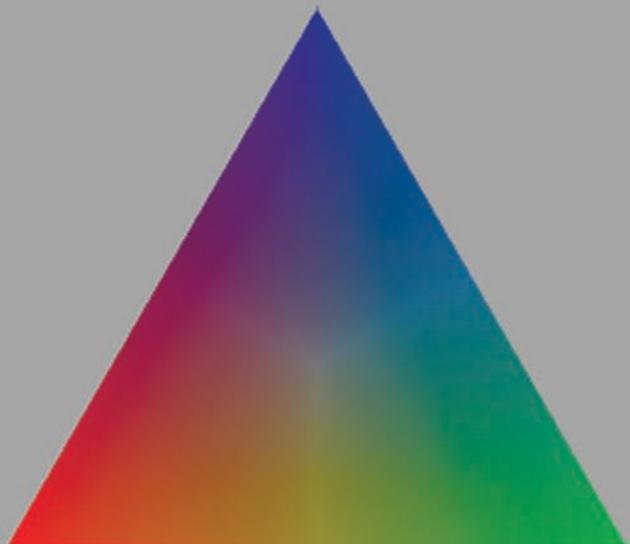


Chapman & Hall/CRC
Interdisciplinary Statistics Series

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

FLEXBIBLE 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

© 2019 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

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)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

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



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Contents

Preface	xi
<hr/>	
1 What are compositional data, and why are they special?	1
1.1 Examples of compositional data	1
1.2 Why are compositional data different from other types of data?	2
1.3 Basic terminology and notation in compositional data analysis	3
1.4 Basic principles of compositional data analysis	5
1.5 Ratios and logratios	6
<hr/>	
2 Geometry and visualization of compositional data	9
2.1 Simple graphics	9
2.2 Geometry in a simplex	12
2.3 Moving out of the simplex	14
2.4 Distances between points in logratio space	15
<hr/>	
3 Logratio transformations	17
3.1 Additive logratio transformations	17
3.2 Centred logratio transformations	18
3.3 Logratios incorporating amalgamations	19
3.4 Isometric logratio transformations	19
3.5 Comparison of logratios in practice	21
3.6 Practical interpretation of logratios	23
<hr/>	
4 Properties and distributions of logratios	25
4.1 Lognormal distribution	25
4.2 Logit function	27
4.3 Additive logistic normal distribution	27
4.4 Logratio variances and covariances	28

4.5	Testing for multivariate normality	30
4.6	When logratios are not normal	31
5	Regression models involving compositional data	33
5.1	Visualizing ratios as a graph	33
5.2	Using simple logratios as predictors	34
5.3	Compositions as responses – total logratio variance	37
5.4	Redundancy analysis	39
6	Dimension reduction using logratio analysis	41
6.1	Weighted principal component analysis	41
6.2	Logratio analysis	42
6.3	Different biplot scaling options	44
6.4	Constrained compositional biplots	46
7	Clustering of compositional data	49
7.1	Logratio distances between rows and between columns	49
7.2	Clustering based on logratio distances	50
7.3	Weighted Ward clustering	50
7.4	Isometric logratio versus amalgamation balances	53
8	Problem of zeros, with some solutions	57
8.1	Zero replacement	57
8.2	Sensitivity to zero replacement	58
8.3	Subcompositional incoherence	59
8.4	Correspondence analysis alternative	60
9	Simplifying the task: variable selection	65
9.1	Explaining total logratio variance	65
9.2	Stepwise selection of logratios	66
9.3	Parsimonious variable selection	68
9.4	Amalgamation logratios as variables for selection	70
9.5	Signal and noise in compositional data	70
10	Case study: Fatty acids of marine amphipods	73
10.1	Introduction	73
10.2	Material and methods	74
10.3	Results	75
10.4	Discussion and conclusion	80

A Appendix: Theory of compositional data analysis	81
A.1 Basic notation	81
A.2 Ratios and logratios	82
A.3 Logratio distance	85
A.4 Logratio variance	85
A.5 Logratio analysis (LRA).....	86
A.6 Principal component analysis (PCA).....	87
A.7 Procrustes analysis	88
A.8 Constrained logratio analysis and redundancy analysis	89
A.9 Permutation tests	89
A.10 Weighted Ward clustering	90
B Appendix: Bibliography of compositional data analysis	91
B.1 Books	91
B.2 Articles	92
B.3 Web resources	95
C Appendix: Computation of compositional data analysis	97
C.1 Simple graphics for compositional data	97
C.2 Logratio transformations	98
C.3 Compositional data modelling	102
C.4 Compositional data analytics.....	103
D Appendix: Glossary of terms	111
E Appendix: Epilogue	115
Index	119



Taylor & Francis

Taylor & Francis Group

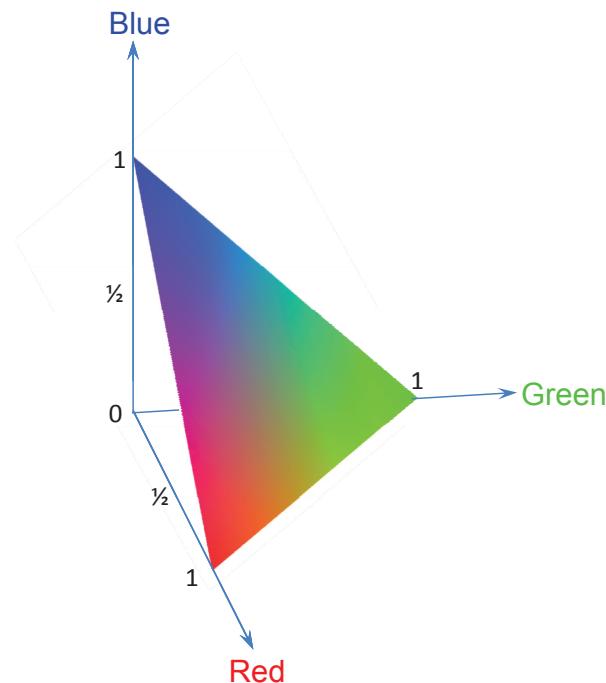
<http://taylorandfrancis.com>

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.