

Contents

1	Grade Data Analysis — A First Look	1
	<i>F. Ruland</i>	
1.1	”Questions” from clients	1
1.2	About ”Grade Models and Methods for Data Analysis”	2
1.3	Addressing the practitioner	3
1.4	Addressing the theorist	5
1.5	Regarding the analysis of data populations	6
1.6	Overview of Grade Data Analysis algorithms	7
1.7	Returning to the clients from the first page	9
1.8	Conclusion — Chapter 1	11
2	The Grade Approach	13
	<i>F. Ruland</i>	
2.1	Introduction	13
2.2	Part 1: Quick start to the understanding of grade concepts	14
2.2.1	A simplified case of the grade approach	14
2.2.2	Examples of data distribution sources	15
2.3	Steps to making a concentration curve	19
2.4	Quick Start summary	27
2.5	Preview of Part 2, and suggestions before your eventual study of the multivariate material	28
2.6	Part 2: Understanding concentration curves	29
2.6.1	Introduction	29
2.6.2	Two identical distributions	31
2.6.3	Cylinder with partitions: cells of equal length, gas in equal proportions	33
2.6.4	Constructing a concentration curve from individual cat- egory segments	35
2.6.5	When proportions <i>do not</i> correspond between distri- butions	36
2.6.6	Using the concentration curve to introduce the concept of overrepresentation	37
2.6.7	Overrepresentation	38
2.6.8	When we manipulate both distributions: gas (unequal proportions) and cylinder (unequal cell sizes)	41
2.6.9	Example application — Winners versus losers in the car sales market	43

2.6.10	Example application — Historic perspective (then vs. now) of car sales market	45
2.6.11	Reordering (prioritizing) categories — and an introduction to the maximal concentration index	47
2.6.12	Part 2 summary	49
2.7	Chapter Summary	50
3	Univariate Lilliputian Model I	51
	<i>T. Kowalczyk, W. Szczesny</i>	
3.1	Introduction	51
3.2	Lilliputian variables and their basic parameters	52
3.2.1	The cdf of a Lilliputian variable	52
3.2.2	The expectation of a Lilliputian variable and the index ar	56
3.2.3	The first moment Lilliputian variable, its variance, and the Gini Index	61
3.2.4	Discontinuity measures	67
3.3	The main equivalence relation which creates the Univariate Lilliputian Model	70
3.3.1	Preliminary definitions and examples	70
3.3.2	Equivalent pairs of random variables	75
3.3.3	Grade transformations of univariate distributions	77
3.4	Grade parameters	80
3.4.1	The parameter ar	80
3.4.2	Normal concentration pattern	82
3.4.3	Likelihood ratio and local concentration	85
3.5	Appendix	87
3.5.1	Monotone grade probability transition function	87
3.5.2	Properties of concentration measures	89
4	Univariate Lilliputian Model II	91
	<i>T. Kowalczyk, E. Pleszczyńska, W. Szczesny</i>	
4.1	Introduction	91
4.2	Lorenz Curve and Gini Index	94
4.2.1	Ratio variables and related concentration curves	94
4.2.2	First moment distribution and Lorenz curve	101
4.2.3	Lorenz Curves with horizontal and/or vertical segments	104
4.2.4	The variable called overrepresentation and its Lorenz curve	106
4.2.5	Diagram of over- and underrepresentation	111
4.2.6	Lorenz Curve and Gini Index for density transform of categorical variables	115
4.3	Order oriented concentration curves	116
4.3.1	Basic definitions	116
4.3.2	The maximal concentration curve and the maximal concentration index	120

