

# Clusteranalyse

Anwendungsorientierte Einführung

Von

Dr. Johann Bacher

R. Oldenbourg Verlag München Wien

---

# INHALTSVERZEICHNIS

---

Vorwort .....	XI
1 Einleitung .....	1
1.1 Primäre Zielsetzung clusteranalytischer Verfahren.....	1
1.2 Homogenität als Grundprinzip der Bildung von Clustern.....	2
1.3 Zuordnungsprinzipien: Unvollständige, deterministische und probabilistische Clusteranalyseverfahren .....	4
1.4 Objektorientierte und variablenorientierte Datenanalyse .....	6
1.5 Ziel und Richtung der Datenanalyse .....	9
1.6 Clusteranalyseverfahren als Datenmodelle.....	16
1.7 Modellprüfung.....	17
1.8 Fehleranalyse.....	19
1.9 Datenanalyse als iterativer Prozeß.....	22
1.10 Aufbau und Zielsetzungen der Arbeit .....	24
1.11 Computerprogramme.....	25
2 Unvollständige Clusteranalyseverfahren .....	27
2.1 Übersicht.....	27
2.2 Die multiple Korrespondenzanalyse .....	33
2.2.1 Ein Anwendungsbeispiel.....	33
2.2.1.1 Faktorenanalytische Interpretation.....	36
2.2.1.2 Clusteranalytische Interpretation .....	44
2.2.1.3 Illustrationsbeispiel für die Darstellung des Kalküls .....	48
2.2.2 Das Modell der multiplen Korrespondenzanalyse .....	49
2.2.2.1 Berechnung der empirischen Zusammenhangsmatrix $G$ .....	50
2.2.2.2 Berechnung der Eigenwerte, Faktorladungen und Koordinatenwerte .....	51
2.2.2.3 Berechnung der Skalenwerte und Interpretation der Koordinaten.....	55
2.2.2.4 Unerwünschter Effekt der Reskalierung der Faktorladungen.....	60
2.2.2.5 Rotation der Faktoren? .....	62
2.2.2.6 Berechnen von Distanzen zwischen den Ausprägungen.....	62

2.2.3	Modellprüfgrößen .....	63
2.2.3.1	Signifikanz der Zusammenhangsstruktur.....	63
2.2.3.2	Die Zahl maximal möglicher und bedeutsamer Dimensionen.....	64
2.2.3.3	Überprüfung der faktorenanalytischen Interpretation .....	65
2.2.3.4	Modellprüfgrößen für die clusteranalytische Interpretation .....	68
2.2.3.5	Schwellenwerte zur Interpretation.....	72
2.3	Nichtmetrische mehrdimensionale Skalierung.....	73
2.3.1	Aufgabenstellung.....	73
2.3.2	Schätzalgorithmus .....	76
2.3.3	Maximale Dimensionszahl .....	83
2.3.4	Scree-Test zur Bestimmung der Dimensionszahl.....	87
2.3.5	Unbekannter Metrikparameter $p$ .....	88
2.3.6	Weitere Modellanpassungsgrößen .....	88
2.3.7	Sozialstruktur und Freizeitverhalten von Kindern.....	90
2.3.7.1	Clusteranalytische Interpretation .....	92
2.3.7.2	Faktorenanalytische Interpretation.....	93
2.3.7.3	Freizeitaktivitäten und Sozialstruktur .....	96
2.4	Weitere räumliche Darstellungsverfahren .....	104
2.4.1	Die bivariate Korrespondenzanalyse .....	104
2.4.1.1	Modellansatz.....	104
2.4.1.2	Freizeitverhalten und Sozialstruktur .....	112
2.4.2	Nominale Faktorenanalyse nach Mc Donald.....	116
2.4.2.1	Der Modellansatz.....	116
2.4.2.2	Dimensionale Analyse der Freizeitaktivitäten von Kindern.....	120
2.4.3	Die Faktorenanalyse .....	122
2.4.3.1	Der Modellansatz.....	122
2.4.3.2	Die R-Faktorenanalyse.....	122
2.4.3.3	Die Q-Faktorenanalyse.....	132
3	Deterministische Clusteranalyseverfahren .....	141
3.1	Einleitende Übersicht.....	141
3.1.1	Überlappende und überlappungsfreie Clusterlösungen.....	141
3.1.2	Grundvorstellungen über die zu bildenden Cluster .....	142
3.1.3	Gemeinsame Algorithmen.....	144
3.1.4	Complete- und Single-Linkage als Basismodelle .....	144
3.1.5	Auswahl eines geeigneten Verfahrens .....	148
3.1.6	Lösungsschritte einer Klassifikationsaufgabe.....	150

