

Numerische Gruppierung und graphische Darstellung von Daten: Ein Methodenvergleich

G. Ohmayer, H. Seiler

Zusammenfassung

Die Fett-, Eiweiß-, Lactose- und Aschegehalte der Milchen von 50 Lebewesen dienten als Daten für eine vergleichende Beschreibung der Methoden Hauptkomponentenanalyse, Biplot, nichtlineare Abbildung, Vernetzungsdiagramm, Dendrogramm, Austauschverfahren und Bestimmung unscharfer Gruppen. Es wird gezeigt, wie graphische Verfahren sowohl die Aufbereitung von Gruppierungsergebnissen als auch zur optischen Wiedergabe der wesentlichsten Datenstrukturen herangezogen werden können. Möglichkeiten zur Abschätzung der Güte einer Gruppierung werden diskutiert.

Summary

Composition of the milk constituents fat, protein, lactose and ash from 50 mammals were used as data for comparing principal component analysis, biplot, non-linear mapping, linkage-maps, dendograms, relocation-techniques and fuzzy partitions. The usefulness of applying graphical methods is demonstrated not only for giving a visual representation of the most important aspects of the data, but also for determining group structures within the data. The possibility of estimating the validity of a particular grouping is discussed.

1. Einleitung

Die Vielfalt der heute bekannten Gruppierungsmethoden, die unter den Begriffen «Clusteranalyse», «automatische Klassifikation» oder «numerische Taxonomie» geführt werden, ist kaum noch zu überblicken. Ständig werden neue Verfahren, zumeist Modifikationen der bekannten Methoden, vorgestellt. In der vorliegenden Arbeit soll deshalb ein Überblick über einige grundsätzliche Gruppierungsmethoden und deren wesentlichste Charakteristika gegeben werden. Auf eine ausführliche Darstellung aller mathematischen Algorithmen wird verzichtet. Vielmehr sollen anhand eines vom Umfang her einfachen, strukturell aber hinreichend komplexen Beispiels die Möglichkeiten und Probleme bei der Anwendung der Methoden in den Vordergrund gestellt werden. Besondere Berücksichtigung erfährt die graphische Darstellung der ermittelten Strukturen.

2. Beschreibung der Daten

Als Demonstrationsbeispiel wurde ein bewußt einfaches, allgemeinverständliches Datenmaterial verwendet (Tabelle 1,

S. 66). Die zu gruppierenden Objekte sind Milchen von Land- und Meeressäugetieren, Primaten sowie *Homo sapiens*. Als Gruppierungskriterien dienten die vier charakteristischen Merkmale Fett, Gesamteiweiß, Lactose und Asche der Muttermilchen. Es sei betont, daß sich alle Resultate hinsichtlich der Ähnlichkeitsbeziehungen dieser Lebewesen ausschließlich auf deren Milchzusammensetzung beziehen. Phylogenetische Verwandtschaftsbeziehungen können nur mit äußerster Vorsicht abgeleitet werden, da die Zusammensetzung der Milch keine konservative Eigenschaft ist.

Die Merkmalswerte sind als «weiche Daten» zu betrachten. Zum einen sind sie Datenerhebungen mehrerer Autoren entnommen, was sie nur beschränkt vergleichbar macht (JENNESS, 1974). Außerdem geben sie, da es sich um durchschnittliche Werte handelt, nicht die mit Sicherheit vorhandene unterschiedliche Varianz einzelner Populationen wieder. Erschwerend kommt hinzu, daß die Milchzusammensetzung in Abhängigkeit von Ernährung, Jahreszeit, Alter der Tiere, Wurfgröße, Laktationszyklus usw. erheblichen Schwankungen unterworfen ist. Trotz dieser Einschränkungen eignen sich die Daten sehr gut für den durchgeführten Methodenvergleich. Bei der sachlichen Interpretation der Ergebnisse soll jedoch die Unschärfe der Ausgangsdaten Berücksichtigung finden.

3. Abbildungsverfahren

Die in der englischsprachigen Literatur unter den Namen «Mapping» bzw. «Ordination-methods» bekannten Verfahren bilden multivariate Daten in einem zweidimensionalen Diagramm ab (EVERITT, 1978). Ziel dieser Verfahren ist es, die einzelnen Beobachtungen so auf Punkte in der Ebene zu verteilen, daß deren tatsächliche Struktur im mehrdimensionalen Raum möglichst gut wiedergegeben wird. Ist die Abbildungsgüte hinreichend, d. h., hält sich der Informationsverlust durch die Datenreduktion in Grenzen, sind solche Abbildungen eine gute Ergänzung zu den rechnerischen Ergebnissen einer Clusteranalyse, da eventuell vorhandene Gruppenstrukturen leicht zu erkennen sind. Man unterscheidet lineare Abbildungsverfahren oder Projektionstechniken, wie zum Beispiel Hauptkomponentenanalyse und Biplot, und nichtlineare Abbildungen (Nonlinear mapping).

3.1 Hauptkomponentenanalyse und Biplot

Ziel der Hauptkomponentenanalyse ist die Bestimmung der wichtigsten Richtungen im Datenraum. Dazu werden die

Tabelle 1. Daten des Demonstrationsbeispiels: Prozentuale Milchzusammensetzung verschiedener Lebewesen

