Psychologische Beiträge, Band 25, 1983, S. 208–237

Probleme und Möglichkeiten bei der Bewertung von Clusteranalyse-Verfahren I. Ein Überblick über einschlägige Evaluationsstudien

W. SCHNEIDER[1] und D. SCHEIBLER[2]

Zusammenfassung, Summary, Résumé

Es wird ein Überblick über Evaluationsstudien gegeben, die sich mit der Validität von Clusteranalyse-Algorithmen befassen. Im Anschluß an die Diskussion möglicher Bewertungskriterien werden Vergleichsuntersuchungen näher analysiert und danach geordnet, ob sie empirische Datensätze, Plasmoden oder Monte-Carlo-Datensätze als Evaluationsgrundlage benutzen. Die Übersicht über komplexer angelegte Monte-Carlo-Studien zeigt die unterschiedliche Qualität der verfügbaren Clusteranalyse-Algorithmen auf, macht andererseits aber auch deutlich, daß bestimmte hierarchisch-agglomerative Verfahren wie etwa die Methoden nach WARD oder LANCE-WILLIAMS bzw. iterativ-partitionierende Prozeduren wie etwa die KMEANS-Algorithmen als relativ robuste Klassifikationsverfahren gelten können.

On the evaluation of clustering algorithms: An integrative review

This paper presents a critical review of research on the evaluation of clustering algorithms. The review includes studies using empirical data sets and studies using so-called „plasmodes" (i. e., empirical data sets with known distributional parameters), but particularly concentrates on investigations using Monte-carlo data sets. Although it turns out to be very difficult to come to a valid evaluation of the various clustering algorithms, hierarchical-agglomerative procedures like WARDs and LANCE-WILLIAMs methods as well as the KMEANS algorithms appear to be most robust.

Procédés de Cluster-analyse

L'article suivant présente une revue des études d'évaluation qui relèvent de validité d'analyse de Cluster-Algorithme. Des expériences de comparaison sont analysées et classées à la suite de la discution de critères de jugement possible. Ceux-ci suivent un ordre donné par l'utilisation de données qui ont été relevées empiriquement, «Plasmoden» ou bien de données empiriquement accumulées selon le principe d'évaluation de Monte-Carlo.

La revue d'étude complexe Monte-Carlo montre la difference qualitative des analyses Cluster algorithme et, d'autre part met clairement en valeur, que des procédés

1 Dr. Wolfgang Schneider, Max-Planck-Institut für psychologische Forschung, Leopoldstr. 24, 8000 München 40.
2 Dipl.-Psych. Dieter Scheibler, Neugasse 7, 6900 Heidelberg.

hierarchiques-agglomeratifs, comme la methode de WARD ou LANCE-WILLIAMS, par exemple – c'est-à-dire des procédés «iteratif-partitionierende» (iteratif comme par exemple, les algorithmes KMEANS, qui peuvent être considéres comme procédés de classification robuste. (Dr. Lohr)

1. Einleitung

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

Wenn im folgenden der Versuch unternommen wird, die Leistungsfähigkeit unterschiedlicher statistischer Klassifikationsverfahren zu bewerten, steht dabei insbesondere ein pragmatischer Aspekt im Vordergrund: der Methodenvergleich soll dem (gelegentlichen) Anwender von Clusteranalysen 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, läßt sich an der besonderen historischen Entwicklung demonstrieren. Wurden die ersten Clusteranalysen noch als ‚Faktorenanalysen des kleinen Mannes' belächelt und wegen des erforderlichen Rechenaufwandes zunächst kaum beachtet, so führte die enorme Weiterentwicklung von Computersystemen in den letzten fünfzehn Jahren zu einer wahren ‚Explosion', was das Interesse an Klassifikationsverfahren anging. Nachdem innerhalb kürzester Zeit mehr als 200 verschiedene Clusteranalyse-Algorithmen verfügbar waren, stieg die Zahl der pro Jahr in wissenschaftlichen Journalen publizierten Untersuchungen zu taxonometrischen Problemen beträchtlich an (so registrierte etwa CORMACK für das Jahr 1971 mehr als 1 000 Veröffentlichungen zu Clusteranalyse-Problemen).

Der Umstand, daß in den unterschiedlichen Anwendungsgebieten (z. B. in der Biologie, den Sozialwissenschaften, der Informatik, Anthropologie, Archäologie und Linguistik) nur die Literatur des nächsten Umfelds rezipiert wurde (meist durch die Dominanz von ‚Schulen' geprägt (s. VOGEL 1975)), förderte gleichzeitig die Ausprägung unterschiedlicher Terminologien (‚jargon') und damit den Ausbau interdisziplinärer Kommunikationsbarrieren (illustrative Beispiele hierzu liefern BLASHFIELD & ALDENDERFER 1978, S. 285). Selbst dem durchaus routinierten Clusteranalysen-Benutzer ist somit häufig nicht bekannt, daß sich mehrere

(manchmal mehr als 7!) semantisch kaum verknüpfbare Etikette auf einen einzigen Algorithmus beziehen[3].

Angesichts der schon erwähnten großen Zahl von verfügbaren Software-Programmen erscheint es ‚weiterhin verwirrend, daß unaufhörlich ‚neue' Algorithmen publiziert werden, ohne daß hier dabei die Notwendigkeit eingesehen wird, Validierungsversuche oder zumindest Vergleiche mit ähnlich strukturierten Programmen zu berichten (Ausnahmen von dieser Regel stellen die neueren Studien von D'ANDRADE (1978) bzw. Von EYE (1977; Von EYE & WIRSING 1978, 1980) dar.

1.2. Kriterien für die Auswahl der zu evaluierenden Algorithmen

Es ist nun in diesem Rahmen weder möglich noch notwendig, die Vielzahl von Einzel-Algorithmen näher darzustellen, da nur vergleichsweise wenige Prozeduren Eingang in populärere Software-Programmpakete gefunden haben und somit häufiger 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 angehören, während bei der divisiven Form genau umgekehrt verfahren wird (die Gesamtheit der Objekte also schrittweise in immer kleinere Klassen ähnlicher Elemente zerlegt wird). In beiden Versionen werden die Ergebnisse auf den unterschiedlichen Klassifikationsebenen in Form von hierarchischen Baumstrukturen (Dendrogrammen) wiedergegeben.

Demgegenüber 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 Clusterhomogenität iterativ (durch Austausch von Einzelelementen) solange zu verbessern versucht, bis ein näher zu definierendes lokales Optimum erreicht worden ist.

Als letzte Gruppe von Klassifikationsverfahren ist vollständigkeitshalber die der sog. *‚clumping'*-Methoden zu erwähnen, die im Unterschied zu den hierarchischen und iterativ partitionierenden Prozeduren überlappende Clusterlösungen zulassen.

Wie detaillierte Literatur-Recherchen (BLASHFIELD 1980; BLASHFIELD & ALDENDERFER 1978) bzw. sorgfältig geplante sog. ‚consumer

3 Eine detaillierte und kurzweilige historische Analyse der Entwicklung des Forschungsprogramms zur Clusteranalyse findet man bei BLASHFIELD (1980).

reports' (ALDENDERFER 1977; BLASHFIELD 1977 a 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 genügt demnach, die aufgrund ihrer enormen Verbreitung in der Biologie und den Sozialwissenschaften geläufigeren 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 berücksichtigen[4].

1.3. Erörterung möglicher Bewertungskriterien

Es bereitet keine sonderliche Mühe, sich verschiedene Gesichtspunkte vorzustellen, nach denen Computerprogramme zur Klassifikation ‚von Objekten miteinander verglichen werden könnten. Neben Ökonomieprinzipien sind so etwa Fragen nach der Flexibilität (z. B. Anzahl der verfügbaren 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 1977 b) ausführlich diskutiert worden sind, soll hier nur kurz auf die wesentlichen Befunde eingegangen werden.

