JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JSFS

Comptes rendus de lecture

Journal de la société française de statistique, tome 143, n° 3-4 (2002), p. 155-162

<http://www.numdam.org/item?id=JSFS_2002__143_3-4_155_0>

COMPTES RENDUS DE LECTURE

Bases de données et statistique

Editeurs Scientifiques :
Annie Morin, Patrick Bosc, Georges Hébrail et Ludovic Lebart
Dunod 2002, 368 pages, ISBN 2 10 005350 7

L'objectif du livre est de faire le point sur les nouvelles préoccupations communes aux statisticiens et aux spécialistes de bases de données avec l'arrivée du *data mining* et l'apparition de très grands ensembles de données. Cet ouvrage collectif est issu des contributions à une Ecole Modulad organisée en 1999, mais il représente bien plus que les « actes » de cette école car les contributions ont été entièrement ré-écrites et articulées entre elles ; le résultat constitue un véritable ouvrage de référence sur ce sujet.

L'ouvrage est divisé en cinq parties présentant les liens actuels entre la statistique et les bases de données. La première partie, intitulée Bases et entrepôts de données, fait le point sur les deux domaines connexes mais distincts que sont les systèmes de gestion de bases de données (SGBD) et les systèmes de recherche d'information (SRI). Les premiers sont utilisés pour des informations fortement structurées tandis que les seconds sont plutôt utilisés en recherche documentaire. Cette partie présente les principes généraux des bases de données relationnelles ainsi que les caractéristiques du langage SQL (P. Bosc). On y trouve aussi des recommandations pratiques pour améliorer les performances d'un SGBD (P. Fresnais). La question de l'indexation dans le cas d'un SRI y est aussi évoquée (C. Berrut) et, dans le dernier chapitre de cette première partie, M.S. Hacid introduit les notions d'entrepôts de données et de systèmes OLAP.

Le seconde partie concerne l'apprentissage et la recherche des règles, plus exactement l'extraction des connaissances à partir des données (ECD). Les principes présentés relèvent de l'analyse des données et de l'apprentissage symbolique. Le premier chapitre de cette partie est consacré à la présentation des graphes d'induction (D. Zighed). L'introduction d'objets symboliques permet de représenter des structures complexes de données auxquelles on peut étendre les méthodes classiques d'analyse de données (Y. Le Chevalier, E. Diday). Par ailleurs, dans le domaine de l'ECD, il est préférable d'obtenir des règles de décision compréhensibles pour le praticien. Cette simplicité peut se faire au détriment de l'optimalité. Le chapitre suivant (Y. Kodratoff) étudie les critères permettant d'apprécier la validité et l'utilité d'une règle produite. Le dernier chapitre (G. Hébrail) présente l'apport des travaux de R. Agrawal et des règles d'association pour traiter des volumes de données importants.

L'utilisation de grands ensembles de données, qu'il s'agisse du nombre de cas ou du nombre de variables, pose de nouveaux problèmes en statistique, non seulement des problèmes de stockage, mais aussi des problèmes statistiques

et numériques liés à l'échelle des données. Cela fait l'objet de la troisième partie de l'ouvrage intitulée « quelques aspects de l'analyse statistique des données ». L. Lebart y expose les nouveaux problèmes d'échelle ainsi que le cas des données manquantes et des fusions de fichiers. M. Quafafou présente une synthèse des différentes méthodes de sélection de variables utilisées en statistique et en intelligence artificielle.

La quatrième partie « traitement des données textuelles » se décompose en quatre chapitres. Les deux premiers, plus méthodologiques, sont respectivement consacrés à l'apprentissage sur des données textuelles (P. Gallinari) et à l'analyse et la visualisation des données mixtes : textuelles/numériques (L. Lebart, M. Rajman). Ces deux chapitres, l'un orienté intelligence artificielle et l'autre méthodes statistiques, sont tout à fait complémentaires. Cette partie se termine par deux chapitres d'application l'un aux bases de données médicales (M. Kerbaol, J.-Y. Bansard) et l'autre aux banques de données bibliographiques (N. Pinhas).