4.3.3	Order oriented Lorenz Curve and inequality (Gini) index	122
4.3.4	Order oriented Lorenz Curve and Gini Index for the density transforms of categorical variables	123
4.3.5	Link with the two-class discriminant analysis	124
4.4	Dual concentration curve	126
4.4.1	Definition of the dual concentration curve and dual Lorenz curve	126
4.4.2	Random variable dual to a ratio variable	129
4.4.3	Dual links between overrepresentation and underrepresentation	130
4.4.4	Towards advantage problems in interpopulation comparisons	132
4.5	Appendix	133
4.5.1	Measurement scales	133
4.5.2	Supplement to Section 4.2 (the inequality measures) . .	135
4.5.3	Supplement to Section 4.3.2 (the maximal concentration measures)	136
4.5.4	Supplement to Section 4.3.3 (the ordered Lorenz Curve and Gini Index)	136
4.5.5	Supplement to Section 4.4.2 (the random variable dual to a ratio variable)	137
4.5.6	Bibliographical remarks to Chapter 3 and 4	137
5	Asymmetry and the inverse concentration set	139
	<i>A. Ciok</i>	
5.1	Introduction	139
5.2	Concentration curves with a common value of the concentration index	140
5.3	Links between asymmetry and opposite orderings	145
5.4	Asymmetry in the Univariate Lilliputian Model	146
5.4.1	Asymmetry curves	146
5.4.2	Asymmetry index	152
5.4.3	Families of curves with special properties	153
5.5	Relative asymmetry	154
5.5.1	Links with measurement scales	154
5.5.2	Relative asymmetry measures	155
5.5.3	Examples	157
5.6	Appendix	162
5.6.1	The inverse concentration set	162
5.6.2	Asymmetry indices	163
5.6.3	Bibliographical remarks	165

6	Discretization and regularity	167
	<i>A. Ciok</i>	
6.1	Introduction	167
6.2	Discretization framework	168
6.3	Optimal discretization for a given number of categories	170
6.4	Ideally regular concentration curves	172
6.5	On the determination of the number of categories	175
6.6	A parametric family of ideally regular Lilliputian curves	178
6.7	Appendix	181
	6.7.1 Optimal discretization	181
	6.7.2 Algorithm of optimal discretization	183
	6.7.3 Bibliographical remarks	184
7	Preliminary concepts of bivariate dependence	185
	<i>T. Kowalczyk and W. Szczesny</i>	
7.1	Introduction	185
7.2	Contingency tables with m rows and k columns	185
7.3	Quadrant dependence	190
7.4	Matrices of ar 's for pairs of profiles. Total positivity of order two	197
7.5	The regression function	205
7.6	The monotone dependence function and the Gini Index	210
7.7	Appendix - Bibliographical remarks	214
8	Dependence Lilliputian Model	217
	<i>T. Kowalczyk, W. Szczesny, W. Wysocki</i>	
8.1	Introduction	217
8.2	Grade bivariate distributions and overrepresentation maps for probability tables	218
8.3	Lilliputian surfaces with uniform marginal distributions	226
8.4	Spearman's ρ and Kendall's τ expressed by volumes and masses in the unit cube	234
8.5	Grade regression functions and related measures	242
8.6	On permuting rows and columns of $m \times k$ probability tables	246
	8.6.1 Maximal grade correlation	246
	8.6.2 Ordered Gini indices for marginal density transforms	250
	8.6.3 Maximal Kendall's τ	256
8.7	The hinged sequences of rows and columns	261
8.8	Appendix: Bibliographical remarks	265
9	Grade Correspondence Analysis and outlier detection	267
	<i>O. Matyja, W. Szczesny</i>	
9.1	Introduction	267
9.2	Algorithms of GCA	268
	9.2.1 GCA algorithm based on Spearman's ρ^*	268
	9.2.2 GCA algorithm based on Kendall's τ	272

9.2.3	GCA algorithm based on τ_{sgn}	274
9.2.4	GCA and a mixture of permuted discretized binormal tables	274
9.2.5	Folds	277
9.3	Algorithm for Smooth Grade Correspondence Analysis (SGCA)	280
9.4	Examples of GCA and SGCA results	282
9.4.1	A mixture of binormals	282
9.4.2	BRIT _{7×7} and CARS _{16×16}	283
9.5	Detection of rows and columns outlying the main trend	286
9.5.1	Scatterplots for rows and for columns	286
9.5.2	Measures of departure from TP ₂	288
9.5.3	Rejecting outlying rows and columns	291
9.6	Appendix - Bibliographical remarks	295
10	Cluster analysis based on GCA	297
	<i>A. Ciok</i>	
10.1	Introduction	297
10.2	Single and double grade clustering	299
10.3	Optimal grade clustering	305
10.4	Cluster analysis in the detection of mixtures	307
10.4.1	Straight and reverse regular structures	307
10.4.2	Survey of small business servicing firms	310
10.4.3	SGCL results for the whole sample	310
10.4.4	SGCL results for the particular branches	312
10.4.5	Some final remarks	314
10.5	Cluster analysis and the detection of an imprecisely defined trend	315
10.5.1	The use of sources of capital by retail trade firms in Poland	315
10.5.2	Typology of firms for the pooled, three-year data	315
10.5.3	Firm typologies for annual data	318
10.5.4	Relationship between the generated firm typology and the firm profitability	319
10.6	On GCCA application to various data sets	321
10.7	Appendix	323
10.7.1	An algorithm for optimal clustering	323
10.7.2	Bibliographical remarks	324
11	Regularity and the number of clusters	325
	<i>A. Ciok</i>	
11.1	Introduction	325
11.2	Generalization of the parabola family from the ULM	325
11.3	The ideal regularity of two-way data tables	330
11.4	Regularity and cluster detection	332
11.5	Cluster detection in finite data tables	335
11.6	Appendix - Bibliographical remarks	338

