Contents

1 Introduction 1
1.1 What Are Compositional Data? 1
1.1.1 Compositions Are Portions of a Total 1
1.1.2 Compositions Are Multivariate by Nature 3
1.1.3 The Total Sum of a Composition Is Irrelevant 3
1.1.4 Practically All Compositions Are Subcompositions 4
1.1.5 The Very Brief History of Compositional Data Analysis 5
1.1.6 Software for Compositional Data Analysis 6
1.2 Getting Started with R 6
1.2.1 Software Needed for the Examples in This Book 6
1.2.2 Installing R and the Extension Packages 7
1.2.3 Basic R Usage 7
1.2.4 Getting Help for Composition Specific Commands 10
1.2.5 Troubleshooting 11
References 11

2 Fundamental Concepts of Compositional Data Analysis 13
2.1 A Practical View to Compositional Concepts 13
2.1.1 Definition of Compositional Data 13
2.1.2 Subcompositions and the Closure Operation 14
2.1.3 Completion of a Composition 15
2.1.4 Compositions as Equivalence Classes 17
2.1.5 Perturbation as a Change of Units 18
2.1.6 Amalgamation 19
2.1.7 Missing Values and Outliers 20
2.2 Principles of Compositional Analysis 20
2.2.1 Scaling Invariance 20
2.2.2 Perturbation Invariance 21
2.2.3 Subcompositional Coherence 22
2.2.4 Permutation Invariance 23

Bibliografische Informationen
http://d-nb.info/103043803X

digitalisiert durch DEUTSCHE NATIONAL BIBLIOTHEK

2.3 Elementary Compositional Graphics 23
2.3.1 Sense and Nonsense of Scatterplots of Components 24
2.3.2 Ternary Diagrams 24
2.3.3 Log-Ratio Scatterplots 27
2.3.4 Bar Plots and Pie Charts 28
2.4 Multivariate Scales 29
2.4.1 Classical Multivariate Vectorial Data (rmult) 31
2.4.2 Positive Data with Absolute Geometry (rplus) 31
2.4.3 Positive Data with Relative Geometry (aplus) 32
2.4.4 Compositional Data with Absolute Geometry (rcomp) 32
2.4.5 Compositional Data with Aitchison Geometry (acomp) 33
2.4.6 Count Compositions (ccomp) 34
2.4.7 Practical Considerations on Scale Selection 34
2.5 The Aitchison Simplex 37
2.5.1 The Simplex and the Closure Operation 37
2.5.2 Perturbation as Compositional Sum 37
2.5.3 Powering as Compositional Scalar Multiplication 39
2.5.4 Compositional Scalar Product, Norm, and Distance 39
2.5.5 The Centered Log-Ratio Transformation (clr) 41
2.5.6 The Isometric Log-Ratio Transformation (ilr) 42
2.5.7 The Additive Log-Ratio Transformation (alr) 44
2.5.8 Geometric Representation of Statistical Results 45
2.5.9 Expectation and Variance in the Simplex 47
References 49

3 Distributions for Random Compositions 51
3.1 Continuous Distribution Models 51
3.1.1 The Normal Distribution on the Simplex 51
3.1.2 Testing for Compositional Normality 53
3.1.3 The Dirichlet Distribution 58
3.1.4 The Aitchison Distribution 61
3.2 Models for Count Compositions 62
3.2.1 The Multinomial Distribution 62
3.2.2 The Multi-Poisson Distribution 64
3.2.3 Double Stochastic Count Distributions 66
3.3 Relations Between Distributions 67
3.3.1 Marginalization Properties 67
3.3.2 Conjugated Priors 70
References 70

4 Descriptive Analysis of Compositional Data 73
4.1 Descriptive Statistics 73
4.1.1 Compositional Mean 74
4.1.2 Metric Variance and Standard Deviation 75
4.1.3 Variation Matrix and Its Relatives 76
4.1.4 Variance Matrices 80
4.1.5 Normal-Based Predictive and Confidence Regions 82

4.2 Exploring Marginals 85
4.2.1 The Three Types of Compositional Marginals 85
4.2.2 Ternary Diagram Matrices for Multidimensional Compositions 87
4.3 Exploring Projections 89
4.3.1 Projections and Balances 89
4.3.2 Balance Bases 91
4.3.3 The Coda-Dendrogram 92
References 93

