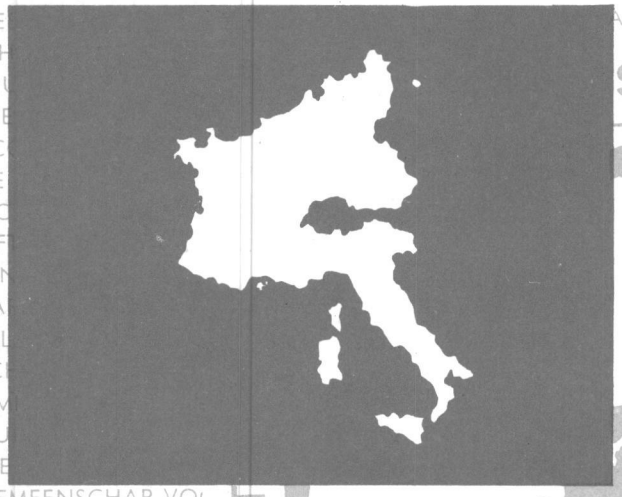
STATISTISCHES AMT
DER EUROPÄISCHEN GEMEINSCHAFTEN

OFFICE STATISTIQUE
DES COMMUNAUTÉS EUROPÉENNES



AGRARSTATISTISCHE STUDIEN

10

Klassifizierung landwirtschaftlicher Betriebe mit Hilfe multivariater statistischer Verfahren

Karl-August Schäffer

ISTITUTO STATISTICO
DELLE COMUNITÀ EUROPEE

BUREAU VOOR DE STATISTIEK
DER EUROPESE GEMEENSCHAPPEN

Dr. Karl-August Schäffer

ordentlicher Professor
an der Universität zu Köln

Klassifizierung landwirtschaftlicher Betriebe mit Hilfe multivariater statistischer Verfahren

Abgeschlossen am 12. September 1969

AGRARSTATISTISCHE HAUSMITTEILUNGEN

(Reihe „Agrarstatistische Studien")

INFORMATIONS INTERNES DE LA STATISTIQUE AGRICOLE

(Série «Études de statistique agricole»)

Das SAEG veröffentlicht im Rahmen seiner „Agrarstatistischen Hausmitteilungen" unter dem Titel „Agrarstatistische Studien" bestimmte Forschungsarbeiten, die in seinem Auftrag und für seine Bedürfnisse durchgeführt wurden. Mit der Zusammenfassung dieser Veröffentlichungen in einer gesonderten Reihe beabsichtigt das Amt, einen möglichst großen Kreis methodisch interessierter Leser zu erreichen.

Die in Frage stehenden Studien wurden Sachverständigen oder Sachverständigengruppen aus den Mitgliedsländern mit dem Ziel anvertraut, eine erschöpfende Analyse einzelner statistischer Probleme zu erlangen, Verbesserungen der Methoden in die Wege zu leiten, eine größere Vergleichbarkeit der vorhandenen Daten zu erzielen und neue Informationsquellen zu erschließen.

Wegen ihres teils sehr spezifischen Charakters werden jedoch nur solche Arbeiten veröffentlicht, die Fragen von gewisser Tragweite behandeln.

Grundsätzlich erscheinen die Studien in französischer und in deutscher Sprache. Falls die Autoren das Original in einer anderen Sprache angefertigt haben, kann das SAEG, je nach dem von den Lesern bekundeten Interesse, die zusätzliche Herausgabe der Originalfassung veranlassen.

Es sei bemerkt, daß für den Inhalt der Studien ausschließlich ihre jeweiligen Autoren verantwortlich sind.

1972

STATISTISCHES AMT
DER EUROPÄISCHEN GEMEINSCHAFTEN

– Agrarstatistik –

Centre Louvigny
Luxemburg

OFFICE STATISTIQUE
DES COMMUNAUTÉS EUROPÉENNES

– Statistique agricole –

Centre Louvigny
Luxembourg

G l i e d e r u n g

1 EINLEITUNG

Eine der Grundaufgaben der Statistik besteht darin, die Vielfalt von realen Erscheinungen übersichtlich darzustellen, mit dem Ziel, ihre Struktur erkennbar zu machen.

Der wichtigste Ansatz zur Lösung dieser Aufgabe geht davon aus, die zu untersuchende Gesamtheit von Einheiten (z.B. die Kreise in einem Lande) in Teilgesamtheiten zu gliedern und für jede dieser Teilgesamtheiten gesondert statistische Ergebnisse zu ermitteln. Eine Aufgliederung der Gesamtheit gibt jedoch nur dann klare Informationen über ihre Struktur, wenn alle einer Teilgesamtheit zugeordneten Einheiten sich bezüglich aller relevanten Eigenschaften möglichst ähnlich sind, die Teilgesamtheiten untereinander dagegen möglichst deutliche Unterschiede aufweisen (wie z.B. landwirtschaftliche Betriebe mit geringem Viehbesatz bzw. mit ausgeprägter Viehhaltung). Solche Teilgesamtheiten, die in sich verhältnismäßig homogen sind, heißen "Klassen", die entsprechende Aufgliederung der Gesamtheit eine "Klassifizierung".

In der Praxis werden Klassifizierungen häufig nur anhand eines einzigen Merkmals vorgenommen, das dafür auf Grund von sachlichen Ueberlegungen ausgewählt wird (z.B. das Bruttosozialprodukt je Einwohner). Selbst wenn zwei oder mehr Merkmale für die Aufgliederung herangezogen werden, bleibt das Ergebnis in vielen Fällen unbefriedigend: Ein Grund dafür liegt darin, daß der Wirklichkeit, die sich statistisch selbst durch viele Merkmale nur annähernd wiedergeben läßt, mit der Gliederung nach einigen wenigen Merkmalen Gewalt angetan wird. Die bei einer solchen Gliederung entstehenden Gruppen von Einheiten sind zwar in den Gliederungsmerkmalen homogen, brauchen aber deshalb nicht auch bezüglich aller übrigen Merkmale homogen zu sein. Dieser Sachverhalt wird weiter dadurch verschärft, daß eine kombinierte Gliederung nach mehreren Merkmalen regelmäßig zu einer unnötig großen Anzahl von Teilgesamtheiten führt, die zum Teil nur ganz schwach besetzt oder leer sind. Infolgedessen bedingt diese Technik auch noch den zusätzlichen Nach-

teil, daß die statistischen Nachweisungen stark aufgesplittert werden, die Auswertung der Ergebnisse also unnötig behindert - wenn nicht verhindert - wird.

Diese Nachteile lassen sich mit den in den letzten Jahren entwickelten Verfahren für die "statistische Klassifizierung" vermeiden. Diese Verfahren berücksichtigen bei der Aufgliederung der Gesamtheit statt einzelner Merkmale die Menge aller relevanten Merkmale. Außerdem gestatten sie es, die Anzahl der Klassen und ihren Umfang auf Grund der aktuellen Daten so festzulegen, daß die statistischen Ergebnisse die Struktur der Gesamtheit in möglichst klarer und übersichtlicher Form aufzeigen.

Der Rechenaufwand für die Verfahren zur statistischen Klassifizierung ist verhältnismäßig groß, sie können deshalb nur mit Hilfe von Rechenanlagen wirtschaftlich angewandt werden.

Die vorliegende Studie gibt zunächst einen allgemeinen Ueberblick über Methoden für die statistische Klassifizierung und ihre Anwendungsbereiche (Abschnitt 2). Ein Verfahren, das für die Klassifizierung von größeren Datenmengen besonders geeignet erscheint, wird in Abschnitt 31 dargestellt. Die praktische Anwendung dieses Verfahrens auf die Klassifizierung von landwirtschaftlichen Betrieben und die dabei gesammelten Erfahrungen sind in Abschnitt 33 dieser Studie beschrieben.

2 METHODOLOGIK DER KLASSIFIZIERUNG

21 AUFGABENSTELLUNG

In der statistischen Literatur wird der Begriff "Klassifizierung" (classification) für drei Aufgabenstellungen benutzt (vgl. M.G. KENDALL, 1966):

1. Das Zuordnungsproblem (Diskriminanzproblem)

   Vorgegeben ist eine Gruppe von Gesamtheiten, für die Stichproben und ferner Informationen über die Zugehörigkeit der Einheiten zu

den Gesamtheiten vorliegen. Das Problem der statistischen Zuordnung besteht dann darin, für weitere Einheiten aus dieser Gruppe von Gesamtheiten allein auf Grund ihrer Merkmalswerte mit möglichst großer Sicherheit jeweils diejenige Gesamtheit zu bestimmen, aus der die Einheiten stammen.

2. Das Aufgliederungsproblem (Klassifizierung im engeren Sinne)

Gegeben sind Einheiten, die zu einer Gesamtheit oder zu mehreren, voneinander verschiedenen Gesamtheiten ("Klassen") gehören, die in sich homogen sind in dem Sinne, daß Einheiten einer Klasse einander ähnlicher sind als Einheiten aus verschiedenen Klassen. Die Fragestellung des Aufgliederungsproblems lautet dann (vgl. R.L. THORNDIKE, 1953):

a) Wie groß ist die Anzahl der Klassen?

b) Wie sind im Fall $k \neq 1$ die Grenzen festzulegen, nach denen die Einheiten in die k Klassen eingeteilt werden?

3. Das Schichtungsproblem (Dissektionsproblem)

Gegeben sind Einheiten einer Gesamtheit. Zu bestimmen ist eine Unterteilung der Gesamtheit in Gruppen mit vorgeschriebenen Eigenschaften.

Eine Lösung für das Zuordnungsproblem ist bereits von R.A. FISHER (1938) angegeben worden; einen Ueberblick über die inzwischen zügig weiter entwickelte Diskriminanzanalyse geben C.R. RAO (1952) und T.W. ANDERSON (1958).

Das Problem der Schichtung nach einem Merkmal ist von T. DALENIUS (1950) und T. DALENIUS - M. GURNEY (1951) nach einem Optimalprinzip allgemein gelöst worden, auch die Arbeiten von D.R. COX (1957) und W.D. FISHER (1958) behandeln dieses Problem. Es ist bemerkenswert, daß M.J. HAGOOD - E.H. BERNERT (1945) bereits ein Verfahren zur Schichtung nach mehreren Variablen angegeben haben, das allerdings nicht auf einem Optimalprinzip aufgebaut ist.

Methodische Untersuchungen zum Aufgliederungsproblem sind dagegen erst verhältnismäßig spät begonnen worden, obwohl der Aufgabe, eine statistisch zweckmäßige Gliederung zu finden, sicher eine fundamentale Bedeutung zukommt. Diese Verzögerung ist vermutlich darauf zurückzuführen, daß es große Schwierigkeiten bereitet, die Aufgabe operabel zu definieren, d.h. insbesondere, sinnvolle Kriterien für das Steuern des Klassifizierungsverfahrens (z.B. nach der "Aehnlichkeit" von Einheiten bezüglich vieler Merkmale) anzugeben. Ein weiteres Hindernis bestand darin, daß eine Klassifizierung von Einheiten nach vielen Merkmalen Rechenkapazität in einem Umfange erfordert, die erst mit der Entwicklung elektronischer Datenverarbeitungsanlagen verfügbar geworden ist (vgl. dazu auch G. H. BALL, 1965).

Etwa ab 1960 sind viele verschiedenartige Verfahren zur Lösung des Problems der Klassifizierung von Einheiten nach mehreren Merkmalen vorgeschlagen worden (entsprechendes gilt auch für das duale Problem - die Gruppierung von Merkmalen -, auf das hier nur hingewiesen wird). Die ältesten Arbeiten auf diesem Gebiet stammen von Biologen; sie haben auch die Bezeichnung "Taxonometrie" für den Problemkreis geprägt.

Die Methoden für die Klassifizierung unterscheiden sich bezüglich ihrer Konstruktionsprinzipien und ihrer Anwendungsbereiche erheblich voneinander. Die wichtigsten Kriterien für die Systematisierung dieser Verfahren werden in den Abschnitten 22 bis 26 dargestellt; vgl. dazu auch G.H. BALL (1965) und P. DAGNELIE (1966).

22 ART DER MERKMALE FUER DIE KLASSIFIZIERUNG

Die Art der verfügbaren Merkmale hat wesentliche Konsequenzen für das Verfahren der Klassifizierung:

Bei den an einer n o m i n a l e n Skale gemessenen klassifikatorischen Merkmalen ist nur die Unterscheidung "gleich" bzw. "ungleich" möglich. Bei den Erwerbstätigen kann z.B. das Merkmal "Wirtschaftsbereich"

die Ausprägungen

"In der Landwirtschaft tätig"
"In der gewerblichen Wirtschaft tätig"
"In sonstigen Wirtschaftsbereichen tätig"

besitzen. Für die Zwecke der Klassifikation werden solche Merkmale allgemein auf Alternativmerkmale beschränkt, weil das Fehlen eines Abstandsmaßes auf einer nominalen Skale andernfalls zu Schwierigkeiten führt. Im Beispiel könnte die obengenannte Dreiteilung durch die Dichotomie

"In der Landwirtschaft tätig"
"Erwerbstätig, aber nicht in der Landwirtschaft"

ersetzt werden; damit ist selbstverständlich ein mehr oder minder großer Verlust an Information verbunden. Ueblicherweise werden die beiden Ausprägungen eines Alternativmerkmals mit "0" und "1" verschlüsselt.