La dernière partie « contexte et mise en œuvre du *data mining* » illustre la complémentarité des deux domaines, bases de données et statistique. Les deux premiers chapitres (M. Léonard et F. Fogelman) sont orientés vers la mise en place de la plate-forme informationnelle et l'intégration des techniques de *data mining*. Dans le dernier chapitre, D. Desbois replace les technologies de l'information dans un contexte historique et fait une synthèse des ressources disponibles sur le web en insistant sur la veille technologique.

Cet ouvrage est accessible à tout spécialiste de la statistique ou/et des bases de données intéressé par le développement du *data mining*. C'est un ouvrage de base qui a le mérite de faire le lien entre deux domaines où les chercheurs gagneraient à collaborer.

Jean-Hugues Chauchat,
Université Lumière - Lyon 2

Économétrie des séries temporelles macroéconomiques et financières

Sandrine Lardic et Valérie Mignon
Economica, 2002, 1 Vol., 405 pages, 37 euros, ISBN : 2 71 784454 6

Depuis les années 1980, l'économétrie a connu de nombreux développements : les problèmes de non stationnarité et leurs conséquences, ainsi que la prise en compte de la non linéarité des processus constituent un progrès indéniable dans le domaine de l'économétrie moderne.

Cet ouvrage fait un point indispensable sur ces méthodes modernes de traitement des séries temporelles. Le titre – réducteur – peut laisser croire au lecteur que les méthodes exposées sont spécifiques d'un certain type de problématiques (macroéconomiques et financières), il n'en est rien. Ce livre a vocation de présenter de manière exhaustive l'ensemble des techniques d'analyse des séries temporelles depuis la classique approche de Box-Jenkins jusqu'aux théories les plus récentes telles que la théorie du chaos ou les modèles à mémoire longue de type ARFIMA.

Si le titre de ce livre fait référence à la macroéconomie et à la finance c'est que d'une part, tous les exemples traités (et ils sont nombreux) sont issus soit de la macroéconomie soit de la finance et que d'autre part, un certain nombre de méthodes sont très adaptées à la spécificité des séries financières (volatilité autorégressive, mémoire longue, ...). Cependant, la portée de cet ouvrage dépasse largement ce simple cadre et concerne tous les champs de l'économie et de la gestion traitant de l'analyse des séries temporelles.

La partie I (chapitres 1 à 4) traite des incontournables références aux définitions de base, aux modèles ARMA et à l'introduction des processus multivariés et à la représentation VAR.

Les chapitres 5 et 6 de la partie II abordent l'économétrie des processus non stationnaires. Il s'agit certainement de la meilleure revue (en français) de tous les tests de racine unitaire. Les auteurs ont une volonté d'exhaustivité et de pédagogie tout à fait remarquable. Une application très détaillée des tests est présentée en fin de chapitre, les sorties informatiques du logiciel Eviews sont commentées. Le lecteur met ainsi immédiatement en pratique la théorie des tests de racine unitaire. Le concept de co-intégration et de modèle à correction d'erreurs fait l'objet du chapitre 6, là encore illustré par des exemples issus du logiciel Eviews.

Enfin, il convient d'insister sur la partie III qui constitue une originalité : trois chapitres consacrés aux processus non linéaires. Cette partie fait un point détaillé sur trois problématiques : les modèles à hétéroscédasticité autorégressive de type ARCH et leurs dérivés (GARCH, EGARCH, IGARCH,...), les processus à mémoire longue ARFIMA et enfin l'économétrie du chaos. Les présentations théoriques et l'ensemble des démonstrations sont très soignées ;

cependant, nous pouvons regretter sur les deux derniers chapitres l'absence ou la quasi absence d'application.

