Überlappende Clusterstrukturen: ein Verfahren zur exploratorischen Datenanalyse

Uehlinger, Hans-Martin

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Überlappende Clusterstrukturen – ein Verfahren zur exploratorischen Datenanalyse

"Clusteranalyse" wird in den empirischen Sozialwissenschaften immer noch häufig mit agglomerativer hierarchischer Clusteranalyse gleichgesetzt. Der Beitrag zeigt auf, daß hierarchische Cluster als Spezialfall überlappender Cluster verstanden werden können. Die Modelle ADCLUS (für ADditive CLUStering) und INDCLUS (für INdividual Differences CLUStering) und deren Annahmen und Implikationen werden aufgezeigt. In einem Beispiel aus der politischen Partizipationsforschung wird das ADCLUS-Modell parallel zu multidimensionaler Skalierung und hierarchischer Clusteranalyse als exploratorisches Datenanalyseverfahren angewendet.

1. Clusteranalytische Ansätze

Bei der Analyse von Ähnlichkeiten bzw. Distanzen mit clusteranalytischen Verfahren stehen in empirischen sozialwissenschaftlichen Studien agglomerative hierarchische Verfahren stark im Vordergrund. Diese Präferenz hierarchischer Verfahren läßt sich wohl eher auf die Einfachheit der Algorithmen und die Verfügbarkeit von Programmen in den Standard-Statistiksoftwarepaketen (SPSS, SAS und BMDP) als auf die Annahme einer entsprechenden latenten Struktur zurückführen.

Die hierarchischen Modelle gehen von der Annahme aus, daß die Objekte (d.h. die Einheiten, die zu Clustern zusammengefaßt werden) eine hierarchische latente Struktur haben. Untersucht man Paare von Clustern, so ist entweder das eine Cluster Teil des anderen Clusters oder aber die beiden Cluster haben keine gemeinsamen Objekte. Die agglomerativen hierarchischen Verfahren beginnen damit, daß jedes Objekt ein einzelnes Cluster bildet. Schrittweise werden die Cluster zu größeren Clustern zusammengefaßt, bis alle Objekte in einem einzigen Cluster vereinigt sind. Daraus resultiert die hierarchische Struktur der Cluster.

Dieses hierarchische Modell kann als Spezialfall clusteranalytischer Modelle gewertet werden, die ein Überlappen von Clustern zulassen. Beim hierarchischen Modell ist die Überlappung auf die Schachtelung begrenzt, gleichzeitig ist die Schachtelung zwingend. Als anderer Spezialfall clusteranalytischer Modelle mit Überlappung lassen sich die Ansätze disjunkter Cluster interpretieren. Bei diesen Ansätzen darf jedes Objekt einem und nur einem Cluster angehören.

2. Die Modelle ADCLUS und INDCLUS

Das von Roger N. Shepard und Phipps Arabie (1979) formulierte ADCLUS-Modell (für ADditive CLUStering) läßt Überlappungen von Clustern in beliebiger Art zu. Ausgangspunkt des Modells ist eine symmetrische Ähnlichkeitsmatrix S. Nach der "taxonomy of measurement data and multidimensional measurement models" von Carroll/Arabie (1980, 608-612) und der Nomenklatur von Young (1987) ist diese Matrix S eine N x N 2-Weg 1-Modus Matrix. Jedes Element s_{ij} der Matrix S gibt die Ähnlichkeit zwischen den Objekten i und j an. Die Matrix muß symmetrisch sein (d.h. $s_{ij} = s_{ji}$) oder vor der Analyse symmetrisiert werden.[1)] Das Modell setzt voraus, daß die Daten zumindest auf Intervallniveau gemessen sind. Indizieren die Daten ursprünglich das Ausmaß der Unähnlichkeit oder Distanz, so sind sie mittels einer linearen Transformation in Ähnlichkeiten umzusetzen. Da davon ausgegangen wird, daß die Daten zumindest auf Intervallniveau gemessen sind, wird die Anpassungsgüte des Modells (goodness-of-fit) durch lineare Transformationen nicht beeinflußt (Carroll/Arabie 1983:158). Die Daten werden allgemein zuerst auf den Wertebereich [0, 1] standardisiert. Die Matrix S muß schließlich vollständig sein, d.h. keines der Elemente s_{ij} darf fehlen.

Die grundlegende Idee des ADCLUS-Modells ist, die Ähnlichkeit s_{ij} zwischen den Objekten i und j als Linearkombination gewichteter Cluster vorherzusagen, zu denen beide Objekte i und j gehören.

$$\hat{s}_{ij} = \sum_{k=1}^{K} w_k p_{ik} p_{jk},$$

wobei w_k = (nicht negatives) Gewicht des k-ten Clusters (k = 1, ..., K)

$$p_{ik} = \begin{cases} 1 & \text{, falls das Objekt i (i = 1, ..., N) zum Cluster k gehört} \\ 0 & \text{sonst.} \end{cases}$$

Das ADCLUS-Modell kann in Matrixnotation wie folgt geschrieben werden:

$$\hat{S} = PWP'$$

wobei $\hat{S}$ = N x N symmetrische Matrix mit den vorhergesagten Ähnlichkeiten $\hat{s}_{ij}$ (mit Werten 1 in der Hauptdiagonalen)

