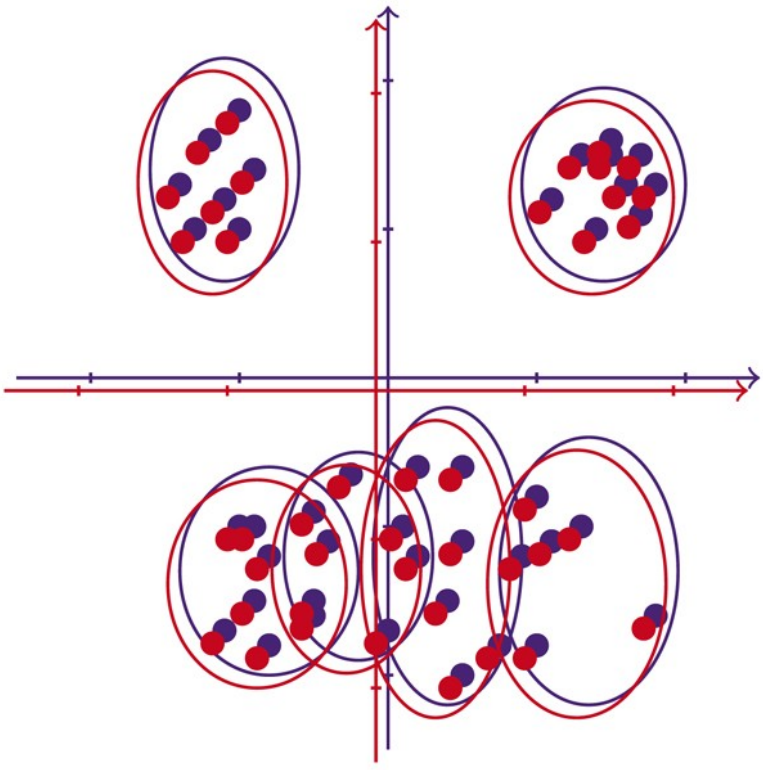
Johann Bacher | Andreas Pöge | Knut Wenzig

Clusteranalyse

Anwendungsorientierte Einführung
in Klassifikationsverfahren

3. Auflage



Oldenbourg

Clusteranalyse

Anwendungsorientierte Einführung in Klassifikationsverfahren

von

Universitätsprofessor
Dr. Johann Bacher
Johannes-Kepler-Universität, Linz

Akademischer Rat
Dr. Andreas Pöge
Universität Bielefeld

Diplom-Sozialwirt
Knut Wenzig
Nationales Bildungspanel, Bamberg

3., ergänzte, vollständig überarbeitete und neu gestaltete Auflage

Oldenbourg Verlag München

www.clusteranalyse.net
3.auflage@clusteranalyse.net

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
oldenbourg.de

Lektorat: Rainer Berger
Herstellung: Anna Grosser
Coverentwurf: Kochan & Partner, München
Umschlagillustration: Cornelia Horn
Gedruckt auf säure- und chlorfreiem Papier
Gesamtherstellung: Grafik + Druck GmbH, München

ISBN 978-3-486-58457-8

Vorwort zur dritten Auflage

Seit der Veröffentlichung der zweiten Auflage sind fast 15 Jahre vergangen, und im Bereich der Clusteranalyse hat es viele Weiterentwicklungen gegeben – die meisten sicherlich im Bereich der probabilistischen Verfahren. Im neuen Autorenteam wurden daher – neben der Umstellung auf LaTeX – die beiden vorausgehenden Auflagen grundlegend überarbeitet, um den aktuellen Entwicklungen Rechnung zu tragen. Beispielhaft genannt seien: Eine Erweiterung der Einleitung um eine Sammlung von Beispielen aus der Forschungspraxis, eine taxative Nennung und Beschreibung von Kriterien für eine gute Klassifikation sowie in Teil I eine klare begriffliche Abgrenzung von Hauptkomponentenmethode und Faktorenanalyse. Ausführlich behandelt werden Datenkonstellationen, für welche die Faktorenanalyse brauchbar oder unbrauchbar ist. Daneben wurde in Teil II das K-Means-Verfahren überarbeitet: Aufgenommen wurde die Methode der multiplen zufälligen Startwerte, die eine entscheidende Weiterentwicklung darstellt und das Problem lokaler Minima weitgehend vermeidet. Dargestellt werden auch Verallgemeinerungen des K-Means-Verfahrens, die andere Distanzfunktionen und Lageparameter nutzen.