Im Hinblick auf den Rechenzeit- und Kernspeicherbedarf sind die iterativen Partitionierungsverfahren gegenüber den hierarchischen agglomerativen Techniken insofern im Vorteil, als bei ihnen nicht die Speicherung der N X N – Ähnlichkeitsmatrix (= N x (n –1) / Elemente), sondern lediglich die Verfügbarkeit der Rohdaten-Matrix sowie einiger relativ kleiner zusätzlicher Matrizen erforderlich ist. Größere Datenmengen lassen sich mit Partitionierungstechniken demnach also weitaus ökonomischer bewältigen.

Demgegenüber sind die hierarchischen Verfahren größtenteils in verbreiteten Software-Programmpaketen (z. B. CLUSTAN 1C von WISHART 1975) verfügbar, die eine bedeutend größere Zahl von Optionen anbieten (so bei CLUSTAN insgesamt 38 Ähnlichkeits-/Unähnlichkeitsmaße und 8 verschiedenen Cluster-Algorithmen).

4 Da eine ausführliche Beschreibung der genannten Verfahren hier aus Platzgründen nicht erfolgen kann, sei der interessierte Leser auf die Einführungen von ANDERBERG, BLASHFIELD & ALDENDERFER, ECKES & ROSSBACH oder VOGEL verwiesen. Ein Appendix mit einer Kurzbeschreibung der wichtigsten Algorithmen wird in einem der nächsten Hefte dieser Zeitschrift veröffentlicht (SCHNEIDER & SCHEIBLER 1983 b).

Wählt 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 (1977 b) 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 überspitzt formuliert ist, kann aus ihr andererseits doch der Schluß gezogen werden, daß aufwendigere Evaluationsstudien zu den oben aufgeführten Vergleichsaspekten kaum lohnen.

Besondere Aufmerksamkeit wird im folgenden dem theoretisch eigentlich fundamentalen Evaluationsgesichtspunkt geschenkt, der zentral mit der Frage verknüpft ist, wie gut unterschiedliche Cluster-Algorithmen bei der Lösung vorgegebener Klassifikationsaufgaben abschneiden. Das Interesse an einer solchen Problemstellung wurde etwa ab 1965 immer stärker erkennbar, wobei die Dringlichkeit von Vergleichsstudien zum einen mit der stetig wachsenden Zahl von verfügbaren Clusteranalyse-Techniken, zum anderen mit dem Unmut darüber begründet wurde, daß in vielen empirischen Untersuchungen die Wahl des jeweiligen CA-Auswertungsinstruments – wenn überhaupt – immer mehr anhand von wenig überzeugenden Argumenten (zufällige lokale Verfügbarkeit, intuitive Annahme) vorgenommen wurde.

Die zur Behebung dieses Wissensdefizits in der Folge durchgeführten Evaluationsstudien lassen sich am einfachsten nach der Art der zum Vergleich verwendeten Datensätze differenzieren: so können (in der chronologisch korrekten Sequenz) Evaluationen anhand von empirischen Datensätzen, anhand von Plasmoden sowie über Monte-Carlo-Methoden unterschieden werden. Ähnlich ging etwa auch MILLIGAN (1981) vor, indem er verschiedene ‚Epochen' der Evaluationsforschung unterschied. Es bleibt allerdings anzumerken, daß hier (wie in den meisten amerikanischen Studien) nicht zwischen Plasmoden- und Monte-Carlo-Methoden unterschieden wurde. Eine ausführlichere Darstellung der wesentlichen Befunde scheint insofern angebracht, als die meisten Studien in schwerer zugänglichen und speziell für Sozialwissenschaftler eher abgelegenen angloamerikanischen Zeitschriften publiziert worden sind.

2. Überblick über die vorliegenden Evaluationsstudien

Als paradigmatisch für den Aufbau der ersten Evaluationsansätze kann die Studie von SNEATH (1966 A) gelten: die drei herangezogenen hierarchisch-agglomerativen Verfahren (single-, complete- und average-link) wurden im Hinblick auf ihre Übereinstimmung bei einer Aufgabe überprüft, die die Klassifikation von 20 sog. ‚random points' erforderte. Alle drei Prozeduren bildeten durchaus ähnliche Cluster (die Korrelationen schwankten zwischen .94 und .97), wobei die complete-link- and average-link-Methoden die größte 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 Möglichkeit, die Adäquanz bzw. Akkuratheit der gefundenen Clusterlösungen zu beurteilen. In einer Reihe von Folgestudien wurde deshalb versucht, gerade dieser Frage gezielt nachzugehen.

2.1. Vergleichsuntersuchungen anhand von empirischen Datensätzen

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 gültig angenommenen a-priori-Gruppierungen erprobt. Die sich in einigen Fällen anschließende diskriminanzanalytische Überprüfung der Clusterlösung sollte im wesentlichen die Ergebnisqualität sichern: wenn sich die clusteranalytisch bestimmten Gruppen auch über ein solchen Verfahren trennen ließen, wurde dies als Beweis dafür gewertet, daß das benutzte Clusteranalyse-Verfahren als valide einzustufen war.

Auf dieser 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-Lösung, während das Taxonome-Programm lediglich eine Hauptgruppe, dafür aber mehrere ‚Klein-Cluster' identifizierte), was wohl nicht zuletzt auf den Tatbestand zurückzuführen war, daß beide Algorithmen unterschiedlichen Kategorien zuzuordnen sind. Während AUKL als partitionierendes Verfahren sämtliche Elemente erfaßt und diese jeweils nur einmal disjunkt klassifiziert, läßt das Taxonome-Programm als ‚clumping technique' (s. o.) sowohl überlappende Cluster als auch die Möglichkeit zu, schwer zuzuordnende Elemente in der Lösung nicht zu berücksichtigen.

Diese Unterschiede arbeitete ALLMER präzise heraus, ohne allerdings den Versuch zu unternehmen, die beiden Verfahren im Hinblick auf die Güte 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 Überlegenheit der (im deutschsprachigen Raum unbekannten) Clusteranalyse von LORR, KLETT & McNAIR (1963) gegenüber 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 übrigen auch bei ALLMER) keine allzu überzeugende Befunde lieferte, zum anderen (wenn auch nicht explizit vermerkt) offensichtlich Korrelationskoeffizienten als Ähnlichkeitsmasse 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 gewählte Ansatz insbesondere wegen seiner Kreativität erwähnenswert. Die Autoren benutzten hierbei die Information, daß bei den genannten Verfahren auf jeder Hierarchie-Ebene eine neue Ähnlichkeitsrangfolge der einzelnen Objektpaare einer Datenstichprobe erstellt wird, die sich mit den originalen Ähnlichkeitsbeziehungen (Distanzen bzw. Korrelationen der Ausgangsmatrix) vergleichen läßt. Der Übereinstimmungsgrad zwischen dem Partitionsrang (d. h. der Hierarchie-Ebene, auf der zwei Objekte erstmals gemeinsam vorkommen) und dem Ähnlichkeitsrang jedes Objektpaars machte Aussagen über das Abschneiden der Clusteranalyse-Verfahren möglich: nur für den Fall einer hohen Entsprechung konnte gefolgert werden, daß die vorgefundene Partitionshierarchie eine angemessene Repräsentation der Daten darstellte. Die Werte für das gewählte Assoziationsmaß (Gamma-Statistik von GOODMAN & KRUSKAL 1954) fielen insgesamt gesehen für das complete-linkage-Verfahren etwas höher aus, ohne allerdings völlig befriedigen zu können.

Dieses Ergebnis stimmt weitgehend mit den Befunden von HUBERT (1974) sowie BAKER & HUBERT (1975) zur Effizienz beider Verfahren überein, wenn auch die letztgenannte Studie eine Differenzierung nahelegt: das single-linkage-Verfahren kommt danach in all denjenigen Fällen zur besseren Lösung, bei denen die ‚wahre' Cluster-Partition ein absolut dominantes Cluster enthält. Leider sind in den Studien von HUBERT und BAKER die Stichprobengrößen für die verwendeten Datenbeispiele äußerst 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, präzise Schlußfolgerungen über die Qualität der überprüften Verfahren zu ziehen, sind wohl entscheidend mit einem spezifischen Phänomen verknüpft: alle mit empirischen (psychologischen) Datensätzen durchgeführten Methodenvergleiche haben nämlich den entscheidenen Nachteil, daß die Datenstruktur nicht genau genug bekannt ist, um als Außenkriterium benutzt werden zu können (vgl. BAUMANN 1973). Wenn man unterstellt, daß die über Klassifikationsverfahren näher zu analysierenden Datensätze 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. inadäquate Rekonstruktion der einzelnen Populationsanteile durch die verschiedenen Cluster-Algorithmen läßt sich somit nicht näher bestimmen.

