© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

Mitt. Bot. LINZ 4/2,87—106

# PRINZIPIEN EINER IMPLEMENTIERUNG+) DES PFLANZENSOZIOLOGISCHEN ZERLEGUNGSPROBLEMS

von FRANZ STOCKINGER und WOLFGANG HOLZNER, Wien

## Vorbemerkungen

Der Einsatz eines Computers zur Auswertung pflanzensoziologischer Vegetationsaufnahmen, die nach der in Mitteleuropa üblichen Methodik erstellt wurden, wird vor allem durch zwei Faktoren bestimmt:

1. Die zu handhabenden Datenmengen sind meist sehr umfangreich.
2. Die Grundbegriffe der ursprünglichen Theorie (BRAUN-BLANQUET 1964) sind insoferne nicht hinreichend scharf definiert, als Fragen innerhalb der Theorie nicht ohne weiteres mit Datenverarbeitungsanlagen bearbeitet werden können.

Die Adäquatheit der Programme, d.h. daß diese tatsächlich die g e s t e l l t e n Aufgaben lösen, wird dadurch gesichert, daß zunächst ein mathematisches Modell der gegebenen Theorie entwickelt und hinsichtlich seiner Adäquatheit zu dieser diskutiert wird. In einem zweiten Schritt werden Algorithmen angegeben, die Fragestellungen innerhalb des Modells lösen. Für die Diskussion der Adäquatheit dieser Algorithmen zum Modell stehen dann wirksame mathematische Hilfsmittel bereit. Der dritte Schritt, vom Algorithmus zu einem Maschinenprogramm, das es gestattet, diesen in endlicher auf Datenmengen großen Umfanges anzuwenden, bringt rückwirkend einschneidende Restriktionen für die Möglichkeiten zur Konstruktion des Modells:

> Es müssen die in der ursprünglichen Theorie wichtigen Fragen durch geeignete Programme in endlicher Zeit und nach Algorithmen, deren Wirkungsweise in den Begriffen des Modells exakt beschreibbar ist, beantwortet werden können. Sobald es nötig ist, Aufnahmen zwischendurch manuell und intuitiv umzusortieren (MOORE 1969), "lästige" Aufnahmen zu streichen u.dgl., so hätte die Konstruktion ihren Zweck nicht erreicht. Damit soll aber nun nicht behauptet werden, daß immer mit dem gesamten vorgegebenen Aufnahmepaket gearbeitet werden muß. Aufgliederungen nach jahreszeitlichen, geographischen und ähnlichen Gesichtspunkten können wichtige Teilresultate liefern.

---

+) Implementierung = Realisierung auf einer Datenverarbeitungsmaschine

Diese Arbeit wurde vom Fonds zur Förderung der wissenschaftlichen Forschung unterstützt.

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

Als Ausgangsmaterial betrachten wir eine hinreichend große Anzahl von Vegetationsaufnahmen, das Aufnahmenfile (oft der Kürze halber nur File genannt). Unter einer Aufnahme verstehen wir, wie üblich, eine Liste von Artnamen, denen je eine Zahl zugeordnet ist, die den Anwesenheitsgrad (Artmächtigkeit, Deckung, Gewicht, ....., im folgenden der Einfachheit halber Deckung genannt), der Art in der Aufnahmefläche angibt.

Die Gesamtheit aller Artennamen, die auf mindestens einer Aufnahme des Files erwähnt werden, heißt Artengesamtheit des Aufnahmenfiles. Gruppen sind beliebige Kombinationen von Artennamen aus dieser Artengesamtheit.

Die Deckungswerte seien in folgendem Sinn Relativzahlen: Ist eine Art in einer Aufnahmefläche stärker vertreten als in einer anderen, so erhält sie dort eine "entsprechend" höhere Deckungszahl.

Für die folgenden Auswertungen ist das hinreichend, d.h. es genügt, wenn aus den Deckungsschätzwerten hervorgeht, wie stark jede Art im Rahmen ihrer Möglichkeiten in der jeweils aufgenommenen Fläche vertreten ist. Damit ist das Problem, Deckungen verschiedener Arten vergleichen zu m ü s s e n , eliminiert und es steht für jede Art potentiell die gesamte Deckungsskala zur Verfügung. Da die Deckungsskala in natürlicher Weise mit "nichts" beginnt, geordnet ist und (Schätz-) Werte begrenzter Genauigkeit enthält, können wir für den allgemeinen Fall annehmen, daß sie die Form 0, 1, 2, ....., k hat.

Es sei hier betont, daß wir damit noch keine Aussage über die Bedeutung der H ö h e der einzelnen Stufen relativ zueinander gemacht haben, obgleich wir sie fortlaufend numerieren.

## A) Das Zerlegungsproblem der Gesellschaftssystematik

Für ein vorgegebenes Aufnahmenfile ist die Artengesamtheit in Gruppen so zu zerlegen, daß

1) die Gruppen voneinander möglichst separiert sind und
2) im einzelnen möglichst "gute Assoziationen" sind.

Für ein praktisch relevantes File (mit einigen hundert Arten) ist die Zahl der kombinatorisch möglichen Zerlegungen der Artengesamtheit in Gruppen von einer Größenordnung, die jedes vollständige Durchsuchen unmöglich macht.

Wir ermitteln daher zunächst diejenigen Gruppen, die für die weitere Konstruktion in Frage kommen. Unter den Kombinationen dieser Gruppen muß dann auch die gesuchte Zerlegung vorkommen.

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

## B) Entwicklung des Modells

### 1) Charakterisierung der essentiellen Gruppen

Im Modell werden bewertete Objekte - die "essentiellen Gruppen" - aufscheinen, die die Eigenschaft besitzen, daß eine "hinreichend große" Menge der am höchsten bewerteten essentiellen Gruppen in ihrer Gesamtheit im "wesentlichen" alle guten Assoziationen repräsentieren. Die Diskussion der Adäquatheit des Modells muß also bei der Adäquatheit der Bewertungsfunktion zur intuitiv vorgegebenen "Güte der Assoziation" beginnen. Sie muß bei einem Vergleich der nach der ursprünglichen Theorie und ihren Methoden gewonnenen Resultate mit den nach dem Modell berechneten enden. (Diese Vergleiche wurden zwar durchgeführt, sollen aber später publiziert werden, um nicht den Rahmen dieser Veröffentlichung zu sprengen.) Im Fall jeder Abweichung bei identischem Ausgangsmaterial ist auf die der ursprünglichen Theorie zugrundeliegenden Vorstellungen zu reflektieren und es sind die als inadäquat erkannten Vorgangsweisen abzuändern. Das Modell muß, um die dadurch geforderte Flexibilität aufzuweisen, daher zunächst Steuerparameter enthalten, deren Variationsbereiche im Verlaufe der die Entwicklung abschließenden Diskussion der Resultate im Idealfall auf einen jeweils festen Wert eingeengt werden können.

