Méthode de base de l'analyse des données

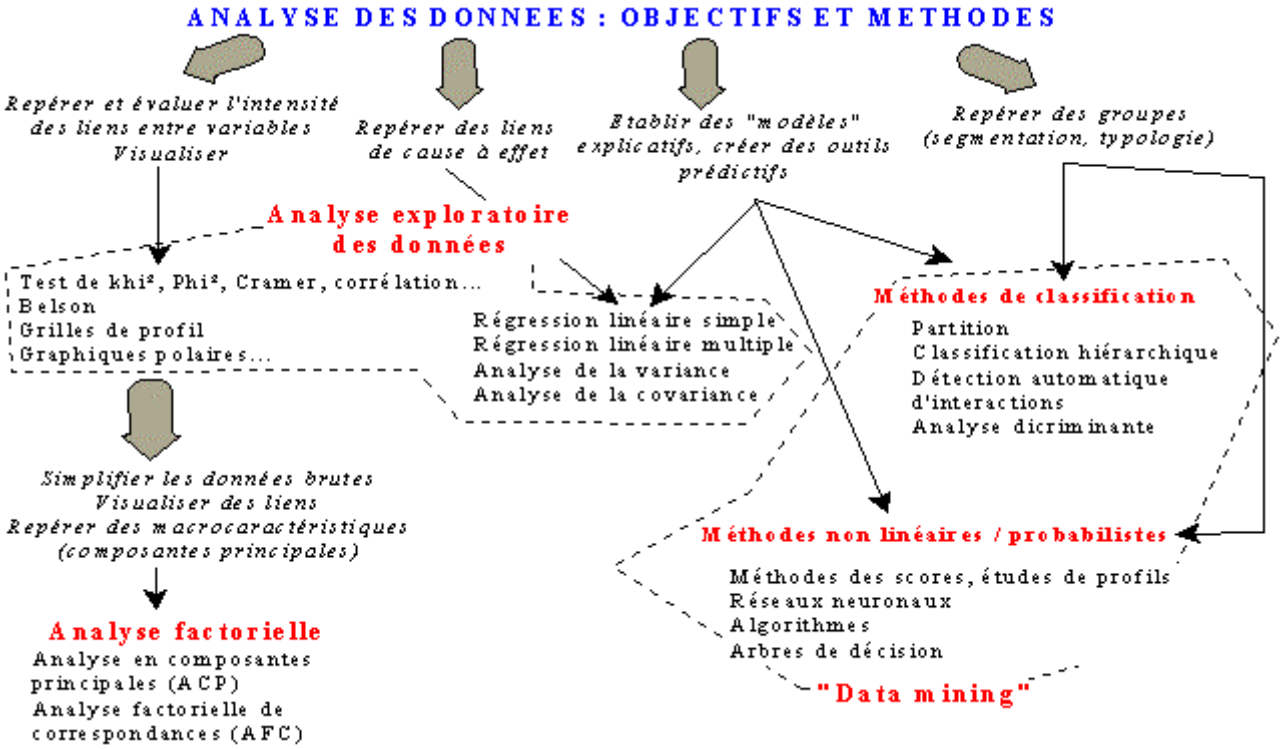
➔ **Auteur :** Michel Jambu Expert en analyse des données et systèmes d'information, professeur à l'université de Paris Dauphine

➔ **Volume** : 418 pages + un CD-Rom d'évaluation du logiciel d'analyse des données StatLab.

➔ **Date de parution** : 1999 ➔ **Editeur** : Eyrolles, collection technique et scientifique des télécommunications

INTERET(S) DE L'OUVRAGE

Selon Michel Jambu, l'analyse des données est l'ensemble des méthodes à partir desquelles on collecte, organise, résume, présente et étudie des données pour en tirer des conclusions et prendre des décisions. Ces méthodes ont beaucoup évolué depuis les années soixante passant du " calcul statistique " à des approches privilégiant la " visualisation interactive des données ". **Au-delà des éléments présentés dans cet ouvrage**, le schéma suivant s'efforce de présenter, sous une forme simplifiée, les principaux domaines de l'analyse des données :



La lecture de cet ouvrage qui nécessite un niveau correct en mathématiques et statistiques comporte trois intérêts :

- les méthodes d'analyse des données sont largement utilisées dans le domaine des études mercatiques en liaison avec l'évolution des technologies de l'informatique et des communications ;
- les méthodes, notamment les plus récentes, sont souvent présentées avec des termes abscons ou des anglicismes (" datamining ", réseaux neuronaux, " scoring ",

" profiling ", etc.) quelquefois pseudo-scientifiques : il est utile de revenir aux fondements mathématiques de ces méthodes afin de comprendre leurs mécanismes de base et leur intérêt en ce qui concerne les applications mercatiques ;
- les processus de calcul sont évidemment automatisés (fonctionnalités présentes dans les logiciels tels que Sphinx, Question, SPSS, etc.) et il n'est donc pas obligatoire de s'attacher au détail des démonstrations mathématiques (abondantes dans l'ouvrage). Toutefois, la compréhension et l'interprétation des résultats fournis par les logiciels (cartes perceptuelles, nombreuses dans l'ouvrage) sont plus faciles pour les personnes qui se sont efforcées de maîtriser les fondements mathématiques de ces méthodes.

