ASU BURO

Introduction à l'Analyse des Données

F. CAILLIEZ et J.P. PAGES

SMASH

F. CAILLIEZ

Centre Technique
Forestier Tropical

J.P. PAGES

Commissariat à
l'Energie Atomique

INTRODUCTION A L'ANALYSE DES DONNEES

Sous la direction de

G. MORLAT

Professeur à l'Institut de Statistique
des Universités de Paris

Avec des contributions de

J .C. AMIARD - J. ANDRES - M.F. BARA - J .M. BRAUN- J. BRENOT
P. CAZES- J. DEHEDIN- B. DIOP- Y. ESCOUFIER- C. GUEGUEN-
N. LACOURLY- J.P. MAILLES- B. MARCHADIER- M. PIET'RI- E. ROY-
G. SAPORITA- F. TESTU- R. THOMAS-

- 1976 –

SOCIETE de MATHEMATIQUES APPLIQUEES et de SCIENCES HUMAINES
9 rue Duban 75016 PARIS

Ce livre est patronné par

l'ASSOCIATION des STATISTICIENS UNIVERSITAIRES (A.S.U)

LE BUREAU UNIVERSITAIRE DE RECHERCHE OPERATIONNELLE (B.U.R.O.)

Remerciements

Cette version électronique de l'ouvrage « Introduction à l'analyse de données » (J.P. PAGES, F. CAILLIEZ) dont la première édition date de 1976 a été rendue possible grâce :

A l'initiative du Centre International de Recherche sur l'Environnement et le Développement (CIRED) - *(UMR 8568 du CNRS)* qui a mené le projet et contribué à la relecture du document électronique

Au service de documentation de l'Ecole Nationale des ponts et chaussée qui a réalisé la scannérisation de l'ouvrage

A des associés fondateurs de la SMASH qui ont assuré la mise en forme du document électronique et contribué à sa relecture

Avertissement

Cette première version électronique de l'ouvrage (2016) intègre les besoins de correction signalés mais non intégrés dans la réédition de 1983.

PREFACE

"Pendant longtemps, j'ai cru que j'étais un statisticien qui s'intéressait aux inférences allant du particulier au général. Mais, "attentif au développement de la statistique mathématique, j'ai trouvé des raisons d'étonnement et de doute".

Ainsi s'exprimait, il y a une dizaine d'années, John W. TUKEY, dans les premières phrases d'un article prophétique publié dans les : **Annals of Mathematical Statistics**, *sous un titre percutant:* **The Future of Data Analysis.**

Depuis cette époque, et tout particulièrement au cours des années les plus récentes, on a vu la gent statisticienne se scinder, grosso modo, en deux classes : la première catégorie est celle des statisticiens d'âge moyen qui ont appris et pratiqué la statistique mathématique classique, celle qui prétend formaliser l'induction, à la suite des statisticiens anglo-saxons, notamment, des années 1900 à 1950. La seconde classe est formée de gens en général plus jeunes, qui ont appris sous la même étiquette de "statistique" des techniques bien différentes, s' appuyant sur un outil mathématique purement algébrique, et visant à décrire, réduire, classer, des observations multidimensionnelles; ceux-là n'ont cure de l'induction et sont volontiers portés à proclamer que le statisticien doit mettre en œuvre ses techniques d'analyse sans faire aucune hypothèse sur les phénomènes observés. Ils pratiquent "l'analyse des données".

Au vrai, ce qui vient d'être dit ne décrit que fort incomplètement les grands courants de la pensée statisticienne. Il y a vingt ou trente ans, à l'époque où fut introduite en France, sous l'impulsion de Georges DARMOIS, la statistique mathématique, les statisticiens se divisaient déjà en deux classes, peut-être bien plus opposées encore dans leurs conceptions que celles d'aujourd'hui.

D'aucuns, que nous appellerons "les anciens", ne pouvaient accepter l'idée d'une inférence qui ne fût point basée sur des connaissances a priori exprimables sous forme de distributions de probabilité concernant les paramètres des lois des *phénomènes étudiés, et sur l'emploi de la formule de BAYES; les autres, que l'on considérait à l'époque comme modernes, mais que nous appellerons ici "les classiques", adoptaient les vues des statisticiens anglo- saxons (R.A. FISCHER, J. NEYMAN, A. WALD) et considéraient comme le premier devoir du statisticien d'établir des méthodes d'inférence qui ne laissent point de place aux idées a priori, sauf pour ce qui regarde la forme analytique des lois de probabilité utilisées (mais c'est peut-être là un a priori non négligeable puisque le statisticien classique se demande si cette forme analytique doit quelque chose aux observations qu'il utilise, témoin le problème des degrés de liberté dans le test d'ajustement du khi deux).*

