Über den Gebrauch mathematisch-statistischer Methoden in der Taxonomie

Autor(en): **Huber, Hans**

Objekttyp: **Article**

Zeitschrift: **Berichte der Schweizerischen Botanischen Gesellschaft = Bulletin de la Société Botanique Suisse**

Band (Jahr): **89 (1979)**

Heft 3-4

PDF erstellt am: **21.09.2024**

Persistenter Link: https://doi.org/10.5169/seals-63120

Ber. Schweiz. Bot. Ges. *89* (3/4), 227–250 (1979)

Über den Gebrauch mathematisch-statistischer Methoden in der Taxonomie

von *Hans Huber*

Manuskript eingegangen am 28. Juni 1979

Einleitung

Da die taxonomischen Einheiten durchwegs Kollektive variabler Objekte sind, wäre die Verwendung biometrisch-statistischer Verfahren zur Beschreibung und Unterscheidung dieser Kollektive durchaus am Platz. Seit R.A.Fisher im Jahre 1936 seine „Discriminant Functions" eingeführt hat, sind zwar eine ganze Reihe von weiteren Verfahren beschrieben worden, welche für Probleme der Klassifikation variabler Objekte brauchbar sind. Trotzdem werden diese Verfahren von den Taxonomen nur selten angewendet, und in den wenigen Fällen, in denen dies doch geschieht, werden dann leider oft statistische Methoden gewählt, welche entweder dem Problem nicht angepasst sind oder die vorhandene Information nur sehr schlecht ausnützen. Durch die geringe Aussagekraft der damit gewonnenen Resultate wird dann der Eindruck erweckt, dass statistische Methoden nicht viel zur Lösung der Probleme der beschreibenden Taxonomie beitragen können.

Für diesen Sachverhalt dürften unter anderem folgende Gründe verantwortlich sein: 1. Die Kenntnis geeigneter Methoden dringt bei der heutigen Spezialisierung der Wissenschaften nur schwer von einem Fachgebiet in ein anderes über. Erschwerend wirkt hierbei die nur den Fachgenossen geläufige Fachsprache mit. 2. Zum Verständnis der einschlägigen Publikationen werden meist ziemlich fortgeschrittene mathematische Kenntnisse vorausgesetzt. 3. Die Anwendung der Verfahren ist meist mit einem recht hohen Rechenaufwand verbunden. 4. Formunterschiede lassen sich oft nur schwer numerisch erfassen.

In der vorliegenden Arbeit soll zunächst durch eine Übersicht über die für taxonomische Probleme brauchbaren biometrisch-statistischen Methoden der Zugang zur einschlägigen Literatur erleichtert werden. Sodann soll ein graphisches Verfahren beschrieben werden, welches mit geringem Rechenaufwand erlaubt, Gliederungen in einem komplexen Formenkreis zu erkennen.

Literaturübersicht

Bei taxonomischen Untersuchungen wird jeweilen an jedem untersuchten Exemplar eine Mehrzahl von Merkmalen beobachtet. Um derartige komplexe Beobachtungen statistisch zu verarbeiten, muss man spezielle Verfahren anwenden, die in elementaren Einführungen in die statistische Methodik höchstens am Rande behandelt werden (siehe z.B. Linder, 1951, Kap. 6). Diese Verfahren werden unter der Bezeichnung „multivariate Verfahren" zusammengefasst.

Zur Einführung in diese Methoden kommen für den Biologen die Werke von Kramer (1972), Cooley & Lohnes (1962), Quenouille (1952), und Seal (1964) in Betracht. Für ein tieferes Eindringen sind die Werke von Anderson (1958), Morrison (1967), Rao (1952, 1965) beizuziehen; von diesen Büchern setzt dasjenige von Morrison am wenigsten Vorkenntnisse voraus. Eine Einführung in die etwas andere Betrachtungsweise der französischen Schule findet man in Cailliez & Pages (1976). Eine besondere Richtung in der multivariaten Statistik befasst sich mit dem Problem, wie man eine Menge von Objekten in Gruppen ähnlicher Objekte gliedern kann. Diese Richtung wird in der angelsächsischen Literatur meist als „Cluster Analysis" bezeichnet. Für den Biologen dürfte die geeignetste Einführung in dieses Gebiet die Bücher von Sokal & Sneath (1963) und Sneath & Sokal (1973) sein. Cole (1969) hat eine Reihe von Beiträgen eines Symposions über diesen Fragenkomplex herausgegeben. Eine Beschreibung der Methoden findet man bei Hartigan (1975). Jardine & Sibson (1971) behandeln den Fragenkomplex vom mathematisch-theoretischen Standpunkt aus, leider ist dieses Buch aber sehr schwer verständlich geschrieben. Einige Gedankengänge daraus sollen daher im folgenden Kapitel anhand eines einfachen Beispiels erläutert werden. Eine neue Darstellung des ganzen Sachgebietes soll demnächst im 2. Band des von Krishnaiah herausgegebenen „Handbook of Statistics" erscheinen. Eine kürzere Übersicht ist von Cormack (1971) verfasst worden. Eine andere Gruppe multivariater Methoden befasst sich mit dem Problem der optimalen Trennung verschiedener Populationen. Eine ausführliche Bibliographie, sowie eine Reihe einschlägiger Arbeiten sind in dem Von Cacoullos (1973) herausgegebenen Band zusammengestellt. Über den neuesten Stand auf diesem Sachgebiet referiert Lachenbruch (1979). Die Anwendung multivariater Methoden bei einer Reihe von biologischen Problemen wird im Buch von Blackith & Reyment (1971) besprochen.

Gruppierung von ähnlichen Objekten

Figur 1 stellt zwei Zweige von *Vaccinium Myrtillus* und einen Zweig von *Lonicera nigra* dar. Die Blätter dieser Zweige kann man als Einzelobjekte betrachten und damit eine Cluster-Analyse durchführen. Da man genau weiss, zu welchem Zweig jedes Blatt gehört, kann auf diese Weise die Brauchbarkeit des gewählten Verfahrens überprüft werden.

Damit eine statistische Bearbeitung überhaupt möglich ist, muss die Form der Blätter irgendwie numerisch erfasst werden. Dies soll zunächst nur durch die Messung von Länge

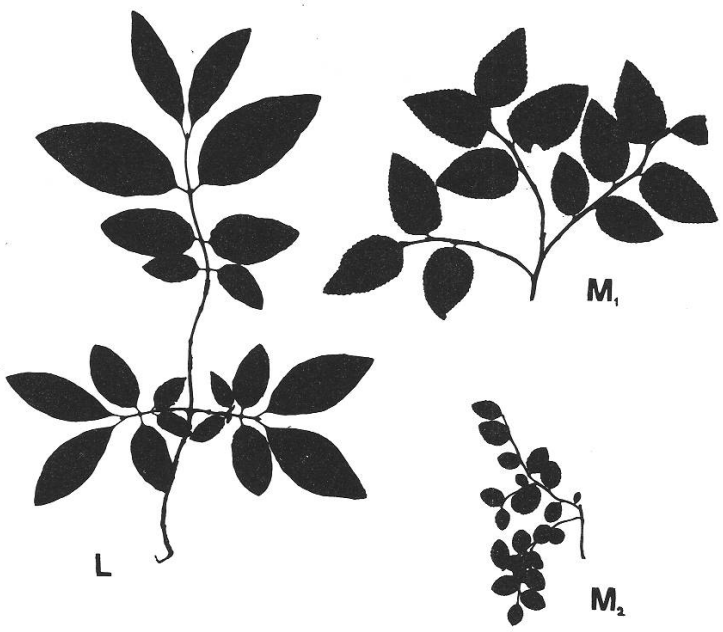

Fig. 1:
Drei Zweige, deren Blätter im Text zur Veranschaulichung gewisser Prinzipien statistischer Klassifizierungsmethoden dienen.
L: *Lonicera nigra*
M_1: grossblättriges *Vaccinium Myrtillus*
M_2: kleinblättriges *Vaccinium Myrtillus*

und Breite eines jeden Blattes geschehen. Die Ergebnisse dieser Messungen lassen sich in ein rechtwinkliges Koordinatensystem eintragen (Fig. 2). Jedem Blatt entspricht in dieser Darstellung ein Punkt. Wie man sieht, nehmen die Punkte, die zu verschiedenen Zweigen gehören, voneinander getrennte Gebiete ein. Daraus folgt, dass die beiden Messungen genügend Information enthalten, um die Blätter der verschiedenen Zweige voneinander zu unterscheiden.

Zur Entscheidung, welche Objekte zu einer Gruppe zu vereinigen sind, benötigt man ein Mass für die Ähnlichkeit zwischen zwei Objekten. Ein oft verwendetes derartiges Ähnlichkeitsmass ist die Distanz zwischen den diesen Objekten zugeordneten Punkten im Koordinatenraum.

Ein Blick auf Fig. 2 zeigt, dass gewisse Punkte, welche zu *Lonicera*-Blättern gehören, nur eine geringe Distanz zu Punkten haben, welche zu *Vaccinium*-Blättern gehören. Andererseits gibt es am gleichen *Lonicera*-Zweig Blätter, deren zugehörige Punkte in der Koordinatenebene weit entfernt voneinander liegen. Eine Gruppierung auf Grund der Distanzen im Koordinatensystem würde infolgedessen zu einer ganz unnatürlichen Einteilung führen.

Wie man auf Fig. 2 sehen kann, lassen sich die Punkte der *Lonicera*-Blätter von den Punkten der *Vaccinium*-Blätter durch eine Gerade (A-A) durch den Nullpunkt trennen. Alle Punkte einer Geraden durch den Nullpunkt des Koordinatensystems haben dasselbe Verhältnis Länge zu Breite. Bei Punkten oberhalb der Geraden ist dieses Verhältnis grösser, und bei Punkten unterhalb der Geraden kleiner als auf der Geraden. Da alle Punkte, welche zu *Lonicera*-Blättern gehören, oberhalb der Geraden A-A liegen, und alle Punkte, welche zu *Vaccinium*-Blättern gehören, unterhalb der Geraden liegen, folgt daraus, dass das Verhältnis der Länge zur Breite bei den *Lonicera*-Blättern grösser ist, als bei den *Vaccinium*-Blättern, ohne dass eine Überschneidung vorkommt. Dieses Verhältnis ist somit ein besseres Mass zur Bestimmung der Zugehörigkeit der einzelnen Blätter, als das Zahlenpaar (Länge, Breite), aus welchem das Verhältnis gebildet wird.

