

# Graphisch gestützte Datenanalyse

Von  
Dr. Rainer Schnell

R. Oldenbourg Verlag München Wien

# Inhalt

<b>1 Grundprinzipien graphischer Datenanalyse</b> . . . . .	1
1.1 Ablauf einer graphisch gestützten Datenanalyse . . . . .	2
1.2 Datenanalyseplots und Wahrnehmungpsychologie . . . . .	4
1.3 Datenanalyseplots und "theoriefrei" Beobachtungen . . . . .	8
<b>2 Univariate Plots</b> . . . . .	11
2.1 Dot-Plots . . . . .	11
2.1.1 Eindimensionale Scatterplots . . . . .	11
2.1.2 Stacked-Dot-Plots . . . . .	12
2.1.3 "Jittered" und "textured" Dot-Plots . . . . .	13
2.1.4 Index-Plots . . . . .	15
2.1.5 Q-Plots . . . . .	16
2.2 Boxplots . . . . .	18
2.3 Histogramme . . . . .	21
2.3.1 Bestimmung der Klasseneinteilung . . . . .	21
2.3.2 "Averaged Shifted Histograms" . . . . .	25
2.3.3 Nichtparametrische Dichteschätzer . . . . .	26
2.3.4 Stem-and-Leaf-Display . . . . .	31
<b>3 Plots für den Vergleich empirischer Verteilungen</b> . . . . .	35
3.1 Back-to-Back-Stem-and-Leaf-Displays . . . . .	35
3.2 Gruppierte Boxplots . . . . .	36
3.2.1 Notched-Boxplots . . . . .	37
3.2.2 Box-Dot-Plots . . . . .	39
3.2.3 Perzentil-Plots mit Kenngrößen . . . . .	42
3.2.4 Q-Q-Plots . . . . .	43
3.2.5 Exkurs: Modellierung der Verteilungsunterschiede zweier Gruppen . . . . .	44
3.3 Dot-Charts . . . . .	46
3.3.1 Dot-Charts für den Vergleich von Kenngrößen . . . . .	47
3.3.2 Paired-Dot-Charts für wiederholte Messungen . . . . .	49
<b>4 Plots zum Vergleich theoretischer und empirischer Verteilungen</b> . . . . .	51
4.1 Probability-Plots ("Theoretical Q-Q-Plots") . . . . .	51
4.1.1 Eigenschaften von Probability-Plots . . . . .	51
4.1.2 Konstruktion von Probability-Plots . . . . .	55
4.1.3 Varianten und Anwendungen der Probability-Plots . . . . .	56
4.1.3.1 Detrended Normal-Probability-Plot . . . . .	57
4.1.3.2 Half-Normal-Probability-Plots . . . . .	58
4.1.3.3 Perzentil-Plots (P-P-Plots) . . . . .	59
4.1.3.4 Stabilisierte Probability-Plots (SP-Plots) . . . . .	60
4.1.4.5 Probability-Plots als Basis für Verteilungstests . . . . .	62
4.1.4.6 $\chi^2$ -Probability-Plots zur Prüfung auf multivariate Normalverteilung . . . . .	63
4.1.4 Multivariate Verallgemeinerungen von Q-Q-Plots . . . . .	64
4.2 Plots für kategorisierte Variablen . . . . .	65
4.2.1 Überlagerte Histogramme . . . . .	65
4.2.2 Hängende Histogramme . . . . .	66
4.2.3 Residuen-Rootogramme und "suspended residual rootogram" . . . . .	66
4.2.4 Poissonness-Plots . . . . .	69

