813*, [https://www.insee.fr/fr/statistiques/4659174].
- Jubénot Marie-Noëlle et Eudes Daniel, *Analyse des données sous R pour les sciences humaines: théories et exemples commentés*, Paris, Ellipses, 2022.
- Klein Grady et Dabney Alan, *The cartoon introduction to statistics*, First edition., New York, Hill and Wang, a Division of Farrar, Straus and Giroux, 2013.
- Leys Christophe, Ley Christophe, Klein Olivier, Bernard Philippe et Licata Laurent, « Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median », *Journal of Experimental Social Psychology*, nº 4, vol. 49, 2013, p. 764-766, [https//doi.org/10.1016/j.jesp.2013.03.013].
- Matejka Justin et Fitzmaurice George, *The Datasaurus data package*, 2025.
- Matejka Justin et Fitzmaurice George, « Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing », Denver Colorado USA, ACM.
- Observatoire du développement Durable, *Sous portail odr — Wiki ODR*, [https://odr.inrae.fr/intranet/carto/cartowiki/index.php/Sous_portail_odr].
- Stigler Stephen M., « Studies in the History of Probability and Statistics. XXXII: Laplace, Fisher, and the discovery of the concept of sufficiency », *Biometrika*, nº 3, vol. 60, 1973, p. 439-445, [https//doi.org/10.1093/biomet/60.3.439].

Kaplan-Meier n.m. [/ka.plan ma.jɛʁ/]

A quoi ça sert ?

Le modèle de Kaplan-Meier vise à estimer la probabilité de survie à différents moments dans le temps. Il répond à la question fondamentale : "Quelle est la probabilité qu'un individu survive (ou reste sans événement) jusqu'au temps t ?". On entend par survie le temps qui s'écoule avant qu'un événement d'intérêt se produise. On pourra aussi parler de "temps jusqu'à l'événement" ou "durée de vie".

Kaplan-Meier est une méthode non-paramétrique qui estime directement la fonction de survie (probabilité de survie jusqu'au temps t) sans faire d'hypothèses sur la distribution sous-jacente.
Ce modèle ne peut analyser qu'une variable à la fois. Pour comparer des groupes on va calculer des courbes de survie séparées, cela nous permettra de comparer le taux de survie du groupe A VS le groupe B.

Kaplan-Meier permet d'estimer des courbes de survie, des probabilités de survie à des temps donnés, des durées médianes de survie. Il offre également la possibilité de comparer la courbe de survie entre différents sous-groupes et d'estimer la significativité des différences observées via des tests statistiques comme le log-rank.

Les calculs de survie tirés de Kaplan-Meier s'appuient sur l'ensemble des observations disponibles, y compris celles dites censurées, c'est-à-dire les individus qui sortent de l'observation avant d'avoir connu l'événement étudié, une situation fréquente dans les enquêtes en coupe transversale menées auprès de personnes d'âges variés.

D'où ça vient ?

L'analyse de survie trouve ses racines dans les tables de mortalité créées par John Graunt en 1662 à Londres (Graunt, 1962). Il analyse les registres de décès pour comprendre les patterns de mortalité urbaine et propose les premières estimations de la probabilité de décès selon l'âge. C'est donc un type d'analyse très ancien. La méthode de Kaplan-Meier elle a été développée par Edward L. Kaplan et Paul Meier en 1958. Leur article fondateur "Nonparametric Estimation from Incomplete Observations" a été publié dans le Journal of the American Statistical Association.

L'enjeu était d'analyser les données de survie avec des observations incomplètes (censurées). Les méthodes existantes étaient limitées sur ce point. Kaplan et Meier proposent une approche non-paramétrique (car elle ne repose pas sur une hypothèse de la distribution des fonctions de survie), qui permet d'estimer la fonction de survie en multipliant des probabilités conditionnelles de survie à chaque temps d'événement observé. Cette méthode du "produit-limite" permet d'utiliser toute l'information disponible, y compris des données censurées.

Mot du praticien

Kaplan-Meier est une approche univariée et descriptive. Elle ne permet pas d'ajuster les estimations simultanément sur plusieurs facteurs ni de quantifier l'effet propre de chacune des variables susceptibles d'être associées au temps de survie. C'est pourquoi elle est souvent utilisée en exploration préliminaire avant des analyses plus complexes comme le modèle de Cox.

Son rôle reste néanmoins fondamental. Kaplan-Meirer offre une description claire et visuelle des patterns de survie, qui sert de base à des analyses de survie plus sophistiquées.

Ressources

- Kaplan E. L. et Meier Paul, « Nonparametric Estimation from Incomplete Observations », in Samuel Kotz et Norman L. Johnson (dir.), *Breakthroughs in Statistics*, New York, NY, Springer New York, 1992, p. 319-337, [https//doi.org/10.1007/978-1-4612-4380-9_25].

Khi-deux n.m [/ki dø/]

Synonymes : test du Khi-deux d'indépendance de Pearson, test d'association du Khi-deux, Chi-square, Khi-deux, Khi², Chi-deux, Chi², χ^2, Chi-Square Test of Independence

A quoi ça sert ?

Le test du Khi-deux ou Khi² présenté dans cette entrée est celui du Khi² d'indépendance[114]. Il s'agit d'un test d'hypothèses (*non paramétrique*) très utilisé en SHS, toute discipline confondue, lorsque nous souhaitons valider l'existence d'une relation entre deux variables catégorielles issues d'un échantillon interrogé par questionnaire. La conclusion établissant l'existence d'un lien traduit le fait que les deux variables sont dépendantes. C'est le cas par exemple quand nous cherchons à nous assurer qu'une opinion ou un comportement politique (ex : être pour ou contre la peine de mort) présente un lien significatif avec le niveau d'éducation des individus ; ou encore si une mesure mise en place pour la préservation d'un site naturel est mieux acceptée pour certaines catégories de revenus des ménages que d'autres.

En pratique, il est nécessaire de réaliser un tableau de contingence (**tableau de fréquences**) entre les deux variables considérées. Puis, nous testons si les effectifs observés dans un échantillon sont significativement différents des valeurs théoriques (attendues), c'est-à-dire dans le cas où il n'y aurait aucun lien entre les deux variables[115]. En effet, les effectifs théoriques correspondent à la répartition des effectifs si leur répartition était uniforme. C'est à dire si aucun des croisements n'était sur-représenté ou sous-représenté. Ainsi, si la différence est suffisamment grande entre cette répartition uniforme théorique et la répartition des effectifs observés (réels), cela permet de mettre en évidence l'existence d'une relation entre les deux variables. Celles-ci ne sont donc pas indépendantes. En situation d'indépendance, les pourcentages en lignes et en colonnes sont identiques pour chaque modalité des variables considérées, et correspondent aux pourcentages marginaux du tableau empirique. En d'autres termes, chaque modalité d'une variable se répartit à l'identique dans les modalités de l'autre, et donc les variables n'ont pas d'effet l'une sur l'autre (exemple : quel que soit le niveau d'éducation, la même proportion d'individus se déclare pour la peine de mort).

D'où ça vient ?

C'est Karl Pearson, statisticien britannique considéré comme l'un des fondateurs des statistiques modernes, qui le premier, proposa en 1900 un test statistique d'ajustement d'une distribution observée à une distribution attendue (Pearson, K., 1900)[116]. Le test du Khi-deux d'indépendance est l'un des trois tests statistiques du Khi-deux développés par Pearson[117].

---

[114] Car il existe différents tests du khi², le Khi² d'ajustement ou d'homogénéité afin de tester si la répartition d'un échantillon entre les modalités d'une même variable est homogène ; le khi² de distribution afin de comparer si les valeurs observées sont égales à des valeurs attendues connues et le khi² d'indépendance présenté ici.

[115] Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.

[116] Pearson Karl, « X. *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling* », *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, nº 302, vol. 50, 1900, p. 157-175, [https//doi.org/10.1080/14786440009463897].

[117] Les autres types de tests du Khi-deux de Pearson, sont le test du Khi-deux d'adéquation à une loi et le test d'homogénéité de distribution.

Ces travaux initiaux ont donné suite à d'autres développements, qui ont contribué à populariser le test du Khi-deux. On peut citer en particulier Ronald Fisher, dont les travaux ont conduit à établir une table qui donne les quantiles de la loi du Khi-deux pour différentes valeurs du nombre de degrés de liberté ou encore à interpréter correctement les résultats du test du Khi-deux (Fisher, R. A., 1922)[118].[La contribution de Fisher, en rendant l'usage du khi-deux plus rigoureux est déterminante, encore aujourd'hui, dans nos pratiques. Il a également permis à d'autres statisticiens de dépasser l'une de ses principales limites : le khi-deux indique s'il y a une association, mais pas sa force (son intensité). Pearson a introduit le coefficient **Phi** pour mesurer l'intensité de l'association dans **un tableau de fréquences** à deux variables (2X2) mais il ne fonctionne que pour les tableaux 2x2. Alexander Tschuprow, en 1939 généralise le Phi de Pearson aux tableaux plus grands, introduisant le T de Tschuprow[119]. En 1946, Harald Cramér propose une variante du T de Tschuprow, le V de Cramer, plus simple à interpréter et adapté aux tableaux carrés[120]. Ainsi, bien que le T de Tschuprow ait été développé en premier, on en parle souvent comme d'une alternative au V de Cramer car c'est ce dernier qui est devenu la référence : il est plus facile à comprendre et interpréter (varie de 0 à 1). Il est utilisé dans des logiciels statistiques populaires (R, SPSS, etc.). Pour un tableau carré, T de Tschuprow et V de Cramer donnent des résultats similaires. Par contre, le T de Tschuprow reste plus fiable pour les tableaux asymétriques (car il évite la surestimation).

Cette entrée présente uniquement le test du Khi-deux d'indépendance, qui peut être mobilisé pour tester l'existence ou non d'une relation entre deux variables A et B, lorsque celles-ci sont de nature **qualitative**, c'est-à-dire nominales ou ordinales, ou dans le cas de variables quantitatives, dont on a regroupé les valeurs en classes.

Le mot du praticien

Le test n'est pas applicable dans n'importe quelles conditions. Tout d'abord, il est nécessaire que les observations soient indépendantes les unes des autres. C'est-à-dire qu'une valeur présente dans notre base de données ne renseigne en rien sur les autres valeurs du reste de notre base de données.

Ensuite, avant de réaliser ce test, il faut s'assurer que les deux variables sur lesquelles porte le test sont des variables de nature qualitative. Si ce n'est pas le cas, il faut regrouper les valeurs dans des classes ou des catégories (la discrétisation).

Un troisième pré-requis concerne les effectifs. Pour que le test du Khi-deux d'indépendance soit valide, il faut une taille d'échantillon suffisamment grande. S'il est difficile de donner une valeur précise pour la taille de l'échantillon (elle s'apprécie en fonction du contexte et de la complexité des données), en ce qui concerne les effectifs théoriques, les conditions sont les suivantes :

- au moins 80% des effectifs théoriques doivent être au moins égaux à 5
- tous les effectifs théoriques doivent être supérieurs à 1.
  Ces deux conditions d'application sont connues comme étant la règle de Cochran.

À noter : les effectifs observés peuvent quant à eux prendre n'importe quelle valeur, y compris la valeur 0.

---

[118] Fisher R. A., « On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P », *Journal of the Royal Statistical Society*, nº 1, vol. 85, 1922, p. 87, [https//doi.org/10.2307/2340521].

[119] Tschuprow a a, *Principles Of The Mathematical Theory Of Correlation.*, 1939.

[120] Cramer Harald, *Mathematical Methods Of Statistics*, Princeton University Press, 1946.

Dans le cas où le test porte sur un faible nombre d'individus, on pourra employer le test exact de Fisher. Il peut s'appliquer sur des petits effectifs et n'a pas de limite sur les effectifs théoriques. Ce test constitue également une alternative pour les tableaux de contingence de taille 2×2 (à savoir 2 colonnes pour les 2 modalités de la variable A et 2 lignes pour les 2 modalités de la variable B). Pour le cas des tableaux 2×2, il est également possible d'appliquer la **correction de continuité de Yates**, tout-à-fait adaptée aux fréquences théoriques faibles, mais celle-ci fait l'objet de nombreux débats dans la littérature et face à la possibilité d'employer le test exact de Fisher, apparaît peu utile (Howell, 2008)[121].

Le test du Khi-deux est très fréquemment utilisé en SHS. Par exemple les auteurs Dehez et al. 2024 utilisent un test de Khi-deux d'indépendance afin d'appréhender les risques liés à la baignade[122]. Ils émettent l'hypothèse d'une relation de dépendance entre le fait de se baigner dans l'océan et l'âge des enquêtés. Le Khi-deux a donc été utilisé afin de tester cette hypothèse. Celui-ci a indiqué une non-indépendance entre ces deux variables, les effectifs se répartissant différemment selon le fait de se baigner et l'âge des participants.

L'utilisation du test du Khi² fait l'objet d'un consensus en SHS. Toutefois, les données en entrée et leur structuration seront différentes, ce qui implique un usage des corrections (telle que la correction de continuité de Yates par exemple)[123] ou des tests alternatifs (Fisher) qui peut différer selon les disciplines.

En limite de test (si la règle de Cochran n'est pas respectée ou tout juste), il est fréquent de réaliser des regroupements de classes afin d'augmenter les effectifs ou de supprimer des modalités du type "sans réponse". Toutefois, si les pré-requis minimums à l'application d'un test du Khi² paraissent faciles à remplir (5 observations par case pour l'effectif théorique), les choses sont moins claires sur les échantillons plus importants, avec une définition de l'importance de la taille de l'échantillon qui varie selon les disciplines. En effet, le khi², comme la majorité des tests statistiques, a tendance à être systématiquement significatif quand l'échantillon est trop important. Les effets de petite taille se retrouvent à être artificiellement mis en avant. Pour pallier ce problème, certaines disciplines (Psychologie, Sociologie, par exemple) opèrent une réduction de leur échantillon pour entrer dans le cadre ou utilisent le V de Cramer (Cramer, H., 1946). En géographie, le T de Tschuprow est également utilisé dans ce contexte. Pour Bergsma, 2013, le T de Tschuprow (Tschuprow, 1925, 1939), étroitement lié et plus ancien, mais moins connu, présente certains avantages théoriques possibles : il est notamment plus robuste pour les tableaux dissymétriques[124].

---

[121] Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.

[122] Dehez Jeoffrey et Lyser Sandrine, « How ocean beach recreational quality fits with safety issues? An analysis of risky behaviours in France », *Journal of Outdoor Recreation and Tourism*, vol. 45, 2024, p. 100711, [https//doi.org/10.1016/j.jort.2023.100711].

[123] Par exemple la correction de continuité de Yates va corriger les résultats obtenus en réduisant la valeur absolue de chaque numérateur de 0.5 unité, puis de l'élever au carré. Une correction est un calcul appliqué à un résultat afin d'affiner ou de préciser celui-ci selon le contexte dans lequel il est obtenu (petit échantillon, etc.). Howell, D. C. (2013). *Méthodes statistiques en sciences humaines* (2e éd.). De Boeck Supérieur.

[124] Bergsma Wicher, « A bias-correction for Cramér's and Tschuprow's », *Journal of the Korean Statistical Society*, nº 3, vol. 42, 2013, p. 323-328, [https//doi.org/10.1016/j.jkss.2012.10.002].

Ressources

- Bergsma Wicher, « A bias-correction for Cramér's and Tschuprow's », *Journal of the Korean Statistical Society*, n° 3, vol. 42, 2013, p. 323-328, [https//doi.org/10.1016/j.jkss.2012.10.002].
- Cramer Harald, *Mathematical Methods Of Statistics*, Princeton University Press, 1946.
- Dehez Jeoffrey et Lyser Sandrine, « How ocean beach recreational quality fits with safety issues? An analysis of risky behaviours in France », *Journal of Outdoor Recreation and Tourism*, vol. 45, 2024, p. 100711, [https//doi.org/10.1016/j.jort.2023.100711].
- Fisher R. A., « On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P », *Journal of the Royal Statistical Society*, n° 1, vol. 85, 1922, p. 87, [https//doi.org/10.2307/2340521].
- Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.
- MetricGate Team, *MetricGate | AI-Powered Statistical Analysis & R Integration*, [https://metricgate.com/docs/tschuprows-t/].
- Pearson Karl, « X. *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling* », *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, n° 302, vol. 50, 1900, p. 157-175, [https//doi.org/10.1080/14786440009463897].
- Tschuprow a a, *Principles Of The Mathematical Theory Of Correlation.*, 1939.

K-Means n.m [keɪ miːnz]

Synonymes : Nuées dynamiques

A quoi ça sert ?

La méthode des K-means est un outil puissant pour identifier des structures cachées dans les données en les regroupant en classes homogènes. C'est une technique largement utilisée en SHS en raison de sa simplicité et de son efficacité. Le but est de diviser un ensemble de données en K classes distincts, où K est un nombre qui va être défini au préalable avant de lancer l'analyse.

L'objectif des K-means est de minimiser la variance intra-classe (c'est-à-dire la somme des distances entre chaque point de données et le centre de sa classe) et de maximiser la variance inter-classe (c'est-à-dire la distance entre les centres des classes). Cette méthode appartient au même « champ » que la **classification ascendante hiérarchique (CAH)**.

D'où ça vient ?

L'algorithme du k-means, proposé par MacQueen en 1967[125], est une technique de clustering (ou partitionnement) non supervisé, utilisée pour regrouper des données en un certain nombre de clusters ou groupes basés sur leurs caractéristiques similaires. Cette méthode appartient au même champ que la **CAH** mais contrairement à la **CAH**, le k-means nécessite de fixer au préalable le nombre de classes désirées.

Pré-requis :

La méthode des K-Means suit 6 grandes étapes

1. Choisir le nombre de classes K dans lesquelles les données seront regroupées.
2. Initialisation des centroids : l'algorithme sélectionne aléatoirement K points dans les données qui serviront de centres initiaux (appelés centroides) pour le nombre K de classes.
3. Assignation des individus aux classes : chaque individu de la base de données est assigné au cluster dont le centroïde est le plus proche.
4. Recalcul des centroides : une fois l'ensemble des individus assignés, le centroïde est recalculé pour chaque cluster en prenant la moyenne des individus qui y sont assignés.
5. Répétition des étapes d'assignation et de recalcul : les étapes 3 et 4 sont répétées jusqu'à ce que les centroides ne changent plus de position (convergence) ou que le changement soit en-dessous d'un seuil déterminé.
6. Résultat final : une fois la convergence atteinte, les individus de la base de données sont regroupés en K classes, chacun avec son propre centroïde.

Le k-means est donc une méthode heuristique, puisque les résultats varient sensiblement en fonction du tirage ou de la sélection des centres de classes initiaux, et surtout, en fonction du nombre d'itérations. Comparativement à la **CAH**, le k-means est tout à fait adapté à la partition de vastes jeux de données, l'inconvénient étant qu'il est nécessaire de connaître au préalable le nombre de

[125] MacQueen J., « Some methods for classification and analysis of multivariate observations », *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, University of California Press, 1967, p. 281-298.

classes. Toutefois, pour pallier cet inconvénient, il est tout à fait possible de réaliser plusieurs k-means (de k=3 à k=15 par exemple), puis d'analyser des mesures de qualité de classification afin de sélectionner le nombre de classes optimal.

Le mot du praticien

En géographie, le k-means a été largement appliqué à des données spatiales, par exemple lors d'études sur les trajectoires de concentration de la pauvreté (Apparicio et al., 2015)[126]. L'objectif des auteurs est de mieux caractériser les trajectoires de changement de la pauvreté dans les quartiers de Montréal. Dans cette étude deux techniques de regroupement, à 20 ans de données de recensement (cinq points dans le temps), ont été appliquées et comparées afin d'identifier des groupes de quartiers ayant suivi une trajectoire similaire de pauvreté entre 1986 et 2006. En français, nous pouvons conseiller la lecture de l'article de Richer et Palmier (2012)[127] qui ont utilisé les k-means pour étudier l'accessibilité aux transports collectifs dans la métropole lilloise.

La méthode des K-means est fréquemment utilisée car elle présente comme avantage d'être simple, efficace et rapide à exécuter. Et son interprétation n'est pas trop complexe. En revanche, cette méthode possède également plusieurs limites importantes. Le nombre de classes doit être spécifié à l'avance, nombre qui peut être difficile à déterminer voire arbitraire. L'algorithme peut converger vers des solutions locales en fonction des centroides initiaux choisis, ce qui fait qu'il est possible d'obtenir des résultats différents en relançant l'analyse. Enfin, les K-means supposent que les clusters sont sphériques et de taille similaire, ce qui peut ne pas toujours correspondre à la réalité.

Pour lutter contre certaines de ses limites, il est possible d'évaluer la qualité du clustering produit par K-means. Pour ce faire différentes mesures peuvent être employées telles que :
- L'inertie intra-classes : la somme des distances au carré entre chaque point et son centroïde. L'objectif est de minimiser cette valeur.
- L'indice de silhouette : une mesure de la séparation entre les clusters. Un score proche de 1 indique que les points sont bien regroupés dans leurs clusters et bien séparés des autres clusters.

Il existe également des variantes des K-Means qui visent à réduire ces limites :
- K-means++ : Une variante de l'algorithme qui améliore l'initialisation des centroides pour obtenir de meilleurs résultats et éviter les solutions locales.
- Mini-batch K-means : Une version de K-means plus efficace pour les grands ensembles de données, où l'algorithme est appliqué à des sous-ensembles aléatoires de données (mini-batches) pour accélérer la convergence.

On peut aussi citer les k-médianes (Jain, Dubes, 1988) ou encore les k-médoïdes (Kaufman, Rousseeuw, 1987). Aussi, des auteurs comme Wong (1982) et Lebart et al. (1995, sect. 2.3) ont proposé une méthode de classification mixte, combinant à la fois les algorithmes d'agrégation autour des centres mobiles et de la **CAH**.

[126] Apparicio Philippe, Riva Mylène et Séguin Anne-Marie, « A comparison of two methods for classifying trajectories: a case study on neighborhood poverty at the intra-metropolitan level in Montreal », *Cybergeo*, , 2015, [https//doi.org/10.4000/cybergeo.27035].
[127] Richer Cyprien et Palmier Patrick, « Mesurer l'accessibilité territoriale par les transports collectifs. Proposition méthodologique appliquée aux pôles d'excellence de Lille Métropole », *Cahiers de géographie du Québec*, n° 158, vol. 56, 2012, p. 31.

Plus généralement, afin de réaliser des classifications, et de synthétiser un jeu de données, ou d'obtenir des profils de participants, il existe différentes méthodes de classification en plus des k-means. Nous avons par exemple déjà cité les classifications de types hiérarchiques telles que les **CA**H. D'autres méthodes impliquant un pré-supposé théorique, et ne sont donc pas exploratoires, contrairement aux méthodes de classification traditionnelles telles que les k-means ou les **CAH** ont le vent en poupe dans certaines disciplines. C'est par exemple le cas des **analyses en classes latentes (ou en profils latents)** qui sont de plus en plus utilisées en Psychologie notamment.

Ressources :

- Apparicio Philippe, Riva Mylène et Séguin Anne-Marie, « A comparison of two methods for classifying trajectories: a case study on neighborhood poverty at the intra-metropolitan level in Montreal », *Cybergeo*, , 2015, [https//doi.org/10.4000/cybergeo.27035].
- Jain Anil K. et Dubes Richard C., *Algorithms for clustering data*, Englewood Cliffs, N.J, Prentice Hall, coll. « Prentice Hall advanced reference series », 1988.
- Kaufman L., Rousseeuw P. J., Mathematics Faculty of et Informatics (Delft), *Clustering by Means of Medoids*, Faculty of Mathematics and Informatics, coll. « Delft University of Technology : reports of the Faculty of Technical Mathematics and Informatics », 1987.
- Lebart Ludovic, Piron Marie et Morineau Alain, *Statistique exploratoire multidimensionnelle: Visualisation et inférences en fouille de données*, 4e éd (2006)., Paris, Dunod, coll. « Sciences SUP », 1995.
- MacQueen J., « Some methods for classification and analysis of multivariate observations », *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, University of California Press, 1967, p. 281-298.
- Richer Cyprien et Palmier Patrick, « Mesurer l'accessibilité territoriale par les transports collectifs. Proposition méthodologique appliquée aux pôles d'excellence de Lille Métropole », *Cahiers de géographie du Québec*, nº 158, vol. 56, 2012, p. 31.
- Wong M. Anthony, « A Hybrid Clustering Method for Identifying High-Density Clusters », *Journal of the American Statistical Association*, nº 380, vol. 77, 1982, p. 841-847, [https//doi.org/10.1080/01621459.1982.10477896].

KMO n.m. [/ka.ɛm.o/]

Synonymes : Measure of sampling adequation, MSA

A quoi ça sert ?

Le test K-M-O (pour Kaiser-Meyer-Olkin) permet de vérifier l'adéquation de l'échantillon sélectionné au modèle factoriel testé. Il va mesurer la part de variance partagée par les différentes variables observées. Si celle-ci est suffisante alors les données pourront être factorisées dans un nombre restreint de facteurs. En revanche, si les variables observées partagent très peu de variance entre elles, alors l'analyse factorielle va nécessiter de considérer un nombre important de facteurs et ne synthétisera pas l'information de façon parcimonieuse. Dans ce dernier cas **l'analyse factorielle** ne sera pas adaptée pour analyser ces données. Afin de faire ce choix et d'évaluer la qualité de l'adéquation des variables sélectionnées à l'échantillon recueilli, Kaiser en 1974 [128] a décrit les différents seuils de l'indice KMO, compris entre 0 et 1 :

- 0.90 et plus : Merveilleux
- 0.80 et plus : Méritoire
- 0.70 et plus : Bien
- 0.60 et plus : Médiocre
- 0.50 et plus : Misérable
- Moins de 0.50 : Inacceptable

D'où ça vient ?

Le test K-M-O (pour Kaiser-Meyer-Olkin) a été développé par Kaiser en 1970, [129] et amélioré par Kaiser et Rice en 1974 [130].

Le mot du praticien

Le KMO n'est pas utilisé dans le cadre d'une AFC ou d'une ACM car il ne s'applique que sur des données quantitative. Le KMO est généralement employé avant la réalisation d'une **Analyse Factorielle Exploratoire** ou d'une **ACP**. Etchepare et al. 2018 présentent le résultat du KMO, sur leurs données, en amont de la réalisation d'une **analyse factorielle exploratoire** visant à valider un outil psychométrique permettant d'évaluer la cognition sociale des personnes en population générale et des patients atteints de schizophténie[131].

---

[128] Kaiser Henry F., « An Index of Factorial Simplicity », *Psychometrika*, n° 1, vol. 39, 1974, p. 31-36, [https//doi.org/10.1007/BF02291575].

[129] Kaiser Henry F., « A Second Generation Little Jiffy », *Psychometrika*, n° 4, vol. 35, 1970, p. 401-415, [https//doi.org/10.1007/BF02291817].

[130] Kaiser Henry F. et Rice John, « Little Jiffy, Mark Iv », *Educational and Psychological Measurement*, n° 1, vol. 34, 1974, p. 111-117, [https//doi.org/10.1177/001316447403400115].

[131] Etchepare Aurore, Roux Solenne, Destaillats Jean-Marc, Cady Florian, Fontanier David, Couhet Geoffroy et Prouteau Antoinette, « Éléments de validation du Protocole d'Évaluation de la Cognition Sociale de Bordeaux (PECS-B) en population générale et dans la schizophrénie », *Annales Médico-psychologiques, revue psychiatrique*, n° 2, vol. 178, 2020, p. 130-136, [https//doi.org/10.1016/j.amp.2018.06.011].

Ressources

- Etchepare Aurore, Roux Solenne, Destaillats Jean-Marc, Cady Florian, Fontanier David, Couhet Geoffroy et Prouteau Antoinette, « Éléments de validation du Protocole d'Évaluation de la Cognition Sociale de Bordeaux (PECS-B) en population générale et dans la schizophrénie », *Annales Médico-psychologiques, revue psychiatrique*, nº 2, vol. 178, 2020, p. 130-136, [https//doi.org/10.1016/j.amp.2018.06.011].
- Kaiser Henry F., « An Index of Factorial Simplicity », *Psychometrika*, nº 1, vol. 39, 1974, p. 31-36, [https//doi.org/10.1007/BF02291575].
- Kaiser Henry F., « A Second Generation Little Jiffy », *Psychometrika*, nº 4, vol. 35, 1970, p. 401-415, [https//doi.org/10.1007/BF02291817].
- Kaiser Henry F. et Rice John, « Little Jiffy, Mark Iv », *Educational and Psychological Measurement*, nº 1, vol. 34, 1974, p. 111-117, [https//doi.org/10.1177/001316447403400115].

Kolmogorov-Smirnov n.m [/ˌkɒlməˈgɒrɒf ˈsmɪrˌnɒvk/]

Synonymes : K-S

A quoi ça sert ?

Afin d'appliquer différents tests statistiques, il est nécessaire d'avoir des informations sur la distribution de nos données. Pour ce faire il est pertinent de vérifier l'adéquation de la distribution des données de notre échantillon à une loi de probabilité déterminée, ce que permet le test de **Kolmogorov-Smirnov (K-S)**[132]. Il peut, par exemple, être employé pour vérifier l'adéquation de la distribution de notre échantillon avec une *distribution normale*. Ce test *non-paramétrique* permet également de comparer deux échantillons afin de déterminer s'ils proviennent d'une même distribution, sous une même *loi de probabilité*. [133].

D'où ça vient ?

Le test de Kolmogorov-Smirnov a été conçu par Andrey Kolmogorov en 1933, puis développé aux comparaisons à deux échantillons par Nikolai Smirnov en 1939.

Mot du praticien

Le Kolmogorov-Smirnov est probablement le test le plus connu pour vérifier la normalité d'une distribution, même s'il ne s'applique pas seulement à la vérification de l'adéquation à une *loi normale*. Il tient compte de la moyenne et de la variance des données. Le test d'Anderson-Darling est une bonne alternative au Kolmogorov-Smirnov. Ce test en est une variante qui donne plus d'importance aux queues de distribution. De ce point de vue, il est plus indiqué dans la phase d'évaluation des données précédant la mise en œuvre d'un test paramétrique (comparaison de moyennes, de variances, etc.).

Ressources

- Kolmogorov A., « Sulla determinazione empirica di una legge di distribuzione », *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, 1933, p. 83-91.
- Smirnov N., « Table for Estimating the Goodness of Fit of Empirical Distributions », *The Annals of Mathematical Statistics*, n° 2, vol. 19, 1948, p. 279-281, [https//doi.org/10.1214/aoms/1177730256].

[132] Kolmogorov A., « Sulla determinazione empirica di una legge di distribuzione », *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, 1933, p. 83-91.

[133] Smirnov N., « Table for Estimating the Goodness of Fit of Empirical Distributions », *The Annals of Mathematical Statistics*, n° 2, vol. 19, 1948, p. 279-281, [https//doi.org/10.1214/aoms/1177730256].

Kruskal-Wallis n.m [/kʁus.kal va.lis/]

Synonymes : ANOVA de Kruskal-Wallis, ANOVA unidirectionnelle sur rangs

A quoi ça sert ?

Le test de Kruskall-Wallis est une alternative *non-paramétrique* au test de l'**ANOVA**. Cela veut dire que son objectif de base va être le même que celui de l'ANOVA, c'est-à-dire comparer entre eux plusieurs groupes (avec un minimum de 2 groupes), toutefois il ne reposera pas sur la moyenne mais sur les rangs.

L'**ANOVA** va être considérée comme plus « puissante » lorsque ses conditions d'applications sont réunies, mais si ces dernières ne sont pas respectées, alors il sera possible d'opter pour le test de Kruskall-Wallis, typiquement en cas de données ordinales, de petits échantillons ou lorsque l'hypothèse de normalité n'est pas respectée. Par ailleurs, le test de Kruskal-Wallis est une généralisation du test de **Mann et Whitney**.

D'où ça vient ?

Le test de Kruskal-Wallis a été proposé presque 30 ans après l'**ANOVA**, publié en 1952 par William Henry Kruskal et Wilson Allen Wallis[134]. Le test de Kruskal-Wallis vient donc combler un manque dans l'analyse statistique non paramétrique et constituer une alternative à l'**ANOVA**.

Pré-requis

Ce test nécessite que la variable comparée entre les différents groupes soit *quantitative*.

Mot du praticien

Ce test est tout-à-fait approprié si l'idée est de comparer un score (comme la performance à un test de mathématiques) entre différents groupes (composés d'élèves ayant reçus des consignes différentes comme par exemple : le groupe A aurait reçu l'instruction que le test était très difficile, le groupe B que le test était très facile et le groupe C aucune consigne particulière).

Ressources

- Kruskal William H. et Wallis W. Allen, « Use of Ranks in One-Criterion Variance Analysis », *Journal of the American Statistical Association*, n° 260, vol. 47, 1952, p. 583-621, [https//doi.org/10.1080/01621459.1952.10483441].

[134] Kruskal William H. et Wallis W. Allen, « Use of Ranks in One-Criterion Variance Analysis », *Journal of the American Statistical Association*, n° 260, vol. 47, 1952, p. 583-621, [https//doi.org/10.1080/01621459.1952.10483441].

LCA n.f. [/ɛl.se.ɑ/]

Synonymes : Latent class analysis, Analyses en classes latentes

A quoi ça sert ?

Les LCA ou Latent Class Analysis constituent un type de classification des observations intégrant des *variables latentes*. Ces classifications sont obtenues sur la base de différents critères rassemblés dans des variables catégorielles. Les LCA permettent ainsi de faire émerger des classes latentes sur des variables catégorielles.

Il est possible de réaliser des analyses de classification avec des variables latentes basées sur des variables continues, mais dans ce cas il faudra employer des **LPA - ou Latent Profil Analysis**.
Le fait que les classes obtenues soient des classes latentes, impliquent l'hypothèses que celles-ci viennent influencer les variables qui les constituent sans être appréhendables directement. Elles nécessitent la définition d'hypothèses a priori permettant d'envisager leur existance. Tout comme dans le cas des **LPA** les analyses en classes latentes nécessitent la structuration en amont des classes envisagées. Les LCA permettront de valider ou non l'hypothèse pré-établie.

D'où ça vient ?

Les LCA ont été introduites par Lazarsfeld en 1950[135] afin d'expliquer des phénomènes sociaux par la subdivision de la population en sous-groupes alors appelés classes latentes.

Mot du praticien

Par exemple, Chen et Yeung en 2025 ont publié un article sur les typologies de résliences familiales durant la crise liée au COVID-19[136]. Les typologies ont été obtenue suite à la réalisation d'Analyses en Classes Latentes. Par ailleurs, il est à noter qu'il est tout-à-fait possible de réaliser des modèles intégrant des covariables ayant un effet sur la classification. Pour en savoir plus, il existe différentes ressources sur le sujet, tels que les articles de (Weller, B. E., Bowen, N. K., & Faubert, S. J., 2020), (Sinha, P., Calfee, C. S., & Delucchi, K. L., 2021) ou encore le livre de Eshima N., (2022).
Afin d'approfondir ces classifications il est généralement intéressant de réaliser des modèles de **régressions** ou des **modèles en équation structurelles**.

Ressources

- Chen Xuejiao et Yeung Wei-Jun Jean, « COVID -19 experiences and family resilience: A latent class analysis », *Journal of Marriage and Family*, nº 1, vol. 87, 2025, p. 280-299, [https//doi.org/10.1111/jomf.13031].
- Eshima Nobuoki, *An Introduction to Latent Class Analysis: Methods and Applications*, Singapore, Springer Singapore, coll. « Behaviormetrics: Quantitative Approaches to Human Behavior », 2022, [https//doi.org/10.1007/978-981-19-0972-6].

[135] Lazarsfeld Paul F., « The logical and mathematical foundation of latent structure analysis », *Measurement and prediction. [Studies in social psychology in World War II. Vol.4.]*, Princeton, NJ, US, Princeton University Press, coll. « Measurement and prediction », 1950, .

[136] Chen Xuejiao et Yeung Wei-Jun Jean, « COVID -19 experiences and family resilience: A latent class analysis », *Journal of Marriage and Family*, nº 1, vol. 87, 2025, p. 280-299, [https//doi.org/10.1111/jomf.13031].

- Hickendorff Marian, Edelsbrunner Peter A., McMullen Jake, Schneider Michael et Trezise Kelly, « Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis », *Learning and Individual Differences*, vol. 66, 2018, p. 4-15, [https//doi.org/10.1016/j.lindif.2017.11.001].
- Lazarsfeld Paul F., « The logical and mathematical foundation of latent structure analysis », *Measurement and prediction. [Studies in social psychology in World War II. Vol.4.]*, Princeton, NJ, US, Princeton University Press, coll. « Measurement and prediction », 1950, .
- Sinha Pratik, Calfee Carolyn S. et Delucchi Kevin L., « Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls », *Critical Care Medicine*, nº 1, vol. 49, 2021, p. e63-e79, [https//doi.org/10.1097/CCM.0000000000004710].
- Weller Bridget E., Bowen Natasha K. et Faubert Sarah J., « Latent Class Analysis: A Guide to Best Practice », *Journal of Black Psychology*, nº 4, vol. 46, 2020, p. 287-311, [https//doi.org/10.1177/0095798420930932].

Lexicométrie n.f. [/lɛk.si.ko.me.tʁi/]

Appelée d'abord lexicométrie, la discipline a vu son nom évoluer en **textométrie** ou logométrie au début des années 2000. Pour en savoir plus sur le sujet voir l'entrée **textométrie**.

Linéarité n.f. [/lineaʁite/]

Synonymes : Relation linéaire

A quoi ça sert ?

La linéarité est l'une des façons de rendre compte de la relation entre deux ou plusieurs variables. La relation linéaire est pré-supposée dans différentes méthodes d'analyse des données telles que par exemple les **corrélations** ou encore les **régressions linéaires**. Il est donc important de vérifier la relation entre les variables du modèle d'analyse avant sa réalisation, car selon le résultat obtenu le choix de la méthode d'analyse sera différente. Par exemple, si avant de réaliser une analyse de régression linéaire, nous observons que la la relation entre les variables n'est pas linéaire mais plutôt quadratique, alors il sera nécessaire de se tourner vers un autre type d'analyse telles que les **régressions polynomiales** par exemple.

D'où ça vient ?

La conception de linéarité a été développée à la fin du XIXe et au début du XXe siècle par Francis Galton et Karl Pearson notamment dans le cadre de corrélations et de régression linéaires[137].

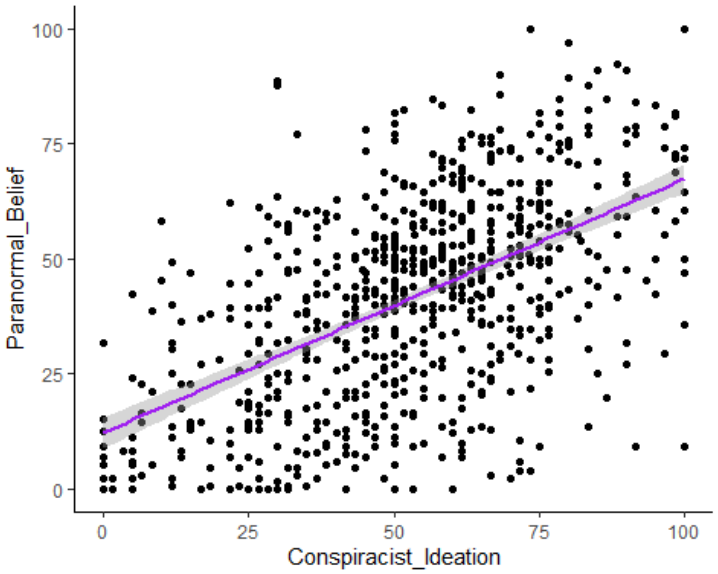
Mot du praticien

Il existe différentes façons de tester l'hypothèse de linéarité entre des variables. Il est par exemple possible de faire appel au **Rainbow Test** développé par Jessica Utts en 1982 [138]. La linéarité peut également faire l'objet d'observations graphiques permettant de vérifier si la droite modélisant la relation linéaire se situe bien au centre du nuage de points modélisant la répartition des observations sélectionnées pour le modèle.

Par exemple, le graphique ci-dessous représente les réponses des participants (chaque point noir) sur les leurs croyances aux phénomènes paranormaux et leur adhésion aux théories du complot. La ligne violette indique la relation linéaire entre ces deux variables. Celle-ci passant au milieu de l'ensemble des points, elle semble donc adaptée pour rendre compte de la relation entre les deux éléments étudiés. Cet aspect graphique peut être approfondi par la comparaison avec une autre forme d'ajustement (quadratique par exemple - voir l'entrée sur les **régressions polynomiales**), l'objectif sera de sélectionner l'ajustement qui correpond le plus à la répartition des observations.

[137] Stanton Jeffrey M., « Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors », *Journal of Statistics Education*, nº 3, vol. 9, 2001, p. 3, [https//doi.org/10.1080/10691898.2001.11910537].

[138] Utts Jessica M., « The rainbow test for lack of fit in regression », *Communications in Statistics - Theory and Methods*, nº 24, vol. 11, 1982, p. 2801-2815, [https//doi.org/10.1080/03610928208828423].



Représentation graphique de la relation entre les croyances dans les phénomènes paranormaux et l'adhésion aux théories du complot, données issues des travaux de Pennycock et al. 2020

La question de la linéarité est essentielle mais régulièrement négligée. C'est sur elle que repose un certain nombre de tests staistiques, or la vérification de la linéarité est de fait une condition préalable à leur utilisation. Il existe en statistique de nombreuses alternatives de tests et de méthodes ne reposant pas nécessairement sur la linéarité.

Ressources

- Nahhas Ramzi W., *5.17 Checking the linearity assumption | Introduction to Regression Methods for Public Health Using R*, 2025.
- Pennycook Gordon, Cheyne James Allan, Koehler Derek J. et Fugelsang Jonathan A., « On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs », *Judgment and Decision Making*, nº 4, vol. 15, 2020, p. 476-498, [https//doi.org/10.1017/S1930297500007439].
- Stanton Jeffrey M., « Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors », *Journal of Statistics Education*, nº 3, vol. 9, 2001, p. 3, [https//doi.org/10.1080/10691898.2001.11910537].
- Tabachnick Barbara et Fidell Linda, *Using Multivariate Statistics*, Pearson International, 2013.
- Utts Jessica M., « The rainbow test for lack of fit in regression », *Communications in Statistics - Theory and Methods*, nº 24, vol. 11, 1982, p. 2801-2815, [https//doi.org/10.1080/03610928208828423].

LPA n.f. [/ɛl.pe.ɑ/]

Synonymes : LPA, Latent Profil Analysis, Analyse en profils latents

A quoi ça sert ?

Les LPA ou Latent Profil Analysis sont une forme de classification des observations basée sur différents critères permettant d'obtenir des profils. Elles font parti de la famille des **modèles en équation structurelles** car elles intègrent la notion de *variable latente*, constitutive des profils obtenus à l'issue de l'analyse. Elles permettent donc de réaliser une classification sur des variables continues qui fera émerger des profils latents.
L'analyse en profils latents repose sur une démarche nécessitant la formulation d'hypothèses a priori. Afin de réaliser une classification en classe latente à partir de données catégorielles, il sera nécessaire de réaliser des **LCA** qui fera l'objet d'une autre entrée. Les profils obtenus à partir de LPA (ou de **LCA**) constituent une variable catégorielle pouvant faire l'objet d'analyses ultérieures.

D'où ça vient ?

Les analyses en profils latents ont été proposés par Gibson en 1962[139], dans la continuité des travaux de Lazarsfled[140] sur le sujet, donc afin de définir des profils mais à partir de variables continues.

