



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

International Journal of Forecasting 22 (2006) 195–199

*international journal
of forecasting*

www.elsevier.com/locate/ijforecast

Book reviews

Nong Ye, *The Handbook of Data Mining*, Lawrence Earlbaum Associates, 2003, US\$149.95, 720 pages

The volume is divided into three parts: “Methodologies of Data Mining” (13 chapters) which surveys various tools such as decision trees, association rules, etc.; “Management of Data Mining” (6 chapters) including data collection and cleaning, etc.; and “Applications of Data Mining” (9 chapters) covering human performance data, text data, geospatial data, bioinformatics, customer relationship management (CRM), computer and network security, image data, and manufacturing quality. The preface states the ambitious goals of this 28-chapter, 44-author, over 700-page tome: to present “comprehensive coverage of data mining concepts, algorithms, methodologies, management issues, and tools, which are all illustrated through simple examples and real-world applications for an easy understanding and mastering of those materials. Necessary materials for data mining are presented and organized coherently in one volume.” After reading all 28 chapters, this reviewer concludes that the goals have been met, but only because the competition is so weak. More specifically, a researcher interested in learning about data mining would be better off buying this handbook than spending an equivalent amount of money on other data mining books, but only because the other available books are not nearly so comprehensive. This could have been a classic volume but for its great downfall: it is so poorly edited. Specifically, most of the chapters are not integrated, and very few of them make use of large data sets, preferring instead to work on toy data sets. For example, Chapter 8 is a tutorial on principal components analysis (PCA). Similar information could be obtained from any multivariate text. What neither a multivariate text nor Chapter 8 addresses is how to apply PCA to large data sets.

As another example, the fourth chapter is a (comprehensive and well-written) discussion of univariate and multivariate control charts, a topic not often mentioned in books on data mining. To this reviewer’s great surprise, this fourth chapter makes no mention of how control charts are used in data mining—not even a reference to another chapter in the handbook. Not until nearly 600 pages later, in Chapter 26, does the reader find any hint of how control charts are used in data mining. At the very least, the fourth chapter should have said, “See Chapter 26 for an application of these methods to data mining.” This is not the fault of the fourth chapter’s author (a former colleague whose authority in this area cannot be questioned), but of the editor. By contrast, Chapter 7 on “Strategies and Methods for Prediction” references Chapters 1, 3 and 17 as appropriate. However, Chapter 7 is an exception: most chapters do not draw on other chapters. It is too much to expect each author to read all the other chapters, but this is precisely one of the main jobs of an editor: to unify the chapters, and this the present editor has failed to do. This, however, is not the major editorial lapse marring this volume.

Most of the chapters make only a bow in the direction of data mining, and so miss the opportunity to instruct on the more important points. In particular, most of the examples in the book do not involve large data sets but are only illustrative examples based on small data sets. Thus, much of the book ignores the large problem of how these methods are applied to large data sets; and is this not a central theme of data mining? How does computing a regression on hundreds of observations differ from computing a regression on millions of observations? Readers of this volume will have no idea. Chapter 12, “Nonlinear Time Series Analysis,” discusses many nonlinear methods, but makes no suggestion of how they might be applied to large data sets. It is hard enough to run a linear regression on a million observations;

even harder to do a logistic regression, and very hard to fit a generalized additive model. This chapter suggests that Lyapunov components from the tangent maps can be done via the QR decomposition—without addressing the issue of whether a QR decomposition can be effected for a matrix with millions of rows and perhaps scores of columns. Certainly, this reviewer cannot compute such a QR decomposition on his computer. It would have been nice for the author of this chapter to let the reader know how to compute Lyapunov exponents for large data sets; after all, he is recommending their use in data mining. Chapter 22, “Techniques for Mining Geospatial Databases,” presents several graphs for detecting outliers; with a million observations, all these graphs would be just clouds of ink, and the authors provide no guidance on how to use these graphical methods with large data sets.

In a rare exception, Chapter 9, “Psychometric Methods,” does tangentially address the issue of large data sets, and in a very powerful way. After discussing the basic latent class model (LCM), the authors of Chapter 9 begin a most instructive series of paragraphs, “To enhance the utility of LCM modeling to handle complex and massive data sets, the basic LCM needs to be expanded.” After which follows an extended discussion of how to so expand the LCM to large data sets. If only the editor had instructed each author to do similarly, to take the time and trouble to specifically relate the topic of the chapter to large data sets, this quality of this handbook would have increased by an order of magnitude. But alas, the editor provided no such instruction.

This could have been a great volume, but it is not. It is useful—for now—but before too long someone will write or edit a much more useful book. While libraries should purchase this volume, individuals would do well to wait for something better to come along.

B.D. McCullough

*Department of Decision Sciences,
Lebow College of Business, Drexel University,
Philadelphia, PA 19104-2875, United States.*

E-mail address: bdmccullough@drexel.edu.

Tel.: +1 215 895 2134.

Michael P. Clements, *Evaluating Econometric Forecasts of Economic and Financial Variables*, Palgrave Texts in Econometrics, 2005, 173 pp, ISBN 1-4039-0173-2 (paperback), £19.99, ISBN 1-4039-0172-4 (hardback), £50

Readers of this journal know the author as an Editor of and frequent writer in *IJF*. He is also well known for the two books on forecasting he wrote jointly with D.F. Hendry. They were reviewed in *IJF* (2000), 425–426, and *IJF* (2001), 133–134. These were rather voluminous books on econometric forecasting in general. Here he has chosen to write a smaller book on a very special subject: *the evaluation of forecasts*.

When I first saw the announcement about this book, what came to my mind were some measures of mean error and a couple of tests. I was quite impressed that the author could fill a whole book on the subject. He gives an exposé of the conventional techniques, but also an overview of recent developments that may not be familiar to an ordinary reader of *IJF*. Briefly, models are getting more complex, and so must the methods for evaluating their forecasts. People are not content with point forecasts anymore, they may want to assess future risk through volatility forecasts, or they may need interval, or even density forecasts. The model does not have to be linear and the loss function of the forecaster may be asymmetric. All these topics are thoroughly discussed in this well-written book. It may be added here that forecast evaluation is now topical in the literature. Three recent articles on evaluation are Granger and Jeon (2003a, 2003b) and Peña and Sánchez (2005).

The first chapter is on point forecasts and contains all the concepts one needs, starting with the regression test for (univariate) rationality. The test is open to criticism from the point of view that regression residuals may be autocorrelated, in which case the t -values of the regression coefficients are biased. Autocorrelation is in itself a sign of irrationality. The author suggests testing the residuals in a second step for correlation over time or with any other series known at the time (multivariate rationality). To test for bias and autocorrelation *simultaneously*, we have suggested whitening forecast errors by regressing them on a constant (bias) and a couple of lags (autocorrelation), see Öller and Barot (2000).