BERICHT
aus dem
PSYCHOLOGISCHEN INSTITUT
DER UNIVERSITÄT HEIDELBERG

Zur Evaluation numerischer Klassifikation:

Probleme bei Vergleich von Clusteranalysen

von

Dieter Scheibler und Wolfgang Schneider

Diskussionspapier Nr. 26 August 1981

6900 HEIDELBERG 1, HAUPTSTR. 47-51, TEL. 06221/547346

Gliederung:

0. Vorbemerkung
I. Einleitung
   1. Grundlegende Schwierigkeiten beim Umgang mit Clusteranalyse-Verfahren
   2. Kriterien fuer die Auswahl der zu evaluierenden Algorithmen
   3. Eroerterung moeglicher Bewertungskriterien
II. Ueberblick ueber die bislang durchgefuehrten Evaluationsstudien
   1. Vergleichsuntersuchungen anhand von empirischen Datensaetzen
   2. Vergleichsuntersuchungen anhand von Plasmoden
   3. Vergleichsuntersuchungen anhand von Monte-Carlo-Untersuchungen
   3.1. Moeglichkeiten zur Beschreibung und Konstruktion theoretischer Cluster-Strukturen
   3.2. Einfach strukturierte Simulationsstudien
   3.3. Komplexer angelegte Monte-Carlo-Untersuchungen
   3.4. Exkurs: Bemerkungen zur Wahl von Aehnlichkeits- und Distanzkoeffizienten
   3.5. Kurzbeschreibung der eigenen Monte-Carlo-Studie
III. Zusammenfassung und Diskussion

Literatur

Anhang

0. Vorbemerkung:

Ungefaehr drei Jahre nach dem Erscheinen unseres ersten Uebersichtsreferats zum Thema (SCHEIBLER & SCHNEIDER, 1978) wird hier nun eine weitere Studie zur Beurteilung von Clusteranalyse-Verfahren vorgelegt. Die Ursachen fuer die relativ rasch erfolgte Erweiterung bzw. Revidierung unserer Einschaetzung lassen sich mit wenigen Worten am besten so zusammenfassen:

Zum einen wurden in den beiden letzten Jahren eine Vielzahl interessanter Evaluationsstudien im anglo-amerikanischen Raum publiziert, deren Befunde es z.T. erforderlich machten, bestimmte Aspekte unserer damaligen Sichtweise zu ueberdenken und vorgenommene Schlussweisen zu relativieren bzw. dadurch zu pruefen, dass eigene Replikationsstudien geplant und durchgefuehrt wurden.

Weiterhin beschaeftigte uns die Frage immer mehr, warum in unserer damals vorgenommenen Simulationsstudie die von BLASHFIELD (1976) publizierten Befunde nur im Trend, nicht aber numerisch einigermassen exakt reproduziert werden konnten, obwohl die Generierung von Monte-Carlo-Stichproben bis ins Detail zu rekonstruieren versucht wurde. Dank der bereitwilligen Unterstuetzung von Prof. BLASHFIELD (Gainesville, Florida) war es uns in der Folge moeglich, die Eigenschaften seiner Datensaetze mit den von uns verwendeten zu vergleichen, um so die entscheidenden Unterschiede in der Stichprobenerzeugung eindeutiger charakterisieren zu koennen. Der Wert dieser Analyse kann insbesondere darin gesehen werden, dass hier allgemeine Probleme des Vergleichs von Monte-Carlo-Studien sichtbar werden, die bei der Wuerdigung von Evaluationsstudien unbedingt zu beachten sind.

Ein letzter Punkt betrifft eine Frage, die wir in unserem ersten Bericht bewusst ausgespart hatten, deren Zeit wir nun aber fuer gekommen halten: neben Problemen der Praezision und Robustheit von Clusteranalyse-Verfahren duerften zunehmend Aspekte der praktischen Handhabbarkeit, des Rechenzeitbedarfs, der Qualitaet von Fehlerdiagnostiken etc. relevant werden. Es spricht einiges dafuer, dass die Phase des sorglosen Umgangs mit Lochkarten, Computer-Papier und -zeit bald vorueber sein wird, so dass bei der Auswahl von

Rechenprozeduren deshalb auch Oekonomieprinzipien immer staerker einkalkuliert werden muessen. Fragen dieser Art wird hier fuer die CA-Programme gezielt nachgegangen (es bleibt in diesem Zusammenhang im uebrigen nur zu hoffen, dass die deutschen Universitaeten nicht zu bald mit Problemen konfrontiert werden, die BENTLER (1980) fuer die Analyse von komplexen Kausalmodellen in amerikanischen Rechenzentren als alltaeglich aufzeigt:

> "Extremely large problems will probably remain out of reach until computer costs come down another order magnitude" (S.439)).

I. Einleitung

I.1. Grundlegende Schwierigkeiten beim Umgang mit Clusteranalyse-Verfahren (CA-Verfahren)

Wenn im folgenden der Versuch unternommen wird, die Leistungsfaehigkeit unterschiedlicher statistischer Klassifikationsverfahren zu bewerten, steht dabei insbesondere ein pragmatischer Aspekt im Vordergrund: der Methodenvergleich soll dem (gelegentlichen) Anwender von Clusteranalysem eine Orientierung bei der Auswahl derjenigen Verfahren geben, die im Hinblick auf die jeweils vorliegende Fragestellung am ehesten geeignet scheinen.

Dass Hilfestellungen dieser Art wohl bei keinem anderen multivariaten Verfahren so dringend erforderlich sind, laesst sich an der besonderen historischen Entwicklung demonstrieren. Wurden die ersten Clusteranalysen noch als 'Faktorenanalysen des kleinen Mannes" belaechelt und wegen des erforderlichen Rechenaufwandes zunaechst kaum beachtet, so fuehrte die enorme Weiterentwicklung von Computersystemen in den letzten fuenfzehn Jahren zu einer wahren 'Explosion', was das Interesse an Klassifikationsverfahren anging. Nachdem innerhalb kuerzester Zeit mehr als 200 verschiedene Clusteranalyse-Algorithmen verfuegbar waren, stieg die Zahl der pro Jahr in wissenschaftlichen Journalen publizierten Untersuchungen zu taxonometrischen Problemen betraechtlich an (so registrierte etwa CORMACK fuer das Jahr 1971 mehr als 1000 Veroeffentlichungen zu Clusteranalyse-Problemen).

Der Umstand, dass in den unterschiedlichen Anwendungsgebieten (z.B. in der Biologie, den Sozialwissenschaften, der Informatik, Amthropologie, Archaeologie und Linguistik) nur die Literatur des naechsten Umfelds rezipiert wurde (meist durch die Dominanz von 'Schulen' gepraegt (s. VOGEL, 1975)), foerderte gleichzeitig die Auspraegung unterschiedlicher Terminologien ('jargon') und damit den Ausbau interdisziplinaerer Kommunikationsbarrieren (illustrative Beispiele hierzu liefern BLASHFIELD & ALDENDEFFER, 1978, S.285). Selbst dem durchaus routinierten Clusteranalysen-Benutzer ist somit haeufig nicht bekannt, dass sich mehrere (manchmal mehr als 7!) semantisch kaum verknuepfbare Etikette auf

einen einzigen Algorithmus beziehen.[1]
Angesichts der schon erwaehnten grossen Zahl von verfuegbaren Software-Programmen erscheint es weiterhin verwirrend, dass unaufhoerlich 'neue' Algorithmen publiziert werden, ohne dass hier dabei die Notwendigkeit eingesehen wird, Validierungsversuche oder zumindest Vergleiche mit aehnlich strukturierten Programmen zu berichten (Ausnahmen von dieser Regel stellen die neueren Studien von D'ANDRADE (1978) bzw. Von EYE (1977) dar).

I.2. Kriterien fuer die Auswahl der zu evaluierenden Algorithmen

Es ist nun in diesem Rahmen weder moeglich noch notwendig, die Vielzahl von Einzel-Algorithmen naeher darzustellen, da nur vergleichsweise wenige Prozeduren Eingang in populaerere Software-Programmpakete gefunden haben und somit haeufiger zur Datenauswertung herangezogen worden sind.
Nach der Einteilung von CORMACK (1971) lassen sich im wesentlichen drei Gruppen von Clusteranalyse-Verfahren unterscheiden:
Hierarchische Klassifikationsverfahren sind in der agglomerativen Variante dadurch charakterisiert, dass nach einer Serie sukzessiver Fusionen von N Elementen (Personen) zu einer Endstufe gelangt wird, in der alle Objekte einer einzigen Klasse angehoeren, waehrend bei der divisiven Form genau umgekehrt verfahren wird (die Gesamtheit der Objekte also schrittweise in immer kleinere Klassen aehnlicher Elemente zerlegt wird). In beiden Versionen werden die Ergebnisse auf den unterschiedlichen Klassifikationsebenen in Form von hierarchischen Baumstrukturen (Dendrogrammen) wiedergegeben.
Demgegenueber sind die mit partitionierenden Techniken gebildeten Cluster nicht hierarchisch, sondern wechselseitig exklusiv organisiert: nach einer Anfangspartition des Datensatzes in eine vorgegebene Zahl von Gruppen wird die Clusterhomogenitaet iterativ (durch Austausch von Einzelele-

---

[1] Eine detaillierte und kurzweilige historische Analyse der Entwicklung des Forschungsprogramms zur Clusteranalyse findet man bei BLASHFIELD (1980)

menten) solange zu verbessern versucht, bis ein naeher zu definierendes lokales Optimum erreicht worden ist.
Als letzte Gruppe von Klasifikationsverfahren ist vollstaendigkeitshalber die der sog. "clumping"-methoden zu erwaehnen, die im Unterschied zu den hierarchischen und iterativ partitionierenden Prozeduren ueberlappende Clusterloesungen zulassen.
Wie detaillierte Literatur-Recherchen (BLASHFIELD, 1980; BLASHFIELD & ALDENDERFER, 1978) bzw. sorgfaeltig geplante sog. "consumer reports" (ALDENDERFER, 1977; BLASHFIELD, 1977a u. b) gezeigt haben, kann auf die Darstellung von "clumping techniques" und divisiven hierarchischen Techniken insofern verzichtet werden, als sie nur in relativ geringem Umfang zur Anwendung gelangt sind. Es genuegt demnach, die aufgrund ihrer enormen Verbreitung in der Biologie und den Sozialwissenschaften gelaeufigeren wichtigsten Varianten der hierarchisch agglomerativen Clusteranalysen (single-linkage-, complete linkage-, average linkage-Methode, Centroid- u. Median-Methode, LANCE-WILLIAMS' flexible beta-Techniken und die WARD-Technik) sowie charakteristische Algorithmen der iterativen Partitionierungstechniken (K-Means, RELOCATE) im Bewertungsprozess zu beruecksichtigen.[4]

Das wegen seines im deutschsprachigen Raum groesseren Bekanntheitsgrades grundsaetzlich interessante Verfahren der Automatischen Klassikation von FABER & NOLLAU (eine Variante der partitionierenden Verfahren) sollte urspruenglich auch in den Kreis der zu evaluierenden Verfahren aufgenommen werden und in einer eigenen Simulationsstudie (SCHEIBLER & SCHNEIDER, 1981) analysiert werden. Da mit diesem Verfahren aufgrund seines enorm hohen Rechenzeit- und Speicherplatz-Bedarfs lediglich relativ kleine Datenmatrizen (selten mehr als 100 Elemente bzw. 2 bis 5 Variablen) verarbeitet werden koennen, war ein Vergleich mit den oben aufgefuehrten, wesentlich schnelleren Algorithmen praktisch nicht moeglich.

---

[4] Da eine ausfuehrliche Beschreibung der genannten Verfahren hier aus Platzgruenden nicht erfolgen kann, sei der interessierte Leser auf die Einfuehrungen von ANDERBERG, BLASHFIELD & ALDENDERFER, ECKES & ROSSBACH oder VOGEL verwiesen. Zur ersten Orientierung kann die im Anhang beigefuegte Kurzbeschreibung dienen

I.3. Eroerterung moeglicher Bewertungskriterien

Es bereitet keine sonderliche Muehe, sich verschiedene Gesichtspunkte vorzustellen, nach denen Computerprogramme zur Klassifikation von Objekten miteinander verglichen werden koennten. Neben Oekonomieprinzipien sind so etwa Fragen nach der Flexibilitaet (z.B. Anzahl der verfuegbaren Optionen), nach den erforderlichen Statistik- bzw. EDV-Vorkenntnissen zur Ingangsetzung/Handhabung des jeweiligen Programms bzw. zur Interpretation des Daten-Outputs (useability) denkbar. Da Probleme dieser Art schon an anderer Stelle (vgl. BLASHFIELD, 1977b) ausfuehrlich diskutiert worden sind, soll hier nur kurz auf die wesentlichen Befunde eingegangen werden.

Im Hinblick auf den Rechenzeit- und Kernspeicherbedarf sind die iterativen Partitionierungsverfahren gegenueber den hierarchischen agglomerativen Techniken insofern im Vorteil, als bei ihnen nicht die Speicherung der N x N - Aehnlichkeitsmatrix (= N x (n - 1) / 2 Elemente), sondern lediglich die Verfuegbarkeit der Rohdaten-Matrix sowie einiger relativ kleiner zusaetzlicher Matrizen erforderlich ist. Groessere Datenmengen lassen sich mit Partitionierungstechniken demnach also weitaus oekonomischer bewaeltigen.

Demgegenueber sind die hierarchischen Verfahren groesstenteils in verbreiteten Software-Programmpaketen (z.B. CLUSTAN 1C von WISHART, 1975) verfuegbar, die eine bedeutend groessere Zahl von Optionen anbieten (so bei CLUSTAN insgesamt 38 Aehnlichkeits-/Unaehnlichkeitsmasse und 8 verschiedene Cluster-Algorithmen).

Waehlt man die sog. 'useability' und damit die erforderlichen Vorkenntnisse (sowie Frustrationstoleranzen) des Anwenders als Vergleichskriterium, so weisen beide Typen von Klassifikationsverfahren im Grunde keinen Unterschied auf. Ihre Anforderungen an die Vorkenntnisse der Benutzer sind derart umfassend, dass BLASHFIELD (1977b) eine relativ pessimistische Beurteilung vornimmt:

> "Those programs were designed for a rather specific audience of cluster analysis users. This audience primarely contained Ph.D level scientists and researchers." (S. 19)

Wenn diese Aussage auch sicherlich etwas ueberspitzt formuliert ist, kann aus ihr andererseits doch der Schluss

gezogen werden, dass aufwendigere Evaluationsstudien zu den oben aufgefuehrten Vergleichsaspekten kaum lohnen.
Besondere Aufmerksamkeit wird im folgenden dem theoretisch eigentlich fundamentalen Evaluationsgesichtspunkt geschenkt, der zentral mit der Frage verknuepft ist, wie gut unterschiedliche Cluster-Algorithmen bei der Loesung vorgegebener Klassifikationsaufgaben abschneiden. Das Interesse an einer solchen Problemstellung wurde etwa ab 1965 immer staerker erkennbar, wobei die Dringlichkeit von Vergleichsstudien zum einen mit der stetig wachsenden Zahl von verfuegbaren Clusteranalyse-Techniken, zum anderem mit dem Unmut darueber begruendet wurde, dass in vielen empirischen Untersuchungen die Wahl des jeweiligen CA-Auswertungsinstruments - wenn ueberhaupt - immer mehr anhand von wenig ueberzeugenden Argumenten (zufaellige lokale Verfuegbarkeit, intuitive Annahmen) vorgenommen wurde.
Die zur Behebung dieses Wissensdefizits in der Folge durchgefuehrten Evaluationsstudien lassen sich am einfachsten nach der Art der zum Vergleich verwendeten Datensaetze differenzieren: so koennen (in der chronologisch korrekten Sequenz) Evaluationen anhand von empirischen Datensaetzen, anhand von Plasmoden sowie ueber Monte-Carlo-Methoden unterschieden werden. Eine ausfuehrlichere Darstellung der wesentlichen Befunde scheint dabei insofern angebracht, als die meisten Studien in schwerer zugaenglichen und speziell fuer Sozialwissenschaftler eher abgelegenen angloamerikanischen Zeitschriften publiziert worden sind.

II. Ueberblick ueber die bislang durchgefuehrten Evaluationsstudien

Als paradigmatisch fuer den Aufbau der ersten Evaluationsansaetze kann die Studie von SNEATH (1966A) gelten: die drei herangezogenen hierarchisch-agglomerativen Verfahren (single-, complete- und average-link) wurden im Hinblick auf ihre Uebereinstimmung bei einer Aufgabe ueberprueft, die die Klassifikation von 20 sog. 'random points' erforderte. Alle drei Prozeduren bildeten durchaus aehnliche Cluster (die Korrelationen schwankten zwischen .94 und .97), wobei die complete-link und average-link- Methoden die groesste Gemeinsamkeit aufwiesen.
Wenn die Analyse von SNEATH auch sicherlich wertvolle Hinweise auf die bei den drei Prozeduren vorfindbaren unterschiedlichen Klassifikationsprozesse liefern konnte, bot sie dennoch keine Moeglichkeit, die Adaequanz bzw. Akkuratheit der gefundenen Clusterloesungen zu beurteilen. In einer Reihe von Folgestudien wurde deshalb versucht, gerade dieser Frage gezielt nachzugehen.

II.1. Vergleichsuntersuchungen anhand von empirischen Datensaetzen

Auch Studien dieser Art waren durch einen relativ einfachen Versuchsaufbau gekennzeichnet. Einige wenige (im Durchschnitt etwa 2-3) Clusteranalyse-Verfahren wurden an einem beliebigen empirischen Datensatz mit als bekannt und gueltig angenommenen a-priori-Gruppierungen erprobt. Die sich in einigen Faellen anschliessende diskriminanzanalytische Ueberpruefung der Clusterloesung sollte im wesentlichen die Ergebnisqualitaet sichern: wenn sich die clusteranalytisch bestimmten Gruppen auch ueber ein solches Verfahren trennen liessen, wurde dies als Beweis dafuer gewertet, dass das benutzte Clusteranalyse-Verfahren als valide einzustufen war.
Auf diese Art und Weise verfuhr ALLMER (1974), um das Taxonome-Programm von CATTELL & COULTER (1966) und das Verfahren der Automatischen Klassifikation (FABER & NOLLAU, 1969) anhand einer Stichprobe von Leistungsmotivations-Dimensionen

deutscher und schweizerischer Leistungssportler zu vergleichen. Beide Verfahren lieferten unterschiedliche Resultate (nur AUKL kam zu einer Zwei-Gruppen-Loesung, waehrend das Taxonome-Programm lediglich eine Hauptgruppe, dafuer aber mehrere 'Kleinst-Cluster' identifizierte), was wohl nicht zuletzt auf den Tatbestand zurueckzufuehren war, dass beide Algorithmen unterschiedlichen Kategorien zuzuordnen sind. Waehrend AUKL als partitionierendes Verfahren saemtliche Elemente erfasst und diese jeweils nur einmal disjunkt klassifiziert, laesst das Taxonome-Programm als 'clumping technique' (s.o.) sowohl ueberlappende Cluster als auch die Moeglichkeit zu, schwer zuzuordnende Elemente in der Loesung nicht zu beruecksichtigen. Diese Unterschiede arbeitete ALLMER praezise heraus, ohne allerdings den Versuch zu unternehmen, die beiden Verfahren im Hinblick auf die Guete des Kategorisierungsprozesses zu bewerten.

In dieser Hinsicht machten die Studien von GOLDSTEIN & LINDEN (1969) bzw. ROGERS & LINDEN (1973) klarere Aussagen: sowohl an psychiatrischen wie auch an psychologischen Daten wurde eine Ueberlegenheit der (im deutschsprachigen Raum unbekannten) Clusteranalyse nach LORR, KLETT & McNAIR (1963) gegenueber der WARD-Technik und einer Hauptkomponentenanalyse (bei GOLDSTEIN & LINDEN) bzw. der WARD-Methode (bei ROGERS & LINDEN) herausgestellt. Kritisch anzumerken bleibt jedoch, dass zum einen die diskriminanzanalytisch vorgenommene Gruppentrennung (wie im uebrigen auch bei ALLMER) keine allzu ueberzeugende Befunde lieferte, zum anderen (wenn auch nicht explizit vermerkt) offensichtlich Korrelationskoeffizienten als Aehnlichkeitsmasse verarbeitet wurden, die bei der WARD-Technik nicht angemessen sind.

