

Clusteranalyse als Strukturanalyse

(Cluster Analysis as Structure Analysis)

Schmidt, H.: Clusteranalyse als Strukturanalyse. (Cluster analysis as structure analysis).
Int.Classif. 16(1989) No.1, p.15-23, 70 refs.

Starting out from a general methodology of data controlled structural analysis, consequences from this methodology are drawn for a suitable utilization of cluster analysis in empirical investigations, with particular attention being paid to the question: "Given my specific query, what method is best suited to my data in my empirical and theoretical contexts?". The basis for answering this question is a hierarchically ordered system of method classes and corresponding structure classes, with target-oriented comparisons of the results yielded by the methods of a given class making it possible to obtain a data-related reply. This, in turn, leads to new starting points for solving the cluster number problem, for cluster-analysis method validation strategies, and for the evaluation and interpretation of cluster solutions.

(Author)

1. Zur Problemstellung

Bereits seit Jahren gibt es hunderte verschiedener Methoden der Clusteranalyse, die sich unterschiedlichen Modelltypenzuordnungen lassen und auch auf unterschiedlichen Voraussetzungen beruhen (Abbildung 1). Diese Vielfalt ist vorwiegend durch die Operationalisierung des sehr allgemeinen heuristischen Problems entstanden, eine Objektmenge so in Cluster aufzuteilen, daß sich die Objekte innerhalb eines Clusters möglichst ähnlich und Objekte verschiedener Cluster möglichst "unähnlich" sind.

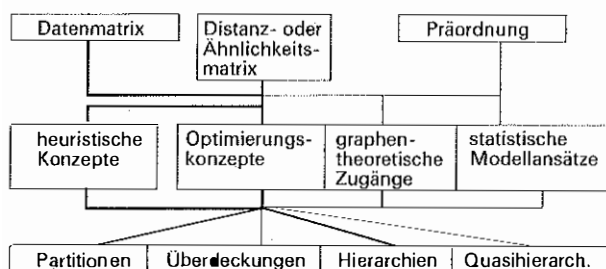


Abb. 1.: Übersicht über Ausgangspunkte, zugrundeliegende Konzepte und Strukturtypen verschiedener Methoden der Clusteranalyse. Dicke Linien stehen für die am häufigsten diskutierten Varianten.

Anwender aus empirischen Wissenschaften stehen damit vor folgendem *Adäquatheitsproblem*: Welche Methode ist meinen Daten unter meiner Fragestellung und in meinem empirischen und theoretischen Kontext ange-

messen? Weder Lehrbüchern noch Monographien zur Clusteranalyse oder Datenanalyse (z.B. 58, 61) noch zur Methodik empirischer Forschung in der Psychologie und anderen Sozialwissenschaften (z.B. 59, 64) geben hier erschöpfende und handhabbare Antworten. Clusteranalyse- und Klassifikationsaufgaben zielen auf Aspekte der Datenstruktur ab, so daß es sich anbietet, über theoretische Fundierungen zu einer allgemeinen Strukturanalyse auch die obige Frage für Clusteranalysen zu behandeln. Dies führt zu weitreichenden Schlußfolgerungen für die Entwicklung, Prüfung und Handhabung clusteranalytischer Methoden.

2. Grundlagen der Strukturanalyse

In (52) haben wir eine allgemeine Methodik für Strukturanalysen entwickelt, die das Adäquatheitsproblem bei datengesteuerten Strukturanalysen löst und zugleich einen Ansatz für eine spezielle Methodik experimenteller Strukturanalyse bietet. Ausgangspunkt ist eine allgemeine Fassung des Strukturbegriffs und eine Diskussion der Problem- und Hypothesenbildung bei Strukturanalysen.

Definition 1: Es seien I und J n - bzw. p -elementige Indexmengen. Weiterhin seien

$Z = \{z_i, i \in I\}$ die Menge der n verschiedenen unterscheidbaren Elemente eines bestimmten Untersuchungsobjekts O ,

$E = \{E_j, j \in J\}$ die Menge derjenigen Eigenschaften E_j der Elemente z_i , die der Untersuchung zugrundeliegen,

$$H = \bigcup_{q=2}^n H_q = \bigcup_{q=2}^n \left(\bigcup_{j \in J} H_{g_1, g_2, \dots, g_q} \right)$$

die nichtleere Menge aller q -stelligen ($2 \leq q \leq n$) Morphismen in Z bzgl. E , wobei

$$H_{ik} = H(z_i, z_k) = \{y_{ik}: z_i \rightarrow z_k\} \subseteq Z^2 \text{ und}$$

$$H_2 = \bigcup_{i,k \in I} H_{ik}, H_{ikm} = H(z_i, z_k, z_m) \subseteq Z^3 \text{ und}$$

$$H_3 = \bigcup_{i,k,m \in I} H_{ikm} \text{ usw.}$$

Dann heißt das geordnete Tripel $S = (Z, E, H)$ die *Struktur von O* und H die O bzgl. E *aufgeprägte Struktur*; Z heißt *Trägermenge* der Struktur S .

Definition 2: Gegeben ist eine Struktur $S = (Z, E, H)$. Jede Abbildung $f: Z \rightarrow (R^j)_{i,j \in I, 1 \leq j \leq q}$, der Menge Z in eine Folge von q -dimensionalen Räumen, so daß für $f(Z)$ in jedem betrachteten Raum eine Morphismenmenge \hat{H} als Schätzung für H existiert, heißt *Strukturanalyse*. Die Menge der Modellbedingungen und Optimalitätskriterien dieser Abbildung heißt *Strukturkonzept von f* , die Menge der numerischen und/oder nichtnumerischen Resultate dieser Abbildung heißt *Strukturmodell von H bzgl. f und E* .

Bemerkungen:

1. Unter *nichtnumerischen* Ergebnissen subsumieren wir solche Möglichkeiten der Ergebnisdarstellung wie (a) Graphen, (b) zwei- und dreidimensionale räumliche Darstellungen von Punktwolken und Vektoren, (c) Cluster-, Simplex- und andere Konfigurationen (12), (d) Liniendiagramme als graphische Skalen zur Abbildung hierarchischer oder verbandsgeordneter Begriffssysteme (70), (e) Feldliniendarstellungen (35) u.ä.

(b) gemäß (15) und zu deren allgemeinpsychologischer (a, b), differentialpsychologischer (b) und Einzelfalldeutung (a, b) (vgl. 28). Ist seine empirische Struktur, so haben wir den Fall der fundamentalen Messung, ist S bereits eine abgeleitete numerische Struktur, so haben wir den Fall der abgeleiteten Messung. Definition 2 kann damit zur Bestimmung eines *allgemeinen Begriffs der Messung* als Modellbildung mit formalisierten (i. a. numerischen) Systemen dienen, der den Meßbegriff klassischer Repräsentationstheorie etc. als Spezialfall enthält.

3. *Beispiel:* Betrachten wir gewöhnliche empirische Distanzmatrizen $D \subseteq H$, wobei $D \subseteq Z \times Z$, und $\bar{D} = \bar{H}$ soll dem Hybridmodell von Carroll/Pruzansky (8) folgen: $\bar{D} = D_1 + D_2 + \dots + D_m + D_{R_q}$, wobei D_1, \dots, D_m Distanzmatrizen von Baumstrukturen und D_{R_q} eine Distanzmatrix eines q-dimensionalen Raumes ist. Dann sind u. a. folgende Spezialfälle ableitbar:

- $D = D_1$, d. h. $i = 1$, wobei D_1 als Modellbedingung die Axiome einer Ultrametrik erfüllen soll. Die Abbildung f bestimmt dann die Klasse der *hierarchischen Clusteranalysen* (31).
- $D = D_2$, d. h. $i = 1$, wobei $D_1(0,1)$ -binäre und der multiplikativen Verknüpfungsregel $d_{ab} = \bigcap_{e \in \{1, \dots, p\}} x_{ae} \cdot x_{be}$ gemäß zerlegt werden kann. X ist die $(0,1)$ -binäre (n, p) -Zuordnungsmatrix der n Elemente x_a, x_b, \dots auf die p Cluster. Die Abbildung f bestimmt dann die Klasse der *partitionierenden Clusteranalysen* (33, 58).
- $D = D_3$, d. h. $i = 1$, und D_3 ist $(0,1)$ -binär und genügt der logisch-multiplikativen Verknüpfungsregel $d_{ab} = \bigcap_{e \in \{1, \dots, p\}} x_{ae} \cap x_{be}$. Die Abbildung f bestimmt dann die Klasse *überlappenden* (nichtdisjunkter, nichthierarchischer) *Clusteranalysen* (33, 48).

