Godelieve Masuy-Stroobant & Rafael Costa (dir.)

Analyser les données en sciences sociales

De la préparation des données à l'analyse multivariée

MPA APM
Méthodes participatives appliquées
Applied participatory methods

P.I.E. Peter Lang

Godelieve Masuy-Stroobant & Rafael Costa (dir.)

Analyser les données en sciences sociales

De la préparation des données à l'analyse multivariée

MPA APM
Méthodes participatives appliquées
Applied participatory methods

P.I.E. Peter Lang

AVANT-PROPOS
L'analyse des données

Godelieve MASUY-STROOBANT

Nos « sociétés de l'information » voient se multiplier les bases de données administratives et les enquêtes *ad hoc* le plus souvent destinées à mieux comprendre et gérer la complexité du social. Nul ne contestera aujourd'hui que le développement des politiques sociales à mettre en place pour faire face au défi du vieillissement de la population, améliorer l'accès des jeunes à l'emploi, permettre aux parents de mieux concilier travail et famille, rompre l'isolement social, éradiquer la pauvreté, soutenir les solidarités intergénérationnelles et interculturelles devrait idéalement se baser sur une analyse approfondie et critique de l'information disponible ou à recueillir. Force est cependant de constater que leur exploitation se limite encore trop souvent à des analyses descriptives **univariées**, voire **bivariées**. Or, en résonance à la complexité du social, l'analyse **multivariée** des données s'impose. Elle permet, non seulement d'affiner l'analyse **descriptive**, mais aussi de mieux **comprendre** ou **expliquer** les mécanismes multiples qui sous-tendent la plupart des phénomènes sociaux.

L'analyse des données nécessite le recours à l'outil statistique, mais ne se résume pas à la statistique : il faut d'abord et avant tout disposer d'une bonne connaissance de la société analysée, de son organisation, des valeurs qu'elle véhicule, et plus précisément du phénomène – ou du problème – social qui fera l'objet de l'analyse. Pour cela, explorer la littérature sur le sujet, ainsi que les théories développées à propos des mécanismes qui en font un problème à résoudre ou un phénomène nécessitant une analyse approfondie, est indispensable. Des informations précises sur le matériau qui sera analysé sont tout aussi importantes : s'agit-il d'un enregistrement imposé par l'administration ? Les informations recueillies font-elles l'objet de vérifications sur la base de documents officiels ? Dans le cas d'enquêtes, connaître le mode de constitution de l'échantillon est essentiel, de même que la population-cible et le taux de participation. S'agit-il d'un auto-questionnaire ou a-t-on eu recours à des enquêteurs ? La participation était-elle obligatoire ? Le contenu du questionnaire ou du registre et sa structure vont déterminer

les possibilités d'analyse, de même que le nombre d'unités d'observation (ménages, individus) qui y ont participé. La connaissance des techniques statistiques et la maîtrise d'un logiciel permettant de les appliquer ne sont donc qu'une partie de l'éventail des compétences de l'analyste des données.

Ce manuel d'analyse des données s'adresse aux chercheuses et chercheurs en sciences sociales, sociologues, politologues, démographes, historiens, épidémiologistes… ayant bénéficié d'un enseignement de base en statistique descriptive et inférentielle et qui maîtrisent un logiciel d'analyse statistique tel que SPSS, SAS, Stata ou R. Il a pour objectif de les **introduire à l'analyse multivariée** en insistant sur la façon dont chacune des techniques statistiques peut les aider à répondre aux questions de recherche qu'ils se posent, tout en tenant compte **du type de variables** disponibles. L'accent est mis sur les **modalités d'application** de ces techniques, les **résultats et mesures** qu'il convient de retenir, la façon de les **présenter** et comment les **interpréter**. Le **savoir-faire** y est privilégié et le recours à des aspects plus formels de la statistique se limite aux éléments indispensables à la compréhension des techniques sélectionnées. Pour les chercheurs qui souhaitent développer leurs compétences, y compris statistiques, il est fait référence, pour chaque technique, à des manuels plus approfondis, mais néanmoins accessibles à des non-statisticiens.