Für die an einer o r d i n a l e n Skale gemessenen komparativen Merkmale ist über die Differenzierung "gleich - ungleich" hinaus eine Relation "größer als" definiert. Beispielsweise sind die Nummern der nach der landwirtschaftlichen Nutzfläche gebildeten Größenklassen von Betrieben als Ausprägungen eines komparativen Merkmals aufzufassen.

Die an einer r a t i o n a l e n Skale gemessenen metrischen Merkmale haben den höchsten Informationsgehalt, weil für sie auch ein Abstandsmaß (und damit zugleich eine Maßeinheit) definiert ist. Bei diesen metrischen Merkmalen besteht andererseits aber das Problem, daß meist nicht für alle Merkmale die gleiche Maßeinheit angewandt werden kann (z.B. wenn die landwirtschaftliche Nutzfläche und der Produktionswert eines landwirtschaftlichen Betriebes für die Klassifikation herangezogen werden sollen). In der Regel wird man deshalb bei metrischen Merkmalen statt der ursprünglichen Meßwerte die dimensionslosen standardisierten Werte für die Analyse heranziehen. Eine solche lineare Transformation bedingt keinen Verlust an statistischer Information, sie kann aber (vgl. G.H. BALL, 1965) die Klassifikation erheblich beein-

flussen. Dagegen zieht die bei einigen Verfahren vorgesehene Reduktion von metrischen Merkmalen auf bloße Alternativmerkmale erhebliche Einbußen an Information nach sich. Aus diesem Grunde ist es methodisch nicht sinnvoll, zur Klassifizierung von Einheiten, für die metrische Merkmale vorliegen, solche Verfahren anzuwenden, die eine Umwandlung in Alternativmerkmale erfordern.

23 DARSTELLUNG UND ZUSAMMENFASSUNG DER DATEN

231 Originaldaten

Es sei p die Anzahl der Merkmale, die für jede der n Einheiten beobachtet ist. Der Merkmalswert des j-ten Merkmals für die i-te Einheit soll mit

$$x_{ji} \qquad \begin{matrix} j = 1,2,\ldots,p; \\ i = 1,2,\ldots,n \end{matrix}$$

bezeichnet werden. Das gesamte Datenmaterial kann somit in Form einer Matrix

$$(1) \qquad \begin{pmatrix} x_{11}\cdots x_{1i}\cdots x_{1n} \\ x_{21}\cdots x_{2i}\cdots x_{2n} \\ \cdot \\ x_{j1}\cdots x_{ji}\cdots x_{jn} \\ \cdot \\ x_{p1}\cdots x_{pi}\cdots x_{pn} \end{pmatrix} =: \underline{X}$$

ausgedrückt werden. Jede Spalte dieser Matrix

$$(2) \qquad \begin{pmatrix} x_{1i} \\ \cdot \\ \cdot \\ \cdot \\ x_{pi} \end{pmatrix} =: \underline{x}_i \quad (i = 1,2,\ldots,n)$$

entspricht den für die i-te Einheit beobachteten Werten der p Merkmale.

Wenn alle Merkmale metrisch sind, läßt sich jede Einheit (bzw. jeder Spaltenvektor $\underline{x}_i$) als ein Punkt in einem p-dimensionalen Raum E_p deuten,

auf dessen rechtwinklig aufeinanderstehenden Koordinatenachsen die Werte der p Merkmale als Koordinaten abgetragen werden.

Die Unterschiede zwischen den im Abschnitt 22 betrachteten drei Arten von Merkmalen lassen sich geometrisch deuten: Falls alle p Merkmale nur der Werte 0 und 1 fähig sind, können in Raum E_p nur die Eckpunkte eines p-dimensionalen Würfels mit der Kantenlänge 1 (insgesamt also höchstens 2^p Punkte) besetzt sein. Bei komparativen und metrischen Merkmalen entfällt diese starke Einschränkung. Wenn alle p Merkmale komparativ oder diskontinuierlich sind, umfaßt der Merkmalsraum eine Menge von Gitterpunkten, deren Umfang erheblich über die Zahl der 2^p Würfeleckpunkte hinausgeht. Falls schließlich alle p Merkmale stetig sind, füllen die möglichen Stichprobenpunkte den Merkmalsraum E_p oder einen durch den Wertebereich der Merkmale abgegrenzten Teil von E_p kontinuierlich aus.

Der Aufgabe, eine Menge von Einheiten zu klassifizieren, entspricht in dieser geometrischen Darstellung das Ziel, die gesamte Menge von Punkten im Raum E_p in Punkthaufen (d.h. verhältnismäßig dicht beieinander liegende Punkte) aufzugliedern.

Jede Klassifizierung (d.h. Aufgliederung) der Menge A von n Einheiten in k Klassen ist als eine Partition der Menge A in Teilmengen $(A_1, A_2, \ldots, A_k)$ aufzufassen, d.h. in Teilmengen, die keine gemeinsamen Punkte besitzen und deren Vereinigung gleich der Menge A ist.

Für weiterführende theoretische Untersuchungen ist es zweckmäßig (vgl. P. SWITZER, 1968), zunächst eine Partition des Raumes E_p in "Schichten" zu definieren und die entsprechenden Grenzen für die Aufgliederung der Einheiten auf Klassen zugrundezulegen.

232 Hauptkomponenten

In der Regel sind die beobachteten Merkmale verhältnismäßig zahlreich und zudem mehr oder weniger stark untereinander korreliert. Die Klassifizierung wird durch beide Umstände erschwert. Diese Schwierigkeiten

lassen sich wesentlich vermindern, wenn die große Zahl korrelierter Merkmale rechnerisch durch eine möglichst kleine Zahl unkorrelierter Größen ersetzt werden kann. Als günstigste Lösung sind die "Hauptkomponenten" zu betrachten (vgl. T.W. ANDERSON, 1958). Sie sind linear so aus den Merkmalen konstruiert, daß eine kleine Zahl von Hauptkomponenten die Varianzen und Kovarianzen zwischen den beobachteten Werten so genau wie möglich reproduzieren.

Geometrisch läßt sich die Transformation in folgender Weise deuten: Sie geht davon aus, daß die Dispersion der Beobachtungswerte im Raum E_p näherungsweise durch ein Ellipsoid beschrieben wird. Die Darstellung dieses Ellipsoides läßt sich dadurch wesentlich vereinfachen, wenn neue Koordinatenachsen bestimmt werden, die mit den Hauptachsen des Ellipsoides zusammenfallen.

Die Formel für das Streuungsellipsoid im ursprünglichen Koordinatensystem lautet

$$(3) \qquad \underline{x}' \, \underline{S} \, \underline{x} = a$$

Dabei ist $a>0$ und $\underline{S}$ gleich der Matrix der Varianzen und Kovarianzen; wenn die Beobachtungswerte jeder Variablen auf die Mittelwerte bezogen sind, gilt

$$(4) \qquad \frac{1}{n-1} \underline{X} \, \underline{X}' = \frac{1}{n-1} \sum_{i=1}^{n} \underline{x}_i \, \underline{x}_i' =: \underline{S}$$

Die Transformation auf die Hauptachsen ergibt sich aus den Eigenvektoren $\underline{c}_k$ und Eigenwerten λ_k der Matrix $\underline{S}$:

$$(5) \qquad \underline{S} \cdot \underline{c}_k = \lambda_k \cdot \underline{c}_k \qquad k = 1,2,\ldots,p$$

Die Einheit $\underline{x}_i$ wird im neuen Koordinatensystem dargestellt durch den Vektor

$$(6) \qquad \underline{y}_i = \underline{C}' \cdot \underline{x}_i \qquad i = 1,2,\ldots,n$$

wobei $\underline{C}$ die aus den Eigenvektoren zusammengesetzte Transformationsmatrix ist. Die Koordinaten im neuen System werden "Hauptkomponenten" genannt.

Der Uebergang in das neue Koordinatensystem bietet einen weiteren Vorteil: In den meisten Fällen nehmen die der Größe nach geordneten Eigenwerte rasch ab. Da den Eigenwerten geometrisch die Quadrate der Achsenlängen des Ellipsoides entsprechen, folgt daraus, daß die Dispersion der Beobachtungswerte im wesentlichen durch zwei oder drei Hauptachsen erfaßt wird; der durch $m<p$ dominante Hauptachsen repräsentierte Varianzanteil ist

$$(7) \qquad \sum_{k=1}^{m} \lambda_k / p$$

Mit Hilfe der Hauptachsentransformation ist es also ohne großen Informationsverlust möglich, die p ursprünglichen Merkmale durch $m<p$ unkorrelierte neue Merkmale zu ersetzen und dadurch die Klassierung wesentlich zu erleichtern. Wenn nämlich nur 3 Einheiten mit je p Merkmalswerten verglichen werden sollen, so müssen dabei insgesamt 3p Werte berücksichtigt werden. Wenn jedoch zuvor die p Merkmale auf m Hauptkomponenten "kondensiert" werden, so sind nur noch 3m Werte zu vergleichen.

233 Orthogonale Faktoren

Für manche Zwecke (vgl. Abschnitt 2414) ist es vorteilhaft, die Maßstäbe im neuen Koordinatensystem so zu strecken, daß aus dem Dispersions-Ellipsoid eine Kugel entsteht. Das läßt sich durch die Transformation auf Faktorenwerte

$$(8) \qquad \underline{z}_i = \underline{A}' \cdot \underline{x}_i$$

erreichen; darin ist $\underline{A}$ die aus den gestreckten Eigenvektoren

$$(9) \qquad \underline{a}_k = \underline{c}_k \sqrt{\lambda_k} \qquad\qquad k = 1,2,\ldots,p$$

zusammengesetzte Matrix ($\lambda_k > 0$ für alle k vorausgesetzt).

Bezeichnet man mit $\underline{\Lambda}$ die aus den Eigenwerten $\lambda_1, \lambda_2, \ldots, \lambda_p$ gebildete Diagonalmatrix, so gilt

$$(10) \qquad \underline{A}' = \underline{\Lambda}^{+\frac{1}{2}} \cdot \underline{C}'$$

24 MASSE FUER DIE KLASSIFIZIERUNG

Die Klassifizierung der Einheiten anhand der vorliegenden multivariablen Daten soll möglichst weitgehend frei von subjektiven Einflüssen sein. Das läßt sich nur dann erreichen, wenn sinnvolle Kriterien für die Steuerung und Bewertung der Klassifizierung angegeben werden können. Grundsätzlich sind zwei Typen von Maßen zu unterscheiden: Maße für die Ähnlichkeit (oder Unähnlichkeit) von Einheiten und Maße für die Güte einer Klassifikation.

241 Maße für die Aehnlichkeit von Einheiten

In der Literatur findet sich eine große Zahl von Aehnlichkeitsmaßen; einen Ueberblick geben R.R. SOKAL - P.H.A. SNEATH (1963) und G.H. BALL (1965). Hier sollen nur die für die weiteren Ueberlegungen relevanten Maße dargestellt werden.

2411 Aehnlichkeitskoeffizienten

Für Alternativmerkmale schlagen SOKAL und SNEATH (1963) den Aehnlichkeitskoeffizienten (matching coefficient) als Maß für den Grad der Uebereinstimmung von zwei Einheiten vor:

$$(11) \qquad M_{hi} = a_{hi}/p$$

wobei a_{hi} die Anzahl der Attribute bedeutet, in denen die Einheiten h und i übereinstimmen.

Ein verwandtes Aehnlichkeitsmaß verwendet R.M. NEEDHAM (1965) für sein Klassifizierungsverfahren. Diese Maße sind für komparative und metrische Merkmale nicht zu empfehlen, weil die Reduktion auf eine Alternative willkürlich ist und vor allem die verfügbare Information nicht voll ausschöpft.

2412 Korrelationskoeffizienten

Für metrische Merkmale kann anstelle dieses Aehnlichkeitskoeffizienten ein K o r r e l a t i o n s k o e f f i z i e n t als Maß für den Grad der Uebereinstimmung angewandt werden. Im Gegensatz zu der üblichen R - Rechnik, bei der die Stärke eines korrelativen Zusammenhangs zwischen zwei Merkmalen ermittelt wird, geht die dazu duale Q - Technik von einer Korrelation zwischen je zwei Einheiten aus. Da die Merkmale in der Regel verschiedene Maßeinheiten besitzen, werden die Merkmalswerte zunächst standardisiert; sie sind dann dimensionslose Zahlen, deren arithmetischer Mittelwert gleich 0 und deren Varianz gleich 1 ist. Der über alle p Merkmale berechnete Korrelationskoeffizient für die h-te und i-te Einheit ist dann proportional zu

$$(12) \qquad q_{hi} = \sum_{j=1}^{p} x_{jh} \cdot x_{ji} = \underline{x}_h' \cdot \underline{x}_i ,$$

wenn man der Einfachheit halber unterstellt, daß die Daten der Matrix $\underline{X}$ bereits standardisiert sind.

Ein auf dieses Uebereinstimmungsmaß aufgebautes Verfahren ist von R.R. SOKAL et al. (1963) beschrieben worden. M.G. KENDALL (1966) hat vorgeschlagen, statt der auf Produktmomenten standardisierter Merkmale aufgebauten Korrelationskoeffizienten die Rangkorrelationskoeffizienten der Originalwerte anzuwenden, weil dieser Koeffizient unabhängig von monotonen Transformationen ist.