CONCEPTS ET IDEES CLES

➜ Composition de l'ouvrage :

- Corps principal : 10 chapitres. Chaque chapitre est suivi d'exercices (sans corrigés).
- index (références bibliographiques) ;
- annexe : CD-Rom StatLab : ensemble de logiciels d'évaluation permettant de pratiquer l'analyse des données sans être statisticien ou informaticien. Le logiciel comporte 30 jeux de données dans des domaines variés (sociologie, économie, sémantique…) à partir desquels les fonctionnalités des logiciels sont mises en œuvre. Trois jeux de données concernent le domaine commercial : performance commerciale d'un réseau de distribution, enquête de satisfaction France Télécom (2 jeux).

➜ Idées principales

La première partie intitulée " analyse élémentaire des données " présente :

- les notions de base de l'analyse des données (chap. 1) :

- objectifs : notamment passage des " données " à " l'information " puis de " l'information " à la " prise de décision " ;
- méthodologie de l'information (étude de l'existant, définition des objectifs, conception des données (notions de variables, entités, périodes), définition des traitements, méthode de collecte, saisie, contrôle, etc. ;
- types de données : tableaux de données (à une ou plusieurs variables, tableaux recodés, tableaux de contingence), variables (chronologiques, logiques ou booléennes, qualitatives à réponses multiples, rang ou classement, préférences, classe) ;

- les domaines de l'élaboration des données (chap. 2) : il s'agit de permettre à l'utilisateur d'avoir des données de bonne qualité, prêtes à l'emploi, définies en fonction des objectifs de l'étude, offrant la possibilité d'effectuer l'analyse et la présentation des résultats en toute confiance. Deux domaines principaux sont abordés :

- la conception des données : recherche de toutes les variables et données informelles ayant, a priori, une relation avec le sujet étudié, définition des populations étudiées (univers et unité statistique de référence, technique d'enquête, etc.), périodes de temps pendant lesquelles les variables doivent être recueillies. Le diagramme d'Ishikawa (ou " causes-effet " ou " arêtes de poisson ") est présenté en détail comme un outil essentiel pour la conception des données ;

- la gestion des données : ensemble des opérations nécessaires du point de vue des utilisateurs des données (accès aux informations, contrôle de la qualité des données, création d'un dictionnaire des données, calculs et gestion des tableaux de données, approches multicritères, etc.).

- les méthodes élémentaires d'analyse des données :

- analyse d'une variable quantitative ou qualitative (chap. 3) : objectif (identifier les éléments essentiels de la répartition des individus associés à une variable), indicateurs numériques de tendance centrale (médiane, moyenne arithmétique, moyenne généralisée…) et de dispersion (amplitude, écart-type, coefficients d'asymétrie…), représentations graphiques (boîtes de dispersion, histogramme, diagramme circulaire…) ;
- analyse de deux variables (chap. 4) : objectif (recherche des relations de " cause à effet " - variable expliquée / variable explicative – ou de dépendance non nécessairement structurée), présentation sous forme graphique et de tableaux de contingence, coefficients de contingence (Pearson, Cramer…), méthode d'étude de l'indépendance (test de χ^2), régression linéaire, moindres carrés, coefficient de corrélation, etc. ;
- analyse conjointe de plusieurs variables (chap. 5) : objectif (étude des relations et des interactions entre plusieurs variables en même temps et non plus seulement deux à deux), approches graphiques (polaire, profil, projection cartographique), analyse d'un tableau de contingence multiple.

La deuxième partie intitulée " L'analyse approfondie des données " présente :

- les fondements de **l'analyse factorielle** (chap. 6) dont l'objectif principal est d'élaborer et de présenter dans un espace euclidien de faible dimension les informations les plus diverses consignées dans des tableaux numériques à double entrée complexes et importants :

- principe du passage d'un ajustement linéaire à une analyse factorielle : résumer un ensemble d'individus en fonction de facteurs communs sous une forme polynomiale,
- technique mathématique (transformation d'un tableau important en une matrice de dimension beaucoup plus petite mais qui conserve la valeur de l'information d'origine : calcul vectoriel).

