Préface au livre : Analyse des données avec R

Au seuil de cet ouvrage sur l'Analyse des Données, il nous paraît utile de faire quelques rappels historiques. Les scientifiques de toutes disciplines ont eu depuis longtemps besoin de traiter de grandes collections de résultats d'expériences ou d'observations. Numériques ou qualitatifs ces résultats devaient être résumés et synthétisés pour être transmis à la communauté scientifique et au public intéressé. Avant l'apparition des ordinateurs, et la création des langages de programmation évolués (en particulier le Fortran, 1954) ces résumés devaient être faits à la main ou avec l'aide de machines à calculer mécaniques. L'analyse des données est née de la convergence des idées du professeur Jean-Paul Benzécri sur le traitement mathématique à appliquer aux tableaux de données et de la mise à disponibilité des premiers ordinateurs dans les écoles et universités.

L'analyse factorielle des correspondances (AFC), inventée pour traiter de vastes données linguistiques, a été le premier maillon d'un grand ensemble de techniques statistiques qui se proposaient d'aborder les tableaux statistiques indépendamment de toute hypothèse mathématique ou probabiliste sur la structure des données. Le premier exposé de la méthode eut lieu en 1963, mais le premier programme de calcul a été mis au point en 1965 par Brigitte Cordier (devenue par la suite Mme Escofier), en Fortran II sur l'ordinateur IBM 1620 de la faculté des Sciences de Rennes. Depuis cette époque la puissance de calcul ainsi que les facultés de stockage des ordinateurs ont fait des progrès fulgurants, mais les logiciels, sans lesquels cette puissance ne peut être exploitée, ont, eux aussi, considérablement évolué. Le langage R est l'une de ces plus emblématiques évolutions.

Logiciel libre, c'est à dire basé sur la coopération de nombreux spécialistes du monde entier, ce langage propose des instructions de très haut niveau orientées vers le calcul mathématique et statistique. Cela permet de réécrire en quelques lignes les programmes qui autrefois ont nécessité des dizaines ou plutôt des centaines d'instructions. Ces logiciels libres et distribués gratuitement (merci Internet !) donnent un nouvel élan à cette tendance générale en informatique : facilité de mise en œuvre, puissance et souplesse des calculs.

L'originalité du présent ouvrage est sa volonté de rendre complètement autonomes

les utilisateurs des méthodes de l'Analyse des données. Cette autonomie est basée sur deux principes liés aux deux disciplines mises en jeu : la statistique multidimensionnelle et l'informatique. En ce qui concerne la statistique les auteurs proposent de nombreux exemples détaillés ; cela permet de montrer quels types de données peuvent être traités par quelle méthode, et ce que l'on peut attendre des résultats. Ceux-ci sont sérieusement interprétés compte tenu des propriétés mathématiques de la méthode appliquée. Ces propriétés mathématiques sont abordées de façon intuitive mais sans négliger d'énoncer les formules les plus importantes. Quant aux aspects informatiques tous les détails sont donnés pour une mise en œuvre complète des méthodes, grâce à la simplicité du langage R déjà mentionnée.
Fruit de la grande expérience des auteurs à Agrocampus Rennes, ce livre, qui est un cours extrêmement pédagogique sur les principales méthodes d'analyse des données, comporte quatre chapitres : Analyse en Composantes Principales (ACP), Analyse Factorielle des Correspondances (AFC), Analyse des Correspondances Multiples (ACM), Méthodes de classification (hiérarchiques et nuées dynamiques). Les exemples traités à la fin de chaque chapitre pour illustrer les méthodes exposées sont issus de données réelles et de domaines très diversifiés : étude des dépenses des ménages français en fonction de caractéristiques comme l'âge, étude sur la température des principales capitales européennes, exemple de données génomiques, répartition des médailles aux jeux olympiques en fonction du pays et de la discipline (ce qui conduit sans hypothèses à des résultats très intéressants), analyse sensorielle sur des vins du Val de Loire, étude des causes de décès des français en fonction de l'âge (découpé en tranches) entre 1979 et 2006 , enquêtes sur la consommation de thé, sur la perception des OGM, analyse textuelle, etc.
En conclusion, ce manuel, qui constitue un guide très clair et très précieux pour analyser des données en liaison avec le logiciel R, s'adresse aux praticiens (étudiants, utilisateurs, etc.) ayant des tableaux de donnée à analyser, et nous lui souhaitons un franc succès.

Pierre CAZES, professeur à l'université Paris IX Dauphine
Maurice ROUX, ancien professeur à l'université Paul Cézanne (Marseille)