4. Das Strukturanalyseproblem besteht also darin, eine geeignete Abbildung $f: Z \rightarrow (R^J)_i$ zu finden, um ein theoretisch nützliches und empirisch tragfähiges Strukturmodell für S zu bilden. Das Strukturkonzept von f hat dabei den Charakter eines Systems von Hypothesen, deren Berechtigung zur Beschreibung des empirischen Materials erst geprüft und begründet werden muß. Dies entspricht eigentlich einer Verallgemeinerung der Repräsentations- und Bedeutsamkeitsprobleme der klassischen Repräsentationstheorie. Offensichtlich (vgl. Beispiel unter 3.) lassen sich die strukturanalytischen Methoden nach Umfang und Explikationsgrad ihres Strukturkonzepts hierarchisch ordnen, womit eine Hierarchie von *Modellstrukturen* korrespondiert. Je höher die Hierarchiestufe, desto weniger expliziert und damit restriktiv ist die Menge korrespondierender Modellstrukturen, die damit gefaßt werden können. Weiterhin ist von einer ganzen Reihe von Methoden bekannt, daß sie bei ganz bestimmten Datenstrukturen zu minimalen Klassifikations- und Strukturierungsfehlern führen, und daß sie den Daten in jedem Falle diese spezifischen Struktureigenschaften aufprägen (z. B. Ketten-Effekt der single-linkage-Methode). Dies führt zu folgender

Definition 3: Eine *Methodenklasse* $F = \{f_i | i = 1, \dots\}$ ist eine Menge strukturanalytischer Methoden f_i mit gleichen Strukturkonzepten und Abbildungseigenschaften. Eine *Strukturklasse* $T = \{t_j | j = 1, \dots\}$ ist eine Menge von Strukturmodellen t_j , die sich durch den gleichen Satz strukturbeschreibender Aussagen einer beliebigen Sprache charakterisieren lassen.

Im allgemeinen wird es bei der Methodenwahl kaum möglich sein, z. B. eine bestimmte Version des Austauschalgorithmus zur Minimierung des Varianzkriteriums aus fachwissenschaftlichen Gründen heraus favorisieren zu können. Man wird in der Hierarchie der Methodenklassen nicht bis zur differenziertesten Stufe hinabsteigen können. Die korrespondierende Strukturklasse wird also stets eine Menge von Strukturmodellen umfassen, aus denen für die Interpretation ein Modell auszuwählen bzw. zu bilden ist. Dieser *Unbestimmtheitseffekt* ist der eigentliche Kern des Adäquatheitsproblems und Gegenstand der Methodik der Strukturanalyse. Ausgangspunkt ist dabei das folgende

Adäquatheitspostulat:

Der Grad der Übereinstimmung der Resultate mehrerer Methoden einer Methodenklasse niedrigen Hierarchieniveaus ist ein Maß für die Adäquatheit der Beschreibung einer konkreten Datenstruktur durch die korrespondierende Strukturklasse, wenn gleichzeitig Divergenzartefakte (vgl. 28) ausgeschlossen wurden.

Im Mittelpunkt der in (52) entwickelten Methodik steht daher die Anwendung mehrerer Methoden einer Methodenklasse auf die Datenmenge. Stimmen die Lösungen hinreichend untereinander und mit dem Strukturkonzept der Methodenklasse überein, kann durch deren Aggregation ein endgültiges, interpretierbares Strukturmodell abgeleitet werden. Andernfalls bilden die verschiedenen Resultate nur verschiedene Diskussionen, die im explorativen Sinne vorsichtig interpretiert werden können, wenn nicht willkürlich eine Dissektion als Strukturmodell gedeutet wird.

3. Clusteranalyse als Strukturanalyse

3.1 Partitionierende Clusteranalysen

Die bekanntesten partitionierenden Clusteranalysemethoden optimieren ein vorgegebenes Gütekriterium, in dem eine (vorgegebene) Anfangspartition iterativ verbessert wird. Die Vielfalt der Herangehensweisen (vgl. Abb. 1) führte zur Entwicklung verschiedenster Gütekriterien, z. B. dem Varianzkriterium, Determinantenkriterium, Spurkriterium, Determinantenproduktkriterium, Kriterium der adaptiven Distanzen, Hotelling's Kriterium, Abstandssummenkriterium, L_1 -Kriterium. Alle diese Kriterien greifen auf die (quantitativen) Daten selbst zurück, basieren jedoch implizit auch auf einer bestimmten, nur für dieses Kriterium charakteristischen Distanzfunktion. Für drei der am besten untersuchten Kriterien sind in Tabelle 1 Abbildungseigenschaften und Aspekte der Strukturkonzepte zusammengefaßt (4, 6, 54, 58).

Alle drei Kriterien sind monoton fallende Funktionen der Clusterzahl. Die Verwendung von Streuungsmatrizen zur Definition der Gütekriterien impliziert ellipsenförmige Cluster; in den meisten praktischen Anwendungen dürfte dies unabhängig vom Robustheitsproblem dem Gesuchten am Nahesten kommen. Die für die Mustererkennung (pattern recognition) typische Problematik evtl. kompliziertester Clusterstrukturen (vgl. z. B. 60) ist für die hier angezielten Anwendungsbereiche nicht relevant. Es wird deutlich, daß die drei Kriterien aus Tabelle 1 Unterstrukturklassen der allgemeinen Strukturklasse "Partitionen" darstellen. Ein und dieselben Optimierungsalgorithmen können dabei allen korrespondierenden Methodenklassen zugeordnet werden, der Unterschied besteht in der Optimierung verschiedener Gütekriterien. Damit wird z. B. die viel diskutierte Frage, ob Minimaldistanz- oder Austauschalgorithmus "besser" im Sinne der Optimalität des Gütekriteriums sind, für unsere Anwendungsbereiche irrelevant. Die so definierten Strukturklassen können weiter zerlegt werden nach der 2. Von der Abbildung f wird nur Eindeutigkeit verlangt. Ist f so gar ein Homomorphismus und $(R^J)_i = R^J$, so führt das zum klassischen Meßbegriff der Repräsentationstheorie. Unterlegen wir dabei allgemein mehrsortige Strukturen, führt das zur Definition von Standardmessungen (a) und zusammengesetzter Messung

| Varianzkriterium | Determinantenkriterium | Kriterium der adapt. Dist. |
|---|---|---|
| 1. Cluster sind Hyperkugeln im R^m | 1. Cluster sind Hyperellipsen mit parallelen Achsen im R^m . | 1. Cluster sind Hyperellipsen beliebiger Ausrichtung im R^m . |
| 2.a.) $n_j \approx N/m$ oder b.) Es existiert ein Punkt $\sigma_0 \in R^m$ so, daß die Mittelpunkte der Cluster K_j um so weiter von σ_0 entfernt sind, je kleiner n_j ist. | 2. wie beim Varianzkriterium | 2. $n_j N/m$ |
| 3. Die Cluster werden (paarweise) durch Hyperebenen getrennt, d. h., die Cluster sind konvex und disjunkt | 3. wie beim Varianzkriterium | 3. Die Cluster werden (paarweise) durch Hyperflächen 2. Ordnung getrennt, d. h. die Cluster sind nicht notwendig konvex |
| 4. Die optimale Partition ist nur invariant gegenüber einem Vorzeichenwechsel beliebiger Variablen. | 4. Die optimale Partition ist invariant gegenüber jeder regulären Transformation der Daten. | 4. wie beim Determinantenkriterium |
| 5. Mit 1. wird impliziert, daß alle Variablen stochastisch sind, bzw. daß euklidische Distanzen zugrundeliegen | 5. Mit 1. wird impliziert, daß die Variablen stochastisch abhängig sind, diese Abhängigkeiten (und die entsprechenden Kovarianzmatrizen) aber in allen Clustern gleich sind, d. h. es liegen Mahalanobisdistanzen zugrunde. | 5. Wie beim Determinantenkriterium, aber es liegen infolge clusterspezifischer Abhängigkeiten clusterspezifische Mahalanobisdistanzen zugrunde. |

Tabelle 1: Eigenschaften ausgewählter Kriterien.