Lebewesen	Fett	Eiweiß	Lactose	Asche
Herrentiere				
Menschen				
1 Mensch	3,8	1,0	7,0	0,2
Menschaffen				
2 Orang-Utan	3,5	1,5	6,0	0,2
3 Schimpanse	3,7	1,2	7,0	0,2
Meerkatzenartige				
4 Zwergmeerkatze	2,9	2,1	7,2	0,3
5 Pavian	5,0	1,6	7,3	0,3
Kralleaffen				
6 Tamarin	3,1	3,8	5,8	0,4
Unpaarhufer				
Pferdeartige				
7 Esel	1,4	2,0	7,4	0,5
8 Hausspferd	1,9	2,5	6,2	0,5
9 Wildpferd	2,2	2,0	6,1	0,4
10 Zebra	2,1	2,3	8,3	0,4
Paarhufer				
Schweine				
11 Wildschwein	6,8	4,8	5,5	1,7
Kamele				
12 Lama	2,4	7,3	6,0	0,5
13 Kamel	5,4	3,9	5,1	0,7
14 Dromedar	4,5	3,6	5,0	0,7
Hirsche				
15 Sikahirsch	19,0	12,4	3,4	1,4
16 Rothirsch	19,7	10,6	2,6	1,4
17 Ren	20,0	9,5	2,6	1,4
Hornträger, Unterfamilie Gazellenartige				
18 Edmigazelle	19,0	12,4	3,3	1,5
19 Thompson Gazelle	19,6	10,5	2,7	1,4
20 Schwarzfersenantilope	20,4	10,8	2,4	1,4
Hornträger, Unterfamilie Rinder				
21 Hausrind	3,7	3,4	4,8	0,7
22 Zebu	4,7	3,2	4,9	0,7
23 Yak	6,5	5,8	4,6	0,9
24 Wasserbüffel	7,4	3,8	4,8	0,8
25 Bison	3,5	4,5	5,1	0,8
Hornträger, Unterfamilie Ziegenartige				
26 Moschusochse	5,4	5,3	4,1	1,1
27 Hausziege	4,5	2,9	4,1	0,8
28 Hausschaf	7,4	5,5	4,8	1,0
Fleischfresser				
Hundeartige				
29 Haushund	12,9	7,9	3,1	1,2
30 Wolf	9,6	9,2	3,4	1,2
31 Kojote	10,7	9,9	3,0	0,9
32 Schakal	10,5	10,0	3,0	1,2
33 Afrik. Wildhund	9,5	9,3	3,5	1,3
Großbären				
34 Schwarzbär	24,5	14,5	0,4	1,8
35 Grizzly Bär	22,3	11,1	0,6	1,5
36 Braunbär	22,6	7,9	2,1	1,4
37 Eisbär	33,1	10,9	0,3	1,4
Ohrerrobben				
38 Nördlicher Seebär	53,3	8,9	0,1	0,5
Nagetiere				
Biber				
39 Biber	11,7	8,1	2,6	1,1
Wühler				
40 Goldhamster	4,9	9,4	4,9	1,4
Mäuse				
41 Wanderratte	10,3	8,4	2,6	1,3
42 Hausmaus	13,1	9,0	3,0	1,3
Hasentiere				
Hasenartige				
43 Hauskaninchen	18,3	13,9	2,1	1,8
44 Florida-Waldkaninchen	13,9	23,7	1,7	1,5
45 Wildkaninchen	17,9	12,5	1,0	2,0
Wale				
Furchenwale				
46 Blauwal	42,3	10,9	1,3	1,4
47 Finnwal	32,4	17,8	0,3	1,0
48 Buckelwal	33,0	12,5	1,1	1,6
Delphine				
49 Delphin	33,0	6,8	1,1	0,7
Insektenfresser				
Igel				
50 Braunbrustigel	10,1	7,2	2,0	2,3

ursprünglichen Merkmale x_1, \dots, x_p ersetzt durch die Hauptkomponenten y_1, \dots, y_r ($r \leq p$), die folgende Eigenschaften besitzen:

a) Die Hauptkomponenten sind Linearkombinationen:

$$y_j = \sum_{k=1}^p a_{jk} x_k \quad (j = 1, \dots, r)$$

b) Sie sind unkorreliert: $r_{yy} = 0$ für $j \neq j'$

c) Sie stellen die Richtungen maximaler Varianzerklärung dar; d. h. in Richtung von y_1 variieren die Daten insgesamt am stärksten, y_2 erklärt im zu y_1 senkrechten Unterraum maximale Varianz, usw.

Dies bedeutet, daß die Projektion der Beobachtungen in die durch die beiden ersten Hauptkomponenten aufgespannte Ebene im „Kleinst-Quadrate-Sinn“ die bestmögliche lineare Abbildung mit planarer Darstellung liefert. Die Berechnung der Hauptkomponenten erfolgt über die Bestimmung der Eigenwerte und Eigenvektoren der Varianz-/Kovarianzmatrix bzw. der Korrelationsmatrix. Während die Eigenvektoren die Koeffizienten der Linearkombination liefern, d. h. die Richtung der Hauptkomponenten bestimmen, kann an den Eigenwerten λ_j die Höhe der Varianzerklärung einzelner Hauptkomponenten abgelesen werden.

Für die Approximationsgüte der zweidimensionalen Darstellung gilt:

$$F = (\lambda_1 + \lambda_2) / \sum_{j=1}^r \lambda_j$$

Im Vergleich zur Hauptkomponentenanalyse werden mit der Biplot-Methode außer den Beobachtungen auch die Merkmale sowie deren Beziehungen zueinander graphisch dargestellt (GABRIEL, 1971). Dieses Verfahren beruht auf folgender Zerlegung der Datenmatrix $X_{n,p}$, welche bedeutet, daß sich jeder Beobachtungswert x_{ik} (ite Beobachtung am kten Merkmal) als Skalarprodukt zwischen g_i (iter Zeilenvektor von G) und h_k (kter Zeilenvektor von H) darstellen läßt:

$$X_{n,p} = G_{n,r} \cdot H_{p,r}^T \Leftrightarrow x_{ik} = g_i^T \cdot h_k \quad (i = 1, \dots, n) \quad (k = 1, \dots, p)$$

Falls nun r – der sog. Rang der Datenmatrix X – gleich 2 ist, lassen sich die Daten ohne Informationsverlust in einem zweidimensionalen Diagramm durch die $n + p$ Vektoren g_i und h_k als Biplot darstellen. Andernfalls ($r > 2$), und dies dürfte in praktischen Anwendungen die Regel sein, wird X durch eine Matrix \tilde{X} , die den Rang 2 besitzt, approximiert und deren Biplot gezeichnet. Dabei kann wiederum ein Maß für die Approximationsgüte berechnet und damit die Relevanz des erzeugten Biplots abgeschätzt werden.

Hinsichtlich der Berechnung der oben erwähnten Zerlegung, die über eine SVD (Singular value decomposition) durchgeführt werden kann, sei auf spezielle Literatur verwiesen (GABRIEL, 1971; GABRIEL et al., 1976).

Zur Interpretation eines Biplots sind noch folgende Punkte zu beachten:

a) die Länge der Vektoren h_k entsprechen den Varianzen ($\|h_k\|^2 \sim s_k^2$), die Winkel zwischen diesen Vektoren den Korrelationen zwischen den Merkmalen ($r_{kk'} \sim \cos \angle h_k h_{k'}$).

b) Die Mahalanobis-Distanzen zwischen den Beobachtungen (wirkliche Abstände unter Berücksichtigung von Korrelationen) werden approximiert durch die Distanzen der Punkte im Biplot.