Ce manuel a été rédigé à l'occasion d'une demande de l'IWEPS[1] de « Mise en place d'outils de modélisation des phénomènes sociaux ». Il est le résultat du travail collectif de chercheurs pratiquant l'analyse des données et d'enseignants qui ont assuré des formations en analyse des données au niveau universitaire en se basant sur leurs propres expériences.

Les techniques multivariées sélectionnées renvoient à trois catégories d'approches :

- Les analyses dimensionnelles : l'**Analyse en composantes principales** (variables quantitatives) et l'**Analyse factorielle des correspondances** et **des correspondances multiples** (variables qualitatives) qui permettent entre autres la construction d'indicateurs ou l'identification de dimensions latentes de l'univers des variables analysées.
- Les analyses de classification : la **Classification hiérarchique de Ward** (variables quantitatives) a été retenue pour sa simplicité. Les analyses de classification ou *cluster analyses* tentent de repérer des regroupements « naturels » d'unités d'observation dans

[1] Institut Wallon de l'Évaluation de la Prospective et de la Statistique, http://www.iweps.be/

l'univers des variables analysées. Elles servent aussi à élaborer des typologies.

- Les analyses de dépendance : la **Régression linéaire multiple** (variables quantitatives et qualitatives) et la **Régression logistique** (variables qualitatives et quantitatives). Ici la variation d'une variable (la variable dépendante) est supposée dépendre de la variation d'une ou de plusieurs autres variables (la ou les variables indépendantes) : les régressions sont utilisées pour prédire la valeur d'une variable dépendante, identifier ses déterminants ou même – sous certaines conditions – la ou les causes de sa variation.

Un préalable indispensable au choix et à l'application de ces méthodes est **l'analyse exploratoire** des données. Cette phase de la recherche met en œuvre toute une palette d'outils : l'**analyse univariée** des variables susceptibles d'être analysées par la suite, l'évaluation de la représentativité de l'enquête, de la qualité de l'information recueillie (par l'analyse des non-réponses, de la cohérence des distributions de fréquences…), ou encore, la **description** et la **représentation graphique** des variables. Il s'agit ici, pour l'analyste, de « faire connaissance » avec ses données, mais aussi de rassembler et valider le matériau qui sera analysé par la suite.

L'évaluation de la cohérence interne des données peut se poursuivre par une première série d'**analyses bivariées**, afin d'explorer les relations simples qui s'établiraient entre les variables. C'est à ce stade de l'analyse que le recours aux **tests statistiques** devient intéressant : ceux-ci servent à repérer ou tester les relations qui s'établissent entre les variables, de même qu'ils permettent – de façon complémentaire – de calculer la **marge d'erreur** qui sous-tend la décision du chercheur à conclure à l'existence – ou non – de la vraisemblance d'une relation entre deux variables. Les tests permettent aussi de décider dans quelle mesure les relations observées à partir d'un échantillon peuvent être généralisées à la population de référence dont est issue la population effectivement enquêtée. C'est pour ces mêmes raisons que les tests statistiques seront aussi utilisés lors du passage à l'analyse **multivariée**, définie ici comme l'analyse simultanée de trois variables au moins.

Il est évident, au vu de l'abondance des manuels statistiques existants – souvent très bien conçus – que pour l'apprentissage de la technique en tant que telle, un manuel supplémentaire ne se justifie pas vraiment. Mais en ce qui concerne la **transmission** d'**expériences de recherche** et de **savoir-faire**, avec tout ce que cela comporte de stratégies de recherche à envisager, de pièges à éviter, de précautions à prendre, les titres sont plus rares. C'est dans cette perspective de partage d'expériences que s'inscrit ce manuel. C'est pourquoi la présentation de

chaque technique d'analyse multivariée est précédée du récit d'une expérience de recherche personnelle, afin d'illustrer et de commenter les raisons qui ont amené le chercheur ou la chercheuse à développer sa propre stratégie de recherche, le contexte de recherche plus global dans lequel s'inscrit cette application particulière, comment il-elle a décidé de présenter les résultats, ainsi que leur interprétation. Suit alors une description plus théorique de la technique en sélectionnant les éléments indispensables à sa compréhension et son application. Enfin, pour « aller plus loin » et inviter le lecteur à approfondir ses compétences, une sélection de références accessibles à des non-spécialistes figure à l'issue de chaque chapitre.