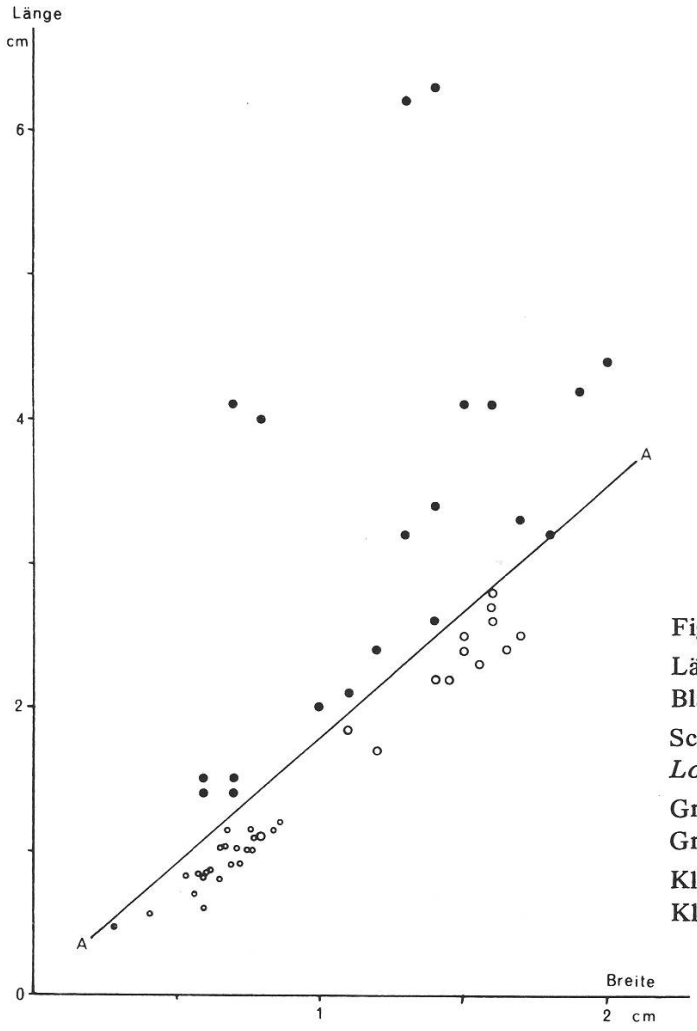

Fig. 2:
Länge und Breite der auf Fig. 1 abgebildeten Blätter.
Schwarze Kreise:
Lonicera nigra
Grosse weisse Kreise:
Grossblättriges *Vaccinium Myrtillus*
Kleine weisse Kreise:
Kleinblättriges *Vaccinium Myrtillus*

Man ersieht aus diesem Beispiel, dass es nicht genügt, wenn die Messungen die notwendige Information enthalten, wesentlich ist auch, in welcher Form diese Information dargestellt wird.

Die Sicherheit der Unterscheidung lässt sich natürlich durch Hinzuziehen weiterer Merkmale verbessern. So sind z.B. die *Vaccinium*-Blätter gezähnt, die *Lonicera*-Blätter hingegen ganzrandig. Man kann nun das Merkmal gezähnt-ganzrandig so kodieren, dass gezähnte Blätter den Wert eins und ganzrandige Blätter den Wert null erhalten. Betrachtet man nun diese Zahlen als dritte Koordinate, so kommen die *Lonicera*-Blätter in eine andere Ebene zu liegen, als die *Vaccinium*-Blätter. Dadurch wird der Abstand der Punkte der beiden Gattungen vergrössert.

Es stellt sich nun aber das Problem, wie gross die Einheit auf der dritten Koordinatenachse, verglichen mit einem Centimeter auf den beiden anderen Achsen, gewählt werden soll. Je grösser man diese Einheit wählt, umso grösser wird auch der Abstand zwischen den Punkten der beiden Gattungen, umso mehr Gewicht bekommt also das Merkmal „gezähnt oder ganzrandig". Das gleiche Problem kann schon auftauchen, wenn gleichartige Masse (z.B. Längen) in ein Koordinatensystem eingetragen werden. Wenn z.B. die eine Koordinate die Länge eines Blattes, und die andere Koordinate den Durchmesser

der Epidermiszellen auf der Blattoberfläche darstellt, wäre es falsch, für beide Achsen denselben Massstab zu wählen, weil der Zelldurchmesser verglichen mit der Blattlänge verschwindend klein ist, und daher gar nicht mehr ins Gewicht fallen würde.

Um zu erreichen, dass alle Merkmale ihr gebührendes Gewicht erhalten, muss für die Blattlänge eine wesentlich längere Strecke als Einheit gewählt werden, als für die Zelldurchmesser. In solchen Fällen wird oft die Standardabweichung aller Messungen der betreffenden Koordinate als Längeneinheit gewählt. Dies hat zur Folge, dass die Streuung in jeder Koordinatenrichtung gleich gross wird. Dies ist jedoch nicht unbedingt wünschenswert, weil gerade in derjenigen Richtung, in welcher grosse Unterschiede zwischen den verschiedenen Gruppen bestehen, eine grosse Streuung vorhanden ist. Die Wahl der Standardabweichung als Einheit bewirkt dann, dass diese Unterschiede verkleinert werden, sodass unähnliche Objekte einander angenähert werden. Dies kann zwar verhindert werden, wenn statt der Standardabweichung zwischen allen Einzelobjekten die Standardabweichung zwischen den Objekten innerhalb der Gruppen als Einheit gewählt wird, was aber erst möglich ist, wenn die Objekte bereits in Gruppen eingeteilt sind. Man muss daher von einer provisorischen Gruppierung ausgehen, um die Standardabweichung innerhalb der Gruppen zu berechnen. Nach Anpassung der Koordinateneinheiten wird dann die Gruppeneinteilung überprüft, was dann wiederum eine Neuberechnung der Standardabweichungen bedingt usw. Derartige iterative Verfahren sind sehr aufwendig, wenn eine grössere Zahl von Objekten klassifiziert werden muss.

Diese Schwierigkeiten können umgangen werden, indem anstelle des Abstands im Koordinatensystem ein anderes Mass für die Ähnlichkeit gewählt wird. Ein derartiges Mass ist von Jardine & Sibson (1971) beschrieben worden. Da die Darstellung dieser Autoren für einen Nicht-Mathematiker kaum verständlich ist, soll im folgenden versucht werden, das Prinzip, auf welchem dieses Ähnlichkeitsmass beruht, anhand des Beispiels der Zweige von *Vaccinium* und *Lonicera* zu erläutern. Zu diesem Zweck betrachten wir die Blätter eines Zweiges als Stichprobe der vom zugehörigen Strauch je gebildeten oder in Zukunft möglicherweise zu bildenden Blätter. Es hat dann einen Sinn, die Wahrscheinlichkeit zu betrachten, dass eine bestimmte Blattform von diesem Strauch verwirklicht wird. Der Einfachheit halber nehmen wir zunächst an, dass die Blattformen in eine endliche Anzahl N von Klassen $F_1, F_2 \ldots F_N$ eingeteilt worden seien. Es sei nun P_{Ai} die Wahrscheinlichkeit, dass ein Blatt aus der Klasse F_i vom Strauch A realisiert wird, und P_{Bi} die Wahrscheinlichkeit, dass ein Blatt aus der gleichen Klasse vom Strauch B realisiert wird. Wenn man nun ein Blatt aus der Klasse F_i vor sich hat, von dem man weiss, dass es entweder vom Strauch A oder vom Strauch B stammt, so ist man umso sicherer, dass es vom Strauch A stammt, je grösser das Verhältnis P_{Ai}/P_{Bi} ist. In der Informationstheorie bezeichnet man die Grösse $\log (P_{Ai}/P_{Bi})$ als diejenige Information zugunsten der Hypothese, dass das Blatt von Strauch A stammt, und nicht von Strauch B, welche in der Kenntnis steckt, dass das Blatt zur Blattform-Klasse F_i gehört (vgl. Kullback 1968). Summiert man die Grösse $P_{AI} \log (P_{Ai}/P_{Bi})$ über alle Klassen F_i, welche vom Strauch A realisiert werden können, so erhält man die mittlere Information einer Blattform des Strauches A zugunsten der Hypothese, dass ein Blatt von Strauch A und nicht von Strauch B stammt. Diese Summe ist dann und nur dann null, wenn P_{Ai} und P_{Bi} für alle Blattformen übereinstimmen. Dann kommt nämlich jede Blattform an beiden Sträuchern gleich häufig vor, sodass die Blattform keine Information zur Beantwortung der

Frage liefert, von welchem der beiden Sträucher ein bestimmtes Blatt stammt. Bildet man die analoge Summe für den Strauch B und addiert die beiden Summen, so erhält man eine Grösse, welche von Kullback „Divergence" genannt wird, weil sie umso mehr von null abweicht, je mehr die Wahrscheinlichkeiten der beiden Sträucher divergieren. Schwierigkeiten entstehen, wenn in einer der Klassen F_i eine der beiden Wahrscheinlichkeiten P_{Ai} oder P_{Bi} null ist, weil die Division nur für von null verschiedene Divisoren definiert ist. Jardine & Sibson verwenden daher im Nenner des Wahrscheinlichkeitsverhältnisses das arithmetische Mittel der beiden Wahrscheinlichkeiten P_{Ai} und P_{Bi}, sodass sie das folgende Mass für die Divergenz zwischen den beiden Sträuchern erhalten, das sie als „Informationsradius" bezeichnen:

$$D(A, B) = \sum_{i=1}^{n} \left[P_{Ai} \log \frac{2 P_{Ai}}{P_{Ai} + P_{Bi}} + P_{Bi} \log \frac{2 P_{Bi}}{P_{Ai} + P_{Bi}} \right] \qquad (1)$$

Betrachtet man anstelle einer endlichen Anzahl von diskreten Klassen von Blattformen eine unendliche Zahl von kontinuierlich ineinander übergehenden Blattformen, so hat man anstelle der Wahrscheinlichkeiten Wahrscheinlichkeitsdichten zu setzen und die Summe ist durch ein Integral zu ersetzen.

Wird ausser der Blattform ein weiteres Merkmal beobachtet, so kann für dieses Merkmal in analoger Weise ebenfalls ein Informationsradius berechnet werden. Wenn dieses Merkmal von der Blattform unabhängig ist, so ergibt die Summe der beiden Informationsradien den Informationsradius zwischen den Sträuchern A und B unter Berücksichtigung beider Merkmale. Auf diese Weise können sowohl quantitative, als auch qualitative Merkmale berücksichtigt werden, sodass ein Divergenzmass erhalten werden kann, welches auf der Gesamtheit der beobachteten Merkmale beruht.

