
Contents

1	Introduction	1
1.1	Data Analysis	1
1.2	Notes on the History of Data Analysis	3
1.2.1	Biometry	4
1.2.2	Era Piscatoria	4
1.2.3	Psychometrics	5
1.2.4	Analysis of Proximities	7
1.2.5	Genesis of Correspondence Analysis	8
1.3	Correspondence Analysis or Principal Components Analysis	9
1.3.1	Similarities of These Two Algorithms	9
1.3.2	Introduction to Principal Components Analysis	10
1.3.3	An Illustrative Example	11
1.3.4	Principal Components Analysis of Globular Clusters	13
1.3.5	Correspondence Analysis of Globular Clusters	14
1.4	R Software for Correspondence Analysis and Clustering	17
1.4.1	Fuzzy or Piecewise Linear Coding	17
1.4.2	Utility for Plotting Axes	18
1.4.3	Correspondence Analysis Program	18
1.4.4	Running the Analysis and Displaying Results	20
1.4.5	Hierarchical Clustering	21
1.4.6	Handling Large Data Sets	27
2	Theory of Correspondence Analysis	29
2.1	Vectors and Projections	29
2.2	Factors	32
2.2.1	Review of Metric Spaces	32
2.2.2	Clouds of Points, Masses, and Inertia	34
2.2.3	Notation for Factors	35
2.2.4	Properties of Factors	36
2.2.5	Properties of Factors: Tensor Notation	36
2.3	Transform	38
2.3.1	Forward Transform	38
2.3.2	Inverse Transform	38
2.3.3	Decomposition of Inertia	38
2.3.4	Relative and Absolute Contributions	39
2.3.5	Reduction of Dimensionality	39

2.3.6	Interpretation of Results	39
2.3.7	Analysis of the Dual Spaces	40
2.3.8	Supplementary Elements	41
2.4	Algebraic Perspective	41
2.4.1	Processing	41
2.4.2	Motivation	41
2.4.3	Operations	42
2.4.4	Axes and Factors	43
2.4.5	Multiple Correspondence Analysis	44
2.4.6	Summary of Correspondence Analysis Properties	46
2.5	Clustering	46
2.5.1	Hierarchical Agglomerative Clustering	46
2.5.2	Minimum Variance Agglomerative Criterion	49
2.5.3	Lance-Williams Dissimilarity Update Formula	49
2.5.4	Reciprocal Nearest Neighbors and Reducibility	52
2.5.5	Nearest-Neighbor Chain Algorithm	53
2.5.6	Minimum Variance Method in Perspective	54
2.5.7	Minimum Variance Method: Mathematical Properties	55
2.5.8	Simultaneous Analysis of Factors and Clusters	57
2.6	Questions	57
2.7	Further R Software for Correspondence Analysis	58
2.7.1	Supplementary Elements	58
2.7.2	FACOR: Interpretation of Factors and Clusters	61
2.7.3	VACOR: Interpretation of Variables and Clusters	64
2.7.4	Hierarchical Clustering in C, Callable from R	67
2.8	Summary	69
3	Input Data Coding	71
3.1	Introduction	71
3.1.1	The Fundamental Role of Coding	72
3.1.2	“Semantic Embedding”	73
3.1.3	Input Data Encodings	75
3.1.4	Input Data Analyzed Without Transformation	76
3.2	From Doubling to Fuzzy Coding and Beyond	77
3.2.1	Doubling	77
3.2.2	Complete Disjunctive Form	79
3.2.3	Fuzzy, Piecewise Linear or Barycentric Coding	80
3.2.4	General Discussion of Data Coding	85
3.2.5	From Fuzzy Coding to Possibility Theory	86
3.3	Assessment of Coding Methods	92
3.4	The Personal Equation and Double Rescaling	98
3.5	Case Study: DNA Exon and Intron Junction Discrimination	99
3.6	Conclusions on Coding	103
3.7	Java Software	104
3.7.1	Running the Java Software	105

4	Examples and Case Studies	111
4.1	Introduction to Analysis of Size and Shape	111
4.1.1	Morphometry of Prehistoric Thai Goblets	111
4.1.2	Software Used	116
4.2	Comparison of Prehistoric and Modern Groups of Canids	118
4.2.1	Software Used	130
4.3	Craniometric Data from Ancient Egyptian Tombs	135
4.3.1	Software Used	139
4.4	Time-Varying Data Analysis: Examples from Economics	140
4.4.1	Imports and Exports of Phosphates	140
4.4.2	Services and Other Sectors in Economic Growth	145
4.5	Financial Modeling and Forecasting	148
4.5.1	Introduction	148
4.5.2	Brownian Motion	149
4.5.3	Granularity of Coding	150
4.5.4	Fingerprinting the Price Movements	158
4.5.5	Conclusions	160
5	Content Analysis of Text	161
5.1	Introduction	161
5.1.1	Accessing Content	161
5.1.2	The Work of J.-P. Benzécri	161
5.1.3	Objectives and Some Findings	163
5.1.4	Outline of the Chapter	164
5.2	Correspondence Analysis	164
5.2.1	Analyzing Data	164
5.2.2	Textual Data Preprocessing	165
5.3	Tool Words: Between Analysis of Form and Analysis of Content	166
5.3.1	Tool Words versus Full Words	166
5.3.2	Tool Words in Various Languages	167
5.3.3	Tool Words versus Metalanguages or Ontologies	168
5.3.4	Refinement of Tool Words	170
5.3.5	Tool Words in Survey Analysis	171
5.3.6	The Text Aggregates Studied	172
5.4	Towards Content Analysis	172
5.4.1	Intra-Document Analysis of Content	172
5.4.2	Comparative Semantics: Diagnosis versus Prognosis	174
5.4.3	Semantics of Connotation and Denotation	175
5.4.4	Discipline-Based Theme Analysis	175
5.4.5	Mapping Cognitive Processes	176
5.4.6	History and Evolution of Ideas	176
5.4.7	Doctrinal Content and Stylistic Expression	177
5.4.8	Interpreting Antinomies Through Cluster Branchings	179
5.4.9	The Hypotheses of Plato on The One	179

5.5	Textual and Documentary Typology	180
5.5.1	Assessing Authorship	180
5.5.2	Further Studies with Tool Words and Miscellaneous Approaches	184
5.6	Conclusion: Methodology in Free Text Analysis	186
5.7	Software for Text Processing	188
5.8	Introduction to the Text Analysis Case Studies	189
5.9	Eight Hypotheses of Parmenides Regarding the One	190
5.10	Comparative Study of Reality, Fable and Dream	197
5.10.1	Aviation Accidents	198
5.10.2	Dream Reports	198
5.10.3	Grimm Fairy Tales	199
5.10.4	Three Jane Austen Novels	199
5.10.5	Set of Texts	200
5.10.6	Tool Words	200
5.10.7	Domain Content Words	201
5.10.8	Analysis of Domains through Content-Oriented Words	205
5.11	Single Document Analysis	207
5.11.1	The Data: Aristotle's <i>Categories</i>	207
5.11.2	Structure of Presentation	210
5.11.3	Evolution of Presentation	214
5.12	Conclusion on Text Analysis Case Studies	220
6	Concluding Remarks	221
	References	223
	Index	229