EUR 1671.d
REPRINT

EUROPÄISCHE ATOMGEMEINSCHAFT - EURATOM

METHODEN DER TAXOMETRIE

von

P. IHM

1964

Gemeinsame Kernforschungsstelle
Forschungsanstalt Ispra - Italien
Zentralstelle für die Verarbeitung wissenschaftlicher Information - CETIS

Sonderdruck aus
TAGUNGSBERICHT DES IBM WORLD TRADE EUROPEAN EDUCATION CENTER
Blaricum - Nederland, November 1962

HINWEIS

Das vorliegende Dokument ist im Rahmen des Forschungsprogramms der Kommission der Europäischen Atomgemeinschaft (EURATOM) ausgearbeitet worden.

Es wird darauf hingewiesen, dass die Euratomkommission, ihre Vertragspartner und alle in deren Namen handelnden Personen :

1° — keine Gewähr dafür übernehmen, dass die in diesem Dokument enthaltenen Informationen richtig und vollständig sind oder dass die Verwendung der in diesem Dokument enthaltenen Informationen oder der in diesem Dokument beschriebenen technischen Anordnungen, Methoden und Verfahren nicht gegen gewerbliche Schutzrechte verstösst;

2° — keine Haftung für die Schäden übernehmen, die infolge der Verwendung der in diesem Dokument enthaltenen Informationen oder der in diesem Dokument beschriebenen technischen Anordnungen, Methoden oder Verfahren entstehen könnten.

EUR 1671.d

REPRINT

METHODEN DER TAXOMETRIE von P. IHM.

Europäische Atomgemeinschaft - EURATOM
Gemeinsame Kernforschungsstelle
Forschungsanstalt Ispra (Italien)
Zentralstelle für die Verarbeitung wissenschaftlicher Information (CETIS)
Sonderdruck aus „Tagungsbericht des IBM World Trade European Education Center" - Blaricum (Nederland), November 1962.

Die Methoden der Taxometrie werden heute überall da angewendet, wo man wegen umfangreichem Material Taxonomie, d.h. die Definition taxonomischer Einheiten und ihre Zusammenfassung zu einem System, mittels elektronischer Rechenautomaten betreiben muss. Nach einer Besprechung des Abstandsbegriffes und einer Definition der Gruppe werden die faktorenanalytische q- und r-Technik diskutiert, die Methode des Gradienten und die Method of Maximum Likelihood. Den Schluss bildet eine Erörterung der auf dem BAYES' schen Theorem beruhenden Methoden.

EUR 1671.d

REPRINT

METHODS OF TAXOMETRY by P. IHM.

European Atomic Energy Community - EURATOM
Joint Nuclear Research Center
Ispra Establishment (Italy)
Scientific Data Processing Center (CETIS)
Reprinted from "Tagungsbericht des IBM World Trade European Education Center" - Blaricum (Nederland), November 1962.

Today, taxometry methods are used wherever it is necessary, on account of the vast amount of material involved, to carry out taxonomy, i.e. the definition of taxonomic units and their reduction to a system, by means of electronic computers. After discussing the distance concept and a definition of "group", the author deals in turn with the factor-analysis q- and r-technique, the gradients method and the method of maximum likelihood. In his conclusion, he comments on the methods based on the Bayes theory.

EUR 1671.d

REPRINT

METHODS OF TAXOMETRY by P. IHM.

European Atomic Energy Community - EURATOM
Joint Nuclear Research Center
Ispra Establishment (Italy)
Scientific Data Processing Center (CETIS)
Reprinted from "Tagungsbericht des IBM World Trade European Education Center" - Blaricum (Nederland), November 1962.

Today, taxometry methods are used wherever it is necessary, on account of the vast amount of material involved, to carry out taxonomy, i.e. the definition of taxonomic units and their reduction to a system, by means of electronic computers. After discussing the distance concept and a definition of "group", the author deals in turn with the factor-analysis q- and r-technique, the gradients method and the method of maximum likelihood. In his conclusion, he comments on the methods based on the Bayes theory.

METHODEN DER TAXOMETRIE

Dr. P. IHM

Euratom CCR
Ispra
Italien

IBM EUROPEAN EDUCATION CENTRE BLARICUM - HOLLAND

ZUSAMMENFASSUNG

Die Methoden der Taxometrie werden heute überall da angewendet, wo man wegen umfangreichem Material Taxonomie, d. h. die Definition taxonomischer Einheiten und ihre Zusammenfassung zu einem System, mittels elektronischer Rechenautomaten betreiben muss. Nach einer Besprechung des Abstandsbegriffes und einer Definition der Gruppe werden die faktorenanalytische q- und r-Technik diskutiert, die Methode des Gradienten und die Method of Maximum Likelihood. Den Schluss bildet eine Erörterung der auf dem BAYESschen Theorem beruhenden Methoden.

