BERICHT
aus dem
PSYCHOLOGISCHEN INSTITUT
DER UNIVERSITÄT HEIDELBERG

Probleme und Ergebnisse bei der Evaluation von Clusteranalyse-Verfahren

Dieter Scheibler und Wolfgang Schneider

Diskussionspapier Nr. 11 Juni 1978

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

PROBLEME UND ERGEBNISSE BEI DER EVALUATION VON CLUSTERANALYSE-VERFAHREN

Dieter Scheibler & Wolfgang Schneider

Diskussionspapier Nr.11

Juni 1978

GLIEDERUNG

1. Einleitung:

Der Begriff Clusteranalyse kennzeichnet eine Reihe von statistischen Verfahren, die Personen, Objekte oder Elemente eines Datensatzes in ähnliche Gruppen zu klassifizieren versuchen. Obwohl verschiedene Techniken zur Gruppierung von Individuen schon länger bekannt waren (vgl. TRYON 1939), setzte erst mit der raschen Weiterentwicklung von EDV-Anlagen innerhalb der letzten Jahrzehnte eine sprunghaft verlaufende Popularisierung dieser multivariaten Methode ein, was zu einer Flut von einschlägigen Publikationen und zur Präsentation von mehr als hundert verschiedenen Algorithmen führte.

Für den relativ unerfahrenen Benutzer von Clusteranalyse-Verfahren macht sich negativ bemerkbar, daß eine einheitliche Nomenklatur der Algorithmen bislang noch fehlt. Die Tatsache, daß sich für ein- und dieselbe Technik mehrere Termini (z.B. single-linkage-method, nearest-neighbour-method, minimum method, linkage analysis) finden lassen, erschwert daher eine vorläufige Kategorisierung der verschiedenen Verfahren. Wesentlich gewichtiger erscheint jedoch ein weiterer Punkt: obwohl bekannt ist, daß die Anwendung verschiedener Clusteranalysen an dem gleichen Datensatz zu unterschiedlichen Ergebnissen führt, mangelt es derzeit immer noch an Methodenstudien, die eine Bewertung der einzelnen Verfahren hinsichtlich ihrer Eignung für spezifische Datenstrukturen zulassen. Die weit verbreitete Praxis, für den zu bearbeitenden Datensatz einen zufällig gerade verfügbaren Algorithmus auszuwählen, wird verständlich, wenn selbst in einer Einführung in die (hierarchische) Clusteranalyse (ROLLETT & BARTRAM 1976) die Selektion des zu beschreibenden Verfahrens nur damit begründet werden kann, daß es sich wachsender Beliebtheit erfreut (vgl. S.21).
Demgegenüber wird in dem vorliegenden Beitrag versucht, die bisher durchgeführten Methodenvergleiche unter Einschluß einer eigenen Untersuchung in ihren Eigenarten zu skizzieren und zu bewerten, da wir mit SIBSON (1971, zit. n. EVERITT 1974, S.97) der folgenden Auffassung sind:

> " The time has come for putting more effort into the efficient implementation of methods which are known to be useful and whose theoretical background is well understood rather than the development of yet more ad-hoc techniques. "

2. Überblick über vorliegende Methodenvergleiche

2.1. Untersuchungen mit empirischen Datensätzen

Diese Form der Vergleichsstudie ist meist durch einen einfachen Versuchsaufbau gekennzeichnet: an einem beliebigen empirischen Datensatz werden mehrere Clusteranalysen durchgeführt, deren Ergebnisse sich diskriminanzanalytisch überprüfen lassen. Die mehr oder minder gelungene Replikation der clusteranalytisch bestimmten Gruppen in der diskriminanzanalytischen Lösung wird als Beweis dafür gewertet, daß die real zugrundeliegende Datenstruktur durch die betreffenden Verfahren rekonstruiert werden konnte.
ALLMER (1974) verglich auf diese Weise das Taxonome-Programm[1)] von CATTELL & COULTER (1966) und das Verfahren der Automatischen Klassifikation (FABER & NOLLAU 1969) anhand einer Stichprobe von Leistungsmotivations-Dimensionen deutscher und schweizerischer Leistungssportler, konnte jedoch aufgrund seiner Ergebnisse keine der beiden Prozeduren eindeutig präferieren. Das Vorgehen von GOLDSTEIN & LINDEN (1969) bzw. ROGERS & LINDEN (1973) unterschied sich nur unwesentlich vom vorher geschilderten Ansatz: sowohl an psychiatrischen als auch an psychologischen Stichproben ergaben beide Untersuchungen eine Überlegenheit der Clusteranalyse nach LORR, KLETT & McNAIR (1963) gegenüber der hierarchischen Gruppierung nach WARD (1963) und einer Hauptkomponentenanalyse (bei GOLDSTEIN & LINDEN) bzw. der WARD-Methode, dem single-linkage- und dem complete-linkage-Verfahren nach JOHNSON (bei ROGERS & LINDEN).
Einen eleganteren Ansatz wählten HUBERT & BAKER (1976), um

---

1) Eine Kurzbeschreibung der einzelnen Techniken findet sich in Anhang A

die Effizienz der single-linkage- mit der der complete-linkage-Methode zu vergleichen: da bei diesen hierarchischen Clusteranalyse-Verfahren auf jeder Hierarchie-Ebene eine neue Ähnlichkeitsrangfolge der einzelnen Objektpaare einer Datenstichprobe erstellt wird, deren originale Ähnlichkeitsbeziehungen (in Form von Korrelations- bzw. Distanzmatrizen) vorliegen, wird für jedes Objektpaar der Partitionsrang (d.h. die Hierarchie-Ebene, auf der beide Objekte erstmals gemeinsam auftauchen) mit seinem Ähnlichkeitsrang über Rangkorrelation in Bezug gesetzt. Nur im Fall einer hohen Übereinstimmung kann gefolgert werden, daß die Partitionshierarchie eine angemessene Repräsentation der Daten darstellt. Durch eine Anwendung dieses Prinzips auf jeder Hierarchie-Ebene läßt sich zusätzlich bestimmen, welche Charakteristika der Ergebnisse interpretierbar sind.
Für den von den Autoren untersuchten Datensatz ergaben sich sowohl für die complete-linkage- als auch für die single-linkage-Methode keine günstigen Korrelationswerte, wenn auch bei der letzteren noch eine schwache Identifikation der zugrunde liegenden Matrix festgestellt werden konnte.
Leider haben alle mit empirischen (psychologischen) Datensätzen durchgeführten Methodenvergleiche den entscheidenden Nachteil, daß die Datenstruktur nicht genau genug bekannt ist, um als Außenkriterium benutzt werden zu können (vgl. BAUMANN 1973). Es muß vielmehr unterstellt werden, daß die der Clusteranalyse zugrunde liegenden Datensätze als 'Daten-Subeinheiten' mehrerer Populationen zu verstehen sind, wobei deren Anzahl und Verteilungsparameter unbekannt bleiben.
Die befriedigende bzw. unbefriedigende Rekonstruktion der einzelnen Populationsanteile durch die verschiedenen Clusteranalysen läßt sich demnach bei Verwendung von empirischen Datensätzen nicht bestimmen.
Weiterhin scheint das übliche Verfahren, (lineare) Diskriminanzanalysen zur Überprüfung der Güte einer Cluster-Lösung zu verwenden, der Problematik nicht angemessen zu sein.
Da die meisten Clusteranalyse-Algorithmen disjunkte Gruppen bilden, die mit einer (nicht geraden) Trennungslinie eindeutig unterscheidbar sind, deutet deshalb das ungünstige Abschneiden einer Diskriminanzanalyse nicht auf die Unangemessenheit der Cluster-Lösung, sondern vielmehr auf die der Diskriminanzanalyse hin. Letztere wäre nur dann sinnvoll, wenn man von der Clusteranalyse zusätzlich fordern würde, daß sie nur solche Cluster bilden dürfe, die über eine (lineare) Trennfunktion gut unterscheidbar sind (dies ist z.B. bei der Automatischen Klassifikation nach FABER & NOLLAU der Fall). Eine Überprüfung der Güte der Cluster-Lösung erübrigt sich dann allerdings.

2.2. Vergleichsstudien anhand von Plasmoden

Als Plasmoden bezeichnet man real gemessene Datensätze mit bekannter Struktur; nach BAUMANN (1971 u. 1973) läßt sich durch Modelluntersuchungen die grundsätzliche (nicht jedoch die generelle) Gültigkeit eines Verfahrens nachweisen.
Für den Fall von KFZ-Daten konnte BAUMANN (1971) eine befriedigende Rekonstruktion der Datenstruktur durch das von ihm beschriebene Taxonome-Programm (CATTELL & COULTER 1966) nachweisen und in einer weiteren Vergleichsuntersuchung (BAUMANN 1973) bestätigen; für das in der späteren Studie mit in den Vergleich aufgenommene Verfahren der Automatischen Klassifikation fanden sich ähnliche Ergebnisse, während die Resultate für eine Konfigurationsfrequenzanalyse (KFA) deutlich abfielen.
Von EYE (1977) verwandte 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 zudem als ausgesprochen anwendungsökonomisch.
Dennoch kann auch diesen Befunden - insbesondere, wenn sie das Taxonome-Programm angehen - keine übermäßige Bedeutung zugemessen werden. HARTMANN (1976a) stellte bei einer Überprüfung einen gewichtigen Fehler dieses Programms fest: die Ergebnisse variieren mit der Eingabe-Reihenfolge des jeweiligen Datensatzes. 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 anhand von Zufallszahlen Populationen bzw. Stichproben mit vorgegebenen Verteilungsparametern erzeugen, um mathematisch-statistische Verfahren adäquater überprüfen zu können.

