Avant-propos

Qu'est-ce que l'analyse des données ?

Dans son acception classique en France, celle utilisée dans ce livre, la terminologie « analyse des données » regroupe un ensemble de méthodes statistiques dont les deux principales caractéristiques sont d'être multidimensionnelles et descriptives.
Le terme multidimensionnel recouvre lui-même deux aspects. Tout d'abord il implique que l'on dispose de la valeur de plusieurs variables pour chaque individu statistique (dans cet avant-propos on se restreint aux données les plus courantes, celles dans lesquelles un ensemble d'individus est décrit par un ensemble de variables). Mais, au-delà de la disponibilité de nombreuses variables pour chaque individu statistique, c'est la volonté de les étudier simultanément qui caractérise une approche multidimensionnelle. Ainsi, aura-t-on recours à l'analyse des données à chaque fois que la notion de profil est pertinente pour considérer un individu, par exemple le profil de réponses d'enquêtés, le profil biométrique de plantes, le profil financier d'entreprises, etc.
D'un point de vue dual, s'il est intéressant de considérer globalement les valeurs des individus pour un ensemble de variables, c'est parce que ces variables sont liées entre elles. Remarquons que l'étude successive de toutes les liaisons entre les variables prises deux à deux ne constitue pas une approche multidimensionnelle au sens strict. Une telle approche implique la prise en compte simultanée de l'ensemble des liaisons entre les variables prises deux à deux. C'est bien ce qui est fait, par exemple dans la mise en évidence de variables synthétiques : une telle variable en représente plusieurs autres, ce qui implique qu'elle soit liée à chacune d'entre elles, ce qui n'est possible que si ces dernières sont elles-mêmes liées entre elles deux à deux. La notion de variable synthétique est donc bien intrinsèquement multidimensionnelle ; elle est un outil puissant de description d'un tableau individus $\times$ variables ; à ces deux titres, elle est un concept clé de l'analyse des données telle que nous l'entendons dans ce livre, à savoir un ensemble de méthodes multidimensionnelles et descriptives.
Cette dernière phrase mérite un commentaire. Le vocable « analyse des données » possède au moins deux sens. Celui précisé ci-dessus et celui, plus large, d'investigation statistique. Ce second sens est un point de vue d'utilisateur ; il est défini par un objectif (analyser des données) et ne stipule rien quant aux méthodes statistiques

mises en œuvre. C'est ce que recouvre le terme anglo-saxon « data analysis ». Le terme « analyse des données », au sens d'un ensemble de méthodes descriptives multidimensionnelles, est plus un point de vue de statisticien. Il a été introduit en France dans les années 60 par Jean-Paul Benzécri et l'adoption de ce terme est sans doute liée au fait que ces méthodes multidimensionnelles sont au cœur de bien des « data analyses ».

A qui s'adresse ce livre ?

Le contenu de ce livre correspond à l'enseignement d'analyse des données proposé à l'ensemble des étudiants d'Agrocampus, *i.e.* toutes filières confondues. Il a été conçu pour des étudiants qui ne se destinent pas aux métiers de la statistique mais qui auront à traiter des données dans le cadre de leurs stages d'abord de leurs emplois ensuite.
Le livre s'adresse donc aux praticiens confrontés à l'analyse statistique de données. Dans cette perspective il est orienté vers les applications ; le formalisme, mathématique a été réduit autant qu'il est possible sachant que les principes généraux qui régissent les méthodes doivent être bien maîtrisés si l'on veut comprendre la signification et la portée des résultats que l'on obtient. Concrètement, le niveau d'une licence scientifique est tout à fait suffisant pour s'approprier tous les concepts introduits.
Sur le plan informatique, une initiation au langage R est suffisante, au moins pour commencer. Il ne s'agit pas à proprement parler d'une compétence en informatique mais plutôt, pour reprendre le mot de R. Delécolle, d'une aptitude à la bureautique scientifique.

Contenu et esprit du livre.

Ce livre est focalisé sur les quatre méthodes fondamentales de l'analyse des données, celles qui ont le plus vaste potentiel d'application : analyse en composantes principales, analyse factorielle des correspondances, analyse des correspondances multiples et classification ascendante hiérarchique. La plus grande place accordée aux méthodes factorielles tient d'une part aux concepts plus nombreux et plus complexes nécessaires à leur bonne utilisation et d'autre part au fait que c'est à travers elles que sont abordées les spécificités des différents types de données.
Pour chaque méthode, la démarche adoptée est la même. Un exemple permet d'introduire la problématique et concrétise presque pas à pas les éléments théoriques. Cet exposé est suivi de plusieurs exemples traités de façon détaillée pour illustrer l'apport de la méthode dans les applications. Tout le long du texte, chaque résultat est accompagné de la commande R qui permet de l'obtenir. Toutes ces commandes sont accessibles à partir de FactoMineR, package R développé par les auteurs. Ainsi, avec cet ouvrage, le lecteur dispose d'un équipement complet (bases théoriques, exemples, logiciels) pour analyser des données multidimensionnelles.

Au terme de ce travail, il nous est agréable de remercier Pierre Cazes et Maurice Roux, membres fondateurs de l'analyse des données en France, pour la relecture attentive du manuscrit.