JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

# JSFS

## Comptes rendus de lecture

*Journal de la société française de statistique*, tome 145, n° 1 (2004), p. 97-105

<http://www.numdam.org/item?id=JSFS_2004__145_1_97_0>

# COMPTES RENDUS DE LECTURE

## Principes d'expérimentation Planification des expériences et analyse de leurs résultats

Pierre Dagnelie
1 vol., 397 pages, les presses agronomiques de Gembloux, Gembloux, 2003,
ISBN 2-87016-069-0
(diffusé en France par Technique et Documentation – Lavoisier)

En 1981, Pierre Dagnelie a publié l'ouvrage qui porte le même titre. En 2001, il avait mis son texte en ligne sur Internet. Cette nouvelle version a été entièrement remaniée et considérablement élargie.

L'ouvrage est constitué de douze chapitres : 1. Le but et les conditions de l'expérience ; 2. Les facteurs et les traitements ou objets ; 3. Les unités expérimentales ; 4. Les observations ; 5. Les expériences complètement aléatoires ; 6. Les expériences en blocs aléatoires complets ; 7. Les expériences en parcelles divisées (*split-plot*) et en bandes croisées (*split-block*) ; 8. Les expériences en carré latin et avec permutations des objets (*cross-over*) ; 9. Les expériences en blocs aléatoires incomplets : expériences non factorielles ; 10. Les expériences en blocs aléatoires incomplets : expériences factorielles ; 11. Les facteurs lieux et temps ; 12. Notions complémentaires. Vingt pages de bibliographie et un index des traductions anglaises complètent l'ouvrage.

Comme toujours dans les ouvrages de Pierre Dagnelie, le texte est d'une remarquable clarté, les exemples sont traités de manière très complète. Les points théoriques délicats, même s'ils ne sont pas développés, sont toujours bien identifiés et le lecteur peut se reporter à l'abondante bibliographie pour compléter ses connaissances.

Le point de vue de l'utilisateur est toujours mis en avant, pour les conditions techniques associées à des suppositions théoriques indispensables pour appliquer les méthodes présentées, mais aussi sur la nature et l'enregistrement des observations, sur les rapports et publications qui doivent accompagner toute étude, sur la sauvegarde des données. Ceux qui ne s'intéressent qu'à la théorie de l'expérimentation et à ses aspects les plus formels perdent souvent de vue ces aspects, pourtant fondamentaux dans toutes les sciences expérimentales. Ils devraient lire cet ouvrage, *a priori* plutôt destiné à des utilisateurs, pour garder un certain contact avec la réalité.

Autre originalité, l'ouvrage est complété par une série de photographies en couleurs accompagnées de légendes détaillées. Elles sont disponibles, avec de nombreuses autres informations sur Internet à l'adresse : www.dagnelie.be.

Tous ces éléments font que l'ouvrage nous paraît être unique dans la littérature non seulement francophone, mais aussi mondiale. D'une lecture aisée, il devrait constituer une référence sur le sujet tant pour les utilisateurs que pour les théoriciens : chacun doit y trouver matière à apprendre, mais aussi à réfléchir.

Richard Tomassone

# Analyse des données

sous la direction de Gérard Govaert
1 vol., 362 pages, Hermès, Lavoisier, Paris, 2003, ISBN 2-7462-0643-9

L'ouvrage présente un état de l'analyse des données par les spécialistes français les plus en pointe dans ce domaine. Sa lecture permettra de voir la multiplicité des outils mathématiques indispensables dans cette branche de la statistique. On peut espérer qu'il montrera que l'analyse des données n'est pas que de l'algèbre linéaire, comme quelques statisticiens semblent encore le croire. L'ouvrage est constitué de onze chapitres, dont voici une brève description.