Wenn auch HUBERT & BAKER (1976) lediglich zwei Clusteralgorithmen (single- und complete-linkage-Verfahren) miteinander verglichen, scheint der dabei gewaehlte Ansatz insbesondere wegen seiner Kreativitaet erwaehnenswert. Die Autoren benutzten hierbei die Information, dass bei den genannten Verfahren auf jeder Hierarchie-Ebene eine neue Aehnlichkeitsrangfolge der einzelnen Objektpaare einer Datenstichprobe erstellt wird, die sich mit den originalen Aehnlichkeitsbeziehungen (Distanzen bzw. Korrelationen der Ausgangsmatrix) vergleichen laesst. Der Uebereinstimmungsgrad

zwischen dem Partitionsrang (d.h. der Hierarchie-Ebene, auf der zwei Objekte erstmals gemeinsam vorkommen) und dem Aehnlichkeitsrang jedes Objektpaars machte Aussagen ueber das Abschneiden der Clusteranalyse-Verfahren moeglich: nur fuer den Fall einer hohen Entsprechung konnte gefolgert werden, dass die vorgefundene Partitionshierarchie eine angemessene Repraesentation der Daten darstellte. Die Werte fuer das gewaehlte Assoziationsmass (Gamma-Statistik von GOODMAN & KRUSKAL, 1954) fielen insgesamt gesehen fuer das conplete-linkage-Verfahren etwas hoeher aus, ohne allerdings voellig befriedigen zu koennen.

Dieses Ergebnis stimmt weitgehend mit den Befunden von HUBERT (1974) sowie BAKER & HUBERT (1975) zur Effizienz beider Verfahren ueberein, wenn auch die letztgenannte Studie eine Differenzierung nahelegt: das single-linkageVerfahren kommt danach in all denjenigen Faellen zur besseren Loesung, bei denen die 'wahre' Cluster-Partition ein absolut dominantes Cluster enthaelt. Leider sind in den Studien von HUBERT und BAKER die Stichprobengroessen fuer die verwendeten Datenbeispiele aeusserst gering ausgefallen (bei BAKER & HUBERT lediglich N = 12 !), so dass schon von daher Verallgemeinerungen schwer fallen.

Die in den geschilderten Untersuchungen mehrfach aufgedeckten Probleme der Autoren, praezise Schlussfolgerungen ueber die Qualitaet der ueberprueften Verfahren zu ziehen, sind wohl entscheidend mit einem spezifischen Phaenomen verknuepft: alle mit empirischen (psychologischen Datensaetzen durchgefuehrten Methodenvergleiche haben naemlich den entscheidenden Nachteil, dass die Datenstruktur nicht genau genug bekannt ist, um als Aussenkriterium benutzt werden zu koennen (vgl. BAUMANN, 1973). Wenn man unterstellt, dass die ueber Klassifikationsverfahren naeher zu analysierenden Datensaetze als 'Daten-Subeinheiten', d.h. als Amalgam mehrerer Populationen zu verstehen sind, so bleiben deren Anzahl bzw. Verteilungsparameter bei Verwendung empirischer Daten unbekannt; die befriedigende bzw. inadaequate Rekonstruktion der einzelnen Populationsanteile durch die verschiedenen Cluster-Algorithmen laesst sich somit nicht naeher bestimmen.

Entgegen der Annahme vieler Autoren scheint auch das

uebliche Verfahren, (lineare) Diskriminanzanalysen zur Ueberpruefung der Guete einer Cluster-Loesung bzw. zur Bestimmung der Loesbarkeit von Klassifikationsaufgaben zu verwenden, der Problematik nicht angemessen zu sein. Da die meisten Clusteranalyse-Verfahren disjunkte Gruppen bilden, die mit einer nicht-geraden Trennungslinie oder -ebene eindeutig separierbar sind, deutet von daher ein unguenstiges Abschneiden der Diskriminanzanalyse nicht notwendigerweise auf die Unangemessenheit der Cluster-Loesung, sondern moeglicherweise eher auf die 'der Diskriminanzanalyse hin. Der Einsatz von Diskriminanzanalysen waere also nur dann denkbar, wenn (wie beispielsweise bei der Automatischen Klassifikation nach FABER & NOLLAU) durch den Cluster-Algorithmus generell nur solche Cluster angestrebt werden, die ueber eine lineare Trennfunktion gut unterschieden werden koennen. Eine Ueberpruefung der Guete der Cluster-Loesung wuerde sich dann allerdings in der Regel eruebrigen.

II.2. Vergleichsuntersuchungen anhand von Plasmoden

Die oben skizzierten Nachteile der mit empirischem Datenmaterial durchgefuehrten Evaluationsstudien veranlassten einige Autoren dazu, real gemessene Datensaetze mit bekannter Struktur, sogenannte Plasmoden, zu verwenden. Nach BAUMANN (1971, 1973) laesst sich durch Modelluntersuchungen dieses Typs die grundsaetzliche (nicht jedoch die generelle) Gueltigkeit eines Verfahrens nachweisen.
Grundsaetzliche Probleme im Hinblick auf den Vergleich unterschiedlicher Plasmodenstudien (aber auch Monte-Carlo-Studien, s.u.) schienen dadurch gegeben zu sein, dass die Uebereinstimmung zwischen vorgegebener Datenstruktur und der reproduzierten Clusterloesung ueber verschiedene Assoziationsmasse (so z.B. Gamma nach GOODMAN & KRUSKAL, RAND's Statistik, COHEN's Kappa, CRAMER's V etc.) erfasst wurde. Da alle diese Masse jedoch ueber Linearfunktionen ineinander ueberfuehrbar sind (vgl. HUBERT & LEVIN, 1976) und die Interkorrelationen dementsprechend hoch ausfallen (so berichteten MILLIGAN & ISAAC (1980) ein r von .97, EDELBROCK & McLAUGHLIN annaehernd identische Befunde fuer Kappa und Rand's Statistik), kann diese Frage im vorliegenden Zusammenhang vernachlaessigt werden.
Eine originelle Methode der Datengewinnung waehlten BARTKO et al. (1971), indem sie eine Gruppe von 100 fiktiven psychiatrischen Patienten zusammenstellten, die (in je gleichem Umfang) fuenf diagnostischen Kategorien zugewiesen wurden. Wenn die Lektuere auch nicht eindeutig erkennen laesst, welche speziellen Cluster-Algorithmen in den Vergleich eingingen, so duerfte es sich neben der complete-linkage--Technik weiterhin um die Optimierungstechnik von RUBIN & FRIEDMAN (1967) sowie um ein Clusteranalyseverfahren nach LORR et al. (s.o.) gehandelt haben (es machte sich hier leider die oben erwaehnte Etikettierungsproblematik sehr nachteilig bemerkbar). Als wesentliches Ergebnis liess sich festhalten, dass die complete-linkage-Technik (und damit das hierarchische Verfahren) den beiden uebrigen Algorithmen eindeutig ueberlegen war: waehrend das Verfahren nach LORR et al. nur einen geringen Prozentsatz der Probanden klassifizieren konnte, versagte das iterative Partitionierungsver-

fahren fast voellig. Interessante Nebenaspekte betrafen die Wahl des Aehnlichkeitskoeffizienten (Euklidsche Distanz vs. Korrelationsmass) und die der zugrundegelegten Datenmatrix (Rohdaten vs. Faktorwerte). Unguenstigere Reproduktionswerte ergaben sich generell fuer die reduzierten Daten und (ueberraschenderweise) gerade bei dem complete-linkage-Verfahren fuer die Euklidschen Distanzwerte.

Als Weiterfuehrung dieses Ansatzes koennen die Arbeiten von MEZZICH (1975, 1978) gelten, der ebenfalls fiktive 'archetypale' Patientengruppen erzeugte und fuer die Validierung insgesamt 10 taxonometrische Methoden (u.a. single-, complete- und average-linkage-Technik als hierarchische Prozeduren, KMEANS als iterativ-partitionierendes Verfahren) heranzog. Wie bei BARTKO et al. erwies sich die complete-linkage-Methode konsistent als bestes Verfahren, wenn Korrelationen als Aehnlichkeitskoeffizienten fungierten. Dieser gerade im Hinblick auf die Selektion von Aehnlichkeitsmassen mit den Lehrbuchmeinungen kaum vereinbare Befund liess sich insofern generalisieren, als Korrelationskoeffizienten - ueber alle verwendeten Verfahren gemittelt - insgesamt besser als die theoretisch eigentlich vorzuziehenden Distanzkoeffizienten abschnitten.

Relativ haeufig fanden bei Plasmodenstudien auch Kraftfahrzeugdaten Verwendung. Am Beispiel einer solchen KFZ-Plasmode konnte BAUMANN (1971) eine befriedigende Rekonstruktion der Datenstruktur durch das von ihm naeher beschriebene Taxonome-Programm (CATTELL & COULTER, 1966) nachweisen und in einer weiteren Vergleichsuntersuchung (BAUMANN, 1973) weitgehend bestaetigen. Waehrend sich fuer das in der spaeteren Studie mit in den Vergleich aufgenommene Verfahren der Automatischen Klassifikation aehnliche Ergebnisse fanden, fielen die Befunde fuer eine Konfigurationsanalyse (KFA) deutlich ab.

Von EYE (1977) verwendete die KFZ-Plasmode von BAUMANN fuer den Vergleich zwischen der Clusteranalyse nach WARD und einem selbst entwickleten Verfahren, der multivariaten automatischen Clustersuchstrategie MACS. Beide Prozeduren bildeten die Datenstruktur gut ab und erwiesen sich zusatzlich als ausgesprochen anwendungsoekonomisch. Unterschiede zwischen beiden Verfahren wurden auf die verschiedenen Aehn-

lichkeitsdefinitionen sowie die spezifischen Datenkonstellationen zurueckgefuehrt.

Wenn auch den vorgestellten Plasmodenstudien insgesamt gesehen mehr Aussagekraft (im Sinne einer grundsaetzlichen Validitaet) als den Untersuchungen anhand von empirischen Datensaetzen zugestanden werden muss, teilen sie mit letzteren doch den spezifischen Nachteil, dass die Methodenvergleiche immer nur auf einer einzigen Datenstichprobe basieren. Bei der Verwendung von "Standard-Datensaetzen" (sog. 'benchmark' data sets) besteht insbesondere die Gefahr,

> "solche Klassifikationsverfahren als besonders leistungsfaehig zu bezeichnen, deren Ergebnisse moeglichst gut mit dem Standardergebnis uebereinstimmen, obwohl diese Klassifikationsverfahren lediglich geeignet sein koennen, Aehnlichkeitsstrukturen dieser Art (und keine anderen) aufzuspueren."
> (VOGEL, 1975, S.40)

Waren es bei Vergleichsuntersuchungen anhand empirischer Stichproben etwa die sog. 'Iris'-Daten von FISHER (1936), die immer wieder zur Analyse herangezogen wurden, so liess sich bei Plasmodenstudien die Tendenz nachweisen, dass gezielt auf KFZ-Daten zurueckgegriffen wurde. Da den so gewonnenen Ergebnissen allenfalls eine eingeschraenkte Validitaet zuerkannt werden konnte, verstaerkte sich in der Folge der Trend, solche Untersuchungsmodi zu praeferieren, bei denen Methodenvergleiche an einer Reihe unterschiedlicher Datenstichproben vorzunehmen waren (s.u.).

Es bleibt hier schliesslich noch erwaehnenswert, dass die Aussagekraft von einigen der oben erwaehnten Studien (ALLMER, 1974; BAUMANN, 1971 und 1973) zusaetzlich dadurch erheblich gemindert wird, dass in ihnen das Taxonome-Programm von CATTELL & COULTER zum Vergleich herangezogen wurde. Dieses deshalb, weil HARTMANN (1976a) bei der Ueberpruefung einen schwerwiegenden Fehler des Programms feststellte: es liess sich zeigen, dass die Ergebnisse mit der Eingabe-Reihenfolge des jeweiligen Datensatzes variierten. Fuer den von HARTMANN (1976b) umgearbeiteten, immens zeitintensiven Algorithmus liegen bislang keine Resultate aus Vergleichsstudien vor.

II.3. Vergleichsuntersuchungen anhand von Monte-Carlo-Studien

Als Monte-Carlo-Studien werden im allgemeinen solche Untersuchungen bezeichnet, die ueber die Ziehung von Zufallszahlen aus Populationen mit bekannten Verteilungsparametern Datenstichproben erzeugen, um das Verhalten von mathematisch-statistischen Verfahren gezielt untersuchen zu koennen.
Bei der Evaluation von Clusteranalyse-Verfahren ueber Monte-Carlo-Studien werden zunaechst mittels Computerprogrammen Datensaetze generiert, d.h. maschinell Stichproben aus mathematischen Populationen gezogen. Diese Populationen setzen sich wiederum aus unterscheidbaren Subpopulationen zusammen, so dass von vornherein feststeht, wie die korrekte Loesung der auf die generierten Datensaetze "angesetzten" Clusteranalyse-Verfahren aussehen muss. Dabei bleibt es dem Untersucher weitgehend selbst ueberlassen, wie er die Populationen definiert. Es ist daher kaum Uebereinstimmung darueber zu erzielen, wie die theoretische Cluster-Struktur eigentlich beschaffen sein sollte (vgl. BAKER & HUBERT, 1975; MILLIGAN, 1978 u. 1980). Im wesentlichen kristallisierten sich zwei unterschiedliche Sichtweisen heraus, die im folgenden kurz skizziert werden.

II.3.1. Moeglichkeiten der Beschreibung und Konstruktion theoretischer Cluster-Strukturen

a) Das sog. "mixture model":

Eine Reihe von Forschern (so BLASHFIELD, 1976; EDELBROCK, 1979; GROSS, 1972; KUYPER & FISHER, 1975 u.a.) bevorzugte eine Defimition, die ein Cluster als Repraesentation von "Mixturen" aus multivariat normalverteilten Populationen auffasst. Waehrend bei realen Daten die Verteilungsparameter des "mixture model" (WOLFE, 1970) mathematisch nicht bestimmbar sind, da die exakte Zahl der Populationen sowie deren Verteilungsparameter unbekannt bleiben, lassen sich ueber Monte-Carlo-Techniken sehr leicht Cluster konstruieren, die den Modellannahmen genuegen. Das Datenmodell impliziert, dass sich die Verteilungsdichte der Merkmalsaus-

praegungen einer (Gesamt-) Population aus den Dichten mehrerer Teilpopulationen zusammensetzt. Die Merkmalsausspraegungen sind durch eine begrenzte Zahl von zugrundeliegenden Dimensionen determiniert, die voneinander unabhaengig sind (unkorrelierte Faktoren), und aus denen sich Korrelationen zwischen einzelnen Merkmalen erklaeren lassen. Eine wesentliche Annahme besagt, dass sich die Teilpopulationen hinsichtlich der Mittelwerte und Varianzen der (innerhalb der Teilpopulationen normalverteilten) Merkmale sowie im Hinblick auf die Korrelation zwischen den einzelnen Merkmalen (und damit auch hinsichtlich der Faktorenstruktur) unterscheiden.
Weiterhin wird meist unterstellt, dass die Merkmale mit einem Messfehler behaftet sind, der sich - gleichverteilt - als Bruchteil der Variablen-Standardabweichung annehmen laesst. Auf der Grundlage dieser Annahmen lassen sich mit einem Computer nahezu beliebig viele Stichproben erzeugen, wobei die Schritte einer auf diesem Datenmodell aufgebauten Monte-Carlo-Studie so ablaufen koennen, wie es exemplarisch in Abb. 1 verdeutlicht wird (die im Flussdiagramm aufgefuehrten 'steps' sind programmtechnisch weitestgehend automatisiert).[1]

---

bitte hier Abb. 1 einfuegen

---

b) die Konzeption eines 'ultrametrischen' Raums:

Einer Anzahl weiterer Untersuchungen (CUNNINGHAM & OGILVIE, 1972; D'ANDRADE, 1978; EVERITT, 1974; MILLIGAN, 1978 u. 1980; MOJENA, 1977 u.a.) wurde eine Definition der theoretischen Clusterstruktur zugrundegelegt, die sich von dem 'mixture'-Ansatz insofern abhob, als sie eine eher geometrische (ultrametrische) und damit raeumliche Konzeptualisierung betonte. Diese nach Auffassung von MILLIGAN auch intuitiv befriedigende Beschreibung der Clusterstruktur schloss sich eng an einen Vorschlag von CORMACK (1971) an, dem zufolge

---

[1] Eine detaillierte Beschreibung der mathematischen Grundlagen findet sich bei SCHEIBLER & SCHNEIDER (1978)

Cluster insbesondere nach den Gesichtspunkten 'externaler Isolation' und 'internaler Kohaesion' konstruiert werden sollten. Damit ist gemeint, dass Elemente verschiedener Cluster durch nichtbesetzte Raeume separiert (externale Isolation) und innerhalb eines Clusters untereinander alle aehnlich sind (vgl. auch die Definition des 'natuerlichen' Clusters bei EVERITT, 1974).

Wenn auch aus den vorgelegten Cluster-Definitionen keine allzu unterschiedlichen Folgerungen fuer den Aufbau der Monte-Carlo-Studien resultieren, ist dennoch darauf hinzuweisen, dass der ultrametrische Ansatz in jedem Fall disjunkte Clusterstrukturen erzeugt, waehrend bei der Verwendung des 'mixture'-Modells in der Regel ueberlappende Cluster-Konfigurationen auf Populations-Ebene vorliegen, die die herangezogenen Klassifikationsverfahren gelegentlich vor unloesbare Aufgaben stellen koennen (vgl. MILLIGAN & ISAAC, 1980). Als Konsequenz solcher (eher seltenen) Ereignisse waeren in den Studien nach dem 'mixture'-Modell insgesamt etwas unguenstigere Rekonstruktionswerte zu erwarten.

Da fuer die vorzunehmende Evaluation aber weniger die absolute Hoehe der Assoziationskoeffizienten als vielmehr die Rangreihe der Clusteranalyse-Verfahren in unterschiedlichen Monte-Carlo-Studien von Bedeutung war, schien eine getrennte Praesentation der beiden Ansaetze nicht sinnvoll zu sein. Im Hinblick auf den Stellenwert der Einzelergebnisse verdiente dagegen die Frage nach dem Komplexitaetsgrad der jeweiligen Untersuchungen groessere Beachtung.

Im folgenden werden deshalb zunaechst kurz die Befunde relativ einfach strukturierter Studien praesentiert, bei denen entweder nur wenige Verfahren verglichen oder kleinere Datenstichproben generiert wurden. Mehr Raum wird anschliessend den komplizierteren und aufwendigeren Vorgehensweisen gewidmet, deren Befunde auch die Grundlage fuer eine summarische Bewertung der einzelnen Algorithmen bilden soll.

