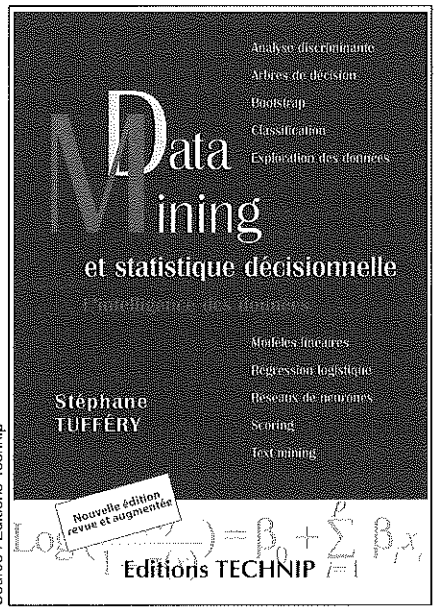
À propos de l'ouvrage « *Data Mining et statistique décisionnelle* », de Stéphane Tufféry

Ouvrage paru aux éditions Technip en 2007

*Jean-Pierre Nakache**

Comme son titre l'indique, cet ouvrage, rigoureux mais agréable à lire, est une synthèse de la connaissance en matière de méthode statistique et de data mining. Il aborde la première avec une approche pratique en utilisant des exemples réels traités à l'aide de logiciels standard (surtout à l'aide du logiciel SAS), en adoptant le point de vue de l'utilisateur mais sans jamais perdre de vue la base théorique. Il aborde le second avec esprit critique et réalisme, sans esquiver les problèmes concrets rencontrés dans le quotidien du statisticien.



Source : Éditions Technip

Le lecteur, s'il est un peu mathématicien, statisticien ou économètre, saura ensuite utiliser un logiciel de statistique ou de data mining pour en tirer des résultats pertinents et utiles pour son entreprise ou son laboratoire. Quant au lecteur non mathématicien, il découvrira, au travers des nombreux exemples cités, un large pan de la société de l'information dans laquelle nous vivons et il pourra, s'il travaille par exemple dans un service de marketing ou de production, s'adresser aux statisticiens pour leur commander en connaissance de cause les études propices à sa propre activité.

Dans la première partie (chapitres 1 à 5), l'auteur introduit le data mining et traite du déroulement d'une étude de data mining, de l'exploration et la préparation des données, de l'utilisation des données commerciales, et fournit un aperçu général des principales méthodes employées.

La deuxième partie (chapitres 6 à 10) est consacrée à l'exposé des méthodes avec, pour chacune d'entre elles, ses principes théoriques, un exemple d'utilisation, les pièges à éviter, son domaine d'application et ses limites.

Dans la troisième partie (chapitres 11 à 15), l'auteur présente des applications diverses (scoring, web mining, text mining), les logiciels de statistique et de data mining, les facteurs de succès d'un projet de data mining et sa mise en œuvre en entreprise.

L'auteur fournit, à la suite de ces trois grandes parties qui composent l'ouvrage : une bibliographie commentée, qui s'ajoute aux renvois dans le corps du texte vers des références spécialisées ; une annexe statistique, précédée de la liste des principales dates de l'histoire de la statistique, qui rappelle les rudiments au néophyte ; et une annexe juridique, qui résume les textes légaux encadrant la pratique de l'analyse statistique des informations nominatives.

Cet ouvrage présente un large éventail de méthodes statistiques descriptives et prédictives : les méthodes classiques de l'analyse des données et de la statistique (classification automatique hiérarchique ou par partitionnement, régression linéaire, régression LOESS, régression PLS, analyse discriminante linéaire, régression logistique, modèles linéaires généralisés...) et les méthodes plus spécifiques au data mining, que sont les arbres de décision, les réseaux de neurones, les séparateurs à vaste marge (« support vector machines »), le bagging, le boosting...