**1. Analyse en composantes principales (Gilbert Saporta, Ndèye Niang) :** la première des méthodes de référence en analyse des données est traitée de façon simple avec une ouverture intéressante (même si elle n'est pas d'une extrême originalité) sur le contrôle statistique ; la maîtrise des procédés de fabrication tient généralement compte de plusieurs caractéristiques à contrôler de manière simultanée. Dans cette présentation, il me semble que l'exemple, qui est une analyse d'eaux minérales à l'aide de leur composition chimique, soulève toutefois une question d'ordre ***plus épistémologique que technique*** dans la manière de l'aborder. Que fait-on des connaissances *a priori* sur le corpus de données ? Tout buveur (d'eau) sait que la composition des eaux gazeuses diffère de celle des eaux plates par une teneur plus élevée de la majorité des éléments (sauf des nitrates). N'est-il pas intéressant voire nécessaire, pour aller plus au fond de l'analyse, d'en tenir compte dès le début ? En traitant les deux types d'eaux séparément ? En faisant une analyse factorielle discriminante ? Le graphique du premier plan factoriel serait alors naturellement interprétable avec une répartition évidente des types ; on éviterait ainsi d'enfoncer quelques portes ouvertes. Je ne pense pas qu'il y ait de loi générale ; mais sans doute faudrait-il inciter l'utilisateur à se poser la question et à y répondre ! Certes la phrase de J-P. Benzecri « *le modèle doit suivre les données et non l'inverse* » est toujours une sorte de dogme en analyse des données, mais la connaissance *a priori* n'est pas synonyme de modèle !

**2. Analyse factorielle des correspondances (Jérôme Pagès) :** ce chapitre est assez remarquable dans la mesure où l'auteur réussit à faire une excellente présentation technique (le formulaire essentiel de l'analyse factorielle des correspondances ou AFC) alliée à une analyse très approfondie des données, ce qui manque un peu au précédent chapitre. L'auteur montre bien comment les multiples possibilités de présentation de l'AFC constituent une des richesses de la méthode. Mais la façon extrêmement détaillée de traiter un tableau d'évaluation sensorielle montre que l'AFC, en des mains expertes, est un outil (une sorte de microscope intellectuel) qui permet d'aller au plus profond d'une interprétation : tout comprendre dans un tableau

de données sans faire dire aux chiffres plus qu'ils ne peuvent ! Ceci soulève une question sous-jacente à l'enseignement de l'analyse des données, comme de toute méthode statistique : qui a la capacité d'enseigner et de maîtriser simultanément la technique et le domaine auquel on l'applique ? L'auteur nous montre ici qu'une même personne peut y parvenir.

**3. Projections révélatrices exploratoires (Henri Caussinus, Anne Ruiz-Gazen) :** l'analyse en composantes principales ne permet pas toujours de mettre en évidence des aspects cachés de la répartition du nuage des individus. Des techniques, pourtant anciennes, développées par Friedman et Tukey (1974) et Friedman et Stuetzle (1981) existent et de rares présentations ont été faites dans des publications en langue française (Tomassone *et al.*, 1988) [1]. L'intérêt de ce chapitre provient de ce que les auteurs indiquent clairement les bases, souvent ambiguës, de cette technique : « *Qu'est-ce qu'une projection intéressante et comment la chercher ? Comment trouver des individus atypiques*? ». Les auteurs présentent quelques indices particulièrement adaptés à cette recherche et indiquent quelles sont les procédures de calcul mises en œuvre pour y parvenir. Ils mettent bien l'accent sur le fait que ces procédures, maintenant accessibles dans des logiciels standard (Matlab, Splus), doivent permettre une utilisation beaucoup plus facile comme appoint aux procédures plus classiques.

**4. Quantification multidimensionnelle (Gérard d'Aubigny) :** représenter ou modéliser des ressemblances (des similarités) pour en extraire des structures interprétables dans un espace de faible dimension est une autre des tâches importantes en analyse des données. La technique de positionnement multidimensionnel (traduction due à Escoufier, 1975 [2] du terme anglais *Multidimensional Scaling* plus connu sous son sigle MDS) est développée depuis un demi siècle. À l'origine surtout utilisée en psychométrie, avec un petit nombre d'objets comparés, elle s'est étendue à un grand nombre de domaines, en particulier en biologie moléculaire où le nombre d'objets peut dépasser plusieurs centaines, avec la nécessité d'adapter les procédures numériques à cette situation. Ce chapitre très complet, bien que l'auteur en limite les extensions, permet d'avoir une très bonne description de l'ensemble des approches, de sa complémentarité avec les techniques de classification hiérarchique. Il insiste bien sur le choix de la dimension de représentation, sur la validation et l'interprétation. L'abondante bibliographie offre au lecteur la possibilité de compléter ses connaissances.