II.3.2. Einfach strukturierte Simulationsstudien

Als typische Beispiele fuer die entweder im Hinblick auf den Umfang der Datenstichproben oder aber die Anzahl der ueberprueften Clusteranalyse-Verfahren limitierten fruehen Simu-

lationsversuche koennen die Arbeiten von EVERITT (1974) bzw. GROSS (1972) gelten.
EVERITT beschraenkte sich auf die Generierung von Zufallsdaten aus bivariaten Normalverteilungen, wobei Sub-Stichproben sowohl aus zwei Populationen (bei Variation von Populationsaehnlichkeit und Stichprobengroesse) als auch einer einzigen Grundgesamtheit gezogen wurden: es interessierte hier das 'Verhalten' verschiedener Clusteranalysen bei einer nicht weiter unterteilbaren Datenkonfiguration.
Wenn auch alle drei verwendeten hierarchischen Clusteranalysen (single-linkage-, Centroid- und WARD-Technik) im erstgenannten Fall relativ guenstig abschnitten, war eine direkte Vergleichbarkeit leider nicht gewaehrleistet, da jeweils unterschiedlich grosse Stichproben verwendet wurden. Nachteilig machte sich insbesondere bemerkbar, dass gut separierbare Stichproben kreiert worden waren, so dass sich die Probleme fuer die CA-Verfahren als zu leicht erwiesen. Wertvolle Hinweise liessen sich dagegen aus den Befunden zur Klassifikation von nicht weiter unterteilbaren Datenmengen ableiten: es zeigte sich, dass alle Prozeduren den Daten eine kuenstliche Struktur aufsetzten und Scheinloesungen praesentierten. Die ueberprueften Verfahren erwiesen sich weiterhin immer dann als suboptimal, wenn Datensaetze konstruiert wurden, die andere als sphaerische Clusterstukturen enthielten.
Waehrend die bei EVERITT vorfindbaren Restriktionen der Stichprobengenerierung Generalisierungen der Befunde erschwerten, verfuhr die Studie von GROSS (1972) insofern etwas variabler, als hier bei der Ziehung der Datensaetze systematische Veraenderungen des Schwierigkeitsparameters intendiert waren. Als Ausgangsbasis diente ein trivariat verteilter, in zwei Subpopulationen zu untergliedernder Datensatz, bei dem neben der Aehnlichkeit der beiden Grundgesamtheiten auch die Groesse der daraus gezogenen Stichproben und deren Groessenverhaeltnis zueinander variiert wurden. Die Rekonstruktionsqualitaet ergab sich aus der Gegenueberstellung von Fehlklassifikations-Raten, wie sie fuer die Population erwartet und fuer die Stichproben verzeichnet worden waren. Leider wurde in dieser Studie mit dem WARDVerfahren nur ein einziger Algorithmus analysiert, fuer den die

Befunde aus ca. 120 Monte-Carlo-Laeufen allerdings durchaus positiv ausfielen: waehrend die Rate der Fehlklassifikationen bei den Stichproben knapp ueber den fuer die Population erwarteten Werten lag, ergaben sich bei ungleichen Gruppengroessen, umfangreicheren Stichproben und unaehnlicheren Populationen insgesamt guenstigere Resultate.
In die Reihe der weniger befriedigenden Monte-Carlo-Studien muss auch die Untersuchung von CUNNINGHAM & OGILVIE (1972) eingeordnet werden. Zwar wurden hier im Unterschied zur Arbeit von GROSS immerhin sechs hierarchische Verfahren (single-, complete-, average-linkage-Verfahren, MedianMethode, Centroid-Methode und WARD-Technik) ueberprueft, doch scheinen die Befunde (leichte Ueberlegenheit der average-linkage-Technik bei sonst gleichwertigen Resultaten) angesichts der schmalen Basis von nur sechs analysierten Datensaetzen kaum generalisierbar zu sein.

II.3.3. Komplexer angelegte Monte-Carlo-Untersuchungen

Im Hinblick auf eine erste Orientierung sind in Tab. 1 die wesentlichen Befunde aus sieben Studien enthalten, in denen aufwendigere Analysen zur Evaluation von Cluster-Algorithmen durchgefuehrt wurden. Waehrend fuer die Studien von BLASHFIELD (Nr. 1), SCHEIBLER & SCHNEIDER (4) und MOJENA (5) jeweils die Mediane der Assoziationskoeffizienten aufgefuehrt sind, wurden die Werte fuer die uebrigen Untersuchungen zusaetzlich ueber unterschiedliche Untersuchungsvarianten hinweg arithmetisch gemittelt. Da die so bestimmten Koeffizienten lediglich als nivellierende Grobindikatoren fuer die Guete der ueberprueften Verfahren gelten koennen, soll im Anschluss zusaetzlich auf differenziertere Teilergebnisse eingegangen werden.

---

bitte hier Tab. 1 einfuegen

Bei der Betrachtung von Tab. 1 faellt zunaechst einmal auf, dass die absolute Hoehe der Koeffizienten fuer die verschiedenen Verfahren betraechtlich variiert, wobei groessere Abweichungen sowohl innerhalb der einzelnen Studien als auch fuer identische Algorithmen ueber verschiedene Untersuchungen hinweg feststellbar sind. Besonders markant wird der Unterschied bei der Evaluation des single-linkage-Verfahrens, dem nach dem Befund von BLASHFIELD (1) kaum mehr als Zufallsklassifikationen zugestanden werden koennen (Kappa = .06), andererseits nach den Ergebnissen von MILLIGAN (6) eine erfreuliche Reproduktionsqualitaet bescheinigt werden muss (RAND's Statistik = .91).
Diese erstaunlichen Diskrepanzen machen deutlich, dass unter dem Etikett 'Monte-Carlo-Studie' durchaus unterschiedliche Modi der Zufallszahlen-Generierung integriert werden. Entgegen der oben geaeusserten Vermutung lassen sich die Unterschiede in den Absolutwerten wohl nicht zwingend auf Unterschiede in der Bestimmung der theoretischen Cluster-Struktur zurueckfuehren: die auf einer geo- bzw. ultrametrischen Clusterkonzeption aufbauenden Untersuchungen (D'ANDRADE,

MILLIGAN, MOJENA) weisen naemlich untereinander ebensolche Differenzen auf, wie sie auch zwischen den auf dem 'mixture'-Ansatz basierenden Studien (BLASHFIELD, KUYPER & FISHER) zu beobachten sind (die Arbeit von SCHEIBLER & SCHNEIDER nimmt insofern eine Sonderstellung ein, als einerseits die Datengenerierung nach dem 'mixture'-Ansatz erfolgte, andererseits aber Diskriminanzanalysen vorgeschaltet wurden (s. Abb. 1), die ueberlappende Clusterkonfigurationen auffinden und eliminieren halfen und von daher auch das Prinzip der 'externalen Isolation' garantierten).

Es scheint weiterhin wenig plausibel, die Differenzen in den Absolutwerten mit der extrem variierenden Anzahl gezogener Zufallsstichproben (von n = 12 bei MOJENA bis n = 766 bei SCHEIBLER & SCHNEIDER) zu begruenden. Ein von SCHEIBLER & SCHNEIDER zusaetzlich vorgenommener Vergleich von 8 Monte-Carlo-Durchlaeufen, die auf jeweils n = 50 Mixturen basierten, ergab eine im Durchschnitt eher geringe Schwankungsbreite der Resultate: die Kappa-Koeffizienten streuten am meisten beim single-linkage-Verfahren (Bandbreite = .16) und am wenigsten bei der LANCE-WILLIAMS-Prozedur (.05) bzw. der WARD-Technik (.06).

Es kann demnach ausgeschlossen werden, dass die grossen Diskrepanzen in der Ausprägung der Kappa-Werte auf Zufallseinfluesse zurueckzufuehren sind.

Eine von den Verfassern durchgefuehrte Re-Analyse der BLASHFIELD-Daten[1] ergab, dass der von BLASHFIELD verwendete Algorithmus zum einen deutlich mehr 'outlier' produzierte, andererseits aber auch zu staerkeren Ueberlappungen zwischen den Unterstichproben fuehrte. Beide Phaenomene bewirken eine z.T. erhebliche Senkung der resultierenden Kappa-Werte.

War also bei BLASHFIELD die prinzipielle Loesbarkeit der Aufgabe, m.a.W. die Entflechtungsmoeglichkeit fuer die Mixturen nicht immer gesichert, so taucht bei der Studie von MILLIGAN (6) ein gerade entgegengesetztes Problem auf: die fuer alle Algorithmen fast ausnahmslos hohen bis sehr hohen Reproduktionswerte lassen darauf schliessen, dass sich die Daten-Elemente als allzugut separierbar und die Aufgaben da-

---

[1] Herrn Prof. Dr. R.K. BLASHFIELD (University of Gainesville, Florida) sei fuer seine Kooperationsbereitschaft, insbesondere fuer die Ueberlassung von 10 Mixturen seiner Simulationsstudien nochmals herzlich gedankt

mit fuer saemtliche Verfahren als zu leicht erwiesen. Dies ist umso bedauerlicher, als der von MILLIGAN vorgestellte Ansatz in seiner Variabilitaet (s.u.) als beispielhaft charakterisiert werden kann.

Nachdem sich die Interpretation der Absolutwerte als problematisch herausgestellt hat, scheint es demnach allein sinnvoll zu sein, die Rangfolge der evaluierten Algorithmen zu betrachten, wie sie sich innerhalb der einzelnen Studien und ueber die Untersuchungen hinweg darstellt. Ein Blick auf die in Klammern wiedergegebenen Rangziffern macht deutlich, dass sich in dieser Hinsicht relativ klare Aussagen treffen lassen. Bei den hierarchischen Verfahren schneidet die single-linkage-Prozedur in sechs von sieben Faellen am schlechtesten ab, und umgekehrt erweist sich die Technik nach WARD ebensooft als (meist) eindeutig ueberlegen. Ein Vergleich der SPEARMAN-Rangkorrelationen fuer die in den Studien (3)-(6) verwendeten sechs hierarchischen Cluster-Algorithmen ergibt fuer die Befunde von KUYPER & FISHER, MOJENA sowie SCHEIBLER & SCHNEIDER eine nahezu perfekte Uebereinstimmung (Koeffizienten von .94 bis 1.0), waehrend die Ergebnisse von MILLIGAN (Uebereinstimmungswerte von -.14 bis .13) sich von den uebrigen enorm unterscheiden. Klammert man hier einmal die (unten noch naeher zu diskutierende) Studie von MILLIGAN aus, so koennen neben der WARD-Methode demnach auch das complete-linkage-Verfahren bzw. das average-link-Verfahren bedingt, das LANCE-WILLIAMS-Verfahren uneingeschraenkt zur Anwendung empfohlen werden, waehrend von der single-linkage-Prozedur, der Centroid- und der Median-Methode in der Regel abzuraten ist.

Aus den Befunden zu den iterativ-partitionierenden Verfahren ist abzuleiten, dass die gewaehlte Startpartition von grosser Bedeutung sein muss: waehrend zufaellig ausgewaehlte Startpartitionen ('random seed points') generell nur mittelmaessige Resultate (s. Werte in Klammern) erbringen, resultieren die besten Reproduktionswerte aus einer Strategie, bei der die Centroide der 'group average-linkage-'Methode als Anfangspartition fuer den iterativen Algorithmus ausgewaehlt werden. Mit einer solchen Vorgehensweise lassen sich Resultate erzielen, die offensichtlich nur von wenigen hierarchischen Verfahren zu erreichen sind.

Die Bedeutsamkeit der ausgewaehlten Startpartition fuer die Effizienz der iterativen Cluster-Algorithmen liess sich auch in einer Pilot-Studie von BLASHFIELD (1977c) nachweisen,* bei der 20 Monte-Carlo-Mixturen Verwendung fanden. Waehrend die Kappa-Werte fuer RELOCATE und ANDER im Mittel kaum differierten (.67 vs. .63), ergaben sich fuer die einzelnen Datensaetze je nach gewaehlter Startpartition enorme Unterschiede, die sich in Kappa-Bandbreiten von .30 bis .94 niederschlugen.

An der hier fuer iterative Verfahren demonstrierten Relevanz moeglicher Wechselwirkungen zwischen Datensatz-Struktur und Anfangspartition zeigt sich gleichzeitig die Notwendigkeit, zusaetzlich zu den in Tab. 1 wiedergegebenen 'overall'-Mittelwerten der aufgefuehrten Verfahren auch ihre spezifischen 'Reaktionen' auf systematisch variierte Besonderheiten von Monte-Carlo-Datensaetzen zu registrieren. Damit sind Moeglichkeiten gegeben, aus der Kenntnis bestimmter Eigenheiten des zu analysierenden Datensatzes heraus auf Algorithmen zurueckzugreifen, die sich in dieser Hinsicht als robust erwiesen haben.

Wie schon oben angedeutet, zeichnet sich der von MILLIGAN (1978, 1980) vorgeschlagene Ansatz insbesondere dadurch aus, dass bei der Daten-Generierung die Eigenschaften der konstruierten Cluster sorgfaeltig kontrolliert wurden. Zusaetzlich zu idealen, fehlerfrei generierten Datensaetzen ('error free parent data sets') gingen fuenf weitere fehlerverzerrte Datenstrukturen in die Analyse ein, die eine Evaluation der Clusteralgorithmen unter Bedingungen ermoeglichten sollte, wie sie wohl im Alltagsgebrauch anzutreffen sind. Die Hinzufuegung von sog. 'outliers' (Elemente, die aufgrund ihrer Distanz zu den Cluster-Centroiden schwer integrierbar sind) erwies sich unter diesem Aspekt ebenso sinnvoll wie die Verzerrung der Distanzwerte (Verfaelschung des Messprozesses).

Weitere Massnahmen bestanden darin, dass 'random noise dimensions' eingebaut wurden, die fuer die 'wahre' Cluster-Struktur keinerlei Bedeutung hatten, weiterhin Berechnungen

---

*Die Ergebnisse sind in Tab. 1 deshalb nicht aufgefuehrt, weil nicht zu rekonstruieren war, welcher der in ANDERBERG (1973) enthaltenen Algorithmen mit der CLUSTAN-Prozedur RELOCATE verglichen wurde

der Aehnlichkeit mit anderen als Euklidschen Distanzmassen vorgenommen bzw. die Variablen standardisiert wurden.
Die Resultate fuer die ueberprueften Clusteranalyse-Techniken lassen sich grob so zusammenfassen, dass die beiden zuletzt erwaehnten Schritte fast keine Abnahme in den Rekonstruktionswerten bewirkten, dass die Hinzufuegung von "random noise dimensions" bei allen Methoden erhebliche Verschlechterungen zur Folge hatten, und dass fuer die beiden uebrigen Massnahmen gravierende Unterschiede zwischen hierarchischen und nicht-hierarchischen Techniken zu registrieren waren: waehrend die Rekonstruktionswerte fuer einige hierarchische Verfahren (complete-link-Methode, WARD-Technik) bei "outlier"-Bedingungen bzw. fehlerverzerrten Distanzen (hier besonders das single-link-Verfahren) staerker abfielen, waren bei den nicht-hierarchischen Prozeduren keinerlei Einbussen zu verzeichnen. Gerade die Befunde zu den "outlier"-Bedingungen muessen jedoch erstaunen, da hier trotz der vielfach bestaetigten "chaining"-Tendenz und der damit verbundenen "outlier"-Anfaelligkeit des single-linkage-Verfahrens keine besonderen Abweichungen dieser Technik auftraten, diese vielmehr bei solchen Verfahren registriert wurden, die laut Fachliteratur in dieser Hinsicht als robust einzuschaetzen sind.
Die schon oben kritisierten Besonderheiten der Datenkonstruktion nach MILLIGAN lassen auch in diesem Punkt Skepsis angeraten sein: im Hinblick auf die "chaining"-Tendenz fallen Befunde wie die von BLASHFIELD (1978; s. Tab. 2) wesentlich hypothesenkonformer aus. Danach sind solche Verkettungstendenzen beim single-linkage-Verfahren sehr ausgepraegt, waehrend sie (mit Ausnahme des average-linkage-Verfahrens) bei den uebrigen analysierten Techniken, insbesondere beim WARD-Verfahren, nicht nachzuweisen sind.

---

bitte hier Tab. 2 einfuegen

Interessante Aufschluesse koennen auch aus der in Tab. 2 bei BLASHFIELD (1978) aufgefuehrten Statistik R-Quadrat (als dem Mass fuer den Zusammenhang zwischen Kappa-Wert und der

Schwierigkeit des Datensatzes) gewonnen werden. Wenn auch bei allen untersuchten Verfahren eine positive Beziehung, also ein guenstigerer Reproduktionswert bei schwierigeren Datenkonstellationen festzustellen ist, laesst sie sich insbesondere fuer das WARD-Verfahren als ausserordentlich eng charakterisieren. Dieser Befund wird im uebrigen auch von BLASHFIELD (1976) bzw. SCHEIBLER & SCHNEIDER (1978) bestaetigt. MOJENA (1977) unterstreicht, dass im Vergleich der hierarchischen Verfahren gerade die WARD-Technik am wenigsten durch den zunehmenden Ueberlappungsgrad der Stichproben bzw. die ansteigende Zahl der zugrundeliegenden Populationen in ihrer Praezision beeintraechtigt wird, was ihn ebenso wie die zuletzt genannten Autoren zur eindeutigen Bevorzugung dieses Verfahrens veranlasst.
Wenn auch McINTYRE & BLASHFIELD (1980) bei ihrer Ueberpruefung von Stabilitaetsniveaus verschiedener Cluster-Loesungen aus "mixture"-Simulationen fuer den WARD-Algorithmus eine negative Korrelation zwischen Ueberlappungsgrad der Stichproben und dem sog. "aggreement kappa"[1] feststellten, bestaetigten auch hier die Befunde den angedeuteten Trend: es liess sich insgesamt ein mittelhoher Zusammenhang zwischen den Stabilitaetswerten und den (generell hohen) Praezisionsmassen ("accuracy kappa") nachweisen, was die Autoren zur Schlussfolgerung fuehrte, dass Stabilitaetskoeffizienten bei dieser Technik eine indirekte Schaetzung dafuer darstellen koennen, wie genau die Cluster-Loesungen die tatsaechliche Datenstruktur rekonstruieren.

Zur Vervollstaendigung des ueberaus komplexen Bildes sind in Tab. 2 die Rangkorrelationen zwischen Kappa-Werten und einigen Statistiken zusammengestellt, die in den Studien von BLASHFIELD sowie SCHEIBLER & SCHNEIDER erhoben wurden. Die Spearman-Koeffizienten verdeutlichen dabei die Relevanz der Element-Anzahl in der gezogenen Gesamtstichprobe (a), der mittleren Element-Anzahl in den Teilstichproben (b), deren Varianz (c) sowie des Ueberlappungsgrades in den Teilstichproben (d). Trotz einiger Abweichungen in Kategorie (d) und

---

[1] dieses Mass fuer die Uebereinstimmung bzw. Stabilitaet von Cluster-Loesungen verschiedener Stichproben aus einer Mixtur wurde ueber eine Prozedur gewonnen, die parallel zu dem bei Regressions- bzw. Diskriminanzanalysen haeufiger verwendeten Kreuzvalidierungs-Paradigma konstruiert war

der tendenziell hoeheren Absolutwerte bei BLASHFIELD kann aus den Ergebnissen insgesamt gefolgert werden, dass die vier variierten Parameter keinen allzu grossen Einfluss auf die Auspraegung der Rekonstruktionswerte nehmen. Die Statistiken a und b (Stichprobengroesse und -repraesentativitaet) wirken sich nur in der Analyse von BLASHFIELD fuer das single-linkage-Verfahren in dem Sinne aus, dass bei zunehmender Groesse der Gesamt-Stichprobe und abnehmender Repraesentativitaet der Einzel-Elemente fuer die Teilstichproben deutlich verringerte Werte zu erwarten sind; Statistik c (spezifische Stichprobenrelationen) hat einen insgesamt zu vernachlaessigenden Einfluss, waehrend der Ueberlappungsgrad der einzelnen Teilstichproben (Statistik d) bei der WARD-Technik, aber auch beim LANCE-WILLIAMS-Verfahren und der Prozedur RELOCATE im oben beschriebenen Sinne wirkt: zunehmende Datenkomplexitaet fuehrt hier zu verbesserten Reproduktionswerten. Dieser Befund kann fuer die drei genannten Verfahren (und nur fuer diese) am Beispiel von aehnlich wirkenden Statistiken (in Tab. 2 nicht aufgefuehrt) validiert werden, die (z.T.) bei BLASHFIELD sowie SCHEIBLER & SCHNEIDER berechnet wurden. Hoehere Ergebnisqualitaet ist danach bei steigender Variablenanzahl (Korrelationen zwischen .62 und .66), bei zunehmender mittlerer Anzahl von Hauptkomponenten (.54 - .59) sowie bei groesserer Varianz in der Anzahl der Hauptkomponenten bei den Teilstichproben (.46 - .47) zu erwarten; die genannten Verfahren schneiden also bei zunehmender Komplexitaet der Kovarianzstrukturen innerhalb und zwischen den Populationen besser ab.

