

# Introduction

---

L'analyse des données est une méthode de statistique descriptive ou exploratoire de données multidimensionnelles (de plus de deux variables) basée sur des calculs relativement compliqués en algèbre linéaire. Les livres qui en traitent, sont souvent très théoriques et analysent à l'aide de logiciels spécialisés les données. Ces logiciels sont des boîtes noires : on entre les données à traiter, on sélectionne la méthode à utiliser et on obtient les résultats sans aucun regard sur les calculs intermédiaires. Aucune interactivité n'est donc possible et l'étudiant qui veut se confronter à des traitements numériques, ou celui qui veut s'initier à ces méthodes, doit se procurer ces logiciels, coûteux pour la plupart d'entre eux, et n'aura de plus aucune interface entre la méthode et les résultats.

L'idée de ce livre est donc toute simple : faire un ouvrage qui rappellera la théorie de façon simple et progressive, faire des « calculs à la main » sur des exemples, créer des classeurs Excel qui permettront l'automatisation des calculs et la comparaison aux calculs effectués manuellement. Ces classeurs seront ensuite généralisés assurant l'interactivité et la reproductibilité des calculs sur d'autres exemples.

Comme pour mes autres ouvrages, références 8 et 9 de la bibliographie, des classeurs d'Excel, figurant sur le CD-Rom joint, sont liés au livre qui permettent de multiples applications sur le modèle étudié : ceci est particulièrement utile pour l'étudiant, le chercheur ou le professeur (ce dernier est souvent confronté pour l'établissement des exemples de son cours à des données qu'il doit traiter à la main et ensuite sur un logiciel, notre méthode lui permet de faire des simulations et lui épargner ainsi un temps considérable...) et ceci dans beaucoup de domaines : des sciences expérimentales aux sciences humaines.

Les classeurs sont protégés, seules les cellules apparaissant en jaune clair ou en orangé à l'écran sont déprotégées (il s'agit de données, de critères ou de paramètres à entrer). En effectuant les modifications sur ces cellules, l'ensemble de la feuille est recalculée et les graphes mis à jour.

Cette interactivité est essentielle, elle permet par exemple de voir :

- les positions des projections sur les axes factoriels et les comparer aux coordonnées calculées : ce qui permet de bien comprendre le changement de repère et les calculs effectués dans la base des vecteurs propres ;
- à partir de quelle origine sont calculées les coordonnées ;
- les transformations des coordonnées selon le choix de la matrice de travail ;
- les proximités entre individus ou variables en effectuant des simulations sur les données initiales ;
- et bien d'autres choses encore qui apparaîtront dans l'ouvrage.

Je me limite aux trois analyses usuelles, l'analyse en composantes principales, l'analyse factorielle des correspondances qui découlent de l'analyse générale. Des classeurs tout prêts permettront la compréhension de ces analyses et leur interprétation.

Cet ouvrage à un but essentiellement didactique et permettra la résolution et l'interprétation de la régression linéaire et orthogonale (chapitre 1) pour 100 individus au plus, de l'analyse générale (chapitre 2), de l'analyse en composantes principales (ACP, chapitre 3) et de l'analyse factorielle des correspondances (AFC, chapitre 4) pour des matrices à diagonaliser de tailles au plus égales à 10, pour au plus 100 individus ou lignes.

L'algèbre linéaire est l'outil de base dans cette théorie. Nous décrirons, dans l'annexe 1, les outils de calcul en algèbre linéaire, disponibles dans Excel et utilisés dans cet ouvrage, avec des exemples d'application. Un classeur interactif (PEDAGO) permettra au lecteur d'effectuer ces calculs et de les confronter à ses propres calculs "à la main".

Deux outils essentiels pour cette étude manquent cruellement au tableur Excel :

- dans la version Excel 2000 (voir Annexe 4 pour la version 97), qui est celle requise pour cet ouvrage, on ne peut inscrire l'étiquette d'un point d'un graphe par un libellé qui le nomme : par exemple si le point « var 1 » a pour coordonnées (1,2) on ne saura inscrire le libellé « var 1 » sur ce point du graphe. Ce problème a été résolu par François Sermier (réf. 14) par la création d'une macro instruction utilisant le langage VBA d'Excel ;
- le calcul des valeurs propres et des vecteurs propres de certaines matrices est à la base de la méthode utilisée dans le traitement des données multidimensionnelles. Il n'existe pas dans la version Excel 2000 de fonction qui permet le calcul des valeurs propres d'une matrice. Nous devons à nouveau à François Sermier la mise en forme dans le langage VBA d'Excel d'une routine, utilisant la méthode de Jacobi, qui crée une fonction personnalisée qui calcule les valeurs propres et les vecteurs propres d'une matrice.

Ces deux outils, sans lesquels cet ouvrage n'aurait pas lieu d'être, conservent à nos classeurs l'interactivité qui nous est si précieuse. L'annexe 2 sera consacrée à leur description.

Nous donnons ci-après un organigramme permettant au lecteur de se repérer selon le niveau d'intervention qu'il désire.

Dans les classeurs que nous vous proposons, nous avons créé des macros instructions qui vous permettront de vous déplacer très facilement. Dans le texte nous avons écrit en caractères gras « **Comic Sans MS** » les instructions que vous devez effectuer sur le classeur Excel, il s'agit essentiellement de cliquer sur des boutons de commande ou des icônes. Les procédures à effectuer qui sont propres à Excel, les noms de référence d'une cellule ou d'une plage de cellules, les noms d'individus ou de variables apparaissant sur les graphes seront généralement écrites dans la police « Courier New ».

~~~~~ Pour un lecteur débutant dans cette théorie, il est évidemment souhaitable qu'il fasse une « lecture linéaire » de l'ouvrage, en ayant sous les yeux l'écran du classeur sur lequel il travaille, de simples coups de clic le mettront en relation avec la feuille concernée. ~~~~~ Néanmoins nous avons prévu dans le texte de cet ouvrage des balises, sous la forme de bandeaux (comme celui en marge de ce paragraphe), en marge du texte, qui isoleront soit

les calculs théoriques, qui peuvent être ignorés en première lecture, soit les applications numériques. Le lecteur initié à ces méthodes pourra commencer son étude au paragraphe 3.7 pour l'analyse en composantes principales où de multiples exemples sont traités et au paragraphe 4.7 pour l'analyse factorielle des correspondances. Des tableaux résumés consignent à la fin de chaque chapitre les notations et les principales formules. Des exemples tout prêts, dans divers domaines, sont « chargeables » directement à partir des classeurs ACP et AFC mais le lecteur pourra aussi entrer directement ses propres données qu'il pourra traiter ainsi de façon extrêmement simple.

Je tiens à remercier François Sermier pour ses encouragements à m'inciter à poursuivre dans ma « méthode d'interactivité » et ses aides par les deux macros instructions mises à notre disposition ; Belaïd Ghermani et Boualem Méliani pour leur relecture attentive, critique et perspicace de la version alpha du manuscrit ; Gildas Brossier et Pierre Cazes qui, par leurs critiques positives et leurs soutiens dans ce projet, ont permis la rédaction définitive et la publication aux éditions PUR de cet ouvrage.

Je suis à la disposition des lecteurs pour toute remarque concernant ce travail. On peut m'écrire à l'adresse :

georgin@univ-paris12.fr