**5. Analyse des données textuelles (Ludovic Lebart) :** ce chapitre qui ne contient aucun formalisme, est sans doute le plus étranger à la majorité des statisticiens. Aucune méthode nouvelle n'est développée ; mais

1. Friedman J.H., Tukey J. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Trans. Comput.*, vol.C-23, 881-889, 1974.
Friedman J.H., Stuetzle W. Projection pursuit regression. *JASA*, 76, 817 823.
Tomassone R., Danzart M, Daudin J-J., Masson J-P. *Discrimination et Classement.* Masson, Paris, 1988.

2. Escoufier Y. Le positionnement multidimensionnel. *RSA*, XXIII(4), 5-14, 1975.

l'assemblage de connaissances connues dans un schéma nouveau (un peu comme celui des pièces d'un puzzle dont on ne connaît pas la figure à laquelle on doit aboutir) fournit une vision enrichie des réponses à une enquête internationale. Ainsi, nous avons une technique capable de nous aider à structurer les réponses à deux questions : « *Quelle est la chose la plus importante pour vous dans la vie ? Quelles autres choses sont très importantes pour vous ?* ». On peut non seulement disposer d'une technique mais aussi savoir dans quelles limites on peut l'employer en évitant toute interprétation erronée. En particulier la technique du *bootstrap* est méthodiquement et fort astucieusement utilisée pour tracer des ellipses de confiance autour des mots et des catégories (par âge et par niveau d'éducation) de répondants. Mais on peut alors rebondir ! (*La façon d'utiliser le bootstrap peut être employée dans des situations complètement différentes et dans d'autres contextes*).

**6. Modélisation statistique de données fonctionnelles (Philippe Besse, Hervé Cardot) :** l'analyse des données classique se limite à l'étude de $n$ observations décrites par $n$ vecteurs d'un espace de dimension $p$. Dans quelques domaines (météorologie, chimie analytique par exemple) ce cadre formel est beaucoup trop restreint : les observations sont des courbes et plus généralement des fonctions. L'automatisation des procédures de mesure ne peut qu'amplifier le nombre de données de ce type. Il est difficile de transposer brutalement les techniques adaptées au traitement des données vectorielles à celui des données fonctionnelles ; il faut donc les adapter. Mais, pour ce faire, il est indispensable de développer un arsenal mathématique assez important, au moins pour aborder le cadre asymptotique. C'est l'objet de ce chapitre de décrire le cadre fonctionnel et de montrer comment des techniques comme la régression, le modèle linéaire classique et le modèle linéaire généralisé, la prévision s'adaptent à ce cadre. Le lecteur travaillant sur ce type de données pourra donc se forger une stratégie d'analyse et pourra rapidement utiliser les programmes écrits en S mis à sa disposition par les auteurs.

**7. Analyse Discriminante (Gilles Celeux) :** fournir en une trentaine de pages une vision synthétique d'une technique dont les premiers travaux publiés datent des années 1920 est un défi courageux, et ce d'autant plus que la technique originelle a donné naissance à une multitude d'autres techniques, parallèles ou complémentaires. Au-delà de l'exposé de cette famille de techniques, l'auteur met bien l'accent sur les dangers d'une utilisation un peu trop automatisée des logiciels du commerce. Ainsi, l'évaluation des performances permet à l'auteur de dire qu'il préfère la *validation croisée* (que j'aurais appelée *jackknife*) au *bootstrap* pour des raisons de simplicité. Il montre bien les caractéristiques des cinq méthodes classiques les plus utilisées (analyse discriminante linéaire, régression logistique, méthodes des k plus proches voisins, arbres de décision et perceptron multicouche) et leurs avantages comparés. La présentation des développements récents est pleine de sagesse et l'auteur note que l'effet de mode n'est pas toujours équilibré par une analyse critique suffisante. Le défi nous semble atteint, même si le lecteur se doit de consulter les publications les plus récentes pour se faire un jugement personnel plus complet.

**8. Les éléments fondamentaux de la classification (Gildas Brossier) :** le défi de ce chapitre est assez voisin de celui du précédent ; l'auteur l'a engagé de manière différente et volontairement plus descriptive. Disons que c'est un exposé de présentation pour un auditoire peu sensibilisé à un formalisme trop poussé ; le seul formalisme concerne les définitions (d'une partition, d'une hiérarchie, d'un arbre additif, d'une pyramide, etc.). L'intérêt principal est sans doute dans les trois dernières pages dans lesquelles l'auteur fournit ce qui doit être la préoccupation essentielle du lecteur : comment choisir la mesure de distance ? Puis, le résultat obtenu, comment en faire une analyse critique en fonction des méthodes employées pour y parvenir. Enfin, même si la remarque est banale et souvent ancienne (je l'entendais énoncée dans les années 1970 !), comment ne pas approuver l'auteur quand il nous dit qu'il est souvent important de faire simultanément une classification et une analyse factorielle ?

