21 Clusteranalyse

Michael Wiedenbeck und Cornelia Züll

GESIS – Leibniz-Institut für Sozialwissenschaften, Mannheim

Zusammenfassung. Clusteranalyse ist ein Verfahren der numerischen Klassifikation für den Fall, dass die Klassen noch nicht (vollständig) bekannt sind und aus Daten erst konstruiert werden müssen. Das Fehlen eines generellen Daten- oder statistischen Modells als formales Gerüst für die Konstruktion von Klassifikationen führt zu einer inzwischen kaum mehr überschaubaren Anzahl von Verfahren zur Entdeckung einer Clusterstruktur. Der Erfolg der Anwendung hängt von der „richtigen" Kombination von Daten und Verfahren ab, die aber – außer bei Simulationsdaten – genau so wenig bekannt ist wie die Clusterstruktur selbst.

Im Folgenden behandeln wir zwei Verfahrensklassen, die dem Anwender seit langem in allen großen Statistikpaketen zur Verfügung stehen: agglomerative hierarchische Verfahren und K-Means. Erstere setzen die Wahl von geeigneten numerischen Differenzmaßen und deren Erweiterung auf Aggregate von Einzelbeobachtungen voraus. Sukzessiv werden Einzelbeobachtungen zu Gruppen, und Gruppen zu größeren Gruppen bis zum Erreichen der Gesamtstichprobe zusammengefasst. Die Anzahl möglicher Cluster muss aus der Abfolge der Differenzmaße nach einem „Ellenbogenkriterium" erschlossen werden. Bei K-Means wird die Anzahl der Cluster vorausgesetzt. Partitionen der Stichprobe werden nach einem Heterogenitätsindex bewertet, der die Homogenität der Cluster und ihre Differenz voneinander misst, und eine Startpartition durch einen Austauschalgorithmus in eine Konfiguration überführt, die dem Minimaldistanzkriterium genügt. Unter allen Partitionen mit dieser Eigenschaft befindet sich diejenige mit minimaler Heterogenität. Abschließend stellen wir das TwoStep-Verfahren (SPSS) dar, das eine Verallgemeinerung der agglomerativ-hierarchischen Verfahren zur Verarbeitung extrem großer Stichprobenumfänge ist.

1 Einführung in das Verfahren

Clusteranalyse ist ein Verfahren der *Mustererkennung* (pattern recognition). Ziel ist die Konstruktion von Typologien anhand von Stichproben von multivariaten Beobachtungen. Der Ansatz der Clusteranalyse setzt voraus, dass diese Stichprobe eine bestimmte Gruppenstruktur aufweist. Diese Struktur ermöglicht, dass sich die Stichprobe in eine Anzahl von Substichproben, so genannte Cluster, aufteilen lässt, deren Einheiten innerhalb der Cluster deutlich größere Ähnlichkeit untereinander besitzen als zwischen verschiedenen Clustern. Ist diese Clustereigenschaft von Substichproben noch ungeklärt, so sprechen wir von Aggregaten.

Die hier dargestellten Verfahren verwenden keine statistischen Modelle. Mit Ausnahme von sehr speziellen nicht-parametrischen Tests stehen daher auch keine statistischen Tests für die Prüfung von Hypothesen zu den Clusterstrukturen zur Verfügung.

S. 525–552 in: Christof Wolf & Henning Best, Hg. (2010). Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften

C. Wolf , H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse,*
DOI 10.1007/978-3-531-92038-2_21,

Basis der Verfahren sind numerische Maße für die paarweise Ähnlichkeit oder Differenz der multivariaten Profile der Einheiten. Auf dieser Grundlage werden die Einheiten in verschiedene Gruppen sortiert, die der o. g. Vorstellung maximaler Homogenität innerhalb der Gruppen und maximaler Heterogenität zwischen den Gruppen entsprechen sollen. Oder, bei einem anderen modus operandi, es werden bereits bestehende Gruppen sukzessiv durch Umsortierung einzelner Einheiten zu optimalen Partitionen verändert.

Die Sortierung der Einheiten wird durch Clusteralgorithmen geleistet, die in großer Vielfalt und Differenziertheit entwickelt wurden, und die zu durchaus unterschiedlichen Resultaten bei ein und derselbe Stichprobe führen. Zu einer Systematik der Algorithmen siehe Theodoridis & Koutroumbas (2003, S. 431 ff.). In den üblichen Statistik-Paketen findet sich davon nur eine relativ kleine Anzahl von Verfahren, über deren Parameter der Anwender allerdings vorab eine Reihe von Entscheidungen treffen muss.

Sind alle Variablen quantitativ und ist die genaue oder ungefähre Anzahl der Cluster einer gesuchten Struktur bekannt, dann ist mit dem K-Means Verfahrens eine direkte Optimierung der Binnenhomogenität und der Zwischenheterogenität im obigen Sinne möglich, wenn man den Algorithmus mit einer plausiblen Partition starten kann. Gibt es aber zur Clusteranzahl keinerlei Informationen, so sind diese aus dem Clusterverfahren selbst abzuleiten. Dies geschieht mit den hierarchisch-agglomerativen Verfahren, bei denen – beginnend mit den einelementigen Aggregaten – sehr viele, sehr kleine (und sehr homogene) überschneidungsfreie Aggregate gebildet werden, die dann zu größeren, möglichst homogenen überschneidungsfreien Aggregaten zusammengefasst werden. Die Heterogenität dieser Aggregate wächst mit ihrem Umfang, bis sie bei der Vereinigung zu einem einzigen Aggregat, der Ausgangsstichprobe, maximal wird. Numerisch wird die wachsende Heterogenität in einer Folge von Kennwerten (Fusionswerte) ausgedrückt, an deren Verlauf sich nach einem „Ellenbogen"-Kriterium – ähnlich wie in der Faktorenanalyse oder in der MDS – eine Clusterzahl bzw. ein Intervall für die Clusteranzahl ablesen lässt (vgl. auch die Kapitel 15 und 17 in diesem Handbuch).

Ist ein geeignetes Differenzmaß gewählt, so laufen sowohl die Algorithmen der hierarchisch-agglomerativen Verfahren als auch K-Means bis zur vollständigen Sortierung aller Fälle durch. Dies führt entweder zur vollständigen Konstruktion einer hierarchischen Folge von Partitionen oder zu einer „optimalen" Partition mit vorgegebener Anzahl von Substichproben. Die Clusteranalyse bietet a priori keine formelle Regel für die Wahl der „richtigen" Clusterzahl. Scheinbare Ausnahmen sind einige heuristisch motivierte Regeln, die sich in Simulationen bewährt haben. Siehe hierzu z. B. Everitt et al. (2001, S.77/103).

Die Bestimmung der Clusteranzahl mit Hilfe der vom Algorithmus berechneten Kennwerte obliegt dem Anwender wie auch die Interpretation der in den gewählten Clustern zusammengefassten Einheiten als Variationen inhaltlich sinnvoller Typen.

1.1 Was ist Clusteranalyse und was sind überhaupt Cluster?

Clusteranalyse von Daten ist der systematische Versuch, Substichproben von untereinander ähnlichen Beobachtungen in einer Stichprobe zu finden, wobei sich diese

Substichproben als Gruppen möglichst deutlich voneinander unterscheiden sollen. Die Gruppen, auch Cluster genannt, sind also nach einem ersten Verständnis durch Homogenität der Beobachtungen innerhalb einer Gruppe und Heterogenität der Beobachtungen zwischen unterschiedlichen Gruppen charakterisiert. Für Clusteranalysen liegen in der Regel Stichproben von Beobachtungseinheiten mit einem einheitlichen Satz von Variablen vor. Die Daten, für die wir uns in dieser Darstellung interessieren, haben also die Form einer Rechtecksmatrix, in der die Zeilen die Beobachtungseinheiten und die Spalten die Variablen repräsentieren. Für Clusteranalysen spielt der Prozess, mit dem die Stichproben generiert werden, (zunächst) eine nachgeordnete Rolle. Hier stehen vielmehr Methoden und Algorithmen zur Sortierung von Beobachtungseinheiten im Vordergrund, die die Einheiten nach Maßgabe ihrer multivariaten Profile gegenseitig zuordnen, sukzessiv zu Gruppen zusammenfassen oder Gruppen von Einheiten in Untergruppen aufspalten, neue Gruppen durch Umordnung von Einheiten definieren etc. Clusteranalyse ist also eine Klasse von Verfahren für die Exploration, Deskription und Sortierung von Daten mit dem Ziel, Gruppenstrukturen im obigen Sinne zu finden.[1] Die hier betrachteten Algorithmen liefern als Resultate Partitionen oder Hierarchien von Partitionen zusammen mit Parametern der einzelnen Schritte der Algorithmen. Dem Anwender obliegt dann die Beurteilung, ob diese Resultate zusammen mit einer substantiellen Theorie zur Identifizierung von Clustern sinnvoll sind.

1.2 Clusterstrukturen

Im Idealfall (für die hier betrachteten Verfahren) zerfällt eine Stichprobe in eine Anzahl von homogenen Clustern, die sich voneinander klar unterscheiden. Es ist aber auch denkbar, dass es eine oder mehrere Gruppen von Beobachtungen gibt, die sich untereinander und vom Rest der Stichprobe deutlich unterscheiden und daher als Cluster anzusehen sind, ohne dass der Rest selbst eine Clusterstruktur besitzt. Eine andere Variante wäre beispielsweise, wenn sich (ein oder mehrere) Cluster in Subcluster aufspalten lassen, d. h. wenn ein Cluster homogen ist verglichen mit der Menge aller Beobachtungen außerhalb seiner selbst, als Substichprobe aber eine Substruktur von Clustern besitzt.

Diese verschiedenen Konfigurationen bezeichnen wir als Clusterstrukturen. Es geht bei der Clusteranalyse nicht allein um das Auffinden einzelner Cluster, sondern auch um die Bestimmung von Clusterstrukturen auf unterschiedlichen Ebenen, also beispielsweise der Bestimmung von Subclustern eines Clusters.

1.3 Algorithmen

Clusteranalyse ist ein Verfahren zur Entdeckung unbekannter Clusterstrukturen. Damit unterscheidet es sich grundsätzlich von Verfahren, bei denen die Gruppenzugehörigkeit

[1] Inzwischen werden auch bestimmte Verfahren der statistischen Modellierung dem Gebiet der Clusteranalyse zugerechnet, die wir an dieser Stelle jedoch nicht diskutieren. Einen breiten Überblick über die unterschiedlichen Formen der Clusteranalyse gibt Bacher (1996).

der Beobachtungen bekannt ist wie z. B. bei der Diskriminanzanalyse (vgl. Kapitel 20 in diesem Handbuch).[2] Bei den hier dargestellten Verfahren werden Algorithmen zur Sortierung einzelner Beobachtungen angewendet, die entweder durch sukzessives Zusammenfassen von Einzelbeobachtungen ein hierarchisches System von Substichproben konstruieren (agglomerative Verfahren) oder durch schrittweise Verbesserung von Partitionen, also überschneidungsfreien Zerlegungen der gegebenen Stichprobe, zu einer in einem bestimmten Sinn optimalen Partition gelangen („K-Means“).