EVERITT (1974) generierte auf diese Art und Weise Zufallsvariablen aus bivariaten Normalverteilungen mit folgenden Varianten: der erste Datensatz rekrutierte sich aus einer Stichprobe, die einer einzigen Population entnommen war; es interessierte das 'Verhalten' verschiedener Clusteranalysen bei einer nicht weiter unterteilbaren Konfiguration. Weitere Datensätze enthielten Stichproben aus zwei Populationen, wobei die Ähnlichkeit der Populationen und die Größe der gezogenen Stichproben jeweils modifiziert wurden. Für ein Optimierungsverfahren (McRAE 1971) ergaben sich insgesamt befriedigende Rekonstruktionswerte; es machte sich allerdings nachteilig bemerkbar, daß die Technik den Daten eine Struktur aufsetzt (sphärische Cluster), anstatt deren tatsächliche Struktur aufzudecken. Von den drei verwendeten hierarchischen Clusteranalysen (single-linkage-, Centroid- und WARD-Methode) schnitt das single-linkage-Verfahren insgesamt am günstigsten ab (die beiden anderen waren jedoch kaum schlechter), was der Autor auf die nicht-überlappenden, gut separierbaren Stichproben zurückführte. Für ein sog. Dichte-Verfahren (Mode analysis) nach WISHART (1969) ergaben sich ebenfalls im wesentlichen befriedigende Ergebnisse. Eine direkte Vergleichbarkeit aller herangezogenen Techniken war leider nicht gewährleistet, da unterschiedlich große Stichproben verwendet wurden. Nachteilig machte sich jedoch vor allem bemerkbar, daß sich die Aufgabe für die verschiedenen Methoden als zu leicht erwies.

Die Untersuchung von GROSS (1972) ist demgegenüber besonders interessant, da hier der Schwierigkeitsparameter für das

überprüfte Verfahren (WARD-Methode) stärker berücksichtigt wurde. Als Ausgangsbasis diente ein trivariat verteilter, in zwei Sub-Populationen zu untergliedernder Datensatz, bei dem neben der Ähnlichkeit der beiden Sub-Populationen (Schwierigkeitsindex) die Größe der daraus gezogenen Stichproben und deren Größenverhältnis zueinander variiert wurden. Um die Güte der Rekonstruktion zu bestimmen, wurden die Wahrscheinlichkeiten der Fehlklassifikationen für Populationen und Stichproben einander gegenübergestellt. Die Ergebnisse von 120 Monte-Carlo-Studien fielen für das WARD-Verfahren durchaus positiv aus: die Rate der Fehlklassifikationen bei den Stichproben war nur unwesentlich höher als die für die Population erwarteten Werte; insgesamt günstigere Resultate ergaben sich bei ungleichen Gruppengrössen, größeren Stichproben und unähnlicheren Populationen.

Den bisher elaboriertesten Ansatz zur Bewertung von verschiedenen Clusteranalysen stellte BLASHFIELD (1976) vor: er griff auf das 'mixture model' (WOLFE 1970) der Clusteranalyse zurück, das jeden zu analysierenden Datensatz als Amalgam von Daten-Subgruppen einer Anzahl von Populationen begreift, wobei die exakte Anzahl und die Verteilungs-Parameter der Populationen unbekannt sind. Die Schwierigkeit, das 'mixture model' mathematisch zu lösen, wird in dieser Untersuchung dadurch umgangen, daß via Monte-Carlo-Studien Populationen mit bekannten Verteilungs-Parametern erzeugt und vermischt werden; die Güte der Clusteranalysen kann daran demonstriert werden, wie adäquat sie die ursprünglichen Populationen rekonstruieren können. Zur Messung der Übereinstimmung zwischen den jeweiligen Cluster-Lösungen und der adäquaten Klassifikation wurde die Statistik Kappa (COHEN 1960) verwendet, die in Äquivalenz zu den meisten Ähnlichkeitsmaßen zwischen 0 und 1 variiert.

Für die vier überprüften hierarchischen Cluster-Verfahren (single-linkage-, complete-linkage-, average-linkage- und WARD-Methode) ergaben sich unterschiedliche Resultate, wie Tab. 1 zeigt: deutlich am besten schnitt die Clusteranalyse

Tab. 1:

Durchschnittliche Kappa-Werte als Maße der Übereinstimmung zwischen Cluster-Lösungen und 'wahren' Klassifikationen [1)]

Clusteranalyse	Kappa-Koeffizient Median	Interquartilber.
single-linkage-Methode	.06	.03 - .10
complete-linkage-Methode	.42	.22 - .58
average-linkage-Methode	.17	.06 - .46
WARD-Methode	.77	.42 - .94

1) nach BLASHFIELD (1976, S.383)

nach WARD ab, gefolgt von der complete-linkage-Technik; die beiden anderen Verfahren fielen dagegen deutlich ab. Diese auf der Basis von 50 'mixtures' gewonnenen Ergebnisse unterstreichen die generelle Überlegenheit der WARD-Methode gegenüber den übrigen drei hierarchischen Verfahren.

So differenziert und faszinierend der Ansatz von BLASHFIELD auch wirken mag; es darf nicht übersehen werden, daß hier nur wenige Algorithmen verglichen werden, von denen zwei keine echten Alternativen zur WARD'schen Technik darstellen. In den letzten Jahren sind jedoch Verfahren entwickelt worden, die sich nicht nur etwa an der single-linkage-Methode

(die inzwischen wohl als antiquiert zu bezeichnen ist) messen mußten. Unter diesem Gesichtspunkt schien es uns angebracht, die Methode von BLASHFIELD aufzugreifen und fortzuführen. Für die Auswahl der in die Analyse mit einzubeziehenden Cluster-Algorithmen waren theoretische und pragmatische Gesichtspunkte gleichermaßen von Bedeutung: zum einen sollten Techniken untersucht werden, die in echter Konkurrenz zum WARD'schen Algorithmus stehen bzw. über deren Eigenschaften noch relativ wenig bekannt ist. Andererseits schien es sinnvoll zu sein, sich auf solche Verfahren zu beschränken, die am Universitätsrechenzentrum Heidelberg implementiert und damit für Benutzer aus dem Psychologischen Institut zugänglich sind.

3. Beschreibung des eigenen Untersuchungsansatzes

3.1. Methode

Wie schon oben erwähnt, orientierten wir uns weitgehend an dem von BLASHFIELD beschriebenen methodischen Ansatz. Aus Gründen der Vergleichbarkeit wurden keine entscheidenden Änderungen vorgenommen, obwohl einige durchaus denkbar schienen.

Voraussetzung für eine Monte-Carlo-Studie ist die Festlegung eines Datenmodells (Annahme über die Beschaffenheit der Daten). BLASHFIELD folgte dabei weitverbreiteten Modellvorstellungen in der Psychologie [2)]. Es wird zunächst davon ausgegangen, daß sich die Verteilungs-Dichte der Merkmalsausprägungen einer (Gesamt-) Population aus den Dichten mehrerer Teilpopulationen zusammensetzt. Die Merkmalsausprägungen sind durch eine begrenzte Zahl von "zugrunde liegenden Dimensionen" determiniert, die voneinander unabhängig sind (unkorrelierte Faktoren), und aus denen sich Korrelationen zwischen einzelnen Merkmalen erklären lassen.

2) Exakte mathematische Darstellung in Anhang B

Entscheidend sind nun die folgenden Annahmen:
Die Teilpopulationen unterscheiden sich hinsichtlich der Mittelwerte und Varianzen der Merkmale sowie im Hinblick auf die Korrelation zwischen den einzelnen Merkmalen (und somit auch hinsichtlich der Faktorenstruktur). Innerhalb der Teilpopulationen sind die Merkmale normalverteilt. Weiterhin wird angenommen, daß sich die Merkmale durch ihre Meßwerte nicht exakt erfassen lassen, diese vielmehr fehlerbehaftet sind (Reliabilität kleiner als 1). In Anlehnung an BLASHFIELD wird der Meßfehler als gleichverteilt im Bereich $\pm$ 0.6 mal der Standardabweichung der Variablen angenommen.

Auf der Grundlage dieser Annahmen lassen sich mit einem Computer nahezu beliebig viele Stichproben erzeugen. Um die Aufgabe für die Cluster-Algorithmen nicht allzu leicht zu machen, werden Grenzwerte für die einzelnen Populationsparameter festgelegt. Innerhalb dieser Grenzen sollen die Ausprägungen der Parameter gleichverteilt sein.