Teil III wurde um den Latent-Gold-Ansatz, der die Modellierung von komplexen Clustermodellen ermöglicht, erweitert. Ergänzt wurde dieser Teil um eine Beschreibung von SPSS-TwoStep und AutoClass. Mit AutoClass wird in die derzeit »boomende« Bayes-Statistik eingeführt. Darüber hinaus wird die Perfomanz von ausgewählter Software verglichen. Im praktisch orientierten Teil IV werden häufig gestellte Anwenderfragen beantwortet und die Klassifikation von Verläufen mittels Optimal Matching, die Bildung von Konsensclustern sowie die formale Gültigkeitsprüfung dargestellt.

In allen Teilen wurden Anwendungsempfehlungen aufgenommen, wobei versucht wurde, einerseits den Praxisanforderungen Rechnung zu tragen und andererseits einfache »Kochbuchrezepte« zu vermeiden. Wie in den bisherigen Auflagen ist das Ziel, eine Darstellung dergestalt zu finden, dass die behandelten Verfahren in ihren Grundlagen von der interessierten Leserin oder dem interessierten Leser nachvollzogen werden können. Für eine wissenschaftliche Analyse erscheint es uns unumgänglich, dass zumindest eine grobe Vorstellung über die Funktionsweise eines Verfahrens vorliegt, bevor ein »Rechne-Knopf« am Computer gedrückt wird.

Über Rückmeldungen und Anregungen zu einzelnen Verfahren und Ausführungen freuen wir uns. Unser Dank gilt Arne Bethmann und Heinz Leitgöb für die Mitarbeit an einzelnen Teilen des Buches. Danken möchten wir darüber hinaus unseren Kolleginnen und Kollegen für Hinweise und Korrekturen sowie dem Oldenbourg-Verlag für seine Geduld bei der Abgabe der Druckvorlage. Die Fertigstellung des Manuskripts hat erhebliche Zeit zu Lasten von familiären Verpflichtungen gebunden. Für das Verständnis hierfür herzlichen Dank an alle.

Linz, Bielefeld, Bamberg 2010

Johann Bacher
Andreas Pöge
Knut Wenzig

Vorwort zur ersten und zweiten Auflage

Clusteranalyseverfahren gewinnen in der Forschungspraxis zunehmend an Bedeutung. Sie werden heute in zahlreichen Wissenschaftsdisziplinen zur Lösung von Klassifikationsaufgaben eingesetzt, in den Sozial- und Wirtschaftswissenschaften beispielsweise zur Identifizierung von unterschiedlichen Lebens- und Konsumstilen oder von Wertorientierungstypen.

Die Arbeit gibt – auf Grundlage der neueren Methodenliteratur – eine systematische Einführung, wie bei der Bestimmung von Typen (Clustern) vorzugehen ist, welche Verfahren dabei zur Verfügung stehen und wie die Ergebnisse zu interpretieren sind. Folgende Verfahren werden behandelt: bivariate und multiple Korrespondenzanalyse, nichtmetrische mehrdimensionale Skalierung, Faktorenanalyseverfahren (nominale Faktorenanalyse nach McDonald, R- und Q-Analyse), hierarchisch-agglomerative Verfahren, Repräsentanten-Verfahren, explorative und konfirmatorische K-Means-Verfahren, explorative und konfirmatorische probabilistische Clusteranalyseverfahren (latente Profilanalyse, Analyse latenter Klassen für nominale, ordinale und gemischte Variablen).