Cet ouvrage apporte une contribution originale au thème de l'analyse des séries temporelles. Malgré la difficulté du sujet et les développements théoriques indispensables, le souci pédagogique des auteurs et les nombreux exemples font de ce livre un ouvrage de référence dans ce domaine. Les économistes ou gestionnaires ayant à traiter des séries temporelles trouveront toutes les réponses à leurs questions et les multiples exemples leur permettront de mieux comprendre la portée de ces méthodes modernes. Pour ma part, je n'ai qu'un regret : l'impossibilité qu'a le lecteur de refaire par lui même les exemples présentés, il est dommage que les auteurs n'aient pas mis à disposition les données statistiques servant d'illustration.

Régis Bourbonnais
Université de Paris-Dauphine

Statistique et économétrie Du modèle linéaire... aux modèles non linéaires

Xavier Guyon
Ellipses, 2002, 205 pages, ISBN 2-7298-0842-6

L'ouvrage de Xavier Guyon concerne les modèles paramétriques «de régression», au sens le plus large, dont l'objectif est d'expliquer une variable aléatoire (endogène) Y par des variables (exogènes) x. Une première partie est consacrée au modèle linéaire : le chapitre 1 donne une présentation générale et le chapitre 2 traite du cas particulier de l'analyse de la variance. Il s'agit de présentations très complètes, avec le souci marqué de montrer à la fois les aspects calculatoires (matriciels) et les aspects géométriques. On trouve ensuite un chapitre sur l'asymptotique du modèle linéaire, une question rarement traitée de façon détaillée dans les autres ouvrages de ce type. Le chapitre 4 traite la question des résidus non sphériques (hétéroscédasticité, corrélation) incluant le cas des régressions empilées (*seemingly unrelated regressions*) ; le chapitre 5 discute du choix et de la validation de modèle ; les chapitres 6 et 7 sont ensuite consacrés respectivement aux régresseurs stochastiques et aux équations linéaires simultanées. Les deux chapitres qui suivent présentent le cas des variables explicatives qualitatives, d'abord binaires (régression logistique, modèle probit,...), puis générales (modèles polytomiques), tandis que le chapitre 10 est consacré au modèle log-linéaire et aux tables de contingence (deux facteurs ou plus, y compris une ouverture sur les modèles graphiques). Les modèles de régression non linéaire sont ensuite présentés, incluant une introduction au modèle linéaire généralisé. Le chapitre 12 est enfin consacré aux modèles de durée, y compris les questions de données censurées, tandis que le chapitre 13 traite de la simulation des variables aléatoires, de la méthode de Monte Carlo et du rééchantillonnage (*bootstrap*). Une annexe rappelle les principaux concepts généraux de probabilités et de statistique ; l'ouvrage se termine avec les solutions des nombreux exercices posés au fil des chapitres.

Au plan pratique, les modèles étudiés sont illustrés et motivés par de très nombreux exemples. Au plan mathématique, les divers résultats concernant les techniques d'estimation et de test sont présentés avec rigueur mais avec un effort pédagogique certain (par exemple la comparaison soigneuse des propriétés selon diverses situations). Le chapitre sur le choix et la validation de modèle est peut-être un peu court : les techniques de pénalisation pour le choix de modèle ne sont que succinctement évoquées, comme les méthodes d'exploration graphique ; mais il est clair qu'il fallait se limiter... En fait, l'ouvrage de Xavier Guyon est extrêmement dense et contient une quantité d'information impressionnante pour son volume. Si certains développements

sont surtout justifiés par les applications économétriques, qui fournissent aussi beaucoup d'exemples, il est clair que l'intérêt de l'ouvrage dépasse largement ce domaine. Je suis sûr qu'il sera utile, comme le souhaite l'auteur, aux étudiants de second cycle d'économétrie ou de mathématiques appliquées (surtout cependant les plus motivés), d'écoles d'ingénieur, etc., ainsi qu'à tous ceux qui, munis des compétences de base suffisantes, souhaitent maîtriser les problèmes de modélisation qu'il traite.