Eine bestimmte Clusterstruktur wird im Allgemeinen nicht gleichmäßig gut von unterschiedlichen Algorithmen identifiziert. Umgekehrt setzt die Anwendung der Clusteranalyse nicht voraus, dass es in einer Stichprobe überhaupt so etwas wie eine Clusterstruktur gibt. In einzelnen Fällen lässt sich vielleicht begründen, ob ein bestimmtes Verfahren angemessen oder vielleicht sogar das einzig sinnvolle Verfahren für die Identifizierung einer bestimmten Clusterstruktur ist. Aber man kann bei der Wahl eines Verfahrens nicht auf Hilfsmittel wie Spezifikationstests oder andere auf einer Verteilungstheorie basierte Tests zurückgreifen. Clusteranalyse ist – zumindest im Sinn der hier betrachteten Verfahren – lediglich eine Klasse von Algorithmen zur Sortierung der Einzelbeobachtungen nach unterschiedlichen Kriterien, die in Form von Verfahrensparametern vom Anwender festzulegen sind. Alternative Parameter lassen sich nur mit Intuition und substanzwissenschaftlichen Überlegungen unter Beachtung vorläufiger Resultate, nicht aber nach (inferenz)statistischen Regeln auswählen.

1.4 Variablenräume

Die Clusteranalyse fasst einzelne Beobachtungen als geometrische Punkte in einem mehrdimensionalen Variablenraum auf und beschreibt ihre gegenseitige Lage durch Distanzen.

Die Auswahl der Variablen haben wir bisher stillschweigend vorausgesetzt. Diese Wahl ist aber zu Beginn der Analyse vom Anwender zu treffen. Rein technisch gesehen ist Clusteranalyse praktisch für jeden Satz von Variablen möglich, wobei eventuell alphanumerische Variablen numerisch codiert und nominal skalierte numerische Variablen in Indikatorvariablen („dummy-Variablen“) transformiert werden müssen.

Die technische Anwendbarkeit garantiert jedoch nicht, dass für jede Wahl von Variablen eine Struktur mit ausgeprägten und sinnvoll interpretierbaren Clustern existiert. Es kann etwa bei gegebenen Daten für einen Satz von Variablen eine bestimmte Clusterstruktur bestehen, in einem anderen Variablenraum dagegen eine andere bzw. eine Struktur ohne ausgeprägte Cluster. Das ist bei verschiedenen Variablensätzen aus unterschiedlichen inhaltlichen Bereichen nicht sonderlich überraschend. Es kann aber auch bei unterschiedlichen Variablen des gleichen inhaltlichen Bereichs auftreten. Die Bestimmung von Clustern erfordert also eine „glückliche“ oder eine mit theoretischen Argumenten gut begründete Wahl der Variablen. Manchmal ist auch das Ausprobieren unterschiedlicher sinnvoller Sätze von Variablen erforderlich, um einer Clusterstruktur auf die Spur zu kommen.

[2] Gelegentlich wird Clusteranalyse daher auch als ein Verfahren des „unobserved learning“, also der Mustererkennung ohne Vorgabe von Mustern, bezeichnet.

In bestimmten Situationen sind Clusterstrukturen allerdings auch mit viel Geschick nicht mit den hier vorgestellten Methoden zu identifizieren, weil jedes ihre Cluster sowohl durch spezifische Beobachtungen als auch durch spezifische Variablen definiert ist. Verfahren zur Identifizierung derartiger Strukturen werden unter dem Begriff „Bimodale Clusteranalyse“ zusammengefasst (siehe hier z. B. Eckes 1991). Sie sind nicht Teil der hier vorgestellten Verfahren. Diese setzen dagegen implizit voraus, dass alle Variablen in gleicher Weise für die Cluster von Bedeutung sind.

1.5 Agglomerative Verfahren

Ähnlichkeitsmaß, Distanz und Index

Agglomerative Verfahren setzen numerische Maße der Ähnlichkeit oder Unähnlichkeit[3] zwischen Paaren von Einzelbeobachtungen als Vergleichskriterien voraus. Andere Verfahren bauen auf einem *Index* auf (siehe dazu Kaufmann & Pape 1984, S. 403 ff.), d. h. einer Maßzahl für die globale Heterogenität einer Partition. Im Fall von Unähnlichkeits- oder Distanzmaßen arbeitet der Algorithmus nach dem folgenden Schema: Ausgehend von der feinsten Zerlegung der gegebenen Stichprobe in das System von einelementigen Teilmengen werden zunächst alle Einheiten paarweise miteinander verglichen, d. h. jedes Paar von Einheiten wird mit dem gewählten Distanzmaß bewertet. Anschließend werden die Paare ihrerseits verglichen und das Paar mit dem kleinsten Wert zu einer neuen Aggregat-Einheit bestehend aus zwei Einheiten zusammengefasst. In der ursprünglichen Partition werden also zwei Einheiten eliminiert und durch ein zweielementiges Aggregat ersetzt. Anschließend wird das Verfahren der Zusammenfassung von Einheiten bzw. Aggregaten analog fortgesetzt, wobei allerdings eine Definition für die Distanz zwischen einer Einheit und einem Aggregat bzw. zwischen zwei Aggregaten vorher festgelegt sein muss.

Die sukzessive Agglomeration setzt also Maße a) für die Distanz zwischen Einzelbeobachtungen und b) zwischen Aggregaten (von Einzelbeobachtungen) bzw. zwischen Aggregaten und Einzelbeobachtungen voraus. Der Anwender muss vor der Analyse eine Wahl zwischen verschiedenen Alternativen für beide Arten von Distanzen treffen. Es gibt a priori keine formalen oder numerischen Kriterien für gute oder sogar optimale Entscheidungen. Allerdings hängen die durch die Agglomeration konstruierten Systeme von Aggregaten teilweise extrem stark von den genannten Maßen ab.

Ist eine Wahl sowohl für die Distanz von Einzelbeobachtungen als auch für die Distanz zwischen Aggregaten getroffen, so wird in jedem Schritt des Verfahrens eine Partition durch Zusammenfassung von zwei Aggregaten der vorangehenden Partition zu einem neuen Aggregat erzeugt. Dazu wird die Matrix der Distanzen zwischen den Aggregaten der jeweils zuletzt konstruierten Partition berechnet und anschließend aus dem Paar der Aggregate ein neues Aggregat gebildet, die sich nach Maßgabe der gewählten Kriterien am ähnlichsten sind. Die Anzahl der Aggregate wird also um

[3] Wir diskutieren im Folgenden der Einfachheit halber nur Unähnlichkeitsmaße und sprechen hier auch von Distanzen. Ähnlichkeitsmaße können in Unähnlichkeitsmaße durch antitone Funktionen transformiert werden.

eins vermindert und der Algorithmus mit einer Neuberechnung der Ähnlichkeitmatrix fortgesetzt.

Das durch Agglomeration konstruierte System von Aggregaten ist ein hierarchisches System von Substichproben, d. h. zwei beliebige Substichproben sind entweder disjunkt, oder eine von den Substichproben ist in der anderen enthalten. Ziel der Clusteranalyse ist es nun, aus diesem System ein Subsystem von Aggregaten, nämliche eine Partition auszuwählen, die möglichst gut der eingangs beschriebenen Forderung nach möglichst großer Homogenität der Einzelbeobachtungen innerhalb der Aggregate und möglichst großer Heterogenität zwischen den Aggregaten entspricht. Aggregate einer Partition, die diesen Anforderungen hinreichend gut genügt, werden als *Cluster* bezeichnet.

Bei Verfahren, die auf einem *Index* aufbauen, d. h. einer Maßzahl für die globale Heterogenität einer Partition, wie beispielsweise beim Ward-Verfahren, verfährt der Algorithmus analog: Es werden sukzessiv Einzelbeobachtungen paarweise zu einem Aggregat und weiter Aggregate paarweise zu einem noch größeren Aggregat vereinigt, sodass auch hier eine Hierarchie von immer „gröberen“ Partitionen konstruiert wird. Aus einer bereits erzeugten Partition wird diejenige Partition durch Vereinigung zweier Aggregate gebildet, bei der nach Maßgabe des Index der geringste Heterogenitätszuwachs auftritt.

Fusionswerte und Dendrogramme

Die Entscheidung für die oben genannte Clusterlösung wird bei agglomerativen Verfahren mit Hilfe des Verlaufs der so genannten Fusionswerte getroffen. Unter einem Fusionswert versteht man die Distanz zwischen denjenigen Aggregaten, die bei einem Schritt des Algorithmus zusammengefasst werden. Für die meisten agglomerativen Verfahren ist die Folge der Fusionswerte monoton wachsend. Man spricht dann von der *Monotonieeigenschaft* des jeweiligen Verfahrens, die intuitiv der Vorstellung entspricht, dass bei der Agglomeration zunehmend heterogenere Aggregate gebildet werden. Stellen wir uns etwa den Idealfall einer Anzahl von Clustern vor, die einerseits sehr homogen sind, bei denen also innerhalb der Cluster die paarweisen Distanzen zwischen den Einzelbeobachtungen sehr klein sind, andererseits aber die paarweisen Distanzen zwischen Beobachtungen oberhalb eines relativ großen Schwellenwerts liegen. Dann wird bei allen üblichen Verfahren die Folge der Fusionswerte zunächst im Bereich „kleiner“ Werte verbleiben (auch wenn die Fusionswerte keine einfachen Funktionen von paarweisen Distanzen sind), und zwar im Verlauf des Algorithmus solange, bis die durch die Cluster definierte Partition durch den Algorithmus selbst generiert wird. Im nächsten Schritt muss dann ein bestimmtes Paar von Aggregaten zu einem neuen Aggregat vereinigt werden. Wegen der großen paarweisen Distanzen zwischen Beobachtungen in verschiedenen Clustern ist dann auch (in diesem Idealfall) das Minimum aller Distanzen zwischen den Clustern groß (verglichen mit den vorangehenden Fusionswerten). Mit anderen Worten: Die als Kurve aufgetragene Folge der Fusionswerte macht an der Stelle, an der nach der Aggregation von Einzelbeobachtungen und Aggregaten innerhalb von Clustern zum ersten Mal zwei *Cluster* zusammengefasst werden, einen „Sprung“.

Man wird also hoffen, dass sich in der Fusionswertekurve der durchgeführten Agglomeration ein derartiger Sprung zeigt: die Aggregate unmittelbar vor dem „Sprung“

werden dann als Cluster identifiziert.[4] Zugleich bedeutet ein solches Bild, dass die Stichprobe vollständig in eine Anzahl von Clustern zerfällt. Die Folge der Fusionswerte ist auch Teil der Information des so genannten Dendrogramms. Dabei handelt es sich um eine Graphik in Form eines „Baums“, von der sich ablesen lässt, welche Einzelbeobachtungen oder Aggregate bei der sukzessiven Agglomeration in welcher Reihenfolge und gemäß welchen Fusionswerten zusammengefasst werden (siehe Abbildung 2 auf Seite 544).

Auch im Dendrogramm lässt sich gegebenenfalls der oben angesprochene „Sprung“ in der Folge der Fusionswerte feststellen. Die dadurch entstehende Lücke („gap“) im Dendrogramm lässt eine einfache Identifizierung der Aggregate zu, die unmittelbar vor dem Sprung gebildet wurden, und die als Cluster interpretierbar sind.

Neben der Identifizierung der Clusterlösung „nach Augenmaß“ gibt es auch einige formale Kriterien, die jedoch nur in wenigen Programmen realisiert sind. Dazu gehören z. B. die in Stata implementieren Stop-Regeln (Everitt et al. 2001, S. 103) oder die Entscheidungsregel in TwoStep.

Da im Dendrogamm – im Prinzip – die gesamte Hierarchie der Partitionen ablesbar ist, lassen sich auch andere als die Clusterstrukturen erkennen, die sich wie oben beschrieben als Partition darstellen lassen, etwa wenn ein Cluster oder allgemeiner ein Aggregat eine Clustersubstruktur aufweist. Beispielsweise erkennt man in Abbildung 2 (S. 544), dass das ganz unten gelegene Cluster 3 in zwei Subcluster von annähernd gleicher Heterogenität zerfällt.