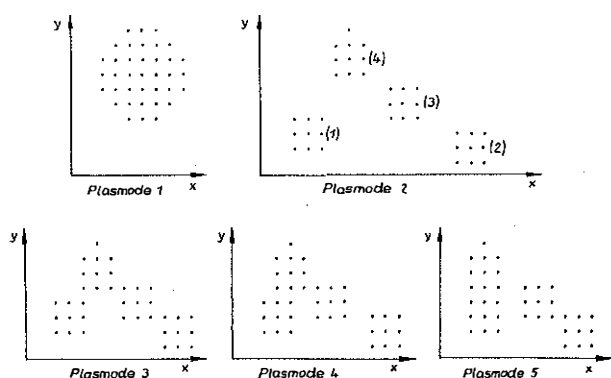


Abb. 2: Plasmoden 1 bis 5

vorgegebenen Clusterzahl, womit sich die Möglichkeit bietet, mittels der Methodik der Strukturanalyse einen Zugang zum Clusterzahlproblem zu finden. Zur Prüfung der Anwendung unserer allgemeinen strukturanalytischen Überlegungen auf Clusteranalysen und zu deren Demonstration führten wir verschiedene Methodenstudien durch. Dazu konstruierten wir fünf Plasmoden zu je 37 Objekten (vgl. Abb. 2), die verschiedene geometrische Eigenschaften aufwiesen. Insbesondere wurde der Abstand der vier Cluster aus Plasmode 2 systematisch variiert.

Damit war zugleich der Einfluß der Separiertheit der Cluster auf die Ergebnisse der Methoden prüfbar. Größere Anschaulichkeit wegen beschränken wir uns auf nur zwei Variablen, was zur prinzipiellen Demonstration genügt, selbst wenn bestimmte Effekte erst bei höherdimensionalen Daten auftreten. In einer zweiten Studie mit denselben Zielstellungen benutzten wir die in (58, S. 158 und 174) verwendeten acht Plasmoden unterschiedlichster Clusterstruktur, was einen weiteren Vergleich mit den Ergebnissen von SPÄTH gestattete.

Zur Durchführung der Studien für das Varianzkriterium implementierten wir unter Benutzung der Subroutinen MINDS 1 und HLCMB 1 (61) das Programm VARKRIST. Weiterhin erstellten wir u.a. mittels Subroutinen aus (58) das Programm KLASSPAS, in welchem ein Austausch- und ein Minimaldistanzalgorithmus zur Minimierung des Varianzkriteriums sowie ein Austausch- und ein modifizierter Minimaldistanzalgorithmus nach (6, S. 222f.) zur Minimierung des Determinantenkriteriums implementiert sind. In beiden Programmen ist also jeweils ein Minimaldistanz- und ein Austauschalgorithmus implementiert, VARKRIST benutzt eine Standardanfangspartition, KLASSPAS generiert (mehrere) zufällige Anfangspartitionen. Der Vergleich der Resultate beider Algorithmen für jede Clusterzahl erfolgt durch Zuordnung der Cluster nach der höchsten Anzahl gemeinsamer Elemente, sodaß insgesamt die Lösungsübereinstimmung maximiert wird, und Effekte durch verschiedene Clusternumerierungen eliminiert werden.

Folgende Schlußfolgerungen lassen sich ziehen:

1. Austausch- und Minimaldistanzalgorithmus stimmen für das Varianzkriterium bei Standardanfangspartitionen für k Cluster vollständig überein, wenn in den Daten k kugelförmige, ungefähr gleichgroße, disjunkte und hinreichend separierte Cluster vorliegen (Tab. 2, Plasmoden 2 und I).
2. Bei weniger klaren Datenstrukturen ist dies nicht mehr gewährleistet. Vollständige Übereinstimmung fehlt (2.1) oder tritt mehrfach auf (2.2). Im Falle von 2.2. läßt sich die Ambivalenz durch Vergleich der Eigenschaften der verschiedenen "guten" Lösungen mit den Eigenschaften der korrespondierenden Strukturklasse aufklären. Das bedeutet hier in erster Linie einen Vergleich der clusterspezifischen Merkmalsabhängigkeiten, jedoch auch der Besetzungsdichten, Clustervarianzen, Clusterabstände u.ä. Bei Plasmode 3 bis 5 sind die Variablen der 4-Cluster-Lösung (aus Tab. 2) jeweils innerhalb dieser Cluster unkorreliert, die Modelladäquatheit der entsprechenden Strukturmodelle ist damit klar. Bei den übrigen

| Clusterzahl k | Plas- mode 1 | Plas- mode 2 | Plas- mode 3 | Plas- mode 4 | Plas- mode 5 |
|------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 2 | 18 & | 2 | 1 x | / x | 1 x |
| 3 | 14 | 6 | 12 | / | 18 |
| 4 | 14 | / | / | / | / |
| 5 | 14 | 4 & | 4 & | 5 | 6 |
| 6 | 14 | 10 & | 5 & | 13 | 13 |
| 7 | 3 | 12 & | 11 | 12 | 14 |
| 8 | 10 & | 13 & | 7 & | 9 | 6 & |
| 9 | 11 & | 12 & | 11 & | 12 & | 15 |
| 10 | 13 & | 15 & | 10 & | 12 & | 7 & |

| Clusterzahl k | I N=37 | II N=41 | III N=44 | IV N=73 | V N=55 | VI N=50 | VII N=52 | VIII N=80 |
|------------------|-----------|------------|-------------|------------|-----------|------------|-------------|--------------|
| 2 | / | 1 x | / x | 33 | 3 | 23 | 5 | 23 |
| 3 | / | 10 | 19 | 27 | 1 x | 17 | 2 | 1 x |
| 4 | 4 | 5 | 3 | 17 | 7 | 1 x | 3 | 5 |
| 5 | 12 | 2 | 8 | 17 | 7 | 13 | 10 | 25 |
| 6 | 11 | 9 | 10 | 13 | 13 | 13 | 1 x | 13 |

Tabelle 2: Ergebnisse des Programms VARKRIST für die fünf Plasmoden aus Abh. 2 (oberer Teil) und die 8 Plasmoden aus (51) (unterer Teil). Angegeben sind die Objektzahlen, die von beiden Algorithmen verschiedenen Clustern zugeordnet wurden, nachdem die Cluster einander nach der höchsten Anzahl gemeinsamer Elemente zugeordnet wurden. &: Mindestens eine Methode erzeugte mindestens ein leeres Cluster. x: Übereinstimmung (fast) ideal, aber Ergebnisse entsprechen nicht der Datenstruktur, da mindestens eins der vorgegebenen Cluster zerrissen wird. Plasmode I: 2 oder 3 disjunkte, ungefähr kugelförmige Cluster. Plasmode III: ein Cluster mit einem randständigen kompakten Kern, das sich kontinuierlich verdünnt, Plasmode VII: 2 überlappende, ellipsoidale Cluster verschiedener Lage im Raum.

| Clusterzahl k | Plasmode 4 (a) (b1) (b2) | Plasmode 5 (a) (b1) (b2) | Plasmode VII (a) (b1) (b2) |
|------------------|-----------------------------|-----------------------------|-------------------------------|
| 2 | 2 20 20 | / 20 7 | / 18 2 |
| 3 | / 20 9 | / 15 10 | / 20 2 |
| 4 | / 20 10/19 | / 15 17 | / 3 6 |
| 5 | / 11 1 | / 4 4/19 | / 11 11/17 |
| 6 | / 7 11/16 | 2 2/15 | 18 13 1 |

| | | | |
|---|------------|-----------|-----------|
| 2 | 1 13 2 | / 14 5 | / 6 1 |
| 3 | 13 9 1 | / 1 2 | 7 7 1/19 |
| 4 | / 20 11/19 | / 18 7/17 | / 2 2/17 |
| 5 | / 12 1/15 | / 11 1/11 | / 2 4/12 |
| 6 | 3 6 1/8 | 8 5 1/6 | 17 1 1/10 |