Henri Caussinus

Modélisation probabiliste et statistique Résumé de cours et annales corrigées

Bernard Garel
Cépadués-éditions, 2002, 204 pages, ISBN 2-85428-590-5

Le livre de Bernard Garel est une introduction aux notions de base du calcul des probabilités et de la statistique mathématique : espaces probabilisés, variables aléatoires réelles à une ou plusieurs dimensions, moments, fonction caractéristique, convergences, estimation (principes, méthodes usuelles) et tests (principes, tests les plus usuels). Le niveau mathématique requis est de l'ordre de bac+3, le cours s'adressant initialement à des étudiants de première année d'une école d'ingénieurs.

Le sous-titre « résumé de cours et annales corrigées » décrit bien le contenu mais risque d'être quelque peu réducteur. En effet, d'une part le résumé de cours, s'il est d'abord synthétique, est suffisamment consistant pour constituer une assise solide de l'ensemble du domaine couvert, d'autre part les exercices proposés ne sont pas de simples contrôles mais des compléments intéressants explorant divers aspects de la modélisation stochastique avec des ouvertures sur la fiabilité, les séries chronologiques, la gestion des stocks, les modèles de capture-recapture, etc. Les solutions sont soignées et pleines d'enseignements.

L'ouvrage est très pédagogique, les aspects mathématiques sont présentés avec rigueur mais sans excessif formalisme, les applications concrètes et les questions de modélisation constituant clairement l'objectif essentiel. Il sera utile non seulement aux élèves ingénieurs auxquels il est d'abord destiné, mais à tout lecteur désireux d'avoir sous la main une bonne introduction aux probabilités et à la statistique inférentielle, sans compter les enseignants qui y trouveront, entre autres, une bonne source d'exercices.

Henri Caussinus

Statistique

Xavier Milhaud
Collection « de la licence à l'agrégation », Belin, 2002, ISBN : 2-7011-2650-9

Une monographie en français par quelqu'un qui aime écrire les choses de manière précise, voilà une bonne idée. Où trouver en effet, réunies dans un même ouvrage, toutes les bases de la statistique depuis l'estimation, les intervalles de confiance, les tests jusqu'à l'optimalité, l'efficacité et l'asymptotique ? Toutes ces notions que l'on a toujours considéré connaître plus ou moins bien mais que l'on découvre ne pas connaître si bien que cela quand on se demande précisément dans quelle classe tel estimateur est vraiment optimal, que l'on a promis d'expliquer à un collègue ce qu'est le test exact de Fisher pour une table de contingence, ou enfin quand on a un cours à préparer pour la semaine prochaine.

C'est un livre très précieux pour l'enseignement car il couvre toutes les bases de la statistique avec de nombreux exemples. Support de cours naturel pour des étudiants un peu matheux (maîtrise de maths ou préparation à l'agrégation), il constitue un ouvrage de référence qu'un enseignant consultera régulièrement. On y trouve en effet un compromis entre d'une part certains ouvrages qui sont trop axés sur la pratique et qui ne conviennent pas pour accrocher, par exemple, des étudiants de la maîtrise de mathématiques et d'autres ouvrages qui sont trop formalisés, utilisent des notions de trop haut niveau ou sont écrit en anglais.

J'ai particulièrement apprécié le chapitre sur la statistique asymptotique et encore plus particulièrement la démonstration de la validité des test de chi-deux d'une hypothèse composite, tests qui sont soit traités sans démonstration, soit avec des démonstrations fort longues par les autres ouvrages.

Du fait de la richesse des exemples et des exercices, ce livre constitue également un complément d'ouvrages plus axés sur la pratique statistique et sur l'utilisation de logiciels. Il s'adresse alors à des praticiens qui désirent comprendre certaines notions statistiques en profondeur.

En conclusion, un livre qui a peu d'équivalent français et qui sera d'un grande utilité pour tout statisticien qui n'est pas allergique à une écriture un tant soit peu mathématique.

Jean-Marc Azaïs