1.6 Wahl der Metriken und Agglomerationsverfahren

Die Durchführung einer Clusteranalyse erfordert neben der Auswahl eines Datensatzes zwei Entscheidungen: Wahl eines Abstands- bzw. eines Ähnlichkeitsmaßes zum Vergleich einzelner Beobachtungen sowie einer Definition für den Abstand bzw. die Ähnlichkeit zweier disjunkter Aggregate von Beobachtungen. Von beiden Entscheidungen kann das Resultat der Analyse, also die Konstruktion der Hierarchie der Aggregate, und damit auch die Identifizierbarkeit von Clustern sehr stark abhängen.

Leider bietet die hierarchisch-agglomerative Clusteranalyse in diesem möglicherweise entscheidenden Punkt zwar eine mitunter verwirrende Fülle von Alternativen, aber keine wirkliche Entscheidungshilfe. Dazu kommt, dass bei jeder Wahl von Abstandsmaß und Agglomerationsverfahren „etwas herauskommt“, d. h. es wird eine Hierarchie von Aggregaten konstruiert und zusammen mit der Folge der Fusionswerte zur Verfügung gestellt. Zeigt das Dendrogramm eine Lücke in den Fusionswerten zwischen Aggregaten, die vor und nach einer bestimmten Stufe des Algorithmus gebildet werden, bzw. weist der Fusionswerteverlauf an dieser Stufe einen „Sprung“ nach oben auf, dann scheint für den Anwender alles in Ordnung zu sein. Er kann die Partition an der Sprungstelle als Clusterlösung wählen und sich an die Interpretation machen.

Er wird in der Regel aber keinen Zusammenhang zwischen seiner Wahl des Abstands und dem Agglomerationsverfahren und dem Auftreten einer Sprungstelle – oder deren

[4] Dieser Idealfall trifft bei „realen“ Daten überwiegend nicht zu. Man wird dann die Aggregate in den Bereichen der Agglomerationsschritte betrachten, für die die Fusionswertekurve eine „beschleunigte“ Steigung zeigt, sich also deutlich nach oben krümmt (siehe Abbildung 1).

Fehlen – im Fusionswerteverlauf herstellen können. Er kann natürlich verschiedene Wahlen treffen, was zu empfehlen ist, und dann unterschiedliche Lösungen sowohl hinsichtlich der Anzahl und der Zusammensetzung der Cluster als auch ihrer Homogenität vergleichen. Das Ausmaß von Übereinstimmungen zwischen zwei Lösungen kann deskriptiv durch Kreuztabellen darstellt werden. In ähnlicher Weise können auch variablenspezifische Varianzen als Indikatoren der Heterogenität zwischen unterschiedlichen Lösungen verglichen werden.

Die generelle Frage, für welche Art von Daten welches Abstandsmaß und welches Agglomerationsverfahren zu wählen ist, ist bisher nicht beantwortet worden. Theoretische Untersuchungen als auch Monte-Carlo-Studien haben nicht zu schlüssigen Regeln geführt (vgl. Everitt et al. 2001, S. 52 ff., 56 ff. und 89). Als positive Standardempfehlungen kann man die Wahl von Single Linkage – wegen der Eigenschaft der Kettenbildung – insbesondere für die Ausreißeranalyse empfehlen. Für die Analyse von Clustern lässt sich Single Linkage nur verwenden, wenn es nicht auf eine generelle Homogenität der Cluster ankommt, sondern, wie etwa in manchen sozialen Netzwerken, auf die Zugehörigkeit zum Cluster infolge indirekter, über eine Kette vermittelter Beziehungen zu entfernter liegenden Einheiten.

Eine weitere Standardregel, die offenbar bereits weithin beachtet wird, ist die Präferenz für Incremental Sum of Squares (Ward) als Agglomerationsverfahren. Die Beliebtheit scheint an der polarisierenden Eigenschaft der quadrierten euklidischen Metrik zu liegen, die benachbarte Beobachtungen mit Abständen < 1 noch näher zusammenrücken lässt und Beobachtungen mit Abständen > 1 noch weiter voneinander entfernt. Weiter wird mit der Summe der quadrierten euklidischen Abstände ein Gesamtmaß für die Heterogenität einer ganzen Partition verwendet. Werden Ausreißer vorher eliminiert, so scheint dieses Verfahren im allgemeinen zu plausiblen Aufteilungen der Stichprobe in homogene Cluster zu gelangen, die durch K-Means, das das gleiche Heterogenitätsmaß verwendet, weiter verbessert werden können.

Weniger kritisch ist aus unserer Sicht die Wahl des Abstandsmaßes, da die meisten Metriken topologisch äquivalent sind, wenn es sich nicht gerade um Ultra-Metriken handelt (siehe 2.1). Dennoch können die Unterschiede zu unterschiedlichen Hierarchien führen, da die Rangordnung von Abständen für verschiedene Metriken unterschiedlich ist. In derartigen Situationen könnte man z. B. die Robustheit einer Wahl durch Anwendung anderer Metriken in weiteren Analysen und durch den Vergleich der Resultate prüfen.

1.7 K-Means (Clusterzentrenanalyse)

Die Grundidee der agglomerativen Verfahren ist die sukzessive Zusammenfassung der einander ähnlichsten Beobachtungseinheiten. Einmal zu Aggregaten zusammengefasste Einheiten werden im Verlauf der Agglomeration nicht mehr in verschiedene Aggregate umsortiert, sondern als ganze in nachfolgenden Schritten zu größeren Aggregaten vereinigt. Dadurch wird das oben beschriebene hierarchische System von Aggregaten erzeugt, die eine sich vergröbernde Folge von Partitionen der Stichprobe bilden.

K-Means optimiert dagegen eine gegebene Partition durch eine Folge von Umsortierungen von Einzelbeobachtungen von einem Aggregat in ein anderes. Die Anzahl der Aggregate bleibt unverändert.

Optimalitätskriterium ist ein Maß für die Heterogenität von Aggregaten und für Partitionen, nämlich die Summe der quadrierten Abstände der Einzelbeobachtungen (Euclidean Sum of Squares, ESS) von den multivariaten Mittelpunkten der Aggregate, zu denen sie jeweils gehören. Dieses Maß ist ein so genannter Index. Je kleiner dieser Index ist, desto homogener sind die Aggregate und desto besser lassen sie sich als Cluster interpretieren. Gesucht ist daher die Partition mit dem kleinsten Index, gegeben die Anzahl der Aggregate.

Der Algorithmus von K-Means sucht nun aber nicht unter der extrem großen Anzahl aller Partitionen mit einer vorgegebenen Anzahl von Aggregaten nach der Partition mit dem kleinsten Index-Wert – dieses Optimierungsproblem ist tatsächlich zu komplex –, sondern beginnend mit einer Startpartition nach einer Partition mit der „Minimum Distanz Eigenschaft“ (MDE). Diese Eigenschaft besagt, dass der Abstand jeder Einzelbeobachtung zum Mittelwert des Aggregats, dem sie angehört, kleiner (oder höchstens gleich) ist als die Abstände zu den Mittelwerten der übrigen Aggregate. Es kann gezeigt werden, dass die MDE eine notwendige Bedingung für eine Partition mit einem minimalen Wert des Index ist.

Erfüllt eine Beobachtung die Bedingung der MDE nicht, dann wird sie in das Aggregat desjenigen Mittelwerts sortiert, dem sie am nächsten liegt. Nach der Umsortierung stimmen die Mittelwerte der Startpartition nicht mehr mit denen der neu konstruierten Partition überein. Mit neu berechneten Mittelwerten werden die Daten dann erneut geprüft und umsortiert. Das Verfahren endet, wenn keine Umsortierungen mehr erforderlich sind. Die zuletzt konstruierte Partition besitzt dann die MDE.

Nun kann es aber mehr als eine Partition mit der MDE geben. Wenn dies zutrifft, dann konvergiert der obige Algorithmus gegen eine Partition, die sowohl von der Startpartition als auch von der Reihenfolge der im Datenfile angeordneten Beobachtungen abhängt. Um also sicher zu gehen, dass K-Means eine indexminimale Lösung erzeugt hat, muss man die Reihenfolge der Beobachtungen und die Startpartition variieren und die nach erneuter Anwendung von K-Means ermittelten Indexwerte vergleichen. Für K-Means gibt es noch die folgenden technischen Varianten:

1. Beim „running means“ werden neue Aggregatmittelwerte nicht erst nach einem vollständigen Durchlauf durch die Daten, sondern bereits nach jeder Umsortierung für die beiden betroffenen Aggregate berechnet. Dadurch wird das Verfahren etwas schneller, was nach unserer Erfahrung aber nicht besonders entscheidend ist.
2. Die Startpartition kann in Form von – frei konstruierten – Beobachtungen als artifizielle Clustermittelpunkte vorgegeben werden. Dies erleichtert die Suche nach unterschiedlichen MDE-Partitionen (die Implementierung von Startmittelwerten in SPSS ist beschrieben in Wiedenbeck & Züll 2001). Mit dem Programm ClustanGraphics (`http://www.clustan.com`) und dem Modul „FocalPoint“ können diese Versuchsrechnungen in großer Anzahl bequem durchführt und hinsichtlich der unterschiedlichen Ergebnisse verglichen werden.

Für K-Means werden häufig zwei Voraussetzungen angegeben: Erstens die Anzahl der Cluster muss von vornherein bekannt sein, und zweitens alle Variablen sind quantitativ. Letzteres heißt, dass zwischen den Einzelbeobachtungen die euklidische Distanz als Abstandsmaß definiert werden kann.

Die erste Voraussetzung kann allerdings abgeschwächt werden. Wenn eine exakte Zahl an Clustern nicht vorgegeben werden kann, dann sollte man das Verfahren für unterschiedliche Vorgaben durchführen, beginnend mit einer minimalen und endend mit einer maximalen Clusterzahl. Für jede Lösung sollte man dann die Werte des Kriteriums (wenn es mehrere MDE-Partitionen gibt, deren Minimum) vergleichen, am besten durch Anlage eines Line-Plots. Auch wenn die Vorgabe nicht mit der wahren Clusterzahl übereinstimmt, konvergiert K-Means zu einer Partition mit der MDE. Die Werte des Kriteriums steigen mit fallender Clusterzahl, und zwar sprunghaft für die Clusterzahl, bei der zum ersten Mal zwei deutlich unterscheidbare Cluster auftreten. Die Anzahl vor der Sprungstelle ist dann ein plausibler Wert für die Clusterzahl, und die zugehörige Partition kann weiter daraufhin untersucht werden, ob sie auch inhaltlich eine Typologie repräsentiert.

Wenn die Clusterzahl unbekannt ist, dann kann man auch vorab agglomerative Clusteranalysen berechnen, daraus Lösungen bestimmen, und diese Lösungen, die im Allgemeinen die MDE nicht besitzen, als Startlösungen von K-Means einsetzen und optimieren, bzw. dies auf ganze Bereiche von agglomerativ gewonnenen Partitionen mit aufeinander folgenden Werten von Clusterzahlen anwenden. Für agglomerative Lösungen nach dem Ward-Verfahren ist diese Vorgehensweise eine geradezu natürliche Ergänzung, da der Index bei Ward und das Kriterium von K-Means übereinstimmen.

1.8 TwoStep-Clusteranalyse

Abschließend stellen wir ein neueres agglomerativ-hierarchisches Verfahren, das TwoStep-Verfahren, vor, das in SPSS seit der Version 11.5 zur Verfügung steht. Die Hersteller nehmen in Anspruch, damit einige wichtige Probleme der angewandten Clusteranalyse in neuer Weise behandeln zu können. In der folgenden Darstellung lehnen wir uns stark an Bacher et al. (2004) an.