- **l'analyse en composantes principales** (chap. 7) dont l'objectif est de représenter graphiquement les relations entre des variables quantitatives afin de visualiser les individus en relation avec les variables. L'analyse de l'espace géométrique obtenu (carte de perception ou " mapping ") permet de donner un sens aux axes factoriels (en fonction de la proximité des variables par rapport aux axes), aux regroupements de données (regroupements d'individus et de variable à divers endroits de la carte), singularités (éloignement de critères ou individus par rapport aux regroupements majoritaires), proximités (entre variables, entre groupes d'individus, entre variables et individus). Les méthodes de calcul sont présentées en détail à partir d'exemples concrets ;

- **l'analyse des correspondances binaires** (chap. 8) est appliquée à des tableaux de fréquences issues du croisement de deux variables qualitatives ou assimilables à des tableaux de correspondances binaires ;
- **l'analyse des correspondances multiples** - qualifiée d'outil privilégié de l'analyse des données par Michel Jambu – (chap. 9) est appliquée à des tableaux issus du croisement de plusieurs variables qualitatives ou quantitatives ;
- **les méthodes de classification** (chap. 10) peuvent être appliquées soit à des individus soit à des variables. La *classification des individus* a pour objectif de construire des classes (ou groupes ou encore segments) d'individus en fonction d'un ensemble de variables qualitatives ou quantitatives afin d'obtenir une vision multidimensionnelle de ceux-ci (et non plus seulement en fonction d'un seul critère à la fois). La *classification des variables* a pour objectif de réduire le nombre de variables d'origine en éliminant les redondances et en ne retenant que les plus représentatives (ou discriminantes). Les principales familles de méthodes de classification sont :

- **les méthodes de partition** fondées sur divers algorithmes (ex. : agrégation autour des centres variables) : il s'agit de chercher un critère de ressemblance entre individus et entre classes qui aboutit à la fois aux classes les plus homogènes possibles (compacité maximum de chaque classe ou groupe) et les plus distinctes les unes par rapport aux autres (séparation maximum entre les classes ou les groupes) ;
- **les méthodes de classification hiérarchiques ascendantes** : le principe de ces méthodes est, dans un premier temps, de créer de petites classes ou groupes ne comportant que des individus très semblables. Les étapes suivantes consistent à créer des classes ou des groupes de moins en moins homogènes par regroupements successifs ;
- **les méthodes de segmentation ou méthodes de classification hiérarchiques descendantes** : le principe général est toujours de déterminer les groupes les plus homogènes possibles. A la différence des autres méthodes de classification, la segmentation privilégie une variable à expliquer par rapport à des variables explicatives :

Méthodes de segmentation	**Variable à expliquer**	**Variables explicatives**
Exploration des liaisons et interactions par segmentation d'un ensemble expérimental (ELISEE) : à chaque étape, division de chaque groupe par deux, utilisation du χ^2 pour repérer les variables explicatives les plus discriminantes	Qualitative	Quantitatives ou qualitatives
Détection automatique d'interactions (AID : automatic interaction detection) : même principe qu'ELISEE avec utilisation du η^2 (corrélation) pour repérer les variables explicatives les plus discriminantes	Quantitative	

- **les méthodes de discrimination ou analyse discriminante** : une variable qualitative ayant permis d'établir une classification déterminée, l'objectif est d'expliquer celle-ci par des variables quantitatives explicatives. Il s'agit de chercher les combinaisons linéaires de variables explicatives qui permettent de séparer au mieux les groupes d'individus (établis à partir de la variable qualitative) et de faire une représentation graphique mettant en valeur les séparations. L'analyse discriminante présente l'intérêt majeur de permettre des approches prédictives ou prévisionnelles particulièrement

utiles en mercatique : prévoir le groupe d'appartenance (ou classe d'affectation) d'un nouvel individu en fonction des variables quantitatives qui le caractérisent.

UTILITE OPERATIONNELLE

	Niveau		Commentaires
Pour la pratique pédagogique	Terminale ACC	-	Revoir les outils de base de l'analyse des données (les plus élémentaires faisant partie des référentiels) ; Comprendre les approches plus complexes pour les collègues intéressés par l'utilisation des outils mathématiques en mercatique.
	BTS action commerciale	-	
	BTS force de vente	-	
Pour la préparation à un concours	Capet interne /externe	+	Révision sur les méthodes descriptives
	Agrégation interne / externe	++	Etude de la partie mathématique, statistique du programme. La maîtrise des fondements de l'analyse des données est très utile pour l'étude de cas du concours externe, pour résoudre certains cas donnés lors des oraux des concours interne et externe ainsi que pour répondre aux éventuelles questions des jurys.
Pour la culture générale professionnelle		++	Mathématiques, statistiques et outils informatique / communication sont des éléments majeurs de la mercatique.