Unter dem Vergesellschaftungsgrad einer Gruppe ist die Idee dessen gemeint, was die Bindungskoeffizienten zu messen versuchen. Es ist nicht zweckmäßig, wie üblich eine Funktion eindeutig festzulegen, welche jeder Gruppe bezüglich eines beliebigen Files genau einen Wert zuordnet. Damit wäre u.a. die Problematik übergangen, was die einzelnen Stufen der Deckungsskala zu bedeuten haben. Es gibt aber durchsichtige Spezialfälle, bei deren Vorliegen leicht zweckmäßige Normierungen für alle zur Messung des Vergesellschaftungsgrades in Frage kommenden Funktionen erstellt werden können. Wir gelangen so zu einer Reihe von Bedingungen, durch die eine Klasse von Funktionen abgegrenzt ist. Daraus wählen wir nun die, die uns für die Implementierung am brauchbarsten erscheinen.

Es liege nun ein Aufnahmenfile vor. Die Aufnahmen sind durchnumeriert (1, 2, ..... nauf), ebenso die Artengesamtheit des Files (1, 2, ..... nart). (x, y, z sollen im Folgenden stets Gruppen, a, b, c ... Arten bezeichnen). $\bar{\bar{a}}$ bedeutet die Anzahl der Aufnahmen, auf denen die Art a mit einer von Null verschiedenen Deckung angeführt ist. d(a,1) bedeutet die Deckung der Art a auf der Aufnahme mit Nummer 1.

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

Definition:

Eine Funktion, die jeder Gruppe eine Zahl zwischen 0 und 1 zuordnet, soll g-Funktion (Geselligkeit) heißen, wenn sie die folgenden Bedingungen erfüllt:

Für z w e i e l e m e n t i g e Gruppen $x = \{a, b\}$ soll gelten:

(G1) $g(x) = 0 \leftrightarrow$ a und b schließen einander aus, d.h. für jede Aufnahme des Files gilt: ist die Dekkung von a nicht Null, so ist die Dekkung von b gleich Null und umgekehrt.

(G2) $g(x) = 1 \leftrightarrow$ Es gibt eine Zahl $\lambda \neq 0$, sodaß für jede Aufnahme des Files gilt: Deckung von a ist gleich Deckung von b mal $\lambda$. Für jede andere Art c : $g(\{a, c\}) = g(\{b, c\})$.

(G3) Falls es Zahlen $c_1$ und $c_2$ (ungleich Null) gibt, sodaß die Deckung von a stets nur $c_1$ oder Null, von b nur $c_2$ oder Null beträgt, dann soll gelten:

$$g(x) = \frac{L}{\sqrt{\bar{\bar{a}} \cdot \bar{\bar{b}}}}$$ wobei L = Anzahl der Aufnahmen,

wo sowohl a als auch b von Null verschiedene Dekkung haben.

Für Gruppen b e l i e b i g e r Elementzahl $x = \{a_1, a_2, \ldots a_n\}$ :

(G4) $g(x) = 0 \leftrightarrow$ es gibt $a_i$ und $a_j$ in x die einander ausschließen (im Sinne von (G1)).

$g(x) = 1 \leftrightarrow$ je zwei Arten aus x sind "vollständig vergesellschaftet" (im Sinne von (G2)).

(G5) Wenn es eine Teilmenge $y \subset x$ gibt mit $g(y) < g(x)$, dann gibt es mindestens eine Teilmenge $z \subset x$ mit $g(z) > g(x)$,

und "im allgemeinen" soll gelten:

ist $1 > g(x) \gg 0$, so sollen etwa gleichviele $y \subset x$ und $z \subset x$ mit $g(y) < g(x)$ und $g(z) > g(x)$ zu erwarten sein.

(G1) ist plausibel; zu (G2) ein Beispiel:

Artengesamtheit: A, B, C, D, E.

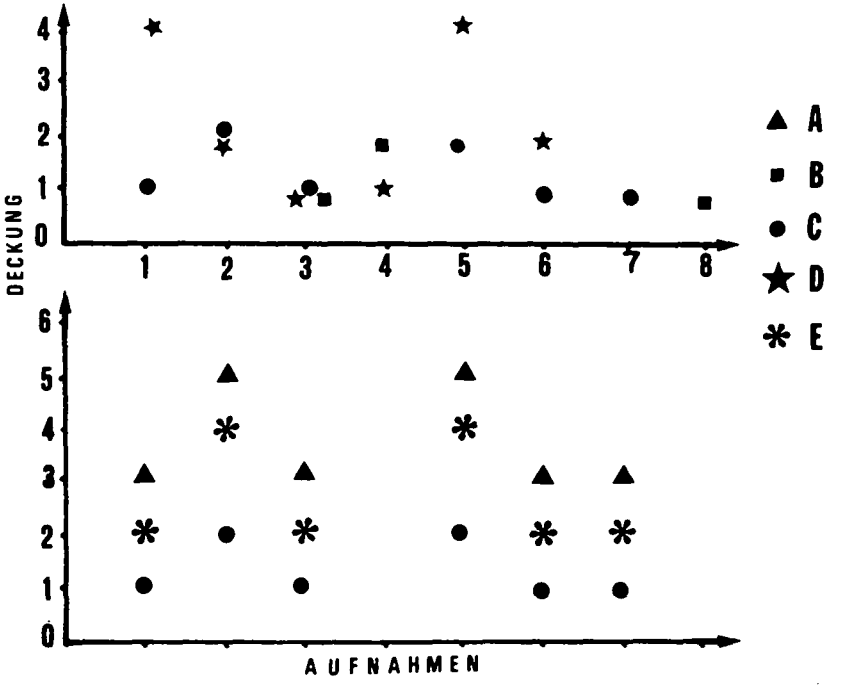
nart = 5, nauf = 8

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

**Aufnahmenfile:**

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 3 | 0 | 1 | 4 | 2 |
| 2 | 5 | 0 | 2 | 2 | 4 |
| 3 | 3 | 1 | 1 | 1 | 2 |
| 4 | 0 | 2 | 0 | 1 | 0 |
| 5 | 5 | 0 | 2 | 4 | 4 |
| 6 | 3 | 0 | 1 | 2 | 2 |
| 7 | 3 | 0 | 1 | 0 | 2 |
| 8 | 0 | 1 | 0 | 0 | 0 |

Betrachten wir eine beliebige Gruppe, z.B. {B, C, D} in einem Diagramm, das die Deckungszahlen gegen die Aufnahmenummern wiedergibt, und vergleichen wir dies mit dem für {A, C, E} :




Im Diagramm für die Gruppe {A, C, E} sind die drei Deckungswerte einander durch das ganze File hindurch proportional. In diesem Fall nimmt wegen (G2) und (G4) jede g-Funktion den Wert 1 an, vorausgesetzt, daß die Deckungsskala in dem Sinne linear ist, daß z.B.