Mit dem TwoStep-Verfahren sind extrem große Datensätze analysierbar, d. h. z. B. Datensätze mit einer Anzahl von Einzelbeobachtungen in der Größenordnung 10^5. Mit dieser Kapazität wird die TwoStep-Clusteranalyse zu einem Verfahren, das für *data mining* eingesetzt werden kann. Diese Leistungsfähigkeit wird durch ein *vorgeschaltetes Präclusterverfahren* ermöglicht. In einer zweiten Stufe wird aus den Präclustern der ersten Stufe in einem hierarchisch-agglomerativen Verfahren ein hierarchisches Mengensystem von Präclustern gebildet, das auch eine Hierarchie der Ausgangsstichprobe ist.

Bei extrem großen Stichprobenumfängen sind Dendrogramme schlicht nicht mehr darstellbar und können somit auch keine Informationen bzgl. der Anzahl von Clustern liefern. Deswegen wird im SPSS-Modul `TwoStep` die Clusteranzahl geschätzt. Sie kann allerdings auch vorgegeben werden.

TwoStep sieht zwei Optionen für die Bestimmung der Distanzen von Einzelbeobachtungen und Aggregaten vor: Sind sämtliche Variablen kontinuierlich, d. h. intervall-

skaliert, dann kann die Distanz sowohl durch die euklidische Metrik als auch durch einen Index definiert werden, der entsprechend der Log-Likelihood unter einem bestimmten Verteilungsmodell gebildet wird. Enthalten die Clustervariablen auch kategoriale Variablen (oder bestehen sie ausschließlich aus kategorialen Variablen), dann ist nur das indexbasierte Abstandsmaß möglich.

2 Mathematisch-statistische Grundlagen

Die folgenden Abschnitte ergänzen die bisherige Beschreibung in einigen formalen Details, diskutieren die mathematischen Eigenschaften einzelner Verfahren und einige mögliche Konsequenzen für ihre Anwendung.

2.1 Hierarchisch-agglomerative Verfahren

Ähnlichkeitsmaße und Distanzen

Eine Clusteranalyse setzt die Definition von numerischen Ähnlichkeitsmaßen bzw. Distanzen zwischen den Beobachtungen der zu analysierenden Gesamtheit voraus. In der Mehrzahl der Analysen werden als Distanzmaße so genannte Metriken verwendet. Eine Metrik ist eine reelle Funktion d auf dem kartesischen Produkt $S \times S$ einer Menge S von Objekten (Beobachtungen) mit den folgenden Eigenschaften:

$$d(i,j) = d(j,i) \geq 0 \qquad \text{für alle } i,j \in S \tag{1a}$$

$$d(i,i) = 0 \qquad \text{für alle } i \in S \tag{1b}$$

$$d(i,j) \leq d(i,k) + d(k,j) \qquad \text{für alle } i,j,k \in S \tag{1c}$$

In Worten: Der Abstand von i nach j ist gleich dem Abstand von j nach i und nicht negativ. Der Abstand jedes Objekts zu sich selbst ist gleich 0, und der direkte Weg von i nach j ist kürzer als der Umweg über k.

Diese Bedingungen erfüllt z. B. die euklidische Metrik, die für multivariate Beobachtungen an Einheiten $i \in S$ mit kontinuierlichen Einzelvariablen $x_{ip}, i \in S, p = 1, \ldots, P$ wie folgt definiert ist:

$$d_{\text{Euklid}}(i,j) = \sqrt{\sum_{p=1}^{P} (x_{ip} - x_{jp})^2} \tag{2}$$

Die euklidische Metrik ist die Metrik, die für drei Variablen der anschaulichen Distanz im dreidimensionalen Raum entspricht. Die Möglichkeiten von Distanzdefinitionen zwischen multivariaten Beobachtungen mit kontinuierlichen Variablen sind damit noch längst nicht ausgeschöpft. Die Menge möglicher – und in den üblichen Softwarepaketen implementierten – Metriken ist (im mathematische Sinne) unendlich groß. Weitere Beispiele findet man in Bacher (1996), Kaufman & Rousseeuw (2005) und Everitt et al. (2001).

Tab. 1: Anzahl der 0/1-Kombinationen bei zwei Beobachtungen

		Beobachtung i		Summe
		1	0	
Beobachtung j	1	a	b	a+b
	0	c	d	c+d
Summe		a+c	b+d	a+b+c+d

Eine genauere vergleichende Diskussion der Eigenschaften der verschiedenen Metriken würde den Rahmen dieser Darstellung sprengen. Allerdings gilt für alle gebräuchlichen Metriken d die Ungleichung

$$k_d d_{\text{Euklid}} \leq d \leq K_d d_{\text{Euklid}} \tag{3}$$

für geeignete Konstanten k_d, $K_d \geq 0$. Die Abschätzung bedeutet eine gewisse Robustheit der Clusterverfahren gegenüber der Wahl von d (so dass für quantitative Variable nichts gegen die Wahl der euklidischen Metrik spricht): Das Verhältnis d/d_{Euklid} wird nach unten durch k_d und nach oben durch K_d beschränkt, „sehr kleine“ d-Abstände bedeuten auch „sehr kleine“ d_{Euklid}-Abstände. Analoges gilt für „sehr große“ Abstände bezüglich d und d_{Euklid}. Die Rangordnung zweier d-Abstände muss aber nicht notwendig die gleiche für die entsprechenden d_{Euklid}-Abstände sein: Aus $d(x_i,x_j) \leq d(x_k,x_i)$ folgt nicht notwendig $d_{\text{Euklid}}(x_i,x_j) \leq d_{\text{Euklid}}(x_k,x_l)$ und umgekehrt. Deswegen können sich die Ergebnisse einer Clusterung bei unterschiedlichen Metriken durchaus unterscheiden, es sei denn, es liegen sehr homogene Cluster vor, die zudem sehr stark voneinander separiert sind.

Wegen der polarisierenden Eigenschaft (siehe oben) wird anstelle der euklidischen Metrik mitunter auch ihr quadrierter Wert als Abstandsmaß verwendet, obwohl dieses Maß keine Metrik ist.

Eine genauere Überlegung wird erforderlich, wenn die Variablen diskret sind. Handelt es sich um ordinal-skalierte Variablen, so kann man ihre Ausprägungen durch die zugehörigen Rangstatistiken ersetzen und dann wie kontinuierliche Variable behandeln.

Kategoriale Variable müssen in mehrere binäre Variable (mit den Ausprägungen 1 für das Vorliegen einer Kategorie und sonst 0) umgewandelt werden, wenn Metriken zur Abstandsbestimmung herangezogen werden sollen.

Bei multivariaten Beobachtungen mit kategorialen Variablen sind auch eine Reihe von Ähnlichkeitsmaßen gebräuchlich, die auf der Auszählung von Übereinstimmungen zweier Beobachtungen in den verschiedenen Variablen beruhen. Betrachten wir als Beispiel ein Set von binären Variablen. Dann lassen sich die Übereinstimmungen und Nicht-Übereinstimmungen wie in Tabelle 1 dargestellt zusammenfassen.

Der einfache Übereinstimmungskoeffizient $s_{ij} = (a+d)/(a+b+c+d)$ setzt die Anzahl der Übereinstimmungen ins Verhältnis zu allen Kombinationen, die bei zwei Beobachtungen auftreten. Dies führt allerdings zu einer fälschlichen Anrechnung der 0-0-Kombinationen als Indikatoren der Ähnlichkeit, wenn die binären Variablen die Ersetzungen von mehrstufigen kategorialen Variablen sind, und die Mehrzahl der 0-0-Kombinationen

nur das gemeinsame Nichtzutreffen eines Merkmals bei beiden Beobachtungen bedeutet. In diesen Fällen ist ein angemessenes Ähnlichkeitsmaß der *Jaccard-Koeffizient* $s_{ij} = {}^{a}/_{(a+b+c)}$, bei dem nur die Übereinstimmungen in der „1“ ins Verhältnis zu allen anderen Kombinationen ohne die 0-0-Kombinationen gesetzt werden. Für eine ausführliche Darstellung der Ähnlichkeitsmaße sei auf Kaufman & Rousseeuw (2005) und Everitt et al. (2001) verwiesen.

Sollen in der Analyse sowohl kontinuierliche als auch diskrete Variablen als Clustervariablen verwendet werden, dann kann ein gemeinsames Distanzmaß oder ein gemeinsames Ähnlichkeitsmaß als eine gewichtete oder ungewichtete Summe aus den Maßen (Ähnlichkeit oder Distanz) der beiden Variablengruppen gebildet werden. Hier wird als Ähnlichkeitsmaß z. B. der Gower-Index verwendet (siehe Everitt et al. 2001, S. 43). Es ist dabei eine offene Frage, mit welchen relativen Gewichten für die beiden Gruppen die Summe gebildet werden sollte.

Gewichtung und Standardisierung von Variablen

Variablen mit unterschiedlichen Varianzen können die Konstruktion von Aggregaten teilweise sehr unterschiedlich beeinflussen. In manchen Fällen kann dies erwünscht sein, vielfach wird man aber bei Clusteranalysen einen größeren Einfluss einiger Variablen gegenüber anderen ausschließen wollen. Die Variablen sollten daher zu Beginn einer Clusteranalyse standardisiert werden. Zwei der häufigsten Möglichkeiten sind die z-Transformation, also die Angleichung aller Varianzen auf 1, und die Angleichung der Wertebereiche.

Diese Transformationen haben die Form von Gewichten. Gewichte können allgemein zur Verstärkung oder Minderung des Einflusses von Variablen verwendet werden. Die Clusteranalyse liefert keine Anhaltspunkte für Gewichtungen.

Erweiterung der Abstandsmaße auf Abstände zwischen Aggregaten

Mit dem ersten Schritt der Agglomeration sind Aggregate erzeugt worden, die zusammen mit den Einzelbeobachtungen weiter sortiert werden müssen. Dadurch wird für die Distanzmatrix die Neuberechung aller Abstände zwischen Einzelbeobachtungen und Aggregaten sowie auch im weiteren Verlauf der Aggregation zwischen Aggregaten notwendig. Für die Definition der neuen Aggregatabstände gibt es unterschiedliche Optionen:

Single Linkage (Nearest Neighbor): Der Abstand zwischen zwei Aggregaten wird als das Minimum aller Abstände zwischen zwei Beobachtungen aus je einem der Aggregate definiert. Dieser Abstand ist also die Länge der kürzesten Verbindung zwischen den Aggregaten. Die Aggregate können daher einen beträchtlichen Durchmesser erreichen, wenn sie als eine Kette von benachbarten Beobachtungen aufgebaut werden. Das Verfahren ist daher für die Identifizierung derartiger Cluster geeignet.

Complete Linkage (Furthest Neighbor): Hier wird als Aggregatabstand die größte Distanz zwischen zwei Beobachtungen aus je einem der beiden Aggregate definiert. Zwar wird damit die Kettenbildung wie bei Single Linkage vermieden, dafür besteht aber das Risiko, dass Teile von zwei Clustern zu einem Aggregat zusammengefasst

werden, der Algorithmus also „natürliche“ Cluster spaltet. Die Anwendung ist daher zweckmäßig, wenn man Cluster mit kleinen Durchmessern erwartet.

Between-Groups Linkage: Der Abstand ist gleich dem Mittelwert aller Distanzen von Inter-Cluster-Paaren von Beobachtungen. Dieses Kriterium stellt einen Kompromiss zwischen Single Linkage und Complete Linkage dar. Seine Verwendung unterstützt im Gegensatz zu den beiden vorhergehenden Verfahren im Prozess der Agglomeration eher die Homogenität bei der Bildung von Aggregaten.