Wenn man dieses Verfahren in der Praxis anwenden will, so muss man die Wahrscheinlichkeitsverteilungen auf Grund von Stichproben schätzen. Am einfachsten ist dies bei qualitativen Merkmalen, also z.B. bei der Frage, ob der Blattrand gezähnt oder ganzrandig ist. Die Schätzung der Wahrscheinlichkeit geschieht dann einfach dadurch, dass man die Zahl der Blätter mit gezähntem Rand durch die Gesamtzahl der beobachteten Blätter dividiert. Wenn bei Strauch A, wie dies bei *Lonicera* der Fall ist, alle Blätter ganzrandig sind, und bei Strauch B, wie bei *Vaccinium,* alle Blätter gezähnt, dann nimmt die Divergenz zwischen den beiden Sträuchern den grösstmöglichen Wert an, nämlich log 2. Sind hingegen bei beiden Sträuchern alle Blätter gezähnt oder bei beiden Sträuchern alle Blätter ganzrandig, so nimmt der Informationsradius den kleinstmöglichen Wert, null, an. Hat man an den zu vergleichenden Objekten lauter qualitative Merkmale beobachtet, von denen man annehmen kann, dass sie bei einem Objekt entweder immer vorhanden sind, oder immer fehlen, so setzt man für jedes Merkmal, in welchem die verglichenen Objekte übereinstimmen, eine null, und für jedes Merkmal, in welchem die Objekte sich unterscheiden, eine eins. Die Summe dieser Nullen und Einsen, multipliziert mit dem Logarithmus von 2, ergibt dann den Informationsradius. Arbeitet man, wie dies in der Informationstheorie üblich ist, mit Logarithmen zur Basis 2, dann ist der Logarithmus von 2 gleich eins, und man erhält ein in der numerischen Taxonomie oft gebrauchtes Distanzmass, die sogenannte „City Block Distance".

Schwieriger wird die Schätzung bei quantitativen Merkmalen, da die Wahrscheinlichkeitsverteilung unendlich viele verschiedene Formen annehmen kann.

Wenn angenommen werden darf, dass die Verteilung durch eine Gauss'sche Normalverteilung beschrieben werden kann, genügt die Schätzung von Mittelwert und Standardabweichung zur Charakterisierung der Verteilung. Der Informationsradius kann dann einer Tabelle aus dem Buch von Jardine & Sibson entnommen werden. Die Addition der Informationsradien mehrerer quantitativer Merkmale ist aber nur dann zulässig, wenn die Merkmale statistisch voneinander unabhängig sind. In unserem Beispiel sind Blattlänge und Blattbreite eindeutig miteinander korreliert, indem längere Blätter im allgemeinen auch eine grössere Breite besitzen. Die Informationsradien für Blattlänge und Blattbreite dürfen daher nicht einfach addiert werden. Durch Berechnung der sog. Hauptkomponenten (principal components) kann man in einem solchen Fall statistisch voneinander unabhängige Variabeln erhalten. Das Verfahren ist in den Büchern über multivariate Statistik beschrieben.

Es ist den Morphologen schon lange bekannt, dass zwei an einem Organ gemessene Strecken oft angenähert auf einer Geraden liegen, wenn sie in ein doppelt logarithmisches Koordinatensystem eingetragen werden (siehe z.B. Pearsall, 1927, Schüepp, 1945, 1963). Man spricht dann von Allometrie. Wenn die Steigung der Geraden gleich eins ist, bedeutet dies, dass die beiden Strecken zueinander in einem konstanten Verhältnis stehen. Wenn man in einem solchen Fall die Koordinatenachsen so dreht, dass die eine Koordinatenachse parallel zu der den Messpunkten angepassten Geraden steht, und die andere Achse senkrecht dazu, so sind die neuen Koordinaten der Messpunkte miteinander praktisch unkorreliert. Wenn man zudem annehmen darf, dass diese Koordinaten praktisch normal verteilt sind, dann dürfen nach der Drehung des Koordinatensystems die Verteilungen in den beiden Koordinatenrichtungen als statistisch voneinander unabhängige Normalverteilungen betrachtet werden, und die Informationsradien, die aus der Tabelle von Jardine & Sibson abgelesen worden sind, dürfen zueinander addiert werden.

Auf Fig. 3 sind die Messungen an den Blättern der beiden *Vaccinium*-Zweige und des *Lonicera*-Zweiges in einem doppelt logarithmischen Koordinatensystem dargestellt. Für jeden Zweig kann eine Gerade mit der Steigung eins gefunden werden, um die sich die Messpunkte fast symmetrisch gruppieren. Für alle Punkte auf einer solchen Geraden gilt die Gleichung:

$$\log 1 = \log b + \text{const.}$$

Daraus folgt:

$$\log 1 - \log b = \log 1/b = \text{const.}$$

d.h. Punkten auf der Geraden entsprechen Blätter mit dem gleichen Verhältnis der Länge zur Breite. Legt man durch einen Punkt der Geraden eine zweite Gerade senkrecht zur ersten, so entsprechen den Punkten dieser zweiten Geraden Blätter mit dem gleichen Produkt Länge mal Breite, also mit annähernd derselben Blattfläche. Wird nun das Koordinatensystem so gedreht, dass die eine Achse parallel zur ersten Geraden, und die andere Achse parallel zur zweiten Geraden liegt, dann entspricht im neuen

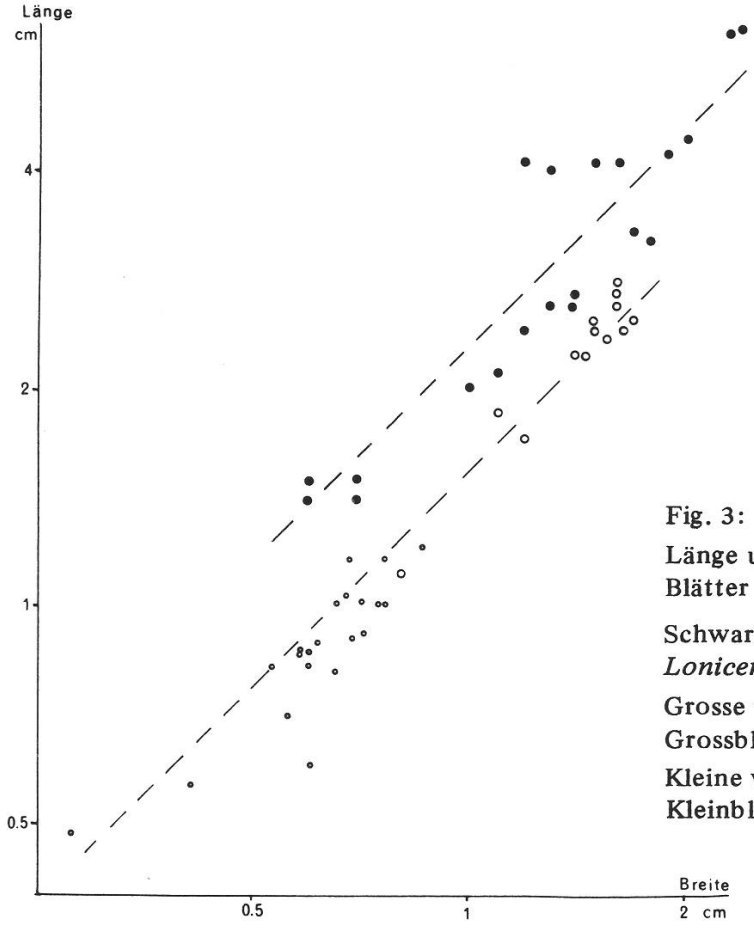

Fig. 3:
Länge und Breite der auf Fig. 1 abgebildeten Blätter in doppelt logarithmischem Massstab.
Schwarze Kreise:
Lonicera nigra
Grosse weisse Kreise:
Grossblättriges *Vaccinium Myrtillus*
Kleine weisse Kreise:
Kleinblättriges *Vaccinium Myrtillus*

Koordinatensystem der einen Koordinate ein bestimmtes Verhältnis Länge zu Breite, und der andern Koordiante ein bestimmtes Produkt Länge mal Breite. Nimmt man nun an, dass die Logarithmen von Blattbreite und Blattlänge normal verteilt seien, dann lässt sich die Wahrscheinlichkeitsverteilung von Blattbreite und Blattlänge mit Hilfe von Mittelwert und Standardabweichung des Logarithmus des Verhältnisses Länge/Breite und des Logarithmus des Produkts Länge mal Breite vollständig beschreiben. Diese vier Grössen genügen auch, um den Informationsradius nach Jardine & Sibson zu bestimmen. In der Tabelle 1 sind die betreffenden Werte für jeden der drei Zweige zusammengestellt. Die Frage, ob die Annahme der Normalverteilung berechtigt ist, kann mit dem Test von Shapiro & Wilk (1965) geprüft werden. Wenn diese Frage verneint werden muss, empfehlen Jardine & Sibson, den Koordinatenraum in Klassen einzuteilen und dann die Anzahl der Messpunkte in jeder dieser Klassen auszuzählen. Der Informationsradius kann dann nach der Formel (1) berechnet werden. Da man in der Praxis selten über sehr grosse Stichproben verfügt, muss diese Einteilung ziemlich grob gewählt werden (vgl. Cochran 1961). Da bei diesem Vorgehen das Ergebnis durch die willkürliche Wahl der Klassengrenzen beeinflusst wird, so ist dieser Weg bloss als ein Notbehelf zu betrachten. Eine andere Möglichkeit besteht darin, dass man versucht, eine geeignete Transformation zu finden, welche die Daten eher normal verteilt macht.

Tabelle 1

Mittelwert und Standardabweichung der Logarithmen des Verhältnisses Länge zu Breite und des Produkts Länge mal Breite der Blätter der drei Zweige von Fig. 1.

	log 1/b Mittel	Standardabw.	log. 1 · b Mittel	Standardabw.
Lonicera (L)	0.3539	0.07841	0.5693	0.3796
Vaccinium, grossbl. (M_1)	0.1924	0.03284	0.4900	0.1986
Vaccinium, kleinbl. (M_2)	0.1461	0.04936	− 0.2529	0.2028

Auf Grund der Werte von Tabelle 1 erhält man die in Tabelle 2 zusammengestellten Informationsradien für die drei möglichen Paarungen von je zwei verschiedenen Zweigen:

Tabelle 2

Informationsradien zwischen je zwei Zweigen von Fig. 1.

Paarung	log 1/b	log 1 · b	Summe
L, M_1	0.7255	0.1658	0.8913
L, M_2	0.8078	0.7201	1.5279
M_1, M_2	0.2248	0.8948	1.1196

Der Informationsradius stellt eine Alternative zur Distanz im Koordiantenraum dar, die in dem Fall angewendet werden kann, dass die zu gruppierenden Objekte selbst Kollektive sind. Sind jedoch Einzelobjekte zu gruppieren, so kommen die Ähnlichkeitsmasse von Goodall (1966) und von Gower (1971) in Frage.

Die Aufgabe besteht nun darin, ähnliche Objekte in Gruppen („Clusters") zusammenzufassen. Zur Lösung dieser Aufgabe ist eine ganze Reihe von Methoden entwickelt worden, die unter der Bezeichnung „Automatische Klassifikation" oder „Cluster Analysis" zusammengefasst werden. Das einfachste Verfahren, das zudem vom mathematischen Standpunkt in einem gewissen Sinne optimal ist (Jardine & Sibson), ist das sogenannte „Single Linkage"-Verfahren. Es besteht darin, dass zunächst die beiden ähnlichsten Objekte vereinigt werden, dann die beiden nächst ähnlichen usw. Auf diese Weise können allmählich auch ziemlich unähnliche Objekte in derselben Gruppe vereinigt sein, wenn sie nämlich durch eine Reihe von Zwischengliedern verbunden sind.