Exemple d'application

Les LPA peuvent par exemple être employées afin de mettre en évidence des profils de comportements et d'attitudes pendant la pandémie de COVID-19, tels que l'ont réalisé Kleitman et al. en 2021 auprès d'un échantillon de participants issus de quatre pays (Australie, Canada, Etats-Unis et Royaume-Uni)[141]. Ces analyses peuvent intégrer des approches plus complexes, telles que les analyses en profils latents multiniveaux. Elles sont notamment utilisées en Sciences de l'Education, comme par exemple Chen et al. en 2023 qui ont mobilisé des LPA multiniveaux afin d'établir des profils latents d'apprentissage autorégulé et d'affect positif des adolescents, auprès d'élèves de différentes écoles, implantées dans différentes régions d'Asie[142]. Il existe également la possibilité d'intégrer une dimension à ces analyses en profils latents (tout comme en **classes latentes**) via les analyse de transition de profil latent (LPTA). Les travaux de Yé et al. en 2021

[139] Gibson W. A., « Extending Latent Class Solutions to Other Variables », *Psychometrika*, n° 1, vol. 27, 1962, p. 73-81, [https//doi.org/10.1007/BF02289666].
[140] Lazarsfeld Paul F., « The logical and mathematical foundation of latent structure analysis », *Measurement and prediction. [Studies in social psychology in World War II. Vol.4.]*, Princeton, NJ, US, Princeton University Press, coll. « Measurement and prediction », 1950, .
[141] Kleitman Sabina, Fullerton Dayna J., Zhang Lisa M., Blanchard Matthew D., Lee Jihyun, Stankov Lazar et Thompson Valerie, « To comply or not comply? A latent profile analysis of behaviours and attitudes during the COVID-19 pandemic », *PLOS ONE*, n° 7, vol. 16, 2021, p. e0255268, [https//doi.org/10.1371/journal.pone.0255268].
[142] Chen Jiangping, Lin Chin-Hsi et Chen Gaowei, « Adolescents' self-regulated and affective learning, teacher support and digital reading literacy: A multilevel latent profile approach », *Computers & Education*, vol. 205, 2023, p. 104883, [https//doi.org/10.1016/j.compedu.2023.104883].

présentent par exemple des profils de résilience face au cancer du sein et les transition entre ces différents profils établis au temps 1 sur les autres temps de mesure[143].

Mot du praticien

Bien qu'il s'agisse d'une démarche hypothético-déductive, il est possible de sélectionner un panel de profil et de choisir celui le plus adapté à nos données sur des critères statistiques (indices statistiques tels que l'entropy, l'AIC, le BIC, etc.). Toutefois, il est il important de ne pas oublier l'intérêt de ce type d'analyse impliquant des variables latentes qui influencent la répartition de nos observations en différents groupes (invariablement appelés classes ou profils). Si l'objectif recherché est une classification purement exploratoire alors les analyses en profils latents ne seront pas adapté et il sera nécessaire de se tourner vers d'autres type de classifications telles que des **CAH** ou des **K-means**.

Enfin, comme dans toutes les analyses multivariées, les LPA nécessitent que les *variables manifestes*, employées pour constituer les profils, soit indépendantes.[144]

Ressources

- Chen Jiangping, Lin Chin-Hsi et Chen Gaowei, « Adolescents' self-regulated and affective learning, teacher support and digital reading literacy: A multilevel latent profile approach », *Computers & Education*, vol. 205, 2023, p. 104883, [https//doi.org/10.1016/j.compedu.2023.104883].
- Eshima Nobuoki, *An Introduction to Latent Class Analysis: Methods and Applications*, Singapore, Springer Singapore, coll. « Behaviormetrics: Quantitative Approaches to Human Behavior », 2022, [https//doi.org/10.1007/978-981-19-0972-6].
- Gibson W. A., « Extending Latent Class Solutions to Other Variables », *Psychometrika*, nº 1, vol. 27, 1962, p. 73-81, [https//doi.org/10.1007/BF02289666].
- Kleitman Sabina, Fullerton Dayna J., Zhang Lisa M., Blanchard Matthew D., Lee Jihyun, Stankov Lazar et Thompson Valerie, « To comply or not comply? A latent profile analysis of behaviours and attitudes during the COVID-19 pandemic », *PLOS ONE*, nº 7, vol. 16, 2021, p. e0255268, [https//doi.org/10.1371/journal.pone.0255268].
- Lazarsfeld Paul F., « The logical and mathematical foundation of latent structure analysis », *Measurement and prediction. [Studies in social psychology in World War II. Vol.4.]*, Princeton, NJ, US, Princeton University Press, coll. « Measurement and prediction », 1950, .
- Ye Zeng Jie, Zhang Zhang, Tang Ying, Liang Jian, Sun Zhe, Hu Guang Yun, Liang Mu Zi et Yu Yuan Liang, « Resilience patterns and transitions in the Be Resilient To Breast Cancer trial: an exploratory latent profile transition analysis », *Psycho-Oncology*, nº 6, vol. 30, 2021, p. 901-909, [https//doi.org/10.1002/pon.5668].

Pour réaliser des LPA sous R, il existe le package tidyLPA :

- Rosenberg Joshua, Beymer Patrick, Anderson Daniel, Van Lissa C. j. et Schmidt Jennifer, « tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-

[143] Ye Zeng Jie, Zhang Zhang, Tang Ying, Liang Jian, Sun Zhe, Hu Guang Yun, Liang Mu Zi et Yu Yuan Liang, « Resilience patterns and transitions in the Be Resilient To Breast Cancer trial: an exploratory latent profile transition analysis », *Psycho-Oncology*, nº 6, vol. 30, 2021, p. 901-909, [https//doi.org/10.1002/pon.5668].

[144] Eshima Nobuoki, *An Introduction to Latent Class Analysis: Methods and Applications*, Singapore, Springer Singapore, coll. « Behaviormetrics: Quantitative Approaches to Human Behavior », 2022, [https//doi.org/10.1007/978-981-19-0972-6].

Source or Commercial Software », *Journal of Open Source Software*, nº 30, vol. 3, 2018, p. 978, [https//doi.org/10.21105/joss.00978].

- Rosenberg Joshua M. et van Lissa Caspar, *Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software*, [https://data-edu.github.io/tidyLPA/].

Mann-Whitney n.m [/man wit.nɛ/]

Synonymes : U de Mann-Whitney, U test, test de somme de rangs de Wilcoxon

A quoi ça sert ?

Le test de Mann et Whitney est une alternative *non-paramétrique* au test du **T de Student**, tout comme le test de **Wilcoxon**. Son objectif est donc de comparer deux échantillons, et sera utilisé lorsque les conditions d'application du **t de Student** ne sont pas remplies et ne permettent pas son application.

Le test de Mann-Whitney va être employé pour comparer 2 échantillons indépendants alors que le test de **Wilcoxon** a été développé pour comparerdes échantillons qu'ils soient indépendants ou appariés.

Il s'agit de deux tests différents, développés par des auteurs différents, néanmoins ces deux tests ont une relation linéaire, la connaissance du résultat de l'un permet d'obtenir le résultat de l'autre. A tel point que la littérature fait mention du test de Mann-Whitney-Wilcoxon ou du Wilcoxon-Mann-Whitney.

D'où ça vient ?

Le test de Mann-Whitney est né en en 1947 suite aux travaux de Henry Berthold Mann et Donald Ransom Whitney[145]. Néanmoins, historiquement le Wilcoxon est légèrement plus ancien car publié en 1945 par Franck Wilcoxon[146]. Et si c'est Henry Berthold Mann et Donald Ransom Whitney qui ont formalisé ce test, c'est Frank Wilcoxon qui va initialement le proposer en 1945 sous le nom de rank-sum test.

Mot du praticien

Le test de Mann-Whitney peut par exemple être employé afin de comparer le salaire des hommes et des femmes dans un secteur d'activité particulier, où la distribution des salaires par groupe ne suivrait pas une *loi normale* ou que la répartition de la variance ne serait pas en situation d'**homoscédasticité**. Il peut aussi être employé afin de comparer le niveau de stress de patients ayant bénéficiés d'une psychothérapie et ceux n'en ayant pas bénéficié (toujours si la distribution de la variable stress par groupe ne se distribue pas normalement ou que la répartition de la variance n'est pas homogène). Ou encore dans les travaux d'Escolà-Gascón et al. 2022 afin de comparer les groupes de participants qui détectent les fakes news et ceux qui ne les détectent pas sur différents aspects psychologiques tels que l'anxiété, la paranoïa et la schizotypie[147].

---

[145] Mann H. B. et Whitney D. R., « On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other », *The Annals of Mathematical Statistics*, n° 1, vol. 18, 1947, p. 50-60, [https//doi.org/10.1214/aoms/1177730491].

[146] Wilcoxon Frank, « Individual Comparisons by Ranking Methods », *Biometrics Bulletin*, n° 6, vol. 1, 1945, p. 80, [https//doi.org/10.2307/3001968].

[147] Escolà-Gascón Álex, Dagnall Neil, Denovan Andrew, Drinkwater Kenneth et Diez-Bosch Miriam, « Who falls for fake news? Psychological and clinical profiling evidence of fake news consumers », *Personality and Individual Differences*, vol. 200, 2023, p. 111893, [https//doi.org/10.1016/j.paid.2022.111893].

Ressources

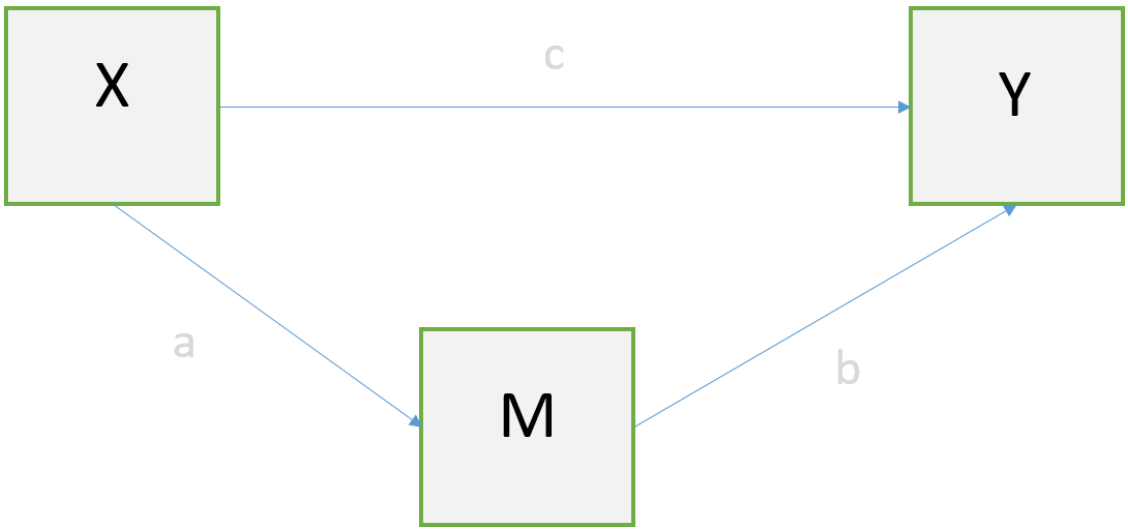
- Escolà-Gascón Álex, Dagnall Neil, Denovan Andrew, Drinkwater Kenneth et Diez-Bosch Miriam, « Who falls for fake news? Psychological and clinical profiling evidence of fake news consumers », *Personality and Individual Differences*, vol. 200, 2023, p. 111893, [https//doi.org/10.1016/j.paid.2022.111893].
- Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.
- Mann H. B. et Whitney D. R., « On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other », *The Annals of Mathematical Statistics*, n° 1, vol. 18, 1947, p. 50-60, [https//doi.org/10.1214/aoms/1177730491].
- Wilcoxon Frank, « Individual Comparisons by Ranking Methods », *Biometrics Bulletin*, n° 6, vol. 1, 1945, p. 80, [https//doi.org/10.2307/3001968].

Médiation n.f. [/me.dja.sjõ/]

Synonymes : Relations médiées, Modèles de médiation

A quoi ça sert ?

Cette entrée est consacrée aux modèles de médiation via des modèles de régression traditionnels. Il existe une autre façon de réaliser des modèles de médiation, en faisant appel aux **modèles en équation structurelles (SEM)**, qui fera l'objet d'une autre entrée.
Les modèles de médiation ont été développés afin d'intégrer une variable considérée comme médiatrice dans une relation entre une variable explicative et une variable à expliquer. La relation entre la *variable explicative* et la *variable à expliquer* serait alors médiée partiellement ou totalement par une (ou plusieurs) autre variable (comme présenté dans le schéma ci-dessous). c représente le lien direct entre la *variable explicative* et la *variable à expliquer*, b le lien entre le médiateur et la *variable à expliquer* et a celui entre la *variable explicative* et le médiateur.



Modèle de médiation simple à 1 médiateur

D'où ça vient ?

Les médiations ont été présentées par Baron et Kenny en 1986 afin de rappeler et de définir l'importance de la différence entre le terme de **modération** et celui de médiation (Baron et Kenny, 1986)[148]. Si les analyses de modération sont relativement récentes (années 1980), les modèles de médiation sont plus anciens car leurs origines remontent au début du XXème siècle, notamment avec les travaux du psychologue expérimental Robert S. Woodworth en 1928[149]. Afin d'améliorer notre compréhension du comportement humain, Woodworth a introduit la notion de médiation en complétant le modèle développé par Pavlov[150] de Stimulus - Réponse, par un modèle Stimulus - Organisme - Réponse (S-O-R), avec l'Organisme comme médiateur de la relation

[148] Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, nº 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].
[149] Woodworth R. S., « How emotions are identified and classified », *Feelings and emotions: The Wittenberg Symposium*, Oxford, England, Clark Univ. Press, 1928, p. 222-227.
[150] Pavlov Ivan P., « The Scientific Investigation of the Psychical Faculties or Processes in the Higher Animals », *Science*, nº 620, vol. 24, 1906, p. 613-619, [https//doi.org/10.1126/science.24.620.613].

Stimulus/Réponse. Les analyses de médiation sont toujours très employées aujourd'hui et se sont complexifiées en incluant notamment la possibilité de médiateurs multiples. Des modèles plus complexes encore ont été pensé et développé, tels que des modérations médiées ou des médiations modérées. Ces modèles ont été développés dans une approche basée sur des régressions, tels que les travaux de Preacher et Hayes sur le sujet[151], mais également dans le cadre de modèles structuraux permettant de traiter cette complexité.

Mot du praticien

Les pré-requis aux analyses de médiations sont les mêmes que celles nécessaires à la réalisation des **modèles de régression**. Il est à noter qu'une médiation peut être réalisée dans un modèle de régression quel qu'il soit : logistique, linéaire, polynomial, etc.

Les modèles de médiation sont très utiles afin de rendre compte de situations fréquemment rencontrées en SHS. Par exemple ils peuvent être employés afin d'appréhender la différence de salaire, ou d'évolution de carrière, entre les hommes et les femmes en intégrant le temps de travail comme médiateur de cette différence. Les femmes étant plus fréquemment à temps partiel que les hommes[152]. Ou encore en Psychologie, Sun et al. ont employé ce type de modèle dans un article paru en 2023 afin de rendre compte du rôle médiateur de la résilience sur la relation entre le soutien social et l'anxiété / la dépression chez des patients porteurs du VIH ou du SIDA, en Chine[153]. Les auteurs ont pu tester leur modèle théoriques grâce au modèle de médiation dégagé suite à différents **modèles de régression**.

Il est à noter qu'une médiation peut être une médiation partielle, dans le cas où la relation entre les variables explicative et la variable à expliquer ne serait pas totalement médiée par la ou les variables médiatrices. Si ces dernières viennent médier totalement la relation entre les variables explicatives et la variable à expliquer alors il s'agira d'une médiation dite totale.

Ressources

- Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, nº 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].
- INSEE, *Temps partiel − Emploi, chômage, revenus du travail*, [https://www.insee.fr/fr/statistiques/7456899?sommaire=7456956#consulter].
- Pavlov Ivan P., « The Scientific Investigation of the Psychical Faculties or Processes in the Higher Animals », *Science*, nº 620, vol. 24, 1906, p. 613-619, [https//doi.org/10.1126/science.24.620.613].

[151] Woodworth R. S., « How emotions are identified and classified », *Feelings and emotions: The Wittenberg Symposium*, Oxford, England, Clark Univ. Press, 1928, p. 222-227.

[152] INSEE, *Temps partiel − Emploi, chômage, revenus du travail*, [https://www.insee.fr/fr/statistiques/7456899?sommaire=7456956#consulter].

[153] Sun Yongbing, Song Bing, Zhen Cheng, Zhang Chao, Cheng Juan et Jiang Tianjun, « The mediating effect of psychological resilience between social support and anxiety/depression in people living with HIV/AIDS–a study from China », *BMC Public Health*, nº 1, vol. 23, 2023, p. 2461, [https//doi.org/10.1186/s12889-023-17403-y].

- Sun Yongbing, Song Bing, Zhen Cheng, Zhang Chao, Cheng Juan et Jiang Tianjun, « The mediating effect of psychological resilience between social support and anxiety/depression in people living with HIV/AIDS–a study from China », *BMC Public Health*, n° 1, vol. 23, 2023, p. 2461, [https//doi.org/10.1186/s12889-023-17403-y].
- Woodworth R. S., « How emotions are identified and classified », *Feelings and emotions: The Wittenberg Symposium*, Oxford, England, Clark Univ. Press, 1928, p. 222-227.

Médiation n.f. [/me.dja.sjõ/]

Synonymes : Relations médiées, Modèles de médiation

A quoi ça sert ?

Cette entrée est consacrée aux modèles de médiation via des modèles en équation structurelles (SEM). Il existe une autre façon de réaliser des modèles de médiation, en faisant appel aux **modèles de régression**, qui fera l'objet d'une autre entrée. Les analyses de médiation nécessitent de décomposer les effets directs (entre la VD et la VI), les effets indirects (en passant par la ou les variable(s) médiatrice(s)) et les effets totaux (impliquant l'ensemble du modèle).

D'où ça vient ?

Cette décomposition a été pensée dès l'origine des modèles en équations structurelles par Sewall Wright en 1934 et ses travaux sur les analyses de parcours[154]. Ce n'est que cinquante ans plus tard, dans les années 1980, que les modèles de médiations vont connaître une certaine popularité avec notamment les travaux de Baron & Kenny, 1986[155] sur les médiations dans le cadre des modèles de régression. En parallèle de ces travaux, des modèles de médiation via les SEM se développent avec notamment les travaux de Jöreskog & Sörbom, 1981[156] ou encore de Bollen, 1987[157]. Les modèles de médiations via les SEM se sont complexifiés afin d'intégrer des relations non-linéaires entre les variables[158].

Mot du praticien

Les pré-requis aux analyses de médiations sont les mêmes que celles nécessaires à la réalisation des **modèles SEM**.

Les modèles de médiation via des modèles SEM permettent de rendre de compte de relations de médiation complexes entre des variables, comme par exemple l'effet médiateur de la cognition sociale entre la neurocognition et le fonctionnement social chez des patients atteints de troubles dépressifs majeur ou de schizophrénie, tels que présenté dans les travaux de Uchino et al en 2023[159]. Cet article relate d'une médiation à un seul médiateur, toutefois la cognition sociale est

---

[154] Bollen Kenneth A., Fisher Zachary, Lilly Adam, Brehm Christopher, Luo Lan, Martinez Alejandro et Ye Ai, « Fifty years of structural equation modeling: A history of generalization, unification, and diffusion », *Social Science Research*, vol. 107, 2022, p. 102769, [https//doi.org/10.1016/j.ssresearch.2022.102769].

[155] Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, nº 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].

[156] Jöreskog K. G., *LISREL V : analysis of linear structural relationships by maximum likelihood and least squares methods*, Version V. Uppsala : University of Uppsala, Dept. of Statistics ; Chicago : distributed by International Educational Services, [1981] ©1981, 1981.

[157] Bollen Kenneth A., « Total, Direct, and Indirect Effects in Structural Equation Models », *Sociological Methodology*, vol. 17, 1987, p. 37, [https//doi.org/10.2307/271028].

[158] Bollen Kenneth A., Fisher Zachary, Lilly Adam, Brehm Christopher, Luo Lan, Martinez Alejandro et Ye Ai, « Fifty years of structural equation modeling: A history of generalization, unification, and diffusion », *Social Science Research*, vol. 107, 2022, p. 102769, [https//doi.org/10.1016/j.ssresearch.2022.102769].

[159] Uchino Takashi, Okubo Ryo, Takubo Youji, Aoki Akiko, Wada Izumi, Hashimoto Naoki, Ikezawa Satoru et Nemoto Takahiro, « Mediation Effects of Social Cognition on the Relationship between Neurocognition and

considérée comme une variable latente et l'approche par les SEM permet de rendre compte de ce construit.

Ressources

- Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, nº 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].
- Bollen Kenneth A., « Total, Direct, and Indirect Effects in Structural Equation Models », *Sociological Methodology*, vol. 17, 1987, p. 37, [https//doi.org/10.2307/271028].
- Bollen Kenneth A., Fisher Zachary, Lilly Adam, Brehm Christopher, Luo Lan, Martinez Alejandro et Ye Ai, « Fifty years of structural equation modeling: A history of generalization, unification, and diffusion », *Social Science Research*, vol. 107, 2022, p. 102769, [https//doi.org/10.1016/j.ssresearch.2022.102769].
- Jöreskog K. G., *LISREL V : analysis of linear structural relationships by maximum likelihood and least squares methods*, Version V. Uppsala : University of Uppsala, Dept. of Statistics ; Chicago : distributed by International Educational Services, [1981] ©1981, 1981.
- Uchino Takashi, Okubo Ryo, Takubo Youji, Aoki Akiko, Wada Izumi, Hashimoto Naoki, Ikezawa Satoru et Nemoto Takahiro, « Mediation Effects of Social Cognition on the Relationship between Neurocognition and Social Functioning in Major Depressive Disorder and Schizophrenia Spectrum Disorders », *Journal of Personalized Medicine*, nº 4, vol. 13, 2023, p. 683, [https//doi.org/10.3390/jpm13040683].

Social Functioning in Major Depressive Disorder and Schizophrenia Spectrum Disorders », *Journal of Personalized Medicine*, nº 4, vol. 13, 2023, p. 683, [https//doi.org/10.3390/jpm13040683].

Modèles mixtes n.m.[/mɔ.dɛl mixt/]

Synonymes : Modèles de régression, Modèles hierarchiques, Modèles multiniveaux

A quoi ça sert ?

Le terme modèles mixtes fait référence, dans le cadre de cette entrée, aux modèles de régression mélangeant des effets fixes et des effets aléatoires. Ces modèles peuvent s'appliquer sur toute forme de régression, donc quelque soit le type de variable à expliquer ou le lien entre les variables (**modèles linéaires**, **logistiques**, **polynomiaux**, etc.). Les effets fixes renvoient à des variables restant constantes au cours du temps, alors que les effets aléatoires impliquent différentes mesures pour une même observation et leurs valeurs sont susceptibles d'évoluer au cours du temps.

D'où ça vient ?

Le premier à conceptualiser la notion d'effets aléatoires est Ronald Fisher en 1919, afin de rendre compte de la corrélation entre des facteurs génétiques héréditaires au sein d'une même famille [160]. A la fin des années 1950[161], puis dans les années 1970[162], les travaux d'Henderson, toujours en biométrie, vont permettre une avancée dans ce champ des statistiques en proposant une approche permettant à des modèles de mixer des effets fixes et des effets aléatoires. Actuellement ce type de modèle est très utilisé dans tous les domaines, en biologie notamment mais également en sciences humaines et sociales.

Mot du praticien

Les modèles mixtes peuvent par exemple être utilisés en Psychologie dans des recherches à mesures répétées visant à évaluer les évolutions émotionnelles et les stratégies de régulation des émotions des participants. C'est notamment le cas des études EMA (Ecological Momentary Assessement). Dans ce type d'étude les participants doivent répondre plusieurs fois par jour pendant plusieurs jours à une batterie de questions telles que : "Quelle émotion ressentez-vous en ce moment ?" ; "Avec quelle personne vous trouvez-vous ?" par exemples. Afin d'isoler l'effet du contexte sur l'émotion ressentie ou sur la stratégie de régulation choisie il sera judicieux de réaliser un modèle mixte comprenant des effets fixes (l'émotion ressentie, le contexte) et des effets aléatoires (le moment de la journée, le jour de la semaine concerné). Les travaux d'Hedeker et al publiés en 2012 illustrent parfaitement cette démarche[163]. En effet, les auteurs ont réalisé une étude de type EMA auprès d'adolescents qu'ils ont analysé grâce à des modèles mixtes. L'objectif de cette étude était de saisir les variations de l'humeur chez les adolescents selon leur consommation de

[160] Fisher R. A., « XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. », *Transactions of the Royal Society of Edinburgh*, n° 2, vol. 52, 1919, p. 399-433, [https//doi.org/10.1017/S0080456800012163].

[161] Henderson C. R., Kempthorne Oscar, Searle S. R. et Von Krosigk C. M., « The Estimation of Environmental and Genetic Trends from Records Subject to Culling », *Biometrics*, n° 2, vol. 15, 1959, p. 192, [https//doi.org/10.2307/2527669].

[162] Henderson C. R., « Best Linear Unbiased Estimation and Prediction under a Selection Model », *Biometrics*, n° 2, vol. 31, 1975, p. 423, [https//doi.org/10.2307/2529430].

[163] Hedeker Donald, Mermelstein Robin J. et Demirtas Hakan, « Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models », *Statistics in Medicine*, n° 27, vol. 31, 2012, p. 3328-3336, [https//doi.org/10.1002/sim.5338].

tabac à 6, 9, 15, 24 et 33 mois d'écarts. Les modèles mixtes leur ont permis de tester leurs hypothèses et de traiter leurs données.
Ce type de modèle peut s'appliquer dans toutes les disciplines des SHS. Par exemple, en Science de l'Education, si nous souhaitons étudier la performance en mathématiques et en français d'élèves d'écoles élémentaires selon différents contextes. Les modèles mixtes, et plus précisément les modèles multiniveaux[164], [165], peuvent permettre d'isoler la performance en mathématiques des élèves selon une instruction de départ donnée (par exemple : le test à accomplir est facile, difficile ou sans consigne particulière), des différences entre les écoles, les classes, qui seront alors considérées comme des effets aléatoires. Ce cas d'emboîtement, un élève est dans une classe, qui est elle-même dans une école, qui est elle même dans une ville peut se traiter par modèles multiniveaux qui sont un cas particulier des modèles mixtes.

Ressources

- Bressoux Pascal, *Modélisation statistique appliquée aux sciences sociales*, De Boeck Supérieur, 2010, [https//doi.org/10.3917/dbu.bress.2010.01].
- Fisher R. A., « XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. », *Transactions of the Royal Society of Edinburgh*, nº 2, vol. 52, 1919, p. 399-433, [https//doi.org/10.1017/S0080456800012163].
- Hedeker Donald, Mermelstein Robin J. et Demirtas Hakan, « Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models », *Statistics in Medicine*, nº 27, vol. 31, 2012, p. 3328-3336, [https//doi.org/10.1002/sim.5338].
- Henderson C. R., « Best Linear Unbiased Estimation and Prediction under a Selection Model », *Biometrics*, nº 2, vol. 31, 1975, p. 423, [https//doi.org/10.2307/2529430].
- Henderson C. R., Kempthorne Oscar, Searle S. R. et Von Krosigk C. M., « The Estimation of Environmental and Genetic Trends from Records Subject to Culling », *Biometrics*, nº 2, vol. 15, 1959, p. 192, [https//doi.org/10.2307/2527669].
- Meteyard Lotte et Davies Robert A. I., « Best practice guidance for linear mixed-effects models in psychological science », *Journal of Memory and Language*, vol. 112, 2020, p. 104092, [https//doi.org/10.1016/j.jml.2020.104092].
- Sommet Nicolas et Morselli Davide, « Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS », *International Review of Social Psychology*, nº 1, vol. 30, 2017, p. 203-218, [https//doi.org/10.5334/irsp.90].
- Tabachnick Barbara et Fidell Linda, *Using Multivariate Statistics*, Pearson International, 2013.

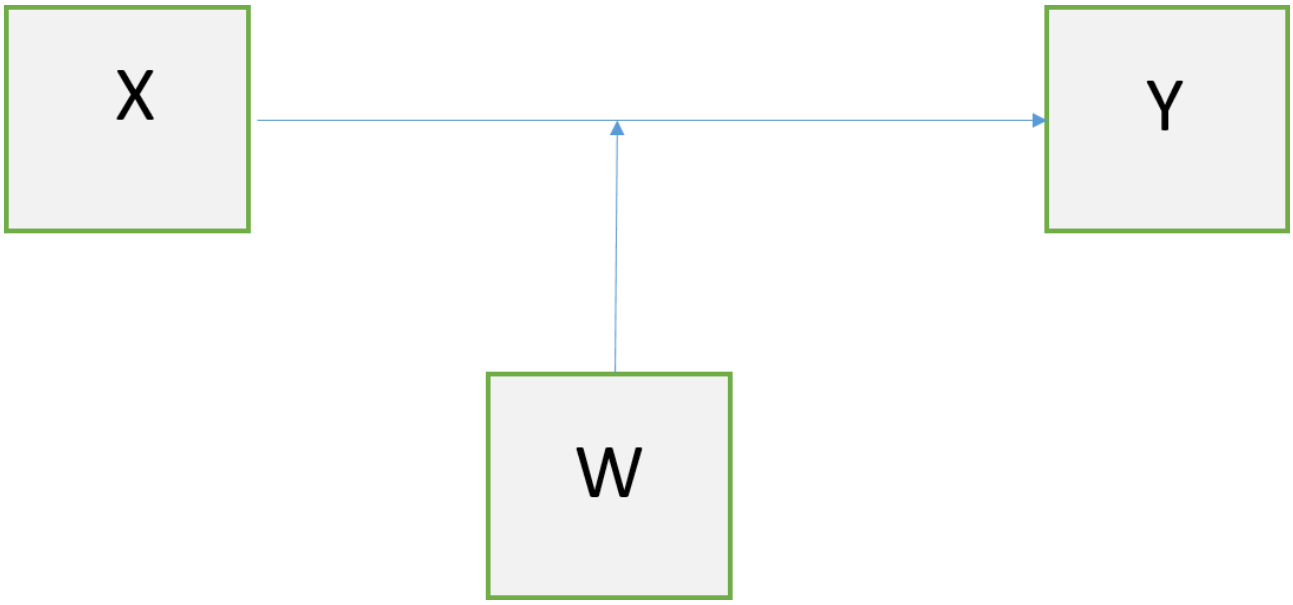
[164] Bressoux Pascal, *Modélisation statistique appliquée aux sciences sociales*, De Boeck Supérieur, 2010, [https//doi.org/10.3917/dbu.bress.2010.01].
[165] Tabachnick Barbara et Fidell Linda, *Using Multivariate Statistics*, Pearson International, 2013.

Modération n.f. [/mɔ.de.ʁa.sjɔ̃ /]

Synonymes : Modèles de modération, Interactions

A quoi ça sert ?

Comme le rappellent Baron & Kenny (1986)[166] un modérateur peut être composé d'une *variable qualitative* ou *quantitative* et il affecte la direction ou la force de la relation entre la *variable explicative* et la *variable à expliquer*. L'étude de la modération est donc l'étude de ce phénomène. La réalisation statistique de la modération s'effectue par une interaction entre la (ou l'une des) *variable explicative* et une troisième variable venant modifier la relation entre la *variable explicative* et la *variable à expliquer*.



Modèle de modération à 1 modérateur (W)

D'où ça vient ?

Les analyses de modération via des modèles de régression sont relativement récentes. Les premières analyses de ce type apparaissent au début des années 1980 dans la littérature, notamment en Psychologie et en Sciences du comportement[167]. Actuellement ces analyses sont toujours employées et se sont complexifiées en incluant notamment la possibilité d'interaction entre plus de 2 variables. Des modèles plus complexes encore ont été pensés et développés, tels que des modérations médiées ou des médiations modérées. Ces modèles ont été développés dans une approche basée sur des régressions, tels que les travaux de Preacher et Hayes sur le sujet[168], mais

[166] Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, nº 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].

[167] Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, nº 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].

[168] Hayes Andrew F. et Little Todd D., *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach*, Second edition., New York London, The Guilford Press, coll. « Methodology in the social sciences », 2018.

également dans le cadre de modèles structuraux permettant de traiter cette complexité, qui fera d'ailleurs l'objet d'une autre entrée.

Mot du praticien

Il est important de centrer et réduire les variables avant d'inclure le terme d'interaction, sinon il y a un risque multicolinéarité. L'ensemble des pré-requis concernant les analyses de modération sont les mêmes que ceux portant sur des **modèles de régression**.

Les modèles de modération sont tout-à-fait adaptés pour répondre à des hypothèses de recherche émises en SHS. Par exemple Teeters et al., en 2015, ont publié un article utilisant des modérations, dans le cadre de régressions, afin de tester le lien entre des attentes spécifiques liées aux jeux d'argent par les étudiants (telles que le gain matériel ou la reconnaissance sociale) et la fréquence de jeu[169]. Les auteurs avaient pour hypothèse que ce lien était modéré par le genre des étudiants.
Les modérations peuvent être simples, comme dans l'exemple cité précédemment ou multiples (avec plusieurs modérateurs) ou encore se combiner avec des variables médiatrices et composer des modérations médiées ou des médiations modérées.

Ressources

- Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, n° 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].
- Hayes Andrew F. et Little Todd D., *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach*, Second edition., New York London, The Guilford Press, coll. « Methodology in the social sciences », 2018.
- Teeters Jenni B., Ginley Meredith K., Whelan James P., Meyers Andrew W. et Pearlson Godfrey D., « The Moderating Effect of Gender on the Relation Between Expectancies and Gambling Frequency Among College Students », *Journal of Gambling Studies*, n° 1, vol. 31, 2015, p. 173-182, [https//doi.org/10.1007/s10899-013-9409-2].

[169] Teeters Jenni B., Ginley Meredith K., Whelan James P., Meyers Andrew W. et Pearlson Godfrey D., « The Moderating Effect of Gender on the Relation Between Expectancies and Gambling Frequency Among College Students », *Journal of Gambling Studies*, n° 1, vol. 31, 2015, p. 173-182, [https//doi.org/10.1007/s10899-013-9409-2].

Modération n.f. [/mɔ.de.ʁa.sjɔ̃ /]

Synonymes : latent interactions, interaction models

A quoi ça sert ?

Cette entrée est consacrée aux modèles de modération via des modèles en équation structurelles (SEM). Il existe une autre façon de réaliser des modèles de modération, en faisant appel aux **modèles de régression**, qui fera l'objet d'une autre entrée. Nous souhaitons tester une modération lorsque nous considérons que le lien entre deux variables va être altéré par les modalités d'une troisième variable. Le lien entre deux variables A et B se voit être modifié selon les valeurs prises par la variable modératrice C, celui-ci pouvant même s'inverser ou disparaître.

D'où ça vient ?

Les modèles de modération ont connu un essor dans les années 1980 dans la mouvance des travaux de Baron et Kenny qui présentent la différence entre **les modèles de médiation** et les modèles de modération dans le cadre de régressions[170]. Ces modèles de modération sont réalisables via des modèles en équation structurelles depuis la complexification de ces modèles dans les années 1980 avec notamment les travaux de Joreskog & Sorbom en 1984[171].

Mot du praticien

Les modèles de modération permettent d'envisager des liens entre des variables faisant l'objet d'hypothèses fréquemment émises en SHS. Par exemple Zhang et al en 2020 ont modélisé des modérations, via des **modèles en équations structurelles**, afin de tester le lien entre le capital social et les fonctions cognitives auprès de la population chinoise[172]. L'hypothèse des auteurs étant que ce lien serait modéré par le niveau d'éducation des participants. Le modèle **SEM** réalisé par les auteurs, et intégrant une médiation, leur ont permis de tester leur hypothèse théorique via cette analyse statistique.

Ressources

- Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, nº 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].
- Cheung Shu Fai, Cheung Sing-Hang, Lau Esther Yuet Ying, Hui C. Harry et Vong Weng Ngai, « Improving an old way to measure moderation effect in standardized units. », *Health Psychology*, nº 7, vol. 41, 2022, p. 502-505, [https//doi.org/10.1037/hea0001188].

[170] Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, nº 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].

[171] Jöreskog Karl G. et Sörbom Dag, *LISREL 6: analysis of linear structural relationships by maximum likelihood, instrumental variables and least squares methods; user's guide*, 4. ed., Mooresville, Ind, Scientific Software, Inc, 1986.

[172] Zhang Chunyu et Liu Liping, « The effect of job crafting to job performance », *Knowledge Management Research & Practice*, nº 2, vol. 19, 2021, p. 253-262, [https//doi.org/10.1080/14778238.2020.1762517].

- Jiang Ge, *Chapter 5 Lavaan Lab 3: Moderation and Conditional Effects | R Cookbook for Structural Equation Modeling*.
- Jöreskog Karl G. et Sörbom Dag, *LISREL 6: analysis of linear structural relationships by maximum likelihood, instrumental variables and least squares methods; user's guide*, 4. ed., Mooresville, Ind, Scientific Software, Inc, 1986.
- Rosseel Yves, « Lavaan: An R Package for Structural Equation Modeling », *Journal of Statistical Software*, vol. 48, 2012, p. 1-36, [https//doi.org/10.18637/jss.v048.i02].
- Slupphaug K. S., *Latent Interaction (and Moderation) Analysis in Structural Equation Models (SEM)*, [https://modsem.org/].
- Zhang Chunyu et Liu Liping, « The effect of job crafting to job performance », *Knowledge Management Research & Practice*, n° 2, vol. 19, 2021, p. 253-262, [https//doi.org/10.1080/14778238.2020.1762517].

Omega n.m. [/o.me.ga/]

Voir entrée **Cohérence Interne**.

Optimal Matching n.m. [/ɔp.ti.mal mat.ʃiŋ/]

Synonymes : Méthode d'appariement optimal

A quoi ça sert ?

L'Optimal Matching (appariement optimal) est une méthode statistique qui permet d'analyser et de comparer des séquences, c'est-à-dire des successions d'états, de comportements ou d'événements dans le temps. La méthode calcule la "distance" entre séquences en mesurant le coût minimal nécessaire pour transformer une séquence en une autre. Le calcul de cette distance repose sur des opérations de modification de bases des séquences :

- Substitution : remplacer un état par un autre (A → B)
- Insertion : ajouter un état dans la séquence
- Suppression : retirer un état de la séquence

Chaque opération est associée à un coût, et l'algorithme trouve la combinaison d'opérations qui minimise le coût total de transformation.

Ainsi cette méthode va nous permettre de :

– Classifier des séquences : Regrouper des parcours similaires en types ou clusters. Par exemple, identifier des "profils types" de carrières professionnelles.
– Mesurer des diversités : Quantifier à quel point les séquences d'un échantillon sont variées ou homogènes.
– Comparer : Comparer la structure des séquences entre groupes (genre, générations, pays), en tenant compte de leur ordre et de leur durée.
– Détecter des patterns : Identifier les transitions les plus fréquentes, les états les plus durables, les séquences atypiques.

D'où ça vient ?

Dans les années 1950 et 1960, les travaux menés dans le domaine de la théorie du codage ont permis d'aboutir aux méthodes d'optimal matching[173]. Ces méthodes ont d'abord été mobilisées en biologie, puis ont été intégrées aux sciences sociales dans les années 1980, 1990 et 2000 par Andrew Abbott[174].

[173] Lesnard Laurent, « Annexe 2. Les méthodes d'appariement optimal », *La famille désarticulée*, Paris cedex 14, Presses Universitaires de France, coll. « Le Lien social », 2009, p. 196-197.
[174] Idem ; Abbott Andrew, « Sequence Analysis: New Methods for Old Ideas », *Annual Review of Sociology*, nº 1, vol. 21, 1995, p. 93-113, [https//doi.org/10.1146/annurev.so.21.080195.000521]. ; Abbott Andrew et Forrest John, « Optimal Matching Methods for Historical Sequences », *Journal of Interdisciplinary History*, nº 3, vol. 16, 1986, p. 471, [https//doi.org/10.2307/204500]. ; Abbott Andrew et Hrycak Alexandra, « Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers », *American Journal of Sociology*, nº 1, vol. 96, 1990, p. 144-185, [https//doi.org/10.1086/229495]. ; Abbott Andrew et Tsay Angela, « Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect », *Sociological Methods & Research*, nº 1, vol. 29, 2000, p. 3-33, [https//doi.org/10.1177/0049124100029001001].

Exemple d'application

L'optimal matching est très utile pour l'étude des évènements qui jalonnent un parcours dans le temps, que ce soit des parcours de vie, des évolutions juridiques, scolaires ou culturelles. Pour cette raison, cette méthode peut être utilisée dans l'ensemble des disciplines des SHS. Abbott et Forrest en 1986 ont rédigé un article mobilisant l'optimal matching en sciences sociales, afin d'étudier les séquences de figures de danses folkloriques en Angleterre au XIXème sicèle dans le but de mettre en évidence les modèles de solidarité en milieu rural[175].

Mots du praticien

Cette méthode est principalement utilisée en sociologie et en démographie, car elle offre la possibilité d'introduire une dimension temporelle dans l'analyse. Elle se révèle ainsi particulièrement utile pour l'étude des parcours de vie (des trajectoires professionnelles, ou encore des séquences familiales), mais également pour l'analyse des comportements, comme par exemple les pratiques de mobilités quotidiennes ou des séquences d'actions répétées.

Cette méthode comporte toutefois plusieurs limites. Son coût computationnel peut être élévé, en particulier si les séquences sont nombreuses et complexes. La préparation des données est cruciale et souvent complexe, puisque l'ordre des séquences est au coeur de l'analyse. Le choix des coûts associés aux différentes transformations constitue un enjeu central, don't la definition est parfois subjective. Enfin, l'application de cette méthode peut conduire à des simplications importantes: certains événements rares ou atypiques peuvent être "écrasés" ou "dilués" dans les comparaisons.

Ressources

- Abbott Andrew, « Sequence Analysis: New Methods for Old Ideas », *Annual Review of Sociology*, nº 1, vol. 21, 1995, p. 93-113, [https//doi.org/10.1146/annurev.so.21.080195.000521].
- Abbott Andrew et Forrest John, « Optimal Matching Methods for Historical Sequences », *Journal of Interdisciplinary History*, nº 3, vol. 16, 1986, p. 471, [https//doi.org/10.2307/204500].
- Abbott Andrew et Hrycak Alexandra, « Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers », *American Journal of Sociology*, nº 1, vol. 96, 1990, p. 144-185, [https//doi.org/10.1086/229495].
- Abbott Andrew et Tsay Angela, « Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect », *Sociological Methods & Research*, nº 1, vol. 29, 2000, p. 3-33, [https//doi.org/10.1177/0049124100029001001].
- Lesnard Laurent, « Annexe 2. Les méthodes d'appariement optimal », *La famille désarticulée*, Paris cedex 14, Presses Universitaires de France, coll. « Le Lien social », 2009, p. 196-197.

[175] Robette Nicolas, « Mesurer la dissemblance entre trajectoires », *L'analyse statistique des trajectoires: Typologies de séquences et autres approches*, Ined Éditions, 2021, , [https//doi.org/10.4000/books.ined.16670]. ; Abbott Andrew et Forrest John, « Optimal Matching Methods for Historical Sequences », *Journal of Interdisciplinary History*, nº 3, vol. 16, 1986, p. 471, [https//doi.org/10.2307/204500].