P = N x K rechteckige Matrix mit binären Elementen (0, 1) p_{ik}

W = K x K Diagonalmatrix mit den (nicht negativen) Gewichten w_k

1983 veröffentlichten J. Douglas Carroll und Phipps Arabie die endgültige Version des INDCLUS-Modells (für INdividual Differences CLUStering) als 3-Weg Generalisierung des ADCLUS-Modells. Im Falle des INDCLUS-Modells ist die Ausgangsmatrix S eine N x N x H 3-Weg 2-Modi Matrix. Ausgehend vom ADCLUS-Modell wird beim INDCLUS-Modell die Ähnlichkeit s^h_{ij} zwischen den Objekten i und j beim Subjekt h als Linearkombination gewichteter Cluster vorhergesagt, zu denen beide Objekte i und j gehören.[2] Entscheidend beim INDCLUS-Modell ist, daß die Clustergewichte spezifisch sind für die H Subjekte.

$$\hat{s}^h_{ij} = \sum_{k=1}^{K} w_{kh} p_{ik} p_{jk},$$

wobei w_{kh} = (nicht negatives) Gewicht des k-ten Clusters (k = 1, ..., K) beim h-ten Subjekt (h = 1, ..., H)

In Matrixnotation ergibt sich dann für das INDCLUS-Modell:

$$\hat{S}^h = PW^hP'.$$

In dieser Notation des ADCLUS- und des INDCLUS-Modells müssen in der K-ten Spalte der Matrix P alle Werte gleich 1 sein, d.h. alle Objekte gehören dem K-ten Cluster an. Dieses universelle Cluster ist notwendig für die Berechnung des Anteils erklärter Varianz (als Maß der Anpassungsgüte des Modells an die Daten) mittels linearer Regression. In der Regressionsrechnung hat dieses universelle Cluster die Bedeutung der additiven Konstante, die auch negativ sein kann.

3. Annahmen und deren Implikationen bei den Modellen ADCLUS und INDCLUS

Bei den Modellen ADCLUS und INDCLUS können die Cluster in beliebiger Form überlappen. Aus theoretisch-inhaltlicher Sicht ist die Annahme überlappender Cluster für sehr viele Studien sinnvoll: Wird die gemeinsame Zugehörigkeit

von zwei Objekten zu einem Cluster als Vorhandensein einer gemeinsamen Eigenschaft interpretiert, so kann ein Objekt unterschiedliche Eigenschaften mit verschiedenen anderen Objekten gemeinsam haben. Die Modelle mit disjunkten Clustern schließen diese Möglichkeit aus; die hierarchischen Modelle auf der anderen Seite beschränken die Überlappung auf Schachtelung.

Die Zusammenfügung von Objekten zu einem Cluster basiert inhaltlich interpretiert auf einer Eigenschaft, die allen Objekten im Cluster gemeinsam ist, während sie den übrigen Objekten fehlt. Jedes Cluster hat nun ein spezifisches Gewicht w_k (k = 1, ..., K) im ADCLUS-Modell bzw. w_{kh} (k= 1, ..., K; h = 1, ..., H) im INDCLUS-Modell. Aus technischer Sicht ist der Gewichtsfaktor w_k bzw. w_{kh} das Ausmaß an Ähnlichkeit zwischen allen Paaren von Objekten, die dem Cluster k angehören, das durch die Eigenschaft k erklärt wird. Die einzelnen Eigenschaften tragen in verschiedenem Maße zur Erklärung der Ähnlichkeit zwischen zwei Objekten i und j bei, wobei dieser Erklärungsbeitrag bei jedem der H Subjekte unterschiedlich ist. In der Literatur finden sich wenig Aussagen über die inhaltliche Bedeutung und Interpretation dieser Gewichtsfaktoren. Arabie/Carroll/DeSarbo/Wind (1981:312) interpretieren das Gewicht als Repräsentation der "salience of the property", Shepard (1980: 397) spricht von "psychological weight".

Die ADCLUS/INDCLUS-Modelle nehmen an, daß ein Objekt i bestimmte Eigenschaften hat, die durch die Zugehörigkeit zu den entsprechenden Clustern repräsentiert werden. Dabei ist die Bedeutung einer Eigenschaft k für alle Objekte, die diese Eigenschaft besitzen, identisch; verschiedene Eigenschaften k haben unterschiedliche Bedeutung. Dies kommt in der Dichotomie der Zugehörigkeit von Objekten zu Clustern zum Ausdruck. Das INDCLUS-Modell geht davon aus, daß die Objekte bei sämtlichen Subjekten identische Eigenschaften besitzen, daß jedoch deren Bedeutung bei den einzelnen Subjekten unterschiedlich ist. Aus inhaltlich-theoretischer Sicht ist jeweils kritisch zu prüfen, ob diese Annahmen gerechtfertigt sind.