Entgegen der Annahme vieler Autoren scheint auch das übliche Verfahren, (lineare) Diskriminanzanalysen zur Überprüfung der Güte einer Cluster-Lösung bzw. zur Bestimmung der Lösbarkeit von Klassifikationsaufgaben zu verwenden, der Problematik oft 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 ungünstiges Abschneiden der Diskriminanzanalyse nicht notwendigerweise auf die Unangemessenheit der Cluster-Lösung, sondern möglicherweise eher auf die der linearen Diskriminanzanalyse hin. Die Verwendung dieses Diskriminanzanalysen-Typs wäre 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 über eine lineare Trennfunktion gut unterschieden werden können. Eine Überprüfung der Güte der Cluster-Lösung würde sich dann allerdings in der Regel erübrigen.

2.2. Vergleichsuntersuchungen anhand von Plasmoden

Die oben skizzierten Nachteile der mit empirischem Datenmaterial durchgeführten Evaluationsstudien veranlaßten einige Autoren dazu, real gemessene Datensätze mit bekannter Struktur, sogenannte Plasmoden, zu verwenden. Nach BAUMANN (1971, 1973) läßt sich durch Modelluntersuchungen dieses Typs die grundsätzliche (nicht jedoch die generelle Gültigkeit eines Verfahrens nachweisen).

Grundsätzliche Probleme im Hinblick auf den Vergleich unterschiedlicher Plasmodenstudien (aber auch Monte-Carlo-Studien, s. u.) schienen dadurch gegeben sein, daß die Übereinstimmung zwischen vorgegebener Datenstruktur und der reproduzierten Clusterlösung über verschiedene

Assoziationsmaße (so z. B. Gamma nach GOODMAN & KRUSKAL, RANDs Statistik, COHENs Kappa, CRAMERs V etc.) erfaßt wurde. Da alle diese Maße jedoch über Linearfunktionen ineinander überführbar sind (vgl. HUBERT & LEVIN 1976) und die Interkorrelationen dementsprechend hoch ausfallen (so berichteten MILLIGAN & ISAAC (1980) ein r von .97, EDELBROCK & McLAUGHLIN annähernd identische Befunde für Kappa und Rands Statistik), kann diese Frage im vorliegenden Zusammenhang vernachlässigt werden.

Eine originelle Methode der Datengewinnung wählten BARTKO et al. (1971), indem sie eine Gruppe von 100 fiktiven psychiatrischen Patienten zusammenstellten, die (in je gleichem Umfang) fünf diagnostischen Kategorien zugewiesen wurden. Wenn die Lektüre auch nicht eindeutig erkennen läßt, welche speziellen Cluster-Algorithmen in den Vergleich eingingen, so dürfte 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 erwähnte Etikettierungsproblematik sehr nachteilig bemerkbar). Als wesentliches Ergebnis ließ sich festhalten, daß die complete-linkage-Technik (und damit das hierarchische Verfahren) den beiden übrigen Algorithmen eindeutig überlegen war: während das Verfahren nach LORR et al. nur einen geringen Prozentsatz der Probanden klassifizieren konnte, versagte das iterative Partitionierungsverfahren fast völlig. Interessante Nebenaspekte betrafen die Wahl des Ähnlichkeitskoeffizienten (Euklidische Distanz vs. Korrelationsmaß) und die der zugrundegelegten Datenmatrix (Rohdaten vs. Faktorwerte). Ungünstigere Reproduktionswerte ergeben sich generell für die reduzierten Daten und (überraschenderweise) gerade bei dem complete-linkage-Verfahren für die Euklidischen Distanzwerte.

Als Weiterführung dieses Ansatzes können die Arbeiten von MEZZICH (1975, 1978) gelten, der ebenfalls fiktive ‚archetypale' Patientengruppen erzeugte und für 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 Ähnlichkeitskoeffizienten fungierten. Dieser gerade im Hinblick auf die Selektion von Ähnlichkeitsmaßen mit den Lehrbuchmeinungen kaum vereinbare Befund ließ ‚sich insofern ‚generalisieren, als Korrelationskoeffizienten (über alle verwendeten Verfahren gemittelt) insgesamt besser als die theoretisch eigentlich vorzuziehenden Distanzkoeffizienten abschnitten.

Relativ häufig fanden bei Plasmodenstudien auch Kraftfahrzeugdaten Verwendung. Am Beispiel einer solchen KFZ-Plasmode konnte BAU-

MANN (1971) eine befriedigende Rekonstruktion der Datenstruktur durch das von ihm näher beschriebene Taxonome-Programm (CATTELL & COULTER 1966) nachweisen und in einer weiteren Vergleichsuntersuchung (BAUMANN 1973) weitgehend bestätigen. Während sich für das in der späteren Studie mit in den Vergleich aufgenommene Verfahren der Automatischen Klassifikation ähnliche Ergebnisse fanden, fielen die Befunde für eine Konfigurationsanalyse (KFA) deutlich ab. Spezielle Probleme dieser Vergleichsstudie[5] ergeben sich daraus, daß BAUMANN die KFA auf vorher faktorisiertes Datenmaterial anwendete. Damit ging der spezifische Vorzug der KFA verloren, der darin besteht, daß Zusammenhänge höherer Ordnung erfaßt werden können. Die Befunde zur KFA sind demnach nur beschränkt interpretierbar.

Von EYE (1977) verwendete die KFZ-Plasmode von BAUMANN für den Vergleich zwischen der Clusteranalyse nach WARD und einem selbst entwickelten Verfahren, der multivariaten automatischen Clustersuchstrategie MACS. Beide Prozeduren bildeten die Datenstruktur gut ab und erwiesen sich zusätzlich als ausgesprochen anwendungsökonomisch. Unterschiede zwischen beiden Verfahren wurden auf die verschiedenen Ähnlichkeitsdefinitionen sowie die spezifischen Datenkonstellationen zurückgeführt.

Wenn auch den vorgestellten Plasmodenstudien insgesamt gesehen mehr Aussagekraft ‚(im Sinne einer grundsätzlichen Validität) als den Untersuchungen anhand von empirischen Datensätzen 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-Datensätzen' (sog. ‚benchmark' data sets) besteht insbesondere die Gefahr,

„solche Klassifikationsverfahren als besonders leistungsfähig zu bezeichnen, deren Ergebnisse möglichst gut mit dem Standardergebnis übereinstimmen, obwohl diese Klassifikationsverfahren lediglich geeignet sein können, Ähnlichkeitsstrukturen dieser Art (und keine anderen) aufzuspüren" (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 zurückgegriffen wurde. Da den so gewonnenen Ergebnissen allenfalls eine eingeschränkte Validität zuerkannt werden konnte, verstärkte sich in der Folge der Trend, solche Untersuchungsmodi zu präferieren, bei denen Methodenvergleiche an einer Reihe unterschiedlicher Datenstichproben vorzunehmen waren (s. u.).

Es bleibt hier schließlich noch erwähnenswert, daß die Aussagekraft von einigen der oben erwähnten Studien (ALLMER 1974: BAUMANN

5 Für diesen Hinweis sind wir Herrn Dr. von Eye sehr dankbar.

1971 und 1973) zusätzlich dadurch erheblich gemindert wird, daß in ihnen das Taxonome-Programm von CATTELL & COULTER zum Vergleich herangezogen wurde. Dieses deshalb, weil HARTMANN (1976a) bei der Überprüfung einen schwerwiegenden Fehler des Programms feststellte: es ließ sich zeigen, daß die Ergebnisse mit der Eingabe-Reihenfolge des jeweiligen Datensatzes variierten. Für den von HARTMANN (1976b) umgearbeiteten, immens zeitintensiven Algorithmus liegen bislang keine Resultate aus Vergleichsstudien vor.

