Journal of Statistical Computation and Simulation
Vol. 75, No. 11, November 2005, 935–939

Taylor & Francis
Taylor & Francis Group

Book review

Nine Excellent Books on Data Mining and Statistical Computations

[1] Data Analysis: An Introduction, by Michael S. Lewis-Beck, 1995, Sage Publication, Thousand Oaks, CA, USA, ISBN: 0 8039 5772 6, pp. 77 + ix. $15.95

[2] The Design and Analysis of Computer Experiments, by Thomas J. Santer, Brian J. Williams, and William I. Notz, 2003, Springer Verlag Publication, New York, USA, ISBN: 0 387 95420 1, pp. 283 + xii. $69.95

[3] Statistical Data Mining and Knowledge Discovery, edited by Hamparsum Bozdogan, 2004, CRC Press Publication, Boca Raton, FL, USA, ISBN: 1 58488 344 8, pp. 588. $99.95

[4] Data Mining: Multimedia, Soft Computing, and Bioinformatics, by Shmita Mitra and Tinku Acharya, 2003, John Wiley and Sons Publication, Hoboken, NJ, USA, ISBN: 0 471 46054 0, pp. 401 + xviii. $89.95

[5] Computational Methods in Statistics and Econometrics, by Hisashi Tanizaki, 2004, Marcel Dekker Inc Publication, New York, USA, ISBN: 0 8247 4804 2, pp. 494 + xix. $150.00

[6] Numerical Issues in Statistical Computing for the Social Scientists, by Micah Altman, Jeff Gill, and Michael P. McDonald, 2004, John Wiley and Sons Publication, Hoboken, NJ, USA, ISBN: 0 471 23633 0, pp. 323 + xv. $89.95

[7] Practical Genetic Algorithms, by Randy L. Haupt and Sue Ellen Haupt, 2004, John Wiley and Sons Inc. Publication, Hoboken, NJ, USA, ISBN: 0 471 45565 2, pp. 253 + xvii. $74.95

[8] Chemometrics: From Basics to Wavelet Transform, by Foo-tim Chau, Yi-zeng Liang, Junbin Gao, Xue-guang Shao, 2004, John Wiley and Sons Inc. Publication, Hoboken, NJ, USA, ISBN: 0 471 20242 8, pp. 316 + xiv. $99.95

[9] Chemometrics: Data Analysis for the Laboratory and Chemical Plant, by Richard G. Brereton, 2003, John Wiley and Sons Inc. Publication, Hoboken, NJ, USA, ISBN: 0 471 48978 6, pp. 489 + xiv. $55.00

In this review, impressive books on an emerging discipline called data mining and its related statistical topic known as computational statistics are reviewed. What is data mining? It is an approach to extract knowledge that is of interest to the user. Data mining is a growing area of research in academia and in industry. Traditional data mining used to be all about flat-file transformations. Soft computing tools consist of fuzzy sets, neural networks, genetic algorithms, and rough sets. Computational methods are an integral part of statistical applications. Without computational methods, data mining will remain more of a philosophy and not a practical science.

Journal of Statistical Computation and Simulation
ISSN 0094-9655 print/ISSN 1563-5163 online © 2005 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/00949650412331320972

The first book, M. S. Lewis-Beck's 'Data Analysis' provides an excellent coverage of *motivating introduction, data gathering and coding techniques, univariate statistics, measures of association for ordinal–nominal–dichotomous data, significance testing, bivariate measures of association, simple regression, prediction, goodness of fit tests, multiple regression including colinearity and nonlinearity,* and *recommendations for further developments.* Wintergreen's college data are explained. Basic knowledge of statistics is necessary to appreciate and understand the contents of this book. Statisticians will find it interesting and useful.

The next book titled 'The Design and Analysis of Computer Experiments' and written by T. J. Santner, B. J. Williams, and W. I. Notz offers an excellent coverage of a wide range of topics such as *motivating introduction, examples of computer models such as evolution of fires, physical versus computer experiments, defining experimental goals, modeling output from computer experiments, Gaussian random function, correlation function, smoothing properties, classes of predictors with their properties, prediction when the correlation function is known or unknown, predictive distributions, prediction for multiple response models, space filling designs, designs based on random samples-distance measures-uniformity, designs based on entropy-optimization criteria, sensitivity analysis,* and *model validation.* The list of notations, mathematical facts, and parametric empirical kriging in appendices are helpful to comprehend the materials. Up-to-date developments about computer experiments are narrated and referenced. The chapter notes increase significantly the readers' understanding of the difficult materials in the book. A batch mode program called PERK written in C language is used in all examples. Computing scientists, simulation engineers, statisticians, and applied mathematicians will find this book extremely useful.