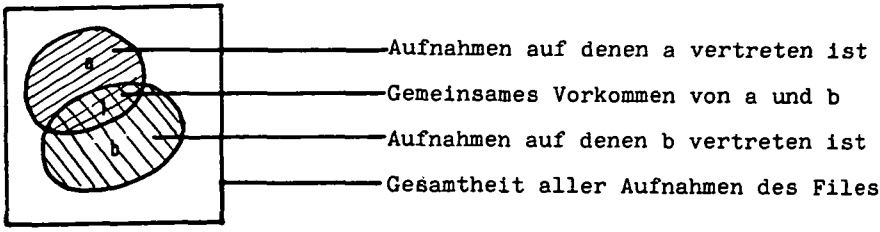
© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

eine doppelt so große Deckungsziffer einen doppelt so hohen "Anwesenheitsgrad" ausdrückt. Dann ist es sinnvoll hier von höchstmöglicher Vergesellschaftung zu sprechen. Diese Linearitätsforderung kann immer durch eine passende Gewichtung der Deckungszahlen erfüllt werden. (Über die Nichtlinearität der Braun-Blanquetschen Artmächtigkeitswerte s. BALASZ, über Gewichtung SCHMID & KUHN).

(G2) legt damit die Bedeutung der Höhe der Stufen in der Dekkungsskala fest: die fortlaufenden Deckungswerte müssen gleichhohen Abstufungen bezüglich der (jeweils) zugrundeliegenden Vorstellung vom Anwesenheitsgrad entsprechen. Beispiel: Wird der Anwesenheitsgrad definiert als "Trockengewicht der Art pro $m^2$", dann ergibt sich eine bezüglich dieses Begriffes vom Anwesenheitsgrad lineare Deckungsskala etwa so:

| | "Deckungswert" |
|---|---|
| 0 $g/m^2$ ..... | 0 |
| über 0 aber höchstens 1 g ..... | 1 |
| über 1 aber höchstens 2 g ..... | 2 |
| über 2 aber höchstens 3 g ..... | 3 |
| über n-1 aber höchstens n g ..... | n |

Um den Bereich geringen Vorkommens zu dehnen, werden häufig Maßzahlen verwendet, die Potenzen zwischen 0 und 1 von obigen Maßzahlen entsprechen. Da nun tatsächlich 1 g Unterschied im Bereich von 0 bis 1 g informationsmäßig bedeutend "mehr ausmacht" als im Bereich von 50 bis 51 g, so liegt es nahe, auch andere Vorstellungen vom Anwesenheitsgrad an ein und demselben File zu testen, z.B. die Braun-Blanquetsche Auffassung. Schließlich werden wir uns für diejenige(n) Definition(en) des Anwesenheitsgrades entscheiden, bei deren Verwendung im Endresultat (üblicherweise) die schärfsten Strukturen zu Tage treten (s.a. CORMACK). (G4) überträgt (G2) auf Gruppen beliebiger Artenzahl. (G3) regelt den Verlauf der g-Funktionen (für Zweiergruppen) im Spezialfall, daß für die Deckungen beider Arten jeweils nur "ja oder nein" gilt. Dann kann der Vergesellschaftungsgrad durch die folgende Figur veranschaulicht werden:



© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

$1/\bar{\bar{a}}$ ist dann die Wahrscheinlichkeit mit der die Art a die Art b mit sich bringt.

$1/\bar{\bar{b}}$ ist dann die Wahrscheinlichkeit mit der die Art b die Art a mit sich bringt.

(G3) setzt fest, daß jede als g-Funktion zugelassene Funktion in diesem Spezialfall gleich dem geometrischen Mittel dieser beiden Wahrscheinlichkeiten ist. Jede g-Funktion geht im Spezialfall konstanter Deckungen in den häufig verwendeten Bildungskoeffizienten

$\frac{n_{1\,2}}{\sqrt{n_1 \; n_2}}$ über. Damit ist die Lesbarkeit der Zahlenwerte gesichert.

(G4) schließt für größere Gruppen aus, daß darin unverträgliche Artenpaare unentdeckt bleiben.

(G5) sichert, daß folgende Definition für jede g-Funktion sinnvoll ist:

Eine Gruppe x heiße e s s e n t i e l l e Gruppe,
wenn 1) $g(x) > 0$ und
2) es keine Gruppe y gibt mit $y \supset x$ und $g(y) > g(x)$.

Man mache sich klar z.B. mit Hilfe der üblichen Tabellenarbeit, daß den nichtessentiellen Gruppen in der Tat bezüglich Vergesellschaftung kein Informationsgehalt zukommt. Es sind sinnlose, beliebige Kombinationen von Artennamen.

(G5) sichert, daß es stets genug essentielle Gruppen gibt. Durch diese Konstruktion werden die üblichen Vergewaltigungen vermieden, wie: Bevorzugung von Gruppen gleicher Größe (Informationstheoretische Methoden, WILLIAMS & LAMBERT); a priori Beschränkung auf disjunkte Gruppen (Cluster Analysis, WISHART, CRAWFORD & WISHART). Vor allem wird keine Treshold+) für die Abgrenzung der essentiellen Gruppen benötigt.

Unter Berücksichtigung von (G1) - (G5) und dem Wunsch nach möglichst guter Implementierbarkeit wählen wir folgende Definitionen:

Für z w e i e l e m e n t i g e Gruppen:

$$g(\{a, b\}) = \frac{\sum_k d(a, k) \cdot d(b, k)}{\sqrt{\sum_k d^2(a, k)} \; \sqrt{\sum_k d^2(b, k)}}$$

(die Summen jeweils über k = 1 ....nauf)

+) Treshold: Schwellenwert

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

Das ist der cosinus des Winkels, der von den zu a und b gehörigen nauf-spaltigen Vektoren der Deckungszahlen eingeschlossen wird. Damit sind (G1) bis (G3) erfüllt.

