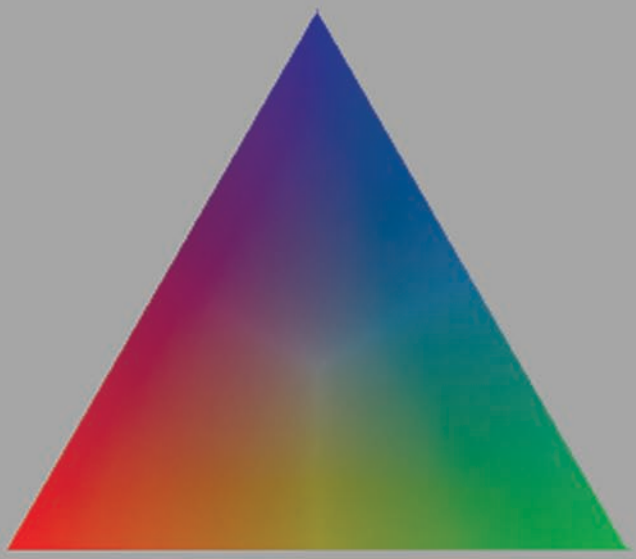
Chapman & Hall/CRC
Interdisciplinary Statistics Series

COMPOSITIONAL DATA ANALYSIS IN PRACTICE



Michael Greenacre

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Compositional Data Analysis in Practice

CHAPMAN & HALL/CRC

Interdisciplinary Statistics Series

Series editors: N. Keiding, B.J.T. Morgan, C.K. Wikle, P. van der Heijden

Recently Published Titles

MENDELIAN RANDOMIZATION: METHODS FOR USING GENETIC VARIANTS IN CAUSAL ESTIMATION
S. Burgess and S.G.Thompson

POWER ANALYSIS OF TRIALS WITH MULTILEVEL DATA
M. Moerbeek and S.Teerenstra

STATISTICAL ANALYSIS OF QUESTIONNAIRES
A UNIFIED APPROACH BASED ON R AND STATA
F. Bartolucci, S. Bacci, and M. Gnaldi

MISSING DATA ANALYSIS IN PRACTICE
T. Raghunathan

SPATIAL POINT PATTERNS
METHODOLOGY AND APPLICATIONS WITH R
A. Baddeley, E Rubak, and R.Turner

CLINICAL TRIALS IN ONCOLOGY, THIRD EDITION
S. Green, J. Benedetti, A. Smith, and J. Crowley

CORRESPONDENCE ANALYSIS IN PRACTICE, THIRD EDITION
M. Greenacre

STATISTICS OF MEDICAL IMAGING
T. Lei

CAPTURE-RECAPTURE METHODS FOR THE SOCIAL AND MEDICAL SCIENCES
D. Böhning, P. G. M. van der Heijden, and J. Bunge

THE DATA BOOK
COLLECTION AND MANAGEMENT OF RESEARCH DATA
Meredith Zozus

MODERN DIRECTIONAL STATISTICS
C. Ley and T. Verdebout

SURVIVAL ANALYSIS WITH INTERVAL-CENSORED DATA
A PRACTICAL APPROACH WITH EXAMPLES IN R, SAS, AND BUGS
K. Bogaerts, A. Komarek, E. Lesaffre

STATISTICAL METHODS IN PSYCHIATRY AND RELATED FIELD
LONGITUDINAL, CLUSTERED AND OTHER REPEAT MEASURES DATA
Ralitza Gueorguieva

FLEXBILE IMPUTATION OF MISSING DATA, SECOND EDITION
Stef van Buuren

COMPOSITIONAL DATA ANALYSIS IN PRACTICE
Michael Greenacre

For more information about this series, please visit: https://www.crcpress.com/go/ids

Compositional Data Analysis in Practice

By
Michael Greenacre

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20180601

International Standard Book Number-13: 978-1-138-31661-4 (Hardback)
International Standard Book Number-13: 978-1-138-31643-0 (Paperback)

In memoriam:
John Aitchison and Paul Lewi,
who both had a clear vision of
compositional data analysis.