5 Linear Models for Compositions 95
5.1 Introduction 95
5.1.1 Classical Linear Regression (Continuous Covariables) 95
5.1.2 Classical Analysis of the Variance (Discrete Covariables) 100
5.1.3 The Different Types of Compositional Regression and ANOVA 102
5.2 Compositions as Independent Variables 103
5.2.1 Example 103
5.2.2 Direct Visualization of the Dependence 103
5.2.3 The Model 106
5.2.4 Estimation of Regression Parameters 107
5.2.5 Displaying the Model 108
5.2.6 Prediction and Predictive Regions 110
5.2.7 Model Checks 112
5.2.8 The Strength of the Relationship 112
5.2.9 Global and Individual Tests 114
5.2.10 Model Diagnostic Plots 115
5.2.11 Checking the Normality Assumptions 117
5.2.12 Checking Constant Variance 118
5.2.13 Robust Regression 118
5.2.14 Quadratic Regression 120
5.3 Compositions as Dependent Variables 122
5.3.1 Example 122
5.3.2 Direct Visualization of the Dependence 125
5.3.3 The Model and Its R Interface 129
5.3.4 Estimation and Representation of the Model Parameters 134
5.3.5 Testing the Influence of Each Variable 138
5.3.6 Comparing Predicted and Observed Values 140
5.3.7 Prediction of New Values 142
5.3.8 Residuals 145
5.3.9 Measures of Compositional Determination 152
5.4 Compositions as Both Dependent and Independent Variables 154
5.4.1 Example 154
5.4.2 Visualization 155

5.4.3 The Composition-to-Composition Regression Model 155
5.4.4 Representation and Visualization of Coefficients 157
5.5 Advanced Considerations 161
5.5.1 Interaction Models 161
5.5.2 Akaike Information Criterion 165
5.5.3 Model Selection 166
5.5.4 Outliers and Heavy-Tailed Distributions 173
References 175

6 Multivariate Statistics 177
6.1 Principal Component Analysis: Exploring Codependence 177
6.1.1 The Singular Value Decomposition 177
6.1.2 Covariance Biplot as Exploratory Tool 178
6.1.3 The Scree Plot and the Form Biplot 183
6.1.4 Loadings and Evolution Plots 184
6.2 Cluster Analysis: Detecting Natural Groups 189
6.2.1 Compositional Distances 190
6.2.2 Agglomeration Techniques 191
6.2.3 Compositional Q-Mode Cluster Analysis 192
6.3 Discriminant Analysis 193
6.4 Other Multivariate Techniques 199
6.4.1 Canonical Correlation 199
6.4.2 Logistic Regression 199
6.4.3 Geostatistics 200
References 206

7 Zeroes, Missings, and Outliers 209
7.1 Descriptive Analysis with and of Missings 209
7.1.1 Types of Zeroes and Missing Values 209
7.1.2 Ternary Diagrams with Missing Values 211
7.1.3 Representing Missing Values Themselves 212
7.1.4 Programming with Missings 213
7.2 Working with Missing Values 214
7.2.1 Imputation 214
7.2.2 Missingness at Random as Projected Information 219
7.2.3 Treating Missing at Random 220
7.2.4 Treating Missings Completely at Random 224
7.2.5 Treating True and Structural Zeros 225
7.2.6 Treating Not Random Missings 225
7.2.7 Treating Values Below the Detection Limit 228
7.2.8 The Physical Meaning of a Detection Limit 232
7.3 Outliers 237
7.3.1 Background on Outliers 237
7.3.2 Types of Outliers 240
7.3.3 Robust Estimation 241
7.3.4 Detection and Classification of Outliers 241

7.4 Descriptive Analysis of Outliers 242
7.4.1 Robust Estimation of Mean and Variance 242
7.4.2 Detecting Outliers 243
7.4.3 Counting Outliers 246
7.4.4 Classifying Outliers 247
7.5 Working with Outliers 249
7.5.1 Random Outliers 249
7.5.2 Failed Measurement or Transmission in Individual Components 249
7.5.3 Subpopulations 250
7.5.4 Polluted Samples 251
7.5.5 Heavy-Tailed Compositional Distributions 252
References 253

Index 255