Tab. 2:

Untere und obere Grenzen für die wichtigsten Populationsparameter

Populationsparameter		kleinster	größter Wert
Mittelwerte (Erwartungswerte)	μ	40	60
Varianzen		3	20
Korr.-Koeffizienten r		-1	+1
arccos (r)		0	2 π
Anzahl d. Sub-Popul. (K)		2	6
Anzahl d. Variablen (p) (i.d. Gesamtpopul.)		3	22
Anzahl d. 'zugrunde liegenden Dimensionen' (q)		2	10

Die Festlegung der Parameter der Gesamtpopulation bzw. der einzelnen Teilpopulationen wird dem Zufall überlassen.

3.2. Ausscheidung ungeeigneter Datensätze

Bei der Generierung der Daten kann es vorkommen, daß die Mittelwerte von Teilstichproben so nahe beieinander liegen, daß die Gruppen selbst mit einer idealen Clusteranalyse nicht mehr getrennt werden können. Eine vollständige Separierung ist schon dann meist nicht mehr möglich, wenn ein oder mehrere Elemente der einen Gruppe näher beim Centroid einer anderen Gruppe liegen als beim eigenen Gruppen-Centroid. Es hängt weitestgehend vom Zufall und nicht von der Güte des Clusteranalyse-Verfahrens ab, wo solche Elemente letztendlich zugeordnet werden. Damit ist es also naheliegend, nur jene Datensätze zur Auswertung zuzulassen, bei denen eine Trennung der Teilstichproben auch tatsächlich möglich ist. Wir entschlossen uns deshalb dazu, jeder clusteranalytischen Auswertung eine Diskriminanzanalyse vorzuschalten. Diese bestimmt bei bekannter Gruppenzugehörigkeit jedes Elementes lineare Trennfunktionen, mit denen eine 'optimale' Unterscheidung der Gruppen möglich ist. Es läßt sich natürlich nicht ausschließen, daß die lineare Diskriminanzanalyse Gruppen nicht trennen kann, obwohl ein anderer Trennungsalgorithmus (z.B. nichtlineare Diskriminanzfunktionen) dazu in der Lage gewesen wäre. Da im vorliegenden Kontext lediglich sichergestellt werden soll, daß eine Trennung möglich ist, läßt sich auch ein suboptimales Verfahren einsetzen. Datensätze, die sich durch die Diskriminanzanalyse nicht eindeutig in die vorgegebenen Teilstichproben aufspalten ließen, wurden besonders markiert. Auf einen vollständigen Ausschluß dieser Datensätze wurde verzichtet, um Vergleichsmöglichkeiten mit den Ergebnissen von BLASHFIELD zu gewährleisten [3].

3) Auch BLASHFIELD hat bei seinen Datensätzen Diskriminanzanalysen durchgeführt, was bei ihm jedoch keine Konsequenz für weitere Berechnungen hat. Sie dienten ihm allein zur groben Abschätzung der Schwierigkeit der an die Clusteranalysen gestellten Aufgaben

Es sind deshalb bei der Datenanalyse mehrere Durchgänge zu unterscheiden. Während beim ersten alle generierten Datensätze (N = 766) eingingen, wurden im zweiten Durchgang nur solche Datensätze berücksichtigt, bei denen die Teilstichproben durch eine Diskriminanzanalyse vollständig trennbar waren (N = 571). Da eines der Clusteranalyse-Verfahren (EUCLID) strenge Einschränkungen bezüglich der Anzahl der Variablen und des Stichprobenumfangs macht, erwies sich ein dritter Durchgang als notwendig, bei dem nur jene Datensätze mit allen Clusteranalysen konfrontiert wurden, die den Restriktionen des Verfahrens entsprachen (N = 97 ; alle Datensätze sind durch Diskriminanzanalyse trennbar).

3.3. Aufgenommene Clusteranalyse-Verfahren

Wie schon oben angedeutet, schien es sinnvoll zu sein, vor allem solche Algorithmen in die Untersuchung aufzunehmen, die am Heidelberger Universitätsrechenzentrum verfügbar sind. Die Wahl fiel (fast zwangsläufig) auf das dort implementierte umfassende Programm-Paket CLUSTAN 1C (WISHART 1975), das einerseits als relativ benutzerfreundlich bezeichnet werden kann und zum anderen die Mehrzahl der ansonsten noch als Einzelprogramme (stand-alone-Programme) im URZ zusätzlich verfügbaren Algorithmen enthält [4].
Für die Monte-Carlo-Studie wurden deshalb die folgenden 10 Clusteranalyse-Verfahren ausgewählt:

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 Ähnlichkeits-Analyse
9) Prozedur RELOCATE
10) Prozedur EUCLID

---

[4] Das in CLUSTAN nicht enthaltene, im URZ verfügbare Programm der Automatischen Klassifikation arbeitet leider fehlerhaft

Die beiden übrigen prinzipiell verfügbaren CLUSTAN-Prozeduren KDEND (JARDINE-SIBSON-Methode) und DNDRITE (Polythetische divisive Methode) konnten wegen ihres enormen Rechenzeit-Kontingents in diesem Rahmen nicht berücksichtigt werden.

Die ausgewählten Clusteranalyse-Verfahren wurden mit bis zu 766 [5] unterschiedlichen Datensätzen konfrontiert, von denen jeder als Stichprobe aus einer Gesamtpopulation definiert wurde, die sich aus mehreren Teilpopulationen zusammensetzte. Von jedem Stichprobenelement (beispielsweise einer Versuchsperson) war bekannt, aus welcher Teilpopulation es entstammte. Die Aufgabe der einzelnen Clusteranalyse-Algorithmen bestand nun darin, die Herkunft jedes Elements zu 'erraten'.
Vor jeder Clusteranalyse wurden die Daten standardisiert. Dies war vor allem deshalb notwendig, weil das für das vorliegende Datenmodell am angemessensten scheinende Euklid'sche Distanzmaß sich bei Verwendung von Rohwerten unbefriedigend verhält (vgl. EVERITT 1974, S.56f.).
Da bei keiner der untersuchten Methoden die günstigste Clusterzahl automatisch bestimmt wird, legten wir diese Zahl für alle Verfahren so fest, daß sie jeweils der Anzahl der Teilpopulationen entsprach.

3.4. Bestimmung der Güte von Cluster-Lösungen

Eine Clusteranalyse hat dann eine optimale Lösung gefunden, wenn die Cluster genau den einzelnen Stichproben aus den definierten Teilpopulationen entsprechen. Eine solche Lösung ist auch von quasi-perfekten Algorithmen dann nicht erreichbar, wenn sich Teilstichproben sehr ähnlich sind und teilweise überlappen (wenn z.B. Elemente aus der Population a dem Mittelwert der Population b näher liegen als dem von a).
In solchen problematischen bzw. schwierigen Fällen weisen manche Cluster-Verfahren hinsichtlich der Fehlklassifikationen erhebliche Unterschiede auf.
Ein Maß für die Güte von Cluster-Lösungen war mit dem Kappa-Koeffizienten von COHEN (1960) gegeben, durch den die Anzahl

---

[5] Die ungerade und verquere Zahl ergibt sich daraus, daß die Generierung der Daten nach einer vorgegebenen Zeit abgebrochen wurde. Das Gesetz der 'schönen Zahl' schien verzichtbar