P. IHM (1964) hat darauf hingewiesen, daß die Q - Technik geometrisch interpretiert werden kann: Die Projektion der Vektoren $\underline{x}_i$ (i = 1,2,...,n) im p-dimensionalen Raum auf die Hyperebene

$$\sum_{j=1}^{p} x_j = 0$$

erzeugt (n-1) dimensionale Vektoren $\underline{y}$. Bezeichnet man die Projektion des Vektors $\underline{x}_i$ für die i-te Einheit mit $\underline{y}_i$, dann sind die Korrelationskoeffizienten q_{hi} gleich dem Kosinus des Winkels zwischen den Vektoren $\underline{y}_h$ und $\underline{y}_i$. Aus dieser Deutung folgert IHM, daß dieses Maß in den Fällen keine einwandfreie Klassifizierung liefert, in denen außer dem Winkel zwischen den Einheiten auch die Abstände relevant sind.

2413 Euklidischer Abstand

Für die Klassifizierung kann statt eines Aehnlichkeitsmaßes auch ein Maß für die Unähnlichkeit von Einheiten zugrundegelegt werden. Solche Maße werden gewöhnlich A b s t a n d s m a ß e genannt. Der ungewogene Euklidische Abstand - genauer gesagt: das Quadrat dieses Abstandes - ist für die zwei Einheiten mit den Nummern h und i wie folgt definiert:

$$(13) \qquad d_{hi}^2 = \sum_{j=1}^{p} (x_{jh} - x_{ji})^2 = (\underline{x}_h - \underline{x}_i)' (\underline{x}_h - \underline{x}_i)$$

Diese Maßzahl entspricht auch geometrisch dem quadratischen Abstand zwischen den Punkten $\underline{x}_h$ und $\underline{x}_i$ im p-dimensionalen Euklidischen Raum E_p. Je näher zwei Punkte beieinander liegen (d.h. je ähnlicher sie sind), desto kleiner fällt der Wert des Abstandsmaßes d_{hi}^2 aus. Ein solches Maß wurde bereits 1926 von K. PEARSON vorgeschlagen; sein "coefficient of racial likeness" ist dem Wert d_{hi}^2 proportional.

Grundsätzlich sollten die Merkmalswerte vor der Berechnung des Abstandsmaßes d_{hi}^2 standardisiert werden, damit es invariant gegenüber Aenderungen in den Maßeinheiten der einzelnen Merkmale wird. Auf diesem Maße ist z.B. das Klassifizierungsverfahren von J. MAC QUEEN (1967) aufgebaut.

Zwischen dem Euklidischen Abstand d_{hi}^2 und den oben betrachteten Maßen für die Aehnlichkeit bestehen (vgl. J.C. GOWER, 1966) folgende Zusammenhänge:

$$(14) \qquad d_{hi}^2 = q_{hh} + q_{ii} - 2q_{hi}$$

bzw.

$$(15) \qquad d_{hi}^2 = p(1 - M_{hi})$$

2414 Verallgemeinerter Abstand

Der Euklidische Abstand hat den großen Vorteil, daß er sehr leicht zu berechnen ist. Von R.A. FISHER wurde jedoch schon 1936 darauf hingewiesen,

daß das Maß einen wesentlichen Nachteil besitzt: Stark kerrelierte Merkmale gehen mit dem gleichen Gewicht in die Maßzahl ein wie schwach korrelierte Merkmale, d.h. stark korrelierte Merkmale erhalten mittelbar ein zu großes Gewicht.

Dieser Nachteil läßt sich vermeiden, wenn statt des Euklidischen Abstandes der von P.C. MAHALANOBIS (1936) vorgeschlagene "verallgemeinerte Abstand"

$$(16) \qquad D^2_{hi} = (\underline{x}_h - \underline{x}_i)' \underline{S}^{-1} (\underline{x}_h - \underline{x}_i)$$

angewandt wird; darin bedeutet $\underline{S}$ die Korrelationsmatrix der standardisierten Merkmalswerte (Vgl. Formel 4) und $\underline{S}^{-1}$ die dazu inverse Matrix (vorausgesetzt, daß der Rang von $\underline{S}$ gleich p ist). Wenn alle beobachteten Merkmale paarweise unkorreliert sind, geht der verallgemeinerte Abstand in den ungewichteten Euklidischen Abstand über, weil dann $\underline{S}$ und $\underline{S}^{-1}$ gleich der Einheitsmatrix sind. Falls sich dagegen die Matrix $\underline{S}$ von der Einheitsmatrix unterscheidet, werden die Werte der einzelnen Merkmale entsprechend ihrer Korrelation mit anderen Merkmalen gewichtet.

Auf Grund einer Anregung, die C.R. RAO (1952) gegeben hat, wurde von VAN DEN DRIESSCHE (1965) ein Verfahren entwickelt, das den verallgemeinerten Abstand zugrundelegt. Eine direkte Anwendung des verallgemeinerten Abstandes zur Klassifizierung ist nicht zweckmäßig, da sie viel Rechenarbeit erfordert. Es gibt jedoch eine Möglichkeit, die theoretischen Vorzüge dieses Maßes mit der einfachen Berechenbarkeit des Euklidischen Abstandes in Einklang zu bringen. Wenn man die Originaldaten für die i-te Einheit vermöge der Transformation

$$(8) \qquad \underline{z}_i = \underline{\Lambda}^{+\frac{1}{2}} \cdot \underline{C}' \cdot \underline{x}_i$$

zunächst auf die Faktorenwerte umrechnet und daraus dann die Euklidischen Abstände ermittelt, so erhält man den verallgemeinerten Abstand D^2_{hi}:

$$(17) \qquad (\underline{z}_h - \underline{z}_i)' (\underline{z}_h - \underline{z}_i)$$
$$= (\underline{x}_h - \underline{x}_i)' \underline{C} \; \underline{\Lambda}^{-\frac{1}{2}} \cdot \underline{\Lambda}^{-\frac{1}{2}} \underline{C}' (\underline{x}_h - \underline{x}_i)$$
$$= (\underline{x}_h - \underline{x}_i)' \cdot \underline{S}^{-1} \cdot (\underline{x}_h - \underline{x}_i)$$

Faßt man nämlich Formel (5) zusammen, so folgt

$$(18) \qquad \underline{S}\,\underline{C} = \underline{C} \cdot \underline{\Lambda}$$

und daraus

$$(19) \qquad \underline{S}^{-1} = \underline{C}\,\underline{\Lambda}^{-1}\underline{C}'$$

weil $\underline{C}$ die Matrix einer orthogonalen Transformation ist, für die

$$(20) \qquad \underline{C}'\underline{C} = \underline{C}\underline{C}' = \underline{I}$$

gilt (aus diesem Grunde bleiben die Euklidischen Abstände zwischen den Einheiten h und i bei der Transformation (6) invariant).

242 Maße für die Güte einer Klassifikation

Nach der allgemeinen Definition in Abschnitt 1 hat eine Klassifizierung die Aufgabe zu erfüllen, daß die einer Klasse zugeordneten Einheiten sich bezüglich aller Merkmale möglichst ähnlich sind, die Klassen untereinander aber möglichst unähnlich sind. Zur Beurteilung der Güte einer Klassifizierung werden also Maße benötigt, die alle p Merkmale berücksichtigen.

2421 Streuungsmaße

Aus dem Instrumentarium der verallgemeinerten Varianzanalyse (vgl. z.B. C.R. RAO, 1952) lassen sich mehrere Maße für die Güte von Klassifizierungen ableiten.

Die Menge von n Einheiten sei in die k Klassen

$$A_1, \ldots, A_h, \ldots, A_k$$

eingeteilt. Bezeichnet man mit n_h die Zahl der Einheiten in der Klasse A_h, so ist

$$(21) \qquad \underline{\bar{x}}_h = \sum_{i \in A_h} \underline{x}_i / n_h$$

der Vektor der Mittelwerte für die p Merkmale in der Klasse A_h und

$$(22) \qquad \underline{\bar{x}} = \sum_{h=1}^{k} \sum_{i \in A_h} \underline{x}_i / n$$

der Vektor der Gesamt - Mittelwerte für alle n Einheiten.

Die p-dimensionale Matrix

$$(23) \qquad \underline{T} = \sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})'$$

ist eine Verallgemeinerung der Gesamtstreuung; entsprechend generalisieren

$$(24) \qquad \underline{B} = \sum_{h=1}^{k} n_h (\bar{\underline{x}}_h - \bar{\underline{x}})(\bar{\underline{x}}_h - \bar{\underline{x}})'$$

die Streuung zwischen den Klassen und

$$(25) \qquad \underline{W} = \sum_{h=1}^{k} \left(\sum_{i \in A_h} (\underline{x}_i - \bar{\underline{x}}_h)(\underline{x}_i - \bar{\underline{x}}_h)' \right)$$

die Streuung innerhalb der Klassen.

Die Matrix $\underline{T}$ läßt sich - ebenso wie im eindimensionalen Fall die Gesamtstreuung - in zwei Komponenten zerlegen:

$$(26) \qquad \underline{T} = \underline{B} + \underline{W}$$

Aus dieser Relation, die für jede beliebige Partition der n Einheiten in k Klassen gilt, lassen sich mehrere Maße für die Güte einer Klassifizierung ableiten (vgl. H.P. FRIEDMAN - J. RUBIN, 1967).

Falls die Anzahl der Merkmale p = 1 ist, beschreibt Formel (26) die übliche Streuungszerlegung skalarer Größen. Es liegt somit nahe, die Partition zu suchen, für die die Streuung W innerhalb der Klassen möglichst klein ist; das ist gleichbedeutend mit der Forderung, den Quotienten T/W zu maximieren.

Für p>1 läßt sich dagegen nicht mehr unmittelbar ein Kriterium angeben, weil dann die Relation (26) eine Matrix-Gleichung ist, als Maß für die Beurteilung der Güte einer Klassifizierung aber ein Skalar benötigt wird.

Eine Möglichkeit für die Reduktion der Matrizen auf einen Skalar besteht darin, die Spur der Matrix $\underline{W}$ (d.h. die Summe ihrer Elemente in der Hauptdiagonalen) als Kriterium zu benutzen:

$$(27) \qquad \mathrm{sp}\underline{W} = \sum_{i=j} w_{ij} \qquad \text{wobei } \underline{W} = (w_{ij})$$

Dies Maß ist einfach zu berechnen, hat aber den Nachteil, daß es abhängig vom Maßstab der Merkmale ist und die Korrelation zwischen den Merkmalen nicht berücksichtigt.

Dieser Nachteil läßt sich vermeiden, wenn man stattdessen das Verhältnis der Determinanten

$$(28) \qquad \frac{|\underline{T}|}{|\underline{W}|} = |\underline{I} + \underline{W}^{-1}\underline{B}|$$

oder die Spur des Matrix-Produktes

$$(29) \qquad sp\left(\underline{W}^{-1} \cdot \underline{B}\right)$$

als Maß verwendet. Beide Maße erfordern höheren Rechenaufwand. Sie sind aber invariant unter nicht-singulären linearen Transformationen der Originalwerte, d.h. sie sind unabhängig vom Maßstab der Merkmale und berücksichtigen überdies die Korrelation zwischen den Merkmalen.

Der reziproke Wert des Maßes (28)

$$(30) \qquad \Lambda = |\underline{W}| \,/\, |\underline{T}|$$

wurde von S.S. WILKS (1932) als Prüfmaß zum Testen der Hypothese vorgeschlagen, daß k normalverteilte Gesamtheiten mit übereinstimmender Kovarianzmatrix sich auch bezüglich der Mittelwertvektoren nicht unterscheiden. Zu Testzwecken ist das Maß jedoch nur dann zu benutzen, wenn die Voraussetzungen über die Verteilung der Merkmale erfüllt sind. Ferner ist zu beachten, daß Λ u.a. von der Zahl k der Klassen abhängig ist; somit sind die Werte für unterschiedliche Klassenzahlen nicht unmittelbar vergleichbar (während $|\underline{T}|$ konstant ist, nimmt $|\underline{W}|$ mit wachsendem k monoton nicht zu).

2422 Informationsmaße

Ein anderer Ansatz zur Konstruktion des Gütekriteriums geht von dem Informationsmaß von SHANNON aus (vgl. C.S. WALLACE - D.M. BOULTON, 1968): Ohne Klassifizierung sind die Daten als n Vektoren von je p Werten darzustellen.

Falls dagegen die Einheiten klassifiziert sind, lassen sich die Informationen übermitteln durch Angabe von

1) der Anzahl der Klassen;
2) ein Verzeichnis der Klassenbezeichnungen;
3) die Bezeichnung der Klasse, zu der die Einheit gehört;
4) die Durchschnittswerte jeder Klasse und
5) den Abweichungen der Einheit von den Durchschnittswerten der Klasse.

Wenn die Einheiten der einzelnen Klassen dicht beisammen liegen, sind die unter 5) genannten Abweichungen klein. In diesem Fall kann der größte Teil der Information sehr viel kürzer wiedergegeben werden als ohne Klassifizierung. Die beste Klassifizierung ist diejenige Partition der n Einheiten, die es ermöglicht, die Information in der kürzesten Form zu übermitteln.

Auf Grund dieser Ueberlegungen haben WALLACE und BOULTON (1968) u.a. ein Informationsmaß für den Fall abgeleitet, daß alle p Variable innerhalb der Klassen unkorreliert sind und jeweils einer p-dimensionalen Normalverteilung folgen.

Für viele praktische Aufgaben dürften die in diesem Ansatz unterstellten Annahmen zu eng sein. Prinzipiell hat das Kriterium viele Vorzüge, insbesondere den, daß die zugrundeliegenden Voraussetzungen explizit anzugeben sind.