Alle Verfahren werden anhand von Beispielen aus der Forschungspraxis behandelt und durchgerechnet. Dabei werden eine Reihe von methodologischen Fragen beantwortet, wie beispielsweise die Frage, welche Konsequenzen eine empirische Standardisierung hat oder wie sich irrelevante Variablen auf die Ergebnisse einer Clusteranalyse auswirken.

Wie jede Arbeit wäre auch das hier vorliegende Buch nicht ohne die Unterstützung meiner akademischen Lehrer, des Oldenbourg-Verlages, der Teilnehmer an unterschiedlichen Kursen zur Datenanalyse sowie meiner Bekannten und Freunden und meiner Familie zustandegekommen. Ihnen allen möchte ich an dieser Stelle herzlichst danken.

Johann Bacher

Inhalt

1 Einleitung

1.1 Zielsetzung clusteranalytischer Verfahren

Primäres Ziel clusteranalytischer Auswertungsverfahren ist, eine Menge von Klassifikationsobjekten in homogene Gruppen (Klassen, Cluster, Typen) zusammenzufassen – oder kurz ausgedrückt – das *Auffinden einer empirischen Klassifikation* (Gruppeneinteilung, Typologie). Von einer empirischen Klassifikation soll dann gesprochen werden, wenn die Klassifikation auf empirischen Beobachtungen, zum Beispiel auf der Grundlage einer Befragung, basiert. Klassifikationsobjekte (siehe Übersichtstabelle 1.1) können sein: Individuen (Personen, Befragte), Aggregate (Organisationen, Nationen, Berufsgruppen usw.) oder Variablen (Merkmale).

Tab. 1.1: *Beispiele für clusteranalytische Auswertungen*

Klassifikationsobjekte	Beispiele für eine clusteranalytische Auswertung
Personen	A) Befragte werden aufgrund ihrer sozialstrukturellen Merkmale in (homogene) soziale Schichten zusammengefasst. B) Befragte werden aufgrund ihrer Lebensstile (Freizeitpräferenzen, Musikgeschmack, Wertorientierungen) in homogene Lebensstilgruppen zusammengefasst.
Aggregate	C) Nationen werden aufgrund ihrer demographischen, wirtschaftlichen und/oder sozialen Entwicklung in homogene Gruppen zusammengefasst. D) Berufe werden aufgrund ihrer Tätigkeitsprofile in homogene Gruppen zusammengefasst.
Variablen	E) Freizeitaktivitäten (Variablen) werden aufgrund ihres gemeinsamen Auftretens bei Personen in homogene Gruppen von Variablen zusammengefasst. F) Indikatoren der demographischen, wirtschaftlichen und sozialen Entwicklung werden aufgrund ihrer Korrelationen bei Nationen in homogene (Variablen-)Gruppen zusammengefasst.

Formal unterscheiden sich die sechs Beispiele der Übersicht darin, dass in den Beispielen A bis D die Zeilen einer Datenmatrix die Klassifikationsobjekte bilden. In den Beispielen E bis F sind dies dagegen die Spalten einer Datenmatrix. Im ersten Fall soll von einer *objektorientierten* Datenanalyse gesprochen werden, im letzten Fall von einer *variablenorientierten* Datenanalyse. Die Bezeichnung Klassifikationsobjekte kann sich auf die Spalten oder Zeilen einer Datenmatrix beziehen und bezeichnet jene Einheiten (Personen, Aggregate oder Variablen), die geclustert werden. Die untersuchten Objekte (Zeilen) müssen dabei keinesfalls mit den Erhebungseinheiten identisch sein. So zum Beispiel können durch Aggregierung über eine oder mehrere Variablen (zum Beispiel Beruf, regionale oder nationale Zugehörigkeit usw.) »neue« Objekte (Aggregate) erzeugt werden, die dann geclustert werden.

