Journal of Statistical Software

September 2005, Volume 14, Book Review 5. http://www.jstatsoft.org/

Reviewer: Jan de Leeuw
University of California at Los Angeles

Correspondence Analysis and Data Coding with Java and R

Fionn Murtagh
Chapman & Hall/CRC, Boca Raton, Florida, 2005.
ISBN 1-58488-528-9. 230 pp. $79.95 (P).
http://www.correspondances.info

This book is difficult to review, because it is at the same time charmingly interesting and maddeningly insular. In the 1970s the school of Jean-Paul Benzécri attained a dominant position in French statistics. This was, to a large extent, because there was a huge vacuum between the probabilists in control of academics and the increasing demands for computerized data analysis. For at least 10 years, French statistics was more or less identical to *Analyse des Données*, or, more specifically, to the dynamic duo *Analyse des Correspondances* and *Taxonomie* (a.k.a. cluster analysis).

But many important developments have taken place since 1980. Benzécri student Michael Greenacre has written extensively about correspondence analysis in widely-read English publications. Books by Benzécri and by his students and co-workers Lebart, Jambu, and Diday have been translated into English. Historical research by Greenacre, Hill, Nishisato and Gifi has pointed out that versions of correspondence analysis appeared previously in the work of Pearson, Fisher, Guttman, Hayashi, Hirschfeld, Lancaster and others. All these developments have managed to make correspondence analysis (under various names) into an internationally recognized statistical technique, which is used extensively, mostly in the social and behavioural sciences. It is no longer true that the technique is the exclusive property of a somewhat esoteric group of French statisticians, and that it is associated with an equally esoteric philosophy of science and data analysis.

The book by Benzécri student Fionn Murtagh we discuss here is, in that specific sense, both revisionary and outdated. It does not go very much beyond the basic framework provided by Benzécri between 1970 and 1980, and it even relies on Benzécri's very limited history of data analysis to place the techniques in a wider context. To be sure, the book is unique and interesting in various aspects. It pays more attention than most English language books on correspondence analysis to coding, updating Pascal code published by Benzécri in 1998. It also discusses, in detail, Benzécri's work on the analysis of free text, with many interesting examples. There is a wonderful foreword by Benzécri himself. The very short index of the book features fanous names such as Jane Austin, Parmenides, and Sokholov.

The book is modern in the sense that it provides us with code and examples done in both R and Java. But again, in this respect it is insular, because there already exist many R packages for correspondence analysis and hierarchical cluster analysis. These alternative packages are not mentioned, let alone compared, and it makes one wonder whether the code in this book was really necessary. Murtagh has also, however briefly, mixed in some references to his later work in neural networks and Kohonen maps.

The particular combination of text, examples, and code in this book make it unique, but very difficult to review. From my point of view it is mainly a historical review of the Benzécri period and the Benzécri toolbox, where the implementation of the toolbox is updated from the old FORTRAN and Pascal code to Java and R. The question is, however, if this is done in the right context, and if the starting point should not have been that Benzécri's work is by now fully integrated in modern multivariate data analysis. From the point of view of statistical software, not much is added to what we already have. It would have been nice, for example, to have code organized into coherent R packages using data objects together wih coding and analysis methods. But the code mainly illustrates the famous adagium that one can write FORTRAN in any language, and it actually seems as if output routines are written in such a way to produce results that maximally resemble line printer output.

And this is, I think, how the book should be viewed. It is an English version, much abbreviated, much better organized and much less idiosyncratic, of the famous two volumes of *L'Analyse des Données* of 1973 and the three volumes of *Pratique de l'Analyse des Données* of 1980. Some small doses of more modern ingredients have been mixed in, but to a large extent while reading this book time stood still.

Reviewer:

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