Wendet man dieses Verfahren auf die dritte Kolonne von Tabelle 2 an, so muss man zuerst den *Lonicera*-Zweig mit dem grossblättrigen Heidelbeerzweig vereinigen, weil dieses Paar den kleinsten Informationsradius (0.8913) hat. Der nächstgrössere Informationsradius (1.1196) gehört zur Paarung der beiden Heidelbeerzweige. Im zweiten Schritt wird daher der kleinblättrige *Vaccinium*-Zweig mit der Gruppe der beiden andern Zweige vereinigt. Weil jetzt alle Objekte miteinander verbunden sind,

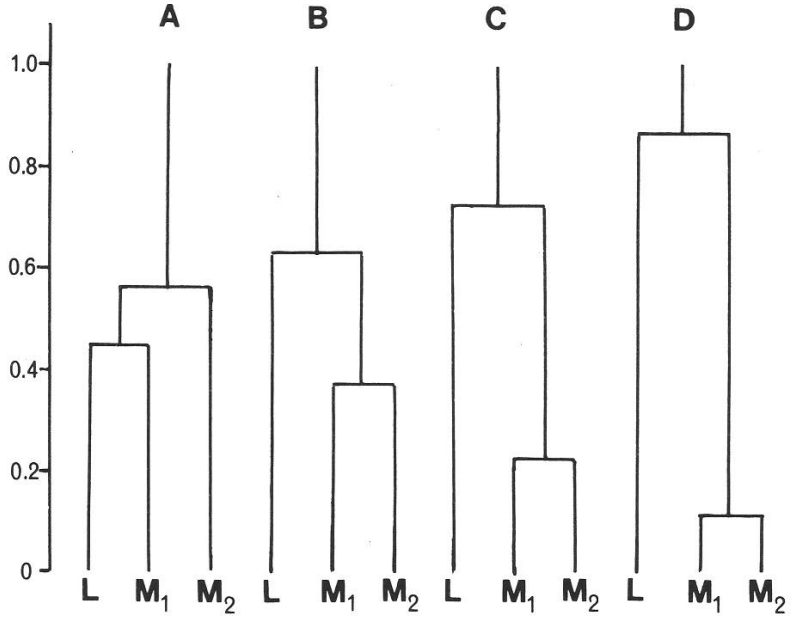

Fig. 4:
Dendrogramme zur Darstellung der Beziehungen zwischen den drei Zweigen von Fig. 1. Erläuterungen im Text.

ist damit das Verfahren abgeschlossen. Das Ergebnis wird meist in der Form eines Dendrogramms (Fig. 4A) dargestellt. Der Massstab links vom Dendrogramm entspricht dem mittleren Informationsradius, d.h. dem Informationsradius, geteilt durch die Anzahl der Merkmale, bei welchem die Vereinigung geschieht. Je weiter oben die Gabelung liegt, welche zwei Objekte verbindet, umso grösser ist also der betreffende Informationsradius, d.h. umso mehr unterscheiden sich die durch die betreffende Gabel verbundenen Objekte.

Das Dendogramm Fig. 4A ist recht unbefriedigend ausgefallen, da in ihm zwei Objekte, welche systematisch in zwei verschiedene Familien gehören, zuerst vereinigt worden sind, und die beiden zur gleichen Art gehörenden Objekte erst nachher. Diese Situation kann verbessert werden, indem weitere Merkmale zugezogen werden, sodass mehr Information zur Klassifizierung der Objekte zur Verfügung steht. In unserem Beispiel ist die Zähnung des Blattrandes ein geeignetes Merkmal, weil alle Blätter des *Lonicera*-Zweiges ganzrandig sind, und alle *Vaccinium*-Blätter gezähnt. Der Informationsradius für dieses Merkmal beträgt daher für eine Paarung eines *Lonicera*-Zweigs mit einem *Vaccinium*-Zweig eins und für die Paarung der beiden *Vaccinium*-Zweige null. Dieser Betrag ist zu den Werten der dritten Kolonne von Tabelle 2 zu addieren, sodass man die Summen der Informationsradien erhält: für die Paarung (L, M_1) 1.8913, für die Paarung (L, M_2) 2.5279 und für die Paarung (M_1, M_2) den Wert 1.1196. Mit diesen Werten erhält man das Dendrogramm Fig. 4B. Die beiden *Vaccinium*-Zweige sind jetzt näher zusammengerückt, der Abstand zu *Lonicera* ist aber noch nicht sehr bedeutend.

Verwendet man nur die Information aus dem Verhältnis der Blattlänge zur Blattbreite (Kolonne 1 von Tabelle 2), so ergibt sich das Dendrogramm Fig. 4C; fügt man die Information der Zähnung des Blattrandes hinzu, erhält man das Dendrogramm Fig. 4D, in welchem die nahe Beziehung der beiden *Vaccinium*-Zweige im Gegensatz zum *Lonicera*-Zweig klar zum Ausdruck kommt. Man kann aus diesem Beispiel verschiedenes lernen: Die Verfahren der automatischen Klassifikation ergeben nicht etwa automatisch eine richtige Darstellung der Verwandtschaftsbeziehungen.

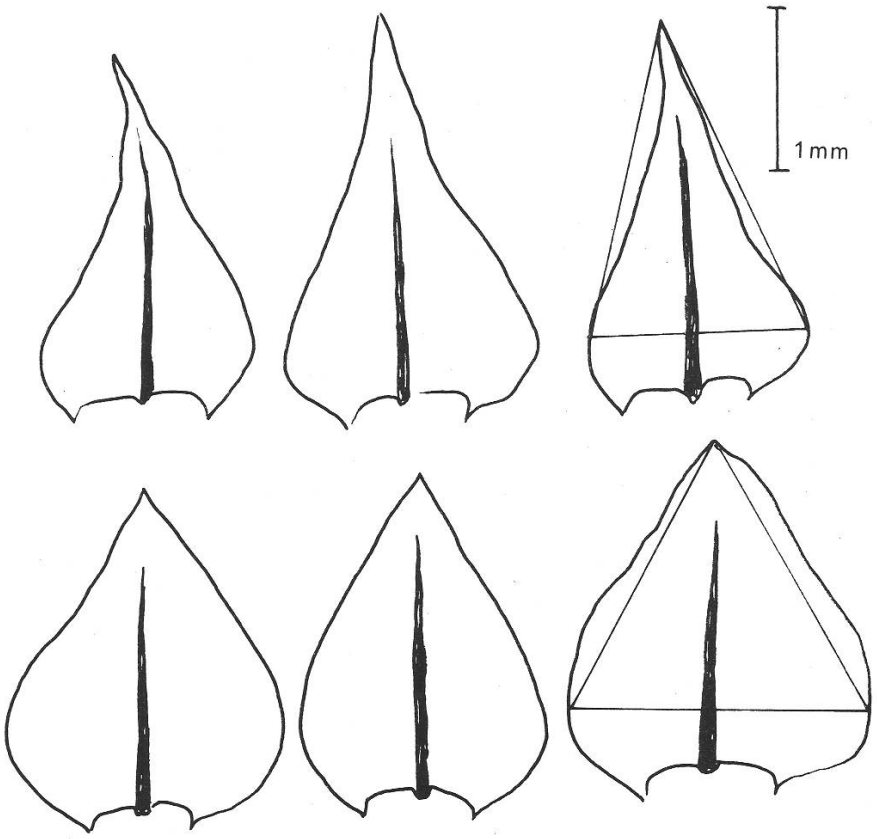

Fig. 5:
Obere Reihe:
Blätter von
Eurhynchium striatum.
Untere Reihe:
Blätter von
Eurhynchium angustirete
(= E. Zetterstedtii).

Voraussetzung dafür, dass brauchbare Resultate erzielt werden, ist vielmehr, dass die wesentliche Information mit den Daten eingegeben wird. Ferner muss man sich darüber klar sein, dass Information, welche für die Klassifikation nicht wesentlich ist, die Verwandtschaftsbeziehungen verschleiern kann. In unserem Beispiel ist die Blattfläche ein solches unwesentliches Merkmal, das ja sehr stark durch die Wachstumsbedingungen am Standort beeinflusst werden kann. Durch Weglassen dieses Merkmals kommen die Verwandtschaftsbeziehungen viel klarer zur Geltung.

Unterscheidung ähnlicher Sippen

Zwei Fragen können mit Hilfe biometrisch-statistischer Methoden untersucht werden, wenn es darum geht, ähnliche Sippen voneinander zu unterscheiden:
1. Liegen überhaupt klar voneinander abgegrenzte Sippen vor? 2. Wie sind die einzelnen Individuen den Sippen zuzuordnen, wenn möglichst wenig Fehlbestimmungen vorkommen sollen? Wie hier im einzelnen vorgegangen werden kann, soll wiederum anhand eines Beispiels erläutert werden.

Störmer (1942) hat die Laubmoosart *Eurhynchium striatum* in zwei Kleinarten aufgespalten, welche sich durch ihre Blattform unterscheiden (Fig. 5). Für die Form mit zugespitzten Blättern behielt er den Namen *E. striatum* bei; die stumpfblättrige

Form nannte er *E. Zetterstedtii.* Nach Koponen (1967) muss die letztere Form *E. angustirete* (Broth.) Koponen heissen. *E. striatum* hat im Norden Europas eine ausgesprochen atlantische Verbreitung und steigt in Norwegen nicht über 300 m in die Höhe (Störmer 1969), die Verbreitung von *E. angustirete* hingegen ist eher kontinental (Koponen 1964). In der Schweiz und den angrenzenden Gebieten ist *E. angustirete* vor allem in der Montanstufe anzutreffen, und steigt nicht unter 600 m hinunter, während *E. striatum* eher in der collinen Stufe vorkommt und höchstens ausnahmsweise über 800 m hinaufsteigt. Es handelt sich also offensichtlich um zwei Sippen mit verschiedenen klimatischen Ansprüchen, sodass deren Unterscheidung für pflanzengeographische und pflanzensoziologische Untersuchungen von Bedeutung ist. Da man oft Formen antrifft, welche schwer einzuordnen sind, hat die Aufspaltung nicht allgemeine Anerkennung gefunden.

Für eine statistische Untersuchung der Unterschiede sind Messungen notwendig. Nach Störmer unterscheiden sich die beiden Arten vor allem durch das Verhältnis der Blattlänge zur Blattbreite, sowie durch den Winkel der Blattspitze. Beim Vergleichen einer grösseren Anzahl von Blättern fällt als weiteres Merkmal auf, dass die Blattkontur bei *E. angustirete* konvex ist, bei *E. striatum* hingegen an der Blattspitze konkav. Konstruiert man ein gleichseitiges Dreieck mit der grössten Blattbreite als Basis und der Blattspitze als Spitze (auf Fig. 5 rechts eingezeichnet), so ist der Spitzenwinkel dieses Dreiecks bei *E. striatum* grösser als der Winkel der Blattspitze, bei *E. angustirete* hingegen ist der Blattspitzenwinkel grösser als der Dreieckswinkel. Zur Bestimmung des Dreieckswinkels genügt die Messung der Länge der grössten Blattbreite und des Abstands der Blattspitze von der grössten Breite: dieser Abstand, dividiert durch die halbe Breite, ergibt den Tangens des halben Dreieckwinkels.