Die ADCLUS/INDCLUS-Modelle sind bewußt als diskrete Modelle formuliert worden. Das Charakteristikum der Dichotomie der Zugehörigkeit von Objekten zu Clustern unterscheidet diese Modelle denn auch von Ansätzen der Faktorenanalyse und der multidimensionalen Skalierung: Wären die Elemente der Matrix P nicht auf die dichotomen Werte 0 und 1 limitiert, so wäre das ADCLUS-Modell im Kern identisch mit dem faktoranalytischen Modell, das Ekman (1954, 1963)

zur Repräsentation der Struktur in Ähnlichkeitsdaten vorgeschlagen hat. Würde die Ähnlichkeitsmatrix S nach dem Hauptkomponentenansatz analysiert, so ergäben sich Schätzungen von P (deren Spalten dann die Eigenvektoren sind) und von W (deren Diagonalelemente die Eigenwerte sind). Das INDCLUS-Modell kann als clusteranalytisches, diskretes Analogon des räumlichen, kontinuierlichen INDSCAL-Modells von Carroll und Chang (1970) interpretiert werden. Die Cluster entsprechen den Dimensionen, die für die H Subjekte unterschiedliche Gewichte haben. Die diskrete Form der Zugehörigkeit von Objekten zu Clustern würde auf das INDSCAL-Modell übertragen bedeuten, daß die Objekte auf jeder Dimension nur an einer von zwei Positionen liegen können.

Bei den ADCLUS/INDCLUS-Modellen wird als Maß für die Anpassungsgüte des Modells an die Daten der Anteil erklärter Varianz VAF (Variance Accounted For) berechnet. Dabei ist der Anteil der Gesamtvarianz für das Modell insgesamt und auch für jedes Cluster zu berechnen. Mirkin (1987:29) zeigt, daß der Wert VAF für ein bestimmtes Cluster gering sein kann, auch wenn das Gewicht w_k bzw. w_{kh} hoch ist. Besondere Bedeutung kommt der Überprüfung der Residuen $s - \hat{s}$ zu.

4. Algorithmen und Computerprogramme für die ADCLUS/INDCLUS-Modelle

Ein Computerprogramm ADCLUS zum ADCLUS-Modell wurde von Shepard und Arabie (1979) entwickelt. Ein Jahr später präsentierten Arabie und Carroll (1980) den MAPCLUS-Algorithmus (für MAthematical Programming CLUStering) mit dem zugehörigen Programm MAPCLUS. MAPCLUS dient wie das Programm ADCLUS der Analyse nach dem ADCLUS-Modell, verbessert aber den algorithmischen Ansatz. Der MAPCLUS-Algorithmus kombiniert einen Alternating Least Squares-Ansatz mit einer Mathematical Programming-Optimierungsroutine. Die Anzahl der Cluster K ist vom Benutzer vorzugeben. Der MAPCLUS-Algorithmus ist detailliert beschrieben in Arabie/Carroll (1980).

Der INDCLUS-Algorithmus generalisiert den MAPCLUS-Algorithmus für das 3-Weg INDCLUS-Modell. Er ist detailliert beschrieben in Carroll/Arabie (1983:159-162). Wird eine 2-Weg 1-Modus Ähnlichkeitsmatrix S mit dem INDCLUS-Algorithmus analysiert, so sollte sich das gleiche Resultat wie mit dem MAPCLUS-Algorithmus ergeben. Das entsprechende Programm INDCLUS ist Teil der Programmbibliothek MDS (2) der AT & T Bell Laboratories.[3)]

Warren S. Sarle (SAS Institute Inc.) schrieb für das INDCLUS-Modell die SAS-Prozedur OVERCLUS, die Teil der SUGI Supplemental Library ist (SAS Institute Inc. 1986:401-421). Der OVERCLUS-Algorithmus verwendet OLS als Optimierungskriterium. Es stehen vier Optimierungstechniken zur Verfügung:
Die Optimierungstechnik 0 besteht aus zwei geschachtelten Schleifen. Die äußere Schleife fügt schrittweise Cluster hinzu, bis die vom Benutzer festgelegte Clusterzahl K erreicht ist. Die einzelnen Cluster werden in der inneren Schleife berechnet. Als erste Objekte eines Clusters werden diejenigen Objekte i und j ausgewählt, die bei der bisherigen Clusterstruktur am schlechtesten erklärt sind (d.h. deren Residualwert $s_{ij} - \hat{s}_{ij}$ am größten ist). Schrittweise werden nun diejenigen Objekte zum neuen Cluster hinzugefügt, die das Optimierungskriterium am stärksten verbessern. Die Bildung des Clusters wird dann abgeschlossen, wenn weitere Objekte das Optimierungskriterium nicht mehr verbessern können. Während der inneren Schleife werden die Gewichte w_k bzw. w_{kh} der davor gebildeten Cluster konstant gehalten.
Die Optimierungstechniken 1 bis 3 unterscheiden sich von der Technik 0 dadurch, daß nach der Bildung eines Clusters in einem weiteren Schritt geprüft wird, ob mit irgendeiner Veränderung der bisherigen Cluster das Optimierungskriterium verbessert werden könnte. Die Techniken 1 bis 3 differieren in bezug auf die erneute Berechnung der Clustergewichte w_k bzw. w_{kh} (vgl. SAS Institute Inc. 1986:403).

5. Illustrative Anwendung des ADCLUS-Modells im Bereich der politischen Partizipationsforschung

Im Forschungsprojekt "Jugend und Staat" sind 1980 knapp fünftausend Jugendliche und junge Erwachsene (16-35 Jahre; Bundesrepublik und West-Berlin) interviewt worden (Schmidtchen 1983). Dabei wurden den Befragten auf Kärtchen 22 Formen der politischen Beteiligung[4)] vorgelegt für die folgenden drei Fragen:

> Wenn Sie politisch in einer Sache, die Ihnen wichtig ist, Einfluß nehmen, Ihren Standpunkt zur Geltung bringen wollen: Welche der Möglichkeiten auf diesen Karten würden Sie dann nutzen, was davon kommt für Sie in Frage?
>
> Und wenn nun die von Ihnen angegebenen Maßnahmen und Aktionen nichts helfen, wenn der Staat und die Behörden einfach taub bleiben und auf nichts eingehen, welche Möglichkeiten kommen dann für Sie

in Frage? Sehen Sie sich diese Karten noch einmal durch und geben Sie mir nochmals alles an, was in dieser Situation für Sie in Frage kommt.