I. EINLEITUNG

Während in der Zwanziger- und Dreissigerjahren dieses Jahrhunderts die Grundlagen der Abstandsberechnung zwischen Populationen durch PEARSON (1926) und MAHALANOBIS (1936) und die der Zuteilung eines Individuums unbekannter Herkunft zu einer von mehreren bekannten Populationen durch FISHER (1936) gelegt wurden, ist die systematische Behandlung der Auffindung von Gruppen in einem Material, dessen Gliederung unbekannt ist, relativ neueren Datums. 1953 wendete STROUD (1953) die faktorenanalytische r-Technik zur Analyse der Systematik von Kalotermes an, später benutzten SOKAL (1958) auf der einen und DRIVER und SCHUESSLER (1957) auf der anderen Seite die q-Technik zu dem gleichen Zweck. Während es sich hier um linear-algebraische Methoden handelt, versuchten andere (HILL 1959, SNEATH 1962, SILVESTRI, TURRI, HILL und GILARDI 1962, TANIMOTO 1958) währenddessen Abstände zwischen den zu klassifizierenden Individuen zu definieren und sie zu Gruppen benachbarter zusammenzufassen. Während bei den faktorenanalytischen Methoden die Individuen Elemente eines normierten Raumes sind, genügt es hier, sie lediglich als die eines metrischen Raumes aufzufassen.

Der Zweck der vorliegenden Beitrages ist, die von uns benutzten und vorgesehenen Methoden sowie die Resultate einiger praktischer Untersuchungen vorzuführen. Gewisse Ergebnisse (BORKO 1962) zeigen, dass die Methoden auch in der Dokumentation von Wert sein können.

II. DIE METRIK UND DER BEGRIFF DER GRUPPE

Wir betrachten eine Menge von Individuen, die wir zu Gruppen zusammenfassen wollen. Diese Individuen bilden Elemente eines metrischen Raumes, d.h. es ist ein Abstand d_{ij} zwischen dem i-ten und j-ten Individuum definiert. In Analogie zum intuitiven Verhalten des Systematikers wird man Individuen zu einer Gruppe zusammenfassen, die untereinander einen geringeren Abstand haben als zu den Individuen anderer Gruppen. Was in der klassischen Systematik oder Taxonomie Intuition ist, soll in der neuen Taxometrie durch objektive Rechenverfahren ersetzt werden. Aus diesem Grunde ist es nötig, den intuitiven Abstandsbegriff durch einen mathematischen zu ersetzen. Praktisch wurde hierzu bisher so vorgegangen, dass die Individuen als Elemente eines nicht notwendig linearen Vektorraumes dargestellt wurden. In diesem Raume wurden Verknüpfungsregeln und eine Metrik definiert, die im allgemeinen (Ausnahme: TANIMOTO 1958) den üblichen Axiomen des metrischen Raumes genügt.

Verschiedenste Methoden und Verhahren, die Gruppen aufzufinden, existieren. Bevor wir sie aber in ihrer Bedeutung untersuchen, haben wir uns zu fragen, nach welchen Gesichtspunkten der Mensch die Metrik, das heisst mittelbar den Verwandschaftsbegriff, festlegt. In den häufigen Fällen typologischer Systeme sind die taxonomischen Einheiten Typen, die gewissen Prinzipien genügen, wir brauchen nur an botanische Systeme mit Gliederung nach essbaren, nicht essbaren, Land- und Wasserpflanzen usw. zu denken oder an die Wortklassen z. B. im Bantu. Es ist aber bekannt, dass die allgemeine Tendenz zu natürlichen Systemen führt, d.h. zu solchen, in denen die Angehörigen der Taxa verschiedene Niveaus in genetischem, speziell phylogenetischen Zusammenhang stehen. Es ist daher nötig, dass in der Taxometrie Abstände so definiert werden, dass sie diese genetischen Zusammenhänge aufdecken. Ein gutes Beispiel hierfür bildet SWADESHs Bemühen, ein Verwandtschaftsmass lexikologischer Art für die Klassifizierung von Sprachen zu finden. Dieser Autor glaubte, in der 'retention rate', d.h. der Proportion der für zwei Sprachen gemeinsamen Wörter, eines gefunden zu haben, welches Funktion der Zeit seit der Trennung dieser beiden Sprachen ist. Wir wissen heute, dass sein Ansatz zu einfach war, um Resultate zu liefern, die die Mehrheit der Linguisten befriedigen (vgl. BERGSLUND u. VOGT 1962). Ein weiteres Beispiel ist MAHALANOBIS' verallgemeinerter Abstand, der im Sinne eines Diffusionsmodelles interpretiert werden kann. Im übrigen scheinen Abstände recht willkürlich definiert worden zu sein. Die meisten auf Abstände beruhenden Methoden arbeiten auch dann noch, wenn man eine andere Abstandsfunktion einführt - die Schwierigkeit liegt aber nicht auf dem Gebiet der Gruppentrennung, wenn man eine Abstandsfunktion hat, sondern in deren sinnvoller Festsetzung.