Schliesslich ist die Länge der Zellen in der Blattspitze bei *E. striatum* grösser als bei *E. angustirete.*

64 Proben aus dem Formenkreis von *Eurhynchium striatum* im weiteren Sinn wurden untersucht. Von jeder dieser Proben wurden 4 Blätter vom Moosstämmchen abgelöst, und daran die folgenden Messungen ausgeführt: Länge und Breite der Blätter, Abstand der grössten Breite von der Blattspitze, Winkel der Blattspitze, Länge von je 4 Zellen in der Blattspitze. Aus diesen Daten wurden für jede Probe die folgenden Werte berechnet:

1. Der Mittelwert des Verhältnisses der Länge zur Breite der Blätter.
2. Mittelwert des Verhältnisses des Abstands der Blattspitze von der grössten Breite zur halben Breite (= Tangens des halben Spitzenwinkels des in Fig. 5 rechts eingezeichneten Dreiecks).
3. Mittelwert des Tangens des halben Blattspitzenwinkels.
4. Mittlere Zellenlänge in der Blattspitze in μm.

Jede Probe ist damit durch 4 Zahlen charakterisiert, welche als Koordinaten eines Punktes in einem 4-dimensionalen Koordinatenraum aufgefasst werden können.

Wenn von jeder dieser 64 Proben bekannt wäre, zu welcher der beiden Kleinarten sie gehört, so könnte man das vorliegende Datenmaterial dazu verwenden, um eine Trennfunktion (Discriminant Function) zu berechnen. Das Ergebnis der Berechnung würde uns dann ermöglichen, weitere Proben, deren Zugehörigkeit nicht bekannt ist, zu bestimmen. Da die Zugehörigkeit der Proben in unserem Fall nicht bekannt ist, soll versucht werden, ausgehend von einer provisorischen Zuordnung eine provi-

sorische Trennfunktion zu berechnen. Diese Trennfunktion soll dann dazu dienen, die Zuordnung der einzelnen Proben zu überprüfen und nötigenfalls zu korrigieren. Die korrigierte Zuteilung der Proben ermöglicht dann wiederum, eine verbesserte Trennfunktion zu berechnen. Dieses Verfahren kann solange fortgesetzt werden, bis keine der Proben mehr ihre Zuordnung wechselt, wenn sie mit der zuletzt berechneten Trennfunktion bewertet wird.

Für die erste provisorische Einteilung sind graphische Darstellungen sehr nützlich. Wenn jede Probe nur durch zwei Messungen repräsentiert wird, kann man mit der Darstellung in der Koordinatenebene auskommen. In dieser Darstellungsweise sind die zusammengehörigen Objekte oft als Punktwolken mehr oder weniger klar erkennbar. Für mehrdimensionale Koordinatenräume ist die Darstellung als „Profil" geeignet (vgl. Hartigan, 1975). Im folgenden soll gezeigt werden, wie die Profildarstellung dazu benutzt werden kann, um auf einfache Weise eine provisorische Trennfunktion zu berechnen.

Ein Profil besteht aus einer Reihe von k parallelen Skalen, von denen jede eine der k Koordinaten repräsentiert. Ein Punkt im k-dimensionalen Koordinatenraum wird durch einen Streckenzug dargestellt, der die Koordianten des betreffenden Punktes miteinander verbindet. Der Massstab der Skalen wird mit Vorteil so gewählt, dass die den Punkten entsprechenden Streckenzüge möglichst weit auseinanderrücken, und die Richtung, in welcher die Koordinatenwerte zunehmen, so, dass möglichst wenig Streckenzüge sich überkreuzen.

Aus den 64 Proben wurde eine Stichprobe von 15 Proben zufällig ausgewählt. Diese 15 Proben sind auf Fig. 6 als Profil dargestellt. Die Reduktion der Anzahl der Proben wurde aus zwei Gründen vorgenommen: 1. lässt sich leider nur eine be-

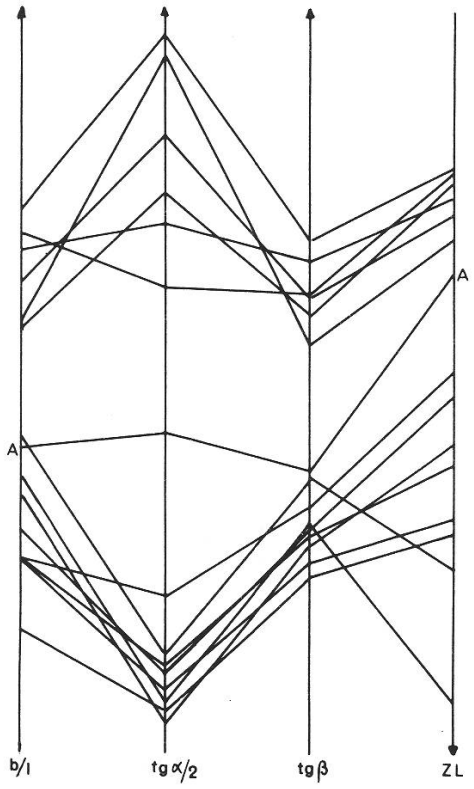

Fig. 6:
Profildarstellung von Messungen an Blättern von 15 Proben von *Eurhynchium striatum* sens. lat. Erläuterungen im Text.

schränkte Anzahl von Punkten in einem Profildiagramm darstellen, weil sonst die Darstellung bald unübersichtlich wird; 2. wenn nur ein Teil des Datenmaterials zur Berechnung der provisorischen Trennfunktion verwendet wird, kann mit dem Rest eine unabhängige statistische Überprüfung der arbiträren provisorischen Einteilung vorgenommen werden.

Die auf Fig. 6 dargestellten Proben lassen sich mit Ausnahme der mit A-A bezeichneten Probe in zwei klar getrennte Gruppen einteilen. Diese Gruppen können nun dazu verwendet werden, um eine provisorische Trennfunktion zu berechnen. Die von R.A. Fisher (1936) vorgeschlagene lineare Trennfunktion hat die folgende Form:

$$Y = b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \qquad (2)$$

Die Koeffizienten b_i erhält man durch Auflösung des folgenden linearen Gleichungssystems (siehe z.B. Linder, 1951, Kap. 64):

$$\begin{aligned} b_1 S_{11} + b_2 S_{12} + \ldots + b_p S_{1p} &= d_1 \\ b_1 S_{21} + b_2 S_{22} + \ldots + b_p S_{2p} &= d_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ b_1 S_{p1} + b_2 S_{p2} + \ldots + b_p S_{pp} &= d_p \end{aligned} \qquad (3)$$

wobei

$$S_{ij} = \sum_A (x_{iA} - \bar{x}_{iA})(x_{jA} - \bar{x}_{jA}) + \sum_B (x_{iB} - \bar{x}_{iB})(x_{jB} - \bar{x}_{jB})$$

x_{iA} = Wert der i-ten Koordinate eines Objekts der Gruppe A

x_{iB} = Wert der i-ten Koordinate eines Objekts der Gruppe B

$\bar{x}_{iA}$ = Mittelwert der i-ten Koordinate in der Gruppe A

$\bar{x}_{iB}$ = Mittelwert der i-ten Koordinate in der Gruppe B

$$d_i = \bar{x}_{iA} - \bar{x}_{iB} = \sum_A x_{iA}/n_A - \sum_B x_{iB}/n_B$$

n_A = Anzahl Objekte in der Gruppe A

n_B = Anzahl Objekte in der Gruppe B

$S_{ij} / (n_A + n_B - 2)$ ist eine Schätzung für das Produkt $\rho_{ij}\sigma_i\sigma_j$ wobei σ_i für die Standardabweichung der i-ten Koordinate, und ρ_{ij} für den Korrelationskoeffizienten zwischen der i-ten und der j-ten Koordinate steht. Wenn die Koordinaten nicht miteinander korreliert sind, vereinfacht sich das Gleichungssystem auf

$$b_i S_{ii} = d_i \qquad (4)$$

Die Rechnung kann nun ganz erheblich vereinfacht werden, indem anstelle der arithmetischen Mittel $\bar{x}_i$ Medianwerte verwendet werden. Der Medianwert einer Stichprobe ist derjenige Wert, welcher von ebensovielen Werten überschritten wie unterboten wird. Bei einer ungeraden Zahl von Werten in der Stichprobe stimmt er mit

dem mittelsten Wert überein, bei einer geraden Anzahl liegt er in der Mitte zwischen den beiden mittleren Werten. In einem Profildiagramm kann er leicht durch Abzählen erhalten werden und eine Schätzung für die Distanz d_i kann man durch Abmessen der Distanz zwischen den Medianwerten der zu unterscheidenden Gruppen erhalten.

Als Mass der Streuung kann die Spannweite w_i zwischen dem grössten und dem kleinsten Wert einer Stichprobe ebenfalls leicht am Diagramm abgelesen werden. Bei normal verteilten Grössen ist die Spannweite im Mittel proportional zur Standardabweichung, resp. zu $\sqrt{S_{ii}}$. Tabellen mit den entsprechenden Proportionalitätsfaktoren sind von Pearson & Hartley (1958) publiziert worden. Für die Schätzung der b_i in Formel (4) ist dieser Proportionalitätsfaktor jedoch unwichtig, da es für die Trennfunktion belanglos ist, wenn alle Koeffizienten b_i mit demselben Faktor multipliziert werden (vgl. z.B. Linder, 1951, S. 241). Wenn die Voraussetzungen für die Gleichungen (4) zutreffen, kann man daher sofort die Trennfunktion (2) so hinschreiben (vgl. Rao 1952, p. 306):

$$y = \frac{d_1}{w_1^2} x_1 + \frac{d_2}{w_2^2} x_2 + \ldots + \frac{d_p}{w_p^2} x_p \qquad (5)$$

Sind jedoch die Koordinaten miteinander korreliert, so benötigt man noch Schätzungen der Korrelationskoeffizienten ρ_{ij}. Auch dies kann aus dem Profildiagramm ohne grossen Rechenaufwand erhalten werden, indem man auszählt, wie oft sich die Linien, welche die Koordinaten der Variablen verbinden, schneiden. Die Zahl der Schnittpunkte sei k. Der Ausdruck

$$t = 1 - 4\,k / (n\,(n-1)) \qquad (6)$$

entspricht dann dem Rang-Korrelationskoeffizienten nach Kendall (1962). Der Ausdruck (6) bedarf einer Korrektur, falls Bindungen (ties) vorkommen. Diese sind auf dem Diagramm daran zu erkennen, dass mehrere Punkte auf einer Achse zusammentreffen. Gehen von einem Punkt auf einer Achse m verschiedene Linien aus, so ist zur Anzahl k die Hälfte der zwischen m Linien möglichen Anzahl Schnittpunkte zu addieren, also m(m-1)/4. Schneiden sich hingegen m Linien zwischen zwei Achsen in einem Punkt, so ist der Beitrag dieses Schnittpunkts zu k gleich m(m-1)/2.