Und zum Schluß: Was davon haben Sie selbst schon gemacht, woran waren Sie schon einmal beteiligt?

Zielsetzung der Analyse ist es, die latenten Strukturen hinter diesen einzelnen Formen der politischen Partizipation zu ermitteln. Ausgangspunkt der Analyse ist die Berechnung der Matrix S mit den Ähnlichkeiten aller Paare von politischen Partizipationsformen. Dabei wird vom größtmöglichen Kreis möglicher zukünftiger bzw. bisheriger Partizipation ausgegangen, d.h. es wird für jede Partizipationsform ermittelt, ob zumindest eine der drei Fragen positiv beantwortet worden ist. Als Ähnlichkeitsmaß dient der Koeffizient ®.[5)]

Die Ähnlichkeitsmatrix S wird parallel mit drei exploratorischen Verfahren analysiert, die der Aufdeckung der latenten Strukturen dienen sollen:
- nichtmetrische multidimensionale Skalierung;
- hierarchische agglomerative Clusteranalyse nach dem Average linkage-Ansatz;
- überlappende Clusteranalyse nach dem ADCLUS-Modell.

Die Analyse folgt damit der allgemeinen Empfehlung, die Ausgangsdaten parallel mit verschiedenen, räumlichen und nicht-räumlichen sowie kontinuierlichen und diskreten Ansätzen zu untersuchen (vgl. etwa Arabie/Carroll/DeSarbo 1987:54).

Die nichtmetrische multidimensionale Skalierung ergibt eine Lösung mit hoher Anpassungsgüte in zwei Dimensionen (vgl. Abbildung 2).[6)] Das Badness of fit-Maß Stress hat bei der zweidimensionalen Lösung den Wert $S_1 = .07$.[7)]

Die hierarchische agglomerative Clusteranalyse nach dem Average linkage-Ansatz ergibt eine klare Strukturierung in mehrere Partizipationscluster (vgl. Abbildung 1). Inhaltlich interpretiert zeigt sich auf der zweitletzten Stufe der Zusammenfügung von Objekten zu Clustern ein Cluster, das alle legalen Partizipationsformen umfaßt, und ein Cluster, das die illegalen Aktivitäten

zusammenschließt. Diese Differenzierung in legale versus illegale Partizipationsformen ist von zentralem Stellenwert für die Strukturierung der politischen Partizipation.

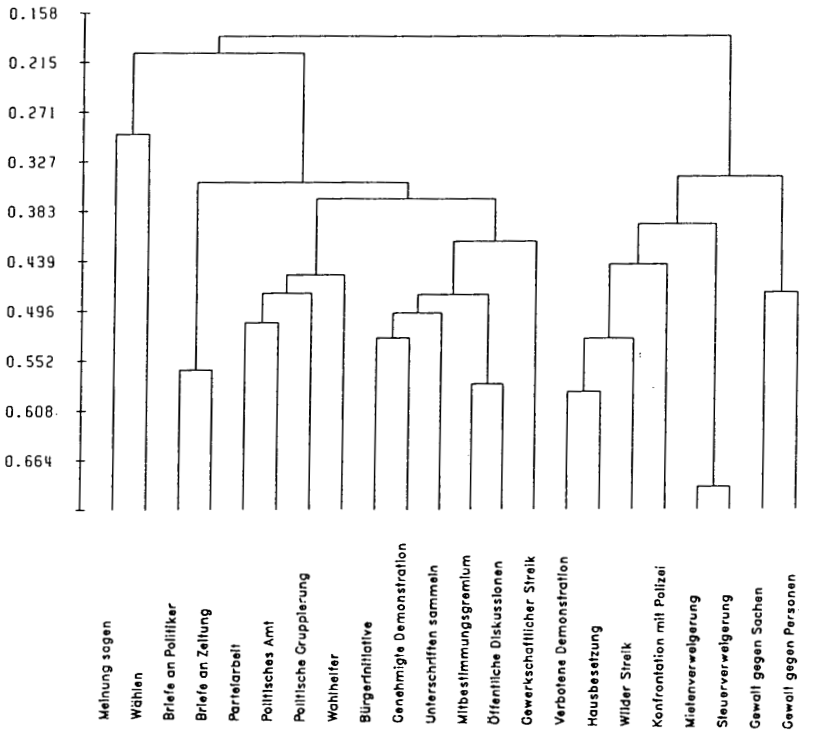

Abbildung 1: Hierarchische agglomerative Clusteranalyse nach dem Average linkage-Ansatz

Die Ähnlichkeit bzw. Divergenz der Lösungsstrukturen der MDS-Analyse und der hierarchischen Clusteranalyse kann visualisiert werden, indem die Clusterstrukturen in die räumliche Lösung eingezeichnet werden (vgl. Abbildung 2). Diese Darstellung zeigt die hierarchische Struktur der Clusterlösung sehr

klar auf. Innerhalb der legalen bzw. illegalen Formen politischer Partizipation ergeben sich Teilcluster, die sich auch räumlich voneinander abgrenzen. Augenscheinlich ist vor allem die Differenzierung zwischen gewaltlosen und gewaltsamen Formen illegaler Aktivitäten.