(G3) bedeutet, daß wir eine Art mit der Projektion ihres Dekkungsvektors auf die Hypersphäre des nauf-dimensionalen Koordinatenraumes identifizieren. Damit ist der cos des Winkels zwischen den Ortsvektoren dieser Projektionspunkte auch geometrisch ein plausibles "Maß der Nähe".

Für Gruppen b e l i e b i g e r Artenzahl:

$$g(\{a_1, \ldots, a_m\}) = \left[\prod_{1 \leq 1<k \leq m} g(a_1, a_k)\right]^{\frac{1}{\binom{m}{2}}}$$

(Fortsetzung der Funktion "Bildung des geometrischen Mittels" mit den Werten der Funktion auf den zweielementigen Teilmengen. Durch die Definition als Produkt ist (G4), durch die Wahl des Exponenten "n über 2" (das ist die Anzahl der auftretenden Faktoren) ist (G5) erfüllt.)

Für diese Funktion wurde eine leistungsfähige Implementierung auf einer IBM 360/44 PS durchgeführt (vorläufig 350 Arten auf beliebig vielen Aufnahmen).

Denken wir uns nun die Menge aller Gruppen nach ihren g-Werten geordnet und streichen alle jene Gruppen, zu denen es solche mit größerem g-Wert gibt, von denen sie umfaßt werden, so bleibt genau die Menge aller essentiellen Gruppen nach g-Werten geordnet über.

Das erste Teilproblem lautet daher: Man ermittle - von der Seite maximaler g-Werte - ein hinreichend großes Stück der nach g-Wert geordneten Menge der essentiellen Gruppen. Auch diese Teilaufgabe ist im direkten Weg, d.h. unter Berechnung der g-Funktion auf jeder kombinatorisch möglichen Gruppe, nicht lösbar. Mit Hilfe geeigneter Approximationsalgorithmen können jedoch hinreichend viele (einige hundert) Gruppen aus dem Anfangsabschnitt der genannten Ordnung erhalten werden. (Erweiterungsalgorithmus: Übergang von n-elementiger Gruppe zu jener n+1-elementigen Gruppe, die von allen - die ursprüngliche Gruppe umfassenden - Gruppen maximalen g-Wert hat; Optimierungsalgorithmus: Austausch jenes Elements in einer Gruppe gegen jenes Element außerhalb der Gruppe, sodaß der g-Wert der Gruppe dabei maximal erhöht wird.)

Ab nun sei eine repräsentative Menge von essentiellen Gruppen gegeben, und es werden Beziehungen dieser Objekte zueinander be-

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

trachtet.

## 2) Charakterisierung separierter Systeme essentieller Gruppen

Zur weiteren Bearbeitung des Zerlegungsproblems müssen wir nun präzisieren, was unter dem Separiertheitsgrad eines Systems von essentiellen Gruppen verstanden wird.

Intuitiv ist klar:

Zwei Gruppen werden genau dann stark separiert heißen, wenn sie einander höchstens bezüglich wenig treuer Arten überlappen.

Haben sie relativ viele eher treue Arten gemeinsam, so sollen sie wenig separiert heißen.

"Treu"wird dabei in dem Sinn verwendet, daß eine Art in einer Gruppe, der sie angehört, umso treuer ist, je stärker sie an die übrigen Gruppenmitglieder "gebunden" ist. Das soll nun erfaßt werden.

Ist $x$ eine Gruppe von $n$ Arten und $a$ eine beliebige Art aus der Artengemeinschaft des Files, so sei die T r e u e der Art $a$ i n der G r u p p e $x$ definiert mit:

$$t(a, x) = \begin{cases} 0 & \text{falls } a \text{ nicht in } x \\ \left[\prod_{b \in x - \{a\}} g(\{a, b\})\right]^{\frac{1}{n-1}} & \text{falls } a \in x \end{cases}$$

Falls $a$ zu $x$ gehört ist demnach die Treue das geometrische Mittel der Bindungen an die übrigen Gruppenmitglieder; sonst 0. (Bemerkung: Falls $x$ essentiell, dann ist $t(a, x) > 0$ gleichbedeutend mit $a \in x$.)

Wir präzisieren nun die oben angedeuteten Forderungen an Funktionen zur Messung des Separiertheitsgrades.

Eine Funktion, die auf jeder vorgegebenen Menge von essentiellen Gruppen einen Wert zwischen 0 und 1 annimmt, soll s-Funktion heißen, wenn für jedes System X essentieller Gruppen $x_1, \ldots, x_r$ gilt:

(S1) $s(X) = 0 \longleftrightarrow$ es gibt $x \in X$ mit $x \subseteq \bigcup_{y \in X - \{x\}} y$

(S2) $s(X) = 1 \longleftrightarrow$ für alle $x, y \in X$ gilt $x \cap y = 0$

(S3) Falls $X = \{x, y\}$ und es gilt $c_1, c_2 \neq 0$, sodaß für alle $a$ aus der Artengesamtheit des Files gilt

$$t(a, x) = \begin{cases} c_1 & \text{für } a \in x \\ 0 & \text{für } a \notin x \end{cases}, \quad t(a, y) = \begin{cases} c_2 & \text{für } a \in y \\ 0 & \text{für } a \notin y \end{cases}$$

dann soll gelten:

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

$$s(X) = \sqrt{(1 - \overline{\overline{x \cap y}} / \overline{\overline{x}})(1 - \overline{\overline{x \cap y}} / \overline{\overline{y}})}$$

($\overline{\overline{z}}$: wenn z eine Menge ist: Anzahl der Elemente von z)

(S4) Seien x, y, z Gruppen. Es gelte

$y \cap x \geqq \{a_1 \ldots a_t\}$

$z \cap x \subseteqq \{b_1 \ldots b_t\}$

und die Numerierung der a und b kann so gewählt werden, daß für alle $i = 1, \ldots, t$ gilt:

$t(a_i, x) \geqq t(b_i, x)$ und $t(a_i, y) \geqq t(b_i, z)$

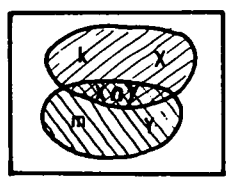
und jedem $a \in y$ sei eindeutig ein $b \in z$ mit $t(a, y) < t(b, z)$ zugeordnet; dann muß $s(x, y) < s(y, x)$ erfüllt sein.