L'auteur fournit des tableaux récapitulatifs très utiles donnant une vue d'ensemble ordonnée du sujet. En amont de la modélisation, l'étape d'exploration et de préparation des données est détaillée avec la mention de tous les tests statistiques permettant de travailler en conciliant rigueur et efficacité. En aval de la modélisation, les méthodes d'amélioration des

* Jean-Pierre Nakache est ingénieur de recherche (CNRS/Inserm U687) et enseignant à l'Institut de statistique de l'Université de Paris (ISUP).

performances (combinaison, stratification et agrégation de modèles) et d'évaluation des performances (courbe ROC, courbe de lift, indice de Gini) sont exposées pour permettre d'aboutir aux meilleurs résultats possibles.

Cette nouvelle édition revue et augmentée suit le plan de l'édition 2005, mais avec plus de 150 pages supplémentaires, les exemples étant nettement plus nombreux. Le chapitre sur la régression logistique est ainsi étoffé, avec un exemple de score entièrement traité sous SAS, mais presque tous les chapitres sont enrichis : au premier chef, ceux qui concernent les tests statistiques, l'analyse factorielle, la régression linéaire, la régression logistique déjà citée, les modèles linéaires généralisés, l'agrégation de modèles et dans une moindre mesure, la classification, les arbres de décision et les « support vector machines ».

Sur tous ces sujets, les aspects statistiques ont été approfondis, tout en étant illustrés par de nombreux exemples numériques, le plus souvent étudiés avec SAS. Certaines fonctionnalités assez récentes de SAS sont utilisées, telles que l'ODS et l'ODS GRAPHICS. L'auteur montre comment l'utilisation de l'ODS et du langage macro de SAS permet d'automatiser les tests à réaliser dans le processus de sélection des variables. Le chapitre sur les logiciels a été actualisé en tenant compte des nouveautés des dernières versions et en présentant en détail les logiciels SAS, SPSS et R, avec un comparatif très précis de SAS et SPSS. Ces deux logiciels sont aussi comparés dans leur mise en œuvre de la régression de Poisson, utile aux actuaires.

Cet ouvrage vise plusieurs publics : tout d'abord, les statisticiens qui travaillent en entreprise (banques, télécommunications, assurance...) et sont confrontés à la construction de modèles prédictifs de scoring. Pour la plupart de ces statisticiens qui sont souvent isolés, sans beaucoup de documentation, l'ouvrage de Stéphane Tufféry leur fournira les bases techniques minimales sous forme compacte.

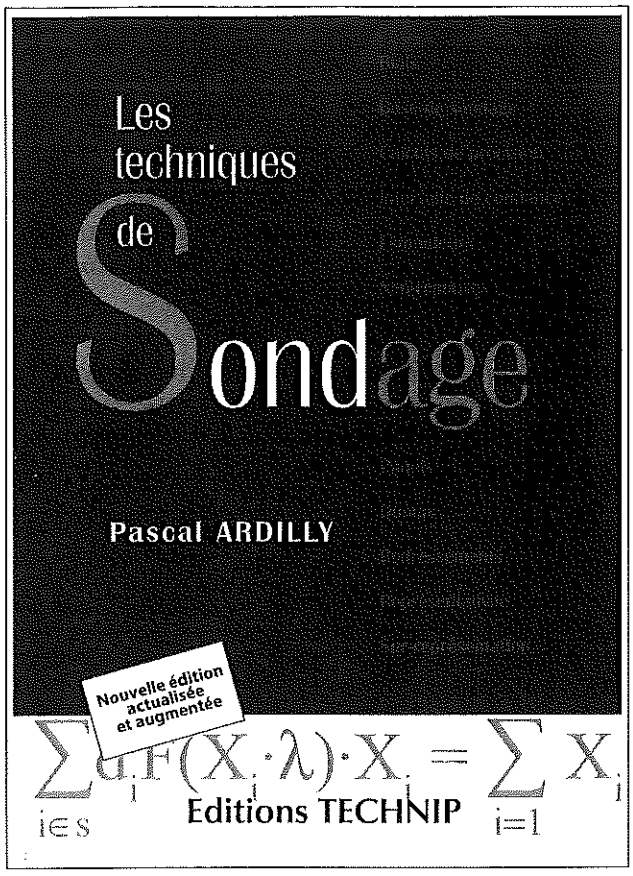
Il s'adresse également aux étudiants et enseignants en statistique appliquée, en ingénierie décisionnelle ou en économétrie, qui y verront une application concrète de leurs cours de statistique. Pour le public universitaire, l'ouvrage aborde d'ailleurs le logiciel R.