Da die bisher fuer die hierarchischen Cluster-Algorithmen vorgelegten Befunde relativ konsistent nur einige wenige Verfahren als akkurat und robust darstellen, andere dagegen regelmaessig als schwach einstufen, scheint hier eine Wahl relativ leicht zu fallen.

Dies scheint allerdings nur so, da EDELBROCK (1979) kuerzlich auf ein Problem hinwies, das als Erklaerung fuer die unterschiedlichen Reproduktionsgenauigkeiten dienen kann und im folgenden etwas detaillierter skizziert werden soll.

EDELBROCK setzt dabei an dem Faktum an, dass sich die bei Monte-Carlo-Studien verwendeten Assoziationskoeffizienten auf eine Hierarchie-Ebene beziehen, bei der die Anzahl der

Cluster der der zugrundegelegten Populationen entspricht und alle Elemente erfasst worden sind. Letzteres wird von ihm insofern als ernstes Problem empfunden, als damit die 'outlier'-Behandlung von Cluster-Algorithmen entscheidend an Bedeutung gewinnt. Verfahren, die 'outlier'-Elemente schon relativ frueh in bestehende Cluster integrieren, sind gegenueber Techniken, die einzelne 'outlier'-Elemente auch in hoeheren Hierarchie-Ebenen als separate Cluster fuehren, dann enorm im Vorteil, wenn eine hundertprozentige Datenrepraesentation fuer die Evaluation gefordert wird (bei letzteren wird schlicht und einfach die inadaequate Hierarchie-Ebene zur Beurteilung ausgewaehlt).

Um eine fairere Behandlung der somit unterschaetzten Algorithmen zu garantieren, schlaegt EDELBROCK eine Prozedur vor, bei der zunaechst wie ueblich diejenige Ebene ueberprueft wird, bei der die Anzahl der Cluster der der zugrundeliegenden Populationen entspricht (also hundertprozentige Objekt-Erfassung), und dann schrittweise niedrigere Hierarchie-Ebenen solange abgetestet werden, bis eine perfekte Rekonstruktion (mit notwendigerweise unvollstaendiger Objekt-Erfassung) erreicht ist.

In einer eigenen empirischen Untersuchung waehlte EDELBROCK (1979) 10 der 50 Daten-Mixturen von BLASHFIELDs Monte-Carlo-Studie aus und verglich die Reproduktionsgenauigkeit des single-link-, complete-link- und average-link-Verfahrens, der Centroid- und der WARD-Methode auf sechs Ebenen der Objekterfassung (es sollten jeweils 100, 95, 90, 80 bzw. 70 Prozent der Elemente klassifiziert werden).

Da bei zunehmend niedrigeren Hierarchie-Ebenen die Zufallswahrscheinlichkeit fuer "richtige" Cluster steigt, wurde zur Kontrolle zusaetzlich ein "random linkage Algorithmus" eingefuehrt, fuer den ebenfalls Kappa-Werte berechnet wurden. Um die Relevanz des verwendeten Aehnlichkeitskoeffizienten abschaetzen zu koennen, kamen fuer die vier erstgenannten Algorithmen sowohl Korrelationskoeffizienten als auch Euklidsche Distanzen zur Anwendung, waehrend die WARD-Technik per definitionem nur mit Distanzen zu benutzen war. Zusaetzlich wurden die Effekte standardisierter vs. nichtstandardisierter Datensaetze untersucht.

---

bitte hier Abb. 2 einfuegen

Wie aus Abb. 2 zu erkennen ist, fallen die Relationen zwischen Akkuratheit und Grad der Element-Erfassung (coverage) fuer die analysierten Algorithmen durchaus unterschiedlich aus. Wenn auch die Reproduktionsgenauigkeit mit zunehmendem Grad der Element-Erfassung generell abnimmt, so muss die konsistente Ueberlegenheit der auf Korrelationskoeffizienten basierenden Loesungen (erneut) verblueffen. So bleibt neben dem Befund, dass fuer die meisten der vielgeschmaehten hierarchischen Verfahren bei einer immerhin 90-prozentigen Objekterfassung zumindest befriedigende Kappa-Werte resultieren, die eigentlich sensationelle Erkenntnis, dass diese Algorithmen auch bei der ueblichen Evaluationsprozedur dann aehnlich gut wie das favorisierte WARD-Verfahren abschneiden, wenn ihnen statt Distanzwerten Korrelationen zugrundeliegen.

EDELBROCK vermag zwar fuer dieses ueberraschende Resultat keine ueberzeugende theoretische Begruendung anzubringen, sieht aber moegliche Verursachungsmomente in der unterschiedlichen Behandlung von "outliers", die bei Benutzung Euklidscher Distanzen durch die simultane Beruecksichtigung von Profil-Form und -steigung (elevation) besonders isoliert werden, bei Verwendung von Korrelationskoeffizienten aber eher zurechtgestutzt und ins Cluster integriert werden koennen, da hier nur die Form des Profils ins Gewicht faellt. Andererseits ist auch das pragmatische Element EDELBROCKs, nach dem die Korrelation wohl eher das geeignete Aehnlichkeitsmass fuer den ueblichen Daten-Typus der Sozialwissenschaften (Fragebogen, Rating-Skalen) darstellt, ebenfalls kaum von der Hand zu weisen.

In einer Folgestudie (EDELBROCK & McLAUGHLIN, 1980) wurde das Problem der Selektion angemessener Aehnlichkeitskoeffizienten erneut aufgegriffen. Fuer das single-, complete- und average-linkage-Verfahren sowie die Centroid-Methode wurde dabei untersucht, ob die Vorgabe von Produkt-Moment-Korrelationen, Euklidschen Distanzen oder sog. "intraclass correlations" (ICCs) unterschiedliche Auswirkungen auf die

Cluster-Loesungen haben wuerden. Das letztgenannte Aehnlichkeitsmass wurde deshalb in die Analyse aufgenommen, weil es im Unterschied zu den Distanz-Metriken und Produkt-Moment-Koeffizienten als sensitiv fuer Unterschiede in Profil-Form, -steigung und -streuung (scatter) eingestuft wurde und insbesondere gut dazu geeignet schien, additiven "bias" zwischen einzelnen Profilen aufzudecken (vgl. BARTKO, 1974 u. 1976). Als Kontrollverfahren diente die nur mit Euklidschen Distanzen handhabbare WARD-Technik. Zusaetzlich interessant erschien die Absicht der Autoren, die Generalisierbarkeit von Ergebnissen aus Monte-Carlo-Studien dadurch zu ueberpruefen, dass zwei unterschiedliche Daten-Generierungs-Modi (multivariat normalverteilte Mixturen nach BLASHFIELD und multivariate Gamma-Mixturen nach MOJENA) miteinander verglichen wurden.

Die Fruchtbarkeit dieser Massnahme liess sich an den Befunden ablesen, die interessante Interaktionen zwischen dem gewaehlten Aehnlichkeitsmass und der Art der Stichproben-Generierung zutage brachten: waehrend Euklidsche Distanzen bei den BLASHFIELD-Daten generell ungenauere Loesungen als die ueblichen Aehnlichkeitsmasse lieferten, waren es bei den Mixturen nach MOJENA die Produkt-Moment-Korrelationen, die insgesamt am schwaechsten abschnitten. EDELBROCK & McLAUGHLIN schliessen daraus, dass bei den multivariat normalverteilten Daten BLASHFIELDs die Form als wesentlicher Profil-Parameter auftritt, waehrend bei den Gamma-Mixturen MOJENAs eher die Profil-Steigung bzw. -streuung relevant wird. Damit scheint klar, dass Charakteristika der Monte-Carlo-Mixturen die Befunde nicht unerheblich beeinflussen, sodass bei Generalisierungsversuchen besondere Vorsicht geboten scheint.

Es ergab sich also in dieser Studie als ein wesentliches Ergebnis, dass sich ueber beide Mixtur-Typen hinweg das average-linkage-Verfahren mit ICC-Koeffizienten als akkuratestes Cluster-Verfahren erwies. Wenn es auch nicht signifikant besser als die WARD-Technik bzw. die Centroid-Methode mit ICC-Mass abschnitt, unterschied es sich bedeutsam von allen uebrigen Kombinationen aus Algorithmen und Aehnlichkeitsmassen. Obwohl das ICC-Mass seine Bewaehrungsprobe damit sicherlich bestanden hat, folgern die Autoren aus ihren Ergebnissen, dass kein einziger Algorithmus fuer alle Anwen-

dungsmoeglichkeiten gleichermassen optimal geeignet ist. Ausschlaggebend sollten deshalb theoretische Vorueberlegungen des Forschers z.B. zur Frage sein, ob er nun eher daran interessiert ist, Cluster vorzugsweise ueber den Profil-Form-Parameter zu definieren, oder ob Subgruppen identifiziert werden sollen, die in Profilform, -steigung und -streuung differieren.

Insgesamt gesehen lassen sich die zuletzt dargestellten Befunde jedenfalls kaum mit der Auffassung MILLIGANs (1980) vereinbaren, der zufolge die Wahl des Cluster-Algorithmus wichtiger als die des Aehnlichkeitsmasses sein sollte.

II.3.4. Exkurs: Bemerkungen zur Wahl von Aehnlichkeits- und Distanzkoeffizienten

Die Darstellung der Monte-Carlo-Studien von EDELBROCK (1979) sowie EDELBROCK & McLAUGHLIN (1980) mag den Eindruck erweckt haben, dass das Ziel adaequater Evaluationsstudien die Suche nach der optimalen Kombination zwischen einem bestimmten Cluster-Algorithmus und einem bestimmten Distanz- oder Aehnlichkeitskoeffizienten sein sollte, bzw. dass man ueber solche Studien auch das optimale Distanz-/Aehnlichkeitsmass ermitteln kann. Dem glauben wir entgegenhalten zu muessen, dass sich die Wahl eines geeigneten Masses in erster Linie aus der Fragestellung ergibt, in deren Rahmen die Clusteranalyse eingesetzt wird, bzw. aus theoretischen Vorueberlegungen. Aus diesem Grund halten wir es auch fuer wenig fruchtbar, die Guete von Distanz/-Aehnlichkeitsmassen (an den gleichen Datensaetzen) zu vergleichen.
Um diese Auffassung zu verdeutlichen, wollen wir zunaechst eine Erklaerung dafuer liefern, dass in den Studien von EDELBROCK einige Clusterverfahren deutlich bessere Resultate erbrachten, wenn als Aehnlichkeitsmass statt der Euklidschen Distanz der Korrelationskoeffizient eingesetzt wurde.

---

bitte hier Abb. 3 einfuegen

Betrachten wir hierzu einmal Abb. 3, in der schematisch die Elemente von Stichproben aus zwei unterschiedlichen Populationen in einem orthogonalen zweidimensionalen Faktorenraum als Punkte dargestellt sind. Bei einer Loesung mit drei Clustern wird man bei den meisten Algorithmen ein Ergebnis erwarten koennen, wie es durch die Einkreisung der Punkte dargestellt ist. Die korrekte Loesung bei zwei Clustern sei durch die gestrichelten Ovale gekennzeichnet.
Waehlt man nun in diesem Beispiel als Distanzkoeffizienten die Euklidsche Distanz, so wuerden z.B. beim single-linkage-Verfahren im naechsten Agglomerationsschritt die Cluster C2 und C3 zusammengefasst werden, da d1 kleiner als d2 ist. Als Endergebnis (Anzahl der Cluster = Anzahl der Populationen) erhielte man also ein grosses Cluster, bestehend aus

fast allen Elementen, und ein kleines Cluster, das nur aus einem einzigen Element, einem sog. 'outlier' bestehen wuerde. Der Vergleich mit der korrekten Loesung ergaebe einen sehr geringen Wert des Guetemasses. Waehlt man nun aber statt der Euklidschen Distanz die Produktmomentkorrelation als Aehnlichkeitskoeffizienten, muss mit einem wesentlich besseren Endergebnis gerechnet werden. Die Korrelation zwischen zwei Elementen entspricht naemlich dem Cosinus des Winkels zwischen den entsprechenden Punkten (mit dem Ursprung des Faktorraums als Spitze des Winkels). Man sieht nun sofort, dass der outlier zu manchen Punkten einen sehr engen Winkel aufweist, d.h. bei diesem Aehnlichkeitsmass ueberhaupt nicht mehr als ein solcher gekennzeichnet ist. In unserem konstruierten Beispiel wuerde dieser Punkt sogar schon auf einer relativ niedrigen Hierarchiestufe dem richtigen Cluster zugeordnet werden.

Die Situation wuerde sich jedoch entscheidend aendern, wenn die Zahl der Populationen deutlich groesser waere als in unserem Beispiel, d.h. wenn weitere Punkteschwaerme in groesserem Abstand um den Ursprung des Faktorraums gruppiert waeren. Dann waere eine Unterscheidung zwischen "aeusseren" und "inneren" Stichproben kaum mehr moeglich; die Ergebnisse muessten demnach deutlich schlechter ausfallen als bei der Wahl der Euklidschen Distanz. Bei den zitierten Untersuchungen schwankte die Zahl der Stichproben (Cluster) zwischen zwei und sechs, war damit also verhaeltnismaessig klein.
Diese Ueberlegungen fuehrte uns zu der Auffassung, dass das (im Schnitt) bessere Abschneiden einzelner Clusteralgorithmen bei der Anwendung der Korrelation als Aehnlichkeitsmass vor allem darauf zurueckzufuehren ist, dass outlier das Resultat nicht mehr in dem Masse beeintraechtigen koennen, wie das bei der Euklidschen Distanz der Fall ist.
Das Beispiel macht aber auch deutlich, dass der Unterschied bzw. die Aehnlichkeit von Elementen bei der Korrelation ganz anders definiert ist (Winkel statt Entfernung) und deshalb nicht einfach ueber das Resultat einer Clusteranalyse bewertet werden kann. Wie schon erwaehnt haengt es von der Fragestellung ab, welcher Distanz-/Aehnlichkeitskoeffizient geeignet ist.
Dazu noch ein simples Beispiel. Angenommen, man wolle Fecht-

ecke nach ihrer Aehnlichkeit gruppieren, d.h. homogene Cluster von Rechtecken bilden. Dazu muss man sich zunaechst einmal ueberlegen, wie die Aehnlichkeit von Rechtecken definiert werden soll. Zwei Rechtecke kann man z.B. dann als aehnlich ansehen, wenn sie die gleiche Form haben, unabhaengig davon, wie gross sie sind. Entscheidend waere in diesem Fall das Verhaeltnis zwischen Laenge und Breite. Man kann Rechtecke aber auch dann als aehnlich definieren, wenn ihr Flaecheninhalt ungefaehr gleich ist, ganz unabhaengig von der Form. Oder auch, wenn sie sich sowohl in der Form als auch in der Flaeche nicht deutlich unterscheiden. Man muss sich also fuer eine Definition entscheiden. Diese Entscheidung sollte inhaltlich begruendet sein und durch theoretische Vorueberlegungen gefunden werden (vgl. dazu auch MEZZICH & SOLOMON, 1980, S.14). Ein geeignetes Distanz- oder Aehnlichkeitsmass - das mag unser Beispiel gezeigt haben - kann demnach nicht ueber Evaluationsstudien gefunden werden.

II.3.5. Beschreibung des eigenen Untersuchungsansatzes

Aufbauend auf eigenen Erfahrungen (SCHEIBLER & SCHNEIDER, 1978) und die ausgiebige Analyse einer Reihe von Monte-Carlo-Studien wurde in einer eigenen Untersuchung (SCHEIBLER & SCHNEIDER, 1981) versucht, nach einer Prozedur zu verfahren, bei der einerseits fruehere Fehler und unnoetiges Beiwerk vermieden und andererseits viele Aspekte beruecksichtigt wurden, die wertvolle Informationen ueber Eigenschaften eines breiten Spektrums von Clusteranalyseverfahren liefern. Dabei schien es unerlaesslich, ein vollkommen neues Untersuchungsdesign zu entwerfen. Wir verzichteten auf die Uebernahme eines schon existierenden Ansatzes, versuchten aber die meisten Pluspunkte frueherer Untersuchungen zu uebernehmen. Trotzdem muss betont werden, dass auch dieser Ansatz nur einen Kompromiss zwischen Idealvorstellungen und der Notwendigkeit darstellt, Abstriche aus technischen, oekonomischen und formalen Gruenden vorzunehmen.

Design

In Anlehnung an die gaengigsten Modellvorstellungen der

psychologischen Methodenlehre (lineare Statistik) waehlten wir als Datenmodell die Multinormalverteilung. Zur Evaluation werden dabei Datensaetze generiert, von denen jeder als Stichprobe aus einer Mixtur (Gesamtpopulation) von multinormalverteilten (Teil-)Populationen definiert ist. Die Teilpopulationen unterscheiden sich hinsichtlich der Mittelwerte und Varianzen der einzelnen normalverteilten Variablen sowie im Hinblick auf die Korrelationen zwischen den Variablen und somit auch hinsichtlich der Faktorenstruktur.

Wir verzichteten darauf, die Variablen mit einem Messfehler (random noise) zu versehen, wie das einzelne Autoren (z.B. BLASHFIELD) tun, da man die Clusteralgorithmen dadurch vor die prinzipiell unloesbare Aufgabe stellt, die "wahren Werte" zu erraten. Wir meinen, dass eine Evaluation nur mit einer eindeutigen Datenstruktur moeglich ist.

Bei der Generierung der Daten durch automatische Ziehung aus den Mischpopulationen ergab sich das Problem, dass bei manchen Datensaetzen die Teilstichproben von fast allen Clusteranalyseverfahren problemlos getrennt werden konnten, auf der anderen Seite aber auch solche Datensaetze existierten, bei denen sich die Teilstichproben so stark ueberlappten, dass eine Trennung faktisch ueberhaupt nicht moeglich war. Diese extremen Datensaetze schlossen wir im neuen Ansatz dadurch aus, dass zunaechst solche Datensaetze eliminiert wurden, die durch das WARDsche Verfahren fehlerfrei getrennt werden konnten (Ausschluss von zu leichten Datensaetzen). In einem zweiten Schritt unterzogen wir die restlichen Datensaetze einer Diskriminanzanalyse und eliminierten all jene, die von diesem Verfahren nicht 100-prozentig separiert werden konnten. Aus dem verbliebenen Pool von Datensaetzen wurde nach einem Quotenplan eine Stichprobe von 200 Datensaetzen gezogen, bei der darauf geachtet wurde, dass eine moeglichst breite Streuung von Datesaetzen unterschiedlichen Schwierigkeitsgrades resultierte (s. Tab. 3).

---

bitte hier Tab. 3 einfuegen

Die Populationsparameter der einzelnen Datensaetze bestimmten wir in Anlehnung an BLASHFIELD durch Zufallsstichprobenziehung aus gleichverteilten Populationen, wobei wir allerdings ein groesseres Spektrum an unterschiedlichen Datensaetzen anstrebten (Tab. 4).

---

bitte hier Tab. 4 einfuegen

Auswahl der untersuchten Clusteranalyse-Verfahren

Unser Ziel bestand darin, eine moeglichst grosse Zahl von den gebraeuchlichsten und bewaehrtesten Clusterverfahren zu evaluieren. Dieses Ziel liess sich aus verschiedenen Gruenden nicht voll erreichen. Die grosse Zahl von generierten Datensaetzen, die fuer eine sinnvolle Monte-Carlo-Untersuchung notwendig ist, legt es nahe, die gesamte Analyse weitestgehend zu automatisieren (anders waere die Arbeit auch nicht zu leisten gewesen). Diese Automatisierung setzt nun aber wieder Grenzen in Bezug auf die Auswahl der Clusteranalyse-Verfahren. Wir mussten uns schon aus rein technischen Gruenden auf solche Verfahren beschraenken, die disjunkte (sich nicht ueberlappende) Cluster generieren und ausserdem keinen allzugrossen Rechenzeitbedarf aufweisen.