Zur Definition einer brauchbaren Metrik braucht man ein Modell. Wir verwenden bei unseren Methoden ein auf dem Diffusionsprinzip beruhendes Modell, bei dem die bedingte Verteilung der niederen taxonomischen Einheiten in einer höheren die n-dimensionale Normalverteilung ist. In vielen praktischen Fällen lässt sich das Material so transformieren, dass mit diesem Modell gearbeitet werden kann. Wir bedienen uns linearer Methoden, die nicht auf der direkten Benutzung eines Abstandes beruhen, und haben die Metrik implizit nur insofern, als wir in einem normierten Raum arbeiten.

III. FAKTORENANALYTISCHE METHODEN

Die faktorenanalytischen Methoden beruhen auf einer Analyse der Kovarianz- oder der Korrelationsmatrix. Bei der r-Methode werden die Korrelationen zwischen den Variablen, bei der q-Methode zwischen den Individuen berechnet. Die q-Methode wird in der Taxometrie hauptsächlich von SOKAL und Mitarbeitern angewendet (vgl. z. B. ROHLFS u. SOKAL 1961), doch benutzten sie auch DRIVER und SCHUESSLER (1957), um kalifornische Indianerstäm-

me zu klassifizieren. Die r-Methode wurde von STROUD (1953) zuerst für den Termitengenus Kalotermes angewandt, ebenso benutzten sie BORKO (1962) und wir selbst.

Die q-Methode hat folgende geometrische Interpretation: Die Individuen werden durch n-dimensionale Vektoren $\underline{x}_k$ repräsentiert. Dieser Vektorraum wird auf die Ebene

$$\sum_{i=1}^{n} x_i = 0$$

projiziert, wobei jedem $\underline{x}_k$ ein (n-1)-dimensionales $\underline{y}_k$ entspricht. Die genannten Korrelationen r_{ij} zwischen den Individuen I_i und I_j, die die Elemente der q-Korrelationsmatrix sind, sind die Kosinus der Winkel zwischen $\underline{y}_i$ und $\underline{y}_j$. Bilden Punkte Gruppen, so sind in den meisten Fällen die Winkel zwischen ihnen klein, sont gross. Eine direkte Inspektion der Korrelationskoeffizienten, wie sie SOKAL und MICHENER (1958) durchführten, erlaubt das Auffinden der Gruppen. Es ist jedoch auch möglich, eine HOTELLINGsche Hauptachsenanalyse durchzuführen. Nach eventueller Drehung lassen sich die Hauptfaktoren als taxonomische Einheiten interpretieren.

Gegen die Methode können Einwände vorgebracht werden. Abb. 1 zeigt einen Fall, in dem nur drei statt vier Gruppen gefunden würden.

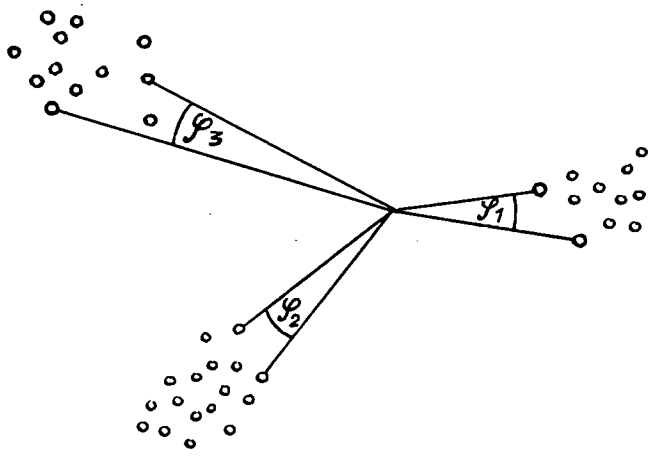

Abb. 1. Fälschliche Trennung in drei Gruppen bei der q-Methode. Drei Winkel sind als Beispiele eingezeichnet.

Allerdings lässt sich dieser Fehler durch gesonderte Untersuchung einer jeden Gruppe leicht korrigieren. Drückt man die Vektoren in Polarkoordinaten aus, so bedeutet die Methode, dass nur die Winkel, nicht aber der Betrag als Information behalten werden. Hatten die Vektoren eine n-dimensionale Normalverteilung, so haben die Winkel eine andere Verteilung was die Verwendung von auf der Normalverteilung beruhenden Methoden ausschliesst.

Die r-Methode besteht geometrisch in einer Projektion der Punkte $\underline{x}_k$ auf eine p-dimensionale Hyperebene, wobei jedem $\underline{x}_k$ ein Bild $\underline{y}_k$ entspricht. Die $\underline{y}_k$ sind durch

$$\underline{y}_k = \underline{U}'\underline{x}_k$$

gegeben, wobei die Zeilen der pxn-Matrix $\underline{U}'$, $\underline{u}_i'$, die Lösungen von

$$\underline{Ru} = \lambda \underline{u}$$

sind, für die die λ_i die p grössten Werte annehmen. $\underline{R}$ ist die Kovarianz- oder Korrelationsmatrix, je nachdem mit welcher man den Unständen entsprechend arbeiten möchte.