25 KONSTRUKTION DER KLASSEN

Das allgemeine Prinzip für die Konstruktion von Klassen besteht darin, die vorgelegten Einheiten anhand der für sie vorliegenden Daten so in Teilmengen aufzugliedern, daß eine passend gewählte Gütefunktion optimiert wird (J. RUBIN, 1967). So können z.B. als Klassen die Teilmengen von Einheiten definiert werden, die ein Maß für die Aehnlichkeit der Einheiten innerhalb der Teilmengen maximieren. Die resultierenden Klassen sind abhängig von dem zugrundegelegten Maßstab und von der Art des Vorgehens. Die Prozedur definiert implizite die Klassen.

251 Theorie der optimalen Klassifikation

Rein theoretisch ist das Vorgehen denkbar, für alle möglichen Partitionen der Menge von n Einheiten in k Teilmengen jeweils das gewählte Gütemaß zu berechnen und so die optimale Partition zu bestimmen (vgl. R.L. THORNDIKE, 1953). Praktisch ist diese Methode jedoch selbst mit Hilfe von leistungsfähigen Rechenanlagen nicht anwendbar, weil die Anzahl der möglichen Partitionen von n Einheiten in k nicht-leere Teilmengen gleich den STIRLING'schen Zahlen zweiter Art sind, die rapide sowohl mit n als auch mit k wachsen: Wenn k = 2 Klassen gebildet werden, gibt es bereits $(2^{n-1}-1)$ Partitionen, d.h. bei nur n = 12 Einheiten wären schon mehr als 2000 Fälle durchzurechnen. Bei k = 3 ist die Anzahl der Partitionen gleich $\frac{1}{2}(3^{n-1}-2^n+1)$, für n = 12 sind das bereits über 86 000 Partitionen (vgl. im einzelnen die Arbeit von J.J. FORTIER - H. SOLOMON, 1966).

252 Approximation der optimalen Klassifikation

In der weitverstreuten Literatur über die Klassifikation sind viele Verfahren beschrieben worden, die versuchen, die optimale Partition ohne Durchmustern aller möglichen Partitionen zu bestimmen oder doch wenigstens eine Partition zu konstruieren, die als näherungsweise optimal gelten kann.

Da die Klassifizierung als Optimierungsaufgabe anzusehen ist, liegt es nahe, die Verfahren nach der Art ihrer Zielfunktion zu gliedern. Wir unterscheiden partielle Verfahren und globale Verfahren, je nachdem, ob die Konstruktion der Klassen von Maßen für die Aehnlichkeit von Einheiten ausgeht oder ob sie ein globales Maß für die Güte der Klassifikation zu maximieren trachtet.

2521 Partielle Klassifizierungsverfahren

Bei den partiellen Verfahren sind grundsätzlich zwei Konstruktionsprinzipien zu unterscheiden (vgl. J.C. GOWER, 1967):

Agglomerative Methoden bilden die Klassen schrittweise durch Zusammenfügen von Einheiten oder zuvor definierten Teilmengen von Einheiten.

Divisive Methoden gehen den entgegengesetzten Weg, indem sie die gesamte Menge von Einheiten schrittweise in Teilmengen aufteilen.

B. KING (1967) hat darauf hingewiesen, daß bei allen schrittweise vorgehenden Methoden die Gefahr besteht, das Optimum zu verfehlen. In dieser Hinsicht bestehen erhebliche Unterschiede zwischen folgenden zwei Verfahrenstypen:

Nicht-iterative Methoden bestimmen jeweils endgültig die Klassenzugehörigkeit von Einheiten, d.h. Einheiten, die einmal einer Klasse zugeordnet worden sind, können aus dieser Klasse nicht wieder ausgesondert werden, auch wenn die Klasse - etwa durch das Zusammenlegen mit einer anderen Klasse - in ihren Eigenschaften wesentliche Aenderungen erfahren hat.

Iterative Methoden sehen Möglichkeiten vor, die Zuteilung von Einheiten zu einer Klasse im Verlauf des Prozesses zu revidieren, wenn dadurch der Wert des Gütekriteriums in Richtung auf das Optimum verändert werden kann.

Eine Gruppe von Klassifizierungsverfahren geht von der (symmetrischen) Matrix

$$(31) \qquad \underline{D} = (d_{hi})$$

der Abstände d_{hi} zwischen allen $n(n-1)/2$ Paaren von Einheiten (bzw. von Maßzahlen für deren Uebereinstimmung) aus und betrachtet diejenigen Einheiten als zu einer Klasse gehörig, deren Abstand untereinander einen vorgegebenen Grenzwert δ nicht übersteigt (bzw. deren Uebereinstimmungsmaß δ nicht unterschreitet). Bei den divisiven Verfahren läßt man δ schrittweise Werte vom Maximalwert des Abstandes bis Null durchlaufen; die Zahl der Klassen wächst dementsprechend von 1 bis n. Im Gegensatz dazu wird bei den agglomerativen Verfahren dieses Typs der

Grenzwert δ sukzessive von Null bis zum Maximalwert variiert; im Anfangsstadium existieren n Klassen mit je einer Einheit, die im Laufe der Prozedur zu einer Klasse mit n Einheiten zusammenschrumpfen. Der Prozeß läßt sich durch ein Dendrogramm anschaulich machen. GOWER (1967) hat darauf hingewiesen, daß die beiden Verfahren ceteris paribus nicht immer zu übereinstimmenden Klassen führen: Bei den divisiven Verfahren sind die nächsten Nachbarn nicht immer in einer Klasse; agglomerative Verfahren neigen dazu, Außenseiter am Beginn des Prozesses abzusplittern. Verfahren dieses Typs sind schon verhältnismäßig früh, u.a. von R.L. THORNDIKE (1953) und P.H.A. SNEATH (1957) entwickelt worden. Diese beiden Methoden arbeiten agglomerativ und nicht-iterativ. Dagegen ist das Verfahren von R.M. NEEDHAM (1965) als divisiv und iterativ zu charakterisieren. Es geht wie folgt vor: Die Menge der n Einheiten wird zunächst willkürlich in zwei Teile zerlegt und über alle Paare von Einheiten, die zu verschiedenen Teilmengen gehören, die Summe ihrer Aehnlichkeitskoeffizienten gebildet. Sukzessive wird dann jedes Element daraufhin untersucht, ob sein Verschieben von einer Teilmenge in die andere die Summe der Aehnlichkeitskoeffizienten vermindert; gegebenenfalls wird es der anderen Teilmenge zugeordnet. Diese Untersuchung wird solange fortgesetzt, wie Versetzungen möglich sind. Die praktische Anwendbarkeit dieses Verfahrens ist dadurch begrenzt, daß das Aehnlichkeitsmaß für alle $n(n-1)/2$ Paare von Einheiten berechnet und für die Klassifizierungsprozedur gespeichert werden muß.

Dieser auf der Abstandsmatrix $\underline{D}$ (bzw. auf einer entsprechenden Matrix von Aehnlichkeitskoeffizienten) beruhende Verfahrenstyp wird unpraktikabel, wenn die Anzahl n der zu klassifizierenden Einheiten groß ist. In diesem Falle sind Verfahren vorzuziehen, bei denen zunächst eine vorgegebene Zahl k von Einheiten ausgewählt wird, die jeweils als "Kristallisationskerne" einer Klasse benutzt werden. Die übrigen (n-k) Einheiten werden einzeln wie folgt behandelt: Die Abstände der gerade betrachteten Einheit zu den Schwerpunkten der k Gruppen werden berechnet und diese Einheit derjenigen Gruppe zugeordnet, von deren Schwerpunkt sie den kleinsten Abstand besitzt; nach dem Hinzufügen eines neuen Punktes wird jeweils der Schwerpunkt der erweiterten Gruppe neu berechnet. Nach dem Zuordnen der letzten Einheit zu einer der Gruppen bricht das Verfahren ab.

Zu seiner Durchführung müssen statt n(n-1)/2 nur noch (n-k)k Abstände berechnet und verglichen werden. Die Reduktion entspricht näherungsweise dem Verhältnis 2k/n; sie wird also umso deutlicher, je größer die Zahl der zu klassifizierenden Einheiten ist. Zu diesem Typ agglomerativer Verfahren gehören die Methoden "ISODATA" von G.H. BALL - D.J. HALL (1965) und das von J. MAC QUEEN (1967) begründete "Verfahren der k Mittelwerte". Beide Methoden verwenden den Euklidischen Abstand als Maßstab. Da die Wahl der Kristallisationskerne das Ergebnis der Klassifizierung erheblich beeinflussen kann, gehen beide Verfahren iterativ vor: Die im ersten Durchlauf ermittelten Schwerpunkte der Klassen werden als Kistallisationspunkte im zweiten Durchlauf benutzt, in dem für jede der n Einheiten die Klassenzugehörigkeit neu festgelegt wird. Jedoch läßt sich auch durch Iterationen die Abhängigkeit des Ergebnisses von der Wahl der Kristallisationspunkte nicht vollständig ausschalten. Es bleibt ferner die Schwäche, daß der Euklidische Abstand bei stark korrelierten Merkmalen zu Verzerrungen führen kann.

Auf dem Euklidischen Abstand beruht auch das von P. SCHNELL (1964) vorgeschlagene divisive Verfahren: Jede Einheit $\underline{x}_i$ wird als ein Punkt im p-dimensionalen Raum betrachtet. Jedem dieser n Punkte wird dann eine glockenförmige Funktion derart zugeordnet, daß ihr Mittelwert gleich $\underline{x}_i$ ist. Eine solche Funktion ist die p-dimensionale Normalverteilung mit der Kovarianzmatrix $\underline{I}$; rechentechnisch einfacher ist die von E. FABER (1968) vorgeschlagene "rationale Glocke"

$$(32) \qquad f(\underline{x} \mid \underline{x}_i, t) = \frac{1}{1+t^2 \mid \underline{x}-\underline{x}_i \mid^2}$$

wobei t ein veränderlicher Parameter ist. Durch Ueberlegerung der n Funktionen ergibt sich die Gesamtfunktion

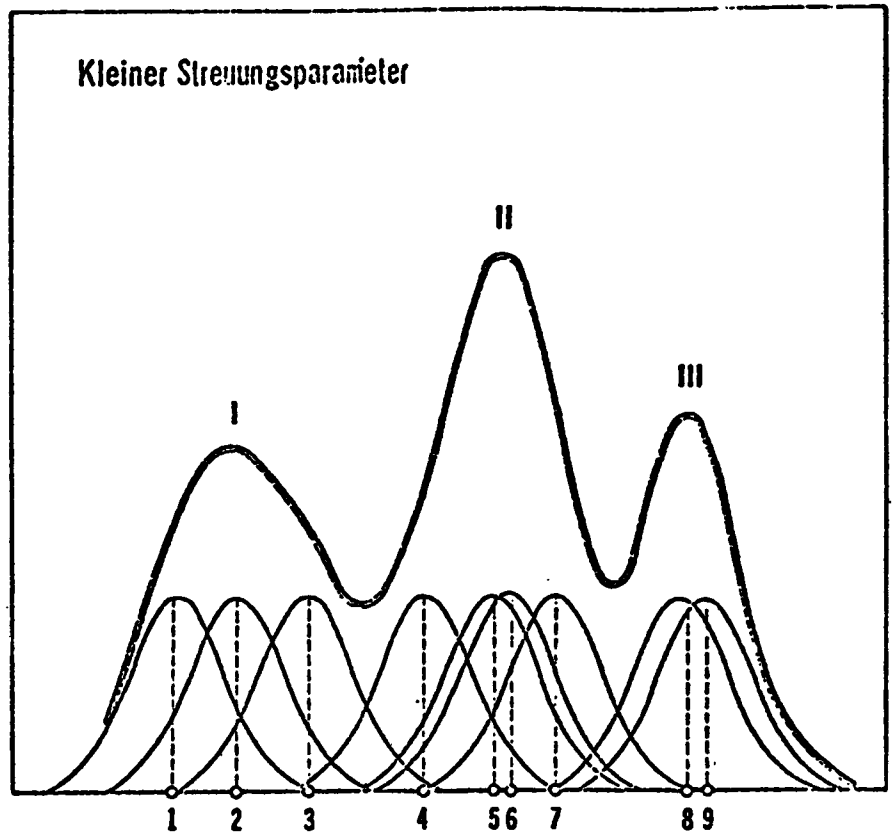
$$(33) \qquad F(\underline{x}/t) = \sum_{i=1}^{n} \frac{1}{1+t \mid \underline{x}-\underline{x}_i \mid^2}$$