**9. Classification et modèle de mélange (Gérard Govaert) :** alors que les techniques classiques de classification automatique ont toujours eu un relent de « bricolage » auprès d'un grand nombre de statisticiens universitaires, les techniques de mélange présentent un aspect plus académique et pour tout dire plus recommandable. Il faut reconnaître que maximiser une vraisemblance à l'aide de l'algorithme EM a le mérite de poser le problème en termes mathématiques simples. Pourtant, elles n'ont pas connu un grand engouement auprès des utilisateurs. La présentation qu'en fait l'auteur est très convaincante ; l'utilisation d'un modèle probabiliste permet de proposer des solutions à des problèmes délicats comme le choix du modèle et du nombre de classes. Il serait sans doute intéressant de traiter plusieurs corpus de données pour comparer les résultats fournis par les différentes approches. Ainsi, on pourrait apprécier les différences et les similitudes des approches. Bien souvent, sans doute, les résultats seraient voisins ; mais quand ils divergeraient on mettrait en évidence les particularités qui permettraient de mieux définir des stratégies devant des situations nouvelles.

**10. Classification automatique de données spatiales (Christophe Ambroise, Mo Dang) :** quand les observations proviennent d'une image ou d'une carte une technique de classification peut (et souvent doit) tenir compte de cette nouvelle information ; ainsi des observations qui se ressemblent peuvent appartenir à des classes différentes dès lors qu'elles sont éloignées dans l'espace. On voit immédiatement qu'il existe un compromis entre une certaine régularité spatiale et une cohésion des classes. Les auteurs présentent les champs de Markov comme modèles de cette approche, modèles les plus adaptés au traitement des données spatiales. Ils ont fait des choix dans leur présentation, l'exposé est particulièrement clair malgré sa complexité technique. On voit bien, en particulier, l'importance relative de l'homogénéité des classes et de la régularité spatiale lorsqu'ils développent la maximisation de la vraisemblance classifiante dans le cas d'un champ de Markov caché. Les exemples sont particulièrement illustratifs.

**11. Data mining et analyse des données (Georges Hébrail, Yves Lechevallier) :** il était indispensable de terminer cet ouvrage par une mise au point du terme *data mining* ou fouille des données qui connaît actuellement une grande diffusion, avec une connotation commerciale pas toujours scientifiquement très convaincante. Les auteurs disent bien que la fouille des données a pour objectif de valoriser les informations des données contenues dans les systèmes d'information des entreprises. La forme du résultat est donc un élément important ; les méthodes et les résultats doivent être utiles et interprétables mais aussi conserver de bonnes propriétés et correspondre à une certaine vision de la réalité. En fait le problème fondamental est que l'utilisateur des résultats n'a en général aucune compétence statistique et qu'il doit appuyer ses décisions sur les résultats fournis par une personne qui a peu, voire pas, de compétence dans la gestion de l'entreprise ! Néanmoins, ceci n'empêche pas les statisticiens de développer des outils souvent complexes pour les fournir. Le « concept » de data mining a au moins le mérite de poser des questions sur l'emploi de méthodes scientifiques par des foules peu rompues à une approche scientifique !

Je pense que ce bilan était utile pour montrer la vitalité de la statistique, donc de l'analyse des données. L'ouvrage sera, à mon sens, fort utile pour tous ceux qui, statisticiens spécialistes d'un domaine, ont besoin d'une mise à jour dans les domaines connexes. Sera-t-il suffisant pour des débutants ? Sans doute pas s'ils se limitent à sa seule lecture. Mais s'ils utilisent l'abondante bibliographie des différents chapitres pour approfondir l'un ou l'autre des domaines, alors oui. Quant aux « simples » utilisateurs qui sauront lire avec attention cet ouvrage (surtout qui devraient le lire si certains formalismes ne les effarent pas !), ils pourront utiliser les différents logiciels avec plus de précaution, car nombre des présentations les mettent en garde sur un emploi aveugle de ces outils.