2.3. Vergleichsuntersuchungen anhand von Monte-Carlo-Studien

Als Monte-Carlo-Studien werden im allgemeinen solche Untersuchungen bezeichnet, die über die Ziehung von Zufallszahlen aus Populationen mit bekannten Verteilungsparametern Datenstichproben erzeugen, um das Verhalten von mathematisch-statistischen Verfahren gezielt untersuchen zu können.

Bei der Evaluation von Clusteranalyse-Verfahren über Monte-Carlo-Studien werden zunächst mittels Computerprogrammen Datensätze 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 Lösung der auf die generierten Datensätze ‚angesetzten' Clusteranalyse-Verfahren aussehen muß. Dabei bleibt es dem Untersucher weitgehend selbst überlassen, wie er die Populationen definiert. Es ist daher kaum Übereinstimmung darüber 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.

2.3.1. Möglichkeiten 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 Definition, die ein Cluster als Repräsentation von ‚Mixturen' aus multivariat normalverteilten Populationen auffaßt. Während 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 über Monte-Carlo-Techniken sehr leicht Cluster konstruieren, die den Modellannahmen genügen. Das Datenmodell impliziert, daß sich die Verteilungsdichte

der Merkmalsausprägungen einer (Gesamt-)Population aus den Dichten mehrerer Teilpopulationen zusammensetzt. Die Merkmalausprägungen sind durch eine begrenzte Zahl von zugrundeliegenden Dimensionen determiniert, die voneinander unabhängig sind (unkorrelierte Faktoren), und aus denen sich Korrelationen zwischen einzelnen Menschen erklären lassen. Eine wesentliche Annahme besagt, daß sich die Teilpopulationen hinsichtlich der Mittelwerte und Varianzen der (innerhalb der Teilpopulationen normalverteilten Merkmalen 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 läßt. Auf der Grundlage dieser Annahme lassen sich mit einem Computer nahezu beliebig viele Stichproben erzeugen[6].

b) Die Konzeption eines ‚ultrametrischen' Raums:

Einer Anzahl weiterer Untersuchungen (CUNNINGHAM & OGILVE 1972; D'ANDRADE 1978; EVERITT 1974; MILIGAN 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 räumliche 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 Cluster insbesondere nach den Gesichtspunkten ‚externaler ‚Isolation' und ‚internaler Kohäsion' konstruiert werden sollten. Damit ist gemeint, daß Elemente verschiedener Cluster durch nichtbesetzte Räume separiert (externale Isolation) und innerhalb eines Clusters untereinander alle ähnlich sind (vgl. auch die Definition des ‚natürlichen' Clusters bei EVERITT 1974).

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

6 Eine detaillierte Beschreibung der mathematischen Grundlage findet sich bei SCHEIBLER & SCHNEIDER (1978).

Da für die vorzunehmende Evaluation aber weniger die absolute Höhe der Assoziationskoeffizienten als vielmehr die Rangreihe der Clusteranalyse-Verfahren in unterschiedlichen Monte-Carlo-Studien von Bedeutung war, schien eine getrennte Präsentation der beiden Ansätze nicht sinnvoll zu sein. Im Hinblick auf den Stellenwert der Einzelergebnisse verdiente dagegen die Frage nach dem Komplexitätsgrad der jeweiligen Untersuchungen größere Beachtung.

Im folgenden werden deshalb zunächst kurz die Befunde relativ einfach strukturierter Studien präsentiert, 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 für eine summarische Bewertung der einzelnen Algorithmen bilden soll.

2.3.2. Einfach strukturierte Simulationsstudien

Als typische Beispiele für die entweder im Hinblick auf den Umfang der Datenstichproben oder aber die Anzahl der überprüften Clusteranalyse-Verfahren limitierten frühen Simulationsversuche können die Arbeiten von EVERITT (1974) bzw. GROSS (1972) gelten.

EVERITT beschränkte sich auf die Generierung von Zufallsdaten aus bivariaten Normalverteilungen, wobei Sub-Stichproben sowohl aus zwei Populationen (bei Variation von Populationsähnlichkeit und Stichprobengröße) 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 günstig abschnitten, war eine direkte Vergleichbarkeit leider nicht gewährleistet, da jeweils unterschiedlich große Stichproben verwendet wurden. Nachteilig machte sich insbesondere bemerkbar, daß gut separierbare Stichproben erzeugt worden waren, so daß sich die Probleme für 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 künstliche Struktur aufsetzten und Scheinlösungen präsentierten. Die überprüften Verfahren erwiesen sich weiterhin immer dann als suboptimal, wenn Datensätze konstruiert wurden, die andere als sphärische Clusterstrukturen erhielten.

Während 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 Datensätze systematische Veränderungen des Schwierigkeits-

parameters intendiert waren. Als Ausgangsbasis diente ein trivariat verteilter, in zwei Subpopulationen zu untergliedernder Datensatz, bei dem neben der Ähnlichkeit der beiden Grundgesamtheiten auch die Größe der daraus gezogenen Stichproben und deren Größenverhältnis zueinander variiert wurden. Die Rekonstruktionsqualität ergab sich aus der Gegenüberstellung von Fehlklassifikations-Raten, wie sie für die Population erwartet und für die Stichproben verzeichnet worden waren. Leider wurde in dieser Studie mit dem WARD-Verfahren nur ein einziger Algorithmus analysiert, für den die Befunde aus ca. 120 Monte-Carlo-Läufen allerdings durchaus positiv ausfielen: während die Rate der Fehlklassifikationen bei den Stichproben knapp über den für die Population erwarteten Werten lag, ergaben sich bei ungleichen Gruppengrößen, umfangreicheren Stichproben und unähnlicheren Populationen insgesamt günstigere Resultate.

In der 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, Median-Methode, Centroid-Methode und WARD-Technik) überprüft, doch scheinen die Befunde (leichte Überlegenheit der average-linkage-Technik bei sonst gleichwertigen Resultaten) angesichts der schmalen Basis von nur sechs analysierten Datensätzen kaum generalisierbar zu sein.

2.3.3. Komplexer angelegte Monte-Carlo-Untersuchungen

a) Analyse der Reproduktionsgenauigkeit bei hundertprozentiger Objekterfassung

Im Hinblick auf eine erste Orientierung sind in Tab. 1 die wesentlichen Befunde aus sieben Studien enthalten, in denen aufwendigere Analyen zur Evaluation von Cluster-Algorithmen durchgeführt wurden. Während für die Studien von BLASHFIELD (Nr. 1), SCHEIBLER & SCHNEIDER (4) und MOJENA (5) jeweils die Mediane der Assoziationskoeffizienten aufgeführt sind, wurden die Werte für die übrigen Untersuchungen zusätzlich über unterschiedliche Untersuchungsvarianten hinweg arithmetisch gemittelt. Da die so bestimmten Koeffizienten lediglich als nivellierende Grobindikatoren für die Güte der überprüften Verfahren gelten können, soll im Anschluß zusätzlich auf differenziertere Teilergebnisse eingegangen werden.