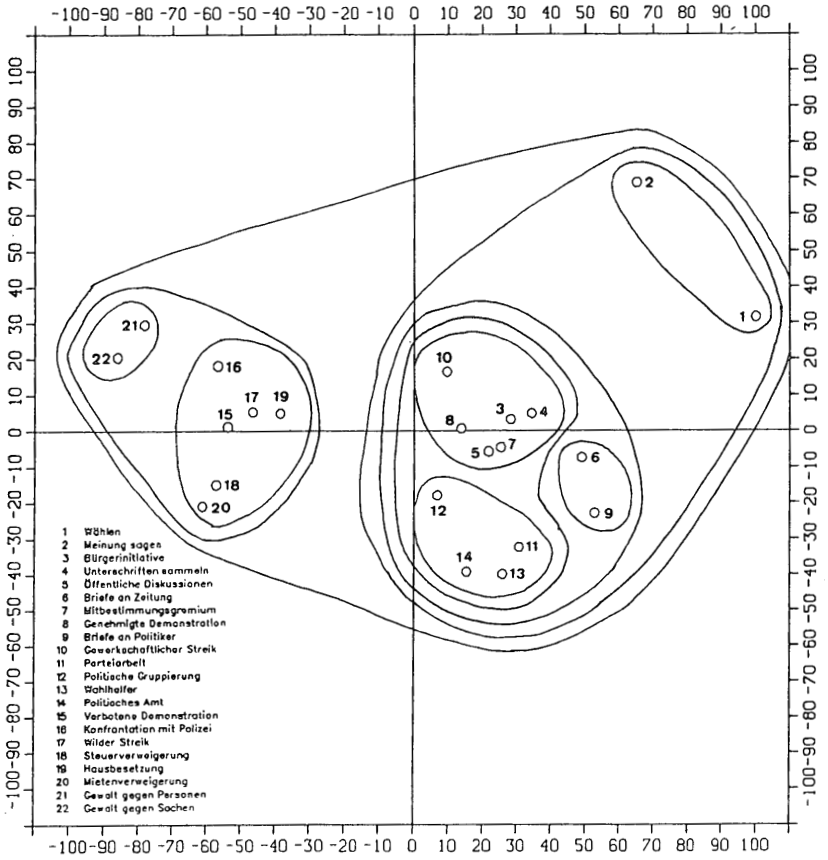

Abbildung 2: Zweidimensionale Lösung der nichtmetrischen multidimensionalen Skalierung/hierarchische Cluster nach dem Average linkage-Ansatz mit $\delta \leq .396$

Bei der Analyse der Matrix S nach dem ADCLUS-Modell[8] ist die Anzahl der Cluster vorzugeben. Da die Anzahl der Cluster bei der exploratorischen Analyse aber zumeist als unbekannt zu werten ist, werden Lösungen mit unterschiedlicher Anzahl Cluster gesucht. Als Entscheidungsgrundlagen für die auszuwählende Anzahl Cluster können die Varianzerklärung und die Interpretierbarkeit herangezogen werden. Die Tabelle 1 gibt die Entwicklung der Varianzerklärung mit steigender Zahl der Cluster wieder.

Tabelle 1: Erklärter Anteil der Gesamtvarianz mit K Clustern im ADCLUS-Modell

Anzahl Cluster[9]	Erklärter Anteil der Gesamtvarianz
3	68.0 %
4	70.2 %
5	80.2 %
6	81.2 %
7	84.8 %
8	88.4 %
9	89.2 %
10	89.5 %
11	91.5 %

Die Gliederung der Objekte in zwei Cluster (neben dem globalen Cluster) ergibt bereits einen erklärten Anteil der Gesamtvarianz von 68 Prozent. Aus der Tabelle 2 geht hervor, welche Partizipationsformen zu diesen Clustern gehören: Während das Cluster 1 alle illegalen Partizipationsformen umfaßt, sind im Cluster 2 die legalen Partizipationsformen (mit Ausnahme der Aktivitäten "Wählen" und "Meinung sagen") enthalten. Dieses Resultat bestätigt die Ergebnisse der nichtmetrischen multidimensionalen Skalierung und der agglomerativen hierarchischen Clusteranalyse insofern, als die kardinale Bedeutung der Trennung zwischen legalen und illegalen politischen Aktivitäten hervorgehoben wird.

Die Hinzunahme eines weiteren Clusters verbessert den erklärten Anteil der Gesamtvarianz nur marginal um 2.2 Prozent (vgl. Tabelle 1). Diesem Cluster gehören die beiden Aktivitäten "Steuerverweigerung" und "Mietenverweigerung" an. Es handelt sich dabei um zwei illegale Partizipationsformen, die somit bereits auch zum Cluster 1 gehören. Dieses Cluster ist deshalb als ganzes Teil eines größeren Clusters; die beiden Cluster stehen in einem geschachtelten Verhältnis als Spezialfall der Überlappung zueinander.