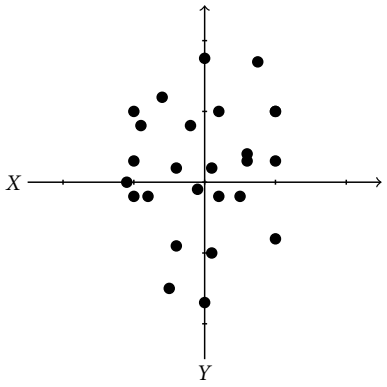
1.2 Homogenität als Grundprinzip der Bildung von Clustern

Jeder Clusterbildung liegt – unabhängig von Unterschieden im Detail – die Grundvorstellung der *Homogenität* bzw. von *»homogenen« Gruppen* zugrunde (Kozelka 1982, S. 6; Sodeur 1974, S. 118–124; u. a). Mit dem Begriff *homogene Gruppe* sind folgende Vorstellungen verbunden:

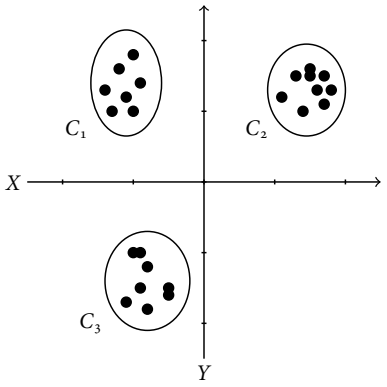
1. Die Klassifikationsobjekte, die einer homogenen Gruppe angehören, sollen untereinander ähnlich sein. Es soll *Homogenität innerhalb der Cluster* vorliegen.
2. Die Klassifikationsobjekte, die unterschiedlichen homogenen Gruppen angehören, sollen verschieden sein. Es soll *Heterogenität zwischen den Clustern* vorliegen.

Diese beiden Grundvorstellungen werden in der Literatur unterschiedlich bezeichnet. Cormack (1971) spricht von *externer Isolierung* (Heterogenität zwischen den Clustern) und *interner Kohäsion* (Homogenität innerhalb der Cluster). Everitt (1980, S. 60) und andere sprechen von *»natürlichen« Clustern*, wenn beide Forderungen erfüllt sind, und beschreiben Cluster als dichte Punktwolken in einem p-dimensionalen Raum, die durch Regionen mit einer geringen Dichte voneinander getrennt sind.[1] Sind die beiden Grundvorstellungen nicht erfüllt, ist es wenig sinnvoll, eine Klassifikation durchzuführen. Abbildung 1.1a verdeutlicht diesen Sachverhalt. Die untersuchten Klassifikationsobjekte bilden in den beiden Variablen X und Y eine große, relativ geschlossene Punktewolke. Eine Aussage der Art, den Daten liegen K Cluster (zum Beispiel $K = 3$) zugrunde, ist nicht sinnvoll. In der Abbildung 1.1b dagegen sind drei Cluster zu erkennen. Sie sind

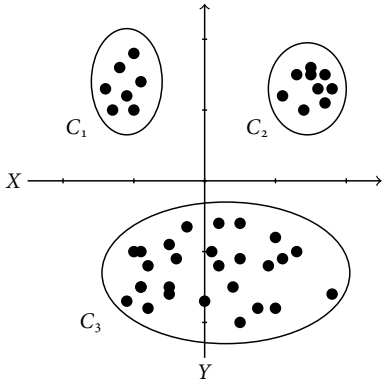
1 Zur Vorstellung von »natürlichen« Clustern siehe auch Aldenderfer und Blashfield (1984, S. 33–34) oder Vogel (1975, S. 16–17).



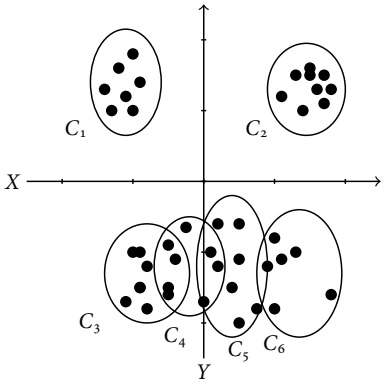
(a) Eine Clusterstruktur ist nicht erkennbar. Die Klassifikationsobjekte bilden eine große Punktewolke.