Es handelt sich hier um die Projektion auf die p Achsen der grössten Varianz. Ist die Binnengruppenkovarianzmatrix bis auf einen Faktor die Einheitsmatrix und gibt es g Gruppen, so ist für p=g-1 die Hyperebene durch die g Gruppenschwerpunkte aufgespannt. Die wesentliche Idee der faktorenanalytischen Behandlung ist aber die, dass p praktisch kleiner als g-1, unter Umständen sogar viel kleiner als g-1 ist. Die von uns verwendeten Methoden bestehen in der Projektion auf eine dreidimensionale Hyperebene, d.h. einen dreidimensionalen Raum. Dort werden zunächst visuell Gruppen provisorisch getrennt und diese erneut der Behandlung unterzogen. Dies kann an Beispielen veranschaulicht werden (siehe dort). Ich werde später zeigen, wie die visuelle Einteilung objektiviert werden kann, doch vorerst noch auf einen Schritt zur Verbesserung eingehen. Wie eingangs erwähnt, ist der MAHALANOBISsche Abstand in vielen Fällen ein vernünftiges Mass für die Verwandschaft. Es ist daher zweckmässig, eine lineare Transformation der Variablen durchzuführen, die den Abstand zwischen den Gruppenschwerpunkten zum MAHALANOBISschen Abstand macht. Sei $\underline{B}$ die Binnengruppenkovarianzmatrix und $\underline{V}$ die Matrix ihrer normalisierten Eigenvektoren, $\underline{L}$ die Diagonalmatrix der zugehörigen Eigenwerte, dann wird bei Nichtsingularit von $\underline{L}$ der $\underline{x}$-raum auf sich selbst vermöge

$$3.1 \qquad \underline{x}^* := \underline{L}^{-1/2}\underline{V}'\underline{x}$$

abgebildet. Es ist dann

$$\left|\underline{x}_i^* - \underline{x}_j^*\right|^2 = (\underline{x}_i - \underline{x}_j)'\,\underline{VL}^{-1}\underline{V}'\,(\underline{x}_i - \underline{x}_j)$$

$$= (\underline{x}_i - \underline{x}_j)'\,\underline{B}^{-1}(\underline{x}_i - \underline{x}_j),$$

der MAHALANOBISsche Abstand. Nach dieser Transformation wird das Verfahren mit den $\underline{x}^*$ wiederholt.

Ein Einwand gegen die Methode ist, dass die visuelle Beurteilung eine Beschränkung auf zwei, höchstens drei Dimensionen verlangt. Damit kann ein Informationsverlust verbunden sein. Durch die im nächsten Kapitel beschriebenen Methoden wird dieser Einwand aber weitgehend ausgeschaltet. SNEATH bemühte sich aufgrund der von ihm definierten Abstände, ein dreidimensionales Modell

zu konstruieren. Er übersah dabei, dass ein abstrakter metrischer Raum nur unter gewissen Bedingungen auf einen EUKLIDischen abstandsgetreu abgebildet werden kann. Die hier beschriebene Projektion hätte ihm hingegen die beste Darstellung im dreidimensionalen Raum geliefert.

Ein Vorteil der Projektion ist, dass sie erlaubt, Hypothesen über die Dispersion in den Gruppen visuell zu prüfen, z. B. ob in allen Gruppen die gleiche Kovarianzmatrix vorliegt, ob Ausreisser vorkommen usw. Ich fasse im übrigen die faktorenanalytische r-Technik aber nur als vorbereitendes Hilfsmittel auf.

IV. EINE METHODE ZUR GLÄTTUNG DER VERTEILUNGEN

Wie die Beispiele gezeigt haben, kommt es bei der visuellen Trennung auf die Identifikation von Punkthäufungszentren an. Hierfür wurde von SCHNELL ein Verfahren entwickelt. Sei eine Stichprobe von N Punkten $\underline{x}_k$ gegeben. Wir betrachten die Funktion

$$f(\underline{x}, \sigma^2) := \sum_{k=1}^{N} e^{-\frac{1}{2\sigma^2}(\underline{x}-\underline{x}_k)'(\underline{x}-\underline{x}_k)}$$

die bis auf einen Faktor die Summe GAUSZscher Dichten mit Erwartungsvektoren $\underline{x}_k$ und Kovarianzmatrix $\sigma^2 I$ ist. Sind alle $\underline{x}_k$ verschieden, so hat $f(\underline{x}, \sigma^2)$ für genügend kleines σ^2 genau N Maxima, ist dagegen σ^2 genügend gross, so nur ein einziges. Dazwischen liegen Werte von σ^2, für die $f(\underline{x}, \sigma^2)$ 1, 2, ... Maxima hat. Für gegebenes σ^2 werden die Maxima mittels der Methode des Gradienten gesucht, ausgehend von jedem Punkte $\underline{x}_k$. Alle $\underline{x}_k$, von denen aus man zum gleichen Maximum kommt, werden als zur gleichen Gruppe gehörig betrachtet. Durch geeignete Programmierung kann das Verfahren erheblich beschleunigt werden. Ist N, die Zahl der Punkte, klein, so hat das Verfahren die Tendenz, Einpunktgruppen zu liefern oder aber nur eine Hauptgruppe. Ist N dagegen gross , so gibt es auch noch bei kleinerem umfangreichere Gruppen. Dies entspricht unserer intuitiven Auffassung von der in der Stichprobe enthaltenen Information: Ist ihr Umfang klein, so können wir wenig verbindliches über Gruppenbildung aussagen und umgekehrt.