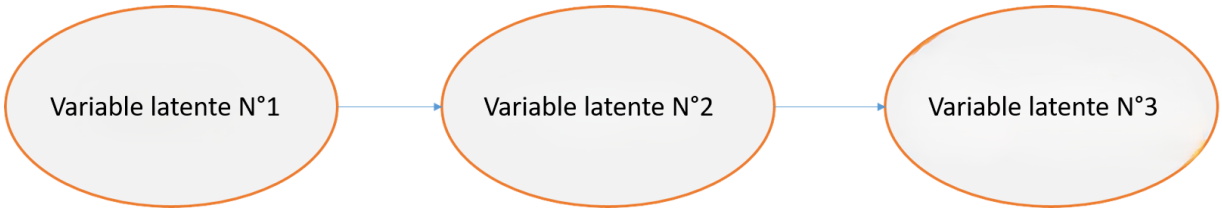
- Lesnard Laurent et Saint Pol Thibaut de, « Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis) », *Bulletin de méthodologie sociologique. Bulletin of sociological methodology*, nº 90, 2006, p. 5-25.
- Robette Nicolas, « Mesurer la dissemblance entre trajectoires », *L'analyse statistique des trajectoires: Typologies de séquences et autres approches*, Ined Éditions, 2021, , [https//doi.org/10.4000/books.ined.16670].

Pistes causales n.f. [/pist kozal/]

Synonymes : Path analysis, SEM, Analyse de parcours

A quoi ça sert ?

Les analyses en pistes causales font parties de la famille des **modèles en équation structurales** encore appelés **SEM**. Elles intègrent donc des *variables manifestes* (mesurées directement auprès de participants) et *latentes* (non-mesurables directement mais qui influencent les variables manifestes). L'intérêt de ces analyses est de tester la présence (ou non) de relations causales entre les différentes *variables latentes* intégrées au modèle. L'idée va être de tester un enchaînement de relations organisées dans un ordre défini par des hypothèses pré-établies. Dans l'illustration ci-dessous, l'hypothèse émise est que la variable latente N°1 a une relation de causalité avec la variable N°2, qui elle-même a une relation de causalité avec la variable N°3. Il y a donc l'hypothèse d'un parcours de relations causales.



D'où ça vient ?

Les modèles en pistes causales figurent parmi les premiers modèles en équations structurelles développés. Ceux-ci ont été créé dans les années 1920 par celui à qui on attribut la paternité des modèles **SEM**, le généticien Sewall Wright. En effet, les premiers développement de Wright portaient sur les analyses de parcours (path analysis) [176],[177].

Mot du praticien

Afin de réaliser des modèles en pistes causales il va être nécessaire de faire attention à différents éléments listés dans les pré-requis de l'entrée générale sur les **SEM**. Selon le type de variable (catégorielle ou continue) ou la taille de l'échantillon, il faudra prendre des précautions particulières à la réalisation de ces modèles.

Un point essentiel, général aux modèles **SEM** auxquels les modèles en pistes causales n'échappent pas, c'est la définition d'hypothèse a priori. Le modèle testé, le parcours causal envisagé, doit avoir été défini théoriquement en amont. Ce type de modèle n'est pas adapté à une démarche exploratoire, si c'est l'objectif visé, il faudra se tourner vers un autre type d'analyse.

---

[176] Wright Sewall, « The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs », *Proceedings of the National Academy of Sciences*, nº 6, vol. 6, 1920, p. 320-332, [https//doi.org/10.1073/pnas.6.6.320].

[177] Wright Sewall, « The Method of Path Coefficients », *The Annals of Mathematical Statistics*, nº 3, vol. 5, 1934, p. 161-215, [https//doi.org/10.1214/aoms/1177732676].

Les modèles en pistes causales sont par exemple employés en Psychologie où ils peuvent rendre compte des relations causales entre la neurocognition (étude des liens entre le cerveau et les processus cognitifs tels que la mémoire, l'attention et la prise de décision) et la cognition sociale (fait référence aux processus cognitifs impliqués dans les interactions sociales), puis entre la cognition sociale et l'alexithymie (la difficulté à identifier et exprimer ses émotions et /ou celles d'autrui). Par exemple Luo et al ont publié un article en 2021 dans lequel ils testent leur modèle théorique avec un modèle en pistes causales[178]. Les auteurs ont cherché à établir un chemin de relations causales entre la cognition (ex. mémoire de travail, vitesse d'exécution, attention, résolution de problèmes, etc.), la cognition des émotions vocales (Dégoût, peur, tristesse, etc.), l'alexithymie, les sympotômes négatifs (tels que la détresse ou l'asocialité), le résultat fonctionnel (ex. gestion des symptômes, des relations sociales, etc.) dans la schizophrénie (figure 3 de l'article). Cette étude, afin d'obtenir des relations causales porte sur des patients schizophrènes et un groupe contrôle de personnes ne souffrant d'aucune pathologie psychiatrique.

Ressources

- Luo Hongge, Zhao Yanli, Fan Fengmei, Fan Hongzhen, Wang Yunhui, Qu Wei, Wang Zhiren, Tan Yunlong, Zhang Xiujun et Tan Shuping, « A bottom-up model of functional outcome in schizophrenia », *Scientific Reports*, nº 1, vol. 11, 2021, p. 7577, [https//doi.org/10.1038/s41598-021-87172-4].
- Wright Sewall, « The Method of Path Coefficients », *The Annals of Mathematical Statistics*, nº 3, vol. 5, 1934, p. 161-215, [https//doi.org/10.1214/aoms/1177732676].
- Wright Sewall, « The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs », *Proceedings of the National Academy of Sciences*, nº 6, vol. 6, 1920, p. 320-332, [https//doi.org/10.1073/pnas.6.6.320].

[178] Luo Hongge, Zhao Yanli, Fan Fengmei, Fan Hongzhen, Wang Yunhui, Qu Wei, Wang Zhiren, Tan Yunlong, Zhang Xiujun et Tan Shuping, « A bottom-up model of functional outcome in schizophrenia », *Scientific Reports*, nº 1, vol. 11, 2021, p. 7577, [https//doi.org/10.1038/s41598-021-87172-4].

Q de Cochran n.m. [/ky də kɔkʁɑ̃/]

Synonymes : Cochran's Q test

A quoi ça sert ?

Le Q de Cochran est employé afin de comparer des fréquences sur des mesures répétées. Il est donc utilisé afin de comparer les modalités de variables catégorielles binaires sur différentes mesures, qu'elles soient temporelles ou d'une autre nature (par exemple : ce peut être pour comparer différentes techniques d'apprentissages testées auprès des mêmes participants).

D'où ça vient ?

Le Q de Cochran a été développé par William Cochran en 1950 [179].

Mot du praticien

Le Q de Cochran ne nécessite pas de pré-requis particuliers à part le fait de comparer les fréquences d'une variable binaire sur des mesures répétées.

Un exemple d'utilisation du Q de Cochran en Psychologie est présenté dans un article publié par Mason et al en 2021[180]. Dans cet article, les auteurs présentent l'utilité du test comme le Q de Cochran dans la mise en évidence de la sursélectivité des stimuli (réponses sélectives) au sein du répertoire verbal d'enfants atteints du trouble du spectre autistique (TSA). L'idée étant de comprendre la sursélectivité des stimuli (et donc dans cet exemple l'accès au répertoire verbal) d'enfants atteints du TSA sur plusieurs temps de mesure.

Ressources

- Cochran W. G., « The Comparison of Percentages in Matched Samples », *Biometrika*, n° 3/4, vol. 37, 1950, p. 256, [https//doi.org/10.2307/2332378].
- Mason Lee, Otero Maria et Andrews Alonzo, « Cochran's Q Test of Stimulus Overselectivity within the Verbal Repertoire of Children with Autism », *Perspectives on Behavior Science*, n° 1, vol. 45, 2022, p. 101-121, [https//doi.org/10.1007/s40614-021-00315-w].

[179] Cochran W. G., « The Comparison of Percentages in Matched Samples », *Biometrika*, n° 3/4, vol. 37, 1950, p. 256, [https//doi.org/10.2307/2332378].

[180] Mason Lee, Otero Maria et Andrews Alonzo, « Cochran's Q Test of Stimulus Overselectivity within the Verbal Repertoire of Children with Autism », *Perspectives on Behavior Science*, n° 1, vol. 45, 2022, p. 101-121, [https//doi.org/10.1007/s40614-021-00315-w].

Rainbowtest n.m. [/ˈreɪnbəʊtɛst/]

Synonymes : Test de linéarité

A quoi ça sert ?

Le Rainbow Test a été développé afin de tester le type de relation présente entre les variables avant de réaliser une modélisation. Il s'agit de vérifier que la relation est bien linéaire et non d'un autre ordre avant d'envisager un modèle linéaire[181].

D'où ça vient ?

Jessica Utts a développé le Rainbow Test en 1982, afin de rendre compte d'un large spectre de problèmes liés à la non-linéarité dans les modèles de **régression linéaires**[182]. L'idée sur laquelle repose un rainbowtest est que même si la relation entre un set de variables n'est pas linéaire, un ajustement linéaire peut malgré tout être obtenu sur des sous-groupes de l'échantillon testé. Il s'agit d'un test d'hypothèse, donc si l'hypothèse nulle est rejetée alors aucun ajustement linéaire n'est envisageable et le modèle linéaire ne devra pas être employé pour rendre compte des relations entre les variables sélectionnées. Ces travaux ont été enrichis par ceux de Williams sur le sujet (Williams, M., N., 2021)[183]. L'auteur présente les risques de réaliser une analyse statistique alors que les conditions ne sont pas adaptées à la méthode sélectionnée et ce dans le cas de variables catégorielles, ordinales ou continues.

Mot du praticien

Le rainbowtest est un très bon complément à la représentation graphique de relations entre des variables avant de réaliser des tests inférentiels envisageant une relation linéaire tels que des **régressions linéaires**.

Ressources

- Hothorn Torsten, Zeileis Achim, Farebrother Richard W., Cummins Clint, Millo Giovanni et Mitchell David, *R: Rainbow Test*, [https://search.r-project.org/CRAN/refmans/lmtest/html/raintest.html].
- Krämer Walter et Sonnberger Harald, *The Linear Regression Model Under Test*, Heidelberg, Physica-Verlag HD, 1986, [https//doi.org/10.1007/978-3-642-95876-2].
- Utts Jessica M., « The rainbow test for lack of fit in regression », *Communications in Statistics - Theory and Methods*, nº 24, vol. 11, 1982, p. 2801-2815, [https//doi.org/10.1080/03610928208828423].
- Williams Matt N., « Levels of measurement and statistical analyses », [https//doi.org/10.31234/osf.io/c5278].

[181] Krämer Walter et Sonnberger Harald, *The Linear Regression Model Under Test*, Heidelberg, Physica-Verlag HD, 1986, [https//doi.org/10.1007/978-3-642-95876-2].

[182] Utts Jessica M., « The rainbow test for lack of fit in regression », *Communications in Statistics - Theory and Methods*, nº 24, vol. 11, 1982, p. 2801-2815, [https//doi.org/10.1080/03610928208828423].

[183] Williams Matt N., « Levels of measurement and statistical analyses », [https//doi.org/10.31234/osf.io/c5278].

Régression linéaire n.f [/ʁegʁesjɔ̃ lineɛʁ/]

Synonymes : Modèle linéaire

A quoi ça sert ?

La régression linéaire est une méthode de modélisation permettant de mettre en évidence des liens corrélationnels ou causaux entre un ensemble de *variables explicatives* (ou *variables indépendantes*) et une *variable à expliquer* (*variable dépendante*). La *variable à expliquer* est nécessairement *quantitative* : continue ou ordinale. Les *variables explicatives* peuvent être aussi bien *quantitatives* que *qualitatives*. Ainsi, on parle aussi de “modèles de régressions linéaires” ou encore de modèles “linéaires”.

D’où ça vient ?

Le premier à utiliser la modélisation par régression est Francis Galton pour ses travaux sur l’hérédité en 1894 (Stanton, 2001)[184]. Karl Pearson s’est inspiré de ces travaux afin de développer le coefficient de corrélation de Pearson r, en 1896, (Stanton, 2001). En 1898, Francis Galton publie un article dans la revue Nature où il présente une méthode incluant la possibilité qu’un phénomène soit influencé par plusieurs facteurs. Cette conceptualisation permet d’envisager les modèles de régression multiples, donc qu’une variable s’explique par plusieurs variables (Stanton, 2001). Galton et Pearson, de l’école anglaise des statisticiens de la fin du XIX^{e} siècle, sont considérés comme les pères de la statistique moderne pour leurs différents travaux sur les corrélations et les modèles de régression notamment (Bingham, N. H., 2000) [185].

Mot du praticien

Lorsqu’il s’agit de rechercher à mettre en évidence des liens entre *variables quantitatives*, nous pouvons employer la **corrélation de Pearson** (il en existe d’autres) ou sa version généralisée : le modèle de régression linéaire, qui permet aussi de prédire la variable réponse sur de nouveaux échantillons de données. Lorsque les *variables explicatives* sont qualitatives, nous pouvons faire appel aux tests de comparaisons de moyennes tels que les **test-t (autrement appelés tests t de Student)** ou les **analyses de variance (ANOVA)**, qui sont en fait des modèles de régressions d’un format particulier (Darlington et Hayes, 2017)[186].

Pré-requis

[184] Stanton Jeffrey M., « Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors », *Journal of Statistics Education*, n° 3, vol. 9, 2001, p. 3, [https//doi.org/10.1080/10691898.2001.11910537].

[185] Bingham N., « Studies in the history of probability and statistics XLVI. Measure into probability: from Lebesgue to Kolmogorov », *Biometrika*, n° 1, vol. 87, 2000, p. 145-156, [https//doi.org/10.1093/biomet/87.1.145].

[186] Darlington Richard B. et Hayes Andrew F., *Regression analysis and linear models: concepts, applications, and implementation*, New York London, The Guilford Press, coll. « Methodology in the social sciences », 2017.

- Normalité de la distribution des résidus : il n'y a pas de nécessité de vérifier absolument cette condition, car la violation de cette condition a très peu d'impact sur les résultats si l'échantillon est suffisamment grand (Knief & Forstmeier, 2021).[187]
- Homogénéité des résidus : les résidus doivent être homogènes en fonction des différentes modalités des variables (si la répartition de l'erreur est bien due au hasard).
- Autocorrélation des résidus : si les résidus ne sont pas distribués au hasard, alors les différentes observations ne sont pas indépendantes, ce qui peut biaiser les résultats.
- Multicolinéarité des résidus : s'il y a une présence de multicolinéarité cela signifie que la part de variance partagée entre différentes variables explicatives (variables indépendantes) est trop importante. Il y a donc une redondance d'information dans notre modèle et donc il y a un risque de surreprésentation d'un phénomène. Ceci qui va fausser les résultats de la régression. Afin de vérifier la multicolinéarité des résidus différents tests existent tels que le Variance Inflation Factor (VIF), ou encore la tolérence. Dans notre exemple, nous présenterons uniquement le VIF. Il n'y a pas de consensus dans la littérature sur le seuil à retenir afin de conclure à une absence de multicolinéarité. Nous reprendrons les standards utilisés en Psychologie indiquant que les VIF doivent être inférieurs à 5 pour conclure à une absence de multicolinéarité (Farrar & Glauber, 1967).[188]
- Linéarité : La relation entre les variables explicatives et la variable expliquée est-elle linéaire ? Elle peut très bien être quadratique ou cubique ou autre, et dans ce cas-là la régression linéaire ne sera pas nécessairement le test le plus adapté pour répondre à l'hypothèse posée.

Au-delà de mettre en évidence des liens entre une variable explicative et une variable à expliquer, tout l'intérêt des modèles de régression réside dans la constitution d'un modèle, comprenant plusieurs variables explicatives. Autrement dit, dans la sélection des variables explicatives. La constitution d'un modèle statistique (de façon générale, mais également pour les régressions) implique de faire des choix car tout ne peut pas être testé. L'explication du phénomène étudié par le modèle alors établi n'est pas neutre et résulte de choix qui doivent être argumentés. En effet, le modèle produit est une version permettant d'apporter des éléments de compréhension d'un phénomène, mais il n'exclu pas que d'autres modèles soient envisageables (Bressoux, 2010) [189]. Toutefois, afin de faire ces choix il existe des procédés statistiques qui peuvent s'avérer éclairant, telles que les approches par comparaisons de modèles notamment (Judd et al., 2018)[190]. Mais ces choix ne peuvent pas être seulement statistiques et doivent nécessairement reposer sur des arguments théoriques. Ceci implique donc une certaine prudence dans l'interprétation des résultats obtenus.

Le choix des variables explicatives à intégrer au modèle de régression repose autant sur des considérations statistiques que théoriques. Dans certains cas, il est par exemple nécessaire de réaliser un modèle composé de variables d'intérêts et de variables contrôles. Les variables d'intérêts sont intégrées dans un modèle, car il est envisagé qu'elles apportent une part d'explication aux

[187] Knief Ulrich et Forstmeier Wolfgang, « Violating the normality assumption may be the lesser of two evils », *Behavior Research Methods*, nº 6, vol. 53, 2021, p. 2576-2590, [https//doi.org/10.3758/s13428-021-01587-5].
[188] Farrar Donald E. et Glauber Robert R., « Multicollinearity in Regression Analysis: The Problem Revisited », *The Review of Economics and Statistics*, nº 1, vol. 49, 1967, p. 92, [https//doi.org/10.2307/1937887].
[189] Bressoux Pascal, *Modélisation statistique appliquée aux sciences sociales*, De Boeck Supérieur, 2010, [https//doi.org/10.3917/dbu.bress.2010.01].
[190] Judd Chalres M., McClelland Gary H., Ryan Carey S., Muller Dominique et Yzerbyt Vincent, *Analyse des données: une approche par comparaison de modèles*, 2e éd., Louvain-la-Neuve [Paris], De Boeck supérieur, coll. « Collection ouvertures psychologiques », 2018.

variations de la variable à expliquer. Les variables contrôles, en revanche, n'ont pas de lien direct avec la variable à expliquer. Elles sont intégrées dans un modèle car leur absence pourraient modifier les résultats obtenus. En effet, dans un modèle de régression, le coefficient de chaque variable s'interprète **"Toutes choses égales par ailleurs"**. En effet, le coefficient représente l'incrément obtenu sur la variable réponse en augmentant d'une unité la variable explicative, et en maintenant constantes toutes les autres. En d'autres termes, à un score "neutre" obtenu sur chacune des autres variables du modèle, voici le résultat de cette variable. Pour obtenir ce score neutre, les variables quantitatives sont ramenées à 0 (si elles sont centrées avant), mais pour les variables catégorielles nous allons plutôt nous intéresser à la répartition entre les catégories et considérer qu'elle est égale. Par exemple, il est souvent observé que les femmes ont des salaires inférieurs aux hommes, mais si ce constat ne s'applique pas à poste égal, l'interprétation sera différente. Soit les hommes ont des postes plus élevés, ou à responsabilité plus élevée, et auquel cas, ils ont un salaire supérieur à celui des femmes. Soit leur salaire est plus élevé que celui des femmes à poste égal et il faudra trouver d'autres facteurs explicatifs à cette inégalité (temps partiel plus élevé chez les femmes, les femmes négocient-elles autant leur salaire que les hommes ?, etc.).

Un exemple de régression linéaire est présenté dans l'article de Sebal et al. 2023, dans lequel les auteurs cherchent à appréhender l'adhésion aux théories conspirationnistes[191]. Ils réalisent notamment une modélisation linéaire afin d'affiner le lien entre les croyances dans les phénomènes paranormaux et l'adhésion aux idées conspirationnistes. Ils observent par exemple que les croyances dans les superstitions (p. ex. les chats noirs portent malheur) sont associées positivement à la perception comme étant vraie de la théorie conspirationniste (résultats présentés dans la table 3 de l'article).

Le modèle de régression linéaire est une méthode extrêmement connue qui constitue souvent un objectif ou une fin en soi, mais il faut noter qu'il existe un grand nombre d'alternatives à cette méthode qui visent les mêmes objectifs.

Ressources

- Bingham N., « Studies in the history of probability and statistics XLVI. Measure into probability: from Lebesgue to Kolmogorov », *Biometrika*, nº 1, vol. 87, 2000, p. 145-156, [https//doi.org/10.1093/biomet/87.1.145].
- Bressoux Pascal, *Modélisation statistique appliquée aux sciences sociales*, De Boeck Supérieur, 2010, [https//doi.org/10.3917/dbu.bress.2010.01].
- Darlington Richard B. et Hayes Andrew F., *Regression analysis and linear models: concepts, applications, and implementation*, New York London, The Guilford Press, coll. « Methodology in the social sciences », 2017.
- Farrar Donald E. et Glauber Robert R., « Multicollinearity in Regression Analysis: The Problem Revisited », *The Review of Economics and Statistics*, nº 1, vol. 49, 1967, p. 92, [https//doi.org/10.2307/1937887].
- Judd Chalres M., McClelland Gary H., Ryan Carey S., Muller Dominique et Yzerbyt Vincent, *Analyse des données: une approche par comparaison de modèles*, 2e éd., Louvain-la-Neuve [Paris], De Boeck supérieur, coll. « Collection ouvertures psychologiques », 2018.

[191] Sebalo Ivan, Ball Linden J., Marsh John E., Morley Andy M., Richardson Beth H., Taylor Paul J. et Threadgold Emma, « Conspiracy theories: why they are believed and how they can be challenged », *Journal of Cognitive Psychology*, nº 4, vol. 35, 2023, p. 383-400, [https//doi.org/10.1080/20445911.2023.2198064].

- Knief Ulrich et Forstmeier Wolfgang, « Violating the normality assumption may be the lesser of two evils », *Behavior Research Methods*, nº 6, vol. 53, 2021, p. 2576-2590, [https//doi.org/10.3758/s13428-021-01587-5].
- Sebalo Ivan, Ball Linden J., Marsh John E., Morley Andy M., Richardson Beth H., Taylor Paul J. et Threadgold Emma, « Conspiracy theories: why they are believed and how they can be challenged », *Journal of Cognitive Psychology*, nº 4, vol. 35, 2023, p. 383-400, [https//doi.org/10.1080/20445911.2023.2198064].
- Stanton Jeffrey M., « Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors », *Journal of Statistics Education*, nº 3, vol. 9, 2001, p. 3, [https//doi.org/10.1080/10691898.2001.11910537].

Régression logistique n.f. [/rɛgʁɛsjɔ̃ lɔʒistik/]

Synonymes : Modèles Logit

A quoi ça sert ?

La régression logistique est une méthode statistique couramment utilisée dans les sciences sociales pour analyser la relation entre une *variable à expliquer* (*dépendante*) catégorielle (ou discrète) et une ou plusieurs *variables explicatives* (*indépendantes*), qu'elles soient de type catégoriel ou quantitatif continu. Contrairement à la régression linéaire, qui prédit des valeurs continues, la régression logistique est spécifiquement utilisée lorsque la *variable dépendante* est discrète, c'est-à-dire lorsqu'elle prend des valeurs distinctes et limitées, comme des catégories ou des classes.
L'objectif principal de la régression logistique est d'expliquer et de prédire la probabilité qu'un individu appartienne à une catégorie particulière de la variable à expliquer, en fonction des caractéristiques de plusieurs variables explicatives. Par exemple, en démographie, nous pouvons nous intéresser à la probabilité d'avoir un enfant (variable à expliquer binaire : « oui » ou « non ») en fonction de diverses caractéristiques sociodémographiques comme l'âge, le revenu, l'origine sociale, le statut d'activité, etc. (variables explicatives).

La régression logistique repose sur une transformation spécifique, appelée fonction logistique ou sigmoïde, qui permet de modéliser les probabilités tout en les contraignant à des valeurs comprises entre 0 et 1 [192] [193].

Les différents types de régression logistique

Il existe trois principaux types de régression logistique [194], dont le choix dépend de la nature de la variable dépendante et de la question de recherche :

1- Régression logistique binaire : utilisée lorsque la variable dépendante est de nature dichotomique (binaire), c'est-à-dire que ses modalités sont de type "oui/non", "réussite/échec" (par exemple : le fait d'avoir un enfant ou non, de voter ou non lors des élections, la réussite ou l'échec à un examen, etc.). Les coefficients de la régression logistique binaire quantifient l'effet de chaque variable explicative sur les chances (odds) que l'événement étudié se produise. Cette approche permet également de générer des prédictions sous forme de probabilités estimées, représentant pour chaque observation la probabilité de l'occurrence de l'événement. L'odds ratio mesure l'effet d'une variable explicative tout en maintenant les autres constantes. Pour **les variables continues** : l'OR représente le changement relatif pour une variation d'une unité. Exemple : un OR de 0,97 pour l'âge signifie que chaque année supplémentaire diminue de 3% les chances d'être au chômage. Pour **les variables catégorielles** : l'OR compare chaque catégorie à une modalité de référence. Exemple : un OR de 2,04 pour les titulaires du Bac (vs Bac+2 et plus) indique qu'ils ont deux fois

[192] Cette probabilité ne peut pas être modélisée par une droite car celle-ci conduirait à des valeurs impossibles (<0 ou >1). La régression logistique est un cas particulier de modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien pour mettre en relation la probabilité de réalisation d'un événement et la combinaison linéaire de variable explicatives.
[193] Le modèle probit peut également être utilisé pour modéliser une variable discrète binaire, mais il repose sur une fonction de répartition normale, contrairement à la régression logistique qui utilise la fonction logistique. Les deux modèles aboutissent à des résultats similaires dans la plupart des cas. Toutefois, lorsque les erreurs suivent une distribution normale, le modèle probit est parfois privilégié.
[194] Il existe d'autres types de modèles logit qui ne sont pas abordés dans cette fiche. Parmi ceux-ci, on peut citer le modèle logit conditionnel, le modèle logit marginal, le modèle logit multiniveau.

plus de risques d'être au chômage. Une fois le modèle établi, il permet de calculer des probabilités prédites individuelles (ex: risque de chômage de 0,17 pour une femme bachelière de 30 ans vivant seule avec enfants), mais aussi visualiser les probabilités moyennes selon différentes caractéristiques et classer les individus selon un seuil de probabilité prédéfini.

2- Régression logistique multinomiale : mobilisée pour une variable dépendante nominale, c'est-à-dire dont les modalités ne peuvent pas être hiérarchisées les unes par rapport aux autres (par exemple : le mode de transport utilisé pour se rendre au travail, le choix d'un régime alimentaire, etc.). Ce modèle est une extension de la régression logistique binaire, il permet de prédire la probabilité qu'une observation appartienne à l'une des modalités de la variable d'intérêt en fonction d'un ensemble de variables explicatives. Son principe repose sur la comparaison de chaque modalité de la variable réponse avec une modalité de référence choisie. Les résultats sont exprimés en termes de cotes (odds), et les odds ratios permettent d'évaluer l'influence de chaque variable explicative sur la probabilité d'appartenance à une catégorie donnée, en comparaison avec la modalité de référence. Pour utiliser une régression logistique multinomiale 3 conditions doivent être vérifiées :

- Indépendance des alternatives (IIA): cette hypothèse stipule que le choix entre deux alternatives ne doit pas être influencé par l'ajout ou la suppression d'autres alternatives. Le test de Hausman-McFadden permet de la vérifier en comparant un modèle complet à un modèle restreint via les log-vraisemblances et l'AIC.
- Linéarité des variables continues : il faut vérifier l'absence de relation non linéaire en ajoutant des termes quadratiques aux variables continues. Si cela améliore significativement l'ajustement, la relation n'est pas strictement linéaire.
- Gérer les valeurs extrêmes.

Pour interpréter les résultats de ce modèle nous allons, comme pour toutes les régressions logistiques, s'appuyer sur les odds ratios. Dans la régression logistique les OR vont mesurer le changement des cotes d'appartenir à une catégorie vs la référence :

- OR > 1 : augmentation de la probabilité d'appartenance
- OR < 1 : diminution de la probabilité
- OR = 1 : aucun impact

Nous retrouverons aussi les probabilités prédites qui complètent les OR en quantifiant les probabilités absolues d'appartenance à chaque catégorie. Les probabilités prédites permettent de mieux comprendre l'impact d'une variable en comparant la probabilité d'appartenance à chaque catégorie, toutes choses égales par ailleurs.

3- Régression logistique ordinale : employée pour modéliser une variable dépendante ordinale, c'est-à-dire dont les modalités sont logiquement ordonnées sur une échelle (par exemple : "d'accord", "ni d'accord, ni pas d'accord", "pas d'accord", etc.). Elle permet d'étudier des degrés d'accord ou de désaccord sur un sujet donné, ou le positionnement des individus sur une échelle socioéconomique ou de valeurs, etc.
Le modèle le plus courant est la régression logistique ordinale cumulative avec odds proportionnels, qui repose sur l'hypothèse que l'effet des variables explicatives reste constant à travers toutes les catégories de la variable dépendante. Un seul coefficient est estimé par variable explicative. Pour vérifier cette hypothèse, nous pouvons utiliser le test de Brant, si l'hypothèse n'est pas respectée, plusieurs alternatives existent :

- Modèle à odds partiels proportionnels (PPO) : permet des effets différenciés pour certaines variables
- Modèle ordinal adjacent : compare les catégories deux à deux
- Modèle multinomial : si l'ordre des catégories est peu pertinent

Dans un modèle de régression logistique ordinale, les odds ratios (OR) mesurent l'effet des variables explicatives sur la probabilité d'appartenir à une catégorie plus élevée de la variable ordonnée, toutes autres caractéristiques étant maintenues constantes dans le modèle. Les OR mesurent l'effet des variables sur la probabilité d'appartenir à une catégorie plus élevée :

- OR < 1 : probabilité plus élevée d'une opinion favorable
- OR > 1 : probabilité plus élevée d'une opinion moins favorable

Les probabilités prédites permettent de mieux comprendre l'effet des variables explicatives sur la répartition des individus entre les différentes catégories de la variable dépendante. Les probabilités prédites vont quantifier directement la probabilité d'appartenir à chaque catégorie selon les caractéristiques individuelles.

D'où ça vient ?

Le concept de fonction logistique remonte au XIX^{e} siècle, lorsque le mathématicien belge Pierre-François Verhulst, élève d'Adolphe Quetelet, propose pour la première fois cette fonction pour modéliser l'accroissement des populations biologiques en prenant en compte les limites imposées par l'environnement [195]. Cette approche a ensuite été adaptée par les statisticiens pour développer la régression logistique. Depuis, elle est devenue un outil central en analyse statistique.

Exemples d'application

Les exemples d'applications sont très nombreux et variés dans toutes les disciplines des SHS (démographie, science politique, psychologie, géographie, épidémiologie, économie...). La régression logistique pourra ainsi servir pour modéliser la probabilité d'avoir un enfant, de se déplacer ou de décéder en fonction de facteurs tels que l'âge, le niveau de diplôme ou le revenu. Elle sera également utile pour prédire la probabilité de voter pour un candidat en fonction de diverses caractéristiques sociodémographiques ou encore prédire la probabilité qu'un individu développe un trouble de l'anxiété en fonction de plusieurs variables, comme les antécédents familiaux, les facteurs environnementaux ou la consommation de certaines substances. D'autres exemples d'application peuvent être mentionnés tels que l'étude des facteurs (altitude, aménagement urbain, type de sol, etc.) qui influencent la probabilité qu'une zone géographique soit sujette à des inondations ou encore pour prédire la probabilité de contracter une maladie en fonction de facteurs de risque tels que l'âge ou le mode de vie.

Mot du praticien

Contrairement à la régression linéaire, qui utilise la méthode des moindres carrés ordinaires pour estimer les paramètres du modèle, la validité de la régression logistique ne dépend pas de la satisfaction des hypothèses de linéarité, d'homoscédasticité ou de normalité des résidus. Toutefois,

[195] Schtickzelle Martial, « Pierre-François Verhulst (1804-1849). La première découverte de la fonction logistique »:, *Population*, n^{o} 3, Vol. 36, 1981, p. 541-556, [https//doi.org/10.3917/popu.p1981.36n3.0556].

pour garantir la validité des résultats estimés par la régression logistique, certaines hypothèses doivent être respectées quel que soit le type de modèle (binaire, multinomial ou ordinal) :

- Absence de multicolinéarité : Il ne doit pas y avoir de fortes relations (corrélations) entre les variables explicatives du modèle, afin d'assurer la stabilité et la fiabilité des estimations des coefficients, ainsi que de permettre une interprétation claire et précise des effets individuels de chaque variable. L'approche la plus classique pour détecter un problème de multicolinéarité consiste à examiner les facteurs d'inflation de la variance (FIV) (ou **variance inflation factor** - **VIF**) qui mesure le degré de corrélation linéaire d'une variable explicative par rapport aux autres. Une autre mesure utilisée est la tolérance, qui est l'inverse du **VIF** (1/VIF). Dans cet exemple, nous calculerons uniquement le **VIF**, qui permet d'évaluer dans quelle mesure la variance d'un coefficient est augmentée en raison de la multicolinéarité.
- Indépendance : Les observations doivent être indépendantes les unes des autres. Cette condition est particulièrement importante, car la dépendance entre les observations peut entraîner des estimations biaisées.
- Log-linéarité: Il doit exister une relation globalement linéaire entre les variables explicatives et la valeur du coefficient du modèle qui lui est associé. En d'autres termes, une augmentation d'une unité d'une variable explicative continue doit entraîner une multiplication constante du facteur de risque (odds ratio), indépendamment de la valeur des autres variables.
- Taille suffisante de l'échantillon : En règle générale, la taille de l'échantillon est déterminée à l'aide du critère événements par variable (EPV), qui définit le nombre minimal d'événements nécessaires par variable explicative pour garantir la fiabilité des résultats. Il est couramment recommandé de disposer d'au moins 10 observations par variable indépendante (EPV ≥ 10) [196], [197], [198]. Par exemple, si le modèle comporte 10 variables explicatives et que la probabilité de la modalité la moins fréquente est de 0.2, l'échantillon doit comporter au moins 500 individus ((10 * 10) / 0.2 = 500).

D'autres hypothèses sont spécifiques à certaines formes de modélisation, comme par exemple l'hypothèse de proportionnalité des risques, dans le cas d'un modèle de régression ordinal qui suppose que l'effet d'une variable explicative sur la variable dépendante reste constant à travers les différentes catégories de la variable dépendante.

Un point particulier concerne la prise en compte des pondérations dans les modèles. En effet, l'inclusion des pondérations dans les modèles statistiques, en particulier dans les modèles de régression logistique, fait l'objet de nombreux débats dans la littérature[199]. L'intégration des

[196] Dans le cas de la régression logistique multinomiale, qui comporte plusieurs catégories pour la variable réponse, il peut être nécessaire de disposer d'un échantillon plus grand afin d'assurer une puissance statistique suffisante pour chaque catégorie, en particulier lorsque certaines catégories sont moins fréquentes.

[197] Hosmer David W. et Lemeshow Stanley, *Applied logistic regression*, New York, Wiley, coll. « Wiley series in probability and mathematical statistics Applied probability and statistics », 1989.

[198] Peduzzi Peter, Concato John, Kemper Elizabeth, Holford Theodore R. et Feinstein Alvan R., « A simulation study of the number of events per variable in logistic regression analysis », *Journal of Clinical Epidemiology*, nº 12, vol. 49, 1996, p. 1373-1379, [https//doi.org/10.1016/S0895-4356(96)00236-3].

[199] Joubert Léo, Van Truoc Olivier Lê, Mercklé Pierre et Tudoux Benoît, « Redresser l'échantillon d'une enquête en ligne : un exemple à partir de l'enquête Vico », *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, nº 1, vol. 158, 2023, p. 143-166, [https//doi.org/10.1177/07591063231160287].
; Davezies Laurent et D'Haultfoeuille Xavier, *Faut-il pondérer ?... Ou l'éternelle question de l'économètre confronté à des données d'enquête - Documents de travail - G2009/06 | Insee*, [https://www.insee.fr/fr/statistiques/1380863] ; Miratrix Luke W., Sekhon Jasjeet S., Theodoridis Alexander G.

pondérations dans la modélisation permet de corriger les biais liés à la non-réponse et de prendre en compte le plan de sondage dans l'analyse, permettant ainsi de mieux refléter la structure de la population cible et d'assurer la représentativité des résultats. Toutefois, l'utilisation des pondérations entraîne généralement une augmentation des erreurs-types des paramètres et réduit la puissance statistique des modèles. En outre, la multiplicité des tests statistiques disponibles pour tenir compte des pondérations dans une analyse de régression suscite des préoccupations quant à leur impact réel sur l'estimation des coefficients et des p-values. Enfin, les pondérations ne permettent pas d'assurer la représentativité de l'échantillon sur les variables inobservées potentiellement correlées au phénomène étudié, ni de tenir compte des interactions complexes entre variables.

Par exemple dans l'enquête ERFI (Étude des relations familiales et intergénérationnelles)[200], des pondérations sont calculées pour ajuster l'échantillon sur les caractéristiques sociodémographiques de la population cible concernant six variables : le croisement entre le sexe et le groupe d'âge, le croisement entre le sexe et le nombre d'habitants du ménage, la catégorie socioprofessionnelle, la nationalité, la taille de l'unité urbaine de résidence, ainsi que la zone d'étude et d'aménagement du territoire de résidence.

Ressources

- Agresti Alan, *Categorical data analysis*, Third edition., Hoboken, New Jersey, Wiley-Interscience, coll. « Wiley series in probability and statistics », 2013.
- Arikan Serkan, Özer Ferah, Şeker Vuslat et Ertaş Güneş, « The Importance of Sample Weights and Plausible Values in Large-Scale Assessments », *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, nº 1, vol. 11, 2020, p. 43-60, [https//doi.org/10.21031/epod.602765].
- Boto-García David, « Good results come to those who weight: on the importance of sampling weights in empirical research using survey data », *Current Issues in Tourism*, nº 2, vol. 27, 2024, p. 268-287, [https//doi.org/10.1080/13683500.2023.2178394].
- Brant Rollin, « Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression », *Biometrics*, nº 4, vol. 46, 1990, p. 1171, [https//doi.org/10.2307/2532457].
- Cheng Simon et Long J. Scott, « Testing for IIA in the Multinomial Logit Model », *Sociological Methods & Research*, nº 4, vol. 35, 2007, p. 583-600, [https//doi.org/10.1177/0049124106292361].
- Davezies Laurent et D'Haultfoeuille Xavier, *Faut-il pondérer ?... Ou l'éternelle question de l'économètre confronté à des données d'enquête - Documents de travail - G2009/06* | *Insee*, [https://www.insee.fr/fr/statistiques/1380863].

et Campos Luis F., « Worth Weighting? How to Think About and Use Weights in Survey Experiments », *Political Analysis*, nº 3, vol. 26, 2018, p. 275-291, [https//doi.org/10.1017/pan.2018.1] ; Boto-García David, « Good results come to those who weight: on the importance of sampling weights in empirical research using survey data », *Current Issues in Tourism*, nº 2, vol. 27, 2024, p. 268-287, [https//doi.org/10.1080/13683500.2023.2178394]. ; Arikan Serkan, Özer Ferah, Şeker Vuslat et Ertaş Güneş, « The Importance of Sample Weights and Plausible Values in Large-Scale Assessments », *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, nº 1, vol. 11, 2020, p. 43-60, [https//doi.org/10.21031/epod.602765]. ; Winship Christopher et Radbill Larry, « Sampling Weights and Regression Analysis », *Sociological Methods & Research*, nº 2, vol. 23, 1994, p. 230-257, [https//doi.org/10.1177/0049124194023002004].

[200] INED, *Enquête ERFI - GGS*, [https://erfi.site.ined.fr/].

- Fagerland Morten W. et Hosmer David W., « How to Test for Goodness of Fit in Ordinal Logistic Regression Models », *The Stata Journal: Promoting communications on statistics and Stata*, n° 3, vol. 17, 2017, p. 668-686, [https//doi.org/10.1177/1536867X1701700308].
- Fullerton Andrew S. et Xu Jun, « The proportional odds with partial proportionality constraints model for ordinal response variables », *Social Science Research*, n° 1, vol. 41, 2012, p. 182-198, [https//doi.org/10.1016/j.ssresearch.2011.09.003].
- Hausman Jerry et McFadden Daniel, « Specification Tests for the Multinomial Logit Model », *Econometrica*, n° 5, vol. 52, 1984, p. 1219, [https//doi.org/10.2307/1910997].
- Hilbe Joseph M., *Logistic Regression Models*, 0 éd., Chapman and Hall/CRC, 2009, [https//doi.org/10.1201/9781420075779].
- Hosmer David W. et Lemeshow Stanley, *Applied logistic regression*, New York, Wiley, coll. « Wiley series in probability and mathematical statistics Applied probability and statistics », 1989.
- Hosmer David W., Lemeshow Stanley et Sturdivant Rodney X., *Applied Logistic Regression*, 1re éd., Wiley, coll. « Wiley Series in Probability and Statistics », 2013, [https//doi.org/10.1002/9781118548387].
- INED, *Enquête ERFI - GGS*, [https://erfi.site.ined.fr/].
- Joubert Léo, Van Truoc Olivier Lê, Mercklé Pierre et Tudoux Benoît, « Redresser l'échantillon d'une enquête en ligne : un exemple à partir de l'enquête Vico », *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, n° 1, vol. 158, 2023, p. 143-166, [https//doi.org/10.1177/07591063231160287].
- McFadden D., *Conditional Logit Analysis of Qualitative Choice Behavior*, Institute of Urban and Regional Development, University of California, 1973.
- Menard Scott, *Logistic Regression: From Introductory to Advanced Concepts and Applications*, 2455 Teller Road, Thousand Oaks California 91320 United States, SAGE Publications, Inc., 2010, [https//doi.org/10.4135/9781483348964].
- Miratrix Luke W., Sekhon Jasjeet S., Theodoridis Alexander G. et Campos Luis F., « Worth Weighting? How to Think About and Use Weights in Survey Experiments », *Political Analysis*, n° 3, vol. 26, 2018, p. 275-291, [https//doi.org/10.1017/pan.2018.1].
- Peduzzi Peter, Concato John, Kemper Elizabeth, Holford Theodore R. et Feinstein Alvan R., « A simulation study of the number of events per variable in logistic regression analysis », *Journal of Clinical Epidemiology*, n° 12, vol. 49, 1996, p. 1373-1379, [https//doi.org/10.1016/S0895-4356(96)00236-3].
- Pregibon Daryl, « Logistic Regression Diagnostics », *The Annals of Statistics*, n° 4, vol. 9, 1981, [https//doi.org/10.1214/aos/1176345513].
- Schtickzelle Martial, « Pierre-François Verhulst (1804-1849). La première découverte de la fonction logistique »:, *Population*, n° 3, Vol. 36, 1981, p. 541-556, [https//doi.org/10.3917/popu.p1981.36n3.0556].
- Sjoberg Daniel, *Function to display multinomial regression models in wide format*, [https://gist.github.com/ddsjoberg/a55afa74ac58e1f895862fcabab62406].
- Werth Rose, *Categorical Regression in Stata and R*, Bookdown, 2022.
- Williams Richard, « Understanding and interpreting generalized ordered logit models », *The Journal of Mathematical Sociology*, n° 1, vol. 40, 2016, p. 7-20, [https//doi.org/10.1080/0022250X.2015.1112384].
- Winship Christopher et Radbill Larry, « Sampling Weights and Regression Analysis », *Sociological Methods & Research*, n° 2, vol. 23, 1994, p. 230-257, [https//doi.org/10.1177/0049124194023002004].

Régressions de panel n.m. [/ʁegʁɛsjɔ̃ də panɛl/]

Synonymes: Modèles de panel

A quoi ça sert ?

Les régressions de panel s'appliquent sur des données de panel. Les données de panel se définissant par un aspect longitudinal, à savoir que pour un même individu (participant, entreprise, etc.) nous disposons de plusieurs observations, c'est-à-dire de plusieurs dates (plusieurs temps de mesure)[201]. Ces données comprennent généralement plusieurs individus. L'avantage des régressions de panel est donc de tenir compte à la fois de données transversales (qui ne changent pas dans le temps) et de données longitudinales (susceptibles de changer avec le temps). Ce type d'analyse permet de prendre en compte les problèmes liés à l'autocorrélation et à l'hétérogénéité inobservée. Ces modèles peuvent répondre aux mêmes questionnements que les **modèles mixtes**.