Folgende kritische Anmerkung zur Anwendung der Hauptkomponentenanalyse und Biplotmethode im Rahmen von

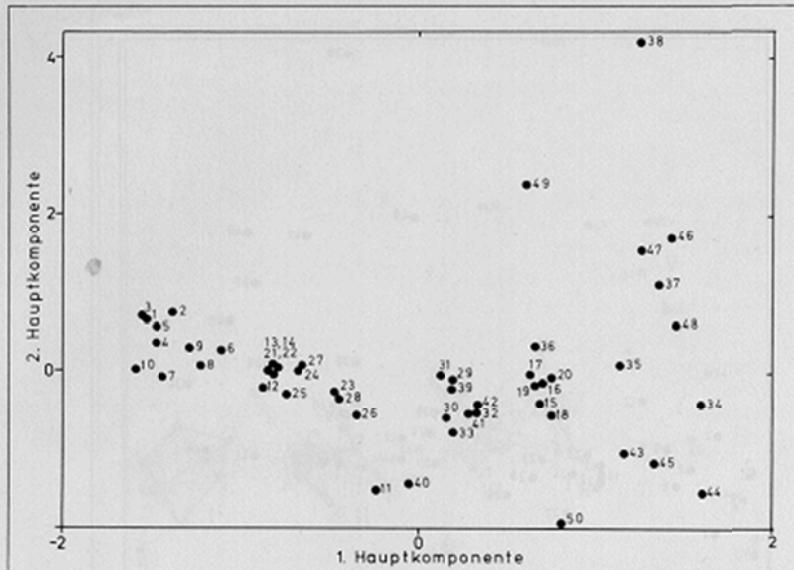


Abb. 1. Abbildung der Ähnlichkeitsbeziehungen von Milchen verschiedener Lebewesen mittels Hauptkomponentenanalyse (Approximationsgüte 92%).

Gruppierungsproblemen sollte berücksichtigt werden: Diese Methoden wurden für den Fall entwickelt, daß homogene Daten im Sinn einer Stichprobe aus nur einer Population vorliegen. Genau dies kann jedoch bei Gruppierungsproblemen nicht unterstellt werden, denn der Anwender vermutet ja gerade die Herkunft der Daten aus verschiedenartigen, allerdings unbekannten Teilpopulationen. Dieser Effekt wird sich im Einzelfall um so negativer auswirken, je mehr die aus allen Daten geschätzte Varianz-/Kovarianzmatrix S – wichtigster Ausgangsparameter der Methoden – von den Varianz-/Kovarianzmatrizen Σ_l ($l = 1, \dots, m$) der m Populationen abweicht. Falls Gleichheit der Matrizen Σ_l unterstellt werden kann ($\Sigma_1 = \Sigma_2, \dots, \Sigma_m = \Sigma$), müßte eine Schätzung von Σ in den Verfahren verwendet werden; diese Schätzung könnte die sog. „Varianz-/Kovarianzmatrix innerhalb der Gruppen“ sein, die nach einer Gruppierung der Daten bestimmt werden kann.

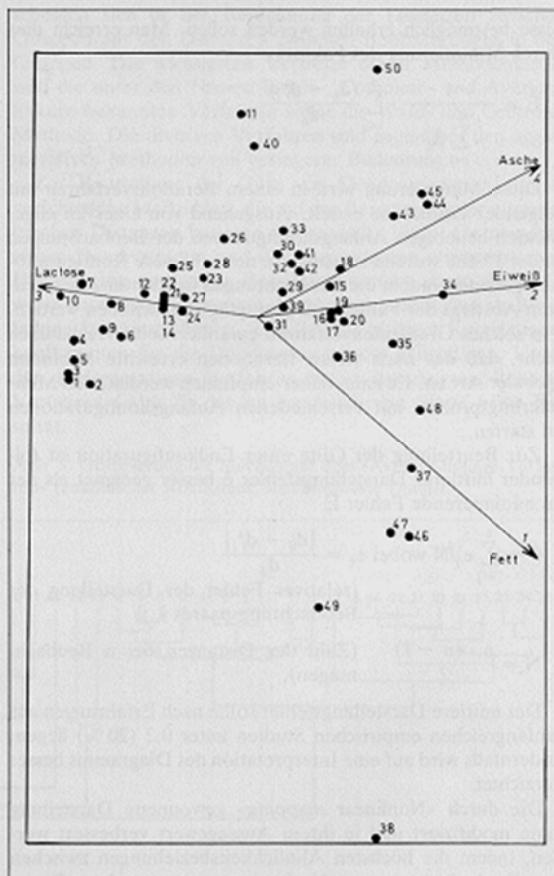
Die Abbildungen 1 und 2 zeigen für unser Anwendungsbeispiel die Ergebnisse der Hauptkomponentenanalyse und des Biplotverfahrens.

Auffallend ist, daß bezüglich der Position der Lebewesen in den beiden Diagrammen kaum Unterschiede bestehen. Im Biplot – hier allerdings berechnet an den auf die Merkmalsvarianzen 1 standardisierten Daten – sieht man aber noch zusätzlich die Wirkung der Merkmale. Es wird deutlich, daß Trockensubstanz und Fett sowie Gesamteiweiß und Asche positiv korreliert sind (Winkel klein), während das Merkmal Lactose zu den erstgenannten eine negative Korrelation aufweist (Winkel zwischen 90° und 180°).

3.2 Nichtlineare Abbildung und Vernetzungsdiagramm

SAMMON beschrieb 1969 die „Nonlinear mapping“ Methode. Das Verfahren setzt Schätzwerte d_{ij} für die Unähnlichkeit zweier Beobachtungen i und j voraus. Welches aus der Palette verfügbarer Distanzmaße adäquat ist, hängt vom aktuellen Anwendungsfall ab (BOCK, 1974). Ziel der nichtlinearen Abbildung ist es, alle Beobachtungen in einem zweidimensionalen Diagramm so zu positionieren, daß deren euklidische Abstände im Diagramm d_{ij}^* möglichst gut mit den Distanzen d_{ij} übereinstimmen. Das bedeutet, daß die Abstandsverhält-

Abb. 2. Abbildung der Ähnlichkeitsbeziehungen mit der Biplot-Methode (Darstellungsgüte = 89,2 %).