(b) Es sind drei Cluster erkennbar. Die Cluster sind homogen und voneinander gut getrennt.



(c) Es lassen sich drei Cluster erkennen. Cluster C_3 ist aber sehr lang gestreckt und wenig homogen.



(d) Im Unterschied zur Abbildung 1.1c wurde das lang gestreckte Cluster in vier überlappende Teilcluster zerlegt.

Abb. 1.1: *Datenkonstellationen mit erkennbarer und nicht erkennbarer Clusterstruktur*

in sich homogen und voneinander verschieden. In der Abbildung 1.1c ist zwar die Vorstellung der Heterogenität zwischen den Clustern erfüllt, Cluster 3 erfüllt aber in einem geringeren Ausmaß die Homogenitätsbedingung innerhalb der Cluster, da das Cluster sehr lang gestreckt ist. Abhängig von dem Gewicht beider Grundvorstellungen bei der Clusterbildung wird man sich entweder für drei Cluster entscheiden oder aus dem lang gestreckten Cluster mehrere überlappende Cluster bilden (siehe Abbildung 1.1d).

Neben diesen beiden Grundvorstellungen werden aus forschungspraktischen, aber auch aus inhaltlichen Gründen weitere Anforderungen an die gesuchte Klassifikation gestellt

(Schlosser 1976, S. 186; Sodeur 1974, S. 125–129), so zum Beispiel, dass die Zahl der Cluster möglichst klein sein sollte (Schlosser 1976, S. 186). Zusammenfassend lassen sich folgende Anforderungen bzw. Kriterien auflisten, die häufig von einer guten Clusterlösung verlangt werden:

1. Die Cluster sollen in sich homogen sein. Objekte, die einem Cluster angehören, sollen zueinander ähnlich sein.
2. Die Cluster sollen voneinander isoliert sein. Objekte, die verschiedenen Clustern angehören, sollen sich voneinander unterscheiden.
3. Die Cluster sollen den Daten gut angepasst sein. Die Klassifikation soll in der Lage sein, die Variation in den Daten zu erklären.
4. Die Cluster sollen stabil sein. Geringfügige Änderungen in den Daten oder im Verfahren sollen in keinen gravierenden Änderungen der Ergebnisse resultieren.
5. Die Cluster sollen inhaltlich gut interpretierbar sein. Den Clustern sollen inhaltlich sinnvolle Namen gegeben werden können. Im Idealfall sollen die Namen aus einer Theorie abgeleitet werden.
6. Die Cluster sollen (inhaltlich) valide sein. Die Cluster sollen mit externen Variablen korrelieren, von denen bekannt ist, dass sie im Zusammenhang mit den Typen stehen, die aber nicht in die Bildung der Cluster eingehen.

Gefordert wird mitunter ferner:

7. Die Zahl der Cluster soll klein und damit überschaubar sein. Angenommen wird, dass dies die inhaltliche Interpretierbarkeit (Kriterium 5) erleichtert und die Stabilität erhöht (Kriterium 4).
8. Die Cluster selbst sollen eine gewisse Mindestgröße haben. Dies soll zur Stabilität (Kriterium 4) beitragen.