Tabelle 3: Resultate des Varianzkriteriums (oberer Teil) und des Determinantenkriteriums (unterer Teil) des Programms KLASSPAS für die Plasmoden 4, 5 und VII. (a): Zahl der Objekte, die von beiden Algorithmen bei den jeweils besten Zielfunktionswerten verschiedenen Clustern zugeordnet werden, nachdem die Cluster einander nach der höchsten Anzahl gemeinsamer Elemente zugeordnet wurden; (b): Zahl der Anfangspartitionen (von insgesamt 20) mit dem geringsten Zeitfunktionswert der Endpartitionen, und zwar (b1) für den Austauschalgorithmus, (b2) für den (modifizierten) Minimaldistanzalgorithmus.

für diese Plasmoden implizierten Lösungen zeigen sich jedoch clusterspezifische Merkmalsabhängigkeiten, d.h. man muß ellipsoidförmige Cluster für die 2-bzw. 3-Cluster-Lösung annehmen. Der (praktisch übliche) alleinige Einsatz des Varianzkriteriums kann hier keine Entscheidung liefern. Die Auswahl der adäquaten Clusterung wird nur durch den zusätzlichen Einsatz anderer Struktur- und Methodenklassen möglich. Bei Plasmode III entsprechenden Besetzungsdichten und Clusterva-

rianzen der 2-Cluster-Lösung nicht der Modellstruktur. Eine detaillierte Prüfung führt zur adäquaten Datenstruktur (vgl. Tab. 2). Im Falle von 2.1 liegt die Notwendigkeit des Einsatzes anderer Struktur- und Methodenklassen erst recht auf der Hand.

3. Die Benutzung mehrerer zufälliger Partitionen bringt bezüglich unserer Fragestellung nach dem Auffinden adäquater Strukturmodelle keinen Gewinn. SPÄTH (58, S. 166) ist zuzustimmen, daß bei kanonischen Strukturen (fast) alle Anfangspartitionen den gleichen Zielfunktionswert erreichen. Wie die in (58) berichteten Resultate sowie unsere Tabelle 3 zeigen, kann jedoch eine Häufung von besten Zielfunktionswerten, v.a. beim Austauschalgorithmus bei verschiedenen Clusterzahlen auftreten, so daß nicht notwendig die Auswahl der adäquaten Lösung aus einer Lösungsserie für mehrere Clusterzahlen erlaubt wird. Auch der Vergleich von Austausch- und Minimaldistanzalgorithmus für die beste von mehreren aus zufälligen Anfangspartitionen erzeugten Lösungen erlaubt nicht die Auswahl der adäquaten Lösung, da unabhängig von der Datenstruktur die Wahrscheinlichkeit einer Übereinstimmung der Resultate erhöht wird (vgl. Tab. 3). Wird einfach die Lösung mit dem besten Zielfunktionswert interpretiert, gibt es überhaupt keinen Hinweis darauf, ob nur eine Dissektion oder eine Struktur aufgedeckt wurde.

Zusammengefaßt: Wenn die Zielstellung Strukturanalyse (und nicht nur Dissektion) für eine Anwendung partitionierender Methoden auf gegebene Daten relevant ist, sollten folgende Schritte ausgearbeitet werden: 1. Durch Explikation der fachwissenschaftlichen Theorie über den zu analysierenden Gegenstand und schrittweisen Vergleich mit Strukturkonzepten bekannter Methoden ist eine geeignete Methodenklasse auszuwählen (bzw. de-

ren Erarbeitung zu beginnen, falls eine solche nicht vorliegt). 2. Wenigstens zwei Methoden der gewählten Methodenklasse sind auf die Daten anzuwenden. 3. Die Resultate dieser Methoden sind auf geeignete Weise untereinander und mit dem Strukturkonzept der Methodenklasse zu vergleichen. 4. Im Falle genügend großer Übereinstimmung Ableitung eines adäquaten Strukturmodells, anderenfalls Rückkehr zu Schritt 1.

Die Art und Weise des Vergleichs, die Baselines für die Übereinstimmungsgüte und die Art und Weise der Aggregation können nur methodenklassenspezifisch festgelegt werden. Dafür sind weitere Methodenuntersuchungen nötig, wie wir sie für das Varianzkriterium oben angedeutet haben. Leider gibt es eine Unmenge von Modellen, deren Abbildungseigenschaften nicht genau bekannt sind und die sich deshalb nicht sinnvoll in ein hierarchisches System von Methodenklassen eingliedern lassen. Darüber hinaus haben verschiedene Untersuchungen (z.B. 11, 23) die mangelnde Reliabilität einzelner Methoden nachgewiesen. CA-Methoden, die zu sensibel auf trivialste Datenänderungen reagieren, oder aber Ergebnisse relativ unabhängig von den konkreten Ausgangsdaten generieren, können jedoch nicht sinnvoll im empirischen Forschungsprozeß Verwendung finden. Derartige Probleme werden in den einschlägigen Programmbeschreibungen der Rechenzentren i.a. nicht sichtbar.

3.2 Hierarchische Clusteranalysen

Divisive hierarchische Methoden werden in der einschlägigen Literatur nur selten behandelt. Zumeist beruft man sich bei einer "Kurzbeschreibung" darauf, daß divisive Methoden weit mehr Rechenzeiten brauchen als agglomerative und Näherungslösungen, wie die von FORTIER/SOLOMON (20), die wenig befriedigend sind (51). Ein Teil divisiver Methoden, wie die "intercolumnar correlational analysis" (45), sind durch zweifelhafte Anwendbarkeit und mangelnde Bekanntheit spezifischer Abbildungseigenschaften gekennzeichnet. Diese Kritikpunkte treffen auf agglomerative Methoden weniger zu. Hier gibt es eine Reihe sehr bekannter Methoden (Single Linkage, Complete Linkage, Average Linkage, Weighted Average Linkage, Median-Verfahren, Centroid-Verfahren, Ward's Verfahren, Flexible Strategie nach LANCE/WILLIAMS (39)), die sehr breit untersucht und angewendet worden sind. Gängige Anwendungspraxis ist dabei die Vorgehensweise, mehrere dieser Methoden zu probieren und das geeignetste Resultat auszuwählen. Wie ist diesem Umstand unter Beachtung der methodischen Anforderungen an eine Strukturanalyse zu begegnen? Es ist aus theoretischen und empirischen Untersuchungen hinreichend bekannt, daß die agglomerativ-hierarchischen Methoden recht unterschiedliche Eigenschaften haben und zu recht unterschiedlichen Ergebnissen führen (z.B. 57). Single Linkage und Complete Linkage basieren z.B. auf direkt entgegengesetzten Verschmelzungsstrategien. Innerhalb der allgemeinen Strukturklasse "Hierarchien" scheinen die agglomerativen Methoden mit verschiedenen, abgrenzbaren Unterklassen zu korrespondieren, die durch wesentlich verschiedene Eigenschaften charakterisierbar sind. Ein Metho-

denvergleich über diese Unterklassen hinweg wird damit nutzlos, selbst wenn bestimmte Seiten der Datenstruktur (z.B. die Existenz von Außenseitern) in mehreren Lösungen zum Vorschein kommt. Genau das bleibt in der oben angeprangerten Anwendungspraxis unberücksichtigt. Ein Ausweg im Sinne unserer Methodik der Strukturanalyse bietet sich durch die Flexible-Strategie nach LANCE/WILLIAMS an, die über Variation eines Parameters eine ganze Verfahrensschar umfaßt und die bekannten Methoden wie Single Linkage simuliert. Durch systematische Variation des Parameters läßt sich (1) die Stabilität einer gefundenen Gruppierung überprüfen, (2) der Charakter des Verfahrens ändern. Damit müßte es möglich sein, wenigstens vier der Strukturklasse "Hierarchien" untergeordnete Strukturklassen zu definieren, die durch die wesentlichen Eigenschaften bestimmter Methoden ausgezeichnet sind:

- (I) 1. "Raumkontrahierend", d.h. Bildung von wenigen großen Clustern,
2. Kettenbildungseffekt, d.h. eine einzige (zufällige) kleine Distanz erzwingt Fusion (Beispiel: Single Linkage)
- (II) 1. "Konservativ", d.h. zwischen Single und Complete Link einzuordnen,
2. weder Kettenbildungs- noch Lassoэффект, da Fusion auf dem Durchschnitt vieler Distanzen beruht,
3. auf alle Distanzmaße anwendbar, für die der Mittelwert eine sinnvolle Größe ist, (Beispiel: Average Linkage)
- (III) 1. Tendiert zur Bildung gleichgroßer Cluster,
2. auf beliebige Distanzmaße anwendbar, weder Lasso- noch Kettenbildungseffekt,
3. entspricht bei quadrierten euklidischen Distanzen einer stufenweisen Minimierung des Varianzkriteriums, (Beispiel: Ward's Methode)
- (IV) 1. "Raumdilatierend", d.h. Bildung vieler kleiner Cluster,
2. Lassoэффект, d.h. eine einzige (zufällige) große Distanz verhindert Fusion (Beispiel: Complete Linkage).