4.3 Exkurs: Datentransformationen . . . . .	71
4.3.1 Anwendungen von Datentransformationen . . . . .	71
4.3.2 Power-Transformationen . . . . .	73
4.3.2.1 Praktische Anwendungen von Power-Transformationen . . . . .	73
4.3.2.2 Symmetrieplots als Transformationshilfsmittel . . . . .	76
4.3.2.3 Maximum-Likelihood-Schätzung des Transformationsparameters . . . . .	78
4.3.2.4 Gematchte Power-Transformationen . . . . .	80
4.3.3 Transformationen für Prozentsätze und Anteile . . . . .	81
4.3.4 Fisher-r-z-Transformation . . . . .	83
<b>5 Scatterplots . . . . .</b>	<b>85</b>
5.1 Konstruktion von Scatterplots . . . . .	85
5.1.1 Achsenkalierung und Korrelationswahrnehmung . . . . .	85
5.1.2 Summen-Differenzen-Plots . . . . .	87
5.1.3 Exkurs: Konstruktion von Zeitreihenplots . . . . .	88
5.1.3.1 "Connected Graphs" . . . . .	89
5.1.3.2 Shape-Parameter . . . . .	89
5.1.3.3 "Median Absolute Slope Procedure" . . . . .	90
5.1.3.4 Step-Plots und Spline-Funktionen . . . . .	91
5.2 Scatterplot-Techniken für große Fallzahlen . . . . .	93
5.2.1 Jittering . . . . .	94
5.2.2 Sunflower-Plots . . . . .	95
5.2.3 Cellulation . . . . .	96
5.2.4 Plots der geschätzten bivariaten Dichteverteilung . . . . .	97
5.3 Informationsangereicherte Scatterplots . . . . .	102
5.3.1 Scatterplot-Smoother . . . . .	102
5.3.1.1 Median-Trace . . . . .	104
5.3.1.2 Kernel-Smoothed-Quantile-Plots . . . . .	106
5.3.1.3 K-NN-Smoother und Running-Line-Smoother . . . . .	108
5.3.1.4 LOWESS . . . . .	109
5.3.1.5 Exkurs: Berechnung von LOWESS . . . . .	112
5.3.1.5 Andere Scatterplot-Smoother . . . . .	114
5.3.2 Scatterplots mit Dichte-Ellipsen . . . . .	116
5.3.3 Scatterplots mit univariaten Randverteilungen . . . . .	121
5.3.4 Influence-Plots . . . . .	122
5.3.5 Voronoi-Plots . . . . .	123
<b>6 Plots für drei- und mehrdimensionale Daten . . . . .</b>	<b>125</b>
6.1 Symbolische Scatterplots . . . . .	125
6.2 Scatterplots mit Icons . . . . .	126
6.3 Dreidimensionale Scatterplots . . . . .	130
6.4 Perspektiv-, Kontur- und Imageplots . . . . .	132
6.4.1 Glättungsmethoden . . . . .	133
6.4.2 Darstellungsmethoden . . . . .	135
6.4.3 Nutzungsmöglichkeiten und einschränkungen . . . . .	138
6.5 Bedingte Scatterplots . . . . .	139
6.5.1 Kategorisierte Scatterplots . . . . .	139
6.5.2 Casement-Plots . . . . .	142
6.5.3 Multiwindow-Plots . . . . .	143
6.5.4 Coplots . . . . .	145
6.6 Scatterplot-Matrizen . . . . .	148
6.7 Andrews-Plots . . . . .	150
6.8 Parallel-Koordinatenplots . . . . .	153