Les données peuvent être analysées selon trois types d'approches :

- Les régressions homogènes ou pooled
- Les modèles à effets aléatoires
- Les modèles à effets fixes

Le choix de l'un de ces modèles dépend des hyothèses de recherche alors posées et des questions auxquelles nous souhaitons répondre.

Les régressions homogènes ou pooled

Ces modèles permettent d'étudier des combinaisons de variations intra-individuelles et inter-individuelles. Ils reposent sur le pré-supposé qu'il n'y a pas d'autocorrélation des résidus.

Modèles à effets fixes

Le modèle à effets fixes contrôle les caractéristiques invariantes des sujets dans le temps, ce qui permet d'examiner les variations intra-sujet. Cette approche est particulièrement utile lorsque l'accent est mis sur l'analyse de l'impact des variables qui changent au fil du temps.

Modèles à effets aléatoires

Le modèle à effets aléatoires suppose, quant à lui, que les effets spécifiques à chaque individu ne sont pas corrélés aux variables indépendantes, ce qui permet d'analyser les variations intra-sujet (facteur idiosyncrasique) et inter-sujet (facteur individuel).
Le choix entre ces modèles dépend de la nature des données et de la question de recherche à traiter.

D'où ça vient ?

Les analyses de données de panel ont été développées dans les années 1960 - 1970 suite à la collecte de données de grande envergure sur une période de temps assez longue. Par exemple,

[201] Bourbonnais Régis, « Chapitre 13. Introduction à l'économétrie des données de panel », *Éco Sup*, 2018, p. 371-387.

après la seconde guerre mondiale le gouvernement américain a mis en place un programme de lutte contre la pauvreté (touchant alors 1 américain sur 5), qui incluait la collecte de données sur un panel de familles, dans le temps, afin d'appréhender les dynamiques du marché du travail notamment[202]. Différents panels ont vu le jour à cette période afin de saisir les évolutions du marché du travail et des conditions de vie des habitants, aux Etats-Unis puis en Europe. C'est pour analyser ces données issues de panels, contenant des informations récoltées auprès de plusieurs miliers de participants, sur plusieurs années, que les modèles de panels ont été développé. Les premiers travaux sur le sujet sont parus dans les années 1970, avec un premier point d'orgue en 1977, lors de la première conférence sur l'analyse des données de panel organisée par l'INSEE[203]. Le champ a continué de s'enrichir de nouveaux développements avec une grande partie des travaux menés en économétrie et plus généralement en économie.

Mot du praticien

Les modèles de panel peuvent être appliqués dans toutes les disciplines des SHS, même s'ils sont principalement employés en économie et en gestion.

En économie, par exemple, ils peuvent être utiles si nous souhaitons étudier les revenus d'un échantillon de ménages sur plusieurs années. En science de l'éducation, les modèles de panel pourront être mobilisés afin d'évaluer la qualité d'une politique éducative sur la performance des élèves dans le temps, sur plusieurs écoles par exemple.

Les pré-requis pour réaliser des régressions de panel sont les mêmes que ceux nécessaires à l'application de modèles de **régression linéaire**. Avant de réaliser ce type de modèles il faudra s'assurer de la linéarité des relations (sauf si le modèle pré-supposé est non-linéaire ou logistique), de l'homoscédasticité (variance constante des erreurs) et de l'absence de multicolinéarité entre les variables indépendantes. Ces conditions d'application viennent s'ajouter aux pré-requis classiques de nettoyage des données, tels que le traitement des valeurs manquantes et des valeurs extrêmes.

Ressources

- Bourbonnais Régis, « Chapitre 13. Introduction à l'économétrie des données de panel », *Éco Sup*, 2018, p. 371-387.
- Hanck Christoph, Arnold Martin, Gerber Alexander et Schmelzer Martin, « 10 Regression with Panel Data | Introduction to Econometrics with R », *Introduction to Econometrics with R*, Bookdown., 2025 .
- Sarafidis Vasilis et Wansbeek Tom, « Celebrating 40 years of panel data analysis: Past, present and future », *Journal of Econometrics*, n° 2, vol. 220, 2021, p. 215-226, [https//doi.org/10.1016/j.jeconom.2020.06.001].

[202] Sarafidis Vasilis et Wansbeek Tom, « Celebrating 40 years of panel data analysis: Past, present and future », *Journal of Econometrics*, n° 2, vol. 220, 2021, p. 215-226, [https//doi.org/10.1016/j.jeconom.2020.06.001].

[203] Sarafidis Vasilis et Wansbeek Tom, « Celebrating 40 years of panel data analysis: Past, present and future », *Journal of Econometrics*, n° 2, vol. 220, 2021, p. 215-226, [https//doi.org/10.1016/j.jeconom.2020.06.001].

Régressions polynomiales n.f. [/ʁe.gʁe.sjɔ̃ pɔ.li.nɔ.mjal/]

Synonymes : Relations non linéaires, régression modulée, modèles non additifs

A quoi ça sert ?

Les modèles de régression polynomiaux permettent de mesurer des relations entre une *variable à expliquer* et des *variables explicatives* qui ne sont pas de nature linéaire, mais quadratiques, cubiques, etc. contrairement au modèle linéaire classique qui étudie une relation linéaire qui lui mesure une relation linéaire entre une VD et une ou des VI. Afin de réaliser ces relations non-linéaires il est nécessaire d'intégrer des polynomes dans l'équation de régression. Ces modèles sont en fait des cas particuliers des modèles linéaires, dans le sens où ceux-ci sont des modèles avec des paramètres linéaires auxquels sont ajoutés des polynômes.

La régression linéaire est donc en réalité un polynôme de degré 1 : $y = \alpha + \beta_1 x + \epsilon$

Les régressions polynomiales de degré 2, ou régressions quadratiques, intègrent donc un degré supplémentaire : $y = \alpha + \beta_1 x + \boldsymbol{\beta_2 x^2} + \epsilon$

Ces régressions peuvent avoir un nombre quelconque de degrés, mais il faudra choisir le modèle qui introduira le moins de biais possible et donc la part d'erreur la plus faible :

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \boldsymbol{\beta_h x^h} + \epsilon$$

Actuellement les régressions polynomiales font l'objet de deux usages. Le premier est de vérifier l'hypothèse de **linéarité** (ou de non-linéarité) entre les variables explicatives et la variable à expliquer. Il s'agira de réaliser un modèle de **régression linéaire**, puis un modèle polynomial (de degré 2) et de comparer les deux modèles. Si le modèle polynomial explique une part de variance substantiellement plus importante que le modèle linéaire, alors l'hypothèse de linéarité peut être rejetée. La seconde application est directement d'émettre l'hypothèse que la relation entre la variable à expliquer et la (ou les) variable(s) explicative(s) n'est pas linéaire mais quadratique (ou cubique, etc.) en s'appuyant sur la littérature existante sur le sujet traité. Dans ce cas il ne sera pas nécessaire de comparer des modèles, il suffira juste d'effectuer le modèle le plus adapté à la relation entre les variables du modèle.

D'où ça vient ?

Le premier modèle polynomial a été développé par Joseph-Diez Gergonne en 1815 [204]. Les travaux de Gergonne ont permis de conceptualiser des modèles de régression polynomiales et donc d'envisager d'autres liens que le lien linéaire entre les variables explicatives et la variable à expliquer.

Mot du praticien

Les pré-requis sont les mêmes que ceux nécessaires à l'application d'une **régression linéaire**, sauf celui concernant la linéarité qui devient inutile.

[204] Stigler, S. M. (1974). "Gergonne's 1815 paper on the design and analysis of polynomial regression experiments". Historia Mathematica. 1 (4): 431–439. https://doi.org/10.1016/0315-0860(74)90033-0

Les régressions polynomiales peuvent s'avérer très utiles afin de rendre compte de relations entre différents éléments en Sciences Humaines et Sociales. La relation linéaire entre les phénomènes ne permet pas d'envisager l'ensemble des relations possibles et contraint la pensée d'hypothèses de recherche qui pourraient être plus adaptées à la réalité du terrain en intrégrant des relations non strictement linéaires. Par exemple, en psychologie, les modèles quadratiques peuvent être employés afin de rendre compte de la relation entre le stress de parents d'enfants atteint du trouble du spectre autistique (TSA) et le retard mental associé de ces enfants. En effet, les enfants atteints du trouble du spectre autistique ayant un retard mental très important ou inexistant ont des parents ayant un taux de stress plus important que le reste de la population d'enfants atteints d'un TSA avec un retard mental moyen. La relation entre le taux de stress des parents et le retard mental des enfants atteints d'un TSA n'est donc pas linéaire mais quadratique. Il sera, dans ce cas, plus adapté de réaliser une régression intégrant des polynomes quadratiques.

Nous pouvons également citer les travaux d'Enders et al. 2024, portant sur la relation entre l'orientation politique et l'adhésion aux théories du complot[205]. Les auteurs émettent l'hypothèse que la relation entre l'orientation politique autorapportée et l'adhésion aux théories complotistes n'est pas linéaire mais curvilinéaire. L'adhésion à ces théories étant plus importante chez les personnes se définissant aux extrêmes de l'échiquier politique, à savoir extrême droite et extrême gauche, et ne suit pas un gradient croissant plus traditionnel entre droite et gauche. Ils ont mené une enquête de grande envergure auprès de participants présents dans une vingtaine de pays occidentaux. Puis afin de rendre compte de la relation émise dans leur hypothèse, les auteurs ont comparé différents modèles de régression : linéaire, quadratique, cubique, polynomial du 4ème degré et polynomial du 5ème degré. Les auteurs ont ensuite comparé la part de variance expliquée ajoutée pour chaque modèle (R^2) afin de sélectionner celui le plus adapté à leurs données et le plus pertinent pour répondre à leur hypothèse.

Ressources

- Bressoux Pascal, *Modélisation statistique appliquée aux sciences sociales*, De Boeck Supérieur, 2010, [https//doi.org/10.3917/dbu.bress.2010.01].
- Darlington Richard B. et Hayes Andrew F., *Regression analysis and linear models: concepts, applications, and implementation*, New York London, The Guilford Press, coll. « Methodology in the social sciences », 2017.
- Enders Adam, Klofstad Casey, Littrell Shane, Miller Joanne, Theocharis Yannis, Uscinski Joseph et Zilinsky Jan, « Left–right political orientations are not systematically related to conspiracism », *Political Psychology*, , 2024, p. pops.13017, [https//doi.org/10.1111/pops.13017].
- Stigler Stephen M., « Gergonne's 1815 paper on the design and analysis of polynomial regression experiments », *Historia Mathematica*, nº 4, vol. 1, 1974, p. 431-439, [https//doi.org/10.1016/0315-0860(74)90033-0].

[205] Enders Adam, Klofstad Casey, Littrell Shane, Miller Joanne, Theocharis Yannis, Uscinski Joseph et Zilinsky Jan, « Left–right political orientations are not systematically related to conspiracism », *Political Psychology*, , 2024, p. pops.13017, [https//doi.org/10.1111/pops.13017].

Régression spatiale n.f [ʁe.gʁe.sjɔ̃ spa.sjal]

Synonymes : régression décalée spatialement, modèle d'économétrie spatiale

A quoi ça sert ?

Les régressions spatiales permettent d'intégrer comme variable la dimension spatiale dans un modèle de régression. C'est donc une méthode qui permet de traiter de la corrélation spatiale dans les modèles de régressions. Les régressions spatiales sont donc particulièrement indiquées lorsque que vous constatez l'existence d'autocorrélation spatiale dans les résidus d'une régression « classique » (par exemple, une **régression linéaire**), ce qui est très souvent le cas dès que les données étudiées ont une dimension spatiale ou si nous postulons une hypothèse sur un effet du spatial dans notre modèle de régression.

L'autocorrélation des résidus dans un modèle classique a pour conséquence que la valeur prise par notre variable dépendante (à expliquer) dans un lieu donné soit fonction des caractéristiques des lieux voisins. Pour traiter cela, la régression spatiale va donc intégrer dans les variables indépendantes du modèle les caractéristiques des voisins. Nous parlons des variables spatialement décalées. Ce décalage peut se faire sur la variable dépendante des voisins, les variables indépendantes ou les résidus. En fonction des caractéristiques du voisin qui seront décalés, les modèles à réaliser seront différents. Certains modèles de régressions spatiales combineront un décalage sur plusieurs de ces caractéristiques en même temps.

D'où ça vient ?

Il existe plusieurs types de modèles de régressions spatiales. Leur conception est directement héritée des hypothèses posées sur l'origine de l'autocorrélation des résidus d'un modèle. Selon Manski (1993) trois effets peuvent être à l'origine de cette corrélation. Il est à noter que Manski s'est appuyé sur l'étude de comportements sociaux pour théoriser ces trois causes. Nous allons donc retrouver :

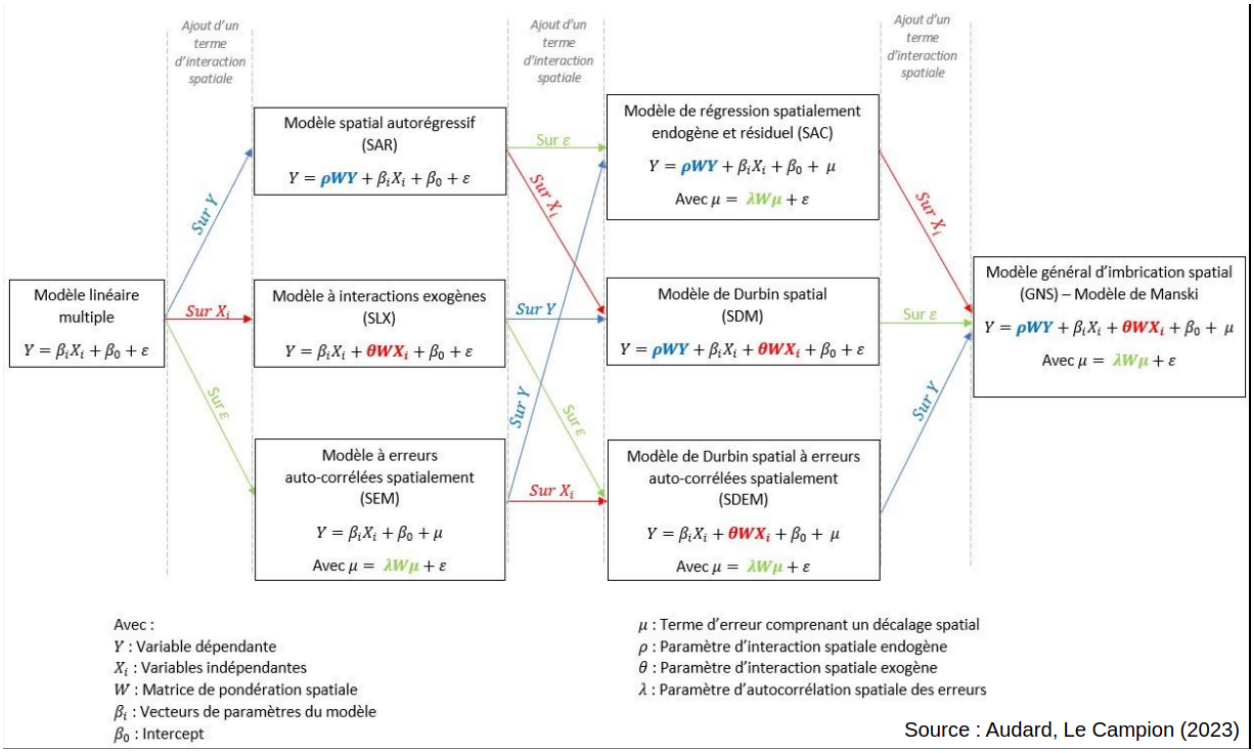
1- Les effets endogènes : la valeur de la *variable dépendante* (VD) prise en un lieu va dépendre de la valeur prise par la variable dépendante dans les lieux voisins. La variable spatialement décalée sera donc la VD. Par exemple, lors de l'étude de la participation électorale dans les bureaux de vote, cette participation dans un bureau de vote dépendra de la participation dans les bureaux de vote voisins. Dans ce cas le modèle approprié est un modèle spatial autorégressif ou modèle SAR, ou encore appelé modèle à variable endogène décalée. L'enjeu est que le paramètre décalé soit significativement différent de 0. Si c'est le cas, cela confirme la dépendance de la VD à son voisinage.

2- Les effets exogènes : la valeur de la variable dépendante prise dans un lieu va dépendre de la valeur prise par les *variables indépendantes* (VI) dans les lieux voisins, toutes choses égales par ailleurs. Par exemple, si nous étudions le prix de vente d'une maison en fonction de la présence d'un jardin et du nombre de chambres, le prix d'une maison va dépendre de la présence d'un jardin et/ou du nombre de chambres des maisons voisines. Il sera alors pertinent d'utiliser un modèle à variables exogènes décalés, aussi appelé à interactions exogènes et abrévié en SLX. L'enjeu est que le paramètre décalé soit significativement différent de 0. Si c'est le cas, cela confirme la dépendance de la VD à son voisinage.

3- Les effets corrélés : la valeur de la variable dépendante prise en un lieu va dépendre d'un facteur spatial inobservé qui influence simultanément l'observation et son voisinage, autrement dit va dépendre des résidus. Par exemple le prix de vente des maisons va dépendre en réalité de la proximité avec la rocade de la ville, variable non prise en compte dans le modèle mais qui se traduit dans les résidus. Le modèle le plus adapté à ce type de cas est le modèle à erreurs spatiales ou SEM[206]. L'enjeu est que le paramètre décalé soit significativement différent de 0. Les erreurs seront donc bien autocorrélées, ce qui implique que notre modèle souffre de l'omission d'une variable expliquant l'effet spatial.

Ces effets peuvent bien sûr se combiner et donner des modèles plus complexes.

Voici une figure représentant la galaxie des modèles de régressions spatiales.



Source : Audard, Le Campion (2022) d'après Elhorst (2010) et Lesage (2009)

Les trois modèles les plus utilisés dans la littérature sont les modèles SAR, SEM et le modèle spatial de Durbin (SDM) qui combine les effets exogènes et endogènes. Le modèle SDM est souvent considéré comme le plus performant, car il fournirait des estimateurs non biaisés parce qu'il tient compte des effets endogènes et exogènes (LeSage et Pace, 2009 ; Elhorst, 2010). Ceci dit ce modèle peut poser un problème de parcimonie. En effet, il sera de fait composé de nombreuses *variables indépendantes*. Toutes les VI initialement présentes dans le modèle, leurs formes décalées ainsi que la VD décalée.

Exemples d'application

Les exemples d'applications sont très nombreux, notamment en économie et en géographie, et sur des sujets très variés, allant de l'étude du lien entre le chômage et le marché de l'emploi, à la relation

[206] A ne pas confondre avec l'acronyme SEM qui renvoie auxmodèles en équations structurelles : Strucural Equations Modeling.

entre les prix immobiliers et les risques industriels, la végétalisation de l'espace urbain ou encore la diffusion des comportements sociaux.
Très souvent dans les différents articles, les différents modèles sont testés et celui avec les meilleures performances est retenu.
Ici, nous vous présenterons un exemple d'application avec un modèle SAR (Spatial lag regression en anglais), tiré de l'article d'Apparicio et al. (2016) qui étudie l'exposition des cyclistes montréalais à la pollution de l'air et au bruit induit par le traffic. Dans cet article, deux modèles ont été construits pour prédire les niveaux d'exposition à la pollution atmosphérique et au bruit.

Les modèles de régression « classique » (des moindres carrés ordinaires dans ce cas) calculés précédemment dans leurs travaux présentaient un problème de dépendance spatiale. Le choix de partir sur des régressions spatiales est donc cohérent. Le choix du modèle SAR se justifie par l'hypothèse des chercheurs que le niveau d'exposition au bruit ou à la pollution de l'air pourrait alors être associé au niveau observé durant les minutes précédentes et suivantes du trajet.

En résumé, les résultats des modèles de régression spatiale montrent que la température, l'humidité et le vent sont négativement associés à la pollution de l'air. Le type de piste ou de voie cyclable utilisé a également un impact sur les niveaux d'exposition à la pollution de l'air et au bruit. Enfin, il n'y a pas d'association significative (au niveau de 5%) entre les prédicteurs du milieu environnant - arbres, bâtiments, parc et mixité d'occupation du sol - et l'exposition à la pollution de l'air.

Le but de cet article est donc également de proposer des solutions pour aider à l'aménagement urbain de la ville de Montréal par exemple en plantant des haies entre les voies cyclables et la rue pour diminuer l'impact du bruit ou encore réserver certaine rue aux cycliste à certains horaires. Ces résultats n'auraient pas pu être mis en évidence avec des modèles de régression classique. La régression spatiale était donc particulièrement indiquée dans cet exemple.

Mot du praticien

La galaxie des différents modèles de régressions spatiales est riche et complexe. Tout l'enjeu est donc de choisir le modèle le plus adapté. Ce qui doit guider notre choix c'est le processus à l'origine de l'auto-corrélation spatiale à traiter dans notre modèle. Or l'identification de ce processus, même en se fiant aux bases théoriques, est très souvent complexe et multifactoriel. Ainsi, pour choisir notre modèle il est nécessaire de se baser sur des critères statistiques. Il est possible d'adopter une méthode ascendante (Florax et al, 2003) où l'idée est de partir du modèle non spatial et de tester la significativité des paramètres décalés pour ensuite choisir le meilleur modèle. Il est également possible de choisir la méthode descendante (Lesage et Pace, 2009) qui obéit à la même logique mais en partant du modèle de Durbin. Floch et Le Saout (2016) ont réalisé une synthèse de l'ensemble de ces questions.

La prise en compte des variables décalées permet de corriger le biais induit par l'autocorrélation spatiale, mais leur lecture peut s'avérer compliquée et surtout ne permet pas d'identifier le processus qui a généré cette autocorrélation spatiale. En effet, cette correction ne permet pas d'appréhender les facteurs locaux qui peuvent générer ces effets, pour cela il y a la **GWR**.

Aujourd'hui, de plus en plus de travaux incluent également une dimension temporelle à la régression spatiale en utilisant des modèles décalés par panel. Pour un bon exemple d'utilisation nous pouvons nous référer à l'article de Gaboriault-Boudreau et al (2019).

Ressources

- Apparicio Philippe, Carrier Mathieu, Gelb Jérémy, Séguin Anne-Marie et Kingham Simon, « Cyclists' exposure to air pollution and road traffic noise in central city neighbourhoods of Montreal », *Journal of Transport Geography*, vol. 57, 2016, p. 63-69, [https//doi.org/10.1016/j.jtrangeo.2016.09.014].
- Audard Frédéric et Le Campion Grégoire, *Initiation - FOrmation à la Statistique, Spatiale*, [https://letg.pages.in2p3.fr/initiation-formation-aux-statistiques-spatiales/ifoss_immo.html].
- Duvivier Chloé, Polèse Mario et Apparicio Philippe, « The location of information technology-led new economy jobs in cities: office parks or cool neighbourhoods? », *Regional Studies*, nº 6, vol. 52, 2018, p. 756-767, [https//doi.org/10.1080/00343404.2017.1322686].
- Elhorst J. Paul, « Applied Spatial Econometrics: Raising the Bar », *Spatial Economic Analysis*, nº 1, vol. 5, 2010, p. 9-28, [https//doi.org/10.1080/17421770903541772].
- Floch J. M. et Le Saout R., *Économétrie spatiale : une introduction pratique - Documents de travail - M2016/06* | *Insee*, [https://www.insee.fr/fr/statistiques/2408659].
- Florax Raymond J. G. M., Folmer Hendrik et Rey Sergio J., « Specification searches in spatial econometrics: the relevance of Hendry's methodology », *Regional Science and Urban Economics*, nº 5, vol. 33, 2003, p. 557-579, [https//doi.org/10.1016/S0166-0462(03)00002-4].
- Gaboriault-Boudreau Maxime, Apparicio Philippe et Brunelle Cédric, « Modélisation de la pauvreté urbaine dans la région métropolitaine de Montréal entre 1986 et 2016: Apport des régressions spatiales par panel », *Cahiers de géographie du Québec*, nº 179-180, vol. 63, 2019, p. 165, [https//doi.org/10.7202/1084230ar].
- Grislain-Letrémy Céline et Katossky Arthur, *Les risques industriels et le prix des logements − Économie et Statistique n° 460-461 - 2013* | *Insee*, [https://www.insee.fr/fr/statistiques/1377429?sommaire=1377437].
- LeSage James et Pace Robert Kelley, *Introduction to Spatial Econometrics*, 0 éd., Chapman and Hall/CRC, 2009, [https//doi.org/10.1201/9781420064254].
- Loonis Vincent et de Bellefon Marie-Pierre, *Manuel d'analyse spatiale* | *Insee*, [https://www.insee.fr/fr/information/3635442].
- Manski Charles F., « Identification of Endogenous Social Effects: The Reflection Problem », *The Review of Economic Studies*, nº 3, vol. 60, 1993, p. 531, [https//doi.org/10.2307/2298123].

SEM n.m. [/sɛm/]

Synonymes : Structural Equations Modeling, Modèles en équations structurelles, Modèles structuraux, MES

Cette entrée est volontairement très généraliste afin de rendre compte de la diversité de la famille des modèles SEM. Des exemples d'applications sont présentés dans les entrées **EFA - Analyses factorielles exploratoires**, **CFA - Analyses factorielles confirmatoires**, **modération**, **médiation** et **LPA - Latent Profil Analysis / LCA - Latent Class Analysis**.

A quoi ça sert ?

L'acronyme SEM pour les termes anglais *Structural Equations Modeling* désigne les modèles en équations structurelles. Ces modèles sont utilisés afin de rendre compte de relations entre des variables observées et des *variables latentes*. Par exemple, les psychologues peuvent être amenés à avoir besoin de mesurer le niveau d'anxiété de leurs patients. Or l'anxiété ne peut pas s'appréhender directement par la question : êtes-vous anxieux ? Oui ou non. Afin de rendre compte de la variable latente anxiété (non observée directement), les psychologues vont demander à leurs patients de répondre à un ensemble de questions (variables manifestes, directement mesurables, et sous l'influence de cette variable latente anxiété), tels que le degré d'accord avec les affirmations suivantes : "Je me sens indécis(e) ?" ou encore "Je me sens effrayé(e)", "Je me sens sûr(e) de moi"[207].

Les modèles SEM impliquent la formalisation d'un modèle théorique *a priori*. En effet, tout l'intérêt de ces modèles est de tester l'adéquation d'un modèle théorique aux données empiriques alors recueillies. Différents indices statistiques ont été développés afin de rendre compte de l'ajustement du modèle testé aux données présentées. Si cet ajustement est bon, nous pourrons alors étudier les relations entre les variables présentes. Ce second objectif porte sur des relations entre variables observées comme c'est le cas pour des modèles de **médiations** ou de **modérations** par exemple. Mais, et c'est là toute la force de ces modèles, il est également possible d'évaluer les relations entre des variables observées et des variables latentes, comme dans le cas d'**Analyses Factorielles Confirmatoires**, de modèles en pistes causales ou encore de classifications telles que les **LCA** Latent Class analysis ou les **LPA** (Latent profils analysis).

D'où ca vient ?

Les premiers modèles en équations structurelles ont été développés dans les années 1920 en biologie. On attribut la paternité de ces modèles au généticien Sewall Wright pour ses travaux en biométrie sur les analyses de parcours (path analysis) [208],[209]. L'objectif de l'analyse de parcours est de rendre compte de relations causales à partir des variations de variables supposées a priori sous l'influence de variables latentes. L'idée est que certains phénomènes ne sont pas mesurables

[207] Ces questions sont extraites de l'échelle STAI permettant d'évaluer l'anxiété, développée par Spielberger et al. en 1970 ; Spielberger Charles D., Gorsuch R. L. et Lushene R. E., « Manual for the State-Trait Anxiety Inventory (STAI) », , 1970.

[208] Wright Sewall, « The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs », *Proceedings of the National Academy of Sciences*, nº 6, vol. 6, 1920, p. 320-332, [https//doi.org/10.1073/pnas.6.6.320].

[209] Wright Sewall, « The Method of Path Coefficients », *The Annals of Mathematical Statistics*, nº 3, vol. 5, 1934, p. 161-215, [https//doi.org/10.1214/aoms/1177732676].

directement mais sont visibles partiellement via certains éléments qui eux sont mesurables. Par exemple, l'anxiété n'est pas appréhendable directement, par contre afin d'évaluer le niveau d'anxiété de leur patient et détecter d'éventuelles souffrances, les psychologues ont développé différents questionnaires abordant des éléments mesurables de l'anxiété tels que la difficulté de prendre des décisions ou la satisfaction de soi, etc. Ces éléments mesurables, manifestes sont influencés par la variable latente anxiété (qui n'est elle pas directement mesurable). En 1928, le père de Sewall Wright, Philip Wright va reprendre le modèle développé par son fils pour l'appliquer à l'économétrie [210]. De cette collaboration familiale émergera l'idée d'ajouter des variables explicatives, ce qui aboutira par la suite au développement des modèles en équations simultanées.

Ces modèles seront enrichis dans les années 1950 en sociologie grâce aux travaux de Lazarsfeld notamment[211], avec le développement des modèles causaux et leur limite théorique indiquant qu'une relation causale n'a de sens que dans le cadre d'une hypothèse formulée a priori. En 1970 en psychologie, Jöreskog intègre l'analyse factorielle à l'analyse de parcours et aux équations simultanées ce qui permet la création des **analyses factorielles confirmatoires**[212]. Les modèles continuent de se complexifier dans les années 1980 avec notamment les travaux de Bentler en Psychologie[213]. Ces travaux impliquent l'ajout de la possibilité de tester des effets indirects et un meilleur contrôle de l'erreur de mesure notamment. Depuis le début des années 2000, et l'avènement de l'informatique, les modèles structuraux se sont encore complexifiés afin de pouvoir appréhender statistiquement des modèles théoriques complexes, que ce soit dans le champ des **analyses factorielles** (Confirmatory factorial analysis), des **régressions** (Mediations / moderations, etc.), des **classifications** (Latent profils analysis - LPA, Latent Class analysis - LCA), des modèles longitudinaux (Growth models), SEM bayésiens, etc.

Mot du praticien

Les modèles SEM peuvent se réaliser avec des variables continues, ordinales et même catégorielles, mais les types de méthode de factorisation choisis seront différents.

La taille de l'échantillon est également un point de vigilance important. En effet, l'échantillon sélectionné ne devra être ni trop petit, ni trop grand. Un échantillon trop petit risque de ne pas détecter l'effet recherché, alors qu'un échantillon trop grand risque de détecter des effets peu intéressants, voir artefactuels. C'est un vrai travail d'équilibriste pour ne pas passer à côté d'un effet essentiel et ne pas gaspiller du temps et de l'argent à recruter un énorme échantillon pour faire ressortir des effets peu intéressants. Le chiffre de 200 / 300 observations est suggéré dans la littérature pour une bonne puissance statistique (avec environ 40 variables) ou encore l'idée d'un ratio de 10 sujets par variables, avec à partir de N=300 une diminution du ratio possible (Kyriazos, T. A., 2018)[214].

---

[210] Wright P. G., *The Tariff on Animal and Vegetable Oils*, Macmillan, 1928.

[211] Lazarsfeld Paul F., « Recent Developments in Latent Structure Analysis », *Sociometry*, n° 4, vol. 18, 1955, p. 391, [https//doi.org/10.2307/2785875].

[212] Jöreskog K. G., « A General Approach to Confirmatory Maximum Likelihood Factor Analysis », *Psychometrika*, n° 2, vol. 34, 1969, p. 183-202, [https//doi.org/10.1007/BF02289343].

[213] Bentler P. M. et Bonett Douglas G., « Significance tests and goodness of fit in the analysis of covariance structures. », *Psychological Bulletin*, n° 3, vol. 88, 1980, p. 588-606, [https//doi.org/10.1037/0033-2909.88.3.588].

[214] Kyriazos Theodoros A., « Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General », *Psychology*, n° 08, vol. 09, 2018, p. 2207-2230, [https//doi.org/10.4236/psych.2018.98126].

Ressources

- Bentler P. M. et Bonett Douglas G., « Significance tests and goodness of fit in the analysis of covariance structures. », *Psychological Bulletin*, nº 3, vol. 88, 1980, p. 588-606, [https//doi.org/10.1037/0033-2909.88.3.588].
- Jöreskog K. G., « A General Approach to Confirmatory Maximum Likelihood Factor Analysis », *Psychometrika*, nº 2, vol. 34, 1969, p. 183-202, [https//doi.org/10.1007/BF02289343].
- Kyriazos Theodoros A., « Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General », *Psychology*, nº 08, vol. 09, 2018, p. 2207-2230, [https//doi.org/10.4236/psych.2018.98126].
- Lazarsfeld Paul F., « Recent Developments in Latent Structure Analysis », *Sociometry*, nº 4, vol. 18, 1955, p. 391, [https//doi.org/10.2307/2785875].
- Mercklé Pierre, « Les méthodes d'équations structurelles (MES) : Pour qui ? Pour quoi faire ? Comment ça marche ? par Alain Lacroux (Vendredis Quanti, 31 janvier 2020) », , 2020, [https//doi.org/10.58079/T4CF].
- Pornprasertmanit Sunthud, Miller Patrick et Schoemann Alexander, *Simsem*, [https://simsem.org/].
- Rosseel Yves, « Lavaan: An R Package for Structural Equation Modeling », *Journal of Statistical Software*, vol. 48, 2012, p. 1-36, [https//doi.org/10.18637/jss.v048.i02].
- Spielberger Charles D., Gorsuch R. L. et Lushene R. E., « Manual for the State-Trait Anxiety Inventory (STAI) », , 1970.
- Tobacyk Jerome J., « A Revised Paranormal Belief Scale », *International Journal of Transpersonal Studies*, nº 1, vol. 23, 2004, p. 94-98, [https//doi.org/10.24972/ijts.2004.23.1.94].
- Whalley Ben, *Just Enough R*.
- Wright P. G., *The Tariff on Animal and Vegetable Oils*, Macmillan, 1928.
- Wright Sewall, « The Method of Path Coefficients », *The Annals of Mathematical Statistics*, nº 3, vol. 5, 1934, p. 161-215, [https//doi.org/10.1214/aoms/1177732676].
- Wright Sewall, « The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs », *Proceedings of the National Academy of Sciences*, nº 6, vol. 6, 1920, p. 320-332, [https//doi.org/10.1073/pnas.6.6.320].

Shapiro-Wilk n.m [/ʃəˈpiɹoʊ wɪlk/]

Synonymes : S-W, test de normalité

A quoi ça sert ?

Afin d'appliquer différents tests (notamment de comparaison de moyennes) nous devons nous assurer que les données suivent une *loi normale* pour chacun des groupes, pour l'ensemble d'une distribution ou pour la différence entre deux distributions issues de deux temps de mesures différents comme dans le cas d'un **test t** pour échantillons appariés par exemple. Ceci est nécessaire afin de s'assurer que la moyenne est bien un indicateur pertinent pour analyser les données concernées. Il s'agit en vérifiant que les données suivent une *loi normale*, de vérifier si la majorité des données est bien rassemblée autour de la valeur moyenne. Si ce n'est pas le cas, si les données sont par exemple réparties en deux pôles, alors la moyenne ne sera pas le meilleur indicateur pour rendre compte de la distribution des données étudiées. Cette vérification de la normalité des distributions peut se faire graphiquement ou par l'emploi de tests, tels que celui de **Shapiro-Wilk** ou de **Kolmogorov-Smirnov (K-S)**.

D'où ça vient ?

Le test de Shapiro-Wilk a été développé en 1965 par Samuel Sanford Shapiro et Martin Wilk (S. S. SHAPIRO, M. B. WILK, 1965)[215]. Il s'agit d'un test d'hypothèse permettant de voir si un échantillon est issu d'une population normalement distribuée ou non.

Mot du praticien

Il est basé sur la statistique W, comparé aux autres tests il est très puissant pour les petits effectifs (n=<50), mais plus faible sur les grands échantillons. Il est considéré comme très fiable car il ne sert qu'à mesurer la normalité d'une distribution. Il est souvent considéré comme le test de normalité de référence.

Ressources

- Shapiro S. S. et Wilk M. B., « An analysis of variance test for normality (complete samples) », *Biometrika*, nº 3-4, vol. 52, 1965, p. 591-611, [https//doi.org/10.1093/biomet/52.3-4.591].

[215] Shapiro S. S. et Wilk M. B., « An analysis of variance test for normality (complete samples) », *Biometrika*, nº 3-4, vol. 52, 1965, p. 591-611, [https//doi.org/10.1093/biomet/52.3-4.591].

SMA n.m [ɛs ɛm a]

Synonyme : Simulation multi-agents

A quoi ça sert ?

La simulation multi-agents ou SMA est utilisée pour modéliser et analyser des phénomènes sociaux complexes en étudiant les interactions entre des individus, des groupes et leur environnement.

Elle repose sur ce qu'on appelle des agents. Chacun des agents va posséder son propre comportement et prendre des « décisions » de manière autonome. Il va évoluer dans un modèle paramétré et interagir avec les autres agents et son environnement, ce qui va faire émerger des dynamiques collectives.

Il faut entendre ici les termes de agents, comportement ou interactions au sens informatique. En effet, un « agent », au sens informatique, doit être pensé comme un programme informatique (Wooldridge, 2012). Ce programme peut être simple ou compliqué (il peut s'agir d'un humain, d'une voiture ou d'une cellule du pancréas). La présence d'au moins deux agents va créer un « système multi-agents ». Dans le contexte de SMA, simuler signifie appeler de manière répétée chaque agent pour qu'il exécute les règles qui le définissent. Au fil de ces itérations, les résultats agrégés du comportement des agents et de leurs interactions peuvent être déterminés pas à pas et être réinjectés dans le comportement de ces mêmes agents.

La simulation multi-agents va donc être utilisée pour comprendre des systèmes complexes en particulier avec beaucoup d'entités, typiquement en géographie c'est l'étude des systèmes urbains ou des réseaux de transport. Cela va également permettre de modéliser des comportements émergents, comment des décisions individuelles de mobilités résidentielles peuvent conduire à des phénomènes d'étalement urbain (Enault, 2012). On pourra également bien sûr utiliser la SMA pour tester des hypothèses ou des scénarios et faire de l'aide à la décision, c'est notamment utilisé pour toutes les questions de gestion des risques environnementaux ou des situations de crise ou encore d'aménagement du territoire. Cela peut également être un bon outil pour intégrer une approche pluridisciplinaire et explorer des phénomènes difficiles à observer directement.

Ceci dit, bien que flexible et évolutive, au départ de la SMA, il aura fallu définir précisément un grand nombre de paramètres. Des règles qui concernent les agents avec leurs attributs, leurs comportements et leurs marges d'action possibles ainsi que l'environnement de la simulation qui peut être une grille, un espace continu statique ou dynamique.

Il existe donc un très grand nombre de modèles différents basés sur des théories différentes en fonction de la discipline qui les a développées, certains modèles sont basés sur des données empiriques et d'autres uniquement théoriques.

D'où ça vient ?

La simulation multi-agents est une méthode déjà ancienne qui émerge dès les années 1950 et 1960. D'abord développée dans le champ de l'informatique, très tôt elle va être utilisée dans les sciences humaines et sociales. Nous pouvons citer par exemple les travaux du sociologue français Raymond Boudon (1973) qui a étudié grâce à la SMA les inégalités scolaires entre les groupes sociaux, ou encore les travaux de Thomas C. Schelling, économiste américain qui recevra en 2005 un prix

Nobel, qui lui a utilisé la simulation multi agent pour démontrer que la ségrégation résidentielle peut apparaître au niveau systémique même si les préférences ethniques des acteurs, pris séparément, ne sont pas discriminatoires. Les travaux de Schelling sont centraux car ils vont introduire la notion de « voisinage » local, qui deviendra centrale dans l'approche multi-agents.

Néanmoins le vrai pionnier en la matière est le géographe suédois Torsten Hägerstrand, fondateur de ce qu'on appelle la *time géographie*. Dès 1965 (T. Hägerstrand, 1965), il publie un article où il réalise une simulation multi-agents, bien qu'il ne la nomme pas encore de cette manière, pour étudier la diffusion en Suède d'innovations agricoles. Depuis, de très nombreux modèles plus ou moins complexes ou spécialisés ont été développés.

Exemples d'application

Il existe de très nombreuses et diverses publications, notamment en géographie, qui utilisent des modèles développés pour la simulation multi-agents. Nous pouvons citer les travaux de Christopher Jankee et de ses collaborateurs (2020) qui ont utilisé la SMA et le modèle Simulation Aéroport Compagnie Aérienne et Territoire (SACAT) pour modéliser les négociations entre aéroports et compagnies aériennes. Ils vont étudier également le rôle que peuvent jouer les collectivités territoriales dans ces négociations et s'appuyer dans le modèle sur la théorie des jeux spatiaux et évolutionnaires.

Bien loin des aéroports, nous pouvons citer les travaux de Thibault Raffaillac et de ses collaborateurs (2023) qui ont développé une interface de SMA pour travailler et réfléchir la gestion concertée des territoires pastoraux de moyenne montagne. Cette interface a pour objectif de visualiser l'impact dans le temps de scénarios d'usage du territoire, et de représenter les dynamiques (p. ex. développement des forêts, présence touristique) à grandes échelles. Dans ces travaux la SMA sert en plus d'outil de base à l'échange et la discussion entre tous les acteurs impliqués dans la gestion et l'aménagement de ces territoires pastoraux.

Hägerstrand, dans sa publication, qui est l'un des premiers exemples d'usage de simulation en SHS, cherche à étudier la diffusion d'une innovation agricole en Suède. Cette innovation concerne la subvention à la pratique du pâturage en Suède. L'hypothèse d'Hägerstrand est que la diffusion de cette innovation dépend fortement d'un processus spatial. L'objectif de sa simulation va donc être de montrer le poids de la spatialité dans la propagation de l'innovation. La diffusion est ainsi un processus qui met en contact des émetteurs et des récepteurs, et il suffit de connaître les principaux canaux de circulation de l'information pour être en mesure de déterminer le déplacement de l'innovation dans l'espace. Partant de l'hypothèse que les contacts entre individus sont le principal vecteur d'information et de diffusion, Torsten Hägerstrand construit une grille d'interaction spatiale, le champ moyen d'information, et modélise le processus de communication et d'adoption des individus. Dans sa grille chaque cellule est habitée par un certain nombre d'agents. Ensuite, il postule que l'innovation se diffuse quand un « agent » ayant déjà adopté l'innovation en parle à un « agent » ne l'ayant pas encore fait. Selon son modèle, la probabilité que cet échange se vérifie dépend de la distance existant entre les cellules où les deux « agents » résident. Hägerstrand utilise des données sur les migrations pour calibrer cette probabilité). En itérant la sélection des partenaires de discussion sur la base de cette règle simple, Hägerstrand produit des formes de concentration spatiale de l'adoption qui ressemblent à la distribution suédoise réelle. Les travaux d'Hägerstrand sont moins connus que ceux de Schelling, mais leur importance est cependant considérable car ils illustrent la possibilité de mettre en relation une simulation multi-agents avec les données réelles, alors que le modèle de Schelling est complètement déconnecté de toute donnée empirique. Par

ailleurs, cette première simulation d'Hägerstrand reste très importante notamment en géographie où elle a été fondatrice et où de nombreux auteurs ont taché de la dupliqué et de l'informatisé (Saint-Julien, 1985, Daudé 2002 et 2004).

Nous trouvons donc de très nombreux exemples, anciens et récents, de publications et de communications scientifiques faisant usage de la simulation multi-agents.

Mot du praticien

Des travaux reposant sur la simulation multi-agents existent donc depuis longtemps en SHS, même s'il semble que la SMA a réussi à davantage se faire une place dans les méthodes d'analyse et de simulation en géographie. Cela tient peut-être à l'influence d'Hägerstrand, à la place que le spatial peut prendre dans la SMA ou la bonne synergie entre l'étude des questions de diffusion centrale en géographie et cette méthode de simulation multi-agents.