(S5) Wenn es ein Teilsystem $Y \subset X$ gibt mit $s(Y) < s(X)$, dann gibt es mindestens ein Teilsystem $Z \subset X$ mit $s(Z) > s(X)$. Im "allgemeinen" soll gelten:
Ist $1 > s(X) \gg 0$, so sollen etwa gleichviele Teilsysteme $Y \subset X$ und $Z \subset X$ mit $s(Y) < s(X)$ und $s(Z) > s(X)$ zu erwarten sein.

Die Bedingungen (S1) und (S2) sind plausibel: Am günstigsten sind Zerlegungen in elementfremde Mengen; jede Gruppe aus der Zerlegung muß irgendwelche Arten enthalten, die in keiner anderen Gruppe der Zerlegung vorkommen.

(S3) bestimmt den Verlauf der Separation in dem Spezialfall, daß innerhalb von zwei Gruppen nur gleichtreue (=gleichwertige) Arten vorkommen. Dann soll die Separation der beiden Gruppen voneinander gleich dem geometrischen Mittel der Teilungsverhältnisse

$k/(k + 1)$ und $m/(m + 1)$ sein.




$k = \overline{\overline{x - y}}$

$l = \overline{\overline{x \cap y}}$

$m = \overline{\overline{y - x}}$

Artengesamtheit des Files

Dadurch erhält s (analog zu g; man vergleiche die Teilungsverhältnisse) eine anschauliche Bedeutung.

(S4) erfaßt qualitativ, daß die Separation geringer wird, wenn

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

die Gruppen treuere Arten gemeinsam haben.

Indessen wurde noch nichts darüber ausgesagt, wieweit weniger treue Arten gegenüber treueren bei der Berechnung des S-Wertes q u a n t i t a t i v vernachlässigt werden sollen.

Wir eliminieren dieses Problem wieder aus den allgemeinen Überlegungen durch eine Steuergröße, mit der wir dann in der Praxis jeden gewünschten Grad der Vernachlässigung der weniger treuen Arten einstellen können. Die günstigsten Werte für diesen Parameter werden dann diejenigen sein, bei deren Verwendung im a l l g e m e i n e n die Strukturen am deutlichsten aufscheinen. Dies, sowie (S1) - (S4) und leichte Implementierbarkeit im Auge definieren wir für <u>Paare</u> $X = \{x, y\}$ von essentiellen Gruppen:

$$s_q(X) = \left[ \frac{\sum_{a \in x - y} t^q(a, x) \cdot \sum_{a \in y - x} t^q(a, y)}{\sum_{a} t^q(a, x) \cdot \sum_{a} t^q(a, y)} \right]^{\frac{1}{2}}$$

und für Systeme b e l i e b i g e r Gruppenanzahl:

$( X = \{x_1, \ldots x_r\} )$

$$s_q(X) = \begin{cases} 0 & \text{wenn es ein } x \in X \text{ gibt mit } x \subseteq \bigcup_{y \in X - \{x\}} y \\ \left[ \prod_{1 \leq u < v \leq r} s_q(\{x_u, x_v\}) \right]^{\frac{1}{\binom{r}{2}}} & \text{sonst.} \end{cases}$$

Es läßt sich leicht nachrechnen, daß (S1) - (S5) durch $s_q$ für jedes $q > 0$ erfüllt sind. $q$ ist ein Steuerparameter der oben besprochenen Art.

Für $q$ nahe bei $0$ (z.B. $q = 0{,}1$) hängt $s_q(X)$ im allgemeinen nur davon ab, ob die Gruppen aus $X$ paarweise elementfremd sind ($s = 1$) oder ob häufig nichtleere Durchschnitte vorkommen ($s \to 0{,}0$). Bei großem $q$ (z.B. $q = 10$) ist in jeder Gruppe praktisch nur die treueste Art entscheidend. Der logisch dazwischen liegende Bereich wird durch passendes $q$ stufenlos erfaßt. Ab nun wird für $q$ ein fester Wert, z.B. $q = 2{,}0$ angenommen und statt $s_q$ wieder einfach $s$ geschrieben.

$s$ ordnet die Menge aller Systeme von essentiellen Gruppen linear. Wegen (S5) ist es daher sinnvoll e s s e n t i e l l e S y s t e m e in Analogie zu den essentiellen Gruppen zu definieren:

Ein System $X$ von essentiellen Gruppen heißt essentiell,

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

wenn es kein System $Y$ aus essentiellen Gruppen gibt mit $Y \supset X$ und $s(Y) > s(X)$.

Für die gesuchte Zerlegung kommt höchstens ein essentielles System in Frage: denn zu jedem nichtessentiellen gibt es laut Definition schließlich mindestens ein essentielles System, das mehr Gruppen enthält, und dennoch im Ganzen besser separiert ist, wodurch das nichtessentielle überflüssig wird.

Durch ähnliche Approximationsschemata wie zur Gewinnung der essentiellen Gruppen können wir uns nun einen Überblick über die essentiellen Systeme verschaffen. Damit erhalten wir explizite Lösungen der eingangs gestellten Aufgabe; laufen aber Gefahr, unter einem Berg von Lösungen begraben zu werden, die in einer gewissen Hinsicht trivial sind.

## C) Relativierung der g- und s-Funktion

Beispiel: Angenommen die Art $a$ ist auf fast allen Aufnahmen des Files mit der Deckung $c_1$ vertreten (sonst 0), die Art $b$ sei ebenfalls sehr häufig mit $c_2$ (sonst 0). Dann trifft die Hypothese von (G3) zu, es muß also

$$g(\{a, b\}) = \frac{1}{\sqrt{\bar{\bar{a}} \cdot \bar{\bar{b}}}}$$

(Bezeichnungen wie bei (G3))

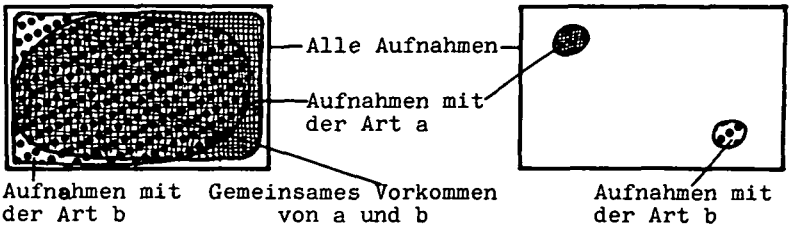
gelten. Da sowohl $a$ als auch $b$ sehr häufig sind, ($\bar{\bar{a}} \sim$ nauf, $\bar{\bar{b}} \sim$ nauf) so wird auch für 1, die Häufigkeit des gemeinsamen Vorkommens von $a$ und $b$, $1 \sim$ nauf gelten. Somit muß in diesem Fall "trivialerweise"

$$g(\{a, b\}) \sim \frac{\text{nauf}}{\sqrt{\text{nauf} \cdot \text{nauf}}} = 1$$

sein.