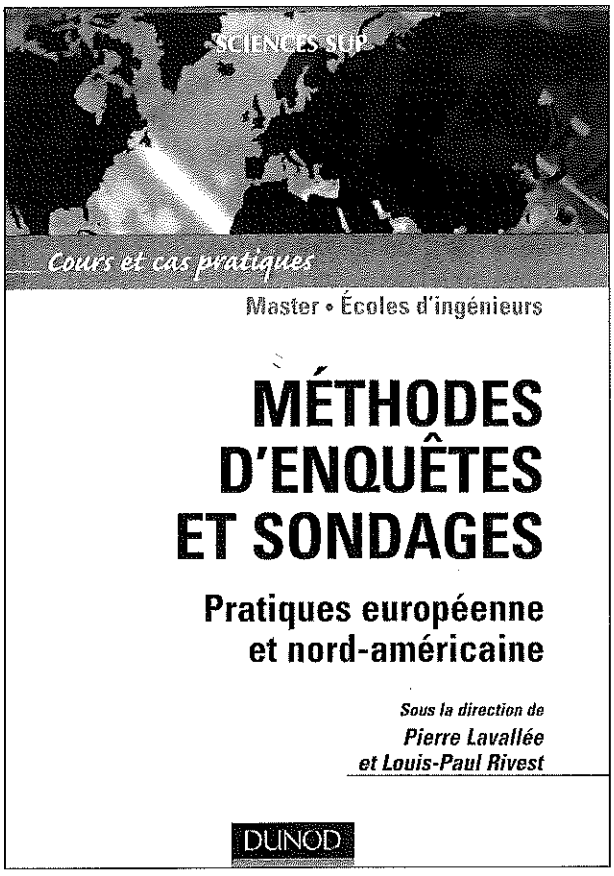
Il intéressera aussi les utilisateurs finaux, en particulier les professionnels du risque et du marketing, qui verront ce que peut leur apporter le data mining et ce qu'ils peuvent demander aux « data miners » et autres statisticiens. Ils seront plus particulièrement concernés par les chapitres 1 à 4, 11 et 12, et trouveront des réponses aux questions qu'ils pourront être amenés à se poser : que faire quand on manque de données ? Qu'est-ce qu'un score générique ? Quelles sont les conditions d'un bon déploiement en entreprise ? Comment évaluer le retour sur investissement ?

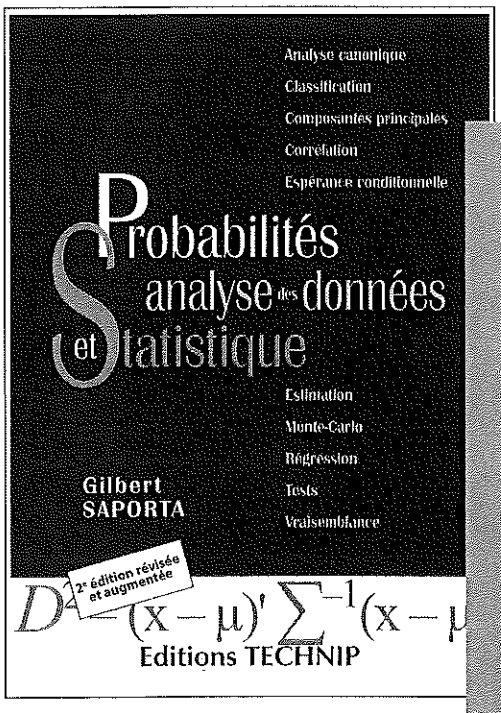
Enfin, cet ouvrage pourra susciter l'intérêt des statisticiens travaillant dans le système statistique public souhaitant avoir une culture générale statistique large, portant sur des outils qu'ils n'utilisent pas nécessairement dans leur environnement professionnel quotidien[1]. ▩