Within-Groups Linkage: Die Definition ist ähnlich zu der von „Between-Groups Linkage“. Für den Mittelwert zwischen Paaren von Beobachtungen werden aber neben den Inter-Cluster-Paaren auch Intra-Cluster-Paare herangezogen. Die mit diesem Kriterium konstruierten Aggregate weisen tendenziell eine noch höhere Homogenität auf als bei Between-Groups Linkage.

Ward: Dieses Verfahren ist indexbasiert. Der Wert des Index für eine Partition ist gleich der Summe der quadrierten euklidischen Abstände der Beobachtungen von den (multivariaten) Mittelwerten der Aggregate $g = 1, \ldots, G : \sum_{g=1}^{G} \sum_{i=1}^{m_g} \sum_{p=1}^{P} (x_{ip,g} - \bar{x}_{p,g})^2$. Der Abstand zweier disjunkter Aggregate ist dann gleich der Differenz aus dem Index für die Partition, bei der beide Aggregate vereinigt sind, und dem Index für die ursprüngliche Partition. Das Ward-Kriterium führt im Vergleich zu anderen Fusionskriterien tendenziell zur Konstruktion von Aggregaten, deren Umfänge ausgeglichener sind.

Dendrogramm und Ultrametrik

Die hierarchisch-agglomerative Clusteranalyse beginnt mit einer Matrix von Distanzen zwischen Paaren von Einzelbeobachtungen oder im Falle indexbasierter Verfahren, wie dem Ward-Verfahren, mit den (eventuell standardisierten oder anderweitig transformierten) Daten der Analysestichprobe. Als Ergebnis liefert sie ein Dendrogramm, also die Folge von Fusionswerten zusammen mit der zugehörigen Hierarchie von Partitionen. Das Resultat besitzt für die oben genannten Aggregatabstände eine bemerkenswerte mathematische Eigenschaft: Es definiert für jede Matrix von Distanzen eine weitere Metrik. Für zwei Einzelbeobachtungen ist der Wert dieser Metrik gleich dem Fusionswert, mit dem die beiden Einzelbeobachtungen in einem Aggregat zusammengeführt werden. Die durch den Algorithmus erzeugte Metrik ist eine *Ultrametrik*. Es gilt für sie die so genannte *verschärfte Dreiecksungleichung*, d. h. in einem System aus drei Beobachtungen ist nicht nur die Summe der Weglängen eines Umwegs über den dritten Fall größer als die direkte Weglänge, sondern bereits mindestens einer der beiden Teilwege des Umwegs ist länger als der direkte Weg. Damit ergibt sich aus Formel (1c):

$$d(i,j) \leq \max\{d(i,k),d(k,j)\} \text{ für alle } i,j,k \in S \tag{4}$$

Wegen dieser eigentümlichen Geometrie sind die ursprünglichen Distanzen und die der Ultrametrik notwendigerweise unterschiedlich. Man könnte auch sagen, dass durch die Agglomeration die ursprüngliche geometrische Anordnung verzerrt wird (siehe dazu Everitt et al. 2001, S. 74 ff.).

Auch bei Ward findet eine bestimmte Verzerrung statt, auch wenn die zugehörige Ultrametrik in der Literatur als „raumerhaltend“ eingestuft wird. Diese Verzerrung wirkt sich etwa so aus, dass in einem Aggregat einer mit Ward konstruierten Partition eine Einzelbeobachtung einen kleineren Abstand zum Mittelwert eines anderen Clusters als zu dem des eigenen haben kann.

2.2 K-Means

Das Kriterium (Index) für K-Means ist die Summe der quadrierten euklidischen Abstände (*E*uclidean *S*um of *S*quares) der einzelnen Beobachtungen vom jeweiligen Aggregatmittelwert:

$$\text{ESS} = \sum_{g=1}^{G} \sum_{i=1}^{m_g} \sum_{p=1}^{P} (x_{ip,g} - \bar{x}_{p,g})^2 \tag{5}$$

Es stimmt mit dem Heterogenitätsindex des Ward-Verfahrens überein.

2.3 TwoStep-Verfahren

Wie oben ausgeführt, basiert die Clusterung nach dem TwoStep-Verfahren auf zwei getrennten Verfahrensstufen. Für eine ausführlichere Darstellung siehe Bacher et al. (2004).

Die beiden Stufen der Clusterung

Step 1 – Präclusterung: In der ersten Stufe des Verfahrens werden die Daten durch eine Prozedur in eine Baumstruktur transformiert, bei der den Knoten der verschiedenen Ebenen bestimmte Statistiken zugeordnet sind (cluster features). Man spricht deshalb von einem cluster feature tree (CFT). Eine detaillierte Beschreibung findet sich in Zhang et al. (1996) und in Chiu et al. (2001). Beim CFT handelt es sich um eine Art Reparametrisierung der Daten unter Berücksichtigung der Ähnlichkeitsverhältnisse zwischen den einzelnen Beobachtungen. Die Endknoten (leaf nodes) repräsentieren die Präcluster, d. h. homogene Cluster von relativ kleinem Umfang.

Step 2 – Agglomeration: In der zweiten Stufe, der Agglomerationsphase, bilden die durch die Endknoten repräsentierten Präcluster die kleinsten Einheiten. Gegenüber dem Umfang der ursprünglichen Stichprobe der einzelnen Beobachtungseinheiten ist die Stichprobe der Präcluster wesentlich kleiner. Mit Hilfe der cluster features lassen sich die Aggregatabstände zwischen den Präclustern unter beiden Optionen, Log-Likelihood oder euklidische Metrik, so berechnen als würde die Berechnung auf den Einzelbeobachtungen fußen. Die Agglomeration folgt dann einem Schema einer indexbasierten Fusion wie z. B. das Ward-Verfahren. Die Log-Likelihood-Option verwendet die folgende Formel für den Index ξ_i des Aggregats i:[5]

[5] Genau genommen handelt es sich bei ξ um einen Index mit negativem Vorzeichen.

$$\xi_i = -n_i \left(\sum_{j=1}^{p} \frac{1}{2} \log(\hat{\sigma}_{ij}^2 + \hat{\sigma}_j^2) - \sum_{k=1}^{q} \sum_{l=1}^{m_k} \hat{\pi}_{ikl} \log(\hat{\pi}_{ikl}) \right) \tag{6}$$

Hierbei sind p die Anzahl der kontinuierlichen und q die Anzahl der diskreten Variablen; n_i ist der Umfang des Aggregats i, $\hat{\sigma}_{ij}^2$ die geschätzte Varianz der kontinuierlichen Variablen j innerhalb des Aggregats i, $\hat{\sigma}_j^2$ die geschätzte Varianz von j in der gesamten Stichprobe, $\hat{\pi}_{ikl}$ die relative Häufigkeit der Kategorie l der Variablen k im Aggregat i und m_k die Anzahl der Ausprägungen der Variablen k.

Der Abstand zweier Aggregate i und s ist dann definiert als

$$d(i,s) = \xi_{i \cup s} - \xi_i - \xi s\,. \tag{7}$$

Zur Interpretation des Index beachte man, dass die Log-Likelihood von unabhängig multinomial-verteilten Variablen, $-n_i \sum_{k=1}^{q} \sum_{l=1}^{m_k} \hat{\pi}_{ikl} \log(\hat{\pi}_{ikl})$, gleich der Entropie der gemeinsamen Verteilung kategorialer Variablen ist, wenn diese stochastisch unabhängig voneinander sind. Wären außerdem die kontinuierlichen Variablen normal und unabhängig voneinander verteilt, so wäre $-n_i \sum_{j=1}^{p} 1/2 \log \hat{\sigma}_{ij}^2$ die Log-Likelihood ihrer gemeinsamen Verteilung unter der Voraussetzung, dass alle Mittelwerte gleich 0 sind. Mit anderen Worten: Würde in den Ausdrücken für ξ_i, ξ_s und $\xi_{i \cup s}$ auf den Term $\hat{\sigma}_j^2$ verzichtet, dann wäre $d(i,s) = \xi_i + \xi_s - \xi_{i \cup s}$ genau die Verminderung der Log-Likelihood für die gesamte Stichprobe, wenn die Aggregate i und s vereinigt würden (die obigen Unabhängigkeitsannahmen vorausgesetzt).

In der gegebenen Form, und da die Unabhängigkeitsvoraussetzungen im allgemeinen nicht gelten, kann ξ_i allerdings nur als ein deskriptives Maß für die Streuung der gemeinsamen Verteilung der kontinuierlichen und kategorialen Variablen gelten; $d(i,s)$ ist dann gleich dem Zuwachs dieses speziellen Streuungsmaßes bei Vereinigung der Aggregate i und s.[6] Als weiteres Argument für die Verwendung der Log-Likelihood wird auch auf die Robustheit des Verfahrens gegenüber der Verletzung der Unabhängigkeitsvoraussetzungen verwiesen (siehe Norusis 2009, S. 361 ff.).

Anzahl der Cluster

Die Clusteranzahl wird in TwoStep nach einer Entscheidungsregel bestimmt, die sich sowohl auf die Sequenz der Verhältnisse des BIC (Bayes Informationskriterium von Schwarz)[7] zwischen aufeinander folgenden Clusteranzahlen als auch auf die Verhältnisse zwischen den Distanzmaßen bezieht.

[6] Es ist daher auch nicht sinnvoll, die Verteilung der Clustervariablen auf die Unabhängigkeitseigenschaften zu testen, da es letztlich nur auf den deskriptiven Aspekt des Streuungsmaßes ankommt.

[7] Das BIC-Kriterium (Bayes Information Criterion) ist ein Kriterium zur Auswahl eines von mehreren parametrischen Modellen, die für die Analyse eines Datensatzes in Frage kommen. Die Formel lautet: $BIC = -2l + \nu \ln(n)$. Hier ist ν gleich der Anzahl der Parameter des für die Formulierung des Index angenommenen Modells, n der Stichprobenumfang und l gleich dem Wert der Log-Likelihood, berechnet für die Maximum-Likelihood-Schätzungen der Parameter. Sind für einen Datensatz zwei unterschiedliche Modelle geschätzt worden, so ist dasjenige mit dem kleineren BIC vorzuziehen.

Daneben ist aber auch die Vorgabe der Clusteranzahl durch den Anwender möglich. Für Details siehe Bacher et al. (2004).

Tabellen und Graphiken zur Beschreibung der Cluster

Neben deskriptiven Darstellungen der Verteilungen der Clustervariablen in Tabellen, die in dieser oder ähnlicher Form auch bei anderen Verfahren ausgegeben werden, sind bei TwoStep verschiedene Typen von Graphiken für die abschließende Beurteilung und Interpretation einer gewählten Lösung nützlich.

Die erste Graphik, die gezeigt wird, bewertet die „Qualität" der Clusterlösung mit einem Umrissmaß bezüglich Kohäsion und Separation. Dabei handelt es sich um den Silhouetten Koeffizienten von Rousseeuw (vgl. Kaufman & Rousseeuw 2005, S. 83 ff.). Je größer der zwischen -1 und 1 liegende Koeffizient ist, desto größer ist die Kohäsion und Separation der betrachteten Clusterstruktur. Wobei man von einer Clusterstruktur erst ab Werten $> 0{,}25$ spricht. Kaufman & Rousseeuw (2005, S. 88) geben in ihrer Monographie eine Tabelle mit Schwellenwerten für die Bewertung der Clusterstruktur an. Neben einer tabellarischen Darstellung der Cluster und der sie beschreibenden Variablen, kann für jede Variable die Verteilung in jedem der Cluster im Vergleich zur Verteilung im Datensatz insgesamt angezeigt werden (siehe Abbildung 8). Zudem ermöglicht SPSS den Aufruf weiterer Graphiken zu Clustervergleichen, mit Maßen für die Wichtigkeit einzelner Variablen für die Clusterbildung.