Weitere Grenzen ergaben sich aus der Verfuegbarkeit einzelner Algorithmen. Bestimmte Verfahren blieben deshalb unberuecksichtigt, weil sie zu grosse Restriktionen im Hinblick auf die Anzahl von Variablen und Objekten setzen, wie z.B. AUKL von FABER & NOLLAU (1969) oder EUCLID von WISHART.

Die wohl umfassendste Sammlung von Clusterverfahren ist in dem schon erwaehnten Programmsystem CLUSTAN 1C (WISHART 1975, 1977) enthalten, aus dem wir alle hierarchisch-agglomerativen Verfahren sowie das partitionierende Verfahren RELOCATE ausgewaehlt haben. Ergaenzend dazu untersuchten wir noch das von SPAETH (1975) hoch bewertete Verfahren KMEANS, das vom Algorithmus her sehr viel Aehnlichkeit mit RELOCATE aufweist.

Folgende zehn Clusteranalyse-Verfahren wurden demnach in unserer Untersuchung beruecksichtigt:

1) single-linkage-Methode
2) complete-linkage-Methode
3) average-linkage-Methode
4) Centroid-Methode
5) Median-Methode
6) WARD-Methode
7) LANCE-WILLIAMS-flexible-beta-Methode
8) McQUITTYs Aehnichkeits-Analyse
9) Prozedur RELOCATE
10) Prozedur KMEANS (nach SPAETH, 1975).

Bei allen Verfahren wurde ausschliesslich die dem Datenmodell alleine angemessene Euklidsche Distanz als Unaehnlichkeitskoeffizient eingesetzt.
Bei dem Verfahren von LANCE-WILLIAMS wurde der Beta-Wert[1] variiert, was uns die Moeglichkeit gab, die optimale Groesse dieses Parameters zu schaetzen.
Die nicht-hierarchischen Verfahren RELOCATE und KMEANS erfordern eine Anfangspartitionierung (vorlaeufige Clusterloesung). Dabei bieten beide Verfahren die Moeglichkeit, entweder eine "Zufallspartitionierung" vom Programm generieren zu lassen oder "von aussen" eine vorlaeufige Clusterloesung einzugeben, die dann von dem Verfahren optimiert wird. Beide Alternativen wurden erprobt, wobei wir bei der zweiten jene Clusterloesung vorgaben, die durch das WARDsche Verfahren gewonnen wurde. Da diese Technik im allgemeinen schon sehr gute Loesungen liefert, interessierte uns die Frage, wieweit diese Loesung ueberhaupt noch verbessert werden kann.

Zu RELOCATE und KMEANS sei noch erwaehnt, dass sich beide Verfahren auf den von MacQUEEN (1967) entwickelten Algorithmus "k-means" stuetzen (denselben Algorithmus verwendet McRAE (1971) in seinem Programm MIKCA). Der entscheidende Unterschied zwischen beiden Techniken besteht darin, dass als Repraesentant fuer ein Cluster bei RELOCATE der Centroid

---

[1] Zur Bedeutung dieses Parameters vgl. die Beschreibung im Anhang, in dem auch die uebrigen Verfahren kurz dargestellt sind

gewaehlt wird, bei KMEANS dagegen der Median des betreffenden Clusters (vgl. im Anhang die Beschreibung beider Verfahren).
Ein weiterer Unterschied ist dadurch gegeben, dass bei KMEANS eine andere Form der Standardisierung der Daten durchgefuehrt wird. Dieses Verfahren faellt damit etwas aus dem Rahmen, der durch die einheitliche z-Transformation bei den anderen Clusteralgorithmen gesetzt wurde. Eine Angleichung von KMEANS waere zwar leicht moeglich gewesen, doch stellt der Programmautor (SPAETH 1975) diese Form der Standardisierung als anderen ueberlegen dar, weshalb wir es dabei beliessen und das Verfahren in der Form testeten, wie es von SPAETH propagiert wird. Durch den Vergleich mit RELOCATE bietet sich damit die Moeglichkeit zu pruefen, inwieweit sich die genannten Unterschiede auswirken.

Bestimmung der Guete der Clusterloesungen

In Anlehnung an BLASHFIELD (1976) waehlten wir als Mass fuer die Guete der Clusterloesungen den in den elaborierteren Monte-Carlo-Studien haeufig verwendeten Kappa-Koeffizienten. Dieses Uebereinstimmungsmass kann wie der Korrelationskoeffizient zwischen -1 und +1 variieren, wobei fuer unsere Fragestellung nur der positive Wertebereich von Bedeutung ist. Ein Kappa-Wert von 1 gibt an, dass vollstaendige Uebereinstimmung zwischen den Clustern und den Teilstichproben besteht, m.a.W. die optimale Clusterloesung gefunden worden ist. Bei einem Kappa von 0 gibt es dagegen ueberhaupt keine Uebereinstimmung zwischen Stichproben- und Clusterstruktur, was als schlechteste Loesung gelten muss.
Wie schon oben angedeutet, liegt ein kritischer Punkt bei der Evaluation von Clusteranalyse-Verfahren in der Existenz von 'outliern', also Einzelelementen, die bei bestimmten Verfahren Mini-Cluster bilden. Diese Verfahren schneiden demnach besonders schlecht ab, wenn sie genausoviele Cluster bilden sollen, wie Teilstichproben vorhanden sind (dies moegen bereits die Ausfuehrungen unter II.3.4. sowie Abb. 3 verdeutlicht haben). Es scheint daher angebracht, solche Mini-Cluster bei der Analyse gesondert zu beruecksichtigen.

Um das Verhalten der einzelnen Verfahren gegenueber solchen

Mini-Clustern genauer zu untersuchen, wurde eine Vorgehensweise gewaehlt, die schrittweise immer mehr Mini-Cluster in die Berechnung der Kappa-Koeffizienten mit einbezieht. Dabei liessen wir von jedem Clusteranalyse-Verfahren zunaechst k+9 Cluster bestimmen (k = Anzahl der Subpopulationen bzw. Teilstichproben), und schlossen die 9 kleinsten Cluster (Rest-Cluster) von der Analyse aus. Es wurden demnach nur die k groessten Cluster mit den k vorgegebenen Teilstichproben verglichen und die Uebereinstimmung per Kappa-Koeffizient ermittelt. Daraufhin wurde die Zahl der Restcluster schrittweise um 1 reduziert, bis die Anzahl der Cluster der Anzahl der Unterstichproben exakt entsprach, und fuer jede der zehn Stufen der Kappa-Koeffizient jeweils neu berechnet.

Zusammenfassende Darstellung der Methode

Vereinfacht dargestellt, verlief unsere Untersuchung fuer jeden Cluster-Algorithmus in folgenden Schritten: zunaechst wurde ein Datensatz ueber einen Zufallsgenerator erzeugt. Dieser Datensatz war als eine Ziehung von k Stichproben aus einer Mischpopulation von k multinormaverteilten Subpopulationen definiert. Der generierte Datensatz wurde von den Algorithmen in einzelne Cluster zerlegt. Dabei lieferte jedes CA-Verfahren 10 Loesungen mit maximal k+9 und minimal k Clustern. Anschliessend wurden alle Clusterloesungen mit den Teilstichproben verglichen und Kappa-Werte zur Bestimmung der Reproduktionsguete berechnet. Diese Prozedur wurde pro Algorithmus 200-mal wiederholt, d.h. es wurden 200 Datensaetze generiert und clusteranalysiert. Die Ergebnisse wurden gespeichert und nach Abschluss der 200 Durchgaenge ausgewertet.
Diese Prozedur unterscheidet sich nur geringfuegig von der in Abb. 1 dargestellten Vorgehensweise. Es entfaellt lediglich die Diskriminanzanalyse, da sie sich aufgrund der Vorselektion der Datensaetze eruebrigt (die vollstaendige Trennbarkeit der Teilstichproben ist gewaehrleistet). Die genannte Prozedur laeuft nahezu vollautomatisch auf der Grossrechenanlage (IBM 370/168) des URZ Heidelberg.[1]

---

[1] Nur am Rande sei erwaehnt, dass diese so simpel klingende Vorgehensweise enorme Anstrengungen bei der Programmierung

Ergebnisse

Das Resultat unserer Analyse ist in Tab. 5 zusammengefasst. Es werden die Mediane von jeweils 200 Kappa-Werten wiedergegeben (je ein Median pro Verfahren und Analyse-Stufe).

---

bitte hier Tab. 5 einfuegen

Aus der Tabelle geht hervor, dass die Ergebnisse der Clusteranalysen umso schlechter ausfallen, je mehr Mini-Cluster aus der Analyse ausgeschlossen werden. Der Verlauf der Kappa-Koeffizienten macht aber auch deutlich, dass es nicht genuegt, nur die 'letzte' Analysestufe zu interpretieren.
Aehnlich wie in den meisten der oben zitierten Untersuchungen schneidet auch in dieser Simulationsstudie das single-linkage-Verfahren am schlechtesten ab. Auf der letzten Analysestufe unterscheidet es sich zwar praktisch kaum von der Median- und Centroid-Methode, doch zeigen jene Verfahren auf den frueheren Stufen durchweg bessere Ergebnisse (wobei ein gewisser Vorsprung der Centroid-Methode erkennbar ist).
Die complete-linkage-Methode liefert bei k Clustern zwar das beste Ergebnis der 5 erstgenannten Verfahren, doch schneidet sie bei steigender Cluster-Anzahl schlechter als die average-linkage-Methode sowie die Median- bzw. Centroid-Technik ab. Insgesamt gesehen wird damit die 'outlier'-Anfaelligkeit der 5 erstgenannten Verfahren bestaetigt. Aus Tab. 5 geht aber auch hervor, dass diese Algorithmen nicht annaehernd an das ueberlegene Resultat der WARD-Methode herankommen. Lediglich der Algorithmus von LANCE & WILLIAMS scheint der WARD-Technik ebenbuertig zu sein (insbesondere fuer beta = -.5). Es faellt auf, dass sich beide Verfahren als aeusserst robust gegenueber outliern erweisen.
Die iterativ-partitionierenden Verfahren RELOCATE und KMEANS erbrachten nur mittelmaessige Ergebnisse, wenn sie mit einer

erforderte (so hat die Entwicklung des Systems bis zur vollen Perfektion mehr als ein Jahr in Anspruch genommen, und auch die nachfolgende Arbeit war mit grossen Rueckschlaegen verbunden)

Zufallspartitionierung starten mussten. Hier zeigt sich der Nachteil ihres heuristischen Algorithmus', der darin zu sehen ist, dass sie sich bei schlechten Startkonfigurationen im lokalen (nicht absoluten) Minimum verfangen. Diese Algorithmen sind demnach nur zu empfehlen, wenn die Anfangspartitionierung hinreichend gut ist, wenn also z.B. die Ergebnisse des WARD-Verfahrens dazu herangezogen werden. Ueberraschend ist allerdings, dass beide Verfahren unter dieser Bedingung nicht besser als die WARD-Technik abschneiden, ja im Schnitt sogar geringfuegig schlechtere Ergebnisse liefern. Dieser Befund sollte aber mit Vorsicht interpretiert werden. Eine genauere Analyse der Einzelergebnisse zeigt naemlich, dass etwa die Haelfte der Kappa-Koeffizienten kleinere und die andere Haelfte gleiche oder groessere Werte enthielt. In allen Faellen, in denen RELOCATE das Ergebnis von WARD (i.d.R. nur geringfuegig) modifizierte, wurde die Summe der Abstandsquadrate innerhalb der Cluster vermindert. Es ist nun allerdings schwer zu sagen, ob ein anderes Minimierungskriterium bessere Ergebnisse liefern wuerde. Bleibt in diesem Zusammenhang noch zu erwaehnen, dass RELOCATE und KMEANS bei sehr schwer separierbaren Stichproben tendenziell besser abschnitten.

Schliesslich sei noch am Rande erwaehnt, dass die in unserer Studie auf der 'letzten' Analysestufe aufgefundenen Mediane der Kappa-Werte fuer das single-linkage-, complete-linkage-, average-linkage-Verfahren sowie den WARD-Algorithmus fast exakt mit den Ergebnissen von BLASHFIELD (1976) uebereinstimmen (s. Tab. 1). Dies ueberrascht vor allem deshalb, weil eine Rekonstruktion der BLASHFIELD-Ergebnisse ueberhaupt nicht angestrebt wurde und einige Unterschiede zwischen den beiden Untersuchungsansaetzen bestehen. Die Aehnlichkeit der Ergebnisse im Hinblick auf die absoluten Kappa-Werte muss daher als Zufall gewertet werden. Allerdings wird damit erneut die von BLASHFIELD gefundene Rangordnung der Verfahren, insbesondere die Ueberlegenheit der WARD-Technik bestaetigt.

III. Zusammenfassung und Diskussion

Die Uebersicht ueber die in den unterschiedlichen wissenschaftlichen Disziplinen durchgefuehrten Versuche zur Bewertung von Clusteranalyseverfahren mag einen ungefaehren Eindruck davon vermittelt haben, welche Schwierigkeiten mit einer eindeutigen Einordnung der gepruefteN Algorithmen verbunden sind.
Es liess sich zunaechst einmal klar herausstellen, dass die Vergleichsuntersuchungen anhand von empirischen Datensaetzen wenig effizient sein konnten: die "wahre" Datenstruktur ist in solchen Faellen einfach zu wenig bekannt, um als Aussenkriterium benutzt werden zu koennen. Nicht viel anders verhaelt es sich mit Evaluationsstudien, die sich auf real gemessene Datensaetze mit bekannter Struktur, m.a.W. auf Plasmoden stuetzen. Wenn solchen Untersuchungen sicherlich auch nach BAUMANN (1973) mehr Aussagekraft im Sinne einer grundsaetzlichen Validitaet zugestanden werden muss, bleibt ihre Generalisierbarkeit angesichts der Vorliebe vieler Forscher, immer wieder auf "klassische" Datensaetze zurueckzugreifen, doch aeusserst umstritten. Argumente fuer eine solche Sichtweise lassen sich (von den Autoren im uebrigen unbeachtet) aus einer Untersuchung von MEZZICH & SOLOMON (1980) gewinnen, bei der insgesamt vier verschiedene Plasmoden-Datensaetze herangezogen wurden, um die Leistungsfaehigkeit von 18 Cluster-Algorithmen zu vergleichen. Aus einer Ergebnisuebersicht (vgl. S. 135) kann fuer mehrere CA-Verfahren nachgewiesen werden, dass sich ihre Rangplatzziffer ueber die vier Datensaetze hinweg z.T. betraechtlich veraendert (so belegt die Prozedur KMEANS z.B. bei der Analyse von "Iris"-Daten den 18. und letzten, bei Daten zum "treatment environment" bei psychiatrischen Patienten dagegen den zweiten Rang).

Als Ausweg aus diesem Dilemma schienen Monte-Carlo-Simulationen geeignet zu sein, bei denen sich ueber die Ziehung von Zufallszahlen Datenstichproben aus Populationen mit bekannten Verteilungsparametern erzeugen lassen, wobei hier nun die Moeglichkeit gegeben ist, (annaehernd) beliebig viele Datenstichproben zu erzeugen und damit das "Verhalten"

von CA-Algorithmen systematischer zu analysieren.
Erste Schritte in diese Richtung waren aber deshalb nur von eingeschraenktem Wert, weil entweder zu wenige Algorithmen in die Untersuchung einbezogen wurden oder aber (aus welchen Gruenden auch immer) immer noch zu wenige Datenstichproben resultierten.
Die Uebersicht ueber die Befunde komplexerer Simulationsstudien (wie sie etwa in Tab. 1 gegeben wird) zeigt demgegenueber im Vergleich zu den erwaehnten Plasmodenstudien einen deutlichen Fortschritt, was die Rangplatz-Varianz von Verfahren in unterschiedlichen Untersuchungen angeht: die (mit einer Ausnahme) nahezu perfekte Korrelation zwischen diesen Rangreihen weist auf ein stabiles Ergebnismuster fuer den Fall hin, dass die Algorithmen alle Elemente klassifizieren muessen. Unter dieser Voraussetzung schneiden von den hierarchisch-agglomerativen Verfahren insbesondere das WARD- sowie das LANCE-WILLIAMS-Verfahren und (mit Abstrichen) die complete-linkage-Methode gut ab, wie sich auch fuer die herangezogenen iterativ-partitionierenden Verfahren guenstige Vergleichswerte registrieren lassen. Fuer letztere ist dies aber immer nur dann der Fall, wenn fuer die Startpartition keine Zufallswerte, sondern die Ergebnisse einer anderen (hierarchischen) Clusteranalyse gewaehlt worden sind.
Wie EDELBROCK als erster zeigen konnte, muss die hier vorgenommene Einstufung der einzelnen Algorithmen jedoch noch differenziert werden, wenn man den Verfahren wirklich gerecht werden will. Insbesondere die "outlier"-anfaelligen Techniken werden naemlich dadurch benachteiligt, dass hier die Anzahl der Cluster gleich der Anzahl der Subpopulationen gesetzt wird und gleichzeitig alle Elemente klassifiziert werden muessen. Wie die Ergebnisse von EDELBROCK und SCHEIBLER & SCHNEIDER (1981) zeigen, werden auch von allgemein schwach eingestuften Verfahren wie der single-linkage-, der average-linkage- und der Median-Methode durchaus befriedigende Loesungen erzielt, wenn nur bestimmte Prozentsaetze der Elemente bei der Klassifikation zu beruecksichtigen sind. Geht man davon aus, dass in der psychologischen Untersuchungspraxis absolut vollstaendige Zuordnungen nur in den wenigsten Faellen erforderlich sind, ist diese Information

ueber das 'Verhalten' von Algorithmen bei der Ausklammerung von outliern sicherlich sehr nuetzlich.
Dies aendert allerdings nichts daran, dass auch bei einer solchen Betrachtungsweise die oben als ueberlegen eingestuften Algorithmen ihren Vorsprung (wenn auch nicht mehr so ausgepraegt) behalten. Es kann also aufgrund der vorliegenden Monte-Carlo-Studien kaum ein Zweifel daran bestehen, dass die WARD-Methode und das LANCE-WILLIAMS-Verfahren sowie die iterativ-partitionierenden Techniken RELOCATE und KMEANS (letztere bei geeigneter Anfangspartitionierung) mit Abstand die robustesten Algorithmen darstellen und damit sicherlich zur Anwendung zu empfehlen sind. Es muss allerdings auch hier die Einschraenkung gemacht werden, dass auch bei den Monte-Carlo-Studien nicht von genereller Validitaet ausgegangen werden kann.
Schon bei der Durchsicht der Ergebnisse von Tab. 1 fiel auf, dass die Absolutwerte der fuer die Bewertung verwendeten Guetemasse in den verschiedenen Studien z.T. erheblich differieren. Diese Unterschiede lassen sich dabei kaum vollstaendig auf die unterschiedliche Definition der theoretischen Cluster-Struktur in den einzelnen Untersuchungen (also die alternative Konzeption von 'mixture model' vs. 'ultrametrischer Raum', s.o.) zurueckfuehren, sondern weisen eher darauf hin, dass unterschiedliche Modi der Zufallsdaten-Generierung benutzt worden sind. Die bei bestimmten Studien (z.B. MILLIGAN, 1980) besonders evidenten systematischen Abweichungen der erzielten Uebereinstimmungs-Werte vom 'Durchschnitt' der Evaluationsstudien machen nachdruecklich darauf aufmerksam, dass unter dem Etikett 'Monte-Carlo-Studie' durchaus unterschiedliche Vorgehensweisen subsummiert sind.
Wenn auch in der eigenen Untersuchung ein sehr breites Spektrum an Datensaetzen untersucht wurde, deren Zahl hinreichend gross schien, um die Guete der Verfahren zuverlaessig schaetzen zu koennen, muss man dennoch mit der Generalisierung der Befunde vorsichtig sein. Strenggenommen erstreckt sich ihr Gueltigkeitsbereich nur auf solche Daten vom selben Typ, d.h. auf Stichproben aus Mischpopulationen, die sich aus multinormalverteilten Subpopulationen zusammensetzen. Es bleibt unklar, wie sich die CA-Verfahren z.B. bei extrem schiefen (assymetrischen) Verteilungen verhalten. Das hier

verwendete Datenmodell erzeugt sphaerische (elliptoide) Teil-Stichproben. Es bleibt damit auch offen, wie die zur sphaerischen Clusterbildung tendierenden Algorithmen (WARD-Methode, LANCE-WILLIAMS-Verfahren und RELOCATE bzw. KMEANS) bei nicht-sphaerischen Stichproben abgeschnitten haetten. Man kann allerdings davon ausgehen, dass die von uns ermittelten "Spitzen-Verfahren" immer dann gute Ergebnisse liefern, wenn die Verteilungen (Dichten) der Subpopulationen eingipflig und nicht extrem schief sind.