Bei der Betrachtung von Tab. 1 fällt zunächst einmal auf, daß die absolute Höhe der Koeffizienten für die verschiedenen Verfahren beträchtlich variiert, wobei größere Abweichungen sowohl innerhalb der einzelnen Studien ,als auch für identische Algorithmen über verschiedene Untersuchungen hinweg feststellbar sind. Besonders markant wird der Unter-

Tab. 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)

	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) CRAMER's V
single-link.	.06 (4)	.52 (3)	.44 (6)	.30 (9)	.37 (6)	.91 (6,5)	.15 (5)
compl.-link.	.42 (2)	.61 (2)	.60 (2)	.68 (6)	.64 (2)	.81 (11)	.33 (3)
aver.-link.	.17 (3)		.55 (3)	.73 (5)	.60 (3)	.95 (4,5)	.27 (4)
Centroid			.51 (5)	.32 (8)	.50 (5)	.87 (10)	
Median-Verf.			.52 (4)	.46 (7)	.55 (4)	.88 (9)	
WARD-Verf.	.77 (1)		.62 (1)	.99 (1)	.84 (1)	.90 (8)	.57 (1)
LANCE-WILL.				.98 (3)		.91 (6,5)	
McQUITTY				.74 (4)			
U-Statistik		.76 (1)					
RELOCATE				.99 (1)			
MacQUEEN						.95 (4,5)	
FORGY's Meth.						97 (1,5)	
JANCEY's M.						96 (3)	
Conv. KMEANS						97 (1,5)	.48 (2)

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

schied bei der Evaluation des single-linkage-Verfahrens, dem nach dem Befund von BLASHFIELD (1) kaum mehr als Zufallsklassifikation zugestanden werden können (Kappa = .06), andererseits nach den Ergebnissen von MILLIGAN (6) eine erfreuliche Reproduktionsqualität bescheinigt werden muß (RANDs 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 geäußerten Vermutung lassen sich die Unterschiede in den Absolutwerten wohl nicht zwingend auf Unterschiede in der Bestimmung der theoretischen Cluster-Struktur zurückführen: die auf einer geo- bzw. ultrametrischen Clusterkonzeption aufbauenden Untersuchungen (D'ANDRADE, MILLIGAN, MOJENA) weisen nämlich untereinander ebensolche Differenzen auf, wie sie auch zwischen den auf dem ‚mixture'-Ansatz basierenden Studien (BLASHFIELD, KUYPER & FISHER) zu beeobachten 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, die überlappende 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 exttrem variierenden Anzahl gezogener Zufallsstichproben (von n = 12 bei MOJENA bis n = 766 bei SCHEIBLER & SCHNEIDER) zu begründen. Ein von SCHEIBLER & SCHNEIDER zusätzlich vorgenommener Vergleich von 8 Monte-Carlo-Durchläufen, 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 Zufallseinflüsse zurückzuführen sind.

Eine von den Verfassern durchgeführte Re-Analyse der BLASHFIELD-Daten[7] ergab, daß der von BLASHFIELD verwendete Algorithmus zum einen deutlich mehr ‚outlier' produzierte, andererseits aber auch zu stärkeren Überlappungen zwischen den Unterstichproben führte. Beide Phänomene bewirken eine z. T. erhebliche Senkung der resultierenden Kappa-Werte.

7 Herrn Prof. Dr. R. K. BLASHFIELD (University of Gainesville, Florida) sei für seine Kooperationsbereitschaft, insbesondere für die Überlassung von 10 Mixturen seiner Simulationsstudien nochmals herzlich gedankt.

War also bei BLASHFIELD die prinzipielle Lösbarkeit der Aufgabe, m.a.W. die Entflechtungsmöglichkeit für die Mixturen nicht immer gesichert, so taucht bei der Studie von MILLIGAN (6) ein gerade entgegengesetztes Problem auf: die für alle Algorithmen fast ausnahmslos hohen bis sehr hohen Reproduktionswerte lassen darauf schließen, daß sich die Daten-Elemente als allzugut separierbar und die Aufgaben damit für sämtliche Verfahren als zu leicht erwiesen. Dies ist umso bedauerlicher, als der von MILLIGAN vorgestellte Ansatz in seiner Variabilität (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 über 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 Fällen am schlechtesten ab, und umgekehrt erweist sich die Technik nach WARD ebenso oft als (meist) eindeutig überlegen. Ein Vergleich der SPEARMAN-Rangkorrelationen für die in den Studien (3)–(6) verwendeten sechs hierarchischen Cluster-Algorithmen ergibt für die Befunde von KUYPER & FISHER, MOJENA sowie SCHEIBLER & SCHNEIDER eine nahezu perfekte Übereinstimmung (Koeffizienten von .94 bis 1.0), während die Ergebnisse von MILLIGAN (Übereinstimmungswerte von –.14 bis .13) sich von den übrigen enorm unterscheiden. Klammert man hier einmal die (unten noch näher zu diskutierende) Studie von MILLIGAN aus, so können neben der WARD-Methode demnach auch das complete-linkage-Verfahren bzw. das average-link-Verfahren bedingt, das LANCE-WILLIAMS-Verfahren uneingeschränkt zur Anwendung empfohlen werden, während 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 gewählte Startpartition von großer Bedeutung sein muß: während zufällig ausgewählte Startpartitionen (‚random seed points') generell nur mittelmäßige Resultate (s. Werte in Klammern) erbringen, resultieren die besten Reproduktionswerte aus einer Strategie, bei der die Centroide der ‚group average-linkage-' Methode als Anfangspartition für den iterativen Algorithmus ausgewählt werden. Mit einer solchen Vorgehensweise lassen sich Resultate erzielen, die offensichtlich nur von wenigen hierarchischen Verfahren zu erreichen sind.

Die Bedeutsamkeit der ausgewählten Startpartition für die Effizienz der iterativen Cluster-Algorithmen ließ sich auch in einer Pilot-Studie von

BLASHFIELD (1977 c) nachweisen[8], bei der 20 Monte-Carlo-Mixturen Verwendung fanden. Während die Kappa-Werte für RELOCATE und ANDER im Mittel kaum differierten (.67 vs. .63), ergaben sich für die einzelnen Datensätze je nach gewählter Startpartition enorme Unterschiede, die sich in Kappa-Bandbreiten von .30 bis .94 niederschlugen.

An der hier für iterative Verfahren demonstrierten Relevanz möglicher Wechselwirkungen zwischen Datensatz-Struktur und Anfangspartition zeigt sich gleichzeitig die Notwendigkeit, zusätzlich zu den in Tab. 1 wiedergegebenen ‚overall'-Mittelwerten der aufgeführten Verfahren auch ihre spezifischen ‚Reaktionen' auf systematisch variierte Besonderheiten von Monte-Carlo-Datensätzen zu registrieren. Damit sind Möglichkeiten gegeben, aus der Kenntnis bestimmter Eigenheiten des zu analysierenden Datensatzes heraus auf Algorithmen zurückzugreifen, 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 sorgfältig kontrolliert wurden. Zusätzlich zu idealen, fehlerfrei generierten Datensätzen (‚error free parent data sets') gingen fünf weitere fehlerverzerrte Datenstrukturen in die Analyse ein, die eine Evaluation der Clusteralgorithmen unter Bedingungen ermöglichen sollten, wie sie wohl im Alltagsgebrauch anzutreffen sind. Die Hinzufügung 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 (Verfälschung des Meßprozesses).

Weitere Maßnahmen bestanden darin, daß ‚random noise dimensions' eingebaut wurden, die für die ‚wahre' Cluster-Struktur keinerlei Bedeutung hatten, weiterhin Berechnungen der Ähnlichkeit mit anderen als Euklidischen Distanzmaßen vorgenommen bzw. die Variablen standardisiert wurden.

Die Resultate für die überprüften Clusteranalyse-Techniken lassen sich grob so zusammenfassen, daß die beiden zuletzt erwähnten Schritte fast keine Abnahme in den Rekonstruktionswerten bewirkten, daß die Hinzufügung von ‚random noise dimensions' bei allen Methoden erhebliche Verschlechterungen zur Folge hatten, und daß für die beiden übrigen Maßnahmen gravierende Unterschiede zwischen hierarchischen und nichthierarchischen Techniken zu registrieren waren: während die Rekonstruktionswerte für einige hierarchische Verfahren (complete-link-Methode,

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

Tab. 2:

Abhängigkeit der Algorithmen-Leistung von Besonderheiten der Datenstruktur

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

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

WARD-Technik) bei ‚outlier'-Bedingungen bzw. fehlerverzerrten Distanzen (hier besonders das single-link-Verfahren) stärker abfielen, waren bei den nicht-hierarchischen Prozeduren keinerlei Einbußen zu verzeichnen. Gerade die Befunde zu den ‚outlier'-Bedingungen müssen jedoch erstaunen, da hier trotz der vielfach bestätigten ‚chaining'-Tendenz und der damit verbundenen ‚outlier'-Anfälligkeit 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 einzuschätzen 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-Verfahren sehr ausgeprägt, während sie (mit Ausnahme des average-linkage-Verfahrens) bei den übrigen analysierten Techniken, insbesondere beim WARD-Verfahren, nicht nachzuweisen sind.