Ceci dit, la méthode SMA fait l'objet de certaines critiques. Outre des limites concernant la complexité de son utilisation et de la création des modèles, du coût computationnel de ces méthodes ou de la sensibilité de ces modèles, se posent plus largement des questions sur la dépendance des résultats à l'égard des valeurs initiales définies pour les agents et leur environnement. À cela s'ajoutent des questionnements sur la reproductibilité et la transparence des résultats.
Pour un aperçu complet, se référer au document de Gianluca Manzo publié dans la revue française de sociologie en 2014.

Ressources

- Boudon Raymond, *L'inégalité des chances : la mobilité sociale dans les sociétés industrielles*, Paris : A. Colin, 1973.
- Daudé Éric, « Apports de la simulation multi-agents à l'étude des processus de diffusion », *Cybergeo*, , 2004, [https//doi.org/10.4000/cybergeo.3835].
- Daudé Eric, *Modélisation de la diffusion d'innovations par la simulation multi-agents. L'exemple d'une innovation en milieu rural.*, thèse de doctorat, Université d'Avignon, 2002.
- Enault Cyril, « Simulation de l'étalement urbain de Dijon en 2030 : approche systémique de la dynamique gravitaire ville-transport », *Cybergeo*, , 2012, [https//doi.org/10.4000/cybergeo.25157].
- Hägerstrand T., *The Propagation of Innovation Waves*, Royal University of Lund, Department of Geography, 1952.
- Hägerstrand Torsten, « A Monte Carlo Approach to Diffusion », *European Journal of Sociology / Archives Européennes de Sociologie*, n° 1, vol. 6, 1965, p. 43-67, [https//doi.org/10.1017/S0003975600001132].
- Jankee Christopher, Carrard Michel, Verel Sébastien et Ramat Éric, « Les conséquences de la réforme aéroportuaire pour les territoires : apports d'une simulation informatique multi-agents », *Cybergeo*, , 2020, [https//doi.org/10.4000/cybergeo.35537].
- Manzo Gianluca, « Potentialités et limites de la simulation multi-agents : une introduction »:, *Revue française de sociologie*, n° 4, Vol. 55, 2014, p. 653-688, [https//doi.org/10.3917/rfs.554.0653].
- Raffaillac Thibault, Boukhelifa Nadia, Crouzat Emilie, Stark Fabien, Müller Jean-Pierre et Lasseur Jacques, « Développement d'une interface de simulation multi-agents pour la gestion concertée des territoires pastoraux de moyenne montagne », Troyes, France, 2023.

- Saint-Julien Thérèse, *La diffusion spatiale des innovations*, Montpellier, Gip Reclus, coll. « Reclus modes d'emploi », 1985.
- Schelling Thomas C., « Dynamic models of segregation† », *The Journal of Mathematical Sociology*, nº 2, vol. 1, 1971, p. 143-186, [https//doi.org/10.1080/0022250X.1971.9989794].
- Wooldridge Michael J., *An introduction to multiagent systems*, 2. ed., repr (1st ed. 2009)., Chichester, Wiley, 2012.

Sphéricité n.f [/sfe.ʁi.si.te/]

Synonymes : Test de sphéricité, Test de Mauchly

A quoi ça sert ?

Le test de sphéricité est utilisé dans le cas de mesures répétées. Afin de comparer des moyennes, par exemple dans le cas d'une analyse de variance (**ANOVA**), il est nécessaire de vérifier l'homogénéité des variances des différences du score entre les conditions (par exemple entre des temps de mesure), il s'agit donc de la **sphéricité des données**. L'objectif est de regarder la variation des scores, chez les mêmes sujets, quand on change de temps de mesure. Nous souhaitons donc que les variances de ces changements soientt homogènes d'une condition à l'autre, d'un temps à l'autre. Ceci est garant d'une comparaison possible entre les conditions (ou temps de mesure). Avec deux conditions, il n'y aurait qu'une seule différence possible (entre T1 et T2), donc il n'y a aucun besoin de vérifier la sphéricité des données (il n'y aurait pas d'autre différence calculable pour vérifier si sa variance est du même ordre que la première). La sphéricité se calcule donc à partir de 3 conditions afin de comparer les 2 variances des différences. Le nombre de condition peut bien sûr être plus important.

D'où ça vient ?

Le test de sphércité aussi appelé test de Mauchly a été développé en 1940 par John Mauchly, (Mauchly, J. W., 1940)[216]. Le F peut cependant être interprété même si la condition de sphéricité n'est pas remplie. Toutefois, l'interpréter sans correction augmente le risque d'obtenir des faux positifs. Afin de corriger cet écueil, Greenhouse et Geisser en 1959 [217] et Huynh et Feldt en 1976 [218] ont développé des corrections permettant d'ajuster le seuil de la p-value et de ne pas produire de faux positifs.

Ressources

- Greenhouse Samuel W. et Geisser Seymour, « On Methods in the Analysis of Profile Data », *Psychometrika*, nº 2, vol. 24, 1959, p. 95-112, [https//doi.org/10.1007/BF02289823].
- Huynh Huynh et Feldt Leonard S., « Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs », *Journal of Educational Statistics*, nº 1, vol. 1, 1976, p. 69, [https//doi.org/10.2307/1164736].
- Mauchly John W., « Significance Test for Sphericity of a Normal n-Variate Distribution », *The Annals of Mathematical Statistics*, nº 2, vol. 11, 1940, p. 204-209, [https//doi.org/10.1214/aoms/1177731915].

[216] Mauchly John W., « Significance Test for Sphericity of a Normal n-Variate Distribution », *The Annals of Mathematical Statistics*, nº 2, vol. 11, 1940, p. 204-209, [https//doi.org/10.1214/aoms/1177731915].

[217] Greenhouse Samuel W. et Geisser Seymour, « On Methods in the Analysis of Profile Data », *Psychometrika*, nº 2, vol. 24, 1959, p. 95-112, [https//doi.org/10.1007/BF02289823].

[218] Huynh Huynh et Feldt Leonard S., « Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs », *Journal of Educational Statistics*, nº 1, vol. 1, 1976, p. 69, [https//doi.org/10.2307/1164736].

Textométrie n.f. [/tɛk.sto.me.tʁi/]

Synonymes : appelée d'abord lexicométrie, la discipline a vu son nom évoluer en textométrie ou logométrie au début des années 2000.

A quoi ça sert ?

Elle permet d'analyser des données textuelles, d'en extraire des informations pertinentes grâce à des méthodes statistiques qui explorent à la fois la structure et le contenu des textes. Par ailleurs, les éléments contextuels (métadonnées) jouent un rôle essentiel dans les traitements textométriques.

Nous parlons de **données textuelles** pour les distinguer des données quantitatives ou qualitatives. Parmi les domaines d'applications les plus courants, l'étude de textes qui de par leur forme est difficile à recoder en variables quantitatives ou qualitatives si ce n'est en neutralisant des spécificités ou différences qui s'avèrent significatives : bases bibliographiques, bases juridiques, archives historiques, questions ouvertes des questionnaires, etc.

La textométrie permet aussi d'analyser des corpus multilingues sans avoir besoin de les traduire au préalable. C'est une méthode qui reste proche des textes et qui met en valeur le langage d'où son intérêt pour les Sciences Humaines et Sociales.

Les méthodes statistiques appliquées en textométrie sont donc nombreuses. Parmi les principales méthodes, nous trouvons :

- Analyse lexicale et fréquentielle (statistiques de fréquences, analyses des spécificités)
- Analyse factorielle et classification (AFC, CAH)
- Modélisation thématique et sémantique (topic modeling, word embeddings)
- Régression et sélection de variables (régressions Lasso et Ridge, Random Forest et SVM)
- Analyses de réseaux (graphes lexicaux, analyse de cooccurrences)

Pour s'y retrouver, il importe d'être au clair sur les objectifs de l'analyse. Et à chaque méthode, ses logiciels ou packages !

D'où ça vient ?

Son histoire a évolué grâce aux contributions de la linguistique, des statistiques et de l'informatique à partir des années 70 en France. La discipline reprend les recherches des linguistes Pierre Guiraud (1954, 1960) et Charles Muller (1968, 1977) en statistique lexicale (évaluation de la richesse du vocabulaire d'un texte, vocabulaire caractéristique d'un texte) mais aussi les méthodes d'analyse des données (analyses factorielles, classifications) mises au point par les statisticiens tels que notamment ceux de Jean-Paul Benzécri (1973)[219]. Le développement des capacités de calcul des ordinateurs a favorisé la diffusion de cette approche statistique du texte.

Ainsi les textes littéraires des grands auteurs ont progressivement cédé la place aux données commerciales, sociologiques, politiques ou financières.

[219] PINCEMIN Bénédicte et HEIDEN Serge, *Qu'est-ce que la textométrie ? Présentation*, [https://txm.gitpages.huma-num.fr/textometrie/Introduction/].

Mot du praticien

La partie logicielle de la méthode textométrique, permet la mémorisation de vastes ensembles de textes et l'efficacité des calculs mais le principe même de la méthode consiste à connaître son corpus pour pouvoir interpréter les résultats obtenus.

Par exemple, dans le cadre de l'analyse textuelle, nous pouvons nous intéresser au dépouillement de corpus particuliers, que constituent les réponses aux questions ouvertes dans les enquêtes socio-économiques, par des méthodes statistiques multidimensionnelles telles que :

- Analyse des correspondances
- Classification
- Caractérisation lexicale
- Analyse factorielle multiple

Analyse des correspondances de données textuelles

Cette méthode permet d'étudier et de visualiser les proximités, les attractions et les répulsions entre documents, entre mots, mais aussi entre documents et mots.
On aura par exemple :

- Des réponses proches à une question si les répondants privilégient ou évitent les mêmes mots
- Des mots proches s'ils se répartissent de la même façon entre les différents répondants
- Les réponses d'une catégorie d'individus (les femmes par ex.) et un mot s'attirent/se repoussent si cette catégorie sur-utilise /sous-utilise ce mot comparé à la moyenne de nos répondants.

L'AC (Analyse des correspondances) est basée sur l'étude du tableau lexical où nous retiendrons les mots suffisamment fréquents pour que la comparaison entre documents ait un sens du point de vue statistique. Elle fournit une représentation simplifiée du Tableau lexical analysé sur un petit nombre d'axes factoriels sans rien perdre d'essentiel. L'ensemble des documents et mots est ainsi cartographié pour dévoiler la structure des données. Les données initiales (les fréquences brutes du tableau lexical) sont centrées et réduites afin de ne pas biaiser les résultats par les mots trop fréquents ou les documents trop longs. Le calcul d'inertie, basé sur la distance du khi-2, mesure l'écart entre la fréquence observée d'un mot dans un document et la fréquence attendue. L'AC cherche à maximiser cette inertie en identifiant des axes factoriels qui expliquent le mieux les variations lexicales dans le corpus. Les graphiques sont validés avec la méthode des bootstrap (Efron, 1979) qui permet d'associer des ellipses de confiance à chaque élément.

Classification en analyse textuelle après une analyse des correspondances

La classification consiste à regrouper les documents en classes non définies à l'avance, de manière à ce que les profils lexicaux des documents d'une même classe soient proches, tout en étant les plus distincts possible d'une classe à l'autre. En phase exploratoire notamment, l'utilisation conjointe de l'AC et d'une méthode de classification ascendante hiérarchique fournit des résultats qui s'enrichissent mutuellement.

Plusieurs types de démarche peuvent être adoptées (descendante, ascendante, hiérarchique, non hiérarchique, supervisée, etc.). En fonction de notre objectif, le mode de constitution des classes

pourra différer. En phase exploratoire d'un corpus de questions ouvertes (enquête par questionnaire), on pourra privilégier l'utilisation conjointe de l'AC et de la classification ascendante hiérarchique (CAH). L'AC servira à structurer l'espace lexical, qui sera ensuite exploité par la CAH pour regrouper les documents de manière pertinente. Les oppositions mises en évidence sur les axes par l'AC deviennent plus claires quand les classes sont représentés sur les plans factoriels.

Au départ, chaque document constitue une classe à part puis on regroupe les deux classes les plus proches à chaque étape jusqu'à la plus grossière avec une seule classe qui réunit tous les documents. Le critère de Ward (minimisation de la variance intra-classe) permet de déterminer les fusions de classes afin qu'elles soient suffisamment homogènes quant à leur vocabulaire. Nous obtenons un arbre hiérarchique (dendogramme) qui montre comment et à quel niveau les documents se regroupent. Nous choisissons un seuil de coupure dans le dendrogramme pour définir la partition la plus pertinente pour notre corpus.

Caractérisation lexicale des parties du corpus après une analyse des correspondances et une classification automatique

Les résultats de l'AC et de la CAH peuvent être enrichis par l'identification des mots sous ou sur-utilisés caractérisant chacune des classes en comparaison avec le corpus dans son ensemble. Cela permet aussi d'identifier quelques documents (les parangons) que nous pouvons considérer comme caractéristiques de leur classe.

Les corpus de réponses ouvertes se prêtent tout particulièrement à cette approche. Avec des réponses/documents usuellement courts, rechercher les documents les caractérisant à un sens. A noter, la classification peut être opérée à partir du contenu verbal des documents mais aussi à partir des variables contextuelles. Les documents d'une même classe auront un contenu lexical similaire ou partageront des caractéristiques contextuelles communes. Les mots caractéristiques de chacune des classes sont reproduits sous forme de tableaux avec pour chacun, le pourcentage des occurrences dans la classe, le pourcentage des occurrences dans l'ensemble du corpus, la fréquence du mot dans la classe, la fréquence du mot dans l'ensemble du corpus, la probabilité critique pour le test du Khi-2 et la valeur-test (traduction de la probabilité).

Analyse factorielle multiple pour tableaux de contingence dans le cadre de l'analyse textuelle (AFMTC)

Cette méthode peut être utilisée pour aborder diverses questions qui se présentent dans une enquête socio-économique. Quelques exemples : l'analyse simultanée de questions ouvertes mais aussi de questions ouvertes et fermées ou même la comparaison de l'évolution d'une réponse dans le temps (y compris dans différentes langues).

L'AFMTC permet d'analyser simultanément un ensemble de tableaux de contingence ayant des lignes-documents homologues. Nous pouvons aussi y ajouter des tableaux d'autres variables. Cette technique peut être une étape préalable pour regrouper des termes apparaissant dans des contextes similaires dans des analyses thématiques ou pour construire des cartes sémantiques. L'idée est de partir d'un tableau multiple, juxtaposant plusieurs tableaux de contingence correspondant à un ensemble de documents (en lignes) décrits par plusieurs groupes de mots (en colonne). Chacun des tableaux est recentré sur ses propres marges ce qui permet de décrire les écarts à l'indépendance intra-tableaux. Les colonnes-mots sont surpondérées pour équilibrer l'importance de chacun des groupes dans la détermination du premier axe global. L'interprétation

passe par la recherche des mots les plus contributifs à l'inertie des axes. Nous déterminons les mots qui s'opposent sur les premiers axes comme dans le cas d'une analyse des correspondances. Des mots apparaissant dans des contextes similaires, même s'ils ne sont pas synonymes, se distinguent et sont considérés comme synonymes distributionnels.

Ressources

- Bécue Bertaut Monique, *Analyse de textuelle avec R*, Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2018.
- Benzécri Jean-Paul, *L'Analyse des données*, Paris Bruxelles Montréal, Dunod, 1973.
- Efron B., « Bootstrap Methods: Another Look at the Jackknife », *The Annals of Statistics*, nº 1, vol. 7, 1979, [https//doi.org/10.1214/aos/1176344552].
- Gentzkow Matthew, Kelly Bryan et Taddy Matt, « Text as Data », *Journal of Economic Literature*, nº 3, vol. 57, 2019, p. 535-574, [https//doi.org/10.1257/jel.20181020].
- Guiraud P., *Les caractères statistiques du vocabulaire: essai de méthodologie*, Presses universitaires de France, 1954.
- Guiraud P., *Problèmes et méthodes de la statistique linguistique*, Presses universitaires de France, 1960.
- Heiden Serge, Magué Jean-Philippe et Pincemin Bénédicte, « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement », Edizioni Universitarie di Lettere Economia Diritto.
- Lebart L. et Salem A., *Statistique Textuelle*, [http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html].
- Lebart Ludovic, Pincemin Bénédicte et Poudat Céline, *Analyse des données textuelles*, Presses de l'Université du Québec, coll. « Mesure et évaluation », 2019.
- Lexicometrica Team, *JADT | Journées d'Analyses statistiques de Données Textuelles*, [http://lexicometrica.univ-paris3.fr/jadt/].
- Marchand Pascal et Ratinaud Pierre, *Faut-il faire des nuages de mots ? — IRaMuTeQ*, [http://www.iramuteq.org/Members/pmarchand/faut-il-faire-des-nuages-de-mots].
- Muller C., *Initiation à la statistique linguistique*, Larousse, 1968.
- Muller C., *Principes et méthodes de statistique lexicale*, Hachette, 1977.
- Pincemin Bénédicte, « La textométrie en question », *Le Français Moderne - Revue de linguistique Française*, nº 1, vol. 88, 2020, p. 26.
- PINCEMIN Bénédicte et HEIDEN Serge, *Qu'est-ce que la textométrie ? Présentation*, [https://txm.gitpages.huma-num.fr/textometrie/Introduction/].

T-test (ou t de Student) n.m [/tiː tɛst/]

Synonymes : Test de Student, Test t

A quoi ça sert ?

Dans la recherche en SHS il peut s'avérer utile de comparer les **moyennes** de deux groupes d'individus. Par exemple, pour comparer le salaire moyen des hommes et des femmes ou encore évaluer quel support de diffusion de l'information, écrit ou audio, permet une meilleure mémorisation des participants. Le test t ou test t de Student permet de comparer deux moyennes, afin de vérifier si l'écart observé est statistiquement significatif ou non.

Le t de Student peut être utilisé à partir d'un échantillon unique – par exemple pour comparer la moyenne d'un échantillon avec une valeur constante, qui peut être issue d'une population générale ou d'un sous-échantillon plus restreint. Il peut être appliqué à des échantillons indépendants (comparer les moyennes issues d'échantillons de population différentes), ou à des échantillons appariés (comparer les moyennes de deux échantillons issus d'une même population).

D'où ça vient ?

Student est le pseudonyme utilisé par William Gosset, statisticien anglais employé par la brasserie Guinness pour améliorer la stabilité de la bière grâce à ses mesures statistiques[220]. Il a par ailleurs développé en 1908 une loi qui porte son nom, la loi de Student à ne pas confondre avec le présent test[221].

Mot du praticien

Avant de réaliser un t de student il faut s'assurer que la moyenne de notre échantillon soit bien représentative de la distribution de la variable, qu'il faudra alors examiner. Si la distribution s'approche de la loi normale, alors la majorité des valeurs sont regroupées autour de la valeur moyenne : la moyenne est donc bien représentative de l'échantillon. Si la distribution de la variable s'écarte de la loi normale, il se peut que celle-ci soit altérée par une (ou plusieurs) valeur extrême qu'il faut corriger pour que la valeur moyenne devienne représentative de l'échantillon. Il est aussi possible que la moyenne ne soit pas l'indicateur le plus adapté pour rendre compte des données, et dans ce cas, il faut se tourner vers un autre type de test tel que celui de **Wilcoxon** par exemple.
Pour les tests t pour échantillons indépendants, il est également nécessaire de s'assurer que les variances des deux groupes sont homogènes (la répartition des valeurs autour de la valeur moyenne est dans les mêmes proportions). Afin de vérifier si les variances sont homogènes (nous parlons aussi de variances égales ou **homoscédasticité**), nous employons le test de Lévène.

Le test t pour échantillon unique peut être employé afin de comparer la moyenne d'une population donnée avec une mesure de référence. Par exemple, il peut être utilisé pour comparer la moyenne d'âge d'un échantillon constitué dans le cadre d'une étude menée aux Etats-Unis et la moyenne d'âge de la population américaine en 2023. Cette comparaison va permettre de vérifier si la moyenne

[220] Boland Philip J., « A Biographical Glimpse of William Sealy Gosset », *The American Statistician*, nº 3, vol. 38, 1984, p. 179-183, [https//doi.org/10.1080/00031305.1984.10483195].
[221] Student, « The Probable Error of a Mean », *Biometrika*, nº 1, vol. 6, 1908, p. 1, [https//doi.org/10.2307/2331554].

d'âge de la population de l'échantillon collecté pour cette étude est similaire à celle de l'ensemble de la population américaine, et donc si l'échantillon est bien représentatif de la population américaine sur ce critère de l'âge.

Le test t pour échantillons indépendants - c'est-à-dire constitué de deux groupes différents et exclusifs - permet de comparer les deux moyennes d'une variable relevées dans chacun des deux groupes. Par exemple, il peut être utilisé afin de comparer la moyenne du score de confiance envers la Science dans deux groupes distingués en fonction de leur tendance politique (Démocrates et Républicains).

Toutefois, si le test t renvoie une différence significative entre les moyennes des deux groupes il sera intéressant de quantifier l'importance de l'effet observé. En effet, La différence observée est-elle statistiquement importante ou non ? Pour répondre à cette question, il va être nécessaire de réaliser un test permettant de mesurer la taille d'effet, tel que le d de Cohen pour mesurer la taille de l'effet observé. Calculer la taille de l'effet observé est tout aussi important que de calculer la présence ou l'absence de l'effet. Car l'effet observé représente-t-il un effet véritablement important ou s'agit-il simplement d'un artefact dû à la taille de l'échantillon ? En effet, plus un échantillon est grand, plus de petis effets peuvent être observés, mais cela ne signifie pas nécessairement qu'ils sont d'une ampleur significative. Il est donc important d'évaluer la taille de l'effet observé en calculant un indice de taille de l'effet observé (taille d'effet). Lorsque nous calculons un test de la famille des tests t, l'indice de taille d'effet le plus couramment utilisé est le d de Cohen [222].

Le test t pour échantillons appariés permet de comparer deux moyennes d'une variable relevée au sein d'une même population sur deux éléments différents. Ce peut aussi être un même élément mais à deux temps de mesures différents. Ce peut également être un élément mesuré au même moment mais dans des conditions différentes auprès d'une même population. Par exemple il peut être employé afin de comparer les réactions des participants avant, pendant et après un évènement marquant tel que la pandémie de COVID-19 ou encore l'élection présidentielle américaine de 2020.

Le t de Student, ou t test, n'est pas un test statistique ayant une complexité très importante. Le t de Student permet de comparer des moyennes 2 à 2, dans le cadre d'échantillons uniques, indépendants ou appariés. En Psychologie, il est très souvent utilisé afin d'obtenir de premiers éléments concernant les données alors récoltées en vue de répondre aux hypothèses émises. Ils sont souvent utilisés avant de réaliser des tests statistiques impliquant des modélisations plus complexes. Mais ils ne sont pas toujours utilisés à des fins descriptives. Les tests de comparaison de moyennes (tout comme les ANOVA) sont aussi utilisés comme résultats finaux d'études.

Ressources

- Boland Philip J., « A Biographical Glimpse of William Sealy Gosset », *The American Statistician*, nº 3, vol. 38, 1984, p. 179-183, [https//doi.org/10.1080/00031305.1984.10483195].
- Cohen Jacob, *Statistical Power Analysis for the Behavioral Sciences*, 0 éd., Routledge, 2013, [https//doi.org/10.4324/9780203771587].

[222] Cohen Jacob, *Statistical Power Analysis for the Behavioral Sciences*, 0 éd., Routledge, 2013, [https//doi.org/10.4324/9780203771587].

- Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.
- Student, « The Probable Error of a Mean », *Biometrika*, n° 1, vol. 6, 1908, p. 1, [https//doi.org/10.2307/2331554].

Tolérance n. f. [/ˈtɒlərəns/]

Synonymes : TOL, tolérance, tolerance, multicolinéarité.

A quoi ça sert ?

La présence de multicolinéarité signifie que la part de variance partagée entre différentes variables explicatives (variables indépendantes) est trop importante. Il y a donc une redondance d'information dans notre modèle et donc il y a un risque de surreprésentation d'un phénomène. Ceci qui va fausser les résultats du test multivarié réalisé. Afin de vérifier la multicolinéarité des résidus différents tests existent tels que le **Variance Inflation Factor (VIF)**, ou encore la tolérance, (Farrar & Glauber, 1967).[223]

D'où ça vient ?

Dans la littérature, il n'est pas simple de retrouver le concepteur à l'origine du test de la Tolérance, celui-ci ayant été cité et développé conjointement par des statisticiens et des économistes. Ces développements ont eu lieu suite aux travaux de Ragnar Frisch en 1934 sur la multicolinéarité et les impacts de celle-ci sur les **modèles de régression linéaires** notamment[224].

Mot du praticien

La tolérance est utilisée afin de vérifier qu'il y a bien une absence de multicolinéarité entre les variables explicatives sélectionnées dans le cadre d'une analyse multivariée tel que des modèles de **régression linéaire.** Dans ce cadre, un test tel que la tolérence va permettre de mettre en évidence sur les variables explicatives partagent une part de variance trop importante ou non. Par exemple nous pouvons envisager un modèle visant à expliquer un taux d'adhésion aux mesures sanitaire dans le cadre de la pandémie de COVID-19. Les variables explicatives étant par exemple, la confiance envers la science, la conformité sociale et la croyance dans des phénomènes paranormaux. Si ces variables explicatives partagent une part de variance commune trop importante risquant de biaiser le modèle, il sera alors nécessaire de sortir l'une de ces variable du modèle et de tester à nouveau la multicolinéarité jusqu'à ce que celle-ci soit satisfaisante. Deux variables (ou plus) partageant trop de variance peuvent en effet surreprésenter un phénomène et fausser les résultats. En SHS, il est une pratique classique d'éviter de mettre dans un même modèle le niveau de diplôme et la catégorie socioprofessionnelle des individus par exemple, ou la dépression et l'estime de soit.

Tout comme pour le **VIF**, il n'y a pas de consensus sur le seuil à conserver concernant la Tolérance et l'identification de la présence de multicolinéarité ou non.

Ressources

- Farrar Donald E. et Glauber Robert R., « Multicollinearity in Regression Analysis: The Problem Revisited », *The Review of Economics and Statistics*, nº 1, vol. 49, 1967, p. 92, [https//doi.org/10.2307/1937887].

[223] Farrar Donald E. et Glauber Robert R., « Multicollinearity in Regression Analysis: The Problem Revisited », *The Review of Economics and Statistics*, nº 1, vol. 49, 1967, p. 92, [https//doi.org/10.2307/1937887].

[224] Frisch Ragnar, *Statistical Confluence Analysis By Means Of Complete Regression System*, 1934.

- Frisch Ragnar, *Statistical Confluence Analysis By Means Of Complete Regression System*, 1934.

T-SNE n.f. [te ɛs ɛn ø]

Synonymes : t-distributed Stochastic Neighbor Embedding

A quoi ça sert ?

L'algorithme t-sne, au même titre que l'algorithme UMAP, fait partie des techniques de machine learning non supervisé qui appartiennent au champ de la réduction de dimension. La méthode la plus courante et la plus connue de réduction de dimension en SHS est l'analyse en composante principale.

Les jeux de données réels sont souvent non-structurées et de grande dimension (images, sons, séries temporelles…). La réduction de dimension a donc deux intérêts principaux : visualiser la structure du nuage des observations et simplifier les données pour des traitements ultérieurs (par ex., classification).

En effet, les méthodes de réduction de dimension visent à simplifier l'étude de points à grande dimension, en mettant à jour les principales caractéristiques de leur organisation. Autrement dit, on va chercher à faire émerger les principales dimensions qui structurent notre base de données d'individus statistiques caractérisés par un grand nombre de variables. Une des forces de ces différentes méthodes est de proposer une représentation lisible en deux dimensions de ce nuage complexe et multidimensionnel que forment nos données.

Pour ce faire, l'ACP va produire de nouvelles variables (les composantes principales) en réalisant une combinaison linéaire des variables existantes, c'est donc une méthode linéaire. En cela, elle peut donc être inadaptée à des données en SHS qui peuvent avoir une nature non linéaire, ce qui arrive en réalité assez souvent. L'algorithme T-SNE va constituer une méthode alternative à l'ACP. Il s'agit d'une méthode probabiliste non linéaire qui cherche à conserver la proximité, la similarité entre nos individus statistiques. Deux points proches dans l'espace d'origine doivent être proches dans l'espace à plus faible dimension.

En revanche, dans la méthode t-sne la distance exacte entre les individus ne sera pas conservée comme c'est le cas dans l'ACP.

Le principe de base de t-SNE est de calculer des distributions de probabilité qui mesurent la similarité entre les points dans l'espace d'origine. Pour chaque point, l'algorithme estime la probabilité qu'un autre point soit son voisin, en utilisant une distribution gaussienne centrée sur ce point. Dans l'espace réduit, t-SNE va alors utiliser une distribution t de Student pour modéliser les distances entre points. Cette distribution a des queues plus épaisses que la distribution gaussienne, ce qui aide à mieux séparer les clusters et à éviter le phénomène de "crowding", où des points éloignés seraient artificiellement rapprochés. Enfin, t-SNE utilise une optimisation itérative pour minimiser la différence entre les distributions de probabilité dans l'espace d'origine et celles de l'espace réduit. Cela permet d'ajuster progressivement la position des points dans l'espace réduit pour mieux représenter les relations d'origine.

t-SNE va donc être particulièrement efficace pour créer des visualisations où les clusters sont bien séparés. Cela peut être utile lorsque l'objectif principal est d'explorer visuellement les relations entre différentes classes ou groupes au sein de nos données.

Il est à noter que l'interprétation de la représentation de t-SNE ne peut pas se faire en interprétant les axes de la figure qui sont en réalité ininterprétable. Il faudra faire appel aux variables associés à nos points où réaliser une analyse complémentaire (classification) pour comprendre ce qui fait les groupes.

D'où ça vient ?

L'algorithme t-SNE, a été développé en 2008 par Geoffrey Hinton et Laurens van der Maaten. Il est donc assez récent mais s'est rapidement diffusée dans différentes disciplines. L'algorithme t-SNE a été utilisée pour de nombreuses applications : traitement automatique de la langue (similarité sémantique entre les mots), analyse de la musique, recherches médicales, bioinformatique, le traitement de signaux, en géographie, etc. Cette méthode est souvent utilisée pour la visualisation de représentations de haut-niveau apprises par un réseau de neurones artificiel.

Exemples d'application

Nous retrouvons depuis quelques années des exemples de publications ayant utilisés l'algorithme t-SNE, dans différentes disciplines, et notamment en géographie des risques, environnementale, etc.
L'exemple présenté ici est un usage particulier de l'algorithme t-SNE en géographie, il a été récompensé en 2019 du grand prix dans le cadre du concours « World Data Visualization Prize »[225]. Les auteurs ont choisi d'utiliser l'algorithme t-SNE pour représenter d'une autre manière les proximités et similarités entre pays. Cet exemple n'est pas issu d'une publication scientifique, mais il illustre le potentiel que peut représenter cette méthode en géographie, et plus largement en SHS.

A partir de la méthode t-SNE, l'objectif des auteurs était de proposer une cartographie alternative des pays du monde fondée sur l'analyse des donnée et l'étude d'une éventuelle similarité entre les pays. Les données provenaient du site « Information is beautiful » et du « World Government Summit ». Pour chaque pays, on retrouve des informations comme la population, l'indice de développement humain, le PIB, l'indice de GINI, la surface, l'espérance de vie. L'algorithme va donc réaliser des clusters de pays en tenant compte de l'ensemble de ces variables. On voit rapidement émerger des clusters, mais t-SNE n'indique pas ce qui fait la proximité/similarité des points regroupés. C'est donc à vous de vous reposer sur vos hypothèses, sur la littérature, sur vos données ou sur des analyses complémentaires pour déterminer ce qui fait ces rapprochements.

Dans cet exemple, à partir de ces données, on retrouve un cluster regroupant ce qu'on appelle les pays développés, un autre composé de la Chine et de l'Inde, qui ont en commun d'être les pays les plus peuplés. On retrouve également des clusters plus compliqués à interpréter ou basés sur des similarités qui semble moins pertinentes (rapprochement du Brésil et de la Russie notamment à cause de leur superficie).

Ainsi, les auteurs, Nikita Rokotyan, Olya Stukova, Dasha Kolmakova ont eu recours à l'algorithme T-SNE pour représenter une base de données très connue en géographie et proposer une nouvelle cartographie des pays[226]. Leur application permet de voir tout l'intérêt de cette méthode qui peut s'avérer particulièrement intéressante et adaptée à des questions de SHS, mais aussi ses

[225] Rokotyan Nikita, Stukova Olya et Kolmakova Dasha, *An alternative data-driven country map*, [https://projects.interacta.io/country-tsne].Rokotyan Nikita, Stukova Olya et Kolmakova Dasha,
[226] Rokotyan Nikita, Stukova Olya et Kolmakova Dasha, *An alternative data-driven country map*, [https://projects.interacta.io/country-tsne].

limites : la nécessité d'accompagner l'utilisation de cette méthode d'une bonne connaissance du sujet.

Mot du praticien

Une des limites importantes de l'algorithme t-SNE est qu'il ne conserve pas du tout la structure globale des données. Cela peut produire des effets « bulles », une distorsion entre des groupes ou encore créer artificiellement des structures locales qui en réalité n'existent pas.
Par ailleurs, certains auteurs soulignent une grande sensibilité de l'algorithme aux choix des paramètres : un choix différent peut entraîner des résultats radicalement différents et parfois trompeurs (Belkina et al 2019).

Une question légitime consiste à se demander pourquoi utiliser l'algorithme t-SNE plutôt que UMAP, souvent considéré comme une amélioration de t-SNE : plus rapide, et capable de préserver à la fois la structure locale et la structure globale.

Cela dit, t-SNE excelle naturellement à séparer des groupes distincts dans les données, ce qui peut être particulièrement utile lorsqu'on cherche à visualiser des classes bien définies. Si nous sommes convaincus de l'existence de clusters bien distincts, que la structure locale est déterminante dans nos données et que la structure globale ne nous intéresse pas, alors t-SNE sera particulièrement pertinente.

Ainsi, t-SNE peut être utile lorsque l'objectif principal est d'explorer visuellement les relations entre différentes classes ou groupes au sein des données. En effet, UMAP en préservant la structure globale, peut parfois le faire au détriment de certaines relation locale.

Par ailleurs, par rapport à l'algorithme UMAP, t-SNE s'avère plus efficace lorsque les distances utilisées sont non euclidiennes, lorsque les échantillons sont de petite taille et il reste généralement moins sensible aux changements de paramètres.

Si t-SNE est une méthode très efficace pour visualiser et identifier des clusters dans nos données, il n'en demeure pas moins qu'il ne s'agit pas d'une méthode de classification. Si un des intérêts des méthodes de réduction de dimension est de servir d'étape préalable à une clusterisation, la pertinence de cette pratique est discutée. Certains auteurs (Chari et al 2023) jugent hasardeux de réaliser une classification à partir des résultats de t-SNE car cette méthode ne conserve pas les distances. Une alternative serait d'utiliser des méthodes se basant sur les densités plutôt que les distances, comme la méthode DBSCAN.

Ressources

- Belkina Anna C., Ciccolella Christopher O., Anno Rina, Halpert Richard, Spidlen Josef et Snyder-Cappione Jennifer E., « Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets », *Nature Communications*, nº 1, vol. 10, 2019, p. 5415, [https//doi.org/10.1038/s41467-019-13055-y].
- Chari Tara et Pachter Lior, « The specious art of single-cell genomics », *PLOS Computational Biology*, nº 8, vol. 19, 2023, p. e1011288, [https//doi.org/10.1371/journal.pcbi.1011288].
- Maaten Laurens van der et Hinton Geoffrey, « Visualizing Data using t-SNE », *Journal of Machine Learning Research*, nº 86, vol. 9, 2008, p. 2579-2605.

- Rokotyan Nikita, Stukova Olya et Kolmakova Dasha, *An alternative data-driven country map*, [https://projects.interacta.io/country-tsne].

UMAP n.f. [y ɛm a pe]

Synonymes : Uniform Manifold Approximation and Projection for Dimension Reduction

A quoi ça sert ?

La méthode UMAP fait partie du champ des méthodes de réduction des dimensions au même titre que la méthode T-SNE ou encore l'ACP qui est en réalité la méthode la plus connue de ce champ. Les méthodes de réduction des dimensions ont pour objectif de simplifier l'étude de points en grandes dimensions, en mettant à jour les caractéristiques principales de cet ensemble de points qui se traduit dans un nuage multidimensionnel. L'idée va donc être de chercher à faire émerger les principales dimensions qui structurent notre base de données d'individus statistiques caractérisés (originellement ?) par un grand nombre de variables. Une des forces de ces différentes méthodes est de proposer une représentation lisible en 2D de ce nuage complexe et multidimensionnel que forme nos données.

La force de l'algorithme UMAP c'est d'être adapté à des données n'ayant pas de relations linéaires. Elle est basée sur la théorie des graphes et la géométrie riemannienne qui construit une représentation topologique des données. L'**ACP** est une méthode linéaire qui va produire de nouvelle variables (les composantes principales) en réalisant une combinaison linéaire des variables existantes et peut donc être inadapté à des données en SHS qui peuvent avoir des relations non linéaires, ce qui arrive souvent. De plus, l'UMAP est conçu pour maintenir la structure locale des données, ce qui signifie qu'il préserve les relations entre les points proches, alors que l'**ACP** se concentre sur la variance globale et peut perdre des informations sur les structures locales importantes.

L'algorithme UMAP va commencer par construire un graphe des voisins pour chaque point. Cela signifie qu'il va identifier pour chaque point les points les plus proches en utilisant une mesure de distance. Le seuil des plus proches voisins est déterminé par l'utilisateur, et la sphère qui est alors composée d'un point et de ses voisins va servir de base pour une distance locale de ce point.
Puis, l'UMAP va utiliser une approche probabiliste pour modéliser la structure des données. En effet, la distance locale est utilisée pour attribuer un poids, une probabilité au lien entre le point considéré et chacun de ses n voisins les plus proches. Ce poids permet de créer un graphe pondéré. Pour optimiser l'algorithme va créer des distributions de probabilité pour les relations entre les points dans l'espace d'origine puis tente de reproduire ces distributions dans l'espace réduit à deux dimensions. La représentation finale produite par UMAP sera la meilleure représentation du graphe pondéré en 2D.

En revanche, contrairement à d'autres méthodes, les axes de la représentation produits par l'UMAP sont ininterprétables, tout simplement car ils ne sont pas porteurs de sens. Il faut imaginer cette représentation comme une carte géographique où les coordonnées géographiques n'auraient aucun intérêt mais où c'est la position des points les uns par rapport aux autres qui compte. La représentation graphique va nous permettre de visuellement identifier des clusters ou des groupes de points qui sont proches les uns des autres. Ces clusters peuvent représenter des catégories ou des classes dans les données d'origine. Par exemple, dans un ensemble de données sur des pays, différents clusters pourraient correspondre à différentes régions du monde. La distance entre les clusters dans l'espace UMAP peut indiquer la similarité entre ces groupes. Des clusters éloignés suggèrent que les groupes sont distincts, tandis que des clusters proches indiquent une certaine similarité ou chevauchement. L'UMAP préserve la structure locale, ce qui signifie que les points qui

sont proches dans l'espace d'origine le resteront dans l'espace réduit. Cela permet d'explorer les relations fines au sein des données. Bien que l'UMAP conserve la structure locale, il peut également donner une idée de la structure globale des données. Par exemple, une forme allongée ou en spirale dans l'espace UMAP peut indiquer une continuité ou une hiérarchie dans les données. Pour mieux interpréter les résultats, il est indispensable de superposer d'autres informations contextuelles, telles que des étiquettes de classe ou des métadonnées, sur le graphique UMAP.

D'où ça vient ?

L'algorithme UMAP est issu de travaux récents. Il a été théorisé et développé en 2018 par Leland Mcinnes[227]. Cet algorithme est souvent considéré comme une amélioration de l'algorithme **t-sne** notamment en matière de rapidité, de représentation et surtout de conservation de la structure globale de nos données.

Exemples d'application

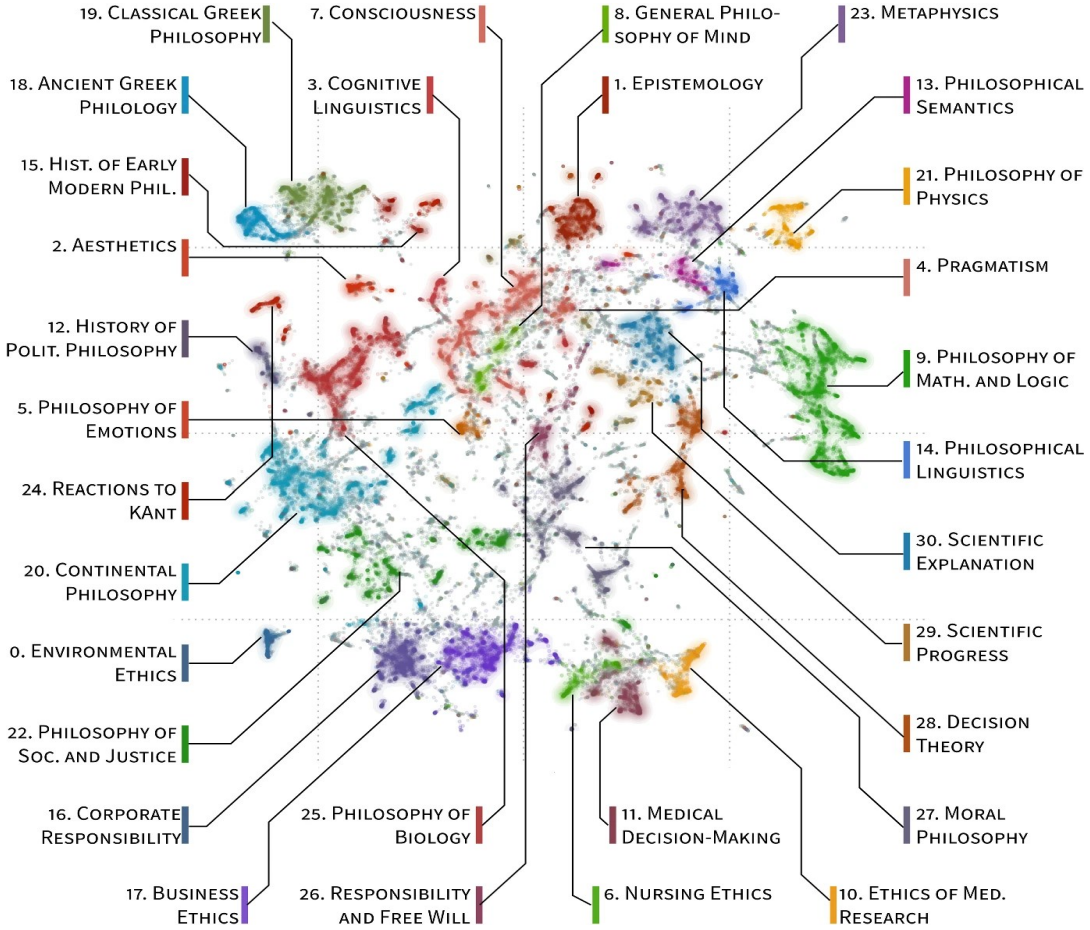
Un des premiers exemples d'utilisation de l'UMAP appliqué aux sciences humaines et sociales a été publié par Maximilian Noichl en 2019[228]. Maximilian Noichl a cartographié les différents courants de pensée en philosophie via les publications. Cette question des courants de pensée en philosophie est apparemment structurante, il y a un vrai questionnement afin de savoir s'il y a un fossé entre la philosophie « analytique » et la philosophie « continentale », ou si la philosophie est divisée plutôt selon les lignes du « naturalisme » et de « l'anti-naturalisme ».

