
Contents

1	Introduction	1
1.1	Introduction	1
1.2	Data Visualization	4
1.3	Research Literature	7
1.4	How Large Is a Large Dataset?	9
1.5	The Effects of Largeness	17
1.5.1	Storage	18
1.5.2	Quality	19
1.5.3	Complexity	20
1.5.4	Speed	20
1.5.5	Analyses	21
1.5.6	Displays	21
1.5.7	Graphical Formats	22
1.6	What Is in This Book	22
1.7	Software	23
1.8	What Is on the Website	24
1.8.1	Files and Code for Figures	24
1.8.2	Links to Software	24
1.8.3	Datasets	25
1.9	Contributing Authors	26

Part I Basics

2	Statistical Graphics	31
2.1	Introduction	31
2.2	Plots for Categorical Data	31
2.2.1	Barcharts and Spineplots for Univariate Categorical Data	32
2.2.2	Mosaic Plots for Multi-dimensional Categorical Data	33
2.3	Plots for Continuous Data	36

2.3.1	Dotplots, Boxplots, and Histograms	36
2.3.2	Scatterplots, Parallel Coordinates, and the Grand Tour	39
2.4	Data on Mixed Scales	44
2.5	Maps	47
2.6	Contour Plots and Image Maps	49
2.7	Time Series Plots	50
2.8	Structure Plots	51
3	Scaling Up Graphics	55
3.1	Introduction	55
3.2	Upscaling as a General Problem in Statistics	55
3.3	Area Plots	56
3.3.1	Histograms	57
3.3.2	Barcharts	58
3.3.3	Mosaic Plots	60
3.4	Point Plots	62
3.4.1	Boxplots	62
3.4.2	Scatterplots	63
3.4.3	Parallel Coordinates	65
3.5	From Areas to Points and Back	67
3.5.1	α -Blending and Tonal Highlighting	69
3.6	Modifying Plots	71
3.7	Summary	72
4	Interacting with Graphics	73
4.1	Introduction	73
4.2	Interaction	74
4.3	Interaction and Data Displays	75
4.3.1	Querying	75
4.3.2	Selection and Linking	77
4.3.3	Selection Sequences	78
4.3.4	Varying Plot Characteristics	82
4.3.5	Interfaces and Interaction	84
4.3.6	Degrees of Linking	86
4.3.7	Warnings and Redmarking	87
4.4	Interaction and Large Datasets	88
4.4.1	Querying	88
4.4.2	Selection, Linking, and Highlighting	89
4.4.3	Varying Plot Characteristics for Large Datasets	92
4.5	New Interactive Tasks	98
4.5.1	Subsetting	98
4.5.2	Aggregation and Recoding	99
4.5.3	Transformations	99
4.5.4	Weighting	99

4.5.5 Managing Screen Layout 101
 4.6 Summary and Future Directions 101

Part II Applications

5 Multivariate Categorical Data — Mosaic Plots 105
 5.1 Introduction 105
 5.2 Area-based Displays 105
 5.2.1 Weighted Displays and Weights in Datasets 107
 5.3 Displays and Techniques in One Dimension 107
 5.3.1 Sorting and Reordering 110
 5.3.2 Grouping, Averaging, and Zooming 111
 5.4 Mosaic Plots 113
 5.4.1 Combinatorics of Mosaic Plots 114
 5.4.2 Cases per Pixel and Pixels per Case 116
 5.4.3 Calibrating the Eye 116
 5.4.4 Gray-shading 119
 5.4.5 Rescaling Binsizes 122
 5.4.6 Rankings 123
 5.5 Summary 123

6 Rotating Plots 125
 6.1 Introduction 125
 6.1.1 Type of Data 126
 6.1.2 Visual Methods for Continuous Variables 127
 6.1.3 Scaling Up Multiple Views for Larger Datasets 128
 6.2 Beginning to Work with a Million Cases 128
 6.2.1 What Happens in GGobi, a Real-time System? 128
 6.2.2 Reducing the Number of Cases 129
 6.2.3 Density Estimation 131
 6.2.4 Screen Real Estate Indexing 134
 6.3 Software System 135
 6.4 Application 137
 6.4.1 Data Description 137
 6.4.2 Viewing a Tour of the Data 137
 6.4.3 Scatterplot Matrix 138
 6.5 Current and Future Developments 140
 6.5.1 Improving the Methods 140
 6.5.2 Software 141
 6.5.3 How Might These Tools Be Used? 141

7	Multivariate Continuous Data — Parallel Coordinates . . .	143
7.1	Introduction	143
7.2	Interpolations and Inner Products	144
7.3	Generalized Parallel Coordinate Geometry	145
7.4	A New Family of Smooth Plots	149
7.5	Examples	150
7.5.1	Automobile Data	150
7.5.2	Hyperspectral Data: Dealing with Massive Datasets	152
7.6	Detecting Second-Order Structures	154
7.7	Summary	155
8	Networks	157
8.1	Introduction	157
8.2	Layout Algorithms	158
8.2.1	Simple Tree Layout	159
8.2.2	Force Layout Methods	161
8.2.3	Individual Node Movement Algorithms	162
8.3	Interactivity	162
8.3.1	Speed Considerations	164
8.3.2	Interaction and Layout	165
8.4	NicheWorks	166
8.5	Example: International Calling Fraud	167
8.6	Languages for Description and Layouts	172
8.6.1	Defining a Graph	172
8.6.2	Graph Specification via VizML	173
8.7	Summary	174
9	Trees	177
9.1	Introduction	177
9.2	Growing Trees for Large Datasets	178
9.2.1	Scalability of the CART Growing Algorithm	179
9.2.2	Scalability of Pruning Methods	181
9.2.3	Statistical Tests and Large Datasets	183
9.2.4	Using Trees for Large Datasets in Practice	184
9.3	Visualization of Large Trees	187
9.3.1	Hierarchical Plots	187
9.3.2	Sectioned Scatterplots	192
9.3.3	Recursive Plots	195
9.4	Forests for Large Datasets	198
9.5	Summary	202

10 Transactions	203
10.1 Introduction and Background	203
10.2 Mice and Elephant Plots and Random Sampling	205
10.3 Biased Sampling	210
10.3.1 Windowed Biased Sampling	211
10.3.2 Box–Cox Biased Sampling	213
10.4 Quantile Window Sampling	215
10.5 Commonality of Flow Rates	221
11 Graphics of a Large Dataset	227
11.1 Introduction	227
11.2 QuickStart Guide	
Data Visualization for Large Datasets	228
11.3 Visualizing the InfoVis 2005 Contest Dataset	229
11.3.1 Preliminaries	229
11.3.2 Variables	230
11.3.3 First Analyses	230
11.3.4 Multivariate Displays	235
11.3.5 Grouping and Selection	239
11.3.6 Special Features	242
11.3.7 Presenting Results	247
11.3.8 Summary	249
References	251
Authors	263
Index	267