Interessante Aufschlüsse können auch aus der in Tab. 2 bei BLASHFIELD (1978) aufgeführten Statistik R-Quadrat (als dem Maß für den Zusammenhang zwischen Kappa-Wert und der Schwierigkeit des Datensatzes) gewonnen werden. Wenn auch bei allen untersuchten Verfahren eine positive Beziehung, also ein günstigerer Reproduktionswert bei schwierigeren Datenkonstellationen festzustellen ist, läßt sie sich insbesondere für das WARD-Verfahren als außerordentlich eng charakterisieren. Dieser Befund wird im übrigen auch von BLASHFIELD (1976) bzw. SCHEIBLER & SCHNEIDER (1978) bestätigt. MOJENA (1977) unterstreicht, daß im Vergleich der hierarchischen Verfahren gerade die WARD-Technik am wenigsten durch den zunehmenden Überlappungsgrad der Stichproben bzw. die ansteigende Zahl der zugrundeliegenden Populationen in ihrer Präzision beeinträchtigt wird, was ihm ebenso wie die zuletzt genannten Autoren zur eindeutigen Bevorzugung dieses Verfahrens veranlaßt.

Wenn auch McINTYRE & BLASHFIELD (1980) bei ihrer Überprüfung von Stabilitätsniveaus verschiedener Cluster-Lösungen aus ‚mixture'-Simulationen für den WARD-Algorithmus eine negative Korrelation zwischen Überlappungsgrad der Stichproben und dem sog. ‚agreement kappa'[9] feststellten, bestätigten auch hier die Befunde den angedeuteten Trend: es ließ sich insgesamt ein mittelhoher Zusammenhang zwischen

9 Dieses Maß für die Übereinstimmung bzw. Stabilität von Cluster-Lösungen verschiedener Stichproben aus einer Mixtur wurde über eine Prozedur gewonnen, die parallel zu dem bei Regressions- bzw. Diskriminanzanalysen häufiger verwendeten Kreuzvalidierungs-Paradigma konstruiert war.

den Stabilitätswerten und den (generell hohen) Präzisionsmaßen (‚accuracy kappa') nachweisen, was die Autoren zur Schlußfolgerung führte, daß Stabilitätskoeffizienten bei dieser Technik eine indirekte Schätzung dafür darstellen können, wie genau die Cluster-Lösungen die tatsächliche Datenstruktur rekonstruieren.

Zur Vervollständigung des überaus komplexen Bildes sind in Tab. 2 die Rangkorrelationen zwischen Kappa-Werten und einigen Statistiken zusammengestellt, die in den Studien von BLASHFIELD (1976) sowie SCHEIBLER & SCHNEIDER (1978) 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 Überlappungsgrades in den Teilstichproben (d). Trotz einiger Abweichungen in Kategorie (d) und der tendenziell höheren Absolutwerte bei BLASHFIELD kann aus den Ergebnissen insgesamt gefolgert werden, daß die vier variierten Parameter keinen allzu großen Einfluß auf die Ausprägung der Rekonstruktionswerte nehmen. Die Statistiken a und b (Stichprobengröße und -repräsentativität) wirken sich nur in der Analyse von BLASHFIELD für das single-linkage-Verfahren in dem Sinne aus, daß bei zunehmender Größe der Gesamt-Stichprobe und abnehmender Repräsentativität der Einzel-Elemente für die Teilstichproben deutlich verringerte Werte zu erwarten sind; Statistik c (spezifische Stichprobenrelationen) hat einen insgesamt zu vernachlässigenden Einfluß, während der Überlappungsgrad 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 Datenkomplexität führt hier zu verbesserten Reproduktionswerten. Dieser Befund kann für die drei genannten Verfahren (und nur für diese) am Beispiel von ähnlich wirkenden Statistiken (in Tab. 2 nicht aufgeführt) validiert werden, die (z. T.) bei BLASHFIELD sowie SCHEIBLER & SCHNEIDER berechnet wurden. Höhere Ergebnisqualität ist danach bei steigender Variablenanzahl (Korrelationen zwischen .62 und .66), bei zunehmender mittlerer Anzahl von Hauptkomponenten (.54–.59) sowie bei größerer Varianz in der Anzahl der Hauptkomponenten bei den Teilstichproben (.46–.47) zu erwarten; die genannten Verfahren schneiden also bei zunehmender Komplexität der Kovarianzstrukturen innerhalb und zwischen den Populationen besser ab.

b) Schrittweise Bewertung der Clusterlösungen auf unterschiedlichen Hierarchie-Ebenen

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

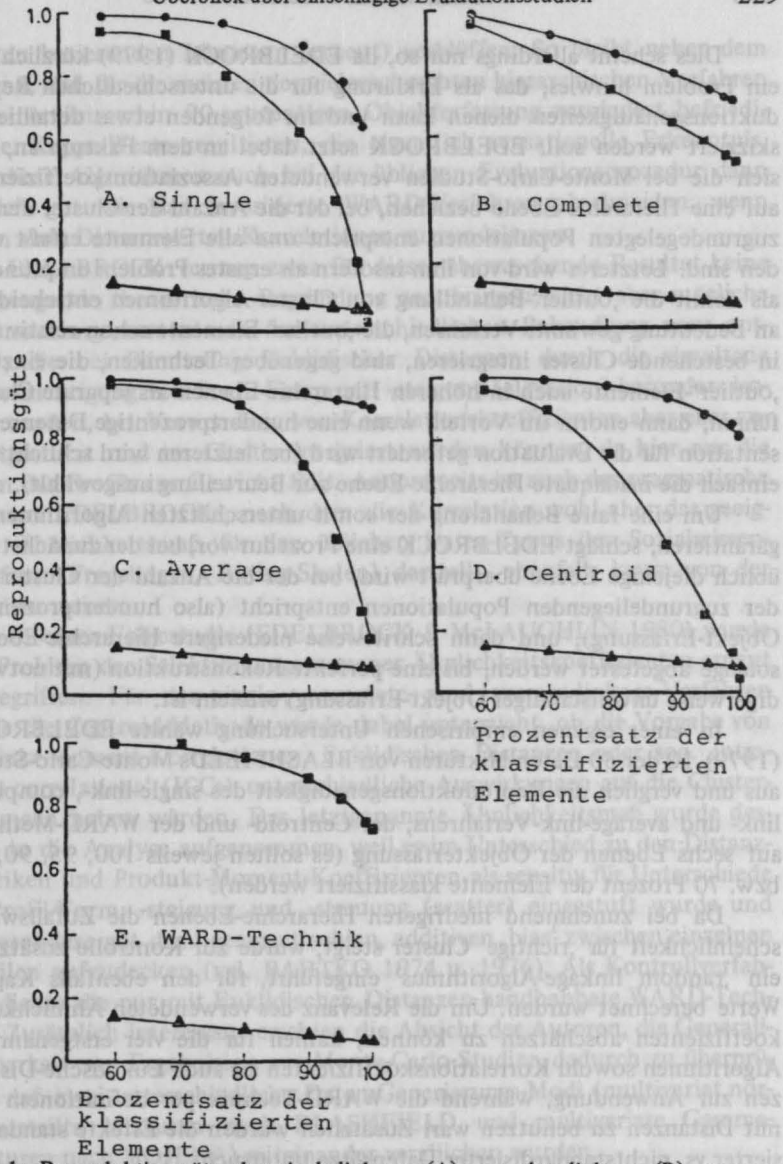

Abb. 1: Reproduktionsgüte des single-linkage- (A), complete-linkage- (B), average-linkage- (C), Centroid- (D) und WARD-Verfahrens (E) für unterschiedliche Prozentsätze zu klassifizierender Elemente nach EDELBROCK (1979). Ergebnisse mit Korr.-Koeffizienten sind mit Kreisen, die mit Distanzkoeffizienten mit Quadraten gekennzeichnet. Für die Resultate eines Zufalls-Algorithmus sind Dreiecke eingesetzt.