3 Beispiel

3.1 Daten- und Variablenauswahl

Im folgenden Beispiel gehen wir der Frage nach, ob es in der Bevölkerung in Bezug auf das Fernsehinteresse verschiedene Typen gibt. Wir erwarten unterschiedliche Interessengruppen bei der Auswahl von Fernsehsendungen. Die Datenbasis bildet der kumulierte ALLBUS. Wir beschränken unsere Analysen auf das Jahr 2004. Dort wurden Fragen zum Interesse an bestimmten Arten von Fernsehsendungen gestellt:

„Ich habe hier Kärtchen, auf denen verschiedene Fernsehsendungen stehen. Bitte sagen Sie mir jeweils, wie stark Sie sich für solche Sendungen interessieren." Gefragt wurde dabei nach „Fernsehshows und Quizsendungen", „Sportsendungen", „Spielfilmen", „Nachrichten", „politischen Magazinen", „Kunst- und Kultursendungen", „Heimatfilmen", „Krimis", „Actionfilmen" und „Unterhaltungsserien".

Die daraus resultierenden Variablen (V385 bis V394) bilden die Basis für unsere Analyse der Interessenstypen. Die Variablen sind codiert als 1 (sehr stark), 2 (stark), 3 (mittel), 4 (wenig), 5 (überhaupt nicht), 0 (TNZ) und 9 (KA). Die beiden letzteren sind als fehlende Werte deklariert.

Nach der Festlegung auf diese Variablen müssen bezüglich der Daten einige Voraussetzungen überprüft werden. Ein Problem, das bei jeder Clusteranalyse zu berücksichtigen ist, ist die Frage von fehlenden Werten. Die hierarchischen Verfahren erlauben in der Regel keine fehlenden Werte, da Abstände zwischen einem gültigen und einem fehlenden Wert nicht berechnet werden können. Dies führt dazu, dass sowohl

Tab. 2: Fusionswerte der letzten Agglomerationsstufen

Agglomeratsstufe	Fusionswerte
2885	21250,154
2886	21564,726
2887	21892,266
2888	22238,568
2889	22599,350
2890	22991,084
2891	23433,203
2892	23878,892
2893	24398,892
2894	24969,969
2895	25576,683
2896	26315,209
2897	27198,526
2898	**28459,961**
2899	**30576,011**
2900	33629,062
2901	37342,111

SPSS wie auch STATA die jeweilige Beobachtung vollständig ausschließen. Hier ist zu entscheiden, ob dieser Ausschluss akzeptabel für die weitere Analyse ist oder ob diese fehlenden Werte in irgendeiner Form ersetzt werden müssen (z. B. durch einen gültigen Wert oberhalb bzw. unterhalb des Wertebereichs der Variablen oder durch Imputation). Die für unsere Analyse verwendeten Allbus-Variablen enthalten in 44 Beobachtungen fehlende Werte: 41 Befragte sehen überhaupt nicht fern. Sie spielen daher bei der Frage nach Typen des Fernsehinteresses keine Rolle und können von der folgenden Analyse ausgeschlossen werden. Drei Befragte haben die Antwort auf alle Fragen nach dem Fernsehinteresse und auch alle weiteren Antworten des Fragebogens verweigert. Nur drei Befragte haben zu einzelnen Sendungen keine Antwort gegeben. Wir haben daher alle Befragten mit fehlenden Werten in den entsprechenden Fragen von der hierarchischen Clusteranalyse ohne Konsequenzen für das Ergebnis ausgeschlossen. Für die Analyse ist der Stichprobenumfang 2902 Beobachtungen.

Eine weitere Frage, die bei der Vorbereitung der Variablen für die Analyse gestellt werden muss, betrifft das Skalenniveau, das bei der Wahl des Ähnlichkeitsmaßes eine Rolle spielt. Wir machen von der üblichen Option Gebrauch, ordinale Variablenausprägungen durch Ihre Ränge zu ersetzen und dann als intervallskaliert zu behandeln. Ein Standardisieren der Variablen ist nicht notwendig, da alle Variablen mit derselben Skala erfasst wurden.

3.2 Analyse

Da wir keine (fundierte theoretische) Kenntnis über die Zahl der zu erwartenden Typen haben, führen wir im ersten Schritt eine hierarchische Clusteranalyse durch.

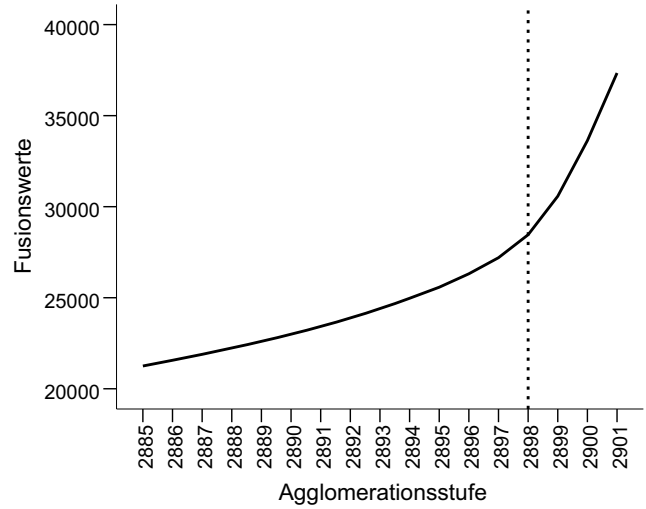

Abb. 1: Line-Plot der Fusionswertekurve

Als Clustermethode verwenden wir Ward und als Ähnlichkeitsmaß die quadrierte euklidische Distanz. Einen ersten Eindruck der Clusterstruktur liefert ein Blick auf die Agglomerationsdaten. In Tabelle 2 sind die Fusionswerte der letzten Agglomerationsstufen zusammengestellt. Auf den letzten Stufen kann man in der Tabelle einen verstärkten Anstieg der Werte erkennen, das heißt hier ist eine Clusterlösung zu verorten. In der Darstellung der Fusionswerte in einem Line-Plot wird dieser Anstieg deutlicher visualisiert (Abbildung 1). Auf Grund dieser Darstellung wird man eine 3- oder 4-Clusterlösung favorisieren.

Da die Dendrogramme in SPSS bei größeren Fallzahlen nicht lesbar sind, zeigen wird neben dem SPSS-Output auch das mit ClustanGraphics erstellte Dendrogramm. In ClustanGraphics kann das Dendrogramm auf die letzten Stufen der Agglomeration verkürzt werden, wie es auch in Stata möglich ist. Das Ergebnis der Analyse der Interessengruppen ist in Abbildung 2 dargestellt. Das Dendrogramm legt die Interpretation von drei Typen nahe.

Zur Illustration werden zusätzlich zum Dendrogramm die Differenzen der Clustermittelpunkte vom Gesamtmittelwert der verschiedenen Variablen in den Clustern als Line-Plot dargestellt (Abbildung 3). Eine hierarchische Clusteranalyse liefert häufig Lösungen auf verschiedenen Hierarchiestufen. In unserem Beispiel wäre auch die Interpretation einer 4-er Lösung denkbar: Das mittlere Cluster mit einem relativ hohen Fusionswert würde dabei in zwei Subcluster zerfallen. Im Folgenden werden wir aber zunächst die 3-Clusterlösung behandeln.

Die drei Cluster können wie folgt beschrieben werden: Cluster 1 (789 Beobachtungen) kann als Gruppe der „politisch und kulturell Interessierten“ beschrieben werden. In dieser Gruppe besteht wenig Interesse an allen anderen Sendungen. Cluster 2 (1266 Beobachtungen) ist die Gruppe der „vielseitig Interessierten“. Besonderes Interesse gilt

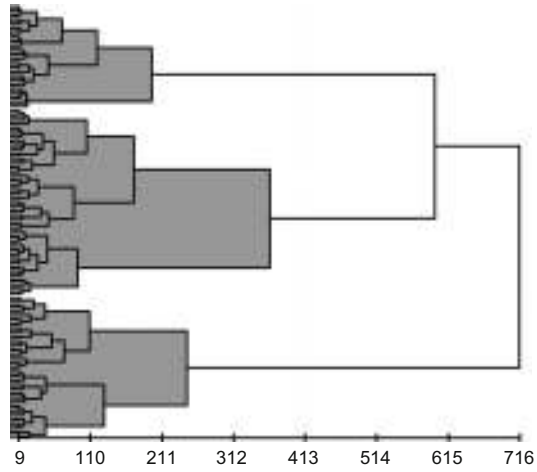

Abb. 2: Dendrogramm der Ward-Clusteranalyse

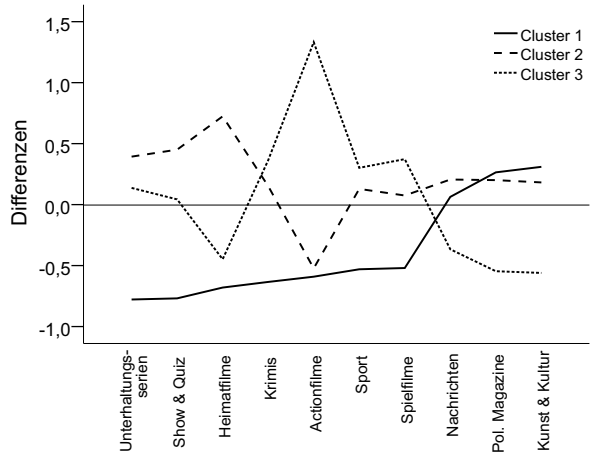

Abb. 3: Line-Plot der Differenzen der Clustermittelwerte vom Gesamtmittelwert

dabei den Heimatfilmen, Shows und Quiz und den Unterhaltungsserien. Actionfilme werden jedoch nicht angesehen. Die Personen in Cluster 3 (847 Beobachtungen) lieben „Spannung“: in erster Linie Actionfilme, aber auch Krimis und Spielfilme.

3.3 Verbesserung der Clusterlösung

Nach dieser ersten Datenexploration soll im nächsten Schritt versucht werden, die Clusterlösung von Abschnitt 3.2 zu optimieren. Dazu gibt es verschiedene Möglichkeiten.

Variablenauswahl

Durch Vergleich von Analyseresultaten mit alternativen Variablensätzen haben wir festgestellt, dass der oben beschriebene Variablensatz für die Interpretation der Clusterlösung gut geeignet ist.

Eliminieren von „Ausreißern“

Bei der Anwendung des Ward-Verfahrens spielen so genannte Ausreißer, d. h. Beobachtungen mit Extremwerten, eine große Rolle. Sie können die Konstruktion einzelner Cluster stark beeinflussen. Tendenziell werden bei Ward Ausreißer zusammen mit den ihnen am nächsten gelegenen Beobachtungen einem Cluster zugeordnet. Dadurch kann insbesondere der Mittelpunkt eines Clusters stark in Richtung der Ausreißer verschoben sein. Es empfiehlt sich daher, die Daten auf solche Problemfälle hin zu untersuchen. Die Clusteranalyse ermöglicht dies mit der Option „Nearest Neighbor“ (Single Linkage). Ausreißer sind diejenigen Beobachtungen, die wegen ihrer Extremwerte oder ihrer Wertekombination von allen anderen Beobachtungen einen auffallend großen Abstand besitzen. Da Single Linkage den Abstand zweier Aggregate als den minimalen Abstand zwischen zwei Beobachtungen der Aggregate definiert, werden Ausreißer erst gegen Ende der Agglomeration bereits konstruierten Aggregaten zugefügt und können so im Dendrogramm identifiziert werden.