Paar häufiger Arten in beliebiger Lage:

Paar im File seltener Arten:



© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

Gruppen, die solche Paare von im File sehr häufigen Arten enthalten, werden daher relativ hohe g-Werte erhalten, und soferne Paare häufiger Arten im File möglich sind, werden diese in dem vom Approximationsverfahren gelieferten Abschnitt der g-Ordnung der essentiellen Gruppen deshalb "trivialerweise" dominieren. Ähnlich werden sich triviale Systeme, die aus kleinen Gruppen bestehen, verhalten.

Die Messungen mittels dieser Funktionen sind also im konkreten Fall von einem Effekt überlagert, dessen Stärke vom Umfang der Meßobjekte innerhalb des zugehörigen "Datenuniversums" abhängt, und der die g-Angaben entwertet und die s-Angaben verschärft, sobald das Datenuniversum kleiner wird (Weglassen unbeteiligter Aufnahmen bzw. Arten).

g- und s-Funktion sind reine Präsenz - Copräsenzmaße (siehe auch CORMACK). Im Sinn der vorgegebenen Theorie kommt aber auch der Coabsenz eine Aussage zu. Wie stark diese zu berücksichtigen ist, hängt vom Gesichtspunkt ab, unter dem das vorliegende File verarbeitet wird (s.u.). Mangelnde Coabsenz wird also abhängig vom Bearbeitungsaspekt an der genannten Trivialisierung der "absolut" gemessenen g- und s-Werte beteiligt sein.

Wir haben, um diesen Effekt aufzuheben, die g- bzw. s-Funktion auf die jeweilige Größe des Datenuniversums zu relativieren.

Eine Meßgröße für Nichttrivialität (im obigen Sinne) der g-Messung (an der Gruppe $\{a, b\}$ ) ist

$$k = \frac{\max(0,\ g(\{a, b\}) - E\,g(\{a, b\}))}{g(\{a, b\}) + E\,g(\{a, b\})}$$

$E\,g(\{a, b\})$ bezeichnet den Erwartungswert von $g$ auf $\{a, b\}$, wenn alle Deckungszahlen von $a$ und $b$ gegeben sind, deren paarweise Zuordnung aber nicht bekannt ist.

Im Beispiel wäre $Eg \sim 1$, also $k \sim 0$.

Der andere Grenzfall ist $Eg \sim 0$ (hohe Coabsenz). Dann gilt, falls $1 \geqq g \gg 0$ wegen

$$k = \frac{\max(0,\ g - Eg)}{g + Eg} \qquad \text{und} \qquad \begin{array}{l} g - Eg \sim g - 0 = g > 0 \\ g + Eg \sim g + 0 = g > 0 \end{array}$$

also $$k \sim \frac{g}{g} = 1 .$$

Eine solche Größe $k$ mit $0 \leqq k \leqq 1$,

$k = 0$ für $g \leqq Eg$

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

$k \sim 1$ für $g > 0$ und $Eg \sim 0$,
die den Trivialitätsgrad der g-Messung relativ zur Aufnahmenzahl mißt,

nennen wir R e l e v a n z k o e f f i z i e n t für g (analog für s). Vergegenwärtigen wir uns die Bedeutung der durch die g- (bzw. s-) Funktion erzeugten Ordnung: den Objekten mit größeren Meßwerten wurde höhere Relevanz zugeschrieben. In dem Maße, in dem der Meßwert aber trivial wird, verringert sich diese Relevanz, das heißt der Wert der g- (bzw. s-) Funktion ist in Abhängigkeit vom zugehörigen Relevanzkoeffizienten durch einen Wert g' (bzw. s') zu ersetzen, für den zunächst gilt:

1) wenn $k = 0$ dann ist $g' = 0$, unabhängig von g.
2) wenn $k \sim 1$ dann ist $g' \sim g$
3) g' wächst im gleichen Sinne wie g und k.

ad 1) $k = 0$ heißt $g \leqq Eg$, es liegt keine Aussage vor.
ad 2) falls $g \gg Eg$ ist nichts zu korrigieren.
ad 3) folgt aus der festgelegten Bedeutung der Richtungen von g', g und k.

Für die Aufgabe, g' als Funktion von g und k darzustellen, kann es keine allgemeingültige eindeutige Lösung geben, da der genaue Verlauf des zu kompensierenden Effektes nicht ein für alle Mal festgelegt werden kann: ob und inwieweit eine Gruppe mit häufigen Arten unterdrückt werden soll, das hängt von der Auffassung über die Lage des bearbeiteten Files in einem größeren ab: Man könnte etwa argumentieren, daß manche Gruppe von häufigen Arten mit (deshalb trivialem) hohem g-Wert, - eben weil schon in diesem Datenuniversum viele Arten inzidieren -, nach Hinzunahme vieler neuer Aufnahmen, auf denen die Arten dieser Gruppe selten sind, den hohen g-Wert behalten werden, dann aber aus mangelnder Coabsenz nicht mehr trivial sind; daß mit dem vorliegenden File gewissermaßen der Verbreitungsbereich besagter Artengruppe bevorzugt erfaßt wurde, was ja auch die Wichtigkeit der Gruppe als Soziologische Enteität keinen Einfluß hätte.

Diese Problematik ist angesichts der hohen Leistungsfähigkeit der zugehörigen Computerprogramme durchaus gegeben, da es möglich ist, lokale Files mit gleicher Artengesamtheit praktisch unbegrenzt zu globalen zusammenzufassen.

Der Grad der Korrektur von g zu g' wird also nicht nur von k abhängen, sondern mindestens noch von einer weiteren Größe, die angibt, wie sehr das vorliegende Aufnahmenfile als Teil eines

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

größeren gedacht wird, d.h. wie stark die Gruppen häufiger Arten "entwertet" werden sollen.