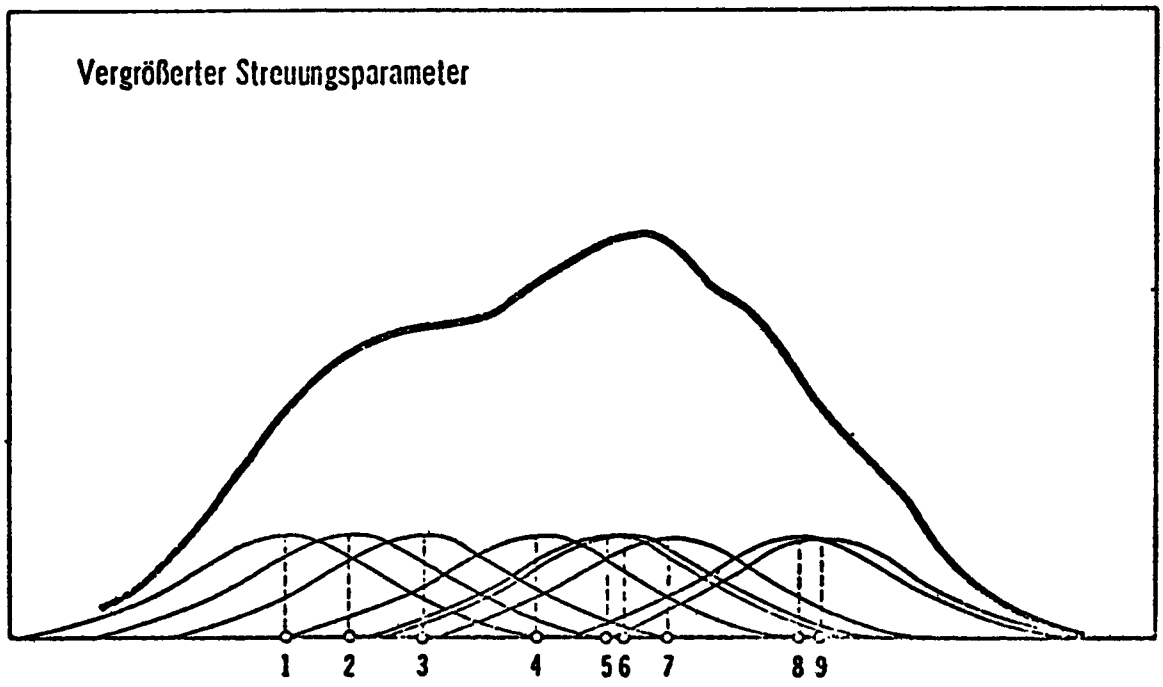
Die Maxima dieser Funktion liegen dort, wo die den Einheiten entsprechenden Punkte dicht beieinander liegen. Die Existenz von Maxima ist somit ein Indiz für Klassen. Dieser Zusammenhang ist aus der Abbildung auf Seite 23 zu ersehen, die auf den Fall p = 1 abgestellt ist:

KLASSIERUNGSVERFAHREN VON P. SCHNELL

Beispiel für den eindimensionalen Fall

——— Einzelne Dichtefunktion ▬▬ Summe der Funktionen





STAT. BUNDESAMT 3567

Die Gesamtfunktion im oberen Teil hat drei Maxima, denen drei Klassen von Einheiten entsprechen. Für die Zuordnung der Einheiten zu den Klassen ist von P. SCHNELL ein Gradientenverfahren angegeben worden. Es sucht für jeden Punkt $\underline{x}_i$ den Weg des steilsten Anstieges zum nächstgelegenen Maximum. Alle Punkte, die zum gleichen Maximum führen, werden in eine Klasse eingeordnet. Im oberen Teil der Abbildung führen z.B. die Punkte Nr. 1, 2 und 3 zum gleichen Maximum und werden dementsprechend zu einer Klasse zusammengefaßt.

Die Zahl k der Klassen hängt vom Parameter t der Funktion $F(\underline{x}|t)$ ab: Für extrem große Werte von t führt jeder Punkt $\underline{x}_i$ zu einem gesonderten Maximum, d.h. die Zahl der Klassen ist gleich n. Verkleinert man den Parameterwert, so verschmelzen immer mehr Maxima, bis schließlich nur noch ein Maximum der Funktion $F(\underline{x}|t)$ existiert, alle Einheiten also in eine Klasse fallen. Ein Beispiel dafür gibt der untere Teil der Abbildung auf Seite 23.

Dieses Klassifizierungsverfahren ist als nicht-iterativ einzustufen, da es bei vorgegebenem Parameterwert t jede Einheit endgültig einer Klasse zuordnet; weil dabei aber alle übrigen Einheiten berücksichtigt werden, entfallen hier weitgehend die Nachteile der nicht-iterativen Verfahren. Andererseits bleibt die Gefahr von Verzerrungen bei hochkorrelierten Merkmalen; sie kann auch durch die aus rechentechnischen Gründen vorteilhafte Transformation der Merkmale auf die dominanten Hauptkomponenten nicht beseitigt werden, weil diese Transformation Euklidische Abstände invariant läßt. Weiter ist zu beachten, daß der Rechenaufwand für die Gradientenmethode steil mit der Zahl der Einheiten wächst.

2522 Globale Klassifizierungsverfahren

Die von BALL und FRIEDMAN (1968) festgestellte Tendenz einer Entwicklung von den ad-hoc-Algorithmen zu einer zusammenhängenden Klassifizierungsmethodik, die mit den bekannten multivariaten Methoden in einer durchschaubaren Interrelation steht, trifft in erster Linie für die globalen Klassifizierungsverfahren zu. Diese Verfahren beruhen nach unserer Definition auf Maßen für die Güte einer Klassifikation und suchen - in recht

verschiedener Weise - diejenige Klassifikation, für die das gewählte Gütemaß den Optimalwert annimmt oder diesem Werte wenigstens sehr nahe kommt.

Das Verfahren von FORTIER und SOLOMON (1966) geht davon aus, eine einfache Zufallsstichprobe aus der Menge aller Partitionen zu ziehen, das Gütemaß für die einzelnen Partitionen in der Stichprobe zu berechnen und die danach "beste" Partition als Klassifikation zu verwenden. Der Umfang der Stichprobe läßt sich so ermitteln, daß die so bestimmte beste Partition in der Stichprobe mit einer vorgegebenen Wahrscheinlichkeit zu der Teilmenge in der Gesamtheit aller Partitionen gehört, die noch als "gut" angesehen werden kann (der Anteil dieser Teilmenge an der Menge aller Partitionen muß vorgegeben werden). Der Nachteil dieses - auf die Klassifizierung von Merkmalen abgestellten - Verfahrens liegt darin, daß die Verteilung des Gütekriteriums im entscheidenden Bereich wie eine Exponentialverteilung verläuft: aus diesem Grunde kann eine Partition, die noch zu den "guten" im Sinne des vorgegebenen Anteilswertes gehört, im Vergleich zur angestrebten optimalen Klassifikation wenig befriedigend sein.

Methodisch interessant ist die von H.D. VINOD (1969) aufgezeigte Möglichkeit, die Aufgabe der Klassifizierung als Problem der ganzzahligen Programmierung mit einer linearen Zielfunktion zu formulieren. Ob dieser Ansatz, der zunächst nur auf Alternativmerkmale abgestellt ist, mit erträglichem Rechenaufwand gelöst werden kann, muß noch untersucht werden.

Das von K.J. JONES (1968) vorgeschlagene Verfahren ist auf dem Kriterium der Modalität

$$(34) \qquad M = \frac{m_4}{m_2^2} - \frac{m_3^2}{m_2^3}$$

aufgebaut; darin bedeuten die Größen m_h ($h = 2,3,4$) die zentrierten Momente einer Variablen. Für eine 2-Punkt-Verteilung ist $M = 1$, für eine Rechteckverteilung gilt $M = 1,8$. Je kleiner also der Koeffizient M ist, desto deutlicher sind Unterschiede zwischen Klassen zu erkennen. Das Verfahren von JONES sucht dementsprechend eine lineare Transformation $\underline{L}$

der p Merkmale in der Datenmatrix $\underline{X}$ derart, daß für die p transformierten Variablen (in den Zeilen der Matrix)

$$(35) \qquad \underline{Y} = \underline{L} \cdot \underline{X}$$

die Optimalforderung

$$(36) \qquad M_j = \text{Min!} \quad \text{für } j = 1,2,\ldots,p$$

erfüllt ist (M_j ist das für die j-te Variable berechnete Modalitäts-Kriterium). Für die Klassifizierung der Einheiten nach den so transformierten Variablen hat JONES jedoch kein einheitliches Verfahren angegeben.

Die Untersuchungen von H.P. FRIEDMAN und J. RUBIN (1967) gehen von den Streuungsmatrizen $\underline{T}$, $\underline{W}$ und $\underline{B}$ (vgl. die Formeln 23, 24 und 25) aus und beruhen auf folgenden Prinzipien für eine Klassifizierung:

a) $\quad \text{sp } \underline{W} = \text{Min!}$

b) $\quad |\underline{T}| \,/\, |\underline{W}| = \text{Max!}$

c) $\quad \text{sp } \underline{W}^{-1}\underline{B} = \text{Max!}$

Gegen das erste Prinzip ist - ebenso wie gegen das damit äquivalente Verfahren von MAC QUEEN - einzuwenden, daß es nur die in der Hauptdiagonale der Matrix $\underline{W}$ stehenden Varianzen der Merkmale berücksichtigt, die Kovarianzen dagegen außer acht läßt. Die Prinzipien b) und c) sind deshalb dem Prinzip a) überlegen, weil sie die Kovarianzen zwischen den Merkmalen in das Klassifizierungsverfahren einbeziehen. Andererseits erfordert das Klassifizieren nach dem Prinzip a) wesentlich weniger Rechnerzeit als die Klassifizierung nach den beiden anderen Prinzipien.

Das Verfahren, eine Partition der vorgelegten Einheiten zu finden, die der Maximalforderung b) oder c) möglichst weitgehend entspricht, geht von einer frei gewählten Anfangspartition aus und versucht, iterativ durch Verschieben von jeweils einer Einheit in andere Klassen den Wert des Gütemaßes zu vergrößern. Nach Erreichen eines lokalen Maximums werden die Einheiten neu den Klassen zugeordnet; dabei dienen die verallgemeinerten Abstände, die auf Grund der Matrix $\underline{W}$ errechnet werden, als

Maßstab. Das Verfahren wird solange alternierend fortgesetzt, bis keine Verbesserung im Gütemaß mehr erzielbar ist. Das Verfahren nach Prinzip b) erfordert weniger Rechenzeit als die auf Prinzip c) aufgebaute Methode und scheint nach den praktischen Erfahrungen von FRIEDMAN und RUBIN (1967) auch bessere Ergebnisse zu liefern.

FRIEDMAN und RUBIN haben ausdrücklich darauf hingewiesen, daß ihr Verfahren nicht mit Sicherheit zum Optimum führt. Dem ist hinzuzufügen, daß die auf den Streuungsmatrizen beruhenden Methoden auch nur dann einigermaßen sinnvolle Ergebnisse liefern können, wenn die Struktur der Daten allein durch die Mittelwerte, Varianzen und Kovarianzen ausreichend beschrieben werden kann (vgl. dazu G.H. BALL, 1965).

26 ANZAHL DER KLASSEN

Die Aufgabe, die dem Material adäquate Anzahl k der Klassen festzulegen, läßt sich nicht allein auf Grund der Information über die Streuung $\underline{B}$ zwischen den Klassen oder der durchschnittlichen Streuung $\underline{W}$ innerhalb der Klassen lösen: Jede Vergrößerung der Anzahl k bewirkt eine Abnahme der Streuung innerhalb der Klassen. R.L. THORNDIKE (1953) hat vorgeschlagen, die Größe sp $\underline{W}$ in Abhängigkeit von k zu betrachten und als adäquate Klassenzahl denjenigen Wert k zu wählen, von dem ab die Funktion verhältnismäßig flach verläuft, die weitere Vergrößerung der Klassenzahl also keine wesentliche Streuungsreduktion mehr einbringt.

FRIEDMAN und RUBIN (1967) haben empfohlen, für verschiedene Klassenzahlen k jeweils den Wert log $|\underline{T}|/|\underline{W}|$ für die günstigste Klassifizierung zu berechnen. Ein Abflachen des Werteverlaufes dieser Größe mit zunehmender Klassenzahl deutet dann darauf hin, daß die dem Datenmaterial adäquate Anzahl von Klassen überschritten ist.

Eine praktisch leichter anwendbare Regel hat P. IHM (1965) unter der Voraussetzung entwickelt, daß innerhalb der Gruppen näherungsweise eine sphärische Dispersion vorliegt. Unter dieser Annahme müssen die der Größe nach geordneten Eigenwerte der Kovarianzmatrix $\underline{S}$ in ihrem Verlauf einen deutlichen Knick aufweisen, weil die Varianz zwischen den Klassen nur die

dominanten Eigenwerte beeinflußt. Daraus folgt, daß die sinnvolle Anzahl k der Klassen gleich der Zahl der Eigenwerte bis zum Knick anzunehmen ist.

Bei manchen Aufgabenstellungen (z.B. im biologischen Bereich) kann es zweckmäßig sein, die Klassifizierung für alle Klassenzahlen k im Bereich $1 \leq k \leq n$ zu ermitteln; eine solche Hierarchie von Klassen gibt Aufschlüsse über die Zusammengehörigkeit der Einheiten.

27 REALE BEURTEILUNG DER KLASSIFIKATION

Außer den formalen Kriterien zur Beurteilung der Güte einer Klassifikation (vgl. Abschnitt 242) sind nach H. BORKO (1965) und E.W. FORGY (1965) folgende reale Gesichtspunkte bei der qualitativen Beurteilung zu berücksichtigen:

Stabilität

Eine Klassifizierung heißt stabil, wenn mehrere Stichproben aus einem Material zu den gleichen Klassen führen. Diese Frage läßt sich untersuchen, indem das Material in eine Experimental-Stichprobe (2/3 des Materials) und in eine Validisierungs-Stichprobe eingeteilt wird, die gesondert nach dem gleichen Verfahren klassifiziert werden. Es ist dann festzustellen, ob die Ergebnisse im wesentlichen übereinstimmen.