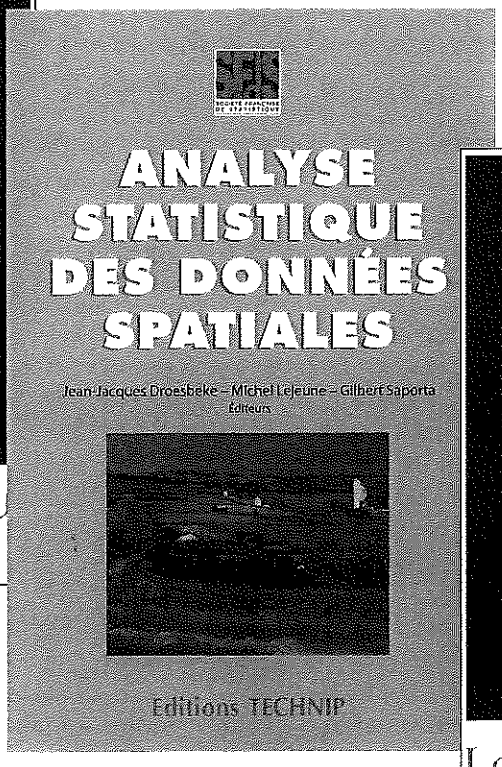
1. À ce sujet, voir aussi les articles parus dans la rubrique « Le métier de statisticien en dehors du système statistique public », *Courrier des statistiques* n° 117-119, année 2006, ainsi que dans le présent numéro [NDLR].

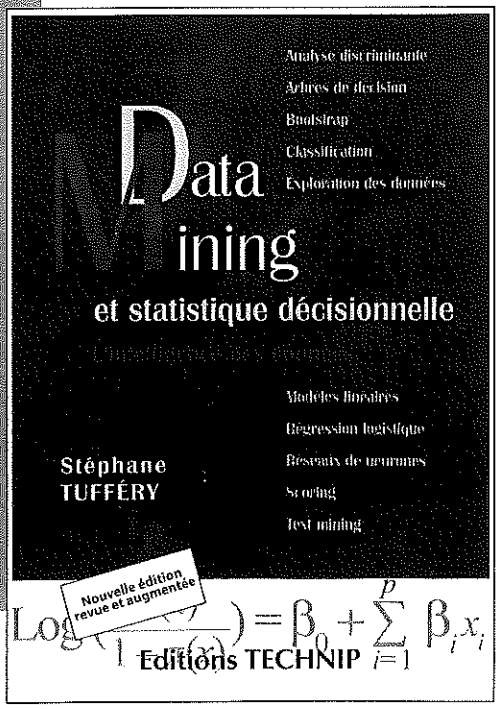
Table des matières


Les techniques de Sondage
Pascal ARDILLY
Nouvelle édition actualisée et augmentée
Editions TECHNIP


SCIENCES SUP
Cours et cas pratiques
Master • Écoles d'ingénieurs
MÉTHODES D'ENQUÊTES ET SONDAGES
Pratiques européenne et nord-américaine
Sous la direction de
Pierre Lavallée et Louis-Paul Rivest
DUNOD


Analyse canonique
Classification
Composantes principales
Corrélation
Espérance conditionnelle
Probabilités analyse des données et Statistique
Estimation
Monte-Carlo
Régression
Tests
Vraisemblance
Gilbert SAPORTA
2e édition révisée et augmentée
Editions TECHNIP


ANALYSE STATISTIQUE DES DONNÉES SPATIALES
Jean-Jacques Droesbeke – Michel Lejeune – Gilbert Saporta
Éditeurs
Editions TECHNIP


Analyse discriminante
Arbres de décision
Bootstrap
Classification
Exploration des données
Data mining
et statistique décisionnelle
Modèles linéaires
Régression logistique
Réseaux de neurones
Scoring
Text mining
Stéphane TUFFÉRY
Nouvelle édition revue et augmentée
Editions TECHNIP