12	Grade approach to the analysis of finite data matrices	339
	<i>W. Szczesny</i>	
12.1	Introduction	339
12.2	Insight Examples	342
12.2.1	The Competitors-Judges Data (C/J Example)	343
12.2.2	The Annual Bonus Data (A/B Example)	349
12.3	Applicability of GCA	353
12.4	A revisit of the univariate data	356
12.5	Finite multivariate datasets and related inequality measures	361
12.5.1	Finite data tables and their grade regression functions	361
12.5.2	Lorenz Surfaces	366
12.5.3	Global differentiation and its decomposition	373
12.5.4	Decomposition of \mathbf{Dif}_X	377
12.6	Transformations of variables	380
12.7	Detection of outliers and decomposition of a dataset	381
13	Inequality measures for multivariate distributions	385
	<i>W. Szczesny</i>	
13.1	Introduction	385
13.2	Inequality measures for multivariate distributions with finite sets of records	389
13.3	Inequality measures for multivariate distributions with non-finite sets of records	394
13.4	Inequality measures for continuous bivariate distributions	398
13.4.1	A pair of independent uniform Lilliputian variables	398
13.4.2	A pair of functionally dependent Lilliputian variables	405
13.4.3	A family of TP_2 distributions from $\mathbb{B}LM$	406
13.4.4	Grade binormal distributions	409
13.5	Inequality measures for grade multinormal distributions	411
13.6	Inequality measures for the Moran distributions	419
13.7	Appendix - link between grade similarity and dissimilarity of two regularly dependent random variables	422
14	Case studies with multivariate data	425
	<i>W. Szczesny, M. Grzegorek</i>	
14.1	Introduction	425
14.2	Case Study 1 - Main Trend of Questionnaire Data	426
14.2.1	The Questionnaire	426
14.2.2	The goal of the analysis	427
14.2.3	The Overrepresentation Map for Main Trend in dataset TOTAL	428
14.2.4	Interpretation of the results (with some general hints)	430
14.3	Case Study 1 - Decomposition of the dataset into regular sub-populations	432
14.3.1	The Overrepresentation Maps for FIT-MT and OUT-MT	433

14.3.2	The grade strip charts for FIT-MT and OUT-MT . . .	434
14.3.3	Two-way ordered clustering	436
14.4	Case Study 2 - Analysis of Engineering Data (Strength of Concrete)	436
14.4.1	The variables:	437
14.4.2	The goal of the analysis	438
14.4.3	The Overrepresentation Map for Main Trend in the dataset TOTAL	438
14.5	Case Study 2 - Decomposition of concrete mixtures into FIT-MT and OUT-MT	439
14.5.1	The Overrepresentation Maps for FIT-MT and OUT-MT	440
14.5.2	The grade strip charts for FIT-MT and OUT-MT . . .	441
14.6	Final remarks for the two case studies	444
14.7	Appendix:	444
14.7.1	Case Study 1 - further details of the analysis	444
14.7.2	Case Study 2 - further details of the analysis	448
14.7.3	Bibliographical remarks	453
15	The GradeStat program	455
	<i>O. Matyja</i>	
15.1	Introduction	455
15.2	Main implemented features	455
15.2.1	Data overview	455
15.2.2	Charts	456
15.2.3	Preprocessing	456
15.2.4	Ordering	457
15.2.5	Clustering	457
	References	459
	Index	468