Wie die unter II.3.4. eingebrachte exemplarische Erlaeuterung gezeigt hat, duerften die unterschiedlichen Eigenschaften der generierten Datensaetze einen wesentlichen Einfluss auf die Frage haben, ob die einzelnen CA-Algorithmen nun besser mit Korrelations- oder Distanzmassen zu kombinieren sind. Es liegt demnach nahe, dass insbesondere die Art der Zufallszahlengenerierung fuer die in einigen neueren Untersuchungen aufgefundenen Resultate verantwortlich zu machen ist, denen zufolge bei der Wahl von Korrelationskoeffizienten guenstigere Befunde zu erwarten sind. Wir gehen davon aus, dass diese Frage aufgrund der bisher durchgefuehrten Evaluationsstudien nicht eindeutig zu beantworten ist.

Wie schon eingangs erwaehnt, wurde die hier vorgelegte Beurteilung von CA-Verfahren nicht zuletzt auch unter dem Aspekt vorgenommen, Aussagen zur praktischen Handhabbarkeit bzw. zur Benutzerfreundlichkeit machen zu koennen. Die hier analysierten Verfahren wurden alle dem Software-Paket CLUSTAN 1C von WISHART (1975) bzw. im Fall von KMEANS aus der Sammlung von "stand-alone"-Programmen bei SPAETH (1975, 1980) entnommen. Sind schon die erforderlichen Vorkenntnisse der Benutzer bei CLUSTAN 1C erheblich zu nennen, so koennen die bei SPAETH (wie auch etwa die bei STEINHAUSEN & LANGER, 1977) publizierten Programme nur mit einigermassen guten FORTRAN-Kenntnissen sachgemaess eingesetzt werden (bei KMEANS ist darueberhinaus nicht auszuschliessen, dass bei bestimmten Datensaetzen Endlosschleifen produziert werden (s. MOEBUS, 1981)).
Gerade fuer den mit EDV-Problemen weniger vertrauten Benutzer stellt es sich nun als geradezu fatal heraus, dass

die grossen und komfortablen Programmsysteme in dieser Hinsicht unzureichend ausgestattet sind. Waehrend im SPSS bis dato kein Cluster-Algorithmus eingebaut ist, stellt das Angebot in den beiden anderen relativ weit verbreiteten Systemen (SAS) und BMDP) allenfalls zweite Wahl dar. Das SAS verfuegt lediglich ueber den complete-linkage-Algorithmus, der den Evaluationsstudien zufolge eher im unteren Durchschnittsbereich lag.[1]

Auch im BMDP wird lediglich ein einziges Verfahren praesentiert, dessen Identifikation nicht ganz leicht faellt, weil es sich um ein eigens konstruiertes und nur ungenau beschriebenes Programm handelt. Es spricht einiges dafuer, dass dieser hierarchisch-agglomerative Algorithmus als Modifikation der Centroid-Methode zu werten ist, die bekanntlich in den Evaluationsstudien zusammen mit der single-linkage-Technik am schlechtesten abschnitt.
Vielleicht moegen diese kritischen Anmerkungen einen Anstoss dazu geben, dass in den genannten benutzerfreundlichen Software-Paketen effizientere CA-Algorithmen installiert werden. Sollte dem in absehbarer Zeit nicht so sein, wird dem auf Ergebnisqualitaet bedachten Anwender kaum etwas anderes uebrig bleiben, als in den sauren Apfel zu beissen und seine Frustrationstoleranz im Sinne des (im URZ HD aufgefundenen) Spruechleins zu testen:

"Die Haelfte seines Lebens
wartet der Programmierer vergebens".

---

[1] Benutzer des URZ Heidelbergs haben allerdings den Vorteil, dass von Dr. Peter Beutel eine SAS-Schnittstelle zum CLUSTAN 1C hergestellt wurde, so dass hier einige Probleme entfallen

Literaturverzeichnis

ALDENDERFER, M.S.A.: A consumer report on cluster analysis software: (2) Hierarchical methods. Report No.2 from N.S.F. grant 74-20007. Gainesville, Florida, July 1977

ALLMER, H.: Taxonome-Programm und Automatische Klassifikation in der Anwendung: eine Vergleichsstudie. Psychologische Beitraege, 1974, 16, 605-617

ANDERBERG, M.R.: Cluster analysis for applications. New York: Academic Press, 1973

BAKER, F.B. & HUBERT, L.J.: Measuring the power of hierarchical cluster analysis. Journal of the American Statistical Association, 1975, 70, 1975

BARTKO, J.J.: A note on the intraclass correlation coefficient as a measure of reliability. Psychological Reports, 1974, 34, 418

BARTKO, J.J.: On various intraclass correlation reliability coefficients. Psychological Bulletin, 1976, 83, 762-765

BARTKO, J.J., STRAUSS, J.S. & CARPENTER, W.T.: An evaluation of taxometric techniques for psychiatric data. Classification Society Bulletin, 1971, 2, 2-28

BAUMANN, U.: Psychologische Taxometrie - Eine Methodenstudie ueber Aehnlichkeitskoeffizienten, Q*-Clusteranalyse, Q-faktorenanalyse. Bern: Huber, 1971

BAUMANN, U.: Die Konfigurationsfrequenzanalyse, ein taxometrisches Verfahren. Psychologische Beitraege, 1973, 15, 153-168

BENTLER, P.M.: Multivariate analysis with latent variables: Causal modeling. Annual Review of Psychology, 1980, 31, 419-456

BLASHFIELD, R.K.: Mixture model tests of cluster analysis: Accuracy of four hierarchical agglomerative methods. Psychological Bulletin, 1976, 83, 377-388

BLASHFIELD, R.K.: A consumer report on cluster analysis software: (3) Iterative partitioning methods. Report No.3 from N.S.F. grant 74-20007. Gainesville, Florida, March 1977 (a)

BLASHFIELD, R.K.: A consumer report on cluster analysis software: (4) Useability. Report No.4 from N.S.F grant 74-20007. Gainesville, Florida, March 1977 (b)

BLASHFIELD, R.K.: Basic features for 20 mixtures. Unpubl. paper, Gainesville, Florida, 1977 (c)

BLASHFIELD, R.K.: Summary from later mixture studies. Unpubl. paper, Gainesville, Florida, 1978

BLASHFIELD, R.K.: The growth of cluster analysis: Tryon, Ward, and Johnson. Multivariate Behavioral Research, 1980, 15, 439-458

BLASHFIELD, R.K. & ALDENDERFER, M.S.: The literature on cluster analysis. Multivariate Behavioral Research, 1978, 13, 271-295

CATTELL, R.B. & COULTER, M.A.: Principles of behavioral taxonomy and mathematical basis of the taxonomic computer program. British Journal of Mathematical and Statistical Psychology, 1966, 19, 237-269

CORMACK, R.M.: A review of classification. Journal of the Royal Statistical Society - A, 1971, 134, 321-367

COHEN, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46

CUNNINGHAM, K.M. & OGILVIE, J.C.: Evaluation of hierarchical grouping techniques: A preliminary study. The Computer Journal, 1972, 15, 209-213

D'ANDRADE, R.G.: U-statistic hierarchical clustering. Psychometrika, 1978, 43, 59-67

ECKES, T. & ROSSBACH, H.: Clusteranalysen. Stuttgart: Kohlhammer, 1980

EDELBROCK, C.: Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. Multivariate Behavioral Research, 1979, 14, 367-384

EDELBROCK, C. & McLAUGHLIN, B.: Hierarchical cluster analysis using intraclass correlations: A mixture model study. Multivariate Behavioral Research, 1980, 15, 299-318

EVERITT, B.: Cluster analysis. London: Heinemann, 1974

FABER, E. & NOLLAU, W.: Ueber ein Verfahren zur automatischen Klassifikation. Schriftenreihe des DRZ, Heft S-6, Darmstadt, 1969

FISHER, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936, 7, 179-188

GOLDSTEIN, S.G. & LINDEN, J.D.: A comparison of multivariate grouping techniques commonly used with profile data.

Multivariate Behavioral Research, 1969, 4, 103-114

GOODMAN, L.A. & KRUSKAL, W.H.: Measures of association for cross-classification. Journal of the American Statistical Association, 1954, 49, 732-764

GROSS, A.C.: A Monte Carlo study of the accuracy of a hierarchical grouping procedure. Multivariate Behavioral Research, 1972, 7, 379-389

HARTMANN, W.: Ueber ein Verfahren der numerischen Taxonomie von Cattell und Coulter. Biometrische Zeitschrift, 1976, 18, 273-290 (a)

HARTMANN, W.: Ueber einen Algorithmus zur Clusteranalyse maximal-kompakter Gruppen und die rechentechnische Realisierung des Verfahrens von Cattell und Coulter. Biometrische Zeitschrift, 1976, 18, 333-349 (b)

HUBERT, L.: Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. Journal of the American Statistical Association, 1974, 69, 698-704

HUBERT, L. & BAKER, F.B.: Data analysis by single-link and complete-link hierarchical clustering. Journal of Educational Statistics, 1976, 1, 87-111

HUBERT, L. & LEVIN, J.R.: Evaluating object set partitions: Free sort analysis and some generalizations. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 459-470

KUYPER, F.K. & FISHER, C.: A Monte Carlo comparison of six clustering procedures. Biometrics, 1975, 31, 777-783

LORR, M., KLETT, C.J. & McNAIR, D.M.: Syndromes of psychosis. New York: Macmillan, 1963

MacQUEEN, J.B.: Some methods for classification and analysis of multivariate observations. 5th Berkeley Symposium on Mathematics, Statistics, and Probability, Vol.1, 1967, 281-297

McINTYRE, R.M. & BLASHFIELD, R.K.: A nearest-centroid technique for evaluating the minimum-variance clustering procedure. Multivariate Behavioral Research, 1980, 15, 225-238

McRAE, D.J.: MICLA: A FORTAN IV iteration k-means cluster analysis program. Behavioral Science, 1971, 16, 423-324

MEZZICH, J.E.: An evaluation of quantitative taxonomic methods. Diss.Abstr.Intern., 1975, 36, 3008-B

MEZZICH, J.E.: Evaluating clustering methods for psychiatric diagnosis. Biological Psychiatry, 1978, 13, 265-281

MEZZICH, J.E. & SOLOMON, H.: Taxonomy and behavioral science - Comparative performance of grouping methods. New York: Academic Press, 1980

MILLIGAN, G.W.: An examination of the effect of six effects of error perturbation on fifteen clustering algorithms. Working Paper Series 78-99, Academic Faculty of Management Sciences, The Ohio State University, December, 1978

MILLIGAN, G.W.: An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 1980, 45, 325-342

MILLIGAN, G.W. & ISAAC, P.D.: The validation of four ultrametric clustering algorithms. Pattern Recognition, 1980, 12, 41-50

MOEBUS, C.: On the nontermination of the k-means clustering algorithm for certain data sets. Unpubl. paper, University of Oldenburg, 1981

MOJENA, R.: Hierarchical grouping methods and stopping rules: An evaluation. The Computer Journal, 1977, 20, 359-363

ROGERS, G. & LINDEN, J.D.: Use of multiple discriminant function analysis in the evaluation of three multivariate grouping techniques. Educational and Psychological Measurement, 1973, 33, 787-802

RAND, W.M.: Objective criteria for evaluating clustering methods. Journal of the American Statistical Association, 1971, 66, 846-850

RUBIN, J. & FRIEDMAN, H.P.: A cluster analysis and taxonomy system for grouping and classifying data - computer program. IBM Corporation, New York Scientific Center, New York, 1967

SCHEIBLER, D. & SCHNEIDER, W.: Probleme und Ergebnisse bei der Evaluation von Clusteranalyse-Verfahren. Bericht aus dem Psychologischen Institut der Universitaet Heidelberg, Nr. 11, Juni 1978

SCHEIBLER, D. & SCHNEIDER, W.: A Monte-Carlo study of the accuracy of cluster analysis algorithms: A comparison of hierarchical and iterative partitioning methods. Unpubl. paper, University of Heidelberg, 1981

SNEATH, P.H.A.: A comparison of different clustering methods

as applied to randomly spaced points. Classification Society Bulletin, 1966, 1, 2-18

SPAETH, H.: Cluster-Analyse-Algorithmen zur Objektklassifikation und Datenreduktion. Muenchen: Oldenbourg, 1975

SPAETH, H.: Cluster analysis algorithms for data reduction and classification of objects. New York: Wiley, 1980

STEINHAUSEN, D. & LANGER, K.: Clusteranalyse. Berlin: de Gruyter, 1977

TRYON, R.C.: Cluster analysis. Ann Arbor: Edward Brothers, 1939

VOGEL, F.: Probleme und Verfahren der automatischen Klassifikation. Goettingen: Vandenhoeck & Ruprecht, 1975

Von Eye, A.: Zum Vergleich zwischen der hierarchischen Clusteranalyse nach WARD und MACS, einer mehrdimensionalen, automatischen Clustersuchstrategie. Psychologische Beitraege, 1977, 19, 201-217

WISHART, D.: Clustan 1C user manual. London: Computer Center, 1975

WISHART, D.: CLUSTAN user manual supplement. Edinburgh, 1977

WOLFE, J.H.: Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, 1970, 5, 329-350

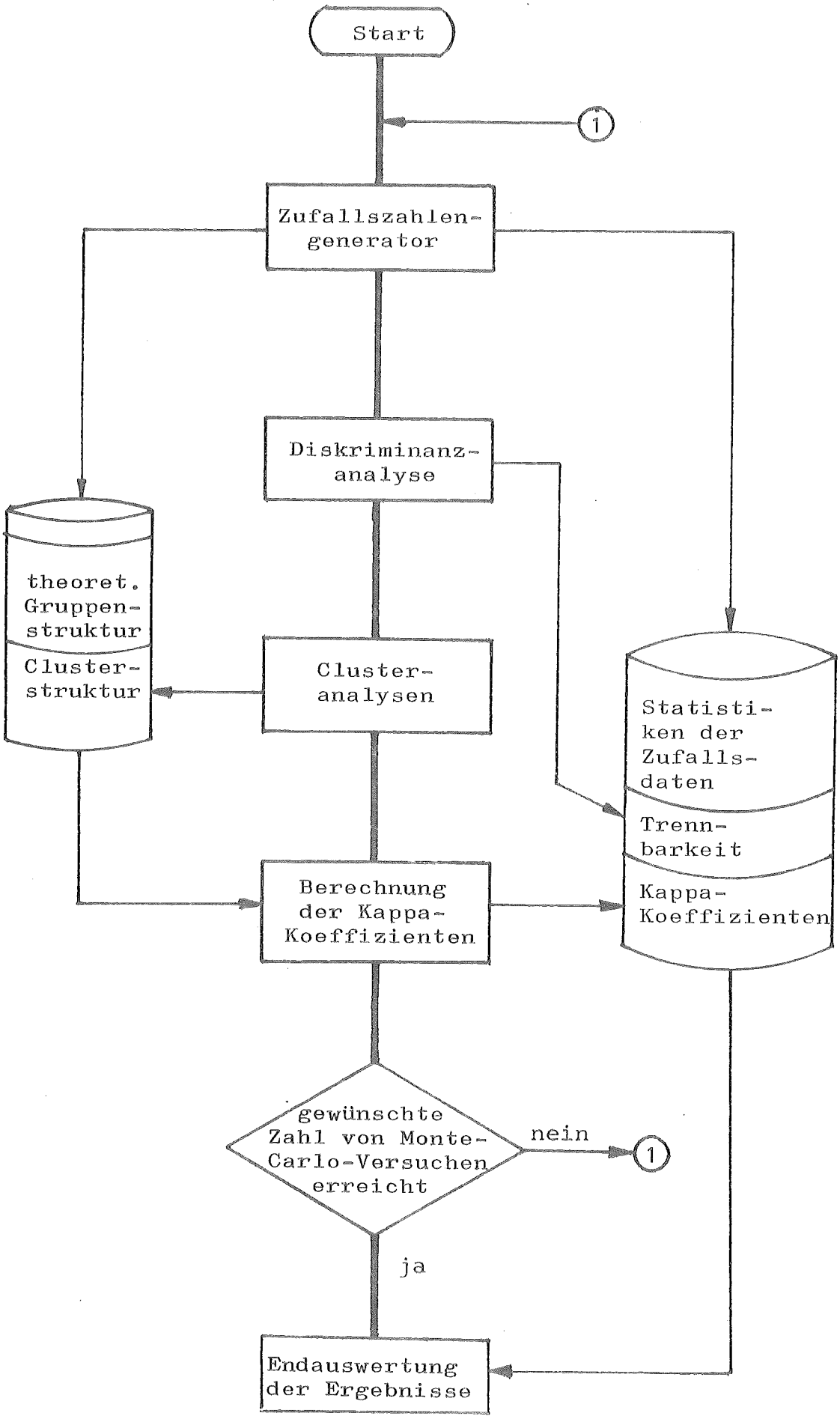
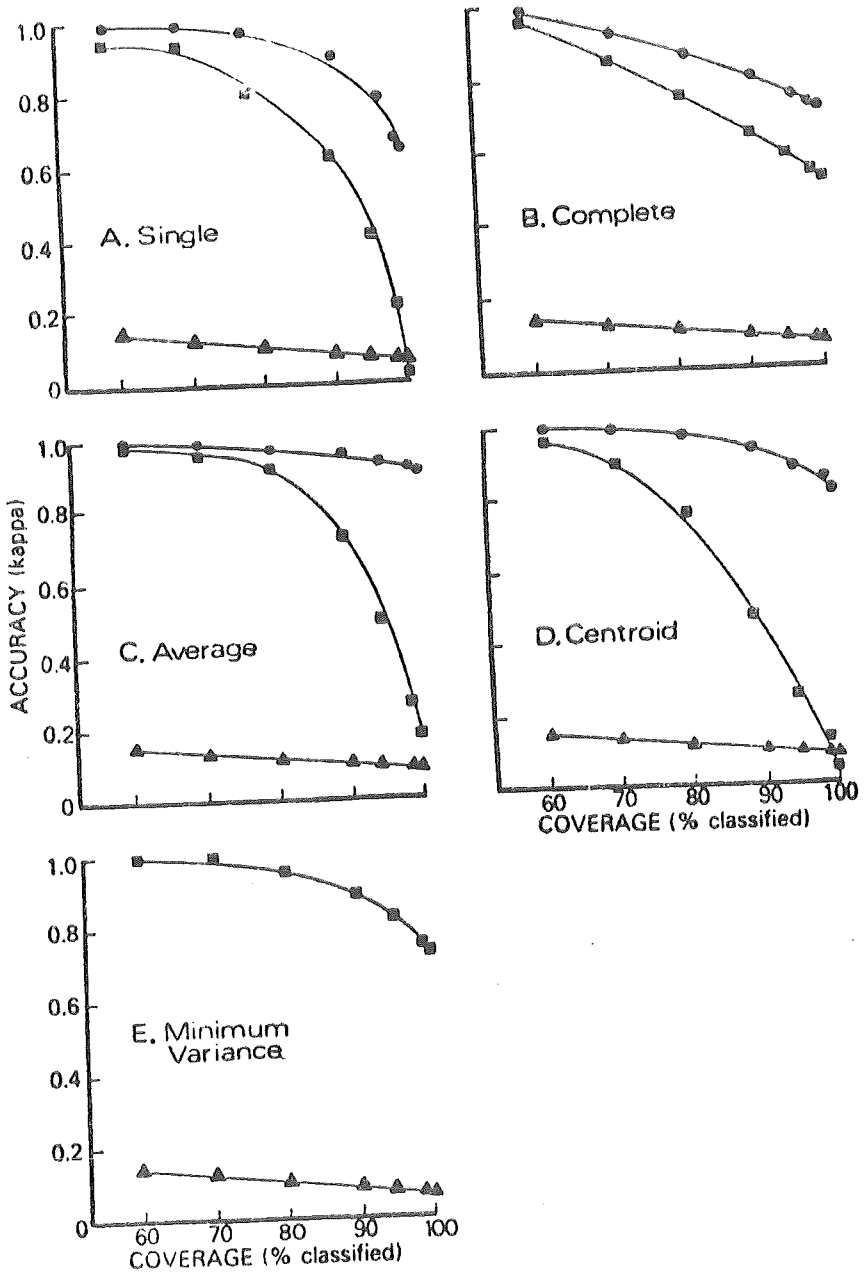