Maximilian Noichl va proposer de répondre à ces questions en utilisant notamment la méthode UMAP. Sa base de données est composée de l'ensemble des articles cités plus de trois fois dans une des revues référencées par Philpapers.org, et présentes sur le Web Of Sciences. Après avoir nettoyé ses données il a construit une matrice de près de 80 000 articles avec au total plus d'1 million de citations.

Voici le résultat de son travail :

[227] McInnes Leland, Healy John et Melville James, « UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction », [https//doi.org/10.48550/arXiv.1802.03426]. ; McInnes Leland, Healy John, Saul Nathaniel et Großberger Lukas, « UMAP: Uniform Manifold Approximation and Projection », *Journal of Open Source Software*, n° 29, vol. 3, 2018, p. 861, [https//doi.org/10.21105/joss.00861].

[228] Noichl Maximilian, « Modeling the structure of recent philosophy », *Synthese*, n° 6, vol. 198, 2021, p. 5089-5100, [https//doi.org/10.1007/s11229-019-02390-8].



Source : Noichl, M. 2021

Chaque point du graphique représente un des articles. Ils sont disposés donc en fonction d'un regroupement umap, qui a positionné les articles ayant des modèles de citation similaires les uns près des autres. Les points sont colorés en fonction de leur groupe, tandis que les points auxquels aucun groupe n'a été attribué sont laissés en gris. Les étiquettes ont été dérivées des termes les plus courants dans les résumés et les titres, ainsi que des publications les plus citées dans les groupes. Pour définir ces clusters, Maximilian Noichl a utilisé un algorithme de clusterisation hDBSCAN sur ces résultats fournis par l'UMAP.

Sa conclusion est que cette approche met l'accent sur une forte interconnexion de nombreux sous-domaines plus ou moins importants de la philosophie. Par ailleurs, ses résultats montrent que l'idée d'un fossé entre la philosophie analytique et la philosophie continentale n'est pas confirmé par la structure de la littérature, bien que la philosophie continentale, contrairement à la philosophie analytique, forme un groupe cohérent de collaboration scientifique.

Mot du praticien

L'UMAP est une méthode qui part bien des égards peut s'avérer très intéressante. Surtout dans le cadre de données multidimensionnelle complexe avec des relations non linéaires de toute sorte. Elle est particulièrement utilisée dans certains champs hors SHS notamment en génomique. Nous pouvons également noter qu'elle fonctionnera avec des données catégorielles qui auront été codé numériquement.

Toutefois cela reste une méthode sensible. En effet, il y a un grand nombre de paramètres (le nombre de voisins, le nombre de dimensions finales, la distance minimale…) et la moindre modification, même légère, va entraîner un résultat différent. Cela peut donc être intéressant de lancer l'UMAP avec différents paramètres et de comparer les résultats obtenus. Par ailleurs, l'UMAP est une méthode non déterministe ce qui implique qu'appliquée deux fois de suite sur les mêmes données avec les mêmes paramètres, elle ne donnera pas exactement les mêmes résultats. Cela peut être déroutant même si la représentation finale ne sera pas radicalement différente.

Concernant les limites de l'algorithme d'UMAP, nous pouvons souligner qu'il fonctionne moins bien sur de petites populations. Il est également par nature moins intéressant sur des des jeux de données pour lesquels la structure globale est plus importante que les structures locales. Si c'est le cas il faudra songer à utiliser une autre méthode.

L'aspect « artisanal », notamment la dimension non déterministe et l'importance des paramètres dans les résultats obtenus peut aussi paraître rédhibitoire pour certaines personnes ou certaines applications. Il peut parfois donner l'impression de forcer le résultat.

Enfin, un des risques de l'UMAP est de faire apparaître des structures au niveau local là où il n'y aurait en réalité que du bruit : il peut donc être risqué de l'appliquer aveuglément, notamment sur des données pour lesquelles on ne pourrait pas contrôler les résultats obtenus a posteriori.

Dans son article fondateur Leland McInnes discute lui même les limites de l'algorithme UMAP dans la section weakness, ainsi que les développements futurs.

Enfin, il faut rappeler que l'UMAP n'est pas une méthode de classification. Si elle fait émerger des groupes, des tendances, il ne s'agit pas d'une méthode de clustering. Il est possible toutefois d'utiliser sur des résultats de l'UMAP des méthodes de classification. Leland McInnes indique qu'à la manière de l'**ACM**, l'UMAP peut être utilisé en prémisse d'une classification, il conseille pour ce faire d'utiliser la méthode DBSCAN qui est basée sur les densités.

Néanmoins, la pertinence de cela est discutée, en effet, certains auteurs (Chari et al 2023)[229] jugent hasardeux de réaliser une classification sur des résultats de l'UMAP. Ceci notamment car cette méthode ne conserve pas les distances et ni la densité. La fluctuation des résultats par son aspect non déterministe et sa sensibilité au moindre changement de paramètre sont effectivement des freins à l'application d'analyses postérieures à l'UMAP. La classification appliquée sur les résultats de l'UMAP pourrait être remise en cause par l'instabilité de l'algorithme qui risque de faire émerger artificiellement des structures locales qui ne seraient donc pas retrouvées par la suite.

Ressources

- Chari Tara et Pachter Lior, « The specious art of single-cell genomics », *PLOS Computational Biology*, nº 8, vol. 19, 2023, p. e1011288, [https//doi.org/10.1371/journal.pcbi.1011288].
- Healy J. et McInnes L., *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — umap 0.5.8 documentation*, [https://umap-learn.readthedocs.io/en/latest/index.html].

[229] Chari Tara et Pachter Lior, « The specious art of single-cell genomics », *PLOS Computational Biology*, nº 8, vol. 19, 2023, p. e1011288, [https//doi.org/10.1371/journal.pcbi.1011288].

- McInnes Leland, Healy John et Melville James, « UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction », [https//doi.org/10.48550/arXiv.1802.03426].
- McInnes Leland, Healy John, Saul Nathaniel et Großberger Lukas, « UMAP: Uniform Manifold Approximation and Projection », *Journal of Open Source Software*, n° 29, vol. 3, 2018, p. 861, [https//doi.org/10.21105/joss.00861].
- Noichl Maximilian, « Modeling the structure of recent philosophy », *Synthese*, n° 6, vol. 198, 2021, p. 5089-5100, [https//doi.org/10.1007/s11229-019-02390-8].

Valeurs aberrantes n.f.p. [/valœʁ aberɑ̃t/]

Synonymes : error outliers

A quoi ça sert ?

Les valeurs abérrantes font référence aux valeurs récoltées qui sont issues d'erreurs de mesure ou d'un mauvais report lors de la collecte des données. Aguinis et al. en 2013 parlent d'*error outliers*[230]. Par exemple, un participant répondant à une enquête en 2025 qui aurait pour année de naissance 1789, traduit nécessairement une erreur, qui peut être liée à une faute de frappe lors de la saisie de la date ou à une saisie intentionnellement inexacte dans le questionnaire. Dans tous les cas, cette valeur est erronée et ne doit pas être conservée dans les données. Elle peut fausser les résultats obtenus en tordant les **indices de tendance centrale** et de **dispersion**, mais ce n'est pas toujours le cas. Par exemple, si la personne avait indiqué être née en 1989, qui est une date de naissance totalement possible pour une personne vivante en 2025, mais que cette date était tout de même erronée, elle serait abbérante mais non **extrême**. Toutefois, elle doit être retirée car elle ne reflète pas la réalité alors observée. Elle pourra être modifiée si le participant peut à nouveau être contacté et accepte de modifier sa réponse. Sinon elle sera considérée comme une **valeur manquante** et devra être traitée comme telle.

Ce type de situation arrive également lorsqu'un instrument de mesure est mal paramétré ou qu'une réponse ne s'enregistre pas correctement. Par exemple, une personne qui mettrait 10 minutes à répondre à une tâche programmée par ordinateur, qui ne nécessite que quelques millisecondes par réponse, devrait être évacuée du reste des données. Celle-ci traduisant une erreur de prise en compte de la réponse du participant.

Mot du praticien

Le terme de valeurs aberrantes est souvent employé pour faire référence à des valeurs qui diffèrent du reste de l'échantillon de données, or c'est un abus de langage. En effet, il est important de différencier les valeurs aberrantes (qui ne peuvent pas se produire et qui relèvent d'une erreur de mesure ou de traitement des données), des **valeurs extrêmes** qui sont des valeurs, qui elles, peuvent se produire, et ne résultent donc pas d'une erreur, mais sont éloignées du reste de l'échantillon pour différentes raisons. Le terme anglais *outlier* renforce cette confusion, car il renvoie aux deux définitions. Toutefois, les auteurs peuvent distinguer les *error outliers* des autres types d'*outliers* (*random*, etc.). Aguinis et al. (2013) proposent 14 définitions des *outliers*[231] et exposent parfaitement la distinction qui existe donc en français entre valeurs aberrantes et **valeurs extrêmes**.

Ressources

- Aguinis Herman, Gottfredson Ryan K. et Joo Harry, « Best-Practice Recommendations for Defining, Identifying, and Handling Outliers », *Organizational Research Methods*, nº 2, vol. 16, 2013, p. 270-301, [https//doi.org/10.1177/1094428112470848].

[230] Aguinis Herman, Gottfredson Ryan K. et Joo Harry, « Best-Practice Recommendations for Defining, Identifying, and Handling Outliers », *Organizational Research Methods*, nº 2, vol. 16, 2013, p. 270-301, [https//doi.org/10.1177/1094428112470848].

[231] Idem

Valeurs extrêmes n.f.p [/va.lœʁ ɛks.tʁɛm/]

Synonymes : random outliers, interesting outliers

A quoi ça sert ?

Toute analyse de données implique la confrontation à des valeurs extrêmes. Une valeur extrême étant la plupart du temps définie en SHS par l'éloignement d'une ou de plusieurs observations du reste de l'échantillon sur un ou plusieurs critères. Toutefois, la définition de ce qu'est une valeur extrême ne fait pas consensus dans la littérature scientifique et Aguinis et al., en 2013, proposent 14 définitions de l'appelation anglaise *outliers*[232]. Les valeurs extrêmes peuvent être produites pour différentes raisons, que ce soit suite à une erreur liée aux outils de mesure, dans la manipulation des données[233], dans l'échantillonnage ou surtout ne pas être une erreur et traduire un apport de nouvelles informations. Dans tous les cas, il faudra accorder une vigilence particulière à ces observations extrêmes dans la mesure où celles-ci peuvent altérer les résultats des analyses. S'il n'y a pas de consens clair dans la littérature sur la suppression systématique ou non des valeurs extrêmes (André, 2022[234]; Karch, 2023[235]), celles-ci nécessitent d'être identifiées afin de pouvoir effectuer des choix éclairés et adaptés aux analyses menées. La littérature est foisonnante sur le sujet, nous citerons notamment les travaux de Bakker, M., & Wicherts, J. M. (2014)[236] et de Leys, C. et al. (2019)[237].

D'où ça vient ?

La conception de valeurs extrêmes remonte au XVI^{e} siècle via les commentaires de Bernouilli sur les "observations discordantes" et la façon de les traiter[238]. Depuis 250 ans, la littérature sur la façon de définir et de traiter les valeurs extrêmes n'a cesser de s'étoffer en parallèle du développement des méthodes d'analyses des données et de leur complexification.

Mot du praticien

---

[232] Aguinis Herman, Gottfredson Ryan K. et Joo Harry, « Best-Practice Recommendations for Defining, Identifying, and Handling Outliers », *Organizational Research Methods*, n^{o} 2, vol. 16, 2013, p. 270-301, [https//doi.org/10.1177/1094428112470848].

[233] Dans ce cas, si les valeurs extrêmes sont produites par des erreurs liées aux outils de mesure ou dans la manipulation des données, ce sont des erreurs et donc des **valeurs aberrantes**. Elles ne peuvent pas se produire, ne renvoient à aucune réalité observée et nécessitent d'être sorties des données.

[234] André Quentin, « Outlier exclusion procedures must be blind to the researcher's hypothesis. », *Journal of Experimental Psychology: General*, n^{o} 1, vol. 151, 2022, p. 213-223, [https//doi.org/10.1037/xge0001069].

[235] Karch Julian D., « Outliers may not be automatically removed. », *Journal of Experimental Psychology: General*, n^{o} 6, vol. 152, 2023, p. 1735-1753, [https//doi.org/10.1037/xge0001357].

[236] Bakker Marjan et Wicherts Jelte M., « Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. », *Psychological Methods*, n^{o} 3, vol. 19, 2014, p. 409-427, [https//doi.org/10.1037/met0000014].

[237] Leys Christophe, Delacre Marie, Mora Youri L., Lakens Daniël et Ley Christophe, « How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration », *International Review of Social Psychology*, n^{o} 1, vol. 32, 2019, p. 5, [https//doi.org/10.5334/irsp.289].

[238] Beckman R. J. et Cook R. D., « Outlier … … …. s », *Technometrics*, n^{o} 2, vol. 25, 1983, p. 119-149, [https//doi.org/10.1080/00401706.1983.10487840].

La littérature et les ressources sur la définition, l'identification, le traitement et la mention des valeurs extrêmes sont particulièrement abondantes. Il n'est pas toujours simple de s'y retrouver. En effet, les impacts de ces observations extrêmes sur les analyses de données, et les méthodes de détection pour les identifier, ne font pas l'objet de consensus clair et précis parmi les spécialistes. Les travaux de Cowell, F., A. & Victoria-Feser, M. en 1996 sont particulièrement éclairants sur cet aspect. Ces auteurs ont publié deux articles en 1996 afin d'indiquer que l'impact des valeurs extrêmes sur les résultats d'analyses de données étaient catastrophiques et négligeables[239]. Ce que souhaitaient ces auteurs c'était souligner l'importance de ne pas traiter les observations extrêmes en appliquant une recette magique, mais d'adapter le traitement à la problématique soulevée par la taille de l'échantillon, les outils de mesure employés, ainsi que les méthodes d'analyses réalisées. Il est donc nécessaire d'adopter une posture réflexive et d'adapter ses choix au cas par cas et surtout de les documenter. Les choix effectués seront l'objet d'un arbitrage qui aurait pu être différent, mais le fait de documenter ces choix permet la reproductibilité des résultats obtenus[240].

Par ailleurs, différentes méthodes de détection des valeurs extrêmes se développent et permettent de croiser les résultats obtenus et de sélectionner la méthode la plus adaptée au cas de figure rencontré. Toutefois, ces différentes méthodes ont souvent été développées sur la base de simulations statistiques et n'ont pas toujours été confrontées à la réalité des données empiriques. Ceci est un élément important à considérer, car si ces méthodes semblent très efficaces sur des données simulées, elles peuvent s'avérer bien peu utiles sur données naturelles.

Ressources

- Aguinis Herman, Gottfredson Ryan K. et Joo Harry, « Best-Practice Recommendations for Defining, Identifying, and Handling Outliers », *Organizational Research Methods*, n° 2, vol. 16, 2013, p. 270-301, [https//doi.org/10.1177/1094428112470848].
- André Quentin, « Outlier exclusion procedures must be blind to the researcher's hypothesis. », *Journal of Experimental Psychology: General*, n° 1, vol. 151, 2022, p. 213-223, [https//doi.org/10.1037/xge0001069].
- Bakker Marjan et Wicherts Jelte M., « Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. », *Psychological Methods*, n° 3, vol. 19, 2014, p. 409-427, [https//doi.org/10.1037/met0000014].
- Beckman R. J. et Cook R. D., « Outlier … … …. s », *Technometrics*, n° 2, vol. 25, 1983, p. 119-149, [https//doi.org/10.1080/00401706.1983.10487840].
- Cowell Frank A. et Victoria-Feser Maria-Pia, « Poverty measurement with contaminated data: A robust approach », *European Economic Review*, n° 9, vol. 40, 1996, p. 1761-1771, [https//doi.org/10.1016/0014-2921(95)00048-8].
- Cowell Frank A. et Victoria-Feser Maria-Pia, « Robustness Properties of Inequality Measures », *Econometrica*, n° 1, vol. 64, 1996, p. 77, [https//doi.org/10.2307/2171925].

[239] Cowell Frank A. et Victoria-Feser Maria-Pia, « Poverty measurement with contaminated data: A robust approach », *European Economic Review*, n° 9, vol. 40, 1996, p. 1761-1771, [https//doi.org/10.1016/0014-2921(95)00048-8].
Cowell Frank A. et Victoria-Feser Maria-Pia, « Robustness Properties of Inequality Measures », *Econometrica*, n° 1, vol. 64, 1996, p. 77, [https//doi.org/10.2307/2171925].

[240] Leys Christophe, Delacre Marie, Mora Youri L., Lakens Daniël et Ley Christophe, « How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration », *International Review of Social Psychology*, n° 1, vol. 32, 2019, p. 5, [https//doi.org/10.5334/irsp.289].

- Hartig F., *DHARMa*, [http://florianhartig.github.io/DHARMa/].
- Hartig F., *Installing, loading and citing the package*, 2024.
- Karch Julian D., « Outliers may not be automatically removed. », *Journal of Experimental Psychology: General*, nº 6, vol. 152, 2023, p. 1735-1753, [https//doi.org/10.1037/xge0001357].
- Leys Christophe, Delacre Marie, Mora Youri L., Lakens Daniël et Ley Christophe, « How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration », *International Review of Social Psychology*, nº 1, vol. 32, 2019, p. 5, [https//doi.org/10.5334/irsp.289].
- Leys Christophe, Klein Olivier, Dominicy Yves et Ley Christophe, « Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance », *Journal of Experimental Social Psychology*, vol. 74, 2018, p. 150-156, [https//doi.org/10.1016/j.jesp.2017.09.011].
- Lüdecke Daniel, Ben-Shachar Mattan, Patil Indrajeet, Waggoner Philip et Makowski Dominique, « Performance: An R Package for Assessment, Comparison and Testing of Statistical Models », *Journal of Open Source Software*, nº 60, vol. 6, 2021, p. 3139, [https//doi.org/10.21105/joss.03139].

Valeurs manquantes n.f.p. [/va.lœʁ mɑ̃.kɑ̃t/]

Synonymes : Données manquantes, données incomplètes

A quoi ça sert ?

Les valeurs manquantes font référence à l'absence de valeur pour une variable donnée (comme par exemple, l'âge, la profession et catégorie sociale, le nombre de mètres parcourus ou encore le niveau de stress perçu) concernant une ou plusieurs observations (individus). Les raisons de ces absences de valeurs sont multiples : non-réponse de la part du participant, erreur de mesure lors de la collecte des données, perte de données ou encore suppression d'une **valeur aberrante**.

Quelle que soit la raison de la présence des valeurs manquantes, il est essentiel de les prendre en compte et de les traiter afin de ne pas altérer les analyses statistiques réalisées ensuite sur les données concernées. En effet, la présence de valeurs manquantes peut altérer la *puissance statistique* des analyses réalisées. Toutefois, il est à noter que la présence de valeurs manquantes réparties de façon aléatoire (Missing At Random – MAR) n'est pas nécessairement problématique selon la taille de l'échantillon séléctionné. En revanche, des valeurs manquantes non distribuées aléatoirement (Missing Not At Random - MNAR) nécessitent d'être traitées, car elles ont des conséquences importantes sur les analyses réalisées (Tabachnick and Fidell, 2014).

Selon la quantité et le type de valeurs manquantes présentes dans la base de données il sera necessaire de soit :

- *Si la base de données comprend une faible proportion de valeurs manquantes* : Supprimer l'ensemble des valeurs du participant (ou des participants) concerné de la base de données[241].
- *Si la base de données comprend une proportion non négligeable de valeurs manquantes* : Imputer les valeurs manquantes pour les participants concernés. Il existe différents algorithmes d'imputation de valeurs manquantes selon le type de données concernés et leurs distributions[242].
- *Si la base de données comprend une proportion trop importante de valeurs manquantes* : De ne pas traiter, ni analyser les données, car les éléments fournis dans la base de données concernés sont trop parcellaires et ne permettent pas de réaliser des analyses fiables et reproductibles.

D'où ça vient ?

Les valeurs manquantes sont inhérentes à la collecte des données. A partir du moment où des données sont collectées, celles-ci peuvent comporter des non-réponses de la part de participants ou encore des erreurs qui seront alors supprimées. Le traitement de ces valeurs manquantes a évolué, au cours du XX^{e} siècle, afin de réduire les biais potentiels lors de la réalisation d'analyses statistiques. Jusqu'aux années 1970 les valeurs manquantes étaient traitées soit par suppression des données du participants, soit par remplacement de la valeur absente par la valeur moyenne de la variable concernée[243]. Au milieu des années 1970 Dempster et al. développent une technique

[241] Tabachnick Barbara et Fidell Linda, *Using Multivariate Statistics*, Pearson International, 2013.
[242] Tabachnick Barbara et Fidell Linda, *Using Multivariate Statistics*, Pearson International, 2013.
[243] Schafer Joseph L. et Graham John W., « Missing data: our view of the state of the art », *Psychological Methods*, no 2, vol. 7, 2002, p. 147-177. ; Schafer Joseph L., « Multiple imputation: a primer », *Statistical Methods in Medical Research*, n^{o} 1, vol. 8, 1999, p. 3-15, [https//doi.org/10.1177/096228029900800102].

permettant de traiter les valeurs manquantes par imputation de valeurs basées sur l'algorithme EM pour Expectation-Maximization[244]. C'est à cette même période que sont introduites les notions de valeurs manquantes distribuées aléatoirement (MCAR, MAR) ou non aléatoirement (NMAR). A la fin des années 1970 et au début des années 1980, Rubin élabore le concept d'imputation multiple alors basé sur des méthodes de simulation[245]. Dans les années 1990, d'autres développements concernant les traitements des valeurs manquantes sont venus enrichir la littérature sur le sujet, notamment selon les méthodes d'analyses envisagées. A l'aube des années 2000 les valeurs manquantes ne sont plus considérées uniquement comme un biais potentiel aux analyses envisagées, mais comme une information en soit, qui peut être traitée comme n'importe quelle valeur. Au cours de ces deux dernières décennies les techniques d'imputation des valeurs manquantes se sont diversifiées et complexifiées, en s'appuyant sur les avancées statistiques, considérant notamment des modèles basés sur la médiane, les k plus proches voisins ou encore les **forêts aléatoires[246]**.

Le mot du praticien

Les valeurs manquantes sont problématiques car la plupart des analyses statistiques n'ont pas été pensées en envisageant leur existance. (Schafer & Graham 2002)[247]. Avant de réaliser des analyses statistiques il est donc nécessaire de traiter les valeurs manquantes, mais le choix du traitement va dépendre de trois éléments : la raison pour laquelle la valeur est manquante, où est située cette valeur dans le jeu de données et la proportion de valeurs manquantes dans l'échantillon considéré[248].

La notion de valeurs manquantes aléatoires (qui découle de la raison pour laquelle une valeur est manquante) est discutée dans la littérature, car ce concept renvoie à des contre-sens et des erreurs d'interprétation[249]. Or selon si les valeurs manquantes sont distribuées aléatoirement dans un échantillon ou non, les méthodes de traitement seront différentes. En effet, les différentes méthodes de traitement des valeurs manquantes ne se valent pas et ne seront pas aussi efficientes, voir peuvent se montrer délétères si elles ne sont pas employées dans des conditions adaptées[250].

Ressources

[244] Dempster A. P., Laird N. M. et Rubin D. B., « Maximum Likelihood from Incomplete Data via the EM Algorithm », *Journal of the Royal Statistical Society. Series B (Methodological)*, nº 1, vol. 39, 1977, p. 1-38.

[245] Schafer Joseph L. et Graham John W., « Missing data: our view of the state of the art », *Psychological Methods*, nº 2, vol. 7, 2002, p. 147-177. ; Rubin Donald B., *Multiple Imputation for Nonresponse in Surveys*, 1re éd., Wiley, coll. « Wiley Series in Probability and Statistics », 1987, [https//doi.org/10.1002/9780470316696].

[246] Ren Lijuan, Wang Tao, Sekhari Seklouli Aicha, Zhang Haiqing et Bouras Abdelaziz, « A review on missing values for main challenges and methods », *Information Systems*, vol. 119, 2023, p. 102268, [https//doi.org/10.1016/j.is.2023.102268].

[247] Schafer Joseph L. et Graham John W.,2002 (Op. Cit.)

[248] Ren Lijuan, Wang Tao, Sekhari Seklouli Aicha, Zhang Haiqing et Bouras Abdelaziz, 2023 (Op. Cit.)

[249] Schafer Joseph L. et Graham John W.,2002 (Op. Cit.)

[250] Schafer Joseph L., « Multiple imputation: a primer », *Statistical Methods in Medical Research*, nº 1, vol. 8, 1999, p. 3-15, [https//doi.org/10.1177/096228029900800102] ; Ren Lijuan, Wang Tao, Sekhari Seklouli Aicha, Zhang Haiqing et Bouras Abdelaziz, 2023 (Op. Cit.)

- Dempster A. P., Laird N. M. et Rubin D. B., « Maximum Likelihood from Incomplete Data via the EM Algorithm », *Journal of the Royal Statistical Society. Series B (Methodological)*, nº 1, vol. 39, 1977, p. 1-38.
- Ren Lijuan, Wang Tao, Sekhari Seklouli Aicha, Zhang Haiqing et Bouras Abdelaziz, « A review on missing values for main challenges and methods », *Information Systems*, vol. 119, 2023, p. 102268, [https//doi.org/10.1016/j.is.2023.102268].
- Rubin Donald B., *Multiple Imputation for Nonresponse in Surveys*, 1re éd., Wiley, coll. « Wiley Series in Probability and Statistics », 1987, [https//doi.org/10.1002/9780470316696].
- Schafer Joseph L., « Multiple imputation: a primer », *Statistical Methods in Medical Research*, nº 1, vol. 8, 1999, p. 3-15, [https//doi.org/10.1177/096228029900800102].
- Schafer Joseph L. et Graham John W., « Missing data: our view of the state of the art », *Psychological Methods*, nº 2, vol. 7, 2002, p. 147-177.
- Tabachnick Barbara et Fidell Linda, *Using Multivariate Statistics*, Pearson International, 2013.

VIF n.m. [/vɪf/]

Synonymes : Variance inflation factor, VIF, Facteur d'Inflation de la Variance

A quoi ça sert ?

Le test Variance Inflation Factor (VIF) permet de vérifier la multicolinéarité des résidus. L'étude de la multicolinéarité est essentielle, car si nous sommes en présence de multicolinéarité cela signifie que la part de variance partagée entre différentes variables explicatives (variables indépendantes) est trop importante. Il y a donc une redondance d'information dans notre modèle et donc il y a un risque de surreprésentation d'un phénomène. Ceci va fausser les résultats d'une analyse multivariée (telle que la régression, l'AFE, etc.). Ainsi, afin de vérifier la multicolinéarité des résidus différents tests existent tels que le Variance Inflation Factor (VIF), ou encore la tolérance.

D'où ça vient ?

Le premier à présenter l'indice de facteur d'inflation de la variance (VIF) est Donald, W., Maquardt en 1970[251] en repartant des travaux de Hoerl et Kennard[252] développés également en 1970.

Exemple d'application

Un exemple d'usage typique du VIF consiste à vérifier qu'il y a bien une absence de multicolinéarité entre les variables explicatives sélectionnées pour un modèle de **régression linéaire.** Nous pouvons par exemple envisager un modèle de régression linéaire visant à saisir les facteurs explicatifs de l'adhésion individuelle aux normes de protection du COVID avec pour variables explicatives : la confiance envers la science, la conformité sociale et la dimension éthique, par exemple. Nous utiliserons donc le test du VIF sur l'ensemble de ces variables explicatives afin de vérifier l'absence de multicolinéarité du modèle proposé. Si tel est le cas la valeur du VIF pour chacune des variables sera inférieur à 5, selon le seuil définit par Farrar et Glauber (Farrar & Glauber, 1967). Nous pourrons donc conclure à une absence de multicolinéarité qui serait problématique.

Le mot du praticien

Il n'y a pas de consensus dans la littérature sur le seuil à retenir afin de conclure à une absence de multicolinéarité. Les standards utilisés en Psychologie, ainsi qu'en Géographie indiquent que les VIF doivent être inférieurs à 5 pour conclure à une absence de multicolinéarité (Farrar & Glauber, 1967).[253] La présentation se base sur un exemple en Psychologie donc les critères retenu afin d'affirmer l'absence de multicolinéarité sont ceux les plus couramment repris dans cette discipline, car comme évoqué il n'y a pas de consensus sur ce critère. Selon les auteurs le VIF peut conclure

[251] Marquardt Donald W., « Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation », *Technometrics*, n° 3, vol. 12, 1970, p. 591-612, [https//doi.org/10.1080/00401706.1970.10488699].

[252] Hoerl Arthur E. et Kennard Robert W., « Ridge Regression: Applications to Nonorthogonal Problems », *Technometrics*, n° 1, vol. 12, 1970, p. 69-82, [https//doi.org/10.1080/00401706.1970.10488635].
Hoerl Arthur E. et Kennard Robert W., « Ridge Regression: Biased Estimation for Nonorthogonal Problems », *Technometrics*, n° 1, vol. 12, 1970, p. 55-67, [https//doi.org/10.1080/00401706.1970.10488634].

[253] Farrar Donald E. et Glauber Robert R., « Multicollinearity in Regression Analysis: The Problem Revisited », *The Review of Economics and Statistics*, n° 1, vol. 49, 1967, p. 92, [https//doi.org/10.2307/1937887].

à une absence de multicoliéarité si celui-ci est inférieur à 10 ou d'autres plus conservateurs vont considérer qu'il faut retenir un VIF inférieur à 2 pour conclure à une absence de multicolinéarité.
Par ailleurs, l'exemple d'application du VIF présenté précédemment est basé sur un modèle de régression linéaire. Il est à noter qu'il est nécessaire de vérifier la présence de multicolinéarité pour toutes les analyses multivariées, que ce soit dans le cadre de modèles de régression que dans celui d'analyses factorielles. Il est en effet tout-à-fait possible de vérifier la multicolinéarité en calculant le VIF en amont d'une analyse factorielle sur le jeu de données alors sélectionné (Kyriazos, T. and Poga, M., 2023)[254].

Ressources

- Farrar Donald E. et Glauber Robert R., « Multicollinearity in Regression Analysis: The Problem Revisited », *The Review of Economics and Statistics*, nº 1, vol. 49, 1967, p. 92, [https//doi.org/10.2307/1937887].
- Hoerl Arthur E. et Kennard Robert W., « Ridge Regression: Applications to Nonorthogonal Problems », *Technometrics*, nº 1, vol. 12, 1970, p. 69-82, [https//doi.org/10.1080/00401706.1970.10488635].
- Hoerl Arthur E. et Kennard Robert W., « Ridge Regression: Biased Estimation for Nonorthogonal Problems », *Technometrics*, nº 1, vol. 12, 1970, p. 55-67, [https//doi.org/10.1080/00401706.1970.10488634].
- Kyriazos Theodoros et Poga Mary, « Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions », *Open Journal of Statistics*, nº 03, vol. 13, 2023, p. 404-424, [https//doi.org/10.4236/ojs.2023.133020].
- Marquardt Donald W., « Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation », *Technometrics*, nº 3, vol. 12, 1970, p. 591-612, [https//doi.org/10.1080/00401706.1970.10488699].

[254] Kyriazos Theodoros et Poga Mary, « Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions », *Open Journal of Statistics*, nº 03, vol. 13, 2023, p. 404-424, [https//doi.org/10.4236/ojs.2023.133020].

Wilcoxon n.m. [/ˈwɪlkɒksən/]

Synonymes : test de la somme des rangs de Wilcoxon

A quoi ça sert ?

Le test de Wilcoxon permet de comparer deux moyennes lorsque les conditions d'application pour un **test t** ou une **ANOVA** ne sont pas remplies. Il s'agit d'un test de la famille des tests non-paramétriques. C'est-à-dire que ces tests ne sont pas basés sur des paramètres tels que la moyenne ou la variance, mais comparent des séries de valeurs deux à deux en se basant sur leurs rangs. Le test de Wilcoxon est souvent assimilé à celui de **Mann-Whitney** car ce sont deux tests non-paramétriques ayant pour objectif de comparer deux échantillons sur une variable quantitative. Ces tests sont effectivement équivalents, toutefois le test de **Mann-Whitney** a été développé pour comparer deux échantillons indépendants alors que le test de Wilcoxon a été créé afin de comparer deux échantillons qu'ils soient indépendants ou appariés (donc composés des mêmes sujets)[255]. Une relation linéaire relie les deux tests et il est courant de trouver dans la littérature ces deux tests sous la même appellation : Mann-Whitney-Wilcoxon ou Wilcoxon-Mann-Whitney.

D'où ça vient ?

Le test de Wilcoxon a été développé par Frank Wilcoxon en 1945[256], afin de comparer les valeurs de deux échantillons dont les paramètres ne seraient pas adaptés à la réalisation d'un test t ou d'une ANOVA.

Mot du praticien

Le test de Wilcoxon peut par exemple être employé pour comparer le temps de reconnaissance des mots selon si ceux-ci sont peu fréquents ou fréquents en français. Les participants étant exposés aux deux types de mots, les échantillons constituant les mots peu fréquents et fréquents sont donc composés des mêmes individus et sont donc des échantillons appariés. De façon générale la répartition des valeurs issues de temps de réaction se suivent rarement une distribution normale, mais plutôt en suivant une distribution gamma[257]. Le test de Wilcoxon peut également être employé sur deux échantillons indépendants, comme le **Mann-Whitney**. Par exemple, nous pouvons appliquer ce test afin de comparer le score à un test de mémoire de deux groupes d'adultes, l'un composé d'adultes jeunes et l'autre d'adultes âgés. Ne pouvant être jeunes et âgés à la fois, ces deux groupes sont donc indépendants.

Ressources

- Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.

[255] Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.

[256] Wilcoxon Frank, « Individual Comparisons by Ranking Methods », *Biometrics Bulletin*, n° 6, vol. 1, 1945, p. 80, [https//doi.org/10.2307/3001968].

[257] Tejo Mauricio, Araya Héctor, Niklitschek-Soto Sebastián et Marmolejo-Ramos Fernando, « Theoretical models of reaction times arising from simple-choice tasks », *Cognitive Neurodynamics*, n° 4, vol. 13, 2019, p. 409-416, [https//doi.org/10.1007/s11571-019-09532-1].

- Tejo Mauricio, Araya Héctor, Niklitschek-Soto Sebastián et Marmolejo-Ramos Fernando, « Theoretical models of reaction times arising from simple-choice tasks », *Cognitive Neurodynamics*, n° 4, vol. 13, 2019, p. 409-416, [https//doi.org/10.1007/s11571-019-09532-1].
- Wilcoxon Frank, « Individual Comparisons by Ranking Methods », *Biometrics Bulletin*, n° 6, vol. 1, 1945, p. 80, [https//doi.org/10.2307/3001968].

Bibliographie

- Abbott Andrew, « Sequence Analysis: New Methods for Old Ideas », *Annual Review of Sociology*, nº 1, vol. 21, 1995, p. 93-113, [https//doi.org/10.1146/annurev.so.21.080195.000521].
- Abbott Andrew et Forrest John, « Optimal Matching Methods for Historical Sequences », *Journal of Interdisciplinary History*, nº 3, vol. 16, 1986, p. 471, [https//doi.org/10.2307/204500].
- Abbott Andrew et Hrycak Alexandra, « Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers », *American Journal of Sociology*, nº 1, vol. 96, 1990, p. 144-185, [https//doi.org/10.1086/229495].
- Abbott Andrew et Tsay Angela, « Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect », *Sociological Methods & Research*, nº 1, vol. 29, 2000, p. 3-33, [https//doi.org/10.1177/0049124100029001001].
- Afin Rifai, Tibor Keresztély et Ilona Cserháti, « Firm performance and markets: survival analysis of medium and large manufacturing enterprises in Indonesia », *Journal of Industrial and Business Economics*, nº 1, vol. 52, 2025, p. 107-151, [https//doi.org/10.1007/s40812-024-00302-7].
- Agresti Alan, *Categorical data analysis*, Third edition., Hoboken, New Jersey, Wiley-Interscience, coll. « Wiley series in probability and statistics », 2013.
- Aguinis Herman, Gottfredson Ryan K. et Joo Harry, « Best-Practice Recommendations for Defining, Identifying, and Handling Outliers », *Organizational Research Methods*, nº 2, vol. 16, 2013, p. 270-301, [https//doi.org/10.1177/1094428112470848].
- Altman Edward I., « FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY », *The Journal of Finance*, nº 4, vol. 23, 1968, p. 589-609, [https//doi.org/10.1111/j.1540-6261.1968.tb00843.x].
- Amand Maëlle, *Tuto@MATE - Les Analyses Factorielles Multiples (AFM)*, [https://mate-shs.cnrs.fr/actions/tutomate/tuto32-les-analyses-factorielles-multiples-afm-amand/].
- Amand Maelle, *A sociophonetic analysis of Tyneside English in the DECTE corpus : the case of FACE, GOAT, PRICE and MOUTH*, thèse de doctorat, Université Paris Cité, 2019.
- André Quentin, « Outlier exclusion procedures must be blind to the researcher's hypothesis. », *Journal of Experimental Psychology: General*, nº 1, vol. 151, 2022, p. 213-223, [https//doi.org/10.1037/xge0001069].
- Anselin Luc, « The Moran scatterplot as an ESDA tool to assess local instability in spatial association », *Spatial Analytical Perspectives on GIS*, Routledge, 1996, .
- Anselin Luc, « Local Indicators of Spatial Association—LISA », *Geographical Analysis*, nº 2, vol. 27, 1995, p. 93-115, [https//doi.org/10.1111/j.1538-4632.1995.tb00338.x].
- Antoine Charles, *Les moyennes*, 1. éd., Paris, Presses Universitaires de France, coll. « Que sais-je? », 1998.
- Antolinos-Basso Diégo, Blanc Nathalie, Chiche Jean et Paddeu Flaminia, « S'engager pour l'environnement dans le Grand Paris : territoires, politiques et inégalités », *Cybergeo: European Journal of Geography*, , 2020, [https//doi.org/10.4000/cybergeo.34544].
- Apparicio Philippe, Carrier Mathieu, Gelb Jérémy, Séguin Anne-Marie et Kingham Simon, « Cyclists' exposure to air pollution and road traffic noise in central city neighbourhoods of Montreal », *Journal of Transport Geography*, vol. 57, 2016, p. 63-69, [https//doi.org/10.1016/j.jtrangeo.2016.09.014].

- Apparicio Philippe, Riva Mylène et Séguin Anne-Marie, « A comparison of two methods for classifying trajectories: a case study on neighborhood poverty at the intra-metropolitan level in Montreal », *Cybergeo*, , 2015, [https//doi.org/10.4000/cybergeo.27035].
- ARCEP, *Baromètre du Numérique*, [https://www.data.gouv.fr/fr/datasets/barometre-du-numerique/informations/].
- Arikan Serkan, Özer Ferah, Şeker Vuslat et Ertaş Güneş, « The Importance of Sample Weights and Plausible Values in Large-Scale Assessments », *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, nº 1, vol. 11, 2020, p. 43-60, [https//doi.org/10.21031/epod.602765].
- Arntz Arnoud, Mensink Kyra, Cox Wouter R., Verhoef Rogier E. J., Van Emmerik Arnold A. P., Rameckers Sophie A., Badenbach Theresa et Grasman Raoul P. P. P., « Dropout from psychological treatment for borderline personality disorder: a multilevel survival meta-analysis », *Psychological Medicine*, nº 3, vol. 53, 2023, p. 668-686, [https//doi.org/10.1017/S0033291722003634].
- Atkinson Anthony C. et Riani Marco, « Distribution Theory and Simulations for Tests of Outliers in Regression », *Journal of Computational and Graphical Statistics*, nº 2, vol. 15, 2006, p. 460-476, [https//doi.org/10.1198/106186006X113593].
- Atkinson Anthony C., Riani Marco et Corbellini Aldo, « The Box–Cox Transformation: Review and Extensions », *Statistical Science*, nº 2, vol. 36, 2021, p. 239-255.
- Audard Frédéric et Le Campion Grégoire, *Initiation - FOrmation à la Statistique, Spatiale*, [https://letg.pages.in2p3.fr/initiation-formation-aux-statistiques-spatiales/ifoss_immo.html].
- Audard Frédéric, Le Campion Grégoire et Pierson Julie, « La régression géographiquement pondérée : GWR », , 2024, [https//doi.org/10.48645/WK1M-HG05].
- Bakker Arthur et Gravemeijer Koeno P. E., « An Historical Phenomenology of Mean and Median », *Educational Studies in Mathematics*, nº 2, vol. 62, 2006, p. 149-168, [https//doi.org/10.1007/s10649-006-7099-8].
- Bakker Marjan et Wicherts Jelte M., « Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. », *Psychological Methods*, nº 3, vol. 19, 2014, p. 409-427, [https//doi.org/10.1037/met0000014].
- Barabási Albert-László, Albert Réka et Jeong Hawoong, « Mean-field theory for scale-free random networks », *Physica A: Statistical Mechanics and its Applications*, nº 1-2, vol. 272, 1999, p. 173-187, [https//doi.org/10.1016/S0378-4371(99)00291-5].
- Bardin Laurence, *L'analyse de contenu*, Presses Universitaires de France, 2013, [https//doi.org/10.3917/puf.bard.2013.01].
- Bardos Mireille, *Analyse discriminante: application au risque et scoring financier*, Paris, Dunod, coll. « Collection Éco sup. Manuel », 2001.
- Barnes J. A., « Class and Committees in a Norwegian Island Parish », *Human Relations*, nº 1, vol. 7, 1954, p. 39-58, [https//doi.org/10.1177/001872675400700102].
- Baron Reuben M. et Kenny David A., « The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. », *Journal of Personality and Social Psychology*, nº 6, vol. 51, 1986, p. 1173-1182, [https//doi.org/10.1037/0022-3514.51.6.1173].
- Beauguitte Laurent, « L'analyse de réseau en sciences sociales. Petit guide pratique ».
- Beauguitte Laurent, *Théorie des graphes et analyse de réseau en géographie : histoire d'un lien faible (1950-1963)*, [https://ouest-edel.univ-nantes.fr/passerelleshs/index.php?id=155].
- Beauguitte Laurent et Ognyanova Katherine, « Visualisation de réseaux avec R », , 2017, [https//doi.org/10.58079/CZE1].