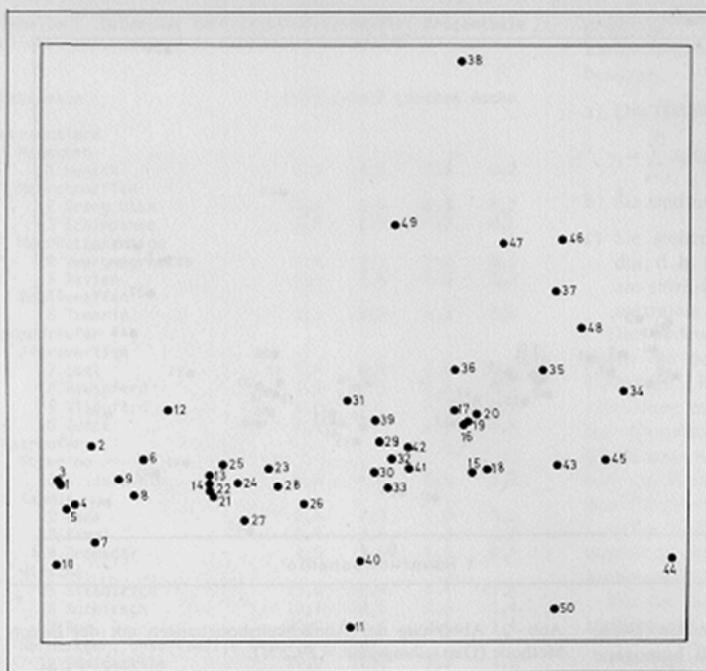


Abb. 3. Darstellung der Milchen mit Hilfe der «nichtlinearen Abbildung» (Darstellungsfehler = 5,3 %).

nisse bestmöglich erhalten werden sollen. Man erreicht dies durch Minimierung des Faktors

$$E = \frac{1}{\sum_{i < j}^n d_{ij}} \cdot \sum_{i < j}^n \frac{(d_{ij} - d_{ij}^*)^2}{d_{ij}}$$

Diese Minimierung wird in einem Iterationsverfahren mit folgender Grundidee erzielt. Ausgehend von einer im allgemeinen beliebigen Anfangskonfiguration der Beobachtungen in der Ebene werden schrittweise neue, bessere Konfigurationen gebildet, indem die Beobachtungen in Richtung des steilsten Abstiegs der Funktion E geringfügig verschoben werden. Ein solches Gradientenverfahren garantiert selbstverständlich nicht, daß das nach vielen Iterationen erreichte Minimum globaler Art ist. Es kann daher empfohlen werden, den Minimierungsprozess mit verschiedenen Anfangskonfigurationen zu starten.

Zur Beurteilung der Güte einer Endkonfiguration ist folgender mittlerer Darstellungsfehler \bar{e} besser geeignet als der zu minimierende Fehler E

$$\bar{e} = \sum_{i < j}^n e_{ij}/N \text{ wobei } e_{ij} = \frac{|d_{ij} - d_{ij}^*|}{d_{ij}}$$

(relativer Fehler der Darstellung des Beobachtungspaares i, j)

$$N = \frac{n \cdot (n - 1)}{2} \quad (\text{Zahl der Distanzen bei } n \text{ Beobachtungen}).$$

Der mittlere Darstellungsfehler sollte nach Erfahrungen aus umfangreichen empirischen Studien unter 0,2 (20 %) liegen; andernfalls wird auf eine Interpretation des Diagramms besser verzichtet.

Die durch «Nonlinear mapping» gewonnene Darstellung kann modifiziert und in ihrem Aussagewert verbessert werden, indem die höchsten Ähnlichkeitsbeziehungen zwischen den Beobachtungen graphisch eingetragen werden. Dieser

Vorschlag geht zurück auf BUSSE (1970) und wurde von OHMAYER et al. (1980) im Zusammenhang mit dem Verfahren «Nonlinear mapping» aufgegriffen. Die Darstellungen wurden Vernetzungsdiagramme oder «Linkage maps» genannt. Voraussetzung ist dabei die Vorgabe einer sinnvollen Schichtung der Distanzen, d. h. die Einteilung der Distanzen in Klassen durch Festlegung sog. Vernetzungsniveaus. Durch Zuordnung einer Linierungsart zu jeder Klasse können die entsprechenden Vernetzungen eingezeichnet werden; Gruppen ähnlicher Beobachtungen werden im «Linkage map» durch hohen Vernetzungsgrad deutlich.

Die Abbildungen 3 und 4 zeigen nichtlineare Abbildung und Vernetzungsdiagramm für die Milchdaten. Der mittlere Darstellungsfehler von 5,3 % ist hinreichend niedrig und bedeutet, daß die Ähnlichkeiten der Lebewesen hinsichtlich der Milchzusammensetzung im Mittel gut durch die Abstände im Diagramm wiedergegeben werden.

4. Gruppierungsverfahren

Die besprochenen Abbildungsverfahren projizieren die Objekte nach verschiedenen Algorithmen in die Ebene, ohne daß eine Entscheidung über Gruppierungen getroffen wird. Die vielfach unter dem Sammelbegriff «Clusteranalyse» bekannten Verfahren gehen einen Schritt weiter und ermitteln als Ergebnis Cluster, d. h. Gruppen von Beobachtungen. Ziel dieser Gruppenbildung ist, daß Beobachtungen innerhalb einer Gruppe möglichst ähnlich (Homogenitätsforderung), Beobachtungen in verschiedenen Gruppen dagegen möglichst unähnlich (Isolationsforderung) sind. Die vielen verfügbaren Methoden (BOCK, 1974; EVERITT, 1974) unterscheiden sich

- durch verschiedene Gewichtung von Homogenität und Isolation
- durch unterschiedliche Gruppierungsstrategien (agglomerative, divisive, iterative, graphentheoretische u.a. Verfahren)

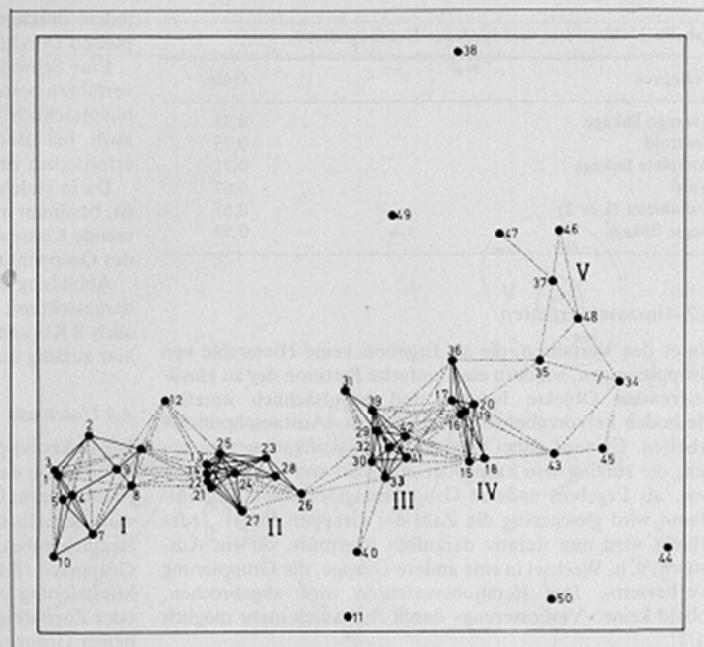


Abb. 4. Vernetzungsdiagramm (Vernetzungsneuauus — $0 \leq d_{ij} \leq 0,15$; $0,15 < d_{ij} \leq 0,25$).