Dies scheint allerdings nur so, da EDELBROCK (1979) kürzlich auf ein Problem hinwies, das als Erklärung für 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 früh in bestehende Cluster integrieren, sind gegenüber Techniken, die einzelne ‚outlier'-Elemente auch in höheren Hierarchie-Ebenen als separate Cluster führen, dann enorm im Vorteil, wenn eine hundertprozentige Datenrepräsentation für die Evaluation gefordert wird (bei letzteren wird schlicht und einfach die inadäquate Hierarchie-Ebene zur Beurteilung ausgewählt).

Um eine faire Behandlung der somit unterschätzten Algorithmen zu garantieren, schlägt EDELBROCK eine Prozedur vor, bei der zunächst wie üblich diejenige Ebene überprüft wird, bei der die Anzahl der Cluster der der zugrundeliegenden Populationen entspricht (also hundertprozentige Objekt-Erfassung), und dann schrittweise niederigere Hierarchie-Ebenen solange abgetestet werden, bis eine perfekte Rekonstruktion (mit notwendigerweise unvollständiger Objekt-Erfassung) erreicht ist.

In einer eigenen empirischen Untersuchung wählte 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 für ‚richtige' Cluster steigt, wurde zur Kontrolle zusätzlich ein ‚random linkage-Algorithmus' eingeführt, für den ebenfalls Kappa-Werte berechnet wurden. Um die Relevanz des verwendeten Ähnlichkeitskoeffizienten abschätzen zu können, kamen für die vier erstgenannten Algorithmen sowohl Korrelationskoeffizienten als auch Euklidsche-Distanzen zur Anwendung, während die WARD-Technik per definitionem nur mit Distanzen zu benutzen war. Zusätzlich wurden die Effekte standardisierter vs. nichtstandardisierter Datensätze untersucht.

Wie aus Abb. 1 zu erkennen ist, fallen die Relationen zwischen Akkuratheit und Grad der Element-Erfassung (coverage) für die analysierten Algorithmen durchaus unterschiedlich aus. Wenn auch die Reproduktionsgenauigkeit mit zunehmendem Grad der Element-Erfassung generell abnimmt, so muß die konsistente Überlegenheit der auf Korrelationskoeffi-

zienten basierenden Lösungen (erneut) verblüffen. So bleibt neben dem Befund, daß für die meisten der vielgeschmähten hierarchischen Verfahren bei einer immerhin 90-prozentigen Objekterfassung zumindest befriedigende Kappa-Werte resultieren, die eigentlich sensationelle Erkenntnis, daß diese Algorithmen auch bei der üblichen Evaluationsprozedur dann ähnlich gut wie das favorisierte WARD-Verfahren abschneiden, wenn ihnen statt Distanzwerten Korrelationen zugrundeliegen.

EDELBROCK vermag zwar für dieses überraschende Resultat keine überzeugende theoretische Begründung anzubringen, sieht aber mögliche Verursachungsmomente in der unterschiedlichen Behandlung von ‚outliers', die bei Benutzung Euklidischer Distanzen durch die simultane Berücksichtigung von Profil-Form und -steigung (elevation) besonders isoliert werden, bei Verwendung von Korrelationskoeffizienten aber eher zurechtgestutzt und ins Cluster integriert werden können, da hier nur die Form des Profils ins Gewicht fällt. Andererseits ist auch das pragmatische Element EDELBROCKs, nach dem die Korrelation wohl eher das geeignete Ähnlichkeitsmaß für den üblichen 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 Ähnlichkeitskoeffizienten erneut aufgegriffen. Für das single-, complete- und average-linkage-Verfahren sowie die Centroid-Methode wurde dabei untersucht, ob die Vorgabe von Produkt-Moment-Korrelationen, Euklidischen Distanzen oder sog. ‚intraclass correlations' (ICCs) unterschiedliche Auswirkungen auf die Cluster-Lösungen haben würden. Das letztgenannte Ähnlichkeitsmaß wurde deshalb in die Analyse aufgenommen, weil es im Unterschied zu den Distanz-Metriken und Produkt-Moment-Koeffizienten als sensitiv für 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 Euklidischen Distanzen handhabbare WARD-Technik. Zusätzlich interessant erschien die Absicht der Autoren, die Generalisierbarkeit von Ergebnissen aus Monte-Carlo-Studien dadurch zu überprüfen, 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 gewählten Ähnlichkeitsmaß und der Art der Stichproben-Generierung zutage brachten: während Euklidische Distanzen bei den BLASHFIELD-Daten generell ungenauere Lösungen als die üblichen Ähnlichkeitsmaße lieferten, waren

es bei den Mixturen nach MOJENA die Produkt-Moment-Korrelationen, die insgesamt am schwächsten abschnitten. EDELBROCK & McLAUGHLIN schließen daraus, daß bei den multivariat normalverteilten Daten BLASHFIELDs die *Form* als wesentlicher Profil-Parameter auftritt, während 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, so daß bei Generalisierungsversuchen besondere Vorsicht geboten scheint.

Es ergab sich also in dieser Studie als ein wesentliches Ergebnis, daß sich über 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 übrigen Kombinationen aus Algorithmen und Ähnlichkeitsmaßen. Obwohl das ICC-Maß seine Bewährungsprobe damit sicherlich bestanden hat, folgern die Autoren aus ihren Ergebnissen, daß kein einziger Algorithmus für alle Anwendungsmöglichkeiten gleichermaßen optimal geeignet ist. Ausschlaggebend sollten deshalb theoretische Vorüberlegungen des Forschers z. B. zur Frage sein, ob er nun eher daran interessiert ist, Cluster vorzugsweise über 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 Ähnlichkeitsmaßes sein sollte.

3. Zusammenfassung und Diskussion

Die Übersicht über die in den unterschiedlichen wissenschaftlichen Disziplinen durchgeführten Versuche zur Bewertung von Clusteranalyse-Verfahren mag einen ungefähren Eindruck davon vermittelt haben, welche Schwierigkeiten mit einer eindeutigen Einordnung der geprüften Algorithmen verbunden sind.

Es ließ sich zunächst einmal klar herausstellen, daß die Vergleichsuntersuchungen anhand von empirischen Datensätzen wenig effizient sein konnten: die ‚wahre' Datenstruktur ist in solchen Fällen einfach zu wenig bekannt, um als Außenkriterium benutzt werden zu können. Nicht viel anders verhält es sich mit Evaluationsstudien, die sich auf real gemessene Datensätze mit bekannter Struktur, m.a.W. auf Plasmoden stützen. Wenn solchen Untersuchungen sicherlich auch nach BAUMANN (1973) mehr Aussagekraft im Sinne einer grundsätzlichen Validität zugestanden werden

muß, bleibt ihre Generalisierbarkeit angesichts der Vorliebe vieler Forscher, immer wieder auf ‚klassische' Datensätze zurückzugreifen, doch äußerst umstritten. Argumente für eine solche Sichtweise lassen sich (von den Autoren im übrigen unbeachtet) aus einer Untersuchung von MEZZICH & SOLOMON (1980) gewinnen, bei der insgesamt vier verschiedene Plasmoden-Datensätze herangezogen wurden, um die Leistungsfähigkeit von 18 Cluster-Algorithmen zu vergleichen. Aus einer Ergebnisübersicht (vgl. S. 135) kann für mehrere CA-Verfahren nachgewiesen werden, daß sich ihre Rangplatzziffer über die vier Datensätze hinweg z. T. beträchtlich verändert (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 über die Ziehung von Zufallszahlen Datenstichproben aus Populationen mit bekannten Verteilungsparametern erzeugen lassen, wobei hier nun die Möglichkeit gegeben ist, (annähernd) beliebig viele Datenstichproben zu erzeugen und damit das ‚Verhalten' von CA-Algorithmen systematischer zu analysieren. Erste Schritte in dieser Richtung waren aber deshalb nur von eingeschränktem Wert, weil entweder zu wenige Algorithmen in die Untersuchung einbezogen wurden oder aber (aus welchen Gründen auch immer) immer noch zu wenige Datenstichproben resultierten.