6.9 Exkurs: Hochinteraktive Graphik ("Dynamic Graphics") . . . . .	158
6.9.1 Basistechniken . . . . .	158
6.9.2 Anwendungen . . . . .	159
6.9.3 Anwendungsprobleme . . . . .	160
6.10 Kognitionspsychologische Grenzen der Plots mehrdimensionaler Daten . . . . .	162
<b>7 Plots dimensionsreduzierender Verfahren . . . . .</b>	<b>163</b>
7.1 Plots in der Hauptkomponentenanalyse . . . . .	163
7.1.1 Berechnung von Hauptkomponenten . . . . .	164
7.1.2 Bestimmung der Zahl der Hauptkomponenten . . . . .	165
7.1.3 PCA als Projektionstechnik . . . . .	167
7.1.4 Exkurs: Plots in der Faktorenanalyse . . . . .	172
7.1.4.1 Graphische Darstellung von Korrelationsmatrizen: RZ-Plots . . . . .	172
7.1.4.2 Residuenanalyse . . . . .	175
7.2 Biplots . . . . .	176
7.2.1 Berechnung des Biplots . . . . .	176
7.2.2 Interpretation des Biplots . . . . .	179
7.2.3 Darstellung großer Fallzahlen . . . . .	182
7.2.4 Varianten des Biplots . . . . .	183
7.2.4.1 Relative Variationsdiagramme (RV-Plots) . . . . .	183
7.2.4.2 Biplots und andere Projektionstechniken . . . . .	186
7.3 Korrespondenzanalyse . . . . .	187
7.3.1 Berechnung einer Korrespondenzanalyse . . . . .	187
7.3.2 Interpretation der CA-Plots . . . . .	190
7.3.3 Graphische Darstellung der Stabilität der Ergebnisse einer CA . . . . .	196
7.3.4 "Multiple Correspondence Analysis" (MCA) . . . . .	198
7.3.5 CA im Vergleich mit anderen multivariaten Analyseverfahren . . . . .	199
7.3.6 Beurteilung der CA als Analysetechnik . . . . .	200
7.4 Weitere Projektionstechniken . . . . .	201
7.4.1. Nonlinear Mapping (NLM) . . . . .	201
7.4.2 Principal Co-Ordinate Analysis . . . . .	202
7.4.3 Sliced-Inverse-Regression (SIR) . . . . .	204
7.4.4 Plots optimaler Scores . . . . .	204
7.4.5 "Small tour" und "Grand tour" . . . . .	205
7.4.6 Exploratory Projection Pursuit (EPP) . . . . .	206
7.5 Vergleich verschiedener Projektionen: Prokrustes-Analyse . . . . .	209
7.6 Interpretation von Projektionsplots . . . . .	213
<b>8 Plots in der multiplen Regression . . . . .</b>	<b>217</b>
8.1 Annahmen der multiplen Regression . . . . .	219
8.2 Überprüfung der Annahmen . . . . .	221
8.2.1 Residuen-Definitionen . . . . .	222
8.2.2 Prüfgrößen für den Einfluß einzelner Beobachtungen . . . . .	223
8.2.3 Plots zur Entdeckung einflußreicher Beobachtungen . . . . .	224
8.2.4 Standard-Scatterplots . . . . .	226
8.2.4.1 Scatterplots aller Variablen . . . . .	227
8.2.4.2 Plot der beobachteten Werte gegen die vorhergesagten Werte . . . . .	228
8.2.4.3 Scatterplots der Residuen gegen die unabhängigen Variablen . . . . .	229
8.2.4.4 Scatterplots der Residuen gegen andere unabhängige Variablen . . . . .	229
8.2.4.5 Scatterplot der Residuen gegen die vorhergesagten Werte . . . . .	231
8.2.4.6 Plots zur Prüfung der Normalverteilung der Residuen . . . . .	233
8.2.5 Spezielle Residuenplots . . . . .	235
8.2.5.1 Partielle-Regressions-Plots . . . . .	235