En fait, si les "anciens" réussissaient bien -au prix du choix de probabilités a priori quelque peu subjectif ou arbitraire, en tout cas prêtant à discussion- à formaliser le raisonnement inductif, les "classiques" traitaient un problème un peu différent, ainsi que J. NEYMAN a eu le mérite de le souligner: il s'agit pour eux de "comportement inductif", et l'on comprend bien comment leurs points de vue mènent à la théorie des décisions statistiques. Cela dit, à l'origine de leur débat, se situe une divergence concernant la signification concrète du concept deprobabilité.

Pour les "anciens ", la probabilité s'applique à tout ce qui est incertain, pour les "classiques" -et il est arrivé à quelques- uns d'entre eux de le proclamer de façon fort explicite-il s'agit d'une notion qui ne peut s'appliquer valablement qu'à des phénomènes susceptibles d'observations répetées, dans des conditions identiques (autant qu'il est possible) de sorte que la probabilité n'est autre chose que l'idéalisation de la notion de fréquence de tel évènement lié aux expériences répétées.

Si la race des "anciens" s'était éteinte ou était en voie de s'éteindre, notre tableau des querelles intestines de la gent statisticienne s'en trouverait simplifié. Mais certains développements théoriques sur l'induction et la décision, de même que de nombreuses possibilités d'application, s'en trouveraient appauvris. Car il s'est toujours trouvé, fort heureusement, même à l'époque de la plus grande gloire de la statistique mathématique classique, quelques statisticiens pour douter de la justesse des conceptions majoritaires. Parmi quelques autres -en petit nombre, il faut le reconnaître- Bruno DE FINETTI qui décrivait à peu près comme ceci l'évolution de la statistique entre 1930 et 1950 :

"On s'est aperçu à un certain moment que la

statistique mathématique était bâtie sur du sable (les probabilités a priori, si discutées), alors on a retiré le sable et la statistique a été bâtie sur rien !".

A partir des années cinquante, cette voix a commencé à trouver quelque écho même chez les statisticiens américains, et ce fut le développement de la statistique "néobayésienne" - qu'il faut bien regarder aussi comme une des branches de la statistique moderne-.

Revenons à l'analyse des données.

Contrairement à la statistique néobayésienne, qui vise à consolider la formalisation de l'induction, en consolidant "le sable" sur lequel s'appuyait la théorie statistique, on peut dire qu'on a retiré, en analyse des données, non seulement le sable, mais tout ce qui reposait dessus, c'est-à-dire le modèle probabiliste.

Dès lors, que reste-t-il de la statistique? Pas grand- chose -aux yeux des classiques ou des néobayésiens- sinon des procédés descriptifs. Et c'est bien de cela qu'il s'agit: l'analyse des données n'est pas autre chose que la forme moderne de la statistique descriptive. Mais, parce qu'on dispose de moyens de calcul un peu plus puissants qu'autrefois, il ne s'agit plus, comme autrefois, d'un chapitre préliminaire, mineur et ennuyeux de la statistique. On peut se livrer à une étude purement descriptive, mais puissante cependant, de tableaux de données multidimensionnelles, dans lesquels on ne pouvait étudier naguère que les marges ou les liaisons par couples. Et la description s'avère si efficace que d'aucuns n'hésitent guère à considérer que désormais l'essentiel de la statistique, voire la statistique toute entière, doit se réduire à l'analyse des données, la statistique mathématique classique n'étant qu'un ensemble de jeux de l'esprit, quelque peu arbitraires, destinés à meubler les loisirs des statisticiens qui à l'époque ne disposaient pas d'ordinateurs pour résoudre des problèmes plus réalistes.

C'est peut-être aller un peu loin. Dans l'article déjà cité, TUKEY écrivait:

"Dans la mesure où certaines parties de la statistique mathématique ne réussissent pas à contribuer, ou ne sont pas destinées à contribuer, même par un chemin long et

tortueux, à la pratique de l'analyse des données, elles doivent être regardées comme des éléments de mathématiques pures et critiquées selon les normes correspondantes. Les chapitres de la statistique mathématique doivent chercher leur justification soit dans l'analyse des données, soit dans les mathématiques pures. Un travail qui n'obéit à aucun de ces maîtres -et certains statisticiens prétendent qu'il en est ainsi pour leur propre travail- ne peut manquer d'être provisoire et voué à l'oubli. Et nous devrons être attentifs à ce qu'il n'entraîne pas dans sa disparition des contributions d'intérêt plus durable".