Wenn man annehmen darf, dass man es mit normal verteilten Daten zu tun hat, kann der Rangkorrelationskoeffizient t dazu verwendet werden, um den Korrelationskoeffizienten ρ_{jj} zu schätzen. Kendall (1962) gibt Formeln für den Erwartungswert und die Varianz von t als Funktion von ρ an. Die Verteilung von t kann durch eine sogenannte Polya-Verteilung approximiert werden (Huber, 1974). Auf Grund dieser Angaben kann man Vertrauensgrenzen für ρ konstruieren. Fig. 7 zeigt solche Grenzen für den Fall, dass 2 Stichproben vom Umfang 6 oder 10 zur Verfügung stehen. Jede Stichprobe liefert dann einen t-Wert, das Diagramm gilt für den Mittelwert aus diesen beiden t-Werten. Die Darstellung zeigt, dass selbst für die geringe Vertrauenswahrscheinlichkeit von 50% die Grenzen recht weit auseinanderliegen, was bedeutet, dass die genaue Lage von ρ ziemlich unsicher ist. Die übliche Methode der Schätzung des Korrelationskoeffizienten („Produkt-Moment-Korrelation") erzielt aber bei so kleinen Stichproben keine wesentlich grössere Präzision der Schätzung.

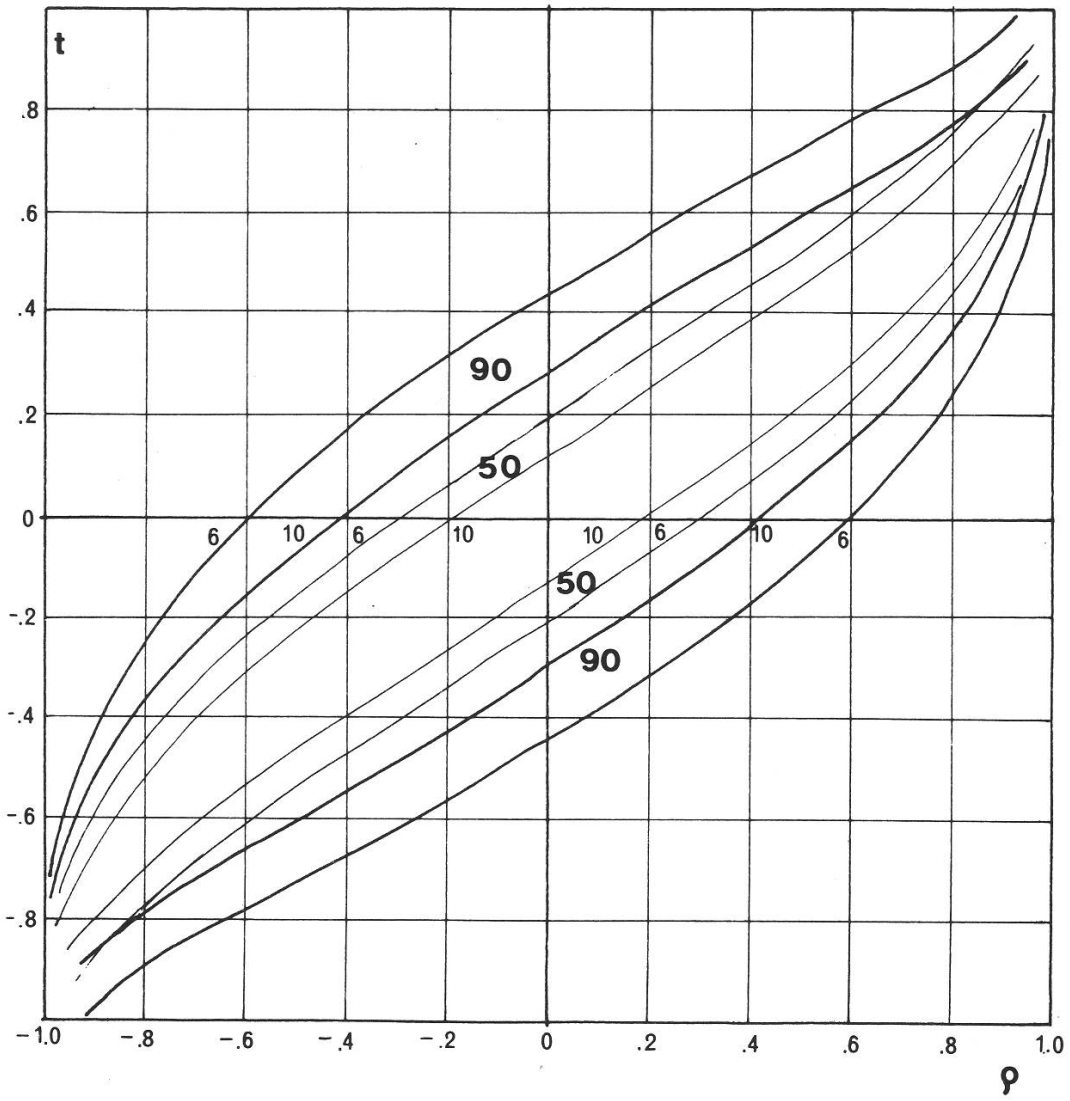

Fig. 7:
Vertrauensgrenzen für den Korrelationskoeffizienten ρ, in Abhängigkeit von t, dem Mittelwert aus zwei Rang-Korrelationskoeffizienten nach Kendall. Umfang der Stichproben 6 oder 10. Vertrauenswahrscheinlichkeiten von 50% und 90%.

Wir haben nun alle Elemente zusammen, um das Gleichungssystem (3) aufzustellen. Die Auflösung eines derartigen Systems war noch bis vor kurzem ohne den Einsatz eines Gross-Computers eine ziemlich mühsame Angelegenheit. Heute gibt es aber bereits Taschenrechner mit einsetzbaren fertigen Programm-Paketen, welche imstande sind, lineare Gleichungssysteme automatisch zu lösen.

In der Tabelle 3 sind die im Profildiagramm (Fig. 6) abgelesenen Werte der Mediane, Spannweiten zwischen grösstem und kleinstem Stichprobenwert, sowie Anzahl Überschneidungen zusammengestellt.

Tabelle 3
Am Profildiagramm (Fig. 6) abgelesene Werte:

a) Medianwerte

		b/1	tg α / 2	tg β	Zell-Länge in µm
Gruppe A	(n = 6)	0.815	0.630	0.536	31.6
Gruppe B	(n = 8)	0.643	0.270	0.374	48.1
Differenz		0.172	0.360	0.162	−16.5

b) Spannweite zwischen grösstem und kleinstem Wert

		b/1	tg α / 2	tg β	Zell-Länge in µm
Gruppe A	(n = 6)	0.107	0.24	0.071	3.9
Gruppe B	(n = 8)	0.116	0.09	0.071	18.2
Mittel		0.1115	0.165	0.071	11.05
Standardabweichung		0.041	0.061	0.026	4.087

c) Überschneidungen
Gruppe A: oberhalb der Diagonale; Gruppe B: unterhalb der Diagonale.

	b/1	tg α / 2	tg β	Zell-Länge
b/1	–	8	2	3
tg α/2	13	–	8	9
tg β	6	7	–	3
Zell-Länge	18	11	12	–

d) Kendall-Rangkorrelation
Gruppe A: oberhalb der Diagonale; Gruppe B: unterhalb der Diagonale.

	b/1	tg α/2	tg β	Zell-Länge
b/1	–	−0.07	0.73	−0.60
tg α/2	0.07	–	−0.07	0.20
tg β	0.57	0.50	–	−0.60
Zell-Länge	0.29	−0.21	−0.14	–

Tabelle 4 enthält die mit Hilfe der Formel (6) aus der Zahl der Überschneidungen berechneten Rangkorrelationskoeffizienten, sowie die entsprechenden auf Fig. 7 abgelesenen Werte des gewöhnlichen Korrelationskoeffizienten.

Tabelle 4
Korrelation zwischen Kendall-Rangkorrelation.

	Gruppe A	Gruppe B	Mittel	ρ
b/1 und tg $\alpha/2$	−0.07	0.07	0.00	0.00
b/1 und tg β	0.73	0.57	0.65	0.83
b/1 und Zell-Länge	−0.60	0.29	−0.16	−0.20
tg $\alpha/2$ und tg β	−0.07	0.50	0.22	0.32
tg $\alpha/2$ und Zell-Länge	0.20	−0.21	0.00	0.00
tg β und Zell-Länge	−0.60	−0.14	−0.37	−0.52

Schätzungen für die Werte von σ erhält man, indem man die mittlere Spannweite durch 2.7 dividiert (Pearson & Hartley, 1958, table 20). Damit sind alle Grundlagen vorhanden, um die Koeffizienten $\rho_{ij}\sigma_i\sigma_j$ des Gleichungssystems (3) zu berechnen. Die erste Zeile von (3) erhält man folgendermassen:

$$
\begin{aligned}
&= \quad 1.00(0.041)\,(0.041) = 0.0016\\
&= \quad 0.00(0.041)\,(0.061) = 0.00\\
&= \quad 0.83(0.041)\,(0.026) = 0.00088\\
&= -0.20(0.041)\,(4.087) = 0.034
\end{aligned}
$$

In analoger Weise werden die übrigen Koeffizienten berechnet, sodass man die folgenden Gleichungen erhält (die Werte sind stark gerundet, da es sich ja nur um Näherungswerte handelt):

$$
\begin{aligned}
0.002\,b_1 \qquad\qquad &+ 0.0009\,b_3 - 0.034\,b_4 &&= \quad 0.17\\
0.004\,b_2 \;+\; &0.0005\,b_3 &&= \quad 0.36\\
0.001\,b_1 + 0.0005\,b_2 \;+\; &0.0007\,b_3 - 0.056\,b_4 &&= \quad 0.16\\
-0.034\,b_1 - \qquad\quad &0.056\,b_3 + 16.7\,b_4 &&= -16.5
\end{aligned}
$$

Die Auflösung dieses Gleichungssystems liefert die folgende Trennfunktion:

$$y = 102\ b/1 + 98\ \text{tg}\,\alpha/2 - 68\ \text{tg}\beta - \text{Zell-Länge in}\ \mu\text{m} \qquad (7)$$

Durch Einsetzen der Messwerte in die Trennfunktion (7) erhält man für jede Probe einen Wert von y. Diese Werte sind in Tabelle 5 in Form eines sog. „stem and leaf plot" (Tukey, 1977) dargestellt.

Tabelle 5

Werte der Trennfunktion (7).
I: Proben, welche zur Berechnung der Trennfunktion gedient haben.
II: übrige Proben.