Abb. 1: Exemplarisches Ablaufschema für eine Monte-Carlo-Simulation



b. 2: Reproduktionsgüte des single-linkage- (A), complete-linkage-, (B), average-linkage- (C), Centroid- (D) und WARD-Verfahrens (E) für unterschiedliche Prozentsätze zu klassifizierender Elemente nach EDELBROCK (1979). Ergebnisse mit Korr.-Koeffizienten sind mit Kreisen, die mit Distanzkoeffizienten mit Quadraten gekennzeichnet. Für die Resultate eines Zufalls-Algorithmus sind Dreiecke eingesetzt

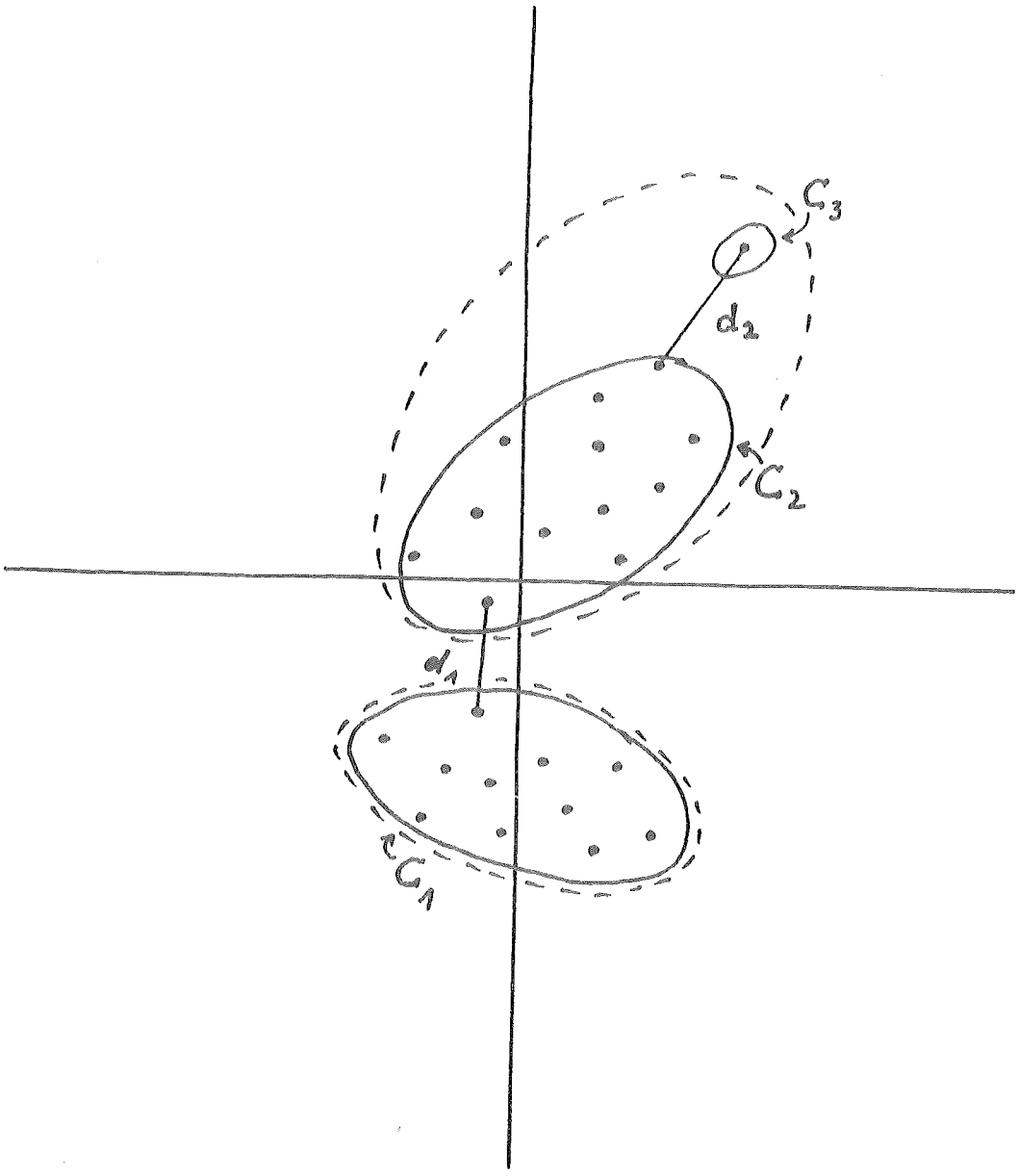


Abb. 3: Veranschaulichung des Effekts von 'outliern' auf das Resultat von Cluster-Algorithmen (z.B. von linkage-Methoden)

ıdie	1 (n=50) Kappa	2 (n=40) Gamma	3 (n=30) RAND	4 (n=766) Kappa	5 (n=12) RAND	6 (n=108) RAND	7 (n=250) CRAMERs V
ıgle-link.	.06(4)	.52 (3)	.44 (6)	.30 (9)	.37 (6)	.91 (6,5)	.15 (5)
ıpl.-link.	.42(2)	.61 (2)	.60 (2)	.68 (6)	.64 (2)	.81 (11)	.33 (3)
ɛr.-link.	.17(3)		.55 (3)	.73 (5)	.60 (3)	.95 (4,5)	.27 (4)
ıtroid			.51 (5)	.32 (8)	.50 (5)	.87 (10)	
ıian-Verf.			.52 (4)	.46 (7)	.55 (4)	.88 (9)	
RD-Verf.	.77(1)		.62 (1)	.99 (1)	.84 (1)	.90 (8)	.57 (1)
NCE-WILL.				.98 (3)		.91 (6,5)	
QUITTY				.74 (4)			
Statistik		.76 (1)					
LOCATE				.99 (1)			
QUEEN						.95 (4,5)	
RGY's Meth.						.97 (1,5)	
NCEY's M.						.96 (3)	
ıv.KMEANS						.97 (1,5)	.48 (2)

Legende:

1 = BLASHFIELD, 1976
2 = D'ANDRADE, 1978 (+)
3 = KUYPER & FISHER, 1975 (+)
4 = SCHEIBLER & SCHNEIDER, 1978
5 = MOJENA, 1977
6 = MILLIGAN, 1980 (+)
7 = BLASHFIELD, 1978 (+)

(+ = Jeweils gemittelte Werte aus verschiedenen Untersuchungsbedingungen)

ɔ. 1: Überblick über die in verschiedenen Monte-Carlo-Studien gefundene Rekonstruktionsqualität der einzelnen Clusteranalyse-Verfahren (es werden jeweils Absolutwerte und - in Klammern - die Rangplätze der Verfahren wiedergegeben)

	SCHEIBLER & SCHNEIDER, 1978 (BLASHFIELD, 1976)				BLASHFIELD, 1978	
	a	b	c	d	R^2	Chaining Index
single-link.	-.18 (-.44)	-.20 (-.43)	.04 (.02)	.00 (-.17)	.32	.91
compl.-link.	-.06 (-.01)	-.03 (-.16)	-.04 (-.03)	.22 (-.49)	.29	.11
aver.-link.	-.07 (.09)	-.08 (-.21)	.08 (.21)	.16 (-.45)	.34	.52
Centroid	-.12	-.20	.05	-.05		
Median	-.14	-.23	.03	-.21		
WARD-Verf.	-.17 (-.11)	.02 (.02)	-.05 (-.25)	.29 (.47)	.62	.01
LANCE-WILL.	.15	.06	-.07	.32		
McQUITTY	-.09	-.05	-.01	.17		
RELOCATE	-.18	.01	-.06	.33		
Conv. KMEANS					.36	.02

Legende:

1) SCHEIBLER & SCHNEIDER bzw. BLASHFIELD, 1976:
Rang-Korrelationskoeffizienten zwischen Kappa und
a) Anzahl der Elemente in der Gesamtstichprobe;
b) Mittlerer Element-Anzahl in den Teilstichprobe
c) Varianz der Elementanzahl in den Teilstichprob
d) Mittelwerte der Varianzen der Einzelvariablen

2) BLASHFIELD, 1978:
R^2 = Maß für den Linearzusammenhang zwischen CRAM V und der Schwierigkeit der Mixturen
Chaining Index = Proportion der Clusterlösungen, denen wenigstens 1 Cluster weniger als 5 Ele. mente enthält

Tab. 2: Abhängigkeit der Algorithmen-Leistung von Besonderheiten der Datenstruktur

Anzahl von Subpopulationen	Anzahl von Datensätzen
2	40
3	40
4	40
5	40
6	40
	N = 200
Anzahl von Variablen	
3 - 10	29
11 - 15	29
16 - 20	58
21 - 25	64
	N = 200

Tab. 3: Verteilung der Zahl von Subpopulationen und Variablen bei den analysierten 200 Datensätzen
(SCHEIBLER & SCHNEIDER, 1981)

	kleinster	(theoretisch) größter Wert
Anzahl der Subpopulationen (k) (gleichverteilt)	2	6
Stichprobengröße pro Subpopulation (gleichverteilt)	5	50
Stichprobengröße pro Datensatz (ungefähr gleichverteilt)	10	300
Anzahl der Variablen (p) (Verteilung nach Tab. 3)	3	25
Anzahl der Faktoren (Hauptkomponenten)	2	10
Mittelwerte (Erwartungswerte) (gleichverteilt)	45.	60.
Varianzen (gleichverteilt)	5.	30.
Korrelationskoeffizienten r	-1.	+1.
arccos (r) (gleichverteilt)	0	2

Tab. 4: Verteilung der Populationsparameter in der Untersuchung von SCHEIBLER & SCHNEIDER (1981)

Mediane der Kappa-Werte auf verschiedenen Hierarchieebenen auf der Basis von 200 Datensätzen

Anzahl der Cluster	single linkage	complete linkage	average linkage	Median-Methode	Centroid-Methode	WARD's-Methode	LANCE-WILLIAMS'-Methode beta = -1.	beta = -.5	beta = -.25	beta = 0 Mc QUITY's Methode	RELOCATE *	KMEANS *
K+9	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	.98
K+8	.89	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	.95
K+7	.86	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	.95
K+6	.82	.98	1.	1.	1.	1.	1.	1.	1.	1.	1.	.92
K+5	.79	.94	.98	1.	1.	1.	1.	1.	1.	1.	1.	.91
K+4	.75	.89	.97	.87	.90	1.	.97	1.	1.	.97	.97	.86
K+3	.70	.82	.95	.82	.86	1.	.95	.98	.98	.95	.96	.85
K+2	.59	.69	.90	.75	.80	.96	.90	.93	.94	.90	.91	.77
K+1	.50	.57	.77	.60	.65	.90	.76	.87	.86	.75	.80	.69
K	.04	.38	.16	.04	.05	.791	.43	.77	.72	.43	.58	.55

K = Anzahl der Teilstichproben

RELOCATE**	.775
KMEANS**	.773

* Anfangspartitionierung durch Zufallsaufteilung der Stichprobe

** Anfangspartitionierung entspricht der WARD-Lösung bei K Clustern

A N H A N G

Übersicht und nähere Erläuterungen zu den verschiedenen Cluster-Analyse-Verfahren

Nach EVERITT (1974, S. 7 ff.) lassen sich Clusteranalysen ganz bestimmten Kategorien zuordnen, die im folgenden näher charakterisiert werden sollen:

1. Hierarchische Clusteranalysen

Es lassen sich bei diesen Techniken agglomerative und divisive Formen unterscheiden. Verfahren, die nach einer Serie von sukzessiven Fussionen der N Elemente (Personen) schließlich zu einer Endstufe gelangen, in der alle Elemente einer einzigen Klasse angehören, werden agglomerativ genannt, während Methoden, die genau umgekehrt vorgehen (nämlich die Gesamtheit der Elemente schrittweise in immer kleinere Klassen von ähnlichen Elementen unterteilen), als divisiv bezeichnet werden.

2. Optimierungs-Partitionierungs-Techniken

Es werden disjunkte (nicht überlappende) Cluster dadurch gebildet, daß ein festgelegtes Kriterium optimiert wird. Die wesentlichen Unterschiede zu den hierarchischen Verfahren bestehen darin, daß zum einen schlechte Anfangspartitionierungen korrigiert, die Elemente demnach re-allokiert werden können, und zum anderen die gewünschte Klassenzahl vom Untersucher a priori festzulegen ist.

3. Dichteverfahren (mode seeking techniques)

Die Elemente werden als Punkte im metrischen Raum aufgefaßt und solche Regionen als Cluster interpretiert, in denen sich die Elemente besonders dicht konzentrieren.

4. 'Clumping Techniques'

Im Unterschied zu den vorher erwähnten Verfahren, die lediglich disjunkte Klassen zulassen, können mit dieser Methode überlappende Cluster (Clumpen) gebildet werden.

5. Andere Verfahren

Hier sind Methoden gemeint, die keiner der vier anderen Kategorien eindeutig zugeordnet werden können.

Kurzbeschreibung der im Manuskript erwähnten Clusteranalyse-Techniken

1. Hierarchisch agglomerative Verfahren

Alle nachfolgend aufgeführten hierarchischen Methoden verfahren nach dem selben Prinzip und unterscheiden sich allein in der Definition der Distanz zwischen zwei Clustern. Alle Verfahren beginnen damit, daß sie zunächst jene beiden Elemente zu einem Cluster verbinden, die am nächsten zu einander liegen. Im folgenden werden schrittweise immer jene zwei Cluster zu einem neuen Cluster zusammengefaßt (agglomeriert), die die geringste Distanz zueinander aufweisen (wobei auch Einzelelemente als Cluster definiert sind). Mit jedem Schritt reduziert sich also die Zahl der Cluster um 1 und es entstehen immer größere aber auch heterogenere Cluster, bis alle Elemente in einem großen Cluster verbunden sind.

Die Definitionen der Ähnlichkeit zwischen Clustern lassen sich für alle unten aufgeführten Verfahren auf eine Rekursionsformel zurückführen:

Wenn die Cluster P und Q zu einem neuen Cluster verbunden werden, dann ist die Ähnlichkeit S(R,P+Q) zwischen dem Cluster R und dem neuen Cluster (P+Q):

$$S(R,P+Q)=AP*S(R,P)+AQ*S(R,Q)+B*S(P,Q)+G*ABS(S(R,P)-S(R,Q)$$

a) Single-linkage-Methode

Bei dieser Technik ist die Distanz zwischen zwei Clustern durch den Abstand jener zwei Elemente aus den beiden Clustern definiert, die die geringste Distanz zueinander aufweisen.

Die Parameter für die Rekursionsformel haben folgende Werte:

AP = AQ = .5; B = 0; G = .5 (bei Ähnlichkeitskoeffizienten)

G = -.5 (bei Distanzkoeffizienten)

b) Complete-linkage-Methode

Bei diesem Ansatz ist die Distanz zwischen zwei Clustern als der Abstand zwischen den am weitesten auseinanderliegenden Elementen aus den jeweiligen Clustern definiert.

Die Parameter der Rekursionsformel haben folgende Werte:

AP = AQ = .5; B = 0; G = -.5 (bei Ähnlichkeitskoeffizienten)

G = .5 (bei Distanzkoeffizienten)

c) Average-linkage-Methode

Diese Methode stellt einen Kompromiß zwischen den beiden vorher beschriebenen Ansätzen dar: Die Distanz zwischen zwei Clustern ist definiert als arithmetisches Mittel der Distanzen zwischen allen Paaren von Elementen, die man aus zwei Elementen jeweils verschiedener Clusterzugehörigkeit bilden kann.

Die Parameter der Rekursionsformel haben hier die Werte:

AP = NP/(NP * NQ); AQ = NQ/NP + NQ; B = G = 0

(ist nur für Distanzkoeffizienten definiert).

d) Centroid-Methode

Das Prinzip dieser Methode läßt sich am leichtesten geometrisch erklären, wenn man sich die Cluster als Punkte-Haufen oder -Wolken in einem euclidischen Raum vorstellt. Die Distanz zwischen zwei Clustern wäre dabei als die Distanz zwischen den Centroiden dieser Cluster definiert.

e) Median-Methode (nach GOWER)

Die Distanz d (R,P+Q) zwischen irgendeinem Cluster R und dem Cluster das durch die Fusion von P und Q gebildet wird, ist bei diesem Verfahren als die Distanz vom Centroid von R zum Mittelpunkt der Strecke definiert, die die Centroide von P und Q verbindet.

Das Verfahren weist große Ähnlichkeit mit der Centroid-Methode auf, unterscheidet sich jedoch in einem wesentlichen Punkt: Die Centroid-Methode hat den Nachteil, daß bei der Fusion zweier extrem unterschiedlich großer Cluster das Centroid der neuen Gruppe sehr nahe an der des größeren Clusters liegt und die charakteristischen Eigenschaften der kleinen Gruppe damit verloren gehen. Dieses Problem besteht bei der Median-Methode nicht, da das Verfahren unabhängig von der Clustergröße arbeitet.

f) Clusteranalyse nach WARD

Auch bei diesem Verfahren werden bei jedem Agglomerationsschritt jene zwei Cluster zu einem neuen verbunden, die sich am ähnlichsten sind. Die zwei Cluster sind am ähnlichsten definiert, deren Vereinigungsmenge die kleinste Varianz von allen bildbaren Clusterverbindungen aufweist.

g) LANCE & WILLIAMS' flexible beta-Methode

Dieses Verfahren kann nicht mehr geometrisch veranschaulicht werden. Es baut auf der oben erwähnten Rekursionsformel auf, wobei ein Teil der Parameter a priori festgesetzt ist während einer (das sogenannte beta) beliebig variiert werden kann. LANCE & WILLIAMS schlugen vor, den Wert von beta auf -0.25 festzusetzen (unsere eigenen Untersuchungen lassen den Wert von -.5 als günstiger erscheinen). In diesem Fall verhält sich das Verfahren im wesentlichen wie die WARD'sche Technik.

Die Parameter der Rekursionsformel für dieses Verfahren lauten:

AP = AQ = (1 - beta)/2; B = beta G = 0

h) Die Methode nach McQUITTY

Dieses Verfahren entspricht dem von LANCE & WILLIAMS, wenn beta = 0 gesetzt wird.