Massive data sets have become the norm of professional life in this internet super-highway age. Such massive data sets pose challenges to existing methodologies. Novel approaches are necessary to extract information from large data sets. Statisticians, computer scientists, marketing researchers, image and speech analysts, and fraud detection experts in profile analysis, the artificial intelligence community, and financial engineers among others have been discovering ways to improve the methodology for large data sets. The third book with the title 'Statistical Data Mining and Knowledge Discovery' edited by statistics expert H. Bozdogan is an excellent source to learn more advanced and up-to-date developments in large data sets data mining, information theory, and knowledge discovery. There are 34 research chapters written by international experts in fields covering a wide range of topics including *Bayesian versus frequents approach to data mining, intelligent statistical data mining with information complexity and genetic algorithms, econometric and statistical data mining with prediction and policy making concepts, data mining strategies for the detection of chemical warfare agents, large contingency tables with applications to disability, partial membership models with application to disability survey, automated scoring of polygraph data, missing value algorithms in decision trees, unsupervised learning from incomplete data using a mixture model approach, improving radial basis function with additional information criteria, Kernel-based techniques for sensor validation in nuclear power plants, data mining with traditional regressions, sliced inverse regression, genetic programming to improve time series prediction, monitoring plat devices with pattern classification methods, consumer preferences with data mining methods, testing structural changes over time of brand attributes perception in market segments, principal component analysis with information complexity, global principal component analysis for distributed data, new metric for categorical data, ordinal logistic modeling, latent class factor analysis versus data mining approaches, cluster effects in complex econometric data, neural network based data mining techniques for steel making, data clustering versus string search, data mining techniques for scientific libraries, software for text mining, intensity from statistical entropy point of view, secondary splitting criterion in classification, self-organizing maps to predict barrier removal, cluster analysis for financial data, data mining in federal* *agencies, evaluation of scientific and technological innovation and progress,* and *semantic conference organizer*, among others. These articles were presented in the C. Warren Neel International Conference on Statistical Data Mining and Knowledge Discovery, Knoxville, Tennessee, during June 22–25, 2002. Key note speeches were given by distinguished statisticians such as Arnold Zellner and Edward J. Wegman, and by distinguished data miners such as Usama Fayyad and Naeem Hashmi. Theoretical and applied statisticians would find this book valuable and thought provoking.

The wonderful 'Data Mining' (by S. Mitra and T. Acharya) explains with examples data mining topics such as *introduction to data mining, knowledge discovery concepts, data compression, information retrieval, text mining, web mining, image mining, classification, clustering, rule mining, string matching, bioinformatics, data warehousing, challenges in data mining, soft computing, fuzzy sets in data mining, neural network ideas in data mining, role of genetic algorithms in data mining, use of wavelets in data mining, hybridizations in data mining, information theory in data compression, memory, source coding, principal components, image compression, text compression, linear order string matching algorithms, classification concepts in data mining, support vector machines, clustering in data mining, distance measures, symbolic objects, categorical data mining, hierarchical symbolic clustering, association rules, rule mining with soft computing, video mining, web mining, multimedia data mining, bioinformatics application of data mining,* and *conclusions with discussions.* Up-to-date articles and books are cited in the reference list. Significant features of this book are illustration of use of soft computing in data mining, data compression principles for lossless memory, and application of data mining tools in bioinformatics. Real life examples are considered for illustrating wonderful data mining concepts. Applied statisticians and probabilists will like this book very much.

H. Tanizaki's 'Computational Methods in Statistics' presents current technological capacities and analytical trends in statistics. With several interesting and exciting examples, this book illustrates statistical modeling, Monte Carlo methods, simulation techniques, nonparametric versus parametric approaches to data analysis, state-space modeling, bias correction of ordinary least squares, auto regressive methods, interpretations of statistical and econometric data. So many computer-intensive procedures have been developed in statistics for data analysis. The first half of this book is devoted to Monte Carlo methods. The second part contains illustration of computer-intensive parametric and nonparametric methods. There are eight excellent chapters covering a wide range of topics such as *basic foundations of mathematical statistics and regression, random number generation of univariate and multivariate distributions, composition method, rejection sampling, importance sampling, metropolis-Hastings algorithms, ratio of uniform methods, Gibbs sampling, comparison of sampling methods, Bayesian method of estimation of heteroscedasticity and autocorrelation parameters, bias correction of ordinary least squares in autoregressive models, state-space modeling, recursive versus non-recursive algorithms, asymptotic relative efficiency, power comparison,* and *independence between two samples.* FORTRAN and C languages have been used in describing the algorithms. The author says in the Preface that he used seven computers to illustrate the contents of the book. To use in the Windows 2000 operating system, free Fortran and C compilers at http://www.penwatcom.org, http://www.cygwin.com, and http://www.delorie.com/djgpp could be downloaded and used. For additional free compilers, visit http://www.vni.com/products/imsl. The book has an accompanying CD-ROM that displays source codes, data sets, demonstration of random numbers, transformation methods, etc. With calculus and programming background, readers can appreciate much better the presentation of the contents and also comprehend the materials easily. Recent articles and books are cited in the references of each chapter. This book is quite fitting for use in a graduate level course on computational statistics.