Wir entscheiden uns für einen Schnitt bei der 2850-ten Stufe mit einem Fusionswert von 6,00. Inklusive dieser Stufe sind 52 Aggregate konstruiert worden, von denen 51 Aggregate bei nachfolgenden Fusionen zu größeren Fusionswerten mit anderen Aggregaten vereinigt werden. Wenn man nun diese 51 Aggregate mit 54 Beobachtungen eliminiert, so verbleiben in der Stichprobe nur die Beobachtungen, die einen „Nachbarn“ in dieser Stichprobe besitzen, dessen Distanz den Fusionswert 6,00 nicht überschreitet. Die eliminierten Beobachtungen haben dagegen *in der reduzierten Stichprobe* keinen Nachbarn in einem Abstand unterhalb dieses Schwellenwerts, wenngleich es durchaus vorkommen kann (siehe Abbildung 4), dass in der Stichprobe der 54 eliminierten Beobachtungen ein Nachbar in einem Abstand unterhalb des Schwellenwerts existiert.

Betrachtet man die Charakteristika der 54 als Ausreißer identifizierten Beobachtungen, wird man keine besondere inhaltliche Orientierung dieser Gruppe feststellen. Es fällt aber auf, dass sie deutlich mehr Antworten in den Extrembereichen der Variablen haben, d. h. bei den Werten 1 (stark interessiert) und/oder 5 (überhaupt nicht interessiert).

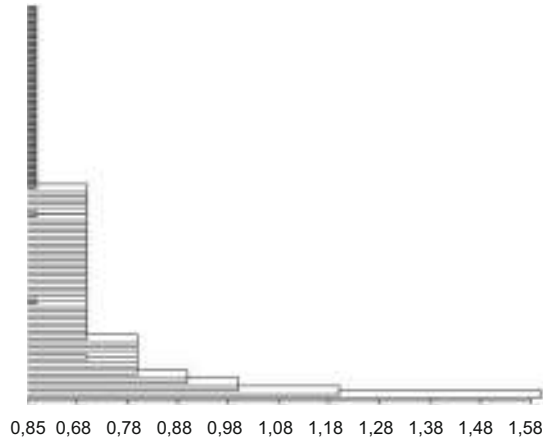

Abb. 4: Dendrogramm der Single Linkage-Lösung

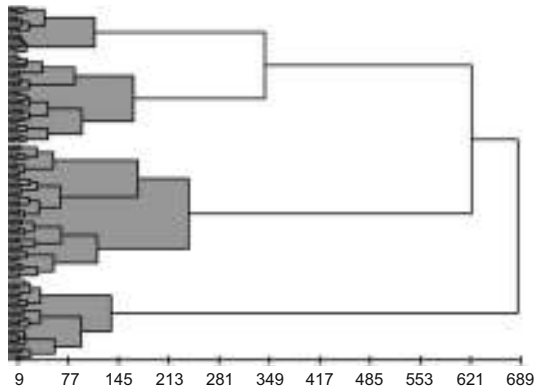

Abb. 5: Clusterlösung nach dem Entfernen der Ausreißer

Die weitere Analyse wird nun mit 2848 Beobachtungen wiederholt. Nun ergibt sich ein anderes Bild als bei der Lösung mit der vollständigen Stichprobe: Es lässt sich an Hand des Dendrogramms eine 4-Clusterlösung identifizieren, die wir an dieser Stelle nicht weiter darstellen, da sie weiter optimiert wird.

Optimierung der Lösung durch eine K-Means-Analyse

Die neue Lösung ohne die vorher ausgeschlossenen Ausreißer kann nun durch ein K-Means-Verfahren weiter optimiert werden. Ausgehend von den vier identifizierten Clustermittelpunkten werden die Daten neu ihrem am nächsten liegenden Clustermittelpunkt zugeordnet (basierend auf der Optimierung des ESS). Die durch das K-Means-Verfahren bereinigte Lösung ist in Abbildung 5 dargestellt.

Die neu gebildeten Cluster können wie folgt beschrieben werden (vergleiche Abbildung 6):

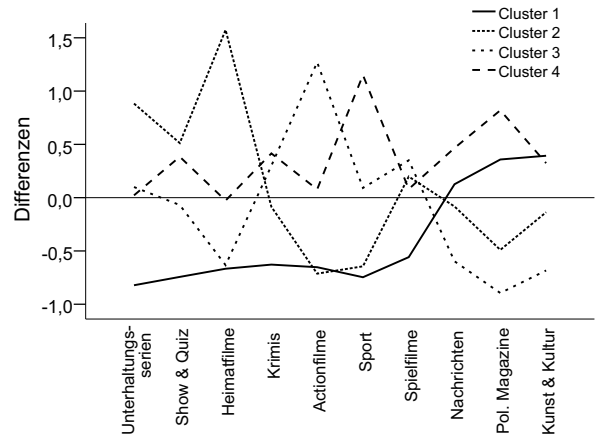

Abb. 6: Differenzen vom Mittelwert nach der K-Means-Analyse

Cluster 1 entspricht dem Cluster 1 der „politisch und kulturell Interessierten“ der ersten Clusterlösung mit nun 763 Beobachtungen. Cluster 2 kann beschrieben werden als die Gruppe, die sich besonders für „Heimatfilme, Shows und Quiz und Unterhaltungsserien“ interessiert (613 Beobachtungen): Spielfilme werden leicht überdurchschnittlich angesehen, Sport und politische Magazine interessieren wenig. Cluster 3 (688 Beobachtungen) entspricht Cluster 3 der ersten Lösung und kann wieder als Cluster der „Spannung-Liebenden“ beschrieben werden: Actionfilme, Krimis und Spielfilme stehen im Zentrum des Interesses. Heimatfilme, Politik und Kultur interessieren eher wenig. In Cluster 4 (784 Beobachtungen) sind die „Vielseitig Interessierten“. Das Hauptinteresse liegt bei Sportsendungen, aber auch in allen anderen Bereichen sind sie eher überdurchschnittlich interessiert. Wenn man die Lösung mit der vorhergehenden (unbereinigten) 3-Cluster vergleicht, fällt auf, dass das Cluster 2 der ersten Lösung „vielseitig Interessierte“ nun in zwei Cluster zerfällt (Cluster 2 und 4, vgl. Tabelle 3). Gleichzeitig werden 162 Beobachtungen aus dem alten Cluster 3 in das neue Cluster 4 übernommen. D. h., dass das Cluster der „vielseitig Interessierten“ nun klarer aufgeteilt wird in die „Heimatfilme-, Shows- und Quiz- und den Unterhaltungsserien-Interessierten“ und die „Sport-Interessierten“. Diese Aufteilung des Clusters 2 der ersten Clusterlösung hatte schon das erste Dendrogramm (Abbildung 2) als eine Option angezeigt.

3.4 Überprüfen der Clusterlösung

Zur Überprüfung der Clusterlösung stehen nur wenige technische Hilfsmittel zur Verfügung. Es gibt insbesondere keine festen Kenngrößen oder Fitmaße, die die Güte der Lösung angeben. Der Permutationstest in ClustanGraphics unterstützt zwar den Anwender bei der Beurteilung des Fusionswerteverlaufs und der Bestimmung einer Clusterzahl, aber letztendlich bleibt dem Anwender nur sein theoretisches Wissen über mögliche Clusterstrukturen, mit dem er die Ergebnisse validieren kann. K-Means in

Tab. 3: Vergleich der ersten (unbereinigten) mit der bereinigten Lösung

		3-Cluster-Lösung			
		1	2	3	Gesamt
4-Cluster-Lösung (Ausreißer-bereinigt) nach K-Means	1	**608**	122	33	763
	2	22	**532**	59	613
	3	76	42	**570**	688
	4	74	**548**	**162**	784
	Gesamt	780	1244	824	2848

SPSS ermöglicht die Ausgabe des Abstandes eines Falles vom Mittelpunkt des Clusters. Diese Angabe kann dabei helfen festzustellen, wie sich die Beobachtungen um den jeweiligen Clustermittelpunkt verteilen. Einen Eindruck über die Stabilität der Lösung vermittelt ein Vergleich der hierarchischen Clusterlösung mit der K-Means-Lösung: Verändern sich die Lösungen grundsätzlich oder werden nur einzelne Beobachtungen verschoben? Eine andere Möglichkeit bietet eine Überprüfung der Stabilität einer gewählten Lösung. Dazu werden die Daten in mehrere zufällige Stichproben zerlegt. Werden bei der Clusteranalyse der verschiedenen Teilstichproben jeweils ähnliche Clusterstrukturen identifiziert?

3.5 TwoStep-Clusteranalyse

Das oben gezeigte Beispiel einer Clusteranalyse wird nun mit dem neben der hierarchischen Clusteranalyse und dem K-Means-Verfahren in SPSS angebotenen TwoStep-Verfahren durchgeführt. Wir verwenden unsere Variablen auch in diesem Beispiel als intervallskalierte Variablen.

Obwohl die Zahl der Beobachtungen (2902) für dieses Verfahren eher klein ist, soll der Einsatz des Verfahrens an diesem Beispiel demonstriert werden. Beobachtungen mit fehlenden Werten auf einzelnen Variablen werden von TwoStep – wie in der hierarchischen Clusteranalyse auch - immer automatisch eliminiert. Im Gegensatz zu den oben vorgestellten Verfahren, wird bei der TwoStep-Analyse die Zahl der Cluster durch das BIC-Maß automatisch bestimmt. In unserem Beispiel wird eine 4-Clusterlösung ermittelt. Die 4-Clusterlösung kann mit Hilfe eines Line-Plots der Abweichungen der Mittelwerte der Variablen vom jeweiligen Gesamtmittelwert beschrieben werden (Abbildung 7):

- Cluster 1 (484 Beobachtungen) kann als „Desinteressierte“ beschrieben werden, die aber bei ihrem Interesse an Heimatfilmen und Unterhaltungsserien im Durchschnitt liegen.
- Cluster 2 (715 Beobachtungen) enthält die „breit Interessierten“. Sie interessieren sich für alle untersuchten Sendungen überdurchschnittlich. Besonders auffallend ist ihr Interesse an Heimatfilmen und Shows und Quizsendungen. Bei Actionfilmen liegen sie dagegen eher im Durchschnitt.

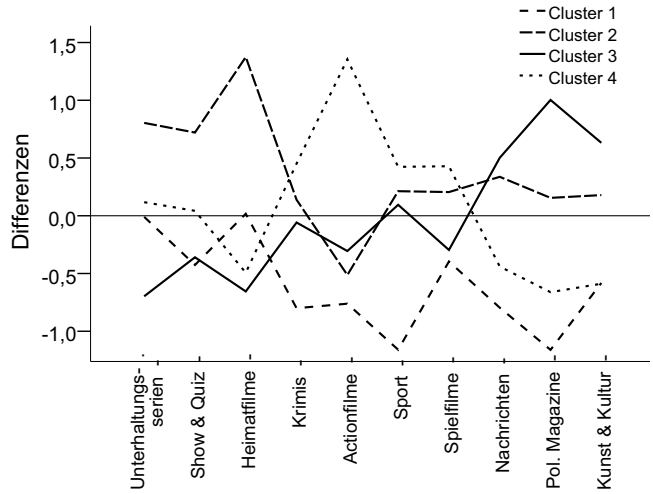

Abb. 7: Distanzen der Mittelwerte vom Gesamtmittelwert

- Cluster 3 (947 Beobachtungen) kann als Cluster der „Informationssuchenden“ beschrieben werden. Nachrichten, politische Magazine und Kunst- und Kultursendungen stehen im Mittelpunkt des Interesses. Bei Sportsendungen liegen sie im Durchschnitt. Alle anderen Sendungen interessieren unterdurchschnittlich wenig.
- Cluster 4 (756 Beobachtungen) enthält alle Beobachtungen der „Spannung Liebenden“ mit sehr großem Interesse an Actionfilmen. Daneben stehen Krimis, Spielfilme und Sportsendungen im Mittelpunkt des Interesses. Das Interesse an allen anderen Sendungen ist eher unterdurchschnittlich. Auffällig ist das große Desinteresse an Politik und Kultur.