Tabelle 2: 3-Cluster-Lösung nach dem ADCLUS-Modell

Cluster	w_k	Typ II Anteil der Gesamtvarianz[10]	Objekte
1	.22	30.3 %	Verbotene Demonstration / Konfrontation mit Polizei / wilder Streik / Steuerverweigerung / Hausbesetzung / Mietenverweigerung / Gewalt gegen Personen / Gewalt gegen Sachen
2	.21	52.7 %	Bürgerinitiative / Unterschriften sammeln / öffentliche Diskussionen / Briefe an Zeitung / Mitbestimmungsgremium / genehmigte Demonstration / Briefe an Politiker / gewerkschaftlicher Streik / Parteiarbeit / politische Gruppierung / Wahlhelfer / politisches Amt
3	.19		alle Objekte

Eine Lösung mit fünf Clustern bringt demgegenüber eine entscheidende Verbesserung der Erklärung der Gesamtvarianz um 10.0 Prozent. Dieser Zuwachs wird dadurch möglich, daß sich die Cluster im ADCLUS-Modell überlappen können. Die Tabelle 3 zeigt die Zugehörigkeit der Objekte zu den fünf Clustern auf; in der Abbildung 3 sind die Cluster in die zweidimensionale MDS-Lösung eingetragen.

Tabelle 3: 5-Cluster-Lösung nach dem ADCLUS-Modell

Cluster	w_k	Typ II Anteil der Gesamtvarianz	Objekte
1	.20	24.0 %	vgl. Tabelle 2
2	.20	49.7 %	vgl. Tabelle 2
3	.33	2.8 %	Steuerverweigerung / Mietenverweigerung
4	.09	10.0 %	Bürgerinitiative / Unterschriften sammeln / öffentliche Diskussionen / Mitbestimmungsgremium / genehmigte Demonstration / gewerkschaftlicher Streik / politische Gruppierung / verbotene Demonstration / Konfrontation mit Polizei / wilder Streik / Steuerverweigerung / Hausbesetzung
5	.17		alle Objekte

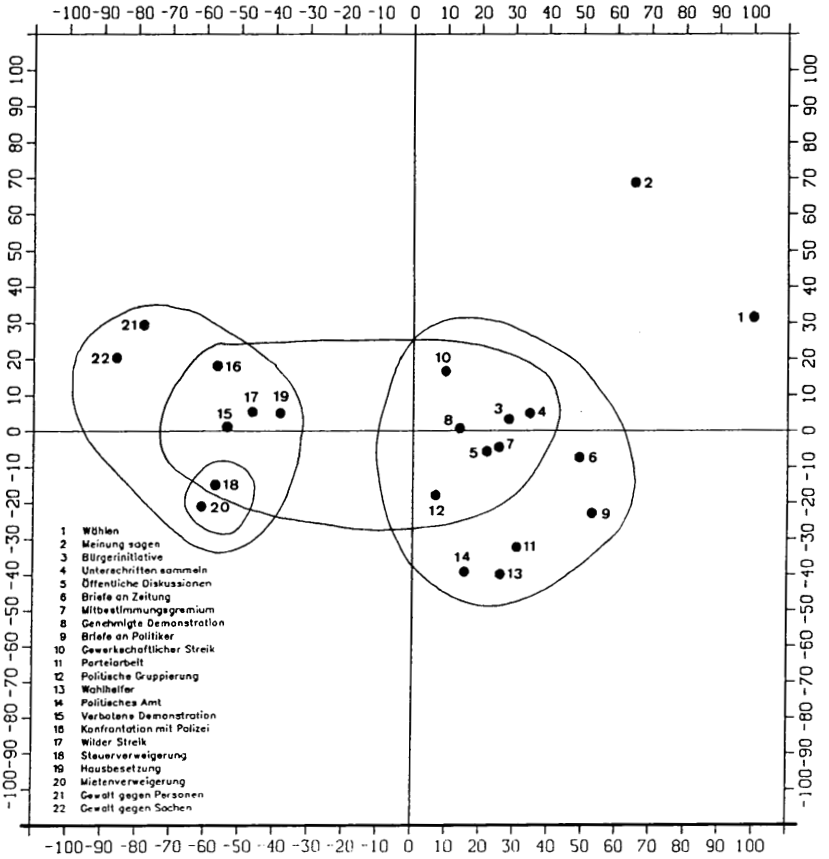

Abbildung 3: Zweidimensionale Lösung der nichtmetrischen multidimensionalen Skalierung / überlappende Cluster nach dem ADCLUS-Modell

Das Cluster 4 umfaßt diejenigen Partizipationsformen, die von Barnes, Kaase et al. (1979) in der Studie "Political action" als "unkonventionelle Aktivitäten" charakterisiert werden. Es handelt sich dabei um legale und um illegale (aber gewaltfreie) Aktivitäten, die auf eine spezifische politische Problemstellung fokussiert sind. Dank des überlappenden Clusteransatzes läßt

sich diese Gruppierung - bei der generellen Dominanz der Trennung zwischen legalen und illegalen Formen - ermitteln. Zum Cluster 1, nicht jedoch zum Cluster 4 gehören die gewaltsamen illegalen Aktivitäten sowie die Mietenverweigerung. Zum Cluster 2, jedoch nicht zum Cluster 4 gehören die parteiorientierten Partizipationsformen ("Parteiarbeit" / "Wahlhelfer" / "politisches Amt") sowie die briefliche Kontaktnahme ("Briefe an Zeitung" / "Briefe an Politiker").[11)]

Die Hinzunahme weiterer Cluster erhöht den erklärten Anteil der Gesamtvarianz jeweils nicht erheblich (vgl. Tabelle 1). Bei insgesamt 11 Clustern werden 91.5 Prozent der Gesamtvarianz erklärt, so daß jedes der sechs zusätzlichen Cluster gegenüber der obigen Lösung durchschnittlich weniger als zwei Prozent der Gesamtvarianz zusätzlich erklärt.