Next we have ‘Numerical Issues in Statistical Computing for the Social Scientist’ by M. Altman, J. Gill, and M. P. McDonald, which illustrates statistical computational topics such as *consequences of numerical inaccuracy, importance of computational statistics, history of computational statistics, motivating examples, rare events counts models, sources of inaccuracy in computations, fundamental concepts, accuracy versus correct inference, algorithmic limitations, evaluating statistical software, robust inference, sensitivity tests, inference for computationally difficult problems, numerical issues in Markov Chain Monte Carlo methods, random number generation, numerical issues in inverting Hessian and generalized inverse matrices, importance resampling, public policy analysis, aliasing, ridge regression, bootstrapping, case study of ecological inference, data perturbation, nonlinear estimation, spatial regression models, convergence issues in logistic regression,* and *recommendation for replications.* Clarity of the presentation is excellent. With a basic background in statistics, readers will appreciate the contents and also understand the material very well. Some amount of programming background will aid the understanding. More resources to the material in this book are found in a web site http://www.hmdc.harvard.edu/\numerical_issues. This book is useful for a graduate level course on statistical computing. Applied statisticians and computer scientists will like this book and find it very useful.

Book number seven is ‘Practical Genetic Algorithms’ by R. L. Haupt and S. E. Haupt, which provides an excellent coverage of *introduction to optimization, root finding versus optimization, minimum seeking algorithms, natural optimization methods, biological optimization, binary versus continuous genetic algorithm, concepts of pairing–mating–mutations-next generation, case studies such as genetic art, word guess, locating emergency response unit, antenna array design, evolution horses, added level of sophistication, advanced application in such as decoding a secret message, robot trajectory planning, stealth design, air pollution receptor modeling, particle swarm optimization, ant colony optimization,* and *evolutionary strategies.* An appendix provides test functions, another appendix explains MATLAB codes and a third displays high performance Fortran codes. A list of glossary adds strength to the book. Statisticians and computing scientists will like this book very much and will benefit greatly from it.

How to extract maximum useful information from data is an aim of data mining. Modeling and processing tools are essential ingredients for this purpose. ‘Chemometrics’, by F. Chau, Y. Liang, J. Gao, and X. Shao, provides an excellent coverage of *introduction to chemometrics, modern analytical chemistry, instrumental response and data processing, white, black, and gray systems, signal processing techniques, wavelets in chemistry, mathematics software for wavelet transform, digital smoothing and filtering methods, moving-window average, Savitsky–Golay filter, Kalman versus spline filtering, Hadamard–Fourier transformation methods, convolution algorithms, data compression, two-dimensional signal processing techniques, principal component analysis, factor analysis, fundamentals of wavelet transform, examples for wavelet function, fast wavelet transform, biorthogonal wavelet transform, application of wavelet transform in chemistry, data denoising, baseline/background removal, resolution enhancement, regression with calibration, classification with pattern recognition, wavelet transform with factor analysis, wavelet neural network, flow injection analysis, spectroscopy,* and *mathematical operations.* An appendix illustrated the needed vector and matrix operations, and another appendix contains descriptions of elementary knowledge of MATLAB. Basic knowledge of calculus is necessary to understand the materials in the book. The presentation is clear. Many examples are given to facilitate a complete understanding. Print and online resources are stated, and websites with MATLAB codes and data sets are given (http://spectroscopynow.com/spy/basehtml/spyH and http://iris4.chem.ohiou.edu). There are several softwares to perform chemometric analyses, and they are MATLAB, SCILAB, MATHEMATICA, MAPLE, MATHCAD, MuPAD, LASTWAVE, MINITAB, EXTRACT, SIRIUS, and XTRICATOR. Statisticians, biochemists, engineers, and health researchers will benefit a lot from this wonderful book.

Lastly, R. G. Brereton's 'Chemometrics' illustrates with examples several chemometrics topics *introduction of chemometrics, descriptions of software, internet sources for further reading, basic principles of experimental design, factorial design, response surface design, mixtures design, simplex optimization, signal processing, sequential signals, linear filters, correlograms and time series analysis, Fourier transform techniques, Kalman filters, Wavelet transform, maximal entropy, Bayesian methods, pattern recognition methods, principal component analysis, unsupervised learning, cluster analysis, supervised learning, calibration, model validation, evolutionary signals, exploratory analysis, preprocessing, determining composition,* and *resolution.* An appendix contains all the necessary mathematical results about vectors, matrices, algorithms, statistical concepts, Excel programs for chemometrics, and Matlab programs for chemometrics. The problems in each exercise are helpful to understand the material in the book. Statisticians, chemical engineers, and computing scientists will find this book valuable and useful.

I learned a lot from these books and hence, highly recommend the above mentioned books to statisticians and professionals in related disciplines.

RAMALINGAM SHANMUGAM
Department of Health Services Research,
Texas State University