Zweckmäßigkeit

Eine Klassifizierung soll zweckmäßig sein in dem Sinne, daß die empirisch aus dem Datenmaterial isolierten Klassen die Struktur der Gesamtheit widerspiegeln. Sofern eine Standardklassifikation der Masse existiert, ist zu prüfen, ob deren wichtigste Klassen mit dem Verfahren ermittelt worden sind.

Interpretierbarkeit

Bei der realen Beurteilung einer Klassifizierung ist schließlich auch ihre Interpretierbarkeit zu berücksichtigen. Die Ausrichtung des Klassifizierungsverfahrens auf optimale Effizienz im Sinne

eines Gütemaßes kann durchaus zu einer Klassifizierung führen, die sachlich nicht sinnvoll zu interpretieren ist. Das gilt z.B. dann, wenn alle Einheiten aus einer einzigen normalverteilten Gesamtheit stammen, die Klassifizierung jedoch zu mehr als einer Klasse führt.

3 KLASSIFIZIERUNG LANDWIRTSCHAFTLICHER BETRIEBE

31 VERFAHREN FUER DIE KLASSIFIZIERUNG

311 Kriterien für die Auswahl des Verfahrens

Bei der Festlegung des Verfahrens zur Klassifizierung landwirtschaftlicher Betriebe sind nach der in Abschnitt 2 dargestellten Methodologie vor allem folgende Gesichtspunkte zu beachten:

1. Für die landwirtschaftlichen Betriebe liegen aus der Agrarstrukturerhebung viele Merkmale vor, die in ihrer Kombination bei der Klassifizierung berücksichtigt werden sollen.
2. Die Merkmale der landwirtschaftlichen Betriebe sind überwiegend metrisch.
3. Die meisten Merkmalspaare sind verhältnismäßig hoch korreliert.
4. Der Rechenaufwand für die Klassifizierung darf nicht sehr hoch sein, damit das Verfahren auf eine große Menge von Betrieben angewandt werden kann.
5. Das Klassifizierungsverfahren soll möglichst wenig von willkürlichen Setzungen abhängen.

Durch die Feststellung 2. werden die für Alternativmerkmale entwickelten Verfahren ausgeschlossen. Punkt 4. hat zur Folge, daß die Methoden von IHM und SCHNELL sowie von FRIEDMAN und RUBIN, die relativ hohen Rechenaufwand erfordern, praktisch für die Klassifizierung von landwirtschaftlichen Betrieben nicht angewandt werden können. Das Verfahren von MAC QUEEN (1967) dürfte dagegen dem Grundsatz nach gut für die vorliegende Aufgabe geeignet sein.

312 Verfahren von MAC QUEEN

Nach G.N. LANCE und W.T. WILLIAMS (1966) müssen für jedes Klassifizierungsverfahren drei Punkte festgelegt werden:

1. Die Anfangsklassifizierung,
2. die Vorschrift für das Zuordnen von Einheiten zu Klassen und
3. die Stopvorschrift für das Beenden des Verfahrens.

Das Verfahren von MAC QUEEN (1967) sieht dafür folgende Regelung vor:

1. Das Verfahren startet auf Grund der Vorgabe einer Klassenzahl mit den k ersten Einheiten, die jeweils als Kristallisationskern einer Klasse verwandt werden. Aus den folgenden (n-k) Einheiten wird nach der Zuordnungsvorschrift die Anfangsklassifizierung konstruiert.
2. Für die Zuordnungsvorschrift sind zwei Parameter, C und R, vorzugeben. Auf Grund von k Kristallisationspunkten $\underline{z}_h$ (h = 1,2,...,k) im p-dimensionalen Raum werden die Einheiten - dargestellt durch Punkte $\underline{x}_i$ (i = 1,2,...,n) in diesem Raum - sukzessive in der vorliegenden Reihenfolge den Klassen zugeordnet: Die Gruppe g, von deren Kristallisationspunkt die Einheit den kleinsten Euklidischen Abstand besitzt, d.h. für die

$$(37) \qquad d_{gi} = \min_{h=1,\ldots,k} (\underline{x}_i-\underline{z}_h)'(\underline{x}_i-\underline{z}_h)$$

gilt, wird festgestellt. Falls $d_{gi} > R$ ist, wird die Einheit als neuer Kristallisationspunkt behandelt und die Zahl der Klassen um 1 erhöht (Verfeinerung der Klassifizierung). Falls dagegen $d_{gi} \leq R$ ausfällt, wird die Einheit $\underline{x}_i$ der Klasse g zugewiesen. Der Kristallisationspunkt dieser Klasse wird nach der Schwerpunktformel

$$(38) \qquad \underline{z}_g^{\vartheta+1} = (\underline{z}_g^{\vartheta} \cdot w_g^{\vartheta} + \underline{x}_i) : (w_g^{\vartheta+1})$$

neu berechnet; dabei ist ϑ die Nummer der Schwerpunktsberechnung und w_g sind ganzzahlige Gewichte, für die im Fall einheitlicher Auswahlwahrscheinlichkeiten der Einheiten

$$(39) \qquad \begin{aligned} w_g^1 &= 1 \\ w_g^{\vartheta+1} &= w_g^{\vartheta}+1 \end{aligned}$$

gilt.

Zu Beginn des Verfahrens und nach jeder Zuweisung einer Einheit wird der minimale Abstand zwischen allen Paaren von Kristallisationspunkten berechnet. Falls

$$(40) \qquad \min_{g,h}(\underline{z}^{\nu}_{g}-\underline{z}^{\nu}_{h})'(\underline{z}^{\nu}_{g}-\underline{z}^{\nu}_{h}) < C$$

ist, werden die entsprechenden Kristallisationspunkte unter Beachtung ihrer Gewichte zusammengelegt und die Zahl der Klassen um 1 vermindert (Vergröberung der Klassifizierung).

Die auf den ersten k Einheiten als Kristallisationspunkten aufgebaute Anfangsklassifikation liefert provisorische Schwerpunkte der Klassen. Für den zweiten Durchlauf werden diese Schwerpunkte als Kristallisationskerne behandelt und alle n Einheiten nach dem dargestellten Verfahren neu den Klassen zugeordnet.

3. Das Verfahren wird nach dem zweiten Durchlauf beendet.

Dieser als "k-Mittelwerte-Verfahren" bezeichnete Prozeß ist übersichtlich, benötigt verhältnismäßig wenig Rechenaufwand und erlaubt wegen seiner sequentiellen Arbeitsweise die Bearbeitung größerer Datenmengen. Diesen Vorteilen stehen folgende Nachteile gegenüber:

a) Das Verfahren benutzt den Euklidischen Abstand als Zuordnungskriterium, d.h. es berücksichtigt nicht die Korrelation zwischen den Merkmalen (vgl. Abschnitt 2414).

b) Die Verwendung der ersten k Einheiten als Kristallisationspunkte ist willkürlich und kann das Ergebnis der Klassifikation erheblich beeinträchtigen.

c) Die Zahl der endgültig gebildeten Klassen hängt von den drei Aktionsparametern k, C und R ab.
Die Vorgabe der Parameter C und R ist mangels theoretischer Anhaltspunkte schwierig und kann zu sachlich unbefriedigenden Klassenzahlen führen.

313 Verallgemeinerung des Verfahrens von MAC QUEEN

Diese Nachteile lassen sich durch folgende Modifikationen des Verfahtens beseitigen bzw. einschränken:

1. Vor dem Start des Verfahrens werden die Merkmalswerte standardisiert, die Eigenwerte und Eigenvektoren der Matrix $\underline{S}$ berechnet (vgl. Abschnitt 232) und schließlich in dem von q dominanten Eigenvektoren aufgespannten Teilraum E_q des Raumes E_p in Form von orthogonalen Faktoren (vgl. Abschnitt 233) dargestellt. Dabei werden die Kristallisationspunkte für die Anfangsklassifikation festgelegt, indem die Einheiten mit Extremalwerten in den ersten m Faktoren und Werten nahe Null herausgesucht werden.

2. Für das Zuordnungsverfahren können entweder die Parameter C und R oder die Grenzen k_{min} und k_{max} eines Bereichs für die gewünschte Klassenzahl $k_{min} \leq k \leq k_{max}$ vorgegeben werden; im zweiten Fall werden die Parameter auf Grund eines Vorschlages von E. KAMMERER in folgender Weise bestimmt:

$$(41) \qquad \begin{aligned} C &= 0{,}5 \cdot s_D \\ R &= 1{,}0 \cdot s_D \end{aligned}$$

   Dabei bezeichnet s_D die Standardabweichung der Abstände zwischen allen Paaren von Kristallisationspunkten.

   Die Zuordnung wird anhand der Abstände zwischen den Faktoren im Raum E_q vorgenommen, die näherungsweise gleich den verallgemeinerten Abständen der Punkte im Raum E_p sind, wenn als Metrik die Invertierte der Streuungsmatrix $\underline{S}$ verwandt wird; bei der Klassifizierung werden also die Korrelationen im Gesamtmaterial berücksichtigt. Im übrigen verläuft der Prozeß analog dem in Abschnitt 312 beschriebenen Verfahren.

3. Nach Abschluß des zweiten Durchlaufs wird geprüft, ob die erreichte Klassenzahl k in den Grenzen k_{min} und k_{max} liegt. Falls $k < k_{min}$ ausgefallen ist, wird das Verfahren mit den verkleinerten Werten $0{,}6 \cdot C$

und R - 0,4(R - C) erneut gestartet. Im Falle $k>k_{max}$ werden der erneuten Klassifizierung die vergrößerten Werte C + 0,4(R - C) und 1,4·R zugrundegelegt. Das Verfahren wird beendet, sobald die Vorschrift $k_{min} \leq k \leq k_{max}$ erfüllt ist.

Das in dieser Weise verallgemeinerte Verfahren ist von E. KAMMERER (1969) programmiert worden. Es enthält außer den dargestellten Arbeitsgängen Plausibilitätstests für die eingegebenen Werte, die Berechnung des Kriteriums $|\underline{W}|/|\underline{T}|$ sowie eine Endausgabe mit Daten für die Interpretation und Beurteilung der Klassifikation.

Die Laufzeit des Programmes ist wesentlich davon abhängig, ob die erreichte Klassenwahl nach Abschluß der Zuordnungsprozedur sofort die Grenzbedingungen erfüllt. In diesem Falle werden zur Klassifizierung von 100 Einheiten anhand von 30 Merkmalen etwa 6 Minuten Rechenzeit auf einer Anlage vom Typ IBM 360/75 benötigt. Der Zeitaufwand steigt mit der Zahl der von der Endbedingung erzwungenen Wiederholungen.

32 UNTERLAGEN UEBER LANDWIRTSCHAFTLICHE BETRIEBE

Für insgesamt 100 landwirtschaftliche Betriebe aus einer Region in Belgien standen außer den Ordnungsangaben der Betriebe folgende Daten aus der EWG-Agrarstrukturerhebung zur Verfügung:

Metrische Merkmale

9 Angaben zur Größe der Betriebe
27 Angaben zur Bodennutzung und über Besitzverhältnisse
22 Angaben zur Viehhaltung und zum Viehverkauf
16 Angaben über landwirtschaftliche Maschinen
7 Angaben über landwirtschaftliche Arbeitskräfte
9 weitere Angaben über die Betriebe.

Alternativmerkmale

30 Angaben über die landwirtschaftlichen Betriebe.

Mit Rücksicht darauf, daß die Klassifizierung vor allem Aufschluß über die Struktur der Betriebe geben soll, wäre es nicht sinnvoll, die Merk-

malswerte unmittelbar für die statistische Analyse heranzuziehen, weil sich dann in erster Linie die absolute Größe der Betriebe in der Klassifizierung niederschlagen würde. Aus diesem Grunde war es erforderlich, die metrischen Merkmale auf Verhältniszahlen umzurechnen.

Für die ersten Klassifizierungen wurden entsprechend dieser Ueberlegung alle 90 metrischen Merkmale relativiert. Es zeigte sich jedoch, daß dann die für die Klassifizierung ebenfalls relevante absolute Größe der landwirtschaftlichen Betriebe unzureichend berücksichtigt wurde. Ferner ergaben diese Klassifizierungen viele isolierte Betriebe, die sich nur bezüglich eines einzigen Merkmals (z.B. Anzahl der gehaltenen Ziegen) von der Menge der übrigen landwirtschaftlichen Betriebe unterschieden; andererseits wurde die Masse der Betriebe nicht zufriedenstellend in Klassen aufgeteilt.

Auf Grund dieser Erfahrungen wurden aus dem Datenmaterial folgende Merkmale herausgezogen (vgl. dazu im einzelnen die Anlage):

A. Betriebsgröße

4 absolute metrische Merkmale

B. Produktionsrichtung der Betriebe

17 relativierte metrische Merkmale

C. Zusatzangaben über Betriebe

2 relativierte metrische Merkmale
4 Alternativmerkmale

D. Betriebswirtschaftliche Kennzahlen

9 relativierte metrische Merkmale.

Parallel zueinander sind Klassifizierungen der Betriebe anhand folgender Merkmalskombinationen durchgeführt worden:

Klassifizierung	Merkmalsgruppen	Anzahl der Merkmale
I	A,B	21
II	A,B,C	27
III	A,B,C,D	36

33 ERGEBNISSE DER KLASSIFIZIERUNG

331 Vorbereinigung des Materials

Beim Relativieren der Daten stellte es sich heraus, daß für insgesamt 4 Betriebe wegen Widersprüchen im Material keine Verhältniszahlen berechnet werden konnten. Diese 4 Betriebe wurden aus den weiteren Untersuchungen ausgeschieden. Insgesamt wurden also nur 96 Betriebe in die Analyse einbezogen.