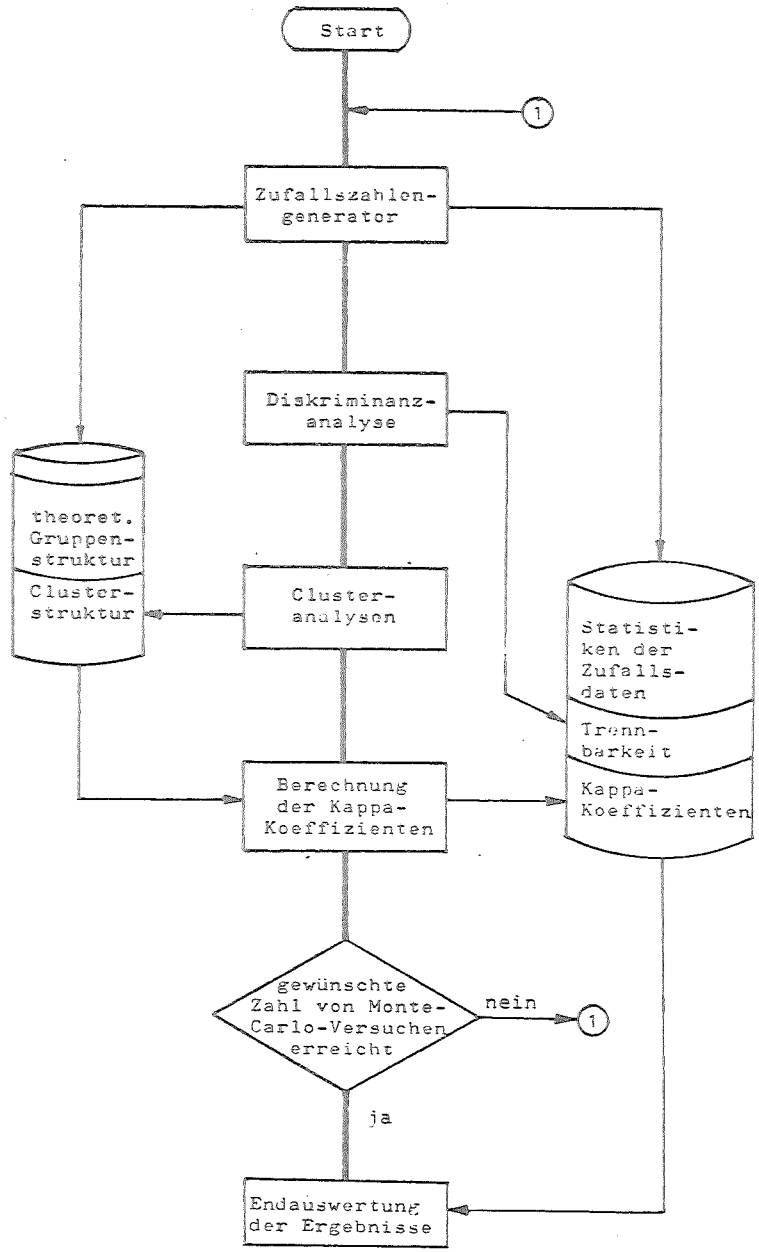
der Fehlklassifikationen bzw. der korrekten Zuteilungen bestimmt werden kann. Seine Werte variieren zwischen 0 und 1, wobei größere Werte bessere Klassifikationen anzeigen.
Um Informationen darüber zu gewinnen, welche Eigenschaften von Populationen oder Stichproben einen Einfluß auf die Höhe des Kappa-Koeffizienten haben, wurden verschiedene Statistiken für die einzelnen erzeugten Datensätze berechnet. Diese Statistiken sollten durch nonparametrische Korrelation zum Kappa-Koeffizienten in Beziehung gesetzt werden.
Bei diesen Statistiken handelte es sich im einzelnen um

a) die Anzahl der Elemente in der Gesamtstichprobe;
b) die mittlere Anzahl von Elementen in den Teilstichproben;
c) die Varianz der Anzahl von Elementen in den Teilstichproben;
d) den Mittelwert der Varianzen der einzelnen Variablen;
e) die Varianz der Varianzen der einzelnen Variablen;
f) die mittlere Anzahl von Hauptkomponenten in den Teilstichproben;
g) die Varianz der Anzahl von Hauptkomponenten;
h) die Anzahl der Variablen und
i) die Anzahl der Cluster (bzw. Teilstichproben).

Es wurde demnach zusätzlich die Abhängigkeit der Cluster-Lösungen in den einzelnen Verfahren von a) der Stichprobengröße, b) der Stichprobenrepräsentativität, c) der spezifischen Stichprobenrelationen, d) dem Überlappungsgrad der einzelnen Stichproben, e) der Tendenz zu sphärischen Clusterbildungen ('Elliptizität' der Populationen), der Komplexität von Kovarianzstrukturen f) innerhalb und g) zwischen den Populationen, h) der Variablenzahl und i) der Teilstichproben-Menge kontrolliert (zur näheren Begründung der Statistiken a) bis g) siehe BLASHFIELD 1976, S.382).

3.5. Zusammenfassende Darstellung der Methode (vgl. Abb. 1)

Es werden zunächst mehrere Zufalls-Datensätze erzeugt, die dem oben beschriebenen Datenmodell (siehe auch Anhang B) folgen. Bei der Generierung eines Datensatzes lassen sich zunächst die Populationsparameter innerhalb bestimmter Grenzen (zufällig) festlegen und anschließend mehrere verschiedene Teilstichproben erzeugen, die zusammen eine Gesamtstichprobe (Datensatz) ergeben. Parallel dazu werden die deskriptiven Statistiken der Population und des Datensatzes berechnet.



Eine für jeden Datensatz vorgenommene Diskriminanzanalyse soll diejenigen Stichproben ermitteln, die nicht eindeutig in die vorgegebenen Teilstichproben separierbar sind. Anschließend wird der jeweilige Datensatz durch die 10 Clusteranalyse-Verfahren ausgewertet. Ein drittes Programm vergleicht die Datenstruktur mit der Clusterstruktur und berechnet den Kappa-Koeffizienten als Maß für die Genauigkeit der von den einzelnen Cluster-Verfahren erstellten Klassifikationen. Diese Schritte werden mehrfach wiederholt, wobei sich für jeden Schritt die Statistiken und Kappa-Koeffizienten registrieren lassen.
In einem abschließenden Schritt werden für jeden Cluster-Algorithmus der Median der Kappa-Koeffizienten und ihre Korrelationen zu den Statistiken berechnet.
Die beschriebene Vorgehensweise wurde programmtechnisch weitestgehend automatisiert.

4. Beschreibung der Ergebnisse

Da die Untersuchung von BLASHFIELD auf einer eher geringen Zahl von Datensätzen aufbaute, waren von vornherein mehr oder weniger große Abweichungen von seinen Ergebnissen zu erwarten. In Tabelle 3 sind die Mittelwerte und Mediane der Kappa-Koeffizienten für die 11 Prozeduren wiedergegeben und den Ergebnissen von BLASHFIELD gegenübergestellt.
Wie schon unter 3.2. erwähnt, basieren die Berechnungen auf unterschiedlich zusammengesetzten Stichproben von Datensätzen, wobei die Gesamtzahl (N_1) 766 beträgt, während die erste reduzierte Größe (N_2 = 571) die über Diskriminanzanalyse ermittelte Zahl eindeutig trennbarer Teilstichproben, die zweite (erheblich) niedrigere Zahl dagegen die Anzahl der mit den Beschränkungen von EUCLID kompatiblen Datensätze wiedergibt.

Tab. 3:
Mediane bzw. Mittelwerte der Kappa-Koeffizienten als Maß für die Güte einzelner Clusteranalysen

Methode	Ergebn. BLASHFIELD N_0=50	Ergebn. d. vorl. Untersuch. N_1=766	N_2=571		N_3=97		
single-linkage	.06	.30	.45	.43	.48	.46	*
complete linkage	.42	.68	.71	.73	.63	.67	**
average linkage	.17	.73	.75	.78	.70	.77	**
Centroid-Methode	-	.32	.38	.32	.45	.42	*
Median-Verfahren	-	.46	.38	.49	.51	.58	*
WARD-Methode	.77	.99	.96	1.	.88	.97	***
LANCE-WILLIAMS I	-	.98	.94	1.	.83	.93	***
LANCE-WILLIAMS II	-	.98	.95	1.	.83	.95	***
McQUITTY-Methode	-	.74	.77	.81	.72	.77	**
RELOCATE	-	.99	.96	1.	.88	.97	***
EUCLID	-	-	-	-	.86	.96	***

Bei den beiden reduzierten Stichproben (N_2 u. N_3) gibt die 1. Zahl den Mittelwert, die 2. den Median wieder

(Die in Tab. 3 ganz zum Schluß eingefügten 'Sterne' sollen eine Grobkategorisierung in schlechte, mittlere und hervorragende Verfahren andeuten).
Es muß erwähnt werden, daß alle Prozeduren bei irgendeinem Datensatz den maximalen Kappa-Koeffizienten von 1 erreichten, was einen deutlichen Unterschied zu den Resultaten von BLASHFIELD darstellt. Am auffälligsten scheinen jedoch die z.T. erheblichen Diskrepanzen im Hinblick auf die durchschnittlichen Kappa-Mediane. Dies überrascht umso mehr, als in der vorliegenden Untersuchung die Vorgehensweise von BLASHFIELD überaus exakt übernommen wurde.
Eine mögliche Erklärung für die Unterschiede sahen wir in der geringen Anzahl von Datensätzen (N=50), die BLASHFIELD in seiner Studie herangezogen hatte. Da bei der Generierung der Datensätze sehr viele Populationsparameter variiert werden, könnten die Abweichungen rein zufälliger Natur sein.