Die Übersicht über die Befunde komplexerer Simulationsstudien (wie sie etwa in Tab. 1 gegeben wird) zeigt demgegenüber im Vergleich zu den erwähnten 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 für den Fall hin, daß die Algorithmen alle Elemente klassifizieren müssen. 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 für die herangezogenen iterativ-partitionierenden Verfahren günstige Vergleichswerte registrieren lassen. Für letztere ist dies aber immer nur dann der Fall, wenn für die Startpartition keine Zufallswerte, sondern die Ergebnisse einer anderen (hierarchischen) Clusteranalyse gewählt 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'-anfälligen Techniken werden nämlich dadurch benachteiligt, daß hier die Anzahl der Cluster gleich der Anzahl der Subpopulationen gesetzt wird und gleichzeitig alle Elemente klassifiziert werden

müssen. Wie die Ergebnisse von EDELBROCK zeigen, werden auch von allgemein schwach eingestuften Verfahren wie der single-linkage-, der average-linkage- und der Median-Methode durchaus befriedigende Lösungen erzielt, wenn nur bestimmte Prozentsätze der Elemente bei der Klassifikation zu berücksichtigen sind. Geht man davon aus, daß in der psychologischen Untersuchungspraxis absolut vollständige Zuordnungen nur in den wenigsten Fällen erforderlich sind, ist diese Information über das ‚Verhalten' von Algorithmen bei der Ausklammerung von outliern sicherlich sehr nützlich.

Schon bei der Durchsicht der Ergebnisse von Tab. 1 fiel auf, daß die Absolutwerte der für die Bewertung verwendeten Gütemaße in den verschiedenen Studien z. T. erheblich differieren. Diese Unterschiede lassen sich dabei kaum vollständig auf die unterschiedliche Definition der theoretischen Cluster-Struktur in den einzelnen Untersuchungen (also die alternative Konzeption von ‚mixture model' vs. ‚ultrametrischer Raum', s. o.) zurückführen, 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 Übereinstimmungs-Werte von ‚Durchschnitt' der Evaluationsstudien machen nachdrücklich darauf aufmerksam, daß unter dem Etikett ‚Monte-Carlo-Studie' durchaus unterschiedliche Vorgehensweisen subsummiert sind.

Wie schon oft näher ausgeführt, dürften die unterschiedlichen Eigenschaften der generierten Datensätze 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 für die in einigen neueren Untersuchungen aufgefundenen Resultate verantwortlich zu machen ist, denen zufolge bei der Wahl von Korrelationskoeffizienten günstigere Befunde zu erwarten sind. Wir gehen davon aus, daß diese Frage aufgrund der bisher durchgeführten Evaluationsstudien nicht eindeutig zu beantworten ist.

Es soll deshalb in einer eigenen Untersuchung (SCHNEIDER & SCHEIBLER, 1983 a) der Versuch gemacht werden, einige der in dieser Literaturübersicht nachgewiesenen Probleme erneut aufzugreifen, um zu einer angemessenen Bewertung der CA-Verfahren zu gelangen.

Literatur

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 Beiträge, 16, 1974, 605–617.

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

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

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

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

Bartko, J. J., J. S. Strauss & W. T. Carpenter: An evaluation of taxometric techniques for psychiatric data. Classification Society Bulletin, 2, 1971, 2–28.

Baumann, U.: Psychologische Taxometrie – Eine Methodenstudie über Ähnlichkeitskoeffizienten ‚Q'-Clusteranalyse, Q-Faktorenanalyse. Bern: Huber 1971.

Baumann, U.: Die Konfigurationsfrequenzanalyse, ein taxometrisches Verfahren. Psychologische Beiträge, 15, 1973, 153–168. P

Blashfield, R. K.: Mixture model tests of cluster analysis: Accuracy of four hierarchical agglomerative methods. Psychological Bulletin, 83, 1976, 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, 15, 1980, 439–458.

Blashfield, R. K. & M. S. Aldenderfer: The literature on cluster analysis. Multivariate Behavioral Research, 113, 1978, 271–295.

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

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

Cunningham, K. M. & J. C. Ogilvie: Evaluation of hierarchical grouping techniques: A preliminary study. The Computer Journal, 15, 1972, 209–213.

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

Eckes, T. & H. Roßbach: Clusteranalysen. Stuttgart: Kohlhammer 1980.

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

Edelbrock, C. & B. McLaughlin: Hierarchical cluster analysis using intraclass correlations: A mixture model study. Multivariate Behavioral Research, 15, 1980, 299–318.

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

Faber, E. & W. Nollau: Über 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, 7, 1936, 179–188.

Goldstein, S. G. & J. D. Linden: A comparison of multivariate grouping techniques commonly used with profile data. Multivariate Behavioral Research, 4, 1969, 103–114.

Goodman, L. A. & W. H. Kruskal: Measures of association for cross-classification. Journal of the Americal Statistical Association, 49, 1954, 732–764.

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

Hartmann, W.: Über ein Verfahren der numerischen Taxonomie von Cattell und Coulter. Biometrische Zeitschrift, 18, 1976, 273–290 (a).

Hartmann, W.: Über einen Algorithmus zur Clusteranalyse maximalkompakter Gruppen und die rechentechnische Realisierung des Verfahrens von Cattell und Coulter. Biometrische Zeitschrift, 18, 1976, 333–349 (b).

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

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

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

Kuyper, F. K. & C. Fisher: A Monte Carlo comparison of six clustering procedures. Biometrics, 31, 1975, 777–783.

Lorr, M., C. J. Klett & D. M. McNair: Syndromes of psychosis. New York: MacMillan 1963.

McIntyre, J. B. & R. K. Blashfield: A nearest-centroid technique for evaluating the minimum-variance clustering procedure. Multivariate Behavioral Research, 15, 1980, 225–238.

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

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

Mezzich, J. E. & H. Solomon: Taxonomy and behaviorial 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, 45, 1980, 325–342.

Milligan, G. W.: A review of Monte Carlo tests of cluster analysis. Multivariate Behavioral Research, 16, 1981, 379–407.

Milligan, G. W. & P. D. Isaac: The validation of four ultrametric clustering algorithms. Pattern Recognition, 12, 1980, 41–50.

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

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

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

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

Schneider, W. & D. Scheibler: Probleme und Möglichkeiten bei der Bewertung von Clusteranalysen. II. Ergebnisse einer Monte-Carlo-Studie. Pschologische Beiträge, 24, 1983, H. 2 (im Druck) (a).

Schneider, W. & D. Scheibler: Probleme und Möglichkeiten bei der Bewertung von Clusteranalysen. III. Appendix: Kurzbeschreibung der verbreitesten Clusteranalyse-Algorithmen. Psychologische Beiträge, 24, 1983, H. 3 (im Druck) (b).

Sneath, P. H. A.: A comparison of different clustering methods as applied to randomly spaced points. Classification Society Bulletin, 1, 1966, 2–18.

Steinhausen, D. & K. Langer: Clusteranalyse. Berlin: de Gruyter 1977.

Vogel, F.: Probleme und Verfahren der automatischen Klassifikation. Göttingen: Vandenhoeck & Ruprecht 1975.

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

Von Eye, A. & M. Wirsing: An attempt of a mathematical foundation and evaluation of MACS, a method for multidimensional automatical cluster detection. Biometrical Journal, 20, 1978, 655–666.

Von Eye, A. & M. Wirsing: Cluster search by enveloping space density maxima. In: M. M. Barritt & D. Wishart (Eds.), COMPSTAT 1980. Vienna: Physica-Verlag 1980.

Wishart, D.: CLUSTAN 1C user manual. London: Computer Center 1975.

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