Das Klassifizierungsprogramm untersucht in der ersten Arbeitsphase alle Merkmale jeweils daraufhin, ob sie für die n Einheiten mindestens zwei verschiedene Werte aufweisen (also Informationen beitragen) oder ob alle Einheiten bezüglich des Merkmals übereinstimmen. Solche Merkmale wurden automatisch ausgeschieden. Im vorliegenden Material handelte es sich um die Merkmale Nr. 9, 13, 14 und 15. Für die weiteren Arbeitsphasen des Programms werden die reduzierten Merkmals-Nummern zugrundegelegt, die in der Anlage ebenfalls angegeben sind.

332 Klassifizierung I

Nach der Vorbereinigung des Materials werden 17 Merkmale für die Klassifizierung I herangezogen. Das Programm errechnet zunächst die standardisierten Werte und anschließend die Matrix der Korrelationskoeffizienten. Die Eigenwerte dieser Matrix geben wesentliche Aufschlüsse darüber, wieviele Hauptkomponenten bzw. Faktoren für die kompakte Darstellung heranzuziehen sind (vgl. Formel (7) in Abschnitt 232) und wie groß die Anzahl der Klassen im Material etwa sein dürfte (vgl. Abschnitt 26).

Eigenwerte der Korrelationsmatrix
für 17 Merkmale landwirtschaftlicher Betriebe

Nummer des Eigenwertes m	Eigenwert λ_m	Anteil der m Hauptkomponenten an der Varianz $\sum_{k=1}^{m} \lambda_k/17$
1	2,84128	0,167
2	2,70419	0,326
3	1,86585	0,436
4	1,55557	0,528
5	1,38240	0,610
6	1,23820	0,680
7	0,89540	0,734
8	0,84326	0,784
9	0,75164	0,828
10	0,73528	0,871
11	0,57805	0,906
12	0,54314	0,938
13	0,47245	0,965
14	0,24799	0,979
15	0,20796	0,990
16	0,08082	0,997
17	0,05623	1,000

Danach repräsentieren 9 Hauptkomponenten über 82 % (11 Hauptkomponenten mehr als 90 %) der gesamten Variabilität; es ist also sicher nicht zweckmäßig, über 11 Hauptkomponenten hinauszugehen. Der Verlauf der Eigenwerte ist stufenförmig: Nach dem 2., 6. und 10. Wert sind jeweils deutliche Abnahmen zu erkennen. Es ist also zu erwarten, daß mindestens 7 echte Klassen im Material existieren werden. Unechte Klassen, die nur durch eine einzelne Einheit repräsentiert werden, sind dabei nicht mitzuzählen, weil sie auch nicht zur Streuungsmatrix $\underline{W}$ beitragen. Schon bei den ersten Klassifizierungsversuchen zeigte sich, daß unter den 100 landwirtschaftlichen Betrieben etwa 16 bis 18 Betriebe sehr stark abweichende Eigenschaften besitzen und somit als unechte Klassen gesondert nachgewiesen werden. Danach ist die Gesamtzahl der Klassen mit mindestens 7 + 16 = 23 anzunehmen. Die weiteren Arbeiten sind deshalb mit dieser Klassenzahl als Untergrenze und 30 Klassen als Obergrenze vorgenommen worden.

Die Klassifizierung anhand der orthogonalen Faktoren hat zu insgesamt 26 Klassen geführt, deren Besetzung die folgende Tabelle zeigt:

Nummer der Klasse	Anzahl der landwirtschaftlichen Betriebe
1	23
2	20
3	9
4	6
5	6
6	5
7	4
8	3
9	3
10 bis 26	je 1
Zusammen	96

Die reale Deutung der Klassen ist eine wichtige Aufgabe, die von dem Rechenprogramm zwar nicht übernommen, aber dort wenigstens erleichtert werden kann. Die Endausgabe enthält u.a. das Minimum, das Maximum, das arithmetische Mittel und die Standardabweichung der standardisierten - und damit unmittelbar vergleichbaren - Merkmalswerte je Merkmal. Für die mit 23 Einheiten besetzte Klasse 1 ist z.B. folgende Tabelle ausgegeben worden:

Merkmal Nr.	Minimum	Mittelwert	Maximum	Standardabweichung
1	-0,64	-0,55	-0,26	0,09
2	-0,81	-0,37	0,89	0,42
3	-0,32	-0,20	0,08	0,08
4	-0,96	-0,69	0,20	0,27
5	-1,00	0,20	0,97	0,53
6	-0,55	0,85	2,52	0,75
7	-0,67	-0,02	2,63	0,97
8	-0,56	-0,56	-0,56	0,00
9	-0,14	-0,10	0,68	0,17
10	-0,66	0,29	1,30	0,52
11	-0,42	-0,31	1,18	0,34
12	-0,87	0,38	2,04	0,80
13	-0,24	0,92	2,36	0,80
14	-0,19	-0,16	0,33	0,13
15	-0,58	0,38	2,03	0,85
16	-0,30	-0,05	1,47	0,58
17	-0,20	-0,10	1,95	0,45

Berücksichtigt man, daß in der Gesamtheit aller in die Klassifizierung einbezogenen Betriebe der Mittelwert gleich 0 und die Standardabweichung gleich 1 ist, so kann man aus der Tabelle folgende Interpretation für die Klasse 1 ableiten:

Klasse 1: Kleine Betriebe (Merkmal 1) mit überdurchschnittlich starkem Anbau von Knollenfrüchten (Merkmal 6) und verhältnismäßig starkem Besatz mit Jungrindern und Masttieren (Merkmal 13)

Ganz entsprechend sind die übrigen echten Klassen zu deuten:

Klasse 2: Kleine Betriebe mit relativ starkem Getreideanbau

Klasse 3: Mittelgroße Gemüseanbau-Betriebe

Klasse 4: Kleine Betriebe mit hohem Rindviehbesatz

Klasse 5: Mittelgroße Getreideanbau-Betriebe

Klasse 6: Große Betriebe mit hohem Produktionswert durch Anbau von Handelsgewächsen

Klasse 7: Kleine Schweinehaltungs-Betriebe

Klasse 8: Kleine Obstanbau-Betriebe

Klasse 9: Große Gemüseanbau-Betriebe

Bei den unechten Klassen handelt es sich um 7 kleine, 5 mittlere und 5 große bzw. sehr große Betriebe, die sich von den übrigen Betrieben bezüglich eines Merkmals oder einiger Merkmale erheblich unterscheiden und deshalb isoliert worden sind. So gehören z.B. 5 verschieden große Betriebe mit überragender Mastgeflügel-Haltung und/oder Legehennen-Haltung zu dieser Gruppe; auch die Merkmale Handelsgewächse und Gemüseanbau haben zur Aussonderung von 4 Betrieben wesentlich beigetragen.

In einer Paralleluntersuchung sind statt der Faktoren- die Hauptkomponenten - Werte für die Klassifizierung der landwirtschaftlichen Betriebe herangezogen worden. Wie theoretisch zu erwarten, ist das Ergebnis der

Vergleichsrechnung weniger günstig: Bei gleicher Gesamtzahl an Klassen sind allein 4 Klassen gebildet worden, in denen je zwei Betriebe zusammengefaßt sind, die zuvor verschiedenen Klassen zugeordnet wurden.

333 Klassifizierung II

Die 23 Variablen der Klassifizierung II enthalten zusätzlich zwei Anteilswerte (Anteil der Fläche im Eigentum, Anteil der familieneigenen Arbeitskräfte) sowie 4 Alternativmerkmale (vgl. dazu auch die Anlage). Die zusätzlichen Informationen haben bewirkt, daß die bezüglich der Alternativmerkmale ungewöhnlichen Betriebe ausgesondert wurden. Im übrigen hat sich wenig an dem im Abschnitt 332 dargestellten Ergebnis geändert.

334 Klassifizierung III

Das Einbeziehen der betriebswirtschaftlichen Kennzahlen in die Informationsbasis der Klassifizierung III hat wesentliche Aenderungen bewirkt. Insgesamt wurden auf Grund der 32 Merkmale 28 Klassen gebildet, von denen jedoch 20 unecht waren in dem Sinne, daß sie jeweils nur einen Betrieb enthielten. Es handelt sich dabei durchweg um die bereits bei der Klassifizierung I isolierten Betriebe, vermehrt um 3 Einheiten, die wegen der zusätzlichen Merkmale ausgesondert worden sind.

Das Ergebnis für die echten Klassen unterscheidet sich dagegen zum Teil sehr erheblich von den bisher betrachteten Resultaten:

Klasse 1 (36 Betriebe)

Betriebe mittlerer Größe mit durchschnittlichen Eigenschaften.

Klasse 2 (18 Betriebe)

Kleine Betriebe mit verhältnismäßig starkem Einsatz von familieneigenen Arbeitskräften und relativ hohem Anteil an tierischen Zugkrafteinheiten.

Klasse 3 (6 Betriebe)

Mittlere Betriebe mit großem Rindviehbesatz und überdurchschnittlichem Einsatz familieneigener Arbeitskräfte.

Klasse 4 (5 Betriebe)

Kleinbetriebe mit ausgeprägter Schweinehaltung und relativ großem Produktionswert.

Klasse 5 (4 Betriebe)

Relativ große Betriebe, deren Betriebsinhaber keine natürliche Person und/oder nicht Betriebsleiter ist.

Klasse 6 (3 Betriebe)

Kleine Betriebe mit starkem Gemüseanbau.

Klasse 7 (2 Betriebe)

Großbetriebe mit hohem Produktionswert durch starken Anbau von Handelsgewächsen usw.

Klasse 8 (2 Betriebe)

Mittlere Betriebe mit starkem Anbau von Knollenfrüchten und wenig familieneigenen Arbeitskräften.

Einige Klassen (z.B. die hier mit Nr. 4, 6 und 7 bezeichneten) entsprechen weitgehend den bei Klassifizierung I ermittelten Gruppen. Die wesentlichen Unterschiede sind vor allem in folgenden Punkten festzustellen:

a) Die Klassifizierung III ist weniger stark nach der Produktionsrichtung der Betriebe orientiert als die Klassifizierung I.

b) Es ist eine große amorphe Klasse (Nr. 1) entstanden.

c) Wegen der Alternativmerkmale ist eine besondere Klasse (Nr. 5) gebildet worden, die mit den übrigen Klassen schlecht vergleichbar ist.

Die Feststellung unter b) läßt sich durch eine Vergrößerung in der Gesamtzahl der Klassen beheben. Dagegen dürfte der unter c) genannte Nachteil nur dadurch zu beheben sein, daß die Alternativmerkmale entweder durch Gliederung der Gesamtmasse vor der Klassifizierung oder durch Unterteilung der Klassen berücksichtigt werden.

4 ZUSAMMENFASSUNG

Die Studie hat folgendes ergeben:

1. Das für die Klassifizierung herangezogene Verfahren muß in methodischer und praktischer Hinsicht an das vorliegende Material und die Aufgabenstellung angepaßt werden.

2. Für die Klassifizierung landwirtschaftlicher Betriebe ist das verallgemeinerte Verfahren von MAC QUEEN geeignet.

3. Die Klassifizierung wird durch die Auswahl der Merkmale stark beeinflußt; Alternativmerkmale sollten vor oder nach der Klassifizierung nach metrischen Merkmalen berücksichtigt werden.

4. Die Untersuchung eines Probematerials von 100 landwirtschaftlichen Betrieben hat zu einer sachlich plausiblen Einteilung in 7 bzw. 8 echte Klassen geführt.

5. Anhand eines erheblich größeren Materials müßte die Frage untersucht werden, ob die verhältnismäßig große Zahl unechter Klassen wesentlich durch den Umfang des Probematerials bedingt ist.