Insgesamt ergeben sich als Lösung der überlappenden Clusteranalyse nach dem ADCLUS-Modell drei Cluster von Bedeutung: Ein Cluster umfaßt alle illegalen Aktivitäten, ein zweites die legalen Aktivitäten (mit Ausnahme von "Wählen" und "Meinung sagen"). Hier kommt die zentrale Bedeutung der Abgrenzung illegaler Partizipationsformen von den legalen zum Ausdruck. Zusätzlich jedoch ist eine Verknüpfung all jener (legalen wie illegal-gewaltfreien) Aktivitäten von Belang, die auf eine spezifische Problemstellung fokussiert sind. Als Beispiel soll von der Partizipationsform "verbotene Demonstration" ausgegangen werden: "Verbotene Demonstration" hat mit "Gewalt gegen Personen" die Eigenschaft der Illegalität, mit "genehmigter Demonstration" dagegen die Eigenschaft der Unkonventionalität oder Fokussierung auf eine spezifische politische Problemstellung gemeinsam. Diese Art unterschiedlicher gemeinsamer Eigenschaften eines Objektes mit anderen Objekten wird erst im Modell überlappender Cluster möglich.

Zusammenfassend haben die parallelen Analysen der Ähnlichkeitsmatrix S der Formen politischer Partizipation nach den Ansätzen der nichtmetrischen multidimensionalen Skalierung, der hierarchischen Clusteranalyse und der überlappenden Clusteranalyse insofern übereinstimmende Resultate ergeben, als stets eine klare Differenzierung zwischen den legalen und den illegalen Aktivitäten erfolgt. Bei der hierarchischen Clusteranalyse wird die Gruppierung der problemspezifischen (legalen wie illegalen) Aktivitäten aufgrund der Modellannahme verdeckt.

6. Schlußfolgerungen

Die ADCLUS/INDCLUS-Modelle sagen die beobachtete Ähnlichkeit s_{ij} zwischen zwei Objekten i und j als Linearkombination gewichteter Cluster vorher, zu denen beide Objekte i und j gehören. Diese Modelle nehmen eine diskrete Zugehörigkeit von Objekt i zu Cluster k an. Es ist bei der jeweiligen Anwendung zu prüfen, inwiefern diese Annahme sinnvoll erscheint. Als Maß für die Bedeutung der einzelnen Cluster ist der erklärte Anteil der Gesamtvarianz und nicht der Gewichtsfaktor w_k bzw. w_{kh} heranzuziehen. Noch unbefriedigend gelöst erscheint die Frage der inhaltlichen Interpretation dieser Gewichtsfaktoren, die aber für die Modelle von zentraler Bedeutung sind.

Die Modelle der überlappenden Cluster beinhalten das Modell hierarchischer Cluster als Spezialfall. Ist die Annahme einer hierarchischen Struktur nicht durch theoretisch-inhaltliche Überlegungen gegeben, so sollten - zumindest parallel zu agglomerativen hierarchischen Clusteranalysen - überlappende Clusteranalysen durchgeführt werden.

Dieser Beitrag wurde von Hans-Martin Uehlinger verfaßt.

Anmerkungen

1. Wayne S. DeSarbo (1982) entwickelte den Algorithmus GENNCLUS (für GENeral Nonhierarchical CLUStering), der nicht-symmetrische Daten zuläßt.
2. Die Elemente h (h = 1, ..., H) des dritten Weges werden in Anlehnung an die entsprechende Terminologie bei der multidimensionalen Skalierung als "Subjekte" bezeichnet.
3. Die Programmbibliothek MDS (2) ist zum Preis von US $ 200 erhältlich bei: Computing Information Service, Room 2F-128A, AT & T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974, USA.
4. Es handelte sich um die folgenden 22 Vorgaben:
   (1) Sich an Wahlen beteiligen
   (2) Seine Meinung sagen, im Bekanntenkreis und am Arbeitsplatz
   (3) Mitarbeit in einer Bürgerinitiative
   (4) Unterschriften sammeln
   (5) Sich in Versammlungen an öffentlichen Diskussionen beteiligen
   (6) Briefe an eine Zeitung senden
   (7) In einem Mitbestimmungsgremium im Betrieb, in der Schule, in der Ausbildungsstätte mitarbeiten
   (8) Teilnahme an einer genehmigten politischen Demonstration
   (9) Briefe an Politiker schreiben
   (10) Teilnahme an einem gewerkschaftlich beschlossenen Streik
   (11) In irgendeine Partei eintreten, aktiv mitarbeiten
   (12) In einer politischen Gruppierung mitmachen
   (13) Als Wahlhelfer Kandidaten unterstützen
   (14) Ein politisches Amt übernehmen