Um einen Überblick über die Variationsbreite der Kappa-Koeffizienten zu gewinnen, wurden aus der Gesamtmenge von Datensätzen 8 Stichproben von je 50 Datensätzen gezogen und getrennt analysiert. Die Ergebnisse sind in Tab. 4 wiedergegeben. Das Verfahren EUCLID wurde wegen seiner Beschränkung auf 10 Variablen und 95 Fälle aus der Analyse ausgeschlossen.
Obwohl die Kappa-Werte - wie zu erwarten - recht deutlich schwanken, ist die Variation nicht so stark ausgeprägt, daß damit die Diskrepanzen zu den Resultaten von BLASHFIELD überzeugend erklärt werden könnten; alle in der vorliegenden Untersuchung erzielten Werte (selbst die Minima) liegen deutlich über den dort angegebenen Resultaten.
Dies weist zunächst nur darauf hin, daß die Zusammensetzung der Daten in beiden Untersuchungen offenbar nicht vergleichbar ist, es zeigt jedoch unserer Auffassung nach, daß grundsätzliche Zweifel an den Ergebnissen von BLASHFIELD geäußert werden müssen. Die mittleren Kappa-Koeffizienten in seiner Studie sind teilweise so niedrig (.06 für single-linkage-, .17 für average-linkage-Methode), daß daraus nur der Schluß gezogen werden kann, die betreffenden Verfahren lieferten kaum bessere Ergebnisse, als es bei einer Zufallszuteilung der Elemente zu einer vorgegebenen Cluster-Zahl zu erwarten

Tab. 4:
Kappa-Koeffizienten der einzelnen Clusteranalyse-Verfahren und zugehörige Schwankungsbreiten für 8 X 50 Monte-Carlo-Studien

Methode	Stichproben mit N=50 Datensätzen								Schwankungsbr.
	1	2	3	4	5	6	7	8	
single-linkage	.37	.43	.40	.31	.41	.46	.38	.28	.18
compl.-linkage	.71	.66	.69	.70	.72	.71	.65	.63	.09
average-link.	.74	.74	.73	.71	.78	.72	.72	.62	.16
Median	.40	.45	.44	.35	.35	.45	.29	.29	.16
Centroid	.43	.49	.49	.41	.47	.53	.40	.42	.13
WARD	.92	.92	.94	.89	.95	.92	.92	.91	.06
LANCE-WILL. I	.90	.88	.89	.87	.92	.89	.87	.91	.05
LANCE-WILL. II	.90	.88	.90	.88	.94	.90	.90	.91	.06
McQUITTY	.76	.70	.68	.70	.75	.77	.67	.61	.16
RELOCATE	.92	.92	.93	.89	.95	.91	.92	.91	.06

Bei LANCE-WILLIAMS I wird beta mit -0.25, bei LANCE-WILLIAMS II mit -.50 angesetzt (vgl. auch Tab. 3); bei allen Verfahren sind jeweils die höchsten und niedrigsten Werte unterstrichen.

wäre. Von daher scheinen die in der vorliegenden Studie gewonnenen Ergebnisse glaubwürdiger und realistischer auszufallen, was auch durch die zusätzliche Untergliederung in 8 Teilstichproben dokumentiert werden kann: die dort festgestellten geringen Schwankungsbreiten weisen darauf hin, daß die Durchschnittswerte der Gesamtstichprobe als stabil und verläßlich zu werten sind.
Die Kappa-Werte der großen Stichproben belegen die relativ schwachen Ergebnisse von single-linkage-, Median- und Centroid-Verfahren sowohl für Mittelwert als auch Median. Von diesen drei schwachen Verfahren hebt sich eine weitere Gruppe mit mittelhohen Kappa-Koeffizienten positiv ab, in die die complete-linkage- und average-linkage-Methode sowie das McQUITTY-Verfahren einzuordnen sind. Als eindeutige Spitzen-Verfahren mit hohen Reproduktionswerten kristallisieren sich dann aber das LANCE-WILLIAMS-Verfahren (in beiden beta-Varian-

ten), die WARD-Methode, RELOCATE und - bei Berücksichtigung der genannten Beschränkungen - auch EUCLID heraus.

Tab. 5:
Rangkorrelationen nach SPEARMAN zwischen den Kappa-Werten der einzelnen Clusteranalysen und den die Daten beschreibenden Statistiken

Methode	Statistiken								
	a	b	c	d	e	f	g	h	i
	-.18	-.20	.04	.00	.10	.18	.21	.27	-.09
single-l.	-.13	-.17	.06	-.13	-.04	-.02	.03	.06	-.05
	-.44	.43	.02	-.17	-.35	-.04	.06		
	-.06	-.03	-.04	.22	.15	.38	.22	.37	-.05
compl.-l.	-.07	-.05	-.09	.12	.01	.23	.00	.18	-.05
	-.01	-.16	-.03	-.49	-.44	.05	.01		
	-.07	-.08	.08	.16	.15	.33	.15	.31	-.04
aver.-l.	-.04	-.08	.05	.07	.02	.20	-.03	.14	-.01
	.09	-.21	.21	-.45	-.27	-.05	.23		
	-.14	-.23	.03	-.21	-.10	-.14	-.16	-.14	-.02
Median	-.13	-.23	.01	-.21	-.11	-.15	-.16	-.16	-.01
	-.12	-.20	.05	-.05	.00	.02	.00	.03	-.01
Centroid	-.12	-.22	.01	-.09	-.04	-.02	-.05	-.03	-.00
	-.17	.02	-.05	.29	.32	.54	.45	.62	-.22
WARD	-.13	.08	-.02	.08	.10	.23	.16	.31	-.21
	-.11	.02	-.25	.47	-.38	.48	.17		
LANCE-	-.16	.06	-.06	.32	.34	.58	.47	.66	-.23
WILL. I	-.11	.12	-.04	.11	.13	.31	.23	.41	-.22
LANCE-	-.14	.06	-.08	.32	.32	.55	.43	.62	-.21
WILL. II	-.09	.13	-.08	.13	.12	.27	.15	.34	-.21
	-.09	-.05	-.01	.17	.16	.35	.21	.36	-.07
McQUITTY	-.10	-.09	.02	.04	-.01	.14	-.02	.12	-.09
	-.18	.01	-.06	.33	.35	.59	.45	.65	-.23
RELOCATE	-.11	.07	-.04	.16	.17	.33	.17	.38	-.20
	-.09	.21	-.12	.12	.05	.34	.23	.32	-.32
EUCLID	-.01	.30	-.17	.09	-.10	.14	-.03	-.05	-.33

Tabelle 5 gibt die Rangkorrelationen wieder, die zwischen den von den Clusteranalysen erzielten Kappa-Werten und den Deskriptivstatistiken [6] berechnet wurden. Pro Verfahren sind maximal drei Korrelations-Reihen dargestellt, wobei sich die erste Zeile auf die Gesamtstichprobe (N_1=766), die zweite Zeile auf die reduzierte Anzahl (N_2=571) und die letzte (falls vorhanden) auf die bei BLASHFIELD angegebenen Werte bezieht.

Auch hier fallen gewisse Diskrepanzen zu den Ergebnissen von BLASHFIELD auf, die wohl in erster Linie auf die unterschiedliche Zusammensetzung der Stichproben zurückzuführen sind. Dennoch muß betont werden, daß auch für die unterschiedlichen Stichproben der vorliegenden Studie gerade bei sehr guten Verfahren erhebliche Abweichungen zu konstatieren sind, die möglicherweise mit der Sensibilität der Statistiken für variierende Aufgabenschwierigkeiten zusammenhängen. Überraschend ist in jedem Falle, daß die Statistik d (Überlappungsgrad der Stichproben) keinerlei Einfluß auf die erzielten Cluster-Lösungen zeigt. Eine mögliche Erklärung liegt im Spezifikum der Datengenerierung: durch die Zufallserzeugung der Varianzen ist damit zu rechnen, daß der Mittelwert der Varianzen bei den einzelnen Stichproben etwa gleich groß ausfällt.

6) Legende zu den Deskriptivstatistiken:
Statistik a: Gesamtzahl der Elemente eines Datensatzes (Mixtur aus mehreren Teilstichproben)
" " b: Mittlere Anzahl von Elementen pro Teilstichp.
" " c: Varianz der Anzahl von Elementen der Teilstichproben
" " d: Mittlere Varianz aller Variablen
" " e: Varianz der Varianzen aller Variablen
" " f: Durchschnittliche Zahl von Hauptkomponenten pro Teilstichprobe
" " g: Varianz der Anzahl von Hauptkomponenten in den einzelnen Teilstichproben
" " h: Anzahl der Variablen
" " i: Anzahl der Teilstichproben (Gruppen, Cluster)

5. Diskussion

In einer abschließenden Betrachtung soll versucht werden, die einzelnen Verfahren nach ihren Ergebnissen in den Monte-Carlo-Studien unter Einbezug der Resultate von BLASHFIELD kritisch zu beurteilen, um Hinweise für das forschungspraktische Vorgehen ableiten zu können.

Einzelbeurteilung der verschiedenen Verfahren

(a) single-linkage-Methode

Es fällt zunächst auf, daß für diese Technik in der vorliegenden Simulationsstudie durchgängig bedeutend höhere Kappa-Werte als bei BLASHFIELD erzielt wurden, was allerdings wenig an der prinzipiellen (relativen) Einordnung des Verfahrens ändert. Die schon bei BLASHFIELD geäußerte Vermutung, daß sich die Prozedur wohl weniger gut zur Klassifikation von Objekten eignet, demgegenüber aber wegen ihrer Tendenz zur Kettenbildung möglicherweise zur Identifikation von sog. 'outliers' herangezogen werden kann, wird auch in der vorliegenden Untersuchung bestätigt. Die negative Korrelation mit der Anzahl der Fälle (Statistik a) begrenzt das Spektrum der vorstellbaren Anwendungsfälle auf Versuchsanordnungen mit geringen Stichprobenquoten, so daß insgesamt festgehalten werden muß, daß die Einordnung der Prozedur in die Gruppe der weniger günstig abschneidenden Techniken sie als Klassifikationsverfahren ungeeignet erscheinen läßt.