Pouvait-on plus clairement accuser les statisticiens de byzantinisme? Personne ne songerait, certes, que ces accusations puissent s'appliquer aux techniques modernes d'analyse des données. Mais faut-il croire pour autant que la statistique mathématique classique est vouée à l'oubli ?

Il n'y a pas très *longtemps que les statisticiens peuvent utiliser des ordinateurs. Il y a encore moins longtemps qu'on a vu se répandre en France les diverses méthodes d'analyse des données multidimensionnelles, principalement grâce à l'impulsion donnée par les travaux du laboratoire de statistique mathématique de J.P. BENZECRI et les services rendus à des utilisateurs les plus variés (comment s'en étonner? L'analyse des données doit rendre service partout où l'on se soucie d'accumuler des observations, et on doit accumuler des observations partout où l'on veut faire œuvre scientifique). Ces services rendus montrent bien que l'analyse des données constitue aujourd'hui, et de loin, la partie la plus immédiatement rentable de la statistique.*

Tout se passe comme si l'absence de moyens de calcul avait longtemps barré une des voies d'exploration les plus fécondes du réel. Ce verrou vient de sauter d'une façon rapide et on est partout encouragé non seulement à exploiter globalement des données nombreuses, mais aussi à en recueillir de nouvelles, ce qui donne naissance à un torrent d'applications efficaces des techniques telles que l'analyse en composantes principales, l'analyse des correspondances, la classification automatique et d'autres encore. Cela permet selon les cas de découvrir dans les phénomènes étudiés des structures directement visibles sur les résultats de l'analyse, alors qu'elles ne l'étaient pas sur les données originelles, ou de retrouver en les précisant des structures que l'on soupçonnait déjà pour telle ou telle raison.

Mais après? Que va-t-il se passer quand on aura *mouliné à l'ordinateur quasiment toutes les données pertinentes à telle discipline?*

L'application des techniques d'analyse des données met en évidence des traits parfois tellement saillants qu'on est tenté d'en rester là; et, en effet, le client venu avec des masses de chiffres repart content avec quelques graphiques et des idées pour un bout de temps. Mais peut-être un jour voudra-t-il prévoir les phénomènes qu'il étudie (et il lui faudra adopter un modèle probabiliste, suggéré sans doute par les résultats de l'étape descriptive), ou éclairer des décisions (et il se posera des problèmes du type de ceux qu'abordait la statistique mathématique classique).

On peut alors penser que les résultats classiques s'avèreront utiles à l'analyse des données -ou plutôt à des prolongements naturels- même si c'est "par un chemin long et tortueux", selon l'expression de TUKEY. Et l'on verra peut-être se développer davantage de recherches visant à proposer des modèles probabilistes précis dans des espaces multidimensionnels -un peu plus sophistiqués que la classique loi de LAPLACE GAUSS- et à permettre l'application des théories de la décision statistique, sur *des exemples moins simples qu'un échantillon de variables normales réduites indépendantes.*

En attendant, et pour quelques années encore, priorité à l'analyse des données.

G. MORLAT.

-1971 -

PRÉFACE A L'ÉDITION DE 1983

Cette Introduction à l'Analyse des Données date déjà de 1976 ; qu'en dire en 1982 ?

Nous ne remettons rien en cause au niveau du langage adopté, tant nous sommes persuadés que le parti pris pour la géométrie et l'exploitation systématique de la dualité était le bon choix. En s'efforçant par exemple, avec le schéma de dualité, de présenter de façon symétrique les individus et les variables, on insistait sur un fait fondamental que les factorialistes du début de ce siècle ne perdaient jamais de vue; en analyse multidimensionnelle, l'univers des variables joue un rôle au moins aussi important dans la stabilité d'un résultat, que l'univers des individus. Si on remet en cause le modèle probabiliste en analyse de données, on tombe donc sur le paradoxe suivant: d'un côté, l'individu n'est plus une simple unité statistique permettant de définir un continuum : la loi de probabilité; de l'autre, les variables, et ceux qui analysent les résultats d'enquête le savent bien, doivent être considérées comme extraites d'un univers dont il est difficile de distinguer, en général, les unités. Mais nous n'avons pas abordé, sauf en régression, ce grand problème de la stabilité et le langage choisi a permis surtout de bien montrer l'importance du choix des métriques et des critères d'optimalité en analyse factorielle.