(15) Teilnahme an einer verbotenen Demonstration
(16) Dem eigenen Standpunkt Nachdruck verleihen, auch wenn es dabei zu einer direkten Konfrontation mit der Polizei, mit der Staatsgewalt kommen sollte
(17) Beteiligung an einem wilden Streik
(18) Weigerung, Steuern oder Stromrechnung zu zahlen
(19) Hausbesetzung, Besetzung von Fabriken, Ämtern
(20) Weigerung, Mieten oder Kreditabzahlungsraten zu zahlen
(21) Für eine Sache kämpfen, auch wenn dazu Gewalt gegen politisch Verantwortliche notwendig ist
(22) Bei einer Demonstration mal richtig Krach schlagen, auch wenn dabei einiges zu Bruch geht

5. Für eine ausführliche Diskussion der operationalen Definition von politischer Partizipation und der Wahl des Ähnlichkeitsmaßes vgl. Uehlinger (1988).
6. Die Analyse erfolgt mit dem Programm MINISSA des Programmpakets MDS(X), Version 3.20.
7. Die Werte von Stress S_1 betragen für die fünf- bis eindimensionalen Lösungen .02 / .03 / .05 / .07 / .14. Der oftmals als Entscheidungshilfe herangezogene Ellbogen im Diagramm Dimensionalität versus Stress liegt somit bei 2 Dimensionen. Das Shepard-Diagramm Distanz versus Ähnlichkeit zeigt über einen Großteil der Werte eine lineare Beziehung. Für Einzelheiten vgl. Uehlinger (1988).
8. Die Analyse erfolgt mit der SAS-Prozedur OVERCLUS von Warren S. Sarle.
9. Bei der Anzahl der Cluster ist zu berücksichtigen, daß gemäß obiger Notation jeweils eines der Cluster alle Objekte, d.h. alle Formen der politischen Partizipation, umfaßt.
10. Der Typ II Anteil der Gesamtvarianz gibt für jedes Cluster denjenigen Teil der Gesamtvarianz an, der durch dieses Cluster erklärt wird, wenn die anderen Cluster des Modells bereits ausgewählt worden sind.
11. Die Vorgabe "Politische Gruppierung" gehört dem Cluster 4 ebenfalls an. Uehlinger (1988:98) diskutiert das Problem der inhaltlichen Interpretation dieser Vorgabe.
Der Nicht-Aufnahme der brieflichen Kontaktnahme in das Cluster 4 wäre inhaltlich detailliert zu erörtern. Vgl. dazu Uehlinger (1988:97-98).

Literatur

Arabie, P./Carroll, J.D., 1980: A mathematical programming approach to fitting the ADCLUS model. Psychometrika 45:211-235.

Arabie, P./Carroll, J.D./DeSarbo, W.S., 1987: Three-way scaling and clustering. Quantitative Applications in the Social Sciences, 65. Newbury Park: Sage.

Arabie, P./Carroll, J.D./DeSarbo, W./Wind, J., 1981: Overlapping clustering: A new method for product positioning. Journal of Marketing Research 18:310-317.

Barnes, S.H./Kaase, M. et al., 1979: Political Action. Mass Participation in Five Western Democracies. Beverly Hills: Sage.

Carroll, J.D./Arabie, P., 1980: Multidimensional scaling. Annual Review of Psychology 31:607-649.

Carroll, J.D./Arabie, P., 1983: INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. Psychometrika 48:157-169.

Carroll, J.D./Chang, J.J., 1970: Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika 35:283-319.

DeSarbo, W., 1982: GENNCLUS: New models for general nonhierarchical clustering analysis. Psychometrika 47:449-475.

Ekman, G., 1954: Dimensions of color vision. Journal of Psychology 38:467-474.

Ekman, G., 1963: A direct method for multidimensional ratio scaling. Psychometrika 28:33-41.

Mirkin, B.G., 1987: Additive clustering and qualitative factor analysis methods for similarity matrices. Journal of Classification 4:7-31.

Oldenbürger, H.-A., 1983: Clusteranalyse. S. 390-439 in: J. Bredenkamp/H. Feger (Hrsg.), Strukturierung und Reduzierung von Daten. Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie I: Forschungsmethoden der Psychologie, Band 4. Göttingen: Verlag für Psychologie/Hogrefe.

SAS Institute Inc., 1986: SUGI Supplemental Library User's Guide, Version 5 Edition. Cary, NC: SAS Institute Inc.

Schmidtchen, G., 1983: Jugend und Staat. Übergänge von der Bürger-Aktivität zur Illegalität. Eine empirische Untersuchung zur Sozialpsychologie der Demokratie. S. 105-437 in: U. Matz/G. Schmidtchen (Hrsg.), Gewalt und Legitimität. Analysen zum Terrorismus, 4/1. Opladen: Westdeutscher Verlag.

Shepard, R.N., 1980: Multidimensional scaling, tree-fitting, and clustering. Science 210:390-398.

Shepard, R.N./Arabie, P., 1979: Additive clustering: Representation of similarities as combinations of discrete overlapping properties. Psychological Review 86:87-123.

Torgerson, W.S., 1986: Scaling and Psychometrika: Spatial and alternative representations of similarity data. Psychometrika 51:57-63.

Uehlinger, H.-M., 1988: Politische Partizipation in der Bundesrepublik. Strukturen und Erklärungsmodelle. Studien zur sozialwissenschaftlichen Forschung, 96. Opladen: Westdeutscher Verlag.

Young, F.W., 1987: Multidimensional Scaling: History, Theory, and Applications. Hillsdale: Erlbaum.