(b) complete-linkage-Verfahren

Auch bei dieser Technik fallen die gegenüber BLASHFIELD weitaus höheren Kappa-Werte auf, denen zufolge eine Einordnung der Clusteranalyse in die Gruppe der Verfahren mittlerer Güte möglich schien.
Weiterhin bleibt erwähnenswert, daß die theoretisch zu erwartende klar negative Korrelation der Prozedur mit Statistik c (und d) ähnlich wie bei BLASHFIELD nicht registriert werden konnte. Während für Statistik d keine nennenswerten korrelativen Beziehungen zu entdecken waren, zeichnete sich für Statistik c lediglich eine leichte Tendenz in der vorhergesagten Richtung ab, ohne daß die Effekte als bedeutsam bezeichnet werden können. Demnach scheint die in der Literatur geäußerte Vermutung, daß dieses Verfahren zur Generierung gleichgroßer Cluster neige, den Fakten nicht zu entsprechen. Insgesamt gesehen lassen sich kaum stichhaltige Argumente für die forschungspraktische Favorisierung der Prozedur finden; von ihrem Einsatz kann in der Regel abgesehen werden.

(c) average-linkage-Verfahren

Entgegen den überraschenden Ergebnissen bei BLASHFIELD schnitt diese Prozedur gegenüber den beiden vorher genannten besser ab, ein Befund, der nicht nur theoretisch zu erwarten war, sondern auch schon in anderen Untersuchungen empirische Bestätigung erfahren hatte. Ebenfalls nicht kompatibel mit den Ergebnissen von BLASHFIELD war das Faktum, daß keine empfindliche Reaktion des Verfahrens (nachweisbar in einer hohen negativen Korrelation) bei starkem Überlappungsgrad der Cluster (Statistik d) ausgemacht werden konnte, wie die Prozedur auch insgesamt nicht die erwarteten korrelativen Beziehungen zu den übrigen berechneten Statistiken aufwies. Die resultierenden Kappa-Werte können als insgesamt befriedigend gewertet werden, lassen jedoch kaum eine andere Einordnung als in die der Verfahren mittlerer Güte zu, was den Rückgriff auf dieses Verfahren kaum sinnvoll erscheinen läßt.

(d) Centroid-Methode

Dieses Verfahren schneidet bei der Überprüfung durch die Monte-Carlo-Studien insgesamt gesehen am schlechtesten ab. Es ist besonders auffallend, daß die Kappa-Werte um so mehr absinken, je größer die Cluster ausfallen (negative Korrelation mit Statistik b). Weiterhin wird hier (wie auch bei den beiden übrigen schlechten Clusteranalyse-Verfahren) die Güte der Cluster-Lösung entscheidend von der Varianz der Variablen (Statistik d) beeinflußt; es bedarf wohl keiner besonderen Erwähnung mehr, daß von dem Einsatz dieser Prozedur dringend abgeraten werden muß.

(e) Median-Verfahren

Die Ergebnisse zu dieser Technik entsprechen der theoretisch fundierten Erwartung (vgl. Anhang A), die im Hinblick auf die Centroid-Methode hier günstigere Werte nahelegt. Die Ver-

besserung fällt jedoch so geringfügig aus, daß das Verfahren dennoch in die Gruppe der schlechtesten Prozeduren eingeordnet werden muß. Ähnlich wie die Centroid-Methode reagiert auch das Median-Verfahren sensibel auf die vorfindbare Clustergröße, was in der negativen Korrelation mit Statistik b zum Ausdruck kommt.

(f) WARD-Methode

Der hohe Durchschnittswert für Kappa zeigt an, daß das WARD-Verfahren (wie schon bei BLASHFIELD demonstriert) eindeutig zur Spitzengruppe der überprüften Prozeduren gezählt werden muß. Es gilt auch hier, daß die Ergebnisse der vorliegenden Untersuchung gegenüber denen von BLASHFIELD deutlich höher liegen, was auf einen systematischen Effekt schließen läßt. Widersprüchliche Befunde liegen zu Statistik e (Varianz der Varianzen) vor: die bei BLASHFIELD erwartete und bestätigte Sensibilität der Technik für diese Statistik (die einen 'bias' zur Generierung sphärischer Cluster nahelegen soll) war mit den vorliegenden Daten nicht nachweisbar.
Demgegenüber blieb die bei BLASHFIELD aufgezeigte positive Korrelation mit Statistik f (Anzahl der Hauptkomponenten) auch für unsere Untersuchung charakteristisch: die Methode arbeitet demnach besser, wenn die Kovarianzstruktur der Daten sehr komplex ausfällt. Insgesamt gesehen erscheint das Verfahren als außerordentlich robust und kann zur Anwendung unbedingt empfohlen werden.

(g) LANCE-WILLIAMS-Verfahren

Dieses in zwei Varianten überprüfte (ebenfalls hierarchische) Clusteranalyse-Verfahren schneidet nur geringfügig schlechter als die WARD-Methode ab und gehört demnach ebenfalls zur Gruppe der Spitzen-Prozeduren.
Am Rande interessant scheint die Tatsache, daß die von LANCE u. WILLIAMS vorgeschlagene Standardversion mit einem beta von -.25 der von uns probeweise aufgenommenen Alternativ-Version mit beta = -.50 insgesamt leicht unterlegen war.
Für die Standardversion gilt zusätzlich, daß sie umso günstiger abschneidet, je unterschiedlicher die Komplexität der Faktorenstruktur in den Clustern ausfällt (positive Korrelation mit Statistik g). Es versteht sich von selbst, daß auch

diese Prozedur als unbedingt empfehlenswert charakterisiert werden muß.

(h) McQUITTY-Verfahren

Bei dieser Technik ließen sich keine großen Besonderheiten registrieren; sie schnitt mittelmäßig ab und zeigte keine auffälligen Affinitäten zu den ausgewählten Statistiken.
Die resultierenden Kappa-Werte fallen im Hinblick auf eine Empfehlung des Verfahrens für forschungspraktische Belange insgesamt gesehen zu niedrig aus.

(i) RELOCATE

Für dieses Verfahren kann die Beurteilung der WARD-Methode übernommen werden; es fällt außerordentlich schwer, Kriterien zu finden, die eine eindeutige Entscheidung für eines der beiden Verfahren beinhalten würden.

(h) EUCLID

Auch für diese Prozedur muß betont werden, daß sie sich prinzipiell für die forschungspraktische Verwendung eignet, wenn auch ihre Limitationen hinsichtlich Variablen- und Versuchspersonenanzahl ihre Attraktivität für größere Forschungsprojekte mindert.
Der erzielte durchschnittliche Kappa-Wert legt es jedenfalls nahe, dieses Verfahren zusammen mit der WARD-Methode, den LANCE-WILLIAMS-Verfahren und RELOCATE in die Gruppe der leistungsstärksten Techniken einzuordnen, die vorbehaltlos für die unterschiedlichsten Untersuchungsfragestellungen einsetzbar scheinen.

Abschließend bleibt festzuhalten, daß eine konsequente Fortführung der hier geschilderten Auswertungsstrategie als nächsten Schritt die Konfrontation dieser bewährten Verfahren mit anderen aus der Literatur bekannten und nicht im CLUSTAN-Paket enthaltenen Prozeduren beinhalten müßte, da so der Kanon wirklich relevanter Klassifikationsverfahren relativ zweifelsfrei bestimmt werden kann.

6. Zusammenfassung

Die inflationäre Entwicklung von Clusteranalyse-Algorithmen macht den Einsatz evaluierender Operationen notwendig, um die wirklich leistungsfähigen Prozeduren ermitteln zu können.

Ein Überblick über die bisher geleisteten Vergleichsstudien deutete auf die methodische Problematik und damit auf die mangelnde Beweiskraft von Plasmodenstudien sowie Untersuchungen anhand empirischer Datensätze hin, während bei den zur Bewertung unterschiedlicher Klassifikationsverfahren besser geeigneten Monte-Carlo-Studien im wesentlichen nur die Untersuchung von BLASHFIELD aufgrund der inhaltlichen und formalen Vorgehensweise beachtenswerte Ergebnisse liefern konnte.

Die in dieser Studie verbliebenen Schwachpunkte (sehr niedrige Zahl von Monte-Carlo-Durchläufen, nur wenige überprüfte Verfahren) legten eine Replikation der Untersuchung nahe, die gerade die beschriebenen Mängel vermeiden sollte. Trotz der engen Anlehnung an das von BLASHFIELD vorgegebene Untersuchungsdesign ließen sich in der Folgestudie nicht alle der dort erzielten Resultate bestätigen, obwohl in wichtigen Punkten Übereinstimmung erzielt werden konnte.