1.3 Clusteranalyseverfahren

Zum Auffinden von Clustern wurde eine Vielzahl von Verfahren entwickelt, für die unterschiedliche Einteilungen und Zusammenfassungen vorgeschlagen wurden. In der vorliegenden Arbeit werden drei große Verfahrensgruppen unterschieden: *unvollständige Clusteranalyseverfahren*, *deterministische Clusteranalyseverfahren* und *probabilistische Clusteranalyseverfahren*. Grundlage der Differenzierung ist die Zuordnung der Klassifikationsobjekte zu den Clustern:

- *Unvollständige Clusteranalyseverfahren* (siehe Teil I): Diese Verfahren werden in der Literatur auch als geometrische Methoden (Gordon 1981, S. 80–120), als Repräsentations-

oder Projektionsverfahren (Jain und Dubes 1988; Opitz 1980) bezeichnet. In dieser Arbeit wurde die Bezeichnung »unvollständige Clusteranalyseverfahren« gewählt, da die Bildung von Clustern und die Zuordnung der Klassifikationsobjekte zu den Clustern von der Anwenderin bei der Interpretation der räumlichen Darstellung vorgenommen werden muss. Die unvollständigen Clusteranalyseverfahren selbst führen nur zu einer räumlichen Darstellung. Die Verfahren können auch dazu verwendet werden, abgeleitete Variablen (zum Beispiel Faktorwerte) zu bilden oder eine metrische Ähnlichkeits- oder Unähnlichkeitsmatrix (zum Beispiel mittels mehrdimensionaler Skalierung) zu schätzen, die anschließend mittels eines hierarchischen Clusteranalyseverfahrens untersucht wird.

- *Deterministische Clusteranalyseverfahren* (siehe Teil II): Die Klassifikationsobjekte werden mit einer Wahrscheinlichkeit von 1 einem oder mehreren Clustern zugeordnet. Es lassen sich zwei Verfahrensgruppen unterscheiden: *hierarchische Verfahren* und *partitionierende Verfahren*. Bei den hierarchischen Verfahren erfolgt die Clusterbildung schrittweise: Bei den sogenannten *hierarchisch-agglomerativen Verfahren* werden aus n Objekten zunächst n Cluster gebildet, aus den n Clustern durch Zusammenfassung der beiden ähnlichsten Cluster dann $n - 1$ Cluster, aus diesen wiederum $n - 2$ Cluster usw. Bei zehn Objekten werden also zunächst zehn Cluster gebildet, aus diesen zehn dann neun, aus den neun dann acht usw. Bei den *divisiven Verfahren* wird umgekehrt vorgegangen: Die n Objekte bilden ein einziges großes Cluster, dieses wird dann in zwei Cluster aufgespaltet, die zwei Cluster dann in drei usw. Bei den *partitionierenden Verfahren* muss dagegen eine Clusterzahl vorgegeben werden. Die Objekte werden dann den Clustern so zugeordnet, dass ein bestimmtes Kriterium maximiert bzw. minimiert wird Das bekannteste partionierende Verfahren ist das K-Means-Verfahren. Es ordnet die n Objekte K Clustern so zu, dass die Varianz in den Clustern minimiert wird.
- *Probabilistische Clusteranalyseverfahren* (siehe Teil III): Die Klassifikationsobjekte werden den Clustern nicht deterministisch mit einer Wahrscheinlichkeit von 1 oder 0, sondern mit einer dazwischen liegenden Wahrscheinlichkeit zugeordnet. Beispielsweise gehört ein Objekt g mit einer Wahrscheinlichkeit von 0,6 dem Cluster 1, mit einer Wahrscheinlichkeit von 0,2 dem Cluster 2 und mit einer Wahrscheinlichkeit von 0,1 dem Cluster 3 usw. an.

Die Beziehung zwischen unvollständigen, deterministischen und probabilistischen Clusteranalyseverfahren können wir uns wie folgt vorstellen: Eine graphische Bildung von Clustern, wie sie bei den unvollständigen Clusteranalyseverfahren durch den Anwender vorgenommen wird, ist nur in einem ein-, zwei- oder maximal dreidimensionalen Raum bei einer kleinen Klassifikationsobjektmenge möglich. Liegt ein höherdimensionaler Raum oder eine größere Klassifikationsobjektmenge vor, sind formale Verfahren zur Clusterbildung erforderlich. Hinsichtlich der Zuordnung der Klassifikationsobjek-