Wir haben das Verfahren so ausgebaut, dass zuerst die faktorenanalytische r-Methode auf die gesamte Kovarianz- oder Korrelationsmatrix angewendet wird, wobei auf eine bis 5-dimensionale Hyperebene projiziert wird, worauf mittels SCHNELLs Methode Punkthäufungen gesucht werden. Eine Gruppe wird dabei durch Verkleinerung von σ^2 noch unterteilt, wenn die Teile sich noch als signifikant verschieden erweisen. Allerdings ist die Testmethode mangels

exakter Verfahren recht Heuristisch; PITMANs Test kann verwendet werden, ist aber auch für eine schnelle Rechenanlage recht aufwendig. Für die nun gefundenen Gruppen wird die Binnengruppenkovarianzmatrix berechnet und die Transformation (3.1) durchgeführt, die Projektion wiederholt, SCHNELLs Verfahren angewandt usw. bis Stabilität des Ergebnisses eintritt.

V. DIE METHODE DER MAXIMALEN LIKELIHOOD

Im Vorausgehenden sind keine expliziten Annahmen über die Verteilung der $\underline{x}$ in den einzelnen Gruppen gemacht worden, lediglich implizit, dass die bedingte Kovarianzmatrix für alle Gruppen gleich ist. Das ist häufig nicht der Fall, ausserdem interessiert man sich häufig für die Angabe einer Wahrscheinlichkeit dafür, dass ein bestimmtes Individuum zu dieser oder jener Gruppe gehört. Dies ist nicht möglich ohne eine konkrete Annahme über die Verteilung. Wir nehmen daher an, dass die bedingte Dichte von $\underline{x}$ in der i-ten Gruppe

$$5.1 \quad f_i(\underline{x}) := f(\underline{x};\underline{m}_i,\underline{C}_i) := \frac{1}{(2\pi)^{n/2}\,|\underline{C}_i|^{1/2}}\, e^{-\frac{1}{2}(\underline{x}-\underline{m}_i)'\,\underline{C}_i^{-1}(\underline{x}-\underline{m}_i)}$$

ist, die totale Dichte für g Gruppen

$$5.2 \quad f(\underline{x}) := \sum_{i=1}^{g} a_i\, f_i(\underline{x})$$

mit

$$5.3 \quad \sum_{i=1}^{g} a_i = 1.$$

Das heisst, dass die bedingte Verteilung in der i-ten Gruppe eine n-dimensionale Normalverteilung mit Erwartungsvektor $\underline{m}_i$ und Kovarianzmatrix $\underline{C}_i$ ist, die totale Verteilung die Überlagerung derartiger Verteilungen mit den Anteilen a_i, d.h. der Wahrscheinlichkeit, dass $\underline{x}$ zur i-ten Gruppe gehört. Die Aufgabe lautet, die Parameter von (5.2) zu schätzen. Dafür kann die Methode der maximalen Likelihood verwendet werden. Wir erhalten für die Elemente t_{ir} von $\underline{m}_i$, $\underline{C}_i$ die Bestimmungsgleichungen

$$0 = \frac{\delta \log L}{\delta t_{ir}} := \frac{\delta}{\delta t_{ir}} \sum_{k=1}^{N} \log f(\underline{x}_k)$$

5.4
$$= \sum_{k=1}^{N} p_i(\underline{x}_k) \frac{1}{f_i(\underline{x}_k)} \frac{\delta}{\delta t_{ir}} f_i(\underline{x}_k)$$

$$= \sum_{k=1}^{N} p_i(\underline{x}_k) \frac{\delta}{\delta t_{ir}} \log f_i(\underline{x}_k)$$

mit

5.5
$$p_{ik} = \frac{a_i f_i(\underline{x}_k)}{f(\underline{x}_k)} .$$

(5.5) ist die a-posteriori-Wahrscheinlichkeit der Hypothese der i-ten Gruppe bei gegebenem $\underline{x}_k$, wie sie sich aus dem BAYESschen Theorem ergibt. Eine Betrachtung von (5.4) zeigt, dass wir die üblichen Schätzformeln für $\underline{m}_i$ und $\underline{C}_i$ erhalten, jedoch mit Gewichten $p_i(\underline{x}_k)$, d.h.