2. Optimierungs-Partitionierungs-Techniken

a) Optimierungstechnik nach MacQUEEN

Bei dieser Partitionierungs-Prozedur wird versucht, die Spur einer Matrix W (Summe der Elemente in der Hauptdiagonalen) zu minimieren, wobei W die gepoolte Kovarianz-Matrix der Abweichungsquadrate und Kreuzprodukte innerhalb der Gruppen darstellt. Die Logik des Vorgehens ist evident, da im Prinzip die Summe der Abweichungsquadrate minimiert wird, was zu möglichst homogenen Gruppen führt. Im Prinzip wird also versucht, die Varianz innerhalb der Cluster zu minimieren. Die Anzahl der Cluster, die gebildet werden sollen, muß ad hoc vorgegeben werden. Außerdem benötigt das Verfahren eine vorläufige Clusterlösung als Startkonfiguration. Danach wird diese Anfangspartitionierung nach folgendem heuristischen Prinzip zu verbessern versucht: Für alle Dateneinheiten (Elemente) werden nacheinander die Distanzen zu allen Cluster-Centroiden berechnet. Wenn der am nächsten gelegene Centroid nicht zu dem Cluster gehört, in dem das betreffende Element liegt, wird es dem anderen (näher gelegenen) Cluster zugeordnet (relociert). Danach werden die Centroide des Clusters, das ein Element verloren hat sowie dessen, das eines dazu bekommen hat neu berechnet, und der Vorgang wird für die weiteren Elemente fortgesetzt. Danach beginnt das Verfahren wieder mit dem ersten Element und setzt die Prozedur so lange fort, bis keine Elemente mehr ausgetauscht werden müssen. Dieses Verfahren ist u.a. in dem Programm MIKCA von McRAE (1971) realisiert.

b) Das Programm RELOCATE von WISHARD (1975)

RELOCATE enthält als Kernstück den Algorithmus von McQUEEN. Es bietet die Möglichkeit, eine Anfangspartitionierung vom Programm generieren zu lassen (Zufallseinteilung) oder die Lösung von einer vorgeschalteten anderen Clusterprozedur zu übernehmen. Wenn RELOCATE eine verbesserte Lösung gefunden hat, besteht die Möglichkeit die Zahl der Cluster ähnlich wie beim WARD'schen Verfahren durch Agglomeration schrittweise zu reduzieren, wobei nach jeder Reduktion die Otimierungsstrategie nach McQUEEN abläuft.

c) Optimierungstechnik nach SPAETH (1975)
Das Programm KMEANS

Dieses Programm baut ebenfalls auf dem Ansatz von McQUEEN auf, berechnet aber nicht die Distanzen der Elemente zu den Centroiden der Cluster, sondern zu deren Medianen. SPAETH empfiehlt, der Prozedur eine Standardisierung der Daten vorzuschalten, und zwar der Gestalt, daß alle Variablen so (linear) transformiert werden, daß sie das gleiche range bzw. gleiche minima und maxima aufweisen.

d) Die Prozedur EUCLID von WISHARD (1975)

Das hier angesprochene Verfahren verwendet die CAUCHY-Methode, um die Quadratsumme S der euclidischen Distanz durch schrittweise Verbesserung einer kontinuierlichen Klassifikationsmatrix Y zu minimieren. Das Element Y_{ik} kann dabei als Wahrscheinlichkeit interpretiert werden, daß der i-te Punkt zum k-ten Cluster gehört.

EUCLID generiert für eine vorgegebene Zahl von Clustern gleichverteilte Zufallswerte für Y, die so skaliert sind, daß die Summe der "Wahrscheinlichkeiten" (fractions) = 1 ist. Durch eine Expotential-Transformation von Y ergeben sich neue Klassifikationsvariablen (kontinuierlich-differenzierbar und monoton-ansteigend).

Eine Folge von Klassifikationsmatrizen wird nun durch sukzessive Applikation der CAUCHY-Korrekturen so generiert, daß bei jedem Schritt ein neues Minimum von S gefunden wird. Die Sequenz wird bei derjenigen Iteration beendet, die für jeden Punkt i einen Wert k so kombiniert, daß Y_{ik} wenigstens .999 ergibt und die partielle Ableitung von S im Hinblick auf Y_{ik} negativ wird.

Die Ergebnisklassifikation wird dann so zu zerlegen versucht, daß durch Re-Allocation von einigen Punkten eine weitere Reduktion von S möglich wird (Verwendung des Prinzips von RELOCATE).

3. Dichte-Verfahren (mode-seeking techniques)

(a) Mode analysis nach WISHARD:

Diese Technik läßt sich der Gruppe der Dichteverfahren zuordnen. Der Algorithmus sieht vor, daß zunächst um jedes Element je ein Raum mit dem Radius R aufgespannt und untersucht wird, wie viele benachbarte Elemente ebenfalls in dieser Sphäre liegen. Wenn der Raum bei bestimmten Individuen K und mehr Elemente enthält, werden diese zu Dichtezentren erklärt. Der Parameter R wird sukzessive vergrößert, wobei mit neuen Dichtezentren verschiedenartig verfahren werden kann: wird das neue Zentrum von allen anderen Dichtepunkten durch eine Distanz getrennt, die größer als R ist, muß ein neues Cluster initiiert werden; ist die Distanz dagegen kleiner als R, sind mehrere Möglichkeiten offen:

1) wenn sich der neue Punkt im Bereich von Dichtezentren befindet, die einen gemeinsamen Clusterkern besitzen, wird er diesem angeschlossen;

2) liegt er im Bereich von Dichtezentren, die unterschiedlichen Clusterkernen angehören, werden die betreffenden Cluster zusammengeschlossen.

Weiterhin wird bei jedem Zyklus die geringste Distanz D zwischen Dichtezentren berechnet, die zu verschiedenen Clustern gehören, und letztere mit einem bestimmten Grenzwert verglichen: falls D diesen Grenzwert nicht übersteigt, werden die betreffenden Cluster kombiniert.

(b) Das Verfahren der Automatischen Klassifikation (AUKLA) nach FABER & NOLLAU:

AUKLA versucht im vollständigen oder reduzierten Datenraum (aufgespannt durch eine begrenzte Zahl von Faktoren) "geometrisch vorgegebene Punktehäufungen innerhalb dieses Datenraums festzustellen und die jeweils zusammengehörigen Elemente zu ermitteln" (FABER & NOLLAU 1969, S. 8).
Die Grundidee des Verfahrens geht auf SCHNELL zurück. Dieser betrachtet die p Meßwerte einer Person (p = Anzahl der Variablen) als Erwartungswerte einer p-dimensionalen Normalverteilung (zum besseren Verständnis sei auf die analoge Betrachtungsweise

in der klassischen Testtheorie hingewiesen, bei der man annimmt, daß aufgrund eines Meßfehlers der wahre Wert mit einer bestimmten Wahrscheinlichkeit - definiert durch eine Normalverteilung mit dem Meßwert als Mittelwert - in einem gewissen Intervall um den gemessenen Wert liegen kann).

Die Verteilungen der Meßwerte mehrerer Personen überlagern sich mehr oder weniger stark, wobei das jeweilige Ausmaß davon bestimmt wird, wie dicht die Werte im Datenraum beieinanderliegen und wie groß die Varianzen dieser Verteilungen sind. Ist die Varianz bei allen Personen gleich null, können sich die Verteilungen verschiedener Personen nicht überlagern (es sei denn, sie hätten identische Meßwerte in allen Variablen).

Erhöht man nun die Varianz allmählich (und für alle Personen gleichmäßig), dann werden zunächst nur solche Punkte im Datenraum (Personen) deutliche Überlagerungen ihrer Verteilungen zeigen, die dicht beieinander liegen.
Das Verfahren AUKLA berechnet eine sog. Belegfunktion, die neben einem Haupt- meist mehrere Nebenmaxima (Gipfel) aufweist. Diese Maxima deuten auf stärkere Überlagerungen von Verteilungen und damit auf Punktkonzentrationen im Datenraum hin. Sie können als Clusterzentren interpretiert werden. Je größer nun die Varianz der Verteilungen ist, desto eher kommt es auch zu ausgeprägten Überlagerungen von Verteilungen um Punkte, die weiter voneinander entfernt sind. Eine stetige Vergrößerung der Varianz führt damit auch gleichzeitig zu einer Verringerung der Anzahl von Maxima in der Belegfunktion, die (in der Endstufe) auf N = 1 reduziert wird. Wenn nun trotz stärkerer Veränderungen der Varianz die Anzahl der Maxima lange Zeit konstant bleibt, deutet dies darauf hin, daß eine (relativ) optimale Anzahl von Clustern gefunden ist, von der man bei der Ergebnisinterpretation ausgehen sollte.

4. 'Clumping techniques'

(a) Das Taxonome-Programm (TAXO) von CATTEL & COUTLER:

Obwohl dieses Verfahren nicht exakt unter Kategorie 4 subsummiert werden kann, wird es hier dennoch dargestellt, weil es als einzige der in diesem Kontext berücksichtigten Prozeduren überlappende Cluster zuläßt.
Die konzeptuelle Grundlage ist in dem Typus-Begriff von CATTELL zu sehen: Als 'Typen' werden relative Maxima in einer mehrdimensionalen Häufigkeitsverteilung definiert. Im ersten Schritt werden aus einer Korrelationsmatrix sich überlappende sog. 'Phenomenal clusters' (PhCl) als Grundeinheiten extrahiert, in denen Elemente gesammelt sind, die sich oberhalb eines beliebig zu wählenden Cutoff-Werts (Grenzwerts) ähneln. Im zweiten Schritt lassen sich die Überlappungsbereiche zweier oder mehrer PhCl als sog. 'nuclear cluster' bestimmen, die wiederum die Voraussetzung für die im dritten Schritt erzeugten 'Segregates' (Verkettungen einzelner oder mehrerer PhCl) bilden, in denen nicht unbedingt alle Elemente einander ähnlich sein müssen. Eine Optimierung der Gruppen-Extraktion kann durch systematische Manipulation des Cut-Off-Werts sowie des Parameters der Überlappungsgröße erreicht werden (genauere Vorschläge zur Optimierung bei BAUMANN 1971, S. 115 f. und 157 f.).

5. Sonstige Verfahren

(a) Die Konfigurationsfrequenzanalyse (KFA):

Das Verfahren analysiert eine endliche Zahl von Variablenmustern (Konfigurationen) für kategoriale Daten (bei n dichotomen Variablen sind beispielsweise 2^n Konfigurationen möglich). Ein 'Typus' läßt sich als signifikant überfrequentierte Konfiguration charakterisieren, ein 'Anti-Typus' dementsprechend als signifikant unterfrequentierte Konfiguration bezeichnen; Personen, die keinem Typus (bzw. Antitypus) zugehören, werden nicht erfaßt. Vor der genaueren Analyse der einzelnen Typen wird jedoch jeweils überprüft, inwieweit die beobachteten von den theoretisch zu erwarteten Konfigurations-Frequenzen abweichen; falls die Nullhypothese (es besteht kein signifikanter Unterschied) nicht widerlegt werden kann, unterbleibt eine weitere Analyse.

(b) Die Clusteranalyse nach LORR et al.:

Der Clustersuchprozeß beginnt bei diesem Verfahren mit einem Auflisten der Code-Nummern (Vpn-Nummern) aller Standardwert-Profile, die mit einem bestimmten Profil X oberhalb eines festgelegten Grenzwertes (Signifikanz-Schranke) korrelieren (gilt für alle X_n). Das Profil mit der längsten Nummernliste eröffnet ein Cluster: es werden sukzessive diejenigen Profile zugeführt, deren durchschnittliche Korrelation mit den Profilen innerhalb des Clusters am größten ist. Ein weiterer Grenzwert gibt an, wann ein Profil (trotz möglicherweise immer noch beträchtlicher Korrelation) als nicht mehr ähnlich gewertet und eliminiert wird, wodurch die Möglichkeit besteht, deutlich separierbare Klassen zu konstruieren. Aus der Residualmatrix wird nun das zweite Cluster initiiert und der Klassifikationsprozeß solange weitergeführt, bis keine Klassen mit wenigstens vier Items mehr gebildet werden können.

(c) Mehrdimensionale automatische Clustersuchstrategie nach von EYE:

Bei diesem Verfahren wird a priori die Homogenität der Klassen dadurch festgelegt, daß alle Cluster im Testraum als gleich große d-dimensionale Wuader (d = Anzahl der Meßwertkategorien) darstellbar sind.
In einem ersten Schritt wird das Konfidenzintervall um das zentrale Element (bei intervallskalierten Daten) bzw. um das Streuungsmaß (bei niedriger skalierten Daten) von allen d Meßwertkategorien errechnet. Danach wird dasjenige Element entweder vorgegeben (bei Hypothesen über die Datenstruktur) oder aber stochastisch aufgesucht, das sich am besten zur Eröffnung eines Clusters eignet, wobei sich dieser Prozeß der sukzessiven Zuordnung solange festsetzt, bis entweder das Maximum der vorgegebenen Klassenzahl erreicht ist oder aber alle Elemente in einem Cluster zusammengefaßt sind. Um festzustellen, ob die entstandenen Gruppierungen stabil bleiben. wird die Prozedur schließlich mit variierenden Quadergrößen wiederholt. Die Kantenlängen bestimmen die Homogenität der Cluster und können in Analogie zu den Hierarchie-Ebenen der hierarchischen Clusteranalysen interpretiert werden.

Tab. 5: Ergebnisse der Monte-Carlo-Studie von SCHEIBLER & SCHNEIDER (1981): Mediane der Kappa-Werte auf verschiedenen Hierarchieebenen auf der Basis von 200 Datensätzen

Anzahl der Cluster	single linkage	complete linkage	average linkage	Medianmethode	Centroidmethode	WARD's Methode	LANCE-WILLIAMS'-Methode: beta = -1.	beta = -.5	beta = -.25	beta = 0 nach QUINN's Methode	RELOCATE*	KMEANS*
k + 9	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	.98
k + 8	.89	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	.95
k + 7	.86	1.	1.	1.	1.	1.	1.	1.	1.	1.	1.	.95
k + 6	.82	.98	1.	1.	1.	1.	1.	1.	1.	1.	1.	.92
k + 5	.79	.94	.98	1.	1.	1.	1.	1.	1.	1.	1.	.91
k + 4	.75	.89	.97	.87	.90	1.	.97	1.	1.	.97	.97	.86
k + 3	.70	.82	.95	.82	.86	1.	.95	.98	.98	.95	.96	.85
k + 2	.59	.69	.90	.75	.80	.96	.90	.93	.94	.90	.91	.77
k + 1	.50	.57	.77	.60	.65	.90	.76	.87	.86	.75	.80	.69
k	.04	.38	.16	.04	.05	.791	.43	.77	.72	.43	.58	.55

RELOCATE**	.775
KMEANS**	.773

k = Anzahl der Teilstichproben

* Anfangspartitionierung durch Zufallsaufteilung der Stichprobe

** Anfangspartitionierung entspricht der WARD-Lösung bei k Clustern

Bisher erschienene Berichte aus dem

Psychologischen Institut der Universität Heidelberg

Diskussionspapier Nr. 1: GROEBEN, N.: Vom behavioralen zum epistemologischen Subjektmodell: Paradigmawechsel in der Psychologie? September 1975

Diskussionspapier Nr. 2: MÖBUS, C. & SIMONS, H.: Zur Fairness psychologischer Intelligenztests gegenüber ethnischen und sozialen Gruppen: Kritik klassischer Konzepte. Oktober 1975

Diskussionspapier Nr. 3: WOTTAWA, H.: Skalenprobleme bei probabilistischen Meßmodellen. März 1976

Diskussionspapier Nr. 4: TREIBER, B. & PETERMANN, F.: Zur Interaktion von Lernermerkmalen und Lehrmethoden: Rekonstruktion und Normierung des ATI-Forschungsprogramms. April 1976

Diskussionspapier Nr. 5: MÖBUS, C. & WALLASCH, R.: Zur Erfassung von Hirnschädigungen bei Kindern: Nichtlineare Entscheidungsregeln auf der Basis von Veränderungsmessungen. August 1976

Diskussionspapier Nr. 6: SCHEELE, B. & GROEBEN, N.: Voraussetzungs- und zielspezifische Anwendung von Konditionierungs- vs. kognitiven Lerntheorien in der klinischen Praxis. Dezember 1976

Diskussionspapier Nr. 7: MÖBUS, C.: Zur Analyse nichtsymmetrischer Ähnlichkeitsurteile: Ein dimensionales Driftmodell, eine Vergleichshypothese, TVERSKY's Kontrastmodell und seine Fokushypothese. Juni 1977

Diskussionspapier Nr. 8: SIMONS, H. & MÖBUS, C.: Veränderung von Berufschancen durch Intelligenztraining. Juli 1977

Diskussionspapier Nr. 9: BRAUNMÜHL, C. v. & GRIMM, H.: Zur Kommunikationspsychologie: Über Versuche der methodischen Konstitution eines genuin humanwissenschaftlichen Forschungsansatzes zur Entwicklung der Verständigungsfähigkeit. November 1977

Diskussionspapier Nr. 10: HOFER, M.: Entwurf einer Heuristik für eine theoretisch geleitete Lehrer- und Erzieherbildung. November 1977

Diskussionspapier Nr. 11: SCHEIBLER, D. & SCHNEIDER, W.: Probleme und Ergebnisse bei der Evaluation von Clusteranalyse-Verfahren. Juni 1978

Diskussionspapier Nr. 12: SCHEELE, B.: Kognitions- und sprachpsychologische Aspekte der Arzt-Patient-Kommunikation. September 1978

Diskussionspapier Nr. 13: TREIBER, B. & SCHNEIDER, W.: Mehrebenenanalyse sozialstruktureller Bedingungen schulischen Lernens. Oktober 1978

Diskussionspapier Nr. 14: AHRENS, H.-J. & KORDY, H.: Möglichkeiten und Grenzen der theoretischen Aussagekraft von multidimensionalen Skalierungen bei der Untersuchung menschlicher Informationsverarbeitung. Teil I: Formale und wissenschaftstheoretische Grundlagen. Februar 1979

Diskussionspapier Nr. 15: GROEBEN, N.: Entwurf eines Utopieprinzips zur Generierung Psychologischer Konstrukte. Juni 1979

Diskussionspapier Nr. 16: WEINERT, F.E. & TREIBER, B.: Socialization and school influences on cognitive development. Juni 1979

Diskussionspapier Nr. 17: GUNDLACH H.: Inventarium der älteren Experimentalapparate im Psychologischen Institut Heidelberg sowie einige historische Bemerkungen. 1978

Diskussionspapier Nr. 18: SCHEELE, B. & GROEBEN, N.: Zur Rekonstruktion von subjektiven Theorien mittlerer Reichweite. Eine Methodik-Kombination von halbstandardisiertem Interview (einschließlich Konfrontationstechnik) und Dialog-Konsens über die Theorie-Rekonstruktion mittels der Struktur-Lege-Technik (SLT). Dezember 1979

Diskussionspapier Nr. 19: GLOGER-TIPPELT, G.: Subjektive Theorie von Frauen über ihre erste Schwangerschaft: Theoretische Konzepte und methodische Möglichkeiten. Januar 1980

Diskussionspapier Nr. 20: KAMMERER, A.: Das Konzept 'psychotherapeutische Strategie' am Beispiel des Problemlösens. Juli 1980

Diskussionspapier Nr. 21: SCHEELE, B.: (unter Mitarbeit von B.Tuschen und C.Maier): Subjektive Theorien über Ironie - als Heuristik für einen wissenschaftlichen Hypothesenkorpus. August 1980

Diskussionspapier Nr. 22: TREIBER, B.: Erklärung von Förderungseffekten in Schulklassen durch Merkmale subjektiver Unterrichtstheorien ihrer Lehrer. Oktober 1980

Diskussionspapier Nr. 23: RÖHRLE, B. & KOMMER, D.: Handlungstheoretische Betrachtungen zur primären Prävention. März 1981

Diskussionspapier Nr. 24: Die Entwicklung des Zahlbegriffs, Bde. 1 und 2. April 1981 (VOIGT, F.)

Diskussionspapier Nr. 25: SCHNEIDER, G. & WEIMER, E.: Aspekte der Kategorisierung städtischer Umwelt - Eine empirische Untersuchung. Juni 1981