On peut aussi penser que la diffusion de la connaissance des méthodes statistiques auprès d'un public très large, qui ne s'améliorera pas avec ce seul ouvrage, devrait être un projet majeur pour tous ceux qui sont des professionnels de la statistique, en particulier les coauteurs de cet ouvrage. On pourrait imaginer, pour ce faire, un complément à l'ouvrage, avec une vision transversale : devant un problème donné (une situation réelle) comment réagiraient les différents auteurs et quelles seraient leur offre de traitement et d'analyse ? Certes, les onze domaines couverts ne se prêteraient pas tous à ce jeu, mais ne pourrait-on le tenter ?

Richard Tomassone

# La Sémiométrie

Ludovic LEBART, Marie PIRON, Jean-François STEINER
1 vol., 228 pages, Dunod, 2003, ISBN 2 10 008105 5

Si, en psychologie sociale, la sémiologie est l'étude des signes comme formes d'expression des sentiments et émotions, la sémiométrie sera naturellement l'étude quantitative de ces signes. L'ouvrage propose un type particulier de quantification, auquel se réfère le néologisme du titre, et en montre l'efficacité. L'outil de mesure est d'abord une liste de mots auxquels les sujets sont invités à attribuer une note selon leur caractère plus ou moins agréable ou désagréable. On obtient ainsi un tableau individus × mots dont l'analyse requiert la méthodologie statistique. Et c'est tout un arsenal de techniques statistiques qui est utilisé pour calibrer l'outil, vérifier sa robustesse, l'utiliser à fin de comparaisons diverses, etc.

Disons-le d'emblée, j'ai pris un grand plaisir à la lecture de ce livre. En un sens, c'est une étude de cas exemplaire où problème substantiel et techniques statistiques sont « tricotés » avec soin, les secondes toujours au service du premier, mais un service extrêmement actif, en permanence susceptible de relancer les questions de fond. Si l'analyse en composantes principales (ACP) est la technique de base, l'étude nécessite de nombreuses autres méthodes ; l'ACP descriptive est d'abord accompagnée de raffinements variés, particulièrement aux niveaux stabilité et validation (ellipsoïdes de confiance, *bootstrap*), mais on voit aussi à l'œuvre bien d'autres techniques, par exemple celle des cartes auto-organisées de Kohonen (voir à ce sujet l'article de Cottrell *et coll.* dans le numéro 144-4, quatrième trimestre 2003, de ce Journal).

Le premier chapitre présente l'outil proposé (choix des mots, techniques d'analyse, ...) et les structures qu'il met en évidence. Le chapitre suivant est consacré aux questions concernant la stabilité des structures ainsi dégagées, stabilité interne de l'outil et stabilité externe dans l'espace (divers pays) et le temps, entre autres. Puisque la méthode est basée sur certaines proximités entre mots, il est important d'analyser si l'outil ainsi construit sort vraiment des structures psychosociologiques ou si celles-ci sont plutôt sémantiques : c'est la question traitée au chapitre trois, tandis que le chapitre quatre discute plus en détail le problème du choix des mots en comparant choix raisonné (l'option choisie) et spontané (ouvert). Le chapitre cinq est une réflexion détaillée et très argumentée sur les effets de taille, notation ou participation, qui sont ici (comme très souvent) caractéristiques du premier axe de l'ACP. Le chapitre six s'éloigne de l'étude proprement statistique pour la compléter de façon fort utile et attrayante par des essais d'interprétation à des niveaux plus littéraires, psychologiques, historiques. Enfin, le chapitre sept donne des ouvertures sur les possibilités d'application qui sont extrêmement variées : par exemple, le lecteur qui se situe à l'une des extrémités du quatrième axe sémiométrique (décryptage dans le chapitre deux !) y trouvera quelques

indications sur ce que peut représenter Dieu pour un homme ou pour une femme, tandis que celui qui se situe de l'autre côté de l'axe en question y comprendra l'intérêt de la méthode dans les études de consommation, donc pour le marketing. Les méthodes statistiques utilisées sont brièvement mais très clairement présentées en annexe et un glossaire complète utilement l'ouvrage.

Je recommande vivement la lecture de cet ouvrage à tous les statisticiens soucieux de comprendre à quoi et comment leur science peut être utilisée, comme à tout honnête homme souhaitant saisir l'utilité de l'analyse statistique sur un exemple élaboré d'intérêt général.

Henri Caussinus