- Beckman R. J. et Cook R. D., « Outlier … … …. s », *Technometrics*, nº 2, vol. 25, 1983, p. 119-149, [https//doi.org/10.1080/00401706.1983.10487840].
- Bécue Bertaut Monique, *Analyse de textuelle avec R*, Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2018.
- Bécue-Bertaut Mónica et Pagès Jérôme, « Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data », *Computational Statistics & Data Analysis*, nº 6, vol. 52, 2008, p. 3255-3268, [https//doi.org/10.1016/j.csda.2007.09.023].
- Belkina Anna C., Ciccolella Christopher O., Anno Rina, Halpert Richard, Spidlen Josef et Snyder-Cappione Jennifer E., « Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets », *Nature Communications*, nº 1, vol. 10, 2019, p. 5415, [https//doi.org/10.1038/s41467-019-13055-y].
- Bellanger Lise, Coulon Arthur et Husi Philippe, « Une méthode de classification ascendante hiérarchique par compromis : hclustcompro », Marseille, France, CIFSD, Mohamed QUAFAFOU.
- Benaze Maud Benichou Duhil de, *Modélisation hybride de réseaux dans un « champ criminel » : contribution des sciences sociales et d'outils logiciels au renseignement criminel*, thèse de doctorat, Université Michel de Montaigne - Bordeaux III, 2023.
- Bentler P. M. et Bonett Douglas G., « Significance tests and goodness of fit in the analysis of covariance structures. », *Psychological Bulletin*, nº 3, vol. 88, 1980, p. 588-606, [https//doi.org/10.1037/0033-2909.88.3.588].
- Benzécri Jean-Paul, « L' analyse des données. 1: La taxinomie », , Paris, Dunod, 1973, .
- Benzécri Jean-Paul, *L'Analyse des données*, Paris Bruxelles Montréal, Dunod, 1973.
- Berelson Bernard, *Content analysis in communication research*, New York, NY, US, Free Press, coll. « Content analysis in communication research », 1952.
- Bergsma Wicher, « A bias-correction for Cramér's and Tschuprow's », *Journal of the Korean Statistical Society*, nº 3, vol. 42, 2013, p. 323-328, [https//doi.org/10.1016/j.jkss.2012.10.002].
- Bertrand Frédéric et Maumy-Bertrand Myriam, *Initiation à la statistique avec R*, 4e éd., Malakoff, Dunod, coll. « Sciences sup », 2023.
- Bidart Claire, Degenne Alain et Grossetti Michel, *La vie en r*[illegible]*seau*, Presses Universitaires de France, 2011, [https//doi.org/10.3917/puf.bidar.2011.01].
- Bingham N., « Studies in the history of probability and statistics XLVI. Measure into probability: from Lebesgue to Kolmogorov », *Biometrika*, nº 1, vol. 87, 2000, p. 145-156, [https//doi.org/10.1093/biomet/87.1.145].
- Blanchet Alain et Gotman Anne, *L'entretien*, 2e éd., nouv. Prés., Suite du tirage., Paris, A. Colin, coll. « Tout le savoir en 128 pages », 2017.
- Blondel Vincent D., Guillaume Jean-Loup, Lambiotte Renaud et Lefebvre Etienne, « Fast unfolding of communities in large networks », *Journal of Statistical Mechanics: Theory and Experiment*, vol. P10008, 2008, p. 1-12, [https//doi.org/10.1088/1742-5468/2008/10/P10008].
- Blum Alain et Biraben Jean-Noël, « Cliff A.D. et Ord J.K. — Spatial processes. Models and applications », *Population*, nº 4, vol. 37, 1982, p. 963-963.
- Boland Philip J., « A Biographical Glimpse of William Sealy Gosset », *The American Statistician*, nº 3, vol. 38, 1984, p. 179-183, [https//doi.org/10.1080/00031305.1984.10483195].
- Bollen Kenneth A., « Total, Direct, and Indirect Effects in Structural Equation Models », *Sociological Methodology*, vol. 17, 1987, p. 37, [https//doi.org/10.2307/271028].
- Bollen Kenneth A., Fisher Zachary, Lilly Adam, Brehm Christopher, Luo Lan, Martinez Alejandro et Ye Ai, « Fifty years of structural equation modeling: A history of generalization,

unification, and diffusion », *Social Science Research*, vol. 107, 2022, p. 102769, [https//doi.org/10.1016/j.ssresearch.2022.102769].
- Bone Jessica K., Fancourt Daisy, Sonke Jill K. et Bu Feifei, « The Changing Relationship Between Hobby Engagement and Substance Use in Young People: Latent Growth Modelling of the Add Health Cohort », *Journal of Youth and Adolescence*, nº 1, vol. 54, 2025, p. 133-145, [https//doi.org/10.1007/s10964-024-02047-x].
- Boto-García David, « Good results come to those who weight: on the importance of sampling weights in empirical research using survey data », *Current Issues in Tourism*, nº 2, vol. 27, 2024, p. 268-287, [https//doi.org/10.1080/13683500.2023.2178394].
- Boudon Raymond, *L'inégalité des chances : la mobilité sociale dans les sociétés industrielles*, Paris : A. Colin, 1973.
- Bourbonnais Régis, « Chapitre 13. Introduction à l'économétrie des données de panel », *Éco Sup*, 2018, p. 371-387.
- Bourdieu Pierre, *La distinction critique sociale du jugement*, Paris, Editions de Minuit : Maison des sciences de l'homme, coll. « Le Sens commun », 2012.
- Box G. E. P. et Cox D. R., « An Analysis of Transformations », *Journal of the Royal Statistical Society Series B: Statistical Methodology*, nº 2, vol. 26, 1964, p. 211-243, [https//doi.org/10.1111/j.2517-6161.1964.tb00553.x].
- Brant Rollin, « Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression », *Biometrics*, nº 4, vol. 46, 1990, p. 1171, [https//doi.org/10.2307/2532457].
- Breiman Leo, « Random Forests », *Machine Learning*, nº 1, vol. 45, 2001, p. 5-32, [https//doi.org/10.1023/A:1010933404324].
- Breiman Leo, Friedman Jerome H., Olshen Richard A. et Stone Charles J., *Classification And Regression Trees*, 1re éd., Routledge, 2017, [https//doi.org/10.1201/9781315139470].
- Bressoux Pascal, *Modélisation statistique appliquée aux sciences sociales*, De Boeck Supérieur, 2010, [https//doi.org/10.3917/dbu.bress.2010.01].
- Breusch T. S., « TESTING FOR AUTOCORRELATION IN DYNAMIC LINEAR MODELS* », *Australian Economic Papers*, nº 31, vol. 17, 1978, p. 334-355, [https//doi.org/10.1111/j.1467-8454.1978.tb00635.x].
- Breusch T. S. et Pagan A. R., « A Simple Test for Heteroscedasticity and Random Coefficient Variation », *Econometrica*, nº 5, vol. 47, 1979, p. 1287, [https//doi.org/10.2307/1911963].
- Brouard Sylvain et Le Hay Viviane, « Les Français et la fiscalité », , 2012, p. 12 p.
- Brown William, « SOME EXPERIMENTAL RESULTS IN THE CORRELATION OF MENTAL ABILITIES[1] », *British Journal of Psychology, 1904-1920*, nº 3, vol. 3, 1910, p. 296-322, [https//doi.org/10.1111/j.2044-8295.1910.tb00207.x].
- Brunsdon Chris, Fotheringham A. Stewart et Charlton Martin E., « Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity », *Geographical Analysis*, nº 4, vol. 28, 1996, p. 281-298, [https//doi.org/10.1111/j.1538-4632.1996.tb00936.x].
- Bulteau Julie, Feuillet Thierry et Le Boennec Rémy, « Spatial Heterogeneity of Sustainable Transportation Offer Values: A Comparative Analysis of Nantes Urban and Periurban/Rural Areas (France) », *Urban Science*, nº 1, vol. 2, 2018, p. 14, [https//doi.org/10.3390/urbansci2010014].
- Carroll Raymond et Ruppert David, *Transformation and Weighting in Regression*, 1st Ed. 1988., Routledge, 2017, [https//doi.org/10.1201/9780203735268].
- CartONG, *Information Management Resource Portal - Learning Corner*, [https://cartong.pages.gitlab.cartong.org/learning-corner/fr/intro_data_analysis].

- Chari Tara et Pachter Lior, « The specious art of single-cell genomics », *PLOS Computational Biology*, nº 8, vol. 19, 2023, p. e1011288, [https//doi.org/10.1371/journal.pcbi.1011288].
- Chavent Marie, Kuentz-Simonet Vanessa, Labenne Amaury et Saracco Jerome, « ClustGeo : Classification Ascendante Hiérarchique (CAH) avec contraintes de proximitté géographique », Lille, France.
- Chavent Marie, Kuentz-Simonet Vanessa et Saracco Jérôme, « Orthogonal rotation in PCAMIX », *Advances in Data Analysis and Classification*, nº 2, vol. 6, 2012, p. 131-146, [https//doi.org/10.1007/s11634-012-0105-3].
- Chen Chao, Liaw Andy et Breiman Leo, *Using Random Forest to Learn Imbalanced Data*, [https://digicoll.lib.berkeley.edu/record/85556].
- Chen Jiangping, Lin Chin-Hsi et Chen Gaowei, « Adolescents' self-regulated and affective learning, teacher support and digital reading literacy: A multilevel latent profile approach », *Computers & Education*, vol. 205, 2023, p. 104883, [https//doi.org/10.1016/j.compedu.2023.104883].
- Chen Xuejiao et Yeung Wei-Jun Jean, « COVID -19 experiences and family resilience: A latent class analysis », *Journal of Marriage and Family*, nº 1, vol. 87, 2025, p. 280-299, [https//doi.org/10.1111/jomf.13031].
- Cheng Simon et Long J. Scott, « Testing for IIA in the Multinomial Logit Model », *Sociological Methods & Research*, nº 4, vol. 35, 2007, p. 583-600, [https//doi.org/10.1177/0049124106292361].
- Cheung Shu Fai, Cheung Sing-Hang, Lau Esther Yuet Ying, Hui C. Harry et Vong Weng Ngai, « Improving an old way to measure moderation effect in standardized units. », *Health Psychology*, nº 7, vol. 41, 2022, p. 502-505, [https//doi.org/10.1037/hea0001188].
- Cho Eunseong et Kim Seonghoon, « Cronbach's Coefficient Alpha: Well Known but Poorly Understood », *Organizational Research Methods*, nº 2, vol. 18, 2015, p. 207-230, [https//doi.org/10.1177/1094428114555994].
- Cibois Philippe, *Les méthodes d'analyse d'enquêtes*, Lyon, ENS Éditions, 2014, [https//doi.org/10.4000/books.enseditions.1443].
- Cibois Philippe, « Les pièges de l'analyse des correspondances », *Histoire & Mesure*, nº 3, vol. 12, 1997, p. 299-320, [https//doi.org/10.3406/hism.1997.1549].
- Cibois Philippe, « L'analyse des correspondances : l'indispensable retour aux données », *Histoire & Mesure*, nº 3, vol. 1, 1986, p. 239-247, [https//doi.org/10.3406/hism.1986.1540].
- Cliff A. D. et Ord K. J., *Spatial autocorrelation*, Pion., London, 1973.
- Cliff A. D. et Ord K. J., « The Problem of Spatial Autocorrelation », in Allen John Scott (dir.), *Studies in regional science*, London, Pion, coll. « London papers in regional science », 1969, p. 25-55.
- Cliff Andrew D. et Ord Keith, « Spatial Autocorrelation: A Review of Existing and New Measures with Applications », *Economic Geography*, vol. 46, 1970, p. 269, [https//doi.org/10.2307/143144].
- Cliff Andrew David et Ord J. K., *Spatial processes: models and applications*, London, Pion, 1981.
- CNRS et Nancy Université, *Centre National de Ressources Textuelles et Lexicales*, [https://www.cnrtl.fr/].
- Cochran W. G., « The Comparison of Percentages in Matched Samples », *Biometrika*, nº 3/4, vol. 37, 1950, p. 256, [https//doi.org/10.2307/2332378].
- Cohen Jacob, *Statistical Power Analysis for the Behavioral Sciences*, 0 éd., Routledge, 2013, [https//doi.org/10.4324/9780203771587].

- Cohen Jacob, « The earth is round (p < .05). », *American Psychologist*, nº 12, vol. 49, 1994, p. 997-1003, [https//doi.org/10.1037/0003-066X.49.12.997].
- Comber Alexis, Brunsdon Christopher, Charlton Martin, Dong Guanpeng, Harris Richard, Lu Binbin, Lü Yihe, Murakami Daisuke, Nakaya Tomoki, Wang Yunqiang et Harris Paul, « A Route Map for Successful Applications of Geographically Weighted Regression », *Geographical Analysis*, nº 1, vol. 55, 2023, p. 155-178, [https//doi.org/10.1111/gean.12316].
- Cornillon Pierre-André, Guyader Arnaud, Husson François, Jégou Nicolas, Josse Julie, Kloareg Maela, Matzner-Lober Eric et Rouvière Laurent, *Statistiques avec R*, 3e éd. revue et Augmentée., Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2012.
- Coulangeon Philippe, « Chapitre 8 - La recomposition des structures sociales du goût et des attitudes culturelles », *Culture de masse et société de classes : Le goût de l'altérité*, Presses Universitaires de France., 2021, p. 241-272.
- Cowell Frank A. et Victoria-Feser Maria-Pia, « Poverty measurement with contaminated data: A robust approach », *European Economic Review*, nº 9, vol. 40, 1996, p. 1761-1771, [https//doi.org/10.1016/0014-2921(95)00048-8].
- Cowell Frank A. et Victoria-Feser Maria-Pia, « Robustness Properties of Inequality Measures », *Econometrica*, nº 1, vol. 64, 1996, p. 77, [https//doi.org/10.2307/2171925].
- Cox D. R., « Regression Models and Life-Tables », *Journal of the Royal Statistical Society Series B: Statistical Methodology*, nº 2, vol. 34, 1972, p. 187-202, [https//doi.org/10.1111/j.2517-6161.1972.tb00899.x].
- Cramer Harald, *Mathematical Methods Of Statistics*, Princeton University Press, 1946.
- Cronbach Lee J., « Coefficient Alpha and the Internal Structure of Tests », *Psychometrika*, nº 3, vol. 16, 1951, p. 297-334, [https//doi.org/10.1007/BF02310555].
- Darlington Richard B. et Hayes Andrew F., *Regression analysis and linear models: concepts, applications, and implementation*, New York London, The Guilford Press, coll. « Methodology in the social sciences », 2017.
- Daudé Éric, « Apports de la simulation multi-agents à l'étude des processus de diffusion », *Cybergeo*, , 2004, [https//doi.org/10.4000/cybergeo.3835].
- Daudé Eric, *Modélisation de la diffusion d'innovations par la simulation multi-agents. L'exemple d'une innovation en milieu rural.*, thèse de doctorat, Université d'Avignon, 2002.
- Davezies Laurent et D'Haultfoeuille Xavier, *Faut-il pondérer ?... Ou l'éternelle question de l'économètre confronté à des données d'enquête - Documents de travail - G2009/06 | Insee*, [https://www.insee.fr/fr/statistiques/1380863].
- Davidson Russell et MacKinnon James G., *Estimation and inference in econometrics*, New York, Oxford University Press, 1993.
- DE BELLEFON Marie-Pierre, LOONIS Vincent et LE GLEUT Renan, *Manuel d'analyse spatiale | Insee*, [https://www.insee.fr/fr/information/3635442].
- De Belsunce Clément, *Tutoriel #6 Statistiques descriptives et tris de données avec R Studio*, PROGEDO.
- De Cezaro Eberhardt Paulo Henrique et Fochezatto Adelar, « Regional Resilience and the Asymmetric Effects of the 2008 Crisis in Brazil: A Survival Model Analysis », *Networks and Spatial Economics*, nº 3, vol. 24, 2024, p. 743-762, [https//doi.org/10.1007/s11067-024-09640-4].
- Dehez Jeoffrey et Lyser Sandrine, « How ocean beach recreational quality fits with safety issues? An analysis of risky behaviours in France », *Journal of Outdoor Recreation and Tourism*, vol. 45, 2024, p. 100711, [https//doi.org/10.1016/j.jort.2023.100711].

- Dempster A. P., Laird N. M. et Rubin D. B., « Maximum Likelihood from Incomplete Data via the EM Algorithm », *Journal of the Royal Statistical Society. Series B (Methodological)*, nº 1, vol. 39, 1977, p. 1-38.
- Déprez Guillaume Roland Michel, Battistelli Adalgisa et Antino Mirko, « Norm and Deviance-Seeking Personal Orientation Scale (NDPOS) Adapted to the Organisational Context », *Psychologica Belgica*, nº 1, vol. 59, 2019, [https//doi.org/10.5334/pb.462].
- DeSante Christopher D., « Revisiting Reliability: The Misuse of Cronbach's Alpha in Political Science ».
- Deville, « Analyses de données chronologiques qualitatives: Comment analyser des calendriers? », *Annales de l'inséé*, nº 45, 1982, p. 45, [https//doi.org/10.2307/20076433].
- Deville, « Méthodes statistiques et numériques de l'analyse harmonique », *Annales de l'inséé*, nº 15, 1974, p. 3, [https//doi.org/10.2307/20075177].
- Deville Jean-Claude, « Analyse harmonique du calendrier de constitution des familles en France. Disparités sociales et évolution de 1920 à 1960 », *Population (French Edition)*, nº 1, vol. 32, 1977, p. 17, [https//doi.org/10.2307/1531590].
- Deville Jean-Claude et Saporta Gilbert, « L'analyse harmonique qualitative », Versailles, France, North-Holland, coll. « Data Analysis and Informatics ».
- Diday E. et Institut National de Recherche en Informatique et en Automatique (dir.), *Data analysis and informatics, V: proceedings of the Fith International Symposium on Data Analysis and Informatics, organised by the Institut National de Recherche en Informatique et en Automatique, Versailles, September 29 - October 2, 1987*, Amsterdam, North-Holland, 1988.
- Dodge Yadolah (dir.), *Statistical data analysis: based on the L1-norm and related methods*, Amsterdam, North-Holland Publ. Co, 1987.
- Dupont Benoît, « La gouvernance polycentrique du cybercrime : les réseaux fragmentés de la coopération internationale », *Cultures & conflits*, nº 102, 2016, p. 95-120, [https//doi.org/10.4000/conflits.19292].
- Durbin J. et Watson G. S., « Testing for Serial Correlation in Least Squares Regression. II », *Biometrika*, nº 1/2, vol. 38, 1951, p. 159, [https//doi.org/10.2307/2332325].
- Duvivier Chloé, Polèse Mario et Apparicio Philippe, « The location of information technology-led new economy jobs in cities: office parks or cool neighbourhoods? », *Regional Studies*, nº 6, vol. 52, 2018, p. 756-767, [https//doi.org/10.1080/00343404.2017.1322686].
- Echegaray Fanny, Roux Solenne, Koleck Michèle, Jouvie Jessika, Artheix-Althabegoity Yvane, Lebourleux Paul, Munuera Caroline et M'bailara Katia, « Development and Preliminary Validation of the Unpleasant and Pleasant Emotion Regulation Assessment (UPER-A) », *European Journal of Psychological Assessment*, , 2024, p. 1015-5759/a000851, [https//doi.org/10.1027/1015-5759/a000851].
- Edwards Ashley A., Joyner Keanan J. et Schatschneider Christopher, « A Simulation Study on the Performance of Different Reliability Estimation Methods », *Educational and Psychological Measurement*, nº 6, vol. 81, 2021, p. 1089-1117, [https//doi.org/10.1177/0013164421994184].
- Efron B., « Bootstrap Methods: Another Look at the Jackknife », *The Annals of Statistics*, nº 1, vol. 7, 1979, [https//doi.org/10.1214/aos/1176344552].
- Elhorst J. Paul, « Applied Spatial Econometrics: Raising the Bar », *Spatial Economic Analysis*, nº 1, vol. 5, 2010, p. 9-28, [https//doi.org/10.1080/17421770903541772].
- Enault Cyril, « Simulation de l'étalement urbain de Dijon en 2030 : approche systémique de la dynamique gravitaire ville-transport », *Cybergeo*, , 2012, [https//doi.org/10.4000/cybergeo.25157].

- Enders Adam, Klofstad Casey, Littrell Shane, Miller Joanne, Theocharis Yannis, Uscinski Joseph et Zilinsky Jan, « Left–right political orientations are not systematically related to conspiracism », *Political Psychology*, , 2024, p. pops.13017, [https//doi.org/10.1111/pops.13017].
- Escofier B. et Pages J., « Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation de vins rouges du Val de Loire », *Revue de Statistique Appliquée*, nº 2, vol. 31, 1983, p. 43-59.
- Escofier Brigitte et Pagès Jérome, *Analyses factorielles simples et multiples. Objectifs méthodes et interprétation*, Dunod, coll. « Sciences Sup », 2008.
- Escolà-Gascón Álex, Dagnall Neil, Denovan Andrew, Drinkwater Kenneth et Diez-Bosch Miriam, « Who falls for fake news? Psychological and clinical profiling evidence of fake news consumers », *Personality and Individual Differences*, vol. 200, 2023, p. 111893, [https//doi.org/10.1016/j.paid.2022.111893].
- Eshima Nobuoki, *An Introduction to Latent Class Analysis: Methods and Applications*, Singapore, Springer Singapore, coll. « Behaviormetrics: Quantitative Approaches to Human Behavior », 2022, [https//doi.org/10.1007/978-981-19-0972-6].
- Etchepare Aurore, Roux Solenne, Destaillats Jean-Marc, Cady Florian, Fontanier David, Couhet Geoffroy et Prouteau Antoinette, « Éléments de validation du Protocole d'Évaluation de la Cognition Sociale de Bordeaux (PECS-B) en population générale et dans la schizophrénie », *Annales Médico-psychologiques, revue psychiatrique*, nº 2, vol. 178, 2020, p. 130-136, [https//doi.org/10.1016/j.amp.2018.06.011].
- Fagerland Morten W. et Hosmer David W., « How to Test for Goodness of Fit in Ordinal Logistic Regression Models », *The Stata Journal: Promoting communications on statistics and Stata*, nº 3, vol. 17, 2017, p. 668-686, [https//doi.org/10.1177/1536867X1701700308].
- Farrar Donald E. et Glauber Robert R., « Multicollinearity in Regression Analysis: The Problem Revisited », *The Review of Economics and Statistics*, nº 1, vol. 49, 1967, p. 92, [https//doi.org/10.2307/1937887].
- Faulon Marie et Sacareau Isabelle, « Tourisme, gestion sociale de l'eau et changement climatique dans un territoire de haute altitude : le massif de l'Everest au Népal », *Revue de géographie alpine*, nº 108-1, 2020, [https//doi.org/10.4000/rga.6759].
- Ferry Mathieu, « Le prix du végétarisme. Légitimité et autonomie culturelle de la caste en Inde contemporaine », *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, nº 1-2, vol. 165-166, 2025, p. 110-153, [https//doi.org/10.1177/07591063251317058].
- Feuillet Thierry, *SIGR 2021 - Atelier analyse spatiale (GWR)*, [https://sigr2021.github.io/gwr/].
- Feuillet Thierry, Coquin Julien, Mercier Denis, Cossart Etienne, Decaulne Armelle, Jónsson Helgi Páll et Sæmundsson Þorsteinn, « Focusing on the spatial non-stationarity of landslide predisposing factors in northern Iceland: Do paraglacial factors vary over space? », *Progress in Physical Geography: Earth and Environment*, nº 3, vol. 38, 2014, p. 354-377, [https//doi.org/10.1177/0309133314528944].
- Feuillet Thierry, Cossart Etienne et Commenges Hadrien, *Manuel de géographie quantitative: Concepts, outils, méthodes*, Armand Colin, 2019, [https//doi.org/10.3917/arco.illet.2019.01].
- Fischer Manfred M., Scholten H. J., Unwin D. et European Science Foundation., *Spatial analytical perspectives on GIS*, London ; Bristol, PA, Taylor & Francis, coll. « GISDATA », 1996.
- Fisher R. A., « THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS », *Annals of Eugenics*, nº 2, vol. 7, 1936, p. 179-188, [https//doi.org/10.1111/j.1469-1809.1936.tb02137.x].

- Fisher, R. A., *Statistical methods for research workers*, 1st éd., Oliver and Boyd, 1925.
- Fisher R. A., « On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P », *Journal of the Royal Statistical Society*, nº 1, vol. 85, 1922, p. 87, [https//doi.org/10.2307/2340521].
- Fisher R. A., « XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. », *Transactions of the Royal Society of Edinburgh*, nº 2, vol. 52, 1919, p. 399-433, [https//doi.org/10.1017/S0080456800012163].
- Fleuret Sébastien et Apparicio Philippe, « Essai de typologie des centres de santé et de services sociaux au Québec », *Canadian Geographies / Géographies canadiennes*, nº 2, vol. 55, 2011, p. 143-157, [https//doi.org/10.1111/j.1541-0064.2010.00318.x].
- Floch J. M. et Le Saout R., *Économétrie spatiale : une introduction pratique - Documents de travail - M2016/06 | Insee*, [https://www.insee.fr/fr/statistiques/2408659].
- Flora David B., « Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates », *Advances in Methods and Practices in Psychological Science*, nº 4, vol. 3, 2020, p. 484-501, [https//doi.org/10.1177/2515245920951747].
- Flora David B. et Flake Jessica K., « The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. », *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, nº 2, vol. 49, 2017, p. 78-88, [https//doi.org/10.1037/cbs0000069].
- Florax Raymond J. G. M., Folmer Hendrik et Rey Sergio J., « Specification searches in spatial econometrics: the relevance of Hendry's methodology », *Regional Science and Urban Economics*, nº 5, vol. 33, 2003, p. 557-579, [https//doi.org/10.1016/S0166-0462(03)00002-4].
- Forsé Michel et Degenne Alain, *Les réseaux sociaux:*, Armand Colin, 2004, [https//doi.org/10.3917/arco.forse.2004.01].
- Fotheringham A. Stewart et Oshan Taylor M., « Geographically weighted regression and multicollinearity: dispelling the myth », *Journal of Geographical Systems*, nº 4, vol. 18, 2016, p. 303-329, [https//doi.org/10.1007/s10109-016-0239-5].
- Fotheringham A. Stewart, Yang Wenbai et Kang Wei, « Multiscale Geographically Weighted Regression (MGWR) », *Annals of the American Association of Geographers*, nº 6, vol. 107, 2017, p. 1247-1265, [https//doi.org/10.1080/24694452.2017.1352480].
- Fotheringham Alexander Stewart, Brunsdon Chris et Charlton Martin, *Geographically weighted regression: the analysis of spatially varying relationships*, Nachdr. der Ausg. 2002., Chichester, Wiley, 2010.
- Fox John et Weisberg Sanford, *An R companion to applied regression*, Third edition., Los Angeles London New Delhi Singapore Washington, DC Melbourne, SAGE, 2019.
- François Axelle, Nolet Anne-Marie et Morselli Carlo, « Sociabilité carcérale et réinsertion »:, *Déviance et Société*, nº 2, Vol. 42, 2018, p. 389-419, [https//doi.org/10.3917/ds.422.0389].
- Freeman Linton C., *The development of social network analysis: a study in the sociology of science*, Vancouver (B.C.), Empirical press, 2004.
- Frisch Ragnar, *Statistical Confluence Analysis By Means Of Complete Regression System*, 1934.
- Fullerton Andrew S. et Xu Jun, « The proportional odds with partial proportionality constraints model for ordinal response variables », *Social Science Research*, nº 1, vol. 41, 2012, p. 182-198, [https//doi.org/10.1016/j.ssresearch.2011.09.003].
- Gaboriault-Boudreau Maxime, Apparicio Philippe et Brunelle Cédric, « Modélisation de la pauvreté urbaine dans la région métropolitaine de Montréal entre 1986 et 2016: Apport des

régressions spatiales par panel », *Cahiers de géographie du Québec*, nº 179-180, vol. 63, 2019, p. 165, [https//doi.org/10.7202/1084230ar].
- Galibourg Antoine, Cussat-Blanc Sylvain, Dumoncel Jean, Telmon Norbert, Monsarrat Paul et Maret Delphine, « Comparison of different machine learning approaches to predict dental age using Demirjian's staging approach », *International Journal of Legal Medicine*, nº 2, vol. 135, 2021, p. 665-675, [https//doi.org/10.1007/s00414-020-02489-5].
- Garrison, L. William, Beauguitte Laurent, Beauguitte Pierre et Gourdon Paul, « William L. Garrison, 1960, Connectivity of the Interstate Highway System.Version bilingue et commentée ».
- Garrison William L., « Connectivity of the interstate highway system », *Papers in Regional Science*, nº 1, vol. 6, 1960, p. 121-137, [https//doi.org/10.1111/j.1435-5597.1960.tb01707.x].
- Gauld Tom, *Tomgauld.com*, [https://www.tomgauld.com].
- Geary R. C., « The Contiguity Ratio and Statistical Mapping », *The Incorporated Statistician*, nº 3, vol. 5, 1954, p. 115, [https//doi.org/10.2307/2986645].
- Gentzkow Matthew, Kelly Bryan et Taddy Matt, « Text as Data », *Journal of Economic Literature*, nº 3, vol. 57, 2019, p. 535-574, [https//doi.org/10.1257/jel.20181020].
- Genuer Robin et Poggi Jean-Michel, *Les forêts aléatoires avec R*, Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2019.
- Getis Arthur et Ord J. K., « The Analysis of Spatial Association by Use of Distance Statistics », *Geographical Analysis*, nº 3, vol. 24, 1992, p. 189-206, [https//doi.org/10.1111/j.1538-4632.1992.tb00261.x].
- Ghosh Abhik, SahaRay Rita, Chakrabarty Sayan et Bhadra Sayan, « Robust Generalised Quadratic Discriminant Analysis », [https//doi.org/10.48550/arXiv.2004.06568].
- Gibson W. A., « Extending Latent Class Solutions to Other Variables », *Psychometrika*, nº 1, vol. 27, 1962, p. 73-81, [https//doi.org/10.1007/BF02289666].
- Glaser Barney G. et Strauss Anselm L., *The discovery of grounded theory: strategies for qualitative research*, 11th printing., New York, Aldine, 1980.
- Godfrey L. G., « Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables », *Econometrica*, nº 6, vol. 46, 1978, p. 1293, [https//doi.org/10.2307/1913829].
- Gourdon Paul, *Tuto@MATE - GEPHI*, [https://mate-shs.cnrs.fr/actions/tutomate/tuto08-gephi-gourdon/].
- Gouvernement du Canada Statistique Canada, *Les statistiques : le pouvoir des données!*, [https://www150.statcan.gc.ca/n1/edu/power-pouvoir/toc-tdm/5214718-fra.htm].
- Graunt John, *Natural and political observations mentioned in a following index, and made upon the bills of mortality by John Graunt ...; with reference to the government, religion, trade, growth, ayre, diseases, and the several changes of the said city.*, 1662.
- Greenhouse Samuel W. et Geisser Seymour, « On Methods in the Analysis of Profile Data », *Psychometrika*, nº 2, vol. 24, 1959, p. 95-112, [https//doi.org/10.1007/BF02289823].
- Grimm Kevin J., Ram Nilam et Estabrook Ryne, *Growth modeling: structural equation and multilevel modeling approaches*, New York, NY, Guilford Press, coll. « Methodology in the social sciences », 2017.
- Grislain-Letrémy Céline et Katossky Arthur, *Les risques industriels et le prix des logements − Économie et Statistique n° 460-461 - 2013 | Insee*, [https://www.insee.fr/fr/statistiques/1377429?sommaire=1377437].
- Guillot Michel, Khlat Myriam, Gansey Romeo, Solignac Matthieu et Elo Irma, « Return Migration Selection and Its Impact on the Migrant Mortality Advantage: New Evidence Using

French Pension Data », *Demography*, n° 5, vol. 60, 2023, p. 1335-1357, [https//doi.org/10.1215/00703370-10938784].
- Guiraud P., *Problèmes et méthodes de la statistique linguistique*, Presses universitaires de France, 1960.
- Guiraud P., *Les caractères statistiques du vocabulaire: essai de méthodologie*, Presses universitaires de France, 1954.
- Guttman Louis, « A Basis for Analyzing Test-Retest Reliability », *Psychometrika*, n° 4, vol. 10, 1945, p. 255-282, [https//doi.org/10.1007/BF02288892].
- Guyot Sylvain, Le Campion Grégoire et Pissoat Olivier, « Diversité et enjeux territoriaux de la mise en art des espaces périphériques dans le monde », *Cybergeo*, , 2020, [https//doi.org/10.4000/cybergeo.35837].
- Hägerstrand T., *The Propagation of Innovation Waves*, Royal University of Lund, Department of Geography, 1952.
- Hägerstrand Torsten, « A Monte Carlo Approach to Diffusion », *European Journal of Sociology / Archives Européennes de Sociologie*, n° 1, vol. 6, 1965, p. 43-67, [https//doi.org/10.1017/S0003975600001132].
- Hakim Nader et Monti Annamaria, « Histoire de la de la pensée juridique et analyse bibliométrique : l'exemple de la circulation des idées entre la France et l'Italie à la Belle Epoque », *Clio@Thémis : Revue électronique d'histoire du droit*, n° 14, 2018, coll. « L'histoire de la pensée juridique : historiographie, actualité et enjeux », [https//doi.org/10.35562/cliothemis.763].
- Hanck Christoph, Arnold Martin, Gerber Alexander et Schmelzer Martin, « 10 Regression with Panel Data | Introduction to Econometrics with R », *Introduction to Econometrics with R*, Bookdown., 2025, .
- Hartig F., *DHARMa*, [http://florianhartig.github.io/DHARMa/].
- Hartig F., *Installing, loading and citing the package*, 2024.
- Hausman Jerry et McFadden Daniel, « Specification Tests for the Multinomial Logit Model », *Econometrica*, n° 5, vol. 52, 1984, p. 1219, [https//doi.org/10.2307/1910997].
- Hayes Andrew F. et Coutts Jacob J., « Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But… », *Communication Methods and Measures*, n° 1, vol. 14, 2020, p. 1-24, [https//doi.org/10.1080/19312458.2020.1718629].
- Hayes Andrew F. et Little Todd D., *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach*, Second edition., New York London, The Guilford Press, coll. « Methodology in the social sciences », 2018.
- Healy Conor et Holtz Yan, *From data to Viz | Find the graphic you need*, [https://www.data-to-viz.com/data-to-viz.com].
- Healy J. et McInnes L., *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — umap 0.5.8 documentation*, [https://umap-learn.readthedocs.io/en/latest/index.html].
- Heath Thomas Little, *A history of Greek mathematics*, Oxford, The Clarendon press, 1921.
- Hedeker Donald, Mermelstein Robin J. et Demirtas Hakan, « Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models », *Statistics in Medicine*, n° 27, vol. 31, 2012, p. 3328-3336, [https//doi.org/10.1002/sim.5338].
- Heiden Serge, Magué Jean-Philippe et Pincemin Bénédicte, « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement », Edizioni Universitarie di Lettere Economia Diritto.

- Henderson C. R., « Best Linear Unbiased Estimation and Prediction under a Selection Model », *Biometrics*, nº 2, vol. 31, 1975, p. 423, [https//doi.org/10.2307/2529430].
- Henderson C. R., Kempthorne Oscar, Searle S. R. et Von Krosigk C. M., « The Estimation of Environmental and Genetic Trends from Records Subject to Culling », *Biometrics*, nº 2, vol. 15, 1959, p. 192, [https//doi.org/10.2307/2527669].
- Hickendorff Marian, Edelsbrunner Peter A., McMullen Jake, Schneider Michael et Trezise Kelly, « Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis », *Learning and Individual Differences*, vol. 66, 2018, p. 4-15, [https//doi.org/10.1016/j.lindif.2017.11.001].
- Hilal Mohamed et Le Gallo Julie, « Carte et modèle statistique pour explorer l'hétérogénéité spatiale », *Traitements et cartographie de l'information géographique*, Londres, Iste éditions, coll. « Géographie et démographie », 2023, .
- Hilbe Joseph M., *Logistic Regression Models*, 0 éd., Chapman and Hall/CRC, 2009, [https//doi.org/10.1201/9781420075779].
- Hoerl Arthur E. et Kennard Robert W., « Ridge Regression: Applications to Nonorthogonal Problems », *Technometrics*, nº 1, vol. 12, 1970, p. 69-82, [https//doi.org/10.1080/00401706.1970.10488635].
- Hoerl Arthur E. et Kennard Robert W., « Ridge Regression: Biased Estimation for Nonorthogonal Problems », *Technometrics*, nº 1, vol. 12, 1970, p. 55-67, [https//doi.org/10.1080/00401706.1970.10488634].
- Holtz Yan, *Python Graph Gallery*, [https://python-graph-gallery.com/].
- Holtz Yan, *The R Graph Gallery – Help and inspiration for R charts*, [https://r-graph-gallery.com/].
- Hornung Roman, « Ordinal Forests », *Journal of Classification*, nº 1, vol. 37, 2020, p. 4-17, [https//doi.org/10.1007/s00357-018-9302-x].
- Hosmer David W. et Lemeshow Stanley, *Applied logistic regression*, New York, Wiley, coll. « Wiley series in probability and mathematical statistics Applied probability and statistics », 1989.
- Hosmer David W., Lemeshow Stanley et Sturdivant Rodney X., *Applied Logistic Regression*, 1re éd., Wiley, coll. « Wiley Series in Probability and Statistics », 2013, [https//doi.org/10.1002/9781118548387].
- Hothorn Torsten, Zeileis Achim, Farebrother Richard W., Cummins Clint, Millo Giovanni et Mitchell David, *R: Rainbow Test*, [https://search.r-project.org/CRAN/refmans/lmtest/html/raintest.html].
- Howell David C., Yzerbyt Vincent, Bestgen Yves et Rogier Marylène, *Méthodes statistiques en sciences humaines*, 2e éd., Bruxelles [Paris], De Boeck, coll. « Ouvertures psychologiques », 2008.
- Hoyt Cyril, « Test Reliability Estimated by Analysis of Variance », *Psychometrika*, nº 3, vol. 6, 1941, p. 153-160, [https//doi.org/10.1007/BF02289270].
- Husson François, Lê Sébastien et Pagès Jérôme, *Analyse de données avec R*, 2e éd. revue et Augmentée., Rennes, Presses universitaires de Rennes, coll. « Pratique de la statistique », 2016.
- Huynh Huynh et Feldt Leonard S., « Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs », *Journal of Educational Statistics*, nº 1, vol. 1, 1976, p. 69, [https//doi.org/10.2307/1164736].
- Iannone Richard, Cheng Joe, Schloerke Barret, Hughes Ellis, Lauer Alexandra, Seo JooYoung, Brevoort Ken et Roy Olivier, *Introduction to Creating gt Tables*, [https://gt.rstudio.com/articles/gt.html].

- INED, *Enquête ERFI - GGS*, [https://erfi.site.ined.fr/].
- INED, *Enquête Trajectoires et Origines 1 - Questionnaire principal TeO*, [https://teo1.site.ined.fr/fr/le_contenu_de_l_enquete/les_grands_themes_traites_dans_le_quetsionnaire/].
- INSEE, *Temps partiel − Emploi, chômage, revenus du travail*, [https://www.insee.fr/fr/statistiques/7456899?sommaire=7456956#consulter].
- INSEE, *Revenu disponible des ménages − Revenus et patrimoine des ménages*, [https://www.insee.fr/fr/statistiques/5371205?sommaire=5371304].
- INSEE, *En 2018, les inégalités de niveau de vie augmentent - Insee Première - N°1813*, [https://www.insee.fr/fr/statistiques/4659174].
- Ioannidis John P. A., « Why Most Published Research Findings Are False », *PLoS Medicine*, nº 8, vol. 2, 2005, p. e124, [https//doi.org/10.1371/journal.pmed.0020124].
- Jain Anil K. et Dubes Richard C., *Algorithms for clustering data*, Englewood Cliffs, N.J, Prentice Hall, coll. « Prentice Hall advanced reference series », 1988.
- Jankee Christopher, Carrard Michel, Verel Sébastien et Ramat Éric, « Les conséquences de la réforme aéroportuaire pour les territoires : apports d'une simulation informatique multi-agents », *Cybergeo*, , 2020, [https//doi.org/10.4000/cybergeo.35537].
- Jiang Ge, *Chapter 5 Lavaan Lab 3: Moderation and Conditional Effects | R Cookbook for Structural Equation Modeling*.
- Jiao Junfeng et Azimian Amin, « Measuring accessibility to grocery stores using radiation model and survival analysis », *Journal of Transport Geography*, vol. 94, 2021, p. 103107, [https//doi.org/10.1016/j.jtrangeo.2021.103107].
- Jöreskog K. G., *LISREL V : analysis of linear structural relationships by maximum likelihood and least squares methods*, Version V. Uppsala : University of Uppsala, Dept. of Statistics ; Chicago : distributed by International Educational Services, [1981] ©1981, 1981.
- Jöreskog K. G., « A General Approach to Confirmatory Maximum Likelihood Factor Analysis », *Psychometrika*, nº 2, vol. 34, 1969, p. 183-202, [https//doi.org/10.1007/BF02289343].
- Jöreskog Karl G. et Sörbom Dag, *LISREL 6: analysis of linear structural relationships by maximum likelihood, instrumental variables and least squares methods; user's guide*, 4. ed., Mooresville, Ind, Scientific Software, Inc, 1986.
- Joubert Léo, Van Truoc Olivier Lê, Mercklé Pierre et Tudoux Benoît, « Redresser l'échantillon d'une enquête en ligne : un exemple à partir de l'enquête Vico », *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, nº 1, vol. 158, 2023, p. 143-166, [https//doi.org/10.1177/07591063231160287].
- Jubénot Marie-Noëlle et Eudes Daniel, *Analyse des données sous R pour les sciences humaines: théories et exemples commentés*, Paris, Ellipses, 2022.
- Judd Chalres M., McClelland Gary H., Ryan Carey S., Muller Dominique et Yzerbyt Vincent, *Analyse des données: une approche par comparaison de modèles*, 2e éd., Louvain-la-Neuve [Paris], De Boeck supérieur, coll. « Collection ouvertures psychologiques », 2018.
- Kabacoff Robert, *Modern Data Visualization with R*.
- Kaiser Henry F., « An Index of Factorial Simplicity », *Psychometrika*, nº 1, vol. 39, 1974, p. 31-36, [https//doi.org/10.1007/BF02291575].
- Kaiser Henry F., « A Second Generation Little Jiffy », *Psychometrika*, nº 4, vol. 35, 1970, p. 401-415, [https//doi.org/10.1007/BF02291817].
- Kaiser Henry F. et Rice John, « Little Jiffy, Mark Iv », *Educational and Psychological Measurement*, nº 1, vol. 34, 1974, p. 111-117, [https//doi.org/10.1177/001316447403400115].

- Kaplan E. L. et Meier Paul, « Nonparametric Estimation from Incomplete Observations », in Samuel Kotz et Norman L. Johnson (dir.), *Breakthroughs in Statistics*, New York, NY, Springer New York, 1992, p. 319-337, [https//doi.org/10.1007/978-1-4612-4380-9_25].
- Karch Julian D., « Outliers may not be automatically removed. », *Journal of Experimental Psychology: General*, nº 6, vol. 152, 2023, p. 1735-1753, [https//doi.org/10.1037/xge0001357].
- Kaufman L., Rousseeuw P. J., Mathematics Faculty of et Informatics (Delft), *Clustering by Means of Medoids*, Faculty of Mathematics and Informatics, coll. « Delft University of Technology : reports of the Faculty of Technical Mathematics and Informatics », 1987.
- Kendall M. G., « A NEW MEASURE OF RANK CORRELATION », *Biometrika*, nº 1-2, vol. 30, 1938, p. 81-93, [https//doi.org/10.1093/biomet/30.1-2.81].
- Klein Grady et Dabney Alan, *The cartoon introduction to statistics*, First edition., New York, Hill and Wang, a Division of Farrar, Straus and Giroux, 2013.
- Kleitman Sabina, Fullerton Dayna J., Zhang Lisa M., Blanchard Matthew D., Lee Jihyun, Stankov Lazar et Thompson Valerie, « To comply or not comply? A latent profile analysis of behaviours and attitudes during the COVID-19 pandemic », *PLOS ONE*, nº 7, vol. 16, 2021, p. e0255268, [https//doi.org/10.1371/journal.pone.0255268].
- Knief Ulrich et Forstmeier Wolfgang, « Violating the normality assumption may be the lesser of two evils », *Behavior Research Methods*, nº 6, vol. 53, 2021, p. 2576-2590, [https//doi.org/10.3758/s13428-021-01587-5].
- Kolmogorov A., « Sulla determinazione empirica di una legge di distribuzione », *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, 1933, p. 83-91.
- Kowalski Charles J., « On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient », *Applied Statistics*, nº 1, vol. 21, 1972, p. 1, [https//doi.org/10.2307/2346598].
- Krämer Walter et Sonnberger Harald, *The Linear Regression Model Under Test*, Heidelberg, Physica-Verlag HD, 1986, [https//doi.org/10.1007/978-3-642-95876-2].
- Kreager Derek A., Schaefer David R., Bouchard Martin, Haynie Dana L., Wakefield Sara, Young Jacob et Zajac Gary, « Toward a Criminology of Inmate Networks », *Justice Quarterly*, nº 6, vol. 33, 2016, p. 1000-1028, [https//doi.org/10.1080/07418825.2015.1016090].
- Kruskal William H. et Wallis W. Allen, « Use of Ranks in One-Criterion Variance Analysis », *Journal of the American Statistical Association*, nº 260, vol. 47, 1952, p. 583-621, [https//doi.org/10.1080/01621459.1952.10483441].
- Kuder G. F. et Richardson M. W., « The Theory of the Estimation of Test Reliability », *Psychometrika*, nº 3, vol. 2, 1937, p. 151-160, [https//doi.org/10.1007/BF02288391].
- Kuentz Simonet V., Lyser Sandrine, Candau Jacqueline, Deuffic Philippe, Chavent Marie et Saracco Jérôme, « Une approche par classification de variables pour la typologie d'observations : le cas d'une enquête agriculture et environnement », *Journal de la Societe Française de Statistique*, nº 2, vol. 154, 2013, p. 37-63.
- Kyriazos Theodoros A., « Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General », *Psychology*, nº 08, vol. 09, 2018, p. 2207-2230, [https//doi.org/10.4236/psych.2018.98126].
- Kyriazos Theodoros et Poga Mary, « Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions », *Open Journal of Statistics*, nº 03, vol. 13, 2023, p. 404-424, [https//doi.org/10.4236/ojs.2023.133020].
- Labenne Amaury, Chavent Marie, Kuentz Simonet V., Rambonilaza Mbolatiana et Saracco Jérôme, « Une extension de l'analyse factorielle multiple pour des groupes de variables mixtes: MFAmix », Toulouse, France.