Mit Hilfe der Flexiblen Strategie stehen dann für jede dieser Strukturklassen (wenigstens) zwei Methoden zur Verfügung und die Adäquatheit eines bestimmten Dendogramms wird überprüfbar. Es eröffnen sich so völlig neue Möglichkeiten für Vergleichsuntersuchungen zu den Methoden und zu empirisch-psychologischen Fragestellungen. Die derzeitige Situation auf diesem Gebiet ist durch unterschiedlichste und unterschiedlich begründete Vergleichskriterien verwirrend und unbefriedigend (vgl. z.B. 25).

4. Konsequenzen

4.1 Zum Clusterzahlproblem

Die Festlegung der "richtigen" Clusterzahl ist ein Kernproblem in den meisten Clusteranalyse-Anwendungen. Die vielfältigen in der Literatur vorgeschlagenen Strategien zur Lösung des Clusterzahlproblems lassen sich vier Richtungen zuordnen:

1. *heuristische Vorschläge*: z.B. deutet die Entstehung leerer Cluster bei partitionierenden Methoden an, daß die falsche Clusterzahl gewählt wurde (43, 58).
2. *graphische Hilfen*: z.B. Plotten der Kriteriumswerte, der Inner-Cluster-Distanzen u.a. über die Clusterzahl analog zum Scree-Test der FA u.ä. (17, 50), evtl. Vergleich dieser Kurven mit Kurven der Kriteriumswerte bei Zufallszahlen (13).
3. *statistische Hilfen*: z.B. mittels verschiedener F-Tests, Likelihood-Ratio-Tests (2, 7, 40, 41, 68).
4. *Verwendung von Außenkriterien*: v.a. Verwendung von nicht in die Clusteranalyse einbezogenen Variablen (67).

Nach (17, 49) liegen die Hauptschwierigkeiten statistischer Kriterien (a) in der Bestimmung geeigneter Nullhypothesen, (b) in der Bestimmung der Stichprobenverteilungen der verschiedenen Distanz- und Ähnlichkeitsmaße, (c) in der Entwicklung flexibler Testverfahren. Außerdem sind die meisten statistischen Tests nur innerhalb enger Grenzen optimal oder erfüllen bestimmte postulierte Voraussetzungen doch nicht (5). Darin liegt auch eine Hauptursache für die immer wiederkehrende Diskussion von "dissection vs. classification" (10, 17, 18) bzw. Cluster-Formation vs. Cluster-Analyse (58). Heuristische und graphische Kriterien sind stark subjektiven Einflüssen unterworfen (4). Darüber hinaus erlaubt die oft benutzte Betrachtung der Kurve der Kriteriumswerte bzw. die Suche nach dem größten Kriteriumswertabfall beim Übergang von einer k -Cluster-Lösung zu einer $(k+1)$ -Cluster-Lösung nicht notwendig die Auswahl der adäquaten Lösung, weil damit höchstens innerhalb einer evtl. inadäquaten Strukturklasse das beste Strukturmodell gefunden werden kann. Folgerichtig implizieren alle Kurven in Abb. 3 die Strukturmodelle mit 4 Clustern als richtige Lösung, d.h. speziell die eindeutig ellipsoidale Clusterstruktur der Plasmode 5 bleibt unerkannt.

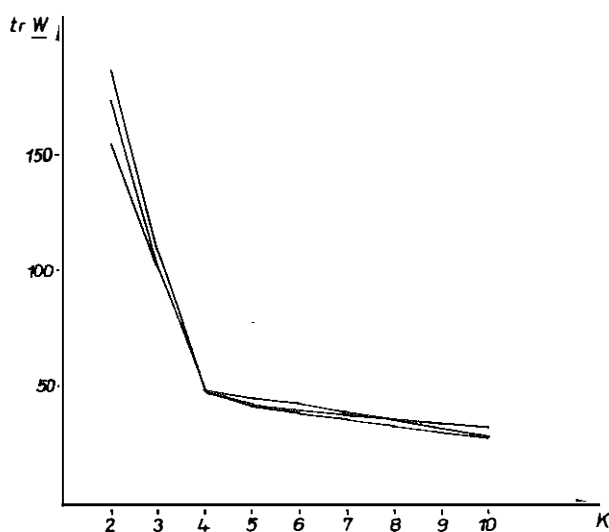


Abb. 3: Kurven der Kriteriumswerte des Austauschalgorithmus für das Varianzkriterium, Programm VARKRIST, Plasmoden 3, 4, 5.

Die Verwendung von Außenvariablen steht vor dem Dilemma, eine Clusterlösung mit Variablen beurteilen zu wollen, die für das jeweilige Strukturanalyseproblem nicht relevant sind oder anderenfalls relevante Variablen nicht in den Analyseprozeß einfließen zu lassen (Divergenzartefakt-Problematik). Eine befriedigende Lösung des Clusterzahlproblems steht also bisher noch aus, was sich in den widersprüchlichen Ergebnissen entsprechender Methodenexperimente niederschlägt. Praktisch wird meist aus mehreren Lösungen verschiedener Clusterzahl, die eventuell alle die Datenstruktur nur inadäquat wiedergeben, mit teilweise ausgefeiltesten statistischen Techniken die Lösung mit der "richtigen" Clusterzahl ausgewählt, *ohne* die Strukturadäquatheit überprüft zu haben. Gemäß obiger Überlegungen ist das Clusterzahlproblem jedoch nur im *Zusammenhang* mit dem Strukturtyp- und dem Methodenproblem lösbar und sollte als ein Aspekt der Adäquatheitsproblematik betrachtet werden. Damit werden scheinbare Paradoxien auflösbar, wie etwa im Beispiel von (58, 167ff.), in welchem die optimalen Partitionen vom Abstand der Punkthaufen abhängen.

4.2 Kritik herkömmlicher Validierungsstrategien für Clusteranalyse-Methoden

Seit Beginn der 70er Jahre wurden immer wieder Studien zur Validierung und Evaluation von Clusteranalyse-Methoden publiziert. Prinzipiell lassen sich je nach Art der verwendeten Daten drei Typen solcher Studien differenzieren: Evaluationen (1) mittels empirischen Datensätzen, (2) mittels Plasmoden und (3) mittels Datensätzen, die über Monte-Carlo-Methoden erzeugt wurden. Empirische Datensätze haben i.a. den Nachteil, daß ihre Datenstruktur nicht zweifelsfrei bekannt ist und sich deshalb nicht als Außenkriterium einer Validierung eignet. Um dies zu umgehen, wählten verschiedene Autoren empirische Datensätze *bekannter* Struktur. Am bekanntesten sind die Kraftfahrzeugdaten von BAUMANN (1) mit der Typeneinteilung der Kraftfahrzeuge als Kriteriumsstruktur geworden. In (29) werden mehrere geschlechtsspezifisch beantwortete MMPI-Items und die Zweiteilung der Vpn-Stichprobe nach dem biologischen Geschlecht als Kriteriumsstruktur benutzt. Solche Versuche bergen offensichtlich stets die Gefahr in sich, daß die intendierte und die tatsächliche benutzte Kriteriumsstruktur nicht identisch sind. Diese Nachteile können nur durch die Konstruktion geeigneter Datensätze mit bekannter Struktur ausgeschaltet werden. Zum einen können diese Daten "per Hand" vom Untersucher konstruiert werden, z.B. im gewöhnlichen zweidimensionalen Merkmalsraum, zum anderen sind in den letzten Jahren Monte-Carlo-Datensimulationen häufig verwendet worden. Bei Monte-Carlo-Plasmoden taucht nun trotz des mathematischen Aufwandes das Problem auf, daß die geometrischen Strukturen nicht mehr genau bekannt sind, sondern nur gewisse Verteilungs-, Überlappungsparameter u.ä. Dies muß u.E. als Nachteil betrachtet werden, der auch die teilweise Unvergleichbarkeit und Widersprüchlichkeit entsprechender Studien mitverursacht. Wir sehen in folgenden Punkten Hauptmängel bis-

herigen Vorgehens bei Validierungsstudien zu Clusteranalyse-Methoden (vgl. auch 46, 53):