- durch verschiedene Gruppierungsformen (disjunktive, hierarchische, unscharfe Gruppierungen)
- durch unterschiedliche A-priori-Vorgaben (Gruppenanzahl bekannt, Verteilungsannahmen u. a.).

An dieser Stelle sollen nur drei Typen von Methoden kurz besprochen und auf den Beispieldatensatz angewendet werden.

4.1 Hierarchische Gruppierungsmethoden

Die hierarchischen Verfahren sind innerhalb der Clusteranalyse die bekanntesten und am häufigsten verwendeten Methoden. Sie liefern als Ergebnis nicht nur eine, sondern eine ganze Hierarchie von Gruppierungen, die graphisch als Dendrogramm dargestellt werden kann. Man unterscheidet aufgrund des verschiedenen Konstruktionsprinzips agglomerative und divisive Verfahren. Während die ersten durch schrittweise Union von ähnlichen Beobachtungen bzw. Gruppen immer größere Gruppen bilden, operieren letztere durch schrittweise Spaltung der jeweils inhomogenen Gruppe in umgekehrter Richtung. Die einzelnen agglomerativen Verfahren unter-

scheiden sich in der Berechnung der Distanzen zwischen Gruppen aus den Distanzen zwischen Beobachtungen in den Gruppen. Die wichtigsten Vertreter dieser Methodenklasse sind die unter den Namen Single-, Complete- und Average-linkage bekannten Verfahren sowie die Ward- und Centroid-Methode. Die divisiven Verfahren sind gegenüber den agglomerativen Methoden von geringerer Bedeutung.

Zur Beurteilung der Güte von Dendrogrammen gibt es verschiedene Maßzahlen, die auf der Berechnung der ultrametrischen Distanzen basieren (OHMAYER, 1982). Da zwischen einem Dendrogramm und der zugeordneten ultrametrischen Distanz eine eindeutige Beziehung besteht, kann beispielsweise deren Korrelation mit den Ausgangsdistanzen als Beurteilungskriterium dienen. Tabelle 2 zeigt diese Korrelationskoeffizienten für die Dendrogramme der Milchdaten. Unter den agglomerativen Verfahren liefert das Average-linkage-Verfahren (Abb. 5) das im beschriebenen Sinne beste Resultat.

Abb. 5. Gruppierung der Milchen mit dem «Average-linkage-Verfahren» (euklidischer Koeffizient, standardisierte Daten).

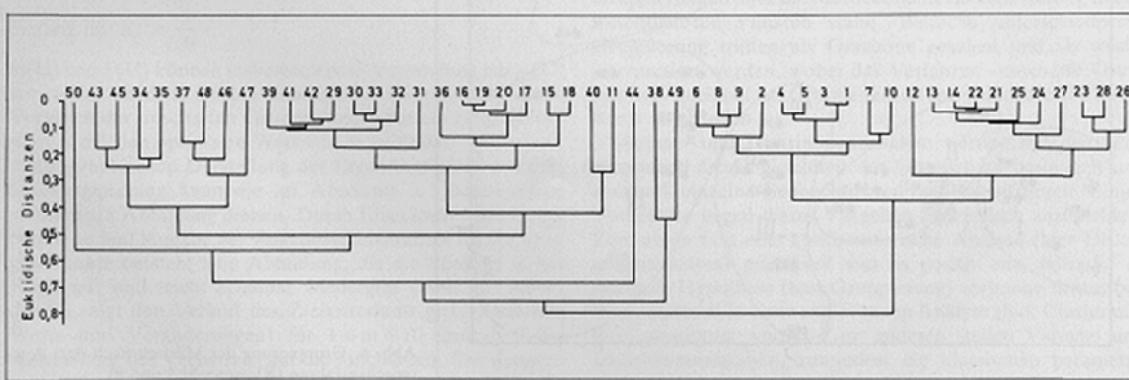


Tabelle 2. Gütwerte verschiedener Dendrogramme.

Verfahren	Güte
Average linkage	0,75
Centroid	0,75
Complete linkage	0,71
Ward	0,67
K-diameter ($k = 2$)	0,67
Single linkage	0,59

4.2 Austauschverfahren

Unter den Verfahren, die als Ergebnis keine Hierarchie von Gruppierungen, sondern eine einfache Partition der zu klassifizierenden Objekte liefern, sind hauptsächlich iterative Methoden hervorzuheben, die nach dem «Austauschprinzip» arbeiten. Es wird dabei eine Anfangsklassifikation vorgegeben, die zufällig sein kann oder aufgrund von Vernetzungen bzw. als Ergebnis anderer Gruppierungsverfahren entsteht. Damit wird gleichzeitig die Zahl der Gruppen fixiert. Jedes Objekt wird nun iterativ daraufhin überprüft, ob ein Austausch, d. h. Wechsel in eine andere Gruppe, die Gruppierung «verbessert». Das Iterationsverfahren wird abgebrochen, sobald keine «Verbesserung» durch Austausch mehr möglich ist.

Die bekannten Verfahren unterscheiden sich insbesondere in der Definition der zu optimierenden «Güte einer Gruppe», wobei im wesentlichen das Varianz-, Determinanten- und Spur-Kriterium zu erwähnen sind. Die einfachste Möglichkeit ist jedoch die Anwendung der «Minimaldistanzregel», nach der ein Austausch immer dann vorzunehmen ist, wenn ein Objekt zum Mittelpunkt einer anderen Gruppe eine geringere Distanz hat als zur eigenen Gruppe. Unterschiede ergeben sich auch durch die Wahl des Zeitpunktes für die Neuberechnung der Gruppenmittelpunkte; diese erfolgt entweder nach

jedem durchgeföhrten Objektausch oder nach jedem kompletten Durchlauf.

Eine Schwierigkeit bei der Anwendung solcher Austauschverfahren besteht in der Vorgabe einer Anfangsklassifikation, hauptsächlich aber in der Festlegung der Gruppenzahl, die auch bei Bildung einer zufälligen Ausgangsklassifikation erforderlich ist.

Da in vielen Fällen die «richtige» Gruppenzahl unbekannt ist, bestimmt man üblicherweise die Gruppierungen für zunehmende Klassenzahlen und entscheidet sich mit Hilfe der Werte des Gruppenkriteriums für eine Klassenzahl.

Abbildung 6 zeigt die Gruppierung der Milchen – graphisch dargestellt mit Hilfe der nichtlinearen Abbildung – mit 5 sowie auch 8 Klassen, da sich letztere als nächstbeste Gruppierung hier zufällig durch Aufspaltung von Gruppen ergibt.