5.6
$$\underline{m}_i = \frac{1}{P_i} \sum_{k=1}^{N} p_i(\underline{x}_k) \underline{x}_k =: \bar{\underline{x}}_i$$

5.7
$$\underline{C}_i = \frac{1}{P_i} \sum_{k=1}^{N} \left\{ p_i(\underline{x}_k) \underline{x}_k \underline{x}_k' - P_i \bar{\underline{x}}_i \bar{\underline{x}}_i' \right\}$$

wobei

$$P_i = \sum_{k=1}^{N} p_i(\underline{x}_k) .$$

Die a_i berechnen sich unter Berücksichtigung von (5.3) bei Verwendung des LAGRANGEschen Multiplikators λ gemäss

$$0 = \frac{\delta L}{\delta a_i} = \frac{\delta}{\delta a_i} \sum_{k=1}^{N} \log f(\underline{x}_k) - \lambda \frac{\delta}{\delta a_i} \sum_{j=1}^{g} a_j$$

$$= \sum_{k=1}^{N} p_i(\underline{x}_k) \frac{1}{a_i} - \lambda .$$

Daraus ergibt sich

$$\lambda a_i = P_i$$

und, damit (5.3) erfüllt ist,

$$\lambda = \sum_{j=1}^{g} P_i =: P$$

folglich

$$5.8 \quad a_i = \frac{P_i}{p} .$$

Wie sich aus (5.5) ergibt enthalten die Gewichte $p_i(\underline{x}_k)$ die zu bestimmenden Parameter, d.h. wir haben in (5.6), (5.7) und (5.8) Gleichungen, die die Schätzwerte implizit enthalten. Praktisch lassen sie sich dadurch lösen, dass provisorische Gewichte $p_i(\underline{x}_k)$ berechnet werden, die in die Gleichungen eingesetzt werden, um neue Werte zu erhalten, die neue Gewichte ergeben usw. Konvergenz tritt nur ein, wenn die ersten Näherungswerte relativ nahe bei den wahren Werten liegen, man muss also schon zu Anfang recht genaue Werte haben. Um sie zu erhalten, kann man die Methoden der Abschnitte 3 und 4 verwenden. Praktische Untersuchungen, die zusammen mit H. Fangmeyer aufgeführt werden, werden an anderer Stelle veröffentlicht.

Ein Nachteil des Verfahrens ist, dass man die Hypothese der Normalität der Verteilungen machen muss. Das ist jedoch nicht so gravierend, da bei grossem n und relativ kleinem p, das ist die Dimension der Hyperebene, durch die Projektion erhaltenen Variablen als Linearkombinationen eine Tendenz zur Normalität aufweisen. Die Konvergenz ist fraglich, wenn sich die Gruppen stark überschneiden.

VI. DAS BAYESSCHE THEOREM

Die in den vorangegangenen Kapiteln beschriebenen Methoden sind mehr oder weniger heuristisch und fassen das zusammen, was bis heute erarbeitet wurde. Sie sind aber vom theoretischen Standpunkt nicht voll befriedigend. So ist noch keine Methode ausgearbeitet worden, die eine Gruppeneinteilung gegen eine andere zu testen erlaubt usw. Meines Erachtens kann hier mittels des BAYESschen Theorems unter Verwendung einer Nutzenfunktion eine bedeutende Verbesserung erreicht werden. Wie im vorhergehenden Kapitel muss man die

wesentliche Annahme machen, dass es sich bei den N Vektoren $\underline{x}_k$ um eine Stichprobe handelt, d.h. dass die $\underline{x}_k$ überhaupt eine Wahrscheinlichkeitsverteilung haben. Das ist selbstverständlich, wenn die Merkmale von Individuum zu Individuum schwanken können, z. B. Länge und Breite von Organen usw. Ist das "Individuum" aber eine taxonomische Einheit, so, ist das begrifflich schwieriger. Die Pflanzenart Iris versicolor ist dann nur eine zufällige Realisation unter vielen anderen. Was wir mittels der Methode der maximalen Likelihood oder mittels des BAYESschen Theorems wahrscheinlich machen, sind also hypothetische taxonomische Einheiten oder lokale Selektionsoptima.

Wir betrachten nun die Menge aller möglichen Systeme. Auf dieser sei ein Wahrscheinlichkeitsmass definiert, ebenso ein zweites Mass, der Nutzen. Da die Systeme im Falle des Modelles von Abschnitt 5 durch die Parameter $\underline{m}_i$, $\underline{C}_i$, und a_i und g definiert werden können, können wir im EUKLIDischen Parameterraum H den Nutzen als Punktfunktion verwenden, sowie die Wahrscheinlichkeitsdichte. Die a-priori-Wahrscheinlichkeit des Systemes hängt von dem Vorwissen ab, der Nutzen kann gegeben sein durch den Umstand, dass ein System mit unscharf getrennten Gruppen für die spätere Bestimmung von Individuen unbekannter Herkunft nicht viel taugt, ob die Definition der taxonomischen Einheiten aufwendig ist oder nicht usw. Sie hängt von Einzelfall ab, und es ist nicht möglich, ein allgemeingültiges Rezept zu geben.
Sei ein Punkt im Parameterraum, der einem System entspricht, durch $\underline{h}$ bezeichnet und sei $dP(\underline{h})$ die a-priori-Wahrscheinlichkeitsverteilung von $\underline{h}$, $n(\underline{h})$ der Nutzen von $\underline{h}$. Nach Erhalt der N Vektoren $\underline{x}_k$ ist die Likelihood von $\underline{h}$ nach (5.2)

$$l(\underline{h};X) := \prod_{k=1}^{N} f(\underline{x}_k) =: \prod_{k=1}^{N} v(\underline{h};\underline{x}_k),$$

wobei X die Stichprobe $\left\{\underline{x}_1,\ldots\underline{x}_N\right\}$ bezeichnet.