Wir verwenden den einfachsten Ansatz der das alles (incl. 1) 2) 3)) leistet:

$$g' = g^{(1 - l)} \cdot k^{l} \qquad 0 \leq l < 1$$

l spielt die Rolle der besprochenen Steuerkonstante.

Im besonderen gilt für

| | | | |
|---|---|---|---|
| $l = 0$ | : | $g' = g$ | Gruppen häufiger Arten dominieren. |
| $l = 1/2$ | : | $g' = \sqrt{g \cdot k}$ | im Bereich der "optimalen" Aufschlüsselung durch Gruppen mit Arten mittlerer Häufigkeit. |
| $l \sim 1$ | : | $g' \sim k$ | Es kommt nur mehr darauf an, wie weit die Messung die Erwartung übertrifft; (statistisch) unerwartete Gruppen seltener Arten dominieren. |

Wegen der Verlagerung des Blickpunktes auf die selteneren Arten bei Erhöhung von l nennen wir diesen Steuerparameter "LUPE".

Analoge Überlegungen gelten für s.

Wir setzen

$$s' = s^{(1 - p)} \cdot k_s^{p}$$

wobei

$$k_s = \left[ \frac{\max(0,\ s^2 - E(s^2))}{s^2 + E(s^2)} \right]^{\frac{1}{2}}$$

Es werden nun die vollständigen Formeln für g' und s' ermittelt. Der Relevanzkoeffizient geht nur in die Definition der g' bzw. s' Werte für Paare explizit ein, und diese Funktion wird dann in üblicher Weise durch Bildung des geometrischen Mittels auf den gesamten Definitionsbereich erweitert.

Damit bleiben für s' und g' die für die Definierbarkeit der essentiellen Systeme bzw. Gruppen benötigten Eigenschaften (S5) und (G5) erhalten, ebenso im Wesentlichen (S1) und (G1) wie (S3) und (G3). In diese Bedingungen sind die Relevanzkoeffizienten formal aufzunehmen.

Statt (G2) und (S2) gelten für die relativierten Funktionen g' und s' :

(G2') $g(\{a, b\}) = 1^{1 - l} \cdot k^{l} = k^{l} \leftrightarrow$ es gibt ein $\lambda$ mit $d(a, j) = \lambda . d(b, j)$ für alle Aufnahmen $j = 1 \ldots$ nauf

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

und analog

(S2') $s'(X) = k_s^p \Longleftrightarrow$ für je zwei Gruppen $x$ und $y$ aus $X$ gilt $x \cap y = 0$.

D.h. mittels $g'$ kann auch zwischen Gruppen mit $g = 1$ unterschieden werden, $s'$ unterscheidet auch im Bereich völlig separierter Systeme. Da in diesem Fall nach dem Relevanzkoeffizienten geordnet wird, bedeutet das eine Ausblendung der trivialen Systeme aus kleinen Gruppen wegen des hohen Erwartungswertes für $s$ und dadurch verringertem $k$.

Die gesuchten Zerlegungen gelangen so in den vom Approximationsverfahren gelieferten ersten Teil der Ordnung der essentiellen Systeme.

Es sind nun $Eg$ und $Es^2$ explizit zu formulieren.
Wegen der Definition von $g(\{a, b\})$ (s.o.) gilt zunächst

$$Eg(\{a, b\}) = \frac{E \sum_k d(a, k) \cdot d(b, k)}{\sqrt{\sum_k d^2(a, k)} \cdot \sqrt{\sum_k d^2(b, k)}},$$

da der Nenner nicht von der paarweisen Zuordnung der $d(a, k)$ zu den $d(b, j)$ mit $j = k$ abhängt. Wir schreiben für $d(a, k)$ nun kurz $a_k$, $n$ für nauf.

Bei gegebenen $a_1 \ldots a_n$ und $b_1 \ldots b_n$ gibt es $n!$ Zuordnungsmöglichkeiten der Form

$$\begin{matrix} a_1 & \ldots\ldots & a_n \\ b_{\pi(1)} & \ldots & b_{\pi(n)} \end{matrix}$$

wobei $\pi$ eine Permutation der Zahlen $1 \ldots n$ ist. $\mathfrak{S}_n$ bezeichne die Menge aller Permutationen von $n$ Zahlen.
Offenbar gilt

$$E \sum_{i=1}^{n} a_i b_i = \frac{1}{n!} \sum_{\pi \in \mathfrak{S}_n} \sum_{i=1}^{n} a_i b_{\pi(i)}$$

Rechts steht eine Summe über Produkte der Gestalt $a_i \cdot b_j$ $i = 1 \ldots n$, $j = 1 \ldots n$, von denen jedes $(n - 1)!$ mal auftritt. (Man überlege sich das z.B. an Hand von $a_1 b_1$ : dieses Produkt erscheint bei jeder Permutation höchstens einmal, und es erscheint bei allen Permutationen $\pi \in \mathfrak{S}_n$ mit $\pi(1) = 1$ und nur bei diesen.

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

Das sind aber genau alle Permutationen aus $\sigma_{n-1}$, also genau $(n-1)!$. Aus Symmetriegründen läßt sich dies für beliebiges $i$ und $j$ wiederholen.)
Also gilt

$$E \sum_{i=1}^{n} a_i b_i = \frac{(n-1)!}{n!} \cdot \sum_{\substack{i=1 \dots n \\ j=1 \dots n}} a_i b_j$$

$$= \frac{1}{n} \sum_{i=1}^{n} a_i \cdot \sum_{i=1}^{n} b_i \quad .$$

Somit wird

$$Eg(\{a, b\}) = \frac{1}{nauf} \cdot \frac{\sum_k d(a, k) \cdot \sum_k d(b, k)}{\sqrt{\sum_k d^2(a, k)} \sqrt{\sum_k d^2(b, k)}}$$

In dieser Form kann $Eg$ implementiert werden.

Wir bestimmen nun $E(s^2)$. $a_x$ bedeute im Folgenden $t^q(a, x)$. Alle $a_x$ und $b_y$ sind gegeben, es sei aber nicht bekannt, welche der Arten $a$ in $y$ und welche der Arten $b$ in $x$ liegen; d.h. in welcher Weise die $a$ mit den $b$ paarweise identisch sind.