4.3 Unscharfe Gruppierungen

Die bisher besprochenen Gruppierungsmethoden geben keine Information darüber, welche Beobachtungen im Zentrum, am Rand oder im Übergangsbereich zwischen Gruppen liegen. Es wird deshalb im folgenden ein unter den Anwendern noch ziemlich unbekanntes Verfahren zur Bestimmung «unscharfer Gruppen» (Fuzzy sets) beschrieben (BOCK, 1979). Durch Minimierung eines Zielkriteriums werden iterativ Gewichts- oder Zugehörigkeitsindizes u_{ik} berechnet, die bei der vorgegebenen Gruppenzahl m für jede Beobachtung i die Höhe der Zugehörigkeit zur Gruppe k prozentual angeben. D. h., es muß für die Matrix $U = (u_{ik})$, welche unscharfe Gruppierung oder unscharfe Partition genannt wird, gelten:

$$0 \leq u_{ik} \leq 1 \text{ für } i = 1, \dots, n \text{ und } \sum_{k=1}^m u_{ik} = 1$$

Durch Vorgabe einer Schranke (z. B. $u_{ik} \geq 0,7$) können Kernpunkte von Gruppen bzw. Rand- und Übergangsbereiche ermittelt werden. Das üblicherweise verwendete Kriterium, welches minimiert wird, ist:

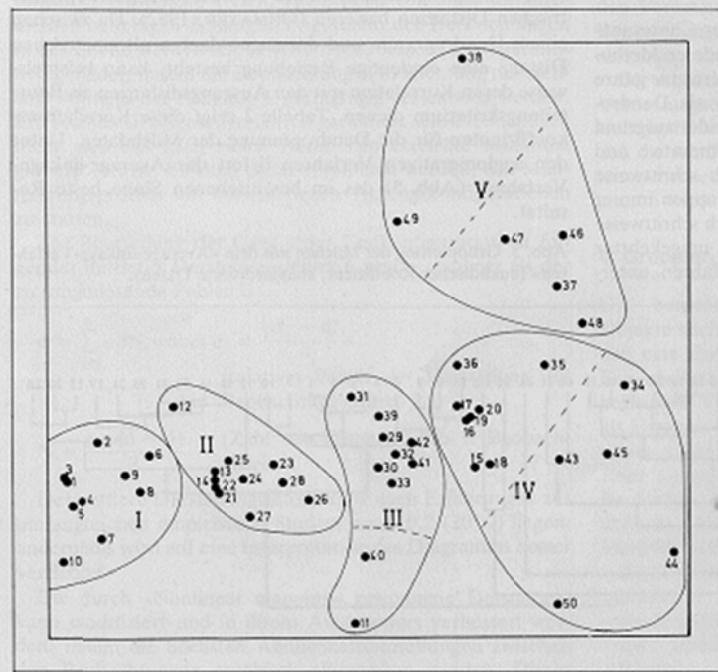
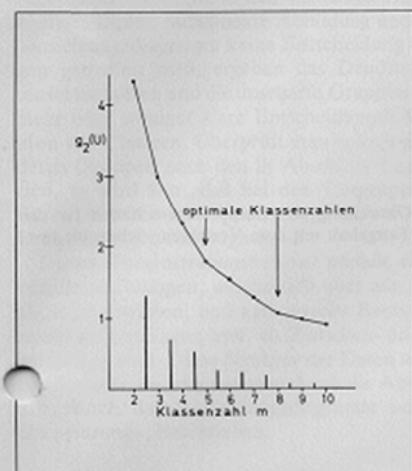


Abb. 6. Gruppierung der Milchen nach dem Austauschverfahren (Klassenzahl 5 und 8).

Abb. 7 (rechts). Darstellung der «unscharfen Gruppierung» für fünf Klassen (Schranken: $u_{ik} \geq 0,8; 0,6; 0,4$).

Abb. 8 (unten). Zielfunktionsverlauf in Abhängigkeit der Gruppenzahl bei «unscharfen Gruppierungen».



$$g_r(U) = \sum_{k=1}^m \sum_{i=1}^n u_{ik}^r \cdot \|x_i - \bar{x}_k\|^2 \text{ mit } \bar{x}_k = \sum_{i=1}^n u_{ik}^r x_i / \sum_{i=1}^n u_{ik}^r$$

Dabei kann über den Parameter r die „Umschärfe“ der Gruppierung gesteuert werden; denn die Wahl von $r = 1$ führt noch zu einer „scharfen“ Partition, d. h. zum Ergebnis $u_{ik} = 0$ oder $= 1$. Erst Werte $r > 1$ liefern unscharfe Gruppierungen. Empfohlen wird die Verwendung der Werte $r = 2$ oder $r = 3$.

Um anzugeben, wie unscharf eine Partition ist, werden das Entropiemaß $H(U)$ bzw. das Fuzzy-Maß $F(U)$ vorgeschlagen:

$$H(U) = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^m u_{ik} \log u_{ik} / \log m,$$

$$F(U) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^m u_{ik}^2$$

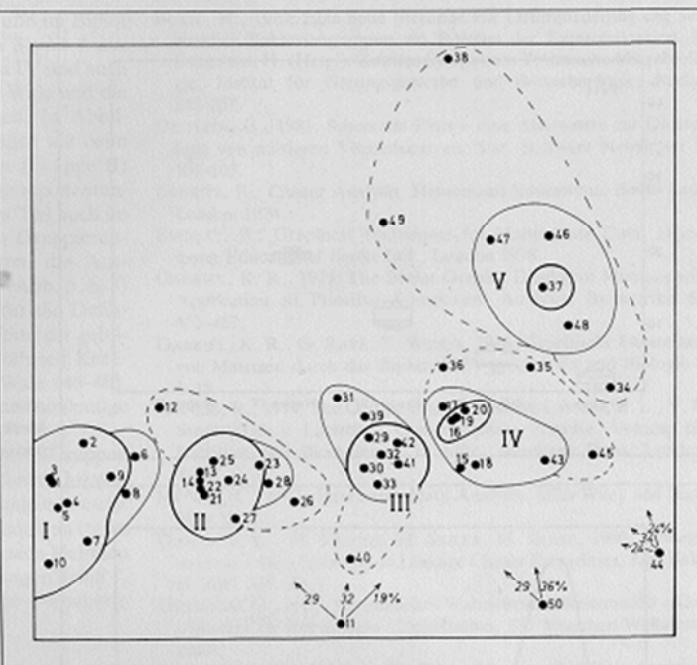
Beide Maße liefern Werte im Intervall $[0,1]$, sind allerdings im folgenden Sinn gegenläufig:

$H(U) = 0 \Leftrightarrow F(U) = 1 \Leftrightarrow U$ ist eine gewöhnliche, d. h. scharfe Gruppierung

$H(U) = 1 \Leftrightarrow F(U) = 0 \Leftrightarrow U$ ist die unschärfste Gruppierung mit $u_{ik} = \frac{1}{m}$

$H(U)$ und $F(U)$ können insbesondere in Verbindung mit $g_r(U)$ verwendet werden, um bei unbekannter Gruppenzahl durch Vergleich der unscharfen Gruppierungen für ein Intervall $m^- \leq m \leq m^+$ den optimalen Wert von m zu finden.