- (1) Nach wie vor sind solche Validierungsstudien als Suche nach *einem* "besten" Verfahren angelegt, obwohl u.a. anerkannt wird, daß die verschiedenen Methoden unterschiedliche Abbildungseigenschaften besitzen.
- (2) Meist wird eine Partition als Kriteriumsstruktur gewählt; geprüft werden v.a. aber hierarchische Verfahren, für die praktisch kein exakter Test zur Auswahl einer Partition existiert.
- (3) Unabhängig vom Strukturtyp der Testdaten werden Methoden verschiedenster Methodenklassen geprüft. Wenn eine Methode unter bestimmten Bedingungen gute Ergebnisse erzielt, ist dennoch nicht gesichert, daß diese Ergebnisse bei der praktischen Anwendung aus einer Serie von Lösungen als gut erkannt werden.
- (4) Teilweise werden Ergebnisse verglichen, ohne die spezifischen Eigenschaften evtl. unterschiedlichster Datensätze in den Vergleich einzubeziehen.

Im Zusammenhang mit der Methodik der Strukturanalyse wird damit eine gezielte Reinterpretation bisheriger Studien unter dem Aspekt adäquater Strukturerkennung nötig. Validierungsstudien sollten prinzipiell methoden- und strukturklassenspezifisch angelegt werden.

4.3 Zur Evaluation von Clusterlösungen

Die Literatur der letzten Jahre enthält eine Fülle von Vorschlägen zur Evaluation einer Clusterlösung; z.T. sind diese identisch mit Tests zur Lösung des Clusterzahlproblems, so daß auch die dort geübte Kritik weiter zutrifft. Solange über die Strukturadäquatheit einer bestimmten Clusterlösung nichts ausgesagt werden kann, ist in praktischen Anwendungen die Benutzung von Evaluationsmaßen und -methoden eigentlich unsinnig, da evtl. inadäquate Strukturmodelle in ihrer "Güte" beurteilt werden sollen. Eine Reihe von Autoren (32, 61, 62) schlagen den Einsatz von Diskriminanz- und Varianzanalysen zur Beurteilung von Clusterlösungen vor. Der standardmäßige Anschluß dieser Methoden an eine Clusteranalyse ist u.E. allerdings kaum sinnvoll, denn: (1) Lösungen von Clusteranalyse-Methoden, die konvexe, durch Hyperebenen getrennte Cluster erzwingen, *müssen* bei nachfolgender (linearer) Diskriminanzanalyse *stets* gut abschneiden. (2) Andere Clusterstrukturen müssen bei linearer Diskriminanzanalyse schlecht abschneiden, wenn die Lösung die tatsächliche Clusterstruktur adäquat beschreibt. (3) die varianzanalytische Prüfung der wechselseitigen Isoliertheit der Cluster zu deren Validierung als *Struktureinheiten* setzt gerade das voraus, was geprüft werden soll: die Übereinstimmung der Clusterlösung mit der tatsächlichen Struktur. Die Signifikanz eines Mittelwertunterschieds ist aber kein Indikator für die Clustervalidität, sondern kann auch eine artifizielle Aufteilung der Objektmenge "bestätigen". Nur unter der Voraussetzung, daß entsprechend unserer Methodik der Strukturanalyse vorgegangen und auf diese Weise Artefakte vermieden wurden, ist der Einsatz von Diskriminanz- und Varianzanalyse nutzbringend. In erster Linie

lassen sich folgende Fragen damit beantworten: (1) wie groß ist die wechselseitige Isoliertheit von Clustern, die mitteln Varianz- oder Determinantenkriterium gefunden wurden? (2) Beinhaltet eine Lösung nach dem Kriterium der adaptiven Distanzen überlappende Cluster? (Wenn nicht, so lassen sich auch ungleich große Ellipsen verschiedener Hauptachsenrichtung durch Hyperebenen trennen.) (3) Wie groß ist die Diskriminiertheit und Isoliertheit von Clustern in *anderen*, nicht in die Clusterbildung eingegangenen Variablen? Mit dieser Frage geht es um das Aufspüren von differentiellen Zusammenhängen zwischen Merkmalsgruppen unter Umgehung der Korrelationsstatistik, die z.B. in (3, 51, 63) stark problematisiert wird. (4) Wie ist der (univariate) Beitrag jeder einzelnen Variablen zur Klassifizierung? Läßt sich die Clusterung eventuell durch die klassifikatorische Wirkung von ein oder zwei "Trenn"-Variablen erklären? Konnte über unsere Methodik der Strukturanalyse die Adäquatheit einer Clusterlösung nachgewiesen werden, so bieten sich verschiedenere Möglichkeiten zur differenzierten Interpretation und nachfolgenden konfirmativen Analyse an. In (44) wird ein Ansatz entwickelt, der die Methode der Kreuzvalidierung für die Clusteranalyse erstmals explizit nutzt. In (9, 14) u.a. werden Permutationsexperimente mit simulierten Datensätzen durchgeführt. In (16, 26, 27, 36, 27) u.a. wird das von HUBERT/SCHULTZ (38) für die Datenanalyse nutzbar gemachte Paradigma der quadratischen Zuordnung als allgemeine nonparametrische Teststrategie weiter entwickelt und konkrete Anwendungsmöglichkeiten aufgezeigt. Diese neuen Ansätze lassen interessante Weiterentwicklungen und neue Einsatzmöglichkeiten für die Datenanalyse allgemein und die Clusteranalyse im Besonderen erwarten.

4.4 Zum Interpretationsproblem

Die Verwendung der Clusteranalyse zur Lösung strukturanalytischer Fragestellungen erfordert die Behandlung der Frage, wie die gewonnene Clusterung der Objektmenge inhaltlich zu interpretieren ist, wie die gefundenen formalen Relationen und Strukturen adäquat in fachwissenschaftlich relevante Relationen und Strukturen "übersetzt" werden können. Die zur Verfügung stehenden Maße zur Interpretation der Lösungen müssen allerdings nicht notwendig zur sachgerechten Beantwortung der inhaltlichen Fragen beitragen. Üblicherweise schließt sich z.B. an eine partitionierende Clusteranalyse ein Profilvergleich der Clustermittelwerte, evtl. noch unter Einschluß der entsprechenden Steuungen, an. Dieses Vorgehen ist allerdings nur dann wirklich sinnvoll, wenn nachgewiesen werden konnte, daß diese Mittelwerte eine adäquate Beschreibung der untersuchten Struktur ermöglichen, d.h., daß die Struktur durch kugelförmige, disjunkte Cluster gekennzeichnet ist. Die Mittelwertvektoren minimieren die Zielfunktionen des Clusteranalysekriteriums, des Determinantenkriteriums wie auch des Kriteriums der adaptiven Distanzen. Daraus ergibt sich aber nicht gleichzeitig deren Eignung zur *Beschreibung* der Clusterung. Bei zwei ellipsoidförmigen, einander wie ein vierflügeliger Propeller überlappenden Clustern im

