Journal of Statistical Software
January 2009, Volume 29, Book Review 7. http://www.jstatsoft.org/

Reviewer: Jan de Leeuw
University of California at Los Angeles

Statistical Methods for Environmental Epidemiology with R

Roger D. Peng and Francesca Dominici
Springer-Verlag, New York, NY, 2008.
ISBN 978-0-387-78166-2. 144 pp. USD 54.95 (P).
http://www.biostat.jhsph.edu/~rpeng/useRbook/

Books in the *Use R!* series are of various different types. There are books which are extended *task views*, dealing with a specific class of R packages, for instance packages for wavelets (Nason 2008), spatial statistics (Bivand, Pebesma, and Gómez-Rubio 2008), or time series analysis (Cowpertwait and Metcalfe 2009). There are books which are extended *vignettes*, dealing with a single package such as **GGobi** (Cook and Swayne 2007) or **lattice** (Sarkar 2008). And there are even books dealing with a single R *function*, such as the one by Ritz and Streibig (2009) on `nls()`. And finally, there are books discussing one or more *case studies*, for example the one by Hahne, Huber, Gentleman, and Falcon (2008) on using **Bioconductor**. The book reviewed here is another such case study, a large and very intersting one, concentrating on actual data analysis and model building.

The book starts with a brief discussion of the most common designs in environmental epidemiology. The data used throughout the book are multivariate time series, in which daily data on air pollution, meteorology, and mortality are collected at multiple sites. There are two huge data sets available, both from studies done at the Johns Hopkins Bloomberg School of Public Health. Chapter 2 discusses the data. The data do not actually come with the R packages **NMMAPSlite** and **MCAPS**, they reside on the Web site http://www.ihapss.jhsph.edu/ and are downloaded as needed. The R packages provide the interface to the data. The **NMMAPS** database has information on about 100 cities over the period 1987–2000, with data for the various causes of mortality and on various pollutant levels. The **MCAPS** database has information about hospitalization in about 200 counties between 1999 and 2002. Data are stored on the server as R dataframes.

The third chapter of the book is a bit of a diversion, but a very pleasant one. It discusses the concept of reproducible research, and some of the R tools that can help with reproducibility. In particular, it discusses the **cacher** (Peng 2008) package that allows the researcher to cache analyses in a database, and to retrieve caches with analyses steps and results. The analyses in the book can all be retrieved from the `ihapps` server. It would have been perfectly fine with me if the book had been about non-local data, caching analyses on a server, and maximizing reproducibility. With an example from environmental health. That is exciting stuff. But

the rest of the book is mostly about traditional statistical model building, although by using these new tools.

Chapter 4 is a brief discussion of statistical issues in estimating health effects from exposure data. It mentions time-varying effects, overdispersion, and hierarchical semi-parametric models. There is just enough here to make it possible to understand what follows in later chapters, but obviously it cannot be an complete, or even an incomplete, introduction to this huge class of problems. In particular Chapter 4 fails to make clear why exactly people working in environmental epidemiology want to build and fit these very complicated regression models. There is some discussion about prediction, but of course the results of most of this type of model building just dissipate in time, new datasets turn out to be quite different, uncertainty is too large, circumstances do change, and no actual predictions are ever made. Fortunately, especially in these dark days, making predictions is not the statistician's reponsibility.

Chapter 5 is interesting because it shows how much one can do with what is commonly called *exploratory data analysis*. I happen to think the distinction between exploratory and confirmatory, or between graphs and models, or between description and inference, is both misleading and annoying, but I realize I am in the minority on this. Anyway, the chapter runs the time series data through a number of filters that produce mostly graphical output and that give insight into the way the data are structured.

Chapter 6 goes into model building and model fitting, and Chapter 7 makes the models bigger by incorporating spatial heterogeneity. What ultimately results is a hierarchical semi-parametric Bayesian Poisson generalized linear/additive model. Going through the various steps in the modeling chapter is undoubtedly useful from a didactic point of view. It illustrates what many statisticians, econometricians and biometricians do for a living, and what the considerations are that make them choose some alternatives over others. Chapter 8 summarizes the model and the results. The model is used to make the graphs look nicer, by smoothing them and providing them with confidence bands.

What makes this book interesting to me is not the precise final form of the regression model, the technical expertise shown in fitting the model, or even the posterior distributions of the parameters. I happen to be interested in air pollution, and a large, clean, and well-organized environmental health database makes me feel all warm inside. Add the way data are handled, stored, manipulated in this book, and the way in which the analyses are cached and can be completely retrieved by anybody who is interested. That, I think, is its most important contribution.

References

Bivand RS, Pebesma EJ, Gómez-Rubio V (2008). *Applied Spatial Data Analysis with R*. Springer-Verlag, New York. ISBN 978-0-387-78170-9.

Cook D, Swayne DF (2007). *Interactive and Dynamic Graphics for Data Analysis – With R and **GGobi***. Springer-Verlag, New York. ISBN 978-0-387-71761-6.

Cowpertwait PSP, Metcalfe A (2009). *Introductory Time Series with R*. Springer-Verlag, New York. ISBN 978-0-387-88697-8.

Hahne F, Huber W, Gentleman R, Falcon S (2008). ***Bioconductor** Case Studies*. Springer-Verlag, New York. ISBN 978-0-387-77239-4.

Nason GP (2008). *Wavelet Methods in Statistics with R*. Springer-Verlag, New York. ISBN 978-0-387-75960-9.

Peng R (2008). "Caching and Distributing Statistical Analyses in R." *Journal of Statistical Software*, **26**(7), 1–24. URL http://www.jstatsoft.org/v26/i07/.

Ritz C, Streibig JC (2009). *Nonlinear Regression with R*. Springer-Verlag, New York. ISBN 978-0-387-09615-5.

Sarkar D (2008). ***lattice**: Multivariate Data Visualization with R*. Springer-Verlag, New York. ISBN 978-0-387-75968-5.

Reviewer:

Jan de Leeuw
University of California at Los Angeles
Department of Statistics
Los Angeles, CA 90095-1554, United States of America
E-mail: deleeuw@stat.ucla.edu
URL: http://gifi.stat.ucla.edu/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 29, Book Review 7 *Published:* 2009-01-05
January 2009