8.2.5.2 Partial-Residual-Plots . . . . .	238
8.2.5.3 CUSUM-Plots . . . . .	239
8.2.5.4 Plots der seriellen Autokorrelation . . . . .	242
8.3 Bedeutung von Plots in der Regressionsdiagnostik . . . . .	243
8.4 Spezielle Plots für lineare Regressionen . . . . .	244
8.4.1 Mallows $C_p$ -Plot . . . . .	244
8.4.2 Ridge-Trace . . . . .	246
8.5 Plots für logistische Regressionen . . . . .	250
8.5.1 Regressionsdiagnostik in logistischen Regressionsmodellen . . . . .	253
8.5.2 Beurteilung der tatsächlichen Leistungsfähigkeit des Modells . . . . .	258
8.5.3 Exkurs: Regressionsdiagnostische Kriterien in der logistischen Regression . . . . .	260
<b>9 Plots in der Varianzanalyse . . . . .</b>	<b>261</b>
9.1 Plots bei der Überprüfung der Annahmen der Varianzanalyse . . . . .	261
9.1.1 Überprüfung der Normalverteilungsannahme . . . . .	262
9.1.2 Varianzhomogenitätsannahme . . . . .	265
9.1.3 Residuen-Diagnostik . . . . .	268
9.1.4 Zusammenfassung . . . . .	270
9.2 Plots als Hilfe bei der Interpretation der Varianzanalyse . . . . .	271
9.2.1 Multiple Mittelwertvergleiche . . . . .	271
9.2.2 Zufälligkeit von Mittelwertdifferenzen . . . . .	274
9.2.3 Interaktionsplots . . . . .	276
9.2.4 Box-Dot-Plots zur Ergebnisdarstellung . . . . .	277
9.2.5 ANOVA-Effects-Plots . . . . .	279
9.2.6 Plot der Mittelwerte nach Faktorlevel . . . . .	283
9.2.7 R-F-Spreadplots . . . . .	284
9.2.8 Profil-Plots . . . . .	285
9.2.9 t-Plots . . . . .	287
9.2.10 Aggregierte Sequenzplots . . . . .	289
9.3 Schlußbemerkung . . . . .	290
<b>10 Plots in der Clusteranalyse . . . . .</b>	<b>291</b>
10.1 Symbolische Darstellung der Datenmatrix . . . . .	291
10.1.1 Manuelle Matrix-Permutation . . . . .	293
10.1.2 Algorithmen zur Matrix-Permutation . . . . .	295
10.2 Symbolische Darstellung der Distanzmatrix . . . . .	297
10.2.1 Shading . . . . .	298
10.2.2 Threshold-Plots . . . . .	299
10.2.3 Median-Distanzen-Plot . . . . .	301
10.3 Fusionsdiagramme . . . . .	303
10.3.1 Dendrogramme . . . . .	303
10.3.1.1 Übereinstimmung des Dendrogramms mit der Distanzmatrix . . . . .	305
10.3.1.2 Vergleich mehrerer Dendrogramme bei Sensitivitätsanalysen . . . . .	306
10.3.2 Icicle-Plots . . . . .	307
10.3.3 Loop-Plots . . . . .	308
10.4 Plots zur Darstellung der Clusterdistanzen . . . . .	309
10.4.1 Fusionsdistanz-Plots . . . . .	309
10.4.2 Cluster-Distanz-Plots . . . . .	310
10.4.3 Objekt-Distanz-Plots . . . . .	311
10.4.4 Silhouetten-Plots . . . . .	312
10.5 Cluster-Profilplots . . . . .	314
10.6 Projektionsplots der Cluster . . . . .	316
10.6.1 Hauptkomponentenplots . . . . .	316

10.6.2 Plots der Diskriminanzfunktion . . . . .	317
10.6.3 Multidimensionale Skalierung und Clusteranalyse . . . . .	318
10.6.4 Projektionsüberprüfungen . . . . .	321
10.6.4.1 Plot der Distanzen gegen die Plotdistanzen . . . . .	322
10.6.4.2 Minimum-Spanning-Trees . . . . .	323
10.6.5 Varianten der Projektionsplots . . . . .	325
10.7 Schlußbemerkung . . . . .	326
<b>11 Datenanalyse-Konzeptionen . . . . .</b>	<b>327</b>
11.1 Explorative Datenanalyse . . . . .	327
11.2 "Explorative" versus "konfirmatorische" Datenanalyse . . . . .	328
11.3 Multivariate graphische Verfahren und "induktives Vorgehen" . . . . .	330
11.4 Einfache und komplexe statistische Analyse . . . . .	333
11.5 Datenanalyse statt der Analyse gegebener Zahlen . . . . .	336
11.6 Datenanalyse statt Statistik: Zur Kritik der Signifikanztests . . . . .	338
11.7 Schlußbemerkung . . . . .	342
<b>Anhang: Existierende Software und Hilfsmittel für eigene Programme . . . . .</b>	<b>343</b>
<b>Literatur . . . . .</b>	<b>347</b>
<b>Index . . . . .</b>	<b>364</b>