2-dimensionalen Merkmalsraum stimmen die Mittelwerte sowohl untereinander als auch mit denen der Gesamtmenge überein, sind also zur Beschreibung der eindeutigen Struktur ungeeignet. Unter der Voraussetzung, daß solche Strukturen überhaupt erkannt werden (was mit dem paradigmatisch verwendeten Varianzkriterium nicht möglich ist), müssen andere Methoden zur Beschreibung der Cluster angewandt werden. Denkbar sind z.B. die clusterspezifischen Kovarianz- bzw. Korrelationsmatrizen, die clusterweise Bestimmung von Regressionsgeraden, die Visualisierung der Struktur in den ersten beiden Hauptkomponenten u.a. Der unkritische Einsatz der Teststatistik zur Prüfung von Mittelwertunterschieden kann trotz "signifikanter" Ergebnisse die Tatsache eines Artefakts verschleiern. Lassen sich Artefakte ausschließen und sind Mittelwertprofile geeignete Beschreibungsmittel, so sollte man deren Möglichkeiten (vgl. 34, 47, 51) auch nutzen. Dasselbe gilt für alle Maße zur Beschreibung von Homogenität und wechselseitiger Isoliertheit der Cluster, z.B. Clusterradien, das Disjunktheitsmaß (55, 56) u.ä. Noch relativ unbekannt ist das Konzept der L- und L-Cluster (30, 39), das gute interpretatorische Möglichkeiten bietet. Weiterhin sind die vielfältigen im deutschen Sprachraum noch kaum bekannten und eingesetzten graphischen Methoden zur Lösungsbeschreibung verwendbar, wie z.B. die Polar- und Kreis-Linien-Diagramme, polygon plots), Metroglyphen, Andrew-Kurven (auch: function plots), Chernoff-Faces, verschiedene Arten von Streudiagrammen u.ä. (vgl. z.B. die Übersichtsreferate in (19, 65, 66).

5. Zum Einsatz der Clusteranalyse in empirischen Untersuchungen

Die Relevanz unseres Ansatzes für praktische empirische Untersuchungen, z.B. in den Sozialwissenschaften, soll hier nun an einem Beispiel skizziert werden. Seit langem wird in der klinisch-psychologischen Forschung versucht, Zusammenhänge von Persönlichkeitsmerkmalen und psychopathologischen Syndromen aufzudecken. Ziel derartiger Bemühungen ist es, über personale Bedingungen der Genese psychischer Störungen zu wirksameren Prophylaxe- und Therapieprogrammen zu kommen. Dabei wurde wiederholt festgestellt, daß es die Suizid-, Sucht- oder Neurosepersönlichkeit nicht gibt. Ein größerer Gewinn besteht daher in der Auffindung von Subgruppen von Neurotikern, Alkoholikern usw., die sich in bestimmten Merkmalskombinationen unterscheiden. Die Therapierelevanz bisheriger Ergebnisse ist aber oft fragwürdig, die in (21) angebotene 4-Cluster-Gruppierung für Alkoholiker ($N = 135$) fordert z.B. allein aus methodischer Sicht folgende Diskussion heraus:

- (1) es ist nicht gesichert, daß die interpretierte Lösung der tatsächlichen Datenstruktur entspricht (es wird nur die Methode nach WARD benutzt)
- (2) eine Reanalyse in (22) mit zwei weiteren CA-Methoden ergab insges. drei bzgl. der Variablenprofile der Cluster, bzgl. der Clustergrößen und bzgl. des Außenkriteriums differentieller Rückfallquoten verschiedene Lösungen.

D.h., es gibt verschiedene Lösungen, die zu unter-

schiedlichen inhaltlichen Interpretationen und diagnostischen sowie therapeutischen Konsequenzen führen. Für relevante Resultate solcher Klassifikation sind also neben einer komplexeren Untersuchungsplanung auch Adäquatheitsfragen bei der Datenauswertung zu berücksichtigen. Letztere sind durch strukturanalytische Herangehensweisen handelbar.

Quellen:

- (1) BAUMANN, U.: Psychologische Taxometrie. Bern: 1971
- (2) BEALE, E.M.L.: Euclidean cluster analysis. In: Bull. Int.Stat.Inst. 43(1969)p.92-94.
- (3) BERG, M.: Zum Problem der Verallgemeinerbarkeit inter- und intraindividuelle Unterschiede. In: SCHRÖDER, H. (Ed.): Psychologie der Persönlichkeit und Persönlichkeitsentwicklung, Berlin: Ges. f. Psychologie der DDR 1982. p.193-201.
- (4) BERGS, S.: Optimalität bei Clusteranalysen. Dissertation, Münster 1981.
- (5) BINDER, D.A.: Bayesian cluster analysis. In: Biometrika 65(1978)p.31-38.
- (6) BOCK, H.H.: Automatische Klassifikation. Göttingen: Vandenhoeck & Ruprecht 1974. p.480.
- (7) BÜTTNER, J.: Zur Clusteranalyse. In: Biom.Z. 17(1975)p.163-179.
- (8) CARROLL, J.D., PRUZANSKY, S.: Discrete and hybrid scaling models. In: LANTERMANN, E.D., FEGER, H. (Eds.): Similarity and Choice. Bern: Huber 1980. p.108-139.
- (9) COHEN, A., GNANADESIKAN, R., KETTENRING, J.R., LANDWEHR, J.M.: Methodological developments in some applications of clustering. In: KRISHNAIAH, P.R. (Ed.): Applications of statistics. Amsterdam: North-Holland 1977. p.142-162.
- (10) CORMACK, R.M.: A review of classification. In: J.Royal Statist. Soc. (Series A) 134(1971)p.321-367.
- (11) DEGENS, P.O., FEDERKIEL, H.: Clusteranalyse in der Medizin. In: EHLERS, C.T., KLAR, R. (Eds.): Informationsverarbeitung in der Medizin, Wege und Irrwege. Berlin: Springer 1979. p.459-464.
- (12) DEGERMANN, R.L.: The geometric representation of some simple structures. In: SHEPARD, R.N. et al (Eds.): Multidimensional Scaling. New York: Seminar Press 1972.
- (13) DOBBENER, R.: Zur Skalen- und Translationsinvarianz von Metriken. In: Int.Classif. 8(1981)No.2, p.64-68.
- (14) DOBBENER, R.: Grundlagen der numerischen Klassifikation anhand gemischter Merkmale. Dissertation, Bamberg 1982.
- (15) DUCAMP, A., FALMAGNE, J.C.: Composite measurement. In: J.of Math.Psychol. 6(1969)p.359-390.
- (16) ECKES, T.: Ein nonparametrischer Test für die Ähnlichkeit zwischen Aufteilungen einer Objektmenge. In: Psychol. Beitr. 24(1982)p.76-85.
- (17) EVERITT, B.S.: Unresolved problems in Cluster Analysis. In: Biometrics 35(1979)p.169-181.
- (18) FLEISS, J.L., LAWLOR, W., PLATMAN, S.R., FIEVE, R.R.: On the use of inverted factor analysis for generating typologies. In: J.Abnorm.Psychol. 77(1971)p.127-132.
- (19) FLURY, B., RIEDWYL, H.: Graphical Representation of Multivariate Data by Means of Asymmetrical Facets. In: J.Am.Stat.Ass. 76(1981)p.757-765.
- (20) FORTIER, J.J., SOLOMON, M.: Clustering procedures. In: KRISHNAIAH, P.R. (Ed.): Multivariate Analysis I. New York: Academic Press 1966. p.493-506.
- (21) FUNKE, J., KOPP, B.: A comparison of 3 methods of cluster analysis: the H means/K means-algorithm, the fuzzy set partition and the procedure of Ward. In: Psycholog. Beitr. 24(1982)p.286-295.
- (22) FUNKE, J., KLEIN, M., SCHELLER, R.: Zur Klassifikation von Alkoholikern durch Persönlichkeitsmerkmale. In: Psycholog. Beitr. 23(1981)p.146-158.
- (23) FUNKHOUSER, G.R.: A note of the reliability of certain clustering algorithms. J.Marketing Res. 20(1983)p.99-103.
- (24) GEDIGA, G.: Ein Abbruchkriterium für hierarchische Clusteranalysen. Psychol. Forschungsberichte, Univ. Osnabrück, Nr.20, 1980a.
- (25) GEDIGA, G.: Hierarchische Klassifikation. Psychol. Forschungsberichte, Univ. Osnabrück, Nr.21, 1980b.
- (26) GEDIGA, G.: Quadratische Zuordnungsprobleme. Ei-