Von den insgesamt 10 untersuchten Prozeduren aus dem CLUSTAN-Programmpaket erreichten immerhin 4 Techniken (WARD-Methode, LANCE-WILLIAMS-Verfahren, RELOCATE und EUCLID) so überzeugende Ergebnisse, daß sie unbedingt zur Anwendung empfohlen werden können.

L i t e r a t u r

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

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

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

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

BOCK, H.H.: Automatische Klassifikation. Göttingen: Vandenhoeck & Ruprecht 1974

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

CATTELL, R.B., COULTER, M.A. & TSUYIOKA, B.: The taxonometric recognition of types and functional emergents. In: R.B. CATTELL (Ed.), Handbook of multivariate experimental psychology. Chicago: R. McNally 1966, 288-329

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

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

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

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

GOWER, J.C.: A comparison of some methods of cluster analysis. Biometrics, 1967, 23, 623-637

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

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

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

HUBERT, L.: Kappa revisited. Psychological Bulletin, 1977, 84, 289-297

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

LORR, M. & RADHAKRISHNAN, B.K.: A comparison of two methods of cluster analysis. Educational and Psychological Measurement, 1967, 27, 47-53

McQUITTY, L.L.: Multiple clusters, types and dimensions from iterative intercolumnar correlational analysis. Multivariate Behavioral Research, 1968, 3, 465-477

McRAE, D.J.: MICKA, a FORTAN IV iterative K-means cluster analysis program. Behavioral Science, 1971, 16, 423-424

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

ROLLETT, B. & BARTRAM, M. (Hrsg.): Einführung in die hierarchische Clusteranalyse. Stuttgart: Klett 1976

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

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, 1977, 19, 201-217

WARD, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 1963, 58, 236-244

WISHART, D.: An algorithm for hierarchical classifications. Biometrics, 1969, 25, 165-170

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

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

Anhang A

Übersicht und nähere Erläuterungen zu den verschiedenen Clusteranalyse-Verfahren

Nach EVERITT (1974, S.7ff.) 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 Fusionen 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 Anfangspartitionen korrigiert, die Elemente demnach re-allokiert werden können, und zum anderen die gewünschte Klassenanzahl vom Untersucher a priori festzulegen ist.

3) Dichte-Verfahren (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 (Klumpen) 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

Die folgende Darstellung verzichtet aus Verständlichkeitsgründen auf eine formale Beschreibung der Algorithmen; stattdessen sollen die Vorgehensweisen der einzelnen Techniken rein inhaltlich skizziert werden, was notwendigerweise zu Übersimplifikationen führt.
Detaillierte formale Darstellungen finden sich bei BOCK 1974 und VOGEL 1975.

ad 1)

(a) single-linkage-Methode:

Diese Technik gehört zu den agglomerativen hierarchischen Verfahren. Es werden in jedem Schritt die beiden Elemente, die sich am nächsten (ähnlichsten) sind, zu einer gemeinsamen Klasse fusioniert. Die Distanz zwischen zwei Gruppen ist definiert durch den Abstand jener zwei Elemente aus den beiden Gruppen, die am dichtesten beieinander liegen.
Der beschriebene Programmablauf kann sich schrittweise solange fortsetzen, bis schließlich alle Elemente in einem einzigen Cluster erfaßt sind. Es besteht jedoch für den Benutzer die Möglichkeit, den Prozeß dann zu stoppen, wenn ein starker Abfall (Diskontinuität) im Fusionskoeffizienten zu beobachten ist.

(b) complete-linkage-Methode:

Sie stellt im Prinzip das genaue Gegenteil der single-link-Technik dar. Bei dieser agglomerativen hierarchischen Clusteranalyse wird die Distanz zwischen den Gruppen als die Entfernung zwischen dem am weitesten auseinanderliegenden Individuenpaar definiert (ansonsten Prozedur wie bei (a)).

(c) average-linkage-Methode:

Auch diese Technik gehört zu den agglomerativen hierarchischen Verfahren; sie stellt einen Kompromiß zwischen den

beiden vorher beschriebenen Ansätzen dar; Distanzen zwischen den Gruppen werden als arithmetisches Mittel der Distanzen aller Individuenpaare definiert, die aus Elementen jeweils verschiedener Gruppenzugehörigkeit gebildet werden können (ansonsten Prozedur wie bei (a) und (b)).

(d) Centroid-Methode:

Bei diesem hierarchischen agglomerativen Verfahren wird unterstellt, daß die Gruppen als Punkte im Euklidschen Raum repräsentiert sind. Die Distanz zwischen den Gruppen läßt sich als die Distanz zwischen den Gruppen-Centroiden definieren. Das Programm fusioniert zunächst die beiden Gruppen mit den geringsten Distanzen und verfährt so schrittweise bis zur letzten Hierarchie-Ebene, in der alle Elemente zu einem einzigen Cluster fusioniert sind.

(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 Linie 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 Gruppen sehr nahe an dem des größeren Clusters liegt und die charakteristischen Eigenschaften der kleinen Gruppen damit verloren gehen. Dieses Problem bestent bei der Median-Methode nicht, da das Verfahren unabhängig von der Cluster-Größe arbeitet.

(f) Clusteranalyse nach WARD:

Dieses ebenfalls agglomerative Verfahren faßt Elemente bzw. Individuen (iterativ) derart zu neuen Clustern zusammen, daß die Varianz innerhalb der neu entstehenden Gruppen ein Minimum ergibt. Wie bei den Verfahren (a) bis (d) kann die am besten zu interpretierende Hierarchiestufe aus dem Verlauf eines Koeffizienten erschlossen werden: der Fehlerwert S

steigt dann nur geringfügig an, wenn nahe beieinanderliegende Elemente zusammengeschlossen werden, erhöht sich jedoch sprunghaft, wenn einem Cluster entfernter liegende Elemente zugeordnet werden.

(g) LANCE-WILLIAMS' flexible_beta:
Dieses Verfahren kann nicht mehr geometrisch interpretiert werden. Es baut auf der Tatsache auf, daß alle bisher genannten hierarchischen Prozeduren mit einer einzigen Rekursionsformel beschrieben werden können und sie sich nur im Hinblick auf die konstanten Parameter in dieser Formel unterscheiden. Bei allen Verfahren sind nun drei Kennwerte a priori festgelegt, während ein weiterer Parameter (beta) variiert werden kann. LANCE & WILLIAMS schlugen vor, den Wert von beta auf -0,25 festzusetzen. In diesem Fall verhält sich die Technik im wesentlichen wie der WARDsche Algorithmus.

(h) Die Methode nach McQUITTY:
Das Verfahren entspricht dem von LANCE & WILLIAMS, wenn beta = 0 gesetzt ist.

ad 2)

(a) Optimierungstechnik nach McRAE:
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. Es werden zunächst k Punkte in einem p-dimensionalen Raum als Anfangsschätzungen der Clusterzentren herausgesucht, dann neue benachbarte Elemente hinzugefügt und immer wieder Neuberechnungen der Clusterzentren angestellt, um die geeignetste Anfangskonfiguration zu finden. Im weiteren Verlauf ist es jederzeit möglich, Re-Allokationen der einzelnen Elemente vorzunehmen: sie werden anderen Clustern zugeordnet, wenn dadurch das Optimierungskriterium (Minimierung der Spur von W) verbessert

werden kann.
Der Untersucher gibt die maximale Anzahl der Gruppen a priori an, wobei sich die Lösungen für jeden Maximalwert danach bewerten lassen, wie gut das Optimierungskriterium erfüllt werden konnte, bzw. welche Cluster-Anzahl damit also zu bevorzugen ist.

(b) Optimierungstechnik nach WISHART (RELOCATE):
Das Verfahren RELOCATE von WISHART (im CLUSTAN-Programmsystem enthalten) zeigt große Ähnlichkeit mit der Prozedur von McRAE. Das Programm beginnt mit einer willkürlichen Einteilung in eine vorgegebene Anzahl von Subgruppen, wobei als Startkonfiguration entweder eine Zufallseinteilung oder das Ergebnis einer vorangehenden Clusteranalyse gewählt werden kann. Die Anfangskonfiguration wird schrittweise so zu verbessern versucht, daß Cluster optimaler Homogenität resultieren (s.o.). Das Programm RELOCATE kann zusätzlich dazu veranlaßt werden, die Anzahl der Cluster mehrfach jeweils um eins zu verringern. Dabei wird nicht mehr von einer willkürlichen Anfangsaufteilung ausgegangen, sondern es werden in Analogie zu den agglomerativen hierarchischen Verfahren jene beiden Cluster zu einem neuen verbunden, die sich am ähnlichsten sind. Danach wird erneut versucht, die Klassifikation durch Re-Allokierung von Elementen zu verbessern.

ad 3)

(a) Mode analysis nach WISHART:
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.7).