Si *nous défendons notre langage, nous sommes les premiers à souligner les insuffisances de l'ouvrage au niveau des méthodes décrites. Il s'agissait, non pas de développer, mais d'introduire de façon rigoureuse et synthétique les principales techniques de régression, d'analyse factorielle et de classification automatique. Malgré les remarques pratiques qui parsèment les différents chapitres, ces introductions se veulent avant tout mathématiques.*

Nous pensons qu'en statistique la révolution informatique est loin encore d'être digérée; le grand ouvrage de l'analyse des données, c'est-à- dire de la statistique moderne, est encore à écrire. Si *nous avons su échapper (pour un temps ?) au discours réducteur, et singulièrement peu adapté à l'approche multidimensionnelle, du statisticien mathématicien « classique», nous sommes encore loin d'une construction logique, structurée... et esthétique. Parmi les travaux qui nous semblent contribuer de façon significative à cette construction, nous plaçons d'abord ceux où l'on s'efforce d'exploiter la diversité à l'intérieur de ce que nous appelons toujours des « échantillons». Cette diversité contient tous les ingrédients qui permettent d'envisager une statistique, qui ne ferait référence au modèle probabiliste a priori, que dans la mesure où il apporterait véritablement quelque chose, compte-tenu de ce que l'on a effectivement à faire.*

F. CAILLIEZ & J.P. PAGES
Novembre 1982

Edition1982
AVIS AUX LECTEURS
Aucun changement n'a été apporté au texte d'origine (édition de 1976)

AVANT PROPOS

Ce livre, qui fait suite à un cours édité en 1971 par le Centre d'Etudes Economiques d'Entreprises (C 3E), ne doit être considéré que comme une introduction à l'analyse des données multidimensionnelles.

En effet, si les bases et les principaux résultats utiles de l'analyse des données sont fournis, les propositions qui sont faites au niveau de la pratique ne prétendent pas à l'universalité. Cette pratique, en pleine évolution, ne saurait être réduite à l'application de quelques techniques standard sur des données toutes préparées; à l'heure actuelle, les tendances sont variées et dépendent du domaine d'application (économie, biologie, psychologie, marketing, ...); les tendances dominantes ne sont pas les mêmes en France et dans les pays anglo-saxons.

Les modifications apportées au cours édité au C 3E n'en ont pas changé l'esprit: nous cherchons à fournir des connaissances solides sur les méthodes de l'analyse des données, aussi bien à des non mathématiciens de formation auxquels les rappels algébriques (chapitres I à V) sont en premier destinés, qu'à des mathématiciens qui trouveront l'occasion de revoir de façon géométrique des connaissances algébriques fondamentales qui apparaîtront ici comme opérationnelles. Aussi ce livre pourra t-il être utile aux praticiens de l'analyse des données recherchant des justifications théoriques et aux étudiants de 2ème ou de 3ème cycle des Universités: son contenu a fait l'objet d'enseignements dans le cadre de la formation permanente notamment au Bureau Universitaire de Recherche Opérationnelle, et dans le cadre universitaire, en particulier à l'Université PARIS VI (cycle supérieur de l'Institut de Statistique des Universités de Paris et 3ème cycle de Statistique Mathématique). Il s'inspire, au niveau de la forme et du langage, à la fois de T.W. ANDERSON et de J.P. BENZECRI.

Un choix a été fait dans la façon de présenter les techniques d'analyse des données : nous utilisons constamment les notions de projecteur, d'application M-symétrique, de codage, et surtout le "schéma de dualité" qui est un instrument de langage efficace permettant de présenter d'un seul jet et sous toutes leurs facettes les techniques relevant de l'algèbre linéaire. P. CAZES et J.P. MAILLES ont largement contribué à ce choix.

Le chapitre 6 est consacré à un rappel de statistique descriptive; on y souligne l'importance de la notion de distance qui assure la cohérence dans le choix de caractéristiques de valeur centrale et de dispersion. Toutes les notions clé permettant d'introduire et de justifier les techniques d'analyse linéaire sont développées au chapitre 7 : notions d'individu et de caractère, codage, schéma de dualité. Au chapitre 8 sont présentées, simultanément grâce à la dualité, l'analyse en composantes principales et l'analyse factorielle sur tableau de distances; on trouve en fin de chapitre quelques théorèmes qui jouent un rôle important en analyse factorielle classique (Spearman, Thurstone, Harman,...) et en régression (théorème de Gauss-Markov). Les exemples développés au chapitre 9 ont pour objectif d'illustrer dans le détail la pratique de l'analyse en composantes principales:

les documents de base que doit fournir tout bon programme sont présentés et commentés. Les autres techniques d'analyse factorielle, qui peuvent d'ailleurs être présentées comme des analyses en composantes principales particulières, conduisent à des documents semblables. Le chapitre 10 qui traite de la régression est particulier dans la mesure où la discussion sur le choix de la métrique, obligeant à s'interroger sur les objectifs d'un "modèle", conduit à osciller entre l'optique "analyse des données" et l'optique "interférentielle"; on trouvera dans ce chapitre une présentation entièrement géométrique du théorème de Gauss-Markov qui joue un rôle fondamental dans l'étude du modèle général linéaire. L'analyse canonique fait l'objet du chapitre 11; elle est assimilée à l'étude des positions relatives de deux sous espaces vectoriels; trois techniques usuelles relèvent de l'analyse canonique: l'analyse canonique classique que l'on utilise parfois pour dresser un bilan des corrélations entre deux paquets de caractères quantitatifs (chapitre 11), l'analyse factorielle discriminante (chapitre 12) qui permet d'étudier les liaisons entre une variable qualitative et un paquet de variables quantitatives et l'analyse factorielle des correspondances, introduite au chapitre 13, pour décrire la liaison entre deux variables qualitatives. A la fin des chapitres 11,12 et 13 on donnera sous forme d'exercices une manière d'étendre l'analyse canonique à l'étude des positions relatives de plus de deux sous espaces vectoriels. Au début du chapitre 14 on dresse un bilan des techniques introduites dans les chapitres précédents; on souligne en particulier que la stratégie à adopter pour l'étude de variables quantitatives est non seulement fonction des objectifs mais aussi du nombre d'observations. Puis on introduit une méthode, basée sur les opérateurs développés par Y. ESCOUFIER, permettant de décrire les proximités entre tableaux individus x caractères ou entre paquets de variables. Le dernier chapitre est consacré à la classification automatique: une grande place est accordée à la discussion sur les indices (proximités entre éléments, entre parties, entre relations binaires); les principales techniques permettant d'obtenir un arbre de classification ou une partition sont passées en revue; à la fin du chapitre on discute des méthodes de segmentation.

Si nous avons exploité en partie les idées développées récemment par Y.ESCOUFIER de l'Université de Montpellier pour l'analyse des tableaux à plusieurs dimensions, nous n'avons pas présenté certains travaux français récents effectués sur l'analyse des données et qui peuvent être considérés comme des aménagements ou des généralisations des techniques décrites; citons parmi ces travaux, outre ceux effectués au sein du Laboratoire de Statistique Mathématique de J.P. BENZECRI, ceux développés à l'Université Paul SABATIER de Toulouse par J.DAUXOIS et A.POUSSE, ceux de G.SAPORTA et de M.MASSON à l'Université de Paris VI, ceux de E.DIDAY et de son équipe à l'IRIA, ceux de J .M. BOUROCHE et de ses collaborateurs, etc ...

F. CAILLIEZ J.P. PAGES

Novembre 1975

TABLE DES MATIERES

CHAPITRE I ENSEMBLES - APPLICATIONS - RELATIONS BINAIRES

CHAPITRE II ESPACES VECTORIELS

CHAPITRE III APPLICATIONS LINEAIRES - MATRICES

CHAPITRE IV DISTANCES ET DISTANCES EUCLIDIENNES

CHAPITRE V PROJECTEURS ASSOCIES A UNE DECOMPOSITION EN SOMME DIRECTE
. Applications linéaires idempotentes
. Applications linéaires M-symétriques

CHAPITRE VI INTRODUCTION A LA DESCRIPTION STATISTIQUE

CHAPITRE VII INTRODUCTION A L'ANALYSE MULTIDIMENSIONNELLE LINEAIRE

CHAPITRE VIII ANALYSE EN COMPOSANTES PRINCIPALESET ET ANALYSE FACTORIELLE SUR TABLEAU DE DISTANCES

CHAPITRE IX DEUX EXEMPLES D'ANALYSE EN COMPOSANTES PRINCIPALES

CHAPITRE X REGRESSION LINEAIRE

CHAPITRE XI ANALYSE CANONIQUE

CHAPITRE XII ANALYSE FACTORIELLE DISCRIMINANTE

CHAPITRE XIII ANALYSE FACTORIELLE DES CORRESPONDANCES

CHAPITRE XIV VERS L'ANALYSE DES TABLEAUX A PLUS DE DEUX DIMENSIONS

CHAPITRE XV CLASSIFICATION AUTOMATIQUE