Zur graphischen Darstellung der Ergebnisse einer unscharfen Gruppierung kann die im Abschnitt 3.2 beschriebene nichtlineare Abbildung dienen. Durch Einzeichnen der Gruppenkerne und Angabe der Zugehörigkeitsindizes für die übrigen Punkte entsteht eine Abbildung, die die Struktur in der Regel gut und leicht erfassbar wiedergibt (Abb. 7). Abbildung 8 zeigt den Verlauf des Zielfunktionsmaßes $g_2(U)$ (absolute Werte und Veränderungen) für $1 \leq m \leq 10$ und $r = 2$ bei Anwendung des Verfahrens auf die Milchdaten. Die Klassenzahlen 5 und 8 erweisen sich auch hier als optimal, da die



Funktion $g_2(U)$ deutliche Knickstellen bei diesen Werten zeigt.

5. Darstellung von Gruppierungsergebnissen

Die meisten Gruppierungsverfahren bilden in jedem Fall, selbst bei homogenem Datenmaterial, Gruppen, was leicht zu Fehlinterpretationen führen kann. Es ist deshalb ratsam, die Gruppierungsergebnisse kritisch zu prüfen. Folgende Maßnahmen bieten sich an, um zufällige und somit falsche Gruppierungen aufzudecken:

- Einsatz mehrerer, möglichst verschiedener Gruppierungsverfahren und Vergleich der Ergebnisse
- Prüfung mittels varianz- und diskriminanzanalytischer Methoden
- Graphische Darstellung der Gruppen-Eigenschaften.

Liegen echte Gruppenstrukturen vor, sind in der Regel die Gruppierungen über die Methodenunterschiede hinweg in den wesentlichsten Punkten stabil. Bereiche unterschiedlicher Gruppierung sollten als Grauzone gesehen und als solche interpretiert werden, wobei das Verfahren «unscharfe Gruppierung» geeignet ist, Gruppenkerne und Randbereiche quantitativ zu erfassen.

Varianz- und Diskriminanzanalyse werden mit dem Ziel eingesetzt, die Ungleichheit der Gruppen zu bestätigen und eventuell einzelne Beobachtungen umzuklassifizieren. Einige Vorbehalte gegen dieses Vorgehen sind jedoch anzumelden. Zum einen fällt eine konfirmatorische Analyse (hier Diskriminanzanalyse) prinzipiell viel zu positiv aus, falls die zu prüfende Hypothese (hier Gruppierung) vorher an demselben Datenmaterial in einer explorativen Analyse (hier Clusteranalyse) gewonnen wurde. Zum anderen stellen Varianz- und Diskriminanzanalyse, zumindest die klassischen parametrischen Verfahren, gewisse Anforderungen an die Daten bezüg-

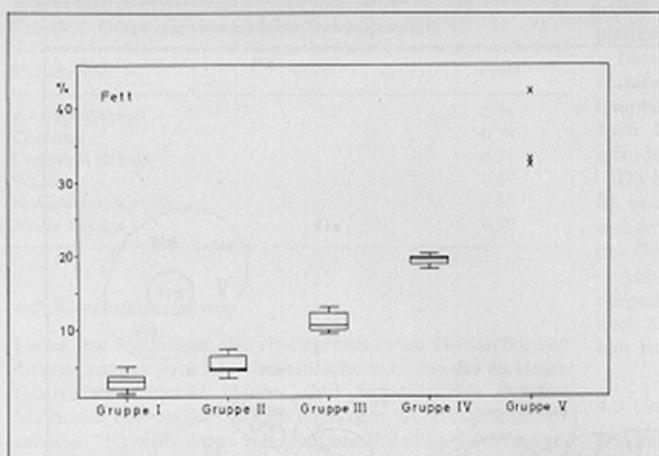


Abb. 9. Darstellung der Gruppeneigenschaften für das Merkmal Fettgehalt mit dem Verfahren «Schematic box-plots».

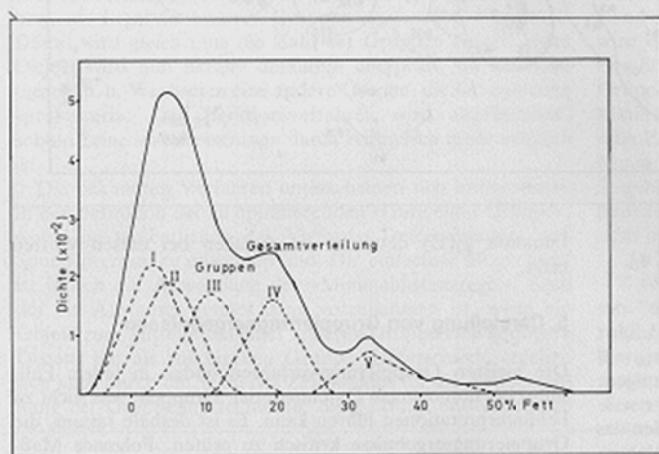


Abb. 10. Darstellung der Gruppeneigenschaften für das Merkmal Fettgehalt mit dem Verfahren «nichtparametrische Dichteschätzung».

lich Normalverteilung, Homoskedastizität etc., was in vielen Fällen sicherlich nicht gegeben ist.

Die graphische Darstellung der Gruppen-Eigenschaften ist in jedem Fall ein sinnvolles Hilfsmittel zur Interpretation der Gruppen und zur Charakterisierung ihrer Unterschiede. Exemplarisch sollen an dieser Stelle zwei Möglichkeiten erläutert werden, die geeignet sind, pro Merkmal die Gruppen-Eigenschaften zu erfassen. Betrachten wir zunächst die «Schematic plots» oder «Box and whisker plots» für die Milchdaten und deren im Abschnitt 4 ermittelte Gruppierung (Abb. 9). Jede Box umfaßt das Intervall 25%- bis 75%-Quantile und kennzeichnet damit den mittleren Datenbereich. Innerhalb jeder Box ist die 50%-Quantile, d. h. der Median, eingezeichnet. Die Definition der Länge der Whiskers (Barthaare) hängt von der gewählten Methode ab (DIETLEIN, 1981). Nach MCNEIL (1977) werden diese Whiskers so festgelegt, daß sie ca. 95 % des Datenbereiches einschließen. Werte, die außerhalb dieser Grenze liegen, werden einzeln geplottet. Eine zweite Form der graphischen Darstellung von Gruppenunterschieden ergibt sich durch Schätzung der Wahrscheinlichkeitsverteilung pro Merkmal und Gruppe. Dazu wird ein nichtparametrisches Verfahren gewählt, mit dem über dreiecksverteilte Kerne die Dichte eines Merkmals geschätzt werden kann (VICTOR, 1978). Abbildung 10 zeigt die geschätzten Dichtefunktionen des Merkmals Fettgehalt für die einzelnen Gruppen sowie für alle Daten.