- Lachenbruch Peter A., Sneeringer Cheryl et Revo Lawrence T., « Robustness of the linear and quadratic discriminant function to certain types of non-normality », *Communications in Statistics*, nº 1, vol. 1, 1973, p. 39-56, [https//doi.org/10.1080/03610927308827006].
- Lancelot A., « L'Orientation du comportement politique », *Traité de science politique*, Paris, Presses universitaires de France, 1985, .
- Lavaud-Legendre Bénédicte, Melançon Guy, Pinaud Bruno, Plessard Cécile et Feron Norbert, *Analyse et visualisation des réseaux criminels*, COMPTRASEC - CNRS - UMR 5114 ; LABRI - UMR 5800.
- Lavaud-Legendre Bénédicte, Plessard Cécile, Desquesnes Gillonne, Proia-Lelouey Nadine, Encrenaz Gaelle et Debruyne Gautier, *Prostitution de mineures - Parcours de vie des individus impliqués dans la prostitution par plans*, CNRS COMPTRASEC UMR 5114.
- Lazarsfeld Paul F., « Recent Developments in Latent Structure Analysis », *Sociometry*, nº 4, vol. 18, 1955, p. 391, [https//doi.org/10.2307/2785875].
- Lazarsfeld Paul F., « The logical and mathematical foundation of latent structure analysis », *Measurement and prediction. [Studies in social psychology in World War II. Vol.4.]*, Princeton, NJ, US, Princeton University Press, coll. « Measurement and prediction », 1950, .
- Lazega Emmanuel, *Réseaux sociaux et structures relationnelles:*, Presses Universitaires de France, coll. « Que sais-je ? », 2007, [https//doi.org/10.3917/puf.lazeg.2007.01].
- Lazega Emmanuel, « Analyse de réseaux d'une organisation collégiale : les avocats d'affaires », *Revue française de sociologie*, nº 4, vol. 33, 1992, p. 559-589, [https//doi.org/10.2307/3322226].
- Le Campion Grégoire, « Analyse des corrélations avec easystats: Guide pratique avec R », , 2021, [https//doi.org/10.48645/QHAV-CB52].
- Le Roux Brigitte et Rouanet Henry, *Multiple Correspondence Analysis*, 2455 Teller Road, Thousand Oaks California 91320 United States of America, SAGE Publications, Inc., 2010, [https//doi.org/10.4135/9781412993906].
- Le Saout Ronan et Floch Jean-Michel, *Économétrie spatiale : une introduction pratique - Documents de travail - M2016/06 | Insee*, [https://www.insee.fr/fr/statistiques/2408659].
- Lebart L. et Salem A., *Statistique Textuelle*, [http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html].
- Lebart Ludovic, Pincemin Bénédicte et Poudat Céline, *Analyse des données textuelles*, Presses de l'Université du Québec, coll. « Mesure et évaluation », 2019.
- Lebart Ludovic, Piron Marie et Morineau Alain, *Statistique exploratoire multidimensionnelle: Visualisation et inférences en fouille de données*, 4e éd (2006)., Paris, Dunod, coll. « Sciences SUP », 1995.
- Lejeune Christophe, *Manuel d'analyse qualitative: Analyser sans compter ni classer*, De Boeck Supérieur, 2019, [https//doi.org/10.3917/dbu.lejeu.2019.01].
- LeSage James et Pace Robert Kelley, *Introduction to Spatial Econometrics*, 0 éd., Chapman and Hall/CRC, 2009, [https//doi.org/10.1201/9781420064254].
- Lesnard Laurent, « Annexe 2. Les méthodes d'appariement optimal », *La famille désarticulée*, Paris cedex 14, Presses Universitaires de France, coll. « Le Lien social », 2009, p. 196-197.
- Lesnard Laurent et Saint Pol Thibaut de, « Introduction aux méthodes d'appariement optimal (Optimal Matching Analysis) », *Bulletin de méthodologie sociologique. Bulletin of sociological methodology*, nº 90, 2006, p. 5-25.
- Lévène H., « Robust Tests for Equality of Variances », *Contributions to probability and statistics; essays in honor of Harold Hotelling*, Stanford, Calif., Stanford University Press, coll. « Stanford studies in mathematics and statistics; 2 », 1960, .

- Lexicometrica Team, *JADT | Journées d'Analyses statistiques de Données Textuelles*, [http://lexicometrica.univ-paris3.fr/jadt/].
- Lexicometrica Team, *LEXICOMETRICA*, [http://lexicometrica.univ-paris3.fr/jadt/].
- Leys Christophe, Delacre Marie, Mora Youri L., Lakens Daniël et Ley Christophe, « How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration », *International Review of Social Psychology*, nº 1, vol. 32, 2019, p. 5, [https//doi.org/10.5334/irsp.289].
- Leys Christophe, Klein Olivier, Dominicy Yves et Ley Christophe, « Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance », *Journal of Experimental Social Psychology*, vol. 74, 2018, p. 150-156, [https//doi.org/10.1016/j.jesp.2017.09.011].
- Leys Christophe, Ley Christophe, Klein Olivier, Bernard Philippe et Licata Laurent, « Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median », *Journal of Experimental Social Psychology*, nº 4, vol. 49, 2013, p. 764-766, [https//doi.org/10.1016/j.jesp.2013.03.013].
- Lin Johnny, *Introduction to Structural Equation Modeling (SEM) in R with lavaan*, [https://stats.oarc.ucla.edu/r/seminars/rsem/].
- Ljung G. M. et Box G. E. P., « On a measure of lack of fit in time series models », *Biometrika*, nº 2, vol. 65, 1978, p. 297-303, [https//doi.org/10.1093/biomet/65.2.297].
- Loh Wei-Yin, « Fifty Years of Classification and Regression Trees », *International Statistical Review*, nº 3, vol. 82, 2014, p. 329-348, [https//doi.org/10.1111/insr.12016].
- Loonis Vincent et de Bellefon Marie-Pierre, *Manuel d'analyse spatiale | Insee*, [https://www.insee.fr/fr/information/3635442].
- Lüdecke Daniel, Ben-Shachar Mattan, Patil Indrajeet, Waggoner Philip et Makowski Dominique, « Performance: An R Package for Assessment, Comparison and Testing of Statistical Models », *Journal of Open Source Software*, nº 60, vol. 6, 2021, p. 3139, [https//doi.org/10.21105/joss.03139].
- Lugu Benjamin, *Exploratory Factor Analysis in R*.
- Luo Hongge, Zhao Yanli, Fan Fengmei, Fan Hongzhen, Wang Yunhui, Qu Wei, Wang Zhiren, Tan Yunlong, Zhang Xiujun et Tan Shuping, « A bottom-up model of functional outcome in schizophrenia », *Scientific Reports*, nº 1, vol. 11, 2021, p. 7577, [https//doi.org/10.1038/s41598-021-87172-4].
- Maaten Laurens van der et Hinton Geoffrey, « Visualizing Data using t-SNE », *Journal of Machine Learning Research*, nº 86, vol. 9, 2008, p. 2579-2605.
- MacQueen J., « Some methods for classification and analysis of multivariate observations », *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, University of California Press, 1967, p. 281-298.
- Makowski Dominique, « The psycho Package: an Efficient and Publishing-Oriented Workflow for Psychological Science », *The Journal of Open Source Software*, nº 22, vol. 3, 2018, p. 470, [https//doi.org/10.21105/joss.00470].
- Mann H. B. et Whitney D. R., « On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other », *The Annals of Mathematical Statistics*, nº 1, vol. 18, 1947, p. 50-60, [https//doi.org/10.1214/aoms/1177730491].
- Manski Charles F., « Identification of Endogenous Social Effects: The Reflection Problem », *The Review of Economic Studies*, nº 3, vol. 60, 1993, p. 531, [https//doi.org/10.2307/2298123].
- Manzo Gianluca, « Potentialités et limites de la simulation multi-agents : une introduction »:, *Revue française de sociologie*, nº 4, Vol. 55, 2014, p. 653-688, [https//doi.org/10.3917/rfs.554.0653].

- Marchand Pascal et Ratinaud Pierre, *Faut-il faire des nuages de mots ? — IRaMuTeQ*, [http://www.iramuteq.org/Members/pmarchand/faut-il-faire-des-nuages-de-mots].
- Maroco João, Silva Dina, Rodrigues Ana, Guerreiro Manuela, Santana Isabel et De Mendonça Alexandre, « Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests », *BMC Research Notes*, nº 1, vol. 4, 2011, p. 299, [https//doi.org/10.1186/1756-0500-4-299].
- Marquardt Donald W., « Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation », *Technometrics*, nº 3, vol. 12, 1970, p. 591-612, [https//doi.org/10.1080/00401706.1970.10488699].
- Martin Angélique, *Traitement des entretiens par analyse de contenu thématique in. Les jeunes, l'insertion et les missions locales du pays d'Auge (Normandie) : les évolutions des représentations sociales entre 1982 et 2017*, thèse de doctorat, Conservatoire national des arts et metiers - CNAM, 2018.
- Mason Lee, Otero Maria et Andrews Alonzo, « Cochran's Q Test of Stimulus Overselectivity within the Verbal Repertoire of Children with Autism », *Perspectives on Behavior Science*, nº 1, vol. 45, 2022, p. 101-121, [https//doi.org/10.1007/s40614-021-00315-w].
- Massé Antoine, *Aide à l'utilisation de R - Se former à R*, [https://sites.google.com/site/rgraphiques/se-former-a-r].
- Matejka Justin et Fitzmaurice George, *The Datasaurus data package*, 2025.
- Matejka Justin et Fitzmaurice George, « Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing », Denver Colorado USA, ACM.
- Mauchly John W., « Significance Test for Sphericity of a Normal n-Variate Distribution », *The Annals of Mathematical Statistics*, nº 2, vol. 11, 1940, p. 204-209, [https//doi.org/10.1214/aoms/1177731915].
- McDonald Roderick P., « Generalizability in Factorable Domains: "Domain Validity and Generalizability"1 », *Educational and Psychological Measurement*, nº 1, vol. 38, 1978, p. 75-79, [https//doi.org/10.1177/001316447803800111].
- McFadden D., *Conditional Logit Analysis of Qualitative Choice Behavior*, Institute of Urban and Regional Development, University of California, 1973.
- Mcgill Robert, Tukey John W. et Larsen Wayne A., « Variations of Box Plots », *The American Statistician*, nº 1, vol. 32, 1978, p. 12-16, [https//doi.org/10.1080/00031305.1978.10479236].
- McGuire Austen, Huffhines Lindsay et Jackson Yo, « The trajectory of PTSD among youth in foster care: A survival analysis examining maltreatment experiences prior to entry into care », *Child Abuse & Neglect*, vol. 115, 2021, p. 105026, [https//doi.org/10.1016/j.chiabu.2021.105026].
- McInnes Leland, Healy John et Melville James, « UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction », [https//doi.org/10.48550/arXiv.1802.03426].
- McInnes Leland, Healy John, Saul Nathaniel et Großberger Lukas, « UMAP: Uniform Manifold Approximation and Projection », *Journal of Open Source Software*, nº 29, vol. 3, 2018, p. 861, [https//doi.org/10.21105/joss.00861].
- Menard Scott, *Logistic Regression: From Introductory to Advanced Concepts and Applications*, 2455 Teller Road, Thousand Oaks California 91320 United States, SAGE Publications, Inc., 2010, [https//doi.org/10.4135/9781483348964].

- Mercklé Pierre, « Les méthodes d'équations structurelles (MES) : Pour qui ? Pour quoi faire ? Comment ça marche ? par Alain Lacroux (Vendredis Quanti, 31 janvier 2020) », , 2020, [https//doi.org/10.58079/T4CF].
- Mercklé Pierre, *La sociologie des réseaux sociaux:*, La Découverte, coll. « Repères », 2011, [https//doi.org/10.3917/dec.merck.2011.01].
- Messu Michel, *L'analyse des relations par opposition*, [https://www.credoc.fr/publications/lanalyse-des-relations-par-opposition].
- Messu Michel, *L'analyse de contenu : premiers éléments de réflexion*, [https://www.credoc.fr/publications/lanalyse-de-contenu-premiers-elements-de-reflexion].
- Meteyard Lotte et Davies Robert A. I., « Best practice guidance for linear mixed-effects models in psychological science », *Journal of Memory and Language*, vol. 112, 2020, p. 104092, [https//doi.org/10.1016/j.jml.2020.104092].
- MetricGate Team, *MetricGate | AI-Powered Statistical Analysis & R Integration*, [https://metricgate.com/docs/tschuprows-t/].
- Miratrix Luke W., Sekhon Jasjeet S., Theodoridis Alexander G. et Campos Luis F., « Worth Weighting? How to Think About and Use Weights in Survey Experiments », *Political Analysis*, nº 3, vol. 26, 2018, p. 275-291, [https//doi.org/10.1017/pan.2018.1].
- Mistral AI, *Frontier AI LLMs, assistants, agents, services | Mistral AI - Le Chat (September 2025 version) [Large language model]*, [https://mistral.ai/].
- Montero Lídia, Mejía-Dorantes Lucía et Barceló Jaume, « The role of life course and gender in mobility patterns: a spatiotemporal sequence analysis in Barcelona », *European Transport Research Review*, nº 1, vol. 15, 2023, p. 44, [https//doi.org/10.1186/s12544-023-00621-1].
- Moran P. A. P., « NOTES ON CONTINUOUS STOCHASTIC PHENOMENA », *Biometrika*, nº 1-2, vol. 37, 1950, p. 17-23, [https//doi.org/10.1093/biomet/37.1-2.17].
- Moreno J. L., *Who shall survive?: A new approach to the problem of human interrelations.*, Washington, Nervous and Mental Disease Publishing Co, 1934, [https//doi.org/10.1037/10648-000].
- Morgan James N. et Sonquist John A., « Problems in the Analysis of Survey Data, and a Proposal », *Journal of the American Statistical Association*, nº 302, vol. 58, 1963, p. 415-434, [https//doi.org/10.1080/01621459.1963.10500855].
- Mosier Charles I., « A Note on Item Analysis and the Criterion of Internal Consistency », *Psychometrika*, nº 4, vol. 1, 1936, p. 275-282, [https//doi.org/10.1007/BF02287879].
- Mucchielli Alex, *Dictionnaire des méthodes qualitatives en sciences humaines et sociales*, Paris, A. Colin Masson, coll. « U », 1996.
- Muller C., *Principes et méthodes de statistique lexicale*, Hachette, 1977.
- Muller C., *Initiation à la statistique linguistique*, Larousse, 1968.
- Muller Jean-Claude, « Comparaison visuelle des cartes et groupements spatiaux », *L'Espace géographique*, nº 1, vol. 6, 1977, p. 59-72, [https//doi.org/10.3406/spgeo.1977.1694].
- Nahhas Ramzi W., *5.17 Checking the linearity assumption | Introduction to Regression Methods for Public Health Using R*, 2025.
- Negura Lilian, « L'analyse de contenu dans l'étude des représentations sociales », *SociologieS*, , 2006, [https//doi.org/10.4000/sociologies.993].
- Nesselroade John R. et Baltes Paul B. (dir.), *Longitudinal research in the study of behavior and development*, New York, NY, Academic Press, 1979.
- Nguyen Mike, *A Guide on Data Analysis*, 2020.
- Noichl Maximilian, « Modeling the structure of recent philosophy », *Synthese*, nº 6, vol. 198, 2021, p. 5089-5100, [https//doi.org/10.1007/s11229-019-02390-8].

- Nuzzo Regina, « Scientific method: Statistical errors », *Nature*, nº 7487, vol. 506, 2014, p. 150-152, [https//doi.org/10.1038/506150a].
- Observatoire du développement Durable, *Sous portail odr — Wiki ODR*, [https://odr.inrae.fr/intranet/carto/cartowiki/index.php/Sous_portail_odr].
- Ognyanova Katya, « Static and dynamic network visualization with R ».
- Oliveau Sébastien, « Autocorrélation spatiale ».
- Oliveau Sébastien, *L'espace compte ! Mesurer les structures spatiales du changement social.*, thèse de doctorat, Université d'Aix-Marseille 1, 2011.
- Oliveau Sébastien, « Autocorrélation spatiale : leçons du changement d'échelle: », *L'Espace géographique*, nº 1, Vol. 39, 2010, p. 51-64, [https//doi.org/10.3917/eg.391.0051].
- Oliveau Sébastien et Doignon Yoann, « La diagonale se vide ? Analyse spatiale exploratoire des décroissances démographiques en France métropolitaine depuis 50 ans », *Cybergeo*, , 2016, [https//doi.org/10.4000/cybergeo.27439].
- Organisation Internationale du Travail, *Revue internationale du Travail*, [https://my.visme.co/view/4dyy1m3y-revue-internationale-du-travail].
- Padilla Miguel A. et Divers Jasmin, « A Comparison of Composite Reliability Estimators: Coefficient Omega Confidence Intervals in the Current Literature », *Educational and Psychological Measurement*, nº 3, vol. 76, 2016, p. 436-453, [https//doi.org/10.1177/0013164415593776].
- Pagès J., « Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes », *Revue de Statistique Appliquée*, nº 4, vol. 50, 2002, p. 5-37.
- Pages J. P., Cailliez F. et Escoufier Y., « Analyse factorielle : un peu d'histoire et de géométrie », *Revue de Statistique Appliquée*, nº 1, vol. 27, 1979, p. 5-28.
- Pagès Jérôme, *Analyse factorielle multiple avec R*, Les Ulis, EDP sciences, coll. « Pratique R », 2013.
- Paillé Pierre, « L'analyse par théorisation ancrée », *Cahiers de recherche sociologique*, nº 23, 1994, p. 147-181, [https//doi.org/10.7202/1002253ar].
- Paillé Pierre et Mucchielli Alex, *L'analyse qualitative en sciences humaines et sociales*, Armand Colin, 2012, [https//doi.org/10.3917/arco.paill.2012.01].
- Pavlov Ivan P., « The Scientific Investigation of the Psychical Faculties or Processes in the Higher Animals », *Science*, nº 620, vol. 24, 1906, p. 613-619, [https//doi.org/10.1126/science.24.620.613].
- Pearson Karl, « LIII. *On lines and planes of closest fit to systems of points in space* », *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, nº 11, vol. 2, 1901, p. 559-572, [https//doi.org/10.1080/14786440109462720].
- Pearson Karl, « X. *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling* », *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, nº 302, vol. 50, 1900, p. 157-175, [https//doi.org/10.1080/14786440009463897].
- Pearson Karl, « VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia », *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, 1896, p. 253-318, [https//doi.org/10.1098/rsta.1896.0007].

- Pearson Karl, « VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia », *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, 1896, p. 253-318, [https//doi.org/10.1098/rsta.1896.0007].

- Peduzzi Peter, Concato John, Kemper Elizabeth, Holford Theodore R. et Feinstein Alvan R., « A simulation study of the number of events per variable in logistic regression analysis », *Journal of Clinical Epidemiology*, nº 12, vol. 49, 1996, p. 1373-1379, [https//doi.org/10.1016/S0895-4356(96)00236-3].
- Pennycook Gordon, Cheyne James Allan, Koehler Derek J. et Fugelsang Jonathan A., « On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs », *Judgment and Decision Making*, nº 4, vol. 15, 2020, p. 476-498, [https//doi.org/10.1017/S1930297500007439].
- Perrineau Pascal, Chiche Jean, Le Roux Brigitte et Rouanet Henry, « L'espace politique des électeurs français à la fin des années 1990. Nouveaux et anciens clivages, hétérogénéité des électoratsé », *Revue française de science politique*, nº 3, vol. 50, 2000, p. 463-488, [https//doi.org/10.3406/rfsp.2000.395484].
- Pilkington Hugo, Feuillet Thierry, Rican Stéphane, Goupil De Bouillé Jeanne, Bouchaud Olivier, Cailhol Johann, Bihan Hélène, Lombrail Pierre et Julia Chantal, « Spatial determinants of excess all-cause mortality during the first wave of the COVID-19 epidemic in France », *BMC Public Health*, nº 1, vol. 21, 2021, p. 2157, [https//doi.org/10.1186/s12889-021-12203-8].
- Pincemin Bénédicte, « La textométrie en question », *Le Français Moderne - Revue de linguistique Française*, nº 1, vol. 88, 2020, p. 26.
- PINCEMIN Bénédicte et HEIDEN Serge, *Qu'est-ce que la textométrie ? Présentation*, [https://txm.gitpages.huma-num.fr/textometrie/Introduction/].
- Playfair (1759–1823) William, *The Commercial and Political Atlas, Representing, By Means of Stained Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, During the Whole of the Eighteenth Century.*, [https://fisherdigitus.library.utoronto.ca/document/7750].
- Ponkilainen Maria, Einiö Elina, Pietiläinen Marjut et Myrskylä Mikko, « Educational Differences in Fertility Among Female Same-Sex Couples in Finland », *Demography*, nº 6, vol. 61, 2024, p. 2053-2079, [https//doi.org/10.1215/00703370-11687583].
- Pornprasertmanit Sunthud, Miller Patrick et Schoemann Alexander, *Simsem*, [https://simsem.org/].
- Pregibon Daryl, « Logistic Regression Diagnostics », *The Annals of Statistics*, nº 4, vol. 9, 1981, [https//doi.org/10.1214/aos/1176345513].
- Prouteau Antoinette, Roux Solenne, Destaillats Jean-Marc et Bergua Valérie, « Profiles of Relationships Between Subjective and Objective Cognition in Schizophrenia: Associations With Quality of Life, Stigmatization, and Mood Factors », *Journal of Cognitive Education and Psychology*, nº 1, vol. 16, 2017, p. 64-76, [https//doi.org/10.1891/1945-8959.16.1.64].
- Pumain Denise, « La composition socio-professionnelle des villes françaises : essai de typologie par analyse des correspondances et classification automatique », *L'Espace géographique*, nº 4, vol. 5, 1976, p. 227-238, [https//doi.org/10.3406/spgeo.1976.1663].
- Raffaillac Thibault, Boukhelifa Nadia, Crouzat Emilie, Stark Fabien, Müller Jean-Pierre et Lasseur Jacques, « Développement d'une interface de simulation multi-agents pour la gestion concertée des territoires pastoraux de moyenne montagne », Troyes, France.
- Ram Nilam et Grimm Kevin, « Using simple and complex growth models to articulate developmental change: Matching theory to method », *International Journal of Behavioral Development*, nº 4, vol. 31, 2007, p. 303-316, [https//doi.org/10.1177/0165025407077751].
- Rao C. Radhakrishna, « Some Statistical Methods for Comparison of Growth Curves », *Biometrics*, nº 1, vol. 14, 1958, p. 1, [https//doi.org/10.2307/2527726].

- Raymond Henri, « Analyse de contenu et entretien non-directif : application au symbolisme de l'habitat », *Revue française de sociologie*, nº 2, vol. 9, 1968, p. 167-179, [https//doi.org/10.2307/3320589].
- Ren Lijuan, Wang Tao, Sekhari Seklouli Aicha, Zhang Haiqing et Bouras Abdelaziz, « A review on missing values for main challenges and methods », *Information Systems*, vol. 119, 2023, p. 102268, [https//doi.org/10.1016/j.is.2023.102268].
- Revelle William, *Psych: Procedures for Psychological, Psychometric, and Personality Research*, 2025.
- Revelle William et Zinbarg Richard E., « Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma », *Psychometrika*, nº 1, vol. 74, 2009, p. 145-154, [https//doi.org/10.1007/s11336-008-9102-z].
- Riani Marco, Atkinson Anthony C. et Corbellini Aldo, « Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression », *Statistical Methods & Applications*, nº 1, vol. 32, 2023, p. 75-102, [https//doi.org/10.1007/s10260-022-00640-7].
- Richer Cyprien et Palmier Patrick, « Mesurer l'accessibilité territoriale par les transports collectifs. Proposition méthodologique appliquée aux pôles d'excellence de Lille Métropole », *Cahiers de géographie du Québec*, nº 158, vol. 56, 2012, p. 31.
- Robette Nicolas, *GDAtools: Geometric Data Analysis in R. version 2.0.*, [https://nicolas-robette.github.io/GDAtools/].
- Robette Nicolas, « Mesurer la dissemblance entre trajectoires », *L'analyse statistique des trajectoires: Typologies de séquences et autres approches*, Ined Éditions, 2021, , [https//doi.org/10.4000/books.ined.16670].
- Rokotyan Nikita, Stukova Olya et Kolmakova Dasha, *An alternative data-driven country map*, [https://projects.interacta.io/country-tsne].
- Rosenberg Joshua, Beymer Patrick, Anderson Daniel, Van Lissa C. j. et Schmidt Jennifer, « tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software », *Journal of Open Source Software*, nº 30, vol. 3, 2018, p. 978, [https//doi.org/10.21105/joss.00978].
- Rosenberg Joshua M. et van Lissa Caspar, *Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software*, [https://data-edu.github.io/tidyLPA/].
- Rosseel Yves, « Lavaan: An R Package for Structural Equation Modeling », *Journal of Statistical Software*, vol. 48, 2012, p. 1-36, [https//doi.org/10.18637/jss.v048.i02].
- Roy Marie-Hélène et Larocque Denis, « Prediction intervals with random forests », *Statistical Methods in Medical Research*, nº 1, vol. 29, 2020, p. 205-229, [https//doi.org/10.1177/0962280219829885].
- Rubin Donald B., *Multiple Imputation for Nonresponse in Surveys*, 1re éd., Wiley, coll. « Wiley Series in Probability and Statistics », 1987, [https//doi.org/10.1002/9780470316696].
- Saint-Julien Thérèse, *La diffusion spatiale des innovations*, Montpellier, Gip Reclus, coll. « Reclus modes d'emploi », 1985.
- Sanders Léna, *L'analyse statistique des données en géographie*, MONTPELLIER, DIFFUSION: LA DOCUMENTATION FRANCAISE,PARIS, ED.GIP RECLUS, coll. « ALIDADE », 1989.
- Saporta Gilbert, *Probabilités, analyse des données et statistique*, 3e éd. révisée., Paris, Éd. Technip, 2011.
- Sarafidis Vasilis et Wansbeek Tom, « Celebrating 40 years of panel data analysis: Past, present and future », *Journal of Econometrics*, nº 2, vol. 220, 2021, p. 215-226, [https//doi.org/10.1016/j.jeconom.2020.06.001].

- Savage Mike, Devine Fiona, Cunningham Niall, Taylor Mark, Li Yaojun, Hjellbrekke Johs, Le Roux Brigitte, Friedman Sam et Miles Andrew, « A New Model of Social Class? Findings from the BBC's Great British Class Survey Experiment », *Sociology*, nº 2, vol. 47, 2013, p. 219-250, [https//doi.org/10.1177/0038038513481128].
- Schafer Joseph L., « Multiple imputation: a primer », *Statistical Methods in Medical Research*, nº 1, vol. 8, 1999, p. 3-15, [https//doi.org/10.1177/096228029900800102].
- Schafer Joseph L. et Graham John W., « Missing data: our view of the state of the art », *Psychological Methods*, nº 2, vol. 7, 2002, p. 147-177.
- Schelling Thomas C., « Dynamic models of segregation† », *The Journal of Mathematical Sociology*, nº 2, vol. 1, 1971, p. 143-186, [https//doi.org/10.1080/0022250X.1971.9989794].
- Schtickzelle Martial, « Pierre-François Verhulst (1804-1849). La première découverte de la fonction logistique »:, *Population*, nº 3, Vol. 36, 1981, p. 541-556, [https//doi.org/10.3917/popu.p1981.36n3.0556].
- Sebalo Ivan, Ball Linden J., Marsh John E., Morley Andy M., Richardson Beth H., Taylor Paul J. et Threadgold Emma, « Conspiracy theories: why they are believed and how they can be challenged », *Journal of Cognitive Psychology*, nº 4, vol. 35, 2023, p. 383-400, [https//doi.org/10.1080/20445911.2023.2198064].
- Setu Sarmistha Paul, Kabir Rasel, Islam Md. Akhtarul, Alauddin Sharlene et Nahar Mst. Tanmin, « Factors associated with time to first birth interval among ever married Bangladeshi women: A comparative analysis on Cox-PH model and parametric models », *PLOS Global Public Health*, nº 12, vol. 4, 2024, p. e0004062, [https//doi.org/10.1371/journal.pgph.0004062].
- Shapiro S. S. et Wilk M. B., « An analysis of variance test for normality (complete samples) », *Biometrika*, nº 3-4, vol. 52, 1965, p. 591-611, [https//doi.org/10.1093/biomet/52.3-4.591].
- Sijtsma Klaas, « On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha », *Psychometrika*, nº 1, vol. 74, 2009, p. 107-120, [https//doi.org/10.1007/s11336-008-9101-0].
- Simon-Bouhet Benoît, *4 Visualiser des données avec ggplot2 | Travaux Pratiques de Biométrie 2*, 2022.
- Simpson E. H., « The Interpretation of Interaction in Contingency Tables », *Journal of the Royal Statistical Society Series B: Statistical Methodology*, nº 2, vol. 13, 1951, p. 238-241, [https//doi.org/10.1111/j.2517-6161.1951.tb00088.x].
- Sinha Pratik, Calfee Carolyn S. et Delucchi Kevin L., « Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls », *Critical Care Medicine*, nº 1, vol. 49, 2021, p. e63-e79, [https//doi.org/10.1097/CCM.0000000000004710].
- Sjoberg Daniel, *Function to display multinomial regression models in wide format*, [https://gist.github.com/ddsjoberg/a55afa74ac58e1f895862fcabab62406].
- Slupphaug K. S., *Latent Interaction (and Moderation) Analysis in Structural Equation Models (SEM)*, [https://modsem.org/].
- Smirnov N., « Table for Estimating the Goodness of Fit of Empirical Distributions », *The Annals of Mathematical Statistics*, nº 2, vol. 19, 1948, p. 279-281, [https//doi.org/10.1214/aoms/1177730256].
- Sommet Nicolas et Morselli Davide, « Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS », *International Review of Social Psychology*, nº 1, vol. 30, 2017, p. 203-218, [https//doi.org/10.5334/irsp.90].
- Spearman C., « CORRELATION CALCULATED FROM FAULTY DATA », *British Journal of Psychology, 1904-1920*, nº 3, vol. 3, 1910, p. 271-295, [https//doi.org/10.1111/j.2044-8295.1910.tb00206.x].

- Spearman C., « The Proof and Measurement of Association between Two Things », *The American Journal of Psychology*, n° 72-101, vol. 15, 1904, [https//doi.org/10.2307/1422689].
- Spielberger Charles D., Gorsuch R. L. et Lushene R. E., « Manual for the State-Trait Anxiety Inventory (STAI) », , 1970.
- Stanton Jeffrey M., « Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors », *Journal of Statistics Education*, n° 3, vol. 9, 2001, p. 3, [https//doi.org/10.1080/10691898.2001.11910537].
- Steiner Markus et Grieder Silvia, *EFAtools: Fast and Flexible Implementations of Exploratory Factor Analysis Tools*, 2020.
- Stigler Stephen M., « Gergonne's 1815 paper on the design and analysis of polynomial regression experiments », *Historia Mathematica*, n° 4, vol. 1, 1974, p. 431-439, [https//doi.org/10.1016/0315-0860(74)90033-0].
- Stigler Stephen M., « Studies in the History of Probability and Statistics. XXXII: Laplace, Fisher, and the discovery of the concept of sufficiency », *Biometrika*, n° 3, vol. 60, 1973, p. 439-445, [https//doi.org/10.1093/biomet/60.3.439].
- Student, « The Probable Error of a Mean », *Biometrika*, n° 1, vol. 6, 1908, p. 1, [https//doi.org/10.2307/2331554].
- Sulik Justin, Deroy Ophelia, Dezecache Guillaume, Newson Martha, Zhao Yi, El Zein Marwa et Tunçgenç Bahar, « Facing the pandemic with trust in science », *Humanities and Social Sciences Communications*, n° 1, vol. 8, 2021, p. 301, [https//doi.org/10.1057/s41599-021-00982-9].
- Sun Yongbing, Song Bing, Zhen Cheng, Zhang Chao, Cheng Juan et Jiang Tianjun, « The mediating effect of psychological resilience between social support and anxiety/depression in people living with HIV/AIDS–a study from China », *BMC Public Health*, n° 1, vol. 23, 2023, p. 2461, [https//doi.org/10.1186/s12889-023-17403-y].
- Tabachnick Barbara et Fidell Linda, *Using Multivariate Statistics*, Pearson International, 2013.
- Teeters Jenni B., Ginley Meredith K., Whelan James P., Meyers Andrew W. et Pearlson Godfrey D., « The Moderating Effect of Gender on the Relation Between Expectancies and Gambling Frequency Among College Students », *Journal of Gambling Studies*, n° 1, vol. 31, 2015, p. 173-182, [https//doi.org/10.1007/s10899-013-9409-2].
- Tejo Mauricio, Araya Héctor, Niklitschek-Soto Sebastián et Marmolejo-Ramos Fernando, « Theoretical models of reaction times arising from simple-choice tasks », *Cognitive Neurodynamics*, n° 4, vol. 13, 2019, p. 409-416, [https//doi.org/10.1007/s11571-019-09532-1].
- Tobacyk Jerome J., « A Revised Paranormal Belief Scale », *International Journal of Transpersonal Studies*, n° 1, vol. 23, 2004, p. 94-98, [https//doi.org/10.24972/ijts.2004.23.1.94].
- Tobler W. R., « A Computer Movie Simulating Urban Growth in the Detroit Region », *Economic Geography*, vol. 46, 1970, p. 234, [https//doi.org/10.2307/143141].
- Tollenaar N. et Van Der Heijden P. G. M., « Which Method Predicts Recidivism Best?: A Comparison of Statistical, Machine Learning and Data Mining Predictive Models », *Journal of the Royal Statistical Society Series A: Statistics in Society*, n° 2, vol. 176, 2013, p. 565-584, [https//doi.org/10.1111/j.1467-985X.2012.01056.x].
- Trang Le Mai, Ha Dinh Thi, Son Dao The, Lan Ninh Thi Hoang et Anh Tran Kim, « Survival analysis of unemployment duration: a case study of Vietnam », *Journal of Education and Work*, n° 1-4, vol. 37, 2024, p. 216-233, [https//doi.org/10.1080/13639080.2024.2383562].
- Tschuprow a a, *Principles Of The Mathematical Theory Of Correlation.*, 1939.

- Tucker Ledyard R., « Determination of Parameters of a Functional Relation by Factor Analysis », *Psychometrika*, nº 1, vol. 23, 1958, p. 19-23, [https//doi.org/10.1007/BF02288975].
- Tukey John Wilder, *Exploratory data analysis*, Springer, 1977.
- Uchino Takashi, Okubo Ryo, Takubo Youji, Aoki Akiko, Wada Izumi, Hashimoto Naoki, Ikezawa Satoru et Nemoto Takahiro, « Mediation Effects of Social Cognition on the Relationship between Neurocognition and Social Functioning in Major Depressive Disorder and Schizophrenia Spectrum Disorders », *Journal of Personalized Medicine*, nº 4, vol. 13, 2023, p. 683, [https//doi.org/10.3390/jpm13040683].
- Upton Graham J. G., Fingleton Bernard et Upton Graham J. G., *Point pattern and quantitative data*, Reprinted., Chichester, Wiley, coll. « Spatial data analysis by example / Graham J. G. Upton; Bernard Fingleton », 1985.
- Utts Jessica M., « The rainbow test for lack of fit in regression », *Communications in Statistics - Theory and Methods*, nº 24, vol. 11, 1982, p. 2801-2815, [https//doi.org/10.1080/03610928208828423].
- Ward Joe H., « Hierarchical Grouping to Optimize an Objective Function », *Journal of the American Statistical Association*, nº 301, vol. 58, 1963, p. 236-244, [https//doi.org/10.1080/01621459.1963.10500845].
- Watkins Marley W., « Exploratory Factor Analysis: A Guide to Best Practice », *Journal of Black Psychology*, nº 3, vol. 44, 2018, p. 219-246, [https//doi.org/10.1177/0095798418771807].
- Watkins Marley W., « The reliability of multidimensional neuropsychological measures: from alpha to omega », *The Clinical Neuropsychologist*, nº 6-7, vol. 31, 2017, p. 1113-1126, [https//doi.org/10.1080/13854046.2017.1317364].
- Watts Duncan et Strogatz Steven, « Collective dynamics of 'small-world' networks », *Nature*, nº 6684, vol. 393, 1998, p. 440-442, [https//doi.org/10.1038/30918].
- Weller Bridget E., Bowen Natasha K. et Faubert Sarah J., « Latent Class Analysis: A Guide to Best Practice », *Journal of Black Psychology*, nº 4, vol. 46, 2020, p. 287-311, [https//doi.org/10.1177/0095798420930932].
- Werth Rose, *Categorical Regression in Stata and R*, Bookdown, 2022.
- Whalley Ben, *Just Enough R*.
- Wheeler David C., « Simultaneous Coefficient Penalization and Model Selection in Geographically Weighted Regression: The Geographically Weighted Lasso », *Environment and Planning A: Economy and Space*, nº 3, vol. 41, 2009, p. 722-742, [https//doi.org/10.1068/a40256].
- Wheeler David C., « Diagnostic Tools and a Remedial Method for Collinearity in Geographically Weighted Regression », *Environment and Planning A: Economy and Space*, nº 10, vol. 39, 2007, p. 2464-2481, [https//doi.org/10.1068/a38325].
- Wheeler David et Tiefelsdorf Michael, « Multicollinearity and correlation among local regression coefficients in geographically weighted regression », *Journal of Geographical Systems*, nº 2, vol. 7, 2005, p. 161-187, [https//doi.org/10.1007/s10109-005-0155-6].
- White Harrison C., Grossetti Michel et Godart Frédéric, *Identité et contrôle: une théorie de l'émergence des formations sociales*, Nouvelle éd. révisée., Paris, Éd. de l'EHESS, coll. « EHESS translations », 2011.
- Wilcoxon Frank, « Individual Comparisons by Ranking Methods », *Biometrics Bulletin*, nº 6, vol. 1, 1945, p. 80, [https//doi.org/10.2307/3001968].
- Wilkinson Leland, *The Grammar of Graphics*, New York, Springer-Verlag, coll. « Statistics and Computing », 2005, [https//doi.org/10.1007/0-387-28695-0].

- Williams Matt N., « Levels of measurement and statistical analyses », [https//doi.org/10.31234/osf.io/c5278].
- Williams Richard, « Understanding and interpreting generalized ordered logit models », *The Journal of Mathematical Sociology*, nº 1, vol. 40, 2016, p. 7-20, [https//doi.org/10.1080/0022250X.2015.1112384].
- Winship Christopher et Radbill Larry, « Sampling Weights and Regression Analysis », *Sociological Methods & Research*, nº 2, vol. 23, 1994, p. 230-257, [https//doi.org/10.1177/0049124194023002004].
- Wishart J., « GROWTH-RATE DETERMINATIONS IN NUTRITION STUDIES WITH THE BACON PIG, AND THEIR ANALYSIS », *Biometrika*, nº 1-2, vol. 30, 1938, p. 16-28, [https//doi.org/10.1093/biomet/30.1-2.16].
- Wong David W., « Exploring Spatial Patterns Using an Expanded Spatial Autocorrelation Framework: Exploring Spatial Patterns », *Geographical Analysis*, nº 3, vol. 43, 2011, p. 327-338, [https//doi.org/10.1111/j.1538-4632.2011.00816.x].
- Wong M. Anthony, « A Hybrid Clustering Method for Identifying High-Density Clusters », *Journal of the American Statistical Association*, nº 380, vol. 77, 1982, p. 841-847, [https//doi.org/10.1080/01621459.1982.10477896].
- Woodworth R. S., « How emotions are identified and classified », *Feelings and emotions: The Wittenberg Symposium*, Oxford, England, Clark Univ. Press, 1928, p. 222-227.
- Wooldridge Michael J., *An introduction to multiagent systems*, 2. ed., repr (1st ed. 2009)., Chichester, Wiley, 2012.
- Wright P. G., *The Tariff on Animal and Vegetable Oils*, Macmillan, 1928.
- Wright Sewall, « The Method of Path Coefficients », *The Annals of Mathematical Statistics*, nº 3, vol. 5, 1934, p. 161-215, [https//doi.org/10.1214/aoms/1177732676].
- Wright Sewall, « THE THEORY OF PATH COEFFICIENTS A REPLY TO NILES'S CRITICISM », *Genetics*, nº 3, vol. 8, 1923, p. 239-255, [https//doi.org/10.1093/genetics/8.3.239].
- Wright Sewall, « The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs », *Proceedings of the National Academy of Sciences*, nº 6, vol. 6, 1920, p. 320-332, [https//doi.org/10.1073/pnas.6.6.320].
- Yaddanapudi LakshmiNarayana, « The American Statistical Association statement on *P* -values explained », *Journal of Anaesthesiology Clinical Pharmacology*, nº 4, vol. 32, 2016, p. 421, [https//doi.org/10.4103/0970-9185.194772].
- Ye Zeng Jie, Zhang Zhang, Tang Ying, Liang Jian, Sun Zhe, Hu Guang Yun, Liang Mu Zi et Yu Yuan Liang, « Resilience patterns and transitions in the Be Resilient To Breast Cancer trial: an exploratory latent profile transition analysis », *Psycho-Oncology*, nº 6, vol. 30, 2021, p. 901-909, [https//doi.org/10.1002/pon.5668].
- Zhang Chunyu et Liu Liping, « The effect of job crafting to job performance », *Knowledge Management Research & Practice*, nº 2, vol. 19, 2021, p. 253-262, [https//doi.org/10.1080/14778238.2020.1762517].
- Zhang Yu et Miller Eric J., « Predicting housing construction period based on a cox proportional hazard model—an empirical study of housing completions in the greater Toronto and Hamilton area », *Environment and Planning B: Urban Analytics and City Science*, nº 6, vol. 50, 2023, p. 1624-1644, [https//doi.org/10.1177/23998083221143386].
- Zhukov Yuri. M., *Applied Spatial Statistics in R*, [https://zhukovyuri.github.io/teaching/].
- Zinbarg Richard E., Revelle William, Yovel Iftah et Li Wen, « Cronbach's α , Revelle's β , and Mcdonald's ω_H : their Relations with Each Other and Two Alternative Conceptualizations of

Reliability », *Psychometrika*, nº 1, vol. 70, 2005, p. 123-133, [https//doi.org/10.1007/s11336-003-0974-7].