Wegen der Unabhängigkeit des Nenners (in der Definitionsformel für $s$) von der Wahl dieser Zuordnungen, genügt es

$$E \left[ \sum_{a \in x - y} a_x \cdot \sum_{b \in y - x} b_y \right] = E \sum_{\substack{a \in Cy \\ b \in Cx}} a_x b_y$$

(Cx: Komplement von $x$ bezüglich der Artengesamtheit des Files.)

zu bestimmen.
(Zur Gültigkeit obiger Gleichung ist zu beachten, daß lt. Definition $a_x = 0$ wenn $a \notin x$.)

Wir haben nun den Erwartungswert einer Teilsumme von einer Summe zu bestimmen, von der alle Summanden bekannt sind; ebenso ist die Wahrscheinlichkeit verfügbar, mit der ein Summand in der Teilsumme auftritt.

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

Darum setzen wir

$$E \sum_{\substack{a \in Cy \\ b \in Cx}} a_x b_y \qquad \sum_{a,\, b} \left[ a_x b_x \cdot w(a \in Cy \text{ und } b \in Cx) \right] =$$

$$= (1 - \bar{\bar{y}}/nart) \cdot (1 - \bar{\bar{x}}/nart) \sum_{a,\, b} a_x b_y \, .$$

Insgesamt also

$$E(s^2) = \frac{(1 - \bar{\bar{y}}/nart).(1 - \bar{\bar{x}}/nart) \cdot \cancel{\sum_a a_x} \cdot \cancel{\sum_b b_y}}{\cancel{\sum_a a_x} \cdot \cancel{\sum_b b_y}}$$

$$E(s^2) = (1 - \bar{\bar{y}}/nart).(1 - \bar{\bar{x}}/nart).$$

Damit sind alle zur Messung der s'- und g'-Werte benötigten Funktionen in implementierbarer Form dargestellt.

Die Implementierung der g'-Funktion erfolgt so, daß zuerst die g'-Werte für alle Paare gespeichert werden. Dazu genügt es, das Aufnahmenfile einmal zu lesen.

Durch die Definition der t- und g'=Werte auf den größeren Gruppen als Produkt genügt es innerhalb der Approximationsalgorithmen einfache Summen zu betrachten, aus denen meist noch konstante Teilsummen abgespalten werden können, womit durch eine entscheidende Ersparnis an Maschinenschritten und somit an Rechenzeit die Durchführung der gestellten Aufgabe erst ermöglicht wird. Analoges gilt für die Implementierung der s'-Funktion, die sich einstweilen noch im Codierungsstadium befindet. Die Implementierung der s'-Funktion und der zugehörigen Approximationsalgorithmen ist durch das Programm "IFFINS" im Wesentlichen abgeschlossen.

## Zusammenfassung

Es wird ein tresholdfreies+), mathematisches Modell der Braun-Blanquetschen Pflanzensoziologie entwickelt mit Rücksicht auf die Entscheidbarkeit der Frage nach einer Zerlegung der Artengrundgesamtheit eines Aufnahmenfiles in eine Anzahl optimal separierter möglichst guter Assoziationen mit Hilfe von Datenverarveitungsanlagen. Ausgehend vom Begriff der Bindung wird zunächst die Menge der für solche Zerlegungen in Frage kommenden Artengruppierungen

+) Schwellenwert frei

© Biologiezentrum Linz/Austria; download unter www.biologiezentrum.at

(die Menge der essentiellen Gruppen) charakterisiert.

Dann wird jene Teilmenge aus der Menge aller Systeme von essentiellen Gruppen charakterisiert, in welcher die gesuchten optimalen Zerlegungen zu erwarten sind.

Ziel der Modellkonstruktion ist es, für Fragen im Rahmen der Theorie (BRAUN-BLANQUET, 1964) eine Grundlage für leistungsfähige, adäquate Datenverarbeitungsmethoden zu legen.

Literaturverzeichnis:

BALASZ, F. (1960): A gyepek botanikai és gazdasági értékelése (Die botanische und wirtschaftliche Bewertung der Rasen). A Keszth. Mez. Akad. Kiadv. 8, 28 S.

BRAUN-BLANQUET, J. (1964): Pflanzensoziologie. Wien - New York, 865 S.

CORMACK, R. M. (1971): A Review of Classification. J. Roy. Statist. Soc., 33, 321 - 367.

CRAWFORD, R. M. M. & WISHART, D. (1967): A rapid multivariate method for the detection and classification of groups of ecologically related species. J. Ecol., 55, 505 - 524.

-- " -- (1968): A rapid classification and ordination method and its application to vegetation mapping. J. Ecol., 56, 385 - 404.

MOORE, J. J. (1969): Zur mathematischen Bestätigung der tabellarischen Abgrenzung der Pflanzengesellschaften. Vortrag a. d. 14. Intern. Symposium d. Intern. Ver. f. Vegkde., Rinteln 1969.

SCHMID, P. u. N. KUHN, (1970): Automatische Ordination von Vegetationsaufnahmen in pflanzensoziologischen Tabellen. Die Naturwissenschaften 57, 462.

STOCKINGER, F. u. HOLZNER, W. (1970): Rationelle Methode zur Auswertung pflanzensoziologischer Aufnahmen mittels Elektronenrechner. Vortrag b. 14. Intern. Symp. d. Intern. Ver. f. Vegkde, Rinteln.

WILLIAMS, W. T. & J. M. LAMBERT (1959): Multivariate methods in plant ecology. I. Association analysis in plant communities. J. Ecol., 47, 83 - 101.

WISHART, D. (1969): Numerical classification method for deriving natural classes. Nature, 221, 97 - 98.

- " - (1969): Clustan IA User Manual. (Computerprogramm, auf Band vervielfältigt.) Edition 1, 1969; Corrected 1970; St. Andrews, Scotland.

Anschrift der Verfasser: Dr. WOLFGANG HOLZNER, FRANZ STOCKINGER, Lehrkanzel für Ökologie und Soziologie der Pflanzen, Hochschule für Bodenkultur, Gregor Mendel-Straße 33, A-1180 Wien

# ZOBODAT - www.zobodat.at

Zoologisch-Botanische Datenbank/Zoological-Botanical Database

Digitale Literatur/Digital Literature

Zeitschrift/Journal: Linzer biologische Beiträge

Jahr/Year: 1972

Band/Volume: 0004_1_2

Autor(en)/Author(s): Stockinger F., Holzner Wolfgang

Artikel/Article: Prinzipien einer Implementierung des pflanzensoziologischen Zerlegungsproblems. 87-106