I		II	
Zehner	Einer	Zehner	Einer
0	5 9	0	2 7
1	3 6 7 8	1	4 5 5 5 7 9
2	2 6	2	0 0 0 2 3 3 3 6 6 9
3		3	2 3 4 8
4		4	2 4 7
5		5	2 9
6		6	2 2 4 5
7	0 1 1 4 8	7	0 0 0 1 3 4 5 5 5 5 7
8	9	8	0 4 4 5 7 8 9
9		9	8
10		10	2
11		11	7
12		12	7

Diese Darstellungsweise vereinigt in sich eine graphische Darstellung mit einer tabellarischen Zusammenstellung des Datenmaterials. Die einzelnen Zahlenwerte werden in Zehner- und Eineranteil aufgespalten, und die Einerziffer wird in der richtigen Zehnerzeile eingetragen. So bedeuten z.B. die 4 Eintragungen in Zeile 1 der Gruppe I rechts vom Strich, dass im Zahlenmaterial die 4 Werte 13, 16, 17 und 18 aufgetreten sind. Die linke Hälfte des Diagramms stellt die 14 Proben dar, welche zur Berechnung der Trennfunktion gedient haben, die rechte Hälfte den Rest der Proben.

Die linke Gruppe wird durch die Trennfunktion sehr deutlich in zwei Teilgruppen getrennt. Dies darf aber nicht als Beweis betrachtet werden, dass zwei klar getrennte Arten vorliegen, da ja die Trennfunktion so berechnet worden ist, dass eine möglichst gute Trennung zustande kommt. Eine entsprechende Aufspaltung findet man aber auch beim Rest der Proben. Da diese Proben bei der Berechnung der Trennfunktion nicht verwendet worden sind, liefern sie eine unabhängige Bestätigung für das Zerfallen in zwei morphologisch verschiedene Sippen. Engelmann & Hartigan (1969) haben einen statistischen Test publiziert, um zu prüfen, ob eine Stichprobe, welche anscheinend in zwei Gruppen zerfällt, in Wirklichkeit aus einer normal verteilten Grundgesamtheit stammen könnte. Da nicht normal verteilte Grundgesamtheiten in der Natur häufig vorkommen, bedeutet die Verwerfung der Hypothese einer normal verteilten Grundgesamtheit noch nicht ohne weiteres, dass damit das Zerfallen in zwei Gruppen nachgewiesen sei. Ein schärferer Test kann folgendermassen durchgeführt werden: wenn tatsächlich zwei getrennte Gruppen vorliegen, so ist zu erwarten, dass im Intervall zwischen den Gruppen die Dichte der y-Werte geringer ist, als in den Gruppenschwerpunkten. Dies kann gegen die „Nullhypothese" getestet werden, dass im ganzen Intervall zwischen dem grössten und dem kleinsten beobachteten y-Wert die Wahrscheinlichkeitsdichte gleich gross ist. Die 14 Proben, die zur Berechnung der provisorischen Trennfunktion gedient haben, ergaben y-Werte

zwischen 5 und 26, sowie zwischen 70 und 89. Die Länge dieser beiden Intervalle beträgt 48% des ganzen Intervalls zwischen 5 und 89. Wenn die Nullhypothese zutrifft, ist daher zu erwarten, dass 52% aller Werte, welche zwischen 5 und 89 liegen, im mittleren Intervall liegen, und nur 48% in den beiden seitlichen Intervallen. 45 der Proben, welche nicht zur Berechnung der provisorischen Trennfunktion gedient hatten, haben y-Werte zwischen 5 und 89. Unter der Nullhypothese ist der Erwartungswert im mittleren Intervall gleich 0.52 x 45 = 23.6, in den beiden seitlichen Intervallen 0.48 x 45 = 21.4, beobachtet werden aber im mittleren Abschnitt 15 und in den beiden seitlichen Intervallen zusammen 30 Werte. Man kann die Wahrscheinlichkeit, dass im mittleren Abschnitt nicht mehr als 15 Werte liegen, wenn die Nullhypothese zutrifft, mit Hilfe der Binominalverteilung berechnen. Diese Wahrscheinlichkeit beträgt 0.0078. Dies bedeutet, dass es unwahrscheinlich ist, das beobachtete Resultat zu erhalten, wenn Gleichverteilung herrscht. Einheitliche Populationen haben aber in der Regel eine mehr oder weniger glockenförmige Verteilung mit erhöhter Wahrscheinlichkeitsdichte in der Mitte. Das beobachtete Ergebnis wird dann noch unwahrscheinlicher. In der Tat erhält man mit dem Test von Engelmann & Hartigan, der von der Nullhypothese einer Normalverteilung ausgeht, eine Wahrscheinlichkeit von weniger als 0.001. Man darf also mit ziemlicher Sicherheit annehmen, dass das vorliegende Material aus einer zweigipfligen Verteilung stammt.

Mit Hilfe der provisorischen Trennfunktion ist es somit gelungen, zu zeigen, dass es sich lohnt, eine bessere Trennung der beiden Komponenten mit Hilfe der aufwendigeren Methode von R.A. Fisher zu versuchen. Man trennt zu diesem Zweck das Material in zwei Gruppen, indem man denjenigen Wert der provisorischen Trennfunktion verwendet, bei welchem sich die deutlichste Lücke zeigt und berechnet für diese Aufteilung dann das Gleichungssystem (3). Die Auflösung dieses Systems liefert die folgende Trennformel:

$$y = -50.3\ b/l + 12.3\ \mathrm{tg}\ \alpha/2 + 207.4\ \mathrm{tg}\ \beta - \text{Zell-Länge} \qquad (8)$$

Die Koeffizienten dieser neuen Trennformel unterscheiden sich ganz wesentlich von denjenigen der provisorischen Formel (7). Verwendet man aber Formel (8) zur Einteilung in zwei Gruppen, so muss nur eine einzige Probe anders eingeteilt werden, als mit der Formel (7). Nach Umteilung dieser Probe kann man das Gleichungssystem (3) erneut berechnen und erhält dann eine dritte Trennformel:

$$y = -18.6\ b/l + 5.58\ \mathrm{tg}\ \alpha/2 + 110.8\ \mathrm{tg}\ \beta - 0.776\ \text{Zell-Länge in } \mu\mathrm{m} \qquad (9)$$

Die Koeffizienten dieser Formel unterscheiden sich nicht mehr stark von denjenigen von Formel (8). Die Verwendung von Formel (9) macht auch keine neuen Umteilungen notwendig. Die Verteilung der y-Werte der Trennfunktion (9) sind auf Fig. 8 in der Form eines Wahrscheinlichkeitsdiagramms dargestellt. Auf der Ordinate sind die y-Werte der Trennfunktion aufgetragen, die Abszissenwerte sind „Probits". Dies sind von der Normalverteilung abgeleitete Grössen. Jeder Wahrscheinlichkeit p ist ein Probit-Wert zugeordnet. Dies ist derjenige Zahlenwert, der von einer Normalverteilung mit dem Mittelwert 5 und der Standardabweichung eins gerade mit der Wahrscheinlichkeit p unterschritten wird. Probit-Werte können einer statistischen Tabellensammlung entnommen werden (z.B. Documenta Geigy-Tabellen, Fisher &

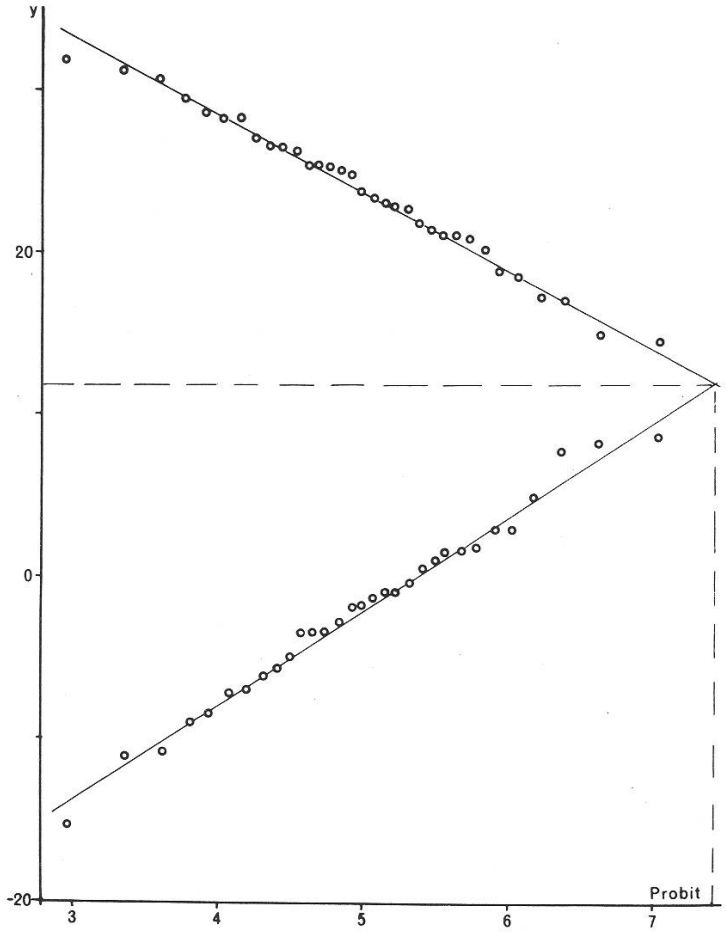

Fig. 8:
Beziehung zwischen dem Wert y der Trennfunktion und Probit-Werten (Wahrscheinlichkeitsdiagramm). Die obere Gerade entspricht den Proben von *Eurhynchium angustirete (= E. Zetterstedtii)*, die untere Gerade den Blättern von *Eurhynchium striatum* sens. strict. Der Schnittpunkt der beiden Geraden ergibt den Trennpunkt, und erlaubt, die Wahrscheinlichkeit einer Fehlklassifikation zu schätzen.

Yates oder Pearson & Hartley). Es gibt aber auch sog. Wahrscheinlichkeitspapier, bei welchem der Wahrscheinlichkeits-Massstab so verzerrt wird, dass man sich den Umweg über die Probit-Werte sparen kann.

Ein Wahrscheinlichkeitsdiagramm wird nun so hergestellt, dass zunächst die n Werte, deren Verteilung dargestellt werden sollen, nach aufsteigender (oder absteigender) Grösse geordnet werden. Dem i-ten derart geordneten Wert wird dann der Probit-Wert zu $p = (i - 1/2) / n$ zugeordnet. Stichproben aus einer normal verteilten Grundgesamtheit ergeben auf diese Weise eine Punkteschar, welche angenähert auf einer Geraden liegen.