6. Interpretation des Demonstrationsbeispiels

Die untersuchten Säugetiere der Ordnungen Primaten, Unpaarhufer, Paarhufer, Fleischfresser, Nagetiere, Hasentiere, Wale und Insektenfresser zeigen im Vernetzungsdiagramm (Abb. 4) eine differenzierte Gliederung. Die Milch von Igel (50), Seebär (38), Delphin (49), Wildschwein (11) und Florida-Waldkaninchen (45) liegen relativ isoliert, während die übrigen Tiere zu Gruppen geordnet sind. Primaten und Unpaarhufer bilden zusammen die Gruppe I. Dies läßt auf eine große quantitative Ähnlichkeit der Milchkomponenten schließen. Im Übergangsbereich von I nach II ist die Lama-Milch (12) lokalisiert. Die übrigen Paarhufer trennen sich in zwei Gruppen; erstere umfaßt die Familie Kamele und die Rinderartigen der Unterfamilien Rinder und Ziegen (Gruppe II), zweitere bilden die Familien Hirsche und Antilopen (Gruppe IV). Die freizehigen Fleischfresser gliedern sich in Hunde und Bären. Die Milchen der Hunde ergeben zusammen mit denen der Nagetiere die Gruppe III; die Milchzusammensetzung der Bären tendiert mehr in Richtung Wale (Gruppe V). Letztere Gattung sowie die Hasentiere bilden nur andeutungsweise Vernetzungsschwerpunkte. Beim Vergleich mit Tabelle 1 wird klar, daß in der nichtlinearen Abbildung (Abb. 3), die als Basis für das Vernetzungsdiagramm dient, im allgemeinen die Milchen von links nach rechts mit zunehmender Gesamtrohmasse geordnet wurden. Die Gliederung

der Objekte in der Hauptkomponentenanalyse und im Biplot (Abb. 1, 2) zeigt gegenüber den Abbildungen 3 und 4 nur geringe Abweichungen. Die Gruppierungen I bis IV sind auch hier relativ gut zu erkennen. Ebenso bilden die Wale und die Hasentiere nur andeutungsweise Gruppierungen. In Abbildung 1 wird das Objekt 12 (Lama-Milch) – ebenso wie beim Austauschverfahren (Abb. 6) – den Paarhufern (Gruppe II) zugeordnet. Während in den Verfahren Hauptkomponentenanalyse, Biplot, nichtlineare Abbildung und zum Teil auch im Vernetzungsdiagramm keine Entscheidung über Gruppierungen getroffen wird, ergeben das Dendrogramm, das Austauschverfahren und die unscharfe Gruppierung (Abb. 5, 6, 7) mehr oder weniger klare Entscheidungshilfen für die Definition von Clustern. Überprüft man jedoch die Güte der gebildeten Gruppen nach den in Abschnitt 5 ausgeführten Kriterien, so wird klar, daß bei den Tiergruppen Wale (46–48), Bären (34–37) und Hasenartige (43–45) eine eindeutige Grenzziehung nicht möglich ist.

Dieses Demonstrationsbeispiel umfaßt einerseits Gruppen von Beobachtungen, welche sich über alle Verfahren hinweg als stabil erweisen, und andererseits Beobachtungen, welche relativ isoliert liegen bzw. als Zwischen- oder Randpunkte zu betrachten sind. Diese Struktur der Daten wird nach Meinung der Autoren am adäquatesten durch die Abbildungen 4 und 7, d.h. durch das Vernetzungsdiagramm und die «unscharfe Gruppierung», beschrieben.

Literatur

- BOCK, H. H.: Automatische Klassifikation. Vandenhoeck & Ruprecht, Göttingen 1974.
- BOCK, H. H., 1979: Clusteranalyse mit unscharfen Partitionen. In BOCK, H. H. (Hrsg.): Klassifikation und Erkenntnis III, Studien zur Klassifikation, Bd. 6, Proc. der 3. Fachtagung der Gesellschaft für Klassifikation e.V., Frankfurt, 137–163.
- BUSSE, M., 1970: Eine neue Methode zur Untergliederung eng verwandter Bakteriengruppen am Beispiel der Enterobakterien. In DELLWEG, H. (Hrsg.): Zweites Symposium Technische Mikrobiologie, Institut für Gärungsgewerbe und Biotechnologie, Berlin, 243–257.
- DIETLEIN, G., 1981: Schematic Plots – eine Alternative zur Darstellung von mittleren Verlaufskurven. Stat. Software Newsletter 7, 100–103.
- EVERITT, B.: Cluster Analysis. Heinemann Educational Books Ltd., London 1974.
- EVERITT, B.: Graphical Techniques for Multivariate Data. Heinemann Educational Books Ltd., London 1978.
- GABRIEL, K. R., 1971: The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika* 58, 453–467.
- GABRIEL, K. R., G. RAVE, E. WEBER, 1976: Graphische Darstellung von Matrizen durch das Biplot. *EDV in Medizin und Biologie* 7, 1–15.
- JENNESS, R., 1974: The Composition of Milk. In LARSON, B. L., V. R. SMITH (Eds.): Lactation. Comprehensive Treatise. Volume III: Nutrition and Biochemistry of Milk, Academic Press, London, 3–107.
- MCNEIL, C., 1977: Interactive Data Analysis. John Wiley and Sons, New York.
- OHMAYER, G., M. PRECHT, H. SEILER, M. BUSSE, 1980: Linkagemaps and their Relations to Linkage Cluster Procedures. *Zbl. Bakt. II. Abt.* 135, 22–37.
- OHMAYER, G., 1982: Ein einfaches Wahrscheinlichkeitsmodell «Klassifikation binärer Daten». Dissertation, TU München-Weihenstephan.
- SAMMON, J. W., 1969: A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, C-18, 401–409.
- VICTOR, N., 1978: Alternativen zum klassischen Histogramm. *Meth. Inform. Med.* 17, 120–126.

Eingegangen am 28. März 1985

Anschrift der Verfasser: Dr. G. Ohmayer, Abt. Mathematik und stat. Methodenlehre, Datenverarbeitungsstelle, TU München-Weihenstephan, D-8050 Freising 12
Dr. H. Seiler, Bakteriologisches Institut der Süddeutschen Versuchs- und Forschungsanstalt für Milchwirtschaft, D-8050 Freising 12