Anlage
Blatt 1

Merkmale für die Klassifizierung von landwirtschaftlichen Betrieben

Merkmal-Nr. Original	Merkmal-Nr. Reduz.	Beschreibung des Merkmals
		A. Betriebsgröße
1	1	Landwirtschaftlich genutzte Fläche
2	2	Anzahl der Vieheinheiten (VE)
3	3	Anzahl der Jahresarbeitseinheiten (JAE)
4	4	Produktionswert
		B. Produktionsrichtung
5	5	Getreide- und Hülsenfrüchte / Ackerland
6	6	Knollenfrüchte / Ackerland
7	7	Handelsgewächse / Ackerland
8	8	Gemüse, Melonen und Erdbeeren / Ackerland
9	-	Blumen und Zierpflanzen / Ackerland
10	9	Schwarzbrache / Ackerland
11	10	Sonstige Wurzelfrüchte / Landw. genutzte Fläche
12	11	Obstanlagen insgesamt / Landw. genutzte Fläche
13	-	Zitrusanlagen / Landw. genutzte Fläche
14	-	Olivenanlagen / Landw. genutzte Fläche
15	-	Rebanlagen / Landw. genutzte Fläche
16	12	Rinder (ohne Jungrinder u. Masttiere) / Landw. genutzte Fläche
17	13	Jungrinder und Masttiere / Landw. gen. Fläche
18	14	Schafe und Ziegen / Landw. genutzte Fläche
19	15	Schweine / Landw. genutzte Fläche
20	16	Mastgeflügel / Landw. genutzte Fläche
21	17	Legehennen / Landw. genutzte Fläche

Anlage
Blatt 2

Merkmal-Nr. Original	Merkmal-Nr. Reduz.	Beschreibung des Merkmals	
		C. Zusatzangaben über Betriebe	
22	18	Landw. genutzte Fläche im Eigentum / Landw. genutzte Fläche insgesamt	
23	19	Jahresarbeitseinheiten familieneigener Arbeitskräfte / JAE insgesamt	
24	20	Erzeugung für den Verkauf	
25	21	Betriebsinhaber ist natürliche Person	ja = 1 nein = 0
26	22	Betriebsinhaber ist Betriebsleiter	
27	23	Außerbetriebl. Tätigkeit des Betriebsleiters	
		D. Betriebswirtschaftliche Kennzahlen	
28	24	Ackerland / Landw. genutzte Fläche	
29	25	Vieheinheiten / Landw. genutzte Fläche	
30	26	Jahresarbeitseinheiten / Landw. gen. Fläche	
31	27	Zugkrafteinheiten / Landw. genutzte Fläche	
32	28	Produktionswert / Jahresarbeitseinheiten	
33	29	Dauergrünland / Landw. genutzte Fläche	
34	30	Unterglasanlagen / Landw. genutzte Fläche	
35	31	Rauhfutterverzehrende Vieheinheiten / Fläche von Grünfutter auf Ackerland, Dauergrünland und sonstigen Wurzelfrüchten	
36	32	Tierische Zugkrafteinheiten / Zugkrafteinheiten insgesamt	

L I T E R A T U R V E R Z E I C H N I S

T.W. ANDERSON (1958)

An introduction to multivariate Statistical analysis
John Wiley, New York - London

G.H. BALL (1965)

Data analysis in the Social Sciences: What about the details?
Am. Fed. Information Processing Soc. Conference 27, 533 - 560

G.H. BALL / D.J. HALL (1965)

ISODATA, a Novel Method of Data Analysis and Pattern Classification
Stanford Research Institute, Menlo Park, California

G.H. BALL / H.P. FRIEDMAN (1968)

On the status of applications of clustering techniques to behavioral sciences data
Proc. Social Statistics Section, Am. Stat. Ass., Washington 1968

H. BORKO (1965)

Research in Computer Based Classification Systems
P. Atherton (Ed.): "Classification Research, Proceedings of the Second International Study Conference", Copenhagen, 220-257

D.R. COX (1957)

Note on grouping
J.Am.Stat. Ass. 52, 543 - 547

P. DAGNELIE (1966)

A propos des différentes méthodes de classification numérique
Rev. Stat. Appliquée 14, 55 - 75

T. DALENIUS (1950)

The problem of optimum stratification
Skandinavisk Aktuarietidskrift, 203 - 213

T. DALENIUS / M. GURNEY (1951)

The problem of optimum stratification II
Skandinavisk Aktuarietidskrift, 133 - 148

E. FABER (1968)

Automatische Klassifikation von Stichproben mit Hilfe elektronischer Rechenanlagen
unveröffentlicht

R.A. FISHER (1936)

The coefficient of racial likeness
J.R. Anthrop. Inst. 66, 57 -

R.A. FISHER (1938)

The statistical utilization of multiple measurements
Ann. Eugen. 9, 238 - 249

W.D. FISHER (1958)

On grouping for maximum homogeneity
J. Am. Stat. Ass. 53, 789 - 798

E.W. FORGY (1965)

Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications
Biometrics 21, 768 - 769

J.J. FORTIER / H. SOLOMON (1966)

Clustering procedures
in: P.R. KRISHNAIAH "Multivariate Analysis", New York - London, 1966

H.P. FRIEDMAN - J. RUBIN (1967)

On some invariant criteria for grouping data
J. Am. Stat. Ass. 62, 1159 - 1178

J.C. GOWER (1966)

Some distance properties of latent root and vector methods used in multivariate analysis
Biometrika 53, 325 - 338

J.C. GOWER (1967)

A comparison of some methods of cluster analysis
Biometrics 23, 623 - 638

M.J. HAGOOD - E.H. BERNERT (1945)

Component indexes as a basis for stratification in sampling
J. Am. Stat. Ass. 40, 330 - 341

P. IHM (1964)

Methoden der Taxonometrie
in: V.F. SERBANESUM (Ed.), IBM Symposion, Blaricum, Holland, 1964

P. IHM (1965)

Automatic classification in anthropology
in: Dell Hymes (Ed.) The use of computers in antropology
Mouton & Co., Den Haag, 358 - 376

K.J. JONES (1968)

Problems of grouping individuals and the method of modality
Behavioral Sciences 13, 496 - 511

M.G. KENDALL (1966)

Discrimination and classification
in: P.R. KRISHNAIAH (Ed.) "Multivariate Analysis",
New York - London, 1966

B. KING (1967)

Step-wise clustering procedures
J. Am. Stat. Ass. 62, 86 - 101

G.N. LANCE / W.T. WILLIAMS (1966)

Computer programs for hierarchical polythetic classification
The Computer Journal 9, 60 - 64

J. MAC QUEEN (1967)

Some methods for classification and analysis of multivariate observations
Proc. Fifth Berkely Symp. 1, 281 297

P.C. MAHALANOBIS (1936)

On the generalized distance in statistics
Proc. Nat. Inst. Sci. India 12, 49 - 55

R.M. NEEDHAM (1965)

Computer methods for classification and grouping
in: Dell Hymes (Ed.) The use of computers in anthropology,
Mouton & Co, Den Haag, 345 - 356

K. PEARSON (1925)

On the coefficient of racial likeness
Biometrika 18, 105 - 117

C.R. RAO (1952)

Advanced statistical methods in biometric research
New York - London

J. RUBIN (1967)

Optimal classification into groups: An approach for solving the taxonomy problem
J. Theoret. Biol. 15, 103 - 144

P. SCHNELL

Eine Methode zur Auffindung von Gruppen
Biometrische Zeitschrift 6, 47 - 48

P.H.A. SNEATH (1957)

The application of computers in taxonomy
J. Gen. Microbiol. 17, 201 - 226

R.R. SOKAL / P.H.A. SNEATH (1963)

Principles of numerical taxonomy
W.H. Freeman, San Francisco

D. STEINER (1965)

Die Faktorenanalyse - Ein modernes statistisches Hilfsmittel des Geographen für die objektive Raumgliederung und Typenbildung
Geographica Helvetica 20, 20 - 34

R. STONE (1960)

A comparison of the economic structure of regions based on the concept of distance
J. Regional Science 2, 1 - 20

P. SWITZER (1968)

Statistical techniques in clustering and pattern recognition
Proc. Social Statistics Section, Am. Stat. Ass., Washington 1968

R.L. THORNDIKE (1953)

Who belongs in the family?
Psychometrika 18, 267 - 276

R. VAN DEN DRIESSCHE (1965)

La recherche des constellations de groupes a partir des distances généralisées D^2 de MAHALANOBIS
Biometrie - Praximetrie 6, 36 - 47

H.D. VINOD (1969)

Integer programming and the theory of grouping
J. Am. Stat. Ass. 64, 506 - 519

C.S. WALLACE / D.M. BOULTON (1968)

An informations measure for classification
The Computer Journal 1968, 185 - 194

S.S. WILKS (1932)

Certain generalizations in the analysis of variance
Biometrika 24, 471 - 494

S.S. WILKS (1962)

Mathematical Statistics
John Wiley, New York - London

AGRARSTATISTISCHE HAUSMITTEILUNGEN

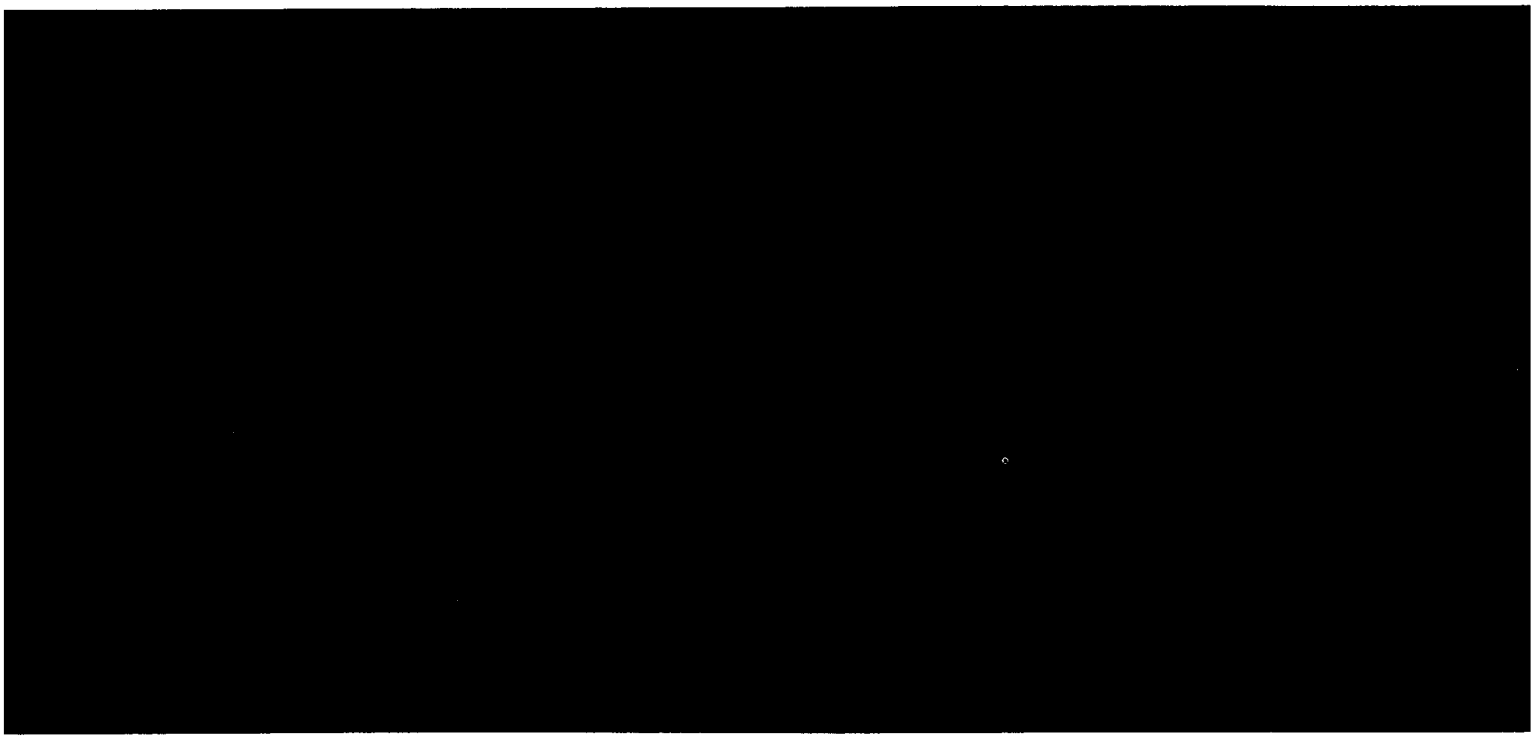
Reihe „Agrarstatistische Studien"

Soweit der Vorrat reicht, werden die Hefte dieser Reihe den an den jeweiligen Themen Interessierten zur Verfügung gestellt. Diesbezügliche Anfragen sind zu richten an: Direktion „Agrarstatistik", Statistisches Amt der Europäischen Gemeinschaften — Postfach 130 — Luxemburg.

		Jahr	Sprachen
Nr. 1	Einfluß der verschiedenen Merkmale des Rinderschlachtkörpers auf seine Preisbestimmung — B.L. DUMONT, J. ARNOUX	1968	F
Nr. 2	Statistische Methoden zur Feststellung der Produktionskapazität der Obstanlagen — G. NEURAY, S. MASSANTE, M. PETRY	1968	D, F
Nr. 3	Die methodischen Probleme bei einer Erhebung der Struktur der Betriebe mit erwerbsmäßigem Anbau von Gartengewächsen — H. STORCK	1968	D, F
Nr. 4	Untersuchung über die Schlachtkörperqualitäten von Rindern in Frankreich — B.L. DUMONT	1969	D, F [1]), N
Nr. 5	Die „Behangdichten-Methode", ein Modell zur Analyse und Prognose von Kernobsterträgen — F. WINTER	1969	D, F
Nr. 6	Die Statistik der Eierpreise in den Mitgliedsländern der EWG — O. STREDLER, H. GOCHT	1969	D, F
Nr. 7	Untersuchung über die Schlachtkörperqualitäten von Rindern in Italien — P.G. BUIATTI	1970	D, F, I
Nr. 8	Modell und Methoden zur Vorausrechnung von Rinderprozessen — H. DIEHL	1970	D, E [2])
Nr. 9	Ein System der Agrarpreisstatistik für die EG — S. GUCKES	1970	D, F
Nr. 10	Klassifizierung landwirtschaftlicher Betriebe mit Hilfe multivariater statistischer Verfahren — K.A. SCHÄFFER		

[1]) Die französische Fassung ist in der Reihe „Statistische Informationen" des Statistischen Amtes der Europäischen Gemeinschaften unter der Nr. 4/1967 veröffentlicht worden.

[2]) Eine sich in Vorbereitung befindende englische Fassung ist nicht für die Veröffentlichung vorgesehen, sie wird nur auf besonderen Wunsch erhältlich sein.



AMT FÜR AMTLICHE VEROFFENTLICHUNGEN DER EUROPÄISCHEN GEMEINSCHAFTEN
Postfach 1003, Luxemburg 1
17516