Die a-posteriori-Wahrscheinlichkeitsverteilung wird jetzt nach dem BAYESschen Theorem

$$6.1 \qquad dP^*(\underline{h};X) := \frac{l(\underline{h};X)dP(\underline{h})}{\int_H l(\underline{h};X)dP(\underline{h})}$$

Soll der Nutzen eines Systems S ausgerechnet werden, so müssen wir uns vor Augen halten, dass ein definitiv aufgestelltes System, mit dem wir arbeiten, ein Idealsystem S_I, nicht einem Element $\underline{h}$ aus H entspricht sondern einer Teilmenge H_I. Dies rührt daher, dass viele $\underline{h}$ ein gleichideales System ergeben, z. B. können die Parameter in gewissen Grenzen variieren, ohne das System in seiner Brauchbarkeit wesentlich zu verändern. Daher ist der relative Nutzen von S_I, $N(S_I)$, durch

$$N(S_I) = \frac{\int_{H_I} n(h)dP^*(\underline{h};X)}{\int_H n(\underline{h})dP^*(\underline{h};X)}$$

gegeben.

Es ist interessant, dass biologische Erkenntnisse die Angabe einer Wahrscheinlichkeit für die a_i ermöglichen. FISHER, CORBET und WILLIAMS (1942) fanden, dass die Zahl r der Arten pro Genus (und allgemein die Zahl niederer taxonomischer Einheiten in höheren) die Verteilung

$$w(r) := \frac{1}{|\log(1-\eta)|} \frac{\eta^r}{r}$$

hat, eine Tatsache, die durch WETTE (1959) eine theoretische Begründung fand. In manchen Fällen liegt nach WETTE auch eine negative Binomialverteilung vor. Anstelle der Likelihood der a_i hat man die des Parameters η.

Hinsichtlich der Anwendung der auf dem BAYESschen Theorems gegebenen Methoden ist noch nichts geschehen. Einmal ist die Berechnung nicht einfach und verlangt kombinatorische Methoden, dann verhindert die Abneigung vieler Wissenschaftler gegen das Theorem dessen Anwendung. Der neuerliche Durchbruch des Subjektivismus und des Dualismus, wie er z. B. von CARNAP (1950) und RICHTER (1954) verfochten wird, dürfte die Übernahme BAYESscher Methoden bewirken. Ich selbst interpretiere nach RICHTER (6. 1) in der Weise, dass $l(\underline{h};X)$ eine von einer objektiven Wahrscheinlichkeitsdichte hergeleitete Likelihood, dP eine subjektive Wahrscheinlichkeitsverteilung ist.