- nige Algorithmen und Programme. Arbeitsberichte Psychol. Meth., Univ. Osnabrück, Nr.7, 1982.
- (27) GEDIGA, G.: Kontingenztafelanalyse und die Stabilität von Clusterstrukturen. Arbeitsberichte Psychol. Meth., Univ. Osnabrück, Nr.10, 1983.
 - (28) GIGERENZER, G.: Messung und Modellbildung in der Psychologie. München: Ernst Reinhardt 1981. p.475.
 - (29) GOLDEN, R.R., MEEHL, P.E.: Detection of Biological Sex: An empirical Test of Cluster Methods. In: Mult.Beh.Res. 15(1980)p.475-496.
 - (30) GORDON, A.D.: Classification. Methods for the exploratory analysis of multivariate data. London 1981.
 - (31) HARTIGAN, J.A.: Representation of similarity matrices by trees. In: J.Amer.Stat.Ass. 62(1967)p.1140-1158.
 - (32) HARTIGAN, J.A.: Clustering Algorithms. New York: Wiley 1975.
 - (33) HARTMANN, W.: Modelle und Verfahren der Clusteranalyse auf der Basis von Skalarprodukt-Relationen. In: Int.Classif. 9(1982)p.129-139.
 - (34) HODAPP, V.: Versuche zur Typenfindung (Cluster-Analyse). In: KOLLER, S. (Ed.): Klinisch-statistische Forschungen. Mainz 1976. p.241-251.
 - (35) HORST, W., MORAWE, G.: Kompartimentierung von Zustandsräumen durch Feldlinien. In: BOCK, H.H. (Ed.): Anwendungen der Klassifikation: Datenanalyse und Numerische Klassifikation. Frankfurt (Main): Indeks Verlag 1984. p.153-161.
 - (36) HUBERT, L.J.: Inference procedures for the evaluation and comparison of proximity matrices. In: FELSENSTEIN, J. (Ed.): Numerical Taxonomy. New York 1982.
 - (37) HUBERT, L.J., GOLLEDGE, R.G.: A heuristic method for the comparison of related structures. In: J.Math.Psychol. 23(1981)p.214-226.
 - (38) HUBERT, L.J., SCHULTZ, J.V.: Quadratic assignment as a general data analysis strategy. Brit. In: J.Math.Statist.Psychol. 29(1976)p.190-241.
 - (39) JARDINE, N.: Towards a general theory of clustering. In: Biometrics 25(1969)p.609-610.
 - (40) KAUFMANN, H.L., ENGELHARDT, K.: Zur Klassenzahlbestimmung bei Mischverteilungsverfahren. In: IHM, P.; DAHLBERG, I. (Eds.): Numerische und Nicht-Numerische Klassifikation zwischen Theorie & Praxis. Frankfurt/Main: Indeks Verlag 1982. p.122-131.
 - (41) KLASTORIN, T.D.: Assessing cluster analytic results. In: J.Marketing Res. 20(1983)p.92-99.
 - (42) LANCE, G.N., WILLIAMS, W.T.: A general theory of classificatory sorting strategies. In: Computer J. 9 (1967) p.373-380.
 - (43) MARRIOTT, F.H.C.: Optimization methods of cluster analysis. In: Biometrika 69(1982)p.417-22.
 - (44) McINTYRE, R.M., BLASHFIELD, R.U.: A pearest-centroid technique for evaluating the minimum-variance clustering procedure. In: Mult.Beh.Res. 15(1980)p.225-238.
 - (45) McQUITTY, L.L.: Multiple clusters, types and dimensions from iterative intercolumnar correlational analysis. In: Mult.Beh.Res. 3(1968)p.465-447.
 - (46) MILLIGAN, G.W.: A review of Monte Carlo tests of cluster analysis. In: Mult.Beh.Res. 16(1981)p.379-407.
 - (47) MÜLLER, L.: Auswertungsmethoden zur Clusteranalyse. Dissertation, Hamburg 1977.
 - (48) OPITZ, O.: Numerische Taxonomie. Stuttgart: G. Fischer-Verlag 1980.
 - (49) PERRUCHET, C.: Significance tests for Clusters: Overview and Comments. In: FELSENSTEIN, J. (Ed.): Numerical Taxonomy. New York 1983.
 - (50) RASCH, D., HERRENDÖRFER, G., BOCK, J., BUSCH, K.: Verfahrensbibliothek Versuchsplanung und -auswertung. Berlin: Dt. Landwirtschaftsverlag 1981. Bd. 3.
 - (51) SCHLOSSER, O.: Sozialwissenschaftliche Zusammenhangs-Analyse und Profil-Cluster-Analyse. Dissertation, Berlin 1975.
 - (52) SCHMIDT, H.: Methodik der Strukturanalyse – ein Beitrag zur Modellbildung in der Psychologie. (To be published in Z.Psychol. 197(1989)No.2.
 - (53) SCHNEIDER, W., SCHEIBLER, D.: Zur Evaluation numerischer Klassifikation: Probleme beim Vergleich von Clusteranalysen. Bericht aus dem Psychologischen Institut der Universität Heidelberg, Nr. 26, 1981.
 - (54) SCOTT, A.J., SYMONS, M.J.: Clustering methods based on likelihood ratio criteria. In: Biometrics 27(1971)p.387-397.
 - (55) SNEATH, P.H.A.: A method for testing the distinctness of clusters: a test of the disjunction of two clusters in Euclidean Space as measured by their overlap. In: J.Int.Assoc.Math.Geol. 9(1977)p.123-143.
 - (56) SNEATH, P.H.A.: Some empirical tests for significance of clusters. In: DIDAY, E., LEBART, L., PAGES, J.P., TOMASSONE, R. (Eds.): Data analysis and information. Amsterdam: North-Holland 1980, p.491-508.
 - (57) SPÄTH, H.: Cluster-Analyse-Algorithmen. München: Oldenbourg 1975. p.217.
 - (58) SPÄTH, H.: Cluster-Formation und -analyse. München: Oldenbourg 1983. 236 p.
 - (59) SPRUNG, L., SPRUNG, H.: Grundlagen der Methodologie und Methodik der Psychologie. Berlin: Dt. Verlag der Wissenschaften 1984. 452 p.
 - (60) STEINHAGEN, H.E., FUCHS, S.: Objekterkennung. Berlin 1976.
 - (61) STEINHAUSEN, D., LANGER, K.: Clusteranalyse. Berlin: de Gruyter 1977.
 - (62) STEINHAUSEN, D., STEINHAUSEN, J.: Clusteranalyse als Instrument der Zielgruppendefinition in der Marktforschung. In: SPÄTH, H. (Ed.): Fallstudien Cluster-Analyse. München: Oldenbourg 1977. p.9-36.
 - (63) STELZL, I.: Fehler und Fallen der Statistik. Bern: Hans Huber 1982.
 - (64) TRAXEL, W.: Grundlagen und Methoden der Psychologie. Bern: Huber 1974.
 - (65) TURNER, D.W.: Graphical methods for representing points in n-dimensional space. 789. Tagung der Amer. Mathemat. Ges., Amherst 1981 (type-written manuscript).
 - (66) WAINER, H.: Multivariate display from Quipus to Faces. Technical Report No.82-32, Educational Testing Service, Princeton 1982.
 - (67) WARD, J.H.: Hierarchical grouping to optimize an objective function. In: J.Amer.Statist.Assoc. 58(1963)p.236-244.
 - (68) WEHNER, T.: Die Methode der "Passiv"-Variablen-Projektion als elementar-statistisches Vorgehen bei der Interpretation einer optimalen Clusterzahl. In: Z.Sozialpsych. 12(1981)No.1, p.42-48.
 - (69) WILLE, R.: Liniendiagramme hierarchischer Begriffssysteme. In: BOCK, H.H. (Ed.): Anwendungen der Klassifikation: Datenanalyse und Numerische Klassifikation. Frankfurt: Indeks Verlag 1984. p.32-51. (Also: Line diagrams of hierarchical concept systems. Int.Classif. 11(1984)No.2, p.77-86)
 - (70) WOLFE, J.H.: Pattern clustering by multivariate mixture analysis. In: Mult.Beh.Res. 5(1970)No.3, p.329-350.

Dr. Helfried Schmidt, Sektion Psychologie, Karl-Marx-Universität, Karl-Marx-Platz, DDR-7010 Leipzig