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.

ad 4)

(a) Das Taxonome-Programm (TAXO) von CATTELL & COULTER:

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 mehrerer 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 Elemen-

te einander ähnlich seir 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.115f. und 157f.).

ad 5)

(a) Die Konfigurationsfrequenzanalyse (KFA):
Das Verfahren analysiert eine endliche Zahl von Variablenmustern (Konfigurationen) für kategoriale Daten (bei n dichotomen Variablen sind beispielsweise 2 hoch 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 Grenzwerts (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 wenigsten 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 Quader (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 fortsetzt, 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.

Nachtrag

ad 2)

(c) Prozedur EUCLID:
Das hier angesprochene Verfahren verwendet die CAUCHY-Methode, um die Quadratsumme S der Euklidschen 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 Summen der 'Wahrscheinlichkeiten' (fractions) = 1 ist. Durch eine Exponential-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-Allokation von einigen Punkten eine weitere Reduktion von S möglich wird (Verwendung des Prinzips von RELOCATE).
Programmbeschränkungen: Max. Anzahl der Fälle = 95, max. Anzahl der Variablen = 10.

(Für dieses neue Verfahren stand den Verfassern lediglich eine vorläufige und knappe Beschreibung zur Verfügung, die keine näheren Ausführungen zuließ).

Anhang B

Mathematische Grundlagen der Monte-Carlo-Studie

1. Datenmodell:

Verteilungsdichte der Gesamtpopulation:

$$f(X) = \sum_{i=1}^{k} (n_i/n) f(X_i)$$

k = Anzahl der Subpopulationen

$$n = \sum_{i=1}^{k} n_i$$

n_i = Anzahl der Entitäten in Subpopulation i

p = Anzahl der Variablen

X = n x p Datenmatrix der Gesamtpopulation

X_i = n_i x p Datenmatrix der Subpopulationen

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_k \end{pmatrix}$$

Verteilungsform der Daten in einer Subpopulation i : Multinormalverteilung mit den Populationsparametern μ_i und Σ_i :

$$F(X_i) = N(\mu_i, \Sigma_i)$$

μ_i = Vektor der Erwartungswerte der p Variablen

Σ_i = p x p Kovarianzmatrix

Die Kovarianzmatrix ist festgelegt durch die Varianzen und die Korrelationsmatrix R der p Variablen:

$\Sigma = SSRSS'$

S = p x p Diagonalmatrix der Standardabweichungen aller Variablen

Die Anzahl q der (positiven) Eigenwerte (= Anzahl der Hauptkomponenten) ist kleiner oder gleich der Variablenanzahl:

$q \leq p$

Die Enddaten (X^*) seien meßfehler-behaftet.
Die Zufallsvariable X^*_{ij} aus der Subpopulation i sei gleichverteilt innerhalb fester Grenzen:

$$X_{ij} - 0{,}6\sigma^2_{ij} \leq X^*_{ij} \leq X_{ij} + 0{,}6\sigma^2_{ij}$$

σ^2_{ij} = Varianz der Variablen j in Subpopulation i.

2. Gewinnung der Daten:

2.1. Generierung der (Sub-) Populationsparameter:

k, n_i, p, q

μ_i = Vektor der p Erwartungswerte in Subpopulation i

σ^2_i = Vektor der p Varianzen in Subpopulation i

R_i = p x p Korrelationsmatrix in Subpopulation i

Es gibt verschiedene Möglichkeiten, auf Grundlage der oben genannten Parameter korrelierte Zufallszahlen (die Korrelation ist durch R festgelegt) zu generieren.
Alle uns bekannten Algorithmen setzen jedoch voraus, daß Σ_i positiv-definit ist. Bei der Erprobung der Algorithmen stellte sich jedoch heraus, daß es nahezu unmöglich war, Korrelationsmatrizen größer als 3 x 3 mit der gewünschten Verteilung der Korrelationskoeffizienten zu generieren, die dieser Forderung (positive Definitheit) entsprechen konnten.
Der unten beschriebene Algorithmus erlaubt es, sehr einfach die Korrelationsmatrix so zu korrigieren, daß sie den Voraussetzungen gerecht wird.

a) Generierung einer beliebigen Korrelationsmatrix R_i (R_i kann auch singulär sein)

b) Faktorisierung der Korrelationsmatrix R_i (zur Vereinfachung werden künftig die Indices weggelassen):

$R = L\Lambda L'$

hierbei bedeutet

Λ → Diagonalmatrix der Eigenwerte von R
L → Matrix der Eigenvektoren von R
L' → Transponierte von L

c) Reduktion der Anzahl von Eigenwerten und Eigenvektoren um die Anzahl der Eigenwerte, die kleiner/gleich 0 sind, bzw. auf die gewünschte Anzahl (q) von Hauptkomponenten. Die reduzierte Eigenwert-Diagonalmatrix sei Λ^*, und die reduzierte Eigenvektormatrix L^*.

d) Bestimmung einer Faktormatrix (Ladungsmatrix)

$A = L^*\Lambda^{*1/2}$

e) Die bereinigte Korrelationsmatrix errechnet sich aus der Faktormatrix (die Korrelationsmatrix wird allerdings für die weiteren Berechnungen nicht mehr benötigt):

$$R^* = AA'$$

2.2. Generierung der Stichproben:

a) Generierung einer $n_i \times q$ Matrix Y
Die Elemente seien Zufallsziehungen aus einer normalverteilten Population mit dem Erwartungswert 0 und der Varianz 1.

b) Generierung einer $n_i \times p$ Matrix F (Fehlermatrix), deren Elemente gleichverteilt zwischen -0,6 und +0,6 liegen.

c) Berechnung der Datenmatrix X_i

$$X_i = YA'S + M + FS$$

wobei S = p x p Diagonalmatrix der Standardabweichungen aller p Variablen und
M = n_i x p Matrix des Mittelwert-Vektors

$$M = \begin{pmatrix} \mu_i \\ \mu_i \\ \mu_i \\ \cdot \\ \cdot \\ \mu_i \end{pmatrix}$$

3. Die Güte der Cluster-Lösung (Kappa-Koeffizient):

Bei der beschriebenen Monte-Carlo-Untersuchung wird den Clusteranalyse-Verfahren bei jedem Durchgang je ein "Gemisch"

aus mehreren Stichproben vorgegeben, die aus (Teil-) Populationen mit hinsichtlich verschiedener Parameter unterschiedlichen Ausprägungen stammen. Aufgabe des Verfahrens ist es, diese Stichproben zu trennen und damit die theoretische Stichprobenstruktur zu reproduzieren.
Eine Clusteranalyse hat dann eine optimale Lösung gefunden, wenn die Anzahl der Cluster gleich der Anzahl der vorgegebenen Teilstichproben ist und wenn je einer Stichprobe je ein Cluster genau entspricht, d.h. wenn sie sich aus denselben Elementen zusammensetzen.
Der Kappa-Koeffizient ist nun ein Maß, das Abweichungen von diesem Idealfall (Kappa = 1) anzeigt. Je geringer die Cluster-Lösung mit der theoretischen Gruppenstruktur übereinstimmt, desto kleiner fällt der Wert von Kappa aus.
Clusteranalyse-Verfahren liefern keine direkten Hinweise dafür, welches gefundene Cluster welcher vorgegebenen Teilstichprobe entspricht. Man kann jedoch davon ausgehen, daß ein Cluster eine Reproduktion einer Stichprobe darstellt, wenn beide sehr viele Elemente gemeinsam haben. Es gilt also, jeder Stichprobe genau ein Cluster zuzuordnen, so daß die Übereinstimmung zwischen allen Stichproben und den zugeordneten Clustern maximal wird. Als Kriterium für eine optimale Zuordnung wählten wir die Gesamtzahl der korrekten Klassifikationen.
Dies soll anhand einer Kreuztabelle veranschaulicht werden:

Cluster-Nr.

Teilstichprobe Nr.	1	2	.....	j	.....	k		
1	n_{11}	n_{12}	. . . . .			n_{1k}	$n_{1.}$	$= \sum_{j=1}^{k}$
2	n_{21}	n_{22}	. . . . .				$n_{2.}$	
.	.	.	. . . . .				. .	
i	.	.	. . . . .	n_{ij}	. . . .		$n_{i.}$	
.	.	.	. . . . .				. .	
k	n_{k1}	.	. . . . .			n_{kk}	$n_{k.}$	
	$n_{.1}$	$n_{.2}$	. . . . .			$n_{.k}$		

$$n_{.j} = \sum_{i=1}^{k} n_{ij}$$

Die beste Übereinstimmung zwischen Clustern und Teilstichproben ist dann gegeben, wenn gilt

$$R_O = \sum_{i=1}^{k} n_{ii} \rightarrow \text{Maximum}$$

Dieses Optimum wurde durch die systematische Vertauschung der Zeilen der Klassifikationsmatrix ermittelt.

Der Kappa-Koeffizient errechnet sich dann folgendermaßen:

$$\kappa = (P_o - P_e) / (1 - P_e)$$

wobei gilt:

$$P_o = R_o/n$$

$$n = \sum_{ij}^{kk} n_{ij}$$

$$P_e = R_e/n$$

$$R_e = \xi(R_o) = 1/n \sum_{i=1}^{k} n_{i.} n_{.i}$$