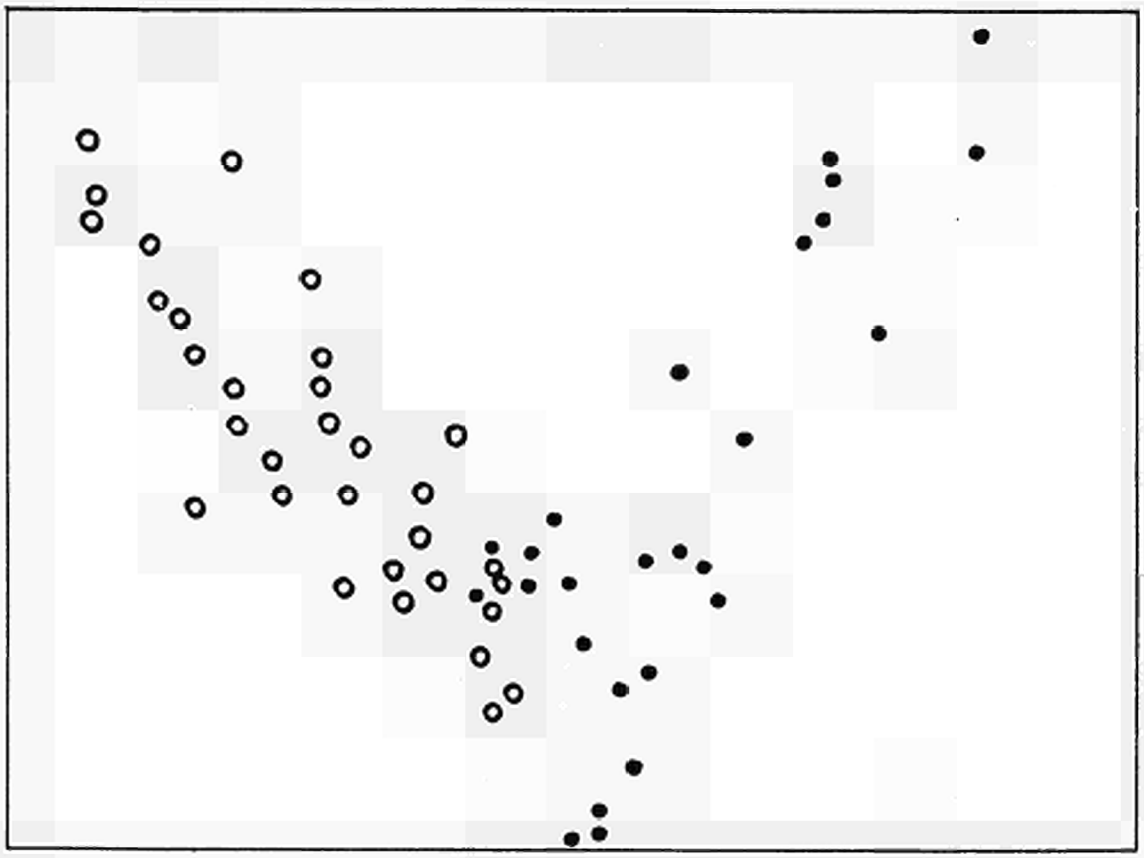

Abb. 2.

Beispiel: Das Wort "Plasma" wurde in 60 Fällen auf seine Umgebung von Vollwörtern (mots pleins d.h. Verben, Substantive, Adjektive und Adverbien) untersucht, wobei die zehn Wörter vor und die zehn Wörter nach "Plasma" untersucht wurden, "Plasma" in 30 Fällen aus der Biologie, in 30 aus der Physik. Die Wörter wurden durch Zusammenlegen von Synonymen, Flexionsformen usw. auf 199 reduziert. Jedes Wort erhielt eine Nummer (1,2...,199), und der Vektor $\underline{x}_k$ hatte die i-te Komponente gleich eins, wenn das Wort Nummer i in der Umgebung des k-ten Vorkommens von "Plasma" war, andernfalls null. Die Resultate der Projektion auf eine Ebene sind in Abb. 2, für jede Gruppe in verschiedener Weise, dargestellt. Man sieht, dass eine visuelle Gruppentrennung sehr wohl die echten Gruppen ergeben hätte.
Weitere Beispiele werden an andere Stellen veröffentlicht.

BIBLIOGRAPHIE

1. BERGSLAND K. u. H. VOGT 1962, On the validity of glottochronology. (With comments), Current Anthropology 3, 115-153.

2. BORKO H. 1962, The construction of an empirically based mathematically derived classification System.
Proc. Spring Joint Computer Conf., 21:279-289.

3. CARNAP R. 1950, Logical foundations of probability.
Chicago 1950.

4. DRIVER H.E. u. K.F. SCHUESSLER, 1957, Factor analysis of ethnographic data. Amer. Anthropologist, 59 , 655-663.

5. FISHER R.A., 1936, The use of multiple measurements in taxonomic problems. Ann. Eng. 7, 179.

6. FISHER R.A., A.S. CORBET u. C.B. WILLIAMS 1942.
The relation between the number of species and the number of individuals in a random sample of an animal population.
J. Animal Ecology 11, 42-58.

7. HILL L.R., 1959. The Adansonian classification and taxonomy.
J. Gen. Microbiol. 6, 318.

8. MAHALANOBIS P.C., 1936, On the generalised distance in statistics.
Proc. Nat. Inst. Sci. Ind., 12, 49.

9. PEARSON K., 1926, On the coefficient of racial likeness.
Biometrika, 18, 105.

10. RICHTER H., 1954, Zur Grundlegung der Wahrscheinlichkeitstheorie: V. Indirekte Theorie.
Math. Annalen 128: 305-339.

11. ROHLFS F.J. u. R.R. SOKAL, 1962. The description of taxonomic relationships by factor analysis.
Syst. Zool. 11, 1-16.

12. SILVESTRI L., M. TURRI, L.R. HILL u. E. GILARDI 1962. A quantitative approach to the systematics of Actinomyces based on overall similarities. "Microbiol. Classification". XII Symp. Soc. Gen. Microbiol. (Cambridge U. Press), 333-360.

13. SNEATH, P.H.A., 1962. The construction of taxonomic groups.
"Microbial Classification". XII Symp. Soc.Gen. Microbiol. (Cambridge U. Press), 289-332.

14. SOKAL R.R. u. C.D. MICHENER, 1958. A statistical method for evaluating systematic relationships.
Univ. Kansas Sci. Bull., 38, 1409-1438.

15. SOKAL R.R., 1958. Quantification of systematic relationships and of phylogenetic trends.
Proc. C. Internat. Congr. Entomol. 1, 409-415.

16. STROUD C.P., 1953. An application of factor analysis to the systematics of Kalotermes.
Syst. Zool. 2, 76-92.

17. TANIMOTO T.T., 1958. An elementary mathematical theory of classification and prediction.
Publ. IBM, New York, 1958.

18. WETTE R., 1959. Zur biomathematischen Begründung der Verteilung der Elemente taxonomischer Einheiten des natürlichen Systems in einer logarithmischen Reihe.
Biometrische Zeitschrift 1, 44-50.

CDNA01671DEC