3.1.7 Ein Anwendungsbeispiel.....	154
3.1.8 Fehlerquellen.....	163
3.1.9 Aufbau des Kapitels.....	173
3.2 Gewichtung und Transformation von Variablen.....	173
3.2.1 Vergleichbarkeit von Klassifikationsmerkmalen.....	173
3.2.2 Lösungsstrategien.....	175
3.2.3 Theoretische und empirische Standardisierung.....	175
3.2.4 Hierarchische Variablen.....	185
3.2.5 Gemischte Variablen.....	186
3.2.6 Standardisierung von Objekten.....	191
3.2.7 Exkurs: Die Problematik einer automatischen Orthogonalisierung.....	194
3.3 Auswahl eines Unähnlichkeits- oder Ähnlichkeitsmaßes.....	198
3.3.1 Dichotome Variablen.....	200
3.3.2 Nominale Variablen.....	210
3.3.3 Ordinale Variablen.....	213
3.3.4 Quantitative Variablen.....	221
3.3.5 A Priori-Prüfung auf Vorhandensein einer Clusterstruktur.....	226
3.3.6 Gewichtung von Variablen und Distanzen, Standardisierung von Objekten.....	228
3.3.7 Fehlende Werte.....	230
3.3.8 Exkurs: Quantifizierung und Konsequenzen der Kategorisierung.....	232
3.4 Nächste-Nachbarn-Verfahren und Mittelwertverfahren.....	238
3.4.1 Der Complete-Linkage als Basismodell.....	239
3.4.1.1 Der hierarchisch agglomerative Algorithmus.....	239
3.4.1.2 Hierarchische Darstellung von Ähnlichkeitsbeziehungen.....	243
3.4.1.3 Maßzahlen zur Bestimmung der Clusterzahl.....	247
3.4.1.4 Zufallstestung des Verschmelzungsschemas.....	250
3.4.1.5 Maßzahlen zur Beurteilung einer bestimmten Clusterlösung.....	253
3.4.2 Der Single-Linkage.....	257
3.4.3 Complete-Linkage für überlappende Cluster.....	261
3.4.4 Verallgemeinerte Nächste-Nachbarn-Verfahren.....	264
3.4.5 Mittelwertverfahren.....	270
3.5 Repräsentanten-Verfahren.....	279
3.5.1 Modellansatz.....	279
3.5.2 Der Algorithmus zur Clusterbildung.....	281
3.5.3 Maßzahlen der Modellanpassung.....	286
3.5.4 Die Wahl der Schwellenwerte.....	291

3.5.5 Ein weiteres Anwendungsbeispiel .....	291
3.5.6 Strategien zur Verwendung der Clusterzugehörigkeit für weitere Analysen.....	295
3.5.7 Weitere Repräsentanten-Verfahren .....	296
3.6 Hierarchische Verfahren zur Konstruktion von Clusterzentren .....	297
3.6.1 Modellansätze und Algorithmen .....	297
3.6.2 Modellprüfgrößen .....	301
3.6.3 Analyse großer Datensätze .....	302
3.7 K-Means-Verfahren.....	308
3.7.1 Modellansatz und Algorithmus .....	308
3.7.2 Modellprüfgrößen .....	316
3.7.2.1 Bestimmung der Clusterzahl.....	316
3.7.2.2 Maßzahlen zur Beurteilung einer bestimmten Clusterlösung.....	323
3.7.2.3 Zufallstestung einer bestimmten Clusterlösung .....	323
3.7.3 Beschreibung und Interpretation der Cluster .....	324
3.7.4 Stabilitäts- und Validitätsprüfung.....	336
3.7.5 Modifikationen des K-Means-Verfahrens.....	338
3.7.5.1 Startwertverfahren .....	339
3.7.5.1.1 Quick-Clustering-Verfahren .....	340
3.7.5.1.2 Stabilitätsprüfung durch Änderung des Startwertverfahrens .....	341
3.7.5.2 Modifikation des Algorithmus .....	344
3.7.5.3 Modifikation der Zuordnungsfunktion .....	345
3.7.6 Konfirmatorisches K-Means-Verfahren.....	348
4 Probabilistische Clusteranalyseverfahren.....	353
4.1 Einleitende Übersicht .....	353
4.2 Latente Profilanalyse .....	357
4.2.1 Modellansatz und Algorithmus .....	357
4.2.2 Modellprüfgrößen .....	365
4.2.2.1 Bestimmung der Klassenzahl .....	365
4.2.2.2 Modellprüfgrößen für eine bestimmte Klassenlösung .....	368
4.2.2.3 Zufallstestung einer Klassenlösung .....	368
4.2.3 Beschreibung und Interpretation einer Klassenlösung.....	369
4.2.4 Konfirmatorische latente Profilanalyse .....	375
4.3 Analyse latenter Klassen für nominalskalierte Variablen .....	379
4.3.1 Modellansatz und Algorithmus .....	379

---

4.3.2 Modellprüfung und Interpretation.....	385
4.3.3 Konfirmatorische Analyse.....	389
4.4 Analyse latenter Klassen für ordinalskalierte Variablen.....	392
4.4.1 Modellansatz und Schätzalgorithmus.....	392
4.4.2 Modellprüfgrößen.....	398
4.4.3 Konfirmatorische Analyse.....	402
4.5 Analyse latenter Klassen für gemischte Variablen.....	402
4.5.1 Modellansatz und Schätzalgorithmus.....	402
4.5.2 Modellprüfung und Interpretation.....	404
4.5.3 Konfirmatorische Analyse.....	407
5. Exkurs: Verwendung aller Variablen für eine Klassifikation?.....	409
Literaturverzeichnis.....	413
Sachregister.....	421