Die TwoStep-Analyse in SPSS bietet Graphiken zur besseren Beurteilung der Clusterlösungen. Da TwoStep ein relativ neues Verfahren in SPSS ist, ist diese Prozedur und die Darstellung ihrer Resultate im Output immer noch im Umbruch. Der folgenden Beschreibung wird SPSS Version 18 zu Grunde gelegt. Die gesamte Ausgabe in dieser Version wird – im Gegensatz zu der sonst in SPSS üblichen Darstellung im SPSS Viewer – in einem sogenannten „Modell Viewer“, d. h. einer speziellen Hypertextstruktur, abgebildet. So kann der Nutzer seine Clusteranalyse durch verschiedene Visualisierungen überprüfen.

Beispielsweise zeigt Abbildung 8 die Verteilung der Variablen „Interesse an Actionfilmen“ in Cluster 4 („Spannung Liebende“) und der Gesamtstichprobe. Man sieht deutlich, dass die Mehrheit der Befragten insgesamt zu den Kategorien 4 und 5 („wenig“ und „überhaupt nicht“) tendiert. Dagegen liegt das Interesse an Actionfilmen im Cluster 4 deutlich höher (Kategorie 2 „stark“ und 3 „mittel“).

Vergleicht man die Lösung der (bereinigten) K-Means-Clusteranalyse mit der Lösung der TwoStep-Clusteranalyse, fällt auf, dass sich die Cluster 3 der TwoStep-Lösung („Informationssuchende“) und Cluster 1 der K-Means-Lösung von der Clusterbeschreibung

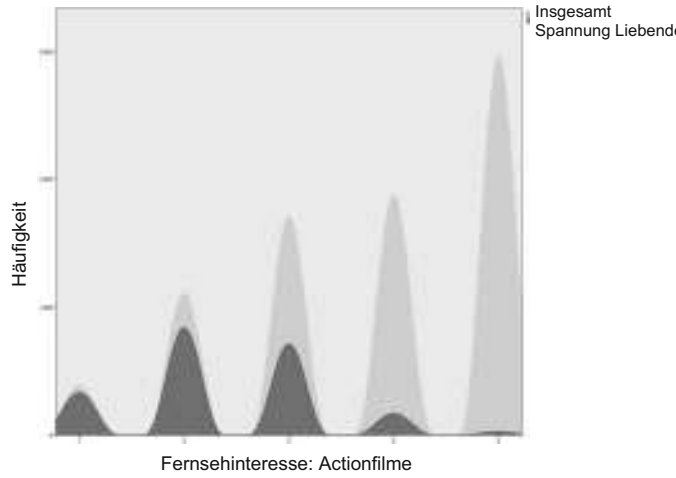

Abb. 8: Verteilung der Variablen „Interesse an Actionfilmen“ in Cluster 4 und den Daten insgesamt

Tab. 4: Vergleich der 4-Clusterlösungen aus K-Means mit der TwoStep-Lösung

		TwoStep-Clusteranalyse				
		1	2	3	4	Gesamt
K-Means basierend auf 4-Cl-Lösung ohne Outlier	1	213	13	**531**	6	763
	2	177	**421**	2	13	613
	3	79	5	13	**591**	688
	4	1	263	395	125	784
	Gesamt	470	702	941	735	2848

her sehr ähnlich sind. Auch ein hoher Anteil der Beobachtungen wird entsprechend gleich zugeordnet (siehe Tabelle 4). Entsprechendes gilt auch für die Cluster 2 der beiden Lösungen. Der Schwerpunkt des Interesses liegt jeweils auf den Heimatfilmen, Shows und Quiz und Unterhaltungsserien. Das Cluster der Spannungsliebenden ist ebenfalls in beiden Lösungen zu identifizieren. Dagegen unterscheiden sich Cluster 4 der K-Means-Lösung deutlich von Cluster 2 der TwoStep-Lösung (vielseitig Interessierte vs. Desinteressierte). Entsprechend heterogen ist auch die Zellverteilung beim Vergleich der beiden Lösungen.

4 Häufige Fehler

Von speziellen „Kunstfehlern“ in der Anwendung der Clusteranalyse lässt sich kaum sprechen, da Clusteranalyse – zumindest bei den hier behandelten Formen – weder ein inferenzstatistisches noch ein datentheoretisches Modell verwendet. Sie ist tatsächlich

nicht viel mehr als eine Sammlung von bestimmten Sortieralgorithmen. Nehmen wir einmal an, dass der Anwender ein einigermaßen sinnvolles Ähnlichkeits- oder Differenzmaß gewählt hat. Dann besteht sein häufigster Fehler vielleicht darin, nicht genügend Skepsis hinsichtlich der Annahme zu hegen, dass in der Stichprobe für den ausgewählten Datensatz tatsächlich eine Clusterstruktur in der Form vorliegt, wie sie von den oben genannten Algorithmen identifiziert werden: als Partition der Stichprobe. Das Dendrogramm bzw. der Fusionswerteverlauf weisen zwar in einer aktuellen Analyse keinen „Sprung“ auf und legen graphisch keine Clusterlösung nahe, dennoch wird dann das Dendrogramm häufig überinterpretiert, um in jedem Fall eine Lösung vorzuweisen.

Ein anderer Fehler ist es, wenn die Exploration der Daten vorschnell abgebrochen wird. Wenn sich bei einer Parameterwahl keine klaren Hinweise auf eine Lösung zeigen, so heißt das zunächst nur, dass sich für die gewählten Variablen und Parameter keine Clusterstruktur identifizieren lässt. Dann könnte es sinnvoll sein, diese Randbedingungen zu variieren. Sind die gewählten Variablen tatsächlich geeignet für eine Typologie? Fehlen vielleicht entscheidende Variablen? Oder enthalten die Analysevariablen überflüssige Variablen (masking variables), deren Beitrag im Differenzmaß die Systematik der Variablen verschleiert, die die Cluster tatsächlich konstituieren? Den Antworten auf diese Fragen kommt man u.U. nur durch mehrere Versuche auf die Spur, bei denen unterschiedliche Variablensätze analysiert werden. Jede solche versuchsweise Analyse muss natürlich auch mit dem Versuch einer inhaltlichen Validierung abgeschlossen werden, bei der durch Line-Plots die inhaltliche Bedeutung der Cluster und die Bedeutung der Variablen für die Clusterbildung veranschaulicht werden sollte.

Zudem kann man die Optionen für die Distanzmaße variieren. Hilfreich ist es auch, die störenden Einflüsse extremer Profile („Ausreißer“) zu eliminieren (siehe dazu 3.3). Aus einer Vielzahl von vollständig durchgeführten Analysen ergibt sich dann möglicherweise eine numerisch tragfähige und inhaltlich sinnvolle Typologie. Schließlich werden in der Clusteranalyse häufig auch Zufallscluster substanziell interpretiert. Die Hypothese, dass in einer gegebenen Stichprobe, bei gegebenen Parametern kein Cluster vorliegt, kann mit dem oben genannten Permutationstest von Wishart (2003) getestet werden. Allerdings ist für bestimmte Clusterstrukturen die Power des Tests gering. In diesen Fällen ist aber das Vorliegen von Clustern im Allgemeinen bereits durch einen charakteristischen Verlauf der Fusionswertekurve indiziert.

5 Literaturempfehlungen

Die Monographie von Everitt et al. (2001) scheint uns hinsichtlich der Stoffauswahl, des systematischen und didaktisch gelungenen Aufbaus und der Beispiele sehr gut geeignet für einen ersten umfassenden Überblick. Der Leser wird mit nahezu allen Arten von Verfahren – und auch verwandten Ansätzen wie z. B. Multidimensionaler Skalierung – in Theorie und Beispielen bekannt gemacht, ohne mit Varianten und technischen Details zu sehr belastet zu werden. Die einzige umfassende Monographie über Clusteranalyse in deutscher Sprache ist das Buch von Bacher (1996). In dieser Monographie sind nahezu alle Verfahrensklassen der Clusteranalyse vertreten, sei es dass es sich um Sortieralgorithmen handelt oder um Anwendung der fuzzy set-Theorie

oder auch um verteilungstheoretische Verfahren wie etwa latent class. Die theoretischen Ausführungen sind reichhaltig mit Anwendungsbeispielen illustriert. Das Buch ist daher gut als Nachschlagewerk und Referenz verwendbar.

Die Dokumentation des statistischen Hintergrunds des TwoStep-Verfahrens ist unbefriedigend. Bacher et al. (2004) versuchen in ihrem Papier eine kritische Würdigung des TwoStep-Verfahrens, die erstens die zu starken Vereinfachungen des Software-Herstellers und zweitens die sehr speziellen theoretischen Darstellungen in den Grundlagenartikeln der Entwickler vermeidet. Leider kann auch in diesem Papier, das uns als einziges dieser Art bekannt ist, der Informationsbedarf des Anwenders nur partiell befriedigt werden.

Literaturverzeichnis

Bacher, J. (1996). *Clusteranalyse. Anwendungsorientierte Einführung.* München: Oldenbourg.

Bacher, J., Wenzig, K., & Vogler, M. (2004). *SPSS TwoStep Cluster - A First Evaluation.* Arbeits- und Diskussionspapiere 2004-2, Universität Erlangen-Nürnberg, Lehrstuhl für Soziologie. Letzter Zugriff 29.03.2010: `http://www.soziologie.wiso.uni-erlangen.de/publikationen/a-u-d-papiere/a_04-02.pdf`.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (S. 263–268). New York: ACM.

Eckes, T. (1991). Bimodale Clusteranalyse: Methoden zur Klassifikation von Elementen zweier Mengen. *Zeitschrift für experimentelle und angewandte Psychologie*, 38, 201–225.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster Analysis.* London: Arnold.

Kaufman, L. & Rousseeuw, P. J. (2005). *Finding Groups in Data.* New York: Wiley.

Kaufmann, H. & Pape, H. (1984). Clusteranalyse. In L. Fahrmeir & A. Hamerle (Hg.), *Multivariate statistische Verfahren.* Berlin: de Gruyter.

Norusis, M. (2009). *SPSS 16.0 Statistical Procedures Companion.* Upper Saddle River: Prentice.

Theodoridis, S. & Koutroumbas, K. (2003). *Pattern Recognition.* Amsterdam: Academic Press, 2. Auflage.

Wiedenbeck, M. & Züll, C. (2001). *Klassifikation mit Clusteranalyse: Grundlegende Techniken hierarchischer und K-means-Verfahren.* ZUMA How-to Reihe 2001, Nr. 10. Letzter Zugriff 29.03.2010: `http://www.gesis.org/forschung-lehre/gesis-publikationen/gesis-reihen/how-to/`.

Wishart, D. (2003). *ClustanGraphics Primer. A Guide to Cluster Analysis.* Edinburgh: Clustan Limited.

Zhang, T., Ramakrishnon, R., & Livny, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large Databases. In H. V. Jagadish & I. S. Mumick (Hg.), *Proceedings of the ACM SIGMOD Conference on Management of Data* (S. 103–114). New York: ACM.