In Fig. 8 ist die eine Gruppe nach aufsteigender Grösse und die andere Gruppe nach absteigender Grösse der y-Werte angeordnet. Diese Anordnung hat den Vorteil, dass sie eine graphische Schätzung des Trennpunkts und der Wahrscheinlichkeit einer Fehlklassifikation erlaubt (Huber, 1964). In beiden Gruppen weicht die Punkteschar nur wenig von einer Geraden ab, was zeigt, dass die y-Werte in beiden Fällen annähernd normal verteilt sind. Die Ordinate des Schnittpunkts der beiden Geraden ergibt nämlich den Trennpunkt, und die Abszisse kann zur Abschätzung der Fehlerwahrscheinlichkeit verwendet werden.

Der Schnittpunkt hat eine Ordinate von 11.0 und eine Abszisse von 7.44. Wenn man 11.0 als Trennpunkt wählt, dann werden alle Proben, welche zur oberen Gruppe gehören, und einen y-Wert kleiner als 11.0 besitzen, und alle Proben, welche zur unteren Gruppe gehören, und einen y-Wert grösser als 11.0 besitzen, falsch

klassifiziert. Wenn die y-Werte normal verteilt sind, dann ist die Wahrscheinlichkeit für das Auftreten solcher Werte gleich gross, wie die Wahrscheinlichkeit eines Wertes grösser als 7.44 bei einer Normalverteilung mit dem Mittelwert 5 und der Standardabweichung 1. Einer Tabelle der Normalverteilung kann man entnehmen, dass diese Wahrscheinlichkeit kleiner als 1% ist.

Eine Abschätzung der Wahrscheinlichkeit von Fehlklassifikationen kann auch auf rechnerischem Weg erhalten werden (siehe z.B. Lachenbruch 1967, Lachenbruch & Mickey 1968). Der Vorteil der graphischen Methode besteht darin, dass durch sie zugleich die Voraussetzung der Normalverteilung überprüft werden kann.

Die lineare Trennfunktion ist optimal, wenn die Daten aus einer multivariaten Normalverteilung stammen, und wenn die Varianzen und die Korrelationskoeffizienten in beiden Gruppen gleich gross sind. Sind diese Voraussetzungen nicht erfüllt, so ist in vielen Fällen die lineare Trennfunktion immer noch brauchbar, manchmal kann allerdings die Zahl der Fehlklassifikationen derart ansteigen, dass komplizierte Verfahren angewandt werden müssen (vgl. Krzanowski 1977). Man kann dann versuchen, die Variablen derart zu transformieren, dass die Voraussetzungen mindestens annähernd erfüllt sind, oder dann berechnet man quadratische Trennfunktionen (Gilbert 1969, Wahl & Kronmal 1977). Letzteres ist vor allem dann am Platz, wenn die Varianzen in den beiden Gruppen verschieden gross sind.

Was in unserem Beispiel durch die Berechnungen gewonnen wurde, ist einerseits der Nachweis, dass sich das Material tatsächlich in zwei morphologisch verschiedene Gruppen zerlegen lässt, und andererseits eine Vorschrift, wie die einzelnen Proben den Gruppen zuzuordnen sind, sodass nur eine kleine Zahl von Fehlklassifikationen zu erwarten ist.

Ob nun diese beiden Gruppen als Arten zu betrachten sind, ist letzten Endes eine biologische Frage, welche auch mit biologischen Methoden zu untersuchen ist. Diese Untersuchung wird aber durch die Trennfunktion sehr erleichtert, weil sie es ermöglicht, mit sicher bestimmten Proben zu arbeiten.

Zusammenfassung

Es wird eine Übersicht über die einschlägige mathematisch-statistische Literatur gegeben.

Anhand eines einfachen Beispiels wird versucht, einige Prinzipien statistischer Klassifizierungsmethoden zu erläutern.

Am Beispiel von 2 Kleinarten der Laubmoosgattung *Eurhynchium (E. striatum* (Hedw.) Schimp. und *E. angustirete* (Broth.) Koponen) wird gezeigt, wie man mit Hilfe von Trennfunktionen (Discriminant Functions) den Grad der Trennung zweier Sippen prüfen, und eine Vorschrift für die Klassifizierung erhalten kann.

Als Hilfsmittel wird ein graphisches Verfahren vorgestellt, welches erlaubt, mit geringem Rechenaufwand Trennfunktionen zu berechnen.

Summary

The use of statistical methods in taxonomy.

A short review of the bibliography is given.

A simple example is used to demonstrate some principles of statistical classification methods.

The problem of discrimination of two moss species (*Eurhynchium striatum* (Hedw.) Schimp. and *E. angustirete* (Broth.) Koponen) is used to demonstrate the application of discriminant functions.

A simple graphical method using profiles is described, which allows a quick calculation of discriminant functions.

Keywords: Numerical Taxonomy, Discriminant Funktions, *Eurhynchium*, Rank Correlation.

Literatur

Anderson T.W. 1958. Introduction to Multivariate Statistical Analysis. John Wiley & Sons, New York. 374 p.

Blackith R.E. and Reyment R.A. 1971. Multivariate Morphometrics. Academic Press, London and New York. 412 p.

Cacoullos T. (Editor) 1973. Discriminant Analysis and Applications. Academic Press, New York and London. 434 p.

Cailliez F. et Pages J.-P. 1976. Introduction à l'Analyse des Données. Société de Mathématiques Appliquées et des Sciences Humaines, Paris. 616 p.

Cochran W.G. 1961. Some Classification Problems with Multivariate Qualitative Data. Biometrics *17*, 10–32.

Cole A.J. (Editor) 1969. Numerical Taxonomy. Academic Press, London and New York. 324 p.

Cooley W.W. and Lohnes P.R. 1962. Multivariate Procedures for the Behavioral Sciences. John Wiley & Sons, New York. 211 p.

Cormack R.M. 1971. A Rewiew of Classification. Proc. Roy. Stat. Soc., Ser. A, *134*, 321–367.

Documenta Geigy 1968. Wissenschaftliche Tabellen, 7. Auflage. Basel. 798 p.

Engelman L. and Hartigan J.A. 1969. Percentage Points of a Test for Clusters. J. Amer. Stat. Assoc. *64*, 1647–1648.

Fisher R.A. 1936. The Use of Multiple Measurements in Taxonomic Problems. Ann. of Eugenics *7*, 179–188.

Fisher R.A. and Yates F. 1957. Statistical Tables for Biological, Agricultural and Medical Research. 5th Ed. Oliver & Boyd, Edinburgh. 138 p.

Gilbert E.S. 1969. The Effect of Unequal Variance-Covariance Matrices on Fisher's Linear Discriminant Function. Biometrics *25*, 505–515.

Goodall D.W. 1966. A New Similarity Index Based on Probability. Biometrics *22*, 882–907.

Gower J.C. 1971. A General Coefficient of Similarity and Some of its Properties. Biometrics *27*, 857–874.

Hartigan J.A. 1975. Clustering Algorithms. John Wiley & Sons, New York. 351 p.

Huber H. 1964. Über statistische Methoden zur Abgrenzung der Arten. Bot. Jb. *83*, 222–249.

– 1974. The Use of Profiles as Rapid Method in Multivariate Analysis. 8th International Biometric Conference, Constanţa, România (Photokopiertes Manuskript). 12 p.

Jardine N. and Sibson R. 1971. Mathematical Taxonomy. John Wiley & Sons, New York, 286 p.

Kendall M.G. 1962. Rank Correlation Methods. 3rd. Ed. Griffin & Co., London. 199 p.

Koponen T. 1964. *Eurhynchium zetterstedtii* Strömer and *E. striatum* (Hedw.) Schimp. in north-western Europe. Ann. Bot. Fenn. *1,* 250–256.
– 1967. *Eurhynchium angustirete* Kop. comb. n. (= *E. Zetterstedtii* Ström.) and its Distribution Pattern. Memor. Soc. F. Fl. Fenn. *43,* 53–59.
Kramer C.Y. 1972. A First Course in Methods of Multivariate Analysis Blacksburg, Va.
Krishnaiah P.R. (Editor) 1979. Handbook of Statistics. Vol. 2: Classification, Pattern Recognition and Reduction of Dimension. North Holland Publishing Co. (im Druck).
Krzanowski W.J. 1977. The Performance of Fisher's Linear Discriminant Function Under Non-Optimal Conditions. Technometrics *19,* 191–200.
Kullback S. 1968. Information Theory and Statistics. Dover Publications, New York. 399 p.
Lachenbruch P.A. 1967. An Almost Unbiased Method of Obtaining Confidence Intervals for the Probability of Misclassification in Discriminant Analysis. Biometrics *23,* 639–645.
– 1979. Discriminant Analysis. Biometrics *35,* 69–85.
– and Mickey M.R. 1968. Estimation of Error Rates in Discriminant Analysis. Technometrics *10,* 1–11.
Linder A. 1951. Statistische Methoden für Naturwissenschafter, Mediziner und Ingenieure. 3. Aufl. Birkhäuser, Basel 484 p.
Morrison D.F. 1967. Multivariate Statistical Methods. McGraw Hill Co., New York. 338 p.
Pearsall W.H. 1927. Growth Studies VI. On the Relative Size of Growing Plant Organs. Ann. Bot. *41,* 549.
Pearson E.S. and Hartley H.O. 1958. Biometrika Tables for Statisticians, Vol. I. 2nd. Ed. Cambridge University Press, Cambridge. 240 p.
Quenouille M.H. 1952. Associated Measurements. Butterworths Scientific Publications, London. 222 p.
Rao C.R. 1952. Advanced Statistical Methods in Biometric Research. John Wiley & Sons, New York. 463 p.
– 1965. Linear Statistical Inference and Its Applications. John Wiley & Sons, New York. 552 p.
Schüepp O. 1945. Statistische Beschreibung der Blattverzweigung bei *Aspidium filix mas* und bei *Delphinium elatum.* Jul. Klaus Stift. Erg. Bd. zu Bd. XX, 328–341.
– 1963. Mathematisches und Botanisches über die Allometrie. Verh. Natf. Ges. Basel *74,* 69–105.
Seal H.L. 1964. Multivariate Statistical Analysis for Biologists. Methuen & Co., London. 207 p.
Shapiro S.S. and Wilk M.B. 1965. An Analysis of Variance Test for Normality. Biometrika *52,* 591–611.
Sneath P.H.A. and Sokal R.R. 1973. Numerical Taxonomy. Freeman & Co., San Francisco, 573 p.
Sokal R.R. and Sneath P.H.A. 1963. Principles of Numerical Taxonomy. Freeman & Co., San Francisco. 359 p.
Störmer P. 1942. *Eurhynchium Zetterstedtii* spec; nov. and *E. striatum* in Norway. Nytt Mag. Naturv. *83,* 79–92.
– 1969. Mosses with a Western and Southern Distribution in Norway. Universitetsforlaget, Oslo. 288 p.
Tukey J.W. 1977. Exploratory Data Analysis. Addison-Wesley Publ. Co. Reading, Mass. 688 p.
Wahl P.W. and Kronmal R.A. 1977. Discriminant Functions when Covariances are Unequal and Sample Sizes are Moderate. Biometrics *33,* 479–484.

Hans Huber
Im Gehracker 2
CH-4125 Riehen