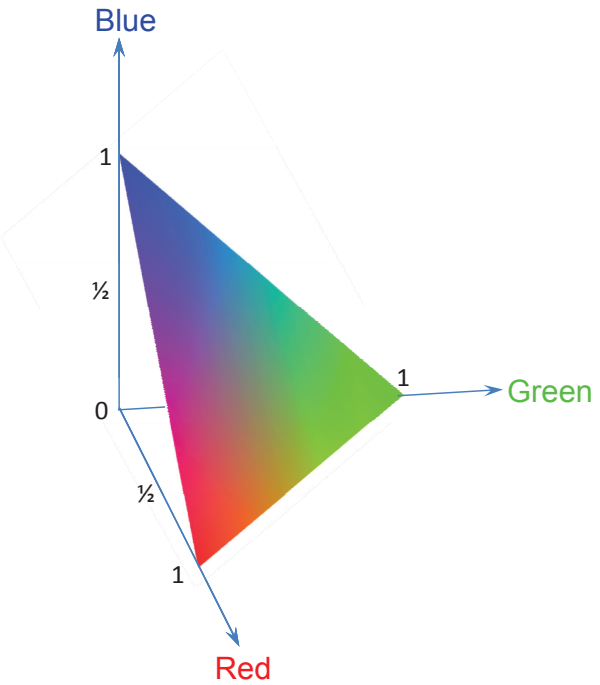
Contents

Preface

Some background

I first became interested in compositional data analysis when I attended a talk by John Aitchison in Girona, Catalonia, in the year 2000. It was one of those life-changing moments that comes to one by sheer good luck. The first slide of John's talk was a blank triangle — I sensed immediately that this talk was going to be interesting!

The first time I learnt about triangular coordinates and the simplex geometry of compositional data was about two decades before meeting Aitchison, in the book *Geological Factor Analysis*, by Karl Jöreskog and co-authors Klovan and Reyment (1978). This book contained a figure showing how trivariate proportions that add up to 1 lie exactly in a planar triangle joining the unit points on three axes, [1 0 0], [0 1 0] and [0 0 1], in three-dimensional space, for example the RGB colour triangle which mixes the three colours red, green and blue in proportions to make any colour.



Since my doctoral studies in France in the 1970s, I had been involved with correspondence analysis, which analyses and visualizes sets of relative frequencies. So I had already realized the relevance of triangular (or ternary) coordinates for displaying trivariate proportions, and generalizations to a three-dimensional tetrahedron for quadrivariate proportions, and so on: in general, a hyper-simplex with n vertices in a space of dimensionality $n - 1$ for n-variate proportions.

I was working mainly with frequency data in the social and environmental sciences and had always considered frequencies relative to their total as the way to "relativize" the data — so I was always working in a simplex, specifically an irregular simplex structured by the chi-square distance. But John Aitchison's talk was about taking compositional data out of the simplex into unrestricted vector space. He did this by expressing data values as pairwise ratios, and then logarithmically transforming them, creating so-called "logratios". This was a revelation to me, and immediately after his talk, with great excitement, I expressed to John the potential of embedding his approach in a biplot of a compositional data matrix.

Of course, this was an obvious idea and not new — indeed, he told me that he had tried to publish an article 10 years earlier in the journal *Applied Statistics*, that it had been turned down and, out of irritation with the criticisms in the reports of the editor and referees, he had just dumped the whole thing. I was so interested that I asked him for that rejected article and the reports and then undertook a complete rewriting of the article. We eventually succeeded in getting it published by the same journal — see Aitchison and Greenacre (2002), listed in Appendix B. John's talk sparked my interest in this field, so closely allied to correspondence analysis. The publication of our joint paper completely cemented my interest and I have been actively working on compositional data analysis ever since.

Almost simultaneously with this immersion into compositional data analysis, I met Paul Lewi, the head of molecular biology at the company Janssen Pharmaceutica in Belgium. I had been aware of Lewi's work as early as the mid 1970s, having seen examples of what he called the *spectral map*, which looked just like a biplot. This was the visualization of a table of positive data that were first logarithmically transformed, then double-centred so that the row and column means were zero, and then this transformed table finally subjected to the usual dimension-reducing steps in forming a joint display of the rows and columns.

My first meeting with Paul, who came to a medical conference in Barcelona in 2002, rekindled my interest in his spectral map, and I soon realized that this method was almost identical to Aitchison and Greenacre's compositional biplot, with one crucial difference: the spectral map included weighting the rows and columns of the table proportional to the row and column marginal totals, just like in correspondence analysis. Paul and I subsequently published the connection between the spectral map, known since the 1970s but not well-known in the statistical literature, and the logratio biplot — see Greenacre and Lewi (2009) in Appendix B. I subsequently made another discovery, that the spectral map (called logratio analysis in this book) and correspondence analysis were not two completely unrelated methods, but actually part of the same family, linked by the Box-Cox power transformation that converges to the logarithmic function as the power parameter tends to zero.