Pour accompagner le chercheur ou l'étudiant dans sa recherche personnelle, il nous a semblé utile de compléter ce manuel par une application des techniques qui y sont exposées. Pour cette première série d'applications, le logiciel SPSS version 10 a été utilisé. Ce logiciel a été privilégié, parce qu'il est souvent préféré par les personnes souhaitant s'initier à la pratique de l'analyse de données. Les exemples sont à chaque fois assortis de la syntaxe[2] utilisée pour les produire. Pour l'élaboration de ces exemples d'application, l'IWEPS nous a autorisés à utiliser une partie de la base de données issue de l'enquête *Identités et capital social en Wallonie*, et nous les en remercions. La « *Pratique de l'analyse des données* »[3] est disponible en ligne à l'adresse: www.uclouvain.be/451259.html

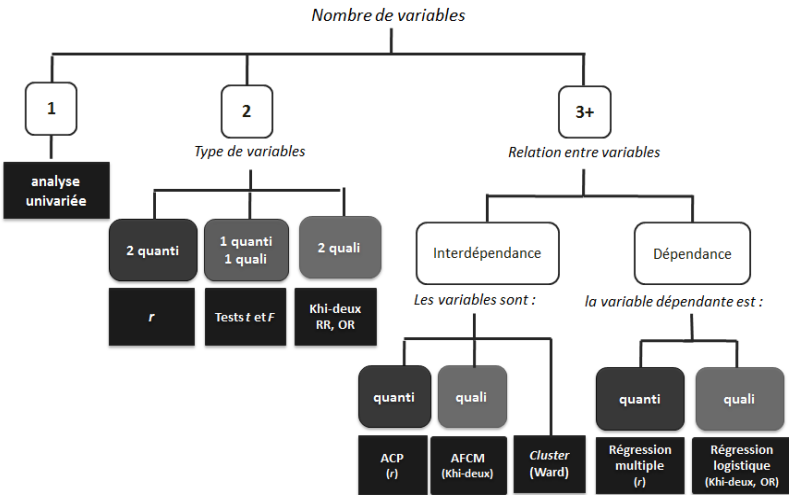
La structure globale du manuel et de son complément « *Pratique de l'analyse des données* », s'articule en deux dimensions : la première est le nombre de variables traitées simultanément (de une à deux, puis trois variables ou plus) et la seconde, le type de variables analysées (**quantitatives**, **qualitatives**) qui conditionne bien évidemment l'éventail des techniques et des tests applicables.

Il est rédigé par Pierre Baudewyns (politologue), Amandine J. Masuy (sociologue), Lorise Moreau (démographe), Ester Rizzi (démographe), Bruno Schoumaker (démographe) et coordonné par Godelieve Masuy-Stroobant (démographe) et Rafael Costa (démographe).

---

2 Les chercheurs n'ayant que peu ou pas d'expérience de SPSS trouveront dans le manuel de Paul Kinnear et Colin Gray (2005). *SPSS facile appliqué à la psychologie et aux sciences sociales*, Bruxelles, éditions de Boeck, un guide leur permettant de s'initier à la manipulation de ce logiciel.

3 Costa R. et Masuy-Stroobant G. (2013). *Pratique de l'analyse des données. SPSS appliqué à l'enquête « Identités et capital social en Wallonie »*, Centre de recherche en démographie et sociétés, UCL.



Il a bénéficié de la relecture attentive et critique de deux collaborateurs : Pierre Baudewyns et Bruno Schoumaker, ainsi que de celle de Philippe Bocquier (UCL), Rébécca Cardelli (IWEPS) et Bernard Masuy (UCL). Ils en non seulement amélioré le contenu et la lisibilité, mais ont aussi « testé » la clarté et la cohérence de l'exposé des méthodes qui y sont présentées.

Enfin, plusieurs générations d'étudiants en ont testé l'approche pédagogique et nous espérons que cet enseignement leur est utile dans leur vie professionnelle.

Louvain-la-Neuve, septembre 2013